

LEHRBUCH

Matthias Plaue

# Data Science

Grundlagen, Statistik und  
maschinelles Lernen



Springer Spektrum

---

## Data Science

---

Matthias Plaue

# Data Science

Grundlagen, Statistik und  
maschinelles Lernen

Matthias Plaue  
MAPEGY GmbH  
Berlin, Deutschland

ISBN 978-3-662-63488-2      ISBN 978-3-662-63489-9 (eBook)  
<https://doi.org/10.1007/978-3-662-63489-9>

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

© Der/die Herausgeber bzw. der/die Autor(en), exklusiv lizenziert durch Springer-Verlag GmbH, DE, ein Teil von Springer Nature 2021

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von allgemein beschreibenden Bezeichnungen, Marken, Unternehmensnamen etc. in diesem Werk bedeutet nicht, dass diese frei durch jedermann benutzt werden dürfen. Die Berechtigung zur Benutzung unterliegt, auch ohne gesonderten Hinweis hierzu, den Regeln des Markenrechts. Die Rechte des jeweiligen Zeicheninhabers sind zu beachten.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Planung/Lektorat: Andreas Ruedinger

Springer Spektrum ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature.

Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Für Katja.

---

## Vorwort

Dieser Band wurde in großen Teilen im Jahr 2020 verfasst – einem Jahr, das den Beginn der COVID-19-Pandemie markiert. In den Medien sind in jener Zeit Datenanalysen allgegenwärtig: Rot gefärbte Landkarten zeigen besonders betroffene Gebiete auf, Graphen von Zeitreihen die Wellen des Infektionsgeschehens. Prognosen auf Grundlage epidemiologischer Modelle werden erstellt, der „R-Wert“ findet Eingang im öffentlichen Sprachgebrauch. Diese Analysen und Prognosen sind teils Grundlage politischer Entscheidungen von großer Tragweite – zu Recht, spiegeln sie doch numerisch gesichert die Faktenlage wider. Dennoch ermöglichen sie aufgrund ihrer statistischen Natur ein Schließen und Handeln nur unter teils großer Unsicherheit.

Zugleich wird das Weltgeschehen seit dem Ende des letzten Jahrhunderts durch den Aufbruch in das Informationszeitalter geprägt. Bedeutende Fortschritte im Bereich der künstlichen Intelligenz ebnen deren Einsatz im Alltag: Suchmaschinen durchforsten das World Wide Web nach relevanten Inhalten und versuchen dabei, die Bedeutung der Nutzereingabe zu erfassen. Algorithmen sortieren automatisch Urlaubsfotos nach Bildinhalten und „smarte“ Haushaltsgeräte werden zunehmend gewöhnlich.

Ob Statistik, „Big Data“ oder maschinelles Lernen: Es kann kaum infrage gestellt werden, dass die Datenwissenschaften auch in den kommenden Jahrzehnten eine bedeutende Rolle spielen werden. Dieses Buch soll künftige Studierende der Data Science dabei unterstützen, einen Einstieg in dieses spannende und zukunftsweisende Gebiet zu finden.

Meiner Meinung nach fußt ein erfolgreiches Studium der Data Science auf zwei wesentlichen Säulen. Zum einen auf der selbstständigen und praktischen Beschäftigung mit der Verarbeitung von Daten, dem „Spielen“ mit Daten, der Programmierung eigener Modelle und dem Durchführen eigener Parameterstudien. Im Web stehen zu diesem Zweck zahlreiche Tutorials, Codebeispiele und Datensätze zur Verfügung. Die andere wichtige Säule besteht in einer hinreichend umfassenden Methodenkenntnis und einem tiefen Verständnis der Prinzipien und Ideen, die den Methoden der Data Science zugrundeliegen. Dieses Buch

## VIII Vorwort

soll in erster Linie dabei helfen, zu einer solchen Kenntnis bzw. einem solchen Verständnis zu gelangen.

Besonderen Dank möchte ich meinem Doktorvater und früheren Weggefährten Mike Scherfner aussprechen. Des Weiteren möchte ich all jenen danken, die mir den Weg in die Datenanalyse bereiteten, bzw. die mich während meiner Laufbahn als Data Scientist begleitet haben oder noch begleiten: Fred Hamprecht, Günter Bärwolff, Hartmut Schwandt und Peter Walde. Alex Dirmeier möchte ich herzlich für das Gegenlesen und seine klugen und wertvollen Anmerkungen danken. Schließlich möchte ich mich bei jenen Lektoren/Lektorinnen und Mitarbeitern/Mitarbeiterinnen des Springer-Verlages kenntlich zeigen, die dieses Projekt unterstützt und ermöglicht haben – mein besonderer Dank gilt dabei Andreas Rüdinger.

Berlin, 29. August 2021

*Matthias Plaue*

---

# Inhaltsverzeichnis

<b>Einführung</b> .....	1
-------------------------	---

---

## Teil I Grundlagen

---

<b>1 Elemente der Datenorganisation</b> .....	11
1.1 Konzeptionelle Datenmodellierung .....	12
1.1.1 Entity-Relationship-Modell .....	12
1.2 Logische Datenmodellierung .....	14
1.2.1 Relationales Datenmodell .....	14
1.2.2 Graphbasierte Datenmodelle .....	17
1.2.3 Hierarchische Datenmodelle .....	20
1.3 Datenqualität .....	22
1.3.1 Datenqualitätsmerkmale .....	23
1.4 Datenbereinigung .....	24
1.4.1 Validierung .....	24
1.4.2 Normierung .....	25
1.4.3 Imputation .....	26
1.4.4 Augmentation .....	28
1.4.5 Deduplikation .....	28
Quellen .....	35
<b>2 Deskriptive Statistik</b> .....	37
2.1 Stichprobe und Merkmale .....	38
2.2 Diagramme .....	40
2.2.1 Säulendiagramme und Histogramme .....	40
2.2.2 Streudiagramme .....	42
2.2.3 Weitere Diagramme .....	44
2.3 Lageparameter .....	45
2.3.1 Arithmetisches Mittel und empirischer Median .....	48
2.3.2 Quantile .....	50
2.3.3 Geometrisches und harmonisches Mittel .....	51

2.4	Streuungsparameter . . . . .	53
2.4.1	Abweichung von Mittelwert oder Median . . . . .	53
2.4.2	Shannon-Entropie . . . . .	54
2.5	Assoziationsparameter . . . . .	56
2.5.1	Empirische Kovarianz und Korrelation . . . . .	57
2.5.2	Rangkorrelationskoeffizienten . . . . .	59
2.5.3	Transinformation und Jaccard-Koeffizient . . . . .	62
	Quellen . . . . .	66

---

**Teil II Stochastik**

<b>3</b>	<b>Wahrscheinlichkeitstheorie</b> . . . . .	69
3.1	Wahrscheinlichkeitsmaße . . . . .	70
3.1.1	Bedingte Wahrscheinlichkeit . . . . .	74
3.1.2	Der Satz von Bayes . . . . .	78
3.2	Zufallsvariablen . . . . .	81
3.2.1	Diskrete und stetige Zufallsvariablen . . . . .	82
3.2.2	Massen- und Dichtefunktionen . . . . .	85
3.2.3	Transformation von Zufallsvariablen . . . . .	88
3.3	Gemeinsame Verteilung von Zufallsvariablen . . . . .	91
3.3.1	Gemeinsame Verteilungs-, Masse- und Dichtefunktionen . . . . .	91
3.3.2	Bedingte Masse- und Dichtefunktionen . . . . .	93
3.3.3	Unabhängige Zufallsvariablen . . . . .	94
3.4	Kennzahlen von Zufallsvariablen . . . . .	96
3.4.1	Median, Erwartungswert und Varianz . . . . .	96
3.4.2	Kovarianz und Korrelation . . . . .	101
3.4.3	Die Tschebyscheff'sche Ungleichung . . . . .	104
3.5	Summen und Produkte von Zufallsvariablen . . . . .	106
3.5.1	Chi-Quadrat- und Student'sche t-Verteilung . . . . .	110
	Quellen . . . . .	113
<b>4</b>	<b>Inferenzstatistik</b> . . . . .	115
4.1	Statistische Modelle . . . . .	116
4.1.1	Modelle diskreter Zufallsvariablen . . . . .	116
4.1.2	Modelle stetiger Zufallsvariablen . . . . .	120
4.2	Gesetze der großen Zahlen . . . . .	123
4.2.1	Bernoulli'sches Gesetz der großen Zahlen . . . . .	126
4.2.2	Tschebyscheff'sches Gesetz der großen Zahlen . . . . .	128
4.2.3	Varianzschätzung und Bessel-Korrektur . . . . .	130
4.2.4	Zentraler Grenzwertsatz von Lindeberg-Lévy . . . . .	132
4.3	Statistische Schätz- und Testverfahren . . . . .	133
4.3.1	Intervallschätzung . . . . .	133
4.3.2	Gauß-Test . . . . .	137
4.3.3	Student'sche Vertrauensintervalle . . . . .	139
4.3.4	Effektstärke . . . . .	142
4.4	Parameter- und Dichteschätzung . . . . .	143

4.4.1	Maximum-Likelihood-Schätzung . . . . .	145
4.4.2	Bayes'sche Parameterschätzung . . . . .	150
4.4.3	Kerndichteschätzung . . . . .	154
4.5	Regressionsanalyse . . . . .	156
4.5.1	Einfache lineare Regression . . . . .	156
4.5.2	Theil-Sen-Verfahren . . . . .	161
4.5.3	Einfache logistische Regression . . . . .	162
	Quellen . . . . .	164
<b>5</b>	<b>Multivariate Statistik . . . . .</b>	<b>165</b>
5.1	Datenmatrizen . . . . .	165
5.2	Abstands- und Ähnlichkeitsmaße . . . . .	167
5.2.1	Metrische Abstands- und Ähnlichkeitsmaße . . . . .	167
5.2.2	Kategoriale und binäre Abstands- und Ähnlichkeitsmaße	170
5.2.3	Abstands- und Ähnlichkeitsmatrizen . . . . .	172
5.3	Multivariate Lage- und Streuungsparameter . . . . .	175
5.3.1	Geometrischer Schwerpunkt und Median, Medoid . . . . .	176
5.3.2	Empirische Kovarianz- und Korrelationsmatrix . . . . .	178
5.4	Zufallsvektoren und -matrizen . . . . .	179
5.4.1	Erwartungswertvektor und Kovarianzmatrix . . . . .	179
5.4.2	Multivariate Normalverteilung . . . . .	181
5.4.3	Multinomialverteilung . . . . .	184
	Quellen . . . . .	186

### Teil III Maschinelles Lernen

<b>6</b>	<b>Überwachtes maschinelles Lernen . . . . .</b>	<b>189</b>
6.1	Elemente des überwachten Lernens . . . . .	191
6.1.1	Verlustfunktionen und empirisches Risiko . . . . .	194
6.1.2	Überanpassung und Unteranpassung . . . . .	196
6.1.3	Training, Modellauswahl und Test . . . . .	200
6.1.4	Numerische Optimierung . . . . .	204
6.2	Regressionsverfahren . . . . .	208
6.2.1	Lineare Regression . . . . .	209
6.2.2	Gauß-Prozess-Regression . . . . .	215
6.3	Klassifikationsverfahren . . . . .	218
6.3.1	Logistische Regression . . . . .	218
6.3.2	Nächste-Nachbarn-Klassifikation . . . . .	223
6.3.3	Bayes'sche Klassifikationsverfahren . . . . .	227
6.4	Künstliche neuronale Netzwerke . . . . .	232
6.4.1	Regression und Klassifikation mittels neuronaler Netzwerke . . . . .	235
6.4.2	Training neuronaler Netzwerke durch Fehlerrückführung . . . . .	239
6.4.3	Convolutional Neural Networks . . . . .	245
	Quellen . . . . .	251

<b>7 Unüberwachtes maschinelles Lernen</b>	255
7.1 Elemente des unüberwachten Lernens	255
7.1.1 Intrinsische Dimension von Daten	256
7.1.2 Topologische Merkmale von Daten	257
7.2 Dimensionsreduktion	259
7.2.1 Hauptkomponentenanalyse	261
7.2.2 Autoencoder	264
7.2.3 Multidimensionale Skalierung	265
7.2.4 T-distributed Stochastic Neighbor Embedding (t-SNE)	269
7.3 Clusteranalyse	272
7.3.1 K-Means-Verfahren	272
7.3.2 Hierarchische Clusteranalyse	276
Quellen	281
<b>8 Maschinelles Lernen in der Anwendung</b>	283
8.1 Anwendungsbeispiele für überwachtes Lernen	283
8.1.1 MNIST: Handschrifterkennung	284
8.1.2 CIFAR-10: Objekterkennung	286
8.1.3 Large Movie Review Dataset: Sentimentanalyse	288
8.2 Anwendungsbeispiele für unüberwachtes Lernen	293
8.2.1 Textanalyse: Themenmodellierung	293
8.2.2 Netzwerkanalyse: Gemeinschaftsstrukturen	294
Quellen	299
<b>Ergänzende Literatur</b>	303
<b>Sachverzeichnis</b>	307

---

## Symbolverzeichnis

$a := b$	Zuweisung, Definition; Objekt $a$ ist erklärt durch Objekt oder Formel $b$
$\mathbb{Z}$	Menge der ganzen Zahlen $\{0, -1, 1, -2, 2, \dots\}$
$\mathbb{N}$	Menge der natürlichen Zahlen $\{0, 1, 2, \dots\}$
$\{0, 1, \dots, D\}$	Menge der ersten $D + 1$ natürlichen Zahlen
$\mathbb{R}$	Menge der reellen Zahlen
$(x_1, \dots, x_D)$	ein $D$ -Tupel von Elementen (z. B. reeller Zahlen)
$\mathbb{R}^D$	Menge aller $D$ -Tupel reeller Zahlen
$]a, b[$ , $[a, b]$ , $]a, b]$	offenes, geschlossenes, halboffenes Intervall mit Intervallgrenzen $a, b \in \mathbb{R}$
$]-\infty, b]$ , $[a, \infty[$	uneigentliche Intervalle
$x \in A$	$x$ ist ein Element der Menge $A$
$\emptyset$	leere Menge
$B \subseteq A$	$B$ ist eine Teilmenge von $A$
$B \subset A$	$B$ ist eine echte Teilmenge von $A$
$\bigcup_{i \in I} A_i$	Vereinigung über endliche oder abzählbare Indexmenge $I$ , z. B. $\bigcup_{i \in \{2,3,5\}} A_i = A_2 \cup A_3 \cup A_5$
$f: A \rightarrow B$ , $x \mapsto f(x)$	$f$ ist eine Abbildung mit Definitionsbereich $A$ , Wertebereich $B$ und Abbildungsvorschrift $x \mapsto f(x)$
$g \circ f$	Komposition/Verknüpfung von Abbildungen: $(g \circ f)(x) = g(f(x))$

$f_\alpha(\cdot   \theta; \beta)$	eine Abbildung $x \mapsto f_\alpha(x   \theta; \beta)$ aus einer Familie von Abbildungen, bestimmt durch diverse weitere Parameter $\alpha, \beta, \theta$
$\{x \in A   x \text{ erfüllt } \mathcal{B}\}$	Menge aller Elemente in $A$ , welche die Bedingung $\mathcal{B}$ erfüllen
$ A $	Anzahl der Elemente einer endlichen Menge $A$
$ B $	Fläche/Volumen/Inhalt eines Bereichs $B \subseteq \mathbb{R}^D$
$ x $	Absolutbetrag einer reellen Zahl $x$
$\operatorname{sgn}(x)$	Vorzeichenfunktion: $\operatorname{sgn}(x) = -1$ , falls $x < 0$ usw.
$[x], \lceil x \rceil$	auf die nächste ganze Zahl abgerundeter/aufgerundeter Wert von $x \in \mathbb{R}$
$x \approx y$	die Größe $x$ ist näherungsweise gleich $y$ (in einem geeigneten Sinn)
$x \gg y$	die Größe $x$ ist sehr viel größer als $y$ (in einem geeigneten Sinn)
$x \cong y$	$x$ und $y$ sind in geeignetem Sinn strukturell identisch
$\sum_{k=1}^K x_k$	Summe der Form $x_1 + x_2 + \dots + x_K$
$\sum_{k \in I} x_k$	Summe über endliche oder abzählbare Indexmenge $I$ , z. B. $\sum_{k \in \{2,3,5\}} x_k = x_2 + x_3 + x_5$
$\prod_{k=1}^K x_k$	Produkt der Form $x_1 \cdot x_2 \cdots x_K$
$\langle x, y \rangle$	Skalarprodukt zweier Vektoren $x$ und $y$
$\ x\ $	Länge/Norm eines Vektors $x$
$f \propto g$	die Vektoren/Funktionen $f \neq 0$ und $g \neq 0$ sind kolinear/linear abhängig, d. h., es gibt einen Skalar/eine Konstante $\lambda$ , sodass $f = \lambda g$
$A^T$	Transponierte einer Matrix $A$ ; $A$ nach Vertauschen von Zeilen und Spalten
$\det(A)$	Determinante einer quadratischen Matrix $A$
$\sphericalangle(x, y)$	Winkel zwischen Vektoren $x$ und $y$
$\operatorname{diag}(d_1, \dots, d_K)$	eine Diagonalmatrix; quadratische Matrix mit den Einträgen $d_1, \dots, d_K$ auf der Diagonalen, alle übrigen Einträge sind null
$\min A, \max A$	Minimum/Maximum (kleinstes/größtes Element) einer (z. B. endlichen) Menge $A \subset \mathbb{R}$
$\inf A, \sup A$	Infimum/Supremum (größte untere/kleinste obere Schranke) einer Menge $A \subseteq \mathbb{R}$

## XVI Symbolverzeichnis

$\inf_{x \in A} f(x)$	Infimum einer Funktion $f$ über der Menge $A$
$\lim_{n \rightarrow \infty} a_n$	Grenzwert einer Folge $(a_n)_{n \in \mathbb{N}}$
$\lim_{u \rightarrow \infty} f(u)$	Grenzwert einer Funktion $f$
$\lim_{u \searrow u_0} f(u)$	rechtsseitiger Grenzwert einer Funktion $f$
$\frac{\partial}{\partial x_2} f(x_1, x_2)$	partielle Ableitung einer (stetig differenzierbaren) Funktion $f$ ; alternative Notation: $\partial_2 f(x_1, x_2)$
$\text{grad } f(x)$	Gradient von $f$ ; Spaltenvektor der ersten Ableitungen $(\partial_i f)$
$\text{Hess } f(x)$	Hesse-Matrix einer (zweimal stetig differenzierbaren) Funktion $f$ : Matrix der zweiten Ableitungen $(\partial_i \partial_j f)$
$Df(x)$	Jacobi-Matrix einer (stetig differenzierbaren) – i. Allg. vektorwertigen – Abbildung $f$ : Matrix der ersten Ableitungen $(\partial_i f_j)$
$\nabla_y f(x, y_1, y_2)$	Nabla-Operator; Gradient bzgl. Teilkomponenten: $\nabla_y f = (\partial_2 f, \partial_3 f)^T$
$D_y f(x, y_1, y_2)$	Jacobi-Matrix bzgl. Teilkomponenten
$\int_a^b f(x) dx$	Integral einer (integrierbaren) Funktion in den Grenzen $a, b \in \mathbb{R}$
$\int_{-\infty}^b f(x) dx$	uneigentliches Integral einer Funktion

---

# Abbildungsverzeichnis

0.1	Liniendiagramm eines Temperaturverlaufs .....	2
1.1	Entity-Relationship-Diagramm .....	13
1.2	Knoten-Kanten-Diagramm .....	18
1.3	Multigraph .....	18
1.4	Property-Graph .....	19
1.5	Azyklischer Graph mit ungerichtetem Kreis .....	20
1.6	Gewurzelter Baum .....	21
1.7	Hierarchieebenen eines gewurzelten Baums .....	21
2.1	Häufigkeitsverteilung von Eigenangaben des Körpergewichts .....	41
2.2	Streudiagramm von Körpergröße und -gewicht .....	42
2.3	Streudiagramm von Broca- und Body-Mass-Index .....	43
2.4	Ringdiagramm .....	45
2.5	Säulendiagramme und Histogramme .....	46
2.6	Heatmap und Chloroplethenkarte .....	47
2.7	Histogramm einer rechtsschiefen Verteilung .....	50
2.8	Streuung um arithmetischen Mittelwert und Median .....	54
2.9	Anscombe-Quartett .....	61
3.1	Mengendiagramme von Wahrscheinlichkeitsmaßen .....	74
3.2	Verteilungsfunktion einer diskreten Zufallsvariable .....	86
3.3	Dichte- und Verteilungsfunktion der Standardnormalverteilung ..	88
3.4	Chi-Quadrat-Verteilung und Student'sche $t$ -Verteilung .....	112
4.1	Massenfunktionen von parametrischen Modellen diskreter Zufallsvariablen .....	124
4.2	Dichtefunktionen von parametrischen Modellen stetiger Zufallsvariablen .....	125
4.3	Zentraler Grenzwertsatz von Lindeberg-Lévy .....	134
4.4	Vertrauensintervalle bei verschiedenen Stichprobenumfang ..	144
4.5	Pareto-Verteilung höherer Einkommen .....	147

4.6	Normalverteilung der Körpergröße . . . . .	148
4.7	Bayes'sche Schätzung eines arithmetischen Mittelwerts . . . . .	153
4.8	Prinzip der Kerndichteschätzung . . . . .	154
4.9	Kerndichteschätzung verschiedener Bandbreite . . . . .	156
4.10	Globale Temperaturentwicklung mit Ausgleichsgerade . . . . .	159
4.11	Vertrauens- und Vorhersagebereich der linearen Regression . . . . .	160
4.12	Theil-Sen-Regression . . . . .	162
5.1	Geometrischer Median und Schwerpunkt; Kovarianzellipse . . . . .	180
5.2	Dichtefunktion einer bivariate Normalverteilung . . . . .	182
6.1	Globale Temperaturentwicklung mit Ausgleichspolynomen verschiedenen Grades . . . . .	197
6.2	Globale Temperaturentwicklung und quadratisches Ausgleichspolynom . . . . .	210
6.3	Gewöhnlicher und stochastischer Gradientenabstieg . . . . .	212
6.4	Globale Temperaturentwicklung und Ausgleichskurve einer Gauß-Prozess-Regression . . . . .	218
6.5	Entscheidungsgrenze der logistischen Regression . . . . .	219
6.6	Logistische Regression mittels polynomiaier Merkmale und Kernel-Methode . . . . .	225
6.7	<i>K</i> -nächste-Nachbarn-Klassifikation . . . . .	226
6.8	Feedforward-Netzwerk . . . . .	233
6.9	Sigmoidfunktion . . . . .	234
6.10	Neuronen und interneuronale Synapsen eines Fadenwurms . . . . .	240
6.11	Klassifikation mit neuronalem Netzwerk . . . . .	241
6.12	FaltungsfILTER in der Bildverarbeitung . . . . .	246
6.13	Architektur des Convolutional Neural Networks VGG-16 . . . . .	250
7.1	Datenpunkte entlang einer Kurve . . . . .	258
7.2	Kreisscheibenüberdeckung . . . . .	260
7.3	„Eigenziffern“ des MNIST-Datensatzes . . . . .	264
7.4	Bildrekonstruktion anhand von Hauptkomponenten . . . . .	264
7.5	Hauptkomponentenanalyse des MNIST-Datensatzes . . . . .	266
7.6	Autoencoding am Beispiel des MNIST-Datensatzes . . . . .	267
7.7	Multidimensionale Skalierung und <i>t</i> -SNE-Verfahren . . . . .	271
7.8	<i>K</i> -Means-Verfahren . . . . .	278
7.9	<i>K</i> -Means-Verfahren nach der Kernel-Methode . . . . .	279
7.10	Hierarchische Clusteranalyse deutscher Städte . . . . .	280
8.1	Kreuzvalidierung eines KNN-Klassifikators . . . . .	285
8.2	Falsch positive Ergebnisse der Klassifikation von MNIST-Ziffern . . . . .	285
8.3	Falsch negative Ergebnisse der Klassifikation von MNIST-Ziffern . . . . .	286
8.4	Falsch positive Ergebnisse der Klassifikation von CIFAR-10-Bildern . . . . .	287
8.5	Falsch negative Ergebnisse der Klassifikation von CIFAR-10-Bildern . . . . .	287

8.6	MNIST- und CIFAR-10-Datensatz . . . . .	289
8.7	Beispielcode: Convolutional Neural Network . . . . .	290
8.8	Themenkarten für Familien- und Science-Fiction-Filme . . . . .	296
8.9	Kooperationsnetzwerk von Schauspielern/Schauspielerinnen . . . . .	297
8.10	Hierarchische Clusteranalyse eines Kooperationsnetzwerks . . . . .	298

---

# Tabellenverzeichnis

1.1	Personendaten . . . . .	14
1.2	Heterogene, nichtnormierte Daten . . . . .	25
1.3	Fehlende Attributwerte . . . . .	27
1.4	Mittelwertimputation . . . . .	27
1.5	Verwendung von Imputationsklassen . . . . .	27
1.6	Datenqualitätsmerkmale . . . . .	34
2.1	Häufigkeit verschiedener Körpergrößen . . . . .	37
2.2	CDC-Stichprobe . . . . .	38
2.3	Merkmaltypen . . . . .	39
2.4	Korrelationsstärken . . . . .	58
2.5	Rangstatistiken, Beispiel „Schulnoten“ . . . . .	60
2.6	Korrelationskennzahlen für das Anscombe-Quartett . . . . .	60
2.7	Kontingenztafel von Parteipräferenz und Meinung zur Notwendigkeit von Migrationskontrolle . . . . .	64
2.8	Jaccard-Koeffizient als Kennzahl für Wählerbindung . . . . .	65
3.1	Stichwortextraktion über Transinformationsgehalt . . . . .	78
4.1	Kritische Werte als Funktion des Vertrauensniveaus . . . . .	135
4.2	Kritische Werte als Funktion des Stichprobenumfangs . . . . .	141
4.3	Effektstärken . . . . .	143
4.4	Aufzeichnung einer Lotterie . . . . .	149
5.1	Maße von Kelch- und Kronblatt von Schwertlilien . . . . .	169
6.1	Kostenmatrix, Beispiel „Spamfilter“ . . . . .	194
6.2	Richtig/falsch positive/negative Ergebnisse . . . . .	201
6.3	Wahrheitsmatrix, Beispiel „Spamfilter“ . . . . .	203
6.4	Schichttypen eines Convolutional Neural Networks . . . . .	248
8.1	Güte verschiedener Klassifikatoren für den MNIST-Datensatz . . . . .	284

## XXIII Tabellenverzeichnis

8.2	Güte verschiedener Klassifikatoren für den CIFAR-10-Datensatz .	286
8.3	Positive/negative Filmrezensionen .....	291
8.4	Positiv konnotierte Texteinheiten, aufwertende Begriffe .....	291
8.5	Negativ konnotierte Texteinheiten, abwertende Begriffe .....	291
8.6	Güte verschiedener Verfahren der Sentimentanalyse .....	292
8.7	Güte naiver Bayes-Klassifikatoren auf Basis von <i>N</i> -Grammen ..	292
8.8	Fehlklassifikationen einer Sentimentanalyse .....	292
8.9	Filmgenretypische Nominalphrasen .....	294



---

## Einführung

**Daten** sind gemäß internationalem Technologiestandard [1] eine „formalisierte Darstellung von Informationen, welche für die Kommunikation, Interpretation oder Verarbeitung geeignet sind<sup>1</sup>“. Eine weitere Charakterisierung liefert der Duden [2]: Daten sind „(durch Beobachtungen, Messungen, statistische Erhebungen u. a. gewonnene) [Zahlen]werte, (auf Beobachtungen, Messungen, statistischen Erhebungen u. a. beruhende) Angaben, formulierbare Befunde“.

Eine zentrale Aufgabe der Datenwissenschaft ist – in der Regel unter Zuhilfenahme von informationstechnischen Mitteln – die Erfassung, Verarbeitung, Interpretation und Kommunikation von Daten mit dem Ziel der Gewinnung von belastbarem und nutzbringendem Wissen.

In den empirischen Wissenschaften, etwa den Naturwissenschaften, ist die Erfassung und Auswertung von Daten seit langer Zeit ein wesentlicher Bestandteil der Erkenntnisgewinnung. Die moderne Physik wäre zum Beispiel ohne ein Zusammenwirken von Theorie und Experiment kaum denkbar: Abweichungen von experimentellen Daten zeigen hier die Grenzen theoretischer Modelle auf. Ein Erfolg der Newton'schen Himmelsmechanik bestand etwa in der genauen Beschreibung und Vorhersage der Bewegungen von Planeten und anderer Himmelskörper im Sonnensystem. Die im 19. Jahrhundert von Urbain Le Verrier genau vermessenen Bahnstörungen des Merkur [3] konnte diese jedoch nicht vollständig erklären – das gelang erst mit der später entwickelten Einstein'schen Theorie der Gravitation, der allgemeinen Relativitätstheorie.

Die Erfassung und Auswertung von Daten ist heute auch im betriebswirtschaftlichen Kontext von großer Bedeutung: Daten können als ein Rohstoff aufgefasst werden, mit denen das Wirtschaftsgut Wissen produziert wird. Datenanalysen dienen auch der Unterstützung von Managemententscheidungen (Schlagworte: **Business Intelligence**, **Business Analytics**). Patentdaten können etwa genutzt werden, um die Patentierungsaktivität und Vernetzung verschiedener Akteure im Wettbewerbsumfeld aufzuzeigen. Diese und weitere Informationen

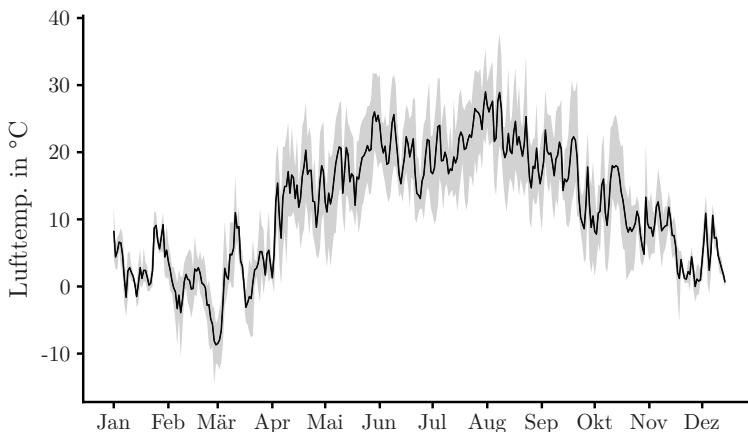
---

<sup>1</sup> *a reinterpretable representation of information in a formalized manner, suitable for communication, interpretation, or processing*

können dann in strategische Entscheidungen, etwa im Innovations- oder Technologiemanagement der Unternehmung, einfließen [4].

Ein weiteres Beispiel sind demografische Untersuchungen. Statistisch verarbeitete Daten dienen etwa der zahlenmäßigen Beschreibung regionaler Unterschiede im Mobilitätsverhalten der Bewohnerinnen und Bewohner (vgl. [5]). Solche Informationen können nutzbringend für die Städteplanung eingesetzt werden.

Daten oder die Ergebnisse einer Datenanalyse können dem Menschen sprachlich kommuniziert werden. Eine weitere Möglichkeit ist die visuelle, grafische Darstellung; ein Beispiel hierfür ist die folgende Abbildung, diese zeigt den zeitlichen Verlauf der im Jahr 2018 in Berlin-Tegel zwei Meter über dem Boden gemessenen Lufttemperatur [6]:



**Abb. 0.1.** Liniendiagramm eines Temperaturverlaufs

Der graue Streifen ist durch die tägliche Schwankung zwischen Minimal- und Maximaltemperatur bestimmt. Insbesondere die folgenden Informationen können auf einem Blick abgelesen werden:

- Ende Februar und Anfang März waren in diesem Jahr die kältesten Tage.
- Offensichtlich ist es im Winter kälter als im Sommer, die Abbildung macht zugleich aber auch deutlich, um wie viel mehr sich im Sommer die Tag- von den Nachttemperaturen unterscheiden: Der graue Streifen ist in diesem Zeitraum deutlich breiter als im Winter.

Dieselben Informationen könnten anhand der etwa tabellarisch aufgestellten Messreihe nicht ohne Weiteres – oder zumindest nicht so effizient – ermittelt werden. Der Bereich der Datenanalyse, welcher sich schwerpunktmäßig mit grafischer Darstellung auseinandersetzt, ist die **Datenvizualisierung**.

Eine weitere Möglichkeit der Kommunikation besteht in der Umwandlung der Daten in akustische nichtsprachliche Signale (**Sonifikation**). Diese ist jedoch weit weniger verbreitet als die grafische Darstellung.

Zusammenfassend lässt sich festhalten:

Eine **Datenanalyse** hat zum Ziel, durch systematische Organisation, statistische Verarbeitung und/oder grafische (ferner: akustische, audiovisuelle) Darstellung von Daten (in diesem Zusammenhang auch **Rohdaten** genannt) relevante Informationen und/oder nutzbringendes Wissen zu gewinnen.

Die folgenden Formen der Datenanalyse sind durch ihre jeweilige Zielsetzung charakterisiert:

- Die **deskriptive Datenanalyse** bzw. Statistik dient der Organisation und zusammenfassenden Darstellung der Daten.
- Durch **explorative Datenanalyse** sollen verborgene Muster in den Daten aufgeklärt werden, etwa um neue Hypothesen zu formulieren.
- Die **Inferenzstatistik** hat zum Ziel, die Beobachtungen durch statistische Modelle zu beschreiben und die Gültigkeit von Hypothesen anhand der Daten zu prüfen.

Werden explorative Datenanalysen „im großen Stil“ durchgeführt, so wird auch von **Data-Mining** gesprochen. Mögliche Beschreibungen von Data-Mining sind „der Prozess der Aufklärung interessanter Muster in enorm großen Datenmengen<sup>2</sup>“ [7, Abschn. 1.8] oder die „[halb] automatische Auswertung großer Datenmengen zur Bestimmung bestimmter Regelmäßigkeiten, Gesetzmäßigkeiten und verborgener Zusammenhänge“ [8]. Eine weitere zentrale Aufgabe der heutigen Datenwissenschaft besteht in der Entwicklung von Algorithmen für intelligente und autonome Systeme mit Methoden des **maschinellen Lernens**: Die Regeln, die das Verhalten solcher Systeme steuern, sind nicht in Gänze explizit vorgegeben, sondern werden im Wesentlichen anhand von Trainingsdaten „erlernt“.

In dem vorliegenden Buch werden strukturwissenschaftliche Methoden und rechnerische Verfahren beschrieben und fundiert, welche geeignet sind, die oben beschriebenen Aufgaben zu erfüllen. Es ist in drei Teile gegliedert. Der erste Teil befasst sich zum einen mit Aspekten der **Datenorganisation**: die konzeptionelle und logische Strukturierung von Daten sowie die Sicherstellung von deren Qualität. Zum anderen wird in das Feld der **deskriptiven Statistik** eingeführt, die der übersichtlichen Darstellung und der Zusammenfassung wesentlicher Charakteristiken der erfassten Daten dient.

Der zweite Teil führt an das mathematische Gebiet der **Stochastik** heran, welches die Wahrscheinlichkeitsrechnung und die Inferenzstatistik umfasst. Die Stochastik stellt wesentliche konzeptionelle Grundlagen und mathematische Werkzeuge für das Schließen unter Unsicherheit bereit.

---

<sup>2</sup> the process of discovering interesting patterns from massive amounts of data

Im letzten Teil des Buches werden dann verschiedene algorithmische Verfahren des maschinellen Lernens vorgestellt, welche wesentlichen Gebrauch von den im zweiten Teil vorgestellten Konzepten machen. Ferner wird auch auf Aspekte der Theorie des maschinellen Lernens eingegangen.

Dieser Band soll die Leserinnen und Leser an wesentliche Themen und Methoden der Data Science in einer inhaltlichen Tiefe heranführen, wie sie für eine Einführung angemessen sind. Am Ende des Buches findet sich eine Zusammenstellung weiterführender und ergänzender Literatur.

Für das erfolgreiche Studium des Buches werden mathematische Kenntnisse vorausgesetzt, wie sie etwa in den ersten zwei bis drei Semestern eines Hochschulstudiums mathematischer, technischer oder naturwissenschaftlicher Richtung gelehrt werden. Im Kern sind dies Kenntnisse der lineare Algebra, Analysis und einiger Aspekte der mehrdimensionalen Analysis. Ein Großteil der zitierten Fachliteratur und der Beispieldatensätze zum Thema Textanalyse ist englischsprachig, sodass entsprechende Sprachkenntnisse von Vorteil sind.

„Übung macht den Meister“ – das gilt natürlich auch oder gerade für ein Studium der Data Science. Obgleich diese Auflage keine expliziten Übungsaufgaben enthält, so werden doch zahlreiche Anwendungsbeispiele auf Grundlage von Daten diskutiert, die frei verfügbar sind. Leserinnen und Leser sind herzlich eingeladen, diese Beispiele mithilfe einer Programmiersprache für statistisches Rechnen ihrer Wahl – zum Beispiel R [9] oder Python [10] – nachzuvollziehen und um eigene Analysen zu erweitern.

Ein Großteil der Abbildungen in diesem Band wurden mithilfe des Datenvisualisierungspakets ggplot2 für R erzeugt [11], händisch erstellte Knoten-Kanten-Diagramme von Graphen und Netzwerken mithilfe des LATEX-Pakets TikZ [12]. Weitere Beispiele von Programmbibliotheken für R, die maßgeblich für die Berechnung und Darstellung der Beispiele verwendet und nicht explizit im Text erwähnt wurden, sind Cairo [13], cowplot [14], dplyr [15], extrafont [16], fast-cluster [17], FNN [18],forcats [19], ggdendro [20], ggforce [21], ggrepel [22], Gmedian [23], gridExtra [24], igraph [25], latex2exp [26], lubridate [27], magick [28], mapproj [29], MASS [30], mlbench [31], mvtnorm [32, 33], neuralnet [34], proxy [35], randomNames [36], reshape2 [37], scales [38], sna [39], sp [40, 41], stringdist [42], stringr [43], tidyverse [44], tidytext [45], tsne [46] und usedist [47].

## Quellen

- [1] ISO Central Secretary. *Information technology – Vocabulary*. Standard ISO/IEC 2382:2015. Genf, Schweiz: International Organization for Standardization, 2015, S. 2121272.
- [2] Dudenredaktion. „*Daten*“ auf Duden online. URL: <https://www.duden.de/node/30506/revision/30535>.
- [3] Clifford M. Will. *Theory and Experiment in Gravitational Physics*. Cambridge University Press, Sep. 2018. DOI: [10.1017/9781316338612](https://doi.org/10.1017/9781316338612).
- [4] Peter Walde u. a. „Erstellung von Technologie- und Wettbewerbsanalysen mithilfe von Big Data“. In: *Wirtschaftsinformatik & Management* 5.2 (Feb. 2013), S. 12–23. DOI: [10.1365/s35764-013-0274-7](https://doi.org/10.1365/s35764-013-0274-7).
- [5] infas Institut für angewandte Sozialwissenschaft GmbH. *Mobilität in Deutschland – MiD*. 2017. URL: <http://www.mobilitaet-in-deutschland.de/publikationen2017.html>.
- [6] Deutscher Wetterdienst – Zentraler Vertrieb Klima und Umwelt. *Klimadaten Deutschland*. Aufgerufen am 01. Apr. 2020. Offenbach. URL: <https://www.dwd.de/DE/leistungen/klimadatendeutschland/klimadatendeutschland.html>.
- [7] Jiawei Han, Micheline Kamber und Jian Pei. *Data Mining: Concepts and Techniques*. 3. Aufl. Elsevier, 2012. ISBN: 9-380-93191-3.
- [8] Dudenredaktion. „*Data-Mining*“ auf Duden online. URL: <https://www.duden.de/node/30498/revision/30527>.
- [9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. 2020. URL: <https://www.R-project.org/>.
- [10] Guido van Rossum und Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, USA: CreateSpace, 2009. ISBN: 1441412697.
- [11] Hadley Wickham. *ggplot2. Elegant Graphics for Data Analysis*. Springer, New York, 2009. DOI: [10.1007/978-0-387-98141-3](https://doi.org/10.1007/978-0-387-98141-3).
- [12] Till Tantau. *The TikZ and PGF Packages. Manual for version 3.1.7*. Nov. 2020. URL: <https://pgf-tikz.github.io/pgf/pgfmanual.pdf>.
- [13] Simon Urbanek und Jeffrey Horner. *Cairo: R Graphics Device using Cairo Graphics Library for Creating High-Quality Bitmap (PNG, JPEG, TIFF), Vector (PDF, SVG, PostScript) and Display (X11 and Win32) Output*. R-Paket, Version 1.5-12.2. 2020. URL: <https://CRAN.R-project.org/package=Cairo>.
- [14] Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R-Paket, Version 1.1.0. 2020. URL: <https://CRAN.R-project.org/package=cowplot>.
- [15] Hadley Wickham u. a. *dplyr: A Grammar of Data Manipulation*. R-Paket, Version 1.0.2. 2020. URL: <https://CRAN.R-project.org/package=dplyr>.
- [16] Winston Chang. *extrafont: Tools for using fonts*. R-Paket, Version 0.17. 2014. URL: <https://CRAN.R-project.org/package=extrafont>.
- [17] Daniel Müllner. „*fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python*“. In: *Journal of Statistical Software* 53.9 (2013), S. 1–18. URL: <http://www.jstatsoft.org/v53/i09/>.

- [18] Alina Beygelzimer u. a. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. R-Paket, Version 1.1.3. 2019. URL: <https://CRAN.R-project.org/package=FNN>.
- [19] Hadley Wickham. *forcats: Tools for Working with Categorical Variables (Factors)*. R-Paket, Version 0.5.0. 2020. URL: <https://CRAN.R-project.org/package=forcats>.
- [20] Andrie de Vries und Brian D. Ripley. *ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'*. R-Paket, Version 0.1.22. 2020. URL: <https://CRAN.R-project.org/package=ggdendro>.
- [21] Thomas Lin Pedersen. *ggforce: Accelerating 'ggplot2'*. R-Paket, Version 0.3.2. 2020. URL: <https://CRAN.R-project.org/package=ggforce>.
- [22] Kamil Slowikowski. *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. R-Paket, Version 0.8.2. 2020. URL: <https://CRAN.R-project.org/package=ggrepel>.
- [23] Herve Cardot. *Gmedian: Geometric Median, k-Median Clustering and Robust Median PCA*. R-Paket, Version 1.2.5. 2020. URL: <https://CRAN.R-project.org/package=Gmedian>.
- [24] Baptiste Auguie. *gridExtra: Miscellaneous Functions for 'Grid' Graphics*. R-Paket, Version 2.3. 2017. URL: <https://CRAN.R-project.org/package=gridExtra>.
- [25] Gabor Csardi und Tamas Nepusz. „The igraph software package for complex network research“. In: *InterJournal Complex Systems* (2006), S. 1695. URL: <https://igraph.org>.
- [26] Stefano Meschiari. *latex2exp: Use LaTeX Expressions in Plots*. R-Paket, Version 0.4.0. 2015. URL: <https://CRAN.R-project.org/package=latex2exp>.
- [27] Garrett Grolemund und Hadley Wickham. „Dates and Times Made Easy with lubridate“. In: *Journal of Statistical Software* 40.3 (2011), S. 1–25. URL: <https://www.jstatsoft.org/v40/i03>.
- [28] Jeroen Ooms. *magick: Advanced Graphics and Image-Processing in R*. R-Paket, Version 2.5.2. 2020. URL: <https://CRAN.R-project.org/package=magick>.
- [29] Doug McIlroy u. a. *mapproj: Map Projections*. R-Paket, Version 1.2.7. 2020. URL: <https://CRAN.R-project.org/package=mapproj>.
- [30] W. N. Venables und B. D. Ripley. *Modern Applied Statistics with S*. 4. Aufl. ISBN 0-387-95457-0. Springer, New York, 2002. DOI: [10.1007/978-0-387-21706-2](https://doi.org/10.1007/978-0-387-21706-2).
- [31] Friedrich Leisch und Evgenia Dimitriadou. *mlbench: Machine Learning Benchmark Problems*. R-Paket, Version 2.1-1. 2010. URL: <https://CRAN.R-project.org/package=mlbench>.
- [32] Alan Genz u. a. *mvtnorm: Multivariate Normal and t Distributions*. R-Paket, Version 1.1-1. 2020. URL: <https://CRAN.R-project.org/package=mvtnorm>.
- [33] Alan Genz und Frank Bretz. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer, Berlin, Heidelberg, 2009. ISBN: 978-3-642-01688-2.

- [34] Stefan Fritsch, Frauke Günther und Marvin N. Wright. *neuralnet: Training of Neural Networks*. R-Paket, Version 1.44.2. 2019. URL: <https://CRAN.R-project.org/package=neuralnet>.
- [35] David Meyer und Christian Buchta. *proxy: Distance and Similarity Measures*. R-Paket, Version 0.4-24. 2020. URL: <https://CRAN.R-project.org/package=proxy>.
- [36] Damian W. Betebenner. *randomNames: Function for Generating Random Names and a Dataset*. R-Paket, Version 1.4-0.0. 2019. URL: <https://cran.r-project.org/package=randomNames>.
- [37] Hadley Wickham. „Reshaping Data with the reshape Package“. In: *Journal of Statistical Software* 21.12 (2007), S. 1–20. URL: <http://www.jstatsoft.org/v21/i12/>.
- [38] Hadley Wickham und Dana Seidel. *scales: Scale Functions for Visualization*. R-Paket, Version 1.1.1. 2020. URL: <https://CRAN.R-project.org/package=scales>.
- [39] Carter T. Butts. *sna: Tools for Social Network Analysis*. R-Paket, Version 2.6. 2020. URL: <https://CRAN.R-project.org/package=sna>.
- [40] Edzer J. Pebesma und Roger S. Bivand. „Classes and methods for spatial data in R“. In: *R News* 5.2 (Nov. 2005), S. 9–13. URL: <https://CRAN.R-project.org/doc/Rnews/>.
- [41] Roger S. Bivand, Edzer Pebesma und Virgilio Gomez-Rubio. *Applied spatial data analysis with R*. 2. Aufl. Springer, New York, 2013. URL: <https://asdar-book.org/>.
- [42] Mark P. J. van der Loo. „The stringdist package for approximate string matching“. In: *The R Journal* 6 (1 2014), S. 111–122. URL: <https://CRAN.R-project.org/package=stringdist>.
- [43] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. R-Paket, Version 1.4.0. 2019. URL: <https://CRAN.R-project.org/package=stringr>.
- [44] Hadley Wickham u. a. „Welcome to the tidyverse“. In: *Journal of Open Source Software* 4.43 (2019), S. 1686. doi: [10.21105/joss.01686](https://doi.org/10.21105/joss.01686).
- [45] Julia Silge und David Robinson. „tidytext: Text Mining and Analysis Using Tidy Data Principles in R“. In: *JOSS* 1.3 (2016). doi: [10.21105/joss.00037](https://doi.org/10.21105/joss.00037).
- [46] Justin Donaldson. *tsne: t-Distributed Stochastic Neighbor Embedding for R (t-SNE)*. R-Paket, Version 0.1-3. 2016. URL: <https://CRAN.R-project.org/package=tsne>.
- [47] Kyle Bittinger. *usedist: Distance Matrix Utilities*. R-Paket, Version 0.4.0. 2020. URL: <https://CRAN.R-project.org/package=usedist>.

## **Teil I**

---

### **Grundlagen**



# 1

---

## Elemente der Datenorganisation

Als **Datenbestand** bezeichnen wir alle erfassten, gespeicherten und für den Zugriff und die Verarbeitung zur Verfügung stehenden Daten. Große Datenbestände sind nicht selten **heterogen**: Sie entstehen durch **Datenintegration**, dem Zusammenführen von Daten aus verschiedenen und verschiedenartigen **Datenquellen**. Bei der Planung des Aufbaus eines großen und/oder heterogenen Datenbestands (Schlagwort: **Big Data**) sowie der Sicherstellung der Qualität der Daten ergeben sich besondere Herausforderungen, auf die wir in diesem Kapitel eingehen.

Zunächst beschreiben wir Ansätze zur Modellierung von Daten. Das vorrangige Ziel der Datenmodellierung besteht darin, den Datenbestand inhaltlich zu erfassen und logisch zu strukturieren.

Es können drei Phasen oder Meilensteine des Datenmodellierungsprozesses unterschieden werden, die jeweils der Beantwortung der folgenden Fragen dienen [1]:

1. **Konzeptionelles Datenmodell:** Welche Objekte der Wissensdomäne sind relevant und sollen durch die Daten beschrieben werden? Welche Eigenschaften haben die Objekte, und in welchen Beziehungen stehen sie zueinander?
2. **Logisches Datenmodell:** Wie sind oder werden die Daten selbst strukturiert (also angeordnet, gruppiert, miteinander verknüpft oder verglichen)?
3. **Physisches Datenmodell:** Wie sieht die konkrete hardware- und softwaretechnische Umsetzung der Datenerfassung und -haltung aus?

Beim konzeptionellen Datenmodell handelt sich um eine systematische Beschreibung der **Semantik** der Daten im Kontext der Wissensdomäne, während beim logischen Datenmodell der Fokus auf deren formale Struktur liegt.

In der Praxis unterscheiden sich diese Phasen der Modellierung insbesondere im Detaillierungsgrad: Das konzeptionelle Datenmodell zielt auf eine globale

Beschreibung der zugrundeliegenden Konzepte ab, während das logische Datenmodell in der Regel eine größere Detailschärfe aufweist. Nicht selten werden beide Phasen aber auch zusammengezogen.

Die physische Datenmodellierung befasst sich schwerpunktmäßig mit den technischen Herausforderungen, die z.B. mit der Erfassung und Validierung, der performanten Verarbeitung und skalierbaren Haltung der Daten mithilfe eines konkreten Datenbankverwaltungssystems verbunden sind. Wir gehen in diesem Buch nicht näher auf diese Herausforderungen des **Data Engineering** ein.

Hiernach stellen wir Konzepte und Verfahren für die Bemessung und Sicherstellung der Qualität der bereitgestellten Daten vor.

## 1.1 Konzeptionelle Datenmodellierung

Eine Verarbeitung von Daten ist nur dann von Nutzen, wenn diese eine Repräsentation von konkreten Inhalten unserer Erfahrungswelt sind. Im Allgemeinen besitzen Daten nicht nur eine formale Beschreibungsebene als numerische Werte, Zeichenketten usw., sondern auch eine inhaltliche.

Daten repräsentieren Informationen über individuell identifizierbare reale oder abstrakte Objekte, in diesem Zusammenhang **Entitäten** oder **Informationsobjekte** genannt.

Bei den Entitäten könnte es sich beispielsweise um Personen in einer Kunden-datenbank handeln und bei den Daten um Namen sowie Adressen des jeweiligen Wohnsitzes.

Handelt es sich bei den durch die Daten beschriebenen Entitäten selbst um Daten, so werden die beschreibenden Daten auch als **Metadaten** bezeichnet. Ein wichtiges Beispiel für Metadaten ist die menschenlesbare Beschreibung der Daten, die **Datendokumentation** (mehr dazu in Abschn. 1.3.1).

Ein **konzeptionelles Datenmodell** ist eine systematische Darstellung der Entitäten, ihrer relevanten Eigenschaften und ihrer Beziehungen untereinander. Ein solches Modell kann z.B. bei der Planung eines Datenanalyseprojekts der Feststellung des Informationsbedarfs dienen.

### 1.1.1 Entity-Relationship-Modell

Eine verbreitete systematische Darstellungsform eines Datenmodells ist das **Entity-Relationship-Modell** (ERM) [2]. Zunächst halten wir fest, dass gleichartige Entitäten durch Abstraktion zu **Entitätstypen** zusammengefasst werden können: So ist etwa jede Person eine **Instanz** des Entitätstyps „Person“.

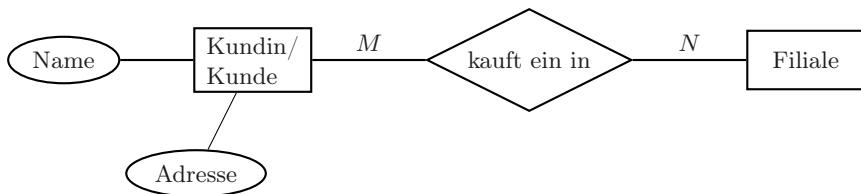
Weiterhin können jeder Entität bestimmte Eigenschaften zugeordnet werden, die zu **Attributen** des jeweiligen Entitätstyps zusammengefasst werden. Attribute des Entitätstyps „Person“ mögen etwa „Vorname“ und „Adresse“ sein.

Eine konkrete Belegung eines Attributs ist ein **Attributwert**. Mögliche Attributwerte für das Attribut „Vorname“ einer Person könnten etwa „Anna“ oder „Karl“ sein. Im Kontext der Statistik sprechen wir anstelle von Attributen bzw. Attributwerten auch von Merkmalen bzw. Merkmalsausprägungen.

Ein Teil der Datenmodellierung besteht in der Aufgabe, eine für die Wissensdomäne und den jeweiligen Anwendungsfall (neudeutsch: *Use Case*) sinnvolle Unterscheidung zwischen Attributen und Entitäten zu erklären. So könnte der Wohnort einer Person auch als Entität aufgefasst werden, die durch die Adresse repräsentiert würde.

Schließlich können Entitäten miteinander in **Beziehung** stehen. So könnten die Personen etwa Kundinnen und Kunden eines Einzelhandelunternehmens sein, die ihre Einkäufe regelmäßig in wenigstens einem Ladengeschäft tätigen. Diese Personen sind mit Verkaufsständen durch Beziehungen vom Typ „kauft ein in“ verknüpft.

Das so umschriebene Datenmodell kann schließlich in einem **Entity-Relationship-Diagramm** wie das folgende zusammengefasst werden:



**Abb. 1.1. Entity-Relationship-Diagramm**

Diese Art von Entity-Relationship-Diagramm wird **Chen-Notation** genannt: Entitätstypen werden durch Rechtecke dargestellt, Attribute durch ovale grafische Elemente, Beziehungstypen durch rautenförmige.

Die Bezeichner „*M*“ bzw. „*N*“ in obigem Diagramm charakterisieren die **Maximalkardinalität** des Beziehungstyps, wie im Folgenden zwischen zwei Entitätstypen *X* und *Y* beschrieben:

- **1:1-Beziehung:** Mit einer Entität vom Typ *Y* steht immer höchstens eine Entität vom Typ *X* in Beziehung und umgekehrt. So steht etwa – von Wechselkennzeichen abgesehen – jedes Kraftfahrzeug mit einem und nur einem Kfz-Kennzeichen in Beziehung.
- **1:N-Beziehung:** Mit einer Entität vom Typ *Y* steht immer höchstens eine Entität vom Typ *X* in Beziehung. Umgekehrt können Entitäten vom Typ *X* auch mit mehr als nur einer Entität vom Typ *Y* in Beziehung stehen. Beispielsweise hat eine Person nur einen angemeldeten Hauptwohnsitz, an diesem können grundsätzlich aber auch mehrere Personen wohnhaft sein.
- **M:N-Beziehung:** Es besteht keine zahlenmäßige Beschränkung in der Anzahl der Entitäten, die miteinander in Beziehung stehen. Ein Beispiel ist der

Beziehungstyp „Kundin/Kunde  $x$  kauft in Filiale  $y$  ein“: In einer Filiale kaufen in der Regel mehrere Personen ein, eine Person kann grundsätzlich auch in mehreren Filialen Geschäfte tätigen.

## 1.2 Logische Datenmodellierung

Ein konzeptionelles Datenmodell dient der Strukturierung der Wissensdomäne; ein Entity-Relationship-Modell ermöglicht dies durch eine Einteilung in Entitäten, Attribute und Beziehungen zwischen den Entitäten. Diesen Objekten der realen Welt stehen Daten gegenüber, die diese repräsentieren. So entspricht etwa der Vorname einer Person einer bestimmten Zeichenkette, der Wohnort könnte in Form von Längen- und Breitengrad als Paar von Gleitkommazahlen gespeichert werden. Um diese Daten einem informationsverarbeitenden System zugänglich zu machen, müssen auch sie in geeigneter Weise strukturiert sein oder werden. Dieses **logische Datenmodell** spiegelt das konzeptionelle Datenmodell strukturell wider.

Wir können ein logisches Datenmodell als Organisation kleinsten Einheiten verstehen:

Ein **Datenfeld** oder **Datenelement** ist die kleinste organisatorische Einheit von Daten. Ein **Datensatz** ist eine Zusammenfassung inhaltlich zusammenhängender Datenelemente, die in gleicher Art und Weise strukturiert (angeordnet, gruppiert, miteinander verknüpft oder verglichen) werden.

Im Folgenden beschreiben wir die Kernaspekte einiger Vertreter logischer Datenmodellierung.

### 1.2.1 Relationales Datenmodell

Eine sehr gebräuchliche Strukturierung von Datensätzen besteht in der Anordnung und Gruppierung von Datenfeldern in tabellarischer Form. Hier ein Beispiel eines Datensatzes von fiktiven personenbezogenen Kundendaten:

ID	Nachname	Vorname	Straße	Hausnr.	PLZ	Stadt
1	Schmidt	Anna	Musterstraße	1	12345	Modellstadt
2	Müller	Robert	Beispielplatz	3	54321	Musterdorf
3	Müller	Karl	Beispielplatz	3	54321	Musterdorf

**Tabelle 1.1.** Personendaten

Die Zeilen eines auf diese Weise strukturierten Datensatzes werden auch **Datentupel** genannt, oft aber auch selbst als Datensatz (im engeren Sinne) bezeichnet.

Das **Schlüsselattribut** „ID“ in der ersten Spalte dient der eindeutigen Identifikation jeder Entität. Eine Nummerierung mittels positiver ganzer Zahlen stellt eine einfache Umsetzung eines solchen Schlüsselattributs dar.

Ein Entity-Relationship-Modell kann wie folgt durch ein **relationales Datenmodell** logisch repräsentiert werden; eine Datentabelle wird in diesem Kontext dann auch als **Relation** bezeichnet:

**Entitäten:** Jeder Entitätstyp entspricht einer Relation. Jede Zeile der Tabelle entspricht einer Entität (in obigem Beispiel einer konkreten Person).

**Attribute:** Die Spalten entsprechen Attributen, und die Attributwerte konstituieren ein Datentupel.

**Beziehungen:** Jeder Beziehungstyp entspricht einer Relation. Jede Zeile der Tabelle entspricht einer Beziehung: Ein Datentupel setzt sich dabei aus den Werten der Schlüsselattribute jener Entitäten zusammen, die miteinander in Beziehung stehen.

