Brückenkurs Mathematik

T. von der Twer

Inhaltsverzeichnis

KAPITEL 1

Elementare Vorbereitungen

1. Grundlegende Zahlenmengen, Grundrechenarten

Unter einer Menge verstehen wir naiv die Zusammenfassung von diversen Objekten zu einem Ganzen. Gehört ein Objekt x zur Menge M, so sagen wir: „x ist Element von M" und schreiben kurz dafür: $x \in M$. Gehört x nicht zu M, so schreiben wir kurz: $x \notin M$. Eine Menge kann man festlegen durch Aufzählen der Elemente, z.B. $M = \{1, 2, 3\}$. Wichtiger ist die Möglichkeit, eine Menge durch eine Eigenschaft zu definieren, z.B. $\{x | x$ ganze Zahl, $-10 \leq x \leq 10\}$. Lies: „Menge aller x, für die gilt: x ist ganze Zahl und..." Gewöhnlich hat man dabei eine Grundmenge anzugeben, aus der die Objekte mit einer gewissen Eigenschaft auszusondern sind. Im Beispiel ist das die Menge der ganzen Zahlen, die man gern mit $\mathbb{Z}$ bezeichnet. Also $\mathbb{Z} = \{.... - 10, -9, -8, ... - 1, 0, 1, ..., 8, 9, 10, ...\}$ (beidseitig ins Unendliche). Dann schreibt man etwa $M = \{z \in \mathbb{Z} |\, z = 3k$ für ein $k \in \mathbb{Z}\}$ (oder dasselbe mit einem Doppelpunkt oder Semikolon anstelle des senkrechten Strichs) für die Menge der durch 3 teilbaren ganzen Zahlen. Also $2 \notin M$, $-36 \in M$. Man mache sich bei dieser symbolischen Bezeichnung klar: *Hinter dem Strich spielt die Musik,* da kommt die wesentliche Eigenschaft, davor steht gar keine Information oder lediglich die Information, in welcher Grundmenge diese Eigenschaft betrachtet wird. (In $\mathbb{R}$,der Menge aller reellen Zahlen wäre Teilbarkeit durch 3 uninteressant, wieso?) Generell sollte man nie einfach „von links nach rechts" lesen - man läuft dann immer Gefahr, die wesentlichen Informationen zu verpassen, nicht richtig zu ordnen usw. Das gilt insbesondere auch für Rechenausdrücke.

Einige Zahlen-Grundmengen werden sehr oft benutzt und haben daher Standard-Bezeichnungen:

$\mathbb{N}$ = $\{1, 2, 3, 4, 5, ...\}$ Menge der natürlichen Zahlen, zuweilen bei 0 beginnend, manchmal auch $\mathbb{N}_0$ geschrieben für $\{0, 1, 2, 3, ...\}$.

$\mathbb{Z}$ = Menge der ganzen Zahlen, also die negativen Zahlen zu $\mathbb{N}_0$ hinzu.

$\mathbb{Q}$ = Menge der rationalen Zahlen $= \left\{ \frac{p}{q} \,|\, p, q \in \mathbb{Z},\ q \neq 0 \right\}$

$\mathbb{R}$ = Menge der reellen Zahlen (sehr schwierig exakt zu definieren - grob gesagt sind das alle Zahlen, die sich beliebig gut durch rationale Zahlen nähern lassen; wichtige Beispiele sind $\sqrt{2}$, π, e. Sie sind reell, aber nicht rational. Man denke an Dezimaldarstellungen - diejenigen mit unendlicher nichtperiodischer Dezimalentwicklung sind reell, aber nicht rational.

Wesentliche Eigenschaften: In $\mathbb{N}$ kann man nicht *allgemein* subtrahieren, aber in $\mathbb{Z}$. In $\mathbb{Z}$ kann man nicht *allgemein* durch Zahlen $\neq 0$ dividieren, aber in $\mathbb{Q}$ kann

man es. $\mathbb{Q}, \mathbb{R}$ und $\mathbb{C}$ sind *Körper*. In allen Körpern beherrscht man das Rechnen über folgende Gesetze (Axiome), die speziell in $\mathbb{Q}$ und $\mathbb{R}$ und definitionsgemäß in allen Körpern erfüllt sind:

$$\begin{aligned} a+(b+c) &= (a+b)+c \\ 0+a &= a \\ -a+a &= 0 \\ a+b &= b+a \\ a\cdot(b\cdot c) &= (a\cdot b)\cdot c \\ a\cdot(b+c) &= a\cdot b+a\cdot c \\ 1\cdot a &= a \\ a^{-1}\cdot a &= 1\ (a\neq 0) \\ a\cdot b &= b\cdot a \end{aligned}$$

Die Allgemeingültigkeit dieser Formeln ist so zu verstehen und anzuwenden, daß man für jeden Buchstaben jede beliebige konkrete Zahlbezeichnung (für eine Zahl des jeweiligen Bereiches), aber auch jeden Rechenausdruck einsetzen kann und damit stets eine gültige Aussage bekommt. (Die Buchstaben mit diesem Zweck in allgemeingültigen Formeln nennt man **freie Variablen**.) Insbesondere folgt daraus das ganze Bruchrechnen, binomische Formeln usw. Noch ein wichtiger Unterschied: $\mathbb{Q}$ und $\mathbb{R}$ sind *angeordnete* Körper. Was Anordnung bei einem Körper bedeutet, folgt im nächsten Abschnitt. Zunächst stellen wir noch die grundlegenden Gesetze der Bruchrechnung zusammen:

DEFINITION 1. $\frac{a}{b}$ *ist für* $b\neq 0$ *definiert als* $a\cdot b^{-1}$.

Der Sinn der Definition ist dieser: $\frac{a}{b}$ ist für $b\neq 0$ die *einzige, also eindeutige* Lösung der Bestimmungsgleichung $x\cdot b=a$. Denn: $\frac{a}{b}b=(ab^{-1})b=a(b^{-1}b)=a\cdot 1=1\cdot a=a$. Und aus $x\cdot b=a$ folgt: $(xb)b^{-1}=ab^{-1}$, also $x=ab^{-1}$. Dagegen hat die Gleichung $x\cdot 0=a$ für $a=0$ jede reelle Zahl als Lösung, für $a\neq 0$ überhaupt keine.

Aus der Definition folgen mit den Körperaxiomen alle diese Grundgesetze der Bruchrechnung:

$$\begin{aligned} \frac{a}{b}+\frac{c}{d} &= \frac{ad+bc}{bd} \\ \frac{a}{b}\cdot\frac{c}{d} &= \frac{ac}{bd} \\ \frac{\frac{a}{b}}{\frac{c}{d}} &= \frac{ad}{bc} \\ \frac{ac}{bc} &= \frac{a}{b} \\ -\frac{a}{b} &= \frac{-a}{b}=\frac{a}{-b} \end{aligned}$$

Bei Doppelbrüchen ist stets darauf zu achten, dass die Klammerung klar ist - was also der Hauptbruchstrich ist. Das wird in obenstehender Formel einmal durch die größere Länge angezeigt, vor allem aber durch die Höhe des Gleichheitszeichens. Beim praktischen Addieren von Brüchen achte man auf die Verwendung des kleinsten gemeinsamen Vielfachen der Nenner als Hauptnenner. Die vierte Regel oben gibt sowohl für das Kürzen wie das Erweitern die Basis. Leiten wir als Beispiel die

Additionsregel her: Zunächst beweise der Leser, dass stets $(ab)^{-1} = a^{-1}b^{-1}$. Dann haben wir:

$$\begin{aligned}\frac{ad+bc}{bd} &= (ad+bc)(bd)^{-1} = (ad+bc)(b^{-1}d^{-1}) \\ &= adb^{-1}d^{-1} + bcb^{-1}d^{-1} = ab^{-1} + cd^{-1} = \frac{a}{b} + \frac{c}{d}.\end{aligned}$$

Dabei haben wir noch freien Gebrauch von Kommutativ- und Assoziativgesetz gemacht, ohne diese Schritte ins Einzelne „aufzudröseln". Weiter nennen wir noch die wichtigen binomischen Formeln und ihre Verallgemeinerungen:

$$\begin{aligned}(a \pm b)^2 &= a^2 \pm 2ab + b^2,\ (a+b)(a-b) = a^2 - b^2 \\ (a+b)^n &= \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i},\ n \in \mathbb{N}. \text{ Dabei sind definiert:} \\ \binom{n}{i} &= \frac{n!}{i!(n-i)!},\ \text{lies: „}n\text{ über }i\text{", wobei} \\ 0! &= 1,\ (n+1)! = (n+1)n!,\ \text{lies } n!\text{: „}n\text{ Fakultät"}, \\ \sum_{i=0}^{n} a_i &= a_0 + ... + a_n.\end{aligned}$$

Bemerkung: Sie sog. rekursive Definition von $n!$ in der dritten Zeile funktioniert so: Die erste Gleichung definiert, was $0!$ ist, die zweite führt den Wert von $(n+1)!$ auf den für $n!$ zurück. Damit ist die Sache für alle natürlichen Zahlen definiert, da sich diese aus der Null durch Nachfolgerbildung *alle* ergeben. Es ist natürlich sofort festzustellen, dass $n!$ für $n \geq 1$ einfach $1 \cdot ... \cdot n$ ist. Also $2! = 2$, $3! = 6$, $4! = 24$, $5! = 120$. Die Binomialkoeffizienten $\binom{n}{i}$ sind dabei genau die, welche Sie aus dem sog. Pascalschen Dreieck kennen:

Exponent	Koeffizienten	Binom ausgeschrieben
$n=0$	1	$(a+b)^0 = 1 \cdot a^0b^0 = 1$
$n=1$	1,1	$(a+b)^1 = 1 \cdot a^0b^1 + 1 \cdot a^1b^0 = a+b$
$n=2$	1,2,1	$(a+b)^2 = 1 \cdot a^0b^2 + 2 \cdot a^1b^1 + 1 \cdot a^2b^0$
$n=3$	1,3,3,1	$(a+b)^3 = b^3 + 3ab^2 + 3a^2b + a^3$
$n=4$	1,4,6,4,1	$(a+b)^4 = b^4 + 4ab^3 + 6a^2b^2 + 4a^3b + a^4$

Die einfache binomische Gleichung hängt eng mit quadratischer Ergänzung, diese wiederum z.B. mit der Lösungsformel für quadratische Gleichungen zusammen: Für $a \neq 0$ haben wir:

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c \text{ (quadratische Ergänzung)},$$

und aus

$$x^2 + px + q = 0,\ p, q \in \mathbb{R}$$

folgt mit quadratischer Ergänzung

$$\begin{aligned}
\left(x+\frac{p}{2}\right)^2-\left(\frac{p}{2}\right)^2+q &= 0, \text{ also} \\
x &= -\frac{p}{2}\pm\sqrt{\left(\frac{p}{2}\right)^2-q}, \text{ 3 Fälle:} \\
\left(\frac{p}{2}\right)^2-q &> 0\text{: Zwei reelle Lösungen,} \\
\left(\frac{p}{2}\right)^2-q &= 0\text{: Einzige Lösung } -\frac{p}{2} \\
\left(\frac{p}{2}\right)^2-q &< 0\text{: Keine reelle Lösung.}
\end{aligned}$$

Anwendungsbeispiele: $2x^2+4x-8=0$ ist zunächst auf Normalform zu bringen: Gleichwertig: $x^2+2x-4=0$, und das hat nach Formel die beiden reellen Lösungen $x_{1,2}=-1\pm\sqrt{1+4}=-1\pm\sqrt{5}$. $x^2+2ax-4=0$ (x: Unbekannte, a: äußerer Parameter - s.u., wie eine Konstante zu behandeln) hat die Lösungen $x_{1,2}(a)=-a\pm\sqrt{a^2+4}$, die natürlich von a abhängen. Auch hier gibt es stets zwei reelle Lösungen, da a^2+4 stets größer als Null ist. Noch ein Beispiel ohne reelle Lösung: $x^2+x+1=0$ würde nach Lösungsformel formal auf die Lösungen $x_{1,2}=-\frac{1}{2}\pm\sqrt{\frac{1}{4}-1}$ führen, aber $\frac{1}{4}-1<0$, daher gibt es *keine reelle Lösung* (wohl Lösungen in komplexen Zahlen).

Abschließend verallgemeinern wir noch die Rechengesetze Distributiv- Assoziativ- und Kommutativgesetz auf mit großem Summenzeichen geschriebene Summen, auch Doppelsummen:

$$\begin{aligned}
\sum_i ca_i &= c\sum_i a_i,\ \sum_{i=1}^{n} c = n\cdot c\ (n\in\mathbb{N}, n\geq 1) \\
\sum_i a_i+\sum_i b_i &= \sum_i (a_i+b_i) \\
\left(\sum_i a_i\right)\cdot\left(\sum_j b_j\right) &= \sum_i\left(\sum_j a_ib_j\right)=\sum_i a_i\left(\sum_j b_j\right) \\
&= \sum_j\left(\sum_i a_ib_j\right)=\sum_j b_j\left(\sum_i a_i\right)=\sum_{i,j} a_ib_j
\end{aligned}$$

Dabei läuft i stets über dieselbe Indexmenge, ebenso j stets über dieselbe Indexmenge, die jedoch nicht mit der für i übereinstimmen muss. Eine Summe über leere Indexmenge ist definitionsgemäß Null, damit stimmt die zweite Formel auch für $n=0$, wenn man die Indexmenge der natürlichen Zahlen ≥ 1, die ≤ 0 sind, richtig als leere Menge versteht. Indizes (so heißt der Plural von *Index*) nimmt man stets ganzzahlig, aber man kann im Prinzip jede endliche Menge nehmen. Verallgemeinerungen werden jeweils aus dem Zusammenhang klar. Konkretisierendes

Beispiel zur letzten Formel:

$$\begin{aligned}(a_1+a_2)(b_1+b_2+b_3) &= a_1b_1+a_1b_2+a_1b_3+a_2b_1+a_2b_2+a_2b_3\\ &= a_1(b_1+b_2+b_3)+a_2(b_1+b_2+b_3)\\ &= b_1a_1+b_1a_2+b_2a_1+b_2a_2+b_3a_1+b_3a_2\\ &= b_1(a_1+a_2)+b_2(a_1+a_2)+b_3(a_1+a_2)\end{aligned}$$

2. Umformen von Rechenausdrücken, Gleichungen und Ungleichungen

Basis des Umgangs mit Rechenausdrücken ist das *gleichwertige Umformen.* Man vollzieht das in Gleichungsketten, macht dabei gewöhnlich mehrere Schritte auf einmal. Konzentrierter Überblick ist wichtig! Z.B. $\frac{(x+y)^3}{x^2-y^2}+\frac{1}{x-y}=\frac{(x+y)^2+1}{x-y}$, $(a+b+c)^2=a^2+b^2+c^2+2ab+2ac+2bc$. Jeder Einzelschritt besteht im Einsetzen in eine als allgemeingültig bekannte Gleichung. Wichtig ist ferner die allgemeingültige Regel, daß man stets Teilausdrücke durch gleichwertige ersetzen darf. Zu den Gesetzen des Rechnens mit Brüchen und Potenzen gesellt sich weiterhin die Nutzung von Formeln zu speziellen Funktionen, z.B. $\ln(3x^5)=\ln(3)+5\ln(x)$ (zum Logarithmus später mehr) sowie Formeln zum Ableiten etc.

Ähnlich funktioniert das gleichwertige Umformen von Gleichungen (gewöhnlich Bestimmungsgleichungen, die man lösen möchte). Dabei nutzt man jedoch nicht nur gleichwertiges Umformen beteiligter Rechenausdrücke, sondern zusätzlich spezielle Rechentechniken für spezielle Gleichungstypen (quadratische Gleichungen, Gleichungen wie $e^{2x+3}=5$). Bewußt sollte man ein wichtiges allgemeines Mittel zur Umformung von Gleichungen verwenden und vorsichtig verstehen - das gilt nur für Gleichungen, nicht für Rechenausdrücke, und es ist vorsichtig und verständnisvoll zu verwenden: „Auf beiden Seiten dasselbe tun". Dies ist zunächst einmal keine korrekt formulierte Regel, z.B. darf man sicher nicht auf beiden Seiten einer Gleichung immer dort, wo eine 2 steht, eine Eins hinsetzen. Hier ist ein besonders gravierender Fehler zu nennen, der von Anfängern *in der Mehrzahl (!)* tatsächlich gemacht wird: Man hat auf einer Seite einer Gleichung einen Bruch stehen, auf der andern Seite eine Summe (etwa auch von Brüchen) und möchte den Kehrwert bilden, da die Unbekannte etwa im Nenner steht: Der rücksichtslose Anfänger bildet gliedweise die Kehrwerte auf beiden Seiten, hat im oberflächlichen Sinne „dasselbe auf beiden Seiten" gemacht und damit völligen Unsinn produziert. Gemeint sein muß genauer: Man darf auf beiden Seiten auf den dort stehenden Rechenausdruck dieselbe Funktion anwenden und erhält eine Gleichung, die aus der ursprünglichen folgt. (Achtung: Nicht immer folgt die ursprüngliche auch aus der so umgeformten, diese Umformungen sind nur dann *gleichwertig*, wenn die angewandte Funktion eine umkehrbare ist.) Beispiel: Aus der Gleichung

$$\begin{aligned}\sqrt{x+2} &= x+1 \text{ folgt}\\ x+2 &= x^2+2x+1.\end{aligned}$$

Die Lösungen der letzteren sind $-\frac{1}{2}\pm\frac{1}{2}\sqrt{5}$. Aber unter $\sqrt{x+2}$ versteht man definitionsgemäß eine positive Zahl. Also ist nur $-\frac{1}{2}+\frac{1}{2}\sqrt{5}$ eine Lösung der ersten Gleichung. Die zweite Gleichung folgt aus der ersten, aber nicht umgekehrt. Aber immerhin konnten wir die einzige Lösung der ersten Gleichung finden, indem wir die zweite lösten und die für die erste Gleichung unbrauchbare Lösung wegwarfen. Das ist ganz typisch so.

Addiert man dagegen auf beiden Seiten einer Gleichung dieselbe Zahl, multipliziert man beide Seiten mit derselben Zahl $\neq 0$, so erhält man eine gleichwertige Umformung. Beispiel: $x + 2 = \frac{3}{7}$, also $x = -\frac{11}{7}$. Es handelt sich um gleichwertige Umformung *der Gleichung*; man beachte: Nicht etwa sind die *Rechenausdrücke* dabei gleichwertig umgeformt worden, $x+2 = x$ ist offenbarer Unsinn. Man vermeide den Anfängerfehler, dies Vorgehen auf Rechenausdrücke fälschlich zu übertragen: $\frac{f(x)}{x^2+1} = 0$ ist gleichwertig zu $f(x) = 0$, aber natürlich $\frac{f(x)}{x^2+1} \neq f(x)$, wenn nur $f(x) \neq 0$.
Basis für den Umgang mit Ungleichungen sind die Gesetze (Axiome) für das Rechnen in angeordneten Körpern (wie $\mathbb{Q}$ und $\mathbb{R}$). Diese lauten ($a \leq b$ bedeutet definitionsgemäß $a < b$ oder $a = b$):

$$\begin{aligned} a &< b \text{ oder } b < a \text{ oder } a = b \text{ (genau eins von diesen dreien trifft zu)} \\ \text{Wenn } a &< b \text{ und } b < c, \text{ dann } a < c \text{ (ebenso für } \leq) \\ \text{Wenn } a &< b, \text{ dann } a + c < b + c \text{ (ebenso für } \leq) \\ \text{Wenn } a &< b \text{ und } c > 0, \text{ dann } ac < bc \end{aligned}$$

Die letzteren beiden Regeln besagen, daß man beim Addieren und Multiplizieren auf beiden Seiten einer Ungleichung analog wie bei Gleichungen verfahren kann. Lediglich hat man darauf zu achten, daß beim Multiplizieren mit einer negativen Zahl auf beiden Seiten Umkehrung des Kleiner-Zeichens eintritt:

$$\text{Wenn } a < b \text{ und } c < 0, \text{ dann } ac > bc.$$

Übrigens kann man diese Regel aus den vorigen bereits herleiten, wir stellen sie nur wegen ihrer Bedeutung zur Vermeidung von Fehlern gesondert heraus. Denn aus $a > 0$ folgt $-a < 0$, da $-a = 0$ ausgeschlossen ist und aus $-a > 0$ und $a > 0$ folgen würde: $0 = -a + a > 0$, im Widerspruch zum ersten Axiom. Es bleibt also nur $-a < 0$ möglich, wenn $a > 0$. Aus $a < b$ und $c < 0$ folgt damit $a(-c) < b(-c)$ und daraus $ac > bc$.

Weitere Folgerungen sind etwa: Wenn $a < b$ und $c < d$, dann $a + c < b + d$; wenn $0 \leq a < b$ und $0 \leq c < d$, dann $ac < bd$.

Zu bemerken ist ferner, dass man wichtige weitere Ungleichungen aus der Kenntnis der Monotonie von Funktionen erhält, z.B. weiß: $0 < x < y$, dann $0 < x^2 < y^2$. Oder: Wenn $0 < x < y$, dann $\frac{1}{y} < \frac{1}{x}$. Übrigens erreicht man mit den feineren Mitteln der Analysis gerade die Herleitung von wichtigen Ungleichungen zur Abschätzung von Fehlern etc., die nicht so einfach einzusehen sind. Zunächst einmal werden wir Ungleichungen vor allem zur beschreibenden Einschränkung von Bereichen antreffen. In dieser Form sollten sie sich von selbst verstehen. Möchte man auch in bereits komplizierteren Fällen wissen, wie ein Bereich aussieht, der durch eine Ungleichung beschrieben wird, z.B. die Menge aller reellen Zahlen x, für die gilt: $x^2 - 2x - 3 < 0$, so wird man vielfach so vorgehen, daß man stattdessen die zugehörige Gleichung löst, im Beispiel also $x^2 - 2x - 3 = 0$, $x_{1,2} = 1, 3$. Dann kennt man die Grenzen und weiß etwa im Beispiel: $x^2 - 2x - 3 < 0$ gilt genau für $1 < x < 3$.

Die schließlich noch wichtigen Rechengesetze werden ausführlich im Kontext der Exponential- und Logarithmusfunktionen besprochen, aber das einfachste für

den Umgang mit rationalen Exponenten nehmen wir hier vorweg:

$$\begin{aligned}
\text{Definition} &: \quad x^0 = 0,\ x^n = \underbrace{x_1 \cdot \ldots \cdot x_n}_{n \text{ mal}},\ \text{für } x \in \mathbb{R},\ n \in \mathbb{N},\ n \geq 1. \\
\text{Definition} &: \quad x^{1/n} = \sqrt[n]{x},\ \text{für } x \geq 0,\ n \in \mathbb{N},\ n \geq 1. \\
\text{Definition} &: \quad x^{-a} = \frac{1}{x^a}.\ \text{Damit kennen wir } x^a \text{ für alle } a \in \mathbb{Q}. \\
\text{Rechengesetze} &: \quad x^a x^b = x^{a+b},\ \left(x^a\right)^b = x^{ab},\ x^a y^a = (xy)^a. \\
\text{Es folgt z.B. auch} &: \quad \frac{x^a}{x^b} = x^{a-b}.
\end{aligned}$$

3. Rollen von Buchstaben

Die nützlichen elementaren Vorkenntnisse gliedern sich in drei Teile, die man zusammenwirken müssen, um zunächst einfachere, dann auch komplexere Probleme behandeln zu lassen: Rechnen, Umgang mit einfachsten geometrischen Gebilden wie Geraden und Parabeln in der Ebene, schließlich der verständige Einsatz von verschiedenen logischen Grundoperationen durch den sinnvollen Gebrauch von Buchstaben (sowohl beim Rechnen als auch bei der Beschreibung geometrischer Sachverhalte). Dazu ein

Beispiel: Warum ist das Rechteck mit dem größten Flächeninhalt bei gegebenem Umfang u ein Quadrat (der Seitenlänge $u/4$)? Wir rechnen dazu aus: Mit den Seitenlängen a, b kommt man auf den Flächeninhalt $F(a,b) = ab$ und hat die Bedingung $2a+2b = u$, also $F(a,b) = a(u-2a)/2 = -a^2 + au/2 = -(a-u/4)^2 + u^2/16$, und nun ist (ohne Differentialrechnung!) klar, dass $F(a,b)$ maximal wird für $a = b = u/4$. Derartige Überlegungen sollte man bei *neuen* Problemen ohne weiteres selbständig anstellen und die zugehörigen Rechnungen praktisch und korrekt ausführen können. Typisch treten jedoch Probleme bei Anfängern gerade in dieser Hinsicht auf, und das liegt daran, dass eben die besagten Vorkenntnisse fehlen. Analysieren wir, was man für das Beispiel tatsächlich benötigt: Zunächst ist u ein **äußerer Parameter**, mit einem beliebig festgesetzten positiven Wert, der durch die ganze Aufgabe gezogen wird. Mit a und b werden die Seitenlängen eines beliebigen Rechtecks bezeichnet. Diese Buchstaben treten jedoch ganz anders auf: Es sind **unabhängige Variablen**, genauer ist nur eine unabhängig, der Wert der anderen mit vorgegebenem Umfang u fixiert. (Das ist der Sinn des Rechenschrittes, der die Variable b eliminiert.) $F(a,b)$ tritt als **abhängige Variable** auf - daher die Funktionsschreibweise '$F(a,b)$'. So weit die Situationsbeschreibung: Flächeninhalt eines Rechtecks in Abhängigkeit von einer Seitenlänge bei vorgeschriebenem Umfang. Nun verlangt die Aufgabe, die variable Seitenlänge so einzurichten, dass der Flächeninhalt maximal wird. Damit tritt ein sehr typischer **Rollenwechsel** ein: a wird zur **Unbestimmten** (oder auch Unbekannten). Die Bedingung lautet: $-(a-u/4)^2 + u^2/16$ soll maximalen Wert erhalten. Diese Bedingung hat eine eindeutige Lösung, da jedes Quadrat reeller Zahlen positiv ist: $a = u/4$. Aus diesem einzigen Wert besteht die **Lösungsmenge (**oder Erfüllungsmenge) unserer Bedingung. Betrachten wir nun die Rechnung näher: Zunächst werden Flächeninhalt und Umfang eines Rechtecks mit den Seitenlängen a, b *allgemein* ausgedrückt. $F = ab$ und $u = 2a + 2b$ sind hier **unter einer speziellen Interpretation der beteiligten Größen allgemeingültige Formeln**. Die darin auftretenden Buchstaben sind **freie Variablen**, d.h. solche, für die man beliebige Werte *im Rahmen*

der vorgegebenen Deutung einsetzen kann und dann stets gültige Aussagen erhält: Etwa für $a = 4$ und $b = 5$ erhält man $F = 20$, aber auch für $F = 3$ und $a = 1$ folgt $b = 3$. In eine solche allgemeingültige Formel kann man jedoch nicht nur konkrete Werte einsetzen, sondern viel wichtiger ist es, allgemeine Ausdrücke einzusetzen. Genau dies geschah oben: Die Gleichung $2a + 2b = u$ wurde nach b aufgelöst zu $b = (u-2a)/2$, und dann wurde dieser Ausdruck für b in die Gleichung $F(a, b) = ab$ eingesetzt, mit dem Resultat $F(a, b) = a(u - 2a)/2$, was nur noch von der einen unabhängigen Variablen a abhängt und auch als $F(a)$ geschrieben werden könnte. Man beachte, dass hier bereits a einen ersten Rollenwechsel von der freien Variablen in der Formel zur unabhängigen Variablen hatte! Der nächste Schritt war das Ausmultiplizieren $a(u - 2a)/2 = -a^2 + au/2$. Dabei wurde die (ohne eine vorgeschriebene Deutung für alle Zahlen) allgemeingültige Rechenformel (Distributivgesetz!) $x(y + z) = xy + xz$ benutzt, wiederum mit Einsetzen von Ausdrücken: $a/2$ für x, u für y und $-2a$ für z. Zur Umordnung wurde noch die Formel $x+y = y+x$ verwandt. Dann wurde quadratische Ergänzung vorgenommen, gemäß der binomischen Formel $(x - y)^2 = x^2 - 2xy + y^2$. Für x setzen wir a ein, für $y : u/4$, dann folgt $(a - u/4)^2 = a^2 - au/2 + u^2/16$. Subtraktion von $u^2/16$ auf beiden Seiten und Multiplikation der Gleichung mit -1 ergibt das Resultat, an dem sich die Lösung der Aufgabe ohne weiteres ab lesen ließ. (Damit sind noch immer nicht alle einzelnen Formelanwendungen genannt.)

Einige weitere Rollen von Buchstaben kamen im Beispiel noch nicht vor:

Konstanten wie π oder die Eulerzahl e, auch Naturkonstanten wie die Lichtgeschwindigkeit bezeichnet man mit Buchstaben. Stets sollte man auch damit Rechnungen ausführen und allenfalls im Endresultat Näherungswerte einsetzen; denn diese Zahlen hören in ihrer Dezimalentwicklung hinter dem Komma nicht auf, und sie sind nicht als Brüche darstellbar. Rechnungen mit Näherungswerten produzieren alsbald untragbare Fehler. (Ähnlich verhält es sich mit $\sqrt{2}$ usw.)

Hilfsvariablen dienen dazu, Rechenausdrücke abzukürzen, zu vereinfachen, gezielt ein Verhalten zu untersuchen oder strategisch Rechnungen auszuführen. Zum Beispiel wird man in der Bestimmungsgleichung $x^4 - 4x^2 + 3 = 0$ die neue Unbestimmte $u = x^2$ einführen, um eine quadratische Gleichung zu erhalten. Oder man setzt im Ausdruck

$$f(x) = \frac{1 + \sqrt{1 + \left(\frac{x}{x+1}\right)^2}}{\sqrt{1 + \left(\frac{x}{x+1}\right)^2}}$$

für die Wurzel die neue unabhängige Variable a und erhält den einfachen Ausdruck $(a + 1)/a$. Daran kann man sofort sehen, dass für $a \geq 1$ nur Werte resultieren im Bereich $]1, 2]$ (das ist das Intervall von 1 bis zwei, wobei 1 ausgeschlossen und 2 eingeschlossen ist). Also hat f genau alle Werte in diesem Intervall, da die Wurzel alle Werte im Bereich $[1, \infty[$ annimmt für $x \in \mathbb{R}$.

Freie Parameter benutzt man gerade in der Naturwissenschaft gern, um Mengen zu beschreiben. Die Lösungsmenge der Gleichung $2x + 3y = 1$ im Reellen ist die Menge aller reellen Zahlenpaare, welche diese Gleichung erfüllen, also $\{(x, y) \,|2x + 3y = 1\}$. Nun rechnet man diese Menge aus, indem man feststellt, dass man mit beliebig festgelegtem $x \in \mathbb{R}$ genau ein Lösungspaar $(x, (1 - 2x)/3)$

erhält. Nun schreibt man gern die Lösungsmenge auf in der Form

$$L(x) = \left(x, \frac{1}{3} - \frac{2}{3}x\right),\ x \in \mathbb{R}.$$

Die Werte des freien Parameters x durchlaufen alle reellen Zahlen, und damit durchläuft $L(x)$ die Lösungsmenge der Gleichung.

Stumme oder gebundene Variablen treten z.B. dann auf, wenn man davon spricht, mit allen Objekten eines Bereiches eine Operation vorzunehmen. Etwa bedeutet der Rechenausdruck

$$\sum_{i=1}^{10} i\ (= 55),$$

dass *alle* ganzen Zahlen von 1 bis 10 zu addieren sind, mit dem Resultat 55. Es wäre also Unfug, zu fragen, welche Zahl i denn nun bedeute. Ebenso wenig wäre eine Gleichung mit dieser Summe nach i aufzulösen. Ebenso bedeutet $\int_0^1 x^2 dx$, dass die Funktion $f(x) = x^2$ im Bereich von 0 bis 1 zu integrieren ist. Wieder spielen *alle* zugehörigen Funktionswerte hinein.

Man wird bemerkt haben, dass die Rollen von Buchstaben eng mit Arten von Gleichungen oder Formeln verknüpft sind, in denen die Buchstaben vorkommen. Der logische Gebrauch dieser Gleichungen muss unbedingt verstanden werden.

4. Arten von Formeln und Gleichungen

Allgemeingültige Formeln: Man setzt beliebige Werte oder Rechenausdrücke für die **freien Variablen** ein und erhält stets eine wahre bzw. allgemeingültige Aussage. Beispiel: Aus der Formel $(x + y)^2 = x^2 + 2xy + y^2$ erhält man durch Einsetzen von $-y$ für y die wiederum allgemeingültige Formel $(x - y)^2 = x^2 - 2xy + y^2$. Man beachte, dass man *jeden* Ausdruck einsetzen darf, nicht etwa wird hier die Behauptung $y = -y$ gemacht, die außer für $y = 0$ falsch ist. Beim Einsetzen ist darauf zu achten, dass man zunächst das Eingesetzte einzuklammern hat und dann zuweilen überlegt diese Klammern ersparen kann, Beispiel: Setzt man oben $a + b$ für y, so kommt $(x + (a + b))^2 = x^2 + 2x(a + b) + (a + b)^2$ heraus - die Klammer um $a + b$ auf der linken Seite ist überflüssig, auf der rechten Seite benötigt man beide Klammern.

Allgemeingültige Formeln unter einer gegebenen Deutung: Die Formel $c^2 = a^2 + b^2$ (Pythagoras) gilt nicht allgemein - setzt man beliebige Zahlen für a, b, c, so wird man eine falsche Gleichung erhalten. Aber unter der Voraussetzung, dass a, b die Kathetenlängen und c die Hypotenusenlänge eines rechtwinkligen Dreiecks sind, gilt die Formel *allgemein.*

Definitorische (allgemeine) Formeln: Man definiert z.B. $f(x) = x^2$ für einen Zusammenhang. Solche definitorischen Formeln sind wie allgemeingültige zu behandeln. Im Beispiel folgt etwa $f(a + b) = (a + b)^2$. (Wieder die Klammer!)

Bestimmungsgleichungen: Eine Gleichung wie $x^2 + 2x - 1 = 0$ ist nicht etwa allgemeingültig. Vielmehr stellt sie eine Bedingung an x, die in diesem Falle von genau den beiden Zahlen $x_{1,2} = -1 \pm \sqrt{2}$ erfüllt wird. Man hat eine Bedingung an eine Größe und sucht die Lösungsmenge. Entscheidend für die logische Behandlung ist die Frage: Welche Buchstaben sind Unbestimmte? Eine Bestimmungsgleichung kann durchaus noch äußere Parameter enthalten, und dann sollte man nicht etwa nach diesen auflösen. Dazu ein

Beispiel: Welche Gerade der Form $y = mx + b$ geht durch die Punkte (x_0, y_0) und (x_1, y_1), und unter welchen Umständen gibt es überhaupt eindeutig eine solche Gerade? In dieser Aufgabe treten m, b als Unbestimmte und $x_{0,1}$ sowie $y_{0,1}$ als äußere Parameter auf. Wir haben ein System von zwei Bestimmungsgleichungen:

$$\begin{aligned} y_0 &= mx_0 + b \\ y_1 &= mx_1 + b \end{aligned}$$

und rechnen nach Subtraktion beider Gleichungen aus:

$$\begin{aligned} m &= \frac{y_1 - y_0}{x_1 - x_0}, \text{ für } x_1 \neq x_0. \\ b &= y_0 - \frac{y_1 - y_0}{x_1 - x_0} x_0. \end{aligned}$$

Für $x_1 \neq x_0$ haben wir also eine eindeutige Lösung, und die hängt in der angegebenen Weise von den äußeren Parametern ab. Der Fall $x_1 = x_0$ bleibt noch zu untersuchen, in dem der Bruch keinen Sinn macht. Für $y_1 \neq y_0$ haben wir eine Gerade senkrecht zur $x-$ Achse, und es gibt *keinen* Ausdruck $y = mx + b$, der die Gerade beschreibt. Für $y_1 = y_0$ dagegen haben wir unendlich viele Geraden der Form $y = mx + b$, welche durch den einen vorgeschriebenen Punkt gehen, unser Bedingunssystem reduziert sich auf die einzige Gleichung $y_0 = mx_0 + b$, und für *jede* Zahl m erhalten wir eine solche Gerade, indem wir $b = y_0 - mx_0$ setzen. Solche Fallunterscheidungen treten typisch bei Problemen mit äußeren Parametern auf.

Man beachte, dass das Gleichungssystem

$$\begin{aligned} y_0 &= mx_0 + b \\ y_1 &= mx_1 + b \end{aligned}$$

mit veränderten Rollen eine völlig andere Aufgabe definiert: Seien nun m, b sowie x_0, x_1 äußere Parameter, y_0, y_1 Unbestimmte. Dann handelt es sich inhaltlich um die Frage, welche $y-$ Werte zu den vorgegebenen $x-$ Werten gehören auf der Geraden mit Steigung m und dem $y-$Achsenabschnitt b. Es ist nichts weiter zu rechnen, die Gleichungen sind in Endform für diese Aufgabe.

Noch einmal zurück zur ersten Aufgabe: Man hat zwei Unbestimmte, m, b. Die **Lösungsmenge des Gleichungssystems** ist dann die Menge alle *Paare* (m, b), welche das System erfüllen. Zusammengefasst lautet das Resultat (typisches Auftreten von Fallunterscheidungen bei äußeren Parametern!):

$$\begin{aligned} \text{Für } x_0 &\neq x_1 \text{ ist } L(x_0, x_1, y_0, y_1) = \left\{ \left(\frac{y_1 - y_0}{x_1 - x_0}, y_0 - \frac{y_1 - y_0}{x_1 - x_0} x_0 \right) \right\}, \\ \text{für } x_0 &= x_1,\ y_0 = y_1 \text{ ist } L\ (x_0, x_0, y_0, y_0) = \{(m, y_0 - mx_0)\,|m \in \mathbb{R}\}, \\ \text{für } x_0 &= x_1,\ y_0 \neq y_1 \text{ ist } L(x_0, x_0, y_0, y_1) = \emptyset \text{ (leere Menge)}. \end{aligned}$$

5. Übersicht zu den Rollen

Nun zur Übersicht über die Rollen von Buchstaben in der Mathematik und ihren Anwendungen in entsprechenden Formeln und Gleichungen:

Art der Gleichung	typische Rollen darin	typische Aktionen
definitorische Formel allgemeingültige Formel (Axiome oder herleitbar)	freie Variablen → unabhängige Var. → abhängige Variablen	Einsetzung von Zahlen und vor allem Rechen- ausdrücken (Termen)
Bestimmungsgleichung (auch System davon)	Unbestimmte → 1 äußere Parameter → 2	1 Auflösen danach 2 Mitschleppen
Lösungsformel bzw. Formel für Lösungen	freie Parameter → 1 äußere Parameter → 2	1 Einsetzen 2 bleiben in ihrer Funktion

Weitere Rollen von Buchstaben in Formeln und Gleichungen

Rolle	Vorkommen	Aktion
Konstante	überall möglich	Mitschleppen, ev. (Näherungs-) Wert einsetzen
Bezeichnung	überall	ein beliebiges gedachtes Objekt wird bezeichnet
Hilfsvariable	überall	ein komplizierter Ausdruck wird abgekürzt, eine Gleichung vereinfacht, usw.
stumme bzw. gebundene Variable	vielfach, insbesondere in Summen und Integralen	keine speziellen Einsetzungen, sondern Durch- führen der zugeh. Aktion mit *allen* Objekten des zugehörigen Bereichs

6. Gleichungen für Geraden und Parabeln in der Ebene

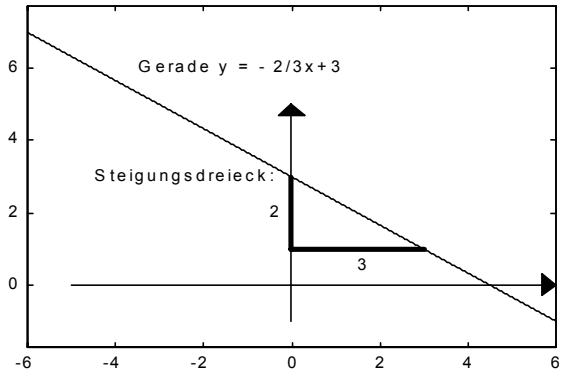
Jede der Formen von Geradengleichungen hat ihre eigene Nützlichkeit:

$$
\begin{aligned}
ax + by &= c, a \neq 0 \text{ oder } b \neq 0 : \text{Bestimmungsgleichung, implizite Form} \\
\frac{x}{a} + \frac{y}{b} &= 1,\ a, b \neq 0 :\ \text{Achsenabschnittsform} \\
y &= mx + b :\ \text{ explizite Form, analog } f(x) = mx + b \text{ (Funktionsform)} \\
y &= y_0 + \frac{y_1 - y_0}{x_1 - x_0}\,(x - x_0)\,,\ x_0 \neq x_1 :\ \text{Zweipunkte-Form} \\
y &= y_0 + m\,(x - x_0)\,,\ x_0 \neq x_1 :\ \text{Punkt-Richtungs-Form}
\end{aligned}
$$

Zugehörige Lösungsmengen (Unbestimmte x, y) - Beispiele für das Aufschreiben unendlicher Lösungsmengen mittels freier Parameter:

$$\begin{aligned} L(x) &= (x, mx+b),\ x \in \mathbb{R},\ \text{für } y = mx+b, \\ L(y) &= \left(\frac{c}{a}, y\right), x \in \mathbb{R},\ \text{für } ax = c,\ a \neq 0\ (\text{Gerade } \parallel y-\text{Achse}) \end{aligned}$$

Man sollte mit dem Graphen einer linearen Funktion der Form $f(x) = mx + b$ umgehen und die Parameter m, b als Steigung und $y-$ Achsenabschnitt geometrisch deuten: Geht man auf der $x-$ Achse von x_0 zu $x_1 \neq x_0$, so gilt für die Funktionswerte $f(x_0) = y_0$ und $f(x_1) = y_1$ *stets:* $m = \frac{y_1 - y_0}{x_1 - x_0}$. Negatives m bedeutet also, dass die Gerade mit steigenden $x-$ Werten *fällt.* Steigung $-\frac{2}{3}$ bedeutet: Geht man auf der $x-$ Achse um 3 nach rechts, so *fällt* der $y-$ Wert um 2. Entsprechend kann man bequem im Schnittpunkt der Geraden mit der $y-$ Achse ein *Steigungsdreieck* ansetzen und damit korrekt die Gerade zeichnen (natürlich immer nur ein Stück, aber man sollte sich klar machen, dass es auch negative Zahlen gibt und dass eine Gerade *beidseitig* unbegrenzt ist). Folgendes Bild illustriert noch einmal die Verhältnisse:



Für Parabeln in der Ebene besprechen wir hier nur die einfachen Formen, die geeignet sind, Parabeln zu beschreiben, welche ihre Achse parallel zur $y-$ Achse haben:

$$\begin{aligned} y &= ax^2 + bx + c,\ a \neq 0: \text{ allgemeine algebraische Form} \\ y &= \alpha\,(x - x_0)^2 + y_0,\ \alpha \neq 0: \text{ Scheitelpunktsform,} \\ &\quad \text{mit Scheitelpunkt } (x_0, y_0) \text{ in Koord.-Darstellung} \end{aligned}$$

Die Scheitelpunktsform entwickelt man leicht durch quadratische Ergänzung aus der ersten Form.

Lösungsmenge der ersten Gleichung (mit x, y als Unbestimmten) ist $L(x) = (x, ax^2 + bx + c), x \in \mathbb{R}$.

7. Motivierende Vorbemerkung

Wir haben im ersten Kapitel recht nützliche Techniken zum elementarsten Umgang mit Mathematik rekapituliert, und die folgenden Kapitel werden etwas weiter reichende mathematische Hilfsmittel auf die Statistik zentriert bereitstellen: Mathematische Grundbegriffe der Statistik und Infinitesimalrechnung als Grundlage aller theoretischen Elemente der Statistik. Diese Vorbemerkung soll ein paar Aspekte der Mathematik nur kurz beschreiben und ein wenig bekanntmachen, damit ein wenig von dem allgemeinen wissenschaftlichen Nutzen erscheine, den man sich mit derlei mathematischem Grundverständnis erschließen kann, damit man also sehe, dass man mit den gelernten Grundbegriffen und Techniken - auch tatsächlich etwas anfangen kann. Man untersucht Phänomene - in der Psychologie ebenso wie in anderen Wissenschaften. Zuweilen lassen sich Phänomene unmittelbar wahrnehmen, sogar verstehen. Wissenschaft nimmt sich der Phänomene an, die dafür zu verborgen oder zu schwierig (komplex) sind. Die Mathematik ist dabei ein unentbehrliches und vielfach sehr weittragendes Mittel dazu, höhere Komplexität zugänglich zu machen. Mathematik erlaubt nicht nur, Dinge genauer quantitativ zu bestimmen, die qualitativ bereits anderweitig verstanden wurden, sondern sie bildet vielfach die Basis dafür, überhaupt zu einer angemessenen Beschreibung einer komplizierten Situation zu gelangen. Daher sollen die folgenden Ausführungen in erster Linie die beschreibende Kraft der Mathematik sichtbar machen und zeigen, dass es „ganz natürlich" ist, mit abstrakten mathematischen Begriffen theoretische „Szenenbilder" zu entwerfen, die als Medium des Verstehens unersetzlich sind. Wenn man einmal gesehen hat, welche grundlegenden geistigen Operation in der Mathematik aufgehoben sind in einer für die allgemeinste Anwendung günstigsten Gestalt, dann ist die Verwunderung darüber nicht mehr allzu groß.

Beginnen wir mit einem ganz allgemeinen Rahmen, in dem alle erdenklichen Phänomene wahrgenommen und formuliert werden: Stets betrachtet man einen Bereich von Objekten (in der Psychologie meist einzelne Menschen oder Gruppen von Menschen) mit wesentlich interessierenden Eigenschaften (Merkmalen). Mathematisch bedeutet dies so viel: Man hat eine Menge von Objekten und gewisse ausgezeichnete Teilmengen (indem man zu einer Eigenschaft die Teilmenge der Objekte bildet, welche diese Eigenschaft haben). Dies ist noch sehr primitiv, wie man von der Mathematik früherer Jahrhunderte her schon sehen kann: Es fehlen noch alle Beziehungen zwischen Objekten (insbesondere zweistellige Relationen, z. B. „kleiner als", aber auch etwa (in einer sozialen Gruppe) „einen höheren Rang haben als", oder sympathischer „Person A mag Person B" oder komplizierter „die Beziehung zwischen Personen A und B ist wesentlich durch C vermittelt"), es fehlen auch alle Operationen (Verknüpfungen), insbesondere die zweistelligen (z. B. Addition), die aus zwei Objekten eindeutig ein Objekt machen. Es ist hier interessant, festzustellen, dass frühe Persönlichkeitspsychologie auf die Betrachtung einstelliger Merkmale beschränkt war (eine Person ist „freundlich", „depressiv", „verbrecherisch" usw.) und dass erst im Laufe eines Entwicklungsprozesses eingesehen wurde, dass dieser begriffliche Rahmen für die Phänomene zu eng ist. Dazu tritt ein weiteres Moment der Verfeinerung, das wiederum auch in der Geschichte der Psychologie gespiegelt auftritt: Es genügt vielfach nicht, von einem Merkmale nur zu sagen, es liege vor oder nicht vor. Natürlicher treten Skalen auf, Grade, in denen eine Eigenschaft zutrifft. Hier bringt die Beschreibung der Grade durch beliebige reelle

Zahlen eine willkommene und bequeme Möglichkeit, auf ideale Weise fein zu unterscheiden (und etwa anschließend sinnvolle Vergröberungen festzulegen). Man sieht jedoch leicht ein, dass dies immer noch nicht genügt: In sehr vielen Fällen benötigt man zur Charakterisierung eines Merkmals zwei oder mehr getrennte Zahlenangaben - so etwas nennen wir einen *Vektor*. (Beispiele: Zwei Zahlangaben - Längen- und Breitengrad etwa - zur Bestimmung eines Ortes auf der Erdoberfläche, drei Zahlangaben für einen Ort oder eine Geschwindigkeit im Raum, viele geordnete Zahlangaben für immer noch einfache physikalische Systeme, ebenso für die Subtestergebnisse eines Menschen bei einem größeren Test.) Damit beherrscht man jedoch nicht nur ein begriffliches Werkzeug für beliebige Genauigkeit - der Wert der quantitativen Beschreibung von Merkmalsausprägungen mittels reeller Zahlen geht viel weiter: Wenn es um die Beschreibung von Zusammenhängen zwischen Merkmalen geht, so hat man mit den mathematischen Funktionen einen riesigen Apparat, beliebige erdenkliche Zusammenhänge zu formulieren, um aus diesen dann etwa bestehende (oder näherungsweise bestehende) im jeweils vorliegenden konkreten Weltausschnitt auszusondern, durch empirische Prüfungen, im besseren Falle jedoch auch durch mathematisch-theoretische Herleitung aus Grundvoraussetzungen, die als besonders bedeutsam und sicher gelten können. Man beachte, dass wir damit unversehens bereits einen Abstraktionsschritt weiter gegangen sind und von Beziehungen zwischen Eigenschaften (oder auch allgemeiner Beziehungen) sprechen. Wir geben zwei einfache Beispiele, diese Bemerkungen zu konkretisieren:

Erstes Beispiel: Man beobachtet grob, dass es so etwas wie Bewegung, Veränderung in der Welt gibt. (Philosophisch hat man tatsächlich eine lange Zeit darüber gestritten, ob es das wirklich gebe oder nur scheinbar, mit Argumenten wie: Es müsse doch ein Bleibendes sein, *das* sich verändere, dies aber ändere sich per definitionem nicht, also sei es gar nichts, was sich ändere, und somit gebe es überhaupt keine Veränderung. Auf der andern Seite stand die unabweisbare Erfahrung der Veränderung.) In nichts löste sich dies Problem damit auf, dass man eine Beschreibung mittels mathematischen Begriffssystems konstruierte: Man stelle alle Zeitpunkte durch reelle Zahlen dar („Zeitachse"), ordne dann jedem Zeitpunkt die Ausprägung des (veränderlichen) Merkmals zu diesem Zeitpunkt (z. B. den Ort zu dieser Zeit, die Temperatur an einem Ort zu dieser Zeit, einen Zustand eines Menschen zu dieser Zeit) zu. Man bildet also eine inhaltlich interpretierte mathematische Funktion (Zeitpunkt t$\mapsto$f(t)=Zustandswert zur Zeit t - man beachte, dass die Zustandswerte Vektoren sein dürfen und meist auch sind). Nun konnte man weiter gehen und etwa feststellen, dass die Weg-Zeit-Funktion für einen frei fallenden Körpern quadratisch ist, nicht linear. Dies konnte man zunächst durch Anpassen an Beobachtungen empirisch finden, später aber auch theoretisch aus der wirkenden Schwerkraft (in deren geeigneter mathematischer Formulierung) herleiten. Damit haben wir den beschriebenen Zusammenhang auf zwei Ebenen: Zunächst werden die beschreibenden Zahlenwerte gekoppelt, dann sogar zwei Beziehungen (Funktionen, nämlich Beschleunigung und Ort als Funktionen der Zeit) in eine Beziehung gesetzt, die sogar völlig allgemein für alle Bewegungen gilt, nicht etwa nur für den freien Fall. Dahinter steckt ein noch bedeutend allgemeinerer Zusammenhang, den man auch bei der Erklärung des Entstehens von Farbeindrücken sowie bei der Wahrscheinlichkeitsrechnung als wesentliches Hilfsmittel benutzt.

Zweites Beispiel: Man stellt fest, dass subjektive Intensität einer Sinneswahrnehmung, eines Lichtes oder Geräusches, nicht einfach proportional zur objektiven

physikalischen Intensität verläuft, und man kann sogar theoretisch je nach (verschiedenartigen) Voraussetzungen eine logarithmische Funktion („Weber-Fechnersches Gesetz") oder auch eine Potenzfunktion herleiten. Daraus erklärt sich dann insbesondere, dass bei kleinen Intensitäten viel feiner unterschieden wird als bei größeren. Logarithmisch: 100 Motorräder klingen subjektiv nicht 100 mal so laut wie eines, sondern der Sprung von 1 auf 10 ist genau so groß wie der Sprung von 10 auf 100.

Mathematik kann von allen erdenklichen Objekten *in strukturellem Zusammenhang* reden, nicht etwa nur von Zahlen oder Beziehungen zwischen ihnen, sondern ebenso von geometrischen Objekten und von logischen Welten. Sie liefert nicht nur ein Begriffssystem, sondern auch ein Anschauungs- und Veranschaulichungssystem. Dies zusammen macht aus, dass Mathematik ein ideales Medium der Modellbildung ist. Wozu braucht man Modellbildung, wenn man sich nicht für die Schönheiten der mathematischen Modelle als solcher interessiert, sondern bestimmte Bereiche der empirischen Realität verstehen will? Zwei wesentliche Gründe gibt es dafür: Einmal ist die Realität stets nur sehr unvollständig bekannt, Reste müssen durch modellhafte Verbindungen gewissermaßen überbrückt werden. Zweitens muss man Modelle bilden, um etwas Einfacheres als die (stets zu komplizierte) Realität haben, um gedankliche Entwicklungen und Prüfungen vornehmen, auch nur interessante Teilbereiche aussondern zu können (Abstraktion!). Schließlich muss man - und das geht nur im Rahmen von Modellvorstellungen - mit Intuition arbeiten, sich erst einmal vorstellen, wie sich eine Sache verhalten *könnte,* welche Sachverhalte *denkbar* wären (mehrere Möglichkeiten, nicht nur das, was auf den ersten Blick als plausibel oder irreführend oft sogar "denknotwendig" erscheint!), um dann herauszufinden, wie sich sich die Sache *tatsächlich* verhält.

Dies ist die Quintessenz des Erfolges aller mathematisierten Wissenschaft: Die Betrachtung wird nicht wie in traditioneller Philosophie darauf gewendet, zu ergründen, *was (dem Wesen nach) die Dinge* sind, sondern *wie die Struktur aussieht, welche die interessierenden Dinge in ihrer Gesamtheit bilden.* Also wird auf die Zusammenhänge zwischen den Objekten geschaut statt auf die Objekte. Von der Ganzheit einer Struktur wiederum macht man sich Modelle, gröbere und feinere, abstraktere und konkretere, auch gleichwertige, die an verschiedene Intuitionen appellieren, an die verschiedene Fähigkeiten des menschlichen Geistes anzukoppeln sind. Insbesondere wird der stark entwickelte Wahrnehmungs- und Anschauungsapparat fruchtbar einbezogen. An Modellen kann man eben Theoretisches wirklich beobachten. Modelle zu bilden, ist überhaupt der wesentliche Schritt in der Entwicklung einer Wissenschaft, und sie kommen auf allen Niveaus vor, vom bloßen ersten Versuch, der vielfach gänzlich verworfen werden muss, bis zum Paradigma für Jahrhunderte oder mehr. Keime der Modellbildung liegen in unserer Anschauung und metaphorischen Übertragungen. Mathematik ist das Medium der Ausarbeitung von Modellen sowie des *vollen* Verständnisses modellhafter Idealisierungen. Auf einer höheren Stufe ist sie längst auch zu einer eigenständigen Quelle von Metaphern geworden.

Diese Bemerkungen sollten mit konkreten Beispielen deutlichere Konturen gewinnen, aber von Anfang an zwei geläufige Missverständnisses ausschließen, die sich gegenseitig in unseliger Weise stützen:

Das erste Missverständnis möchte ich das *mechanistische* nennen: Mathematik ist danach eine Technik, ein Mechanismus, um "wie auf Schienen" sichere Ergebnisse auszurechnen. Man steckt seine Daten hinein, und ohne Nachdenken kommen die Resultate. Der Interpretationsarbeit ist man enthoben, und gerade darin wird der

Wert der Mathematik gesehen, Objektivität und Exaktheit. Das zweite Missverständnis schließt unmittelbar daran, es ist das von der beschränkten Anwendbarkeit der Mathematik: Mathematik könne als ihrerseits exakt und mechanisch nur auf Exaktes und Mechanisches angewandt werden, man denkt etwa an physikalische Mechanik. Dies zu überdenken, sollte bereits das oben angeführte Beispiel von Bewegung, Veränderung anregen. In diesem Zuge sieht man auch noch eine Beschränkung auf das Quantitative. Weiter stellt man sich vor, dass sowohl die Fragestellungen als auch die Wege zur Beantwortung eindeutig fixiert sein müssen, *dann* könne man das mechanische System der Mathematik darauf loslassen und Antwort erhalten, nur eben in allen interessanteren Fällen nicht. Dazu wäre so viel zu sagen, dass es mehrere Vorlesungen füllen würde. Es sei nur festgestellt, dass diese Vorstellungen völlig absurd sind, und zur Anregung bemerkt: Weder die Ziele, noch die Wege liegen in der Mathematik vorher fest, sondern sie entwickeln sich, wie auch sonst. Mathematisches Verständnis wird in der Regel erst dann erreicht, wenn verschiedene Wege zum selben Resultat gesehen und verstanden wurden. Viele mathematische Strukturen, die mit dem Blick auf einen bestimmten mathematischen Problemkreis entwickelt wurden, erweisen sich notorisch später als fruchtbar für ganz andere Gebiete, auch in empirischen Wissenschaften. Schließlich eröffnet Mathematik vielfach erst dem Blick für das Bestehen *verschiedener* Möglichkeiten. Bereits einfachste formale Logik zeigt oft, dass es unter gewissen Bedingungen noch nicht feststeht, ob eine Aussage gilt oder nicht, und entwickelte mathematische Logik hat das tiefe philosophische Resultat erbracht, dass es kein korrektes formales System geben kann, das alle mathematische Wahrheit enthält. Gödel hat formal bewiesen, dass das Sehen von Wahrheit in der Mathematik auf ständig neue Ideen angewiesen ist. Menschenverstand" zugänglich, nicht erst dem mathematischen Talent. Allerdings will die erstere durchaus verbreitete Gabe auch ein wenig mehr als alltäglich üblich strapaziert und geübt werden. Weiter fügt es sich günstig, dass dabei gleichzeitig auch die allgemeine Fähigkeit gestärkt wird, mit komplizierteren sowie abstrakteren Sachverhalten umzugehen - die abstrakten Grundbegriffe sowie quantitative Beschreibung als solche haben ihre Bedeutung weit über die Statistik hinaus. Insbesondere soll dieser Kurs verdeutlichen, dass Mathematik eine immense beschreibende Kraft besitzt (die ihre Erfolge bei Anwendungen auf nahezu allen Gebieten der Wissenschaft erklärt), und dass Mathematik nicht nur für speziell Begabte, den individuellen Eigenschaften ergeben, welche nicht.) Eine noch höhere Stufe wird bei entwickelter Wissenschaft stets wichtig: Beziehungen von Beziehungen: Wie sieht das Gefüge der Beziehungen in einer sozialen Gruppe aus, z.B.: Welche Eigenschaften einer Beziehung zieht welche einer andern nach sich, usw. Oder: Wie sieht die gesamte Kommunikationsstruktur einer Gruppe aus? So gelangt man schließlich zur Betrachtung ganzer Strukturen (Mengen von Objekten mit ihren interessierenden Beziehungen) und zu Beziehungen zwischen Strukturen (etwa beim Verhältnis zweier sozialer Gruppen zueinander). Endlich werden umfassendere, abstraktere Strukturen gebildet, bei denen die Objekte bereits Strukturen sind. (Beispiel: Ein Netz von sozialen Institutionen.) Die Mathematik hat nicht nur zuerst einen solchen Begriffsaufbau benutzt, sie hat ihn auch völlig abstrakt dargestellt - für jeden derartigen Begriffsaufbau, von beliebigen Begriffen eines beliebigen Feldes - und wichtige Resultate dazu geliefert. Darüber hinaus werden mathematische Begriffe und Strukturen stets dann bedeutsam, wenn man eine komplexe Struktur genauer beschreiben möchte. Z.B. wird eine Analyse des Baus von Kommunikationsstrukturen ohne ausdrückliche Verwendung von Mathematik

nicht gelingen. Stellen wir zur Anregung dies Problem: Man beschreibe Kommunikationsstrukturen geeignet so, dass darin Begriffe wie „zentralistischer Charakter" präzise und adäquat ausgedrückt werden können.

Diese Bemerkungen sollten nur andeuten, dass man tendenziell für interessantere und tiefere Resultate mit der Mathematik zu tun bekommt, und ein Gegengewicht zur verbreiteten Auffassung bilden, Mathematik sei gerade für das Kompliziertere, Lebendigere nicht verwendbar. Diese Auffassung ist vielfach mit dem Aberwitz verbunden, bei der Rede von besonders Kompliziertem mit außerordentlich primitiven Strukturen auskommen zu können. Selbstverständlich müssen wir für diesen Kurs zu Einfacherem zurückschalten, und es lohnt sich durchaus, auf der ersten der genannten Stufen wieder zu beginnen, bei den (einstelligen) Eigenschaften von interessierenden Objekten, nicht nur wegen ihrer Einfachheit, sondern auch, weil es bereits hier eine notwendige und nicht selbstverständlich verfügbare Verfeinerung gibt: Von qualitativen Begriffen („weiblich-nicht weiblich", „klein-mittelgroß-groß") zu quantitativen Begriffen (Größen): Statt zu sagen, jemand sehe sehr viel fern, wäre es viel informativer, mitzuteilen, wie viele Stunden der betreffende Mensch im Mittel täglich fernsieht, noch informativer, wie sich das auf einzelne Sendungstypen aufgliedert. Variablen sind Merkmale, deren jeweilige Ausprägung durch eine Zahlangabe (oder auch eine geordnete Folge mehrerer Zahlangaben) zu beschreiben ist. Man beachte, dass qualitative Begriffe nicht etwa eine höhere Dignität besitzen, sondern einfach nur einen primitiveren Spezialfall darstellen: Für „weiblich-männlich" kommt man z.B. mit den Zahlen 1,0 aus, und alle Statistik, die man für Variablen betreibt, ist insbesondere auch für soch einfache Spezialfälle anwendbar und gültig. Was Variablen attraktiv macht für diesen Kurs, ist nicht nur ihre Allgegenwart in allen Bereichen und die innewohnende Kraft verfeinerter Beschreibung, sondern auch die Tatsache, dass sie selbst ein besonders einfaches Beispiel des allgemeinsten und nützlichsten Begriffs der Mathematik darstellen: Es handelt sich um den Begriff der Zuordnung (Abbildung, bei Zahlen auch gern „Funktion" genannt). Bei der Variablen „Lebensalter (in einer Population von Menschen)" wird *jedem* Menschen dieser Population *sein* Lebensalter zugeordnet (in ganzen vollendeten Jahren oder feiner, wie man will). Ebenso für „Anzahl der Schuljahre", „Anzahl der (pro Jahr z.B.) gelesenen Bücher" usw. Welche fundamentale Rolle dieser Begriff der Zuordnung mittlerweile bei der Beschreibung von komplizierteren Sachverhalten spielt, ist kaum zu ermessen. Insbesondere können wir den gesamten Begriffsapparat der Statistik mit diesem Begriff aufbauen, beginnend mit dem grundlegenden Begriff der Statistik überhaupt: „Verteilung einer Variablen". Dabei ordnet man (im einfachsten Fall) jedem Zahlenwert die relative Häufigkeit zu, mit der er als Wert der Variablen auftritt. Es bereitet dem Anfänger gewöhnlich Schwierigkeiten, von einzelnen Zahlangaben zur Zuordnung aufzusteigen, aber dafür auch mit Erreichen dieser Stufe ein großer Schritt getan. Der Zugang ist eröffnet zur Beschreibung zeitlicher Entwicklungen, auch anschaulich durch Kurven, allgemeiner zur Beschreibung von Zusammenhängen zwischen Variablen (die man idealtypisch durch mathematische Funktionen beschreibt und deren Abweichung vom Idealtypus man wiederum mit Statistik in den Griff bekommt). Wir steigen damit wieder auf zu den Beziehungen von Beziehungen. Im einzelnen sollen die Beziehungen zwischen Stichproben und Gesamtpopulation und die „Korrelation von Variablen" genauer betrachtet und ausgeführt werden.

KAPITEL 2

Deskriptive Statistik

1. Der Begriff „Variable" (oder „Größe")

1.1. Merkmale und Merkmalsausprägungen. Psychologisch interessierende Merkmale von Personen sind z.B.: Lebensthematik, soziale Einbindung, Erfolgsorientiertheit / Misserfolgsorientiertheit, Introversion /Extroversion, Deprivationen (verschiedenster Arten), aber auch so etwas wie eine spezielle oder eine komplexere Fähigkeit oder berufliche Eignung, oder Merkmale der Sinneswahrnehmung. Alle diese Merkmale haben *verschiedene Ausprägungen*: Manchmal gibt man nur einen Typus an wie „geeignet/ungeeignet", vielfach jedoch konstruiert man eine feinere Skala von zahlenmäßig auszudrückenden Ausprägungen, und dann fragt sich, ob man eigentlich nur eine Ordnungsbeziehung hat oder nicht einmal eine solche, oder ob sogar die Differenzen von Zahlwerten etwas bedeuten. Das Merkmal „Geschlecht" etwa hat nur ganze zwei Ausprägungen, keine Differenzbildung, keine Ordnung. Wenigstens eine Ordnung wird man bei Merkmalen wie Introverion haben, dagegen wird man (eindimensionale) Fähigkeiten oft feiner messen können, so dass auch Differenzen von Werten etwas Genaues bedeuten. Schließlich wird man für komplexe Merkmale (z.B. komplexere Fähigkeiten) mehrere Dimensionen benötigen, also die jeweilige Ausprägung durch eine *geordnete Folge* von Zahlen, also einen *Vektor* beschreiben. Dann hat man ein Geflecht mehrerer Variablen (für die einzelnen Dimensionen) zu betrachten, und später werden wir unter den Namen „Regression" und „Korrelation" etwas zu diesem Sektor kennenlernen. Bei all den wichtigen Differenzierungen: Stets kann man die Ausprägungen zahlenmäßig erfassen, stets kann man sinnvoll Statistik betreiben. Es sei insbesondere darauf verwiesen, dass auch bei einem Hauptinteresse an *einzelnen Individuen* Statistik keineswegs bedeutungslos wird: Einmal ist anhand des Verhaltens großer Zahlen von Individuen auch einzuschätzen, was unter welchen Bedingungen bei einem *einzelnen* zu erwarten ist (oder welche Wahrscheinlichkeiten für bestimmte Entwicklungen bestehen). Zum andern kann auch die Betrachtung eines einzelnen Individuums über gewisse Zeiträume selbst ein wiederholtes Zufallsexperiment darstellen: Ein Individuum wird z.B. nicht in jeder Situation mit einer Aufgabe bestimmten Typs erfolgreich umgehen, aber es gibt einen Erwartungswert für dies Individuum.

Halten wir fest: Die Ausprägungen eindimensionaler Merkmale lassen sich stets durch Zahlen beschreiben. Ob dabei feinere Strukturen sinnvoll sind, hängt von der Willkürlichkeit einer solchen Zahlbeschreibung ab. Die Ausprägungen mehrdimensionaler Merkmale lassen sich dagegen stets mit Folgen von Zahlenwerten beschreiben. Wir werden nun über weiteste Strecken bei eindimensionalen Merkmalen bleiben.

Nachdem wir den Schritt vollzogen haben, die Ausprägungen von Merkmalen durch Zahlen zu beschreiben, liegt es nahe, von dem schwerfälligen Sprachgebrauch

„Merkmal", „Merkmalsausprägungen" überzugehen zu „Variable" (oder „Größe"), „Werte der Variablen - oder der Größe". Man achte jedoch auf die Unterscheidung und merke sich simpel: Eine Variable ist ein Ding, das mehrere Werte haben kann. Natürlich interessiert uns gerade die Variation der Werte - hätten alle Individuen denselben Wert, so könnte man damit nichts unterscheiden, also nichts Interessantes beschreiben. Unbeschadet dessen ist es mathematisch zweckmäßig, auch als Grenzfälle Variablen mit konstantem Wert zu betrachten, diese sind einfach günstig bei manchen Rechnungen. Wir schließen diesen Fall daher ein und nicht aus. Von überragender Bedeutung für das Grundverständnis ist die nun folgende wesentlich genauere mathematische Erklärung des Begriffs der Variablen.

