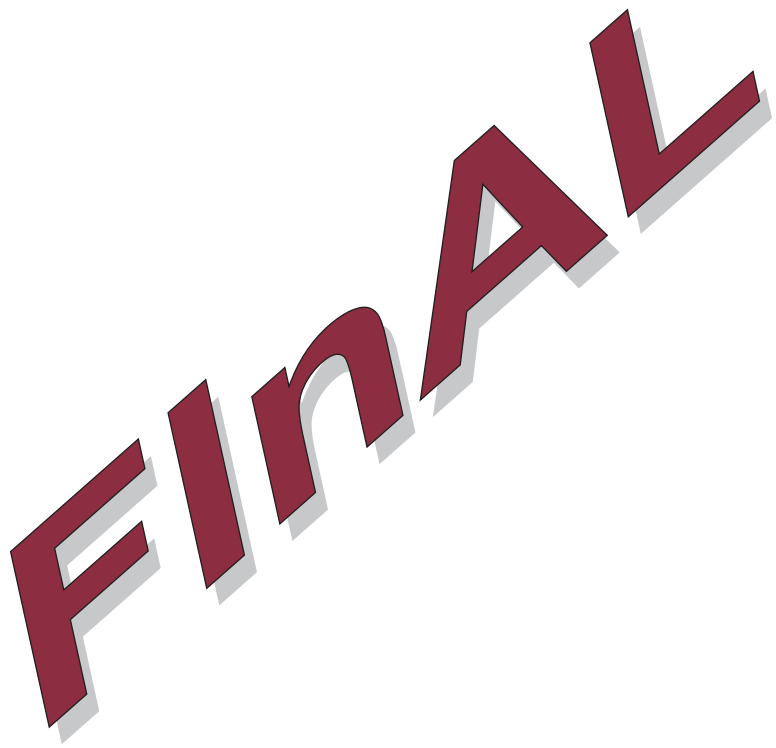
23. Jahrgang, Heft 1, Oktober 2013, ISSN 0939-8821



Mathematik für Wirtschaftsinformatik

Ulrich Hoffmann

Technical Reports and Working Papers
Leuphana Universität Lüneburg

Hrsg. der Schriftreihe FINAL: Ulrich Hoffmann
Scharnhorststraße 1, D-21335 Lüneburg

Mathematik für Wirtschaftsinformatik

Prof. Dr. rer. nat. Ulrich Hoffmann

Überarbeitete Fassung Oktober 2013

Das vorliegende Skript dient als begleitende Unterlage für die Veranstaltungen *Mathematik für Wirtschaftsinformatik* und weiterer Veranstaltungen an der Leuphana Universität Lüneburg mit inhaltlichem Bezug zur Mathematik für Wirtschaftsinformatiker und Informatiker. Die Durcharbeitung des Skripts ersetzt nicht den Besuch der Veranstaltungen, da dort zusätzlich Zusammenhänge, ergänzende Sachverhalte und im Skript nicht ausgeführte Inhalte behandelt werden.

Dieses Skript ist eine überarbeitete Fassung der Veröffentlichung (mit Copyright):
Hoffmann, U: **Mathematik für Wirtschaftsinformatiker und Informatiker & Übungen**, FINAL 17:2, 2007, ISSN 0939-8821.
Gegenüber dieser Auflage wurden einige Fehler berichtigt und Beweisideen ergänzt. Die Kapitel zur Einführung der Integralrechnung und zur Wahrscheinlichkeitstheorie und Statistik sind neu hinzugekommen.

Vorwort

Das vorliegende Skript dient als Unterlage für Veranstaltungen zur Mathematik für Wirtschaftsinformatiker und Informatiker in den ersten Studiensemestern an der Leuphana Universität Lüneburg. Es werden Schulkenntnisse der Mathematik und der Wille, sich mit mehr oder weniger abstrakten Sachverhalten auseinandersetzen zu wollen, vorausgesetzt.

Die Themenauswahl erfolgte mit Blick auf die Anwendungen in der Informatik und der Wirtschaftsinformatik. Schwerpunkte dieser Anwendungen sind die in späteren Semestern behandelten Analyseverfahren in der Theorie der Datenstrukturen und Algorithmen, der Theoretischen Informatik, insbesondere der Angewandten Komplexitätstheorie und der Kryptologie. Dem für die Vorgehensweise und Argumentationstechnik in der Informatik besonders wichtigen Prinzip der vollständigen Induktion wird ein eigenes Kapitel gewidmet und in späteren Kapiteln an mehreren auch nichttrivialen Beispielen dargestellt. Darüber hinaus werden Grundlagen vermittelt, die dem Verständnis wirtschaftswissenschaftlicher Zusammenhänge dienen. So wurden Themen aus verschiedenen Gebieten der Mathematik ausgewählt, deren Inhalte im späteren Studium von Belang sind: aus der elementaren Zahlentheorie, der Kombinatorik, der Analysis, der Linearen Algebra und der Wahrscheinlichkeitstheorie und der Statistik. Übungsaufgaben mit Lösungen zu den einzelnen Kapiteln ergänzen das Skript in einem gesonderten Text.

Die Durcharbeitung des Skripts ersetzt nicht den Besuch der Veranstaltungen zur Mathematik, da dort zusätzlich wichtige Zusammenhänge, Beispiele und mathematische Beweise, die dem Verständnis der mathematischen Sätze dienen, erläutert und ergänzende Sachverhalte behandelt werden. Das Skript verzichtet gelegentlich auf die Darstellung der mathematischen Beweise, wenn sie über den Rahmen des Textes hinausgehen. Ansonsten werden die in den Sätzen dargestellten Sachverhalte auch formal begründet.

Der Leser wird schnell feststellen, dass sich der Schwierigkeitsgrad bei der Erarbeitung der einzelnen Themen im Verlauf des Textes erhöht. Dieses vielen mathematischen Texten inhärente Problem sollte nicht davor abschrecken, sich weiter und tiefergehend mit mathematischen Sachverhalten zu beschäftigen. Vielmehr sollte es Anreiz sein, sich mit Geduld und Durchhaltewillen auch schwierige Inhalte anzueignen. Der intellektuelle Gewinn sollte dabei nicht unterschätzt werden.

Literaturauswahl zur begleitenden Lektüre

Aigner, M.: **Diskrete Mathematik**, 6. Aufl., Vieweg+Teubner, 2006.

Bartholomé, A.; Rung, J.; Kern, H.: **Zahlentheorie für Einsteiger**, 7. Aufl.,Vieweg+Teubner, 2010.

Beutelspacher, A.: **Lineare Algebra**, 7. Aufl.,Vieweg+Teubner, 2009.

Beutelspacher, A.; Neumann, H.B.; Schwarzpaul, T.: **Kryptografie in Theorie und Praxis**, 2. Aufl., Vieweg+Teubner, 2010.

Bornemann, F.: **Konkrete Analysis**, Springer, 2008.

Fisz, M.: **Wahrscheinlichkeitsrechnung und mathematische Statistik**, 11. Aufl., Deutscher Verlag der Wissenschaften, 1989.

Hachenberger, D.: **Mathematik für Informatiker**, 2. Aufl., Pearson Studium, 2008.

Haggarty, R.: **Diskrete Mathematik für Informatiker**, Pearson Studium, 2004.

Hartmann, P.: **Mathematik für Informatiker**, 5. Aufl., Vieweg+Teubner, 2012.

Henze, N.: **Stochastik für Einsteiger**, 9. Aufl., Vieweg+Teubner, 2012.

Meinel, C.; Mundhenk, M.: **Mathematische Grundlagen der Informatik**, 5. Aufl., Vieweg+Teubner, 2011.

Purkert, W.: **Brückenkurs Mathematik für Wirtschaftswissenschafter**, 7. Aufl., Vieweg+Teubner, 2011.

Schira, J.: **Statistische Methoden der VWL und BWL**, 4. Aufl., Pearson Studium, 2012.

Sydsæter, K.; Hammond, P.: **Mathematik für Wirtschaftswissenschaftler**, 3. Aufl., Pearson Studium, 2009.

Sydsæter, K.; Hammond, P.; Seierstad, A.; Strøm, A.: **Further Mathematics for Economic Analysis**, Pearson Education, 2005.

Zwerenz, K.: **Statistik verstehen mit Excel**, 2. Aufl., Oldenbourg, 2008.

Weiterführende mathematische Werke:

Graham, R.L.; Knuth, D.E.; Patashnik, O.: **Concrete Mathematics**, 2. Aufl., Addison-Wesley, 2001.

Maurer, S.B.; Ralston, A.: **Discrete Algorithmic Mathematics**, 3. Aufl., Addison-Wesley, 2004.

Yan, S.Y.: **Number Theory for Computing**, 2. Aufl., Springer, 2010.

Inhaltsverzeichnis

1 Grundlegende Definitionen, Bezeichnungen und Sachverhalte

In diesem Kapitel werden grundlegende Definitionen, Bezeichnungen und Regeln aus verschiedenen Gebieten der Mathematik zusammengestellt. Dabei wird eine gewisse Vertrautheit mit der Symbolik der Mathematik vorausgesetzt. Exemplarisch für die axiomatische Einführung einer mathematischen Grundstruktur wird der systematische Aufbau des Zahlensystems von den natürlichen bis zu den komplexen Zahlen beschrieben.

Die Mathematik begründet ihre Theorien formal jeweils durch ein System von **Axiomen**, d.h. Grundaussagen, die in einem Teil der mathematischen Welt als Basisbausteine dienen, um aus ihnen Aussagen und Erkenntnisse über diesen Teil der mathematischen Welt abzuleiten. Der Ableitungsvorgang wird durch definierte **logische Schlussregeln** gesteuert. So wurde versucht, den Aufbau der gesamten Mathematik, insbesondere den Aufbau des Zahlensystems und der Mengenlehre, streng axiomatisch zu begründen. Die Entwicklung entsprechender Axiomensysteme bzw. die Erkenntnis über Grenzen der Möglichkeiten dieses Ansatzes können als überragende Ergebnisse der mathematischen Forschung des 19. und 20. Jahrhunderts angesehen werden. Leider sprengt eine formale Behandlung dieser Themen den Rahmen einer universitären Anfängerveranstaltung, so dass im folgenden ausgewählte Themen aus einzelnen Teilgebieten der Mathematik, die für die spätere Informatikausbildung von Bedeutung sind, eher anschaulich und intuitiv beschrieben werden. Natürlich werden auch hierbei Präzision in den Begriffen und formale Korrektheit in der Argumentation versucht.

1.1 Mengen

Die Definition des Begriffs **Menge** gehört zu den grundlegenden Bausteinen der Mathematik. Die formale Behandlung ist den Grundlagen der Mathematik vorbehalten. Georg Cantor (1845 – 1918), der Begründer der Mengenlehre, hat den Begriff der Menge anschaulich folgendermaßen definiert:

> *Eine Menge ist eine Zusammenfassung von bestimmten, wohlunterscheidbaren Objekten unserer Anschauung oder unseres Denkens zu einem Ganzen.*

Eine Menge A kann durch die **Aufzählung der in ihr enthaltenen Elemente** beschrieben werden, z.B. $A = \{ a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, r, t, u, v, w, x, y, z, \}$ als die Menge der Buchstaben unseres Alphabets in Kleinschreibung ohne Umlaute. Falls die einzelnen Elemente der Aufzählung allgemein geläufig sind, beschränkt man sich häufig auf die Angabe der ersten und letzten Elemente wie in $A = \{ a, b, c, \ldots, x, y, z \}$ bzw. auf die Angabe

der ersten Elemente, z.B. $A = \{ 4, 6, 8, ... \}$ als die Menge aller geraden Zahlen, die größer oder gleich 4 sind.

Zu beachten ist, dass es bei der Aufzählung auf die Reihenfolge der Elemente einer Menge nicht ankommt und dass gleiche Elemente in der Aufzählung nur einmal angegeben werden.

Sehr häufig wird eine Menge **durch charakteristische Eigenschaften beschrieben**, die jedem ihrer Elemente zukommen, und zwar in der Form $M = \{ x \mid x \text{ hat die Eigenschaften } ... \}$, z.B. $L = \{ z \mid z \text{ ist Lösung der Gleichung } x^2 + 2x - 3 = 0 \}$ oder in aufzählender Schreibweise $L = \{ -3,\ 1 \}$.

Liegt ein Element *a* in der Menge *A* (ist *a* in der Menge *A* enthalten), so wird $a \in A$ geschrieben; liegt *a* nicht in *A*, so wird $a \notin A$ geschrieben.

Die **leere Menge**, bezeichnet mit $\varnothing$, ist diejenige Menge, die kein Element enthält.

Besitzen alle Elemente einer Menge *A* auch die Eigenschaften, durch die die Elemente einer Menge *B* gekennzeichnet sind, so ist *A* **Teilmenge** von *B*, geschrieben $A \subseteq B$. Enthält *B* mindestens ein Element, das nicht in *A* vorkommt, so ist *A* **echte Teilmenge** von *B*, geschrieben $A \subset B$.

Werden also die Elemente von *B* durch die Eigenschaften E_1, E_2, ..., E_n charakterisiert, d.h. $B = \{ x \mid x \text{ hat die Eigenschaften } E_1, ..., E_n \}$, und werden die Elemente von *A* durch die Eigenschaften E_1', ..., E_m' charakterisiert, d.h. $A = \{ x \mid x \text{ hat die Eigenschaften } E_1', ..., E_m' \}$, wobei alle Eigenschaften E_1, E_2, ..., E_n unter den Eigenschaften E_1', ..., E_m' vorkommen oder sich aus den Eigenschaften E_1', ..., E_m' durch logische Schlüsse ableiten lassen, so ist $A \subseteq B$. Die Elemente einer Teilmenge *A* einer Menge *B* werden also in der Regel durch mehr Eigenschaften charakterisiert als die Elemente der Obermenge *B*.

Für jede Menge *A* gelten die beiden Teilmengenbeziehungen $\varnothing \subseteq A$ und $A \subseteq A$.

Zwei Mengen *A* und *B* sind **gleich**, geschrieben $A = B$, wenn für jedes Element $a \in A$ auch $a \in B$ und für jedes $b \in B$ auch $b \in A$ gilt, wenn also sowohl $A \subseteq B$ als auch $B \subseteq A$ gelten.

Die **Vereinigung** der Mengen A und B, geschrieben $A \cup B$, besteht aus den Elementen, die in A oder in B (oder in beiden Mengen) liegen: $A \cup B = \{ x \mid x \in A \text{ oder } x \in B \}$.

Der **Schnitt** der Mengen A und B, geschrieben $A \cap B$, besteht aus den Elementen, die sowohl in A als auch in B liegen: $A \cap B = \{ x \mid x \in A \text{ und } x \in B \}$.

Die **Differenz** der Mengen B und A, geschrieben $B \setminus A$ besteht aus den Elementen, die in B, aber nicht in A liegen: $B \setminus A = \{ x \mid x \in B \text{ und } x \notin A \}$.

Ist $A \subseteq B$, so ist das **Komplement** der Menge A **bezüglich** der Menge B, geschrieben $\overline{A}^B$, definiert durch $\overline{A}^B = \{ x \mid x \in B \text{ und } x \notin A \}$.

Offensichtlich ist (für $A \subseteq B$) $\overline{A}^B = B \setminus A$.

Für eine Menge A bezeichnet $|A|$ die **Anzahl der Elemente (**oder die **Mächtigkeit)** von A.

Mit $\mathbf{P}(A)$ wird die **Potenzmenge** der Menge A bezeichnet, die aus allen Teilmengen der Menge A besteht, d.h. $\mathbf{P}(A) = \{ L \mid L \subseteq A \}$.

Beispielsweise lautet für $A = \{1, 2, 3\}$ die Potenzmenge
$\mathbf{P}(A) = \{ \varnothing, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\} \}$.

Für Mengen A_1, A_2, ..., A_n wird das **kartesische Produkt** definiert als
$A_1 \times A_2 \times \ldots \times A_n = \{ (a_1, a_2, \ldots, a_n) \mid a_1 \in A_1, a_2 \in A_2, \ldots, a_n \in A_n \}$.

Ein Element $(a_1, a_2, \ldots, a_n) \in A_1 \times A_2 \times \ldots \times A_n$ wird als ***n*-Tupel** bezeichnet. Bei einem 2-Tupel spricht man auch von einem **Paar**.

Eine ***n*-stellige Relation** R auf einer Menge M ist eine Teilmenge von $\underbrace{M \times M \times \ldots \times M}_{n\text{-mal}}$. Bei einer zweistelligen Relation $R \subseteq M \times M$ schreibt man anstelle von $(a, b) \in R$ häufig aRb.

Die grundlegenden Rechenregeln für Operationen mit Mengen werden in folgendem Satz zusammengefasst:

Satz 1.1-1:

Es seien im folgenden A, B und C Mengen. Dann gilt:

(i) $A \cup \varnothing = A$, $A \cap \varnothing = \varnothing$.

(ii) $A \cap B \subseteq A$, $A \cap B \subseteq B$, $A \subseteq A \cup B$, $B \subseteq A \cup B$.

(iii) $A \cup B = B \cup A$, $A \cap B = B \cap A$ (**Kommutativgesetze**).

(iv) $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$ (**Assoziativgesetze**); diese Regeln rechtfertigen die Schreibweisen
$A \cup B \cup C = A \cup (B \cup C) = (A \cup B) \cup C$ und
$A \cap B \cap C = A \cap (B \cap C) = (A \cap B) \cap C$.

(v) $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$, die rechte Seite ist eine **disjunkte Zerlegung** von $A \cup B$. Dabei heißt eine Zerlegung $M = M_1 \cup M_2$ der Menge M disjunkt, wenn $M_1 \cap M_2 = \varnothing$ ist.

(vi) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (**Distributivgesetze**)

(vii) $A \cap (A \cup B) = A$, $A \cup (A \cap B) = A$.

(viii) Ist $A \subseteq C$, so ist $\left(\overline{A}^C\right)^C = A$.

(ix) Sind $A \subseteq C$ und $B \subseteq C$, so gelten
$\overline{(A \cap B)}^C = \overline{A}^C \cup \overline{B}^C$, $\overline{(A \cup B)}^C = \overline{A}^C \cap \overline{B}^C$ (**Regeln von de Morgan**).

(x) Sind $A \subseteq C$ und $B \subseteq C$, so folgt aus $A \subseteq B$ die Beziehung $\overline{B}^C \subseteq \overline{A}^C$ und umgekehrt.

Bemerkung: Die Voraussetzungen $A \subseteq C$ bzw. $B \subseteq C$ in den Aussagen (viii) – (x) wurden nur gemacht, weil das Komplement einer Menge A relativ zu einer die Menge A umfassenden Menge C definiert wurde.

Diese Eigenschaften folgen direkt aus den Definitionen; ihr Nachweis wird dem Leser als Übung überlassen.

1.2 Aussagen und deren logische Verknüpfung

Logische Aussagen in der Mathematik werden wie Mengen streng axiomatisch definiert. Auf diesen Ansatz soll hier ebenfalls zugunsten eines intuitiven Ansatzes verzichtet werden.

Unter einer **mathematisch logischen Aussage** versteht man einen Satz (in einem logischen System), der entweder WAHR oder FALSCH ist (den Wahrheitswert WAHR oder FALSCH besitzt, umgangssprachlich: wahr oder falsch ist). Beispielsweise ist

- „*13 ist eine Primzahl*“ eine Aussage mit Wahrheitswert WAHR („eine wahre Aussage“)
- „$\sqrt{2}$ *ist eine rationale Zahl*“ eine Aussage mit Wahrheitswert FALSCH („eine falsche Aussage“)
- „*Jede gerade natürliche Zahl, die größer als 2 ist, lässt sich als Summe zweier Primzahlen darstellen*“ eine Aussage, deren Wahrheitswert noch nicht bekannt ist, die aber einen der beiden Wahrheitswerte WAHR oder FALSCH besitzt.

Der Satz „*Dieser Satz hat den Wahrheitswert* FALSCH“ ist keine mathematische Aussage, da er weder den Wahrheitswert WAHR noch FALSCH haben kann. Derartige Sätze, die eine selbstbezogene semantische Aussage versuchen, heißen **Paradoxien**.

Sind P und Q Aussagen, so kann man sie mit Hilfe der **logischen Junktoren** $\wedge$ („und“), $\vee$ („oder“) bzw. $\neg$ („nicht“) zu neuen Aussagen $(P \wedge Q)$, $(P \vee Q)$ bzw. $(\neg P)$ zusammensetzen. Dabei kann man häufig auf die Klammern verzichten, wenn man annimmt, dass der Junktor $\neg$ stärker als der Junktor $\wedge$ und dieser stärker als der Junktor $\vee$ bindet. Die Wahrheitswerte der zusammengesetzten Aussagen ergeben sich aus den Wahrheitswerten der Teile gemäß folgender **Wahrheitstafeln**:

Wahrheitswerte von			
P	Q	$(P \wedge Q)$	$(P \vee Q)$
FALSCH	FALSCH	FALSCH	FALSCH
FALSCH	WAHR	FALSCH	WAHR
WAHR	FALSCH	FALSCH	WAHR
WAHR	WAHR	WAHR	WAHR
Bezeichnung		**Konjunktion**	**Disjunktion**

Wahrheitswerte von	
P	$(\neg P)$
FALSCH	WAHR
WAHR	FALSCH
Bezeichnung	**Negation**

Neben diesen drei Junktoren werden in logischen Aussagen häufig noch die Junktoren $\Rightarrow$ („impliziert“, „hat zur Folge“, „aus ... folgt ...“), $\Leftrightarrow$ („... ist gleichbedeutend mit ...“, „... gilt genau dann wenn ... gilt“) und $\oplus$ („exklusives oder“, in der englischsprachigen Fachliteratur auch XOR) verwendet, die durch folgende Wahrheitstafeln definiert sind:

Wahrheitswerte von				
P	Q	$(P \Rightarrow Q)$	$(P \Leftrightarrow Q)$	$(P \oplus Q)$
FALSCH	FALSCH	WAHR	WAHR	FALSCH
FALSCH	WAHR	WAHR	FALSCH	WAHR
WAHR	FALSCH	FALSCH	FALSCH	WAHR
WAHR	WAHR	WAHR	WAHR	FALSCH
Bezeichnung		**Implikation**	**Äquivalenz**	**Antivalenz**

Um den Wahrheitswert einer komplexen Aussage zu ermitteln, die aus durch Junktoren verbundenen Teilaussagen besteht, werden alle Kombinationen von Wahrheitswerten in die Grundaussagen, d.h. in die Teilaussagen, die keine Junktoren erhalten, eingesetzt und durch Anwendung obiger Wahrheitstafeln die sich ergebenden Wahrheitswerte unter Beachtung der Klammersetzung bzw. Bindung der Junktoren ermittelt.

Beispiele:

P	Q	$(P \Rightarrow Q)$	$\Leftrightarrow$	$((\neg P)$	$\wedge$	$Q)$
FALSCH	FALSCH	WAHR	FALSCH	W	F	F
FALSCH	WAHR	WAHR	WAHR	W	W	W
WAHR	FALSCH	FALSCH	WAHR	F	F	F
WAHR	WAHR	WAHR	FALSCH	F	F	W

P	Q	$(P \Rightarrow Q)$	$\Leftrightarrow$	$((\neg P)$	$\vee$	$Q)$
FALSCH	FALSCH	WAHR	WAHR	W	W	F
FALSCH	WAHR	WAHR	WAHR	W	W	W
WAHR	FALSCH	FALSCH	WAHR	F	F	F
WAHR	WAHR	WAHR	WAHR	F	W	W

Eine zusammengesetzte Aussage, deren Wahrheitswert bei allen möglichen Belegungen mit Wahrheitswerten der Teilaussagen, die keine Junktoren enthalten, stets WAHR ist, heißt **Tautologie**.

Beispielweise sind die Aussagen $(P \vee (\neg P))$ und $(P \Rightarrow Q) \Leftrightarrow ((\neg P) \vee Q)$ Tautologien, nicht aber $(P \Rightarrow Q) \Leftrightarrow ((\neg P) \wedge Q)$.

Das Beispiel $(P \Rightarrow Q) \Leftrightarrow ((\neg P) \vee Q)$ zeigt, dass man den Junktor $\Rightarrow$ durch die Junktoren $\neg$ und $\vee$ ausdrücken kann, indem man in einer Aussage jedes Auftreten einer Teilaussage der Form $(P \Rightarrow Q)$ durch die Teilaussage $((\neg P) \vee Q)$ ersetzt. Der Junktor $\Rightarrow$ ist im Grunde also überflüssig, nur vereinfacht er die Struktur der Aussagen und erhöht damit ihre Lesbarkeit. Der folgende Satz zeigt, dass sich auch die anderen Junktoren $\wedge$, $\Leftrightarrow$ und $\oplus$ mit Hilfe der Junktoren $\vee$ und $\neg$ ausdrücken lassen, da sich jeweils links und rechts des Junktors $\Leftrightarrow$ die gleichen Wahrheitswerte ergeben. Teil (v) zeigt, dass man auch $\wedge$ und $\neg$ nehmen kann, um alle Junktoren $\vee$, $\Rightarrow$, $\Leftrightarrow$ und $\oplus$ auszudrücken.

Satz 1.2-1:

Es seien P und Q Aussagen. Dann sind die folgenden Aussagen Tautologien.

(i) $(P \wedge Q) \Leftrightarrow (\neg((\neg P) \vee (\neg Q)))$,
in vereinfachter Schreibweise: $(P \wedge Q) \Leftrightarrow \neg(\neg P \vee \neg Q)$.

(ii) $(P \Rightarrow Q) \Leftrightarrow ((\neg P) \vee Q)$,
in vereinfachter Schreibweise: $(P \Rightarrow Q) \Leftrightarrow (\neg P \vee Q)$.

(iii) $(P \Leftrightarrow Q) \Leftrightarrow ((P \Rightarrow Q) \wedge (Q \Rightarrow P))$,

$(P \Leftrightarrow Q) \Leftrightarrow (((\neg P) \vee Q) \wedge ((\neg Q) \vee P))$,
in vereinfachter Schreibweise: $(P \Leftrightarrow Q) \Leftrightarrow ((\neg P \vee Q) \wedge (\neg Q \vee P))$,

$(P \Leftrightarrow Q) \Leftrightarrow (\neg((\neg((\neg P) \vee Q)) \vee (\neg((\neg Q) \vee P))))$,
in vereinfachter Schreibweise: $(P \Leftrightarrow Q) \Leftrightarrow \neg((\neg(\neg P \vee Q)) \vee (\neg(\neg Q \vee P)))$.

(iv) $(P \oplus Q) \Leftrightarrow ((P \wedge (\neg Q)) \vee ((\neg P) \wedge Q))$,
in vereinfachter Schreibweise: $(P \oplus Q) \Leftrightarrow ((P \wedge \neg Q) \vee (\neg P \wedge Q))$,

$(P \oplus Q) \Leftrightarrow ((\neg((\neg P) \vee Q)) \vee (\neg((\neg Q) \vee P)))$,
in vereinfachter Schreibweise: $(P \oplus Q) \Leftrightarrow (\neg(\neg P \vee Q) \vee \neg(\neg Q \vee P))$.

(v) $(P \vee Q) \Leftrightarrow (\neg((\neg P) \wedge (\neg Q)))$,
in vereinfachter Schreibweise: $(P \vee Q) \Leftrightarrow \neg(\neg P \wedge \neg Q)$.

(vi) $(P \wedge (P \Rightarrow Q)) \Rightarrow Q$ (**Modus ponens**).

Der Nachweis dieser Aussagen erfolgt durch Auswertung der entsprechenden Wahrheitstafeln.

Bemerkung: Man kommt sogar mit nur einem Junktor aus, um alle anderen Junktoren auszudrücken. Dazu wird der Junktor $\uparrow$ durch folgende Wahrheitstafel definiert:

Wahrheitswerte von		
P	Q	$(P \uparrow Q)$
FALSCH	FALSCH	WAHR
FALSCH	WAHR	WAHR
WAHR	FALSCH	WAHR
WAHR	WAHR	FALSCH
Bezeichnung		**Sheffer-Operation**

Es gilt $(P \uparrow Q) \Leftrightarrow (\neg(P \wedge Q))$, und damit ist $(\neg P)$ gleichwertig mit (lässt sich ausdrücken durch) $(P \uparrow P)$, und es ist $(P \wedge Q)$ gleichwertig mit $((P \uparrow Q) \uparrow (P \uparrow Q))$.

Der folgende Satz zeigt strukturelle Äquivalenzen zwischen Sätzen der elementaren Mengenlehre (Satz 1.1-1) und der Aussagenlogik.

Satz 1.2-2:

Es seien P, Q und R Aussagen. Die Aussage W habe den Wahrheitswert WAHR, die Aussage F habe den Wahrheitswert FALSCH.

Dann sind die folgenden Aussagen Tautologien.

(i) $(P \vee F) \Leftrightarrow P$, $P \wedge F \Leftrightarrow F$.

(ii) $(P \wedge Q) \Rightarrow P$, $(P \wedge Q) \Rightarrow Q$,
$P \Rightarrow (P \vee Q)$, $Q \Rightarrow (P \vee Q)$.

(iii) $(P \vee Q) \Leftrightarrow (Q \vee P)$, $(P \wedge Q) \Leftrightarrow (Q \wedge P)$
(**Kommutativgesetze**).

Satz 1.1-1:

Es seien im folgenden A, B und C Mengen.

Dann gilt:

(i) $A \cup \varnothing = A$, $A \cap \varnothing = \varnothing$.

(ii) $A \cap B \subseteq A$, $A \cap B \subseteq B$,
$A \subseteq A \cup B$, $B \subseteq A \cup B$.

(iii) $A \cup B = B \cup A$, $A \cap B = B \cap A$
(Kommutativgesetze).

../..

(iv) $(P \vee (Q \vee R)) \Leftrightarrow ((P \vee Q) \vee R)$, $(P \wedge (Q \wedge R)) \Leftrightarrow ((P \wedge Q) \wedge R)$ (**Assoziativgesetze**); diese Regeln rechtfertigen die Schreibweisen $(P \vee Q \vee R)$ anstelle von $(P \vee (Q \vee R))$ und $(P \wedge Q \wedge R)$ anstelle von $(P \wedge (Q \wedge R))$.	(iv) $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$ (Assoziativgesetze).
(v) $(P \vee Q) \Leftrightarrow ((P \wedge \neg Q) \vee (P \wedge Q) \vee (Q \wedge \neg P))$.	(v) $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$.
(vi) $(P \wedge (Q \vee R)) \Leftrightarrow ((P \wedge Q) \vee (P \wedge R))$, $(P \vee (Q \wedge R)) \Leftrightarrow ((P \vee Q) \wedge (P \vee R))$ (**Distributivgesetze**)	(vi) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (Distributivgesetze)
(vii) $(P \wedge (P \vee Q)) \Leftrightarrow P$, $(P \vee (P \wedge Q)) \Leftrightarrow P$.	(vii) $A \cap (A \cup B) = A$, $A \cup (A \cap B) = A$.
(viii) $(\neg(\neg P)) \Leftrightarrow P$.	(viii) Ist $A \subseteq C$, so ist $(\overline{A^C})^C = A$.
(ix) $(\neg(P \wedge Q)) \Leftrightarrow (\neg P \vee \neg Q)$, $(\neg(P \vee Q)) \Leftrightarrow (\neg P \wedge \neg Q)$ (**Regeln von de Morgan**)	(ix) Sind $A \subseteq C$ und $B \subseteq C$, so gelten $\overline{(A \cap B)}^C = \overline{A}^C \cup \overline{B}^C$, $\overline{(A \cup B)}^C = \overline{A}^C \cap \overline{B}^C$ (Regeln von de Morgan)
(x) $(P \Rightarrow Q) \Leftrightarrow (\neg Q \Rightarrow \neg P)$	(x) Sind $A \subseteq C$ und $B \subseteq C$, so folgt aus $A \subseteq B$ die Beziehung $\overline{B}^C \subseteq \overline{A}^C$ und umgekehrt.

Allgemein gilt, dass ein Satz der elementaren Mengenlehre, der nur die Relationenzeichen = und $\subseteq$ und die Operatoren $\cup, \cap$ bzw. $^{-C}$ (Komplement) verwendet, eine korrespondierende logische Aussage besitzt und umgekehrt, indem man Ersetzungen gemäß folgender Tabelle vornimmt.

Elementare Mengenlehre	=	$\subseteq$	$\cup$	$\cap$	$^{-C}$
Aussagenlogik	$\Leftrightarrow$	$\Rightarrow$	$\vee$	$\wedge$	$\neg$

Die Erweiterung der Aussagenlogik führt auf die **Prädikatenlogik (erster Stufe)**, in der logische Sätze formuliert werden, die die **Quantoren** $\forall$ („für alle ...“) und $\exists$ („es gibt ...“) enthalten können, wobei über „freie Variablen“ in Formeln quantifiziert wird.

Beispiel:
$\forall x \left(\left(\left(x \in \mathbf{N} \right) \wedge \left(x > 1 \right) \right) \Rightarrow \left(\exists p \left(\left(p \text{ ist Primzahl} \right) \wedge \left(p \text{ teilt } x \right) \right) \right) \right)$.

Auf eine weiterführende Einführung in die mathematische Logik soll hier jedoch verzichtet werden.

1.3 Beweistechniken

Um den Wahrheitswert einer komplexen Aussage zu ermitteln, die aus Teilaussagen besteht, die durch Junktoren verbundenen sind, wird ein **mathematischer Beweis** angeführt. Dieser besteht aus einer Aneinanderreihung logischer Schlüsse, die genau spezifizierten Schlussregeln folgen und jederzeit eindeutig nachvollziehbar sind (zumindest sein sollten). Die Grundlage aller Beweise in einem theoretischen System ist eine **Menge von Axiomen**, die als wahr angenommen werden und eine „vernünftige“ Basierung der zugrundeliegenden Theorie bilden. Außerdem gibt es eine **endliche Menge von Schlussregeln**, die es erlauben, aus Aussagen, die bereits als wahr erkannt wurden (dazu gehören die Axiome, deren Wahrheitswert als WAHR angenommen wird), neue wahre Aussagen herzuleiten.

Formal ist ein Beweis eines mathematischen Satzes P eine Aneinanderreihung

$p_1, p_2, ..., p_n$

von Aussagen, wobei am Anfang Axiome oder bereits bewiesene mathematische Sätze, d.h. Sätze mit dem Wahrheitswert WAHR, stehen, und p_n gleich P ist. Den Übergang von einer Zeile p_i zur Zeile p_{i+1} erhält man beispielsweise, indem man eine Tautologie, etwa aus Satz 1.2-2, anwendet: Ist p_i die linke Seite einer Tautologie der Form $Q \Leftrightarrow R$, d.h. p_i hat dieselbe Struktur wie Q, dann ist p_{i+1} gleich R. Eine weitere Möglichkeit der Beweisführung besteht in der Anwendung der als **Modus ponens** bekannten Tautologie $\left(P \wedge \left(P \Rightarrow Q \right) \right) \Rightarrow Q$ (Satz 1.2-1 (vi)). Ist p_j mit $j < i$ eine Zeile, die mit der Aussage P gleichzusetzen ist und hat p_i die Form $\left(P \Rightarrow Q \right)$, dann ist p_{i+1} gleich Q. Wenn man also die Aussage P in Zeile p_j bereits als Aussage mit dem Wahrheitswert WAHR erkannt hat und in Zeile p_i feststellt, dass die Gültigkeit der Aussage P die Gültigkeit der Aussage Q impliziert, dann ist Q eine Aussage mit dem Wahrheitswert WAHR.

Ohne an dieser Stelle genauer auf den formalen Vorgang des Beweisens in der Mathematik einzugehen, werden einige mögliche Vorgehensweisen bei Beweisführung von Aussagen beschrieben.

A. **Direkter Beweis**:

Die oben beschriebene Vorgehensweise ist ein Beispiel für einen direkten Beweis.

Häufig treten mathematische Aussagen der Form $(P \Rightarrow Q)$ auf. Für einen Beweis dieser Aussage kann man folgendermaßen vorgehen:

Man nimmt an, dass *P* den Wahrheitswert WAHR besitzt (dass *P* wahr ist). Durch eine „geeignete" Argumentation (Anwendung logischer Schlüsse) zeigt man, dass dann auch *Q* den Wahrheitswert WAHR hat (dass *Q* wahr ist).

Beispiel:

Es ist folgende Aussage zu beweisen:
Ist p eine Primzahl[1]*, die größer oder gleich 5 ist, so ist* $p^2 - 1$ *durch 24 teilbar.*
Die Aussage besitzt die Form $(P \Rightarrow Q)$ mit
P: *p ist eine Primzahl, die größer oder gleich 5 ist*
und
Q: $p^2 - 1$ *ist durch 24 teilbar.*

Für den Beweis kann folgendermaßen argumentiert werden:
Es sei *p* eine Primzahl mit $p \geq 5$ (Aussage *P* wird als WAHR angenommen). Die Zahlen $p-1$, p und $p+1$ sind drei aufeinanderfolgende Zahlen. Genau eine von ihnen ist durch 3 teilbar. Da *p* Primzahl ist, also nur durch ±1 oder ±*p* teilbar ist und $p \geq 5$ vorausgesetzt wird, kommt nur eine der Zahlen $p-1$ oder $p+1$ in Frage. Da *p* Primzahl und $p \geq 5$ ist, sind sowohl $p-1$ als auch $p+1$ gerade Zahlen, also durch 2 teilbar. Eine von zwei aufeinanderfolgenden geraden Zahlen lässt sich durch 4 teilen. Insgesamt ergibt sich: Eine der Zahlen $p-1$ oder $p+1$ ist durch 3 teilbar, eine der Zahlen $p-1$ oder $p+1$ ist durch 4 teilbar, und die andere der beiden Zahlen ist durch 2 teilbar. Daher wird $p^2 - 1 = (p-1) \cdot (p+1)$ durch 24 geteilt (Aussage *Q* hat den Wahrheitswert WAHR).

[1] Eine Primzahl *p* ist eine natürliche Zahl mit $p \geq 2$, die nur durch ±1 oder ±*p* (ohne Rest) teilbar ist.

B. **Indirekter Beweis**:

Zum Beweis der Aussage $(P \Rightarrow Q)$ zeigt man $(\neg Q \Rightarrow \neg P)$, weil eventuell die Argumentation in dieser Richtung einfacher ist. Der indirekte Beweis beruht auf der Tautologie $(P \Rightarrow Q) \Leftrightarrow (\neg Q \Rightarrow \neg P)$ (Satz 1.2-2 (x)).

Beispiel:

Es ist folgende Aussage zu beweisen:

Ist a^2 ungerade, wobei a eine ganze Zahl ist, dann ist a ungerade.

Die Aussage besitzt die Form $(P \Rightarrow Q)$ mit
P: *a^2 ist ungerade*
und
Q: *a ist ungerade*.

Für den Beweis kann folgendermaßen argumentiert werden:
Es sei *a* gerade (Die Aussage $\neg Q$ wird als WAHR angenommen). Das bedeutet $a = 2 \cdot a'$ mit einer ganzen Zahl a'. Dann ist $a^2 = 4 \cdot a'^2$ gerade (Die Aussage $\neg P$ hat den Wahrheitswert WAHR).

C. **Beweis einer Äquivalenz**:

Um die Aussage $(P \Leftrightarrow Q)$ zu beweisen, sind zwei „Richtungen“ zu zeigen, nämlich ein Beweis für $(P \Rightarrow Q)$ und ein Beweis für $(Q \Rightarrow P)$.

Häufig treten Äquivalenzen auch in der Form $(P \Leftrightarrow Q)$ und $(Q \Leftrightarrow R)$. In diesem Fall kann man anstelle der vier Beweise für $(P \Rightarrow Q)$, $(Q \Rightarrow P)$, $(Q \Rightarrow R)$ und $(R \Rightarrow Q)$ auch drei Beweise, nämlich für $(P \Rightarrow Q)$, $(Q \Rightarrow R)$ und $(R \Rightarrow P)$, erbringen.

D. **Beweis durch Widerspruch**:

Zum Beweis der Gültigkeit einer Aussage *P* nimmt man an, dass $\neg P$ den Wahrheitswert WAHR besitzt. Durch eine „geeignete“ Argumentation (Anwendung logischer Schlüsse) zeigt man von einer Aussage *Q*, deren Wahrheitswert vorher bereits als WAHR erkannt wurde, dass diese dann den Wahrheitswert FALSCH besitzt. Man zeigt also die Gültigkeit von

$(\neg P \Rightarrow (Q \wedge \neg Q))$. Diese Aussage kann jedoch aufgrund des Wahrheitswerts einer Implikation und wegen der Tatsache, dass $(Q \wedge \neg Q)$ immer den Wahrheitswert FALSCH besitzt, nur dann gültig sein, wenn $\neg P$ den Wahrheitswert FALSCH bzw. P den Wahrheitswert WAHR besitzt.

Beispiel:

Zu zeigen ist die Gültigkeit der Aussage

$\sqrt{2}$ ist keine rationale Zahl bzw. *$\sqrt{2}$ lässt sich nicht als Bruch p/q mit natürlichen Zahlen p und q darstellen*.

Die Aussage P lautet hier: *$\sqrt{2}$ lässt sich nicht als Bruch p/q mit natürlichen Zahlen p und q darstellen*.

Es wird die Aussage $\neg P$ als WAHR angenommen, d.h. $\sqrt{2} = p/q$ mit natürlichen Zahlen p und q. In einem Bruch p/q kann man Faktoren, die in p und q gemeinsam auftreten, heraus kürzen, d.h. p und q haben (bis auf 1) keinen gemeinsamen Faktor (das ist die Aussage Q: *p und q haben (bis auf 1) keinen gemeinsamen Faktor*; Q besitzt den Wahrheitswert WAHR).
Aus $\sqrt{2} = p/q$ folgt $q \cdot \sqrt{2} = p$ und $2 \cdot q^2 = p^2$, also teilt 2 den Wert p^2. Da 2 Primzahl ist, teilt 2 die Zahl p, d.h. $p = 2 \cdot p'$. Damit folgt $2 \cdot q^2 = 4 \cdot p'^2$ und $q^2 = 2 \cdot p'^2$. Jetzt ergibt sich, dass 2 den Wert q^2 und damit die Zahl q teilt (wieder, weil 2 Primzahl ist). Insgesamt ist die Zahl 2 Faktor sowohl von p als auch von q; die Aussage Q hat also den Wahrheitswert FALSCH.

E. **Beweis durch vollständige Induktion**:

Zum Beweis von Aussagen über natürliche Zahlen wird häufig die Beweismethode der vollständigen Induktion eingesetzt. Diese Methode beruht auf einer charakteristischen Eigenschaft der natürlichen Zahlen und wird in Kapitel 1.5 behandelt.

1.4 Algebraische Grundstrukturen und Zahlensysteme

Im Folgenden wird ein Überblick über den Aufbau der wichtigsten Zahlensysteme gegeben, und es werden einige grundlegende algebraische Strukturen definiert.

Grundlage aller hier beschriebenen Zahlensysteme ist die **Menge der natürlichen Zahlen**. Diese wird durch ein Axiomensystem beschrieben, d.h. durch ein Regelsystem, das die Menge der natürlichen Zahlen eindeutig durch ihre Eigenschaften definiert:

Axiom 1: $0 \in \mathbf{N}$

Axiom 2: Für jedes $n \in \mathbf{N}$ gibt es ein als **Nachfolger** (Nachfolgerzahl) bezeichnetes Element $n' \in \mathbf{N}$.

Axiom 3: 0 ist nicht Nachfolger eines Elements $n \in \mathbf{N}$, d.h. es gibt kein $n \in \mathbf{N}$ mit $n' = 0$.

Axiom 4: Unterschiedliche Elemente n und m haben unterschiedliche Nachfolger. Gleichbedeutend damit ist: Sind die Nachfolger n' und m' zweier natürlicher Zahlen gleich, so sind die Zahlen n und m ebenfalls gleich.

Axiom 5: Eine Menge M natürlicher Zahlen, die die Zahl 0 und mit jeder Zahl n auch ihren Nachfolger n' enthält, ist mit **N** identisch.

N ist das einzige „Modell“ dieses Axiomensystems.

Statt $0'$ schreibt man auch 1, statt $0''$ schreibt man 2, statt $0'''$ schreibt man 3 usw. Der n-te Nachfolger der 0 ist die natürliche Zahl n. Insbesondere ist eine natürliche Zahl $n \neq 0$ eine Nachfolgerzahl.

Man schreibt üblicherweise

$\mathbf{N} = \{ 0, 1, 2, 3, 4, ... \}$.

Aufbauend auf den so definierten Grundeigenschaften der natürlichen Zahlen werden **arithmetische Operationen** auf den natürlichen Zahlen eingeführt:

Die **Addition** + wird über den Nachfolger n' einer natürlichen Zahl n definiert durch die („rekursiven“) Regeln

$m + 0 = m$,
$m + n' = (m + n)'$ für jede natürliche Zahl m und jede natürliche Zahl n.

Damit ist beispielsweise $2 + 3 = 2 + 0''' = (2 + 0'')' = ((2 + 0')')' = (((2 + 0)')')' = 2''' = (0'')''' = 5$.

Auf ähnliche Weise und durch Zurückführung auf die Addition wird die **Multiplikation** $\cdot$ eingeführt:

$m \cdot 0 = 0$,
$m \cdot n' = (m \cdot n) + m$ für jede natürliche Zahl m und jede natürliche Zahl n.

Damit ist $7 \cdot 3 = 7 \cdot 2' = (7 \cdot 2) + 7 = (7 \cdot 1') + 7 = ((7 \cdot 1) + 7) + 7 = (7 + 7 + 7) = 21$.

Die so eingeführten arithmetischen Operationen genügen wichtigen Gesetzmäßigkeiten:

Es gilt für jede natürliche Zahl n, für jede natürliche Zahl m und für jede natürliche Zahl k:

$k + (m + n) = (k + m) + n$,
$k \cdot (m \cdot n) = (k \cdot m) \cdot n$ (**Assoziativgesetze**)

Das Assoziativgesetz besagt, dass es bei gleichen Operatoren auf die Reihenfolge der Klammerung nicht ankommt; sie wird daher in der Regel weggelassen.

$k + m = m + k$,
$k \cdot m = m \cdot k$ (**Kommutativgesetze**)

$k \cdot (m + n) = k \cdot m + k \cdot n$ (**Distributivgesetz**)

Der Nachweis dieser Regeln erfolgt durch Rückführung auf obige Definitionen. Beispielsweise wird das Kommutativgesetz der Addition wie folgt gezeigt:

Die Gültigkeit des Kommutativgesetzes wird zunächst für $k = 0$ und alle $m \in \mathbf{N}$ nachgewiesen:

Für $m = 0$ ist $0 + m = 0 + 0 = m + 0$.

Für eine Nachfolgerzahl m' ist nach Definition der Addition und der Tatsache, dass das Kommutativgesetz für $k = 0$ und m bereits nachgewiesen wurde,

$0 + m' = (0 + m)'$ (nach Definition der Addition)
$= (m + 0)'$ (für m bereits gezeigt)
$= m'$ (nach Definition der Addition)
$= m' + 0$ (nach Definition der Addition).

Aus der Tatsache, dass das Kommutativgesetz für k bereits gezeigt wurde, wird seine Gültigkeit für k' und alle $m \in \mathbf{N}$ nachgewiesen:

Für $m = 0$ ist
$k' + 0 = k'$ (nach Definition der Addition)
$= (k + 0)'$ (nach Definition der Addition)
$= (0 + k)'$ (für k bereits nachgewiesen)
$= 0 + k'$ (nach Definition der Addition).

Für eine Nachfolgerzahl m' ist nach Definition der Addition und der Tatsache, dass das Kommutativgesetz für k und m bereits nachgewiesen wurde,
$k' + m' = (k' + m)'$ (nach Definition der Addition)
$= (m + k')'$ (für m bereits gezeigt)
$= (m + k)''$ (nach Definition der Addition)
$= (k + m)''$ (für k bereits gezeigt)
$= (k + m')'$ (nach Definition der Addition)
$= (m' + k)'$ (für k bereits gezeigt)
$= m' + k'$ (nach Definition der Addition).

Ähnlich geht es mit dem Kommutativgesetz der Multiplikation:

Die Gültigkeit des Kommutativgesetzes wird zunächst für $k = 0$ und alle $m \in \mathbf{N}$ gezeigt:

Für $m = 0$ ist $0 \cdot m = 0 \cdot 0 = 0 = m \cdot 0$.

Für eine Nachfolgerzahl m' ist nach Definition der Multiplikation und der Tatsache, dass das Kommutativgesetz für $k = 0$ und m bereits nachgewiesen wurde,
$0 \cdot m' = (0 \cdot m) + 0$ (nach Definition der Multiplikation)
$= (m \cdot 0) + 0$ (für m bereits gezeigt)
$= m \cdot 0$ (nach Definition der Addition)
$= 0$ (nach Definition der Multiplikation)
$= m' \cdot 0$ (nach Definition der Multiplikation).

Aus der Tatsache, dass das Kommutativgesetz für k bereits gezeigt wurde, wird seine Gültigkeit für k' und alle $m \in \mathbf{N}$ nachgewiesen:

Für $m = 0$ ist

$k' \cdot 0 = 0$ (nach Definition der Multiplikation)
$= 0 + 0$ (nach Definition der Addition)
$= (k \cdot 0) + 0$ (nach Definition der Multiplikation)
$= (0 \cdot k) + 0$ (für k bereits nachgewiesen)
$= 0 \cdot k'$ (nach Definition der Multiplikation).

Für eine Nachfolgerzahl m' ist nach Definition der Multiplikation und der Tatsache, dass das Kommutativgesetz für k und m bereits nachgewiesen wurde,

$k' \cdot m' = (k' \cdot m) + k'$ (nach Definition der Multiplikation)
$= ((k' \cdot m) + k)'$ (nach Definition der Addition)
$= ((m \cdot k') + k)'$ (für m bereits nachgewiesen)
$= ((m \cdot k) + m + k)'$ (nach Definition der Multiplikation)
$= ((k \cdot m) + k + m)'$ (wegen der Kommutativität der Addition und da die Kommutativität für k bereits nachgewiesen ist)
$= ((k \cdot m') + m)'$ (nach Definition der Multiplikation)
$= (k \cdot m') + m'$ (nach Definition der Addition)
$= (m' \cdot k) + m'$ (für k bereits nachgewiesen)
$= m' \cdot k'$ (nach Definition der Multiplikation).

Zu beachten ist hierbei, dass die Gültigkeit der Assoziativgesetze als bereits nachgewiesen vorausgesetzt wird.

Satz 1.4-1:

Es seien $n \in \mathbf{N}$ und $m \in \mathbf{N}$. Dann gilt

(i) $n \cdot 1 = n$.

(ii) Für $n \neq 0$ und $m \neq 0$ ist $n \cdot m \neq 0$. Die Gleichung $n \cdot m = 0$ impliziert $n = 0$ oder $m = 0$.

(iii) Die Gleichung $n + x = n$ mit $x \in \mathbf{N}$ ist nur für $x = 0$ erfüllt. Die Gleichung $n + m = 0$ ist nur für $n = m = 0$ erfüllt.

Laut Definition der Multiplikation und der Addition ist
$n \cdot 1 = n \cdot 0' = (n \cdot 0) + n = 0 + n = n + 0 = n$.

Für (ii) seien n und m Nachfolgerzahlen, d.h. $n = k'$ und $m = l'$, dann ist

$n \cdot m = k' \cdot l' = (k' \cdot l) + k' = ((k' \cdot l) + k)'$, also eine Nachfolgerzahl. Mit Axiom 3 folgt die Behauptung.

(iii) sieht man folgendermaßen:
Für $n = 0$ ist $0 = 0 + x = x$. Ist *n* eine Nachfolgerzahl, $n = k'$, und ist $k + y = k$ mit $y \in \mathbf{N}$ nur mit $y = 0$ möglich, dann folgt aus $k' = k' + x = x + k' = (x + k)'$ mit Axiom 4 $k = x + k$ und $x = 0$.
In der zweiten Gleichung in (iii) ist bei $n = 0$: $0 = 0 + m = m$. Wäre *n* eine Nachfolgerzahl, $n = k'$, dann ist $0 = k' + m = m + k' = (m + k)'$. Das ist gemäß Axiom 3 nicht möglich. Daher gilt $n = 0$ und $m = 0$.

Die natürliche Zahl *n* heißt **kleiner als** die natürliche Zahl *m*, geschrieben $n < m$, wenn die Gleichung $n + x = m$ eine Lösung $x \in \mathbf{N}$ mit $x \neq 0$ besitzt. Man schreibt $n \leq m$, wenn $n < m$ oder $n = m$ gilt. Durch diese Festlegungen wird eine **totale Ordnungsrelation** auf den natürlichen Zahlen definiert.

Erläuterung:

Eine zweistellige Relation $\lhd$ auf einer Menge *M* heißt **partielle Ordnungsrelation**, wenn für jedes $a \in M$, jedes $b \in M$ und jedes $c \in M$ gilt:

(i) $a \lhd a$ (**Reflexivität**)

(ii) aus $a \lhd b$ und $b \lhd a$ folgt $a = b$ (**Antisymmetrie**)

(iii) aus $a \lhd b$ und $b \lhd c$ folgt $a \lhd c$ (**Transitivität**).

Eine partielle Ordnungsrelation heißt **totale Ordnungsrelation**, wenn für jedes $a \in M$ und für jedes $b \in M$ zusätzlich gilt:

$a \lhd b$ oder $b \lhd a$ (**Vergleichbarkeit**).

Dass es sich bei der durch $\leq$ definierten Relation um eine partielle Ordnungsrelation handelt, sieht man wie folgt:

Die Reflexivität $n \leq n$ ist wegen $n = n$ offensichtlich.

Für den Nachweis der Antisymmetrie setzt man $n \le m$ und $m \le n$ voraus. Das bedeutet $n + x = m$ mit $x \in \mathbf{N}$ und $m + y = n$ mit $y \in \mathbf{N}$. Ersetzt man in der ersten Gleichung *n* durch die Angabe der zweiten Gleichung, so erhält man $m + (y + x) = m$. Mit Satz 1.4-1 (iii) folgt $y + x = 0$ und $y = x = 0$.

Aus $n \le m$ und $m \le k$ folgt $n + x = m$ mit $x \in \mathbf{N}$ und $m + y = k$ mit $y \in \mathbf{N}$. Die erste Gleichung wird in die zweite Gleichung eingesetzt, und man erhält mit der Assoziativität der Addition $k = m + y = (n + x) + y = n + (x + y)$. Das zeigt $n \le k$, und damit gilt die Transitivität der Relation $\le$.

Die so definierten arithmetischen Operationen erlauben nur wenige wirkliche Rechenmanipulationen. Operationen wie Differenzen- oder Quotientenbildung als Umkehroperationen der Addition bzw. der Multiplikation sind nur sehr eingeschränkt möglich, wenn man fordert, dass jeweils das Ergebnis einer dieser Operationen wieder ein Element aus **N** ergibt. Daher bildet man auf der Basis der natürlichen Zahlen einen Zahlenbereich, in den man **N** einbetten kann, und zwar so, dass die arithmetischen Operationen und die definierte Ordnungsrelation auf **N** fortgesetzt werden:

Man definiert auf der Menge $\mathbf{N} \times \mathbf{N}$ der Paare natürlicher Zahlen eine zweistellige Relation $\approx$ durch

$(n, m) \approx (k, l)$ genau dann, wenn $n + l = m + k$ gilt (hier ist die definierte Addition auf den natürlichen Zahlen gemeint).

Diese Relation ist eine Äquivalenzrelation auf der Menge $\mathbf{N} \times \mathbf{N}$ (siehe unten).

Erläuterung:

Eine zweistellige Relation ~ auf einer Menge *M* heißt **Äquivalenzrelation**, wenn für jedes $a \in M$, jedes $b \in M$ und jedes $c \in M$ gilt:

(i) $a \sim a$ (**Reflexivität**)

(ii) aus $a \sim b$ folgt $b \sim a$ (**Symmetrie**)

(iii) aus $a \sim b$ und $b \sim c$ folgt $a \sim c$ (**Transitivität**).

Für $a \in M$ bezeichnet $[a]_{\sim} = \{ b \mid b \sim a \}$ die zu *a* gehörende **Äquivalenzklasse**.

Der folgende Satz führt einige wichtige Eigenschaften einer Äquivalenzrelation auf:

Satz 1.4-2:

Es sei ~ eine Äquivalenzrelation auf der Menge *M*, $a \in M$ und $b \in M$. Dann gilt:

(i) Es ist $a \sim b$ genau dann, wenn $[a]_\sim = [b]_\sim$ gilt.

(ii) Es gilt entweder $[a]_\sim = [b]_\sim$ oder $[a]_\sim \cap [b]_\sim = \varnothing$.

(iii) $\bigcup_{a \in M} [a]_\sim = M$ (auf der linken Seite des Gleichheitszeichens steht die Vereinigung aller Äquivalenzklassen, die man mit Elementen aus *M* bilden kann).

Die Gültigkeit der Aussagen sieht man wie folgt:
Es gelte $a \sim b$, und es sei $x \in [a]_\sim$. Dann ist $x \sim a$ und mit der Transitivität auch $x \sim b$. Also ist $x \in [b]_\sim$, insgesamt $[a]_\sim \subseteq [b]_\sim$. Ist umgekehrt $x \in [b]_\sim$, also $x \sim b$, dann ist mit der Reflexivität und der Transitivität $b \sim a$ und $x \sim a$, d.h. $[b]_\sim \subseteq [a]_\sim$.

Damit folgt aus $x \in [a]_\sim \cap [b]_\sim$ nacheinander: $x \sim a$, $x \sim b$, $a \sim x$ und $a \sim b$, also $[a]_\sim = [b]_\sim$. Besitzen also zwei Äquivalenzklassen ein gemeinsames Element, dann sind sie gleich.

$\bigcup_{a \in M} [a]_\sim \subseteq M$, da Äquivalenzklassen aus Elementen aus *M* bestehen. Ist umgekehrt $x \in M$, dann ist wegen der Reflexivität $x \sim x$ und $x \in [x]_\sim$ und damit $x \in \bigcup_{a \in M} [a]_\sim$.

Die durch „$(n,m) \approx (k,l)$ genau dann, wenn $n + l = m + k$ gilt“ definierte Relation ist eine Äquivalenzrelation auf der Menge $\mathbf{N} \times \mathbf{N}$: Dazu sind die Eigenschaften Reflexivität, Symmetrie und Transitivität nachzuweisen:

Wegen $n + m = m + n$ ist $(n,m) \approx (n,m)$ (Reflexivität).

Es gelte $(n,m) \approx (k,l)$, d.h. $n + l = m + k$. Dann gilt auch $k + m = l + n$, also $(k,l) \approx (n,m)$ (Symmetrie).

Aus $(n,m) \approx (k,l)$ und $(k,l) \approx (g,h)$ folgt $n + l = m + k$ und $k + h = l + g$. Daher gilt $n + l + k + h = m + k + l + g$. Die linke Seite kann geschrieben werden als $(n + h) + (l + k)$, die rechte Seite als $(m + g) + (l + k)$. Aus Axiom 4 der natürlichen Zahlen folgt daraus $n + h = m + g$, d.h. $(n,m) \approx (g,h)$ (Transitivität).

Die Menge der Äquivalenzklassen zu dieser Relation wird als **Menge Z der ganzen Zahlen** bezeichnet:

$$\mathbf{Z} = \left\{ [(n,m)]_{\approx} \mid \; n \in \mathbf{N} \text{ und } m \in \mathbf{N} \right\}.$$

Es sei $(n,m) \in \mathbf{N} \times \mathbf{N}$.

Für $n < m$ sei $x \in \mathbf{N}$ die Lösung der Gleichung $n + x = m$ (anschaulich $x = m - n > 0$). Dann ist $(n,m) \approx (0,x)$ und $[(n,m)]_{\approx} = [(0,x)]_{\approx}$.

Für $m \le n$ sei $y \in \mathbf{N}$ die Lösung der Gleichung $m + y = n$ (anschaulich $y = n - m \ge 0$). Dann ist $(n,m) \approx (y,0)$ und $[(n,m)]_{\approx} = [(y,0)]_{\approx}$.

Daher kann man auch

$$\mathbf{Z} = \left\{ [(n,0)]_{\approx} \mid \; n \in \mathbf{N} \right\} \cup \left\{ [(0,n)]_{\approx} \mid \; n \in \mathbf{N}, n \neq 0 \right\}$$

schreiben. Die Menge der natürlichen Zahlen lässt sich in **Z** einbetten, etwa durch die Vorschrift: Die natürliche Zahl $n \in \mathbf{N}$ wird mit der Äquivalenzklasse $[(n,0)]_{\approx}$ identifiziert. Es ist dann $\mathbf{N} \approx \left\{ [(n,0)]_{\approx} \mid \; n \in \mathbf{N} \right\}$, und es wird (obwohl mathematisch inkorrekt) $\mathbf{N} \subset \mathbf{Z}$ geschrieben.

Auf **Z** werden nun die arithmetischen Operationen Addition, geschrieben $+_{\mathbf{Z}}$, und Multiplikation, geschrieben $\cdot_{\mathbf{Z}}$, definiert:

$$[(n,0)]_{\approx} +_{\mathbf{Z}} [(m,0)]_{\approx} = [(n+m,0)]_{\approx},$$
$$[(0,n)]_{\approx} +_{\mathbf{Z}} [(0,m)]_{\approx} = [(0,n+m)]_{\approx},$$
$$[(n,0)]_{\approx} +_{\mathbf{Z}} [(0,m)]_{\approx} = \begin{cases} [(x,0)]_{\approx} & \text{falls } m \le n \text{ ist; hierbei ist } x \text{ die Lösung der Gleichung } m + x = n \\ [(0,x)]_{\approx} & \text{falls } n < m \text{ ist; hierbei ist } x \text{ die Lösung der Gleichung } n + x = m. \end{cases}$$

Beispielsweise ist $[(3,0)]_{\approx} +_{\mathbf{Z}} [(0,4)]_{\approx} = [(0,1)]_{\approx}$ und $[(4,0)]_{\approx} +_{\mathbf{Z}} [(0,3)]_{\approx} = [(1,0)]_{\approx}$.

$[(n,m)]_{\approx} \cdot_{\mathbf{Z}} [(k,l)]_{\approx} = [(n \cdot k + m \cdot l, n \cdot l + m \cdot k)]_{\approx}$. Hierbei ist zu beachten, dass in den Paaren (n,m) und (k,l) jeweils mindestens ein Wert gleich 0 ist.

Beispielsweise ist $[(3,0)]_{\approx} \cdot_{\mathbf{Z}} [(4,0)]_{\approx} = [(12,0)]_{\approx}$ und $[(3,0)]_{\approx} \cdot_{\mathbf{Z}} [(0,4)]_{\approx} = [(0,12)]_{\approx}$.

Man kann sich davon überzeugen, dass für die so definierten Operationen die Assoziativgesetze, Kommutativgesetze und Distributivgesetze gelten.

Für jede ganze Zahl $[(n,0)]_{\approx}$ bzw. jede ganze Zahl $[(0,m)]_{\approx}$ gelten die Beziehungen
$[(n,0)]_{\approx} +_{\mathbf{Z}} [(0,0)]_{\approx} = [(n,0)]_{\approx}$ bzw. $[(0,m)]_{\approx} +_{\mathbf{Z}} [(0,0)]_{\approx} = [(0,m)]_{\approx}$
und
$[(n,0)]_{\approx} \cdot_{\mathbf{Z}} [(1,0)]_{\approx} = [(n,0)]_{\approx}$ bzw. $[(0,m)]_{\approx} \cdot_{\mathbf{Z}} [(1,0)]_{\approx} = [(0,m)]_{\approx}$.

In der so definierten Menge **Z** ist nun jede Gleichung der Form $[(n,m)]_{\approx} +_{\mathbf{Z}} [(x_1,x_2)]_{\approx} = [(k,l)]_{\approx}$ durch $[(x_1,x_2)]_{\approx} = [(m,n)]_{\approx} +_{\mathbf{Z}} [(k,l)]_{\approx}$ lösbar:

Für $m = 0$ ist

$$\begin{aligned}
[(n,m)]_{\approx} +_{\mathbf{Z}} [(x_1,x_2)]_{\approx} &= [(n,0)]_{\approx} +_{\mathbf{Z}} [(x_1,x_2)]_{\approx} \\
&= [(n,0)]_{\approx} +_{\mathbf{Z}} ([(0,n)]_{\approx} +_{\mathbf{Z}} [(k,l)]_{\approx}) \\
&= ([(n,0)]_{\approx} +_{\mathbf{Z}} [(0,n)]_{\approx}) +_{\mathbf{Z}} [(k,l)]_{\approx} \\
&= [(0,0)]_{\approx} +_{\mathbf{Z}} [(k,l)]_{\approx} \\
&= [(k,l)]_{\approx}.
\end{aligned}$$

Für $n = 0$ zeigt man die Gleichung entsprechend.

Diese Aussage bedeutet, dass es zu jeder ganzen Zahl $[(n,m)]_{\approx} \in \mathbf{Z}$ eine **additiv inverse** Zahl, nämlich $[(m,n)]_{\approx}$ mit $[(n,m)]_{\approx} +_{\mathbf{Z}} [(m,n)]_{\approx} = [(0,0)]_{\approx}$ gibt.

Die Gleichung $[(n,m)]_{\approx} \cdot_{\mathbf{Z}} [(x_1,x_2)]_{\approx} = [(0,0)]_{\approx}$ mit $[(n,m)]_{\approx} \neq [(0,0)]_{\approx}$ (hierbei ist wieder mindestens einer der Werte *n* oder *m* gleich 0) hat als einzige Lösung $[(x_1,x_2)]_{\approx} = [(0,0)]_{\approx}$: Ist etwa $n \neq 0$ und $[(n,0)]_{\approx} \cdot_{\mathbf{Z}} [(x_1,x_2)]_{\approx} = [(0,0)]_{\approx}$, dann gilt $n \cdot x_1 = 0$ und $n \cdot x_2 = 0$ und damit $x_1 = 0$ und $x_2 = 0$ (in den natürlichen Zahlen folgt aus $n \cdot x_1 = 0$: $n = 0$ oder $x_1 = 0$, siehe oben). Genauso argumentiert man bei $[(0,m)]_{\approx} \cdot_{\mathbf{Z}} [(x_1,x_2)]_{\approx} = [(0,0)]_{\approx}$ mit $m \neq 0$.

Die Ordnungsrelation $\leq$ auf den natürlichen Zahlen wird zu einer Ordnungsrelation $\leq_{\mathbf{Z}}$ auf den ganzen Zahlen fortgesetzt:

$[(n,0)]_{\approx} \leq_{\mathbf{Z}} [(m,0)]_{\approx}$ genau dann, wenn $n \leq m$ (in den natürlichen Zahlen) gilt;

$[(0,n)]_{\approx} \leq_{\mathbf{Z}} [(m,0)]_{\approx}$;
$[(0,n)]_{\approx} \leq_{\mathbf{Z}} [(0,m)]_{\approx}$ genau dann, wenn $m \leq n$ (in den natürlichen Zahlen) gilt.

Mit diesen Definitionen bildet die algebraische Struktur $(\mathbf{Z}, +_{\mathbf{Z}}, \cdot_{\mathbf{Z}}, [(0,0)]_{\approx}, [(1,0)]_{\approx})$ einen nullteilerfreien kommunikativen Ring mit 1.

Erläuterung:

Eine algebraische Struktur $(G, \circ)$ mit der Menge G und der zweistelligen Operation $\circ$ heißt **Gruppe**, wenn gilt:

(i) $a \circ b \in G$ für jedes $a \in G$ und jedes $b \in G$ (**Abgeschlossenheit der Operation** $\circ$)

(ii) $a \circ (b \circ c) = (a \circ b) \circ c$ für jedes $a \in G$, jedes $b \in G$ und jedes $c \in G$ (**Assoziativität**)

(iii) es gibt ein Element $e \in G$ mit der Eigenschaft $e \circ a = a \circ e = a$ für jedes $a \in G$ (**Existenz eines neutralen Elements**)

(iv) für jedes $a \in G$ gibt es ein Element $a^{-1} \in G$ mit $a \circ a^{-1} = a^{-1} \circ a = e$; dieses Element heißt **inverses Element** zu a.

Das neutrale Element e einer Gruppe wird meist in die Angabe der Gruppe mit aufgenommen: $(G, \circ, e)$.

Eine Gruppe $(G, \circ)$ heißt **kommutative Gruppe**, wenn zusätzlich gilt:
$a \circ b = b \circ a$ für jedes $a \in G$ und für jedes $b \in G$.

../..

Eine algebraische Struktur $(R, \oplus, \otimes)$ heißt **Ring**, wenn gilt:

(i) $(R, \oplus, 0)$ ist eine kommutative Gruppe (mit neutralem Element $0 \in R$). Das zu $a \in R$ bezüglich der Operation $\oplus$ (additiv) inverse Element wird mit $-a$ bezeichnet.

(ii) $a \otimes b \in R$ für jedes $a \in R$ und jedes $b \in R$ (**Abgeschlossenheit der Operation** $\otimes$)

(iii) $a \otimes (b \otimes c) = (a \otimes b) \otimes c$ für jedes $a \in R$, jedes $b \in R$ und jedes $c \in R$ (**Assoziativität der Operation** $\otimes$)

(iv) $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$ und $(b \oplus c) \otimes a = (b \otimes a) \oplus (c \otimes a)$ für jedes $a \in R$, jedes $b \in R$ und jedes $c \in R$ (**Distributivität der Operation** $\otimes$ **über die Operation** $\oplus$)

Ein Ring $(R, \oplus, \otimes)$ heißt **Ring mit 1**, wenn es ein Element $1 \in R$ gibt mit $a \otimes 1 = 1 \otimes a = a$ für jedes $a \in R$ (**Existenz eines neutralen Elements bezüglich der Operation** $\otimes$).

Ein Ring $(R, \oplus, \otimes)$ heißt **kommutativer Ring**, wenn zusätzlich gilt:
$a \otimes b = b \otimes a$ für jedes $a \in R$ und für jedes $b \in R$.

Ein Ring $(R, \oplus, \otimes)$ heißt **nullteilerfreier Ring**, wenn zusätzlich gilt:
Die Gleichung $a \otimes x = 0$ besitzt für jedes $a \in R$ mit $a \neq 0$ nur die Lösung $x = 0$.

Für den Ring $(R, \oplus, \otimes)$ mit 1 werden die neutralen Elemente mit in die Angabe des Rings aufgenommen: $(R, \oplus, \otimes, 0, 1)$.

Einige häufig verwendete Rechenregeln folgen unmittelbar aus diesen Axiomen:

Satz 1.4-3:

(i) Das neutrale Element e einer Gruppe $(G, \circ, e)$ ist eindeutig bestimmt.
Das zu $a \in G$ inverse Element $a^{-1} \in G$ ist eindeutig bestimmt.
Für $a \in G$ und $b \in G$ ist $(a \circ b)^{-1} = b^{-1} \circ a^{-1}$.

(ii) Ist $(R, \oplus, \otimes, 0, 1)$ ein Ring mit 1 mit bezüglich der Operation $\oplus$ neutralem Element $0 \in R$ und bezüglich der Operation $\otimes$ neutralem Element $1 \in R$, dann gilt für $a \in R$ und $b \in R$ mit den additiv inversen Elementen $-a$ und $-b$:
$-(a \oplus b) = (-b) \oplus (-a) = (-a) \oplus (-b)$,
$0 \otimes a = a \otimes 0 = 0$,
$-a = (-1) \otimes a$,
$(-a) \otimes (-b) = a \otimes b$.

Die Eindeutigkeit des neutralen Elements einer Gruppe $(G, \circ, e)$ sieht man wie folgt: Sind $e_1 \in G$ und $e_2 \in G$ neutrale Elemente in G, dann ist

$e_1 = e_1 \circ e_2$ (da e_2 neutrales Element in G ist)
$= e_2$ (da e_1 neutrales Element in G ist).

Die Eindeutigkeit des zu $a \in G$ inversen Elements ergibt sich aus:
Ist $b \in G$ invers zu $a \in G$, dann ist $b = b \circ e = b \circ (a \circ a^{-1}) = (b \circ a) \circ a^{-1} = e \circ a^{-1} = a^{-1}$.
Damit ist $(a \circ b) \circ (b^{-1} \circ a^{-1}) = a \circ (b \circ b^{-1}) \circ a^{-1} = a \circ e \circ a^{-1} = a \circ a^{-1} = e$, also wegen der Eindeutigkeit inverser Elemente $(a \circ b)^{-1} = (b^{-1} \circ a^{-1})$.

Die Aussagen in (ii) lassen sich nachrechnen:
Mit (i) und der Kommutativität der Operation $\oplus$ folgt $-(a \oplus b) = (-b) \oplus (-a) = (-a) \oplus (-b)$.
$(0 \otimes a) \oplus a = (0 \otimes a) \oplus (1 \otimes a) = (0 \oplus 1) \otimes a = 1 \otimes a = a$; wegen der Eindeutigkeit des neutralen Elements bezüglich der Operation $\oplus$ ist daher $(0 \otimes a) = 0$. Entsprechend ergibt sich $a \otimes 0 = 0$.
$((-1) \otimes a) \oplus a = ((-1) \otimes a) \oplus (1 \otimes a) = (-1 \oplus 1) \otimes a = 0 \otimes a = 0$; wegen der Eindeutigkeit inverser Elemente bezüglich der Operation $\oplus$ ist daher $-a = (-1) \otimes a$.
Für die letzte Gleichung in (ii) zeigt man, dass $(-a) \otimes (-b)$ additiv invers zu $-(a \otimes b)$ ist; nach Definition ist $a \otimes b$ invers zu $-(a \otimes b)$; wegen der Eindeutigkeit inverser Elemente bezüglich der Operation $\oplus$ ist dann $(-a) \otimes (-b) = a \otimes b$:

$$\begin{aligned}((-a)\otimes(-b))\oplus -(a\otimes b) &= ((-a)\otimes(-b))\oplus((-1)\otimes(a\otimes b))\\ &= ((-a)\otimes(-b))\oplus((-a)\otimes b)\\ &= (-a)\otimes(-b\oplus b)\\ &= (-a)\otimes 0\\ &= 0\ .\end{aligned}$$

Wie oben beschrieben, lässt sich **N** in **Z** durch die Identifizierung von $n \in \mathbf{N}$ mit $[(n,0)]_{\approx}$ einbetten. Anstelle von $[(0,n)]_{\approx}$, dem zu $[(n,0)]_{\approx}$ additiv inversen Element, schreibt man auch $-n$. In diesem Sinne bilden die ganzen Zahlen $\{\,[(n,0)]_{\approx} \mid n \in \mathbf{N}\,\}$ die **positiven ganzen Zahlen** (einschließlich 0), entsprechend den natürlichen Zahlen **N**, und $\{\,[(0,n)]_{\approx} \mid n \in \mathbf{N}, n \neq 0\,\}$ die **negativen ganzen Zahlen**.

Eine interessante Eigenschaft folgt direkt aus der Definition der Multiplikation ganzer Zahlen: Quadrate ganzer Zahlen ergeben immer positive ganze Zahlen:

$[(n,0)]_{\approx} \cdot_{\mathbf{Z}} [(n,0)]_{\approx} = [(n\cdot n,0)]_{\approx}$ und $[(0,n)]_{\approx} \cdot_{\mathbf{Z}} [(0,n)]_{\approx} = [(n\cdot n,0)]_{\approx}$.

Zur Vereinfachung wird im Folgenden

$$\begin{aligned}\mathbf{Z} &= \mathbf{N} \cup \{-n \mid n \in \mathbf{N}\}\\ &= \{0,1,-1,2,-2,3,-3,\ldots\}\end{aligned}$$

geschrieben. Die in **Z** definierte Addition $+_{\mathbf{Z}}$ bzw. die in **Z** definierte Multiplikation $\cdot_{\mathbf{Z}}$ wird wieder mit + bzw. $\cdot$ bezeichnet, die Ordnungsrelation $\leq_{\mathbf{Z}}$ vereinfacht mit $\leq$.

Z erlaubt bereits eine Vielzahl interessanter arithmetischer Operationen, jedoch ist „richtiges Rechnen“, d.h. auch **Division (Umkehrung der Multiplikation)** nicht immer möglich. Daher wird **Z** auf die **Menge der rationalen Zahlen** erweitert. Auch diese Erweiterung erfolgt wieder formal über die Definition einer geeigneten Äquivalenzrelation auf Paaren, dieses Mal von ganzen Zahlen, und Übergang auf die zugehörigen Äquivalenzklassen und Einbettung von **Z** in die Menge dieser Äquivalenzklassen:

Im Folgenden seien $a \in \mathbf{Z}, b \in \mathbf{Z}, c \in \mathbf{Z}, d \in \mathbf{Z}, b \neq 0$ und $d \neq 0$.

Man definiert auf der Menge $\mathbf{Z} \times \mathbf{Z}$ der Paare ganzer Zahlen eine Relation ~ durch

$(a,b) \sim (c,d)$ genau dann, wenn $a \cdot d = c \cdot b$ gilt (hier ist die definierte Multiplikation $\cdot$ auf den ganze Zahlen gemeint).

Die bezüglich dieser Äquivalenzrelation zum Paar (a,b) mit $b \neq 0$ ganzer Zahlen gehörende Äquivalenzklasse wird mit $\frac{a}{b}$ bezeichnet. Die rationalen Zahlen sind dann genau die Menge dieser Äquivalenzklassen:

$$\mathbf{Q} = \left\{ \frac{a}{b} \,\middle|\, a \in \mathbf{Z} \text{ und } b \in \mathbf{Z} \text{ und } b \neq 0 \right\}.$$

Mit Satz 1.4-2 (i) ist $\frac{a}{b} = \frac{c}{d}$ genau dann, wenn $a \cdot d = c \cdot b$ (in $\mathbf{Z}$) gilt.

Damit sind $\frac{a}{b} = \frac{(-a)}{(-b)}$, $\frac{0}{b} = \frac{0}{1}$ und $\frac{b}{b} = \frac{1}{1}$.

Die Menge der ganzen Zahlen ist in der Menge der rationalen Zahlen eingebettet:

$$\left\{ \frac{a}{1} \,\middle|\, a \in \mathbf{Z} \right\} \subset \mathbf{Q} \text{ und } \left\{ \frac{a}{1} \,\middle|\, a \in \mathbf{Z} \right\} \approx \mathbf{Z}.$$

Daher schreiben wir $\mathbf{Z} \subset \mathbf{Q}$ (obwohl auch diese Aussage wieder mathematisch nicht korrekt ist).

Auf $\mathbf{Q}$ lassen sich eine Addition $+_{\mathbf{Q}}$ und eine Multiplikation $\cdot_{\mathbf{Q}}$ definieren, die die entsprechenden Operationen auf $\mathbf{Z}$ fortsetzen:

Für $\frac{a}{b} \in \mathbf{Q}$ und $\frac{c}{d} \in \mathbf{Q}$ wird definiert:

$$\frac{a}{b} +_{\mathbf{Q}} \frac{c}{d} = \frac{a \cdot d + c \cdot b}{b \cdot d} \text{ und}$$
$$\frac{a}{b} \cdot_{\mathbf{Q}} \frac{c}{d} = \frac{a \cdot c}{b \cdot d}.$$

Die rationale Zahl $\frac{0}{1}$ ist neutrales Element der Addition, die rationale Zahl $\frac{1}{1}$ neutrales Element der Multiplikation:

$$\frac{a}{b} +_{\mathbf{Q}} \frac{0}{1} = \frac{a \cdot 1 + b \cdot 0}{b \cdot 1} = \frac{a}{b}, \quad \frac{a}{b} \cdot_{\mathbf{Q}} \frac{1}{1} = \frac{a \cdot 1}{b \cdot 1} = \frac{a}{b}.$$

Zu jeder rationalen Zahl $\frac{a}{b} \in \mathbf{Q}$ gibt es eine eindeutige **additiv inverse Zahl**, geschrieben $-\frac{a}{b}$, mit $\frac{a}{b} +_{\mathbf{Q}} -\frac{a}{b} = \frac{0}{1}$, nämlich $-\frac{a}{b} = \frac{-a}{b} = \frac{a}{-b}$:

$$\begin{aligned} \frac{a}{b} +_{\mathbf{Q}} -\frac{a}{b} &= \frac{a \cdot b + b \cdot (-a)}{b \cdot b} \\ &= \frac{b \cdot a + b \cdot (-a)}{b \cdot b} \\ &= \frac{b \cdot (a + (-a))}{b \cdot b} \\ &= \frac{0}{b \cdot b} = \frac{0}{1} \end{aligned}$$

Zu jeder rationalen Zahl $r = \frac{a}{b}$ mit $a \neq 0$ gibt es eine eindeutige **multiplikativ inverse Zahl**, geschrieben r^{-1}, mit $r \cdot_{\mathbf{Q}} r^{-1} = \frac{1}{1}$. Es ist $r^{-1} = \left(\frac{a}{b}\right)^{-1} = \frac{b}{a}$:

$$\frac{a}{b} \cdot_{\mathbf{Q}} \frac{b}{a} = \frac{a \cdot b}{b \cdot a} = \frac{a \cdot b}{a \cdot b} = \frac{1}{1}.$$

Unter der **Division** r/s in **Q** zweier Zahlen $r = \frac{a}{b}$ und $s = \frac{c}{d}$ mit $c \neq 0$ versteht man die Multiplikation von r mit s^{-1}:

$$r/s = r \cdot_{\mathbf{Q}} s^{-1} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a}{b} \cdot_{\mathbf{Q}} \left(\frac{c}{d}\right)^{-1} = \frac{a}{b} \cdot_{\mathbf{Q}} \frac{d}{c} = \frac{a \cdot d}{b \cdot c}.$$

Die Ordnungsrelation $\leq$ wird von **Z** auf **Q** erweitert, geschrieben $\leq_{\mathbf{Q}}$:

Es ist $\frac{0}{1} \leq_{\mathbf{Q}} \frac{a}{b}$ genau dann, wenn $((0 \leq a) \text{ und } (0 < b))$ oder $((a \leq 0) \text{ und } (b < 0))$ gelten. Außerdem ist $\frac{a}{b} \leq_{\mathbf{Q}} \frac{c}{d}$ genau dann, wenn $\frac{0}{1} \leq_{\mathbf{Q}} \frac{c}{d} +_{\mathbf{Q}} \left(-\frac{a}{b}\right)$ gilt.

Die **positiven rationalen Zahlen** (einschließlich $\frac{0}{1}$) in **Q** sind die Werte $\frac{a}{b}$ mit $\frac{0}{1} \leq_{\mathbf{Q}} \frac{a}{b}$, die **negativen rationalen Zahlen** die Werte $\frac{a}{b}$ mit $\frac{a}{b} \leq_{\mathbf{Q}} \frac{0}{1}$ und $a \neq 0$. Auch in den rationalen zahlen sind Quadrate positiv: $\frac{a}{b} \cdot_{\mathbf{Q}} \frac{a}{b} = \frac{a \cdot a}{b \cdot b}$, und es ist $0 \leq a \cdot a$ und $0 < b \cdot b$ (in **Z**, siehe oben).

Mit diesen Festlegungen bildet die algebraische Struktur $\left(\mathbf{Q}, +_{\mathbf{Q}}, \cdot_{\mathbf{Q}}, \frac{0}{1}, \frac{1}{1}\right)$ mit der Ordnungsrelation $\leq_{\mathbf{Q}}$ einen angeordneten Körper (der Nachweis der einzelnen Eigenschaften wird als Übung empfohlen).

Erläuterung:

Eine algebraische Struktur $(K, \oplus, \otimes)$ heißt **Körper**, wenn gilt:

(i) $(K, \oplus, \otimes, 0, 1)$ ist ein kommutativer Ring mit 1

(ii) für jedes $a \in K$ mit $a \neq 0$ gibt es ein Element $a^{-1} \in K$ mit $a \otimes a^{-1} = 1$; dieses Element heißt **multipliktives inverses Element** zu *a*.

$(K, \oplus, 0)$ ist also eine kommutative Gruppe, die additive Gruppe des Körpers,
$(K \setminus \{ 0 \}, \otimes, 1)$ ist eine kommutative Gruppe, die multiplikative Gruppe des Körpers, und es gelten die Distributivgesetze.

Der Körper $(K, \oplus, \otimes)$ heißt **angeordneter Körper**, wenn es eine totale Ordnungsrelation $\lhd$ auf *K* gibt mit folgenden Eigenschaften:

(i) für jedes $x \in K$, für jedes $y \in K$ und für jedes $z \in K$ gilt:
aus $x \lhd y$ folgt $x \oplus z \lhd y \oplus z$

(ii) für jedes $x \in K$ und für jedes $y \in K$ gilt:
aus $0 \lhd x$ und $0 \lhd y$ folgt $0 \lhd x \otimes y$.

Aus den Definitionen folgt

Satz 1.4-4:

Es sei $(K, \oplus, \otimes)$ ein angeordneter Körper mit der Ordnungsrelation $\lhd$ und $a \in K$. Das bezüglich der Operation $\oplus$ inverse Element wird wieder mit $-a$ und das bei $a \neq 0$ bezüglich der Operation $\otimes$ inverse Element mit a^{-1} bezeichnet. Dann gilt:

(i) Für $a \neq 0$ ist $(-a)^{-1} = -(a^{-1})$;
$(-1) \otimes (-1) = 1$;
diese Aussagen gelten in jedem Körper (unabhängig von einer Anordnung).

(ii) Bei $0 \lhd a$ ist $-a \lhd 0$; bei $a \lhd 0$ ist $0 \lhd -a$. Insbesondere ist $0 \lhd 1$ und $-1 \lhd 0$.

(iii) $0 \lhd a \otimes a$.

Für (i) wird zunächst $(-1) \otimes (-1) = 1$ gezeigt (hierzu wird die Anordnung von *K* nicht verwendet): Für $a \in K$ ist mit Satz 1.4-3 (ii)
$((-1) \otimes (-1)) \otimes a = (-1) \otimes ((-1) \otimes a) = (-1) \otimes (-a) = -(-a) = a$, also wegen der Eindeutigkeit des multiplikativ neutralen Elements $(-1) \otimes (-1) = 1$. Daraus ergibt sich, dass $-(a^{-1})$ multiplikativ invers zu $-a$ ist; folglich ist dann $(-a)^{-1} = -(a^{-1})$:
$-(a^{-1}) \otimes (-a) = (-1) \otimes a^{-1} \otimes (-1) \otimes a = (-1) \otimes (-1) \otimes a^{-1} \otimes a = 1$.

In (ii) folgt bei $0 \lhd a$: $-a = 0 \oplus (-a) \lhd a \oplus (-a) = 0$; entsprechend folgt bei $a \lhd 0$:
$0 = a \oplus (-a) \lhd 0 \oplus (-a) = -a$. Die Aussage $0 \lhd 1$ ergibt sich aus (iii) mit $a = 1$.

In (iii) folgt aus der Anordnungseigenschaft bei $0 \lhd a$ offensichtlich $0 \lhd a \otimes a$. Für $a \lhd 0$ ist $0 \lhd -a$ und $0 \lhd (-a) \otimes (-a) = (-1) \otimes a \otimes (-1) \otimes a = (-1) \otimes (-1) \otimes a \otimes a = a \otimes a$.

Zur Vereinfachung werde wieder die in **Q** definierte Addition $+_{\mathbf{Q}}$ bzw. die in **Q** definierte Multiplikation $\cdot_{\mathbf{Q}}$ mit + bzw. $\cdot$ bezeichnet, die Ordnungsrelation $\leq_{\mathbf{Q}}$ vereinfacht mit $\leq$. Außerdem wird anstelle von $\frac{a}{1}$ vereinfacht *a* geschrieben.

In $(\mathbf{Q}, +, \cdot, 0, 1)$ sind die wichtigsten arithmetischen Operationen möglich. Jedoch fehlt der Ordnungsrelation auf **Q** eine wichtige Eigenschaft, nämlich die Vollständigkeit. Beispielswei-

se sind die Elemente der Menge $\left\{ q \,\middle|\, q \in \mathbf{Q} \text{ und } \left(q \le 0 \text{ oder } q^2 \le 2\right) \right\}$ wohl nach oben beschränkt, z.B. durch $r = 3/2$, es gibt in **Q** aber keine *kleinste* obere Schranke für die Elemente dieser Menge (denn **Q** enthält kein Element, dessen Quadrat 2 ergibt). Daher wird die Menge der rationalen Zahlen so erweitert, dass die arithmetischen Operationen und die Ordnungsrelation von **Q** fortgesetzt werden und zusätzlich jede nichtleere nach oben beschränkte Menge eine kleinste obere Schranke besitzt. Im Beispiel der Menge
$\left\{ q \,\middle|\, q \in \mathbf{Q} \text{ und } \left(q \le 0 \text{ oder } q^2 \le 2\right) \right\}$ wird diese kleinste obere Schranke mit $\sqrt{2}$ bezeichnet.

Das Resultat ist der Körper $\left(\mathbf{R}, +, \cdot, 0, 1\right)$ der **reellen Zahlen**. Der Erweiterungsprozess kann auf verschiedene Weisen unter topologischen Aspekten vollzogen werden (z.B. Dedekind-Schnitte, mittels Fundamentalfolgen, Intervallschachtelung oder durch Dezimalbruchentwicklung). Exemplarisch wird hier die Methode der Dedekind-Schnitte angegeben:

Erläuterung:

Eine Teilmenge $S \subseteq \mathbf{Q}$ heißt (**Dedekind-) Schnitt**, wenn gilt:

(i) $S \neq \varnothing$ und $S \neq \mathbf{Q}$

(ii) Für jedes $r \in S$ ist die Menge $\left\{ q \,\middle|\, q \in \mathbf{Q} \text{ und } q \le r \right\}$ eine echte Teilmenge von *S*.

Bedingung (ii) beinhaltet zwei Eigenschaften:

(i') Ist $r \in S$ und $q \in \mathbf{Q}$ mit $q \le r$, so ist auch $q \in S$ (**Abgeschlossenheit nach unten**)

(ii') Ist $r \in S$, so gibt es ein $p \in S$ mit $r < p$ (**Nichtexistenz eines Maximums**).

Jeder Schnitt *S* ist nach oben beschränkt: Denn wäre *S* nicht nach oben beschränkt, so gäbe es für jedes $q \in \mathbf{Q}$ ein $r \in S$ mit $q \le r$. Mit (i') folgt $q \in S$, also $\mathbf{Q} \subseteq S$ und damit $S = \mathbf{Q}$, was aber gemäß (i) ausgeschlossen ist.

Die Menge **R** der reellen Zahlen ist die Menge aller (Dedekind-) Schnitte.

Im Folgenden werden Schnitte mit kleinen griechischen Buchstaben (α, β, …) bezeichnet.

Die Menge der rationalen Zahlen lässt sich in **R** einbetten: Mit $r \in \mathbf{Q}$ wird der Schnitt $\{x \mid x \in \mathbf{Q} \text{ und } x < r\}$ identifiziert:

$\{x \mid x \in \mathbf{Q} \text{ und } x < r\} \subseteq \mathbf{R}$ und $\{x \mid x \in \mathbf{Q} \text{ und } x < r\} \approx \mathbf{Q}$.

Dass es sich für jedes $r \in \mathbf{Q}$ bei $\{x \mid x \in \mathbf{Q} \text{ und } x < r\}$ um einen Schnitt handelt, ist klar: offensichtlich gilt (i); für den Nachweis von (ii) sei $p \in \{x \mid x \in \mathbf{Q} \text{ und } x < r\}$, dann ist $\{q \mid q \in \mathbf{Q} \text{ und } q \le p\}$ eine echte Teilmenge von $\{x \mid x \in \mathbf{Q} \text{ und } x < r\}$, da etwa $\frac{p+r}{2} \in \{x \mid x \in \mathbf{Q} \text{ und } x < r\}$, aber $\frac{p+r}{2} \notin \{q \mid q \in \mathbf{Q} \text{ und } q \le p\}$ gilt.

Daher schreiben wir $\mathbf{Q} \subset \mathbf{R}$ (obwohl diese Aussage mathematisch nicht korrekt ist).

Sind $\alpha \in \mathbf{R}$ und $\beta \in \mathbf{R}$ verschiedene Schnitte, so gibt es ein $q \in \mathbf{Q}$ mit $q \in \beta \setminus \alpha$ oder $q \in \alpha \setminus \beta$. Im ersten Fall ist gemäß (ii) $\{r \mid r \in \mathbf{Q} \text{ und } r \le q\} \subset \beta$; außerdem ist $\alpha \subseteq \{r \mid r \in \mathbf{Q} \text{ und } r \le q\}$ (denn mit $s \in \alpha$ und $s > q$ ist nach (ii) auch $q \in \alpha$, was aber im ersten Fall gerade nicht gilt); daher ist $\alpha \subset \beta$. Im zweiten Fall folgt entsprechend $\beta \subset \alpha$.

Insgesamt wird auf der Menge der Schnitte, d.h. den reellen Zahlen, durch die Mengeninklusion $\subseteq$ eine totale Ordnungsrelation $\le_{\mathbf{R}}$ definiert, die die Ordnungsrelation $\le$ auf **Q** fortsetzt, d.h. für $r_1 \in \mathbf{Q}$ und $r_2 \in \mathbf{Q}$ mit $r_1 < r_2$ ist $\{x \mid x \in \mathbf{Q} \text{ und } x < r_1\} <_{\mathbf{R}} \{x \mid x \in \mathbf{Q} \text{ und } x < r_2\}$.

Diese Fortsetzung der Ordnungsrelation auf die Schnitte in **Q**, d.h. auf die reellen Zahlen, hat die Eigenschaft, dass jede nichtleere nach oben beschränkt Teilmenge reeller Zahlen eine bezüglich $\le_{\mathbf{R}}$ kleinste obere Schranke (**Supremum**) in **R** besitzt (**Vollständigkeit der Ordnungsrelation**).

Diese Eigenschaft sieht man wie folgt:
Es sei $B \subseteq \mathbf{R}$ nichtleer und nach oben beschränkt, etwa durch $\gamma \in \mathbf{R}$. Dann gilt für jedes $\beta \in B$: $\beta \le_{\mathbf{R}} \gamma$ bzw. definitionsgemäß $\beta \subseteq \gamma$. Es sei $b = \bigcup_{\beta \in B} \beta$. Offensichtlich ist $b \subseteq \gamma$ und $b \neq \varnothing$ und $b \neq \mathbf{Q}$. Die Menge b erfüllt also Bedingung (i) in der Definition eines Schnitts.
Bedingung (ii) wird auch erfüllt: Sei $r \in b$, d.h. $r \in \beta$ für ein $\beta \in B$. Da für β Bedingung (ii) gilt, ist $\{q \mid q \in \mathbf{Q} \text{ und } q \le r\}$ eine echte Teilmenge von β und damit von b.
Die reelle Zahl b ist eine obere Schranke für B: Denn ist σ eine weitere obere Schranke für B und $x \in b$, dann gibt es ein $\beta \in B$ mit $x \in \beta$. Wegen $\beta \subseteq \sigma$ gilt $x \in \sigma$. Damit ist $b \subseteq \sigma$, d.h.

mit b ist eine kleinste obere Schranke von B gefunden, die selbst wieder ein Schnitt, d.h. eine reelle Zahl ist.

Damit hat der Schnitt $\{q \mid q \in \mathbf{Q} \text{ mit } q \le 0\} \cup \{q \mid q \in \mathbf{Q} \text{ mit } q > 0 \text{ und } q^2 < 2\}$ eine kleinste obere Schranke, die allerdings nicht in **Q** liegt, nämlich die als $\sqrt{2}$ bezeichnete reelle Zahl.

Entsprechend lässt sich zeigen, dass jede nichtleere nach unten beschränkt Teilmenge reeller Zahlen eine bezüglich $\le_{\mathbf{R}}$ größte untere Schranke (**Infimum**) in **R** besitzt.

Statt $\le_{\mathbf{R}}$ wird im Folgenden $\le$ geschrieben.

Die Operationen der Addition und der Multiplikation lassen sich von **Q** auf die Menge der Schnitte, d.h. **R**, fortsetzen, so dass insgesamt $(\mathbf{R}, +_{\mathbf{R}}, \cdot_{\mathbf{R}}, 0_{\mathbf{R}}, 1_{\mathbf{R}})$ zu einem vollständig angeordneten Körper wird (einige Details des Nachweises dieser Behauptung werden hier aus Platzgründen nicht dargestellt):

Die Addition $+_{\mathbf{R}}$ zweier reeller Zahlen (Schnitte) α und β wird definiert durch
$\alpha +_{\mathbf{R}} \beta = \{x + y \mid x \in \alpha \text{ und } y \in \beta\}$. Mit $x + y$ ist hier die Addition in **Q** gemeint.

Es lässt sich zeigen, dass hierdurch wieder ein Schnitt definiert wird, d.h. dass die obigen Bedingungen (i) und (ii) erfüllt sind. Außerdem bildet $(\mathbf{R}, +_{\mathbf{R}}, 0_{\mathbf{R}})$ eine kommutative Gruppe mit neutralem Element (der Addition) $0_{\mathbf{R}} = \{x \mid x \in \mathbf{Q} \text{ und } x < 0\}$.

Dass $0_{\mathbf{R}}$ das neutrale Element der Addition ist, d.h. dass für jedes $\alpha \in \mathbf{R}$ die Gleichung $\alpha +_{\mathbf{R}} 0_{\mathbf{R}} = \alpha$ gilt, sieht man wie folgt:
Ist $x \in \alpha +_{\mathbf{R}} 0_{\mathbf{R}}$, dann hat x die Form $x = a + s$ mit $a \in \alpha$ und $s < 0$. Insbesondere ist $x = a + s < a$, und mit (i') folgt $x \in \alpha$.
Ist umgekehrt $x \in \alpha$, so ist wegen (ii) die Menge $\{q \mid q \in \mathbf{Q} \text{ und } q \le x\}$ eine echte Teilmenge von α. Das bedeutet die Existenz eines Elements $x_1 \in \alpha$ mit $x_1 > x$. Es ist $x - x_1 < 0$, d.h. $x - x_1 \in 0_{\mathbf{R}}$, und $x_1 + (x - x_1) \in \alpha +_{\mathbf{R}} 0_{\mathbf{R}}$, d.h. $x \in \alpha +_{\mathbf{R}} 0_{\mathbf{R}}$.

Das additiv inverse Element zu $\alpha \in \mathbf{R}$ ist
$-\alpha = \{p \mid p \in \mathbf{Q} \text{ und es gibt } r \in \mathbf{Q} \text{ mit } r > 0 \text{ und } -p - r \notin \alpha\}$.
Es gilt nämlich $\alpha +_{\mathbf{R}} (-\alpha) = 0_{\mathbf{R}}$:
Es sei $x \in \alpha +_{\mathbf{R}} (-\alpha)$, also $x = a + p$ mit $a \in \alpha$ und $p \in -\alpha$. Zu p gibt es ein $r \in \mathbf{Q}$ mit $r > 0$ und $-p - r \notin \alpha$. Wäre $a + p \ge 0$, so ist $a \ge -p$ und mit (i') $-p \in \alpha$. Wegen $-p - r < -p$ ergibt sich der Widerspruch $-p - r \in \alpha$. Daher gilt $a + p < 0$, d.h. $x \in 0_{\mathbf{R}}$.

Es sei umgekehrt $x \in 0_{\mathbf{R}}$, d.h. $x < 0$. Wegen (i) existiert ein $y \in \mathbf{Q}$ mit $y \notin \alpha$, und es existiert ein $z \in \alpha$. Die Menge $M = \{ m \mid m \in \mathbf{N} \text{ und } y + m \cdot x \in \alpha \}$ ist nicht leer (denn für $m > \frac{z-y}{x}$ gilt $y + m \cdot x < z$; man beachte, dass $x < 0$ ist und sich dadurch die Ungleichung „umdreht"; mit (i') folgt $y + m \cdot x \in \alpha$). Folglich hat M als Menge aus natürlichen Zahlen ein kleinstes Element $n_0 \in M$. Dann gilt $y + n_0 \cdot x \in \alpha$ und $y + (n_0 - 1) \cdot x \notin \alpha$. Ist $x_0 = y + (n_0 - 1) \cdot x$ nicht das kleinste Element in $\mathbf{Q} \setminus \alpha$, dann gibt es $r > 0$ mit $x_0 - r \notin \alpha$. Das bedeutet $-x_0 \in (-\alpha)$. Der Wert $x_0 + x = y + (n_0 - 1) \cdot x + x = y + n_0 \cdot x$ liegt in α und damit $(x_0 + x) - x_0 = x$ in $\alpha +_{\mathbf{R}} (-\alpha)$. Ist $y + (n_0 - 1) \cdot x$ das kleinste Element in $\mathbf{Q} \setminus \alpha$, so setzt man $y' = y - x/2$. Da $x < 0$ gilt, ist $y' > y$. Außerdem ist $y' + n_0 \cdot x \in \alpha$ und $y' + (n_0 - 1) \cdot x \notin \alpha$, und $y' + (n_0 - 1) \cdot x$ ist nicht kleinstes Element in $\mathbf{Q} \setminus \alpha$. Wie oben zeigt man $x \in \alpha +_{\mathbf{R}} (-\alpha)$.

Die Multiplikation auf **R** wird folgendermaßen definiert:

Für $\alpha > 0_{\mathbf{R}}$ und $\beta > 0_{\mathbf{R}}$ ist $\alpha \cdot_{\mathbf{R}} \beta = \left\{ p \;\middle|\; \begin{array}{l} p \in \mathbf{Q} \text{ und es gibt } r \in \alpha \text{ mit } r > 0 \\ \text{und es gibt } s \in \beta \text{ mit } s > 0 \text{ und } p \le r \cdot s \end{array} \right\}$; mit $r \cdot s$ ist hier die Multiplikation in **Q** gemeint. Diese Multiplikation wird auf ganz **R** fortgesetzt durch $\alpha \cdot_{\mathbf{R}} 0_{\mathbf{R}} = 0_{\mathbf{R}} \cdot_{\mathbf{R}} \alpha = 0_{\mathbf{R}}$ und

$$\alpha \cdot_{\mathbf{R}} \beta = \begin{cases} (-\alpha) \cdot_{\mathbf{R}} (-\beta) & \text{für } \alpha < 0_{\mathbf{R}} \text{ und } \beta < 0_{\mathbf{R}} \\ -((-\alpha) \cdot_{\mathbf{R}} \beta) & \text{für } \alpha < 0_{\mathbf{R}} \text{ und } \beta > 0_{\mathbf{R}} \\ -(\alpha \cdot_{\mathbf{R}} (-\beta)) & \text{für } \alpha > 0_{\mathbf{R}} \text{ und } \beta < 0_{\mathbf{R}} \end{cases}.$$

Auch hier lässt sich zeigen, dass $\alpha \cdot_{\mathbf{R}} \beta$ wieder ein Schnitt, d.h. eine reelle Zahl ist.

Das neutrale Element der Multiplikation ist $1_{\mathbf{R}} = \{ x \mid x \in \mathbf{Q} \text{ und } x < 1 \}$.

Exemplarisch wird hier $\alpha \cdot_{\mathbf{R}} 1_{\mathbf{R}} = \alpha$ für $\alpha > 0_{\mathbf{R}}$ gezeigt:
Nach Definition ist $0_{\mathbf{R}} < 1_{\mathbf{R}}$. Ist $x \in \alpha \cdot_{\mathbf{R}} 1_{\mathbf{R}}$, dann gibt es $r \in \alpha$ mit $r > 0$ und $s > 0$ mit $s < 1$ und $x \le r \cdot s < r$. Mit (i') folgt $x \in \alpha$.
Es sei umgekehrt $x \in \alpha$. Ist $x > 0$, so ist wegen (ii) die Menge $\{ q \mid q \in \mathbf{Q} \text{ und } q \le x \}$ eine echte Teilmenge von α. Das bedeutet die Existenz eines Elements $x_1 \in \alpha$ mit $x_1 > x > 0$. Für $r = x/x_1$ ist $0 < r < 1$, insbesondere $r \in 1_{\mathbf{R}}$, und $x = x_1 \cdot r$, also $x \in \alpha \cdot_{\mathbf{R}} 1_{\mathbf{R}}$. Ist $x \le 0$, so gibt es wegen $\alpha > 0_{\mathbf{R}}$ ein $x' \in \alpha$ mit $x' > 0$ (man beachte, dass hier auch (ii) verwendet wurde). Hiermit folgt $x' \in \alpha \cdot_{\mathbf{R}} 1_{\mathbf{R}}$. Nach (ii) ist $\{ q \mid q \in \mathbf{Q} \text{ und } q \le x' \} \subset \alpha \cdot_{\mathbf{R}} 1_{\mathbf{R}}$ und wegen $x \in \{ q \mid q \in \mathbf{Q} \text{ und } q \le x' \}$ ebenfalls $x \in \alpha \cdot_{\mathbf{R}} 1_{\mathbf{R}}$.

Zu $\alpha > 0_{\mathbf{R}}$ invers bezüglich der Multiplikation ist der Schnitt

$\left\{ p \,\middle|\, \left(p \in \mathbf{Q} \text{ und } p \le 0\right) \text{oder} \left(p \in \mathbf{Q} \text{ und } p > 0 \text{ und } \frac{1}{p} \notin \alpha \right) \right\}$, der mit α^{-1} oder mit $1/\alpha$ bezeichnet wird.

Zu $\alpha < 0_{\mathbf{R}}$ invers bezüglich der Multiplikation ist $-(-\alpha)^{-1} = -\left(1/(-\alpha)\right)$.

Dazu soll exemplarisch für $\alpha > 0_{\mathbf{R}}$ die Gleichung $\alpha \cdot_{\mathbf{R}} 1/\alpha = 1_{\mathbf{R}}$ gezeigt werden:

Nach Definition ist $0_{\mathbf{R}} < 1/\alpha$.

Es sei $x \in \alpha \cdot_{\mathbf{R}} \frac{1}{\alpha}$. Dann gibt es $a \in \alpha$ mit $a > 0$ und es gibt $p > 0$ mit $\frac{1}{p} \notin \alpha$ und $x \le a \cdot p$.

Für $x \le 0$ ist $x \in 1_{\mathbf{R}}$. Ist $x > 0$, dann ist $p < \frac{1}{a}$; denn wäre $p \ge \frac{1}{a}$, so wäre $\frac{1}{p} \le a$ und damit $\frac{1}{p} \in \alpha$. Damit ergibt sich $x \le a \cdot p < a \cdot \frac{1}{a} = 1$, also $x \in 1_{\mathbf{R}}$.

Sei umgekehrt $x \in 1_{\mathbf{R}}$, d.h. $x < 1$. Es wird der Fall $x = 1 - \delta$ mit $0 < \delta < 1$ betrachtet. Für ein derartiges x wird $x \in \alpha \cdot_{\mathbf{R}} 1/\alpha$ gezeigt (ist $x \in 1_{\mathbf{R}}$ mit $x \le 1 - \delta$, so ist wegen (ii) dann ebenfalls $x \in \alpha \cdot_{\mathbf{R}} 1/\alpha$):

Da $\alpha \neq \varnothing$ ist, gibt es $a \in \alpha$, und man kann wegen $\alpha > 0_{\mathbf{R}}$ mit (ii) annehmen, dass $0 < a$ gilt. Es wird $b = \frac{\delta \cdot a \cdot (1-\delta)}{2-\delta} > 0$ gesetzt. Wegen $b < a$ ist $b \in \alpha$. Da α beschränkt ist, gibt es eine kleinste natürliche Zahl n_0 mit $\frac{(n_0+1)\cdot b}{1-\delta} \notin \alpha$. Es sei $r = \frac{n_0 \cdot b}{1-\delta}$, also $r \in \alpha$. Außerdem ist $n_0 > \frac{2\cdot(1-\delta)}{\delta} > 1$ (denn andernfalls wäre $\frac{(n_0+1)\cdot b}{1-\delta} \le \left(\frac{2\cdot(1-\delta)}{\delta} + 1\right) \cdot \frac{b}{(1-\delta)} = a$ im Widerspruch zu $\frac{(n_0+1)\cdot b}{1-\delta} \notin \alpha$). Damit ergibt sich $\frac{b}{1-\delta} < r \cdot \frac{\delta/2}{1-\delta}$. Mit dieser Abschätzung folgt $\frac{(n_0+1)\cdot b}{1-\delta} = r + \frac{b}{1-\delta} < r \cdot \left(1 + \frac{\delta/2}{1-\delta}\right) = r \cdot \frac{1-\delta/2}{1-\delta} < \frac{r}{1-\delta}$. Man sieht, dass $\frac{r}{1-\delta} \notin \alpha$, also $\frac{1-\delta}{r} \in 1/\alpha$ ist. Damit ist $x = r \cdot \frac{1-\delta}{r}$ und $x \in \alpha \cdot_{\mathbf{R}} 1/\alpha$.

Hiermit ist $(\mathbf{R} \setminus \{0_{\mathbf{R}}\}, \cdot_{\mathbf{R}}, 1_{\mathbf{R}})$ eine kommutative Gruppe.

Bezüglich Addition $+_{\mathbf{R}}$ und Multiplikation $\cdot_{\mathbf{R}}$ gelten die Distributivgesetze, so dass $(\mathbf{R}, +_{\mathbf{R}}, \cdot_{\mathbf{R}}, 0_{\mathbf{R}}, 1_{\mathbf{R}})$ ein Körper ist. Mit der Ordnungsrelation $\le_{\mathbf{R}}$ bzw. $\subseteq$ wird $(\mathbf{R}, +_{\mathbf{R}}, \cdot_{\mathbf{R}})$ zu

einem vollständig angeordneten Körper. Es lässt sich zeigen, dass er strukturell der einzige angeordnete Körper ist.

Elemente $\alpha \in \mathbf{R}$ mit $\alpha <_{\mathbf{R}} 0_{\mathbf{R}}$ heißen **negative reelle Zahlen**, Elemente $\alpha \in \mathbf{R}$ mit $0_{\mathbf{R}} <_{\mathbf{R}} \alpha$ **positive reelle Zahlen**.

Zur Vereinfachung wird für den Körper der reellen Zahlen im Folgenden wieder $(\mathbf{R}, +, \cdot, 0, 1)$ geschrieben, und es wird mit reellen Zahlen so gerechnet, wie man es aus der Schule gewohnt ist.

Die Menge **R** \ **Q** wird als die **Menge der irrationalen Zahlen** bezeichnet. Eine irrationale Zahl heißt **algebraische Zahl**, wenn sie Lösung einer Gleichung der Form

$a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + \ldots a_1 \cdot x + a_0 = 0$ ist, wobei $n \geq 1$, $a_i \in \mathbf{Z}$ für $i = 1, \ldots, n$ und $a_n \neq 0$ gelten.

Beispielsweise ist $\sqrt{2}$ als Lösung der Gleichung $x^2 - 2 = 0$ eine algebraische Zahl.

Eine irrationale Zahl, die nicht algebraisch ist, heißt **transzendente Zahl**.

Zu den transzendenten Zahlen gehören π und *e*.

Leider sind auch in $(\mathbf{R}, +, \cdot, 0, 1)$ noch nicht alle arithmetischen Operationen möglich. So besitzt die Gleichung $x^2 + 1 = 0$ keine Lösung $x \in \mathbf{R}$. Daher erweitert man den Zahlbereich **R** (unter Wahrung der arithmetischen Operationen):

Die **imaginäre Zahl** *i* wird durch die Eigenschaft $i^2 = -1$ definiert. Dann ist die **Menge der komplexen Zahlen** definiert durch

$$\mathbf{C} = \{\, a + b \cdot i \mid a \in \mathbf{R} \text{ und } b \in \mathbf{R} \,\}.$$

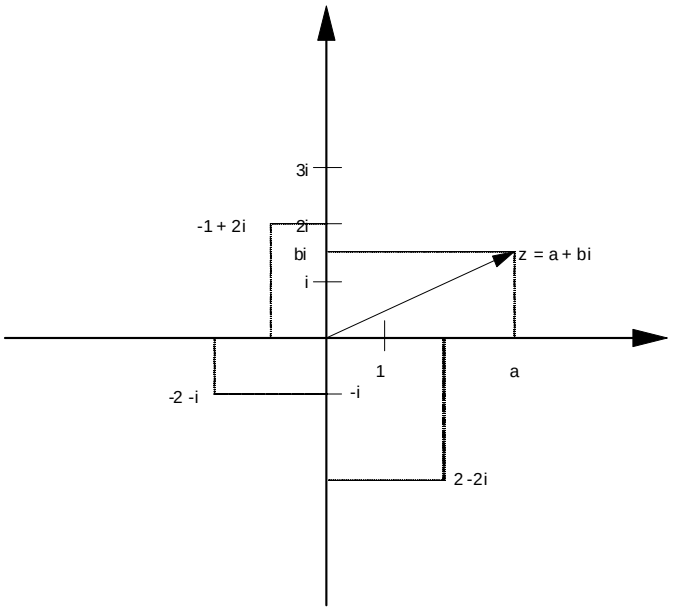
In der Darstellung einer komplexen Zahl in der Form $z = a + b \cdot i$ dienen die arithmetischen Symbole „+“ bzw. „$\cdot$“ lediglich als Trennzeichen zwischen den Komponenten *a* und $b \cdot i$ bzw. *b* und *i*. Es werden keine arithmetischen Operationen ausgeführt. Für $z = a + b \cdot i$ heißt *a* der **Realteil** und *b* der **Imaginärteil** von *z*.

Die Menge der reellen Zahlen ist in der Menge der komplexen Zahlen eingebettet:

$\{\ r+0\cdot i \mid\ r\in \mathbf{R}\ \}\subset \mathbf{C}$ und $\{\ r+0\cdot i \mid\ r\in \mathbf{R}\ \}\approx \mathbf{R}$.

Daher schreiben wir $\mathbf{R}\subset \mathbf{C}$ (obwohl auch diese Aussage mathematisch nicht korrekt ist).

Die komplexen Zahlen lassen sich als Punkte in einer Ebene mit rechtwinkligem Koordinatensystem, der **komplexen Ebene**, darstellen. Dabei wird für eine komplexe Zahl $z=a+bi$ ihr Realteil *a* auf der horizontalen Achse abgetragen, ihr Imaginärteil *b* auf der vertikalen Achse. Die folgende Abbildung zeigt die komplexen Zahlen $z=a+b\cdot i$, $-1+2\cdot i$, $-2-i$ und $2-2\cdot i$.



Der **Betrag** $|z|$ **der komplexen Zahl** $z=a+b\cdot i$ ist geometrisch durch die Länge der Verbindungslinie des Punkts (0, 0) der komplexen Ebene mit dem Punkt (*a*, *b*) definiert:

$$|z| = |\ a+b\cdot i\ | = \sqrt{a^2+b^2}\ .$$

Die arithmetischen Operationen $+_{\mathbf{C}}$ (Addition) und $\cdot_{\mathbf{C}}$ (Multiplikation) auf den komplexen Zahlen werden definiert durch

$$(a+b\cdot i)+_{\mathbf{C}}(c+d\cdot i)=(a+c)+(b+d)\cdot i \text{ und}$$
$$(a+b\cdot i)\cdot_{\mathbf{C}}(c+d\cdot i)=(a\cdot c-b\cdot d)+(a\cdot d+b\cdot c)\cdot i\ .$$

In $(a+c)+(b+d)\cdot i$ bzw. $(a\cdot c-b\cdot d)+(a\cdot d+b\cdot c)\cdot i$ bedeuten „$a+c$“ und „$b+d$“ bzw. „$a\cdot c-b\cdot d$“ und „$a\cdot d+b\cdot c$“ arithmetische Operationen in den reellen Zahlen.

Das neutrale Element der Addition ist die komplexe Zahl $0+0\cdot i$, das neutrale Element der Multiplikation ist $1+0\cdot i$.

Wie man leicht nachrechnet, ist zu einer komplexen Zahl $a+b\cdot i$ additiv invers die komplexe Zahl $-(a+b\cdot i)=-a+(-b)\cdot i$. Multiplikativ invers ist $(a+b\cdot i)^{-1}=\frac{a}{a^2+b^2}+\frac{-b}{a^2+b^2}\cdot i$:

$$\begin{aligned}(a+b\cdot i)\cdot_{\mathbf{C}}(a+b\cdot i)^{-1} &= (a+b\cdot i)\cdot_{\mathbf{C}}\left(\frac{a}{a^2+b^2}+\frac{-b}{a^2+b^2}\cdot i\right)\\ &= \left(\frac{a^2}{a^2+b^2}-\frac{-b^2}{a^2+b^2}\right)+\left(\frac{-a\cdot b}{a^2+b^2}+\frac{b\cdot a}{a^2+b^2}\right)\cdot i\\ &= 1+0\cdot i \quad .\end{aligned}$$

Die Division zweier komplexer Zahlen $a+b\cdot i$ und $c+d\cdot i$ mit $c\neq 0$ oder $d\neq 0$ wird auf die Multiplikation zurückgeführt:

$$\begin{aligned}(a+b\cdot i)/_{\mathbf{C}}(c+d\cdot i) &= (a+b\cdot i)\cdot_{\mathbf{C}}(c+d\cdot i)^{-1}\\ &= (a+b\cdot i)\cdot_{\mathbf{C}}\left(\frac{c}{c^2+d^2}+\frac{-d}{c^2+d^2}\cdot i\right)\\ &= \frac{a\cdot c+b\cdot d}{c^2+d^2}+\frac{b\cdot c+a\cdot d}{c^2+d^2}\cdot i \quad .\end{aligned}$$

Mit diesen Operationen bildet auch $(\mathbf{C},+_{\mathbf{C}},\cdot_{\mathbf{C}},0,1)$ einen Körper. Auch hier wird wieder zur Vereinfachung der Schreibweise an den arithmetischen Operationen das Subskript weggelassen.

Zu beachten ist, dass die Ordnungsrelation der reellen Zahlen, die $(\mathbf{R},+,\cdot,0,1)$ zu einem vollständig angeordneten Körper macht, nicht auf die komplexen Zahlen fortgesetzt wird; denn der Körper $(\mathbf{C},+_{\mathbf{C}},\cdot_{\mathbf{C}},0,1)$ lässt sich nicht anordnen: In einem angeordneten Körper K gilt nach Satz 1.4-4 (iii) $0\lhd a\otimes a$ für jedes $a\in K$, d.h. Quadrate sind positiv; in $\mathbf{C}$ ist nach Definition $i^2=(0+1\cdot i)\cdot_{\mathbf{C}}(0+1\cdot i)=-1+0\cdot i$, und $-1+0\cdot i=-1$ als Element in $\mathbf{R}$ ist negativ.

Insgesamt gilt (mathematisch nicht korrekt): $\mathbf{N}\subset\mathbf{Z}\subset\mathbf{Q}\subset\mathbf{R}\subset\mathbf{C}$. Die jeweiligen arithmetischen Operationen $+,-,\cdot,/$ und die Ordnungsrelation $\leq$, soweit sie in den einzelnen Zahlensystemen überhaupt definiert sind, werden für alle Zahlensysteme gleich bezeichnet.

1.5 Vollständige Induktion

Das vorliegende Kapitel behandelt eine der wichtigsten Beweismethoden, wenn es um Aussagen über natürliche Zahlen oder um Aussagen über Mengen geht, die strukturell äquivalent zu den natürlichen Zahlen sind: die vollständige Induktion. Wegen der großen Bedeutung dieser Methode für die Informatik werden in diesem Kapitel ausnahmsweise die durchgeführten Beweise in den Beispielen explizit angegeben.

Es sei $A(n)$ eine Aussage über die natürliche Zahl $n \in \mathbf{N}$, d.h. die von n abhängt. Es soll gezeigt werden, dass diese Aussage für alle natürlichen Zahlen $n \geq n_0$ gilt.

Häufig ist $n_0 = 0$; dann soll $A(n)$ für alle natürlichen Zahlen n bewiesen werden.

Nach der **Beweismethode der vollständigen Induktion** geht man wie folgt vor:

1. Man zeigt die Gültigkeit der Aussage $A(n_0)$ (**Induktionsanfang**)

2. Man beweist die Gültigkeit der Implikation $(A(n) \Rightarrow A(n+1))$ (**Induktionsschluss**).

Aus Axiom 5 der natürlichen Zahlen in Kapitel 1.4 („Eine Menge M natürlicher Zahlen, die die Zahl 0 und mit jeder Zahl n auch ihren Nachfolger n' enthält, ist mit **N** identisch.“) folgt dann, dass $A(n)$ für alle natürlichen Zahlen gilt. Hier soll nur der Spezialfall $n_0 = 0$ gezeigt werden. Dazu setzt man $M = \{ m \mid A(m) \text{ gilt für } m \}$. Der Induktionsanfang besagt $n_0 \in M$; der Induktionsschluss besagt: wenn $n \in M$ ist, dann ist auch der Nachfolger $n' \in M$; Axiom 5 besagt nun gerade, dass $M = \mathbf{N}$ gilt, d.h. dass $A(n)$ für alle natürlichen Zahlen n gilt[2].

Die Beweismethode der vollständigen Induktion wird in unterschiedlichen Varianten an einigen Beispielen erläutert:

Beispiel:

Zu beweisen ist die Aussage

[2] Der Fall $n_0 > 0$ verläuft analog, indem man eine Variante (Folgerung) von Axiom 5 verwendet, nämlich: „Eine Menge M natürlicher Zahlen, die die Zahl $n_0 \in \mathbf{N}$ und mit jeder Zahl n auch ihren Nachfolger n' enthält, ist mit $\{ n \mid n \in \mathbf{N} \text{ und } n \geq n_0 \}$ identisch.“.

Für alle $n \in \mathbf{N}$ gilt: $0+1+2+...+n = \frac{n \cdot (n+1)}{2}$.

Induktionsanfang: Die Aussage gilt für $n = 0$, denn auf der linken Seite des Gleichheitszeichens steht nur 0, und auf der rechten Seite des Gleichheitszeichens steht $\frac{0 \cdot 1}{2}$; beide Seiten ergeben denselben Wert.

Induktionsschluss: Für $n \in \mathbf{N}$ gelte: $0+1+2+...+n = \frac{n \cdot (n+1)}{2}$. Zu zeigen ist, dass dann diese Formel auch für die natürliche Zahl $n + 1$ gilt. Das lässt sich aber leicht nachrechnen: Auf der linken Seite des Gleichheitszeichens steht

$$0+1+2+...+n+(n+1).$$

Für diese Summe gilt (da die Formel für n als gültig vorausgesetzt wird):

$$0+1+2+...+n+(n+1) = \frac{n \cdot (n+1)}{2} + (n+1) = \frac{n \cdot (n+1) + 2 \cdot (n+1)}{2} = \frac{n^2 + 3 \cdot n + 2}{2};$$

auf der rechten Seite des Gleichheitszeichens steht $\frac{(n+1) \cdot (n+2)}{2} = \frac{n^2 + 3 \cdot n + 2}{2}$;

beide Seiten sind gleich, also gilt die Formel auch für die natürliche Zahl $n + 1$.

Beispiel:

Zu beweisen ist die Aussage

Ist A eine endliche Menge mit n Elementen, dann enthält die Potenzmenge $\mathbf{P}(A)$ 2^n viele Elemente (d.h. eine Menge mit n Elementen besitzt 2^n viele Teilmengen).

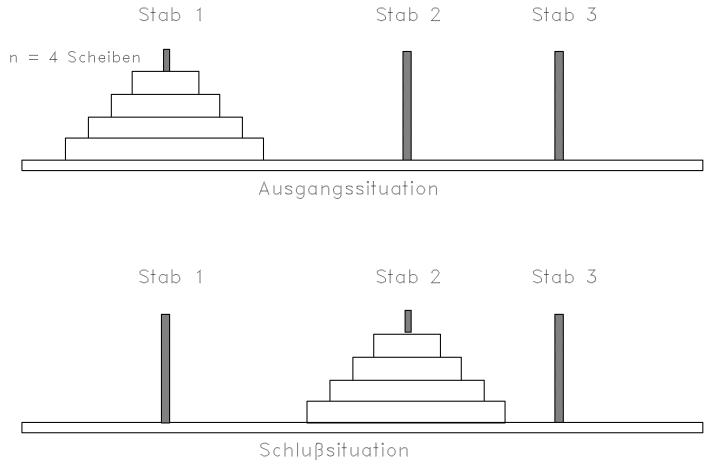
Induktionsanfang: Die Aussage gilt für $n = 0$, denn wenn A kein Element besitzt, dann ist $A = \varnothing$; andererseits ist $\mathbf{P}(\varnothing) = \{B | B \subseteq \varnothing\} = \{\varnothing\}$, und diese Menge enthält $1 = 2^0$ viele Elemente.

Induktionsschluss: Die Aussage gelte für Mengen A mit n Elementen, d.h. $|\mathbf{P}(A)| = 2^{|A|} = 2^n$. Es sei B eine Menge mit $n+1$ Elementen. Zu zeigen ist, dass die Anzahl der Teilmengen von B gleich 2^{n+1} ist: Da $|B| = n+1 > 0$ ist, gibt es ein Element $b \in B$. Die Teil-

mengen C von B werden danach klassifiziert, ob sie das Element b enthalten oder nicht. Jede Teilmenge $C \subseteq B$, mit $b \notin C$ ist Teilmenge von $B \setminus \{ b \}$. Da diese Menge genau n Elemente enthält, gibt es 2^n viele Teilmengen $C \subseteq B$ mit $b \notin C$. Jeder Teilmenge $C \subseteq B$ mit $b \in C$ entspricht genau eine Teilmenge $C' \subseteq B$ mit $b \notin C'$, nämlich $C' = C \setminus \{ b \}$. Umgekehrt entspricht jeder Teilmenge $C' \subseteq B$ mit $b \notin C'$ eine Teilmenge $C \subseteq B$ mit $b \in C$, nämlich $C = C' \cup \{ b \}$. Daher gibt es genauso viele Teilmengen $C \subseteq B$ mit $b \in C$ wie Teilmengen $C \subseteq B$ mit $b \notin C$, nämlich 2^n viele. Insgesamt ist die Anzahl an Teilmengen von B gleich $2 \cdot 2^n = 2^{n+1}$.

Beispiel:

Beim Spiel der **Türme von Hanoi** sind drei Stapel mit Bezeichnern A, B und C gegeben. Auf dem Stapel A befinden sich n Scheiben, die übereinander in absteigender Größe liegen; die beiden anderen Stapel sind leer. Die Aufgabe besteht darin, die Scheiben vom Ausgangsstapel A auf einen der anderen Stapel, etwa B, zu bewegen, wobei jeweils die Scheiben nur einzeln bewegt werden dürfen. Der dritte Stapel C kann als Zwischenablage verwendet werden. Dabei ist die Randbedingung, niemals eine größere auf eine kleinere Scheibe zu legen, einzuhalten. Wieviele Scheiben müssen genau bewegt werden?



Es bezeichne $T(n)$ die minimale Anzahl an Bewegungen, um n Scheiben von einem Stapel auf einen anderen Stapel unter Zuhilfenahme des dritten Stapels zu bewegen.

Offensichtlich ist $T(0) = 0$, $T(1) = 1$, $T(2) = 3$.

Folgende Lösungsstrategie führt zum Ziel: Man ignoriere zunächst die größte (unterste) Scheibe auf Stapel A und bringe die oberen $n-1$ Scheiben vom Stapel A zum Stapel C unter Zuhilfenahme des Stapels B als Zwischenspeicher unter Beachtung obiger Randbedingung.

Anschließend bewege man die auf Stapel *A* verbliebene größte Scheibe zum Stapel *B*, er ja jetzt leer ist. Dann bringe man die $n-1$ Scheiben vom Stapel *C* zum Stapel *B* unter Zuhilfenahme des Stapels *A* als Zwischenspeicher unter Beachtung obiger Randbedingung.

Der folgende (Pascal-) Programmausschnitt realisiert diese Startegie:

```
CONST max = ...;

TYPE disk_typ = 0 .. max;
     stab_typ = 1 .. 3;

PROCEDURE move (anz: disk_typ;
                a  : stab_typ;
                b  : stab_typ;
                c  : stab_typ);
  { move bewegt Scheiben, deren Anzahl in anz angegeben wird,
    vom Stab a zum Stab b, wobei der Stab c als Zwischenspeicher
    verwendet wird                                              }

    BEGIN { move }
      IF anz > 0 THEN BEGIN
                        move (anz - 1, a, c, b);
                        Writeln ('Lege eine Scheibe von ', a,
                                                  ' nach ', b, '.');
                        move (anz - 1, c, b, a)
                      END
    END   { move };
```

Es gilt $T(0)=0$ und $T(n)\le 2\cdot T(n-1)+1$ für $n>0$.

Man kommt nicht mit weniger Scheibenbewegungen aus; denn um an die größte Scheibe auf dem Ausgangsstapel heranzukommen und diese vom Ausgangsstapel auf einen anderen Stapel zu bewegen, müssen zuvor $n-1$ kleinere Scheiben auf einem einzigen Stapel liegen. Dann kann die größte Scheibe bewegt werden (mindestens einmal), und dann müssen noch einmal $n-1$ kleinere Scheiben bewegt werden. Das bedeutet $T(n)\ge 2\cdot T(n-1)+1$.

Insgesamt gilt $T(0)=0$ und $T(n)=2\cdot T(n-1)+1$ für $n>0$.

Es bleibt die Bestimmung von $T(n)$ in alleiniger Abhängigkeit von *n* (und nicht von $T(n-1)$). Dazu werde einige kleinere Werte für *n* ausprobiert:

$T(0)=0$,

$T(1)=2\cdot T(0)+1=2\cdot 0+1=1$,

$T(2)=2\cdot T(1)+1=2\cdot 1+1=3$,

$T(3) = 2 \cdot T(2) + 1 = 2 \cdot 3 + 1 = 7$,
$T(4) = 2 \cdot T(3) + 1 = 2 \cdot 7 + 1 = 15$,
$T(5) = 2 \cdot T(4) + 1 = 2 \cdot 15 + 1 = 31$,
$T(6) = 2 \cdot T(5) + 1 = 2 \cdot 31 + 1 = 63$.

Die Vermutung liegt nahe, dass $T(n) = 2^n - 1$ für alle $n \in \mathbf{N}$ gilt (für $n = 0, ..., 6$ wurde es explizit ausgerechnet. Für die übrigen $n \in \mathbf{N}$ wird die Vermutung durch vollständige Induktion nachgewiesen:

Induktionsanfang: Die Aussage gilt für $n = 0, ..., 6$ (siehe oben).

Induktionsschluss: Die Aussage gelte bis $n \in \mathbf{N}$. Dann ist

$$\begin{aligned} T(n+1) &= 2 \cdot T(n) + 1 && \text{(nach Definition von } T(n)\text{)} \\ &= 2 \cdot \left(2^n - 1\right) + 1 && \text{(Voraussetzung im Induktionsschluss)} \\ &= 2^{n+1} - 2 + 1 \\ &= 2^{n+1} - 1. \end{aligned}$$

Beispiel:

Ein wichtiges Suchverfahren in der Informatik ist die **Binärsuche**:

Gegeben sei ein Feld `t[1]`, ..., `t[`n`]` mit ganzzahligen Einträgen (allgemeiner: mit bezüglich einer Ordnungsrelation vergleichbaren Einträgen), die nach aufsteigender Größe sortiert sind, d.h. es gilt `t[1]` ≤ `t[2]` ≤ ... ≤ `t[`$n-1$`]` ≤ `t[`n`]`. Die Aufgabe besteht darin festzustellen, ob ein vorgegebener Wert a unter `t[1]`, ..., `t[`n`]` vorkommt und in diesem Fall den Index i zu ermitteln, für den a = t[i] gilt. Anstelle das Feld linear von Anfang bis eventuell zum Ende zu durchsuchen, kann man folgendermaßen vorgehen:

Zunächst wird das mittlere Element `t[mitte]` geprüft (bei einer geraden Anzahl von Elementen ist das mittlere Element das erste Element der zweiten Feldhälfte). Ist es gleich a, so ist der gesuchte Feldindex gefunden, und die Suche ist beendet. Andernfalls liegt a, wenn es überhaupt im Feld vorkommt, im vorderen Feldabschnitt, falls a < `t[mitte]` ist, oder im hinteren Feldabschnitt, falls a > `t[mitte]` ist. Die Entscheidung, in welchem Feldabschnitt weiterzusuchen ist, kann jetzt getroffen werden. Gleichzeitig wird durch diese Entscheidung die andere Hälfte aller potentiell auf Übereinstimmung mit a zu überprüfenden Feldelement ausgeschlossen. Im Feldabschnitt, der weiter zu überprüfen ist, wird nach dem gleichen Prinzip (also rekursiv) verfahren. Unter Umständen muss die Suche fortgesetzt werden, bis ein noch zu überprüfender Feldabschnitt nur noch ein Feldelement enthält.

Eine (Pascal-) Implementierung der Binärsuche lautet wie folgt:

```
CONST n = ...;

TYPE Tarray = ARRAY [1..n] OF INTEGER;

FUNCTION Binaersuche (t   : Tarray;
                      a   : INTEGER;
                      von : INTEGER;
                      bis : INTEGER) : INTEGER;

VAR mitte : INTEGER;

BEGIN { Binaersuche }
  Binaersuche := -1;
  IF von < bis
  THEN BEGIN { der Feldausschnitt
               t[von] , ... t[bis]
               enthält mindestens 2 Elemente                  }
         mitte := von + ((bis - von + 1) DIV 2);
         IF a = t[mitte]          { ← }
         THEN Binaersuche := mitte
         ELSE BEGIN
                IF a < t[mitte]  { ← }
                THEN Binaersuche := Binaersuche (t, a, von, mitte-1)
                ELSE Binaersuche
                              := Binaersuche (t, a, mitte + 1, bis);
              END
       END
  ELSE BEGIN
         IF a = t[von]
         THEN Binaersuche := von;
       END;
END   { Binaersuche };
```

Der Aufruf zum Durchsuchen des Feldes `t[1]`, ..., `t[`n`]` nach dem Element a lautet
`Binaersuche (t,` a`, 1,` n`);`

Der Rechenaufwand der Binärsuche ist proportional zur Anzahl der Vergleiche in der mit ← gekennzeichneten Zeilen. Zur Vereinfachung der Analyse des Rechenaufwands wird $n = 2^m - 1$ angenommen (hat n nicht diese Form, dann ergibt sich eine ähnliche Abschätzung). Der Wert n beschreibt die Anzahl der Elemente des zu durchsuchenden Felds. In diesem Fall ist

`mitte =` $1 + \left((n - 1 + 1)$ `DIV 2`$\right) = 1 + \left(\left(2^m - 1\right)$ `DIV 2`$\right) = 2^{m-1}$,

d.h. wenn *a* nicht in der Mitte des Felds vorkommt, enthält das Anfangsstück des Felds `t[1]`, ..., `t[mitte - 1]` $2^{m-1}-1$ viele Elemente bzw. das Endstück t[mitte + 1], ..., t[*n*] ebenfalls $2^{m-1}-1$ viele Elemente; die Suche wird in einem dieser Abschnitte fortgeführt.

Mit $B(n)$ wird die Anzahl der Vergleiche des Elements *a* mit einem Feldelement bezeichnet, wenn das Feld *n* Elemente enthält. Dann gilt

$$B(n) = B\left(2^m - 1\right) \le B\left(2^{m-1} - 1\right) + 2,$$
$$B(1) = B\left(2^1 - 1\right) = 1.$$

Um diese Ungleichungen nur in Abhängigkeit von $n = 2^m - 1$ bzw. von *m* auszudrücken, werden einige Werte für *m* ausprobiert:

$$B(1) = B\left(2^1 - 1\right) = 1,$$
$$B(3) = B\left(2^2 - 1\right) \le B\left(2^1 - 1\right) + 2 = 3,$$
$$B(7) = B\left(2^3 - 1\right) \le B\left(2^2 - 1\right) + 2 \le 3 + 2 = 5,$$
$$B(15) = B\left(2^4 - 1\right) \le B\left(2^3 - 1\right) + 2 \le 5 + 2 = 7,$$
$$B(31) = B\left(2^5 - 1\right) \le B\left(2^4 - 1\right) + 2 \le 7 + 2 = 9,$$
$$B(63) = B\left(2^6 - 1\right) \le B\left(2^5 - 1\right) + 2 \le 9 + 2 = 11.$$

Die Vermutung liegt nahe, dass folgende Gesetzmäßigkeit gilt:

$B\left(2^m - 1\right) \le 2 \cdot m - 1$ für jedes $m \in \mathbf{N}$ mit $m \ge 1$.

Diese Gesetzmäßigkeit wird durch vollständige Induktion bezogen auf *m* („über *m*“) bewiesen:

Induktionsanfang: Die Aussage gilt für $m = 1, ..., 6$ (siehe oben).

Induktionsschluss: Die Aussage gelte bis zur natürlichen Zahl $m \in \mathbf{N}$. Dann ist

$$\begin{aligned} B(2^{m+1} - 1) &\le B(2^m - 1) + 2 && \text{(aus dem Algorithmus)} \\ &\le \left(2 \cdot m - 1\right) + 2 && \text{(Voraussetzung im Induktionsschluss)} \\ &= 2 \cdot \left(m + 1\right) - 1. \end{aligned}$$

Dieses Ergebnis zeigt beispielsweise, dass die Binärsuche in einem Feld mit $n = 2.147.483.647 = 2^{31} - 1$ Elementen maximal nur 61 Feldelementvergleiche benötigt.

Weitere Beispiele für Beweise durch vollständige Induktion finden sich in den folgenden Kapiteln.