Zur Illustration betrachten wir wieder den Datenbestand eines fiktiven Einzelhandelunternehmens. Dieser bestehe noch aus folgender Liste von Verkaufsstationalen:

ID	Name	Stadt
1	Günstigland	Musterdorf
2	Günstigland	Modellstadt

Der Beziehungstyp „kauft ein in“ könnte etwa durch folgende Relation realisiert werden:

Kundin/ Kunde	ID Filiale
1	2
2	1
2	2
3	1

Die Kundin oder der Kunde mit dem Schlüsselattribut „1“ kauft in der Filiale mit dem Schlüsselattribut „2“ ein usw.

Relationen können durch die Anwendung von Operationen der sogenannten **relationalen Algebra** miteinander verknüpft werden. Um diese Operationen zu erklären, ist es hilfreich, eine Relation mathematisch als Menge von Datentupeln aufzufassen. Eine Relation  $R$  mit  $D = D(R)$  Spalten bzw. Attributen mit Wertebereichen  $R_1, \dots, R_D$  sowie  $N$  Zeilen bzw. Datentupeln kann damit wie folgt geschrieben werden:

$$R = \{(r^1_1, \dots, r^1_D), (r^2_1, \dots, r^2_D), \dots, (r^N_1, \dots, r^N_D)\}$$

Dabei ist  $r^n_d \in R_d$  für alle  $d \in \{1, \dots, D\}$  und  $n \in \{1, \dots, N\}$  der Eintrag in der  $n$ -ten Zeile und der  $d$ -ten Spalte der Relation. Der Zeilenindex wird im Folgenden eine untergeordnete Rolle spielen, daher schreiben wir kurz:  $R = \{(r_1, \dots, r_D)\}$ .

Zwei Relationen  $R$  und  $S$  heißen **typkompatibel**, falls sie durch identische Attribute beschrieben werden: Die Anzahl der Attribute  $D$  beider Relationen ist identisch und es gilt  $R_1 = S_1, R_2 = S_2, \dots, R_D = S_D$ .

Für typkompatible Relationen  $R$  und  $S$  können sinnvoll die üblichen Mengenoperationen definiert werden: **Vereinigung**  $R \cup S$ , **Schnitt**  $R \cap S$ , **Differenz**  $R \setminus S$ .

Für beliebige, nicht notwendigerweise typkompatible Relationen ist auch das **kartesische Produkt** wohldefiniert:

$$R \times S = \{(r_1, \dots, r_{D(R)}, s_1, \dots, s_{D(S)})\}$$

Die Auswahl bestimmter Attribute bzw. Spalten wird **Projektion** genannt:

$$R = \{(r_1, \dots, r_D)\} \mapsto \{(r_{\iota(1)}, \dots, r_{\iota(K)})\} = \pi_{\iota(1), \dots, \iota(K)}(R)$$

mit  $\iota: \{1, \dots, K\} \rightarrow \{1, \dots, D\}$ ,  $K \leq D$ .

Eine Auswahl von Zeilen auf Grundlage einer an die Attributwerte zu stellende Bedingung  $\mathcal{B}$  wird **Selektion** genannt:

$$R[\mathcal{B}] = \{r \in R | r \text{ erfüllt } \mathcal{B}\}$$

Der **Verbund** bzw. **Join** ist eine Hintereinanderausführung von kartesischem Produkt und Selektion:

$$R \bowtie_{\mathcal{B}} S = (R \times S)[\mathcal{B}]$$

Für ein konkretes Beispiel im Umgang mit obigen Operationen betrachten erneut die Relationen, welche die Entitätstypen „Kundin/Kunde“ ( $=: K$ ) und „Filialen“ ( $=: F$ ) sowie den Beziehungstyp „kaufte ein in“ ( $=: R$ ) repräsentieren. Wir wollen eine neue Relation konstruieren, die Nachname und Wohnort jener Kundinnen und Kunden auflistet, die in Filialen am Standort „Modellstadt“ ( $=: x$ ) einkaufen.

Dies kann z. B. wie folgt erreicht werden:

1. Verbund der der einzelnen Relationen über die Schlüsselattribute. Wir erhalten die neue Relation  $K \bowtie_{k_1=r_1} R \bowtie_{r_2=f_1} F$ .
2. Selektion nach den gewünschten Filialen:  $K \bowtie_{k_1=r_1} R \bowtie_{r_2=f_1} F[f_3 = x]$ .

3. Projektion auf die Attribute „Nachname“ und „Stadt“ der Personen (2. und 7. Spalte); wir erhalten schließlich:

$$\pi_{2,7}(K \bowtie_{k_1=r_1} R \bowtie_{r_2=f_1} F[f_3 = x])$$

In der Umsetzung würden diese Operationen in der Regel mithilfe der standardisierten und weit verbreiteten Datenbanksprache SQL (Structured Query Language) formuliert. Diese verfügt über eine vergleichsweise einfache und „sprechende“ Syntax. Eine in SQL geschriebene Datenbankabfrage für obiges Beispiel könnte etwa wie folgt aussehen:

```
SELECT
    K.Nachname,
    K.Stadt
FROM Kunden K
JOIN kauft_ein_in R ON K.ID = R.ID_Kunde
JOIN Filialen F ON F.ID = R.ID_Filiale
WHERE F.Stadt = 'Modellstadt';
```

SQL ermöglicht die Umsetzung weiterer wichtiger Operationen über die der relationalen Algebra hinaus. So können Entitäten auch **gruppiert** und deren Attributwerte **aggregiert** werden. Eine Datenbankabfrage zur Ermittlung der Gesamtzahl der Kundinnen und Kunden jeder Filiale könnte etwa so aussehen:

```
SELECT
    F.ID, F.Name, F.Stadt,
    COUNT(DISTINCT K.ID) AS Anzahl_Kunden
FROM Kunden K
JOIN kauft_ein_in R ON K.ID = R.ID_Kunde
JOIN Filialen F ON F.ID = R.ID_Filiale
GROUP BY F.ID;
```

Datenbankverwaltungssysteme, mit denen relationale Datenmodelle umgesetzt werden können, sind zum Beispiel MySQL [3], MariaDB [4] und PostgreSQL [5].

## 1.2.2 Graphbasierte Datenmodelle

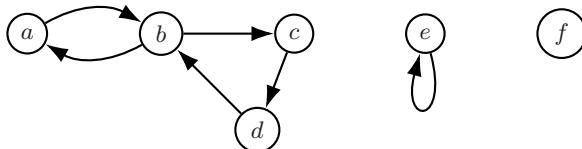
Wesentliche mathematische Grundlage des relationalen Datenmodells ist die Vorstellung von Relationen als Mengen von Datentupeln. Diese sind über die Operationen der relationalen Algebra miteinander verknüpft. Graphbasierten Datenmodellen liegt hingegen der folgende mathematische Begriff zugrunde:

Ein **gerichteter Graph**  $G$  ist ein Paar  $G = (V, E)$ , bestehend aus einer endlichen Menge  $V$  und einer Menge  $E \subseteq V \times V$ .

Die Elemente von  $V$  heißen **Knoten**, die Elemente von  $E$  **gerichtete Kanten**.

Die hier verwendete Bedeutung des Begriffs „Graph“ darf nicht mit der eines Funktionsgraphen verwechselt werden. Wir können uns eine gerichtete Kante  $(u, v) \in E$  so vorstellen, dass diese den **Startknoten**  $u \in V$  mit dem **Endknoten**  $v \in V$  verbindet.

Die Interpretation der Kanten als gerichtete Verbindungen zwischen Knoten ermöglicht die Darstellung eines gerichteten Graphen als ein **Knoten-Kanten-Diagramm** wie das folgende:



**Abb. 1.2.** Knoten-Kanten-Diagramm

Die Abbildung zeigt den gerichteten Graph  $G = (V, E)$  mit Knotenmenge  $V = \{a, \dots, f\}$  und den Kanten  $E = \{(a, b), (b, a), (b, c), (c, d), (d, b), (e, e)\}$ .

Ein **ungerichteter Graph** unterscheidet sich von einem gerichteten im Wesentlichen dadurch, dass die Kanten nicht gerichtet sind, zwischen Start- und Endknoten also nicht unterschieden wird.

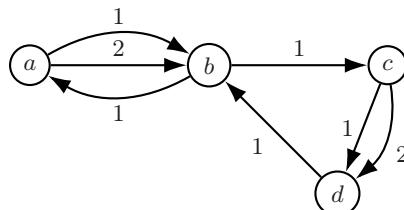
In der Praxis der Datenmodellierung sind mit „Graphen“ eigentlich Multigraphen gemeint, bei diesen können Knoten durch mehr als nur eine Kante miteinander verbunden sein.

Ein **gerichteter Multigraph**  $G = (V, E)$  besteht aus einer endlichen Menge von Knoten  $V$  und einer Menge durchnummerierter gerichteter Kanten

$$E \subseteq \{(u, v, n) | u, v \in V, n \in \{1, 2, 3, \dots\}\}.$$

„Durchnummeriert“ bedeutet genauer: Für alle  $u, v \in V$  und  $n \in \mathbb{N}, n > 1$  gilt: Wenn  $(u, v, n)$  eine Kante ist, dann ist auch  $(u, v, n - 1)$  eine Kante.

Anstelle von  $(u, v, n)$  für eine nummerierte Kante können wir auch  $(u, v)_n$  schreiben, um die Nummer deutlicher von Start- und Endknoten in der Notation abzugrenzen. Wir illustrieren das Konzept anhand eines konkreten Beispiels:



**Abb. 1.3.** Multigraph

Die Abbildung zeigt ein Knoten-Kanten-Diagramm des Multigraphen mit den Knoten  $V = \{a, \dots, e\}$  und den Kanten:

$$E = \{(a, b)_1, (a, b)_2, (b, a)_1, (b, c)_1, (c, d)_1, (c, d)_2, (d, b)_1\}$$

Ein Entity-Relationship-Modell kann wie folgt in eine logische Struktur überführt werden, die **Property-Graph-Modell** genannt wird [6, Kap. 3]:

**Entitäten:** Jede Entität entspricht einem Knoten. Entitätstypen werden durch sogenannte **Labels** unterschieden. So trüge zum Beispiel ein Knoten, der eine Person repräsentiert, das Label „Person“. Einem Knoten können grundsätzlich mehrere Labels zugeordnet werden.

**Attribute:** Diese werden mit Knoten oder Kanten in Form von Schlüssel-Wert-Paaren hinterlegt, den **Properties**. Eine Property wäre beispielsweise ("Vorname" : "Karl").

**Beziehungen:** Jede Beziehung zwischen zwei Entitäten entspricht einer Kante zwischen den entsprechenden Knoten. Beziehungstypen werden durch Labels voneinander unterschieden.

Während Relationen in Form von Tabellen für den Menschen lesbar dargestellt werden, eignet sich für die Darstellung eines Property-Graphen ein Knoten-Kanten-Diagramm wie das folgende:

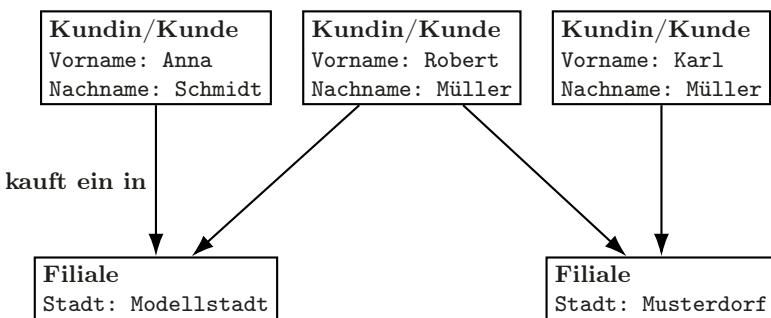


Abb. 1.4. *Property-Graph*

Ein Beispiel für ein graphbasiertes Datenbankverwaltungssystem ist neo4j [7]. Eine Abfrage mit der Datenbanksprache Cypher in neo4j könnte wie folgt aussehen:

```

MATCH (K:Kunde) - [:kaufte_ein_in] -> (F:Filiale)
WHERE F.Stadt = 'Modellstadt'
RETURN K.Name, K.Stadt
  
```

### 1.2.3 Hierarchische Datenmodelle

Hierarchische Datenmodelle bauen auf dem mathematischen Konzept des Baums auf. Bäume sind Graphen von spezieller Form: Ausgehend von einem bestimmten Knoten – der „Wurzel“ – gibt es nur eine Richtung, in der wir entlang der Kanten gedanklich laufen können, hin zu den „Blättern“. Bevor wir Bäume formal definieren können, müssen wir erst noch ein paar weitere Begriffsbestimmungen aus der Graphentheorie durchführen.

Sei  $G = (V, E)$  ein gerichteter Graph. Ein **gerichteter Kantenzug** mit Startknoten  $u \in V$  und Endknoten  $w \in V$  der Länge  $N \geq 1$  ist eine alternierende Folge von Knoten und Kanten  $u = v_0, e_1, v_1, e_2, \dots, v_{N-1}, e_N, v_N = w$ , sodass  $e_n = (v_{n-1}, v_n)$  für alle  $n \in \{1, \dots, N\}$ .

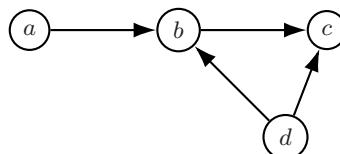
Bei einem **ungerichteten Kantenzug** kann  $e_n = (v_{n-1}, v_n)$  oder  $e_n = (v_n, v_{n-1})$  gelten, d. h., ein ungerichteter Kantenzug kann (ganz oder in Teilen) auch entgegen der Richtung von Kanten verlaufen.

Ein (**un-**)**gerichteter Pfad** ist ein (**un-**)gerichteter Kantenzug, bei dem alle Wegknoten  $u, v_1, \dots, v_N, w$  paarweise verschieden sind.

Ein (**un-**)**gerichteter Kreis** ist ein (**un-**)gerichteter Kantenzug, bei dem Start- und Endknoten identisch sind,  $u = w$ , alle übrigen Wegknoten  $v_1, \dots, v_N$  jedoch paarweise verschieden.

Graphen, die keinen gerichteten Kreis enthalten, werden als **azyklisch** bezeichnet.

Der unten abgebildete Graph ist ein Beispiel für einen azyklischen Graphen: Er enthält keinen gerichteten Kreis, es gibt also keine geschlossene Bahn mit gleichem Start- wie Endknoten, der in Richtung der Pfeile durchlaufen werden könnte. Der Graph enthält jedoch einen ungerichteten Kreis mit Knotenfolge  $b, c, d, b$ .



**Abb. 1.5.** Azyklischer Graph mit ungerichtetem Kreis

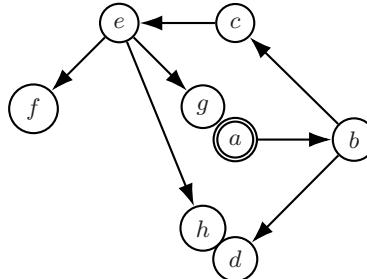
Azyklische Graphen werden uns im Abschn. 6.4 der Beschreibung von künstlichen neuronalen Netzwerken dienen.

Schließlich können wir die Struktur definieren, die hierarchisch strukturierten Daten zugrundeliegt.

Ein **gewurzelter Baum** ist ein gerichteter Graph  $G = (V, E)$ , der keinen ungerichteten Kreis enthält, zusammen mit einem ausgezeichneten Knoten  $r \in V$ , der **Wurzel** von  $G$ , sodass für alle  $v \in V$  mit  $v \neq r$  ein gerichteter Pfad von  $r$  nach  $v$  existiert.

Die **Blätter** eines Baums sind jene von der Wurzel verschiedene Knoten, die nur einen Nachbarn haben.

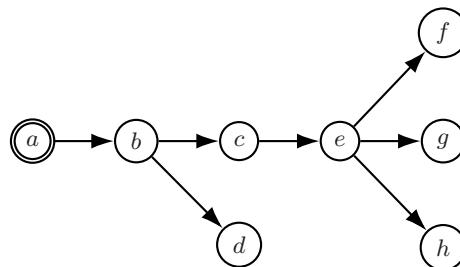
Das folgende Knoten-Kanten-Diagramm zeigt ein Beispiel für einen Baum mit Wurzel  $a$ :



**Abb. 1.6.** *Gewurzelter Baum*

Jedem Knoten kann durch die (eindeutig bestimmte) Länge des Pfades vom Wurzelknoten aus eine Hierarchieebene zugeordnet werden: auf der „nullten“ Ebene befindet sich nur der Wurzelknoten  $a$ , auf der ersten Ebene der Knoten  $b$ , die dritte Ebene besteht aus den Knoten  $c$  und  $d$  usw. Die Blätter durch die Knoten  $d$ ,  $f$  und  $h$  gegeben.

Das folgende Diagramm zeigt denselben Baum noch einmal – in dieser Darstellung tritt die hierarchische Struktur jedoch deutlicher zutage, da Knoten in derselben Hierarchieebene auf gleicher Höhe abgebildet sind:



**Abb. 1.7.** *Hierarchieebenen eines gewurzelten Baums*

Bäume dienen auch der Darstellung der Ergebnisse des Verfahrens der hierarchischen Clusteranalyse, welches wir in Abschn. 7.3.2 behandeln werden.

**Anwendungsbeispiel.** Dies ist ein kleiner Ausschnitt einer sogenannten XML-Datei (Extensible Markup Language), die durch einen RSS-Nachrichtenfeed übermittelt wurde [8]:

```
<?xml version="1.0" encoding="UTF-8"?>
<rss xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:content="http://purl.org/rss/1.0/modules/content/"
      xmlns:atom="http://www.w3.org/2005/Atom" version="2.0"
      xmlns:cc="http://cyber.law.harvard.edu/rss/
      creativeCommonsRssModule.html">
  <channel>
    <title><![CDATA[Towards Data Science - Medium]]></title>
    <description><![CDATA[A Medium publication sharing
      concepts, ideas, and codes. - Medium]]></description>
    <link>https://towardsdatascience.com?source=rss
    ----7f60cf5620c9---4</link>
    <item>
      <title><![CDATA[Why Psychologists can be great
        Data Scientists]]></title>
      <guid isPermaLink="false">https://medium.com/p/
      970552b5223</guid>
      <category><![CDATA[careers]]></category>
      <category><![CDATA[machine-learning]]></category>
      <dc:creator><![CDATA[Maarten Grootendorst]]>
      </dc:creator>
      <pubDate>Fri, 18 Sep 2020 14:00:02 GMT</pubDate>
      <atom:updated>2020-09-18T14:00:02.568Z</atom:updated>
    </item>
    <item>
      ...
    </item>
  </channel>
</rss>
```

XML-Dateien sind Beispiele für hierarchisch strukturierte Daten. Jeder Knoten des **Strukturbauums** wird in diesem Kontext als **Element** bezeichnet. Jedes Element sowie die mit diesem Element verbundenen Elemente der darunterliegenden Hierarchieebenen werden durch Start- und Endtags der Form `<Elementname>` bzw. `</Elementname>` umschlossen. Die Wurzel wird in diesem Beispiel durch die Tags `<rss>...</rss>` repräsentiert, welche alle weiteren Elemente umschließen.

## 1.3 Datenqualität

Eine zentrale Aufgabe der Datenwissenschaft ist es, belastbares und nützliches Wissen aus Daten zu generieren. Ein qualitativ hochwertiger Datenbestand ist eine Voraussetzung für die Erfüllung dieses Auftrags.

### 1.3.1 Datenqualitätsmerkmale

Die Data Management Association UK schlägt eine Bewertung der inhaltlichen Qualität eines Datenbestands anhand von sechs zentralen Merkmalen vor [9]: **Vollständigkeit, Konsistenz, Gültigkeit, Eindeutigkeit, Richtigkeit** und **Aktualität**. In Tab. 1.6 sind diese Merkmale aufgeführt. Neben einer Erläuterung des jeweiligen Qualitätsmerkmals in Form einer prüfenden Frage finden sich in der Tabelle auch Beispiele für typische **Datenfehler**. Ein **Data-Profiling** dient der Erhebung bestimmter Kennzahlen, welche der quantitativen Bewertung der Datenqualität dienen. In der letzten Spalte der Tabelle sind Ansätze für ein solches Data-Profiling aufgeführt, bezogen auf das Negativbeispiel.

Die **Qualität von Daten** bemisst, inwieweit diese bestimmungsgemäß verwendet werden können. Insbesondere ist bei der Bewertung der Qualität, etwa bei der Gewichtung der verschiedenen Datenqualitätsmerkmale, stets der vorliegende Anwendungsfall zu berücksichtigen.

Beispielsweise ist für eine automatisierte Weiterverarbeitung die Gültigkeit der Daten oft von herausragender Bedeutung, da ungültige Eingaben außerhalb des Definitionsbereichs von Funktionen zu Softwarefehlern führen können, wenn diese nicht durch eine geeignete Ausnahmebehandlung abgefangen werden. Werden die Informationen jedoch nur angezeigt und von einem Menschen beurteilt, so spielt die Gültigkeit der Daten eine ähnlich große Rolle wie deren Richtigkeit.

Ein meist einschlägiges Qualitätsmerkmal besteht in der Vollständigkeit der Metadaten, insbesondere der **Datendokumentation**. Eine gute Datendokumentation stellt Reproduzierbarkeit der Ergebnisse und Transparenz des Analysevorgangs sicher. Diese mag etwa die folgenden Informationen enthalten:

- Kontext der Datenerfassung, Hypothesen und Ziele des Datenanalyseprojekts;
- inhaltliche Bedeutung der Daten, ein konzeptionelles Datenmodell, z. B. in Form eines Entity-Relationship-Diagramms;
- eine Darstellung des logischen Datenmodells und menschenlesbare Übersetzung von Bezeichnern, z. B. im Kontext sozialwissenschaftlicher Umfragen der **Codeplan**, d. h., eine Übersetzung von Variablennamen zu den Fragen;
- Darstellung der Datenqualität, der Ergebnisse eines Data-Profileings, der evtl. vorgenommenen Maßnahmen zur Datenbereinigung (siehe nächster Abschnitt);
- Beschreibung des Ortes, Zeitpunkts und der Modalität der Datenerfassung: verwendete Methoden, Software, Messapparate, erschlossene Datenquellen usw.;

- Informationen zur **Datenversionierung**, Unterschiede zwischen verschiedenen Versionen;
- technisch bedeutsame Informationen, etwa zu Schnittstellen für den Zugang, physische Größe der Dateien;
- Nutzungsbedingungen, vorgenommene Datenschutzmaßnahmen.

## 1.4 Datenbereinigung

Im Folgenden wollen wir ein paar einfache Ansätze zur Identifikation und Behandlung von Datenfehlern vorstellen. Meist empfiehlt es sich, über eine solche **Datenvorverarbeitung** Buch zu führen und die Rohdaten im Datenbestand zu erhalten; eine Rückverfolgung der **Datenherkunft** sollte gewährleistet bleiben.

### 1.4.1 Validierung

Formal bzw. syntaktisch inkorrekte, ungültige Attributwerte können unter Umständen zu Fehlfunktionen der datenverarbeitenden Software führen. Unplausible Werte außerhalb eines zulässigen Wertebereichs sind zudem in der Lage, die Ergebnisse von Datenanalysen zu verzerren.

Ungültige Attributwerte können mithilfe von Validierungsregeln identifiziert werden. Solche Regeln können sich zum einen auf die Syntax beziehen. So könnte etwa gefordert werden, dass ein Geburtsdatum im Format JJJJ-MM-TT abgespeichert wird, dabei steht „JJJJ“ für eine vierstellige Jahreszahl, „MM“ für eine zweistellige Monatszahl (mit ggf. führender Null) und „TT“ für eine zweistellige Tageszahl. Bzgl. dieser Validierungsregel wäre 2010-12-01 ein gültiges Datum, 10-12-01 oder 2010-12-1 wären es jedoch nicht.

Zum anderen können Werte als ungültig erkannt werden, wenn diese außerhalb eines festgelegten Gültigkeitsbereichs liegen. So wäre 2010-23-01 nach obiger Validierungsregel ein gültiges Datum – nicht jedoch, wenn zusätzlich  $1 \leq MM \leq 12$  gefordert würde.

Darüber hinaus können notwendige Bedingungen für die Richtigkeit des Attributwerts geprüft werden, also deren Plausibilität. So könnten Geburtsdaten als ungültig markiert werden, die zum Zeitpunkt der Erstellung des Datensatzes in der Zukunft lagen oder unzulässig weit in der Vergangenheit.

In vielen Fällen besteht die Behandlung ungültiger oder unplausibler Attributwerte in der Entfernung des jeweiligen Werts. In anderen Fällen kann eine Transformation durchgeführt werden, um den Datenfehler zu beheben. Z. B. könnten Datumsangaben der Form JJJJ-MM-D in das gültige Schema JJJJ-MM-DD überführt werden, indem eine führende Null eingefügt würde: JJJJ-MM-0D. Das ist natürlich nur ein empfehlenswertes Vorgehen, wenn sich

aus der Kenntnis der Datenherkunft und des datengenerierenden Prozesses ableiten lässt, dass das Ergebnis dann nicht nur gültig, sondern in der Regel auch richtig ist.

Wir fassen zusammen:

Ungültige oder unplausible Attributwerte können durch Prüfung von **Validierungs- und Plausibilitätsregeln** identifiziert werden. Ungültige Werte können im Rahmen der **Datenvalidierung** entfernt oder in ein gültiges Format überführt werden.

Für eine Implementierung sind oftmals sogenannte **reguläre Ausdrücke** hilfreich, welche weite Verbreitung in Programmiersprachen und Standardbibliotheken finden [10]. Mithilfe regulärer Ausdrücke können syntaktische Muster in Zeichenketten gesucht und ggf. ersetzt werden.

### 1.4.2 Normierung

Zentrale Aufgabe der Datenintegration ist die Zusammenführung von Daten aus verschiedenen Datenquellen. Wesentliche Herausforderungen entstehen dabei durch die Heterogenität der Quellen, etwa hinsichtlich der verwendeten Datenmodelle oder der technischen Umsetzung in der Datenbereitstellung (z. B. verschiedene Dateiformate oder Schnittstellen). Mithin ist eine wichtige Voraussetzung für die erfolgreiche Datenintegration in der Regel eine **Datenharmonisierung**, welche eine Verringerung der vorhandenen Heterogenität zum Ziel hat.

Eine wichtige Teilaufgabe der Harmonisierung eines Datenbestands besteht in der Normierung der Daten, sodass diese in einer einheitlichen Syntax und mit denselben Maßeinheiten vorliegen. Hier ein (fiktives) Beispiel für Daten mit unterschiedlicher Syntax bzw. Maßeinheit:

Name	Rosi Schmidt	Schmidt, Rosi
Telefon-Nr.	+4989-32168	+4989/32 16 8
Geburtsdatum	07.12.1981	1981-12-07
Organisation	Muster GmbH	Muster Gesellschaft mbH
Körpergröße	180 cm	1,80 m

**Tabelle 1.2.** Heterogene, nichtnormierte Daten

Um die Datumsangaben zu normieren, könnte als Zielsyntax zum Beispiel JJJJ-MM-TT gewählt werden, wobei die Buchstaben für Jahres-, Monats- und Tageszahl stehen. Andere Datumsangaben wie etwa solche von der Form TT.MM.JJJJ können in der offensichtlichen Weise auf diese Syntax abgebildet werden.

Ähnlich kann vorgegangen werden, um Angaben für die Körpergröße anzugeleichen: Alle Angaben von Körpergrößen über Datenquellen hinweg würden in dieselbe Maßeinheit, beispielsweise Meter, umgerechnet.

Wir fassen zusammen:

Eine **Datennormierung** hat zum Ziel, eine heterogene Syntax oder einen heterogenen Gebrauch von Maßeinheiten anzugleichen. Zu diesem Zweck wird eine Zielsyntax bzw. gemeinsame Maßeinheit festgelegt, in die alle Werte gleichartiger Attribute aufeinander abgebildet bzw. umgerechnet werden.

### 1.4.3 Imputation

Fehlende Attributwerte können zu einer Verzerrung der Ergebnisse einer Datenanalyse führen. Das Aufkommen fehlender Werte im Datensatz kann vielfältige Gründe haben. Bei einer Umfrage könnte die befragte Person etwa die Auskunft verweigert haben. Im Falle einer physikalischen Messung mag das Messgerät in seiner Funktion beeinträchtigt gewesen sein. Fehlende Werte können auch entstehen, wenn im Rahmen einer Datenvalidierung ungültige Werte verworfen und entfernt wurden.

Eine **Datenimputation** dient der Vervollständigung fehlender Attributwerte im Datensatz durch Einsetzen möglichst plausibler Werte.

Das vorrangige Ziel ist dabei nicht, für jedes Datentupel die korrekten Attributwerte zu ermitteln, die anstelle der fehlenden stehen müssten. Vielmehr geht es um eine Ergänzung mit dem Ziel, eine Verzerrung der Analyseergebnisse zu minimieren.

Eine Alternative zur Imputation und ein einfaches wie häufig eingesetztes Verfahren ist die **Complete-Case-Analyse**, bei der Datentupel mit fehlenden Werten von der Analyse ausgeschlossen werden.

#### Imputation durch Lageparameter

Auch wenn für manche Informationsobjekte einem bestimmten Attribut kein Wert zugeordnet ist, kann anhand der übrigen, vorhandenen Werte desselben Attributs ein Durchschnittswert ermittelt werden. Allgemein bekannt und gebräuchlich ist das arithmetische Mittel für einen solchen Durchschnittswert, im Abschn. 2.3 stellen wir noch andere solche Lageparameter wie z. B. den Modus oder den Median vor. Die fehlenden Attributwerte können durch diesen Lageparameter ersetzt werden, insbesondere bei Verwendung des arithmetischen Mittelwerts wird von **Mittelwertimputation** gesprochen.

**Anwendungsbeispiel.** Die folgende Tabelle ist ein kleiner Ausschnitt der Ergebnisse einer telefonischen Umfrage unter US-Bürgerinnen und -Bürgern:

Geschlecht	Alter in Jahren	Körpergröße in cm	Körpergewicht in kg
männlich	35	170	98
männlich	66	NA	NA
weiblich	67	155	NA
:	:	:	:

**Tabelle 1.3.** Fehlende Attributwerte

Der gesamte Datensatz wird von den US-Gesundheitsbehörden (CDC) zur Verfügung gestellt [11]. Die Attributbelegungen „NA“ stehen dabei für fehlende Daten: Hier hatte die oder der Befragte keine, ungültige oder unzureichende Angaben gemacht.

Die durchschnittliche Körpergröße und das durchschnittliche Körpergewicht, ermittelt anhand der vorhandenen Werte im gesamten Datensatz, sind 170 cm bzw. 82 kg. Der mittelwertimputierte Datensatz wäre folglich:

Geschlecht	Alter in Jahren	Körpergröße in cm	Körpergewicht in kg
männlich	35	170	98
männlich	66	170	82
weiblich	67	155	82

**Tabelle 1.4.** Mittelwertimputation

Der Datensatz kann auch in zweckmäßige **Imputationsklassen** partitioniert werden, auf die das Imputationsverfahren jeweils angewendet wird. In obigem Beispiel könnten anstelle von Mittelwerten über den gesamten Datensatz fehlende Daten auch durch den Mittelwert nach Geschlecht ersetzt werden. In diesem Fall ergibt sich die folgende imputierte Datenmatrix:

Geschlecht	Alter in Jahren	Körpergröße in cm	Körpergewicht in kg
männlich	35	170	98
männlich	66	178	90
weiblich	67	155	75

**Tabelle 1.5.** Verwendung von Imputationsklassen

### Imputation mittels Regressions- und Klassifikationsverfahren

In den Abschnitten zur Inferenzstatistik und zum maschinellen Lernen (Kap. 4 und 6) werden wir verschiedene Verfahren vorstellen, mit denen ein unbekannter, fehlender Wert eines Attributs (in diesem Zusammenhang Zielgröße genannt) durch die übrigen Attributwerte derselben Entität (den Merkmalen

oder Einflussgrößen) vorhergesagt werden kann. Offenbar können solche Verfahren auch für eine Datenimputation herangezogen werden. Tatsächlich kann die oben erklärte Mittelwertimputation als ein primitives Regressionsverfahren aufgefasst werden. Sogenannte **Hot-Deck-Verfahren** ersetzen die fehlenden Attributwerte durch entsprechende Werte ansonsten ähnlicher Informationsobjekte desselben Datensatzes. Auch hierbei handelt es sich letztlich um ein Regressionsverfahren (ein Nächste-Nachbarn-Verfahren, siehe Abschn. 6.3.2 und vgl. etwa [12]).

#### 1.4.4 Augmentation

In manchen Fällen kann es sinnvoll oder erforderlich sein, den Datenbestand um neue Datensätze zu erweitern, da dessen Abdeckung für den jeweiligen Anwendungsfall nicht ausreichend ist. Eine solche Erweiterung kann grundsätzlich durch erneute Datenerhebung oder die Erschließung neuer Datenquellen erreicht werden.

Insbesondere im Bereich des maschinellen Lernens werden aber auch **synthetische Daten** verwendet, um mehr Daten für das Training der Algorithmen zur Verfügung zu stellen. Auf diese Weise soll auch einer sogenannten Überanpassung (siehe Abschn. 6.1.2) entgegengewirkt werden.

Eine **Datenaugmentation** bezeichnet eine Erweiterung des Datenbestands um künstlich erzeugte Datensätze. In der Regel werden diese Datensätze durch geeignete Transformation vorhandener Datensätze erzeugt.

Ähnlich wie bei der Imputation geht es bei der Augmentation nicht um eine Ergänzung um korrekte Daten. Vielmehr soll eine Verzerrung der Analyseergebnisse aufgrund von „Datenmangel“ vermieden bzw. die Güte eines Verfahrens für maschinelles Lernen erhöht werden.

- In der Bild- oder Videoanalyse können die vorhandenen Bilder diversen Transformationen der Bildverarbeitung unterzogen werden, um neue Daten zu generieren [13]: An geometrischen Transformationen können diese eine Scherung, einen Zoom, eine Spiegelung oder eine Drehung darstellen. Weiterhin können Bildschärfe, Kontrast, Helligkeit oder Farbtemperatur verändert werden, um zusätzliche Bilder zu erzeugen.
- In der Textanalyse können Wörter mithilfe eines Thesaurus durch ihre Synonyme ersetzt werden [14].

#### 1.4.5 Deduplikation

Eine besondere Herausforderung für die Datenharmonisierung stellen **unstrukturierte Daten** dar, die keiner festen Syntax folgen. So könnte etwa das Datum händisch in ein freies Textfeld eingegeben worden sein, sodass die Zeichenketten „07. Dez. 1981“, „7. Dezember 81“, „07.12.1981“ usw. sämtlich für dasselbe

Datum stehen. In diesem Fall müssen komplexere Regeln für eine Extraktion von Monat, Tag und Jahr entwickelt werden.

Insbesondere bei Organisationsnamen ist es nicht immer einfach, diese auf eine normierte Form zu bringen. Die folgende Liste stammt aus der Patentdatenbank PATSTAT [15, 16] und zeigt eine kleine Auswahl von Schreibweisen, mit denen der Technologiekonzern Siemens AG als Anmelder eines Patents auftreten kann:

SIEMENES AKTIENGESELLSCHAFT; Siemens; Siemens A.G.; Siemens AG;  
 SIEMENS AG (DE); SIEMENS AG, 8000 MUENCHEN, DE;  
 SIEMENS AG, 80333 MUENCHEN, DE; Siemens Akteingesellschaft;  
 Siemens Aktiengellschaft; Siemens Aktiengesellscahft;  
 Siemens Aktiengesesellschaft; SIMENS AKTSIENGEZELL'SHAFT

Die Gründe für die Heterogenität sind vielfältig: Tippfehler, abkürzende Schreibweisen, dem Namen hinzugefügte Adressangaben oder eine Transkription.

Selbst die Anzahl der von der Siemens AG angemeldeten Patente zu ermitteln wird so zu einer nichttrivialen Aufgabe, da die Organisation nicht in eindeutiger Weise repräsentiert ist.

Eine **Datendeduplikation** dient dem Auffinden der Datensätze, welche jeweils eine Entität beschreiben, mit dem Ziel, diese zu einem einzelnen repräsentativen Datensatz zu fusionieren.

Im Kern handelt es sich bei vielen gängigen automatisierten Verfahren der Duplikaterkennung und -fusion um eine sogenannte Clusteranalyse: Es sollen hinreichend ähnliche Informationsobjekte gruppiert werden. Mehr zum Thema Clusteranalyse kann im Abschn. 7.3 in Erfahrung gebracht werden. Neben dem Begriff der Datendeduplikation sind auch die Bezeichnungen **Objektidentifikation**, **Entity-Resolution** oder **Record-Linkage** geläufig.

### **Abstands- und Ähnlichkeitsmaße für Zeichenketten**

Ein Ansatz der Datendeduplikation besteht in der Fusionierung von Datensätzen, die in geeignetem Sinn eine hinreichend große Ähnlichkeit bzw. einen hinreichend kleinen „Abstand“ aufweisen. Im Abschn. 5.2 werden wir ein paar Abstands- und Ähnlichkeitsmaße vorstellen, mit denen Datentupel verglichen werden können. An dieser Stelle möchten wir dem Thema vorwiegend greifen und derartige Kennzahlen für den Vergleich von Zeichenketten bzw. Symbolfolgen vorstellen, die sich insbesondere für die Erkennung von Duplikaten in Listen von Organisations- oder Personennamen eignen können. Kennzahlen dieser Art finden auch in anderen Kontexten Anwendung, etwa im Information Retrieval für **unscharfe Suchen** (auch Fuzzy-Suchen genannt) oder in der Bioinformatik für den Vergleich von DNA-Sequenzen [17].

Eine sehr einfache, aber auch recht grobe Definition für den Abstand zwischen zwei Zeichenketten  $a$  und  $b$  ist die **diskrete Metrik**:

$$\delta(a, b) = \begin{cases} 0 & \text{falls } a = b \\ 1 & \text{falls } a \neq b \end{cases}$$

Diese Abstandsdefinition ist insofern universell, als dass sie auch auf andere Datentypen angewendet werden kann. Sie lässt jedoch offensichtlich keinen graduellen Vergleich zwischen den Objekten zu: Die Zeichenketten  $a = \text{SIEMENS}$  und  $b = \text{IEMENS}$  sind unter der diskreten Metrik „ebenso verschieden“ wie die Zeichenketten  $a = \text{SIEMENS}$  und  $c = \text{GRUNDIG}$ . Intuitiv scheint aber klar, dass die Zeichenketten  $a$  und  $b$  eine größere Ähnlichkeit aufweisen als die Zeichenketten  $a$  und  $c$ . Ein sinnvolles Abstandsmaß sollte diesen Umstand numerisch widerspiegeln.

Um solche weniger groben Abstandsmaße erklären zu können, halten wir zunächst fest, auf welche Weise Zeichenketten ineinander umgewandelt werden können.

Auf Zeichenketten können folgende elementare **Editieroperationen** angewendet werden:

**Löschen eines Zeichens:** SIEMENS  $\mapsto$  SIMENS

**Einfügen eines Zeichens:** SIEMENS  $\mapsto$  SIEMENES

**Ersetzen eines Zeichens:** SIEMENS  $\mapsto$  SOEMENS

**Vertauschen zweier benachbarter Zeichen:** SIEMENS  $\mapsto$  SEIMENS

Eine gegebene Zeichenkette kann durch Hintereinanderausführung geeigneter elementarer Editieroperationen in jede beliebige andere Zeichenkette umgewandelt werden. Die Grundidee von **Editierabstandsmaßen** besteht darin, die Anzahl der elementaren Operationen zu ermitteln, die notwendig sind, um zwei Zeichenketten ineinander zu überführen. Je mehr Operationen notwendig sind, als um so verschiedener können die Zeichenketten erachtet werden.

Seien  $a$  und  $b$  Zeichenketten.

Die **Levenshtein-Distanz**<sup>1</sup> zwischen  $a$  und  $b$  ist durch die Mindestanzahl an Einfüge-, Lösch- und Ersetzoperationen gegeben, die notwendig sind, um  $a$  in  $b$  zu überführen.

Die **Damerau-Levenshtein-Distanz** zwischen  $a$  und  $b$  ist durch die Mindestanzahl an Vertausch-, Einfüge-, Lösch- und Ersetzoperationen gegeben, die notwendig sind, um  $a$  in  $b$  zu überführen.

Die Levenshtein-Distanz zwischen  $a = \text{SIEMENS}$  und  $b = \text{SOEMEN}$  ist zum Beispiel zwei, da wenigstens eine Ersetz- und eine Löschoperation notwendig ist,

---

<sup>1</sup> Nach Wladimir Lewenstein (\* 1935 – † 2017), „Levenshtein“ ist eine englische Transkription des Namens.

um  $b$  aus  $a$  zu erhalten. Umgekehrt ist wenigstens eine Ersetz- und eine Einfügeoperation notwendig, um  $a$  aus  $b$  zu erhalten.

Wir bezeichnen die Längen der Zeichenketten mit  $|a|$  bzw.  $|b|$ , also z. B.  $|\text{xyYz}| = 4$ . Sei weiterhin  $\text{lev}_{i,j}(a, b)$  die Levenshtein-Distanz zwischen den ersten  $i$  Zeichen von  $a$  und den ersten  $j$  Zeichen von  $b$ . Für die konkrete Berechnung der vollständigen Levenshtein-Distanz  $\text{lev}(a, b) = \text{lev}_{|a|, |b|}(a, b)$  kann das folgende rekursive Rechenschema herangezogen werden [18, Theorem 2]:

$$\text{lev}_{i,j}(a, b) = \begin{cases} \max(i, j) & \text{falls } \min(i, j) = 0, \\ \min \begin{cases} \text{lev}_{i-1,j}(a, b) + 1 \\ \text{lev}_{i,j-1}(a, b) + 1 \\ \text{lev}_{i-1,j-1}(a, b) + \delta(a_i, b_j) \end{cases} & \text{sonst} \end{cases}$$

Dabei ist  $a_i$  das Zeichen an der  $i$ -ten Stelle von  $a$  und  $b_j$  das Zeichen an der  $j$ -ten Stelle von  $b$  sowie

$$\delta(a_i, b_j) = \begin{cases} 0 & \text{falls } a_i = b_j \\ 1 & \text{falls } a_i \neq b_j \end{cases}$$

Es gilt  $0 \leq \text{lev}(a, b) \leq \max\{|a|, |b|\}$ , sodass eine **normierte Levenshtein-Distanz** (mit höchstens dem Wert eins) wie folgt definiert werden kann:

$$\text{lev}_{\text{norm}}(a, b) = \frac{\text{lev}(a, b)}{\max\{|a|, |b|\}}$$

Die Kennzahl  $1 - \text{lev}_{\text{norm}}(a, b)$  kann dann als Maß für die Ähnlichkeit beider Zeichenketten verstanden werden: Identische Zeichenketten haben eine maximal mögliche Levenshtein-Ähnlichkeit von eins, „maximal verschiedene“ Zeichenketten haben ein verschwindendes Ähnlichkeitsmaß.

Die Damerau-Levenshtein-Distanz ist stets höchstens so groß wie die Levenshtein-Distanz, da jede Vertauschoperation das Produkt von zwei Ersetzoperationen ist, z. B. **SIEMENS**  $\mapsto$  **SIIMENS**  $\mapsto$  **SEIMENS**.

Während die Levenshtein-Distanz bzw. -Ähnlichkeit für eine Vielzahl verschiedener Anwendungen eingesetzt wird, sind die im Folgenden beschriebenen Ähnlichkeitsmaße insbesondere für den Vergleich von Personennamen entwickelt worden [19].

Für zwei Zeichenketten  $a$  und  $b$  der Länge  $|a|$  bzw.  $|b|$  führen wir zunächst folgende Konstruktion durch: Es ist  $a \cap b = a_{i_1} \cdots a_{i_m}$  mit  $i_1 < \cdots < i_m$  die Zeichenkette, die durch schrittweise, vom Beginn der Zeichenkette ausgehende Auswahl von Zeichen  $a_i$  aus  $a$  hervorgeht, für die es ein korrespondierendes Zeichen  $b_j$  mit  $a_i = b_j$  und  $|j - i| \leq \frac{1}{2} \max\{|a|, |b|\} - 1$  gibt. Bei Wiederholungen von Zeichen innerhalb einer Zeichenkette ist eine eindeutige Korrespondenz durch vom Beginn der Zeichenkette ausgehende Auswahl der  $b_j$  herzustellen.

Analog definieren wir  $b \cap a$ . Notwendigerweise besteht  $b \cap a$  aus denselben Zeichen wie  $a \cap b$ : jene sowohl in  $a$  als auch  $b$  enthaltenen Zeichen, die nicht zu großen Abstand voneinander haben. Im Allgemeinen können sich  $a \cap b$  und  $b \cap a$  jedoch in der Reihenfolge der Zeichen unterscheiden.

Für zwei Zeichenketten  $a$  und  $b$  sei  $m = |a \cap b| = |b \cap a|$ , und  $t$  sei die Mindestanzahl an Vertauschungen nicht notwendigerweise benachbarter Zeichen, die erforderlich ist, um  $a \cap b$  in  $b \cap a$  überzuführen.

Die **Jaro-Ähnlichkeit** zwischen  $a$  und  $b$  ist dann wie folgt gegeben:

$$\text{jaro}(a, b) = \begin{cases} 0 & \text{falls } m = 0 \\ \frac{1}{3} \cdot \left( \frac{m}{|a|} + \frac{m}{|b|} + \frac{m-t}{m} \right) & \text{sonst} \end{cases}$$

Wir wollen als konkretes Rechenbeispiel die Jaro-Ähnlichkeit zwischen den Zeichenketten  $a = \text{SIEMENS}$  und  $b = \text{SIMESN}$  ermitteln. Es gilt  $\frac{1}{2} \max\{|a|, |b|\} - 1 = \frac{1}{2} \max\{7, 6\} - 1 = 2,5$ , es werden also nur Zeichen mit einem Abstand von höchstens zwei in Korrespondenz gesetzt. Es ergeben sich  $a \cap b = \text{SIEMNS}$  und  $b \cap a = \text{SIMESN}$ , folglich  $m = 6$  und  $t = 2$ . Schließlich:

$$\text{jaro}(a, b) = \frac{1}{3} \cdot \left( \frac{6}{6} + \frac{6}{7} + \frac{6-2}{6} \right) \approx 0,84$$

Seien  $a$  und  $b$  Zeichenketten der Länge  $|a|$  bzw.  $|b|$ , und  $L \in \mathbb{N}, p \in \mathbb{R}$  mit  $L \geq 1$  und  $0 \leq p \leq \frac{1}{L}$  sind fest gewählte Parameter. Für den Fall  $L = 0$  ist die Wahl von  $p$  beliebig.

Sei  $l = l(a, b) \in \mathbb{N}$  die größte Zahl, sodass  $l \leq \min\{|a|, |b|\}$  gilt und die ersten  $l$  Zeichen von  $a$  und  $b$  als Zeichenketten übereinstimmen:  $a_1 \dots a_l = b_1 \dots b_l$ .

Dann ist die **Jaro-Winkler-Ähnlichkeit** wie folgt gegeben:

$$\text{jw}_{p,L}(a, b) = \text{jaro}(a, b) + l(a, b) \cdot p \cdot (1 - \text{jaro}(a, b))$$

Typische Werte für die Parameter der Jaro-Winkler-Ähnlichkeit sind  $L = 4$  und  $p = 0,1$ . Die Grundidee der Winkler-Korrektur besteht darin, dem Abgleich der ersten  $L$  Zeichen ein größeres Gewicht zu geben. Beispielsweise sind die Jaro-Ähnlichkeiten zwischen  $a = \text{SIEMENS}$  und  $b = \text{SIEMENX}$  bzw.  $c = \text{SXEMENS}$  identisch:  $\text{jaro}(a, b) = \text{jaro}(a, c) \approx 0,90$ . Die Jaro-Winkler-Ähnlichkeiten sind jedoch verschieden, da  $l(a, b) = 4$  und  $l(a, c) = 1$ :

$$\begin{aligned} \text{jw}_{0,1;4}(a, b) &= 0,90 + 4 \cdot 0,1 \cdot (1 - 0,90) = 0,94 \\ \text{jw}_{0,1;4}(a, c) &= 0,90 + 1 \cdot 0,1 \cdot (1 - 0,90) = 0,91 \end{aligned}$$

Wir skizzieren schließlich, wie Abstandsmaße zwischen Zeichenketten für eine Deduplikation genutzt werden können und ziehen hierzu wieder das Beispiel

einer Liste von Firmennamen heran. Zunächst einmal können Bezeichner der Gesellschaftsform wie „GmbH“ oder „Aktiengesellschaft“ entfernt oder normiert werden, etwa mittels regulärer Ausdrücke unter Berücksichtigung häufig vor kommender Tipp- und Rechtschreibfehler. Diese Vorverarbeitung mag auf eine Liste von Organisationsnamen wie die folgende führen:

**SIEMENES; Siemens; SIMENS; Bosch; BOSH**

Wir gehen des Weiteren davon aus, dass Groß- oder Kleinschreibung keine Rolle spielt und ersetzen alle Zeichen durch die entsprechenden großgeschriebenen Buchstaben.

Ein paarweiser Vergleich der auf diese Weise normierten Zeichenketten mittels der normierten Levenshtein-Ähnlichkeit führt auf die folgende Matrix:

	<b>SIEMENES</b>	<b>SIEMENS</b>	<b>SIMENS</b>	<b>BOSCH</b>	<b>BOSH</b>
<b>SIEMENES</b>	1,00	0,88	0,75	0,00	0,00
<b>SIEMENS</b>	0,88	1,00	0,86	0,00	0,00
<b>SIMENS</b>	0,75	0,86	1,00	0,00	0,00
<b>BOSCH</b>	0,00	0,00	0,00	1,00	0,80
<b>BOSH</b>	0,00	0,00	0,00	0,80	1,00

Würde etwa die Regel eingeführt, dass alle Entitäten mit einer Levenshtein-Ähnlichkeit von wenigstens 0,80 zusammengeführt werden, so erhielten wir die Paarungen (**SIEMENES**, **SIEMENS**), (**SIEMENS**, **SIMENS**) und (**BOSCH**, **BOSH**). Die entsprechenden Datensätze würden gemäß einer solchen Regel zusammengeführt.

Auf diese Weise würden – über deren Ähnlichkeit zu **SIEMENS** – auch **SIEMENES** und **SIMENS** zusammengeführt, obgleich diese Paarung der Regel nicht genügt. Dieser Umstand kann aber auch genutzt werden, um einen repräsentativen Wert auszuwählen, etwa durch Bestimmen eines **Medoids** in der Gruppe  $S = \{\text{SIEMENES}, \text{SIEMENS}, \text{SIMENS}\}$ . Ein Medoid ist ein Objekt  $s \in S$ , dass die Summe über die Ähnlichkeiten maximiert bzw. die Summe über die Abstände  $\sum_{t \in S} \text{lev}_{\text{norm}}(s, t)$  minimiert, siehe dazu auch Abschn. 5.3.1. Im vorliegenden Fall ist der Medoid durch den (korrekten) Wert **SIEMENS** gegeben.

Neben den Namen können weitere im Datensatz enthaltene Informationen für eine Deduplikation herangezogen werden, wie etwa der Hauptfirmensitz oder das Gründungsjahr, falls vorhanden. Einen allgemeinen methodischen Rahmen für die Datendeduplikation mit Mitteln der Statistik ist das 1969 formulierte **Fellegi-Sunter-Modell** [20, 21]. Dieses Modell ist äquivalent zum Verfahren der naiven Bayes-Klassifikation, das wir in Abschn. 6.3.3 besprechen, und es werden heute vermehrt auch andere Verfahren des maschinellen Lernens für ein Record-Linkage eingesetzt [22].

	<b>Prüfung</b>	<b>Beispiel für Datenfehler</b>	<b>Data-Profiling</b>
<b>Vollständigkeit</b>	Wurden alle relevanten Daten erhoben und gespeichert?	Bei einem Teil der Personen in der Kundendatenbank fehlt eine Adressangabe.	prozentualer Anteil der Personen mit fehlender Adressangabe
<b>Konsistenz</b>	Sind redundant gespeicherte Daten konsistent, oder sind diese widersprüchlich?	Die am Filialort gespeicherten Kundenadressen unterscheiden sich von jenen an einem zentralen Speicherort.	prozentualer Anteil von Personen mit abweichenden Adressen
<b>Gültigkeit</b>	Sind die Daten formal korrekt? D.h., sind die Daten syntaktisch korrekt, vom korrekten Typ, innerhalb des korrekten Wertebereichs?	Die Kundenadressen enthalten Postleitzahlen mit ungültigen Zeichen, etwa Buchstaben: „123Q5“.	<p>prozentualer Anteil der Personen mit ungültiger Postleitzahl:</p> <ul style="list-style-type: none"> <li>• durch Abgleich mit einer vollständigen und korrekten Liste von Postleitzahlen oder</li> <li>• durch Abgleich mit einem festen Muster, beispielsweise „fünfstellig, nur Ziffern“</li> </ul>
<b>Eindeutigkeit</b>	Ist es möglich, für alle Entitäten in eindeutiger Weise die ihr zugehörigen Daten zu identifizieren?	Derselben Person wurden mehrere verschiedene Schlüsselattribute zugewiesen.	Identifikation potenzieller Duplikate, siehe Abschn. 1.4.5; prozentualer Anteil der Personen mit vorhandenen Duplikaten
<b>Richtigkeit</b>	Geben die Daten den wahren Stand der Informationen wieder, oder befinden sie den Name oder Adresse fehlerhaft erheben im Widerspruch mit der Realität?	Für einige Kundinnen oder Kunden wurden Name oder Adresse fehlerhaft erheben.	Händische Prüfung der Richtigkeit der Daten in einer Stichprobe, etwa durch telefonische Umfrage; prozentualer Anteil der Personen mit fehlerhaften Angaben
<b>Aktualität</b>	Sind die Daten aktuell oder veraltet? D.h., geben die Daten den aktuellen Stand der Informationen über die Entitäten wieder?	Ein Teil der Personen ist verzogen, die Adressen wurden jedoch nicht aktualisiert.	Anteil von Personen, deren Adressdaten seit geraumer Zeit nicht aktualisiert wurden

Tabelle 1.6. Datenqualitätsmerkmale

## Quellen

- [1] „Interim Report: ANSI/X3/SPARC Study Group on Data Base Management Systems 75-02-08“. In: *Bulletin of ACM SIGMOD* 7.2 (1975). Hrsg. von Thomas B. Steel, Jr.
- [2] Peter Pin-Shan Chen. „The entity-relationship model—toward a unified view of data“. In: *ACM Transactions on Database Systems* 1.1 (März 1976), S. 9–36. DOI: [10.1145/320434.320440](https://doi.org/10.1145/320434.320440).
- [3] Michael Widenius, Davis Axmark und Paul DuBois. *MySQL Reference Manual*. 1. Aufl. Sebastopol, USA: O'Reilly & Associates, 2002. ISBN: 0596002653.
- [4] MariaDB Corporation. *MariaDB Server Documentation*. Aufgerufen am 29. Aug. 2021. URL: <https://mariadb.com/kb/en/documentation/>.
- [5] PostgreSQL Development Team. *PostgreSQL Documentation*. Aufgerufen am 10. Juli 2020. URL: <https://www.postgresql.org/docs/>.
- [6] Ian Robinson, Jim Webber und Emil Eifrem. *Graph Databases*. 2. Aufl. Sebastopol, USA: O'Reilly, 2015.
- [7] Neo4j Team. *The Neo4j Operations Manual v4.1*. Aufgerufen am 11. Juli 2020. URL: <https://neo4j.com/docs/pdf/neo4j-operations-manual-4.1.pdf>.
- [8] *RSS-Feed von towards data science*. Aufgerufen am 18. Sep. 2020. URL: <https://towardsdatascience.com/feed>.
- [9] Nicola Askham u. a. *The Six Primary Dimensions for Data Quality Assessment*. Techn. Ber. Data Management Association UK, Okt. 2013.
- [10] Jeffrey E. F. Friedl. *Reguläre Ausdrücke*. 3. Aufl. O'Reilly, Okt. 2012. ISBN: 978-3-897-21720-1.
- [11] CDC Population Health Surveillance Branch. *Behavioral Risk Factor Surveillance System (BRFSS) Survey Data 2018*. Aufgerufen am 01. Feb. 2020. URL: <https://www.cdc.gov/brfss/>.
- [12] Dieter William Hermann Joenssen. „Hot-Deck-Verfahren zur Imputation fehlender Daten: Auswirkungen des Donor-Limits“. Diss. Technische Universität Ilmenau, 2015.
- [13] Agnieszka Mikołajczyk und Michał Grochowski. „Data augmentation for improving deep learning in image classification problem“. In: *2018 International Interdisciplinary PhD Workshop (IIPhDW)*. IEEE, Mai 2018. DOI: [10.1109/iiphdw.2018.8388338](https://doi.org/10.1109/iiphdw.2018.8388338).
- [14] Xiang Zhang, Junbo Zhao und Yann LeCun. „Character-Level Convolutional Networks for Text Classification“. In: *28th International Conference on Neural Information Processing Systems, Montreal, Canada*. Bd. 1. NIPS'15. MIT Press, 2015, S. 649–657.
- [15] Gaétan de Rassenfosse, Hélène Dernis und Geert Boedt. „An Introduction to the Patstat Database with Example Queries“. In: *Australian Economic Review* 47.3 (Sep. 2014), S. 395–408. DOI: [10.1111/1467-8462.12073](https://doi.org/10.1111/1467-8462.12073).
- [16] Europäisches Patentamt. *PATSTAT Global - Ausgabe Herbst 2018 (Mustertaten)*. Aufgerufen am 20. Sep. 2020. URL: [https://www.epo.org/searching-for-patents/business/patstat\\_de.html](https://www.epo.org/searching-for-patents/business/patstat_de.html).

- [17] Bonnie Berger, Michael S. Waterman und Yun William Yu. „Levenshtein Distance, Sequence Comparison and Biological Database Search“. In: *IEEE Transactions on Information Theory* (2020). Early Access. DOI: [10.1109/tit.2020.2996543](https://doi.org/10.1109/tit.2020.2996543).
- [18] Robert A. Wagner und Michael J. Fischer. „The String-to-String Correction Problem“. In: *Journal of the ACM* 21.1 (Jan. 1974), S. 168–173. doi: [10.1145/321796.321811](https://doi.org/10.1145/321796.321811).
- [19] William E. Winkler. *Overview of Record Linkage and Current Research Directions*. Techn. Ber. Washington: U.S. Census Bureau, Statistical Research Division, 2006.
- [20] H. B. Newcombe u. a. „Automatic Linkage of Vital Records: Computers can be used to extract 'follow-up' statistics of families from files of routine records“. In: *Science* 130.3381 (Okt. 1959), S. 954–959. doi: [10.1126/science.130.3381.954](https://doi.org/10.1126/science.130.3381.954).
- [21] Ivan P. Fellegi und Alan B. Sunter. „A Theory for Record Linkage“. In: *Journal of the American Statistical Association* 64.328 (1969), S. 1183–1210. doi: [10.1080/01621459.1969.10501049](https://doi.org/10.1080/01621459.1969.10501049).
- [22] D. Randall Wilson. „Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage“. In: *2011 International Joint Conference on Neural Networks, San Diego, USA*. IEEE, Juli 2011. doi: [10.1109/ijcnn.2011.6033192](https://doi.org/10.1109/ijcnn.2011.6033192).