1.2. Der mathematische Begriff der Variablen. Man beachte, dass „Variable" noch für Buchstaben in Formeln wie $(x+y)^2 = x^2+2xy+y^2$ gebraucht wird, dass dies jedoch ein *anderer* Begriff ist! Es gibt noch weitere mathematische Verwendungsweisen des Wortes „Variable" - man beachte: für verschiedene Begriffe! Zum Beispiel kennen Sie „unabhängige" und „abhängige" Variable. Damit kommen wir der Sache schon näher, wir werden nämlich sehen, dass „Variablen" der Statistik (im Grundbegriff!) sinngemäß nichts anderes als *Beispiele* für abhängige Variablen sind. Dies ist nun ein ziemlich altväterlicher Sprachgebrauch, den wir sogleich moderner erklären werden, mit dem Abbildungsbegriff. (Später lasse man sich nicht dadurch verwirren, dass im Kontext mit Regression wieder „Variablen" der Statistik als abhängige und unabhängige zu betrachten sind. Vielmehr erkenne man darin gerade die Tugend der Mathematik, dass in hundertfältiger Variation ihre Grundbegriffe auch auf höhreren Ebenen immer wieder anwendbar sind!) Knüpfen wir an den Gebrauch von Zahlen zur Beschreibung von Merkmalsausprägungen: Jemand ist 20 Jahre alt (in ganzen vollendeten Jahren), sieht täglich im Durchschnitt (hoffentlich nur) eine halbe Stunde fern. Zu einem andern Menschen der zu betrachtenden Gruppe mögen andere Werte gehören. Das ergäbe eine lange Liste. Wir wollen darüber reden, was hier und in unzähligen weiteren Beispielen geschieht, wollen außerdem praktischere Information als derartig lange Listen bereitstellen. Dazu muss man nur dies erfassen: Es wird eine Menge von Objekten (etwa Menschen einer bestimmten Altersgruppe in einem Gebiet zu einer Zeit) betrachtet, in der Statistik „Population" genannt. *Jedem* dieser Objekte wird *genau ein* Zahlenwert zugeordnet (die Ausprägung des interessierenden Merkmals). Eine solche Zuordnung nennt man eine *Variable*. Man erfasst sie auf abstrakterer Ebene völlig mit drei Informationen:

- Definitionsbereich = Menge der Objekte, denen etwas zugeordnet wird
- Wertebereich, eine Menge von Objekten, zu der alle zugeordneten Objekte gehören (bei Variablen stets die Menge der reellen Zahlen, die wir $\mathbb{R}$ nennen)
- Eine genaue Zuordnungsvorschrift: Sie besagt als allgemeine Regel, welches Objekt jedem einzelnen Mitglied des Definitionsbereiches zugeordnet wird.

Die in der Mathematik übliche symbolische Beschreibung einer solchen *Zuordnung* oder *Abbildung* sieht so aus:

$$\begin{array}{rccc} f: & A & \to & B \\ & a & \mapsto & f(a) \end{array}$$

Die erste Zeile ist zu lesen: „f geht von A nach B", A ist also Definitionsbereich und B Wertebereich. Die zweite Zeile liest man: „(Dem beliebigen Element) $a \in A$

wird das Objekt $f(a)$ (lies: „f von a") zugeordnet. Man beachte vor allem den Unterschied zwischen f - das ist die Abbildung selbst - und $f(a)$, das ist ein Element aus B.

DEFINITION 2. *Eine Abbildung von einer Menge A in eine Menge B ist eine Zuordnung, die jedem Element von A genau ein Element von B zuordnet.*

Erstes Beispiel:

$$\begin{array}{rccc} f: & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & x^2 \end{array}$$

Dies ist die Quadratfunktion, deren Graph die bekannte Parabel in Normalgestalt ergibt. Man beachte, dass diese zwei Zeilen die gesamte Liste $f(1) = 1, f(2) = 4, f(-3) = 9, f(2/3) = 4/9$,... erfasst, die niemals zu einem Ende käme, in der auch Einträge für Zahlen wie π nicht mit exaktem Wert anzugeben wären. Besonders wichtig ist es, dass auch allgemeinere Aussagen in der Liste stecken wie
$f(x^2 + y) = (x^2 + y)^2 = x^4 + 2x^2y + y^2$. (Letztere Gleichung mit binomischer Formel.)
Zweites Beispiel:

$$\begin{array}{rccc} X: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & X(\omega) \end{array}$$

Dies ist keine konkrete Abbildung, sondern nur das definitorische Schema (mit den üblichen Bezeichnungen) für eine Variable im Sinne der Statistik. Ein konkretes Beispiel könnte so aussehen:

$$\begin{array}{rccc} \mathit{Alter}: & \mathbb{W} & \to & \mathbb{R} \\ & S & \mapsto & \mathit{Alter} \text{ von } S \text{ in voll. Jahren} \end{array}$$

(Der Definitionsbereich $\mathbb{W}$ wäre noch zu präzisieren, etwa: „Menge aller eingeschriebenen Wuppertaler Studenten am 16.10.2000". Hier wäre also $X = \mathit{Alter}$ in voll. Jahren in der genannten Population von Studenten - so ausführlich muss das auch jeder Sozialwissenschaftler in Worten angeben! Man merke sich: Der Definitionsbereich einer Variablen in der Statistik heißt „Population". Dabei ist es wichtig, „Alter" nicht als eine Zahl zu verstehen, sondern als die gesamte Zuordnung. Das Alter eines bestimmten Studenten ist $\mathit{Alter}(S)$ - im altertümlichen Sprachgebrauch wäre genauer $A(S)$ die abhängige Variable, und S wäre die unabhängige. Aber wir wollen das genauer und einfacher so verstehen: S bezeichnet ein *beliebiges* Element von $\mathbb{W}$, $A(S)$ ist dann eine eindeitig gestimmte Zahl. A ist dagegen *keine Zahl* und auch nicht etwa die Menge aller Zahlen $A(S)$ für S aus $\mathbb{W}$ - diese nennt man korrekt die Menge aller Werte von A oder Bild von A, und die Identifikation von A damit führt zu völlig Unsinnigem (ungeachtet dessen steht das so in miserablen deutschsprachigen Statistikbüchern). Man sieht den wichtigen Unterschied zwischen A und $A(S)$ sehr schön daran, dass es sinnvoll ist, von einem Mittelwert und einer Streuung von *Alter* als einer Variablen zu sprechen. Eine konstante Zahl dagegen hat allenfalls sich selbst als Mittelwert und Streuung Null, das gibt also keine Information extra. In den Sozialwissenschaften würde man das letzte Beispiel etwa so in Worten benennen (und dasselbe meinen, nämlich die Zuordnung und

keine einzelne Zahl!): „Alter in vollendeten Jahren bei den Wuppertaler Studenten, die am 16.10.2000 eingeschrieben waren". Genau genommen wird allerdings A erst zu einer Variablen im Sinne der Statistik, indem man so etwas wie Zählungen von Häufigkeiten beginnt oder abstrakter Wahrscheinlichkeiten dafür angeben kann, dass Werte in gewissen Bereichen auftreten. Das ist gemeint mit der „Verteilung von A", worauf wir noch ausführlich kommen werden. Jedenfalls wird man sich nunmehr nicht mehr zu sehr wundern, dass man die Variablen in der Statistik oft unter dem ausführlicheren Namen der „Zufallsvariablen" oder „zufälligen Größe" findet. Diesem Schema folgen alle bereits gegebenen und alle nachfolgenden Beispiele. Wir heben daher noch einmal hervor:

DEFINITION 3. *Eine Variable ist eine Abbildung von irgendeiner Menge Ω in die Menge $\mathbb{R}$ der reellen Zahlen. Symbolisch: $X : \Omega \to \mathbb{R}$. Mit $X(\omega)$ wird der Wert der Variablen bei einem beliebigen Populationsmitglied ω aus Ω bezeichnet. (Zusatz: Zu einer Variablen im eigentlichen Sinne der Wahrscheinlichkeitstheorie oder Statistik wird X erst dadurch, dass man weiter noch einen Wahrscheinlichkeitsbegriff auf Ω hat. Eine einfache Realisierung eines solchen Begriffes bei endlicher Population Ω besteht darin, dass man die relativen Häufigkeiten betrachtet, mit denen Werte von X aus beliebigen Bereichen vorkommen, z.B. auch die relative Häufigkeit, mit der ganz bestimmte Werte in der Population vorkommen.)*

Ein wichtiger Punkt sei noch hervorgehoben, der mir selbst bei aufrichtig bemühten und keineswegs unfähigen fortgeschrittenen Studenten noch wiederholt begegnete: Es genügt einfach nicht, eine Variable mit der Zuordnungsvorschrift allein zu definieren, etwa als „Alter in vollendeten Jahren": Es macht keinen Sinn, etwa die Menge aller Objekte zu betrachten, denen man ein Alter zuordnen kann, alle Tiere und Untertassen und die Planeten wären dabei. Es macht dagegen wohl einen Sinn, etwa eine physiologische Variable auf einer Population von „Kranken" und einer Population von „Gesunden" zu unterscheiden und die Verteilungen miteinander zu vergleichen - so etwas nutzt man für diagnostische Zwecke. Man mache sich klar: Wir haben in solchen Fällen mit zwei verschiedenen Variablen zu tun, und dafür genügen die verschiedenen Populationen, wenn auch die Zuordnungsvorschrift dieselbe ist. (Wenn wir zu den Stichproben kommen, müssen wir sogar die (neue) Population aller Stichproben (aus einer vorgegebenen Population) festen Umfangs ins Auge fassen!)

2. Die elementaren Grundbegriffe der deskriptiven Statistik

Eine Vorbemerkung zum Zusatz „deskriptiv" (d.h. beschreibend): Wir wollen nur die wesentlichen statistischen Eigenschaften einer Variablen (auf ihrer „Gesamtpopulation" oder einfach „Population" Ω beschreiben, vorerst noch nicht anhand unvollständiger Information durch eine Stichprobe auf diese Eigenschaften schließen („schließende Statistik" oder „Inferenzstatistik") - vgl. dazu Kapitel 4.

2.1. Variablen mit wenigen Werten und ihre Verteilung, Mittelwert und Streuung. Den wichtigsten Begriff haben wir bereits eingeführt: Statistik handelt von Variablen, also Abbildungen $X : \Omega \to \mathbb{R}$, wobei Ω eine beliebige Menge von irgendwelchen Objekten ist. (Alle zugehörigen Überlegungen lassen sich gewöhnlich recht einfach auf vektorielle Variablen verallgemeinern, bei denen die Werte eben nicht einfach Zahlen, sondern Vektoren sind.)

Wir bemerkten, dass die volle Information über eine Variable X gewöhnlich in einer fürchterlichen Liste $(\omega \mid X(\omega))$ bestünde, bei unendlich großem Ω schließlich überhaupt nicht gegeben werden könnte. Das ist ganz anders bei mathematischen Abbildungen, für die man einen Rechenausdruck besitzt, wie $f(x) = x^2$. Darin steckt *alle* Information über die Quadratfunktion. Es stellt sich daher folgendes

Problem: Kann man im Falle einer statistischen Variablen zu einer ökonomisch eleganten mathematischen Beschreibung gelangen?

Offenbar geht das nicht, solange man die *volle* Information fordert, z.B. welche Körperlänge jeder einzelne Bundesbürger hat. Man muss daher Information wegwerfen, die vielleicht nicht gar so wichtig ist. Genau dies bewirkt auf geeignete Weise die *entscheidende statistische Abstraktion:*

Man fragt nur noch danach, *welche Werte* von einer Variablen X *wie häufig* vorkommen. Diese Information nennt man *die Verteilung* von X.

Wir nehmen *für diesen Abschnitt* zwei wesentliche Beschränkungen für unsere Betrachtungen vor (erstere wurde schon im Titel vermerkt):

- Wir reden nur von der Verteilung einer Variablen X in der *Gesamtpopulation* Ω, *nicht von Stichproben.*
- Wir setzen voraus, dass der Definitionsbereich Ω von X *endlich* ist (Spezialfall von „diskret verteilten Variablen", bei denen die Werte diskret auseinander liegen, kein Kontinuum bilden).

DEFINITION 4. *Sei $X : \Omega \to \mathbb{R}$ eine Variable mit endlichem Ω. Dann ist die Verteilung von X definitionsgemäß folgende Abbildung:*

$$\begin{array}{rccl} f_X: & \mathbb{R} & \to & \mathbb{R} \\ & a & \mapsto & \textit{relative Häufigkeit, mit der } a \textit{ als } X\textit{-Wert vorkommt.} \end{array}$$

Eine solche Verteilung kann man mittels eines Stabdiagramms veranschaulichen. Man beachte: Nunmehr steht die unabhängige Variable a der Funktion f_X für eine beliebige reelle Zahl, und $f_X(a)$ ist die relative Häufigkeit, mit welcher dieser Wert a als Wert der Variablen X auftritt. Kommt a gar nicht vor, so ist die absolute Häufigkeit Null, also auch die relative - das ist die absolute geteilt durch die Anzahl aller Populationsmitglieder. (Die prozentuale Häufigkeit ist die relative mal hundert, oder auch: Die relative Häufigkeit ist die Häufigkeit pro 1 wie die prozentuale die pro hundert ist.)

Beispiel: In einer bestimmten Population von Studenten könnte man folgende Verteilung der Semesterzahlen gefunden haben (Tabelle zu f_X, wobei X die Variable „Semesterzahl in dieser Poplation" ist):

Semesterzahl	0	1	2	3	4	5	6	7
relative Häufigkeit	0.18	0.13	0.1	0.1	0.09	0.08	0.08	0.07
Semesterzahl	8	9	10	11				
relative Häufigkeit	0.07	0.05	0.04	0.01				

Hier ist ein Stabdiagramm dazu:

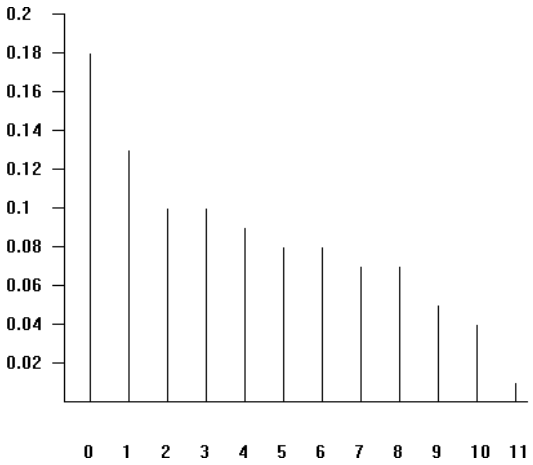


Abb. 1: Stabdiagramm der relativen Häufigkeiten, welche auf die einzelnen Variablenwerte entfallen - hier Semesterzahlen

Für manche Zwecke möchte man noch weiter reduzieren; tatsächlich genügen vielfach bereits zwei Zahlangaben, um eine Verteilung hinlänglich genau zu beschreiben, Mittelwert und Streuung. Der Mittelwert hat eine anschauliche Bedeutung im Rahmen eines Stabdiagramms: Man stelle sich die Stäbe vor mit einem Gewicht, das proportional zu ihrer Länge ist. Der Punkt auf der Größenachse (waagerechten Achse, also der Achse für die Werte der Variablen), an dem man unterstützen muss, um das Ganze im Gleichgewicht zu halten, das ist der Mittelwert. Rechnerisch bekommt man ihn so, dass man alle Einzelwerte mit ihrer relativen Häufigkeit multipliziert, das Ganze dann addiert. Dies kann man auch noch auf andere Weise verstehen: Man nehme die einzelnen X-Werte, zu den einzelnen Populationsmitgliedern, mit ihren Wiederholungen, bilde von diesen Werten das arithmetische Mittel, also die Summe von allen durch den Populationsumfang geteilt. Dann kommt dasselbe Resultat; denn in unserer Liste der relativen Häufigkeiten hat man einfach die wiederholten Werte zusammengefasst. Beispiel:

Mit Einzelwerten 0,0,0,1,1,2,2,3,3 hätte man arithmetisches Mittel $12/9 = 4/3$. Die Werte 0,1,2,3 hätten der Reihe nach relative Häufigkeiten: 1/3, 2/9, 2/9, 2/9, mit der Summe der Produkte „Wert mal relative Häufigkeit" erhielte man: $0 \cdot 1/3 + 1 \cdot 2/9 + 2 \cdot 2/9 + 3 \cdot 2/9 = 12/9 = 4/3$. Die Übereinstimmung beruht auf dem Distributivgesetz: $(3 + 3)/9 = (2 \cdot 3)/9 = 3 \cdot (2/9)$. Dies ist der Beitrag zum Mittelwert, der von den beiden Dreien kommt, zuerst in der Version, wie man das arithmetische Mittel der Einzelwerte berechnet, zuletzt in der Version: Wert mal relative Häufigkeit.

Wir fassen in einer definitorischen Formel zusammen:

DEFINITION 5 (Mittelwert einer Variablen). *(Vorausgesetzt ist wiederum, dass X nur endlich viele Werte annehme.) Der Mittelwert einer Variablen X mit der Verteilung f_X lautet:*

$\mu(X) = \sum_{X-Werte\ a} a \cdot f_X(a)$. *(Auch einfach nur „$\mu$", wenn die Variable klar ist.)*
Bemerkung: Sind mit x_i, $i = 1...n$, alle einzelnen Werte der Populationsmitglieder mit Wiederholungen aufgezählt, so gilt:
$\mu(X) = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Es sei bemerkt, dass die erste Formel auch dann verwendbar ist, wenn es nur endlich viele $X-$ Werte a gibt, für welche $f_X(a) \neq 0$. Insbesondere kann die Formel auch verallgemeinert angewandt werden, wenn es sich um Wahrscheinlichkeiten (statt relative Häufigkeiten) handelt. In beiden Hinsichten ist die zweite Formel weniger allgemein, sie setzt sowohl die Endlichkeit der Population als auch (beim Arbeiten mit Wahrscheinlichkeiten) die gleiche Wahrscheinlichkeit für jedes Populationsmitglied voraus. (Man vergleiche spätere Ausführungen, Stichworte: Zufällige Variablen, Verteilungen mit Wahrscheinlichkeitsdichten.)

In unserem tabellarischen Semesterzahl-Beispiel erhält man den Wert $\mu = 3.88$.

Nun zur zweiten charakteristischen Zahl für eine Verteilung: Die Streuung ist ein Maß für die Breite der Verteilung, die sich anschaulich auch als Breite des Stabdiagrammbildes darstellt. Aber man nimmt dazu nicht etwa die Breite des Intervalls, in dem die Werte liegen, sondern man gewichtet wieder: Wenige „Ausreißer" erhöhen die Streuung nicht nennenswert, wohl aber das Auftreten entlegener Werte mit nennenswerter relativer Häufigkeit. Intuitiv läge es nahe, den mittleren absoluten Abstand der Einzelwerte vom Mittelwert zu nehmen, doch wählt man in aller Regel ein etwas anderes Maß: Man bildet den Mittelwert der *quadrierten* Differenzen zum Mittelwert, anschließend zieht man die Wurzel, um ein Maß zu bekommen, das als Breite auf der Größenachse interpretierbar ist. Der Grund ist folgender: Wenn man gewisse Elemente der Wahrscheinlichkeitsrechnung und Statistik mathematisch tiefer durchdenkt, so stößt man bei der Beschreibung von mathematisch idealen Verteilungen wie Normalverteilungen und weiteren sehr häufig auf diese „Streuung" als Parameter. Sie ist also systematisch wichtiger. Im übrigen haben quadratische Differenzen noch den Vorteil der Differenzierbarkeit, was sie auch sonst für Fehlermessungen günstiger macht als absolute Differenzen.

Wiederum fassen wir in einer definitorischen Formel zusammen:

DEFINITION 6 (Varianz und Streuung einer Variablen). *Eine Variable X mit der Verteilung f_X hat folgende Streuung:*

$\sigma(X) = \sqrt{\sum_{X-Werte\ a} (a - \mu(X))^2 \cdot f_X(a)}$. *(Oder einfach nur „$\sigma$".)*
Bemerkung: Das Ganze ohne die Wurzel ist zuweilen nützlich:
$\sigma^2(X)$ heißt Varianz von X (symbolisch auch: „Var(X)").
Bemerkung: Mit den Einzelwerten $x_1, ..., x_n$ sieht die Varianz so aus:
$\sigma^2(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu(X))^2$.

(Dieselben Bemerkungen zum Gültigkeitsbereich wie bei der vorigen Definition.)

In unserem Tabellen-Beispiel erhält man: $\sigma^2(X) = 9.87$ (gerundet), und die Streuung ist $\sigma(X) = 3.14$.

Wir wollen etwas genauer charakterisieren, was die Streuung besagt. Zunächst einmal kann man leicht feststellen, dass der Übergang von einer Variablen X zur Variablen $2X$ bewirkt, dass die Streuung sich verdoppelt: $\sigma(2X) = 2\sigma(X)$. Das folgt sofort aus der Rechenformel. Das Verteilungsbild zur Variablen $2X$ ist gegenüber dem zur Verteilung von X einfach mit Faktor 2 längs der Größenachse (waagerechten Achse) gestreckt. σ ist also wirklich ein Maß für die Breite. Andererseits ist σ nicht in allen Fällen unmittelbar geometrisch interpretierbar. Genau gibt σ eine Beschränkung für die relative Häufigkeit der Variablenwerte, die einen vorgebbaren Mindestabstand von μ haben. Zum Beispiel können außerhalb des Intervalls $\mu \pm 2\sigma$ *höchstens* 25% der Population liegen, mit mathematischer Sicherheit. Allerdings kann man die Faustregel angeben, dass im allgemeinen sogar weniger als 10% außerhalb dieses Intervalls liegen, mithin über 90% im Bereich $\mu \pm 2\sigma$ liegen. Bei Normalverteilungen sind es sogar ziemlich genau 95%.

Somit gibt die Streuung (insbesondere) an, in welchem Bereich um den Mittelwert der Löwenanteil der Population liegt, nämlich $\mu \pm 2\sigma$; Achtung: 2σ, nicht σ! Diese Aussage ist viel nützlicher als die Angabe eines Riesenintervalls, in dem mit Sicherheit alle Variablenwerte liegen. Schauen wir zur Konkretisierung noch einmal an, was in unserem Beispiel herauskommt: Bei der betrachteten Semesterzahlverteilung reicht das Intervall $\mu \pm 2\sigma$ von 0 bis 10, diese einschließend. Nur der Wert 11 liegt außerhalb, also nur 1% der Population in diesem Fall.

Man bedenke jedoch stets, dass μ und σ im allgemeinen nicht ausreichen, die gesamte Verteilung zu rekonstruieren - dies ist nur zuweilen der Fall, wenn man schon weiß, dass es sich um eine Verteilung eines gewissen Typs handelt, die tatsächlich vollständig mit diesen Parametern bestimmt ist. So verhält es sich z.B. bei Normalverteilungen.

Hat man die einzelnen Werte $x_1, ..., x_n$ einer Variablen X bei verschiedenen Populationsmitgliedern beobachtet, so definiert man auch:

DEFINITION 7 (arithmetisches Mittel der Werte $x_1, ..., x_n$).

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + ... + x_n}{n}$$

Dass diese Definition formal genau wie die von $\mu(X)$ lautet, verführt den Anfänger immer wieder dazu, beide miteinander zu verwechseln. Man beachte dazu genau: Die Bezeichnung $\overline{x}$ steht für das arithmetische Mittel einer bloßen Stichprobe aus der Population. Erst dann, wenn die Stichprobe (was realistisch in den meisten Fällen nicht so ist) die gesamte Population umfasst, fallen beide zusammen, im Normalfalle ist eine Stichprobe viel kleiner als die Population, und dann liefert $\overline{x}$ einen sehr wichtigen Schätzwert für $\mu(X)$, auf den wir noch sehr genau eingehen werden. (Analog wird ein aus einer Stichprobe zu gewinnender Schätzwert s für die Streuung einzuführen sein, dazu später mehr.) Gerade dann, wenn man „nur" eine Stichprobe besitzt und die Qualität dieses Schätzwertes $\overline{x}$ für $\mu(X)$ diskutiert, wird diese Unterscheidung unerlässlich. Weiter bereitet es dem Anfänger dann Schwierigkeiten, abstrakt genug zu denken, dass $\mu(X)$, also in Worten das „Populationsmittel" *existiert*, auch wenn man es mangels der Kenntnis aller Einzelwerte *nicht konkret kennt oder ausrechnen kann* (!).

2.2. Variablen mit vielen Werten oder sogar einem Kontinuum von Werten: Histogramme und die Idee der Dichte und Verteilungsfunktion. Wir sind noch nicht fertig mit der einfachsten Beschreibung von Verteilungen; was

sollte man mit einem Stabdiagramm beginnen, wenn eine Variable sehr viele Werte annimmt? Nehmen wir den Extremfall, dass man in einer riesigen Population jeden Wert nur einmal bekommt. Dann haben alle diese Werte dieselbe winzige relative Häufigkeit, und man erhält eine Art Rasenteppich als Stabdiagramm, nur stehen die Grashalme verschieden dicht verschiedenen Stellen. Dies wäre nicht nur sehr aufwendig zu zeichnen (per Computer ginge es natürlich wieder leicht), man hätte auch wenig davon. Aber das Problem führt zu einem tieferen Begriff, der für die gesamte theoretische Wahrscheinlichkeitstheorie und Statistik von größter Bedeutung ist, zum Begriff der Dichte. Die folgenden praktischen Ausführungen wollen auch dorthin geleiten, nicht nur weitere graphische Darstellungen von Verteilungen einführen.

Ein Beispiel: Wenn wir von einer großen Population von Jugendlichen die mittlere tägliche Fernsehdauer (unser X hier) erhoben haben, so werden wir etwa gruppieren: 0 bis unter 1 Stunde, 1 bis unter 2 Stunden, usw., bis 5 bis unter 6 Stunden. (Realistisch würde man weiter gehen müssen.) Dann werden wir die relativen Häufigkeiten zu diesen Klassen bilden und etwa zu folgender (unrealistischen!) Tabelle kommen:

Klasse	$0- < 1$	$1- < 2$	$2- < 3$	$3- < 4$	$4- < 5$	$5- < 6$
relative Häufigk.	0.15	0.4	0.2	0.14	0.1	0.01

Solche Daten kommen sehr oft vor. Stellen wir zunächst klar, dass man bei solcher Gruppierung eine Vergröberung der Verteilung von X vorgenommen hat: Z.B. weiß man in unserem Falle nicht, welcher Populationsanteil auf das Stundenintervall von 0 bis 1/2 entfällt - man wird vermuten, dass es sich um weniger als 0.075 handelt, aber die Zahl kennt man nicht. Das Vorgehen besteht nun einfach darin: Man betrachtet vereinfachend die Intervalle als völlig gleichmäßig besetzt und stellt die gruppierte Verteilung dann in einem Histogramm dar, auf folgende Weise:

Histogrammkonstruktion zu gruppierten Daten:Über jedem der Intervalle der Klasseneinteilung errichtet man einen Kasten, dessen Höhe der zugehörigen relativen Häufigkeit geteilt durch Kastenbreite entspricht und als Dichte zu interpretieren ist. Bei dieser Konstruktion entsprechen die relativen Häufigkeiten der Klassen den-Flächeninhalten der Kästen. (Den Kastenhöhen entsprechen sie zugleich nur dann, wenn die Kästen alle dieselbe Breite haben.)

In unserem Beispiel erhält man das folgende Bild.

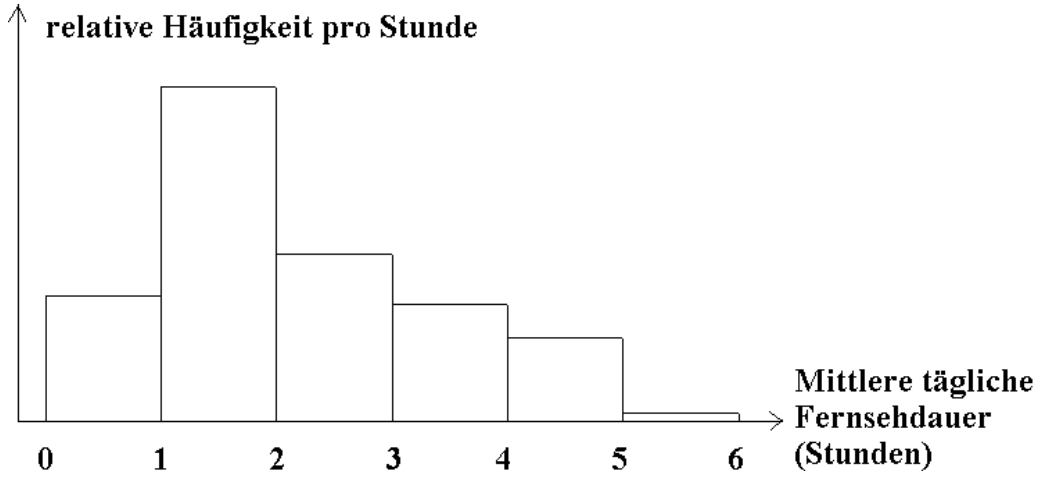


Abb. 2: Histogramm gruppierter Daten: Den relativen Häufigkeiten entsprechen die Flächeninhalte - im allgemeinen nicht die Höhen - das nur im Falle immer gleicher Klassenbreiten. Die Höhen repräsentieren Dichtewerte.

In *diesem* Falle könnte man wegen der gleichbleibenden Intervallbreite sogar auch die relativen Häufigkeiten der Tabelle ablesen. Anders sähe es aus, wenn man etwa in der Gruppierung die letzten beiden Intervalle zusammengefasst hätte zu: 4-6 Stunden: relative Häufigkeit 0.11. Ein Kasten über dem Intervall [4,6], dessen Höhe dem Wert 0.11 entspräche, würde den Anteil im Bild völlig verzerren, dem Intervall [4,6] mehr als das Doppelte von dem Gewicht geben, das es zuvor hatte. Die oben gegebene Vorschrift zur Konstruktion von Histogrammen korrigiert das automatisch: Division durch Klassenbreite ergibt eine Höhe, die dem Wert 0.055 entspricht, und dann erhält der Kasten über dem Intervall [4,6] denselben Flächeninhalt wie zuvor die beiden Kästen zu [4,5], [4,6] zusammen. Nun erkennt man auch, dass die Kastenhöhe allgemein abstrakter als Dichte zu interpretieren ist (hier: „relative Häufigkeit pro Jahr"). Der Wert 0.055 ist genau diese Dichte im Bereich [4,6], und er ist nicht als relative Häufigkeit deutbar.

Die Idee der Dichte führt auch theoretisch weiter: Wir können die Beschränkung auf endlich viele Werte und gar endliche Populationen fallenlassen und uns vorstellen, dass *jeder* Wert möglich ist; allerdings können wir nicht mehr von relativen Häufigkeiten reden, mit denen *einzelne* Werte vorkommen, sondern nur noch von relativen Häufigkeiten (oder besser noch: Wahrscheinlichkeiten) für *Intervalle (oder allgemeinere Bereiche)* von Werten. Für den einzelnen Wert haben wir stattdessen die Dichte. Weiter können wir verallgemeinern auf beliebige Dichtefunktionen - sie müssen nicht mehr stückweise konstant sein, sondern es können ihre Graphen beliebige Kurven sein. Nur sollte der Gesamtflächeninhalt unter der Kurve existieren (als Integral) und den Wert 1 haben. Dann lassen sich als Flächeninhalte (Integrale) alle interessierenden Wahrscheinlichkeiten für beliebige Bereiche ausrechnen. Ebenso lassen sich μ und σ über Integrale berechnen. Tatsächlich spielt diese Verallgemeinerung eine immense theoretische Rolle: Wir werden sehen, dass man mathematisch gewisse „Idealverteilungen" (z.B. die Normalverteilungen)

über Dichtefunktionen definieren und berechnen kann, die sich als außerordentlich wichtig auch für ganz praktische endliche Dinge erweisen, weil eben letztere in vielen Fällen systematisch voraussagbar sich einem solchen Idealtyp stark annähern. In diesem Zusammenhang wird eine weitere Darstellung wichtig, die *kumulierte.* Ihre besondere Tugend besteht darin, *universell* anwendbar zu sein, völlig gleichgültig, ob es sich um eine Variable mit diskreter oder kontinuierlicher Verteilung handelt, oder konkreter gesagt, ob die Verteilung durch ein Stabdiagramm oder eine Dichte gegeben ist. Dabei werden die Wahrscheinlichkeiten (oder relativen Häufigkeiten) aufsummiert, und zwar so: Man ordnet jeder reellen Zahl a die relative Häufigkeit (oder allgemeiner Wahrscheinlichkeit) für Werte $\underline{\leq a}$ zu. Das ist die entscheidende kleine Veränderung gegenüber der früher betrachteten Verteilung f_X. Bei ihr stand „=" statt „≤". Übrigens ist diese kumulierte Verteilung, die man „Verteilungsfunktion" nennt (das ist also ein terminus technicus!), auch manchmal in praktischer Hinsicht hilfreich: Histogramme eignen sich nicht bei Intervallen sehr verschiedener Breiten, da wegen der Flächendarstellung der Häufigkeiten Längen nicht verzerrt werden dürfen. Bei der Verteilungsfunktion stellen wieder die (an der vertikalen Achse abzulesenden) Funktionswerte die relativen Häufigkeiten dar, und die Intervallbreiten auf der Abszisse (horizontalen Achse) dürfen beliebig ungleichmäßig gedehnt oder gestaucht werden. (Denken Sie etwa an die Mitgliederzahlen in der Population der Vereine, da kommen sehr kleine und riesige vor. Das wird man nur mit der Verteilungsfunktion gut darstellen können.) Wir heben die Definition der Verteilungsfunktion zu einer Variable noch einmal in systematisch vollständiger Form hervor und geben anschließend wesentlich verschiedene Beispiele für Tabellen und Graphen von Verteilungsfunktionen. Außerdem gelangen wir bei der Erläuterung des zweiten der Beispiele zum überaus wichtigen Zusammenhang zwischen Dichte und Verteilungsfunktion.

DEFINITION 8. *Sei X eine Variable mit beliebiger Verteilung. Dann ist die Verteilungsfunktion von X folgende Abbildung:*

$$\begin{array}{rccl} F_X: & \mathbb{R} & \to & \mathbb{R} \\ & a & \mapsto & \begin{array}{l} relative\ Häufigkeit \\ der\ X-Werte \leq a. \end{array} \end{array}$$

(Allgemeiner steht „Wahrscheinlichkeit" für „relative Häufigkeit".)

Man sollte sich diese Definition merken! Eingangs wurde *untechnisch* formuliert, die Verteilung einer Variablen gebe die Information darüber, welche Werte wie häufig vorkommen. Anschließend haben wir Stabdiagramme und Histogramme kennengelernt, welche je auf ihre Weise diese Information geben und jeweils nur für bestimmte Klassen von Verteilungen brauchbar sind. Nunmehr verfügen wir mit der Verteilungsfunktion über eine *technische* Fassung des Begriffs der Verteilung, welche *für alle Fälle brauchbar* und überdies *stets praktisch geeignet graphisch darstellbar ist* (wie die folgenden Abbildungen zeigen werden). Das sollte man zu schätzen wissen.

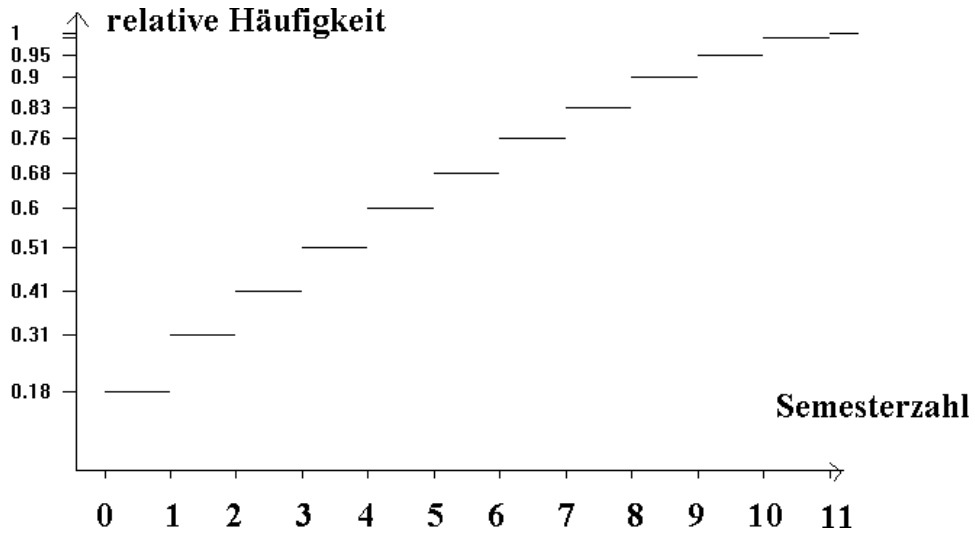


Abb. 3: Graphische Darstellung der Verteilungsfunktion zum Beispiel des ersten Stabdiagramms (Abb.1, S.11)

Zu beobachten sind die Treppen: Nur an den vorkommenden (hier ganzzahligen) Werten 0 bis 11 springt die Verteilungsfunktion, dazwischen kommt nichts hinzu, die Verteilungsfunktion bleibt also konstant. Vor dem ersten Wert (hier 0) hat sie konstant den Wert 0, nach dem letzten (11) konstant den Wert 1.

Das zweite Beispiel (Abb. 4 auf der nächsten Seite) zeigt die Verteilungsfunktion zum Histogramm von Abb. 2 (S. 13).

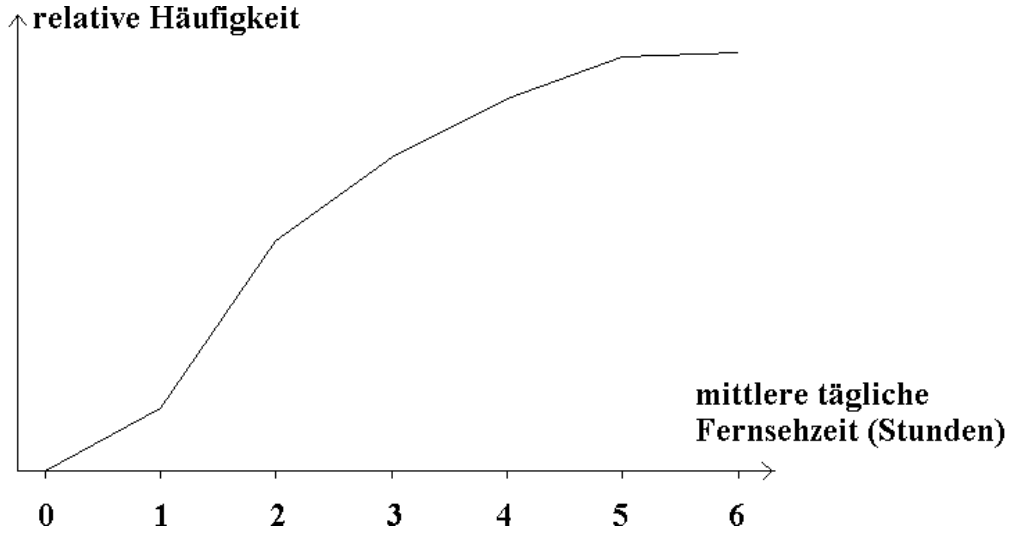


Abb. 4: Verteilungsfunktion zum früher gegebenen Histogramm (Abb. 2, S. 15)

Hier ist schon so etwas wie eine Kurve zu sehen, aber deutlich handelt es sich um einen stückweise geraden Polygonzug. Das liegt an dem linearen Anstieg innerhalb der Gruppierungsintervalle: Dort ist die Dichte konstant, also die Steigung der Verteilungsfunktion konstant. Dass die Dichte (das ist hier die stückweise konstante Funktion (graphisch: Treppe), welche die Histogrammkastenhöhen gibt) die Steigung der Verteilungsfunktion angibt, sieht man in unserem Beispiel folgendermaßen:

Nehmen wir das Intervall von 0 bis 1. Die Dichte ist konstant 0.15 (pro Stunde), da die Intervallbreite 1 beträgt. Wir haben $F_X(1) - F_X(0) = 0.15$, und $(F_X(1) - F_X(0))/1$ ist die *mittlere* Steigung von F_X auf [0,1] . Wenn man den $X - Wert$ von 0 bis 1 ansteigen lässt, kommt in gleichen Stücken stets der gleiche Flächeninhalt hinzu, da das Histogramm hier ein einziges Rechteck bildet. Folglich muss F_X auf [0,1] linear ansteigen, und 0.15, die mittlere Steigung von F_X auf [0,1], ist sogar die *konstante* Steigung der Verteilungsfunktion auf diesem Intervall. Das Histogramm gibt das Auf und Ab dieser stückweise konstanten Steigungen. Eine Dichte ist stets positiv (≥ 0), kann aber wachsen und fallen, eine Verteilungsfunktion hat stets Werte nur in [0,1] und ist stets monoton wachsend. Wir verdeutlichen noch einmal in allgemeinerer Form die Rolle der Klassenbreiten:

Sei eine Dichtefunktion auf dem Intervall [a,b] konstant, $a < b$, mit Wert c. (Man denke an [a,b] als ein Gruppierungsintervall (beliebiger Breite) bei einem Histogramm.

Dann ist c die Kastenhöhe:

$$c = \frac{relative\ Häufigkeit\ zu\ [a,b]}{b-a}$$

Aber der Zähler lässt sich auch ausdrücken als $F_X(b) - F_X(a)$, und es gilt:

$$Mittlere\ Steigung\ von\ F_X\ auf\ [a,b] = \frac{F_X(b) - F_X(a)}{b-a}$$

(Dies gilt allgemein für jede Funktion.)

Es kommt heraus, da diese mittlere Steigung zugleich die (konstante) Steigung von F_X auf [a,b] ist, an jeder Stelle des Intervalls:

Satz 1. *Bei einer Verteilung mit (stückweise stetiger) Dichtefunktion f gilt für jede Stelle: Wert der Dichte = Steigung der Verteilungsfunktion, also:*

$$F'_X(a) = f(a)\ für\ jeden\ Wert\ a.$$

Umgekehrt ist die Verteilungsfunktion zur Dichte f daher diejenige Stammfunktion von f, deren Werte im Bereich [0,1] liegen.

Dies Resultat gilt völlig allgemein, auch für Dichtefunktionen, deren Graphen glatte Kurven sind. Idee: Man stelle sich Histogramme mit immer kleineren Intervallbreiten, immer feinerer Gruppierung vor, die sich der Kurve annähern. Aus der mittleren Steigung wird dann die Ableitung, die lokale Steigung in jedem einzelnen Punkt. Dies ist nichts anderes als der Inhalt des Hauptsatzes der Differential- und Integralrechnung, ausgesprochen speziell für positive Funktionen, die einen gesamten Flächeninhalt 1 mit der x-Achse bilden, d.h. Dichtefunktionen.

Nun wollen wir dies in Aktion sehen bei einer mathematisch idealen Verteilung, die zugleich die für praktische Zwecke wichtigste ist, der Standard-Normalverteilung

mit $\mu = 0$, $\sigma = 1$. Die folgende Graphik zeigt Dichte und Verteilungsfunktion zusammen.

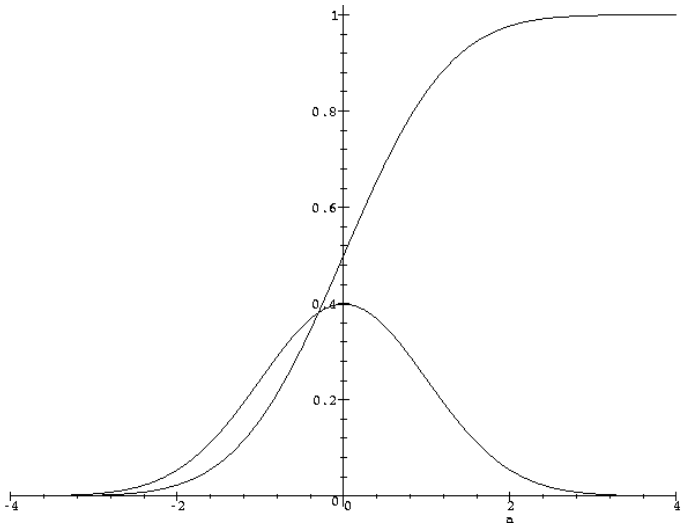


Abb. 5: Normalverteilung zu $\mu = 0, \sigma = 1$: Dichte (Glockenkurve) und Verteilungsfunktion (Sigmoid)

Abschließend schalten wir noch einmal zurück zu den Histogrammen und besprechen die Berechnung von μ und σ bei gruppierten Daten oder Histogrammen:

Bei gruppierten Daten kann man diese Verteilungsparameter nur näherungsweise angeben, und zwar zweckmäßig über die Idealisierung zu der Verteilung, die *genau* dem Histogramm entspricht, also mit stückweise konstanter Dichte. *Für diese Idealisierung* kann man leicht μ und σ *genau* angeben: Die Mittelwerte der Klassen entsprechen offenbar den Mittelpunkten der Klassenintervalle, und nun hat man (ganz gemäß Definition 3) einfach deren mit den relativen Häufigkeiten gewichtetes Mittel zu bilden, um μ zu erhalten. (Wenn man alle Einzelwerte, die in ein solches Intervall fallen, an der Intervallmitte ansiedeln würde, so erhielte man denselben Wert.) Wie steht es mit σ? Gewöhnlich macht man sich nicht einmal die Mühe, diesen Wert auch nur für die stückweise Gleichverteilung (also Histogrammverteilung) anzugeben, sondern benutzt einfach die Konstruktion, alle Einzelwerte in die jeweilige Klassenmitte zu verlegen. Dann ergibt die alte Formel: Varianz = mit den relativen Klassenhäufigkeiten gewichtetes Mittel der quadratischen Differenzen der Klassenmitten zu μ. Die Histogrammverteilung hat eine etwas größere Varianz als diesen Wert, da die quadratischen Entfernungen zur entlegeneren Intervallhälfte überwiegen. Aber für diesen Wert hätte man nicht nur komplizierter zu rechnen, er wäre auch in der Regel die schlechtere Näherung der Varianz der ursprünglichen Verteilung; denn meist hat man schiefe Glockenformen - so auch in unserem Beispiel, und dann gibt die Histogrammverteilung einen übertrieben großen Wert. Fassen wir zusammen:

Näherungsweise Berechnung von μ und σ bei gruppierten Verteilungsdaten:

$$\mu \approx \sum_{Klassen} Klassenmitte \cdot relative\ Häufigkeit\ zur\ Klasse$$

$$\sigma \approx \sqrt{\sum_{Klassen} (Klassenmitte - \mu)^2 \cdot relative\ Häufigkeit\ zur\ Klasse}$$

(Für μ ist hier natürlich der Näherungswert der 1. *Zeile einzusetzen.)*

2.3. Der Median als Alternative zum Mittelwert. In manchen Situationen erweist sich der Mittelwert als ungünstig für den Zweck der Beschreibung eines „Zentrums" einer Verteilung. Meist geht es dabei um das Problem weniger „Ausreißer", die durchaus (sogar einzeln!) in der Lage sind, das arithmetische Mittel gewaltig zu verschieben. Ein Beispiel: Nehmen wir eine Einkommensverteilung in einer Population von eintausend Menschen, deren Monatseinkommen sich mit sehr geringer Streuung um 4000 DM bewegen, sagen wir vereinfachend: Alle diese haben genau dies Einkommen. Nun fügen wir dieser Population einen einzigen Menschen hinzu mit Monatseinkommen 10 Millionen DM. Das arithmetische Mittel der Einkommensgröße wird dann von 4000 DM auf sage und schreibe knapp 14 000 DM angehoben! Natürlich sagt dieser Wert korrekt, was das arithmetische Mittel immer sagt, z.B., wie viel jeder Einzelne bekommen könnte, wenn gleichmäßig aufgeteilt würde. Aber man wird bemängeln, dass dieser Wert gerade im Niemandsland liegt, sogar eine überhaupt in der Bevölkerung nicht vertretene Größenordnung hat. Will man auf so etwas wie einen „typischen" Wert hinaus, dann ist bei sehr schiefen Verteilungen das arithmetische Mittel unbrauchbar. Gerade in solchen Situationen erweist sich der (oder besser „ein") Medianwert als günstig: Das ist ein (im allgemeinen nicht eindeutig bestimmter, aber doch hinreichend genau anzugebender) Wert auf der Größenskala mit der Eigenschaft: Die Hälfte der Population liegt unter diesem Wert, die andere darüber. In unserem Beispiel ist klar: Der Medianwert ist 4000 DM. Er gibt also das Typische für die Population - der „Ausreißer" beeinflusst diesen Wert überhaupt nicht. Generell gilt, dass einige Exoten kaum eine Veränderung im Median bewirken. Stets ist ein Median brauchbar als „typischer" Wert, *wenn* die Verteilung so etwas wie eine (eventuell schiefe) Glocke ist - und diese Form ist die bei weitem häufigste. Bei U-förmiger Verteilung ergibt natürlich auch ein Median nichts „Typisches".

Hier ein Beispiel zur Nichteindeutigkeit eines Medians: Liegen die Einzelwerte 0, 0, 1, 1, 1, 2, 2, 2, 3, 3 vor, so ist *jeder* Wert zwischen 1 und 2 als Median brauchbar. Man sollte dann auch nur sagen, der Median liege zwischen 1 und 2. Krampfhafte Versuche (sie existieren), mit irgendwelchen Regeln eine Eindeutigkeit zu erzwingen, sind völlig willkürlich und lohnen nicht. Bei einer ungeraden Zahl von Werten nimmt man den im Sinne der Anzahl „mittleren", also den sechsten bei 11 Werten, usw., aber auch hier könnte man sich getrost mit der Angabe eines Bereichs begnügen. In natürlicher Weise eindeutig definiert ist der Median bei einer Verteilung, die durch eine Dichte gegeben ist: Dann sind die relativen Häufigkeiten, die auf Bereiche von Größenwerten entfallen, durch die entsprechenden Flächeninhalte unter der Dichtekurve gegeben, und man findet genau eine Stelle, an welcher der gesamte Flächeninhalt unter der Dichtekurve halbiert wird. Dort hat die Verteilungsfunktion genau den Wert 0.5, der Median ist also definitionsgemäß exakt

diese Stelle. Für Verteilungen mit Dichtefunktion können wir daher den Median als (eindeutig bestimmte) Lösung der Gleichung $F_X(a) = 0.5$ definieren.

Es sei noch erwähnt, dass man bei „Mittelungen", die durch so etwas wie einen demokratischen Entscheidungsprozess entstehen, aus den erwähnten Gründen eher auf einen Median als auf den Mittelwert stoßen wird. Lassen wir zum Beispiel die Leute auf Zettel schreiben, wie viel Geld die Gesellschaft ausgeben sollte für einen bestimmten Zweck, dann ist die Sache nach unten begrenzt durch den Wert 0, nach oben aber offen. Exoten oder Witzbolde könnten erdenklich große Zahlen hinschreiben. Das arithmetische Mittel geriete in absurde Höhen. Ein Median wäre hier viel vernünftiger, auch „demokratischer", und tatsächlich wird so ein Wert als „gerechter" empfunden und setzt sich eher gesellschaftlich durch.

KAPITEL 3

Elementare Wahrscheinlichkeitsrechnung

1. Der Begriff der Wahrscheinlichkeit

1.1. Relative Häufigkeit und Wahrscheinlichkeit. Wir wollen den abstrakteren Begriff der Wahrscheinlichkeit vom vertrauteren der relativen Häufigkeit her entwickeln. Zunächst einmal brauchen wir einen generellen Rahmen, in dem eine Rede von „Wahrscheinlichkeit" erst sinnvoll werden kann: Stets muss ein (im Prinzip beliebig) wiederholbares Zufallsexperiment vorliegen. Im interessanten Fall sind bei der Durchführung eines solchen Experiments verschiedene möglich, wir begnügen uns zunächst mit endlich vielen Ausgängen. Zwei klassische Beispiele, an denen man die Sache gut verstehen kann:

Würfeln mit einem gewöhnlichen Würfel: Die möglichen Ausgänge sind die Augenzahlen 1,2,3,4,5,6.

Zufälliges Ziehen (also nach „blindem Mischen") einer Kugel aus einer Urne: Die möglichen Ausgänge sind die einzelnen Kugeln (die wir etwa numeriert haben mögen).

Beobachten Sie: Das zweite Beispiel enthält das erste strukturell als Spezialfall! (Man nehme eine Urne mit 6 Kugeln, und bei Wiederholen des Experiments ist die gezogene Kugel stets zurückzulegen und neu zu mischen.)

Bleiben wir beim zweiten Beispiel, das schon eine gewisse Allgemeinheit besitzt. Betrachten wir nun ein Merkmal in der Menge der Kugeln, allgemeiner gesprochen: bei den möglichen Ausgängen, zum Beispiel seien von den insgesamt n Kugeln k rot, der Rest weiß. Das Merkmal „rot" ist also in der Urne mit einer relativen Häufigkeit von k/n vertreten. Wenn man nun fragt, wie wahrscheinlich es nun sei, beim zufä lligen Ziehen einer Kugel eine rote zu ziehen, so wird fast jeder antworten, diese Wahrscheinlichkeit sei gerade diese relative Häufigkeit, k/n. Diese Antwort ist auch völlig korrekt. Aber nicht jedem ist dabei klar, was diese Aussage bedeutet. Was ist überhaupt gemeint mit: „Das Ereignis ... (das bei Durchführung eines bestimmten Experiments eintreten kann) hat die Wahrscheinlichkeit ... (eine Zahl aus [0,1])"? Jedenfalls meint diese Aussage nicht unmittelbar eine relative Häufigkeit, diese Übereinstimmung ist allenfalls ein *Resultat* mathematischer Überlegung. Sondern sie zielt auf *künftig (bei langer Wiederholung des Experiments) zu Erwartendes ab.* In unserm Beispiel besagt sie: Bei oftmaliger Wiederholung des Ziehens wird die relative Häufigkeit *der Fälle, in denen eine rote Kugel gezogen wurde (*unter allen Ziehungen), *nahe bei k/n liegen.* Dies ist eine Voraussage, und wir können empirisch prüfen und sogar mathematisch berechnen, wie gut sie ist. Wir halten dies als naive, informelle (aber überaus nützliche) Definition des Wahrscheinlichkeitsbegriffs fest:

Gegeben sei ein Zufallsexperiment, dazu ein Ereignis A, das bei diesem Experiment eintreten kann (oder auch vielleicht nicht). Dann gilt folgende Grundtatsache: Bei oftmaliger Durchführung des Experiments setzt sich eine relative Häufigkeit durch, mit der A beobachtet wird. Diese ideale relative Häufigkeit, um welche die beobachteten relativen Eintrittshäufgkeiten mit tendenziell immer kleineren Abständen pendeln, heißt Wahrscheinlichkeit von A, symbolisch: P(A).

Diese Definition ist auch in solchen Fällen brauchbar, in denen nicht bereits eine relative Häufigkeit zugrundeliegt, wie in unserem Urnenbeispiel.

Die beschriebene Grundtatsache wollen wir einmal am Werke sehen in einem Beispiel (vgl. Abb. 6, nächste Seite): Hier wurde (Computer-simuliert) 1000 mal gewürfelt und für jede Durchführungszahl i ($1 \leq i \leq 1000$ - waagerechte Achse) aufgetragen (senkrechte Achse), mit welcher relativen Häufigkeit eine gerade Augenzahl beobachtet wurde bis einschließlich zu diesem Versuch. Man sieht, dass die Wahrscheinlichkeit 1/2 so gut wie nie genau erreicht wird, die Abweichungen auch stellenweise wieder größer werden können, aber doch auf lange Sicht stets kleiner (tatsächlich sogar beliebig klein) werden. Dagegen sind bei wenigen Versuchen durchaus große Abweichungen zu möglich (und auch zu erwarten). Insbesondere kann die beobachtete relative Häufigkeit nach einem einzigen Versuch nur den Wert 0 oder aber 1 haben.

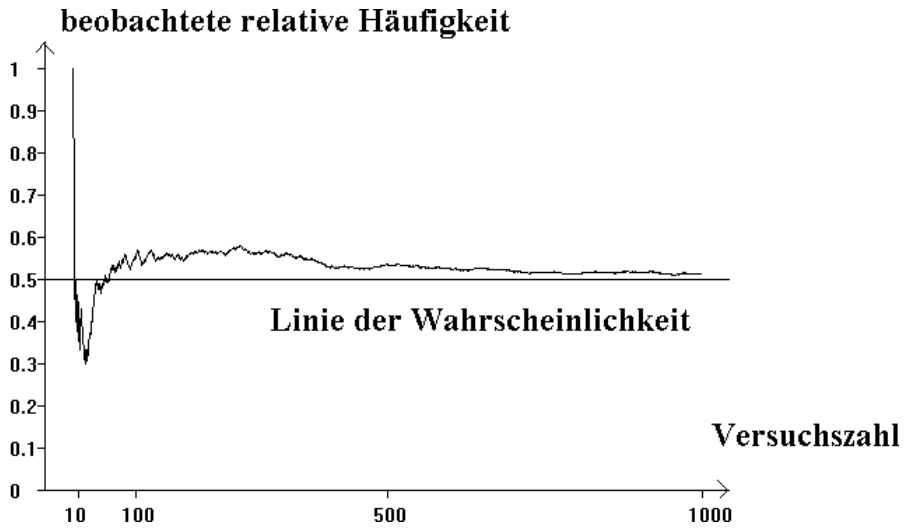


Abb. 6: Näherung empirischer relativer Häufigkeiten an eine ideale Wahrscheinlichkeit (hier mit Wert 1/2)

Die nächste Graphik zeigt noch etwas mehr: Dort wurden nicht nur mehrere Versuchsreihen aufgenommen, die immer das gleiche Bild zeigen - nur starten einige bei dem ganz falschen Wert 0, einige bei 1. Außerdem wurden Kurven mit eingezeichnet, die angeben, wie nahe bei der Wahrscheinlichkeit eine beobachtete relative Häufigkeit bei der jeweiligen Zahl der Versuche mit einer Wahrscheinlichkeit von 0.95 (oder 95%) liegen sollte. (Beobachten Sie, dass in einigen wenigen Fällen diese

Kurven überschritten wurden.) Zur Ermittlung dieser Kurven vgl. den späteren Abschnitt zur Anwendung der Normalverteilungen.

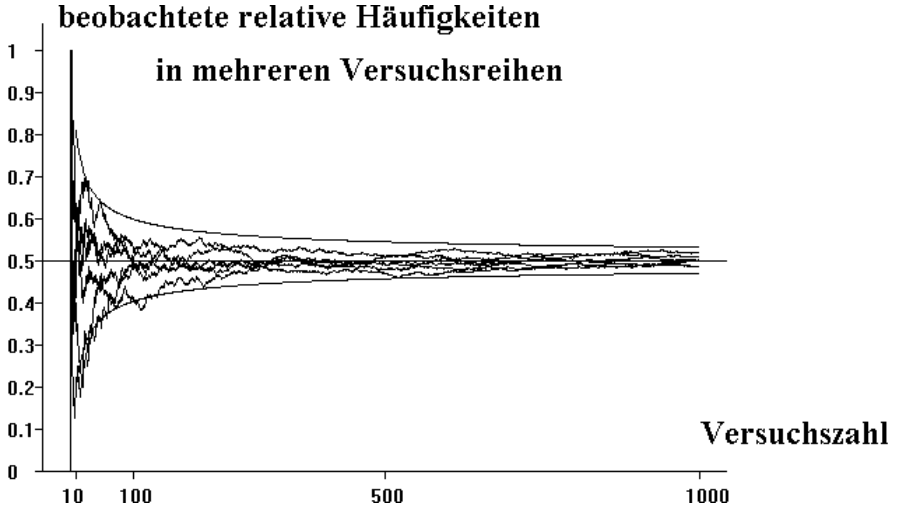


Abb. 7: Wahrscheinlichkeit und beobachtete relative Häufigkeiten, mit Begrenzungskurven, die mit 95% Wahrscheinlichkeit eingehalten werden.

Was wir hier beobachtet haben, ist die Wirksamkeit des „Gesetzes der großen Zahl", grob formuliert: Die empirischen relativen Häufigkeiten nähern sich (immer besser und sicherer) den Wahrscheinlichkeiten.

1.2. Der abstrakte Wahrscheinlichkeitsbegriff. Mathematisch geht man nunmehr so vor: Die Wahrscheinlichkeiten sollen so etwas wie „ideale relative Häufigkeiten" sein, daher muss man von ihnen verlangen, dass sie sich rechnerisch auch wie relative Häufigkeiten benehmen. Unmittelbar ergibt die in der folgenden Definition niedergelegten axiomatischen Anforderungen an den Wahrscheinlichkeitsbegriff, wenn man eine kleine geistige Vorbereitung getroffen hat: *Jedem Ereignis* soll *seine* Wahrscheinlichkeit (eine Zahl aus [0,1]) zugeordnet werden. Aber wie kann man *alle denkbaren Ereignisse* fassen? Es genügt, dies für ein beliebiges fest gegebenes Zufallsexperiment zu schaffen. Dabei hat man eine bestimmte Menge Ω möglicher Ausgänge. Nunmehr kann man jedes Ereignis so formulieren:

„Der zufällig herauskommende Ausgang ω hat die Eigenschaft E". Dies lässt sich umformulieren zu:

„Der zufällig herauskommende Ausgang ω ist Element der Menge $\{\omega \in \Omega \mid \omega$ hat die Eigenschaft E$\}$."

Einziger Bedeutungsträger in diesem Satz ist die Menge $\{\omega \in \Omega \mid \omega$ hat E$\}$. Somit kann jedes Ereignis als eine Teilmenge von Ω codiert werden, und mit der Menge aller teilmengen von Ω, $\mathcal{P}(\Omega)$, haben wir jedenfalls alle denkbaren Ereignisse erfasst.

Man prüfe das Verständnis dieser Bemerkungen, indem man folgende Ereignisse beim Würfeln (mit nur einem gewöhnlichen Würfel) verbal formuliert: $\{2, 4, 6\}, \{1\}$ und das Ereignis „Es kommt eine Zahl unter Drei heraus" als Menge umschreibt.

DEFINITION 9 (Begriff der Wahrscheinlichkeitsfunktion). *Sei Ω endlich. (Man denke an Ω als Menge aller möglichen Ausgänge eines Zufallsexperiments). Dann heißt eine Funktion P Wahrscheinlichkeitsfunktion über Ω, wenn sie folgende Eigenschaften besitzt:*

(1.1)
$$\boxed{\begin{array}{l} P : \mathcal{P}(\Omega) \to [0,1] \textit{(Lies P(A): „Wahrscheinlichkeit von A".)} \\ P(\Omega) = 1 \\ P(A \cup B) = P(A) + P(B), \textit{ wenn } A \cap B = \varnothing, \textit{ für alle } A, B \in \mathcal{P}(\Omega) \end{array}}$$

In unserem beschränkten Fall von endlichem Ω sind alle Teilmengen von Ω Ereignisse. Der Begriff ist jedoch sogar auf überabzählbare Ω verallgemeinerbar, allerdings kann P dann *nicht mehr auf der gesamten Potenzmenge von Ω definiert werden.* Insbesondere benötigt man die Verallgemeinerung der Summenformel auf abzählbar unendlich viele Mengen, die vereinigt werden, zu: $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$, wenn die Mengen A_i paarweise leeren Durchschnitt haben. (Auf der rechten Seite steht eine unendliche Reihe.)

Unmittelbar gewinnt man aus diesen Axiomen die Folgerungen:

(1.2)

$$\boxed{\begin{array}{ll} (i) & P(\overline{A}) = 1 - P(A) \\ (ii) & P(A \cap B) = P(A) + P(B) - P(A \cup B) \textit{ (stets!)} \\ (iii) & P(A) = \sum_{\omega \in \Omega} P(\{\omega\}) \textit{ (Für endliche, aber auch abzählbare } \Omega.) \\ (iv) & \begin{array}{l} \textit{Spezialfall von } (iii) : \textit{Sind alle Ausgänge gleich wahrscheinlich, so gilt:} \\ P(A) = \textit{relative Häufigkeit von } A \textit{ in } \Omega = \frac{\textit{Anzahl der „günstigen" Fälle}}{\textit{Anzahl der möglichen Fälle}} \end{array} \end{array}}$$

Insbesondere *(iii)* wirft ein Licht darauf, was am Wahrscheinlichkeitsbegriff unbefriedigend geklungen haben sollte: Er sagt nämlich nicht, wie man Wahrscheinlichkeiten ausrechnen kann, sondern nur, wie Wahrscheinlichkeiten miteinander zusammenhängen. Damit sind wir bei der geistigen Hauptfigur der Mathematik: Über die Zusammenhänge der Dinge miteinander das Wesentliche über diese Dinge herauszufinden. Aus den Zusammenhängen, die in den Axiomen stehen, folgert man mit (iii), dass man zum Ausrechnen von Wahrscheinlichkeiten lediglich die Wahrscheinlichkeiten der einzelnen Ausgänge selbst, genauer der Elementarereignisse $\{\omega\}$, zu kennen braucht. Nun sollte man verstehen, dass der *allgemeine* Wahrscheinlichkeitsbegriff nicht mehr sagen kann, da er z.B. für einen symmetrischen Würfel, aber auch noch für den schiefsten Würfel gelten sollte. Allenfalls bei zusätzlicher Information über den Einzelfall kann man erwarten, die Wahrscheinlichkeiten der Elementarereignisse spezifizieren zu können. Ein besonders einfacher Fall: Man hat endlich viele, sagen wir Anzahl n, die alle gleich wahrscheinlich sind. (Symmetrischer Würfel, „zufälliges Ziehen" usw.) Dann ist klar, dass alle die Wahrscheinlichkeit $1/n$ haben, und man gelangt zu (iv). Für kompliziertere Fälle liefert das Gesetz der großen Zahl einen empirischen Zugang: Man kann die relative Häufigkeit beobachten, mit der ein Ereignis eintritt, und damit die Wahrscheinlichkeit recht sicher und genau annähern. Die zwei grundlegenden theoretischen Methoden zur Gewinnung von Wahrscheinlichkeiten werden an den elementarsten und praktisch wichtigsten Beispielen in den Abschnitten 2.1 bis 2.4 dieses Kapitels gezeigt.

1.3. Von den Variablen zu den Zufallsvariablen (oder „zufälligen Größen"). Die meisten Wahrscheinlichkeiten, nach denen zu fragen interessant ist, beziehen sich auf Variaben, deren Werte man bei Zufallsexperimenten beobachten kann: Wie wahrscheinlich ist es, unter 1000 zufällig gezogenen Leuten wenigstens 300 mit einem gewissen Merkmal zu finden? Andererseits kann man auch Ereignisse, die nur eintreten oder nicht eintreten können, mit Werten 1,0 einer Größe beschreiben, so dass dieser Gesichtspunkt allgemeiner ist.

Wir stellen die begriffliche Verbindung her zwischen den eingangs betrachteten Variablen mit ihren Verteilungen (dort ging es um relative Häufigkeiten von Werten) zu den Zufallsvariablen, deren Werte man jeweils bei einem Zufallsexperiment beobachten kann, und ihren Verteilungen (dabei geht es um Wahrscheinlichkeiten von Werten). Dazu folgende

Definition 10.

Eine Zufallsvariable X ist eine Abbildung $\Omega \to \mathbb{R}$, wobei über Ω eine Wahrscheinlichkeitsfunktion P gegeben ist.
Wenn Ω höchstens abzählbar ist, so ist die Verteilung von X ist folgende Abbildung:

$$\begin{array}{rccl} f_X: & \mathbb{R} & \to & \mathbb{R} \\ & a & \mapsto & P(X=a) := P(\{\omega \in \Omega \mid X(\omega) = a\}) \end{array}$$

Im ganz allgemeinen Fall hat man die stets brauchbare Verteilungsfunktion von X:

$$\begin{array}{rccl} F_X: & \mathbb{R} & \to & \mathbb{R} \\ & a & \mapsto & P(X \leq a) := P(\{\omega \in \Omega \mid X(\omega) \leq a\}) \end{array}$$

Man beachte: In den bisherigen Definitionen der Verteilung (nichtkumulativ bzw. kumulativ) ist jeweils einfach „relative Häufigkeit" durch „Wahrscheinlichkeit" zu ersetzen. Es ist sehr nützlich, sich zu merken, was Ausdrücke wie $P(X \leq a)$, $P(X > a)$, $P(X = a)$ bedeuten.

Zur Einübung in die Notation ein einfaches Beispiel: $X = Augensumme\ beim\ Würfeln\ mit\ zwei\ Würfeln.$ Wir haben hier:

$\Omega = \{(a, b) \in \mathbb{N} \times \mathbb{N} \mid 1 \leq a, b \leq 6\}$, also Menge der Paare natürlicher Zahlen bis 6.

$P(X = 2) = 1/36$, $P(X = 3) = 1/18$, $P(X = 4) = 1/12$, $P(X = 5) = 1/9$, $P(X = 6) = 5/36$, $P(X = 7) = 1/6$, $P(X = 8) = 5/36$, $P(X = 9) = 1/9$, $P(X = 10) = 1/12$, $P(X = 11) = 1/18$, $P(X = 12) = 1/36$.