Die Methode der vollständigen Induktion erfordert die sorgfältige Formulierung der zu beweisenden Aussage $A(n)$. Dabei muss man die zu beweisende Aussage bereits kennen (raten) und sie dann mit Hilfe des Induktionsanfangs und des Induktionsschlusses beweisen (verifizieren).

Folgende Hinweise sollten beachtet werden:

1. Der Induktionsanfang ist wichtig. Fehlt er, kann der Beweis fehlschlagen.

2. Im Induktionschluss lautet die Annahme „Es gelte $A(n)$“, d.h. die Gültigkeit von $A(n)$ wird nur für ein $n \in \mathbf{N}$ angenommen und nicht für *alle* $n \in \mathbf{N}$.

3. Bei der Durchführung des Induktionsschlusses muss die Annahme der Gültigkeit von $A(n)$ auch verwendet, d.h. in die Argumentation eingebaut werden.

1.6 Endliche Summen

Häufig hat man es mit **Summen mit einer endlichen Anzahl von Summanden** zu tun, die alle jeweils nach einem ähnlichen Schema aufgebaut sind, etwa

$$S = a_1 + a_2 + a_3 + \ \ldots\ a_{n-1} + a_n .$$

Für diese Summe schreibt man abkürzend $S = \sum_{i=1}^{n} a_i$.

In die „Formel“ a_i wird nacheinander $i = 1$, $i = 2$, ... , $i = n-1$ und $i = n$ eingesetzt, und die so erhaltenen Summanden werden aufsummiert. Die Berechnung von S könnte also in einer Programmiersprache wie folgt formuliert werden (Pascal-Pseudocode):

```
S := 0;
FOR i := 1 TO n DO
  S := S + ai;
```

bzw.

```
S := 0;
i := 1;
WHILE i <= n DO
  BEGIN
    S := S + ai;
    i := i + 1;
  END;
```

Beispiel:

Es sei $a_i = 3i^2 + 1$. Dann ist

$$\sum_{i=1}^{4} a_i = \sum_{i=1}^{4} (3i^2 + 1) = (3 \cdot \mathbf{1}^2 + 1) + (3 \cdot \mathbf{2}^2 + 1) + (3 \cdot \mathbf{3}^2 + 1) + (3 \cdot \mathbf{4}^2 + 1) = 4 + 13 + 28 + 49 = 94.$$

Häufig beginnt eine Summe nicht mit dem kleinsten Index $i = 1$, sondern mit einer anderen ganzen Zahl (auch negative Zahlen sind zugelassen), so dass man es allgemein mit einer endlichen Summe der Form $S = \sum_{i=k}^{n} a_i$ zu tun hat. Hierin heißt i der **Summationsindex**, die Zahl k die **Summationsuntergrenze** und die Zahl n die **Summationsobergrenze**.

Die Summe $S = \sum_{i=k}^{n} a_i$ enthält $n - k + 1$ viele Summanden.

In der Darstellung der Summe $S = \sum_{i=k}^{n} a_i$ wird deutlich, wie die einzelnen Summanden aufgebaut sind, nämlich gemäß einer Formel $a_i = a(i)$. Die Summe S ist nicht nur von den einzelnen Summanden, sondern auch von der Summationsuntergrenze und –obergrenze abhängig, d.h. $S = S(k, n)$. Die Darstellung $S(k, n) = \sum_{i=k}^{n} a_i$ zeigt nicht den Wert der Summe in Abhängigkeit von der Summationsuntergrenze k und der Summationsobergrenze n. Eine Aufgabe besteht daher in der Berechnung des Werts der Summe in Abhängigkeit von den Summationsgrenzen (**Berechnung der Summe *S*(*k*, *n*) in geschlossener Form**).

Beispiel:

Die Summe $S(1, n) = \sum_{i=1}^{n} (3i^2 + 1)$ hat den Wert $S(1, n) = \frac{n(2 + (2n+1)(n+1))}{2}$. Bei $n = 4$ ergibt sich $S(1, 4) = 94$.

Eine endliche Summe lässt sich in Teilsummen zerlegen, die ihrerseits wieder mit jeweils einem Summenzeichen zusammengefasst werden können, z.B.

$$\begin{aligned}
&a_1 + a_2 + \ \dots\ + a_{k-1} + a_k + a_{k+1} + a_{k+2} + \ \dots\ + a_{j-1} + a_j + a_{j+1} + a_{j+2} + \ \dots\ + a_{n-1} + a_n \\
&= a_1 + a_2 + \ \dots\ + a_{k-1} + \left(\sum_{i=k}^{j} a_i \right) + a_{j+1} + a_{j+2} + \ \dots\ + a_{n-1} + a_n \\
&= \left(\sum_{i=1}^{k-1} a_i \right) + \left(\sum_{i=k}^{j} a_i \right) + \left(\sum_{i=j+1}^{n} a_i \right).
\end{aligned}$$

Die *Bezeichnung i* des Summationsindex kann beliebig geändert werden: $\sum_{i=k}^{n} a_i = \sum_{\mu=k}^{n} a_\mu$.

Anstelle von $\sum_{i=k}^{n} a_i$ schreibt man auch $\sum_{k \le i \le n} a_i$.

Ist I eine beliebige Menge (**Indexmenge**), so ist $\sum_{i \in I} a_i$ die Summe, die man dadurch enthält, dass man nacheinander a_i für jedes $i \in I$ bildet und die einzelnen Summanden aufaddiert. Auf die Reihenfolge, in der man die einzelnen Indizes $i \in I$ betrachtet, kommt es nicht an.

Beispiel:

Die Summe der Quadrate aller geraden Zahlen zwischen 4 und 12 ist

$$\begin{aligned}
\sum_{i \in \{4,6,8,10,12\}} i^2 &= 4^2 + 6^2 + 8^2 + 10^2 + 12^2 \\
&= (2 \cdot 2)^2 + (2 \cdot 3)^2 + (2 \cdot 4)^2 + (2 \cdot 5)^2 + (2 \cdot 6)^2 \\
&= \sum_{i=2}^{6} (2i)^2 \\
&= \sum_{i=2}^{6} 4i^2 = 4 \cdot 2^2 + 4 \cdot 3^2 + 4 \cdot 4^2 + 4 \cdot 5^2 + 4 \cdot 6^2 \\
&= 4 \cdot \sum_{i=2}^{6} i^2 = 360.
\end{aligned}$$

Aus der Darstellung der Summenberechnung mit Hilfe des oben angegebenen Pascal-Pseudocodes sieht man, dass die Summe über eine leere Anzahl von Summanden gleich 0 ist:

$\sum_{i \in \emptyset} a_i = 0$ und $\sum_{i=k}^{n} a_i = 0$ für $k > n$.

Der folgende Satz fasst einfache Rechenregeln mit endlichen Summen zusammen.

Satz 1.6-1:

(i) Ist c eine Konstante, die vom Summationsindex nicht abhängt, so ist

$$\sum_{i\in I}(c\cdot a_i)=c\cdot\sum_{i\in I}a_i\ .$$

(ii) $$\sum_{i\in I}(a_i\pm b_i)=\left(\sum_{i\in I}a_i\right)\pm\left(\sum_{i\in I}b_i\right).$$

(iii) Ist c eine Konstante, die vom Summationsindex nicht abhängt, so ist

$$\sum_{i=1}^{n}c=n\cdot c \text{ und } \sum_{i=k}^{n}c=(n-k+1)\cdot c\ .$$

(iv) Es sei $k\in\mathbf{N}$ mit $1\le k\le n$. Dann ist

$$\sum_{i=1}^{n}a_i=\sum_{i=k}^{n+k-1}a_{i-k+1}\ (\textbf{Indexverschiebung}).$$

(v) $$\left(\sum_{i=1}^{n}a_i\right)\cdot\left(\sum_{j=1}^{m}b_j\right)=(a_1+\dots a_n)\cdot(b_1+\dots+b_m)$$

$$=a_1\cdot(b_1+\dots b_m)+\dots a_n\cdot(b_1+\dots+b_m)$$

$$=a_1\cdot\left(\sum_{j=1}^{m}b_j\right)+\dots+a_n\cdot\left(\sum_{j=1}^{m}b_j\right)$$

$$=\sum_{i=1}^{n}\left(a_i\cdot\left(\sum_{j=1}^{m}b_j\right)\right)$$

$$=\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{m}a_i\cdot b_j\right)\right),$$

$$\left(\sum_{i\in I}a_i\right)\cdot\left(\sum_{j\in J}b_j\right)=\sum_{i\in I,j\in J}(a_i\cdot b_j).$$

Satz 1.6-2:

(i) Die Summe aller natürlichen Zahlen bis zur Zahl *n* ist gleich

$$\sum_{i=0}^{n} i = 1 + 2 + \ldots (n-1) + n = \frac{n \cdot (n+1)}{2}.$$

Die Summe aller *geraden* natürlichen Zahlen bis zur Zahl *n* ist gleich

$$\sum_{\substack{0 \le i \le n \\ \text{und} \\ i \text{ ist gerade}}} i = \lfloor n/2 \rfloor \cdot (\lfloor n/2 \rfloor + 1).$$

Die Summe aller *ungeraden* natürlichen Zahlen bis zur Zahl *n* ist gleich

$$\sum_{\substack{0 \le i \le n \\ \text{und} \\ i \text{ ist ungerade}}} i = \left\lfloor \frac{n+1}{2} \right\rfloor^2.$$

(ii) Es sei $q \in \mathbf{R}$ eine Konstante. Dann ist

$$\sum_{i=0}^{n} q^i = 1 + q + q^2 + q^3 + \ldots + q^{n-1} + q^n$$

$$= \begin{cases} n+1 & \text{für } q = 1 \\ \dfrac{1-q^{n+1}}{1-q} = \dfrac{q^{n+1}-1}{q-1} & \text{für } q \neq 1 \end{cases}$$

Spezialfall: $q = 2$: $\quad \sum_{i=0}^{n} 2^i = 1 + 2 + 4 + \ \ldots\ 2^n = 2^{n+1} - 1.$

$$\sum_{i=0}^{n} i \cdot q^i = q + 2 \cdot q^2 + 3 \cdot q^3 + \ldots + (n-1) \cdot q^{n-1} + n \cdot q^n$$

$$= \begin{cases} \dfrac{n \cdot (n+1)}{2} & \text{für } q = 1 \\ \dfrac{q - (n+1) \cdot q^{n+1} + n \cdot q^{n+2}}{(1-q)^2} & \text{für } q \neq 1 \end{cases}$$

Spezialfall: $q = 2$: $\quad \sum_{i=0}^{n} i \cdot 2^i = (n-1) \cdot 2^{n+1} + 2.$

../..

$$\sum_{i=0}^{n} i^2 \cdot q^i = q + 4 \cdot q^2 + 9 \cdot q^3 + \ldots + (n-1)^2 \cdot q^{n-1} + n^2 \cdot q^n$$

$$= \begin{cases} \dfrac{n \cdot (n+1) \cdot (2 \cdot n+1)}{6} & \text{für } q = 1 \\ \dfrac{q \cdot (1+q) - (n+1)^2 \cdot q^{n+1} + \left((n+2)^2 - 2\right) \cdot q^{n+2} + (2 \cdot n+1) \cdot q^{n+3}}{(1-q)^3} & \text{für } q \neq 1 \end{cases}$$

(iii) $$\sum_{i=2}^{n} \frac{1}{i \cdot (i-1)} = \frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \ldots + \frac{1}{n \cdot (n-1)} = 1 - \frac{1}{n} \quad .$$

$$\sum_{i=1}^{n} \frac{1}{i \cdot (i+1)} = \frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \ldots + \frac{1}{n \cdot (n+1)} = 1 - \frac{1}{n+1} \quad .$$

(iv) $$\sum_{i=0}^{n} i^2 = 1 + 4 + 9 + \ldots + (n-1)^2 + n^2 = \frac{n \cdot (n+1) \cdot (2 \cdot n+1)}{6} \quad .$$

Für die Herleitung der einzelnen Aussagen werden teilweise Beweistechniken eingesetzt, die auch auf andere Formeln übertragbar sind. Diese Beweistechniken sollen exemplarisch erläutert werden.

Für die Herleitung der Aussage (i) wird unterschieden, ob *n* gerade oder ungerade ist. Ist *n* gerade, etwa $n = 2 \cdot k$, dann ist

$$\sum_{i=0}^{n} i = 1 + 2 + \ldots (n-1) + n$$

$$= \sum_{i=0}^{k} i + \sum_{i=k+1}^{2 \cdot k} i$$

$$= \underbrace{(1 + 2 + \ldots k)}_{1.\text{Summe}, k \text{ Summanden}} + \underbrace{(k+1) + (k+2) + \ldots (k+k)}_{2.\text{Summe } k \text{ Summanden}}.$$

Die Summationsreihenfolge wird geändert: Aus der 1. Summe wird der 1. Summand zum letzten Summanden in der 2. Summe addiert. Das Ergebnis ist $2 \cdot k + 1$. Die Summe des zweiten Summanden in der 1. Summe mit dem vorletzten der 2. Summe lautet $(2 \cdot k - 1) + 2 = 2 \cdot k + 1$. Allgemein ist die Summe des *i*-ten Summanden in der 1. Summe *i*-letzten Summanden der 2. Summe gleich $i + (2 \cdot k - (i-1)) = 2 \cdot k + 1$. Diese Summenbildung kann man sooft vornehmen, wie es Summanden in der 1. Bzw. 2. Summe gibt, nämlich *k*-mal. Daher ist $\sum_{i=0}^{n} i = k \cdot (2 \cdot k + 1) = \frac{n}{2} \cdot (n+1) = \frac{n \cdot (n+1)}{2}$.

Ist *n* ungerade, etwa $n = 2 \cdot k + 1$, dann ist

$$\begin{aligned}\sum_{i=0}^{n} i &= \sum_{i=0}^{2\cdot k+1} i \\ &= \sum_{i=0}^{2\cdot k} i + \underbrace{(2\cdot k+1)}_{i=2\cdot k+1} \\ &= \frac{(2\cdot k)\cdot(2\cdot k+1)}{2} + (2\cdot k+1) \\ &= \frac{(n-1)\cdot n + 2\cdot n}{2} \\ &= \frac{n\cdot(n+1)}{2}.\end{aligned}$$

Die Summe $\sum_{\substack{0\le i\le n \\ \text{und} \\ i \text{ ist gerade}}} i$ der geraden natürlichen Zahlen bis zur Zahl *n* ergibt sich wie folgt:

Ist *n* gerade, etwa $n = 2\cdot k$, dann ist

$$\sum_{\substack{0\le i\le n \\ \text{und} \\ i \text{ ist gerade}}} i = \sum_{j=0}^{k} 2\cdot j = 2\cdot\sum_{j=0}^{k} j = k\cdot(k+1) = \frac{n}{2}\cdot\left(\frac{n}{2}+1\right) = \left\lfloor\frac{n}{2}\right\rfloor\cdot\left(\left\lfloor\frac{n}{2}\right\rfloor+1\right).$$

Ist *n* ungerade, etwa $n = 2\cdot k+1$, dann ist

$$\sum_{\substack{0\le i\le n \\ \text{und} \\ i \text{ ist gerade}}} i = \sum_{\substack{0\le i\le n-1 \\ \text{und} \\ i \text{ ist gerade}}} i = \frac{n-1}{2}\cdot\left(\frac{n-1}{2}+1\right) = \left\lfloor\frac{n}{2}\right\rfloor\cdot\left(\left\lfloor\frac{n}{2}\right\rfloor+1\right).$$

Die Summe $\sum_{\substack{0\le i\le n \\ \text{und} \\ i \text{ ist ungerade}}} i$ der ungeraden natürlichen Zahlen bis zur Zahl *n* ergibt sich wie folgt:

$$\sum_{\substack{0\le i\le n \\ \text{und} \\ i \text{ ist ungerade}}} i = \sum_{0\le i\le n} i - \sum_{\substack{0\le i\le n \\ \text{und} \\ i \text{ ist gerade}}} i = \frac{n\cdot(n+1)}{2} - \left\lfloor\frac{n}{2}\right\rfloor\cdot\left(\left\lfloor\frac{n}{2}\right\rfloor+1\right).$$

Ist *n* gerade, so ist $\frac{n\cdot(n+1)}{2} - \left\lfloor\frac{n}{2}\right\rfloor\cdot\left(\left\lfloor\frac{n}{2}\right\rfloor+1\right) = \frac{n\cdot(n+1)}{2} - \frac{n}{2}\cdot\left(\frac{n}{2}+1\right) = \frac{n}{2}\cdot\left(\frac{n}{2}\right)$. Für gerades *n* ist $\left\lfloor\frac{n+1}{2}\right\rfloor = \frac{n}{2}$; damit folgt das Ergebnis.

Ist *n* ungerade, so ist

$\frac{n\cdot(n+1)}{2} - \left\lfloor\frac{n}{2}\right\rfloor\cdot\left(\left\lfloor\frac{n}{2}\right\rfloor+1\right) = \frac{n\cdot(n+1)}{2} - \frac{n-1}{2}\cdot\left(\frac{n-1}{2}+1\right) = \frac{n^2+2\cdot n+1}{4} = \left(\frac{n+1}{2}\right)^2$. Für ungerades *n*, etwa $n = 2\cdot k+1$, ist $\left\lfloor\frac{n+1}{2}\right\rfloor = \left\lfloor\frac{2\cdot k+2}{2}\right\rfloor = k+1$ und $\frac{n+1}{2} = \frac{2\cdot k+2}{2} = k+1$. Daher gilt die Formel auch für ungerades *n*.

Bei der Herleitung von (ii) für $q \neq 1$ kann eine spezielle Methode (**Pertubationsmethode**) eingesetzt werden:

Zur Berechnung von $S(n) = \sum_{i=0}^{n} a_i$ wird die Summe um einen weiteren Summanden a_{n+1} ergänzt: Dann ist $S(n) + a_{n+1} = \sum_{i=0}^{n} a_i + a_{n+1} = a_0 + \sum_{i=1}^{n+1} a_i = a_0 + \sum_{i=0}^{n} a_{i+1}$. Dieser Ansatz wird hier gewählt:

$S(n) = \sum_{i=0}^{n} q^i$, $a_i = q^i$, $a_0 = q^0 = 1$, $a_{i+1} = q^{i+1} = q \cdot q^i$,

$S(n) + q^{n+1} = 1 + \sum_{i=0}^{n} q \cdot q^i = 1 + q \cdot S(n)$. Auflösung nach $S(n)$ liefert das Ergebnis

$$S(n) = \frac{1 - q^{n+1}}{1 - q} = \frac{q^{n+1} - 1}{q - 1} \quad .$$

Ist $S(n) = \sum_{i=0}^{n} i \cdot q^i$ mit $q \neq 1$, dann ist $a_i = i \cdot q^i$, $a_0 = 0 \cdot q^0 = 0$, $a_{i+1} = (i+1) \cdot q^{i+1}$,

$S(n) + (n+1) \cdot q^{n+1} = 0 + \sum_{i=0}^{n} (i+1) \cdot q^{i+1} = q \cdot \sum_{i=0}^{n} i \cdot q^i + q \cdot \sum_{i=0}^{n} q^i = q \cdot S(n) + q \cdot \frac{1 - q^{n+1}}{1 - q}$. Diese Formel wird nach $S(n)$ aufgelöst.

Ist $S(n) = \sum_{i=0}^{n} i^2 \cdot q^i$ mit $q \neq 1$, dann ist $a_i = i^2 \cdot q^i$, $a_0 = 0 \cdot q^0 = 0$, $a_{i+1} = (i+1)^2 \cdot q^{i+1}$,

$$S(n) + (n+1)^2 \cdot q^{n+1} = 0 + \sum_{i=0}^{n} (i+1)^2 \cdot q^{i+1} = q \cdot \sum_{i=0}^{n} i^2 \cdot q^i + 2 \cdot q \cdot \sum_{i=0}^{n} i \cdot q^i + q \cdot \sum_{i=0}^{n} q^i$$

$$= q \cdot S(n) + 2 \cdot q \cdot \frac{q - (n+1) \cdot q^{n+1} + n \cdot q^{n+2}}{(1-q)^2} + q \cdot \frac{1 - q^{n+1}}{1 - q} \quad .$$

Auflösung der Formel nach $S(n)$ ergibt das Ergebnis. Für $q = 1$ wird die Formel (iv), siehe auch unten, genommen.

Zur Herleitung der Formel in (iii) wird die Methode der **Partialbruchzerlegung** eingesetzt. Dazu wird der Summand $\frac{1}{i \cdot (i-1)}$ in eine Summe $\frac{A}{i} + \frac{B}{i-1}$ zerlegt, und es werden die Werte A und B bestimmt: $\frac{1}{i \cdot (i-1)} = \frac{A}{i} + \frac{B}{i-1} = \frac{A \cdot (i-1) + B \cdot i}{i \cdot (i-1)} = \frac{(A+B) \cdot i - A}{i \cdot (i-1)}$. Diese Gleichung gilt für alle $i \geq 2$. Daher ist $A + B = 0$ und $A = -1$, also $B = 1$. Damit ergibt sich

$$\sum_{i=2}^{n} \frac{1}{i \cdot (i-1)} = \sum_{i=2}^{n} \left(\frac{1}{i-1} - \frac{1}{i} \right) = \sum_{i=2}^{n} \frac{1}{i-1} - \sum_{i=2}^{n} \frac{1}{i} = \sum_{i=1}^{n-1} \frac{1}{i} - \sum_{i=2}^{n} \frac{1}{i} = 1 - \frac{1}{n}.$$

$$\sum_{i=1}^{n} \frac{1}{i \cdot (i+1)} = \sum_{i=2}^{n+1} \frac{1}{(i-1) \cdot i} = 1 - \frac{1}{n+1}.$$

Für (iv) könnte man wieder den Einsatz der Pertubationsmethode versuchen. Mit

$S(n)=\sum_{i=0}^{n} i^2$, $a_i = i^2$, $a_0 = 0$, $a_{i+1} = (i+1)^2$ ist

$$S(n)+(n+1)^2 = \sum_{i=0}^{n}(i+1)^2 = \sum_{i=0}^{n} i^2 + \sum_{i=0}^{n}(2\cdot i+1) = S(n) + \sum_{i=0}^{n}(2\cdot i+1).$$

Die Auflösung nach $S(n)$ gelingt nicht, da sich dieser Ausdruck auf beiden Seiten aufhebt. Es bleibt aber bei der Summe der Quadrate (die sich leider aufhebt) die Summe $\sum_{i=0}^{n} i$ übrig. Vielleicht könnte man die Summe der Kubikzahlen nach dem Ansatz der Pertubationsmethode zu berechnen versuchen und hoffen, dass sich diese Summe dann auch aufhebt und die Summe der Quadrate übrig bleibt. Man setzt also $Q(n)=\sum_{i=0}^{n} i^3$ und berechnet

$$Q(n)+(n+1)^3 = \sum_{i=0}^{n}(i+1)^3 = \sum_{i=0}^{n}(i^3+3\cdot i^2+3\cdot i+1) = \sum_{i=0}^{n} i^3 + 3\cdot\sum_{i=0}^{n} i^2 + 3\cdot\sum_{i=0}^{n} i + (n+1)$$

$$= Q(n) + 3\cdot\sum_{i=0}^{n} i^2 + 3\cdot\frac{n\cdot(n+1)}{2} + (n+1).$$

Diese Gleichung wird nach $\sum_{i=0}^{n} i^2$ aufgelöst.

1.7 Elementare Ungleichungen

Der **Betrag einer reellen Zahl** $a \in \mathbf{R}$ wird definiert durch

$$|a| = \begin{cases} a & \text{für } a \geq 0 \\ -a & \text{für } a < 0 \end{cases} .$$

Dann gelten folgende Regeln:

Satz 1.7-1:

Es seien $a \in \mathbf{R}$, $b \in \mathbf{R}$ und $c \in \mathbf{R}$. Dann gilt:

(i) $|a| \geq 0$ und $|a| = |-a|$.

(ii) $|a\cdot b| = |a|\cdot|b|$.

(iii) $|a-b| = |b-a|$.

../..

(iv) $|a-b| \le |a-c| + |c-b|$ (**Dreiecksungleichung**).

(v) $|a-b| \le |a| + |b|$, $|a+b| \le |a| + |b|$.

(vi) $\big||a| - |b|\big| \le |a-b|$.

(vii) Für $a \cdot b \ge 0$ ist $\sqrt{a \cdot b} \le \dfrac{a+b}{2}$.

Teil (i) folgt unmittelbar aus der Definition.

Teil (ii) sieht man folgendermaßen:
Gilt $a \ge 0$ und $b \ge 0$, so ist $|a \cdot b| = a \cdot b = |a| \cdot |b|$. Gilt $a \ge 0$ und $b < 0$, dann ist $|a \cdot b| = -(a \cdot b) = a \cdot (-b) = |a| \cdot |b|$; genauso argumentiert man bei $a < 0$ und $b \ge 0$. Gilt $a < 0$ und $b < 0$, dann ist $|a \cdot b| = a \cdot b = (-a) \cdot (-b) = |a| \cdot |b|$.

Teil (iii) folgt aus (ii): $|a-b| = |(-1) \cdot (b-a)| = |(-1)| \cdot |b-a| = |b-a|$.

Für die Verifikation von Teil (iv) werden mehrere Fälle unterschieden und (iii) verwendet:
Ist $a \le c \le b$, dann ist $|a-b| = |b-a| = b-a = b-c+c-a = |b-c| + |c-a| = |a-c| + |c-b|$.
Ist $a \le b \le c$, dann ist $|a-c| = c-a = b-a+c-b = |b-a| + |c-b| = |a-b| + |c-b|$, also
$|a-b| = |a-c| - |c-b| \le |a-c| \le |a-c| + |c-b|$.
Ist $c \le a \le b$, dann ist $|b-c| = b-c = b-a+a-c = |b-a| + |a-c| = |a-b| + |a-c|$, also
$|a-b| = |b-c| - |a-c| = |c-b| - |a-c| \le |c-b| \le |c-b| + |a-c|$.
Für $b < a$ wird ähnlich abhängig von der Lage von *c* relativ zu *a* und *b* argumentiert.

Setzt man in (iv) $c = 0$, so erhält man die erste Ungleichung in Teil (v). Die zweite Ungleichung erhält man aus der ersten Ungleichung: $|a+b| = |a-(-b)| \le |a| + |-b| = |a| + |b|$.

In Teil (vi) trifft man wieder Fallunterscheidungen und wendet Teil (v) an:
Ist $0 \le a \le b$ oder $0 \le b \le a$, so ist $\big||a| - |b|\big| = |a-b|$. Entsprechend ergibt sich bei $a \le b < 0$ oder $b \le a < 0$: $\big||a| - |b|\big| = |-a-(-b)| = |b-a| = |a-b|$.
Für $b \le 0 \le a$ ist $\big||a| - |b|\big| = |a-(-b)| = |a+b| \le |a| + |b| = a-b = |a-b|$, da $a-b \ge 0$ ist. Aus Symmetriegründen gilt die Behauptung auch für $a \le 0 \le b$.

Teil (vii) folgt aus der binomischen Formel: Aus $0 \le (a-b)^2 = a^2 - 2 \cdot a \cdot b + b^2$ ergibt sich nacheinander $4 \cdot a \cdot b \le a^2 + 2 \cdot a \cdot b + b^2 = (a+b)^2$, $a \cdot b \le \frac{(a+b)^2}{4}$, $\sqrt{a \cdot b} \le \frac{a+b}{2}$.

Häufig werden Betragsgleichungen in Ungleichungen umgesetzt:

Satz 1.7-2:

Es seien $x \in \mathbf{R}$, $a \in \mathbf{R}$ und $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$. Dann gilt:

Die Ungleichung $|x-a| < \varepsilon$ ist gleichbedeutend mit $a - \varepsilon < x < a + \varepsilon$;
die Ungleichung $|x-a| \le \varepsilon$ ist gleichbedeutend mit $a - \varepsilon \le x \le a + \varepsilon$.

Die Aussagen verifiziert man durch Unterscheidung aller möglichen Fälle:
Ist $x < a$, dann ist bei $|x-a| < \varepsilon$: $|x-a| = -(x-a) < \varepsilon$, also $x > a - \varepsilon$;
$x \le x + 2 \cdot |x-a| = x - 2 \cdot (x-a) = -x + a + a = a + (-(x-a)) < a + \varepsilon$. Ist umgekehrt
$a - \varepsilon < x < a + \varepsilon$, dann ist $-x + a < \varepsilon$, d.h. $|x-a| = -(x-a) < \varepsilon$.
Die Verifikation im Fall $x \ge a$ verläuft analog.

Es seien $a \in \mathbf{R}$ und $b \in \mathbf{R}$ reelle Zahlen mit $a \le b$. Die Menge

$[a,b] = \{\ x \mid x \in \mathbf{R} \text{ und } a \le x \le b\ \}$ heißt **abgeschlossenes Intervall** von a bis b,
$]a,b[= \{\ x \mid x \in \mathbf{R} \text{ und } a < x < b\ \}$ heißt **offenes Intervall** von a bis b,
$[a,b[= \{\ x \mid x \in \mathbf{R} \text{ und } a \le x < b\ \}$ heißt **halboffenes Intervall** von a bis b,
$]a,b] = \{\ x \mid x \in \mathbf{R} \text{ und } a < x \le b\ \}$ heißt **halboffenes Intervall** von a bis b.

Unter einem **Intervall** wird ein abgeschlossenes oder offenes oder halboffenes Intervall verstanden. Zusätzlich zu den Intervallen mit reellwertigen Begrenzungspunkten werden folgende Intervalle definiert:

$]-\infty, a] = \{\ x \mid x \in \mathbf{R} \text{ und } x \le a\ \}$,
$]-\infty, a[= \{\ x \mid x \in \mathbf{R} \text{ und } x < a\ \}$,
$[a, \infty[= \{\ x \mid x \in \mathbf{R} \text{ und } a \le x\ \}$,
$]a, \infty[= \{\ x \mid x \in \mathbf{R} \text{ und } a < x\ \}$ und
$]-\infty, \infty[= \mathbf{R}$.

Mit Satz 1.7-2 ist dann $]a,b[=\left\{x \;\middle|\; x\in\mathbf{R} \text{ und } \left|x-\frac{a+b}{2}\right|<\frac{|a-b|}{2}\right\}$

und $[a,b]=\left\{x \;\middle|\; x\in\mathbf{R} \text{ und } \left|x-\frac{a+b}{2}\right|\le\frac{|a-b|}{2}\right\}$.

Satz 1.7-3:

Die folgenden Aussagen (a) und (b) sind gleichbedeutend:

(a) Die Teilmenge $I\subseteq\mathbf{R}$ ist ein Intervall.
und
(b) Für jedes $x_0\in I$ und jedes $x_1\in I$ ist $[x_0,x_1]\subseteq I$.

Der Nachweis der Korrektheit dieser Aussage erfolgt in zwei Richtungen:

„$(a)\Rightarrow(b)$“:
Es werden alle Möglichkeiten für *I* überprüft. Exemplarisch wird der Fall $I=[a,b[$ gezeigt: Ist $x\in[x_0,x_1]$, dann ist $a\le x_0\le x\le x_1<b$, also $x\in I$.

„$(b)\Rightarrow(a)$“:
Auch hier werden alle Möglichkeiten für *I* überprüft.
Ist beispielsweise $I\neq\varnothing$ und nach oben und unten beschränkt, dann wird *a* als die größte untere Schranke von *I* und *b* als die kleinste obere Schranke von *I* gesetzt; beide Werte *a* und *b* existieren nach Konstruktion von **R**. Jetzt wird danach unterschieden, ob $a\in I$, $a\notin I$, $b\in I$ bzw. $b\notin I$ gilt. Es sei etwa $a\in I$ und $b\notin I$; in den anderen Fällen argumentiert man entsprechend. Für $x\in I$ gilt $a\le x\le b$, und $x=b$ ist wegen $b\notin I$ nicht möglich. Das bedeutet $I\subseteq[a,b[$. Gilt umgekehrt $x\in[a,b[$, dann ist $a\le x<b$. Wäre für jedes $y\in I$ $y\le x$, dann wäre *x* obere Schranke von *I* mit $x<b$ im Widerspruch zur Wahl von *b*. Also gibt es $y\in I$ mit $a\le x<y$. Setz man in (b) $x_0=a$ und $x_1=y$, so folgt $[a,y]\subseteq I$ und damit $x\in I$. Das bedeutet $[a,b[\subseteq I$.
Die Überprüfung der übrigen Fälle für *I* erfolgt auf ähnliche Weise.

2 Abbildungen

Abbildungen stellen Beziehungen zwischen Mengen A und B her. Sie können als Spezialisierung des Konzepts der Relationen zwischen Mengen definiert werden.

2.1 Allgemeines

In Kapitel 1.4 wurden die Begriffe Äquivalenzrelation und der Ordnungsrelation eingeführt. Diese sind Spezialisierungen des allgemeineren Begriffs der Relation:

Es seien A und B zwei Mengen. Eine Teilmenge $R \subseteq A \times B$ heißt **Relation** zwischen A und B. Eine Relation besteht also aus Paaren (a, b) mit $a \in A$ und $b \in B$.

Eine Relation $R \subseteq A \times B$ heißt **linkstotal**, wenn es zu jedem $a \in A$ ein $b \in B$ mit $(a, b) \in R$ gibt. Bei einer linkstotalen Relation R kommen alle Elemente von A als erste Komponenten in den Paaren in R vor. Zu $a \in A$ kann es auch mehrere $b_1 \in B$, ..., $b_m \in B$ geben mit $(a, b_1) \in R$, ..., $(a, b_m) \in R$.

Eine Relation $R \subseteq A \times B$ heißt **rechtstotal**, wenn es zu jedem $b \in B$ ein $a \in A$ mit $(a, b) \in R$ gibt. Bei einer rechtstotalen Relation R kommen alle Elemente von B als zweite Komponenten in den Paaren in R vor. Das Element $a \in A$, das es zu $b \in B$ mit $(a, b) \in R$ gibt, muss auch hier nicht eindeutig bestimmt sein, d.h. es kann zu $b \in B$ mehrere $a_1 \in A$, ..., $a_n \in A$ geben mit $(a_1, b) \in R$, ..., $(a_n, b) \in R$.

Eine Relation $R \subseteq A \times B$ heißt **linkseindeutig**, wenn gilt: aus $(a_1, b) \in R$ und $(a_2, b) \in R$ folgt $a_1 = a_2$. Bei einer linkseindeutigen Relation gilt dann: Sind $(a_1, b_1) \in R$ und $(a_2, b_2) \in R$ und gilt $a_1 \neq a_2$, so ist auch $b_1 \neq b_2$.

Eine Relation $R \subseteq A \times B$ heißt **rechteindeutig**, wenn gilt: aus $(a, b_1) \in R$ und $(a, b_2) \in R$ folgt $b_1 = b_2$. Bei einer rechtseindeutigen Relation gilt dann: Sind $(a_1, b_1) \in R$ und $(a_2, b_2) \in R$ und gilt $b_1 \neq b_2$, so ist auch $a_1 \neq a_2$.

Eine Relation $f \subseteq A \times B$ heißt **Abbildung** von A nach B, wenn f linkstotal und rechtseindeutig ist. Gleichbedeutend damit ist folgende Formulierung:

$f \subseteq A \times B$ ist eine Abbildung, wenn es zu jedem $a \in A$ genau ein $b \in B$ gibt mit $(a, b) \in f$.

Da dieses eindeutig bestimmt Element $b \in B$ „zu $a \in A$ gehört“, schreibt man anstelle von $(a,b) \in f$ auch $b = f(a)$ und bezeichnet es als **Bild** von *a* unter *f*. Häufig gibt es eine Rechenvorschrift, nach der zu gegebenem $a \in A$ das Bild $f(a)$ zu bestimmen ist, etwa $f(a) = a^3 - 3a^2 + 2$.

Dann wird eine Abbildung *f* von *A* nach *B* beschrieben durch

$$f: \begin{cases} A \to B \\ a \to f(a) \end{cases}$$

oder auch in der Form

$f: A \to B$, $f(a) = \ ...$

Die Menge *A* heißt **Definitionsbereich** von *f*, die Menge $W(f) = \{b \mid b \in B, \text{ und es gibt } a \in A \text{ mit } f(a) = b\}$ heißt **Wertebereich** von *f*. Es ist $W(f) \subseteq B$. Anstelle von $W(f)$ schreibt man auch $f(A)$.

Für eine Funktion $f: A \to B$ ist also der Wertebereich gleich

$f(A) = \{b \mid b \in B, \text{ und es gibt } a \in A \text{ mit } f(a) = b\}$.

Die Angabe $f: A \to B$ legt fest, dass einem Element vom (Daten-) Typ, der „charakteristisch“ für *A* ist, jeweils genau ein Element vom (Daten-) Typ, der „charakteristisch“ für *B* ist, zugeordnet wird. Beispielsweise könnte die Menge *A* aus Objekten vom Objekttyp *T* und die Menge *B* aus natürlichen Zahlen bestehen. Dann legt die Angabe $f: A \to B$ fest, dass jedem Objekt vom Objekttyp *T* in der Menge *A* durch *f* eine natürliche Zahl, die beispielsweise als Primärschlüsselwert interpretierbar ist, zugeordnet wird. Die Angabe $f(a) = \ ...$ beschreibt, wie diese Zuordnung für jedes Element $a \in A$ geschieht.

Das Bild eines Elements $a \in A$ unter *f* ist eindeutig bestimmt, und es gilt $|f(a)| = 1$ für jedes $a \in A$. Andererseits kann es durchaus Werte a_1 und a_2 mit $a_1 \neq a_2$ und $f(a_1) = f(a_2)$ geben; beispielsweise ist für die durch $f(x) = x^2$ für $x \in \mathbf{R}$ definierte Abbildung $f(-2) = f(2) = 4$.

Die Menge $f^{-1}(b) = \{a \mid a \in A \text{ und } f(a) = b\}$ wird als **Urbild** von *b* unter *f* bezeichnet.

Eine Abbildung $f: A \to B$ mit $A \subseteq \mathbf{R}$ und $W(f) \subseteq \mathbf{R}$ heißt **reelle Funktion einer Veränderlichen**.

Beispiele:

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

$$g: \begin{cases} \mathbf{R} \setminus \{0\} & \to & \mathbf{R} \\ x & \to & 3/x \end{cases}$$

$$h: \begin{cases} [0,\infty[& \to & \mathbf{R} \\ x & \to & 1-e^{-x} \end{cases}$$

$$id_A: \begin{cases} A & \to & A \\ x & \to & x \end{cases}$$ **Identität auf *A***

$$par_a: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & ax(x-1) \end{cases}$$ **Parabel**

$$F: \begin{cases}]-10,\infty[& \to & \mathbf{R} \\ x & \to & \begin{cases} x^2 & \text{für } -10 < x \le 2 \\ x^3 - x & \text{für } 2 < x \le 20 \\ 2x+8 & \text{für } x > 20 \end{cases} \end{cases}$$

$$f_! : \begin{cases} \mathbf{N} & \to & \mathbf{N} \\ n & \to & \begin{cases} 1 & \text{für } n = 0 \\ n \cdot f_!(n-1) & \text{für } n > 0 \end{cases} \end{cases}$$ **Fakultätsfunktion**

Die hier aufgeführte Definition der Fakultätsfunktion zeigt die Form einer **rekursiven Definition**. Rekursive Funktionsdefinitionen werden häufig angewandt, wenn der Definitionsbereich der Funktion die natürlichen Zahlen oder eine Teilmenge der natürlichen Zahlen ist. Für den kleinsten Wert *n* des Definitionsbereich bzw. für mehrere der kleinsten Werte wird $f(n)$ direkt angegeben. Für größere Werte *n* wird $f(n)$ als arithmetischer Ausdruck, der *n*, eventuell kleinere Werte *m* und Funktionswerte $f(m)$ mit $m < n$ enthält.

Die Fakultätsfunktion kann auch nicht-rekursiv definiert werden:

$$f_! : \begin{cases} \mathbf{N} & \to & \mathbf{N} \\ n & \to & \begin{cases} 1 & \text{für } n = 0 \\ 1 \cdot \ldots \cdot (n-1) \cdot n & \text{für } n > 0 \end{cases} \end{cases}$$

Ein weiteres Beispiel einer rekursiven Funktion mit der zugehörigen nicht-rekursiven Definition ist die Fibonnacci-Funktion, die einen nichttrivialen Zusammenhang zwischen beiden Formen der Definition zeigt (siehe Kapitel 5.10):

$$fib: \begin{cases} \mathbf{N} & \to & \mathbf{N} \\ n & \to & \begin{cases} n & \text{für } n = 0 \text{ und } n = 1 \\ fib(n-1) + fib(n-2) & \text{für } n \geq 2 \end{cases} \end{cases} \quad \text{bzw.}$$

$$fib(n) = \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right) \quad \text{für } n \geq 0 .$$

Im Folgenden werden hauptsächlich reelle Funktionen betrachtet. Gelegentlich wird auf die Angabe des Definitionsbereichs einer Abbildung verzichtet; dann wird implizit immer die größte Teilmenge von **R** genommen, für die die Abbildungsvorschrift definiert ist.

Für eine Abbildung $f: A \to B$ heißt die Menge $\{(a, f(a)) \mid a \in A\}$ **Graph** der Abbildung *f*.

Sind $f: A \to B$ und $g: B \to C$ zwei Abbildungen, dann heißt die Abbildung $h: A \to C$ mit $h(a) = g(f(a))$ die **Komposition (Zusammensetzung)** der Abbildungen *f* und *g*, geschrieben $h = g \circ f$.

Es ist $W(g \circ f) \subseteq W(g) \subseteq C$, und i.a. gilt $g \circ f \neq f \circ g$.

Beispiel:

$$f: \begin{cases} \mathbf{R} \setminus \{-1\} & \to & \mathbf{R} \\ x & \to & 1/(1+x) \end{cases} \qquad g: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

$$g \circ f: \begin{cases} \mathbf{R} \setminus \{-1\} & \to & \mathbf{R} \\ x & \to & 1/(1+x)^2 \end{cases}$$

$$f \circ g: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & 1/(1+x^2) \end{cases}$$

2.2 Grundlegende Eigenschaften von Abbildungen

Eine Abbildung $f: A \to B$ heißt **surjektive Abbildung (Surjektion)**, wenn sie rechtstotal ist.

Die Abbildung $f: A \to B$ sei surjektiv. Dann ist $f(A) = B$, d.h. der Wertebereich von f umfasst ganz B. Für jedes $b \in B$ ist also $\left|f^{-1}(b)\right| \geq 1$, d.h. es gibt mindestens ein $a \in A$ mit $f(a) = b$ (eventuell gibt es mehrere Werte $a \in A$, die auf b abgebildet werden).

Beispiel:

Die Abbildung

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

ist nicht surjektiv, da es zu keiner negativen Zahl $y \in \mathbf{R}$ einen Wert $x \in \mathbf{R}$ gibt mit $f(x) = x^2 = y < 0$. Durch Einschränkung der Zielmenge kann man jedoch die Surjektivität erzwingen. Beispielsweise ist $f_0: \begin{cases} \mathbf{R} & \to & \mathbf{R}_{\geq 0} \\ x & \to & x^2 \end{cases}$ surjektiv.

Eine Abbildung $f: A \to B$ heißt **injektive Abbildung (Injektion)**, wenn sie linkseindeutig ist.

Beispiel:

Die Abbildung

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

ist nicht injektiv, da für $x_1 = -1$ und $x_2 = 1$ offensichtlich $x_1 \neq x_2$ ist, aber $f(x_1) = f(-1) = (-1)^2 = 1 = f(1) = f(x_2)$ ist.

Die Injektivität kann man durch Einschränkung des Definitionsbereichs erzwingen. Beispielsweise ist die Funktion $f_1: \begin{cases} \mathbf{R}_{\geq 0} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$ injektiv.

Die Injektivität einer Funktion kann an ihrem Graphen abgelesen werden: Jede Parallele zur x-Achse schneidet den Graphen einer injektiven Funktion in höchstens einem Punkt.

Eine Abbildung heißt **bijektive Abbildung (Bijektion)**, wenn sie sowohl surjektiv als auch injektiv ist.

Die Begriffe bezüglich Relationen und Abbildungen fasst folgende Tabelle zusammen.

Typ der Relation				
linkstotal	x	x	x	x
rechtstotal		x		x
linkseindeutig			x	x
rechtseindeutig	x	x	x	x
	Abbildung	Surjektion	Injektion	Bijektion

Satz 2.2-1:

Es sei $f : A \to B$ eine bijektive Abbildung. Dann gilt:

(i) Für jedes $b \in B$ gibt es genau ein $a_b \in A$ mit $b = f(a_b)$.

(ii) Es gibt eine eindeutig bestimmte Abbildung $g : B \to A$ mit $g(b) = a_b$; außerdem gilt für jedes $a \in A$: $g(f(a)) = a$ und für jedes $b \in B$: $f(g(b)) = b$.

In (i) folgt die Existenz eines Elements $a_b \in A$ mit $b = f(a_b)$ aus der Surjektivität von *f*; die Eindeutigkeit folgt aus der Injektivität.

Die Aussage in Satz 2.2-1 (i) kann man so interpretieren, dass es eine Eins-zu-Eins-Beziehung zwischen den Elementen der Menge *A* und der Menge *B* gibt.

Ist $f : A \to B$ eine bijektive Abbildung, so heißt die gemäß Satz 2.2-1 (ii) existierende Funktion $g : B \to A$ die **Umkehrabbildung** von *f* und wird mit f^{-1} bezeichnet. Es gilt:
$f^{-1} \circ f = id_A$ und $f \circ f^{-1} = id_B$, d.h. $\left(f^{-1} \circ f\right)(a) = a$ und $\left(f \circ f^{-1}\right)(b) = b$.

Beim Graph einer bijektiven Abbildung $f : \mathbf{R} \to \mathbf{R}$ vollzieht sich der Übergang zur Umkehrfunktion f^{-1} durch Spiegelung an der Winkelhalbierenden (45°-Linie).

Beispiele:

Die Abbildung

$$F: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & ax+b \end{cases}$$

ist für festes $a \in \mathbf{R}$ mit $a \neq 0$ und festem $b \in \mathbf{R}$ bijektiv und hat die Umkehrfunktion

$$\overset{-1}{F}: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ y & \to & \frac{1}{a} y - \frac{b}{a} \end{cases} .$$

Im allgemeinen ist eine Abbildung der Form

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & ax^3 + bx^2 + cx + d \end{cases}$$

mit festen rellen Werten *a*, *b*, *c* und *d* nicht bijektiv.

Die Abbildung

$$f_1: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

ist weder injektiv noch surjektiv. Jedoch sind die Abbildungen

$f_2: \begin{cases} \mathbf{R}_{\geq 0} & \to & \mathbf{R}_{\geq 0} \\ x & \to & x^2 \end{cases}$ und $f_3: \begin{cases} \mathbf{R}_{\leq 0} & \to & \mathbf{R}_{\geq 0} \\ x & \to & x^2 \end{cases}$ jeweils bijektiv mit den durch $\overset{-1}{f}_2(y) = +\sqrt{y}$ bzw. $\overset{-1}{f}_3(y) = -\sqrt{y}$ definierten Umkehrabbildungen.

Häufig wird bereits für eine injektive Abbildung $f: A \to B$, die nicht notwendigerweise surjektiv ist, die Umkehrabbildung $\overset{-1}{f}$ definiert, und zwar nur für die Werte $b \in B$ aus dem Wertebereich von *f*: $\overset{-1}{f}: f(X) \to X$.

Satz 2.2-2:

Es seien $f : A \to B$ und $g : B \to C$ Abbildungen. Dann gilt:

(i) Sind f und g surjektiv, dann ist auch $g \circ f$ surjektiv.

(ii) Sind f und g injektiv, dann ist auch $g \circ f$ injektiv.

(iii) Sind f und g bijektiv, dann ist auch $g \circ f$ bijektiv. In diesem Fall gilt $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

(i) und (ii) und der erste Teil von (iii) werden direkt mit den entsprechenden Definitionen nachgeprüft. Für den zweiten Teil von (iii) zeigt man, dass $f^{-1} \circ g^{-1}$ die Umkehrabbildung zu $g \circ f$ ist: Dazu seien $a \in A$, $b \in B$ mit $b = f(a)$ und $c \in C$ mit $c = g(b) = g(f(a))$. Es gilt $f^{-1} \circ g^{-1}(g(f(a))) = f^{-1}\left(g^{-1}(g(f(a)))\right) = f^{-1}(f(a)) = a$.

Für **endliche** Mengen A und B kann man die Existenz surjektiver, injektiver und bijektiver Abbildungen zwischen A und B folgendermaßen entscheiden:

Satz 2.2-3:

Die Mengen A und B seien endliche Mengen mit $|A| = n$ und $|B| = m$. Dann gilt:

(i) Es gibt genau dann eine surjektive Abbildung $f : A \to B$, wenn $n \geq m$ ist.

(ii) Es gibt genau dann eine injektive Abbildung $f : A \to B$, wenn $n \leq m$ ist.

(iii) Es gibt genau dann eine bijektive Abbildung $f : A \to B$, wenn $n = m$ ist.

Für (i) sind zwei Beweisrichtungen zu zeigen, nämlich

(i1) Wenn es eine surjektive Abbildung $f : A \to B$ gibt, dann ist $|A| \geq |B|$.

(i2) Wenn $|A| \geq |B|$ ist, dann kann man eine surjektive Abbildung $f : A \to B$ definieren.

Zu (i1): Es sei $b \in B$. Wegen der Surjektivität von f ist $\left|f^{-1}(b)\right| \geq 1$. Da f als Abbildung rechtseindeutig ist, gilt für $b_1 \in B$ und $b_2 \in B$ mit $b_1 \neq b_2$: $\left|f^{-1}(b_1)\right| \cap \left|f^{-1}(b_2)\right| = \varnothing$. Für jedes $b \in B$ wählt man einen „Repräsentanten“ $a_b \in f^{-1}(b)$. Diese sind paarweise verschieden. Da $\{ a_b \mid b \in B \} \subseteq A$ ist, folgt $|B| = |\{ a_b \mid b \in B \}| = \sum_{a_b} 1 \leq \sum_{a \in A} 1 = |A|$.

Zu (i2): Es sei $A = \{ a_1, ..., a_n \}$ und $B = \{ b_1, ..., b_m \}$ mit $n \geq m$. Die Abbildung

$$f : \begin{cases} A & \to & B \\ a_i & \to & \begin{cases} b_i & \text{für } i = 1, ..., m \\ b_m & \text{für } i = m+1, ..., n \end{cases} \end{cases}$$ ist surjektiv.

Auch für (ii) sind zwei Beweisrichtungen zu zeigen, nämlich

(ii1) Wenn es eine injektive Abbildung $f : A \to B$ gibt, dann ist $|A| \leq |B|$.

(ii2) Wenn $|A| \leq |B|$ ist, dann kann man eine injektive Abbildung $f : A \to B$ definieren.

Zu (ii1): Injektivität bedeutet: Sind $a_1 \in A$ und $a_2 \in A$ mit $a_1 \neq a_2$, dann ist $f(a_1) \neq f(a_2)$. Ist $A = \{ a_1, ..., a_n \}$ und $B' = \{ f(a_1), ..., f(a_n) \}$, dann ist $B' \subseteq B$ und $|A| = |B'| \leq |B|$.

Zu (ii2): Es sei $A = \{ a_1, ..., a_n \}$ und $B = \{ b_1, ..., b_m \}$ mit $n \leq m$. Die Abbildung

$$f : \begin{cases} A & \to & B \\ a_i & \to & b_i \end{cases}$$ ist injektiv.

Bei (iii) folgt aus der Existenz einer bijektiven Abbildung $f : A \to B$, dass wegen der Injektivität von f $n \leq m$ und wegen der Surjektivität von f $n \geq m$ ist, also $n = m$ gilt. Ist umgekehrt $n = m$, so ist die Abbildung $f : \begin{cases} A & \to & B \\ a_i & \to & b_i \end{cases}$ bijektiv.

Satz 2.2-3 lässt den Schluss zu, dass bei Abbildungen zwischen endlichen Mengen mit derselben Elementanzahl die Begriffe Injektivität, Surjektivität und Bijektivität zusammenfallen :

Satz 2.2-4:

Die Mengen A und B seien endliche Mengen mit derselben Elementanzahl $|A| = |B|$. Es sei $f : A \to B$ eine Abbildung. Dann gilt:

Die folgenden drei Aussagen (a), (b) und (c) sind äquivalent:

(a) f ist eine injektive Abbildung.

(b) f ist eine surjektive Abbildung.

(c) f ist eine bijektive Abbildung.

Die Äquivalenz von (a) und (b) ergibt sich wie folgt (die Äquivalenz zu (c) ergibt sich dann aus der Definition der Bijektivität): Ist f eine injektive Abbildung, dann ist $|f(A)| = |A| = |B|$. Wegen $f(A) \subseteq B$ hat daher jedes $b \in B$ ein Urbild, d.h. f ist surjektiv. Ist umgekehrt f nicht injektiv, dann gibt es $a_1 \in A$ und $a_2 \in A$ mit $a_1 \neq a_2$ und $f(a_1) = f(a_2)$. Daher ist $|f(A)| < |A| = |B|$, und f ist nicht surjektiv.

Satz 2.2-3 (iii) besagt für endliche Mengen, dass sie genau dann gleichmächtig sind, wenn es zwischen ihnen eine bijektive Abbildung gibt. Dieser Ansatz lässt sich auf unendliche Mengen übertragen:

Zwei unendliche Mengen A und B heißen **gleichmächtig**, wenn es eine bijektive Abbildung $f : A \to B$ gibt. Wegen der Existenz der bijektiven Umkehrfunktion g zu f ist diese Definition gleichbedeutend mit der Existenz einer bijektiven Abbildung $g : B \to A$.

Bei unendlichen Mengen A und B tritt die folgende Situation auf, die sich am Beispiel der Vorgängerfunktion *pred* auf den natürlichen Zahlen verdeutlichen lässt:
$$pred : \begin{cases} \mathbf{N}_{>0} & \to & \mathbf{N} \\ n & \to & n-1 \end{cases}.$$

Die Funktion ist für alle natürlichen Zahlen $n \geq 1$ definiert . Aus Axiom 2 der natürlichen Zahlen (siehe Kapitel 1.4) folgt, dass sie surjektiv ist; zu beachten ist, dass 0 im Definitionsbereich nicht enthalten ist. Axiom 4 besagt, dass sie injektiv ist. Es handelt es sich hierbei also um eine bijektive Abbildung, d.h. die Menge **N** der natürlichen Zahlen ist gleichmächtig mit

der echten Teilmenge $\mathbf{N}_{>0} = \mathbf{N} \setminus \{ 0 \}$. Dieses Phänomen erlaubt es, die Endlichkeit bzw. Unendlichkeit einer Menge exakt zu definieren:

Eine Menge A ist **endlich von der Mächtigkeit** n, wenn es eine bijektive Abbildung $f : \{0, ..., n-1\} \to A$ gibt, d.h. man kann die Elemente in A mit den natürlichen Zahlen 0, ..., $n-1$ durchnumerieren: $A = \{ f(0), ..., f(n-1) \} = \{ a_0, ..., a_{n-1} \}$. Hierbei ist $f(i) \neq f(j)$ bzw. $a_i \neq a_j$ für $i \neq j$. Ist B eine echte Teilmenge von A, dann kann es nach Satz 2.2-3(iii) keine bijektive Abbildung $f : B \to A$ geben. Diese Eigenschaft unterscheidet eine endliche Menge von einer Menge mit unendlicher Mächtigkeit.

Eine Menge A ist **von der Mächtigkeit unendlich**, wenn es eine bijektive Abbildung $f : B \to A$ zwischen einer echten Teilmenge $B \subset A$ und A gibt.

Eine Menge heißt **abzählbar**, wenn sie entweder endlich oder gleichmächtig zu den natürlichen Zahlen ist. Eine unendliche Menge, die nicht abzählbar ist, heißt **überabzählbar**.

Satz 2.2-5:

(i) Es gibt eine bijektive Abbildung $h_{\mathbf{Z}} : \mathbf{N} \to \mathbf{Z}$, und es gibt eine bijektive Abbildung $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$. Die Mengen **N**, **Z** und **Q** sind daher abzählbar.

(ii) Die Menge **N** der natürlichen Zahlen lässt sich nicht auf die Menge **R** der reellen Zahlen bijektiv abbilden. Die Menge **R** ist daher überabzählbar.

Für Satz 2.2-5 (i) sind zwei bijektive Abbildungen $h_{\mathbf{Z}} : \mathbf{N} \to \mathbf{Z}$ und $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$ anzugeben.

Die Abbildung $h_{\mathbf{Z}} : \mathbf{N} \to \mathbf{Z}$ wird definiert durch

$$h_{\mathbf{Z}} : \begin{cases} \mathbf{N} & \to & \mathbf{Z} \\ n & \to & \begin{cases} -n/2 & \text{falls } n \text{ gerade ist} \\ (n+1)/2 & \text{falls } n \text{ ungerade ist} \end{cases} \end{cases} .$$

Einige Werte dieser Abbildung sind

n	0	1	2	3	4	5	6	7	8	9	10	11
$h_{\mathbf{Z}}(n)$	0	1	−1	2	−2	3	−3	4	−4	5	−5	6

Die Injektivität von $h_{\mathbf{Z}}$ sieht man folgendermaßen: Es seien $n_1 \in \mathbf{N}$ und $n_2 \in \mathbf{N}$ mit $n_1 \neq n_2$. Ist n_1 gerade und n_2 ungerade, dann ist $h_{\mathbf{Z}}(n_1) \leq 0 < h_{\mathbf{Z}}(n_2)$, insbesondere $h_{\mathbf{Z}}(n_1) \neq h_{\mathbf{Z}}(n_2)$. Ist

n_1 ungerade und n_2 gerade, dann ist $h_{\mathbf{Z}}(n_2) \le 0 < h_{\mathbf{Z}}(n_1)$. Sind n_1 und n_2 beide gerade, dann ist $h_{\mathbf{Z}}(n_1) = -n_1/2 \neq -n_2/2 = h_{\mathbf{Z}}(n_2)$. Sind n_1 und n_2 beide ungerade, dann ist $h_{\mathbf{Z}}(n_1) = (n_1+1)/2 \neq (n_2+1)/2 = h_{\mathbf{Z}}(n_2)$.

Zum Nachweis der Surjektivität sei $z \in \mathbf{Z}$. Ist $z \le 0$, dann ist $n = -2 \cdot z$ in $\mathbf{N}$ und $h_{\mathbf{Z}}(n) = h_{\mathbf{Z}}(-2 \cdot z) = \dfrac{-(-2 \cdot z)}{2} = z$. Ist $z > 0$, dann ist $n = 2 \cdot z - 1$ in $\mathbf{N}$ und

$$h_{\mathbf{Z}}(n) = h_{\mathbf{Z}}(2 \cdot z - 1) = \frac{(2 \cdot z - 1) + 1}{2} = z.$$

Die Abbildung $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$ wird in zwei Schritten konstruiert. Zunächst wird eine bijektive Abbildung $f_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}_{\ge 0}$ angegeben, die dann zu einer bijektiven Abbildung $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$ erweitert wird:

Die rationalen Zahlen $\mathbf{Q}_{>0} = \left\{ \dfrac{r}{t} \mid r \in \mathbf{N}_{>0} \text{ und } t \in \mathbf{N}_{>0} \right\}$ kann man sich in ein unendliches Zahlenschema eingetragen denken, das aus Zeilen und Spalten besteht. In der ersten Zeile stehen alle Zahlen $1/1$, $1/2$, $1/3$, $1/4$, ... ; die zweite Zeile enthält $2/1$, $2/2$, $2/3$, $2/4$, ...; die *i*-te Zeile enthält $i/1$, $i/2$, $i/3$, $i/4$, ... Dann steht die Zahl $\dfrac{r}{t}$ in der *r*-ten Zeile und *t*-ten Spalte. Dieses Zahlenschema wird durchnumeriert: $1/1$ erhält die Nummer 1 ($1/1$ ist die einzige rationale Zahl $\dfrac{r}{t} \in \mathbf{Q}_{>0}$ mit $r + t = 2$). Dann kommen alle Zahlen $\dfrac{r}{t} \in \mathbf{Q}_{>0}$ mit $r + t = 3$, nach aufsteigenden Zählern geordnet (das sind $1/2$ und $2/1$). Anschließend kommen alle $\dfrac{r}{t} \in \mathbf{Q}_{>0}$ mit $r + t = 4$, nach aufsteigenden Zählern geordnet (das sind $1/3$, $2/2$ und $3/1$) usw. Die folgende Tabelle zeigt einige kleine rationale Zahlen mit ihren Nummern.

m	r/t mit $r \ge 1$, $t \ge 1$ und $r + t = m$	Anzahl	Nummern
2	1/1	1	1
3	1/2 2/1	2	2 3
4	1/3 2/2 3/1	3	4 5 6
5	1/4 2/3 3/2 4/1	4	7 8 9 10
6	1/5 2/4 3/3 4/2 5/1	5	11 12 13 14 15

Es gibt genau $m-1$ rationale Zahlen $\dfrac{r}{t} \in \mathbf{Q}_{>0}$ mit $r + t = m$, nämlich $1/(m-1)$, $2/(m-2)$, $3/(m-3)$, ..., $(m-1)/1$. Vor diesem „Block“ von Zahlen liegen alle Zahlen $\dfrac{u}{v} \in \mathbf{Q}_{>0}$ mit $u \ge 1$, $v \ge 1$ und $u + v = i$ mit $2 \le i \le m-1$. Daher bekommt $1/(m-1)$ die Nummer

$\sum_{i=2}^{m-1}(i-1)+1=\sum_{i=1}^{m-2} i+1=\frac{(m-1)\cdot(m-2)}{2}+1$.

Die Zahl $k/(m-k)$ für $k=1,...,m-1$ bekommt die Nummer $\frac{(m-1)\cdot(m-2)}{2}+k$, d.h. die natürlichen Zahlen

$$\frac{(m-1)\cdot(m-2)}{2}+1,\ \frac{(m-1)\cdot(m-2)}{2}+2,\ ...,\ \frac{(m-1)\cdot(m-2)}{2}+m-1=\frac{(m-1)\cdot m}{2}$$

numerieren die Zahlen $\frac{r}{t}\in\mathbf{Q}_{>0}$ mit $r+t=m$.

Die Abbildung $f_{\mathbf{Q}}:\mathbf{N}\to\mathbf{Q}_{\geq 0}$ wird nun wie folgt definiert:

$f_{\mathbf{Q}}(0)=0$.

Für $n>0$ gibt es eine eindeutig bestimmte Zahl $m\in\mathbf{N}_{>0}$ mit $\frac{(m-1)\cdot(m-2)}{2}<n\leq\frac{(m-1)\cdot m}{2}$.

Es gilt nämlich: Für $m\in\mathbf{N}_{>0}$ mit $m\geq 2$ ist die nächste auf $\frac{(m-1)\cdot(m-2)}{2}$ folgende Zahl der Form $\frac{(k-1)\cdot(k-2)}{2}$ mit $k\in\mathbf{N}_{>0}$ die Zahl $\frac{((m+1)-1)\cdot((m+1)-2)}{2}=\frac{m\cdot(m-1)}{2}$. Daher liegt *n* eindeutig zwischen $\frac{(m-1)\cdot(m-2)}{2}$ und $\frac{(m-1)\cdot m}{2}$.

m	1	2	3	4	5	6	7	8	9	10	11	12
$\frac{(m-1)\cdot(m-2)}{2}$	0	0	1	3	6	10	15	21	28	36	45	55
$\frac{(m-1)\cdot m}{2}$	0	1	3	6	10	15	21	28	36	45	55	66

Für $n=17$ ist $m=7$, für $n=21$ ist $m=7$, für $n=22$ ist $m=8$.

Die Anzahl der Werte *n* mit $\frac{(m-1)\cdot(m-2)}{2}<n\leq\frac{(m-1)\cdot m}{2}$ beträgt $m-1$.

Es wird $r=n-\frac{(m-1)\cdot(m-2)}{2}$ und $t=m-r$ gesetzt.

Da $(m-1)\cdot(m-2)$ gerade ist und $\frac{(m-1)\cdot(m-2)}{2}<n$ gilt, ist $r\in\mathbf{N}_{>0}$. Außerdem ist

$r=n-\frac{(m-1)\cdot(m-2)}{2}\leq\frac{(m-1)\cdot m}{2}-\frac{(m-1)\cdot(m-2)}{2}=m-1$ und damit auch $t\in\mathbf{N}_{>0}$.

Es wird $f_{\mathbf{Q}}(n)$ durch

$f_{\mathbf{Q}}(n)=\frac{r}{t}$ (in dieser ungekürzten Darstellung)

definiert. Die folgende Tabelle zeigt die Ergebnisse $f_{\mathbf{Q}}(n)$ für die $m-1$ Werte *n* mit

$\frac{(m-1)\cdot(m-2)}{2} < n \le \frac{(m-1)\cdot m}{2}$:

n	$\frac{(m-1)\cdot(m-2)}{2}+1$	$\frac{(m-1)\cdot(m-2)}{2}+2$	...	$\frac{(m-1)\cdot m}{2}-1$	$\frac{(m-1)\cdot m}{2}$
r	1	2	...	$m-2$	$m-1$
t	$m-1$	$m-2$	...	2	1
$f_{\mathbf{Q}}(n)$	$1/(m-1)$	$2/(m-2)$		$(m-2)/2$	$(m-1)/1$

Offensichtlich gilt für diese n jeweils $f_{\mathbf{Q}}(n)=\frac{r}{t}$ mit $1 \le r \le m-1$, $1 \le t \le m-1$ und $r+t=m$.

Die so definierte Abbildung $f_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}_{\ge 0}$ ist bijektiv:

Zum Nachweis der Surjektivität sei $q \in \mathbf{Q}_{\ge 0}$. Für $q=0$ wird $n=0$ gesetzt; nach Definition ist $f_{\mathbf{Q}}(n)=f_{\mathbf{Q}}(0)=0=q$. Für $q>0$, etwa $q=r/t$ mit $r \in \mathbf{N}$, $t \in \mathbf{N}$, $r>0$ und $t>0$, sei $m=r+t \ge 2$. Es wird $n=r+\frac{(r+t-1)\cdot(r+t-2)}{2}=r+\frac{(m-1)\cdot(m-2)}{2}$ gesetzt. Dann ist $n \in \mathbf{N}$ und $\frac{(m-1)\cdot(m-2)}{2} < n = m-t+\frac{(m-1)\cdot(m-2)}{2} \le m-1+\frac{(m-1)\cdot(m-2)}{2}=\frac{m\cdot(m-1)}{2}$.

Nach Definition von $f_{\mathbf{Q}}$ ist $f_{\mathbf{Q}}(n)=\frac{r}{t}$.

Zum Nachweis der Injektivität seien $n_1 \in \mathbf{N}$ und $n_2 \in \mathbf{N}$ mit $n_1 \ne n_2$. Gilt $\frac{(m-1)\cdot(m-2)}{2} < n_i \le \frac{m\cdot(m-1)}{2}$ für $i=1$, $i=2$ und $m \in \mathbf{N}_{>0}$, dann ist mit $r_i = n_i - \frac{(m-1)\cdot(m-2)}{2}$ und $t_i = m - r_i$ für $i=1$, $i=2$: $r_1 \ne r_2$ und $f_{\mathbf{Q}}(n_1)=\frac{r_1}{t_1} \ne \frac{r_2}{t_2}=f_{\mathbf{Q}}(n_2)$.

Gilt $\frac{(m_1-1)\cdot(m_1-2)}{2} < n_1 \le \frac{m_1\cdot(m_1-1)}{2}$ und $\frac{(m_2-1)\cdot(m_2-2)}{2} < n_2 \le \frac{m_2\cdot(m_2-1)}{2}$ mit $m_1 \ne m_2$ und setzt man $r_i = n_i - \frac{(m_i-1)\cdot(m_i-2)}{2}$ und $t_i = m_i - r_i$ für $i=1$, $i=2$, dann folgt im Fall $r_1 \ne r_2$ wegen der ungekürzten Darstellung von $f_{\mathbf{Q}}$: $f_{\mathbf{Q}}(n_1)=\frac{r_1}{t_1} \ne \frac{r_2}{t_2}=f_{\mathbf{Q}}(n_2)$; ist $r_1=r_2$, dann ist $t_1 \ne t_2$ und $f_{\mathbf{Q}}(n_1) \ne f_{\mathbf{Q}}(n_2)$.

Die gesuchte Bijektion $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$ ergibt sich zu

$$h_{\mathbf{Q}} : \begin{cases} \mathbf{N} & \to & \mathbf{Q} \\ n & \to & \begin{cases} -f_{\mathbf{Q}}(n/2) & \text{falls } n \text{ gerade ist} \\ f_{\mathbf{Q}}((n+1)/2) & \text{falls } n \text{ ungerade ist} \end{cases} \end{cases}.$$

Mit den geraden natürlichen Zahlen werden also die nichtpositiven rationalen Zahlen, mit den ungeraden natürlichen Zahlen die positiven rationalen Zahlen numeriert:

n	0	1	2	3	4	5	6	7	8	9	10	11	…
i mit $n = 2 \cdot i$	0		1		2		3		4		5		…
	Numerierung der nichtpositiven rationalen Zahlen mittels $-f_{\mathbf{Q}}(n/2) = -f_{\mathbf{Q}}(i)$												
i mit $n = 2 \cdot i - 1$		1		2		3		4		5		6	…
	Numerierung der positiven rationalen Zahlen mittels $f_{\mathbf{Q}}((n+1)/2) = f_{\mathbf{Q}}(i)$												

Für Satz 2.2-5 (ii) wird angenommen, dass es eine bijektive Abbildungen $h_{\mathbf{R}} : \mathbf{N} \to \mathbf{R}$ gibt. Diese Annahme muss auf einen Widerspruch führen. Da dieses Vorgehen eine „klassische" Beweismethode auch insbesondere der Theoretischen Informatik ist und auch in die populärwissenschaftliche mathematische Literatur Eingang gefunden hat, soll die Beweisidee hier skizziert werden:

Es seien $n_1 < n_2 < n_3 < \ldots$ diejenigen $n_i \in \mathbf{N}$, für die $h_{\mathbf{R}}(n_i) \in \mathbf{R}$ mit $0 \le h_{\mathbf{R}}(n_i) \le 1$ ist. Jedes $r \in \mathbf{R}$ mit $0 \le r \le 1$ hat eine eindeutige Nummer n_i und lautet in Dezimalschreibweise
$r = 0, d_{n_i,-1} d_{n_i,-2} d_{n_i,-3} \ldots$
mit den Dezimalziffern $d_{n_i,-j} \in \{0,1,2,3,4,5,6,7,8,9\}$. Dabei ist $1 = 0,\overline{9}\ldots$ (siehe Kapitel 5.1).

Es wird eine reelle Zahl $\bar{r}$ mit $0 \le \bar{r} \le 1$ durch folgende Vorschrift konstruiert:
Die erste Dezimalziffer von $\bar{r}$ nach dem Komma lautet $9 - d_{n_1,-1}$ (es wird aus der reellen Zahl mit der Nummer n_1 die erste Dezimalziffer nach dem Komma genommen und von 9 abgezogen; dadurch entsteht wieder eine Ziffer zwischen 0 und 9); es ist $9 - d_{n_1,-1} \ne d_{n_1,-1}$. Die zweite Dezimalziffer von $\bar{r}$ nach dem Komma lautet $9 - d_{n_2,-2}$ (es wird aus der reellen Zahl mit der Nummer n_2 die zweite Dezimalziffer nach dem Komma genommen und von 9 abgezogen); es ist $9 - d_{n_2,-2} \ne d_{n_2,-2}$. Allgemein: die *j*-te Dezimalziffer von $\bar{r}$ nach dem Komma lautet $9 - d_{n_j,-j}$; es ist $9 - d_{n_j,-j} \ne d_{n_j,-j}$.

Da $\bar{r}$ eine reelle Zahl mit $0 \le \bar{r} \le 1$ ist und $h_{\mathbf{R}}$ als bijektiv angenommen wurde, gibt es einen Wert $n_k \in \mathbf{N}$ mit $h_{\mathbf{R}}(n_k) = \bar{r}$ ($\bar{r}$ ist die reelle Zahl mit der Nummer n_k):
$\bar{r} = 0, d_{n_k,-1} d_{n_k,-2} d_{n_k,-3} \ldots d_{n_k,-k} \ldots$
Die *k*-te Dezimalziffer von $\bar{r}$ nach dem Komma ist $d_{n_k,-k}$. Nach Konstruktion von $\bar{r}$ lautet die *k*-te Dezimalziffer von $\bar{r}$ nach dem Komma jedoch $9 - d_{n_k,-k}$, und es ist $9 - d_{n_k,-k} \ne d_{n_k,-k}$. Daher kann es keinen Wert $n_k \in \mathbf{N}$ mit $h_{\mathbf{R}}(n_k) = \bar{r}$ geben, und die Annahme der Existenz einer bijektiven Abbildung $h_{\mathbf{R}} : \mathbf{N} \to \mathbf{R}$ ist falsch.

3 Ausgewählte Themen der elementaren Zahlentheorie

In diesem Kapitel werden einige für die Informatik grundlegende und wichtige Themen der elementaren Zahlentheorie behandelt. Neben der Tatsache, dass sie zum mathematischen Basiswissen in jeder Disziplin gehören, haben diese Themen in den letzten Jahren insbesondere in der Kryptologie zunehmende Bedeutung erlangt.

3.1 Primzahlen

Es seien $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ ganze Zahlen mit $b \neq 0$. Die Zahl a heißt durch b **teilbar** (b **teilt** a), geschrieben $b \mid a$, wenn es ein $b' \in \mathbf{Z}$ gibt mit $a = b \cdot b'$.

Der folgende Satz führt einige wichtige Teilbarkeitsregeln ganzer Zahlen auf:

Satz 3.1-1:

(i) Gilt $c \mid b$ und $b \mid a$, so gilt auch $c \mid a$.

(ii) Gilt $b_1 \mid a_1$ und $b_2 \mid a_2$, so gilt auch $b_1 b_2 \mid a_1 a_2$.

(iii) Gilt $b \mid a_1$ und $b \mid a_2$, so gilt für jedes $x \in \mathbf{Z}$ und für jedes $y \in \mathbf{Z}$: $b \mid (x a_1 + y a_2)$.

(iv) Gilt $b \mid a$ und $a \mid b$, so ist $a = b$ oder $a = -b$.

Die Teile (i) bis (iii) lassen sich durch Zurückführen auf obige Definition direkt verifizieren. Für (iv) setzt man $b \mid a$ und $a \mid b$ voraus, d.h. $a = b \cdot b'$ mit $b' \in \mathbf{Z}$ und $b = a \cdot a'$ mit $a' \in \mathbf{Z}$. Dann ist $a = a \cdot a' \cdot b'$, also $a' \cdot b' = 1$ und folglich $a' = \pm 1$ und $b' = \pm 1$.

Bemerkung: Da trivialerweise immer $a \mid a$ gilt, definiert wegen Satz 3.1-1 (i) und (iv) die durch
„$(n, m) \in R$ genau dann, wenn $n \mid m$ gilt“
definierte Relation eine partielle Ordnungsrelation (siehe Kapitel 1.4) auf $\mathbf{N} \times \mathbf{N}$.

Eine wichtige Teilmenge der natürlichen Zahlen ist die Menge **P** der **Primzahlen**:

$\mathbf{P} = \{\ p \mid p \in \mathbf{N},\ p \geq 2,$ und die einzigen Teiler von p sind 1 und $p\ \}$
$= \{\ 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, \ldots\ \}.$

Der folgende Satz zeigt, dass die Primzahlen als Grundbausteine der natürlichen Zahlen und damit des gesamten Zahlensystems angesehen werden können.

Satz 3.1-2:

Jedes $n \in \mathbf{N}$ mit $n \geq 2$ lässt sich in ein **Produkt aus Primzahlpotenzen** zerlegen, d.h.

$$n = p_1^{e_1} \cdot p_2^{e_2} \cdot \ldots \cdot p_r^{e_r}$$

mit Primzahlen p_1, p_2, ..., p_r und natürlichen Zahlen $e_1 \geq 1$, $e_2 \geq 1$, ..., $e_r \geq 1$. Diese Zerlegung ist (bis auf die Reihenfolge der Primzahlpotenzen) eindeutig.

Die Zerlegung in Primfaktoren lässt sich leicht durch vollständige Induktion zeigen:
Für $n = 2$ ist die Behauptung klar. Sie gelte für $n \geq 2$.
Ist $n+1$ Primzahl, dann gilt die Behauptung auch für $n+1$.
Ist $n+1$ keine Primzahl, dann besitzt sie einen Teiler a mit $1 < a < n+1$ und einen Teiler b mit $1 < b < n+1$ und $n+1 = a \cdot b$. Nach Induktionsvoraussetzung sind a und b Produkte von Primzahlpotenzen und damit auch $n+1$.

Auch die Eindeutigkeit der Zerlegung (bis auf die Reihenfolge) kann durch vollständige Induktion nachgewiesen werden:
Für $n = 2$ ist die Behauptung klar. Sie gelte für alle natürlichen Zahlen m mit $2 \leq m \leq n$.
Besitzt $n+1$ zwei unterschiedliche Darstellungen von Primzahlpotenzen, die sich nicht nur in der Reihenfolge der Faktoren unterscheiden, etwa $n+1 = p_1^{e_1} \cdot p_2^{e_2} \cdot \ldots \cdot p_r^{e_r} = q_1^{e_1'} \cdot q_2^{e_2'} \cdot \ldots \cdot q_s^{e_s'}$, dann ist jedes p_i von jedem q_j verschieden (denn andernfalls kann man dieselbe Primzahl in beiden Darstellungen weglassen und erhielte eine Zahl m mit $2 \leq m \leq n$, die der Induktionsvoraussetzung widerspricht). Insbesondere ist $p_1 \neq q_1$, und man kann $p_1 < q_1$ annehmen. Es seien $a = p_1^{e_1 - 1} \cdot p_2^{e_2} \cdot \ldots \cdot p_r^{e_r}$ und $b = q_1^{e_1' - 1} \cdot q_2^{e_2'} \cdot \ldots \cdot q_s^{e_s'}$, d.h. a ergibt sich durch Fortlassen der Primzahl p_1 aus $n+1$ und b durch Fortlassen der Primzahl q_1 aus $n+1$. Für die Zahl $m = n+1-p_1 \cdot b$ gilt $m = p_1 \cdot (a-b) = (q_1 - p_1) \cdot b$ und $2 \leq m \leq n$.
Auf die Zahlen m, $q_1 - p_1$ und b ist die Induktionsvoraussetzung anwendbar, und aus den eindeutigen Zerlegungen in Primzahlpotenzen von $q_1 - p_1$ und b erhält man durch Multiplikation die eindeutige Zerlegung in Primzahlpotenzen von m. Wegen $m = p_1 \cdot (a-b)$ kommt p_1 in der

Zerlegung von *m* vor und damit auch in den Zerlegungen von $q_1 - p_1$ oder *b*. Wegen $p_1 \neq q_j$ für $j = 1, ..., s$ scheidet die zweite Möglichkeit aus. Es folgt $p_1 \mid (q_1 - p_1)$ und wegen $p_1 \mid p_1$ der Widerspruch $p_1 \mid q_1$.

Beispielsweise ist $600 = 2^3 \cdot 3 \cdot 5^2$.

Der folgende Satz fasst einige wichtige Eigenschaften von Primzahlen zusammen:

Satz 3.1-3:

(i) Es gibt unendlich viele Primzahlen.

(ii) Es gibt beliebig große Abstände zwischen zwei aufeinanderfolgenden Primzahlen.

(iii) Es gibt unendlich viele Paare $(p, p+2)$, die beide Primzahlen sind (**Primzahlzwillinge**)

(iv) Ist $2^n + 1$ eine Primzahl, so ist *n* eine Zweierpotenz.

(v) Ist $2^n - 1$ eine Primzahl, so ist *n* eine Primzahl.

Zum Nachweis der Aussage (i) wird angenommen, dass es nur endlich viele Primzahlen, etwa $p_1, ..., p_n$, gibt. Dann wird $m = p_1 \cdot ... \cdot p_n + 1$ gesetzt, und es gilt $m > \max\{ p_1, ..., p_n \}$. Dann kann aber *m* nach Annahme keine Primzahl sein. Eine der Primzahlen $p_1, ..., p_n$, etwa p_i, muss aber nach Satz 3.1-2 ein Teiler von *m* sein. Das hätte aber $p_i \mid 1$ zur Folge, ein Widerspruch zu $p_i \geq 2$.

Für den Nachweis von Aussage (ii) werden *n* aufeinanderfolgende natürliche Zahlen angegeben, die sämtlich keine Primzahlen sind: $(n+1)!+2$, $(n+1)!+3$, …, $(n+1)!+(n+1)$. Diese Zahlen sind nacheinander jeweils durch 2, 3, …, $n+1$ teilbar.

Aussage (iii) erfordert weitergehende Betrachtungen aus der Zahlentheorie.

Für den Nachweis von (iv) wird angenommen, dass *n* keine Zweierpotenz ist. Dann hat *n* einen ungeraden Teiler *a*, d.h. $n = a \cdot b$ und $a > 1$ und $b \geq 1$. Setzt man in Satz 1.6-2(ii) $q = -(2^b)$ und den oberen Laufindex der Summe auf den Wert $a - 1$, so ergibt sich

$$\sum_{i=0}^{a-1} \left(-\left(2^b\right)\right)^i = 1 - \left(2^b\right) + \left(2^b\right)^2 - ... \pm ... + \left(2^b\right)^{a-1} = \frac{1 - \left(-\left(2^b\right)\right)^a}{1 - \left(-\left(2^b\right)\right)} = \frac{1 + 2^{a \cdot b}}{1 + 2^b} = \frac{1 + 2^n}{1 + 2^b}.$$

Daher teilt $1+2^b$ die Zahl 2^n+1; diese ist somit keine Primzahl.

Der Nachweis von (v) erfolgt auf ähnliche Weise: Ist *n* keine Primzahl, so hat *n* die Form $n = a \cdot b$ mit $1 < a < n$ und $1 < b < n$. Es gilt

$$1 < \left(2^a\right)^0 + \left(2^a\right)^1 + \ldots + \left(2^a\right)^{b-1} = \sum_{i=0}^{b-1} \left(2^a\right)^i = \frac{\left(2^a\right)^b - 1}{2^a - 1} = \frac{2^n - 1}{2^a - 1} \; .$$

Daher ist $2^n - 1$ keine Primzahl.

In der Praxis der Kryptographie werden ständig große Primzahlen benötigt (mit einer Stellenzahl von mehr als 150 Dezimalstellen). Dabei sind neben Primzahlen, deren Ziffernfolgen keinen festen Gesetzmäßigkeiten unterliegen, Primzahlen der Form 2^n+1 und 2^n-1 besonders interessant. Diese haben nämlich eine sehr einfache Binärdarstellung ($2^n+1 = 1\underbrace{0\ldots0}_{(n-1)\text{-mal}}1$, $2^n-1 = \underbrace{1\ldots1}_{n\text{-mal}}$). Wegen Satz 3.1-3 (iv) kann man die Suche nach sehr großen Primzahlen der Form 2^n+1 auf diejenigen *n* beschränken, die die Form $n = 2^m$ haben, d.h. auf Zahlen der Form $2^n+1 = 2^{2^m}+1$. Zahlen der Form $2^{2^m}+1$ heißen **Fermat**-Zahlen. Beispielsweise sind die Fermat-Zahlen

$2^{2^0}+1 = 3$, $2^{2^1}+1 = 5$, $2^{2^2}+1 = 17$, $2^{2^3}+1 = 257$, $2^{2^4}+1 = 65.537$

Primzahlen. Nicht jede Fermat-Zahl ist jedoch eine Primzahl, wie das Beispiel

$2^{2^5}+1 = 641 \cdot 6\,700\,417$

zeigt. Man kennt heute 230 Fermat-Zahlen, die nachweislich keine Primzahlen sind.

Satz 3.1-1(v) sagt *nicht*, dass jede Zahl der Form 2^p-1 mit einer Primzahl *p* selbst Primzahl ist. Die Zahlen der Form 2^p-1 mit einer Primzahl *p* heißen **Mersenne**-Zahlen. Nicht jede Mersenne-Zahl ist Primzahl. Beispielsweise sind die Zahlen

$2^2-1 = 3$, $2^3-1 = 7$, $2^5-1 = 31$, $2^7-1 = 127$, $2^{13}-1 = 8191$

Primzahlen, nicht aber $2^{11}-1 = 2047 = 23 \cdot 89$. Die bisher bekannten größten Primzahlen sind Mersenne-Zahlen.

Rekordprimzahlen nach Jahren (aus Wikipedia)

Zahl	Anzahl der Dezimalziffern	Jahr	Entdecker (genutzter Computer)
$2^{17}-1$	6	1588	Cataldi
$2^{19}-1$	6	1588	Cataldi
$2^{31}-1$	10	1772	Euler
$(2^{59}-1)/179951$	13	1867	Landry
$2^{127}-1$	39	1876	Lucas
$(2^{148}+1)/17$	44	1951	Ferrier
$180\cdot(2^{127}-1)^2+1$	79	1951	Miller & Wheeler (EDSAC1)
$2^{521}-1$	157	1952	Robinson (SWAC)
$2^{607}-1$	183	1952	Robinson (SWAC)
$2^{1.279}-1$	386	1952	Robinson (SWAC)
$2^{2.203}-1$	664	1952	Robinson (SWAC)
$2^{2.281}-1$	687	1952	Robinson (SWAC)
$2^{3.217}-1$	969	1957	Riesel (BESK)
$2^{4.423}-1$	1.332	1961	Hurwitz (IBM7090)
$2^{9.689}-1$	2.917	1963	Gillies (ILLIAC 2)
$2^{9.941}-1$	2.993	1963	Gillies (ILLIAC 2)
$2^{11.213}-1$	3.376	1963	Gillies (ILLIAC 2)
$2^{19.937}-1$	6.002	1971	Tuckerman (IBM360/91)
$2^{21.701}-1$	6.533	1978	Noll & Nickel (CDC Cyber 174)
$2^{23.209}-1$	6.987	1979	Noll (CDC Cyber 174)
$2^{44.497}-1$	13.395	1979	Nelson & Slowinski (Cray 1)
$2^{86.243}-1$	25.962	1982	Slowinski (Cray 1)
$2^{132.049}-1$	39.751	1983	Slowinski (Cray X-MP)
$2^{216.091}-1$	65.050	1985	Slowinski (Cray X-MP/24)
$2^{216.193}-1$	65.087	1989	„Amdahler Sechs“ (Amdahl 1200)
$2^{756.839}-1$	227.832	1992	Slowinski & Gage (Cray 2)
$2^{859.433}-1$	258.716	1994	Slowinski & Gage (Cray C90)
$2^{1.257.787}-1$	378.632	1996	Slowinski & Gage (Cray T94)
$2^{1.398.269}-1$	420.921	1996	Armengaud, Woltman (GIMPS, Pentium 90 MHz)
$2^{2.976.221}-1$	895.932	1997	Spence, Woltman (GIMPS, Pentium 100 MHz)
$2^{3.021.377}-1$	909.526	1998	Clarkson, Woltman, Kurowski (GIMPS, Pentium 200 MHz)
$2^{6.972.593}-1$	2.098.960	1999	Hajratwala, Woltman, Kurowski (GIMPS, Pentium 350 MHz)
$2^{13.466.917}-1$	4.053.946	2001	Cameron, Woltman, Kurowski (GIMPS, Athlon 800 MHz)
$2^{20.996.011}-1$	6.320.430	2003	Shafer (GIMPS, Pentium 4 2 GHz)
$2^{24.036.583}-1$	7.235.733	2004	Findley (GIMPS, Pentium 4 2,4 GHz)
$2^{25.964.951}-1$	7.816.230	2005	Nowak (GIMPS, Pentium 4 2,4 GHz)
$2^{30.402.457}-1$	9.152.052	2005	Cooper, Boone (GIMPS, Pentium 4 3 GHz)
$2^{32.582.657}-1$	9.808.358	2006	Cooper, Boone (GIMPS, Pentium 4 3 GHz)
$2^{43.112.609}-1$	12.978.189	2008	Smith, Woltman, Kurowski, et al. (GIMPS, Core 2 Duo 2,4 GHz)
$2^{57.885.161}-1$	17.425.170	2013	Cooper, Woltman, Kurowski, et al. (GIMPS)

Einer der wichtigsten Sätze der Zahlentheorie beschreibt die Anzahl der Primzahlen unterhalb einer vorgegebenen Grenze *x*:

Es sei $\pi(x)$ die Anzahl der Primzahlen, die $\leq x$ sind, d.h. $\pi(x) = \sum_{\substack{p \in \mathbf{P} \\ p \leq x}} 1$.

Mit p_n werde die *n*-te Primzahl bezeichnet: $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, ...

Satz 3.1-4:

(i) Es gilt $\lim_{x \to \infty} \frac{\pi(x) \cdot \ln(x)}{x} = 1$, d.h. $\pi(x) \sim \frac{x}{\ln(x)}$ (für große *x*).

(ii) Für $x \geq 67$ ist $\ln(x) - 3/2 < \frac{x}{\pi(x)} < \ln(x) - 1/2$.

(iii) Für $n \geq 20$ ist $n \cdot \left(\ln(n) + \ln(\ln(n)) - 3/2\right) < p_n < n \cdot \left(\ln(n) + \ln(\ln(n)) - 1/2\right)$.

Auf der Grundlage dieser Sätze lässt sich ein sehr effizientes Verfahren zur Erzeugung von (großen) Zahlen angeben, die mit beliebig großer Wahrscheinlichkeit Primzahlen sind. Dabei wird in Kauf genommen, dass das Verfahren eine Zahl eventuell als Primzahl einstuft, die keine Primzahl ist. Die Fehlerwahrscheinlichkeit dieser falschen Entscheidung kann jedoch auf einfache Weise beliebig klein gehalten werden. Man spricht hier von einem **probabilistischen Verfahren (nach dem Monte-Carlo-Prinzip)**.

3.2 Modulare Artihmetik

Es sei $n \in \mathbf{N}$ eine natürliche Zahl mit $n \geq 1$. Auf den ganzen Zahlen **Z** wird durch die folgende Festlegung eine Relation $\equiv$ definiert:

Die Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ heißen **kongruent modulo** *n*, geschrieben $a \equiv b \pmod{n}$, genau dann, wenn gilt: die Zahl *n* teilt $a - b$.

Anders ausgedrückt: $a \equiv b \pmod{n}$ gilt genau dann, wenn es ein $k \in \mathbf{Z}$ mit $a - b = k \cdot n$ gibt.

Beispielsweise gilt

$21 \equiv 0 \,(\mathrm{mod}\, 7)$, $22 \equiv 1 \,(\mathrm{mod}\, 7)$, $23 \equiv 2 \,(\mathrm{mod}\, 7)$, $24 \equiv 3 \,(\mathrm{mod}\, 7)$, $25 \equiv 4 \,(\mathrm{mod}\, 7)$,
$26 \equiv 5 \,(\mathrm{mod}\, 7)$, $27 \equiv 6 \,(\mathrm{mod}\, 7)$,
$28 \equiv 0 \,(\mathrm{mod}\, 7)$, $28 \equiv 21 \,(\mathrm{mod}\, 7)$,
$29 \equiv 1 \,(\mathrm{mod}\, 7)$, $29 \equiv 22 \,(\mathrm{mod}\, 7)$.

Satz 3.2-1:

Es sei $n \in \mathbf{N}$ eine natürliche Zahl mit $n \geq 1$. Die Relation $\equiv$ ist eine Äquivalenzrelation auf den ganzen Zahlen **Z**, d.h. es gilt:

(i) $a \equiv a \,(\mathrm{mod}\, n)$ für jedes $a \in \mathbf{Z}$

(ii) Aus $a \equiv b \,(\mathrm{mod}\, n)$ folgt $b \equiv a \,(\mathrm{mod}\, n)$

(iii) Aus $a \equiv b \,(\mathrm{mod}\, n)$ und $b \equiv c \,(\mathrm{mod}\, n)$ folgt $a \equiv c \,(\mathrm{mod}\, n)$.

Durch Anwendung der Definition lassen sich diese Aussagen nachweisen.

Für $a \in \mathbf{Z}$ bezeichnet $[a]_n = \{ b \mid b \in \mathbf{Z} \text{ und } a \equiv b \,(\mathrm{mod}\, n) \}$ die zu a gehörende **Restklasse (mod n)**.

Beispielsweise ist
$[3]_7 = \{ 3, 10, 17, 24, 31, \ldots \} \cup \{ -4, -11, -18, -25, \ldots \}$
$= \{ m \mid \text{es gibt } k \in \mathbf{Z} \text{ mit } m = k \cdot 7 + 3 \}$.

Allgemein ist für $a \in \mathbf{Z}$

$[a]_n = \{ b \mid b \in \mathbf{Z} \text{ und } a \equiv b \,(\mathrm{mod}\, n) \}$
$= \{ b \mid \text{es gibt } k \in \mathbf{Z} \text{ mit } b = k \cdot n + a \}$.

Da die Relation $\equiv$ eine Äquivalenzrelation ist, folgt mit Satz 1.4-2:

Satz 3.2-2:

Es sei $n \in \mathbf{N}$ eine natürliche Zahl mit $n \geq 1$. Dann gilt:

(i) Es gilt $a \equiv b \,(\mathrm{mod}\, n)$ genau dann, wenn $[a]_n = [b]_n$ ist.

(ii) Jeweils zwei Restklassen $[a]_n$ und $[b]_n$ sind entweder gleich oder disjunkt.

(iii) Es gibt genau n disjunkte Restklassen modulo n, nämlich $[0]_n$, $[1]_n$, $[2]_n$, ..., $[n-1]_n$, und es gilt $\bigcup_{a=0}^{n-1} [a]_n = \mathbf{Z}$.

Jede Restklasse $[a]_n$ besteht aus unendlich vielen Elementen, nämlich aus allen Elementen der Form $k \cdot n + a$ mit $k \in \mathbf{Z}$. Für ein festes $k \in \mathbf{Z}$ sind (wegen Satz 3.2-2 (i)) die Restklassen $[a]_n$ und $[k \cdot n + a]_n$ gleich, d.h. jede Zahl der Form $(k \cdot n + a) \in [a]_n$ repräsentiert die Restklasse $[a]_n$. Man kann daher in jeder Restklasse $[a]_n$ eine Zahl a' mit folgenden Eigenschaften (i) und (ii) finden:

(i) $0 \leq a' < n$
(ii) $a' \equiv a \,(\mathrm{mod}\, n)$, d.h. $[a']_n = [a]_n$.

Für positives $a \in \mathbf{Z}$ erhält man dieses Element a' beispielsweise dadurch, dass man von a so lange den Wert n abzieht, bis die Bedingung $0 \leq a' < n$ erfüllt ist. Für negatives $a \in \mathbf{Z}$ wird der Wert n sukzessive addiert. Dieser kleinste Wert a' mit $0 \leq a' < n$ ist der **Rest bei der ganzzahligen Division** von a durch n und wird mit

$a \bmod n$

bezeichnet.

Beispielsweise ist wegen $3 = 45 - 7 - 7 - 7 - 7 - 7 - 7$: $45 \bmod 7 = 3$ und $[45]_7 = [3]_7$ und $5 = -16 + 7 + 7 + 7$: $-16 \bmod 7 = 5$ und $[-16]_7 = [5]_7$.

Es gilt also:

$0 \leq (a \bmod n) \leq n - 1$ und $[(a \bmod n)]_n = [a]_n$.

Für positives $a \in \mathbf{Z}$ ist nach Konstruktion $(a \bmod n) = a - \lfloor a/n \rfloor \cdot n$;
für negatives $a \in \mathbf{Z}$ ist $(a \bmod n) = a + \lfloor a/n \rfloor \cdot n$.

Beispiele:

$(21 \bmod 7) = 0$, $(28 \bmod 7) = 0$,
$(22 \bmod 7) = 1$, $(29 \bmod 7) = 1$,
$(27 \bmod 7) = 6$, $(6 \bmod 7) = 6$, $(-1 \bmod 7) = 6$.

Für zwei Restklassen $[a]_n$ und $[b]_n$ gilt:

Sind $a_1 \in [a]_n$ und $a_2 \in [a]_n$ bzw. $b_1 \in [b]_n$ und $b_2 \in [b]_n$, dann ist
$a_1 + b_1 \equiv a_2 + b_2 \equiv a + b \pmod n$, d.h. $[a_1 + b_1]_n = [a_2 + b_2]_n = [a + b]_n$.
Entsprechend gilt $a_1 \cdot b_1 \equiv a_2 \cdot b_2 \equiv a \cdot b \pmod n$.

Daher kann man auf eindeutige Weise **arithmetische Operationen** $+_n$ und $\cdot_n$ **auf den Restklassen** (modulo *n*) definieren:

$[a]_n +_n [b]_n = [a + b]_n$ und $[a]_n \cdot_n [b]_n = [a \cdot b]_n$.

Man nimmt also aus jeder Restklasse $[a]_n$ bzw. $[b]_n$ ein beliebiges Element $a_1 \in [a]_n$ bzw. $b_1 \in [b]_n$ und bildet $[a_1 + b_1]_n = [a + b]_n$. Entsprechendes gilt für die Multiplikation. Insbesondere folgt hieraus:

Satz 3.2-3:

(i) $(a \bmod n) + (b \bmod n) \equiv (a + b) \bmod n \pmod n$,
$[a]_n +_n [b]_n = [(a + b) \bmod n]_n$.

(ii) $(a \bmod n) \cdot (b \bmod n) \equiv (a \cdot b) \bmod n \pmod n$,
$[a]_n \cdot_n [b]_n = [(a \cdot b) \bmod n]_n$.

(iii) $b \cdot (a \bmod n) \equiv (a \cdot b) \bmod n \pmod n$.

Die Teile (i) und (iii) sollen hier formal gezeigt werden (Teil (ii) wird ähnlich wie Teil (i) verifiziert):

Da $(a \bmod n) \equiv a \pmod n$ ist, gilt $(a \bmod n) \in [a]_n$. Entsprechend ist $(b \bmod n) \in [b]_n$. Daher ist $[a]_n +_n [b]_n = [(a \bmod n) + (b \bmod n)]_n$ und $[a]_n +_n [b]_n = [a+b]_n = [(a+b) \bmod n]_n$, insgesamt $[(a \bmod n) + (b \bmod n)]_n = [(a+b) \bmod n]_n$ und $(a \bmod n) + (b \bmod n) \equiv (a+b) \bmod n \pmod n$.

Nach Definition ist $(a \bmod n) = a - k \cdot n$ mit $k \in \mathbf{Z}$. Dann ist
$b \cdot (a \bmod n) = a \cdot b - k \cdot b \cdot n = (a \cdot b \bmod n) - k' \cdot n$ mit $k' \in \mathbf{Z}$, d.h.
$b \cdot (a \bmod n) \equiv (a \cdot b) \bmod n \pmod n$.

Beispiele:

$[3]_7 +_7 [6]_7 = [(3+6) \bmod 7]_7 = [2]_7$,
$[3]_7 \cdot_7 [5]_7 = [(3 \cdot 5) \bmod 7]_7 = [1]_7$,
$[3]_7 \cdot_{12} [4]_{12} = [(3 \cdot 4) \bmod 12]_{12} = [0]_{12}$.

Mit $\mathbf{Z}/n\mathbf{Z}$ wird die Menge $\{[0]_n, [1]_n, ..., [n-1]_n\}$ bezeichnet. Häufig findet man auch die Bezeichnung $\mathbf{Z}/n\mathbf{Z} = \{0, 1, ..., n-1\}$ und meint damit die Restklassen modulo *n*. Zusammen mit den oben definierten arithmetischen Operationen auf Restklassen weist die endliche Menge $\mathbf{Z}/n\mathbf{Z}$ sehr ähnliche Eigenschaften zu der unendlichen Menge $\mathbf{Z}$ auf, wie man durch Nachrechnen nachweist:

Satz 3.2-4:

$(\mathbf{Z}/n\mathbf{Z}, +_n, \cdot_n)$ bildet einen kommutativen Ring mit 1. Das neutrale Element der Addition ist $[0]_n = \{a \mid a = k \cdot n \text{ mit } k \in \mathbf{Z}\}$, das neutrale Element der Multiplikation ist $[1]_n = \{a \mid a = k \cdot n + 1 \text{ mit } k \in \mathbf{Z}\}$. Das bezüglich der Addition $+_n$ inverse Element zur Restklasse $[a]_n$ ist die Restklasse $-[a]_n = [-a]_n = [n-a]_n$.

Eine Restklasse $[a]_n$ besitzt bezüglich der Multiplikation $\cdot_n$ genau dann ein inverses Element $[a]_n^{-1}$, wenn $ggT(a,n) = 1$ ist (zur Definition von $ggT(a,n)$ und zur Bestimmung der inversen Restklasse in diesem Fall siehe Kapitel 3.3).

Beispiele:

Die Additions- und Multiplikationstabellen von
$(\mathbf{Z}/7\mathbf{Z}, +_7, \cdot_7) = (\{[0]_7, [1]_7, [2]_7, [3]_7, [4]_7, [5]_7, [6]_7\}, +_7, \cdot_7)$ lauten (statt $[a]_7$ wird zur Vereinfachung *a* geschrieben):

$+_7$	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

$\cdot_7$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	1	3	5
3	3	6	2	5	1	4
4	4	1	5	2	6	3
5	5	3	1	6	4	2
6	6	5	4	3	2	1

In $(\mathbf{Z}/7\mathbf{Z}, +_7, \cdot_7)$ ist das inverse Element bezüglich der Addition zum Element $[3]_7$ das Element $-[3]_7 = [4]_7$ und das inverse Element bezüglich der Multiplikation zum Element $[3]_7$ das Element $[3]_7^{-1} = [5]_7$.

Die Additions- und Multiplikationstabellen von $(\mathbf{Z}/12\mathbf{Z}, +_{12}, \cdot_{12})$ lauten (statt $[a]_{12}$ wird zur Vereinfachung wieder *a* geschrieben):

$+_{12}$	0	1	2	3	4	5	6	7	8	9	10	11
0	0	1	2	3	4	5	6	7	8	9	10	11
1	1	2	3	4	5	6	7	8	9	10	11	0
2	2	3	4	5	6	7	8	9	10	11	0	1
3	3	4	5	6	7	8	9	10	11	0	1	2
4	4	5	6	7	8	9	10	11	0	1	2	3
5	5	6	7	8	9	10	11	0	1	2	3	4
6	6	7	8	9	10	11	0	1	2	3	4	5
7	7	8	9	10	11	0	1	2	3	4	5	6
8	8	9	10	11	0	1	2	3	4	5	6	7
9	9	10	11	0	1	2	3	4	5	6	7	8
10	10	11	0	1	2	3	4	5	6	7	8	9
11	11	0	1	2	3	4	5	6	7	8	9	10

$\cdot_{12}$	1	2	3	4	5	6	7	8	9	10	11
1	1	2	3	4	5	6	7	8	9	10	11
2	2	4	6	8	10	0	2	4	6	8	10
3	3	6	9	0	3	6	9	0	3	6	9
4	4	8	0	4	8	0	4	8	0	4	8
5	5	10	3	8	1	6	11	4	9	2	7
6	6	0	6	0	6	0	6	0	6	0	6
7	7	2	9	4	11	6	1	8	3	10	5
8	8	4	0	8	4	0	8	4	0	8	4
9	9	6	3	0	9	6	3	0	9	6	3
10	10	8	6	4	2	0	10	8	6	4	2
11	11	10	9	8	7	6	5	4	3	2	1

In $(\mathbf{Z}/12\mathbf{Z}, +_{12}, \cdot_{12})$ hat das Element $[3]_{12}$ wegen $[3]_{12} \cdot_{12} [4]_{12} = [(3 \cdot 4) \bmod 12]_{12} = [0]_{12}$ kein inverses Element bezüglich der Multiplikation.

3.3 Der Euklidische Algorithmus

Es seien $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$. Besitzt $d \in \mathbf{Z}$ die Eigenschaften $d \mid a$ und $d \mid b$, dann heißt d **gemeinsamer Teiler** von a und b. Besitzt jeder gemeinsame Teiler c von a und b die Eigenschaft $c \mid d$, dann heißt d **größter gemeinsamer Teiler** von a und b und wird mit $ggT(a,b)$ bezeichnet.

Zur Bestimmung des größten gemeinsamen Teilers zweier Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ könnte man diese gemäß Satz 3.1-2 in ihre Primfaktoren zerlegen und alle gemeinsamen Primfaktoren in ihrer gemeinsamen Vielfachheit herausziehen. Beispielsweise ist $792 = 2^3 \cdot 3^2 \cdot 11$ und $84 = 2^2 \cdot 3 \cdot 7$, d.h. $ggT(792,84) = 2^2 \cdot 3 = 12$. Dieses Verfahren (Schulmethode) ist höchstens für kleine Zahlen praktisch einsetzbar; denn für große Zahlen (in der Praxis mit mehr als 100 Dezimalstellen) stößt man auf Effizienzgrenzen. Als äußerst effizient zur Bestimmung des größten gemeinsamen Teilers zweier ganzer Zahlen hat sich der **Euklidische Algorithmus** erwiesen (Euklid, um 325 v. Chr.). Dieses Verfahren geht läuft folgendermaßen ab:

Für die beiden Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ kann $a \geq b$ angenommen werden. Da es bei der Bestimmung von Teilern nicht auf das Vorzeichen ankommt, kann weiterhin $b > 0$ angenommen werden, so dass insgesamt $a \geq b > 0$ ist. Es werden ganze Zahlen m_1 und r_1 bestimmt mit

$a = m_1 \cdot b + r_1$ und $0 \leq r_1 < b$.

Durch die Festlegung $0 \le r_1 < b$ sind m_1 und r_1 eindeutig bestimmt: $r_1 = a \bmod b$ und $m_1 = \lfloor a/b \rfloor$. Für $r_1 = 0$ endet das Verfahren hier, und es ist $ggT(a,b) = b$. Ansonsten werden ganze Zahlen m_2 und r_2 bestimmt mit

$b = m_2 \cdot r_1 + r_2$ und $0 \le r_2 < r_1$.

Man sieht, dass *b* die Rolle von *a* und r_1 die Rolle von *b* übernimmt. Wieder sind durch die Festlegung $0 \le r_2 < r_1$ die Werte m_2 und r_2 eindeutig bestimmt. Für $r_2 = 0$ endet das Verfahren hier, und es ist $ggT(a,b) = r_1$ (das muss man natürlich mathematisch beweisen). Ansonsten werden ganze Zahlen m_3 und r_3 bestimmt mit

$r_1 = m_3 \cdot r_2 + r_3$ und $0 \le r_3 < r_2$.

Man sieht, dass r_1 die Rolle von *b* und r_2 die Rolle von r_1 übernimmt. Das Verfahren wird so lange fortgesetzt, bis zum ersten Mal der Rest $r_n = 0$ entsteht. Insgesamt lassen sich die einzelnen Schritte wie folgt zusammenfassen:

Man bestimmt ganze Zahlen m_1 und r_1 mit	$a = m_1 \cdot b + r_1$ und $0 < r_1 < b$.
Man bestimmt ganze Zahlen m_2 und r_2 mit	$b = m_2 \cdot r_1 + r_2$ und $0 < r_2 < r_1$.
Man bestimmt ganze Zahlen m_3 und r_3 mit	$r_1 = m_3 \cdot r_2 + r_3$ und $0 < r_3 < r_2$.
usw.	
Man bestimmt ganze Zahlen m_n und r_n mit	$r_{n-2} = m_n \cdot r_{n-1} + r_n$ und $0 < r_n < r_{n-1}$.
Fortsetzung des Verfahrens, bis	$r_{n-1} = m_{n+1} \cdot r_n + 0$ gilt.

Es gilt $b > r_1 > r_2 > ... > r_{n-1} > r_n > 0$, d.h. die Reste $r_1, r_2, ..., r_n$ werden immer kleiner, so dass das Verfahren abbricht.

Liest man das Schema von unten nach oben, so sieht man, dass r_n die Zahl r_{n-1} teilt (letzte Zeile); aus der vorletzten Zeile erkennt man, dass r_n dann auch r_{n-2} teilt usw. Bis zur zweiten Zeile folgt, dass r_n die Werte r_2 und r_1 und damit *b* teilt. Die 1. Zeile liefert schließlich die Teilbarkeit von *a* durch r_n.

Das Lesen des Schemas von oben nach unten zeigt, dass ein gemeinsamer Teiler von *a* und *b* alle Werte r_i für $i = 1, ..., n$ teilt. Daher gilt:

Satz 3.3-1:

Das beschriebene Verfahren bestimmt den größten gemeinsamen Teiler zweier ganzer Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ mit $b \neq 0$, und zwar gilt

$ggT(a,b) = r_n$,

d.h. der größte gemeinsame Teiler von *a* und *b* ist der letzte von 0 verschiedene Rest.

Die folgende Pascal-Funktion `ggT` ist eine Implementierung des Verfahrens; sie bestimmt den größten gemeinsamen Teiler der als Parameter übergebenen ganzen Zahlen *a* und *b*. Die Anzahl der von ihm ausgeführten arithmetischen Operationen ist proportional zur Länge der Zahlendarstellung von *a* und *b*.

```
FUNCTION ggT (a : INTEGER; b : INTEGER) : INTEGER;

VAR   r : INTEGER;
      s : INTEGER;
      t : INTEGER;
      m : INTEGER;

BEGIN { ggT }
   r := b;
   s := a;

   WHILE r <> 0 DO
     BEGIN
       { t und s aus der vorherigen Iteration neu besetzen }
       t := s;
       s := r;
       { bilde t =  m * s + r }
       m := t DIV s;
       r := t - m*s
     END;

   { der größte gemeinsame Teiler ist der letzte von 0
     verschiedene Rest                                  }

   ggT := s

END   { ggT };
```

Beispiele:

a: 28 *b*: 15 t = m * s + r	*a*: 198 *b*: 84 t = m * s + r	*a*: 84 *b*: 198 t = m * s + r
28 = 1 * 15 + 13	198 = 2 * 84 + 30	84 = 0 * 198 + 84
15 = 1 * 13 + 2	84 = 2 * 30 + 24	198 = 2 * 84 + 30
13 = 6 * 2 + 1	30 = 1 * 24 + 6	84 = 2 * 30 + 24
2 = 2 * 1 + 0	24 = 4 * 6 + 0	30 = 1 * 24 + 6
		24 = 4 * 6 + 0
ggT (28, 15) = 1	ggT (198, 84) = 6	ggT (84, 198) = 6

Das obige Zahlenschema (eine typische Zeile *i* ist hinzugefügt)

$a = m_1 \cdot b + r_1$ und $0 < r_1 < b$, Zeile 1

$b = m_2 \cdot r_1 + r_2$ und $0 < r_2 < r_1$, Zeile 2

$r_1 = m_3 \cdot r_2 + r_3$ und $0 < r_3 < r_2$, Zeile 3

...

$r_{i-2} = m_i \cdot r_{i-1} + r_i$ und $0 < r_i < r_{i-1}$, Zeile *i*

...

$r_{n-2} = m_n \cdot r_{n-1} + r_n$ und $0 < r_n < r_{n-1}$, Zeile *n*

$r_{n-1} = m_{n+1} \cdot r_n + 0$, Zeile $n+1$

$ggT(a,b) = r_n$

Mit $r_1 = a \bmod b$ ergibt sich unmittelbar

Satz 3.3-2:

Für zwei Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ ist $ggT(a,b) = \begin{cases} a & \text{für } b = 0 \\ ggT(b, a \bmod b) & \text{für } b \neq 0 \end{cases}$.

Löst man in dem Zahlenschema die Zeilen *i* für $i = 1, ..., n$ nach r_i auf, so lassen sich ganze Zahlen $a_1, ..., a_n$ und $b_1, ..., b_n$ definieren, für die gilt:

Zeile 1: $r_1 = a - m_1 \cdot b$

$= 1 \cdot a + (-m_1) \cdot b$

$= a_1 \cdot a + b_1 \cdot b$ mit $a_1 = 1$, $b_1 = -m_1$.

Zeile 2: $r_2 = b - m_2 \cdot r_1$

$= b - m_2 \cdot (a - m_1 \cdot b)$

$= -m_2 \cdot a + (1 + m_1 \cdot m_2) \cdot b$

$= a_2 \cdot a + b_2 \cdot b$ mit $a_2 = -m_2$, $b_2 = 1 + m_1 \cdot m_2$.

Angenommen, in allen Zeilen $l = 1, ..., i-1$ ließe sich der jeweilige Rest r_l in der Form

$r_l = a_l \cdot a + b_l \cdot b$

schreiben. Dann geht das auch in Zeile *i*:

Zeile *i*: $r_i = r_{i-2} - m_i \cdot r_{i-1}$

$= a_{i-2} \cdot a + b_{i-2} \cdot b - m_i \cdot (a_{i-1} \cdot a + b_{i-1} \cdot b)$

$= (a_{i-2} - m_i \cdot a_{i-1}) \cdot a + (b_{i-2} - m_i \cdot b_{i-1}) \cdot b$

$= a_i \cdot a + b_i \cdot b$ mit $a_i = a_{i-2} - m_i \cdot a_{i-1}$, $b_i = b_{i-2} - m_i \cdot b_{i-1}$.

Insbesondere

Zeile *n*: $ggT(a,b) = r_n = a_n \cdot a + b_n \cdot b$.

Die Folgen $a_1, ..., a_n$ und $b_1, ..., b_n$ werden also rekursiv definiert durch:

$a_{-1} = 1$, $a_0 = 0$, $a_i = a_{i-2} - m_i \cdot a_{i-1}$ für $i = 1, ..., n$

$b_{-1} = 0$, $b_0 = 1$, $b_i = b_{i-2} - m_i \cdot b_{i-1}$ für $i = 1, ..., n$.

Die Berechnung dieser beiden Folgen kann in den Euklidischen Algorithmus direkt eingebaut werden. Die Pascal-Funktion `ggt` wird erweitert zur Funktion `invers` (die Wahl des Prozedurbezeichners ergibt sich aus den anschließenden Bemerkungen zu Satz 3.3-5).

```
PROCEDURE  invers (a : LONGINT; b : LONGINT;
                   VAR a_inv : LONGINT;
                   VAR b_inv : LONGINT;
                   VAR ggt   : LONGINT);

{ die Funktion berechnet zu a und b ganze
  Zahlen a_inv und b_inv mit a*a_inv + b*b_inv = ggT(a, b) }

VAR    r        : LONGINT;
       s        : LONGINT;
       t        : LONGINT;
       m        : LONGINT;
       a_min_2  : LONGINT;
       a_min_1  : LONGINT;
       b_min_2  : LONGINT;
       b_min_1  : LONGINT;
       store    : LONGINT;

BEGIN
  r       := b;
  s       := a;
  a_min_2 := 1;
  a_min_1 := 0;
  b_min_2 := 0;
  b_min_1 := 1;

WHILE r <> 0 DO
  BEGIN
    { t und s aus der vorigen Iteration neu besetzen }
    t := s;
    s := r;

    { bilde t = m * s + r }
    m := t DIV s;
    r := t - m*s;

    store   := a_min_2;
    a_min_2 := a_min_1;
    a_min_1 := store - m * a_min_1;
    store   := b_min_2;
    b_min_2 := b_min_1;
    b_min_1 := store - m * b_min_1
   END;

{ der ggT (a, m) ist der letzte von 0 verschiedene Rest,
   d. h. der gegenwärtige Wert von s                    }
   ggT := s;
   a_inv := a_min_2;
   b_inv := b_min_2

END { invers };
```

Satz 3.3-3:

Zu zwei Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ gibt es eindeutig bestimmte Zahlen $a' \in \mathbf{Z}$ und $b' \in \mathbf{Z}$ mit

$a \cdot a' + b \cdot b' = ggT(a,b)$.

Die Zahlen a' und b' lassen sich mit der Pascal-Funktion `invers`, einer Erweiterung des Euklidischen Algorithmus, bestimmen.

Der folgende Satz stellt einige wichtige Eigenschaften des *ggT* zusammen:

Satz 3.3-4:

(i) Es sei $d = ggT(a,b)$. Dann gibt es Zahlen $a_1 \in \mathbf{Z}$ und $b_1 \in \mathbf{Z}$ mit $a = d \cdot a_1$ und $b = d \cdot b_1$ und $ggT(a_1,b_1) = 1$.

(ii) Es gilt $ggT(a,b) = 1$ genau dann, wenn es Zahlen $x \in \mathbf{Z}$ und $y \in \mathbf{Z}$ gibt mit $a \cdot x + b \cdot y = 1$.

(iii) Es sei $ggT(a,b) = 1$. Falls *a* das Produkt $b \cdot c$ teilt, dann teilt *a* die Zahl *c*.

(iv) Es sei $ggT(a,b) = 1$. Falls *a* die Zahl *c* teilt und *b* die Zahl *c* teilt, dann teilt $a \cdot b$ die Zahl *c*.

Nur Teil (ii) bedarf einer kurzen Erläuterung: Ist $ggT(a,b) = 1$, so besagt Satz 3.3-3: Es gibt eindeutig bestimmte Zahlen $a' \in \mathbf{Z}$ und $b' \in \mathbf{Z}$ mit $a \cdot a' + b \cdot b' = ggT(a,b) = 1$. Gilt umgekehrt $a \cdot x + b \cdot y = 1$ mit $x \in \mathbf{Z}$ und $y \in \mathbf{Z}$, so teilt $ggT(a,b)$ den Wert 1, daher ist $ggT(a,b) = 1$.

In vielen Anwendungen spielen **lineare Kongruenzen** eine wichtige Rolle. Dabei handelt es sich um Gleichungen der Form $a \cdot x \equiv b \pmod{n}$, wobei *a* und *b* vorgegebene ganze Zahlen sind und $n > 1$ eine natürliche Zahl ist. Gesucht wird nach einer ganzzahligen Lösung *x*. Die folgenden Sätze sagen aus, wann eine lineare Kongruenz lösbar ist. In diesem Fall lassen sich die Lösungen mit Hilfe der angegebenen Prozedur `invers`, d.h. im wesentlichen mit Hilfe des Euklidischen Algorithmus bestimmen.

Satz 3.3-5:

Es sei $ggT(a,n)=1$.

Dann hat die lineare Kongruenz $a\cdot x\equiv b\,(\mathrm{mod}\,n)$ eine Lösung. Alle Lösungen sind kongruent modulo *n*. Man sagt daher, dass die lineare Kongruenz $a\cdot x\equiv b\,(\mathrm{mod}\,n)$ in diesem Fall genau eine Lösung modulo *n* besitzt.

Nach Satz 3.3-3 lassen sich zu *a* und *n* mit der Prozedur `invers` Zahlen a' und n' finden, für die $a\cdot a'+n\cdot n'=ggT(a,n)=1$ gilt. Die gesuchte Lösung lautet dann $x=\left((a'\cdot b)\,\mathrm{mod}\,n\right)$. Diese Lösung ist modulo *n* eindeutig.

In Satz 3.2-4 wird behauptet, dass eine Restklasse $[a]_n$ genau dann ein bezüglich der Multiplikation $\cdot_n$ inverses Element $[a]_n^{-1}$ besitzt, wenn $ggT(a,n)=1$ gilt. Dazu bestimmt man wie oben die Zahlen a' und n' mit $a\cdot a'+n\cdot n'=ggT(a,n)=1$. Wegen $a\cdot a'\equiv 1\,(\mathrm{mod}\,n)$ gilt $[a\cdot a']_n=[a]_n\cdot_n[a']_n=[1]_n$. Daher kann man $[a]_n^{-1}=[a']_n=[a'\,\mathrm{mod}\,n]_n$ setzen.

Eine Verallgemeinerung von Satz 3.3-5 ist der folgende Satz.

Satz 3.3-6:

Es sei $ggT(a,n)=d$. Dann hat die lineare Kongruenz $a\cdot x\equiv b\,(\mathrm{mod}\,n)$ genau dann Lösungen, wenn *d* ein Teiler von *b* ist.

Die Aussage des Satzes ergibt sich aus folgenden Überlegungen:
Hat die lineare Kongruenz $a\cdot x\equiv b\,(\mathrm{mod}\,n)$ eine Lösung *x*, so gibt es ein $l\in\mathbf{Z}$ mit $a\cdot x-b=l\cdot n$ bzw. $a\cdot x-l\cdot n=b$. Daher teilt $ggT(a,n)$ die Zahl *b*. Ist umgekehrt $ggT(a,n)=d$ ein Teiler von *b*, so ist $b=d\cdot b_1$. Es werden Zahlen a' und n' bestimmt, für die $a\cdot a'+n\cdot n'=ggT(a,n)=d$ gilt. Dann ist $b_1\cdot a\cdot a'+b_1\cdot n\cdot n'=b_1\cdot d=b$ und $a\cdot(b_1\cdot a')\equiv b\,(\mathrm{mod}\,n)$.

Ist *d* ein Teiler von *b*, so gilt $b=d\cdot b_1$ mit einer ganzen Zahl b_1. Alle Lösungen der linearen Kongruenz $a\cdot x\equiv b\,(\mathrm{mod}\,n)$ erhält man folgendermaßen:
Nach Satz 3.3-4 (i) gibt es Zahlen $a_1\in\mathbf{Z}$ und $n_1\in\mathbf{Z}$ mit $a=d\cdot a_1$ und $n=d\cdot n_1$ und $ggT(a_1,n_1)=1$. Nach Satz 3.3-5 wird die modulo n_1 eindeutige Lösung *y* der linearen Kon-

gruenz $a_1 \cdot y \equiv b_1 \pmod{n_1}$ bestimmt. Alle Lösungen (es sind genau d viele) der linearen Kongruenz $a \cdot x \equiv b \pmod{n}$ lauten dann:
$y,\ y+n_1,\ y+2\cdot n_1,\ ...,\ y+(d-1)\cdot n_1$.
In der Tat gilt $a\cdot(y+i\cdot n_1)\equiv b \pmod{n}$ für $i=0,...,d-1$: Nach Konstruktion ist $a_1\cdot(y+i\cdot n_1)\equiv a_1\cdot y\equiv b_1 \pmod{n_1}$. Daher teilt $n=d\cdot n_1$ den Wert $d\cdot(a_1\cdot(y+i\cdot n_1)-b_1)$ und weiter den Wert $d\cdot a_1\cdot y - d\cdot b_1$. Das bedeutet $a\cdot y\equiv b \pmod{n}$.

3.4 Weitere ausgewählte Ergebnisse der elementaren Zahlentheorie

Die **Eulersche φ-Funktion (Phi-Funktion)** wird für jede natürliche Zahl $n\geq 1$ definiert durch die Anzahl der natürlichen Zahlen a zwischen 1 und n (einschließlich) mit $ggT(a,n)=1$, d.h.

$$\varphi(n)=\sum_{\substack{a\\ 1\leq a\leq n,\\ ggT(a,n)=1}} 1\ .$$

Satz 3.4-1:

(i) Ist p eine Primzahl, dann ist $\varphi(p)=p-1$. Gilt umgekehrt $\varphi(n)=n-1$, dann ist n eine Primzahl.

(ii) Ist p eine Primzahl, dann ist $\varphi(p^k)=p^k-p^{k-1}$.

(iii) Für natürliche Zahlen n und m mit $ggT(n,m)=1$ ist $\varphi(n\cdot m)=\varphi(n)\cdot\varphi(m)$.

(iv) $$\varphi(n)=n\cdot\prod_{\substack{p \text{ teilt } n\\ p \text{ ist Primzahl}}}\left(1-1/p\right).$$

Teil (i) folgt unmittelbar aus der Definition.

Für Teil (ii) wird die Menge $M=\{1,2,3,...,p^k\}$ betrachtet, die p^k viele Elemente enthält. Gesucht ist die Anzahl der Zahlen a in M mit $ggT(a,p^k)=1$. Es gilt nun:
Für die Zahl a ist genau dann $ggT(a,p^k)\neq 1$, wenn a ein Vielfaches von p ist. Zur Begründung: Gilt $ggT(a,p^k)=d\neq 1$, dann teilt d die Zahl p^k, hat also die Form $d=p^{k'}$ mit $k'\geq 1$,

und a hat die Form $a = d \cdot a' = p^{k'} \cdot a'$. Daher ist a ein Vielfaches von p. Ist umgekehrt a ein Vielfaches von p, etwa $a = a' \cdot p$, dann ist $ggT(a,p^k) = ggT(a' \cdot p, p^k) \geq p \neq 1$.
Die Vielfachen von p in M sind $1 \cdot p, 2 \cdot p, ..., \left(p^{k-1} - 1\right) \cdot p, p^{k-1} \cdot p$; das sind p^{k-1} viele Werte. Daher ist $\varphi(p^k) = p^k - p^{k-1}$.

Für Teil (iii) kann man folgendermaßen argumentieren:
Wegen $ggT(n,m) = 1$ gibt es nach Satz 3.3-4(ii) eindeutig bestimmte Zahlen n' und m' mit $n \cdot n' + m \cdot m' = 1$. Es wird eine Abbildung f definiert durch

$$f : \begin{cases} (\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/m\mathbf{Z}) & \to \quad \mathbf{Z}/(n \cdot m)\mathbf{Z} \\ ([x]_n, [y]_m) & \to \quad [y + m \cdot m' \cdot (x - y)]_{n \cdot m} \end{cases}.$$

Gilt $f(([x_1]_n, [y_1]_m)) = f(([x_2]_n, [y_2]_m))$, dann ist
$[y_1 + m \cdot m' \cdot (x_1 - y_1)]_{n \cdot m} = [y_2 + m \cdot m' \cdot (x_2 - y_2)]_{n \cdot m}$, d.h.
$y_1 + m \cdot m' \cdot (x_1 - y_1) \equiv y_2 + m \cdot m' \cdot (x_2 - y_2) \mod(n \cdot m)$ bzw.
$y_1 - y_2 \equiv m \cdot m' \cdot ((x_2 - y_2) - (x_1 - y_1)) \mod(n \cdot m)$, also
$y_1 - y_2 - m \cdot m' \cdot ((x_2 - y_2) - (x_1 - y_1)) = t \cdot n \cdot m$ für ein $t \in \mathbf{Z}$. Daher ist $y_1 - y_2 \equiv 0 \mod(m)$, $y_1 \equiv y_2 \mod(m)$ und $[y_1]_m = [y_2]_m$.
Mit $m \cdot m' = 1 - n \cdot n'$ folgt weiter $f(([x]_n, [y]_m)) = [y + m \cdot m' \cdot (x - y)]_{n \cdot m} = [x + n \cdot n' \cdot (y - x)]_{n \cdot m}$.
Damit ergibt sich (wie mit y_1 und y_2) aus $f(([x_1]_n, [y_1]_m)) = f(([x_2]_n, [y_2]_m))$: $[x_1]_n = [x_2]_n$. Die Abbildung f ist also injektiv und (da Definitions- und Zielmenge gleichmächtig sind) sogar bijektiv (vgl. Satz 2.2-4). Die Umkehrabbildung zu f lautet

$$f^{-1} : \begin{cases} \mathbf{Z}/(n \cdot m)\mathbf{Z} & \to \quad (\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/m\mathbf{Z}) \\ [z]_{n \cdot m} & \to \quad ([z \bmod n]_n, [z \bmod m]_m) \end{cases}:$$

Ist nämlich $(z \bmod n) = z - t \cdot n$ bzw. $(z \bmod m) = z - s \cdot m$ mit $t \in \mathbf{Z}$ bzw. $s \in \mathbf{Z}$, dann ist

$$\begin{aligned} f(([z \bmod n]_n, [z \bmod m]_m)) &= [z - s \cdot m + m \cdot m' \cdot (z - t \cdot n - (z - s \cdot m))]_{n \cdot m} \\ &= [z - s \cdot m \cdot (1 - m \cdot m') - m \cdot m' \cdot t \cdot n]_{n \cdot m} \\ &= [z - s \cdot m \cdot n \cdot n' - m \cdot m' \cdot t \cdot n]_{n \cdot m} \\ &= [z]_{n \cdot m} \ . \end{aligned}$$

Die Anzahl der Restklassenpaare $([x]_n, [y]_m) \in (\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/m\mathbf{Z})$ mit $ggT(x,n) = 1$ und $ggT(y,m) = 1$ ist $\varphi(n) \cdot \varphi(m)$. Die Anzahl der Restklassen $[z]_{n \cdot m} \in \mathbf{Z}/(n \cdot m)\mathbf{Z}$ mit $ggT(z, n \cdot m) = 1$ ist $\varphi(n \cdot m)$.
Die Abbildung f bildet jedes Restklassenpaar $([x]_n, [y]_m) \in (\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/m\mathbf{Z})$ mit $ggT(x,n) = 1$ und $ggT(y,m) = 1$ auf eine Restklasse $[z]_{n \cdot m} \in \mathbf{Z}/(n \cdot m)\mathbf{Z}$ mit $ggT(z, n \cdot m) = 1$ ab:
Es seien $[x]_n \in \mathbf{Z}/n\mathbf{Z}$ mit $ggT(x,n) = 1$ und $[y]_m \in \mathbf{Z}/m\mathbf{Z}$ mit $ggT(y,m) = 1$. Dann ist
$f(([x]_n, [y]_m)) = [y + m \cdot m' \cdot (x - y)]_{n \cdot m} = [x + n \cdot n' \cdot (y - x)]_{n \cdot m}$. Es sei
$d = ggT(y + m \cdot m' \cdot (x - y), n \cdot m)$, also $y + m \cdot m' \cdot (x - y) = d \cdot t_1$ und $n \cdot m = d \cdot t_2$ mit $t_1 \in \mathbf{Z}$ und $t_2 \in \mathbf{Z}$. Dann folgt nacheinander

$(y + m \cdot m' \cdot (x - y)) \cdot t_2 = d \cdot t_1 \cdot t_2 = t_1 \cdot n \cdot m$,
$(y + m \cdot m' \cdot (x - y)) \cdot t_2 = (x + n \cdot n' \cdot (y - x)) \cdot t_2 = t_1 \cdot n \cdot m$,
$y \cdot t_2 \equiv 0 \mod(m)$, $x \cdot t_2 \equiv 0 \mod(n)$.
Da $ggT(x,n) = 1$ und $ggT(y,m) = 1$ gelten, sind $[x]_n$ in $\mathbf{Z}/n\mathbf{Z}$ und $[y]_m$ in $\mathbf{Z}/m\mathbf{Z}$ nach Satz 3.2-4 invertierbar. Daher folgt aus $y \cdot t_2 \equiv 0 \mod(m)$ (bzw. $[y]_m \cdot_m [t_2]_m = [0]_m$) und aus $x \cdot t_2 \equiv 0 \mod(n)$ (bzw. $[x]_n \cdot_n [t_2]_n = [0]_n$) : $t_2 \equiv 0 \mod(m)$ und $t_2 \equiv 0 \mod(n)$.[3] Da $ggT(n,m) = 1$ ist, ergibt sich $t_2 = k \cdot n \cdot m$ für ein $k \in \mathbf{Z}$. Damit folgt nacheinander $n \cdot m = d \cdot t_2 = d \cdot k \cdot n \cdot m$, $d \cdot k = 1$ und $d = ggT(y + m \cdot m' \cdot (x - y), n \cdot m) = 1$.
Das bedeutet $\varphi(n) \cdot \varphi(m) \le \varphi(n \cdot m)$.
Es sei umgekehrt $[z]_{n \cdot m} \in \mathbf{Z}/(n \cdot m)\mathbf{Z}$ mit $ggT(z, n \cdot m) = 1$. Dann ist $[z]_{n \cdot m}$ in $\mathbf{Z}/(n \cdot m)\mathbf{Z}$ nach Satz 3.2-4 invertierbar, d.h. es gibt $[z']_{n \cdot m} \in \mathbf{Z}/(n \cdot m)\mathbf{Z}$ mit $[z]_{n \cdot m} \cdot_{n \cdot m} [z']_{n \cdot m} = [1]_{n \cdot m}$, also $z \cdot z' \equiv 1 \mod(n \cdot m)$. Damit folgt $z \cdot z' \equiv 1 \mod(n)$ und $z \cdot z' \equiv 1 \mod(m)$ und weiter $(z \bmod n) \cdot z' \equiv 1 \mod(n)$ und $(z \bmod m) \cdot z' \equiv 1 \mod(m)$. Durch Übergang auf die Restklassen ergibt sich $[z \bmod n]_n \cdot_n [z']_n = [1]_n$ und $[z \bmod m]_m \cdot_m [z']_m = [1]_m$. Die Restklasse $[z \bmod n]_n$ ist also in $\mathbf{Z}/n\mathbf{Z}$ und die Restklasse $[z \bmod m]_m$ in $\mathbf{Z}/m\mathbf{Z}$ invertierbar. Mit Satz 3.2-4 folgt $ggT((z \bmod n), n) = 1$ und $ggT((z \bmod m), m) = 1$. Die Restklasse $[z]_{n \cdot m}$ wird also durch $\overset{-1}{f}$ auf ein Restklassenpaar $([z \bmod n]_n, [z \bmod m]_m)$ abgebildet, das zum jeweiligen Modul teilerfremd ist. Daher gilt $\varphi(n \cdot m) \le \varphi(n) \cdot \varphi(m)$.

Teil (iv) folgt aus (iii) und Satz 3.1-2: Die Zahl *n* wird in ihre Primzahlpotenzen zerlegt: $n = p_1^{e_1} \cdot p_2^{e_2} \cdot \ldots \cdot p_r^{e_r}$. Dann ist

$$\begin{aligned}
\varphi(n) &= \varphi\left(p_1^{e_1} \cdot p_2^{e_2} \cdot \ldots \cdot p_r^{e_r}\right) \\
&= \varphi\left(p_1^{e_1}\right) \cdot \varphi\left(p_2^{e_2}\right) \cdot \ldots \cdot \varphi\left(p_r^{e_r}\right) \\
&= \left(p_1^{e_1} - p_1^{e_1 - 1}\right) \cdot \left(p_2^{e_2} - p_2^{e_2 - 1}\right) \cdot \ldots \cdot \left(p_r^{e_r} - p_r^{e_r - 1}\right) \\
&= p_1^{e_1} \cdot \left(1 - 1/p_1\right) \cdot p_2^{e_2} \cdot \left(1 - 1/p_2\right) \cdot \ldots \cdot p_r^{e_r} \cdot \left(1 - 1/p_r\right) \\
&= n \cdot \prod_{\substack{p \text{ teilt } n \\ p \text{ ist Primzahl}}} \left(1 - 1/p\right).
\end{aligned}$$

Der folgende Satz (Satz von Euler) ist wichtig für die Korrektheit des Public Key Encryption-Verfahrens RSA:

---

[3] In jedem Ring gilt $a \otimes 0 = a \otimes (a \oplus (-a)) = a \otimes a \oplus a \otimes (-a) = a \otimes a \oplus (-a \otimes a) = 0$. Genauso folgt $0 \otimes a = 0$. Besitzt *a* in dem Ring ein multiplikativ inverses Element a^{-1}, dann folgt für jedes Element *b* des Rings aus $a \otimes b = 0$: $0 = a^{-1} \otimes 0 = a^{-1} \otimes (a \otimes b) = (a^{-1} \otimes a) \otimes b = 1 \otimes b = b$.

Satz 3.4-2:

Es seien a und n natürliche Zahlen mit $ggT(a,n)=1$. Dann ist $a^{\varphi(n)} \equiv 1 \,(\mathrm{mod}\, n)$.

Zum Beweisdes Satzes seien a und n natürliche Zahlen mit $ggT(a,n)=1$. Außerdem sei a_1 eine natürliche Zahl mit $0 \le a_1 < n$ und $ggT(a_1,n)=1$. Dann gilt auch $ggT(a \cdot a_1, n)=1$; denn: Es sei $d = ggT(a \cdot a_1, n)$. Es gilt $d \mid a \cdot a_1$ und $d \mid n$. Aus Satz 3.3-3 folgt die Existenz zweier ganzer Zahlen a_1' und n' mit $a_1' \cdot a_1 + n' \cdot n = 1$. Daher ist $a_1' \cdot a \cdot a_1 + n' \cdot a \cdot n = a$. Damit folgt $d \mid a$. Wegen $ggT(a,n)=1$ ist $d=1$.

Die Menge $M = \{ a_1, ..., a_{\varphi(n)} \}$ bestehe aus allen natürlichen Zahlen a_i mit $1 \le a_i < n$ und $ggT(a_i,n)=1$. Für jedes dieser a_i gilt $ggT(a \cdot a_i, n)=1$ und damit auch $ggT(a \cdot a_i \bmod n, n)=1$: Es ist nämlich $(a \cdot a_i \bmod n) = a \cdot a_i - k \cdot n$ für ein $k \in \mathbf{Z}$; ist d ein Teiler von $a \cdot a_i \bmod n$ und von n, dann ist d auch ein Teiler von $a \cdot a_i$ und damit $d=1$.
Weiterhin gilt für $i \ne j$: $(a \cdot a_i \bmod n) \ne (a \cdot a_j \bmod n)$: Ist nämlich $(a \cdot a_i \bmod n) = (a \cdot a_j \bmod n)$, dann ist $a \cdot a_i \equiv a \cdot a_j \bmod n$, und n teilt das Produkt $a \cdot (a_i - a_j)$; wegen $ggT(a,n)=1$ und $-n < a_i - a_j < n$ ist $a_i = a_j$ bzw. $i=j$.