## 2

---

# Deskriptive Statistik

Wir haben durch unsere alltägliche Erfahrung ein intuitives Verständnis davon, welche Körpergröße für Menschen in der Bevölkerung typisch ist: In weiten Teilen der Welt sind erwachsene Menschen typischerweise zwischen 1,60 m und 1,80 m groß, während Menschen mit einer Körpergröße von mehr als zwei Metern eher die Seltenheit sind.

Indem eine **Häufigkeitsverteilung** der Körpergröße angeben wird, kann diese Intuition mit Zahlen untermauert werden:

Körpergröße $l$	Anzahl Personen/ absolute Häufigkeit	relative Häufigkeit
$l < 1,60 \text{ m}$	81.199	19 %
$1,60 \text{ m} \leq l < 1,80 \text{ m}$	260.433	62 %
$1,80 \text{ m} \leq l < 2,00 \text{ m}$	77.462	18 %
$2,00 \text{ m} \leq l$	770	<1 %

**Tabelle 2.1.** Häufigkeit verschiedener Körpergrößen

Die Zahlen basieren auf einem von den US-Gesundheitsbehörden (CDC) erhobenen Datensatz, in dem unter anderem die Körpergröße von mehr als 340.000 Personen aufgelistet ist [1].

Ein Blick auf die Angaben für die Häufigkeiten in der Tabelle zeigt, dass in der Tat mehr als die Hälfte der im Rahmen der Studie befragten Personen eine Körpergröße zwischen 1,60 m und 1,80 m angegeben haben. Ein vorrangiges Ziel der **deskriptiven Statistik** oder **beschreibenden Statistik** ist es, mithilfe von Kennzahlen und grafischen Darstellungen wesentliche Charakteristiken von Häufigkeitsverteilungen herauszuarbeiten.

## 2.1 Stichprobe und Merkmale

Ist ein Datensatz Gegenstand statistischer Untersuchungen, so wird dieser auch als **Stichprobe** bezeichnet. Hier ist ein kleiner Ausschnitt aus der CDC-Stichprobe:

Geschlecht	Einkommen-klasse	Körpergröße in m	Körpergewicht in kg
weiblich	6	1,62	59
weiblich	4	1,66	91
weiblich	3	1,47	64
männlich	3	1,79	86
:	:	:	:

**Tabelle 2.2.** *CDC-Stichprobe*

Die Einkommenklasse nimmt hierbei diskrete Werte von 1 bis 8 an und bezeichnet in aufsteigender Reihenfolge ein bestimmtes Intervall an Haushaltseinkommen (z. B. steht die Klasse „3“ für ein Haushaltseinkommen von 15.000 US-Dollar oder mehr, jedoch weniger als 20.000 US-Dollar).

Die Variablen, welche die Spalten eines solchen Datensatzes in tabellarischer Form befüllen, hatten wir schon als Attribute bezeichnet. Im Rahmen der Statistik und des maschinellen Lernens werden diese auch **Merkmale** genannt (neudeutsch auch: *Features*). Ein von einem Merkmal angenommener Wert ist eine **Merkmalsausprägung**. Alle grundsätzlich möglichen Merkmalsausprägungen bilden zusammen den **Merkmalsraum**.

Merkmale, die nur eine begrenzte Anzahl von Ausprägungen annehmen können und über keine natürliche Anordnung verfügen, werden als **nominale Merkmale** bezeichnet. Das Geschlecht einer Person ist beispielsweise ein nominales Merkmal dieser Person.

Merkmale, die über eine begrenzte Anzahl von Ausprägungen verfügen und sich in natürlicher Weise anordnen lassen, werden als **ordinale Merkmale** bezeichnet. Zum Beispiel sind die Einkommenklassen in natürlicher Weise angeordnet, denn eine Person in der Einkommenklasse 6 verdient mehr als eine Person in der Einkommenklasse 3.

Ein Merkmal, das nominal oder ordinal ist, wird auch als **kategorial** oder **qualitativ** bezeichnet.

Merkmale, die einen sinnvollen zahlenmäßigen Vergleich zulassen, werden als **metrische** oder **quantitative Merkmale** bezeichnet. Zum Beispiel ist die Körpergröße ein metrisches Merkmal, denn Aussagen wie „Diese Person ist 10 cm größer als jene Person“ können sinnvoll getroffen werden.

Die Zeilen der tabellarisch erfassten Stichprobe werden gerne fälschlich ebenfalls als „Stichproben“ bezeichnet. Besser ist es, von **statistischen Einheiten**,

**Merksalsträgern, Beobachtungen oder Entitäten** zu sprechen – oder einfach konkret zu benennen: Im vorliegenden Fall können wir z. B. von „Personen“ sprechen. Die Anzahl der Zeilen bzw. Beobachtungen wird **Stichprobenumfang** oder **Stichprobengröße** genannt.

Eine Spalte stellt eine Folge von Ausprägungen eines Merkmals in der Stichprobe dar. Wir sprechen von einer **Stichprobenliste**, im Fall eines metrischen Merkmals auch von einem **Stichprobenvektor**.

Die Menge  $\Omega$  aller potenziellen Merksalsträger (also nicht nur jene, die durch die Stichprobe erfasst sind) bezeichnen wir auch als **Grundgesamtheit**. Im vorliegenden Fall ist die Grundgesamtheit durch alle erwachsenen US-Amerikaner bzw. deren Haushalte gegeben. Jede Teilmenge  $\Omega_0 \subseteq \Omega$  der Grundgesamtheit (also auch die Stichprobe) ist eine **Teilgesamtheit**. Handelt es sich bei der Teilgesamtheit konkret um eine Personengruppe, so wird – etwa im Rahmen einer demografischen Studie – auch von einer **Kohorte** gesprochen.

Jede vorgegebene Teilgesamtheit definiert in eindeutiger Weise ein **dichotomes** oder **binäres Merkmal**, also ein Merkmal mit den möglichen Ausprägungen  $\{0, 1\}$ . Dabei wird einem Merksalsträger der Wert „1“ zugeordnet, wenn dieser zur Teilgesamtheit gehört, und ein Wert von „0“, wenn dieser nicht zur Teilgesamtheit gehört. Umgekehrt definiert jedes binäre Merkmal in eindeutiger Weise eine Teilgesamtheit: all jene statistischen Einheiten, denen der Wert „1“ zugeordnet ist. In diesem Sinne sind Teilgesamtheiten und binäre Merkmale als Charakteristiken äquivalent.

Alle oben aufgeführten Beispiele von Merkmalen werden jeweils als **univariat** bezeichnet, denn mit jedem Merksalsträger nimmt das Merkmal jeweils nur einen Wert an. Wir können aber auch mehrere univariate Merkmale zu einem **multivariaten** Merkmal zusammenfassen.

Wir sprechen dann auch von einer **Merksaltsliste**, und gerade bei metrischen Merkmalen auch von einem **Merksaltsvektor** oder **Datenpunkt**: Beispielsweise kann für jede Person der Merksaltsvektor (Körpergröße in m, Körpergewicht in kg) betrachtet werden. Ein solches multivariates Merkmal mit genau zwei Einträgen wird auch **bivariat** genannt.

Insgesamt können die Variablen in dem CDC-Datensatz also wie folgt charakterisiert werden:

Merkmal	Merksaltsraum	Merksalstyp
Geschlecht	{weiblich, männlich, divers}	nominal
Einkommenklasse	$\{1, 2, \dots, 8\}$	ordinal
Körpergröße (Körpergröße, Körpergewicht)	$]0, \infty[ \subset \mathbb{R}$ $]0, \infty[ \times ]0, \infty[$	metrisch metrisch, bivariat

**Tabelle 2.3.** Merksalstypen

Schließlich können Teilgesamtheiten durch Bedingungen an Merkmale erklärt werden. Beispiele wären die Kohorte der männlichen Personen  $\Omega_{\text{männl.}} = \{\omega | \text{Gechlecht}(\omega) = \text{männlich}\}$  oder die Kohorte der Personen mit einem Body-Mass-Index (BMI) zwischen  $25 \text{ kg/m}^2$  und  $30 \text{ kg/m}^2$ :

$$\Omega_{\text{präadipös}} = \left\{ \omega \mid 25 \leq \frac{\text{Körpergewicht}(\omega)}{\text{Körpergröße}(\omega)^2} < 30 \right\}$$

## 2.2 Diagramme

In der Einführung hatten wir bereits eine Form von statistischem Diagramm vorgestellt, das **Liniendiagramm** (Abb. 0.1). Es ähnelt dem in einem der folgendem Abschnitte besprochenen Streudiagramm, jedoch werden aufeinanderfolgende Datenpunkte noch mit einem Streckenzug verbunden, um einen Anstieg oder Abfall im Wert der Ordinate hervorzuheben. Diese Art Diagramm eignet sich besonders für die Darstellung von **Zeitreihen**, also zeitlich geordneten Daten.

Der Einsatz statistischer Diagramme eignet sich gut für **explorative Datenanalysen**. Diese dienen dem Aufdecken von Mustern und Zusammenhängen in den Daten sowie dem Aufstellen neuer Hypothesen, die anhand der Daten getestet werden können.

### 2.2.1 Säulendiagramme und Histogramme

Um uns – im wahrsten Sinne des Wortes – ein Bild von einer gegebenen Häufigkeitsverteilung zu machen, können wir diese in einer Form von Diagramm darstellen, welches im Fall kategorialer Merkmale **Säulendiagramm** und im Fall metrischer Merkmale **Histogramm** genannt wird. In Abb. 2.5 sind ein paar Säulendiagramme bzw. Histogramme für Merkmale des CDC-Datensatzes dargestellt.

Bei kategorialen Merkmalen wie Geschlecht oder Einkommensklasse gibt die Höhe jedes Rechtecks die relative Häufigkeit der jeweiligen Merkmalsausprägung an. Es kann auch die absolute Häufigkeit aufgetragen werden, dies entspricht lediglich einer gemeinsamen Skalierung aller Säulenhöhen. Der Breite der Rechtecke kommt keine besondere Bedeutung zu.

Für metrische Merkmale wie Körpergröße oder Gewicht muss erst eine geeignete **Klasseneinteilung** des Merkmalsraums gewählt werden, um die Beobachtungen innerhalb jeder Klasse auszählen und so die Häufigkeit ermitteln zu können. Hier spielt insbesondere die **Klassenbreite** eine Rolle, also die Breite jedes Rechtecks. Zum Beispiel wurde als Klassenbreite für das Histogramm der Körpergröße genau 2,54 cm gewählt. Dies entspricht einem Zoll, der anglo-amerikanischen Maßeinheit, in der die Angaben ursprünglich gemacht wurden.

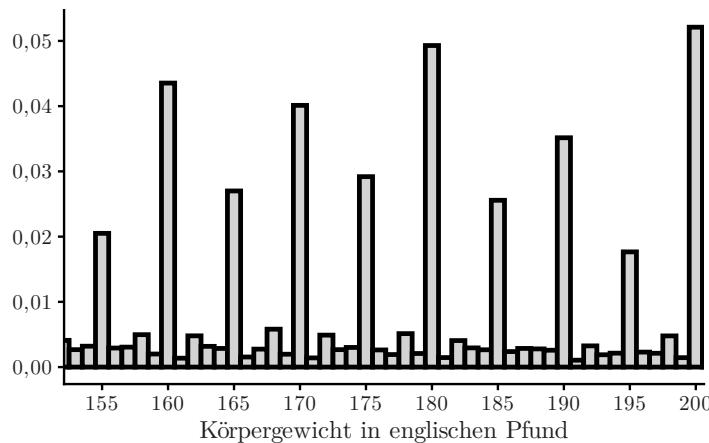
In diesem Fall wird die Häufigkeit durch die *Fläche* jedes Rechtecks repräsentiert, die Höhe ergibt sich aus der Häufigkeit dividiert durch die Klassenbreite. Grundsätzlich können Histogramme auch eine variable Klassenbreite aufweisen.

An Säulendiagrammen und Histogrammen können bereits wesentliche Charakteristiken von Häufigkeitsverteilungen abgelesen werden:

- Welche Merkmalsausprägung ist am häufigsten, welche Werte eines metrischen Merkmals können als typisch angesehen werden?
- Kommen alle Ausprägungen gleich häufig vor, haben also eine große Streuung? Oder sind sie um einen typischen Wert herum konzentriert?

Statistische Kennzahlen helfen uns, diese Charakteristiken einer rechnerischen Analyse zugänglich zu machen: Dies wird durch die in den späteren Abschnitten beschriebenen Lage- bzw. Streuungsparametern geleistet.

**Anwendungsbeispiel.** Vielleicht fällt Ihnen am Histogramm der Häufigkeitsverteilung des Körpergewichts in der CDC-Studie auf, dass dieses in regelmäßigen Abständen deutlich erkennbare Lücken aufweist. Ein Ausschnitt des Histogramms in Einheiten des angloamerikanischen Maßsystems gibt näheren Aufschluss:



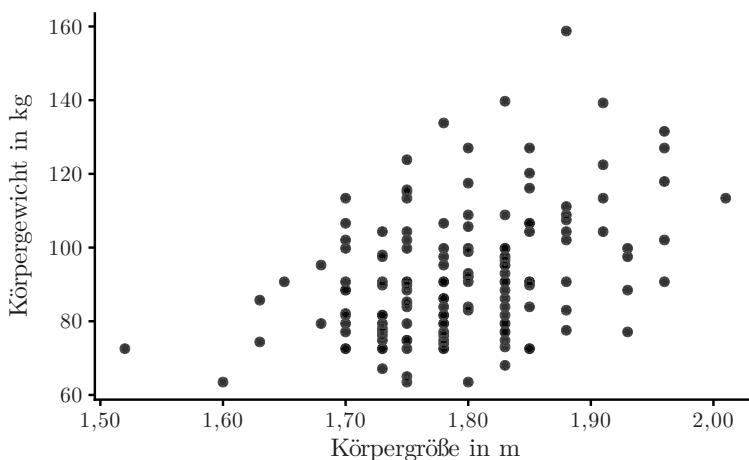
**Abb. 2.1.** Häufigkeitsverteilung von Eigenangaben des Körpergewichts

Die Daten wurden mittels einer telefonischen Umfrage unter US-Bürgerinnen und -Bürgern erhoben. Die Teilnehmer/-innen gaben ihr Gewicht selten auf das Pfund genau, sondern gerundete Werte an. Aus diesem Grund sind Angaben, die durch fünf teilbar sind, überrepräsentiert.

## 2.2.2 Streudiagramme

Merkmale sind oftmals nicht unabhängig voneinander, es können funktionale Zusammenhänge zwischen ihnen bestehen. Zusammenhänge zwischen zwei metrischen Merkmalen können grafisch durch **Streudiagramme** dargestellt werden. Diese Diagramme dienen dazu, bekannte Zusammenhänge zu bestätigen oder im Rahmen einer explorativen Datenanalyse bisher unbekannte Zusammenhänge aufzuklären.

Für ein Streudiagramm werden die gepaarten (also jeweils zu einer Beobachtung gehörigen) Ausprägungen als Datenpunkte in einem Achsenkreuz aufeinander aufgetragen. Das folgende Beispiel zeigt Körpergröße und -gewicht für 150 zufällig ausgewählte männliche Befragte der CDC-Studie. Erwartungsgemäß ist eine allgemeine Tendenz erkennbar, nach der größere Menschen auch ein höheres Gewicht haben. Dennoch gibt es auch bei gleicher Körpergröße eine große Streuung des Körbergewichts unter den Befragten. Die Größe ist also nicht allein ausschlaggebend für das Gewicht – auch dieses Ergebnis deckt sich mit unserer Alltagserfahrung:



**Abb. 2.2.** Streudiagramm von Körpergröße und -gewicht

**Anwendungsbeispiel.** Als weiteres illustratives Beispiel wollen wir anhand des CDC-Datensatzes den Body-Mass-Index und den Broca-Index miteinander vergleichen. Diese Maßzahlen dienen einer groben Bestimmung dessen, was noch als ein gesundes Körbergewicht angesehen werden kann. Der BMI berechnet sich über die folgende Formel:

$$\text{BMI} = \frac{\text{Körbergewicht}}{(\text{Körpergröße})^2}$$

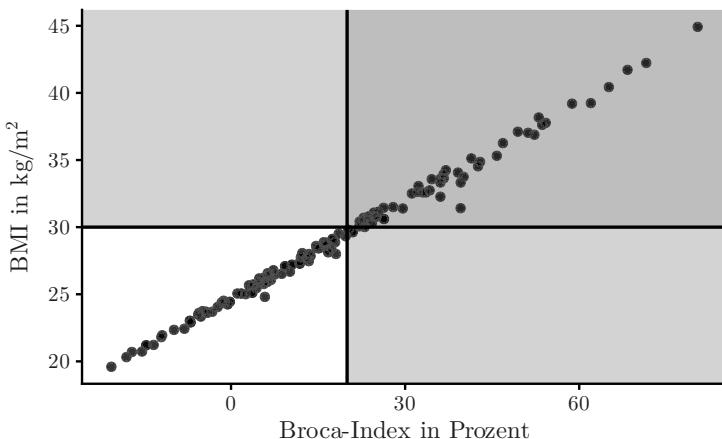
Gewöhnlich wird der BMI in Einheiten von Kilogramm pro Quadratmeter angegeben. Erwachsene mit einem BMI von mehr als  $30 \frac{\text{kg}}{\text{m}^2}$  gelten laut Weltgesundheitsorganisation im Allgemeinen als fettleibig.

Das Normalgewicht nach Broca wird wie folgt berechnet:

$$\text{Normalgewicht nach Broca in kg} = \text{Körpergröße in cm} - 100$$

Der Broca-Index ist die prozentuale Abweichung des tatsächlichen Körpergewichts von diesem Richtwert. Personen mit einer Abweichung von mehr als +20 % können als adipös angesehen werden.

Das folgende Diagramm zeigt eine Auftragung des BMI auf den Broca-Index für die weiter oben ausgewählten männlichen Befragten der CDC-Studie.



**Abb. 2.3.** Streudiagramm von Broca- und Body-Mass-Index

Die grauen Bereiche zeigen zusätzlich die „kritischen“ Bereiche von  $\text{BMI} > 30 \frac{\text{kg}}{\text{m}^2}$  bzw.  $\text{Broca-Index} > 20\%$  an. Beide Maßzahlen zeigen eine starke Abhängigkeit voneinander: Obgleich der Broca-Index gegenüber dem BMI „aus der Mode“ gekommen ist, treffen beide Kennzahlen sehr ähnliche Aussagen. Auch die Kennzeichnungen von Fettleibigkeit sind miteinander verträglich: Es finden sich nur wenige Beobachtungen in den hellgrauen Bereichen links oben oder rechts unten, welche auf eine unstimmige Einschätzung hinweisen.

Neben der grafischen Darstellung können Zusammenhänge zwischen Merkmalen zahlenmäßig messbar gemacht werden. Dies ist Aufgabe von Assoziations- und Korrelationsparametern, auf die wir in Abschn. 2.5 näher eingehen wollen.

Darüber hinaus können die Zusammenhänge statistisch modelliert werden. Die Abhängigkeit obiger Merkmale scheint linear zu sein: Die Datenpunkte liegen zwar nicht genau, aber doch annähernd auf einer Geraden. Die optimale

Position und Orientierung einer solchen Ausgleichsgeraden kann durch eine Regressionsanalyse ermittelt werden – siehe Abschn. 4.5.1.

### 2.2.3 Weitere Diagramme

#### Heatmap-Diagramm

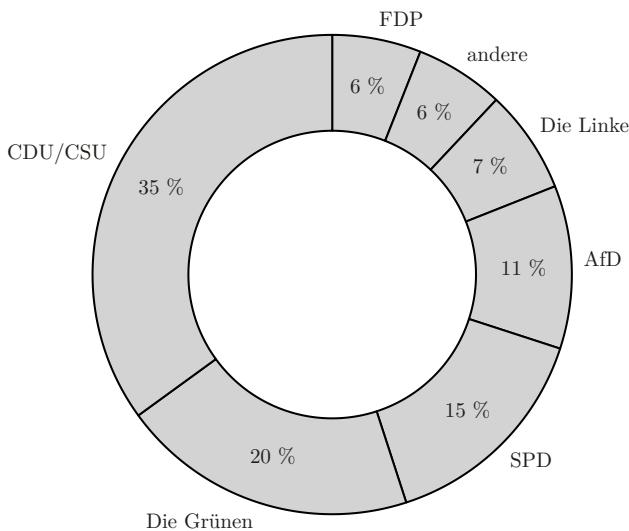
Ein **Heatmap-Diagramm** dient der Visualisierung von Kennzahlen, die in einer Matrix angeordnet sind. Dabei entspricht jeder Zahlenwert einer farbcodierten Kachel. Abb. 2.6 zeigt links ein solches Heatmap-Diagramm auf Grundlage von Daten der ALLBUS-Umfrage 2018 [2]. Das Diagramm stellt die Präferenz für eine politische Partei in Deutschland der Beantwortung folgender Frage gegenüber: „Was halten Sie von folgender Aussage: Der Zuzug von Flüchtlingen sollte unterbunden werden.“ Die Farbsättigung der Kacheln zeigt den Anteil der Personen mit gegebener Parteipräferenz an: Jede Zeile stellt also die Verteilung der Häufigkeit von Antworten von Personen mit der jeweiligen Parteipräferenz dar. Auf diese Weise können die Verteilungen in den jeweiligen Kohorten schnell und einprägsam miteinander verglichen werden.

Eine weitere Variante der Heatmap ist die **Choroplethenkarte**, bei der geografische Gebiete eingefärbt werden. Eine solche Karte sehen Sie in Abb. 2.6 rechts. Sie zeigt den Anteil der Befragten der ALLBUS-Studie, deren Haus über eine Gegensprechanlage verfügt, aufgeschlüsselt nach Bundesland. Anhand der Karte ist sofort ersichtlich, dass dieser Anteil in Baden-Württemberg am größten ist. Erste Nachforschungen ergeben, dass ein führender Hersteller für Hauskommunikationssysteme seinen Sitz in diesem Bundesland hat. Der Verdacht, dass diese zahlenmäßige Verbindung aufgrund eines kausalen Zusammenhangs besteht, müsste freilich durch weitere Daten erhärtet werden.

#### Kreisdiagramm

Eine weitere Methode Anteile darzustellen ist das **Kreisdiagramm**. Dabei werden die Anteile als Sektoren eines Kreises dargestellt, ähnlich Kuchen- oder Pizzastücken. Obwohl in öffentlichen Publikationen weit verbreitet, gilt die Eignung des Kreisdiagramms als effektive Visualisierungsform in Fachkreisen als umstritten [3]. Ein Kreisdiagramm sollte nach Möglichkeit nur verwendet werden, wenn es nicht zu viele Merkmalsausprägungen bzw. „Kuchenstücke“ darstellte. Von der Verwendung dreidimensionaler Kreisdiagramme ist in jedem Fall abzusehen.

Das **Ringdiagramm** ist eine Variante des Kreisdiagramms, das folgende zeigt die Parteipräferenz in Deutschland gemäß der sogenannten Sonntagsfrage (laut infratest dimap, Stand: 16. Oktober 2020 [4]):



**Abb. 2.4.** Ringdiagramm

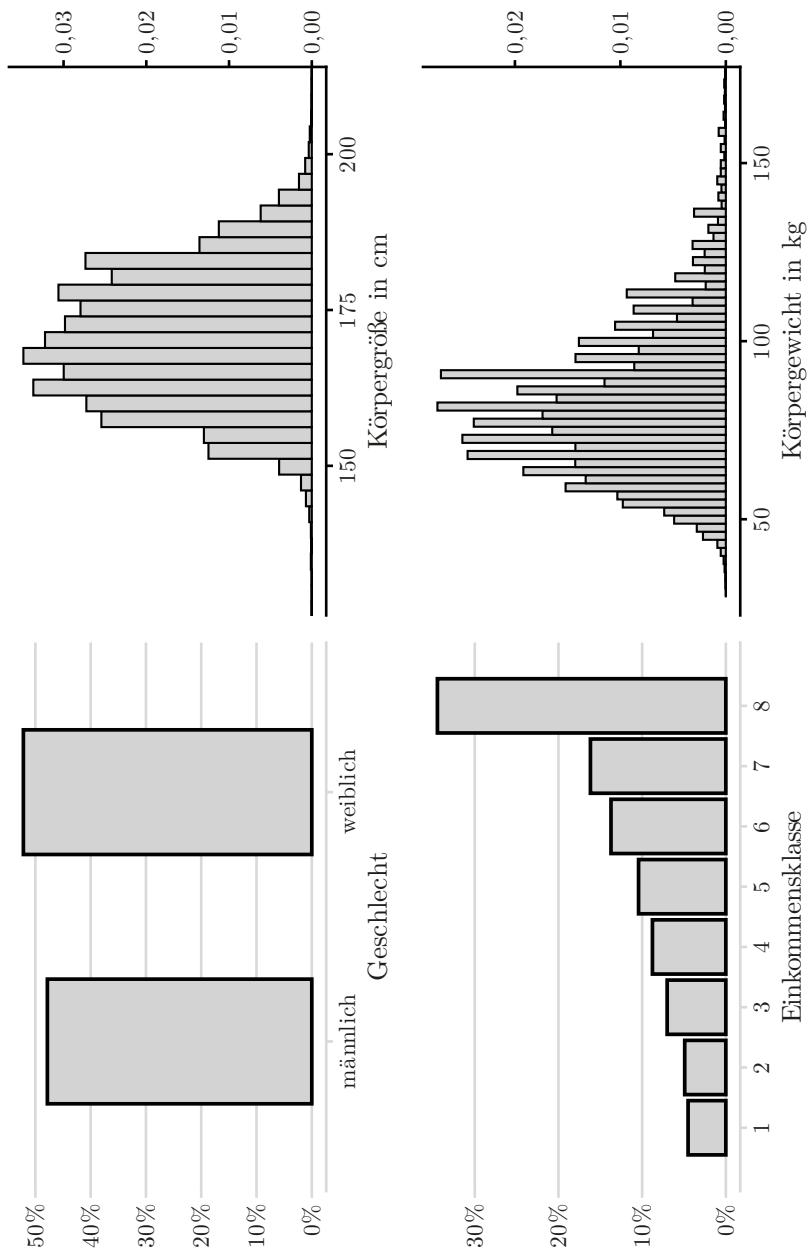
## 2.3 Lageparameter

Lageparameter dienen der Beantwortung der Frage, welche der in der Stichprobe vorkommenden Merkmalsausprägungen „typisch“ sind.

Der Tab. 2.1 am Anfang des Kapitels kann leicht entnommen werden, dass die meisten Menschen eine Körpergröße zwischen 160 und 180 cm aufweisen: 62 %, das ist ein größerer Anteil als in jeder anderen der angegebenen Größenkohorten. Daher kann der Wertebereich zwischen 160 und 180 cm als „typisch“ angesehen werden, denn Menschen mit einer Körpergröße in diesem Bereich werden am häufigsten angetroffen. Anders ausgedrückt: Die Häufigkeitsverteilung nimmt dort ein Maximum an. Ein solches Maximum wird auch als ein **Modus** der Verteilung bezeichnet. In der Praxis wird oft von „dem“ Modus gesprochen, obwohl dieser nicht eindeutig bestimmt sein muss: Zum einen können verschiedene Merkmalsausprägungen mit genau derselben oder fast derselben Häufigkeit vorkommen. Zum anderen kann es vorkommen, dass ein Histogramm mehrere *lokale* Maxima aufweist, in diesem Fall heißt die Verteilung **multimodal**, andernfalls **unimodal**. In Abb. 2.8 ist das Histogramm einer multimodalen Verteilung dargestellt.

Bei kategorialen Merkmalen muss für die Bestimmung des Modus einfach nur gezählt werden, welche Merkmalsausprägung am häufigsten vorkommt. Bei metrischen Merkmalen hängt der Modus offensichtlich von der Klasseneinteilung ab, die der Auszählung der Häufigkeiten zugrundegelegt wird.

**Abb. 2.5.** Häufigkeitsverteilung von Merkmalen der CDC-Studie: Säulendiagramme (links) und Histogramme (rechts)



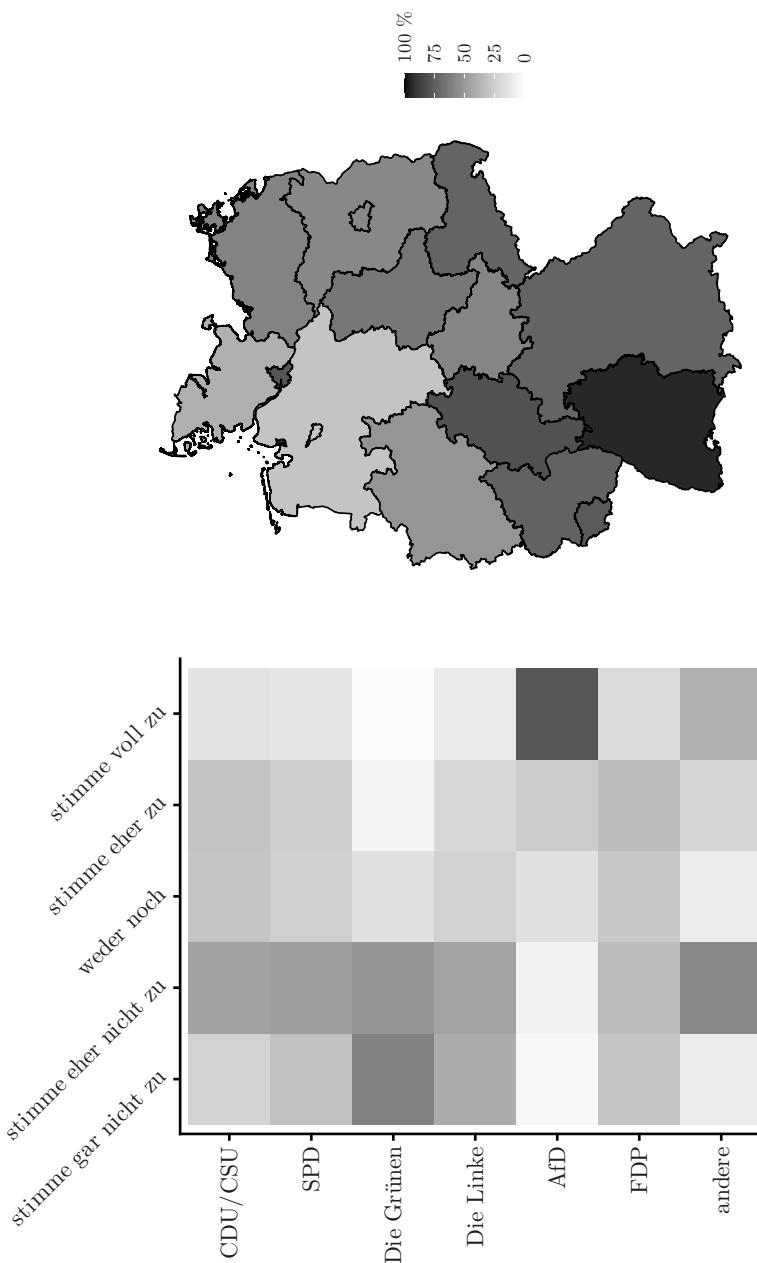


Abb. 2.6. Heatmap (links): Parteipräferenz und Meinung zur Notwendigkeit von Migrationskontrolle; Chloroplethenkarte (rechts): Vorhandensein einer Gegen sprechanlage

### 2.3.1 Arithmetisches Mittel und empirischer Median

Neben dem Modus gehören die folgenden Lageparameter zu den gebräuchlichsten.

Für einen metrischen Stichprobe  $x = (x_1, x_2, \dots, x_N) \in \mathbb{R}^N$  ist das **arithmetische Mittel** wie folgt erklärt:

$$\bar{x}_{\text{arithm}} = \frac{1}{N} (x_1 + x_2 + \dots + x_N) = \frac{1}{N} \sum_{n=1}^N x_n$$

Angenommen, die Merkmalsausprägungen wurden aufsteigend sortiert:  $x_1 \leq x_2 \leq \dots \leq x_N$ . Dann ist der **empirische Median** von  $x$  als der mittlere Wert dieser sortierten Folge definiert:

$$\bar{x}_{\text{median}} = \begin{cases} x_{\frac{N+1}{2}} & \text{falls } N \text{ ungerade} \\ \frac{1}{2} \cdot (x_{\frac{N}{2}} + x_{\frac{N}{2}+1}) & \text{falls } N \text{ gerade} \end{cases}$$

Ein Median kann auch für ordinale Merkmale berechnet werden. Allerdings gibt es dann im Allgemeinen keine sinnvolle Konvention mehr, mit der im Fall einer geraden Anzahl von Beobachtungen ein eindeutiges Ergebnis erzwungen werden kann. So wären z. B. für eine Stichprobe  $x = (1, 5, 6, 8)$  von Einkommensklassen sowohl  $\bar{x}_{\text{median}} = 5$  als auch  $\bar{x}_{\text{median}} = 6$  gültige Mediane.

In der Praxis würde dieses Ergebnis unter Umständen dennoch mit  $\bar{x}_{\text{median}} = 5,5$  beschrieben, obwohl keine gebrochenzahligen Einkommensklassen definiert sind. Sogar eine Berechnung des arithmetischen Mittel für – streng genommen – ordinale Merkmale kann in manchen Fällen sinnvoll sein. Beispielsweise könnte im Verlaufe eines Würfelspiels im Mittel die Augenzahl 3,14 fallen – obwohl ein Würfel natürlich keine Seite hat, der eine solche gebrochenzahlige Augenzahl entsprechen würde.

Für den Median schreiben wir auch:

$$\text{median}(x) = \underset{n \in \{1, \dots, N\}}{\text{median}} (x_n) = \bar{x}_{\text{median}}$$

Das arithmetische Mittel (auch arithmetischer Mittelwert genannt) ist sehr gebräuchlich, daher wird das Subskript im Folgenden oft weglassen, wenn kein anderer Lageparameter gemeint ist:  $\bar{x} = \bar{x}_{\text{arithm}}$ . Weitere mögliche Schreibweisen sind auch die folgenden:

$$\mu(x) = \langle x \rangle = \langle x_n \rangle_{n \in \{1, \dots, N\}} = \bar{x}_{\text{arithm}}$$

Eine weitere Bezeichnung für das arithmetische Mittel ist **empirischer Erwartungswert**. Das Adjektiv „empirisch“ wird uns in der deskriptiven Statistik noch öfter begegnen. Empirisch erhobene Parameter haben ihre – meist

gleichnamigen – Entsprechungen in der später diskutierten Wahrscheinlichkeitstheorie, die sich von idealisierten Zufallsgrößen ableiten. Die empirischen Kennzahlen werden dann im Rahmen der Inferenzstatistik als Schätzungen dieser „theoretischen“ Kennzahlen aufgefasst. Mitunter werden wir die Kennzeichnung „empirisch“ der Kürze halber aber fortlassen.

**Anwendungsbeispiel.** Das arithmetische Mittel der im Rahmen der CDC-Studie erhobenen Körpergröße beträgt 1,70 m. Der Median beträgt ebenfalls 1,70 m.

Obwohl das arithmetische Mittel so gebräuchlich ist, hat der Median eine Eigenschaft, die ihn als Lageparameter für manche Anwendungen überlegen macht: Er ist **robust** gegenüber **Ausreißern**. Das bedeutet, dass er seinen Wert nicht oder nur wenig ändert, wenn sich manche Werte in der Stichprobe wesentlich von den übrigen unterscheiden. Ausreißer können die Folge fehlerhafter Daten sein, sodass eine solche Robustheit wünschenswert sein kann.

Wir illustrieren diese Eigenschaft anhand eines Beispiels und berechnen für  $x = (-1,0; -1,0; 0,0; 1,0; 2,0; 100,0)$  den arithmetischen Mittelwert und den Median:

$$\bar{x} = \frac{1}{6} \cdot (-1,0 - 1,0 + 0,0 + 1,0 + 2,0 + 100,0) \approx 16,8$$

$$\bar{x}_{\text{median}} = \frac{1}{2} \cdot (x_3 + x_4) = \frac{1}{2} \cdot (0,0 + 1,0) = 0,5$$

Der Ausreißer  $x_6 = 100,0$  bestimmt also wesentlich den Wert des arithmetischen Mittels, nicht jedoch den Wert des Medians.

Bei sogenannten **schiefen Verteilungen** kann die relative Häufigkeit von Werten, die kleiner bzw. größer als der arithmetische Mittelwert sind, wesentlich von 50 % abweichen. Der Median ist aber gerade so konstruiert, dass stets 50 % aller Stichprobenwerte kleiner bzw. größer als dieser sind.

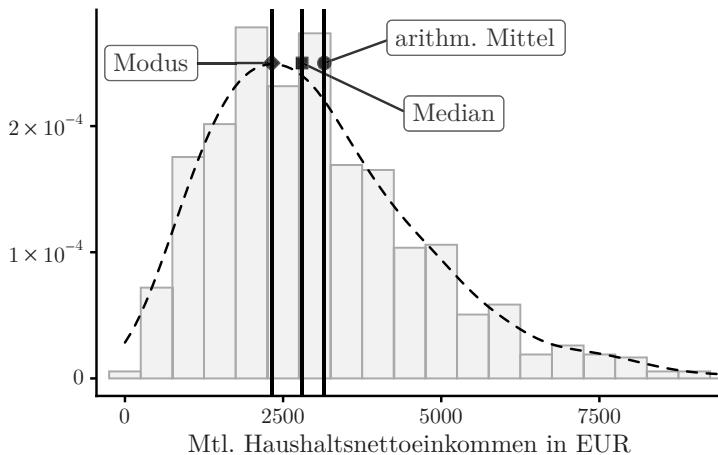
**Anwendungsbeispiel.** Einkommenstatistiken führen oft auf schiefe Verteilungen. So beziffert sich gemäß der ALLBUS-Studie das mittlere monatliche Nettohaushaltseinkommen in Deutschland im Jahr 2018 auf 3150 EUR. Tatsächlich haben jedoch mehr als die Hälfte (60 %) der Haushalte ein niedrigeres Einkommen. Das Medianeinkommen ist daher ein besserer Richtwert, dieses beträgt 2800 EUR.

Für die Bewertung der Schiefe oder Asymmetrie einer Häufigkeitsverteilung gibt es verschiedene Kennzahlen. Eine *Faustregel* (vgl. [5]) ist die folgende.

Eine unimodale Verteilung von Werten  $x_1, \dots, x_N$  mit Modus  $\bar{x}_{\text{mod}}$ , Median  $\bar{x}_{\text{median}}$  und arithmetischen Mittelwert  $\bar{x}$  kann wie folgt charakterisiert werden; sie ist

**symmetrisch**, falls  $\bar{x}_{\text{mod}} \approx \bar{x}_{\text{median}} \approx \bar{x}$ ,  
**linksschief**, falls  $\bar{x}_{\text{mod}} > \bar{x}_{\text{median}} > \bar{x}$ ,  
**rechtsschief**, falls  $\bar{x}_{\text{mod}} < \bar{x}_{\text{median}} < \bar{x}$ .

Die Verteilung der Einkommen in der ALLBUS-Studie ist rechtsschief, wie an folgender Abbildung abgelesen werden kann:



**Abb. 2.7.** Histogramm einer rechtsschiefen Verteilung

Der Modus wurde über das Maximum einer sogenannten Kerndichteschätzung der Verteilung ermittelt (gestrichelte Kurve). Wie eine solche berechnet wird, erklären wir in Abschn. 4.4.3.

### 2.3.2 Quantile

Eine weitere wichtige Klasse von Lagemaßen sind Quantile.

Sei  $0 < \alpha < 1$  und  $x = (x_1, \dots, x_N)$  eine Stichprobe von metrischen oder ordinalen Merkmalen. Ein **empirisches Quantil der Ordnung  $\alpha$**  oder **empirisches  $\alpha$ -Quantil** ist ein Wert  $Q_\alpha(x)$  der Stichprobe unter bzw. über dem ein bestimmter Anteil von Beobachtungen vorgefunden werden kann:

1. Mindestens  $\alpha \cdot N$  der Datenpunkte sind kleiner oder gleich  $Q_\alpha(x)$ , und
2. mindestens  $(1 - \alpha) \cdot N$  der Datenpunkte sind größer oder gleich  $Q_\alpha(x)$ .

Das empirische  $\alpha$ -Quantil ist nur dann nicht eindeutig bestimmt, wenn  $\alpha \cdot N$  ganzzahlig ist. Andernfalls ist das Quantil auf eindeutige Weise durch  $Q_\alpha(x) = x_{[\alpha N] + 1}$  gegeben, wobei die Stichprobe vorher aufsteigend sortiert wurde und  $[\alpha N]$  der ganzzahlige abgerundete Wert von  $\alpha \cdot N$  ist.

In der Praxis ist es andernfalls – wie schon beim Median – üblich, durch Bildung des arithmetischen Mittelwerts der Kandidaten Eindeutigkeit zu erzwingen:

$$Q_\alpha(x) = \begin{cases} x_{\lfloor \alpha N \rfloor + 1} & \text{falls } \alpha \cdot N \notin \mathbb{N} \\ \frac{1}{2} \cdot (x_{\alpha N} + x_{\alpha N + 1}) & \text{falls } \alpha \cdot N \in \mathbb{N} \end{cases}$$

Der Median ist gerade mit dem 0,5-Quantil identisch:  $\bar{x}_{\text{median}} = Q_{0,5}(x)$ . Weitere Quantile, die einen besonderen Namen verdienen, sind das untere und obere **Quartil**:  $Q_{0,25}$  und  $Q_{0,75}$ .

**Anwendungsbeispiel.** Das untere Quartil der im Rahmen der CDC-Studie erhobenen Körpergröße beträgt 1,63 m, das obere 1,78 m.

### 2.3.3 Geometrisches und harmonisches Mittel

Die folgenden Lagemaße haben besondere Einsatzbereiche, in denen das arithmetischen Mittel auf keinen sinnvollen Durchschnittswert führte.

Für eine Stichprobe von *positiven* metrischen Werten  $x = (x_1, x_2, \dots, x_N) \in ]0, \infty[^N$  wird definiert:

Das **geometrische Mittel**

$$\bar{x}_{\text{geom}} = (x_1 \cdot x_2 \cdots x_N)^{\frac{1}{N}} = \left( \prod_{n=1}^N x_n \right)^{\frac{1}{N}}$$

und das **harmonische Mittel**

$$\bar{x}_{\text{harm}} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}} = N \cdot \left( \sum_{n=1}^N \frac{1}{x_n} \right)^{-1}.$$

Es kann gezeigt werden, dass – für eine beliebige Stichprobe  $x = (x_1, \dots, x_N)$  von positiven Werten – die verschiedenen Lageparameter folgenden allgemeinen Ungleichungen genügen:

$$0 < \min\{x_n\} \leq \bar{x}_{\text{harm}} \leq \bar{x}_{\text{geom}} \leq \bar{x}_{\text{arithm}} \leq \bar{x}_{\text{quadr}} \leq \max\{x_n\}$$

Dabei ist  $\bar{x}_{\text{quadr}}$  der bislang noch nicht erwähnte **quadratische Mittelwert**:

$$\bar{x}_{\text{quadr}} = \frac{1}{\sqrt{N}} \cdot \sqrt{\sum_{n=1}^N x_n^2}$$

Das geometrische Mittel findet insbesondere Verwendung in der Mittelung von Wachstumsfaktoren, wie das folgende Rechenbeispiel illustrieren soll: Eine Geldanlage von 1.000 EUR vermehrt sich variabel: im ersten Jahr mit

$p_1 = +20,0\%$  Gewinn, im zweiten Jahr mit  $p_2 = -20,0\%$  Verlust, im dritten Jahr mit  $p_3 = +10,0\%$  Gewinn. Die entsprechenden Wachstumsfaktoren sind  $W = (1,200; 0,800; 1,100)$ . Am Ende steht also der folgende Betrag in EUR:

$$1000 \cdot 1,200 \cdot 0,800 \cdot 1,100 \approx 1000 \cdot 1,056$$

Der gesamte Wachstumsfaktor liegt folglich bei 1,056, das entspricht einer gesamten Wachstumsrate von  $+5,6\%$ . Was ist jedoch der durchschnittliche, jährliche Wachstumsfaktor? Das ist derjenige Faktor  $\bar{W}_?$ , der einer gedachten festen Verzinsung mit demselben Endergebnis zugrunde liegt:

$$1000 \cdot \bar{W}_? \cdot \bar{W}_? \cdot \bar{W}_? = 1000 \cdot 1,200 \cdot 0,800 \cdot 1,100$$

Es ist nicht so schwer zu sehen, dass dieser durchschnittliche Wachstumsfaktor gerade dem geometrischen Mittel der einzelnen Wachstumsfaktoren entspricht:

$$\bar{W}_? = \bar{W}_{\text{geom}} = (1,200 \cdot 0,800 \cdot 1,100)^{\frac{1}{3}} \approx 1,018$$

Das ergibt eine durchschnittliche jährliche Wachstumsrate von  $1,8\%$ . Eine naive Verwendung des arithmetischen Mittels hätte zu einem abweichenden und damit sachlich falschen Ergebnis geführt:  $\bar{W}_{\text{arithm}} = 1,033$ ,  $\bar{p}_{\text{arithm}} = 3,3\%$ .

Das harmonische Mittel wird uns später bei der Definition des  $F_1$ -Maßes wiederbegegnen, einer Kennzahl für die Beurteilung von Klassifikationsalgorithmen (siehe Abschn. 6.1.3). Eine andere Anwendung besteht in der Mitteilung von Geschwindigkeiten, wie das folgende Rechenbeispiel illustrieren soll: Anna fährt morgens mit dem öffentlichen Bus zur Arbeit (Entfernung  $\Delta s = 5 \text{ km}$ ), bei wenig Verkehr und mit einer Durchschnittsgeschwindigkeit von  $v_1 = 40 \frac{\text{km}}{\text{h}}$ . Die Durchschnittsgeschwindigkeit auf der Rückfahrt nach Hause beträgt nur  $v_2 = 10 \frac{\text{km}}{\text{h}}$ , da sich der Bus durch den Verkehr kämpfen muss.

Mit welcher durchschnittlichen Geschwindigkeit  $\bar{v}_?$  hat Anna bzw. der Bus die Strecke auf Hin- und Rückweg,  $2 \cdot \Delta s$ , zurückgelegt? Diese Geschwindigkeit entspricht der zurückgelegten Strecke geteilt durch die insgesamt benötigte Zeit  $\Delta t$ :  $\bar{v}_? = \frac{2 \cdot \Delta s}{\Delta t}$ . Für den Hinweg benötigt der Bus die Zeit  $\Delta t_1 = \frac{\Delta s}{v_1}$ , für den Rückweg die Zeit  $\Delta t_2 = \frac{\Delta s}{v_2}$ . Damit ergibt sich:

$$\bar{v}_? = \frac{2 \cdot \Delta s}{\Delta t_1 + \Delta t_2} = \frac{2 \cdot \Delta s}{\frac{\Delta s}{v_1} + \frac{\Delta s}{v_2}} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}} = \bar{v}_{\text{harm}}$$

In konkreten Zahlen:  $\bar{v}_{\text{harm}} = 16 \frac{\text{km}}{\text{h}}$ . Das arithmetische Mittel hat einen höheren Wert:  $\bar{v}_{\text{arithm}} = 25 \frac{\text{km}}{\text{h}}$ . Das liegt daran, dass Anna wesentlich mehr Zeit für die Rückfahrt als für die Hinfahrt benötigt. Daher sitzt sie auch längere Zeit im langsam fahrenden Bus, sodass sie sich insgesamt langsamer fortbewegt als durch das arithmetische Mittel gegeben.

## 2.4 Streuungsparameter

Lageparameter sollen einen typischen Wert einer Verteilung widerspiegeln. Allerdings kann die Aussagekraft eines solchen Parameters variieren; lax formuliert kann der ermittelte Wert „mehr oder weniger typisch“ sein. Beispielsweise gaben etwa 55 % der Befragten der CDC-Studie an, weiblichen Geschlechts zu sein, während 45 % angaben, männlichen Geschlechts zu sein. Damit ist „weiblich“ streng genommen der Modus des Merkmals „Geschlecht“. Dennoch ist diese Merkmalsausprägung kaum als typisch zu bezeichnen, denn der Anteil von Personen männlichen Geschlechts ist in etwa gleich.

Andererseits kann die durchschnittliche Körpergröße von 1,70 m durchaus als typischer Wert erachtet werden, denn aus Erfahrung oder der Gestalt des Histogramms können wir ableiten, dass die Größe von vergleichweise wenigen Personen in signifikantem Maße von diesem Mittelwert abweicht.

**Streuungsparameter** geben an, inwieweit Merkmalsausprägungen mit gleicher Häufigkeit vorkommen bzw. quantifizieren das Ausmaß von deren Streuung um einen Lageparameter herum.

### 2.4.1 Abweichung von Mittelwert oder Median

Ein einfaches Maß für die Streuung metrischer Werte ist die **Spannweite**, gegeben durch die Differenz zwischen größter und kleinster Merkmalsausprägung in der Stichprobe. Bereits einzelne Ausreißer können diesen Wert allerdings erheblich verändern. Ein Streuungsparameter, der von der Grundidee der Spannweite ähnelt aber größere Robustheit aufweist, ist der **Quartilsabstand**. Dieser ist durch die Differenz von oberem und unterem Quartil gegeben:  $Q_{0,75} - Q_{0,25}$ .

**Anwendungsbeispiel.** Der Quartilsabstand der im Rahmen der CDC-Studie erhobenen Körpergröße beträgt  $178 \text{ cm} - 163 \text{ cm} = 15 \text{ cm}$ .

Den folgenden Streuungsparameter liegt die Idee zugrunde, die durchschnittliche Abweichung vom Durchschnittswert zu bemessen.

Für einen Stichprobenvektor  $x = (x_1, x_2, \dots, x_N) \in \mathbb{R}^N$  ist die **empirische Varianz** die mittlere quadrierte Abweichung vom arithmetischen Mittelwert:

$$s^2(x) = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2$$

Die **empirische Standardabweichung** ist die Quadratwurzel der Varianz:  $s(x) = \sqrt{s^2(x)}$ .

Die **mittlere Abweichung vom Median** ist:

$$s_{\text{median}}(x) = \frac{1}{N} \sum_{n=1}^N |x_n - \bar{x}_{\text{median}}|$$

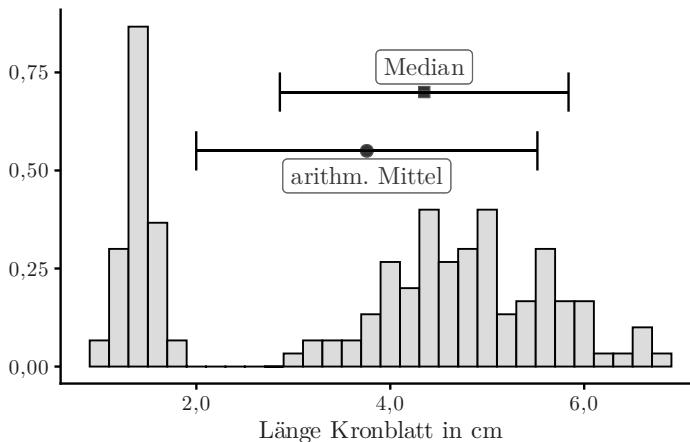
**Anwendungsbeispiel.** Die Standardabweichung der im Rahmen der CDC-Studie erhobenen Körpergröße beträgt 11 cm, die mittlere Abweichung vom Median ist 9 cm.

Neben der obigen Definition gibt es auch noch die **korrigierte empirische Varianz**, deren Verwendung sich insbesondere für kleine Stichproben empfiehlt:

$$s_{\text{kor}}^2(x) = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2 = \frac{N}{N-1} \cdot s^2(x)$$

Für große Stichproben liefern beide Definitionen jedoch näherungsweise dieselben Ergebnisse. Was es mit der Bedeutung der korrigierten Varianz auf sich hat, erfahren Sie in Abschn. 4.2.3.

**Anwendungsbeispiel.** Das folgende Histogramm gibt Aufschluss über die Verteilung der Länge des Kronblattes von Schwertliliengewächsen [6] sowie deren arithmetischer Mittelwert und Median:



**Abb. 2.8.** Streuung um arithmetischen Mittelwert und Median

Der horizontale Balken gibt jeweils das Intervall  $[\bar{x} - s(x), \bar{x} + s(x)]$  bzw. zwischen  $\bar{x}_{\text{median}} \pm s_{\text{median}}(x)$  an. Anscheinend ist die Häufigkeitsverteilung multimodal: Es können zwei Maxima der Verteilung, ja sogar zwei separierte Gruppen ausgemacht werden. Tatsächlich gehören diese zu verschiedenen Arten der Pflanze. Für eine genauere Analyse mag es daher sinnvoll sein, Lage- und Streuparameter nach diesen Gruppen getrennt zu evaluieren.

## 2.4.2 Shannon-Entropie

Für ein kategoriales Merkmal mit  $K$  möglichen Ausprägungen können wir die relativen Häufigkeiten  $f_1, \dots, f_K$  des Vorkommens jeder Ausprägung erhalten,

indem wir die absoluten Häufigkeiten  $n_1, \dots, n_K$  durch die Größe der Stichprobe  $N$  teilen. Die absoluten Häufigkeiten summieren sich stets zu  $N$ , die relativen zu eins:

$$\sum_{k=1}^K f_k = \sum_{k=1}^K \frac{n_k}{N} = \frac{1}{N} \cdot \sum_{k=1}^K n_k = \frac{1}{N} \cdot N = 1$$

Die Merkmalsausprägung mit größter Häufigkeit, also maximalem Wert für  $f_k$  bzw.  $n_k$ , ist der Lageparameter, den wir bereits unter der Bezeichnung Modus kennenlernten. In einem Säulendiagramm entspricht er der höchsten Säule. Die folgende Kennzahl ist hingegen ein Maß dafür, inwieweit alle Säulen eine ähnliche Höhe besitzen.

Die **empirische Shannon-Entropie** einer Stichprobe kategorialer Merkmalsausprägungen der Größe  $N$  mit relativen Häufigkeiten  $f_1, \dots, f_K$  der  $K$  möglichen Ausprägungen ist wie folgt gegeben:

$$H(x) = - \sum_{k=1}^K f_k \log_b(f_k)$$

Dabei ist  $b$  die Basis des Logarithmus, weiterhin vereinbaren wir „ $0 \cdot \log_b 0 = 0$ “: Summanden mit verschwindender relativer Häufigkeit  $f_k$  werden gleich null gesetzt.

Die **normierte Shannon-Entropie** ist wie folgt gegeben:

$$H_{\text{norm}}(x) = \frac{1}{\log_b(K)} \cdot H(x)$$

Übliche Werte für die Basis des Logarithmus sind  $b = 2$  (binärer Logarithmus) oder  $b = e = 2,718\dots$  (natürlicher Logarithmus). Die verwendete Basis des Logarithmus bestimmt die Maßeinheit, in welcher die Entropie angegeben wird; im Falle von  $b = 2$  entspricht diese der allgemein bekannten Informatiuneinheit Bit. Für die Definition der *normierten* Shannon-Entropie spielt die Wahl der Basis bzw. Maßeinheit keine Rolle. Im Folgenden werden wir stets den natürlichen Logarithmus zugrundelegen.

Die normierte Shannon-Entropie variiert für verschiedene Häufigkeitsverteilungen zwischen null und eins. Sie ist minimal, wenn in der Stichprobe effektiv nur eine Merkmalsausprägung vorkommt. Das kann z. B. die erste sein, es gilt dann  $f_1 = 1$  und ansonsten  $f_k = 0$  für alle  $k \in \{2, \dots, K\}$ :

$$H_{\text{norm}}(x) = - \frac{1}{\ln(K)} \sum_{k=1}^K f_k \ln(f_k) = - \frac{1}{\ln(K)} \cdot 1 \cdot \ln(1) = 0$$

Kommen die Merkmalsausprägungen in der Stichprobe mit gleicher Häufigkeit vor, so liegt das andere Extrem einer **Gleichverteilung** vor. Es gilt dann

$f_k = 1/K$  für alle  $k \in \{1, \dots, K\}$  und die Shannon-Entropie ist maximal:

$$\begin{aligned} H_{\text{norm}}(x) &= -\frac{1}{\ln(K)} \sum_{k=1}^K f_k \ln(f_k) = -\frac{1}{\ln(K)} \sum_{k=1}^K \frac{1}{K} \ln\left(\frac{1}{K}\right) \\ &= -\frac{1}{\ln(K)} \cdot K \cdot \frac{1}{K} \cdot (-\ln(K)) = 1 \end{aligned}$$

Die Umkehrung obiger Aussagen gilt ebenfalls, so zeigt etwa eine maximale Entropie stets Gleichverteilung an. Die Shannon-Entropie kann somit als ein Indikator dafür aufgefasst werden, in welchem Maße die Häufigkeiten der Merkmalsausprägungen streuen: Ist sie nahe null, so ist die Mehrheit der Ausprägungen in der Stichprobe durch den Modus des Merkmals bestimmt. Ist die Shannon-Entropie annähernd maximal, so kommen die möglichen Ausprägungen mit vergleichbarer Häufigkeit vor.

**Anwendungsbeispiel.** Die normierte Shannon-Entropie der Geschlechterverteilung in der CDC-Umfrage beträgt:

$$H_{\text{norm}}(\text{Geschlecht}) = -\frac{1}{\log(2)} (0,55 \cdot \log(0,55) + 0,45 \cdot \log(0,45)) \approx 0,993$$

Die Entropie der Verteilung der Einkommenklassen ist geringer, da manche Klassen deutlich häufiger vorkommen als andere. Eine Berechnung ergibt  $H_{\text{norm}}(\text{Einkommen}) \approx 0,893$ .

## 2.5 Assoziationsparameter

Merkmale sind oftmals nicht unabhängig voneinander, es können funktionale Zusammenhänge zwischen ihnen bestehen. Stellen wir uns zum Beispiel eine Firma vor, die Werkstücke rechteckiger Form produziert und – etwa als Qualitätssicherungsmaßnahme – für eine Stichprobe dieser Werkstücke die Größe vermisst. Selbst wenn die hergestellten Teile in der Größe variieren, so gilt für jedes: Die Fläche ist das Produkt seiner Seitenlängen. Das Beispiel erscheint trivial, weil der Zusammenhang aus geometrischen Überlegungen sofort folgt. Allerdings ist dies nicht immer der Fall, es kann auch *verborgene* funktionale Zusammenhänge zwischen Merkmalen geben, die nicht so unmittelbar ersichtlich sind. Zum anderen spiegeln auch sehr starke Assoziationen keine streng deterministische Abhängigkeit zwischen den Beobachtungen wider. Beispielsweise besteht ein durch viele Studien nachgewiesener Zusammenhang zwischen der Inhalation von Tabakrauch und der Entwicklung von Krebserkrankungen [7]. Dennoch erkrankt nicht *jeder* Raucher an Krebs, und manche – wenn auch wenige – Lungenkrebspatienten erweisen sich als Nichtraucher.