2. Drei wichtige Verteilungstypen

Im ersten Kapitel haben wir den Begriff der Verteilung eingeführt, der sich vollkommen von relativen Häufigkeiten zu Wahrscheinlichkeiten verallgemeinert. Insbesondere verstehen wir nunmehr stets allgemein unter der Verteilungsfunktion F_X einer Zufallsvariablen X die Funktion von $\mathbb{R}$ nach $[0, 1]$ mit der definierenden Gleichung $F_X(a) = P(X \leq a)$, in Worten: $F_X(a)$ ist die Wahrscheinlichkeit dafür, dass ein $X-$ Wert $\leq a$ herauskommt bei Durchführung des zu X gehörigen Zufallsexperiments. Nun fragt sich, wie man an diese Wahrscheinlichkeiten herankommt. Hat man für das Zufallsexperiment ein genaues (oder näherungsweise zutreffendes mathematisches Modell, so kann man die zugehörigen allgemeinen mathematischen Resultate nutzen, in denen die Wahrscheinlichkeiten gerade zugänglich gemacht

werden (über Formeln, eventuell noch Tabellen, heutzutage vielfach in Computerprogrammen direkt nutzbar). (Andere Möglichkeiten bestehen im zusätzlichen oder auch alleinigen Einsatz empirischer Stichproben.) Wir stellen nunmehr die drei elementarsten mathematischen Modelle vor, die besonders häufig und auch vielfältig zu nutzen sind. Davon sind zwei Verteilungstypen diskret, in diesen Fällen mit nur endlich vielen möglichen Werten der Zufallsvariablen, welche so verteilt sind, dazu gesellt sich der wichtigste stetige (kontinuierliche) Verteilungstyp: Normalverteilung. So verteilte Variablen können **alle** reellen Zahlen als Werte haben. Das kommt so *genau* empirisch also nicht vor, dennoch ist dieser Typ der gerade für *praktische* Anwendungen wichtigste überhaupt, weil viele empirische Variablen diesem mathematischen Idealtyp *sehr nahe* kommen.

2.1. Binomialverteilungen. Eine Verteilungsform ist in der Regel auf einen bestimmten Situationstyp zugeschnitten, und das System (die Situation) ist zu konkretisieren durch einen oder mehrere Parameter. So auch in unserem Beispiel (und bei allen weiteren interessanten Verteilungstypen). Stellen wir folgendes Problem: Bei einem (beliebigen) Zufallsexperiment trete ein gewisses Ereignis mit Wahrscheinlichkeit p ein. Man führt das Experiment n mal durch. Mit welcher Wahrscheinlichkeit tritt dabei das Ereignis genau k mal ein? Diese Wahrscheinlichkeit hängt sicher von p, n, k ab. Aber es ist nützlich, p und n als äußere Parameter zu betrachten, k dagegen als unabhängige Variable; denn k ist ein beliebiger ganzzahliger Wert von 0 bis n, er variiert noch, wenn n und p bereits festgelegt sind. Zum Beispiel: Welche Wahrscheinlichkeit hat man für k Sechsen bei 100 Würfen? Hier sind p und n fixiert, aber k möchte man durchaus noch variieren. Die vollständige Antwort kann mit einer Formel gegeben werden. Aber solch eine Formel kann man immer nachschlagen - wichtiger ist das Verständnis der *Situationen*, in denen die Formel gültig ist. (Weiter unterscheide man dann noch, wann sie praktisch oder nur sehr mühsam bis unmöglich zu verwenden ist.)

DEFINITION 11.

Eine Variable X heißt p-Bernoulli-verteilt, wenn sie nur die Werte 1,0 annimmt, und zwar den Wert 1 mit Wahrscheinlichkeit p.
Eine Variable heißt $(n,p)-$binomialverteilt, wenn sie Summe von n unabhängigen p-Bernoulli-Variablen ist.
Wesentlich anschaulichere Beschreibung der binomialverteilten Variablen: Die (n,p)-binomialverteilten Variablen sind gerade die mit folgender Struktur: X = Trefferanzahl bei n unabhängigen Versuchen, wobei die Trefferwahrscheinlichkeit in jedem Versuch p beträgt.

Hier ist die Formel für die (nichtkumulierte) Verteilung:

SATZ 2. *Sei X (n,p)-binomialverteilt. Dann gilt:*

$$f_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ \text{für } k = 0, 1, ...n \tag{2.1}$$

Dabei ist das Symbol $\binom{n}{k}$ zu lesen : „n über k”,

$$\text{Berechnung} : \binom{n}{k} = \frac{n!}{k!(n-k)!}, \ \text{wobei } n! = 1 \cdot ... \cdot n \ (n \geq 1), \ 0! = 1,$$

lies „n Fakultät".

Beispiel: Mit welcher Wahrscheinlichkeit hat man bei 12 Würfen mit einem Würfel genau drei Sechsen? Die Antwort ist: $\binom{12}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^9 = 0.1974$.

Der Beweis der Formel ist nicht besonders schwierig. Man benötigt drei Überlegungen:

Bei unabhängigen Ereignissen ergibt sich die Wahrscheinlichkeit der „und"-Verbindung durch Multiplikation der Einzelwahrscheinlichkeiten. Beispiel: Die Wahrscheinlichkeit dafür, zwei mal hintereinander eine Sechs zu würfeln, ist $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. Man beachte, wie sich die möglichen Fälle multiplizieren zu $6 \cdot 6 = 36$ und wie diese wieder gleich wahrscheinlich sind. Das ist das Wesentliche bei der Unabhängigkeit: Tritt das eine Ereignis ein, so ändert sich damit nicht die Wahrscheinlichkeit für das andere. Wenn man von einem Menschen das Geschlecht kennt, so ändern sich damit die Wahrscheinlichkeiten für gewisse Merkmale (Körperlängen, bevorzugter Typ von Literatur, auch von Sportarten), für andere nicht (IQ z.B.). Folglich ist $p^k(1-p)^{n-k}$ die Wahrscheinlichkeit dafür, die k „Treffer" und $n-k$ „Nieten" in einer ganz bestimmten Reihenfolge (gleichgültig welcher) zu erhalten.

Zweite Überlegung: Wie viele Anordnungen von k „Treffern" auf n Plätzen gibt es? (Gerade so oft hat man den Wert $p^k(1-p)^{n-k}$ zu nehmen, da sich diese Ereignisse, dass sich die Treffer in ganz bestimmter Anordnung ereignen, gegenseitig ausschließen und daher für die „oder"-Verbindung die Wahrscheinlichkeiten zu addieren sind (Axiom!). Wählen wir die k Plätze, so haben wir n Möglichkeiten für den ersten, dann unabhängig $n-1$ für den zweiten, zusammen also $n(n-1)$ Möglichkeiten, weiter geht es mit $n-2$ für den dritten usw., bis $n-k+1$ für den k-ten. Das macht $n(n-1)\cdot\ldots\cdot(n-k+1)$. Das ist jedoch noch nicht die gesuchte Zahl; wir müssen die Reihenfolge noch loswerden - wir wollten k Plätze, keinen ersten, zweiten, ..., k-ten, plastischer: k gleichberechtigte Vorsitzende aus n Leuten, nicht einen ersten, zweiten,..., k-ten Vorsitzenden.

Dritte Überlegung: Es bleibt noch zu teilen durch die Anzahl der Reihenfolgen, in die man k Objekte bringen kann. Das aber ist $k! = k(k-1)\cdot\ldots\cdot 1$, also das Produkt aller Zahlen von 1 bis k ($k \geq 1$). Nehmen wir nämlich an, wir wüssten schon die Anzahl der Möglichkeiten, m Dinge anzuordnen. Überlegen wir dann, wie viele es für $m+1$ sind: Fixieren wir eines der Objekte - wir können es als erstes ... bis letztes nehmen. Das sind $m+1$ Möglichkeiten. Unabhängig können wir die restlichen m Objekte *untereinander* auf so viele Weisen anordnen, wie wir schon wissen. Somit haben wir folgendes Resultat:

Anzahl der Möglichkeiten, $m+1$ Dinge anzuordnen = $(m+1)\cdot$ Anzahl der Möglichkeiten, m Dinge anzuordnen. Für $m=1$ kommt eine Anordnung, für $m=2$ kommen also $2 \cdot 1 = 2$ mögliche Anordnungen, für $m=3$ werden es also $3 \cdot 2$. Allgemein für k also $k!$ mögliche Anordnungen. (Diese Schlussweise, von einer ganzen Anfangszahl mit einem Schema von jeder beliebigen zur nachfolgenden überzugehen und damit das Resultat für alle ganzen Zahlen von der Anfangszahl ab zu haben, nennt man *vollständige Induktion.)* Auch der Grenzfall $0! = 1$ stimmt, es gibt bei rechter Deutung *eine* Möglichkeit, die leere Menge von Objekten anzuordnen: Keiner ist der Erste/Letzte, also im Nichtstun besteht das Anordnen. Zusammen haben wir folgenden

SATZ 3. $\binom{n}{k}$ = *Anzahl der Möglichkeiten, k Dinge aus n Dingen auszuwählen. Dabei ist* $0 \leq k \leq n$, k *und* n *sind als ganze Zahlen vorauszusetzen.*

Dieser Satz gilt auch für $k = 0$. Denn es gibt genau eine Möglichkeit, 0 Dinge aus k Dingen auszuwählen - man lässt alle ungewählt. Man könnte auch sagen, dass man alle n als Liegengelassene auswählt. Und offensichtlich gibt es genau eine Möglichkeit, n Dinge aus n Dingen auszuwählen. Damit sollte auch klar sein, dass stets $\binom{n}{k} = \binom{n}{n-k}$ gilt.

(Wir benutzten tatsächlich Satz 3, um Satz 2 zu zeigen, aber Satz 3 ist auch in weiteren Situationen nützlich.)

Man wird leicht bemerken, dass für große Zahlen n solche Zahlen wie $\binom{n}{k}$ sehr groß werden können, z.B. auf keinen Taschenrechner mehr passen. Erst recht wird dann eine Frage nach der Wahrscheinlichkeit von *höchstens* k Treffern unbequem: Alle Werte $P(X = i), i = 0...n$ wären dann zu addieren. Gerade in diesen Fällen hilft die Näherung durch eine Normalverteilung. Aber auch für viel praktischerer Fragen sind die Normalverteilungen wichtig: Wie gut ist es, wenn man den unbekannten Mittelwert einer Größe durch ein arithmetisches Mittel nähert, das man in einer Stichprobe fand? So besprechen wir im übernächsten die Normalverteilungen selbst, anschließend die benötigte Technik zum Umgang mit Mittelwert und Streuung sowie praktische Anwendungen.

2.2. Hypergeometrische Verteilungen. Eng verwandt mit den Binomialverteilungen sind die sogenannten hypergeometrischen. Ähnlichkeit und Unterschied soll an entsprechenden Urnenmodellen veranschaulicht werden: Zieht man aus eine Urne n Kugeln, legt jedoch dabei die gezogene Kugel stets zurück, um erneut die Kugeln in der Urne zu mischen, so ist die Variable „Trefferzahl" (d.h. Anzahl der gezogenen Kugeln mit einer bestimmten Eigenschaft) binomialverteilt, mit den Parametern $p =$ Anteil der Trefferkugeln in der Urne, $n =$ Anzahl der gezogenen Kugeln. In diesem Falle kommt es offenbar nicht darauf an, wie viele Kugeln absolut in der Urne sind. Zieht man dagegen n Kugeln auf einmal heraus (oder einzeln, aber eben ohne Zurücklegen) und betrachtet wieder die Variable „Trefferzahl", so bemerkt man, dass die absolute Anzahl N der Kugeln in der Urne ein bedeutsamer Parameter ist. Wir leiten die Wahrscheinlichkeitsfunktion her, welche die Wahrscheinlichkeiten für die verschiedenen möglichen Trefferzahlen angibt: Es werden n Kugeln aus N gezogen, natürlich $n \leq N$. Das macht $\binom{N}{n}$ mögliche Fälle. Die günstigen Fälle für k Treffer haben folgende Anzahl: $\binom{K}{k} \cdot \binom{N-K}{n-k}$, wobei K die absolute Anzahl der „Trefferkugeln" in der Urne ist. Denn aus den Trefferkugeln sind genau k auszuwählen, aus den übrigen „Nietenkugeln" unabhängig $n-k$. Nach der Formel „Wahrscheinlichkeit = Anzahl der günstigen geteilt durch Anzahl der möglichen Fälle" (anwendbar bei lauter gleichwahrscheinlichen Fällen!) haben wir also:

Satz 4. *Sei X die Variable „Trefferzahl" beim Ziehen (ohne Zurücklegen) von n Kugeln aus einer Urne mit N Kugeln, davon K „Trefferkugeln". Dann heißt X $(N, K, n)-$ hypergeometrisch verteilt, und es gilt für $0 \leq n \leq N$, $0 \leq k \leq K$:*

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Vergleichen wir mit der Binomialverteilung mit demselben Wert n und $p = K/N$. Zunächst fällt auf, dass für $k > K$ die hypergeometrisch verteilte Trefferzahlgröße die Wahrscheinlichkeit für k Treffer gleich Null ist - für die entsprechende binomialverteilte gilt das nicht. Auch die andern Werte werden anders. Beispiel:

$N = 10$, $K = 5$, $n = 5$, $k = 2$. Binomialverteilung mit entsprechendem $p = 1/2$ ergibt die Wahrscheinlichkeit $\binom{5}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = 0.3125$, hypergeometrische Verteilung ergibt: $\binom{5}{2}\binom{5}{3}/\binom{10}{5} = 0.396$. Im Gegensatz dazu wird man finden, dass für N *viel* größer als n keine nennenswerten Unterschiede auftreten.

2.3. Eine Anwendung der hypergeometrischen Verteilungen: Fisher's exakter Test. Wir betrachten folgende Situation: Auf einer Population haben wir zwei Eigenschaften (sagen wir z.B. „Student" und „Internetbenutzer"). Das sind zwei Variablen mit jeweils zwei Ausprägungen „ja", „nein" bzw. „1", „0". Dann interessiert man sich häufig dafür, ob die eine mit der andern „irgendwie zu tun hat" oder nicht - man denke nicht an „Ursache - Wirkung"! Genauer (und viel allgemeiner als das leidige Schema von Ursache und Wirkung!) fragt man danach, ob die Verteilung der Ausprägungen der einen Eigenschaft dieselbe ist, gleichgültig, welche Ausprägung der andern Eigenschaft vorliegt. Dann nennt man die Merkmale „unabhängig" (im statistisch-wahrscheinlichkeitstheoretischen Sinne!), sonst „abhängig". Beispiele: Bei mehrmaligem Würfeln ist die nachfolgende Augenzahl unabhängig von allen vorangehenden, aber man erwartet, dass das Merkmal „Internetbenutzer" nicht unabhängig vom Merkmal „Student" ist, und zwar in dem Sinne, dass der Anteil der Internetnutzer bei den Studenten höher liegt als in der sonstigen Bevölkerung, *nicht in dem Sinne, dass jeder Student Internetnutzer wäre* (oder jeder Nichtstudent kein solcher)! Man sucht nun einen solchen Sachverhalt von Abhängigkeit empirisch zu ermitteln über eine Zufallsstichprobe. Im Beispiel möge man etwa gefunden haben (die Einträge bedeuten absolute Häufigkeiten):

	Internetnutzer	Kein Internetnutzer
Student	6	5
Kein Student	3	10

Wir setzen nun hypothetisch einmal voraus, diese Merkmale seien völlig unabhängig, und stellen uns das konkret so vor: Bei den 24 Leuten wurden 9 rein zufällig als „Internetbenutzer" ausgewählt, und von diesen 9 gerieten rein zufällig 6 in die Gruppe der 11 Studenten und 3 nur in die Gruppe der Nichtstudenten. Das könnte ja erst einmal so passiert sein: Wählen Sie 20 Leute zufällig aus, so werden Sie auch nicht genau 10 Frauen dabei haben! Aber, so fragen wir weiter: Ist es nicht sehr unwahrscheinlich, dass ein solch krasses Missverhältnis rein zufällig entsteht, wenn der Sachverhalt der Unabhängigkeit wie vorausgesetzt besteht? Genau diese Wahrscheinlichkeitsfrage beantworten wir mit Einsetzen der hypergeometrischen Verteilung, die genau unserem Unabhängigkeitsmodell entspricht; zunächst: Wie wahrscheinlich ist es, *genau die beobachtete* Tafel von Häufigkeiten zu erhalten? Aus 24 Leuten, davon 11 Studenten, wurden rein zufällig 9 ausgewählt als Internetnutzer, das macht $\binom{24}{9}$ gleichwahrscheinliche Fälle. Die Anzahl der günstigen Fälle für unsere Tafel ergibt sich daraus, dass unter den 9 Internetnutzern genau 6 Studenten „gezogen" wurden und 3 Nichtstudenten, das macht $\binom{11}{6}\binom{13}{3}$, also haben wir genau die Wahrscheinlichkeit

$$\frac{\binom{11}{6}\binom{13}{3}}{\binom{24}{9}} \approx 0.1$$

dafür, genau die obenstehende Tafel zu beobachten. Aber unsere Frage ist die nach der Wahrscheinlichkeit dafür, ein mindestens so starkes Missverhältnis zu beobachten, wir haben also die Wahrscheinlichkeiten aller Tafeln zu addieren, welche ein

solches darstellen. Es sind dies die folgenden:

(2.2)
$$\begin{array}{|c|c|}\hline 6&5\\\hline 3&10\\\hline\end{array},\ \begin{array}{|c|c|}\hline 7&4\\\hline 2&11\\\hline\end{array},\ \begin{array}{|c|c|}\hline 8&3\\\hline 1&12\\\hline\end{array},\ \begin{array}{|c|c|}\hline 9&2\\\hline 0&13\\\hline\end{array},$$
$$\begin{array}{|c|c|}\hline 2&9\\\hline 7&6\\\hline\end{array},\ \begin{array}{|c|c|}\hline 1&10\\\hline 8&5\\\hline\end{array},\ \begin{array}{|c|c|}\hline 0&11\\\hline 9&4\\\hline\end{array}.$$

Man beachte: Die zweite Reihe von Tafeln repräsentiert ein mindestens ebenso deutliches Missverhältnis wie beobachtet, nur nach der andern Seite: Unabhängigkeit bedeutet ja, dass auch nicht etwa unter den Nichtstudenten die Internetnutzer überrepräsentiert wären. Genauer ist einzusehen, dass man nicht etwa mit der Tafel

$$\begin{array}{|c|c|}\hline 3&8\\\hline 6&7\\\hline\end{array}$$

zu beginnen hätte: Bei dieser Tafel ist der Anteil der Internetnutzer unter den Studenten $3/11$, unter den Nichtstudenten $6/13$, die absolute Differenz zwischen diesen beiden Zahlen (die laut Hypothese als gleich *zu erwarten* wären), ist $6/13 - 3/11 = 27/143$. Bei der tatsächlich beobachteten Tafel haben wir eine Diskrepanz von $6/11 - 3/13 = 45/143$, und diese Zahl ist größer als $27/143$. Stellen Sie analog fest, dass dagegen die Tafel

$$\begin{array}{|c|c|}\hline 2&9\\\hline 7&6\\\hline\end{array}$$

ein größeres Missverhältnis als $45/143$ darstellt. Das Aufaddieren aller Wahrscheinlichkeiten für alle Tafeln aus 2.2 ergibt:

$$\frac{1}{\binom{24}{9}}\left(\binom{11}{6}\binom{13}{3}+\binom{11}{7}\binom{13}{2}+\binom{11}{8}\binom{13}{1}+\binom{11}{9}\binom{13}{0}\right.$$
$$\left.+\binom{11}{2}\binom{13}{7}+\binom{11}{1}\binom{13}{8}+\binom{11}{0}\binom{13}{9}\right),\ \text{das ist etwa } 0.21.$$

Nun zur Bewertung: Es ist nicht *sehr* unwahrscheinlich, bei reinem Zufall ein solches Missverhältnis zu erhalten wie beobachtet oder ein größeres. Wir haben also nicht gerade so etwas Seltenes wie einen Lotto-Hauptgewinn beobachtet. Damit ist die Hypothese nicht stark erschüttert, trotz des deutlichen Missverhältnisses in der Stichprobe. Das bedeutet aber keineswegs, dass wir die Hypothese glauben und als wahr annehmen sollten, vielmehr werden wir uns erklären: Die Stichprobe war recht klein, und größere Stichproben könnten die Hypothese sehr wohl noch zu Fall bringen. Eine Beobachtung etwa wie

$$\begin{array}{|c|c|}\hline 60&50\\\hline 30&100\\\hline\end{array}$$

zeigt dieselbe Diskrepanz, aber eine solche wäre *viel unwahrscheinlicher als* $1/5$ (das Resultat ist praktisch Null) unter der Hypothese, dass sie zufällig herauskäme, also die Merkmale in der gesamten Population unabhängig wären. Allerdings wäre es nur mittels eines Computerprogramms möglich, diese Wahrscheinlichkeit nach dem oben gezeigten Muster auszurechnen, weil es zu viele Summanden gibt und zudem die auftretenden Zahlen $\binom{n}{k}$ zu groß werden. Später lernen wir dazu den sogenannten χ^2- Test kennen, der gerade im Falle hoher Einträge eine sehr akkurate Näherung ergibt. Mittels des Computers kann man auch ohne weiteres

die Verallgemeinerung auf Merkmale mit mehr als zwei Ausprägungen noch nach obenstehendem Muster rechnen.

2.4. Die Normalverteilungen. Es handelt sich um die bekannte Glockenform (im Dichtebild), allerdings ist es eine ganz bestimmte Glockenform mit ihren durch bloßes Strecken, Stauchen und Verschieben gegebenen Modifikationen. (Mathematisch kann man eine unendliche Fülle völlig andersartiger Glockenkurven bilden.) Die folgende Definition benötigt man zum praktischen Umgang nicht, sie soll nur klarstellen, dass es sich um eine durch zwei Parameter definiterte feste Familie von Dichtefunktionen handelt. Außerdem werden die weiterhin oft zu benutzenden Bezeichnungen insbesondere für die zugehörigen Verteilungsfunktionen eingeführt.

DEFINITION 12. *Für jede Zahl $\mu \in \mathbb{R}$ und jede Zahl $\sigma > 0$ definiert man folgende Dichte:*

$$\boxed{\varphi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} \ (x \in \mathbb{R}))}, \tag{2.3}$$

$\varphi_{\mu,\sigma}$ = Dichte der $(\mu, \sigma)-$Normalverteilung. Die zugehörige Verteilungsfunktion nennen wir $\Phi_{\mu,\sigma}$. Also:

$$\boxed{\begin{array}{ll} \Phi_{\mu,\sigma}(a) & = \text{Flächeninhalt unter dem Graphen der Dichte } \varphi_{\mu,\sigma} \text{ links von } a \\ & (\text{als Integral zu berechnen}) \ (\text{für alle } a \in \mathbb{R}). \end{array}}$$

Speziell für $\mu = 0, \sigma = 1$ ergibt sich die Standard-Normalverteilung, und man schreibet manchmal kurz: φ für $\varphi_{0,1}$, Φ für $\Phi_{0,1}$. Ist X eine $(\mu, \sigma)-$ normalverteilte Größe, so ist mit dieser Definition:

$$\boxed{P(X \leq a) = \Phi_{\mu,\sigma}(a) = \int_{-\infty}^{a} \varphi_{\mu,\sigma}(x)dx \ (\text{für alle } a \in \mathbb{R}).} \tag{2.4}$$

Natürlich gilt mit Satz 1 wieder: $\Phi'_{\mu,\sigma} = \varphi_{\mu,\sigma}$. (Vgl. Abb. 5 für die Standard-Normalverteilung.)

Der folgende Satz 4 erklärt die überragende Bedeutung der Normalverteilungen:

SATZ 5.

$$\boxed{\begin{array}{l} \textit{Zentraler Grenzwertsatz (untechnische Formulierung)}: \\ \textit{Lange Summen unabhängiger Variablen sind annähernd normalverteilt,} \\ \textit{wenn nur die Streuungen der summierten Variablen in endlichen Grenzen} \\ \textit{und oberhalb einer festen Zahl} > 0 \textit{ bleiben, außerdem die dritten zentralen} \\ \textit{Momente (wie Varianzen gebildet, nur mit dritter Potenz) beschränkt bleiben.} \end{array}}$$

Insbesondere sind all diese Bedingungen erfüllt im praktischen Hauptfall der Anwendung, dass es sich bei den Summanden um unabhängige Kopien ein und derselben Variable handelt, wie man bei Stichproben hat (genau genommen müsste man sie „mit Zurücklegen" ziehen, um Unabhängigkeit zu wahren, aber das kann man getrost verletzen, wenn der Umfang gering gegen den Populationsumfang ist).

Bemerkung: Was „hinreichend lang" ist, muss man aus Erfahrung lernen. Faustregel für nicht allzu anti-normalverteilte zu summierende Variablen (also nicht extrem schief oder U-förmig): Länge 10 ist schon ziemlich gut. Um einen Eindruck davon zu geben, betrachten wir folgende Abbildung 8, welche die Entwicklung von einer Gleichverteilung auf dem Intervall $[-1, 1]$ zur Normalverteilung durch Summenbildung unabhängiger Einzelwerte für verschiedene Summenlängen zeigt. Damit die Verteilungen nicht immer breiter werden, dividieren wir dabei die Summen der Länge n stets durch n, zeigen also jeweils die Verteilung der Variablen: Arithmetisches Mittel von n zufällig ausgewählten Werten einer auf $[-1, 1]$ gleichverteilten Größe.

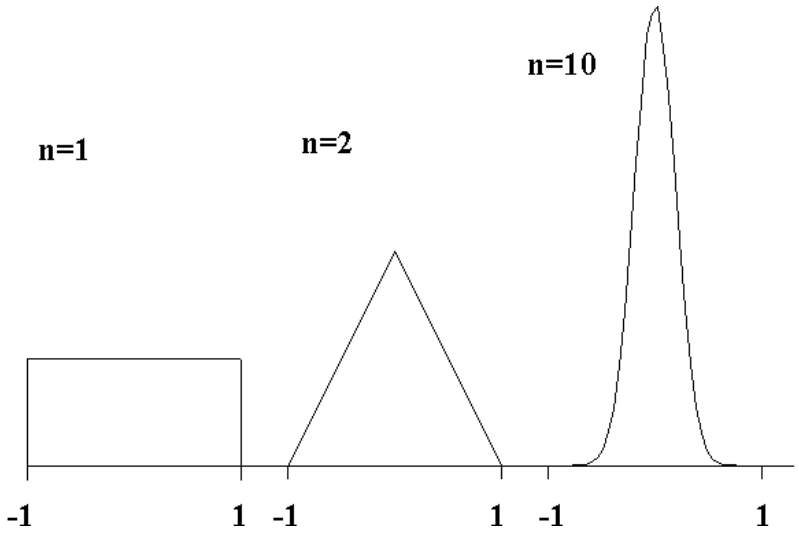


Abb. 8: Die sukzessive Näherung an die Normalverteilung durch Summenbildung unabhängiger Variablen, hier z.B. gleichverteilter Variablen auf $[-1, 1]$ - n gibt die Zahl der unabhängigen Summanden.

Hier sind drei weitere Illustrationen des Zentralen Grenzwertsatzes: Binomialverteilte Variablen sind für große n lange Summen unabhängiger Kopien ein und derselben Bernoulli-Variablen. Somit sollte bei p, das nicht zu weit von $1/2$ liegt, sehr schnell, sonst (wegen der zunächst starken Asymmetrie) langsamer, also erst für größere n, eine Normalverteilung angenähert erscheinen.

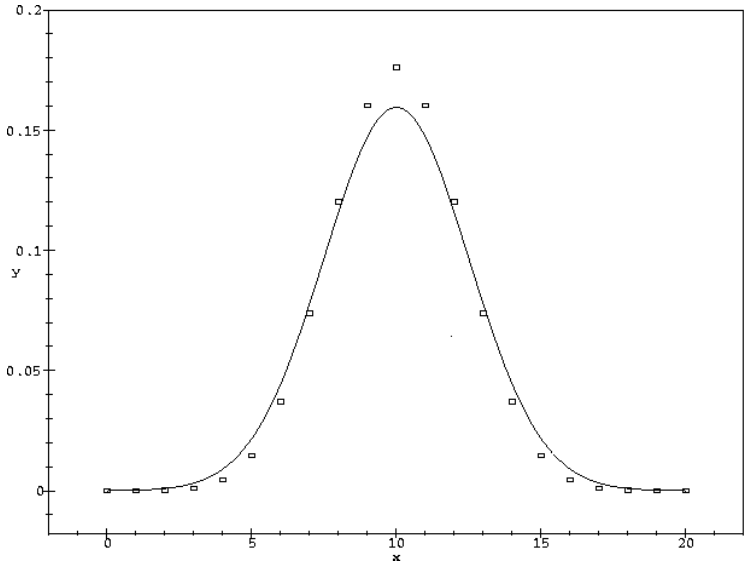


Abb. 9: Bereits recht gute Annäherung an eine Normalverteilung durch die Binomialverteilung mit $n = 20$, $p = 1/2$ (die Quadrate geben die Werte der Wahrscheinlichkeitsfunktion)

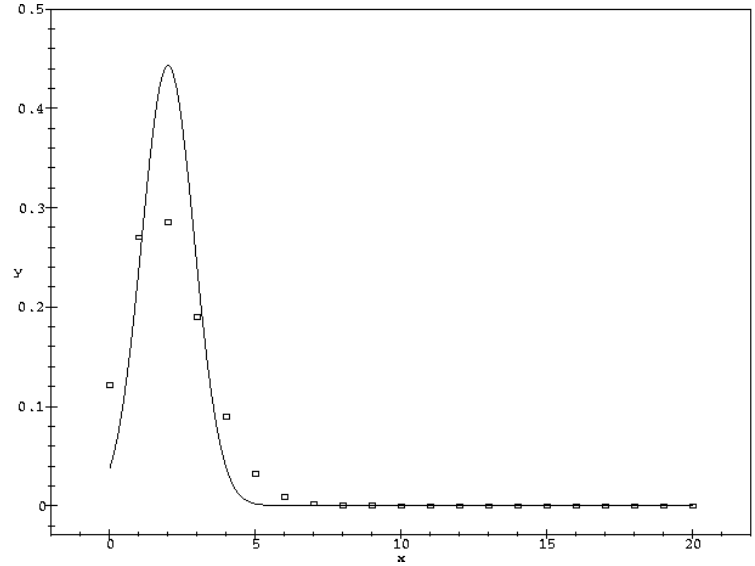


Abb. 10: Weniger gute Annäherung an eine Normalverteilung mit der schiefen Binomialverteilung zu $n = 20$, $p = 1/10$

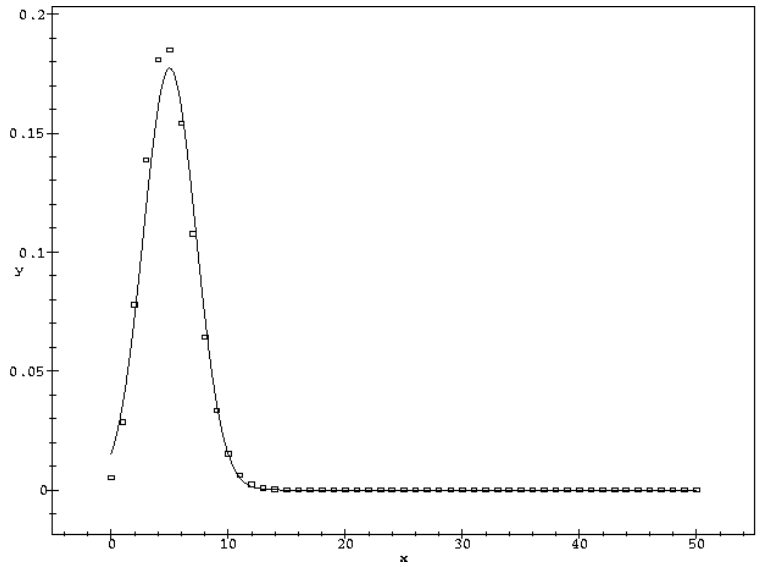


Abb. 11: Auch bei schiefer Binomialverteilung gute Annäherung an eine Normalverteilung mit wachsendem n, hier $n = 50$, wieder $p = 1/10$.

Dabei ist natürlich zu beachten, dass man immer nur eine ordentliche Annäherung an die jeweils *passende* Normalverteilung erhält, d.h. diejenige mit den richtigen Werten μ, σ. (Vgl. den nächsten Abschnitt 3. für das Rechnen mit μ, σ.) Auch die hypergeometrischen Verteilungen nähern sich Normalverteilungen an, wenn nur N recht viel größer als n ist und n ebenfalls relativ zum Abstand von K/N zu $1/2$ nicht zu klein. Hier zwei Beispiele:

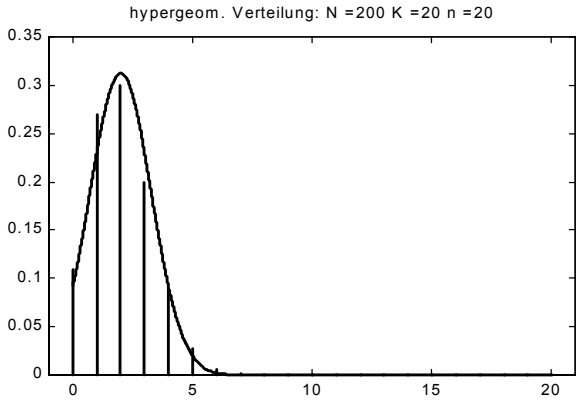


Abb. 12: Annäherung an die entsprechende Normalverteilung ist schon nicht schlecht bei der hypergeometrischen Verteilung zu $N = 200$, $K = 20$, $n = 20$.

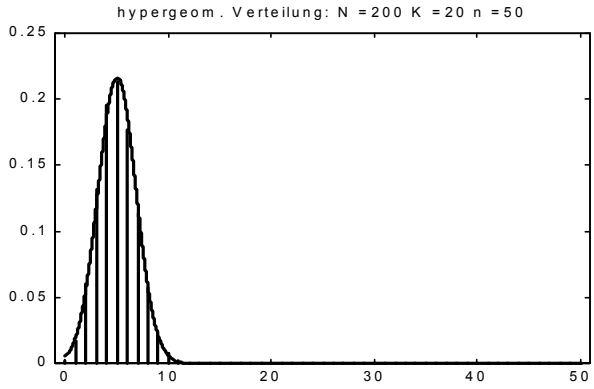


Abb. 13: Für $n = 50$ (weiterhin $N = 200$, $K = 20$) wird die Annäherung an die entsprechende Normalverteilung auch für die hypergeometrische deutlich besser.

Zwei Umstände machen den Zentralen Grenzwertsatz so interessant für mannigfachste Anwendungen: Einmal erklärt er, warum man so viele Normalverteilungen als „natürliche" in der Welt antrifft: Immer dann, wenn sich unabhängig gewisse Variablen mitteln, um einen Wert zu bestimmen (Erbschaften, Ernährung usw. zur Bestimmung der Körperlänge eines Erwachsenen; Erbschaften und Anregungen und mehr zur Entwicklung einer Intelligenz), dann kristallisiert sich eine Normalverteilung heraus. Zweitens: Einige wiederkehrend zu benutzende Verteilungen von großer praktischer Bedeutung erweisen sich sofort als annähernde Normalverteilungen. Vor allem sind Mittelungsgrößen zu nennen: Man geht aus von einer Größe X, betrachtet nun aber Stichproben aus deren Population Ω, Stichproben eines festen Umfang n. So entsteht eine neue Größe, indem man jeder Stichprobe ihr arithmetisches Mittel der $X-$ Werte zuordnet. Diese Variable nennen wir $\overline{X}$. Man beachte, dass der Stichprobenumfang in dieser Notation unterdrückt ist. Man teilt ihn gewöhnlich in Worten oder Zusätzen wie „$n = ...$" mit. Einen beobachteten Wert von $\overline{X}$ bezeichnet man konsequent mit $\overline{x}$ und benutzt ihn als Schätzwert für den meist unbekannten Wert $\mu(X)$. Gewöhnlich kennt man ja nicht sämtliche X-Werte der Gesamtpopulation. Nun wissen wir aber mit dem Zentralen Grenzwertsatz, dass für hinreichend große n die Variable $\overline{X}$ annähernd normalverteilt ist (für das arithmetische Mittel bildet man eine Summe unabhängiger Einzelwerte, und die anschließende Division durch n ergibt nur ein Stauchen/Strecken im Verteilungsbild, ändert daher nichts am Normalverteilungscharakter, wie auch im Beispiel illustriert), und somit werden wir mit der praktischen Beherrschung der Normalverteilungen aussagen können, mit welcher Sicherheit (oder Wahrscheinlichkeit) bei einer solchen Schätzung der Fehler höchstens wie groß ist. Wir können uns also von der Qualität der Schätzung ein *quantitativ genaues* Bild machen, *ohne μ zu kennen!* Auf diese Weise führt uns die Normalverteilung auch in die „Schließende Statistik" (auch „Inferenzstatistik" genannt).

3. Bedingte Wahrscheinlichkeiten und Unabhängigkeit

Es gibt verschiedene Begriffe von „Unabhängigkeit". In unserem Kontext kommt es vor allem darauf an, den wahrscheinlichkeitstheoretischen Begriff der Unabhängigkeit richtig zu verstehen und insbesondere nicht mit dem logischen zu verwechseln. Eine Aussage A hängt logisch von einer andern Aussage B ab, wenn

man aus B entweder A oder aber „*nicht* A" schließen kann. Zum Beispiel: Wenn der erste Wurf mit einem Würfel eine Drei ergibt (A), dann muss die Augensumme nach zwei Würfen einen Wert unter 10 ergeben (B). B ist also logisch abhängig von A. Abhängigkeit im wahrscheinlichkeitstheoretischen Sinne ist dagegen eine schwächere Beziehung, man kann sie auch als Verallgemeinerung logischer Abhängigkeit auffassen: Die Ereignisse A („erster Wurf eine Drei") und C („Augensumme nach 2 Würfen unter 9") sind nicht logisch abhängig: Aus A kann man weder C noch „*nicht* C" logisch folgern. Aber unter der Bedingung A ist C wahrscheinlicher als C ohne Voraussetzung irgendeiner Bedingung wäre; dies bedeutet genau: C ist im wahrscheinlichkeitstheoretischen Sinne abhängig von A. Diesen Begriff wollen wir nun genauer quantitativ klären. Dazu definieren wir:

DEFINITION 13. *Für* $P(B) \neq 0$ *wird definiert:*

$$P(A\,|\,B) := \frac{P(A \cap B)}{P(B)} \text{ (lies: Wahrscheinlichkeit von } A \text{ bedingt durch } B).$$

Zur inhaltlichen Interpretation: Man schaut B als eingeschränkte Menge von möglichen Ausgängen an und bestimmt in diesem Rahmen die Wahrscheinlichkeit von A. Selbstverständlich sind nunmehr die „günstigen Fälle" diejenigen in $A \cap B$. Auf keinen Fall interpretiere man B als „Ursache von A" oder auch nur als zeitlich auf A folgend.

Wichtige Bemerkung zur Definition: Man verwendet diese Definition fast nie, um eine bedingte Wahrscheinlichkeit zu berechnen, vielmehr ergeben sich die Werte bedingter Wahrscheinlichkeiten meist unmittelbar. Dagegen ist sie von theoretischer Bedeutung für weitere Zusammenhänge und Formeln (s.u.). Die Division durch $P(B)$ ergibt sich gerade daraus, dass B als neue Menge Ω zu betrachten ist, und es sollte $P(B\,|\,B) = 1$ sein.

Nun können wir vorläufig A und B *im wahrscheinlichkeitstheoretischen Sinne unabhängig* nennen, wenn $P(A\,|\,B) = P(A)$, falls $P(B) \neq 0$. Beispiel: Zwei mal wird gewürfelt. A : „Erster Wurf eine Drei", B : „Zweiter Wurf eine Vier". Offenbar ist $P(A\,|\,B) = 1/6 = P(A)$, also A und B unabhängig. (Man beachte: B folgt sogar zeitlich auf A,was dem Sinn von $P(A\,|\,B)$ keinen Abbruch tut. Ausführliche Berechnung von $P(A\,|\,B)$: Im verkleinerten Topf B sind die Ausgänge $(1,4),(2,4),(3,4),(4,4),(5,4),(6,4)$. Das sind 6 gleichwahrscheinliche. In $A \cap B$ ist nur $(3,4)$, also ein günstiger Fall, macht Wahrscheinlichkeit $1/6$. Dagegen sind B und das folgende Ereignis C abhängig: C : „Die Augensumme beider Würfe ist 5". Denn offenbar $P(C\,|\,B) = 1/6$, während $P(C) = 4/36 = 1/9$. (Man zähle das nach.)

Wir kommen zur theoretischen Bedeutung der bedingten Wahrscheinlichkeiten:

SATZ 6. *Es gilt allgemein für* $P(B) \neq 0$:

$$P(A \cap B) = P(A\,|\,B)P(B). \tag{3.1}$$

Dazu ist nur die definierende Gleichung für $P(A\,|\,B)$ mit $P(B)$ zu multiplizieren. Diese Formel ist so häufig direkt anwendbar wie man direkt an $P(A\,|\,B)$ herankommt. Beispiel: Man zieht aus einer Urne mit 10 Kugeln, davon 5 rot, zwei Kugeln ohne Zurücklegen. Sei A das Ereignis: „Die zweite Kugel ist rot", B das Ereignis: „Die erste Kugel ist rot". Dann ist $P(A \cap B) = \frac{4}{10} \cdot \frac{5}{10} = \frac{1}{5}$; denn nach dem Herausziehen einer roten Kugel verbleiben 9 Kugeln, davon 4 rote, so dass die Wahrscheinlichkeit für das Ziehen einer zweiten roten Kugel 4/10

beträgt. Weiter können wir folgern, dass bei unabhängigen Ereignissen A, B stets $P(A \cap B) = P(A)P(B)$ ist, da $P(A\,|B) = P(A)$ in diesem Falle gilt. Um nun die Randfälle $P(B) = 0$ noch mitzunehmen, in denen diese Formel ebenfalls gilt, *definiert* man:

DEFINITION 14. *Zwei Ereignisse A und B (im Rahmen eines Zufallsexperiments) heißen unabhängig, wenn*

$$P(A \cap B) = P(A)P(B).$$

(Man beachte, dass dies für $P(B) \neq 0$ gleichwertig zu $P(A\,|B) = P(A)$ ist.)

Eine sehr wichtige Verallgemeinerung dieses Begriffs auf Variablen kann man daraus entwickeln: Die für eine Variable X interessierenden Ereignisse lauten $X \leq a$, $a \in \mathbb{R}$. Denn deren Wahrscheinlichkeiten bestimmen bereits die Verteilungsfunktion. Somit kann man in typisch mathematischer Weise den Begriff der Unabhängigkeit von Variablen auf den von Ereignissen zurückführen:

DEFINITION 15. *Zwei Variablen X, Y heißen unabhängig, wenn für jedes Paar (a, b) reeller Zahlen gilt: Die Ereignisse $X \leq a$ und $Y \leq b$ sind unabhängig.*

Beispiel: Körperlänge und Körpergewicht (auf der Population der Menschen) sind sicher nicht unabhängig; zwar gibt es kurze Dicke und lange Dünne, aber wenn die Körperlänge unter einer Grenze wie 160 cm bleibt, so ist jedenfalls die Wahrscheinlichkeit für ein Körpergewicht unter 60 kg erhöht, d.h. größer als unter allen Leuten. Ein wichtiger Teil der Statistik (vgl. das letzte Kapitel) beschäftigt sich damit, die Abhängigkeiten zwischen mehreren Variablen zu beschreiben.

Wir kommen zu zwei weiteren recht wichtigen Formeln der Wahrscheinlichkeitsrechnung, die mit bedingten Wahrscheinlichkeiten arbeiten:

SATZ 7 (Bayessche Formel und Formel von der totalen Wahrscheinlichkeit). *Es seien $P(A), P(B) \neq 0$. Dann gilt folgende Bayessche Formel:*

$$P(B\,|A) = \frac{P(A\,|B)P(B)}{P(A)} \tag{3.2}$$

Sei $\{B_1, ..., B_n\}$ eine Klasseneinteilung von Ω, d.h.

$$\Omega = \bigcup_{i=1}^{n} B_i \text{ und } B_i \cap B_j = \emptyset \text{ für } i \neq j,\ 1 \leq i, j \leq n.$$

Dann gilt folgende Formel von der totalen Wahrscheinlichkeit:

$$P(A) = \sum_{i=1}^{n} P(A\,|B_i)P(B_i). \tag{3.3}$$

Die Formeln sind ungeachtet ihrer Nützlichkeit sehr einfach zu zeigen: Die erste ergibt sich sofort aus $P(A \cap B) = P(A\,|B)P(B)$ und $P(B \cap A) = P(B\,|A)P(A)$ nach vorigem Satz, aber $A \cap B = B \cap A$. Damit $P(A\,|B)P(B) = P(B\,|A)P(A)$, nun teile man diese Gleichung durch $P(B)$. Der Nutzen der Formel besteht darin, dass man bedingte Wahrscheinlichkeiten „herumdrehen" kann: Man stößt vielfach auf Situationen, in denen $P(B\,|A)$ inhaltlich interessiert, aber nut $P(A\,|B)$ empirisch zugänglich ist. Etwa A : Auftreten eines Symptoms, B : Vorliegen einer bestimmten Krankheit. Diagnostisch wichtig ist $P(B\,|A)$, aber nur $P(A\,|B)$ ist empirisch zugänglich. Ebenso sind $P(A\,|\overline{B})$ sowie $P(B)$ zugänglich, und so kann man mittels der zweiten Formel den Nenner ausrechnen als $P(A) = P(A\,|B)P(B) + P(A\,|\overline{B})P(\overline{B})$.

Mit der Bayesschen Formel hat man dann $P(B\,|A\,)$, also im Beispiel die Wahrscheinlichkeit dafür, dass jemand die Krankheit hat, bei dem das Symptom auftritt. Es gibt noch einen theoretisch weiter reichenden Zweig von Bayesscher Statistik, der eine recht große Bedeutung erlangt hat und insbesondere eine weiterführende Art des induktiven Schließens aus Empirischem darstellt, und auch hier steht dies Herumdrehen bedingter Wahrscheinlichkeiten am Anfang der Überlegungen: Stellen Sie sich vor, Sie haben verschiedene Theorien $T_1, ...T_n$, sagen wir Wahrscheinlichkeitsmodelle, als mögliche Erklärungen für einen Phänomenbestand, und fragen sich, welche dieser Theorien am besten passt zu allen Beobachtungsdaten B (als ein komplexes Ereignis aufgefasst). Nun können Sie $P(B\,|T_i\,)$ jeweils bestimmen, d.h. die Wahrscheinlichkeit für die Beobachtungen bei Voraussetzung der Theorie T_i, $1 \leq i \leq n$. Ein primitiveres Prinzip (man nennt es „Maximum Likelihood") würde nun einfach eine solche Theorie bevorzugen, bei der B maximale Wahrscheinlichkeit erhält. Das elaboriertere Bayesprinzip lautet stattdessen: Man fasse die Gültigkeit einer Theorie wiederum als ein Ereignis auf, schreibe also den Theorien selbst Wahrscheinlichkeiten zu, und zwar bedingte durch die Beobachtung, gemäß Bayesscher Formel:

$$P(T_i\,|B\,) = \frac{P(B\,|T_i\,)P(T_i)}{P(B)}.$$

Interessant sind hier die Ausgangswahrscheinlichkeiten $P(T_i)$ für die Theorien; setzt man sie alle gleich, so landet man wieder bei Maximum Likelihood: $P(T_i\,|B\,)$ wird am höchsten für die Theorien, für die $P(B\,|T_i\,)$ maximal ist. Aber man kann auch Vorerfahrungen oder plausible Annahmen einfließen lassen, um diese Ausgangswahrscheinlichkeiten differenzierter anzusetzen.

Zur Begründung der Formel von der totalen Wahrscheinlichkeit ist nur zu bemerken:

$$P(A) = P\,(\Omega \cap A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A\,|B_i\,)P(B_i)$$

Dabei wurde benutzt, dass $\Omega \cap A = \bigcup_{i=1}^{n} (A \cap B_i)$ und die Summenformel 1.1 aus den Axiomen sowie Formel 3.1. Die Formel 3.3 ist nicht etwa nur im Bayes-Zusammenhang zu nutzen, sondern elementar tritt häufig die Situation auf, dass man bedingte Wahrscheinlichkeiten zur Verfügung hat und eine unbedingte sachlich interessiert: Zum Beispiel ist die relative Häufigkeit einer Eigenschaft bei allen Bundesbürgern auf diese Weise zu ermitteln, wenn man die relativen Häufigkeiten für die einzelnen Bundesländer und dazu die Bevölkerungsanteile der Bundesländer an der gesamten Republik kennt.

4. Das Rechnen mit μ und σ

Es gibt einige sehr nützliche Formeln für das Rechnen mit μ und σ, die insbesondere im Zusammenhang mit der Normalverteilung sehr wichtig sind - erinnern wir uns daran, dass Kenntnis von μ, σ bei einer Normalverteilung ausreicht, um jegliche Wahrscheinlichkeitsfrage zu beantworten. Wir stellen diese Formeln in zwei Blöcken zusammen - der erste umfasst nur stets gültige Formeln, der zweite solche, die gewissen Voraussetzungen unterliegen.

SATZ 8. *Es gelten stets folgende Formeln - seien X, Y Variablen auf derselben Population, $a \in \mathbb{R}$:*

$$\left.\begin{array}{c}\mu(aX) = a\mu(X) \\ \mu(X+Y) = \mu(X) + \mu(Y)\end{array}\right\} \text{ Linearität von } \mu \tag{4.1}$$

$$\begin{aligned}\sigma^2(aX) &= a^2\sigma^2(X),\ \sigma(aX) = |a|\,\sigma(X),\ \sigma(X+a) = \sigma(X) \\ \sigma^2(X) &= \mu\left(X^2\right) - \mu^2(X).\end{aligned}$$

Eine wichtige Folgerung: Für $\sigma(X) \neq 0$ hat man

$$\begin{aligned}\mu\left(\frac{X-\mu(X)}{\sigma(X)}\right) &= 0 \\ \sigma\left(\frac{X-\mu(X)}{\sigma(X)}\right) &= 1.\end{aligned}$$

Man kann also jede Variable mit nichtverschwindender Varianz durch diese lineare Transformation der Standardisierung auf Mittelwert Null und Varianz (damit auch Streuung) 1 bringen.

Alle diese Formeln sind sehr leicht nachzurechnen. Die Formel $\sigma^2(X) = \mu\left(X^2\right) - \mu^2(X)$ wird weiter unten allgemeiner für die Kovarianz bewiesen.

Zunächst ist noch ein Spezialfall von Abhängigkeit zweier Variablen zu definieren, damit die Voraussetzungen für weitere Formeln angemessen formuliert werde können. Die zugehörigen Überlegungen ergeben eine weitere nützliche Formel für die Varianz.

Rechnet man $\sigma^2(X+Y)$ aus, so erhält man

$$\begin{aligned}\sigma^2(X+Y) &= \mu\left((X+Y)^2\right) - \mu^2(X+Y) \\ &= \mu\left(X^2\right) + \mu\left(Y^2\right) + 2\mu(XY) - \mu^2(X) - \mu^2(Y) - 2\mu(X)\mu(Y) \\ &= \sigma^2(X) + \sigma^2(Y) + 2\left(\mu(XY) - \mu(X)\mu(Y)\right).\end{aligned}$$

Man sieht daran, dass sich im allgemeinen die Varianzen der Variablen nicht zur Varianz der Summe addieren, sondern ein Zusatzterm entsteht. Dieser wird noch im Zusammenhang mit linearer Regression interessieren, und daher heben wir ihn mit einer gesonderten Definition hervor:

DEFINITION 16. *Die Kovarianz der Variablen X, Y, welche auf derselben Population Ω definiert seien, ist definiert als*

$$\begin{aligned}Cov(X,Y) \;:\; &= \mu\left((X-\mu(X))(Y-\mu(Y))\right). \text{ Es folgt:} \\ &\text{Man schreibt auch } \sigma^2(X,Y) \text{ für } Cov(X,Y).\end{aligned}$$

Man beachte die Analogie zur Varianz: Setzt man $X = Y$, so ist $Cov(X,X) = \sigma^2(X)$. Allerdings kann eine Kovarianz bei verschiedenen Variablen durchaus negativ sein. Der folgende einfache Satz zeigt die wichtigen Rechengesetze für die Kovarianz:

SATZ 9. *Es gelten folgende Formeln für die Kovarianz - es seien dabei X, Y, Z Variablen auf derselben Population und a eine beliebige reelle Zahl:*

(4.2)

$$\begin{array}{lcl} (i) & \left.\begin{array}{c} Cov(aX,Y) = aCov(X,Y) \\ Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z) \end{array}\right\} & \textit{Linearität der Kovarianz im ersten Schlitz, macht Bilinearität mit (ii)} \\ (ii) & Cov(X,Y) = Cov(Y,X) & \\ (iii) & Cov(X,Y) = \mu(XY) - \mu(X)\mu(Y). & \end{array}$$

Die Gleichungen (i) folgen sofort aus Kommutativgesetzen für $+, \cdot$ und dem Distributivgesetz für die reellen Zahlen sowie der Linearität von μ. Zeigen wir exemplarisch die zweite Gleichung von (i), die sich bei Gebrauch eben auch wie ein Distributivgesetz „anfühlt":

$$\begin{array}{rcl} Cov(X+Y,Z) & = & \mu\left[(X+Y-\mu(X+Y))\cdot(Z-\mu(Z))\right] \\ & = & \mu\left[(X-\mu(X))\cdot(Z-\mu(Z)) + (Y-\mu(Y)\cdot(Z-\mu(Z))\right] \\ (\text{Linearität von } \mu) & = & \mu\left[(X-\mu(X))\,(Z-\mu(Z))\right] + \mu\left[(Y-\mu(Y)\,(Z-\mu(Z))\right] \\ & = & Cov(X,Z) + Cov(Y,Z) \end{array}$$

Die Gleichung (ii) folgt sofort aus dem Kommutativgesetz für die Multiplikation reeller Zahlen. Man beachte, dass (i) und (ii) zusammen auch ergeben, dass $\sigma^2(X, aY) = a\sigma^2(X,Y)$ und $\sigma^2(X, Y+Z) = \sigma^2(X,Z) + \sigma^2(Y,Z)$, d.h. die Linearität von σ^2 im 2. Schlitz, was man dann zusammen Bilinearität nennt.

Die Gleichung (iii) kann man so beweisen (damit haben wir insbesondere auch die oben angeführte entsprechende Formel für die Varianz bewiesen, da diese nur den Spezialfall $X = Y$ darstellt):

$$\begin{array}{rcl} \mu\left((X-\mu(X))\,(Y-\mu(Y))\right) & = & \mu\left(XY - X\mu(Y) - Y\mu(X) + \mu(X)\mu(Y)\right) \\ & = & \mu(XY) - 2\mu(X)\mu(Y) + \mu(X)\mu(Y) \\ & = & \mu(XY) - \mu(X)\mu(Y). \end{array}$$

Die Sache funktioniert also mit einfachem Ausrechnen der zusammengesetzten Variablen und anschließender Nutzung der Linearität von μ.

Wir haben oben gesehen, dass die naiv zu erwartende Formel $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y)$ genau dann gilt, wenn $Cov(X,Y) = 0$, wenn also der Zusatzterm verschwindet. Später (im Abschnitt über lineare Regression) werden wir genauer verstehen, dass dies gleichbedeutend mit *linearer Unabhängigkeit* von X, Y ist und was diese *lineare* Unabhängigkeit bedeutet. Man achte genau darauf, dass der mathematische Sprachgebrauch hier einfach rein sprachlich ordentlich ist: Unabhängigkeit der Variablen impliziert lineare Unabhängigkeit, aber nicht umgekehrt. Lineare Abhängigkeit dagegen impliziert Abhängigkeit überhaupt, aber nicht umgekehrt. Nunmehr *definieren* wir einfach nur:

DEFINITION 17. *Zwei Variablen X, Y heißen linear unabhängig, wenn ihre Kovarianz verschwindet, also $Cov(X,Y) = 0$ gilt.*

Wir haben den folgenden einfachen Satz, der besagt, dass Unabhängigkeit (überhaupt) die lineare Unabhängigkeit impliziert:

SATZ 10. *Sind X, Y unabhängig, so sind X, Y auch linear unabhängig, also $Cov(X,Y) = 0$.*

Zum Beweis ist zu zeigen: Unter der Voraussetzung der Unabhängigkeit gilt $\mu(XY) = \mu(X)\mu(Y)$. Da wir nur für Variablen X mit nur endlich vielen Werten $a \in \mathbb{R}$, an denen $f_X(a)$ nicht verschwindet, $\mu(X)$ definiert haben, können wir die Sache natürlich auch nur für diesen Fall beweisen, sie ist aber allgemeiner gültig. Wir haben unter der genannten Voraussetzung für X und Y:

$$\begin{aligned}
\mu(XY) &= \sum_{a,b\in\mathbb{R}} abP(X=a \text{ und } Y=b) \\
\text{(Unabhängigkeit von } X,Y) &= \sum_{a,b\in\mathbb{R}} abP(X=a)P(Y=b) \\
\text{(Distributivgesetz, Rechnen mit } \Sigma) &= \sum_{a\in\mathbb{R}} aP(X=a)\sum_{b\in\mathbb{R}} bP(Y=b) \\
&= \mu(X)\mu(Y)
\end{aligned}$$

Damit folgen mit den vorangehenden Argumenten sofort die weiterführenden Formeln für μ, σ:

SATZ 11. *Alle folgenden Formeln gelten* ***unter Voraussetzung der linearen Unabhängigkeit*** *von X, Y, also insbesondere auch dann, wenn X, Y überhaupt unabhängig sind:*

$$\begin{array}{ll}
(i) & \mu(XY) = \mu(X)\mu(Y) \\
(ii) & \sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) \\
(iii) & \sigma(X+Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)}
\end{array} \tag{4.3}$$

Wiederholtes Anwenden von (ii) ergibt die wichtige verallgemeinerte Formel: Sind $X_1, ..., X_n$ linear unabhängig (und dafür genügt bereits die paarweise lineare Unabhängigkeit dieser Variablen, was nicht etwa rein logisch selbstverständlich ist, aber eben für die ***lineare*** *Unabhängigkeit von Variablen - übrigens nicht für die Unabhängigkeit schlechthin! - zutrifft), so hat man*

$$\begin{aligned}
(iv) \quad \sigma^2\left(\sum_{i=1}^n X_i\right) &= \sum_{i=1}^n \sigma^2(X_i) \text{ und daher} \\
(v) \quad \sigma\left(\sum_{i=1}^n X_i\right) &= \sqrt{\sum_{i=1}^n \sigma^2(X_i)}
\end{aligned} \tag{4.4}$$

Man vermeide Stümpereien wie die Anwendung dieser Formeln, ohne dass die entscheidende Voraussetzung erfüllt ist, oder wie das lineare Rechnen mit der Wurzel - es ist $\sqrt{a+b} \neq \sqrt{a} + \sqrt{b}$, wenn nur a, b beide von Null verschieden sind.

Abschließend wollen wir noch die bereitgestellten Formeln dazu nutzen, Mittelwert und Varianz für binomialverteilte und hypergeometrisch verteilte sowie für Stichprobenmittelgrößen herauszupräparieren. Das ist insbesondere dafür nützlich, auf die betreffenden Variablen dann die Normalverteilungen anwenden zu können, für die man bekanntlich gerade μ, σ wissen muss.

Vorab sei noch einmal daran erinnert, dass zu einer Varialen X mit $\overline{X}$ bei vorausgesetztem Stichprobenumfang n (in der Notation tritt er nicht auf) die folgende Größe gemeint ist: *Jeder Stichprobe* aus Ω (wobei Ω die Population zu X ist) wird das arithmetische Mittel der darin vorzufindenden $X-$ Werte zugeordnet. Man beachte also: Die Population zu $\overline{X}$ ist die Menge aller Teilmengen aus Ω vom

Umfang n! Wir fassen die angesprochenen nützlichen Resultate alle zu einem Block zusammen:

SATZ 12. *(i) Sei X $(n,p)-$ binomialverteilt. Dann gilt:*

$$\begin{aligned}\mu(X) &= np\\ \sigma^2(X) &= np(1-p).\end{aligned}$$

(ii) Sei X $(N,K,n)-$ hypergeometrisch verteilt und $N>1$. Dann gilt:

$$\begin{aligned}\mu(X) &= n\frac{K}{N}\\ \sigma^2(X) &= n\frac{K}{N}\cdot\frac{N-K}{N}\cdot\frac{N-n}{N-1}\end{aligned}$$

(iii) Sei X in beliebiger Weise verteilt. Dann gilt für die Stichprobenmittelgröße $\overline{X}$ zum Stichprobenumfang $n\geq 1$:

$$\begin{aligned}\mu\left(\overline{X}\right) &= \mu(X)\\ \sigma\left(\overline{X}\right) &= \frac{\sigma(X)}{\sqrt{n}},\ \textit{entsprechend } \sigma^2\left(\overline{X}\right)=\frac{\sigma^2(X)}{n}.\end{aligned}$$

Zur Begründung von (i): Man rechnet einfach aus, dass für $n=1$ herauskommt: $\mu=p$ und $\sigma^2=p(1-p)$. Nunmehr fasst man eine $(n,p)-$ binomialverteilte Variable X als Summe von n unabhängigen $(1,p)-$ binomialverteilten (bzw. $p-$ Bernoulliverteilten) Variablen X_i auf. Dann ist $\mu(X)=\mu(\sum_{i=1}^n X_i)=\sum_{i=1}^n\mu(X_i)=np$, da $\mu(X_i)=p$ für $1\leq i\leq n$. Ebenso wird die Varianz (wegen der Unabhängigkeit, also insbesondere linearen Unabhängigkeit der X_i) n mal so groß, mit Formel 4.4. Völlig analog läuft die Begründung für (iii); denn man hat $\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i$, wobei jede der Variablen X_i dieselbe Verteilung wie X hat und zudem diese Variablen unabhängig sind. Also hat man

$$\begin{aligned}\mu\left(\overline{X}\right) &= \frac{1}{n}\sum_{i=1}^n\mu(X_i)=\frac{1}{n}n\mu(X)=\mu(X)\ \text{und}\\ \sigma^2\left(\overline{X}\right) &= \frac{1}{n^2}\sum_{i=1}^n\sigma^2(X_i)=\frac{1}{n^2}n\sigma^2(X)=\frac{\sigma^2(X)}{n}.\end{aligned}$$

(Für das Herausziehen eines Faktors $1/n^2$ bei der Varianz vgl. 4.1.)

Verbleibt noch die Berechnung für die hypergeometrische Verteilung. Sie ist für die Varianz ein wenig schwieriger, da keine Unabhängigkeit (auch keine lineare) vorliegt. Wir begnügen uns zunächst für μ mit der Bemerkung, dass die Formel für μ unter (ii) genau der unter (i) entspricht, da K/N eben dem Parameter p bei der Binomialverteilung entspricht. Weiter sehen wir die Ähnlichkeit bei σ^2: Der Faktor $(N-K)/N$ entspricht gerade dem Faktor $1-p$ bei der Varianz der Binomialverteilung. Nun kommt ein neuer hinzu, $(N-n)/(N-1)$. Offenbar ist der vernachlässigbar (wegen eines Wertes ≈ 1), wenn n viel kleiner als N ist. Das sollte einleuchten, da die hypergeometrische Verteilung sich dann der entsprechenden Binomialverteilung annähert. Wenn aber n nicht winzig gegen N ist, so sollte plausibel sein, dass man eine geringere Streuung bekommt als bei der entsprechenden Binomialverteilung, denken wir uns insbesondere den Extremfall $n=N$; dann werden alle Kugeln aus der Urne gezogen, und wir erhalten stets K als Trefferkugelzahl. Genau dies zeigt die Varianzformel, es kommt in diesem Falle Varianz Null

heraus. In derartigen Fällen liegt natürlich keine Näherung an eine Normalverteilung vor. Man beachte, dass der Fall $N = 1$ keine Schwierigkeiten macht, weil man dann auch höchstens $n = 1$ haben, nämlich eine einzige Kugel herausziehen kann, und es liegt ein besonders triviales Bernoulli-Experiment vor mit $p = 0$ oder $p = 1$, je nach dem, ob $K = 0$ oder $K = 1$.

Für diejenigen Leser, welche damit nicht zufrieden sind, folgt hier noch der Beweis beider Formeln:

Zu μ: Wir fassen das Experiment zur $(N, K, n)-$ hypergeometrisch verteilten Trefferzahlvariablen X so auf: n mal wird je eine Kugel aus der Urne ohne Zurücklegen gezogen. Wir definieren Variablen X_i so, dass X_i den Wert 1 erhält, wenn die $i-$ te gezogene Kugel eine Trefferkugel ist, sonst erhält X_i den Wert Null. Wir haben dann $X = \sum_{i=1}^{n} X_i$. Folglich ist $\mu(X) = \sum_{i=1}^{n} \mu(X_i)$. Wir zeigen nunmehr, dass $\mu(X_i) = K/N$, für alle i, $1 \leq i \leq n$. Damit folgt dann sofort die Behauptung. Offensichtlich gilt $\mu(X_1) = K/N$, da wir nur eine Kugel aus der Urne ziehen und sich X_1 als $(K/N)-$ Bernoulligröße verhält. Nun setzen wir voraus, dass bereits für $k < n$ gelte: $\mu(X_i) = K/N$ für alle $i \leq k$. Wir haben unter dieser Voraussetzung (vollständige Induktion) zu zeigen, dass auch $\mu(X_{k+1}) = K/N$ gilt - dann ist $\mu(X_i) = K/N$ für alle $i \leq n$ bewiesen. Nach der Induktionsvoraussetzung sind nach k Zügen im Mittel $\mu(\sum_{i=1}^{k} X_i) = k \cdot K/N$ Trefferkugeln gezogen. Im Mittel sind also noch $K - k \cdot K/N = K(N-k)/N$ Trefferkugeln vorhanden, unter noch insgesamt $N - k$ Kugeln. Die Wahrscheinlichkeit, eine Trefferkugel im $(k+1)$. Zug zu erhalten, ist also

$$\frac{K(N-k)}{N(N-k)} = \frac{K}{N}.$$

Also ist dies der Erwartungswert $\mu(X_{k+1})$.

Zu σ^2: Wir machen uns zunächst klar - X und X_i bedeuten dasselbe wie oben, dass

$$(*)\ \sigma^2(X) = \sigma^2\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma^2(X_i) + \sum_{1 \leq i,j \leq n, i \neq j}^{n} Cov(X_i, X_j),$$

in Verallgemeinerung der Formel $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + 2Cov(X,Y)$. Nun stellen wir aber fest, dass für alle $i \neq j, 1 \leq i, j \leq n$ der Wert von $Cov(X_i, X_j)$ stets derselbe ist, und zwar

$$Cov(X_i, X_j) = \frac{K}{N} \cdot \frac{K-N}{N(N-1)}.$$

Denn (sei $i \neq j$) $X_i X_j$ hat genau dann den Wert 1, wenn beide Kugeln Trefferkugeln sind. Die Wahrscheinlichkeit dafür, dass die $i-$ te eine solche ist, beträgt K/N, wie oben eingesehen, und wenn die $i-$ te eine Trefferkugel ist, dann hat man eine bedingte Wahrscheinlichkeit von $(K-1)/(N-1)$ dafür, dass auch X_j den Wert 1 erhält. Also nimmt $X_i X_j$ den Wert 1 genau mit der Wahrscheinlichkeit

$$\frac{K}{N} \cdot \frac{K-1}{N-1}$$

an, und somit ist

$$\begin{aligned} Cov(X_iX_k) &= \mu(X_iX_k) - \mu(X_i)\mu(X_k) \\ &= \frac{K}{N}\cdot\frac{K-1}{N-1} - \frac{K}{N}\cdot\frac{K}{N} \\ &= \frac{K}{N}\cdot\frac{K-N}{N(N-1)}. \end{aligned}$$

Wie zu erwarten, ist sie im allgemeinen negativ! Weiter stellt man fest dass

$$\sigma^2(X_i) = \frac{K}{N}\cdot\frac{N-K}{N-K} \text{ für } 1 \le i \le n,$$

man denke stets an die einfache Bernoulli-Situation mit $p = K/N$.