Damit ergibt sich $M = \{ a_1, ..., a_{\varphi(n)} \} = \{ (a \cdot a_1 \bmod n), ..., (a \cdot a_{\varphi(n)} \bmod n) \}$ und $a_1 \cdot ... \cdot a_{\varphi(n)} = (a \cdot a_1 \bmod n) \cdot ... \cdot (a \cdot a_{\varphi(n)} \bmod n)$. Mit Satz 3.2-3 (iii) folgt $(a \cdot a_1 \bmod n) \cdot ... \cdot (a \cdot a_{\varphi(n)} \bmod n) \equiv a^{\varphi(n)} \cdot (a_1 \bmod n) \cdot ... \cdot (a_{\varphi(n)} \bmod n) \,(\mathrm{mod}\, n)$. Da $a_i \equiv (a_i \bmod n) (\mathrm{mod}\, n)$ ist, folgt $a_1 \cdot ... \cdot a_{\varphi(n)} \equiv a^{\varphi(n)} \cdot a_1 \cdot ... \cdot a_{\varphi(n)} \bmod n$ und $(a^{\varphi(n)} - 1) \cdot (a_1 \cdot ... \cdot a_{\varphi(n)}) \equiv 0 \bmod n$. Wegen $ggT(a_i,n)=1$ ist $(a^{\varphi(n)} - 1) \equiv 0 \bmod n$.

Der folgende Satz (Satz von Fermat) ist ein Spezialfall von Satz 3.4-2:

Satz 3.4-3:

(i) Es sei a eine natürliche Zahl und p eine Primzahl mit $ggT(a,p)=1$. Dann ist $a^{p-1} \equiv 1 \,(\mathrm{mod}\, p)$.

(ii) Es sei a eine natürliche Zahl und n eine ungerade natürliche Zahl mit $ggT(a,n)=1$. Gilt *nicht* $a^{n-1} \equiv 1 \,(\mathrm{mod}\, n)$, dann ist n keine Primzahl.

Mit Satz 3.3-5 wurde das zu einer Restklasse $[a]_n$ bezüglich der Multiplikation $\cdot_n$ inverse Element $[a]_n^{-1}$ bestimmt, falls $ggT(a,n)=1$ gilt. Nach Satz 3.4-2 gilt (bei $ggT(a,n)=1$): $a \cdot a^{\varphi(n)-1} = a^{\varphi(n)} \equiv 1 \pmod{n}$. Daher ist $[a]_n^{-1} = \left[a^{\varphi(n)-1}\right]_n$. Dieses Ergebnis führt (in Erweiterung von Satz 3.3-5) auf

Satz 3.4-4:

Es sei $ggT(a,n)=1$. Dann hat die lineare Kongruenz $a \cdot x \equiv b \pmod{n}$ genau eine Lösung modulo *n*, nämlich $x = \left[b \cdot a^{\varphi(n)-1}\right]_n$, d.h. für alle Lösungen *x* der linearen Kongruenz $a \cdot x \equiv b \pmod{n}$ gilt $x \equiv b \cdot a^{\varphi(n)-1} \pmod{n}$.

3.5 Anwendung in der Kryptologie

Die folgende Abbildung zeigt das grundlegende Szenario, in dem kryptographische Verfahren zur Datenverschlüsselung und –entschlüsselung eingesetzt werden.

Vertrauliche Daten werden von einem Sender *A* zu einem Empfänger *B* gesandt. Fragen der korrekten Datenübertragung sollen in diesem Zusammenhang ausgeklammert werden. Es soll lediglich garantiert werden, dass ein unberechtigter Dritter, der die Daten während der Übertragungsphase eventuell mithört, diese inhaltlich nicht interpretieren kann. Dieser „Angreifer“ auf das Übertragungssystem wird als **Kryptoanalytiker** bezeichnet; seine Tätigkeit heißt **Kryptoanalyse**. Zum Schutz werden die Daten vor ihrer Übertragung vom Sender verschlüsselt. Die unverschlüsselten Daten werden als **Klartext** bezeichnet, die verschlüsselten Daten als **Schlüsseltext (Chiffretext)**. Der Schlüsseltext wird zum Empfänger gesendet und dort von diesem entschlüsselt, so dass er wieder den Klartext erhält. Zwischen Sender und Empfänger sind also **Absprachen** über das verwendete Verschlüsselungs- und Entschlüsselungsverfahren notwendig.

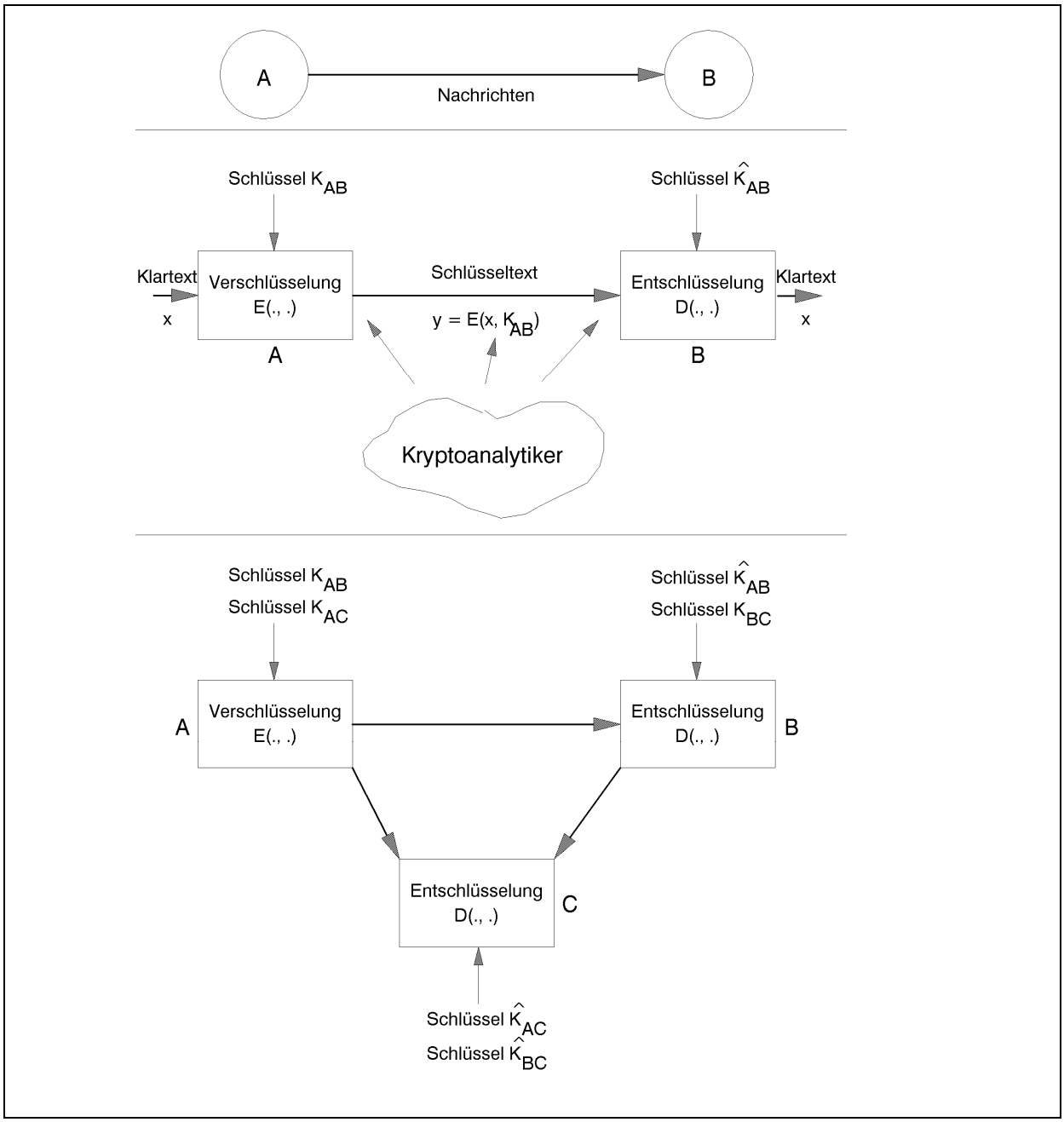


Abbildung: Datenverschlüsselung und -entschlüsselung

Die im folgenden beschriebenen Verfahren zur Verschlüsselung und Entschlüsselung von Daten zwischen einem Sender *A* und einem Empfänger *B* bestehen formal aus mehreren Teilen:

1. Mit dem **Verschlüsselungsalgorithmus** *E* (encryption, Verschlüsselung) verschlüsselt der Sender *A* Klartexte. Der Verschlüsselungsalgorithmus *E* hat zwei Eingabeparameter, nämlich einen Klartext *x* und einen **Schlüssel** (key) K_{AB}, mit dem **alle Klartexte**, die **von *A* nach *B* laufen**, verschlüsselt werden. Für eine Kommunikationsbeziehung von *A* an einen Empfänger *C* mit $C \neq B$ wird ein Schlüssel $K_{AC} \neq K_{AB}$, aber derselbe Verschlüsselungsalgorithmus *E* verwendet.

   Der aus einem Klartext *x* entstehende Schlüsseltext ist $y = E(x, K_{AB})$.

2. Beim Empfänger *B* ankommende Schlüsseltexte werden von ihm mit Hilfe des **Entschlüsselungsalgorithmus** *D* (decryption, Entschlüsselung) entschlüsselt. Der Entschlüsselungsalgorithmus *D* hat ebenfalls zwei Eingabeparameter, nämlich einen Schlüsseltext *y* und einen Schlüssel $\hat{K}_{AB}$, mit dem alle Nachrichten, die von *A* nach *B* laufen, entschlüsselt werden. Zwischen den Algorithmen *E* und *D* und den Schlüsseln K_{AB} und $\hat{K}_{AB}$ besteht die Beziehung

   $$D\left(E\left(x, K_{AB}\right), \hat{K}_{AB}\right) = x,$$

   d.h. der gesendete Klartext kann aus dem empfangenen Schlüsseltext bei korrekter Verwendung der Verfahren wieder gewonnen werden. Der Sender *A* muss den vom Empfänger *B* eingesetzten Schlüssel $\hat{K}_{AB}$ zum Entschlüsseln eines Schlüsseltextes nicht notwendigerweise kennen.

3. Das Schlüsselpaar $\left(K_{AB}, \hat{K}_{AB}\right)$ wird für die Ver- bzw. Entschlüsselung aller Klartexte verwendet, die von *A* nach *B* laufen. Für die umgekehrte Kommunikationsrichtung ist eventuell ein anderes Schlüsselpaar erforderlich ebenso für die Kommunikation zwischen anderen Teilnehmern.

Einige **grundlegende Anforderungen an ein kryptographisches Verfahren** sind:

(i) Die Berechnung von $y = E\left(x, K_{AB}\right)$ aus *x* und K_{AB} (Verschlüsselung) muss vom Sender auf einfache Weise, d.h. mit geringem algorithmischen Aufwand, durchführbar sein. Außerdem sollte der Schlüsseltext $y = E\left(x, K_{AB}\right)$ nicht wesentlich länger als der zugehörige Klartext *x* sein. Natürlich wird dabei vorausgesetzt, dass der Sender über geeignete Rechenkapazität verfügt.

(ii) Die Berechnung von *x* aus einer empfangenen Nachricht der Form $y = E\left(x, K_{AB}\right)$ mit Hilfe von $\hat{K}_{AB}$ (Entschlüsselung) muss vom Empfänger ebenfalls auf einfache Weise, d.h. mit geringem algorithmischen Aufwand, durchführbar sein. Auch hier wird vorausgesetzt, dass der Empfänger über geeignete Rechenkapazität verfügt.

(iii) Ohne Kenntnis des Schlüssels $\hat{K}_{AB}$ zum Entschlüsseln ist es „unmöglich“, aus $y = E\left(x, K_{AB}\right)$ auf den Klartext *x* zu schließen. Systematisches Probieren aller Werte, die als eventuelle Schlüssel $\hat{K}_{AB}$ in Frage kommen, ist mit einem derart großen algorithmischen Aufwand verbunden, dass diese experimentelle Suche praktisch nicht durchführbar ist.

(iv) Die Verschlüsselungs- bzw. Entschlüsselungsalgorithmen E bzw. D sollten nicht geheim gehalten werden. Abgesehen davon, dass eine Geheimhaltung wahrscheinlich nur temporär möglich ist, wird durch eine Offenlegung von E und D erreicht, dass die Verfahren mathematisch analysiert und eventuelle Schwachstellen aufgedeckt und behoben werden können. Zusätzlich lässt sich eine korrekte Implementierung der Verfahren verifizieren.

Die **Angriffe** durch einen unbefugten Kryptoanalytiker **auf ein Verschlüsselungsverfahren** lassen sich in verschiedene Gruppen einteilen:

- Der Kryptoanalytiker versucht, aus der Kenntnis von $y = E(x, K_{AB})$ den Klartext x zu erhalten. Man nennt diesen Angriff **Cipher-text-only-Attacke**. Diese Form der Attacke ist die schwierigste, denn im Normalfall hat man wenig Informationen darüber, welchen Inhalt der Klartext x aufweist. Gleichzeitig ist sie aber auch diejenige, die in der Praxis am häufigsten vorkommt. Ein anderes Ziel eines Kryptoanalytikers bei einer Cipher-text-only-Attacke ist die Ermittlung des Schlüssels $\hat{K}_{AB}$ zum Entschlüsseln aus der Kenntnis einer oder mehrerer verschlüsselter Nachrichten. Damit können dann spätere verschlüsselte Texte entschlüsselt werden.

- Der Kryptoanalytiker kennt eine von ihm nicht beeinflußte Auswahl von Klartexten x_1, ..., x_n mit den zugehörigen Schlüsseltexten $E(x_1, K_{AB})$, ..., $E(x_n, K_{AB})$ und versucht daraus, das Schlüsselpaar $(K_{AB}, \hat{K}_{AB})$ abzuleiten. Man nennt diesen Angriff **Known-plaintext-Attacke**. Eine derartige Attacke ist häufig dann möglich, wenn sich Nachrichten oder Teile davon wiederholen. Wenn Klartexte beispielsweise immer denselben Briefkopf oder dieselbe Anrede verwenden, sind zumindest Teile eines Klartextes bekannt.

- Der Kryptoanalytiker kann selbst eine Auswahl von Klartexten x_1, ..., x_n vorschlagen und sieht die zugehörigen Schlüsseltexte $E(x_1, K_{AB})$, ..., $E(x_n, K_{AB})$. Er wählt die Klartexte so, dass er daraus eventuell leicht auf das verwendete Schlüsselpaar $(K_{AB}, \hat{K}_{AB})$ schließen kann. Man nennt diesen Angriff **Chosen-plaintext-Attacke**. Ein gutes kryptographisches Verfahren muss gegen Chosen-plaintext-Attacke resistent sein.

- Der Kryptoanalytiker kennt das Verschlüsselungsverfahren E und das Entschlüsselungsverfahren D einschließlich des verwendeten Schlüssels K_{AB} zum Verschlüsseln eines Klartextes (eine typische Situation der Public-Key-Encryption-Verfahren). Er verfügt über viel Zeit und Rechnerleistung, um aus dieser Kenntnis den Schlüssel $\hat{K}_{AB}$ zu ermitteln.

Bei einer Chosen-Plaintext-Attacke kann man versuchen, systematisch alle möglichen Schlüssel $\hat{K}_{AB}$ auszuprobieren (ein in der Praxis durchaus gängiger Ansatz). Dabei hofft man natürlich, schon nach wenigen Versuchen auf den richtigen Schlüssel zu stoßen. Eine derartige Attacke heißt **Brute–Force–Attacke**. Man muss sich jedoch darüber im klaren sein, dass die Anzahl auszuprobierender Schlüssel exponentiell wächst. Geht man davon aus, dass der Schlüssel $\hat{K}_{AB}$ eine Binärzahl der Länge *n* ist, so gibt es 2^n viele Kandidaten für $\hat{K}_{AB}$. Um eine Vorstellung von der Größenordnung dieser Zahl zu bekommen, wird angenommen, dass die Erzeugung und das Ausprobieren eines einzigen Schlüssels nur 10^{-9} Sekunden benötigt. Dann dauert eine Brute–Force–Attacke bei einer Schlüssellänge von 56 Bits (eine heute nicht mehr als sicher angesehene Schlüssellänge), d.h. das Durchprobieren sämtlicher $2^{56} \approx 7{,}20576 \cdot 10^{16}$ verschiedener Schlüssel, insgesamt mehr als 8,34 Tage benötigt. Bei einer Schlüssellänge von 64 Bits braucht man dann bereits etwa 584 Jahre, um alle Schlüssel zu erzeugen. Nimmt man an, dass bei einer Schlüssellänge von 56 Bits alle Schlüssel in nur 1 Sekunde ausprobiert werden können, dann ergeben sich die folgenden Werte:

Schlüssellänge *n*	Anzahl an Schlüsseln	Aufwand zur Erzeugung aller 2^n Schlüssel
56 Bits	$7{,}20576 \cdot 10^{16}$	1 Sekunde (angenommen)
64 Bits	$1{,}84467 \cdot 10^{19}$	4 Minuten 16 Sekunden
80 Bits	$1{,}20893 \cdot 10^{24}$	194 Tage
112 Bits	$5{,}19230 \cdot 10^{33}$	$\approx 2{,}285 \cdot 10^{9}$ Jahre
128 Bits	$3{,}40282 \cdot 10^{38}$	$\approx 1{,}497 \cdot 10^{14}$ Jahre
192 Bits	$6{,}27710 \cdot 10^{57}$	$\approx 2{,}7623 \cdot 10^{33}$ Jahre
256 Bits	$1{,}15792 \cdot 10^{77}$	$\approx 5{,}0956 \cdot 10^{52}$ Jahre

Nimmt man an, dass die Erzeugung eines Schlüssels in einem Rechner die Zeit *t* benötigt, so beträgt der Aufwand zur Erzeugung aller Schlüssel der Länge *n* die Zeit $t \cdot 2^n$. Die minimale Zeit für einen Schaltvorgang in einem Rechner beträgt aus physikalischen Gründen (u.a. weil sich Elektronen mit einer Geschwindigkeit bewegen, die die Lichtgeschwindigkeit nicht überschreitet) mindestens $t_m \approx 5{,}6 \cdot 10^{-33}$ Sekunden. Setzt man für *t* diesen Wert ein, so beträgt die Dauer in einer Brute-Force-Attacke bei einer Schlüssellänge von 128 Bits allein zur Erzeugung aller 2^{128} Schlüssel immer noch mehr als 22 Tage. Das zeigt, dass eine Brute-Force-Attacke auf die Güte der Verschlüsselung nur unter massiver Parallelisierung sinnvoll ist, indem die Menge aller zu probierender Schlüssel auf eine Vielzahl gleichzeitig agierender Kryptoanalytiker aufgeteilt wird.

Bei den **symmetrischen Kryptologieverfahren** werden für jede Kommunikationsbeziehung zwischen einem Sender *A* und einem Empfänger *B* zum Ver- und Entschlüsseln dieselben Schlüssel verwendet, d.h. es gilt $K_{AB} = \hat{K}_{AB}$. Der Sender verwendet den Schlüssel, um die

Nachricht zu verschlüsseln und der Empfänger, um diese zu entschlüsseln. Folglich muss sowohl der Sender als auch der Empfänger denselben Schlüssel K_{AB} kennen und gegenüber Dritten, auch anderen Kommunikationsteilnehmern, geheimhalten. Aus diesem Grund bietet sich an, den Schlüssel K_{AB} auch für die Kommunikationsrichtung von *B* nach *A* zu verwenden. Es gilt dann $K_{AB} = \hat{K}_{AB} = K_{BA} = \hat{K}_{BA}$.

Bei den **asymmetrischen Kryptologieverfahren** werden verschiedene Schlüssel zum Verschlüsseln und Entschlüsseln der Nachrichten verwendet. Eine für die Praxis bedeutende Klasse asymmetrischer Verschlüsselungsverfahren bilden die **öffentlichen Verschlüsselungsverfahren (PKE-Verfahren, public key encryption)**. Zunächst soll das allgemeine Prinzip eines PKE-Verfahrens am Nachrichtenaustausch zwischen einem Sender *A* und einem Empfänger *B* und weiteren Teilnehmern *C* erläutert werden.

Jeder Kommunikationsteilnehmer *B*, der von anderen Kommunikationsteilnehmern verschlüsselte Nachrichten empfangen möchte, stellt einen Schlüssel c_B in einem **öffentlichen Register** bereit, auf das alle Kommunikationsteilnehmer zugreifen können. Der Schlüssel c_B („codieren“) dient allen Kommunikationsteilnehmern zur Verschlüsselung von Nachrichten, die an *B* gesendet werden. Zusätzlich besitzt jeder Empfänger *B* einen **geheimen Schlüssel** d_B („decodieren“), mit dem er alle Nachrichten entschlüsselt, die an ihn gesandt wurden.

Oben wurde der Schlüssel zum Verschlüsseln einer Nachricht von *A* nach *B* mit K_{AB} bezeichnet. Um auszudrücken, dass alle Kommunikationsteilnehmer denselben Schlüssel für Nachrichten an *B* verwenden, wird er hier c_B (anstelle von K_{AB} bzw. K_{CB}) geschrieben. Entsprechend wird hier nicht die allgemeine Bezeichnung $\hat{K}_{AB}$ für den Schlüssel zum Entschlüsseln einer Nachricht verwendet, die *B* von einem Kommunikationsteilnehmer *A* empfangen hat, sondern d_B, da *B* den Schlüssel d_B zum Entschlüsseln aller Nachrichten an ihn, unabhängig vom Absender, verwendet.

Das Eintragen des öffentlichen Schlüssels c_B in das Register unterliegt keiner Geheimhaltung, da dieser Schlüssel ja sowieso öffentlich ist. Das Problem der Schlüsselverteilung wie bei symmetrischen Verfahren stellt sich hier nicht.

Ein Klartext *x*, der von *A* nach *B* verschlüsselt gesandt werden soll, wird von *A* in den Schlüsseltext $y = E(x, c_B)$ transformiert. Eine von *B* empfangene verschlüsselte Nachricht *y* wird von *B* in $D(y, d_B)$ entschlüsselt.

Die folgende Abbildung zeigt drei Kommunikationsteilnehmer *A*, *B* und *C* mit den jeweiligen Schlüsseln.

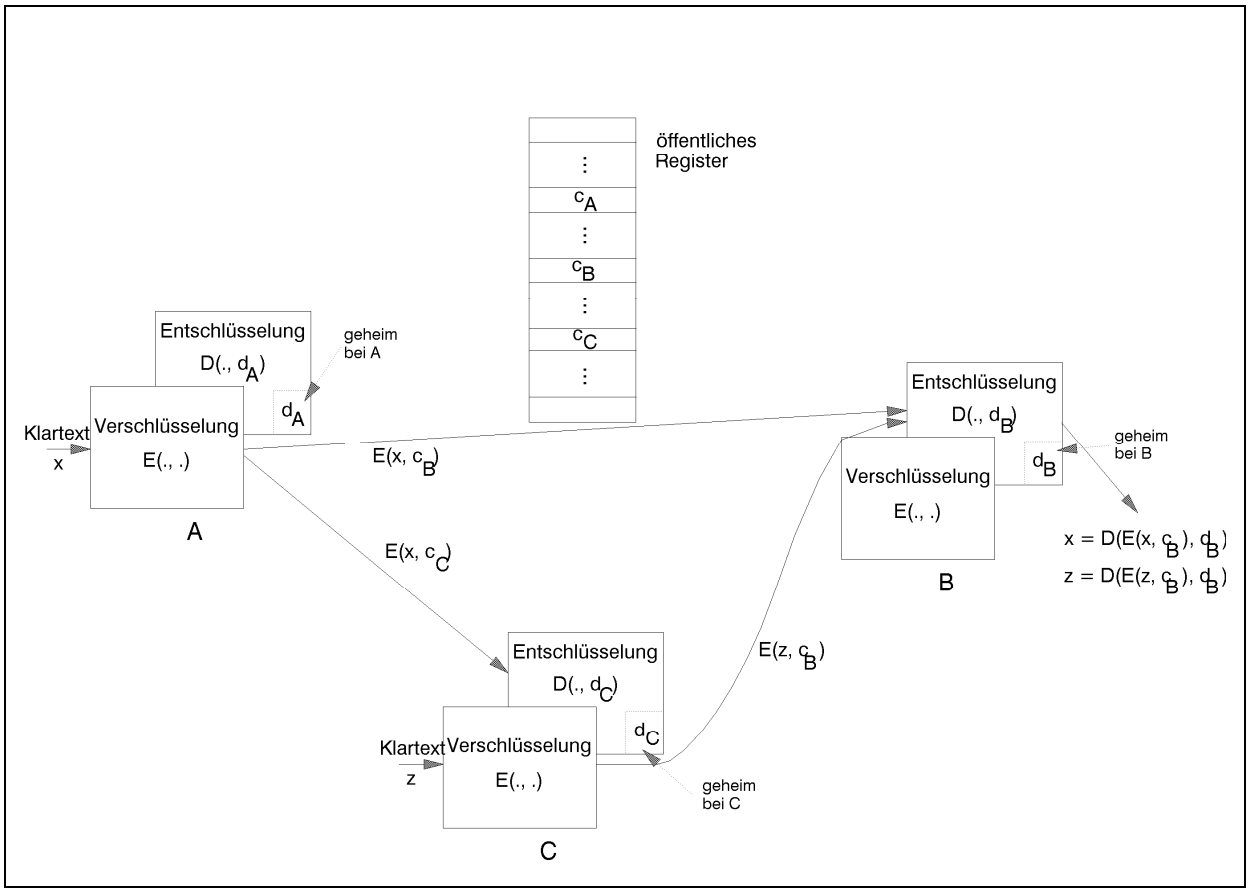


Abbildung: Verschlüsselung und Entschlüsselung mit asymmetrischen Verfahren

Verschlüsselungs- und Entschlüsselungsverfahren müssen folgende Bedingungen erfüllen:

(i) Ein Empfänger *B* kann eine empfangene verschlüsselte Nachricht mit seinem Schlüssel korrekt entschlüsseln, d.h. $D(E(x, c_B), d_B) = x$.

(ii) Die Verschlüsselung einer Nachricht, d.h. die Berechnung von $E(x, c_B)$, und die Entschlüsselung einer Nachricht bei Kenntnis des Schlüssels d_B, d.h. die Berechnung von $D(y, d_B)$, sind mit geringem Rechenaufwand durchzuführen.

(iii) Aus der Kenntnis eines öffentlichen Schlüssels c_B zum Verschlüsseln der Nachrichten an einen Empfänger *B* kann man „nicht leicht“ auf den bei *B* geheim gehaltenen Schlüssel d_B schließen. Die Forderung, eine Berechnung „nicht leicht“ durchführen zu können, wird mathematisch exakt durch den Begriff „**intractable**“ umschrieben, der ausdrückt, dass es zur Berechnung (beweisbar) keinen schnell ausführbaren Algorithmus gibt.

(iv) Ohne d_B zu kennen, kann ein Kryptoanalytiker aus einem Schlüsseltext $E(x, c_B)$ nicht leicht *x* ermitteln. Die Verschlüsselungsfunktion $E(.,.)$ stellt eine sogenannte **Einweg-**

funktion mit Falltür dar. Erst wenn man die geheime Zusatzinformation d_B (die **Falltürinformation**) kennt, kann man die zu *E* inverse Funktion leicht berechnen.

(v) Zur Realisierung eines Unterschriftenprotokolls wird zusätzlich die Vertauschbarkeit der Verschlüsselung und Entschlüsselung gefordert. Neben der in (i) formulierten Bedingung $D(E(x, c_B), d_B) = x$ gilt auch $E(D(y, d_B), c_B) = y$.

In der Literatur sind eine Reihe von PKE-Verfahren veröffentlicht. Ihre Sicherheit ist mit Einschränkungen mathematisch beweisbar und hat sich in der Praxis bewährt; die Einschränkung bezieht sich auf eine bisher unbewiesene mathematische Vermutung bezüglich der Komplexität nichtdeterministischer Rechenverfahren (das sogenannte P-NP-Problem bzw. gewisser zahlentheoretischer Problemstellungen).

Zur Beschreibung eines PKA-Verfahrens muss angegeben werden, wie ein Kommunikationsteilnehmer *B* seinen geheimen Schlüssel d_B und seinen öffentlichen Schlüssel c_B festlegt, und wie die Verschlüsselungs- bzw. Entschlüsselungsalgorithmen $E(.,.)$ bzw. $D(.,.)$ definiert sind.

Das bekannteste PKE-Verfahren wird nach seinen Entdeckern Rivest, Shamir und Adleman ***RSA*-Verfahren** genannt[4]. Es bietet bei sorgfältiger Auswahl einiger im Verfahren frei wählbarer Parameter und entsprechender Implementierung eine sehr hohe Sicherheit. Es ist ein rein **softwaremäßig implementiertes Verfahren**. Dadurch ist seine Verschlüsselungs- bzw. Entschlüsselungsgeschwindigkeit etwa um den Faktor 1.000 langsamer als beispielsweise bei *DES* (gängiges symmetrisches Verfahren, das hardwaremäßig implementierbar ist). Es erfordert eine besondere Arithmetik natürlicher Zahlen mit sehr großen Stellenzahlen. Daher eignet es sich zur Verschlüsselung langer Nachrichten bzw. zur online-Verschlüsselung nur begrenzt. Ein „Hybrid"-Verfahren, das den Vorteil der Schnelligkeit von *Tripel-DES* bzw. *IDEA* beim Verschlüsseln und Entschlüsseln mit der Sicherheit von *RSA* verbinden, ist das seit 1991 über das Internet als Shareware verbreitete ***PGP*-Verfahren** (**pretty good privacy**), das sich besonders im E-Mail-Bereich bewährt hat.

Im folgenden werden einige Details des *RSA*-Verfahrens beschrieben. Das Verfahren beruht auf mathematischen, insbesondere zahlentheoretischen Grundlagen, Erkenntnissen der Komplexitätstheorie und dem Einsatz sehr großer Zahlen (mehr als 300 Dezimalstellen).

Im RSA-Verfahren wird die Modulo-Arithmetik für ganze Zahlen (im folgenden nur natürliche Zahlen, d.h. nicht-negative ganze Zahlen) eingesetzt. Neben Additionen und Multiplikationen wird auch die Exponentiation verwendet.

---

[4] Rivest, M.; Shamir, A.; Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems, Comm. ACM, 21, S.120-126, 1978.

Die Festlegung des öffentlichen Schlüssels c_B für die Verschlüsselung von Klartexten an einen Empfänger *B* und des geheimen Schlüssels d_B zum Entschlüsseln eines Schlüsseltextes beim Empfänger *B* verläuft im *RSA*-Verfahren wie folgt.

1. Es werden zwei verschiedene sehr große Primzahlen *p* und *q* ausgewählt, z.B. in der Größenordnung von 150 Dezimalstellen. Dann wird die Zahl $n = p \cdot q$ gebildet. Die Zahl *n* hat dann mindestens 300 Dezimalstellen, d.h. etwa 1.000 Binärstellen; in der Praxis wählt man *p* und *q* so, dass die Zahl *n* eine Binärstellenzahl von 1.024 aufweist.

   Zum Auffinden von Primzahlen in dieser Größenordnung und zum Testen auf Primzahleigenschaft kennt man schnelle Verfahren.

2. Für *B* wird eine Zufallszahl $d > \max\{p, q\}$ ausgewählt, die mit $\varphi(n) = (p-1)\cdot(q-1)$ keinen gemeinsamen Teiler ausser 1 besitzt. Der Wert *d* darf nicht zu klein sein, da er Teil des geheim gehalten Schlüssels zum Entschlüsseln ist und daher von einem Kryptoanalytiker nicht durch systematisches Probieren gefunden werden darf.

3. Mit Hilfe der Erweiterung des Euklidischen Algorithmus (Funktion `invers`) ermittelt man zu *d* und $\varphi(n)$ zwei Zahlen d' und f' mit $d \cdot d' + \varphi(n) \cdot f' = 1$ und setzt $e = d' \bmod \varphi(n)$. Dann ist $0 < e < \varphi(n)$ und $e = d' - k \cdot \varphi(n)$ mit einem Wert $k \in \mathbf{Z}$. Es ist $e \cdot d = (d' - k \cdot \varphi(n)) \cdot d = 1 - f' \cdot \varphi(n) - d \cdot k \cdot \varphi(n) = 1 - f \cdot \varphi(n)$ mit $f = f' + d \cdot k$. Insgesamt gilt

   $0 < e < \varphi(n)$, $e \cdot d + f \cdot \varphi(n) = 1$, $e \cdot d \equiv 1 \,(\mathrm{mod}\, \varphi(n))$ und $(e \cdot d \bmod \varphi(n)) = 1$.

   Außerdem ist nach Satz 3.3-4(ii) $ggT(e, \varphi(n)) = 1$.

4. Der von *B* **geheim gehaltene Schlüssel zum Entschlüsseln** von Nachrichten, **die an *B* gesendet werden**, besteht aus der Zahlenfolge $d_B = [d, p, q, \varphi(n)]$. Zum Entschlüsseln wird nur *d* verwendet, es ist jedoch unbedingt erforderlich, die Werte *p*, *q* und $\varphi(n)$ ebenfalls geheim zu halten, da ein Kryptoanalytiker aus der Kenntnis des öffentlichen Schlüssels (siehe 5.) und aus der Kenntnis eines der Werte *p*, *q* oder $\varphi(n)$ den geheimen Schlüsselteil *d* leicht ermitteln kann (siehe unten). Die Zahlenfolge $d_B = [d, p, q, \varphi(n)]$ stellt die Falltürinformation dar.

5. Der in das öffentliche Register eingetragene Schlüssel zum Verschlüsseln aller Nachrichten an *B* ist die Zahlenfolge $c_B = [e,\ n]$.

Die **Vorschrift zur Verschlüsselung** von Nachrichten an *B* lautet:

Ein eventuell sehr lange Klartext *x* wird als Binärmuster aufgefasst und in Blöcke x_i mit jeweils $\lfloor \log_2(n) \rfloor$ vielen Stellen aufgeteilt: $x = x_1 x_2 \ldots x_r$. Eventuell wird dabei der letzte Teilblock x_r mit binären Nullen aufgefüllt. Jeder Teilblock x_i kann als Binärzahl interpretiert werden, die einen Wert $0 \le x_i < 2^{\log_2(n)} = n$ hat. Der Klartext *x* wird blockweise verschlüsselt; die einzelnen verschlüsselten Klartextblöcke werden dann wieder zu einem Schlüsseltext *y* zusammengesetzt:

Die Verschlüsselung eines Blockes x_i lautet
$y_i = E(x_i, c_B) = E(x_i, [e, n]) = (x_i^e \bmod n)$.

Diese Zahl kann wieder als Binärmuster mit $\lfloor \log_2(n) \rfloor$ vielen Stellen aufgefasst werden.

Die Hintereinanderreihung aller so entstandenen Binärmuster $y_1, y_2, ..., y_r$ ergibt den zu *x* gehörenden Schlüsseltexte $y = E(x, c_B) = E(x_1, c_B) E(x_2, c_B) \ldots E(x_r, c_B)$.

Es gibt sehr effiziente Algorithmen zur Berechnung von $y_i = (x_i^e \bmod n)$, so dass die Verschlüsselung schnell erfolgen kann.

Eine bei *B* ankommende verschlüsselte Nachricht *y* wird zur **Entschlüsselung** als Binärmuster interpretiert und in einzelne Blöcke mit $\lfloor \log_2(n) \rfloor$ vielen Stellen zerlegt, d.h. $y = y_1 y_2 \ldots y_r$. Jeder Block y_i wird einzeln nach folgender Vorschrift entschlüsselt:
$D(y_i, d_B) = D(y_i, [d, p, q, \phi(n)]) = (y_i^d \bmod n)$.
Die so entstehenden Zahlen werden als Bitmuster mit jeweils mit $\lfloor \log_2(n) \rfloor$ vielen Stellen interpretiert und durch Hintereinanderreihung zum entschlüsselten Text zusammengesetzt. Der algorithmische Aufwand zur Entschlüsselung ist wie bei der Verschlüsselung klein.

Die Korrektheit des Verfahrens, nämlich $D(E(x_i, c_B), d_B) = ((x_i^e)^d \bmod n) = x_i$, folgt aus Satz 3.4-2:

Dazu werden 3 Fälle unterschieden:

1. Fall: Weder *p* noch *q* teilen x_i. Dann gilt $ggT(x_i, n) = 1$ und mit Satz 3.4-2: $x_i^{\varphi(n)} \equiv 1 \,(\bmod n)$. Nach Konstruktion ist $e \cdot d + f \cdot \varphi(n) = 1$ bzw.

$e \cdot d = 1 + (-f) \cdot \varphi(n)$. Also $(x_i^e)^d \equiv x_i^{1+(-f)\cdot\varphi(n)} \equiv x_i \cdot (x_i^{\varphi(n)})^{(-f)} \equiv x_i \pmod{n}$. Da $x_i < n$ ist, ergibt sich $((x_i^e)^d \bmod n) = x_i$.

2. Fall: Genau eine der Zahlen p oder q teilt x_i; es sei dieses p. Dann gilt (wieder mit Satz 3.4-2): $x_i^{\varphi(q)} \equiv 1 \pmod{q}$. Damit folgt nacheinander $x_i^{(-f)\cdot(p-1)\cdot\varphi(q)} \equiv 1 \pmod{q}$, $x_i \cdot x_i^{(-f)\cdot(p-1)\cdot\varphi(q)} = x_i^{1+(-f)\cdot(p-1)\cdot\varphi(q)} = x_i^{e\cdot d} \equiv x_i \pmod{q}$, d.h. q teilt $x_i^{e\cdot d} - x_i$. Da nach Fallannahme die Zahl p den Wert x_i teilt, teilt p auch $x_i^{e\cdot d}$, und daher teilt p den Wert $x_i^{e\cdot d} - x_i$. Mit Satz 3.3-4 (iv) folgt: $n = p \cdot q$ teilt $x_i^{e\cdot d} - x_i$, d.h. $x_i^{e\cdot d} \equiv x_i \pmod{n}$. Da $x_i < n$ ist, ergibt sich wie im 1. Fall: $((x_i^e)^d \bmod n) = x_i$.

3. Fall: Beide Zahlen p und q teilen x_i. Dann teilen p und q den Wert $x_i^{e\cdot d} - x_i$ und mit Satz 3.3-4 (iv) folgt: $n = p \cdot q$ teilt $x_i^{e\cdot d} - x_i$, d.h. $x_i^{e\cdot d} \equiv x_i \pmod{n}$. Da $x_i < n$ ist, ergibt sich wie im 1. Fall: $((x_i^e)^d \bmod n) = x_i$.

Es gilt außerdem die Symmetriegleichung $E(D(y_i, d_B), c_B) = ((y_i^d)^e \bmod n) = y_i$, so dass das *RSA*-Verfahren für ein digitales Unterschriftenprotokoll geeignet ist.

Bei der Konstruktion des geheimen Schlüssels $d_B = [d, p, q, \varphi(n)]$ besteht eine gewisse Freiheit bezüglich der Wahl der einzelnen Komponenten. Beispielsweise kann man den Wert *d* so groß wählen, dass er von einem Kryptoanalytiker nicht leicht durch systematisches Testen gefunden werden kann. Der Exponent *e* zum Verschlüsseln eines Klartextes an *B* ist nach Wahl von *d* eindeutig bestimmt. Der umgekehrte Weg, nämlich erst *e* zu wählen, und zwar so, dass *e* und $\varphi(n) = (p-1)\cdot(q-1)$ teilerfremd sind, und dann mit Hilfe des Euklidischen Algorithmus *d* zu ermitteln, ist ebenfalls möglich. Auf diese Weise kann man für *e* einen „günstigen" Wert nehmen. Als günstige Werte haben sich die Primzahlen $e = 3$, $e = 17$ und $e = 65.537$ erwiesen, da diese Fermat-Zahlen in ihrer Binärdarstellung nur jeweils zwei binäre Einsen haben und damit die in der Verschlüsselung auszuführende Exponentiation sehr schnell abläuft.

Die folgenden **Empfehlungen bezüglich der im Verfahren auszuwählenden Zahlen** zielen auf die Gewährung eines hohen Sicherheitsniveaus des *RSA*-Verfahrens.

- Die Primzahlen *p* und *q* sollten „zufällig" gewählt und nicht etwa einer Primzahltabelle entnommen werden und auch keine spezielle funktionale Form (etwa $2^{2^k} - 1$) aufweisen.

- Die Primzahlen *p* und *q* sollten nicht zu dicht zusammenliegen.

- Die Primzahlen *p* und *q* sollten so gewählt werden, dass $p-1$ und $q-1$ keine großen gemeinsamen Faktoren besitzen.

- Die Primzahlen *p* und *q* sollten so gewählt werden, dass $\varphi(n)=(p-1)\cdot(q-1)$ nicht nur kleine Primfaktoren enthält.

- Der Wert *d* sollte nicht zu klein sein, damit man ihn nicht durch systematisches Testen ermitteln kann.

- Verschiedene Kommunikationspartner sollten nicht denselben Wert oder einen zu kleinen Wert für *e* nehmen.

- Die Klartexte (hier als numerische Werte aufgefasst) $x=1$ und $x=n-1$ werden auf sich selbst verschlüsselt, d.h. in diesen Fällen gilt $E(x,c_B)=x$. Dasselbe Fixpunktverhalten der Funktion *E* zeigt sich, wenn $e-1$ ein gemeinsames Vielfaches von $p-1$ und $q-1$ ist, etwa $e-1=\varphi(n)/2$. Dann gilt sogar für jeden Klartext *x* die Gleichung $E(x,c_B)=x$. In diesem Fall ist eine andere Wahl von *d* angeraten.

Da beim *RSA*-Verfahren alle Komponenten bis auf den geheimen Schlüssel $d_B=[d,p,q,\varphi(n)]$ öffentlich sind, bietet es für einen Kryptoanalytiker Angriffspunkte. Ein Kryptoanalytiker ist prinzipiell nur an der Kenntnis des Schlüsselteils *d* des geheimen Schlüssels d_B interessiert, wobei er beide Teile *e* und *n* des öffentlichen Schlüssels c_B kennt. Folgende Überlegungen zeigen, dass es erforderlich ist, neben *d* auch die Werte *p*, *q* und $\varphi(n)$ geheimzuhalten.

Kennt der Kryptoanalytiker die Werte *e, n* (aus dem öffentlichen Register) und $\varphi(n)$, dann kann er mit Hilfe des Euklidischen Algorithmus zwei Zahlen *a* und *b* berechnen, für die die Beziehung $e\cdot a+\varphi(n)\cdot b=1$ gilt (hierbei ist zu beachten, dass nach Konstruktion des Verfahrens *e* und $\varphi(n)$ teilerfremd sind). Damit folgt nacheinander $e\cdot a+\varphi(n)\cdot b=e\cdot d+f\cdot\varphi(n)$, $e\cdot(a-d)=(f-b)\cdot\varphi(n)$ und $a\equiv d \pmod{\varphi(n)}$. Wegen $0<d<\varphi(n)$ ist $d=(a \bmod \varphi(n))$.

Kennt der Kryptoanalytiker die Werte *e, n* (aus dem öffentlichen Register) und mindestens einen der Werte *p* oder *q*, etwa *p*, dann kann er wegen $\varphi(n)=(p-1)\cdot(q-1)$ und $n=p\cdot q$ bzw. $\varphi(n)=(p-1)\cdot(n/(p-1))$ sofort $\varphi(n)$ und damit *d* ermitteln.

Offensichtlich ist die Geheimhaltung von $\varphi(n)$ wesentlich. Natürlich könnte der Kryptoanalytiker versuchen, den Wert $\varphi(n)$ direkt aus dem öffentlichen Schlüssel $c_B = [e,\ n]$ zu gewinnen. Falls ihm dieses mit geringem Rechenaufwand gelänge, hätte er gleichzeitig einen schnellen Algorithmus, um die Zahl n in ihre Primfaktoren p und q zu zerlegen: Er berechnet nacheinander $z = \varphi(n) - n - 1$, $y = \sqrt{z^2 - 4 \cdot n}$, $q = 1/2 \cdot (-z - y)$ und $p = n/q$. Daher ist die schnelle Berechnung von $\varphi(n)$ aus $c_B = [e,\ n]$ gleichbedeutend mit der schnellen Primfaktorisierung von n. Andererseits kennt man bis heute kein schnelles Verfahren, um n zu faktorisieren. Die schnellsten bisher bekannten Verfahren zur Zerlegung einer Zahl n in ihre Primfaktoren haben eine Laufzeit, die proportional zu $L(n) = e^{\sqrt{\ln(n) \cdot \ln(\ln(n))}}$ ist. Die folgende Tabelle zeigt einige Werte von $L(n)$.

n	10^{50}	10^{100}	10^{150}	10^{200}	10^{250}	10^{300}
$L(n)$	$1{,}42 \cdot 10^{10}$	$2{,}34 \cdot 10^{15}$	$3{,}26 \cdot 10^{19}$	$1{,}20 \cdot 10^{23}$	$1{,}86 \cdot 10^{26}$	$1{,}53 \cdot 10^{29}$

Wäre man heute technisch in der Lage, Rechengeschwindigkeit von 10^{12} Operationen pro Sekunde zu realisieren, würde die Faktorisierung einer 200-stellige Zahl immer noch etwa 1.000 Jahre erfordern, die Faktorisierung einer 300-stelligen Zahl sogar mehr als 10^6 viele Jahrtausende. Heutige Schlüssellängen von 1.024 Bits bzw. ca. 300 Dezimalstellen erscheinen daher heute sicher.

Des weiteren könnte der Kryptoanalytiker versuchen, den Wert d direkt aus dem öffentlichen Schlüssel $c_B = [e,\ n]$ zu ermitteln. Es lässt sich zeigen, dass ein schneller Algorithmus zur Ermittlung von d aus $c_B = [e,\ n]$ in einen schnellen (probabilistischen) Algorithmus umgewandelt werden kann, der mit beliebig großer Wahrscheinlichkeit die Zahl n in ihre Primfaktoren p und q korrekt zerlegt. Eine Brute-Force-Attacke, in der alle möglichen Werte für d systematisch probiert werden, verspricht darüber hinaus wegen der großen Schlüssellänge (Stellenzahl von n) keinen Erfolg.

Zusammenfassend kann man feststellen, dass die Garantie der Sicherheit des *RSA*-Verfahrens darauf zurückzuführen ist, dass kein schnelles Verfahren bekannt ist, das eine gegebene natürliche Zahl in ihre Primfaktoren zerlegt. Sollte ein derartiges Verfahren für das Faktorisierungsproblem gefunden werden, ist das *RSA*-Verfahren nicht mehr sicher.

Die bisher in diesem Kapitel beschriebenen Methoden beruhen auf der Anwendung zahlentheoretischer Erkenntnisse, die im wesentlichen im 18. Jahrhundert entdeckt wurden. Die Überlegungen zum Laufzeitverhalten der beteiligten Algorithmen stammen aus den letzten 30 Jahren des 20. Jahrhunderts. Seit etwa 1987 findet eine Theorie, deren Grundlagen zum Ende des 19. Jahrhunderts gelegt wurden, beim Entwurf kryptographischer Verfahren verstärkt

Anwendung. Diese Kryptographie-Verfahren setzen zur Verschlüsselung die **Arithmetik elliptischer Kurven über endlichen Körpern** ein. Da zum Verständnis dieser Methoden jedoch weitergehende mathematische Kenntnisse erforderlich sind, wird auf deren Darstellung hier verzichtet.

4 Ausgewählte Themen der Kombinatorik

Die Kombinatorik befasst sich im wesentlichen mit dem Abzählen endlicher Mengen und damit verwandter Fragestellungen. Die in diesem Kapitel behandelte Themenauswahl gehört zum mathematischen Handwerkszeug, das in vielen Teilgebieten der Mathematik und Informatik benötigt wird. Insbesondere in der Wahrscheinlichkeitsrechnung (diskrete Wahrscheinlichkeitsverteilungen) werden die Themen weiter vertieft.

4.1 Binomialkoeffizienten

Es seien $n \in \mathbf{N}$, $x \in \mathbf{R}$ und $y \in \mathbf{R}$. Der aus der Schule bekannte binomische Lehrsatz besagt

$$(x+y)^2 = x^2 + 2 \cdot x \cdot y + y^2 .$$

Wie man leicht nachrechnet, ist

$$(x+y)^3 = x^3 + 3 \cdot x^2 \cdot y + 3 \cdot x \cdot y^2 + y^3 ,$$
$$(x+y)^4 = x^4 + 4 \cdot x^3 \cdot y + 6 \cdot x^2 \cdot y^2 + 4 \cdot x \cdot y^3 + y^4 .$$

Es soll nun die allgemeine Form von $(x+y)^n$ als ausgeschriebene Summe hergeleitet werden (das könnte wieder formal nach dem Induktionsprinzip geschehen; hier soll die Herleitung etwas informeller beschrieben werden). Durch vollständige Induktion kann man zeigen, dass die Summanden in der ausgeschriebenen Summe von $(x+y)^n$ die Form $k_{i,j} \cdot x^i \cdot y^j$ mit $i + j = n$ haben. Der Faktor $k_{i,j}$ im Summanden $k_{i,j} \cdot x^i \cdot y^{n-i}$ der ausgeschriebenen Summe von $(x+y)^n$ heißt **Binomialkoeffizient** $\binom{n}{i}$, gesprochen „*n* über *i*“, d.h.

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} \cdot x^i \cdot y^{n-i} = \binom{n}{0} \cdot y^n + \binom{n}{1} \cdot x \cdot y^{n-1} + \binom{n}{2} \cdot x^2 \cdot y^{n-2} + \ldots + \binom{n}{n} \cdot x^n .$$

Das vorliegende Kapitel untersucht Eigenschaften und Interpretationen der Binomialkoeffizienten.

Offensichtlich gilt $\binom{n}{0} = 1$ und $\binom{n}{n} = 1$.

Aus

$$\begin{aligned}(x+y)^n &= \sum_{i=0}^{n} \binom{n}{i} \cdot x^i \cdot y^{n-i} = \binom{n}{0} \cdot y^n + \binom{n}{1} \cdot x \cdot y^{n-1} + \binom{n}{2} \cdot x^2 \cdot y^{n-2} + \ldots + \binom{n}{n} \cdot x^n \\ &= (y+x)^n \\ &= \sum_{i=0}^{n} \binom{n}{i} \cdot y^i \cdot x^{n-i} = \binom{n}{0} \cdot x^n + \binom{n}{1} \cdot y \cdot x^{n-1} + \binom{n}{2} \cdot y^2 \cdot x^{n-2} + \ldots + \binom{n}{n} \cdot y^n\end{aligned}$$

folgt durch Koeffizientenvergleich die Symmetrie der Binomialkoeffizienten:

$$\binom{n}{i} = \binom{n}{n-i} \text{ für } i \in \mathbf{N} \text{ mit } 0 \le i \le n.$$

Es ist $(x+y)^n = (x+y)^{n-1} \cdot (x+y) = (x+y)^{n-1} \cdot x + (x+y)^{n-1} \cdot y$. Aus dieser Darstellung kann man ablesen, wie der Faktor $\binom{n}{i}$ im Summanden $\binom{n}{i} \cdot x^i \cdot y^{n-i}$ der ausgeschriebenen Summe von $(x+y)^n$ entsteht: man sucht in der ausgeschriebenen Summe von $(x+y)^{n-1}$ denjenigen Summanden, in dem x mit der Potenz $i-1$ und y mit der Potenz $n-i$ steht, und denjenigen Summanden in der ausgeschriebenen Summe von $(x+y)^{n-1}$, in dem x mit der Potenz i und y mit der Potenz $n-i-1$ steht. Diese Summanden sind

$$\binom{n-1}{i-1} \cdot x^{i-1} \cdot y^{n-i} = \binom{n-1}{i-1} \cdot x^{i-1} \cdot y^{(n-1)-(i-1)} \quad \text{und} \quad \binom{n-1}{i} \cdot x^i \cdot y^{n-i-1} = \binom{n-1}{i} \cdot x^i \cdot y^{(n-1)-i}.$$

Wird der erste Summand mit x und der zweite Summand mit y multipliziert und anschließend beide Summanden addiert, entsteht

$$\begin{aligned}\binom{n-1}{i-1} \cdot x^{i-1} \cdot y^{n-i} \cdot x + \binom{n-1}{i} \cdot x^i \cdot y^{n-i-1} \cdot y &= \left(\binom{n-1}{i-1} + \binom{n-1}{i}\right) \cdot x^i \cdot y^{n-i} \\ &= \binom{n}{i} \cdot x^i \cdot y^{n-i}.\end{aligned}$$

Damit ergibt sich

Satz 4.1-1:

Für jedes $n \in \mathbf{N}$ und für jedes $i \in \mathbf{N}$ ist

$$\binom{n}{0} = 1, \binom{n}{n} = 1,$$

$$\binom{n}{i} = \binom{n-1}{i-1} + \binom{n-1}{i} \text{ mit } 0 < i < n.$$

Mit Hilfe der Rekursionsformel in Satz 4.1-1 lassen sich die Binomialkoeffizienten berechnen. Die einzelnen Werte können in einem Schema in Dreicksform (**Pascal'sche Dreieck**) angeordnet werden. Dabei steht in der *n*-ten Zeile und der *i*-ten Spalte der Wert $\binom{n}{i}$ für $n \geq 0$ und $0 \leq i \leq n$. Dieser Eintrag ist die Summe, die sich aus dem direkt drüber stehenden Eintrag $\binom{n-1}{i}$ und dem Eintrag $\binom{n-1}{i-1}$ links davon ergibt. Der Anfang des Pascal'schen Dreiecks lautet:

	Spalte 0					Spalte *i* = 5				
Zeile 0	1									
	1	1								
	1	2	1							
	1	3	3	1						
	1	4	6	4	1					
	1	5	10	10	5	1				
	1	6	15	20	15	6	1			
	1	7	21	35	35	21	7	1		
Zeile *n* = 8	1	8	28	56	70	56	28	8	1	
	1	9	36	84	126	126	84	36	9	1
					...	...				

Der folgende Satz beschreibt, wie der Wert eines Binomialkoeffizienten $\binom{n}{i}$ direkt in Abhängigkeit von *n* und *i* ausgedrückt werden kann:

Satz 4.1-2:

Für jedes $n \in \mathbf{N}$ und für jedes $i \in \mathbf{N}$ mit $0 \le i \le n$ ist

$$\binom{n}{i} = \frac{n!}{i! \cdot (n-i)!}.$$

Der Beweis soll hier als Beispiel eines Beweises durch vollständige Induktion angegeben werden:

Für $n = 0$ ist $\binom{n}{i} = \binom{0}{0} = 1$ und $\frac{n!}{i! \cdot (n-i)!} = \frac{0!}{0! \cdot 0!} = 1$.

Es wird angenommen, dass die Formel in Satz 4.1-2 für ein $n \in \mathbf{N}$ und für jedes $i \in \mathbf{N}$ mit $0 \le i \le n$ gilt. Zu zeigen ist, dass aus dieser Annahme die Gültigkeit der Formel auch für $n+1$ und für jedes $i \in \mathbf{N}$ mit $0 \le i \le n+1$ folgt.

Für $i = 0$ ist $\binom{n+1}{i} = \binom{n+1}{0} = 1$ und $\frac{(n+1)!}{i! \cdot (n+1-i)!} = \frac{(n+1)!}{0! \cdot (n+1)!} = 1$.

Für $i = n+1$ ist $\binom{n+1}{i} = \binom{n+1}{n+1} = 1$ und $\frac{(n+1)!}{(n+1)! \cdot 0!} = 1$.

Für $0 < i < n+1$ verwendet man die Rekursionsformel aus Satz 4.1-1:

$$\binom{n+1}{i} = \binom{n}{i-1} + \binom{n}{i} \qquad \text{(nach Satz 4.1-1)}$$

$$= \frac{n!}{(i-1)! \cdot (n-i+1)!} + \frac{n!}{i! \cdot (n-i)!} \qquad \text{(nach Induktionsannahme)}$$

$$= \frac{n! \cdot i + n! \cdot (n-i+1)}{i! \cdot (n-i+1)!}$$

$$= \frac{n! \cdot (i+n-i+1)}{i! \cdot (n-i+1)!}$$

$$= \frac{(n+1)!}{i! \cdot (n+1-i)!}.$$

Die Formel gilt also auch für $n+1$.

Im Pascal'schen Dreieck kann man einige Gesetzmäßigkeiten der Binomialkoeffizienten verifizieren, die direkt aus der Definition $(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} \cdot x^i \cdot y^{n-i}$ bzw. aus den Formeln in den

vorherigen Sätzen folgen. Beispielsweise hat die Summe aller Binomialkoeffizienten in der *n*-ten Zeile des Pascal'schen Dreiecks den Wert 2^n; die Summe der Binomialkoeffizienten $\binom{n}{i}$ in Zeile *n* mit $n \geq 1$ und geradem *i* ist gleich der Summe der Binomialkoeffizienten $\binom{n}{i}$ in Zeile *n* mit ungeradem *i*; summiert man in der *i*-ten Spalte alle Binomialkoeffizienten bis zur Zeile *n*, so erhält man wieder einen Binomialkoeffizienten, nämlich $\binom{n+1}{i+1}$; summiert man alle Binomialkoeffizienten ab Zeile $n-i$ und Spalte 0 diagonal (von links oben nach rechts unten) bis zur Zeile *n* und Spalte *i*, so ist das Ergebnis der Binomialkoeffizient $\binom{n+1}{i}$. Diese und weitere Eigenschaften der Binomialkoeffizienten werden im folgenden Satz zusammengefasst.

Satz 4.1-3:

Es sei $n \in \mathbf{N}$. Dann gilt:

(i) $\sum_{i=0}^{n} \binom{n}{i} \cdot x^i = (1+x)^n$ für jedes $x \in \mathbf{R}$.

(ii) $\sum_{i=0}^{n} \binom{n}{i} = 2^n$ (Summe über die *n*-te Zeile im Pascal'schen Dreieck).

(iii) $\sum_{\substack{i=0 \\ (i \bmod 2)=0}}^{n} \binom{n}{i} = \sum_{\substack{i=0 \\ (i \bmod 2)=1}}^{n} \binom{n}{i}$ bzw. $\sum_{i=0}^{n} (-1)^i \binom{n}{i} = 0$ für $n \geq 1$

(Die Summe der Binomialkoeffizienten mit geradem *i* ist gleich der Summe der Binomialkoeffizienten mit ungeradem *i*).

(iv) $\sum_{k=i}^{n} \binom{k}{i} = \binom{n+1}{i+1}$ für $i \in \mathbf{N}$ mit $0 \leq i \leq n$

(Summe der *i*-ten Spalte bis zur Zeile *n* im Pascal'schen Dreieck).

../..

(v) $\sum_{k=0}^{i}\binom{n-i+k}{k}=\binom{n+1}{i}$ für $i \in \mathbf{N}$ mit $0 \le i \le n$

(Summe der Binomialkoeffizienten ab Zeile $n-i$ und Spalte 0 im Pascal'schen Dreieck diagonal abwärts bis zur Zeile *n* und Spalte *i*).

(vi) $\binom{n}{i}=\binom{n}{n-i}$ für $i \in \mathbf{N}$ mit $0 \le i \le n$

(Symmetrie der Binomialkoeffizienten).

(vii) $\binom{n}{i}=\frac{n}{i}\cdot\binom{n-1}{i-1}$ für $i \in \mathbf{N}$ mit $0 < i \le n$

und $(n-i)\cdot\binom{n}{i}=n\cdot\binom{n-1}{i}$ für $i \in \mathbf{N}$ mit $0 \le i \le n$.

(viii) Durch $F_0=0, F_1=1$ und $F_n=F_{n-1}+F_{n-2}$ wird die Folge der Fibonacci-Zahlen definiert (siehe Kapitel 5.10); die ersten dreizehn Fibonacci-Zahlen lauten
0, 1, 1, 2, 3, 5, 8,13, 21, 34, 55, 89, 144, …
$\sum_{k=0}^{\lfloor n/2 \rfloor}\binom{n-k}{k}=F_{n+1}$

(Summe der Binomialkoeffizienten ab Zeile n und Spalte 0 im Pascal'schen Dreieck diagonal aufwärts bis zur Diagonalen).

Die Formel in Satz 4.1-3 (i) erhält man, indem man in der Definitionsgleichung $(x+y)^n=\sum_{i=0}^{n}\binom{n}{i}\cdot x^i\cdot y^{n-i}$ den Wert $y=1$ setzt. Die Formel in Satz 4.1-3 (ii) erhält man aus der Definitionsgleichung für $x=y=1$. Die Formel $\sum_{i=0}^{n}(-1)^i\binom{n}{i}=0$ in Satz 4.1-3 (iii) erhält man aus der Definitionsgleichung für $x=-1$ und $y=1$; bringt man in dieser Formel die Summanden $(-1)^i\binom{n}{i}$ mit ungeradem *i* auf die rechte Seite der Gleichung, so ergibt sich die erste Formel in Satz 4.1-3 (iii).

Die Formeln in Satz 4.1-3 (vi) und (vii) ergeben sich unmittelbar aus Satz 4.1-2.

Die Formeln in Satz 4.1-3 (iv) und (v) können durch vollständige Induktion bewiesen werden oder durch direkte wiederholte Anwendung der Rekursionsgleichung aus Satz 4.1-1. Für die Formel aus Satz 4.1-3 (iv) ergibt sich ausgehend von der rechten Seite der Gleichung:

$$\binom{n+1}{i+1} = \binom{n}{i+1} + \binom{n}{i} \qquad \text{(mit Satz 4.1-1)}$$

$$= \left(\binom{n-1}{i+1} + \binom{n-1}{i}\right) + \binom{n}{i} \quad \text{(mit Satz 4.1-1, angewandt auf den ersten Binomialkoeffizienten)}$$

$$= \left(\binom{n-2}{i+1} + \binom{n-2}{i}\right) + \binom{n-1}{i} + \binom{n}{i}$$

$$= \binom{n-l}{i+1} + \sum_{k=0}^{l} \binom{n-k}{i} \qquad \text{(allgemeine Form; mit } l = n-i-1\text{:)}$$

$$= \binom{i+1}{i+1} + \sum_{k=0}^{n-i-1} \binom{n-k}{i}$$

$$= 1 + \binom{n}{i} + \binom{n-1}{i} + \ldots + \binom{i+1}{i}$$

$$= \binom{i}{i} + \binom{n}{i} + \binom{n-1}{i} + \ldots + \binom{i+1}{i}$$

$$= \sum_{k=i}^{n} \binom{k}{i} \quad .$$

Für die Formel aus Satz 4.1-3 (v) ergibt sich wieder ausgehend von der rechten Seite der Gleichung:

$$\binom{n+1}{i} = \binom{n}{i} + \binom{n}{i-1} \qquad \text{(mit Satz 4.1-1)}$$

$$= \binom{n}{i} + \left(\binom{n-1}{i-1} + \binom{n-1}{i-2}\right) \quad \text{(mit Satz 4.1-1, angewandt auf den zweiten Binomialkoeffizienten)}$$

$$= \binom{n}{i} + \binom{n-1}{i-1} + \left(\binom{n-2}{i-2} + \binom{n-2}{i-3}\right)$$

$$=$$

$$\ldots$$

$$= \sum_{k=0}^{l} \binom{n-k}{i-k} + \binom{n-l}{i-l-1} \qquad \text{(allgemeine Form; mit } l = i-1\text{:)}$$

$$= \sum_{k=0}^{i-1} \binom{n-k}{i-k} + \binom{n-i+1}{0}$$

$$= 1 + \binom{n}{i} + \binom{n-1}{i-1} + \ldots + \binom{n-i+1}{1}$$

$$= \binom{n-i}{0} + \binom{n}{i} + \binom{n-1}{i-1} + \ldots + \binom{n-i+1}{1} = \sum_{k=0}^{i} \binom{n-i+k}{k}$$

Die Formel in 4.1-3 (viii) lässt sich durch vollständige Induktion zeigen:
Für $n=0$ ist $\sum_{k=0}^{0}\binom{0-k}{k}=\binom{0}{0}=1=F_1$; für $n=1$ ist $\sum_{k=0}^{0}\binom{1-k}{k}=\binom{1}{0}=1=F_2$; für $n=2$ ist $\sum_{k=0}^{1}\binom{2-k}{k}=\binom{2}{0}+\binom{1}{1}=1+1=2=F_3$; für $n=3$ ist $\sum_{k=0}^{1}\binom{3-k}{k}=\binom{3}{0}+\binom{2}{1}=1+2=3=F_4$.
Die Aussage gelte für $n\geq 3$. Für $n+1$ sieht man die Gültigkeit der Formel dann wie folgt:

$$\begin{aligned}\sum_{k=0}^{\lfloor (n+1)/2\rfloor}\binom{n+1-k}{k} &= 1+\sum_{k=1}^{\lfloor (n+1)/2\rfloor}\binom{n+1-k}{k}\\ &= 1+\sum_{k=1}^{\lfloor (n+1)/2\rfloor}\binom{n-k}{k}+\sum_{k=1}^{\lfloor (n+1)/2\rfloor}\binom{n-k}{k-1}\\ &= \sum_{k=0}^{\lfloor (n+1)/2\rfloor}\binom{n-k}{k}+\sum_{k=0}^{\lfloor (n+1)/2\rfloor-1}\binom{n-k-1}{k}\ .\end{aligned}$$

Ist *n* gerade, etwa $n=2\cdot m$, dann ist $\lfloor (n+1)/2\rfloor=\lfloor (2\cdot m+1)/2\rfloor=m=n/2=\lfloor n/2\rfloor$ und $\lfloor (n+1)/2\rfloor-1=m-1=\lfloor (2\cdot m-1)/2\rfloor=\lfloor (n-1)/2\rfloor$; daher ergibt sich für die Summe mit der Induktionsannahme

$$\sum_{k=0}^{\lfloor (n+1)/2\rfloor}\binom{n-k}{k}+\sum_{k=0}^{\lfloor (n+1)/2\rfloor-1}\binom{n-k-1}{k}=\sum_{k=0}^{\lfloor n/2\rfloor}\binom{n-k}{k}+\sum_{k=0}^{\lfloor (n-1)/2\rfloor}\binom{n-1-k}{k}=F_{n+1}+F_n=F_{n+2}\ .$$

Ist *n* ungerade, etwa $n=2\cdot m+1$, dann ist in der ersten Summe der letzte aufzusummierende Binomialkoeffizient (bei $k=\lfloor (n+1)/2\rfloor=\lfloor (2\cdot m+2)/2\rfloor=m+1$) formal gleich $\binom{m}{m+1}$; daher läuft *k* in dieser Summe nur bis $m=\lfloor (2\cdot m+1)/2\rfloor=\lfloor n/2\rfloor$; außerdem ist $\lfloor (n+1)/2\rfloor-1=m=\lfloor (2\cdot m)/2\rfloor=\lfloor (n-1)/2\rfloor$; insgesamt ergibt sich mit der Induktionsannahme auch hier

$$\sum_{k=0}^{\lfloor (n+1)/2\rfloor}\binom{n-k}{k}+\sum_{k=0}^{\lfloor (n+1)/2\rfloor-1}\binom{n-k-1}{k}=\sum_{k=0}^{\lfloor n/2\rfloor}\binom{n-k}{k}+\sum_{k=0}^{\lfloor (n-1)/2\rfloor}\binom{n-1-k}{k}=F_{n+1}+F_n=F_{n+2}\ .$$

Die Binomialkoeffizienten kommen nicht nur als Faktor $k_{i,j}$ im Summanden $k_{i,j}\cdot x^i\cdot y^{n-i}$ der ausgeschriebenen Summe von $(x+y)^n$ vor, sondern auch in vielen praktischen Abzählproblemen. In Kapitel 1.5 wird gezeigt, dass eine endliche Menge *A* mit *n* Elementen genau 2^n viele Teilmengen besitzt. Es soll nun untersucht werden, wie viele Teilmengen *A* hat, die aus genau *k* Elementen (mit $0\leq k\leq n$) bestehen.

Es sei $A = \{a_1, ..., a_n\}$. Unter einer **Permutation** von A versteht man eine feste Anordnung der Elemente von A. Beispielsweise sind alle Permutationen der Menge $A = \{1,2,3,4\}$ die Anordnungen

1 2 3 4	4 1 3 2	3 1 2 4	4 3 2 1	2 3 1 4	4 2 1 3
1 2 4 3	1 4 3 2	3 1 4 2	3 4 2 1	2 3 4 1	2 4 1 3
1 4 2 3	1 3 4 2	3 4 1 2	3 2 4 1	2 4 3 1	2 1 4 3
4 1 2 3	1 3 2 4	4 3 1 2	3 2 1 4	4 2 3 1	2 1 3 4.

Eine Permutation der Menge $A = \{a_1, ..., a_n\}$ ist also ein Tupel $(a_{i_1}, ..., a_{i_n})$ mit $a_{i_j} \in A$ für $j = 1, ..., n$ und $a_{i_j} \neq a_{i_l}$ für $i_j \neq i_l$. Um die Anzahl aller Permutationen von A zu bestimmen, betrachtet man alle derartigen Tupel $(a_{i_1}, ..., a_{i_n})$: Für a_{i_1} gibt es n mögliche Werte, nämlich $a_1, ..., a_n$. Hat man sich für eine Möglichkeit entschieden, bleiben für a_{i_2} noch $n-1$ Möglichkeiten. Für a_{i_3} bleiben nach Auswahl für a_{i_1} und a_{i_2} noch $n-2$ Möglichkeiten usw. Für a_{i_n} bleibt nach Festlegung der ersten $n-1$ Elemente nur noch 1 Möglichkeit. Insgesamt hat gibt es also $n \cdot (n-1) \cdot ... \cdot 1 = n!$ viele Möglichkeiten zur Bildung einer Permutation einer n-elementigen Menge.

Unter einer ***k*-Permutation** von A versteht man ein Tupel $(a_{i_1}, ..., a_{i_k})$ mit $a_{i_j} \in A$ für $j = 1, ..., k$ und $a_{i_j} \neq a_{i_l}$ für $i_j \neq i_l$. Beispielsweise sind alle 2-Permutationen von $A = \{1,2,3,4\}$ die Anordnungen

1 2	2 1	3 1	4 1
1 3	2 3	3 2	4 2
1 4	2 4	3 4	4 3

Um die Anzahl aller k-Permutationen von A zu bestimmen, betrachtet man wieder alle derartigen Tupel $(a_{i_1}, ..., a_{i_k})$: Für a_{i_1} gibt es n mögliche Werte, nämlich $a_1, ..., a_n$. Hat man sich für eine Möglichkeit entschieden, bleiben für a_{i_2} noch $n-1$ Möglichkeiten. Für a_{i_3} bleiben nach Auswahl für a_{i_1} und a_{i_2} noch $n-2$ Möglichkeiten usw. Für a_{i_k} bleibt nach Festlegung der ersten $k-1$ Elemente noch $n-(k-1) = n-k+1$ Möglichkeit. Insgesamt hat gibt es also

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

viele Möglichkeiten zur Bildung einer *k*-Permutation einer *n*-elementigen Menge.

Für jede Teilmengen $B \subseteq A$ mit $|B| = k$, etwa $B = \{a_{i_1}, ..., a_{i_k}\}$, gilt $a_{i_j} \neq a_{i_l}$ für $i_j \neq i_l$. Jede Permutation von *B* (davon gibt es *k*! viele) ist eine *k*-Permutation von *A*. Verschiedene Teilmengen $B_1 \subseteq A$ und $B_2 \subseteq A$ mit $|B_1| = |B_2| = k$ ergeben paarweise verschiedene *k*-Permutation von *A*, da die *k*-Permutationen, die aus B_1 entstanden sind, mindestens ein unterschiedliches Element zu den *k*-Permutationen enthalten, die aus B_2 entstanden sind, und die Permutationen von B_1 bzw. von B_2 sind untereinander paarweise verschieden. Umgekehrt gibt es zu jeder *k*-Permutation von *A* eine *k*-elementige Teilmengen von *A*, nämlich die Menge ihrer Elemente. Bezeichnet $C_{n,k}$ die Anzahl *k*-elementiger Teilmengen von *A*, so gilt daher:

$$C_{n,k} \cdot k! = \frac{n!}{(n-k)!}.$$

Satz 4.1-4:

Es sei *A* eine endliche Menge mit $|A| = n$. Die Anzahl der Teilmengen $B \subseteq A$ mit $|B| = k$ für $0 \leq k \leq n$ ist

$$\binom{n}{k} = \frac{n!}{k!\cdot(n-k)!}.$$

Damit kann man auch noch einmal die Formel aus Satz 4.1-3 (ii) $\sum_{i=0}^{n} \binom{n}{i} = 2^n$ verifizieren: Links steht die Anzahl aller Teilmengen einer *n*-elementigen Menge, aufgeteilt nach den *i*-elementigen Teilmengen für $0 \leq i \leq n$, von der bereits früher gezeigt wurde, dass sie gleich 2^n ist.

4.2 Abbildungen zwischen endlichen Mengen

In diesem Kapitel seien A und B endliche Mengen mit $|A| = n$ und $|B| = m$: $A = \{a_1, ..., a_n\}$, $B = \{b_1, ..., b_m\}$.

Es sollen die Anzahlen der Abbildungen, der injektiven, surjektiven und bijektiven Abbildungen $f : A \to B$ ermittelt werden.

A. Anzahl der Abbildungen $f : A \to B$

Jede Abbildung $f : A \to B$ kann in Form einer endlichen Tabelle notiert werden:

a_i	a_1	...	a_n
$f(a_i)$	$f(a_1)$	...	$f(a_n)$

Dabei genügt das Notieren der Funktionswerte in der Reihenfolge $f(a_1)$, ..., $f(a_n)$ bzw. als n-Tupel $(f(a_1), ..., f(a_n))$.

Verschiedene Abbildungen führen zu verschiedenen n-Tupeln. Umgekehrt kann man jedes n-Tupel $(b_1, ..., b_n)$ mit $b_i \in B$ für $1 \le i \le n$ als eine Abbildung $f : A \to B$ auffassen, nämlich als die durch $f(a_i) = b_i$ definierte Abbildung, und unterschiedliche n-Tupel beschreiben unterschiedliche Abbildungen. Daher ist die Anzahl der Abbildungen $f : A \to B$ gleich der Anzahl der n-Tupeln $(b_1, ..., b_n)$ mit $b_i \in B$ für $1 \le i \le n$.

Satz 4.2-1:

Es seien A und B endliche Mengen mit $|A| = n$ und $|B| = m$. Dann ist die Anzahl der Abbildungen $f : A \to B$ gleich m^n.

B. Anzahl bijektiver Abbildungen $f : A \to B$

Nach Satz 2.2-3 gilt im Falle bijektiver Abbildungen $f : A \to B$ bezüglich der Mächtigkeiten: $n = m$. Jede bijektive Abbildung $f : A \to B$ bzw. in Tupel-Schreibweise $(f(a_1), ..., f(a_n))$ beschreibt daher eine Permutation der Menge *B*. Verschiedene bijektive Abbildungen führen zu verschiedenen Permutationen. Umgekehrt kann jede Permutation $(b_{i_1}, ..., b_{i_n})$ von *B* mit einer bijektiven Abbildung $f : A \to B$, nämlich mit $f : \begin{cases} A & \to & B \\ a_j & \to & b_{i_j} \end{cases}$ gleichgesetzt werden.

Die Anzahl bijektiver Abbildungen $f : A \to B$ ist daher gleich der Anzahl der Permutationen der Menge *B*. Mit den Überlegungen am Ende von Kapitel 4.1 folgt

Satz 4.2-2:

Es seien *A* und *B* endliche Mengen mit $|A| = |B| = n$. Dann ist die Anzahl der bijektiven Abbildungen $f : A \to B$ gleich *n*!.

C. Anzahl injektiver Abbildungen $f : A \to B$

Nach Satz 2.2-3 gilt im Falle injektiver Abbildungen $f : A \to B$ bezüglich der Mächtigkeiten: $n \le m$. Der Wertebereich $W(f) = \{ b \mid b \in B, \text{und es gibt } a \in A \text{ mit } f(a) = b \}$ ist eine *n*-elementige Teilmenge von *B*, und die Abbildung $g : \begin{cases} A & \to & W(f) \\ a & \to & f(a) \end{cases}$ ist eine Bijektion zwischen *A* und W(*f*). Daher ist W(*f*) eine *n*-Permutation von *B*. Umgekehrt ist kann man jede *n*-Permutation von *B*, etwa $(b_{i_1}, ..., b_{i_n})$ mit der injektiven Abbildung $f : \begin{cases} A & \to & B \\ a_j & \to & b_{i_j} \end{cases}$ identifizieren. Mit den Überlegungen am Ende von Kapitel 4.1 folgt

Satz 4.2-3:

Es seien *A* und *B* endliche Mengen mit $|A| = n$ und $|B| = m$ und $n \le m$. Dann ist die Anzahl injektiver Abbildungen $f : A \to B$ gleich $\frac{m!}{(m-n)!} = n! \cdot \binom{m}{n}$.

D. Anzahl surjektiver Abbildungen $f : A \to B$

Nach Satz 2.2-3 gilt im Falle surjektiver Abbildungen $f : A \to B$ bezüglich der Mächtigkeiten: $n \geq m$. Die Bestimmung der Anzahl surjektiver Abbildungen zwischen A und B ist schwieriger und benötigt Hilfsmittel, die in Kapitel 4.3 bereitgestellt werden. Das Ergebnis soll aber der Vollständigkeit halber hier bereits zitiert werden.

Satz 4.2-4:

Es seien A und B endliche Mengen mit $|A| = n$ und $|B| = m$ und $n \geq m$. Dann ist die Anzahl surjektiver Abbildungen $f : A \to B$ gleich $\sum_{i=0}^{m} (-1)^{m-i} \cdot \binom{m}{i} \cdot i^n$.

4.3 Das Prinzip von Inklusion und Exklusion

Es seien A und B wieder endliche Mengen mit $|A| = n$ und $|B| = m$. Nach Satz 1.1-1 (v) ist $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$ eine disjunkte Zerlegung von $A \cup B$. Daraus folgt für die Mächtigkeit von $A \cup B$: $|A \cup B| = |A \setminus B| + |A \cap B| + |B \setminus A|$.

Es sei $|A \cap B| = i$. Dann ist

$$\begin{aligned} |A \cup B| &= |A \setminus B| + |A \cap B| + |B \setminus A| \\ &= (n - i) + i + (m - i) \\ &= n + m - i \\ &= |A| + |B| - |A \cap B|. \end{aligned}$$

Bei drei Mengen A, B und C mit $|A| = n$, $|B| = m$ und $|C| = k$ lauten die entsprechenden Formeln:

Die Menge $A \cup B \cup C$ lässt sich disjunkt zerlegen in

$$\begin{aligned} A \cup B \cup C = {} & ((A \setminus B) \setminus C) \cup ((B \setminus A) \setminus C) \cup ((C \setminus A) \setminus B) \\ & \cup ((A \cap B) \setminus C) \\ & \cup ((B \cap C) \setminus A) \\ & \cup ((A \cap C) \setminus B) \\ & \cup (A \cap B \cap C). \end{aligned}$$

Mit $i_1 = |(A \cap B) \setminus C|$, $i_2 = |(B \cap C) \setminus A|$, $i_3 = |(A \cap C) \setminus B|$ und $i_4 = |A \cap B \cap C|$ folgt daraus:

$$\begin{aligned}|A \cup B \cup C| &= n - (i_1 + i_3 + i_4) + m - (i_1 + i_2 + i_4) + k - (i_2 + i_3 + i_4) \\ &\quad + i_1 + i_2 + i_3 + i_4 \\ &= n + m + k - (i_1 + i_4) - (i_2 + i_4) - (i_3 + i_4) + i_4 \\ &= |A| + |B| + |C| - |A \cap B| - |B \cap C| - |A \cap C| + |A \cap B \cap C|.\end{aligned}$$

Zur Berechnung von $|A \cup B \cup C|$ wäre das Ergebnis $|A| + |B| + |C|$ korrekt, wenn die drei Mengen disjunkt wären (Inklusion). Andernfalls zählt man die Elemente, die in jeweils zwei Mengen liegen, doppelt, und man muss die Anzahl der zwei Mengen gemeinsamen Elemente wieder abziehen (Exklusion). Allerdings hat man dadurch die Elemente „vergessen“, die in allen drei Mengen liegen, so dass diese wieder hinzugezählt werden müssen (Inklusion). Dieses **Prinzip der Inklusion und Exklusion** lässt sich auf eine endliche Anzahl von Mengen erweitern.

Es seien A_1, ..., A_l Teilmengen einer endlichen Menge *M*. Mit $\bigcup_{i=1}^{l} A_i$ wird $A_1 \cup ... \cup A_l$ abgekürzt; mit $\bigcap_{i \in I} A_i$ mit $I \subseteq \{1, ..., l\}$ wird der Schnitt derjenigen Mengen A_i bezeichnet, für deren Index $i \in I$ gilt.

Satz 4.3-1:

Es seien A_1, ..., A_l Teilmengen einer endlichen Menge *M*, $l \geq 1$ Dann gilt:

$$\left| \bigcup_{i=1}^{l} A_i \right| = \sum_{\substack{I \subseteq \{1, ..., l\} \\ \text{und } I \neq \emptyset}} (-1)^{1+|I|} \cdot \left| \bigcap_{i \in I} A_i \right|.$$

Die Summation geht über alle nichtleeren Teilmengen *I* der Indexmenge $\{1, ..., l\}$.

Der Satz kann durch vollständige Induktion bewiesen werden. Zuvor soll er auf den obigen Fall dreier Mengen angewendet werden ($A_1 = A$, $A_2 = B$ und $A_3 = C$). Hierbei ist $\{1, ..., l\} = \{1, 2, 3\}$. Alle nichtleeren Teilmengen der Indexmenge sind $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$ und $\{1, 2, 3\}$.

Die einzelnen Summanden lauten

für $I = \{1\}$: $(-1)^{1+|I|} \cdot \left| \bigcap_{i \in I} A_i \right| = (-1)^{1+1} \cdot |A_1| = |A_1|$,

für $I = \{2\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i\in I} A_i\right| = (-1)^{1+1} \cdot |A_2| = |A_2|$,

für $I = \{3\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i\in I} A_i\right| = (-1)^{1+1} \cdot |A_3| = |A_3|$,

für $I = \{1,2\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i\in I} A_i\right| = (-1)^{1+2} \cdot |A_1 \cap A_2| = -|A_1 \cap A_2|$,

für $I = \{1,3\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i\in I} A_i\right| = (-1)^{1+2} \cdot |A_1 \cap A_3| = -|A_1 \cap A_3|$,

für $I = \{2,3\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i\in I} A_i\right| = (-1)^{1+2} \cdot |A_2 \cap A_3| = -|A_2 \cap A_3|$,

für $I = \{1,2,3\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i\in I} A_i\right| = (-1)^{1+3} \cdot |A_1 \cap A_2 \cap A_3| = |A_1 \cap A_2 \cap A_3|$.

Der Beweis der Formel in Satz 4.3-1 erfolgt durch vollständige Induktion über die Anzahl l der beteiligten Teilmengen:

Der Induktionsanfang für $l = 1$, $l = 2$ und $l = 3$ ist offensichtlich bzw. wurde in den obigen Beispielen gezeigt. Die Formel gelte für $l \geq 3$. Zu zeigen ist, dass aus dieser Annahme ihre Gültigkeit auch für $l+1$ folgt.

Es sei $A = \bigcup_{i=1}^{l} A_i$. Dann ist gemäß Induktionsanfang

$$\left|\bigcup_{i=1}^{l+1} A_i\right| = |A \cup A_{l+1}| = |A| + |A_{l+1}| - |A \cap A_{l+1}|.$$

Nach Induktionsannahme ist

$$|A| = \left|\bigcup_{i=1}^{l} A_i\right| = \sum_{\substack{I\subseteq\{1,\ldots,l\} \\ \text{und } I\neq\emptyset}} (-1)^{1+|I|} \cdot \left|\bigcap_{i\in I} A_i\right|.$$

Die in Satz 1.1-1 (vi) formulierten Distributivgesetze gelten auch für mehr als drei Mengen, so dass gilt:

$$A \cap A_{l+1} = \left(\bigcup_{i=1}^{l} A_i\right) \cap A_{l+1} = \bigcup_{i=1}^{l} (A_i \cap A_{l+1}).$$

Hierbei handelt es sich also um die Vereinigung von l Mengen der Form $A_i \cap A_{l+1}$, so dass die Induktionsannahme anwendbar ist:

$$|A \cap A_{l+1}| = \left|\bigcup_{i=1}^{l} (A_i \cap A_{l+1})\right| = \sum_{\substack{I\subseteq\{1,\ldots,l\} \\ \text{und } I\neq\emptyset}} (-1)^{1+|I|} \cdot \left|\bigcap_{i\in I} (A_i \cap A_{l+1})\right|.$$

Insgesamt ergibt sich

$$\left|\bigcup_{i=1}^{l+1} A_i\right| = \sum_{\substack{I\subseteq\{1,...,l\}\\ \text{und } I\neq\varnothing}} (-1)^{1+|I|}\cdot\left|\bigcap_{i\in I} A_i\right| + |A_{l+1}| - \sum_{\substack{I\subseteq\{1,...,l\}\\ \text{und } I\neq\varnothing}} (-1)^{1+|I|}\cdot\left|\bigcap_{i\in I}(A_i\cap A_{l+1})\right|.$$

In der ersten Summe lautet die Bedingung des Laufindex „$I\subseteq\{1,...,l\}$ und $I\neq\varnothing$“. Man verändert nichts, wenn man diese durch „$I\subseteq\{1,...,l+1\}$ und $I\neq\varnothing$ und $l+1\notin I$“ ersetzt. In der zweiten Summe lautet die Bedingung des Laufindex ebenfalls „$I\subseteq\{1,...,l\}$ und $I\neq\varnothing$“, allerdings taucht als Summand nicht $\left|\bigcap_{i\in I} A_i\right|$, sondern $\left|\bigcap_{i\in I}(A_i\cap A_{l+1})\right|$ auf, d.h. die durch den jeweiligen Laufindex *I* bestimmten Mengen A_i werden noch mit A_{l+1} geschnitten; man kann also zu jedem Laufindex *I* noch den Index $l+1$ hinzunehmen. In der zweiten Summe wird die Bedingung „$I\subseteq\{1,...,l\}$ und $I\neq\varnothing$“ des Laufindex durch „$I\subseteq\{1,...,l+1\}$ und $I\neq\varnothing$ und $l+1\in I$“ ersetzt und die Summanden entsprechend angepasst; der Summand $|A_{l+1}|$ kann in die Summe mit aufgenommen werden ($I=\{l+1\}$). Damit wird die letzte Gleichung zu

$$\begin{aligned}\left|\bigcup_{i=1}^{l+1} A_i\right| &= \sum_{\substack{I\subseteq\{1,...,l+1\}\\ \text{und } I\neq\varnothing\\ \text{und } l+1\notin I}} (-1)^{1+|I|}\cdot\left|\bigcap_{i\in I} A_i\right| + \sum_{\substack{I\subseteq\{1,...,l+1\}\\ \text{und } I\neq\varnothing\\ \text{und } l+1\in I}} (-1)^{1+|I|}\cdot\left|\bigcap_{i\in I} A_i\right|\\ &= \sum_{\substack{I\subseteq\{1,...,l+1\}\\ \text{und } I\neq\varnothing}} (-1)^{1+|I|}\cdot\left|\bigcap_{i\in I} A_i\right|.\end{aligned}$$

Die Formel in Satz 4.3-1 gilt also auch für $l+1$ Teilmengen.

Eine direkte Folgerung aus Satz 4.3-1 ist der folgende

Satz 4.3-2:

Es seien $A_1, ..., A_l$ Teilmengen einer endlichen Menge *M*, $l\geq 1$ Dann ist die Anzahl der $x\in M$, die in keiner der Mengen $A_1, ..., A_l$ liegen, gleich

$$|M| + \sum_{\substack{I\subseteq\{1,...,l\}\\ \text{und } I\neq\varnothing}} (-1)^{|I|}\cdot\left|\bigcap_{i\in I} A_i\right|.$$

Die Formel ergibt sich wie folgt:

$$\left|M \setminus \bigcup_{i=1}^{l} A_i\right| = |M| - \left|\bigcup_{i=1}^{l} A_i\right|$$

$$= |M| - \sum_{\substack{I \subseteq \{1,\ldots,l\} \\ \text{und } I \neq \varnothing}} (-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right| \quad \text{nach Satz 4.3-1}$$

$$= |M| + \sum_{\substack{I \subseteq \{1,\ldots,l\} \\ \text{und } I \neq \varnothing}} (-1)^{|I|} \cdot \left|\bigcap_{i \in I} A_i\right|.$$

Beide Sätze finden ihre Anwendung in vielen Teilen der Mathematik.

Als Beispiel dient der Beweis von Satz 4.2-4:

Es seien A und B endliche Mengen mit $|A| = n$ und $|B| = m$ und $n \geq m$, $A = \{a_1, \ldots, a_n\}$, $B = \{b_1, \ldots, b_m\}$. Die Mengen A_i werden definiert durch
$A_i = \{ f \mid f : A \to B$ und es gibt kein $a \in A$ mit $f(a) = b_i \}$ für $i = 1, \ldots, m$.
Eine Abbildung $f : A \to B$ ist damit genau dann surjektiv, wenn f in keiner der Mengen A_i für $i = 1, \ldots, m$ enthalten ist.

Es sei $M = \{ f \mid f : A \to B \}$. Nach Satz 4.3-1 ist $|M| = m^n$.

In der Terminologie von Satz 4.3-2 wird die Anzahl der $f \in M$ gesucht, die in keiner der Mengen $A_1, \ldots, A_m$ liegen. Diese Anzahl ist

$$|M| + \sum_{\substack{I \subseteq \{1,\ldots,m\} \\ \text{und } I \neq \varnothing}} (-1)^{|I|} \cdot \left|\bigcap_{i \in I} A_i\right| = m^n + \sum_{\substack{I \subseteq \{1,\ldots,m\} \\ \text{und } I \neq \varnothing}} (-1)^{|I|} \cdot \left|\bigcap_{i \in I} A_i\right|.$$

Für $I \subseteq \{1, \ldots, m\}$ mit $I \neq \varnothing$ wird $\bigcap_{i \in I} A_i$ betrachtet. Für jedes $f \in \bigcap_{i \in I} A_i$ gilt: es gibt kein $a \in A$, das durch f auf b_i abgebildet wird, wobei $i \in I$ ist. Also ist f eine Abbildung $f : A \to B \setminus \{ b_i \mid i \in I \}$.

Ist umgekehrt f eine Abbildung mit $f : A \to B \setminus \{ b_i \mid i \in I \}$, so ist $f \in \bigcap_{i \in I} A_i$. Daher ist nach Satz 4.2-1

$$\left|\bigcap_{i \in I} A_i\right| = \left|\{ f \mid f : A \to B \setminus \{ b_i \mid i \in I \} \}\right| = (m - |I|)^n.$$

Nach Satz 4.1-4 gibt es $\binom{m}{|I|}$ viele Teilmengen $I \subseteq \{1,...,m\}$ mit Mächtigkeit $|I|$. Damit wird

$$
\begin{aligned}
|M| + \sum_{\substack{I \subseteq \{1,...,m\} \\ \text{und } I \neq \varnothing}} (-1)^{|I|} \cdot \left| \bigcap_{i \in I} A_i \right| &= m^n + \sum_{\substack{I \subseteq \{1,...,m\} \\ \text{und } I \neq \varnothing}} (-1)^{|I|} \cdot \left| \bigcap_{i \in I} A_i \right| \\
&= m^n + \sum_{\substack{I \subseteq \{1,...,m\} \\ \text{und } I \neq \varnothing}} (-1)^{|I|} \cdot (m - |I|)^n \\
&= m^n + \sum_{k=1}^{m} (-1)^k \cdot \binom{m}{k} (m-k)^n \\
&= \sum_{k=0}^{m} (-1)^k \cdot \binom{m}{k} (m-k)^n \\
&= \sum_{i=0}^{m} (-1)^{m-i} \cdot \binom{m}{m-i} i^n \qquad \text{(mit der Indextransformation } i = m-k) \\
&= \sum_{i=0}^{m} (-1)^{m-i} \cdot \binom{m}{i} i^n \qquad \text{(gemäß Satz 4.1-3 (vi)).}
\end{aligned}
$$

Das ist das Ergebnis aus Satz 4.2-4.

5 Ausgewählte Themen der Analysis

Die für praktische Anwendungen wichtigsten Themen der Analysis werden im mathematischen Schulunterricht behandelt. Daher kann das vorliegende Kapitel als Wiederholung und Vertiefung dieser Themen betrachtet werden. Darüber hinaus werden weiterführende Themen wie Taylorpolynome und erzeugende Funktionen und ihre Anwendungen behandelt.

5.1 Folgen und Reihen

Wird jedes $n \in \mathbf{N}$ nach einer bestimmten Vorschrift eine reelle Zahl $a_n \in \mathbf{R}$ zugeordnet, so entsteht eine reellwertige Zahlenfolge $a_0, a_1, a_2, ...$ Sie wird mit $(a_n)_{n\in\mathbf{N}}$ bezeichnet. a_n heißt auch ***n*-tes Folgenglied von** $(a_n)_{n\in\mathbf{N}}$.

In der Regel stellt a_n eine von n abhängige Formel dar. Beispielsweise ist für $a_n = \sqrt{n+1} - \sqrt{n}$ die entsprechende Folge $(a_n)_{n\in\mathbf{N}} = \left(\sqrt{n+1} - \sqrt{n}\right)_{n\in\mathbf{N}}$.

Die Definition einer Folge kann auf unterschiedliche Weise geschehen, beispielsweise:

$a_0 = 1,\ a_1 = 2,\ a_2 = 4,\ a_3 = 8, ..., a_n = 2^n, ...$

oder

$(a_n)_{n\in\mathbf{N}} = (2^n)_{n\in\mathbf{N}}$

oder als **rekursive Definition**

$a_0 = 1,\ a_n = 2 \cdot a_{n-1}$ für $n \geq 1$.

Eine Zahl $a \in \mathbf{R}$ heißt **Grenzwert (Limes)** der Folge $(a_n)_{n\in\mathbf{N}}$, wenn folgender Sachverhalt gilt:

Für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ gibt es eine (eventuell von ε abhängige) natürliche Zahl $n_0 = n_0(\varepsilon)$ mit der Eigenschaft:
Für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ gilt $|a_n - a| < \varepsilon$.

Die Folge heißt dann **gegen** *a* **konvergent** (sie **konvergiert gegen** *a*), und man schreibt $a = \lim\limits_{n\to\infty} a_n$ bzw. $a_n \to a$ für $n \to \infty$.

Die Konvergenz der Folge $(a_n)_{n\in\mathbf{N}}$ gegen den Wert $a \in \mathbf{R}$ bedeutet anschaulich, dass bei Vorgabe eines beliebig kleinen Werts $\varepsilon > 0$ *alle* Folgenglieder a_n, bis auf höchstens endlich viele Ausnahmen („*fast alle*“ Folgenglieder), „dicht“ bei *a*, genauer einen Abstand von *a* haben, der kleiner als ε ist. Verkleinert man ε auf den Wert $\varepsilon' < \varepsilon$, so steigt eventuell die Anzahl der Ausnahmen, die nicht dicht bei *a* liegen; es bleiben aber weiterhin höchstens endlich viele Ausnahmen.

Beispiel:

Die Folge $(a_n)_{n\in\mathbf{N}} = \left(\sqrt{n+1} - \sqrt{n}\right)_{n\in\mathbf{N}}$ konvergiert gegen 0: Gibt man $\varepsilon > 0$ vor und setzt $n_0 = n_0(\varepsilon) = \left\lceil \left(\frac{1}{2\cdot\varepsilon}\right)^2 \right\rceil + 1$, so gilt für $n \geq n_0$:

$$\begin{aligned}
\left|\left(\sqrt{n+1} - \sqrt{n}\right) - 0\right| &= \sqrt{n+1} - \sqrt{n} \\
&= \frac{\left(\sqrt{n+1} - \sqrt{n}\right)\left(\sqrt{n+1} + \sqrt{n}\right)}{\sqrt{n+1} + \sqrt{n}} \\
&= \frac{1}{\sqrt{n+1} + \sqrt{n}} \\
&< \frac{1}{2\cdot\sqrt{n}} \qquad \text{(wegen } \sqrt{n+1} > \sqrt{n}\text{)} \\
&\leq \frac{1}{2\cdot\sqrt{n_0}} \\
&< \frac{\sqrt{(2\cdot\varepsilon)^2}}{2} = \varepsilon \qquad \text{(nach Wahl von } n_0\text{)}.
\end{aligned}$$

Wie man sieht, ist es nicht immer ganz leicht, bei Vorgabe von $\varepsilon > 0$ die passende Zahl $n_0(\varepsilon)$ zu finden, von der an alle Folgenglieder dicht beim Grenzwert *a* liegen, den man zudem bereits kennen oder vermuten muss. Häufig berechnet man die Werte a_n einiger Folgenglieder für wachsende Werte von *n*. Falls man sieht, dass sich diese Werte einer Größe *a* annähern, stellt man die allgemeine Ungleichung $|a_n - a| < \varepsilon$ auf und versucht diese nach *n* aufzulösen. Ergibt sich dabei eine Ungleichung der Form „$n > \ldots$“, dann kann man als $n_0(\varepsilon)$ die kleinste natürliche Zahl wählen, von der an die Ungleichung „$n > \ldots$“ gilt:

Beispiel:

Für die Folge $(a_n)_{n\in\mathbf{N}} = \left(\frac{2^n + (-1)^n}{2^n}\right)_{n\in\mathbf{N}}$ werden zunächst einige Werte berechnet:

n	0	1	2	3	4	5	10	15	100
a_n	2	1/2	5/4	7/8	17/16	31/32	1025/1024	32 767/32 768	1,00000000000000000000000000008

Die Vermutung liegt nahe, dass diese Folge gegen $a = 1$ konvergiert, also wird versucht, die Ungleichung $|a_n - 1| < \varepsilon$ nach *n* aufzulösen:

$$|a_n - 1| = \left|\frac{2^n + (-1)^n}{2^n} - 1\right| = \left|\frac{(-1)^n}{2^n}\right| = \frac{1}{2^n} < \varepsilon \text{, d.h. } 2^n > 1/\varepsilon .$$

Man kann die nächstgrößere 2er-Potenz oberhalb $1/\varepsilon$ wählen, und n_0 ist dann der Exponent in dieser 2er-Potenz. Für jedes $n \geq n_0$ ist dann $2^n \geq 2^{n_0} > 1/\varepsilon$ bzw. $|a_n - 1| = 1/2^n \leq 1/2^{n_0} < \varepsilon$.

Eine Folge, die nicht konvergent ist, also keinen Grenzwert hat, heißt **divergent**.

Für eine divergente Folge gilt
entweder:

(i) es gibt mindestens zwei reelle Zahlen $a \in \mathbf{R}$ und $b \in \mathbf{R}$ mit $a \neq b$, so dass es sowohl beliebig dicht bei *a* als auch beliebig dicht bei *b* unendlich viele Folgenglieder gibt

oder:

(ii) die Werte a_n werden mit wachsendem *n* beliebig groß bzw. beliebig klein (Schreibweise dann $a_n \to \infty$, $\lim\limits_{n\to\infty} a_n = \infty$, bzw. $a_n \to -\infty$, $\lim\limits_{n\to\infty} a_n = -\infty$).

Im Fall (i) sagt man, die Folge habe mindestens zwei verschiedene (reelle) **Häufungspunkte**.

Beispielsweise besitzt die Folge $(a_n)_{n\in\mathbf{N}} = \left((-1)^n\right)_{n\in\mathbf{N}}$ die beiden Häufungspunkte $+1$ und -1 und ist daher nicht konvergent.

Eine gegen eine reelle Zahl *a* konvergente Folge besitzt nur den einzigen Häufungspunkt *a*, der dann auch Grenzwert der Folge ist.

Satz 5.1-1:

(i) Jede Folge $(a_n)_{n\in\mathbf{N}}$, die gegen einen Grenzwert konvergiert, ist beschränkt, d.h. es gibt eine Konstante $C \in \mathbf{R}$ mit $|a_n| < C$ für alle $n \in \mathbf{N}$.

(ii) Konvergiert die Folge $(a_n)_{n\in\mathbf{N}}$ gegen $a \in \mathbf{R}$, so ist *a* eindeutig bestimmt.

Teil (i) ist anschaulich klar. Formal sieht man dessen Gültigkeit so:
Konvergiert die Folge $(a_n)_{n\in\mathbf{N}}$ gegen $a \in \mathbf{R}$, so gibt es für $\varepsilon = 1$ eine natürliche Zahl n_0, so dass für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ $|a_n - a| < \varepsilon = 1$ gilt. Dann ist $|a_n| \leq \max\{|a_0|, ..., |a_{n_0-1}|\}$ für $n < n_0$ und $|a_n| = |a_n - a + a| \leq |a_n - a| + |a| < 1 + |a|$ für $n \geq n_0$. Also gilt für alle $n \in \mathbf{N}$:
$|a_n| \leq \max\{|a_0|, ..., |a_{n_0-1}|, 1 + |a|\}$.

Teil (ii) sieht man folgendermaßen:
Konvergiert die Folge $(a_n)_{n\in\mathbf{N}}$ gegen $a \in \mathbf{R}$ und gegen $b \in \mathbf{R}$ mit $a \neq b$, so gibt es für $\varepsilon = 1/2 \cdot |b - a| > 0$ eine natürliche Zahl n_0, so dass für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ $|a_n - a| < \varepsilon$ gilt. Außerdem gibt es eine natürliche Zahl n_1, so dass für jedes $n \in \mathbf{N}$ mit $n \geq n_1$ $|a_n - b| < \varepsilon$ gilt. Für $n \geq \max\{n_0, n_1\}$ ergibt sich der Widerspruch
$|b - a| = |b - a_n + a_n - a| \leq |b - a_n| + |a_n - a| < 2 \cdot \varepsilon = |b - a|$.

Die Limesbildung und das Rechnen mit arithmetischen Ausdrücken ist häufig miteinander vertauschbar. So kann man den Grenzwert einer Folge, deren Folgenglieder sich als arithmetischer Ausdruck (gebildet mit den Operatoren +, − , · und /) von Folgenglieder konvergenter Folgen darstellen lassen, dadurch berechnen, dass man die Limesbildung in den arithmetischen Ausdruck hineinzieht: Man berechnet die Grenzwerte der einzelnen Teile und verknüpft diese dann gemäß dem arithmetischen Ausdruck. Konstante Faktoren, die nicht von *n* abhängen, kann man jeweils vor den Limes ziehen. Die Grundlage dieses Kalküls liefert der folgende Satz.

Satz 5.1-2:

Es seien $(a_n)_{n\in\mathbf{N}}$ bzw. $(b_n)_{n\in\mathbf{N}}$ zwei konvergente Folgen mit den Grenzwerten *a* bzw. *b*. Dann gilt:

(i) $\lim\limits_{n\to\infty}(a_n \pm b_n) = a \pm b$,
$\lim\limits_{n\to\infty}(a_n \cdot b_n) = a \cdot b$.

(ii) Für $r \in \mathbf{R}$ ist $\lim\limits_{n\to\infty}(r \cdot a_n) = r \cdot \lim\limits_{n\to\infty} a_n = r \cdot a$.

(iii) Gilt $b \neq 0$ und $b_n \neq 0$ für alle $n \in \mathbf{N}$, so ist $\lim\limits_{n\to\infty}\left(\frac{a_n}{b_n}\right) = \frac{a}{b}$.

(iv) Aus $\lim\limits_{n\to\infty} a_n = a$ kann man $\lim\limits_{n\to\infty}|a_n| = |a|$ schließen.
Für $a = 0$ gilt auch die Umkehrung: $\lim\limits_{n\to\infty}|a_n| = 0$ impliziert $\lim\limits_{n\to\infty} a_n = 0$.

(v) Jede (fast überall) konstante Folge $(a_n)_{n\in\mathbf{N}}$ konvergiert, genauer:
Ist $a_n = a$ für (fast) alle $n \in \mathbf{N}$, so ist
$\lim\limits_{n\to\infty} a_n = a$.

Exemplarisch sollen die Teile (i), (iii) und (iv) bewiesen werden:

Nach Definition der Konvergenz gibt es zu $\varepsilon > 0$ eine natürliche Zahl n_0, so dass für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ $|a_n - a| < \varepsilon/2$ gilt. Entsprechend gibt es eine natürliche Zahl n_1 mit $|b_n - b| < \varepsilon/2$ für jedes $n \in \mathbf{N}$ mit $n \geq n_1$. Für $n \geq \max\{n_0, n_1\}$ ist
$|a_n + b_n - (a+b)| \leq |a_n - a| + |b_n - b| < \varepsilon$. Für die Folge $(a_n - b_n)_{n\in\mathbf{N}}$ wird entsprechend argumentiert.

Es sei $\varepsilon > 0$ gegeben. Da die Folge $(a_n)_{n\in\mathbf{N}}$ konvergiert, gibt es nach Satz 5.1-1 (i) eine Konstante $C > 0$ mit $|a_n| < C$ für alle $n \in \mathbf{N}$. Es wird $\varepsilon_1 = \frac{\varepsilon}{2 \cdot C} > 0$ und $\varepsilon_2 = \frac{\varepsilon}{2 \cdot (1 + |b|)} > 0$ gesetzt. Dann gibt es natürliche Zahlen n_0 und n_1 mit $|a_n - a| < \varepsilon_2$ und $|b_n - b| < \varepsilon_1$ für $n \geq \max\{n_0, n_1\}$. Für diese *n* gilt:

$$\begin{aligned}|a_n \cdot b_n - a \cdot b| &= |a_n \cdot b_n - b \cdot a_n + b \cdot a_n - a \cdot b| \\ &\le |a_n| \cdot |b_n - b| + |b| \cdot |a_n - a| \\ &< C \cdot \varepsilon_1 + (|b| + 1) \cdot \varepsilon_2 \\ &= \varepsilon/2 + \varepsilon/2 = \varepsilon.\end{aligned}$$

Für Teil (iii) wird zunächst der Spezialfall $a_n = 1$ und $a = 1$ für alle $n \in \mathbf{N}$ untersucht. Der allgemeine Fall folgt dann aus (i).
Es sei $\varepsilon > 0$. Aufgrund der Konvergenz der Folge $(b_n)_{n\in\mathbf{N}}$ gegen b gibt eine natürliche Zahl n_1 mit $|b_n - b| < \frac{|b|}{2}$ für jedes $n \in \mathbf{N}$ mit $n \ge n_1$. Daraus folgt $|b_n| > \frac{|b|}{2}$. Weiterhin gibt es eine natürliche Zahl n_0 mit $|b_n - b| < \frac{\varepsilon \cdot |b|^2}{2}$ für jedes $n \in \mathbf{N}$ mit $n \ge n_0$. Dann gilt für $n \ge \max\{n_0, n_1\}$: $\left|\frac{1}{b_n} - \frac{1}{b}\right| = \left|\frac{b - b_n}{b_n \cdot b}\right| = \frac{|b - b_n|}{|b_n| \cdot |b|} < \frac{\varepsilon \cdot |b|^2 \cdot 2}{2 \cdot |b|^2} = \varepsilon.$

Für (iv) verwendet man Satz 1.7-1 (vi):
Es ist $\big||a_n| - |a|\big| \le |a_n - a|$. Gilt also $|a_n - a| < \varepsilon$, dann auch $\big||a_n| - |a|\big| < \varepsilon$. Für $a = 0$ ist $|a_n - a| = |a_n - 0| = \big||a_n| - 0\big| = \big||a_n| - a\big|$.

„Ähnlich“ aussehende Folgen verhalten sich bezüglich der Konvergenz häufig sehr unterschiedlich: So ist die durch $a_n = \frac{2^n + (-2)^n}{2^n}$ definierte Folge $(a_n)_{n\in\mathbf{N}}$ nicht konvergent. Dagegen konvergiert die durch $b_n = \frac{2^n + (-2)^n}{3^n}$ definierte Folge $(b_n)_{n\in\mathbf{N}}$ gegen 0.

Häufig ist das Konvergenzverhalten einer Folge $(a_n)_{n\in\mathbf{N}}$ zu untersuchen. Ist eine Zahl $a \in \mathbf{R}$ „verdächtig“, Grenzwert der Folge $(a_n)_{n\in\mathbf{N}}$ zu sein, so lässt sich durch Rückgriff auf die Definition des Limesbegriffs nachprüfen, ob die Folge tatsächlich konvergiert, und zwar gegen a, d.h. ob $\lim\limits_{n\to\infty} a_n = a$ gilt. Kann man einer Folge, ohne von ihrem möglichen Grenzwert etwas zu wissen, ansehen, dass sie konvergiert? Die folgenden Sätze liefern einige **Konvergenzkriterien für Folgen**. Daneben gibt es eine Reihe weiterer Konvergenzkriterien, die mit Hinweis auf die angegebene Literatur hier nicht angeführt werden.

Satz 5.1-3:

Es seien $(a_n)_{n\in\mathbf{N}}$ bzw. $(b_n)_{n\in\mathbf{N}}$ zwei konvergente Folgen mit den Grenzwerten *a* bzw. *b*. Dann gilt:

(i) Es seien $(a_n)_{n\in\mathbf{N}}$ und $(b_n)_{n\in\mathbf{N}}$ zwei konvergente Folgen mit demselben Grenzwert $a \in \mathbf{R}$. Für fast alle Folgenglieder c_n der Folge $(c_n)_{n\in\mathbf{N}}$ gelte $a_n \le c_n \le b_n$. Dann konvergiert auch die Folge $(c_n)_{n\in\mathbf{N}}$, und zwar zum selben Grenzwert *a*.

(ii) Jede nach oben beschränkte und monoton wachsende Folge konvergiert, und ihr Limes ist gleich der kleinsten oberen Schranke (**Supremum**) ihrer Wertemenge.

Jede nach unten beschränkte und monoton fallende Folge konvergiert, und ihr Limes ist gleich der größten unteren Schranke (**Infimum**) ihrer Wertemenge.

Eine unbeschränkte, monoton wachsende bzw. monoton fallende Folge strebt gegen ∞ bzw. $-\infty$.

Bemerkung: Nicht jede beschränkte Folge ist konvergent.

(iii) Jede Umordnung und jede Teilfolge einer konvergenten Folge ist ebenfalls konvergent mit demselben Grenzwert. Dasselbe gilt, wenn man endlich viele Folgenglieder einer konvergenten Folge abändert.

Teil (i) ist wieder anschaulich klar. Formal lässt sich die Aussage wie folgt nachweisen:
Es sei $\varepsilon > 0$ und n_0 eine natürliche Zahl, so dass für jedes $n \in \mathbf{N}$ mit $n \ge n_0$ $|a_n - a| < \varepsilon/6$ gilt. Entsprechend gibt es eine natürliche Zahl n_1 mit $|b_n - a| < \varepsilon/6$ für jedes $n \in \mathbf{N}$ mit $n \ge n_1$. Für $n \ge \max\{n_0, n_1\}$ ist $|a_n - b_n| = |a_n - a + a - b_n| \le |a_n - a| + |b_n - a| < 2 \cdot \varepsilon/6 = \varepsilon/3$ und

$$|c_n - a| = |c_n - a_n + a_n - b_n + b_n - a| \le |c_n - a_n| + |a_n - b_n| + |b_n - a| \le |b_n - a_n| + |a_n - b_n| + |b_n - a|$$
$$\le 2 \cdot \varepsilon/3 + \varepsilon/6 < \varepsilon.$$

Für (ii) wird nur der erste Teil dargestellt:
Es sei $(a_n)_{n\in\mathbf{N}}$ eine beschränkte und monoton wachsende Folge. Mit *s* werde die kleinste obere Schranke der Menge $M = \{a_n \mid n \in \mathbf{N}\}$ aller Folgenglieder bezeichnet. Weiterhin sei $\varepsilon > 0$ vorgegeben. Dann ist $a_n \le s < s + \varepsilon$. Der Wert $s - \varepsilon$ ist keine obere Schranke von *M*. Daher gibt es eine natürliche Zahl n_0, so dass $a_{n_0} > s - \varepsilon$ gilt. Für jedes $n \in \mathbf{N}$ mit $n \ge n_0$ gilt wegen

der Monotonie der Folge ebenfalls $a_n > s - \varepsilon$, insgesamt $s - \varepsilon < a_n < s + \varepsilon$ bzw. $|a_n - s| < \varepsilon$. Also konvergiert $(a_n)_{n \in \mathbf{N}}$ gegen *s*.

Teil (iii) lässt sich durch Umnumerierung der Folgenglieder zeigen. Auf Details soll hier verzichtet werden.

Das folgende **Konvergenzkriterium von Cauchy** gibt eine notwendige und hinreichende Eigenschaft der Konvergenz einer Folge an:

Satz 5.1-4:

Eine Folge $(a_n)_{n \in \mathbf{N}}$ ist genau dann konvergent, wenn es zu jedem $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ eine natürliche Zahl $n_0 = n_0(\varepsilon)$ gibt, so dass gilt:

$|a_n - a_m| < \varepsilon$ für jedes $n \in \mathbf{N}$ und jedes $m \in \mathbf{N}$ mit $n \geq n_0$ und $m \geq n_0$.

Bemerkung: Ohne Beschränkung der Allgemeinheit kann angenommen werden, dass $m > n$ gilt. Dann kann die Ungleichung auch in der Form $|a_{n+k} - a_n| < \varepsilon$ für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ und jedes $k \in \mathbf{N}$ geschrieben werden.

Der Satz beinhaltet zwei Beweisrichtungen:

Die Folge $(a_n)_{n \in \mathbf{N}}$ sei konvergent gegen *a*. Dann gibt es zu $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ eine natürliche Zahl n_0, so dass für jedes $n \in \mathbf{N}$ und jedes $m \in \mathbf{N}$ mit $n \geq n_0$ und $m \geq n_0$ gilt: $|a_n - a| < \varepsilon/2$ und $|a_m - a| < \varepsilon/2$. Damit folgt $|a_n - a_m| = |a_n - a + a - a_m| \leq |a_n - a| + |a_m - a| < \varepsilon$.

Für die umgekehrte Richtung soll die Beweisidee nur skizziert werden: Gibt es zu $\varepsilon > 0$ eine natürliche Zahl n_0 mit $|a_n - a_m| < \varepsilon/2$ für jedes $n \in \mathbf{N}$ und jedes $m \in \mathbf{N}$ mit $n \geq n_0$ und $m \geq n_0$, so ist die Folge $(|a_n|)_{n \in \mathbf{N}}$ beschränkt; denn $|a_n - a_{n_0}| < \varepsilon/2$ für alle $n \in \mathbf{N}$ mit $n \geq n_0$ impliziert $|a_n| < |a_{n_0}| + \varepsilon/2$; daher ist die größere der beiden Zahlen $\max\{|a_n| \mid n < n_0\}$ und $|a_{n_0}| + \varepsilon/2$ eine Schranke für die Werte der Folge $(|a_n|)_{n \in \mathbf{N}}$. Als beschränkte Folge besitzt sie einen Häufungspunkt *a* (hier ohne Beweis), also einen Wert, an dem unendlich viele Folgenglieder beliebig nahe liegen. Für unendlich viele $m \in \mathbf{N}$ gilt $|a_m - a| < \varepsilon/2$. Man wählt ein derartiges $m \geq n_0$ und erhält für $n \geq n_0$: $|a_n - a| = |a_n - a_m + a_m - a| \leq |a_n - a_m| + |a_m - a| < \varepsilon$.

Satz 5.1-5:

(i) Es sei $q \in \mathbf{R}$. Dann gilt

$$\lim_{n\to\infty} q^n = \begin{cases} 0 & \text{für } -1 < q < 1 \\ 1 & \text{für } q = 1 \end{cases};$$

für $q > 1$ und $q \le -1$ ist $\left(q^n\right)_{n\in\mathbf{N}}$ divergent.

Ist $k \in \mathbf{N}$ und $|q| < 1$, so ist $\lim_{n\to\infty}(n^k \cdot q^n) = 0$.

(ii) Es sei $a \in \mathbf{R}$. Dann ist

$$\lim_{n\to\infty} n^a = \begin{cases} 0 & \text{für } a < 0 \\ 1 & \text{für } a = 0 \\ \infty & \text{für } a > 0 \end{cases}.$$

(iii) Für jedes $a \in \mathbf{R}$ ist

$$\lim_{n\to\infty} \frac{a^n}{n!} = 0.$$

(iv) $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e = 2.7182818284590\ldots;$

e heißt **Eulersche Konstante**.

(v) Die Folge $\left(\sqrt{n}\right)_{n\in\mathbf{N}}$ divergiert: mit wachsendem *n* werden die Folgenglieder beliebig groß. Hingegen werden die Zuwächse von einem Folgenglied zum nächsten mit wachsendem *n* beliebig klein; denn die Folge

$$\left(\sqrt{n+1} - \sqrt{n}\right)_{n\in\mathbf{N}}$$

konvergiert gegen 0.

(vi) Die Folge $\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right)_{n\in\mathbf{N}_{>0}}$ divergiert.

Für Teil (i) sei zunächst $0 < q < 1$. Die Folge ist monoton fallend und nach unten beispielsweise durch 0 beschränkt. Daher ist sie gemäß Satz 5.1-3 (ii) konvergent, etwa gegen den Grenzwert *a*. Aus $\left(q^n\right)_{n\in\mathbf{N}}$ wird die neue Folge $\left(q^{n+1}\right)_{n\in\mathbf{N}}$ gebildet, die durch Fortlassen des ersten Folgenglieds von $\left(q^n\right)_{n\in\mathbf{N}}$ entsteht. Diese konvergiert ebenfalls gegen *a*. Daher ist $a = \lim\limits_{n\to\infty} q^n = \lim\limits_{n\to\infty} q^{n+1} = q \cdot \lim\limits_{n\to\infty} q^n = q \cdot a$, also $a \cdot (1-q) = 0$. Wegen $0 < q < 1$ ist $a = 0$.

Es werde jetzt $n > \left(\sqrt[k]{1/q} - 1\right)^{-1}$ gewählt. Dann folgt nacheinander $\frac{1}{n} < \sqrt[k]{1/q} - 1$, $1 + \frac{1}{n} < \sqrt[k]{1/q}$, $\left(1 + \frac{1}{n}\right)^k < 1/q$ und $\frac{(n+1)^k \cdot q^{n+1}}{n^k \cdot q^n} = \left(1 + \frac{1}{n}\right)^k \cdot q < 1$. Das zeigt, dass die Folge $\left(n^k \cdot q^n\right)_{n\in\mathbf{N}}$ fast überall monoton fällt; sie ist nach unten beschränkt; also konvergiert sie. Der „vermutete" Grenzwert ist 0. Diese Vermutung lässt sich bestätigen:
Für $k = 1$ sei *b* der Grenzwert von $\left(n \cdot q^n\right)_{n\in\mathbf{N}}$. Dann ist *b* ebenfalls der Grenzwert von $\left((n+1) \cdot q^{n+1}\right)_{n\in\mathbf{N}}$, und es gilt $b = \lim\limits_{n\to\infty}\left(n \cdot q^n\right) = \lim\limits_{n\to\infty}\left((n+1) \cdot q^{n+1}\right) = q \cdot \lim\limits_{n\to\infty}\left(n \cdot q^n\right) + q \cdot \lim\limits_{n\to\infty} q^n = q \cdot b$, also $b = 0$. Für $k > 1$ kann man wie folgt argumentieren: Es sei $\varepsilon > 0$ und $n_0 \in \mathbf{N}$ so gewählt, dass für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ gilt: $n \cdot \left(\sqrt[k]{q}\right)^n < \sqrt[k]{\varepsilon}$ (das ist möglich, da $\lim\limits_{n\to\infty}\left(n \cdot \left(\sqrt[k]{q}\right)^n\right) = 0$ ist). Dann ist $n^k \cdot q^n = \left(n \cdot \left(\sqrt[k]{q}\right)^n\right)^k < \left(\sqrt[k]{\varepsilon}\right)^k = \varepsilon$.

Ist $-1 < q < 0$, so konvergieren die Folgen $\left(|q|^n\right)_{n\in\mathbf{N}}$ bzw. $\left(n^k \cdot |q|^n\right)_{n\in\mathbf{N}}$ gegen 0. Nach Satz 5.1-2 (iv) gilt $\lim\limits_{n\to\infty} q^n = 0$ bzw. $\lim\limits_{n\to\infty}\left(n^k \cdot q^n\right) = 0$.

Für Teil (ii) mit $a < 0$ sei zu $\varepsilon > 0$ die natürliche Zahl $n_0 \in \mathbf{N}$ so gewählt, dass für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ gilt: $\frac{1}{n} < \varepsilon^{1/|a|}$. Für diese *n* ergibt sich dann $n^a = \left(\frac{1}{n}\right)^{|a|} < \left(\varepsilon^{1/|a|}\right)^{|a|} = \varepsilon$. Das bedeutet Konvergenz gegen 0.

Für Teil (iii) wird eine Fallunterscheidung nach der Größe von *a* getroffen:
Für $0 \leq a \leq 1$ ist $0 = \frac{0^n}{n!} \leq \frac{a^n}{n!} \leq \frac{1^n}{n!} = \frac{1}{n!}$. Mit Satz 5.1-3 (i) folgt $\lim\limits_{n\to\infty} \frac{a^n}{n!} = 0$.
Für $a > 1$ sei $n_1 \in \mathbf{N}$ diejenige eindeutig bestimmte natürliche Zahl mit $n_1 \leq a < n_1 + 1$. Für $n \geq n_1 + 1$ gilt dann $0 \leq \frac{a^n}{n!} = \frac{a \cdot a \cdot \ldots \cdot a \cdot a \cdot \ldots a}{1 \cdot 2 \cdot \ldots n_1 \cdot (n_1+1) \cdot \ldots \cdot n} \leq C \cdot \frac{a^{n-n_1}}{(n_1+1)^{n-n_1}} = C \cdot \left(\frac{a}{n_1+1}\right)^{n-n_1}$ mit der

von n unabhängigen Konstanten $C = \frac{a^{n_1}}{(n_1)!}$. Da $\frac{a}{n_1+1} < 1$ ist, konvergiert gemäß Teil (i) die rechte Seite gegen 0, und damit ist $\lim_{n\to\infty} \frac{a^n}{n!} = 0$.

Für $a < 0$ ist $\lim_{n\to\infty} \frac{|a|^n}{n!} = 0$, und nach Satz 5.1-2 (iv) ist $\lim_{n\to\infty} \frac{a^n}{n!} = 0$.

Für Teil (iv) wird gezeigt, dass die Folge $\left(\left(1+\frac{1}{n}\right)^n\right)_{n\in\mathbf{N}_{>0}}$ monoton wächst und beschränkt ist.

Die Berechnung des Grenzwerts erfolgt dann am Ende des Kapitels.
Einige Zahlenbeispiele machen die Monotonie und Beschränktheit der Folge plausibel (die Zahlenwerte sind in den letzten Stellen teilweise gerundet):

n	1	2	3	4	5	6	100	10.000	1.000.000
a_n	2	2,25	2,3703	2,4414	2,4883	2,5216	2,7048	2,718146	2,7182804693

Wie man leicht durch vollständige Induktion zeigt, gilt für $x \in \mathbf{R}$ mit $x \geq -1$ und $n \in \mathbf{N}$ die Bernoulli'sche Ungleichung: $(1+x)^n \geq 1 + n \cdot x$.

Es sei $a_n = \left(1+\frac{1}{n}\right)^n$. Dann ist

$\frac{a_{n+1}}{a_n} = \frac{(n+2)^{n+1}}{(n+1)^{n+1}} \cdot \frac{n^n}{(n+1)^n} = \frac{n+2}{n+1} \cdot \left(\frac{(n+2)\cdot n}{(n+1)^2}\right)^n = \left(1+\frac{1}{n+1}\right) \cdot \left(1-\frac{1}{(n+1)^2}\right)^n$. Mit der Bernoulli'schen Ungleichung folgt

$\frac{a_{n+1}}{a_n} = \left(1+\frac{1}{n+1}\right) \cdot \left(1-\frac{1}{(n+1)^2}\right)^n \geq \left(1+\frac{1}{n+1}\right) \cdot \left(1-\frac{n}{(n+1)^2}\right) = 1 + \frac{1}{(n+1)^3} > 1$, also $a_{n+1} > a_n$.

Die Beschränkung folgt aus

$\left(1+\frac{1}{n}\right)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot \frac{1}{n^k} = 1 + 1 + \sum_{k=2}^{n} \binom{n}{k} \cdot \frac{1}{n^k}$. Für $k \geq 2$ ist

$\binom{n}{k} \cdot \frac{1}{n^k} = \frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{k!\cdot n^k} = \frac{1}{k!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \frac{n-k+1}{n} < \frac{1}{k!} = \frac{1}{1\cdot 2\cdot\ldots\cdot k} \leq \frac{1}{2^{k-1}}$.

Insgesamt ist daher

$\left(1+\frac{1}{n}\right)^n = 2 + \sum_{k=2}^{n} \binom{n}{k} \cdot \frac{1}{n^k} < 2 + \sum_{k=2}^{n} \frac{1}{2^{k-1}} = 2 + \sum_{k=1}^{n-1} \frac{1}{2^k} = 2 + \left(\frac{1-(1/2)^n}{1-1/2} - 1\right) = 3 - \left(\frac{1}{2}\right)^{n-1} < 3$.

Die Aussage in Teil (v) wurde am Anfang des Kapitels nachgewiesen.

In Teil (v) wird ein Folgenglied $a_n = 1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n}$ in Teilsummen zerlegt. Dazu sei *l* so bestimmt, dass $2^{l-1} \le n < 2^l$ ist. Dann ist $a_n = (1) + \left(\frac{1}{2} + \frac{1}{3}\right) + \left(\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7}\right) + ... + \left(\frac{1}{2^{l-1}} + ... + \frac{1}{n}\right)$. Die *k*-te Teilsumme für $1 \le k < l-1$, lautet $t_k = \left(\frac{1}{2^{k-1}} + \frac{1}{2^{k-1}+1} + ... + \frac{1}{2^k - 1}\right)$; sie hat 2^{k-1} viele Summanden, und es ist $\frac{1}{2} = 2^{k-1} \cdot \frac{1}{2^k} < t_k \le 2^{k-1} \cdot \frac{1}{2^{k-1}} = 1$. Damit lässt sich a_n abschätzen: $\frac{l-1}{2} < a_n = \sum_{k=1}^{l-1} t_k + \left(\frac{1}{2^{l-1}} + ... + \frac{1}{n}\right) \le \sum_{k=1}^{l} t_k \le l$. Mit $n \to \infty$ geht auch $\frac{l-1}{2} \to \infty$, daher divergiert $(a_n)_{n \in \mathbf{N}_{>0}}$.

Das Beispiel in Satz 5.1-5 (vi) zeigt eine Folge in einer speziellen Form: Das *n*-te Folgenglied ist selbst eine Summe aus einer endlichen Anzahl von Summanden, die aus einer Folge stammen, die nach einer einheitlichen Gesetzmäßigkeit aufgebaut ist. Derartige Folgen sollen nun genauer betrachtet werden.

Zur Zahlenfolge $(a_n)_{n \in \mathbf{N}}$ wird eine neue Zahlenfolge $(s_n)_{n \in \mathbf{N}}$ durch

$$s_n = \sum_{i=0}^{n} a_i$$

definiert. Der Wert s_n heißt ***n*-te Partialsumme** von $(a_n)_{n \in \mathbf{N}}$:

$s_0 = a_0$,
$s_1 = a_0 + a_1$,
...
$s_n = a_0 + a_1 + \ \ldots\ a_n = s_{n-1} + a_n$.

Falls der Grenzwert $\lim_{n\to\infty} s_n = \lim_{n\to\infty}\left(\sum_{i=0}^{n} a_i\right)$ existiert, so heißt er **unendliche Reihe der Folge** $(a_n)_{n\in\mathbf{N}}$ und wird mit $\sum_{i=0}^{\infty} a_i$ bezeichnet. Gelegentlich schreibt man auch $\sum_{i=0}^{\infty} a_i = a_0 + a_1 + a_2 + \ldots$

Satz 5.1-6:

(i) Der Grenzwert $\sum_{i=0}^{\infty} a_i$ existiert genau dann, wenn es zu jedem $\varepsilon > 0$ eine natürliche Zahl n_0 gibt, so dass $\left|\sum_{i=n+1}^{n+k} a_i\right| < \varepsilon$ für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ und jedes $k \in \mathbf{N}$ gilt.

(ii) Existiert der Grenzwert $\sum_{i=0}^{\infty} a_i$, so konvergiert die Folge $(a_n)_{n\in\mathbf{N}}$ gegen 0, d.h. aus $\sum_{i=0}^{\infty} a_i < \infty$ folgt $\lim_{n\to\infty} a_n = 0$.

In Aussage (i) ist $\sum_{i=0}^{\infty} a_i$ der Grenzwert der Folge der Partialsummen: $\sum_{i=0}^{\infty} a_i = \lim_{n\to\infty} s_n = \lim_{n\to\infty}\left(\sum_{i=0}^{n} a_i\right)$. Setzt man in Satz 5.1-4 die Folge der Partialsummen ein, so ist $|s_{n+k} - s_n| = \left|\sum_{i=0}^{n+k} a_i - \sum_{i=0}^{n} a_i\right| = \left|\sum_{i=n+1}^{n+k} a_i\right| < \varepsilon$.

Setzt man in (i) $k = 1$, so erhält man (ii).

Bemerkung: Die Umkehrung von (ii) gilt i.a. nicht, d.h. aus der Konvergenz der Folge $(a_n)_{n\in\mathbf{N}}$ gegen 0 folgt i.a. nicht die Existenz von $\sum_{i=0}^{\infty} a_i$, wie das Beispiel der Folge $(a_n)_{n\in\mathbf{N}} = \left(1/n\right)_{n\in\mathbf{N}}$ zeigt: die Folge $\left(1/n\right)_{n\in\mathbf{N}}$ konvergiert gegen 0, aber die Folge der n-ten Partialsummen $(s_n)_{n\in\mathbf{N}} = \left(\sum_{i\_1}^{n} 1/i\right)$ konvergiert nicht (Satz 5.1-5 (vi)).

Der folgende technische Satz ergibt sich unmittelbar aus den Sätzen zur Arithmetik von Grenzwerten:

Satz 5.1-7:

Es seien $\sum_{i=0}^{\infty} a_i$ und $\sum_{i=0}^{\infty} b_i$ konvergent. Dann gilt:

(i) $\sum_{i=0}^{\infty} (c \cdot a_i) = c \cdot \sum_{i=0}^{\infty} a_i$ für jedes $c \in \mathbf{R}$.

(ii) $\sum_{i=0}^{\infty} (a_i \pm b_i) = \sum_{i=0}^{\infty} a_i \pm \sum_{i=0}^{\infty} b_i$.

Zur **Berechnung** von $\sum_{i=0}^{\infty} a_i$ sind folgende Schritte erforderlich:

1. Schritt:

Man bildet die *n*-te Partialsumme $s_n = \sum_{i=0}^{n} a_i$. Falls möglich, findet man hierfür einen geschlossenen Ausdruck, der von *n* abhängt.

2. *Schritt:*

Man vollzieht den Grenzübergang $n \to \infty$ und erhält $\lim_{n\to\infty} s_n = \lim_{n\to\infty} \left(\sum_{i=0}^{n} a_i \right) = \sum_{i=0}^{\infty} a_i$.

Es sei $m \in \mathbf{N}$. Unter $\sum_{i=m}^{\infty} a_i$ versteht man den Grenzwert der Folge $\left(\sum_{i=m}^{n} a_i \right)_{n \in \mathbf{N}, n \geq m}$.

Satz 5.1-8:

Falls $\sum_{i=0}^{\infty} a_i$ existiert, so existiert auch $\sum_{i=m}^{\infty} a_i$ für jedes $m \in \mathbf{N}$, und es gilt

$$\sum_{i=m}^{\infty} a_i = \sum_{i=0}^{\infty} a_i - \sum_{i=0}^{m-1} a_i .$$

Der folgende Satz liefert einige Beispiele. Gerade Teil (iii) zeigt dabei, dass ähnlich erscheinende Reihen ganz unterschiedliches Konvergenzverhalten aufweisen.

Satz 5.1-9:

(i) Für $q \in \mathbf{R}$ mit $-1 < q < 1$ ist

$$\sum_{i=0}^{\infty} q^i = \frac{1}{1-q},$$

$$\sum_{i=1}^{\infty} q^i = \frac{q}{1-q},$$

$$\sum_{i=0}^{\infty} i \cdot q^i = \frac{q}{(1-q)^2},$$

$$\sum_{i=0}^{\infty} i^2 \cdot q^i = \frac{q \cdot (1+q)}{(1-q)^3}.$$

(ii) Die Reihen $\sum_{i=0}^{\infty} (-1)^i$ und $\sum_{i=1}^{\infty} \frac{1}{i}$ existieren nicht (sind divergent).

(iii) $$\sum_{i=1}^{\infty} (-1)^{i-1} \cdot \frac{1}{i} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + (-1)^{n-1} \cdot \frac{1}{n} \pm \ldots = \ln(2)$$

$$\approx 0{,}6931471805599$$

$$\sum_{i=1}^{\infty} (-1)^{i-1} \cdot \frac{1}{2i-1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ \ldots \ + (-1)^{n-1} \cdot \frac{1}{2n-1} \pm \ \ldots \ = \frac{\pi}{4}$$

$$\approx \ 0{,}7853981633974$$

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \ \ldots \ \frac{1}{n^2} + \ \ldots \ = \frac{\pi^2}{6} \approx \ 1{,}644934066848$$

$$\sum_{i=1}^{\infty} \frac{1}{i \cdot (i+1)} = \sum_{i=2}^{\infty} \frac{1}{i \cdot (i-1)} = 1.$$

(iv) $\sum_{i=1}^{\infty} \frac{1}{i^{\alpha}}$ konvergiert für jedes $\alpha \in \mathbf{R}$ mit $\alpha > 1$.

Teil (i) folgt direkt aus den Ergebnissen aus Satz 1.6-2 und den Rechenregeln mit Grenzwerten in Satz 5.1-2. Beispielsweise ist

$\sum_{i=0}^{\infty} i \cdot q^i = \lim_{n\to\infty} \frac{q-(n+1)\cdot q^{n+1}+n\cdot q^{n+2}}{(1-q)^2} = \frac{q}{(1-q)^2}$, da die Teile $(n+1)\cdot q^{n+1}$ und $n\cdot q^{n+2}$ jeweils gegen 0 konvergieren.

In Teil (ii) lautet die n-te Partialsumme $\sum_{i=0}^{n}(-1)^i = \begin{cases} 1 & \text{für gerades } n \\ 0 & \text{für ungerades } n \end{cases}$. Die Folge der Partialsummen hat also zwei unterschiedliche Häufungspunkte und ist daher nicht konvergent.

Der Wert der ersten Reihe in Teil (iii) wird in Kapitel 5.9 berechnet. Zur Herleitung der zweiten und dritten Reihe in Teil (iii) und zu der Aussage inTeil (iv) werden teilweise Hilfsmittel der Mathematik benötigt, die hier nicht behandelt werden. Die 2. Aussage in Teil (iii) folgt unmittelbar aus Satz 1.6-2 (iii).

Die rationalen Zahlen werden in Kapitel 1.4 durch

$$\mathbf{Q} = \left\{ m/n \mid m \in \mathbf{Z} \text{ und } n \in \mathbf{Z} \text{ und } n \neq 0 \right\}$$

definiert. Im Folgenden wird eine Charakterisierung mit Hilfe ihrer Dezimalbruchentwicklung gegeben.

Es sei $r \in \mathbf{R}$ eine reelle Zahl mit $0 \le r < 1$, deren Dezimalbruchentwicklung nach endlich vielen Stellen nur noch aus den Ziffern 0 besteht, d.h. nach endlich vielen Stellen abbricht:

$r = [0, d_{-1}d_{-2}\ldots d_{-m}]_{10} = \sum_{i=1}^{m} d_{-i}\cdot 10^{-i}$ mit $d_{-i} \in \{0,1,2,3,4,5,6,7,8,9\}$ für $i = 1,\ldots,m$.

Da jeder Summand $d_{-i}\cdot 10^{-i} = \frac{d_{-i}}{10^i}$ eine rationale Zahl ist, gilt auch $r \in \mathbf{Q}$. In diesem Fall ist $r = [0, d_{-1}d_{-2}\ldots d_{-m}]_{10} = \frac{[d_{-1}d_{-2}\ldots d_{-m}]_{10}}{10^m}$ (im Zähler steht die natürliche Zahl, deren Dezimalzahldarstellung aus der Ziffernfolge der Dezimalbruchentwicklung besteht, und im Nenner steht die Zahl, deren Dezimaldarstellung von links gelesen aus einer Ziffer 1, gefolgt von m Ziffern 0 besteht).