**Assoziationsparameter** dienen dazu, bekannte Zusammenhänge zwischen den Merkmalen einer Stichprobe zu bestätigen oder verborgene Zusammenhänge aufzuklären. Dies geschieht durch den Vergleich von Stichprobenlisten

$x = (x_1, \dots, x_N)$  und  $y = (y_1, \dots, y_N)$  der jeweiligen Merkmale. Es werden dabei sinnvollerweise immer nur Stichprobenlisten in derselben Stichprobe verglichen; die entsprechenden Ausprägungen müssen gemeinsam erhoben worden sein und jeweils zu derselben Beobachtung gehören. Wir sagen dazu kurz, dass die Stichprobenlisten **gepaart** sind.

### 2.5.1 Empirische Kovarianz und Korrelation

Die im Folgenden definierten Kennzahlen können verwendet werden, um einen *linearen* funktionalen Zusammenhang zwischen metrischen Merkmalen aufzuzeigen: Haben die Parameter betragsmäßig einen hohen Wert, dann ist der lineare Anstieg eines Werts aus der einen Stichprobenliste stets oder oft mit dem linearen Anstieg oder Abfall des Werts aus der anderen Stichprobenliste verbunden.

Für zwei gepaarte metrische Stichprobenvektoren  $x = (x_1, \dots, x_N)$  und  $y = (y_1, \dots, y_N)$  ist deren **empirische Kovarianz** wie folgt gegeben:

$$s(x, y) = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x}) \cdot (y_n - \bar{y})$$

Der **empirische Korrelationskoeffizient nach Bravais-Pearson** ist:

$$r(x, y) = \frac{s(x, y)}{s(x) \cdot s(y)},$$

wobei  $s(x) \neq 0$  und  $s(y) \neq 0$ .

Die empirische Varianz kann als Kovarianz einer Stichprobe mit sich selbst gedeutet werden: Es gilt nämlich  $s(x, x) = s^2(x)$  für alle  $x \in \mathbb{R}^N$ . Der Korrelationskoeffizient nach Bravais-Pearson wird auch kürzer als Pearson-Korrelation bezeichnet.

Wir wollen nun die Interpretation der Kovarianz bzw. Pearson-Korrelation verdeutlichen. Zu diesem Zweck halten wir zunächst die folgenden Eigenschaften des arithmetischen Mittelwerts  $\mu(\cdot)$  und der empirischen Varianz  $s^2(\cdot)$  fest, die nicht schwer nachzuweisen sind:

$$\begin{aligned}\mu(mx + c) &= m\mu(x) + c, \\ s^2(mx + c) &= m^2 s^2(x)\end{aligned}$$

für alle  $x \in \mathbb{R}^N$  und  $m, c \in \mathbb{R}$ . Für die empirische Standardabweichung gilt mithin  $s(mx + c) = \sqrt{s^2(mx + c)} = \sqrt{m^2 s^2(x)} = |m| \cdot s(x)$ .

Nehmen wir einmal an, dass die Gleichung  $y = mx + c$  mit  $m \neq 0$  gilt: Bei Auftragen der  $y$ -Werte über die  $x$ -Werte in einem Streudiagramm würden diese *genau* auf einer Geraden mit der Steigung  $m$  und dem  $y$ -Achsenabschnitt  $c$

liegen. Es besteht also ein idealer linearer Zusammenhang zwischen den beiden Merkmalen. In diesem Fall ergibt sich für die Kovarianz:

$$\begin{aligned}s(x, y) &= s(x, mx + c) = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x}) \cdot (mx_n + c - m\bar{x} - c) \\&= m \cdot \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x}) \cdot (x_n - \bar{x}) = m \cdot s(x, x) \\&= m \cdot s^2(x)\end{aligned}$$

Damit dann weiterhin für den Korrelationskoeffizienten ( $m \neq 0$ ):

$$\begin{aligned}r(x, y) &= \frac{s(x, mx + c)}{s(x) \cdot s(mx + c)} = \frac{ms^2(x)}{|m|(s(x))^2} \\&= \text{sgn}(m) = \begin{cases} 1 & \text{falls } m > 0 \\ -1 & \text{falls } m < 0 \end{cases}\end{aligned}$$

Steigen die  $y$ -Werte linear mit  $x$  an, so ist die Korrelation  $r(x, y) = +1$ ; fallen sie linear mit  $x$ , so ist die Korrelation  $r(x, y) = -1$ . Dies entspricht einer perfekten positiven bzw. negativen linearen Korrelation beider Merkmale. Wir kommen am Ende von Abschn. 4.5.1 detaillierter auf den Zusammenhang zwischen Pearson-Korrelation und linearer Abhängigkeit zu sprechen.

Neben diesen Extremwerten können die folgenden Interpretationen der Effektstärke für gewisse Größenordnungen des Bravais-Pearson-Koeffizienten angegeben werden [8, S. 79]:

$ r(x, y) $	0,0	0,1	0,3	0,5
Korrelation	keine	schwach	moderat	stark

Tabelle 2.4. Korrelationsstärken

**Anwendungsbeispiel.** Anhand von  $N = 164.798$  männlichen Befragten der CDC-Studie lässt sich ein Pearson-Koeffizient zwischen Body-Mass-Index und Broca-Index ermitteln. Im Streudiagramm in Abb. 2.3 liegen die Datenpunkte näherungsweise auf einer Geraden. Daher lässt sich vermuten, dass eine starke positive Korrelation vorliegt und  $r \approx 1$  gilt. Tatsächlich gilt  $r = 0,995$ . Zwischen Körpergewicht und BMI beträgt der Korrelationskoeffizient nur  $r = 0,872$ , was aber immer noch auf eine sehr starke Korrelation hindeutet. Körpergröße und -gewicht zeigen mit  $r = 0,387$  eine moderate Korrelation. Körpergröße und BMI sind nur unwesentlich miteinander linear korreliert:  $r = -0,094$ .

### 2.5.2 Rangkorrelationskoeffizienten

Die im Folgenden definierten Parameter können dazu verwendet werden, einen funktionalen Zusammenhang zwischen metrischen Merkmalen aufzuzeigen, der monoton ist: Der Anstieg eines Werts aus der einen Stichprobenliste ist stets oder oft mit dem Anstieg oder Abfall des Werts aus der anderen Stichprobenliste verbunden. Dieser Zusammenhang muss – im Gegensatz zum Anwendungsbereich der Pearson-Korrelation – nicht zwangsläufig linear sein.

Sei  $x = (x_1, \dots, x_N)$  eine Stichprobe eines metrischen oder ordinalen Merkmals. Wir können die Beobachtungen absteigend sortieren:

$$x_{\iota(1)} \geq x_{\iota(2)} \geq \dots \geq x_{\iota(N)},$$

mit einer Permutation  $\iota: \{1, \dots, N\} \rightarrow \{1, \dots, N\}$ . Die Zahlen  $\text{rg}(x) = (\iota(1), \iota(2), \dots, \iota(N))$  werden die **Ränge** der Beobachtungen genannt: Die Beobachtung mit dem größten Wert erhält den Rang 1, die Beobachtung mit dem kleinsten Wert den Rang  $N$ . Die Konvention, nach der der kleinste Wert im Rang an erster Stelle steht, ist ebenfalls üblich.

Wenn keine Merkmalsausprägung in der Stichprobe mehrmals vorkommt, ist der Rang auf diese Weise eindeutig definiert. Andernfalls kann Eindeutigkeit erzwungen werden, indem wir fordern, dass die zuerst in der Stichprobe vorkommende Ausprägung kleineren Rang hat.

Allerdings haben auch dann identische Merkmalsausprägungen immer noch verschiedenen Rang, was in vielen Fällen keine wünschenswerte Eigenschaft ist. Daher korrigieren wir die Rangzahlen wie folgt: Identischen Beobachtungen wird der arithmetische Mittelwert ihrer Ränge zugeordnet, das Ergebnis bezeichnen wir mit  $\bar{\text{rg}}(x)$ .

Eine weitere nützliche Definition ist der **Prozentrang** einer Beobachtung  $x_n$ :

$$\%-\text{rg}(x_n) = \frac{1}{N} \cdot |\{m \in \{1, \dots, N\} | x_m \leq x_n\}|$$

Der größte Wert der Stichprobe erhält also immer den maximalen Prozentrang von 100 %.

Wir betrachten ein (fiktives) Beispiel zur Illustration der unterschiedlichen Definitionen eines Rangs. Fünf Schülerinnen und Schüler erhalten am Ende des Schuljahres die folgenden Schulnoten:  $x = (\mathbf{2}, \mathbf{1}, \mathbf{2}, \mathbf{1}, \mathbf{3})$ . Wir interpretieren diese Sequenz als Ausprägungen eines ordinalen Merkmals mit der Ordnungsrelation  $\mathbf{6} < \mathbf{5} < \dots < \mathbf{1}$ . Die entsprechenden Ränge sind dann wie folgt gegeben:

$n$	$x_n$	$\text{rg}(x_n)$	$\bar{\text{rg}}(x_n)$	%-rg( $x_n$ )
1	<b>2</b>	3	3,5	60 %
2	<b>1</b>	1	1,5	100 %
3	<b>2</b>	4	3,5	60 %
4	<b>1</b>	2	1,5	100 %
5	<b>3</b>	5	5	20 %

**Tabelle 2.5.** Rangstatistiken, Beispiel „Schulnoten“

Seien zwei gepaarte metrische oder ordinale Stichproben  $x = (x_1, \dots, x_N)$  und  $y = (y_1, \dots, y_N)$  gegeben.

Der **Rangkorrelationskoeffizient nach Spearman** ist wie folgt definiert:

$$\rho(x, y) = r(\bar{\text{rg}}(x), \bar{\text{rg}}(y))$$

Dabei ist  $r(\cdot, \cdot)$  ist der Korrelationskoeffizient nach Bravais-Pearson.

Der **Rangkorrelationskoeffizient nach Kendall** ist:

$$\tau(x, y) = \frac{1}{N(N-1)} \sum_{k,l \in \{1, \dots, N\}} \text{sgn}(x_l - x_k) \cdot \text{sgn}(y_l - y_k)$$

Um die Charakteristiken dieser Assoziationsmaße aufzuzeigen, berechnen wir sie für eine Sammlung von vier synthetischen Datensätzen, die **Anscombe-Quartett** genannt wird und in Abb. 2.9 dargestellt ist [9].

Eine Berechnung der einzelnen Korrelationskennzahlen für die Stichprobenvektoren  $x^{(i)}$  und  $y^{(i)}$ ,  $i = 1, 2, 3, 4$ , führt auf die folgenden Ergebnisse:

$i$	Pearson	Spearman	Kendall
1	0,82	0,82	0,63
2	0,82	0,69	0,56
3	0,82	0,99	0,96
4	0,82	0,50	0,18

**Tabelle 2.6.** Korrelationskennzahlen für das Anscombe-Quartett

Zunächst beobachten wir, dass für alle Datensätze – obwohl augenscheinlich von sehr unterschiedlicher Verteilung – der Korrelationskoeffizient nach Bravais-Pearson identisch ist. Die Rangkorrelationskoeffizienten für den zweiten Datensatz sind hingegen geringer, da die  $y$ -Werte nicht monoton mit den  $x$ -Werten wachsen oder fallen. Darüber hinaus ist die Rangkorrelation gegenüber einzelnen Ausreißern robuster, wie das dritte Beispiel zeigt. Das vierte Beispiel offenbart aber auch eine Schwierigkeit: Die Mehrheit der  $x$ -Werte ist hier identisch oder nahezu identisch. Jede kleine Variation in diesen Werten führte auf beliebige Ränge und folglich auf ebenso beliebige Rangkorrelationsmaße.

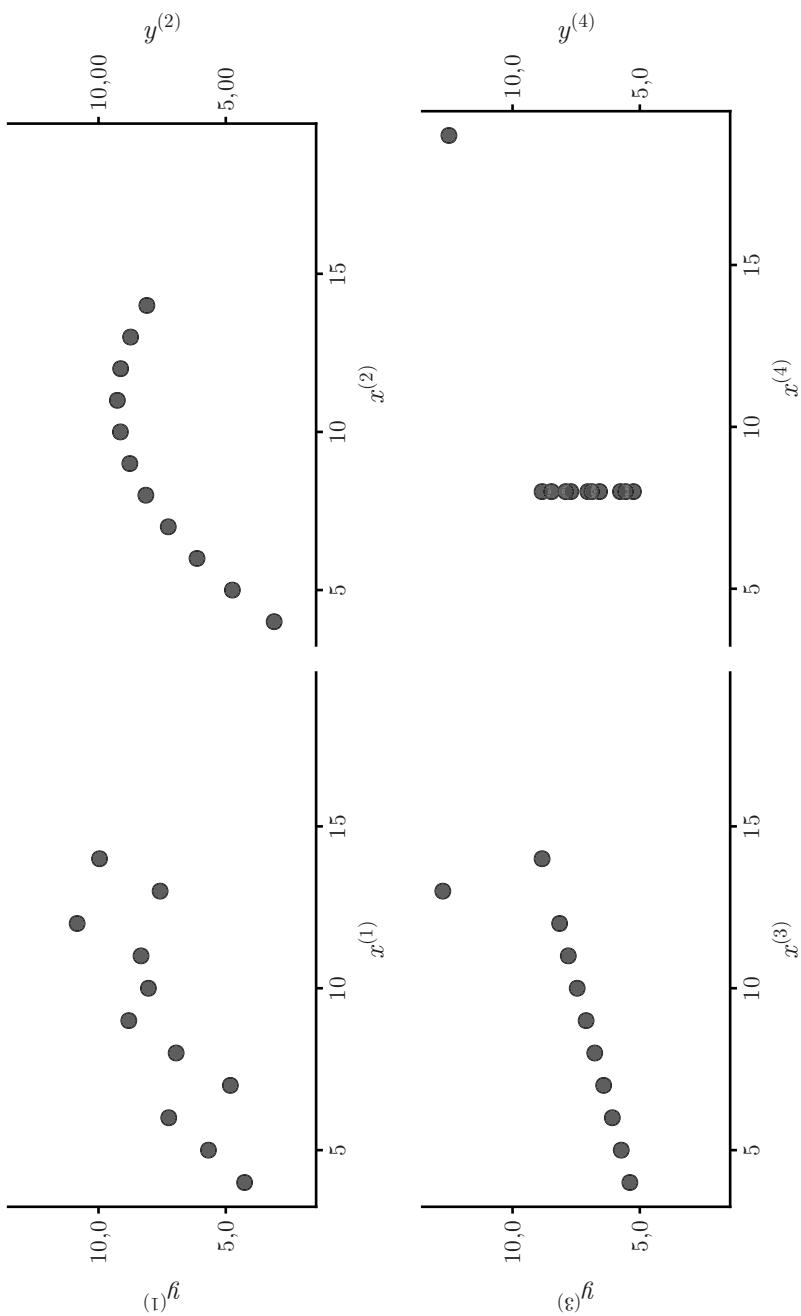


Abb. 2.9. Anscombe-Quartett: vier verschiedene Datensätze mit identischer Pearson-Korrelation

### 2.5.3 Transinformation und Jaccard-Koeffizient

Eine Teilgesamtheit, in der eine Ausprägung eines kategorialen Merkmals mit hoher Häufigkeit vorkommt, zeigt vielleicht zugleich eine signifikant hohe oder niedrige Häufigkeit der Ausprägung eines anderen Merkmals. Im Folgenden werden Assoziationsparameter vorgestellt, die helfen können, solche Zusammenhänge aufzudecken oder zu bestätigen.

Seien  $x = (x_1, \dots, x_N)$  und  $y = (y_1, \dots, y_N)$  gepaarte Stichproben zweier kategorialer Merkmale mit  $K$  bzw.  $L$  möglichen Ausprägungen.

Seien weiterhin  $0 \leq f_{kl} \leq 1$  mit  $k \in \{1, \dots, K\}$  und  $l \in \{1, \dots, L\}$  die relativen Häufigkeiten, mit denen die Merkmalsausprägungen gemeinsam auftreten: Es ist also  $f_{kl}$  die relative Häufigkeit, mit der  $x_n$  die  $k$ -te Merkmalsausprägung und  $y_n$  zugleich die  $l$ -te Merkmalsausprägung annimmt.

Die **empirische gemeinsame Shannon-Entropie** ist dann wie folgt gegeben (wie üblich vereinbaren wir „ $0 \cdot \ln 0 = 0$ “):

$$H(x, y) = - \sum_{k=1}^K \sum_{l=1}^L f_{kl} \cdot \ln(f_{kl})$$

Die **empirische Transinformation** (im Englischen: *mutual information*, MI) der beiden Merkmale ist die folgende Größe:

$$\text{MI}(x, y) = \sum_{k=1}^K \sum_{l=1}^L f_{kl} \cdot \ln \left( \frac{f_{kl}}{f_{k\bullet} \cdot f_{\bullet l}} \right)$$

mit

$$f_{k\bullet} = \sum_{j=1}^L f_{kj}, \quad f_{\bullet l} = \sum_{i=1}^K f_{il}$$

für  $k \in \{1, \dots, K\}$  und  $l \in \{1, \dots, L\}$ .

Die **normierte Transinformation** ermittelt sich wie folgt:

$$\text{MI}_{\text{norm}}(x, y) = \frac{\text{MI}(x, y)}{H(x, y)}$$

für  $H(x, y) > 0$ , andernfalls kann  $\text{MI}_{\text{norm}}(x, y) = 1$  gesetzt werden.

Die Größen  $f_{1\bullet}, \dots, f_{K\bullet}$  und  $f_{\bullet 1}, \dots, f_{\bullet L}$  sind gerade die relativen Häufigkeiten der einzelnen Merkmale, die wir hier als **Randsummen** über die **gemeinsamen Häufigkeiten** dargestellt haben. Es ist üblich, die gemeinsamen Häufigkeiten und **Randhäufigkeiten** in einer sogenannten **Kontingenztafel** zusammenzufassen:

$f_{11}$	$f_{12}$	$\dots$	$f_{1L}$	$  f_{1\bullet}$
$f_{21}$	$f_{22}$	$\dots$	$f_{2L}$	$  f_{2\bullet}$
$\vdots$			$\vdots$	$\vdots$
$f_{K1}$	$f_{K2}$	$\dots$	$f_{KL}$	$  f_{K\bullet}$
$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet L}$	

Weiterhin gilt zu beachten, dass sich alle Häufigkeiten zu 100 % summieren:

$$\sum_{k=1}^K f_{k\bullet} = \sum_{l=1}^L f_{\bullet l} = \sum_{k=1}^K \sum_{l=1}^L f_{kl} = 1$$

Die Entropie oder Transinformation einer Stichprobe  $x$  bezüglich ihrer selbst ist gerade ihre Shannon-Entropie, d. h., es gilt  $MI(x, x) = H(x, x) = H(x)$ . In diesem Fall haben wir nämlich  $f_{kl} = 0$  für  $k \neq l$  und  $f_{kl} = f_{k\bullet} = f_{\bullet l}$  für  $k = l$ . Mithin gilt  $MI_{\text{norm}}(x, x) = 1$ .

Es kann gezeigt werden, dass die normierte Transinformation allein Werte zwischen null und eins annimmt. Wie eben angedeutet, entspricht ein Wert von eins der maximal möglichen Assoziation. Eine verschwindende Transinformation entspricht hingegen der Bedingung  $f_{kl} = f_{k\bullet} \cdot f_{\bullet l}$ , d. h., die Häufigkeit gemeinsamen Auftretens der Merkmalsausprägungen ist das Produkt der Häufigkeit des Auftretens der einzelnen Ausprägungen. Interpretieren wir die Häufigkeiten als empirische Wahrscheinlichkeiten, so werden wir später sehen (in Abschn. 3.1.1), dass diese Bedingung eine Aussage über die Unabhängigkeit der Merkmale darstellt: Wie oft die  $k$ -te Ausprägung des einen Merkmals vorkommt, hat keinen Einfluss darauf, wie häufig die  $l$ -te Ausprägung des anderen Merkmals ist.

**Anwendungsbeispiel.** Im Folgenden ist eine auf der ALLBUS-Umfrage basierende Kontingenztafel angeführt. Diese stellt die Präferenz für eine politische Partei der Beantwortung folgender Frage gegenüber: „Was halten Sie von folgender Aussage: Der Zuzug von Flüchtlingen sollte unterbunden werden.“

Die Transinformation zwischen beiden Merkmalen beträgt  $MI_{\text{norm}} = 0,035$ . Das hört sich zunächst einmal nach einem kleinen Wert an, daher hier ein Vergleich: Die Transinformation zwischen der Parteienpräferenz und dem Vorhandensein einer Gegensprechanlage im Haus der Befragten oder des Befragten ist wesentlich geringer und beträgt  $MI_{\text{norm}} = 0,0026$ .

Alle Angaben sind in Prozent:

Partei \ Zustimmung	gar nicht	eher nicht	weder noch	eher	voll	$\Sigma$
<b>CDU/CSU</b>	5,9	13,0	8,0	8,3	3,8	39,0
<b>SPD</b>	5,9	9,3	4,6	4,7	2,3	26,9
<b>Die Grünen</b>	6,4	5,4	1,5	0,5	0,2	13,9
<b>Die Linke</b>	2,7	3,0	1,4	1,2	0,7	9,0
<b>AfD</b>	0,2	0,3	0,6	1,1	3,7	5,8
<b>FDP</b>	1,0	1,1	0,9	1,1	0,6	4,7
<b>andere</b>	0,1	0,3	0,1	0,1	0,2	0,7
$\Sigma$	22,1	32,3	17,1	17,0	11,5	

**Tabelle 2.7.** Kontingenztafel von Parteipräferenz und Meinung zur Notwendigkeit von Migrationskontrolle

Seien  $x = (x_1, \dots, x_N)$  und  $y = (y_1, \dots, y_N)$  gepaarte Stichproben zweier binärer Merkmale:  $x_n, y_n \in \{0, 1\}$  für alle  $n \in \{1, \dots, N\}$ .

Sei  $f_{11}$  die Häufigkeit des gemeinsamen Auftretens positiver Merkmalsausprägung, also die Häufigkeit von  $x_n = y_n = 1$ . Weiter sei  $f_{1\bullet}$  die Häufigkeit von  $x_n = 1$  und  $f_{\bullet 1}$  die Häufigkeit von  $y_n = 1$ .

Der **Jaccard-Koeffizient** ist dann wie folgt gegeben:

$$J(x, y) = \frac{f_{11}}{f_{\bullet 1} + f_{1\bullet} - f_{11}},$$

falls nicht gerade  $x = y = 0$  gilt; in diesem Fall ist der Jaccard-Koeffizient nicht definiert.

In Analogie zur Transinformation haben wir den Jaccard-Koeffizienten über die Häufigkeiten definiert. Dabei ist es unerheblich, ob die relativen oder absoluten Häufigkeiten verwendet werden; die Stichprobengröße kürzt sich heraus. Der Jaccard-Koeffizient kann aber auch wie folgt geschrieben werden:

$$J(x, y) = \frac{|\{d | x_d = 1 \text{ und } y_d = 1\}|}{|\{d | x_d = 1 \text{ oder } y_d = 1\}|}$$

Eine weitere, noch einfachere Schreibweise ergibt sich aus der Interpretation eines binären Merkmals als eine Teilgesamtheit. Zur Erinnerung: Ein Merkmalsträger gehört zu dieser Teilgesamtheit genau dann, wenn er eine positive Merkmalsausprägung aufweist. Der Jaccard-Koeffizient zweier Teilgesamtheiten  $S$  und  $T$  ergibt sich somit wie folgt:

$$J(S, T) = \frac{|S \cap T|}{|S \cup T|}$$

Eine alternative, in der Praxis oft gebrauchte Formel ist die folgende:

$$J(S, T) = \frac{|S \cap T|}{|S| + |T| - |S \cap T|}$$

Der Jaccard-Koeffizient nimmt Werte zwischen null und eins an, wobei  $J(S, T) = 0$  einer verschwindenden Überschneidung entspricht:  $S \cap T = \emptyset$ . Es gilt genau dann  $J(S, T) = 1$ , wenn beide Teilgesamtheiten identisch sind:  $S = T$ .

**Anwendungsbeispiel.** Bei der deutschen Bundestagswahl 2013 haben 16,89 Millionen Wählerinnen und Wähler, welche bis 2017 nicht weggezogen oder verstorben waren, ihre Stimme für die Union aus CDU und CSU abgegeben. Ohne Neuwähler/-innen und Zuzüge waren es 2017 bei der darauffolgenden Wahl 14,77 Millionen. 11,09 Millionen Wähler/-innen haben in beiden Jahren ihre Stimme der Union gegeben [10]. Der Jaccard-Koeffizient zwischen Wähler/-innen der Union 2013 und solchen 2017 ist somit:

$$J(\text{CDU/CSU 2013}, \text{CDU/CSU 2017}) = \frac{11,09}{16,89 + 14,77 - 11,09} \approx 54\%$$

Für weitere Parteien sowie die Kohorte der Nichtwähler/-innen ergibt sich der so berechnete Jaccard-Koeffizient wie folgt:

Partei	Jaccard-Koeff. Wählerschaft 2013/2017
CDU/CSU	54 %
SPD	43 %
Die Grünen	34 %
Die Linke	36 %
AfD	23 %
FDP	22 %
Nichtwähler/-innen	48 %

**Tabelle 2.8.** Jaccard-Koeffizient als Kennzahl für Wählerbindung

Wir können diesen Wert für jede Partei als eine Kennzahl interpretieren, die neben konstantem Stimmenanteil auch auf konstante Wählerbindung hindeutet: Läge der Wert bei 0 %, würde dies bedeuten, dass 2017 ausschließlich Wechselwähler/-innen die Partei wählten. Ein Wert von 100 % würde bedeuten, dass die Wählerschaft 2013 und 2017 identisch war; es fand weder eine Abwanderung noch eine Zuwanderung statt.

## Quellen

- [1] CDC Population Health Surveillance Branch. *Behavioral Risk Factor Surveillance System (BRFSS) Survey Data 2018*. Aufgerufen am 01. Feb. 2020. URL: <https://www.cdc.gov/brfss/>.
- [2] GESIS-Leibniz-Institut für Sozialwissenschaften. *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2018*. 2019. doi: [10.4232/1.13250](https://doi.org/10.4232/1.13250).
- [3] Harri Siirtola. „The Cost of Pie Charts“. In: *23rd International Conference Information Visualisation (IV)*. 2019, S. 151–156. doi: [10.1109/IV.2019.00034](https://doi.org/10.1109/IV.2019.00034).
- [4] infratest dimap Gesellschaft für Trend- und Wahlforschung mbH. *Ergebnisse der Sonntagsfrage (bundesweit) vom 16. Oktober 2020*. Aufgerufen am 18. Okt. 2020. URL: <https://www.infratest-dimap.de/umfragen-analysen/bundesweit/sonntagsfrage/>.
- [5] Paul T. von Hippel. „Mean, Median, and Skew: Correcting a Textbook Rule“. In: *Journal of Statistics Education* 13.2 (Jan. 2005). doi: [10.1080/10691898.2005.11910556](https://doi.org/10.1080/10691898.2005.11910556).
- [6] Ronald Aylmer Fisher. „The use of multiple measurements in taxonomic problems“. In: *Annals of Eugenics* 7.2 (Sep. 1936), S. 179–188. doi: [10.1111/j.1469-1809.1936.tb02137.x](https://doi.org/10.1111/j.1469-1809.1936.tb02137.x).
- [7] Robert N. Proctor. „Tobacco and the global lung cancer epidemic“. In: *Nature Reviews Cancer* 1.1 (Okt. 2001), S. 82–86. doi: [10.1038/35094091](https://doi.org/10.1038/35094091).
- [8] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. 2. Aufl. New Jersey, USA: Lawrence Earlbaum Associates, 1988. ISBN: 0-8058-0283-5.
- [9] Francis John Anscombe. „Graphs in Statistical Analysis“. In: *The American Statistician* 27.1 (Feb. 1973), S. 17–21.
- [10] infratest dimap Gesellschaft für Trend- und Wahlforschung mbH. *Bundestagswahl 2017 Deutschland Wählerwanderungen*. Aufgerufen am 10. Juni 2020. URL: <https://wahl.tagesschau.de/wahlen/2017-09-24-BT-DE/analyse-wanderung.shtml>.

## **Teil II**

---

### **Stochastik**



## Wahrscheinlichkeitstheorie

Mittels der deskriptiven Statistik können wir detaillierte Aussagen über die Häufigkeitsverteilung einer Stichprobe anstellen. Eine wesentliche Erkenntnis ist, dass diese Aussagen oft Rückschlüsse auf die Grundgesamtheit zulassen. Beispielsweise können wir allein aufgrund der Datenlage folgern, dass ein Mensch nicht auf eine Größe von drei Metern heranwachsen kann.

Eine etwas praxisnähere Prognose ist die politische Wahlumfrage: Einer Stichprobe von Wahlberechtigten wird bereits vor der eigentlichen Wahl dazu befragt, welche Partei sie wählen wird. Die Hoffnung ist, dass die relative Häufigkeit, mit der eine Partei in der Stichprobe gewählt würde, am Wahltag von *allen* Wahlberechtigten näherungsweise mit derselben Häufigkeit gewählt wird.

Darüber hinaus kann die relative Häufigkeit auch als Kennzahl für die Wahrscheinlichkeit interpretiert werden, einen/eine Wähler/-in einer Partei in der Grundgesamtheit bei zufälliger Auswahl anzutreffen. Aus diesem Grund wird die relative Häufigkeit auch als **empirische Wahrscheinlichkeit** bezeichnet.

Eine Aufgabe der **Stochastik** ist es, ein solches Vorgehen mathematisch im Detail zu begründen. Das Gebiet kann in zwei Untergebiete eingeteilt werden:

- Die **Wahrscheinlichkeitstheorie**, welche Gegenstand dieses Kapitels ist, befasst sich mit der mathematischen Definition und Untersuchung des Begriffes der Wahrscheinlichkeit. Ein zentraler Untersuchungsgegenstand sind Maßzahlen, deren Werte nicht exakt vorgegeben, sondern Unsicherheiten unterworfen sind. Genauer bedeutet das: Für eine solche Maßzahl kann nur eine Wahrscheinlichkeit dafür angegeben werden, dass sie Werte in einem bestimmten Bereich annimmt. Solche Maßzahlen werden **Zufallsvariablen** genannt.
- Die **Inferenzstatistik**, welche im nächsten Kapitel behandelt wird, baut auf der deskriptiven Statistik und der Wahrscheinlichkeitstheorie auf. Ihr liegt die Annahme zugrunde, dass statistische Beobachtungen und Kennzahlen wie Häufigkeiten, Mittelwerte usw. gerade **Realisierungen** von Zufallsvariablen sind. Sie untersucht, inwieweit Charakteristiken solcher Zu-

fallsvariablen anhand von Stichproben geschätzt werden können. Die Inferenzstatistik ermöglicht es insbesondere, die Unsicherheiten einer solchen Schätzung zu quantifizieren.

Auch der Bereich des maschinellen Lernens basiert wesentlich auf stochastischen Überlegungen, z. B. bei der Klassifikation: So quantifizieren Wahrscheinlichkeiten die Häufigkeit des Vorkommens einer Klasse („Anteil der Fotografien im Trainingsdatensatz, die eine Katze zeigen“) oder die Unsicherheit im Klassifikationsergebnis („Mit hoher Wahrscheinlichkeit zeigt dieses Foto im Testdatensatz eine Katze“).

### 3.1 Wahrscheinlichkeitsmaße

Als Beispiele für Vorgänge mit unsicherem Ausgang werden klassischerweise **Zufallsexperimente** herangezogen. Solche Experimente können – zumindest prinzipiell – unter identischen Bedingungen beliebig oft wiederholt werden. Ein Beispiel für ein solches Zufallsexperiment ist das Werfen einer Münze: Am Ende zeigt die Seite mit Bild („Kopf“) nach oben, oder die Seite mit „Zahl“. Ein weiteres Beispiel ist der Wurf eines sechseitigen Spielwürfels.

Bei diesen Experimenten können wir uns die möglichen eindeutigen Ausgänge zu einer Menge zusammengefasst vorstellen, dem **Ergebnisraum** oder **Stichprobenraum**  $\Omega = \{\text{Zahl}, \text{Kopf}\}$  bzw.  $\Omega = \{\square, \blacksquare, \boxdot, \blacksquare\boxdot, \boxtimes, \blacksquare\boxtimes\}$ .

Die relative Häufigkeit, mit der ein Ergebnis bei einem (hinreichend oft) wiederholt durchgeführten Zufallsexperiment auftritt, können wir als die Wahrscheinlichkeit für das Auftreten dieses Ereignisses interpretieren. Diese Interpretation wird **frequentistischer Wahrscheinlichkeitsbegriff** genannt und ist in den Naturwissenschaften verbreitet, in denen reproduzierbare Experimente wesentlich für den empirischen Erkenntnisgewinn sind. Beispiel einer frequentistischen Aussage: „Wir können beobachten, dass rund 50 % aller Atomkerne des radioaktiven Isotops Jod-131 nach acht Tagen zerfallen sind. Folglich liegt die Wahrscheinlichkeit, dass ein einzelner solcher Atomkern nach dieser Zeit zerfallen ist, bei 50 %.“

Von dieser Interpretation kann der **Bayes'sche Wahrscheinlichkeitsbegriff** abgegrenzt werden. Dieser sieht Wahrscheinlichkeit als ein (subjektives) Maß für die Plausibilität eines Ereignisses an, ohne dass dieses ein Ergebnis eines beliebig reproduzierbaren Experiments sein muss. Beispielsweise wäre es somit auch sinnvoll zu sagen: „Die Regenwahrscheinlichkeit für den morgigen Tag beträgt 75 %.“ Im Vordergrund stehen hier das Schließen und Handeln unter Unsicherheit mit dem Ziel einer Maximierung des erwarteten Nutzens („Nehme ich einen Regenschirm mit oder nicht?“).

Es gibt noch weitere Ansätze, dem Begriff der Wahrscheinlichkeit eine universelle Bedeutung zukommen zu lassen [1]. Unabhängig davon kann dieser mathematisch klar definiert werden. Die Grundidee ist dabei, Teilmengen des Ergebnisraums als potenzielle Ausgänge eines zufälligen Vorgangs oder einer

zufälligen Auswahl aus einer statistischen Grundgesamtheit aufzufassen. Ein Maß für die Wahrscheinlichkeit, dass eines der Ergebnisse realisiert wird, ist dann eine Zahl zwischen null und eins.

Ein **Wahrscheinlichkeitsmaß** ist eine Abbildung  $\Pr(\cdot)$ , die bestimmten Teilmengen  $A \subseteq \Omega$  eines Ergebnisraums  $\Omega$  eine nichtnegative reelle Zahl zuordnet und die folgenden Eigenschaften hat:

- (1)  $\Pr(\emptyset) = 0$  und  $\Pr(\Omega) = 1$
- (2) Für jede endliche oder abzählbar unendliche Familie paarweise disjunkter Mengen  $(A_i)_{i \in I}$  gilt:

$$\Pr\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} \Pr(A_i)$$

Obige Bedingungen bilden die Grundlage für die **Axiomatisierung der Wahrscheinlichkeitsrechnung nach Kolmogoroff**. Eine genauere mathematische Analyse zeigt, dass es nicht immer möglich oder sinnvoll ist, jeder beliebigen Teilmenge von  $\Omega$  eine wohldefinierte Wahrscheinlichkeit zuzuordnen. Teilmengen des Ergebnisraums, denen jedoch eine Wahrscheinlichkeit zugeordnet werden kann, heißen **Ereignisse** oder werden **messbar** genannt. Ereignisse, die genau ein Element enthalten, werden als **Elementarereignisse** bezeichnet.

Als konkretes Beispiel wollen wir ein sinnvolles Wahrscheinlichkeitsmaß konstruieren, mit dem ein Würfelspiel beschrieben werden kann. Wenn ein sechsseitiger Würfel geworfen wird, so soll es der Spielerin oder dem Spieler unmöglich sein vorherzusagen, welche der sechs möglichen Augenzahlen fällt. Dennoch kann eingeschätzt werden, wie wahrscheinlich es ist, dass eine bestimmte Augenzahl fällt. Ohne weitere Informationen muss sie oder er davon ausgehen, dass jedes mögliche Ergebnis gleichwahrscheinlich ist:

$$\Pr(\{\omega\}) = \frac{1}{6}$$

für alle  $\omega \in \Omega = \{\square, \blacksquare, \boxtimes, \blacksquare, \boxtimes, \blacksquare\}$ . Die Wahrscheinlichkeiten aller übrigen Ereignisse ergeben sich durch Anwendung der obigen Rechenregeln. Beispielsweise ist die Wahrscheinlichkeit, eine gerade Augenzahl zu werfen:

$$\begin{aligned} \Pr(\{\omega | \omega \text{ entspricht einer geraden Augenzahl}\}) &= \\ \Pr(\{\blacksquare, \boxtimes, \blacksquare\}) &= \\ \Pr(\{\blacksquare\} \cup \{\boxtimes\} \cup \{\blacksquare\}) &= \\ \Pr(\{\blacksquare\}) + \Pr(\{\boxtimes\}) + \Pr(\{\blacksquare\}) &= \\ \frac{1}{6} + \frac{1}{6} + \frac{1}{6} &= \frac{1}{2} \end{aligned}$$

In Ermangelung besseren Wissens haben wir hier jedem Elementarereignis  $\{\omega\} \subset \Omega$  dieselbe Wahrscheinlichkeit zugeordnet. Im Allgemeinen berufen wir

uns bei dieser Vorgehensweise auf das **Indifferenzprinzip**: Wenn keine weiteren Informationen vorliegen, sollten die mögliche Ergebnisse als **gleichverteilt** angenommen werden sollten.

Für endliche Ergebnisräume mit Gleichverteilung gilt  $\Pr(\{\omega\}) = \frac{1}{|\Omega|}$ , woraus unmittelbar die **Laplace'sche Formel** folgt:

$$\Pr(A) = \frac{|A|}{|\Omega|}$$

für alle Ereignisse  $A \subseteq \Omega$ . Mit dieser Formel lässt sich die Wahrscheinlichkeit in obigem Beispiel auch wie folgt ermitteln:

$$\Pr(\{\boxdot, \boxtimes, \boxbl\}) = \frac{|\{\boxdot, \boxtimes, \boxbl\}|}{|\{\boxdot, \boxsquare, \boxtimes, \boxbl, \boxbl\}|} = \frac{3}{6} = \frac{1}{2}$$

Ein weiteres Beispiel: Eine Dartspielerin wirft mit einem Pfeil auf eine Dartscheibe mit Radius  $R > 0$ . Eine sinnvolle mathematische Modellierung dieser Situation ist über den Ergebnisraum  $\Omega = \{(\omega_1, \omega_2) \in \mathbb{R}^2 | (\omega_1)^2 + (\omega_2)^2 \leq R^2\}$  gegeben, also einer Kreisscheibe in der Ebene. Jeder Punkt in der Kreisschreibe entspricht einem möglichen Treffer des Dartpfeils.

Ziel der Spielerin ist es, den Pfeil so zu werfen, dass er möglichst nah am Mittelpunkt  $(0, 0)$  landet. Die Einschläge einer perfekten Spielerin würden folgendem Wahrscheinlichkeitsmaß genügen, welches **Dirac-Maß** genannt wird:

$$\Pr_0(A) = \begin{cases} 1 & \text{falls } (0, 0) \in A \\ 0 & \text{sonst} \end{cases}$$

Für Elementarereignisse  $\{\omega\} \in \Omega$  gilt insbesondere:

$$\Pr_0(\{\omega\}) = \begin{cases} 1 & \text{falls } \omega = (0, 0) \\ 0 & \text{sonst} \end{cases}$$

Die perfekte Spielerin trifft also mit Wahrscheinlichkeit 1 den Mittelpunkt.

Eine Spielerin, die zwar die Dartscheibe trifft, ansonsten aber keine Kontrolle über das Ziel des Pfeils hat, entspricht dem anderen Extrem, einer Gleichverteilung:

$$\Pr_R(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{\pi R^2}$$

Dabei haben wir die Fläche eines Bereichs  $B \subseteq \Omega$  mit  $|B|$  bezeichnet. Es gilt zu beachten, dass eine Vorhersage darüber, ob ein einzelner Punkt auf der Dartscheibe getroffen wird, nicht mehr sinnvoll ist: Es gilt nämlich  $\Pr_R(\{\omega\}) = 0$  für alle  $\omega \in \Omega$ . Mithin kann das Wahrscheinlichkeitsmaß nicht über dessen Wert für Elementarereignisse definiert werden, wie das bei endlichen Ergebnisräumen noch möglich ist.

Eine realistischere Situation als die vorhergehenden ist vielleicht die, dass die Spielerin nicht perfekt, aber doch sehr gut ist und daher mit 90-prozentiger Wahrscheinlichkeit im „Bull’s Eye“  $B$  mit Radius  $r$ ,  $0 < r < R$ , trifft:

$$\Pr_r(A) = 0,9 \cdot \frac{|A \cap B(r)|}{|B(r)|} + 0,1 \cdot \frac{|A \cap (\Omega \setminus B(r))|}{|\Omega \setminus B(r)|}$$

mit  $B(r) = \{(\omega_1, \omega_2) \in \mathbb{R}^2 | (\omega_1)^2 + (\omega_2)^2 \leq r^2\}$ .

**Rechenregeln für Wahrscheinlichkeiten.** Für ein festes Wahrscheinlichkeitsmaß  $\Pr(\cdot)$  gilt, wobei  $A, B \subseteq \Omega$  beliebige Ereignisse sind:

- (1)  $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- (2)  $\Pr(A \setminus B) = \Pr(A \cup B) - \Pr(B)$ , insbesondere  $\Pr(\Omega \setminus A) = 1 - \Pr(A)$

Das Ereignis  $\Omega \setminus A$  wird das zu  $A$  **komplementäre Ereignis** genannt. Für dieses wollen wir auch  $\neg A$  schreiben.

Eine Beispielrechnung zu Punkt (1): Die Wahrscheinlichkeit, eine gerade Augenzahl oder eine durch drei teilbare Augenzahl ( $A = \{\square, \blacksquare, \blacksquare\}$  oder  $B = \{\blacksquare, \blacksquare\}$ ) zu würfeln, beträgt

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &= \frac{3}{6} + \frac{2}{6} - \Pr(\{\blacksquare\}) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} \\ &= \frac{2}{3} \approx 0,67\end{aligned}$$

Natürlich können wir das in diesem Fall auch leicht direkt ausrechnen:

$$\Pr(A \cup B) = \Pr(\{\square, \blacksquare, \blacksquare, \blacksquare\}) = \frac{4}{6} = \frac{2}{3}$$

Zu (2): Die Wahrscheinlichkeit, *keine* Augenzahl von zwei zu würfeln, beträgt

$$\Pr(\neg C) = \Pr(\Omega \setminus C) = 1 - \frac{1}{6} = \frac{5}{6} \approx 0,83,$$

wobei  $C = \{\square\}$ .

**Abschätzungen für Wahrscheinlichkeiten.** Für beliebige Ereignisse  $A, B \subseteq \Omega$  gilt:

- (1)  $\Pr(A) \leq \Pr(B)$  falls  $A \subseteq B$
- (2)  $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$
- (3)  $\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$

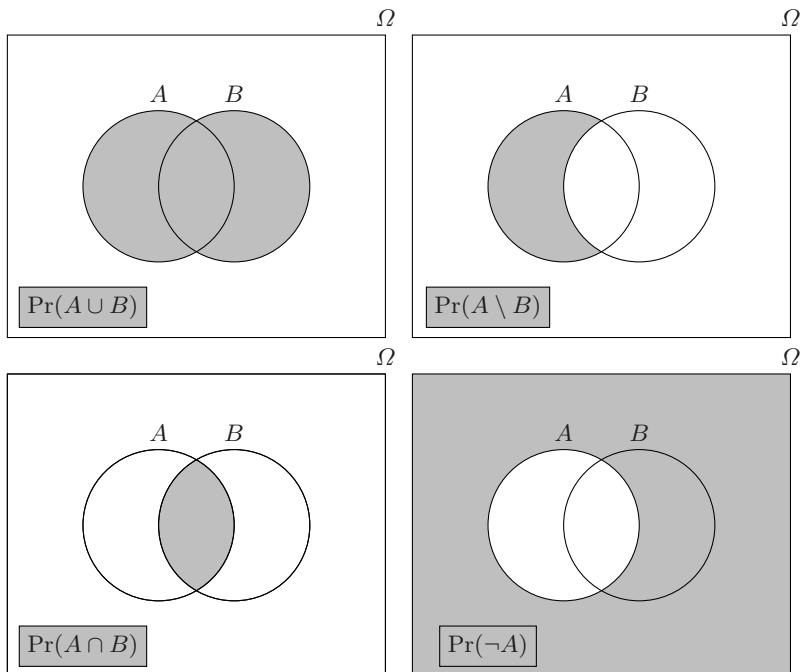
Die zweite Abschätzung ist nur für kleine Wahrscheinlichkeiten oder Häufigkeiten nützlich (da andernfalls die rechte Seite größer 1 werden kann), die dritte nur für große (da andernfalls die rechte Seite kleiner 0 werden kann).

**Anwendungsbeispiel.** Zu (1): Der Anteil der deutschen Haushalte mit höchstens zwei eigenen Pkws ist sicher mindestens so groß wie der Anteil der Haushalte mit höchstens einem Pkw.

Zu (2): In 19 % aller deutschen Haushalte lebt ein Hund, in 23 % eine Katze (Stand 2018 [2]). Daraus können wir folgern, dass in höchstens 42 % der Haushalte ein Hund oder eine Katze leben.

Zu (3): In Deutschland verfügen etwa 77 % aller Haushalte über ein eigenes Auto und 97 % über ein Mobiltelefon (Stand 2019 [3]). Daraus können wir schließen, dass mindestens 74 % sowohl über ein eigenes Auto als auch ein Mobiltelefon verfügen.

Da sich Wahrscheinlichkeitsmaße in vielerlei Hinsicht so wie andere Inhalts- oder Größenmaße von Mengen verhalten, können die obigen und ähnliche Sachverhalte über Mengendiagramme wie die folgenden veranschaulicht werden:



**Abb. 3.1.** Mengendiagramme von Wahrscheinlichkeitsmaßen

### 3.1.1 Bedingte Wahrscheinlichkeit

Eine wichtige Beobachtung ist die Tatsache, dass die Einschätzung der Wahrscheinlichkeit eines Ereignisses von der Kenntnis zusätzlicher Informationen abhängig sein kann: Angenommen, die Spielleiterin eines Würfelspiels würfelt verdeckt und teilt einem Spieler mit, dass das Ergebnis eine gerade Augenzahl ist, das Ergebnis also nur zwei, vier oder sechs sein kann. Mehr Informationen

werden ihm nicht zugänglich gemacht. Aus Sicht des Spielers ist die Wahrscheinlichkeit, dass die geworfene Augenzahl beispielsweise gleich drei ist, nun nicht mehr ein Sechstel – sondern offensichtlich gleich null.

Wie groß ist unter diesen Umständen die Wahrscheinlichkeit, dass die Augenzahl gleich zwei ist,  $A = \{\square\}$ ? Immerhin ist dieses Ergebnis nach wie vor möglich. Die Menge der möglichen Ausgänge des Zufallsexperiments hat sich durch die Zusatzinformation effektiv von  $\Omega = \{\square, \square\square, \square\square\square\}$  auf  $B = \{\square, \square\square, \square\square\square\}$  verkleinert. Daher gilt nach der Laplace'schen Formel für die gesuchte **bedingte Wahrscheinlichkeit**:

$$\Pr(A|B) = \Pr(\{\square\}|\{\square, \square\square, \square\square\square\}) = \frac{|\{\square\}|}{|\{\square, \square\square, \square\square\square\}|} = \frac{1}{3}$$

Die Laplace'sche Formel gilt nur für Gleichverteilungen über endlichen Wahrscheinlichkeitsräumen. Im Allgemeinen ist die bedingte Wahrscheinlichkeit wie folgt definiert.

Seien  $A, B \subseteq \Omega$  Ereignisse mit  $\Pr(B) > 0$ . Dann wird

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

die **Wahrscheinlichkeit von  $A$  unter der Bedingung  $B$**  genannt.

Wir können uns leicht davon zu überzeugen, dass diese Definition konsistent mit der weiter oben gewonnenen Intuition ist. Wenn die Laplace'sche Formel Gültigkeit hat, ist die bedingte Wahrscheinlichkeit der Anteil des Ereignisses  $A$  an einer reduzierten Anzahl der möglichen Ausgänge  $B$ :

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{|A \cap B|}{|B|}$$

Zur Illustration des Konzepts der bedingten Wahrscheinlichkeit betrachten wir das folgende Würfelspiel: Anna würfelt verdeckt und teilt Robert mit, ob die Augenzahl gerade oder ungerade ist. Robert muss daraufhin raten, ob die Augenzahl eine Primzahl ist oder nicht. Robert möchte seine Gewinnchancen maximieren. Wir legen fest:

$$B = \{\omega | \omega \text{ entspricht einer geraden Augenzahl}\} = \{\square, \square\square, \square\square\square\}$$

$$A = \{\omega | \omega \text{ entspricht einer Primzahl}\} = \{\square, \square\square, \square\square\square\}$$

Wenn Anna ansagt, dass die Augenzahl gerade ist, so ergibt sich für die Hypothese, dass sie auch prim ist, aus Sicht von Robert die folgende Wahrscheinlichkeit:

$$\begin{aligned} \Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(\{\square\})}{\Pr(\{\square, \square\square, \square\square\square\})} \\ &= \frac{1/6}{3/6} = \frac{1}{3} \approx 0,33 \end{aligned}$$

In diesem Fall ist die Wahrscheinlichkeit, dass die Augenzahl prim ist, lediglich 33 %. Daher sollte Robert die Ansage machen, dass das Ergebnis *nicht* prim ist. Die Gewinnwahrscheinlichkeit beträgt dann 67 %.

Falls Anna ansagt, dass die Augenzahl ungerade ist, ergibt sich hingegen:

$$\begin{aligned}\Pr(A|\neg B) &= \frac{\Pr(A \cap (\Omega \setminus B))}{\Pr(\Omega \setminus B)} = \frac{\Pr(\{\boxdot, \boxtimes\})}{1 - \Pr(B)} \\ &= \frac{2/6}{1/2} = \frac{2}{3} \approx 0,67\end{aligned}$$

In diesem Fall sollte Robert ansagen, dass das Ergebnis eine Primzahl ist.

Zwei Ereignisse  $A$  und  $B$  werden **unabhängig** genannt, wenn gilt:

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

Unter der Annahme, dass die Wahrscheinlichkeiten von  $A$  und  $B$  strikt positiv sind, kann die Bedingung der Unabhängigkeit auch wie folgt geschrieben werden:

$$\Pr(A|B) = \Pr(A) \text{ bzw. } \Pr(B|A) = \Pr(B)$$

Das Auftreten von  $B$  hat also keinen Einfluss auf die Wahrscheinlichkeit des Auftretens von  $A$  (und umgekehrt), am Beispiel: Anna würfelt verdeckt und teilt Robert mit, ob die Augenzahl gerade oder ungerade ist. Robert muss daraufhin raten, ob die Augenzahl durch drei teilbar ist oder nicht:

$$B = \{\omega | \omega \text{ entspricht einer geraden Augenzahl}\} = \{\boxdot, \boxtimes, \boxblacksquare\}$$

$$A = \{\omega | \omega \text{ entspricht einer durch drei teilbaren Augenzahl}\} = \{\boxdot, \boxblacksquare\}$$

Es gilt:

$$\begin{aligned}\Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(\{\boxblacksquare\})}{\Pr(\{\boxdot, \boxtimes, \boxblacksquare\})} \\ &= \frac{1/6}{3/6} = \frac{1}{3} = \Pr(A)\end{aligned}$$

Robert hat folglich keine nützlichen Informationen durch Annas Aussage gewonnen: Er sollte stets raten, dass die Zahl *nicht* durch drei teilbar ist, denn es gilt  $\Pr(\neg A) = \frac{2}{3} > \Pr(A)$ .

Eine Kennzahl, welche als Maß für die Abhängigkeit zweier Ereignisse interpretiert werden kann, ist die folgende:

Der **Transinformationsgehalt** (engl. *pointwise mutual information*, PMI) zweier Ereignisse  $A$  und  $B$  ist wie folgt gegeben:

$$\text{pmi}(A, B) = \log \left( \frac{\Pr(A \cap B)}{\Pr(A) \cdot \Pr(B)} \right)$$

Die Wahl der Basis des Logarithmus spielt inhaltlich eine untergeordnete, bei geeigneter Normierung (siehe unten) gar keine Rolle. Im Zweifel kann der natürliche Logarithmus verwendet werden. Die Kennzahl kann aus der Bedingung abgeleitet werden, dass sich bei Vorliegen von  $B$  die Wahrscheinlichkeit eines Eintreffens von  $A$  erhöht:

$$\begin{aligned}\Pr(A|B) > \Pr(A) &\Leftrightarrow \log\left(\frac{\Pr(A|B)}{\Pr(A)}\right) > 0 \\ &\Leftrightarrow \log\left(\frac{\Pr(A \cap B)}{\Pr(A) \cdot \Pr(B)}\right) > 0 \\ &\Leftrightarrow \text{pmi}(A, B) > 0\end{aligned}$$

Alle auftretenden Wahrscheinlichkeiten müssen zunächst als strikt positiv vorausgesetzt werden. Der Transinformationsgehalt zweier Ereignisse verschwindet genau dann, wenn diese unabhängig sind.

Oft wird diese Kennzahl noch in geeigneter Weise normiert, z. B. wie folgt:

$$\begin{aligned}\text{pmi}_{\text{norm}}(A, B) &= \log\left(\frac{\Pr(A \cap B)}{\Pr(A) \cdot \Pr(B)}\right) \cdot (-\log(\Pr(A \cap B)))^{-1} \\ &= \frac{\log(\Pr(A) \cdot \Pr(B))}{\log(\Pr(A \cap B))} - 1\end{aligned}$$

Für alle Ereignisse  $A$  und  $B$  gilt  $-1 < \text{pmi}_{\text{norm}}(A, B) \leq 1$ .

Der Grenzfall  $\text{pmi}_{\text{norm}}(A, B) \rightarrow (-1)$  kann als der Fall interpretiert werden, in dem  $A$  und  $B$  mit verschwindender Wahrscheinlichkeit überhaupt gemeinsam auftreten,  $\Pr(A \cap B) = 0$ .

Der Fall  $\text{pmi}_{\text{norm}}(A, B) = 1$  entspricht einer maximal möglichen Abhängigkeit der Ereignisse:  $\Pr(B) = \Pr(A) = \Pr(A \cap B)$ . In diesem Fall sind  $A$  und  $B$  bis auf eine Menge mit verschwindendem Wahrscheinlichkeitsmaß identisch:

$$\begin{aligned}\Pr(A \setminus B) &= \Pr(A \cup B) - \Pr(B) \\ &= \Pr(A) + \Pr(B) - \Pr(A \cap B) - \Pr(B) = 0, \\ \Pr(B \setminus A) &= 0\end{aligned}$$

**Anwendungsbeispiel.** Der Transinformationsgehalt ist in der Computerlinguistik (neudeutsch auch: *Natural Language Processing*) ein gebräuchliches Assoziationsmaß. Ist ein Korpus von Dokumenten (oder Dokumentteilen) gegeben, so kann die Assoziation eines Wortes (oder einer Phrase) mit einem Teilkorpus wie folgt bewertet werden:

$$\begin{aligned}\text{pmi}(\text{Teilkorpus}, \text{Wort}) &= \\ &\log\left(\frac{\text{Relative Häufigkeit des Worts im Teilkorpus}}{\text{Relative Häufigkeit des Worts im Gesamtkorpus}}\right)\end{aligned}$$

Kommt das Wort im Teilkorpus häufiger vor als im Gesamtkorpus, so hat diese Kennzahl einen positiven Wert; die Interpretation ist in diesem Fall, dass das mit dem Wort assoziierte Thema im Kontext des Teilkorpus relevant ist.

Eine Untersuchung des Korpus der Wahlprogramme für die Landtagswahl 2018 in Bayern führt auf folgende Wörter mit besonders hohem (normierten) Transinformationsgehalt [4]:

CSU	$\text{pmi}_{\text{norm}}$	SPD	$\text{pmi}_{\text{norm}}$
invest	0,359	ausbildungssqualität	0,189
rückführungen	0,359	energiequellen	0,189
familiengeld	0,348	lfv	0,189
baukindergeld	0,336	rentenversicherung	0,189
digitalbonus	0,336	kindergrundsicherung	0,186

Die Grünen	$\text{pmi}_{\text{norm}}$	Die Linken	$\text{pmi}_{\text{norm}}$
lebensgrundlagen	0,195	jobcenter	0,151
lsbtiq	0,194	mieterinnen	0,149
bio	0,177	patientinnen	0,149
fahrkarte	0,177	entbindungspfleger	0,147
antidiskriminierung	0,174	migrantinnen	0,147

Tabelle 3.1. Stichwortextraktion über Transinformationsgehalt

### 3.1.2 Der Satz von Bayes

Im letzten Abschnitt haben wir gesehen, wie die Kenntnis neuer Informationen die subjektive Wahrscheinlichkeit eines Ereignisses beeinflussen kann. Dieser Vorgang kann als Lernprozess unter Unsicherheit verstanden werden, der als Basis für weitere Entscheidungen dienen kann: Erfähre ein Würfelspieler, dass eine gerade Augenzahl als Ergebnis sehr wahrscheinlich ist, so würde er in Folge z. B. weniger Geld auf einen Ausgang von „drei“ wetten.

Ein solches Lernen unter Unsicherheit kann genauer wie folgt beschrieben werden: Die Wahrscheinlichkeit für die **Hypothese  $H$**  nimmt zunächst eine bestimmte **A-priori-Wahrscheinlichkeit**  $\text{Pr}(H)$  an. Die Kenntnis neuer Beobachtungen oder Feststellung der aktuellen Datenlage durch Erhebung **empirischer Daten  $E$**  (engl. *evidence*) führt zu einer Korrektur der A-priori-Wahrscheinlichkeit: Eine Berücksichtigung der neuen Informationen führt dann auf die **A-posteriori-Wahrscheinlichkeit**  $\text{Pr}(H|E)$ . In obigem Beispiel entspricht die Hypothese  $H$  dem Ereignis „eine Augenzahl von drei wird geworfen“ und  $\text{Pr}(E)$  der Wahrscheinlichkeit von „die Augenzahl ist gerade“.