Insgesamt erhalten wir damit aus (*), da es genau $n(n-1)$ Zahlenpaare (i,j) mit $1 \le i,j \le n$ und $i \ne j$ gibt, dass

$$\begin{aligned} \sigma^2(X) &= n\frac{K}{N}\left(1-\frac{K}{N}\right) + n(n-1)\frac{K}{N}\cdot\frac{K-N}{N(N-1)} \\ &= n\frac{K}{N}\left(\frac{N-K}{N} - (n-1)\frac{N-K}{N(N-1)}\right) \\ &= n\frac{K}{N}\cdot\frac{N-K}{N}\left(1-\frac{n-1}{N-1}\right) \\ &= n\frac{K}{N}\cdot\frac{N-K}{N}\left(\frac{N-1-(n-1)}{N-1}\right) \\ &= n\frac{K}{N}\cdot\frac{N-K}{N}\cdot\frac{N-n}{N-1}. \end{aligned}$$

KAPITEL 4

Vertrauensintervalle und Hypothesentests

Bisher haben wir in gewissen Situationen untersucht, welche Wahrscheinlichkeiten für interessierende Ereignisse bestehen. Es wurde dabei stets vorausgesetzt, dass jeweils die Modellbeschreibung zutraf, somit auch die auf dieser Beschreibung fußende mathematische Wahrscheinlichkeitsberechnung. Außerdem haben wir empirische Verteilungen im Sinne deskriptiver Statistik einfach beschrieben. Beide Enden fassen wir nunmehr zu einer neuen Sichtweise zusammen: Stellen wir uns vor, dass eine empirische Situation zu beschreiben sei, dass wir aber weder ein gültiges mathematisches Modell dafür haben, aus dem einfach zu folgern wäre, noch eine vollständige empirische Datenerfassung, die dann einfach zu komprimieren wäre. Natürlich benötigen wir wenigstens *einige* Daten - sonst werden wir schwerlich begründete Aussagen machen können, aber wir nehmen nunmehr nur an, im Besitze einer Stichprobe aus der Population zu sein, dass für diese Stichprobe die Werte einer Variablen X (oder auch mehrerer Variablen $X, Y, ...$) uns bekannt seien. Es liegt naiv nahe, die Verhältnisse innerhalb der Stichprobe auf die Gesamtpopulation zu übertragen, aber da liegt das Problem: Inwieweit ist diese Übertragung erlaubt, d.h. genau und sicher genug? Das werden im Folgenden die beiden Momente sein, nach denen stets zu fragen ist: Sicherheit und Genauigkeit. Ein Beispiel: Wenn wir von 1000 zufällig ausgewählten Bundesrepublikanern wissen, ob sie Blutgruppe A oder eine andere haben, so wird mit hoher Sicherheit, d.h. Wahrscheinlichkeit der in dieser Stichprobe angetroffene Anteil von Blutgruppe A in einem gewissen (nicht allzu weiten) Abstand vom betreffenden Anteil in der Gesamtbevölkerung sein. Sicherheit und zugleich Genauigkeit werden besser sein können als bei einer Stichprobe von nur 100 Bürgern. Aber wie sicher und wie genau? Gerade dies wollen wir *quantitativ* bestimmen. Das ist zunächst einmal der Beitrag der „Schließenden Statistik" oder Inferenzstatistik, die eben von Stichproben mit gewisser angebbarer Sicherheit und Genauigkeit auf die Population schließt.

Zunächst einmal wollen wir klarstellen, dass man im allgemeinen nur völlig unbrauchbare, weil viel zu ungenaue Aussagen machen kann, wenn man *absolute* Sicherheit verlangt: Haben wir etwa die Daten von 1000 Bundesbürgern und dabei 30% mit Blutgruppe A angetroffen, so wissen wir sicher nur, dass mindestens 300/80000000, also knapp 4 Millionstel der Bundesbürger diese Blutgruppe hat. Verzichten wir dagegen auf absolute Sicherheit und begnügen uns mit einer hohen Wahrscheinlichkeit unserer Aussage, so können wir dagegen eine viel brauchbarere, weil genauere Aussage machen. Im vorliegenden Fall z.B. bekommt man mit den Mitteln dieses Kapitels heraus, dass mit einer Sicherheit von 0.99 oder 99% gilt: Der Anteil der Leute mit Blutgruppe A liegt im Bereich $[0.26; 0.34]$, also zwischen 26 und 34 Prozent. Verlangen wir eine höhere Sicherheit, also etwa 0.999 oder 99.9%, so können wir nur eine ungenauere Aussage treffen, ber immerhin noch sagen: Zwischen 25% und 35%. Das mag wie Zauberei anmuten, da wir doch nur

einen verschwindenden Teil der Population mit unserer Stichprobe untersucht haben, es ist aber reine Wahrscheinlichkeitsrechnung, und die erweist, dass in die Berechnung der Wahrscheinlichkeit oder Sicherheit zu vorgegebener Genauigkeit (bzw. umgekehrt) keineswegs das Verhältnis des Stichprobenumfangs zum Populationsumfang eingeht, sondern allein der absolute Stichprobenumfang entscheidet. Allerdings muss hier angemerkt werden, dass die Gewinnung einer „sauberen" Zufallsstichprobe aus einer Riesenpopulation nicht ganz einfach praktisch durchzuführen ist. Über diesen Weg schlägt das Problem denn doch noch zu, es kann aber immerhin wirksam angegangen werden im Gegensatz zu einer Generalerfassung der Bevölkerung.

Im Beispiel haben wir zu verschiedenen Sicherheiten zwei Vertrauensintervalle angegeben für die Lage eines unbekannten Populationsmittelwertes (nämlich der Bernoulligröße, welche jedem Populationsmitglied den Wert 1 zuordnet, falls es Blutgruppe A hat, sonst den Wert 0 - der Mittelwert μ dieser Größe ist $p =$ relative Häufigkeit der Blutgruppe A in der Bevölkerung). Ähnlich gelagert sind Probleme der Form, dass eine Hypothese H_0 über eine Variable X in einer Population Ω empirisch zu testen ist anhand einer Zufallsstichprobe von $X-$ Werten. - Eine Hypothese ist zunächst einmal eine Aussage, und zwar eine solche, deren Wahrheit fraglich ist (aber möglichst auch interessant!). (Es kann sich natürlich auch um eine Hypothese über eine Beziehung zwischen mehreren Variablen handeln, dann müssten all diese in der Stichprobe erhoben sein.) Man kann dann H_0 nicht direkt und sicher überprüfen - dazu müsste man den überwältigenden Teil der Population kennen, aber man kann aus H_0 folgern, dass ein Ereignis B in der Stichprobe mit hoher Wahrscheinlichkeit zu beobachten sein müsste, und dann die Hypothese H_0 als ihrerseits sehr unwahrscheinlich verwerfen, wenn man in der Stichprobe $\overline{B}$ beobachtet hat. Geben wir auch dazu ein Beispiel, diesmal eines, das wir bereits mit den bisherigen Mitteln rechnen können: Sei H_0 die Hypothese, unter den Wuppertaler Studenten gebe es höchstens 5%, die ihr Studium ausschließlich mit eigener Werktätigkeit finanzieren. In einer Zufallsstichprobe von 10 Studenten fanden Sie aber 3, welche dies tun. In der Stichprobe liegt der Anteil also deutlich höher (0.3), als die Hypothese für die Population sagt (0.05). Können wir das auf die Population übertragen, also die Hypothese verwerfen? Es fragt sich hier - wie immer in diesem Kontext, ob die Abweichung zufällig so hoch sein könnte oder ob man eher annehmen sollte, dass andere Verhältnisse in der Population vorliegen. Eine Abweichung gibt es praktisch immer - in unserem Falle wären genau 5% in der Stichprobe niemals zu beobachten! Stellen wir uns nun auf den intuitiven Standpunkt, dass uns die Beobachtung von 3 Selbstfinanzierern unter 10 Leuten reicht, die Hypothese zu verwerfen. Sicher hätten wir dann auch bei 4,5,...,10 beobachteten Selbstfinanzierern verworfen. Das ist nun das Ereignis $\overline{B}$: Trefferzahl mindestens 3 (unter 10 zufällig Ausgewählten). Welche Wahrscheinlichkeit lässt sich aus der Hypothese H_0 für B folgern? Das ist gemäß Binomialverteilung $\sum_{k=0}^{2} \binom{10}{k} 0.05^k \cdot 0.95^{10-k} = 0.9885$. Setzen wir die Hypothese als wahr voraus, so erhalten wir also eine Wahrscheinlichkeit von fast 99% für B, also von kaum mehr als einem Prozent für $\overline{B}$, man würde also sagen, dass die Beobachtung gegen die Hypothese spreche. Natürlich sind wir nicht absolut *sicher*, dass die Hypothese falsch sein müsse - sie schließt das Beobachtete nicht aus, aber sollte uns gerade so Unwahrscheinliches passiert sein? Man beschreibt die Entscheidung gegen die Hypothese in solchem Falle so: „Die Hypothese wird auf dem Signifikanzniveau 1% (oder Niveau 0.01) verworfen". Das

bedeutet: Eine mindestens so große Abweichung zwischen Beobachtung und Hypothese wie tatsächlich beobachtet hat laut Hypothese eine Wahrscheinlichkeit unter 1%. (Im Beispiel reichte es nicht ganz dafür, sondern nur zum Niveau 0.012.) Das Signifikanzniveau gibt also an, mit welcher Wahrscheinlichkeit folgender sogenannter Fehler erster Art begangen wird: Die Hypothese wird verworfen, ist aber wahr. Genauer handelt es sich um die Wahrscheinlichkeit dafür, dass ein zum Verwerfen führendes Beobachtungsergebnis herauskommt, *wenn die Hypothese wahr ist,* also unter der Bedingung der Hypothese. Wichtig ist die Spezifikation des Ereignisses B vorab; denn sonst könnte man einfach irgendeine Besonderheit des beobachteten Resultats auswählen, die laut Hypothese sehr unwahrscheinlich wäre, aber eben nicht laut wegen des Inhalts der Hypothese, sondern weil sie stets sehr unwahrscheinlich ist. Beispiel: Es soll ein Würfel getestet werden daraufhin, ob er ein Sechstel Wahrscheinlichkeit für Sechsen hat. Unter 600 Würfen beobachtet man genau 100 Sechsen. Das ist aber sehr unwahrscheinlich (obgleich genau der Erwartungswert laut Hypothese), Wahrscheinlichkeit $\binom{600}{100}\left(\frac{1}{6}\right)^{100}\left(\frac{5}{6}\right)^{500} \approx 0.044$. Merke: Das Konkrete ist immer beliebig unwahrscheinlich, es spricht nicht gegen eine allgemeine Hypothese. Im Beispiel würde man allerdings sehen, dass alle denkbaren Alternativen zur Hypothese das Beobachtete noch unwahrscheinlicher machen würden.

Man kann zu allen erdenklichen Sachverhalten Hypothesen aufstellen - natürlich sollte man das vernünftigerweise erst nach gewisser Erfahrung und Kenntnis tun. Nicht alle davon sind statistisch zu prüfen, es gibt auch andere, z.B. die Hypothese, die Erde sei eine Kugel: Diese wurde vor beinahe 2000 Jahren auf raffinierte Weise dadurch geprüft, dass man die Einfallswinkel der Sonnenenstrahlen an zwei 1000 km entfernten Orten maß - man konnte sogar den Erdradius erstaunlich genau damit angeben. Statistisch zu prüfen sind zunächst einmal nur solche Hypothesen, die sich auf Zufallsvariablen beziehen. Aber auch von solchen Hypothesen gibt es eine große Vielfalt, aus der wir nur einige der wichtigsten Beispiele bringen. Ebenso vielfältig ist das Bestimmen von Vertrauensintervallen für unbekannte Parameterwerte von Verteilungen. Wir gehen nun so vor: Zunächst werden diese beiden Aufgabenstellungen abstrakt beschrieben, anschließend werden sie durchgeführt für den einfachsten Fall, in dem es um einen unbekannten Populationsmittelwert $\mu(X)$ geht sowie um den Vergleich zweier Populationsmittelwerte $\mu(X_{|\Omega_1})$ und $\mu(X_{|\Omega_2})$.

1. Abstrakte Beschreibung der Aufgaben

1.1. Das Schätzen eines unbekannten Verteilungsparameters. Man hat eine Variable X und möchte einen Parameter, nennen wir ihn allgemein *par*, der Verteilung von X anhand einer Stichprobe schätzen, etwa $\mu(X)$ oder $\sigma(X)$ oder den Median von X. Dann sucht man sich eine Schätzvariable $\mathcal{S}$, von der ein Wert anhand der Stichprobe berechnet und damit indirekt „beobachtet" werden kann. Mittels der mathematischen Theorie zum wahrscheinlichkeitstheoretischen Verhalten von $\mathcal{S}$ bestimmt man dann ein Vertrauensintervall zu $\mathcal{S}$ der Art: „Der Wert von $\mathcal{S}$ liegt mit Sicherheit w im Bereich $par \pm \varepsilon$". Dann weiß man auch, dass der unbekannte Parameter *par* vom beobachteten Wert $\mathfrak{s}$ von $\mathcal{S}$ nur höchstens den Abstand ε hat, mit eben der Sicherheit w, die eine Wahrscheinlichkeit nahe 1 sein sollte.

1.2. Statistisches Testen einer Hypothese H_0 anhand einer Stichprobe.

1. Man formuliert sorgfältig die zu prüfende Hypothese H_0.

2. Man bestimmt ein Signifikanzniveau α, das ist eine vorgeschriebene kleine Wahrscheinlichkeit, mit der beim nunmehr zu besprechenden Verfahren herauskommt, dass die Hypothese *fälschlich verworfen,* also der Fehler 1. Art begangen wird: Diese Wahrscheinlichkeit ist unter Voraussetzung der Hypothese zu verstehen.
3. Man bestimmt *unter Voraussetzung der Hypothese* ein Vertrauensintervall für eine sogenannte Prüfgröße T, d.h. einen Bereich, in dem ein zufällig beobachteter Wert von T mit der Wahrscheinlichkeit $1-\alpha$ liegt. (Das Ereignis B aus der Einleitung ist dann gerade „$a \leq T \leq b$" oder auch „$a \leq T$", „$T \leq b$". Der Bereich ist also im allgemeinen ein Intervall. Dies Ereignis sagt die Hypothese mit hoher Wahrscheinlichkeit voraus. Die Form des Bereiches richtet sich nach der Form der Hypothese H_0, s.u. unter „Einseitige und zweiseitige Hypothesenformulierung".)
4. Man beobachtet einen Wert t der Testvariablen T (gewöhnlich anhand einer Stichprobe).
5. Entscheidung: Liegt t nicht im Bereich aus 2., also nicht im Vertrauensintervall, sondern im Rest, dem sog. „Verwerfungsbereich", so ist also $\overline{B}$ eingetreten, und die Hypothese H_0 wird auf dem Niveau α verworfen. (Das bedeutet: Man ist recht sicher, dass H_0 falsch sein muss, folglich die Verneinung H_1 von H_0 wahr. Man sagt daher auch *mit Recht*, H_1 werde *angenommen* - eben als wahr angenommen. Liegt t dagegen im Vertrauensintervall, so sage man präzise: Die Beobachtung reicht nicht aus, um die Hypothese H_0 mit hoher Sicherheit $(1-\alpha)$ (bzw. geringer Irrtumswahrscheinlichkeit α) zu verwerfen. *Es ist eine leider immer wieder begegnende komplette Fehlinterpretation,* dies zu verwechseln damit, H_0 sei nunmehr als wahr „anzunehmen", oder gar, H_0 sei daher mit Sicherheit $1-\alpha$ wahr.

Wir werden das Schema ausführlich vor allem im Beispiel der Hypothesen über Populationsmittelwerte ausfüllen. Hier begnügen wir uns damit, ausdrücklich zu begründen, warum die zuletzt genannte Fehlinterpretation wirklich kompletter Unsinn ist: Stellen Sie sich vor, H_0 sei die Hypothese, die mittlere tägliche Fernsehzeit bei Jugendlichen liege bei 2 Stunden. Sie haben in einer Stichprobe ein Mittel von 2.5 Stunden beobachtet, aber das Vertrauensintervall zu 0.99 reiche laut Hypothese bis 2.6 Stunden. Sie können also nicht auf Niveau 0.01 verwerfen. Aber auf dem Niveau 0.05 könnten Sie verwerfen, das Vertrauensintervall dafür reiche nur bis 2.4. Dann hätten Sie immer noch die Sicherheit von 0.95 dafür, dass die Hypothese *falsch sei* und *nicht etwa* eine hohe Sicherheit dafür (gar 0.99), dass die Hypothese *richtig* sei. Genau dies ist aber der Inhalt jener Fehlinterpretation. Auch in der schwächeren Form, man „nehme die Hypothese H_0 an", bleibt sie Unsinn. Man sollte einen intuitiven Begriff von „schwachen" und „starken" Tests haben: Bei kleinem Stichprobenumfang liegt ein schwacher Test vor - fällt eine Hypothese bereits bei einem solchen um, so ist sie ziemlich sicher falsch. Fällt sie aber dabei nicht, so hat sie erst einen sehr schwachen Test bestanden - ein stärkerer könnte sie durchaus noch zu Fall bringen, es gibt noch lange keinen Grund dafür, sie „anzunehmen" oder zu glauben. Besteht eine Hypothese dagegen einen starken Test, so kann man sehr wohl glauben, dass die Hypothese zumindest nicht allzu falsch sein kann.

Das ergibt nur scheinbar ein Dilemma: Häufig werden wir eine Hypothese H im Auge haben, von der wir die Wahrheit zeigen wollen. Bei statistischem Test

von $H_0 = H$ kann man nach dem Vorangehenden nicht so verfahren, dass man nur H_0 nicht verwirft. Stattdessen kann man die Sache oft so angehen: Man bildet die Verneinung „nicht H" und nennt *diese* H_0, steckt sie in einen statistischen Test. Kommt man zum Verwerfen von H_0 auf vernünftigem Niveau, so ist H_0 also ziemlich sicher falsch und H_1 ziemlich sicher wahr. Das Verfahren klappt nicht immer (s.u.), aber im ungünstigen Fall kann man durch einen starken Test unter Kontrolle des Fehlers 2. Art (Hypothese ist falsch - mit spezifizierter Abweichung von der Wahrheit, wird aber fälschlich nicht verworfen) doch zu einem quantifizierten Bestätigungsresultat kommen. Allerdings wird man vielfach einfacher vorgehen können und etwa ein Vertrauensintervall angeben für den fraglichen Verteilungsparameter (ein solcher ist vielfach der Gegenstand einer Hypothese). Hat man etwa für die mittlere Fernsehzeit der Jugendlichen ein sehr sicheres Vertrauensintervall von 2 bis 2.3 Stunden bestimmt, so kann eine Hypothese H des Inhalts, dieser Mittelwert liege bei 2.25, nicht allzu falsch sein, also im Rahmen einer allenfalls anzustrebenden Genauigkeit richtig. Auch eine Angabe von 2.5 wäre noch korrekt, wenn Differenzen unter 0.5 Stunden nicht interessieren.

2. Konkretisierung für den Fall eines unbekannten $\mu(X)$

2.1. Bestimmung eines Vertrauensintervalls für einen unbekannten Populationsmittelwert $\mu(X)$. Die naheliegend zu verwendende Schätzgröße $\mathcal{S}$ ist in diesem Fall $\overline{X}$. Der Grund ist folgender: Man hat $\mu(\overline{X}) = \mu(X)$, und $\sigma(\overline{X}) = \sigma(X)/\sqrt{n}$,wie im vorigen Abschnitt eingesehen. Die Werte von $\overline{X}$ tendieren also zu $\mu(X)$, desto stärker, je größer der Stichprobenumfang n wird. Mit einer Stichprobe $x_1, ..., x_n$ von $X-$ Werten können wir davon den Wert beobachten:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Ebenso wichtig: Die Variable $\overline{X}$ ist selbst bei bescheidenen Stichprobenumfängen bereits näherungsweise normalverteilt, wie aus dem Zentralen Grenzwertsatz folgt, so dass man also die zugehörige Wahrscheinlichkeitsrechnung gut beherrscht, sobald man μ, σ kennt. Daher lautet das zweiseitige Vertrauensintervall zur Sicherheit w für $\mu(X)$:

$$\overline{x} \pm z_{w+(1-w)/2} \cdot \frac{\sigma(X)}{\sqrt{n}} \tag{2.1}$$

Dabei ist z_w zu einer Wahrscheinlichkeit w die Grenze z, für die $\Phi_{0,1}(z) = w$ gilt, also die in der Tabelle zu w abzulesende Grenze. Die Begründung zur Formel: Zunächst haben wir für die Variable $\overline{X}$ das Vertrauensintervall

$$\mu(X) \pm z_{w+(1-w)/2} \cdot \frac{\sigma(X)}{\sqrt{n}}. \tag{2.2}$$

Denn

$$\begin{aligned}
&\Phi_{\mu(X),\frac{\sigma(X)}{\sqrt{n}}}\left(\mu(X)+z_{w+(1-w)/2}\cdot\frac{\sigma(X)}{\sqrt{n}}\right)\\
&-\Phi_{\mu(X),\frac{\sigma(X)}{\sqrt{n}}}\left(\mu(X)-z_{w+(1-w)/2}\cdot\frac{\sigma(X)}{\sqrt{n}}\right)\\
=\;& 1-2\Phi_{\mu(X),\frac{\sigma(X)}{\sqrt{n}}}\left(\mu(X)-z_{w+(1-w)/2}\cdot\frac{\sigma(X)}{\sqrt{n}}\right)\\
=\;& 1-2\Phi_{0,1}(z_{w+(1-w)/2})=w.
\end{aligned}$$

Nunmehr die Überlegung: Formel 2.2 bedeutet, dass ein beliebig „gezogener" Wert von $\overline{X}$ mit Wahrscheinlichkeit w um höchstens den Betrag $z_{w+(1-w)/2}\cdot\frac{\sigma(X)}{\sqrt{n}}$ von $\mu(\overline{X})=\mu(X)$ abweicht. Also weicht auch $\mu(X)$ um höchstens diesen Betrag von dem beobachteten Wert $\overline{x}$ ab, mit der Sicherheit w. Daher Formel 2.1.

Nun hat die gewonnene Lösung des Problems einen Schönheitsfehler: Sie verlangt Eingabe des Wertes $\sigma(X)$, der natürlich ebenso unbekannt ist wie $\mu(X)$. Diese Schwierigkeit kann man auf zwei Weisen überwinden:

- 1. Möglichkeit: Man ersetzt $\sigma(X)$ durch eine obere Schranke $m\geq\sigma(X)$, von der man theoretisch/empirisch weiß, dass $\sigma(X)$ darunter bleiben muss. Allerdings hat man dann ein Vertrauensintervall zu irgendeiner Wahrscheinlichkeit $w'\geq w$ bestimmt. Allerdings hat man sich, um nur Wahrscheinlichkeit w zu behaupten, eventuell eine zu große Ungenauigkeit eingehandelt - die halbe Breite des Vertrauensintervalls heißt nun $z_{w+(1-w)/2}\cdot\frac{m}{\sqrt{n}}$ und ist vielleicht unnötig groß. Ein Beispiel: Sei X eine $p-$ Bernoulli-Größe, mit unbekanntem $\mu(X)=p$. Dann ist $\overline{X}$ die Größe: Relative Häufigkeit der Treffer in der Stichprobe. Nun gilt $\sigma(X)=\sqrt{p(1-p)}$, und das ist stets höchstens $\sqrt{\frac{1}{2}\cdot\frac{1}{2}}=\frac{1}{2}$. In diesem Falle kann man daher $m=1/2$ setzen. Hat man also z.B. in einer Stichprobe vom Umfang 100 genau 30 Treffer gefunden, so erhält man mit $0.3\pm 2.58\frac{1}{2\sqrt{100}}$, also etwa $[0.17;0.47]$ als 99%-Vertrauensintervall.
- 2. Möglichkeit: Man ersetzt den unbekannten Wert $\sigma(X)$ durch den geeigneten Schätzwert

$$s(X):=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}. \tag{2.3}$$

Entsprechend lautet der Schätzwert für die Streuung $\sigma(\overline{X})$, mit der wir letztlich arbeiten:

$$s(\overline{X}):=\frac{s(X)}{\sqrt{n}}. \tag{2.4}$$

Man beachte: Das geht wie die Berechnung der Streuung innerhalb der Stichprobe, diese als Population aufgefasst - anders nur: Der Faktor $1/(n-1)$ anstelle von $1/n$. (Man kann zeigen, dass mit $1/n$ die Streuung systematisch unterschätzt würde.) In diesem Falle ersetzt man also $\sigma(X)$ durch den Wert einer neuen Zufallsgröße $S(X)$, und dieser Wert kann zufällig eben größer oder kleiner als $\sigma(X)$ sein. Aus diesem Grunde ergibt $\overline{x}\pm z_{w+(1-w)/2}s(X)/\sqrt{n}$ kein exaktes Vertrauensintervall mehr zur Wahrscheinlichkeit w, und die Abweichung hängt natürlich von der Sicherheit des Schätzers $S(X)$ ab, die mit

steigendem Stichprobenumfang besser wird. Aber es gibt eine einparametrige Schar von Verteilungen, in der gerade der Parameter n vorkommt und die gerade die Kompensation für das Einsetzen von $s(X)$ für $\sigma(X)$ leistet: Das sind die (wiederum symmetrisch um den Mittelwert 0 liegenden!) $t-$ Verteilungen mit jeweils $1, 2, ...$ Freiheitsgraden, wobei gilt:

$$\textit{Anzahl der Freiheitsgrade} = n - 1.$$

Nun läuft das Verfahren sehr einfach: Man hat mit der $t-$ Tabelle anstelle der $z-$ Tabelle zu arbeiten, also das Resultat:

Das zweiseitige Vertrauensintervall zur Wahrscheinlichkeit w
für unbekannten Mittelwert $\mu(X)$ lautet
$\overline{x} \pm t^{n-1}_{w+(1-w)/2} s(\overline{X})$ (Index $n-1$ für die Freiheitsgrade, auch „df" genannt).

Man wird in einer groben Tabelle ab $n-1 = 100$ bzw. 200 keinen Unterschied zur Standard-Normalverteilung mehr bemerken, dagegen große Unterschiede bei kleinen Stichprobenumfängen n. Für dasselbe Beispiel wir oben hätte man das Vertrauensintervall $0.3 \pm t^{99}_{0.995}\sqrt{\frac{0.3 \cdot 0.7}{99}}$, das ist $0.3 \pm 2.63\sqrt{\frac{0.3 \cdot 0.7}{99}}$, also ungefähr $[0.18, 0.42]$. Man sieht: Der benötigte $t-$ Wert ist mit 2.63 etwas höher als der $z-$Wert 2.58. Aber der Schätzwert $s(X)$ wird kleiner als die obere Schranke $1/2$, und wir erhalten ein etwas günstigeres Vertrauensintervall als oben. Wir haben dabei folgendes nützliche Resultat benutzt, das stets für Bernoulligrößen X gilt:

Für Bernoulligrößen X gilt stets: $s(\overline{X}) = \dfrac{s(X)}{\sqrt{n}} = \sqrt{\dfrac{\overline{x}(1-\overline{x})}{n-1}}$.

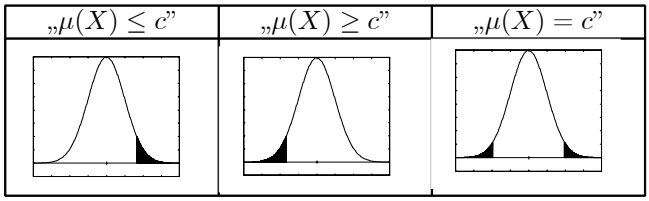
Dabei ist $\overline{x}$ die in der Stichprobe beobachtete relative Häufigkeit der Treffer. (In andern Fällen muss man $s(X)$ aufgrund der beobachteten Einzelwerte ausrechnen.)

2.2. Test einer Hypothese über einen unbekannten Populationsmittelwert $\mu(X)$.

- Formulierungen von H_0:

Einseitig		Zweiseitig
„$\mu(X) \leq c$"	„$\mu(X) \geq c$"	„$\mu(X) = c$"

„c" bedeutet dabei einfach den beliebigen Grenzwert bzw. Wert, der in der jeweiligen Hypothese behauptet wird.
- Vorgabe des Signifikanzniveaus α
- Zugehörige Formen von Vertrauensintervallen zu $1 - \alpha$ und entsprechend Verwerfungsbereichen zu α - die schwarz markierten Teile zeigen die jeweils zum Verwerfungsbereich gehörigen Wahrscheinlichkeiten - stets c in der Mitte:



- Berechnung der Vertrauensintervallgrenzen:

Hypothese	„$\mu(X) \leq c$"	„$\mu(X) \geq c$"	„$\mu(X) = c$"
Vertr.-Int. bei bek. $\sigma(\overline{X})$	$]-\infty, c + z_{1-\alpha} \cdot \sigma(\overline{X})]$	$[c + z_{\alpha} \cdot \sigma(\overline{X}), \infty[$	$c \pm z_{1-\alpha/2} \cdot \sigma(\overline{X})$
gewöhnlich: Vert.-Int. mit $s(\overline{X})$ u. $t-$ Vert.	$]-\infty, c + t^{n-1}_{1-\alpha} \cdot s(\overline{X})]$	$[c + t^{n-1}_{\alpha} \cdot s(\overline{X}), \infty[$	$c \pm t^{n-1}_{1-\alpha/2} \cdot s(\overline{X})$

Man beachte, dass $z_\alpha = -z_{1-\alpha}$ und $t^{n-1}_\alpha = -t^{n-1}_{1-\alpha}$. Ferner sehe man ein, dass $\Phi_{0,1}(z_{1-\alpha/2}) - \Phi_{0,1}(-z_{1-\alpha/2}) = 1 - \alpha$ für $1 - \alpha/2 \geq 0.5$ (analog für $t-$ Verteilung), und α sollte im Kontext eine kleine Wahrscheinlichkeit sein. Erinnerung: $\sigma(\overline{X}) = \sigma(X)/\sqrt{n}$ mit n = Stichprobenumfang. Man stößt auch hier im allgemeinen auf das bereits oben behandelte Problem des unbekannten Wertes von $\sigma(X)$, und man löst es wie oben, arbeitet also bei Einsetzen von $s(\overline{X}) = s(X)/\sqrt{n}$ für $\sigma(\overline{X})$ mit $t^{n-1}_{\dots}$anstelle von $z_{\dots}$. Das steht in der dritten Zeile. Speziell z.B. im Falle einer Bernoulligröße X hat man jedoch den günstigen Umstand, dass die Hypothese über $\mu(X)$ bereits den Wert von $\sigma(X)$ impliziert - dann ist der zu nutzen, und man kann gemäß der zweiten Zeile des Schemas verfahren.

- Entscheidung: Verwerfen auf Niveau α, wenn $\overline{x}$ nicht im Vertrauensbereich, also im Verwerfungsbereich liegt. Wenn dagegen $\overline{x}$ *im Vertrauensintervall* liegt, so sage man *nicht* (s.o.), die Hypothese H_0 werde „angenommen" - darum wird auch das Vertrauensintervall laut Hypothese hier *nicht „Annahmebereich"* genannt.

2.2.1. *Eine Modifikation des Schemas.* Zuweilen zieht man es vor, kein Standard-Signifikanzniveau wie $\alpha = 0.05, 0.01, 0.001$ vorzugeben, sondern umgekehrt danach zu fragen, auf welchem Signifikanzniveau man die Hypothese anhand des Beobachteten gerade noch verwerfen könnte. Man berechnet also nicht die Vertrauensgrenze zu α, sondern den Wert α, welcher zur Beobachtung $\overline{x}$ als Vertrauensgrenze gehört. Dann gibt man das Resultat in der Form „$\alpha = ...$", „$\alpha < ...$" oder „$p = ...$" an (wobei *in diesem Kontext* „p" dasselbe wie „α" bedeutet) - so z.B. in vielen Publikationen und auch Computerausdrucken von Statistikprogrammen.

2.2.2. *Beispiele.*

1. X sei die Variable „Intelligenzquotient" (im Sinne einer bestimmten Messung) in einer speziellen Population von Kindern, H_0 sei die Hypothese „$\mu(X) \geq 110$", beobachtet habe man in einer Stichprobe vom Umfang 100 aus jener Population $\overline{x} = 106$ und $s(X) = 10$.
   (a) Erste Version: Test von H_0 auf vorgegebenem Niveau $\alpha = 0.01$. Man errechnet mittels der benötigten Zahl $t^{99}_{0.01} = -2.365$ die Vertrauensgrenze zu.—

$$110 - 2.365\frac{10}{\sqrt{100}} = 107.64$$

   und verwirft also H_0 auf dem Niveau $\alpha = 0.01$. Entsprechend ist man recht sicher, dass gilt: $\mu(X) < 110$.

(b) Zweite Version: Angabe des besten Niveaus, auf dem man H_0 mittels der Beobachtung gerade noch verwerfen kann. Man löst die Gleichung

$$110 + t\frac{10}{\sqrt{100}} = 106$$

mit dem Resultat $t = -4$ und fragt nunmehr nach der Wahrscheinlichkeit, die links von diesem Wert sitzt. Zu $t^{99} = -4$ gehört eine Wahrscheinlichkeit $\alpha < 0.0000612$, man kann also auf dem (sehr guten) Niveau 0.0000612 die Hypothese H_0 verwerfen und ist in der Tat *sehr sicher*, dass $\mu(X) < 110$. Dies hätte man bei der ersten Version nicht so genau gesehen, sondern nur grob geahnt, dass man das Niveau 0.01 bei weitem übertreffen könnte.

2. H_0 sei die Hypothese, dass in einer gewissen Population eine Eigenschaft mit relativer Häufigkeit 0.3 anzutreffen sei, also H_0 : „$\mu(X) = 0.3$". Hier ist X die Bernoulligröße, welche jedem Populationsmitglied mit der besagten Eigenschaft die Zahl Eins zuordnet, jedem sonstigen die Zahl Null. Man habe in einer Stichprobe vom Umfang 50 genau 25 mit jener Eigenschaft angetroffen. Das bedeutet: $\overline{x} = 0.5$. Man beachte: Die Hypothese impliziert, dass $\sigma(X) = \sqrt{0.3 \cdot 0.7}$, also $\sigma(\overline{X}) = \sqrt{0.3 \cdot 0.7}/100$. Somit haben wir hier *nicht* mit $t-$ Verteilung zu arbeiten, sondern mit (laut Hypothese) korrekter Streuung, also mit Normalverteilung, da wir für die Berechnung die Hypothese zur Voraussetzung nehmen. Wir besprechen wieder die oben beschriebenen Versionen, woraus in diesem Falle jedoch 4 Versionen werden, da wir einmal mit der Variablen der relativen Stichprobenhäufigkeit, das andere Mal mit der binomialverteilten Variablen der absoluten Stichprobenhäufigkeit arbeiten werden - letzteres Verfahren hat den Vorzug, dass man die Rechnung mit Näherung durch Normalverteilung mittels Stetigkeitskorrektur genauer hinbekommt.

(a) Erste Version: $\alpha = 0.05$. Wir errechnen das zweiseitige Vertrauensintervall

$$0.3 \pm 1.96\frac{\sqrt{0.3 \cdot 0.7}}{\sqrt{50}}, \text{ also } [0.17; 0.43].$$

Der beobachtete Wert 0.5 fällt deutlich heraus, wir verwerfen also auf dem Niveau 0.05.

(b) Zweite Version: Wir bestimmen das Signifikanzniveau, auf dem wir verwerfen können, lösen also die Gleichung

$$0.3 + z\frac{\sqrt{0.3 \cdot 0.7}}{\sqrt{50}} = 0.5$$

und finden $z = 3.0861$. Das gesuchte Niveau ist doppelt so groß wie die Wahrscheinlichkeit, die auf den Bereich oberhalb von z entfällt (bei der Standard-Normalverteilung). Also

$$\alpha = 2\Phi_{0,1}(-3.0861) \approx 0.00203.$$

Man kann also auf einem Niveau nahe 2/1000 verwerfen.

(c) Dritte Version: Wie die erste, nur mit Binomialverteilung: Sei Y die Variable „Trefferzahl" auf der Population aller Stichproben des Umfangs 50. Sie ist laut Hypothese $(50, 0.3)-$ binomialverteilt. Mit

$\alpha = 0.05$ (analog zu a.) ergibt sich das Vertrauensintervall für Y laut Hypothese:

$$15 \pm 1.96\sqrt{50 \cdot 0.3 \cdot 0.7}, \text{ also } [8.6489; 21.351].$$

Mit Stetigkeitskorrekur ist die Obergrenze als 21 anzusetzen (hätte man einen Wert *über 21.5* errechnet, so wäre mit dieser Korrektur 22 anzusetzen gewesen). Der Befund lautet wiederum, dass auf dem Niveau $\alpha = 0.05$ zu verwerfen ist.

(d) Vierte Version: Wie die zweite, wieder jedoch mit Binomialverteilung. Das Niveau α ist die Wahrscheinlichkeit dafür, mindestens 25 oder höchstens 5 Treffer zu bekommen bei einer Binomialverteilung mit $n = 50, p = 0.3$. Näherung durch Normalverteilung ergibt dafür *mit Stetigkeitskorrektur*:

$$\begin{aligned} \alpha &= \Phi_{15;\sqrt{50\cdot 0.3\cdot 0.7}}(5.5) + 1 - \Phi_{15;\sqrt{50\cdot 0.3\cdot 0.7}}(24.5) \\ &= 2\Phi_{15;\sqrt{50\cdot 0.3\cdot 0.7}}(5.5) = 2\Phi_{0,1}\left(\frac{5.5-15}{\sqrt{50\cdot 0.3\cdot 0.7}}\right) \approx 0.0034. \end{aligned}$$

Diese Wahrscheinlichkeit ist zwar auch klein, aber deutlich höher (und korrekter!) als die in der zweiten Version (b.) gegebene. (Wenn man hier ohne Stetigkeitskorrektur mit den Grenzen $5, 25$ arbeiten würde, käme genau dasselbe Resultat wie in b.)

Bemerkung: Man hat bei der Variante, die aus der Beobachtung das erreichbare Niveau berechnet, grundsätzlich die Möglichkeit, die Vertrauensintervallform so zu ändern, dass man t bzw. z als Unbekannte setzt (so in b.) und danach auflöst, oder aber die Wahrscheinlichkeit des Verwerfungsbereichs direkt zu berechnen (so in d.) - beide ergeben dasselbe Resultat.

Bemerkung: Feinere Wertetabellen als die üblichen zur Normalverteilung und $t-$ Verteilung erhält man besonders bequem mit dem Computer - davon wurde in den Beispielen oben freier Gebrauch gemacht.

2.3. Test einer Hypothese über die Differenz zweier unbekannter Populationsmittelwerte. Vielfach interessiert die Frage, ob eine Variable X in einer Teilpopulation A einen höheren / niedrigeren / gleichen Mittelwert wie in einer anderen Teilpopulation B habe. Genauer: Die Variable X sei auf Ω definiert, A und B disjunkte Klassen von Ω, also $A, B \subseteq \Omega$ und $A \cap B = \varnothing$. (Nicht notwendig $A \cup B = \Omega$.) Dann betrachten wir die *verschiedenen Variablen* $X_{|A}$ und $X_{|B}$, das sind die Einschränkungen von X jeweils auf die Definitionsbereiche A, B, also z.B. $X_{|A} : A \to \Omega$, $a \mapsto X(a)$. Man denke etwa an eine medizinisch relevante Variable X und an A, B als Teilpopulationen von Kranken verschiedener Krankheiten usw. Nun möchte man eine Hypothese der Form „$\mu(X_{|A}) \leq [=, \geq] \mu(X_{|B})$" testen, in diesem Sinne einen Mittelwertvergleich anstellen. Der empirische Hintergrund sollte ausgemacht werden von einer Stichprobe eines Umfangs n_A von Messungen in A und einer unabhängige Stichprobe eines (nicht notwenig gleichen) Umfangs n_B aus B. (Die Unabhängigkeit ergibt sich hier normalerweise aus der Grundsituation, dass $A \cap B = \varnothing$. Anders liegt die Sache bei sogenannten „verbundenen" Stichproben, wenn man etwa dieselben Menschen zu verschiedenen Zeitpunkten beobachtet, vor und nach einem Training z.B. Jedenfalls wollen wir so vorsichtig sein, diese Unabhängigkeit der Stichproben ausdrücklich zu fordern.)

Die Lösung des Problems besteht in der Zurückführung auf den Fall einer einzigen Variablen, der oben bereits behandelt wurde. Allerdings ergibt sich noch ein gesondertes kleines technisches Problem mit der Streuungsschätzung.

Die Zurückführung: Man nutzt die Gleichwertigkeit von z.B. „$\mu(X) \leq \mu(Y)$" mit „$\mu(X) - \mu(Y) \leq 0$", dann $\mu(X) - \mu(Y) = \mu(X - Y)$, schließlich $\mu(X) = \mu(\overline{X})$ und erhält - den dritten Fall kann man sich schenken, da man nur A und B zu vertauschen braucht, um erstere Version wieder zu erhalten:

	Hypothese einseitig	Hypothese zweiseitig
Hypothese	„$\mu(X_{\mid A}) \leq \mu(X_{\mid B})$"	„$\mu(X_{\mid A}) = \mu(X_{\mid B})$"
gleichwertige Umformulierung	„$\mu(\overline{X_{\mid A}} - \overline{X_{\mid B}}) \leq 0$"	„$\mu(\overline{X_{\mid A}} - \overline{X_{\mid B}}) = 0$"

Setzen wir nun $Y := \overline{X_{\mid A}} - \overline{X_{\mid B}}$, so haben wir die Vergleichs-Hypothesen auf eine über Y zurückgeführt, welche die Form einer Hypothese über einen einzelnen Mittelwert hat. Man beachte jedoch, dass wir zu Y nur einen einzigen Wert beobachtet haben, der sich allerdings aus Mittelwerten zusammensetzt - Y wird daher bereits kleine Streuung haben.

Bemerkung zu sinnvoller Verallgemeinerung der Hypothesenformulierung: Sachlich ist es in aller Regel völlig uninteressant, etwa statistisch festzustellen, eine Studentenpopulation erbringe höhere Leistungen als eine andere. Vielmehr wird man auf eine Aussage der Form hinauswollen: „Gruppe B ist im Mittel um mehr als die Notendifferenz c im Mittel schlechter als Gruppe A", wobei c eine sachlich interessierende Differenz bedeutet. Diese Aussage mit gewisser Sicherheit behaupten zu können, das heißt, die Verneinung H_0: „Gruppe B ist im Mittel um höchstens c besser als A" auf einem guten Niveau α verwerfen zu können. H_0 bedeutet, wenn wir mit X die Variable „Note" bezeichnen: „$\mu(X_{\mid A}) - \mu(X_{\mid B}) \leq c$", also „$\mu(Y) \leq c$" ($c > 0$, und geringere Notenwerte sind die besseren). Dies bedeutet nun keinerlei neues technisches Problem, es macht nichts aus, ob eine Hypothese „$\mu(Y) \leq 0$" oder „$\mu(Y) \leq c$" mit einem beliebigen Wert c lautet. Es ist eine furchtbare Krankheit, diesen simplen Sachverhalt nicht zu sehen und dann lauter sachlich nichtssagende Hypothesen zu testen, wozu auch die unselige Bezeichnung von H_0 als „Nullhypothese" verführen mag. Weitere Verwirrung wird dadurch gestiftet, dass „signifikant" - auf Deutsch: „bedeutsam" *zwei verschiedene Bedeutungen* in unserem Kontext besitzt, einmal: „statistisch bedeutsam" im Sinne dessen, dass eine beobachtete Abweichung mit hoher Wahrscheinlichkeit nicht auf Zufall beruht, sodann „inhaltlich bedeutsame" Differenz. Es ist eine Unsitte, lediglich auf die erstere Bedeutung zu sehen und gar noch zu meinen, sie schließe letztere ein: Man kann völlig nichtssagende Behauptungen mit hoher statistischer Signifikanz versehen, aber man sollte sich selbstverständlich bemühen, inhaltlich wichtige Aussagen statistisch zu erhärten. Wir formulieren daher die benötigte Verallgemeinerung noch einmal ausdrücklich, die insbesondere für einseitige Fragestellungen von Bedeutung ist und die eben den Einbau inhaltlicher Bedeutsamkeit erlaubt:

	Hypothese (einseitig)
Hypothese	„$\mu(X_{\mid A}) \leq \mu(X_{\mid B}) + c$"
gleichwertige Umformulierung	„$\mu(\overline{X_{\mid A}} - \overline{X_{\mid B}}) \leq c$"

Grundlegend für das Rechnen sind nun folgende Feststellungen:

- $Y = \overline{X_{|A}} - \overline{X_{|B}}$ ist näherungsweise normalverteilt (wegen des Satzes, dass Summen unabhängiger normalverteilter Variablen normalverteilt sind), da $\overline{X_{|A}}$ und $\overline{X_{|B}}$ unabhängig und (zumindest näherungsweise) normalverteilt sind.
- $$\begin{aligned} \mu(Y) &= \mu\left(X_{|A}\right) - \mu\left(X_{|B}\right). \\ \sigma(Y) &= \sqrt{\frac{\sigma^2\left(X_{|A}\right)}{n_A} + \frac{\sigma^2\left(X_{|B}\right)}{n_B}}. \end{aligned}$$

  Dies folgt sofort aus der Linearität von μ und den Varianzformeln für $\overline{X}$ und für (linear) unabhängige Summen. Weiter folgt, dass man mit den Stichproben folgenden Schätzwert für $\sigma(Y)$ bei (normalerweise) unbekannten $\sigma\left(X_{|A}\right)$ und $\sigma\left(X_{|B}\right)$ hat:

  $$s(Y) = \sqrt{\frac{s^2\left(X_{|A}\right)}{n_A} + \frac{s^2\left(X_{|B}\right)}{n_B}},$$

  mit den oben eingeführten Schätzwerten für $\sigma^2\left(X_{|A}\right)$ und $\sigma^2\left(X_{|B}\right)$.

Bemerkung: In der Literatur findet man - leider - vielfach statt des letzteren $s(Y)$ einen anderen Streuungsschätzer, der nach bestandenem Test der Hypothese „$\sigma\left(X_{|A}\right) = \sigma\left(X_{|B}\right)$" durch Zusammenwerfen der Stichproben für eine gemeinsame Streuungsschätzung produziert wird. Anschließend wird ein $t-$ Test mit Einsetzen dieses Schätzers durchgeführt, analog dem oben beschriebenen für einfache Stichproben. Dies Verfahren ist jedoch nicht nur logisch völlig unsauber (man nimmt die Hypothese „$\sigma\left(X_{|A}\right) = \sigma\left(X_{|B}\right)$" als wahr an, nur weil man sie nicht hat verwerfen können), es liefert auch gerade in den Fällen kleinerer Stichprobenumfänge häufig völlig falsche Resultate (wie man mittels kleiner Computerexperimente leicht nachweist), wenn die Streuungen tatsächlich verschieden sind - und damit ist durchaus zu rechnen; es ist z.B. ganz typisch, dass eine Teilpopulation mit einem höheren $X-$ Mittelwert auch eine höhere Streuung aufweist. Daher werden wir dies ganze Verfahren mit keiner Silbe weiter beschreiben, das vielfach noch gar als einzige Methode („der $t-$ Test") erwähnt wird (!) und kompliziert, schlecht und völlig überflüssig ist. Wir werden stattdessen eine sehr bequeme und etwas ungenauere sowie eine immer noch recht bequeme und wirklich genaue Lösung des Problems angeben, die beide durchaus in guter Literatur bekannt sind.

2.3.1. *Erstes (ungenaueres) Verfahren: Anwenden der Normalverteilung.* Dies Verfahren zeigt brauchbare (bei hohen Stichprobenumfängen ausgezeichnete) Resultate, sobald die Stichprobenumfänge nicht zu klein sind (beide über 40 als grobe Faustregel). Zum Test der Hypothese, die wir stets in der Form „$\mu(Y) \leq [=]0$" formulieren können, auf Niveau α hat man folgende Vertrauensintervalle:

Hypothese	„$\mu(Y) \leq c$"	„$\mu(Y) = c$"
Vertrauensintervall	$c + z_{1-\alpha} s(Y)$	$c \pm z_{1-\alpha/2} s(Y)$

Wie zuvor ist auf Niveau α zu verwerfen, wenn der beobachtete Wert $y = \overline{x_A} - \overline{x_B}$ der Mittelwertdifferenzen außerhalb des betreffenden Vertrauensintervalls liegt. (Ebenso kann man wieder die entsprechende Gleichung nach z auflösen, um das Niveau zu bestimmen, auf dem man bei gegebener Beobachtung verwerfen könnte.)

Beispiel: H_0: „Gruppe A von Jugendlichen sieht im Mittel höchstens 1/2 Stunde täglich länger fern als Gruppe B" ($A \cap B = \varnothing$). Bezeichnen wir mit X

die Variable der mittleren täglichen Fernsehzeit für jede Person. Nun möge man beobachtet haben: $\overline{x_A} = 2.5$ Stunden, $s_A = 3/4$ Stunde, bei $n_A = 80$, $\overline{x_B} = 1.8$ Stunden, $s_B = 1/3$ Stunde, bei $n_B = 100$. Wir berechnen

$$s(Y) = s(\overline{X_{|A}} - \overline{X_{|B}}) = \sqrt{\frac{9/16}{80} + \frac{1/9}{100}} = 0.09$$

und erhalten die Gleichung

$$\frac{1}{2} + z\sqrt{\frac{9/16}{80} + \frac{1/9}{100}} = 0.7 (= \overline{x_A} - \overline{x_B}).$$

Das ergibt $z = 2.216\,4$, und $\alpha = \Phi_{0,1}(-2.21649) = 0.0133$, man könnte also auf 5%-Niveau verwerfen, hat jedoch das 1%-Niveau knapp verfehlt. Das bedeutet: Man ist recht sicher, dass Gruppe A im Mittel täglich mehr als 1/2 Stunde länger fernsieht als Gruppe B.

2.3.2. *Genaueres Verfahren: Anwenden der $t-$ Verteilungen.* Es ist tatsächlich kaum etwas zu ändern, lediglich ist eine etwas verwickelte Formel für die korrekte df = Freiheitsgrade zu verwenden, die im allgemeinen keine ganze Zahl mehr ergibt. Natürlich macht das kein Problem: Die zugehörigen $t-$ Verteilungen existieren, und man kann bei Benutzung von Tabellen die nächstkleinere ganze Zahl von Freiheitsgraden nehmen. Mit dem Computer hat man keinerlei Mehraufwand, da ein ordentliches Programm diese verallgemeinerten $t-$ Verteilungen besitzt.

Hier die Formel zur Berechnung der Freiheitsgrade; dabei bedeuten $n_{A,B}$ wie oben eingeführt die Stichprobenumfänge und sind mit $s_{A,B}$ die Streuungsschätzungen $s\left(X_{|A}\right), s\left(X_{|B}\right)$ abgekürzt - man beachte, dass man ohnehin $n_A, n_B > 1$ haben muss, weil es sonst keine Streuungsschätzungen $s_{A,B}$ gibt:

$$df = \frac{\left(s_A^2/n_A + s_B^2/n_B\right)^2}{\left(s_A^2/n_A\right)^2/(n_A - 1) + \left(s_B^2/n_B\right)^2/(n_B - 1)}.$$

Die Formel bewirkt, dass für $n_A = n_B$ und $s_A = s_B$ herauskommt: $df = n_A + n_B - 2$. Für sehr verschiedene Umfänge oder auch allein bei sehr verschiedenen $s_{A,B}$ kommt man dagegen kaum über den geringeren der beiden Stichprobenumfänge hinaus. Mit der so ausgerechneten Zahl df hat man dann folgende Vertrauensintervalle:

Hypothese	„$\mu(Y) \leq c$"	„$\mu(Y) = c$"
Vertrauensintervall	$c + t^{df}_{1-\alpha} s(Y)$	$c \pm t^{df}_{1-\alpha/2} s(Y)$

.

Beispiel: Wir behandeln dasselbe Fernsehdauer-Beispiel wie oben mit der verfeinerten Methode: Zusätzlich haben wir nur zu berechnen:

$$df = \frac{(9/(16 \cdot 80) + 1/(9 \cdot 100))^2}{(9/(16 \cdot 80))^2/79 + (1/(9 \cdot 100))^2/99} = 103.87.$$

Nunmehr ist die Wahrscheinlichkeit rechtsseitig von $t^{103.87} = 2.2164$ (die Gleichung bleibt dieselbe!) bzw. linksseitig von -2.2164 zur $t-$ Verteilung mit 103.87 Freiheitsgraden zu bestimmen, das ist etwa 0.144. Wie für solche Stichprobenumfänge versprochen, ist der Unterschied zum Resultat mit der einfachen Methode klein (das war 0.133). Aber dies (etwas ungünstigere) Ergebnis ist korrekter, und bei wesentlich kleineren Stichprobenumfängen können die Unterschiede drastisch werden.

KAPITEL 5

Reelle Funktionen

1. Elementare mathematische Funktionen und ihre Graphen

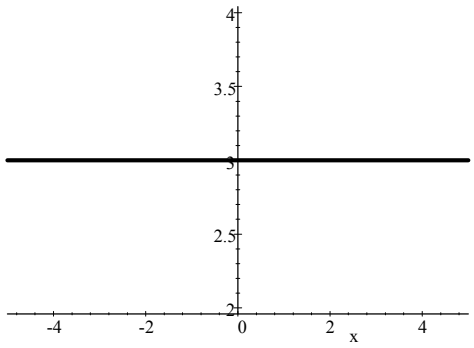
In diesem Abschnitt wollen wir einerseits einen Grundbaukasten zur Bildung von Funktionen und Funktionsgraphen vorstellen und dabei noch nicht die Ableitung benutzen. Andererseits sollte allgemein das Lesen von Graphen gestärkt werden. Bei der Beschreibung inhaltlicher Zusammenhänge spielen Funktionen und ihre Graphen mittlerweile in den meisten Wissenschaften eine hervorragende Rolle. Es handelt sich einfach um eines der nächstliegenden Mittel, Komplexes auf einfache Weise zu sagen. Hinzu kommt noch die Bedeutung der Funktionen für die theoretische Seite: Soll ein Zusammenhang aus Grundprinzipien hergeleitet werden, so ist die explizite mathematische Handhabung von Funktionen unerlässlich.

1.1. Einige Grundfunktionen. Grundfunktionen bilden eine Art von „Atomen" für den aufzubauenden Bereich von nützlichen Funktionen. Mit Verknüpfungen werden daraus kompliziertere gebaut, entsprechend werden die graphischen Eigenschaften verknüpft. Die Eigenschaften der folgenden immer wieder wichtigen Funktionen sollten daher gut bekannt sein und behalten werden. Es sei bemerkt, dass man für weiterführende Zwecke stets neue Grundfunktionen bereitstellen kann, z.B. findet man mehr in einer ausführlicheren Formelsammlung bzw. einem Kompendium in Taschenbuchform.

- Konstanten

$$\begin{array}{rccc} f_c: & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & c \end{array}$$

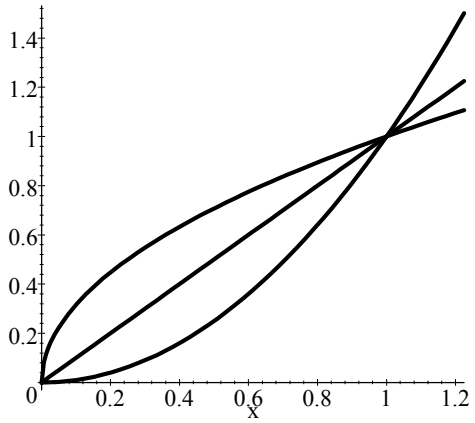
Die Graphen sehen sehr einfach aus, es sind Parallelen zur x-Achse, hier z.B. zu $c = 3$:



- Potenzfunktionen: Für $a \in \mathbb{R}$, $a > 0$, definiert man die *Potenzfunktion zum festen Exponenten* a (mit $\mathbb{R}_{\geq 0}$ bezeichnen wir kurz die Menge $\{x \in \mathbb{R} \mid x \geq 0\}$):

$$\begin{array}{rccc} g_a: & \mathbb{R}_{\geq 0} & \to & \mathbb{R}_{\geq 0} \\ & x & \mapsto & x^a \end{array}$$

Die Graphen sehen so aus, hier speziell zu $a = 2$, $a = 1/2$, dazu die Gerade $y = x$ für $a = 1$:



Man beachte: g_a ist stets Umkehrfunktion zu $g_{\left(\frac{1}{a}\right)}$, d.h. $g_a(g_{\left(\frac{1}{a}\right)}(x)) = x$, und die beiden Graphen entstehen durch Spiegelung an der Geraden $y = x$ auseinander.

Stets erscheint für $a > 1$ qualitativ etwas Ähnliches, eine Kurve mit wachsender Steigung, ebenso für $a < 1$ eine Kurve mit fallender Steigung.

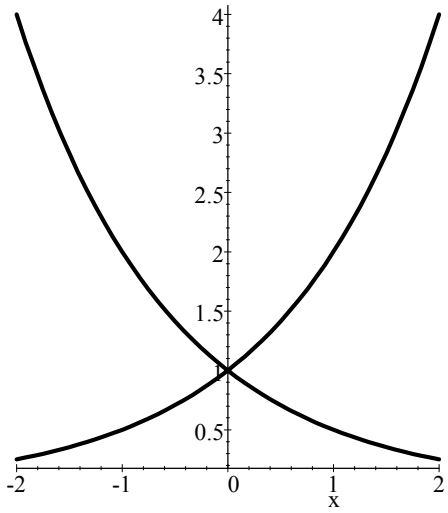
- Exponential- und Logarithmusfunktionen:

Zu jeder Basis $a > 0, a \neq 1$ definiert man die Exponentialfunktion zur Basis a:

$$\begin{array}{rccc} \exp_a: & \mathbb{R} & \to & \mathbb{R}_{>0} \\ & x & \mapsto & a^x \end{array}$$

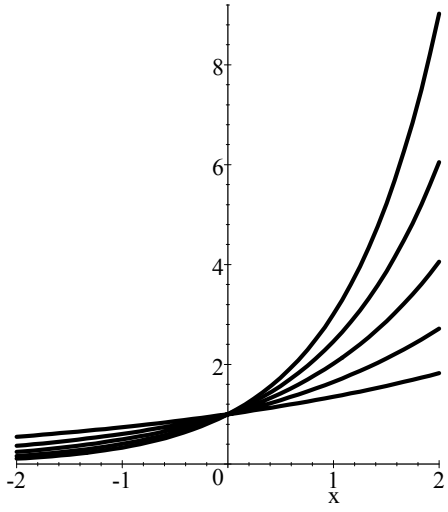
Hier ist die Basis fest, der Exponent die unabhängige Variable, gerade umgekehrt wie bei den Potenzfunktionen! Die Exponentialfunktionen wachsen (für $a > 1$) viel schneller als die Potenzfunktionen.

Man beachte, dass nur Werte > 0 herauskommen. Die Graphen sehen so aus, hier für $a = 2, a = 1/2$:



Man beachte, dass sich für $a > 1, a < 1$ stets dieselben qualitativen Bilder ergeben und dass $\exp_a(x) = \exp_{\left(\frac{1}{a}\right)}(-x)$ für alle $x \in \mathbb{R}$, so dass die Graphen durch Spiegelung an der $y-$ Achse auseinander hervorgehen. Wesentlich ist nun die folgende Beobachtung: $\exp_a$ ist für $a > 1$ eine extrem schnell wachsende Funktion, mit $x \to \infty$ geht sie sehr schnell nach ∞, mit $x \to -\infty$ sehr schnell nach Null. (Man denke an: $2^{10} = 1024$, $2^{20} = 1.0486 \times 10^6$, $2^{-10} = 1/1024$, $2^{-20} = 1/2^{20} = 9.5367 \times 10^{-7}$.)

Nun kann man einmal die Schar aller Exponentialfunktionen $\exp_a$ mit $a > 1$ betrachten, hier sind einige aufgezeichnet:



Alle steigen monoton (auch die im gezeichneten Bereich langsamen werden schließlich beliebig steil!), doch ist die Steigung kontinuierlich veränderbar durch

Variation des Parameters a. Nun kann man sich vorstellen, dass es genau einen Parameterwert gibt, der an der Stelle $x = 0$ die Graphensteigung (Tangentensteigung) 1 ergibt. Wählt man a zu groß, wird die Steigung an dieser Stelle größer, mit zu kleinem a eben kleiner. Der spezielle Wert liegt bei ≈ 2.71 und wird e genannt. Also:

Die Zahl e wird definiert durch die Eigenschaft:

$$\exp_e'(0) = \exp_e(0) = 1.$$

Übrigens folgt sogleich daraus, wie wir später einsehen werden:

$$\exp_e'(x) = \exp_e(x) \text{ für } \textit{alle } x \in \mathbb{R}.$$

Wegen dieser einfachen Eigenschaft nennt man die Basis e die „natürliche", und man bezeichnet die zugehörige Exponentialfunktion $\exp_e$ kurz nur mit „exp".

Die Exponentialfunktionen sind alle streng monoton, folglich umkehrbar, und hier sind die Umkehrfunktionen, wiederum für alle $a > 0$, $a \neq 1$:

$$\begin{array}{rccl} \log_a: & \mathbb{R}_{>0} & \to & \mathbb{R} \\ & x & \mapsto & \log_a(x) = \textit{die eindeutige Lösung } y \textit{ von } a^y = x. \end{array}$$

Es sollte klar sein, dass alle Logarithmusfunktionen nur im Bereich $x > 0$ definiert sind.

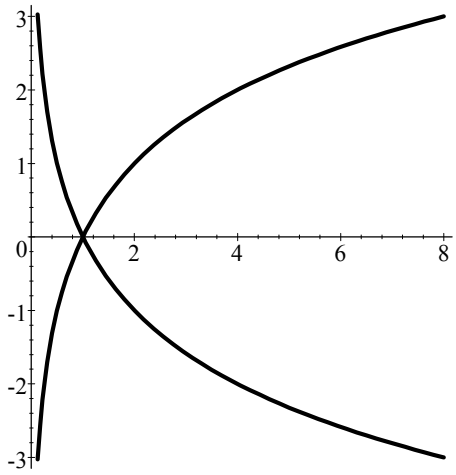
Man beachte den logischen Trick, mit dem man verbal zu jeder umkehrbaren Funktion f die Umkehrfunktion f^{-1} definieren kann:

$$f^{-1}(x) = \textit{dasjenige } y, \textit{ so dass } f(y) = x \textit{ gilt}.$$

(Mehr zu den Umkehrfunktionen im nächsten Abschnitt.)

Man hebt die Logarithmusfunktion $\log_e$ zur natürlichen Basis e heraus und nennt sie kurz „ln" („logarithmus naturalis", „natürlicher Logarithmus"). (Mit „ln" ist immer dies gemeint. Bei der Bezeichnung „log" dagegen wissen Sie nie (außer durch Kontext), woran Sie sind, es könnte ln sein (Computer), aber auch $\log_{10}$ (Feld-, Wald- und Wiesen- Bezeichnung), oder auch in informationstheoretischem Zusammenhang $\log_2$.)

Natürlich entstehen die Graphen der Logarithmusfunktionen aus denen der zugehörigen Exponentialfunktionen durch Spiegelung an der Geraden $y = x$. Hier sind die Graphen zu $\log_2$, $\log_{\left(\frac{1}{2}\right)}$zur Illustration, ersterer ist die steigende Kurve:



Diese beiden sind spiegelbildlich bezüglich der x-Achse, was aus $\log_a(x) = -\log_{(\frac{1}{a})}(x)$ oder auch direkt aus der Symmetrie der Graphen der zugehörigen Exponentialfunktionen bezüglich der y-Achse folgt. Bemerkenswert ist der Pol bei $x = 0$: Dort gehen die Funktionswerte nach $-\infty$ bzw. ∞. Er entspricht der Asymptote $x = 0$ für die Exponentialfunktionen.

Das Rechnen mit den Exponentialfunktionen und Logarithmusfunktionen ebenso wie das mit den Potenzfunktionen wird - naheliegend - beherrscht von den Regeln des Rechnens mit Potenzen. Aber es wird nützlich sein, die resultierenden Formeln für die Exponentialfunktion exp und die Logarithmusfunktion ln (ohne Verlust beschränken wir uns auf die natürliche Basis, das vereinfacht die Schreibweise, und beim Rechnen kommt man voll und ganz damit aus) noch einmal zusammenzustellen, zumal da die Konsequenzen für ln dem Anfänger nicht ohne weiteres klar sind:

Formeln für exp und entsprechend für ln

Stets sind x, y beliebige reelle Zahlen, a, b beliebige Zahlen > 0, $d > 0$, $d \neq 1$

$e^0 = \exp(0) = 1$	$\ln(1) = 0$
$e^x \cdot e^y = e^{x+y}$	$\ln(a \cdot b) = \ln(a) + \ln(b)$
$(e^x)^y = e^{x \cdot y}$	$\ln(a^y) = y \cdot \ln(a)$
$d^x = e^{x \cdot \ln(d)}$	$\log_d(b) = \frac{\ln(b)}{\ln(d)}$

Die ersten drei Formeln kennen wir vom Potenzrechnen, man sollte nur einsehen, dass auf der rechten Seite lediglich durch Logarithmen völlig Gleichwertiges zur linken Seite ausgedrückt wird; nehmen wir z.B. die dritte Formel, $\ln(a^y) = y \cdot \ln(a)$. Mit $a > 0$ haben wir eine (eindeutige) Zahl x mit $a = e^x$. Somit ist $\ln(a) = x$, die rechte Seite der Gleichung also $y \cdot x$. Weiter ist (die linke Seite der Gleichung) $\ln(a^y) = \ln((e^x)^y) = \ln(e^{x \cdot y})$, gemäß der exp-Formel auf der linken Seite. Aber nach Definition von ln haben wir $\ln(e^{x \cdot y}) = x \cdot y$. Somit kommt auf beiden Seiten der Gleichung (stets) dasselbe Resultat, die Gleichung gilt.

1.2. Das Verknüpfen von Funktionen zu neuen Funktionen. Ausgehend von unseren Grundfunktionen können wir mit den hier einzuführenden Funktionsbildungsprozessen einen riesigen und für fast alle praktischen Zwecke ausreichenden Funktionenvorrat beschaffen. (Für manche weitergehenden Zwecke würde man noch ein paar weitere Grundfunktionen benötigen, nennen wir die trigonometrischen.)

Die Funktionsbildungsprozesse (oder Operationen, die aus Funktionen neue Funktionen machen):

- Addition, Subtraktion, Multiplikation und Division von Funktionen:

Seien $f : A \to \mathbb{R}$ und $g : A \to \mathbb{R}$ zwei Funktionen, $A \subseteq \mathbb{R}$. Dann ist definiert:

$$\begin{array}{rccc} f+g: & A & \to & \mathbb{R} \\ & x & \mapsto & f(x)+g(x) \end{array}$$

Also: $f+g$ ist eine neue Funktion, und zwar ist nach Definition: $(f+g)(x) = f(x)+g(x)$. Völlig analog definiert man $f-g$ und $f \cdot g$. Auch $\frac{f}{g}$ ist ähnlich, allerdings darf dann g nirgends den Wert 0 annehmen, andernfalls muss man den Definitionsbereich von $\frac{f}{g}$ weiter einschränken. Kurzum kann man nicht annehmen, dass $\frac{f}{g}$ auf ganz A definiert ist, wenn f und g es sind.

- Hintereinanderschaltung von Funktionen:

Seien $f : A \to \mathbb{R}$ und $g : B \to \mathbb{R}$ zwei Funktionen, $A \subseteq \mathbb{R}$, $f(A) \subseteq B$. $f(A)$ ist dabei definiert als Menge aller Werte $\{f(a) \mid a \in A\}$. Das ist also die Menge aller Werte von f, da A der Definitionsbereich von f ist. Die Menge aller Werte einer Abbildung f nennt man auch $Bild(f)$ oder internationaler $Im(f)$. Wir wollen f und g hintereinanderschalten, zuerst f, dann g anwenden: Wir nehmen ein $x \in A$, wenden darauf f an, erhalten $f(x)$, wenden auf dies Zwischenergebnis g an und landen bei $g(f(x))$. Voraussetzen müssen wir dabei, dass $f(x)$ im Definitionsbereich von g liegt, andernfalls scheitern wir mit dem zweiten Schritt, können g gar nicht auf das Zwischenergebnis anwenden. Damit das stets klappt, muss offenbar gerade $f(A)$ Teilmenge des Definitionsbereiches von g sein, was wir mit $f(A) \subseteq B$ formulierten. Somit haben wir unter den angegebenen Voraussetzungen:

$$\begin{array}{rccc} g \circ f: & A & \to & \mathbb{R} \\ & x & \mapsto & g(f(x)) \end{array}$$

„$g \circ f$" („Kringel") liest man „g hinter f". Das ist wieder eine Funktion, und das Resultat von deren Anwendung auf x ist mit dem Rechenausdruck $g(f(x))$ beschreib- und berechenbar. Man nennt auch gern die zuerst anzuwendende Funktion (hier f) „innere Funktion", die danach kommende „äußere Funktion", eben weil das im geschachtelten Rechenausdruck $g(f(x))$ so aussieht.