Es sei nun $r \in \mathbf{R}$ eine reelle Zahl mit $0 \le r < 1$, deren Dezimalbruchentwicklung aus einem nichtperiodischen und einem periodischen Teil besteht, etwa

$$r = \left[0, d_{-1}d_{-2} \ldots d_{-m}\overline{d_{-m-1}d_{-m-2} \ldots d_{-m-k}} \ldots\right]_{10}.$$

Hierbei wiederholt sich die Ziffernfolge $d_{-m-1}d_{-m-2} \ldots d_{-m-k}$ beliebig oft, und es kommen keine nichtperiodischen Abschnitte in der Dezimalziffernfolge mehr vor. Zu beachten ist, dass man mit der Periodennotation nur endlich viele Dezimalziffern notieren muss, obwohl die Zahl in ihrer Dezimaldarstellung unendlich viele Ziffern benötigt. Es ist

$$\begin{aligned} r &= \left[0, d_{-1}d_{-2} \ldots d_{-m}\overline{d_{-m-1}d_{-m-2} \ldots d_{-m-k}} \ldots\right]_{10} \\ &= \left[0, d_{-1}d_{-2} \ldots d_{-m}\right]_{10} + \left[0, \underbrace{00 \ldots 0}_{m-mal}\overline{d_{-m-1}d_{-m-2} \ldots d_{-m-k}} \ldots\right]_{10}. \end{aligned}$$

Der zweite Summand hat den Wert

$$\begin{aligned} &\left[0, \underbrace{00 \ldots 0}_{m-mal}\overline{d_{-m-1}d_{-m-2} \ldots d_{-m-k}} \ldots\right]_{10} \\ &\quad = 1/10^m \cdot \left[0, \overline{d_{-m-1}d_{-m-2} \ldots d_{-m-k}} \ldots\right]_{10} \\ &\quad = 1/10^m \cdot \sum_{i=1}^{\infty} \frac{\left[d_{-m-1}d_{-m-2} \ldots d_{-m-k}\right]_{10}}{\left(10^k\right)^i} \qquad \text{(im Zähler die Dezimalzahl } d_{-m-1}d_{-m-2} \ldots d_{-m-k}\text{)} \\ &\quad = 1/10^m \cdot \left[d_{-m-1}d_{-m-2} \ldots d_{-m-k}\right]_{10} \cdot \sum_{i=1}^{\infty} \frac{1}{\left(10^k\right)^i} \\ &\quad = 1/10^m \cdot \left[d_{-m-1}d_{-m-2} \ldots d_{-m-k}\right]_{10} \cdot \frac{\left(1/10\right)^k}{1-\left(1/10\right)^k} \qquad \text{(nach Satz 5.1-9 (i))} \\ &\quad = 1/10^m \cdot \left[d_{-m-1}d_{-m-2} \ldots d_{-m-k}\right]_{10} \cdot \frac{1}{10^k - 1}. \end{aligned}$$

Insgesamt ergibt sich

$$\begin{aligned} r &= \left[0, d_{-1}d_{-2} \ldots d_{-m}\overline{d_{-m-1}d_{-m-2} \ldots d_{-m-k}} \ldots\right]_{10} \\ &= \frac{\left[d_{-1}d_{-2} \ldots d_{-m}\right]_{10}}{10^m} + \frac{\left[d_{-m-1}d_{-m-2} \ldots d_{-m-k}\right]_{10}}{10^m \cdot \left(10^k - 1\right)}, \end{aligned}$$

(die Dezimaldarstellung der Zahl $10^m \cdot \left(10^k - 1\right)$ besteht von links gelesen aus k Ziffern 9, gefolgt von m Ziffern 0). Auch hier gilt wieder $r \in \mathbf{Q}$.

Bemerkung: Die Zahl $0,\overline{9}...$ hat den Wert 1; denn mit $m=1$ und $k=1$ ist

$$0,\overline{9}... = 0,9\overline{9}... = \frac{9}{10} + \frac{9}{10\cdot(10-1)} = 1 .$$

Es sei umgekehrt die rationale Zahl $r \in \mathbf{Q}$ mit $0 \le r < 1$ in gekürzter Form $r = \frac{a}{b}$ mit $a \in \mathbf{N}$ und $b \in \mathbf{N}$ mit $b \ge 1$ gegeben. Es ist $ggT(a,b) = 1$. Im Folgenden wird gezeigt, dass r

entweder als
endliche nichtperiodische Dezimalbruchentwicklung
oder als
rein periodische Dezimalbruchentwicklung
oder als
endliche nichtperiodische Ziffernfolge, gefolgt von einer endlichen periodischen Ziffernfolge

dargestellt werden kann. In den letzten beiden Fällen ist die Periodenlänge kleiner oder gleich $b-1$.

Für den Nachweis dieser Aussage werden drei Fälle unterschieden:

1. Fall: b hat die Form $b = 2^i \cdot 5^j$ mit $i + j = m \ge 1$.

Dann teilt b die Zahl $2^{i+j} \cdot 5^{i+j} = 10^{i+j} = 10^m$, d.h. $10^m = b \cdot b'$ mit einer natürlichen Zahl b'. Der Bruch $\frac{a}{b}$ lässt sich umschreiben in $\frac{a}{b} = \frac{a \cdot b'}{10^m}$. Da $r = \frac{a}{b} < 1$ gilt, ist $0 \le a \cdot b' < 10^m$. Die Darstellung von $a \cdot b'$ als Dezimalzahl hat endliche Länge und lautet $a \cdot b' = \sum_{l=0}^{m-1} d_l \cdot 10^l$ mit den Dezimalziffern d_0, ..., d_{m-1}; zu beachten ist hierbei, dass $a \cdot b'$ als Dezimalziffernfolge $a \cdot b' = [d_{m-1} \ldots d_0]_{10}$ lautet. Damit ergibt sich

$$\frac{a}{b} = \frac{a \cdot b'}{10^m} = \frac{\sum_{l=0}^{m-1} d_l \cdot 10^l}{10^m} = \sum_{l=0}^{m-1} d_l \cdot 10^{l-m} = \sum_{l=1}^{m} d_{m-l} \cdot 10^{-l} ,$$

also die endliche nichtperiodische Dezimalbruchentwicklung $\frac{a}{b} = [0, d_{m-1} \ldots d_0]_{10}$.

2. Fall: $ggT(b,10) = 1$

Nach Satz 3.4-2 gilt in diesem Fall $10^{\varphi(b)} \equiv 1 \pmod{b}$. Es gibt also eine kleinste natürliche Zahl $m \le \varphi(b)$, so dass b die Zahl $10^m - 1$ teilt. Wie im 1. Fall ist

$\frac{a}{b} = \frac{a \cdot b'}{10^m - 1}$ mit $b \cdot b' = 10^m - 1$. Wegen $\frac{a}{b} < 1$ ist $0 \le a \cdot b' < 10^m - 1$; die Darstellung von $a \cdot b'$ als Dezimalzahl hat wieder endliche Länge und lautet $a \cdot b' = \sum_{l=0}^{m-1} d_l \cdot 10^l$ mit den Dezimalziffern d_0, ..., d_{m-1}; zu beachten ist auch hier, dass $a \cdot b'$ als Dezimalziffernfolge $a \cdot b' = [d_{m-1} \dots d_0]_{10}$ lautet. Nach Satz 5.1-9 (i) gilt $\frac{1}{10^m - 1} = \sum_{i=1}^{\infty} \left((1/10)^m \right)^i$ und damit

$$\begin{aligned}
\frac{a}{b} &= \frac{a \cdot b'}{10^m - 1} = \left(\sum_{l=0}^{m-1} d_l \cdot 10^l \right) \cdot \left(\sum_{i=1}^{\infty} 10^{-m \cdot i} \right) \\
&= d_0 \cdot \left(\sum_{i=1}^{\infty} 10^{-m \cdot i} \right) + d_1 \cdot 10 \cdot \left(\sum_{i=1}^{\infty} 10^{-m \cdot i} \right) + \dots + d_{m-1} \cdot 10^{m-1} \cdot \left(\sum_{i=1}^{\infty} 10^{-m \cdot i} \right) \\
&= \left(\sum_{i=1}^{\infty} d_0 \cdot 10^{-m \cdot i} \right) + \left(\sum_{i=1}^{\infty} d_1 \cdot 10^{-m \cdot i + 1} \right) + \dots + \left(\sum_{i=1}^{\infty} d_{m-1} 10^{-m \cdot i + m - 1} \right) \\
&= \sum_{i=1}^{\infty} \left(d_0 \cdot 10^{-m \cdot i} + d_1 \cdot 10^{-m \cdot i + 1} + \dots + d_{m-1} 10^{-m \cdot i + m - 1} \right) \\
&= \sum_{i=1}^{\infty} \left(\sum_{l=0}^{m-1} d_l \cdot 10^{l - m \cdot i} \right) \\
&= \sum_{i=1}^{\infty} \left(\sum_{l=0}^{m-1} d_l \cdot 10^{l - m - (i-1) \cdot m} \right) \\
&= \sum_{i=1}^{\infty} 10^{-(i-1) \cdot m} \cdot \left(\sum_{l=0}^{m-1} d_l \cdot 10^{l-m} \right) \\
&= \sum_{i=1}^{\infty} 10^{-(i-1) \cdot m} \cdot \left(\sum_{l=1}^{m} d_{m-l} \cdot 10^{-l} \right).
\end{aligned}$$

Diese Zahl in der Darstellung als Dezimalbruchentwicklung lautet

$$\frac{a}{b} = \left[0, \underbrace{d_{m-1} \dots d_0}_{i=1} \underbrace{d_{m-1} \dots d_0}_{i=2} \underbrace{d_{m-1} \dots d_0}_{i=3} \dots \right]_{10} = \left[0, \overline{d_{m-1} \dots d_0} \dots \right]_{10},$$

ist also ein rein periodischer Dezimalbruch.

Offensichtlich ist (wegen $b \ge 3$) $m \le \varphi(b) \le b - 1$. Das bedeutet, dass die Periodenlänge *m* kleiner oder gleich $b - 1$ ist.

3. Fall: *b* hat die Form $b = 2^i \cdot 5^j \cdot b'$ mit $i + j = m \ge 1$, $b' > 1$ und $ggT(b', 10) = 1$. Wie im 1. Fall lässt sich der Bruch $\frac{a}{b}$ erweitern zu $\frac{a}{b} = \frac{a \cdot 2^j \cdot 5^i}{2^{i+j} \cdot 5^{i+j} \cdot b'} = \frac{a \cdot 2^j \cdot 5^i}{10^m \cdot b'}$ mit $2^{i+j} \cdot 5^{i+j} = 10^{i+j} = 10^m$. Gilt $\frac{a \cdot 2^j \cdot 5^i}{b'} \ge 1$, dann kann man den Zähler in der Form $a \cdot 2^j \cdot 5^i = k \cdot b' + a'$ mit $k \in \mathbf{N}$ und $0 \le a' = \left(a \cdot 2^j \cdot 5^i \right) \bmod b' < b'$ schreiben, d.h.

$\frac{a \cdot 2^j \cdot 5^i}{b'} = k + \frac{a'}{b'}$, $0 \le \frac{a'}{b'} < 1$. Ist $\frac{a \cdot 2^j \cdot 5^i}{b'} < 1$, dann hat der Bruch ebenfalls die Form $\frac{a \cdot 2^j \cdot 5^i}{b'} = k + \frac{a'}{b'}$ (mit $k = 0$ und $a' = a \cdot 2^j \cdot 5^i$). In beiden Fällen ergibt sich aus $ggT(a,b) = 1$ und $ggT(b',10) = 1$, dass auch $ggT(a',b') = 1$ ist; denn für $\frac{a \cdot 2^j \cdot 5^i}{b'} \ge 1$ folgt mit Satz 3.3-2:

$ggT(a',b') = ggT((a \cdot 2^j \cdot 5^i) \bmod b', b') = ggT(a \cdot 2^j \cdot 5^i, b') = 1$, und für $\frac{a \cdot 2^j \cdot 5^i}{b'} < 1$ ist bereits $ggT(a',b') = ggT(a \cdot 2^j \cdot 5^i, b') = 1$. Die Dezimalzahldarstellung von k sei $k = \sum_{l=0}^{h-1} d_l \cdot 10^l = [d_{h-1} \dots d_0]_{10}$, die Darstellung von $\frac{a'}{b'}$ als Dezimalbruch gemäß dem 2. Fall sei $\frac{a'}{b'} = [0, \overline{d'_{n-1} \dots d'_0} \dots]_{10}$ Dann hat $\frac{a \cdot 2^j \cdot 5^i}{b'} = k + \frac{a'}{b'}$ die Dezimaldarstellung $\frac{a \cdot 2^j \cdot 5^i}{b'} = k + \frac{a'}{b'} = [d_{h-1} \dots d_0, \overline{d'_{n-1} \dots d'_0} \dots]_{10}$. Die Darstellung von $\frac{a}{b}$ als Dezimalbruch enthält genau dieselbe Ziffernfolge, nur ist das Komma um m Stellen nach links geschoben, d.h. die Dezimalbruchdarstellung von $\frac{a}{b}$ besteht aus einer endlichen nichtperiodischen Ziffernfolge, gefolgt von einer endlichen periodischen Ziffernfolge. Auch hier ist die Periodenlänge kleiner oder gleich $b' - 1 \le b - 1$.

Die Betrachtung bezieht sich auf rationale Zahlen $r \in \mathbf{Q}$ mit $0 \le r < 1$. Sie ist natürlich auf ganz **Q** erweiterbar:

$$\mathbf{Q} = \{z + r \mid z \in \mathbf{Z} \text{ und } (r = [0, d_{-1} d_{-2} \dots d_{-m}]_{10} \text{ oder } r = [0, d_{-1} d_{-2} \dots d_{-m} \overline{d_{-m-1} d_{-m-2} \dots d_{-m-k}} \dots]_{10})\},$$

d.h. **Q** besteht aus allen Zahlen, deren gebrochener Anteil in Dezimaldarstellung entweder nach endlich vielen Dezimalziffern abbricht (es folgen nur noch Nullen) oder unendlich periodisch endet. Irrationalen Zahlen haben demzufolge einen gebrochenen Anteil in Dezimaldarstellung, der unendlich und nichtperiodisch ist.

Eine Reihe $\sum_{i=0}^{\infty} a_i$ heißt **absolut konvergent**, wenn die Reihe $\sum_{i=0}^{\infty} |a_i|$ konvergiert.

Für jedes $n \in \mathbf{N}$ und jedes $k \in \mathbf{N}$ ist $\left|\sum_{i=n+1}^{n+k} a_i\right| \le \sum_{i=n+1}^{n+k} |a_i|$. Hieraus ergibt sich mit Satz 5.1-6 (i) unmittelbar, dass jede absolut konvergente Reihe auch konvergiert, d.h. aus der Konvergenz von $\sum_{i=0}^{\infty} |a_i|$ folgt die Konvergenz von $\sum_{i=0}^{\infty} a_i$. Mit $n = 0$ und $k \to \infty$ ergibt sich $\left|\sum_{i=0}^{\infty} a_i\right| \le \sum_{i=0}^{\infty} |a_i|$.

Unter bestimmten Voraussetzungen kann man Reihen miteinander multiplizieren, jedenfalls dann, wenn sie absolut konvergent sind. Die folgende (nichtmathematische) Darstellung liefert die Motivation für die etwas komplizierte Indizierung in den auftretenden Reihen. Zwei Reihen $\sum_{i=0}^{\infty} a_i$ und $\sum_{i=0}^{\infty} b_i$ werden miteinander multipliziert, indem die einzelnen Summanden nach der Summe ihrer Indizes sortiert werden:

$$(a_0 + a_1 + a_2 + a_3 + \ldots a_n + \ldots) \cdot (b_0 + b_1 + b_2 + b_3 + \ldots b_n + \ldots)$$

$$= \underbrace{a_0 \cdot b_0}_{\text{Indexsumme } 0} + \underbrace{a_0 \cdot b_1 + a_1 \cdot b_0}_{\text{Indexsumme } 1} + \underbrace{a_0 \cdot b_2 + a_1 \cdot b_1 + a_2 \cdot b_0}_{\text{Indexsumme } 2} + \ldots + \underbrace{\sum_{j=0}^{n} a_j \cdot b_{n-j}}_{\text{Indexsumme } n} + \ldots$$

Satz 5.1-10:

Sind die Reihen $\sum_{i=0}^{\infty} a_i = a$ und $\sum_{i=0}^{\infty} b_i = b$ absolut konvergent, so ist auch die Reihe $\sum_{i=0}^{\infty} \left(\sum_{k=0}^{i} a_k \cdot b_{i-k} \right)$ absolut konvergent, und es gilt

$$\sum_{i=0}^{\infty} \left(\sum_{k=0}^{i} a_k \cdot b_{i-k} \right) = \left(\sum_{i=0}^{\infty} a_i \right) \cdot \left(\sum_{i=0}^{\infty} b_i \right) = a \cdot b \,.$$

Auf den technisch aufwendigen Beweis wird hier verzichtet.

Im allgemeinen ist die Bestimmung des Konvergenzverhaltens einer Reihe nicht einfach, so dass sich die Frage nach **Konvergenzkriterien für Reihen** stellt. Ein „Negativkriterium“ liefert Satz 5.1-6: Falls für eine Reihe $\sum_{i=0}^{\infty} a_i$ die Folge $(a_n)_{n \in \mathbf{N}}$ *nicht* gegen 0 konvergiert, existiert auch der Grenzwert $\sum_{i=0}^{\infty} a_i$ nicht. Zwei „Positivkriterien“ für absolute Konvergenz sind im folgenden Satz zusammengefasst.

Satz 5.1-11:

(i) (**Majorantenkriterium**)

Ist $\sum_{i=0}^{\infty} b_i$ absolut konvergent und gilt $|a_i| \le |b_i|$ für (fast) alle $i \in \mathbf{N}$, so ist auch die Reihe $\sum_{i=0}^{\infty} a_i$ absolut konvergent.

(ii) (**Quotientenkriterium**)

Die Reihe $\sum_{i=0}^{\infty} a_i$ besitze die Eigenschaft, dass ab einem Index $n_0 \in \mathbf{N}$ stets $\left|\frac{a_{n+1}}{a_n}\right| \le q$ für einen Wert q mit $0 < q < 1$ gilt. Dann ist die Reihe $\sum_{i=0}^{\infty} a_i$ absolut konvergent. Ist ab einem Index stets $\left|\frac{a_{n+1}}{a_n}\right| > 1$, so ist die Reihe divergent.

Es sei $B = \sum_{i=0}^{\infty} |b_i|$. In Aussage (i) wird die Existenz von B vorausgesetzt. Damit ergibt sich $\sum_{i=0}^{n} |a_i| \le \sum_{i=0}^{n} |b_i| \le B$, d.h. die Folge $\left(\sum_{i=0}^{n} |a_i|\right)_{n \in \mathbf{N}}$ ist monoton wachsend und nach oben beschränkt. Nach Satz 5.1-3 (ii) konvergiert daher $\sum_{i=0}^{\infty} a_i$ absolut.

In Aussage (ii) gelte $\left|\frac{a_{n+1}}{a_n}\right| \le q$ für $n \ge n_0$, also $|a_{n+1}| \le q \cdot |a_n| \le \ldots \le q^{n-n_0+1} \cdot |a_{n_0}|$. Für $n \ge n_0 + 1$ ist daher $|a_n| \le q^n \cdot |a_{n_0}| \cdot q^{-n_0}$. Die Reihe $\sum_{i=0}^{\infty} \left(q^i \cdot |a_{n_0}| \cdot q^{-n_0}\right)$ konvergiert nach Satz 5.1-9 (i) absolut und damit nach (i) auch $\sum_{i=0}^{\infty} a_i$.

Es sei $x \in \mathbf{R}$. Die Reihe $\sum_{i=0}^{\infty} \frac{x^i}{i!}$ ist absolut konvergent; denn nach Satz 5.1-11 (ii) ist $\left|\frac{x^{n+1}/(n+1)!}{x^n/n!}\right| = \left|\frac{x}{n+1}\right| = \frac{|x|}{n+1} < 1/2$ für jedes $n \in \mathbf{N}$ mit $n > 2 \cdot |x| - 1$. Daher ist die Definition der folgenden Funktion sinnvoll.

Die Funktion

$$\exp : \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & \sum_{i=0}^{\infty} \frac{x^i}{i!} \end{cases}$$

heißt **Exponentialfunktion**. Wichtige Eigenschaften der Exponentialfunktion und verwandter Funktionen werden in Kapitel 5.5 behandelt.

In Satz 5.1-5 (iv) wurde die Eulersche Konstante als $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e = 2.7182818284590\ldots$ definiert. Wie man direkt nachrechnet, gilt $\binom{n}{k}\cdot\frac{1}{n^k} \le \frac{1}{k!}$. Daher ist (siehe Kapitel 4.1)

$$\left(1+\frac{1}{n}\right)^n = \sum_{k=0}^{n}\binom{n}{k}\cdot\frac{1}{n^k} \le \sum_{k=0}^{n}\frac{1}{k!} .$$

Daraus folgt $e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \le \lim_{n\to\infty}\left(\sum_{k=0}^{n}\frac{1}{k!}\right) = \sum_{k=0}^{\infty}\frac{1}{k!}$. Es sei umgekehrt $n > m \ge 1$. Dann ist

$$\left(1+\frac{1}{n}\right)^n > \sum_{k=0}^{m}\binom{n}{k}\cdot\frac{1}{n^k} = \sum_{k=0}^{m}\left(\frac{1}{k!}\cdot\frac{n}{n}\cdot\frac{n-1}{n}\cdot\ldots\cdot\frac{n-k+1}{n}\right) = \sum_{k=0}^{m}\left(\frac{1}{k!}\cdot 1\cdot\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\right).$$

Durch Limesbildung folgt $e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \ge \sum_{k=0}^{m}\frac{1}{k!}$ und durch erneute Limesbildung $e \ge \lim_{m\to\infty}\sum_{k=0}^{m}\frac{1}{k!} = \sum_{k=0}^{\infty}\frac{1}{k!}$. Insgesamt ergibt sich

$$e = \sum_{i=0}^{\infty}\frac{1}{i!} = \exp(1) .$$

Es sei $x \in \mathbf{R}$. Die obige Abschätzung lässt sich auf $\left(1+\frac{x}{n}\right)^n$ übertragen:

$$\sum_{k=0}^{m}\left(\frac{x^k}{k!}\cdot 1\cdot\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\right) < \left(1+\frac{x}{n}\right)^n \le \sum_{k=0}^{n}\frac{x^k}{k!} \text{ und}$$

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \sum_{k=0}^{\infty}\frac{x^k}{k!} = \exp(x).$$

Es lässt sich nun leicht zeigen, dass *e* keine rationale Zahl ist. Denn angenommen, *e* ließe sich als Bruch $e = p/q$ mit $p \in \mathbf{N}_{>0}$ und $q \in \mathbf{N}_{>0}$, dann folgt $e = \sum_{i=0}^{\infty} \frac{1}{i!} = \sum_{i=0}^{q} \frac{1}{i!} + \sum_{i=q+1}^{\infty} \frac{1}{i!}$ und $q!\left(e - \sum_{i=0}^{q} \frac{1}{i!}\right) = \sum_{i=q+1}^{\infty} \frac{q!}{i!} = \sum_{i=1}^{\infty} \frac{q!}{(q+i)!}$. Die linke Seite der letzten Gleichung ist gleich $p \cdot (q-1)! - \left(\frac{q!}{0!} + \frac{q!}{1!} + \ldots + \frac{q!}{q!}\right)$, also eine ganze Zahl. Da die rechte Seite positiv ist, ist die linke Seite aus $\mathbf{N}_{>0}$. Für die Summanden der rechten Seite gilt $\frac{q!}{(q+i)!} = \frac{1}{(q+1)\cdot \ldots \cdot (q+i)} < \left(\frac{1}{2}\right)^i$, also $0 < \sum_{i=q+1}^{\infty} \frac{q!}{i!} < \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i = \frac{1}{1-1/2} - 1 = 1$. Dieses Ergebnis widerspricht der Folgerung, dass der Wert aus aus $\mathbf{N}_{>0}$ ist. Daher ist *e* irrational.

Die Exponentialfunktion, die durch die Reihe $\sum_{i=0}^{\infty} \frac{x^i}{i!}$ definiert wird, kann man durch eine endliche Summe und einen Restfehlerterm darstellen, dessen Größe man abschätzen kann:

Satz 5.1-12:

Es sei $n \in \mathbf{N}$ und $x \in \mathbf{R}$ mit $|x| \le 1 + \frac{n}{2}$. Dann gilt mit $R_{n+1}(x) = \sum_{i=n+1}^{\infty} \frac{x^i}{i!}$:

$$\exp(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!} = \sum_{i=0}^{n} \frac{x^i}{i!} + R_{n+1}(x) \text{ und } |R_{n+1}(x)| \le \frac{2 \cdot |x|^{n+1}}{(n+1)!}.$$

Die Abschätzung des Restterms sieht man wie folgt:

Wegen $|x| \le 1 + \frac{n}{2} = \frac{n+2}{2}$ ist $\frac{|x|}{n+2+i} \le \frac{|x|}{n+2} \le \frac{1}{2}$ für jedes $i \ge 0$.

$$|R_{n+1}(x)| \le \sum_{i=n+1}^{\infty} \frac{|x|^i}{i!} = \frac{|x|^{n+1}}{(n+1)!} \cdot \sum_{i=n+1}^{\infty} \frac{|x|^{i-(n+1)} \cdot (n+1)!}{i!} = \frac{|x|^{n+1}}{(n+1)!} \cdot \left(\frac{|x|^0}{1} + \frac{|x|^1}{n+2} + \frac{|x|^2}{(n+2)\cdot(n+3)} + \ldots\right)$$

$$\le \frac{|x|^{n+1}}{(n+1)!} \cdot \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^i = \frac{2 \cdot |x|^{n+1}}{(n+1)!}.$$

Für $x = 1$ und $n = 10$ liefert Satz 5.1-12 die Abschätzung $R_{11}(1) < 0{,}00000006$ und damit 2,71828180 < *e* < 2,71828186.

Der folgende Satz gibt ein Konvergenzkriterium für Reihen mit **alternierenden Folgengliedern**, d.h. jeweils aufeinanderfolgende Folgenglieder wechseln ihr Vorzeichen. Außerdem wird eine Abschätzung des Restterms gegeben, wenn die Reihe nach dem *n*-ten Folgenglied abgebrochen wird.

Satz 5.1-13 (Leibnizkriterium):

Es sei $(a_n)_{n\in\mathbf{N}}$ eine Folge reeller Zahlen mit $a_n > 0$ und $a_{n+1} \le a_n$ für alle $n \in \mathbf{N}$ und mit $\lim\limits_{n\to\infty} a_n = 0$. Dann konvergiert die Reihe $\sum\limits_{i=0}^{\infty}(-1)^i \cdot a_i$, und es gilt

$$\sum_{i=n+1}^{\infty}(-1)^i \cdot a_i = \lambda_n \cdot (-1)^{n+1} a_{n+1} \text{ mit } 0 < \lambda_n \le 1.$$

Für die Folge der Partialsummen mit geraden Index $(s_{2\cdot n})_{n\in\mathbf{N}}$ gilt

$$s_{2\cdot n} = \sum_{i=0}^{2\cdot n}(-1)^i \cdot a_i = a_0 - a_1 + a_2 - a_3 + \ldots + a_{2\cdot n-2} - a_{2\cdot n-1} + a_{2\cdot n}$$
$$= (a_0 - a_1) + (a_2 - a_3) + \ldots + (a_{2\cdot n-2} - a_{2\cdot n-1}) + a_{2\cdot n} \ge a_{2\cdot n} > 0 \quad \text{und}$$

$s_{2\cdot n+2} - s_{2\cdot n} = a_{2\cdot n+2} - a_{2\cdot n+1} \le 0$, d.h. $s_{2\cdot n+2} \le s_{2\cdot n}$. Für die Folge der Partialsummen mit ungeraden Index $(s_{2\cdot n+1})_{n\in\mathbf{N}}$ gilt entsprechend

$$s_{2\cdot n+1} = \sum_{i=0}^{2\cdot n+1}(-1)^i \cdot a_i = a_0 - a_1 + a_2 - a_3 + \ldots + a_{2\cdot n} - a_{2\cdot n+1}$$
$$= (a_0 - a_1) + (a_2 - a_3) + \ldots + (a_{2\cdot n} - a_{2\cdot n+1}) \ge 0 \quad \text{und}$$

$s_{2\cdot n+3} - s_{2\cdot n+1} = -a_{2\cdot n+3} + a_{2\cdot n+2} \ge 0$, d.h. $s_{2\cdot n+3} \ge s_{2\cdot n+1}$.

Außerdem ist $0 < s_{2\cdot n+1} = s_{2\cdot n} - a_{2\cdot n+1} \le s_{2\cdot n} = a_0 + (-a_1 + a_2) - \ldots + (-a_{2\cdot n-1} + a_{2\cdot n}) \le a_0$. Nach Satz 5.1-3 konvergieren $(s_{2\cdot n})_{n\in\mathbf{N}}$ und $(s_{2\cdot n+1})_{n\in\mathbf{N}}$. Es ist

$\lim\limits_{n\to\infty}(s_{2\cdot n+1} - s_{2\cdot n}) = \lim\limits_{n\to\infty}(-a_{2\cdot n+1}) = -\lim\limits_{n\to\infty}(a_{2\cdot n+1}) = 0$, also ist $\lim\limits_{n\to\infty} s_{2\cdot n+1} = \lim\limits_{n\to\infty} s_{2\cdot n}$, und damit existiert $S = \sum\limits_{i=0}^{\infty}(-1)^i \cdot a_i = \lim\limits_{n\to\infty} s_n$, und es gilt $0 < S \le a_0$, also $S = \lambda \cdot a_0$ mit $0 < \lambda \le 1$.

Der Restterm $\sum\limits_{i=n+1}^{\infty}(-1)^i \cdot a_i$ ist nichts anderes als eine Reihe mit alternierenden Folgengliedern, die mit dem Index $n+1$ beginnt: $\sum\limits_{i=n+1}^{\infty}(-1)^i \cdot a_i = \sum\limits_{i=0}^{\infty} a_i'$ mit $a_i' = (-1)^{i+n+1} a_{i+n+1}$. Die Überlegungen gelten auch für diese Reihe, so dass $\sum\limits_{i=n+1}^{\infty}(-1)^i \cdot a_i = \sum\limits_{i=0}^{\infty} a_i' = \lambda_n \cdot a_0' = \lambda_n \cdot (-1)^{n+1} \cdot a_{n+1}$ mit $0 < \lambda_n \le 1$ gilt.

5.2 Eigenschaften reeller Funktionen einer Veränderlichen

Im vorliegenden Kapitel werden eine Reihe wichtiger Definitionen zusammengestellt, die Eigenschaften reeller Funktionen einer Veränderlichen beschreiben.

Im Folgenden sei $f: X \to \mathbf{R}$, $X \subseteq \mathbf{R}$, und $I \subseteq \mathbf{R}$ sei ein Intervall (siehe Kapitel 1.7).

f heißt **auf *I* monoton steigend** (bzw. **monoton fallend)**, wenn für $x_1 \in I$ und $x_2 \in I$ gilt:

Ist $x_1 < x_2$, so ist $f(x_1) \leq f(x_2)$
(bzw.
ist $x_1 < x_2$, so ist $f(x_1) \geq f(x_2)$).

Der Graph einer monoton steigenden Funktion fällt also mit wachsenden *x*-Werten nicht ab; der Graph einer monoton fallenden Funktion steigt also mit wachsenden *x*-Werten nicht.

f heißt **auf *I* streng monoton steigend** (bzw. **streng monoton fallend**), wenn für $x_1 \in I$ und $x_2 \in I$ gilt:

Ist $x_1 < x_2$, so ist $f(x_1) < f(x_2)$
(bzw.
ist $x_1 < x_2$, so ist $f(x_1) > f(x_2)$).

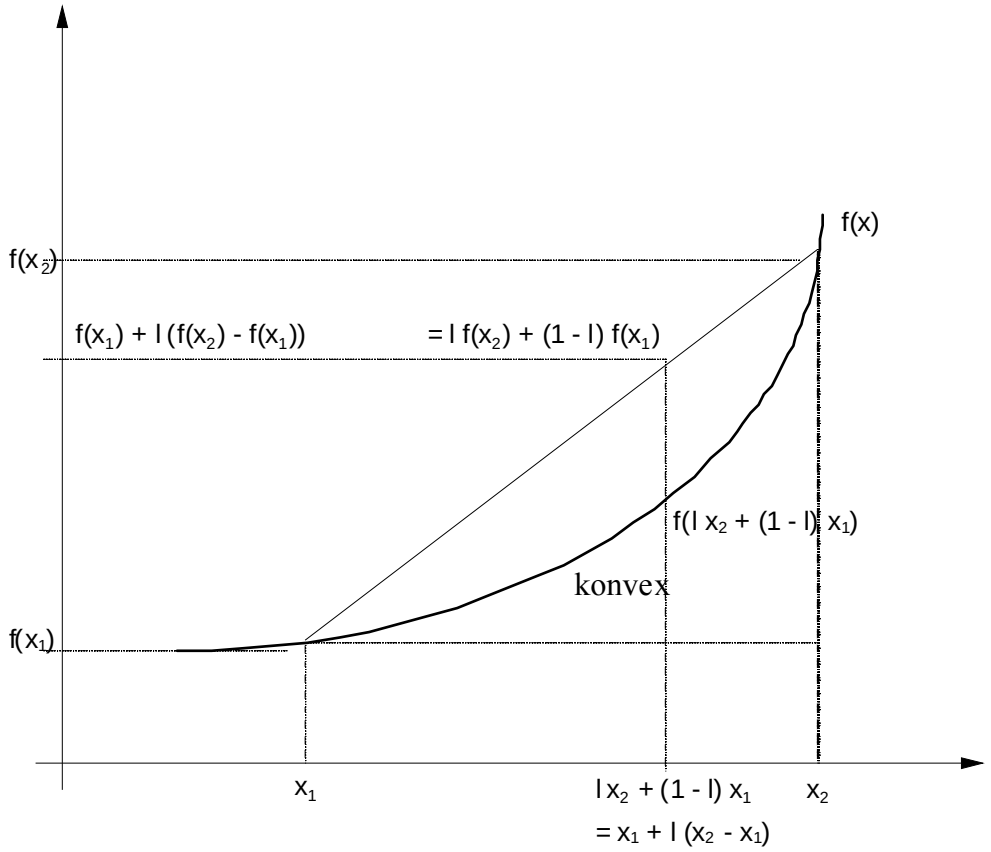
f heißt **auf *I* beschränkt**, wenn es ein $c \in \mathbf{R}_{\geq 0}$ gibt, so dass für jedes $x \in I$ gilt: $|f(x)| \leq c$.

Der Graph einer beschränkten Funktion verläuft also weder oberhalb von *c* noch unterhalb von $-c$.

f heißt **auf *I* nach oben beschränkt** (bzw. **nach unten beschränkt**), wenn es ein $c \in \mathbf{R}$ gibt, so dass für jedes $x \in I$ gilt: $f(x) \leq c$ bzw. $f(x) \geq c$.

f heißt **konvex über *I***, wenn für $x_1 \in I$ und $x_2 \in I$ mit $x_1 \neq x_2$ und für jedes $l \in \mathbf{R}$ mit $0 < l < 1$ gilt:

$$f(l \cdot x_2 + (1-l) \cdot x_1) \leq l \cdot f(x_2) + (1-l) \cdot f(x_1).$$



Nimmt man also zwei beliebige verschiedene Werte $x_1 \in I$ und $x_2 \in I$ und verbindet die Punkte $(x_1, f(x_1))$ und $(x_2, f(x_2))$ des Graphen einer über *I* konvexen Funktion durch eine gerade Linie, so verläuft der Graph zwischen $(x_1, f(x_1))$ und $(x_2, f(x_2))$ unterhalb dieser Verbindungslinie. Betrachtet man diese Verbindungslinie als Annäherung an den Graphen der Funktion zwischen $(x_1, f(x_1))$ und $(x_2, f(x_2))$, so macht man einen Approximationsfehler in Richtung größerer Werte, d.h. die Approximation liefert zu große Werte.

f heißt **konkav über *I***, wenn für $x_1 \in I$ und $x_2 \in I$ mit $x_1 \neq x_2$ und für jedes $l \in \mathbf{R}$ mit $0 < l < 1$ gilt:

$$f(l \cdot x_2 + (1-l) \cdot x_1) \geq l \cdot f(x_2) + (1-l) \cdot f(x_1).$$

Bei einer konkaven Funktion verläuft der Graph oberhalb der entsprechenden Verbindungslinie. Betrachtet man auch hier wieder die Verbindungslinie zwischen den Punkten $(x_1, f(x_1))$

und $(x_2, f(x_2))$ als Annäherung an den Graphen der Funktion, so liefert sie hier zu kleine Werte.

f heißt **streng konvex über *I***, wenn für $x_1 \in I$ und $x_2 \in I$ mit $x_1 \neq x_2$ und für jedes $l \in \mathbf{R}$ mit $0 < l < 1$ gilt:

$$f(l \cdot x_2 + (1-l) \cdot x_1) < l \cdot f(x_2) + (1-l) \cdot f(x_1).$$

f heißt **streng konkav über *I***, wenn für $x_1 \in I$ und $x_2 \in I$ mit $x_1 \neq x_2$ und für jedes $l \in \mathbf{R}$ mit $0 < l < 1$ gilt:

$$f(l \cdot x_2 + (1-l) \cdot x_1) > l \cdot f(x_2) + (1-l) \cdot f(x_1).$$

Eine Funktion kann in Teilintervallen ihres Definitionsbereichs (streng) konvex und in anderen Teilintervallen (streng) konkav sein.

Die Funktion $f: X \to \mathbf{R}$, $X \subseteq \mathbf{R}$ heißt **stetig im Punkt** $x_0 \in X$, wenn gilt:
Für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ gibt es ein $\delta > 0$, das von ε und x_0 abhängen kann (d.h. $\delta = \delta(\varepsilon, x_0)$), mit folgender Eigenschaft:

Für jedes $x \in X$ mit $|x - x_0| < \delta$ ist $|f(x) - f(x_0)| < \varepsilon$.

Die Funktion $f: X \to \mathbf{R}$ heißt **stetig in** $D \subseteq X$, wenn f in jedem Punkt $x_0 \in D$ stetig ist.

Für einen Wert $x \in \mathbf{R}$ und ein $\varepsilon > 0$ bezeichnet man das offene Intervall

$$U_\varepsilon(x) = \{ z \mid x - \varepsilon < z < x + \varepsilon \} = \{ z \mid |x - z| < \varepsilon \}$$

als **ε-Umgebung** von *x*. Mit dieser Bezeichnung bedeutet die Stetigkeit einer Funktion $f: X \to \mathbf{R}$ in einem Punkt $x_0 \in X$:

Zu jeder ε-Umgebung $U(f(x_0), \varepsilon)$ von $f(x_0)$ gibt es eine (von ε und x_0 abhängige) δ-Umgebung $U(x_0, \delta)$ von x_0, die durch f ganz nach $U(f(x_0), \varepsilon)$ abgebildet wird, d.h. für die

$$f(U(x_0,\delta)) \subseteq U(f(x_0),\varepsilon)$$

gilt. Anschaulich heißt dieses, dass für ein Argument *x*, das sich „nahe bei“ x_0 befindet (in der δ-Umgebung $U(x_0,\delta)$ von x_0), der Funktionswert $f(x)$ „nahe bei“ $f(x_0)$ liegt (in der ε-Umgebung $U(f(x_0),\varepsilon)$ von $f(x_0)$). Eine „sehr kleine Änderung“ des Arguments, d.h. der Übergang von x_0 zu *x* mit $|x-x_0|<\delta$, führt zu einer „sehr kleinen Änderung“ von $f(x_0)$, d.h. der Funktionswert $f(x)$ erfüllt $|f(x)-f(x_0)|<\varepsilon$. Insbesondere macht der Graph der Funktion an der Stelle bzw. „nahe“ der Stelle x_0 keinen Sprung. Graphen stetiger Funktionen lassen sich in einem Zuge zeichnen, ohne den Zeichenstift abzusetzen. Der Graph einer in $x_0 \in X$ stetigen Funktion weist in $(x_0, f(x_0))$ ***keine Sprungstelle*** auf.

Ändert sich die Funktion *f* in der Nähe von x_0 langsam, so wird man keine Mühe haben, zu vorgegebenem $\varepsilon > 0$ ein passendes $\delta > 0$ zu finden; ändert sie sich rasch, so wird man δ entsprechend klein wählen müssen.

Beispiele:

Die Funktion

$$f: \begin{cases} \mathbf{R}_{>0} & \to & \mathbf{R} \\ x & \to & 1/x \end{cases}$$

ist stetig in jedem Punkt $x_0 \in \mathbf{R}_{>0}$: Zu $\varepsilon > 0$ kann man $\delta = \delta(\varepsilon, x_0) = \dfrac{\varepsilon \cdot x_0^2}{1+\varepsilon \cdot x_0} > 0$ nehmen. Ist nämlich $x \in X$ mit $|x-x_0|<\delta$, so ist

$$|f(x)-f(x_0)| = |1/x - 1/x_0| = \left|\frac{x_0-x}{x\cdot x_0}\right| < \frac{\delta}{x\cdot x_0} < \frac{\delta}{x_0\cdot(x_0-\delta)}.$$

Die letzte Ungleichung ergibt sich aus der Annahme $|x-x_0|<\delta$, die gleichbedeutend ist mit $x_0-\delta < x < x_0+\delta$ (also gilt insbesondere $x_0-\delta < x$ bzw. $1/x < 1/(x_0-\delta)$). Setzt man $\delta = \dfrac{\varepsilon \cdot x_0^2}{1+\varepsilon \cdot x_0}$ ein, so sieht man $\dfrac{\delta}{x_0\cdot(x_0-\delta)} = \varepsilon$, also insgesamt $|f(x)-f(x_0)|<\varepsilon$.

Die Funktion

$$f: \begin{cases} \mathbf{R}_{\geq 0} & \to & \mathbf{R} \\ x & \to & \sqrt{x} \end{cases}$$

ist stetig in jedem Punkt $x_0 \in \mathbf{R}_{\geq 0}$. Zu $\varepsilon > 0$ und $x_0 \geq 0$ wähle man z.B. $\delta = \delta(\varepsilon, x_0) = \varepsilon^2$. Man beachte, dass δ hier nur von ε und nicht von x_0 abhängt. Ist nämlich $x \in X$ mit $|x - x_0| < \delta$ und $x \neq x_0$ (für $x = x_0$ gilt sowieso $|f(x) - f(x_0)| = 0 < \varepsilon$):

$$|f(x) - f(x_0)| = \left|\sqrt{x} - \sqrt{x_0}\right| = \left|\frac{\left(\sqrt{x} - \sqrt{x_0}\right) \cdot \left(\sqrt{x} + \sqrt{x_0}\right)}{\left(\sqrt{x} + \sqrt{x_0}\right)}\right| = \frac{|x - x_0|}{\sqrt{x} + \sqrt{x_0}} \leq \frac{|x - x_0|}{\sqrt{|x - x_0|}}.$$

Die letzte Ungleichung folgt aus der binomischen Formel: Sind $a \in \mathbf{R}_{\geq 0}$ und $b \in \mathbf{R}_{\geq 0}$ reelle Zahlen, so gilt:

$$a + b \leq a + b + 2 \cdot \sqrt{a} \cdot \sqrt{b} = \left(\sqrt{a} + \sqrt{b}\right)^2;$$

außerdem gilt $|a - b| \leq a + b$ und damit $\sqrt{|a - b|} \leq \sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$.

Die obige Ungleichung wird fortgesetzt:

$$|f(x) - f(x_0)| \leq \frac{|x - x_0|}{\sqrt{|x - x_0|}} = \sqrt{|x - x_0|} < \sqrt{\delta} = \sqrt{\varepsilon^2} = \varepsilon.$$

Wie im Fall der Konvergenz ist das zu vorgegebenem $\varepsilon > 0$ „passende“ $\delta > 0$ nicht immer leicht zu finden; hier ist häufig mathematische Phantasie gefragt. Man kann beispielsweise die konkrete Angabe von δ zunächst offen lassen und versuchen, die Ungleichung $|f(x) - f(x_0)| < \varepsilon$ so umzuformen, dass dort der Ausdruck $|x - x_0|$ vorkommt, von dem man dann ja annimmt, dass er kleiner als δ ist. Dann versucht man, die so entstandene Ungleichung nach δ aufzulösen. Das folgende Beispiel soll die Vorgehensweise erläutern.

Die Funktion

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

ist in jedem Punkt $x_0 \in \mathbf{R}$ stetig. Zu $\varepsilon > 0$ und $x_0 \in \mathbf{R}$ setzt man beispielsweise $\delta = \delta(\varepsilon, x_0) = ?$ Es wird zunächst $|f(x) - f(x_0)|$ in Abhängigkeit von $|x - x_0|$ und x_0 bestimmt:

$$|f(x) - f(x_0)| = |x^2 - x_0^2| = |(x - x_0) \cdot (x + x_0)| = |x - x_0| \cdot |x + x_0|$$
$$\leq |x - x_0| \cdot (|x| + |x_0|) < \delta \cdot (2 \cdot |x_0| + \delta).$$

In der letzten Ungleichung wurden die später zu treffende Annahme $|x - x_0| < \delta$ und die aus dieser Ungleichung leicht nachzurechnende Folgerung $|x| < |x_0| + \delta$ bereits verwendet. Wie ist also δ zu wählen, damit $\delta \cdot (2 \cdot |x_0| + \delta) < \varepsilon$ ist (hier reicht auch „=“ anstelle von „<“, da ja in der Ungleichungskette bereits „<“ vorkommt)?

$$\delta \cdot (2 \cdot |x_0| + \delta) = \delta^2 + 2 \cdot \delta \cdot |x_0| + |x_0|^2 - |x_0|^2 = (\delta + |x_0|)^2 - |x_0|^2 = \varepsilon$$

ergibt

$$\delta = \delta(\varepsilon, x_0) = \sqrt{|x_0|^2 + \varepsilon} - |x_0|.$$

Da der Ausdruck unter dem Wurzelzeichen größer als $|x_0|^2$ ist, ist auch $\delta > 0$. Wählt man den so angegebenen Wert von δ, so folgt aus $|x - x_0| < \delta$ die Ungleichung $|f(x) - f(x_0)| < \varepsilon$.

Die Funktion $f: X \to \mathbf{R}$ heißt **gleichmäßig stetig in** $D \subseteq X$, wenn es für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ ein $\delta > 0$ gibt, das höchstens von ε abhängt (d.h. $\delta = \delta(\varepsilon)$), mit folgender Eigenschaft:

Für jedes $x \in D$ und für jedes $y \in D$ mit $|x - y| < \delta$ ist $|f(x) - f(y)| < \varepsilon$.

Jede in $D \subseteq X$ gleichmäßig stetige Funktion ist dort natürlich auch stetig. Es gibt jedoch stetige Funktionen, die nicht gleichmäßig stetig sind, wie das Beispiel der Funktion $f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$ zeigt:

Angenommen, f wäre gleichmäßig stetig. Es sei $\varepsilon > 0$ vorgegeben, und $\delta > 0$ sei der zugehörige Wert aus der Definition der gleichmäßigen Stetigkeit. Es wird ein Wert δ' gewählt mit $0 < \delta' < \delta$ und eine große reelle Zahl $x \in \mathbf{R}$ mit $\delta' \cdot (2 \cdot x + \delta') \geq \varepsilon$ (einen derartigen Wert x findet man immer, da der Ausdruck links beliebig groß gemacht werden kann). Außerdem wird $y = x + \delta'$ gesetzt. Dann ist $|x - y| = |x - (x + \delta')| = \delta' < \delta$, aber

$$|f(x) - f(y)| = |x^2 - y^2| = |y^2 - x^2| = |(x + \delta')^2 - x^2| = |2 \cdot x \cdot \delta' + \delta'^2| = \delta' \cdot (2 \cdot x + \delta') \geq \varepsilon .$$

Es lässt sich zeigen, dass eine auf $\mathbf{R}$ gleichmäßig stetige Funktion „nicht zu schnell wächst", nämlich höchstens wie eine lineare Funktion (genauer: Ist $f: \mathbf{R} \to \mathbf{R}$ gleichmäßig stetig, so gibt es eine Konstante $C > 0$ mit $|f(x)| \leq C \cdot (1 + |x|)$).

Satz 5.2-1:

Die folgenden beiden Aussagen (a) und (b) sind gleichbedeutend:

(a) $f: X \to \mathbf{R}$ ist an der Stelle $x_0 \in X$ stetig.

und

(b) Für jede Folge $(x_n)_{n \in \mathbf{N}}$ mit $\lim\limits_{n \to \infty} x_n = x_0$ gilt $\lim\limits_{n \to \infty} f(x_n) = f(x_0)$.

Um die Äquivalenz beider Aussagen zu beweisen, wird zunächst „(a) $\Rightarrow$ (b)" gezeigt:
Die Funktion $f: X \to \mathbf{R}$ sei an der Stelle $x_0 \in X$ stetig, und es sei $(x_n)_{n \in \mathbf{N}}$ eine Folge mit $\lim\limits_{n \to \infty} x_n = x_0$. Zu $\varepsilon > 0$ gibt es $\delta > 0$ mit $|f(x) - f(x_0)| < \varepsilon$ für jedes $x \in X$ mit $|x - x_0| < \delta$. Aufgrund der Konvergenz der Folge $(x_n)_{n \in \mathbf{N}}$ gibt es zu δ ein $n_0 \in \mathbf{N}$ mit $|x_n - x_0| < \delta$ für jedes $n \geq n_0$. Daher ist $|f(x_n) - f(x_0)| < \varepsilon$ für jedes $n \geq n_0$, d.h. die Folge $(f(x_n))_{n \in \mathbf{N}}$ konvergiert gegen $f(x_0)$.

Für die Umkehrung „(b) $\Rightarrow$ (a)" wird (b) als gültig vorausgesetzt, aber angenommen, dass f in x_0 nicht stetig ist. Das bedeutet, dass es $\varepsilon > 0$ gibt, so dass für alle $\delta > 0$ ein $x_\delta \in X$ mit $|x_\delta - x_0| < \delta$, aber $|f(x_\delta) - f(x_0)| \geq \varepsilon$ existiert. Es werden alle δ der Form $\delta = 1/n$ mit $n \geq 1$ betrachtet. Zu jedem $n \in \mathbf{N}$ mit $n \geq 1$ gibt es also ein $x_{1/n} \in X$ mit $|x_{1/n} - x_0| < 1/n$, aber $|f(x_{1/n}) - f(x_0)| \geq \varepsilon$. Die Folge $(x_{1/n})_{n \in \mathbf{N},\, n \geq 1}$ konvergiert gegen x_0. Nach Voraussetzung (b)

konvergiert $\left(f(x_{1/n})\right)_{n\in\mathbf{N},\, n\geq 1}$ gegen $f(x_0)$, d.h. $\left|f(x_{1/n}) - f(x_0)\right| < \varepsilon$ für fast alle $n \in \mathbf{N}$ im Widerspruch zur Konstruktion von $\left(x_{1/n}\right)_{n\in\mathbf{N},\, n\geq 1}$.

Mit Hilfe des Satzes 5.2-1 lässt sich häufig nachweisen, dass eine Funktion in einem Punkt $x_0 \in X$ *nicht* stetig ist. Dazu braucht man nur eine einzige Folge $\left(x_n\right)_{n\in\mathbf{N}}$ anzugeben, die gegen $x_0 \in X$ konvergiert, deren Bildwerte unter f aber nicht gegen $f(x_0)$ gehen.

Beispiel:

Die Funktion

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & \begin{cases} x & \text{für } x < 1 \\ x+1 & \text{für } x \geq 1 \end{cases} \end{cases}$$

ist in $x_0 = 1$ nicht stetig. Dazu betrachte man die Folge $\left(x_n\right)_{n\in\mathbf{N}}$ mit $x_n = 1 - 1/(n+1)$. Es gilt $\lim\limits_{n\to\infty} x_n = 1$. Andererseits ist $f(x_n) = 1 - 1/(n+1)$ und $f(x_0) = f(1) = 2$, also $\lim\limits_{n\to\infty} f(x_n) \neq f(x_0)$.

Satz 5.2-2:

Sind $f: X \to \mathbf{R}$ und $g: X \to \mathbf{R}$ mit $X \subseteq \mathbf{R}$ stetig, so auch die folgenden Abbildungen:

(i) $$f+g: \begin{cases} X & \to & \mathbf{R} \\ x & \to & f(x)+g(x) \end{cases}$$

(ii) $$f\cdot g: \begin{cases} X & \to & \mathbf{R} \\ x & \to & f(x)\cdot g(x) \end{cases}$$

(iii) $$|f|: \begin{cases} X & \to & \mathbf{R} \\ x & \to & |f(x)| \end{cases}$$

(iv) $$c\cdot f: \begin{cases} X & \to & \mathbf{R} \\ x & \to & c\cdot f(x) \end{cases} \quad \text{mit } c \in \mathbf{R}$$

../..

(v) Ist $g(x_0) \neq 0$, so ist $f / g: \begin{cases} X & \to & \mathbf{R} \\ x & \to & \dfrac{f(x)}{g(x)} \end{cases}$ in $x_0 \in \mathbf{R}$ stetig.

(vi) Mit f und g ist auch $g \circ f$ stetig.

(vii) Es sei $X \subseteq \mathbf{R}$ ein Intervall und $f: X \to \mathbf{R}$ eine streng monoton steigende/fallende stetige Funktion. Dann ist $f(X)$ ein Intervall, $f: X \to f(X)$ ist bijektiv, und die Umkehrfunktion $\overset{-1}{f}: f(X) \to X$ ist streng monoton steigend/fallend und stetig.

Die Teile (i) – (v) ergeben sich mit Satz 5.2-1, jeweils zusammen mit den Regeln aus Satz 5.1-2. Die Teile (vi) und (vii) verifiziert man durch Anwendung von Satz 5.2-1.

Teil (vi) wendet Satz 5.2-1 an: Es sei $x_0 \in X$ und $(x_n)_{n \in \mathbf{N}}$ eine Folge mit $\lim\limits_{n \to \infty} x_n = x_0$, dann gilt wegen der Stetigkeit von f $\lim\limits_{n \to \infty} f(x_n) = f(x_0)$. Wegen der Stetigkeit von g ist $\lim\limits_{n \to \infty} g(f(x_n)) = g(f(x_0))$. Daher ist $g \circ f$ in x_0 stetig.

Teil (vii) bedarf einer genaueren Betrachtung, hier für den Fall, dass f streng monoton steigend ist:
Dazu seien und $y_1 \in f(X)$, etwa $y_0 \in f(x_0)$ und $y_1 \in f(x_1)$ mit $x_0 \in X$ und $x_1 \in X$ und $x_0 \leq x_1$. Wegen der Monotonie von f ist $y_0 = f(x_0) \leq f(x_1) = y_1$. Es sei $y \in [y_0, y_1]$. Im folgenden Satz 5.2-3 (ii) wird nachgewiesen, dass es zu $y \in \mathbf{R}$ mit $y_0 \leq y \leq y_1$ ein $x \in X$ mit $x_0 \leq x \leq x_1$ und $f(x) = y$ gibt[5]. Das bedeutet $y \in f(X)$ bzw. $[y_0, y_1] \subseteq f(X)$. Aus Satz 1.7-3 folgt, dass $f(X)$ ein Intervall ist.
Sind $x_0 \in X$ und $x_1 \in X$ mit $x_0 \neq x_1$, dann ist wegen der strengen Monotonie von f im Fall $x_0 < x_1$ auch $f(x_0) < f(x_1)$, und im Fall $x_0 > x_1$ auch $f(x_0) > f(x_1)$, also $f(x_0) \neq f(x_1)$. Daher ist f injektiv (und $f: X \to f(X)$ nach Definition surjektiv), und es existiert die Umkehrabbildung $\overset{-1}{f}: f(X) \to X$.
Diese ist streng monoton steigend: Dazu seien $y_0 \in f(X)$ und $y_1 \in f(X)$, etwa $y_0 \in f(x_0)$ und $y_1 \in f(x_1)$ mit $x_0 \in X$ und $x_1 \in X$ und $y_0 < y_1$. Wäre $x_0 \geq x_1$, so folgte wegen der Monotonie von f ist $y_0 = f(x_0) \geq f(x_1) = y_1$. Also ist $x_0 = \overset{-1}{f}(y_0) < \overset{-1}{f}(y_1) = x_1$.

[5] Da für Satz 5.2-3 (ii) die Aussage in Satz 5.2-2 (vii) nicht verwendet wird, ist dieser „Vorgriff“ zulässig.

Die Stetigkeit von $\overset{-1}{f}: f(X) \to X$ sieht man wie folgt:
Es sei $y_0 \in f(X)$ und $(y_n)_{n\in\mathbf{N}}$ eine Folge in $f(X)$ eine Folge mit $\lim\limits_{n\to\infty} y_n = y_0$. Es sei $x_0 = \overset{-1}{f}(y_0)$ und $x_n = \overset{-1}{f}(y_n)$, also $\lim\limits_{n\to\infty} f(x_n) = f(x_0)$. Gemäß Satz 5.2-1 ist $\lim\limits_{n\to\infty} x_n = x_0$ zu zeigen. Falls $(x_n)_{n\in\mathbf{N}}$ nicht gegen x_0 konvergiert, so gibt es ein $\varepsilon > 0$, so dass für unendlich viele $n \in \mathbf{N}$ $|x_n - x_0| \geq \varepsilon$ gilt. Die Menge dieser natürlichen Zahlen (Indizes) sei *M*. Definiert man $M_{\geq} = \{n \mid x_n \geq x_0 + \varepsilon\}$ und $M_{\leq} = \{n \mid x_n \leq x_0 + \varepsilon\}$, so ist $M = M_{\geq} \cup M_{\leq}$, und mindestens eine der Mengen $M_{\geq}$ oder $M_{\leq}$ hat unendlich viele Elemente, etwa $M_{\geq}$. Für jedes $n \in M_{\geq}$ ist (wegen der strengen Monotonie von *f*) $f(x_n) \geq f(x_0 + \varepsilon) > f(x_0)$. Das bedeutet: es gibt $\varepsilon' > 0$, nämlich $\varepsilon' = f(x_0 + \varepsilon) - f(x_0)$, so dass für unendlich viele $n \in \mathbf{N}$, nämlich für jedes $n \in M_{\geq}$, die Abschätzung $|f(x_n) - f(x_0)| \geq f(x_0 + \varepsilon) - f(x_0) = \varepsilon'$ gilt. Diese Schlussfolgerung steht im Widerspruch zu $\lim\limits_{n\to\infty} f(x_n) = f(x_0)$.

Satz 5.2-2 (vii) begründet die **Existenz der Wurzelfunktionen in den reellen Zahlen**, d.h. die Bildung einer beliebigen Wurzel aus einer reellen Zahl. Dazu wird für $n \in \mathbf{N}$ mit $n \geq 1$ die Funktion

$$f_n : \begin{cases} \mathbf{R}_{\geq 0} & \to & \mathbf{R}_{\geq 0} \\ x & \to & x^n \end{cases}$$

betrachtet. Sie ist auf $\mathbf{R}_{\geq 0}$ stetig und streng monoton steigend. Nach Satz 5.2-2 (vii) ist sie bijektiv, und die Umkehrabbildung ist streng monoton steigend und stetig. Sie wird als ***n*-te Wurzelfunktion** bezeichnet:

$$\overset{-1}{f_n} : \begin{cases} \mathbf{R}_{\geq 0} & \to & \mathbf{R}_{\geq 0} \\ x & \to & \overset{-1}{f_n}(x) = \sqrt[n]{x} \end{cases}$$

$\sqrt[n]{x}$ ist derjenige Wert $y \in \mathbf{R}$, für den $y^n = x$ gilt.

Der folgende Satz, der in 5.2-2 (vii) bereits verwendet wurde, drückt noch einmal aus, dass der Graph einer stetigen Funktion keine Sprünge aufweist.

Satz 5.2-3: (Zwischenwertsatz)

(i) Es seien $a \in \mathbf{R}$ und $b \in \mathbf{R}$ reelle Zahlen mit $a \le b$. Die Funktion $f : [a,b] \to \mathbf{R}$ sei im Intervall $[a,b]$ stetig. Außerdem gelte $f(a) \le 0 \le f(b)$. Dann gibt es ein $\gamma \in [a,b]$ mit $f(\gamma) = 0$.

(ii) Es seien $a \in \mathbf{R}$ und $b \in \mathbf{R}$ reelle Zahlen mit $a < b$. Die Funktion $f : [a,b] \to \mathbf{R}$ sei im Intervall $[a,b]$ stetig. Es gelte $f(a) < f(b)$. Außerdem sei $c \in \mathbf{R}$ mit $f(a) \le c \le f(b)$. Dann gibt es ein $\gamma \in [a,b]$ mit $f(\gamma) = c$.

Aussage (i) ist richtig, wenn $f(a) = 0$ oder $f(b) = 0$ ist. Es gelte daher $f(a) < 0 < f(b)$. Es werden zwei Folgen $(x_n)_{n\in\mathbf{N}}$ und $(y_n)_{n\in\mathbf{N}}$ rekursiv definiert:

$x_0 = a$, $y_0 = b$,

$$x_{n+1} = \begin{cases} \dfrac{x_n + y_n}{2} & \text{falls } f\left(\dfrac{x_n + y_n}{2}\right) < 0 \\ x_n & \text{falls } f\left(\dfrac{x_n + y_n}{2}\right) \ge 0 \end{cases}, \quad y_{n+1} = \begin{cases} y_n & \text{falls } f\left(\dfrac{x_n + y_n}{2}\right) < 0 \\ \dfrac{x_n + y_n}{2} & \text{falls } f\left(\dfrac{x_n + y_n}{2}\right) \ge 0 \end{cases}.$$

Offensichtlich gilt $[x_{n+1}, y_{n+1}] \subseteq [x_n, y_n]$ und $y_{n+1} - x_{n+1} = \dfrac{y_n - x_n}{2} = \ldots = \dfrac{b-a}{2^n}$. Die Folgen $(x_n)_{n\in\mathbf{N}}$ und $(y_n)_{n\in\mathbf{N}}$ definieren eine Intervallschachtelung, deren Länge gegen 0 konvergiert. Die linke und die rechte Intervallgrenzen konvergieren daher gegen eine Zahl $\gamma \in \mathbf{R}$: $\gamma = \lim\limits_{n\to\infty} x_n = \lim\limits_{n\to\infty} y_n$. Mit Satz 5.2-1 folgt $f(\gamma) = \lim\limits_{n\to\infty} f(x_n) = \lim\limits_{n\to\infty} f(y_n)$. Nach Konstruktion ist $f(x_n) \le 0$ und damit $f(\gamma) \le 0$; ebenso gilt nach Konstruktion $f(y_n) \ge 0$ und damit $f(\gamma) \ge 0$, insgesamt $f(\gamma) = 0$.

Für den Nachweis von (ii) wird $g : [a,b] \to \mathbf{R}$ durch $g(x) = f(x) - c$ definiert. Dann ist g stetig, und es gilt $g(a) = f(a) - c \le 0$ und $g(b) = f(b) - c \ge 0$. Gemäß Teil (i) gibt es ein $\gamma \in [a,b]$ mit $g(\gamma) = 0$, d.h. $f(\gamma) = c$.

Es seien $X \subseteq \mathbf{R}$ und $f : X \to \mathbf{R}$ eine Funktion. Das Element $x_0 \in X$ heißt **Nullstelle** von f, wenn $f(x_0) = 0$ gilt.

Die Funktion $f: X \to \mathbf{R}$ besitzt im Punkt $x_0 \in \mathbf{R}$ den (endlichen) **Grenzwert** $f_0 \in \mathbf{R}$, wenn gilt:

Für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ gibt es ein (von ε abhängiges) $\delta = \delta(\varepsilon)$ mit folgender Eigenschaft:

Für jedes $x \in X$ mit $|x - x_0| < \delta$ ist $|f(x) - f_0| < \varepsilon$.

Zu beachten ist, dass der Wert x_0 nicht zum Definitionsbereich von *f* gehören muss.

Wählt man eine beliebig kleine ε-Umgebung von f_0, so findet man immer eine δ-Umgebung von x_0, die durch *f* komplett in diese ε-Umgebung abgebildet wird. In jeder beliebig kleinen ε-Umgebung von f_0 findet man Bildpunkte (unter *f*), deren Urbilder nahe bei x_0 liegen.

Schreibweise: $\lim_{x \to x_0} f(x) = f_0$.

Die Funktion $f: X \to \mathbf{R}$ besitzt in $x_P \in \mathbf{R}$ einen **Pol**, wenn gilt:

Für jedes $K \in \mathbf{R}$ mit $K > 0$ gibt es ein (von *K* abhängiges) $\delta = \delta(K)$ mit folgender Eigenschaft:

Für jedes $x \in X$ mit $|x - x_P| < \delta$ ist $|f(x)| > K$.

Die Funktionswerte wachsen über jede Grenze, wenn man sich dem Wert x_P nähert. Dabei ist zu beachten, dass x_P nicht zu *X* gehört.

Schreibweise: $\lim_{x \to x_P} f(x) = \pm\infty$.

Unmittelbar aus der Definition folgt

Satz 5.2-4:

Die folgenden beiden Aussagen (a) und (b) sind gleichbedeutend:

(a) $\lim\limits_{x \to x_P} f(x) = 0$

und

(b) Die durch $\left(1/f\right)(x) = 1/f(x)$ definierte Funktion besitzt bei x_P einen Pol.

Für den Nachweis von „$(\mathrm{a}) \Rightarrow (\mathrm{b})$“ gelte $\lim\limits_{x \to x_P} f(x) = 0$, und es sei $K \in \mathbf{R}$ mit $K > 0$. Zu $\varepsilon = 1/K$ gibt es $\delta > 0$, so dass für $x \in X$ mit $|x - x_p| < \delta$ gilt: $|f(x) - 0| = |f(x)| < \varepsilon = 1/K$. Dann ist $\left|\left(1/f\right)(x)\right| = \left|1/f(x)\right| > K$, d.h. $1/f$ besitzt bei x_P einen Pol.

Für die umgekehrte Implikation „$(\mathrm{b}) \Rightarrow (\mathrm{a})$“ wird entsprechend argumentiert.

Die Funktion $f: X \to \mathbf{R}$ hat für $x \to \infty$ die **Asymptote** $s: X \to \mathbf{R}$, wenn gilt: Für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ gibt es eine (von ε abhängige) Konstante $C = C(\varepsilon) > 0$ mit folgender Eigenschaft:

Für jedes $x \in X$ mit $x > C(\varepsilon)$ ist $|f(x) - s(x)| < \varepsilon$.

Der Funktionsverlauf von f nähert sich beliebig dicht dem Funktionsverlauf von s an, wenn man x nur genügend groß wählt.

Schreibweise: $\lim\limits_{x \to \infty} |f(x) - s(x)| = 0$ bzw. $\lim\limits_{x \to \infty} f(x) = s(x)$.

Entsprechend hat die Funktion $f: X \to \mathbf{R}$ hat für $x \to -\infty$ die **Asymptote** $s: X \to \mathbf{R}$, wenn gilt: Für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ gibt es eine (von ε abhängige) Konstante $C = C(\varepsilon) > 0$ mit folgender Eigenschaft:

Für jedes $x \in X$ mit $x < 0$ und $|x| > C(\varepsilon)$ ist $|f(x) - s(x)| < \varepsilon$.

Schreibweise: $\lim\limits_{x \to -\infty} |f(x) - s(x)| = 0$ bzw. $\lim\limits_{x \to -\infty} f(x) = s(x)$.

5.3 Polynome

Ein Polynom ist eine Funktion $p : \mathbf{R} \to \mathbf{R}$, zu deren Berechnung man mit den Rechenoperationen Addition, Subtraktion und Multiplikation auskommt.

Beispielsweise wird durch $p(x) = (x-1)\cdot(x^2+5) - \sqrt{2}\cdot x^7 = -\sqrt{2}\cdot x^7 + x^3 - x^2 + 5\cdot x - 5$ ein Polynom definiert.

Polynome lassen sich immer auf eine „standardisierte“ Form bringen:

Eine Funktion

$$p : \begin{cases} \mathbf{R} \to \mathbf{R} \\ x \to a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + \ldots a_1 \cdot x + a_0 \end{cases}$$

mit reellen Konstanten $a_n, a_{n-1}, \ \ldots \ , a_1, a_0$ und $a_n \neq 0$ heißt **Polynom vom Grad *n***.

Für $p(x) = a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + \ldots a_1 \cdot x + a_0$ schreibt man wie üblich $p(x) = \sum_{i=0}^{n} a_i \cdot x^i$.

Beispiele:

Die durch $p(x) = (x-1)\cdot(x^2+5) - \sqrt{2}\cdot x^7 = -\sqrt{2}\cdot x^7 + x^3 - x^2 + 5\cdot x - 5$ definierte Funktion ist ein Polynom vom Grad 7.

Die durch $p(x) = 5\cdot(x^2 - x)\cdot x^2 + \sqrt{2{,}5}$ definierte Funktion ist ein Polynom vom Grad 4.

Die durch $p(x) = 3\cdot x^2 - \sqrt{x} + 9$ definierte Funktion ist kein Polynom.

Polynome vom Grad 0:

$p(x) = a_0 = const.$

Hier wird auch $a_0 = 0$ zugelassen.

Der Graph eines Polynoms vom Grad 0 ist eine Gerade, die im (x, y)-Koordinatensystem parallel zur x-Achse verläuft und die y-Achse im Punkt $(0, a_0)$ schneidet.

Polynome vom Grad 1:

$p(x) = a_1 \cdot x + a_0$ mit $a_1 \neq 0$

Die einzige Nullstelle ist $x_0 = -\frac{a_0}{a_1}$.

Der Graph eines Polynoms 1. Grades ist eine Gerade und schneidet im (x, y)-Koordinatensystem die y-Achse im Punkt $(0, a_0)$.

Polynome vom Grad 2:

$p(x) = a_2 \cdot x^2 + a_1 \cdot x + a_0$ mit $a_2 \neq 0$

Es gibt zwei oder eine oder keine reelle Nullstelle. Die Nullstellen berechnen sich zu

$$x_{01,02} = -\frac{a_1}{2 \cdot a_2} \pm \sqrt{\frac{{a_1}^2}{4 \cdot {a_2}^2} - \frac{a_0}{a_2}}.$$

Diese sind nur dann reellwertig, wenn ${a_1}^2 \geq 4 \cdot a_2 \cdot a_0$ ist.

Ein häufig auftretender Spezialfall ist das Polynom der Form $p(x) = x^2 + p \cdot x + q$ mit $p \in \mathrm{R}$ und $q \in \mathbf{R}$. Dieses Polynom hat die Nullstellen

$$x_{01,02} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}.$$

Die Bedingung für die Reellwertigkeit der Nullstellen lautet $p^2 - 4 \cdot q \geq 0$.

Der Graph eines Polynoms 2.Grades ist eine Parabel, die für $a_2 > 0$ nach oben und für $a_2 < 0$ nach unten geöffnet ist.

Ist $a_2 > 0$ (bzw. $a_2 < 0$), so wird der minimale (bzw. maximale) Wert des Polynoms $p(x) = a_2 \cdot x^2 + a_1 \cdot x + a_0$ an der Stelle

$$x_S = -\frac{a_1}{2 \cdot a_2}$$

angenommen; der Funktionswert lautet dabei $p(x_S) = a_0 - \frac{a_1^{\,2}}{4 \cdot a_2}$.

Im Spezialfall $p(x) = x^2 + p \cdot x + q$ lauten die entsprechenden Werte

$x_S = -p/2$ und $p(x_S) = q - \left(p/2\right)^2$.

Polynome vom Grad ≥ 3:

Für Polynome 3. und 4. Grades gibt es noch eine geschlossene Formel zur Nullstellenbestimmung, für Polynome höheren Grades i.a. nicht.

Satz 5.3-1:

(i) $p(x) = a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + ... + a_1 \cdot x + a_0 = \sum_{i=0}^{n} a_i \cdot x^i$ sei ein Polynom vom Grad n und x_0 eine Nullstelle von p (d.h. $p(x_0) = 0$). Dann gibt es ein Polynom p_1 vom Grad $n-1$ mit

$$p(x) = (x - x_0) \cdot p_1(x).$$

Man kann also den **Linearfaktor** $x - x_0$ aus $p(x)$ ausklammern.

Im Spezialfall $p(x) = x^n - a^n$ mit $a \neq 0$ lautet eine Nullstelle $x_0 = a$. Es ist

$$p(x) = x^n - a^n = (x - a) \cdot \sum_{i=0}^{n-1} a^{n-i-1} \cdot x^i .$$

(ii) Ein Polynom vom Grad n hat höchstens n viele reelle Nullstellen.

(iii) Ein Polynom von *ungeradem* Grad hat mindestens eine Nullstelle.

Teil (i) sieht man durch Rechnen:

Es ist im Spezialfall $p(x) = x^n - a^n$ die rechte Seite der Gleichung

$$(x-a)\cdot\sum_{i=0}^{n-1} a^{n-i-1}\cdot x^i = \sum_{i=0}^{n-1} a^{n-(i+1)}\cdot x^{i+1} - \sum_{i=0}^{n-1} a^{n-i}\cdot x^i = \sum_{i=1}^{n} a^{n-i}\cdot x^i - \sum_{i=0}^{n-1} a^{n-i}\cdot x^i = x^n - a^n .$$

Es sei x_0 eine Nullstelle von *p*. Für $x_0 = 0$ ist $0 = p(0) = a_0$ und

$p(x) = a_n\cdot x^n + a_{n-1}\cdot x^{n-1} + ... + a_1\cdot x = (x-0)\cdot\sum_{i=1}^{n} a_i\cdot x^{i-1}$. Wegen $a_n \neq 0$ ist $p_1(x) = \sum_{i=1}^{n} a_i\cdot x^{i-1}$

ein Polynom vom Grad $n-1$.

Für $x_0 \neq 0$ ist

$$\begin{aligned} p(x) &= p(x) - p(x_0) \\ &= \sum_{i=0}^{n} a_i\cdot x^i - \sum_{i=0}^{n} a_i\cdot x_0^i \\ &= \sum_{i=1}^{n} a_i\cdot\left(x^i - x_0^i\right) \qquad \text{(gemäß Spezialfall mit } a = x_0\text{)} \\ &= (x-x_0)\cdot\sum_{i=1}^{n} a_i\cdot\left(x^{i-1} + x_0\cdot x^{i-2} + x_0^2\cdot x^{i-3} + ... + x_0^{i-2}\cdot x + x_0^{i-1}\right). \end{aligned}$$

Setzt man $p_1(x) = \sum_{i=1}^{n} a_i\cdot\left(x^{i-1} + x_0\cdot x^{i-2} + x_0^2\cdot x^{i-3} + ... + x_0^{i-2}\cdot x + x_0^{i-1}\right)$, so sieht man, dass es sich wegen $a_n \neq 0$ um ein Polynom vom Grad $n-1$ handelt.

Der Vorgang des Ausklammerns eines Linearfaktors kann höchstens *n*-mal durchgeführt werden (Teil (ii)).

Teil (iii) kann mit Satz 5.2-3 nachgewiesen werden:

Es sei $p(x) = \sum_{i=0}^{n} a_i\cdot x^i$ mit ungeradem *n* und $a_n \neq 0$. Es wird für $x \neq 0$

$f(x) = 1 + \frac{a_{n-1}}{a_n\cdot x} + ... + \frac{a_1}{a_n\cdot x^{n-1}} + \frac{a_0}{a_n\cdot x^n}$ gesetzt. Dann ist $a_n\cdot x^n\cdot f(x) = p(x)$. Es gilt:

$\lim_{x\to\infty} f(x) = 1 = \lim_{x\to-\infty} f(x)$. Bei negativem a_n ist daher $\lim_{x\to\infty} p(x) = -\infty$ und $\lim_{x\to-\infty} p(x) = \infty$; bei positvem $a_n \neq 0$ ist $\lim_{x\to\infty} p(x) = \infty$ und $\lim_{x\to-\infty} p(x) = -\infty$. Es gibt also $a \in \mathbf{R}$ und $b \in \mathbf{R}$ mit $p(a) < 0$ und $p(b) > 0$ bzw. $p(a) > 0$ und $p(b) < 0$, je nach Vorzeichen von a_n. Mit Satz 5.2-3 folgt die Aussage.

Satz 5.3-2:

Das Polynom $p(x) = \sum_{i=0}^{n} a_i \cdot x^i$ vom Grad n habe die reellen Nullstellen $x_{01}, \ldots, x_{0m}$; hierbei werden mehrfache reelle Nullstellen jeweils auch mehrfach aufgeführt. Dann gilt

$$p(x) = (x - x_{01}) \cdot \ldots \cdot (x - x_{0m}) \cdot p_g(x)$$

mit einem Polynom $p_g(x)$ von geradem Grad $2 \cdot k$, das keine reellen Nullstellen hat. Außerdem ist $n = m + 2 \cdot k$.

5.4 Gebrochen rationale Funktionen

Eine Funktion der Form

$$f: \begin{cases} X & \to & \mathbf{R} \\ x & \to & \dfrac{p(x)}{q(x)} \end{cases}$$

mit $X \subseteq \mathbf{R}$ und den Polynomen $p(x) = \sum_{i=0}^{n} a_i \cdot x^i$ und $q(x) = \sum_{j=0}^{m} b_j \cdot x^j$ und $b_m \neq 0$ heißt **gebrochen rationale Funktion**.

An den Nullstellen von q ist f nicht definiert, d.h. der Definitionsbereich von f lautet

$$D(f) = \mathbf{R} \setminus \{ x_0 \mid q(x_0) = 0 \}.$$

Skizzierung einer gebrochen rationalen Funktion *f*:

1. Schritt:

Es werden alle Nullstellen von p und alle Nullstellen von q bestimmt. Alle diese Nullstellen seien $x_{01}, \ldots, x_{0l}$.

Die Nullstellen von q gehören nicht zum Definitionsbereich von f.

Für jede dieser Nullstellen x_{0i} von p und q wird der 2. Schritt durchgeführt.

2. Schritt:

Es werden 3 mögliche Fälle unterschieden:

1. Fall: x_{0i} ist eine Nullstelle von p, aber nicht von q:

$p(x_{0i}) = 0$ und $q(x_{0i}) \neq 0$

Es gilt

$$f(x_{0i}) = \frac{p(x_{0i})}{q(x_{0i})} = \frac{0}{q(x_{0i})} = 0,$$

d.h. x_{0i} ist eine Nullstelle von f.

2. Fall: x_{0i} ist keine Nullstelle von p, aber eine Nullstelle von q:

$p(x_{0i}) \neq 0$ und $q(x_{0i}) = 0$

Zu beachten ist, dass f für x_{0i} nicht definiert ist.

Es gilt

$$\lim_{x \to x_{0i}} 1/f(x) = \frac{\lim\limits_{x \to x_{0i}} q(x)}{\lim\limits_{x \to x_{0i}} p(x)} = \frac{0}{p(x_{0i})} = 0\,,$$

d.h. f besitzt bei x_{0i} einen Pol.

3. Fall: x_{0i} ist sowohl eine Nullstelle von p, als auch eine Nullstelle von q:

$p(x_{0i}) = 0$ und $q(x_{0i}) = 0$

Zu beachten ist, dass f für x_{0i} nicht definiert ist.

Ist x_{0i} eine *r*-fache Nullstelle von *p* und eine *s*-fache Nullstelle von *q*, dann gilt

$$f(x)=\frac{(x-x_{0i})^r\cdot p_1(x)}{(x-x_{0i})^s\cdot q_1(x)} \text{ mit } p_1(x_{0i})\neq 0 \text{ und } q_1(x_{0i})\neq 0 .$$

Fall 3a: $r > s$

$$\lim_{x\to x_{0i}} f(x)=\lim_{x\to x_{0i}}(x-x_{0i})^{r-s}\cdot\frac{p_1(x_{0i})}{q_1(x_{0i})}=0$$

Fall 3b: $r = s$

$$\lim_{x\to x_{0i}} f(x)=\frac{p_1(x_{0i})}{q_1(x_{0i})}\neq 0$$

In beiden Fällen nennt man x_{0i} eine **behebbare Unstetigkeitsstelle** von *f*, da man *f* stetig nach x_{0i} fortsetzen kann.

Fall 3c: $r < s$

$$\lim_{x\to x_{0i}} 1/f(x)=\lim_{x\to x_{0i}}\frac{(x-x_{0i})^{s-r}\cdot q_1(x)}{p_1(x)}=0 \text{, d.h. } f \text{ hat bei } x_{0i} \text{ einen Pol.}$$

3. Schritt:

Es wird das Verhalten von $f(x)$ bei $x\to\pm\infty$ untersucht.

Es ist $f(x)=\frac{p(x)}{q(x)}$, $p(x)=\sum_{i=0}^{n}a_i\cdot x^i$, $q(x)=\sum_{j=0}^{m}b_j\cdot x^j$ und $b_m\neq 0$.

Fall 4a: Der Grad von *q* ist größer als der Grad von *p*, d.h. $m > n$.

$$\begin{aligned}\lim_{x\to\infty} f(x)&=\lim_{x\to\infty}\frac{a_n\cdot x^n+a_{n-1}\cdot x^{n-1}+\ldots+a_1\cdot x+a_0}{b_m\cdot x^m+b_{m-1}\cdot x^{m-1}+\ldots+b_1\cdot x+b_0}\\&=\lim_{x\to\infty}\frac{a_n\cdot 1/x^{m-n}+a_{n-1}\cdot 1/x^{m-(n-1)}+\ldots+a_1\cdot 1/x^{m-1}+a_0\cdot 1/x^m}{b_m\cdot 1+b_{m-1}\cdot 1/x+\ldots+b_1\cdot 1/x^{m-1}+b_0\cdot 1/x^m}\\&=0,\end{aligned}$$

d.h. f hat bei $x \to \pm\infty$ die Asymptote $s(x) = 0$ (x-Achse).

Fall 4b: Der Grad von q ist nicht größer als der Grad von p, d.h. $m \le n$.

Durch Ausdividieren (**Polynomdivision**) von $p(x) \big/ q(x)$ erhält man auf eindeutige Weise Polynome $s(x)$ und $r(x)$ mit $f(x) = s(x) + \dfrac{r(x)}{q(x)}$. Hierbei hat $s(x)$ den Grad $n - m$ und $r(x)$ einen kleineren Grad als $q(x)$, und es gilt:

$\lim\limits_{x \to \pm\infty} f(x) = s(x)$, d.h. f hat bei $x \to \pm\infty$ die Asymptote $s(x)$.

5.5 Exponential- und Logarithmusfunktion

In Kapitel 5.1 wird die Exponentialfunktion

$$\exp : \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & \sum\limits_{i=0}^{\infty} \dfrac{x^i}{i!} \end{cases}$$

definiert. Zunächst werden einige Eigenschaften dieser Funktion untersucht.

Ein wichtiges Ergebnis, nämlich

$$\exp(1) = \sum_{i=0}^{\infty} \frac{1}{i!} = e = 2{,}718281\ldots,$$

wird in Kapitel 5.1 hergeleitet. Außerdem gilt $\sum\limits_{i=0}^{\infty} \dfrac{x^i}{i!} = \underbrace{1}_{i=0} + \sum\limits_{i=1}^{\infty} \dfrac{x^i}{i!}$, und daher $\exp(0) = 1$.

Mit Satz 5.1-10 ergibt sich

$$\begin{aligned}\exp(x)\cdot\exp(y) &= \left(\sum_{i=0}^{\infty}\frac{x^i}{i!}\right)\cdot\left(\sum_{i=0}^{\infty}\frac{y^i}{i!}\right)\\ &= \sum_{i=0}^{\infty}\left(\sum_{k=0}^{i}\frac{x^k}{k!}\cdot\frac{y^{i-k}}{(i-k)!}\right)\\ &= \sum_{i=0}^{\infty}\left(\frac{1}{i!}\cdot\sum_{k=0}^{i}\frac{i!}{k!\cdot(i-k)!}\cdot x^k\cdot y^{i-k}\right)\\ &= \sum_{i=0}^{\infty}\left(\frac{1}{i!}\cdot(x+y)^i\right)\\ &= \exp(x+y)\end{aligned}$$

Für jedes $r\in\mathbf{R}$ gilt daher: $\exp(r)=\exp\left(\frac{r}{2}+\frac{r}{2}\right)=\left(\exp\left(\frac{r}{2}\right)\right)^2\geq 0$. Wegen $1=\exp(0)=\exp(r-r)=\exp(r)\cdot\exp(-r)$ ist sogar $\exp(r)>0$. Daher kann der Wertebereich der Exponentialfunktion auf $\mathbf{R}_{>0}$ eingeschränkt werden.

Für jedes $n\in\mathbf{N}$ lässt sich der Wert der Exponentialfunktion folgendermaßen berechnen:
$\exp(n)=\exp(\underbrace{1+1+\ldots+1}_{n-\text{mal}})=(\exp(1))^n=e^n$.

Dieses Ergebnis bedeutet, dass man zur Ermittlung des Werts von $\exp(n)$ anstelle der Grenzwertberechnung $\sum_{i=0}^{\infty}\frac{n^i}{i!}$ in $\mathbf{R}$ das n-fache Produkt der reellen Zahl $e\in\mathbf{R}$ bildet. Zur Berechnung des Werts von $\exp(n)$ mit Hilfe eines Computers, in dem reelle Zahlen nur approximiert werden können, wird man eher die Reihenentwicklung $\sum_{i=0}^{\infty}\frac{n^i}{i!}$ verwenden und diese nach einer endlichen Anzahl Summanden, entsprechend der vorgegebenen Genauigkeit zur Darstellung reeller Zahlen, abbrechen.

Für eine negative ganze Zahl $m\in\mathbf{Z}$, $m=-n$ mit $n\in\mathbf{N}$, ist wegen
$\exp(m)\cdot\exp(n)=\exp(-n)\cdot\exp(n)=\exp(0)=1$: $\exp(-n)=(\exp(n))^{-1}$ und
$\exp(m)=\exp(-n)=(\exp(n))^{-1}=(e^n)^{-1}=e^{-n}=e^m$.

Für die vorletzte Gleichung $\left((e^n)^{-1}=e^{-n}\right)$ wurden die Regeln der wiederholten Produktbildung in einer kommutativen Gruppe $(G,\circ)$ verwendet: Für $a\in G$ mit dem zu a inversen Element a^{-1} und dem neutralen Element 1 (hier: G = multiplikative Gruppe von $\mathbf{R}$ und $a=e$) und $n\in\mathbf{N}$ wird definiert:
$a^0=1$, $a^n=\underbrace{a\circ\ldots\circ a}_{n-\text{mal}}$ und $a^{-n}=\underbrace{a^{-1}\circ\ldots\circ a^{-1}}_{n-\text{mal}}=(a^{-1})^n$ für $n\geq 1$.

Damit ist wegen $a^{-n} \circ a^n = \underbrace{a^{-1} \circ \ldots \circ a^{-1}}_{n-\text{mal}} \circ \underbrace{a \circ \ldots \circ a}_{n-\text{mal}} = 1$: $a^{-n} = \left(a^n\right)^{-1}$.

Weiter wird in G für $n \in \mathbf{N}$ und $m \in \mathbf{N}_{>0}$ definiert: Ist für $b \in G$ $a = b^m$, dann wird b mit $a^{\frac{1}{m}}$ bezeichnet; entsprechend wird b mit $a^{\frac{n}{m}}$ bezeichnet, wenn $b^m = a^n$ gilt.

Mit diesen Bezeichnungen ist $a_1^{\frac{n}{m}} \circ a_2^{\frac{n}{m}} = \left(a_1 \circ a_2\right)^{\frac{n}{m}}$: Es sei $b_1 = a_1^{\frac{n}{m}}$, $b_2 = a_2^{\frac{n}{m}}$ und $c = \left(a_1 \circ a_2\right)^{\frac{n}{m}}$. Dann ist $b_1^m = a_1^n$, $b_2^m = a_2^n$ und $c^m = \left(a_1 \circ a_2\right)^n = a_1^n \circ a_2^n = b_1^m \circ b_2^m = \left(b_1 \circ b_2\right)^m$ und damit $c = \left(b_1 \circ b_2\right)^{\frac{m}{m}} = b_1 \circ b_2$.

Außerdem gilt wegen $1^m = 1^n$: $1^{\frac{n}{m}} = 1$.

Schließlich wird $a^{-\frac{n}{m}} = \left(a^{-1}\right)^{\frac{n}{m}}$ definiert. Dann ist $a^{-\frac{n}{m}} = \left(a^{\frac{n}{m}}\right)^{-1}$; denn

$$a^{-\frac{n}{m}} \circ a^{\frac{n}{m}} = \left(a^{-1}\right)^{\frac{n}{m}} \circ a^{\frac{n}{m}} = \left(a \circ a^{-1}\right)^{\frac{n}{m}} = 1^{\frac{n}{m}} = 1.$$

Für eine rationale Zahl $q = \frac{n}{m}$ mit $n \in \mathbf{N}$ und $m \in \mathbf{N}_{>0}$ ist

$$\left(\exp\left(\frac{n}{m}\right)\right)^m = \exp\left(\underbrace{\frac{n}{m} + \frac{n}{m} + \ldots + \frac{n}{m}}_{m\text{-mal}}\right) = \exp(n) = e^n, \text{ also}$$

$$\exp\left(\frac{n}{m}\right) = e^{\frac{n}{m}}.$$

Für eine rationale Zahl $q < 0$, $p = -q > 0$, ist wegen

$\exp(q) \cdot \exp(p) = \exp(q) \cdot \exp(-q) = \exp(q - q) = \exp(0) = 1$: $\exp(q) = \left(\exp(-q)\right)^{-1}$ und daher

$\exp(q) = \left(\exp(-q)\right)^{-1} = \left(e^{-q}\right)^{-1} = e^{-(-q)} = e^q$.

Insgesamt ist also für jedes $x \in \mathbf{Q}$ gezeigt: $\exp(x) = e^x$.

Aufgrund dieses Ergebnisses verwendet man für alle $x \in \mathbf{R}$ anstelle von $\exp(x)$ die Bezeichnung e^x; zu beachten ist, dass dieses für $x \in \mathbf{Q}$ bewiesen wurde, für $x \in \mathbf{R} \setminus \mathbf{Q}$ stellt es eine abkürzende Schreibweise für den Grenzwert der Reihe $\sum_{i=0}^{\infty} \frac{x^i}{i!}$ dar.

Diese und weitere Ergebnisse fasst folgender Satz zusammen:

Satz 5.5-1:

(i) Die Exponentialfunktion

$$\exp: \begin{cases} \mathbf{R} & \to &]0,\infty[\\ x & \to & \sum_{i=0}^{\infty} \frac{x^i}{i!} \end{cases}$$

ist streng monoton steigend, bijektiv und stetig und erfüllt die Funktionalgleichung

$$\exp(x+y) = \exp(x) \cdot \exp(y) \text{ bzw. } e^{x+y} = e^x \cdot e^y .$$

(ii) Es seien $x \in \mathbf{R}$ und $y \in \mathbf{R}$. Dann gilt

$$\exp(0) = 1,$$
$$\exp(1) = e,$$
$$\exp(x-y) = \exp(x) \cdot (\exp(y))^{-1} \text{ bzw. } e^{x-y} = \frac{e^x}{e^y}.$$

Die Gültigkeit der Funktionalgleichung in Teil (i) wurde oben bereits gezeigt.
Die strenge Monotonie der Exponentialfunktion zeigt man in zwei Schritten: Zunächst ist für $z \in \mathbf{R}$ mit $z > 0$ ist $\exp(z) = 1 + z + \sum_{i=2}^{\infty} \frac{z^i}{i!} > 1$. Für $x \in \mathbf{R}$ und $y \in \mathbf{R}$ mit $x < y$ ist $z = y - x > 0$ und $\exp(y) = \exp(y - x + x) = \exp(y - x) \cdot \exp(x) = \exp(z) \cdot \exp(x) > \exp(x)$. Damit ist die Exponentialfunktion streng monoton steigend und auch injektiv.
Zum Nachweis der Stetigkeit nutzt man die Abschätzung $|\exp(x) - 1| \le 2 \cdot |x|$ für $|x| \le 1$, die aus Satz 5.1-12 (dort mit $n = 0$) folgt, und Satz 5.2-1:
Es sei $x_0 \in \mathbf{R}$ und $(x_n)_{n \in \mathbf{N}}$ eine Folge mit $\lim_{n\to\infty} x_n = x_0$. Zu zeigen ist: $\lim_{n\to\infty} \exp(x_n) = \exp(x_0)$.
Es gilt $|x_n - x_0| \le 1$ ab einem Index $n_0 \in \mathbf{N}$. Hiermit ist $0 < |\exp(x_n - x_0) - 1| \le 2 \cdot |x_n - x_0|$ und $\lim_{n\to\infty} (\exp(x_n - x_0) - 1) = 0$ bzw. $\lim_{n\to\infty} (\exp(x_n - x_0)) = 1$. Mit der Funktionalgleichung folgt
$\lim_{n\to\infty} \exp(x_n) = \lim_{n\to\infty} (\exp(x_n - x_0 + x_0)) = \lim_{n\to\infty} (\exp(x_n - x_0) \cdot \exp(x_0)) = \exp(x_0)$.

Die Surjektivität der Exponentialfunktion folgt mit dem Zwischenwertsatz Satz 5.2-3: Es sei $y \in \mathbf{R}_{>0}$. Zu zeigen ist: es gibt $x \in \mathbf{R}$ mit $\exp(x) = y$.

Für $y = 1$ wird $x = 0$ genommen.

Ist $y > 1$, so setzt man $a = 0$ und $b = n$, so dass $\exp(n) > y$ ist. Diese Wahl für *b* ist möglich, da für $x \geq 0$ wegen $\exp(x) = 1 + x + \sum_{i=2}^{\infty} \frac{x^i}{i!} \geq 1 + x$ das Grenzverhalten $\lim_{x \to \infty} \exp(x) = \infty$ vorliegt. Damit ist $\exp(a) = 1 < \exp(b)$ und $\exp(a) < y < \exp(b)$. Aus der Stetigkeit der Exponentialfunktion folgt mit Satz 5.2-3 (ii) die Existenz von $x \in [a, b]$ mit $\exp(x) = y$.

Ist $0 < y < 1$, so ist $1/y > 1$. Dann gibt es $x \in \mathbf{R}$ mit $\exp(x) = 1/y$ und $\exp(-x) = (\exp(x))^{-1} = y$.

Die letzte Gleichung in Teil (ii) ergibt sich aus $\exp(y) \cdot \exp(-y) = \exp(y - y) = \exp(0) = 1$, also $\exp(-y) = (\exp(y))^{-1}$: $\exp(x - y) = \exp(x + (-y)) = \exp(x) \cdot (\exp(y))^{-1}$.

Aufgrund von Satz 2.2-1 gibt es zur Exponentialfunktion eine eindeutig bestimmte Umkehrfunktion, die stetig, bijektiv und streng monoton steigend ist (Satz 2.2-1 und Satz 5.2-1 (vii)). Diese Funktion heißt **natürlicher Logarithmus** und wird mit ln bezeichnet:

$$\ln : \begin{cases}]0, \infty[& \to & \mathbf{R} \\ x & \to & \ln(x) \end{cases}$$

Es ist $y = \ln(x)$ genau dann, wenn $x = \exp(y) = e^y$ ist. Weiter gilt:

Mit $z = \ln(x)$ und $z' = \ln(y)$ ist $\exp(z + z') = \exp(z) \cdot \exp(z') = x \cdot y$, also

$$\ln(\exp(z + z')) = z + z' = \ln(x) + \ln(y) = \ln(x \cdot y).$$

Die Ergebnisse und weitere Eigenschaften des natürlichen Logarithmus sind im folgenden Satz zusammengefasst.

Satz 5.5-2:

(i) Die natürliche Logarithmusfunktion

$$\ln : \begin{cases}]0,\infty[& \to \quad \mathbf{R} \\ x & \to \quad \ln(x) \end{cases}$$

ist streng monoton steigend, bijektiv und stetig und erfüllt die Funktionalgleichung

$$\ln(x \cdot y) = \ln(x) + \ln(y) .$$

(ii) Es seien $x \in \mathbf{R}_{>0}$ und $y \in \mathbf{R}_{>0}$. Dann gilt

$$\ln(1) = 0 ,$$
$$\ln(e) = 1 ,$$
$$\ln(x) < 0 \text{ für } x < 1 , \ \ln(x) > 0 \text{ für } x > 1 ,$$
$$\ln(x/y) = \ln(x) - \ln(y) ,$$
$$\ln(x^n) = n \cdot \ln(x) \text{ für jedes } n \in \mathbf{Z} ,$$
$$\ln(e^x) = x \text{ und } \exp(\ln(x)) = x \text{ bzw. } e^{\ln(x)} = x .$$

Es sei $a \in \mathbf{R}$ mit $a > 0$. Dann heißt die Funktion

$$\exp_a : \begin{cases} \mathbf{R} & \to \quad]0,\infty[\\ x & \to \quad \exp(\ln(a) \cdot x) = e^{\ln(a) \cdot x} \end{cases}$$

Exponentialfunktion zur Basis *a*. Statt $\exp_a(x)$ schreibt man a^x.

Satz 5.5-3:

Es sei $a \in \mathbf{R}$ mit $a > 0$.

(i) Die Exponentialfunktion zur Basis *a*

$$\exp_a : \begin{cases} \mathbf{R} & \to &]0, \infty[\\ x & \to & \exp(\ln(a) \cdot x) = e^{\ln(a) \cdot x} \end{cases}$$

ist stetig. Für $a > 1$ ist sie streng monoton steigend, für $a < 1$ ist sie streng monoton fallend, für $a = 1$ ist sie konstant 1.

Für $a \neq 1$ ist die Exponentialfunktion zur Basis *a* bijektiv.

(ii) Es seien $x \in \mathbf{R}$ und $y \in \mathbf{R}$. Dann gilt

$$\exp_a(0) = 1 \quad \text{bzw.} \quad a^0 = 1,$$
$$\exp_a(1) = a \quad \text{bzw.} \quad a^1 = a,$$
$$\exp_a(x + y) = \exp_a(x) \cdot \exp_a(y) \quad \text{bzw.} \quad a^{x+y} = a^x \cdot a^y,$$
$$\exp_a(x - y) = \exp_a(x) \cdot (\exp_a(y))^{-1} \quad \text{bzw.} \quad a^{x-y} = \frac{a^x}{a^y},$$
$$(\exp_a(x))^y = \exp_a(x \cdot y) \quad \text{bzw.} \quad (a^x)^y = a^{x \cdot y}.$$

Die Stetigkeit in Teil (i) folgt aus Satz 5.2-2 (vi).
Zum Nachweis der Monotonie sei zunächst $a > 1$. Dann ist $\ln(a) > 0$. Für $x_1 < x_2$ ist $\ln(a) \cdot x_1 < \ln(a) \cdot x_2$ und wegen der strengen Monotonie der Exponentialfunktion $\exp_a(x_1) = \exp(\ln(a) \cdot x_1) < \exp(\ln(a) \cdot x_2) = \exp_a(x_2)$. Ist $a < 1$, dann ist $\ln(a) < 0$ und $x_1 < x_2$ impliziert $\ln(a) \cdot x_1 > \ln(a) \cdot x_2$ und daher
$\exp_a(x_1) = \exp(\ln(a) \cdot x_1) > \exp(\ln(a) \cdot x_2) = \exp_a(x_2)$.

Die letzte Gleichung in Teil (ii) soll verifiziert werden:

$$
\begin{aligned}
\left(\exp_a(x)\right)^y &= \left(\exp\left(\ln(a)\cdot x\right)\right)^y && \text{(nach Definition der Exponentialfunktion zur Basis } a\text{)}\\
&= \exp\left(\ln\left(\exp\left(\ln(a)\cdot x\right)\right)\cdot y\right) && \text{(nach Definition der Exponentialfunktion zur Basis } \exp\left(\ln(a)\cdot x\right)\text{)}\\
&= \exp\left(\ln(a)\cdot x\cdot y\right) && \text{(da der natürliche Logarithmus und die Exponentialfunktion zueinander invers sind)}\\
&= \exp_a(x\cdot y) && \text{(nach Definition der Exponentialfunktion zur Basis } a\text{).}
\end{aligned}
$$

Für $x \in \mathbf{R}$ und $y \in \mathbf{R}$ mit $y > 0$ lässt sich nun auch der Ausdruck y^x sinnvoll definieren:

$$y^x = \exp\left(\ln(y)\cdot x\right) = e^{\ln(y)\cdot x}.$$

Beispielsweise ist $1^x = e^{\ln(1)\cdot x} = e^{0\cdot x} = 1$, und Werte wie $\pi^{\sqrt{2}}$ oder 2^e machen einen Sinn.

Für $a \in \mathbf{R}$ mit $a > 0$ und $a \neq 1$ heißt die zur Exponentialfunktion $\exp_a$ existierende Umkehrfunktion $\overset{-1}{\exp}_a$, die **Logarithmusfunktion zur Basis *a*** und wird mit $\log_a$ bezeichnet:

$$\log_a : \begin{cases}]0,\infty[& \to & \mathbf{R} \\ x & \to & \log_a(x) \end{cases}.$$

Satz 5.5-4:

Es sei $a \in \mathbf{R}$ mit $a > 0$ und $a \neq 1$.

(i) Die Logarithmusfunktion zur Basis *a*

$$\log_a : \begin{cases}]0,\infty[& \to & \mathbf{R} \\ x & \to & \log_a(x) \end{cases}$$

ist streng monoton steigend, bijektiv und stetig und erfüllt die Funktionalgleichung

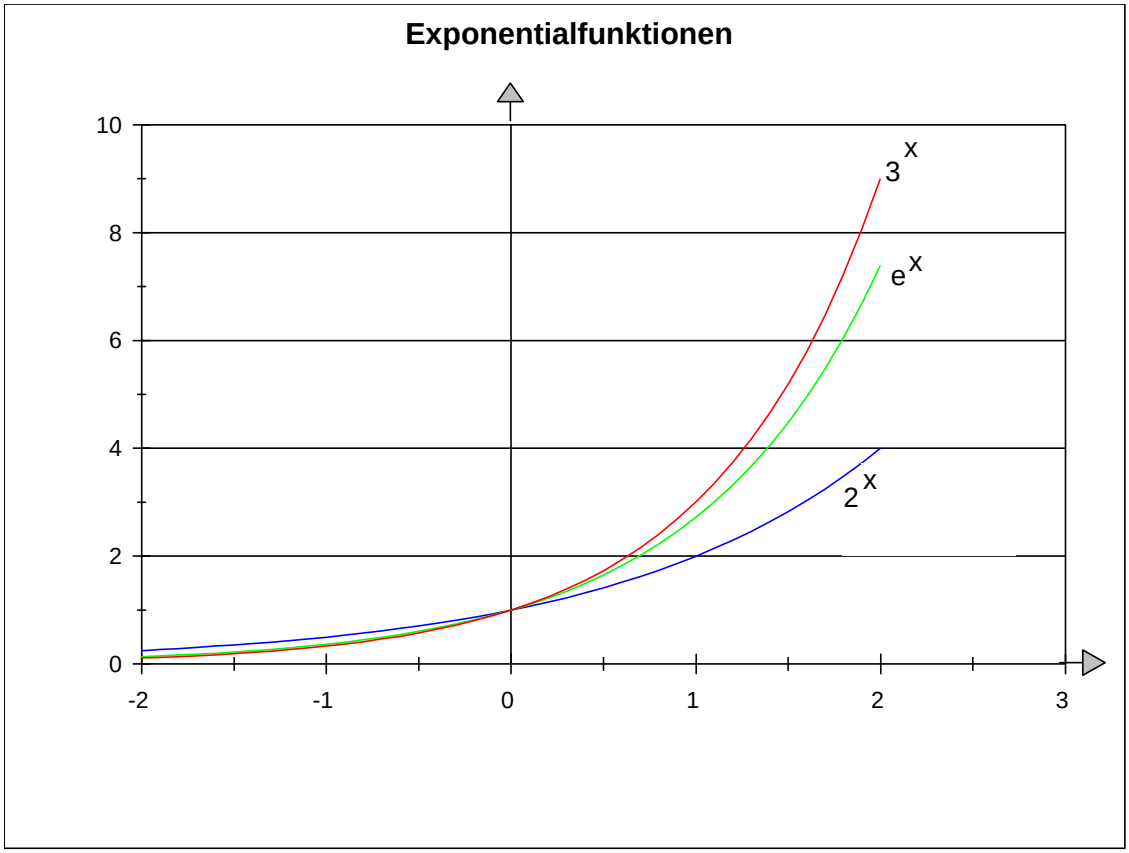
$$\log_a(x\cdot y) = \log_a(x) + \log_a(y).$$

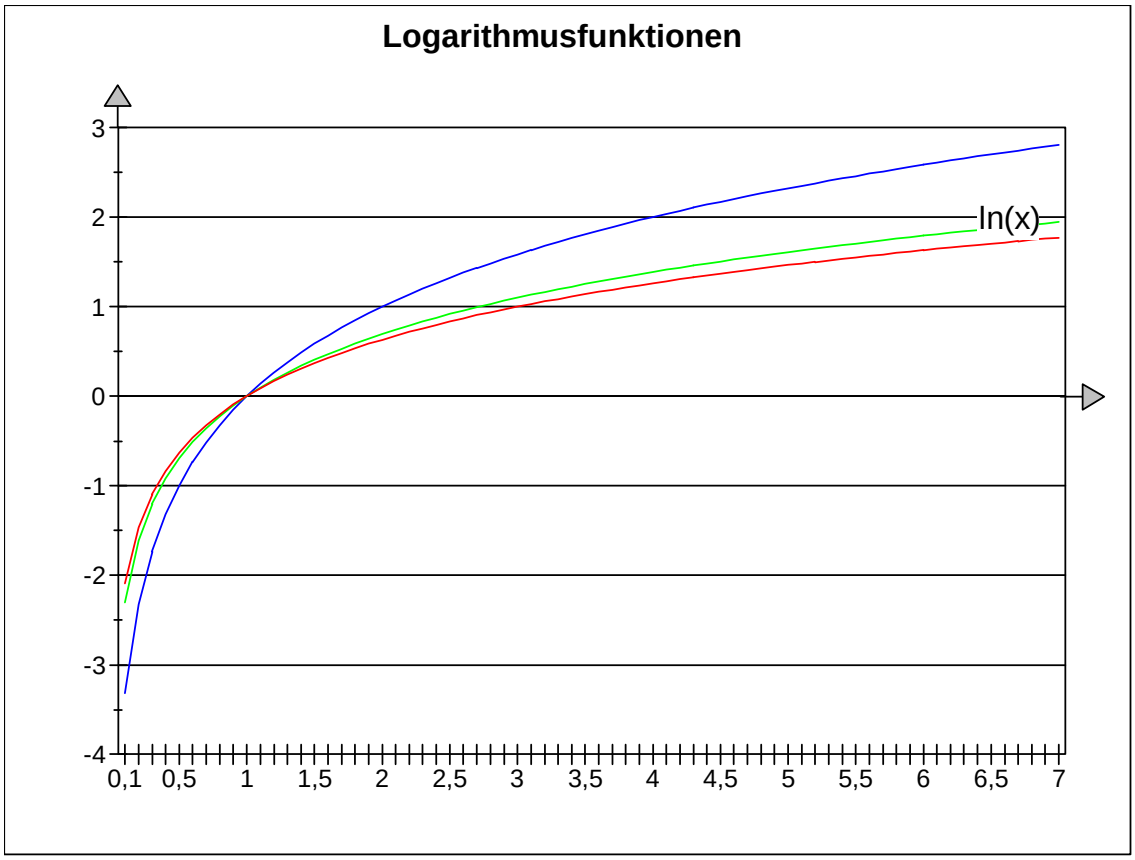
../..

(ii) Es seien $x \in \mathbf{R}_{>0}$ und $y \in \mathbf{R}_{>0}$. Dann gilt

$$\log_a(1) = 0,$$
$$\log_a(a) = 1,$$
$$\log_a(x) < 0 \text{ für } x < 1,\ \log_a(x) > 0 \text{ für } x > 1,$$
$$\log_a(x/y) = \log_a(x) - \log_a(y),$$
$$\log_a(x^m) = m \cdot \log_a(x) \text{ für jedes } m \in \mathbf{Z},$$
$$\log_a(a^x) = x \text{ und } a^{\log_a(x)} = x.$$

Die folgenden Abbildungen zeigen die Verläufe einiger Exponential- und Logarithmusfunktionen.





Die Exponential- und Logarithmusfunktionen zu unterschiedlichen Basen $a \in \mathbf{R}$ mit $a > 0$ und $a \neq 1$ und $b \in \mathbf{R}$ mit $b > 0$ und $b \neq 1$ lassen sich ineinander umrechnen.

Satz 5.5-5:

Es seien $a \in \mathbf{R}$ und $b \in \mathbf{R}$ mit $a > 0$, $a \neq 1$, $b > 0$ und $b \neq 1$.

(i) Den Zusammenhang zwischen verschiedenen Exponentialfunktionen stellt die Gleichung

$$\exp_a(x) = \left(\exp_b(x)\right)^{\log_b(a)} \quad \text{bzw.} \quad a^x = \left(b^x\right)^{\log_b(a)}$$

her. Verschiedene Exponentialfunktionen unterscheiden sich also durch potenzierte Werte.

(ii) Der Zusammenhang zwischen verschiedenen Logarithmusfunktionen wird durch die Gleichung

$$\log_a(x) = \frac{1}{\log_b(a)} \cdot \log_b(x) = \frac{\ln(x)}{\ln(a)}$$

beschrieben. Verschiedene Logarithmusfunktionen unterscheiden sich also durch konstante Faktoren.

(iii) $\log_a\left(b^x\right) = x \cdot \log_a(b)$.

Die Herleitung dieser Gleichungen kann als gute Übung zum Umgang mit Exponential- und Logarithmusausdrücken angesehen werden:

Zunächst wird der zweite Teil der Gleichung in (ii) gezeigt. Diese beschreibt, wie sich $z = \log_a(x)$ mit Hilfe des natürlichen Logarithmus ausdrücken lässt. Aus $z = \log_a(x)$ folgt nacheinander

$\exp_a(z) = x$

$\exp(\ln(a) \cdot z) = x$ (Definition der Exponentialfunktion zur Basis *a*),

$\ln(a) \cdot z = \ln(x)$ (Übergang zur Umkehrfunktion, dem natürlichen Logarithmus),

$z = \log_a(x) = \frac{\ln(x)}{\ln(a)}$.

Daraus folgt direkt der erste Teil der Gleichung in (ii):

$$\log_a(x) = \frac{\ln(x)}{\ln(a)} = \frac{\ln(x)\cdot\ln(b)}{\ln(a)\cdot\ln(b)} = \frac{\ln(x)}{\ln(b)}\cdot\frac{\ln(b)}{\ln(a)} = \log_b(x)\cdot\frac{1}{\log_b(a)}.$$

Die letzte Gleichung entsteht durch Setzen von $x = a$ und $a = b$ in der Formel $\log_a(x) = \frac{\ln(x)}{\ln(a)}$.

Gleichung (i) beschreibt, wie sich die Exponentialfunktion zur Basis *a* durch die Exponentialfunktion zur Basis *b* ausdrücken lässt. Dazu wird die Gleichung

$$\exp_a(x) = \left(\exp_b(y)\right) \text{bzw. } a^x = b^y$$

zunächst auf die ursprüngliche Definition zurückgeführt, dann nach *y* aufgelöst und die Gleichung in (ii) verwendet:

$\exp(\ln(a)\cdot x) = \left(\exp(\ln(b)\cdot y)\right)$ (Definition der Exponentialfunktion zur Basis *a* bzw. zur Basis *b*),

$\ln(a)\cdot x = \ln(b)\cdot y$ (Übergang zur inversen Funktion, dem natürlichen Logarithmus),

$y = \frac{\ln(a)}{\ln(b)}\cdot x = \log_b(a)\cdot x$ (aus (ii))

$\exp_a(x) = \left(\exp_b(\log_b(a)\cdot x)\right) = \left(\exp_b(x)\right)^{\log_b(a)}$ (Satz 5.5-3 (ii)).

Gleichung (iii) ist eine Verallgemeinerung der Gleichung in 5.5-4 (ii) auf alle reellen Zahlen. Mit der Gleichung aus (ii) ergibt sich:

$$\log_a\left(b^x\right) = \frac{\ln\left(b^x\right)}{\ln(a)} = \frac{\ln\left(\exp\left(\ln(b)\cdot x\right)\right)}{\ln(a)} = \frac{\ln(b)\cdot x}{\ln(a)} = x\cdot\log_a(b).$$

Im folgenden sei $a > 1$. Die Exponentialfunktion zur Basis *a* steigt bei wachsendem *x* schnell an. Es gilt nämlich $\exp_a(x+1) = a\cdot\exp_a(x)$ bzw. $a^{x+1} = a\cdot a^x$, d.h. bei Vergrößerung des Argumentwerts um 1 vergrößert sich der Funktionswert um den Faktor *a*.

Hingegen wachsen die entsprechenden Logarithmusfunktionen sehr langsam. Es gilt nämlich $\lim_{x\to\infty}\left(\log_a(x+1) - \log_a(x)\right) = 0$, d.h. obwohl die Logarithmusfunktion bei wachsendem Argumentwert gegen ∞ strebt, nehmen die Funktionswerte letztlich nur noch geringfügig zu:

Es gilt nämlich wegen der Stetigkeit der Logarithmusfunktion (mit Satz 5.2-2):

$$\lim_{n\to\infty}\left(\log_a(n+1)-\log_a(n)\right)=\lim_{n\to\infty}\left(\log_a\left(\frac{n+1}{n}\right)\right)=\log_a\left(\lim_{n\to\infty}\left(\frac{n+1}{n}\right)\right)=\log_a(1)=0\,.$$

Das Wachstumsverhalten der Exponential- und Logarithmusfunktionen im Vergleich mit Polynomen und Wurzelfunktionen zeigt der folgende Satz, dessen Beweis sich aus Überlegungen ergibt, die in Kapitel 5.7 angestellt werden.

Satz 5.5-6:

Es sei $a\in\mathbf{R}, a>1$.

(i) Es sei $p(x)$ ein Polynom. Dann gilt:

$$\lim_{x\to\infty}\frac{|p(x)|}{a^x}=0\,,$$

d.h. die Exponentialfunktionen wachsen schneller als alle Polynome.

(ii) Für jedes $m\in\mathbf{N}$ ist

$$\lim_{x\to\infty}\frac{\left(\log_a(x)\right)^m}{x}=0\,.$$

Man sieht, dass selbst Potenzen von Logarithmusfunktionen im Verhältnis zu Polynomen (sogar zu Polynomen 1. Grades) langsamer wachsen.

(iii) Für jedes $m\in\mathbf{N}$ ist

$$\lim_{x\to\infty}\frac{\log_a(x)}{\sqrt[m]{x}}=0\,.$$

Man sieht, dass Logarithmusfunktionen im Verhältnis zu Wurzelfunktionen langsamer wachsen.

Die folgende Tabelle zeigt fünf Funktionen $h_i : \mathbf{R}_{>0} \to \mathbf{R}$, $i = 1, \ldots, 5$ und einige ausgewählte (gerundete) Funktionswerte.

Spalte 1	Spalte 2	Spalte 3	Spalte 4	Spalte 5
i	$h_i(x)$	$h_i(10)$	$h_i(100)$	$h_i(1000)$
1	$\log_2(x)$	3,3219	6,6439	9,9658
2	$\sqrt{x}$	3,1623	10	31,6228
3	x	10	100	1000
4	x^2	100	10.000	1.000.000
5	2^x	1024	$1{,}2676506 \cdot 10^{30}$	$> 10^{693}$

Die folgende Tabelle zeigt noch einmal die fünf Funktionen $h_i : \mathbf{R}_{>0} \to \mathbf{R}$, $i = 1, \ldots, 5$. Es sei $y_0 > 0$ ein fester Wert. Die dritte Spalte zeigt für jede der fünf Funktionen *x*-Werte x_i mit $h_i(x_i) = y_0$. In der vierten Spalte sind diejenigen *x*-Werte $\overline{x_i}$ aufgeführt, für die $h_i\left(\overline{x_i}\right) = 10 \cdot y_0$ gilt, d.h. dort ist angegeben, auf welchen Wert man x_i vergrößern muss, damit der Funktionswert auf den 10-fachen Wert wächst. Wie man sieht, muss bei der Logartihmusfunktion wegen ihres langsamen Wachstums der *x*-Wert stark vergrößert werden, während bei der schnell anwachsenden Exponentialfunktion nur eine additive konstante Steigerung um ca. 3,3 erforderlich ist.

Spalte 1	Spalte 2	Spalte 3	Spalte 4
i	$h_i(x)$	x_i mit $h_i(x_i) = y_0$	$\overline{x_i}$ mit $h_i\left(\overline{x_i}\right) = 10 \cdot y_0$
1	$\log_2(x)$	x_1	$\left(x_1\right)^{10}$
2	$\sqrt{x}$	x_2	$100 \cdot x_2$
3	x	x_3	$10 \cdot x_3$
4	x^2	x_4	$\approx 3{,}162 \cdot x_4$
5	2^x	x_5	$\approx x_5 + 3{,}322$

Die Logarithmusfunktion zu einer Basis $B > 1$ gibt u.a. näherungsweise an, wieviele Ziffern benötigt werden, um eine natürliche Zahl im Zahlensystem zur Basis *B* darzustellen:

Gegeben sei die Zahl $n \in \mathbf{N}$ mit $n > 0$. Sie benötige $m = m(n, B)$ signifikante Stellen zur Darstellung im Zahlensystem zur Basis *B*, d.h.

$$n = \sum_{i=0}^{m-1} a_i \cdot B^i \text{ mit } a_i \in \{0,\ 1,\ \ldots,\ B-1\} \text{ und } a_{m-1} \neq 0\,.$$

Es ist $B^{m-1} \le n < B^m$ und folglich $m-1 \le \log_B(n) < m$. Daraus ergibt sich für die Anzahl der benötigten Stellen, um eine Zahl *n* im Zahlensystem zur Basis *B* darzustellen,

$$m(n,B) = \lfloor \log_B(n) \rfloor + 1 = \lceil \log_B(n+1) \rceil.$$

Die Anzahl an Dezimalziffern zur Darstellung einer Zahl *n* beträgt demnach $\lfloor \log_{10}(n) \rfloor + 1$, an Binärziffern $\lfloor \log_2(n) \rfloor + 1$ und an Sedezimalziffern $\lfloor \log_{16}(n) \rfloor + 1$.

Die folgende Tabelle zeigt die Zusammenhänge an benötigten Stellen zur Darstellung einer Zahl *n* in den in der Informatik üblichen Zahlensystemen.

Stellenzahl im		
Dezimalsystem $B = 10$	Binärsystem $B = 2$	Sedezimalsystem $B = 16$
m	zwischen $\lfloor c_{10,2} \cdot m \rfloor - 3$ und $\lceil c_{10,2} \cdot m \rceil$ mit $c_{10,2} = 1/\log_{10}(2) \approx 3{,}3219281$	Zwischen $\lfloor c_{10,16} \cdot m \rfloor - 1$ und $\lceil c_{10,16} \cdot m \rceil$ mit $c_{10,16} = 1/\log_{10}(16) \approx 0{,}830482$
Zwischen $\lfloor c_{2,10} \cdot m \rfloor - 1$ und $\lceil c_{2,10} \cdot m \rceil$ mit $c_{2,10} = \log_{10}(2) \approx 0{,}30103$	m	$\left\lceil \frac{m}{4} \right\rceil$
zwischen $\lfloor c_{16,10} \cdot m \rfloor - 1$ und $\lceil c_{16,10} \cdot m \rceil$ mit $c_{16,10} = \log_{10}(16) \approx 1{,}20412$	$4m$	m

Hierbei ist $\lceil x \rceil$ der nach oben auf die nächstgrößere ganze Zahl aufgerundete Wert von *x* und $\lfloor x \rfloor$ der auf die nächstkleinere ganze Zahl abgerundete Wert von *x*.

Werden zwei Zahlen $n_1 \in \mathbf{N}$ und $n_2 \in \mathbf{N}$ mit $n_1 \ge n_2$ addiert, vergrößert sich u.U. die Stellenzahl der Summe im Vergleich zur Stellenzahl von n_1. Ohne die Logarithmusfunktion bemü-

hen zu müssen, kann man die Stellenzahl der Summe $n_1 + n_2$ im Verhältnis zur Stellenzahl von n_1 abschätzen:

Die Zahl n_1 besitze *m* signifikante Stellen, d.h.

$$n_1 = \sum_{i=0}^{m-1} a_i \cdot B^i \text{ mit } a_i \in \{0,1,...,B-1\} \text{ und } a_{m-1} \neq 0 .$$

Der ungünstigste Fall liegt vor, wenn n_1 und n_2 möglichst groß sind, wenn also in n_1 alle Ziffern den Wert $B-1$ haben und $n_1 = n_2$ ist. Dann ist

$$n_1 = \sum_{i=0}^{m-1} (B-1) \cdot B^i = (B-1) \cdot \sum_{i=0}^{m-1} B^i = (B-1) \cdot \frac{B^m - 1}{B-1} = B^m - 1 \text{ und}$$

$$n_1 + n_2 = 2 \cdot \left(B^m - 1\right) = 1 \cdot B^m + \left(B^m - 2\right).$$

Diese Zahl belegt (in der Darstellung im Zahlensystem zur Basis *B*) $m+1$ Stellen:

$$n_1 + n_2 = \left[1 \underbrace{(B-1)...(B-1)}_{(m\text{-}1)\text{-mal}} (B-2) \right]_B .$$

Bei der Addition zweier natürlicher Zahlen nimmt also die Stellenzahl der Summe um höchstens eine Stelle (bezüglich der Stellenzahl der größeren Zahl) zu.

Bei der Multiplikation kann man eine ähnliche Betrachtung durchführen. Wieder liegt der ungünstigste Fall vor, wenn $n_1 = n_2 = B^m - 1$ ist. Dann ist

$$n_1 \cdot n_2 = \left(B^m - 1\right)^2 = B^{2 \cdot m} - 2 \cdot B^m + 1 = B^m \cdot \left(B^m - 2\right) + 1 .$$

Diese Zahl hat folgende Darstellung im Zahlensystem zur Basis *B*:

$$n_1 \cdot n_2 = \left[\underbrace{(B-1)...(B-1)}_{(m\text{-}1)\text{-mal}} (B-2) \underbrace{0...0}_{(m\text{-}1)\text{-mal}} 1 \right]_B ,$$

belegt also $2 \cdot m$ viele Stellen. Bei der Multiplikation zweier natürlicher Zahlen verdoppelt sich die Stellenzahl also höchstens (bezogen auf die Stellenzahl der größeren Zahl).

5.6 Einführung in die Differentialrechnung

Bei der Untersuchung des Kurvenverlaufs einer Funktion *f* ist es häufig notwendig zu wissen, wie sich der Wert von $f(x)$ ändert, wenn man sich von einem festen Wert x_0 „um einen kleinen Betrag“ bis zum Wert $x > x_0$ entfernt. Man vergleicht dabei die Änderung von

$$\Delta y = f(x) - f(x_0)$$

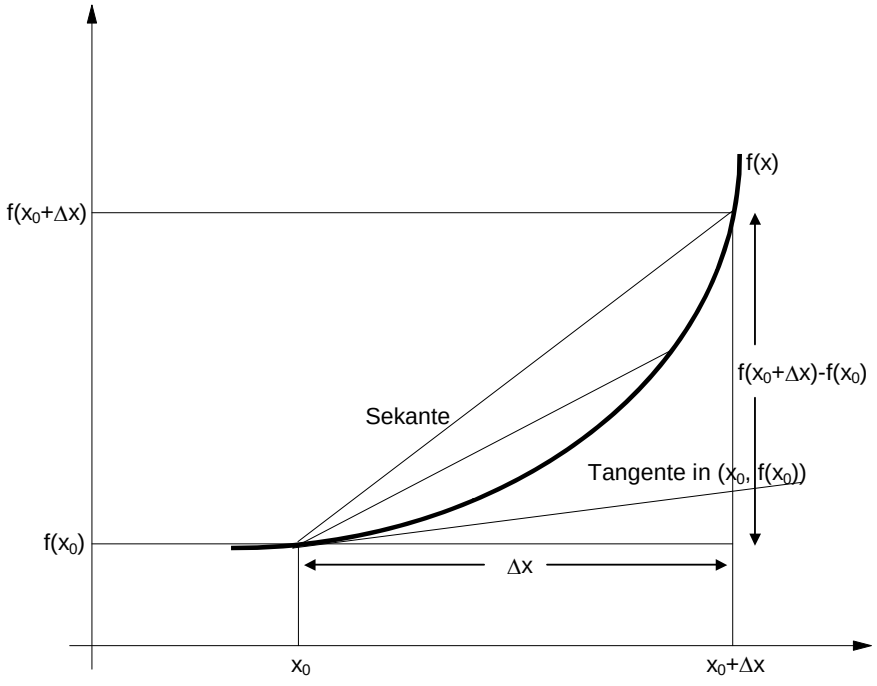
mit der Änderung von

$$\Delta x = x - x_0$$

und bildet den **Differenzenquotienten**

$$\frac{\Delta y}{\Delta x} = \frac{f(x) - f(x_0)}{x - x_0} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}.$$

Geht man „nahe genug“ an x_0 heran, so wird bei vielen Funktionen der Differenzenquotient unabhängig von Δx und beschreibt dann eine charakteristische quantitative Eigenschaft der Funktion *f* im Punkt x_0: die **Steigung der Funktion *f* im Punkt** x_0 .



Im folgenden sei wieder $X \subseteq \mathbf{R}$ und $f: X \to \mathbf{R}$ eine Funktion.

Die Funktion $f: X \to \mathbf{R}$ heißt **an der Stelle** $x_0 \in X$ **differenzierbar**, wenn der Grenzwert

$$\lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

existiert. Dieser Grenzwert heißt **Ableitung von *f* an der Stelle** x_0.

Übliche Schreibweisen für die Ableitung von *f* an der Stelle x_0 sind:

$$\lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x},$$

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

$$\left.\frac{df(x)}{dx}\right|_{x=x_0},$$

$$f'(x_0).$$

Existiert dieser Grenzwert für jedes $x_0 \in X$, so heißt *f* (nach *x*) **differenzierbar**. $f'(x)$ ist eine Funktion von *x*.

Der Differenzenquotient

$$\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

gibt die durchschnittliche Veränderung im Intervall $[x_0, x_0 + \Delta x]$ an und ist von x_0 und Δx abhängig. Er ist gleich der Steigung der Sekante zwischen den Punkten $(x_0, f(x_0))$ und $(x_0 + \Delta x, f(x_0 + \Delta x))$ des Graphen von *f*. Nach dem Grenzübergang $\Delta x \to 0$ ist der Quotient gleich der Steigung der Tangente an den Graphen von *f* im Punkt $(x_0, f(x_0))$ und ist nur von x_0 abhängig.

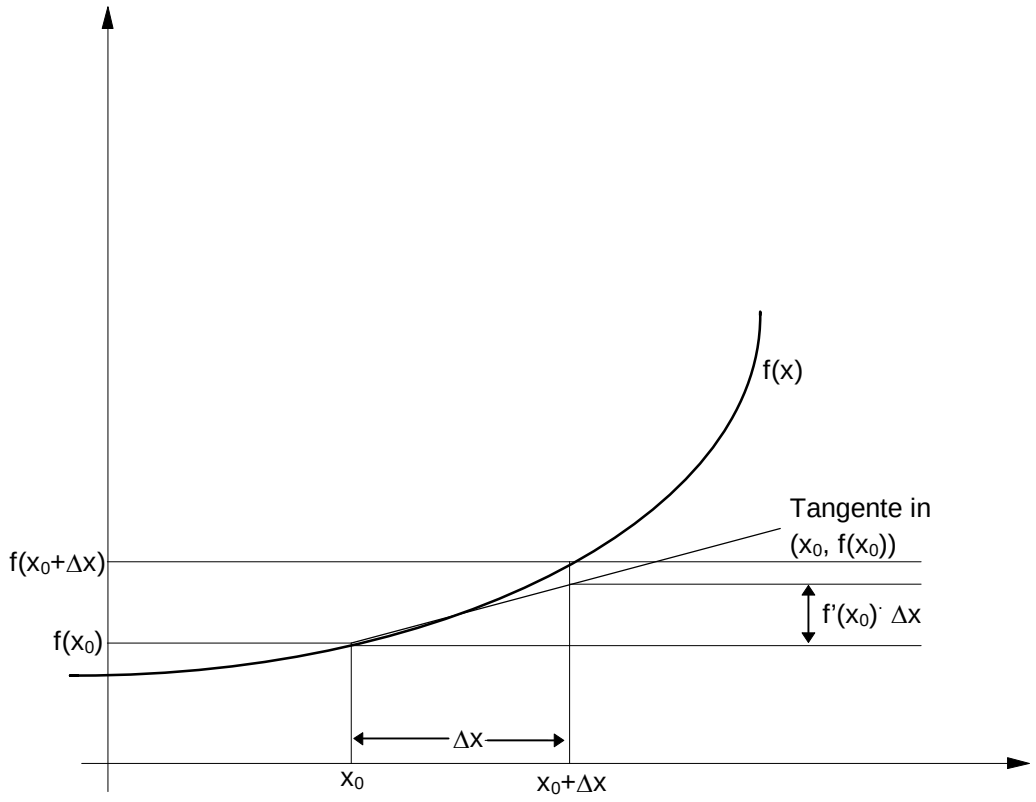
Die Tangente an den Graphen von f im Punkt $\left(x_0, f(x_0)\right)$ hat die Geradengleichung

$$y_T(x) = f'(x_0) \cdot (x - x_0) + f(x_0) .$$

Im Punkt $x_0 + \Delta x$ hat die Tangente also den Wert

$$y_T(x_0 + \Delta x) = f'(x_0) \cdot \Delta x + f(x_0) .$$

Der Wert $f'(x_0) \cdot \Delta x$ gibt also eine ***gute Näherung für die Veränderung von f*** von $f(x_0)$ bis zu $f(x_0 + \Delta x)$, wenn sich x_0 um einen kleinen Wert Δx ändert; diese Änderung ist proportional zu Δx (mit dem Proportionalitätsfaktor $f'(x_0)$).



Satz 5.6-1:

Ist $f: X \to \mathbf{R}$ in $x_0 \in X$ differenzierbar, so ist f in x_0 stetig.

Die Umkehrung gilt im allgemeinen nicht, d.h. aus der Stetigkeit einer Funktion in einem Punkt x_0 folgt i.a. nicht die Differenzierbarkeit in x_0.

Hierzu ist nach Satz 5.2-1 zu zeigen, dass für jede Folge $(x_n)_{n\in\mathbf{N}}$ mit $\lim\limits_{n\to\infty} x_n = x_0$ auch $\lim\limits_{n\to\infty} f(x_n) = f(x_0)$ gilt. Es sei $\varepsilon > 0$ und $n_0 \in \mathbf{N}$, so dass für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ die Ungleichung $|x_n - x_0| < \varepsilon$ gilt. Da f in x_0 differenzierbar ist, existiert der Grenzwert $\lim\limits_{x\to x_0} \dfrac{f(x)-f(x_0)}{x-x_0}$. Für Folgenglieder x_n mit $n \geq n_0$ ist daher $\left|\dfrac{f(x_n)-f(x_0)}{x_n-x_0}\right|$ beschränkt, etwa durch $C > 0$. Es gibt $n_1 \in \mathbf{N}$, so dass für jedes $n \in \mathbf{N}$ mit $n \geq n_1$ die Ungleichung $|x_n - x_0| < \varepsilon/C$ gilt. Für $n \geq \max\{n_0, n_1\}$ ist $|f(x_n) - f(x_0)| \leq C \cdot |x_n - x_0| < \varepsilon$. Also gilt $\lim\limits_{n\to\infty} f(x_n) = f(x_0)$.

Ein Beispiel für eine Funktion, die überall stetig, aber nicht überall differenzierbar ist, ist die Betragsfunktion

$$f_{|\,|}: \begin{cases} \mathbf{R} \to \mathbf{R} \\ x \to |x| \end{cases} .$$

Es gilt

$$\lim_{h\to 0} \frac{f_{|\,|}(x+h) - f_{|\,|}(x)}{h}\bigg|_{x=0} = \lim_{h\to 0} \frac{f_{|\,|}(h)}{h} = \lim_{h\to 0} \frac{|h|}{h} = \begin{cases} +1 & \text{für } h > 0 \\ -1 & \text{für } h < 0 \end{cases}$$

Der Grenzwert existiert also nicht, d.h. f ist in $x_0 = 0$ nicht differenzierbar (aber stetig).

Satz 5.6-2:

Die Funktionen $f: X \to \mathbf{R}$ und $g: X \to \mathbf{R}$ seien differenzierbar. Dann gilt:

(i) $\dfrac{d}{dx}\left(a \cdot f(x) + b \cdot g(x)\right) = a \cdot \dfrac{df(x)}{dx} + b \cdot \dfrac{dg(x)}{dx}$,

$\left(a \cdot f(x) + b \cdot g(x)\right)' = a \cdot f'(x) + b \cdot g'(x)$.

Hierbei sind *a* und *b* Konstanten, die insbesondere nicht von *x* abhängig sind.

(ii) $\dfrac{d}{dx}\left(f(x) \cdot g(x)\right) = \dfrac{df(x)}{dx} \cdot g(x) + f(x) \cdot \dfrac{dg(x)}{dx}$,

$\left(f(x) \cdot g(x)\right)' = f'(x) \cdot g(x) + f(x) \cdot g'(x)$

(**Produktregel**)

(iii) Für $g(x) \neq 0$ gilt:

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{\dfrac{df(x)}{dx} \cdot g(x) - f(x) \cdot \dfrac{dg(x)}{dx}}{\left(g(x)\right)^2},$$

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{\left(g(x)\right)^2}$$

(**Quotientenregel**)

(iv) $\dfrac{d}{dx}\left(f(g(x))\right) = \left.\dfrac{df(y)}{dy}\right|_{y=g(x)} \cdot \dfrac{dg(x)}{dx}$,

$\left(f(g(x))\right)' = f'(g(x)) \cdot g'(x)$

(**Kettenregel**)

(v) Hat f die Umkehrfunktion $\overset{-1}{f}$ und ist $f'(x_0) \neq 0$ für $x_0 \in X$, so ist für $y_0 = f(x_0)$:

$$\left.\frac{d}{dy}\overset{-1}{f}(y)\right|_{y=y_0} = \frac{1}{\left.\dfrac{df(x)}{dx}\right|_{x=x_0}},$$

$$\left[\overset{-1}{f}\left(f(x_0)\right)\right]' = \frac{1}{f'(x_0)}.$$

Teil (i) lässt sich direkt aus der Definition der Differenzierbarkeit ableiten.

Für Teil (ii) wird berechnet:

$$\frac{(f\cdot g)(x_0+\Delta x)-(f\cdot g)(x_0)}{\Delta x}=\frac{f(x_0+\Delta x)\cdot g(x_0+\Delta x)-f(x_0)\cdot g(x_0)}{\Delta x}$$
$$=\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}\cdot g(x_0+\Delta x)+f(x_0)\cdot\frac{g(x_0+\Delta x)-g(x_0)}{\Delta x}\ .$$

Daher ist $(f\cdot g)'(x_0)=\lim\limits_{\Delta x\to 0}\dfrac{(f\cdot g)(x_0+\Delta x)-(f\cdot g)(x_0)}{\Delta x}=f'(x_0)\cdot g(x_0)+f(x_0)\cdot g'(x_0)$.

Der Nachweis für (iii) erfolgt in zwei Schritten:

Mit $\left(\frac{1}{g}\right)(x)=\frac{1}{g(x)}$ ist

$$\frac{\left(\frac{1}{g}\right)(x_0+\Delta x)-\left(\frac{1}{g}\right)(x_0)}{\Delta x}=\frac{\frac{1}{g(x_0+\Delta x)}-\frac{1}{g(x_0)}}{\Delta x}$$
$$=-\frac{1}{g(x_0+\Delta x)\cdot g(x_0)}\cdot\frac{g(x_0+\Delta x)-g(x_0)}{\Delta x}\ .$$

Daher ist $\left(\frac{1}{g}\right)'(x_0)=\lim\limits_{\Delta x\to 0}\dfrac{\left(\frac{1}{g}\right)(x_0+\Delta x)-\left(\frac{1}{g}\right)(x_0)}{\Delta x}=-\frac{1}{(g(x_0))^2}\cdot g'(x_0)$.