Der Satz von Bayes stellt diese Zusammenhänge mathematisch klar.

**Satz von Bayes.** Für Ereignisse  $H, E$  mit  $\Pr(H) > 0, \Pr(E) > 0$  gilt

$$\Pr(H|E) = \frac{\Pr(E|H)}{\Pr(E)} \cdot \Pr(H)$$

Die A-posteriori-Wahrscheinlichkeit  $\Pr(H|E)$  ergibt sich aus der A-priori-Wahrscheinlichkeit  $\Pr(H)$  gerade durch Multiplikation eines Faktors, der proportional zur **Likelihood**  $\Pr(E|H)$  ist, also der Wahrscheinlichkeit für eine Beobachtung der empirischen Daten unter der Hypothese.

Der Beweis des Satzes ist nicht schwierig und folgt sofort aus der Definition der bedingten Wahrscheinlichkeit:

$$\Pr(H|E) = \frac{\Pr(H \cap E)}{\Pr(E)} = \frac{\Pr(E|H) \cdot \Pr(H)}{\Pr(E)}$$

Eine für viele Rechnungen nützliche alternative Fassung des Satzes von Bayes ist die folgende:

$$\Pr(H|E) = \frac{\Pr(E|H)}{\Pr(E|H) \cdot \Pr(H) + \Pr(E|\neg H) \cdot \Pr(\neg H)} \cdot \Pr(H)$$

**Anwendungsbeispiel.** In der Medizin werden bei der Bewertung von diagnostischen Tests in der Regel die **Sensitivität** und die **Spezifität** angegeben. Diese Tests dienen etwa dem Nachweis einer Erkrankung oder der Infektion mit einem bestimmten Erreger. Die Sensitivität ist der Anteil der kranken bzw. infizierten Patienten, bei denen der Test positiv ausfällt. Die Spezifität ist der Anteil der gesunden Patienten, bei denen der Test (korrekterweise) negativ ausfällt. Wir deuten diese Anteile als empirische Wahrscheinlichkeiten. Damit gilt:

$$\text{Sensitivität} = \Pr(E|H)$$

$$\text{Spezifität} = \Pr(\neg E|\neg H) = 1 - \Pr(E|\neg H)$$

Dabei haben wir die Hypothese „Patient ist krank“ mit  $H$  und die empirische Information „Test fällt positiv aus“ mit  $E$  abgekürzt.

PCR-Tests zum Nachweis einer Infektion mit dem Erreger SARS-CoV-2, dem Verursacher der Erkrankung COVID-19, verfügen über eine Sensitivität von wenigstens 70 % und eine Spezifität von mutmaßlich 95 % (Stand: Juni 2020 [5]).

Angenommen, ein Patient weist einen positiven Test auf: Was ist die Wahrscheinlichkeit  $\Pr(H|E)$ , dass der Patient auch tatsächlich mit dem Erreger infiziert ist, auf den der Test ansprechen soll?

Diese Wahrscheinlichkeit hängt entscheidend von der Verbreitung des Erregers unter den Testpersonen ab. Diese Verbreitung wird **Prävalenz** genannt

und bestimmt die A-priori-Wahrscheinlichkeit. Bei Vorstellung in einer Hausarztpraxis gehen wir von einer Prävalenz von  $\Pr(H) = 3\%$  aus. Damit ergibt sich die gesuchte A-posteriori-Wahrscheinlichkeit für eine tatsächliche Infektion mit SARS-CoV-2 bei erstem positivem Testergebnis:

$$\begin{aligned}\Pr(H|E) &= \frac{\Pr(E|H)}{\Pr(E|H) \cdot \Pr(H) + \Pr(E|\neg H) \cdot \Pr(\neg H)} \cdot \Pr(H) \\ &= \frac{0,7}{0,7 \cdot 0,03 + (1 - 0,95) \cdot (1 - 0,03)} \cdot 0,03 \\ &\approx 30\%\end{aligned}$$

Umgekehrt können wir fragen, mit welcher Wahrscheinlichkeit ein Patient mit negativem Testergebnis auch tatsächlich nicht infiziert ist:

$$\begin{aligned}\Pr(\neg H|\neg E) &= \frac{\Pr(\neg E|\neg H)}{\Pr(\neg E|\neg H) \cdot \Pr(\neg H) + \Pr(\neg E|H) \cdot \Pr(H)} \cdot \Pr(\neg H) \\ &= \frac{0,95}{0,95 \cdot (1 - 0,03) + (1 - 0,7) \cdot 0,03} \cdot (1 - 0,03) \\ &\approx 99\%\end{aligned}$$

Der Satz von Bayes bildet die Grundlage für einige Verfahren des maschinellen Lernens. Hierbei wird der Satz so interpretiert, dass eine initiale Überzeugung des lernenden Algorithmus über die Wahrscheinlichkeit einer Hypothese  $H$  durch  $\Pr(H)$  gegeben ist. Diese Überzeugung wird dann bei Vorliegen neu erhobener Daten  $E$  aktualisiert, sie beträgt nach dem Lernprozess  $\Pr(H|E)$ . Liegen dem Algorithmus wiederholt neue Daten vor, können weitere Anpassungen vorgenommen werden.

Wir betrachten nun ein Rechenbeispiel, das die grundlegende Funktionsweise eines auf Bayes'scher Inferenz basierenden Algorithmus für die Filterung von E-Mail-Spam illustrieren soll. Dieser Algorithmus könnte mit den folgenden Daten gefüttert werden:

- Generell sind 45 % aller E-Mail-Nachrichten als „Spam“ zu bewerten. Die A-priori-Wahrscheinlichkeit für die Hypothese „diese Nachricht ist Spam“ beträgt somit  $\Pr(H) = 0,45$ .
- Unter all diesen Spamm Nachrichten haben 5 % das Wort „Viagra“ in der Betreffzeile, die Likelihood unter dieser Datenlage beträgt also  $\Pr(E|H) = 0,05$ .
- Unter allen übrigen Nachrichten hat nur ein verschwindend geringer Anteil das Wort „Viagra“ im Betreff:  $\Pr(E|\neg H) = 0,001$ .

Insgesamt ergibt sich daraus die folgende A-posteriori-Wahrscheinlichkeit:

$$\begin{aligned}
 \Pr(H|E) &= \frac{\Pr(E|H)}{\Pr(E|H) \cdot \Pr(H) + \Pr(E|\neg H) \cdot \Pr(\neg H)} \cdot \Pr(H) \\
 &= \frac{0,05}{0,05 \cdot 0,45 + 0,001 \cdot (1 - 0,45)} \cdot 0,45 \\
 &\approx 98\%
 \end{aligned}$$

Das heißt: Wenn das Wort „Viagra“ im E-Mail-Betreff enthalten ist, ist der Algorithmus der Überzeugung, dass es sich bei dieser Nachricht mit hoher Wahrscheinlichkeit um Spam handelt.

Sei nun die Datenlage  $E$  dadurch gegeben, dass die E-Mail-Nachricht von einer Adresse aus versendet wurde, die im Adressbuch des Benutzers eingetragen ist. Ein verschwindend geringer Anteil der Spammnachrichten stammt von bekannten Absendern:  $\Pr(E|H) = 0,001$ . Demgegenüber stammt ein deutlich höherer Anteil der übrigen Nachrichten von bekannten Absendern:  $\Pr(E|\neg H) = 0,2$ . Aus diesen Statistiken ergibt sich:

$$\begin{aligned}
 \Pr(H|E) &= \frac{\Pr(E|H)}{\Pr(E|H) \cdot \Pr(H) + \Pr(E|\neg H) \cdot \Pr(\neg H)} \cdot \Pr(H) \\
 &= \frac{0,001}{0,001 \cdot 0,45 + 0,2 \cdot (1 - 0,45)} \cdot 0,45 \\
 &\approx 0,4\%
 \end{aligned}$$

Folglich kommt der Algorithmus zu dem Schluss, dass eine E-Mail von bekanntem Absender mit sehr hoher Wahrscheinlichkeit *kein* Spam ist.

Eine Bemerkung zu Hypothesen mit A-priori-Wahrscheinlichkeit eins bzw. null: Gilt  $\Pr(H) = 1$ , so gilt auch stets  $\Pr(H|E) = 1$  ungeachtet der Datenlage  $E$ . Dies lässt sich so interpretieren, dass eine a priori als richtig angenommene Hypothese unumstößlich ist. Ebenfalls ist die Bayes'sche Methode dann nicht „lernfähig“, wenn  $\Pr(H) = 0$  gilt.

Auch eine Datenlage mit einer Wahrscheinlichkeit von eins ist problematisch: Aus  $\Pr(E) = 1$  folgt stets  $\Pr(H|E) = \Pr(H)$ , eine solche empirische Grundlage ist in diesem Sinne also immer schwach. Daher sollten bei der Modellbildung im Allgemeinen extreme Wahrscheinlichkeiten vermieden werden: Dies wird auch die **Cromwell'sche Regel** genannt.

## 3.2 Zufallsvariablen

Betrachten wir den Würfelwurf einmal aus einer etwas grundlegenderen Perspektive, als einen komplexen mechanischen Vorgang. Der Würfel hat zu jedem Zeitpunkt während des Wurfs eine definierte Position und Orientierung im Raum. Was den Vorgang zu einem Zufallsexperiment macht, ist unsere Unkenntnis über die genaue Flugbahn und Orientierung des Würfels.

Genaugenommen besteht der Ergebnisraum  $\Omega$  also aus viel mehr Zuständen als nur den sechs Seiten des Würfels, die am Ende jeden Wurfs zuoberst liegen

können. Es wäre jedoch ein ungeheuer kompliziertes und unpraktikables Unterfangen, ein Wahrscheinlichkeitsmaß über alle möglichen Flugbahnen und Orientierungen zu bestimmen. Letztlich sind wir nämlich nur an der Augenzahl interessiert – diesen Umstand können wir formal durch eine Abbildung  $X$  darstellen, die **Zufallsvariable** oder **Zufallsgröße** genannt wird:  $X: \Omega \rightarrow \{1, 2, \dots, 6\}$ .

Es wird sich zeigen, dass für die Zwecke der Inferenzstatistik die Untersuchung der Verteilung der Werte der studierten Zufallsvariablen wesentlich und eine genaue Kenntnis des zugrundeliegenden Ergebnisraums letztlich unerheblich ist.

### 3.2.1 Diskrete und stetige Zufallsvariablen

Eine Zufallsvariable ist eine Größe, die Werte in einem gegebenem, begrenzten Bereich nicht mit Sicherheit, sondern nur mit einer gewissen Wahrscheinlichkeit annimmt. Formal führt diese Überlegung auf die folgende Definition.

Eine (**reellwertige**) **Zufallsvariable**  $X$  ist eine Funktion auf dem Ergebnisraum  $X: \Omega \rightarrow \mathbb{R}$  mit der Eigenschaft, dass für jeden Wert  $u \in \mathbb{R}$  die Menge  $X^{-1}([-\infty, u])$  messbar ist.

Ein bestimmter von einer Zufallsvariablen angenommener Wert  $u = X(\omega)$  wird auch als eine **Realisierung** dieser Zufallsgröße bezeichnet. Die Menge aller grundsätzlich möglichen Realisierungen ist durch die Bildmenge  $\text{Bild}(X) = X(\Omega)$  gegeben. Eine Grundannahme der Inferenzstatistik besteht darin, dass die Beobachtungen in einer Stichprobe Realisierungen von Zufallsvariablen sind.

In Worten ist die Größe  $\Pr(X^{-1}([-\infty, u]))$  „die Wahrscheinlichkeit, dass die Zufallsvariable  $X$  einen Wert annimmt, der kleiner oder gleich  $u$  ist.“ Daher schreiben wir etwas eingängiger auch einfach:

$$\Pr(X \leq u) := \Pr(X^{-1}([-\infty, u]))$$

Ebenso steht  $\Pr(X = u)$  für  $\Pr(X^{-1}(\{u\}))$ . Analog notieren wir für andere Mengen und Arten von Intervallen, z. B. schreiben wir für ein Intervall  $[a, b] \subset \mathbb{R}$  statt  $\Pr(X^{-1}([a, b]))$  auch  $\Pr(X \in [a, b])$  oder  $\Pr(a \leq X \leq b)$ .

Die **Verteilungsfunktion** einer Zufallsvariablen  $X$  ist wie folgt definiert:

$$F_X: \mathbb{R} \rightarrow [0, 1], F_X(u) = \Pr(X \leq u)$$

Die Verteilungsfunktion gibt also die Wahrscheinlichkeit an, mit der eine Zufallsgröße Werte bis zu einem Schwellenwert annimmt. Darüber hinaus kann bei Kenntnis der Verteilungsfunktion auch die Wahrscheinlichkeit berechnet werden, mit der die Zufallsvariable Werte in einem bestimmten Intervall annimmt; es gilt nämlich:  $\Pr(a < X \leq b) = F_X(b) - F_X(a)$  für alle  $a, b \in \mathbb{R}$  mit  $a \leq b$ .

Es kann gezeigt werden, dass eine Verteilungsfunktion immer auch die folgenden Eigenschaften erfüllt.

**Eigenschaften der Verteilungsfunktion.** Sei  $X$  eine Zufallsvariable. Dann gilt für deren Verteilungsfunktion  $F_X(\cdot)$  stets:

1.  $F_X$  ist monoton steigend.
2.  $F_X$  ist rechtsseitig stetig, d. h.:  $\lim_{u \searrow u_0} F_X(u) = u_0$  für alle  $u_0 \in \mathbb{R}$ .
3. Es gilt  $\lim_{u \rightarrow -\infty} F_X(u) = 0$  und  $\lim_{u \rightarrow \infty} F_X(u) = 1$ .

Ähnlich wie kategoriale und metrische Merkmale können auch Zufallsvariablen danach unterschieden werden, ob sie eine diskrete Anzahl von Werten (z. B.  $\{0, 1, 2, \dots\}$ ) oder ein ganzes Kontinuum von Werten (z. B. im Intervall  $[0, 1]$ ) annehmen. Diese Unterscheidung kann wie folgt charakterisiert werden.

Eine Zufallsvariable heißt **diskret**, wenn diese höchstens endlich viele oder abzählbar unendlich viele Werte annehmen kann.

Eine Zufallsvariable heißt **stetig**, wenn deren Verteilungsfunktion stetig ist.

Eine Zufallsvariable ist **absolut stetig**, wenn deren Verteilungsfunktion stetig und zudem mit Ausnahme höchstens endlich vieler Stellen stetig differenzierbar ist.

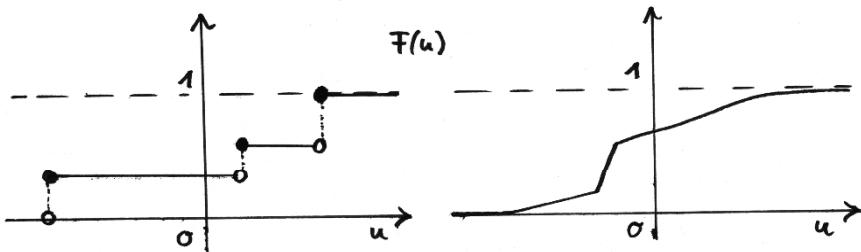
Ein paar Bemerkungen zu den mathematischen Grundlagen: Obige Bedingung an Absolutstetigkeit kann abgeschwächt werden, in der Praxis der Datenanalyse ist die angegebene Charakterisierung in der Regel ausreichend. Eine genauere mathematische Analyse zeigt außerdem, dass jede „hinreichend reguläre“ Zufallsvariable als die Summe einer diskreten und einer absolutstetigen Zufallsvariablen dargestellt werden kann (**Lebesgue-Zerlegung**, vgl. [6, Satz 12.1.14]). In diesem Sinne sind mit obigen Verteilungsfunktionen, von „pathologischen Fällen“ abgesehen, bereits alle wesentlichen aufgelistet.

Im Folgenden werden wir anstelle von „absolut stetig“ einfach „stetig“ sagen, da wir es nur mit dieser Art von stetiger Zufallsvariable zu tun haben werden. Für eine diskrete Zufallsvariable  $X$  erklären wir deren **Träger** wie folgt:

$$\text{supp}(X) = \{u \in \mathbb{R} | \Pr(X = u) > 0\}$$

Die Verteilungsfunktion einer diskreten Zufallsvariablen ist notwendigerweise eine Treppenfunktion, also abschnittsweise konstant, wobei die Sprungstellen durch die Elemente des Trägers gegeben sind.

Zusammen mit den weiter oben angegebenen allgemeinen Eigenschaften von Verteilungsfunktionen sind wir in der Lage, typische Verteilungsfunktionen diskreter bzw. stetiger Zufallsvariablen (rechts) zu skizzieren:



Gewinn und Verlust bei Glücksspielen sind Beispiele von Größen, die durch Zufallsvariablen modelliert werden können, am Beispiel: Robert gewinnt 5,00 EUR beim Wurf der Augenzahl 6, ansonsten verliert er 1,00 EUR. Die Auszahlung in EUR ist eine Zufallsvariable  $X$  mit

$$X(\omega) = \begin{cases} -1,00 & \text{falls } \omega \in \{\square, \blacksquare, \blacksquare, \blacksquare, \blacksquare\} \\ 5,00 & \text{falls } \omega = \blacksquare \end{cases}$$

Diese Zufallsgröße nimmt nur die zwei Werte  $t_0 = -1,00$  oder  $t_1 = 5,00$ , es handelt sich also um eine diskrete Zufallsvariable. Die zugehörige Verteilungsfunktion ist die folgende stückweise konstante Funktion:

$$F_X(u) = \begin{cases} 0 & \text{falls } u < -1,00 \\ \frac{5}{6} & \text{falls } -1,00 \leq u < 5,00 \\ 1 & \text{falls } 5,00 \leq u \end{cases}$$

Wir kehren zurück zum Beispiel der Dartspielerin. Als Zufallsvariable betrachten wir den Abstand des geworfenen Pfeils zum Mittelpunkt der Dartscheibe:  $X: \Omega \rightarrow \mathbb{R}$ ,  $X(\omega_1, \omega_2) = \sqrt{(\omega_1)^2 + (\omega_2)^2}$ . Die Verteilungsfunktion einer Zufallsvariablen hängt natürlich auch von dem zugrundeliegenden Wahrscheinlichkeitsmaß ab. Für die perfekte Dartspielerin (entspricht dem Dirac-Maß) ist  $X$  diskret mit folgender Verteilungsfunktion:

$$F_X^{(0)}(u) = \begin{cases} 0 & \text{falls } u < 0 \\ 1 & \text{falls } 0 \leq u \end{cases}$$

Werden alle Bereiche auf der Scheibe mit gleicher Wahrscheinlichkeit getroffen, ist die Zufallsvariable stetig und wie folgt verteilt:

$$F_X^{(R)}(u) = \begin{cases} 0 & \text{falls } u < 0 \\ \frac{u}{R} & \text{falls } 0 \leq u < R \\ 1 & \text{falls } R \leq u \end{cases}$$

Für das letzte im Beispiel angegebene Wahrscheinlichkeitsmaß (Treffer im „Bull's Eye“  $u \leq r$  in 90 % aller Fälle) haben wir ebenfalls eine stetige Verteilung:

$$F_X^{(r)}(u) = \begin{cases} 0 & \text{falls } u < 0 \\ 0,9 \cdot \frac{u}{r} & \text{falls } 0 \leq u < r \\ 0,9 + 0,1 \cdot \frac{u-r}{R-r} & \text{falls } r \leq u < R \\ 1 & \text{falls } R \leq u \end{cases}$$

### 3.2.2 Massen- und Dichtefunktionen

Ein zentraler Untersuchungsgegenstand der deskriptiven Statistik sind Häufigkeitsverteilungen, grafisch als Säulendiagramme bzw. Histogramme dargestellt. Diese bemessen die Häufigkeit des Vorkommens der Ausprägungen einer kategorialen Variable in der Stichprobe bzw. die Häufigkeit, mit der ein metrisches Merkmal Werte in einem bestimmten Bereich annimmt. Die in diesem Abschnitt vorgestellten Funktionen haben eine analoge Bedeutung in der Wahrscheinlichkeitstheorie: Massenfunktionen dienen der Angabe der Wahrscheinlichkeit, mit denen eine diskrete Zufallsvariable bestimmte Werte annimmt. Dichtefunktionen bemessen, mit welcher Wahrscheinlichkeit eine stetige Zufallsvariable Werte in einem bestimmten Bereich annimmt.

#### Massenfunktionen diskreter Zufallsvariablen

Eine diskrete Zufallsvariable ist durch die Wahrscheinlichkeiten charakterisiert, mit denen die Werte in deren Träger angenommen werden.

Die **Wahrscheinlichkeitsmassenfunktion** oder kurz **Massenfunktion** einer diskreten Zufallsvariable  $X$  ist wie folgt gegeben:

$$p_X: \text{supp}(X) \rightarrow [0, 1], p_X(u) = \Pr(X = u)$$

Die Verteilungsfunktion kann aus der Massenfunktion wie folgt rekonstruiert werden:

$$F_X(u) = \sum_{\substack{\kappa \leq u, \\ \kappa \in \text{supp}(X)}} p_X(\kappa)$$

Umgekehrt wollen wir jede Abbildung  $p: T \rightarrow [0, 1]$  bei höchstens abzählbar unendlich großer Definitionsmenge  $T = \{t_0, t_1 \dots\} \subset \mathbb{R}$ , welche die Bedingung

$$\sum_{\kappa \in T} p(\kappa) = 1$$

erfüllt, eine Massenfunktion nennen. Eine alternative Bezeichnung ist **Zähldichte**.

Als Beispiel betrachten wir wieder die stückweise konstante Verteilungsfunktion:

$$F(u) = \begin{cases} 0 & \text{falls } u < -1,00 \\ \frac{5}{6} & \text{falls } -1,00 \leq u < 5,00 \\ 1 & \text{falls } 5,00 \leq u \end{cases}$$

Der Graph dieser Funktion sieht wie folgt aus:

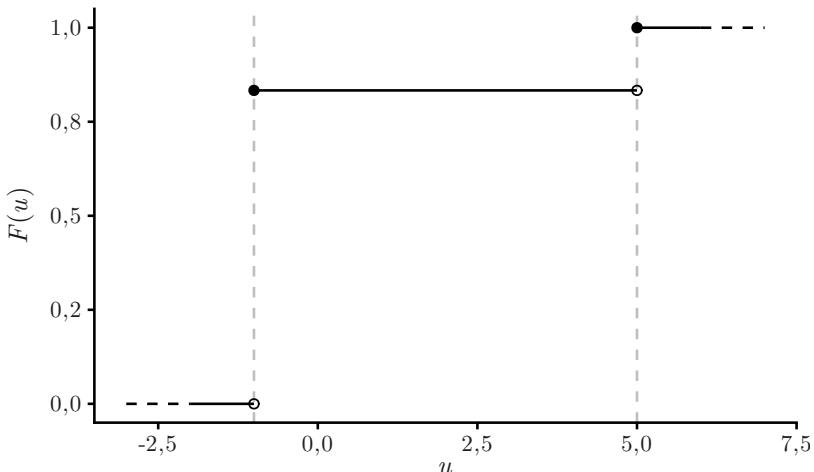


Abb. 3.2. Verteilungsfunktion einer diskreten Zufallsvariable

Die Position der einzelnen Stufen ist durch die Sprungstellen  $t_0 = -1,00$  und  $t_1 = 5,00$  bestimmt. Die zugehörige Massenfunktion gibt einfach die Höhe der Stufen wieder:

$$p: \{-1,00; 5,00\} \rightarrow [0, 1], p(u) = \begin{cases} \frac{5}{6} & \text{falls } u = -1,00 \\ \frac{1}{6} & \text{falls } u = 5,00 \end{cases}$$

### Dichtefunktionen stetiger Zufallsvariablen

Die Verteilungsfunktion einer diskreten Zufallsvariablen kann als Summe über die Massenfunktion dargestellt werden. Bei stetigen Zufallsvariablen ist eine ähnliche Konstruktion möglich, bei der anstelle einer Summe jedoch ein Integral steht.

Eine **Wahrscheinlichkeitsdichtefunktion**  $p_X: \mathbb{R} \rightarrow [0, \infty[$  einer stetigen Zufallsvariablen  $X$  erfüllt die folgende Bedingung:

$$F_X(u) = \int_{-\infty}^u p_X(\xi) d\xi$$

Statt Wahrscheinlichkeitsdichtefunktion sagen wir auch kurz **Wahrscheinlichkeitsdichte**, **Dichtefunktion** oder noch kürzer einfach „Dichte“. Für – in unserem vereinfachten Sinne – stetige Zufallsvariablen  $X$  ist deren Verteilungsfunktion  $F_X$  abschnittsweise stetig differenzierbar, daher existiert stets eine Dichte,

und diese ist an jeder Stelle  $u \in \mathbb{R}$ , an der  $F_X$  differenzierbar ist, eindeutig durch die Ableitung gegeben:  $p_X(u) = \frac{d}{du} F_X(u)$ .

Bei bekannter Dichtefunktion können wir recht komfortabel die Wahrscheinlichkeit berechnen, mit der eine stetige Zufallsvariable Werte in einem bestimmten Intervall annimmt:

$$\Pr(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b p_X(\xi) d\xi$$

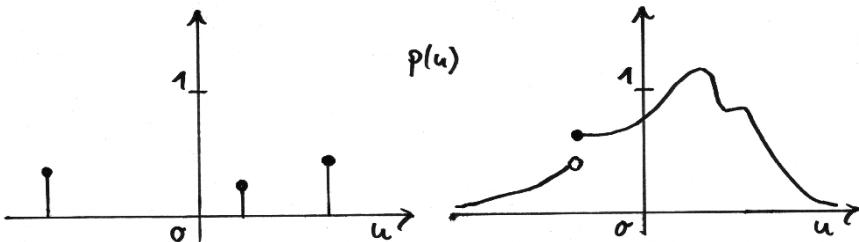
Umgekehrt können wir jede stückweise stetige Funktion  $p: \mathbb{R} \rightarrow [0, \infty[$ , welche die Bedingung

$$\int_{-\infty}^{\infty} p(\xi) d\xi = 1$$

erfüllt, eine Dichtefunktion nennen. Dabei ist das uneigentliche Integral über ganz  $\mathbb{R}$  wie folgt erklärt:

$$\begin{aligned} \int_{-\infty}^{\infty} p(\xi) d\xi &= \int_{-\infty}^0 p(\xi) d\xi + \int_0^{\infty} p(\xi) dx \\ &= \lim_{u \rightarrow -\infty} \int_u^0 p(\xi) d\xi + \lim_{u \rightarrow \infty} \int_0^u p(\xi) d\xi \end{aligned}$$

Mithin müssen beide Teilintegrale existieren. Notwendigerweise gilt daher für eine Dichtefunktion stets:  $\lim_{u \rightarrow -\infty} p(u) = \lim_{u \rightarrow \infty} p(u) = 0$ . Im Gegensatz zu einer Massenfunktion ist es durchaus möglich, dass für eine Dichtefunktion an einer Stelle  $p(u) > 1$  gilt. Eine typische Massenfunktion und Dichtefunktion (rechts) können wie folgt skizziert werden:



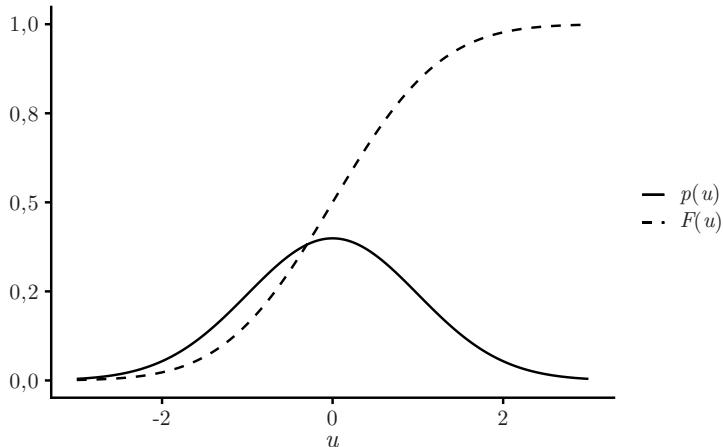
Für die Tatsache, dass die stetige bzw. diskrete Zufallsvariable  $X$  gemäß einer Dichte- bzw. Massenfunktion  $p(\cdot)$  verteilt ist, wird auch die Schreibweise  $X \sim p(\cdot)$  gebraucht.

Eine wichtiges Beispiel für eine Dichtefunktion ist die folgende.

Die **Standardnormalverteilung** ist durch die folgende Dichtefunktion gegeben:

$$p: \mathbb{R} \rightarrow [0, \infty[, p(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

Die folgende Abbildung zeigt den Funktionsgraphen der Dichte sowie der zugehörigen Verteilungsfunktion  $F(\cdot)$ :



**Abb. 3.3.** Dichte- und Verteilungsfunktion der Standardnormalverteilung

Das Integral dieser Funktion über alle reellen Zahlen ist tatsächlich gleich eins, dies ergibt sich aus dem Wert des **Gauß'schen Fehlerintegrals** (siehe z. B. [7, S. 99]):

$$\int_{-\infty}^{\infty} e^{-\xi^2} d\xi = \sqrt{\pi}$$

Die Wahrscheinlichkeit, dass eine standardnormalverteilte Zufallsvariable  $X$  einen Wert z. B. zwischen  $-1,96$  und  $+1,96$  annimmt, ist durch das folgende Integral gegeben:

$$\Pr(-1,96 \leq X \leq 1,96) = \int_{-1,96}^{1,96} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} d\xi$$

Im Gegensatz zum Integral über die gesamte Zahlengerade lässt sich dessen Wert nicht analytisch, sondern nur näherungsweise mit Methoden der Numerik berechnen, er ergibt sich zu  $\Pr(-1,96 \leq X \leq 1,96) \approx 0,95$ .

### 3.2.3 Transformation von Zufallsvariablen

Ist eine Zufallsvariable  $X: \Omega \rightarrow \mathbb{R}$  gegeben, so können wir auch Funktionen dieser Variablen untersuchen. Genauer gesagt: Haben wir eine hinreichend reguläre (z. B. stückweise stetige) Funktion auf der Bildmenge von  $X$  gegeben,  $f: I \rightarrow \mathbb{R}$  mit  $\text{Bild}(X) \subseteq I$ , so können wir die **transformierte Zufallsvariable**  $f(X)$  betrachten:

$$f(X): \Omega \rightarrow \mathbb{R}, f(X)(\omega) := (f \circ X)(\omega)$$

Die Werte der transformierten Zufallsvariablen  $Y = f(X)$  folgen einer anderen Verteilung als die Werte der ursprünglichen Variablen:

$$F_Y(u) = \Pr(Y \leq u) = \Pr(f(X) \leq u) = \Pr(X \in f^{-1}([-\infty, u]))$$

### Transformation diskreter Zufallsvariablen

Im diskreten Fall kann die Frage, wie sich die Werte der transformierten Zufallsvariablen verteilen, direkt beantwortet werden.

**Transformierte Massenfunktion.** Wenn  $X$  eine diskrete Zufallsvariable ist, dann ist  $f(X)$  ebenfalls diskret verteilt und die Massenfunktion wie folgt gegeben:

$$p_{f(X)}: f(\text{supp}(X)) \rightarrow [0, 1], p_{f(X)}(v) = \sum_{\kappa \in f^{-1}(v)} p_X(\kappa)$$

Etwas detaillierter ergibt sich das aus folgender kurzen Rechnung:

$$\begin{aligned} p_{f(X)}(v) &= \Pr(f(X) = v) = \Pr(X \in f^{-1}(v)) = \sum_{\kappa \in f^{-1}(v)} \Pr(X = \kappa) \\ &= \sum_{\kappa \in f^{-1}(v)} p_X(\kappa) \end{aligned}$$

für alle  $v \in \text{supp}(f(X)) = f(\text{supp}(X))$ .

Wir verdeutlichen anhand eines Beispiels. Sei  $X$  eine Zufallsvariable mit folgender Verteilung:

$$\Pr(X = u) = \begin{cases} 0,1 & \text{falls } u = -3 \\ 0,2 & \text{falls } u = 0 \\ 0,7 & \text{falls } u = 3 \\ 0 & \text{sonst} \end{cases}$$

Sei nun weiterhin  $f: \mathbb{R} \rightarrow \mathbb{R}$ ,  $f(u) = u^2$  die quadratische Funktion. Die Werte der Zufallsvariable  $f(X) = X^2$  verteilen sich dann wie folgt:

$$\begin{aligned} \Pr(X^2 = v) &= \begin{cases} \Pr(X = 0) & \text{falls } v = 0 \\ \Pr(X = -3) + \Pr(X = 3) & \text{falls } v = 9 \\ 0 & \text{sonst} \end{cases} \\ &= \begin{cases} 0,2 & \text{falls } v = 0 \\ 0,8 & \text{falls } v = 9 \\ 0 & \text{sonst} \end{cases} \end{aligned}$$

### Transformation stetiger Zufallsvariablen

Wir befassen uns nun mit der Berechnung der Dichtefunktion von Transformationen einer stetigen Zufallsvariable. Hierbei gilt zu beachten, dass die trans-

formierte Zufallsvariable nicht zwangsläufig wieder stetig sein muss – beispielsweise würde die Nullfunktion,  $f(u) = 0$  für alle  $u$ , stets auf die diskrete Zufallsvariable  $f(X) = 0$  führen.

Falls jedoch  $f: \mathbb{R} \rightarrow \mathbb{R}$  eine stetig differenzierbare, monoton steigende Funktion mit positiver Ableitung darstellt, so gilt für  $f(X)$  und alle  $u \in \mathbb{R}$ :

$$\begin{aligned}\Pr(f(X) \leq u) &= \Pr(X \leq f^{-1}(u)) \\ &= \int_{-\infty}^{f^{-1}(u)} p_X(\xi) d\xi\end{aligned}$$

Und weiter mit der Substitution  $\xi(t) = f^{-1}(t)$ :

$$\Pr(f(X) \leq u) = \int_{-\infty}^u p_X(f^{-1}(t)) \cdot \frac{d}{dt}(f^{-1}(t)) dt = \int_{-\infty}^u \frac{p_X(f^{-1}(t))}{f'(f^{-1}(t))} dt$$

Falls  $f$  monoton *fallend* ist, so haben wir dieselbe Rechnung mit umgekehrten Vorzeichen. Falls also die Transformation auf ganz  $\mathbb{R}$  monoton steigend oder fallend ist, so sehen wir im Integranden bereits eine handliche Formel für die Dichtefunktion von  $f(X)$ :

$$p_{f(X)}: \mathbb{R} \rightarrow [0, \infty[, p_{f(X)}(v) = \frac{p_X(f^{-1}(v))}{|f'(f^{-1}(v))|}$$

Einfache aber wichtige Beispiele stellen affin-lineare Funktionen dar; die Dichtefunktion von  $Y = m \cdot X + c$  mit Konstanten  $m \in \mathbb{R}, c \in \mathbb{R}, m \neq 0$ , ergibt sich gemäß der obigen Formel zu:

$$p_Y(v) = \frac{1}{|m|} \cdot p_X\left(\frac{v - c}{m}\right)$$

Ist  $f$  nicht monoton, sind Überlegungen nötig, wie die Transformation abschnittsweise zusammengesetzt werden kann. Diese Überlegungen führen auf die folgende Formel.

**Transformierte Dichtefunktion.** Seien  $X$  eine stetige Zufallsvariable und  $f: I \rightarrow \mathbb{R}$  mit  $\text{Bild}(X) \subseteq I$  eine stetig differenzierbare Funktion mit nichtverschwindender Ableitung.

Die Dichtefunktion der transformierten Zufallsvariablen  $f(X)$  ist dann wie folgt gegeben:

$$p_{f(X)}: \mathbb{R} \rightarrow [0, \infty[, p_{f(X)}(v) = \sum_{k=1}^K |g'_k(v)| \cdot p_X(g_k(v))$$

Dabei sind  $g_1(\cdot), \dots, g_K(\cdot)$  alle Lösungen der Gleichung  $f(g(v)) = v$ .

Als Rechenbeispiel betrachten wir wieder die quadratische Funktion  $f(u) = u^2$ . Die Stelle  $u = 0$  müssen wir zunächst ausschließen, da hier  $f'(u) = 0$  gilt. Da es sich aber bloß um eine isolierte Stelle handelt, können wir dort später einfach einen geeigneten Wert festlegen. Wir wollen eine Zufallsvariable  $X$  transformieren, die standardnormalverteilt ist:

$$p_X: \mathbb{R} \rightarrow [0, \infty[, p_X(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}u^2}$$

Die Gleichung  $f(g(v)) = v \Leftrightarrow (g(v))^2 = v$  hat zwei Lösungen, falls  $v > 0$ , nämlich  $g_1(v) = -\sqrt{v}$  und  $g_2(v) = \sqrt{v}$ . Folglich gilt für die Dichtefunktion von  $X^2$ :

$$\begin{aligned} p_{X^2}(v) &= |g'_1(v)| \cdot p_X(g_1(v)) + |g'_2(v)| \cdot p_X(g_2(v)) \\ &= \frac{1}{2\sqrt{v}} \cdot (p_X(-\sqrt{v}) + p_X(\sqrt{v})) = \frac{1}{\sqrt{v}} \cdot p_X(\sqrt{v}) = \\ &= \frac{1}{\sqrt{2\pi v}} \cdot e^{-\frac{1}{2}v} \end{aligned}$$

Offensichtlich gibt es keine reelle Lösung für  $(g(v))^2 = v$  für  $v < 0$ , in diesem Fall ergibt sich eine leere Summe und diese ist einfach null. Zusammengesetzt ergibt sich:

$$p_{X^2}: \mathbb{R} \rightarrow [0, \infty[, p_{X^2}(v) = \begin{cases} 0 & \text{falls } v \leq 0 \\ \frac{1}{\sqrt{2\pi v}} \cdot e^{-\frac{1}{2}v} & \text{falls } v > 0 \end{cases}$$

Diese Verteilung wird **Chi-Quadrat-Verteilung mit einem Freiheitsgrad** genannt. Wie die Chi-Quadrat-Verteilung mit beliebig vielen Freiheitsgraden aussieht, erklären wir im Abschn. 3.5.1.

### 3.3 Gemeinsame Verteilung von Zufallsvariablen

In der Regel bestehen funktionale Abhängigkeiten zwischen Zufallsgrößen. Eine wichtige Aufgabe der Inferenzstatistik besteht in der Modellierung derartiger Zusammenhänge und ihrer Beschreibung aufgrund vorliegender Daten – mehr dazu im Abschn. 4.5 zum Thema der Regressionsanalyse.

#### 3.3.1 Gemeinsame Verteilungs-, Masse- und Dichtefunktionen

Für zwei Zufallsvariablen  $X, Y$  (und sinngemäß für mehr als zwei Zufallsvariablen bzw. Arten von Intervallen) schreiben wir:

$$\Pr(X \leq u, Y \leq v) := \Pr(X^{-1}([-\infty, u]) \cap Y^{-1}([-\infty, v]))$$

In Worten ausgedrückt ist das die Wahrscheinlichkeit, mit der die Zufallsvariablen zugleich bestimmte Schwellenwerte nicht überschreiten.

Die **gemeinsame Verteilungsfunktion** einer Anzahl von Zufallsvariablen  $X_1, \dots, X_D$  ist durch

$F_{X_1, \dots, X_D} : \mathbb{R}^D \rightarrow \mathbb{R}$ ,  $F_{X_1, \dots, X_D}(u_1, \dots, u_D) = \Pr(X_1 \leq u_1, \dots, X_D \leq u_D)$  gegeben.

Sind  $X_1, \dots, X_D$  diskrete Zufallsvariablen, so ist ihre **gemeinsame Massenfunktion** wie folgt gegeben:

$$p_{X_1, \dots, X_D}(u_1, \dots, u_D) = \Pr(X_1 = u_1, \dots, X_D = u_D)$$

für alle  $u_1 \in \text{supp}(X_1), \dots, u_D \in \text{supp}(X_D)$ .

Sind  $X_1, \dots, X_D$  stetige Zufallsvariablen, so ergibt das Mehrfachintegral über eine **gemeinsame Dichtefunktion** die gemeinsame Verteilungsfunktion:

$$F_{X_1, \dots, X_D}(u_1, \dots, u_D) = \int_{-\infty}^{u_1} \cdots \int_{-\infty}^{u_D} p_{X_1, \dots, X_D}(\xi_1, \dots, \xi_D) d\xi_1 \cdots d\xi_D$$

für alle  $u_1, \dots, u_D \in \mathbb{R}$ . An allen Stellen, an denen die entsprechenden partiellen Ableitungen existieren, gilt

$$p_X(u_1, \dots, u_D) = \frac{\partial^D F_X}{\partial u_1 \cdots \partial u_D}(u_1, \dots, u_D)$$

Bei diskreten Zufallsgrößen können die einzelnen Massenfunktionen aus der gemeinsamen Massenfunktion durch Summation über die übrigen Variablen rekonstruiert werden, am Beispiel zweier Variablen  $X, Y$ :

$$\begin{aligned} p_X(u) &= \sum_{\kappa \in \text{supp}(Y)} p_{X,Y}(u, \kappa) \\ p_Y(v) &= \sum_{\kappa \in \text{supp}(X)} p_{X,Y}(\kappa, v) \end{aligned}$$

Die Dichtefunktionen stetiger Variablen können als **Randdichten** aus ihrer gemeinsamen Dichtefunktion „herausintegriert“ werden:

$$\begin{aligned} p_X(u) &= \int_{-\infty}^{\infty} p_{X,Y}(u, \xi) d\xi \\ p_Y(v) &= \int_{-\infty}^{\infty} p_{X,Y}(\xi, v) d\xi \end{aligned}$$

Schließlich kann auch die gemeinsame Verteilung einer diskreten und einer stetigen Zufallsvariablen betrachtet werden.

Eine **gemischte Wahrscheinlichkeitsfunktion** für eine stetige Zufallsvariable  $X$  und eine diskrete Zufallsvariablen  $Y$  ist eine Funktion  $p_{X,Y}: \mathbb{R} \times \text{supp}(Y) \rightarrow [0, \infty[$  mit der Eigenschaft

$$F_{X,Y}(u, v) = \Pr(X \leq u, Y \leq v) = \sum_{\substack{\kappa \leq v, \\ \kappa \in \text{supp}(Y)}} \int_{-\infty}^u p_{X,Y}(\xi, \kappa) d\xi$$

für alle  $u \in \mathbb{R}$ ,  $v \in \text{supp}(Y)$ .

In diesem Fall lassen sich die Randverteilungen durch Summation bzw. Integration nach der anderen Variablen ermitteln:

$$\begin{aligned} p_X(u) &= \sum_{\kappa \in \text{supp}(Y)} p_{X,Y}(u, \kappa) \\ p_Y(v) &= \int_{-\infty}^{\infty} p_{X,Y}(\xi, v) d\xi \end{aligned}$$

Dieses Konzept kann sinngemäß auf eine beliebige (endliche) Anzahl diskreter und stetiger Zufallsgrößen verallgemeinert werden.

### 3.3.2 Bedingte Massen- und Dichtefunktionen

Der Begriff bedingter Wahrscheinlichkeit kann sinngemäß auf Massen- und Dichtefunktionen übertragen werden. Eine bedingte Massen- oder Dichtefunktion beschreibt die Abhängigkeit der Zufallsvariablen  $Y$  von der Zufallsvariablen  $X$ . Die Rekonstruktion bedingter Verteilungen aus den Daten ist Gegenstand von Verfahren der Regressionsanalyse – siehe Abschn. 4.5.

Für zwei diskrete Zufallsvariablen  $X, Y$  und eine Realisierung  $u \in \text{supp}(X)$  von  $X$  ist

$$p_{Y|X}: \text{supp}(Y) \rightarrow [0, 1], v \mapsto p_{Y|X}(v|u) = \frac{p_{X,Y}(u, v)}{p_X(u)},$$

die **Massenfunktion von  $Y$  unter der Bedingung  $X = u$** .

Die bedingte Massenfunktion beschreibt einfach eine Wahrscheinlichkeit für das Auftreten eines Werts von  $Y$  unter der Bedingung, dass  $X$  einen bestimmten Wert annimmt:

$$p_{Y|X}(v|u) = \frac{p_{X,Y}(u, v)}{p_X(u)} = \frac{\Pr(X = u, Y = v)}{\Pr(X = u)} = \Pr(Y = v | X = u)$$

Die Schwierigkeit besteht in der Betrachtung stetiger Zufallsgrößen, für die stets  $\Pr(X = u) = 0$  gilt, sodass die bedingte Wahrscheinlichkeit nicht definiert ist. Dennoch können wir das Konzept auch hier anwenden.

Für zwei stetige Zufallsvariablen  $X, Y$  und eine Realisierung  $u \in \mathbb{R}$  von  $X$  mit  $p_X(u) > 0$  ist

$$p_{Y|X}: \mathbb{R} \rightarrow [0, \infty[, v \mapsto p_{Y|X}(v|u) = \frac{p_{X,Y}(u, v)}{p_X(u)},$$

die **Dichtefunktion von  $Y$  unter der Bedingung  $X = u$** .

Die bedingte Dichtefunktion erfüllt in der Tat die definierenden Eigenschaften einer Dichtefunktion, denn zum einen gilt offensichtlich  $p_{Y|X}(v|u) \geq 0$  für alle  $v \in \mathbb{R}$ , zum anderen:

$$\int_{-\infty}^{\infty} p_{Y|X}(\xi|u) d\xi = \frac{1}{p_X(u)} \cdot \int_{-\infty}^{\infty} p_{X,Y}(u, \xi) d\xi = \frac{1}{p_X(u)} \cdot p_X(u) = 1$$

Analog können wir bedingte Massen- und Dichtefunktionen auch für andere Kombinationen diskreter und stetiger Zufallsvariablen konstruieren. Wenn  $Y$  eine diskrete und  $X$  eine stetige Zufallsgröße ist, so ergibt sich die Massenfunktion von  $Y$  unter der Bedingung  $X = x$  aus der gemischten Wahrscheinlichkeitsfunktion:

$$p_{Y|X}(\cdot|u): \text{supp}(Y) \rightarrow [0, \infty[, p_{Y|X}(v|u) = \frac{p_{X,Y}(u, v)}{p_X(u)}$$

Auch in diesem Fall können wir die bedingte Massenfunktion nicht direkt als eine bedingte Wahrscheinlichkeit interpretieren. Wir werden sie dennoch als solche notieren, also  $\Pr(Y = v|X = u) := p_{Y|X}(v|u)$ , denn im Grenzwert kleiner Intervalle  $X \in [u, u + h]$  gilt, wo  $p_X(\cdot)$  und  $p_{X,Y}(\cdot, v)$  an der Stelle  $u$  stetig sind:

$$\begin{aligned} \lim_{h \searrow 0} \Pr(Y = v|u \leq X \leq u + h) &= \lim_{h \searrow 0} \frac{\Pr(Y = v, u \leq X \leq u + h)}{\Pr(u \leq X \leq u + h)} \\ &= \lim_{h \searrow 0} \frac{F_{X,Y}(u + h, v) - F_{X,Y}(u, v)}{F_X(u + h) - F_X(u)} \\ &= \lim_{h \searrow 0} \frac{h^{-1} \cdot (F_{X,Y}(u + h, v) - F_{X,Y}(u, v))}{h^{-1} \cdot (F_X(u + h) - F_X(u))} \\ &= \frac{p_{X,Y}(u, v)}{p_X(u)} \end{aligned}$$

### 3.3.3 Unabhängige Zufallsvariablen

Wir können zwei Zufallsvariablen  $X$  und  $Y$  als unabhängig voneinander ansehen, wenn für alle Intervalle  $[a, b]$  und  $[c, d]$  die Ereignisse  $X^{-1}([a, b])$  und  $Y^{-1}([c, d])$  unabhängig sind:

$$\Pr(a \leq X \leq b, c \leq Y \leq d) = \Pr(a \leq X \leq b) \cdot \Pr(c \leq Y \leq d)$$

Alle “vernünftigen” Mengen  $A, B$  können als abzählbare Vereinigung von Intervallen geschrieben werden, daher können wir für unabhängige Zufallsvariablen grundsätzlich annehmen:

$$\Pr(X \in A, Y \in B) = \Pr(X \in A) \cdot \Pr(Y \in B)$$

Da die Verteilung der Werte der Zufallsvariablen über Intervallen wiederum durch die Verteilungsfunktion gegeben ist, lässt sich die Bedingung der Unabhängigkeit wie folgt charakterisieren.

Die Zufallsvariablen  $X_1, \dots, X_D$  heißen **unabhängig (voneinander)**, falls für alle  $u_1, \dots, u_D \in \mathbb{R}$  gilt:

$$F_{X_1, \dots, X_D}(u_1, \dots, u_D) = F_{X_1}(u_1) \cdot F_{X_2}(u_2) \cdots F_{X_D}(u_D) = \prod_{d=1}^D F_{X_d}(u_d)$$

Die Bedingung der Unabhängigkeit überträgt sich sinngemäß auf Massen- und Dichtefunktionen:

**Unabhängigkeit diskreter oder stetiger Zufallsvariablen.** Diskrete Zufallsvariablen  $X_1, \dots, X_D$  sind genau dann unabhängig voneinander, wenn gilt:

$$p_{X_1, \dots, X_D}(u_1, \dots, u_D) = \prod_{d=1}^D p_{X_d}(u_d)$$

für alle  $u_1 \in \text{supp}(X_1), \dots, u_D \in \text{supp}(X_D)$ .

Stetige Zufallsvariablen  $X_1, \dots, X_D$  sind genau dann unabhängig voneinander, wenn gilt:

$$p_{X_1, \dots, X_D}(u_1, \dots, u_D) = \prod_{d=1}^D p_{X_d}(u_d)$$

an allen Stellen  $u_1, \dots, u_D \in \mathbb{R}$ , an denen die jeweilige Dichtefunktion stetig ist.

Eine weitere wichtige Aussage ist die folgende:

**Transformationen unabhängiger Zufallsvariablen.** Wenn zwei Zufallsvariablen  $X$  und  $Y$  unabhängig sind, dann sind auch die transformierten Variablen  $f(X)$  und  $g(Y)$  unabhängig.

Das Resultat ergibt sich wie folgt:

$$\begin{aligned} F_{f(X), f(Y)}(u, v) &= \Pr(X \in f^{-1}(-\infty, u], Y \in g^{-1}(-\infty, v])) \\ &= \Pr(X \in f^{-1}(-\infty, u]) \cdot \Pr(Y \in g^{-1}(-\infty, v])) \end{aligned}$$

$$\begin{aligned}
&= \Pr(f(X) \in ]-\infty, u]) \cdot \Pr(g(Y) \in ]-\infty, v]) \\
&= F_{f(X)}(u) \cdot F_{g(Y)}(v)
\end{aligned}$$

### 3.4 Kennzahlen von Zufallsvariablen

Masse- und Dichtefunktionen entsprechen den Häufigkeitsverteilungen in der deskriptiven Statistik: Diskrete Zufallsvariablen produzieren die Ausprägungen kategorialer Merkmale und stetige Zufallsvariablen die Werte metrischer Merkmale. Unsere Erwartung ist, dass bei einer hinreichend großen Stichprobe das Säulendiagramm eine Repräsentation der zugrundeliegenden Massenfunktion bzw. das Histogramm eine gute Näherung der Dichtefunktion ist.

Auch statistische Kennzahlen wie Mittelwerte oder Streuungsparameter finden ihre Entsprechung in der Wahrscheinlichkeitstheorie.

#### 3.4.1 Median, Erwartungswert und Varianz

Im Folgenden definieren wir Lage- und Streuungsparameter von Zufallsvariablen.

Sei  $X$  eine Zufallsvariable mit Verteilungsfunktion  $F_X: \mathbb{R} \rightarrow [0, 1]$ . Die **Quantilfunktion** von  $X$  ist wie folgt gegeben:

$$Q[X]: \mathbb{R} \rightarrow [0, 1], Q[X](r) = \inf\{u \in \mathbb{R} | F_X(u) \geq r\}$$

Sei  $0 < \alpha < 1$ . Ein  **$\alpha$ -Quantil** von  $X$  ist jede Zahl  $q_\alpha$  mit

$$\Pr(X \leq q_\alpha) \geq \alpha \text{ und } \Pr(X \geq q_\alpha) \geq 1 - \alpha$$

Insbesondere ist  $Q[X](\alpha)$  ein  $\alpha$ -Quantil. Ein **Median** von  $X$  ist ein  $\alpha$ -Quantil mit  $\alpha = 1/2$ .

Ähnlich wie in der deskriptiven Statistik sind subtile Definitionen notwendig, da es grundsätzlich mehrere Mediane oder  $\alpha$ -Quantile geben kann. Wir können aber auch hier Eindeutigkeit erzwingen, indem wir salopp von *dem* Quantil  $q_\alpha[X] := Q[X](\alpha)$  und *dem* Median  $m[X] := Q[X](1/2)$  sprechen.

Das  $\alpha$ -Quantil einer stetigen Zufallsvariable kann über die Wahrscheinlichkeitsdichte  $p_X(\cdot)$  ausgedrückt werden, es genügt der folgenden Formel:

$$\int_{-\infty}^{q_\alpha[X]} p_X(\xi) d\xi = \alpha$$

Speziell für den Median:

$$\int_{-\infty}^{m[X]} p_X(\xi) d\xi = \frac{1}{2}$$

Sei  $X$  eine diskrete oder stetige Zufallsvariable mit der Massenfunktion  $p_X : \text{supp}(X) \rightarrow [0, 1]$  bzw. einer Dichtefunktion  $p_X : \mathbb{R} \rightarrow [0, \infty[$ .

Der **Erwartungswert**  $E[X]$  der Zufallsgröße ist dann wie folgt erklärt:

$$E[X] = \sum_{\kappa \in \text{supp}(X)} \kappa \cdot p_X(\kappa)$$

bzw.

$$E[X] = \int_{-\infty}^{\infty} \xi \cdot p_X(\xi) d\xi$$

Erwartungswert und Median von Zufallsvariablen sind eng verwandt mit dem arithmetischen Mittel bzw. dem empirischen Median der deskriptiven Statistik. Sie geben Werte an, welche eine durchschnittliche oder typische Realisierung einer Zufallsvariablen darstellen.

Für das Würfelspiel mit  $t_0 = 5$  EUR Gewinn bei einer Augenzahl von sechs, ansonsten  $t_1 = 1$  EUR Verlust ist der erwartete Gewinn zum Beispiel:

$$E[X] = t_0 \cdot \Pr(X = t_0) + t_1 \cdot \Pr(X = t_1) = (-1) \cdot \frac{5}{6} + 5 \cdot \frac{1}{6} = 0$$

Im Schnitt würde Spieler Robert auf Dauer weder Geld verlieren, noch Geld gewinnen. Es gilt jedoch zu beachten, dass bei einmaligem Spiel sein Verlustrisiko wesentlich größer als die Gewinnwahrscheinlichkeit ist:  $5/6$  gegenüber  $1/6$ , also fünfmal höher. Dieser Umstand wird durch den im Gewinnfall wesentlich höheren Auszahlbetrag ausgeglichen.

Im Gegensatz zum arithmetischen Mittel muss der Erwartungswert jedoch nicht immer existieren oder endlich sein. Das sogenannte **Sankt-Petersburg-Paradoxon** führt zum Beispiel auf eine diskrete Zufallsvariable, für die der Erwartungswert nicht existiert: Spielleiterin Anna wirft eine Münze so oft hintereinander, bis „Kopf“ zuoberst liegt. Passiert dies nach  $k$  Würfen, so erhält Robert  $2^{k-1}$  Euro ausbezahlt: Fällt „Kopf“ bereits beim ersten Wurf, erhält er 1,00 EUR, fällt „Kopf“ erst beim zweiten Wurf, 2,00 EUR usw.; der potenzielle Gewinn verdoppelt sich mit jedem Wurf von „Zahl“. Stellen mit jedem Wurf die potenziellen Ausgänge „Kopf“ und „Zahl“ gleich wahrscheinliche und von den anderen Würfen unabhängige Ereignisse dar, so ist der Gewinn eine Zufallsvariable  $X$  mit der folgenden Massenfunktion:

$$p_X(2^{k-1}) = \left(\frac{1}{2}\right)^k, k \in \{1, 2, \dots\}$$

Der erwartete Gewinn beträgt folglich:

$$E[X] = \sum_{k=1}^{\infty} 2^{k-1} \cdot \left(\frac{1}{2}\right)^k = \sum_{k=1}^{\infty} \frac{1}{2} = \infty$$

Das Paradoxon besteht in der Beobachtung, dass Robert, würde er allein den erwarteten Gewinn zugrundelegen, eine beliebig hohe Gebühr für die Teilnahme an dem Spiel akzeptieren würde.

Eine stetige Dichtefunktion, die auf keinen wohldefinierten Erwartungswert führt, ist die folgende:

$$p(u) = \frac{1}{\pi} \cdot \frac{1}{u^2 + 1}$$

Dem uneigentlichen Integral

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \cdot \frac{\xi}{\xi^2 + 1} d\xi$$

kann nämlich kein endlicher Wert zugeordnet werden.

Für die Berechnung des Erwartungswerts von transformierten Zufallsvariablen gilt glücklicherweise eine einfache Berechnungsvorschrift.

**Erwartungswert transformierter Zufallsvariablen.** Sei eine diskrete oder stetige Zufallsvariable  $X$  sowie eine (hinreichend reguläre, nichtkonstante) Funktion  $f: I \rightarrow \mathbb{R}$  mit  $\text{Bild}(X) \subseteq I$  gegeben.

Der Erwartungswert der transformierten Zufallsvariablen  $f(X)$  kann dann wie folgt berechnet werden:

$$E[f(X)] = \sum_{\kappa \in \text{supp}(X)} f(\kappa) \cdot p_X(\kappa).$$

bzw.

$$E[f(X)] = \int_{-\infty}^{\infty} f(\xi) \cdot p_X(\xi) d\xi$$

Gelegentlich werden diese Formeln auch für eine *Definition* des Erwartungswerts herangezogen. Sie können aber – oder sollten sogar – als Lehrsätze aufgefasst werden, die aus den Transformationsformeln für Masse- und Dichtefunktionen folgen. Wir zeigen dies am Beispiel einer stetigen Zufallsvariablen  $X$  und einer monoton steigenden Transformation  $f: \mathbb{R} \rightarrow \mathbb{R}$  mit positiver Ableitung:

$$\begin{aligned} E[f(X)] &= \int_{-\infty}^{\infty} \xi \cdot p_{f(X)}(\xi) d\xi \\ &= \int_{-\infty}^{\infty} \xi \cdot p_X(f^{-1}(\xi)) \cdot \left( \frac{d}{d\xi} f^{-1}(\xi) \right) d\xi \\ &= \int_{-\infty}^{\infty} f(t) \cdot p_X(t) dt \end{aligned}$$

Dabei wurde die Substitution  $\xi(t) = f(t)$  vorgenommen.