Kringel ist eine neue Art von Produkt von Funktionen, die tatsächlich assoziativ ist, d.h. $(h \circ g) \circ f = h \circ (g \circ f)$, so dass Ausdrücke wie $h \circ g \circ f$ sinnvoll sind. *Aber Kringel ist nicht kommutativ, im allgemeinen ist* $g \circ f \neq f \circ g$, wie man sich an Beispielen aus dem Leben klarmache. (Ziehen Sie einmal zuerst die Strümpfe, dann die Schuhe aus!) Hier ein Beispiel mit reellen Funktionen: $f(x) = x^2$, $g(x) = 2^x$, $x \in \mathbb{R}$. Dann ist $g(f(3)) = 2^9$, $f(g(3)) = 8^2$, also völlig verschieden. Das andere Produkt von Funktionen, die Multiplikation $f \cdot g$, ist dagegen kommutativ, weil das Produkt bei den Zahlen es ist. Nun kommt eine schwerwiegende

Verwechslungsgefahr:

Bei Zahlen $a \neq 0$ ist $a^{-1} = \frac{1}{a}$, das Inverse bezüglich der Multiplikation, mit der charakteristischen Eigenschaft: $a^{-1} \cdot a = 1$, das neutrale Element bezüglich der Multiplikation mit seiner charakteristischen Eigenschaft: $1 \cdot a = a$, für alle a. *Bei Funktionen dagegen* bezeichnet man mit f^{-1} nicht etwa das Inverse von f bezüglich der Multiplikation $\cdot$, sondern bezüglich der Kringel-Multiplikation! Nun ist aber das neutrale Element hier die Funktion id, mit $id(x) = x$ (identische Funktion). Also ist die charakteristische Eigenschaft von f^{-1}die folgende: $f^{-1} \circ f = id$, und das heißt: $f^{-1}(f(x)) = x$, für alle x aus dem Definitionsbereich von f. Das heißt: f^{-1} macht die Operation f rückgängig, ist die Umkehrfunktion von f. Das geht natürlich nicht, wenn f nicht injektiv ist, also zwei Argumente x_1, x_2, $x_1 \neq x_2$, zusammenwirft zu $f(x_1) = f(x_2)$. Dann müsste automatisch gelten: $f^{-1}(f(x_1)) = f^{-1}(f(x_2))$, kurzum kann keine Umkehrfunktion f^{-1} von f existieren. Außerdem verlangt man noch, dass f^{-1} auf dem ganzen Wertebereich von f definiert sei, und dafür muss f surjektiv sein, d.h. die Bilder von f müssen den ganzen Wertebereich ausfüllen. Dann gilt auch $f(f^{-1}(x)) = x$ für alle x aus dem Definitionsbereich von f^{-1}, d.h. aus dem Wertebereich von f. (Ein solches x ist nach unseren Voraussetzungen eindeutig als $x = f(a)$ darstellbar, und wir haben $f^{-1}(f(a)) = a$, also $f(f^{-1}(x)) = f(a) = x$. Automatisch ist f die Umkehrfunktion von f^{-1}, und auch f^{-1} ist sowohl injektiv als surjektiv, d.h. bijektiv. Zusammengefasst:

- Umkehrfunktion zu einer bijektiven Funktion $f : A \to B$:

$$\begin{array}{rccl} f^{-1}: & B & \to & A \\ & x & \mapsto & \textit{das } y \in A \textit{ mit } f(y) = x \end{array}$$

Beispiel: Zu $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $f(x) = x^2$, ist $f^{-1} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ zu definieren durch $f^{-1}(x) = \sqrt{x}$. In Worten: Ziehen der Quadratwurzel ist die Umkehroperation zum Quadrieren, und umgekehrt. Analog verhält es sich mit exp und ln. (Stetige umkehrbare Funktionen müssen stets streng monoton wachsend oder streng monoton fallend sein. Stetigkeit bedeutet grob: Unterscheiden sich die Urbilder nur wenig, so auch die Bilder; insbesondere darf es keine Sprünge oder Pole geben.)

Hier ist eine völlig allgemeine Standardanwendung der Umkehrfunktionen:

Nehmen wir f als umkehrbar an, und wir wollen ein Gleichung der Form

$$f(x) = c$$

lösen, c gegeben, x Unbekannte. Dann ist diese Gleichung äquivalent zu

$$x = f^{-1}(f(x)) = f^{-1}(c).$$

Wir können x also ausrechnen mittels Anwendung der Umkehrfunktion auf das vorgegebene Bild c.

Beispiel: $e^x = 3$ hat die Lösung $x = \ln(3)$. $2^{x-1} = 3$ ist gleichwertig zu $(x-1)\ln(2) = \ln(3)$, also $x = \frac{\ln(3)}{\ln(2)} + 1$. (Unbekannte im Exponenten kann man stets da herunterholen durch Anwendung der natürlichen Logarithmusfunktion.)

Schauen wir nun einmal, welche Funktionen man aus unseren Grundfunktionen mit welchen Verknüpfungen bekommt, und lassen wir dabei eine wichtige Klassifikation der Funktionen erscheinen:

- Aus *id* und den Konstanten entstehen mittels der Addition sowie Multiplikation beschränkt auf „Funktion mal Zahl" allein die linearen Funktionen. ($mx + b$ ist offenbar bildbar, und Addition solcher Ausdrücke sowie ihre Multiplikation mit einer Zahl führt nie darüber hinaus.)
- Aus *id* und den Konstanten entstehen mittels Addition und Multiplikation genau alle Polynome. (Man bildet x^n für beliebiges $n \in \mathbb{N}$ und kann mit Zahlen multiplizieren sowie addieren, also jedes Polynom $\sum_{i=0}^{n} a_i x^i$ bilden. Addition und Multiplikation führen dann nicht mehr über die Polynome hinaus.)
- Aus *id* und den Konstanten entstehen mittels Addition, Multiplikation und Division genau alle gebrochen rationalen Funktionen. (Man kann offenbar alle Polynome und dann auch Quotienten von Polynomen bilden, das sind aber die gebrochen rationalen Funktionen. Die Operationen führen dann nicht mehr darüber hinaus.)
- Wir haben mit den Potenzfunktionen zu gebrochenen Exponenten auch algebraische und mit beliebigen reellen Exponenten sowie mit exp, ln auch transzendente Funktionen als Grundfunktionen und können damit bereits sehr komplizierte Exemplare bilden.

Ein recht fundamentaler Punkt ist nunmehr zu besprechen, das Lesen von komplizierteren Rechenausdrücken:

Man lese nie „von links nach rechts", sondern rekonstruiere stets den baumartigen Funktionsbildungsprozess aus den Grundfunktionen. Wir führen das an einem Beispiel vor:

$$\frac{e^{x^2-2x}}{2x+1}$$

Letzter Aufbauschritt war Quotientenbildung, und so zerfällt die Sache in

$$e^{x^2-2x} \quad 2x+1.$$

Der rechte Teil ist einfach, trotzdem analysierbar als Summe, zerfällt in $2x$ und 1. $2x$ zerfällt wieder in ein Produkt, $2 \cdot x$, dessen Bestandteile nun aber Atome, nämlich Rechenausdrücke von Grundfunktionen sind.

Der linke Teil

$$e^{x^2-2x}$$

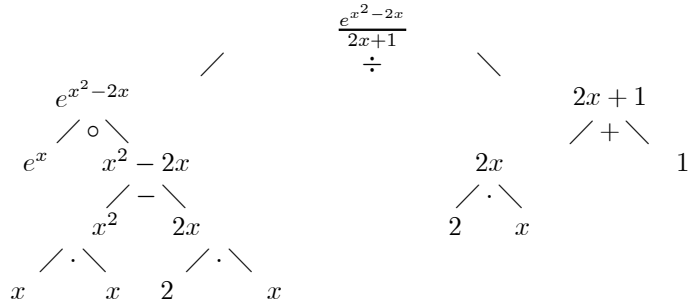
zerfällt mit dem Aufbauschritt „Hintereinanderschaltung" in folgende:

$$e^x \quad x^2 - 2x,$$

links der Rechenausdruck der „äußeren" Funktion, rechts der „inneren" Funktion.

Die äußere ist hier Grundfunktion, die innere analysierbar als Differenz, deren Bestandteile jeweils Produkte von Grundfunktionen sind. Fassen wir das Ganze

noch einmal zu folgendem baumartigen Graphen zusammen:



Beachten Sie: Die Analyse führt stets auf Grundfunktionen (im Beispiel die Potenzfunktion mit Exponenten 1 (Rechenausdruck x), Konstanten sowie die Exponentialfunktion). Achtung: Wir arbeiteten hier der Anschaulichkeit und Ökonomie halber stets mit Rechenausdrücken für die jeweiligen Funktionen. Aber natürlich ist die Operation $\circ$ nicht für Rechenausdrücke definiert, gemeint ist die Hintereinanderschaltung von Funktionen, deren Rechenausdrücke in den Baumknotenpunkten stehen. Ferner ist bei $\circ$ sowie Quotientenbildung $\div$ auf die Reihenfolge zu sehen: Wir nannten hier den Ausdruck der äußeren Funktion links, den der inneren rechts, ferner beim Quotienten den Zähler links.

Bemerkung: Eine solche Analyse ist keineswegs eindeutig, z.B. kann man $\frac{1}{\sqrt{2x}}$ als Hintereinanderschaltung $g(f(x))$ mit $g(x) = 1/x$, $f(x) = \sqrt{2x}$ auffassen oder auch als Hintereinanderschaltung $k(h(x))$ mit $k(x) = 1/\sqrt{x}$ und $h(x) = 2x$, schließlich auch als Quotienten $u(x)/t(x)$ mit $u(x) = 1$, $t(x) = \sqrt{2x}$. Es kommt durchaus auf den Kontext an, welche dieser Möglichkeiten die praktischste ist. Korrekt sind sie alle, und jede liefert stets korrekte Resultate, sei es bei der Graphenkonstruktion, beim Ableiten oder Integrieren.

Die beschriebene Analyse ist sehr nützlich und sogar unerlässlich: Typisch will man von einer Funktion wissen, wie der Graph aussieht, wie die erste Ableitung zu berechnen ist. Wenn man das für die Grundfunktionen beherrscht und weiß, wie sich die Sache jeweils bei Verknüpfungen verhält (also so etwas wie Verknüpfungsregeln hat), so beherrscht man das Gewünschte im Prinzip für jede komplex zusammengesetzte Funktion. Unsere Funktionsbildung ist ein Beispiel für eine äußerst wichtige Denkfigur, die Rekursion: Komplexeres wird aus einfachen Atomen aufgebaut mittels regelhafter Prozesse. Ein anderes Beispiel tritt bei den natürlichen Zahlen auf: Wenn man weiß, dass $0 + n = n$ und $(m + 1) + n = (m + n) + 1$, beides für alle $n, m \in \mathbb{N}$, so hat man die Addition für *alle* natürlichen Zahlen definiert!

1.3. Konstruktion der Graphen zusammengesetzter Funktionen aus denen der Bestandteile.

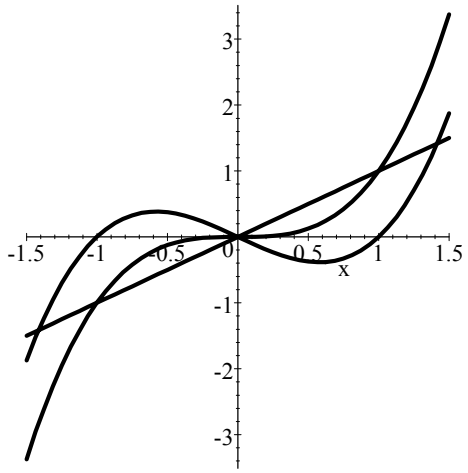
1.3.1. *Geometrische Pendants zu den Verknüpfungen von Funktionen.*

- Summe und Differenz von Funktionen

Wir nehmen an, man kenne bereits die Graphen von f und g und wolle den Graphen von $f + g$ (grob qualitativ) daraus konstruieren. Für $f + g$ ist an jeder Stelle x der Wert $f(x)+g(x)$ zu bilden. Nun stellen sich die Zahlwerte $f(x)$, $g(x)$ als gerichtete Strecken dar (gerade so, wie man Zahlen auf der Zahlengeraden darstellt), und $f(x) + g(x)$ kann man daraus graphisch bestimmten (Aneinandersetzen oder

Abziehen, je nach Richtungen). Auf diese Weise entsteht der Graph von $f + g$ durch Überlagerung der einzelnen Graphen. Entscheidend ist es nun, dass man diese graphische Operation nicht nur punktweise vornimmt, sondern über ganze Bereiche hin überblickt, was sich daraus ergibt. Insbesondere kann man erkennen, welches Vorzeichen zu $f + g$ (in einem Bereich) gehört, ob $f + g$ dort steigt oder fällt, usw. Ebenso leicht kann man $f - g$ bilden.

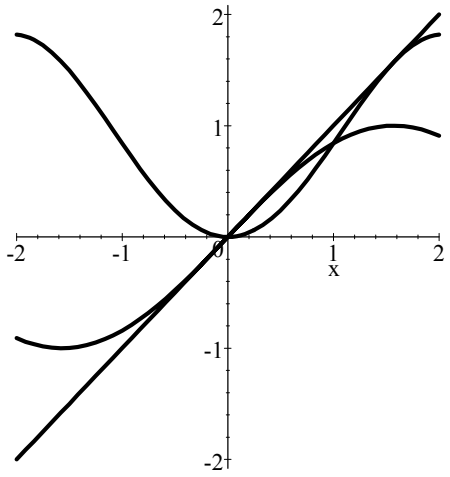
Hier ist ein Beispiel. Man sollte erkennen, welche Kurve die „Summenkurve" der beiden andern ist, welche man als „Differenzkurve" von welchen auffassen kann (für unsere Zwecke hier sind die Zahlenwerte völlig irrelevant und sollten ignoriert werden.



Die Kurve mit den drei Nullstellen „plus" die gerade ergibt die andere Kurve, diese „minus" die Gerade ergibt die erstere.

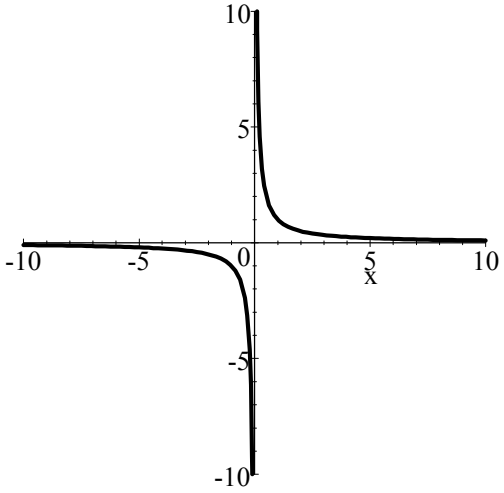
- Produkt und Quotient von Funktionen

Hier ist die Verknüpfung nicht geometrisch so präzise nachzuvollziehen, man sieht das Produkt zweier als Strecken dargestellter Zahlen nicht. Dennoch lässt sich wieder das Verhalten von $f \cdot g$ (bzw. f/g) an den Graphen von f und g *grob qualitativ* ablesen: Man sieht das Vorzeichen, auch wieder Monotonie. (Steigt in einem Bereich sowohl f als auch g, so tut $f \cdot g$ das auch, usw.) Ein Beispiel:

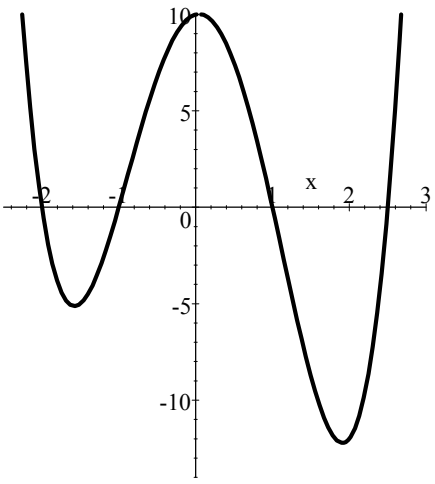


(Die vasenförmige Kurve ist „Produkt" der beiden anderen.)

Interessante Phänomene ergeben sich durch Quotientenbildung, insbesondere Pole und waagerechte Asymptoten, hier das einfachste Beispiel $f(x) = 1/x$ $(x \neq 0)$:



Man erkennt, dass für $x \to 0$, $x > 0$, $f(x)$ gegen ∞ geht. (Die Funktion hat einen Pol, eine senkrechte Asymptote, bei $x = 0$.) Für $x \to \infty$ hat man $f(x) \to 0$. Auch die x-Achse ist eine Asymptote der Funktion. Bei Polynomen treten diese Phänomene nicht auf, ein Polynom sieht stets so aus, dass nach einigen glatten Schwingungen die Werte nach Unendlich gehen (positiv oder negativ), typisch so:



Bei Produkten und Quotienten spielt es eine große Rolle, welcher Bestandteil *lokal* - d.h. in einer bestimmten Umgebung - jeweils dominiert. Wenn z.B. der eine Faktor nach 0, der andere nach ∞ tendiert, so ist es entscheidend, welcher das drastischer, schneller tut - er bestimmt, wie sich das Produkt entwickelt. Natürlich wird auch bereits bei Summen das Dominanzphänomen wichtig: Etwa bei $x - x^3$ dominiert der erste Summand in der Umgebung von $x = 0$, der zweite bei $\pm\infty$. Hier sind die wichtigsten Dominanzregeln, die man durch Analyse mittels der ersten Ableitung herausbekommt. Sie gelten für die Umgebung von ∞:

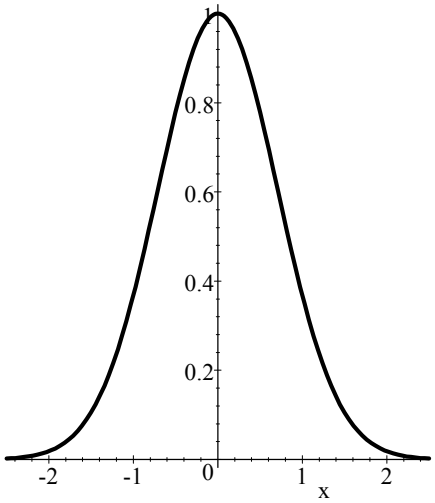
Jedes Polynom höheren Grades dominiert über jedes niederen Grades.
Die Exponentialfunktion dominiert über jede Potenzfunktion.
Jede Potenzfunktion dominiert über die Logarithmusfunktion.

Einige Anwendungen:
Von $\frac{x^2+100x}{x^3}$ weiß man sofort, dass der Wert für $x \to \infty$ gegen 0 geht.
$\frac{x^{1000}}{e^x} \to 0$ für $x \to \infty$. $\frac{\ln(x)}{x^{0.0001}} \to 0$ für $x \to \infty$.
Insbesondere erhält man Aufschluss über mögliche Vereinfachungen eines Rechenausdrucks für bestimmte Umgebungen, z.B. kann man in $x^2 + 1$ für große $|x|$ getrost die 1 vergessen, in $x + e^x$ für große positive x den Summanden x, für sehr kleine x (nahe $-\infty$) dagegen den Summanden e^x.

- Hintereinanderschaltung

Diese Operation ist graphisch zunächst etwas mühsam: Man liest auf dem Graphen von f zu x den Wert $f(x)$ ab, legt diesen auf die x-Achse und liest dann den Wert $g(f(x))$ mittels des Graphen von g ab. Praktisch geht man günstig so vor, dass man sich auf die Werte, welche der Graph von f zeigt, noch die Funktion g (deren Graphen man etwa getrennt aufgemalt hat) angewandt denkt. Man hat hier auch wieder ganz nützliche Regeln: Wenn z.B. f und g beide monoton steigen, so auch $g \circ f$. Steigt die eine, während die ander fällt, so fällt auch $g \circ f$. Fallen beide, so steigt $g \circ f$ (wie man sich überlege!).

Wir überlegen den Graphen zu e^{-x^2}: Die innere Funktion mit dem Ausdruck $-x^2$ fällt ab $x = 0$, mit wachsenden x. Folglich fällt dort der Graph von e^{-x^2}. Da $-x^2$ gegen $-\infty$ geht für $x \to \infty$, schneller als $-x$, muss e^{-x^2} (schnell) gegen Null gehen. Weiter überlegt man, dass e^{-x^2} denselben Wert für x wie für $-x$ ergibt, der Graph muss also spiegelsymmetrisch zur y-Achse liegen. (Solch eine Funktion nennt man „gerade".) Insgesamt resultiert folgende Figur:



1.3.2. *Ein nützlicher Fragenkatalog für die Konstruktion von Graphen zu gegebenen Rechenausdrücken.* Hier sammeln wir noch einmal die Aspekte, die typisch nützlich werden (und es bereits bei den wenigen besprochenen Beispielen wurden). Tatsächlich ergibt sich auch eine Reihenfolge, in der die Überlegungen natürlicherweise ablaufen, vom Gröberen zum Feineren hin. Allerdings sei bemerkt, dass nicht etwa in jedem Falle immer dieselben Punkte zu beachten sind. Vielmehr sollte man im Einzelfall sehen, was besonders wichtig und interessant ist und was nicht. (Also nicht immer krampfhaft alle Felder ausfüllen! Beispielsweise ist die Frage nach Asymptoten bei Polynomen nicht besonders intelligent.) Dabei kann es zu verfeinerten Fragestellungen kommen, für die man feineren Werkzeuges bedarf. An erster Stelle ist dazu die Ableitung zu nennen, die wir u.a. auch für feinere lokale Graphenbetrachtung nutzen werden. Auch die erwähnten Standard-Dominanzen sind von der Art, die wir im Augenblick nur als festes Wissen aufnehmen, ausgestattet jedoch mit einem gewissen quantitativen Grundverständnis.-Dass x^3 für große x nichts ist gegen x^4, dass 2^x für große x ($x = 10$ ist natürlich zu wenig!) den Wert x^{10} geradezu erschlägt, so etwas sollte man schon ganz konkret verstehen. Nun folgt der Katalog; manchmal werden dabei abstrahierbare Regeln genannt. Man sollte sie als einfache jederzeit zu produzierende Überlegungen verstehen, nicht etwa als Stoff zum Auswendiglernen.

- **Definitionsbereich**

Zwar gehört dessen Angabe zur Definition einer Funktion. Aber wenn man nur einen Rechenausdruck hat, so liegt die Frage nahe, in welchem Bereich der

überhaupt definiert ist, so dass man die zugehörige Funktion mit maximal möglichem Definitionsbereich bildet. Das kann durchaus von inhaltlichem Interesse sein, wenn man zunächst einmal engere Anwendungen im Sinne hatte. Vielleicht ergibt sich in größerem Zusammenhang Sinnvolles daraus. Hinzu kommt, dass man etwa typisch bei gebrochen rationalen Funktionen oder allgemeiner Quotienten routinemäßig nachsehen sollte, wie viele Nullstellen der Nenner hat und wo diese (gegebenenfalls: ungefähr) liegen. Anschließend ist die interessantere Frage zu stellen, wie sich die Funktion dort verhält. Denn es kann sich um Polstellen handeln (senkrechte Asymptoten), aber auch um Konvergenz zu einem endlichen Wert, mit welchem die Funktion stetig zu ergänzen wäre. Die Frage lautet nämlich: Geht der Zähler an der betreffenden Stelle nicht oder langsamer als der Nenner nach Null? Dann ist es ein Pol. Geht er genau so schnell nach Null? Dann kann der Wert des Bruchs gegen irgendeine Konstante gehen. Oder geht der Zähler dort schneller nach Null als der Nenner? Dann geht der Bruch nach Null! (Man merke also: „Definitionslücke" genügt nicht, die weit interessantere Frage ist damit nicht einmal gestellt.)

- **Standard-Symmetrien**

Ist f gerade oder ungerade (spiegelsymmetrisch zur y-Achse / punktsymmetrisch zum Koordinatenursprung, rechnerisch: $f(x) = f(-x)$ (für alle x) / $f(-x) = -f(x)$ (für alle x))?

Man beachte, dass man die benötigte Information um die Hälfte reduziert hat, wenn man eine solche Symmetrie feststellen konnte. Selbstverständlich kann man auch Symmetrien um andere Achsen oder Punkte begegnen, allerdings sind die nicht am Rechenausdruck so gut sichtbar wie die Standard-Symmetrien. Vielmehr erkennt man sie besonders günstig dann, wenn man den Graphen aus einem mit einfacherer Symmetrie in kontrollierter Weise entstehen lassen kann, vgl. z.B. 6.3.2. Für die Standardsymmetrien hat man auch einfache Zusammensetzungsregeln: Eine Summe gerader [ungerader] Funktionen ist sicher gerade [ungerade], und wie bei ganzen Zahlen: „gerade mal gerade = gerade", „ungerade mal ungerade = gerade", „gerade mal ungerade = ungerade". Ist f gerade, so ist $g \circ f$ sicher gerade, gleichgültig, welche Eigenschaften g hat.

- **Vorzeichen (stückweise)**

Welches Vorzeichen hat f in einem Intervall? (Das kann man einfach aus den Vorzeichen der Bestandteile ermitteln, zuweilen ist noch das quantitative Überwiegen eines Bestandteiles heranzuziehen. Beispiele: $g(x) + h(x)$ ist sicher positiv, wenn beide Summanden es sind, ebenso $g(x) \cdot h(x)$, und letzterer Ausdruck ist negativ genau dann, wenn die Faktoren verschiedene Vorzeichen haben - und das überblickt man graphisch über ganze Bereiche. $g(f(x))$ ist sicher positiv, wenn g nur positive Werte annimmt.)

- **Monotonie (stückweise)**

Vielfach ist es leicht, aus dem Monotonieverhalten (also Steigen oder Fallen) der Bestandteile auf das der Zusammensetzung zu schließen. Steigen beide Summanden, so die Summe, ebenso für Produkte. Sind f und g in einem Bereich beide monoton steigend oder beide fallend, dann ist auch $g \circ f$ dort stets monoton steigend. Steigt die eine, während die andere fällt, so fällt die Hintereinanderschaltung. Natürlich gibt es unbequeme Fälle: f steigt, g fällt, dann ist zunächst unklar, was die Summe oder das Produkt tun. Dafür hat man wieder notfalls Dominanzregeln

und letztlich die erste Ableitung (deren Vorzeichen allein entscheidet über Steigen und Fallen der Originalfunktion!)

- **Dominanz (lokal an einer Stelle (auch $\pm\infty$)**

Die Frage lautet, ob einer der Bestandteile eines zusammengesetzten Ausdrucks sich an einer Stelle qualitativ völlig durchsetzt, also den Graphenverlauf allein entscheidet. Besonders beim Problem des Gesamtverhaltens für große Beträge von x ist das von Interesse, aber auch beim Auftreten von Polen. Man gewöhne sich (nach Betrachten einiger Beispiele) einfach daran, Unerhebliches wegzulassen und damit Rechenausdrücke bedeutend zu vereinfachen, ohne das wesentliche Verhalten (in einer gewissen Umgebung!) zu ändern. Z.B. wird man aus $\sqrt{x^2+1}$ für große $|x|$ ohne weiteres $\sqrt{x^2}=|x|$ machen und daran das asymptotische Verhalten erkennen. Dagegen spielt der Summand 1 sicher in der Umgebung von $x=0$ die dominierende rolle, so dass man hier geradezu vereinfachen kann zur Konstanten $\sqrt{1}$. Wiederum sind quantitativ feinere Dominanzfragen recht gut mittels der ersten Ableitung zu behandeln.

1.3.3. *Leichte (lineare) Veränderungen von Graphen.* Am letzten Beispiel konnte man sehen, dass das Hintereinanderschalten einer Funktion mit einer anderen ein gegenüber den zwei Bestandteilen völlig neuartiges Bild ergeben kann. Genauer: Der Graph von $g\circ f$ und der Graph von $f\circ g$ sehen im allgemeinen ganz anders aus als der von f. In einem speziellen Falle ist das nicht so, nämlich dann, wenn g eine *nichtkonstante lineare* Funktion ist: Dann setzt sich stets das Bild von f durch. Dies ist eine praktisch sehr wichtige Angelegenheit, aus zwei Gründen:

Einmal kann man mit den Grundfunktionen selbst oder anderen recht einfach gebauten Funktionen für Beschreibungszwecke *direkt* nicht besonders viel anfangen, sondern muss sie an jeweilige Situationen anpassen. In vielen Fällen kommt man dabei mit einer einfachen Vor- und/oder Nachschaltung einer linearen Funktion aus. Gerade so können wir aus der soeben besprochenen Funktion

$$f(x)=e^{-x^2}$$

in einem ersten Schritt die Funktion

$$g(x)=e^{-\left(\frac{1}{\sqrt{2}}x\right)^2}$$

bilden, als

$$g=f\circ l_1,\ \text{oder konkreter } g(x)=f(l_1(x)),$$

mit

$$l_1(x)=\frac{1}{\sqrt{2}}x.$$

Nun haben wir vereinfacht

$$g(x)=e^{-\frac{1}{2}x^2},$$

und daraus gewinnen wir durch

$$l_2\circ g,$$

mit

$$l_2(x) = \frac{1}{\sqrt{2\pi}}x,$$

die Standard-Normalverteilungsdichte

$$\varphi_{0,1}(x) = l_2(g(x)) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}.$$

Das ist also nur eine lineare Modifikation oder Transformation von $f(x) = e^{-x^2}$.

Nun geht es weiter: Aus der Standard-Normalverteilungsdichte $\varphi_{0,1}$ erhalten wir wiederum durch Vor- und Nachschalten geeigneter linearer Funktionen alle Normalverteilungsdichten: Mit $l_{\mu,\sigma}(x) = \frac{x-\mu}{\sigma}$ (das ist dasselbe wie $\frac{1}{\sigma}x - \frac{\mu}{\sigma}$) entsteht

$$\varphi_{0,1}(l_{\mu,\sigma}(x)) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

schließlich schalten wir noch $l(x) = \frac{1}{\sigma}x$ davor, und es kommt heraus:

$$(l \circ \varphi_{0,1} \circ l_{\mu,\sigma})(x) = l(\varphi_{0,1}(l_{\mu,\sigma}(x))) = \frac{1}{\sigma \cdot \sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Damit sind wir tatsächlich angelangt bei der $(\mu, \sigma)-$ Normalverteilungsdichte:

$$\varphi_{\mu,\sigma}(x) = l(\varphi_{0,1}(l_{\mu,\sigma}(x))).$$

Selbstverständlich hätten wir das auch direkt aus der Ausgangsfunktion $f(x) = e^{-x^2}$ gewinnen können, indem wir nur je eine lineare Funktion vor- und nachgeschaltet hätten. (Das Hintereinanderschalten zweier linearer Funktionen ergibt wieder eine lineare.)

Wir halten fest: Anpassen an μ und σ bedeutet nur lineares Transformieren.

Lineares Transformieren spielt auch sonst eine wichtige praktische Rolle: Wenn man in einer Einheit für eine interessierende Größe unbequeme Zahlenwerte bekommt, so transformiert man sie gewöhnlich linear, geht also zu einer anderen Einheit über, so dass die Werte viel größer oder kleiner werden, und manchmal zieht man auch noch etwas ab, um etwa eine Skala zwischen 0 und einem Maximum zu erhalten, oder auch, um einen mittleren Wert bei Null zu erreichen. Selbstverständlich handelt es sich um eben denselben Vorgang, wenn man eine graphische Darstellung passend dimensioniert.

Wir wollen nun einmal kurz in einer Tabelle zusammenfassen, was die linearen Transformationen mit den Rechenausdrücken und mit den Graphen anstellen.

Vorbemerkung: Festbleibende Streckungszentren sind jeweils die Achsen!

Ausgangsausdruck $f(x)$	*Ausgangsgraph von* f
Modifizierter Ausdruck	*Geometrische Operation, welche ausgehend vom Graphen von* f *zum zugehörigen neuen Graphen führt*
$f(x-a)$ (*Dieser Ausdruck entsteht aus* $f(x)$ *durch Einsetzen von* $x-a$ *für* x.)	*Verschieben des Graphen von* f *um* a *nach rechts (!) längs der x-Achse*
$f(a \cdot x),\ a \neq 0$	*Stauchen (!) des Graphen von* f *um* a *längs der x-Achse*
$f(x)+a$	*Verschieben des Graphen von* f *um* a *nach oben längs der y-Achse*
$a \cdot f(x),\ a \neq 0$	*Strecken des Graphen von* f *um* a *längs der y-Achse*

Folgende Dinge sollten dabei beobachtet werden:

- Die Bedeutung der geometrischen Operationen für negative a

„Verschieben nach rechts um a" bedeutet für negatives a: Verschieben nach links mit $|a|$.

„Verschieben nach oben um a" bedeutet für negatives a: Verschieben nach unten mit $|a|$.

„Stauchen längs der x-Achse mit negativem a" bedeutet: Stauchen mit $|a|$ und Spiegeln an der $y-$ Achse. (Beides hintereinander, die Reihenfolge spielt *hier* keine Rolle.)

„Strecken längs der $y-$ Achse mit negativem a" bedeutet: Strecken mit $|a|$ und Spiegeln an der $x-$ Achse. (Beides hintereinander, die Reihenfolge spielt wiederum keine Rolle.)

- Strecken und Stauchen mit positiven Werten > 1 und < 1

Ein Stauchen mit einem Wert a bedeutet dasselbe wie ein Strecken mit dem Wert $1/a$, ein Stauchen mit $1/a$ dasselbe wie ein Strecken mit a.

- Was wäre mit $a = 0$ in den Streckungs- oder Stauchungsfällen?

Nun, auch dafür kommt bei guter Interpretation das Richtige heraus: Man bekäme mit der Bildung des Ausdrucks $f(0 \cdot x)$ die konstante Funktion mit dem Wert $f(0)$, und darum schlossen wir diesen Fall aus, weil die Gestalt des Graphen von f eben nicht bewahrt bleibt. Für $0 \cdot f(x)$ ergäbe sich die konstante Funktion mit Wert 0.

- Zusammenhang zwischen Vor- bzw. Nachschaltung einer linearen Funktion und den Achsen, auf die sich die zugehörigen geometrischen Operationen beziehen

Die innere lineare Funktion bewirkt stets eine auf die x-Achse bezogene Streckungs- und/oder (zu „und" vgl. den nächsten Punkt!) Verschiebungsoperation, und dabei sind die Richtungen der naiven (und falschen!) Anschauung (auf der Zahlengeraden bedeutet doch „-1" ein Verschieben nach *links*, ein Multiplizieren ein *Strecken!)* gerade entgegengesetzt. Dieselbe Intuition stimmt dagegen für die Operationen bezüglich der $y-$ Achse, die von den äußeren linearen Funktionen bewirkt werden. (Ja, da ist es auch richtig, weil nur mit den Funktionswerten auf der $y-$ Zahlengeraden gearbeitet wird. Auf der $x-$ Achse heißt es dagegen: Indem man den

Funktionswert von f an der Stelle $x-1$ (z.B.) bildet und an der Stelle x einträgt, verschiebt man den Graphen von f tatsächlich nach *rechts*. Analog: Wenn man stets den f-Wert von $2x$ bei x einträgt, so *staucht* man den Graphen von f um 2 zusammen.)

- Kombinationen der Einzelvorgänge und Vertauschbarkeit bei den Operationen

Man kann alle besprochenen Einzelvorgänge kombinieren und etwa den Rechenausdruck $cf(ax+b)+d$ bilden. Will man mit den einzelnen besprochenen Einsetzungen (auf der Seite der Rechenausdrücke) und geometrischen Operationen (auf der Seite der Graphen logisch folgern, was in solchem Falle insgesamt passiert, so muss man die einzelnen Vorgänge (auf beiden Seiten) sorgfältig hintereinander ausführen und alle Veränderungen nicht etwa auf den Ausgangszustand, sondern auf die jeweils erreichten Zwischenresultate beziehen. Dann ergibt sich das richtige Resultat, sonst entstehen leicht Fehler. Wir fragen also zunächst, durch welche Einsetzungen der atomaren Art der Ausdruck

$$cf(ax+b)+d$$

aus $f(x)$ entsteht:

1.)	$f(x+b)$	(Einsetzen von $x+b$ für x in den Ausdruck $f(x)$.)
2.)	$f(ax+b)$	(Einsetzen von ax für x in den Ausdruck $f(x+b)$.)
3.)	$cf(ax+b)$	(Multiplizieren der Werte von $f(ax+b)$ mit c.)
4.)	$cf(ax+b)+d$	(Addieren von d auf die Werte von $f(ax+b)$.)

Man macht sich leicht klar, dass 1.) und 2.) nicht vertauscht werden dürfen, sonst entstünde nämlich $f\left(a(x+b)\right)$. Ebenso sind 3.) und 4.) nicht vertauschbar, sonst bekäme man $c(d+f(ax+b))$. Dagegen dürfen die (Einsetzungs-) Operationen 1.) und 2.) beliebig mit 3.) und 4.) vertauscht werden, ohne dass sich das Endresultat ändert.

Das ergibt gemäß unserer Tabelle folgende Reihenfolge der geometrischen Operationen (jeweils auf das Erreichte anzuwenden!), die schließlich vom Graphen von f zu dem von g führt, mit $g(x)=cf(ax+b)+d$:

1.) Verschieben um b nach links.
2.) Stauchen mit a längs der x-Achse.
3.) Strecken mit c längs der y-Achse.
4.) Verschieben um d nach oben.

Man beachte, dass bei 2.) und 3.) je nach Vorzeichen auch Spiegelungen an den Achsen involviert sein können.

Beispiele: Wie entsteht der Graph der $(\mu,\sigma)-$ Normalverteilungsdichte aus dem zu $\mu=0,\sigma=1$? Es ist

$$\varphi_{\mu,\sigma}(x)=\frac{1}{\sigma}\varphi_{0,1}\left(\frac{x-\mu}{\sigma}\right),$$

also:

1.) Strecken längs der x-Achse mit σ.
2.) Verschieben um μ nach rechts.
3.) Strecken mit $\frac{1}{\sigma}$ längs der $y-$ Achse (oder Stauchen mit σ).

Umgekehrt: Wie sieht der Rechenausdruck für eine Parabel (entwickelt aus dem Graphen von $f(x) = x^2$) aus, die nach unten geöffnet ist, ihren Scheitelpunkt in $(2, 3)$ hat und „halb so schnell geht", d.h. deren innere Gestalt sich durch eine Streckung längs der x-Achse mit Faktor 2 aus der Normalparabel ergibt? Hier sind die geometrischen Operationen klar:

1.) Strecken längs der $x-$ Achse mit 2.
2.) Verschieben um 2 nach rechts.
3.) Spiegeln an der $y-$ Achse.
4.) Verschieben um 3 nach oben.

Der Rechenausdruck ergibt sich durch die aus der Tafel zu entnehmenden Einsetzungen (wieder jeweils an den Zwischenergebnissen auszuführen!):

$$-\left(\frac{x-2}{2}\right)^2 + 3.$$

Ein grobes Fazit:

Typisch findet man in Anwendungssituationen leichte Veränderungen von Grundfunktionen oder recht einfachen Zusammensetzungen vor. Das äußert sich im Auftreten irgenwelcher Konstanten als Vorfaktoren oder Summanden, innen oder außen. In all diesen Fällen sollte man sich die Vereinfachung vorstellen, die durch Weglassen von all diesem Kram entsteht. Dann weiß man die Grundgestalt, die zum Ganzen gehört, und kann (so genau und fein wie die Situation erfordert) die Modifikationen anbringen, die von den Zusätzen erzeugt werden. Zum Beispiel sollte man beim Anblick von $e^{\alpha x}$ denken: Wie exp, nur gestreckt bzw. gestaucht längs x-Achse, eventuell gespiegelt an y-Achse. Allerdings achte man auf einen Umstand wirklich genau: Die Veränderungen müssen am Ausdruck ganz innen bzw. ganz außen angebracht sein, nur dann handelt es sich (sicher) um eine einfache lineare Transformation. Ansonsten können kleine Konstanten, irgendwo „dazwischen" angebracht, sehr große qualitative Veränderungen hervorrufen. Ein Beispiel:

$\frac{x^2-3}{x^2}$ hat einen Pol bei $x = 0$. Dagegen hat $\frac{x^2-3}{x^2+1}$ überhaupt keinen Pol, da nach Anbringen des Summanden 1 der Nenner ohne (reelle) Nullstelle ist. (Das Verhalten für große Beträge von x ist in beiden Ausdrücken noch dasselbe). Es handelt sich keineswegs um eine lineare Transformation, die graphisch mit Schieben, Strecken, Spiegeln zu bewerkstelligen wäre. Entsprechend gelingt es nicht, den einen Ausdruck aus dem andern zu erzeugen durch Vor- und Nachschalten einer linearen Funktion.

2. Ableitung (Differentiation) von Funktionen

2.1. Die Idee der Ableitung. Zwei verschiedene Ideen führen unmittelbar zu ein und demselben Begriff der Ableitung, die geometrische Vorstellung einer Tangente an einen Funktionsgraphen in einem vorgesetzten Punkte sowie die Absicht, in einem Punkt eine optimale Näherung einer nichtlinearen Funktion durch eine lineare vorzunehmen. Erstere Idee ist allgemeiner bekannt als Vorwissen, letztere ist hingegen die systematisch wichtigere.

- Die erstere Idee: Eine Funktion f sowie der Punkt $(x_0, f(x_0))$ seien vorgegeben, der Graph von f verlaufe glatt in diesem Punkt. Dann sollte es eine Tangente an den Graphen von f im Punkt $(x_0, f(x_0))$ geben, und wir fragen, wie man deren Steigung ausrechnen kann. Die geometrische Idee dazu ist

es, einfach eine Folge von Punkten $(x_i, f(x_i))_{i\in\mathbb{N}\setminus\{0\}}$ zu nehmen, wobei $(x_i)_i$ gegen x_0 konvergiert - sonst ist die Folge beliebig, und dann eine Folge von Sekanten durch die zwei Punkte $(x_0, f(x_0)), (x_i, f(x_i))$ zu bestimmen, die geometrisch gegen die Tangente konvergieren sollte, so dass sich deren Steigung demnach als Grenzwert der Sekantensteigungen ergibt (zur Illustration vgl. die Abbildung, nachdem der Gebrauch von „Δx" eingeführt ist.)

Damit man überhaupt solch beliebige Sekantenfolgen bilden kann, so setzen wir voraus, dass f in einer beidseitigen Umgebung von x_0 definiert sei.

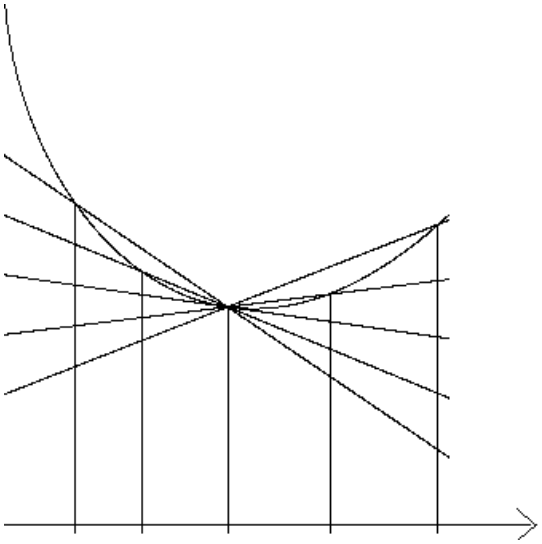
Da es auf die Folge (x_i) nicht ankommt, so sagt man, man lasse x gegen x_0 gehen und betrachte dafür den Grenzwert der Sekantensteigungen, also:

$$\lim_{\substack{x\to x_0\\ x\neq x_0}} \frac{f(x)-f(x_0)}{x-x_0}$$

Dies schreibt man lieber anders, da die Differenz $x-x_0$ gerade die systematisch wichtige Sache ist (nenne sie Δx - man lese das auf keinen Fall „Delta mal x", sondern „Delta x"; das wird wie ein einziger Buchstabe gebraucht, eine untrennbare Einheit, gerade so wie bei Indizes! Insbesondere bedeutet Δx^2, dass die Zahl Δx quadriert wird, es bedarf dazu keiner Klammer wie $(\Delta x)^2$. Gemeint ist inhaltlich immer irgendeine kleine Differenz von x-Werten):

$$\lim_{\Delta x\to 0\ (\Delta x\neq 0)} \frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}.$$

Dieser Grenzwert ergibt nicht nur im Falle, dass der Graph von f in x_0 glatt verläuft, die richtige Tangentensteigung, sondern zusätzlich bildet seine Existenz das adäquate Kriterium von Glattheit, also: Der Grenzwert der Sekantensteigungen existiert genau dann, wenn der Graph von f in x_0 glatt ist. Hier sieht man, wie die Tangente als Grenzsekante entsteht (man stelle sich Δx_1, Δx_2, Δx_3, Δx_4... als Folge vor, die gegen Null geht - die Sekanten gehen dann gegen die Tangente):



$x_0+\Delta x_2$ $x_0+\Delta x_4$ x_0 $x_0+\Delta x_3$ $x_0+\Delta x_1$

Insgesamt landen wir bei folgender Definition:

DEFINITION 18. *Eine Funktion f sei in einer Umgebung von x_0 definiert. f heißt differenzierbar in x_0 genau dann, wenn der Grenzwert*

$$\lim_{\Delta x \to 0 \ (\Delta x \neq 0)} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

existiert. Dessen (eindeutig bestimmter) Wert heißt dann $f'(x_0)$, „Ableitung von f an der Stelle x_0". Die Ableitungsfunktion von f ist dann die Funktion $x \mapsto f'(x)$, definiert an allen Stellen x, wo f differenzierbar ist. Es sei noch einmal die Formel hervorgehoben:

$$f'(x) = \lim_{\Delta x \to 0 \ (\Delta x \neq 0)} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \text{ (wenn existent).} \tag{2.1}$$

Beispiel: Nehmen wir einmal für die Funktion $f(x) = x^2$ an der Stelle $x_0 = 3$ folgende konkrete Folge von Δx-Werten und geben dazu die Sekantensteigungen:

Δx	$\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x} = \frac{(3+\Delta x)^2-3^2}{\Delta x}$
0.1	6.1
0.01	6.01
0.001	6.001
0.0001	6.0001
0.00001	6.00001

Man wird hier *empirisch* bereits einsehen, dass $f'(x_0)$ in unserem Falle den Wert 6 haben wird. *Mathematisch* freilich muss man beweisen, dass das wirklich der Grenzwert ist. Nun, im Beispiel ist das wirklich leicht:

$$\frac{(3 + \Delta x)^2 - 3^2}{\Delta x} = \frac{6\Delta x + (\Delta x)^2}{\Delta x} = 6 + \Delta x \ (\to 6 \text{ für } \Delta x \to 0, \text{ klar!})$$

(Hier immer $\Delta x \neq 0$ für die Folgenbildung!) Sogar allgemein für jedes x_0 anstelle von 3 ist das einfach:

$$\frac{(x_0 + \Delta x)^2 - 3^2}{\Delta x} = \frac{2x_0\Delta x + (\Delta x)^2}{\Delta x} = 2x_0 + \Delta x \ (\to 2x_0 \text{ für } \Delta x \to 0)$$

Wir sehen damit, dass für unsere Quadratfunktion $f(x) = x^2$ gilt: $f'(x) = 2x$, für alle $x \in \mathbb{R}$.

- Die zweite Idee: Lokale lineare Approximation einer Funktion, Näherung 1. Ordnung

Diese Idee gibt mehr inhaltlichen Hintergrund, ihr Gesichtspunkt ist überhaupt ein sehr allgemeiner und bedeutender, für theoretische sowie für Anwendungszwecke. Schließlich werden wir sehen, dass diese Idee viel verallgemeinerungsfähiger ist, und zwar sowohl hinsichtlich der Dimensionen (→ Funktionen mehrerer unabhängiger Variablen) als auch hinsichtlich der Fortsetzung im Grade der Näherung, die bis hin zur exakten Reihendarstellung komplizierter Funktionen führt. Um gerade diese letztere Fortsetzung zu verdeutlichen, beginnen wir mit einer Vorstufe, die noch vor der Ableitung liegt, der Näherung 0. Ordnung.

Stellen wir uns vor, dass man eine komplizierte Funktion hat, sagen wir z.B. $\exp(x) = e^x$, und dass man an einer Stelle x_0 ihren Wert kenne, im Beispiel wäre das bei $x_0 = 0$ der Fall. Nun wollen wir wissen, was $\exp(x_0 + \Delta x)$ für kleine $|\Delta x|$

ist, und es genüge uns eine gute Näherung. Für eine in x_0 stetige Funktion (das bedeutet ja gerade: Unterscheidet sich x nur wenig von x_0, so unterscheidet sich auch $f(x)$ nur wenig von $f(x_0)$) liegt es nahe, zu sagen: $f(x_0 + \Delta x) \approx f(x_0)$. Tatsächlich klappt das auch im Beispiel so ordentlich, wie man das erwarten kann:

Δx	$e^{\Delta x}$	absolute Differenz zum Näherungswert $e^0 = 1$
0.1	1.1052	0.1052
0.01	1.0101	0.0101
-0.01	0.99005	0.00995

Fassen wir zusammen:

Näherung 0. Ordnung einer in x_0 stetigen Funktion in einer Umgebung von x_0

$$f(x_0 + \Delta x) \approx f(x_0),$$

geschrieben als Gleichung mit Rest- oder Fehlerglied:

$$f(x_0 + \Delta x) = f(x_0) + R_{f,x_0}(\Delta x).$$

Die Näherung geschieht mit einem Polynom 0. Grades (d.h. einfach mit einer Konstanten), daher „0. Näherung". Wenn die Funktion f stetig in x_0 ist, so erfüllt das Restglied folgende Bedingung:

Restgliedbedingung (oder Fehlerbedingung) 0. *Ordnung:*

$$|R(\Delta x)| \to 0 \text{ für } \Delta x \to 0.$$

(Hier darf auch die konstante Nullfolge als Δx-Folge genommen werden. (Konsequenz: $R(0) = 0$.) Gleichwertig könnte man auch Stetigkeit von R in 0 fordern (mit der Konsequenz $R(0) = 0$) und sich für die Δx-Folgen mit $\Delta x \neq 0$ begnügen.)

Umgekehrt: Fordert man diese in sich vernünftige Bedingung für den auftretenden Rest R bei *irgendeiner* Näherung 0. Ordnung um x_0 bei irgendeiner Funktion f,

$$f(x_0 + \Delta x) = c + R(\Delta x),$$

so folgt daraus:

$$\begin{array}{ll} (i) & f \text{ ist stetig in } x_0 \\ (ii) & c = f(x_0) \end{array}$$

Denn laut Restbedingung hat man für die konstante Folge 0,0,... als Δx-Folge (oder gleich mit $R(0) = 0$): $f(x_0) = c$, da $R(0) = 0$, und für $|\Delta x|$ klein genug hat man $|R(\Delta x)|$ so klein, wie man will, also $f(x_0 + \Delta x)$ so nahe bei c, wie man will. Das bedeutet aber genau die Stetigkeit von f in x_0.

Eine vernünftige Näherung (welche die entsprechende Restbedingung erfüllt) 0. Ordnung von f um x_0 existiert also genau dann, wenn f in x_0 stetig ist. Zusätzlich ist die Näherung eindeutig bestimmt, die Konstante muss gerade $f(x_0)$ sein.

Wir kommen nunmehr zur Näherung 1. Ordnung. Nach unserer Vorbereitung sollte der Ansatz klar sein: Wir wählen zur Näherung ein Polynom 1. Grades, also:

$$f(x_0 + \Delta x) = a + b \cdot \Delta x + R(\Delta x).$$

Das mag etwas befremden, da die unabhängige Variable x (links als $x_0 + \Delta x$ geschrieben) auch als unabhängige Variable der linearen Funktion rechts auftreten sollte, so dass man $\alpha+\beta(x_0+\Delta x)$ erwartet. Aber $\alpha+\beta(x_0+\Delta x) = (\alpha+\beta x_0)+\beta\Delta x$, also kann man das ohne weiteres als lineare Funktion in Δx umschreiben.

Wir fragen uns zunächst, wie man die Restbedingung zu formulieren hätte. Man sollte erwarten, dass die Bedingung 0. Ordnung zu verschärfen wäre: Mit einem Polynom 1. Grades sollte man eine bessere Näherung hinbekommen als mit einer Konstanten. Mit der Bedingung 0. Ordnung bekommt man sofort wie oben, dass $a = f(x_0)$ werden muss, und das setzen wir sofort in unseren Ansatz ein:

$$f(x_0 + \Delta x) = f(x_0) + b \cdot \Delta x + R(\Delta x).$$

Nun sieht man, dass die Bedingung 0. Ordnung weiter nichts über b hergibt: Mit in x_0 stetigem f ist diese Bedingung für *jeden* Wert von b erfüllt, da der Ausdruck $b\Delta x$ nach Null geht für $\Delta x \to 0$. Man braucht also eine schärfere Bedingung, um sinnvoll einen Wert von b zu bestimmen. Nun haben die Zahlen Δx kleine Beträge, wir denken an beliebig kleine Umgebungen von x_0. Daher liegt es nahe, dass man für die (aufgeblasenen) Reste $R(\Delta x)/\Delta x$ ($\Delta x \neq 0$) fordert, dass sie gegen 0 gehen für $\Delta x \to 0$. Das liegt natürlich auch geometrisch nahe; denn für eine gute lokale Näherung eines Funktionsgraphen durch eine Gerade wird man erwarten, dass sich bei beliebiger Vergrößerung der Stelle immer stärker das Bild einer Geraden herauskristallisiert. Die Vergrößerung bedeutet aber, dass man die Fehler entsprechend durch Division durch Δx vergrößert. Daher formulieren wir folgendermaßen:

DEFINITION 19. ***Tangentenzerlegung einer Funktion um eine Stelle*** x_0 ***und Restbedingung 1. Ordnung***

$$f(x_0 + \Delta x) = f(x_0) + b \cdot \Delta x + R(\Delta x)$$

heißt Tangentenzerlegung von f an der Stelle x_0 $f(x_0+\Delta x) = f(x_0)+b\cdot\Delta x$ genau dann, wenn die folgende Restbedingung 1. Ordnung erfüllt ist:

$$\lim_{\Delta x \to 0, \Delta x \neq 0} \frac{R(\Delta x)}{\Delta x} = 0.$$

Wir wollen nun sehen, dass damit die Zahl b eindeutig bestimmt ist, wenn diese Restbedingung überhaupt erfüllbar ist durch irgendeine Wahl von b. Dafür brauchen wir nur die obenstehende Gleichung gleichwertig für $\Delta x \neq 0$ folgendermaßen umzuformulieren:

$$\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} = b + \frac{R(\Delta x)}{\Delta x}.$$

Lassen wir nun Δx gegen Null gehen, so kommt mit der Restbedingung unsere alte Definition der 1. Ableitung gemäß der ersten Idee:

$$b = \lim_{\Delta x \to 0, \Delta x \neq 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}.$$

Also $b = f'(x_0)$. Und ein Grenzwert ist eindeutig bestimmt, wenn er denn existiert. Denn in beliebig kleiner vorgebbarer Umgebung dieses Wertes müssen von einem Index an alle Folgenglieder liegen, was offensichtlich nicht für zwei verschiedene Werte gelten kann, da man um sie hinreichend kleine Kreise ziehen kann, die sich nicht überschneiden.

Wir haben damit folgenden

SATZ 13. *f besitzt eine Tangentenzerlegung an der Stelle x_0 genau dann, wenn die erste Ableitung von f an der Stelle x_0 existiert, und die Tangentenzerlegung lautet damit*

$$f(x_0 + \Delta x) = f(x_0) + f'(x_0) \cdot \Delta x + R(\Delta x).$$

Das Fehlerglied, der Rest R, erfüllt dann die Bedingung

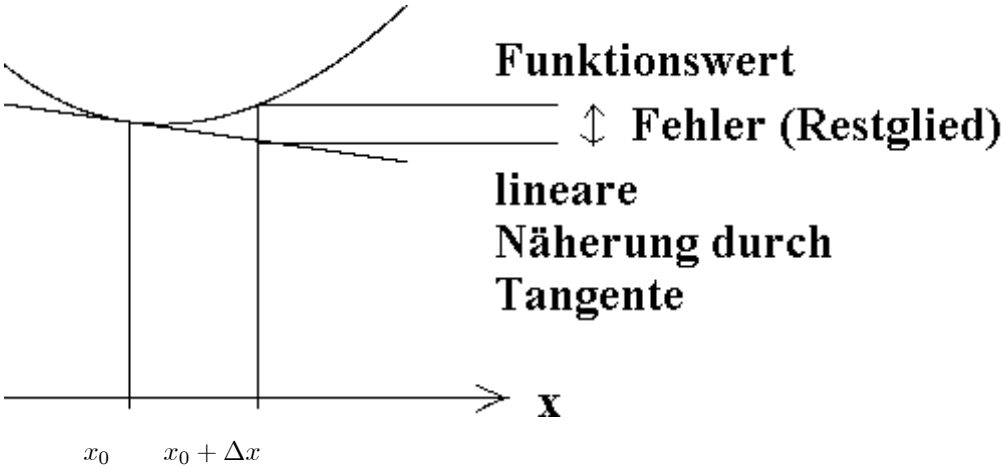
$$\lim_{\Delta x \to 0, \Delta x \neq 0} \frac{R(\Delta x)}{\Delta x} = 0, oder \frac{R(\Delta x)}{\Delta x} \to 0 für \Delta x \underset{\Delta x \neq 0}{\to} 0.$$

Die Näherung

$$f(x_0 + \Delta x) \approx f(x_0) + f'(x_0) \cdot \Delta x$$

heißt dann ***Näherung 1. Ordnung****. (Sie existiert also nur im Falle der Differenzierbarkeit von f an der Stelle x_0.)*

Hier ist eine graphische Illustration dieser Idee:



Umgekehrt hat man auch: Wenn für eine Näherung durch ein Polynom 1. Grades der Gestalt

$$f(x_0 + \Delta x) = a + b \cdot \Delta x + R(\Delta x), \text{ für alle } \Delta x \text{ in einer Umgebung von Null}$$

diese Restbedingung gilt und zusätzlich $R(0) = 0$, dann ist f in x_0 differenzierbar und insbesondere auch stetig, und es gilt $a = f(x_0), b = f'(x_0)$. Denn mit der Gleichung für $\Delta x = 0$ hat man $f(x_0) = a + R(0) = a$ (nur dafür brauchen wir $R(0) = 0$). Nun folgert man wie oben geschehen, dass $f'(x_0)$ existiert und $b = f'(x_0)$. Die Stetigkeit folgt sofort aus der Differenzierbarkeit (ohne weitere Zusatzvoraussetzungen), da mit

$$\lim_{\Delta x \to 0, \Delta x \neq 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} = f'(x_0)$$

insbesondere gilt:

$$\lim_{\Delta x \to 0, \Delta x \neq 0} (f(x_0 + \Delta x) - f(x_0)) = 0, \text{ oder } \lim_{\Delta x \to 0, \Delta x \neq 0} f(x_0 + \Delta x) = f(x_0),$$

was eine der Formulierungen der Stetigkeit von f in x_0 ist.

Beispiel:

Wir wollen nun einmal wiederum am Beispiel der Exponentialfunktion mit $x_0 = 0$ sehen, wie das funktioniert - wir haben ja bereits in die Definition der Zahl e hineingesteckt, dass $\exp'(0) = \exp(0) = 1$:

Die Tangentenzerlegung für exp an der Stelle $x_0 = 0$ lautet:

$$e^{0+\Delta x} = e^0 + e^0 \Delta x + R(\Delta x) = 1 + \Delta x + R(\Delta x),$$

die Näherung 1. Ordnung ist also (Weglassen des Restgliedes!):

$$e^{\Delta x} \approx 1 + \Delta x.$$

Das ergibt:

Δx	$e^{\Delta x}$	absolute Differenz zum Näherungswert $1 + \Delta x$
0.1	$e^{0.1} : 1.1052$	0.0052
0.01	$e^{0.01}$	0.000050167
-0.01	$e^{-0.01}$	0.000049834

Vergleichen Sie mit den Ergebnissen der Näherung 0. Ordnung: Diese ist viel besser, man hat etwa die doppelte Stellenzahl hinter dem Komma korrekt.

2.2. Rezept zum Auffinden der ersten Ableitung mittels Tangentenzerlegung. Aus den Ausführungen der zweiten Idee zur Ableitung ergibt sich eine sehr praktische Methode, Rechenausdrücke für die erste Ableitung von Funktionen *herzuleiten, auch ohne sie zuvor zu kennen.* Das beruht darauf: Hat man eine Zerlegung

$$f(x_0 + \Delta x) = f(x_0) + b\Delta x + R(\Delta x)$$

und erfüllt R die Restbedingung

$$\lim_{\Delta x \to 0, \Delta x \neq 0} \frac{R(\Delta x)}{\Delta x} = 0,$$

so gilt stets:

$$b = f'(x_0).$$

Das ist meist viel leichter als die Berechnung des Grenzwertes der Differenzenquotienten, Beispiele:

$$(x+\Delta x)^3 = x^3 + 3x^2\Delta x + 3x\Delta x^2 + \Delta x^3,$$

und der Restterm

$$3x\Delta x^2 + \Delta x^3$$

hat offenbar die Eigenschaft, noch durch Δx geteilt gegen Null zu gehen, also kann man die Ableitung

$$f'(x) = 3x^2, \text{ für } f(x) = x^3$$

unmittelbar ablesen als den Faktor in der Zerlegung vor Δx. Damit ist sowohl die Existenz der Ableitung (für beliebiges x) als auch ihre Gestalt gezeigt!

Als nächstes Beispiel zeigen wir, dass die Exponentialfunktion überall differenzierbar ist und überall $\exp'(x) = \exp(x)$ gilt. Wir setzen dabei lediglich $\exp'(0) = 1$ und die allgemeine Gleichung $e^{x+y} = e^x \cdot e^y$ voraus. Das ergibt, wenn wir für $e^{\Delta x}$ die Tangentenzerlegung von exp an der stelle $x_0 = 0$ einsetzen (s.o.):

$$e^{x+\Delta x} = e^x \cdot e^{\Delta x} = e^x(1 + \Delta x + R(\Delta x)) = e^x + e^x\Delta x + R(\Delta x).$$

Nun ist der Restterm laut Definition der von der Tangentenzerlegung um Null, also in Ordnung, und wir lesen einfach den Faktor bei Δx ab:

$$\exp'(x) = \exp(x), \text{ für alle } x \in \mathbb{R}.$$

2.3. Der einfache Aufbau aller Ableitungen mittels des rekursiven Aufbaus aller Funktionen. Tatsächlich braucht man auch 2.2 nur für kompliziertere theoretische Überlegungen, wenn man es einmal auf die Grundfunktionen angewandt und zur Herleitung der Regeln für das Zusammensetzen der Ableitungen der Bestandteile zu den Ableitungen von Zusammensetzungen strapaziert hat. Aus letzterem ergeben sich dann einfache Ableitungsregeln, mit denen man komplizierte Ableitungen zusammenbasteln kann.

2.3.1. *Die Ableitungen der Grundfunktionen.* Zunächst führen wir eine Notation ein, die es erlaubt, von Ableitungen von Rechenausdrücken zu reden, ohne immer ein Symbol wie f für die jeweilige Funktion einzuführen:

$$\frac{d}{dx}f(x) := f'(x).$$

Das erinnert an den „Differentialquotienten" $\frac{dy}{dx}$, nur, dass wir statt „y" stets „$f(x)$" schreiben und außerdem $\frac{d}{dx}$ (lies: „d nach dx") als Operator vor den Ausdruck $f(x)$ setzen.

Hier ist die praktische Tabelle mit den Ableitungen unserer Grundfunktionen:

$$\begin{aligned} \frac{d}{dx}x^a &= ax^{a-1},\ a\in\mathbb{R} \text{ beliebig} \\ \frac{d}{dx}e^x &= e^x \\ \frac{d}{dx}\ln(x) &= \frac{1}{x} \end{aligned} \tag{2.2}$$

Die zweite Formel haben wir oben schon eingesehen. Die dritte werden wir mittels einer Regel (Ableitung von Umkehrfunktionen, 2.6) aus der zweiten herleiten, und ebenso werden wir (Kettenregel, 2.5) die erste Formel in voller Allgemeinheit aus der zweiten herleiten. Wir sollten nur noch herausstellen, wie allgemein die erste Grundfunktion ist:

Insbesondere hat man

$$\begin{aligned} \frac{d}{dx}x^0 &= 0 \\ \frac{d}{dx}x^n &= nx^{n-1}\ (n\in\mathbb{N}, n>0) \\ \frac{d}{dx}\sqrt{x} &= \frac{d}{dx}x^{\frac{1}{2}} = \frac{1}{2}x^{-\frac{1}{2}} = \frac{1}{2\sqrt{x}} \\ \frac{d}{dx}\sqrt[5]{x^4} &= \frac{d}{dx}x^{\frac{4}{5}} = \frac{4}{5}x^{-\frac{1}{5}} \\ \frac{d}{dx}x^{-\frac{3}{8}} &= -\frac{3}{8}x^{-\frac{11}{8}}. \end{aligned}$$

Natürlich sollte man es als geometrisch selbstverständlich ansehen, dass eine konstante Funktion Ableitung Konstante Null hat. (Genau genommen wäre der Rechenausdruck $0\cdot x^{-1}$ an der Stelle Null nicht definiert, aber die Ableitung der Funktion $f(x)=x^0=1$ existiert natürlich auch bei $x=0$, mit dem Wert 0.)

2.3.2. *Die Ableitungsregeln für Zusammensetzungen.* Für diese haben wir etwas mehr zu tun. Sie lauten stets genauer: Wenn die Ableitungen der Bestandteile (an den jeweils betreffenden Stellen, die aus der Formel ersichtlich sind) existieren, so auch die Ableitung der Zusammensetzungen, und letztere hat dann den angegebenen Wert.

Linearität des Ableitungsoperators:

$$\begin{aligned} \frac{d}{dx}(cf(x)) &= c\frac{d}{dx}f(x),\ \text{kürzer } (cf)'=cf' \\ \frac{d}{dx}(f(x)+g(x)) &= \frac{d}{dx}f(x)+\frac{d}{dx}g(x),\ \text{kürzer } (f+g)'=f'+g' \end{aligned} \tag{2.3}$$

Produkt- und Quotientenregel:

$$\begin{aligned} (f\cdot g)' &= f'g+g'f \\ \left(\frac{f}{g}\right)' &= \frac{f'g-g'f}{g^2}. \end{aligned} \tag{2.4}$$

Kettenregel (für Hintereinanderschaltungen):

$$\frac{d}{dx}g(f(x)) = g'(f(x)) \cdot f'(x), \text{ kürzer: } (g \circ f)' = (g' \circ f) \cdot f'. \tag{2.5}$$

Regel für Umkehrfunktionen:

$$\left(f^{-1}\right)'(f(x)) = \frac{1}{f'(x)}, \text{ wenn } f'(x) \neq 0. \tag{2.6}$$

Bevor wir die komplizierteren der benötigten Herleitungen geben (nur diese ohnehin), wollen wir zunächst einige Anwendungsbeispiele anschauen, insbesondere zuerst auch den Rest zu den Grundfunktionen erledigen:

Beispiele:

- Die Grundregel

$$\frac{d}{dx}x^a = ax^{a-1}$$

sieht man so: Zunächst hat man für $x > 0$ (so dass man $\ln(x)$ bilden kann):

$$\frac{d}{dx}x^a = \frac{d}{dx}(e^{\ln(x)})^a = \frac{d}{dx}e^{a\cdot\ln(x)} = \frac{a}{x}e^{a\cdot\ln(x)} = \frac{a}{x}x^a = ax^{a-1}.$$

Nur Umformungen und einmal die Kettenregel waren beteiligt.