Im zweiten Schritt wird (ii) auf das Ergebnis des ersten Schritts angewandt:

$$\left(\frac{f}{g}\right)'=\left(f\cdot\frac{1}{g}\right)'=f'\cdot\frac{1}{g}+f\cdot\left(\frac{1}{g}\right)'=\frac{f'}{g}-\frac{f\cdot g'}{g^2}=\frac{f'\cdot g-f\cdot g'}{g^2}.$$

Zum Nachweis von (iv) wird definiert:

$y_0=g(x_0)$ und

$$f^*(y)=\begin{cases}\dfrac{f(y)-f(y_0)}{y-y_0} & \text{für } y\neq y_0\\ f'(y_0) & \text{für } y=y_0\end{cases}\ .$$

Es ist $f(y)-f(y_0)=(y-y_0)\cdot f^*(y)$ bzw. $f(y)=f(y_0)+(y-y_0)\cdot f^*(y)$.

$$(f\circ g)(x_0+\Delta x)-(f\circ g)(x_0)=f(g(x_0+\Delta x))-f(g(x_0))$$
$$=f(y_0)+(g(x_0+\Delta x)-y_0)\cdot f^*(g(x_0+\Delta x))-f(g(x_0))$$
$$=(g(x_0+\Delta x)-g(x_0))\cdot f^*(g(x_0+\Delta x))\ .$$

Damit ist

$$\lim_{\Delta x\to 0}\frac{(f\circ g)(x_0+\Delta x)-(f\circ g)(x_0)}{\Delta x}=\lim_{\Delta x\to 0}\frac{\big(g(x_0+\Delta x)-g(x_0)\big)\cdot f^*\big(g(x_0+\Delta x)\big)}{\Delta x}$$
$$=g'(x_0)\cdot f^*(y_0)$$
$$=f'\big(g(x_0)\big)\cdot g'(x_0)\ .$$

Für (v) wird $x_1=\overset{-1}{f}(y_0+\Delta y)$ gesetzt. Dann ist gilt für $\Delta y\to 0:\ x_1\to x_0$.

$$\lim_{\Delta y\to 0}\frac{\overset{-1}{f}(y_0+\Delta y)-\overset{-1}{f}(y_0)}{\Delta y}=\lim_{\Delta y\to 0}\frac{x_1-x_0}{(y_0+\Delta y)-y_0}=\lim_{\Delta y\to 0}\frac{1}{\frac{f(x_1)-f(x_0)}{x_1-x_0}}=\lim_{x_1\to x_0}\frac{1}{\frac{f(x_1)-f(x_0)}{x_1-x_0}}=\frac{1}{f'(x_0)}.$$

Für einige grundlegende Beispiele soll die jeweilige Ableitung in einem Punkt x_0 des Definitionsbereichs berechnet werden. Dazu wird entweder der Quotient $\dfrac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$ und dann der Grenzübergang $\Delta x\to 0$ vollzogen oder es werden die Regeln des Satzes 5.5-2 mit bereits bekannten Ableitungen verwendet.

Beispiel:

Für die durch $f(x)=x^n$ mit $n\in\mathbf{N}$ definierte Funktion ist

$$\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}=\frac{(x_0+\Delta x)^n-x_0^n}{\Delta x}$$
$$=\frac{\sum_{i=0}^{n}\binom{n}{i}\cdot x_0^{n-i}\cdot\Delta x^i-x_0^n}{\Delta x}$$
$$=\frac{x_0^n+\Delta x\cdot\sum_{i=1}^{n}\binom{n}{i}\cdot x_0^{n-i}\cdot\Delta x^{i-1}-x_0^n}{\Delta x}$$
$$=\sum_{i=1}^{n}\binom{n}{i}\cdot x_0^{n-i}\cdot\Delta x^{i-1}$$
$$=\underbrace{n\cdot x_0^{n-1}}_{i=1}+\Delta x\cdot\sum_{i=2}^{n}\binom{n}{i}\cdot x_0^{n-i}\cdot\Delta x^{i-2},\ \text{ also}$$

$$\lim_{\Delta x\to 0}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}=n\cdot x_0^{n-1}\ \text{ bzw.}$$

$$\left(x^n\right)' = n \cdot x^{n-1}.$$

Beispiel:

Für die durch $\overset{-1}{f}(x) = \sqrt[n]{x} = x^{1/n}$ mit $n \in \mathbf{N}$ definierte Funktion ist nach Satz 5.6-2 (v) (da $\overset{-1}{f}(x)$ die Umkehrfunktion zu der Funktion $f(x) = x^n$ des vorherigen Beispiels ist) mit $y_0 = f(x_0)$, d.h. $x_0 = \overset{-1}{f}(y_0) = \sqrt[n]{y_0}$,

$$\frac{d}{dy}\overset{-1}{f}(y)\bigg|_{y=y_0} = \frac{1}{\left.\frac{df(x)}{dx}\right|_{x=x_0}} = \frac{1}{n \cdot x_0^{n-1}} = \frac{1}{n \cdot \left(\sqrt[n]{y_0}\right)^{n-1}} = \frac{1}{n \cdot \left(y_0^{1/n}\right)^{n-1}} = \frac{1}{n \cdot y_0^{1-1/n}} = \frac{1}{n \cdot y_0^{-(1/n-1)}} = \frac{1}{n} \cdot y_0^{1/n-1}$$

bzw.

$$\left(x^{1/n}\right)' = \frac{1}{n} \cdot x^{\frac{1}{n}-1}.$$

Beispiel:

Für die durch $g(x) = x^q$ mit $q \in \mathbf{Q}$ und $q > 0$, etwa $q = \frac{n}{m}$ mit $n \in \mathbf{N}$ und $m \in \mathbf{N}_{>0}$, definierte Funktion ist gemäß Kettenregel (Satz 5.6-2 (iv)):

$$\left(x^{\frac{n}{m}}\right)' = \frac{d}{dx}\left(x^{\frac{n}{m}}\right) = \frac{d}{dx}\left(x^n\right)^{\frac{1}{m}} = \frac{1}{m} \cdot \left(x^n\right)^{\frac{1}{m}-1} \cdot n \cdot x^{n-1} = \frac{n}{m} \cdot x^{\frac{n}{m}-1}.$$

Beispiel:

Die Berechnung der Ableitung der Exponentialfunktion $\exp(x) = e^x$ erfolgt wieder direkt über die Definition der Ableitung. Dazu zunächst eine Vorbemerkung: Gemäß Satz 5.1-12 ist für einen kleinen Wert $|\Delta x|$:

$$\exp(\Delta x) = \sum_{i=0}^{1} \frac{\Delta x^i}{i!} + R_2(\Delta x) = 1 + \Delta x + R_2(\Delta x) \text{ mit } |R_2(\Delta x)| \le \frac{2 \cdot |\Delta x|^2}{2!} = |\Delta x|^2.$$

Damit ist

$$\frac{\exp(x_0+\Delta x)-\exp(x_0)}{\Delta x}=\frac{\exp(x_0)\cdot\exp(\Delta x)-\exp(x_0)}{\Delta x}=\exp(x_0)\cdot\frac{\exp(\Delta x)-1}{\Delta x},$$

und der Grenzübergang ergibt

$$\begin{aligned}\lim_{\Delta x\to 0}\frac{\exp(x_0+\Delta x)-\exp(x_0)}{\Delta x}&=\lim_{\Delta x\to 0}\exp(x_0)\cdot\frac{\exp(\Delta x)-1}{\Delta x}\\&=\exp(x_0)\cdot\lim_{\Delta x\to 0}\frac{\exp(\Delta x)-1}{\Delta x}\\&=\exp(x_0)\cdot\lim_{\Delta x\to 0}\frac{(\Delta x+R_2(\Delta x))}{\Delta x}\\&=\exp(x_0)\cdot\left(1+\lim_{\Delta x\to 0}\frac{R_2(\Delta x)}{\Delta x}\right).\end{aligned}$$

Wegen $\left|\frac{R_2(\Delta x)}{\Delta x}\right|\le|\Delta x|$ folgt damit

$$\left(\exp(x)\right)'=\left(e^x\right)'=e^x=\exp(x).$$

Wegen $a^x=\exp(\ln(a)\cdot x)$ ist mit der Kettenregel (Satz 5.6-2 (iv)):

$$\left(\exp_a(x)\right)'=\left(\exp(\ln(a)\cdot x)\right)'=\exp(\ln(a)\cdot x)\cdot\ln(a)=\ln(a)\cdot a^x.$$

Beispiel:

Die Ableitung des natürlichen Logarithmus als Umkehrfunktion der Exponentialfunktion wird wieder mit Hilfe von Satz 5.6-2 (v) berechnet:

Es sei $y_0=\exp(x_0)$, d.h. $x_0=\ln(y_0)$.

$$\frac{d}{dy}\overset{-1}{\exp}(y)\Big|_{y=y_0}=\frac{1}{\frac{d\exp(x)}{dx}\Big|_{x=x_0}}=\frac{1}{\exp(x_0)}=\frac{1}{y_0},\text{ also}$$

$$\left(\ln(x)\right)'=\frac{1}{x}.$$

Die Ableitung der Logarithmusfunktion zu einer Basis $a>1$ lautet nun

$$\left(\log_a(x)\right)' = \left(\frac{\ln(x)}{\ln(a)}\right)' = \frac{1}{\ln(a)\cdot x}.$$

Beispiel:

Es kann nun auch die Ableitung der durch $h(x) = x^r$ mit $r \in \mathbf{R}$ bestimmten Funktion ermittelt mit Hilfe der Kettenregel (Satz 5.6-2 (iv)) werden:

$$\left(x^r\right)' = \left(e^{\ln(x)\cdot r}\right)' = e^{\ln(x)\cdot r}\cdot r\cdot\frac{1}{x} = r\cdot x^r\cdot\frac{1}{x} = r\cdot x^{r-1}.$$

Beispiel:

Die durch $k(x) = \left(x^2+1\right)^x$ gegebene Funktion hat die Eigenschaft, dass x sowohl in der „Basis“ als auch im Exponenten vorkommt. In diesem Fall wendet man den Trick des Logarithmierens mit Satz 5.5-5 (iii) an:

$$\ln\left(k(x)\right) = x\cdot\ln\left(x^2+1\right);$$

die Ableitung der linken Seite lautet:

$$\left(\ln\left(k(x)\right)\right)' = \frac{\left(k(x)\right)'}{k(x)},$$

die Ableitung der rechten Seite lautet:

$$\left(x\cdot\ln\left(x^2+1\right)\right)' = 1\cdot\ln\left(x^2+1\right) + x\cdot\frac{2\cdot x}{x^2+1};$$

beide Seiten werden gleichgesetzt und die entstandene Gleichung nach $\left(k(x)\right)'$ aufgelöst:

$$\frac{\left(k(x)\right)'}{k(x)} = \ln\left(x^2+1\right) + \frac{2\cdot x^2}{x^2+1},$$

$$\left(k(x)\right)' = \left(x^2+1\right)^x\cdot\left(\ln\left(x^2+1\right) + \frac{2\cdot x^2}{x^2+1}\right).$$

Die folgende Tabelle fasst die Ergebnisse der Beispiele zusammen.

$f(x)$	$f'(x)$
x^r, $r \in \mathbf{R}$	$r \cdot x^{r-1}$
$c = const.$	0
$\sum_{i=0}^{n} a_i \cdot x^i$	$\sum_{i=0}^{n} i \cdot a_i \cdot x^{i-1}$
$1/x^n$	$-n/x^{n+1}$
$\sqrt{h(x)}$	$\frac{h'(x)}{2\sqrt{h(x)}}$
$\ln(x) = \log_e(x)$	$1/x$
$\ln(h(x))$	$\frac{h'(x)}{h(x)}$
$\log_a(x)$	$\frac{1}{x \cdot \ln(a)}$
e^x	e^x
a^x, $a > 0$, $a = const.$	$a^x \cdot \ln(a)$
$e^{h(x)}$	$h'(x) \cdot e^{h(x)}$
$a^{h(x)}$, $a > 0$, $a = const.$	$h'(x) \cdot a^{h(x)} \cdot \ln(a)$

Die Funktion $f: X \to \mathbf{R}$ sei differenzierbar (und damit auch stetig). Dann ist $f': X \to \mathbf{R}$ ebenfalls eine Funktion, die aber nicht unbedingt differenzierbar oder stetig sein muss. Ist sie jedoch differenzierbar, so kann man

$$\frac{df'(x)}{dx}$$

bilden und nennt dieses die 2. Ableitung von *f*.

Allgemein werden **Ableitungen höherer Ordnung** wie folgt definiert:

Es ist

$f^{(0)}(x) = f(x)$;

ist die $(n-1)$-te Ableitung der Funktion $f: X \to \mathbf{R}$ im Intervall $I \subseteq X$ differenzierbar, so ist die *n*-te Ableitung von *f* gegeben durch

$$f^{(n)}(x) = \frac{d}{dx} f^{(n-1)}(x).$$

Existieren für *f* alle Ableitungen bis zur *n*-ten Ableitung, so heißt *f* ***n*-mal differenzierbar**.

Beispiel:

Für die durch $f(x) = x \cdot e^x$ definierte Funktion lauten die ersten beiden Ableitungen:

$$f'(x) = 1 \cdot e^x + x \cdot e^x = (1+x) \cdot e^x,$$
$$f''(x) = 1 \cdot e^x + (1+x) \cdot e^x = (2+x) \cdot e^x.$$

Zu vermuten ist, dass die *n*-te Ableitung $f^{(n)}(x) = (n+x) \cdot e^x$ lautet. Für $n = 0,1,2$ stimmt dieses, und die Vermutung gelte für $n \geq 2$. Die $(n+1)$-te Ableitung ist dann

$$f^{(n+1)}(x) = \left((n+x) \cdot e^x\right)' = 1 \cdot e^x + (n+x) \cdot e^x = (n+1+x) \cdot e^x, \text{ d.h.}$$

die Vermutung gilt für jedes $n \in \mathbf{N}$.

Beispiel:

Für das Polynom $p(x) = x^m$ lauten alle Ableitungen:

$$p^{(n)}(x) = \begin{cases} m \cdot (m-1) \cdot \ldots \cdot (m-n+1) \cdot x^{m-n} & \text{für } n \leq m \\ 0 & \text{für } n > m. \end{cases}$$

Ableitungen höherer Ordnung werden insbesondere zur Untersuchung des Kurvenverlaufs von Graphen zu reellen Funktionen (**Kurvendiskussion**) eingesetzt. Diesem Thema ist der Rest des Kapitels gewidmet.

Die Funktion $f: X \to \mathbf{R}$ hat an der Stelle $x_0 \in X$ ein **(lokales) Maximum**, wenn es eine ε-Umgebung $U(x_0, \varepsilon) = \{ x \mid |x - x_0| < \varepsilon \} = \{ x \mid x_0 - \varepsilon < x < x_0 + \varepsilon \}$ von x_0 gibt, so dass für alle $x \in U(x_0, \varepsilon)$ mit $x \neq x_0$ gilt: $f(x) < f(x_0)$.

Die Funktion $f: X \to \mathbf{R}$ hat an der Stelle $x_0 \in X$ ein **(lokales) Minimum**, wenn es eine ε-Umgebung $U(x_0, \varepsilon) = \{ x \mid |x - x_0| < \varepsilon \} = \{ x \mid x_0 - \varepsilon < x < x_0 + \varepsilon \}$ von x_0 gibt, so dass für alle $x \in U(x_0, \varepsilon)$ mit $x \neq x_0$ gilt: $f(x) > f(x_0)$.

Unter einem **(lokalen) Extremwert** versteht man ein lokales Maximum oder ein lokales Minimum.

Satz 5.6-3:

Die Funktion $f: X \to \mathbf{R}$ sei in einer ε-Umgebung $U(x_0, \varepsilon) = \{ x \mid |x - x_0| < \varepsilon \} = \{ x \mid x_0 - \varepsilon < x < x_0 + \varepsilon \}$ von $x_0 \in X$ differenzierbar. Hat f in x_0 einen lokalen Extremwert, so ist $f'(x_0) = 0$.

Diese Aussage lässt sich wie folgt beweisen:
Der Extremwert sei ein Maximum. Dann gilt für $x \in X$ mit $x_0 - \varepsilon < x < x_0$: $x - x_0 < 0$ und $f(x) - f(x_0) < 0$, also $\frac{f(x) - f(x_0)}{x - x_0} > 0$. Für $x \in X$ mit $x_0 < x < x_0 + \varepsilon$ ist $x - x_0 > 0$ und $f(x) - f(x_0) < 0$, also $\frac{f(x) - f(x_0)}{x - x_0} < 0$. Da f bei x_0 differenzierbar ist, existiert der Grenzwert $\lim\limits_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$ und ist gleich $f'(x_0)$. Es ist $f'(x_0) = \lim\limits_{\substack{x \to x_0 \\ x < x_0}} \frac{f(x) - f(x_0)}{x - x_0} \geq 0$ und $f'(x_0) = \lim\limits_{\substack{x \to x_0 \\ x < x_0}} \frac{f(x) - f(x_0)}{x - x_0} \leq 0$, also $f'(x_0) = 0$.

Ist X ein Intervall, dann gilt die Aussage des Satzes 5.6-3 nur für innere Punkte von X, bei denen die Funktion differenzierbar ist. **Kandidaten für Extremwerte** sind daher:

- die Randpunkte von X
- die inneren Punkte von X; hier ist die erste Ableitung gleich 0

– die Punkte von X, in denen die Funktion nicht differenzierbar ist.

Satz 5.6-4:

Die Funktion $f:[a,b]\to\mathbf{R}$ sei stetig und in $]a,b[$ differenzierbar. Weiterhin gelte $f(a)=f(b)=0$. Dann gibt es ein $x_0\in]a,b[$ mit $f'(x_0)=0$.

Für konstantes f ist die Aussage klar. Ansonsten besitzt f in $]a,b[$ einen lokalen Extremwert bei $x_0\in]a,b[$, für den nach Satz 5.6-3 $f'(x_0)=0$ gilt.

Satz 5.6-5: (Mittelwertsätze der Differentialrechnung)

(i) Die Funktion $f:[a,b]\to\mathbf{R}$ mit $a<b$ sei stetig und in $]a,b[$ differenzierbar. Dann gibt es ein $x_0\in]a,b[$ mit

$$f'(x_0)=\frac{f(b)-f(a)}{b-a}.$$

(ii) Die Funktionen $f:[a,b]\to\mathbf{R}$ und $g:[a,b]\to\mathbf{R}$ mit $a<b$ seien stetig und in $]a,b[$ differenzierbar, und es gelte $g'(x)\neq 0$ für $x\in]a,b[$. Dann gibt es ein $x_0\in]a,b[$ mit

$$\frac{f'(x_0)}{g'(x_0)}=\frac{f(b)-f(a)}{g(b)-g(a)}.$$

Teil (i) ist ein Spezialfall von Teil (ii), nämlich mit $g(x)=x$.

Teil (ii) folgt aus Satz 5.6-4:
Zunächst ist $g(b)-g(a)\neq 0$; denn sonst gäbe es nach Satz 5.6-4 ein $x_0\in]a,b[$ mit $g'(x_0)=0$.
Für die Funktion $h:[a,b]\to\mathbf{R}$ mit $h(x)=f(x)-(g(x)-g(a))\cdot\frac{f(b)-f(a)}{g(b)-g(a)}$ gilt $h(a)=f(a)$ und $h(b)=f(b)-(g(b)-g(a))\cdot\frac{f(b)-f(a)}{g(b)-g(a)}=f(a)=h(a)$. Daher gibt es $x_0\in]a,b[$ mit $h'(x_0)=0$. Es ist $h'(x)=f'(x)-g'(x)\cdot\frac{f(b)-f(a)}{g(b)-g(a)}$, also $\frac{f'(x_0)}{g'(x_0)}=\frac{f(b)-f(a)}{g(b)-g(a)}$.

Bemerkung: Den Wert $x_0 \in]a,b[$ kann man in der Form $x_0 = a + \lambda \cdot (b-a)$ mit $0 < \lambda < 1$ schreiben. Dann lautet die Formel in Satz 5.6-5 (i):

$f'(a + \lambda \cdot (b-a)) = \dfrac{f(b) - f(a)}{b-a}$ mit $0 < \lambda < 1$, und äquivalent

$f'(b + (1-\lambda) \cdot (a-b)) = \dfrac{f(a) - f(b)}{a-b}$ mit $0 < 1-\lambda < 1$. Die Voraussetzung $a < b$ kann entfallen: Man setzt $h = b - a$ und erhält für einen Wert λ mit $0 < \lambda < 1$:

$f(a+h) = f(a) + h \cdot f'(a + \lambda \cdot h)$ mit $h \neq 0$ (positiv oder negativ). Damit lässt sich die Aussage in Satz 5.6-5 (i) folgendermaßen formulieren:

Es sei $f : [x_0, x_1] \to \mathbf{R}$ mit $x_0 < x_1$ stetig und in $]x_0, x_1[$ differenzierbar. Es gelte $a \in [x_0, x_1]$ und $a + h \in [x_0, x_1]$ mit $h \neq 0$. Dann gibt es einen Wert λ mit $0 < \lambda < 1$ und $f(a+h) = f(a) + h \cdot f'(a + \lambda \cdot h)$.

Die Formel in Satz 5.6-5 (ii) lautet entsprechend:

$f(a+h) - f(a) = \dfrac{g(a+h) - g(a)}{g'(a + \lambda \cdot h)} \cdot f'(a + \lambda \cdot h)$ mit $0 < \lambda < 1$.

Satz 5.6-6:

Ist die Funktion $f : X \to \mathbf{R}$ im Intervall $I \subseteq X$ differenzierbar, so sind folgende Aussagen (a) und (b) gleichbedeutend:

(a) f ist in I monoton fallend (bzw. steigend).
und
(b) Für jedes $x \in I$ gilt $f'(x) \leq 0$ (bzw. $f'(x) \geq 0$).

Für „monoton fallend“ kann man wie folgt argumentieren:

Für $x \in I$ und $x + \Delta x \in I$ ist $f(x + \Delta x) - f(x) = \begin{cases} \leq 0 & \text{für } \Delta x > 0 \\ \geq 0 & \text{für } \Delta x < 0 \end{cases}$. Daher ist

$f'(x) = \lim\limits_{\Delta x \to 0} \dfrac{f(x + \Delta x) - f(x)}{\Delta x} \leq 0$.

Es gelte umgekehrt $f'(x) \leq 0$ für jedes $x \in I$. Es seien $x_1 \in I$ und $x_2 \in I$ mit $x_1 < x_2$. Dann gibt es gemäß Satz 5.6-5(i) ein $x_0 \in I$ mit $x_1 < x_0 < x_2$ und $f'(x_0) = \dfrac{f(x_2) - f(x_1)}{x_2 - x_1} \leq 0$. Daher ist $f(x_2) - f(x_1) \leq 0$ bzw. $f(x_2) \leq f(x_1)$, d.h. f ist monoton fallend.

Satz 5.6-7:

Es seien $f: X \to \mathbf{R}$ und $g: X \to \mathbf{R}$ im Intervall $I \subseteq X$ differenzierbare Funktionen.

(i) Gilt $f'(x) = 0$ für jedes $x \in I$, dann ist f auf I konstant.

(ii) Es gelte $f'(x) = g'(x)$ für jedes $x \in I$. Dann gibt es eine Konstante $C \in \mathbf{R}$ mit $f(x) = g(x) + C$ für jedes $x \in I$.

Für (i) argumentiert man wieder mit Satz 5.6-5:
Angenommen f ist auf I nicht konstant. Dann gibt es $x_1 \in I$ und $x_2 \in I$ mit $x_1 < x_2$ und $f(x_1) \neq f(x_2)$. Gemäß Satz 5.6-5(i) gibt es ein $x_0 \in I$ mit $x_1 < x_0 < x_2$ und $f'(x_0) = \dfrac{f(x_2) - f(x_1)}{x_2 - x_1}$. Da nach Voraussetzung $f'(x_0) = 0$ gilt, folgt aber $f(x_1) = f(x_2)$.

Teil (ii) folgt aus (i); denn die Funktion $f - g$ hat die Ableitung 0.

Zu Beginn des Kapitels 5.2 wird der Begriff Konvexität bzw. Konkavität einer Funktion auf einem Intervall I eingeführt. Anschaulich bedeutet Konvexität einer Funktion f, dass der Graph der Funktion mit wachsenden x-Werten linksgekrümmt ist. Das besagt, dass die Steigung der Tangente an den Graphen in jedem Punkt $(x, f(x))$, d.h. die Ableitung $f'(x)$, monoton wächst. Entsprechend bedeutet Konkavität, dass die Ableitung $f'(x)$ monoton fällt. Ähnlich wie bei Satz 5.6-6 kann man die Aussagen des folgenden Satzes begründen.

Satz 5.6-8:

Ist die Funktion $f: X \to \mathbf{R}$ im Intervall $I \subseteq X$ zweimal differenzierbar, so sind folgende Aussagen (a) bis (c) gleichbedeutend:

(a) f ist in I konvex (bzw. konkav).
und
(b) Die Ableitung $f'(x)$ ist in I monoton steigend (bzw. monoton fallend).
und
(c) Für jedes $x \in I$ ist $f''(x) \geq 0$ (bzw. $f''(x) \leq 0$).

Beispiel:

Zu untersuchen ist das Krümmungsverhalten des durch

$$p(x) = -\frac{1}{10} \cdot x^5 + x^3$$

definierten Polynoms. Seine zweite Ableitung lautet

$$p''(x) = -2 \cdot x^3 + 6 \cdot x = -2 \cdot x \cdot \left(x^2 - 3\right).$$

Es ist

$p''(x) \le 0$ genau dann wenn $(x \ge 0) \wedge \left(x^2 - 3 \ge 0\right)$ oder $(x \le 0) \wedge \left(x^2 - 3 \le 0\right)$ gilt. Im ersten Fall gilt $(x \ge 0) \wedge \left((x \ge \sqrt{3}) \vee (x \le -\sqrt{3})\right)$, also $x \ge \sqrt{3}$. Im zweiten Fall ist $(x \le 0) \wedge \left(\left(-\sqrt{3} \le x \le \sqrt{3}\right)\right)$, also $-\sqrt{3} \le x \le 0$. Für $x \ge \sqrt{3}$ oder für $-\sqrt{3} \le x \le 0$ ist also p konkav, für alle übrigen Bereiche konvex.

Die Funktion $f: X \to \mathbf{R}$ hat an der Stelle x_W einen **Wendepunkt**, wenn es ε-Umgebung $U(x_W, \varepsilon) = \{x \mid |x - x_W| < \varepsilon\} = \{x \mid x_W - \varepsilon < x < x_W + \varepsilon\}$ von x_W gibt, so dass f für jedes $x \in U(x_W, \varepsilon)$ mit $x_W - \varepsilon < x < x_W$ streng konvex und für jedes $x \in U(x_W, \varepsilon)$ mit $x_W < x < x_W + \varepsilon$ streng konkav ist bzw. für jedes $x \in U(x_W, \varepsilon)$ mit $x_W - \varepsilon < x < x_W$ streng konkav und für jedes $x \in U(x_W, \varepsilon)$ mit $x_W < x < x_W + \varepsilon$ streng konvex ist.

Aus der Definition und Satz 5.6-8 folgt unmittelbar:

Satz 5.6-9:

Die Funktion $f: X \to \mathbf{R}$ sei in einer ε-Umgebung $U(x_0, \varepsilon) = \{x \mid |x - x_0| < \varepsilon\} = \{x \mid x_0 - \varepsilon < x < x_0 + \varepsilon\}$ von $x_0 \in X$ zweimal differenzierbar. Hat f in x_0 einen Wendepunkt, so ist $f''(x_0) = 0$.

Bei der Untersuchung des Funktionsverlaufs einer Funktion $f: X \to \mathbf{R}$ einem Intervall I (**Kurvendiskussion**) sind meist folgende Werte von Interesse:

- der Definitonsbereich (ist f für alle $x \in I$ definiert?)
- die Nullstellen von f in I
- die Unstetigkeitsstellen von f in I
- die Stellen, an denen f differenzierbar ist
- die Extremwerte von f in I
- die Wendepunkte von f in I
- das Krümmungsverhalten (konvex/konkav) von f in I

Zur Untersuchung der Extremwerte reicht es nicht aus, allein diejenigen Werte x zu bestimmen, für die $f'(x) = 0$ gilt. Gemäß Satz 5.6-3 hat wohl bei einem Extremwert die erste Ableitung den Wert 0, aber nicht jedes x mit $f'(x) = 0$ ist ein Extremwert, wie folgendes Beispiel zeigt:

Für die Funktion f mit $f(x) = x^3$ gilt $f'(0) = 0$, aber bei 0 liegt kein Extremwert, sondern ein Wendpunkt (mit waagerechter Tangente) vor.

Entsprechend reicht es zur Untersuchung der Wendepunkte nicht aus, allein diejenigen Werte x zu bestimmen, für die $f''(x) = 0$ gilt. Gemäß Satz 5.6-9 bei einem Wendepunkt die zweite Ableitung den Wert 0, aber nicht jedes x mit $f''(x) = 0$ ist ein Wendpunkt, wie folgendes Beispiel zeigt:

Für die Funktion f mit $f(x) = x^4$ gilt $f''(0) = 0$, aber bei 0 liegt kein Wendepunkt, sondern Extremwert vor.

Der folgende Satz, der hier ohne Beweis angeführt wird, gibt Auskunft über die Bestimmung von Extremwerten und Wendepunkte.

Satz 5.6-10:

Die Funktion $f: X \to \mathbf{R}$ sei an der Stelle $x_0 \in X$ mindestens *n*-mal differenzierbar.

Ist

$f^{(k)}(x_0) = 0$ für $k = 1, \ldots, n-1$ und
$f^{(n)}(x_0) \neq 0$,
und ist *n gerade*,

so hat *f* an der Stelle x_0 einen Extremwert, und zwar ein (lokales) Maximum, wenn $f^{(n)}(x_0) < 0$ ist bzw. ein (lokales) Minimum, wenn $f^{(n)}(x_0) > 0$ ist.

Ist

$f^{(k)}(x_0) = 0$ für $k = 2, \ldots, n-1$ und
$f^{(n)}(x_0) \neq 0$
und ist *n ungerade*,

so hat *f* an der Stelle einen Wendepunkt; die Krümmung wechselt von konvex nach konkav, wenn $f^{(n)}(x_0) < 0$ ist; sie wechselt von konkav nach konvex, wenn $f^{(n)}(x_0) > 0$ ist. Gilt zusätzlich $f'(x_0) = 0$, so liegt ein Wendepunkt mit waagerechter Tangente (**Sattelpunkt**) vor.

Das folgende Beispiel untersucht den Kurvenverlauf des Graphen zum Polynom, das durch

$$p(x) = 0{,}2 \cdot x^5 - x^3 + 1$$

definiert wird. Die Ableitungen lauten:

$p'(x) = x^4 - 3 \cdot x^2$,
$p''(x) = 4 \cdot x^3 - 6 \cdot x$,
$p'''(x) = 12 \cdot x^2 - 6$,
$p^{(4)}(x) = 24 \cdot x$,
$p^{(5)}(x) = 24$,

$p^{(6)}(x) = 0$.

Die erste Ableitung $p'(x)$ hat die Nullstellen $x_{0,1} = 0$, $x_{0,2} = \sqrt{3}$, $x_{0,3} = -\sqrt{3}$. Diese Werte werden in die höheren Ableitungen eingesetzt:

$p''(x_{0,1}) = p''(0) = 0$, $p'''(x_{0,1}) = p'''(0) = -6$, also liegt hier ein Wendeüunkt mit waagerechter Tangente (Sattelpunkt) vor.

$p''(x_{0,2}) = p''(\sqrt{3}) = 6 \cdot \sqrt{3} > 0$, also liegt hier ein lokales Minimum vor.

$p''(x_{0,3}) = p''(-\sqrt{3}) = -6 \cdot \sqrt{3} < 0$, also liegt hier ein lokales Maximum vor.

Die Nullstellen der zweiten Ableitung lauten $x_{0,4} = x_{0,1} = 0$, $x_{0,5} = \sqrt{3/2}$, $x_{0,6} = -\sqrt{3/2}$. Die Werte $x_{0,5}$ und $x_{0,6}$ in die dritte Ableitung eingesetzt ergeben $p'''(x_{0,5}) = p'''(x_{0,6}) = 12$. Es liegen also an diesen Werten Wendepunkte vor.

5.7 Die Regeln von de l'Hospital

Häufig sind Grenzwerte der Form

$$\lim_{x \to x_0} f(x)$$

zu berechnen, wobei

$$f(x) = \frac{g(x)}{h(x)} \text{ und } \lim_{x \to x_0} g(x) = \lim_{x \to x_0} h(x) = 0$$

gelten. Hierbei muss x_0 nicht im Definitionsbereich von *f* liegen bzw. es wird auch ∞ anstelle von x_0 zugelassen. In diesen Fällen sind die folgenden Sätze von Bedeutung (**Regeln von de l'Hospital**, 1661-1704):

Satz 5.7-1:

Gegeben seien die Funktionen $g: X \to \mathbf{R}$ und $h: X \to \mathbf{R}$, $X \subseteq \mathbf{R}$. Für $x_0 \in \mathbf{R}$ und $\varepsilon > 0$ bezeichne $\widetilde{U}_\varepsilon$ die Menge $\{x \mid 0 < |x - x_0| < \varepsilon\}$ (ε-Umgebung von x_0 mit Ausnahme von x_0) bzw. das Intervall $\{x \mid a < x < x_0\}$ bzw. das Intervall $\{x \mid x_0 < x < b\}$, wobei für *a* oder *b* die reelle Zahl ε oder auch ∞ stehen können. Die Funktionen *g* und *h* seien in $\widetilde{U}_\varepsilon$ definiert und (*n* + 1)-mal differenzierbar. Ferner gelte

$$(*) \quad \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} g(x_0) = \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} g'(x_0) = \ldots = \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} g^{(n)}(x_0) = 0,$$

$$\lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} h(x_0) = \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} h'(x_0) = \ldots = \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} h^{(n)}(x_0) = 0 \text{ und } h^{(n+1)}(x) \neq 0 \text{ für } x \in \widetilde{U}_\varepsilon.$$

Dann gilt:

Existiert $\lim\limits_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} \dfrac{g^{(n+1)}(x)}{h^{(n+1)}(x)}$, dann ist

$$\lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} \frac{g(x)}{h(x)} = \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} \frac{g'(x)}{h'(x)} = \ldots = \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} \frac{g^{(n)}(x)}{h^{(n)}(x)} = \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} \frac{g^{(n+1)}(x)}{h^{(n+1)}(x)} \quad .$$

Es werden also Zähler- und Nennerfunktion getrennt abgeleitet.

Die Gültigkeit des Satzes soll hier für $n = 0$ gezeigt werden. Durch wiederholte Anwendung folgt dann die im Satz formulierte allgemeine Aussage.

Setzt man $G(x) = \begin{cases} 0 & \text{für } x = x_0 \\ g(x) & \text{für } x \neq x_0 \end{cases}$ und $H(x) = \begin{cases} 0 & \text{für } x = x_0 \\ h(x) & \text{für } x \neq x_0 \end{cases}$, dann folgt mit Satz 5.6-5(ii) (und den anschließenden Bemerkungen):

$\dfrac{G(x)}{H(x)} = \dfrac{G(x) - G(x_0)}{H(x) - H(x_0)} = \dfrac{G'(x_0 + \lambda \cdot (x - x_0))}{H'(x_0 + \lambda \cdot (x - x_0))}$ für einen Wert λ mit $0 < \lambda < 1$. Mit $x \to x_0$ geht $x_0 + \lambda \cdot (x - x_0) \to x_0$. Daher ist

$$\lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} \frac{g(x)}{h(x)} = \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} \frac{G(x)}{H(x)} = \lim_{\substack{x_0 + \lambda \cdot (x - x_0) \to x_0 \\ x_0 + \lambda \cdot (x - x_0) \in \widetilde{U}_\varepsilon}} \frac{G'(x_0 + \lambda \cdot (x - x_0))}{H'(x_0 + \lambda \cdot (x - x_0))} = \lim_{\substack{x \to x_0 \\ x \in \widetilde{U}_\varepsilon}} \frac{g'(x)}{h'(x)}.$$

Beispiele:

$$\lim_{x\to 1}\frac{x^n-1}{x-1}=\lim_{x\to 1}\frac{n\cdot x^{n-1}}{1}=n\,.$$

Dieses Ergebnis erhält man natürlich auch aus der Gleichung aus Satz 5.3-1:

$$x^n-1=(x-1)\cdot\sum_{i=0}^{n-1}x^i\,,\ \text{also}\ \lim_{x\to 1}\frac{x^n-1}{x-1}=\lim_{x\to 1}\sum_{i=0}^{n-1}x^i=n\,.$$

$$\lim_{x\to 0}\frac{e^x-1}{x}=\lim_{x\to 0}\frac{e^x}{1}=1\,.$$

Auch dieses Ergebnis lässt sich anders erzielen:

$$e^x=\exp(x)=\sum_{i=0}^{\infty}\frac{x^i}{i!}=1+\sum_{i=1}^{\infty}\frac{x^i}{i!}=1+x\cdot\sum_{i=1}^{\infty}\frac{x^{i-1}}{i!}\ \text{ und}$$

$$\frac{e^x-1}{x}=\frac{x\cdot\sum_{i=1}^{\infty}\frac{x^{i-1}}{i!}}{x}=\sum_{i=0}^{\infty}\frac{x^i}{(i+1)!}=1+\sum_{i=1}^{\infty}\frac{x^i}{(i+1)!}\,,\ \text{also}\ \lim_{x\to 0}\frac{e^x-1}{x}=\lim_{x\to 0}\left(1+\sum_{i=1}^{\infty}\frac{x^i}{(i+1)!}\right)=1\,.$$

Satz 5.7-1 gilt auch für Grenzwerte der Form $\lim\limits_{x\to\infty}\dfrac{g(x)}{h(x)}$ und $\lim\limits_{x\to-\infty}\dfrac{g(x)}{h(x)}$ (hier nur für den ersten Fall formuliert):

Satz 5.7-2:

Gegeben seien die Funktionen $g:X\to\mathbf{R}$ und $h:X\to\mathbf{R}$, $X\subseteq\mathbf{R}$. Sie seien im Intervall $I=\{x\mid a<x<\infty\}$ für eine reelle Zahl a definiert und $(n+1)$-mal differenzierbar. Ferner gelte

(∗) $\lim\limits_{x\to\infty}g(x_0)=\lim\limits_{x\to\infty}g'(x_0)=\ldots=\lim\limits_{x\to\infty}g^{(n)}(x_0)=0\,,$

$\lim\limits_{x\to\infty}h(x_0)=\lim\limits_{x\to\infty}h'(x_0)=\ldots=\lim\limits_{x\to\infty}h^{(n)}(x_0)=0$ und $h^{(n+1)}(x)\neq 0$ für $x\in I$.

Dann gilt:

Existiert $\lim\limits_{x\to\infty}\dfrac{g^{(n+1)}(x)}{h^{(n+1)}(x)}$, dann ist

$$\lim_{x\to\infty}\frac{g(x)}{h(x)}=\lim_{x\to\infty}\frac{g'(x)}{h'(x)}=\ldots=\lim_{x\to\infty}\frac{g^{(n)}(x)}{h^{(n)}(x)}=\lim_{x\to\infty}\frac{g^{(n+1)}(x)}{h^{(n+1)}(x)}\quad.$$

Die Aussage des Satzes (hier nur wieder für $n = 0$ ausgeführt) erhält man, indem man die Variablentransformation $t = 1/x$ durchführt. Ein Grenzübergang $x \to \infty$ entspricht dann einem Grenzübergang $t \to 0$ mit $t > 0$:

$$\lim_{x\to\infty} \frac{g(x)}{h(x)} = \lim_{\substack{t\to 0\\ t>0}} \frac{g(1/t)}{h(1/t)}$$

$$= \lim_{\substack{t\to 0\\ t>0}} \frac{\frac{d}{dt} g(1/t)}{\frac{d}{dt} h(1/t)} \qquad \text{(gemäß Satz 5.7 - 1)}$$

$$= \lim_{\substack{t\to 0\\ t>0}} \frac{-g'(1/t)\cdot t^{-2}}{-h'(1/t)\cdot t^{-2}} = \lim_{\substack{t\to 0\\ t>0}} \frac{g'(1/t)}{h'(1/t)} = \lim_{x\to\infty} \frac{g'(x)}{h'(x)} \quad .$$

Die hier zitierten Sätze stehen für den „Fall 0/0". Sie gelten auch, wenn die Bedingung $(*)$ durch

$$(**) \quad \lim_{\substack{x\to x_0\\ x\in\tilde{U}_\varepsilon}} g(x_0) = \lim_{\substack{x\to x_0\\ x\in\tilde{U}_\varepsilon}} g'(x_0) = \ldots = \lim_{\substack{x\to x_0\\ x\in\tilde{U}_\varepsilon}} g^{(n)}(x_0) = \infty ,$$

$$\lim_{\substack{x\to x_0\\ x\in\tilde{U}_\varepsilon}} h(x_0) = \lim_{\substack{x\to x_0\\ x\in\tilde{U}_\varepsilon}} h'(x_0) = \ldots = \lim_{\substack{x\to x_0\\ x\in\tilde{U}_\varepsilon}} h^{(n)}(x_0) = \infty \text{ und } h^{(n+1)}(x) \neq 0 \text{ für } x \in \tilde{U}_\varepsilon .$$

ersetzt wird („Fall ∞/∞").

Die Regeln von de l'Hospital ermöglicht auch die Berechnung unbestimmter Ausdrücke der Art $\infty - \infty$, $0\cdot\infty$, 1^∞, ∞^0, 0^0. Diese werden zunächst so umgeformt, dass der „Fall 0/0" oder der „Fall ∞/∞" entsteht. Die folgende Tabelle gibt die Umformungen auf den „Fall 0/0" an:

Typ	Funktion	Umformung	„Fall 0/0“
$\infty-\infty$	$g(x)-h(x)$	-	$\frac{1/h(x)-1/g(x)}{1/h(x)\cdot 1/g(x)}$
$\infty-\infty$	$g(x)-h(x)$	Exponentieren	$\frac{1/e^{h(x)}}{1/e^{g(x)}}$
$0\cdot\infty$	$g(x)\cdot h(x)$	-	$\frac{g(x)}{1/h(x)}$
1^{∞}	$g(x)^{h(x)}$	Logarithmieren	$\frac{\ln(g(x))}{1/h(x)}$
∞^{0}	$g(x)^{h(x)}$	Logarithmieren	$\frac{h(x)}{1/\ln(g(x))}$
0^{0}	$g(x)^{h(x)}$	Logarithmieren	$\frac{h(x)}{1/\ln(g(x))}$

Beispiel:

Es soll $\lim\limits_{x\to 0}(1+x)^{1/x}$ bestimmt werden. Dieser Grenzwert ist vom Typ „1^{∞}“ mit den Funktionen $g(x)=1+x$ und $h(x)=1/x$. Logarithmieren der gesamten Funktion ergibt

$$\ln\left(g(x)^{h(x)}\right)=h(x)\cdot\ln\left(g(x)\right)=\frac{\ln(g(x))}{1/h(x)}=\frac{\ln(1+x)}{x}.$$

Jetzt liegt der „Fall 0/0“ vor:

$$\lim_{x\to 0}\frac{\ln(1+x)}{x}=\lim_{x\to 0}\frac{1/(1+x)}{1}=1;$$

die Logarithmierung wird durch Exponentiation wieder rückgängig gemacht, also

$$\lim_{x\to 0}(1+x)^{1/x}=e^{1}=e.$$

Beispiel:

In Satz 5.5-6 (ii) wird formuliert, dass für jedes $a \in \mathbf{R}$ mit $a > 1$ und jedes Polynom $p(x)$

$$\lim_{x\to\infty} \frac{|p(x)|}{a^x} = 0$$

gilt. Mit Hilfe der Regeln von de l'Hospital für den „Fall ∞ / ∞" lässt sich dieses verifizieren:

Es sei $p(x)$ ein Polynom vom Grade *n*, d.h. $p(x) = \sum_{i=0}^{n} a_i \cdot x^i$ mit $a_n \neq 0$. Es erfolgt eine Beschränkung auf den Fall $p(x) \geq 0$, so dass in der Limesbetrachtung auf die Betragsstriche verzichtet werden kann. Dann ist

$$\lim_{x\to\infty} \frac{p(x)}{a^x} = \lim_{x\to\infty} \frac{(p(x))^{(n)}}{(a^x)^{(n)}} = \lim_{x\to\infty} \frac{n! \cdot a_n}{a^x \cdot (\ln(a))^n} = 0 .$$

Beispiel:

In Satz 5.5-6 (ii) wird formuliert, dass für jedes $a \in \mathbf{R}$ mit $a > 1$ und $m \in \mathbf{N}$

$$\lim_{x\to\infty} \frac{(\log_a(x))^m}{x} = 0$$

gilt. Dieser Grenzwert ist ebenfalls ein Beispiel für den „Fall ∞ / ∞"; er wird verifiziert durch

$$\begin{aligned}
\lim_{x\to\infty} \frac{(\log_a(x))^m}{x} &= \lim_{x\to\infty} \frac{m \cdot (\log_a(x))^{m-1} \Big/ x \cdot \ln(a)}{1} = \lim_{x\to\infty} \frac{m \cdot (\log_a(x))^{m-1}}{x \cdot \ln(a)} \\
&= \lim_{x\to\infty} \frac{m \cdot (m-1) \cdot (\log_a(x))^{m-2} \Big/ x \cdot \ln(a)}{\ln(a)} = \lim_{x\to\infty} \frac{m \cdot (m-1) \cdot (\log_a(x))^{m-2}}{x \cdot (\ln(a))^2} \\
&= \ldots = \lim_{x\to\infty} \frac{m \cdot (m-1) \cdot \ldots \cdot (m-(m-2)) \cdot (\log_a(x))^{m-(m-1)}}{x \cdot (\ln(a))^{m-1}} \\
&= \lim_{x\to\infty} \frac{m!}{x \cdot (\ln(a))^m} = 0 \quad .
\end{aligned}$$

5.8 Das Newton-Verfahren

Bei der Lösung von Gleichungen kommt es häufig vor, dass eine explizite Auflösung nach der unbekannten Größe nicht möglich ist. Man ist dann an einer numerischen Lösung interessiert. Ähnliche Probleme ergeben sich bei der numerischen Bestimmung von Nullstellen von Funktionen.

Gegeben sei eine Funktion $f: X \to \mathbf{R}$, die auf einem Intervall $I = [a, b]$, $I \subseteq X$ mindestens zweimal differenzierbar mit stetiger 2. Ableitung ist. Außerdem seien folgende Bedingungen 1. bis 4. erfüllt:

1. $f(a) \cdot f(b) < 0$, d.h. f hat im Intervall I eine Nullstelle (das ergibt sich aus der Stetigkeit von f und der Tatsache, dass f im Intervall I das Vorzeichen wechselt, sie Satz 5.2-3(i)).

2. $f'(x) \neq 0$ für jedes $x \in I$, d.h. die Nullstelle ist eindeutig (da in I kein Extremwert von f existiert).

3. $f''(x) \leq 0$ oder $f''(x) \geq 0$ für jedes $x \in I$, d.h. f ist entweder konkav oder konvex auf I.

4. Bezeichnet c denjenigen Endpunkt von $[a, b]$, für den $|f'(x)|$ kleiner ist als am anderen Endpunkt, so gilt

   $$\left| \frac{f(c)}{f'(c)} \right| \leq b - a .$$

   Die Bedingung sichert die Konvergenz des Verfahrens.

Gesucht wird eine Lösung der Gleichung

$f(x) = 0$ mit $x \in I$.

Bei diesen Voraussetzungen über f approximiert das folgende Verfahren die gesucht Lösung $x_0 \in I$ (für die dann $f(x_0) = 0$ gilt):

Man wählt einen beliebigen Punkt $a_0 \in I$.
Man berechnet für $n = 0, 1, 2, ...$

$$a_{n+1} = a_n - \frac{f(a_n)}{f'(a_n)},$$

bis sich aufeinanderfolgende Werte von a_{n+1} und a_n nur noch „wenig“ unterscheiden (weniger als eine vorgegebene Genauigkeitsschranke).

Die so definierte Folge $(a_n)_{n\in\mathbf{N}}$ approximiert die gesuchte Lösung $x_0 \in I$ der Gleichung $f(x) = 0$.

Anschaulich lässt sich das Verfahren wie folgt darstellen. Nach Wahl von a_0 konstruiert man die Tangente im Punkt $(a_0, f(a_0))$ an den Graphen von *f*. Die Tangenten besitzt die Gleichung $y_T(x) = f(a_0) + (x - a_0)\cdot f'(a_0)$. Der Wert a_1 ist der Schnittpunkt dieser Tangente mit der *x*-Achse. Mit diesem Wert (anstelle von a_0) iteriert man das Verfahren.

Beispiele:

Zur Bestimmung der Quadratwurzel $\sqrt{c}$ einer reellen Zahl $c > 0$ wählt man

$$f(x) = x^2 - c.$$

Die Folge $(a_n)_{n\in\mathbf{N}}$ lautet hierbei:

$$a_0 = c/2,$$
$$a_{n+1} = \frac{1}{2}\cdot\left(a_n + \frac{c}{a_n}\right),\ n \in \mathbf{N}.$$

Zur Bestimmung der beliebigen Wurzel $\sqrt[k]{c} = c^{1/k}$ mit $k \in \mathbf{N}_{>0}$ einer reellen Zahl $c > 0$ wählt man

$$f(x) = x^k - c.$$

Die Folge $(a_n)_{n\in\mathbf{N}}$ lautet hierbei:

$$a_0 > 0 \text{ beliebig},$$
$$a_{n+1} = \left(1 - \frac{1}{k}\right)\cdot a_n + \frac{1}{k}\cdot c \cdot a_n^{1-k},\ n \in \mathbf{N}.$$

Zur Berechnung des inversen Werts $1/c$ einer reellen Zahl $c > 0$ sind keine Divisionen erforderlich: Man wählt

$$f(x) = \frac{1}{x} - c\,.$$

Die Folge $(a_n)_{n \in \mathbf{N}}$ lautet hierbei:

$a_0 =$ beliebig mit $0 < a_0 < 2c^{-1}$ (Schätzung),
$a_{n+1} = a_n \cdot (2 - c \cdot a_n)\,,\ n \in \mathbf{N}$.

Das Newton-Verfahren ist robust gegen Rundungsfehler. Ein Iterationsschritt im Verfahren, d.h. die Berechnung eines weiteren Werts a_{n+1}, verwendet nur den Wert a_n und nicht vorherige Werte, etwa a_{n-1}, a_{n-2}, ..., a_0. Der Wert a_{n+1} hängt also nur von a_n ab. Derartige „einstellige" Iterationsverfahren haben den Vorteil, dass sich Rundungsfehler nicht akkumulieren.

Außerdem zeigt das Newton-Verfahren ein gutes Konvergenzverhalten („quadratische Konvergenz"), d.h. nach wenigen Iterationsschritten bekommt man bereits eine gute Näherung an die gesuchte Lösung.

Beispiel:

Es wird eine Nullstelle der durch $f(x) = x^3 - x^2 + 2 \cdot x + 5$ gegebenen Funktion gesucht. Es ist $f'(x) = 3 \cdot x^2 - 2 \cdot x + 2$ und $f''(x) = 6 \cdot x - 2$. Als „Suchintervall" für eine Nullstelle kann $I = [-2,\ -1]$ genommen werden. Hierfür sind alle obigen Bedingungen 1. bis 4. erfüllt. Als Startwert der Iteration wird $a_0 = -1{,}5$ gewählt. Die folgende Tabelle zeigt das Ergebnis nach dem Newton-Verfahren nach 6 Iterationsschritten (ermittelt mit einem Tabellenkalkulationsprogramm). Zusätzlich ist das Ergebnis angegeben, das das Tabellenkalkulationsprogramm mit der eingebauten „Berechne-für"-Funktion bei 100 Iterationen liefert. Offensichtlich ist hier das Newton-Verfahren bei weitem überlegen.

n	a_n	$f(a_n)$	$f'(a_n)$
0	− 1,5	− 3,625	11,75
1	− 1,1914893617021	− 0,49411980004431	8,6419194205523
2	− 1,1343122722156	− 0,014768016090808	8,1286175371279
3	− 1,1324954791357	$-1{,}4526940122457 \cdot 10^{-05}$	8,1126289890596
4	− 1,1324936884782	$-1{,}4100025400726 \cdot 10^{-11}$	8,1126132402848
5	− 1,1324936884764	$-7{,}6501305290577 \cdot 10^{-16}$	8,1126132402696
6	− 1,1324936884764	$-7{,}6501305290577 \cdot 10^{-16}$	8,1126132402696

Ergebnis der eingebauten „Berechne-für“-Funktion bei 100 Iterationen:

$x_0 = -1{,}1324940415747 \qquad f(x_0) = -2{,}8645504939842 \cdot 10^{-06}$

5.9 Taylorpolynome

Im folgenden sei $f: X \to \mathbf{R}$ mit $X \subseteq \mathbf{R}$ eine „genügend oft“ differenzierbare Funktion. Der Wert $x_0 \in X$ sei ein festgewählter Punkt. Der Funktionsverlauf von f soll durch eine Folge $\left(T_n(x; x_0; f)\right)_{n \in \mathbf{N}}$ „einfacherer“ Funktionen angenähert werden, die mit f im Punkt $\left(x_0, f(x_0)\right)$ übereinstimmen und folgenden Bedingungen genügen:

$T_n(x; x_0; f)$ für $n \geq 0$ ist dasjenige Polynom n-ten Grades, das mit f an der Stelle x_0 übereinstimmt und dessen sämtliche Ableitungen bis zur n-ten Ableitung mit den entsprechenden Ableitungen von f bei x_0 übereinstimmen. Zur Vereinfachung der Rechnung wird dabei nicht der Ansatz $T_n(x; x_0; f) = \sum_{i=0}^{n} b_i \cdot x^i$ gewählt, sondern

$T_n(x; x_0; f) = a_n \cdot (x - x_0)^n + a_{n-1} \cdot (x - x_0)^{n-1} + \ldots + a_1 \cdot (x - x_0) + a_0$ mit

$T_n(x_0; x_0; f) = f(x_0),$

$T_n'(x_0; x_0; f) = f'(x_0),$

...

$T_n^{(n)}(x_0; x_0; f) = f^{(n)}(x_0).$

Zur Berechnung der Koeffizienten $a_0, \ldots, a_n$ wird $T_n(x; x_0; f)$ nacheinander nach x abgeleitet und x_0 eingesetzt:

$T_n^{(0)}(x; x_0; f)\big|_{x=x_0} = T_n(x_0; x_0; f) = a_0 = f(x_0),$

$T_n'(x; x_0; f)\big|_{x=x_0} = n \cdot a_n \cdot (x - x_0)^{n-1} + (n-1) \cdot a_{n-1} \cdot (x - x_0)^{n-2} + \ldots + a_1\big|_{x=x_0} = a_1 = f'(x_0),$

$$T_n''(x;x_0;f)\big|_{x=x_0} = n\cdot(n-1)\cdot a_n\cdot(x-x_0)^{n-2} + (n-1)\cdot(n-2)\cdot a_{n-1}\cdot(x-x_0)^{n-3} + \ldots + 2\cdot a_2\Big|_{x=x_0}$$
$$= 2\cdot a_2 = f''(x_0), \text{ also } a_2 = \frac{1}{2}\cdot f''(x_0),$$

...

$$T_n^{(n-1)}(x;x_0;f)\big|_{x=x_0} = n\cdot(n-1)\cdot\ldots\cdot 2\cdot a_n\cdot(x-x_0) + (n-1)!\cdot a_{n-1}\big|_{x=x_0} = (n-1)!\cdot a_{n-1}$$
$$= f^{(n-1)}(x_0), \text{ also } a_{n-1} = \frac{1}{(n-1)!}\cdot f^{(n-1)}(x_0),$$

$$T_n^{(n)}(x;x_0;f)\big|_{x=x_0} = n\cdot(n-1)\cdot\ldots\cdot 2\cdot 1\cdot a_n\big|_{x=x_0} = n!\cdot a_n$$
$$= f^{(n)}(x_0), \text{ also } a_n = \frac{1}{n!}\cdot f^{(n)}(x_0).$$

Insgesamt ergibt sich also

$$T_n(x;x_0;f) = \sum_{i=0}^{n}\frac{f^{(i)}(x_0)}{i!}\cdot(x-x_0)^i .$$

Die Polynome $T_n(x;x_0;f)$ für $n=0$ und $n=1$ lauten:

$n=0$: $T_0(x;x_0;f) = f(x_0)$, ist also die konstante Funktion mit Wert $f(x_0)$.

$n=1$: $T_1(x;x_0;f) = f(x_0) + f'(x_0)\cdot(x-x_0)$, d.h. T_1 ist die Tangente an den Graphen von f im Punkt $(x_0, f(x_0))$.

Selbstverständlich ist $f(x)$ in der Regel nicht gleich $T_n(x;x_0;f)$, sondern es gilt

$$f(x) = T_n(x;x_0;f) + R_n(x;x_0;f)$$

mit einer als Restglied bezeichneten Funktion $R_n(x;x_0;f)$.

Satz 5.9-1:

Die Funktion $f: X \to \mathbf{R}$ sei in einer ε-Umgebung $U(x_0, \varepsilon) = \{ x \mid |x - x_0| < \varepsilon \} \subseteq X$ von $x_0 \in X$ $(n+1)$-mal differenzierbar. Dann gilt für alle $x \in U(x_0, \varepsilon)$:

$$f(x) = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!} \cdot (x - x_0)^i + R_n(x; x_0; f).$$

Die Summe

$$T_n(x; x_0; f) = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!} \cdot (x - x_0)^i$$

heißt **Taylorpolynom *n*-ter Ordnung von *f* an der Stelle** x_0; $R_n(x; x_0; f)$ heißt **Restglied des Taylorpolynoms *n*-ter Ordnung von *f* an der Stelle** x_0. Die Darstellung von $f(x)$ in der Form $f(x) = T_n(x; x_0; f) + R_n(x; x_0; f)$ nennt man **Taylorentwicklung von *f*** an der Stelle x_0**.**

Für das Restglied gilt:

$$R_n(x; x_0; f) = \frac{1}{(n+1)!} \cdot f^{(n+1)}(z) \cdot (x - x_0)^{n+1}.$$

Dabei ist z ein Wert mit $x_0 < z < x$, falls $x_0 < x$ ist, bzw. mit $x < z < x_0$, falls $x < x_0$ ist. Für $x = x_0$ ist $R_n(x) = 0$.

Das Restglied lässt sich auch in der Form

$$R_n(x; x_0; f) = \frac{1}{(n+1)!} \cdot f^{(n+1)}(x_0 + \lambda \cdot (x - x_0)) \cdot (x - x_0)^{n+1} \text{ mit } 0 < \lambda < 1$$

schreiben.

Eine alternative Restglieddarstellung lautet

$$R_n(x; x_0; f) = \frac{1}{n!} f^{(n+1)}(x_0 + \lambda \cdot (x - x_0)) \cdot (1 - \lambda)^n \cdot (x - x_0)^{n+1} \text{ mit } 0 < \lambda < 1.$$

Es bleibt zunächst zu zeigen, dass $R_n(x;x_0;f)=\frac{1}{(n+1)!}f^{(n+1)}(z)\cdot(x-x_0)^{n+1}$ für einen Wert z zwischen x und x_0 ist.

Es sei $x\neq x_0$. Zur Abkürzung wird $R(x)=R_n(x;x_0;f)$ und $T(x)=T_n(x;x_0;f)$ gesetzt. Dann ist $f(x)=T(x)+R(x)$. Mit $r(x)=\frac{R(x)}{(x-x_0)^{n+1}}$ ist $f(x)=T(x)+r(x)\cdot(x-x_0)^{n+1}$. Durch

$$h(y)=f(y)-T(y)-r(x)\cdot(y-x_0)^{n+1}$$

wird eine zwischen x und x_0 $(n+1)$-mal differenzierbare Funktion definiert, für die $h(x_0)=0$ und $h(x)=0$ gilt. Nach Konstruktion von T gilt $f^{(i)}(x_0)=T^{(i)}(x_0)$ für $i=0,...,n$. Daher ist $h^{(i)}(x_0)=0$ für $i=0,...,n$.

Nach Satz 5.6-4 gibt es zwischen x und x_0 einen Wert z_1 mit $h'(z_1)=0$. Nochmalige Anwendung des Satzes 5.6-4 liefert einen Wert z_2 zwischen z_1 und x_0 mit $h''(z_2)=0$. Die Anwendung des Satzes 5.6-4 lässt sich fortsetzen, und es entsteht eine Folge z_1, z_2, …, z_n und z_{n+1}, der zwischen z_n und x_0 liegt, mit $h^{(n+1)}(z_{n+1})=0$. Da T ein Polynom n-ten Grades ist, gilt $h^{(n+1)}(y)=f^{(n+1)}(y)-(n+1)!\cdot r(x)$, also mit $z=z_{n+1}$: $f^{(n+1)}(z)=(n+1)!\cdot r(x)$ bzw.

$$R(x)=\frac{1}{(n+1)!}\cdot f^{(n+1)}(z)\cdot(x-x_0)^{n+1}.$$

Die Entwicklung der alternativen Restglieddarstellung

$R_n(x;x_0;f)=\frac{1}{n!}f^{(n+1)}(x_0+\lambda\cdot(x-x_0))\cdot(1-\lambda)^n\cdot(x-x_0)^{n+1}$ mit $0<\lambda<1$ ergibt sich aus folgender Überlegung:

Setzt man $F(t)=f(x)-\sum_{i=0}^{n}\frac{f^{(i)}(t)}{i!}\cdot(x-t)^i$ und $G(t)=(x-t)$, dann gibt es nach Satz 5.6-5(ii) einen Wert z zwischen x und x_0, etwa $z=x_0+\lambda\cdot(x-x_0)$ mit $0<\lambda<1$, mit

$\left.\frac{d/dt\,F(t)}{d/dt\,G(t)}\right|_{t=z}=\frac{F(x_0)-F(x)}{G(x_0)-G(x)}$. Es ist

$$\begin{aligned}
d/dt\,F(t)&=d/dt\left(f(x)-f(t)-\sum_{i=1}^{n}\frac{f^{(i)}(t)}{i!}\cdot(x-t)^i\right)\\
&=-\left(f'(t)+\sum_{i=1}^{n}\left(\frac{f^{(i+1)}(t)}{i!}\cdot(x-t)^i-\frac{f^{(i)}(t)}{i!}i\cdot(x-t)^{i-1}\right)\right)\\
&=-\left(f'(t)+\sum_{i=1}^{n}\frac{f^{(i+1)}(t)}{i!}\cdot(x-t)^i-\sum_{i=0}^{n-1}\frac{f^{(i+1)}(t)}{i!}\cdot(x-t)^i\right)\\
&=-\frac{f^{(n+1)}(t)}{n!}\cdot(x-t)^n \quad \text{und}
\end{aligned}$$

$d/dt\, G(t) = -1$. Weiter ist $F(x) = 0$, $G(x) = 0$, $F(x_0) - F(x) = R_n(x; x_0; f)$ und $G(x_0) - G(x) = x - x_0$. Damit ergibt sich

$$\frac{F(x_0) - F(x)}{G(x_0) - G(x)} = \frac{R_n(x; x_0; f)}{x - x_0} = \frac{f^{(n+1)}(z)}{n!} \cdot (x - z)^n$$

$$= \frac{f^{(n+1)}(x_0 + \lambda \cdot (x - x_0))}{n!} \cdot ((x - x_0) \cdot (1 - \lambda))^n \quad ,$$

$$R_n(x; x_0; f) = \frac{1}{n!} f^{(n+1)}(x_0 + \lambda \cdot (x - x_0)) \cdot (1 - \lambda)^n \cdot (x - x_0)^{n+1} .$$

Beispiel:

Es soll die Taylorentwicklung für die Funktion

$$f(x) = e^x$$

an der Stelle $x_0 = 0$ berechnet werden. Dabei sollen nur die Ableitungen von $f(x)$ verwendet werden. Da für alle Ableitungen $f^{(i)}(x) = e^x$ gilt und $e^0 = 1$ ist, ergibt sich für das *n*-te Taylorpolynom von $f(x) = e^x$ an der Stelle $x_0 = 0$:

$$T_n(x; 0; e^x) = \sum_{i=0}^{n} \frac{1}{i!} \cdot x^i = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots + \frac{x^n}{n!} .$$

Mit Satz 5.7-2 ist

$$e^x = \sum_{i=0}^{n} \frac{1}{i!} \cdot x^i + R_n(x; 0; e^x) = \sum_{i=0}^{n} \frac{1}{i!} \cdot x^i + \frac{1}{(n+1)!} \cdot e^z \cdot x^{n+1} \text{ mit } 0 < z < x \text{ für } x > 0 \text{ und}$$

$$x < z < 0 \text{ für } x < 0 .$$

Nun gilt:

$$\lim_{n \to \infty} R_n(x; 0; e^x) = \lim_{n \to \infty} \frac{1}{(n+1)!} \cdot e^z \cdot x^{n+1} = 0 \text{; denn}$$

für $x > 0$ ist $0 \le R_n(x; 0; e^x) < \frac{1}{(n+1)!} \cdot e^x \cdot x^{n+1}$, da $0 < z < x$ ist, und damit

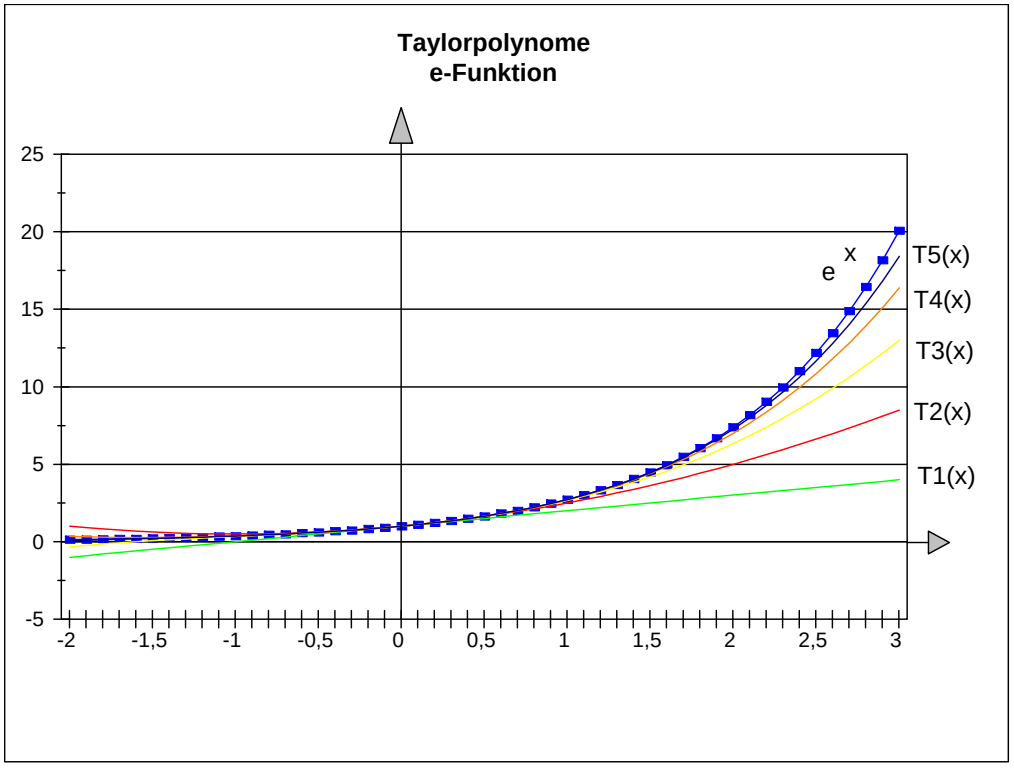
$$0 \le \lim_{n \to \infty} R_n(x; 0; e^x) \le \lim_{n \to \infty} \left(\frac{1}{(n+1)!} \cdot e^x \cdot x^{n+1} \right) = 0 \quad \text{(siehe Satz 5.1-5 (iii)).}$$

Für $x < 0$ wird $\left|R_n\left(x;0;e^x\right)\right| = \left|\frac{1}{(n+1)!}\cdot e^z \cdot x^{n+1}\right| = \frac{1}{(n+1)!}\cdot e^z \cdot \left|x^{n+1}\right| < \frac{\left|x^{n+1}\right|}{(n+1)!}$ betrachtet (die letzte Ungleichung ergibt sich aus $e^z < e^0 = 1$): hier ergibt sich mit demselben Argument $\lim\limits_{n\to\infty}\left|R_n\left(x;0;e^x\right)\right| = 0$, und mit Satz 5.1-2 (v) folgt $\lim\limits_{n\to\infty} R_n\left(x;0;e^x\right) = 0$.

Insgesamt ist

$$e^x = \sum_{i=0}^{\infty} \frac{1}{i!}\cdot x^i \quad \text{für } x \in \mathbf{R}.$$

Dieses ist ein Ergebnis, das nicht überrascht; denn so wurde die Exponentialfunktion ja definiert. Zu beachten ist aber, dass bei der Taylorentwicklung allein von der Tatsache Gebrauch gemacht wurde, dass $f^{(i)}(x) = f(x) = e^x$ und $e^0 = 1$ ist. Die folgende Abbildung zeigt den Verlauf der Exponentialfunktion und ihrer Taylorpolynome an der Stelle $x_0 = 0$ für $0 \le n \le 5$:



Um auch ein „numerisches Gefühl“ für die Qualität der Approximation der Exponentialfunktion durch ihr n-tes Taylorpolynom zu bekommen, werden in folgender Tabelle Werte des dritten Taylorpolynoms zur Exponentialfunktion an der Stelle $x_0 = 0$ dicht an der Entwicklungsstelle und weit entfernt von der Entwicklungsstelle mit numerisch ermittelten Werten

von e^x verglichen. Man sieht dabei, dass für Werte, die dicht bei $x_0 = 0$ liegen, bereits $1+x$ häufig eine befriedigende Annäherung an e^x liefert. Für große Werte von *x* ist ein *n*-tes Taylorpolynom mit einem großen Wert von *n* zu nehmen.

x	$T_3\left(x; 0; e^x\right) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$	e^x (letzte Stelle gerundet)
1	$2,\overline{6}...$	2,71828182845904523536028747135 27
1/2	$1,6458\overline{3}...$	1,64872127070012814684865078781 42
1/10	$1,1051\overline{6}...$	1,10517091807564762481170782649 02
1/100	$1,0100501\overline{6}...$	1,01005016708416805754216545690 29
20	$1554,\overline{3}...$	485 165 195,4097902779691068305415 4

Insbesondere gilt

$$e = \sum_{i=0}^{\infty} \frac{1}{i!} = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \ \ldots \ + \frac{1}{n!} + \sum_{i=n+1}^{\infty} \frac{1}{i!}.$$

Wegen $a^x = e^{\ln(a)\cdot x}$ ist

$$a^x = \sum_{i=0}^{\infty} \frac{\left(x \cdot \ln(a)\right)^i}{i!} \quad \text{für } x \in \mathbf{R}.$$

Beispiel:

Es soll nun die Taylorentwicklung der natürlichen Logarithmusfunktion $\ln(x)$ hergeleitet werden. Die Wahl $x_0 = 0$ der Entwicklungsstelle ist hierbei nicht möglich, da $\ln(x)\big|_{x=x_0}$ nicht definiert ist. Andererseits ist eine Taylorentwicklung an der Entwicklungsstelle $x_0 = 0$ besonders einfach. Es wird daher zunächst die durch

$$f(x) = \begin{cases} \mathbf{R}_{>-1} & \to \quad \mathbf{R} \\ x & \to \quad \ln(1+x) \end{cases}$$

definierte Funktion in eine Taylorreihe an der Stelle $x_0 = 0$ entwickelt:

Es ist

$$f'(x)=\frac{1}{1+x}=(1+x)^{-1},\quad f''(x)=-(1+x)^{-2},\quad f'''(x)=2\cdot(1+x)^{-3},$$
$$f^{(i)}(x)=(-1)^{i-1}\cdot(i-1)!\cdot(1+x)^{-i}\ \text{und damit}$$

$$T_n(x;0;\ln(1+x))=\frac{\ln(1+0)}{0!}+\sum_{i=1}^{n}\frac{(-1)^{i-1}\cdot(i-1)!\cdot(1+0)^{-i}}{i!}\cdot x^i$$
$$=\sum_{i=1}^{n}\frac{(-1)^{i-1}}{i}\cdot x^i=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\ldots+\frac{(-1)^{n-1}}{n}\cdot x^n$$

Mit Satz 5.7-2 ist

$$\ln(1+x)=\sum_{i=1}^{n}\frac{(-1)^{i-1}}{i}\cdot x^i+R_n(x;0;\ln(1+x))=\sum_{i=1}^{n}\frac{(-1)^{i-1}}{i}\cdot x^i+\frac{(-1)^n}{(n+1)}\cdot(1+z)^{-(n+1)}\cdot x^{n+1}$$
$$\text{mit } 0<z<x \text{ für } x>0 \text{ und}$$
$$x<z<0 \text{ für } -1<x<0.$$

Das Restglied

$$R_n(x;0;\ln(1+x))=\frac{(-1)^n}{(n+1)}\cdot(1+z)^{-(n+1)}\cdot x^{n+1}$$

lässt sich betragsmäßig abschätzen:

Für $0<z<x$ gilt (wegen $1+z>1$ bzw. $\frac{1}{1+z}<1$) $|R_n(x;0;\ln(1+x))|<\frac{x^{n+1}}{n+1}$. Für $0\le x\le 1$ folgt daher $\lim\limits_{n\to\infty}|R_n(x;0;\ln(1+x))|=0$.

Für $-1/2<x<z<0$ ist (wegen $0<1-|x|=1-(-x)=1+x<1+z<1$ bzw. $0<\frac{1}{1+z}<\frac{1}{1-|x|}$)

$|R_n(x;0;\ln(1+x))|=\frac{|x|^{n+1}}{(n+1)}\cdot|(1+z)|^{-(n+1)}<\frac{1}{n+1}\cdot\left(\frac{|x|}{1-|x|}\right)^{n+1}$. Wegen $|x|<1/2$ ist $\frac{|x|}{1-|x|}<1$, so dass auch in diesem Fall $\lim\limits_{n\to\infty}|R_n(x;0;\ln(1+x))|=0$ gilt.

Für $-1<x\le -1/2$ verwendet man die alternative Restgliedabschätzung in Satz 5.9-1:

$$R_n(x;0;\ln(1+x))=(-1)^n\cdot\frac{(1-\lambda)^n}{(1+\lambda\cdot x)^{n+1}}\cdot x^{n+1}\ \text{mit}\ 0<\lambda<1.$$

Es ist $0<\frac{1-\lambda}{1+\lambda\cdot x}=\frac{1-\lambda}{1-\lambda\cdot|x|}<1$ und $\frac{1}{1+\lambda\cdot x}<\frac{1}{1-|x|}$ und damit

$\left|R_n(x;0;\ln(1+x))\right| \le \frac{1}{(1-|x|)^{n+1}} \cdot |x|^{n+1}$. Daraus folgt auch hier $\lim_{n\to\infty}\left|R_n(x;0;\ln(1+x))\right| = 0$.

Insgesamt ist

$$\ln(1+x) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot x^i$$
$$= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \ldots + \frac{(-1)^{n+1}}{n} x^n + \sum_{i=n+1}^{\infty} \frac{(-1)^{i+1}}{i} x^i \quad \text{für } x \in \mathbf{R} \text{ mit } -1 < x \le 1.$$

Die Taylorentwicklung für $f(x) = \ln(x)$ für $x > 0$ erhält man aus der Substitution $z = x - 1$, $x = z + 1$:

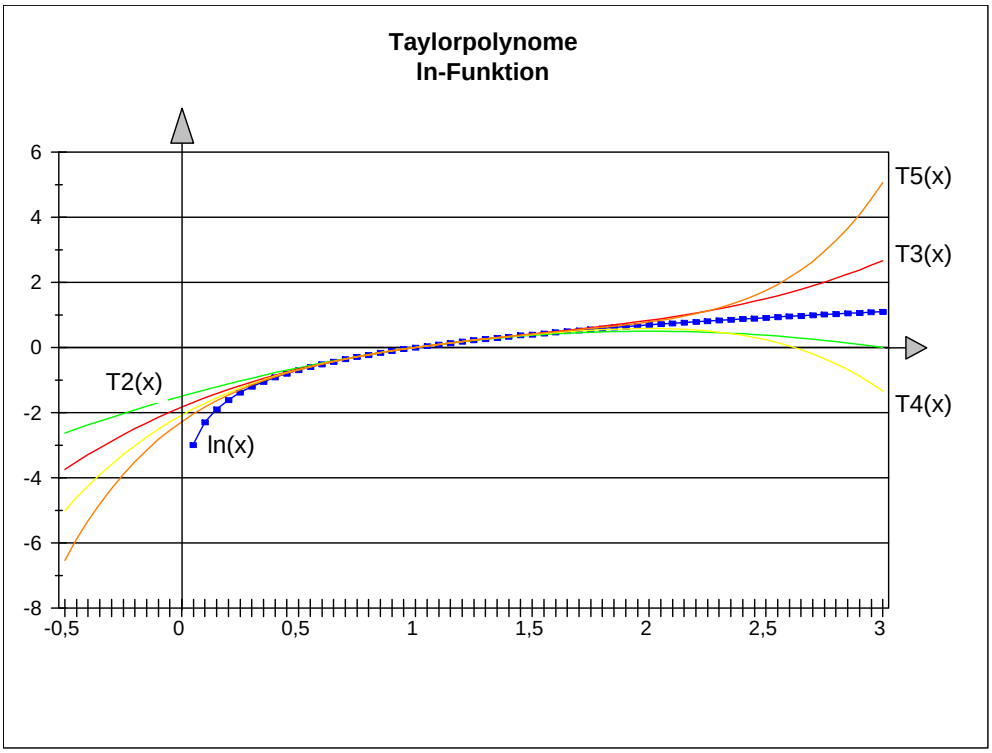
$$\ln(x) = \ln(z+1) = \sum_{i=1}^{n} \frac{(-1)^{i-1}}{i} \cdot (x-1)^i + R_n(x-1;0;\ln(1+z)) \text{ und}$$

$$\ln(x) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot (x-1)^i$$
$$= x - 1 - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \ldots + \frac{(-1)^{n+1}}{n}(x-1)^n + \sum_{i=n+1}^{\infty} \frac{(-1)^{i+1}}{i}(x-1)^i$$
für $x \in \mathbf{R}$ mit $0 < x \le 2$.

Daraus ergibt sich das Ergebnis aus Satz 5.1-9 (iii)

$$\ln(2) = \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i}$$
$$= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \ \ldots\ + \frac{(-1)^{n+1}}{n} + \sum_{i=n+1}^{\infty} \frac{(-1)^{i+1}}{i} \approx 0{,}6931471.$$

Die folgende Abbildung zeigt den Verlauf der natürlichen Logarithmusfunktion und ihrer Taylorpolynome an der Stelle $x_0 = 1$ für $0 \le n \le 5$; man sieht sehr schön die Approximation der Taylorpolynome für $x \in \mathbf{R}$ mit $0 < x \le 2$.



Es soll nun die Größe von n berechnet werden, die ausreicht, damit $T_n(x;0;\ln(1+x))$ eine Approximation an $\ln(1+x)$ liefert, die einer vorgegebenen Genauigkeit genügt. Hierbei sei $0 \le x \le 1$. Soll also $T_n(x;0;\ln(1+x))$ in der Dezimalentwicklung bis zur m-ten Nachkommastelle genau sein, so kann man folgendermaßen vorgehen: Die Dezimalentwicklung von $\ln(1+x)$ bzw. von $T_n(x;0;\ln(1+x))$, die bis zur m-ten Nachkommastelle identisch sind, seien

$$\ln(1+x) = [d_k d_{k-1} \dots d_1 d_0, d_{-1} d_{-2} \dots d_{-m} d_{-m-1} d_{-m-2} \dots]_{10} \quad \text{bzw.}$$

$$T_n(x;0;\ln(1+x)) = [d_k d_{k-1} \dots d_1 d_0, d_{-1} d_{-2} \dots d_{-m} c_{-m-1} c_{-m-2} \dots]_{10}.$$

Dann ist

$$\begin{aligned} R_n(x;0;\ln(1+x)) &= \ln(1+x) - T_n(x;0;\ln(1+x)) \\ &= 0,\underbrace{0 \dots 0}_{m-\text{mal}} + (d_{-m-1} - c_{-m-1}) \cdot 10^{-(m+1)} + (d_{-m-2} - c_{-m-2}) \cdot 10^{-(m+2)} + \dots \\ &= \sum_{i=m+1}^{\infty} (d_{-i} - c_{-i}) \cdot 10^{-i}, \end{aligned}$$

$(-9) \cdot \sum_{i=m+1}^{\infty} 10^{-i} \le R_n(x;0;\ln(1+x)) \le 9 \cdot \sum_{i=m+1}^{\infty} 10^{-i}$, d.h. wegen $\sum_{i=m+1}^{\infty} 10^{-i} = \frac{1}{9} \cdot 10^{-m}$:

$-10^{-m} \le R_n(x;0;\ln(1+x)) \le 10^{-m}$ bzw. $0 \le \left|R_n(x;0;\ln(1+x))\right| \le 10^{-m}$.

Im angenommenen Fall $0 \le x \le 1$ ist $0 \le \left|R_n(x;0;\ln(1+x))\right| \le \frac{x^{n+1}}{n+1}$. Die Anforderung an n lautet also

$$\frac{x^{n+1}}{n+1} \le 10^{-m}.$$

Soll beispielsweise $\ln(1{,}5)$ auf 7 Nachkommastellen genau durch das n-te Taylorpolynom angegeben werden, so wird n so bestimmt, dass $\frac{1/2^{n+1}}{n+1} \le 10^{-7}$, d.h. $(n+1)\cdot 2^{n+1} \ge 10^7$ ist; $(n+1)\cdot 2^{n+1}$ muss also mindestens 8 Dezimalstellen aufweisen. Die folgende Tabelle listet einige Werte für $(n+1)\cdot 2^{n+1}$ auf:

n	$(n+1)\cdot 2^{n+1}$
1	8
2	24
...	...
10	22.528
11	49.152
...	...
17	4.718.592
18	9.961.472
19	10.971.520

In diesem Beispiel ist also $n \ge 19$ zu wählen.

Bemerkung:

ln(1,5) =
0,40546510810816438197801311546435
(letzte Stelle gerundet)

Zur Berechnung von ln(x) für $x \in \mathbf{R}$ mit $0 < x \le 2$ kann man also ein n-tes Taylorpolynom mit genügend großem n als Approximation an ln(x) nehmen. Im allgemeinen (auch für $x > 2$) funktioniert dieser Ansatz nicht, da das Restglied nicht gegen 0 konvergiert. Nun gilt aber für $|h| < 1$:

$$\ln(1+h) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot h^i \quad \text{und} \quad \ln(1-h) = \ln(1+(-h)) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot (-h)^i = \sum_{i=1}^{\infty} \frac{-1}{i} \cdot h^i \quad \text{und damit}$$

$$\ln\left(\frac{1+h}{1-h}\right) = \ln(1+h) - \ln(1-h) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot h^i - \sum_{i=1}^{\infty} \frac{-1}{i} \cdot h^i = 2 \cdot \sum_{i=0}^{\infty} \frac{1}{2 \cdot i + 1} \cdot h^{2 \cdot i + 1}$$

$$= 2 \cdot \left(h + \frac{h^3}{3} + \frac{h^5}{5} + \ldots + \frac{h^{2 \cdot n+1}}{2 \cdot n + 1} \right) + 2 \cdot \sum_{i=n+1}^{\infty} \frac{h^{2 \cdot i+1}}{2 \cdot i + 1}.$$

Für $x \in \mathbf{R}$ mit $x > 0$ ist $-1 < \frac{x-1}{x+1} < 1$. Setzt man $h = \frac{x-1}{x+1}$, so ist $|h| < 1$ und $x = \frac{1+h}{1-h}$. Damit ergibt sich

$$\ln(x) = 2 \cdot \left(h + \frac{h^3}{3} + \frac{h^5}{5} + \ldots + \frac{h^{2 \cdot n+1}}{2 \cdot n+1} \right) + 2 \cdot \sum_{i=n+1}^{\infty} \frac{h^{2 \cdot i+1}}{2 \cdot i+1} \Bigg|_{h=\frac{x-1}{x+1}} \quad \text{für } x \in \mathbf{R}.$$

Beispiel:

Für $m \in \mathbf{N}$ ist

$$(1 \pm x)^m = \sum_{i=0}^{m} (\pm\ 1)^i \, \frac{m \cdot (m-1) \cdot \ \ldots \ \cdot (m-i+1)}{i!} x^i = \sum_{i=0}^{m} (\pm\ 1)^i \binom{m}{i} x^i \quad \text{für } x \in \mathbf{R}.$$

Diese Formel ist ein Spezialfall der allgemeineren Formel

$$(a \pm b)^m = \sum_{i=0}^{m} (\pm\ 1)^i \, \frac{m \cdot (m-1) \cdot \ \ldots \ \cdot (m-i+1)}{i!} a^i b^{m-i} = \sum_{i=0}^{m} (\pm\ 1)^i \binom{m}{i} a^i b^{m-i}$$

für $a \in \mathbf{R}$, $b \in \mathbf{R}$.

Für $m \in \mathbf{R} \setminus \mathbf{N}$ mit $m > 0$ ist

$$(1 \pm x)^m = \sum_{i=0}^{\infty} (\pm\ 1)^i \, \frac{m \cdot (m-1) \cdot \ \ldots \ \cdot (m-i+1)}{i!} x^i$$

$$= 1 \pm mx + \frac{m(m-1)}{2} x^2 \pm \frac{m(m-1)(m-2)}{6} x^3 + \sum_{i=4}^{\infty} (\pm\ 1)^i \, \frac{m(m-1) \cdot \ldots \cdot (m-i+1)}{i!} x^i$$

für $x \in \mathbf{R}$ mit $-1 \le x \le 1$.

Beispiel:

Zur Berechnung eines Wertes der Form $\frac{1{,}000001}{0{,}999999^2}$ mit großer Genauigkeit kann die Taylorentwicklung einer „geeigneten“ Funktion herangezogen werden. Mit

$$f : \begin{cases} \mathbf{R}_{\neq 1} & \to & \mathbf{R} \\ x & \to & \frac{1+x}{(1-x)^2} \end{cases}$$

ist $\frac{1{,}000001}{0{,}999999^2} = f\left(10^{-6}\right)$. Das *n*-te Taylorpolynom der Funktion *f* wird ermittelt; dazu werden einige Ableitungen ermittelt, um daraus auf die Form der *i*-ten Ableitung zu schließen:

$$f^{(0)}(x) = f(x) = \frac{1+x}{(1-x)^2},$$

$$f'(x) = \frac{1\cdot(1-x)^2 - (1+x)\cdot(-2)\cdot(1-x)}{(1-x)^4} = \frac{3+x}{(1-x)^3},$$

$$f''(x) = \frac{(1-x)^3 - (3+x)\cdot(-3)\cdot(1-x)^2}{(1-x)^6} = \frac{10+2\cdot x}{(1-x)^4} = \frac{2\cdot(5+x)}{(1-x)^4},$$

$$f'''(x) = \frac{2\cdot(1-x)^4 - 2\cdot(5+x)\cdot(-4)\cdot(1-x)^3}{(1-x)^8} = \frac{42+6\cdot x}{(1-x)^5} = \frac{6\cdot(7+x)}{(1-x)^5};$$

die *i*-te Ableitung lautet also (das kann man durch vollständige Induktion beweisen):

$$f^{(i)}(x) = \frac{i!\cdot(2\cdot i+1+x)}{(1-x)^{i+2}}.$$

Damit gilt für das *n*-te Taylorpolynom an der Stelle $x_0 = 0$:

$$T_n(x;0;f(x)) = \sum_{i=0}^{n} \frac{i!\cdot(2\cdot i+1)}{i!}\cdot x^i = \sum_{i=0}^{n}(2\cdot i+1)\cdot x^i = 1+3\cdot x+5\cdot x^2+7\cdot x^3+\ldots+(2\cdot n+1)\cdot x^n.$$

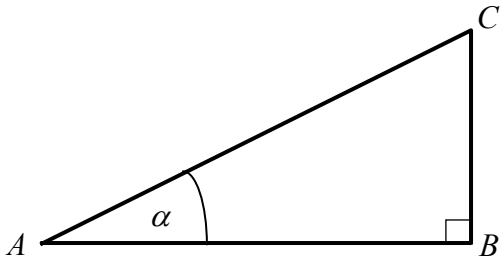
Für $-1 < x < 1$ konvergiert das Restglied $R_n(x;0;f(x)) = (2\cdot n+3+z)\cdot x^{n+1}$ gegen 0, so dass gilt:

$$\frac{1+x}{(1-x)^2} = \sum_{i=0}^{\infty}(2\cdot i+1)\cdot x^i \quad \text{für } -1 < x < 1.$$

Damit ist $\frac{1{,}000001}{0{,}999999^2} = \left.\frac{1+x}{(1-x)^2}\right|_{x=10^{-6}} = 1{,}000003000005000007000009000011000013 0...$

Beispiel:

Aus der Schule sind die trigonometrischen Funktionen Sinus-, Kosinus-, Tangens-, Kotangensfunktion und weitere daraus abgeleitete Funktionen bekannt. Grundlagen bilden die Sinus- und Kosinusfunktion, die über Längenverhältnisse in einem rechtwinkligen Dreieck definiert werden:



Der **Sinus eines Winkels** α in einem rechtwinkligen Dreieck (siehe Abbildung) ist definiert als das Verhältnis der Länge der Seite *BC* zur Länge der Seite *AC*. Bezeichnet man die Länge einer Seite *XY* mit $\overline{XY}$, dann ist $\sin(\alpha)=\frac{\overline{BC}}{\overline{AC}}$.

Der **Kosinus eines Winkels** α in einem rechtwinkligen Dreieck ist definiert als das Verhältnis der Länge der Seite *AB* zur Länge der Seite *AC*: $\cos(\alpha)=\frac{\overline{AB}}{\overline{AC}}$.

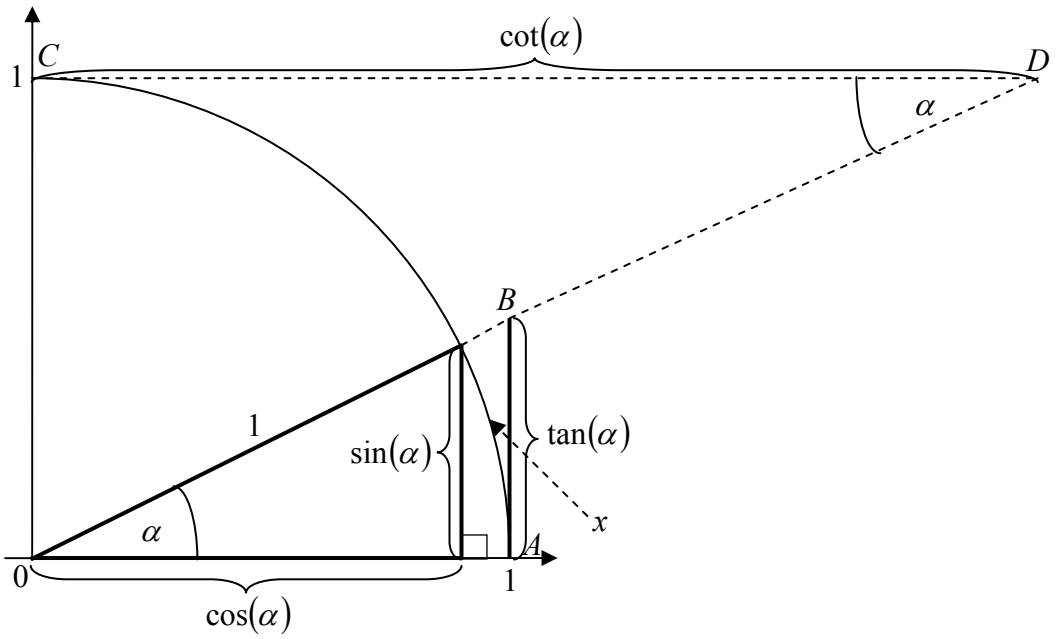
Der **Tangens eines Winkels** α in einem rechtwinkligen Dreieck ist definiert als das Verhältnis der Länge der Seite *BC* zur Länge der Seite *AB*: $\tan(\alpha)=\frac{\overline{BC}}{\overline{AB}}=\frac{\sin(\alpha)}{\cos(a)}$.

Der **Kotangens eines Winkels** α in einem rechtwinkligen Dreieck ist definiert als das Verhältnis der Länge der Seite *AB* zur Länge der Seite *BC*: $\cot(\alpha)=\frac{\overline{AB}}{\overline{BC}}=\frac{1}{\tan(a)}$.

Normiert man die Strecke $\overline{AC}$ zur Länge 1, so findet man die einzelnen Werte geometrisch aus folgender Abbildung, in der das definierende Dreieck in den Einheitskreis (Kreis mit Radius 1 um den Nullpunkt) eingezeichnet ist.

Dabei folgt aus dem Strahlensatz $\frac{\overline{AC}}{\sin(\alpha)}=\frac{1}{\cos(\alpha)}$, also $\overline{AC}=\frac{\sin(\alpha)}{\cos(\alpha)}=\tan(\alpha)$.

Aus der Ähnlichkeit der Dreiecke 0*DC* und *BA*0 folgt $\frac{\overline{DC}}{1}=\frac{1}{\tan(\alpha)}$, also $\overline{DC}=\cot(\alpha)$.



Die Winkelfunktionen lassen sich auch in Abhängigkeit der Länge x des Bogenstücks ausdrücken, das durch den Winkel auf dem Einheitskreis bestimmt wird. Die Umrechnung zwischen Winkel und Bogenstücklänge erfolgt über einen festen Faktor, der hier nicht weiter detailliert werden soll. Es ist dann $\sin(\alpha) = \sin(x)$; entsprechend für die anderen Funktionen.

Offensichtlich gelten folgende Beziehungen (hier werden die Argumente der Funktionen in Abhängigkeit der Bogenlänge angegeben):

$\sin(0) = 0$, $\cos(0) = 1$,
$(\sin(x))^2 + (\cos(x))^2 = 1$,
$\sin(x) \le x \le \tan(x)$.

Aus der letzten Gleichung ergibt sich wegen $\tan(x) = \sin(x)/\cos(x)$ bei $x \neq 0$: $x \cdot \cos(x) \le \sin(x) \le x$ und $\cos(x) \le \dfrac{\sin(x)}{x} \le 1$. Diese Ungleichung führt auf die später verwendete Limesbeziehung

$$\lim_{x \to 0} \frac{\sin(x)}{x} = 1.$$

Eine ähnliche später verwendete Limesbeziehung lässt sich für die Kosinusfunktion herleiten:

$$\begin{aligned}\frac{\cos(x)-1}{x} &= \frac{(\cos(x)-1)\cdot(\cos(x)+1)}{x\cdot(\cos(x)+1)}\\ &= \frac{(\cos(x))^2-1}{x\cdot(\cos(x)+1)}\\ &= -\frac{(\sin(x))^2}{x\cdot(\cos(x)+1)}\\ &= -\frac{\sin(x)}{x}\cdot\frac{\sin(x)}{\cos(x)+1},\end{aligned}$$

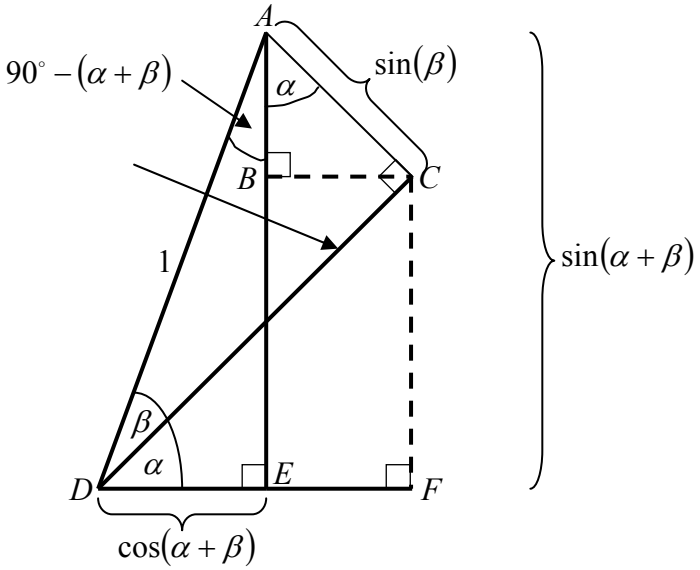
$$\lim_{x\to 0}\frac{\cos(x)-1}{x}=0\,.$$

Schließlich sollen noch die Additionstheoreme für die Sinus- und Kosinusfunktion (geometrisch) hergeleitet werden. Hierbei werden wieder Beziehungen im Dreieck betrachtet.

Es gilt

$$\begin{aligned}\sin(\alpha+\beta) &= \sin(\alpha)\cdot\cos(\beta)+\sin(\beta)\cdot\cos(\alpha),\\ \cos(\alpha+\beta) &= \cos(\alpha)\cdot\cos(\beta)-\sin(\alpha)\cdot\sin(\beta).\end{aligned}$$

Zur Herleitung wird folgende Abbildung betrachtet:



Die eingezeichneten Größen ergeben sich unmittelbar aus der Definition der Sinus- und Kosinusfunktion. Betrachtet man das Dreieck *ADE*, so erkennt man, dass der linke Winkel bei *A* die Größe $180^\circ - 90^\circ - (\alpha+\beta) = 90^\circ - (\alpha+\beta)$ besitzt. Aus der Winkelsumme 180° im Dreieck *ADC* ergibt sich die Größe $180^\circ - \beta - 90^\circ - \left(90^\circ - (\alpha+\beta)\right) = \alpha$ für den anderen Winkel bei

A. Aus $\sin(\alpha) = \overline{BC}/\sin(\beta)$ folgt $\overline{BC} = \sin(\alpha)\cdot\sin(\beta)$; aus $\cos(\alpha) = \overline{AB}/\sin(\beta)$ folgt $\overline{AB} = \cos(\alpha)\cdot\sin(\beta)$; aus $\sin(\alpha) = \overline{CF}/\cos(\beta)$ folgt $\overline{CF} = \sin(\alpha)\cdot\cos(\beta)$;
aus $\cos(\alpha) = \overline{DF}/\cos(\beta)$ folgt $\overline{DF} = \cos(\alpha)\cdot\cos(\beta)$.
Insgesamt ergibt sich
$\sin(\alpha+\beta) = \overline{CF} + \overline{AB} = \sin(\alpha)\cdot\cos(\beta) + \sin(\beta)\cdot\cos(\alpha)$ und
$\cos(\alpha+\beta) = \overline{DF} - \overline{BC} = \cos(\alpha)\cdot\cos(\beta) - \sin(\alpha)\cdot\sin(\beta)$.

Auch in diesen Gleichungen können wieder die Winkel durch entsprechende Bogenlängen im Einheitskreis ersetzt werden. Damit lassen sich nun die Ableitung der Sinus- und der Kosinusfunktion bestimmen, wobei die oben hergeleiteten Limesbeziehungen eingesetzt werden:

$$\begin{aligned}
\left.\frac{d}{dx}\sin(x)\right|_{x=x_0} &= \lim_{\Delta x\to 0}\frac{\sin(x_0+\Delta x)-\sin(x_0)}{\Delta x}\\
&= \lim_{\Delta x\to 0}\frac{\sin(x_0)\cdot\cos(\Delta x)+\sin(x\Delta)\cdot\cos(x_0)-\sin(x_0)}{\Delta x}\\
&= \lim_{\Delta x\to 0}\frac{\sin(x_0)\cdot(\cos(\Delta x)-1)}{\Delta x} + \lim_{\Delta x\to 0}\frac{\sin(x\Delta)\cdot\cos(x_0)}{\Delta x}\\
&= \sin(x_0)\cdot\lim_{\Delta x\to 0}\frac{\cos(\Delta x)-1}{\Delta x} + \cos(x_0)\cdot\lim_{\Delta x\to 0}\frac{\sin(x\Delta)}{\Delta x}\\
&= \cos(x_0)\ .
\end{aligned}$$

$$\begin{aligned}
\left.\frac{d}{dx}\cos(x)\right|_{x=x_0} &= \lim_{\Delta x\to 0}\frac{\cos(x_0+\Delta x)-\cos(x_0)}{\Delta x}\\
&= \lim_{\Delta x\to 0}\frac{\cos(x_0)\cdot\cos(\Delta x)-\sin(x_0)\cdot\sin(\Delta x)-\cos(x_0)}{\Delta x}\\
&= \lim_{\Delta x\to 0}\frac{\cos(x_0)\cdot(\cos(\Delta x)-1)}{\Delta x} - \lim_{\Delta x\to 0}\frac{\sin(x_0)\cdot\sin(x\Delta)}{\Delta x}\\
&= \cos(x_0)\cdot\lim_{\Delta x\to 0}\frac{\cos(\Delta x)-1}{\Delta x} - \sin(x_0)\cdot\lim_{\Delta x\to 0}\frac{\sin(x\Delta)}{\Delta x}\\
&= -\sin(x_0)\ .
\end{aligned}$$

Die Sinus- und die Kosinusfunktion bilden also ein Paar von (*g*, *h*) von Funktionen, die auf ganz **R** definiert und differenzierbar sind und die folgende Eigenschaften besitzen:

$g(0) = 0, \qquad h(0) = 1,$
$g'(x) = h(x), \quad h'(x) = -g(x).$

Mit dieser Betrachtungsweise erhält man einen analytischen Zugang zu den Winkelfunktionen, ohne ihre geometrische Herleitung zu bemühen. Im Folgenden wird gezeigt, dass diese wenigen Annahmen ausreichen, um *g* und *h* eindeutig festzulegen, und dass beide Funktionen genau denselben Gesetzmäßigkeiten folgen wie die Sinus- und Kosinusfunktion. Darüber hinaus ermöglicht dieser Ansatz einen nicht-geometrischen Weg, um π zu definieren.

Aus den obigen Annahmen folgt

$g(0) = 0$,
$g'(0) = h(0) = 1$,
$g''(0) = h'(0) = -g(0) = 0$,
$g'''(0) = -g'(0) = -h(0) = -1$,
$g^{(4)}(0) = -h'(0) = g(0) = 0$, d.h.

$$g^{(i)}(0) = \begin{cases} 0 & \text{für } (i \bmod 2) = 0 \\ 1 & \text{für } (i \bmod 4) = 1 \\ -1 & \text{für } (i \bmod 4) = 3 \end{cases}$$

Das $(2 \cdot n + 1)-$ te Taylorpolynom von *g* an der Stelle $x_0 = 0$ lautet:

$$T_{2 \cdot n+1}(x; 0; g(x)) = \sum_{i=0}^{n} \frac{(-1)^i}{(2 \cdot i + 1)!} \cdot x^{2 \cdot i + 1}.$$

Das Restglied konvergiert für jedes $x \in \mathbf{R}$ gegen 0 (die Begründung erfolgt wie bei der Exponentialfunktion), so dass

$$g(x) = \sum_{i=0}^{\infty} \frac{(-1)^i}{(2 \cdot i + 1)!} \cdot x^{2 \cdot i + 1}$$

gilt. Im vorliegenden Fall darf man die Bildung der unendlichen Reihe und die Ableitungsoperation miteinander vertauschen, so dass

$$h(x) = g'(x) = \sum_{i=0}^{\infty} \frac{(-1)^i \cdot (2 \cdot i + 1)}{(2 \cdot i + 1)!} \cdot x^{2 \cdot i} = \sum_{i=0}^{\infty} \frac{(-1)^i}{(2 \cdot i)!} \cdot x^{2 \cdot i}$$

Gilt (man hätte die Taylorentwicklung auch direkt herleiten können).

Die Ausgangsbedingungen $g(0) = 0$, $h(0) = 1$, $g'(x) = h(x)$ und $h'(x) = -g(x)$ legen die Funktionen *g* und *h* eindeutig fest. Dieser Ansatz bildet die Grundlage einer **analytischen Definition** der **Sinusfunktion** und der **Kosinusfunktion**:

$$\sin(x)=\sum_{i=0}^{\infty}\frac{(-1)^i}{(2\cdot i+1)!}\cdot x^{2\cdot i+1}=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\ldots+\frac{(-1)^n}{(2\cdot n+1)!}\cdot x^{2\cdot n+1}+\sum_{i=n+1}^{\infty}\frac{(-1)^i}{(2\cdot i+1)!}\cdot x^{2\cdot i+1}$$

für $x\in\mathbf{R}$.

$$\cos(x)=\sum_{i=0}^{\infty}\frac{(-1)^i}{(2\cdot i)!}\cdot x^{2\cdot i}=1-\frac{x^2}{2}+\frac{x^4}{4!}-\frac{x^6}{6!}+\ldots+\frac{(-1)^n}{(2\cdot n)!}\cdot x^{2\cdot n}+\sum_{i=n+1}^{\infty}\frac{(-1)^i}{(2\cdot i)!}\cdot x^{2\cdot i}$$

für $x\in\mathbf{R}$.

Es zeigt sich, dass diese Funktionen dieselben Eigenschaften besitzen wie die über geometrische Betrachtungen am rechtwinkligen Dreieck eingeführten Sinus- und Kosinusfunktionen. Einige wichtige Formeln werden hier zusammengestellt:

Aus der Definition folgt unmittelbar

$$\sin(-x)=-\sin(x),$$
$$\cos(-x)=\cos(x),$$
$$\frac{d}{dx}\sin(x)=\cos(x),$$
$$\frac{d}{dx}\cos(x)=-\sin(x).$$

Definiert man für festes $y\in\mathbf{R}$ die Funktion f durch $f(x)=\cos(x+y)$ und entwickelt diese bei $x_0=0$ in eine Taylorreihe, d.h. $f(x)=\sum_{k=0}^{\infty}\frac{f^{(k)}(0)}{k!}\cdot x^k$, so erhält man mit

$f''(x)=\frac{d}{dx}(-\sin(x+y))=-\cos(x+y)=-f(x)$:

Für $k\in\mathbf{N}$ ist $f^{(2\cdot k)}(x)=(-1)^k\cdot f(x)$ und $f^{(2\cdot k+1)}(x)=(-1)^k\cdot f'(x)$, also

$$f(x)=\cos(x+y)=f(0)\cdot\sum_{k=0}^{\infty}\frac{(-1)^k}{(2\cdot k)!}\cdot x^{2\cdot k}+f'(0)\cdot\sum_{k=0}^{\infty}\frac{(-1)^k}{(2\cdot k+1)!}\cdot x^{2\cdot k+1}$$
$$=f(0)\cdot\cos(x)+f'(0)\cdot\sin(x),$$

also

$$\cos(x+y)=\cos(x)\cdot\cos(y)-\sin(x)\cdot\sin(y).$$

Durch Ableiten von $f(x)=f(0)\cdot\cos(x)+f'(0)\cdot\sin(x)$ erhält man
$-\sin(x+y)=-f(0)\cdot\sin(x)+f'(0)\cdot\cos(x)$ und

$$\sin(x+y)=\sin(x)\cdot\cos(y)+\cos(x)\cdot\sin(y).$$

Damit ergibt sich

$\cos(x-y)=\cos(x)\cdot\cos(y)+\sin(x)\cdot\sin(y)$ und
$\sin(x-y)=\sin(x)\cdot\cos(y)-\cos(x)\cdot\sin(y)$.

Mit $x=y$ und $\cos(0)=1$ folgt der **Satz des Pythagoras**:

$(\sin(x))^2+(\cos(x))^2=1$.

Weiterhin ist mit $x=y$ $\cos(x+x)=\cos(2\cdot x)$, also

$\cos(2\cdot x)=(\cos(x))^2-(\sin(x))^2$ und
$\sin(2\cdot x)=2\cdot\sin(x)\cdot\cos(x)$.

Die Taylorentwicklung des Kosinus $\cos(x)=\sum_{i=0}^{\infty}\frac{(-1)^i}{(2\cdot i)!}\cdot x^{2\cdot i}$ ist eine unendliche Reihe mit alternierenden Folgengliedern, und es gilt für $i\geq 1$ und $0<x<3$ wie man leicht nachrechnet: $\frac{x^{2\cdot i}}{(2\cdot i)!}>\frac{x^{2\cdot(i+1)}}{(2\cdot(i+1))!}$. Damit ergibt sich mit Satz 5-1.13
$1-\frac{x^2}{2}<\cos(x)\leq 1-\frac{x^2}{2}+\frac{x^4}{24}$ für $0<x<3$.

Das Polynom $p(x)=1-\frac{x^2}{2}$ hat die Nullstellen $x_{p,1}=-\sqrt{2}$ und $x_{p,2}=\sqrt{2}$. Die Nullstellen des Polynoms $q(x)=1-\frac{x^2}{2}+\frac{x^4}{24}$ erhält man mit der Substitution $z=x^2$: $z_{q,1}=6-2\cdot\sqrt{3}$ und $z_{q,2}=6+2\cdot\sqrt{3}$ sind Nullstellen von $1-\frac{z}{2}+\frac{z^2}{24}$, und damit sind $x_{q,1}=-\sqrt{6+2\cdot\sqrt{3}}$, $x_{q,2}=-\sqrt{6-2\cdot\sqrt{3}}$, $x_{q,3}=\sqrt{6-2\cdot\sqrt{3}}$ und $x_{q,4}=\sqrt{6+2\cdot\sqrt{3}}$ die Nullstellen von *q* (nach aufsteigender Größe sortiert). Das bedeutet $0<\cos(\sqrt{2})$ und $\cos\left(\sqrt{6-2\cdot\sqrt{3}}\right)\leq 0$. Nach Satz 5.2-3 gibt es (mindestens) eine Nullstelle x_0 der Kosinusfunktion im Intervall $\left[\sqrt{2},\sqrt{6-2\cdot\sqrt{3}}\right]$.

Die **kleinste Nullstelle** der Kosinusfunktion $\cos(x)$ für $x>0$ **definiert den Wert** $\pi/2$.

Insgesamt $\cos(\pi/2)=0$ und $\pi\leq 2\cdot\sqrt{6-2\cdot\sqrt{3}}\approx 3{,}1849$.

Mit dem Satz des Pythagoras folgt $1=(\sin(\pi/2))^2+(\cos(\pi/2))^2=(\sin(\pi/2))^2$, also $\sin(\pi/2)=\pm 1$. Für $x\in\mathbf{R}$ mit $0\le x<\pi/2$ ist $\cos(x)>0$ und wegen $\frac{d}{dx}\sin(x)=\cos(x)$ ist die Sinusfunktion für diese Werte *x* streng monoton steigend. Die Stetigkeit der Sinusfunktion lässt daher nur den Wert $\sin(\pi/2)=1$ zu.

Aus $\sin(x+y)=\sin(x)\cdot\cos(y)+\cos(x)\cdot\sin(y)$ folgt:

$$\sin(x+\pi/2)=\sin(x)\cdot\cos(\pi/2)+\cos(x)\cdot\sin(\pi/2)=\cos(x)$$
$$\sin(x+\pi/2)=\cos(x)=\cos(-x)=\sin(\pi/2-x).$$

Ersetzt man in der letzten Formel *x* durch $x+\pi/2$, so erhält man

$$\sin(x+\pi)=\sin(-x)=-\sin(x).$$

Hier wird *x* durch $x+\pi$ ersetzt, und man erhält

$$\sin(x+2\cdot\pi)=-\sin(x+\pi)=\sin(x).$$

Die Nullstellen der Sinusfunktion sind die Werte $x=k\cdot\pi$ mit $k\in\mathbf{Z}$.

Entsprechend ergibt sich die Herleitung von

$$\cos(x+2\cdot\pi)=\cos(x).$$

Die Nullstellen der Kosinusfunktion sind die Werte $x=\frac{\pi}{2}+k\cdot\pi$ mit $k\in\mathbf{Z}$.

Über die Sinus- und Kosinusfunktion definiert man wieder die übrigen trigonometrischen Funktionen:

Tangensfunktion $\tan(x)=\frac{\sin(x)}{\cos(x)}$ für $x\in\mathbf{R}$ mit $x\neq\frac{\pi}{2}+k\cdot\pi$ für $k\in\mathbf{Z}$,

Kotangensfunktion $\cot(x)=\frac{\cos(x)}{\sin(x)}$ für $x\in\mathbf{R}$ mit $x\neq k\cdot\pi$ für $k\in\mathbf{Z}$.

Auf weitere Details wird hier mit Verweis auf die angegebene Literatur verzichtet.

Vergleicht man die Taylorreihenentwicklung der Kosinus- und der Sinusfunktion mit derjenigen der Exponentialfunktion, so fällt die große Ähnlichkeit auf. Es gelten folgende Zusammenhänge zwischen der Exponential-, der Sinus- und der Kosinusfunktion:

Erweitert man die Definition der Exponentialfunktion von den reellen Zahlen auf die komplexen Zahlen (siehe Kapitel 1.4), so erhält man die komplexwertige Exponentialfunktion

$$\exp: \begin{cases} \mathbf{C} & \to & \mathbf{C} \\ z & \to & \sum_{k=0}^{\infty} \frac{z^k}{k!} \end{cases} .$$

Die unendliche Reihe in dieser Definition wird wieder über endliche Partialsummen definiert, wobei hierbei alle Operationen in den komplexen Zahlen ausgeführt werden. Die bei den Konvergenzbetrachtungen auftretenden Betragswerte sind dann wegen $|z| = |a + bi| = \sqrt{a^2 + b^2}$ reelle Zahlen, so dass Konvergenzbetrachtungen von den komplexen Zahlen auf die reellen Zahlen übertragen werden und in **R** stattfinden. Es lässt sich zeigen, dass auch hier wieder $\sum_{k=0}^{\infty} \frac{z^k}{k!}$ absolut konvergiert, so dass der Grenzwert $\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!}$ mit $z \in \mathbf{C}$ in der Tat existiert. Zur Abkürzung schreibt man anstelle von $\exp(z)$ wieder e^z und meint damit den Grenzwert $\sum_{k=0}^{\infty} \frac{z^k}{k!}$.

Nun gilt für die imaginäre Zahl i: $i^2 = -1$. Für $x \in \mathbf{R}$ ergibt sich:

$$\begin{aligned} \frac{e^{i\cdot x} - e^{-i\cdot x}}{2} &= \frac{1}{2} \cdot \left(\sum_{k=0}^{\infty} \frac{(i\cdot x)^k}{k!} - \sum_{k=0}^{\infty} \frac{(-i\cdot x)^k}{k!} \right) \\ &= \frac{1}{2} \cdot \sum_{k=0}^{\infty} \frac{(i\cdot x)^k - (-i\cdot x)^k}{k!} \\ &= \frac{1}{2} \cdot \sum_{k=0}^{\infty} \frac{2\cdot (i\cdot x)^{2\cdot k+1}}{(2\cdot k+1)!} = i \cdot \sum_{k=0}^{\infty} \frac{i^{2\cdot k} \cdot x^{2\cdot k+1}}{(2\cdot k+1)!} = i \cdot \sum_{k=0}^{\infty} \frac{(-1)^k \cdot x^{2\cdot k+1}}{(2\cdot k+1)!} = i \cdot \sin(x) \quad , \end{aligned}$$

$$\begin{aligned} \frac{e^{i\cdot x} + e^{-i\cdot x}}{2} &= \frac{1}{2} \cdot \left(\sum_{k=0}^{\infty} \frac{(i\cdot x)^k}{k!} + \sum_{k=0}^{\infty} \frac{(-i\cdot x)^k}{k!} \right) \\ &= \frac{1}{2} \cdot \sum_{k=0}^{\infty} \frac{(i\cdot x)^k + (-i\cdot x)^k}{i!} \\ &= \frac{1}{2} \cdot \sum_{k=0}^{\infty} \frac{2\cdot (i\cdot x)^{2\cdot k}}{(2\cdot k)!} = \sum_{k=0}^{\infty} \frac{(-1)^k \cdot x^{2\cdot k}}{(2\cdot k)!} = \cos(x) \quad . \end{aligned}$$

Damit folgt

$$e^{i\cdot x} = \frac{e^{i\cdot x} + e^{-i\cdot x}}{2} + \frac{e^{i\cdot x} - e^{-i\cdot x}}{2} = \cos(x) + i\cdot\sin(x)\,.$$

Für $x \in \mathbf{R}$ ist also $\cos(x)$ der Realteil und $\sin(x)$ der Imaginärteil von $e^{i\cdot x}$.

Insbesondere ist $\left|e^{i\cdot x}\right| = \left|\cos(x) + i\cdot\sin(x)\right| = \sqrt{\left(\cos(x)\right)^2 + \left(\sin(x)\right)^2} = \sqrt{1} = 1$.

Die in Satz 5.5-1 für reelle Zahlen x und y definierte Funktionalgleichung $\exp(x+y) = \exp(x)\cdot\exp(y)$ lässt sich auch für die komplexwertige Exponentialfunktion nachweisen, so dass insgesamt für $z \in \mathbf{C}$, etwa $z = a + b\cdot i$ mit $a \in \mathbf{R}$ und $b \in \mathbf{R}$, gilt:

$$\exp(z) = \exp\left(a + i\cdot b\right) = \exp(a)\cdot\exp\left(i\cdot b\right) = \exp(a)\cdot\left(\cos(b) + i\cdot\sin(b)\right).$$

Auch die komplexe Exponentialfunktion ist wie die Sinus- und Kosinusfunktion periodisch:

$$\begin{aligned}\exp\left(i\cdot x\right) &= \cos(x) + i\cdot\sin(x)\\ &= \cos(x + 2\cdot\pi) + i\cdot\sin(x + 2\cdot\pi) = \exp\left(i\cdot\left(x + 2\cdot\pi\right)\right) = \exp\left(i\cdot x + i\cdot 2\cdot\pi\right)\,.\end{aligned}$$

5.10 Fibonacci-Zahlen

In Kapitel 2.1 werden die Fibonacci-Zahlen als Funktion definiert:

$$fib: \begin{cases} \mathbf{N} & \to & \mathbf{N} \\ n & \to & \begin{cases} n & \text{für } n = 0 \text{ und } n = 1 \\ fib(n-1) + fib(n-2) & \text{für } n \geq 2. \end{cases} \end{cases}$$

Die Fibonacci-Zahlen spielen in vielen Teilen der Mathematik und der Informatik (z.B. bei der Laufzeitberechnung des Zugriffs auf Daten, die in Form höhenbalancierter Bäume gespeichert sind) eine wichtige Rolle.

Zur Vereinfachung der Darstellung wird $F_n = fib(n)$ gesetzt, d.h. $\left(F_n\right)_{n\in\mathbf{N}}$ ist die Folge der Fibonacci-Zahlen. Die ersten elf Fibonacci-Zahlen lauten:

n	0	1	2	3	4	5	6	7	8	9	10
F_n	0	1	1	2	3	5	8	13	21	34	55

Gemäß obiger Definition ist

$$F_n = \begin{cases} n & \text{für } n = 0 \text{ oder } n = 1 \\ F_{n-2} + F_{n-1} & \text{für } n \geq 2. \end{cases}$$

Zur Berechnung der n-ten Fibonacci-Zahlen kann man z.B. folgende Pascal-Funktion einsetzen, die wohl elegant ist, aber schlechtes Laufzeitverhalten zeigt:

```
FUNCTION fib_1 (n : INTEGER) : INTEGER;

BEGIN { fib_1 }
  IF n < 0 THEN Exit;
  CASE OF n
  0    : fib_1 := 0;
  1    : fib_1 := 1;
  ELSE fib_1 := fib_1(n-2) + fib_1(n-1);
  END;
END   { fib_1 };
```

Optimales Laufzeitverhalten zeigt folgende Pascal-Funktion, die zur Berechnung der n-ten Fibonacci-Zahl F_n nacheinander alle Fibonacci-Zahlen F_i mit $0 \leq i < n$ nach dem Prinzip der Dynamischen Programmierung berechnet:

```
FUNCTION fib_2 (n : INTEGER) : INTEGER;

VAR f_n1, f_n2, f_n : INTEGER;
    idx             : INTEGER;

BEGIN { fib_2 }
   IF n < 0 THEN Exit;
   CASE OF n
   0    : fib_2 := 0;
   1    : fib_2 := 1;
   ELSE BEGIN
          f_n2 := 0;
          f_n1 := 1;
          FOR idx := n DOWNTO 2 DO
            BEGIN
              f_n  := f_n2 + f_n1;
              f_n2 := f_n1;
              f_n1 := f_n;
            END;
          fib_2 := f_n;
        END;
   END { CASE };
END   { fib_2 };
```

Die Fibonacci-Zahlen sind rekursiv definiert, und es ist wünschenswert, den Wert der *n*-ten Fibonacci-Zahl direkt in Abhängigkeit von *n* zu erhalten. Hier hilft eine spezielle mathematische Methode, die Methode der erzeugenden Funktionen, weiter, die hier nur auf das vorliegende Beispiel angewandt werden soll.

Zunächst fasst man die beiden Fälle der definierenden Gleichung der Fibonacci-Zahlen zu einer Gleichung zusammen. Dazu definiert man

$$F_{-1} = F_{-2} = 0$$

und erhält

$F_n = F_{n-1} + F_{n-2} + a_n$ für jedes $n \in \mathbf{N}$; hierbei ist $a_1 = 1$ und $a_n = 0$ für $n \neq 1$.

Beide Seiten werden mit x^n multipliziert und alle Werte aufaddiert; hier ist nicht gesagt, welchen Wert *x* annimmt, und auch Fragen der Konvergenz spielen zunächst keine Rolle. Man erhält:

$$\begin{aligned}\sum_{n=0}^{\infty} F_n \cdot x^n &= \sum_{n=0}^{\infty} F_{n-1} \cdot x^n + \sum_{n=0}^{\infty} F_{n-2} \cdot x^n + \sum_{n=0}^{\infty} a_n \cdot x^n \\ &= \sum_{n=0}^{\infty} F_n \cdot x^{n+1} + \sum_{n=0}^{\infty} F_n \cdot x^{n+2} + x \qquad \text{wegen } F_{-1} = F_{-2} = 0 \\ &= x \cdot \sum_{n=0}^{\infty} F_n \cdot x^n + x^2 \cdot \sum_{n=0}^{\infty} F_n \cdot x^n + x.\end{aligned}$$

Setzt man $F(x) = \sum_{n=0}^{\infty} F_n \cdot x^n$, so erhält man die Gleichung

$$F(x) = x \cdot F(x) + x^2 \cdot F(x) + x,$$

die man nach $F(x)$ auflöst:

$$F(x) = \frac{x}{1 - x - x^2}.$$

Diese Funktion erfüllt für $x_0 = 0$ die Voraussetzungen von Satz 5.7-2, so dass man versuchen könnte, die Taylorentwicklung an der Stelle $x_0 = 0$ herzuleiten. Wieder unter der Annahme, dass Konvergenz vorliegt, ergäbe sich dann:

$$F(x) = \sum_{i=0}^{\infty} \frac{F^{(i)}(0)}{i!} \cdot x^i = \sum_{n=0}^{\infty} F_n \cdot x^n \ .$$

Ein Koeffizientenvergleich lieferte $F_n = \frac{F^{(n)}(0)}{n!}$. Dieser Weg ist jedoch mühsam, da die Ableiungen $F^{(i)}(x)$ schwierig zu bestimmen sind. Daher wird ein anderer Weg beschritten:

Es werden Zahlen A, B, α und β mit Hilfe der Partialbruchzerlegung bestimmt, für die gilt:

$$F(x) = \frac{x}{1-x-x^2} = \frac{A}{1-\alpha \cdot x} + \frac{B}{1-\beta \cdot x}.$$

$$\begin{aligned} \frac{A}{1-\alpha \cdot x} + \frac{B}{1-\beta \cdot x} &= \frac{A \cdot (1-\beta \cdot x) + B \cdot (1-\alpha \cdot x)}{(1-\alpha \cdot x) \cdot (1-\beta \cdot x)} \\ &= \frac{A - A \cdot \beta \cdot x + B - B \cdot \alpha \cdot x}{(1-\alpha \cdot x) \cdot (1-\beta \cdot x)} \\ &= \frac{x}{1-x-x^2}. \end{aligned}$$

Der Koeffizientenvergleich ergibt:

$A + B - (A \cdot \beta + B \cdot \alpha) \cdot x = x$ und
$(1-\alpha \cdot x) \cdot (1-\beta \cdot x) = 1-x-x^2$.

Mit $x = 0$ folgt aus der ersten Gleichung

$A + B = 0$ bzw. $A = -B$.

Zur Bestimmung von α und β werden die Nullstellen von $1-x-x^2$ bestimmt. Diese lauten (siehe Kapitel 5.3):

$$x_{01} = -\frac{1}{2} \cdot (1+\sqrt{5}) \text{ und } x_{02} = -\frac{1}{2} \cdot (1-\sqrt{5}).$$

Damit ist

$$
\begin{aligned}
1-x-x^2 &= -(x-x_{01})\cdot(x-x_{02}) \\
&= -(x_{01}-x)\cdot(x_{02}-x) \\
&= -x_{01}\cdot x_{02}\cdot\left(1-\frac{x}{x_{01}}\right)\cdot\left(1-\frac{x}{x_{02}}\right) \\
&= \left(1-\frac{x}{x_{01}}\right)\cdot\left(1-\frac{x}{x_{02}}\right) \qquad \text{wegen } -x_{01}\cdot x_{02}=1 \\
&= (1-\alpha\cdot x)\cdot(1-\beta\cdot x)
\end{aligned}
$$

und folglich

$$
\begin{aligned}
\alpha &= \frac{1}{x_{01}} = -2\cdot\frac{1}{1+\sqrt{5}} = -2\cdot\frac{1-\sqrt{5}}{(1+\sqrt{5})\cdot(1-\sqrt{5})} = \frac{1}{2}\cdot(1-\sqrt{5}) \\
&\approx -0{,}618034
\end{aligned}
$$

und

$$
\begin{aligned}
\beta &= \frac{1}{x_{02}} = -2\cdot\frac{1}{1-\sqrt{5}} = \frac{1}{2}\cdot(1+\sqrt{5}) \\
&\approx 1{,}618034.
\end{aligned}
$$

Diese Werte für α und β werden in die Gleichung $A+B-(A\cdot\beta+B\cdot\alpha)\cdot x = x$ eingesetzt, wobei $A+B=0$ bzw. $A=-B$ bereits bekannt ist, und der Wert $x=1$ genommen wird. Man erhält

$$1 = -A\cdot\beta + A\cdot\alpha = A\cdot(\alpha-\beta) = A\cdot(-\sqrt{5}) \text{ bzw.}$$

$$A = -\frac{1}{\sqrt{5}} \text{ und } B = \frac{1}{\sqrt{5}}.$$

Damit ist

$$
\begin{aligned}
F(x) &= \frac{x}{1-x-x^2} \\
&= \frac{A}{1-\alpha\cdot x} + \frac{B}{1-\beta\cdot x} \\
&= \frac{1}{\sqrt{5}}\cdot\left(\frac{1}{1-\beta\cdot x} - \frac{1}{1-\alpha\cdot x}\right).
\end{aligned}
$$

Nun ist nach Satz 5.1-9 (i) für $|\alpha\cdot x|<1$ bzw. $|\beta\cdot x|<1$:

$$\frac{1}{1-\alpha\cdot x}=\sum_{i=0}^{\infty}(\alpha\cdot x)^i \text{ bzw. } \frac{1}{1-\beta\cdot x}=\sum_{i=0}^{\infty}(\beta\cdot x)^i .$$

und insgesamt

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{5}}\cdot\left(\frac{1}{1-\beta\cdot x}-\frac{1}{1-\alpha\cdot x}\right) \\ &= \sum_{n=0}^{\infty}\left(\frac{1}{\sqrt{5}}\cdot\beta^n-\frac{1}{\sqrt{5}}\cdot\alpha^n\right)\cdot x^n \\ &= \sum_{n=0}^{\infty}F_n\cdot x^n . \end{aligned}$$

Bemerkung: Diese Gleichung gilt wegen der Konvergenzanforderungen $|\alpha\cdot x|<1$ und $|\beta\cdot x|<1$ für jedes $x\in\mathbf{R}$ mit $|x|<\min\left\{|1/\alpha|,\|1/\beta\|\right\}=|1/\beta|<0{,}618034$, eine Tatsache, die hier nicht von Belang ist.

Der Koeffizientenvergleich liefert:

Die durch

$$F_n=\begin{cases} n & \text{für } n=0 \text{ oder } n=1 \\ F_{n-2}+F_{n-1} & \text{für } n\geq 2 \end{cases}$$

definierten Fibonacci-Zahlen erfüllen die Gleichung

$$F_n=\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n\right) \text{ für jedes } n\in\mathbf{N} .$$

5.11 Erzeugende Funktionen

Im vorliegenden Kapitel wird die in Kapitel 5.10 erwähnte Methode der erzeugenden Funktionen beschrieben und im Kapitel 5.12 auf weitere interessante Fragestellungen der Informatik angewendet.

Es sei $(g_n)_{n\in\mathbf{N}}$ eine Folge reeller Zahlen. Die **erzeugende Funktion** G der Folge $(g_n)_{n\in\mathbf{N}}$ wird durch $G(z)=\sum_{n=0}^{\infty} g_n \cdot z^n$ definiert. Hierbei ist z eine „formale“ (komplexe oder reelle) Variable.

Die erzeugende Funktion fasst die gesamte Information über die Folge $(g_n)_{n\in\mathbf{N}}$ in einem einzigen arithmetischen Ausdruck zusammen. Die z-Potenzen separieren dabei die einzelnen Folgenglieder. Natürlich kann man nach Konvergenzbedingungen in Abhängigkeit von z fragen. Im vorliegenden Zusammenhang spielen diese aber eine untergeordnete Rolle.

Aus der erzeugenden Funktion $G(z)=\sum_{n=0}^{\infty} g_n \cdot z^n$ lassen sich die einzelnen Folgenglieder zurückgewinnen: Es gilt nämlich (vgl. Kapitel 5.6) für die i-te Ableitung von $G(z)$:

$G^{(i)}(z)=\sum_{n=i}^{\infty} n\cdot(n-1)\cdot\ldots\cdot(n-i+1)\cdot g_n \cdot z^{n-i}$, also $G^{(n)}(0)=n!\cdot g_n$ bzw. $g_n=\frac{1}{n!}\cdot G^{(n)}(0)$.

Daher sind zwei Folgen mit derselben erzeugenden Funktion identisch.

In Kapitel 5.10 wird die erzeugende Funktion der Fibonacci-Folge $(F_n)_{n\in\mathbf{N}}$ berechnet zu $F(z)=\sum_{n=0}^{\infty} F_n \cdot z^n = \frac{z}{1-z-z^2}$. Hierbei ist zur Berechnung die Tatsache, dass diese Reihe nur für $|z|<2\cdot 1/\left(1+\sqrt{5}\right)\approx 0{,}618034$ konvergiert, ohne Belang. Die erzeugende Funktion der Fibonacci-Folge wird in eine Potenzreihe umgewandelt mit dem Ergebnis $F(z)=\frac{z}{1-z-z^2}=\sum_{n=0}^{\infty}\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n\right)\cdot z^n$. Daraus lässt sich die n-te Fibonacci-Zahl in geschlossener Form ablesen: $F_n=\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n\right)$.

Satz 5.11-1:

Es seien $(f_n)_{n\in\mathbf{N}}$ bzw. $(g_n)_{n\in\mathbf{N}}$ zwei Folgen mit den erzeugenden Funktionen $F(z)$ bzw. $G(z)$ und α und β Konstanten.

(i) Die erzeugende Funktion der Folge $(\alpha\cdot f_n+\beta\cdot g_n)_{n\in\mathbf{N}}$ lautet

$$\sum_{n=0}^{\infty}(\alpha\cdot f_n+\beta\cdot g_n)\cdot z^n=\alpha\cdot F(z)+\beta\cdot G(z).$$

(ii) Die erzeugende Funktion der Folge, die man erhält, indem man $(g_n)_{n\in\mathbf{N}}$ um m Plätze nach rechts verschiebt, also der Folge $\left(\underbrace{0,...,0}_{m},g_0,g_1,...\right)$, lautet

$$\sum_{n=m}^{\infty}g_{n-m}\cdot z^n=\sum_{n=0}^{\infty}g_n\cdot z^{n+m}=z^m\cdot\sum_{n=0}^{\infty}g_n\cdot z^n=z^m\cdot G(z).$$

(iii) Die erzeugende Funktion der Folge, die man erhält, indem man $(g_n)_{n\in\mathbf{N}}$ um m Plätze nach links verschiebt, also der Folge $(g_m,g_{m+1},g_{m+2},...)$, lautet

$$\sum_{n=0}^{\infty}g_{n+m}\cdot z^n=\frac{G(z)-g_0-g_1\cdot z-g_2\cdot z^2-...-g_{m-1}\cdot z^{m-1}}{z^m}.$$

(iv) Die erzeugende Funktion der Folge $(\alpha^n\cdot g_n)_{n\in\mathbf{N}}$ lautet

$$\sum_{n=0}^{\infty}\alpha^n\cdot g_n\cdot z^n=\sum_{n=0}^{\infty}g_n\cdot(\alpha\cdot z)^n=G(\alpha\cdot z).$$

../..

(v) Die erzeugende Funktion der Folge $\big((n+1)\cdot g_{n+1}\big)_{n\in\mathbf{N}} = \big(g_1, 2\cdot g_2, 3\cdot g_3, ...\big)$ lautet $\sum_{n=0}^{\infty}(n+1)\cdot g_{n+1}\cdot z^n = G'(z)$.

(vi) Die erzeugende Funktion der Folge $\big(n\cdot g_n\big)_{n\in\mathbf{N}} = \big(0, g_1, 2\cdot g_2, 3\cdot g_3, ...\big)$ lautet $\sum_{n=0}^{\infty} n\cdot g_n\cdot z^n = z\cdot G'(z)$.

(vii) Die erzeugende Funktion der Folge $\left(\frac{1}{n}\cdot g_{n-1}\right)_{n\in\mathbf{N}} = \left(0, g_0, \frac{1}{2}\cdot g_1, \frac{1}{3}\cdot g_2, ...\right)$, wobei das 0-te Folgenglied den Wert 0 hat, lautet $\sum_{n=1}^{\infty}\frac{1}{n}\cdot g_{n-1}\cdot z^n = \int_0^z G(t)dt$.

(viii) Die Faltung der Folgen $\big(f_n\big)_{n\in\mathbf{N}}$ und $\big(g_n\big)_{n\in\mathbf{N}}$, d.h. die Folge $\left(\sum_{k=0}^{n} f_k\cdot g_{n-k}\right)_{n\in\mathbf{N}}$, hat die erzeugende Funktion $\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} f_k\cdot g_{n-k}\right)\cdot z^n = F(z)\cdot G(z)$.

(ix) Die Folge, deren Glieder die *n*-ten Partialsummen der Folge $\big(g_n\big)_{n\in\mathbf{N}}$ bilden, d.h. die Folge $\big(g_0, g_0+g_1, g_0+g_1+g_2, ...\big) = \left(\sum_{k=0}^{n} g_k\right)_{n\in\mathbf{N}}$ hat die erzeugende Funktion $\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} g_k\right)\cdot z^n = \frac{1}{1-z}\cdot G(z)$.

(x) Die erzeugende Funktion der Folge, die aus $\big(g_n\big)_{n\in\mathbf{N}}$ entsteht, indem man die Folgenglieder mit ungeradem Index durch 0 ersetzt, d.h. der Folge $\big(g_0, 0, g_2, 0, g_4, 0, ...\big)$, lautet $\sum_{n=0}^{\infty} g_{2\cdot n}\cdot z^{2\cdot n} = \frac{G(z)+G(-z)}{2}$.

Die erzeugende Funktion der Folge, die aus $\big(g_n\big)_{n\in\mathbf{N}}$ entsteht, indem man die Folgenglieder mit geradem Index durch 0 ersetzt, d.h. der Folge $\big(0, g_1, 0, g_3, 0, g_5, 0, ...\big)$, lautet $\sum_{n=0}^{\infty} g_{2\cdot n+1}\cdot z^{2\cdot n+1} = \frac{G(z)-G(-z)}{2}$.

Die Aussagen (i) – (iv) sind unmittelbar einsichtig.

Aussage (v) ergibt sich aus

$$G'(z) = \left(\sum_{n=0}^{\infty} g_n \cdot z^n \right)' = g_1 + 2 \cdot g_2 \cdot z + 3 \cdot g_3 \cdot z^2 + \ldots$$
$$= \sum_{n=0}^{\infty} (n+1) \cdot g_{n+1} \cdot z^n \quad .$$

Aussage (vi) folgt aus (ii), indem man die Folge aus (v) um einen Platz nach rechts verschiebt.

Aussage (vii) ergibt sich wie folgt (hier wird benutzt, dass unter der Voraussetzung der Konvergenz die Operationen der Integral- und Summenbildung vertauschbar sind; hinzu kommt ein wenig elementares Schulwissen):
$\int_0^z G(t)dt = \int_0^z \left(\sum_{n=0}^{\infty} g_n \cdot t^n \right) dt = \sum_{n=0}^{\infty} g_n \cdot \int_0^z t^n dt = \sum_{n=0}^{\infty} g_n \cdot \frac{z^{n+1}}{n+1} = \sum_{n=1}^{\infty} \frac{1}{n} \cdot g_{n-1} \cdot z^n + 0 \cdot z^0$. Dazu gehört die Folge $\left(0, g_0, \frac{1}{2} \cdot g_1, \frac{1}{3} \cdot g_2, \ldots \right)$.

Für Aussage (viii) ist
$$F(z) \cdot G(z) = \left(\sum_{n=0}^{\infty} f_n \cdot z^n \right) \cdot \left(\sum_{n=0}^{\infty} g_n \cdot z^n \right)$$
$$= f_0 \cdot g_0 + (f_0 \cdot g_1 + f_1 \cdot g_0) \cdot z + (f_0 \cdot g_2 + f_1 \cdot g_1 + f_2 \cdot g_0) \cdot z^2 + \ldots$$
$$= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} f_k \cdot g_{n-k} \right) \cdot z^n$$
zu beachten.

Die erzeugende Funktion der konstanten Folge $(1,1,1,\ldots) = (1)_{n \in \mathbf{N}}$ lautet $F(z) = \sum_{n=0}^{\infty} z^n = \frac{1}{1-z}$.

Daher ist mit (viii) $\frac{1}{1-z} \cdot G(z)$ die erzeugende Funktion der Faltung von $(1,1,1,\ldots) = (1)_{n \in \mathbf{N}}$ und $(g_n)_{n \in \mathbf{N}}$. Das n-te Folgenglied der Faltung lautet $\sum_{k=0}^{n} 1 \cdot g_{n-k} = \sum_{k=0}^{n} g_k$. Das zeigt die Aussage (ix).