Die Formeln für die Berechnung von Erwartungswerten gelten auch sinngemäß für Funktionen von Zufallsvariablen über der gemeinsamen Verteilung:

**Erwartungswert gemeinsam transformierter Zufallsvariablen.** Seien zwei diskrete bzw. stetige Zufallsvariablen  $X, Y$  sowie eine (hinreichend reguläre, nichtkonstante) Funktion  $f: I \times J \rightarrow \mathbb{R}$  mit  $\text{Bild}(X) \subseteq I$ ,  $\text{Bild}(Y) \subseteq J$  gegeben.

Der Erwartungswert der Zufallsvariablen  $f(X, Y)$  kann dann wie folgt berechnet werden:

$$E[f(X, Y)] = \sum_{\kappa_1 \in \text{supp}(X)} \sum_{\kappa_2 \in \text{supp}(Y)} f(\kappa_1, \kappa_2) \cdot p_{X,Y}(\kappa_1, \kappa_2).$$

bzw.

$$E[f(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi_1, \xi_2) \cdot p_{X,Y}(\xi_1, \xi_2) d\xi_1 d\xi_2$$

Aus den bislang bekannten Formeln für den Erwartungswert können wichtige Eigenschaften leicht abgeleitet werden.

**Eigenschaften des Erwartungswerts.** Der Erwartungswert ist eine lineare Abbildungsvorschrift:

$$E[a \cdot X + b \cdot Y + c] = a \cdot E[X] + b \cdot E[Y] + c$$

für Zufallsvariablen  $X, Y$  und Zahlen  $a, b, c \in \mathbb{R}$ .

Zudem gilt die folgende Monotonieeigenschaft. Falls  $\Pr(Y - X \geq 0) = 0$  gilt, dann folgt:

$$E[Y] \geq E[X]$$

Die Bedingung  $\Pr(Y - X \geq 0) = 0$  ist zum Beispiel erfüllt, wenn  $Y(\omega) \geq X(\omega)$  für alle  $\omega \in \Omega$  gilt. Insbesondere ist der Erwartungswert einer nichtnegativen Zufallsvariablen ebenfalls nicht negativ:  $Y \geq 0 \Rightarrow E[Y] \geq 0$ .

Es ist wichtig sich einzuprägen, dass trotz der Linearitätseigenschaft im Allgemeinen  $E[f(X)] \neq f(E[X])$  gilt.

Die folgenden Kennzahlen sind analog zu den entsprechenden empirischen Streuungsparametern.

Die **Varianz** einer Zufallsvariablen  $X$  ist durch den erwarteten quadrierten Abstand zum Erwartungswert gegeben:

$$\sigma^2[X] = E[(X - E[X])^2]$$

Die **Standardabweichung** ist durch  $\sigma[X] = \sqrt{\sigma^2[X]}$  gegeben.

Eine alternative Schreibweise ist  $\text{var}[X] = \sigma^2[X]$ . Die Varianz ist im Gegensatz zum Erwartungswert kein lineares Funktional, es gilt jedoch

$$\sigma^2[m \cdot X + c] = m^2 \cdot \sigma^2[X]$$

für alle  $m, c \in \mathbb{R}$ .

Als Rechenbeispiel ermitteln wir Erwartungswert und Varianz einer standardnormalverteilten Zufallsvariablen  $X$ . Der Erwartungswert ist null, da er sich durch ein Integral über eine ungerade Funktion mit  $f(-\xi) = -f(\xi)$  berechnet:

$$E[X] = \int_{-\infty}^{\infty} \xi \cdot p_X(\xi) d\xi = \int_{-\infty}^{\infty} \frac{\xi}{\sqrt{2\pi}} \cdot e^{-\frac{\xi^2}{2}} d\xi = 0$$

Für die Varianz gilt somit vermöge einer Produktintegration:

$$\begin{aligned} \text{var}[X] &= E[(X - E[X])^2] = E[X^2] \\ &= \int_{-\infty}^{\infty} \xi^2 \cdot p_X(\xi) d\xi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \xi \cdot \left( \xi \cdot e^{-\frac{\xi^2}{2}} \right) d\xi \\ &= -\frac{\xi}{\sqrt{2\pi}} \cdot e^{-\frac{\xi^2}{2}} \Big|_{\xi=-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{\xi^2}{2}} d\xi \\ &= 1 \end{aligned}$$

Weiterhin wollen wir lineare Funktionen der standardnormalverteilten Zufallsvariable betrachten:  $Y = \sigma \cdot X + \mu$  mit  $\mu, \sigma \in \mathbb{R}$ ,  $\sigma > 0$ . Gemäß der Transformationsformeln für die Dichtefunktion gilt:

$$p_Y(u) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(u-\mu)^2}{2\sigma^2}}$$

für alle  $u \in \mathbb{R}$ . Um Erwartungswert und Varianz von  $Y$  zu ermitteln, müssen wir uns nicht noch einmal den Kopf über Integrationsregeln zerbrechen, sondern nutzen direkt das Verhalten der Kennzahlen unter solchen Transformationen aus:

$$\begin{aligned} E[Y] &= E[\sigma \cdot X + \mu] = \sigma \cdot E[X] + \mu = \mu, \\ \text{var}[Y] &= \text{var}[\sigma \cdot X + \mu] = \sigma^2 \cdot \text{var}[X] = \sigma^2 \end{aligned}$$

Eine Zufallsvariable, die wie  $Y$  verteilt ist, nennen wir **normalverteilt mit Erwartungswert  $\mu$  und Varianz  $\sigma^2$** . Für die Dichtefunktion schreiben wir:

$$Y \sim p_Y(\cdot) = \mathcal{N}(\cdot | \mu, \sigma^2)$$

Für  $\mu = 0$  und verschiedene Werte von  $\sigma$  ist die Normalverteilung in Abb. 4.2 links oben dargestellt. Mit kleinerem Wert für  $\sigma$  wird der Funktionsgraph, der auch als **Gauß'sche Glockenkurve** bezeichnet wird, schmäler: Die Werte der normalverteilten Zufallsvariable weisen eine geringere Streuung auf. Die Standardnormalverteilung ist gerade die Normalverteilung mit Erwartungswert null und Varianz eins:  $X \sim \mathcal{N}(\cdot | 0, 1)$ .

Schließlich kann auch die Entropie von Zufallsvariablen definiert werden.

Die **Shannon-Entropie** einer diskreten Zufallsvariablen  $X$  ist wie folgt gegeben:

$$H[X] = - \sum_{\kappa \in \text{supp}(X)} p_X(\kappa) \cdot \ln(p_X(\kappa))$$

Die **differenzielle Entropie** einer stetigen Zufallsvariablen  $X$ :

$$H[X] = - \int_{-\infty}^{\infty} p_X(\xi) \cdot \ln(p_X(\xi)) d\xi$$

Wie üblich gilt dabei die Vereinbarung „ $0 \cdot \ln 0 = 0$ “.

### 3.4.2 Kovarianz und Korrelation

Die folgenden Kennzahlen entsprechen den gleichnamigen Assoziationsparametern der deskriptiven Statistik.

Die **Kovarianz** zweier Zufallsvariablen  $X, Y$  ist wie folgt gegeben:

$$\sigma[X, Y] = E[(X - E[X]) \cdot (Y - E[Y])]$$

Die **Korrelation** ist durch folgende Größe gegeben:

$$\rho[X, Y] = \frac{\sigma[X, Y]}{\sigma[X] \cdot \sigma[Y]},$$

falls  $\sigma[X] > 0$  und  $\sigma[Y] > 0$ .

Die Zufallsvariablen  $X$  und  $Y$  heißen **unkorreliert**, falls  $\sigma[X, Y] = 0$  gilt.

Eine alternative Schreibweise ist  $\text{cov}[X, Y] = \sigma[X, Y]$ . Die Kovarianz einer Zufallsvariablen mit sich selbst ist gerade deren Varianz:  $\sigma[X, X] = \sigma^2[X]$ .

Eine alternative, oft gebrauchte Berechnungsvorschrift ergibt sich durch ausmultiplizieren unter Beachtung der Linearität der Erwartungswerts:

$$\begin{aligned} \sigma[X, Y] &= E[X \cdot Y] - E[E[X] \cdot Y] - E[E[Y] \cdot X] + E[X] \cdot E[Y] \\ &= E[X \cdot Y] - E[X] \cdot E[Y] \end{aligned}$$

Insbesondere gilt für die Varianz:

$$\sigma^2[X] = E[X^2] - (E[X])^2$$

Als konkretes Rechenbeispiel betrachten wir zwei stetige Zufallsvariablen  $X$  und  $Y$  mit folgender gemeinsamen Dichtefunktion:

$$p_{X,Y}(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{u^2 - 2\rho \cdot u \cdot v + v^2}{2(1-\rho^2)}\right)$$

mit dem zusätzlichen Parameter  $\rho \in \mathbb{R}$ ,  $0 \leq \rho < 1$ . Dies ist ein Beispiel für eine multivariate Normalverteilung – mehr zu solchen Verteilungen in Abschn. 5.4.2. Wir wollen die Kovarianz von  $X$  und  $Y$  berechnen,  $\sigma[X, Y] = E[X \cdot Y] - E[X] \cdot E[Y]$ . Hierfür verwenden wir die Formel für die Berechnung eines allgemeinen Gauß'schen Integrals:

$$\int_{-\infty}^{\infty} e^{-a\xi^2 + b\xi + c} d\xi = \sqrt{\frac{\pi}{a}} \cdot e^{\frac{b^2}{4a} + c}$$

für alle  $a, b, c \in \mathbb{R}$  mit  $a > 0$ . Die Randdichte von  $Y$  ergibt sich nach dieser Formel wie folgt:

$$\begin{aligned} p_Y(v) &= \int_{-\infty}^{\infty} p_{X,Y}(\xi, v) d\xi \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{\xi^2}{2(1-\rho^2)} + \frac{\rho v}{(1-\rho^2)} \cdot \xi - \frac{v^2}{2(1-\rho^2)}\right) d\xi \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} = \mathcal{N}(v|0, 1) \end{aligned}$$

Das ist eine Standardnormalverteilung; die mit  $X$  assoziierte Dichtefunktion aus Symmetriegründen ebenfalls. Folglich gilt  $E[X] = E[Y] = 0$ .

Jetzt müssen wir bloß noch den Erwartungswerts des Produkts berechnen.

$$\begin{aligned} \sigma[X, Y] &= E[X \cdot Y] - E[X] \cdot E[Y] = E[X \cdot Y] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \xi_1 \cdot \xi_2 \cdot p_{X,Y}(\xi_1, \xi_2) d\xi_1 d\xi_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\xi_1 \cdot \xi_2}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{(\xi_1)^2 - 2\rho \cdot \xi_1 \cdot \xi_2 + (\xi_2)^2}{2(1-\rho^2)}\right) d\xi_1 d\xi_2 \end{aligned}$$

Die folgende lineare Transformation mit Determinante  $\det(D\xi(s, t)) = \sqrt{1-\rho^2}$  führt zu einer Entkopplung der Variablen:

$$\xi(s, t) = \begin{pmatrix} \xi_1(s, t) \\ \xi_2(s, t) \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{1+\rho} & -\sqrt{1-\rho} \\ \sqrt{1+\rho} & \sqrt{1-\rho} \end{pmatrix} \cdot \begin{pmatrix} s \\ t \end{pmatrix}$$

Das Integral transformiert sich dann nämlich über ein paar algebraische Umformungen zu folgendem:

$$\begin{aligned} \sigma[X, Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \xi_1(s, t) \cdot \xi_2(s, t) \cdot p_{X,Y}(\psi(s, t)) \cdot \det(D\xi(s, t)) ds dt \\ &= \frac{1}{4\pi} \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ((1+\rho) \cdot s^2 - (1-\rho)t^2) \cdot e^{-\frac{1}{2}(s^2+t^2)} \cdot ds dt \end{aligned}$$

Dieses kann nun schrittweise – erst nach der einen, dann nach der anderen Variable – integriert werden, das Ergebnis ist schließlich:

$$\sigma[X, Y] = \rho$$

Um Eigenschaften und Bedeutung der Kovarianz aus Anwendungssicht etwas näher zu beleuchten, betrachten wir eine fiktive Firma, die quadratische Werkstücke mit der mittleren Seitenlänge  $\bar{x} = 1,00 \text{ m}$  herstellt. Nach dem Produktionsprozess werden die Seiten des Werkstücks noch einmal genau vermessen, und im Schnitt wird eine Abweichung der Seitenlänge von  $s(x) = 0,05 \text{ m}$  beobachtet.

Wir gehen davon aus, dass die Seitenlänge als Zufallsvariable  $X$  mit Erwartungswert  $E[X] = \bar{x}$  und Varianz  $\sigma^2[X] = s^2(x)$  modelliert werden kann. Wir werden später in Abschn. 4.2 sehen, inwieweit eine solche Annahme gerechtfertigt sein kann.

Damit ergibt sich, dass die durchschnittliche Fläche eines Werkstücks *nicht*, wie naiv zu erwarten wäre,  $(E[X])^2 = 1,00 \text{ m}^2$  beträgt, sondern tatsächlich geringfügig *mehr*, selbst wenn die Seitenlängen symmetrisch um den Erwartungswert verteilt sind:

$$E[X^2] = \sigma^2[X] + (E[X])^2 = 1,0025 \text{ m}^2$$

Dies ist jedoch nur dann gegeben, wenn der Herstellungsprozess zwar zu Variationen in der Seitenlänge führt, beide Seitenlängen jedoch stets genau identisch wären. Wenn davon nicht ausgegangen werden kann, besteht ein realistisches Modell wohl eher in der Annahme, dass die Seitenlängen Realisierungen von verschiedenen Zufallsvariablen  $X$  und  $Y$  sind, die gleichen Erwartungswert haben, jedoch unkorreliert sind, d. h., es gilt  $E[X \cdot Y] = E[X] \cdot E[Y]$ . Die durchschnittliche Fläche der Werkstücke beträgt dann tatsächlich  $E[X \cdot Y] = 1,00 \text{ m}^2$ .

Weiterhin ist der folgende Zusammenhang zwischen Unabhängigkeit und Korrelation von Zufallsvariablen von Bedeutung.

**Korrelation unabhängiger Zufallsvariablen.** Unabhängige Zufallsvariablen sind paarweise unkorreliert.

Für zwei stetige Zufallsvariablen  $X, Y$  ergibt sich das etwa wie folgt:

$$\begin{aligned} E[X \cdot Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \xi_1 \cdot \xi_2 \cdot p_{X,Y}(\xi_1, \xi_2) d\xi_1 d\xi_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \xi_1 \cdot \xi_2 \cdot p_X(\xi_1) \cdot p_Y(\xi_2) d\xi_1 d\xi_2 \\ &= \left( \int_{-\infty}^{\infty} \xi_1 \cdot p_X(\xi_1) d\xi_1 \right) \cdot \left( \int_{-\infty}^{\infty} \xi_2 \cdot p_Y(\xi_2) d\xi_2 \right) \\ &= E[X] \cdot E[Y] \end{aligned}$$

Die Umkehrung gilt im Allgemeinen *nicht*: Es gibt unkorrelierte Zufallsvariablen, die nicht unabhängig sind. Für ein Gegenbeispiel betrachten wir eine stetige Variable  $X$  mit folgender Dichtefunktion, einer Gleichverteilung auf dem Intervall  $[-2, 2]$ :

$$p_X(u) = \begin{cases} \frac{1}{4} & \text{falls } -2 \leq u \leq 2 \\ 0 & \text{sonst} \end{cases}$$

Die Zufallsvariablen  $X$  und  $Y = X^2$  sind nicht unabhängig, denn auf der einen Seite gilt:

$$\begin{aligned} F_{X,Y}(1, 1) &= \Pr(X \leq 1, X^2 \leq 1) = \Pr(X \leq 1, -1 \leq X \leq 1) \\ &= \Pr(-1 \leq X \leq 1) = \frac{1}{2} \end{aligned}$$

Auf der anderen Seite gilt:

$$\begin{aligned} F_X(1) \cdot F_Y(1) &= \Pr(X \leq 1) \cdot \Pr(-1 \leq X \leq 1) = \frac{3}{4} \cdot \frac{1}{2} \\ &\neq F_{X,Y}(1, 1) \end{aligned}$$

Dennoch sind die Variablen unkorreliert, da ihre Kovarianz verschwindet:

$$\begin{aligned} \sigma[X, Y] &= E[X \cdot Y] - E[X] \cdot E[Y] \\ &= E[X^3] - E[X] \cdot E[X^2] \\ &= \frac{1}{4} \int_{-2}^2 \xi^3 d\xi - \left( \frac{1}{4} \int_{-2}^2 \xi d\xi \right) \cdot \left( \frac{1}{4} \int_{-2}^2 \xi^2 d\xi \right) \\ &= 0 \end{aligned}$$

### 3.4.3 Die Tschebyscheff'sche Ungleichung

Unsere geleistete Vorarbeit wird durch das folgende Resultat belohnt, welches weitreichende Konsequenzen hat.

**Tschebyscheff'sche Ungleichung.** Für jede Zufallsvariable  $X$ , bei der Erwartungswert  $E[X]$  und Varianz  $\sigma^2[X]$  existieren und endlich sind, gilt:

$$\Pr(|X - E[X]| \geq r) \leq \frac{\sigma^2[X]}{r^2}$$

für alle  $r \in \mathbb{R}$ ,  $r > 0$ .

Damit ergibt sich ein sehr bedeutsamer Zusammenhang: Je kleiner die Varianz, desto unwahrscheinlicher wird es, dass die Zufallsvariable Werte annimmt, die weit entfernt vom Erwartungswert liegen. Diese Aussage deckt sich mit unserer intuitiven Vorstellung einer Kennzahl für die Streuung.

Eine alternative Form der Ungleichung kann durch Einsetzen von  $r = z\sigma[X]$  erhalten werden:

$$\Pr(|X - E[X]| \geq z \cdot \sigma[X]) \leq \frac{1}{z^2}$$

für alle  $z \in \mathbb{R}$ ,  $z > 1$ .

Beispielsweise ist die Wahrscheinlichkeit, dass eine Zufallsvariable einen Wert annimmt, der um mehr als sechs Standardabweichungen vom Erwartungswert abweicht, geringer als 3 %. Oder anders herum ausgedrückt: Die Wahrscheinlichkeit, eine Realisierung der Zufallsvariablen in einem Abstand von höchstens sechs Standardabweichungen vom Erwartungswert zu finden, ist sehr hoch und beträgt wenigstens 97 %. Dabei spielt es keine Rolle, wie die Verteilung der Variablen aussieht, solange Erwartungswert und Varianz endliche Werte darstellen.

Die Tschebyscheff'sche Ungleichung kann wie folgt bewiesen werden. Zunächst setzen wir der Übersichtlichkeit der Formeln halber  $\mu = E[X]$  und  $\sigma = \sigma[X]$ .

Darüber hinaus führen wir das Konzept der **Indikatorfunktion** ein. Für jedes Ereignis  $A \subseteq \Omega$  können wir diese wie folgt definieren:

$$I_A: \Omega \rightarrow \mathbb{R}, I_A(\omega) = \begin{cases} 1 & \text{falls } \omega \in A \\ 0 & \text{sonst} \end{cases}$$

Die Indikatorfunktion ist eine diskrete Zufallsvariable mit folgender Verteilungsfunktion:

$$F_{I_A}(u) = \begin{cases} 0 & \text{falls } u < 1 \\ \Pr(A) & \text{falls } 1 \leq u \end{cases}$$

Daraus ergibt sich, dass gerade  $E[I_A] = \Pr(A)$  gilt. Jede Wahrscheinlichkeit kann daher auch als Erwartungswert einer Zufallsvariable aufgefasst werden.

Wir wenden das Konzept auf  $A := \{\omega \in \Omega \mid |X(\omega) - \mu| \geq r\}$  an und betrachten  $Y = I_A$ , es gilt somit:

$$E[Y] = \Pr(A) = \Pr(|X(\omega) - \mu| \geq r)$$

Auf der anderen Seite gilt für alle  $\omega \in \Omega$ :  $|X(\omega) - \mu|^2 \geq r^2 \cdot Y(\omega)$ , und folglich gemäß Monotonie und Linearität des Erwartungswerts:

$$\sigma^2 = E[|X - \mu|^2] \geq E[r^2 Y] = r^2 E[Y] = r^2 \cdot \Pr(|X(\omega) - \mu| \geq r)$$

Teilen von rechter und linker Seite durch  $r^2$  liefert das gewünschte Ergebnis.

Eine weitere nützliche Form der Tschebyscheff'schen Ungleichung ist

$$\Pr(|X - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}$$

für alle  $\varepsilon \in \mathbb{R}$ ,  $\varepsilon > 0$ .

### 3.5 Summen und Produkte von Zufallsvariablen

Für diskrete Zufallsvariablen  $X, Y$  können wir die Massenfunktion von deren Summe  $Z = X + Y$  ermitteln. Zunächst stellen wir fest, dass der Träger von  $Z$  durch

$$\text{supp}(Z) = \text{supp}(X) + \text{supp}(Y) := \{u + v \mid u \in \text{supp}(X), v \in \text{supp}(Y)\}$$

gegeben ist. Weiterhin gilt:

$$\Pr(X + Y = u) = \sum_{\kappa \in \text{supp}(X)} \Pr(X = \kappa, Y = u - \kappa)$$

**Massenfunktion der Summe diskreter Zufallsvariablen.** Für diskrete Zufallsvariablen  $X, Y$  gilt:

$$p_{X+Y}(u) = \sum_{\kappa \in \text{supp}(X)} p_{X,Y}(\kappa, u - \kappa)$$

für alle  $u \in \text{supp}(X + Y) = \text{supp}(X) + \text{supp}(Y)$ .

Falls  $X$  und  $Y$  zudem noch unabhängige Zufallsvariablen sind, gilt:

$$p_{X+Y}(u) = \sum_{\kappa \in \text{supp}(X)} p_X(\kappa) \cdot p_Y(u - \kappa)$$

Wir wollen nun die Dichtefunktion der Summe  $Z = X + Y$  zweier stetiger Zufallsvariablen  $X$  und  $Y$  berechnen. Auch in diesem Fall können wir eine Formel für die resultierende Dichtefunktion angeben. Die Verteilung der Summe kann zunächst über die gemeinsame Verteilung von  $X$  und  $Y$  wie folgt ausgedrückt werden:

$$F_Z(u) = \Pr(X + Y \leq u) = \iint_{B(u)} p_{X,Y}(\xi_1, \xi_2) d\xi_1 d\xi_2$$

Dabei ist  $B(u)$  der Integrationsbereich  $\{(\xi_1, \xi_2) \in \mathbb{R}^2 \mid \xi_1 + \xi_2 \leq u\}$ . Als Mehrfachintegral geschrieben ergibt sich mit anschließender Substitution von  $\xi_2(t) = t - \xi_1$  und Vertauschen der Integralreihenfolge:

$$\begin{aligned} F_Z(u) &= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{u-\xi_1} p_{X,Y}(\xi_1, \xi_2) d\xi_2 \right) d\xi_1 \\ &= \int_{-\infty}^{\infty} \left( \int_{-\infty}^u p_{X,Y}(\xi_1, t - \xi_1) dt \right) d\xi_1 \\ &= \int_{-\infty}^u \left( \int_{-\infty}^{\infty} p_{X,Y}(\xi_1, t - \xi_1) d\xi_1 \right) dt \end{aligned}$$

Die Dichtefunktion ist die Ableitung der Verteilungsfunktion  $p_Z(u) = \frac{d}{du} F_Z(u)$ , also der innere Integrand.

**Dichtefunktion der Summe stetiger Zufallsvariablen.** Für stetige Zufallsvariablen  $X, Y$  ist die Dichtefunktion von deren Summe wie folgt gegeben:

$$p_{X+Y}: \mathbb{R} \rightarrow [0, \infty[, p_{X+Y}(u) = \int_{-\infty}^{\infty} p_{X,Y}(\xi, u - \xi) d\xi$$

Ein interessanter Spezialfall ist dann gegeben, wenn  $X$  und  $Y$  unabhängige Zufallsvariablen sind. In diesem Fall gilt:

$$p_{X+Y}(u) = \int_{-\infty}^{\infty} p_X(\xi) \cdot p_Y(u - \xi) d\xi$$

Allgemein wird die Operation

$$(f * g)(u) := \int_{-\infty}^{\infty} f(\xi) \cdot g(u - \xi) d\xi$$

für integrierbare Funktionen  $f, g: \mathbb{R} \rightarrow \mathbb{R}$  eine **Faltung** genannt. Mit dieser Sprechweise gilt also: Die Dichtefunktion der Summe zweier unabhängiger stetiger Zufallsvariablen ist durch die Faltung der einzelnen Dichtefunktionen gegeben,  $p_{X+Y} = p_X * p_Y$ .

Als Rechenbeispiel wollen wir die Dichtefunktion der Summe zweier unabhängiger Zufallsvariablen  $X$  und  $Y$  berechnen, die jeweils normalverteilt mit Erwartungswert null und Varianz  $\sigma^2$  bzw.  $\tau^2$  sind:

$$p_X(u) = \mathcal{N}(u|0, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{u^2}{2\sigma^2}}, p_Y(u) = \mathcal{N}(u|0, \tau^2) = \frac{1}{\sqrt{2\pi}\tau} \cdot e^{-\frac{u^2}{2\tau^2}}$$

mit  $\sigma, \tau \in \mathbb{R}$ ,  $\sigma, \tau > 0$ . Das Faltungsintegral führt wieder auf ein allgemeines Gauß'sches Integral:

$$\begin{aligned} p_{X+Y}(u) &= \frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} e^{-\frac{\xi^2}{2\sigma^2}} \cdot e^{-\frac{(u-\xi)^2}{2\tau^2}} d\xi \\ &= \frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} \exp\left(-\frac{\sigma^2 + \tau^2}{2\sigma^2\tau^2}\xi^2 + \frac{u}{\tau^2}\xi - \frac{u^2}{2\tau^2}\right) d\xi \\ &= \frac{1}{\sqrt{2\pi} \cdot \sqrt{\sigma^2 + \tau^2}} \cdot e^{-\frac{u^2}{2(\sigma^2 + \tau^2)}} \\ &= \mathcal{N}(u|0, \sigma^2 + \tau^2) \end{aligned}$$

Die Summe der normalverteilten Zufallsvariablen mit Erwartungswert null ist also wiederum normalverteilt mit Erwartungswert null. Die Varianz berechnet sich über die Summe der Varianzen – tatsächlich gilt das auch für die Erwartungswerte, ganz allgemein:

$$\begin{aligned} p_{X_1} &= \mathcal{N}(\cdot | \mu_1, \sigma_1^2), p_{X_2} = \mathcal{N}(\cdot | \mu_2, \sigma_2^2) \Rightarrow \\ p_{X_1+X_2} &= \mathcal{N}(\cdot | \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \end{aligned}$$

Als weitere Anwendung der Faltungsformel wollen wir die Dichtefunktion der Summe zweier unabhängiger Zufallsvariablen  $X$  und  $Y$  berechnen, die beide gemäß der Chi-Quadrat-Verteilung mit einem Freiheitsgrad verteilt sind:

$$\chi_1^2: \mathbb{R} \rightarrow [0, \infty[, \chi_1^2(v) = \begin{cases} 0 & \text{falls } v \leq 0 \\ \frac{1}{\sqrt{2\pi v}} \cdot e^{-\frac{1}{2}v} & \text{falls } v > 0 \end{cases}$$

Die Faltung dieser Funktion mit sich selbst berechnet sich wie folgt:

$$p_{X+Y}(u) = \int_{-\infty}^{\infty} \chi_1^2(\xi) \cdot \chi_1^2(u - \xi) d\xi$$

Es gilt  $p_{X+Y}(u) = 0$  für  $u \leq 0$ , wir können für die Rechnung daher  $u > 0$  annehmen. Der Integrand verschwindet, wenn  $\xi \leq 0$  oder  $u \leq \xi$ . Folglich, mit der Substitution  $\xi(t) = \frac{u}{2} \cdot (t + 1)$ :

$$\begin{aligned} p_{X+Y}(u) &= \frac{e^{-\frac{1}{2}u}}{2\pi} \int_0^u \frac{1}{\sqrt{\xi \cdot (u - \xi)}} d\xi = \frac{e^{-\frac{1}{2}u}}{2\pi} \int_{-1}^1 \frac{1}{\sqrt{1 - t^2}} dt \\ &= \frac{e^{-\frac{1}{2}u}}{2\pi} \cdot \arcsin t \Big|_{t=-1}^1 = \frac{1}{2} \cdot e^{-\frac{1}{2}u} \end{aligned}$$

für  $u > 0$ .

Wir erinnern daran, dass die Chi-Quadrat-Verteilung mit der Verteilung des Quadrats einer standardnormalverteilten Variablen identisch ist, somit entspricht obige Verteilung der Summe zweier solcher Quadrate: Dies ist die Chi-Quadrat-Verteilung mit zwei Freiheitsgraden.

Für Produkte von unabhängigen stetigen Zufallsvariablen gilt hingegen die folgende Formel.

**Dichtefunktion des Produkts stetiger Zufallsvariablen.** Für zwei stetige Zufallsvariablen  $X, Y$  gilt:

$$p_{X \cdot Y}(u) = \int_{-\infty}^{\infty} \frac{1}{|\xi|} \cdot p_{X,Y} \left( \xi, \frac{u}{\xi} \right) d\xi$$

Der Beweis funktioniert ähnlich wie bei der Summe von Zufallsvariablen. Zunächst gilt für die Verteilungsfunktion:

$$F_{X \cdot Y}(u) = \Pr(X \cdot Y \leq u) = \iint_{B(u)} p_{X,Y}(\xi_1, \xi_2) d\xi_1 d\xi_2$$

mit  $B(u) = \{(\xi_1, \xi_2) \in \mathbb{R}^2 | \xi_1 \cdot \xi_2 \leq u\}$ . Wir zerlegen den Integrationsbereich in zwei Bereiche:

$$\begin{aligned} B_+(u) &= \left\{ (\xi_1, \xi_2) \in \mathbb{R}^2 \middle| \xi_2 \leq \frac{u}{\xi_1}, \xi_1 > 0 \right\}, \\ B_-(u) &= \left\{ (\xi_1, \xi_2) \in \mathbb{R}^2 \middle| \xi_2 \geq \frac{u}{\xi_1}, \xi_1 < 0 \right\} \end{aligned}$$

Mit der Substitution  $\xi_2(t) = \frac{t}{\xi_1}$  folgt:

$$\begin{aligned} \iint_{B_+(u)} p_{X,Y}(\xi_1, \xi_2) d\xi_1 d\xi_2 &= \int_0^\infty \left( \int_{-\infty}^{u/\xi_1} p_{X,Y}(\xi_1, \xi_2) d\xi_2 \right) d\xi_1 \\ &= \int_{-\infty}^u \left( \int_0^\infty \frac{1}{\xi_1} \cdot p_{X,Y}\left(\xi_1, \frac{t}{\xi_1}\right) d\xi_1 \right) dt \end{aligned}$$

Auf ähnliche Weise erzielen wir das Ergebnis:

$$\iint_{B_-(u)} p_{X,Y}(\xi_1, \xi_2) d\xi_1 d\xi_2 = \int_u^\infty \left( \int_{-\infty}^0 \frac{1}{\xi_1} \cdot p_{X,Y}\left(\xi_1, \frac{t}{\xi_1}\right) d\xi_1 \right) dt$$

Daher:

$$\begin{aligned} p_{X,Y}(u) &= \frac{d}{du} F_{X,Y}(u) \\ &= \int_0^\infty \frac{1}{\xi} \cdot p_{X,Y}\left(\xi, \frac{u}{\xi}\right) d\xi - \int_{-\infty}^0 \frac{1}{\xi} \cdot p_{X,Y}\left(\xi, \frac{u}{\xi}\right) d\xi \\ &= \int_{-\infty}^\infty \frac{1}{|\xi|} \cdot p_{X,Y}\left(\xi, \frac{u}{\xi}\right) d\xi \end{aligned}$$

Als eine Anwendung der Produktformel betrachten wir eine standardnormalverteilte Zufallsvariable  $Z$  sowie eine Zufallsvariable  $Y$ , die gemäß der Chi-Quadrat-Verteilung mit zwei Freiheitsgraden verteilt ist, also:

$$p_Y(u) = \begin{cases} 0 & \text{falls } u \leq 0 \\ 1/2 \cdot e^{-u/2} & \text{falls } u > 0 \end{cases}$$

Wir nehmen an, dass  $Z$  und  $Y$  unabhängig sind und wollen die Verteilung der Zufallsvariablen

$$T = \frac{\sqrt{2} \cdot Z}{\sqrt{Y}} = \sqrt{2} \cdot Z \cdot Y^{-\frac{1}{2}}$$

bestimmen. Zunächst einmal, mit  $f(u) = u^{-1/2}$ :

$$\begin{aligned} p_{Y^{-1/2}}(v) &= p_{f(Y)}(v) = \frac{p_Y(f^{-1}(v))}{|f'(f^{-1}(v))|} \\ &= \begin{cases} 0 & \text{falls } v \leq 0 \\ \frac{1}{v^3} \cdot e^{-\frac{1}{2v^2}} & \text{falls } v > 0 \end{cases} \end{aligned}$$

Aus der Produktformel folgt unter Voraussetzung der Unabhängigkeit mit der Substitution  $\xi(t) = \sqrt{1+u^2} \cdot t^{-1}$ :

$$\begin{aligned} p_{Z \cdot Y^{-1/2}}(u) &= \int_{-\infty}^\infty \frac{1}{|\xi|} \cdot p_Y(\xi) \cdot p_Z\left(\frac{u}{\xi}\right) d\xi \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{1}{\xi^4} \cdot e^{-\frac{1+u^2}{\xi^2}} d\xi \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \cdot (1+u^2)^{-\frac{3}{2}} \cdot \int_0^\infty t^2 \cdot e^{-\frac{1}{2}t^2} dt \\
&= \frac{1}{2} \cdot (1+u^2)^{-\frac{3}{2}}
\end{aligned}$$

Und schließlich mit der Transformationsformel für lineare Funktionen:

$$p_{\sqrt{2} \cdot Z \cdot Y^{-1/2}}(u) = \frac{1}{\sqrt{2}} \cdot p_{Z \cdot Y^{-1/2}}\left(\frac{u}{\sqrt{2}}\right) = \frac{1}{2\sqrt{2}} \cdot \left(1 + \frac{u^2}{2}\right)^{-\frac{3}{2}}$$

Das ist die sogenannte Student'sche  $t$ -Verteilung<sup>1</sup> mit zwei Freiheitsgraden.

### 3.5.1 Chi-Quadrat- und Student'sche $t$ -Verteilung

In den Rechenbeispielen der vorigen Abschnitte haben wir die Dichtefunktionen von  $Y_1 = (X_1)^2$  und  $Y_2 = (X_1)^2 + (X_2)^2$  für standardnormalverteilte und unabhängige Zufallsvariablen  $X_1, X_2$  berechnet. In Verallgemeinerung dieser Resultate ergibt sich für beliebig lange Summen der Quadrate standardnormalverteilter Zufallsgrößen die folgende Verteilung.

Die Summe  $Y_N = (X_1)^2 + \dots + (X_N)^2$  der Quadrate unabhängiger standardnormalverteilter Zufallsvariablen  $X_1, \dots, X_N$  folgt einer **Chi-Quadrat-Verteilung mit  $N$  Freiheitsgraden**:

$$\begin{aligned}
p_{Y_N}(\cdot) &= \chi_N^2(\cdot): \mathbb{R} \rightarrow [0, \infty[, \\
\chi_N^2(u) &= \begin{cases} 0 & \text{falls } u \leq 0 \\ 2^{-\frac{N}{2}} \cdot (\Gamma(\frac{N}{2}))^{-1} \cdot u^{\frac{N}{2}-1} \cdot e^{-\frac{u}{2}} & \text{falls } u > 0 \end{cases}
\end{aligned}$$

Dabei ist

$$\Gamma: ]0, \infty[ \rightarrow \mathbb{R}, \Gamma(z) = \int_0^\infty \xi^{z-1} \cdot e^{-\xi} d\xi$$

die **Gammafunktion**.

Für jede natürliche Zahl  $n \in \mathbb{N}$  gilt

$$\Gamma(n+1) = n!,$$

dabei ist  $n! = n \cdots 2 \cdot 1$  für jede natürliche Zahl  $n$  deren **Fakultät**, per Definition gilt  $0! = 1$ . Die Gammafunktion erweitert also den Definitionsbereich der Fakultät über ganzzahlige Werte hinaus. Speziell für halbzahlig Werte gilt:

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{(2n)! \cdot \sqrt{\pi}}{n! \cdot 4^n}$$

Abb. 3.4 zeigt oben den Funktionsgraphen der Chi-Quadrat-Verteilung für ein paar ausgesuchte Werte für die Anzahl der Freiheitsgrade.

---

<sup>1</sup> Eigentlich William Sealy Gosset (\* 1876 – † 1937), „Student“ ist ein Pseudonym.

Für  $N = 1$  und  $N = 2$  hatten wir die Chi-Quadrat-Verteilung bereits berechnet, Einsetzen in die obige Formel überzeugt uns von deren Richtigkeit für diese Fälle. Ansonsten können wir vollständige Induktion anwenden. Entscheidend ist dabei der nichtkonstante Teil, der konstante Vorfaktor ergibt sich letztlich aus der Normierungsbedingung  $\int_{-\infty}^{\infty} \chi_N^2(\xi) d\xi = 1$ . Angenommen, die Formel ist korrekt für  $\chi_N^2(u)$  mit  $u > 0$ , dann folgt mit der Substitution  $\xi(t) = \frac{u}{2} \cdot (t+1)$  ganz ähnlich wie im schon berechneten Fall  $\chi_2^2 = \chi_1^2 * \chi_1^2$ :

$$\begin{aligned} (\chi_{N+1}^2)(u) &= (\chi_N^2 * \chi_1^2)(u) \propto \int_0^u \xi^{\frac{N}{2}-1} \cdot e^{-\frac{\xi}{2}} \cdot (u - \xi)^{-\frac{1}{2}} \cdot e^{-\frac{u-\xi}{2}} d\xi \\ &= \left(\frac{u}{2}\right)^{\frac{N+1}{2}-1} \cdot e^{-\frac{u}{2}} \cdot \int_{-1}^1 (1+t)^{\frac{N}{2}-1} \cdot (1-t)^{-\frac{1}{2}} dt \\ &\propto \left(\frac{u}{2}\right)^{\frac{N+1}{2}-1} \cdot e^{-\frac{u}{2}} \end{aligned}$$

Das ist aber gerade die nachzuweisende Formel für  $N + 1$  Freiheitsgrade.

Für zwei Freiheitsgrade haben wir die Student'sche  $t$ -Verteilung bereits vorgestellt. Für eine beliebige Anzahl von Freiheitsgraden ist sie wie folgt gegeben.

Seien  $Z$  eine standardnormalverteilte Zufallsvariable und  $Y_N$  eine Chi-Quadrat-verteilte Zufallsvariable mit  $N$  Freiheitsgraden. Wenn  $Z$  und  $Y_N$  unabhängig sind, dann folgt die Zufallsvariable

$$T_N = \frac{\sqrt{N} \cdot Z}{\sqrt{Y_N}}$$

einer **Student'schen  $t$ -Verteilung mit  $N$  Freiheitsgraden**:

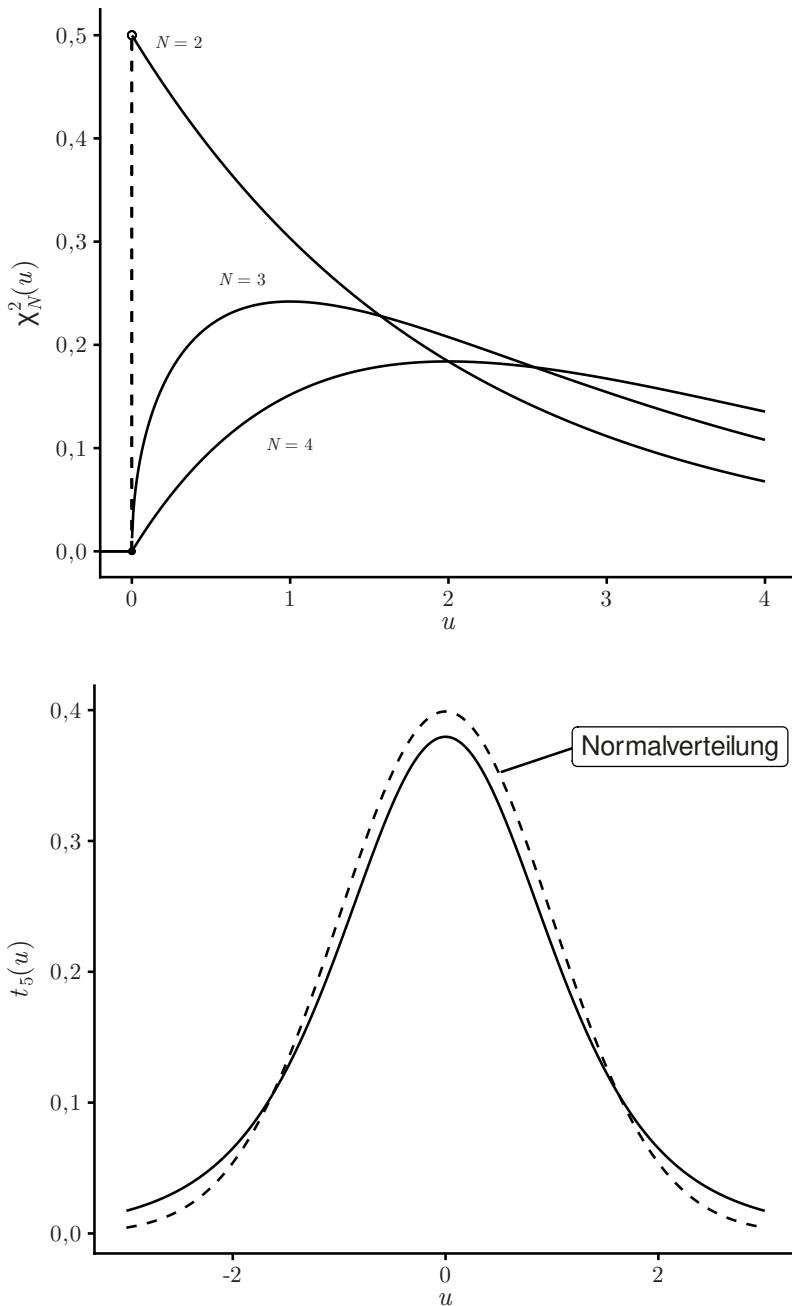
$$p_{T_N}(\cdot) = t_N(\cdot): \mathbb{R} \rightarrow [0, \infty[, t_N(u) = \frac{\Gamma(\frac{N+1}{2})}{\sqrt{N\pi} \cdot \Gamma(\frac{N}{2})} \cdot \left(1 + \frac{u^2}{N}\right)^{-\frac{N+1}{2}}$$

Auf eine genauere Herleitung verzichten wir nun. Die Verteilung ist von besonderer Bedeutung in der statistischen Testtheorie, mehr zu ihrer Anwendung im Abschn. 4.3.3.

Abb. 3.4 zeigt unten eine Student'sche  $t$ -Verteilung für  $N = 5$  Freiheitsgrade. Die  $t$ -Verteilung ist der Standardnormalverteilung sehr ähnlich, und im Grenzwert einer großen Anzahl von Freiheitsgraden gilt:

$$\lim_{N \rightarrow \infty} t_N(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

für alle  $u \in \mathbb{R}$ . Für hinreichend große Werte von  $N$  sind beide Verteilungen in der Praxis kaum unterscheidbar, eine übliche Faustregel ist  $N > 30$ .



**Abb. 3.4.** Chi-Quadrat-Verteilung (oben) und Student'sche  $t$ -Verteilung (unten)

## Quellen

- [1] Alan Hájek. „Interpretations of Probability“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta. Herbst 2019. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/>.
- [2] SKOPOS Institut für Markt- und Kommunikationsforschung GmbH & Co. KG. *Der deutsche Heimtiermarkt 2018 – Struktur und Umsatzdaten*. Aufgerufen am 10. Juli 2020. URL: [https://www.zzf.de/fileadmin/files/ZZF/Marktdaten/ZZF\\_IVH\\_Folder\\_2018\\_Deutscher\\_Heimtiermarkt\\_und\\_Heimtierpopulation\\_UPDATE.pdf](https://www.zzf.de/fileadmin/files/ZZF/Marktdaten/ZZF_IVH_Folder_2018_Deutscher_Heimtiermarkt_und_Heimtierpopulation_UPDATE.pdf).
- [3] Statistisches Bundesamt (Destatis). „Laufende Wirtschaftsrechnungen Ausstattungen privater Haushalte mit ausgewählten Gebrauchsgütern“. In: *Fachserie 15.2* (Dez. 2019), S. 12.
- [4] *Wahlprogramme für die Landtagswahl 2018 in Bayern*. Aufgerufen am 09. Juni 2020. URL: <https://www.wahlen.info/landtagswahl/bayern-wahlprogramme/>.
- [5] Ralf L. Schlenger. „PCR-Tests auf SARS-CoV-2: Ergebnisse richtig interpretieren“. In: *Dtsch Arztebl* 117.24 (Juni 2020), A-1194 / B-1010. URL: <https://www.aerzteblatt.de/int/article.asp?id=214370>.
- [6] Klaus D. Schmidt. *Maß und Wahrscheinlichkeit*. 2. Aufl. Springer, Berlin, Heidelberg, 2011. ISBN: 978-3-642-21026-6. DOI: [10.1007/978-3-642-21026-6](https://doi.org/10.1007/978-3-642-21026-6).
- [7] Matthias Plaue und Mike Scherfner. *Mathematik für das Bachelorstudium II*. Springer Spektrum, Berlin, Heidelberg, 2019. ISBN: 978-3-8274-2068-8. DOI: [10.1007/978-3-8274-2557-7](https://doi.org/10.1007/978-3-8274-2557-7).



## Inferenzstatistik

Eine Grundidee der **Inferenzstatistik** oder **schließenden Statistik** besteht in der Annahme, dass Ausprägungen von Merkmalen in Stichproben Realisierungen von Zufallsvariablen sind. Bei einem Münzwurf ist zum Beispiel die Annahme, dass der Ausgang des Zufallsexperiments durch eine diskrete Zufallsvariable  $X_1$  bestimmt wird, die mit gleicher Wahrscheinlichkeit die Werte Null („Zahl“) oder Eins („Kopf“) annimmt:  $\Pr(X_1 = 0) = \Pr(X_1 = 1) = 1/2$ . Wird das Experiment unter identischen Bedingungen wiederholt, so können wir annehmen, dass dessen Ausgang von einer von  $X_1$  unabhängigen **Stichprobenvariablen**  $X_2$  beschrieben werden kann, welche dieselbe Verteilung aufweist:  $\Pr(X_2 = 0) = \Pr(X_2 = 1) = 1/2$ .

Führen wir das Experiment oft genug durch, so wird die Folge von Beobachtungen von „Kopf“ oder „Zahl“ zum Gegenstand der deskriptiven Statistik: Eine Stichprobe von Ausprägungen eines binären Merkmals. Unsere Erwartung ist, dass beide Ausprägungen mit etwa derselben Häufigkeit vorkommen:  $f_1 = 1 - f_0 \approx 0,5$ .

Unsere Erwartungen gehen jedoch über den rein deskriptiven Aspekt hinaus. Würden wir etwa sehr verschiedene Häufigkeiten beobachten, z. B. eine Häufigkeit des Vorkommens von „Kopf“ mit 70 % gegenüber „Zahl“ mit 30 %, so würden wir – wenn der Stichprobenumfang hinreichend groß ist – davon überzeugt, dass wir unsere ursprüngliche Modellannahme gleichwahrscheinlichen Auftritts anpassen müssen. Wir erlauben uns also auch umgekehrt Rückschlüsse von der Statistik auf die zugrundeliegende, wahre Verteilung.

Das obige Beispiel beschreibt die relative Häufigkeit als **Schätzfunktion** der Wahrscheinlichkeit. Weitere wichtige Schätzfunktionen sind der arithmetische Mittelwert als Schätzer des Erwartungswerts und die empirische Varianz als Schätzer der (theoretischen) Varianz.

Intuitiv wächst unser Vertrauen in die Schätzungen mit dem Stichprobenumfang: Würde oben beschriebenes Ungleichgewicht nur unter zehn Münzwürfen gefunden werden – also sieben Würfe von „Kopf“ und drei von „Zahl“ – so

mag uns das noch nicht davon überzeugen, dass es sich um eine unfaire Trickmünze handelt. Es besteht immer noch die realistische Möglichkeit, dass die sogenannte **Nullhypothese** einer fairen Münze der Wahrheit entspricht. Umgangssprachlich formuliert könnte es sich bei dem Ausgang des Experiments auch um „reinen Zufall“ gehandelt haben. Bei siebzsigmaligem Vorkommen von „Kopf“ unter hundert Würfen würde die Alternativhypothese einer unfairen Münze zu einer wesentlich stärkeren Überzeugung!

Die Inferenzstatistik ermöglicht uns, diese Intuition von einem „Gesetz der großen Zahlen“ mathematisch zu präzisieren. Aus diesen Erkenntnissen können statistische Verfahren abgeleitet werden, mit denen nicht nur eine bloße Schätzung von Kennzahlen wie Wahrscheinlichkeit, Erwartungswert und Varianz möglich ist, sondern auch eine Bemessung unseres Vertrauens in die Richtigkeit der Schätzung.

Des Weiteren befassen wir uns in diesem Kapitel mit dem Thema der statistischen Modellierung: Hierbei geht es im Kern um die Anpassung der Parameter von Familien von Massen- bzw. Dichtefunktionen anhand einer Stichprobe, so dass diese der beobachteten Häufigkeitsverteilung möglichst genau entsprechen.

## 4.1 Statistische Modelle

Für die Zwecke der statistischen Inferenz sind wir besonders an Familien von Massen- und Dichtefunktionen interessiert, also solchen, deren Wert  $p(u)$  noch von einer Reihe von Parametern  $\theta_1, \dots, \theta_K$  abhängt, was wir mit der Schreibweise  $p(u|\theta_1, \dots, \theta_K)$  ausdrücken. Eine solche Familie nennen wir auch ein **(parametrisches) statistisches Modell**. Von einem statistischen Modell wird in der Regel gefordert:

$$p(\cdot | \theta_1, \dots, \theta_K) = p(\cdot | \nu_1, \dots, \nu_K) \Rightarrow \theta_1 = \nu_1, \dots, \theta_K = \nu_K$$

Die Parameter bestimmen jede Massen- bzw. Dichtefunktion innerhalb der Modelfamilie also in eindeutiger Weise: Es ist ausgeschlossen, dass zwei verschiedene Parameterbelegungen auf dieselbe Verteilung führen.

### 4.1.1 Modelle diskreter Zufallsvariablen

Die folgenden statistischen Modelle stellen sämtlich Familien von Massenfunktionen über den natürlichen Zahlen dar: Für jede Belegung der Parameter handelt es sich um eine nichtnegative Funktion  $p: \mathbb{N} \rightarrow \mathbb{R}$ ,  $k \mapsto p(k)$ , deren Werte sich zu eins summieren. Säulendiagramme für eine Auswahl von Parametern sind in Abb. 4.1 dargestellt.

**Diskrete Gleichverteilung über einem Abschnitt der natürlichen Zahlen.**

$$\mathcal{U}(k|L) = \begin{cases} \frac{1}{L} & \text{falls } k \in \{1, \dots, L\} \\ 0 & \text{sonst} \end{cases}$$

mit  $L \in \mathbb{N}$ ,  $L > 0$ .

Wesentliche Charakterisierung einer Gleichverteilung ist, dass jeder mögliche Wert gleichwahrscheinlich ist. Wir können daher auch gleichverteilte Zufallsvariablen mit anderen Trägern als  $\{1, \dots, L\}$  betrachten. Für eine gleichverteilte diskrete Zufallsvariable  $X$  mit Träger  $\text{supp}(X) = \{t_0, \dots, t_{L-1}\}$  berechnet sich der Erwartungswert etwa wie folgt:

$$E[X] = \sum_{\kappa \in \text{supp}(X)} \kappa \cdot \frac{1}{L} = \frac{1}{L} \sum_{k=0}^{L-1} t_k$$

Formal entspricht der Erwartungswert also dem arithmetischen Mittelwert der möglichen Realisierungen der gleichverteilten Variable.

Die Varianz lässt sich ebenso über die aus der deskriptiven Statistik bekannten Formel ermitteln:

$$\sigma^2[X] = E[(X - E[X])]^2 = \frac{1}{L} \sum_{l=0}^{L-1} \left( t_l - \frac{1}{L} \sum_{k=0}^{L-1} t_k \right)^2$$

Das bereits erwähnte Indifferenzprinzip besagt, dass – wenn keine weiteren Informationen vorliegen – mögliche Ergebnisse als gleichverteilt angenommen werden sollten.

Unter allen diskreten Verteilungen, die eine bestimmte endliche Anzahl von verschiedenen Werte annehmen kann, ist die Gleichverteilung die eindeutig bestimmte Verteilung mit maximaler Shannon-Entropie.

**Bernoulli-Verteilung.**

$$\mathcal{B}(k|p, 1) = \begin{cases} 1-p & \text{falls } k = 0 \\ p & \text{falls } k = 1 \\ 0 & \text{sonst} \end{cases}$$

mit  $p \in [0, 1]$ .

Die Bernoulli-Verteilung beschreibt ein sogenanntes **Bernoulli-Experiment**. Dies ist ein Vorgang, der nur zwei Ausgänge zulässt,  $k = 0$  („Fehlschlag“) und  $k = 1$  („Erfolg“). Folglich gibt der Parameter  $p$  die Erfolgswahrscheinlichkeit an.

Das klassische Beispiel für ein Bernoulli-Experiment ist der Münzwurf: Am Ende zeigt die Seite mit Bild („Kopf“) nach oben ( $k = 1$ ) oder die Seite mit „Zahl“ ( $k = 0$ ). Bei einer gewöhnlichen Münze können wir davon ausgehen, dass beide Ausgänge etwa gleich wahrscheinlich sind:  $p \approx \frac{1}{2}$ . Im Allgemeinen könnte es sich aber auch um eine unfaire Trickmünze handeln, bei der z. B. „Kopf“ signifikant wahrscheinlicher als „Zahl“ ist, sodass der Parameter  $p$  auch andere Werte annehmen kann.

Im Sinne der deskriptiven Statistik sind die Realisierungen einer Bernoulli-Variable die Ausprägungen eines binären Merkmals.

Erwartungswert und Varianz einer Bernoulli-verteilten Zufallsvariablen  $X$  berechnen sich wie folgt:

$$\begin{aligned} E[X] &= 0 \cdot (1 - p) + 1 \cdot p = p, \\ \sigma^2[X] &= (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p = p(1 - p) \end{aligned}$$

Die folgende Verteilung beschreibt die Wahrscheinlichkeit, unter einer Reihe von  $N$  Münzwürfen genau  $k$ -mal „Kopf“ vorzufinden. Sie wird in Abschn. 4.2.1 eine bedeutende Rolle spielen, wenn wir den Zusammenhang zwischen der Häufigkeit einer Beobachtung und der Wahrscheinlichkeit für deren Auftreten herstellen.

### Binomialverteilung.

$$\mathcal{B}(k|p, N) = \begin{cases} \binom{N}{k} p^k (1 - p)^{N-k} & \text{falls } k \in \{0, \dots, N\} \\ 0 & \text{sonst} \end{cases}$$

mit  $p \in [0, 1]$ ,  $N \in \mathbb{N}$ .

Für den Fall, dass  $p = 0$  oder  $p = 1$ , soll die Konvention „ $0^0 = 0$ “ gelten. Mit „ $\binom{N}{k}$ “ ist hier der **Binomialkoeffizient** gemeint, der wie folgt definiert ist:

$$\binom{N}{k} = \frac{N!}{k!(N - k)!}$$

Dabei ist  $n! = n \cdot (n - 1) \cdots 2 \cdot 1$  für jede natürliche Zahl  $n$  deren Fakultät, per Definition gilt  $0! = 1$ . Der Binomialkoeffizient entspringt kombinatorischen Überlegungen und entspricht der Anzahl aller Teilmengen mit  $k$  Elementen einer Obermenge mit  $N$  Elementen. Oder im Rahmen eines Zufallsexperiments mit  $N$  Münzwürfen: Die Anzahl der möglichen Ausgänge des Experiments, bei denen genau  $k$ -mal „Kopf“ vorkommt.

Die Bernoulli-Verteilung ist mit der Binomialverteilung für den Fall  $N = 1$  identisch. Für Erwartungswert und Varianz einer binomial verteilten Zufallsvariable  $X \sim \mathcal{B}(\cdot | p, N)$  gilt:

$$E[X] = Np, \sigma^2[X] = Np(1 - p)$$

**Geometrische Verteilung.**

$$\text{Geom}(k|p) = p(1-p)^k$$

mit  $p \in \mathbb{R}$ ,  $0 < p \leq 1$ .

Die Wahrscheinlichkeit, dass sich in einer Reihe von Münzwürfen das erstmalige Auftreten von „Kopf“ nach genau  $k$ -maligem Wurf von „Zahl“ ereignet, ist durch  $\text{Geom}(k|p)$  gegeben.

Für Erwartungswert und Varianz einer geometrisch verteilten Zufallsvariable  $X \sim \text{Geom}(\cdot|p)$  gilt:

$$E[X] = \frac{1-p}{p}, \quad \sigma^2[X] = \frac{1-p}{p^2}$$

**Poisson-Verteilung.**

$$\text{Pois}(k|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

mit  $\lambda \in \mathbb{R}$ ,  $\lambda > 0$ .

Für Erwartungswert und Varianz einer geometrisch verteilten Zufallsvariable  $X \sim \text{Pois}(\cdot|\lambda)$  gilt:

$$E[X] = \sigma^2[X] = \lambda$$

Die Poisson-Verteilung ergibt sich aus der Binomialverteilung im Grenzwert sehr vieler Bernoulli-Experimente mit Erfolgswahrscheinlichkeit  $p = \frac{\lambda}{N}$ . Dies zeigt die folgende Rechnung:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{B}\left(k \left| \frac{\lambda}{N}, N\right.\right) &= \lim_{N \rightarrow \infty} \left[ \frac{N!}{k!(N-k)!} \cdot \left(\frac{\lambda}{N}\right)^k \cdot \left(1 - \frac{\lambda}{N}\right)^{N-k} \right] \\ &= \frac{\lambda^k}{k!} \cdot \lim_{N \rightarrow \infty} \left[ \frac{N(N-1)\cdots(N-k+1)}{N^k} \cdot \left(1 - \frac{\lambda}{N}\right)^{N-k} \right] \\ &= \frac{\lambda^k}{k!} \cdot \lim_{N \rightarrow \infty} \left[ \left(1 - \frac{\lambda}{N}\right)^{-k} \left(1 - \frac{\lambda}{N}\right)^N \right] \\ &= \frac{\lambda^k}{k!} \cdot \lim_{N \rightarrow \infty} \left(1 - \frac{\lambda}{N}\right)^N = \frac{\lambda^k}{k!} e^{-\lambda} = \text{Pois}(k|\lambda) \end{aligned}$$

Hierbei verwendeten wir unter anderem die für alle  $x \in \mathbb{R}$  gültige Identität  $e^x = \lim_{N \rightarrow \infty} \left(1 + \frac{x}{N}\right)^N$ . Der Parameter  $\lambda$  kann dann als durchschnittliche Häufigkeit für das Eintreffen von „Erfolg“ gedeutet werden.