Tatsächlich bilden $x = 0$ und $x < 0$ Sonderfälle: Für allgemeine Exponenten a ist x^a überhaupt nicht definiert für $x \leq 0$, und dafür macht dann auch die Ableitung keinen Sinn. Für $x = 0$ ist zwar x^a mit $1 > a > 0$ definiert, aber das hat bei $x = 0$ unendliche Steigung und somit keine Ableitung. Genau dies gibt der hergeleitete Ausdruck auch wieder. Mit $a > 1$ hat man $0^a = 0$, die Ableitung existiert als einseitige mit dem richtigen Wert Null, den der Ausdruck wiederum angibt. Natürlich ist x^a zuweilen auch für negative Werte von x definiert, aber dann ergibt sich wiederum der richtige Wert, weil es sich um eine (positive oder negative) ungerade ganze Zahl a handeln muss. Kein Problem: Die Funktion $x \mapsto x^a$ ist dann ungerade, die Ableitung also gerade (kann man leicht herleiten!), und somit stimmt die Formel wieder, da $x \mapsto ax^{a-1}$ mit ungeradem a tatsächlich eine gerade Funktion ist, also ist die Ableitung überall korrekt, wenn sie es auf der positiven Seite war.

- Die Ableitung der Logarithmusfunktion

Die Umkehrfunktions-Regel gibt her:

$$\left(\exp^{-1}\right)'(\exp(x)) = \ln'(e^x) = \frac{1}{\exp(x)} = \frac{1}{e^x}.$$

Dabei ist x beliebig, somit e^x eine beliebige Zahl > 0. Folglich gilt:

$$\ln'(x) = \frac{1}{x} \text{ für alle } x > 0.$$

(Verwenden Sie zunächst einen Buchstaben a für e^x aus der darüber stehenden Formel, das gilt dann für alle $a > 0$, und nun denken Sie daran, dass der Buchstabe in Ausdrücken „für alle...", „es gibt..." keine Rolle spielt.)

- Einfache Beispiele zur Anwendung der Kettenregel:

$$\frac{d}{dx}(x^2+1)^{70} = 140x\left(x^2+1\right)^{69}.$$

$$\frac{d}{dx}\ln(x^2+1) = \frac{2x}{x^2+1}.$$

$$\frac{d}{dx}a^x = \frac{d}{dx}e^{x\cdot\ln(a)} = \ln(a)\cdot e^{x\cdot\ln(a)} = \ln(a)\cdot a^x,\ a>0.$$

$$\frac{d}{dx}\log_a(x) = \frac{d}{dx}\frac{\ln(x)}{\ln(a)} = \frac{1}{\ln(a)}\ln'(x) = \frac{1}{x\cdot\ln(a)},\ a>0.$$

(Man beachte, dass man für den konstanten (!) Nenner $\ln(a)$ nicht die Quotientenregel benötigt (obgleich sie selbstverständlich auch das richtige Resultat brächte, nur viel zu umständlich), sondern einfach das Stehenbleiben eines konstanten Faktors - hier $\frac{1}{\ln(a)}$ - beim Ableiten (gemäß Linearität) benutzen kann.

- Ein Beispiel zur iterierten Verwendung der Kettenregel für Mehrfachschachtelungen:

$$\begin{aligned}\frac{d}{dx}e^{\sqrt{x^2+1}} &= \exp'\left(\sqrt{x^2+1}\right)\cdot\frac{d}{dx}\sqrt{x^2+1} = \exp\left(\sqrt{x^2+1}\right)\cdot\frac{2x}{2\sqrt{x^2+1}}\\ &= \frac{x}{\sqrt{x^2+1}}e^{\sqrt{x^2+1}}.\end{aligned}$$

Man sieht also, wie das logisch funktioniert: Man wendet einmal die Kettenregel an, dabei sieht man, dass die innere Funktion wiederum eine Schachtelung ist, und hat die Kettenregel erneut zur Ableitung dieser inneren Funktion anzuwenden. Mit etwas Übung würde man den Ausdruck $\frac{d}{dx}\sqrt{x^2+1}$ nicht mehr erst hinschreiben, sondern gleich dessen Ergebnis.

2.3.3. *Die Herleitung zweier ausgewählter Regeln: Kettenregel und Quotientenregel.* Zur Herleitung der Kettenregel braucht man lediglich die Voraussetzungen, dass f an der Stelle x_0, g an der Stelle $f(x_0)$ differenzierbar seien, in die Existenz von Tangentenzerlegungen umzusetzen und anschließend das Rezept von 2.2 zu befolgen, d.h. den Ausdruck $(g\circ f)(x_0+\Delta x)$ für die zu differenzierende Funktion zu schlachten. Man beachte, dass wir die komische Kettenregel-Formel auf diese Weise *finden* werden (nicht eine uns von Autoritäten eingeflößte Aussage bestätigen).

Die Voraussetzungen lauten:

(i) $f(x_0+\Delta x) = f(x_0)+f'(x_0)\Delta x + R(\Delta x)\Delta x,\ R(\Delta x)\to 0$ für $\Delta x\to 0$.
(ii) $g(f(x_0)+\Delta y) = g(f(x_0))+g'(f(x_0))\Delta y + S(\Delta y)\Delta y,\ S(\Delta y)\to 0$ für $\Delta y\to 0$.

Wir haben dabei in beiden Fällen die Resttermbedingung 1. Ordnung etwas umformuliert: Sie lautete oben: Rest geteilt durch Δx geht gegen Null. Nennen wir den Rest $R_1(\Delta x)$. Dann können wir auch gleichwertig formulieren, dass der Restterm $R_1(\Delta x)$ eine Darstellung der Form $R(\Delta x)\cdot\Delta x$ besitzt mit der Eigenschaft, dass $R(\Delta x)\to 0$ für $\Delta x\to 0$. Denn ein solcher Restterm geht offenbar durch Δx geteilt nach Null für $\Delta x\to 0$, und umgekehrt kann man zu gegebenem $R_1(\Delta x)$, der

die frühere Bedingung erfüllt, einfach definieren: $R(\Delta x) := R_1(\Delta x)/\Delta x$ und hat damit die neue Form. (Analog für (ii).) Weiterer Hinweis: Es ist kein prinzipieller Unterschied zwischen Δx und Δy, die Gleichungen gelten für jeweils alle Zahlen, wobei die Restbedingungen etwas Interessantes nur für kleine derartige Zahlen aussagen. Jetzt zerlegen wir:

$$\begin{aligned}
& g(f(x_0+\Delta x)\\
[\ = \ & (g\circ f)(x_0+\Delta x),\ \text{also das Gewünschte})]\\
= \ & g(f(x_0)+\underbrace{f'(x_0)\Delta x+R(\Delta x)\Delta x}_{\Delta y \text{ nennen wir das}}) \ (\text{für inneren Ausdruck } (i) \text{ benutzt})\\
= \ & g(f(x_0)+\Delta y)\\
= \ & g(f(x_0))+g'(f(x_0))\Delta y+S(\Delta y)\Delta y \ (\text{benutze } (ii))\\
= \ & g(f(x_0))+g'(f(x_0))(f'(x_0)\Delta x+R(\Delta x)\Delta x)+S(\Delta y)\Delta y \ (\Delta y \text{ eingesetzt})\\
= \ & g(f(x_0))+\underbrace{g'(f(x_0))f'(x_0)}_{\text{Faktor bei } \Delta x,\ \text{gesuchte Ableitung!}}\Delta x+\underbrace{g'(f(x_0))R(\Delta x)\Delta x+S(\Delta y)\Delta y}_{\text{Das sollte der Restterm sein.}}
\end{aligned}$$

Wir lesen die Ableitung als Faktor bei Δx ab, nachdem wir uns überzeugt haben, dass der Restterm in Ordnung ist: Er lautet

$$\begin{aligned}
T(\Delta x) &= g'(f(x_0))R(\Delta x)\Delta x+S(\Delta y)\Delta y\\
&= g'(f(x_0))R(\Delta x)\Delta x+S(\Delta y)(f'(x_0)\Delta x+R(\Delta x)\Delta x)\\
&= [g'(f(x_0))R(\Delta x)+S(\Delta y)(f'(x_0)+R(\Delta x))]\Delta x
\end{aligned}$$

Damit ist $T(\Delta x)$ zerlegt in ein Produkt mit einem Faktor Δx; zu zeigen ist nur noch, dass der andere Faktor nach Null geht für Δx nach Null. Dieser andere Faktor ist folgende Funktion $U(\Delta x)$ von Δx:

$$U(\Delta x)=g'(f(x_0))R(\Delta x)+S(\Delta y)(f'(x_0)+R(\Delta x)).$$

Nun ist $U(\Delta x)$ eine Summe, und es genügt, wenn wir von beiden Summanden zeigen, dass sie für $\Delta x\to 0$ nach Null gehen:

$$g'(f(x_0))R(\Delta x)\to 0 \text{ für } \Delta x\to 0,$$

da $R(\Delta x)$ nach Voraussetzung diese Eigenschaft hat und $g'(f(x_0))$ nach Voraussetzung als endliche Zahl existiert (eine Nullfolge mal einer Konstanten ergibt wieder eine Nullfolge).

Der zweite Summand lautet

$$\begin{aligned}
&S(\Delta y)(f'(x_0)+R(\Delta x)),\ \text{ausführlicher als Funktion von } \Delta x\\
&S\left(f'(x_0)\Delta x+R(\Delta x)\Delta x\right)\left(f'(x_0)+R(\Delta x)\right).
\end{aligned}$$

Das ist ein Produkt, dessen zweiter Faktor nach $f'(x_0)$ geht für $\Delta x\to 0$, weil nach Voraussetzung $R(\Delta x)$ dafür nach Null geht. Wir müssen also zeigen, dass der erste Faktor nach Null geht. Das ist aber der Fall, weil mit $\Delta x\to 0$ auch $\Delta y=f'(x_0)\Delta x+R(\Delta x)\Delta x$ gegen Null geht und nach der zweiten Voraussetzung damit auch $S(\Delta y)$ nach Null geht.

Herleitung der Quotientenregel: Zunächst genügt es, die Regel speziell nur für Funktionen der Form

$$f(x) = \frac{1}{g(x)}$$

zu zeigen. (Den Rest besorge man mit der Produktregel, die man sich leicht bei Bedarf herleiten kann.) Außerdem können wir (ohne Verlust) noch weiter spezialisieren auf den Fall, dass $g(x) > 0$ im relevanten Intervall. - Sonst gehe man über zu $-g$ und benutze die Linearität der Ableitung. Wir bilden nun

$$h(x) = \ln(f(x)) = -\ln(g(x))$$

und leiten dies mittels der schon bereitliegenden Kettenregel ab:

$$h'(x) = \frac{f'(x)}{f(x)} = -\frac{g'(x)}{g(x)}.$$

Das ergibt

$$f'(x) = -\frac{g'(x) \cdot f(x)}{g(x)} = -\frac{g'(x)}{g^2(x)}.$$

Damit ist gezeigt, dass $\frac{d}{dx}\frac{1}{g(x)} = -\frac{g'(x)}{g^2(x)}$, und mit einer Anwendung der Produktregel kommt man nun auf die Formel $\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{f'(x)g(x)-g'(x)f(x)}{g^2(x)}$.

2.4. Anwendung der Ableitung auf die quantitative Bestimmung lokaler Extrema. Eine Funktion wie $f(x) = x^3$ (mit Definitionsbereich $\mathbb{R}$ verstanden) hat keine *globalen* Extrema, d.h. keinen minimalen und keinen maximalen Wert *bezogen auf den gesamten Definitionsbereich.* Denn es kommen beliebig große und beliebig kleine Werte heraus. Auch $g(x) = x^3 - x$ hat keine *globalen* Extrema, wohl aber *lokale, d.h. auf beliebig kleine Umgebungen bezogen.* (f dagegen besitzt auch keine lokalen Extrema.) Das wissen wir von der groben Skizze des Graphen von g. Nunmehr stellen wir das Problem, die betreffenden Stellen genau zu lokalisieren. Es genügt für alle Zwecke, die zugehörigen Abszissenwerte (x-Werte) bestimmen zu können, dann lassen sich die zugehörigen (minimalen oder maximalen) Funktionswerte, an denen man interessiert ist, ohne weiteres finden. Das lokale Verständnis erweist sich als recht fruchtbar, da es mittels der Ableitung bewältigt werden kann und zu für sich interessanten lokalen Extrema führt oder auch gegebenenfalls zu globalen. Daher definieren wir noch einmal ausführlicher:

DEFINITION 20. *f hat an der Stelle x_0 ein lokales Minimum [bzw. lokales Maximum], wenn es ein (beliebig kleines!) Intervall $]x_0 - \varepsilon, x_0 + \varepsilon[$, $\varepsilon > 0$, um x_0 herum gibt, so dass für alle $x \in]x_0 - \varepsilon, x_0 + \varepsilon[$, d.h. $|x - x_0| < \varepsilon$, gilt:*

$$f(x) - f(x_0) \geq 0 \text{ [bzw. } f(x) - f(x_0) \leq 0].$$

Es ist leicht auszurechnen, aber anschaulich noch einfacher zu sehen, dass im Falle der Differenzierbarkeit von f an der Stelle x_0 die Tangente an den Graphen von f im Punkte $(x_0, f(x_0))$ die Steigung Null besitzen muss, wenn f an dieser Stelle ein lokales Extremum hat. Also:

SATZ 14. *Wenn $f'(x_0)$ existiert und f an der Stelle x_0 ein lokales Extremum hat, dann gilt*

$$f'(x) = 0.$$

Zur Anwendung dieser Aussage achte man unbedingt auf die Richtung des „wenn-so"-Pfeiles: Aus $f(x_0) \neq 0$ kann man (bei Existenz der Ableitung) schließen, dass f an der Stelle x_0 *keinen* Extremwert haben kann. (*Nicht aber folgt etwa aus* $f'(x_0) = 0$, dass dort ein Extremwert vorliegt. Man merke sich dazu das Gegenbeispiel $f(x) = x^3$: An der Stelle $x_0 = 0$ verschwindet die Ableitung, aber es handelt sich um keinen Extremwert, sondern um einen Sattel.)

$f'(x_0) = 0$ ist nur eine *notwendige,* aber keine *hinreichende* Bedingung für die Existenz eines lokalen Extremwertes. Das bedeutet: Die Nullstellen der ersten Ableitung bilden nur erst *Kandidaten für Extremstellen.* Man mache sich jedoch den Wert dieser Aussage klar: Nachdem zunächst einmal *jede* reelle Zahl in Frage kam, zumindest jedoch ein Kontinuum, so hat man in aller Regel nunmehr endlich viele Kandidaten (allenfalls abzählbar viele). Wie kommt man zur endgültigen Entscheidung? Hier sind zwei Möglichkeiten, zu denen man eher greifen sollte als zur zweiten Ableitung:

- Man hat bereits eine grobe Skizze und weiß, dass es mindestens k lokale Extremstellen geben muss, und findet nun genau k Nullstellen der ersten Ableitung: Dann müssen diese Nullstellen die Abszissenwerte dieser Extremstellen sein!
- Man betrachtet die Ableitung - und zwar nur ihr Vorzeichen - in einer Umgebung von einer fraglichen Stelle x_0 mit der Eigenschaft $f'(x_0) = 0$: Wenn f' an dieser Stelle einen *Vorzeichenwechsel erleidet, dann handelt es sich um ein Extremum, sonst nicht!* Man kann sogar sagen, ob es sich im ersteren Falle um ein Minimum oder Maximum handelt: Bei Wechsel des Vorzeichens von negativ auf positiv (von links nach rechts gesehen) handelt es sich um ein Minimum, wie man sich veranschauliche. Bei umgekehrter Reihenfolge ist es ein Maximum. (Man beachte, dass bei einer in einem Intervall konstanten Funktion Minimum und Maximum dasselbe sind und die Ableitung konstant Null ist, diesen Fall wollten wir hier als trivialen nicht besonders betrachten. Daher interessieren hier nur echt positive oder negative Ableitungen abseits von x_0.)

Beispiel:
$f(x) = x^3 - x$; die Ableitung wird Null für $3x^2 - 1 = 0$, also $x_1 = -1/\sqrt{3}$, $x_2 = 1/\sqrt{3}$. Bereits aus der groben Skizze erkennt man sofort, dass an der ersten Stelle ein lokales Maximum liegt, an der zweiten ein lokales Mimimum. Alternativ könnten wir auch rein rechnerisch (über das Vorzeichen der Ableitung) bequem zu diesem Resultat gelangen: An der Stelle $x_1 = -1/\sqrt{3}$ beobachtet man den Wechsel $+|-$. Offensichtlich ist nämlich $f'(x) > 0$ für $x < x_1$, und offensichtlich gilt $f'(x) < 0$ für Werte x, die *ein wenig* größer sind als x_1 (nämlich noch kleiner als x_2). Klar liegt bei x_2 der Wechsel $-|+$ vor.

2.5. Globale (stückweise) Monotonieeigenschaften und zugehörige globale (stückweise) Eigenschaften der ersten Ableitung. Bisher haben wir noch nicht globale Eigenschaften der Ableitung ausgenutzt, d.h. solche Eigenschaften, die über ein ganzes Intervall hinweg bestehen. Man hat dazu folgenden

SATZ 15.

Wenn $f'(x) > 0$ *für alle* $x \in [a,b]$, *dann ist* f *auf* $[a,b]$ *streng monoton steigend.*
Wenn $f'(x) < 0$ *für alle* $x \in [a,b]$, *dann ist* f *auf* $[a,b]$ *streng monoton fallend.*
Wenn $f'(x) = 0$ *für alle* $x \in [a,b]$, *dann ist* f *auf* $[a,b]$ *konstant. Insbesondere folgt aus* $f' = g'$ *auf* $[a,b]$, *dass* $f = g + c$, *mit einer Konstanten* c, *auf* $[a,b]$ *gilt.*

Der Beweis dieser Resultate ist ziemlich raffiniert und wird hier ausgelassen, aber immerhin sollte man sich die anschauliche Bedeutung klarmachen und die Aussagen also höchst plausibel finden.

3. Integration

3.1. Die Idee des Integrals. Wie bei der Ableitung hat man zunächst zwei Ebenen zu unterscheiden: Was ist die inhaltliche Bedeutung eines Integrals, und wie rechnet man es aus? Weiter noch ergibt sich wieder die Verzweigung Zahl/Funktion. Hier geht es zunächst um die Idee des *bestimmten* Integrals, noch nicht um dessen Berechnung. Allerdings wird diese Idee direkt zu einer näherungsweisen Berechnung führen, die auch dann immer noch wichtig bleibt, wenn man einige Integrale exakt zu berechnen gelernt hat.

Idee des Integrals: Mittelwert einer Funktion auf einem Intervall, und Zusammenhang mit den Flächeninhalten der Flächen zwischen dem Graphen der Funktion und der x-Achse

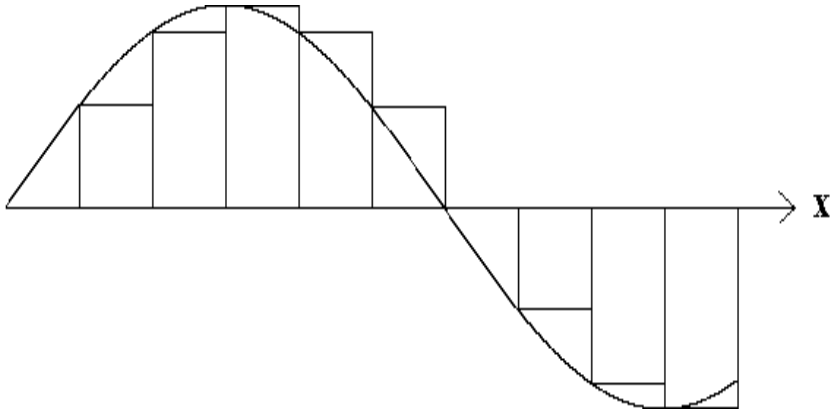
Wie würde man versuchen, so etwas wie den Mittelwert einer Funktion f auf dem Intervall $[a,b]$ (wir setzen zunächst $a < b$ voraus, vgl. aber Formel (3.1)) näherungsweise auszurechnen? (Man beachte, dass es hier um einen Mittelwert von *überabzählbar vielen Funktionswerten,* also von einem Kontinuum geht. Dennoch liegt die einfache Idee nahe, ein einfaches arithmetisches Mittel von endlich vielen Werten zur Näherung zu gebrauchen: Man nimmt Werte $a = x_0$, $x_1 = x_0 + \Delta x$, $x_2 = x_1 + \Delta x, ..., x_n = b$, unterteilt das Intervall also mit $n-1$ Zwischenpunkten gleichmäßig, und dann erwartet man vernünftigerweise (man beachte: $b - a = n \cdot \Delta x$), wenn nur n hinreichend groß ist:

$$\text{Mittel von } f \text{ auf } [a,b] = \overline{f_{[a,b]}} \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i) = \frac{1}{b-a}\sum_{i=1}^{n} f(x_i)\Delta x.$$

Vernünftig ist diese Erwartung selbstverständlich nur dann, wenn f einigermaßen anständig ist, sagen wir stückweise stetig (mit endlich vielen Sprüngen). Im letzteren Ausdruck hat die Summe ohne den Faktor eine eigenständige Bedeutung:

$$\sum_{i=1}^{n} f(x_i)\Delta x \approx \text{Inhalt der Fläche zwischen dem Graphen von } f \text{ und der } x\text{-Achse im Intervall } [a,b] \textit{ mit Orientierungsvorzeichen.}$$

„Mit Orientierungsvorzeichen" bedeutet: Flächen oberhalb der x-Achse werden positiv gezählt, solche unterhalt der x-Achse negativ. Folgende Skizze verdeutlicht diesen Sachverhalt:



Wir wollen bemerken, dass es in solchem Zusammenhang nicht auf eine gleichmäßige Unterteilung ankommt. Entscheidend ist es nur, dass die Länge des längsten der Teilintervalle nach Null geht und dabei mit den Summen immer derselbe Grenzwert herauskommt: So definiert man die Existenz von Integral bzw. Mittelwert. $\overline{f}_{[a,b]}$.

DEFINITION 21. *Wir nennen nun den genauen Inhalt der Fläche zwischen dem Graphen von f und der x-Achse im Intervall $[a, b]$ mit Orientierungsvorzeichen:*

$$\int_a^b f(x)dx = \text{ bestimmtes Integral über } f \text{ in den Grenzen von } a \text{ bis } b.$$

Damit haben wir folgenden Zusammenhang zwischen Mittelwert und Flächeninhalt mit Orientierungsvorzeichen, der auch unmittelbar für sich einleuchtet:

$$\overline{f_{[a,b]}} = \frac{1}{b-a} \int_a^b f(x)dx$$

Wir kennen bereits Möglichkeiten, beide Seiten näherungsweise auszurechnen. Aber dieser Zusammenhang wird uns unmittelbar auch zu einer Möglichkeit exakter Berechnung führen, die allerdings nur für gewisse Funktionen praktisch gangbar sein wird, vgl. den nächsten Abschnitt. Hier wollen wir zunächst noch die Idee verallgemeinern und zeigen, wie typisch Integrale aus inhaltlichen Aufgabenstellungen resultieren: Man hat eine Größe, die näherungsweise durch Summation über kleine Stücke eines Kontinuums (das kann eindimensional sein, wie hier zunächst ausschließlich betrachtet, aber auch zwei- , drei- oder noch höher dimensional) berechnet werden kann. Dann ist diese Größe durch ein Integral zu berechnen. - Dann ist einfach das Summenzeichen durch das Integralzeichen (ein stilisiertes Summenzeichen!) zu ersetzen, das Δx durch dx, was als „unendlich kleines Δx" intuitiv aufzufassen ist. Dazu zwei Beispiele:

Ein Weg wird mit veränderlicher Geschwindigkeit zurückgelegt. Für kleine Zeitabschnitte betrachtet man die Geschwindigkeit als konstant und rechnet „Länge des zurückgelegten Weges = Geschwindigkeit mal Zeit", dann summiert man,

um eine Näherung der Länge des insgesamt zurückgelegten Weges zu erhalten mit $v(t) = Geschwindigkeit\ zur\ Zeit\ t$, $t_0 = Anfangszeitpunkt$, $t_1 = Endzeitpunkt$:

$$\begin{aligned} Weglänge &\approx \sum_{t_0}^{t_1} v(t_i)\Delta t, \text{ also} \\ Weglänge &= \int_{t_0}^{t_1} v(t)dt. \end{aligned}$$

Zweites Beispiel (eigentlich gleich mehrere, wir zeigen hier auf, wie die wichtigsten Integrale in der Wahrscheinlichkeitsrechnung bzw. Statistik auftreten): Wenn wir eine Dichtefunktion f haben, welche die Verteilung einer Größe X beschreibt, dann ist (wieder die Idee, mit stückweise konstanter Dichte auf kleinen Intervallen zu rechnen, also wie bei Histogrammen):

$$\begin{aligned} P(a \leq X \leq b) &\approx \sum f(x_i)\Delta x, \text{ also} \\ P(a \leq X \leq b) &= \int_a^b f(x)dx. \end{aligned}$$

Das ist übrigens genau der Flächeninhalt der Fläche, die im Intervall $[a, b]$ zwischen dem Graphen von f und der x-Achse eingeschlossen wird. Denn Dichtewerte sind stets positiv. Übrigens verwendet man dieselbe Idee, um etwa die Masse bei Material inhomogener Dichte auszurechnen etc. Natürlich macht man diesen Sachverhalt in der folgenden Form gerade zur *Definition* des Begriffs: „Die Größe X ist mit der Dichte f verteilt" (aber mit der vorigen Überlegung verstehen wir den Sinn dieser Definition):

Definition 22. *X ist mit der Dichte f verteilt genau dann, wenn für alle $a \in \mathbb{R}$ gilt:*

$$P(X \leq a) = \int_{-\infty}^{a} f(x)dx.$$

Ähnlich finden wir heraus, wie man den Erwartungswert von X mit der Dichte f auszurechnen hat:

$$\begin{aligned} \mu(X) &\approx \sum x_i \cdot f(x_i) \cdot \Delta x, \text{ also} \\ \mu(X) &= \int_{-\infty}^{\infty} x \cdot f(x)dx. \end{aligned}$$

Zur ersten Zeile: Wieder nimmt man sowohl die Dichte als auch die Größenwerte näherungsweise als auf kleinen Intervallen konstant an und rechnet dann „Summe der Produkte aus Größenwert mal dessen Wahrscheinlichkeit", und letztere Wahrscheinlichkeit, genauer die für Größenwerte im kleinen Intervall der Breite Δx, ist eben näherungsweise $f(x_i) \cdot \Delta x$. Eine Bemerkung zu den unendlichen Grenzen beim Integral: Wenn f nur im Bereich $[a, b]$ Werte $\neq 0$ annimmt, so kann man natürlich diese Grenzen nehmen, aber die angegebene Formel gilt dann ebenfalls, weil f

außerhalb den Wert Null hat und ein Integral über die Nullfunktion in beliebigem (auch unendlichem) Intervall den Wert Null hat. Im allgemeinen weiß man aber nicht, welches Intervall relevant ist, außerdem tritt bei mathematischen Idealverteilungen wie der Normalverteilung ganz $\mathbb{R}$ auf. Stets gilt dann die angegebene Formel.

Nunmehr ist es nicht schwierig, auch die Varianz einer mit Dichte f verteilten Größe X als Integral zu verstehen:

$$\begin{aligned}
\sigma^2(X) &= \int_{-\infty}^{\infty} (x-\mu(X))^2 f(x)dx. \text{ Zum Rechnen ist vielfach bequemer:} \\
\sigma^2(X) &= \int_{-\infty}^{\infty} x^2 f(x)dx - (\mu(X))^2. \text{ (Auch hier gilt die bekannte Formel} \\
\sigma^2(X) &= \mu(X^2) - (\mu(X))^2.)
\end{aligned}$$

Bemerkung: Man kann Dichtefunktionen konstruieren, bei denen das Mittelwertsintegral oder das Varianzintegral nicht existieren, aber bei praktisch wichtigen Dichten verhält sich das nicht so.

3.2. Grundlage der exakten Berechnung von bestimmten Integralen. Wie angekündigt, führt der schöne Zusammenhang

$$\overline{f_{[a,b]}} = \frac{1}{b-a} \int_a^b f(x)dx$$

zum Ziel. Dazu nehmen wir an, wir hätten eine Funktion F mit der Eigenschaft $F' = f$. An jeder Stelle x gibt f also die Steigung von F. Wir betrachten nunmehr die mittlere Steigung von F auf $[a, b]$. Das ist einerseits der bekannte Differenzenquotient:

$$\text{mittlere Steigung von } F \text{ auf } [a,b] = \frac{F(b)-F(a)}{b-a}.$$

Andererseits haben wir

$$\text{mittlere Steigung von } F \text{ auf } [a,b] = \overline{f_{[a,b]}}.$$

Das ergibt zusammen:

$$\overline{f_{[a,b]}} = \frac{F(b)-F(a)}{b-a}.$$

Daraus folgt mit

$$\overline{f_{[a,b]}} = \frac{1}{b-a} \int_a^b f(x)dx,$$

dass gilt:

SATZ 16 (Hauptsatz der Differential- und Integralrechnung). *Wenn f auf $[a,b]$ stetig („stückweise" genügt) ist und F eine Stammfunktion von f auf $[a,b]$ (d.h. $F' = f$ auf $[a,b]$), dann gilt:*

$$\int_a^b f(x)dx = F(b) - F(a), \text{ und entsprechend} \tag{3.1}$$

$$\overline{f_{[a,b]}} = \frac{F(b)-F(a)}{b-a}.$$

Weiter gilt auch noch, dass unter der gennanten Voraussetzung für f stets eine Stammfunktion F zu f existiert. (Allerdings ist es vielfach unmöglich, einen üblichen Rechenausdruck für F zu berechnen, einfach, weil ein solcher nicht existiert!)

3.3. Praktische Berechnung von bestimmten und unbestimmten Integralen: Formeln und Beispiele. Grundlage ist der Hauptsatz. Seine Anwendung besteht natürlicherweise in zwei Schritten: Zuerst sucht man eine Stammfunktion F zu f, dann setzt man die Grenzen des Integrals ein und bildet die Differenz. Zu diesem Zweck ist es nützlich, folgende (übliche!) Notation einzuführen:

$$[F(x)]_a^b = F(b) - F(a),$$

und nun rechnet man z.B.:

$$\int_1^2 x^2 dx = \left[\frac{x^3}{3}\right]_1^2 = \frac{2^3}{3} - \frac{1^3}{3} = \frac{7}{3}.$$

Hier sind zwei Formeln zum Umgang mit den Grenzen bei bestimmten Integralen. Die erste versteht sich unmittelbar aus der Flächendeutung, die zweite aus der Bedeutung der Orientierung:

$$\int_a^b f(x)dx + \int_b^c f(x)dx = \int_a^c f(x)dx \tag{3.2}$$

$$\int_a^b f(x)dx = -\int_b^a f(x)dx.$$

Ansonsten geht es im wesentlichen um das Auffinden von Stammfunktionen zu einer gegebenen Funktion. Dabei beachten wir, dass mit F auch stets $F + c$, c eine Konstante, eine Stammfunktion von f ist, dass man damit aber auch *alle* Stammfunktionen von f beschrieben hat (Forlgerung aus dem dritten Teil von Satz 15). Im folgenden Text wollen wir das ewige Wiederholen von „$+c$" vermeiden und müssen dann nur daran denken, dass wir stets nur *eine* Stammfunktion beschreiben. Dazu ist folgende Notation in Gebrauch:

$$\int f(x)dx = F(x), \text{ mit } F' = f. \text{ (Unbestimmtes Integral)}$$

Der Vorgang verläuft wie bei den Ableitungsformeln: Formeln für Grundintegrale und für zusammengesetzte Funktionen, für letztere aber wesentlich unvollständiger bzw. gestörter.

3.3.1. *Grundintegrale.* Durch Umkehren der zugehörigen Ableitungsregeln erhalten wir unmittelbar:

$$\begin{aligned} \int x^a dx &= \frac{x^{a+1}}{a+1}, \text{ für } a \neq -1, \\ \int \frac{1}{x} dx &= \ln|x|, \\ \int e^x dx &= e^x. \end{aligned} \tag{3.3}$$

Wir fügen noch hinzu, für eine Herleitung vgl. Beispiel zu Formel (3.4):

$$\int \ln(x) dx = x\ln(x) - x.$$

3.3.2. *Integrationsregeln für Zusammensetzungen von Funktionen.* Linearität des Integrals:

$$\begin{aligned} \int (f(x) + g(x))\, dx &= \int f(x)dx + \int g(x)dx, \\ \int cf(x)dx &= c\int f(x)dx. \end{aligned} \tag{3.4}$$

Damit hat man z.B.

$$\int \left(3 + 2x + 5e^x - \frac{4}{x}\right) dx = 3x + 2\frac{x^2}{2} + 5e^x - 4\ln|x|.$$

Man integriert also in Summen mit konstanten Faktoren einfach gliedweise, ähnlich wie beim Ableiten.

(Nur sehr eingeschränkt taugliche) Produktregel: „Partielle Integration":

$$\int F(x)g(x)dx = F(x)G(x) - \int f(x)G(x)dx. \tag{3.5}$$

Es bleibt also ein Integral übrig, das nur hoffentlich lösbar ist. Typisch ist das anzuwenden, wenn ein Faktor ein Polynom, der andere Exponential- oder Logarithmusfunktion (beide allenfalls leicht verändert) ist, Anwendungsbeispiel:

$$\int \ln(x)dx = \int (\ln(x)\cdot 1)dx = \ln(x)\cdot x - \int \frac{1}{x} x dx = x\ln(x) - x.$$

Es ist sehr einfach, die Regel aus der Produktregel des Differenzierens herzuleiten, sie ist einfach die Umkehrung davon:

$$\begin{aligned} (FG)' &= fG + Fg, \text{ also} \\ \int (f(x)G(x) + F(x)g(x))dx &= F(x)G(x), \text{ daher mit } \mathbf{??} \\ \int f(x)G(x)dx + \int F(x)g(x)dx &= F(x)G(x), \text{ nun erstes Integral hinüber.} \end{aligned}$$

$\frac{1}{\alpha}$-Regel für lineare Transformationen von Funktionen, zu denen man schon Stammfunktionen kennt:

$$\int f(\alpha x + \beta)dx = \frac{1}{\alpha} F(\alpha x + \beta), \text{ wobei } \alpha \neq 0 \text{ und } F' = f \text{ sind.} \tag{3.6}$$

Anwendungsbeispiele:

$$\begin{aligned}\int (2x+3)^{100}dx &= \frac{1}{2}\cdot\frac{(2x+3)^{101}}{101},\\ \int \frac{1}{2x+3}dx &= \frac{1}{2}\ln|2x+3|,\\ \int 2^x dx &= \int e^{x\ln(2)}dx = \frac{1}{\ln(2)}e^{x\ln(2)} = \frac{1}{\ln(2)}\cdot 2^x.\end{aligned}$$

Auch diese Regel bestätigt man unmittelbar durch Bildung von

$$\frac{d}{dx}\frac{1}{\alpha}F(\alpha x+\beta) = \frac{1}{\alpha}\alpha f(\alpha x+\beta) = f(\alpha x+\beta).$$

Umkehrung der Kettenregel (oder Substitutionsregel) - wir schreiben das dx hier links, das ist oft nützlich, insbesondere auch bei mehrfachen Integralen:

$$\int dx f'(x)g(f(x)) = \int du g(u) = G(f(x)),\ \text{mit } G' = g. \tag{3.7}$$

Wieder erhält man durch Ableiten unmittelbar die Bestätigung:

$$\frac{d}{dx}G(f(x)) = g(f(x))f'(x).$$

Der eingefügte Zwischenschritt ist sehr nützlich: Für $f(x)$ wird die neue Integrationsvariable u eingeführt, und man hat die Ersetzungszeilen:

$$\begin{aligned}u &= f(x),\\ du &= f'(x)dx,\ \text{gemäß } \frac{du}{dx} = f'(x).\end{aligned}$$

Nun kann man formal ersetzen:

$$\int \underbrace{dx f'(x)}_{du}\underbrace{g(f(x))}_{g(u)} = \int du g(u) = G(u) \underbrace{=}_{Rückeinsetzen} G(f(x)).$$

Anwendungsbeispiel:

$$\int dx\frac{\ln x}{x} = \int du\cdot u = \frac{u^2}{2} = \frac{1}{2}\ln^2(x),$$

wie man folgendermaßen sieht:

$$\begin{aligned}u &= \ln(x),\ g(u) = u,\\ du &= \frac{1}{x}dx.\end{aligned}$$

Weiteres Anwendungsbeispiel:

$$\begin{aligned}\int dx\cdot x\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} &= -\int dx(-x)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} = -\int du\frac{1}{\sqrt{2\pi}}e^u\\ &= -\frac{1}{\sqrt{2\pi}}e^u = -\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2},\end{aligned}$$

mit der Substitution

$$\begin{aligned}u &= -\frac{1}{2}x^2,\ g(u) = e^u,\\ du &= -x.\end{aligned}$$

Damit hat man

$$\int_0^\infty x\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx = \lim_{b\to\infty}\int_0^b x\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx = \lim_{b\to\infty} -\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}b^2} = 0,$$

ebenso ergibt sich

$$\int_{-\infty}^0 x\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx = \lim_{a\to-\infty} -\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}a^2} = 0,$$

daher ist der Erwartungswert einer standard-normalverteilten Größe

$$\int_{-\infty}^\infty x\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx = \int_{-\infty}^0 x\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx + \int_0^\infty x\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx = 0.$$

Wir haben die Gelegenheit benutzt, nebenbei zu zeigen, wie man ein Integral mit unendlicher Grenze als Grenzwert von Integralen mit endlichen Grenzen berechnet. Aber das Integral, das man hauptsächlich bei der Standard-Normalverteilung benötigt, nämlich

$$\int dx\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2},$$

gerade das kann man nicht ausrechnen, sondern nur die Werte nähern wie in Abschnitt 3.1 ausgeführt!

3.4. Anwendung der Integralrechnung auf alles Interessierende bei einer durch Dichte gegebenen Wahrscheinlichkeitsverteilung. Erstes Beispiel:

Es sei die Größe X verteilt mit der Dichte

$$f(x) = \begin{cases} \frac{1}{2\sqrt{x}} \text{ für } 0 < a \leq 1, \\ 0 \text{ sonst.} \end{cases}$$

Wir berechnen die Verteilungsfunktion von X, für Werte $0 \leq a \leq 1$:

$$F_X(a) = P(X \leq a) = \int_0^a dx\frac{1}{2\sqrt{x}} = \left[\sqrt{x}\right]_0^a = \sqrt{a}.$$

Somit haben wir insgesamt:

$$F_X(a) = \begin{cases} \sqrt{a} \text{ für } 0 \leq a \leq 1, \\ 0 \text{ für } a < 0, \\ 1 \text{ für } a > 1. \end{cases}$$

Wir berechnen den Median, das ist definitionsgemäß die Lösung der Gleichung

$$P(X \leq a) = \sqrt{a} = \frac{1}{2}.$$

Damit liegt der Median von X bei $\frac{1}{4}$.
Wir berechnen den Erwartungswert von X:

$$\mu(X) = \int_0^1 dx \cdot x \cdot \frac{1}{2\sqrt{x}} = \int_0^1 dx\frac{\sqrt{x}}{2} = \left[\frac{1}{2}\cdot\frac{x^{\frac{3}{2}}}{\frac{3}{2}}\right]_0^1 = \frac{1}{3}.$$

Wie bei der Verteilungsform zu erwarten, liegt er rechts vom Median.

Schließlich berechnen wir noch die Varianz von X:

$$\sigma^2(X) = \int_0^1 dx \cdot x^2 \cdot \frac{1}{2\sqrt{x}} - \frac{1}{9} = \int_0^1 dx \frac{x^{\frac{3}{2}}}{2} - \frac{1}{9} = \frac{1}{5} - \frac{1}{9} = \frac{4}{45}.$$

Zweites Beispiel: Die Familie der λ-Exponentialverteilungen, $\lambda > 0$:
Die zugehörige Dichte ist

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & 0 \leq x, \\ 0 \text{ sonst.} \end{cases}$$

Die zugehörigen Verteilungsfunktionen sind

$$F_\lambda(a) = \begin{cases} 1 - e^{-\lambda a}, & 0 \leq a, \\ 0 \text{ sonst.} \end{cases}$$

Denn

$$\int_0^a dx \lambda e^{-\lambda x} = \left[-\lambda \frac{1}{\lambda} e^{-\lambda x} \right]_0^a = -e^{-\lambda a} - (-e^0) = 1 - e^{-\lambda a}.$$

Median zum Parameter λ ist die Lösung a von

$$1 - e^{-\lambda a} = \frac{1}{2}, \text{das ist } a = \frac{\ln(2)}{\lambda}.$$

Der Erwartungswert zum Parameter λ ist

$$\int_0^\infty dx \cdot x \cdot \lambda e^{-\lambda x} = \left[-x \cdot e^{-\lambda x} \right]_0^\infty - \int_0^\infty dx(-e^{-\lambda x}) = 0 + \frac{1}{\lambda} = \frac{1}{\lambda}.$$

Die Varianz zum Parameter λ ist

$$\int_0^\infty dx \cdot x^2 \cdot \lambda e^{-\lambda x} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Den Wert des Integrals findet man hier, indem man zwei mal die partielle Integration anwendet, um den Faktor x^2 schließlich auf eine Konstante herunterzubringen:

$$\begin{aligned} \int dx \cdot x^2 \lambda e^{-\lambda x} &= -x^2 e^{-\lambda x} - \int dx \left(-2x e^{-\lambda x} \right) \\ &= -x^2 e^{-\lambda x} - \frac{2}{\lambda} x e^{-\lambda x} - \int dx \left(-\frac{2}{\lambda} e^{-\lambda x} \right) \\ &= -x^2 e^{-\lambda x} - \frac{2}{\lambda} x e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x}. \text{ Damit ist} \\ \int_0^\infty dx \cdot x^2 \cdot \lambda e^{-\lambda x} &= \frac{2}{\lambda^2}. \end{aligned}$$

Zur inhaltlichen Bedeutung der Exponentialverteilungen:

Gehen wir aus von einer λ-Poisson-verteilten Größe, das ist eine Trefferzahlgröße. Jedoch hat man nicht eine diskrete Zahl von Versuchen (bei binomialverteilten Größen sind es z.B. endlich viele, mit fester Anzahl n). Stattdessen hat man auf einem Kontinuum, z.B. einem endlichen Zeitintervall, eine bestimmte Zahl (nämlich λ) von Treffern *zu erwarten,* λ wird also im allgemeinen eine beliebige positive reelle Zahl sein. λ ist dann die erwartete Trefferzahl pro Zeiteinheit, sagt z.B., wie oft

im Mittel pro Minute ein Neuron „feuert". Eine Poisson-verteilte Größe hat definitionsgemäß die weitere Eigenschaft, dass es völlig unabhängig von vorangehenden Treffern oder deren Ausbleiben ist, ob in einem Augenblick ein Treffer stattfindet. (Bei realen Prozessen wie dem Feuern von Neuronen oder auch dem Fallen eines Fußballtores ist das nicht genau erfüllt, aber dennoch stellt die Poissonverteilung eine sehr gute mathematische Idealisierung dar, die sehr genaue Ergebnisse zu liefern vermag.) Für eine λ-Poisson-verteilte Größe X hat man folgende Wahrscheinlichkeitsformel:

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k \in \mathbb{N} \text{ (einschließlich 0)}.$$

Man beachte, dass auf dem Kontinuum beliebig viele Treffer passieren können, weshalb es keine obere Begrenzung für k gibt. Lediglich werden sehr hohe Trefferzahlen immer unwahrscheinlicher, aber stets bleibt auch bei kleinem λ diese Wahrscheinlichkeit größer als Null. Weiter kommen nur natürliche Zahlen als Werte von X in Frage, die Größe X ist diskret. Nun ist mit jeder solchen λ-Poisson-verteilten Größe X eine weitere Größe verbunden, die λ-exponentialverteilt ist. Es handelt sich um Y = Wartezeit bis zur Beobachtung des ersten Poisson-Treffers. Dabei ist zu verstehen, dass die Zeit erst einmal unbeschränkt ist (bei X hatten wir eine feste Zeitspanne). Man lässt den Prozess im einzelnen Versuch so lange laufen, bis der der erste Treffer kommt, und beobachtet die benötigte Zeit. Das ist der Wert der Größe Y. Diese Größe Y ist nun λ-exponentialverteilt. Erwartet man also pro Fußballspiel (90 Min.) 4 Tore, dann ist $\lambda = 4/90$ (erwartete Trefferzahl pro Zeiteinheit), und die Wahrscheinlichkeit dafür, dass man höchstens 10 Minuten warten muss, bis man das erste Tor sieht, hat den Wert:

$$\int_0^{10} e^{-\frac{4}{90}x}dx = \left[1 - e^{-\frac{4}{90}x}\right]_0^{10} = 1 - e^{-\frac{4}{9}} \approx 0.36.$$

Allgemein fassen wir noch einmal zusammen:

Die Wahrscheinlichkeit dafür, dass die Wartezeit bis zum ersten λ-Poissontreffer höchstens t beträgt, lautet:

$$\int_0^t dx e^{-\lambda x} = 1 - e^{-\lambda t}.$$

Das wichtigste Beispiel für ein durch Dichte gegebene Verteilungen sind natürlich die Normalverteilungen. Man kann zu (μ, σ) wirklich nachrechnen, dass die Integrale für Mittelwert und Varianz μ und σ^2 ergeben. Aber prinzipiell lässt sich die Verteilungsfunktion nicht ausrechnen zu einem einfachen Ausdruck, der weder ein Integral, noch eine unendliche Reihe enthält. Dafür hat man die numerische Tabelle als Ersatz! Weitere wichtige Dichten (wieder ganze Familien) sind: Die χ^2-Verteilungen mit Freiheitsgraden 1,2,3,... (das ist der einfache Parameter dabei), ebenso die einparametrige Schar der t-Verteilungen (wieder heißt der Parameter „Freiheitsgrade" und ist eine natürliche Zahl ≥ 1), schließlich die zweiparametrige Schar (dabei handelt es sich um ein Paar von Freiheitsgraden) der F-Verteilungen. (Man beachte, dass „F" in diesem Kontext ein Eigenname ist, ebenso wie χ^2, t und „normal".) Damit sind auch schon alle Verteilungen aufgezählt, die man in Standard-Tests verwandt findet, und in allen Fällen handelt es sich um zu schwierige

Integrale, so dass man Näherungswerte aus Tabellen (oder alternativ vom Computer geliefert) benötigt. Eine Ausnahme macht da nur die χ^2-Verteilung mit *genau zwei* Freiheitsgraden, das ist dasselbe wie die Exponentialverteilung mit $\lambda = 1$.

3.5. Zur Herleitung der Poissonverteilungen und der zugehörigen Exponentialverteilungen. Dafür benötigt man deutlich die Mittel der reellen Analysis, so dass wir auch für die diskreten Poissonverteilungen erst jetzt gerüstet sind. Dafür genügen einfache Grenzwertbetrachtungen, während wir für die Exponentialverteilungen, genauer für deren Verteilungsfunktionen, eine Differentialgleichung aufstellen werden anhand einer Wahrscheinlichkeitsbetrachtung, anschließend wird uns die Integralrechnung lehren, wie wir diese Differentialgleichung lösen können.

Zu den Poissonverteilungen:

Die Idee ist einfach die einer Approximation über unsere bekannten Binomialverteilungen: Stellen wir uns das kontinuierliche Medium (sagen wir der Einfachheit halber: die Zeit) in n gleichlange kleine Stückchen zerhackt vor: Dann können wir sagen, die Zahl der $\lambda-$ Poissontreffer sei etwa dasselbe wie die Anzahl der Binomialtreffer bei n Versuchen mit Einzeltrefferwahrscheinlichkeit p, so dass $\lambda = np$ gilt. Die Idee ist also, $n \to \infty$ und gleichzeitig $p \to 0$ derart gehen zu lassen, dass $np = \lambda$ dabei konstant bleibt mit dem beliebig vorgelegten Wert $\lambda > 0$. Dann wird die Wahrscheinlichkeit, genau k Poissontreffer zu erhalten (bei erwarteten λ), der folgende Grenzwert sein:

$$P(X = k) = \lim_{n\to\infty,\ p\to 0,\ np=\lambda} \binom{n}{k} p^k (1-p)^{n-k},\ \text{für } X\ \lambda - Poisson - verteilt.$$

Diesen Grenzwert rechnen wir nun einfach aus: Ersetzen von p durch λ/n ergibt:

$$\begin{aligned}\binom{n}{k} p^k (1-p)^{n-k} &= \frac{n!}{k!\,(n-k)!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} = \frac{n!}{(n-k)!n^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdot\ldots\cdot(n-k+1)}{n^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k}.\end{aligned}$$

Nunmehr beobachten wir, dass (beachte: k ist ein beliebiger fester Wert, während wir nun n nach Unendlich gehen lassen wie beschrieben - das p sind wir los) der erste Faktor gegen 1 geht für $n \to \infty$. Begründung: Jeder Term $(n-r)/n = 1 - r/n$ geht offenbar nach Eins für $n \to \infty$, für $r = 0, 1, ..., k-1$. Damit geht auch das Produkt von den vorhandenen k derartigen Faktoren nach Eins. Die zweite Beobachtung ist:

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-k} = e^{-\lambda}.$$

Dazu bilden wir

$$\ln\left(\left(1-\frac{\lambda}{n}\right)^{n-k}\right) = (n-k)\ln\left(1-\frac{\lambda}{n}\right)$$

und schreiben nunmehr für $\ln\left(1-\frac{\lambda}{n}\right)$ die Tangentenzerlegung mit $x_0 = 1$ hin - ln ist an der Stelle 1 differenzierbar! Das sieht so aus:

$$\begin{aligned}(n-k)\ln\left(1-\frac{\lambda}{n}\right) &= (n-k)\left(\ln(1)+1\cdot\left(-\frac{\lambda}{n}\right)+R\left(-\frac{\lambda}{n}\right)\right)\\ &= \frac{(n-k)(-\lambda)}{n}+(n-k)\,R\left(-\frac{\lambda}{n}\right).\end{aligned}$$

Man beachte: $\ln(1) = 0$. Schauen wir die Summanden an: Der erste geht für $n \to \infty$ nach $-\lambda$, der zweite nach Null; denn das Restglied $R\left(-\frac{\lambda}{n}\right)$ geteilt durch $-\lambda/n$ geht nach Restgliedeigenschaft der Tangentenzerlegung nach Null, also haben wir

$$\lim_{n\to\infty}\frac{R\left(-\frac{\lambda}{n}\right)}{-\frac{\lambda}{n}} = \lim_{n\to\infty}\frac{nR\left(-\frac{\lambda}{n}\right)}{-\lambda} = 0.$$

Damit gilt aber auch

$$\lim_{n\to\infty} nR\left(-\frac{\lambda}{n}\right) = 0, \text{ und erst recht } \lim_{n\to\infty}(n-k)R\left(-\frac{\lambda}{n}\right) = 0.$$

Nun erinnern wir uns, dass wir suchten: $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-k}$. Da die Logarithmen gegen $-\lambda$ gehen, ist dieser Limes also gleich $e^{-\lambda}$. (Dazu argumentieren wir einfach mit der Stetigkeit der Exponentialfunktion.) Fassen wir zusammen, so haben wir die versprochene Wahrscheinlichkeitsformel für alle Poissonverteilungen:

$$P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}, \text{ für } X\ \lambda - Poisson - verteilt.$$

Herleitung der Exponentialverteilungsdichten (und zugehörigen Verteilungsfunktionen):

Sei ein beliebiger Poissonparameter $\lambda > 0$ vorausgesetzt. Sei X die Wartezeit bis zum ersten Treffer. Bezeichnen wir mit F die *Verteilungsfunktion* von X. Dann haben wir folgende Gleichung:

$$F(x+\Delta x) - F(x) = (1-F(x))\cdot\left(1-e^{-\lambda\Delta x}\right).$$

Denn auf der linken Seite steht die Wahrscheinlichkeit dafür, dass wir in der Zeit von 0 bis x keinen Treffer beobachten *und* in der Zeit von x bis $x+\Delta x$ *mindestens* einen Treffer beobachten. Wegen der für die Poissontreffer strukturell vorausgesetzten Unabhängigkeit sind diese zu multiplizieren. Aber die erste ist gerade $1-F(x)$, da $F(x) = P(X \le x)$. Und die zweite ist gerade die Wahrscheinlichkeit dafür, dass wir *nicht keinen* Poissontreffer in der Zeitspanne Δx beobachten, das ist aber nach Poissonformel gerade $1-e^{-\lambda\Delta x}$, wozu wir nur zu beachten haben, dass die erwartete Trefferzahl für die Zeitspanne Δx gerade $\lambda\Delta x$ ist.

Unser Plan ist es nun, die gewonnene Gleichung durch Δx zu teilen und anschließend Δx nach Null gehen zu lassen. Dann steht auf der linken Seite $F'(x)$. Daher interessiert uns:

$$\lim_{\Delta x\to 0}\frac{1-e^{-\lambda\Delta x}}{\Delta x} = -\lim_{\Delta x\to 0}\frac{e^{-\lambda\Delta x}-1}{\Delta x} = \lambda.$$

Denn

$$\frac{e^{-\lambda\Delta x}-1}{\Delta x} = \frac{e^{-\lambda\Delta x}-e^{-\lambda\cdot 0}}{\Delta x}$$

ist Differenzenquotient für die Funktion $g(x) = e^{-\lambda x}$ an der Stelle $x_0 = 0$, und der Limes für $\Delta x \to 0$ ist einfach $g'(0) = -\lambda e^{-\lambda \cdot 0} = -\lambda$. Damit haben wir, Δx nach Null gehen lassend:

$$F'(x) = (1 - F(x))\,\lambda.$$

Dies ist eine Differentialgleichung für $F(x)$. Sie hat unendlich viele Lösungen, aber wir wissen, dass $F(0) = 0$ sein muss, und daraus erhalten wir eine eindeutige Lösung wie folgt: Aus der Differentialgleichung folgt:

$$\int_0^a \frac{F'(x)}{1 - F(x)} dx = \int_0^a \lambda = \lambda a.$$

Das Integral auf der linken Seite berechnen wir so:

$$\int_0^a \frac{F'(x)}{1 - F(x)} dx = \int_{F(0)}^{F(a)} \frac{du}{1 - u} = -[\ln(1 - u)]_0^{F(a)} = -\ln(1 - F(a)).$$

Wir benutzten $F(0) = 0$ und $0 \leq F(a) < 1$ (so dass wir die Betragstriche bei der Stammfunktion weglassen konnten). Zusammen haben wir

$$\begin{aligned} \ln(1 - F(a)) &= -\lambda a, \text{ also} \\ F(a) &= 1 - e^{-\lambda a} \ (a \geq 0). \end{aligned}$$

Die Ableitung ist $f(a) = F'(a) = \lambda e^{-\lambda a}$, wieder für $a \geq 0$.

KAPITEL 6

Regression und Korrelation

1. Exakte funktionsartige Zusammenhänge zwischen Variablen, linear und nichtlinear

Eine sprachliche Vorbemerkung ist notwendig: Der Begriff der Unabhängigkeit tritt in der Mathematik auf vielfältigste Weise auf, und bisher verwandten wir ihn bereits häufig in zwei *völlig verschiedenen Bedeutungen:* Variablen X, Y (auch mehr noch als zwei) heißen (*im statistischen Sinne*) unabhängig, wenn stets gilt, dass $P(X \leq a \text{ und } Y \leq b \text{ und...}) = P(X \leq a) \cdot P(Y \leq b) \cdot ....$ Dagegen abzugrenzen ist der Sprachgebrauch, dass bei Funktionen die unabhängige Variable für ein beliebiges Element des Definitionsbereiches steht und die abhängige Variable der davon exakt abhängige Funktionswert ist. Nun ist es eine Besonderheit gerade dieses Kapitels, dass *beide Bedeutungen* ständig nebeneinander vorkommen. Das liegt genau darin begründet, dass hier ins Auge gefasst wird, eine Variable X im Sinne der Statistik als unabhängige Variable und eine andere statistische Variable Y als abhängige Variable, also $Y = f(X)$ aufzufassen, mindestens näherungsweise. Daher werden wir stets ausführlicher die statistische Unabhängigkeit mit dem Adjektiv „statistisch" versehen. Steht bloß „unabhängig", so ist das einfach im Sinne von „unabhängige Variable" gemeint. Ferner tritt besonders auch die „lineare" Unabhängigkeit von Variablen auf, was stets die Abschwächung der statistischen Unabhängigkeit bedeutet (bei zwei Variablen X, Y: $Cov(X, Y) = 0$).

Wir setzen stets voraus, dass wir so etwas wie einen funktionalen Zusammenhang zwischen zwei Variablen X und Y betrachten, die *denselben* Definitionsbereich Ω haben. Wir haben gesehen, dass solche Variablen *statistisch unabhängig* sein können, dann ergibt die Kenntnis des X-Wertes keinerlei Information über den Y-Wert. Antipodisch dazu steht der Fall, dass der Y-Wert durch den X-Wert *eindeutig* bestimmt ist; das heißt: Es gibt eine Funktion f, so dass stets gilt: Y-Wert $= f(X\text{-Wert})$. Wir können damit X als unabhängige Variable, Y als abhängige Variable auffassen. (Dies ist der einfachste Fall, in komplexeren Situationen wird man stets mehrere unabhängige und auch abhängige haben.) Also genau: Für alle $\omega \in \Omega$ gilt: $Y(\omega) = f(X(\omega))$. Kürzer formuliert man das gern so: $Y = f(X)$, exakter müsste es „$Y = f$ hinter X geschaltet" heißen. f ist dabei eine Funktion $\mathbb{R} \to \mathbb{R}$, und sie rechnet die Y-Werte aus den X-Werten aus, man kann sie nicht etwa auf eine Variable wie X *anwenden.* Man beachte weiter, dass eine solche funktionale Abhängigkeit eine *Richtung* hat; gibt es eine in der einen Richtung, so gibt es nicht zwangsläufig auch eine für die andere Richtung. Das hängt daran, ob die Funktion im relevanten Gebiet umkehrbar ist.

Beispiele für exakte funktionale Abhängigkeit:

$X =$ Preis in DM, $Y =$ Preis in Dollar (bei einem festgelegten Wechselkurs). In diesem Falle gilt $Y = c \cdot X$ [detaillierter gesagt noch einmal: Für jede Ware ω gilt: $Y(\omega) = f(X(\omega))$], der Zusammenhang wird vermittelt durch eine lineare

Funktion, sogar Proportionalität. Mit dem X-Wert liegt der Y-Wert eindeutig fest. Ein ähnliches Beispiel: X = Temperatur in Celsius, Y = Temperatur in Fahrenheit. Wieder hat man eine lineare Funktion, allerdings mit einer additiven Konstanten.

Ein nichtlinearer (aber auch exakter funktionaler) Zusammenhang: Y = zurückgelegter Weg bei einem freien Fall, X = abgelaufene Zeit nach „Loslassen". Hier hat man $Y = \frac{1}{2}X^2$. Das ist noch extrem einfach, es gibt wichtige wesentlich komplexere funktionale Zusammenhänge.

Wir wollen in der Statistik jedoch auf *inexakte* (aber dennoch „ungefähr" funktionale) Zusammenhänge hinaus, und so etwas besprechen wir im Prinzip zu Beginn des Abschnittes 2., im Einzelnen für den einfachsten linearen Fall in 2.1, verallgemeinert in 2.2. Zum Verständnis ist es jedoch wichtig, erst einmal zu wissen, wie ein exakter funktionaler Zusammenhang aussieht. Das haben wir für den Fall einer unabhängigen und abhängigen Variablen besprochen. Wenigstens ein Beispiel wollen wir noch dafür bringen, dass man zwei unabhängige Variablen hat:

Population: Alle rechtwinkligen Dreiecke, X_1 = Länge der ersten Kathete (unwichtig, wie man das festlegt, für unser Beispiel ist das auch gleichgültig), X_2 = Länge der zweiten Kathete, Y = Länge der Hypotenuse. Dann können wir nach dem Satz des Pythagoras Y als Funktion von X_1 und X_2 folgendermaßen ausrechnen: $Y = X_1^2 + X_2^2$. Die Hypotenusenlänge (als abhängige Variable) ist also eine nichtlineare Funktion beider Kathetenlängen (als unabhängiger Variablen). Ein solcher Zusammenhang ist stets theoretischer Art, und man nennt so etwas daher auch „Modell". Das Modell des funktionalen Zusammenhangs lässt sich in einer Gleichung ausdrücken, und daher findet man zuweilen den etwas irreführenden und unglücklichen Ausdruck, diese Gleichung sei das Modell. Man lasse sich aber nicht verleiten, zu glauben, diese simplen Modelle seien alles, was die Mathematik an Modellen zu bieten hätte! Aber selbst beim funktionalen Zusammenhang handelt es sich um ein „Modell" durchaus in einem tieferen Sinne: Es wird mit einer mathematischen Rechenoperation aus den Werten der Einflussvariablen ein Wert produziert, den etwa die Natur unter den Bedingungen, welche durch die Werte der Einflussvariablen beschrieben werden, auf eine ganz andere Weise produziert, mit unendlichen Komplikationen. Es ist eine gewaltige metaphorische Übertragung, wenn man sagt, die Natur „rechne".

Wir fassen noch einmal zusammen:

Gleichung für das Modell eines funktionalen Zusammenhangs zwischen einer abhängigen Variablen Y und unabhängigen Variablen $\mathbf{X_1}, ..., \mathbf{X_n}$:

$$\mathbf{Y} = \mathbf{f}(\mathbf{X_1}, ..., \mathbf{X_n}) \tag{1.1}$$

Man beachte: Es ist nichts über die Art von f gesagt, diese Funktion könnte beliebig einfach oder kompliziert sein, und sie ist stets nur im konkreten Einzelfall gegebener Variablen zu spezifizieren, beispielsweise lautet das *lineare* Modell mit nur einer unabhängigen Variablen: $Y = aX + b$.

In jedem Einzelfall stellen sich folgende Grundfragen: Gibt es überhaupt einen derartigen Zusammenhang zwischen den interessierenden Variablen, und wenn ja, welcher Art ist er dann, welche Funktion f beschreibt diesen Zusammenhang korrekt (oder doch wenigstens: genau genug - dann kommen mehrere verschiedene Funktionen in Frage, die sich durchaus zuweilen nach rationalen Prinzipien auswählen lassen)? (Diese Grundfragen werden sich bei Betrachtung *inexakter* funktionaler Zusammenhänge ein wenig modifizieren.) Diese Grundfragen sind je nach

theoretischem oder empirischem Zusammenhang in völlig verschiedenartigen Weisen anzugehen, das kann mathematisch-theoretisch deduzierend sein, aber in empirischer Wissenschaft auch stark anhand gegebenen Datenmaterials vorgehen (dazu mehr im nächsten Abschnitt).

Abschließend bemerken wir noch, dass man im Zusammenhang mit solchen Modellen gern die unabhängigen Variablen auch „Prädiktorvariablen" oder einfach „Prädiktoren" nennt, das heißt „voraussagende Variablen" - man denkt dabei an den Zweck, den Wert einer interessierenden „Zielvariablen" (so nennt man dann gern die abhängige Variable) *vorauszusagen*, wenn man die Werte der Prädiktorvariablen kennt. Beispiel: Man möchte aus sozialwissenschaftlichen Daten über die soziale Umgebung eines Kindes dessen späteren Berufserfolg etc. voraussagen. So etwas geht natürlich nicht exakt, aber immerhin kann man einige Variablen als wesentliche EinflussVariablen erkennen und auch genauer erweisen. (Dieser Aspekt wird im nächsten Abschnitt vertieft.)

2. Inexakte (statistische) funktionsartige Zusammenhänge

Es sollte naheliegen, die Inexaktheit einfach dadurch ins Spiel zu bringen, dass man aus der Modellgleichung 1.1 eine Ungefähr-Gleichung macht, also formuliert:

$$\mathbf{Y \approx f(X_1, ..., X_n)}.$$

Dies trifft jedoch nicht stets das Gewünschte, einerseits möchte man spezifizieren, wie gut denn das „ungefähr" ist (auf manchen Gebieten für mancherlei Zwecke verlangt man sehr kleine Abweichungen), welche Fehler man also zu erwarten hat, andererseits möchte man zumal auf Gebieten, in denen die meisten Einflussvariablen auf eine Zielvariable unbekannt sind und bleiben, z.B. auf dem der Psychologie oder Sozialwissenschaft, vernünftigerweise nicht gleich darauf hinaus, einen Wert aus Werten von Einflussvariablen gleich mit guter Genauigkeit zu reproduzieren, sondern man gibt sich durchaus mit einem großen Fehler zufrieden („ungefähr" ist dann viel zu viel gesagt). Aber man möchte dann wenigstens wissen und beschreiben, dass die betrachteten Einflussvariablen den Wert der Zielvariablen in *nennenswertem Maße* erklären, also einen guten Teil davon. ein wenig bemerkenswert, aber durchaus typisch: Die Mathematik stellt ein Mittel bereit, das gleichermaßen geeignet für beide so verschieden erscheinede Zwecke ist. Der Kunstgriff ist sehr einfach und besteht darin, aus der „ungefähr"-Gleichung eine genaue Gleichung zu machen und das nicht Aufgehende gesondert als damit implizit definierte Fehlervariable oder Rest aufzuführen (statistisches Gegenstück zur Formel 1.1):

$$\mathbf{Y = f(X_1, ..., X_n)+E}. \tag{2.1}$$

(E für „error", d.h. Fehler.) E ist also *definitionsgemäß* die Differenz $Y - f(X_1, ..., X_n)$. Die Fehlervariable zu beschreiben, das heißt die Qualität des inexakten funktionalen Zusammenhangs zu beschreiben. Der exakte Fall ist gleich mit enthalten: Dabei ist einfach die Fehlervariable die Konstante mit Wert Null. Wir wollen nunmehr sehen, wie man sowohl für hohe Genauigkeit als auch für geringe etwas Substantielles über E aussagen kann. Im ersteren Fall würde man etwa als Resultat befriedigend finden, der Wert von E betrage stets (oder mit spezifizierter

hoher Wahrscheinlichkeit) nur ein Prozent (oder gar ein Tausendstel etc.) vom Sollwert, dem Y-Wert. Im letzteren Falle, z.B. in der Sozialwissenschaft oder Psychologie, formuliert man typisch Resultate der Form: „Der Anteil der Varianz der Zielvariablen Y, der durch den funktionalen Zusammenhang erkärt wird (also durch den Wert $f(X_1, ..., X_n)$), beträgt ...". Ebenso gut kann man das ausdrücken durch den Anteil, den die Varianz von E an der Varianz von Y hat, also durch den Quotienten $\sigma^2(E)/\sigma^2(Y)$. (Genauer ist dies dann der Fall, wenn die Variablen $f(X_1, ...X_n)$ und E linear unabhängig sind, so dass gilt: $\sigma^2(Y) = \sigma^2(f(X_1, ..., X_n)) + \sigma^2(E)$.) Somit lohnt es in beiden Fällen, die Eigenschaften von E zu beschreiben, um die Qualität der versuchten funktionalen Erklärung zu erfassen. Nun stellen wir das Grundproblem konkreter für den Fall einer empirischen Wissenschaft, die in der Situation steht, dass man keine tieferen theoretischen Mittel zur Hand hat, sondern sich stark an gegebene empirischen Daten halten muss:

Grundsituation und Grundproblem:

Man hat empirische Daten der Form $(x_1^{(i)}, ...x_n^{(i)}, y^{(i)})$, $1 \leq i \leq n$, d.h. n Beispiele für Werte der Prädiktoren und den zugehörigen Wert der Zielvariablen. Gesucht ist eine Funktion f, die für die gegebenen Beispiele (die gegebene Stichprobe) den besten funktionalen Zusammenhang darstellt. Dies Problem ist durchaus nicht ganz einfach, vielmehr ist man zunächst darauf angewiesen, aus der Durchsicht der Daten eine günstige Klasse möglicher Funktionen, gewöhnlich mit Parametern definiert, auszusuchen und auszuprobieren. Niemals weiß man (sofern man nicht über tiefere theoretische Prinzipien verfügt), ob es noch bessere gibt oder nicht. Allerdings verhilft die Mathematik dann zu zweierlei: Erstens kann man mit mathematischer Extremwertrechnung die optimale Funktion aus der anvisierten Klasse aussondern (tatächlich ausrechnen), zweitens dann beurteilen, wie weit man damit gekommen ist. Reicht das Resultat für die gegebenen Zwecke aus, so ist es gut, andernfalls hat man immerhin die Möglichkeit, sich eine günstigere Klasse von Modellfunktionen auszudenken. Man kommt also auch zu einem vernünftigen Urteil darüber, was die ausprobierte Klasse taugt.