Aussage (x) rechnet man direkt nach:
$G(z) + G(-z) = \sum_{n=0}^{\infty} g_n \cdot \left(1 + (-1)^n\right) \cdot z^n = \sum_{\substack{n=0 \\ n \text{ ist gerade}}}^{\infty} g_n \cdot 2 \cdot z^n = 2 \cdot \sum_{n=0}^{\infty} g_{2 \cdot n} \cdot z^{2 \cdot n}$. Entsprechend ist

$G(z) - G(-z) = \sum_{n=0}^{\infty} g_n \cdot \left(1 - (-1)^n\right) \cdot z^n = \sum_{\substack{n=0 \\ n \text{ ist ungerade}}}^{\infty} g_n \cdot 2 \cdot z^n = 2 \cdot \sum_{n=0}^{\infty} g_{2 \cdot n+1} \cdot z^{2 \cdot n+1}$.

Die folgende Zusammenstellung zeigt einige wichtige Beispiele von Folgen mit ihren erzeugenden Funktionen. Auch hier sollen Fragen des Konvergenzverhaltens bezüglich der formalen Variablen z unberücksichtigt bleiben.

Folge $(g_n)_{n\in\mathbf{N}}$	erzeugende Funktion	geschlossene Form
(i) $g_0 = 1$, $g_n = 0$ für $n \geq 1$: $(1, 0, 0, 0, \ldots)$	$1 \cdot z^0$	$G(z) = 1$
(ii) $g_m = 1$, $g_n = 0$ für $n \neq m$, $m \in \mathbf{N}$ fest: $\left(0, \ldots, 0, \underbrace{1}_{\text{Position } m}, 0, 0, 0, \ldots\right)$	$1 \cdot z^m$	$G(z) = z^m$
(iii) $g_n = 1$ für $n \in \mathbf{N}$: $(1, 1, 1, 1, \ldots)$	$\sum_{n=0}^{\infty} z^n$	$G(z) = \frac{1}{1-z}$
(iv) $g_{2\cdot i} = 1$, $g_{2\cdot i+1} = -1$ für $i \in \mathbf{N}$: $(1, -1, 1, -1, \ldots)$	$\sum_{n=0}^{\infty} (-1)^n \cdot z^n$	$G(z) = \frac{1}{1+z}$
(v) $g_{2\cdot i} = 1$, $g_{2\cdot i+1} = 0$ für $i \in \mathbf{N}$: $(1, 0, 1, 0, 1, \ldots)$	$\sum_{n=0}^{\infty} z^{2\cdot n}$	$G(z) = \frac{1}{1-z^2}$
(vi) $g_{m\cdot n} = 1$ für $n \in \mathbf{N}$, $g_i = 0$ sonst: $(1, 0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots)$	$\sum_{n=0}^{\infty} z^{m\cdot n}$	$G(z) = \frac{1}{1-z^m}$
(vii) $g_n = n+1$ für $n \in \mathbf{N}$: $(1, 2, 3, 4, \ldots)$	$\sum_{n=0}^{\infty} (n+1) \cdot z^n$	$G(z) = \frac{1}{(1-z)^2}$
(viii) $g_n = c^n$ für $n \in \mathbf{N}$, $c \in \mathbf{R}$ fest: $(1, c, c^2, c^3, \ldots)$	$\sum_{n=0}^{\infty} c^n \cdot z^n$	$G(z) = \frac{1}{1-c\cdot z}$
(ix) $g_n = \binom{m}{n}$ für $n \in \mathbf{N}$, $m \in \mathbf{N}$ fest: $\left(1, m, \binom{m}{2}, \binom{m}{3}, \ldots, m, 1, 0, 0, \ldots\right)$	$\sum_{n=0}^{\infty} \binom{m}{n} \cdot z^n$	$G(z) = (1+z)^m$
(x) $g_n = \binom{m+n-1}{n}$ für $n \in \mathbf{N}$, $m \in \mathbf{N}$, $m \geq 1$ fest: $\left(1, m, \binom{m+1}{2}, \binom{m+2}{3}, \ldots\right)$	$\sum_{n=0}^{\infty} \binom{m+n-1}{n} \cdot z^n$	$G(z) = \frac{1}{(1-z)^m}$

../..

(xi) $g_n = \binom{m+n}{n}$ für $n \in \mathbf{N}$, $m \in \mathbf{N}$ fest: $\left(1, \binom{m+1}{2}, \binom{m+2}{3}, \binom{m+3}{3}, \ldots\right)$	$\sum_{n=0}^{\infty} \binom{m+n}{n} \cdot z^n$	$G(z) = \frac{1}{(1-z)^{m+1}}$
(xii) $g_0 = 0$, $g_n = 1/n$ für $n \geq 1$: $(0, 1, 1/2, 1/3, 1/4 \ldots)$	$\sum_{n=1}^{\infty} 1/n \cdot z^n$	$G(z) = \ln\left(\frac{1}{1-z}\right)$
(xiii) $g_0 = 0$, $g_n = (-1)^{n+1} \cdot 1/n$ für $n \geq 1$: $(0, 1, -1/2, 1/3, -1/4 \ldots)$	$\sum_{n=1}^{\infty} (-1)^{n+1} \cdot 1/n \cdot z^n$	$G(z) = \ln(1+z)$
(xiv) $g_0 = 0$, $g_n = 1/(n!)$ für $n \geq 1$: $(1, 1, 1/2, 1/6, 1/24, 1/120, \ldots)$	$\sum_{n=0}^{\infty} 1/(n!) \cdot z^n$	$G(z) = e^z$

Die Berechnung der geschlossenen Form der erzeugenden Funktion in den Beispielen (i) – (vi) und (viii) ist klar bzw. ergibt sich aus Satz 5.1-9. Beispiel (vii) kann ebenfalls mit Satz 5.1-9 oder auch mit Satz 5.11-1 (v) begründet werden: Setzt man dort $g_n = 1$, so folgt mit Beispiel (iii):

$$\sum_{n=0}^{\infty} (n+1) \cdot z^n = \sum_{n=0}^{\infty} (n+1) \cdot 1 \cdot z^n = \left(\frac{1}{1-z}\right)' = \frac{1}{(1-z)^2}.$$

Beispiel (ix) ist die binomische Formel.

Beispiel (x) lässt sich durch Induktion über *m* zeigen:

Für $m = 1$ ist $\sum_{n=0}^{\infty} \binom{1+n-1}{n} \cdot z^n = \sum_{n=0}^{\infty} z^n = 1/(1-z) = 1/(1-z)^1$. Die Behauptung gelte für $m \geq 1$.

Dann ist $\sum_{n=0}^{\infty} \binom{m+1+n-1}{n} \cdot z^n = \sum_{n=0}^{\infty} \binom{m+n}{n} \cdot z^n$. Nach Satz 4.1-3 (v) (dort für *i* den Wert *n* und für *n* den Ausdruck $m+n-1$ einsetzen) ist $\binom{m+n}{n} = \sum_{k=0}^{n} \binom{m-1+k}{k}$. Setzt man $g_n = \binom{m+n-1}{n}$, so ist $\sum_{n=0}^{\infty} \binom{m+n}{n} \cdot z^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \binom{m-1+k}{k} \right) \cdot z^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} g_k \right) \cdot z^n$. Nach Induktionsvoraussetzung lautet die erzeugende Funktion der Folge $(g_n)_{n \in \mathbf{N}}$:

$G(z) = \sum_{n=0}^{\infty} \binom{m+n-1}{n} \cdot z^n = \frac{1}{(1-z)^m}$. Mit Satz 5.11-1 (ix) folgt

$$\sum_{n=0}^{\infty} \binom{m+n}{n} \cdot z^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} g_k \right) \cdot z^n = \frac{1}{1-z} \cdot G(z) = \frac{1}{(1-z)^{m+1}}.$$

Beispiel (xi) ergibt sich aus (x) wie im Induktionsschritt zu (x).

Die Folge $(0,1,1/2,1/3,1/4...)$ in Beispiel (xii) hat die erzeugende Funktion $\sum_{n=1}^{\infty} 1/n \cdot z^n = \sum_{n=1}^{\infty} 1/n \cdot 1 \cdot z^n = \sum_{n=1}^{\infty} 1/n \cdot g_{n-1} \cdot z^n$ mit der konstanten Folge $(g_n)_{n\in\mathbf{N}} = (1)_{n\in\mathbf{N}}$. Die Folge $(g_n)_{n\in\mathbf{N}}$ hat nach (iii) die erzeugende Funktion $G(z) = \frac{1}{1-z}$. Mit Satz 5.11-1 (vii) folgt

$$\sum_{n=1}^{\infty} 1/n \cdot z^n = \sum_{n=1}^{\infty} \frac{1}{n} \cdot g_{n-1} \cdot z^n = \int_0^z G(t)dt = \int_0^z \frac{1}{1-t} dt = \int_1^{1-z} \left(-\frac{1}{x}\right) dx = -\ln(1-z) = \ln\left(\frac{1}{1-z}\right).$$

Die Folge $(0,1,-1/2,1/3,-1/4...)$ in Beispiel (xiii) hat die erzeugende Funktion

$$\sum_{n=1}^{\infty} (-1)^{n+1} \cdot 1/n \cdot z^n = -\sum_{n=1}^{\infty} 1/n \cdot (-z)^n = -\ln\left(\frac{1}{1+z}\right) = \ln(1+z).$$

Beispiel (xiv) ist die Taylorentwicklung der Exponentialfunktion e^z.

Erzeugende Funktionen stellen ein sehr gutes Werkzeug bei der **Auflösung von rekursiv definierten Folgen** zur Verfügung.

Dabei geht man wie folgt vor:

Gegeben sei die Folge $(g_n)_{n\in\mathbf{N}}$ über ein rekursives Gleichungssystem. Gesucht ist eine geschlossene Darstellung von g_n in Abhängigkeit von *n*.

1. Schritt: Man schreibt eine einzige Gleichung auf, die g_n mit Hilfe anderer Folgenglieder definiert. Diese Gleichung sollte für alle *n* gelten; eventuell muss man Folgenglieder mit negativen Indizes formal anfügen, für die dann $g_{-1} = g_{-2} = ... = 0$ gesetzt wird.

2. Schritt: Beide Seiten der Gleichung aus dem 1. Schritt werden nacheinander mit z^n multipliziert und jeweils aufsummiert, so dass auf der linken Seite der Gleichung die erzeugende Funktion $G(z) = \sum_{n=0}^{\infty} g_n \cdot z^n$ entsteht. Die rechte Seite der Gleichung wird so umgeformt, dass sie ebenfalls einen arithmetischen Ausdruck in $G(z)$ darstellt.

3. Schritt: Die Gleichung aus dem 2. Schritt wird nach $G(z)$ aufgelöst.

4. Schritt: $G(z)$ wird in eine formale Potenzreihe entwickelt. Der Koeffizient von z^n ist g_n in geschlossener Form.

Bemerkung: Der komplexeste Schritt ist im allgemeinen der 4. Schritt.

Zur Illustration wird das Verfahren an der Darstellung der Fibonacci-Zahlen in geschlossener Form gezeigt. Die Einzelheiten sind in Kapitel 5.10 bereits ausgeführt.

Die Folge $(F_n)_{n\in\mathbf{N}}$ der Fibonacci-Zahlen ist rekursiv definiert durch

$$F_n = \begin{cases} n & \text{für } n = 0 \text{ oder } n = 1 \\ F_{n-1} + F_{n-2} & \text{für } n \geq 2 \end{cases} .$$

Die Ergebnisse der einzelnen Schritte lauten:

1. Schritt: Mit $F_{-1} = F_{-2} = 0$ ist $F_n = F_{n-1} + F_{n-2} + a_n$ für jedes $n \in \mathbf{N}$; hierbei ist $a_1 = 1$ und $a_n = 0$ für $n \neq 1$.

2. Schritt:

$$\begin{aligned} F(z) &= \sum_{n=0}^{\infty} F_n \cdot z^n = \sum_{n=0}^{\infty} F_{n-1} \cdot z^n + \sum_{n=0}^{\infty} F_{n-2} \cdot z^n + \sum_{n=0}^{\infty} a_n \cdot z^n \\ &= \sum_{n=0}^{\infty} F_n \cdot z^{n+1} + \sum_{n=0}^{\infty} F_n \cdot z^{n+2} + z \\ &= z \cdot \sum_{n=0}^{\infty} F_n \cdot z^n + z^2 \cdot \sum_{n=0}^{\infty} F_n \cdot z^n + z \\ &= z \cdot F(z) + z^2 F(z) + z \end{aligned}$$

3. Schritt: $$F(z) = \frac{z}{1 - z - z^2}$$

4. Schritt:

$$\begin{aligned} F(z) &= \sum_{n=0}^{\infty} \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right) \cdot z^n \\ &= \sum_{n=0}^{\infty} F_n \cdot z^n . \end{aligned}$$

5.12 Anzahlbetrachtungen in Binärbäumen

Eine der wichtigsten Datenstrukturen, die in der Informatik vorkommen, sind Binärbäume. Dazu einige einführende Definitionen:

Ein **gerichteter Graph** $G = (V, E)$ besteht aus einer endlichen Menge $V = \{v_1, \ldots, v_n\}$ von **Knoten** (vertices) und einer endlichen Menge $E = \{e_1, \ldots, e_k\} \subseteq V \times V$ von **Kanten** (edges).

Die Kante $e = (v_i, v_j)$ läuft von v_i nach v_j (verbindet v_i mit v_j). Der Knoten v_i heißt **Anfangsknoten** der Kante $e = (v_i, v_j)$, der Knoten v_j **Endknoten** von $e = (v_i, v_j)$. Zu einem Knoten $v \in V$ heißt $pred(v) = \{v' \mid (v', v) \in E\}$ die **Menge der direkten Vorgänger** von v, $succ(v) = \{v' \mid (v, v') \in E\}$ die **Menge der direkten Nachfolger** von v.

Ein **Binärbaum** $B_n = (V, E)$ mit n Knoten wird durch folgende Eigenschaften 1. – 4. charakterisiert:

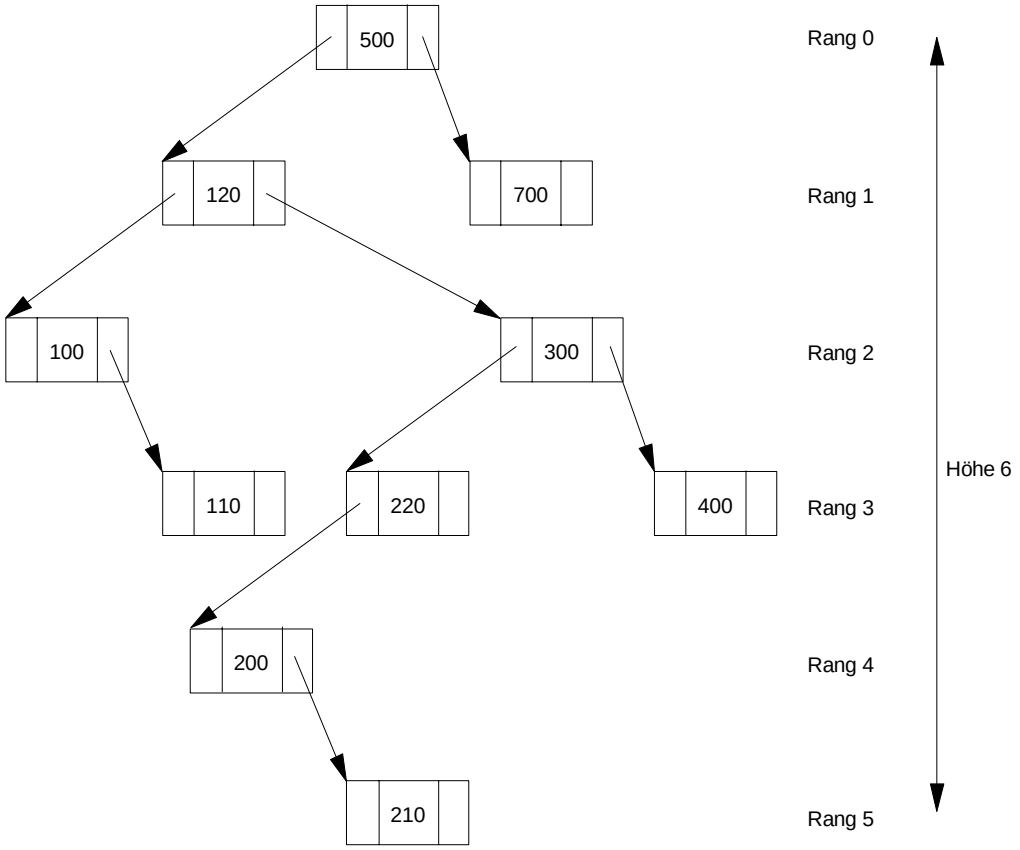
1. Entweder ist $n \geq 1$ und $|V| = n \geq 1$ und $|E| = n - 1$,
   oder es ist $n = 0$ und $V = E = \varnothing$ (**leerer Baum**)

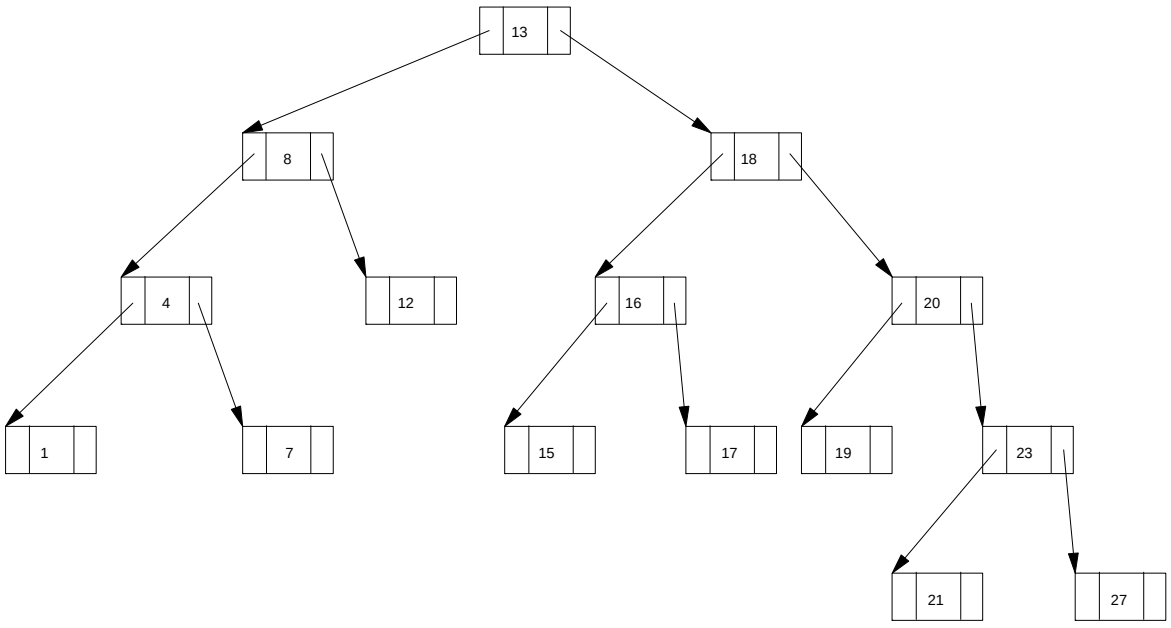
2. Bei $n \geq 1$ gibt es genau einen Knoten $r \in V$, dessen Menge direkter Vorgänger leer ist; dieser Knoten heißt **Wurzel** von B_n.

3. Bei $n \geq 1$ besteht die Menge der direkten Vorgänger eines jeden Knotens, der nicht die Wurzel ist, aus genau einem Element.

4. Bei $n \geq 1$ besteht die Menge der direkten Nachfolger eines jeden Knotens aus einem Element oder zwei Elementen oder ist leer. Ein Knoten, dessen Menge der direkten Nachfolger leer ist, heißt **Blatt**.

Beispiele:


500
120
700
100
300
110
220
400
200
210
Rang 0
Rang 1
Rang 2
Rang 3
Rang 4
Rang 5
Höhe 6


13
8
18
4
12
16
20
1
7
15
17
19
23
21
27

In einem Binärbaum $B = (V, E)$ gibt es für jeden Knoten $v \in V$ genau einen **Pfad** von der Wurzel r zu v, d.h. es gibt eine Folge $\big((a_0, a_1), (a_1, a_2), \ldots, (a_{m-1}, a_m)\big)$ mit $r = a_0$, $v = a_m$ und $(a_{i-1}, a_i) \in E$ für $i = 1, ..., m$. Der Wert m gibt die **Länge des Pfads** an. Um den Knoten v von der Wurzel aus über die Kanten des Pfads zu erreichen, werden m Kanten durchlaufen. Diese Länge wird auch als **Rang des Knotens** v bezeichnet.

Der Rang eines Knotens lässt sich auch folgendermaßen definieren:

1. Die Wurzel hat den Rang 0.
2. Ist v ein Knoten im Baum mit Rang $r-1$ und w ein direkter Nachfolger von v, so hat w den Rang r.

Unter der **Höhe eines Binärbaums** versteht man den maximal vorkommenden Rang eines Blattes + 1.

Der zweite Binärbaum zeichnet sich dadurch aus, dass sich die Höhen der Teilbäume, die von einem Knoten ausgehen, höchstens um 1 unterscheiden. Bäume mit dieser Eigenschaft heißen **AVL-Bäume**.

In einem Binärbaum bilden alle Knoten mit demselben Rang ein **Niveau des Baums**. Das Niveau 0 eines Binärbaums enthält genau einen Knoten, nämlich die Wurzel. Das Niveau 1 enthält mindestens 1 und höchstens 2 Knoten. Das Niveau j enthält höchstens doppelt so viele Knoten wie das Niveau $j-1$. Daher gilt:

Satz 5.12-1:

(i) Das Niveau $j \geq 0$ eines Binärbaums enthält mindestens einen und höchstens 2^j Knoten. Die Anzahl der Knoten vom Niveau 0 bis zum Niveau j (einschließlich) beträgt mindestens $j+1$ Knoten und höchstens $\sum_{i=0}^{j} 2^j = 2^{j+1} - 1$ Knoten.

(ii) Ein Binärbaum hat maximale Höhe, wenn jedes Niveau genau einen Knoten enthält. Er hat minimale Höhe, wenn jedes Niveau eine maximale Anzahl von Knoten enthält. Also gilt für die Höhe $h(B_n)$ eines Binärbaums mit n Knoten:

$$\lfloor \log_2(n) \rfloor + 1 = \lceil \log_2(n+1) \rceil \leq h(B_n) \leq n .$$

(iii) Für die Höhe $h(B_n)$ eines AVL-Baums mit n Knoten gilt

$$\lceil \log_2(n+1) \rceil \leq h(B_n) < 1{,}4404201 \cdot \log_2(n+2).$$

Aussage (i) ergibt sich durch vollständige Induktion.

Aussage (ii) ergibt sich aus folgenden Überlegungen: Die obere Abschätzung $h(B_n) \le n$ ist offensichtlich. Für die untere Abschätzung betrachtet man einen Binärbaum mit *n* Knoten und minimaler Höhe *h*. Jedes Niveau, bis eventuell das höchste Niveau *m*, ist vollständig gefüllt. Es ist $h = m+1$. Bis zum Niveau $m-1$ enthält der Binärbaum gemäß (i) insgesamt $2^{m-1+1}-1 = 2^m - 1$ viele Knoten, auf Niveau *m* sind es mindestens einer und höchstens 2^m. Daraus folgt: $2^m - 1 + 1 \le n \le 2^m - 1 + 2^m$, also $2^m \le n \le 2^{m+1} - 1$ und damit $m < \log_2(n+1) \le m+1$, d.h. $\lceil \log_2(n+1) \rceil = m+1 = h$. Für einen beliebigen Binärbaum mit *n* Knoten gilt daher $\lceil \log_2(n+1) \rceil \le h(B_n)$.

Aussage (iii) zeigt, dass die Höhe eines AVL-Baums durch eine logarithmischen Größenordnung, gemessen in der Anzahl der Knoten bewegt. Aussage (iii) lässt sich mit Hilfe der Methode der erzeugenden Funktion (siehe Ende des Kapitels 5.11) nachweisen. Dazu werden die dort beschriebenen 4 Schritte durchgeführt.

Da ein Binärbaum, dessen Niveaus, bis eventuell das höchste Niveau *m*, vollständig gefüllt sind, ein AVL-Baum ist, folgt die untere Abschätzung $\lceil \log_2(n+1) \rceil \le h(B_n)$.

Es sei ein AVL-Baum mit Höhe $h+1$ gegeben, der eine minimale Knotenanzahl enthält. Dann sind unter der Wurzel zwei Binärbäume mit Höhen *h* und $h-1$ mit jeweils minimaler Knotenanzahl. Es sei K_h die minimale Knotenanzahl eines AVL-Baums bei Höhe *h*. Dann gilt:

$K_0 = 0$, $K_1 = 1$, $K_h = K_{h-1} + K_{h-2} + 1$ für $h \ge 2$. Die erzeugende Funktion der Folge $(K_h)_{h \in \mathbf{N}}$ sei $K(z)$.

1. Schritt: $K_h = K_{h-1} + K_{h-2} + 1 + a_h$ für $h \ge 0$ mit $a_0 = -1$, $a_i = 0$ für $i \ge 1$.

2. Schritt:

$$\begin{aligned} K(z) &= \sum_{h=0}^{\infty} K_h \cdot z^h = \sum_{h=1}^{\infty} K_{h-1} \cdot z^h + \sum_{h=2}^{\infty} K_{h-2} \cdot z^h + \sum_{h=0}^{\infty} z^h - 1 \\ &= \sum_{h=0}^{\infty} K_h \cdot z^{h+1} + \sum_{h=0}^{\infty} K_h \cdot z^{h+2} + \sum_{h=0}^{\infty} z^h - 1 \\ &= z \cdot K(z) + z^2 \cdot K(z) + \frac{1}{1-z} - 1 \quad . \end{aligned}$$

3. Schritt: $K(z) = \frac{z}{(1-z)\cdot(1-z-z^2)} = F(z)\cdot\frac{1}{1-z}$; hierbei ist $F(z)$ die erzeugende Funktion der Folge der Fibonacci-Zahlen. Die Folge $(K_h)_{h\in\mathbf{N}}$ ist also nach Satz 5.11-1 (viii) die Faltung der Folge der Fibonacci-Zahlen $(F_n)_{n\in\mathbf{N}}$ mit der konstanten Folge $(1,1,1,...)$.

4. Schritt: Dieser erübrigt sich, da die Lösung im 3. Schritt bereits ermittelt wurde: $K_h = \sum_{k=0}^{n} F_k \cdot 1 = \sum_{k=0}^{n} F_k = F_{n+2} - 1$. Die letzte Gleichung zeigt man beispielsweise durch vollständige Induktion (Übungsaufgabe 5.30 (b)).

Für einen AVL-Baum mit Höhe *h* und *n* Knoten ist
$n \geq$ minimale Knotenanzahl bei Höhe $h = F_{h+2} - 1$ bzw. $F_{h+2} \leq n+1$. In Kapitel 5.10 wird $F_n = \frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right)$ für jedes $n \in \mathbf{N}$ ermittelt. Da F_n eine natürliche Zahl ist, folgt $F_n = \left\lfloor \frac{1}{\sqrt{5}}\cdot\left(\frac{1+\sqrt{5}}{2}\right)^n + \frac{1}{2} \right\rfloor$. Eingesetzt in die obige Ungleichung ergibt sich
$F_{h+2} = \left\lfloor \frac{1}{\sqrt{5}}\cdot\left(\frac{1+\sqrt{5}}{2}\right)^{h+2} + \frac{1}{2} \right\rfloor \leq n+1$ und $\frac{1}{\sqrt{5}}\cdot\left(\frac{1+\sqrt{5}}{2}\right)^{h+2} + \frac{1}{2} \leq n+2$. Löst man diese Ungleichung nach *h* auf, folgt unter Verwendung von Satz 5.5-5:
$h < 1{,}4404201\cdot\log_2(n+2) - 0{,}327724 < 1{,}4404201\cdot\log_2(n+2)$.

Satz 5.12-2:

(i) Die Anzahl strukturell verschiedener Binärbäume mit *n* Knoten mit $n \geq 0$ beträgt

$$\frac{1}{n+1}\cdot\binom{2\cdot n}{n} = \frac{4^n}{(n+1)\cdot\sqrt{\pi\cdot n}} + C \quad \text{mit einer reellen Konstanten } C > 0.$$

(ii) Die *mittlere* Anzahl von Knoten, die von der Wurzel aus bis zur Erreichung eines beliebigen Knotens eines Binärbaums mit *n* Knoten (gemittelt über alle *n* Knoten) besucht werden, d.h. der *mittlere „Abstand" eines Knotens von der Wurzel* in einem Binärbaum mit *n* Knoten, ist $C'\cdot\sqrt{\pi n} + C''$ mit reellen Konstanten $C' > 0$ und $C'' > 0$. Im günstigsten Fall (wenn also alle Niveaus voll besetzt sind) ist der größte Abstand eines Knotens von der Wurzel in einem Binärbaum mit *n* Knoten gleich $\lfloor \log_2(n) \rfloor + 1 = \lceil \log_2(n+1) \rceil \approx \log_2(n)$, im ungünstigsten Fall ist dieser Abstand gleich *n*.

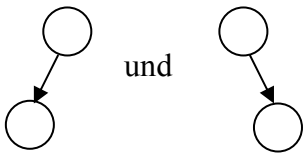
Beide Aussagen mit Hilfe des am Ende von Kapitel 5.11 beschriebenen Vefahrens bewiesen werden.

Es sei b_n für $n \in \mathbf{N}$ die Anzahl strukturell verschiedener Binärbäume mit *n* Knoten. $B(z)$ sei die erzeugende Funktion der Folge $(b_n)_{n\in\mathbf{N}}$. Für kleine Werte von *n* kann man b_n direkt angeben:

$b_0 = 1$: der einzige Binärbaum ohne Knoten ist der leere Baum;

$b_1 = 1$: der einzige Binärbaum mit genau 1 Knoten ist der Baum, der nur aus der Wurzel besteht;

$b_2 = 2$: die beiden Binärbäume mit genau 2 Knoten sind:



Für $n \geq 1$ besteht ein Binärbaum mit *n* Knoten aus der Wurzel und zwei Binärbäumen mit zusammen $n-1$ Knoten, die an den beiden Nachfolgerpositionen der Wurzel beginnen. An der linken Nachfolgerposition befindet sich ein Binärbaum mit *k* vielen Knoten, an der rechten Nachfolgerposition ein Binärbaum mit $n-1-k$ vielen Knoten. Daher gilt

$$b_n = \sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} \text{ für } n \geq 1.$$

1. Schritt: Alle Gleichungen werden in einer einzigen Gleichung zusammengefasst:

$$b_n = \sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} + a_n \text{ für } n \in \mathbf{N}\text{; hierbei ist } a_0 = 1 \text{ und } a_n = 0 \text{ für } n \geq 1.$$

2. Schritt: Beide Seiten werden mit z^n multipliziert und alle Werte aufaddiert. Man erhält:

$$
\begin{aligned}
B(z) &= \sum_{n=0}^{\infty} b_n \cdot z^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} + a_n \right) \cdot z^n \\
&= \sum_{n=1}^{\infty} \left(\sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} \right) \cdot z^n + 1 \\
&= z \cdot \sum_{n=1}^{\infty} \left(\sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} \right) \cdot z^{n-1} + 1 \\
&= z \cdot \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} b_k \cdot b_{n-k} \right) \cdot z^n + 1 \\
&= z \cdot \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} b_k \cdot b_{n-k} \cdot z^k \cdot z^{n-k} \right) + 1 \\
&= z \cdot \left(\sum_{n=0}^{\infty} b_n \cdot z^n \right) \cdot \left(\sum_{n=0}^{\infty} b_n \cdot z^n \right) + 1 \qquad \text{nach Satz 5.1-10} \\
&= z \cdot \left(B(z) \right)^2 + 1 \quad .
\end{aligned}
$$

3. Schritt: Es gilt $\left(B(z)\right)^2 - \frac{1}{z} \cdot B(z) + \frac{1}{z} = 0$ bzw.

$$
\begin{aligned}
B(z) &= \frac{1}{2 \cdot z} \pm \sqrt{\frac{1}{4 \cdot z^2} - \frac{1}{z}} \\
&= \frac{1 \pm \sqrt{1 - 4 \cdot z}}{2 \cdot z} \quad .
\end{aligned}
$$

Falls $B(z) = \frac{1 + \sqrt{1 - 4 \cdot z}}{2 \cdot z}$ gilt, so ergibt sich der Widerspruch $B(0) = b_0 = \infty$, also ist $B(z) = \frac{1 - \sqrt{1 - 4 \cdot z}}{2 \cdot z}$.

4. Schritt: Mit der Regel von de l'Hospital gilt

$$
b_0 = B(0) = \left. \frac{1 - \sqrt{1 - 4 \cdot z}}{2 \cdot z} \right|_{z=0} = \left. \frac{-(1/2) \cdot (-4) \cdot \left(1 / \sqrt{1 - 4 \cdot z}\right)}{2} \right|_{z=0} = 1 \, .
$$

Aus der in Kapitel 5.10 hergeleiten Formel

$$
(1 \pm x)^m = \sum_{i=0}^{\infty} (\pm 1)^i \, \frac{m \cdot (m-1) \cdot \ldots \cdot (m-i+1)}{i!} x^i \, ,
$$

wobei $m = 1/2$ und anstelle von x der Wert $-4 \cdot z$ gesetzt wird, ergibt sich

$$
\begin{aligned}
1 - \sqrt{1 - 4 \cdot z} &= 1 - \sum_{i=0}^{\infty} \frac{(1/2) \cdot (-1/2) \cdot \ldots \cdot (1/2 - i + 1)}{i!} \cdot (-4 \cdot z)^i \\
&= 1 - \left(1 + (-4 \cdot z) \cdot \sum_{i=1}^{\infty} \frac{(1/2) \cdot (-1/2) \cdot \ldots \cdot (1/2 - i + 1)}{i!} \cdot (-4 \cdot z)^{i-1} \right) .
\end{aligned}
$$

Damit ist

$$B(z) = \frac{1-\sqrt{1-4\cdot z}}{2\cdot z}$$
$$= 2\cdot\sum_{i=1}^{\infty}\frac{(1/2)\cdot(-1/2)\cdot\ldots\cdot(1/2-i+1)}{i!}\cdot(-4\cdot z)^{i-1}.$$

Um diese etwas „unangenehm“ aussehende unendliche Reihe zu vereinfachen, wird folgende Nebenrechnung durchgeführt:

$$\frac{(1/2)\cdot(-1/2)\cdot\ldots\cdot(1/2-i+1)}{i!} = \frac{(1/2)\cdot(-1/2)\cdot(-3/2)\ldots\cdot\left(-\frac{2\cdot i-3}{2}\right)}{i!}$$
$$= \frac{\left(\frac{1}{2}\right)^{i}\cdot(-1)^{i-1}\cdot 1\cdot 3\cdot\ldots\cdot(2\cdot i-3)}{i!}\cdot\frac{2\cdot 4\cdot\ldots\cdot(2\cdot i-2)}{2\cdot 4\cdot\ldots\cdot(2\cdot i-2)}$$
$$= \frac{\left(\frac{1}{2}\right)^{i}\cdot(-1)^{i-1}\cdot(2\cdot i-2)!}{i!\cdot 2^{i-1}\cdot(i-1)!}$$
$$= \frac{1}{2}\cdot\left(-\frac{1}{4}\right)^{i-1}\cdot\frac{1}{i}\cdot\binom{2\cdot i-2}{i-1}.$$

Damit ist

$$B(z) = 2\cdot\sum_{i=1}^{\infty}\frac{(1/2)\cdot(-1/2)\cdot\ldots\cdot(1/2-i+1)}{i!}\cdot(-4\cdot z)^{i-1}.$$
$$= \sum_{i=1}^{\infty}\left(-\frac{1}{4}\right)^{i-1}\cdot\frac{1}{i}\cdot\binom{2\cdot i-2}{i-1}\cdot(-4\cdot z)^{i-1}$$
$$= \sum_{i=0}^{\infty}\left(-\frac{1}{4}\right)^{i}\cdot\frac{1}{i+1}\cdot\binom{2\cdot i}{i}\cdot(-4\cdot z)^{i}$$
$$= \sum_{i=0}^{\infty}\frac{1}{i+1}\cdot\binom{2\cdot i}{i}\cdot z^{i}$$
$$= \sum_{n=0}^{\infty}\frac{1}{n+1}\cdot\binom{2\cdot n}{n}\cdot z^{n}.$$

Der Koeffizientenvergleich in $B(z)=\sum_{n=0}^{\infty}b_n\cdot z^n=\sum_{n=0}^{\infty}\frac{1}{n+1}\cdot\binom{2\cdot n}{n}\cdot z^n$ ergibt

$$b_n = \frac{1}{n+1}\cdot\binom{2\cdot n}{n}.$$

Damit ist der erste Teil der Formel in Satz 5.12-2 (i) bewiesen. Für den zweiten Teil wird die Stirling’sche Formel (siehe angegebene Literatur)

$$n!\sim\sqrt{2\cdot\pi\cdot n}\cdot\left(\frac{n}{e}\right)^{n}$$

zusammen mit Satz 4.1-2 eingesetzt. Die Stirling'sche Formel ist eine sehr gute Approximation; der relative Fehler ist etwa $1/(12 \cdot n)$. Damit ergibt sich:

$$\begin{aligned}\frac{1}{n+1} \cdot \binom{2 \cdot n}{n} &\sim \frac{1}{n+1} \cdot \frac{(2 \cdot n)!}{n! \cdot n!} \\ &= \frac{\sqrt{4 \cdot \pi \cdot n} \cdot (2 \cdot n)^{2 \cdot n} \cdot e^{2 \cdot n}}{(n+1) \cdot e^{2 \cdot n} \cdot (2 \cdot \pi \cdot n) \cdot n^{2 \cdot n}} \\ &= \frac{4^n}{(n+1) \cdot \sqrt{\pi \cdot n}}.\end{aligned}$$

Für einen Knoten *v* in einem Baum *G* mit *n* Knoten bezeichne $r(v)$ den Rang des Knotens *v*, d.h. die Anzahl der Kanten, die von der Wurzel aus durchlaufen werden, um *v* zu erreichen. Um *v* von der Wurzel aus zu erreichen, werden $r(v)+1$ Knoten durchlaufen. Es sei $I(G) = \sum_v (r(v)+1)$, d.h. die Summe aller in *G* möglichen Pfadlängen, gemessen in der Anzahl besuchter Knoten.

Der linke Teilbaum unterhalb der Wurzel des Baum *G* sei G_l; der rechte Teilbaum unterhalb der Wurzel sei G_r. G_l oder G_r können auch leer sein. Für einen Knoten *v* in G_l sei $r_l(v)$ der Rang von *v* bezogen auf G_l. Entsprechend bezeichne $r_r(v)$ für einen Knoten *v* in G_r den Rang von *v* bezogen auf G_r. Dann ist für einen Knoten *v* in G_l $r_l(v) = r(v) - 1$. Es gilt daher:

$$\begin{aligned}I(G) &= \sum_{v \text{ in } G} (r(v)+1) \\ &= \sum_{v \text{ in } G_l} (r_l(v)+2) + \sum_{v \text{ in } G_r} (r_r(v)+2) + 1 \qquad \text{(der Rang der Wurzel ist 1)} \\ &= I(G_l) + \sum_{v \text{ in } G_l} 1 + I(G_r) + \sum_{v \text{ in } G_r} 1 + 1 \\ &= I(G_l) + I(G_r) + n \qquad \text{für } n > 0,\end{aligned}$$

$I(G) = 0$ für $n = 0$.

Es sei $I_n = \sum_{\substack{G \text{ ist ein Binärbaum} \\ \text{mit } n \text{ Knoten}}} I(G)$, d.h. die Summe aller möglichen Pfadlängen in Binärbäumen mit *n* Knoten, gemessen in der Anzahl besuchter Knoten.

Mit Hilfe des am Ende von Kapitel 5.11 beschriebenen Vefahrens wird eine geschlossene Formel für I_n hergeleitet. Es bezeichne $I(z)$ die erzeugende Funktion der Folge $(I_n)_{n \in \mathbf{N}}$.

1. Schritt:

$$I_n = \sum_{\substack{G \text{ ist ein Binärbaum} \\ \text{mit } n \text{ Knoten}}} I(G) \qquad \text{für } n \geq 1, \text{ Aufteilung nach linken Teilbäumen:}$$

$$= \sum_{\substack{G \text{ ist ein Binärbaum} \\ \text{mit } n \text{ Knoten}}} \sum_{i=0}^{n-1} \left(\underbrace{I(G_l)}_{\substack{G_l \text{ enthält} \\ i \text{ Knoten}}} + \underbrace{I(G_r)}_{\substack{G_r \text{ enthält} \\ n-i-1 \text{ Knoten}}} + n \right)$$

$$= \sum_{i=0}^{n-1} \left(I_i \cdot b_{n-i-1} + I_{n-i-1} \cdot b_i \right) + n \cdot b_n \quad .$$

Hierbei bezeichnet b_n für $n \in \mathbf{N}$ wieder die Anzahl strukturell verschiedener Binärbäume mit n Knoten.

2. Schritt:

$$I(z) = \sum_{n=1}^{\infty} I_n \cdot z^n$$

$$= \sum_{n=1}^{\infty} \left(\sum_{i=0}^{n-1} \left(I_i \cdot b_{n-i-1} + I_{n-i-1} \cdot b_i \right) + n \cdot b_n \right) \cdot z^n$$

$$= \sum_{n=1}^{\infty} \left(\sum_{i=0}^{n-1} \left(I_i \cdot b_{n-i-1} + I_{n-i-1} \cdot b_i \right) \cdot z^n + n \cdot b_n \cdot z^n \right)$$

$$= \sum_{n=1}^{\infty} \left(\sum_{i=0}^{n-1} \left(I_i \cdot b_{n-i-1} \right) \cdot z^n \right) + \sum_{n=1}^{\infty} \left(\sum_{i=0}^{n-1} \left(I_{n-i-1} \cdot b_i \right) \cdot z^n \right) + \sum_{n=1}^{\infty} n \cdot b_n \cdot z^n$$

Der Ausdruck $f_{n-1} = \sum_{i=0}^{n-1} I_i \cdot b_{n-i-1}$ ist das $(n-1)$-te Folgenglied der Faltung der Folgen $(I_n)_{n \in \mathbf{N}}$ und $(b_n)_{n \in \mathbf{N}}$. Dasselbe gilt für $\sum_{i=0}^{n-1} I_{n-i-1} \cdot b_i = f_{n-1}$. Daher ist

$$I(z) = \sum_{n=1}^{\infty} f_{n-1} \cdot z^n + \sum_{n=1}^{\infty} f_{n-1} \cdot z^n + \sum_{n=1}^{\infty} n \cdot b_n \cdot z^n \qquad \text{mit Satz 5.11-1}$$

$$= \sum_{n=0}^{\infty} f_n \cdot z^{n+1} + \sum_{n=0}^{\infty} f_n \cdot z^{n+1} + \sum_{n=1}^{\infty} n \cdot b_n \cdot z^n$$

$$= 2 \cdot z \cdot I(z) \cdot B(z) + z \cdot B'(z) \quad .$$

3. Schritt: $I(z) = \dfrac{1}{1 - 2 \cdot z \cdot B(z)} \cdot z \cdot B'(z)$.

Mit $B(z) = \dfrac{1 - \sqrt{1 - 4 \cdot z}}{2 \cdot z}$ und $B'(z) = \dfrac{1 - \sqrt{1 - 4 \cdot z} - 2 \cdot z}{2 \cdot z^2 \cdot \sqrt{1 - 4 \cdot z}}$ ist

$$I(z) = \frac{2 \cdot z - \sqrt{1 - 4 \cdot z} + 1 - 4 \cdot z}{2 \cdot z \cdot (1 - 4 \cdot z)}$$

$$= \frac{1}{1 - 4 \cdot z} - \frac{1}{2 \cdot z \cdot \sqrt{1 - 4 \cdot z}} + \frac{1}{2 \cdot z} \quad .$$

4. Schritt:

$$I(z) = \frac{1}{1-4\cdot z} - \frac{1}{2\cdot z\cdot\sqrt{1-4\cdot z}} + \frac{1}{2\cdot z}$$
$$= \sum_{n=0}^{\infty} 4^n \cdot z^n - \frac{1}{2\cdot z}\cdot\left(\frac{1}{\sqrt{1-4\cdot z}} - 1\right).$$

Das Beispiel $\sum_{n=0}^{\infty}\binom{m+n-1}{n}\cdot z^n = \frac{1}{(1-z)^m}$ aus Kapitel 5.10 ist auf Werte $m \in \mathbf{R}$ verallgemeinerbar. Damit ist

$$\frac{1}{\sqrt{1-4\cdot z}} = \sum_{n=0}^{\infty}\binom{n-1/2}{n}\cdot(4\cdot z)^n = \sum_{n=1}^{\infty}\binom{n-1/2}{n}\cdot(4\cdot z)^n + 1 \text{ und}$$

$$\frac{1}{2\cdot z}\cdot\left(\frac{1}{\sqrt{1-4\cdot z}} - 1\right) = \sum_{n=1}^{\infty}\binom{n-1/2}{n}\cdot 2\cdot(4\cdot z)^{n-1}$$
$$= 2\cdot\sum_{n=1}^{\infty}\binom{2\cdot n}{n}\cdot\frac{4^{n-1}}{4^n}\cdot z^{n-1}.$$

Hierbei wurde die Identität $\binom{n-1/2}{n} = \binom{2\cdot n}{n} \Big/ 2^{2\cdot n}$ verwendet. Mit $\binom{2\cdot n}{n} = \frac{(2\cdot n-1)\cdot 2}{n}\cdot\binom{2\cdot(n-1)}{n-1}$ folgt weiter:

$$2\cdot\sum_{n=1}^{\infty}\binom{2\cdot n}{n}\cdot\frac{4^{n-1}}{4^n}\cdot z^{n-1} = \frac{1}{2}\cdot\sum_{n=1}^{\infty}\binom{2\cdot n}{n}\cdot z^{n-1}$$
$$= \sum_{n=1}^{\infty}\binom{2\cdot(n-1)}{n-1}\cdot\frac{2\cdot n-1}{n}\cdot z^{n-1}$$
$$= \sum_{n=0}^{\infty}\binom{2\cdot n}{n}\cdot\frac{2\cdot n+1}{n+1}\cdot z^n$$
$$= \sum_{n=0}^{\infty}(2\cdot n+1)\cdot b_n\cdot z^n.$$

Insgesamt ist damit

$$I(z) = \sum_{n=0}^{\infty}\left(4^n - (2\cdot n+1)\cdot b_n\right)\cdot z^n \text{ und } I_n = 4^n - (2\cdot n+1)\cdot b_n.$$

Der Ausdruck $I_n \Big/ n\cdot b_n$ ist die mittlere Anzahl an Knoten, die in Binärbäumen mit n Knoten von der Wurzel aus zu einem Knoten durchlaufen werden. Mit dem gerade hergeleiteten Ergebnis folgt $I_n \Big/ n\cdot b_n = \frac{4^n}{n\cdot b_n} - \frac{2\cdot n+1}{n}$. Mit der Stirlingschen Formel ist

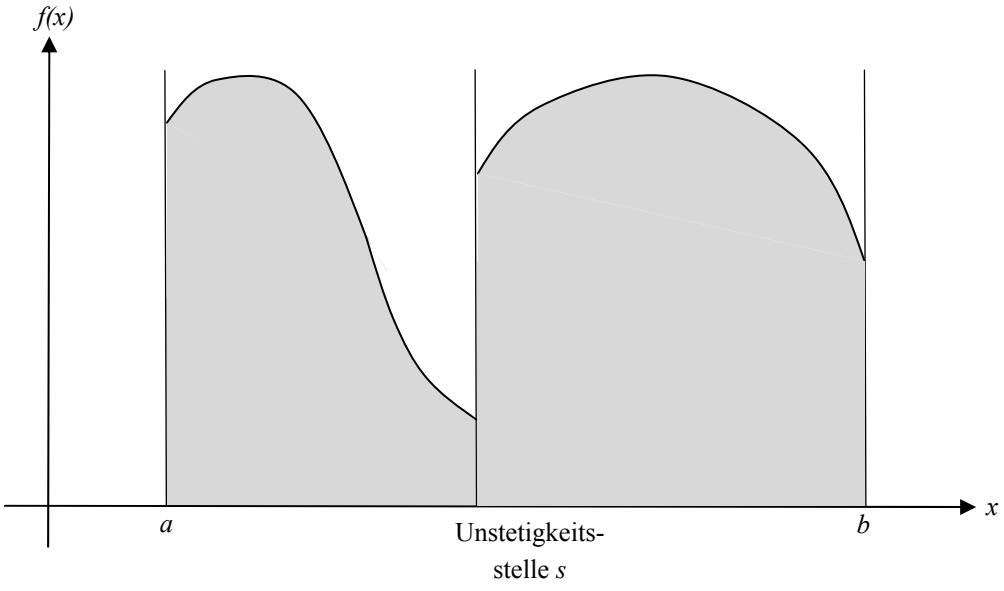
$$b_n = \frac{1}{n+1}\cdot\binom{2\cdot n}{n} \sim \frac{4^n}{(n+1)\cdot\sqrt{\pi\cdot n}} \text{ und}$$

$I_n / n \cdot b_n \sim \sqrt{\pi \cdot n} \cdot \frac{n+1}{n} - 2 + \frac{1}{n} \le C' \cdot \sqrt{\pi \cdot n} + C''$ mit reellen Konstanten $C' > 0$ und $C'' > 0$.

5.13 Einführung in die Integralrechnung

Die grundlegende Aufgabe der Integralrechnung ist die Berechnung von Flächeninhalten von Gebieten im zweidimensionalen Raum bzw. von Volumina in höherdimensionalen Bereichen, die durch Graphen entsprechender Funktionen begrenzt werden. Weitere Anwendungen finden sich in der Bestimmung von Längen von Kurven, in der Statistik bei der Bestimmung von Momenten (Erwartungswert, Varianz usw.) bei stetigen Verteilungen und in zahlreichen anderen Gebieten der Mathematik. Ein interessanter Zusammenhang wird durch den Hauptsatz der Differential- und Integralrechnung formuliert: Differenzieren und Integrieren sind zueinander inverse Operationen.

Im folgenden soll der Flächeninhalt zwischen dem Graphen einer Funktion $f : I \to \mathbf{R}$ in einem Intervall $I \subseteq \mathbf{R}$, der *x*-Achse und einer linken und einer rechten jeweils senkrecht auf der *x*-Achse an den Intervallgrenzen stehenden Begrenzungslinie bestimmt werden. Dabei kann angenommen werden, dass das Intervall die Form $I = [a, b]$ mit $a \in \mathbf{R}$ und $b \in \mathbf{R}$ hat und dass *f* auf *I* zumindest stückweise stetig ist. Das bedeutet, dass es in *I* höchstens endliche viele Unstetigkeitsstellen von *f* gibt und dass für jede Unstetigkeitsstelle $s \in I$ die Grenzwerte $\lim\limits_{\substack{x \to s \\ x < s}} f(x)$ und $\lim\limits_{\substack{x \to s \\ x > s}} f(x)$ existieren (siehe nachfolgende Abbildung). Die letzte Bedingung bedeutet, dass *f* an einer Unstetigkeitsstelle nicht etwa gegen Unendlich strebt.



Die Lösung dieser Aufgabe führt auf einen speziellen Integralbegriff, das **Riemann'sche Integral**. Es handelt sich dabei um einen Spezialfall verschiedener Integralbegriffe, die im Falle stückweise stetiger Funktionen jedoch zu einem übereinstimmenden Ergebnis kommen.

Die folgende Betrachtung kann für jedes Teilintervall von *I*, in dem *I* keine Unstetigkeitsstellen besitzt, getrennt erfolgen, so dass gleich angenommen werden kann, dass *f* im gesamten Intervall *I* stetig ist.

Das Intervall *I* wird durch $k+1$ Punkte $x_0, ..., x_k$ in *k* Teilintervalle zerlegt; diese **Zerlegung** werde mit *Z* bezeichnet:

$a = x_0 < x_1 < x_2 < ... x_{k-1} < x_k = b$.

Die Länge des *i*-ten Teilintervalls für $i = 1, ..., k$ ist $\Delta x_i = x_i - x_{i-1}$.

Unter der **Feinheit der Zerlegung** *Z* versteht man den Wert $\Delta Z = \max\left\{ \Delta x_i \mid i = 1, ... k \right\}$.

Für $i = 1, ..., k$ sei $a_i \in [x_{i-1}, x_i]$. Dann heißt

$$\sum_{i=1}^{k} f(a_i) \cdot \Delta x_i$$

eine **Riemann'sche Summe** von *f* zu *Z*.

Eine Zerlegung Z_2 ist eine **Verfeinerung** der Zerlegung Z_1, wenn jeder Teilungspunkt von Z_1 auch Teilungspunkte von Z_2 ist. Das bedeutet, dass die Teilungspunkte von Z_2 die Teilintervalle von Z_1 in der Regel noch einmal unterteilen.

Es sei $(Z_n)_{n \in \mathbf{N}}$ eine Folge von Zerlegungen von $[a, b]$ mit der Eigenschaft, dass Z_{n+1} eine Verfeinerung von Z_n ist und dass $\lim_{n \to \infty} \Delta Z_n = 0$ gilt. S_n sei eine Riemann'sche Summe zu Z_n. Die Anzahl der Teilintervalle der Zerlegung Z_n sei $k(n)$. Wegen der Stetigkeit besitzt *f* im Intervall $[x_{i-1}, x_i]$ jeweils ein Minimum m_i und ein Maximum M_i. Mit

$$U_n = \sum_{i=1}^{k(n)} m_i \cdot \Delta x_i \text{ und } O_n = \sum_{i=1}^{k(n)} M_i \cdot \Delta x_i$$

gilt

$U_n \le S_n \le O_n$.

Anschaulich stellt U_n für eine Funktion $f:[a,b]\to\mathbf{R}$ mit $f(x)\geq 0$ eine untere Abschätzung und O_n eine obere Abschätzung für den gesuchten Flächeninhalt dar.

Die Zerlegung Z_n laute $a=x_0<x_1<x_2<\ldots x_{k(n)-1}<x_{k(n)}=b$, die Zerlegung Z_{n+1} sei $a=y_0<y_1<y_2<\ldots y_{k(n+1)-1}<y_{k(n+1)}=b$. Das Teilintervall $[x_{i-1},x_i]$ werde durch Z_{n+1} zerlegt in $x_{i-1}=y_l<y_{l+1}<\ldots y_{l+t}=x_i$. Für $j=1,\ldots,t$ sei m'_{l+j} das Minimum von f im Intervall $[y_{l+j-1},y_{l+j}]$; entsprechend sei M'_{l+j} das Maximum von f im Intervall $[y_{l+j-1},y_{l+j}]$. Dann ist

$$m_i\cdot\Delta x_i=m_i\cdot\sum_{j=1}^{t}(y_{l+j}-y_{l+j-1})\leq\sum_{j=1}^{t}m'_{l+j}(y_{l+j}-y_{l+j-1}),$$

$$M_i\cdot\Delta x_i=M_i\cdot\sum_{j=1}^{t}(y_{l+j}-y_{l+j-1})\geq\sum_{j=1}^{t}M'_{l+j}(y_{l+j}-y_{l+j-1}),$$

und damit

$U_n\leq U_{n+1}$ und $O_n\geq O_{n+1}$,

$U_n\leq U_{n+1}\leq O_{n+1}\leq O_n\leq O_1$ und $U_1\leq U_{n+1}\leq O_{n+1}\leq O_n$.

Das bedeutet, dass die Folge $(U_n)_{n\in\mathbf{N}}$ monoton steigt und nach oben beschränkt ist, während die Folge $(O_n)_{n\in\mathbf{N}}$ monoton fällt und nach unten beschränkt ist. Nach Satz 5.1-3 konvergieren diese Folgen daher, d.h. es existieren $U=\lim\limits_{n\to\infty}U_n$ und $O=\lim\limits_{n\to\infty}O_n$, und es gilt $U\leq O$. Mit weitergehenden Überlegungen zu stetigen Funktionen lässt sich sogar $U=O$ zeigen. Da für die Folge $(S_n)_{n\in\mathbf{N}}$ der Riemann'schen Summen $U_n\leq S_n\leq O_n$ gilt, konvergiert nach Satz 5.1-3 diese Folge gegen denselben Grenzwert. Weitere Überlegungen zeigen, dass dieser Grenzwert unabhängig von der speziellen Zerlegung des Intervalls $[a,b]$ ist, solange die oben genannten Bedingungen eingehalten werden.

Zusammenfassend gilt, ohne hier im einzelnen auf weitere technische Beweisschritte einzugehen:

Satz 5.13-1:

Die Funktion $f:[a,b]\to \mathbf{R}$ sei stückweise stetig. Es sei $(Z_n)_{n\in\mathbf{N}}$ eine Folge von Zerlegungen von $[a,b]$ mit der Eigenschaft $\lim\limits_{n\to\infty}\Delta Z_n = 0$ und S_n eine Riemann'sche Summe zu Z_n. Dann existiert der Grenzwert $\lim\limits_{n\to\infty} S_n$ und ist von der speziellen Folge von Zerlegungen unabhängig.

Dieser Grenzwert heißt das **bestimmte Integral** von *f* über $[a,b]$ und wird mit

$$\int_a^b f(x)dx$$

bezeichnet. Die Werte *a* und *b* heißen **Integrationsgrenzen**, die Funktion *f* heißt **Integrand** und *x* **Integrationsvariable**. Hierbei ist die Benennung der Integrationsvariablen beliebig, etwa

$$\int_a^b f(x)dx = \int_a^b f(t)dt\ .$$

Die Funktion *f* heißt über $I=[a,b]$ **integrierbar**, wenn $\int\limits_a^b f(x)dx$ existiert.

Man definiert für $a<b$ die Integrale $\int\limits_b^a f(x)dx = -\int\limits_a^b f(x)dx$ und $\int\limits_a^a f(x)dx = 0$.

Satz 5.13-2:

Es seien f und g über einem Intervall I integrierbare Funktionen. Es gelte $a \in I$, $b \in I$, $c \in I$ und $\lambda \in \mathbf{R}$. Dann gilt:

(i) $\int_a^b \lambda dx = \lambda \cdot (b-a)$.

(ii) $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$.

(iii) $\int_a^b \big(f(x)+g(x)\big)dx = \int_a^b f(x)dx + \int_a^b g(x)dx$.

(iv) $\int_a^b \lambda \cdot f(x)dx = \lambda \cdot \int_a^b f(x)dx$.

(v) Ist $f(x) \le g(x)$ für alle $x \in [a,b]$, so gilt $\int_a^b f(x)dx \le \int_a^b g(x)dx$.

Die Aussagen ergeben sich aus der Definition des Integralbegriffs und der Rechenregeln für Grenzwerte.

Der gesuchte **Flächeninhalt** zwischen dem Graphen einer Funktion $f: I \to \mathbf{R}$ in einem Intervall $I = [a,b]$, der x-Achse und einer linken und einer rechten jeweils senkrecht auf der x-Achse an den Intervallgrenzen stehenden Begrenzungslinie wird durch $\left|\int_a^b f(x)dx\right|$ bestimmt.

Es muss der *Betrag* des bestimmten Integrals genommen werden, da Flächen, die unterhalb der x-Achse liegen, wegen $f(x) \le 0$ negativ gerechnet werden.

Zur konkreten Berechnung von Integralen werden noch einige mathematische Sätze benötigt.

Satz 5.13-3: (Mittelwertsatz der Integralrechnung)

Es sei f auf $[a,b]$ stetig. Dann gibt es ein Element $y \in [a,b]$ mit

$$\int_a^b f(x)dx = (b-a)\cdot f(y).$$

Die Aussage dieses Satzes folgt aus dem Zwischenwertsatz (Satz 5.2-3):

Da f auf $[a,b]$ stetig ist, besitzt es dort ein Minimum $m = f(x_1)$ und ein Maximum $M = f(x_2)$. Es sei $\lambda = \frac{1}{b-a}\cdot\int_a^b f(x)dx$. Für alle $x \in [a,b]$ gilt $m \le f(x) \le M$ und damit

$\int_a^b m dx \le \int_a^b f(x)dx \le \int_a^b M dx$ bzw. mit Satz 5.13-2

$m\cdot(b-a) = \int_a^b m dx \le \int_a^b f(x)dx = \lambda\cdot(b-a) \le \int_a^b M dx = M\cdot(b-a)$, also $m \le \lambda \le M$. Nach Satz 5.2-3 (ii) gibt es ein y zwischen x_1 und x_2, also $y \in [a,b]$, mit $f(y) = \lambda$, d.h. $\int_a^b f(x)dx = (b-a)\cdot f(y)$.

Es sei I ein Intervall. Eine differenzierbare Funktion $F: I \to \mathbf{R}$ heißt **Stammfunktion** von $f: I \to \mathbf{R}$, wenn $F'(x) = f(x)$ für alle $x \in I$ gilt.

Satz 5.13-4:

Es seien F und G Stammfunktionen von f. Dann ist $F(x) - G(x) = C$ mit einer Konstanten $C \in \mathbf{R}$ eine Stammfunktion von f.

Alle Stammfunktionen von f unterscheiden sich nur um eine Konstante.

Die Aussage folgt aus der Tatsache, dass $(F(x)-G(x))' = F'(x) - G'(x) = f(x) - f(x) = 0$ für alle $x \in I$ gilt und dann durch Anwendung von Satz 5.6-7 (ii).

Der folgende Satz stellt einen Zusammenhang zwischen den Operationen Differenzieren und Integrieren her und liefert das Werkzeug zur Berechnung von Integralen.

Satz 5.13-5: Hauptsatz der Differential- und Integralrechnung

Es sei *I* ein Intervall, $a \in I$ und $f : I \to \mathbf{R}$ eine stetige Funktion. Dann ist die durch

$$F_a(x) = \int_a^x f(t)dt$$

definierte Funktion $F_a : I \to \mathbf{R}$ eine Stammfunktion von *f*, d.h.

$$F_a'(x) = \frac{d}{dx}\int_a^x f(t)dt = f(x).$$

Für jede Stammfunktion *F* von *f* gilt $F(x) = F_a(x) + C$ für ein $C \in \mathbf{R}$ und

$$\int_a^b f(x)dx = F(b) - F(a).$$

Zum Beweis der Aussage ist zunächst zu zeigen, dass $F_a'(x) = \lim\limits_{\Delta x \to 0} \dfrac{F_a(x+\Delta x) - F_a(x)}{\Delta x} = f(x)$ gilt. Dazu wird zunächst der Zähler berechnet:

$$F_a(x+\Delta x) - F_a(x) = \int_a^{x+\Delta x} f(t)dt - \int_a^x f(t)dt = \int_x^{x+\Delta x} f(t)dt = f(y)\cdot\Delta x$$

mit einem Wert $y \in [x, x+\Delta x]$ (siehe Satz 5.13-3). Damit gilt

$$F_a'(x) = \lim_{\Delta x \to 0} \frac{F_a(x+\Delta x) - F_a(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(y)\cdot\Delta x}{\Delta x} = \lim_{\Delta x \to 0} f(y).$$

Mit $\Delta x \to 0$ geht $y \to x$, also $F_a'(x) = \lim\limits_{\Delta x \to 0} f(y) = f(x)$.

Es sei *F* eine beliebige Stammfunktion von *f*. Nach Satz 5.13-4 ist dann $F(x) = F_a(x) + C$ für ein $C \in \mathbf{R}$. Dann ist $F(b) = F_a(b) + C$, und wegen $F_a(a) = \int_a^a f(t)dt = 0$ ist

$F(a) = F_a(a) + C = C$. Insgesamt ergibt sich $\int_a^b f(t)dt = F_a(b) = F(b) - C = F(b) - F(a)$.

Um ein bestimmtes Integral $\int_a^b f(x)dx$ zu berechnen, geht man also folgendermaßen vor:

1. Man sucht eine Stammfunktion F von f, also eine Funktion mit $F'(x) = f(x)$.

2. Man bildet $F(b) - F(a)$.

Satz 5.13-6:

Es seien f und g über einem Intervall I integrierbare Funktionen. Es gelte $a \in I$ und $b \in I$. Dann gilt:

(i) $\int_a^b \big(f(x) \cdot g'(x)\big)dx = \big(f(b) \cdot g(b) - f(a) \cdot g(a)\big) - \int_a^b \big(f'(x) \cdot g(x)\big)dx$

(**Partielle Integration**)

(ii) $\int_a^b \big(f\big(g(x)\big) \cdot g'(x)\big)dx = \int_{g(a)}^{g(b)} f(x)dx$

(Substitutionsregel)

Es gilt $(f(x) \cdot g(x))' = f'(x) \cdot g(x) + f(x) \cdot g'(x)$. Also ist $(f \cdot g)$ eine Stammfunktion von $f' \cdot g + f \cdot g'$. Nach Satz 5.13-5 ist $\int_a^b \big(f'(x) \cdot g(x) + f(x) \cdot g'(x)\big)dx = \big(f(b) \cdot g(b) - f(a) \cdot g(a)\big)$.

Nach Satz 5.13-2 (iii) ist $\int_a^b \big(f'(x) \cdot g(x) + f(x) \cdot g'(x)\big)dx = \int_a^b \big(f'(x) \cdot g(x)\big)dx + \int_a^b \big(f(x) \cdot g'(x)\big)dx$.

Damit folgt die Aussage von (i).

Es sei F eine Stammfunktion von f. Es gilt gemäß Satz 5.6-2 (iv) $\big(F(g(x))\big)' = F'(g(x)) \cdot g'(x) = f(g(x)) \cdot g'(x)$, also ist $F \circ g$ eine Stammfunktion von $\big(f \circ g\big) \cdot g'$. Damit ist mit Satz 5.13-5 $\int_a^b \big(f\big(g(x)\big) \cdot g'(x)\big)dx = F\big(g(b)\big) - F\big(g(a)\big)$.

Nach Satz 5.13-5 ist andererseits $\int_{g(a)}^{g(b)} f(x)dx = F\big(g(b)\big) - F\big(g(a)\big)$. Damit folgt die Aussage in (ii).

Zur Vereinfachung wird folgende Schreibweise eingeführt:

Für eine Funktion $f: X \to \mathbf{R}$ und $a \in \mathbf{R}$ und $b \in \mathbf{R}$ ist bedeutet $f(x)|_{x=a}^{x=b}$ die Differenz $f(b) - f(a)$. Mit dieser Schreibweise lautet beispielsweise die Aussage aus Satz 5.13-6 (i):
$\int_a^b \big(f(x) \cdot g'(x)\big)dx = \big(f \cdot g\big)(x)|_{x=a}^{x=b} - \int_a^b \big(f'(x) \cdot g(x)\big)dx$.

Ist F eine Stammfunktion von f, dann ist $\int_a^b f(x)dx = F(x)|_{x=a}^{x=b}$.

Im folgenden werden eine Reihe von Beispielen behandelt, die auf der Anwendung der vorherigen Sätze und Definitionen beruhen.

Beispiele:

1. Für reelle Zahlen n mit $n \neq -1$ ist wegen $\frac{d}{dx}\frac{x^{n+1}}{n+1} = x^n$:
$$\int_a^b x^n dx = \frac{x^{n+1}}{n+1}\Big|_{x=a}^{x=b}.$$

2. Es ist für $x > 0$ $\frac{d}{dx}\ln(x) = \frac{1}{x}$. Für $x < 0$ ist $\frac{d}{dx}\ln(|x|) = \frac{d}{dx}\ln(-x) = -\frac{1}{-x} = \frac{1}{x}$. Daher gilt bei $0 \notin [a, b]$
$$\int_a^b \frac{1}{x}dx = \ln(|x|)\Big|_{x=a}^{x=b}.$$

3. Es sei $\lambda \in \mathbf{R}$ mit $\lambda \neq 0$. Dann ist $\frac{d}{dx}\frac{e^{\lambda \cdot x}}{\lambda} = e^{\lambda \cdot x}$ und damit
$$\int_a^b e^{\lambda \cdot x}dx = \left(\frac{1}{\lambda} \cdot e^x\right)\Big|_{x=a}^{x=b}.$$

4. Für reelle Zahlen *n* mit $n \neq -1$ ist wegen $\frac{d}{dx}\frac{(\ln(x))^{n+1}}{n+1} = \frac{(\ln(x))^n}{x}$:

$$\int_a^b \frac{(\ln(x))^n}{x}\,dx = \frac{(\ln(x))^{n+1}}{n+1}\Bigg|_{x=a}^{x=b}.$$

5. Setzt man in der Regel $\int_a^b (f(x)\cdot g'(x))dx = f(x)|_{x=a}^{x=b} - \int_a^b (f'(x)\cdot g(x))dx$ (partielle Integration) $f(x) = \ln(x)$ und $g(x) = x$, so erhält man

$$\int_a^b \ln(x)dx = (\ln(x)\cdot x)\Big|_{x=a}^{x=b} - \int_a^b \left(\frac{x}{x}\right)dx = (\ln(x)\cdot x - x)\Big|_{x=a}^{x=b} = (x\cdot(\ln(x)-1))\Big|_{x=a}^{x=b}.$$

6. Mit Hilfe der partiellen Integration (dort $f(x) = x^n$ und $g(x) = e^x$) erhält man

$$\int_a^b (x^n\cdot e^x)dx = (x^n\cdot e^x)\Big|_{x=a}^{x=b} - n\cdot\int_a^b (x^{n-1}\cdot e^x)dx.$$

Mit dieser Rekursionsformel berechnet man beispielsweise

$$\begin{aligned}\int_a^b (x^2\cdot e^x)dx &= (x^2\cdot e^x)\Big|_{x=a}^{x=b} - 2\cdot\int_a^b (x\cdot e^x)dx\\ &= (x^2\cdot e^x)\Big|_{x=a}^{x=b} - 2\cdot\left((x\cdot e^x)\Big|_{x=a}^{x=b} - 1\cdot\int_a^b (x^0\cdot e^x)dx\right)\\ &= (x^2\cdot e^x)\Big|_{x=a}^{x=b} - 2\cdot((x-1)\cdot e^x)\Big|_{x=a}^{x=b}\\ &= ((x^2-2\cdot x-1)\cdot e^x)\Big|_{x=a}^{x=b}.\end{aligned}$$

7. Es seien $\lambda \in \mathbf{R}$ mit $\lambda \neq 0$ und $\mu \in \mathbf{R}$. Mit Hilfe der Substitutionsregel aus Satz 5.13-6 lässt sich $\int_a^b \left(x\cdot\sqrt{\lambda\cdot x^2+\mu}\right)dx$ bestimmen: Setzt man $f(x) = \sqrt{x}$ und $g(x) = \lambda\cdot x^2 + \mu$, dann ist $f(g(x))\cdot g'(x) = 2\cdot\lambda\cdot x\cdot\sqrt{\lambda\cdot x^2+\mu}$, also

$$\int_a^b \left(x\cdot\sqrt{\lambda\cdot x^2+\mu}\right)dx = \frac{1}{2\cdot\lambda}\cdot\int_{g(a)}^{g(b)} x^{1/2}dx = \frac{1}{2\cdot\lambda}\cdot\left(x\cdot\sqrt{x}\right)\Big|_{x=\lambda\cdot a^2+\mu}^{x=\lambda\cdot b^2+\mu}.$$

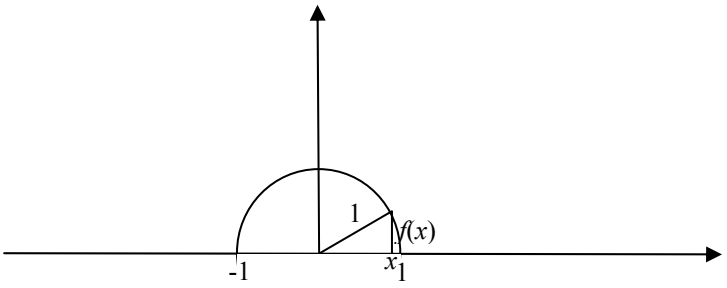
8. Es sei $h: X \to \mathbf{R}$ eine stetige Funktion. Dann lassen sich mit Hilfe der Substitutionsregel aus Satz 5.13-6 Integrale der Form $\int_a^b \frac{h'(x)}{h(x)}dx$ berechnen. Setzt man nämlich $f(x)=\frac{1}{x}$ und $g(x)=h(x)$, dann ist $\left(f\left(g(x)\right)\cdot g'(x)\right)=\frac{h'(x)}{h(x)}$, also

$$\int_a^b \frac{h'(x)}{h(x)}dx = \int_{h(a)}^{h(b)} \frac{1}{x}dx = \ln(|x|)\Big|_{x=h(a)}^{x=h(b)} = \ln(|h(x)|)\Big|_{x=a}^{x=b} = \ln\left(\left|\frac{h(b)}{h(a)}\right|\right).$$

Beispielsweise ist

$$\int_a^b \frac{6\cdot x}{3\cdot x^2+7}dx = \ln\left(\frac{3\cdot b^2+7}{3\cdot a^2+7}\right).$$

9. Das Beispiel greift die am Ende von Kapitel 9 angeführten und aus der Schulmathematik bekannten Sinus- und Kosinusfunktion auf:
Es gilt $\sin'(x)=\cos(x)$, $\cos'(x)=-\sin(x)$ für $x\in\mathbf{R}$ und
$\sin(i\cdot\pi)=0$ und $\cos((2\cdot i+1)\cdot\pi/2)=0$ für $i\in\mathbf{Z}$, $\sin(-\pi/2)=-1$, $\sin(\pi/2)=1$.
Zeichnet man in einem *x*-*y*-Koordinatensystem einen Halbkreis mit Mittelpunkt $(0,0)$ und Radius 1, so werden die Punkte des Halbkreises für $-1\le x\le 1$ durch Koordinaten (x,y) beschrieben, für die $x^2+y^2=1$ gilt. Die *y*-Koordinate ist also eine Funktion $f(x)$ mit $f(x)=\sqrt{1-x^2}$. Die Fläche des Halbkreises beträgt daher $\int_{-1}^{1}\sqrt{1-x^2}dx$.



Wählt man in der Substitutionsregel aus Satz 5.13-6 $\int_a^b \left(f\left(g(x)\right)\cdot g'(x)\right)dx = \int_{g(a)}^{g(b)} f(x)dx$ eine umkehrbare Funktion *g*, so kann man diese Regel auch als $\int_{g^{-1}(\alpha)}^{g^{-1}(\beta)} \left(f\left(g(x)\right)\cdot g'(x)\right)dx = \int_\alpha^\beta f(x)dx$ lesen. Die Sinusfunktion ist für $x\in[-\pi/2,\pi/2]$ umkehrbar. Damit ist mit $g(x)=\sin(x)$ und $\cos(x)=\sqrt{1-(\sin(x))^2}$

$$\int_{-1}^{1}\sqrt{1-x^2}\,dx=\int_{-1}^{1}f(x)dx=\int_{\sin(-1)}^{\sin(1)}\left(\sqrt{1-\sin^2(x)}\cdot\cos(x)\right)dx=\int_{-\pi/2}^{\pi/2}(\cos(x))^2\,dx\,.$$

Dieses Integral lässt sich mit partieller Integration lösen:

$$\begin{aligned}\int_{-\pi/2}^{\pi/2}(\cos(x))^2\,dx&=\sin(x)\cdot\cos(x)\big|_{x=-\pi/2}^{x=\pi/2}-\int_{-\pi/2}^{\pi/2}-(\sin(x))^2\,dx\\&=\sin(x)\cdot\cos(x)\big|_{x=-\pi/2}^{x=\pi/2}+\int_{-\pi/2}^{\pi/2}\left(1-(\cos(x))^2\right)dx\\&=\sin(x)\cdot\cos(x)\big|_{x=-\pi/2}^{x=\pi/2}+\int_{-\pi/2}^{\pi/2}1dx-\int_{-\pi/2}^{\pi/2}(\cos(x))^2\,dx\;.\end{aligned}$$

Insgesamt ist

$$\begin{aligned}\int_{-\pi/2}^{\pi/2}(\cos(x))^2\,dx&=\frac{1}{2}\cdot\left(\sin(x)\cdot\cos(x)+x\right)\big|_{x=-\pi/2}^{x=\pi/2}\\&=\frac{1}{2}\cdot\left(\sin(\pi/2)\cdot\cos(\pi/2)+\pi/2\right)-\frac{1}{2}\cdot\left(\sin(-\pi/2)\cdot\cos(-\pi/2)-\pi/2\right)\\&=\pi/2\quad.\end{aligned}$$

Die Zahl π kann also als Fläche eines Kreises mit Radius 1 definiert werden.

In Kapitel 5.3 werden Polynome $p(x)$ behandelt. Diese sind stetig und integrierbar. Beispiel 1 und Satz 5.13-2 liefern das Ergebnis für $\int_a^b p(x)dx$.

In den folgenden Ausführungen wird eine Methode zur Integration gebrochen rationaler Funktionen beschrieben.

Eine gebrochen rationale Funktion (siehe Kapitel 5.4) der Form $f(x)=p(x)/q(x)$ mit Polynomen *p* und *q* ist an allen Stellen bis auf die Nullstellen von *q* definiert und stetig, also integrierbar. Ist der Grad von *q* nicht größer als der Grad von *p*, kann man *f* in eindeutiger Weise in der Form $f(x)=s(x)+\frac{r(x)}{q(x)}$ mit Polynomen *s* und *r* schreiben, wobei *r* einen kleineren Grad als *q* besitzt. Dann ist $\int_a^b f(x)dx=\int_a^b\left(s(x)+\frac{r(x)}{q(x)}\right)dx=\int_a^b s(x)dx+\int_a^b\frac{r(x)}{q(x)}dx$. Die Berechnung von $\int_a^b s(x)dx$ ist oben beschrieben. Sind $x_{01},\ldots,x_{0l}$ die reellen Nullstellen von *q*, wobei mehrfache Nullstellen jeweils auch mehrfach aufgeführt sind, kann man *q* in der Form $q(x)=(x-x_{01})\cdot\ldots\cdot(x-x_{0l})\cdot q_g(x)$ mit einem Polynom $q_g(x)$ von geradem Grad, das keine reellen Nullstellen hat, schreiben. Mit dem algebraisch etwas aufwendigen Verfahren der Par-

tialbruchzerlegung lässt sich $r(x)/q(x)$ in eine Summe von Brüchen zerlegen, die die Form $\frac{A}{(x-B)^{k+1}}$ oder $\frac{A\cdot x+B}{(x^2+2\cdot C\cdot x+D)^{k+1}}$ mit $k\in\mathbf{N}$ und $C^2<D$ haben. Jeden dieser Summanden kann man getrennt integrieren. Es ist

$$\int_a^b \frac{A}{(x-B)^{k+1}} = \begin{cases} A\cdot\ln(|x-B|)\Big|_{x=a}^{x=b} & \text{für } k=0 \\ -A\cdot\dfrac{1}{k\cdot(x-B)^k}\Big|_{x=a}^{x=b} & \text{für } k>0 \end{cases}.$$

Den Summanden $\frac{A\cdot x+B}{(x^2+2\cdot C\cdot x+D)^{k+1}}$ kann man noch einmal zerlegen:

$$\frac{A\cdot x+B}{(x^2+2\cdot C\cdot x+D)^{k+1}} = \frac{A}{2}\cdot\frac{2\cdot(x+C)}{(x^2+2\cdot C\cdot x+D)^{k+1}} + \frac{B-A\cdot C}{(x^2+2\cdot C\cdot x+D)^{k+1}}.$$

Mit der Technik aus Beispiel 8 ist für $k=0$ mit $h(x)=x^2+2\cdot C\cdot x+D$ und $h'(x)=2\cdot(x+C)$

$$\int_a^b \frac{2\cdot(x+C)}{x^2+2\cdot C\cdot x+D}dx = \ln(|x^2+2\cdot C\cdot x+D|)\Big|_{x=a}^{x=b}.$$

Für $k\geq 1$ ist mit $h(x)=x^{-(k+1)}$ und $g(x)=x^2+2\cdot C\cdot x+D$, $g'(x)=2\cdot(x+C)$:

$$\int_a^b \frac{2\cdot(x+C)}{(x^2+2\cdot C\cdot x+D)^{k+1}}dx = \int_a^b h(g(x))\cdot g'(x)dx = \int_{g(a)}^{g(b)} h(x)dx = -\frac{1}{k\cdot(x^2+2\cdot C\cdot x+D)^k}\Big|_{x=a}^{x=b}.$$

Daher bleibt nur $\int_a^b \frac{1}{(x^2+2\cdot C\cdot x+D)^{k+1}}dx$ zu behandeln. Dazu wird $t(x)=\frac{x+C}{\sqrt{D-C^2}}$ und $h(x)=\frac{1}{(x^2+1)^{k+1}}$ gesetzt.

Dann ist $t'(x)=(D-C^2)^{-1/2}$ und $(t(x))^2+1=\frac{(x+C)^2+D-C^2}{D-C^2}=\frac{x^2+2\cdot C\cdot x+D}{D-C^2}$, also

$$\begin{aligned}\int_a^b \frac{1}{(x^2+2\cdot C\cdot x+D)^{k+1}}dx &= (D-C^2)^{k+1}\cdot\int_a^b \frac{1}{((t(x))^2+1)^{k+1}}dx \\ &= (D-C^2)^{k+1+1/2}\cdot\int_a^b h(t(x))\cdot t'(x)dx \\ &= (D-C^2)^{k+1+1/2}\cdot\int_{h(a)}^{h(b)} h(x)dx\end{aligned}.$$

Es bleibt die Bestimmung des Integrals $\int_{h(a)}^{h(b)} h(x)dx == \int_{\alpha}^{\beta} \frac{1}{\left(x^2+1\right)^{k+1}} dx$ mit $\alpha = h(a)$ und $\beta = h(b)$. Durch Anwendung der partiellen Integration ist

$$\int_{\alpha}^{\beta} \frac{1}{\left(x^2+1\right)^{k}} dx = \left(\frac{1}{\left(x^2+1\right)^{k}} \cdot x \right)\Bigg|_{x=\alpha}^{x=\beta} - \int_{\alpha}^{\beta} \left(\frac{d}{dx}\left(\frac{1}{\left(x^2+1\right)^{k}} \right) \cdot x \right) dx$$

$$= \left(\frac{1}{\left(x^2+1\right)^{k}} \cdot x \right)\Bigg|_{x=\alpha}^{x=\beta} + 2 \cdot k \cdot \int_{\alpha}^{\beta} \frac{x^2}{\left(x^2+1\right)^{k+1}} dx \ .$$

Wegen $\frac{x^2}{\left(x^2+1\right)^{k+1}} = \frac{1}{\left(x^2+1\right)^{k}} - \frac{1}{\left(x^2+1\right)^{k+1}}$ ergibt sich

$$2 \cdot k \cdot \int_{\alpha}^{\beta} \frac{1}{\left(x^2+1\right)^{k+1}} dx = \left(\frac{1}{\left(x^2+1\right)^{k}} \cdot x \right)\Bigg|_{x=\alpha}^{x=\beta} + (2 \cdot k - 1) \cdot \int_{\alpha}^{\beta} \frac{1}{\left(x^2+1\right)^{k}} dx$$

und damit eine Rekursionsformel für $\int_{\alpha}^{\beta} \frac{1}{\left(x^2+1\right)^{k+1}} dx$. Diese bleibt für $k = 1$ zu bestimmen. Aus der Behandlung der trigonometrischen Funktionen, die hier ausgeklammert wurde, kennt man deren Ableitungen. Für die Umkehrfunktion *arctg* der Tangensfunktion, der Arcustangensfunktion, gilt $\frac{d}{dy} arctg(y) = \frac{1}{1+y^2}$. Daher ist

$$\int_{\alpha}^{\beta} \frac{1}{x^2+1} dx = arctg(x)\Big|_{x=\alpha}^{x=\beta} .$$

Bisher wurden Integrale der Form $\int_a^b f(x)dx$ behandelt, für die f in $[a,b]$ wenigstens stückweise stetig ist. Es soll nun der Fall behandelt werden, dass f an einer oder beiden Intervallgrenzen einen Pol besitzt oder dass die Intervallgrenzen gegen ∞ oder $-\infty$ laufen. Dieser Ansatz führt auf den Begriff des uneigentlichen Integrals:

Falls der Grenzwert $\int_a^{\infty} f(x)dx = \lim_{t\to\infty} \int_a^t f(x)dx$ existiert, wird er als **uneigentliches Integral** bezeichnet. Entsprechend wird das uneigentliche Integral $\int_{-\infty}^{\infty} f(x)dx = \lim_{t\to\infty} \int_{-t}^{t} f(x)dx$ definiert.

Besitzt *f* bei *a* einen Pol, dann ist das uneigentliche Integral $\int_a^b f(x)dx = \lim\limits_{\substack{\varepsilon\to 0\\ \varepsilon>0}} \int_{a+\varepsilon}^b f(x)dx$. Für einen Pol bei *b* wird das uneigentliche Integral $\int_a^b f(x)dx = \lim\limits_{\substack{\varepsilon\to 0\\ \varepsilon>0}} \int_a^{b-\varepsilon} f(x)dx$ definiert.

Beispiele:

1. $\int_1^\infty \frac{1}{x^2}dx = \lim\limits_{t\to\infty}\int_1^t \frac{1}{x^2}dx = \lim\limits_{t\to\infty}\left(-\frac{1}{x}\right)\Big|_{x=1}^{x=t} = \lim\limits_{t\to\infty}\left(-\frac{1}{t}+1\right) = 1$.

2. Der Grenzwert $\int_1^\infty \frac{1}{x}dx$ existiert nicht; denn $\lim\limits_{t\to\infty}\int_1^t \frac{1}{x}dx = \lim\limits_{t\to\infty}\big(\ln(x)\big)\big|_{x=1}^{x=t} = \lim\limits_{t\to\infty}\big(\ln(t)-\ln(1)\big)$.

3. In der Wahrscheinlichkeitsrechnung wird die Exponentialverteilung behandelt. Eine exponentialverteilte Zufallsvariable *X* hat den Erwartungswert $\int_0^\infty a\cdot x\cdot e^{-a\cdot x}dx$ mit einem Parameter $a>0$: Mit partieller Integration (hier $f(x)=x$, $g'(x)=e^{-a\cdot x}$, also $g(x)=-\frac{1}{a}\cdot e^{-a\cdot x}$) ist

$$\int_0^t a\cdot x\cdot e^{-a\cdot x}dx = a\cdot\left(\left(-x\cdot\frac{1}{a}\cdot e^{-a\cdot x}\right)\Big|_{x=0}^{x=t} - \int_0^t\left(-\frac{1}{a}\cdot e^{-a\cdot x}\right)dx\right)$$

$$= \left(-x\cdot e^{-a\cdot x}\right)\Big|_{x=0}^{x=t} + \left(-\frac{1}{a}\cdot e^{-a\cdot x}\right)\Big|_{x=0}^{x=t}$$

$$= -t\cdot e^{-a\cdot t} + \left(-\frac{1}{a}\cdot e^{-a\cdot t}+\frac{1}{a}\right) .$$

Damit folgt $\int_0^\infty a\cdot x\cdot e^{-a\cdot x}dx = \lim\limits_{t\to\infty}\int_0^t a\cdot x\cdot e^{-a\cdot x}dx = \frac{1}{a}$.

4. Die durch $f(x)=\frac{1}{\sqrt{|x|}}$ definierte Funktion hat bei $x=0$ einen Pol. Damit ist $\int_{-1}^1 \frac{1}{\sqrt{|x|}}dx$ die Summe zweier uneigentlicher Integrale: $\int_{-1}^1 \frac{1}{\sqrt{|x|}}dx = \int_{-1}^0 \frac{1}{\sqrt{|x|}}dx + \int_0^1 \frac{1}{\sqrt{|x|}}dx$. Eine

Stammfunktion von $g(x)=\frac{1}{\sqrt{x}}$ ist $G(x)=2\cdot\sqrt{x}$; eine Stammfunktion von $h(x)=\frac{1}{\sqrt{-x}}$ ist $H(x)=-2\cdot\sqrt{-x}$. Damit ist

$$\int_0^1 \frac{1}{\sqrt{|x|}}dx=\int_0^1 \frac{1}{\sqrt{x}}dx=\lim_{\substack{\varepsilon\to 0\\ \varepsilon>0}}\int_\varepsilon^1 \frac{1}{\sqrt{x}}dx=\lim_{\substack{\varepsilon\to 0\\ \varepsilon>0}}\left(\left(2\cdot\sqrt{x}\right)\Big|_{x=\varepsilon}^{x=1}\right)=2,$$

$$\int_{-1}^0 \frac{1}{\sqrt{|x|}}dx=\lim_{\substack{\varepsilon\to 0\\ \varepsilon>0}}\int_{-1}^{-\varepsilon} \frac{1}{\sqrt{-x}}dx=\lim_{\substack{\varepsilon\to 0\\ \varepsilon>0}}\left(\left(-2\cdot\sqrt{-x}\right)\Big|_{x=-1}^{x=-\varepsilon}\right)=2 \text{ und}$$

$$\int_{-1}^1 \frac{1}{\sqrt{|x|}}dx=4.$$

6 Das Lösen linearer Gleichungssysteme

Das vorliegende Kapitel wählt aus einem Gebiet der Mathematik, der Linearen Algebra, ein spezielles Thema aus, nämlich die Behandlung eines effizienten Verfahrens zur Lösung linearer Gleichungssysteme, wie sie in vielen Anwendungen der Mathematik vorkommen. Damit verbunden ist das Invertieren von Matrizen (die Begriffe werden im Laufe des Kapitels erläutert).

Die Lineare Algebra als zugrundeliegende Theorie hat sich zu einem der wichtigsten Teilgebiete der Mathematik entwickelt. Eine auch nur einführende Darstellung dieser Theorie würde jedoch den Rahmen dieses Textes sprengen. Daher wird in diesem Kapitel auf die Darstellung der Beweise weitgehend verzichtet.

6.1 Matrizen und Vektoren

Ein rechteckiges Zahlenschema aus reellen Zahlen mit *m* Zeilen und *n* Spalten

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,j} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,j} & \dots & a_{2,n} \\ . & . & \dots & . & \dots & . \\ a_{i,1} & a_{i,2} & \dots & a_{i,j} & \dots & a_{i,n} \\ . & . & \dots & . & \dots & . \\ a_{m,1} & a_{m,2} & \dots & a_{m,j} & \dots & a_{m,n} \end{bmatrix} = \mathbf{A}_{(m,n)} = \left[a_{i,j}\right]_{i=1,\dots,m;\, j=1,\dots,n}$$

heißt **reellwertige Matrix vom Typ (*m, n*)**. Im Schnittpunkt der **Zeile** *i* und der **Spalte** *j* steht das **Matrixelement** $a_{i,j} \in \mathbf{R}$. Der erste Index gibt die Zeilennummer, der zweite Index die Spaltennummer an. Im folgenden werden Matrizen durch fett gedruckte Buchstaben bezeichnet.

Zwei Matrizen $\mathbf{A}_{(m,n)} = \left[a_{i,j}\right]$ und $\mathbf{B}_{(r,s)} = \left[b_{l,k}\right]$ sind gleich, wenn sie vom selben Typ sind, d.h. $m = r$ und $n = s$, und sie elementweise gleich sind, d.h. wenn $a_{i,j} = b_{i,j}$ für $i = 1, \dots, m$ und $j = 1, \dots, n$ gilt.

Eine Matrix vom Typ (*n*, *n*) heißt **quadratische Matrix**.

Eine Matrix, deren sämtliche Elemente 0 sind, heißt **Nullmatrix**; sie wird mit **0** bezeichnet.

Die quadratische Matrix

$$\mathbf{I}_{(n,n)} = \begin{bmatrix} 1 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 1 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ . & . & ... & . & ... & . & & & & \\ 0 & 0 & 0 & ... & 1 & 0 & 0 & ... & 0 & 0 \\ . & . & ... & . & ... & . & & & & \\ 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 1 \end{bmatrix}$$

vom Typ (n, n), die in der Diagonalen die Zahlen 1 und sonst nur Nullen enthält, heißt **Einheitsmatrix vom Typ (n, n)**. Es ist

$$\mathbf{I}_{(n,n)} = \left\lfloor \delta_{i,j} \right\rfloor \text{ mit } \delta_{i,j} = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases}.$$

Eine Matrix vom Typ (1, n) heißt **Zeilenvektor** der Länge n. Eine Matrix vom Typ (m, 1) heißt **Spaltenvektor** der Länge m. In beiden Fällen verzichtet man meist auf die doppelte Indizierung:

Ein Zeilenvektor wird geschrieben als

$$\vec{a} = \left\lfloor a_1 \quad a_2 \quad ... \quad a_j \quad ... \quad a_n \right\rfloor.$$

Ein Spaltenvektor wird geschrieben als

$$\vec{b} = \begin{bmatrix} b_1 \\ . \\ . \\ . \\ b_i \\ . \\ . \\ . \\ b_m \end{bmatrix}.$$

Es sollen nun **Rechenoperationen auf Matrizen** definiert werden:

Es seien **A** und **B** zwei Matrizen *vom (gleichen) Typ* (*m, n*).

Die **Summe** von **A** und **B** ist die Matrix $\mathbf{C} = [c_{i,j}] = \mathbf{A} + \mathbf{B}$ mit $c_{i,j} = a_{i,j} + b_{i,j}$.

Die **Differenz** von **A** und **B** ist die Matrix $\mathbf{C} = [c_{i,j}] = \mathbf{A} - \mathbf{B}$ mit $c_{i,j} = a_{i,j} - b_{i,j}$.

Die Summe (Differenz) zweier Matrizen vom Typ (*m*, *n*) ist wieder vom Typ (*m*, *n*). Man erhält sie also, indem man die Elemente an den sich entsprechenden Positionen addiert (subtrahiert).

Es sei $k \in \mathbf{R}$. Das **Skalarprodukt** der Matrix **A** mit (dem Skalar) *k* ist die Matrix $\mathbf{D} = [d_{i,j}] = k \cdot \mathbf{A} = \mathbf{A} \cdot k$ mit $d_{i,j} = k \cdot a_{i,j} = a_{i,j} \cdot k$.

Das Skalarprodukt einer Matrix **A** vom Typ (*m*, *n*) mit einer reellen Zahl ist wieder eine Matrix vom Typ (*m*, *n*). Bei der Bildung des Skalarprodukts einer Matrix mit einer Zahl werden also alle Matrixelemente mit dieser Zahl multipliziert.

Es seien **A**, **B** und **C** Matrizen gleichen Typs, $k \in \mathbf{R}$ und $h \in \mathbf{R}$. Dann gelten folgende Regeln:

$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$,
$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$,
$\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$,
Mit $-\mathbf{A} = (-1) \cdot \mathbf{A}$ ist $\mathbf{A} - \mathbf{A} = \mathbf{0}$,

$k \cdot (\mathbf{A} + \mathbf{B}) = k \cdot \mathbf{A} + k \cdot \mathbf{B}$,
$(k + h) \cdot \mathbf{A} = k \cdot \mathbf{A} + h \cdot \mathbf{A}$,
$(k \cdot h) \cdot \mathbf{A} = k \cdot (h \cdot \mathbf{A})$,
$1 \cdot \mathbf{A} = \mathbf{A}$.

Die Menge der Matrizen vom (gleichen) Typ (*m*, *n*) mit der definierten Addition von Matrizen und der Multiplikation von reellen Zahlen mit Matrizen bildet einen **Vektorraum** über **R**.

Erläuterung:

Eine algebraische Struktur $(V, \oplus, K, \cdot)$ heißt **Vektorraum über *K***, wenn gilt:

(i) $(V, \oplus)$ ist eine kommutative Gruppe

(ii) K ist ein Körper (**Skalarkörper**)

(iii) die Abbildung $\cdot : \begin{cases} K \times V & \to & V \\ (k, \mathbf{v}) & \to & k \cdot \mathbf{v} \end{cases}$ genügt den folgenden Regeln:

für jedes $k \in K$, für jedes $l \in K$, für jedes $\mathbf{v} \in V$ und jedes $\mathbf{w} \in V$ gilt

$k \cdot (l \cdot \mathbf{v}) = (k \cdot l) \cdot \mathbf{v}$,

$1 \cdot \mathbf{v} = \mathbf{v}$,

$k \cdot (\mathbf{v} \oplus \mathbf{w}) = (k \cdot \mathbf{v}) \oplus (k \cdot \mathbf{w})$,

$(k + l) \cdot \mathbf{v} = (k \cdot \mathbf{v}) \oplus (l \cdot \mathbf{v})$.

Das **Produkt** der beiden Matrizen $\mathbf{A}_{(m,n)}$ und $\mathbf{B}_{(n,k)}$ ist nur dann definiert, wenn der erste Faktor $\mathbf{A}_{(m,n)}$ genauso viele Spalten wie der zweite Faktor $\mathbf{B}_{(n,k)}$ Zeilen hat. Das Produkt ist eine Matrix $\mathbf{C}_{(m,k)} = [c_{r,s}] = \mathbf{A} \cdot \mathbf{B}$ vom Typ (m, k) (mit der Zeilenanzahl von **A** und der Spaltenanzahl von **B**) mit

$$c_{r,s} = \sum_{i=1}^{n} a_{r,i} \cdot b_{i,s} \quad \text{für } r = 1, ..., m,\ s = 1, ..., n.$$

Durch Nachrechnen verifiziert man die Gültigkeit von

$\mathbf{A} \cdot (\mathbf{B} \pm \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} \pm \mathbf{A} \cdot \mathbf{C}$,

$(\mathbf{A} \pm \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot \mathbf{C} \pm \mathbf{B} \cdot \mathbf{C}$.

Im allgemeinen ist (selbst für quadratische Matrizen) $\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}$.

Eine Matrix

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & ... & a_{1,j} & ... & a_{1,n} \\ a_{2,1} & a_{2,2} & ... & a_{2,j} & ... & a_{2,n} \\ . & . & ... & . & ... & . \\ a_{i,1} & a_{i,2} & ... & a_{i,j} & ... & a_{i,n} \\ . & . & ... & . & ... & . \\ a_{m,1} & a_{m,2} & ... & a_{m,j} & ... & a_{m,n} \end{bmatrix} = \mathbf{A}_{(m,n)} = \left[a_{i,j}\right]_{i=1,...,m;\, j=1,...,n}$$

kann man als Menge $\{\vec{a}_1, \vec{a}_2, ..., \vec{a}_i, ..., \vec{a}_m\}$ ihrer Zeilenvektoren mit

$$\vec{a}_i = \begin{bmatrix} a_{i,1} & a_{i,2} & ... & a_{i,j} & ... & a_{i,n} \end{bmatrix} \text{ für } i = 1, ..., m$$

bzw. als Menge $\{\vec{b}_1, \vec{b}_2, ..., \vec{b}_j, ..., \vec{b}_n\}$ ihrer Spaltenvektoren mit

$$\vec{b}_j = \begin{bmatrix} a_{1,j} \\ a_{2,j} \\ . \\ . \\ a_{i,j} \\ . \\ . \\ . \\ a_{m,j} \end{bmatrix} \text{ für } j = 1, ..., n$$

auffassen.

Eine Menge $\{\vec{a}_1, ..., \vec{a}_r\}$ von Vektoren heißt **linear unabhängig**, wenn gilt:

Aus der Gleichung

$k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$ mit $k_i \in \mathbf{R}$ für $i = 1, ..., r$ folgt $k_1 = ... = k_r = 0$.

Andernfalls heißt $\{\vec{a}_1, ..., \vec{a}_r\}$ **linear abhängig**.

Um zu überprüfen, ob eine Menge von Vektoren linear unabhängig ist, stellt man also die „Vektorgleichung“ $k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$ auf, wobei die reellen Zahlen $k_1, ..., k_r$ zunächst

„Unbekannte“ sind, und zeigt dann, dass diese Gleichung nur gültig sein kann, wenn alle Unbekannten $k_1, ..., k_r$ gleich 0 sind. Kann man andererseits die Gleichung $k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$ aufstellen, wobei mindestens eine der Zahlen $k_1, ..., k_r$ von 0 verschieden ist, so sind die Vektoren linear abhängig.

Sind die Vektoren $\vec{a}_1, ..., \vec{a}_r$ jeweils Spaltenvektoren mit *m* Komponenten, so ist die Vektorgleichung $k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$ ein **Gleichungssystem** mit *m* Zeilen.

Ein Vektor $\vec{a}$ ist eine **Linearkombination** der Vektoren $\vec{a}_1, ..., \vec{a}_n$, wenn es Zahlen $k_1 \in \mathbf{R}$, ..., $k_n \in \mathbf{R}$ gibt mit

$\vec{a} = k_1 \cdot \vec{a}_1 + ... + k_n \cdot \vec{a}_n$.

In diesem Fall gilt die Vektorgleichung

$k_1 \cdot \vec{a}_1 + ... + k_n \cdot \vec{a}_n - 1 \cdot \vec{a} = \mathbf{0}$, d.h.

die Menge $\{\vec{a}_1, ..., \vec{a}_r, \vec{a}\}$ ist nicht linear unabhängig. Ist umgekehrt die Menge $\{\vec{a}_1, ..., \vec{a}_r\}$ linear abhängig, so sind in der Vektorgleichung

$k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$

nicht alle Skalare gleich 0, etwa $k_j \neq 0$. Es ist dann

$k_j \cdot \vec{a}_j = -k_1 \cdot \vec{a}_1 - ... - k_{j-1} \cdot \vec{a}_{j-1} - k_{j+1} \cdot \vec{a}_{j+1} - ... - k_r \cdot \vec{a}_r$, also

$$\vec{a}_j = \left(-\frac{k_1}{k_j}\right) \cdot \vec{a}_1 + ... + \left(-\frac{k_{j-1}}{k_j}\right) \cdot \vec{a}_{j-1} + \left(-\frac{k_{j+1}}{k_j}\right) \cdot \vec{a}_{j+1} + ... + \left(-\frac{k_r}{k_j}\right) \cdot \vec{a}_r .$$

Insgesamt ergibt sich damit

Satz 6.1-1:

Es sei $n \geq 2$.

Die Vektoren $\vec{a}_1, ..., \vec{a}_n$ sind genau dann linear abhängig, wenn sich wenigstens ein Vektor dieser Menge als Linearkombination der anderen Vektoren dieser Menge darstellen lässt.

Unter dem **Zeilenrang $r_Z(\mathbf{A})$ einer Matrix $\mathbf{A} = \mathbf{A}_{(m,n)}$** versteht man die Maximalzahl linear unabhängiger Zeilen (-vektoren). Unter dem **Spaltenrang $r_S(\mathbf{A})$ einer Matrix $\mathbf{A} = \mathbf{A}_{(m,n)}$** versteht man die Maximalzahl linear unabhängiger Spalten (-vektoren).

Offensichtlich gilt $r_Z(\mathbf{A}) \leq m$ und $r_S(\mathbf{A}) \leq n$.

Der Beweis des folgenden Satzes erfordert eine Reihe weiterführender Überlegungen und einen ziemlich trickreichen Umgang mit den beteiligten Indizes.

Satz 6.1-2:

Für jede Matrix **A** gilt:

$r_Z(\mathbf{A}) = r_S(\mathbf{A})$.

Wegen Satz 6.1-2 kann man **Rang $r(\mathbf{A})$ einer Matrix A** durch $r(\mathbf{A}) = r_Z(\mathbf{A}) = r_S(\mathbf{A})$ definieren. Ist $\mathbf{A} = \mathbf{A}_{(m,n)}$, d.h. **A** besitzt *m* Zeilen und *n* Spalten, dann ist $r(\mathbf{A}) \leq \min\{n, m\}$.

Der folgende Satz beschreibt Operationen, die auf die Zeilen bzw. Spalten einer Matrix anwendbar sind, ohne ihren Rang zu ändern. Diese Operationen sind Grundlage des Verfahrens zur Lösung linearer Gleichungssysteme

Satz 6.1-3:

Gegeben sei die Matrix

$$\mathbf{A} = \mathbf{A}_{(m,n)} = \begin{bmatrix} \vec{a}_1 \\ . \\ . \\ . \\ \vec{a}_m \end{bmatrix} = \begin{bmatrix} \vec{b}_1 & \dots & \vec{b}_n \end{bmatrix}.$$

(Die Vektoren $\vec{a}_1, ..., \vec{a}_m$ sind die Zeilen der Matrix **A**, die Vektoren $\vec{b}_1, ..., \vec{b}_n$ sind die Spalten von **A**.)

Dann gilt:

Zeilenrang und Spaltenrang von **A** ändern sich nicht, wenn man die Matrix **A** einer der folgenden **elementaren Umformungen** unterwirft:

(z1) Zwei Zeilen von **A** werden vertauscht.

(z2) Eine Zeile $\vec{a}_i$ von **A** wird ersetzt durch $\vec{a}_i + k \cdot \vec{a}_j$, wobei $k \in \mathbf{R}$, $1 \le i \le m$, $1 \le j \le m$ und $i \ne j$ gilt.
(Auf $\vec{a}_i$ wird ein Vielfaches einer anderen Zeile addiert.)

(z3) Eine Zeile $\vec{a}_i$ von **A** wird ersetzt durch $k \cdot \vec{a}_i$, wobei $k \in \mathbf{R} \setminus \{0\}$ und $1 \le i \le m$ gilt.
($\vec{a}_i$ wird um ein Vielfaches verändert.)

(s1) Zwei Spalten von **A** werden vertauscht.

(s2) Eine Spalte $\vec{b}_i$ von **A** wird ersetzt durch $\vec{b}_i + k \cdot \vec{b}_j$, wobei $k \in \mathbf{R}$, $1 \le i \le n$, $1 \le j \le n$ und $i \ne j$ gilt.
(Auf $\vec{b}_i$ wird ein Vielfaches einer anderen Spalte addiert.)

(s3) Eine Spalte $\vec{b}_i$ von **A** wird ersetzt durch $k \cdot \vec{b}_i$, wobei $k \in \mathbf{R} \setminus \{0\}$ und $1 \le i \le n$ gilt.
($\vec{b}_i$ wird um ein Vielfaches verändert.)

6.2 Lineare Gleichungssysteme

Eine Menge von *m* Gleichungen in *n* Variablen der Form

$$
\begin{array}{ccccccccccccc}
a_{1,1}\cdot x_1 & + & a_{1,2}\cdot x_2 & + & \ldots & + & a_{1,j}\cdot x_j & + & \ldots & + & a_{1,n}\cdot x_n & = & b_1 \\
a_{2,1}\cdot x_1 & + & a_{2,2}\cdot x_2 & + & \ldots & + & a_{2,j}\cdot x_j & + & \ldots & + & a_{2,n}\cdot x_n & = & b_2 \\
 & & \vdots & & & & & & & & & & \\
a_{i,1}\cdot x_1 & + & a_{i,2}\cdot x_2 & + & \ldots & + & a_{i,j}\cdot x_j & + & \ldots & + & a_{i,n}\cdot x_n & = & b_i \\
 & & \vdots & & & & & & & & & & \\
a_{m,1}\cdot x_1 & + & a_{m,2}\cdot x_2 & + & \ldots & + & a_{m,j}\cdot x_j & + & \ldots & + & a_{m,n}\cdot x_n & = & b_m
\end{array}
$$

heißt **lineares Gleichungssystem (in den Variablen** $x_1, \ldots, x_n$**)**. Abgekürzt lässt es sich schreiben als

$$\mathbf{A}_{(m,n)} \cdot \vec{x}_{(n,1)} = \vec{b}_{(m,1)}.$$

Die Matrix $\mathbf{A} = \mathbf{A}_{(m,n)} = \begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,j} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,j} & \ldots & a_{2,n} \\ \vdots & & & & & \\ a_{i,1} & a_{i,2} & \ldots & a_{i,j} & \ldots & a_{i,n} \\ \vdots & & & & & \\ a_{m,1} & a_{m,2} & \ldots & a_{m,j} & \ldots & a_{m,n} \end{bmatrix}$ heißt **Koeffizientenmatrix**.

Die Elemente der Koeffizientenmatrix und des Vektors $\vec{b} = \vec{b}_{(m,1)}$ sind vorgegebene reelle Zahlen.

Jeder Vektor $\vec{y} = \vec{y}_{(n,1)} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$ mit $\mathbf{A}_{(m,n)} \cdot \vec{y}_{(n,1)} = \vec{b}_{(m,1)}$ heißt **Lösung des linearen Gleichungssystems**.

Ein lineares Gleichungssystem $\mathbf{A}_{(m,n)} \cdot \vec{x}_{(n,1)} = \vec{b}_{(m,1)}$ heißt **homogen**, wenn $b_1 = b_2 = ... = b_m = 0$ ist. Andernfalls heißt es **inhomogen**.

Im linearen Gleichungssystem $\mathbf{A}_{(m,n)} \cdot \vec{x}_{(n,1)} = \vec{b}_{(m,1)}$ heißt die Matrix

$$\left[\mathbf{A}_{(m,n)} \middle| \vec{b}_{(m,1)}\right] = \left[\mathbf{A} \middle| \vec{b}\right] = \left[\begin{array}{ccccccc} a_{1,1} & a_{1,2} & ... & a_{1,j} & ... & a_{1,n} & \mid b_1 \\ a_{2,1} & a_{2,2} & ... & a_{2,j} & ... & a_{2,n} & \mid b_2 \\ . & & & & & & . \; . \\ . & & & & & & . \; . \\ . & & & & & & . \; . \\ a_{i,1} & a_{i,2} & ... & a_{i,j} & ... & a_{i,n} & \mid b_i \\ . & & & & & & . \; . \\ . & & & & & & . \; . \\ . & & & & & & . \; . \\ a_{m,1} & a_{m,2} & ... & a_{m,j} & ... & a_{m,n} & \mid b_m \end{array}\right]$$

die **erweiterte Koeffizientenmatrix**.

Es stellt sich die Frage nach der **Lösbarkeit eines linearen Gleichungssystems** (existiert überhaupt eine Lösung? Ist die Lösung eindeutig bestimmt?)

Gegeben sei das lineare Gleichungssystem $\mathbf{A}_{(m,n)} \cdot \vec{x}_{(n,1)} = \vec{b}_{(m,1)}$. Gesucht wird eine Lösung

$$\vec{y} = \vec{y}_{(n,1)} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}.$$

Im folgenden wird die Typangabe zur Vereinfachung der Schreibweise weggelassen, so dass das Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ lautet.

Der folgende Satz gibt an, in welchen Fällen ein lineares Gleichungssystem lösbar ist und wieviele Lösungen in diesem Fall existieren.

Satz 6.2-1:
Das lineare Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ ist genau dann lösbar, wenn der Rang der Koeffizientenmatrix gleich dem Rang der erweiterten Koeffizientenmatrix ist, d.h. wenn $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right])$ gilt. Ist $r(\mathbf{A}) < r(\left[\mathbf{A} \middle| \vec{b}\right])$, so heißt das Gleichungssystem **inkonsistent** (und ist nicht lösbar).

Für das lineare Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ mit einer *m*-zeiligen und *n*-spaltigen Koeffizientenmatrix $\mathbf{A} = \mathbf{A}_{(m,n)}$ gelte $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right]) = r$, so dass das **Gleichungssystem lösbar** ist. Es ist $r \le m$ und $r \le n$.

Ist $r < m$ (= Anzahl der Zeilen bzw. Gleichungen), so ist das Gleichungssystem lösbar, aber $m - r$ Gleichungen sind „überflüssig“, genauer: **redundant**, da sie Linearkombinationen der übrigen Gleichungen sind.

Die Anzahl der Lösungen des Gleichungssystems hängt davon ab, wie sich der Rang *r* zu der Anzahl *n* der Variablen verhält:

Ist $r < n$ (= Anzahl der Spalten bzw. Variablen), so sind $n - r$ Spaltenvektoren Linearkombinationen der anderen Spaltenvektoren. Das System ist lösbar, jedoch mit $n - r$ **freien Variablen**, denen beliebige reelle Werte zugeordnet werden können. Es gibt also **unendlich viele Lösungen**. Die Werte, die den übrigen *r* Variablen zugeordnet werden, hängen von den zugeordneten Werten der freien Variablen ab.

Ist $r = n$ (= Anzahl der Spalten bzw. Variablen), so ist $n \le m$ ($m - n$ Gleichungen sind redundant). Das System ist **eindeutig lösbar**, d.h. es gibt genau eine Lösung.

Ist die Koeffizientenmatrix **A** eines linearen Gleichungssystems quadratisch, d.h. $n = m$, d.h. es gibt soviele Gleichungen wie Variablen, dann gilt:
Für $r(\mathbf{A}) < r(\left[\mathbf{A} \middle| \vec{b}\right]) \le n$ gibt es keine Lösung;
für $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right]) = n$ gibt es genau eine Lösung;
für $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right]) < n$ gibt es unendlich viele Lösungen.

In einem homogenen linearen Gleichungssystem ist immer $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right])$. Es gibt dann wenigstens eine Lösung (nämlich die **triviale Lösung** $y_1 = y_2 = ... = y_n = 0$). Ist zudem $r(\mathbf{A}) = n$, dann gibt es nur diese Lösung. Ist $r(\mathbf{A}) = r < n$, dann gibt es weitere Lösungen mit $n - r$ freien Variablen. Ist $m < n$, d.h. es gibt weniger Gleichungen als Unbekannte, dann ist auch $r(\mathbf{A}) < n$. Für $m = n$ gibt es nur dann mehr als die triviale Lösung, wenn $r(\mathbf{A}) < n$ ist.

Im folgenden wird eine **Methode zur Lösung eines linearen Gleichungssystems (Gaußscher Algorithmus)** und damit einhergehend eine **Methode zur Bestimmung des Rangs einer Matrix** vorgestellt.

Gegeben sei das lineare Gleichungssystem

$$
\begin{array}{lllllllllllll}
a_{1,1}\cdot x_1 & + & a_{1,2}\cdot x_2 & + & \ldots & + & a_{1,j}\cdot x_j & + & \ldots & + & a_{1,n}\cdot x_n & = & b_1 \\
a_{2,1}\cdot x_1 & + & a_{2,2}\cdot x_2 & + & \ldots & + & a_{2,j}\cdot x_j & + & \ldots & + & a_{2,n}\cdot x_n & = & b_2 \\
\vdots & & & & & & & & & & & & \\
a_{i,1}\cdot x_1 & + & a_{i,2}\cdot x_2 & + & \ldots & + & a_{i,j}\cdot x_j & + & \ldots & + & a_{i,n}\cdot x_n & = & b_i \\
\vdots & & & & & & & & & & & & \\
a_{m,1}\cdot x_1 & + & a_{m,2}\cdot x_2 & + & \ldots & + & a_{m,j}\cdot x_j & + & \ldots & + & a_{m,n}\cdot x_n & = & b_m
\end{array}
$$

bzw.

$$\mathbf{A}\cdot\vec{x}=\vec{b}\,.$$

Hierbei sei mindestens einer der Werte $a_{i,1}$ in der ersten Spalte von 0 verschieden; denn sonst käme x_1 im Gleichungssystem gar nicht vor. Die erweiterte Koeffizientenmatrix sei wieder

$$
\left[\mathbf{A}_{(m,n)}\middle|\,\vec{b}_{(m,1)}\right]=\left[\mathbf{A}\middle|\,\vec{b}\right]=\left[\begin{array}{ccccccc|c}
a_{1,1} & a_{1,2} & \ldots & a_{1,j} & \ldots & a_{1,n} & & b_1 \\
a_{2,1} & a_{2,2} & \ldots & a_{2,j} & \ldots & a_{2,n} & & b_2 \\
\vdots & & & & & & & \vdots \\
a_{i,1} & a_{i,2} & \ldots & a_{i,j} & \ldots & a_{i,n} & & b_i \\
\vdots & & & & & & & \vdots \\
a_{m,1} & a_{m,2} & \ldots & a_{m,j} & \ldots & a_{m,n} & & b_m
\end{array}\right]=\left[\begin{array}{c|c}
\vec{a}_1 & b_1 \\
\vec{a}_2 & b_2 \\
\vdots & \vdots \\
\vec{a}_i & b_i \\
\vdots & \vdots \\
\vec{a}_m & b_m
\end{array}\right]\,.
$$

Sie wird durch elementare Umformungen in eine „Treppenmatrix“ (siehe unten) umgewandelt, aus der man dann die Lösung des Gleichungssystems ablesen kann. Bei diesem Umformungsvorgang wird schrittweise eine Folge von Matrizen $\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(r)}$ erzeugt, die alle den-

selben Rang wie $\left[\mathbf{A}\middle|\vec{b}\right]$ haben. Hierbei ist $\mathbf{A}^{(i)}$ das Ergebnis der Umformung von $\mathbf{A}^{(i-1)}$ nach dem *i*-ten Schritt ($i = 1, ..., r$).

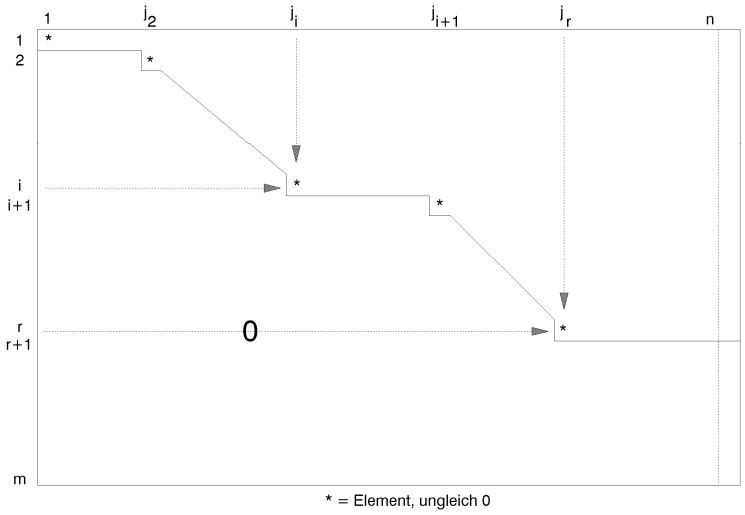
Um die Einträge von $\mathbf{A}^{(i)}$ von den Einträgen der übrigen Matrizen unterscheiden zu können, wird

$$\mathbf{A}^{(i)} = \left[\begin{array}{ccccccc} a_{1,1}^{(i)} & a_{1,2}^{(i)} & \dots & a_{1,j}^{(i)} & \dots & a_{1,n}^{(i)} & \left| \; b_1^{(i)} \right. \\ a_{2,1}^{(i)} & a_{2,2}^{(i)} & \dots & a_{2,j}^{(i)} & \dots & a_{2,n}^{(i)} & \left| \; b_2^{(i)} \right. \\ \cdot & & & & & & \cdot \quad \cdot \\ \cdot & & & & & & \cdot \quad \cdot \\ \cdot & & & & & & \cdot \quad \cdot \\ a_{i,1}^{(i)} & a_{i,2}^{(i)} & \dots & a_{i,j}^{(i)} & \dots & a_{i,n}^{(i)} & \left| \; b_i^{(i)} \right. \\ \cdot & & & & & & \cdot \quad \cdot \\ \cdot & & & & & & \cdot \quad \cdot \\ \cdot & & & & & & \cdot \quad \cdot \\ a_{m,1}^{(i)} & a_{m,2}^{(i)} & \dots & a_{m,j}^{(i)} & \dots & a_{m,n}^{(i)} & \left| \; b_m^{(i)} \right. \end{array}\right] = \left[\begin{array}{ccc} \vec{a}_1^{(i)} & \left| \right. & b_1^{(i)} \\ \vec{a}_2^{(i)} & \left| \right. & b_2^{(i)} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \vec{a}_i^{(i)} & \left| \right. & b_i^{(i)} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \vec{a}_m^{(i)} & \left| \right. & b_m^{(i)} \end{array}\right]$$

gesetzt.

Zusätzlich wird eine Folge von Spaltennummern $j_1, ..., j_r$ erzeugt, deren Bedeutung aus dem Zusammenhang klar wird.

Die Matrix $\mathbf{A}^{(r)}$, die nach dem *r*-ten Umformungsvorgang entstanden ist, hat die Form



1. Schritt:

In der 1. Spalte von $\left[\mathbf{A}\middle|\,\vec{b}\right]$ wird von oben nach unten das erste von 0 verschiedene Element, d.h. ein Element der Form $a_{s,1} \neq 0$, gesucht.

Es wird $p := a_{s,1}$ gesetzt. Man nennt *p* das **Pivot-Element (im 1. Schritt)**.

Ist $s > 1$, so wird die erste Zeile $\left[\vec{a}_1 \middle|\, b_1\right]$ von $\left[\mathbf{A}\middle|\,\vec{b}\right]$ mit der *s*-ten Zeile ausgetauscht; ist bereits $a_{1,1} \neq 0$ (d.h. $s = 1$), so findet kein Austausch statt. Die erste Zeile der durch den eventuellen Zeilenaustausch entstandenen Matrix werde wieder mit $\left[\vec{a}_1 \middle|\, b_1\right]$ bezeichnet; entsprechend erhält die ursprünglich erste und nun an der *s*-ten Position stehende Zeile wieder die Bezeichnung $\left[\vec{a}_s \middle|\, b_s\right]$. Insbesondere ist mit dieser Numerierung $p = a_{1,1}$.

Für $k = 2, ..., m$ wird anschließend die Zeile $\left[\vec{a}_k \middle|\, b_k\right]$ durch

$$-a_{k,1} \cdot \left[\vec{a}_1 \middle|\, b_1\right] + p \cdot \left[\vec{a}_k \middle|\, b_k\right]$$

ersetzt.

$\mathbf{A}^{(1)}$ ist die so aus $\left[\mathbf{A}\middle|\,\vec{b}\right]$ entstandene Matrix. Es wird $j_1 := 1$ gesetzt.

Ergebnis: Alle Zeilen von $\mathbf{A}^{(1)}$ ab Zeile 2 enthalten mindestens in der ersten Spalte den Wert 0; es gilt außerdem $a_{1,j_1}^{(1)} \neq 0$. Eventuell sind auch in der zweiten und einigen folgenden Spalten von Zeile 2 abwärts ausschließlich die Werte 0 entstanden.

Es wird $i = 2$ gesetzt und im *i*-ten Schritt fortgefahren.

i-ter Schritt für $1 < i \leq m$ *:*

Die Matrix $\mathbf{A}^{(i-1)}$ sei bereits bestimmt. Sie hat die Form

$$\mathbf{A}^{(i-1)} = \left[\begin{array}{cccccccccccc|c}
a_{1,1}^{(i-1)} & a_{1,2}^{(i-1)} & \ldots & . & . & . & . & . & . & . & \ldots & a_{1,n}^{(i-1)} & b_1^{(i-1)} \\
0 & 0 & \ldots & 0 & a_{2,j_2}^{(i-1)} & a_{2,j_2+1}^{(i-1)} & . & . & . & . & \ldots & a_{2,n}^{(i-1)} & b_2^{(i-1)} \\
. & . & \ldots & . & . & . & . & . & . & . & \ldots & . & . \\
0 & 0 & \ldots & . & . & . & . & 0 & a_{i-1,j_{i-1}}^{(i-1)} & a_{i-1,j_{i-1}+1}^{(i-1)} & \ldots & a_{i-1,n}^{(i-1)} & b_{i-1}^{(i-1)} \\
0 & 0 & \ldots & . & . & . & . & 0 & 0 & a_{i,j_{i-1}+1}^{(i-1)} & \ldots & a_{i,n}^{(i-1)} & b_i^{(i-1)} \\
0 & 0 & \ldots & . & . & . & . & 0 & 0 & a_{i+1,j_{i-1}+1}^{(i-1)} & \ldots & a_{i+1,n}^{(i-1)} & b_{i+1}^{(i-1)} \\
. & . & \ldots & . & . & . & . & . & . & . & \ldots & . & . \\
0 & 0 & \ldots & . & . & . & . & 0 & 0 & a_{m,j_{i-1}+1}^{(i-1)} & \ldots & a_{m,n}^{(i-1)} & b_m^{(i-1)}
\end{array}\right]$$

$$= \left[\begin{array}{c|c}
\vec{a}_1^{(i-1)} & b_1^{(i-1)} \\
\vec{a}_2^{(i-1)} & b_2^{(i-1)} \\
. & . \\
\vec{a}_{i-1}^{(i-1)} & b_{i-1}^{(i-1)} \\
\vec{a}_i^{(i-1)} & b_i^{(i-1)} \\
\vec{a}_{i+1}^{(i-1)} & b_{i+1}^{(i-1)} \\
. & . \\
\vec{a}_m^{(i-1)} & b_m^{(i-1)}
\end{array}\right]$$

Es gilt für jede Zeile k mit $1 \le k \le i-1$:

- alle Elemente bis zur Spalte $j_k - 1$ (einschließlich) sind gleich 0
- $a_{k,j_k}^{(i-1} \neq 0$
- alle Elemente in der Teilmatrix, die durch die Zeilen i und m und die Spalten 1 und j_{i-1} (einschließlich) begrenzt wird, sind gleich 0.

In der Teilmatrix

$$\left[\begin{array}{ccc|c}
a_{i,j_{i-1}+1}^{(i-1)} & \ldots & a_{i,n}^{(i-1)} & b_i^{(i-1)} \\
a_{i+1,j_{i-1}+1}^{(i-1)} & \ldots & a_{i+1,n}^{(i-1)} & b_{i+1}^{(i-1)} \\
. & \ldots & . & . \\
a_{m,j_{i-1}+1}^{(i-1)} & \ldots & a_{m,n}^{(i-1)} & b_m^{(i-1)}
\end{array}\right]$$

(das ist der untere rechte Teil) wird von links nach rechts gehend diejenige Spalte bestimmt, die zum ersten Mal Einträge enthält, die nicht sämtlich gleich 0 sind (hierbei wird die Zeilen- und Spaltennumerierung aus der Matrix übernommen):

Es wird also $j = j_{i-1} + 1$ gesetzt und die Bedingung

$$a_{i,j}^{(i-1)} = a_{i+1,j}^{(i-1)} = \ldots = a_{m,j}^{(i-1)} = 0$$

geprüft. Gilt diese Bedingung und ist $j < n$ (= Anzahl der Spalten von **A**), so wird *j* um 1 erhöht und die Bedingung erneut geprüft; gilt die Bedingung und ist bereits $j = n$, so ist das Verfahren beendet.

Im folgenden sei *j* der kleinste Wert, für den die Bedingung nicht gilt, d.h. die Teilmatrix

$$\left[\begin{array}{ccc|c} a_{i,j_{i-1}+1}^{(i-1)} & \ldots & a_{i,n}^{(i-1)} & b_i^{(i-1)} \\ a_{i+1,j_{i-1}+1}^{(i-1)} & \ldots & a_{i+1,n}^{(i-1)} & b_{i+1}^{(i-1)} \\ . & \ldots & . & . \\ a_{m,j_{i-1}+1}^{(i-1)} & \ldots & a_{m,n}^{(i-1)} & b_m^{(i-1)} \end{array}\right]$$

enthält in der Spalte *j* ein Element, das ungleich 0 ist. Die kleinste Zeilennummer, für die das zutrifft, laute *s*, d.h.

$$\begin{aligned} & a_{i,j_{i-1}+1}^{(i-1)} = a_{i+1,j_{i-1}+1}^{(i-1)} = \ldots a_{m,j_{i-1}+1}^{(i-1)} \\ & = \ldots \\ & = a_{i,j-1}^{(i-1)} = a_{i+1,j-1}^{(i-1)} = \ldots a_{m,j-1}^{(i-1)} \\ & = a_{i,j}^{(i-1)} = a_{i+1,j}^{(i-1)} = \ldots a_{s-1,j}^{(i-1)} \\ & = 0 \\ & \text{und } a_{s,j}^{(i-1)} \neq 0. \end{aligned}$$

Es wird $p = a_{s,j}^{(i-1)}$ gesetzt. Der Wert *p* heißt **Pivot-Element (im *i*-ten Schritt)**.

Die *i*-te Zeile von $\mathbf{A}^{(i-1)}$ wird mit der *s*-ten Zeile ausgetauscht (und die Numerierungen der Zeilen wie im ersten Schritt angepaßt).

Es wird $j_i = j$ gesetzt.

Für $k = i+1, \ldots, m$ wird nun Zeile $\left[\vec{a}_k^{(i-1)} \middle| b_k^{(i-1)}\right]$ durch

$$-a_{k,j_i} \cdot \left[\vec{a}_i^{(i-1)} \middle| b_i^{(i-1)}\right] + p \cdot \left[\vec{a}_k^{(i-1)} \middle| b_k^{(i-1)}\right]$$

ersetzt.

Die so entstandene Matrix ist $\mathbf{A}^{(i)}$.

Es wird *i* um 1 erhöht und der *i*-te Schritt mit diesem neuen Wert für *i* wiederholt.

Nach dem *r*-ten Schritt hat die Matrix $\mathbf{A}^{(r)}$ die oben dargestellte Form

$$\mathbf{A}^{(r)} = \left[\begin{array}{ccccccccc|c} a_{1,j_1}^{(r)} & \ldots & . & . & . & . & \ldots & a_{1,n}^{(r)} & | & b_1^{(r)} \\ 0 & \ldots & 0 & a_{2,j_2}^{(r)} & . & . & \ldots & a_{2,n}^{(r)} & | & b_2^{(r)} \\ . & \ldots & . & . & . & . & \ldots & . & | & . \\ 0 & \ldots & 0 & . & 0 & a_{r,j_r}^{(r)} & \ldots & a_{r,n}^{(r)} & | & b_r^{(r)} \\ 0 & \ldots & 0 & . & 0 & . & \ldots & 0 & | & b_{r+1}^{(r)} \\ . & \ldots & . & . & . & . & \ldots & 0 & | & . \\ 0 & \ldots & 0 & . & 0 & . & \ldots & 0 & | & b_m^{(r)} \end{array}\right]$$

mit $1 = j_1 < j_2 < \ldots < j_r$ und $a_{i,j_i}^{(r)} \neq 0$ für $i = 1, \ldots, r$.

Ist $b_{r+1}^{(r)} = \ldots b_m^{(r)} = 0$, so ist $r(\mathbf{A}) = r\left(\left[\mathbf{A} \middle| \vec{b}\right]\right) = r\left(\mathbf{A}^{(r)}\right) = r$, und das Gleichungssystem ist lösbar. Andernfalls ist das Gleichungssystem nicht lösbar.

Im folgenden sei $r(\mathbf{A}) = r\left(\left[\mathbf{A} \middle| \vec{b}\right]\right) = r\left(\mathbf{A}^{(r)}\right) = r$.

Es gilt:

Das ursprüngliche Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ und das *r*-zeilige Gleichungssystem $\mathbf{A}^{(r)} \cdot \vec{x} = \vec{b}^{(r)}$ haben dieselbe Lösung, da nur elementare Umformungen durchgeführt wurden. In ausgeschriebener Form (ohne die Zeilen, die nur Nullen enthalten) lautet

$\mathbf{A}^{(r)} \cdot \vec{x} = \vec{b}^{(r)}$:

$$\begin{array}{ccccccccccc} a_{1,1}^{(r)} \cdot x_1 & + & & & \ldots & & & + & a_{1,n}^{(r)} \cdot x_n & = & b_1^{(r)} \\ & & a_{2,j_2}^{(r)} \cdot x_2 & + & & \ldots & & + & a_{2,n}^{(r)} \cdot x_n & = & b_2^{(r)} \\ & & & & . \quad . \quad . \quad . \quad . \quad . & & & . & . & . & \\ & & & & a_{r,j_r}^{(r)} \cdot x_{j_r} & & & & a_{r,n}^{(r)} \cdot x_n & = & b_r^{(r)} \end{array}$$

Die Zeilen dieser Matrix (es sind die ersten *r* Zeilen von $\mathbf{A}^{(r)}$) werden von ***unten nach oben*** bearbeitet, und dabei werden den Variablen $x_n, x_{n-1}, \ldots, x_1$ Werte zugeordnet:

(r) Bearbeitung der Zeile mit der Nummer r:

Den Variablen $x_{j_r+1}, x_{j_r+2}, ..., x_n$ werden beliebige Werte aus **R** zugewiesen (freie Variablen):

$$x_{j_r+1} := u_{j_r+1}, \quad x_{j_r+2} := u_{j_r+2}, \quad ... \quad , x_n := u_n .$$

x_{j_r} wird aus der letzten Gleichung berechnet:

$$x_{j_r} = \frac{1}{a_{r,j_r}^{(r)}} \cdot \left(b_r^{(r)} - \sum_{k=j_r+1}^{n} a_{r,k}^{(r)} \cdot x_k \right) = \frac{1}{a_{r,j_r}^{(r)}} \cdot \left(b_r^{(r)} - \sum_{k=j_r+1}^{n} a_{r,k}^{(r)} \cdot u_k \right).$$

(i) Bearbeitung der Zeile mit der Nummer i mit $1 \le i < r$:

Die Zeilen $i+1, ..., r$ seien bereits bearbeitet. Die Variablen, die bisher entweder als freie oder berechnete Variablen ermittelt wurden, seien in aufsteigender Numerierung $x_k, x_{k+1}, ..., x_n$.

Den Variablen $x_{j_i+1}, ..., x_{k-1}$ werden wieder beliebige Werte aus **R** zugewiesen (freie Variablen):

$$x_{j_i+1} = u_{j_i+1}, \quad ... \quad , x_{k-1} := u_{k-1} .$$

x_{j_i} wird berechnet zu:

$$x_{j_i} = \frac{1}{a_{i,j_i}^{(r)}} \cdot \left(b_i^{(r)} - \sum_{k=j_i+1}^{n} a_{r,k}^{(r)} \cdot x_k \right).$$

Beispiel:

Das Gleichungssystem

$$\begin{array}{rcrcrcrcrcrcrcrcrcrcr}
-4 & x_1 & +4 & x_2 & -8 & x_3 & -24 & x_4 & -44 & x_5 & +4 & x_6 & -56 & x_7 & -44 & x_8 & = & -24 \\
3 & x_1 & -3 & x_2 & +6 & x_3 & +18 & x_4 & +30 & x_5 & -9 & x_6 & +42 & x_7 & +24 & x_8 & = & 15 \\
2 & x_1 & -2 & x_2 & +4 & x_3 & +10 & x_4 & +16 & x_5 & -4 & x_6 & +20 & x_7 & +12 & x_8 & = & 8 \\
-2 & x_1 & +2 & x_2 & -4 & x_3 & -12 & x_4 & -18 & x_5 & +10 & x_6 & -28 & x_7 & -10 & x_8 & = & -8 \\
2 & x_1 & -2 & x_2 & +4 & x_3 & +10 & x_4 & +18 & x_5 & & & +20 & x_7 & +18 & x_8 & = & 10
\end{array}$$

hat die erweiterte Koeffizientenmatrix

$$\left[\begin{array}{rrrrrrrr|r} -4 & 4 & -8 & -24 & -44 & 4 & -56 & -44 & -24 \\ 3 & -3 & 6 & 18 & 30 & -9 & 42 & 24 & 15 \\ 2 & -2 & 4 & 10 & 16 & -4 & 20 & 12 & 8 \\ -2 & 2 & -4 & -12 & -18 & 10 & -28 & -10 & -8 \\ 2 & -2 & 4 & 10 & 18 & 0 & 20 & 18 & 10 \end{array}\right]$$

1. Schritt:

$p = a_{1,1} = -4 \neq 0$; die *k*-te Zeile für $k = 2, 3, 4, 5$ wird ersetzt durch $-a_{k,1} \cdot (1.\,\text{Zeile}) - 4 \cdot (k\text{-te Zeile})$. Das ergibt:

$$\left[\begin{array}{rrrrrrrr|r} -4 & 4 & -8 & -24 & -44 & 4 & -56 & -44 & -24 \\ 0 & 0 & 0 & 0 & 12 & 24 & 0 & 36 & 12 \\ 0 & 0 & 0 & 8 & 24 & 8 & 32 & 40 & 16 \\ 0 & 0 & 0 & 0 & -16 & -32 & 0 & -48 & -16 \\ 0 & 0 & 0 & 8 & 16 & -8 & 32 & 16 & 8 \end{array}\right]$$

Um die Größen der Zahlen zu reduzieren, werden die einzelnen Zeilen jeweils durch einen geeigneten Faktor dividiert, z.B. wird die 1. Zeile durch –4, die 2. Zeile durch 12, die 3. Zeile durch 8, die 4. Zeile durch –16 und die 5. Zeile durch 8 dividiert, und es entsteht:

$$\mathbf{A}^{(1)} = \left[\begin{array}{rrrrrrrr|r} 1 & -1 & 2 & 6 & 11 & -1 & 14 & 11 & 6 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 & 3 & 1 & 4 & 5 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 & 2 & -1 & 4 & 2 & 1 \end{array}\right]$$

$j_1 = 1$

i-ter Schritt für i = 2:

Es ist $j_2 = 4$, $s = 3$, $p = 1$. Die 2. Zeile von $\mathbf{A}^{(1)}$ wird mit der 3. Zeile ausgetauscht, und es ergibt sich

$$\left[\begin{array}{cccccccc|c} 1 & -1 & 2 & 6 & 11 & -1 & 14 & 11 & 6 \\ 0 & 0 & 0 & 1 & 3 & 1 & 4 & 5 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 & 2 & -1 & 4 & 2 & 1 \end{array}\right]$$

Für $k = 3, 4, 5$ wird die *k*-te Zeile ersetzt durch

$$-a_{k,4} \cdot (2.\,\text{Zeile}) + 1 \cdot (k\text{-te Zeile}).$$

Damit ergibt sich

$$\mathbf{A}^{(2)} = \left[\begin{array}{cccccccc|c} 1 & -1 & 2 & 6 & 11 & -1 & 14 & 11 & 6 \\ 0 & 0 & 0 & 1 & 3 & 1 & 4 & 5 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & -1 & -2 & 0 & -3 & -1 \end{array}\right]$$

i-ter Schritt für i = 3:

Es ist $j_3 = 5,\ s = 3,\ p = 1$.