### 4.1.2 Modelle stetiger Zufallsvariablen

Im Folgenden sind Modelle für stetige Zufallsvariablen aufgelistet: Für jede Belegung der Parameter handelt es sich um eine nichtnegative Funktion  $p: \mathbb{R} \rightarrow \mathbb{R}$ ,  $u \mapsto p(u)$ , deren Werte sich zu eins aufintegrieren. Grafische Abbildungen der Wahrscheinlichkeitsdichtefunktionen für ausgewählte Parameter finden sich in Abb. 4.2.

**Stetige Gleichverteilung.**

$$\mathcal{U}(u|a, b) = \begin{cases} \frac{1}{b-a} & \text{falls } u \in [a, b] \\ 0 & \text{sonst} \end{cases}$$

mit  $a, b \in \mathbb{R}$ ,  $a < b$ .

Die stetige Gleichverteilung können wir uns als Grenzwert der diskreten Gleichverteilung für viele Sprungstellen in einem festen Intervall vorstellen: Seien die Zufallsvariablen  $X_L$  diskret gleichverteilt mit jeweiliger Verteilungsfunktion  $F_L(\cdot) := F_{X_L}(\cdot)$ , welche  $L$  äquidistanten Sprungstellen in den fest gewählten Intervallgrenzen  $a$  und  $b$  aufweist:  $t_l = a + \frac{l}{L}(b - a)$ ,  $l \in \{0, \dots, L - 1\}$ .

An den Sprungstellen nimmt die jeweilige Verteilungsfunktion die Werte  $F_L(t_l) = \frac{t_l - a}{b - a}$  an. Für beliebige Argumente  $u$  mit  $a \leq u < b$  gilt im Grenzwert einer immer feineren Unterteilung des Intervalls:

$$\begin{aligned} F(u|a, b) &:= \lim_{L \rightarrow \infty} F_L(u) = \lim_{L \rightarrow \infty} \frac{1}{b - a} \cdot \left( u - a - \min_{t_l \leq u} \{u - t_l\} \right) \\ &= \frac{u - a}{b - a} \end{aligned}$$

wegen

$$0 \leq \min_{t_l \leq u} \{u - t_l\} \leq \frac{1}{2L(b - a)} \rightarrow 0.$$

Für  $u < a$  ist  $F(u|a, b) = 0$  und für  $u \geq b$  gilt  $F(u|a, b) = 1$ . Das ist aber gerade die Verteilungsfunktion der *stetigen* Gleichverteilung:

$$F(u|a, b) = \begin{cases} 0 & \text{falls } u < a \\ \frac{u - a}{b - a} & \text{falls } a \leq u < b \\ 1 & \text{falls } b \leq u \end{cases} = \int_{-\infty}^u \mathcal{U}(\xi|a, b) d\xi$$

Erwartungswert und Varianz einer stetig gleichverteilten Zufallsvariablen  $X \sim \mathcal{U}(\cdot|a, b)$  sind wie folgt gegeben:

$$E[X] = \frac{1}{2}(b + a), \sigma^2[X] = \frac{1}{12}(b - a)^2$$

### Normalverteilung, Gauß-Verteilung.

$$\mathcal{N}(u|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{u-\mu}{\sigma})^2}$$

mit  $\mu, \sigma \in \mathbb{R}, \sigma > 0$ .

Erwartungswert und Varianz einer normalverteilten Zufallsvariablen  $X \sim \mathcal{N}(\cdot|\mu, \sigma^2)$  sind durch den Lage- bzw. Skalenparameter  $\mu$  bzw.  $\sigma$  bestimmt:

$$E[X] = \mu, \sigma^2[X] = \sigma^2$$

Die Normalverteilung mit Erwartungswert  $\mu = 0$  und Varianz  $\sigma^2 = 1$  wird **Standardnormalverteilung** genannt.

Eine wichtige Eigenschaft der Klasse der normalverteilten Zufallsvariablen besteht darin, dass diese abgeschlossen gegenüber Linearkombinationen ist: Sind  $X_1, \dots, X_N$  normalverteilt, dann ist auch  $Y = \sum_{n=1}^N a_n X_n$  mit beliebigen Zahlen  $a_1, \dots, a_N \in \mathbb{R}$  normalverteilt (wenn nicht alle Koeffizienten gleich null sind, versteht sich). Sind die Zufallsvariablen zudem noch unabhängig, dann gilt genauer:

$$Y = \sum_{n=1}^N a_n X_n \sim \mathcal{N}(\cdot|\mu_Y, \sigma_Y^2) \text{ mit } \mu_Y = \sum_{n=1}^N a_n \mu_n \text{ und } \sigma_Y^2 = \sum_{n=1}^N (a_n \sigma_n)^2,$$

wobei  $\mu_1, \dots, \mu_N$  die Erwartungswerte und  $\sigma_1^2, \dots, \sigma_N^2$  die Varianzen von  $X_1, \dots, X_N$  sind.

Sind  $X_1, \dots, X_N$  nicht nur unabhängig, sondern folgen auch derselben Normalverteilung mit identischem Erwartungswert  $\mu$  und Varianz  $\sigma^2$ , dann gilt insbesondere für deren Summe:

$$\sum_{n=1}^N X_n = X_1 + \dots + X_N \sim \mathcal{N}(\cdot|N \cdot \mu, N \cdot \sigma^2)$$

Später wird für uns auch die folgende Zufallsvariable von besonderem Interesse sein:

$$Z = \sqrt{N} \cdot \frac{\frac{1}{N} \sum_{n=1}^N X_n - \mu}{\sigma} = \frac{-N\mu + \sum_{n=1}^N X_n}{\sqrt{N}\sigma}$$

Unter den gegebenen Voraussetzungen unabhängiger normalverteilter Zufallsgrößen ist diese stets standardnormalverteilt:

$$\begin{aligned} p_Z(u) &= \sqrt{N}\sigma \cdot \mathcal{N}\left(\sqrt{N}\sigma \cdot u + N \cdot \mu | N \cdot \mu, N \cdot \sigma^2\right) \\ &= \sqrt{N}\sigma \cdot \frac{1}{\sqrt{2\pi}\sqrt{N}\sigma} \exp\left(-\frac{1}{2} \left(\frac{\sqrt{N}\sigma u + N\mu - N\mu}{\sqrt{N}\sigma}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} = \mathcal{N}(u|0, 1) \end{aligned}$$

Der später besprochene zentrale Grenzwertsatz macht die überraschende Aussage, dass für große  $N$  die Variable  $Z$  unter recht allgemeinen Umständen auch dann annähernd normalverteilt ist, wenn  $X_1, \dots, X_N$  es nicht sind. Aufgrund dieser Sonderrolle kommt der Normalverteilung eine besondere Bedeutung zu.

Eine weitere Charakterisierung der Gauß-Verteilung kann ebenfalls dazu geeignet sein, ihre häufige Verwendung in der Praxis zu begründen: Unter allen Verteilungen von stetigen Zufallsvariablen mit fest vorgegebenen Werten für Erwartungswert und Varianz handelt es sich um die Verteilung mit maximaler differenzieller Entropie.<sup>1</sup>

Das Indifferenzprinzip besagt, dass bei vollständig fehlenden Informationen eine Gleichverteilung der möglichen Beobachtungen vorausgesetzt werden sollte. Eine Verallgemeinerung des Indifferenzprinzips stellt das **Prinzip der maximalen Entropie** dar. Dieses fordert, bei mangelnden Informationen über eine Verteilung jene mit maximaler Entropie zu wählen. Folglich geht die Normalverteilung aus diesem Prinzip hervor, wenn nur Erwartungswert und Varianz einer stetigen Zufallsgröße bekannt sind.

### Cauchy-Lorentz-Verteilung.

$$\mathcal{L}(u|x_0, \gamma) = \frac{1}{\pi\gamma} \cdot \frac{1}{1 + \left(\frac{u-x_0}{\gamma}\right)^2}$$

mit  $x_0, \gamma \in \mathbb{R}, \gamma > 0$ .

Die Cauchy-Lorentz-Verteilung (auch kürzer: Cauchy-Verteilung) ist ein Beispiel für eine Klasse stetiger Verteilungen, bei der weder Erwartungswert noch Varianz existieren. Es gilt zwar für alle Parameterbelegungen  $x_0, \gamma$  die notwendige Normierungsbedingung einer Dichtefunktion:

$$\begin{aligned} \int_{-\infty}^{\infty} \mathcal{L}(\xi|x_0, \gamma) d\xi &= \int_{-\infty}^{\infty} \frac{1}{\pi\gamma} \cdot \frac{1}{1 + \left(\frac{\xi-x_0}{\gamma}\right)^2} d\xi = \frac{1}{\pi} \cdot \int_{-\infty}^{\infty} \frac{1}{1+t^2} dt \\ &= \frac{1}{\pi} \cdot \left( \int_{-\infty}^0 \frac{1}{1+t^2} dt + \int_0^{\infty} \frac{1}{1+t^2} dt \right) \\ &= \frac{1}{\pi} \cdot \left( \lim_{t \rightarrow -\infty} (\arctan(0) - \arctan(t)) + \right. \\ &\quad \left. \lim_{t \rightarrow \infty} (\arctan(t) - \arctan(0)) \right) \\ &= \frac{1}{\pi} \cdot \left( 0 - \left(-\frac{\pi}{2}\right) + \frac{\pi}{2} - 0 \right) = 1 \end{aligned}$$

<sup>1</sup> Diese Eigenschaft der Normalverteilung lässt sich mit Methoden der Variationsrechnung nachweisen: Es gilt, dass Funktional  $H[p] = - \int_{-\infty}^{\infty} p(\xi) \cdot \ln(p(\xi)) d\xi$  unter den Nebenbedingungen  $\int_{-\infty}^{\infty} p(\xi) d\xi = 1$ ,  $\int_{-\infty}^{\infty} \xi \cdot p(\xi) d\xi = \mu$  und  $\int_{-\infty}^{\infty} (\xi - \mu)^2 \cdot p(x) d\xi = \sigma^2$  zu maximieren.

Die entsprechenden uneigentlichen Integrale für Erwartungswert und Varianz sind jedoch nicht konvergent. Allerdings existieren Median und Modus der Cauchy-Verteilung, beide sind durch den Lageparameter  $x_0$  gegeben.

Obwohl der Graph der Dichtefunktion jenem der Normalverteilung (einer „Gauß-Glocke“) ähnlich sieht (siehe Abb. 4.2 links), ist sie im Gegensatz zu dieser eine **endlastige Wahrscheinlichkeitsdichte**: Werte weit ab vom Median  $x_0$  (also **Ausreißer**) einer Cauchy-verteilten Zufallsvariable sind sehr viel wahrscheinlicher als bei einer Gauß-verteilten Zufallsvariable.

### Pareto-Verteilung.

$$\text{Par}(u|x_{\min}, \alpha) = \begin{cases} \alpha x_{\min}^\alpha \cdot u^{-(\alpha+1)} & \text{falls } u \geq x_{\min} \\ 0 & \text{sonst} \end{cases}$$

mit  $x_{\min}, \alpha \in ]0, \infty[$ .

Eine Pareto-verteilte Zufallsvariable  $X \sim \text{Par}(\cdot | x_{\min}, \alpha)$  besitzt den folgenden Erwartungswert:

$$E[X] = \begin{cases} \infty & \text{falls } 0 < \alpha \leq 1 \\ \frac{\alpha x_{\min}}{\alpha - 1} & \text{falls } 1 < \alpha \end{cases}$$

Die Varianz ist wie folgt gegeben:

$$\sigma^2[X] = \begin{cases} \infty & \text{falls } 0 < \alpha \leq 2 \\ \frac{\alpha(x_{\min})^2}{(\alpha-1)(\alpha-2)} & \text{falls } 2 < \alpha \end{cases}$$

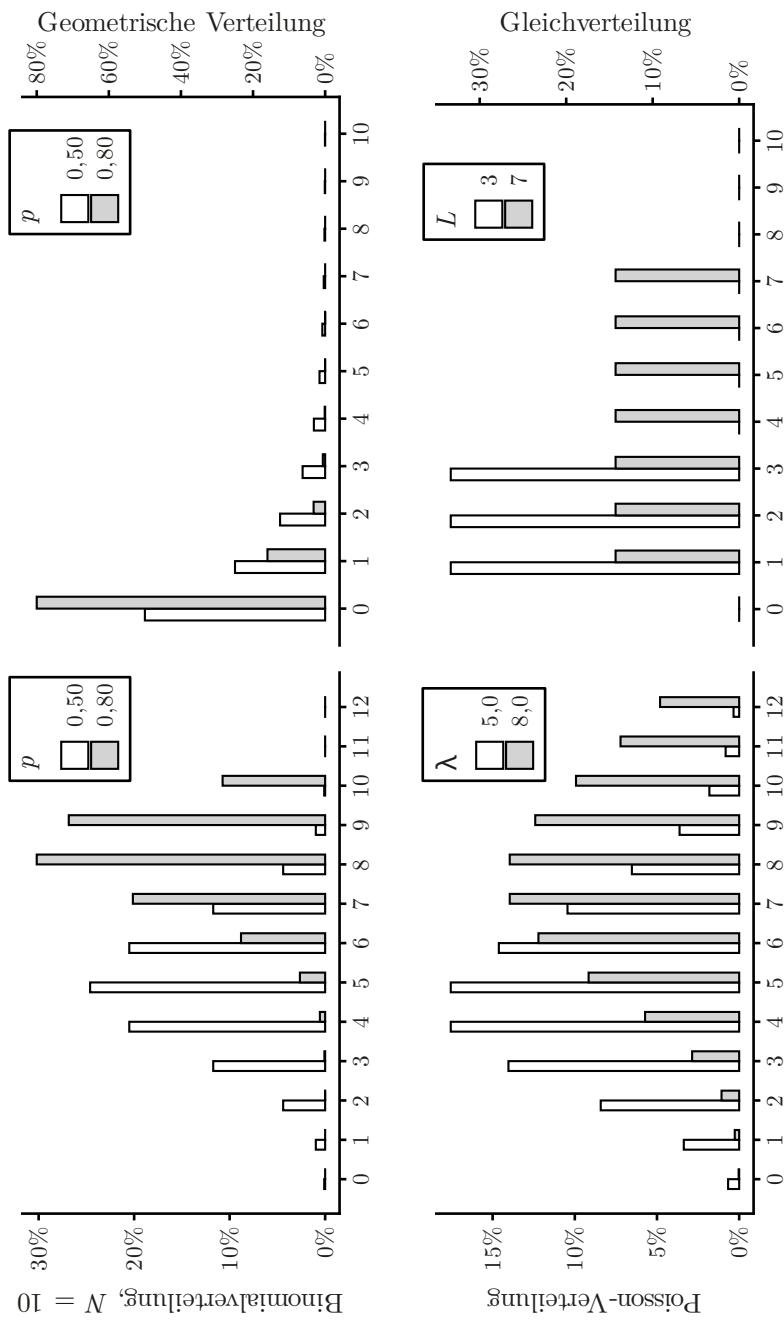
## 4.2 Gesetze der großen Zahlen

Intuitiv ist die relative Häufigkeit einer Merkmalsausprägung in einer Stichprobe eine gute Schätzung für die Wahrscheinlichkeit, dass sich gerade diese Beobachtung ereignet. Wir können beispielsweise einen Spielwürfel viele Male werfen und werden in der Regel feststellen, dass die Augenzahl sechs in etwa mit der relativen Häufigkeit  $1/6 \approx 0,167$  fällt. Daraus schließen wir, dass auch die Wahrscheinlichkeit für dieses Ereignis näherungsweise bei 0,167 liegt.

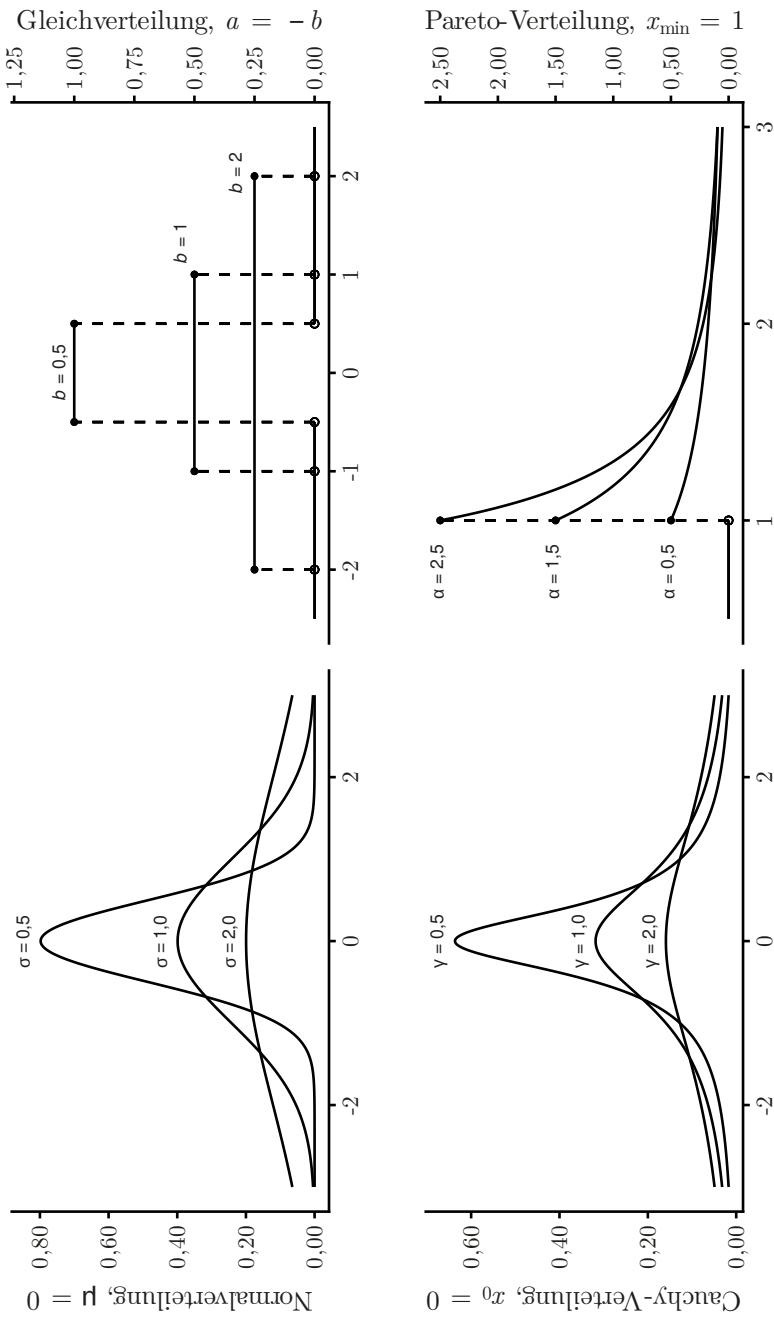
Gleiches gilt für den Erwartungswert und andere Eigenschaften von Wahrscheinlichkeitsverteilungen: Unsere Hoffnung ist, diese durch entsprechende Kennzahlen einer beobachteten Häufigkeitsverteilung sinnvoll schätzen zu können.

Intuitiv lässt sich auch die Tatsache erschließen, dass solche Schätzungen um so genauer werden, je mehr Beobachtungen vorliegen, je größer also die Stichprobe ist. Unsere Intuition werden wir im Folgenden mit mathematischen Argumenten unterfüttern und präzisieren. Ergebnisse dieser Art sind als Gesetze der großen Zahlen bekannt.

Abb. 4.1. Massenfunktionen von parametrischen Modellen diskreter Zufallsvariablen



**Abb. 4.2.** Dichtefunktionen von parametrischen Modellen stetiger Zufallsvariablen



### 4.2.1 Bernoulli'sches Gesetz der großen Zahlen

Wir möchten im Folgenden zeigen, dass unter recht allgemeinen Umständen die Wahrscheinlichkeit eines Ereignisses über die relative Häufigkeit wiederholter Beobachtung gut geschätzt werden kann. Der entsprechende Lehrsatz wird auch **Bernoulli'sches Gesetz der großen Zahlen** genannt. Besonders wichtig ist hierbei, dass wir dann auch in der Lage sein werden, Kennzahlen für die Güte dieser Schätzung anzugeben.

Wir betrachten eine diskrete Zufallsvariable  $X_1$ , die mit der Wahrscheinlichkeit  $0 \leq p \leq 1$  den Wert Eins und mit Wahrscheinlichkeit  $1 - p$  den Wert Null annimmt. Folglich genügt die Variable einer Bernoulli-Verteilung. Wir können uns der Anschaulichkeit halber einen Münzwurf mit nicht notwendigerweise fairer Münze vorstellen: Die Ereignisse  $(X_1)^{-1}(1)$  und  $(X_1)^{-1}(0)$  entsprechen den möglichen Ausgängen „Kopf“ bzw. „Zahl“, welche wir auch als „Erfolg“ oder „Misserfolg“ bezeichnen.

Weiterhin gehen wir davon aus, dass bei Wiederholung des Münzwurfs unter denselben Bedingungen die möglichen Ausgänge mit denselben Wahrscheinlichkeiten auftreten, und dass diese als Ereignisse unabhängig vom ersten Münzwurf sind. Stellen wir uns die Ausgänge des zweiten Münzwurfs als Realisierungen einer Zufallsvariablen  $X_2$  vor, so bedeutet das:

$$\begin{aligned}\Pr(X_1 = 1) &= \Pr(X_2 = 1) = p, \\ \Pr(X_1 = 1, X_2 = 1) &= \Pr(X_1 = 1) \cdot \Pr(X_2 = 1) = p^2 \\ \Pr(X_1 = 1, X_2 = 0) &= p \cdot (1 - p) \\ \Pr(X_1 = 0, X_2 = 1) &= (1 - p) \cdot p \\ \Pr(X_1 = 0, X_2 = 0) &= (1 - p) \cdot (1 - p)\end{aligned}$$

Nach insgesamt  $N$  Münzwürfen, beschrieben durch unabhängige und identisch Bernoulli-verteilte Zufallsvariablen  $X_1, \dots, X_N$ , ist die absolute Häufigkeit des Auftretens von „Kopf“ durch die Zufallsvariable  $n(N) := \sum_{l=1}^N X_l$  gegeben. Wir wollen die Verteilung von  $n(N)$  ermitteln. Die Wahrscheinlichkeit für eine ganz bestimmte Kombination mit  $k$ -maligem Vorkommen von „Kopf“ und  $(N - k)$ -maligem Vorkommen von „Zahl“ ist zunächst einmal:

$$\Pr(X_1 = 1, \dots, X_k = 1, X_{k+1} = 0, \dots, X_N = 0) = p^k (1 - p)^{N-k}$$

Durch kombinatorische Überlegungen kann ermittelt werden, dass es insgesamt  $\binom{N}{k} = \frac{N!}{k!(N-k)!}$  mögliche Kombinationen mit genau  $k$  Würfen von „Kopf“ gibt. Daher ist die Wahrscheinlichkeit für  $k$ -maligen Erfolg, ungeachtet der genauen Reihenfolge der Ausgänge der einzelnen Bernoulli-Experimente:

$$\Pr(n(N) = k) = \binom{N}{k} \cdot p^k (1 - p)^{N-k}$$

Folglich ist  $n(N)$  eine binomial verteilte Zufallsgröße. Der Erwartungswert der relativen Häufigkeit  $f(N) = N^{-1} \cdot n(N)$  des Vorkommens von „Kopf“ bei  $N$  Würfen bestimmt sich somit wie folgt:

$$E[f(N)] = E\left[\frac{n(N)}{N}\right] = \frac{1}{N} \cdot E[n(N)] = \frac{1}{N} \cdot Np = p$$

Das bedeutet, dass der Erwartungswert der relativen Häufigkeit des Erfolgs tatsächlich der Erfolgswahrscheinlichkeit der Einzelereignisse entspricht.

Hierzu wird auch gesagt: Die relative Häufigkeit  $f(N)$  ist ein **erwartungstreuer Schätzer** des Parameters  $p$ .

Um eine Güte für diese Schätzung zu ermitteln, benötigen wir die Varianz:

$$\sigma^2[f(N)] = \frac{1}{N^2} \cdot \sigma^2[n(N)] = \frac{1}{N^2} \cdot Np(1-p) = \frac{p(1-p)}{N}$$

Wir können nun die Tschebyscheff'sche Ungleichung anwenden:

$$\Pr(|f(N) - E[f(N)]| \geq r) \leq \frac{\sigma^2[f(N)]}{r^2},$$

folglich

$$\Pr(|f(N) - p| \geq r) \leq \frac{p(1-p)}{Nr^2} \leq \frac{1}{4Nr^2}.$$

Die praktischen Konsequenzen dieser Abschätzung können anhand eines Beispiels verdeutlicht werden. Angenommen, wir wollen die Wahrscheinlichkeit  $p$  für das Auftreten des Ereignisses „Kopf“ bis auf einen maximalen Fehler von  $r = 0,05$  genau abschätzen. Da es sich um einen zufälligen Prozess handelt, können wir uns nie ganz sicher sein, dass diese Fehlerschranke eingehalten wird: Wir könnten das Pech haben, dass hundert Mal hintereinander „Kopf“ fällt, obwohl die Münze fair ist. Wir können jedoch die Wahrscheinlichkeit für einen solchen „Unfall“ abschätzen und zum Beispiel verlangen, dass diese höchstens  $\alpha = 1\%$  beträgt. Damit ergibt sich eine minimale Stichprobengröße, die diese Parameter sicherstellt:  $N \geq \frac{1}{4\alpha r^2}$ . Bei den vorliegenden Beispielwerten bedeutet das  $N \geq 10.000$ .

Zusammengefasst: Wird die Münze 10.000 Mal geworfen, so werden die beobachtete relative Häufigkeit und die tatsächliche Wahrscheinlichkeit eines Vorkommens von „Kopf“ mit einer Wahrscheinlichkeit von 99 % höchstens um  $\pm 0,05$  voneinander abweichen. Wir können uns bei dieser Stichprobengröße also sehr sicher sein, dass die Schätzung hinreichend genau ist.

Darüber hinaus folgt aus der Tschebyscheff'schen Ungleichung für alle  $\varepsilon > 0$ :

$$\lim_{N \rightarrow \infty} \Pr(|f(N) - p| < \varepsilon) = 1$$

Das heißt, bei vorgegebener Fehlerschranke wird eine zu große Abweichung immer unwahrscheinlicher, im Grenzfall beliebig großer Stichprobe wird eine präzise Schätzung **fast sicher**. Hierzu wird auch gesagt: Die relative Häufigkeit  $f(N)$  ist ein **konsistenter Schätzer** des Parameters  $p$ .

Zusammengefasst ergibt sich somit:

**Bernoulli'sches Gesetz der großen Zahlen.** Die relative Häufigkeit des Erfolgs einer Reihe von unabhängigen Bernoulli-Experimenten ist ein erwartungstreuer und konsistenter Schätzer für die Erfolgswahrscheinlichkeit des Experiments.

**Anwendungsbeispiel.** Insgesamt haben  $N = 436.323$  der für die CDC-Studie [1] befragten Personen eine Angabe zu ihrem Geschlecht gemacht. Von diesen Personen gaben  $n_1 = 238.911$  an, weiblichen Geschlechts zu sein. Das entspricht einer relativen Häufigkeit bzw. empirischen Wahrscheinlichkeit von  $f = \frac{n_1}{N} = \frac{238.911}{436.323} = 54,8\%$ .

Den *tatsächlichen* Anteil erwachsener Personen weiblichen Geschlechts unter US-Bürgern könnte mit absoluter statistischer Genauigkeit nur durch eine sogenannte **Vollerhebung** ermittelt werden, bei der das Geschlecht aller etwa 300 Millionen US-Bürger ermittelt würde.

Die Wahrscheinlichkeit, dass unsere Angabe von  $f = 54,8\%$  um mehr als  $\Delta f = 1,0\%$  vom tatsächlichen Anteil abweicht, beträgt aus Sicht der statistischen Schätztheorie jedoch höchstens  $\alpha = \frac{1}{4N(\Delta f)^2} = 0,57\%$ . Es gilt jedoch auch zu beachten, dass die Wahrscheinlichkeit, dass die Schätzung um mehr als  $\Delta f = 0,1\%$  daneben liegt, bereits bei  $\alpha = 57\%$  liegt. Mithin ist es im Ergebnis nicht sinnvoll, eine auf drei geltende Ziffern genaue Angabe zu machen; es sollte vielmehr  $f = 55\%$  geschrieben werden.

Unabhängige Statistiken führen dennoch auf andere Zahlen für den Anteil der Personen weiblichen Geschlechts in der US-Bevölkerung, mit Abweichungen weit jenseits der oben berechneten statistischen Fehler. So kommt eine Statistik der Vereinten Nationen zu dem Schluss, dass der Anteil bei  $f = 50,5\%$  liegt (Stand 2018 [2]). Obwohl statistische Fehlerabschätzungen von großem Nutzen sein können, sollte stets beachtet werden, dass auch systematische Einflüsse wie Studiendesign oder Datenqualität entscheidend sein können.

#### 4.2.2 Tschebyscheff'sches Gesetz der großen Zahlen

Ebenso wie die relative Häufigkeit (unter bestimmten Voraussetzungen) eine geeignete Schätzfunktion für die Wahrscheinlichkeit darstellt, so ist das arithmetische Mittel eine geeignete Schätzfunktion für den Erwartungswert. Nicht umsonst wird das arithmetische Mittel daher auch empirischer Erwartungswert genannt.

Es ist wichtig zu verstehen, dass diese Schätzung selbst wiederum eine Zufallsvariable darstellt: Angenommen, wir werfen einen Spielwürfel 50-mal und notieren die mittlere Augenzahl. Wenn wir dieses Experiment wiederholen, so erhalten wir mit großer Wahrscheinlichkeit einen etwas anderen Wert. Wiederholen wir es viele Male, so werden die notierten arithmetischen Mittel einer Verteilung entsprechen. Die notierten Mittelwerte werden sich gerade für große

Stichproben trotzdem nicht wesentlich unterscheiden. Dies ist die Kernaussage des **Tschebyscheff'schen Gesetzes der großen Zahlen**, das wir im Folgenden zeigen wollen.

Wir gehen davon aus, dass jede der Merkmalsausprägungen  $x_1, \dots, x_N$  in einer Stichprobe, deren arithmetischen Mittelwert wir bestimmen wollen, eine Realisierung einer entsprechenden Zufallsvariable  $X_n$ ,  $n = 1, \dots, N$  darstellt. Die  $X_n$  werden daher auch **Stichprobenvariablen** genannt.

Eine übliche Grundannahme ist, dass die Stichprobenvariablen identische Verteilungsfunktion haben und unabhängig voneinander sind. Dies ist in vielen Fällen eine sinnvolle Annahme, wenn die Beobachtungen die Folge von identisch präparierten und voneinander unabhängigen Zufallsexperimenten oder -prozessen sind oder eine zufällige Auswahl aus einer statistischen Grundgesamtheit darstellen. Es sollte jedoch beachtet werden, dass diese Annahme dennoch – auch oder gerade in der Praxis – nicht immer gerechtfertigt ist.

Außerdem gehen wir davon aus, dass der (für alle  $X_n$  notwendigerweise identische) Erwartungswert  $\mu = E[X_n]$  existiert, ebenso wie die (endliche) Varianz  $\sigma^2 = \sigma^2[X_n]$ .

Der arithmetische Mittelwert  $\bar{x}$  der  $N$  Beobachtungen  $x_1, \dots, x_N$  ist eine Realisierung der folgenden Zufallsvariable:

$$\bar{X} = \bar{X}(N) = \frac{1}{N} \sum_{n=1}^N X_n$$

Wir fassen den arithmetischen Mittelwert nun also als eine Schätzfunktion in den Stichprobenvariablen auf. Damit wird dieser bei einer Belegung mit spezifischen  $X_1, \dots, X_N$  selbst zu einer Zufallsvariablen, deren Erwartungswert und Varianz wir bestimmen wollen.

Aus der Linearität des Erwartungswerts folgt sofort:

$$E[\bar{X}(N)] = E\left[\frac{1}{N} \sum_{n=1}^N X_n\right] = \frac{1}{N} \sum_{n=1}^N E[X_n] = \mu$$

Das arithmetische Mittel ist also ein erwartungstreuer Schätzer des Erwartungswerts  $\mu$ .

Da die Stichprobenvariablen als unabhängig vorausgesetzt werden, sind diese außerdem paarweise unkorreliert:

$$E[X_k \cdot X_l] = E[X_k] \cdot E[X_l]$$

für alle  $k, l = 1, \dots, N$  mit  $k \neq l$ .

Es ist nicht schwer einzusehen, dass für paarweise unkorrelierte Zufallsvariablen die Varianz wie folgt dargestellt werden kann; alle Kreuzterme verschwinden:

$$\sigma^2(X_1 + \dots + X_N) = \sigma^2(X_1) + \dots + \sigma^2(X_N)$$

Damit ergibt sich für die Varianz der Schätzfunktion des arithmetischen Mittelwerts:

$$\sigma^2[\bar{X}(N)] = \sigma^2 \left[ \frac{1}{N} \sum_{n=1}^N X_n \right] = \frac{1}{N^2} \sum_{n=1}^N \sigma^2[X_n] = \frac{\sigma^2}{N}$$

Wir können nun wieder die Tschebyscheff'sche Ungleichung anwenden:

$$\begin{aligned} \Pr(|\bar{X}(N) - E[\bar{X}(N)]| < \varepsilon) &\geq 1 - \frac{\sigma[\bar{X}(N)]^2}{\varepsilon^2} \Rightarrow \\ \Pr(|\bar{X}(N) - \mu| < \varepsilon) &\geq 1 - \frac{\sigma^2}{\varepsilon^2 N} \Rightarrow \\ \lim_{N \rightarrow \infty} \Pr(|\bar{X}(N) - \mu| < \varepsilon) &= 1 \end{aligned}$$

Dies zeigt, dass der arithmetische Mittelwert einen konsistenten Schätzer des Erwartungswerts darstellt: Für große Stichproben ist es unwahrscheinlich, dass sich der Erwartungswert der zugrundeliegenden Verteilung und das arithmetische Mittel der Stichprobe wesentlich unterscheiden.

Zusammengefasst gilt also:

**Tschebyscheff'sches Gesetz der großen Zahlen.** Seien  $X_1, \dots, X_N$  unabhängige und identisch verteilte Stichprobenvariablen mit endlichem Erwartungswert  $\mu$  und endlicher Varianz.

Der arithmetische Mittelwert  $\bar{X} = \frac{1}{N}(X_1 + \dots + X_N)$  ist dann ein erwartungstreuer und konsistenter Schätzer des Erwartungswerts  $\mu$ .

#### 4.2.3 Varianzschätzung und Bessel-Korrektur

Ähnliche Überlegungen wie für das arithmetische Mittel bzw. den Erwartungswert können wir für die empirische bzw. theoretische Varianz anstellen. Die Schätzfunktion der empirischen Varianz einer Reihe von Stichprobenvariablen  $X_1, \dots, X_N$  ist die folgende Zufallsvariable:

$$S^2 = S^2(N) = \frac{1}{N} \sum_{n=1}^N (X_n - \bar{X}(N))^2$$

Deren Erwartungswert berechnet sich wie folgt:

$$\begin{aligned} E[S^2] &= \frac{1}{N} \sum_{n=1}^N E[(X_n - \bar{X})^2] \\ &= \frac{1}{N} \sum_{n=1}^N (E[X_n^2] - 2 \cdot E[X_n \cdot \bar{X}] + E[\bar{X}^2]) \\ &= \frac{1}{N} \sum_{n=1}^N (\mu^2 + \sigma^2) - \frac{2}{N} \sum_{n=1}^N E \left[ \frac{X_n}{N} \sum_{k=1}^N X_k \right] + E \left[ \left( \frac{1}{N} \sum_{k=1}^N X_k \right)^2 \right] \end{aligned}$$

$$\begin{aligned}
&= \mu^2 + \sigma^2 - \frac{2}{N^2} \sum_{n=1}^N (\mu^2 + \sigma^2 + (N-1)\mu^2) + \\
&\quad + \frac{1}{N^2} (N(\mu^2 + \sigma^2) + (N^2 - N)\mu^2) \\
&= \left(1 - \frac{1}{N}\right) \cdot \sigma^2
\end{aligned}$$

Dabei wurde von folgender Beziehung Gebrauch gemacht, welche aus der paarweisen Unkorreliertheit der  $X_n$  folgt:

$$E[X_k \cdot X_n] = \begin{cases} E[X_n^2] = \mu^2 + \sigma^2 & \text{falls } k = n \\ E[X_n] \cdot E[X_k] = \mu^2 & \text{falls } k \neq n \end{cases}$$

Der Erwartungswert der empirischen Varianz entspricht also *nicht* exakt der Varianz. Daher wird stattdessen mitunter auch die **korrigierte empirische Varianz** verwendet:

$$S_{\text{kor}}^2(N) = \frac{1}{N-1} \sum_{n=1}^N (X_n - \bar{X}(N))^2 = \frac{N}{N-1} S^2(N)$$

Der Faktor „ $N/(N-1)$ “ wird auch als **Bessel-Korrektur** bezeichnet. Die korrigierte Varianz ist im Gegensatz zur unkorrigierten Varianz ein erwartungstreuer Schätzer, d. h., es gilt  $E[S_{\text{kor}}^2(N)] = \sigma^2$ .

Beide Schätzer liefern für große Stichproben jedoch ähnliche Werte und sind mithin **asymptotisch erwartungstreu**:

$$E[S_{\text{kor}}^2(N)] = \lim_{N \rightarrow \infty} E[S^2(N)] = \sigma^2$$

Ebenso wie das arithmetische Mittel und die empirische Wahrscheinlichkeit sind beide Schätzer konsistent:

$$\lim_{N \rightarrow \infty} \Pr(|S^2(N) - \sigma^2| < \varepsilon) = \lim_{N \rightarrow \infty} \Pr(|S_{\text{kor}}^2(N) - \sigma^2| < \varepsilon) = 1$$

**Varianzschätzung.** Seien  $X_1, \dots, X_N$  unabhängige und identisch verteilte Stichprobenvariablen mit endlichem Erwartungswert und endlicher Varianz  $\sigma^2$ .

Die empirische Varianz  $S^2 = \frac{1}{N} \sum_{n=1}^N (X_n - \bar{X})^2$  ist ein asymptotisch erwartungstreuer und konsistenter Schätzer, die korrigierte empirische Varianz  $S_{\text{kor}}^2 = \frac{1}{N-1} \sum_{n=1}^N (X_n - \bar{X})^2$  ein erwartungstreuer und konsistenter Schätzer der Varianz  $\sigma^2$ .

#### 4.2.4 Zentraler Grenzwertsatz von Lindeberg-Lévy

Das Tschebyscheff'sche Gesetz der großen Zahlen besagt, dass unter relativ allgemeinen Voraussetzungen der arithmetische Mittelwert einer hinreichend großen Stichprobe den Erwartungswert der zugrundeliegenden Zufallsgröße sehr wahrscheinlich mit hoher Genauigkeit wiedergibt. Erstaunlicherweise können wir sogar die Verteilung angeben, mit der erhobene Mittelwerte großer Stichproben streuen:

**Zentraler Grenzwertsatz von Lindeberg-Lévy.** Seien  $X_1, \dots, X_N$  unabhängige und identisch verteilte Stichprobenvariablen mit endlichem Erwartungswert  $\mu$  und endlicher Varianz  $\sigma^2 > 0$ .

Dann ist der Mittelwert  $\bar{X}(N) = \frac{1}{N}(X_1 + \dots + X_N)$  großer Stichproben annähernd normalverteilt:

$$\lim_{N \rightarrow \infty} \Pr \left( a \leq \sqrt{N} \cdot \frac{\bar{X}(N) - \mu}{\sigma} \leq b \right) = \int_a^b \mathcal{N}(\xi | 0, 1) d\xi$$

Im Allgemeinen sagt man, eine Folge von Zufallsvariablen  $Z_1, Z_2, \dots$  **konvergiert in Verteilung** gegen die Zufallsvariable  $Z$ , falls die jeweiligen Verteilungsfunktionen punktweise gegen die Verteilung von  $Z$  konvergieren:  $\lim_{N \rightarrow \infty} F_{Z_N}(u) = F_Z(u)$  für alle  $u \in \mathbb{R}$ , in abgekürzter Notation:  $Z_N \xrightarrow{D} Z$ . In dieser Sprechweise sagt der zentrale Grenzwertsatz aus, dass die Folge von Zufallsvariablen

$$Z_N := \sqrt{N} \sigma^{-1} (\bar{X}(N) - \mu)$$

stets in Verteilung gegen eine standardnormalverteilte Zufallsvariable  $Z$  konvergiert:

$$Z_N \xrightarrow{D} Z \text{ mit } Z \sim \mathcal{N}(\cdot | 0, 1)$$

Die Kernaussage ist, dass bei wiederholter Erhebung von hinreichend großen Stichproben die Mittelwerte dieser Stichproben normalverteilt sind. Eine praktische Bedeutung des zentralen Grenzwertsatzes besteht darin, dass er präzise Aussagen über statistische Genauigkeit ermöglicht. Ein häufiges Missverständnis ist, dass dieser Satz die Grundlage dafür darstellen soll, dass in der Praxis viele empirische Verteilungen durch eine Normalverteilung approximiert werden. Dies ist jedoch nicht der Fall.

Der Beweis des Satzes ist letztlich nicht schwer, erfordert aber weitere analytische Hilfsmittel, auf die wir in diesem Buch nicht näher eingehen (für einen Beweis siehe z. B. [3, Satz 17.3.1]). Wir wollen uns lediglich anhand eines numerischen Beispiels von den praktischen Konsequenzen überzeugen. Wir können uns den folgenden Vorgang vorstellen: Wir ziehen gemäß einer festen Wahrscheinlichkeitsdichte  $p(\cdot)$  wiederholt Stichproben  $x^{(1)}, x^{(2)}, \dots$  von hinreichen-der Größe  $N$ , also:

$$\begin{aligned} x^{(1)} &= \left( x_1^{(1)}, \dots, x_N^{(1)} \right), \\ x^{(2)} &= \left( x_1^{(2)}, \dots, x_N^{(2)} \right), \dots \end{aligned}$$

Selbst wenn wir nichts über die Verteilung  $p(\cdot)$  wissen (außer, dass sie endlichen Erwartungswert und endliche Varianz besitzt), so können wir doch davon überzeugt sein, dass die zugehörigen arithmetischen Mittelwerte  $\bar{x}^{(1)}, \bar{x}^{(2)}, \dots$  einer Normalverteilung folgen.

In Abb. 4.3 ist links oben eine beispielhafte Dichtefunktion  $p(\cdot)$  mit Erwartungswert  $\mu_0 \approx 0,64$  und Standardabweichung  $\sigma_0 \approx 1,80$  dargestellt.

Die folgenden Abbildungen zeigen Histogramme über sehr viele Mittelwerte  $\bar{x}^{(1)}, \bar{x}^{(2)}, \dots$  für verschiedene Stichproben der Größe  $N$ , welche durch numerische Simulation gemäß der Verteilung  $p(\cdot)$  gezogen wurden. Außerdem ist die theoretische Grenzverteilung zum Vergleich abgebildet, also eine Gauß-Verteilung mit Lageparameter  $\mu = \mu_0$  und Streuungsparameter  $\sigma = \frac{\sigma_0}{\sqrt{N}}$ .

Im Falle  $N = 1$  können wir nicht erwarten, dass der Grenzwertsatz einschlägig wird, denn dann reproduzieren wir lediglich die ursprüngliche Verteilung. Doch bereits für  $N = 5$  zeigt die Verteilung der arithmetischen Mittelwerte  $\bar{x}^{(k)}$  die typische Glockenform der Gauß-Verteilung. Für  $N = 20$  ist diese deutlich schmäler; dieses Verhalten demonstriert nochmals das Tschebyscheff'sche Gesetz der großen Zahlen: Die arithmetischen Mittelwerte streuen für große Stichproben nur noch in geringem Maße um den tatsächlichen Erwartungswert herum.

## 4.3 Statistische Schätz- und Testverfahren

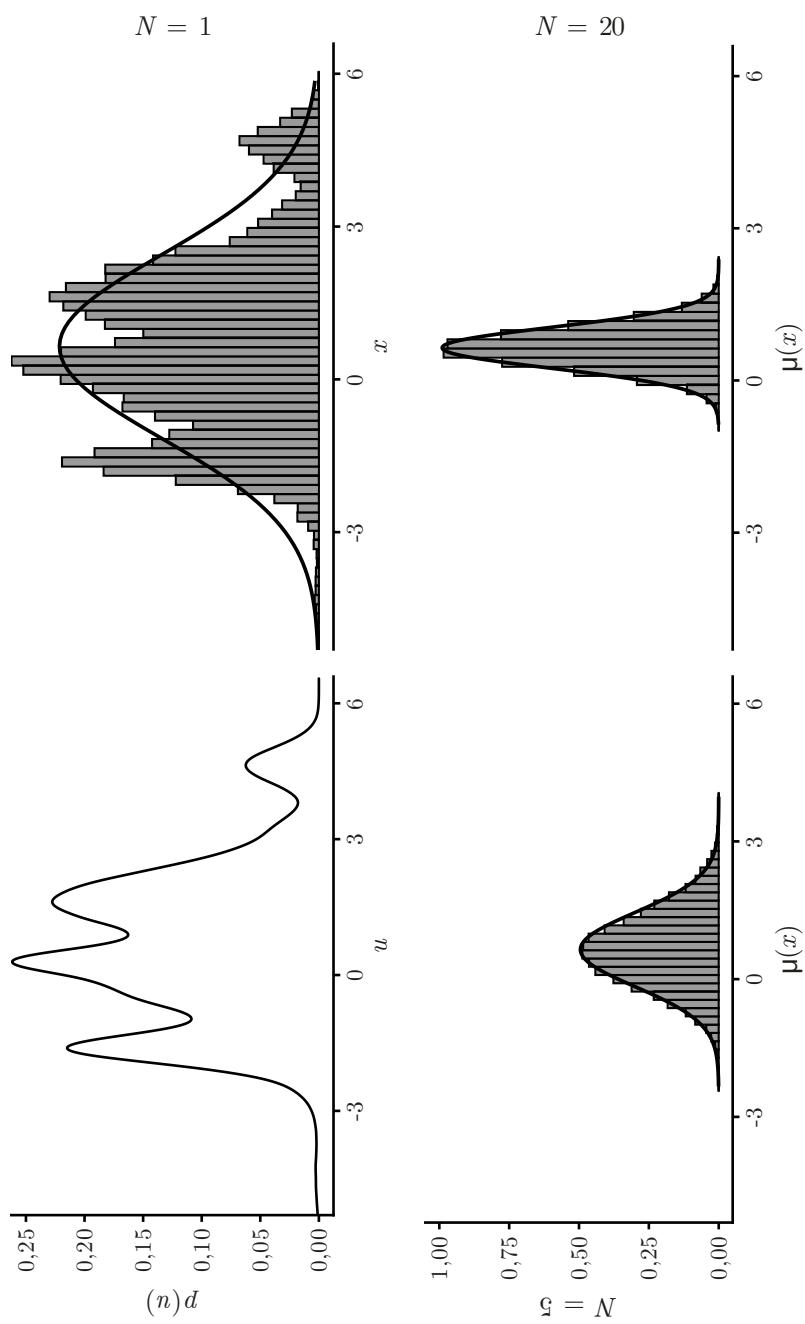
Wir sahen bei der Herleitung der Gesetze der großen Zahlen: Fassen wir die Erhebung einer Stichprobe als wiederholtes Zufallsexperiment auf, so können wir Angaben über die Streuung von geschätzten Parametern wie dem arithmetischen Mittelwert bei wiederholter Erhebung treffen. Mithin können wir untersuchen:

- Welchen Genauigkeitsangaben kann bei vorgegebener Irrtumswahrscheinlichkeit noch Vertrauen geschenkt werden?
- Welche Abweichungen einer Parameterschätzung von einem vorgegebenen Wert können als statistisch signifikant erachtet werden?

### 4.3.1 Intervallschätzung

Der zentrale Grenzwertsatz ist von großer Bedeutung, da er uns anstelle einer bloßen **Punktschätzung** erlaubt, ein **Vertrauensintervall** anzugeben, auch **Konfidenzintervall** genannt. Das bedeutet, anstelle eines einzelnen Schätzwertes geben wir ein Intervall an, von dem wir uns hinreichend sicher sein können, dass es den wahren Parameter enthält.

**Abb. 4.3.** Zentraler Grenzwertsatz von Lindeberg-Lévy: Verteilung der arithmetischen Mittelwerte mit wachsendem Stichprobenumfang



Angenommen, wir wollen annehmen dürfen, dass unser Schätzer des Mittelwerts für hinreichend große Stichproben mit 95-prozentiger Sicherheit korrekt ist. Wir fordern also für das Vertrauensintervall  $[\bar{x}]_\gamma = [\bar{x}_{\min}, \bar{x}_{\max}]$  mit vorher fest gewähltem **Vertrauensniveau**  $\gamma = 0,95$ :

$$\lim_{N \rightarrow \infty} \Pr(\bar{X}(N) \in [\bar{x}_{\min}, \bar{x}_{\max}]) = \gamma$$

Wir machen folgenden Ansatz mit der uns zunächst noch unbekannten Zahl  $z \in \mathbb{R}$ ,  $z > 0$ :

$$\gamma = \lim_{N \rightarrow \infty} \Pr\left(\bar{X}(N) \in \left[\mu - z \cdot \frac{\sigma}{\sqrt{N}}, \mu + z \cdot \frac{\sigma}{\sqrt{N}}\right]\right)$$

Aus dem zentralen Grenzwertsatz ergibt sich:

$$\begin{aligned}\gamma &= \lim_{N \rightarrow \infty} \Pr\left(-z \leq \sqrt{N} \cdot \frac{\bar{X}(N) - \mu}{\sigma} \leq z\right) \\ &= \int_{-z}^z \mathcal{N}(\xi | 0, 1) d\xi\end{aligned}$$

Damit bestimmt sich  $z = z(\gamma)$  gerade über die Integralgrenzen, zwischen denen sich unter der Standardnormalverteilung eine Fläche von  $\gamma$  ergibt. Mithin gilt

$$z(\gamma) = \Phi^{-1}\left(\frac{1 + \gamma}{2}\right),$$

wobei  $\Phi(\cdot)$  die Verteilungsfunktion der Standardnormalverteilung ist:

$$\Phi(u) = \int_{-\infty}^u \mathcal{N}(\xi | 0, 1) d\xi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{\xi^2}{2}} d\xi$$

für alle  $u \in \mathbb{R}$ . Werte von  $\Phi(u)$  können numerisch oder aus Tabellen bestimmt werden, es gilt insbesondere  $z(0,95) = \Phi^{-1}(0,975) \approx 1,96$ .

Ein paar Werte für das Vertrauensniveau sind im Folgenden tabellarisch aufgeführt:

$\gamma$	90,0 %	95,0 %	99,0 %	99,9 %
$z(\gamma)$	1,64	1,96	2,58	3,29
$z^*(\gamma)$	1,28	1,64	2,33	3,09

**Tabelle 4.1.** Kritische Werte als Funktion des Vertrauensniveaus

Der **Intervallschätzer des Erwartungswerts** zum Vertrauensniveau  $\gamma$  auf Grundlage einer hinreichend großen Stichprobe  $x = (x_1, \dots, x_N)$  ist wie folgt gegeben:

$$[\bar{x}]_\gamma = \left[ \bar{x} - z(\gamma) \cdot \frac{s(x)}{\sqrt{N}}, \bar{x} + z(\gamma) \cdot \frac{s(x)}{\sqrt{N}} \right]$$

Dabei haben wir den Mittelwert  $\mu$  und die Standardabweichung  $\sigma$  durch ihre empirischen Schätzer ersetzt, was für hinreichend große Stichproben eine sinnvolle Näherung darstellt. Dies kann durch den sogenannten **Satz von Slutsky** genauer begründet werden: Falls  $A_N \xrightarrow{D} a$  und  $B_N \xrightarrow{D} b$  mit Konstanten  $a, b \in \mathbb{R}$ , so gilt  $A_N \cdot Z_N + B_N \xrightarrow{D} aZ + b$ , falls  $Z_N \xrightarrow{D} Z$ . Wir können also (für hinreichend große Stichproben) konstante Summanden und Faktoren wie  $\mu$  und  $\sigma$  durch ihre konsistenten Schätzer ersetzen. Verbreitete Faustregeln für „hinreichend großen“ Stichprobenumfang sind  $N > 30$  oder  $N > 50$ .

Anstelle des Vertrauensniveaus kann auch das **Signifikanzniveau** bzw. die **Irrtumswahrscheinlichkeit**  $\alpha = 1 - \gamma$  angegeben werden.

**Anwendungsbeispiel.** Das 99,9 %-Vertrauensintervall für den empirischen Erwartungswert der durch die CDC-Studie erhobenen Körpergröße männlicher Befragter ist wie folgt gegeben:

$$[\mu(\text{Körpergröße})]_{0,999} = [177,98 \text{ cm}; 178,10 \text{ cm}]$$

Diese hohe statistische Genauigkeit ist insbesondere auf die Größe der Stichprobe  $N$  zurückzuführen, es wurden mehr als 190.000 männliche Personen befragt.

Wir wollen demonstrieren, wie es sich mit Schätzungen von geringerem Stichprobenumfang verhält. Zu diesem Zweck erheben wir hundert Mal zufällig eine Reihe von  $N = 50$  Werten für die Körpergröße und berechnen das zugehörige 95 %-Vertrauensintervall. Das Ergebnis ist in Abb. 4.4 oben zu sehen: Die meisten der Vertrauensintervalle, dargestellt als senkrechte Fehlerbalken, enthalten auch den wahren Wert von 178 cm, dargestellt als horizontale gestrichelte Linie. Fünf der Intervalle enthalten ihn jedoch nicht – das ist von der Konstruktion her erwartbar und entspricht der vorgegebenen Irrtumswahrscheinlichkeit von  $\alpha = 5\%$ : Insbesondere auf niedrigem Vertrauensniveau besteht immer die Möglichkeit, dass die Intervallschätzung den wahren Erwartungswert bzw. den wahren Mittelwert der Grundgesamtheit nicht trifft.

Die im vorigen Beispiel untersuchte Stichprobe der Körpergröße männlicher Befragter weist eine empirische Standardabweichung von  $s(x) = 7,85 \text{ cm}$  auf. Ohne die Form der Verteilung genauer zu kennen, können wir allein aus der Tschebyscheff'schen Ungleichung folgern, dass wenigstens  $\delta = 95\%$  der Datenpunkte in folgendem Intervall um den Erwartungswert liegen:

$$\left[ \bar{x} - \frac{1}{\sqrt{1-\delta}} \cdot s(x), \bar{x} + \frac{1}{\sqrt{1-\delta}} \cdot s(x) \right] = \left[ 178 \text{ cm} \pm \frac{1}{\sqrt{0,05}} \cdot 7,85 \text{ cm} \right] \\ = [143 \text{ cm}, 213 \text{ cm}]$$

Eine genauere Schätzung kann durch die Ermittlung von Quantilen erhalten werden:

$$[Q_{0,025}(x), Q_{0,975}(x)] = [163 \text{ cm}, 193 \text{ cm}]$$

Intervalle dieser Art werden **Vorhersage-** oder **Prognoseintervalle** genannt. Die untersuchte Größe mag mehr oder weniger breit streuen, die meisten Werte werden aber in dem Intervall angenommen. Prognoseintervalle sollten nicht mit den weiter oben beschriebenen Vertrauensintervallen verwechselt werden. Letztere stellen einen Bereich dar, in dem mit hoher Wahrscheinlichkeit eine Kennzahl der Verteilung wie etwa der Erwartungswert oder ein anderer Modellparameter vermutet wird.

**Anwendungsbeispiel.** Das 95 %-Vertrauensintervall für das mittlere monatliche Nettohaushaltseinkommen gemäß der ALLBUS-Studie [4] ist wie folgt gegeben:

$$[\mu(\text{Einkommen})]_{0,95} = [3066 \text{ EUR}, 3236 \text{ EUR}]$$

In diesem Fall sind die statistischen Ungenauigkeiten in der Schätzung nicht mehr vernachlässigbar: Insgesamt machten nur  $N = 2530$  Personen eine Angabe zu ihrem Einkommen bei einer hohen empirischen Standardabweichung von 2179 EUR.

### 4.3.2 Gauß-Test

Eine weitere wichtige Anwendung des zentralen Grenzwertsatzes sind Hypothesentests. Wir betrachten dazu zunächst die folgende Fragestellung: Unter welchen Umständen können wir auf Grundlage statistischer Hinweise hinreichend sicher sein, dass der Erwartungswert  $E[X]$  einer Zufallsgröße  $X$  verschwindet? In vielen Anwendungsfällen stellt dies die sogenannte **Nullhypothese** dar, welche wir zugunsten einer **Arbeitshypothese** oder **Alternativhypothese** widerlegen wollen. Auf einem Vertrauensniveau von  $\gamma$  gilt die Nullhypothese „ $E[X] = 0$ “ als widerlegt, wenn das arithmetische Mittel  $\bar{x}$  der Stichprobe in folgendem Intervall *nicht* enthalten ist:

$$\left[ -z(\gamma) \cdot \frac{s(x)}{\sqrt{N}}, z(\gamma) \cdot \frac{s(x)}{\sqrt{N}} \right]$$

Hierbei gehen wir wieder davon aus, dass die Stichprobe  $x = (x_1, \dots, x_N)$  hinreichend groß ist und eine Folge von Realisierungen von unabhängigen Zufallsvariablen mit derselben Verteilung wie  $X$  darstellt. Ist die Nullhypothese dadurch gegeben, dass der Erwartungswert einen bestimmten Wert  $\mu$  annimmt, der nicht notwendigerweise null ist, so ist das obige Intervall wie folgt anzupassen:

$$\left[ \mu - z(\gamma) \cdot \frac{s(x)}{\sqrt{N}}, \mu + z(\gamma) \cdot \frac{s(x)}{\sqrt{N}} \right]$$

Wir betrachten das zu Beginn des Kapitels genannte Beispiel der möglicherweise unfairen Münze. Das arithmetische Mittel des binären Merkmals mit den

Ausprägungen  $x_n = 1$  für „Kopf“ und  $x_n = 0$  für „Zahl“ ist gerade die relative Häufigkeit  $f$  des Vorkommens von „Kopf“:

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n = f$$

Die Erfolgswahrscheinlichkeit  $p$  ist durch den Erwartungswert der zugrundeliegenden Bernoulli-verteilten Zufallsvariablen geben:  $E[X] = p \approx f$ .