Nunmehr wenden wir uns der Lösung des Optimalitätsproblems speziell für eine *lineare* Funktion f zu und beschränken uns zunächst auf eine einzige unabhängige Variable. (Für Verallgemeinerungen vgl. Abschnitt 2.2.)

2.1. Linearer Zusammenhang zwischen zwei Variablen: Regressionsgerade. Wir betrachten hier speziell das lineare Modell (also mit linearer Funktion f) für nur eine einzige Prädiktorvariable, die zugehörige Gleichung lautet:

$$Y = aX + b + E. \tag{2.2}$$

Wir setzen in den folgenden Abschnitten *generell voraus*, dass $\sigma^2(Y) \neq 0$ und $\sigma^2(X) \neq 0$. Wäre nämlich $\sigma^2(Y) = 0$, so wäre Y eine Konstante, und wir können Y perfekt durch diese Konstante b mit $a = 0$ wiedergeben, das ganze Problem wäre ins völlig Uninteressante trivialisiert. Wäre dagegen $\sigma^2(X) = 0$, so wäre X eine Konstante, ohne Varation, und der Wert von X könnte keinerlei Information über variierende Werte von Y geben, es wäre also von vornherein unmöglich, Y auch nur zu einem kleinen Teil als Funktion von X vorauszusagen, unser Projekt wäre von vornherein gescheitert.

Es seien X und Y nun zwei beliebige Variablen. Wie sollte man dann die Parameter a und b optimal wählen, so dass von der Variation der Y-Werte so

viel wie möglich durch die lineare Funktion von X erklärt wird, der Fehler E also minimalisiert? Wie wäre „Optimum" präziser zu fassen? Ein sehr einfacher Ansatz dazu besteht darin, folgende geringfügigen und selbstverständlichen Forderungen an die Wahl von a und b zu stellen:

$$\begin{aligned} 1)\ \mu(E) &= 0 \\ 2)\ Cov(aX+b,E) &= 0. \text{ (Für } a \neq 0 \text{ gleichwertig: } Cov(X,Y)=0.) \end{aligned}$$

Erinnerung: $Cov(X,Y) = \mu[(X-\mu(X))\cdot(Y-\mu(Y))]$.

Zur Selbstverständlichkeit: Wäre $\mu(E) \neq 0$, so könnte man sofort den Mittelwert $\mu(E)$ der Konstanten b zuschlagen und hätte dann eine neue Fehlervariable mit Mittelwert 0. Die zweite Forderung ist deswegen vernünftig, weil eben der ganze lineare Zusammenhang zwischen X und Y in $aX+b$ enthalten sein sollte, eine weitere lineare Abhängigkeit zwischen E und $aX+b$ (gleichwertig zwischen E und X, wenn nur $a \neq 0$) also nicht mehr bestehen sollte. Allerdings werden wir erst mit dem hier herzuleitenden Resultat diese Bedeutung der Forderung 2) an die Kovarianz erkennen. Dagegen wissen wir bereits, dass 2) die Bedeutung einer Abschwächung der *statistischen* Unabhängigkeit zwischen E und X besitzt. Erstaunlich ist es, dass diese schwachen Forderungen genügen, um bereits a und b eindeutig zu bestimmen! Später werden wir auch sehen, in welchem Sinne damit die Fehler minimalisiert werden und die mittels 1), 2) ausgesonderte Lösung optimal ist.

Wir folgern nunmehr aus 1) und 2), wie a und b aussehen müssen:

Kovarianz Null bedeutet, dass Summenbildung und Varianz verträglich sind (vgl. den Abschnitt über das Rechnen mit Mittelwerten und Varianzen, den wir auch für weitere Rechnungen hier stark benutzen müssen), also haben wir:

$$\sigma^2(Y) = \sigma^2(aX+b) + \sigma^2(E) = \sigma^2(aX+b) + \sigma^2(Y-aX-b). \text{ (Def. von } E!)$$

Nun gilt ganz allgemein, dass $\sigma^2(X) = \mu((X-\mu(X))^2) = \mu(X^2) - \mu^2(X)$ (dabei ist $\mu^2(X) = (\mu(X))^2$), auch hat man das Analogon dazu für die Kovarianz: $Cov(X,Y) = \mu(XY) - \mu(X)\mu(Y)$, also erhalten wir unter Benutzung der Tatsache, dass Addition einer Konstanten zu einer Variablen nichts an der Varianz ändert:

$$\begin{aligned} \mu(Y^2) - \mu^2(Y) &= a^2\sigma^2(X) + \mu((Y-aX)^2) - \mu^2(Y-aX) \\ &= a^2\mu(X^2) - a^2\mu^2(X) + \mu(Y^2) - 2a\mu(XY) + a^2\mu(X^2) \\ &\quad -\mu^2(Y) + 2a\mu(X)\mu(Y) - a^2\mu^2(X) \\ &= \mu(Y^2) - \mu^2(Y) \\ &\quad -2a\mu(XY) + 2a\mu(X)\mu(Y) + 2a^2\mu(X^2) - 2a^2\mu^2(X) \\ &= \mu(Y^2) - \mu^2(Y) - 2aCov(X,Y) + 2a^2\sigma^2(X). \end{aligned}$$

Damit resultiert die Gleichung

$$a = \frac{Cov(X,Y)}{\sigma^2(X)}. \text{ (Wir setzten } \sigma^2(X) \neq 0 \text{ voraus!)}$$

Außerdem ergibt die erste Forderung sofort

$$\mu(E) = \mu(Y-aX-b) = \mu(Y) - a\mu(X) - b = 0,$$

also

$$b = \mu(Y) - a\mu(X).$$

Bemerkung: Eleganter wird die Rechnung, wenn man sie unter der Voraussetzung $\mu(X) = \mu(Y) = 0$ ausführt und von da aus theoretisch bequem auf den allgemeinen Fall schließt, was dem Anfänger jedoch eine zusätzliche Schwierigkeit bedeuten kann.
Damit haben wir die Parameter der (besten, s.u. Abschnitt 2.1.3) linearen Funktion zur Voraussage der Y-Werte aus den X-Werten bestimmt. Man nennt die zugehörige Gerade **Regressionsgerade**, und die soeben berechneten Parameter heißen **Regressionsparameter**. Um die Richtung der Voraussage mit zu bezeichnen, notiert man auch gern in unseren Rechenergebnissen (der Punkt dient hier nur zur Trennung und kann auch fehlen, man denke hier nicht an „mal"!):

$$a_{Y\cdot X} = \frac{Cov(X,Y)}{\sigma^2(X)},\ b_{Y\cdot X} = \mu(Y) - a\mu(X). \tag{2.3}$$

Man beachte, dass für die andere Voraussagerichtung (von Y auf X) *andere* Werte $a_{X\cdot Y}$, $b_{X\cdot Y}$ herauskommen, natürlich mit den analogen Formeln, die durch Vertauschen der Buchstaben X, Y entstehen.

Eine erste Bemerkung zur Interpretation des Resultats für a: Wenn $a = 0$ herauskommt, so bedeutet dass: Die Variation von X trägt linear nichts zur Erklärung der Variation von Y bei, d.h. X und Y sind linear unabhängig. Aber $a = 0$ bedeutet nach unserer Formel, dass $Cov(X,Y) = 0$. Damit haben wir erklärt, warum diese Kovarianzbedingung mit Recht „lineare Unabhängigkeit" genannt wird.

2.1.1. *Der Korrelationskoeffizient $\rho(X,Y)$ als Maß für die Stärke des linearen Zusammenhangs zwischen X und Y.* Wir schauen nach der Qualität der linearen Voraussage der Y-Werte aus den X-Werten, d.h. danach, wie hoch die Varianz des Fehlers, $\sigma^2(E)$, relativ zur Varianz der Zielvariablen Y ist. Wieder ist unsere Varianzzerlegung gemäß Forderung 2) entscheidend:

$$\sigma^2(Y) = \sigma^2(aX + b) + \sigma^2(E) = a^2\sigma^2(X) + \sigma^2(E).$$

Wir dividieren diese Gleichung durch $\sigma^2(Y)$ (das ist nach Voraussetzung $\neq 0$) und erhalten, indem wir den ausgerechneten Wert für a einsetzen:

$$1 = \frac{Cov^2(X,Y)\sigma^2(X)}{\sigma^2(Y)\sigma^4(X)} + \frac{\sigma^2(E)}{\sigma^2(Y)} = \frac{Cov^2(X,Y)}{\sigma^2(Y)\sigma^2(X)} + \frac{\sigma^2(E)}{\sigma^2(Y)}.$$

Hier tritt im ersten Summanden eine interessante Größe auf, durch die man offenbar den Anteil der Fehlervarianz an der Varanz von Y ausdrücken kann:

Definition 23 (Korrelationskoeffizient zweier Variablen). *Der Korrelationskoeffizient der Variablen X, Y mit nichtverschwindenden Streuungen, bezeichnet mit $\rho(X,Y)$, ist definiert als*

$$\rho(X,Y) := \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{2.4}$$

Damit haben wir folgende Zusammenhänge: $\rho^2(X,Y)$ (der erste Summand oben) ist der Anteil der Varianz der Zielvariablen Y, welcher durch lineare Regression auf X erklärt ist, und entsprechend ist

$$1 - \rho^2(X,Y) = \frac{\sigma^2(E)}{\sigma^2(Y)}. \tag{2.5}$$

der Anteil der Varianz von Y, der *nicht* durch den *linearen* Zusammenhang aus X erklärt wird.

Folgerung: Es gilt stets

$$-1 \leq \rho(X,Y) \leq 1.$$

Denn in der vorigen Formel steht auf der rechten Seite eine Zahl ≥ 0, somit muss $\rho^2(X,Y) \leq 1$ gelten, also kann der Korrelationskoeffizient nur Werte im Bereich $[-1,1]$ annehmen.

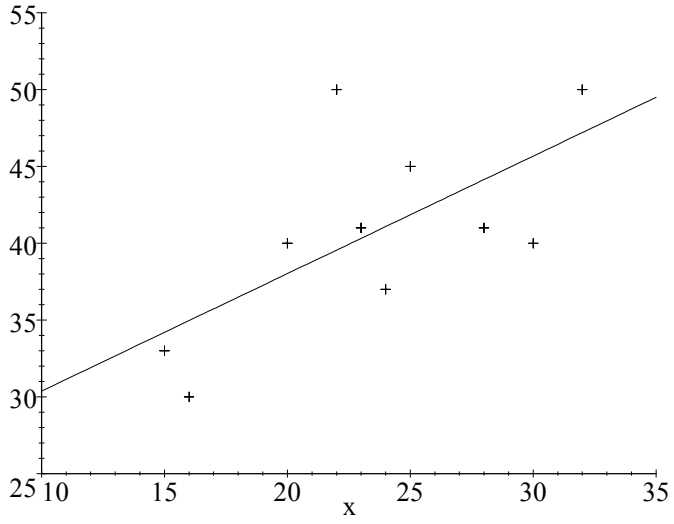
Übrigens kann man (mit kleiner Ungenauigkeit) $\sigma(E)$ als mittleren Fehler der Voraussage der Y-Werte durch die X-Werte interpretieren.

Beispiel:

Wir nehmen eine Gruppe von 10 Studenten als Gesamtpopulation (wir reden also von μ, ρ, a, b statt von Schätzwerten $\overline{x}, \overline{y}, r, \widehat{a}, \widehat{b}$ - vgl. zu diesen den nächsten Abschnitt 2.1.2), die Variable X sei die Punktezahl in einer Statistikklausur, die Variable Y die Punktezahl in einer Methodenklausur. Die zugehörige Wertetabelle aus den Wertepaaren (X-Wert, Y-Wert) („bivariat" nennt man das in gewissen Kreisen gern, „multivariat", wenn es noch mehr als zwei sind, also bei mehreren Prädiktoren) sehe so aus:

$(20,40), (25,45), (16,30), (30,40), (22,50), (15,33), (28,41), (32,50), (23,41), (24,37)$.

Zunächst sollte man den zugehörigen Punkteschwarm in der (X,Y)-Ebene aufzeichnen, das ergibt die Kreuze in folgender Graphik:



Man sieht im Beispiel, dass es Leute gab, die bei der einen Klausur relativ gut und bei der anderen nicht so gut waren, aber immerhin ist ein linearer Trend deutlich erkennbar (in vergleichbaren Beispielen kommt übrigens meist ein noch

viel besserer linearer Zusammenhang als hier heraus). Wir können folgende Werte ausrechnen - das geht auch schon bequem mit jedem Taschenrechner, der Paardaten aufzunehmen fähig ist):

$$\rho = 0.65692.$$

(Vielfach kann man 0.8 und mehr bei solchen Beispielen sehen.) Das ist immerhin etwas. Der Anteil der Varianz von Y, der durch lineare Regression auf X erklärt wird, ist also $\rho^2 = 0.43154$ oder etwa 43%. Die Regressionsparameter kann man ebenfalls auf einem Taschenrechner ablesen, wenn auch nicht in dieser genauen Form von Brüchen:

$$\begin{aligned} a &= \frac{13}{17}, \\ b &= \frac{1932}{85}, \text{ die Regressionsgerade lautet also} \\ Y &= \frac{13}{17}X + \frac{1932}{85}. \end{aligned}$$

Sie wurde in der Graphik oben bereits eingezeichnet. (Rechnung sowie Zeichnung anzufertigen ist äußerst langweilig und sollte man am besten einem geeigneten Computerprogramm überlassen, man muss dazu lediglich lernen, in welcher Form die Daten einzugeben sind, das ist bei verschiedenen Programmen immer wieder anders und natürlich erst recht äußerst langweilig.)

Bemerkung 1: Dieselben Werte erhält man als Schätzwerte für die richtigen Werte, wenn es sich nur um eine Stichprobe handelt, allerdings müssen die Stichproben recht groß sein, wenn sie gut, also mit einiger Sicherheit einigermaßen genau sein sollen, vgl. dazu den nächsten Abschnitt 2.1.2.

Bemerkung 2: Wir wollen hier einmal quantitativ nachschauen, wie genau die Voraussage der Y-Werte durch die X-Werte in unserem Beispiel ist. (Wir wissen natürlich bereits, dass die Regressionsgerade die beste lineare Vorhersage gibt, und wir sehen im Beispiel am Punkteschwarm keinen Anlass, es mit einer nichtlinearen Funktion zu versuchen.) Dazu betrachten wir den mittleren quadratischen Fehler der Vorhersage, das ist einfach

$$\frac{1}{10}\sum_{i=1}^{10}\left(y_i - \frac{13}{17}x_i - \frac{1932}{85}\right)^2 = \sigma^2(E) = \sigma^2(Y)\cdot(1-\rho^2).$$

Das brauchen wir also gar nicht konkret auszurechnen, vielmehr lesen wir $\sigma^2(Y)$ sofort aus dem Rechner ab, und ρ^2 haben wir bereits. Es ergibt sich in unserem Beispiel der mittlere quadratische Fehler der linearen Vorhersage

$$38.01\cdot(1-0.65692^2) = 21.607.$$

Mit leichter Ungenauigkeit können wir die Wurzel davon, das ist natürlich $\sigma(E)$, als mittleren absoluten Fehler dieser Vorhersage angeben, also etwa 4.65. Das bedeutet also, dass wir im Mittel die Punktezahl der zweiten Klausur mit nur etwa 4.65 Punkten Fehler, das ist etwa 10% von der Größenordnung der vorauszusagenden Werte, voraussagen können, und das ist recht gut, es macht keinen Notenunterschied. Im *Einzelfall* liegt die Voraussage natürlich auch einmal gründlich daneben! Der mittlere absolute Fehler ist natürlich *genau genommen* nicht dasselbe wie die Wurzel aus dem mittleren quadratischen Fehler, im Beispiel wollen einmal die Werte vergleichen:

Der mittlere absolute Voraussagefehler im Beispiel beträgt

$$\frac{1}{10}\sum_{i=1}^{10}\left|y_i - \frac{13}{17}x_i - \frac{1932}{85}\right| \approx 3.812.$$

Wir sehen also, dass der Unterschied zur Wurzel aus dem mittleren quadratischen Fehler nicht allzu groß ist. Dafür haben wir mit dem quadratischen Fehler (der Varianz) eine Größe, mit der sich systematische theoretische Rechnungen wesentlich leichter und übersichtlicher gestalten als solche mit Absolutwerten - alle allgemeinen Resultate insbesondere dieser Abschnitte wären damit nicht annähernd möglich!

2.1.2. *Schätzwerte für die Regressionsparameter und den Korrelationskoeffizienten anhand von Stichproben.* Alle interessierenden Zahlwerte berechnen sich durch Kovarianz und Varianzen, und so liegt es nahe, die übliche Varianzschätzung anhand von Stichproben einzusetzen, die man analog auch für Kovarianzen durchführt:

$$\begin{aligned} s^2(X) &= \widehat{\sigma^2(X)} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}, \\ \widehat{Cov(X,Y)} &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}. \end{aligned}$$

Damit erhalten wir (man beachte, dass sich die Nenner $n-1$ wegkürzen):

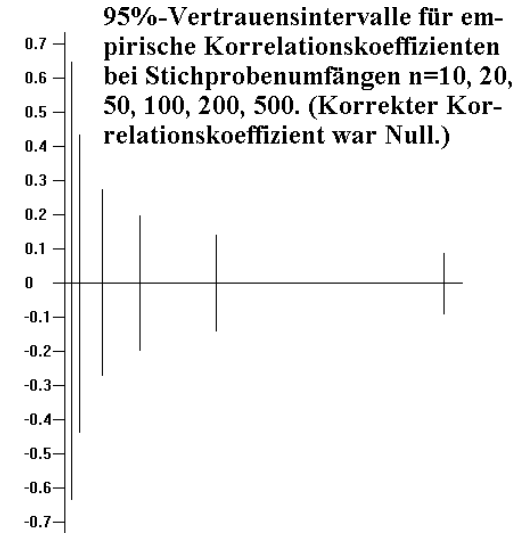
$$\begin{aligned} r(X,Y) &= \widehat{\rho(X,Y)} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}, \qquad (2.6) \\ \widehat{a_{Y\cdot X}} &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}. \end{aligned}$$

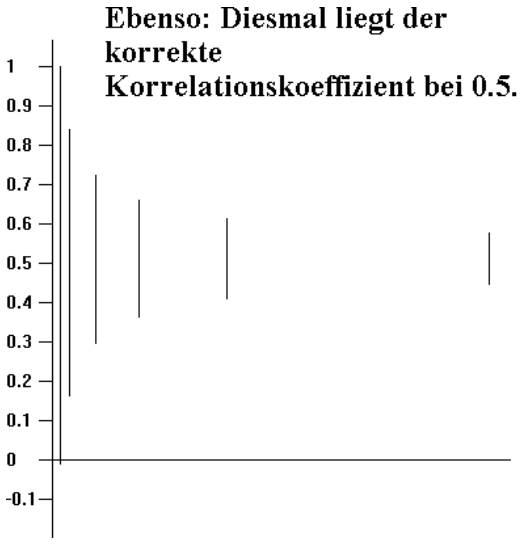
Man nennt diese auch den „empirischen" Korrelationskoeffizienten und die „empirische" Regressionssteigung. Bemerkung: Wenn die x_i, $1 \le i \le n$, alle Werte der Gesamtpopulation sind, zu den einzelnen Populationsmitgliedern, mit eventuellen Wiederholungen also, dann kommen bei diesen Formeln natürlich ρ und a heraus!

Wie genau ist die Schätzung eines Korrelationskoeffizienten (und entsprechend der Regressionssteigung) anhand einer Stichprobe?

Folgende Graphiken zeigen auf, dass man für vernünftige Qualität unangenehm hohe Stichprobenumfänge benötigt: Für verschiedene Stichprobenumfänge zeigen sie die Verteilung der empririschen Korrelationskoeffizienten und insbesondere die Breite eines 95%-Vertrauensintervalls für den Korrelationskoeffizienten. Dabei haben wir die idealtypische Situation zugrundegelegt, dass in Wirklichkeit gilt: $Y = X + E$, mit einer normalverteilten Variablen X und einer von X *statistisch* unabhängigen und ebenfalls normalverteilten Variablen E, deren Mittelwert Null ist. (Man vermute den Grund für die halbwegs deprimierenden Ergebnisse also nicht etwa darin, dass die betrachteten Variablen überhaupt das Korrelationskonzept nicht korrekt anwenden lassen, im Gegenteil handelt es sich um den mathematisch idealen Fall der Anwendbarkeit, der so genau für empirische Variablen

niemals vorliegt.) Den richtigen Korrelationskoeffizienten ρ kennen wir natürlich, wir können ihn über $\sigma^2(E)/\sigma^2(Y) = 1 - \rho^2$ beliebig einstellen. Wir wählen im Beispiel den Wert $\rho = 0$ und im zweiten den Wert $\rho = 0.5$ (das ist in *manchen* Zusammenhängen (Eignungstests etc.) schon hoch).





2.1.3. *Die Regressionsgerade als Lösung eines Extremwertproblems.* Hier leiten wir die Regressionsgerade noch einmal her als „optimale" Gerade zu einem Punkteschwarm (x_i, y_i), $1 \leq i \leq n$, dabei lernt man die wichtige und verbreitete „Kleinste-Quadrate-Methode" (engl. 'least squares') kennen. Für eine Gerade mit (beliebigen) Parametern a, b hat man den Voraussagefehler $y_i - ax_i - b$ für das i-te

Beobachtungspaar. Damit ist

$$f(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2 = \text{ quadratischer Gesamtfehler bei Vorhersage von } Y \text{ durch die Variable } aX+b$$

ein vernünftiges Maß für die Qualität der Vorhersage. Die beste Gerade wäre demnach gerade die, für welche $f(a,b)$ minimal wird. (Wir werden gleich sehen, dass es eine einzige Gerade gibt mit minimalem quadratischem Gesamtfehler.) Dies ist eine sehr einfache Extremwertaufgabe, ein wenig erschwerdend ist es nur, dass die Funktion f, deren Wert minimalisiert werden soll, eine Funktion von zwei unabhängigen Variablen ist. Für den vorliegenden konkreten Fall können wir dies Problem umgehen, stellen seine Lösung jedoch im nächsten Abschnitt 2.2 vor, der sich dem verallgemeinerten Regressionsproblem widmet.

Lösung des Extremwertproblems:

Zunächst setzen wir mit derselben Begründung wie oben wieder $b = \mu(Y) - a\mu(X)$, was wir hier konkreter $\overline{y} - a\overline{x}$ schreiben können, da wir ausschließlich die vorgegebenen Wertepaare betrachten wollen. Dann reduziert sich unser Problem auf die Aufgabe,

$$g(a) = \sum_{i=1}^{n}(y_i - ax_i - \overline{y} + a\overline{x})^2$$

zu minimalisieren. Wir bilden die erste Ableitung und setzen sie gleich Null:

$$g'(a) = 2\sum_{i=1}^{n}(\overline{x} - x_i)(y_i - ax_i - \overline{y} + a\overline{x}) = 0.$$

Das ist eine simple lineare Gleichung für a, man hat lediglich die Glieder mit dem Faktor a zu isolieren, auszuklammern und kann auflösen (fleißig $\sum x_i = n\overline{x}$ benutzen!) zu:

$$a = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x}^2} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}.$$

Die letzte Gleichung ergibt sich daraus, dass man wieder $\sigma^2(X) = \mu(X^2) - \mu^2(X)$ und das Analogon für die Kovarianz benutzt, was man leicht konkret ausrechnet. Das Resultat ist dasselbe wie oben. Man sieht leicht, dass es sich wirklich um ein absolutes Minimum handelt, da g eine quadratische Funktion zu einer nach oben geöffneten Parabel ist. Wir haben also zu einem endlichen Punkteschwarm die im Sinne des Kleinsten-Quadrate-Kriteriums optimale Gerade ausgerechnet.

Man kann das Problem auch abstrakter fassen und eine Variable $aX+b$ derart suchen, dass die Varianz der Fehlervariable E - das ist eine Funktion von a und b - minimalisiert wird, also wieder ausgehen von

$$\begin{aligned} Y &= aX + b + E, \text{ also} \\ E &= Y - aX - b, \text{ zu minimalisieren ist daher:} \\ g(a) &= \sigma^2(E) = \sigma^2(Y - aX) = \mu((Y-aX)^2) - \mu^2(Y - aX) \\ &= a^2\sigma^2(X) - 2aCov(X,Y). \end{aligned}$$

Man hat sofort

$$g'(a) = -2Cov(X,Y) + 2a\sigma^2(X).$$

Null wird das genau für

$$a = \frac{Cov(X,Y)}{\sigma^2(X)},$$

und auch das ergibt sofort - und am elegantesten - die alte Formel. Interessant ist es dabei, dass nichts über die auch nur lineare Unabhängigkeit zwischen X und E vorausgesetzt wurde, dass man diese vielmehr als Konsequenz der Minimalität von $\sigma^2(E)$ mit dem ausgerechneten Parameter a erhält.

2.2. Verallgemeinerung des Problems auf mehrere unabhängige Variablen und auf nichtlineare Regression. Wir behandeln zunächst den *linearen* Fall mit mehreren unabhängigen Veränderlichen, also die Modellgleichung:

$$Y = a_1X_1 + a_2X_2 + ... + a_nX_n + E. \tag{2.7}$$

Man beachte, dass eine additive Konstante hier überflüssig ist, weil man dafür etwa eine Variable (Erinnerung: Das klingt komisch, aber wir waren immer in der Lage, Variablen mit nur einem Wert - also Streuung Null - ebenso zu behandeln wie alle anderen!) X_n = Konstante mit Wert 1 zusätzlich zu den anvisierten Variablen hinzufügen kann. Dann ist a_n die additive Konstante. Nunmehr ist die zu minimalisierende Varianz von E eine Funktion von den n Parametern $a_1, ..., a_n$. Diese sind also nunmehr unsere unabhängigen Variablen der Funktion

$$g(a_1, ..., a_n) = \sigma^2(Y - \sum_{i=1}^{n} a_iX_i).$$

Gesucht ist die (wieder stellt sich heraus, dass es eine einzige gibt) Parameterfolge, für die das minimal wird. Es liegt nahe, dass man dazu auch die Ableitung benutzt, die wir aber für Funktionen mehrer Veränderlicher noch gar nicht kennen. Die Idee dazu können wir jedoch anhand unseres Extremwertproblems leicht einführen:

Stellen wir uns vor, für die konkrete Folge $(b_1, ..., b_n)$ hätten wir einen *lokalen* Extremwert von g, d.h. für alle $(a_1, ..., a_n)$ in einer kleinen Umgebung von $(b_1, ..., b_n)$ gälte $g(a_1, ..., a_n) \geq g(b_1, ..., b_n)$. Dann gilt *insbesondere* für a_1 in einer kleinen Umgebung von b_1, dass $g(a_1, b_2, ..., b_n) \geq g(b_1, ..., b_n)$. Das heißt aber, dass an der Stelle b_1 ein lokaler Extremwert der Funktion $h(a) = g(a, b_2, ..., b_n)$ liegt. Also können wir folgern, dass $h'(b_1) = 0$ gilt. Das ist eine Gleichung mit den Unbekannten $b_1, ..., b_n$. Dasselbe können wir nun für jede Variable von g durchführen, und so erhalten wir n Gleichungen für unsere n Unbekannten. Die Idee war also, aus einer Funktion mit n unabhängigen Veränderlichen n Funktionen mit nur einer zu machen und deren Ableitungen zu betrachten. Man variiert also immer nur eine Variable und betrachtet die anderen als konstant. Dafür hat man nun eine bequeme Notation und verbale Bezeichnung:

Partielle Ableitungen von einer Funktion $g{:}(a_1, ..., a_n) \mapsto g(a_1, ..., a_n)$:

$$\begin{aligned} \frac{\partial}{\partial a_i} g(a_1, ..., a_n) &= h_i'(a_i), \text{ wobei} \\ h_i(a_i) &= g(a_1, ..., a_n), \text{ worin nunmehr alle } a_j,\ j \neq i, \\ &\quad \text{als äußere Parameter betrachtet werden.} \end{aligned} \tag{2.8}$$

Notwendige Bedingung für das Auftreten von lokalen Extremwerten bei einer Funktion mehrerer Veränderlicher ist demnach das Verschwinden aller partiellen Ableitungen (sofern diese existieren). Übrigens errechnet man aus den partiellen Ableitungen auch die Gesamtableitung zur linearen Approximation, Existenz der letzteren ist aber noch etwas mehr als bloße Existenz der partiellen Ableitungen.

Wie sieht nun das Gleichungssystem in unserem Beispielproblem aus, das durch diesen Prozess entsteht? Glücklicherweise ist es viel einfacher als man allgemein fände, nämlich ein *lineares (*quadratisches*)* Gleichungssystem, das man stets eindeutig lösen kann, wenn nur die Gleichungen unabhängig (wieder ein anderer Begriff von Unabhängigkeit - keine Gleichung soll aus den übrigen folgen, und die Gleichungen sollen miteinander verträglich sein) sind. (Bei Daten zu Prädiktorvariablen läuft das in unserem Beispiel auf eine gewisse (das ist übrigens *wieder* eine andere Unabhängigkeit, nämlich die lineare im Sinne der Vektorraumtheorie!) Unabhängigkeit der einzelnen Variablen hinaus. Klar ist z.B., dass man nicht erwarten kann, Y eindeutig als $a_2X + a_2X$ darzustellen (hier wäre $X_1 = X_2$). Mehr und besser noch: In unserem linearen Spezialfall gelangt man mit der eindeutigen Lösung auch zu einem absoluten Minimum, also Optimum.

Das lineare Gleichungssystem für die Modellgleichung
$$\mathbf{Y = a_1X_1 + ... + a_nX_n + E}$$

Tatsächlich lässt sich diese sehr leicht ausrechnen. Man hat mit der Forderung $\mu(E) = 0$ (unter den Prädiktorvariablen sei eine Konstante untergebracht, dann ist das unproblematisch):

$$\begin{aligned} g(a_1, ..., a_n) &= \sigma^2\left(Y - \sum_{i=1}^{n} a_iX_i\right) = \mu\left(\left(Y - \sum_{i=1}^{n} a_iX_i\right)^2\right) \text{ (mit } \mu(E) = 0) \\ &= \mu\left(Y^2 - 2\sum a_iX_iY + \left(\sum a_iX_i\right)^2\right) \\ &= \mu(Y^2) - 2\sum a_i\mu(X_iY) + \sum_{i\neq j} a_ia_j\mu(X_iX_j) + \sum_{i=1}^{n} a_i^2\mu\left(X_i^2\right) \end{aligned}$$

Dies ist zu minimalisieren. Dies ergibt das System der folgenden n linearen Gleichungen, von denen die i-te lautet:

$$\frac{\partial}{\partial a_i}g(a_1, ..., a_n) = -2\mu(X_iY) + 2\sum_{i\neq j} a_j\mu(X_iX_j) + 2a_i\mu\left(X_i^2\right) = 0.$$

Ordentlich als lineares Gleichungssystem hingeschrieben, nach Division durch 2:

$$\sum_{i\neq j} a_j\mu(X_iX_j) + a_i\mu\left(X_i^2\right) = \mu(X_iY),\ 1 \leq i \leq n.$$

In der konkreteren vertrauteren Form als lineares Gleichungssystem mit Pünktchen geschrieben:

$$\begin{aligned} a_1\mu(X_1^2) + a_2\mu(X_1X_2) + a_3\mu(X_1X_3) + ... + a_n\mu(X_1X_n) &= \mu(X_1Y) \\ a_1\mu(X_2X_1) + a_2\mu(X_2^2) + a_3\mu(X_2X_3) + ... + a_n\mu(X_2X_n) &= \mu(X_1Y) \\ \vdots &= \vdots \text{ (weitere Zeilen)} \\ a_1\mu(X_nX_1) + a_2\mu(X_nX_2) + a_3\mu(X_nX_3) + ... + a_n\mu(X_n^2) &= \mu(X_nY). \end{aligned}$$

Wenn die Variablen nur endliche Population haben oder aber mit Stichproben gearbeitet wird, dann stimmt die Lösung dieses Gleichungssystems wieder mit der Kleinsten-Quadrate-Lösung überein, und das Gleichungssystem erhält die konkretere Gestalt, nach Weglassen der Faktoren $\frac{1}{n}$: Wir bezeichnen mit $x_{i,k}$ (man könnte das auch ohne Komma schreiben, das Komma dient nur zur Verdeutlichung dessen, dass es sich um einen Doppelindex handelt, nicht etwa um ein Produkt) den Wert der Variablen X_i beim k-ten Populations- bzw. Stichprobenmitglied, mit y_k den Wert der Variablen Y bei eben diesem Mitglied. Dabei laufe k von 1 bis $N =$ Populations- bzw. Stichprobenumfang. (Diese Zahl N müssen wir sorgsam unterscheiden von n, der Anzahl der benutzten unabhängigen Variablen. Normalerweise wird n sehr klein gegen N sein.)

$$\begin{aligned}
a_1\sum_{i=1}^{N}x_{1,i}^2 + a_2\sum_{i=1}^{N}x_{1,i}x_{2,i} + a_3\sum_{i=1}^{N}x_{1,i}x_{3,i} + ... + a_n\sum_{i=1}^{N}x_{1,i}x_{n,i} &= \sum_{i=1}^{N}x_{1,i}y_i\\
a_1\sum_{i=1}^{N}x_{2,i}x_{1,i} + a_2\sum_{i=1}^{N}x_{2,i}^2 + a_3\sum_{i=1}^{N}x_{2,i}x_{3,i} + ... + a_n\sum_{i=1}^{N}x_{2,i}x_{n,i} &= \sum_{i=1}^{N}x_{2,i}y_i\\
&\vdots = \vdots\\
a_1\sum_{i=1}^{N}x_{n,i}x_{1,i} + a_2\sum_{i=1}^{N}x_{n,i}x_{2,i} + a_3\sum_{i=1}^{N}x_{n,i}x_{3,i} + ... + a_n\sum_{i=1}^{N}x_{n,i}^2 &= \sum_{i=1}^{N}x_{n,i}y_i.
\end{aligned}$$

Manchmal wird hier zwecks größerer Übersichtlichkeit die Notation $[x_ix_j]$ für $\sum_{k=1}^{N}x_{i,k}x_{j,k}$ bzw. $[x_iy]$ für $\sum_{k=1}^{N}x_{i,k}y_k$ verwandt. In der i-ten Zeile steht also auf der rechten Seite $[x_iy]$, auf der linken bei der Unbekannten a_j der Faktor (Koeffizient) $[x_ix_j]$. Übrigens lösen Computerprogramme diese Gleichungssysteme bequem auf, nachdem man in geeigneter Weise die Vektoren $(x_{1,k}, x_{2,k}, x_{3,k}..., x_{n,k}, y_k)$, $1 \le k \le N$, eingegeben hat.

Zur *Konkretisierung* schauen wir einmal nach, was wir mit $n = 2$, $X_1 = X$, $X_2 = 1$ bekommen. Wir schreiben auch a für a_1 und b für a_2. Natürlich ist das genau das einfache Modell der linearen Regression mit einer unabhängigen Variablen, das wir oben besprachen, und selbstverständlich sollte das alte Ergebnis herauskommen. Aber an diesem kleinen Beispiel kann man schon das Funktionieren des verallgemeinerten Ansatzes beobachten. Das Gleichungssystem lautet konkret (mit den Summen, nicht den Erwartungswerten) - man beachte, dass stets $x_{2,k} = 1$, da X_2 konstante Größe mit Wert 1 ist, außerdem $x_{1,k}$ nunmehr vereinfacht x_k heißt:

$$\begin{aligned}
a\sum_{k=1}^{N}x_k^2 + b\sum_{k=1}^{N}x_k &= \sum_{k=1}^{N}x_ky_k\\
a\sum_{k=1}^{N}x_k + b\sum_{k=1}^{N}1 &= \sum_{k=1}^{N}y_k \text{ (beachte: } \sum_{k=1}^{N}1 = N).
\end{aligned}$$

Dies ist ein übersichtliches (lineares ohnehin) (2×2)-Gleichungssystem, und wir lösen es auf: Die zweite Zeile ergibt sofort

$$b = \frac{1}{N}\sum_{k=1}^{N}y_k - a\cdot\frac{1}{N}\sum_{k=1}^{N}x_k = \overline{y} - a\overline{x}.$$

Eingesetzt in die erste Gleichung:

$$a\left(\left(\sum_{k=1}^{N} x_k^2\right) - \overline{x}\sum_{k=1}^{N} x_k\right) + \overline{y}\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} x_k y_k$$

ergibt das

$$a = \frac{\sum x_k y_k - N\overline{xy}}{\sum x_k^2 - N\overline{x}^2} = \frac{\sum (x_k - \overline{x})(y_k - \overline{y})}{\sum (x_k - \overline{x})^2}.$$

Tatsächlich das alte Resultat. (Darum ging es natürlich nicht, das wurde oben eleganter hergeleitet, sondern es war nur um ein konkretes lineares Gleichungssystem zu tun, wie es sich bei mehreren unabhängigen Variablen ergibt.)

Zuvor fragten wir im Falle von nur einer unabhängigen Variablen X danach, welcher Anteil der Varianz von einer Zielvariablen Y durch lineare Regression auf X erklärt ist, und erhielten als Antwort: $\rho^2(X, Y)$ beschreibt genau diesen Anteil. Analoges ist auch bei mehreren unabhängigen Variablen vorzunehmen:

DEFINITION 24 (multipler Korrelationskoeffizient). *Der multiple Korrelationskoeffizient für die Regression von Y auf $X_1, ..., X_n$ wird mit $R(Y|X_1, ..., X_n)$ bezeichnet und ist folgendermaßen definiert: Sei*

$$Y = \sum_{i=1}^{n} a_i X_i + c + E$$

die Regressionsgleichung mit den zuvor berechneten Regressionskoeffizienten, dann ist

$$\widehat{Y} := \sum_{i=1}^{n} a_i X_i + c, \text{ manchmal ausführlicher } \widehat{Y}(X_1, ..., X_n) \text{ geschrieben,}$$

die lineare Schätzgröße für Y aus $X_1, ..., X_n$, und man definiert

$$R(Y|X_1, ..., X_n) := \rho(Y, \widehat{Y}).$$

(Das ist der gewöhnliche einfache Korrelationskoeffizient zwischen diesen beiden Variablen.)

Bemerkung zur Bezeichnung: Eigentlich sollte man großes griechisches „R" verwenden, das sieht aber so aus: P, was für Wahrscheinlichkeitsfunktion vergeben ist. Für die empirische Schätzung sollte man daher konsequent „$\widehat{R}$" verwenden, was sich wie oben unter Einsetzen der empirischen Regressionskoeffizienten $\widehat{a_i}$ mittels $r(Y, \widehat{Y}_{empir})$ (empirischer Korrelationskoeffizient, s.o. 2.1.2) ergibt, mit $\widehat{Y}_{empir} = \sum_{i=1}^{n} \widehat{a_i} X_i + c$.

SATZ 17. *Man hat mit den Bezeichnungen der Definition:*

$$R^2(Y|X_1, ..., X_n) = \frac{\sigma^2\left(\widehat{Y}\right)}{\sigma^2(Y)}.$$

Das ist der Anteil der Varianz von Y, der durch lineare Regression auf $X_1, ..., X_n$ erklärt wird.

Der Grund ist einfach der, dass die Fehlervariable E wieder linear unabhängig ist von $\widehat{Y}$ und sich somit wie im einfachen Fall die Varianz von Y zerlegt als $\sigma^2(Y) = \sigma^2(\widehat{Y}) + \sigma^2(E)$.

Verallgemeinerung auf nichtlineare Regression

Tatsächlich kann man mit dem gelösten Problem für den linearen Spezialfall auch nichtlineare Abhängigkeiten funktional darstellen. Dazu bedient man sich des einfachen Kunstgriffs, nichtlineare Funktionen $Z_i = f_i(X_i)$ der Prädiktorvariablen als neue Prädiktorvariablen Z_i zu nehmen und dann *für diese* den *linearen Ansatz*

$$Y = a_1 Z_1 + ... + a_n Z_n + E$$

zu machen. Man rechnet wie oben für den verallgemeinerten linearen Ansatz gezeigt und hat damit Y mit dem Fehler E als nichtlineare Funktion der X_i dargestellt. Selbstverständlich kann man auch allgemeiner Prädiktoren der Form $f_i(X_1, ..., X_n)$, mit einer völlig freien Anzahl dieser Funktionen f_i (also nicht notwendig n) versuchen. Natürlich besteht das Hauptproblem darin, geeignete nichtlineare Funktionen zu finden, was man keineswegs schematisieren kann. Manchmal hat man Erfolg damit, ein Polynom gewissen Grades in den $X_1, ..., X_n$ zu verwenden. Der Ansatz sieht dann z.B. bei zwei Variablen für ein Polynom zweiten Grades so aus:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_1^2 + a_4 X_2^2 + a_5 X_1 X_2 + E.$$

(Also Summe der Exponenten zu den unabhängigen Variablen ≤ 2.) In der gewöhnlichen Varianzanalyse handelt es sich übrigens genau um diesen Fall, noch abgespeckt um die Terme mit den quadrierten Variablen. Außer dem linearen Anteil steht dann lediglich noch das Produkt $X_1 X_2$ (mit einem Faktor) als „Interaktionsterm".

Bemerkungen zu Wahl und eventuell Revision des Ansatzes

Der Möglichkeiten sind viele, wörtlich unendlich viele. Aber ein generelles Prinzip ist wichtig zu beachten: Man hat Daten (Vektoren der Werte der Prädiktoren und der Zielvariablen), dazu weitere mögliche Daten aus der Population. Idee des Ganzen muss es sein, diese Datenmenge übersichtlich, möglichst einfach darzustellen. Dabei interessieren genau genommen nur die Werte der Zielgröße - die Prädiktoren und zugehörige Parameter sollen ja gerade diese Werte rekonstruieren. Das heißt aber, man sollte möglichst wenige Prädiktoren, damit möglichst wenige zu bestimmende Parameter benutzen. Dazu eine wichtige generelle Einsicht: Wenn ein einfaches Modell (also mit wenigen Parametern) recht gut auf eine Stichprobe passt, so hat man gute Aussicht, dass es auch einigermaßen zu weiteren Daten passt, die noch nicht in der Stichprobe enthalten waren. Man kann also eine recht gute Generalisierungsfähigkeit (von den Beispielen der Stichprobe auf die Gesamtpopulation) erwarten. Dagegen ist es eine Binsenweisheit, dass man mit allzu vielen Parametern eine Fülle von ganz verschiedenen Möglichkeiten hat, die Daten einer Stichprobe recht genau, ja im Prinzip beliebig genau zu reproduzieren. Aber einmal hat man damit keine gute Reduktion - das Modell selbst enthält mit den vielen Parametern viele Daten, außerdem kann man sicher sein, dass Verallgemeinerung auf weitere Daten nicht gelingt - bei denen gehen die Voraussageresultate der verschiedenen zur Stichprobe passenden Modelle dann weit auseinander. An solchen Modellen besteht demnach aus doppeltem Grund keinerlei vernünftiges Interesse. Noch eine Bemerkung: Hat man ein Modell gerechnet, so schaue man nach, welche der Koeffizienten sehr nahe bei Null liegen und damit zur Voraussage der Zielgröße kaum etwas beitragen - die kann man dann zwecks weiterer Reduktion gut weglassen. Schließlich muss noch bemerkt werden, dass die Sicherheit der Parameterbestimmung durch Stichproben (Parameterschätzung) mit demselben Problem

behaftet ist wie das Schätzen von Korrelationskoeffizienten, das wir oben illustrierten. Bei relativ kleinen Stichproben werden die Vertrauensintervalle unbrauchbar groß. Noch ein abschließender Rat: Wer zu diesem und ähnlichen Themen ernsthaft mehr als hier dargestellt lernen möchte, wende sich an amerikanische Literatur, und zwar den Teil, der selbstverständlich mit Vektoren und Matrizen arbeitet, und lerne die mathematischen Prärequisiten!

3. Korrelation/Regression und „Klassische Testtheorie"

Wir reden von Tests, die irgendeine Eignung, Fähigkeit usw. „messen", wir denken auch an IQ-Tests. „Klassische Testtheorie" hört sich formidabel an, dahinter steckt jedoch nur ein einfaches Modell solcher Tests, dessen Anwendung auf eine simple Übungsaufgabe zum Thema Regression und Korrelation hinausläuft.

Das Modell: Wenn man an einer Versuchsperson einen Test durchführt, so ermittelt man eine Testpunktzahl (die könnte auch ein beliebiger Dezimalbruch sein). Dieser Zahlenwert gibt jedoch nicht den „wahren Wert", sondern den wahren Wert plus einen Zufallsfehlerwert. In diesem Zusammenhang ist nun keineswegs zu mystifizieren, was der „wahre" Wert (für die besagte Versuchsperson) sei: Es ist einfach der Erwartungswert (Mittelwert) der Testpunktzahl bei einer beliebigen Durchführung des Tests bei dieser Person. Diese Einfachheit der begrifflichen Beschreibung bedeutet allerdings nicht, dass es leicht wäre, an diesen Erwartungswert heranzukommen - man müsste sehr oft den Test *(statistisch!) unabhängig* mit dieser Person durchführen und das arithmetische Mittel nehmen - und hätte immer noch nur einen Näherungswert. Gerade dieser Punkt wird später eine nützliche kleine Übungsaufgabe stellen.

Es ist nun zum Verständnis äußerst wichtig, nicht von einzelnen Zahlen zu reden, sondern von *Zufallsvariablen*. Wir haben gelernt: Eine Zufallsvariable ordnet jedem Element einer fixierten Population eindeutig eine Zahl zu. In unserer Situation: Wir haben die *Testvariable* T, die jeder Person ihre Testpunktzahl zuordnet, welche bei einer beliebigen Durchführung des Tests mit dieser Person beobachtet werden kann. Weiter haben wir die Zufallsvariable W, welche jeder Person ihren Erwartungswert für T bei allen Durchführungen an dieser person zuordnet. Die Variable W ordnet also jeder Person eindeutig einen (wenn auch schwer zugänglichen) Zahlenwert zu, während die Variable T jedem Paar ($Person$, $Versuch$ Nr....) eindeutig einen Zahlenwert zuordnet, aber auch bei Einschränkung auf eine einzige Person noch eine Zufallvariable bleibt, die viele Werte annehmen kann. Für die folgenden Betrachtungen sollten T und W dieselbe Population haben. Das lässt sich leicht dadurch bewerkstelligen, dass man W ebenfalls jedem Paar ($Person$, $Versuch$ Nr. ...) eine Zahl zuordnen lässt, nämlich einfach den oben beschriebenen „wahren Wert" zur Person (eben denselben für jede Versuchs-Nummer). Nun können wir das Modell der Klassischen Testtheorie beschreiben, das der Modellvorstellung der mit einem Zufallsfehler behafteten physikalischen Messung entlehnt ist: T ist eine Messung, die W mit dem Zufallsfehler E misst. Also in jedem Einzelfall $(Person\ \omega, Testdurchführung\ Nr.\ n) : T(\omega, n) = W(\omega, n) + E(\omega, n)$. Der Wert $E(\omega, n) = T(\omega, n) - W(\omega, n)$ ist der beobachtete Fehler des Messwertes $T(\omega, n)$ gegenüber dem zu messenden „wahren Wert" $W(\omega, n)$ - letzterer hängt natürlich nach unserer Konstruktion nur von ω wirklich ab. Wichtig ist es, diese Gleichung auf dem Niveau der Zufallsvariablen zu nehmen, also Abbildungen:

$$T = W + E.$$

Gerade dies Regressionsmodell (E linear unabhängig von W, $\mu(E) = 0$, $\sigma^2(E) = \sigma^2(E|W = w)$ für alle Werte w von W) ist das Grundmodell der Klassischen Testtheorie, wobei das Idealbild des allgemeinen Regressionsmodells vorausgesetzt wird: T, W, E normalverteilt. Ferner wird noch vorausgesetzt, dass bei (statistisch) unabhängiger Testwiederholung die Fehlervariablen statistisch unabhängig voneinander seien. Das hat zunächst die Konsequenz, dass man mit oftmaliger Wiederholung des Tests beliebig nahe an den jeweiligen Wert von W herankommt (wie auch sonst bei Mittelung statistisch unabhängiger Messwerte). Nun kommen wir zum ersten Standardproblem: Wenn man nichts tun kann als Testwerte nehmen, wie kann man dann zur Genauigkeit eines solchen Tests - d.h. natürlich: zur Fehlervarianz $\sigma^2(E)$ - eine Aussage machen? (Man beachte: Es ist niemand da, der in jedem Einzelfall einer Testdurchführung den zugehörigen „wahren Wert" bereitstellt! Außerdem sind auch viele statistisch unabhängige Testwiederholungen bei einer einzigen Person recht unrealistisch. Es wird sich jedoch zeigen, dass man sich tatsächlich am eigenen Schopf aus dem Sumpfe ziehen kann:

3.1. Das Problem, die Genauigkeit eines Tests empirisch zu bestimmen. Wir stellen uns vor, dass wir mit jedem Probanden nur eine einzige statistisch unabhängige Testwiederholung durchführen und nennen die Testpunktzahlvariable für die erste Durchführung T_1, für die zweite T_2. Empirisch können wir nunmehr aufgrund der beobachteten Wertepaare (Wert von T_1, Wert von T_2) $r_{T_1T_2}$ als Näherungswert für $\rho(T_1, T_2)$ bestimmen. Nun stellen wir folgende kleine rein mathematische Überlegung an:

$$\begin{aligned}\rho(T_1,T_2) &= \frac{Cov(T_1,T_2)}{\sigma(T_1)\sigma(T_2)} = \frac{Cov(W+E_1,W+E_2)}{\sigma^2(T)} = \frac{Cov(W,W)}{\sigma^2(T)} = \frac{\sigma^2(W)}{\sigma^2(T)} \\ &= \rho^2(W,T). \\ \sigma^2(E) &= \sigma^2(T)(1-\rho^2(W,T)) = \sigma^2(T)(1-\rho(T_1,T_2)).\end{aligned}$$

Die zweite Zeile ist einfach die Formel für die Varianzzerlegung für jede beliebige Regressionsgleichung, nur ist aufgrund der ersten Zeile $\rho^2(W, T)$ durch $\rho(T_1, T_2)$ ersetzt worden. Nun die Interpretation: $\rho(T_1, T_2)$ ist die Korrelation zwischen Test und Retest (Testwiederholung), und diese heißt **Reliabilität** (Zuverlässigkeit) des Tests. Das ist intuitiv verständlich: Ist diese Korrelation hoch (d. h. nahe bei 1 - positiv ist sie jedenfalls nach Konstruktion), so misst man mit dem Test bei Wiederholung mit hoher Wahrscheinlichkeit keine stark voneinander abweichenden Werte. Das mathematische Resultat bedeutet nun aber, dass man über diese Reliabilität und die Varianz von T - diese sind beide durch einfache Testwiederholung an einer Stichprobe empirisch zu ermitteln! - auch an die Varianz von E und damit an die mittlere quadratische Abweichung zwischen T und W herankommt: $\sigma^2(E) = \sigma^2(T)(1 - \rho(T_1, T_2))$. Die unmittelbare praktische Konsequenz: Kennen wir gute empirische Näherungswerte für $\sigma^2(T)$ und $\rho(T_1, T_2)$, so wissen wir nach einer Messung eines Testwertes t bei einem Probanden, in welchem Bereich dessen „wahrer Wert" liegt. Genauer: Wir können ein Vertrauensintervall für den „wahren Wert" geben, wie wir das bei jeder Normalverteilung tun. Dafür benötigen wir die Streuung, in diesem Falle $\sigma(E)$ - für die haben wir aber nun eine Formel.

Beispiel: Ein IQ-Test habe eine Reliabilität von 0.9, und die Testwerte mögen mit 15 Punkten streuen, also $\sigma(T) = 15$. Wir haben bei einem Probanden mit einmaliger Durchführung des IQ-Tests den Wert 115 ermittelt. In welchem Bereich liegt der „wahre IQ-Wert" dieser Person mit 95% Sicherheit? (Noch einmal: Der

„wahre IQ-Wert" dieser Person ist definiert als Erwartungswert aller Testwiederholungen bei dieser Person - nennen wir sie I, etwas mathematisch ausgedrückt: $\mu(T|\omega = I)$. Wir fragen also nicht, ob der Test „wirkliche Intelligenz" oder Intelligenz in angemessener Weise messe, sondern lassen den Begriff der Intelligenz hier einmal operationalisiert sein als das, was dieser Test misst (im Sinne des Mittelwertes für jede Person). Wir fragen nur nach der Abweichung der einzelnen Testergebnisse (bei dieser Person) von jenem Mittelwert (für diese Person).

Antwort: Wir haben in diesem Beispiel $\sigma(E) = 15(1 - 0.9) = 1.5$. Der wahre Wert der Person liegt also mit 95% Sicherheit im Bereich $115 \pm 1.96 \cdot 1.5$, also zwischen 112 und 118. Wir können also darauf vertrauen, dass der gemessene Wert höchstens um drei Punkte von dem „wahren Wert" der Person abweicht.

3.2. Ein weiterer Gesichtspunkt zum praktischen Umgang mit der Reliabilität eines Tests: Testverlängerung bzw. -verkürzung. Wenn wir einen sehr zuverlässigen Test haben, so können wir überlegen, ob er bei einer praktisch wünschbaren Verkürzung immer noch ausreichende Reliabilität besitzt. Umgekehrt kann bei einer unbefriedigenden Zuverlässigkeit über eine Verlängerung des Tests nachgedacht werden, um die Zuverlässigkeit in befriedigender Weise zu steigern. Dazu denken wir uns den Tests aus „Items" aufgebaut, von denen wir wegnehmen oder hinzufügen können. Das Problem läuft klar auf folgende Übungsaufgabe hinaus: Wie ändert sich bei Vervielfachung mit Wert k ($k < 1$ bedeutet Verkürzung, $k > 1$ Verlängerung des Tests) die Reliabilität? Wir nennen nunmehr die Testvariable T wie bisher und bezeichnen die Testvariable zum mit Faktor k multiplizierten Test mit $T^{(k)}$. Wir wollen die Reliabilität von $T^{(k)}$ durch die von T ausdrücken - wie bisher bezeichnen wir die Nummer der Testdurchführung mit unterem Index (1 oder 2). Tatsächlich macht die Sache auch für gebrochene Werte von k Sinn, aber der Einfachheit halber begnügen wir uns mit ganzzahligem $k > 1$ (für entsprechende Verkürzung mit Faktor $1/k$ können wir Entsprechendes dann folgern): Wir können die Testvariable $T^{(k)}$ dann darstellen als $T_1 + ... + T_k$, sie benimmt sich wie einer Summe von k Wiederholungen des Ausgangstests mit Testvariable T. Nun soll aber $T^{(k)}$ zwei mal durchgeführt werden (Test und Retest), damit wir von der Reliabilität von $T^{(k)}$ reden können. Dazu bezeichnen wir die Durchführungen mit $T_1^{(k)}$, $T_2^{(k)}$ und stellen sie strukturiert so dar:

$$\begin{aligned} T_1^{(k)} &= T_{1,1} + ... + T_{k,1}, \\ T_2^{(k)} &= T_{1,2} + ... + T_{k,2}. \end{aligned}$$

Wir berechnen die interessierende Reliabilität des mit Faktor k verlängerten Tests:

$$\begin{aligned} \rho(T_1^{(k)}, T_2^{(k)}) &= \frac{Cov(T_{1,1} + ... + T_{k,1}, T_{1,2} + ... + T_{k,2})}{\sigma^2(T_{1,1} + ... + T_{k,1})} \\ &= \frac{k^2 Cov(T_1, T_2)}{k\sigma^2(T) + k(k-1)Cov(T_1, T_2)} \\ &= \frac{k\rho(T_1, T_2)}{1 + (k-1)\rho(T_1, T_2)}. \end{aligned}$$

Man kann also die Reliabilität von $T^{(k)}$ durch die von T und den Faktor k ausdrücken. Die Formel zeigt, dass man natürlich durch Steigerung von k nur dem Wert 1 sich nähern kann, ohne ihn zu erreichen, falls $\rho(T_1, , T_2) < 1$. Für den Fall der Verkürzung drehen wir den Spiess einfach um und nutzen die eben hergeleitete

Formel, worin wir einfach $\rho(T_1^{(k)}, T_2^{(k)})$ als bekannte Reliabilität des "Langtests" nehmen und $\rho(T_1, T_2)$ als Unbekannte auffassen. Wir lösen nach $\rho(T_1, T_2)$ auf und erhalten tatsächlich:

$$\rho(T_1, T_2) = \frac{\frac{1}{k}\rho(T_1^{(k)}, T_2^{(k)})}{1 + \left(\frac{1}{k} - 1\right)\rho(T_1^{(k)}, T_2^{(k)})}.$$

3.3. Bemerkungen zum Begriff der Validität eines Tests. Intuitiv bedeutet „Validität" so etwas wie Gültigkeit, man meint, dass ein Test wirklich das misst, was er messen soll - und nichts sonst. Klassisches Beispiel: Man möchte das Bewegungstalent eines Menschen messen und misst stattdessen, wie weit er bestimmte Übungen eintrainiert hat. Ein Test könnte zuverlässig messen (in sich konsistent), mit geringem Zufallsfehler, jedoch stets inhaltlich völlig Unbeabsichtigtes messen. (Um diese Frage rankt sich z.B. jede Diskussion um „Kulturabhängigkeit" von Intelligenztests.) Diese Frage haben wir im vorigen Abschnitt nicht berücksichtigt sondern stattdessen angenommen, der personale Test-Mittelwert W sei das zu Messende. Über Definition eines geeigneten sogenannten „Außenkriteriums" C kann man sich der Validitätsfrage nähern: Zum Beispiel nimmt man für C die tatsächlich Arbeitsleistung bei einer bestimmten Arbeit, wenn T eine Testvariable ist, welche die spätere tatsächliche (Langzeit-) Arbeitsleistung vorhersagen soll. C hat im allgemeinen den Nachteil, dass man nur mühsam Werte davon bekommt. Aber immerhin kann man über die Korrelation $\rho(T, C)$ so etwas wie Validität von T bezüglich des Außenkriteriums C sinnvoll definieren und diese Validität empirisch ermitteln. Eine hohe Validität (eines Tests bezüglich eines Ausßenkriteriums) impliziert eine hohe Reliabilität - aber nicht umgekehrt. Bei hoher Validität korrelieren Test wie Retest hoch mit C, also korrelieren beide untereinander hoch. Quantitativ ergibt sich genauer: Wenn ein lineares Regressionsmodell

$$T = \alpha C + \beta + E$$

gilt mit hohem Korrelationskoeffizienten $\rho(T, C)$, so errechnet man analog zur Rechung oben bezüglich der Reliabiltät, *wenn* man C als nur von der Person abhängig betrachten kann, nicht von der Testdurchführung, weil dann $T_1 = \alpha C + E_1$, $T_2 = \alpha C + E$, und *wenn* man weiter E_1, E_2 als statistisch unabhängig voraussetzt:

$$\rho(T_1, T_2) = \frac{\alpha^2\sigma^2(C)}{\sigma^2(T)} = \alpha^2\rho^2(T, C). \ (*)$$

Die Reliabilität wird also durch α und die Validität bestimmt. Umgekehrt aber: Ist die Reliabilität hoch, so braucht kein solches Regressionsmodell $T = \alpha C + \beta + E$ zu gelten mit hohem $\rho(T, C)$; denn die hohe Reliabilität könnte auch ganz anders zustandekommen als über eine Korrelation mit C. In diesem Zusammenhang wird nun vielfach eine mystifizierende Rede über ein sogenanntes „Verdünnungsparadoxon" erhoben: Man steigere die Reliabilität und damit die Validität, letztere im Falle $\alpha = 1$ beliebig nahe an 1, im Falle $|\alpha| < 1$ gar über 1? Das ist natürlich grober Unfug, und es ist überhaupt kein Paradoxon vorhanden, man muss nur ein wenig genau nehmen, wovon man redet. Selbstverständlich verändern sich bei Testverlängerung der Regressionskoeffizient und die Validität genau derart, dass die obenstehende Reliabilitätsformel für Testvervielfachung herauskommt. Man sollte also nicht erst so tun, als könnte man links die Reliabilität erhöhen, rechts aber immer dieselbe *alte* Validität des ursprünglichen Tests bewahren. Weiter sollte klar sein: Wenn die Voraussetzung der statistischen Unabhängigkeit von E_1, E_2 nicht

zutrifft, dann ist die Gleichung $(*)$ falsch, und sehr wohl kann dann $\rho(T_1, T_2)$ viel größer als $\alpha^2\rho^2(T, C)$ sein.

Interessant ist folgende Frage: Kann man die Korrelation zwischen den personalen Testmittelwerten und dem Außenkriterium ausrechnen aus empirisch Zugänglichem? Kann man das auch in dem Falle tun, dass zusätzlich die Ermittlung der Werte des Ausßenkriteriums mit einem Zufallsfehler behaftet ist? Die Antwort ist positiv und wird durch die folgende Verdünnungsformel gegeben. Allerdings sollte man auch in diesem Kontext klarer als üblich sprechen und stets verdeutlichen, von welchern Variablen die Rede ist; $\rho(T, C)$ ist und bleibt, was es ist, es wird nicht größer durch eine mystifizierende Anwendung der Verdünnungsformel. Aber $\rho(W, T)$ (W wie oben die personale Mittelwertsgröße zu T) kann sehr wohl größer als $\rho(T, C)$ sein, und es ist sinnvoll, unter Validität eines Tests bezüglich eines Außenkriteriums den linearen Zusammenhang zwischen den personalen Mittelwerten (des Tests wie des Außenkriteriums) zu verstehen, also um die Abschwächung zu bereinigen, welche die sichtbare Korrelation zwischen T und C aufgrund der Zufallsfehler zeigt gegenüber der Korrelation zwischen W und U (wobei wir nunmehr unter C eine Messung des Außenkriteriums mit Zufallsfehler und unter U die ideale zugehörige Mittelwertsgröße verstehen wollen, deren Wert man sich jeweils nähert, wenn man die Messung C an ein und demselben Probanden oft wiederholt.)

3.4. Das Zurückrechnen einer durch Zufallsfehler abgeschwächten Korrelation: Verdünnungsformel. Noch einmal seien die Voraussetzungen betont: W und U seien zwei Variablen, deren Korrelation man kennenlernen möchte (wir denken insbesondere an den Fall: Idealer Testwert und idealer Außenkriteriumswert, Wiederholung bei derselben Person ergäbe also fehlerlos denselben Wert). Nun kann man diese aber nicht unmittelbar beobachten, stattdessen verfügt man nur im Sinne der Klassischen Testtheorie über Messvariablen T und C mit jeweiligem Zufallsfehler und kann auch empirisch nur $\rho(T, C)$ beobachten (d.h. durch $r(T, C)$ nähern). Intuitiv leuchtet ein, dass $\rho(T, C)$ kleiner als $\rho(W, U)$ sein sollte. Aber man kann die genaue quantitative Beziehung ausrechnen, wenn man ordentlich die Situation mit folgender Voraussetzung beschrieben hat:

$$\begin{aligned} T &= W + E \\ C &= U + F, \end{aligned}$$

wobei E linear unabhängig von W, $\mu(E) = 0$, F linear unabhängig von U, $\mu(F) = 0$, E statistisch unabhängig von U, F und F statistisch unabhängig von W. Man hat dann:

$$\begin{aligned} \rho(T, C) &= \frac{Cov(W + E, U + F)}{\sigma(T)\sigma(C)} = \frac{Cov(W, U)}{\sigma(T)\sigma(C)} = \frac{\rho(W, U)}{\sigma(T)/\sigma(W) \cdot \sigma(C)/\sigma(U)} \\ &= \rho(W, U) \cdot \frac{\sigma(W)}{\sigma(T)} \cdot \frac{\sigma(U)}{\sigma(C)} = \rho(W, U) \cdot \sqrt{\rho^2(W, T)}\sqrt{\rho^2(U, C)} \\ &= \rho(W, U) \cdot \sqrt{\rho(T_1, T_2)} \cdot \sqrt{\rho(C_1, C_2)}. \end{aligned}$$

Man kann also $\rho(W, U)$ folgendermaßen über die (zugänglichen!) Reliabilitäten der Messungen T und C sowie die zugängliche Korrelation $\rho(T, C)$ ausrechnen:

$$\rho(W, U) = \frac{\rho(T, C)}{\sqrt{\rho(T_1, T_2)} \cdot \sqrt{\rho(C_1, C_2)}}.$$

Der Nenner ist kleiner als 1, wenn nur eine der Reliabilitäten unter 1 liegt. Man sollte jedoch stets beachten: $\rho(T, C)$ bleibt, was es ist. Wenn man nun auf eine hohe Validität im Sinne der Korrelation der „wahren Variablen" kommt, so wird ein wenig reliabler Test und eine wenig reliable Messung des Außenkriteriums nicht praktisch nützlich!

KAPITEL 7

Eine Einführung in die multivariate Statistik

Multivariate Statistik behandelt mehrere Variablen $X_1, ..., X_n$ auf einmal. So betrachtet man z.B. die gemeinsame Verteilungsfunktion

$$F(\alpha_1, ..., \alpha_n) = P(X_1 \leq \alpha_1 \text{ und } X_2 \leq \alpha_2 \text{ und...und } X_n \leq \alpha_n).$$

Ist das ganze Variablensystem unabhängig, so kann man sie als Produkt der einzelnen eindimensionalen Verteilungsfunktionen erhalten. Ebenso wäre dann die zugehörige Dichte einfach das Produkt der eindimensionalen Dichtefunktionen. Interessant für Anwendungen ist jedoch gerade der Fall eines Abhängigkeitsgeflechtes eines Satzes von Variablen zu einem Thema. Ein Problem aus diesem Kreis besprachen wir bereits: Multiple Regression und Korrelation. Allgemeiner wäre die Aufgabe ins Auge zu fassen, wenigstens das *lineare* Beziehungsgeflecht eines Satzes von Variablen zu beschreiben - das reicht schon für sehr viele praktische Fälle aus. (Dem werden wir uns im zweiten Abschnitt dieses Kapitels widmen.) Weitere Standardthemen aus diesem Bereich sind: Hauptkomponentenanalyse („principal component analysis", kurz PCA), multidimensionale Skalierung („multidimensional scaling") - vgl. dazu Abschnitte 3 und 4. Faktorenanalyse gehört auch zu diesem Kreis, aber wir besprechen sie nicht, da wir ehrlich gesagt nicht viel davon halten; es gibt jedoch weitere interessante, insbesondere seien Clusteranalyse und Diskriminanzanalyse erwähnt, die den Rahmen einer kleinen Einführung überspannen würden. Allen diesen Methoden ist eines gemeinsam: Man hat eine Folge von Variablen zu betrachten und zu beschreiben, dazu geht es gerade zunächst um deren *lineare* Beziehungen; daher sind stets grundlegend die Methoden der Vektorrechnung bzw. Linearen Algebra: Eine ordentliche (gar nicht einmal sehr umfängliche) Auffassung der Anfangsgründe dieses Gebietes ist unerlässliche, aber auch schon völlig ausreichende Voraussetzung für das Verstehen in multivariater Statistik. Daher beginnen wir mit einem Abschnitt über diese Grundlagen.

1. Vektorrechnung und Lineare Algebra

1.1. Motivierung: Woher Vektoren? Wir stellen uns Variablen $X_1, ..., X_n$ vor, die alle auf derselben Population Ω definiert sind. Man kann nun die endliche Folge dieser Variablen $(X_1, ..., X_n)$ bilden und sie als *eine vektorielle Variable* auf demselben Ω auffassen, deren Werte eben nicht Zahlen, sondern Folgen von reellen Zahlen der Länge n sind: Jedem Individuum $\omega \in \Omega$ kommt dann als Wert die Folge $(X_1(\omega), ..., X_n(\omega))$ zu. - das ist ein Vektor der Dimension n. Beispiel: Testpunktzahlen für je einen sprachlichen, einen mathematischen und einen historischen Test ($n = 3$). Ebenso bildet $(X_1, ..., X_n)$ einen Vektor von Variablen. Vektoren bezeichnet man gern mit einem Pfeil über den Buchstaben, also $\vec{X}$ für $(X_1, ..., X_n)$, $\vec{x}$ für einen Vektor von Zahlen $(x_1, ..., x_n)$. Man unterscheide Folgen $(x_1, ..., x_n)$

von Mengen $\{x_1, ..., x_n\}$ - bei den Folgen kommt es auf die Reihenfolge an, und Wiederholungen gleicher Elemente sind nicht etwa wegzulassen!