Für $k = 4, 5$ wird die *k*-te Zeile ersetzt durch

$$-a_{k,5} \cdot (3.\,\text{Zeile}) + 1 \cdot (k\text{-te Zeile})$$

Damit ergibt sich

$$\mathbf{A}^{(3)} = \left[\begin{array}{cccccccc|c} 1 & -1 & 2 & 6 & 11 & -1 & 14 & 11 & 6 \\ 0 & 0 & 0 & 1 & 3 & 1 & 4 & 5 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

i-ter Schritt für i = 4:

Für $j = j_3 + 1\ \ (= 6), ..., 8\ \ (= n)$ gilt jeweils $a_{4,j}^{(3)} = a_{5,j}^{(3)} = 0$, so dass das Verfahren abbricht. Es ist $r = 3$.

Es wird daher gesetzt:

$x_6 = u_6, \quad x_7 = u_7, \quad x_8 = u_8,$

$$x_5 = \frac{1}{1} \cdot \left(1 - \left(2 \cdot x_6 + 0 \cdot x_7 + 3 \cdot x_8\right)\right) = 1 - 2 \cdot u_6 - 3 \cdot u_8,$$

$$x_4 = \frac{1}{1} \cdot \left(2 - \left(3 \cdot x_5 + 1 \cdot x_6 + 4 \cdot x_7 + 5 \cdot x_8\right)\right) = -1 + 5 \cdot u_6 - 4 \cdot u_7 + 4 \cdot u_8,$$

$x_2 = u_2, \quad x_3 = u_3,$

$$x_1 = \frac{1}{1} \cdot \left(6 - \left(-1 \cdot x_2 + 2 \cdot x_3 + 6 \cdot x_4 + 11 \cdot x_5 - 1 \cdot x_6 + 14 \cdot x_7 + 11 \cdot x_8\right)\right)$$
$$= 1 + u_2 - 2 \cdot u_3 - 7 \cdot u_6 + 10 \cdot u_7 - 2 \cdot u_8.$$

Die (unendlich große) Lösungsmenge des Gleichungssystems ist also

$$L = \left\{ \vec{X}_{(8,1)} \;\middle|\; \vec{X} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + u_2 \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + u_3 \cdot \begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + u_6 \cdot \begin{bmatrix} -7 \\ 0 \\ 0 \\ 5 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + u_7 \cdot \begin{bmatrix} 10 \\ 0 \\ 0 \\ -4 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + u_8 \cdot \begin{bmatrix} -2 \\ 0 \\ 0 \\ 4 \\ -3 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$
$$\text{mit beliebigen reellen Zahlen } u_2, u_3, u_6, u_7 \text{ und } u_8$$

6.3 Invertieren von Matrizen

Eine quadratische Matrix $\mathbf{A}_{(n,n)}$ vom Typ (*n*, *n*) heißt **regulär**, wenn $r\left(\mathbf{A}_{(n,n)}\right) = n$ ist. Sie heißt **singulär**, wenn $r\left(\mathbf{A}_{(n,n)}\right) < n$ gilt.

Es sei $\mathbf{A}_{(n,n)}$ eine *quadratische* Matrix vom Typ (*n*, *n*). Gibt es eine Matrix $\mathbf{B}_{(n,n)}$ vom Typ (n, n) mit $\mathbf{A}_{(n,n)} \cdot \mathbf{B}_{(n,n)} = \mathbf{I}_{(n,n)}$, dann heißt $\mathbf{B}_{(n,n)}$ die zu $\mathbf{A}_{(n,n)}$ **inverse Matrix** und wird mit $\mathbf{A}^{-1}_{(n,n)}$ bezeichnet.

Zur Erinnerung: Mit $\mathbf{I}_{(n,n)}$ wird die quadratische Matrix bezeichnet, die in der Diagonalen die Zahlen 1 und sonst nur Nullen enthält (Einheitsmatrix).

Satz 6.3-1:

A und **B** seien quadratische Matrizen, zu denen jeweils die inversen Matrizen $\mathbf{A}^{-1}$ und $\mathbf{B}^{-1}$ existieren. Dann gilt:

(i) Aus $\mathbf{B}\cdot\mathbf{A}=\mathbf{A}$ folgt $\mathbf{B}=\mathbf{I}$.

(ii) $\mathbf{I}=\mathbf{A}\cdot\mathbf{A}^{-1}=\mathbf{A}^{-1}\cdot\mathbf{A}$.

(iii) $\mathbf{A}^{-1}$ ist eindeutig bestimmt.

(iv) $\left(\mathbf{A}^{-1}\right)^{-1}=\mathbf{A}$.

(v) $\left(\mathbf{A}\cdot\mathbf{B}\right)^{-1}=\mathbf{B}^{-1}\cdot\mathbf{A}^{-1}$.

(vi) $\left(k\cdot\mathbf{A}\right)^{-1}=1/k\cdot\mathbf{A}^{-1}$ für $k\in\mathbf{R}_{\neq 0}$.

Zunächst überzeugt man sich durch Nachrechnen, dass für quadratische Matrizen **A**, **B** und **C** die Assoziativität $\mathbf{A}\cdot\left(\mathbf{B}\cdot\mathbf{C}\right)=\left(\mathbf{A}\cdot\mathbf{B}\right)\cdot\mathbf{C}$ gilt, d.h. die Klammerung bei einer Folge von Matrixmultiplikationen ist irrelevant.

(i) ergibt sich folgendermaßen: Ist $\mathbf{B}\cdot\mathbf{A}=\mathbf{A}$, dann ist $\mathbf{B}\cdot\mathbf{A}\cdot\mathbf{A}^{-1}=\mathbf{A}\cdot\mathbf{A}^{-1}=\mathbf{I}$ und $\mathbf{B}\cdot\mathbf{A}\cdot\mathbf{A}^{-1}=\mathbf{B}$.

Damit folgt (ii): $\mathbf{A}^{-1}=\mathbf{A}^{-1}\cdot\left(\mathbf{A}\cdot\mathbf{A}^{-1}\right)=\left(\mathbf{A}^{-1}\cdot\mathbf{A}\right)\cdot\mathbf{A}^{-1}$, und mit (i) ist $\mathbf{A}^{-1}\cdot\mathbf{A}=\mathbf{I}=\mathbf{A}\cdot\mathbf{A}^{-1}$.

Die Eindeutigkeit der inversen Matrix in (iii) sieht man wie folgt: Ist **B** eine zu **A** inverse Matrix, d.h. ist $\mathbf{A}\cdot\mathbf{B}=\mathbf{I}$, dann ist $\mathbf{A}^{-1}=\mathbf{A}^{-1}\cdot\left(\mathbf{A}\cdot\mathbf{B}\right)=\left(\mathbf{A}^{-1}\cdot\mathbf{A}\right)\cdot\mathbf{B}=\left(\mathbf{A}\cdot\mathbf{A}^{-1}\right)\cdot\mathbf{B}=\mathbf{B}$.

Nach Definition ist $\mathbf{I}=\mathbf{A}\cdot\mathbf{A}^{-1}$ und $\mathbf{A}\cdot\mathbf{A}^{-1}=\mathbf{A}^{-1}\cdot\mathbf{A}$. Also ist $\mathbf{A}^{-1}$ invertierbar, und wegen der Eindeutigkeit der inversen Matrix ist $\left(\mathbf{A}^{-1}\right)^{-1}=\mathbf{A}$.

Mit (iii) ergibt sich (v): $\left(\mathbf{A}\cdot\mathbf{B}\right)\cdot\left(\mathbf{B}^{-1}\cdot\mathbf{A}^{-1}\right)=\mathbf{A}\cdot\left(\mathbf{B}\cdot\mathbf{B}^{-1}\right)\cdot\mathbf{A}^{-1}=\mathbf{A}\cdot\mathbf{A}^{-1}=\mathbf{I}$, also $\left(\mathbf{A}\cdot\mathbf{B}\right)^{-1}=\mathbf{B}^{-1}\cdot\mathbf{A}^{-1}$.

Entsprechend sieht man die Gültigkeit von (v): $(k \cdot \mathbf{A}) \cdot (1/k \cdot \mathbf{A}^{-1}) = \mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}$, also $(k \cdot \mathbf{A})^{-1} = 1/k \cdot \mathbf{A}^{-1}$.

Satz 6.3-2:

Für eine quadratische Matrix **A** sind folgende Aussagen (a) und (b) äquivalent:

(a) **A** ist eine reguläre Matrix.

(b) Zu **A** existiert die inverse Matrix $\mathbf{A}^{-1}$.

Den Beweis dieses Satzes, der einen tieferen Einstieg in die Lineare Algebra erfordert, findet man in der angegebenen Literatur[6].

Satz 6.3-3:

Die Lösung der Matrixgleichung $\mathbf{A} \cdot \mathbf{X} = \mathbf{B}$ mit einer regulären Matrix **A** lautet $\mathbf{X} = \mathbf{A}^{-1} \cdot \mathbf{B}$.

Die Berechnung der inversen Matrix zu einer gegebenen quadratischen regulären Matrix **A** heißt **Invertieren der Matrix A**.

Die quadratische Matrix $\mathbf{A}_{(n,n)} = \lfloor a_{i,j} \rfloor_{(n,n)}$ sei regulär. Man kann zeigen, dass man die Zeilen einer regulären Matrix so vertauschen kann, dass nach dem Austausch alle Elemente der Diagonalen von 0 verschieden sind. Daher kann man gleich für $a_{i,i} \neq 0$ für $i = 1, ..., n$ voraussetzen.

Die Matrix $\mathbf{X}_{(n,n)} = \lfloor x_{i,j} \rfloor_{(n,n)}$ sei in Spaltenschreibweise:

$\mathbf{X} = [\vec{x}_1, ..., \vec{x}_n]$.

Der Vektor $\vec{e}_i$ für $i = 1, ..., n$ sei der Spaltenvektor, der in der *i*-ten Zeile eine 1 und sonst nur Nullen hat.

---

6 Beutelspacher, A.: **Lineare Algebra**, 7. Aufl.,Vieweg+Teubner, 2009.

Zur Invertierung der Matrix **A** sind simultan die *n* linearen Gleichungssysteme

$$\mathbf{A} \cdot \vec{x}_1 = \vec{e}_1, \ldots, \mathbf{A} \cdot \vec{x}_n = \vec{e}_n$$

zu lösen. Diese Gleichungssysteme kann man zu einem Gleichungssystem

$$\mathbf{A}_{(n,n)} \cdot \mathbf{X}_{(n,n)} = \mathbf{I}_{(n,n)}$$

zusammenfassen und mit einer Variante des Gaußschen Verfahrens lösen (anstelle des Vektors $\vec{b}_{(m,1)}$ steht jetzt die Einheitsmatrix $\mathbf{I}_{(n,n)}$):

Die zu **A** gehörende erweiterte Matrix hat die Form

$$[\mathbf{A} \mid \mathbf{I}] = \begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} & \mid & 1 & 0 & 0 & \ldots & 0 & 0 \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} & \mid & 0 & 1 & 0 & \ldots & 0 & 0 \\ . & . & \ldots & . & \mid & . & . & . & \ldots & . & . \\ a_{n,1} & a_{n,2} & \ldots & a_{n,n} & \mid & 0 & 0 & 0 & \ldots & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \vec{a}_1 & \mid & 1 & 0 & 0 & \ldots & 0 & 0 \\ \vec{a}_1 & \mid & 0 & 1 & 0 & \ldots & 0 & 0 \\ . & \mid & . & . & . & \ldots & . & . \\ \vec{a}_n & \mid & 0 & 0 & 0 & \ldots & 0 & 1 \end{bmatrix}$$

Sie wird schrittweise durch elementare Umformungen in eine „erweiterte Diagonalmatrix“ umgewandelt. Dabei wird eine Folge von Matrizen $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(n)}$ erzeugt; $\mathbf{A}^{(i)}$ ist das Ergebnis nach dem *i*-ten Schritt:

1. Schritt:

Es ist $a_{1,1} \neq 0$. Die 1. Zeile der erweiterten Matrix wird durch $a_{1,1}$ geteilt. Anschließend wird für $k = 2, \ldots, n$ die mit $-a_{k,1}$ multiplizierte 1. Zeile zur *k*-ten Zeile addiert. Das Ergebnis ist $\mathbf{A}^{(1)}$. Das Element in der 1. Spalte und der 1. Zeile ist gleich 1; alle Elemente der 1. Spalte ab Zeile 2 sind gleich 0.

Anschließend wird $i = 2$ gesetzt.

i-ter Schritt für $1 < i \leq n$:

Die Matrix $\mathbf{A}^{(i-1)}$ sei bereits ermittelt. Sie hat die Form

$$\mathbf{A}^{(i-1)} = \left[\begin{array}{cccccc|cccc|ccc}
1 & 0 & 0 & \dots & 0 & 0 & a_{1,i}^{(i-1)} & a_{1,i+1}^{(i-1)} & \dots & a_{1,n}^{(i-1)} & u_{1,1}^{(i-1)} & \dots & u_{1,n}^{(i-1)} \\
0 & 1 & 0 & \dots & 0 & 0 & a_{2,i}^{(i-1)} & a_{2,i+1}^{(i-1)} & \dots & a_{2,n}^{(i-1)} & u_{2,1}^{(i-1)} & \dots & u_{2,n}^{(i-1)} \\
. & . & . & \dots & . & . & . & . & \dots & . & . & \dots & . \\
0 & 0 & 0 & \dots & 0 & 1 & a_{i-1,i}^{(i-1)} & a_{i-1,i+1}^{(i-1)} & \dots & a_{i-1,n}^{(i-1)} & u_{i-1,1}^{(i-1)} & \dots & u_{i-1,n}^{(i-1)} \\
0 & 0 & 0 & \dots & 0 & 0 & a_{i,i}^{(i-1)} & a_{i,i+1}^{(i-1)} & \dots & a_{i,n}^{(i-1)} & u_{i,1}^{(i-1)} & \dots & u_{i,n}^{(i-1)} \\
0 & 0 & 0 & \dots & 0 & 0 & a_{i+1,i}^{(i-1)} & a_{i+1,i+1}^{(i-1)} & \dots & a_{i+1,n}^{(i-1)} & u_{i+1,1}^{(i-1)} & \dots & u_{i+1,n}^{(i-1)} \\
. & . & . & \dots & . & . & . & . & \dots & . & . & \dots & . \\
0 & 0 & 0 & \dots & 0 & 0 & a_{n,i}^{(i-1)} & a_{n,i+1}^{(i-1)} & \dots & a_{n,n}^{(i-1)} & u_{n,1}^{(i-1)} & \dots & u_{n,n}^{(i-1)}
\end{array}\right]$$

$$= \left[\begin{array}{c|c}
\vec{a}_1^{(i-1)} & \vec{u}_1^{(i-1)} \\
\vec{a}_2^{(i-1)} & \vec{u}_2^{(i-1)} \\
. & . \\
\vec{a}_{i-1}^{(i-1)} & \vec{u}_{i-1}^{(i-1)} \\
\vec{a}_i^{(i-1)} & \vec{u}_i^{(i-1)} \\
\vec{a}_{i+1}^{(i-1)} & \vec{u}_{i+1}^{(i-1)} \\
. & . \\
\vec{a}_n^{(i-1)} & \vec{u}_n^{(i-1)}
\end{array}\right]$$

Man kann $a_{i,i}^{(i-1)} \neq 0$ annehmen; ansonsten gibt es wegen der Regularität von **A** ein $s \in \{i, i+1, ..., n\}$ mit $a_{s,i}^{(i-1)} \neq 0$, und die *s*-te Zeile wird mit der *i*-ten Zeile ausgetauscht.

Die *i*-te Zeile wird durch $a_{i,i}^{(i-1)}$ geteilt, so dass sie in der *i*-ten Spalte (im Diagonalelement) den Wert 1 hat. Anschließend wird für $k = 1, ..., i-1, i+1, ..., n$ die mit $-a_{k,i}^{(i-1)}$ multiplizierte *i*-te Zeile zur *k*-ten Zeile addiert. Zu beachten ist, dass hierbei sowohl Zeilen behandelt werden, die oberhalb der *i*-ten Zeile stehen ($k = 1, ..., i-1$), als auch Zeilen, die unterhalb der *i*-ten Zeile stehen ($k = i+1, ..., n$).

$\mathbf{A}^{(i)}$ ist die so entstandene Matrix.

Es wird *i* um1 erhöht und der *i*-te Schritt mit diesem neuen Wert für *i* wiederholt.

$\mathbf{A}^{(n)}$ hat die Form

$$\mathbf{A}^{(n)} = \left[\begin{array}{cccccccccc|ccc} 1 & 0 & 0 & ... & 0 & 0 & 0 & 0 & ... & 0 & u_{1,1}^{(n)} & ... & u_{1,n}^{(n)} \\ 0 & 1 & 0 & ... & 0 & 0 & 0 & 0 & ... & 0 & u_{2,1}^{(n)} & ... & u_{2,n}^{(n)} \\ . & . & . & ... & . & . & . & . & ... & . & . & ... & . \\ 0 & 0 & 0 & ... & 0 & 1 & 0 & 0 & ... & 0 & u_{i-1,1}^{(n)} & ... & u_{i-1,n}^{(n)} \\ 0 & 0 & 0 & ... & 0 & 0 & 1 & 0 & ... & 0 & u_{i,1}^{(n)} & ... & u_{i,n}^{(n)} \\ 0 & 0 & 0 & ... & 0 & 0 & 0 & 1 & ... & 0 & u_{i+1,1}^{(n)} & ... & u_{i+1,n}^{(n)} \\ . & . & . & ... & . & . & . & . & ... & . & . & ... & . \\ 0 & 0 & 0 & ... & 0 & 0 & 0 & 0 & ... & 1 & u_{n,1}^{(n)} & ... & u_{n,n}^{(n)} \end{array}\right]$$

$$= \left[\mathbf{I}_{(n,n)} \quad | \quad \mathbf{U}_{(n,n)}\right]$$

Es gilt $\mathbf{U}_{(n,n)} = \mathbf{A}^{-1}$.

Beispiel:

Bestimmung der zu $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$ inversen Matrix:

Die folgenden Matrizen sind die Ergebnisse nach den Schritten :

$$\mathbf{A}^{(1)} = \begin{bmatrix} 1 & 2 & 3 & | & 1 & 0 & 0 \\ 0 & -1 & -4 & | & -2 & 1 & 0 \\ 0 & 0 & -1 & | & -1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A}^{(2)} = \begin{bmatrix} 1 & 0 & -5 & | & -3 & 2 & 0 \\ 0 & 1 & 4 & | & 2 & -1 & 0 \\ 0 & 0 & -1 & | & -1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A}^{(3)} = \begin{bmatrix} 1 & 0 & 0 & | & 2 & 2 & -5 \\ 0 & 1 & 0 & | & -2 & -1 & 4 \\ 0 & 0 & 1 & | & 1 & 0 & -1 \end{bmatrix}$$

Es gilt $\mathbf{A}^{-1} = \begin{bmatrix} 2 & 2 & -5 \\ -2 & -1 & 4 \\ 1 & 0 & -1 \end{bmatrix}$.

7 Ausgewählte Themen der Wahrscheinlichkeitstheorie und Statistik

Das vorliegende Kapitel mit seinen Unterkapiteln gibt eine knappe Einführung in die Anwendungen der Schließenden Statistik und stellt dazu zunächst erforderliche Hilfsmittel aus der Wahrscheinlichkeitsrechnung zur Verfügung. Die einzelnen Sätze werden wieder bis auf wenige Ausnahmen, die weiterführende mathematische Betrachtungen erfordern, durch Argumente untermauert[7]. Die mathematischen Hilfsmittel sind entweder elementar oder ergeben sich aus den Ausführungen in Kapitel 5 und seinen Unterkapiteln.

Auf Zahlenbeispiele wird mit Verweis auf die angegebene Literatur weitgehend verzichtet. Das Kapitel dient vielmehr als Leitfaden, um sich das Themengebiet (auch selbständig) zu erarbeiten.

7.1 Bezeichnungen und Ergebnisse aus der Wahrscheinlichkeitstheorie

In diesem Kapitel werden Bezeichnungen und Ergebnisse aus der Wahrscheinlichkeitsrechnung zusammengestellt, wie sie in einer einführenden Veranstaltung behandelt werden.

Mit Ω werde eine **Menge von Elementareignissen** bezeichnet. Ω wird auch **Ereignisraum** genannt. Bei der Durchführung eines **Zufallsexperiments** (z.B. Münzwurf, Würfeln, Ziehen eines produzierten Teils aus einem Produktionsvorgang) wird ein Elementarereignis ausgewählt. Das Zufallsexperiment wird nach einer bestimmten Vorschrift durchgeführt und kann unter den gleichen Rahmenbedingungen beliebig oft wiederholt werden. Welches Elementarereignis jeweils gewählt wird, ist nicht vorhersagbar. Lediglich die Menge der sich gegenseitig ausschließenden Versuchsausgänge ist bekannt.

Ist Ω endlich oder abzählbar unendlich, so handelt es sich um einen **diskreten Ereignisraum**.

[7] Die ausgelassenen Beweise findet man in Fisz, M.: **Wahrscheinlichkeitsrechnung und mathematische Statistik**, 11. Aufl., Deutscher Verlag der Wissenschaften, 1989.

Ein Mengensystem $\mathbf{E}(\Omega) \subseteq \mathbf{P}(\Omega)$ heißt **Ereignismenge mit dem Ereignisraum** Ω, wenn folgende Bedingungen erfüllt sind:

(i) $\Omega \in \mathbf{E}(\Omega)$

(ii) Mit $A \in \mathbf{E}(\Omega)$ gilt auch $\overline{A}^{\Omega} \in \mathbf{E}(\Omega)$

(iii) Mit $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$ gilt auch $A \cup B \in \mathbf{E}(\Omega)$

(iv) Ist $(A_i)_{i \in I}$ eine Folge von abzählbar vielen Mengen $A_i \in \mathbf{E}(\Omega)$ für $i \in I$, $I \subseteq \mathbf{N}$, so ist auch $\bigcup_{i \in I} A_i \in \mathbf{E}(\Omega)$.

Mit $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$ gilt auch $A \cap B \in \mathbf{E}(\Omega)$; diese Aussage folgt aus einer mehrfachen Anwendung der Regeln (i) und (ii) auf $A \cap B = \overline{\overline{A}^{\Omega} \cup \overline{B}^{\Omega}}^{\Omega}$.

Eine Menge $A \in \mathbf{E}(\Omega)$, d.h. $A \subseteq \Omega$, heißt **(zufälliges) Ereignis**. Das Ereignis A **tritt bei einem Zufallsexperiment ein**, wenn das ausgewählte Elementarereignis Element von A ist. Die Menge Ω heißt **sicheres Ereignis**, die Menge $\varnothing$ **unmögliches Ereignis**.

Eine Abbildung P, die jedem Ereignis $A \in \mathbf{E}(\Omega)$ eine reelle Zahl zuordnet, heißt **Wahrscheinlichkeitsmaß** auf Ω, und $P(A)$ heißt **Wahrscheinlichkeit des Ereignisses** A, wenn gilt:

(i) $0 \le P(A) \le 1$ für jedes Ereignis $A \in \mathbf{E}(\Omega)$

(ii) $P(\Omega) = 1$

(iii) $P(A \cup B) = P(A) + P(B)$ für disjunkte Ereignisse $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$

(iv) Ist $(A_i)_{i \in I}$ eine Folge abzählbar vieler paarweise disjunkter Mengen $A_i \in \mathbf{E}(\Omega)$ für $i \in I$, $I \subseteq \mathbf{N}$, so ist $P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i)$.

Ist die Menge der Elementarereignisse endlich, so kann man in (iv) nur endlich viele paarweise disjunkte Ereignisse $A_i \subseteq \Omega$ bilden.

Beispiele:

1. Für $\Omega = \{a_1, ..., a_n\}$ und $\mathbf{E}(\Omega) = \mathbf{P}(\Omega)$ wird durch $P(\{a_i\}) = 1/n$ ein Wahrscheinlichkeitsmaß mit $P(A) = \frac{|A|}{n}$ festgelegt (**diskrete Gleichverteilung**).

2. Für $\Omega = \mathbf{N}$ und $\mathbf{E}(\Omega) = \mathbf{P}(\Omega)$ wird durch $P(\{n\}) = \left(\frac{1}{2}\right)^{n+1}$ ein Wahrscheinlichkeitsmaß festgelegt: Dazu ist lediglich die Bedingung (ii) zu prüfen:
$$P(\mathbf{N}) = P\left(\bigcup_{n\in\mathbf{N}} \{n\}\right) = \sum_{n\in\mathbf{N}} P(\{n\}) = \sum_{n\in\mathbf{N}} \left(\frac{1}{2}\right)^{n+1} = \frac{1}{2} \cdot \frac{1}{1 - 1/2} = 1.$$

Direkt aus den obigen Bedingungen ergibt sich

Satz 7.1-1:

Es sei Ω eine Menge von Elementarereignissen mit der Ereignismenge $\mathbf{E}(\Omega)$ und P ein Wahrscheinlichkeitsmaß auf Ω. Dann gilt:

(i) $P\left(\overline{A}^{\Omega}\right) = 1 - P(A)$ für jedes Ereignis $A \in \mathbf{E}(\Omega)$.

(ii) $P(\varnothing) = 0$.

(iii) $P(A \setminus B) = P(A) - P(B)$ für Ereignisse $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$ mit $B \subseteq A$.

(iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ für Ereignisse $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$.

(v) $P(A) \le P(B)$ für Ereignisse $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$ mit $A \subseteq B$.

Es seien $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$ Ereignisse eines Ereignisraums Ω, und es gelte $P(B) > 0$. Dann heißt $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ die **bedingte Wahrscheinlichkeit von *A* unter der Bedingung *B*** (**bedingte Wahrscheinlichkeit von *A*, wenn *B* eingetroffen ist**).

Bei festem $B \subseteq \Omega$ wird auf diese Weise ein Wahrscheinlichkeitsmaß auf Ω definiert. Dazu sind die obigen Bedingungen (i) bis (iv) nachzuweisen:

Es gilt $P(A \mid B) = \frac{P(A \cap B)}{P(B)} \geq 0$ und mit Satz 7.1-1 (v) wegen $A \cap B \subseteq B$ auch $\frac{P(A \cap B)}{P(B)} \leq 1$.

Weiter ist $P(\Omega \mid B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$. Sind $A_1 \in \mathbf{E}(\Omega)$ und $A_2 \in \mathbf{E}(\Omega)$ disjunkte Ereignisse, so ist

$$P(A_1 \cup A_2 \mid B) = \frac{P((A_1 \cup A_2) \cap B)}{P(B)} = \frac{P((A_1 \cap B) \cup (A_2 \cap B))}{P(B)} = \frac{P(A_1 \cap B) + P(A_2 \cap B)}{P(B)}$$
$$= P(A_1 \mid B) + P(A_2 \mid B).$$

Bedingung (iv) lässt sich auf ähnliche Art nachweisen.

Bedingte Wahrscheinlichkeiten lassen sich in vielen Anwendungen häufig einfacher als „absolute“ Wahrscheinlichkeiten berechnen, da dazu vorausgesetzt werden kann, dass die in Frage kommende Bedingung *B* bereits eingetroffen ist. Man besitzt also Zusatzinformationen in Form des eingetretenen Ereignisses *B*, um die Wahrscheinlichkeit eines Ereignisses *A* zu bewerten.

Aus der Definition der bedingten Wahrscheinlichkeit folgt unmittelbar:

Satz 7.1-2:

Es seien $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$ Ereignisse eines Ereignisraums Ω, und es sei $P(B) > 0$. Dann gilt:

$$P(A \cap B) = P(A \mid B) \cdot P(B).$$

Die Ereignisse $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$ heißen **(stochastisch) unabhängig**, wenn $P(A \cap B) = P(A) \cdot P(B)$ gilt. In diesem Fall gilt $P(A \mid B) = P(A)$ und $P(B \mid A) = P(B)$.

Die Ereignisse $A_1 \in \mathbf{E}(\Omega)$, …, $A_n \in \mathbf{E}(\Omega)$ heißen **(stochastisch) unabhängig**, wenn für beliebige Indizes k_1, …, k_s mit $1 \leq k_1 < \ldots < k_s \leq n$ die Beziehung
$P(A_{k_1} \cap \ldots \cap A_{k_s}) = P(A_{k_1}) \cdot \ldots \cdot P(A_{k_s})$ gilt.

Die abzählbar vielen Ereignisse $A_1 \in \mathbf{E}(\Omega)$, $A_2 \in \mathbf{E}(\Omega)$, $A_3 \in \mathbf{E}(\Omega)$, … heißen **(stochastisch) unabhängig**, wenn für jedes $n = 2, 3, 4, ...$ die Ereignisse A_1, …, A_n (stochastisch) unabhängig sind.

Der folgende Satz ist Grundlage vieler Berechnungen von Wahrscheinlichkeiten und weiteren Kenngrößen von Verteilungen (siehe unten).

Satz 7.1-3:

Die Ereignisse $B_1 \in \mathbf{E}(\Omega)$, …, $B_n \in \mathbf{E}(\Omega)$ seien eine paarweise disjunkte Zerlegung des Ereignisraums Ω, d.h. $\bigcup_{i=1}^{n} B_i = \Omega$ und $B_i \cap B_j = \varnothing$ für $i \neq j$. Dann gilt für das Ereignis $A \in \mathbf{E}(\Omega)$:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i).$$

Die Aussage folgt aus Satz 7.1-2 und den Eigenschaften des Wahrscheinlichkeitsmaßes *P*: Es ist $A = \bigcup_{i=1}^{n} (A \cap B_i)$ und $(A \cap B_i) \cap (A \cap B_j) = \varnothing$ für $i \neq j$. Daher gilt

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i).$$

Es seien $A \in \mathbf{E}(\Omega)$ und $B \in \mathbf{E}(\Omega)$ Ereignisse eines Ereignisraums Ω mit $P(A) > 0$ und $P(B) > 0$. Dann ist

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}.$$

Die bedingte Wahrscheinlichkeit $P(A \mid B)$ lässt sich also mit Hilfe der bedingten Wahrscheinlichkeit $P(B \mid A)$ ausdrücken. Wendet man auf $P(B)$ Satz 7.1-3 an, so erhält man den folgenden Satz.

Satz 7.1-4 (Satz von Bayes):

Die Ereignisse $A_1 \in \mathbf{E}(\Omega)$, …, $A_n \in \mathbf{E}(\Omega)$ seien eine paarweise disjunkte Zerlegung des Ereignisraums Ω, d.h. $\bigcup_{i=1}^{n} A_i = \Omega$ und $A_i \cap A_j = \varnothing$ für $i \neq j$. Dann gilt für das Ereignis $B \in \mathbf{E}(\Omega)$ mit $P(B) > 0$ und jedes $A_i \in \mathbf{E}(\Omega)$:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)}.$$

7.2 Zufallsvariablen

Es sei $\mathbf{E}(\Omega)$ eine Ereignismenge mit dem Ereignisraum Ω. Eine Abbildung $X : \Omega \to \mathbf{R}$ heißt **Zufallsvariable**, wenn zu jedem $r \in \mathbf{R}$ die Menge $A_r = \{ e \mid e \in \Omega \text{ und } X(e) \leq r \}$ in $\mathbf{E}(\Omega)$ ist.

Ist der Wertebereich einer Zufallsvariablen X endlich oder abzählbar unendlich, so heißt X **diskrete Zufallsvariable**, ansonsten **stetige Zufallsvariable**.

Unter der Wahrscheinlichkeit, dass die Zufallsvariable

- einen Wert $\leq a$ mit $a \in \mathbf{R}$ annimmt, geschrieben $P(X \leq a)$, versteht man $P(\{ e \mid e \in \Omega \text{ und } X(e) \leq a \})$
- den Wert $a \in \mathbf{R}$ annimmt, geschrieben $P(X = a)$, versteht man $P(\{ e \mid e \in \Omega \text{ und } X(e) = a \})$
- einen Wert zwischen a und b mit $a \in \mathbf{R}$ und $b \in \mathbf{R}$ annimmt, geschrieben $P(a < X \leq b)$, versteht man $P(\{ e \mid e \in \Omega \text{ und } a < X(e) \leq b \})$.

Die Funktion $F_X : \mathbf{R} \to [0,1]$ mit $F_X(x) = P(X \leq x) = P(\{ e \mid e \in \Omega \text{ und } X(e) \leq x \})$ heißt **Verteilungsfunktion** der Zufallsvariablen X. Das Subskript X wird fortgelassen, wenn der Zusammenhang klar ist.

Satz 7.2-1:

Für die Verteilungsfunktion einer F_X einer Zufallsvariablen X gilt:

(i) F_X ist an jeder Stelle $x \in \mathbf{R}$ zumindest rechtsseitig stetig, d.h. $\lim_{\Delta x \to 0} F_X(x + \Delta x) = F_X(x)$.

(ii) F_X ist monoton steigend, d.h. für $a \in \mathbf{R}$ und $b \in \mathbf{R}$ mit $a \le b$ ist $F_X(a) \le F_X(b)$.

(iii) F_X hat die Grenzwerte $\lim_{x \to -\infty} F_X(x) = 0$ und $\lim_{x \to \infty} F_X(x) = 1$.

Im folgenden werden Definitionen angeführt, die sich in der Darstellung für diskrete und stetige Zufallsvariablen unterscheiden.

Die Zufallsvariable X sei ***diskret***, d.h. sie nimmt endlich oder abzählbar unendlich viele Werte x_i an. Dann heißt die Funktion

$f_X : \mathbf{R} \to [0,1]$ mit $f_X(x) = P(X = x) = \begin{cases} P(X = x_i) & \text{für } x = x_i \\ 0 & \text{sonst} \end{cases}$ **(Wahrscheinlichkeits-) Massenfunktion** von X.

Offensichtlich gilt $0 \le f_X(x) \le 1$ für jedes $x \in \mathbf{R}$ und $\sum_{\text{alle } i} f_X(x_i) = 1$. Den Zusammenhang zwischen der Massenfunktion f_X und der Verteilungsfunktion F_X beschreibt
$F_X(x) = P(X \le x) = \sum_{x_i \le x} f_X(x)$ und $f_X(x) = F_X(x) - \lim_{\Delta x \to 0} F_X(x - \Delta x)$.

Die Zufallsvariable X sei ***stetig*** mit Verteilungsfunktion F_X. Die erste Ableitung f_X von F_X heißt **(Wahrscheinlichkeits-) Dichtefunktion** von X:
$f_X = \frac{d}{dx} F_X(x)$.

Aus der Definition und mit Satz 7.2-1 (ii) und (iii) folgt $f_X(x) \ge 0$ und $\int_{-\infty}^{\infty} f_X(x) dx = 1$.

Zu beachten ist, dass eine Dichtefunktion keine Wahrscheinlichkeit beschreibt und daher im Einzelfall Werte x mit $f_X(x) > 1$ besitzen kann.

Den Zusammenhang zwischen der Dichtefunktion f_X und der Verteilungsfunktion F_X beschreibt

$$P(X \le x) = F_X(x) = \int_{-\infty}^{x} f_X(t)dt \text{ und } P(a < X \le b) = F_X(b) - F_X(a) = \int_{a}^{b} f_X(t)dt .$$

Wegen $P(X = a) = \int_{a}^{a} f_X(t)dt = 0$ ist

$$P(a < X < b) = P(a \le X < b) = P(a \le X \le b) = P(a < X \le b) = F_X(b) - F_X(a).$$

Für eine ***diskrete Zufallsvariable*** X mit den Werten x_i und der Massenfunktion $f_X(x)$ definiert man den **Erwartungswert** von X durch

$$\mathrm{E}[X] = \sum_{\text{alle } i} x_i \cdot P(X = x_i) = \sum_{\text{alle } i} x_i \cdot f_X(x_i).$$

Für eine ***stetige Zufallsvariable*** X mit der Dichtefunktion $f_X(x)$ ist der Erwartungswert definiert durch

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} (x \cdot f_X(x))dx .$$

Natürlich müssen Summe bzw. Integral existieren.

Den **Erwartungswert einer Funktion** $g(X)$ **einer Zufallsvariablen** X definiert man entsprechend durch

$\mathrm{E}[g(X)] = \sum_{\text{alle } i} g(x_i) \cdot P(X = x_i) = \sum_{\text{alle } i} g(x_i) \cdot f_X(x_i)$, falls X ***diskret*** mit den Werten x_i und der Massenfunktion $f_X(x)$ ist,

bzw. durch

$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} (g(x) \cdot f_X(x))dx$, falls X ***stetig*** mit der Dichtefunktion $f_X(x)$ ist.

Unterwirft man eine Zufallsvariable X einer linearen Transformation $a + b \cdot X$, so erhält man:

Satz 7.2-2:

Die Zufallsvariable X habe den Erwartungswert $\mathrm{E}[X]$. Es seien $a \in \mathbf{R}$ und $b \in \mathbf{R}$. Dann gilt:

$$\mathrm{E}[a + b \cdot X] = a + b \cdot \mathrm{E}[X].$$

Die Herleitung sei hier nur für den stetigen Fall dargestellt:

$$\begin{aligned}
\mathrm{E}[a + b \cdot X] &= \int_{-\infty}^{\infty} ((a + b \cdot x) \cdot f_X(x)) dx \\
&= \int_{-\infty}^{\infty} (a \cdot f_X(x)) dx + \int_{-\infty}^{\infty} (b \cdot x \cdot f_X(x)) dx \\
&= a \cdot \int_{-\infty}^{\infty} f_X(x) dx + b \cdot \int_{-\infty}^{\infty} (x \cdot f_X(x)) dx \\
&= a + b \cdot \mathrm{E}[X] .
\end{aligned}$$

Die **Varianz** einer Zufallsvariablen X wird durch $\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$ definiert. Auch hier müssen natürlich die entsprechenden Summen bzw. Integrale existieren. Die **Standardabweichung** ist $+\sqrt{\mathrm{Var}(X)}$.

Satz 7.2-3:

Die Zufallsvariable X habe den Erwartungswert $\mathrm{E}[X]$ und die Varianz $\mathrm{Var}(X)$. Dann gilt:

(i) $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$.

(ii) $\mathrm{Var}(a + b \cdot X) = b^2 \cdot \mathrm{Var}(X)$ mit $a \in \mathbf{R}$ und $b \in \mathbf{R}$.

(i) rechnet man nach:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2 - 2 \cdot X \cdot \mathrm{E}[X] + (\mathrm{E}[X])^2\right] \\
&= \mathrm{E}[X^2] - 2 \cdot \mathrm{E}[X] \cdot \mathrm{E}[X] + (\mathrm{E}[X])^2 \\
&= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 .
\end{aligned}$$

(ii) rechnet man in drei Schritten nach:
$\mathrm{Var}(a+X) = \mathrm{E}\left[\left((a+X)-\mathrm{E}[a+X]\right)^2\right] = \mathrm{E}\left[\left((a+X)-a-\mathrm{E}[X]\right)^2\right] = \mathrm{E}\left[\left(X-\mathrm{E}[X]\right)^2\right] = \mathrm{Var}(X)$,
$\mathrm{Var}(b\cdot X) = \mathrm{E}\left[\left(b\cdot X-\mathrm{E}[b\cdot X]\right)^2\right] = \mathrm{E}\left[\left(b\cdot\left(X-\mathrm{E}[X]\right)\right)^2\right] = b^2\cdot\mathrm{E}\left[\left(X-\mathrm{E}[X]\right)^2\right] = b^2\cdot\mathrm{Var}(X)$ und
$\mathrm{Var}(a+b\cdot X) = \mathrm{Var}(b\cdot X) = b^2\cdot\mathrm{Var}(X)$.

Der folgende Satz liefert zwei „nützliche“ Abschätzungen:

Satz 7.2-4:

(i) Für die Zufallsvariable X gelte $X(e)\geq 0$ für alle $e\in\Omega$, und sie habe den Erwartungswert $\mathrm{E}[X]$. Dann gilt für $t\in\mathbf{R}$ mit $t>0$ (**Markoffsche Ungleichung**):
$P(X\geq t)\leq \mathrm{E}[X]/t$.

(ii) Die Zufallsvariable X habe den Erwartungswert $\mathrm{E}[X]$ und die Varianz $\mathrm{Var}(X)$. Dann gilt (**Tschebyscheffsche Ungleichung**):
$P\left(\left|X-\mathrm{E}[X]\right|\geq t\right)\leq \mathrm{Var}(X)/t^2$.

(iii) Die Zufallsvariable X habe den Erwartungswert $\mathrm{E}[X]$ und die Varianz $\mathrm{Var}(X)$. Dann gilt:
$P\left(\left|X-\mathrm{E}[X]\right|\geq \sqrt{\mathrm{Var}(X)}\cdot t\right)\leq 1/t^2$ und $P\left(\left|X-\mathrm{E}[X]\right|<\sqrt{\mathrm{Var}(X)}\cdot t\right)\geq 1-1/t^2$.

(i) ergibt sich wie folgt: Die Zufallsvariable $Y:\Omega\to\mathbf{R}$ sei auf demselben Ereignisraum Ω wie X durch $Y(e)=\begin{cases} t & \text{für } X(e)\geq t \\ 0 & \text{sonst}\end{cases}$ definiert. Dann ist $Y(e)\leq X(e)$ für jedes $e\in\Omega$ und damit $\mathrm{E}[Y]\leq\mathrm{E}[X]$. Es ist $\mathrm{E}[Y]=0\cdot P(X<t)+t\cdot P(X\geq t)\leq \mathrm{E}[X]$.

(ii) wird hier nur für eine stetige Zufallsvariable mit Dichtefunktion f_X gezeigt; für eine diskrete Zufallsvariable verläuft der Nachweis entsprechend. Es ist

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left[\left(X-\mathrm{E}[X]\right)^2\right] \\ &= \int_{-\infty}^{\infty}\left(x-\mathrm{E}[X]\right)^2\cdot f_X(x)\,dx \\ &= \int_{-\infty}^{\mathrm{E}[X]-t}\left(x-\mathrm{E}[X]\right)^2\cdot f_X(x)\,dx + \int_{\mathrm{E}[X]-t}^{\mathrm{E}[X]+t}\left(x-\mathrm{E}[X]\right)^2\cdot f_X(x)\,dx + \int_{\mathrm{E}[X]+t}^{\infty}\left(x-\mathrm{E}[X]\right)^2\cdot f_X(x)\,dx\end{aligned}$$

$$
\begin{aligned}
&\geq \int_{-\infty}^{\mathrm{E}[X]-t} (x-\mathrm{E}[X])^2 \cdot f_X(x)dx + \int_{\mathrm{E}[X]+t}^{\infty} (x-\mathrm{E}[X])^2 \cdot f_X(x)dx \\
&\geq \int_{-\infty}^{\mathrm{E}[X]-t} t^2 \cdot f_X(x)dx + \int_{\mathrm{E}[X]+t}^{\infty} t^2 \cdot f_X(x)dx \qquad \text{Ungleichung (*)} \\
&= t^2 \cdot \left(P(X \leq \mathrm{E}[X]-t) + P(X \geq \mathrm{E}[X]+t)\right) \\
&= t^2 \cdot P\left(|X-\mathrm{E}[X]| \geq t\right).
\end{aligned}
$$

Ungleichung (*) gilt, da im ersten Integral $-\infty < x \leq \mathrm{E}[X]-t$ und damit $x-\mathrm{E}[X] \leq -t$ und $(x-\mathrm{E}[X])^2 \geq t^2$ ist; im zweiten Integral gilt $\mathrm{E}[X]+t \leq x < \infty$ und damit $x-\mathrm{E}[X] \geq t$ und $(x-\mathrm{E}[X])^2 \geq t^2$.

(iii) ist eine Umformulierung der Tschebyscheffschen Ungleichung.

In vielen Anwendungsfällen ist eine Zufallsvariable Y als Funktion einer Zufallsvariablen X definiert, d.h. $Y = g(X)$, wobei die Verteilung (Massenfunktion bzw. Dichtefunktion bzw. Verteilungsfunktion) von X bekannt oder vorgegeben ist. Der folgende Satz beschreibt den Sachverhalt für zwei wichtige Beispiele.

Satz 7.2-5:

Die Zufallsvariabel X habe die Verteilungsfunktion $F_X(x) = P(X \le x)$.

(i) Die Zufallsvariable $Y = a \cdot X + b$ mit Konstanten $a \in \mathbf{R}$, $a \neq 0$, und $b \in \mathbf{R}$ hat die Verteilungsfunktion

$$F_Y(y) = P(Y \le y) = F_{a \cdot X + b}(y) = \begin{cases} F_X\left(\frac{y-b}{a}\right) & \text{für } a > 0 \\ 1 - F_X\left(\frac{y-b}{a}\right) & \text{für } a < 0 \end{cases} .$$

Ist X eine stetige Zufallsvariable mit Dichtefunktion $f_X(x)$, so besitzt Y die Dichtefunktion $f_Y(y) = \frac{1}{|a|} \cdot f_X\left(\frac{y-b}{a}\right)$.

(ii) Die Zufallsvariable X sei stetig mit Dichtefunktion $f_X(x)$. Dann hat die Zufallsvariable $Y = X^2$ die Verteilungsfunktion

$$F_Y(y) = P(Y \le y) = F_{X^2}(y) = \begin{cases} 0 & \text{für } y \le 0 \\ F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right) & \text{für } y > 0 \end{cases}$$

und die Dichtefunktion

$$f_Y(y) = \begin{cases} 0 & \text{für } y \le 0 \\ \dfrac{f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)}{2 \cdot \sqrt{y}} & \text{für } y > 0 \end{cases} .$$

In (i) ist für $a > 0$ $F_Y(y) = P(a \cdot X + b \le y) = P\left(X \le \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$. Für $a < 0$ ist $F_Y(y) = P(a \cdot X + b \le y) = P\left(X \ge \frac{y-b}{a}\right) = 1 - P\left(X < \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right)$. Die Aussage über die Dichtefunktion erhält man durch die Bildung der Ableitung:

$f_Y(y) = F_Y'(y) = \frac{1}{a} \cdot F_X'\left(\frac{y-b}{a}\right) = \frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right)$ für $a > 0$ und

$f_Y(y) = F_Y'(y) = -\frac{1}{a} \cdot F_X'\left(\frac{y-b}{a}\right) = -\frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right)$ für $a < 0$. Beide Fälle zusammengefasst ergeben $f_Y(y) = \frac{1}{|a|} \cdot f_X\left(\frac{y-b}{a}\right)$.

Die Aussage in (ii) sieht man wie folgt:

$$F_Y(y) = P(X^2 \le y) = \begin{cases} 0 & \text{für } y \le 0 \\ P\left(-\sqrt{y} \le X \le \sqrt{y}\right) = F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right) & \text{für } y > 0 \end{cases} ,$$

$$f_Y(y)=\begin{cases}0 & \text{für } y\le 0\\ \dfrac{F_X'\left(\sqrt{y}\right)+F_X'\left(-\sqrt{y}\right)}{2\cdot\sqrt{y}}=\dfrac{f_X\left(\sqrt{y}\right)+f_X\left(-\sqrt{y}\right)}{2\cdot\sqrt{y}} & \text{für } y>0\end{cases}.$$

7.3 Mehrdimensionale Zufallsvariablen

Es sei $\mathbf{E}(\Omega)$ eine Ereignismenge mit dem Ereignisraum Ω. Eine Abbildung $(X_1,...,X_n):\Omega\to\mathbf{R}^n$ mit Zufallsvariablen $X_1,...,X_n$ heißt ***n*-dimensionale Zufallsvariable**. Im vorliegenden Kapitel wird $n=2$ genommen; alle Ergebnisse lassen sich aber auch auf höhere Dimensionen übertragen. Zur Schreibvereinfachung wird für (X_1,X_2) der Zufallsvektor (X,Y) geschrieben.

Sind die Komponenten *X* und *Y* einer zweidimensionalen Zufallsvariablen ***diskrete Zufallsvariablen***, so heißt die Funktion

$$f_{(X,Y)}:\mathbf{R}^2\to[0,1] \text{ mit}$$
$$f_{(X,Y)}(x,y)=P(X=x \text{ und } Y=y)$$
$$=P\left(\{e\mid e\in\Omega \text{ und } X(e)=x\}\cap\{e\mid e\in\Omega \text{ und } Y(e)=y\}\right)$$

gemeinsame (Wahrscheinlichkeits-) Massenfunktion von *X* und *Y*.

Nimmt *X* die (endlich oder abzählbar unendlich vielen) Werte x_i und *Y* die (endlich oder abzählbar unendlich vielen) Werte y_j an, so ist analog zum eindimensionalen Fall $0\le f_{(X,Y)}(x,y)\le 1$ für jedes $x\in\mathbf{R}$ und jedes $y\in\mathbf{R}$ und $\sum_{\text{alle } i}\sum_{\text{alle } j} f_{(X,Y)}(x_i,y_j)=1$.

Die Komponenten *X* und *Y* einer zweidimensionalen Zufallsvariablen seien ***stetig***. Es seien $a\in\mathbf{R}$, $b\in\mathbf{R}$, $c\in\mathbf{R}$ und $d\in\mathbf{R}$ mit $a<b$ und $c<d$. Die Funktion $f_{(X,Y)}:\mathbf{R}^2\to\mathbf{R}_{\ge 0}$ mit

$$\int_a^b\int_c^d f_{(X,Y)}(x,y)\,dy\,dx=P(a<X\le b \text{ und } c<Y\le d)$$
$$=P\left(\{e\mid e\in\Omega \text{ und } a<X(e)\le b\}\cap\{e\mid e\in\Omega \text{ und } c<Y(e)\le d\}\right)$$

heißt **gemeinsame (Wahrscheinlichkeits-) Dichtefunktion** von *X* und *Y*.

Offensichtlich gilt $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\,dy\,dx = 1$.

Die **gemeinsame Verteilungsfunktion** von X und Y wird definiert durch

$$F_{(X,Y)}(x,y) = P(X \le x \text{ und } Y \le y)$$
$$= P\left(\left\{ e \,\middle|\, e \in \Omega \text{ und } X(e) \le x \right\} \cap \left\{ e \,\middle|\, e \in \Omega \text{ und } Y(e) \le y \right\}\right) .$$

Im ***diskreten Fall*** ist $F_{(X,Y)}(x,y) = \sum_{x_i \le x} \sum_{y_j \le y} f_{(X,Y)}(x_i, y_j)$.

Im ***stetigen Fall*** ist $F_{(X,Y)}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{(X,Y)}(u,v)\,dv\,du$.

Der **Erwartungswert einer Funktion** g der zweidimensionalen Zufallsvariablen (X,Y) wird

im ***diskreten Fall*** durch $\mathrm{E}[g(X,Y)] = \sum_{\text{alle } x_i} \sum_{\text{alle } y_j} g(x_i, y_j) \cdot f_{(X,Y)}(x_i, y_j)$

und ***im stetigen Fall*** durch $\mathrm{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) \cdot f_{(X,Y)}(x,y)\,dx\,dy$ definiert.

Die **Massenfunktion der Randverteilung** von X bzw. von Y wird ***im diskreten Fall*** mit Hilfe der gemeinsamen Massenfunktion $f_{(X,Y)}$ definiert durch

$$f_X(x) = P(X = x) = P\left(\left\{ e \,\middle|\, e \in \Omega \text{ und } X(e) = x \right\}\right) = \sum_{\text{alle } j} f_{(X,Y)}(x, y_j)$$

bzw.

$$f_Y(y) = P(Y = y) = P\left(\left\{ e \,\middle|\, e \in \Omega \text{ und } Y(e) = y \right\}\right) = \sum_{\text{alle } i} f_{(X,Y)}(x_i, y) .$$

Die **Dichtefunktion der Randverteilung** von X bzw. von Y wird ***im stetigen Fall*** ebenfalls mit Hilfe der gemeinsamen Dichtefunktion $f_{(X,Y)}$ definiert durch

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\,dy$$

bzw.

$$f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\,dx .$$

Die Randverteilungen lassen sich aus der gemeinsamen Massen- bzw. Dichtefunktion $f_{(X,Y)}$ ermitteln. Umgekehrt legen die Randverteilungen eine gemeinsame Verteilung *nicht* fest.

Satz 7.3-1:

Für die ***diskreten*** Zufallsvariablen X bzw. Y sei $\mu_X = \sum_{\text{alle } x_i} x_i \cdot f_X(x_i)$ bzw. $\mu_Y = \sum_{\text{alle } y_j} y_j \cdot f_Y(y_j)$. Dann ist $\mathrm{E}[X] = \sum_{\text{alle } x_i} \sum_{\text{alle } y_j} x_i \cdot f_{(X,Y)}(x_i, y_j) = \mu_X$ bzw.

$\mathrm{E}[Y] = \sum_{\text{alle } y_j} \sum_{\text{alle } x_i} y_j \cdot f_{(X,Y)}(x_i, y_j) = \mu_Y$.

Für die ***stetigen*** Zufallsvariablen X bzw. Y sei $\mu_X = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$ bzw. $\mu_Y = \int_{-\infty}^{\infty} y \cdot f_Y(y)dy$. Dann ist $\mathrm{E}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_{(X,Y)}(x, y)dxdy = \mu_X$ bzw.

$\mathrm{E}[Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \cdot f_{(X,Y)}(x, y)dxdy = \mu_Y$.

Die Erwartungswerte werden also über die Erwartungswerte der Randverteilungen berechnet.

Für die Varianzen gelten entsprechende Aussagen.

Für den stetigen Fall sieht man die Aussage wie folgt (im diskreten Fall verläuft die Argumentation genauso):

Mit $g(X,Y) = X$ ist

$$\begin{aligned}\mathrm{E}[g(X,Y)] = \mathrm{E}[X] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_{(X,Y)}(x, y)dxdy \\ &= \int_{-\infty}^{\infty} x \cdot \left(\int_{-\infty}^{\infty} f_{(X,Y)}(x, y)dy \right) dx \\ &= \int_{-\infty}^{\infty} x \cdot f_X(x)dx \\ &= \mu_X \quad .\end{aligned}$$

In Kapitel 7.1 wird die bedingte Wahrscheinlichkeit eines Ereignisses A unter der Bedingung B (bedingte Wahrscheinlichkeit von A, wenn B eingetroffen ist) durch $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ definiert. Dieses Konzept lässt sich auf bedingte Verteilungen übertragen:

Die **Massenfunktion** f_1 **der bedingten Verteilung** von X **unter der Bedingung** $Y = y_j$ (bedingte Verteilung von X, wenn $Y = y_j$ gilt) wird ***im diskreten Fall*** über die gemeinsame Dichtefunktion $f_{(X,Y)}$ und die Randverteilung f_Y von Y definiert durch

$$f_1(x \mid y_j) = \frac{f_{(X,Y)}(x, y_j)}{f_Y(y_j)}.$$

Entsprechend ist die **Massenfunktion** f_2 **der bedingten Verteilung** von Y **unter der Bedingung** $X = x_i$ (bedingte Verteilung von Y, wenn $X = x_i$ gilt) ***im diskreten Fall*** über die gemeinsame Dichtefunktion $f_{(X,Y)}$ und die Randverteilung f_X von X definiert durch

$$f_2(y \mid x_i) = \frac{f_{(X,Y)}(x_i, y)}{f_X(x_i)}.$$

Im ***stetigen Fall*** lauten die Definitionen:

Die **Dichtefunktion der bedingten Verteilung** von X **für festes** y der Zufallsvariablen Y wird durch die Dichtefunktion

$$f_1(x \mid y) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)}$$

definiert.

Die **Dichtefunktion der bedingten Verteilung** von Y für **festes** x der Zufallsvariablen X wird durch die Dichtefunktion

$$f_2(y \mid x) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

definiert.

Aus diesen Definitionen lassen sich bedingte Erwartungswerte ableiten:

Im diskreten Fall ist der **bedingte Erwartungswert** von X **für einen festen Wert** $Y = y_j$ gleich

$$\mathrm{E}[X \mid Y = y_j] = \sum_{\text{alle } i} x_i \cdot f_1(x_i \mid y_j) = \sum_{\text{alle } i} x_i \cdot \frac{f_{(X,Y)}(x_i, y_j)}{f_Y(y_j)}.$$

Entsprechend ist der **bedingte Erwartungswert** von Y **für einen festen Wert** $X = x_i$ gleich

$$\mathrm{E}[Y \mid X = x_i] = \sum_{\text{alle } j} y_j \cdot f_2(y_j \mid x_i) = \sum_{\text{alle } j} y_j \cdot \frac{f_{(X,Y)}(x_i, y_j)}{f_X(x_i)}.$$

Im stetigen Fall ist der **bedingte Erwartungswert** von X **für einen festen Wert** y der Zufallsvariablen Y gleich

$$\mathrm{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \cdot f_1(x \mid y)\,dx = \int_{-\infty}^{\infty} x \cdot \frac{f_{(X,Y)}(x, y)}{f_Y(y)}\,dx.$$

Entsprechend ist der **bedingte Erwartungswert** von Y **für einen festen Wert** x der Zufallsvariablen X gleich

$$\mathrm{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y \cdot f_2(y \mid x)\,dy = \int_{-\infty}^{\infty} y \cdot \frac{f_{(X,Y)}(x, y)}{f_X(x)}\,dy.$$

Zu beachten ist, dass bedingte Erwartungswerte wieder Zufallsvariablen darstellen: $\mathrm{E}[X \mid Y = y]$ ist eine Funktion von y, d.h. der Zufallsvariablen Y; $\mathrm{E}[Y \mid X = x]$ ist eine Funktion von x, d.h. der Zufallsvariablen X.

Der folgende Satz wird in der Informatik häufig verwendet, um mittleres Verhalten von Algorithmen abzuschätzen. So lässt sich die mittlere Laufzeit $\mathrm{E}[T]$ eines Algorithmus eventuell aus Erwartungswerten bestimmen, die die Laufzeit beschreiben, wenn eine andere Zufallsvariable Y, die in dem Algorithmus vorkommt, einen festen Wert y annimmt, d.h. aus $\mathrm{E}[T \mid Y = y]$.

Satz 7.3-2:

Es ist $\mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X \mid Y = y]]$.

Im ***diskreten Fall*** ist $\mathrm{E}[X] = \sum_{\text{alle } j} \mathrm{E}[X \mid Y = y_j] \cdot P(Y = y_j)$.

Im ***stetigen Fall*** ist $\mathrm{E}[X] = \int_{-\infty}^{\infty} (\mathrm{E}[X \mid Y = y] \cdot f_Y(y))\,dy$.

Für den stetigen Fall soll die Aussage verifiziert werden (der diskrete Fall ergibt sich auf analoge Weise):

$$\begin{aligned}
\int_{-\infty}^{\infty}\left(\mathrm{E}[X \mid Y=y]\cdot f_Y(y)\right)dy &= \int_{-\infty}^{\infty}\left(\left(\int_{-\infty}^{\infty} x\cdot f_1(x \mid y)dx\right)\cdot f_Y(y)\right)dy \\
&= \int_{-\infty}^{\infty}\left(\left(\int_{-\infty}^{\infty} x\cdot \frac{f_{(X,Y)}(x,y)}{f_Y(y)}dx\right)\cdot f_Y(y)\right)dy \\
&= \int_{-\infty}^{\infty}\left(\left(\int_{-\infty}^{\infty} x\cdot f_{(X,Y)}(x,y)dx\right)\right)dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot f_{(X,Y)}(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot f_{(X,Y)}(x,y)dydx \\
&= \int_{-\infty}^{\infty}\left(x\cdot\left(\int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dy\right)\right)dx \\
&\int_{-\infty}^{\infty} x\cdot f_X(x)dx \\
&= \mathrm{E}[X] \ .
\end{aligned}$$

Die Zufallsvariablen X und Y heißen **(stochastisch) unabhängig**, wenn die gemeinsame Massen- bzw. Dichtefunktion $f_{(X,Y)}$ das Produkt der Randverteilungen ist:
$f_{(X,Y)}(x,y) = f_X(x)\cdot f_Y(y)$.

Im Fall der Unabhängigkeit gilt für die gemeinsame Verteilungsfunktion

$F_{(X,Y)}(x,y) = F_X(x)\cdot F_Y(y)$.

Diese Aussage lässt sich leicht verifizieren (hier nur für den stetigen Fall):

$$
\begin{aligned}
F_{(X,Y)}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{(X,Y)}(u,v)\,dv\,du \\
&= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(u)\cdot f_v(v)\,dv\,du \\
&= \int_{-\infty}^{x} f_X(u)\cdot\left(\int_{-\infty}^{y} f_v(v)\,dv\right)du \\
&= \int_{-\infty}^{x} f_X(u)\cdot F_Y(y)\,du \\
&= F_Y(y)\cdot\int_{-\infty}^{x} f_X(u)\,du \\
&= F_Y(y)\cdot F_X(x) = F_X(x)\cdot F_Y(y)\ .
\end{aligned}
$$

Für die bedingten Verteilungen gilt im Fall der Unabhängigkeit:

$$f_1(x\mid y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} = \frac{f_X(x)\cdot f_Y(y)}{f_Y(y)} = f_X(x)$$

und

$$f_2(y\mid x) = \frac{f_{(X,Y)}(x,y)}{f_X(x)} = \frac{f_X(x)\cdot f_Y(y)}{f_X(x)} = f_Y(y).$$

Die bedingten Verteilungen sind im Fall der Unabhängigkeit also die Randverteilungen.

Der folgende Satz lässt sich bei der Erhebung von Stichproben anwenden, wenn die Stichprobenanzahl selbst eine Zufallsvariable darstellt.

Satz 7.3-3:

Die Zufallsvariable N nehme als Werte nur natürliche Zahlen ungleich 0 an. Sie sei von den Zufallsvariablen X_i für $i \geq 1$ unabhängig und habe den Erwartungswert $\mathrm{E}[N]$. Dann gilt für die durch $Y = \sum_{i=1}^{N} X_i$ definierte Zufallsvariable:

Ist $\sum_{i=1}^{\infty} P(N \geq i)\cdot \mathrm{E}[|X_i|] < \infty$, dann existiert der Erwartungswert von Y, und es ist

$$\mathrm{E}[Y] = \sum_{i=1}^{\infty} P(N \geq i)\cdot \mathrm{E}[X_i].$$

Besitzen alle X_i dieselbe Verteilung mit Erwartungswert $\mathrm{E}[X_i] = \mathrm{E}[X]$, dann ist $\mathrm{E}[Y] = \mathrm{E}[X]\cdot\mathrm{E}[N]$.

Mit Satz 7.3-2 ist (unter Berücksichtigung der Bedingung $\sum_{i=1}^{\infty} P(N \geq i) \cdot \mathrm{E}[|X_i|] < \infty$)

$$\mathrm{E}[Y] = \sum_{n=1}^{\infty} P(N=n) \cdot \mathrm{E}[Y \mid N=n] = \sum_{n=1}^{\infty} P(N=n) \cdot \mathrm{E}\left[\sum_{i=1}^{n} X_i\right]$$

$$= \sum_{i=1}^{\infty} \mathrm{E}[X_i] \cdot \sum_{n=i}^{\infty} P(N=n) = \sum_{i=1}^{\infty} P(N \geq i) \cdot \mathrm{E}[X_i] \,.$$

Ist $\mathrm{E}[X_i] = \mathrm{E}[X]$, dann ist $\mathrm{E}[Y] = \mathrm{E}[X] \cdot \sum_{i=1}^{\infty} P(N \geq i) = \mathrm{E}[X] \cdot \sum_{i=1}^{\infty} i \cdot P(N=i) = \mathrm{E}[X] \cdot \mathrm{E}[N]$.

7.4 Kovarianz und Korrelationskoeffizient

Die **Kovarianz** einer zweidimensionalen Zufallsvariablen (X, Y) wird definiert durch $\mathrm{Kov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y])]$. Sie beschreibt eine Maßzahl für den stochastischen Zusammenhang zwischen den beiden Komponenten.

Sind die Komponenten X und Y einer zweidimensionalen Zufallsvariablen ***diskrete Zufallsvariablen*** mit Werten x_i bzw. y_j und gemeinsamer Massenfunktion

$f_{(X,Y)}(x, y) = P(X = x \text{ und } Y = y)$, so ist

$$\mathrm{Kov}(X, Y) = \sum_{\text{alle } i} \sum_{\text{alle } j} (x_i - \mathrm{E}[X]) \cdot (y_j - \mathrm{E}[Y]) \cdot f_{(X,Y)}(x_i, y_j)$$

$$= \sum_{\text{alle } i} \sum_{\text{alle } j} (x_i - \mathrm{E}[X]) \cdot (y_j - \mathrm{E}[Y]) \cdot P(X = x_i \text{ und } Y = y_j) \,.$$

Sind X und Y ***stetige Zufallsvariablen*** mit gemeinsamer Dichtefunktion $f_{(X,Y)}(x, y)$, so ist

$$\mathrm{Kov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mathrm{E}[X]) \cdot (y - \mathrm{E}[Y]) \cdot f_{(X,Y)}(x, y) \, dx \, dy \,.$$

Offensichtlich ist $\mathrm{Kov}(X, X) = \mathrm{Var}(X)$.

Satz 7.4-1:

Es seien X und Y Zufallsvariablen, für die $\mathrm{E}[X]$ und $\mathrm{E}[Y]$ existieren. Dann gilt:

(i) Sind X und Y stochastisch unabhängig, so ist $\mathrm{Kov}(X, Y) = 0$.

(ii) $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y] + \mathrm{Kov}(X, Y)$.

(iii) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Kov}(X, Y)$,
$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2 \cdot \mathrm{Kov}(X, Y)$.

Die Aussagen sollen hier nur für den stetigen Fall dargestellt werden: Mit Satz 7.2-2 und 7.3-1 ist

$$\begin{aligned}\mathrm{Kov}(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mathrm{E}[X])\cdot(y-\mathrm{E}[Y])\cdot f_{(X,Y)}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mathrm{E}[X])\cdot(y-\mathrm{E}[Y])\cdot f_X(x)\cdot f_Y(y)\,dx\,dy \quad \text{wegen der stochastischen Unabhängigkeit} \\ &= \int_{-\infty}^{\infty}(x-\mathrm{E}[X])\cdot f_X(x)\cdot\left(\int_{-\infty}^{\infty}(y-\mathrm{E}[Y])\cdot f_Y(y)\,dy\right)dx \\ &= 0 \quad ,\end{aligned}$$

denn $\int_{-\infty}^{\infty}(y-\mathrm{E}[Y])\cdot f_Y(y)\,dy = \int_{-\infty}^{\infty} y\cdot f_Y(y)\,dy - \mathrm{E}[Y]\cdot\int_{-\infty}^{\infty} f_Y(y)\,dy = \mathrm{E}[Y]-\mathrm{E}[Y]=0$.

Durch Ausmultiplizieren erhält man

$$\mathrm{Kov}(X,Y) = \mathrm{E}[(X-\mathrm{E}[X])\cdot(Y-\mathrm{E}[Y])] = \mathrm{E}[X\cdot Y]-\mathrm{E}[X]\cdot\mathrm{E}[Y].$$

$$\begin{aligned}\mathrm{Var}(X\pm Y) &= \mathrm{E}\left[((X\pm Y)-\mathrm{E}[X\pm Y])^2\right] \\ &= \mathrm{E}\left[((X-\mathrm{E}[X])\pm(Y-\mathrm{E}[Y]))^2\right] \\ &= \mathrm{E}\left[(X-\mathrm{E}[X])^2 \pm 2\cdot(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])+(Y-\mathrm{E}[Y])^2\right] \\ &= \mathrm{Var}(X)+\mathrm{Var}(Y)\pm 2\cdot\mathrm{Kov}(X,Y)\ .\end{aligned}$$

Existieren die Varianzen der Zufallsvariablen X und Y und sind jeweils von Null verschieden, so kann man ein Maß für ihre stochastische Abhängigkeit durch den **Korrelationskoeffizienten** von X und Y definieren:

$$\rho_{(X,Y)} = \frac{\mathrm{Kov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\cdot\sqrt{\mathrm{Var}(Y)}}\ .$$

Satz 7.4-2:

Es seien X und Y Zufallsvariablen, für die die jeweiligen Varianzen existieren und von Null verschieden sind. Dann gilt:

(i) $-1 \le \rho_{(X,Y)} = \dfrac{\mathrm{Kov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\cdot\sqrt{\mathrm{Var}(Y)}} \le 1$.

(ii) $\rho_{(X,Y)} = \pm 1$ gilt genau dann, wenn $P(Y = a\cdot X + b) = 1$ mit $a \in \mathbf{R}$ und $b \in \mathbf{R}$ ist.

Die Aussage in (ii) lässt folgende Interpretation zu: Gilt $\rho_{(X,Y)} = \pm 1$, dann ist die Wahrscheinlichkeit dafür, dass die Zufallsvariable (X, Y) Werte in der (x, y)-Ebene annimmt, die auf einer Geraden liegen, gleich 1. Dieser Fall stellt also den größten Gegensatz zur stochastischen Unabhängigkeit der Zufallsvariablen *X* und *Y* dar. Ist $\rho_{(X,Y)} = +1$, dann entsprechen größere Werte der einen Zufallsvariablen auch größeren Werten der anderen Zufallsvariablen. Bei $\rho_{(X,Y)} = -1$ entsprechen größeren Werten der einen Zufallsvariablen kleinere Werte der anderen Zufallsvariablen.

Zum Nachweis der beiden Aussagen in Satz 7.4-2 wird die Funktion
$h(s,t) = \mathrm{E}\left[(s \cdot (X - \mathrm{E}[X]) + t \cdot (Y - \mathrm{E}[Y]))^2\right]$ betrachtet.
Es gilt $h(s,t) \geq 0$ und $h(s,t) = s^2 \cdot \mathrm{Var}(X) + 2 \cdot s \cdot t \cdot \mathrm{Kov}(X,Y) + t^2 \cdot \mathrm{Var}(Y)$.
Die Nullstellen von *h* bezüglich *s* lauten wegen $\mathrm{Var}(X) > 0$:

$$s_{01,02} = -t \cdot \frac{\mathrm{Kov}(X,Y)}{\mathrm{Var}(X)} \pm \sqrt{\frac{t^2 \cdot (\mathrm{Kov}(X,Y))^2 - t^2 \cdot \mathrm{Var}(X) \cdot \mathrm{Var}(Y)}{(\mathrm{Var}(X))^2}} \quad .$$

h hat bezüglich *s* bei $t \neq 0$ nur höchstens eine Nullstelle (bei zwei Nullstellen würde *h* zwischen den Nullstellen oder jenseits der Nullstellen negative Werte annehmen). Daher ist
$t^2 \cdot (\mathrm{Kov}(X,Y))^2 - t^2 \cdot \mathrm{Var}(X) \cdot \mathrm{Var}(Y) \leq 0$ (= 0 bedeutet eine Nullstelle, < 0 bedeutet keine Nullstelle). Daraus folgt $(\mathrm{Kov}(X,Y))^2 \leq \mathrm{Var}(X) \cdot \mathrm{Var}(Y)$ bzw.
$-\sqrt{\mathrm{Var}(X)} \cdot \sqrt{\mathrm{Var}(Y)} \leq (\mathrm{Kov}(X,Y)) \leq \sqrt{\mathrm{Var}(X)} \cdot \sqrt{\mathrm{Var}(Y)}$ und
$-1 \leq \rho_{(X,Y)} = \dfrac{\mathrm{Kov}(X,Y)}{\sqrt{\mathrm{Var}(X)} \cdot \sqrt{\mathrm{Var}(Y)}} \leq 1$. Das ist Aussage (i).

Zum Nachweis von Aussage (ii) wird zunächst $\rho_{(X,Y)} = \pm 1$, also $\rho_{(X,Y)}{}^2 = 1$ angenommen. Daher ist $(\mathrm{Kov}(X,Y))^2 - \mathrm{Var}(X) \cdot \mathrm{Var}(Y) = 0$, und die Funktion
$h(s,t) = \mathrm{E}\left[(s \cdot (X - \mathrm{E}[X]) + t \cdot (Y - \mathrm{E}[Y]))^2\right]$ (siehe oben) hat eine Nullstelle (s_0, t_0) für $t_0 \neq 0$, d.h. $h(s_0, t_0) = \mathrm{E}\left[(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]))^2\right] = 0$. Es gilt

$$\begin{aligned}
&\mathrm{E}\left[(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]))^2\right] \\
&= \mathrm{E}\left[(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]))^2 \,|\, s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]) = 0\right] \\
&\qquad \cdot P(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]) = 0) \\
&\quad + \mathrm{E}\left[(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]))^2 \,|\, s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]) \neq 0\right] \\
&\qquad \cdot P(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]) \neq 0) \\
&= \mathrm{E}\left[(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]))^2 \,|\, s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]) \neq 0\right] \\
&\qquad \cdot P(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]) \neq 0) \quad .
\end{aligned}$$

Damit dieser Ausdruck den Wert 0 annimmt, muss

$P(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]) \neq 0) = 0$ bzw. $P(s_0 \cdot (X - \mathrm{E}[X]) + t_0 \cdot (Y - \mathrm{E}[Y]) = 0) = 1$ gelten.

Mit $t_0 \neq 0$ folgt $P\left(Y = -\frac{s_0}{t_0} \cdot X + \frac{\mathrm{E}[Y] \cdot t_0 + \mathrm{E}[X] \cdot s_0}{t_0} \right) = 1$.

Gilt umgekehrt $P(Y = a \cdot X + b) = 1$ mit $a \in \mathbf{R}$ und $b \in \mathbf{R}$, dann ist $P(Y \neq a \cdot X + b) = 0$. Damit ergibt sich

$$\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}[Y \mid Y = a \cdot X + b] \cdot P(Y = a \cdot X + b) + \mathrm{E}[Y \mid Y \neq a \cdot X + b] \cdot P(Y \neq a \cdot X + b) \\
&= \mathrm{E}[a \cdot X + b] = a \cdot \mathrm{E}[X] + b \quad , \\
\mathrm{Var}(Y) &= \mathrm{E}\left[(Y - \mathrm{E}[Y])^2\right] \\
&= \mathrm{E}\left[(Y - \mathrm{E}[Y])^2 \mid Y = a \cdot X + b\right] \cdot P(Y = a \cdot X + b) \\
&\quad + \mathrm{E}\left[(Y - \mathrm{E}[Y])^2 \mid Y \neq a \cdot X + b\right] \cdot P(Y \neq a \cdot X + b) \\
&= \mathrm{E}\left[(a \cdot X + b - \mathrm{E}[Y])^2\right] \\
&= \mathrm{E}\left[a^2 \cdot (X - \mathrm{E}[X])^2\right] \\
&= a^2 \cdot \mathrm{Var}(\mathrm{X})
\end{aligned}$$

und

$$\begin{aligned}
\mathrm{Kov}(X, Y) &= \mathrm{E}[(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y])] \\
&= \mathrm{E}\left[(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y]) \mid Y = a \cdot X + b\right] \cdot P(Y = a \cdot X + b) \\
&\quad + \mathrm{E}\left[(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y]) \mid Y \neq a \cdot X + b\right] \cdot P(Y \neq a \cdot X + b) \\
&= \mathrm{E}[(X - \mathrm{E}[X]) \cdot (a \cdot X + b - \mathrm{E}[Y])] \\
&= \mathrm{E}[(X - \mathrm{E}[X]) \cdot (a \cdot X + b - (a \cdot \mathrm{E}[X] + b))] \\
&= a \cdot \mathrm{Var}(X) \quad .
\end{aligned}$$

Damit folgt $\rho_{(X,Y)}{}^2 = \frac{(\mathrm{Kov}(X, Y))^2}{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)} = 1$ und $\rho_{(X,Y)} = \pm 1$.

7.5 Bemerkungen zu erzeugenden und charakteristischen Funktionen

In Kapitel 5.11 wurde die erzeugende Funktion G einer Folge $(g_n)_{n \in \mathbf{N}}$ durch $G(z) = \sum_{n=0}^{\infty} g_n \cdot z^n$ definiert. Die erzeugende Funktion fasst die Informationen über unendlich viele Folgenglieder g_n mit $n \in \mathbf{N}$ in einem einzigen arithmetischen Ausdruck zusammen. Umgekehrt lassen sich aus der erzeugenden Funktion $G(z) = \sum_{n=0}^{\infty} g_n \cdot z^n$ die einzelnen Folgenglieder durch die Beziehung (siehe Kapitel 5.11) $g_n = \frac{1}{n!} \cdot G^{(n)}(0)$ zurückgewinnen.

Die Idee der erzeugenden Funktion lässt sich in folgendem Sinne auf die Wahrscheinlichkeitsrechnung übertragen: Die Informationen über die möglichen Ausprägungen der Zufallsvariablen X werden in einer einzigen eindeutig bestimmten Funktion zusammengefasst; diese Funktion legt auf eindeutige Weise die Verteilung von X fest.

Es sei X eine diskrete bzw. stetige Zufallsvariable mit Massenfunktion bzw. Dichtefunktion f_X. Dann heißt $mef_X(t) = \mathrm{E}\left[e^{t \cdot X}\right]$ die **momenterzeugende Funktion der Zufallsvariablen** X.

Im ***diskreten*** Fall ist $mef_X(t) = \sum_{\text{alle } i} e^{t \cdot x_i} \cdot f_X(x_i) = \sum_{\text{alle } i} e^{t \cdot x_i} \cdot P(X = x_i)$;

im ***stetigen*** Fall ist $mef_X(t) = \int_{-\infty}^{\infty} \left(e^{t \cdot x} \cdot f_X(x)\right) dx$.

Die momenterzeugende Funktion der Zufallsvariablen muss nicht immer existieren, da Summe bzw. Integral eventuell nicht konvergieren.

Satz 7.5-1:

Es seien X und Y Zufallsvariablen, für die die momenteerzeugenden Funktionen $mef_X(t)$ und $mef_Y(t)$ in einer Umgebung von $t = 0$ existieren. Dann gilt:

(i) X und Y besitzen genau dann dieselbe Verteilungsfunktion, wenn $mef_X = mef_Y$ gilt.

(ii) Existieren die ***k*-ten Momente** $\mathrm{E}\left[X^k\right]$ der Zufallsvariablen X für $k = 1, ..., n$, dann gilt $\mathrm{E}\left[X^k\right] = \left.\frac{d^k}{dt^k} mef_X(t)\right|_{t=0} = mef_X^{(k)}(0)$.

Insbesondere gilt $\mathrm{E}[X] = mef_X'(0)$ und $\mathrm{Var}(X) = mef_X''(0) - \left(mef_X'(0)\right)^2$.

(iii) Sind X und Y stochastisch unabhängig, so gilt $mef_{X+Y}(t) = mef_X(t) \cdot mef_Y(t)$.

Besitzen in (i) X und Y dieselbe Verteilungsfunktion, dann ist $mef_X = mef_Y$. Existieren mef_X und mef_Y und gilt $mef_X = mef_Y$, dann lassen sich nach (ii) die Momente beider Zufallsvariablen berechnen, und diese stimmen überein. In Fisz, M.: **Wahrscheinlichkeitsrechnung und mathematische Statistik**, 11. Aufl., Deutscher Verlag der Wissenschaften, 1989. wird gezeigt, dass dann die Verteilungen von X und Y identisch sind.

Für (ii) wird die *k*-te Ableitung von mef_X nach *t* gebildet (hier nur für den diskreten Fall; der stetige Fall ergibt sich entsprechend). Dabei wird vorausgesetzt, dass diese Ableitungen existiert, d.h. insbesondere dass Summen- und Ableitungsbildung vertauscht werden können.

$$\begin{aligned} mef_X^{(k)}(t) &= \frac{d^k}{dt^k} mef_X(t) \\ &= \sum_{\text{alle } i} \frac{d^k}{dt^k}\left(e^{t\cdot x_i} \cdot P(X = x_i)\right) \\ &= \sum_{\text{alle } i} x_i^k \cdot e^{t\cdot x_i} \cdot P(X = x_i) \\ &= \mathrm{E}\left[X^k \cdot e^{t\cdot X}\right] \end{aligned}$$

und damit $mef_X^{(n)}(0) = \mathrm{E}\left[X^n\right]$.

(iii) folgt aus der Eigenschaft des Erwartungswertoperators bei stochastischer Unabhängigkeit der beteiligten Zufallsvariablen. Dabei ist zu beachten, dass mit der stochastischen Unabhängigkeit von *X* und *Y* auch die Zufallsvariablen $e^{t\cdot X}$ und $e^{t\cdot Y}$ stochastisch unabhängig sind.
$mef_{X+Y}(t) = \mathrm{E}\left[e^{t\cdot(X+Y)}\right] = \mathrm{E}\left[e^{t\cdot X} \cdot e^{t\cdot Y}\right] = \mathrm{E}\left[e^{t\cdot X}\right] \cdot \mathrm{E}\left[e^{t\cdot Y}\right] = mef_X(t) \cdot mef_Y(t)$.

Eine noch bedeutendere Rolle als die momenteerzeugende Funktion spielt die **charakteristische Funktion** cf_X einer Zufallsvariablen *X*. Diese ist definiert durch[8]

$$cf_X(t) = \mathrm{E}\left[e^{i\cdot t\cdot X}\right] = \begin{cases} \sum\limits_{\text{alle } i} e^{i\cdot t\cdot x_i} \cdot P(X = x_i) & \text{falls } X \text{ diskret ist} \\ \int\limits_{-\infty}^{\infty} \left(e^{i\cdot t\cdot x} \cdot f_X(x)\right) dx & \text{falls } X \text{ stetig mit Dichtefunktion } f_X \text{ ist} \end{cases}.$$

Wegen $e^{i\cdot y} = \cos(y) + i\cdot\sin(y)$ (vgl. Kapitel 5.9) ist $\left|e^{i\cdot y}\right| = \sqrt{(\cos(y))^2 + (\sin(y))^2} = 1$. Außerdem gilt im diskreten Fall $\sum\limits_{\text{alle } i} P(X = x_i) = 1$ bzw. im stetigen Fall $\int\limits_{-\infty}^{\infty} f_X(x)dx = 1$, so dass $cf_X(t)$ absolut und gleichmäßig konvergiert[9] und eine stetige Funktion von *t* darstellt, insbesondere immer definiert ist.

Die charakteristische Funktion einer Zufallsvariablen weist eine Reihe interessanter Eigenschaften auf. Einige der wichtigsten Eigenschaften werden im folgenden zitiert. Da zu deren Herleitung meist tiefergehende mathematische Kenntnisse aus der Analysis erforderlich sind, die hier aus Platzgründen nicht behandelt werden können, wird für Details der Herleitungen auf die angegebene Literatur verwiesen.

---

[8] Hierbei ist *i* die durch $i^2 = -1$ definierte komplexe Zahl.

[9] Zu den Begriffen vergleiche man in der angegebenen Literatur die entsprechenden weiterführenden Kapitel aus der Analysis.

Satz 7.5-2:

Es seien X und Y Zufallsvariablen mit charakteristischer Funktion cf_X bzw. cf_Y.

(i) Es gilt $cf_X(0)=1$ und $|cf_X(t)|\le 1$.
$cf_X(t)=\mathrm{E}[\cos(t\cdot X)]+i\cdot\mathrm{E}[\sin(t\cdot X)]$, $cf_X(-t)=\mathrm{E}[\cos(t\cdot X)]-i\cdot\mathrm{E}[\sin(t\cdot X)]$.

(ii) Existiert die ersten n Momente $\mathrm{E}[X^k]$ der Zufallsvariablen X für $k=1,...,n$, dann gilt $\mathrm{E}[X^k]=\frac{1}{i^k}\cdot\frac{d^k}{dt^k}cf_X(t)\Big|_{t=0}=\frac{cf_X^{(k)}(0)}{i^k}$ für $k=1,...,n$.

(iii) Die charakteristische Funktion der Zufallsvariablen $Y=a\cdot X+b$ mit $a\in\mathbf{R}$ und $b\in\mathbf{R}$ lautet $cf_Y(t)=cf_{a\cdot X+b}(t)=e^{i\cdot b\cdot t}\cdot cf_X(a\cdot t)$.

(iv) Sind die Zufallsvariablen X und Y stochastisch unabhängig, so gilt $cf_{X+Y}(t)=cf_X(t)\cdot cf_Y(t)$.

(i) folgt aus der Definition der charakteristischen Funktion: $cf_X(0)=\mathrm{E}[1]=1$ und
$|cf_X(t)|=\left|\mathrm{E}[e^{i\cdot t\cdot X}]\right|\le\mathrm{E}\left[\left|e^{i\cdot t\cdot X}\right|\right]=\mathrm{E}[1]=1$.
$cf_X(t)=\mathrm{E}[e^{i\cdot t\cdot X}]=\mathrm{E}[\cos(t\cdot X)+i\cdot\sin(t\cdot X)]=\mathrm{E}[\cos(t\cdot X)]+i\cdot\mathrm{E}[\sin(t\cdot X)]$,
$cf_X(-t)=\mathrm{E}[e^{-i\cdot t\cdot X}]$
$=\mathrm{E}[\cos(-t\cdot X)+i\cdot\sin(-t\cdot X)]$
$=\mathrm{E}[\cos(t\cdot X)-i\cdot\sin(t\cdot X)]=\mathrm{E}[\cos(t\cdot X)]-i\cdot\mathrm{E}[\sin(t\cdot X)]$.

Bei (ii) benötigt man die Vertauschbarkeit von Summenbildung (im diskreten Fall) bzw. Integralbildung (im stetigen Fall) mit dem Ableitungsoperator. Diese wird durch die Annahme über die Existenz der ersten n Momente gewährleistet.

Zu bemerken ist, dass die Umkehrung dieser Aussage nicht gilt: Es gibt Zufallsvariablen, deren charakteristische Funktion an der Stelle $t=0$ differenzierbar ist, aber deren Erwartungswert (erstes Moment) nicht existiert.

Die charakteristische Funktion in (iii) ergibt sich aus
$cf_{a\cdot X+b}(t)=\mathrm{E}[e^{i\cdot t\cdot(a\cdot X+b)}]=e^{i\cdot b\cdot t}\cdot\mathrm{E}[e^{i\cdot t\cdot a\cdot X}]=e^{i\cdot b\cdot t}\cdot cf_X(a\cdot t)$.

In (iv) ist zu beachten, dass mit der stochastischen Unabhängigkeit von X und Y auch die Zufallsvariablen $e^{i\cdot t\cdot X}$ und $e^{i\cdot t\cdot Y}$ stochastisch unabhängig sind. Dann ist

$$cf_{X+Y}(t) = \mathrm{E}\left[e^{i\cdot t\cdot(X+Y)}\right] = \mathrm{E}\left[e^{i\cdot t\cdot X}\cdot e^{i\cdot t\cdot Y}\right] = \mathrm{E}\left[e^{i\cdot t\cdot X}\right]\cdot \mathrm{E}\left[e^{i\cdot t\cdot Y}\right] = cf_X(t)\cdot cf_Y(t).$$

Die charakteristische Funktion ist aus der Verteilungsfunktion einer Zufallsvariable berechenbar. Der folgende Satz beschreibt die Umkehrung dieses Sachverhalts: Durch die charakteristische Funktion ist die Verteilungsfunktion einer Zufallsvariablen eindeutig bestimmt.

Satz 7.5-3:

Es seien F_X bzw. cf_X die Verteilungsfunktion bzw. charakteristische Funktion einer Zufallsvariablen *X*. Es seien $a+h$ und $a-h$ für $h>0$ Stetigkeitsstellen der Verteilungsfunktion F_X. Dann gilt:

$$F_X(a+h) - F_X(a-h) = \lim_{T\to\infty} \frac{1}{\pi}\cdot \int_{-T}^{T}\left(\frac{\sin(h\cdot t)}{t}\cdot e^{-i\cdot t\cdot a}\cdot cf_X(t)\right)dt\,.$$

Es seien x_1 und x_2 Stetigkeitsstellen der Verteilungsfunktion F_X. Dann lässt sich aus der Formel in Satz 7.5-3 die Differenz $F_X(x_2) - F_X(x_1) = P(x_1 \le X < x_2)$ berechnen, indem man $a = (x_1 + x_2)/2$ und $h = (x_2 - x_1)/2$) setzt. Hält man $x_2 = x$ fest und vollzieht den Grenzübergang $x_1 \to -\infty$, so konvergiert die Folge der Differenzen $F_X(x) - F_X(x_1)$, die aus der charakteristischen Funktion bestimmt werden kann, gegen $F_X(x)$. Damit ist $F_X(x)$ in jeder Stetigkeitsstelle und damit überall bestimmt.

Ist die charakteristische Funktion cf_X der Zufallsvariablen *X* überall absolut integrierbar, d.h. existiert das Integral $\int_{-\infty}^{\infty} |cf_X(t)|dt$, dann existiert das in Satz 7.5-3 angegebene Integral. Weiter gilt dann $\frac{F_X(x+h) - F_X(x-h)}{2\cdot h} = \frac{1}{2\cdot\pi}\cdot \int_{-\infty}^{\infty}\left(\frac{\sin(h\cdot t)}{h\cdot t}\cdot e^{-i\cdot t\cdot x}\cdot cf_X(t)\right)dt$. Man kann jetzt den Grenzübergang $h\to 0$ auf der linken Seite vollziehen und erhält die Dichtefunktion der Zufallsvariablen *X*. Auf der rechten Seite kann man Integralbildung und Grenzübergang vertauschen. In Kapitel 5.9 wird $\lim_{h\to 0} \sin(h\cdot t)/h\cdot t = 1$ gezeigt, und damit folgt

Satz 7.5-4:

Es sei cf_X die charakteristische Funktion einer Zufallsvariablen *X*. Das Integral $\int_{-\infty}^{\infty} |cf_X(t)| dt$ existiere. Dann besitzt *X* die stetige Dichtefunktion $f_X(x)$, die sich in der Form

$$F_X'(x) = f_X(x) = \frac{1}{2 \cdot \pi} \cdot \int_{-\infty}^{\infty} \left(e^{-i \cdot t \cdot a} \cdot cf_X(t) \right) dt$$

darstellen lässt.

Der folgende Satz stellt ein zentrales Hilfsmittel bei der Untersuchung von Folgen von Verteilungen und Konvergenz gegen eine Grenzverteilung dar.

Satz 7.5-5:

(i) Konvergiert die Folge $(F_n(x))_{n \in \mathbf{N}}$ von Verteilungsfunktionen gegen die Verteilungsfunktion $F(x)$, so konvergiert die entsprechende Folge $(cf_n(t))_{n \in \mathbf{N}}$ von charakteristischen Funktionen an jeder Stelle *t* mit $-\infty < t < \infty$ gegen die Funktion $cf(t)$, wobei $cf(t)$ die charakteristische Funktion der Grenzverteilung $F(x)$ ist.

(ii) Konvergiert die Folge $(cf_n(t))_{n \in \mathbf{N}}$ von charakteristischen Funktionen an jeder Stelle *t* mit $-\infty < t < \infty$ gegen die Funktion $cf(t)$, die in einem gewissen Intervall $]-\tau, \tau[$ stetig ist, so konvergiert die entsprechende Folge $(F_n(x))_{n \in \mathbf{N}}$ von Verteilungsfunktionen gegen die Verteilungsfunktion $F(x)$, die die charakteristische Funktion $cf(t)$ besitzt.

In (i) ist es wichtig, dass $F(x)$ eine Verteilungsfunktion ist; andernfalls kann die Konvergenz von $(cf_n(t))_{n \in \mathbf{N}}$ nicht gesichert werden.

In (ii) ist die Stetigkeit in einem gewissen Intervall $]-\tau, \tau[$ notwendig; die Stetigkeit im Punkt $t = 0$ genügt nicht, um zu sichern, dass $F(x)$ eine Verteilungsfunktion ist.

7.6 Beispiele von Verteilungen

In diesem Kapitel werden Beispiele für wichtige Verteilungen beschrieben. Dabei wird im Einzelnen angegeben, welcher Sachverhalt durch eine Zufallsvariable modelliert wird. Für die angegebenen Massen- bzw. Dichtefunktionen der Verteilungen wird gezeigt, dass sie den Anforderungen

$0 \le f_X(x) \le 1$ und $\sum_{\text{alle } i} f_X(x_i) = 1$ (bei einer Massenfunktion f_X) bzw.

$0 \le f_X(x)$ und $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (bei einer Dichtefunktion f_X)

erfüllen. Außerdem werden wichtige Parameter wie Erwartungswert und Varianz und die charakteristische Funktion berechnet.

Tabellen zu den jeweiligen Verteilungsfunktionen findet man in der angegebenen Literatur.

Diskrete Gleichverteilung

Modell:
Die Zufallsvariable X nimmt endlich viele Werte $x_1, ..., x_n$ mit $x_1 < ... < x_n$ jeweils mit gleicher Wahrscheinlichkeit $p = 1/n$ an.

Massenfunktion:

$$f_X(x) = P(X = x) = \begin{cases} P(X = x_i) = 1/n & \text{für } x = x_i, i = 1, ..., n \\ 0 & \text{sonst} \end{cases}$$

Wegen $0 \le f_X(x) \le 1$ und $\sum_{\text{alle } i} f_X(x_i) = n \cdot 1/n = 1$ handelt es sich offensichtlich um eine Massenfunktion.

Verteilungsfunktion:

$$F_X(x) = P(X \le x) = \begin{cases} 0 & \text{für } x < x_1 \\ i/n & \text{für } x_i \le x < x_{i+1}, i = 1, ..., n-1 \\ 1 & \text{für } x \ge x_n \end{cases}$$

Ein häufig auftretender Spezialfall liegt vor, wenn X die Werte 1, …, n annimmt.

Massenfunktion:

$$f_X(x) = P(X = x) = \begin{cases} P(X = i) = 1/n & \text{für } i = 1, ..., n \\ 0 & \text{sonst} \end{cases}$$

Verteilungsfunktion:

$$F_X(x) = P(X \le x) = \begin{cases} 0 & \text{für } x < 1 \\ i/n & \text{für } i \le x < i+1, i = 1, ..., n-1 \\ 1 & \text{für } x \ge n \end{cases}$$

Erwartungswert: $\mathrm{E}[X] = \frac{n+1}{2}$ ***Varianz:*** $\mathrm{Var}(X) = \frac{n^2-1}{12}$

Diese Werte ergeben sich wie folgt:

$\mathrm{E}[X] = \frac{1}{n} \cdot \sum_{i=1}^{n} i = \frac{n+1}{2}$ (mit Satz 1.6-2(i)).

$$\begin{aligned} \mathrm{Var}(X) &= \frac{1}{n} \cdot \sum_{i=1}^{n} i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^{n} i \right)^2 && \text{(mit Satz 7.2 - 3(i))} \\ &= \frac{(n+1) \cdot (2 \cdot n + 1)}{6} - \frac{(n+1)^2}{4} && \text{(mit Satz 1.6 - 2(i), (iv))} \\ &= \frac{n^2 - 1}{12} \quad . \end{aligned}$$

charakteristische Funktion:

$$cf_X(t) = \frac{1}{n} \cdot \sum_{j=1}^{n} e^{i \cdot t \cdot j}$$

Momente:

$$\mathrm{E}[X^k] = \frac{cf_X^{(k)}(0)}{i^k} = \frac{1}{n} \cdot \sum_{j=1}^{n} j^k \qquad \text{(mit Satz 7.5-2(ii))}$$

Stetige Gleichverteilung

Modell:

Die Dichtefunktion f_X der Zufallsvariablen X nimmt auf dem endlichen Intervall $[a, b]$ mit $a < b$ einen konstanten Wert $c > 0$ an; außerhalb dieses Intervalls ist sie konstant 0. Um $\int_{-\infty}^{\infty} f_X(t)dt = 1$ zu erfüllen, gilt $1 = \int_{-\infty}^{\infty} f_X(t)dt = \int_a^b c\,dt = c \cdot (b-a)$ und $c = 1/(b-a)$.

Dichtefunktion:

$$f_X(x) = \begin{cases} 1/(b-a) & \text{für } a \le x \le b \\ 0 & \text{sonst} \end{cases}$$

Verteilungsfunktion:

$$F_X(x) = P(X \le x) = \begin{cases} 0 & \text{für } x < a \\ \dfrac{x-a}{b-a} & \text{für } a \le x \le b \\ 1 & \text{für } x > b \end{cases}$$

Erwartungswert: $\mathrm{E}[X] = \dfrac{a+b}{2}$ ***Varianz:*** $\mathrm{Var}(X) = \dfrac{(b-a)^2}{12}$

Die Werte ergeben sich wie folgt:

$$\mathrm{E}[X] = \int_a^b \left(x \cdot \frac{1}{b-a} \right) dx = \frac{1}{b-a} \cdot \left(\frac{x^2}{2} \Big|_{x=a}^{x=b} \right) = \frac{a+b}{2}.$$

$$\begin{aligned} \mathrm{Var}(X) &= \int_a^b \left(x^2 \cdot \frac{1}{b-a} \right) dx - \left(\frac{a+b}{2} \right)^2 \qquad \text{(mit Satz 7.2 - 3(i))} \\ &= \frac{1}{(b-a)\cdot 3} \cdot \left(b^3 - a^3 \right) - \frac{(a+b)^2}{4} \\ &= \frac{4 \cdot \left(b^3 - a^3 \right)}{12 \cdot (b-a)} - \frac{3 \cdot (b-a) \cdot (a+b)^2}{12 \cdot (b-a)} \\ &= \frac{b^3 - 3 \cdot a \cdot b^2 + 3 \cdot a^2 \cdot b - a^3}{12 \cdot (b-a)} \\ &= \frac{b^3 - 3 \cdot a \cdot b^2 + 3 \cdot a^2 \cdot b - a^3}{12 \cdot (b-a)} \\ &= \frac{(b-a)^2}{12} \qquad \text{(siehe Kapitel 4.1).} \end{aligned}$$

charakteristische Funktion:

$$cf_X(t) = e^{i \cdot t \cdot m} \cdot \frac{\sin(t \cdot h)}{t \cdot h}$$

Zur Berechnung wird $m = \dfrac{a+b}{2}$ gesetzt, d.h. m ist der Mittelpunkt des Intervalls $[a, b]$. Außerdem sei $h = \dfrac{b-a}{2}$ die halbe Länge des Intervalls. Dann ist $[a, b] = [m-h, m+h]$. Damit ergibt sich

$$
\begin{aligned}
cf_X(t) &= \int_a^b \left(e^{i\cdot t\cdot x}\cdot 1/(b-a)\right)dx \\
&= \frac{1}{2\cdot h}\cdot \int_{m-h}^{m+h} e^{i\cdot t\cdot x}dx \\
&= \frac{1}{2\cdot h}\cdot \frac{e^{i\cdot t\cdot x}}{i\cdot t}\Bigg|_{x=m-h}^{x=m+h} \\
&= \frac{1}{2\cdot h}\cdot \frac{e^{i\cdot t\cdot (m+h)} - e^{i\cdot t\cdot (m-h)}}{i\cdot t} \\
&= \frac{1}{2\cdot h}\cdot e^{i\cdot t\cdot m}\cdot \frac{e^{i\cdot t\cdot h} - e^{i\cdot t\cdot (-h)}}{i\cdot t} \\
&= e^{i\cdot t\cdot m}\cdot \frac{\sin(t\cdot h)}{t\cdot h} \qquad \text{(siehe Kapitel 5.9).}
\end{aligned}
$$

Momente:

$$\mathrm{E}\left[X^k\right] = \frac{1}{b-a}\cdot\frac{b^{k+1}-a^{k+1}}{k+1}$$

Das Ergebnis folgt direkt aus der Definition des *k*-ten Moments:

$$\mathrm{E}\left[X^k\right] = \int_a^b \left(x^k\cdot 1/(b-a)\right)dx = \frac{1}{b-a}\cdot\int_a^b x^k dx = \frac{1}{b-a}\cdot\frac{b^{k+1}-a^{k+1}}{k+1}.$$

Einpunktverteilung:

Modell:

Die Zufallsvariable *X* nimmt nur die Werte 0 und 1 an. Der Wert1 wird als **Treffer** (**Erfolg**) eines Zufallsexperiments bezeichnet. Ein Treffer tritt mit Trefferwahrscheinlichkeit *p* ein.

Massenfunktion:

$$f_X(x) = P(X=x) = \begin{cases} 1-p & \text{für } x=0 \\ p & \text{für } x=1 \\ 0 & \text{sonst} \end{cases}$$

Hierbei handelt es sich offensichtlich um eine Massenfunktion.

Verteilungsfunktion:

$$F_X(x) = P(X\le x) = \begin{cases} 0 & \text{für } x<0 \\ 1-p & \text{für } 0\le x<1 \\ 1 & \text{für } 1\le x \end{cases}$$

Erwartungswert: $\mathrm{E}[X] = p$ ***Varianz:*** $\mathrm{Var}(X) = p \cdot (1-p)$

Denn $\mathrm{E}[X] = 0 \cdot (1-p) + 1 \cdot p = p$ und $\mathrm{Var}[X] = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = p \cdot (1-p)$.

charakteristische Funktion:

$cf_X(t) = 1 + p \cdot \left(e^{i \cdot t} - 1\right)$

Denn $cf_X(t) = e^{i \cdot t \cdot 0} \cdot (1-p) + e^{i \cdot t} \cdot p = 1 + p \cdot \left(e^{i \cdot t} - 1\right)$.

Momente:

$\mathrm{E}[X^k] = p$

Denn $\mathrm{E}[X^k] = 0^k \cdot (1-p) + 1^k \cdot p = p$.

Binomialverteilung:

Modell:

Ein Experiment wird *n*-mal hintereinander durchgeführt. Bei jeder Ausführung kann ein Treffer auftreten (Wert 1) oder nicht (Wert 0). In jedem Experiment ist die Trefferwahrscheinlichkeit gleich *p*. Die Trefferwahrscheinlichkeit des *i*-ten Experiments hängt nicht von den Ausgängen der vorherigen Experimente ab. Die Zufallsvariable *X* zählt die Anzahl der Treffer, d.h. X nimmt die Werte 0, ... *n* an.

Eine entsprechende Modellvorstellung wird durch das **Urnenmodell mit Zurücklegen** beschrieben: In einer Urne befinden sich *N* Kugeln, und zwar *w* weiße und $N - w$ schwarze Kugeln. Der Anteil der weißen Kugeln ist $p = w/N$. Es werden nacheinander *n* Kugeln gezogen. Dabei wird jeweils die Färbung der gezogenen Kugel notiert und anschließend diese sofort wieder in die Urne zurückgelegt. Die Zufallsvariable *X* zählt die Anzahl der gezogenen weißen Kugeln, d.h. *X* nimmt die Werte 0, ... *n* an.

Massenfunktion:

$$f_X(k) = P(X = k) = \begin{cases} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} & \text{für } k = 0, ..., n \\ 0 & \text{sonst} \end{cases}$$

Es ist $0 \le f_X(k)$ und $\sum_{k=0}^{n} f_X(k) = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \left(p + (1-p)\right)^n = 1$ und damit insbesondere auch $f_X(k) \le 1$.

Verteilungsfunktion:

$$F_X(x) = P(X \le x) = \sum_{0 \le k \le x} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Erwartungswert: $\mathrm{E}[X] = n \cdot p$ ***Varianz:*** $\mathrm{Var}(X) = n \cdot p \cdot (1-p)$

Zur Berechnung wird mehrmals Satz 4.1-3(vii) verwendet:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \\
&= n \cdot p \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} \cdot p^{k-1} \cdot (1-p)^{n-1-(k-1)} \\
&= n \cdot p \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} \cdot p^{k} \cdot (1-p)^{n-1-k} \\
&= n \cdot p \cdot (p + (1-p))^{n-1} \\
&= n \cdot p \ ,
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{k=0}^{n} k^2 \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} - (n \cdot p)^2 \\
&= \sum_{k=0}^{n} k \cdot (k-1) \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} + \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} - (n \cdot p)^2 \\
&= n \cdot (n-1) \cdot p^2 \cdot \sum_{k=2}^{n} \binom{n-2}{k-2} \cdot p^{k-2} \cdot (1-p)^{n-2-(k-2)} + n \cdot p \cdot (1 - n \cdot p) \\
&= n \cdot (n-1) \cdot p^2 \cdot \sum_{k=0}^{n-2} \binom{n-2}{k} \cdot p^{k} \cdot (1-p)^{n-2-k} + n \cdot p \cdot (1 - n \cdot p) \\
&= n \cdot (n-1) \cdot p^2 + n \cdot p \cdot (1 - n \cdot p) \\
&= n \cdot p \cdot (1-p) \ .
\end{aligned}$$

charakteristische Funktion:

$$cf_X(t) = \left(1 + p \cdot \left(e^{i \cdot t} - 1\right)\right)^n$$

Denn $cf_X(t) = \sum_{k=0}^{n} e^{i \cdot t \cdot k} \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \left(p \cdot e^{i \cdot t} + 1 - p\right)^n = \left(1 + p \cdot \left(e^{i \cdot t} - 1\right)\right)^n$.

Momente:

$\mathrm{E}[X^k] = \dfrac{cf_X^{(k)}(0)}{i^k}$ (mit Satz 7.5-2(ii)).

Insbesondere

$$\mathrm{E}[X] = cf_X'(0)/i = \frac{n\cdot\left(1+p\cdot\left(e^{i\cdot t}-1\right)\right)^{n-1}\cdot i\cdot p\cdot e^{i\cdot t}\Big|_{t=o}}{i} = n\cdot p \text{ und}$$

$$\mathrm{E}[X^2] = cf_X''(0)/i^2 = \frac{n\cdot p\cdot i^2\cdot e^{i\cdot t}\cdot\left(p\cdot(n-1)\cdot\left(1+p\cdot\left(e^{i\cdot t}-1\right)\right)^{n-2}\cdot e^{i\cdot t}+\left(1+p\cdot\left(e^{i\cdot t}-1\right)\right)^{n-1}\right)\Big|_{t=o}}{i^2}$$

$$= n\cdot p + n\cdot(n-1)\cdot p^2 \ .$$

Hypergeometrische Verteilung:

Modell:

Das **Urnenmodell ohne Zurücklegen** beschreibt folgende Situation: In einer Urne befinden sich N Kugeln, und zwar w weiße und $N-w$ schwarze Kugeln. Der Anteil der weißen Kugeln ist $p = w/N$. Es werden nacheinander n Kugeln gezogen. Bei jeder Ziehung wird die Färbung der Kugel notiert und anschließend die gezogene Kugel nicht wieder in die Urne zurückgelegt. Die Zufallsvariable X zählt die Anzahl der gezogenen weißen Kugeln. Ist $n \le w$, dann nimmt X maximal den Wert n; ist $n > w$, dann ist der maximale Wert von X gleich $w = N\cdot p$. Daher nimmt X einen Wert k mit $k \le \min\{n, N\cdot p\}$ an. Entsprechend kann für $n \le N-w$ (gleich Anzahl der schwarzen Kugeln) bei jedem Zug eine schwarze Kugel gezogen werden, d.h. die Variable X kann den Wert $k=0$ annehmen; für $n > N-w$ können höchstens $N-w$ schwarze Kugeln gezogen werden, also werden mindestens $n-(N-w) = n-N\cdot(1-p)$ weiße Kugeln gezogen. Daher nimmt X die Werte k mit $\max\{0, n-N\cdot(1-p)\} \le k \le \min\{n, N\cdot p\}$ an.

Die Anzahl der Möglichkeiten, bei den n gezogenen Kugeln genau k weiße Kugeln zu haben, beträgt $\binom{n}{k}$. Daher beträgt die Wahrscheinlichkeit, in den ersten k Zügen ohne Zurücklegen jeweils eine weiße und in den nächsten $n-k$ Zügen jeweils eine schwarze Kugel zu ziehen, $\binom{n}{k}\cdot\frac{w}{N}\cdot\frac{w-1}{N-1}\cdot\ldots\cdot\frac{w-(k-1)}{N-(k-1)}\cdot\frac{N-w}{N-k}\cdot\frac{N-w-1}{N-k-1}\cdot\ldots\cdot\frac{N-w-(n-k-1)}{N-k-(n-k-1)}$. Dieser Ausdruck ist auch die Wahrscheinlichkeit, dass man k weiße und $n-k$ schwarze Kugeln in einer beliebigen Reihenfolge zieht, da die Reihenfolge des Ziehens nur Einfluss auf die Faktorenanordnung des Zählers hat. Daher ist die Wahrscheinlichkeit, bei n Ziehungen ohne Zurücklegen k weiße und $n-k$ schwarze Kugeln zu ziehen gleich (siehe Kapitel 4.1)

$$\binom{n}{k}\cdot\frac{w}{N}\cdot\frac{w-1}{N-1}\cdot\ldots\cdot\frac{w-(k-1)}{N-(k-1)}\cdot\frac{N-w}{N-k}\cdot\frac{N-w-1}{N-k-1}\cdot\ldots\cdot\frac{N-w-(n-k-1)}{N-k-(n-k-1)}$$

$$=\frac{n!}{k!\cdot(n-k)!}\cdot\frac{w\cdot(w-1)\cdot\ldots\cdot(w-k+1)\cdot(N-w)\cdot(N-w-1)\cdot\ldots\cdot(N-w-(n-k)+1)}{N\cdot(N-1)\cdot\ldots\cdot(N-n+1)}$$

$$=\frac{(w\cdot(w-1)\cdot\ldots\cdot(w-k+1)/k!)\cdot((N-w)\cdot(N-w-1)\cdot\ldots\cdot(N-w-(n-k)+1)/(n-k)!)}{(N\cdot(N-1)\cdot\ldots\cdot(N-n+1))/n!}$$

$$=\frac{\binom{w}{w-k}\cdot\binom{N-w}{n-k}}{\binom{N}{N-n}}$$

$$=\frac{\binom{w}{k}\cdot\binom{N-w}{n-k}}{\binom{N}{n}}\quad.$$

Massenfunktion:

$$f_X(k)=P(X=k)=\begin{cases}\dfrac{\binom{w}{k}\cdot\binom{N-w}{n-k}}{\binom{N}{n}} & \text{für } \max\{\,0,\ n-N\cdot(1-p)\,\}\le k\le\min\{\,n,\ N\cdot p\,\}\\ 0 & \text{sonst}\end{cases}$$

Hierdurch wird eine Massenfunktion definiert: Es ist $f_X(k)\ge 0$. Zum Nachweis, dass die Summe aller definierten Werte gleich 1 ist, wird die Formel $\binom{a+b}{l}=\sum_{i=0}^{l}\binom{a}{i}\cdot\binom{b}{l-i}$ mit $a\in\mathbf{N}$, $b\in\mathbf{N}$ und $l\in\mathbf{N}$ verwendet. Diese Formel erklärt sich wie folgt: Sind *A* und *B* disjunkte Mengen mit $|A|=a$ und $|B|=b$. Dann steht auf der linken Seite des Gleichheitszeichens die Anzahl *l*-elementiger Teilmengen von $A\cup B$. Es sei *C* eine *l*-elementige Teilmengen von $A\cup B$. Hat *C* mit *A* genau *i* Elemente gemeinsam, dann hat *C* mit *B* genau $l-i$ Elemente gemeinsam. Das bedeutet, dass man jeder *l*-elementigen Teilmenge von $A\cup B$, die genau *i* Elemente aus *A* enthält, alle $(l-i)$-elementigen Teilmengen von *B* zuordnen kann. Das sind genau $\binom{b}{l-i}$ viele. Daher gilt $\binom{a+b}{l}=\sum_{i=0}^{l}\binom{a}{i}\cdot\binom{b}{l-i}$. Damit ist

$$\sum_k f_X(k)=1\Big/\binom{N}{n}\cdot\sum_k\binom{w}{k}\cdot\binom{N-w}{n-k}=1$$ und insbesondere auch $f_X(k)\le 1$.

Verteilungsfunktion:

$$F_X(x)=P(X\le x)=\sum_{k\le x}f_X(k)$$

Erwartungswert: $\mathrm{E}[X] = n \cdot p$ ***Varianz:*** $\mathrm{Var}(X) = n \cdot p \cdot (1-p) \cdot \frac{N-n}{N-1}$

Obige Formel wird auf den Erwartungswert und die Varianz angewandt:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{k=0}^{n} k \cdot \binom{w}{k} \cdot \binom{N-w}{n-k} \Big/ \binom{N}{n} \\
&= 1 \Big/ \binom{N}{n} \cdot w \cdot \sum_{k=1}^{n} \binom{w-1}{k-1} \cdot \binom{N-w}{n-k} \\
&= 1 \Big/ \binom{N}{n} \cdot w \cdot \sum_{k=0}^{n-1} \binom{w-1}{k} \cdot \binom{N-w}{n-1-k} \\
&= 1 \Big/ \binom{N}{n} \cdot w \cdot \binom{N-1}{n-1} \\
&= w \cdot \frac{n}{N} \\
&= n \cdot p
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{k=0}^{n} k^2 \cdot f_X(k) - (n \cdot p)^2 \\
&= \sum_{k=0}^{n} k \cdot (k-1) \cdot f_X(k) + \sum_{k=0}^{n} k \cdot f_X(k) - (n \cdot p)^2 \\
&= \sum_{k=0}^{n} k \cdot (k-1) \cdot \binom{w}{k} \cdot \binom{N-w}{n-k} \Big/ \binom{N}{n} + n \cdot p \cdot (1 - n \cdot p) \\
&= 1 \Big/ \binom{N}{n} \cdot w \cdot (w-1) \cdot \sum_{k=2}^{n} \binom{w-2}{k-2} \cdot \binom{N-w}{n-k} + n \cdot p \cdot (1 - n \cdot p) \\
&= 1 \Big/ \binom{N}{n} \cdot w \cdot (w-1) \cdot \sum_{k=0}^{n-2} \binom{w-2}{k} \cdot \binom{N-w}{n-2-k} + n \cdot p \cdot (1 - n \cdot p) \\
&= 1 \Big/ \binom{N}{n} \cdot w \cdot (w-1) \cdot \binom{N-2}{n-2} + n \cdot p \cdot (1 - n \cdot p) \qquad (\text{mit } w = p \cdot N) \\
&= \frac{p \cdot N \cdot (p \cdot N - 1) \cdot (n-1) \cdot n}{(N-1) \cdot N} + n \cdot p \cdot (1 - n \cdot p) \\
&= \frac{n \cdot p \cdot N - n \cdot p^2 \cdot N + n^2 \cdot p^2 - n \cdot p^2}{(N-1)} \\
&= n \cdot p \cdot (1-p) \cdot \frac{N-n}{N-1}
\end{aligned}$$

Offensichtlich haben Binomialverteilung und hypergeometrische Verteilung denselben Erwartungswert. Die Varianzen unterscheiden sich um den Faktor $\frac{N-n}{N-1}$, der für $n > 1$ kleiner als 1 ist, d.h. die Varianz ist bei Ziehen ohne Zurücklegen kleiner. Bei festem Umfang N der Gesamtheit und großem Stichprobenumfang, d.h. $n \to N$, geht die Varianz gegen 0. Bei fes-

tem *n* und großem Umfang der Gesamtheit, d.h. $N \to \infty$, geht die Varianz gegen die Varianz der Binomialverteilung, d.h. Ziehen mit und ohne Zurücklegen zeigen dann dasselbe statistische Verhalten. Als Faustregel findet sich für diese Situation in der Literatur die Angabe $N > 20 \cdot n$.

Poissonverteilung:

Modell:

Ein Experiment wird *n*-mal hintereinander durchgeführt. Bei jeder Ausführung kann ein Treffer auftreten (Wert 1) oder nicht (Wert 0). Die Trefferwahrscheinlichkeit des *i*-ten Experiments hängt nicht von den Ausgängen der vorherigen Experimente ab. Die Zufallsvariable *X* zählt die Anzahl der Treffer, d.h. *X* nimmt die Werte 0, ... *n* an. Das Experiment wird sehr oft durchgeführt: $n \to \infty$. In jedem Experiment ist die Trefferwahrscheinlichkeit *p* sehr klein: $p \to 0$.

Die Zufallsvariable *X* ist also binomialverteilt. Dabei wird der Grenzübergang $n \to \infty$ und $p \to 0$ so durchgeführt, dass der Erwartungswert $n \cdot p$ konstant bleibt. Daher ist mit $n \to \infty$ auch $p \to 0$ erfüllt.

Es wird $\lambda = n \cdot p$ gesetzt, d.h. $p = \lambda / n$ (die hier beschriebene Verteilung gilt auch dann, wenn $\lim_{n\to\infty}(n \cdot p) = \lambda$ ist). Dann ist für $k = 1, ..., n$

$$f_X(k) = \lim_{n\to\infty}\left(\binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}\right)$$

$$= \lim_{n\to\infty}\left(\frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k}\right) .$$

Der erste Faktor ist gleich $\frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{n^k} = 1 \cdot \prod_{i=1}^{k-1}\left(1 - \frac{i}{n}\right)$. Beim Grenzübergang $n \to \infty$ gilt daher $\lim_{n\to\infty} \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{n^k} = 1$, $\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^n = \lim_{n\to\infty}\left(1 + \frac{-\lambda}{n}\right)^n = e^{-\lambda}$ (siehe Kapitel 5.1) und $\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^{-k} = 1$. Insgesamt ist $f_X(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$.

Massenfunktion:

$$f_X(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \text{ für } k \in \mathbf{N}$$

Es ist $f_X(k) \geq 0$, $\sum_{k=0}^{\infty} f_X(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot e^{-\lambda} = e^{-\lambda} \cdot e^{\lambda} = 1$ und $f_X(k) \leq 1$.

Verteilungsfunktion:

$$F_X(x) = P(X \le x) = \sum_{k \le \lfloor x \rfloor} f_X(k)$$

Erwartungswert: $\mathrm{E}[X] = \lambda$ ***Varianz:*** $\mathrm{Var}(X) = \lambda$

Zur Berechnung wird $G(z) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot z^k = e^{\lambda \cdot z}$ gesetzt. Nach Satz 5.11-1(vi) ist

$\sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} \cdot z^k = z \cdot G'(z) = z \cdot \lambda \cdot e^{\lambda \cdot z}$, also (mit $z = 1$) $\sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} = \lambda \cdot e^{\lambda}$. Damit ist

$$\mathrm{E}[X] = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} = e^{-\lambda} \cdot \lambda \cdot e^{\lambda} = \lambda .$$

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{k=0}^{\infty} k^2 \cdot f_X(k) - \lambda^2 \\
&= \lambda \cdot e^{-\lambda} \cdot \sum_{k=1}^{\infty} k \cdot \frac{\lambda^{k-1}}{(k-1)!} - \lambda^2 \\
&= \lambda \cdot e^{-\lambda} \cdot \sum_{k=0}^{\infty} (k+1) \cdot \frac{\lambda^k}{k!} - \lambda^2 \\
&= \lambda \cdot e^{-\lambda} \cdot \left(\sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} + \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right) - \lambda^2 \\
&= \lambda \cdot e^{-\lambda} \cdot \left(\lambda \cdot e^{\lambda} + e^{\lambda} \right) - \lambda^2 \\
&= \lambda \quad .
\end{aligned}$$

charakteristische Funktion:

$$cf_X(t) = e^{\lambda \cdot (e^{i \cdot t} - 1)}$$

$$\begin{aligned}
\text{Denn } cf_X(t) &= \sum_{k=0}^{n} e^{i \cdot t \cdot k} \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \\
&= e^{-\lambda} \cdot \sum_{k=0}^{n} \frac{\left(e^{i \cdot t} \cdot \lambda \right)^k}{k!} \\
&= e^{-\lambda} \cdot e^{\lambda \cdot e^{i \cdot t}} \\
&= e^{\lambda \cdot (e^{i \cdot t} - 1)} \quad .
\end{aligned}$$

Momente:

$$\mathrm{E}[X^k] = cf_X^{(k)}(0) / i^k$$

Geometrische Verteilung:

Modell:
Es wird ein Experiment durchgeführt, bei dem jeweils ein Treffer oder ein Nichttreffer auftreten kann, und zwar solange, bis zum ersten Male ein Treffer vorliegt. Die Trefferwahrscheinlichkeit ist gleich p. Die Trefferwahrscheinlichkeit des i-ten Experiments hängt nicht von den Ausgängen der vorherigen Experimente ab. Die Zufallsvariable X zählt die Anzahl der Nichttreffer bis zum ersten Treffer, d.h. X nimmt als Werte alle natürlichen Zahlen an.

Massenfunktion:
$f_X(k) = (1-p)^k \cdot p$ für $k \in \mathbf{N}$

Offensichtlich ist $0 \le f_X(k) \le 1$ und $\sum_{k=0}^{\infty} p \cdot (1-p)^k = p \cdot \frac{1}{1-(1-p)} = 1$.

Verteilungsfunktion:
$F_X(x) = P(X \le x) = 1 - (1-p)^{\lfloor x \rfloor + 1}$

Mit Satz 1.6-2(ii) gilt:
$F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} p \cdot (1-p)^k = p \cdot \frac{1-(1-p)^{\lfloor x \rfloor + 1}}{1-(1-p)} = 1 - (1-p)^{\lfloor x \rfloor + 1}$.

Erwartungswert: $\mathrm{E}[X] = \frac{1-p}{p}$ ***Varianz:*** $\mathrm{Var}(X) = \frac{(1-p)}{p^2}$

Mit Satz 5.1-9(i) folgt
$\mathrm{E}[X] = \sum_{k=0}^{\infty} k \cdot p \cdot (1-p)^k = p \cdot \frac{1-p}{(1-(1-p))^2} = \frac{1-p}{p}$.

Zur Berechnung der Varianz $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$ wird das 2-te Moment mit Hilfe der charakteristischen Funktion (siehe unten) berechnet.
$\mathrm{Var}(X) = p \cdot \sum_{k=0}^{\infty} k^2 \cdot (1-p)^k - \left(\frac{1-p}{p}\right)^2 = p \cdot \frac{(1-p)\cdot(2-p)}{p^3} - \left(\frac{1-p}{p}\right)^2 = \frac{(1-p)}{p^2}$.

charakteristische Funktion:
$cf_X(t) = \sum_{k=0}^{\infty} e^{i\cdot t\cdot k} \cdot p \cdot (1-p)^k = \frac{p}{1-e^{i\cdot t}\cdot(1-p)} = p \cdot (1-e^{i\cdot t}\cdot(1-p))^{-1}$,

$cf_X'(t) = p \cdot (-1) \cdot (1-p) \cdot e^{i\cdot t} \cdot i \cdot (-1) \cdot (1-e^{i\cdot t}\cdot(1-p))^{-2}$

$= p \cdot (1-p) \cdot i \cdot e^{i\cdot t} \cdot (1-e^{i\cdot t}\cdot(1-p))^{-2}$,

$$
\begin{aligned}
cf_X''(t) &= p\cdot(1-p)\cdot i\cdot e^{i\cdot t}\cdot\left(1-e^{i\cdot t}\cdot(1-p)\right)^{-2}\\
&= p\cdot(1-p)\cdot i\cdot\left(i\cdot e^{i\cdot t}\cdot\left(1-e^{i\cdot t}\cdot(1-p)\right)^{-2}+e^{i\cdot t}\cdot(-2)\cdot\left(1-e^{i\cdot t}\cdot(1-p)\right)^{-3}\cdot(-1)\cdot(1-p)\cdot i\cdot e^{i\cdot t}\right)\\
&= p\cdot(1-p)\cdot i^2\cdot e^{i\cdot t}\cdot\left(\frac{2\cdot e^{i\cdot t}\cdot(1-p)}{\left(1-e^{i\cdot t}\cdot(1-p)\right)^3}+\frac{1}{\left(1-e^{i\cdot t}\cdot(1-p)\right)^2}\right)\ ,
\end{aligned}
$$

Momente:

$\mathrm{E}\left[X^k\right]=cf_X^{(k)}(0)/i^k$, insbesondere

$\mathrm{E}[X]=cf_X'(0)/i=p\cdot(1-p)\cdot p^{-2}=(1-p)/p$,

$\mathrm{E}\left[X^2\right]=cf_X''(0)/i^2=p\cdot(1-p)\cdot\left(\frac{2\cdot(1-p)}{p^3}+\frac{1}{p^2}\right)=(1-p)\cdot\frac{2-p}{p^2}$.

Exponentialverteilung:

Modell:

Die Zufallsvariable X beschreibt die Dauer eines Ereignisses, für das kleine Werte „sehr wahrscheinlich“ sind und größere Werte nur mit exponentiell abnehmender Wahrscheinlichkeit auftreten. Dabei können beliebige nichtnegative Werte auftreten.

Dichtefunktion:

$f_X(x)=\lambda\cdot e^{-\lambda\cdot x}$ für $x\in\mathbf{R}$ mit $x\geq 0$; hierbei ist $\lambda>0$ ein fester Parameter.

Es ist $f_X(x)\geq 0$ und $\int_0^\infty\left(\lambda\cdot e^{-\lambda\cdot t}\right)dt=\lim_{h\to\infty}\int_0^h\left(\lambda\cdot e^{-\lambda\cdot t}\right)dt=\lambda\cdot\lim_{h\to\infty}\left(\frac{1}{-\lambda}\cdot e^{-\lambda\cdot t}\Big|_{t=0}^{t=h}\right)=1$.

Verteilungsfunktion:

$$F_X(x)=\int_0^x\left(\lambda\cdot e^{-\lambda\cdot t}\right)dt=\lambda\cdot\frac{1}{-\lambda}\cdot e^{-\lambda\cdot t}\Big|_{t=0}^{t=x}=1-e^{-\lambda\cdot t}$$

Erwartungswert: $\mathrm{E}[X]=1/\lambda$ ***Varianz:*** $\mathrm{Var}(X)=(1/\lambda)^2$

Mit Beispiel 3. am Ende von Kapitel 5.13 ergibt sich $\mathrm{E}[X]=\int_0^\infty\left(t\cdot\lambda\cdot e^{-\lambda\cdot t}\right)dt=1/\lambda$.

Zur Berechnung der Varianz wird Satz 7.5-2(ii) herangezogen:

$\mathrm{Var}(X)=\mathrm{E}\left[X^2\right]-(\mathrm{E}[X])^2$, wobei das 2-te Moment mit Hilfe der charakteristischen Funktion (siehe unten) berechnet wird.

$\mathrm{Var}(X)=2/\lambda^2-(1/\lambda)^2=(1/\lambda)^2$.

charakteristische Funktion:

$$cf_X(t) = \frac{\lambda}{\lambda - i \cdot t} = \frac{1}{1 - \frac{i \cdot t}{\lambda}}$$

Denn

$$cf_X(t) = \int_0^\infty \left(e^{i \cdot t \cdot x} \lambda \cdot e^{-\lambda \cdot x}\right) dx = \lim_{h \to \infty} \int_0^h \left(e^{i \cdot t \cdot x} \lambda \cdot e^{-\lambda \cdot x}\right) dx = \lambda \cdot \lim_{h \to \infty} \left(\frac{1}{i \cdot t - \lambda} e^{-x \cdot (\lambda - i \cdot t)} \Big|_{x=0}^{x=h} \right) = \frac{\lambda}{\lambda - i \cdot t} .$$

Damit ist $cf_X^{(k)}(t) = \frac{k!}{\lambda^k} \cdot i^k \cdot \frac{1}{\left(1 - \frac{i \cdot t}{\lambda}\right)^k}$.

Momente:

$\mathrm{E}\left[X^k\right] = cf_X^{(k)}(0) / i^k = \frac{k!}{\lambda^k}$, insbesondere

$\mathrm{E}[X] = 1/\lambda$,

$\mathrm{E}\left[X^2\right] = 2/\lambda^2$.

Normalverteilung:

Die Normalverteilung ist die wichtigste Verteilung in der Statistik. Viele empirisch beobachtbare Merkmale sind normalverteilt. Die besondere Bedeutung der Normalverteilung zeigt sich aber besonders in den Grenzwertsätzen der Statistik.

Dichtefunktion:

$f_X(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x - \mu}{\sigma}\right)^2}$ für $x \in \mathbf{R}$. Hierbei sind $\sigma > 0$ und μ feste Parameter.

Die durch $Y = \frac{X - \mu}{\sigma}$ definierte Zufallsvariable hat nach Satz 7.2-5(ii) (dort mit $a = 1/\sigma$ und $b = -\mu/\sigma$) die Dichtefunktion $f_Y(y) = \frac{1}{1/\sigma} \cdot \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x - \mu}{\sigma}\right)^2} \Big|_{x = \frac{y - (-\mu/\sigma)}{1/\sigma}} = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot y^2}$.

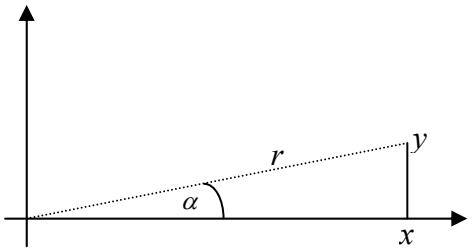
Man nennt Y **normiert (0,1)-normalverteilt**.

Der Nachweis, dass es sich bei der durch $f(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot x^2}$ definierten Funktion um eine Dichtefunktion handelt, d.h. dass $\int_{-\infty}^{\infty} f(x) dx = 1$ gilt, erfordert einige Hilfsmittel aus der Analy-

sis. Dazu wird zunächst $I = \int_{-\infty}^{\infty} e^{-\frac{1}{2}\cdot x^2}\,dx$ bestimmt. Dieses Integral ist nicht elementar integrierbar. Jedoch führt folgender Ansatz weiter:

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}\cdot x^2}\,dx\right)\cdot\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}\cdot y^2}\,dy\right) = \int_{-\infty-\infty}^{\infty\infty} e^{-\frac{1}{2}\cdot x^2}\cdot e^{-\frac{1}{2}\cdot y^2}\,dxdy = \int_{-\infty-\infty}^{\infty\infty} e^{-\frac{1}{2}\cdot\left(x^2+y^2\right)}\,dxdy\,.$$

Die Integration geht über alle Punkte $(x,y)\in\mathbf{R}\times\mathbf{R}$. Ein derartiger Punkt (mit Ausnahme von $(0,0)$) lässt sich nicht nur eindeutig durch seine Koordinaten *x* und *y* darstellen, sondern auch eindeutig durch seinen Abstand *r* zum Punkt $(0,0)$ und durch den Winkel α, den die Verbindungslinie von $(0,0)$ zu (x,y) und die positive *x*-Achse bildet:



Es gilt $(x,y)=(r\cdot\cos(\alpha), r\cdot\sin(\alpha))$ mit $r>0$ und $0\le\alpha\le 2\cdot\pi$. Als Erweiterung der Substitutionsregel aus Satz 5.13-6(ii) auf mehrdimensionale Integrale (hier auf den zweidimensionalen Raum) gilt (siehe [SHSS]):

Es sei $x=g_1(r,\alpha)$ und $y=g_2(r,\alpha)$. Dann gilt (unter bestimmten „vernünftigen“ Bedingungen an den Integrationsbereich und die beteiligten Funktionen)

$$\iint_A h(x,y)dxdy = \iint_{A'}\left(|D|\cdot h(g_1(r,\alpha), g_2(r,\alpha))\right)drd\alpha$$

mit $D=\frac{\partial g_1(r,\alpha)}{\partial r}\cdot\frac{\partial g_2(r,\alpha)}{\partial\alpha}-\frac{\partial g_2(r,\alpha)}{\partial r}\cdot\frac{\partial g_1(r,\alpha)}{\partial\alpha}$ (hier stehen die jeweiligen partiellen Ableitungen); A' ist die Menge in (r,α)-Koordinaten, die der Menge *A* in (x,y)-Koordinaten entspricht.

Im vorliegenden Fall ist $h(x,y)=e^{-\frac{1}{2}\cdot\left(x^2+y^2\right)}$, $D=\cos(\alpha)\cdot r\cdot\cos(\alpha)-\sin(\alpha)\cdot(-r\cdot\sin(\alpha))=r$. Mit $A_n=\left\{(x,y)\,\middle|\, x^2+y^2\le n^2\right\}$ ist $A_n'=\left\{(r,\alpha)\,\middle|\, r\le n \text{ und } 0\le\alpha\le 2\cdot\pi\right\}$ und

$$\begin{aligned}\iint_{A_n} e^{-\frac{1}{2}\cdot\left(x^2+y^2\right)}dxdy &= \int_0^{2\cdot\pi}\left(\int_0^n r\cdot e^{-\frac{1}{2}\cdot r^2}\,dr\right)d\alpha\\ &= \int_0^{2\cdot\pi}\left(-e^{-\frac{1}{2}\cdot r^2}\Big|_{r=0}^{r=n}\right)d\alpha\\ &= 2\cdot\pi\cdot\left(-e^{-\frac{1}{2}\cdot n^2}+1\right)\,.\end{aligned}$$

Damit ist $I^2 = \lim\limits_{n\to\infty}\left(\iint_{A_n} e^{-\frac{1}{2}\cdot\left(x^2+y^2\right)}dxdy\right) = \lim\limits_{n\to\infty}\left(2\cdot\pi\cdot\left(-e^{-n^2}+1\right)\right) = 2\cdot\pi$ und

$$\int\limits_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\cdot x^2}\right)dx = 1.$$

Mit der Substitutionsregel aus Satz 5.13-6(ii) ist ebenfalls

$$\int\limits_{-\infty}^{\infty} f_X(x)dx = \int\limits_{-\infty}^{\infty}\left(\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}\right)dx = \int\limits_{-\infty}^{\infty}\left(f_Y\left(\frac{x-\mu}{\sigma}\right)\cdot\left(\frac{x-\mu}{\sigma}\right)'\right)dx = \int\limits_{-\infty}^{\infty} f_Y(x)dx = 1.$$

Die Dichtefunktion $f_X(x) = \frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}$ ist symmetrisch zum Punkt $x=\mu$ und hat folgende Eigenschaften:

Durch Bildung der ersten und zweiten Ableitung findet man Extremwert und Wendepunkte:

$$f_X'(x) = -\frac{1}{\sigma^2\cdot\sqrt{2\cdot\pi}}\cdot\frac{x-\mu}{\sigma}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2},\quad f_X''(x) = \frac{1}{\sigma^3\cdot\sqrt{2\cdot\pi}}\cdot\left(\left(\frac{x-\mu}{\sigma}\right)^2-1\right)\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Der Extremwert liegt bei $x=\mu$; wegen $f_X''(\mu) = -\frac{1}{\sigma^3\cdot\sqrt{2\cdot\pi}} < 0$ handelt es sich um ein Maximum. Die Wendepunkte liegen bei $x_{1,2} = \mu\pm\sigma$.

Verteilungsfunktion:

$$F_X(x) = \int\limits_{-\infty}^{x}\left(\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{t-\mu}{\sigma}\right)^2}\right)dt$$

Erwartungswert: $\mathrm{E}[X] = \mu$ ***Varianz:*** $\mathrm{Var}(X) = \sigma^2$

Man bezeichnet X auch als (μ,σ^2)**-normalverteilt**.

Der Erwartungswert der normiert (0, 1)-normalverteilten Zufallsvariable Y berechnet sich zu

$$\mathrm{E}[Y] = \int\limits_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\cdot\pi}}\cdot x\cdot e^{-\frac{1}{2}\cdot x^2}\right)dx = \lim_{t\to\infty}\int\limits_{-t}^{t}\left(\frac{1}{\sqrt{2\cdot\pi}}\cdot x\cdot e^{-\frac{1}{2}\cdot x^2}\right)dx = \lim_{t\to\infty}\left(-\frac{1}{\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\cdot x^2}\Big|_{x=-t}^{x=t}\right) = 0.$$

Die Varianz der normiert (0, 1)-normalverteilten Zufallsvariable Y berechnet sich zu $\mathrm{Var}(Y) = \mathrm{E}[Y^2] = 1$. Das zweite Moment wird dabei mit Hilfe der charakteristischen Funktion berechnet (siehe unten). Der Erwartungswert von $X = \sigma\cdot Y+\mu$ lautet daher $\mathrm{E}[X] = \mu$ und die Varianz $\mathrm{Var}(X) = \sigma^2$.

charakteristische Funktion:

Die charakteristische Funktion der normiert (0, 1)-normalverteilten Zufallsvariable Y lautet

$$cf_Y(t) = e^{-\frac{1}{2}\cdot t^2}$$

Denn

$$cf_Y(t) = \int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\cdot\pi}}\cdot e^{i\cdot t\cdot x}\cdot e^{-\frac{1}{2}\cdot x^2}\right)dx$$

$$= \frac{1}{\sqrt{2\cdot\pi}}\cdot\int_{-\infty}^{\infty}\left(e^{i\cdot t\cdot x}\cdot e^{-\frac{1}{2}\cdot x^2}\right)dx = \frac{1}{\sqrt{2\cdot\pi}}\cdot\int_{-\infty}^{\infty}\left(e^{-\frac{(x-i\cdot t)^2}{2}}\cdot e^{-\frac{1}{2}\cdot t^2}\right)dx = e^{-\frac{1}{2}\cdot t^2} \quad .$$

Mit Satz 7.5-2(iii) gilt dann für $X = \sigma\cdot Y + \mu$: $cf_X(t) = e^{i\cdot\mu\cdot t}\cdot e^{-\frac{1}{2}\cdot\sigma^2\cdot t^2}$.