1.2. Die Vektorräume $\mathbb{R}^n$, und das anschauliche Verständnis von Vektoren.

DEFINITION 25. *Der Vektorraum $\mathbb{R}^n$ ist die Menge aller $n-$ Tupel reeller Zahlen (oder Folgen der Länge n reeller Zahlen)*

$$\mathbb{R}^n := \left\{ \left(\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right) \middle| \, x_i \in \mathbb{R}, \ 1 \leq i \leq n \right\}$$

(Vektoren als Spalten geschrieben! Kurzbezeichnung für Vektoren: $\vec{x}$ usw.), zusammen mit folgenden Operationen (Verknüpfungen):

$$\left(\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right) + \left(\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \right) := \left(\begin{array}{c} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{array} \right)$$

(Vektoraddition) und

$$\alpha \left(\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right) := \left(\begin{array}{c} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{array} \right), \ \alpha \in \mathbb{R}.$$

Diese Operationen heißen die linearen (Vektorraum-) Operationen.

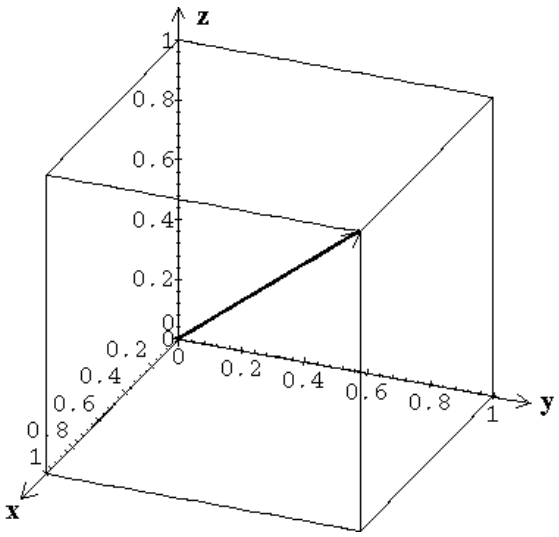
Beispiel (Vektoren platzsparend als Zeilenvektoren geschrieben):

$$3\,(1,2,-2) - \frac{1}{2}\,(2,-1,3) = \left(2, \frac{13}{2}, -\frac{15}{2}\right) = \frac{1}{2}\,(4,13,-15)\,.$$

Es gelten, wie man leicht nachprüft, ganz die vom Rechnen mit reellen Zahlen gewohnten Regeln, z.B. hat man $\alpha\,(\vec{x}+\vec{y}) = \alpha\vec{x}+\alpha\vec{y}$, $(\alpha+\beta)\,\vec{x} = \alpha\vec{x}+\beta\vec{x}$, $\alpha\,(\beta\vec{x}) = (\alpha\beta)\,\vec{x}$, $(-\alpha)\,\vec{x} = \alpha\,(-\vec{x})\,$, $1 \cdot \vec{x} = \vec{x}$. Der Vektor mit lauter Einträgen Null heißt Nullvektor und wir d mit $\vec{0}$ bezeichnet. Es gilt damit $\vec{0}+\vec{x} = \vec{x}$. Der Vektor $-\vec{x}$ hat alle Einträge von $\vec{x}$ mit umgedrehtem Vorzeichen, und man hat $-\vec{x}+\vec{x} = \vec{0}$. Einzige Ausnahme vom Gewohnten: Es macht generell keinen Sinn, durch einen Vektor zu dividieren. Weiter ist die (hier weggelassene) Multiplikation eines Vektors mit einem anderen eine neue (nicht zu den linearen Operationen gehörige) Operation mit einem gewissen geometrischen Sinn. Beim sogenannten Skalarprodukt kommt dabei eine Zahl heraus und nicht etwa ein Vektor.

Wir kommen zum anschaulichen Verständnis von Vektoren, das unmittelbar nur in Dimensionen 1 bis 3 gelingt, aber doch (irgendwie) auf höhere Dimensionen sich überträgt.Für $n = 2, 3$ sieht das so aus: Im anschaulichen zweidimensionalen Punktraum E^2 (stellen Sie sich im dreidimensionalen Anschauungsraum eine Ebene vor, zum Beispiel die, welche durch dies gerade hingelegte Blatt gegeben ist) führen Sie ein Koordinatensystem ein: Typisch (so bei kartesischem System) zwei senkrecht aufeinander stehende Achsen mit gleicher Einheitenaufteilung. Dann machen Sie

aus einem Zahlenpaar wie $(2,-3)$ in gewohnter Weise einen Punkt. Allgemein können Sie auf beide genannten Bedingungen für das Koordinatensystem verzichten, dann funktioniert die Sache immer noch. (Machen Sie sich das durch eine Skizze klar!) Sie haben nunmehr folgende Veranschaulichungen für ihr Zahlenpaar (oder 2-Tupel): Punkt in der Ebene, oder auch Vektorpfeil, der im Ursprung beginnt und in diesem Punkt endet. Diesen Vektor(pfeil) nennt man auch den Ortsvektor jenes Punktes. Schließlich können Sie auch noch alle Vektorpfeile, die gleich lang und parallel sind, miteinander identifizieren, sogenannte freie Vektoren bilden, und auch sie zur Interpretation der Zahlenpaare heranziehen. Alles Gesagte funktioniert auch in drei Dimensionen, wenn man ein dreibeiniges Koordinatensystem verwendet. Folgende Skizze sollte das verständlich machen:



Sie sehen drei Achsen ($x-, y-, z-Achse$), die Sie sich im vorliegenden Falle als senkrecht aufeinander und mit gleichen Einheiten versehen vorstellen (kartesisches System) - was wiederum nicht notwendig wäre (es genügt, wenn die Achsen nicht auf einer Ebene liegen!). Der Beispielvektor ist das Zahlentripel (1,1,1), und Sie sehen den zugehörigen Punkt und den zugehörigen Ortsvektor als Pfeil. Weiter ist der achsenparallele Würfel eingezeichnet, der die Strecke vom Koordinatenursprung bis zum Punkt mit der Koordinatendarstellung (1,1,1) zur Diagonalen hat. (Bei beliebigen Koordinaten erhält man einen Quader, bei nichtkartesischen Systemen so etwas wie einen „schiefen Quader", genannt „Spat".)

Wir können demnach reelle Zahlentupel ganz ähnlich im Raum entsprechender Dimension veranschaulichen wie reelle Zahlen auf der Zahlengeraden. Dabei benutzen wir folgende Sprechweisen:

Zum Punkt P im $n-$ dimensionalen Raum gehört *bezüglich eines Koordinatensystems* K umkehrbar eindeutig die Koordinatendarstellung

$$\begin{aligned}(x_1^K(P),...,x_n^K(P)) &= \textit{Koordinatendarstellung von } P \textit{ bezüglich } K \\ &= \overrightarrow{x}_P^K = \textit{Koordinatendarstellung des Ortsvektors } \overrightarrow{x}_P \textit{ bezüglich } K.\end{aligned}$$

Dem zugehörigen freien Vektor verpasst man bezüglich K wiederum dieselbe Koordinatendarstellung.

Bemerkung: Es gibt auch unendlich-dimensionale Vektorräume, aber hier setzen wir stets endliche Dimension voraus.

Insbesondere sollte man sich vorstellen, dass die Folgen der Werte von vielen Variablen bei einer Stichprobe von Individuen einen Punkt in einem (eventuell hochdimensionalen) Raum darstellen. Dann wird man insbesondere danach fragen, ob man das Wesentliche davon schon in einem Raum wesentlich geringerer Dimension darstellen und eventuell sogar aufmalen kann. (Dieser Aspekt wird bei der Behandlung der PCA in Abschnitt 3 mitbehandelt werden.) Ein weiterer grundlegender Gesichtspunkt ist der des Abstandes zwischen Punkten (oder vektoriellen Werten): Man wird gewisse Individuen benachbart, andere Gruppen in größeren Abständen sehen. Hier kann man sehr einfach die Pythagorasformel verallgemeinern:

1.3. Längen von Vektoren und Abstände zwischen Punkten. Wir setzen hier ein kartesisches Koordinatensystem voraus, also rechtwinklige Achsen und gleiche Achseneinheiten! Zunächst einmal ist klar:

DEFINITION 26. *Der Abstand zwischen zwei Punkten P, Q im E^n ist dasselbe wie die Länge des Vektors $\vec{x}_Q - \vec{x}_P$.*

In Dimension $n = 2$ können wir klar sehen, dass die Länge des Vektors $(a, b) \in \mathbb{R}^2$ (oder der Abstand dieses Punktes vom Ursprung $(0, 0)$) dasselbe ist wie $\sqrt{a^2 + b^2}$, nach Pythagoras. Denn wir haben ein Dreieck mit rechtem Winkel, dessen eine Kathete die Länge $|a|$ und dessen andere Kathete die Länge $|b|$ hat. Dies verallgemeinert man erfolgreich mit:

DEFINITION 27. *Die Länge des Vektors aus $\mathbb{R}^n$ ist allgemein:*

$$\left| \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right| := \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Beispiel: $|(2, -3, 4)| = \sqrt{4 + 9 + 16} = \sqrt{29}$.

1.4. Matrizen und lineare Abbildungen zwischen Vektorräumen.

1.4.1. *Motivierung: Woher Matrizen?* Nehmen wir einen üblichen Test: Der besteht aus mehreren Punkten („items") $Item_1, ..., Item_n$, und wenn wir ihn mit Personen $Person_1, ..., Person_m$ durchführen, so haben wir automatisch eine $(m \times n)-$ Matrix mit m Zeilen und n Spalten, deren Eintrag in Zeile i und Spalte j lautet: Punktzahl von Person Nr. i bei Item Nr. j. (Diese Matrix schlachtet man dann in der Faktorenanalyse, um das Wesentliche mit viel weniger Zahlangaben beschreiben zu können.)

Ein zweites Beispiel (mit dem wir uns noch ausführlicher beschäftigen werden): Unser Eingangsproblem lautete, das Geflecht der linearen Beziehungen zwischen den Variablen $X_1, ..., X_n$ zu beschreiben. Wie sich zeigen wird, trägt folgender einfache naheliegende Ansatz dafür bereits sehr weit: Man betrachtet alle paarweisen Kovarianzen $Cov(X_i, X_j)$, $1 \leq i, j \leq n$, und fasst diese naheliegend zu einer $(n \times n)-$ Matrix zusammen, in deren $i-$ ter Zeile und $j-$ ter Spalte der Eintrag $Cov(X_i, X_j)$ steht. Ganz analog kann man die zugehörige Korrelationsmatrix bilden, deren Einträge eben $\rho(X_i, X_j)$ lauten.

1.4.2. *Definition reellwertiger Matrizen und Schreibweisen.*

DEFINITION 28 (bequeme Bezeichnung für Matrizen). *Eine $(m \times n)-$ Matrix ist allgemein ein rechteckiges Zahlenschema (hier stets: reeller Zahlen) mit m Zeilen und n Spalten:*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Dabei sind folgende Bezeichnungen nützlich: Der Eintrag in der $i-$ ten Zeile und $j-$ ten Spalte wird gern mit a_{ij} bezeichnet, die ganze Matrix mit zugehörigem Großbuchstaben. Aber diese Notation ist immer noch sehr schwerfällig, eleganter benutzt man:

$$\begin{aligned} A &= (a_{ij})_{ij}\,,\ (1 \le i \le m, 1 \le j \le n), \\ (A)_{ij} &: = a_{ij}. \end{aligned}$$

Die erste Bezeichnung erlaubt es, aus den Einträgen a_{ij} die Doppelfolge, d.h. die Matrix selbst zu bilden. Letztere Bezeichnung ermöglicht den Übergang von der ganzen Matrix zu den Einträgen.

Das führen wir noch einmal am Beispiel einer Kovarianzmatrix aus:

DEFINITION 29. *Die Kovarianzmatrix einer vektoriellen Variablen $\vec{X} = (X_1, ..., X_n)$ ist definiert als*

$$Cov\left(\vec{X}\right) := \begin{pmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_1, X_n) \\ \vdots & \vdots & & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \cdots & Cov(X_n, X_n) \end{pmatrix}.$$

Das ist also ein quadratisches Zahlenschema, bei dem der Eintrag in der $i-$ ten Zeile und der $j-$ ten Spalte $Cov(X_i, X_j)$ lautet. Mit der eingeführten eleganteren Notation:

$$Cov\left(\vec{X}\right) := (Cov(X_i, X_j)_{ij}\,,\ (1 \le i, j \le n).$$

Bemerkung: Offenbar stehen in der Hauptdiagonalen (von links oben nach rechts unten) der Kovarianzmatrix die Varianzen der Variablen, da $Cov(X_i, X_i) = \sigma^2(X_i)$ gilt. Ferner ist die Matrix symmetrisch, d.h. sie wird in sich überführt bei Spiegelung der Einträge an der Hauptdiagonalen, da $Cov(X_i, X_j) = Cov(X_j, X_i)$. Ferner ist die Matrix quadratisch, d.h. sie hat ebenso viele Zeilen wie Spalten.

Wir werden noch sehen, dass die Indexschreibweisen insbesondere für die Operationen mit Matrizen sehr praktisch sind.

1.4.3. *$(m \times n)-$Matrizen als lineare Abbildungen $\mathbb{R}^n \to \mathbb{R}^m$.* Eine lineare Abbildung ist nichts anderes als ein Vektorraumhomomorphismus, d.h. die Abbildung vertauscht mit den linearen Operationen (das sind die zur Vektorraumstruktur gehörigen Operationen):

DEFINITION 30. *Eine Abbildung $f : \mathbb{R}^n \to \mathbb{R}^m$ heißt genau dann linear, wenn folgende Bedingungen für alle $\vec{x}, \vec{y} \in \mathbb{R}^n$ und alle $\alpha \in \mathbb{R}$ erfüllt sind:*

$$\begin{aligned} f(\vec{x}+\vec{y}) &= f(\vec{x}) + f(\vec{y}), \\ f(\alpha\vec{x}) &= \alpha f(\vec{x}). \end{aligned}$$

Beispiel: Die Abbildung $f((x,y)) = (2x-3y, -3x+2y)$ für $(x,y) \in \mathbb{R}^2$ ist linear, aber bereits die Abbildung $g((x,y)) = (2x-3y+1, -3x+2y)$ ist es nicht mehr (keine additiven Konstanten zulässig!), da $g\left(\vec{0}\right) = (1,0)$, aber es müsste mit der zweiten Bedingung $g\left(\vec{0}\right) = g\left(0 \cdot \vec{0}\right) = 0g\left(\vec{0}\right) = \vec{0}$ sein. (Wir nutzten wiederum platzsparend Zeilennotation.)

Nun ist es mittels der Matrizen leicht, explizit *alle* linearen Abbildungen $\mathbb{R}^n \to \mathbb{R}^m$ zu beschreiben, und man kann die Anwendung einer solchen Abbildung auf einen Vektor stets als Multiplikation „Matrix mal Vektor" gewinnen:

DEFINITION 31 (Matrix mal Vektor). *Sei A eine $(m \times n)-$ Matrix. Dann definiert A folgende Abbildung:*

$$\begin{aligned} A: \quad & \mathbb{R}^n \to \mathbb{R}^m \\ & \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} := \\ & \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + ...a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + ...a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + ...a_{mn}x_n \end{pmatrix}. \end{aligned}$$

Kurzbezeichnung: $\vec{x} \mapsto A\vec{x}$.

Wesentlich elegantere Version mit Indexschreibweise und großem Summenzeichen:

$$\begin{aligned} A: \quad & \mathbb{R}^n \to \mathbb{R}^m \\ & (x_j)_{1 \le j \le n} \mapsto (a_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}} (x_i)_{1 \le j \le n} := \left(\sum_{j=1}^{n} a_{ij} x_j \right)_{1 \le i \le m} \end{aligned}$$

Beispiel:

$$\begin{pmatrix} -1 & 2 \\ 3 & 5 \\ 4 & -6 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 14 \\ 6 \end{pmatrix}.$$

Hinweis: Man lege den Spaltenvektor umklappend auf die Zeilen, multipliziere übereinanderliegende Zahlen und summiere jeweils für eine Zeile darüber. Erste Resultatkomponente also: $(-1) \cdot 3 + 2 \cdot 1 = -1$.

SATZ 18. *Jede Abbildung der Form $\vec{x} \mapsto A\vec{x}$ ist linear, und jede lineare Abbildung $\mathbb{R}^n \to \mathbb{R}^m$ wird in dieser Weise durch eine Matrix gegeben.*

Wir begründen nur den weniger banalen Teil - die Linearität von A als Abbildung ist banal auszurechnen: Sei $f : \mathbb{R}^n \to \mathbb{R}^m$ linear. Wir bezeichnen nun mit $\vec{e}_j$ den Vektor aus $\mathbb{R}^n$, dessen $j-$ te Komponente den Wert 1 hat und dessen andere Komponenten alle Null sind. Dann haben wir offenbar für den *beliebigen* Vektor

${}^t(x_1,...,x_n) = \vec{x} \in \mathbb{R}^n$ (gemeint ist mit dem hochgestellten ^{t}, dass diese Zeile als Spalte zu schreiben ist - vgl. später: Transposition von Matrizen): $\vec{x} = \sum x_j \vec{e}_j$, also mit der Lineartität: $f(\vec{x}) = \sum x_j f(\vec{e}_j)$. Wir sehen also, dass mit den Bildern der Einheitsvektoren $\vec{e}_j$ bereits alle anderen eindeutig auszurechnen sind, wenn eine Abbildung linear ist (!). Jetzt brauchen wir nur die Matrix A zu f so anzugeben, dass $A\vec{e}_j = f(\vec{e}_j)$ für alle j gilt. Dazu genügt die Beobachtung: $B\vec{e}_j = j-$ ter Spaltenvektor von B, für jede Matrix B. Wir schreiben daher in die Spalten der gesuchten Matrix A die Bildvektoren $f(\vec{e}_j)$ und haben die verlangte Matrix A.

1.4.4. *Die Hintereinanderschaltung linearer Abbildungen und das Produkt von Matrizen.* Wenn wir zwei lineare Abbildungen $f : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}^s$ haben, so können wir mit $(g \circ f)(\vec{x}) := g(f(\vec{x}))$ die Hintereinanderschaltung bilden und logisch einsehen, dass diese wieder linear sein muss. Also wird auch sie durch eine Matrix gegeben, fragt sich nur, wie man diese Matrix aus den Matrizen zu f, g bestimmen kann. Die Antwort gibt folgender Satz:

SATZ 19. *Wird die lineare Abbildung $f : \mathbb{R}^n \to \mathbb{R}^m$ durch die $(m \times n)-$ Matrix A und die lineare Abbildung $g : \mathbb{R}^m \to \mathbb{R}^s$ durch die $(s \times m)-$ Matrix B gegeben, so wird $g \circ f$ durch die wie folgt definierte Produktmatrix BA dargestellt:*

DEFINITION 32 (Produkt zweier Matrizen). *Sei $A = (a_{ij})_{ij}$ eine $(m \times n)-$ Matrix und $B = (b_{ki})_{ki}$ eine $(s \times m)-$ Matrix. Dann ist definiert:*

$$BA = (b_{ki})_{ki}\,(a_{ij})_{ij} = \left(\sum_{i=1}^{m} b_{ki}a_{ij}\right)_{kj}.$$

Das ist nach Definition eine $(s \times n)-$ Matrix.

Bemerkung zur Definition: Bei unpassenden Dimensionen ist das Matrizenprodukt nicht definiert, zur Bildung von BA ist erforderlich: Zeilenzahl von $A =$ Spaltenzahl von B. Matrix mal Vektor kann als Spezialfall angesehen werden, wenn man den Vektor als Matrix mit nur einer Spalte betrachtet. Andererseits kann man nach operativem Verständnis von Matrix mal Vektor leicht zu Matrix mal Matrix verallgemeinern: Zur Bildung von BA wende man der Reihe nach B auf jeden Spaltenvektor von A an und schreibe die Resultate als Spalten der Matrix BA. Genau das verlangt obenstehende formale Definition.

Man probiere diese Beschreibung, aber auch die formale Definition aus an folgendem Beispiel:

$$\begin{pmatrix} 1 & -2 \\ 3 & 4 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} 3 & -2 & 1 \\ -1 & 2 & -4 \end{pmatrix} = \begin{pmatrix} 5 & -6 & 9 \\ 5 & 2 & -13 \\ 9 & -10 & 14 \end{pmatrix}.$$

1.4.5. *Inversenbildung bei quadratischen (d.h. $(n \times n)-$) Matrizen.* Wir kennen bereits inverse Abbildungen, nämlich Umkehrabbildungen: $f^{-1}(f(x)) = x$. f^{-1} macht dabei die Operation f rückgängig, z.B. $\ln(e^x) = x$. Nun sind quadratische $(n \times n)-$ Matrizen nichts als lineare Abbildungen $\mathbb{R}^n \to \mathbb{R}^n$. Wenn nun A die lineare Abbildung $f : \mathbb{R}^n \to \mathbb{R}^n$ darstellt und B die inverse Abbildung f^{-1}, dann muss gelten: $BA = E_n$, wobei E_n die $(n \times n)-$ Einheitsmatrix ist, mit lauter Einsen auf der Hauptdiagonalen, sonst Nullen. Denn man hat $E_n\vec{x} = \vec{x}$, diese Matrix stellt also die identische Abbildung dar, und es soll ja gerade $BA\vec{x} = \vec{x}$ gelten, so dass Anwendung von B die Anwendung von A rückgängig macht. Daher definieren wir:

DEFINITION 33. *B heißt Inverse zur $(n \times n)-$ Matrix A, wenn $BA = E_n$ gilt. Wenn B mit dieser Eigenschaft existiert, so ist B eindeutig bestimmt und wird A^{-1} genannt.*

Die behauptete Aussage folgt sofort daraus, dass die Umkehrabbildung einer umkehrbaren Abbildung eindeutig bestimmt ist. Nun ist aber nicht jede Abbildung umkehrbar, auch nicht jede lineare. Aber wir merken uns: Eine quadratische Matrix ist *in aller Regel* umkehrbar (invertierbar). Wir nennen zwei Charakterisierungen der Invertierbarkeit ohne Beweis:

SATZ 20. *Eine quadratische Matrix A ist genau dann invertierbar, wenn das lineare Gleichungssystem $A\vec{x} = \vec{0}$ als einzige Lösung $\vec{x} = \vec{0}$ besitzt. Ebenso ist A genau dann invertierbar, wenn die Determinante von A verschieden von Null ist.*

Da quadratische Matrizen in aller Regel invertierbar sind, heißen die invertierbaren auch „regulär" (die andern „singulär").

Bemerkung zur Berechnung der Inversen zu einer gegebenen quadratischen Matrix A: Die Bedingung $A^{-1}A = E_n$ läuft unmittelbar auf das simultane Lösen von n linearen $(n \times n)-$ Gleichungssystemen hinaus. Das kann (nicht nur numerisch, sondern auch symbolisch-exakt) auch vom Computer getan werden. (Für Riesensysteme, etwa $n = 100000$, braucht man natürlich ausgefuchstere numerische Methoden, aber im Rahmen der in der Psychologie auftretenden multivariaten Statistik bleibt alles im harmlosen Bereich, und Standardprogramme tun den Dienst.)

Beispiel: Zu

$$C = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix} \text{ lautet}$$

$$C^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Rechnen Sie nach, dass tatsächlich gilt:

$$C^{-1}C = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = E_3.$$

1.4.6. *Orthogonale Matrizen bzw. lineare Abbildungen, und Transposition von Matrizen.* Oben stellten wir alle linearen Abbildungen $\mathbb{R}^n \to \mathbb{R}^m$ durch $(m \times n)-$ Matrizen dar. Nunmehr fragen wir nach den Abbildungen, die Längen und Winkel erhalten (es genügt die Erhaltung der Längen, wie man zeigen kann). Dazu lässt sich einsehen, dass eine solche Abbildung insbesondere linear sein muss: Längen- und Winkel erhaltende Abbildungen sind also spezielle lineare Abbildungen. Man nennt sie orthogonale Abbildungen, weil sie insbesondere senkrecht aufeinander stehende Vektoren immer auf solche abbildet. Weiter kann man zeigen, dass eine solche Abbildung $\mathbb{R}^n \to \mathbb{R}^m$ nur für $m \geq n$ möglich ist. Denn sonst muss ein Vektor $\vec{x} \neq \vec{0}$ auf $\vec{0}$ abgebildet werden, so dass dessen Länge damit nicht erhalten bleibt. Nun kann man aber eine orthogonale Abbildung $\mathbb{R}^n \to \mathbb{R}^m$ mit $m \geq n$ immer auffassen als eine Hintereinanderschaltung von zuerst einer orthogonalen Abbildung $\mathbb{R}^n \to \mathbb{R}^n$ mit eventueller anschließenden Einbettung von $\mathbb{R}^n$ in $\mathbb{R}^m$. Damit sind wir bei der Frage angelangt, welche $(n \times n)-$ Matrizen denn orthogonale Abbildungen darstellen. (Entsprechende Matrizen nennt man dann auch orthogonale Matrizen).

Die Antwort ist sehr einfach: Wir haben schon gesehen, dass die Spaltenvektoren einer Matrix genau die Bilder der Einheitsvektoren ergeben. Nun sind diese alle der Länge 1 und stehen paarweise senkrecht aufeinander (mit dem üblichen Euklidischen Skalarprodukt $(x_i)_i (y_i)_i := \sum_i x_i y_i$ sind zwei Vektoren genau senkrecht aufeinender, wenn dies Skalarprodukt von ihnen verschwindet). Also kann eine Matrix $(n \times n) -$ Matrix A nur dann orthogonal sein, wenn ihre Spaltenvektoren alle die Länge 1 haben und paarweise senkrecht aufeinander stehen. Man kann nun wiederum zeigen, dass dies auch genügt. Wir wiederholen die Definition und formulieren das Fazit:

DEFINITION 34. *Eine lineare Abbildung $\mathbb{R}^n \to \mathbb{R}^n$ heißt orthogonal, wenn sie Längen und Winkel erhält. (Dafür genügt es bereits, dass alle Längen erhalten bleiben, also stets gilt:*

$$f(|\vec{x}| = |f(\vec{x})|.)$$

Bemerkung: Man formuliert gern die Bedingung als Erhaltung des Skalarprodukts.

In diesem Kontext tritt eine sehr einfache Matrizenoperation auf, das ist die Transposition. Diese kann man für beliebige $(m \times n) -$ Matrizen bilden:

DEFINITION 35. *Die Transponierte der Matrix $(m \times n) -$ Matrix $A = (a_{ij})_{ij}$ erhält man, indem man die Spalten von A als Zeilen schreibt. Die Transponierte von A ist also eine $(n \times m) -$ Matrix. Man schreibt ${}^tA$ für die Transponierte von A. Also formal:*

$$\begin{aligned} {}^t\left((a_{ij})_{ij}\right) &: = (a_{ji})_{ji}, \text{ und entsprechend} \\ \left({}^tA\right)_{ji} &: = (A)_{ij}. \end{aligned}$$

Beispiel:

$$\text{Für } A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \text{ ist } {}^tA = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}.$$

Bemerkung: Wenn eine quadratische Matrix symmetrisch ist (bezüglich Spiegelung der Einträge an der Hauptdiagonalen), so ist sie offenbar gleich ihrer Transponierten. So ist es bei der Kovarianzmatrix.

SATZ 21. *Eine lineare Abbildung $\mathbb{R}^n \to \mathbb{R}^n$, durch eine Matrix $(n \times n) - A$ gegeben, ist genau dann orthogonal, wenn die Spaltenvektoren von A alle Betrag 1 haben und paarweise senkrecht aufeinander stehen. Dann gilt dasselbe stets auch für die Zeilenvektoren, und man hat*

$$A^{-1} = {}^tA.$$

Man beachte, dass diese Gleichung *nur für orthogonale Abbildungen* gilt, nicht etwa allgemein für Inversenbildung!

Beispiel: Eine orthogonale $(2 \times 2) -$ Matrix ist z.B.

$$A = \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix}.$$

Betrachten wir nämlich den ersten Spaltenvektor, so sehen wir, dass dessen Betrag lautet: $\sqrt{1/2 + 1/2} = 1$, analog für den zweiten Spaltenvektor. Das Skalarprodukt

der beiden Spaltenvektoren ist Null (erste Komponente mal erste plus zweite mal zweite), also stehen sie senkrecht aufeinander, wie man auch beim Einzeichnen in ein kartesisches System zu sehen bekäme. Und wie vom Satz versprochen ist tatsächlich die Transponierte gleich der Inversen, es gilt nämlich

$$ {}^tAA = \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \\ -\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. $$

1.4.7. *Eigenvektoren und Eigenwerte einer quadratischen (also $(n \times n)-$) Matrix.* Vielfach ist es bei einer $(n \times n)-$ Matrix interessant (z.B. in den ersten beiden der nachfolgenden Anwendungen wird das eine große Rolle spielen), danach zu fragen, ob es Vektoren $\vec{x} \neq \vec{0}$ gibt mit der Eigenschaft, dass ihr Bild unter A einfach ein Vielfaches von $\vec{x}$ ist, also gilt $A\vec{x} = \lambda\vec{x}$ mit $\lambda \in \mathbb{R}$. Dann heißt λ ein *Eigenwert von A* und $\vec{x}$ ein *Eigenvektor von A.* Dazu gibt es ein grundlegendes Resultat, das wir nutzen werden:

Satz 22. *Eine symmetrische $(n \times n)-$ Matrix A, die positiv definit ist, d.h. ${}^t\vec{x}A\vec{x}$ ist für jeden Spaltenvektor $\vec{x} \neq \vec{0}$ eine positive Zahl > 0, hat stets n Eigenwerte $\lambda_1, ..., \lambda_n$, alle > 0, derart dass für eine orthogonale Matrix B gilt:*

$$ {}^tBAB = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \lambda_n \end{pmatrix}. $$

Letztere nennt man eine Diagonalmatrix; sie hat außerhalb der Hauptdiagonalen nur Einträge Null. (Nach dem vorigen Satz ist dabei ${}^tB = B^{-1}$.) Dabei ist zusätzlich der $i-$ te Spaltenvektor von B Eigenvektoren von A zum Eigenwert λ_i, für $1 \leq i \leq n$.

Dieser Satz ist für uns hier nützlich, weil Kovarianzmatrizen in allen interessanten Fällen diese Eigenschaft haben - lediglich könnte es passieren, dass Eigenwerte 0 noch auftreten, was bedeuten würde, dass man eine oder mehrere der Variablen X_i gleich weglassen könnte, da sie sich in der Form $X_i = \sum_{j \neq i} \alpha_j X_j$ schreiben lassen.

1.4.8. *Das abstrakte Rechnen mit Matrizen, mit den Operationen der Addition, Multiplikation, Inversenbildung.*

Definition 36. *Die Summe zweier $(m \times n)-$ Matrizen $A = (a_{ij})_{ij}$ und $B = (b_{ij})_{ij}$ lautet*

$$ A + B := (a_{ij} + b_{ij})_{ij}. $$

Man addiert also einfach komponentenweise (gerade so wie bei Vektoren).

Definition 37. *Das Produkt einer reellen Zahl λ mit einer Matrix $(a_{ij})_{ij}$ ist die Matrix $(\lambda a_{ij})_{ij}$. (Es wird also jeder Matrixeintrag mit λ multipliziert.*

Bemerkung: Mit diesen Operationen bilden alle $(m \times n)-$ Matrizen einen Vektorraum.

Nun hat man aber auch noch die Multiplikation von Matrizen. Mit den entstehenden Rechenausdrücken rechnet man wie vom Zahlenrechnen her gewohnt, mit zwei wesentlichen *Ausnahmen:* 1) Im allgemeinen ist $AB \neq BA$ (das kennen wir als das Phänomen, dass man die Reihenfolge bei Hintereinanderschaltungen von Abbildungen in aller Regel nicht umkehren darf). 2) Division durch eine

Matrix hat nur bei quadratischen Matrizen Sinn, die invertierbar sind, und man schreibt A^{-1}, nicht $1/A$. Ansonsten gelten die gewohnten Assoziativgesetze wie $(A+B)+C = A+(B+C)$, $(AB)C = A(BC)$, vor allem die Distributivgesetze $A(B+C) = AB+AC$, $(A+B)C = AC+BC$, ferner $A+B = B+A$. Die Einheitsmatrix ist (bei quadratischen Matrizen) das neutrale Element bezüglich der Multiplikation, die Nullmatrix (lauter Nullen) das neutrale bei der Addition. Schließlich beachte man, dass $(AB)^{-1} = B^{-1}A^{-1}$ (wenn A^{-1}, B^{-1} existieren). Das folgt aus der Logik des Hintereinanderschaltens und Umkehrens von Abbildungen: Wendet man zuerst B, dann A an, so wird das rückgängig gemacht, indem man zuerst A umkehrt, dann B umkehrt. Man kann auch rechnen, unter Voraussetzung lauter invertierbarer $(n \times n) -$ Matrizen:

$$\left(B^{-1}A^{-1}\right)(AB) = B^{-1}\left(A^{-1}A\right)B = B^{-1}E_nB = B^{-1}B = E_n.$$

Dieselbe Sache ergibt sich auch für die oben beschriebene Transposition:

$${}^t(AB) = {}^tB{}^tA$$

Nach dieser Einführung in Vektorrechnung und Lineare Algebra sind wir in der Lage, die wichtigsten Anwendungen der Linearen Algebra in der multivariaten Statistik zu verstehen. Wir bringen nun drei verschiedene Anwendungen in den nächsten drei Abschnitten.

2. Kovarianzmatrix, multiple und Partialkorrelationen

Wir stellen uns vor, zu einem Thema hätten wir Variablen $X_1, ..., X_n$ (alle auf derselben Population Ω) gebildet und wollten nunmehr als Ganzes das Geflecht der linearen Beziehungen zwischen diesen Variablen überschauen. Ein Beispiel wäre: Eine komplexe Fähigkeit besteht aus mehreren einzelnen Fähigkeiten, und die X_i sind zugehörige Testvariablen. Insbesondere werden wir folgende Fragen stellen:

- Wie lauten die multiplen Korrelationen von einer der Variablen bezüglich aller anderen?

Diese Frage haben wir bereits beantwortet, aber unsere bisherige Lösung des Problems sah vor, dass wir jeweils eine der Variablen vornehmen und durch die übrigen linear zu erklären versuchen. Das müssten wir dann n mal tun - viel langweilige Arbeit. Wir suchen nach einer Methode, das alles auf einen Schlag zu erledigen.

- Das zweite Problem kommt neu hinzu: Zuweilen hat man eine Korrelation zwischen zwei Variablen, sagen wir X, Y, die einfach zu erklären ist aus beider Korrelationen mit einer dritten Variablen („Y-Modell"). Geht die Korrelation zwischen X, Y ganz in dieser Beziehung zu Z auf? Man möchte also herausfinden, welcher Zusammenhang noch zwischen X und Y besteht, wenn man den zu Z „herausrechnet". Dazu bildet man den sogleich zu definierenden „Partialkorrelationskoeffizienten" $\rho(X, Y|Z)$. Allgemeiner stellt sich das Problem bei unserer Ausgangssituation von Variablen $X_1, ..., X_n$ so: Man finde alle Partialkorrelationen $\rho(X_i, X_j|X_1, ..., X_n$ ohne $X_i, X_j)$.

Wir werden nun eine Methode angeben - sie besteht in der genaueren Betrachtung der Kovarianzmatrix $Cov\left(\vec{X}\right) = Cov\left((X_1, ..., X_n)\right)$, die wir bereits einführten, wie man *beide Probleme* global auf einen Schlag mittels mathematischer Standardoperationen lösen kann, ohne also jeweils einzeln all die Rechnungen vornehmen

zu müssen. Zuvor jedoch müssen wir erst einmal klären, wie das Herauspartialisieren von Korrelationen sinngemäß zu verstehen, zu definieren und im Einzelnen zu berechnen wäre.

2.1. Die Idee der Partialkorrelation. Man kann sicherlich nicht einfach $\rho(X,Z)$ und $\rho(Y,Z)$ von $\rho(X,Y)$ abziehen, um sinngemäß die letzteren Korrelationen aus der ersteren „herauszurechnen" und so etwas wie $\rho(X,Y|Z)$ zu bestimmen. Aber das Abziehen von Variablen macht Sinn: Wir erklären beide Variablen X und Y zunächst linear aus Z, bilden also $\widehat{X}(Z) = a_{X.Z}Z + b_{X.Z}$ sowie $\widehat{Y}(Z) = a_{Y.Z}Z + b_{Y.Z}$ mit den jeweiligen Regressionskoeffizienten. Nun können wir sagen, was „übrigbleibt":

$$\begin{aligned}\widetilde{X} &: = X - \widehat{X}(Z),\\ \widetilde{Y} &: = Y - \widehat{Y}(Z).\end{aligned}$$

Damit haben wir den (linearen) Anteil von Z aus X und Y herausgerechnet und können $\rho(X,Y|Z)$ einfach als $\rho\left(\widetilde{X},\widetilde{Y}\right)$ definieren. Man kann nun folgende Formel angeben, welche diesen Koeffizienten auf die absoluten einfachen Paarkoeffizienten zurückführt:

$$\rho(X,Y|Z) = \frac{\rho(X,Y) - \rho(X,Z)\rho(Y,Z)}{\sqrt{\left(1-\rho^2(X,Z)\right)\left(1-\rho^2(Y,Z)\right)}}.$$

Das sieht bereits recht kompliziert aus - dazu ist es noch ziemlich unnütz für unser viel allgemeineres Problem, auf einen Schlag alle $\rho(X_i,X_j|X_1,...,X_n \text{ ohne } X_i,X_j)$ zu finden für gegebene $X_1,...,X_n$. Aber die oben angegebene Idee des Herausrechnens können wir ohne weiteres übertragen und in folgende Definition gießen:

Definition 38 (Partialkorrelation). *Der partielle Korrelationskoeffizient lautet allgemein*

$$\rho\left(X,Y|Z_1,...,Z_n\right) := \rho\left(\widehat{X}\left(Z_1,...,Z_n\right),\widehat{Y}\left(Z_1,...,Z_n\right)\right).$$

Dabei sind $\widehat{X}\left(Z_1,...,Z_n\right)$, $\widehat{Y}\left(Z_1,...,Z_n\right)$ die oben eingeführten linearen Schätzgrößen, und auf der rechten Seite steht der gewöhnliche einfache Korrelationskoeffizient. Noch einmal: $\rho\left(X,Y|Z_1,...,Z_n\right)$ beschreibt die Korrelation zwischen X und Y, beider lineare Beziehung zu $Z_1,...,Z_n$ herausgerechnet.

Wir kehren zurück zu unseren beiden Ausgangsproblemen, das lineare Beziehungsgeflecht eines vorgelegten Satzes $X_1,...,X_n$ von Variablen auf einen Schlag zu beschreiben. Wir könnten ein wenig hoffen (ohne dass dies irgendwie selbstverständlich wäre!), dass unsere Aufgabe mittels der paarweisen Kovarianzen $Cov(X_i,X_j)$ lösbar sein möge, die alle in der Kovarianzmatrix versammelt sind.

2.2. Die Lösung der Ausgangsprobleme über die Inverse der Kovarianzmatrix. Nunmehr können wir sogleich das Hauptresultat formulieren und nutzen - man beachte, dass es für jede Anzahl n der Variablen gilt:

Satz 23. *Sei $C = \left(Cov\left(X_i,X_j\right)\right)_{ij}$ die Kovarianzmatrix von $\vec{X} = \left(X_1,...,X_n\right)$. Es existiere die Inverse C^{-1} von C. Dann gilt für $i \neq j$:*

$$(i)\ \rho\left(X_i,X_j\,|X_1,...,X_n \text{ ohne } X_i,X_j\right) = -\frac{\left(C^{-1}\right)_{ij}}{\sqrt{\left(C^{-1}\right)_{ii}\left(C^{-1}\right)_{jj}}},$$

(Erinnerung: $\left(C^{-1}\right)_{ij}$ *ist das Element der Matrix* C^{-1} *in der* $i-$ *ten Zeile und* $j-$ *ten Spalte bezeichnen - Achtung, das ist* nicht etwa *der Kehrwert von* C_{ij}*!) Weiter haben wir mit den Diagonalelementen der Matrix* C^{-1}*:*

$$(ii)\ 1-R^2\left(X_i\,|X_1...X_n\ \textit{ohne}\ X_i\right)=\frac{1}{\left(C^{-1}\right)_{ii}\cdot\left(C_{ii}\right)}.$$

(Man beachte, dass $C_{ii}=\sigma^2\left(X_i\right)$ *ist und wiederum, dass nicht etwa* $\left(C^{-1}\right)_{ii}$ *der Kehrwert von* C_{ii} *ist - dann käme unsinnigerweise stets* $R^2=0$ *heraus!)*

Bemerkungen: Aussage (i) sagt nichts über $\rho\left(X_i,X_i\,|X_1,...,X_n\text{ ohne }X_i\right)$, das ist auch nicht nötig: Selbstverständlich sind diese Partialkorrelationen trivial und haben den Wert 1 ebenso wie $\rho(X,X)=1$ stets gilt. Mit (ii) ist natürlich auch $|R|$ bestimmt - das Vorzeichen von R wird in den meisten Fällen klar sein. Insbesondere kann man ohne weitere Rechnung angeben, welcher Anteil der Varianz von X_i durch lineare Regression auf die übrigen Variablen erklärt und nicht erklärt ist - die Werte sind jeweils $R^2\left(X_i\,|X_1...X_n\text{ ohne }X_i\right)=1-1/\left(\left(C^{-1}\right)_{ii}\cdot C_{ii}\right)$ bzw. $1-R^2\left(...\right)=1/\left(\left(C^{-1}\right)_{ii}\cdot C_{ii}\right)$.

Modellbeispiel: Seien X,Y,Z alle unabhängig und $(0,1)-$ normalverteilt, sei $X_1=X$, $X_2=X+Y$, $X_3=X+Y+Z$. Wir sind hier ohne weiteres in der Lage (bilineares Rechnen mit Cov!), die Kovarianzmatrix direkt auszurechnen, sie lautet

$$C=Cov((X_1,X_2,X_3))=\begin{pmatrix}1&1&1\\1&2&2\\1&2&3\end{pmatrix}.$$

Damit haben wir, wie Sie hoffentlich oben in gerade diesem beispiel nachgerechnet haben, die Inverse

$$C^{-1}=\begin{pmatrix}1&1&1\\1&2&2\\1&2&3\end{pmatrix}^{-1}=\begin{pmatrix}2&-1&0\\-1&2&-1\\0&-1&1\end{pmatrix}.$$

Um die Partialkorrelationskoeffizienten zu erhalten, haben wir laut Satz zu bilden:

$$\begin{pmatrix}\times&\frac{1}{2}&0\\1/\sqrt{4}=\frac{1}{2}&\times&1/\sqrt{2}\approx0.707\\0&1/\sqrt{2}\approx0.707&\times\end{pmatrix}.$$

(Die Kreuzchen stehen in der Diagonalen, die uns hier nicht interessiert und bei voller Anwendung der Normierung und Vorzeichenumkehr auf nichtssagende Werte -1 in der Hauptdiagonalen hinausliefe.)

Nunmehr berechnen wir aus C^{-1} die drei durch multiple Korrelation mit den jeweils anderen beiden Variablen *nicht erklärten Varianzanteil*, wozu wir die Varianzen benötigen:

$$\begin{aligned}1-R^2\left(X_1|X_2,X_3\right)&=\frac{1}{\left(C^{-1}\right)_{11}\cdot\sigma^2(X_1)}=\frac{1}{2},\\1-R^2\left(X_2|X_1,X_3\right)&=\frac{1}{\left(C^{-1}\right)_{22}\cdot\sigma^2(X_2)}=\frac{1}{4},\\1-R^2\left(X_3|X_1,X_2\right)&=\frac{1}{\left(C^{-1}\right)_{33}\cdot\sigma^2(X_3)}=\frac{1}{3}.\end{aligned}$$

Letzteres Resultat erhält man übrigens auch nach Bildung der Korrelationsmatrix $Corr\left(X_1,X_2,X_3\right)=\left(\rho\left(X_i,X_j\right)\right)_{ij}$, $1\leq i,j\leq3$, worin anstelle der Kovarianzen

die zu den Korrelationskoeffizienten normierten Kovarianzen stehen, analog zur Kovarianzmatrix („jeder mit jedem"). Tun wir dies, so erhalten wir im Beispiel:

$$Corr\,(X_1, X_2, X_3) = \begin{pmatrix} 1 & 1/\sqrt{2} & 1/\sqrt{3} \\ 1/\sqrt{2} & 1 & 2/\sqrt{6} = \sqrt{2}/\sqrt{3} \\ 1/\sqrt{3} & \sqrt{2}/\sqrt{3} & 1 \end{pmatrix}.$$

Nun invertieren wir diese Matrix und erhalten:

$$(Corr\,(X_1, X_2, X_3))^{-1} = \begin{pmatrix} 2 & -\sqrt{2} & 0 \\ -\sqrt{2} & 4 & -\sqrt{2}\sqrt{3} \\ 0 & -\sqrt{2}\sqrt{3} & 3 \end{pmatrix}.$$

Die Reziproken der Elemente auf der Hauptdiagonalen ergeben 1/2, 1/4, 1/3, das sind genau die zuvor anders bestimmten durch multiple Korrelation unerklärten Varianzanteile.

Bemerkung: Man weiß natürlich bei empirisch auftretenden Variablen die benötigten Kovarianzen nicht genau, dann wissen wir sie aber anhand einer (hinreichend großen!) Stichprobe zu schätzen nach dem vorangehenden Kapitel, und so arbeitet man eben mit der empirischen Kovarianzmatrix, d.h. mit ihrem Schätzwert aus lauter geschätzten Kovarianzen, invertiert diese Matrix und holt entsprechende Schätzwerte für die interessierenden Partialkorrelationen und durch multiple Korrelationen erklärten Varianzanteile heraus.

3. Zweite Anwendung: Hauptkomponentenanalyse (PCA)

Diese Anwendung geht ebenfalls wieder von der Kovarianzmatrix $C = Cov\left(\vec{X}\right)$ aus, aber die Zielrichtung und Problemstellung sind andere. Hier fragen wir:

- Normalerweise werden die betrachteten Variablen X_i korreliert sein. Kann man diese Variablen durch neue Variablen Y_i, die nur einfache lineare Kombinationen der alten sind, also z.B. $Y_1 = 2X_1 - 3X_2 + 4X_3$, derart ersetzen, dass der neue Satz von Variablen völlig unkorreliert ist? (Die neuen Variablen liefern dann ebenso viel Information wie die alten, aber sie liefern die Information getrennt und ohne die Redundanzen und Überlappungen, welche bei den alten vorhanden sind. - Natürlich sind im allgemeinen die neuen Variablen nich in nichtlinearer Weise abhängig voneinander, aber doch in wesentlich geringerem Umfang und bei multivariater Normalverteilung von $\vec{X}$ gar überhaupt nicht mehr.
- Kann man, ohne allzu viel Information zu verlieren, gewisse der neuen Variablen Y_i einfach weglassen und damit die Dimension des Raumes, in dem man die Population anordnen und die Lage der einzelnen Individuen anschauen kann, wesentlich reduzieren, gar auf wirklich anschaubare 3 Dimensionen?

Beide Fragen laufen darauf hinaus, wie man seinen Variablensatz ohne allzu viel Mühe wesentlich ökonomischer anlegen kann. Die nachfolgende Hauptkomponentenanalyse wird beide Probleme lösen.

3.1. Zum Grundverständnis der PCA. Wir behandeln zunächst ausschließlich die zuerst genannte Aufgabe: Die Variablen $X_1, ..., X_n$ sind derart linear zu transformieren, dass das entstehende Variablensystem $Y_1, ...Y_n$ keinerlei Korrelationen mehr aufweist (damit wären die ursprünglichen Variablen „dekorreliert", wie man sagt).

Das Einführen der neuen Variablen ist gleichbedeutend damit, dass man ein neues Koordinatensystem einführt. Die Unkorreliertheit der neuen Variablen $\vec{Y}$ bedeutet nun gerade, dass die mit ihnen gebildete Kovarianzmatrix $Cov\left(\vec{Y}\right)$ eine Diagonalmatrix wird: In der Hauptdiagonalen stehen die Varianzen, und alle andern Einträge (die Paarkovarianzen) haben den Wert Null. Wir präparieren heraus, welche Bedingung die Transformation von $\vec{X}$ genau erfüllen muss, damit die Kovarianzmatrix diagonal wird:

SATZ 24. *Sei* $\vec{X} = {}^t(X_1, ..., X_n)$ *gegeben. Sei* $\vec{Y} = {}^t(Y_1, ..., Y_n)$ *folgende lineare Transformation von* $\vec{X}$*:*

$$Y_i = \sum_{j=1}^{n} a_{ij} X_j, \text{ für } i = 1, ..., n.$$

Seien weiter $C_{\vec{X}} = Cov\left(\vec{X}\right)$, $C_{\vec{Y}} = Cov\left(\vec{Y}\right)$. *Dann gilt folgende Aussage:*

$C_{\vec{Y}}$ *ist diagonal genau dann, wenn alle Spaltenvektoren von* ${}^tA$,

$$\vec{a}_i = \begin{pmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{in} \end{pmatrix},$$

Eigenvektoren von $C_{\vec{X}}$ *sind, also* $C_{\vec{X}}\vec{a}_i = \lambda_i \vec{a}_i$ *mit Zahlen* λ_i.

Weiter gilt, dass in diesem Falle, dass also eine dieser gleichwertigen Bedingungen zutrifft, die Eigenwerte λ_i *gerade die Einträge auf der Diagonalen der Diagonalmatrix* $C_{\vec{Y}}$ *sind.*

(Erläuterung: Die Matrix $A = (a_{ij})_{ij}$ macht also jeweils aus dem Vektorwert von $\vec{X}$ den Vektorwert von $\vec{Y}$, oder kurz $\vec{Y} = A\vec{X}$. Wir müssen hier festlegen, dass $\vec{X}$, $\vec{Y}$ Spaltenvektoren sind, daher die Transpositionszeichen an den Zeilenvektoren oben.)

Beweis: Mit der Bilinearität der Kovarianz haben wir:

$$\begin{aligned} Cov\left(\vec{Y}\right) &= \left(Cov\left(Y_{ik}\right)\right)_{ik} = \left(Cov\left(\sum_{j=1}^{n} a_{ij}X_j, \sum_{j=1}^{n} a_{kj}X_j\right)\right)_{ik} \\ &= \left(\sum_{j,s} a_{ij}a_{ks}Cov\left(X_j, X_s\right)\right)_{ik} = \left(\sum_{j,s} a_{ij}Cov\left(X_j, X_s\right)a_{ks}\right)_{ik} \\ &= ACov\left(\vec{X}\right)^t A. \end{aligned}$$

Man beachte nur, dass bei a_{ks} die Indexreihenfolge falsch für das Matrizenprodukt ist, was wir genau mit Setzung der Transponierten rechts richtig gestellt haben. Nach dem Satz aus der Linearen Algebra wissen wir nun, dass $Cov\left(\vec{Y}\right)$ genau dann die Diagonalmatrix mit $\lambda_1, ...\lambda_n$ ist, wenn die Spaltenvektoren von ${}^tA$ (das entspricht B aus dem Satz!) die zugehörigen Eigenvektoren von $Cov\left(\vec{X}\right)$ sind. Damit ist der Satz bewiesen.

Ferner wissen wir nach dem genannten Resultat, dass es eine sogar orthogonale Matrix B mit ${}^tBCov\left(\vec{X}\right)B$ = Diagonalmatrix stets gibt. Wir setzen also als

unsere Transformationsmatrix $A := {}^tB$ und kennen damit unsere neuen Variablen $Y_i = \sum_j a_{ij}X_j$. Ebenfalls können wir mit der inversen Matrix ${}^tA$ aus den $Y-$ Werten wieder die $\vec{X}-$ Werte berechnen mit $\vec{X} = {}^tA\vec{Y}$ oder ausführlicher $X_i = \sum_j a_{ji}Y_j$. (Beachte die Reihenfolge der Indizes!) Damit haben wir den

SATZ 25. *Es existiert stets eine orthogonale Transformation $A = (a_{ij})_{ij}$, mit der man gegebene Variablen $\vec{X} = (X_1, ..., X_n)$ durch Definition neuer Variablen $Y_i = \sum_j a_{ij}X_j$ dekorrelieren kann, dass also die neuen Variablen paarweise linear unabhängig sind, ihre Kovarianzen für Paare verschiedener davon also verschwinden und damit die Korrelationskoeffizienten. (Die Transformation kann man auch als Drehung des Koordinatensystems auffassen.)*

Damit ist das erste Problem gelöst. Nun wird man fragen, wie man denn die Eigenwerte und Eigenvektoren zu einer Kovarianzmatrix berechnen kann. Das ist ziemlich aufwendig, aber doch im Prinzip nicht schwierig, insbesondere können Standardprogramme das numerisch befriedigend ausführen. Somit hat man ein effizientes Verfahren, die Sache durchzuführen.

Wir kommen nunmehr zum zweiten Problem, dem einer möglichen Dimensionsreduktion, also Reduktion der Anzahl der Variablen. Dies Problem löst sich im Rahmen der Diagonalisierung ganz einfach. Dazu müssen wir nur die Eigenwerte λ_i interpretieren: Es gilt $\lambda_i = \sigma^2(Y_i)$, da die Diagonalmatrix der λ_i mit den Hauptdiagonalen gleich $Cov\left(\vec{Y}\right)$ ist, somit in der Hauptdiagonalen die Varianzen stehen. Daher ist $\sum_i \lambda_i^2 = \sum_i \sigma^2(Y_i)$ als Gesamtvarianz in dem Sinne zu interpretieren, dass dieser Wert den mittleren quadratischen Abstand der Punkte $\vec{Y}(\omega)$ für alle $\omega \in \Omega$ ergibt. (Denn nach dem oben eingeführten Abstandsbegriff ist der quadrierte Abstand zweier Punkte gerade die Quadratsumme der Koordinatendifferenzen.) Ein Teil der Variablen, sagen wir also $Y_1, ..., Y_k$ mit $k < n$, „erklärt" also den Anteil $\sum_{i\leq k} \lambda_i^2 / \sum_{i\leq n} \lambda_i^2$ von dieser Gesamtvarianz, und man kann sich sicher dazu verstehen, von 80 Variablen 75 wegzulassen, wenn die übrigbleibenden fünf Variablen bereits 95% der Varianz erklären. Eine Methode, den Schnitt zu bestimmen, besteht darin, die Funktion $i \mapsto \lambda_i$ (oder auch λ_i^2) aufzutragen, nachdem man die Variablen Y_i so angeordnet hat, dass die Eigenwerte λ_i eine fallende Folge bilden. Den Schnitt setzt man dann gern an einer Stelle, wo diese Funktion einen deutlichen Knick hat, wenn man einen solchen findet. Beispiel: Für eine empirische Stichprobe vom Umfang 1000 für die Variablen $X_1 = 9Z_1$, Z_1 standardnormalverteilt, $X_2 = X_1 + 9Z_2$, Z_2 standard-normalverteilt und unabhängig von Z_1, $X_3 = X_1 + X_2 + 8Z_3$ usw. bis $X_{10} = X_1 + ... + X_9 + Z_{10}$ finden wir eine sehr unübersichtliche Kovarianzmatrix. Nach Diagonalisierung erhalten wir für die Transformierte $\vec{Y}$ die diagonale Kovarianzmatrix mit den folgenden Einträgen in der Hauptdiagonalen:

$$2529.3, 1574, 542, 319, 227, 144, 98, 48, 22, 5.$$

Dann ist klar: Die ersten beiden der neuen Variablen, also Y_1, Y_2, erklären bereits den Anteil

$$\frac{2529.3 + 1574}{2529.3 + 1574 + 542 + 319 + 227 + 144 + 98 + 48 + 22 + 5} = 0.74493$$

der Gesamtvarianz, also bereits fast 75%. Mit der dritten zusammen ergibt sich schon etwa 84%, erst mit der sechsten erreicht man etwa 95%. Die beschriebene Grafik des Abfallens der Eigenwerte sieht in unserm Falle so aus:

Das „Knie" bei 2 ist deutlich, und man würde nur die ersten beiden Variablen Y_1, Y_2 nehmen, allenfalls noch die dritte. Nun zeigen wir noch den Punkteschwarm für diese beiden Variablen und sehen anschaulich, dass wirklich dekorreliert ist (horizontale Achse: Y_1- Werte, vertikale Achse: Y_2- Werte) - das sieht genau so aus wie ein Zufallsbild unabhängiger normalverteilter Variablen, und das ist es auch:

Ganz bewusst haben wir hier gleiche Einheiten auf den Achsen verwandt, um herauszubringen, dass Y_1 bereits deutlich größere Varianz hat als Y_2. Das fällt dann in den weiteren Dimensionen weiter drastisch. Man beachte: All dies liefert der Computer auf Knopfdruck - von Hand oder auch mit einem Taschenrechner hätte man schier endlos dafür zu rechnen! (Etwa eine halbe Million Operationen mit langen Dezimalzahlen!) Natürlich kann man auch die Transformationsmatrix ablesen, in unserm Falle ist

$$Y_1 = 0.0966X_1 + 0.18680X_2 + 0.2552X_3 + ...0.3752X_9 + 0.3754X_{10}.$$

Das können wir durchaus ein wenig intuitiv verstehen: Die letzten der $X-$ Variablen haben größte Varianz und werden stärker gewichtet. Dagegen sehen die Koeffizienten zu Y_2 so aus, dass die ersten $X-$ Variablen starke negative Gewichte erhalten, die letzten ansteigende positive.

4. Dritte Anwendung: Multidimensionale Skalierung (MDS)

Vorbemerkung: Der englische Name lautet „multidimensional scaling". Die zu analysierende Situation sieht so aus: Man hat n Objekte, die auf irgendeine Art paarweise als „ähnlicher" oder „unähnlicher" charakterisiert sind. Es ist dabei ganz gleichgültig, wie hoch die subjektiven oder objektiven Anteile dieser Charakterisierung aussehen. Man kann die Ähnlichkeit von Individuen zum Beispiel auf einer subjektiven Skala bewerten lassen, man könnte auch die Ähnlichkeit zweier Aufgaben durch den Korrelationskoeffizienten der erreichten Punktezahlen messen. Oder es könnten objektive Kriterien einen Ähnlichkeitsbegriff festlegen - denken Sie an ein Abzählen von Übereinstimmungen in einer festgelegten Reihe von objektiven Merkmalen, oder auch an quantitative Verfeinerungen. Analog kann man statt von Ähnlichkeiten von Abständen oder Distanzen reden. Skaliert man Ähnlichkeiten so, dass ihre Werte zwischen 0 und 1 liegen, so kann man mit $d(P,Q) = 1-\ddot{a}(P,Q)$ (mit „P", „Q" werden hier die beliebigen Objekte des betrachteten Bereichs bezeichnet) aus dem Ähnlichkeitsbegriff $\ddot{a}$ einen Abstandsbegriff d machen. Ist $\ddot{a}$ symmetrisch und gilt $\ddot{a}(P,P) = 1$, so gelingt es sogar nach kleiner Variation, unsere Axiome für einen Abstandsbegriff zu erfüllen.

Hier ist die *Idee des Multiddensional Scaling:* Man fasst die zu vergleichenden Objekte als Punkte eines Vektorraums mit dem euklidischen (oder einem andern) Abstandsbegriff auf. Diese Punkte identifiziert man wiederum mit ihren Koordinatentupeln. Dabei ist es natürlich ganz gleichgültig, durch was für einen $n-Tupel-Vektor$ ein Objekt dargestellt wird - es ist ja auch im Prinzip egal, was für ein Koordinatensystem man wählt. *Entscheidend ist nur, die Vektoren so zu wählen, dass ihre Abstände paarweise* (wir werden nur den euklidischen betrachten) möglichst genau den wie oben beschrieben irgendwie (gegebenenfalls über Ähnlichkeitswerte) ermittelten Abständen zwischen den ursprünglichen Objekten entsprechen. Nun ist es eine Trivialität, dass man diese Abstände ganz genau darstellen kann, indem man für unsere n Objekte einen $(n-1)-$ dimensionalen Darstellungsraum wählt: Zwei Punkte kann man offenbar auf einer Geraden in beliebigen Abstand setzen (Dimension 1), einen dritten Punkt dazu in einer Ebene so unterbringen, dass er in vorgeschriebenen Abständen zu den ersten beiden steht, usw. Lediglich müssen die Abstandsaxiome erfüllt sein. Für n bekommt man allgemein ein lösbares Gleichungssystem zu sehen. Aber die leitende Grundidee für alle multivariaten Methoden ist eine Kopplung aus Veranschaulichung und Vereinfachung. Und so lautet die entscheidende Frage bei MDS: In welcher *kleinsten* Dimension kann man noch mit einer kleinen, vernachlässigbaren Ungenauigkeit die vorgegebenen Abstände zwischen den ursprünglichen n Objekten repräsentieren? Dazu geht man so vor: Man wählt eine Dimension $dd < n-1$ und wählt Vektoren $\overrightarrow{x}_1, ..., \overrightarrow{x}_n$ in $\mathbb{R}^{dd}$ (Erinnerung: $\overrightarrow{x}_i$ entspricht P_i, dem Objekt Nummer i, $1 \leq i \leq n$) *derart, dass die absoluten (oder auch quadrierten) Differenzen zwischen* $d_{ij} = d(P_i, P_j)$ und $\delta_{ij} = |\overrightarrow{x}_i - \overrightarrow{x}_i|$ *im Mittel minimal werden.* (Man beachte: Die Distanzmatrix $D = (d_{ij})_{1\leq i,j\leq n}$ beschreibt unsere gesamte Ausgangslage!) Es wird also ein Minimum der Funktion (hier für quadrierte Differenzen)

$$f(\overrightarrow{x}_1, ..., \overrightarrow{x}_n) = \sum_{i,j=1}^{n} (d_{ij} - \delta_{ij})^2$$

gesucht. (Das ist also eine Funktion von $dd\cdot n$ unabhängigen Variablen, oder auch von einer Matrix.) Gesucht ist wie immer die Stelle im Urbild von f, also die Matrix,

wofür die Sache minimal wird. Glücklicherweise kann das von einem Computerprogramm leicht erledigt werden. Man sucht nun einfach, für welche Dimension dd beim Minimum ein noch sehr kleiner Funktionswert unter f herauskommt, so dass bei $dd-1$ aber der Wert für das dortige Minimum deutlich größer und unbefriedigend ist. (Das ist natürlich nicht immer ganz eindeutig, aber doch in der Regel recht klar.) Diese Dimension dd wählt man für die MDS-Darstellung. Nunmehr kann man eventuell in einem Raum recht kleiner Dimension dd (wie 2 oder 3) eine Reihe von Punkten anschauen, welche die Objekte vertreten, und die gesamte Ähnlichkeitsstruktur überschauen - Cluster, geometrische Anordnung usw. Und davon hätte man nichts gesehen bei Stieren auf die Ähnlichkeitsmatrix oder Distanzmatrix für die ursprünglichen Objekte, zumal bei einer großen Anzahl von ihnen. Zum Verständnis und zum Beleg dafür, dass dabei etwas ganz Ansehnliches resultieren *kann*, folgende Illustration:

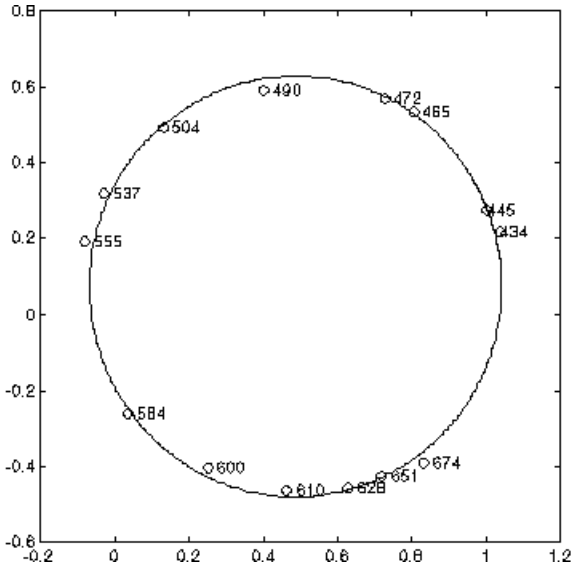
Beispiel: Die Ähnlichkeits- oder Distanzstruktur der subjektiven Wahrnehmung von Farben bei Lichtern:

Einer Reihe von Individuen wurden Lichter (gleicher Helligkeit und Sattheit) verschiedener Wellenlängen gezeigt. (Versuch von Ekman 1954 - tatsächlich wurden die Versuchspersonen gebeten, auf einer Skala von 0 bis 4 die Ähnlichkeit zu bewerten, dann wurde über die Versuchspersonen gemittelt und auf [0,1] skaliert. Es wurden farbige Plättchen verwendet, also eigentlich nicht Lichter, die es lieber hätten sein sollen, sondern Oberflächenfarben, jedoch ziemlich „reine", künstliche (Munsell-Chips). Wir geben hier die aus $d(P,Q) = 1 - \ddot{a}(P,Q)$ resultierende Distanzmatrix - die Objekte (Lichter) werden mit den Wellenlängen in Nanometern bezeichnet:

	434	445	465	472	490	504	537	555	584	600	610	628	651	674
434	0	0.14	0.58	0.58	0.82	0.94	0.93	0.96	0.98	0.93	0.91	0.88	0.87	0.84
445	0	0	0.5	0.56	0.78	0.91	0.93	0.93	0.98	0.96	0.93	0.89	0.87	0.86
465			0	0.19	0.53	0.83	0.9	0.92	0.98	0.99	0.98	0.99	0.95	0.97
472				0	0.46	0.75	0.9	0.91	0.98	0.99	1	0.99	0.98	0.96
490					0	0.39	0.69	0.74	0.93	0.98	0.98	0.99	0.98	1
504						0	0.38	0.65	0.86	0.92	0.98	0.98	0.98	0.99
537							0	0.27	0.78	0.86	0.95	0.98	0.98	1
555								0	0.67	0.81	0.96	0.97	0.98	0.98
584									0	0.42	0.63	0.73	0.8	0.77
600										0	0.26	0.5	0.59	0.72
610											0	0.24	0.38	0.45
628												0	0.15	0.32
651													0	0.24
674														0

Beachten sie, dass die nicht eingefüllten Elemente einfach symmetrisch zu den andern liegen, es ist ja $d_{ij} = d(P_i, P_j) = d(P_j, P_i) = d_{ji}$. Im hier vorliegenden Beispiel ist das auch von der Datengewinnung her sofort der Fall, da paarweise verglichen wurde. Bei unsymmetrischem Versuchsaufbau könnte man $d_{ij} \neq d_{ji}$ erhalten und dann $d'_{ij} = \frac{1}{2}(d_{ij} + d_{ji})$ bilden. Hier sehen Sie nun, was durch MDS aus dieser wenig übersichtlichen Matrix wird (ich habe das selber mit dem Computer berechnet und Varianten hinsichtlich des Abstandsbegriffes und Optimierungskriteriums betrachtet. Die Ergebnisse waren alle außerordentlich ähnlich und sahen so aus:

Dimension $dd = 2$ ist angemessen, und graphisch hat das *errechnete* Resultat folgende Gestalt - an den Punkten steht mit der Wellenlänge das zugehörige Objekt bezeichnet, zusätzlich ist ein exakter Kreis eingezeichnet:



Die Skalenwerte bedeuten nichts, außer in folgender Hinsicht: Die Skaleneinteilungen sind völlig gleich, so dass die sichtbaren Abstände den euklidisch zu errechnenden Abständen genau entsprechen, die ihrerseits wiederum die Abstände aus der vorgegebenen Distanzmatrix sehr genau wiedergeben.

Was wir hier sehen, ist der berühmte Farbkreis, auf den Newton schon mit ganz anderen und viel intuitiveren Überlegungen kam: Rot und Violett, die Enden des sichtbaren Spektrums, sind Antipoden hinsichtlich der Wellenlängen, biegen sich aber für uns zusammen, und die einfachste Weise, das zu realisieren, ist tatsächlich das Zusammenbiegen zum Kreis. Mit MDS kam man quantitativ auf einen Kreis, durch Anpassung an reale Daten. Tatsächlich könnte man auch sehen, dass die Abstände auf dem Kreis nicht so ganz den Wellenlängenabständen entsprechen. Übrigens beobachtet man bei Wahl höherer Darstellungsdimension $dd > 2$ keine drastischen Sprünge mehr in der erreichten Genauigkeit der Abstandswiedergabe, dafür aber eine deutliche Verunklärung des Bildes. Für $dd = 1$ dagegen erzielt man überhaupt keine angemessene Darstellung der Abstände, wie man leicht einsieht.

Selbstverständlich gibt es eine Fülle weiterer Anwendungsbeispiele, in denen man auf ordentliche Strukturen stößt, auch etwa aus dem Bereich der Sozialpsychologie.

Index