Momente:

Die Momente der normiert (0, 1)-normalverteilten Zufallsvariable Y lauten $\mathrm{E}[Y^k] = cf_Y^{(k)}(0)/i^k$, insbesondere

$$\mathrm{E}[Y] = cf_Y'(0)/i = -t\cdot e^{-\frac{1}{2}\cdot t^2}\Big|_{t=0}\Big/ i = 0,$$

$$\mathrm{E}[Y^2] = cf_Y''(0)/i^2 = \left(e^{-\frac{1}{2}\cdot t^2}\cdot\left(t^2-1\right)\right)\Big|_{t=0}\Big/ i^2 = 1.$$

Die Verteilung von X ist stark um den Erwartungswert μ konzentriert. Dazu wird $P\left(|X-\mu| \le k\cdot\sigma\right)$ mit $k\in\mathbf{N}_{>0}$ berechnet:

$$P\left(|X-\mu| \le k\cdot\sigma\right) = P\left(\frac{|X-\mu|}{\sigma} \le k\right)$$

$$= P\left(|Y| \le k\right) = P\left(-k \le Y \le k\right)$$

$$= P\left(Y \le k\right) - \left(1 - P\left(Y \le k\right)\right) \qquad \text{(denn } P\left(Y \le -k\right) = P\left(Y > k\right) = 1 - P\left(Y \le k\right)\text{)}$$

$$= 2\cdot P\left(Y \le k\right) - 1$$

mit der normiert (0, 1)-normalverteilten Zufallsvariable Y. Die Werte $P\left(Y \le k\right)$ lassen sich den entsprechenden Tabellen entnehmen. Einige der Werte lauten:

k	$P\left(\lvert X-\mu\rvert \le k\cdot\sigma\right)$
1	0,6826
2	0,9544
3	0,9974

Die n Zufallsvariablen X_i für $i = 1, \ldots, n$ seien stochastisch unabhängig und $\left(\mu_i, \sigma_i^2\right)$-normalverteilt. Die Zufallsvariable $\sum_{i=1}^{n} X_i$ ist dann $\left(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2\right)$-normalverteilt.

Die Zufallsvariable $\frac{1}{n}\cdot\sum_{i=1}^{n} X_i$ ist $\left(\frac{1}{n}\cdot\sum_{i=1}^{n}\mu_i, \frac{1}{n^2}\cdot\sum_{i=1}^{n}\sigma_i^2\right)$-normalverteilt.

Die Aussage wird hier für $n=2$ gezeigt. Die charakteristische Funktion von X_1 lautet $cf_{X_1}(t)=e^{i\cdot\mu_1\cdot t}\cdot e^{-\frac{1}{2}\cdot\sigma_1^2\cdot t^2}$, die von X_2 entsprechend $cf_{X_2}(t)=e^{i\cdot\mu_2\cdot t}\cdot e^{-\frac{1}{2}\cdot\sigma_2^2\cdot t^2}$. Die charakteristische Funktion von X_1+X_2 lautet gemäß Satz 7.5-2(iv)
$cf_{X_1+X_2}(t)=e^{i\cdot\mu_1\cdot t}\cdot e^{-\frac{1}{2}\cdot\sigma_1^2\cdot t^2}\cdot e^{i\cdot\mu_2\cdot t}\cdot e^{-\frac{1}{2}\cdot\sigma_2^2\cdot t^2}=e^{i\cdot(\mu_1+\mu_2)\cdot t}\cdot e^{-\frac{1}{2}\cdot\left(\sigma_1^2+\sigma_2^2\right)t^2}$. Dieses ist die charakteristische Funktion einer $\left(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2\right)$-normalverteilten Zufallsvariablen. Die zweite Aussage folgt mit den Sätzen 7.2-2 und 7.2-3(ii).

Gamma-Verteilung:

Modell:
Die Gamma-Verteilung spielt in den Anwendungen der Schätz- und Testtheorie der Statistik eine wichtige Rolle. Die Verteilung ist eng verknüpft mit der durch

$$\Gamma(p)=\int_0^{\infty}\left(x^{p-1}\cdot e^{-x}\right)dx \quad \text{für} \quad p\in\mathbf{R}_{>0}$$

definierten **Gamma-Fuktion**.

Bevor die Dichtefunktion der Gamma-Verteilung definiert wird, werden einige Eigenschaften der Gamma-Funktion angegeben:

(i) $\Gamma(p)=\int_0^{\infty}\left(x^{p-1}\cdot e^{-x}\right)dx$ ist für $p\in\mathbf{R}_{>0}$ definiert, d.h. das Integral ist endlich.

Die Konvergenz des Integrals sieht man wie folgt:
$0\le\int_0^{\infty}\left(x^{p-1}\cdot e^{-x}\right)dx=\int_0^{1}\left(x^{p-1}\cdot e^{-x}\right)dx+\int_1^{\infty}\left(x^{p-1}\cdot e^{-x}\right)dx$. Im ersten Integral auf der rechten Seite ist $0<e^{-x}\le 1$; daher ist $\int_0^{1}\left(x^{p-1}\cdot e^{-x}\right)dx\le\int_0^{1}\left(x^{p-1}\right)dx=1/p\cdot x^p\Big|_{x=0}^{x=1}=1/p$. Für den Nachweis der Konvergenz des zweiten Integrals auf der rechten Seite sei $n=\lceil p-1\rceil$.Nach Satz 5.5.6-(i) zusammen mit den Überlegungen in den Beispielen von Kapitel 5.7 gilt $\lim_{x\to\infty}\left(x^{n+2}\cdot e^{-x}\right)=0$; daher gibt es eine Zahl $x_0\in\mathbf{R}$, so dass für $x\ge x_0$ die Abschätzung $x^{n+2}\cdot e^{-x}<1$ und damit $x^n\cdot e^{-x}<1/x^2$ gilt. Daher ist

$$\int_1^\infty \left(x^{p-1}\cdot e^{-x}\right)dx \le \int_1^\infty \left(x^n\cdot e^{-x}\right)dx$$

$$= \int_1^{x_0} \left(x^n\cdot e^{-x}\right)dx + \int_{x_0}^\infty \left(x^n\cdot e^{-x}\right)dx$$

$$\le \int_1^{x_0} \left(x^n\cdot e^{-x}\right)dx + \lim_{k\to\infty}\int_{x_0}^k 1/x^2\,dx$$

$$= \int_1^{x_0} \left(x^n\cdot e^{-x}\right)dx + \lim_{k\to\infty}\left(-1/x\big|_{x=1}^k\right)$$

$$= \int_1^{x_0} \left(x^n\cdot e^{-x}\right)dx + 1/x_0 \quad .$$

(ii) $\Gamma(p+1) = p\cdot\Gamma(p)$ für $p\in\mathbf{R}_{>0}$.

Denn $\Gamma(p+1) = \int_0^\infty \left(x^p\cdot e^{-x}\right)dx = \left(-e^{-x}\cdot x^p\right)\Big|_{x=0}^{x=\infty} + p\cdot\int_0^\infty \left(x^{p-1}\cdot e^{-x}\right)dx = p\cdot\Gamma(p).$

(iii) $\Gamma(1)=1$ und $\Gamma(n+1)=n!$ für $n\in\mathbf{N}_{>0}$.

Denn $\Gamma(1) = \int_0^\infty e^{-x}dx = \lim_{t\to\infty}\left(-e^{-x}\right)\Big|_{x=0}^{x=t} = 1$; mit vollständiger Induktion folgt

$\Gamma(n+2) = (n+1)\cdot\Gamma(n+1) = (n+1)\cdot n! = (n+1)!$.

(iv) $\Gamma(1/2)=\sqrt{\pi}$ und $\Gamma\left(n+\frac{1}{2}\right) = \frac{1\cdot 3\cdot\ldots\cdot(2\cdot n-1)}{2^n}\cdot\sqrt{\pi} = \frac{(2\cdot n-1)!}{2^{2\cdot n-1}\cdot(n-1)!}\cdot\sqrt{\pi}$.

Auch hier erfolgt der Nachweis durch vollständige Induktion: Für $n=0$ ist mit der Substitution $x=t^2/2$, d.h. $dx=tdt$, und Satz 5.13.6 (ii)

$\Gamma(1/2) = \int_0^\infty \left(x^{-1/2}\cdot e^{-x}\right)dx = \int_0^\infty \left(\left(\frac{t^2}{2}\right)^{-1/2}\cdot e^{-t^2/2}\cdot t\right)dt = \sqrt{2}\cdot\int_0^\infty \left(e^{-t^2/2}\right)dt = \sqrt{2}\cdot\frac{1}{2}\cdot\int_{-\infty}^\infty \left(e^{-t^2/2}\right)dt$. Für die Dichtefunktion einer $(0,1)$-normalverteilten Zufallsvariablen Y gilt (siehe oben) $\int_{-\infty}^\infty \left(\frac{1}{\sqrt{2\cdot\pi}}\cdot e^{-t^2/2}\right)dt = 1$. Damit ist $\Gamma(1/2)=\sqrt{\pi}$. Der Induktionsschritt erfolgt mit (ii):

$$\Gamma\left(n+1+\frac{1}{2}\right)=\Gamma\left(n+\frac{1}{2}+1\right)$$
$$=\left(n+\frac{1}{2}\right)\cdot\Gamma\left(n+\frac{1}{2}\right)$$
$$=\frac{2\cdot n+1}{2}\cdot\frac{1\cdot 3\cdot\ldots\cdot(2\cdot n-1)}{2^n}\cdot\sqrt{\pi}=\frac{1\cdot 3\cdot\ldots\cdot(2\cdot n-1)\cdot(2\cdot(n+1)-1)}{2^{n+1}}\cdot\sqrt{\pi}\quad .$$

(v) Es sei $\lambda>0$. Dann gilt $\frac{\Gamma(p)}{\lambda^p}=\int_0^\infty\left(x^{p-1}\cdot e^{-\lambda\cdot x}\right)dx$.

Diese Gleichung ist auch für ein komplexes $\lambda=\lambda_1+i\cdot\lambda_2$ richtig.

Diese Gleichung erhält man, wenn man in der Definitionsgleichung der Gamma-Funktion $y=\frac{x}{\lambda}$ substituiert: $\Gamma(p)=\int_0^\infty\left(x^{p-1}\cdot e^{-x}\right)dx=\lambda^{p-1}\cdot\int_0^\infty\left(x^{p-1}\cdot e^{-\lambda\cdot y}\cdot\lambda\right)dy$.

Dichtefunktion:

Eine Zufallsvariable X mit Gamma-Verteilung mit Parametern $\alpha>0$ und $\lambda>0$ wird über die Dichtefunktion

$f_X(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}\cdot x^{\alpha-1}\cdot e^{-\lambda\cdot x}$ für $x>0$ definiert.

Die obige Eigenschaft (v) der Gamma-Funktion zeigt, dass es sich um eine Dichtefunktion handelt.

Verteilungsfunktion:

$$F_X(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}\cdot\int_0^x\left(x^{\alpha-1}\cdot e^{-\lambda\cdot x}\right)dx$$

Erwartungswert: $\mathrm{E}[X]=\alpha/\lambda$ ***Varianz:*** $Var(X)=\alpha/\lambda^2$

Der Erwartungswert berechnet sich mit der Substitution $y=\lambda\cdot x$, d.h. $dy=\lambda\cdot dx$ zu

$$\mathrm{E}[X]=\frac{\lambda^\alpha}{\Gamma(\alpha)}\cdot\int_0^\infty\left(x^\alpha\cdot e^{-\lambda\cdot x}\right)dx=\frac{1}{\Gamma(\alpha)}\cdot\int_0^\infty\left((\lambda\cdot x)^\alpha\cdot e^{-\lambda\cdot x}\right)dx$$
$$=\frac{1}{\Gamma(\alpha)}\cdot\int_0^\infty\left(y^\alpha\cdot e^{-y}\cdot\frac{1}{\lambda}\right)dy=\frac{\Gamma(\alpha+1)}{\lambda\cdot\Gamma(\alpha)}=\frac{\alpha}{\lambda}\quad .$$

Die Varianz ergibt sich wieder mit Hilfe der charakteristischen Funktion (siehe unten) und des zweiten Moments (siehe Satz 7.5-2(ii)):

$$Var(X)=\mathrm{E}[X^2]-(\mathrm{E}[X])^2=\frac{\alpha\cdot(\alpha+1)}{\lambda^2}-\left(\frac{\alpha}{\lambda}\right)^2=\frac{\alpha}{\lambda^2}.$$

charakteristische Funktion:

$$cf_X(t) = \int_{-\infty}^{\infty} \left(e^{i\cdot t\cdot x} \cdot f_X(x)\right) dx \qquad \text{(wegen } f_X(x) = 0 \text{ für } x \le 0\text{)}$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \int_0^{\infty} \left(e^{i\cdot t\cdot x} \cdot x^{\alpha-1} \cdot e^{-\lambda\cdot x}\right) dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \int_0^{\infty} \left(x^{\alpha-1} \cdot e^{-(\lambda - i\cdot t)\cdot x}\right) dx$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha)}{(\lambda - i\cdot t)^\alpha}$$

$$= \frac{1}{\left(1 - \frac{i\cdot t}{\lambda}\right)^\alpha} \quad .$$

Die *k*-te Ableitung lautet $cf_X^{(k)}(t) = \frac{\alpha\cdot(\alpha+1)\cdot\ldots\cdot(\alpha+k-1)}{\lambda^k} \cdot i^k \cdot \frac{1}{\left(1 - \frac{i\cdot t}{\lambda}\right)^{\alpha+k}}$ für $k \in \mathbf{N}$.

Momente:

$$\mathrm{E}\left[X^k\right] = cf_X^{(k)}(0) / i^k = \frac{\alpha\cdot(\alpha+1)\cdot\ldots\cdot(\alpha+k-1)}{\lambda^k}.$$

Mit $\alpha = 1$ erhält man als Spezialfall einer Gamma-Verteilung die Exponentialverteilung.

Ist X_1 Gamma-verteilt mit Parametern α_1 und λ und ist X_2 Gamma-verteilt mit Parametern α_2 und λ und sind X_1 und X_2 stochastisch unabhängig, dann ist $Y = X_1 + X_2$ Gamma-verteilt mit Parametern $\alpha_1 + \alpha_2$ und λ: Die charakteristische Funktion von *Y* lautet nämlich:

$$cf_{X_1+X_2}(t) = \frac{1}{\left(1 - \frac{i\cdot t}{\lambda}\right)^{\alpha_1}} \cdot \frac{1}{\left(1 - \frac{i\cdot t}{\lambda}\right)^{\alpha_2}} = \frac{1}{\left(1 - \frac{i\cdot t}{\lambda}\right)^{\alpha_1+\alpha_1}} \quad .$$

Quadrierte Normalverteilung:

Modell:

Die Zufallsvariable ist definiert durch $X = Y^2$ mit einer $(0,1)$-normalverteilten Zufallsvariablen *Y*.

Dichtefunktion:

Nach Satz 7.2.5 (ii) lautet die Dichtefunktion von *X* für $x > 0$ $f_X(x) = \frac{f_Y\left(\sqrt{x}\right) + f_Y\left(-\sqrt{x}\right)}{2\cdot\sqrt{x}}$ mit

$f_Y(x) = \frac{1}{\sqrt{2\cdot\pi}} \cdot e^{-\frac{1}{2}\cdot x^2}$, also

$f_X(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sqrt{x}} \cdot e^{-\frac{1}{2} \cdot x}$. Das ist die Dichtefunktion der Gamma-Verteilung mit $\alpha = 1/2$ und $\lambda = 1/2$.

Erwartungswert: $\mathrm{E}[X] = 1$ ***Varianz:*** $Var(X) = 2$

charakteristische Funktion:

$$cf_X(t) = \frac{1}{\sqrt{1 - 2 \cdot i \cdot t}}$$

Momente:

$$\mathrm{E}[X^k] = cf_X^{(k)}(0) / i^k = 2^k \cdot 1/2 \cdot (1/2 + 1) \cdot \ldots \cdot (1/2 + k - 1) = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2 \cdot k - 1)$$

χ^2-Verteilung (chi-Quadrat-Verteilung):

Modell:

Es seien $X_1, \ldots, X_n$ stochastisch unabhängige jeweils $(0,1)$-normalverteilte Zufallsvariablen und $Y = X_1^2 + \ldots X_n^2 = \sum_{i=1}^{n} X_i^2$. Dann sind die Zufallsvariablen $X_1^2, \ldots, X_n^2$ jeweils Gamma-verteilt mit $\alpha = 1/2$ und $\lambda = 1/2$ (siehe oben). *Y* ist daher Gamma-verteilt mit $\alpha = n/2$ und $\lambda = 1/2$ (siehe Bemerkungen am Ende der Beschreibung der Gamma-Verteilung).

Die Verteilung von *Y*, die in der Anwendung der Stichprobentheorie genutzt wird, heißt **χ^2-Verteilung (chi-Quadrat-Verteilung) mit *n* Freiheitsgraden.**

Aus der Definition folgt unmittelbar folgende Aussage:

Ist Y_n χ^2-verteilt mit *n* Freiheitsgraden und Y_m χ^2-verteilt mit *m* Freiheitsgraden, so ist $Y_n + Y_m$ χ^2-verteilt mit $n + m$ Freiheitsgraden.

Dichtefunktion:

$$f_X(x) = \frac{\lambda^\alpha}{2^{n/2} \cdot \Gamma(n/2)} \cdot x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}} \text{ für } x > 0$$

Erwartungswert: $\mathrm{E}[X] = n$ ***Varianz:*** $Var(X) = 2 \cdot n$

charakteristische Funktion:

$$cf_X(t) = \frac{1}{(1-2\cdot i\cdot t)^{n/2}}$$

Momente:

$$\mathrm{E}\left[X^k\right] = cf_X^{(k)}(0)/i^k = n\cdot(n+2)\cdot(n+4)\cdot\ldots\cdot(n+2\cdot(k-1)).$$

STUDENT-t-Verteilung:

Modell:

Die Zufallsvariablen Y_n und Z seien stochastisch unabhängig. Y_n sei χ^2-verteilt mit n Freiheitsgraden, Z $(0,1)$-normalverteilt. Dann heißt die Zufallsvariable

$$T_n = \frac{Z}{\sqrt{\frac{1}{n}\cdot Y_n}}$$

t-verteilt mit *n* Freiheitsgraden.

Die t-Verteilung spielt eine wichtige Rolle in der Stichproben- und Schätztheorie.

Die folgenden Angaben werden nur der Vollständigkeit angeführt. Die Herleitung findet man beispielsweise in Fisz, M.: **Wahrscheinlichkeitsrechnung und mathematische Statistik**, 11. Aufl., Deutscher Verlag der Wissenschaften, 1989..

Dichtefunktion:

$$f_{T_n}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\cdot\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad \text{für } -\infty < x < \infty.$$

Erwartungswert: $\mathrm{E}[T_n] = 0$ ***Varianz:*** $Var(T_n) = \frac{n}{n-2}$

Die t-Verteilung ist symmetrisch zu 0 und ähnelt der Normalverteilung. Für große Werte von n, etwa $n > 30$, wird die Dichtefunktion gut durch die Dichtefunktion der (0, 1)-Normalverteilung approximiert. Die Werte der Verteilungsfunktion findet man in der Literatur tabelliert.

F-Verteilung:

Modell:

Die Zufallsvariable X_m sei sei χ^2-verteilt mit *m* Freiheitsgraden, die Zufallsvariable X_n sei sei χ^2-verteilt mit *n* Freiheitsgraden. Dann heißt die Zufallsvariable

$$F_n^m = \frac{\frac{1}{m} \cdot X_m}{\frac{1}{n} \cdot X_n}$$ **F-verteilt mit *m* und *n* Freiheitsgraden.**

Die F-Verteilung spielt eine wichtige Rolle in der Stichprobentheorie. Der Vollständigkeit halber werden die folgenden Werte angeführt:

Erwartungswert: $\mathrm{E}\left[F_n^m\right] = \frac{n}{n-2}$ bei $n > 2$

Varianz: $Var\left(F_n^m\right) = \frac{2 \cdot n^2 \cdot (m+n-2)}{m \cdot (n-2)^2 \cdot (n-4)}$ bei $n > 4$.

7.7 Grenzwertsätze

Grenzwertsätze spielen eine wichtige Rolle in der Anwendung der Stichproben- und Testtheorie. Bei den Grenzwertsätzen zeigt sich insbesondere die Bedeutung der Normalverteilung.

Gegeben sei eine *n*-dimensionale Zufallsvariable $(X_1, ..., X_n)$, deren einzelne Komponenten stochastisch unabhängig und identisch verteilt sind. Jedes X_i für $i = 1, ..., n$ habe den Erwartungswert $\mathrm{E}[X_i] = \mu$ und die Varianz $Var(X_i) = \sigma^2$. In der Praxis entsteht eine derartige Zufallsvariable in unterschiedlichen Situationen:

- Aus einer Grundgesamtheit wird eine Stichprobe vom Umfang *n* gezogen. Das Stichprobenergebnis $x_1, ..., x_n$ ist dann die Realisierung einer *n*-dimensionalen Zufallsvariablen $(X_1, ..., X_n)$. Die Grundgesamtheit hat dabei den Erwartungswert $\mathrm{E}[X] = \mu$ und die Varianz $Var(X) = \sigma^2$. Die stochastische Unabhängigkeit der einzelnen Komponenten wird dadurch gegeben, dass die Stichprobe mit Zurücklegen genommen wird, d.h. dass jedes gezogene Stichprobenelement x_i in die Gesamtheit zurückgelegt wird, bevor das nächste Stichprobenelement x_{i+1} ermittelt wird.

– Ein Zufallsexperiment, das eine Zufallsvariable X mit Erwartungswert $\mathrm{E}[X]=\mu$ und Varianz $Var(X)=\sigma^2$ testet, wird n-mal durchgeführt. Das Ergebnis jedes Experiments erzeugt die Zufallsvariable X_i für $i=1,...,n$.

Für die folgenden Betrachtungen werden die Bezeichnungen S_n und $\overline{X}_n$ definiert:

Die **Summe der Zufallsvariablen** $X_1,...,X_n$ ist die Zufallsvariable $S_n=\sum_{i=1}^{n} X_i$.

Das **arithmetische Mittel der Zufallsvariablen** $X_1,...,X_n$ ist die Zufallsvariable $\overline{X}_n=\frac{1}{n}\cdot\sum_{i=1}^{n} X_i$.

Der folgende Satz besagt, dass der Erwartungswert des arithmetischen Mittels der Stichprobe dem Erwartungswert der Grundgesamtheit entspricht und die Varianz des arithmetischen Mittels der Stichprobe mit dem Faktor $1/n$ kleiner ist als die Varianz der Grundgesamtheit.

Satz 7.7-1:

Die Zufallsvariablen $X_1,...,X_n$ seien stochastisch unabhängig und identisch verteilt mit Erwartungswert $\mathrm{E}[X_i]=\mu$ und Varianz $\mathrm{Var}(X_i)=\sigma^2$. Dann gilt:

$$\mathrm{E}[\overline{X}_n]=\mu \text{ und } \mathrm{Var}(\overline{X}_n)=\frac{1}{n}\cdot\sigma^2.$$

Die Aussagen ergeben sich wie folgt:

$$\mathrm{E}[\overline{X}_n]=\mathrm{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n} X_i\right]=\frac{1}{n}\cdot\sum_{i=1}^{n}\mathrm{E}[X_i]=\frac{1}{n}\cdot n\cdot\mu=\mu \text{ und (mit Satz 7.2-3(ii) und Satz 7.4-1)}$$

$$\mathrm{Var}(\overline{X}_n)=\mathrm{Var}\left(\frac{1}{n}\cdot\sum_{i=1}^{n} X_i\right)=\frac{1}{n^2}\cdot n\cdot\sigma^2=\frac{1}{n}\cdot\sigma^2.$$

Im Folgenden werden Folgen $(X_n)_{n\in\mathbf{N}}$ von Zufallsvariablen betrachtet und dann die Eigenschaften der zugehörigen Folge des arithmetischen Mittels beschrieben.

Der folgende Satz zeigt, dass mit wachsendem Stichprobenumfang das arithmetische Mittel der Stichprobe mit „sehr großer“ Wahrscheinlichkeit in ein kleines Intervall um den Erwartungswert des zu untersuchenden Merkmals fällt.

Satz 7.7-2: (Schwaches Gesetz der großen Zahlen)

(i) Die Zufallsvariablen der Folge $(X_n)_{n\in\mathbf{N}_{>0}}$ seien paarweise unkorreliert mit existierenden Erwartungswerten $\mathrm{E}[X_i]=\mu_i$ und Varianzen $\mathrm{Var}(X_i)=\sigma_i^2$. Es gelte weiterhin $\lim\limits_{n\to\infty}\dfrac{\sum\limits_{i=1}^{n}\sigma_i^2}{n^2}=0$. Dann gilt für jedes $\varepsilon>0$:

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n-\frac{\mu_1+\ldots+\mu_n}{n}\right|\geq\varepsilon\right)=0.$$

(ii) Die Zufallsvariablen der Folge $(X_n)_{n\in\mathbf{N}_{>0}}$ seien stochastisch unabhängig mit identischem Erwartungswert $\mathrm{E}[X_i]=\mu$ und existierenden Varianzen $\mathrm{Var}(X_i)=\sigma^2$. Dann gilt für jedes $\varepsilon>0$:

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n-\mu\right|\geq\varepsilon\right)=0.$$

(iii) Die Zufallsvariablen der Folge $(X_n)_{n\in\mathbf{N}_{>0}}$ seien stochastisch unabhängig mit identischer Verteilung und Erwartungswert $\mathrm{E}[X_i]=\mu$. Dann gilt für jedes $\varepsilon>0$:

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n-\mu\right|\geq\varepsilon\right)=0.$$

Bemerkungen: (1) Man beachte, dass Satz 7.7-2 (ii) ***nicht*** aussagt, dass $\overline{X}_n$ mit wachsendem *n* im Sinne der üblichen Konvergenz gegen μ konvergiert.

(2) Die Zufallsvariablen X_i müssen in (ii) nicht dieselbe Verteilung besitzen. Es werden lediglich identische Erwartungswerte und Varianzen vorausgesetzt.

(3) Teil (iii) ist anwendbar, wenn alle Zufallsvariablen X_i dieselbe Verteilung besitzen. Über die Existenz von Varianzen wird hierbei nichts vorausgesetzt.

Aussage (i) folgt aus Satz 7.2-4 (ii): Es ist $\mathrm{E}\left[\overline{X}_n\right]=\mathrm{E}\left[\dfrac{1}{n}\cdot\sum\limits_{i=1}^{n}X_i\right]=\dfrac{1}{n}\cdot\sum\limits_{i=1}^{n}\mathrm{E}[X_i]=\dfrac{\mu_1+\ldots+\mu_n}{n}$,

also $0\leq P\left(\left|\overline{X}_n-\dfrac{\mu_1+\ldots+\mu_n}{n}\right|\geq\varepsilon\right)\leq\dfrac{\mathrm{Var}\left(\overline{X}_n\right)}{\varepsilon^2}=\dfrac{\frac{1}{n^2}\cdot\sum\limits_{i=1}^{n}\sigma_i^2}{\varepsilon^2}$. Damit ist

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \frac{\mu_1 + \ldots + \mu_n}{n}\right| \geq \varepsilon\right) = 0 .$$

Die Aussage (ii) des Satzes ist ein Spezialfall von (i): Die stochastische Unabhängigkeit impliziert die Unkorreliertheit der Zufallsvariablen; die Bedingung $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \sigma_i^2}{n^2} = 0$ ist wegen $\sigma_i^2 = \sigma^2$ und $\frac{\sum_{i=1}^{n} \sigma_i^2}{n^2} = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$ erfüllt.

Den Beweis von (iii), der weitere Hilfsmittel aus der Analysis erfordert, findet man in Fisz, M.: **Wahrscheinlichkeitsrechnung und mathematische Statistik**, 11. Aufl., Deutscher Verlag der Wissenschaften, 1989..

Man sagt, eine Folge $(Z_n)_{n\in\mathbf{N}}$ von Zufallsvariablen **konvergiert stochastisch gegen 0**, wenn für jedes $\varepsilon > 0$ die Beziehung $\lim_{n\to\infty} P(|Z_n| \geq \varepsilon) = 0$ gilt. Gleichbedeutend damit ist $\lim_{n\to\infty} P(|Z_n| < \varepsilon) = 1$.

Eine Folge $(Z_n)_{n\in\mathbf{N}}$ von Zufallsvariablen **konvergiert fast überall (mit Wahrscheinlichkeit 1) gegen 0**, wenn $P(\lim_{n\to\infty} Z_n = 0) = 1$ gilt.

Das schwache Gesetz großer Zahlen in Satz 7.7-2 (i) besagt demnach, dass $\left(\overline{X}_n - \frac{\mu_1 + \ldots + \mu_n}{n}\right)_{n\in\mathbf{N}_{>0}}$ stochastisch gegen 0 konvergiert. In Satz 7.7-2 (ii) wird ausgesagt, dass $(\overline{X}_n - \mu)_{n\in\mathbf{N}_{>0}}$ stochastisch gegen 0 konvergiert.

Der folgende Satz wird hier nur zitiert.

Satz 7.7-3:

Konvergiert eine Folge $(Z_n)_{n\in\mathbf{N}}$ von Zufallsvariablen fast überall gegen 0, so konvergiert sie stochastisch gegen 0, d.h. $P(\lim_{n\to\infty} Z_n = 0) = 1$ impliziert $\lim_{n\to\infty} P(|Z_n| \geq \varepsilon) = 0$ für jedes $\varepsilon > 0$.

Die Umkehrung gilt nicht, wie das Beispiel der Folge $(Z_n)_{n\in\mathbf{N}_{>0}}$ mit

$$Z_n = \begin{cases} 1 & \text{mit Wahrscheinlichkeit } 1/n \\ 0 & \text{mit Wahrscheinlichkeit } 1-1/n \end{cases}$$

zeigt: Es sei $\varepsilon > 0$. Dann ist $P(|Z_n| \geq \varepsilon) = P(Z_n = 1) = 1/n$, also $\lim_{n\to\infty} P(|Z_n| \geq \varepsilon) = 0$, d.h. $(Z_n)_{n\in\mathbf{N}_{>0}}$ konvergiert stochastisch gegen 0. Aber mit $A_n = \{Z_n = 1\}$ ist $P(A_n) = 1/n$ und $\sum_{i=1}^{\infty} P(A_n) = \infty$. Es lässt sich zeigen, dass die Wahrscheinlichkeit des Eintretens von unendlich vielen Ereignissen A_n gleich 1 ist. Die Wahrscheinlichkeit für die Existenz einer Teilfolge von $(Z_n)_{n\in\mathbf{N}_{>0}}$, die nicht gegen 0 konvergiert, ist somit gleich 1, im Widerspruch zu $P(\lim_{n\to\infty} Z_n = 0) = 1$.

Auch auf den Beweis des folgenden Satzes wird mit Hinweis auf die angegebene Literatur verzichtet.

Satz 7.7-4: (Starkes Gesetz großer Zahlen)

Die Zufallsvariablen der Folge $(X_n)_{n\in\mathbf{N}_{>0}}$ seien unabhängig mit existierenden Erwartungswerten $\mathrm{E}[X_i] = \mu_i$ und Varianzen $\mathrm{Var}(X_i) = \sigma_i^2$, und es gelte $\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty$. Dann gilt $P\left(\lim_{n\to\infty} \left| \overline{X}_n - \frac{\mu_1 + \ldots + \mu_n}{n} \right| = 0\right) = 1$.

Satz 7.7-4 besagt also, dass unter den gegebenen Voraussetzungen, insbesondere wenn die Varianzen „nicht zu schnell“ wachsen (hier $\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty$), die Folge der Zufallsvariablen $\left(\overline{X}_n - \frac{\mu_1 + \ldots + \mu_n}{n}\right)_{n\in\mathbf{N}}$ nicht nur stochastisch gegen 0 (schwaches Gesetz großer Zahlen), sondern sogar fast überall gegen 0 (starkes Gesetz großer Zahlen) konvergiert.

Ein Experiment wird *n*-mal hintereinander durchgeführt. Bei jeder Ausführung kann ein Treffer auftreten (Wert 1) oder nicht (Wert 0). Das Ergebnis des *i*-ten Experiments ist eine Zufallsvariable X_i. Die Trefferwahrscheinlichkeit des *i*-ten Experiments hänge nicht von den Ausgängen der vorherigen Experimente ab; in diesem Fall sind die einzelnen Zufallsvariablen der Folge $(X_n)_{n\in\mathbf{N}_{>0}}$ stochastisch unabhängig. In jedem Experiment sei die Trefferwahrscheinlichkeit gleich p, $\mathrm{E}[X_i] = p$. Man nennt dieses Experiment auch **Bernoullisches Versuchs-**

schema. Dann ist die Zufallsvariable $S_n = \sum_{i=1}^{n} X_i$, die die Anzahl der Treffer nach *n*-maliger Versuchsausführung beschreibt, binomialverteilt (siehe Kapitel 7.6) mit Erwartungswert $\mathrm{E}[S_n] = n \cdot p$ und Varianz $\mathrm{Var}(S_n) = n \cdot p \cdot (1-p)$. Die Zufallsvariable $\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ beschreibt dann die relative Trefferhäufigkeit der Versuchsreihe. Es ist $\mathrm{E}[\overline{X}_n] = p$ und Varianz $\mathrm{Var}(\overline{X}_n) = p \cdot (1-p)/n$.

Das schwache Gesetz großer Zahlen (Satz 7.7-2 (ii)) besagt nun, dass die Wahrscheinlichkeit, dass mit wachsender Versuchsanzahl die relative Trefferhäufigkeit dicht bei *p* liegt, gegen 1 konvergiert, genauer: für jedes $\varepsilon > 0$ ist $\lim_{n\to\infty} P(|\overline{X}_n - p| < \varepsilon) = 1$. Das starke Gesetz großer Zahlen (Satz 7.7-4) besagt, dass die relative Trefferhäufigkeit mit wachsender Versuchsanzahl „mit Sicherheit“ gegen *p* konvergiert: $P(\lim_{n\to\infty} |\overline{X}_n - p| = 0) = P(\lim_{n\to\infty} \overline{X}_n = p) = 1$.

Es sei $p = 1/2$. Wie hoch muss die Versuchsanzahl liegen, damit mit 99%-iger Sicherheit die relative Trefferhäufigkeit zwischen 0,45 und 0,55 liegt? Die Antwort liefert Satz 7.7-2 (ii) bzw. Satz 7.7-4 (ii): $P(|\overline{X}_n - 1/2| \geq 0{,}05) \leq \frac{0{,}5 \cdot 0{,}5}{n \cdot 0{,}05^2} \leq 0{,}01$ impliziert $n \geq 10.000$.

Die Realisierung einer Zufallsvariablen *X* mit Verteilungsfunktion F_X werde *n*-mal beobachtet, d.h. man nimmt eine Stichprobe $x_1, ..., x_n$ vom Umfang *n*, wobei die einzelnen Stichprobenwerte unabhängig voneinander ermittelt werden. Dabei kann es vorkommen, dass manche Stichprobenwerte gleich sind. Die Stichprobenwerte werden der Größe nach aufsteigend geordnet, so dass $x_1 \leq x_2 \leq ... \leq x_n$ angenommen werden kann. Die **empirische Verteilungsfunktion** $H_n(x)$ wird definiert durch

$$H_n(x) = \begin{cases} 0 & \text{für } x < x_1 \\ \dfrac{m}{n} & \text{für } x \geq x_1; \text{ dabei ist } m \text{ der größte Index mit } x_m \leq x \end{cases}.$$

Dann ist $n \cdot H_n(x)$ gleich der Anzahl der Werte x_i mit $x_i \leq x$.
Die Funktion $H_n(x)$ ist eine Treppenfunktion: Sind $x_{i_1}, ..., x_{i_k}$ mit $i_1 = 1$ und $x_1 = x_{i_1} < x_{i_2} < ... < x_{i_k}$ die *unterschiedlichen* Werte in der Stichprobe und $h_{i_1}, ..., h_{i_k}$ die Häufigkeiten der Werte $x_{i_1}, ..., x_{i_k}$, dann ist

$$H_n(x) = \begin{cases} 0 & \text{für } x < x_{i_1} = x_1 \\ \sum_{j=1}^{l} \frac{h_{i_j}}{n} & \text{für } x_{i_l} \le x < x_{i_{l+1}}, 1 \le l < k \\ 1 & \text{für } x \ge x_{i_k} \end{cases} .$$

Es wird zusätzlich angenommen, dass F_X eine stetige Verteilungsfunktion ist. Dann ist die Wahrscheinlichkeit, dass zwei Stichprobenwerte gleich sind, gleich 0: Der *i*-te Stichprobenwert für $i = 1, ..., n$ werde durch die Zufallsvariable X_i beschrieben. Wegen der Unabhängigkeit der Erhebung der Stichprobenwerte und der Stetigkeit der zugrundeliegenden Verteilung ist $P(X_i = x \text{ und } X_j = x) = P(X_i = x) \cdot P(X_j = x) = 0 \cdot 0 = 0$. Es kann also für alle Stichprobenwerte $x_1 < x_2 < ... < x_n$ angenommen werden. Es sei *x* ein fester Wert. Dann ist
$P(X_i \le x) = F_X(x)$ für $i = 1, ..., n$ und

$$P(n \cdot H_n(x) = m) = \binom{n}{m} \cdot (F_X(x))^m \cdot (1 - F_X(x))^{n-m} .$$

Hierbei handelt es sich um das oben beschriebene Bernoullische Versuchsschema, in dem Treffer („$X_i \le x$“) und Nichttreffer gezählt werden; die Trefferwahrscheinlichkeit beträgt $F_X(x)$; die Zufallsvariable $n \cdot H_n(x)$ beschreibt die Anzahl der Treffer. Für die relative Trefferhäufigkeit der Versuchsreihe $1/n \cdot (n \cdot H_n(x)) = H_n(x)$ gilt dann mit dem starken Gesetz großer Zahlen (siehe oben):
$P\left(\lim_{n\to\infty} |H_n(x) - F_X(x)| = 0\right) = P\left(\lim_{n\to\infty} H_n(x) = F_X(x)\right) = 1$.

Der folgende Satz besagt, dass diese Beziehung *gleichmäßig* für alle $x \in \mathbf{R}$ gilt, wenn die Anzahl der Stichprobenelemente über alle Grenzen wächst. Das bedeutet, dass man bei einer genügend großen Stichprobe mit Wahrscheinlichkeit 1 aus der empirischen Verteilungsfunktion eine ausführliche Information über die gesamte Verteilungsfunktion $F_X(x)$ erhält.

Satz 7.7-5: (Hauptsatz der Statistik)

Es werde eine Stichprobe $x_1, ..., x_n$ vom Umfang *n* zu einem Merkmal *X* erhoben, das die Verteilungsfunktion F_X besitzt. $H_n(x)$ bezeichne die empirische Verteilungsfunktion zur Stichprobe, d.h. $n \cdot H_n(x)$ ist gleich der Anzahl der Werte x_i mit $x_i \le x$. Es sei $\Delta_n = \sup\{|H_n(x) - F_X(x)| \mid -\infty < x < \infty\}$. Dann gilt:
$P\left(\lim_{n\to\infty} \Delta_n = 0\right) = 1$.

Zur Erinnerung: Das Supremum einer Zahlenmenge ist ihre kleinste obere Schranke.

In den Gesetzen der großen Zahlen (Satz 7.7-2 und Satz 7.7-4) wird ausgesagt, dass sich der Erwartungswert des arithmetischen Mittels von *n* unabhängigen identisch verteilten Zufallsvariablen $X_1, ..., X_n$ in gewisser Weise um den Erwartungswert der gemeinsamen Verteilung gruppiert, je größer die Versuchsreihe oder der Stichprobenumfang ist. Über die Verteilungsfunktion des arithmetischen Mittels oder der Summe der Zufallsvariablen wird jedoch nichts ausgesagt. Es ist auch nicht immer einfach, aus $X_1, ..., X_n$ die Verteilungsfunktion von $\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ und $S_n = \sum_{i=1}^{n} X_i$ zu ermitteln. Ist zudem die Verteilung der X_i unbekannt, so ist die Berechnung der Verteilung des arithmetischen Mittels oder der Summe der Zufallsvariablen unmöglich. Der folgende Satz schafft hier Abhilfe. Er besagt nämlich, dass die Verteilung des arithmetischen Mittels oder der Summe der Zufallsvariablen gegen die Normalverteilung konvergiert. Hier zeigt sich die besondere Rolle, die der Normalverteilung in Theorie und Praxis zukommt.

Satz 7.7-6: (Zentraler Grenzwertsatz)

Es seien $X_1, ..., X_n$ stochastisch unabhängige und identisch verteilte Zufallsvariablen mit Erwartungswert $\mathrm{E}[X_i] = \mu$ und Varianz $\mathrm{Var}(X_i) = \sigma^2$. Mit $F_n(z)$ werde die Verteilungsfunktion der Zufallsvariablen

$$Z_n = \frac{S_n - n \cdot \mu}{\sigma \cdot \sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}}$$

bezeichnet. Dann gilt:

$$\lim_{n \to \infty} F_n(z) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^{z} e^{-\frac{1}{2} \cdot t^2} \, dt .$$

Die Beweisidee für diesen erstaunlichen Sachverhalt soll skizziert werden:

Es ist $Z_n = \frac{S_n - n \cdot \mu}{\sigma \cdot \sqrt{n}} = \frac{1}{\sigma \cdot \sqrt{n}} \cdot \sum_{i=1}^{n} (X_i - \mu)$. Die durch $Y_i = X_i - \mu$ definierten Zufallsvariablen sind identisch verteilt mit $\mathrm{E}[Y_i] = 0$ und $\mathrm{Var}(Y_i) = \mathrm{E}[Y_i^2] = \sigma^2$; ihre charakteristische Funktion sei $cf_{Y_i}(t)$. Nach Satz 7.5-2 (ii) und (iv) lautet die charakteristische Funktion von Z_n:

$$cf_{Z_n}(t) = \left(cf_{Y_i}\left(\frac{t}{\sigma \cdot \sqrt{n}} \right) \right)^n .$$

Es ist $cf_{Y_i}(t) = \mathrm{E}[e^{i \cdot t \cdot Y_i}]$. Im stetigen Fall (mit Dichtefunktion f_{Y_i}) ist $cf_{Y_i}(t) = \int_{-\infty}^{\infty} \left(e^{i \cdot t \cdot x} \cdot f_{Y_i}(x) \right) dx$.

Im diskreten Fall gilt eine entsprechende Formel. Die Reihenentwicklung von $e^{i \cdot t \cdot x}$ lautet

$e^{i\cdot t\cdot x}=\sum_{n=0}^{\infty}\frac{(i\cdot t\cdot x)^n}{n!}=1+i\cdot t\cdot x-\frac{(t\cdot x)^2}{2}+R_3(i\cdot t\cdot x)$. Damit ist (denn man darf die einzelnen Operationen Integral- und Reihenbildung hier vertauschen)

$$cf_{Y_i}(t)=\int_{-\infty}^{\infty}\left(\left(1+i\cdot t\cdot x-\frac{(t\cdot x)^2}{2}+R_3(i\cdot t\cdot x)\right)\cdot f_{Y_i}(x)\right)dx$$

$$=1+i\cdot t\cdot \mathrm{E}[Y_i]-\frac{t^2}{2}\cdot \mathrm{E}[Y_i^2]+h(t)$$

$$=1-\frac{t^2}{2}\cdot\sigma^2+h(t)\quad .$$

Hierbei ist $h(t)$ eine Funktion mit der Eigenschaft $\lim_{t\to 0}\frac{h(t)}{t^2}=0$.

Setzt man diesen Wert in $cf_{Z_n}(t)$ ein, so erhält man:

$$cf_{Z_n}(t)=\left(cf_{Y_i}\left(\frac{t}{\sigma\cdot\sqrt{n}}\right)\right)^n=\left(1-\frac{t^2}{2\cdot n}+h\left(\frac{t}{\sigma\cdot\sqrt{n}}\right)\right)^n \text{ mit } \lim_{t/(\sigma\cdot\sqrt{n})\to 0}\frac{h\left(\frac{t}{\sigma\cdot\sqrt{n}}\right)}{t^2/(\sigma^2\cdot n)}=0.$$

Für festes *t* ist daher $\lim_{n\to\infty} n\cdot h\left(\frac{t}{\sigma\cdot\sqrt{n}}\right)=0$. Setzt man $u=-\frac{t^2}{2\cdot n}+h\left(\frac{t}{\sigma\cdot\sqrt{n}}\right)$, so ist $\ln(cf_{Z_n}(t))=n\cdot\ln(1+u)$. Die Reihenentwicklung für $\ln(1+u)$ (siehe Kapitel 5.9; diese gilt auch für komplexe Zahlen) liefert „für genügend kleine Werte“ von $t/(\sigma\cdot\sqrt{n})$:

$\ln(cf_{Z_n}(t))=n\cdot\left(-\frac{t^2}{2\cdot n}+h\left(\frac{t}{\sigma\cdot\sqrt{n}}\right)+g(t)\right)$ mit $\lim_{n\to\infty} n\cdot g(t)=0$ (hier argumentiert man ähnlich wie mit *h*). Damit folgt $\lim_{n\to\infty}\ln(cf_{Z_n}(t))=-t^2/2$ bzw. $\lim_{n\to\infty} cf_{Z_n}(t)=e^{-t^2/2}$. Die rechte Seite ist die charakteristische Funktion einer normiert (0, 1)-normalverteilten Zufallsvariable (siehe Kapitel 7.6). Mit Satz 7.5-5 (ii) ergibt sich die Aussage des Satzes 7.7-6.

Satz 7.7-6 lässt sich umformulieren zu

Satz 7.7-7:

Es seien $X_1,...,X_n$ stochastisch unabhängige und identisch verteilte Zufallsvariablen mit Erwartungswert $\mathrm{E}[X_i]=\mu$ und Varianz $\mathrm{Var}(X_i)=\sigma^2$. Für $s_1\in\mathbf{R}$ und $s_2\in\mathbf{R}$ gelte $s_1<s_2$. Dann gilt:

$$\lim_{n\to\infty}P(s_1<S_n<s_2)=\frac{1}{\sqrt{2\cdot\pi}}\cdot\int_{z_1}^{z_2}e^{-\frac{1}{2}\cdot t^2}\,dt \text{ mit } z_1=\frac{s_1-n\cdot\mu}{\sigma\cdot\sqrt{n}} \text{ und } z_2=\frac{s_2-n\cdot\mu}{\sigma\cdot\sqrt{n}}.$$

Es ist nämlich $P(s_1<S_n<s_2)=P\left(z_1<\frac{S_n-n\cdot\mu}{\sigma\cdot\sqrt{n}}<z_2\right)$ und mit Satz 7.7-6

$$\lim_{n\to\infty} P\left(z_1 < \frac{S_n - n\cdot\mu}{\sigma\cdot\sqrt{n}} < z_2 \right) = \lim_{n\to\infty}\left(F_n(z_2) - F_n(z_1)\right) = \frac{1}{\sqrt{2\cdot\pi}} \cdot \int_{z_1}^{z_2} e^{-\frac{1}{2}\cdot t^2}\, dt \,.$$

7.8 Punktschätzungen

In einer Grundgesamtheit soll die Verteilung eines Merkmals *X* ermittelt werden. Häufig ist die Grundgesamtheit sehr groß, oder die vollständige Information über das Merkmal kann nur über eine Totalerhebung ermittelt werden. Dann begnügt man sich mit einer Stichprobe von kleinerem Umfang, um auf die Charakteristika des Merkmals zu schließen. Es gibt eine Reihe von Methoden, die beschreiben, wie Stichproben zu erheben sind, um die gewünschten Schlüsse über das Merkmal ziehen zu können. Dazu gehören insbesondere Zufallsstichproben, die durch das Urnenmodell (Ziehen mit oder ohne Zurücklegen) beschrieben werden.

Im Folgenden werden Methoden beschrieben, um stochastische Parameter wie Erwartungswert und Varianz eines metrischen Merkmals *X* zu schätzen.

Der Parameter *u* des Merkmals *X* sei unbekannt und soll durch eine Zufallsstichprobe geschätzt werden. Es wird dazu eine Zufallsstichprobe vom Umfang *n* gemäß dem **Urnenmodell mit Zurücklegen** genommen und aus den beobachteten Werten $x_1, ..., x_n$ auf den Wert des zu untersuchenden Parameters *u* geschlossen. Hierbei wird jedes einzelne Stichprobenelement nach seiner Entnahme aus der Grundgesamtheit zurückgelegt, so dass es potenziell auch bei der nächsten Entnahme von Stichprobenelementen ausgewählt werden kann. Man hat es also hier mit einem *n*-dimensionalen Zufallsvektor $(X_1, ..., X_n)$ zu tun, dessen einzelne Komponenten stochastisch unabhängig und identisch verteilt sind. Mit Hilfe des Zufallsvektors $(X_1, ..., X_n)$ bzw. der Beobachtung $x_1, ..., x_n$ wird die **Schätzfunktion** $U_n = g(X_1, ..., X_n)$ mit einer „geeigneten" Funktion *g* zur Schätzung des Parameters definiert. Eine einzelne **Punktschätzung** für den zu untersuchenden Parameter *u* ist dann der aus der Beobachtung gewonnene Wert $g(x_1, ..., x_n)$. Dieser wird meistens vom wahren Wert *u* abweichen. Die Funktion *g* muss so gewählt werden, dass trotzdem akzeptable Schätzungen erreicht werden können.

Durch Vergrößerung des Stichprobenumfangs *n* erhält man eine Folge $(U_n)_{n\in\mathbf{N}}$ von Schätzfunktionen für den zu untersuchenden Parameter *u*. Diese heißt **konsistent**, wenn $\lim_{n\to\infty} P(|U_n - u| \geq \varepsilon) = 0$ für jedes $\varepsilon > 0$ gilt. Mit wachsendem Stichprobenumfang verringert sich die Wahrscheinlichkeit, dass die Schätzung vom wahren Parameterwert abweicht, bzw. es steigt die Wahrscheinlichkeit der Genauigkeit der Schätzung.

Die Schätzfunktion U_n heißt **erwartungstreu**, wenn $\mathrm{E}[U_n] = u$ gilt. „Im Mittel“ (genommen über alle Stichproben $x_1, ..., x_n$) wird also für eine erwartungstreue Schätzfunktion der wahre Wert getroffen.

Die Folge $(U_n)_{n \in \mathbf{N}}$ von Schätzfunktionen heißt **asymptotisch erwartungstreu**, wenn $\lim_{n \to \infty} \mathrm{E}[U_n] = u$ gilt.

Als **Schätzfehler** wird der Wert $g(X_1, ..., X_n) - u$ definiert. Der Erwartungswert des Schätzfehlers heißt **Verzerrung (Bias)**. Eine Schätzfunktion mit Verzerrung 0, z.B. eine erwartungstreue Schätzung, heißt **unverzerrt**; eine Schätzfunktion mit einer Verzerrung, die ungleich 0 ist, heißt **verzerrt**.

Eine erwartungstreue (unverzerrte) Schätzfunktion U heißt **wirksamer** als eine erwartungstreue (unverzerrte) Schätzfunktion V, wenn $\mathrm{Var}(U) < \mathrm{Var}(V)$ ist.

Punktschätzung für den Erwartungswert $\mathrm{E}[X] = \mu$:

Der Erwartungswert sei unbekannt und soll durch eine Zufallsstichprobe vom Umfang n gemäß dem Urnenmodell mit Zurücklegen geschätzt werden. Aus den beobachteten Werte $x_1, ..., x_n$ wird als Punktschätzung für den Erwartungswert $\hat{\mu} = g(x_1, ..., x_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ berechnet, d.h. die Schätzfunktion für den Erwartungswert lautet $U_n = \overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$.

Gemäß Satz 7.7-1 ist $\mathrm{E}[\overline{X}_n] = \mu$, d.h. die Schätzfunktion $\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ ist erwartungstreu.

Nach Satz 7.7-2(i) ist die Folge $(\overline{X}_n)_{n \in \mathbf{N}}$ konsistent. Das bedeutet, dass mit steigendem Stichprobenumfang die Genauigkeit der Schätzung mit hoher Wahrscheinlichkeit zunimmt.

Die Varianz der Schätzfunktion ist $\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{1}{n} \cdot \sigma^2$.

Punktschätzung für die Varianz $\mathrm{Var}(X)=\sigma^2$ bei bekanntem Erwartungswert $\mathrm{E}[X]=\mu$:

Aus den beobachteten Werte $x_1,...,x_n$ der Stichprobe wird als Punktschätzung für die Varianz der Wert $\hat{\sigma}^2 = g(x_1,...,x_n) = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\mu)^2$ berechnet, d.h. die Schätzfunktion lautet $U_n = \frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\mu)^2$.

Die Schätzfunktion ist erwartungstreu:
$\mathrm{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\mu)^2\right] = \frac{1}{n}\cdot\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\mu)^2\right] = \frac{1}{n}\cdot n\cdot\sigma^2 = \sigma^2$.

Weiterhin gilt $\lim_{n\to\infty} P\left(\left|\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\mu)^2-\sigma^2\right|\geq\varepsilon\right)=0$ für jedes $\varepsilon>0$:

Es ist $\sigma^2 = \mathrm{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\mu)^2\right]$. Daher gilt mit Satz 7.2-4(ii) und wegen der stochastischen Unabhängigkeit der Zufallsvariablen $(X_i-\mu)^2$ mit $v=\mathrm{Var}\left((X_i-\mu)^2\right)$:

$$P\left(\left|\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\mu)^2-\sigma^2\right|\geq\varepsilon\right)\leq\frac{\mathrm{Var}\left(\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\mu)^2\right)}{\varepsilon^2}=\frac{\frac{1}{n^2}\cdot\sum_{i=1}^{n}\mathrm{Var}\left((X_i-\mu)^2\right)}{\varepsilon^2}=\frac{v}{n\cdot\varepsilon^2}.$$

Daher ist die Folge $\left(\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\mu)^2\right)_{n\in\mathbf{N}}$ konsistent.

Die Voraussetzung über die Kenntnis des Erwartungswerts ist jedoch in der Praxis unrealistisch. Daher wird auch dieser geschätzt.

Punktschätzung für die Varianz $\mathrm{Var}(X)=\sigma^2$ bei unbekanntem Erwartungswert $\mathrm{E}[X]=\mu$:

Aus den beobachteten Werte $x_1,...,x_n$ der Stichprobe wird als Punktschätzung für die Varianz der Wert $\hat{\sigma}^2 = g(x_1,...,x_n) = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\hat{\mu})^2$ mit $\hat{\mu}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i$ berechnet, d.h. die Schätzfunktion lautet $U_n = \frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$.

Diese Schätzfunktion ist nicht erwartungstreu:

Es ist $\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\mu\right)^2-\left(\overline{X}_n-\mu\right)^2$, denn

$$\begin{aligned}\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\mu\right)^2-\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2&=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i^{\,2}-2\cdot X_i\cdot\mu+\mu^2-X_i^{\,2}+2\cdot X_i\cdot\overline{X}_n-\overline{X}_n^{\,2}\right)\\&=-2\cdot\mu\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}X_i+\mu^2+2\cdot\overline{X}_n\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}X_i-\overline{X}_n^{\,2}\\&=\overline{X}_n^{\,2}-2\cdot\mu\cdot\overline{X}_n+\mu^2\\&=\left(\overline{X}_n-\mu\right)^2\quad.\end{aligned}$$

Damit ergibt sich

$$\begin{aligned}\mathrm{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2\right]&=\mathrm{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right]-\mathrm{E}\left[\left(\overline{X}_n-\mu\right)^2\right]\\&=\frac{1}{n}\cdot\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i-\mu\right)^2\right]-\mathrm{E}\left[\left(\overline{X}_n-\mu\right)^2\right]\\&=\sigma^2-\mathrm{Var}\left(\overline{X}_n\right)\qquad\text{wegen }\mathrm{Var}\left(X_i\right)=\sigma^2\text{ und }\mu=\mathrm{E}\left[\overline{X}_n\right]\\&=\sigma^2-1/n\cdot\sigma^2\qquad\text{mit Satz 7.7 -1}\\&=\frac{n-1}{n}\cdot\sigma^2\quad.\end{aligned}$$

Das bedeutet, dass durch die Schätzfunktion $\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2$ die Varianz $\mathrm{Var}\left(X\right)=\sigma^2$ systematisch mit Verzerrung $-1/n\cdot\sigma^2$ unterschätzt wird. Die Schätzung ist jedoch asymptotisch erwartungstreu.

Eine erwartungstreue Schätzung für die Varianz bei unbekanntem Erwartungswert erhält man durch die Schätzfunktion $\frac{n}{n-1}\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2=\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2$ bzw. durch die aus der Stichprobe ermittelten Punktschätzung $\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right)^2$ mit $\hat{\mu}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i$.

Auch hier lässt sich die Konsistenz der Schätzfunktion zeigen.

Punktschätzung für den unbekannten Anteilswert *p* eines Merkmals in der Grundgesamtheit:

Das Merkmal *X* trete mit einem unbekannten Anteil *p* in der Grundgesamtheit auf. In einer Stichprobe $x_1,...,x_n$ wird gezählt, wie häufig das Merkmal *X* vorkommt. Als Schätzung für *p*

wird $\frac{1}{n}\cdot\sum_{i=1}^{n} x_i$ genommen, d.h. die Schätzfunktion lautet $\frac{1}{n}\cdot\sum_{i=1}^{n} X_i$, wobei $\mathrm{E}[X_i]=p$ und $\mathrm{Var}(X_i)=p\cdot(1-p)$ gilt.

Diese Schätzung ist erwartungstreu: $\mathrm{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n} X_i\right]=\frac{1}{n}\cdot n\cdot\mathrm{E}[X_i]=p$. Außerdem ist sie konsistent.

Die Grundgesamtheit enthalte *N* Elemente, die Stichprobe enthält *n* Elemente. Ist *N* gegenüber *n* groß (in einem noch zu präzisierenden Sinn, etwa, wenn *N* unendlich groß ist), so ist das Urnenmodell mit Zurücklegen gerechtfertigt; die Unabhängigkeit der einzelnen Stichprobenelemente kann wegen des Zurücklegens angenommen werden. Ist jedoch *N* gegenüber *n* klein oder auch endlich, so stellt das Urnenmodell mit Zurücklegen nur eine Annäherung an eine unabhängige Stichprobenerhebung dar. Es sollen daher die Urnenmodelle mit und ohne Zurücklegen zur Schätzung des Erwartungswerts μ des Merkmals *X* verglichen werden. Die Grundgesamtheit sei endlich, etwa $\{a_1,...,a_N\}$.

Eine Stichprobe $x_1,...,x_n$ im Urnenmodell mit Zurücklegen ist die Realisierung des Zufallsvektors $(X_1,...,X_n)$, dessen Komponenten unabhängig sind. Als Schätzfunktion für μ werde $\overline{X}_n=\frac{1}{n}\cdot\sum_{i=1}^{n} X_i$ mit dem Schätzwert $\hat{\mu}_x=\frac{1}{n}\cdot\sum_{i=1}^{n} x_i$ genommen. Eine Stichprobe $z_1,...,z_n$ im Urnenmodell ohne Zurücklegen ist die Realisierung eines Zufallsvektors $(Z_1,...,Z_n)$, dessen Komponenten jetzt jedoch nicht mehr unabhängig sind. Als Schätzfunktion für μ werde auch in diesem Modell $\overline{Z}_n=\frac{1}{n}\cdot\sum_{i=1}^{n} Z_i$ mit dem Schätzwert $\hat{\mu}_z=\frac{1}{n}\cdot\sum_{i=1}^{n} z_i$ genommen.

Jedes Element der Grundgesamtheit habe die gleiche Wahrscheinlichkeit $1/N$, um beim Ziehen in die Stichprobe zu gelangen. Es gilt

$$\mu=\mathrm{E}[X]=\frac{1}{N}\cdot\sum_{i=1}^{N} a_i\,. \qquad \text{(Formel (i))}$$

Die Varianz von *X* ist

$$\sigma^2=\mathrm{Var}(X)=\frac{1}{N}\cdot\sum_{i=1}^{N}(a_i-\mu)^2=\frac{1}{N}\cdot\sum_{i=1}^{N} a_i^2-\mu^2=\mathrm{E}[X^2]-\mu^2\,. \qquad \text{(Formel (ii))}$$

Im Urnenmodell mit Zurücklegen ist die Varianz der Stichprobe gleich

$$\mathrm{Var}(\overline{X}_n)=\frac{1}{n}\cdot\sigma^2\,. \qquad \text{(Formel (iii))}$$

Im Urnenmodell ohne Zurücklegen ist die Varianz der Stichprobe gleich

$$\mathrm{Var}\left(\overline{Z}_n\right)=\mathrm{E}\left[\left(\overline{Z}_n-\mathrm{E}\left[\overline{Z}_n\right]\right)^2\right]=\mathrm{E}\left[\overline{Z}_n{}^2\right]-\left(\mathrm{E}\left[\overline{Z}_n\right]\right)^2. \qquad \text{(Formel (iv))}$$

Diese wird wie folgt ermittelt.

Der Erwartungswert von $\overline{Z}_n$ ist

$$\mathrm{E}\left[\overline{Z}_n\right]=\frac{1}{n}\cdot\sum_{i=1}^{n}\mathrm{E}\left[Z_i\right]=\mu. \qquad \text{(Formel (v))}$$

$$\mathrm{E}\left[\overline{Z}_n{}^2\right]=\mathrm{E}\left[\left(\frac{1}{n}\cdot\sum_{i=1}^{n}Z_i\right)^2\right]=\mathrm{E}\left[\frac{1}{n^2}\cdot\sum_{i=1}^{n}Z_i^2\right]+\frac{2}{n^2}\cdot\mathrm{E}\left[\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}Z_i\cdot Z_k\right]. \qquad \text{(Formel (vi))}$$

Hierbei wird die durch vollständige Induktion über die Elementanzahl *n* zu beweisende Identität $\left(\sum_{i=1}^{n}b_i\right)^2=\sum_{i=1}^{n}b_i^2+2\cdot\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}b_i\cdot b_k$ verwendet.

Der erste Summand auf der rechten Seite in Formel (vi) ist

$$\mathrm{E}\left[\frac{1}{n^2}\cdot\sum_{i=1}^{n}Z_i^2\right]=\frac{1}{n^2}\cdot\sum_{i=1}^{n}\mathrm{E}\left[Z_i^2\right]=\frac{1}{n^2}\cdot\sum_{i=1}^{n}\mathrm{E}\left[X^2\right]=\frac{1}{n}\cdot\left(\sigma^2+\mu^2\right). \qquad \text{(Formel (vii))}$$

Aus Formel (i) folgt $\sum_{i=1}^{N}\left(a_i-\mu\right)=0$. Durch Quadrieren erhält man

$\sum_{i=1}^{N}\left(a_i-\mu\right)^2+2\cdot\sum_{i=1}^{N-1}\sum_{k=i+1}^{N}\left(a_i-\mu\right)\cdot\left(a_k-\mu\right)=0$. Mit Formel (ii) erhält man

$2\cdot\sum_{i=1}^{N-1}\sum_{k=i+1}^{N}\left(a_i-\mu\right)\cdot\left(a_k-\mu\right)=-\sum_{i=1}^{N}\left(a_i-\mu\right)^2=-N\cdot\sigma^2$. Daraus ergibt sich für $i\neq k$

$\mathrm{E}\left[\left(X_i-\mu\right)\cdot\left(X_k-\mu\right)\right]=\frac{1}{L}\cdot\sum_{i=1}^{N-1}\sum_{k=i+1}^{N}\left(a_i-\mu\right)\cdot\left(a_k-\mu\right)$ mit

$L=\text{Anzahl der Werte}\ a_i-\mu=\sum_{i=1}^{N-1}\sum_{k=i+1}^{N}1=\sum_{i=1}^{N-1}\left(N-\left(i+1\right)+1\right)=\sum_{i=1}^{N-1}i=\frac{N\cdot\left(N-1\right)}{2}$. Also ist

$$\mathrm{E}\left[\left(X_i-\mu\right)\cdot\left(X_k-\mu\right)\right]=-\frac{\sigma^2}{N-1}. \qquad \text{(Formel (viii))}$$

Da $\mathrm{E}\left[Z_i\cdot Z_k\right]=\mathrm{E}\left[\left(Z_i-\mu\right)\cdot\left(Z_k-\mu\right)\right]+\mu^2$ ist, erhält man für den zweiten Summanden auf der rechten Seite in Formel (vi):

$$\begin{aligned}\frac{2}{n^2}\cdot\mathrm{E}\left[\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}Z_i\cdot Z_k\right]&=\frac{2}{n^2}\cdot\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}\mathrm{E}\left[Z_i\cdot Z_k\right]\\&=\frac{2}{n^2}\cdot\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}\left(\mathrm{E}\left[\left(Z_i-\mu\right)\cdot\left(Z_k-\mu\right)\right]+\mu^2\right)\\&=\frac{n-1}{n}\cdot\mu^2-\frac{n-1}{n\cdot\left(N-1\right)}\cdot\sigma^2\quad.\end{aligned} \qquad \text{(Formel (ix))}$$

Aus den Formeln (vi), (vii) und (ix) erhält man $\mathrm{E}\left[\overline{Z}_n{}^2\right]=\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}+\mu^2$ und schließlich

$$\mathrm{Var}\left(\overline{Z}_n\right)=\mathrm{E}\left[\overline{Z}_n{}^2\right]-\left(\mathrm{E}\left[\overline{Z}_n\right]\right)^2=\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}. \qquad \text{(Formel (x))}$$

Für $n>1$ ist $\frac{N-n}{N-1}<1$, d.h. $\mathrm{Var}(\overline{Z}_n)<\mathrm{Var}(\overline{X}_n)$, d.h. das Urnenmodell ohne Zurücklegen gibt eine wirksamere Schätzung für den Erwartungswert als das Urnenmodell mit Zurücklegen.

Wird die Varianz des Merkmals *X* im Urnenmodell ohne Zurücklegen aus einer beobachteten Stichprobe $z_1,...,z_n$ (Zufallsvektor $(Z_1,...,Z_n)$) durch $\hat{\sigma}_z^{\,2}=\frac{1}{n}\cdot\sum_{i=1}^{n}(z_i-\hat{\mu}_z)^2$ mit $\hat{\mu}_z=\frac{1}{n}\cdot\sum_{i=1}^{n}z_i$ geschätzt, so ist auch diese Schätzfunktion nicht erwartungstreu; denn

$$\begin{aligned}\mathrm{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n}(Z_i-\overline{Z}_n)^2\right]&=\frac{1}{n}\cdot\sum_{i=1}^{n}\mathrm{E}\left[(Z_i-\mu)^2\right]-\mathrm{E}\left[(\overline{Z}_n-\mu)^2\right]\\&=\sigma^2-\mathrm{Var}(\overline{X}_n)\qquad\text{(siehe oben)}\\&=\sigma^2-\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}\\&=\frac{N}{N-1}\cdot\frac{n-1}{n}\cdot\sigma^2\quad.\end{aligned}$$

Eine erwartungstreue Schätzung für die Varianz bei unbekanntem Erwartungswert im Urnenmodell mit Zurücklegen erhält man durch die Schätzfunktion

$\frac{n}{n-1}\cdot\frac{N-1}{N}\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\overline{X}_n)^2=\frac{N-1}{N\cdot(n-1)}\cdot\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$ bzw. durch die aus der Stichprobe ermittelten Punktschätzung $\frac{N-1}{N\cdot(n-1)}\cdot\sum_{i=1}^{n}(z_i-\hat{\mu}_z)^2$ mit $\hat{\mu}_z=\frac{1}{n}\cdot\sum_{i=1}^{n}z_i$.

Abschließend werden exemplarisch zwei Methoden zur „geeigneten“ Festgelegung von Schätzfunktionen beschrieben.

Methode der kleinsten Quadrate:

Soll etwa ein unbekannter Mittelwert μ geschätzt werden, so kann bei gegebener Stichprobe $x_1,...,x_n$ derjenige Wert $\hat{\mu}$ als Schätzwert genommen werden, für den die Summe der quadrierten Abstände zu den Stichprobenwerten minimal ist. Es wird also $\sum_{i=1}^{n}(x_i-\hat{\mu})^2$ bezüglich $\hat{\mu}$ minimiert. Dazu wird die erste Ableitung von $\sum_{i=1}^{n}(x_i-\hat{\mu})^2$ gleich 0 gesetzt und nach $\hat{\mu}$ aufgelöst:

$-2\cdot\sum_{i=1}^{n}(x_i-\hat{\mu})=0$ impliziert $\hat{\mu}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i$.

Das ergibt die obige Schätzfunktion für den Erwartungswert.

Maximum-Likelihood-Methode:

Die Verteilung des Merkmals *X* in der Grundgesamtheit sei bis auf einen zu schätzenden Parameter *u* in Form der Massen- bzw. Dichtefunktion bekannt. Diese hängt auch von *u* ab, d.h. sie kann als $f_X(x,u)$ geschrieben werden. Es wird eine Stichprobe $x_1,...,x_n$ nach dem Urnenmodell mit Zurücklegen genommen. Diese entspricht der Realisierung des *n*-dimensionalen Zufallsvektors $(X_1,...,X_n)$. Dessen Massen- bzw. Dichtefunktion lautet $f(x_1,...,x_n,u)$. Es wird nun für *u* derjenige Wert genommen, bei dem die Stichprobenwerte $x_1,...,x_n$ den größte Massen- bzw. Dichtewert besitzen. Dieser Schätzwert für *u* ist somit der plausibelste Wert für die Stichprobenwerte $x_1,...,x_n$.

Es sei *X* normalverteilt mit Erwartungswert μ und Varianz σ^2. Beide seien unbekannt, d.h. der gesuchte Parameter ist $u=(\mu,\sigma)$. Die Dichtefunktion der Stichprobe lautet

$$f(x_1,...,x_n,\mu,\sigma)=f_{X_1}(x_1,\mu,\sigma)\cdot...\cdot f_{X_n}(x_n,\mu,\sigma)$$

$$=\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x_1-\mu}{\sigma}\right)^2}\cdot...\cdot\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x_n-\mu}{\sigma}\right)^2}$$

$$=\frac{1}{\sigma^n\cdot\left(\sqrt{2\cdot\pi}\right)^n}\cdot e^{-\frac{1}{2\cdot\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2}\quad.$$

Dieser Ausdruck ist bezüglich $u=(\mu,\sigma)$ zu maximieren. Da der Berechnungsvorgang kompliziert ist, wird stattdessen der Logarithmus des Ausdrucks maximiert; aufgrund der Monotonie der Logarithmusfunktion hat er das Maximum an derselben Stelle wie der nichtlogarithmierte Ausdruck, und es gilt $\frac{d}{dx}\ln(f(x))=\frac{f'(x)}{f(x)}=0$ genau dann, wenn $f'(x)=0$ ist.

Es ist $\ln(f(x_1,...,x_n,\mu,\sigma))=-n\cdot\ln(\sigma)-n\cdot\ln\left(\sqrt{2\cdot\pi}\right)-\frac{1}{2\cdot\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2$.

Partielles Ableiten nach μ und σ und Nullsetzen beider Gleichungen führt auf die Schätzungen $\hat{\mu}$ für den Erwartungswert und $\hat{\sigma}^2$ für die Varianz:

$$\frac{\partial}{\partial\mu}\ln(f(x_1,...,x_n,\mu,\sigma))=-\frac{1}{2\cdot\sigma^2}\cdot2\cdot(-1)\cdot\sum_{i=1}^{n}(x_i-\mu)=0\,,$$

$$\frac{\partial}{\partial\sigma}\ln(f(x_1,...,x_n,\mu,\sigma))=-n\cdot\frac{1}{\sigma}-\frac{1}{2\cdot\sigma^3}\cdot(-2)\cdot\sum_{i=1}^{n}(x_i-\mu)^2=0\,.$$

Aus der ersten der beiden Gleichungen folgt $\sum_{i=1}^{n}(x_i-\mu)=0$ und $\hat{\mu}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i$.

Dieser Wert wird in die zweite Gleichung eingesetzt und diese nach σ aufgelöst mit dem Ergebnis $\hat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \hat{\mu})^2$. Diese Schätzfunktion ist asymptotisch erwartungstreu (siehe oben).

7.9 Intervallschätzungen

Bei der Punktschätzung eines Parameters der Verteilung eines Merkmals *X* in einer Grundgesamtheit ist es nicht sicher, dass der geschätzte Wert mit dem tatsächlichen Wert übereinstimmt. Aufgrund einer Stichprobe möchte man aber sagen, dass der tatsächliche Parameterwert „mit hoher Wahrscheinlichkeit" nahe bei dem Ergebnis liegt, das man aus der Stichprobe ermittelt. Man sucht daher ein Intervall, das aufgrund der Stichprobe und nach Vorgabe einer Wahrscheinlichkeit den tatsächlichen Parameterwert mit dieser Wahrscheinlichkeit enthält.

Es sei $1-\alpha$ (etwa $1-\alpha = 0{,}95$ oder $1-\alpha = 0{,}99$) eine vorgegebene Wahrscheinlichkeit. Der *n*-dimensionale Zufallsvektor $(X_1, ..., X_n)$ bestehe aus unabhängigen, identisch verteilten Zufallsvariablen mit einem unbekannten Parameter *u*. Für die Stichprobenfunktionen $U_1 = g_1(X_1, ..., X_n)$ und $U_2 = g_2(X_1, ..., X_n)$ gelte $P(U_1 < u < U_2) = 1-\alpha$. Dann heißt das stochastische Intervall $[U_1, U_2]$ ein **Konfidenzintervall** (**Vertrauensintervall**) für *u* mit **Konfidenzniveau** (**Vertrauensniveau**) $1-\alpha$.

Mit jeder Stichprobe $x_1, ..., x_n$, d.h. jeder Realisierung des Zufallsvektors $(X_1, ..., X_n)$, erhält man Realisierungen $u_1 = g_1(x_1, ..., x_n)$ und $u_2 = g_2(x_1, ..., x_n)$ der Zufallsvariablen U_1 und U_2, d.h. jeweils ein Intervall $[u_1, u_2]$. Man kann davon ausgehen, dass $100 \cdot (1-\alpha)\%$ aller so berechneten Intervalle den unbekannten Parameter *u* enthalten.

Im folgenden wird die (0, 1)-Normalverteilung eingesetzt, d.h. die Verteilung mit Dichtefunktion $f_{(0,1)-norm}(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot x^2}$ bzw. Verteilungsfunktion

$$F_{(0,1)-norm}(x) = \int_{-\infty}^{x} \left(\frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot t^2} \right) dt \,.$$

Das Merkmal *X* habe den Erwartungswert $\mathrm{E}[X] = \mu$ und die Varianz $\mathrm{Var}(X) = \sigma^2$. Für die Zufallsstichprobe $(X_1, ..., X_n)$ gilt dann nach Satz 7.7-1 und 7.7-6 mit $\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$:

(1) $\mathrm{E}\left[\overline{X}_n\right] = \mu$ und $\mathrm{Var}(\overline{X}_n) = \frac{1}{n} \cdot \sigma^2$

(2) $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ ist für große n annähernd (0, 1)-normalverteilt, d.h.

$$P\left(-z < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < z\right) \approx F_{(0,1)-norm}(z) - F_{(0,1)-norm}(-z)$$
$$= F_{(0,1)-norm}(z) - \left(1 - F_{(0,1)-norm}(z)\right)$$
$$= 2 \cdot F_{(0,1)-norm}(z) - 1$$

bzw. $P\left(\overline{X}_n - z \cdot \sigma/\sqrt{n} < \mu < \overline{X}_n + z \cdot \sigma/\sqrt{n}\right) \approx 2 \cdot F_{(0,1)-norm}(z) - 1$.

Konfidenzintervall für den unbekannten Erwartungswert μ bei gegebener Varianz σ^2

Zur Konstruktion eines Konfidenzintervalls für μ bei gegebener Varianz σ^2 werden folgende Schritte ausgeführt:

1. Schritt: Man wählt ein Vertrauensniveau $1-\alpha$.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $F_{(0,1)-norm}(x)$ der (0, 1)-Normalverteilung bestimmt man denjenigen Wert $z_{1-\alpha/2}$ mit $2 \cdot F_{(0,1)-norm}(z_{1-\alpha/2}) - 1 = 1 - \alpha$ bzw. $F_{(0,1)-norm}(z_{1-\alpha/2}) = \frac{1 + (1-\alpha)}{2} = 1 - \frac{\alpha}{2}$.

Beispielsweise gelten für die üblichen Vertrauensniveaus folgende Werte $z_{1-\alpha/2}$:

$1-\alpha$	0,90	0,95	0,99	0,999
$z_{1-\alpha/2}$	1,65	1,96	2,58	3,29

3. Schritt: Man berechnet den Mittelwert $\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ der Stichprobenwerte $x_1, ..., x_n$.

4. Schritt: Das Konfidenzintervall für μ lautet $\left[\overline{x} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n}\,,\, \overline{x} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n}\right]$.

Das Konfizenzintervall hat die Länge $2 \cdot z_{1-\alpha/2} \cdot \sigma/\sqrt{n}$; diese hängt vom Vertrauensniveau, der Varianz und dem Stichprobenumfang ab. Ist die vorgegebene Varianz groß, so wird auch das Konfidenzintervall groß sein. Möchte man die Länge des Konfidenzintervalls halbieren, kann man dieses durch Vervierfachung des Stichprobenumfangs erreichen. Erhöht man das Vertrauensniveau, d.h. die Sicherheit, so vergrößert sich entsprechend die Länge des Konfidenzintervalls, d.h. die Genauigkeit der Intervallschätzung verringert sich.

Im allgemeinen ist die Varianz σ^2 nicht bekannt. Für große Stichprobenumfänge (Faustregel $n > 30$) kann man die erwartungstreue Punktschätzung (siehe Kapitel 7.8) $\hat{\sigma}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$ verwenden. Der dadurch entstehende Fehler kann meist vernachlässigt werden. Für kleine Stichprobenumfänge kann man keine Schätzung der Varianz verwenden.

Für den Fall, dass *X* selbst ***normalverteilt*** ist, kann man folgendermaßen vorgehen:

Konfidenzintervall für den unbekannten Erwartungswert μ bei unbekannter Varianz σ^2 (bei normalverteiltem Merkmal):

Mit $\bar{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i^2 - \bar{X}_n^2$ und der erwartungstreuen Schätzung $\frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{n \cdot S^2}{n-1}$ für $\mathrm{Var}(X)$ ist

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X}_n)^2} \Big/ \sqrt{n}}$$

der Quotient zweier Zufallsvariablen (und nicht etwa eine Normierung der normalverteilten Zufallsvariablen $\bar{X}_n$ auf Erwartungswert 0 und Varianz 1). Es ist

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X}_n)^2} \Big/ \sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \cdot \frac{\sigma}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2}} = \frac{\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}}{\sqrt{\frac{1}{n-1} \cdot \frac{n \cdot S^2}{\sigma^2}}}.$$

Im Zähler dieses Bruchs steht eine (0, 1)-normalverteilte Zufallsvariable. Die Zufallsvariable $\frac{n \cdot S^2}{\sigma^2}$ im Nenner des Bruchs ist χ^2-verteilt mit $n-1$ Freiheitsgraden. Dieses wird durch folgende Überlegung plausibel: In Kapitel 7.8 wurde die Identität

$S^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \mu)^2 - (\bar{X}_n - \mu)^2$ gezeigt. Daraus folgt

$$\frac{n \cdot S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 - \left(\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \right)^2 \quad \text{bzw.} \quad \frac{n \cdot S^2}{\sigma^2} + \left(\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2.$$

Auf der rechten Seite des Gleichheitszeichens steht eine χ^2-verteilte Zufallsvariable mit *n* Freiheitsgraden; der zweite Summand auf der linken Seite ist eine χ^2-verteilte Zufallsvariable mit 1 Freiheitsgrad; also ist $\frac{n \cdot S^2}{\sigma^2}$ χ^2-verteilt mit $n-1$ Freiheitsgraden (siehe Kapitel 7.6).

Daher ist $\dfrac{\overline{X}-\mu}{\sqrt{\dfrac{1}{n-1}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\Big/\sqrt{n}}$ t-verteilt mit $n-1$ Freiheitsgraden.

Es werden folgende Schritte ausgeführt:

1. Schritt: Man wählt ein Vertrauensniveau $1-\alpha$.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $T_{n-1}(x)$ der t-Verteilung mit $n-1$ Freiheitsgraden bestimmt man denjenigen Wert $z_{1-\alpha/2}$ mit $2\cdot T_{n-1}(z_{1-\alpha/2})-1=1-\alpha$ bzw. $T_{n-1}(z_{1-\alpha/2})=\dfrac{1+(1-\alpha)}{2}=1-\dfrac{\alpha}{2}$.

3. Schritt: Man berechnet den Mittelwert $\overline{x}=\dfrac{1}{n}\cdot\sum_{i=1}^{n}x_i$ der Stichprobenwerte $x_1,...,x_n$ und den Wert $\overline{\sigma}^2=\dfrac{1}{n-1}\cdot\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$.

4. Schritt: Das Konfidenzintervall für μ lautet $\left[\overline{x}-z_{1-\alpha/2}\cdot\overline{\sigma}/\sqrt{n}\,,\,\overline{x}+z_{1-\alpha/2}\cdot\overline{\sigma}/\sqrt{n}\right]$.

Konfidenzintervall für die unbekannte Varianz σ^2:

Die Zufallsvariable $\dfrac{n\cdot S^2}{\sigma^2}$ ist χ^2-verteilt mit $n-1$ Freiheitsgraden; hierbei ist $S^2=\dfrac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2$ und $\overline{X}_n=\dfrac{1}{n}\cdot\sum_{i=1}^{n}X_i$. Ein Konfidenzintervall für die unbekannte Varianz σ^2 erhält man aus der Forderung, dass die Wahrscheinlichkeit, mit der $\dfrac{n\cdot S^2}{\sigma^2}$ in das Konfidenzintervall fällt, gleich $1-\alpha$ ist und die restliche Wahrscheinlichkeitsmasse α zu gleichen Teilen aufgeteilt wird:

Es sei $F_{\chi^2}(x)$ die Verteilungsfunktion der χ^2-Verteilung. Für den Wert $z_{\alpha/2}$ gelte $F_{\chi^2}\left(z_{\alpha/2}\right)=\alpha/2$, für den Wert $z_{1-\alpha/2}$ gelte $F_{\chi^2}\left(z_{1-\alpha/2}\right)=1-\alpha/2$. Dann lautet die Forderung zur Bestimmung des Konfidenzintervalls $P\left(z_{\alpha/2}\le\dfrac{n\cdot S^2}{\sigma^2}\le z_{1-\alpha/2}\right)=1-\alpha$.

Offensichtlich müssen hier keine Annahmen über den Mittelwert getroffen werden.

Es werden folgende Schritte ausgeführt:

1. Schritt: Man wählt ein Vertrauensniveau $1-\alpha$.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $F_{\chi^2}(x)$ der χ^2-Verteilung mit $n-1$ Freiheitsgraden bestimmt man diejenigen Werte $z_{\alpha/2}$ mit $F_{\chi^2}\left(z_{\alpha/2}\right)=\alpha/2$ und $z_{1-\alpha/2}$ mit $F_{\chi^2}\left(z_{1-\alpha/2}\right)=1-\alpha/2$.

3. Schritt: Man berechnet den Mittelwert $\bar{x}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i$ der Stichprobenwerte $x_1,...,x_n$ und den Wert $\bar{s}^2=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$.

4. Schritt: Das Konfidenzintervall für σ^2 lautet dann $\left[\frac{n\cdot\bar{s}^2}{z_{1-\alpha/2}},\frac{n\cdot\bar{s}^2}{z_{\alpha/2}}\right]$.

Diese Methode ist anwendbar für Konfidenzintervalle für Varianzen bei kleinen Stichproben einer normalverteilten Grundgesamtheit. Der Erwartungswert der mit $n-1$ Freiheitsgraden χ^2-verteilten Zufallsvariablen $\frac{n\cdot S^2}{\sigma^2}$ ist $\mathrm{E}\left[\frac{n\cdot S^2}{\sigma^2}\right]=n-1$, und die Varianz ist $\mathrm{Var}\left(\frac{n\cdot S^2}{\sigma^2}\right)=2\cdot(n-1)$. Für große Stichprobenumfänge konvergiert die χ^2-Verteilung gegen die Normalverteilung, d.h. die auf Erwartungswert 0 und Varianz 1 normierte Zufallsvariable $\frac{n\cdot S^2/\sigma^2-(n-1)}{\sqrt{2\cdot(n-1)}}$ ist annähernd (0, 1)-normalverteilt. Ein Konfidenzintervall zum Vertrauensniveau $1-\alpha$ erhält man aus der Forderung

$$P\left(-z_{1-\alpha/2}\le\frac{n\cdot S^2/\sigma^2-(n-1)}{\sqrt{2\cdot(n-1)}}\le z_{1-\alpha/2}\right)=1-\alpha$$

mit der Stelle $z_{1-\alpha/2}$, an der $F_{(0,1)-norm}(z_{1-\alpha/2})=1-\frac{\alpha}{2}$ ist. Das Konfidenzintervall für σ^2 zum Vertrauensniveau $1-\alpha$ lautet dann

$$\left[\frac{n\cdot\bar{s}^2}{(n-1)+z_{1-\alpha/2}\cdot\sqrt{2\cdot(n-1)}},\frac{n\cdot\bar{s}^2}{(n-1)-z_{1-\alpha/2}\cdot\sqrt{2\cdot(n-1)}}\right].$$

7.10 Hypothesentests

Bei der Konstruktion eines Konfidenzintervalls wurde aus den Werten einer Stichprobe auf die Lage eines unbekannten Parameters geschlossen, die mit vorgegebener, d.h. in der Regel hoher Wahrscheinlichkeit zutrifft. Im vorliegenden Kapitel wird eine **Hypothese** (Vermutung) über den Wert eines unbekannten Parameters oder über eine unbekannte Verteilung

aufgestellt und aufgrund einer Stichprobe, d.h. eines **statistischen Tests**, entschieden, ob die **Hypothese beibehalten** werden kann oder ob die **Hypothese verworfen** werden muss. Es wird keine Aussage darüber getroffen, ob die Hypothese wahr oder falsch ist. Natürlich wird der Schluss aus der Stichprobe auf das Beibehalten bzw. Verwerfen der Hypothese so gezogen, dass die Wahrscheinlichkeiten, eine richtige Hypothese zu verwerfen und eine falsche Hypothese anzunehmen, gering sind.

Man unterscheidet eine Reihe statistischer Testverfahren je nach Art der aufgestellten Hypothese:

- Ein **Parametertest** überprüft eine Hypothese über den Wert eines oder mehrerer unbekannter Parameter einer Grundgesamtheit.
- Ein **Verteilungs-** oder **Anpassungstest** überprüft Annahmen über Wahrscheinlichkeitsverteilungen in einer Grundgesamtheit.
- Bei einem **Unabhängigkeitstest** werden Hypothesen über die stochastische Unabhängigkeit bzw. Abhängigkeit von Zufallsvariablen oder Merkmalen in einer Grundgesamtheit getestet.

Zunächst werden **Parametertests** beschrieben.

Es wird eine Hypothese über den Zahlenwert u_0 eines unbekannten Parameters u einer Verteilung aufgestellt. Die Hypothese kann sich aus früheren statistischen Beobachtungen, theoretischen Überlegungen, Plausibilitätsüberlegungen oder anderen Betrachtungen ergeben. Diese Hypothese über den Wert u_0 von u wird als **Nullhypothese** (**Ausgangshypothese**) H_0 bezeichnet. Die **Gegenhypothese** (**Alternativhypothese**) wird mit H_1 bezeichnet. Je nach Fragestellung sind verschiedene Formen der Null- und Gegenhypothese üblich:

- $H_0: u = u_0$ gegen $H_1: u \neq u_0$ (zweiseitige Fragestellung)
- $H_0: u \leq u_0$ gegen $H_1: u > u_0$ (einseitige Fragestellung)
- $H_0: u \geq u_0$ gegen $H_1: u < u_0$ (einseitige Fragestellung).

Man zieht eine Stichprobe $x_1, ..., x_n$, berechnet daraus den Wert einer Stichprobenfunktion und trifft die Testentscheidung, H_0 beizubehalten oder H_0 abzulehnen. Dabei können viele Ergebnisse zu Überlegungen aus Kapitel 7.9 verwendet werden.

Die Testentscheidung und die Realität können differieren, so dass fehlerhafte Testentscheidungen auftreten:

Der **Fehler 1. Art (α-Fehler)** ist die Wahrscheinlichkeit, H_0 abzulehnen, obwohl H_0 wahr ist, d.h. $\alpha = P(H_0 \text{ ablehnen} \mid H_0 \text{ ist wahr})$.

Der **Fehler 2. Art (β-Fehler)** ist die Wahrscheinlichkeit, H_0 beizubehalten, obwohl H_0 falsch ist, d.h. $\beta = P(H_0 \text{ beibehalten} \mid H_0 \text{ ist falsch})$.

		Realität	
		H_0 ist wahr	H_0 ist falsch
Testentscheidung aufgrund einer Stichprobe	H_0 beibehalten	*korrekt*	Fehler 2. Art, β-Fehler
	H_0 ablehnen	Fehler 1. Art, α-Fehler	*korrekt*

Ziel des Hypothesentests ist es, den Fehler 1. Art möglichst klein zu halten und dabei den Fehler 2. Art nicht zu groß werden zu lassen.

Im Extremfall könnte man sich aufgrund der Stichprobe immer für H_0 entscheiden (dann macht man keinen Fehler 1. Art, d.h. $\alpha = 0$); dann ist allerdings $\beta = 1$.

Üblicherweise wird der Fehler 1. Art vorgegeben, etwa $\alpha = 0{,}05$ oder $\alpha = 0{,}01$. Der Fehler 2. Art müsste dann berechnet werden, was sich im Einzelfall als schwierig erweisen kann.

Hypothesentest des Erwartungswerts μ bei bekannter Varianz σ^2 und großem Stichprobenumfang

Der unbekannte Parameter sei der Erwartungswert μ eines metrischen Merkmals *X* in einer Grundgesamtheit. Die Varianz σ^2 sei bekannt, etwa aus früheren Hypothesentests oder aufgrund einer numerischen Vorgabe. Die Nullhypothese lautet ***bei zweiseitiger Fragestellung***

$H_0: \mu = \mu_0$.

Der Mittelwert $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ einer genommenen Stichprobe $x_1, ..., x_n$, d.h. die Realisierung einer Zufallsstichprobe $(X_1, ..., X_n)$, wird vom hypothetischen Wert μ_0 meist abweichen. Han-

delt es sich um eine ***große Stichprobe*** ($n \geq 30$), so ist der Stichprobenmittelwert $\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ annähernd normalverteilt (siehe Kapitel 7.7):

$$P\left(\left| \frac{\overline{X}_n - \mu_0}{\sigma / \sqrt{n}} \right| \leq z \,\middle|\, \mu = \mu_0 \right) = P\left(-z \leq \frac{\overline{X}_n - \mu_0}{\sigma / \sqrt{n}} \leq z \,\middle|\, \mu = \mu_0 \right)$$
$$\approx F_{(0,1)-norm}(z) - F_{(0,1)-norm}(-z) = 2 \cdot F_{(0,1)-norm}(z) - 1 \quad .$$

Daher bietet sich folgender Hypothesentest an:

1. Schritt: Man wählt den Fehler 1. Art α.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $F_{(0,1)-norm}(x)$ der (0, 1)-Normalverteilung bestimmt man denjenigen Wert $z_{1-\alpha/2}$ mit $2 \cdot F_{(0,1)-norm}(z_{1-\alpha/2}) - 1 = 1 - \alpha$ bzw. $F_{(0,1)-norm}(z_{1-\alpha/2}) = \frac{1 + (1-\alpha)}{2} = 1 - \frac{\alpha}{2}$.

   Beispielsweise gelten für die üblichen Werte α folgende Werte $z_{1-\alpha/2}$:

α	0,1	0,05	0,01	0,001
$z_{1-\alpha/2}$	1,65	1,96	2,58	3,29

3. Schritt: Man berechnet den Mittelwert $\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ der Stichprobenwerte $x_1, ..., x_n$.

4. Schritt: Bei $\left| \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} \right| \leq z_{1-\alpha/2}$ wird H_0 beibehalten, bei $\left| \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} \right| > z_{1-\alpha/2}$ wird H_0 abgelehnt.

Man sieht:

$$P\left(\text{falsche Entscheidung} \,\middle|\, \mu = \mu_0\right) = P\left(\left| \frac{\overline{X}_n - \mu_0}{\sigma / \sqrt{n}} \right| > z_{1-\alpha/2} \,\middle|\, \mu = \mu_0 \right)$$
$$= 1 - P\left(\left| \frac{\overline{X}_n - \mu_0}{\sigma / \sqrt{n}} \right| \leq z_{1-\alpha/2} \,\middle|\, \mu = \mu_0 \right)$$
$$= 1 - \left(2 \cdot F_{(0,1)-norm}(z_{1-\alpha/2}) - 1 \right)$$
$$= 1 - \left(2 \cdot \left(1 - \frac{\alpha}{2} \right) - 1 \right)$$
$$= \alpha \quad .$$

Der Fehler 1. Art wird also eingehalten.

Bei ***einseitiger Fragestellung*** lautet die Nullhypothese

$H_0: \ \mu \le \mu_0$.

Dann gilt $P\left(\dfrac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \le z \,\middle|\, \mu = \mu_0 \right) \approx F_{(0,1)-norm}(z)$. Zu beachten ist, dass eigentlich die bedingte Wahrscheinlichkeit $P\left(\dfrac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \le z \,\middle|\, \mu \le \mu_0 \right)$ genommen werde müsste; dies ist jedoch rechnerisch komplex, und man begnügt sich mit einer Annäherung an den Fehler 1. Art (Höchstwert für den Fehler 1. Art).

Der Hypothesentest lautet nun:

1. Schritt: Man wählt den Fehler 1. Art α.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $F_{(0,1)-norm}(x)$ der (0, 1)-Normalverteilung bestimmt man denjenigen Wert $z_{1-\alpha}$ mit $F_{(0,1)-norm}(z_{1-\alpha}) = 1-\alpha$.
Beispielsweise gelten für die üblichen Werte α folgende Werte $z_{1-\alpha}$:

α	0,1	0,05	0,01	0,001
$z_{1-\alpha}$	1,28	1,64	2,33	3,08

3. Schritt: Man berechnet den Mittelwert $\bar{x} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} x_i$ der Stichprobenwerte $x_1, ..., x_n$.

4. Schritt: Bei $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \le z_{1-\alpha}$ wird H_0 beibehalten, bei $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$ wird H_0 abgelehnt.

Wieder sieht man: $P\left(\text{falsche Entscheidung} \,\middle|\, \mu = \mu_0\right) = \alpha$.

Bei der ***einseitiger Fragestellung*** mit Nullhypothese

$H_0: \ \mu \ge \mu_0$

wird der obige Schritt 4 ersetzt durch

4. Schritt: Bei $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \ge z_{1-\alpha}$ wird H_0 beibehalten, bei $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha}$ wird H_0 abgelehnt.

Hypothesentest des Erwartungswerts μ bei unbekannter Varianz σ^2 und großem Stichprobenumfang

Ist die Varianz σ^2 unbekannt, und ist der Stichprobenumfang groß ($n \geq 30$), so kann man für σ^2 die erwartungstreue Schätzung (siehe Kapitel 7.8) $\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$ nehmen.

Hypothesentest des Erwartungswerts μ bei unbekannter Varianz σ^2 und kleinem Stichprobenumfang und normalverteiltem Merkmal der Grundgesamtheit

In diesem Fall gilt (siehe Kapitel 7.9 bei der Konstruktion des entsprechenden Konfidenzintervalls):

Die Größe $\dfrac{\overline{X} - \mu_0}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2} \Big/ \sqrt{n}}$ ist t-verteilt mit $n-1$ Freiheitsgraden.

Der Hypothesentest lautet nun:

1. Schritt: Man wählt den Fehler 1. Art α.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $T_{n-1}(x)$ der t-Verteilung mit $n-1$ Freiheitsgraden bestimmt man denjenigen Wert $z_{1-\alpha/2}$ mit $2 \cdot T_{n-1}(z_{1-\alpha/2}) - 1 = 1 - \alpha$ bzw. $T_{n-1}(z_{1-\alpha/2}) = \dfrac{1 + (1-\alpha)}{2} = 1 - \dfrac{\alpha}{2}$.

3. Schritt: Man berechnet den Mittelwert $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ der Stichprobenwerte $x_1, ..., x_n$.

4. Schritt: Bei $\dfrac{\bar{x} - \mu_0}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \Big/ \sqrt{n}} \leq z_{1-\alpha/2}$ wird H_0 beibehalten,

bei $\dfrac{\bar{x} - \mu_0}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \Big/ \sqrt{n}} > z_{1-\alpha/2}$ wird H_0 abgelehnt.

Wieder sieht man: $P(\text{falsche Entscheidung} \mid \mu = \mu_0) = \alpha$.

Bei einseitiger Fragestellung wird entsprechend vorgegangen.

Hypothesentest der unbekannten Varianz σ^2

Es ist die Nullhypothese zu überprüfen, dass die unbekannte Varianz der Grundgesamtheit einen definierten Wert annimmt:

$H_0: \sigma^2 = \sigma_0^2$.

Dazu wird die Zufallsvariable $\frac{n \cdot S^2}{\sigma_0^{\,2}}$ mit $S^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$ und $\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ herangezogen. Diese ist χ^2-verteilt mit $n-1$ Freiheitsgraden (siehe Kapitel 7.9).

Der Hypothesentest lautet:

1. Schritt: Man wählt den Fehler 1. Art α.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $F_{\chi^2}(x)$ der χ^2-Verteilung mit $n-1$ Freiheitsgraden bestimmt man diejenigen Werte $z_{\alpha/2}$ mit $F_{\chi^2}\left(z_{\alpha/2}\right) = \alpha/2$ und $z_{1-\alpha/2}$ mit $F_{\chi^2}\left(z_{1-\alpha/2}\right) = 1-\alpha/2$.

3. Schritt: Man berechnet den Mittelwert $\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ der Stichprobenwerte $x_1, ..., x_n$ und den Wert $s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2$.

4. Schritt: Bei $\frac{\sigma_0^{\,2} \cdot z_{\alpha/2}}{n} \le s^2 \le \frac{\sigma_0^{\,2} \cdot z_{1-\alpha/2}}{n}$ wird H_0 beibehalten,

bei $s^2 > \frac{\sigma_0^{\,2} \cdot z_{1-\alpha/2}}{n}$ oder $s^2 < \frac{\sigma_0^{\,2} \cdot z_{\alpha/2}}{n}$ wird H_0 abgelehnt.

Wieder sieht man:

$$
\begin{aligned}
P\left(\text{falsche Entscheidung} \,\middle|\, \sigma^2 = \sigma_0^2\right) &= 1 - P\left(\frac{\sigma_0^{\,2} \cdot z_{\alpha/2}}{n} \le s^2 \le \frac{\sigma_0^{\,2} \cdot z_{1-\alpha/2}}{n} \,\middle|\, \sigma^2 = \sigma_0^2\right) \\
&= 1 - \left(F_{\chi^2}\left(z_{1-\alpha/2}\right) - F_{\chi^2}\left(z_{\alpha/2}\right)\right) \\
&= 1 - \left(1 - \alpha/2 - \alpha/2\right) \\
&= \alpha \quad .
\end{aligned}
$$

Ist die Nullhypothese zu überprüfen, dass die unbekannte Varianz der Grundgesamtheit einen definierten Wert nicht überschreitet, d.h.

$H_0: \sigma^2 \le \sigma_0^2$,

so wird im 2. Schritt derjenige Wert $z_{\alpha/2}$ mit $F_{\chi^2}(z_{1-\alpha}) = 1-\alpha$ bestimmt; die Entscheidung im 4. Schritt lautet nun:

Bei $s^2 \le \frac{{\sigma_0}^2 \cdot z_{1-\alpha}}{n}$ wird H_0 beibehalten, bei $s^2 > \frac{{\sigma_0}^2 \cdot z_{1-\alpha}}{n}$ wird H_0 abgelehnt.

Entsprechend wird verfahren, wenn die Nullhypothese

$H_0: \sigma^2 \ge \sigma_0^2$

lautet.

Hypothesentest für den Vergleich von Erwartungswerten

Es werden zwei Stichproben $x_1, ..., x_{n_1}$ und $y_1, ..., y_{n_2}$ als Realisierung der Zufallsvektoren $(X_1, ..., X_{n_1})$ und $(Y_1, ..., Y_{n_2})$ unabhängig voneinander genommen, und man möchte feststellen, ob sie derselben Grundgesamtheit entstammen bzw. ob die Grundgesamtheiten wenigstens denselben Erwartungswert aufweisen. Hierbei seien die beiden Erwartungswerte μ_1 bzw. μ_2.

Es sei $\overline{X}_{n_1} = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} X_i$ und $\overline{Y}_{n_2} = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} Y_i$.

Die Nullhypothese lautet

$H_0: \mu_1 = \mu_2$.

Sind die Stichprobenumfänge n_1 und n_2 groß oder sind die Merkmale in den Grundgesamtheiten normalverteilt mit den Varianzen σ_1^2 bzw. σ_2^2, so ist die Testgröße $\overline{X}_{n_1} - \overline{Y}_{n_2}$ (asymptotisch) normalverteilt mit dem Erwartungswert $\mu_1 - \mu_2$ und der Varianz $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Kennt man die Varianzen σ_1^2 bzw. σ_2^2 nicht, so kann man sie bei ***großen Stichproben*** durch die erwartungstreuen Schätzwerte $s_1^2 = \frac{1}{n_1 - 1} \cdot \sum_{i=1}^{n_1} (x_i - \overline{x})^2$ und $s_2^2 = \frac{1}{n_2 - 1} \cdot \sum_{i=1}^{n_2} (y_i - \overline{y})^2$ mit $\overline{x} = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} x_i$ und $\overline{y} = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} y_i$ ersetzen.

Der Hypothesentest lautet:

1. Schritt: Man wählt den Fehler 1. Art α.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $F_{(0,1)-norm}(x)$ der (0, 1)-Normalverteilung bestimmt man denjenigen Wert $z_{1-\alpha/2}$ mit $2 \cdot F_{(0,1)-norm}(z_{1-\alpha/2}) - 1 = 1 - \alpha$ bzw. $F_{(0,1)-norm}(z_{1-\alpha/2}) = \frac{1+(1-\alpha)}{2} = 1 - \frac{\alpha}{2}$.

3. Schritt: Man berechnet die Mittelwerte $\bar{x} = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} x_i$ und $\bar{y} = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} y_i$ der Stichprobenwerte.

4. Schritt: Bei $\left| \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} \right| \le z_{1-\alpha/2}$ wird H_0 beibehalten, bei $\left| \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} \right| > z_{1-\alpha/2}$ wird H_0 abgelehnt.

Sind die ***Stichprobenumfänge klein und die Merkmale in den Grundgesamtheiten normalverteilt*** und gilt zusätzlich $\sigma_1^2 = \sigma_2^2 = \sigma^2$, so wird die Testgröße

$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\hat{\sigma}^2 / n_1 + \hat{\sigma}^2 / n_2}} = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\hat{\sigma} \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}}$ genommen. Diese ist t-verteilt mit $n_1 + n_2 - 2$ Freiheitsgraden. Dabei wird $\hat{\sigma}^2$ aus den Beobachtungen beider Stichproben geschätzt: Mit $t_1^2 = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} (x_i - \bar{x})^2$ und $t_2^2 = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} (y_i - \bar{y})^2$ wird $\hat{\sigma}^2 = \frac{n_1 \cdot t_1^{\,2} + n_2 \cdot t_2^2}{n_1 + n_2 - 2}$ gesetzt. Im 2. Schritt bestimmt man aus einer Tabelle der Verteilungsfunktion $T_{n_1+n_2-2}(x)$ der t-Verteilung mit $n_1 + n_2 - 2$ Freiheitsgraden denjenigen Wert $z_{1-\alpha/2}$ mit $2 \cdot T_{n_1+n_2-2}(z_{1-\alpha/2}) - 1 = 1 - \alpha$ bzw. $T_{n_1+n_2-2}(z_{1-\alpha/2}) = \frac{1+(1-\alpha)}{2} = 1 - \frac{\alpha}{2}$. Im 4. Schritt wird bei $\left| \frac{\bar{x} - \bar{y}}{\hat{\sigma} \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}} \right| \le z_{1-\alpha/2}$ H_0 beibehalten; bei $\left| \frac{\bar{x} - \bar{y}}{\hat{\sigma} \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}} \right| > z_{1-\alpha/2}$ wird H_0 abgelehnt.

Hypothesentest für den Vergleich von Varianzen

Es werden zwei Stichproben $x_1,...,x_{n_1}$ und $y_1,...,y_{n_2}$ als Realisierung der Zufallsvektoren $(X_1,...,X_{n_1})$ und $(Y_1,...,Y_{n_2})$ unabhängig voneinander genommen, und man möchte feststellen, ob sie derselben Grundgesamtheit entstammen bzw. ob sie wenigstens Grundgesamtheiten mit denselben Varianzen entnommen wurden.

Die Nullhypothese lautet

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2$.

Es sei $\overline{X}_{n_1} = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} X_i$ und $\overline{Y}_{n_2} = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} Y_i$. Die Zufallsvariable $\frac{n_1 \cdot S_1^2}{\sigma^2}$ ist χ^2-verteilt mit $n_1 - 1$ Freiheitsgraden; hierbei ist $S_1^2 = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} \left(X_i - \overline{X}_{n_1}\right)^2$. Die Zufallsvariable $\frac{n_2 \cdot S_2^2}{\sigma^2}$ ist χ^2-verteilt mit $n_2 - 1$ Freiheitsgraden; hierbei ist $S_2^2 = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} \left(Y_i - \overline{Y}_{n_2}\right)^2$. Die Zufallsvariable

$F_{n_2-1}^{n_1-1} = \frac{\frac{n_1}{n_1-1} \cdot S_1^2}{\frac{n_2}{n_2-1} \cdot S_2^2}$ ist F-verteilt mit den Freiheitsgraden $n_1 - 1$ und $n_2 - 1$.

Es werden folgende Schritte ausgeführt:

1. Schritt: Man wählt ein Vertrauensniveau $1 - \alpha$.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $F_{F_{n_2-1}^{n_1-1}}$ der F-Verteilung mit den Freiheitsgraden $n_1 - 1$ und $n_2 - 1$ bestimmt man diejenigen Werte $z_{\alpha/2}$ mit $F_{F_{n_2-1}^{n_1-1}}(z_{\alpha/2}) = \alpha/2$ und $z_{1-\alpha/2}$ mit $F_{F_{n_2-1}^{n_1-1}}(z_{1-\alpha/2}) = 1 - \alpha/2$.

3. Schritt: Man berechnet die Werte $s_1^2 = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} (x_i - \bar{x})^2$ und $s_2^2 = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} (y_i - \bar{y})^2$ mit $\bar{x} = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} x_i$ und $\bar{y} = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} y_i$.

4. Schritt: Bei $z_{\alpha/2} \le \dfrac{\dfrac{n_1}{n_1-1}\cdot s_1^2}{\dfrac{n_2}{n_2-1}\cdot s_2^2} \le z_{1-\alpha/2}$ wird H_0 beibehalten; bei $\dfrac{\dfrac{n_1}{n_1-1}\cdot s_1^2}{\dfrac{n_2}{n_2-1}\cdot s_2^2} < z_{\alpha/2}$ oder $\dfrac{\dfrac{n_1}{n_1-1}\cdot s_1^2}{\dfrac{n_2}{n_2-1}\cdot s_2^2} > z_{1-\alpha/2}$ wird H_0 abgelehnt.

Anpassungstest

Bei einem Anpassungstest prüft man, ob ein Merkmal *X* in einer Grundgesamtheit einer vorgegebenen Verteilung genügt. Das Merkmal besitze die unbekannte Verteilungsfunktion *F*. Man untersucht also nicht einzelne Parameter wie Erwartungswert oder Varianz der Verteilung, sondern stellt eine Hypothese über die gesamte Verteilung auf. Man bezeichnet einen derartigen Test auch als **nichtparametrischen Test**. Ein prominenter Vertreter von Anpassungstests ist der χ^2**-Anpassungstest**.

Die Nullhypothese lautet in diesem Fall:

$H_0: F = F_0$, wobei F_0 eine vollständig bekannte Verteilungsfunktion ist.

Man geht wie folgt vor, um die Nullhypothese zu bestätigen oder zu verwerfen.

Man zerlegt die Menge der reellen Zahlen in *r* paarweise disjunkte Mengen S_k. Es sei π_k die Wahrscheinlichkeit, dass *X* einen Wert in S_k annimmt. Nimmt man für S_k beispielsweise das Intervall $[a_k, a_{k+1})$, dann ist $\pi_k = F_0(a_{k+1}) - F_0(a_k)$ für $k = 1, ..., r-1$.

Es wird eine unabhängige Stichprobe $x_1, ..., x_n$ genommen. Die Anzahl der Beobachtungen im Intervall $[a_k, a_{k+1})$ sei n_k. Für festes k ist n_k binomialverteilt. Die Testgröße $\chi^2 = \sum_{k=1}^{r} \dfrac{(n_k - n\cdot\pi_k)^2}{n\cdot\pi_k}$ ist für großes *n* (Faustregel $n\cdot\pi_k \ge 10$) asymptotisch χ^2-verteilt mit $r-1$ Freiheitsgraden.

Im Einzelnen führt man die folgenden Schritte aus:

1. Schritt: Man wählt den Fehler 1. Art α.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $F_{\chi^2}(x)$ der χ^2-Verteilung mit $r-1$ Freiheitsgraden bestimmt man denjenigen Wert $z_{1-\alpha}$ mit $F_{\chi^2}(z_{1-\alpha})=1-\alpha$.

3. Schritt: Man berechnet $\sum_{k=1}^{r}\frac{(n_k-n\cdot\pi_k)^2}{n\cdot\pi_k}$.

4. Schritt: Bei $\sum_{k=1}^{r}\frac{(n_k-n\cdot\pi_k)^2}{n\cdot\pi_k}\le z_{1-\alpha}$ wird H_0 beibehalten;

bei $\sum_{k=1}^{r}\frac{(n_k-n\cdot\pi_k)^2}{n\cdot\pi_k}> z_{1-\alpha}$ wird H_0 abgelehnt.

Ist die hypothetische Verteilungsfunktion F_0 nicht vollständig bekannt – man weiß etwa nur, dass sie einer Klasse von Verteilungsfunktionen angehört, jedoch fehlt die Kenntnis einiger ihrer Parameter – , so kann man eventuell nach dem Maximum-Likelihood-Prinzip (siehe Kapitel 7.8) verfahren. Weiß man etwa, dass die Verteilungsfunktion F_0 über die Dichtefunktion $f_0(\lambda_1,...,\lambda_m)$ definiert ist, wobei $\lambda_1,...,\lambda_m$ unbekannt sind, so nimmt man eine unabhängige Stichprobe $x_1,...,x_n$ und bildet $L=\prod_{i=1}^{n}f_0(x_i,\lambda_1,...,\lambda_m)$. Die unbekannten Parameterwerte $\lambda_1,...,\lambda_m$ werden geschätzt, und zwar durch die Lösungen, die sich aus den Gleichungen $\frac{\partial}{\partial\lambda_i}\ln(L)=0$ für $i=1,...,m$ ergeben. Anschließend kann man wie im obigen Anpassungstest fortfahren.

Unabhängigkeitstests

Die Elemente einer Stichprobe werden im Hinblick auf zwei Merkmale *X* und *Y* klassifiziert. Der Bereich der möglichen Werte von *X* wird in *r* Gruppen, der Bereich der möglichen Werte von *Y* wird in *s* Gruppen eingeteilt. Es wird eine Stichprobe vom Umfang *n* genommen.

Mit $n_{i,k}$ für $i=1,...,r$ und $k=1,...,s$ werde die Anzahl der Stichprobenelemente, die zur *i*-ten Gruppe bezüglich des Merkmals *X* und zur *k*-ten Gruppe bezüglich des Merkmals *Y* gehören.

Die Anzahl der Stichprobenelemente in Gruppe *i* bezüglich des Merkmals *X* ist dann gleich $n_{i,\bullet}=\sum_{k=1}^{s}n_{i,k}$. Die Anzahl der Stichprobenelemente in Gruppe *k* bezüglich des Merkmals *Y* ist gleich $n_{\bullet,k}=\sum_{i=1}^{r}n_{i,k}$. Außerdem gilt $n=\sum_{i=1}^{r}\sum_{k=1}^{s}n_{i,k}$.

Anzahl der Stichprobenelemente	Gruppe 1 bzgl. Y	Gruppe 2 bzgl. Y	. . .	Gruppe s bzgl. Y	zusammen $n_{i,\bullet}$
Gruppe 1 bzgl. X	$n_{1,1}$	$n_{1,2}$	. . .	$n_{1,s}$	$n_{1,\bullet}$
Gruppe 2 bzgl. X	$n_{2,1}$	$n_{2,2}$	. . .	$n_{1,s}$	$n_{2,\bullet}$
. . .	. . .	. . .	. . . .	. . .	. . .
Gruppe r bzgl. X	$n_{r,1}$	$n_{r,2}$	. . .	$n_{r,s}$	$n_{r,\bullet}$
zusammen $n_{\bullet,k}$	$n_{\bullet,1}$	$n_{\bullet,2}$	. . .	$n_{\bullet,s}$	n

Es wird die folgende Nullhypothese aufgestellt:

H_0: Die Merkmale X und Y sind stochastisch unabhängig.

Bezeichnet $p_{i,k}$ für $i = 1, ..., r$ und $k = 1, ..., s$ die Wahrscheinlichkeit, dass ein zufällig ausgewähltes Element bezüglich des Merkmals X zur i-ten Gruppe und bezüglich des Merkmals Y zur k-ten Gruppe gehört, und bezeichnen $p_{i,\bullet}$ und $p_{\bullet,k}$ die entsprechenden Randwahrscheinlichkeiten, dann ist die Hypothese H_0 äquivalent zu $p_{i,k} = p_{i,\bullet} \cdot p_{\bullet,k}$ für $i = 1, ..., r$ und $k = 1, ..., s$, wobei $\sum_{i=1}^{r} p_{i,\bullet} = \sum_{k=1}^{s} p_{\bullet,k} = 1$ gilt.

Die Hypothese H_0 sagt nichts über die $r + s$ vielen Werte $p_{i,\bullet}$ und $p_{\bullet,k}$ aus. Zwei dieser Werte lassen sich aus den beiden Gleichungen $\sum_{i=1}^{r} p_{i,\bullet} = 1$ und $\sum_{k=1}^{s} p_{\bullet,k} = 1$ bestimmen, so dass $r + s - 2$ Werte unbekannt sind. Die unbekannten Parameter $p_{i,\bullet}$ und $p_{\bullet,k}$ werden gemäß dem Maximum-Likelihood-Prinzip ermittelt:

Bei Gültigkeit von H_0 ist

$$L = \prod_{i=1}^{r}\prod_{k=1}^{s} p_{i,k}^{n_{i,k}} = \prod_{i=1}^{r}\prod_{k=1}^{s}\left(p_{i,\bullet} \cdot p_{\bullet,k}\right)^{n_{i,k}} = \prod_{i=1}^{r}\prod_{k=1}^{s}\left(p_{i,\bullet}^{n_{i,k}} \cdot p_{\bullet,k}^{n_{i,k}}\right) = \prod_{i=1}^{r} p_{i,\bullet}^{n_{i,\bullet}} \cdot \prod_{k=1}^{s} p_{\bullet,k}^{n_{\bullet,k}}.$$

$p_{r,\bullet}$ und $p_{\bullet,s}$ ergeben sich aus $\sum_{i=1}^{r} p_{i,\bullet} = 1$ bzw. $\sum_{k=1}^{s} p_{\bullet,k} = 1$ zu

$$p_{r,\bullet} = 1 - \sum_{i=1}^{r-1} p_{i,\bullet} \text{ und } p_{\bullet,s} = 1 - \sum_{k=1}^{s-1} p_{\bullet,k}.$$

Damit ist $L = \left(1 - \sum_{i=1}^{r-1} p_{i,\bullet}\right)^{n_{r,\bullet}} \cdot \left(1 - \sum_{k=1}^{s-1} p_{\bullet,k}\right)^{n_{\bullet,s}} \cdot \prod_{i=1}^{r-1} p_{i,\bullet}^{n_{i,\bullet}} \cdot \prod_{k=1}^{s-1} p_{\bullet,k}^{n_{\bullet,k}}$ und

$$\ln(L) = n_{r,\bullet} \cdot \ln\left(1 - \sum_{i=1}^{r-1} p_{i,\bullet}\right) + n_{\bullet,s} \cdot \ln\left(1 - \sum_{k=1}^{s-1} p_{\bullet,k}\right) + \sum_{i=1}^{r-1} n_{i,\bullet} \cdot \ln\left(p_{i,\bullet}\right) + \sum_{k=1}^{s-1} n_{\bullet,k} \cdot \ln\left(p_{\bullet,k}\right).$$

Daraus ergibt sich das Gleichungssystem

$$\frac{\partial}{\partial p_{i,\bullet}} \ln(L) = -\frac{n_{r,\bullet}}{1-\sum_{i=1}^{r-1} p_{i,\bullet}} + \frac{n_{i,\bullet}}{p_{i,\bullet}} = \frac{n_{i,\bullet}}{p_{i,\bullet}} - \frac{n_{r,\bullet}}{p_{r,\bullet}} = 0 \text{ für } i = 1, ..., r-1 ,$$

$$\frac{\partial}{\partial p_{\bullet,k}} \ln(L) = -\frac{n_{\bullet,s}}{1-\sum_{k=1}^{s-1} p_{\bullet,k}} + \frac{n_{\bullet,k}}{p_{\bullet,k}} = \frac{n_{\bullet,k}}{p_{\bullet,k}} - \frac{n_{\bullet,s}}{p_{\bullet,s}} = 0 \text{ für } k = 1, ..., s-1 .$$

Setzt man $A = \frac{n_{r,\bullet}}{p_{r,\bullet}}$ und $B = \frac{n_{\bullet,s}}{p_{\bullet,s}}$, so folgt $p_{i,\bullet} = \frac{n_{i,\bullet}}{A}$ für $i = 1, ..., r$ und $p_{\bullet,k} = \frac{n_{\bullet,k}}{B}$ für $k = 1, ..., s$.

Es ist $1 = \sum_{i=1}^{r} p_{i,\bullet} = \frac{\sum_{i=1}^{r} n_{i,\bullet}}{A} = \frac{n}{A}$ und $1 = \sum_{k=1}^{s} p_{\bullet,k} = \frac{\sum_{k=1}^{r} n_{\bullet,k}}{B} = \frac{n}{B}$, also $A = B = n$.

Insgesamt ergibt sich

$$p_{i,\bullet} = \frac{n_{i,\bullet}}{n} \text{ für } i = 1, ..., r \text{ und } p_{\bullet,k} = \frac{n_{\bullet,k}}{n} \text{ für } k = 1, ..., s .$$

Wie im obigen Anpassungstest (dort lautet die Testgröße $\chi^2 = \sum_{k=1}^{r} \frac{(n_k - n \cdot \pi_k)^2}{n \cdot \pi_k}$) wird jetzt die Testgröße $\chi^2 = \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{(n_{i,k} - n \cdot p_{i,k})^2}{n \cdot p_{i,k}}$ gewählt. Bei Gültigkeit der Nullhypothese ist $p_{i,k} = p_{i,\bullet} \cdot p_{\bullet,k}$; die Werte und $p_{i,\bullet}$ für $i = 1, ..., r$ und $p_{\bullet,k}$ für $k = 1, ..., s$ wurden nach dem Maximum-Likelihood-Prinzip bestimmt. Damit ist

$$\chi^2 = \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{(n_{i,k} - n \cdot p_{i,\bullet} \cdot p_{\bullet,k})^2}{n \cdot p_{i,\bullet} \cdot p_{\bullet,k}} = \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{i,k} - \frac{n_{i,\bullet} \cdot n_{\bullet,k}}{n}\right)^2}{\frac{n_{i,\bullet} \cdot n_{\bullet,k}}{n}} = n \cdot \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{i,k} - \frac{n_{i,\bullet} \cdot n_{\bullet,k}}{n}\right)^2}{n_{i,\bullet} \cdot n_{\bullet,k}} .$$ Da aus der Stichprobe $r + s - 2$ Parameter $p_{i,\bullet}$ für $i = 1, ..., r-1$ und $p_{\bullet,k}$ für $k = 1, ..., s-1$ bestimmt wurden (die beiden Parameter $p_{r,\bullet}$ und $p_{\bullet,s}$ liegen damit fest), besitzt diese Testgröße asymptotisch eine χ^2-Verteilung mit $r \cdot s - (r + s - 2) - 1 = (r-1) \cdot (s-1)$ Freiheitsgraden.

Der so konstruierte Unabhängigkeitstest lautet:

1. Schritt: Man wählt den Fehler 1. Art α.

2. Schritt: Aus einer Tabelle der Verteilungsfunktion $F_{\chi^2}(x)$ der χ^2-Verteilung mit $(r-1) \cdot (s-1)$ Freiheitsgraden bestimmt man denjenigen Wert $z_{1-\alpha}$ mit $F_{\chi^2}(z_{1-\alpha}) = 1 - \alpha$.

3. Schritt: Man berechnet $\chi^2 = n \cdot \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{i,k} - \frac{n_{i,\bullet} \cdot n_{\bullet,k}}{n} \right)^2}{n_{i,\bullet} \cdot n_{\bullet,k}}$.

4. Schritt: Bei $n \cdot \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{i,k} - \frac{n_{i,\bullet} \cdot n_{\bullet,k}}{n} \right)^2}{n_{i,\bullet} \cdot n_{\bullet,k}} \leq z_{1-\alpha}$ wird H_0 beibehalten;

bei $n \cdot \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{i,k} - \frac{n_{i,\bullet} \cdot n_{\bullet,k}}{n} \right)^2}{n_{i,\bullet} \cdot n_{\bullet,k}} > z_{1-\alpha}$ wird H_0 abgelehnt.

In der Reihe FINAL sind bisher erschienen:

1. Jahrgang 1991:

1. Hinrich E. G. Bonin; Softwaretechnik, Heft 1, 1991 (ersetzt durch Heft 2, 1992).
2. Hinrich E. G. Bonin (Herausgeber); Konturen der Verwaltungsinformatik, Heft 2, 1991 (überarbeitet und erschienen im Wissenschaftsverlag, Bibliographisches Institut & F. A. Brockhaus AG, Mannheim 1992, ISBN 3-411-15671-6).

2. Jahrgang 1992:

1. Hinrich E. G. Bonin; Produktionshilfen zur Softwaretechnik --- Computer-Aided Software Engineering --- CASE, Materialien zum Seminar 1992, Heft 1, 1992.
2. Hinrich E. G. Bonin; Arbeitstechniken für die Softwareentwicklung, Heft 2, 1992 (3. überarbeitete Auflage Februar 1994), PDF-Format.
3. Hinrich E. G. Bonin; Object-Orientedness --- a New Boxologie, Heft 3, 1992.
4. Hinrich E. G. Bonin; Objekt-orientierte Analyse, Entwurf und Programmierung, Materialien zum Seminar 1992, Heft 4, 1992.
5. Hinrich E. G. Bonin; Kooperative Produktion von Dokumenten, Materialien zum Seminar 1992, Heft 5, 1992.

3. Jahrgang 1993:

1. Hinrich E. G. Bonin; Systems Engineering in Public Administration, Proceedings IFIP TC8/ WG8.5: Governmental and Municipal Information Systems, March 3--5, 1993, Lüneburg, Heft 1, 1993 (überarbeitet und erschienen bei North-Holland, IFIP Transactions A-36, ISSN 0926-5473).
2. Antje Binder, Ralf Linhart, Jürgen Schultz, Frank Sperschneider, Thomas True, Bernd Willenbockel; COTEXT --- ein Prototyp für die kooperative Produktion von Dokumenten, 19. März 1993, Heft 2, 1993.
3. Gareth Harries; An Introduction to Artificial Intelligence, April 1993, Heft 3, 1993.
4. Jens Benecke, Jürgen Grothmann, Mark Hilmer, Manfred Hölzen, Heiko Köster, Peter Mattfeld, Andre Peters, Harald Weiss; ConFusion --- Das Produkt des AWÖ-Projektes 1992/93, 1. August 1993, Heft 4, 1993.
5. Hinrich E. G. Bonin; The Joy of Computer Science --- Skript zur Vorlesung EDV ---, September 1993, Heft 5, 1993 (4. ergänzte Auflage März 1995).
6. Hans-Joachim Blanke; UNIX to UNIX Copy --- Interactive application for installation and configuration of UUCP ---, Oktober 1993, Heft 6, 1993.

4. Jahrgang 1994:

1. Andre Peters, Harald Weiss; COMO 1.0 --- Programmierumgebung für die Sprache COBOL --- Benutzerhandbuch, Februar 1994, Heft 1, 1994.
2. Manfred Hölzen; UNIX-Mail --- Schnelleinstieg und Handbuch ---, März 1994, Heft 2, 1994.
3. Norbert Kröger, Roland Seen; EBrain --- Documentation of the 1994 AWÖ-Project Prototype ---, June 11, 1994, Heft 3, 1994.
4. Dirk Mayer, Rainer Saalfeld; ADLATUS --- Documentation of the 1994 AWÖ-Project Prototype -- -, July 26, 1994, Heft 4, 1994.
5. Ulrich Hoffmann; Datenverarbeitungssystem 1, September 1994, Heft 5, 1994. (2. überarbeitete Auflage Dezember 1994).
6. Karl Goede; EDV-gestützte Kommunikation und Hochschulorganisation, Oktober 1994, Heft 6 (Teil 1), 1994.
7. Ulrich Hoffmann; Zur Situation der Informatik, Oktober 1994, Heft 6 (Teil 2), 1994.

5. Jahrgang 1995:

1. Horst Meyer-Wachsmuth; Systemprogrammierung 1, Januar 1995, Heft 1, 1995.
2. Ulrich Hoffmann; Datenverarbeitungssystem 2, Februar 1995, Heft 2, 1995.
3. Michael Guder / Kersten Kalischefski / Jörg Meier / Ralf Stöver / Cheikh Zeine; OFFICE-LINK --- Das Produkt des AWÖ-Projektes 1994/95, März 1995, Heft 3, 1995.
4. Dieter Riebesehl; Lineare Optimierung und Operations Research, März 1995, Heft 4, 1995.
5. Jürgen Mattern / Mark Hilmer; Sicherheitsrahmen einer UTM-Anwendung, April 1995, Heft 5, 1995.
6. Hinrich E. G. Bonin; Publizieren im World-Wide Web --- HyperText Markup Language und die Kunst der Programmierung ---, Mai 1995, Heft 6, 1995.
7. Dieter Riebesehl; Einführung in Grundlagen der theoretischen Informatik, Juli 1995, Heft 7, 1995.
8. Jürgen Jacobs; Anwendungsprogrammierung mit Embedded-SQL, August 1995, Heft 8, 1995.
9. Ulrich Hoffmann; Systemnahe Programmierung, September 1995, Heft 9, 1995 (ersetzt durch Heft 1, 1999).
10. Klaus Lindner; Neuere statistische Ergebnisse, Dezember 1995, Heft 10, 1995.

6. Jahrgang 1996:

1. Jürgen Jacobs / Dieter Riebesehl; Computergestütztes Repetitorium der Elementarmathematik, Februar 1996, Heft 1, 1996.
2. Hinrich E. G. Bonin; "Schlanker Staat" & Informatik, März 1996, Heft 2, 1996.
3. Jürgen Jacobs; Datenmodellierung mit dem Entity-Relationship-Ansatz, Mai 1996, Heft 3, 1996.
4. Ulrich Hoffmann; Systemnahe Programmierung, (2. überarbeitete Auflage von Heft 9, 1995), September 1996, Heft 4, 1996 (ersetzt durch Heft 1, 1999).
5. Dieter Riebesehl; Prolog und relationale Datenbanken als Grundlagen zur Implementierung einer NF2-Datenbank (Sommer 1995), November 1996, Heft 5, 1996.

7. Jahrgang 1997:

1. Jan Binge, Hinrich E. G. Bonin, Volker Neumann, Ingo Stadtsholte, Jürgen Utz; Intranet-/Internet- Technologie für die Öffentliche Verwaltung --- Das AWÖ-Projekt im WS96/97 --- (Anwendungen in der Öffentlichen Verwaltung), Februar 1997, Heft 1, 1997.
2. Hinrich E. G. Bonin; Auswirkungen des Java-Konzeptes für Verwaltungen, FTVI'97, Oktober 1997, Heft 2, 1997.

8. Jahrgang 1998:

1. Hinrich E. G. Bonin; Der Java-Coach, Heft 1, Oktober 1998, (CD-ROM, PDF-Format; aktuelle Fassung).
2. Hinrich E. G. Bonin (Hrsg.); Anwendungsentwicklung WS 1997/98 --- Programmierbeispiele in COBOL & Java mit Oracle, Dokumentation in HTML und tcl/tk, September 1998, Heft 2, 1998 (CD-ROM).
3. Hinrich E. G. Bonin (Hrsg); Anwendungsentwicklung SS 1998 --- Innovator, SNiFF+, Java, Tools, Oktober 1998, Heft 3, 1998 (CD-ROM).
4. Hinrich E. G. Bonin (Hrsg); Anwendungsentwicklung WS 1998 --- Innovator, SNiFF+, Java, Mail und andere Tools, November 1998, Heft 4, 1998 (CD-ROM).
5. Hinrich E. G. Bonin; Persistente Objekte --- Der Elchtest für ein Java-Programm, Dezember 1998, Heft 5, 1998 (CD-ROM).

9. Jahrgang 1999:

1. Ulrich Hoffmann; Systemnahe Programmierung (3. überarbeitete Auflage von Heft 9, 1995), Juli 1999, Heft 1, 1999 (CD-ROM und Papierform), Postscript-Format, zip-Postscript-Format, PDF-Format und zip-PDF-Format.

10. Jahrgang 2000:

1. Hinrich E. G. Bonin; Citizen Relationship Management, September 2000, Heft 1, 2000 (CD-ROM und Papierform), PDF-Format.
2. Hinrich E. G. Bonin; WI>DATA --- Eine Einführung in die Wirtschaftsinformatik auf der Basis der Web_Technologie, September 2000, Heft 2, 2000 (CD-ROM und Papierform), PDF-Format.
3. Ulrich Hoffmann; Angewandte Komplexitätstheorie, November 2000, Heft 3, 2000 (CD-ROM und Papierform), PDF-Format.
4. Hinrich E. G. Bonin; Der kleine XMLer, Dezember 2000, Heft 4, 2000 (CD-ROM und Papierform), PDF-Format, aktuelle Fassung.

11. Jahrgang 2001:

1. Hinrich E. G. Bonin (Hrsg.): 4. SAP-Anwenderforum der FHNON, März 2001, (CD-ROM und Papierform), Downloads & Videos.
2. J. Jacobs / G. Weinrich; Bonitätsklassifikation kleiner Unternehmen mit multivariater linear Diskriminanzanalyse und Neuronalen Netzen; Mai 2001, Heft 2, 2001, (CD-ROM und Papierform), PDF-Format und MS Word DOC-Format
3. K. Lindner; Simultanttestprozedur für globale Nullhypothesen bei beliebiger Abhängigkeitsstruktur der Einzeltests, September 2001, Heft 3, 2001 (CD-ROM und Papierform).

12. Jahrgang 2002:

1. Hinrich E. G. Bonin: Aspect-Oriented Software Development. März 2002, Heft 1, 2002 (CD-ROM und Papierform), PDF-Format.
2. Hinrich E. G. Bonin: WAP & WML --- Das Projekt Jagdzeit ---. April 2002, Heft 2, 2002 (CD-ROM und Papierform), PDF-Format.
3. Ulrich Hoffmann: Ausgewählte Kapitel der Theoretischen Informatik (CD-ROM und Papierform), PDF-Format.
4. Jürgen Jacobs / Dieter Riebesehl; Computergestütztes Repetitorium der Elementarmathematik, September 2002, Heft 4, 2002 (CD-ROM und Papierform), PDF-Format.
5. Verschiedene Referenten; 3. Praxisforum "Systemintegration", 18.10.2002, Oktober 2002, Heft 5, 2002 (CD-ROM und Papierform), Praxisforum.html (Web-Site).

13. Jahrgang 2003:

1. Ulrich Hoffmann; Ausgewählte Kapitel der Theoretischen Informatik; Heft 1, 2003, (CD-ROM und Papierform) PDF-Format.
2. Dieter Riebesehl; Mathematik 1, Heft 2, 2003, (CD-ROM und Papierform) PDF-Format.
3. Ulrich Hoffmann; Mathematik 1, Heft 3, 2003, (CD-ROM und Papierform) PDF-Format und Übungen.
4. Verschiedene Autoren; Zukunft von Verwaltung und Informatik, Festschrift für Heinrich Reinermann, Heft 4, 2003, (CD-ROM und Papierform) PDF-Format.

14. Jahrgang 2004:

1. Jürgen Jacobs; Multilayer Neural Networks; Heft 1, 2004, (CD-ROM und Papierform) PDF-Format.

15. Jahrgang 2005:

1. Ulrich Hoffmann; Mathematik für Wirtschaftsinformatiker; Heft 1, 2005, (CD-ROM und Papierform) PDF-Format.
2. Ulrich Hoffmann; Übungen & Lösungen zur Mathematik für Wirtschaftsinformatiker; Heft 1, 2005, (CD-ROM und Papierform) PDF-Format.
3. Ulrich Hoffmann; Datenstrukturen & Algorithmen; Heft 2, 2005, (CD-ROM und Papierform) PDF-Format.

16. Jahrgang 2006:

1. Hinrich E. G. Bonin; Systemanalyse für Softwaresysteme; Heft 1, August 2006, (CD-ROM und Papierform) PDF-Format.
2. Hinrich E. G. Bonin; Faszination Programmierung; Heft 2, August 2006, (CD-ROM und Papierform) PDF-Format.
3. Dieter Riebesehl; Strukturanalogien in Datenmodellen, Heft 3, Dezember 2006, (CD-ROM und Papierform) PDF-Format.

17. Jahrgang 2007:

1. Ulrich Hoffmann; Ausgewählte Kapitel der Theoretischen Informatik; Heft 1, August 2007, (CD-ROM und Papierform) PDF-Format.
2. Ulrich Hoffmann; Mathematik für Wirtschaftsinformatiker und Informatiker; Heft 2, August 2007, (CD-ROM und Papierform) PDF-Format.
3. Hinrich E. G. Bonin; Der Java-Coach, Heft 3, September 2007, (CD-ROM und Papierform) PDF-Format.
4. Jürgen Jacobs; Dichteprognose autoregressiver Zeitreihen, Heft 4, September 2007, (CD-ROM und Papierform) PDF-Format.

18. Jahrgang 2008:

1. Verschiedene Autoren; Festschrift für Prof. Dr. Meyer-Wachsmuth; Heft 1, Juli 2008, (CD-ROM und Papierform) PDF-Format.
2. Ulrich Hoffmann; Ausgewählte Kapitel der Theoretischen Informatik; Heft 2, Dezember 2008, (CD-ROM und Papierform) PDF-Format.

19. Jahrgang 2009:

1. Verschiedene Autoren; Festschrift für Prof. Dr. Goede; Heft 1, August 2009, (CD-ROM und Papierform) PDF-Format.

20. Jahrgang 2010:

1. Hinrich E. G. Bonin; Konstrukte, Konstruktionen, Konstruktionsempfehlungen – Programmieren in LISP; Heft 1, März 2010, (CD-ROM und Papierform) PDF-Format.
2. Verschiedene Autoren; Festschrift für Prof. Dr. Bonin; Heft 2, April 2010, (CD-ROM und Papierform) PDF-Format.
3. Verschiedene Autoren; Frühwarnindikatoren und Risikomanagement,
   1. Forschungssymposiuman der Leuphana Universität Lüneburg, Oktober 2009; Heft 3, April 2010, (CD-ROM und Papierform) PDF-Format.

21. Jahrgang 2011:

1. Verschiedene Autoren; Frühwarnindikatoren und Risikomanagement,
2. Forschungssymposium an der Leuphana Universität Lüneburg, November 2010; Heft 1, Februar 2011, (CD-ROM und Papierform) PDF-Format.

22. Jahrgang 2012:

1. Andreas Mastel, Jürgen Jacobs; Mining User-Generated Financial Content to Predict Stock Price Movements, Heft 1, Dezember 2012, (CD-ROM und Papierform) PDF-Format.

23. Jahrgang 2013:

1. Ulrich Hoffmann; Mathematik für Wirtschaftsinformatik, Heft 1, Oktober 2013, (Papierform) PDF-Format.

Herausgeber der Schriftenreihe FINAL:
Prof. Dr. Ulrich Hoffmann
Leuphana Universität Lüneburg, Scharnhorststraße 1, D-21335 Lüneburg, Germany
email: ulrich.hoffmann@uni.leuphana.de

Verlag:
Eigenverlag (Fotographische Vervielfältigung), Leuphana Universität Lüneburg (vormals Fachhochschule Nordostniedersachsen)

Erscheinungsweise:
ca. 4 Hefte pro Jahr.
Für unverlangt eingesendete Manuskripte wird nicht gehaftet. Sie sind aber willkommen.

Digitales FlnAL-Archiv:
http://www.leuphana.de/institute/iwi/final.html