Die Nullhypothese entspricht der Behauptung, die Münze sei fair, also „ $E[X] = 1/2$ “. Wir legen ein Vertrauensniveau von  $\gamma = 0,95$  fest und beobachten nach zehn Würfen, dass die Münze sieben Mal mit „Kopf“ zuoberst zum Liegen kommt. Mit diesen konkreten Werten wird das Vertrauensintervall zu:

$$\left[ 0,5 - 1,96 \cdot \frac{0,42}{\sqrt{10}}; 0,5 + 1,96 \cdot \frac{0,42}{\sqrt{10}} \right] = [0,24; 0,76]$$

Der beobachtete Anteil von 0,7 liegt noch innerhalb der Intervallgrenzen: Die Nullhypothese einer fairen Münze kann auf dem gegebenen Vertrauensniveau daher nicht abgelehnt werden. Die Stichprobengröße ist relativ klein, daher empfiehlt sich eigentlich die Anwendung eines **Student'schen t-Tests**, auf den wir im nächsten Abschnitt eingehen. Dieser Test korrigierte den Wert  $z(0,95) \approx 1,96$  auf  $z_9(0,95) \approx 2,26$  – das Prinzip bleibt jedoch das gleiche.

Beobachteten wir hingegen siebzig von hundert Münzwürfen mit dem Ausgang von „Kopf“, so ergäbe sich das folgende Vertrauensintervall:

$$\left[ 0,5 - 1,96 \cdot \frac{0,46}{\sqrt{100}}; 0,5 + 1,96 \cdot \frac{0,46}{\sqrt{100}} \right] = [0,41; 0,59]$$

In diesem Fall sollte die Nullhypothese abgelehnt und – auf dem gegebenen Vertrauensniveau – die Münze als unfair erkannt werden.

Das oben skizzierte Verfahren entspricht dem sogenannten **Einstichproben-Gauß-Test**. Oft stellt sich eher die Frage, ob die Differenz der Erwartungswerte zweier Zufallsgrößen verschwindet, oder ob ein Erwartungswert kleiner oder größer als der andere ist. Hierfür werden die Mittelwerte zweier Stichproben miteinander verglichen. Die Differenz von zwei normalverteilten und unabhängigen Zufallsvariablen  $X \sim \mathcal{N}(\cdot | \mu_X, \sigma_X^2)$  und  $Y \sim \mathcal{N}(\cdot | \mu_Y, \sigma_Y^2)$  ist wiederum normalverteilt:

$$Y - X \sim \mathcal{N}(\cdot | \mu_Y - \mu_X, \sigma_X^2 + \sigma_Y^2)$$

Angewandt auf die Grenzverteilung der arithmetischen Mittelwerte ergeben sich letztlich die folgenden Testkriterien.

**Zweistichproben-Gauß-Test.** Gegeben seien mit  $x = (x_1, \dots, x_{N_x})$  und  $y = (y_1, \dots, y_{N_y})$  zwei hinreichend große Stichprobenvektoren, die wir als Realisierungen von jeweils identisch verteilten StichprobenvARIABLEN  $X_1, \dots, X_{N_x}$  bzw.  $Y_1, \dots, Y_{N_y}$  mit jeweils endlichem Erwartungswert und

endlicher Varianz ansehen dürfen, wobei  $X_1, \dots, X_{N_x}, Y_1, \dots, Y_{N_y}$  sämtlich unabhängig sind.

**Zweiseitiger Gauß-Test.** Auf einem Vertrauensniveau von  $0 < \gamma < 1$  ist die Nullhypothese  $E[X] = E[Y]$  abzulehnen, falls gilt:

$$|\bar{y} - \bar{x}| > z(\gamma) \cdot \sqrt{\frac{s^2(x)}{N_x} + \frac{s^2(y)}{N_y}}$$

mit dem **kritischen Wert**  $z(\gamma) = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$ .

**Einseitiger Gauß-Test.** Auf einem Vertrauensniveau von  $0 < \gamma < 1$  ist die Nullhypothese  $E[X] \leq E[Y]$  abzulehnen, falls gilt:

$$\bar{y} - \bar{x} > z^*(\gamma) \cdot \sqrt{\frac{s^2(x)}{N_x} + \frac{s^2(y)}{N_y}}$$

mit  $z^*(\gamma) = \Phi^{-1}(\gamma)$ . Für  $\gamma = 0,95$  gilt z. B.  $z^*(\gamma) = 1,645$ .

**Anwendungsbeispiel.** Das mittlere monatliche Nettoeinkommen der Einpersonenhaushalte in Deutschland beziffert sich gemäß der ALLBUS-Studie auf 1668 EUR. Es gibt jedoch einen deutlichen Unterschied zwischen weiblichen und männlichen Befragten von 1535 EUR gegenüber 1788 EUR.

Wir vermuten daher, dass Frauen in Deutschland über ein geringeres Nettoeinkommen verfügen als Männer und möchten nachweisen, dass die Datenlage mit dieser Hypothese konsistent ist. Insgesamt machten  $N_y = 269$  Männer und  $N_x = 267$  Frauen eine Angabe. Die jeweiligen Standardabweichungen sind  $s(y) = 1083$  EUR und  $s(x) = 827$  EUR. Der einseitige Gauß-Test auf dem Vertrauensniveau 95 % fordert eine Ablehnung der Nullhypothese („Frauen verdienen mindestens soviel wie Männer“), wenn die beobachtete Einkommensdifferenz größer als der folgende Wert ist:

$$1,645 \cdot \sqrt{\frac{s^2(x)}{N_x} + \frac{s^2(y)}{N_y}} = 1,645 \cdot \sqrt{\frac{(827 \text{ EUR})^2}{169} + \frac{(1083 \text{ EUR})^2}{267}} \\ \approx 137 \text{ EUR}$$

Der Wert liegt unter der beobachteten Einkommensdifferenz von 253 EUR. Die Datenlage bietet somit hinreichenden Anlass, um die Nullhypothese abzulehnen.

### 4.3.3 Student'sche Vertrauensintervalle

Bislang sind wir für die Ermittlung von Vertrauensintervallen stets von einer „hinreichend großen“ Stichprobe ausgegangen. Wir wollen hier kurz ein

Verfahren vorstellen, dessen Anwendung sich (auch) bei kleineren Stichproben empfiehlt.

Wesentlicher Ausgangspunkt war stets – bei bekannter Standardabweichung  $\sigma$  – der zentrale Grenzwertsatz:

$$\lim_{N \rightarrow \infty} \Pr \left( -z \leq \sqrt{N} \cdot \frac{\bar{X}(N) - \mu}{\sigma} \leq z \right) = \int_{-z}^z \mathcal{N}(\xi | 0, 1) d\xi$$

Nun ist die Standardabweichung in der Regel unbekannt und muss genauso wie der Mittelwert aus den Daten geschätzt werden. Glücklicherweise gilt auch dann, wenn wir die wahre Standardabweichung durch ihre Schätzfunktion ersetzen:

$$\lim_{N \rightarrow \infty} \Pr \left( -z \leq \sqrt{N} \cdot \frac{\bar{X}(N) - \mu}{S_{\text{kor}}(N)} \leq z \right) = \int_{-z}^z \mathcal{N}(\xi | 0, 1) d\xi$$

Sind  $\bar{X}(N)$  und  $S_{\text{kor}}(N)$  voneinander abhängige Zufallsvariablen, so besteht wenig Hoffnung, für kleine Stichproben genauere Aussagen über die Verteilung der einzugrenzenden Testgröße zu treffen. Wenn wir andererseits davon ausgehen, dass sie unabhängig sind, dann kennen wir deren Verteilung sogar exakt.

**Bedingungen an eine  $t$ -verteilte Testgröße.** Seien  $X_1, \dots, X_N$ ,  $N > 1$ , unabhängige und identisch verteilte Stichprobenvariablen mit endlichem Erwartungswert  $\mu$  und endlicher Varianz  $\sigma^2$ .

Falls die Schätzer für den Erwartungswert  $\bar{X}(N)$  und die Varianz  $S_{\text{kor}}^2(N)$  unabhängige Zufallsvariablen darstellen, dann folgt die Testgröße

$$T(N) = \sqrt{N} \cdot \frac{\bar{X}(N) - \mu}{S_{\text{kor}}(N)}$$

einer Student'schen  $t$ -Verteilung mit  $N - 1$  Freiheitsgraden.

Wir skizzieren kurz einen Beweisweg. Der sogenannte **Satz von Greary** bzw. dessen Verallgemeinerung durch Lukacs [5, 6] sagt aus, dass die Schätzer für Erwartungswert und Varianz dann und nur dann unabhängige Zufallsvariablen sind, wenn die Stichprobenvariablen  $X_1, \dots, X_N$  einer Normalverteilung folgen. Die Bedingung ist also – vielleicht überraschenderweise – recht restriktiv und schränkt die mögliche Verteilung der untersuchten Daten stark ein.

Die Stichprobenvariable der Testgröße kann wie folgt geschrieben werden:

$$T(N) = \sqrt{N-1} \cdot \frac{Z}{\sqrt{Y_{N-1}}}$$

mit

$$Z = \sqrt{N} \cdot \frac{\bar{X}(N) - \mu}{\sigma} \quad \text{und} \quad Y_{N-1} = \frac{(N-1) \cdot S_{\text{kor}}^2(N)}{\sigma^2}.$$

Da die Stichprobenvariablen exakt einer Normalverteilung folgen, ist  $Z$  standardnormalverteilt, das hatten wir bereits in Abschn. 4.1.2 festgestellt.

Wir definieren die folgende neue Zufallsvariablen:

$$\tilde{X}_n := \frac{1}{\sqrt{n \cdot (n+1)\sigma}} \cdot \left( -n \cdot X_{n+1} + \sum_{k=1}^n X_k \right)$$

für alle  $n \in \{1, \dots, N-1\}$ . Es ist nicht schwer zu sehen, dass  $\tilde{X}_1, \dots, \tilde{X}_N$  standardnormalverteilt sind. Außerdem kann deren Unabhängigkeit nachgewiesen werden, und es gilt:

$$Y_{N-1} = \sum_{n=1}^{N-1} (\tilde{X}_n)^2$$

Daher ist  $Y_{N-1}$  gemäß einer Chi-Quadrat-Verteilung mit  $N-1$  Freiheitsgraden verteilt, und somit  $T(N)$  gemäß einer  $t$ -Verteilung mit  $N-1$  Freiheitsgraden.

Bezeichnen wir mit  $\Phi_{N-1}(\cdot)$  die entsprechende Verteilungsfunktion:

$$\Phi_{N-1}(u) = \int_{-\infty}^u t_{N-1}(\xi) d\xi$$

für  $u \in \mathbb{R}$ . Die in vorigen Abschnitten konstruierten Vertrauensintervalle für Intervallschätzungen und Hypothesentests können angepasst werden, indem anstelle der Normalverteilung eine  $t$ -Verteilung verwendet wird. Auf gegebenem Vertrauensniveau  $\gamma$  sind die kritischen Werte bei einem Stichprobenumfang  $N$  dann wie folgt gegeben:

$$z_{N-1}(\gamma) = \Phi_{N-1}^{-1}\left(\frac{1+\gamma}{2}\right) \text{ bzw. } z_{N-1}^*(\gamma) = \Phi_{N-1}^{-1}(\gamma)$$

Für hinreichend große Stichproben führen diese kritischen Werte in der Praxis auf dieselben Vertrauensintervalle wie die beim Gauß-Test verwendeten, da die  $t$ -Verteilung im Grenzwert vieler Freiheitsgrade mit der Normalverteilung identisch ist:

$$\lim_{N \rightarrow \infty} z_{N-1}(\gamma) = z(\gamma), \quad \lim_{N \rightarrow \infty} z_{N-1}^*(\gamma) = z^*(\gamma)$$

Auf einem 95 %-Vertrauensniveau ergeben sich zum Beispiel die folgenden kritischen Werte als Funktion des Stichprobenumfangs:

$N$	5	15	30	50	100	1000	$\infty$
$z_{N-1}(0,95)$	2,78	2,14	2,05	2,01	1,98	1,96	1,96
$z_{N-1}^*(0,95)$	2,13	1,76	1,70	1,68	1,66	1,65	1,64

**Tabelle 4.2.** Kritische Werte als Funktion des Stichprobenumfangs

**Anwendungsbeispiel.** Die Körpergröße männlicher Personen der CDC-Studie kann annähernd als normalverteilt angenommen werden, vgl. hierzu Abb. 4.6. Der große Stichprobenumfang von mehr als 190.000 Personen ermöglicht eine sehr genaue Schätzung der durchschnittlichen Körpergröße von 178 cm. Wir wollen aufzeigen, inwieweit dieser Wert vermöge der Student'schen Testgröße auch mittels einer deutlich kleineren Stichprobe geschätzt werden könnte. Hierzu erheben wir hundert Mal zufällig eine Reihe von fünf Werten  $x_1, \dots, x_5$  für die Körpergröße und berechnen das zugehörige 95 %-Vertrauensintervall:

$$\left[ \bar{x} \pm z_{\alpha/2}(0,95) \cdot \frac{s(x)}{\sqrt{5}} \right] \approx [\bar{x} \pm 1,24 \cdot s(x)]$$

Die meisten der Vertrauensintervalle enthalten auch den wahren Wert von 178 cm, siehe Abb. 4.4 unten. Wir sehen auch deutlich, warum wir die Größe des Vertrauensintervalls gegenüber dem Gauß-Test anpassen müssen: Die aus den sehr kleinen Stichproben geschätzte Standardabweichung ist ebenfalls starken Schwankungen unterworfen. Vier der hundert Intervallschätzungen enthalten den wahren Wert nicht, dies spiegelt die gewählte Irrtumswahrscheinlichkeit von  $\alpha = 5\%$  wieder.

#### 4.3.4 Effektstärke

Insbesondere für sehr große Stichproben kann es vorkommen, dass sich regelmäßig statistisch signifikante Effekte nachweisen lassen. Dies sagt jedoch noch nichts über die **Effektstärke** aus.

**Anwendungsbeispiel.** Anhand des CDC-Datensatzes können Statistiken für bestimmte US-Bundesstaaten erstellt werden. Vergleichen wir die durchschnittliche Körpergröße männlicher Befragter in Rhode Island  $\bar{x}$  mit jenen in New York  $\bar{y}$ , so ergibt sich folgende Teststatistik für den zweiseitigen Gauß-Test auf einem Vertrauensniveau von 95 %:

$$1,96 \cdot \sqrt{\frac{s^2(x)}{N_x} + \frac{s^2(y)}{N_y}} = 1,96 \cdot \sqrt{\frac{(7,68 \text{ cm})^2}{2391} + \frac{(8,23 \text{ cm})^2}{15843}} \\ \approx 0,33 \text{ cm}$$

Dieser Wert liegt unterhalb der beobachteten Differenz von  $|\bar{y} - \bar{x}| = 0,44 \text{ cm}$ . Die Differenz ist also statistisch signifikant – dennoch ist sie sehr klein und daher von geringer Bedeutung.

In vielen Fällen genügt es, die Differenz zweier Mittelwerte in natürlichen Einheiten anzugeben; im obigen Beispiel wurde sie in metrischen Einheiten der Länge gemessen. Eine weitere Möglichkeit besteht in einer Angabe in Einheiten, die einem Vielfachen der Standardabweichung entsprechen.

Für zwei Stichproben  $x = (x_1, \dots, x_{N_x})$  und  $y = (y_1, \dots, y_{N_y})$  ist die **Cohen'sche Effektstärke** (auch „Cohens d“ genannt) eine Maßzahl für die praktische Relevanz eines statistischen Effekts und wie folgt definiert:

$$d(y, x) = \frac{\bar{y} - \bar{x}}{s_{\text{pool}}(x, y)}$$

mit der **gepoolten Varianz** [7, S. 67]:

$$s_{\text{pool}}^2(x, y) = \frac{N_x s^2(x) + N_y s^2(y)}{N_x + N_y - 2}$$

**Anwendungsbeispiel.** Die im Beispiel weiter oben beobachtete Differenz von  $|\bar{y} - \bar{x}| = 0,44$  cm entspricht einem Wert für die Cohen'sche Effektstärke von  $d(y, x) = 0,05$ . Eine Berechnung für einen Vergleich der mittleren Körpergröße von Befragten in Puerto Rico  $\bar{z}$  mit jenen in New York ergibt  $d(y, z) = 0,50$ , dies entspricht einer Differenz in metrischen Einheiten von  $\bar{y} - \bar{z} = 4,1$  cm.

Faustregeln für die Interpretation von Werten für die Cohen'sche Effektstärke sind in der folgenden Tabelle festgehalten [8]:

$ d(y, x) $	0,01	0,02	0,5	0,8	1,2	2,0
<b>Effektstärke</b>	sehr klein	klein	mittel	groß	sehr groß	enorm

**Tabelle 4.3.** Effektstärken

## 4.4 Parameter- und Dichteschätzung

Für eine Stichprobe  $x = (x_1, \dots, x_N) \in \mathbb{R}^N$  ist die **empirische Verteilungsfunktion** durch den Anteil der Datenpunkte unterhalb eines bestimmten Werts gegeben:

$$\hat{F}: \mathbb{R} \rightarrow [0, 1], \hat{F}(u) = \frac{1}{N} \cdot |\{m \in \{1, \dots, N\} | x_m \leq u\}|$$

Diese ist eng verwandt mit dem Prozentrang (siehe Abschn. 2.5.2), denn es gilt gerade  $\hat{F}(x_n) = \%-\text{rg}(x_n)$  für alle  $n \in \{1, \dots, N\}$ .

Das Histogramm der Häufigkeitsverteilung, basierend auf einer Einteilung der reellen Zahlen in Intervalle  $]u_k, u_{k+1}] \subset \mathbb{R}$ ,  $k \in \mathbb{Z}$ , kann wie folgt mittels der empirischen Verteilungsfunktion ausgedrückt werden:

$$\hat{p}: \mathbb{R} \rightarrow [0, \infty[, \hat{p}(u) = \frac{\hat{F}(u_{k+1}) - \hat{F}(u_k)}{u_{k+1} - u_k} \text{ für alle } u \in ]u_k, u_{k+1}]$$

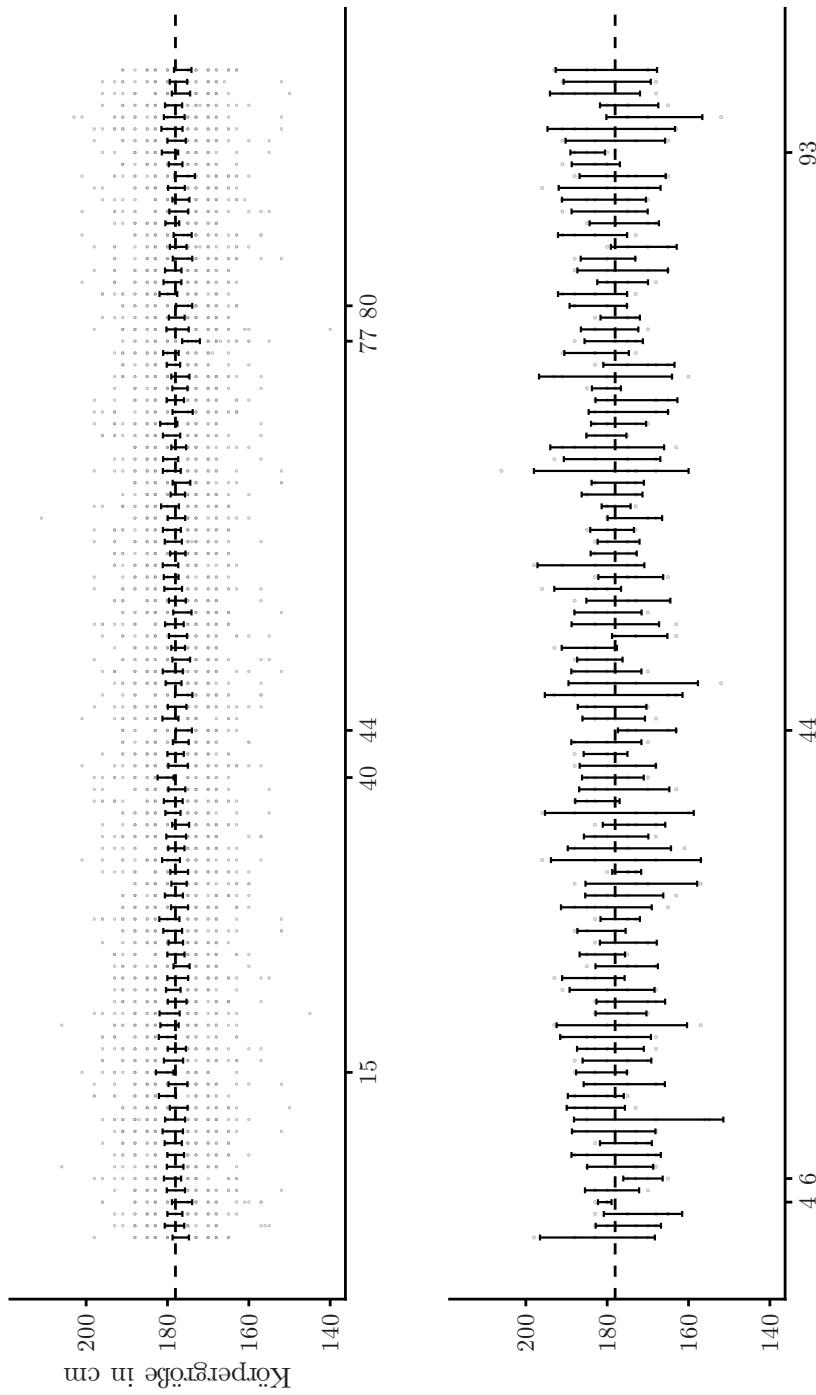


Abb. 4.4. 95 %-Vertrauensintervalle bei Stichprobenumfang  $N = 50$  (oben) und  $N = 5$  (unten)

Der **Satz von Gliwenko-Cantelli** ist ein weiteres Gesetz der großen Zahlen, auf das wir in diesem Band nicht näher eingehen (siehe [3, Satz 15.3.3] für Details): Es garantiert, dass die empirische Verteilungsfunktion in geeignetem Sinn gegen die wahre Verteilungsfunktion konvergiert. Das Histogramm wiederum stellt unter gewissen Bedingungen, die an die Klassenbreite zu stellen sind, und für stetige Zufallsvariablen einen konsistenten Schätzer der Dichtefunktion dar.

Wir können das Histogramm daher als eine empirische Dichtefunktion auffassen. Im Folgenden werden wir weitere Methoden der Dichteschätzung vorstellen. Diese basieren zum Teil auf der Annahme, dass den Beobachtungen ein bestimmtes statistisches Modell zugrundeliegt; die Aufgabe der Dichteschätzung besteht dann darin, die optimalen Parameter anhand der Daten zu bestimmen.

#### 4.4.1 Maximum-Likelihood-Schätzung

Angenommen, wir haben gute Gründe anzunehmen, dass die Häufigkeitsverteilung eines Merkmals in einer Stichprobe einem bestimmten parametrischen statistischen Modell folgt. Wir wollen die Parameter des Modells so bestimmen, dass das Modell die Daten möglichst präzise beschreibt.

Hierzu betrachten wir für einen Stichprobenvektor  $x = (x_1, \dots, x_N)$  und ein statistisches Modell  $p(\cdot | \theta_1, \dots, \theta_K)$  die **Likelihood-Funktion**:

$$\begin{aligned} L(\theta_1, \dots, \theta_K | x_1, \dots, x_N) &= p(x_1 | \theta_1, \dots, \theta_K) \cdots p(x_N | \theta_1, \dots, \theta_K) \\ &= \prod_{n=1}^N p(x_n | \theta_1, \dots, \theta_K) \end{aligned}$$

Wie so oft gehen wir auch hier wieder davon aus, dass jede Merkmalsausprägung eine Realisierung einer Zufallsvariablen ist, und alle diese Variablen haben die gleiche Verteilung und sind unabhängig voneinander. Dann ist die Likelihood-Funktion gerade die gemeinsame Wahrscheinlichkeitsfunktion, ausgewertet an der durch die Stichprobe realisierten Stelle.

Eine sinnvolles Kriterium für die Auswahl optimaler Parameter  $\hat{\theta}_1, \dots, \hat{\theta}_K$  ist eine Maximierung der Likelihood-Funktion: Das sind dann jene Parameter, welche die Beobachtungen unter der Modellannahme am wahrscheinlichsten machen.

Für viele gebräuchliche Modelle vereinfacht sich die Rechnung, wenn statt der Likelihood-Funktion die **Log-Likelihood-Funktion** verwendet wird, welche dieselben Maximalstellen aufweist:

$$\begin{aligned}
\ell(\theta_1, \dots, \theta_K | x_1, \dots, x_N) &= \ln(L(\theta_1, \dots, \theta_K | x_1, \dots, x_N)) \\
&= \ln\left(\prod_{n=1}^N p(x_n | \theta_1, \dots, \theta_K)\right) \\
&= \sum_{n=1}^N \ln(p(x_n | \theta_1, \dots, \theta_K))
\end{aligned}$$

Um die Notation nicht unnötig zu überfrachten, unterschlagen wir im Folgenden in der Regel die Abhängigkeit von den Stichprobenwerten, die wir ohnehin als vorgegeben annehmen:  $\ell(\theta_1, \dots, \theta_K) = \ell(\theta_1, \dots, \theta_K | x_1, \dots, x_N)$ .

Zusammengefasst ergibt sich das folgende Verfahren.

Seien  $p(\cdot | \theta_1, \dots, \theta_K)$  ein statistisches Modell und  $x = (x_1, \dots, x_N)$  eine Stichprobe. Die **Maximum-Likelihood-Schätzung**  $\hat{\theta}_1, \dots, \hat{\theta}_K$  der Parameter ist durch die Maximalstelle der Log-Likelihood-Funktion

$$\ell(\theta_1, \dots, \theta_K) = \sum_{n=1}^N \ln(p(x_n | \theta_1, \dots, \theta_K))$$

gegeben.

Die geschätzte Dichte ist dann die Funktion  $\hat{p}(\cdot) = p(\cdot | \hat{\theta}_1, \dots, \hat{\theta}_K)$ .

Für verschwindende Likelihood, also  $p(x_n | \theta_1, \dots, \theta_K) = 0$  für wenigstens ein  $n \in \{1, \dots, N\}$ , können wir formal  $\ell(\theta_1, \dots, \theta_K) = -\infty$  festlegen.

Wir demonstrieren das Verfahren zunächst am Beispiel der Pareto-Verteilung; wir wollen deren Parameter aus den Daten schätzen. Wir können voraussetzen, dass in der Stichprobe alle Werte größer als der Parameter  $x_{\min}$  sind, da die Verteilungsfunktion für kleinere Werte verschwindet. Die Log-Likelihood-Funktion ist dann wie folgt gegeben:

$$\begin{aligned}
\ell(x_{\min}, \alpha) &= \sum_{n=1}^N \ln(\text{Par}(x_n | x_{\min}, \alpha)) \\
&= \sum_{n=1}^N \ln\left(\frac{\alpha x_{\min}^\alpha}{x_n^{\alpha+1}}\right) \\
&= N \ln \alpha + N \alpha \ln x_{\min} - (\alpha + 1) \sum_{n=1}^N \ln x_n
\end{aligned}$$

Die Likelihood-Funktion wächst mit  $x_{\min}$ . Daher wollen wir den Parameter so groß wählen, dass er gerade noch mit der Stichprobe vereinbart werden kann:

$$\hat{x}_{\min} = \min_{n \in \{1, \dots, N\}} x_n$$

Die partielle Ableitung nach dem zweiten Parameter  $\alpha$  berechnet sich zu:

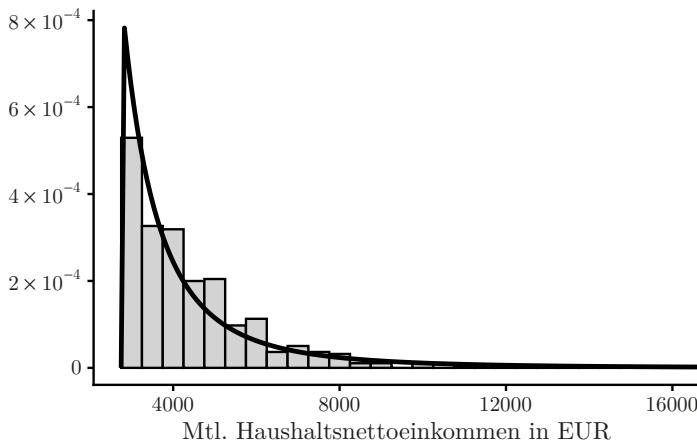
$$\frac{\partial \ell}{\partial \alpha}(\hat{x}_{\min}, \alpha) = \frac{N}{\alpha} + N \ln \hat{x}_{\min} - \sum_{n=1}^N \ln x_n$$

Die Ableitung ist monoton fallend mit  $\alpha$ , daher befindet sich das Maximum  $(\hat{x}_{\min}, \hat{\alpha})$  der Log-Likelihood-Funktion an dessen Nullstelle:

$$\hat{\alpha} = \frac{N}{\sum_{n=1}^N \ln x_n - N \ln \hat{x}_{\min}} = \frac{1}{\mu \left( \ln \frac{x}{\hat{x}_{\min}} \right)}$$

Dabei ist mit  $\mu(\cdot)$  der arithmetische Mittelwert gemeint.

**Anwendungsbeispiel.** Wir glauben, dass alle höheren Einkommen (mehr als 2700 EUR) in der ALLBUS-Studie durch eine Pareto-Verteilung modelliert werden können. Setzen wir die Stichprobenwerte in obige Formel ein, erhalten wir  $x_{\min} = 2750$  EUR,  $\alpha = 2,36$ . In folgender Abbildung ist ein Histogramm der geschätzten Pareto-Dichte gegenübergestellt:



**Abb. 4.5.** Pareto-Verteilung höherer Einkommen

Wird eine Normalverteilung als Modell zugrundegelegt, so stellt sich die Log-Likelihood-Funktion bei gegebener Stichprobe  $x = (x_1, \dots, x_N)$  wie folgt dar:

$$\begin{aligned} \ell(\mu, \sigma) &= \sum_{n=1}^N \ln(\mathcal{N}(x_n | \mu, \sigma^2)) \\ &= \sum_{n=1}^N \ln \left( \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot \frac{(x_n - \mu)^2}{\sigma^2}} \right) \\ &= -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - N \ln(\sigma) - \frac{N}{2} \ln(2\pi) \end{aligned}$$

Der Gradient der Log-Likelihood-Funktion ist wie folgt gegeben:

$$\text{grad } \ell(\mu, \sigma) = \begin{pmatrix} \frac{\partial \ell}{\partial \mu}(\mu, \sigma) \\ \frac{\partial \ell}{\partial \sigma}(\mu, \sigma) \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu) \\ \frac{N}{\sigma^3} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{\sigma} \end{pmatrix}$$

Das gesuchte Maximum  $(\hat{\mu}, \hat{\sigma})$  findet sich notwendigerweise an der Nullstelle des Gradienten, daraus ergibt sich:

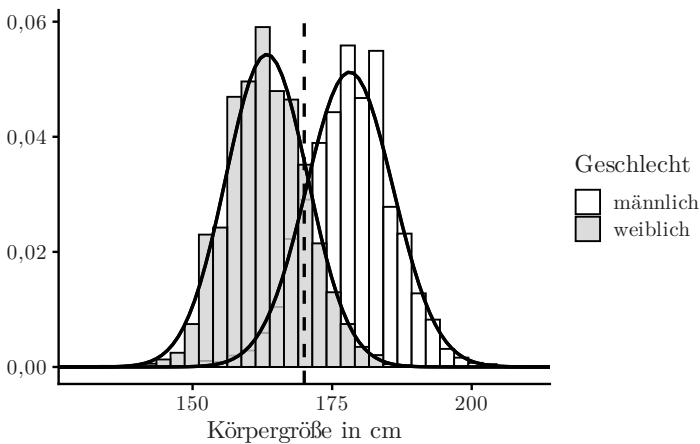
$$\begin{aligned}\hat{\mu} &= \frac{1}{N} \sum_{n=1}^N x_n = \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2 = s^2(x)\end{aligned}$$

Durch Untersuchung der zweiten Ableitungen (der Hesse'schen Matrix) kann bestätigt werden, dass diese Parameter tatsächlich eine Maximalstelle darstellen.

**Anwendungsbeispiel.** Wir glauben, dass die Körpergröße in der CDC-Studie der Kohorten von weiblichen bzw. männlichen Befragten jeweils normalverteilt sind. Es ergibt sich:

$$\begin{aligned}\hat{\mu}(\text{Körpergröße|weiblich}) &= 163 \text{ cm}, & \hat{\sigma}(\text{Körpergröße|weiblich}) &= 7,3 \text{ cm} \\ \hat{\mu}(\text{Körpergröße|männlich}) &= 178 \text{ cm}, & \hat{\sigma}(\text{Körpergröße|männlich}) &= 7,8 \text{ cm}\end{aligned}$$

In folgender Abbildung sind die Histogramme den Normalverteilungen mit geschätzten Parametern gegenübergestellt:



**Abb. 4.6.** Normalverteilung der Körpergröße

Die senkrechte gestrichelte Linie liegt bei einer KörpergröÙe von 170 cm und grenzt die Kohorten auf Basis der KörpergröÙe voneinander ab. In Ab-

schn. 6.3.3 kommen wir noch einmal auf dieses Beispiel zurück und erklären, wie diese sogenannte Entscheidungsgrenze berechnet wird.

Die Maximum-Likelihood-Methode kann auch auf diskrete Modelle angewendet werden, wie wir mit folgendem Beispiel illustrieren wollen: Wir nehmen an einer Lotterie teil, bei der farbige Bälle aus einem nicht einsehbaren Topf mit Zurücklegen gezogen werden. Die Anzahl der Bälle ist ebenso unbekannt wie die Anzahl der Farben. Wir glauben jedoch, dass die Wahrscheinlichkeit für das Ziehen der Farben gleichverteilt ist. Nummerieren wir die Farben von 1 bis  $K$  durch, ist die Annahme folglich:

$$\Pr(\{\text{Farbe Nr. } k \text{ wird gezogen}\}) = \mathcal{U}(k|K) = \begin{cases} \frac{1}{K} & \text{falls } k \in \{1, \dots, K\} \\ 0 & \text{sonst} \end{cases}$$

Wir beobachten  $N$  Lotteriezessionen und notieren die Ergebnisse  $k_1, \dots, k_N \in \{1, 2, \dots\}$ , wobei jeder zuvor noch nicht beobachteten Farbe  $k_{n+1}$  die Nummer  $\max_n\{k_n\} + 1$  zugeordnet wird, beispielsweise:

$n$	1	2	3	4	5	6	7	8	...
Farbe	rot	rot	gelb	blau	rot	gelb	grün	grün	...
$k_n$	1	1	2	3	1	2	4	4	...

**Tabelle 4.4.** Aufzeichnung einer Lotterie

Die Log-Likelihood-Funktion ist wie folgt gegeben:

$$\begin{aligned} \ell(K) &= \sum_{n=1}^N \ln(\mathcal{U}(k_n|K)) \\ &= \begin{cases} -\infty & \text{falls } k_n > K \text{ für wenigstens ein } n \in \{1, \dots, N\} \\ -N \ln K & \text{sonst} \end{cases} \end{aligned}$$

Der Fall  $\max_n\{k_n\} > K$  entspricht der Situation, in der mehr als  $K$  verschiedene Farben gezogen würden. Ein solches Modell wäre mit den Beobachtungen in keinem Fall vereinbar, wir können also gleich  $k_n \leq K$  annehmen.

Der Term  $\ell(K) = -N \ln K$  fällt streng monoton mit  $K$  ab. Daher müssen wir  $K$  so klein wie möglich wählen, ohne der Bedingung  $k_n \leq K$  zu widersprechen. Damit bleibt nur eine Wahl als Maximum-Likelihood-Schätzung für  $K$ :

$$\hat{K} = \max_{n \in \{1, \dots, N\}} k_n$$

Zugleich ist dies auch ein Schätzer für die Anzahl der Farben im Lostopf: Wenn wir beobachten, dass  $\hat{K}$  verschiedene Farben gezogen wurden, so ist die im Sinne einer Maximum-Likelihood-Schätzung optimale Annahme, dass sich insgesamt  $\hat{K}$  verschiedene Farben im Lostopf befinden.

Diese Schätzung ist allerdings nicht erwartungstreu. Fassen wir  $\hat{K} = \hat{K}(N)$  als Funktion in Stichprobenvariablen  $X_1, \dots, X_N$  mit Gleichverteilung  $\mathcal{U}(\cdot | K_0)$  auf, wobei  $K_0$  die wahre Anzahl verschiedener Farben im Lostopf bezeichnet, so können wir deren Erwartungswert berechnen. Die Wahrscheinlichkeit, dass die  $k$ -te Farbe in einer Stichprobe der Größe  $N$  auftaucht, beträgt:

$$\begin{aligned} 1 - \Pr(X_1 \neq k, \dots, X_N \neq k) &= 1 - \prod_{n=1}^N \Pr(X_n \neq k) \\ &= 1 - \prod_{n=1}^N (1 - \Pr(X_n = k)) \\ &= 1 - \left(1 - \frac{1}{K_0}\right)^N \end{aligned}$$

Der Erwartungswert der Anzahl der gezogenen Farben ist daher:

$$\begin{aligned} E[\hat{K}(N)] &= \sum_{k=1}^{K_0} \left(1 - \left(1 - \frac{1}{K_0}\right)^N\right) \\ &= K_0 - K_0 \cdot \left(1 - \frac{1}{K_0}\right)^N \end{aligned}$$

Die Maximum-Likelihood-Schätzung  $\hat{K}$  liefert also stets einen kleineren als den wahren Wert  $K_0$ . Da die Differenz vom unbekannten Wert  $K_0$  abhängt, können wir sie in der Praxis jedoch für eine Korrektur nicht einfach zur Schätzung addieren.

#### 4.4.2 Bayes'sche Parameterschätzung

Bei der Maximum-Likelihood-Methode wird davon ausgegangen, dass die Parameter des statistischen Modells für die Häufigkeiten, mit denen bestimmte Stichprobenwerte vorkommen, bestimmend sind. Aus den beobachteten Häufigkeiten werden diese Parameter dann rekonstruiert. Diese Schätzung hat dann zwar nur endliche Genauigkeit – dennoch geht die Methode von einem festen Parameterwert aus, deren genauer Wert der Statistikerin bzw. dem Statistiker schlicht unbekannt ist.

Konzeptionell ist es im Rahmen der Maximum-Likelihood-Schätzung also wenig sinnvoll, von einer Wahrscheinlichkeit zu sprechen, mit der die Parameter bestimmte Werte annehmen; die Parameter selbst sind keine Zufallsvariablen. Eine alternative Sichtweise zu dieser frequentistischen Statistik vertritt die **Bayes'sche Statistik**: Letztere macht sich die Annahme zunutze, dass die statistischen Parameter selbst Wahrscheinlichkeitsverteilungen genügen, die wiederum von sogenannten Hyperparametern abhängig sein können.

Um diesen Gedanken in sinnvoller Weise anwenden zu können, modifizieren wir die Likelihood-Funktion wie folgt.

Die sogenannte **A-posteriori-Verteilung** der Parameter eines statistischen Modells  $p(\cdot | \theta_1, \dots, \theta_K)$  ist proportional zum Produkt von Likelihood-Funktion und **A-priori-Verteilung**  $p_{\text{prior}}(\cdot | \alpha_1, \dots, \alpha_K)$ :

$$p_{\text{post}}(\theta_1, \dots, \theta_K | x_1, \dots, x_N; \alpha_1, \dots, \alpha_K) \\ \propto \left( \prod_{n=1}^N p(x_n | \theta_1, \dots, \theta_K) \right) \cdot p_{\text{prior}}(\theta_1, \dots, \theta_K | \alpha_1, \dots, \alpha_K)$$

Dabei sind vorgegeben: Die Stichprobe  $x = (x_1, \dots, x_N)$  sowie die **Hyperparameter**  $\alpha_1, \dots, \alpha_K$ .

Um aus der rechten Seite der obigen Formel eine Massen- bzw. Dichtefunktion zu gewinnen, muss diese noch mit einem Normierungsfaktor multipliziert werden, damit Integral bzw. Summe über die Parameter gleich eins ist. Mithin muss die A-priori-Verteilung nicht unbedingt eine Massen- oder Dichtefunktion sein, solange die rechte Seite geeignet normiert werden kann. In diesem Fall wird von einer **uneigentlichen A-priori-Verteilung** gesprochen.

Im Endergebnis erhalten wir also keine Punktschätzung  $\hat{\theta}_1, \dots, \hat{\theta}_K$  der Parameter, sondern eine gemeinsame Verteilung über mögliche Werte  $\theta_1, \dots, \theta_K$ .

Die A-priori-Verteilung modelliert Informationen, die bereits vor Erhebung der Stichprobe über die Parameter vorhanden sein mögen. Beispielsweise könnte die Spielerin eines Glücksspiels davon überzeugt sein, dass eine bestimmte Münze mit hoher Wahrscheinlichkeit fair ist. Anders ausgedrückt: A priori wird angenommen, dass sich der Parameter der entsprechenden Bernoulli-Verteilung mit hoher Wahrscheinlichkeit wenig von  $p = 1/2$  unterscheidet. Entsprechend würde die Spielerin für eine Bayes'sche Analyse ihrer Gewinnchancen eine A-priori-Verteilung mit Lageparameter  $\mu = 1/2$  ansetzen sowie einem Streuungsparameter, der ihre initiale Unsicherheit über diese Information wiederspiegelt. Vielleicht ist der Spieelleiter besonders vertrauenswürdig und die Münze sieht zunächst nicht ungewöhnlich aus, dann wird dieser Streuungsparameter klein sein. Durch Multiplikation der initialen A-priori-Verteilung mit der von der tatsächlichen Datenlage (d. h., der beobachteten Folge von Münzwürfen) abhängigen Likelihood-Funktion erhält die Spielerin dann eine aktualisierte Überzeugung von dem Wert für  $p$ , welche dann durch die A-posteriori-Verteilung gegeben ist.

Durch das angeführte Beispiel sollte deutlich werden, warum Bayes'schen Ansätzen im Bereich des maschinellen Lernens eine besondere Bedeutung zukommt: Die Methoden können direkt als eine Art von Lernprozess aufgefasst werden, bei der Informationen in Form von Parametern statistischer Modelle durch neue Daten fortlaufend aktualisiert werden können.

Aus der A-posteriori-Verteilung können wir eine Punktschätzung gewinnen, indem wir – analog wie beim Maximum-Likelihood-Verfahren – deren Maximalstelle ermitteln. In diesem Fall wird von einer **Maximum-a-posteriori-**

**Schätzung** (MAP) gesprochen. Wird als (uneigentliche) A-priori-Verteilung eine konstante Funktion gewählt, so entspricht das der Situation, in der wir keine A-priori-Informationen über die zu schätzenden Parameter haben: Vor Erhebung der Stichprobe sind alle Parameterwerte gleich wahrscheinlich, wir sprechen dann auch von einer **nichtinformativen A-priori-Verteilung**. In diesem Fall ist die A-posteriori-Verteilung proportional zur Likelihood-Funktion, und Maximum-a-posteriori- und Maximum-Likelihood-Schätzer stimmen überein.

Wir wollen die wesentlichen Charakteristiken des Bayes'schen Verfahrens der Parameterschätzung anhand eines speziellen Falls verdeutlichen (siehe auch [9, Abschnitt 3.4.1]). Sei dazu eine Stichprobe  $x = (x_1, \dots, x_N)$  gegeben, von der wir ausgehen, dass diese normalverteilt ist. Der Einfachheit halber nehmen wir weiterhin an, dass der Streuungsparameter  $\sigma$  bekannt ist und  $\sigma_0 > 0$  beträgt; es muss also nur noch der Lageparameter  $\mu$  geschätzt werden:  $p(\cdot | \mu) = \mathcal{N}(\cdot | \mu, \sigma_0^2)$ . Als A-priori-Verteilung für  $\mu$  wählen wir ebenfalls eine Gauß-Verteilung, mit Lageparameter  $\mu_0$  und Streuungsparameter  $\Delta_0$ .

$$\begin{aligned} p_{\text{post}}(\mu | x; \mu_0, \Delta_0) &\propto \left( \prod_{n=1}^N p(x_n | \mu) \right) \cdot p_{\text{prior}}(\mu | \mu_0, \Delta_0) \\ &= \left( \prod_{n=1}^N \mathcal{N}(x_n | \mu, \sigma_0^2) \right) \cdot \mathcal{N}(\mu | \mu_0, \Delta_0^2) \\ &\propto e^{-\frac{(\mu-\mu_0)^2}{2(\Delta_0)^2}} \cdot \prod_{n=1}^N e^{-\frac{(x_n-\mu)^2}{2(\sigma_0)^2}} \end{aligned}$$

Der funktionalen Form sehen wir an, dass die A-posteriori-Verteilung wiederum eine Normalverteilung sein muss:

$$p_{\text{post}}(\mu | x; \mu_0, \Delta_0) = \frac{1}{\sqrt{2\pi}\Delta_N} \cdot e^{-\frac{(\mu-\hat{\mu}_N)^2}{2\Delta_N^2}}$$

Eine etwas längere Rechnung führt auf Lage- und Streuungsparameter dieser Verteilung:

$$\begin{aligned} \hat{\mu}_N &= \frac{N\Delta_0^2 \cdot \bar{x} + \sigma_0^2 \cdot \mu_0}{N\Delta_0^2 + \sigma_0^2}, \\ \Delta_N &= \frac{\Delta_0 \sigma_0}{\sqrt{N\Delta_0^2 + \sigma_0^2}}. \end{aligned}$$

Die Maximum-a-posteriori-Schätzung von  $\mu$  ist gerade durch  $\hat{\mu}_N$  gegeben. Anstelle einer solchen Punktschätzung kann an dieser Stelle auch in sinnvoller Weise ein Intervallschätzer angegeben werden:

$$[\bar{x}]_{\text{MAP}, \gamma} = [\hat{\mu}_N - z(\gamma) \cdot \Delta_N, \hat{\mu}_N + z(\gamma) \cdot \Delta_N]$$

mit  $0 < \gamma < 1$  und z. B.  $z(0,95) = 1,96$ . Mit der Wahrscheinlichkeit  $\gamma$  nimmt der Parameter einen Wert in diesem Intervall an. In der Bayes'schen Statistik

wird ein solches Intervall **Glaubwürdigkeitsintervall** oder **Kredibilitätsintervall** genannt, um den Begriff vom Vertrauensintervall der frequentistischen Statistik abzugrenzen. Die Interpretation von letzterem ist, dass die Schätzungen innerhalb des Intervalls variieren können, der wahre Parameter jedoch einen festen, deterministischen Wert hat.

Ungeachtet der verschiedenen Interpretationen stimmen für große Stichproben die Bayes'schen Intervallschätzungen mit der in Abschn. 4.3.1 beschriebenen frequentistischen Schätzung näherungsweise überein. Dies folgt zum einen aus der Tatsache, dass Maximum-Likelihood-Schätzung (= arithmetischer Mittelwert  $\bar{x}$ ) und Maximum-a-posteriori-Schätzung  $\hat{\mu}_N$  für große Stichproben näherungsweise übereinstimmen:

$$\lim_{N \rightarrow \infty} \hat{\mu}_N = \bar{x}$$

Zum anderen sind auch die Folgen der Intervallgrenzen asymptotisch gleich:

$$\lim_{N \rightarrow \infty} \frac{\sigma_0/\sqrt{N}}{\Delta_N} = 1$$

**Anwendungsbeispiel.** Wir nehmen an, dass die Körpergröße männlicher Umfrageteilnehmer der CDC-Studie normalverteilt mit Standardabweichung  $\sigma_0 = 7,8$  cm ist. Wir nehmen weiterhin an, der Mittelwert  $\mu$  sei unbekannt, aber eine A-priori-Schätzung von  $\mu_0 = 175$  cm mit der Genauigkeit  $\Delta_0 = 5,0$  cm sei möglich. Diese A-priori-Schätzung kann durch die Feststellung neuer Daten verbessert werden:

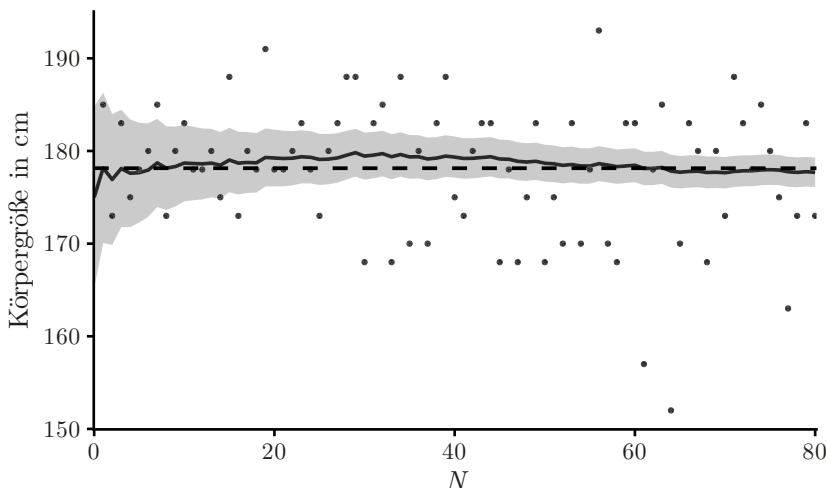


Abb. 4.7. Bayes'sche Schätzung eines arithmetischen Mittelwerts

Obige Abbildung zeigt die Ergebnisse einer Maximum-a-posteriori-Schätzung auf Grundlage eines wachsenden Ausschnitts aus der Stichprobe mit zugehörigem 95 %-Glaubwürdigkeitsintervall. Die gestrichelte Linie zeigt zum Vergleich den arithmetischen Mittelwert von 178 cm über der gesamten Stichprobe.

#### 4.4.3 Kerndichteschätzung

Eine weitere Methode, um die Gestalt einer Wahrscheinlichkeitsdichtefunktion aus einer Stichprobe zu schätzen, besteht in dem folgenden Ansatz.

Sei  $x_1, \dots, x_N$  eine Stichprobe metrischer Werte. Die **Kerndichteschätzung** mit den **Bandbreiten**  $h_1, \dots, h_N \in ]0, \infty[$  ist wie folgt gegeben:

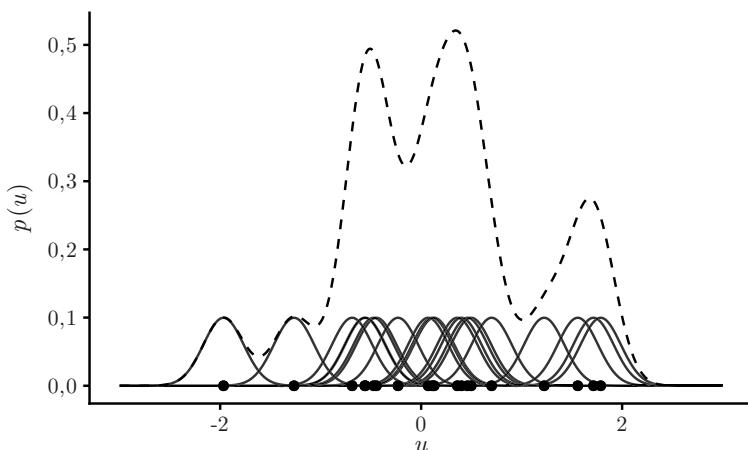
$$p_{h_1, \dots, h_N}(u) = \sum_{n=1}^N \frac{1}{Nh_n} K\left(\frac{u - x_n}{h_n}\right)$$

für alle  $u \in \mathbb{R}$ . Dabei ist der **Kern**  $K: \mathbb{R} \rightarrow [0, \infty[$  eine fest gewählte Funktion mit  $\int_{-\infty}^{\infty} K(\xi) d\xi = 1$ .

Eine beliebte Wahl für den Kern ist eine Gauß'sche Funktion:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

Wir können uns das Verfahren so vorstellen, dass an jeder Stelle, an der ein Datenpunkt sitzt, eine kleine „Gauß-Glocke“ platziert wird. Die Gesamtdichte besteht dann in der Summe aller Gauß-Glocken und dort, wo sich Datenpunkte häufen, wird diese Summe groß sein:



**Abb. 4.8.** Prinzip der Kerndichteschätzung

Die Kerndichteschätzung ähnelt einem Histogramm (siehe das Anwendungsbeispiel bzw. die Abbildung weiter unten) – ein Vorteil besteht jedoch darin, dass das Resultat keine Treppenfunktion, sondern bei Verwendung eines entsprechenden Kerns eine stetige oder sogar glatte, differenzierbare Funktion darstellt.

Nach Konstruktion gilt  $p_{h_1, \dots, h_N}(u) \geq 0$  für alle  $u \in \mathbb{R}$ , außerdem:

$$\begin{aligned} \int_{-\infty}^{\infty} p_{h_1, \dots, h_N}(\xi) d\xi &= \frac{1}{N} \sum_{n=1}^N \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{\xi - x_n}{h_n}\right) d\xi \\ &= \frac{1}{N} \sum_{n=1}^N \frac{1}{h_n} \int_{-\infty}^{\infty} h_n \cdot K(y_n) dy_n \\ &= \frac{1}{N} \sum_{n=1}^N \int_{-\infty}^{\infty} K(y_n) dy_n = 1 \end{aligned}$$

Dabei wurde im Integral für jeden Summanden die Substitution  $y_n = \frac{\xi - x_n}{h_n}$  vorgenommen. Daher ist  $p_{h_1, \dots, h_N}(\cdot)$  tatsächlich stets eine Wahrscheinlichkeitsdichte.

Erfüllt der Kern zudem noch die Bedingung  $\int_{-\infty}^{\infty} \xi \cdot K(\xi) d\xi = 0$  (dies setzen wir im Folgenden stets voraus), so zeigt eine ähnliche Rechnung:

$$\int_{-\infty}^{\infty} \xi \cdot p_{h_1, \dots, h_N}(\xi) d\xi = \bar{x}$$

Die geschätzte Dichtefunktion reproduziert in diesem Fall also den arithmetischen Mittelwert:  $E[X \sim p_{h_1, \dots, h_N}(\cdot)] = \bar{x}$ . Die aus der Kerndichteschätzung gewonnene Varianz weicht jedoch stets von der empirischen Varianz ab, es gilt:

$$\sigma^2[X \sim p_{h_1, \dots, h_N}(\cdot)] = s^2(x) + \int_{-\infty}^{\infty} \xi^2 \cdot K(\xi) d\xi \cdot \frac{1}{N} \sum_{n=1}^N h_n^2$$

Die Parameter  $h_1, \dots, h_N$  sind Hyperparameter in dem Sinn, dass diese nicht ohne Weiteres durch die Maximum-Likelihood-Methode bestimmt werden können; dieser Ansatz würde auf verschwindende Bandbreiten führen.

Die Kerndichteschätzung kann als Approximation des Histogramms – eine unstetige Treppenfunktion mit Sprungstellen – durch eine stetige Funktion ohne Sprungstellen aufgefasst werden. Daher wird die Methode mitunter auch als **Kernglättung** bezeichnet. Kleinere Werte für die Bandbreite führen zu einer größeren Anpassung an ein Histogramm mit geringer Klassenbreite. Höhere Werte für die Bandbreite bewirken hingegen eine stärkere lokale Mittelung und folglich Glättung der geschätzten Dichte.

**Anwendungsbeispiel.** Die folgende Abbildung zeigt zwei Kerndichteschätzungen auf Grundlage der Häufigkeitsverteilung des Körpergewichts im CDC-Datensatz. Es wurden eine konstante Bandbreite von  $h = 5,0 \text{ kg} = h_1 = h_2 = \dots = h_N$  (durchgezogene Linie) bzw.  $h = 1,2 \text{ kg}$  (getrichelte Linie) zugrundegelegt und ein Gauß'scher Kern verwendet. Zum Vergleich ist auch das Histogramm mit Klassenbreite 2,27 kg zu sehen.

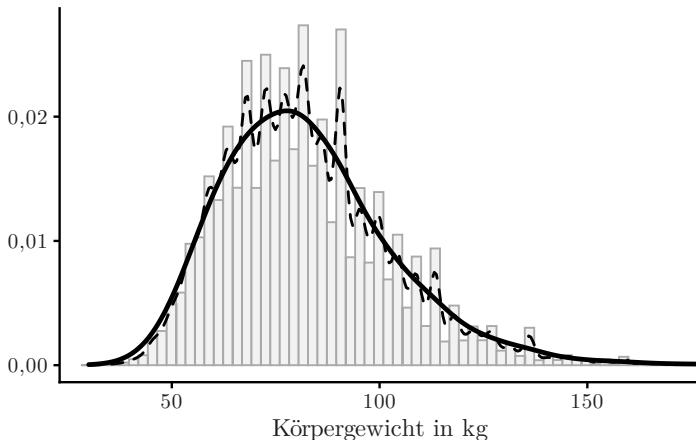


Abb. 4.9. Kerndichteschätzung verschiedener Bandbreite

## 4.5 Regressionsanalyse

Im vorigen Abschnitt betrachteten wir parametrisierte Familien von Massen- und Dichtefunktionen, die der Beschreibung der Häufigkeitsverteilung einzelner Zufallsgrößen dienen. Die Regressionsanalyse dient der Untersuchung von mehreren, in der Regel funktional assoziierten Zufallsgrößen: Es gilt, funktionale Abhängigkeiten zwischen den Größen zu modellieren und die Parameter der Modelle anhand von Korrelationsmustern in ihren Ausprägungen zu schätzen.

### 4.5.1 Einfache lineare Regression

Wir betrachten eine Folge von gepaarten Beobachtungen  $(x_1, y_1), \dots, (x_N, y_N)$ . Wie üblich stellen wir uns diese als Realisierungen von unabhängigen und identisch verteilten Stichprobenvariablen  $X_1, X_2, \dots, X_N$  bzw. von unabhängigen und identisch verteilten Stichprobenvariablen  $Y_1, Y_2, \dots, Y_N$  vor. Wir erweitern diese Folgen gedanklich um eine Stichprobenvariable  $X_* = X_{N+1}$  bzw.  $Y_* = Y_{N+1}$ , welche etwa einer noch nicht gemachten Beobachtung oder einem noch nicht durchgeföhrten Zufallsexperiment entspricht. Ziel soll es sein, anhand der bisher tatsächlich gemachten Beobachtungen eine geschätzte **Prognose**  $\hat{y}_* = \hat{f}(x_*)$  für  $y_*$  unter der Bedingung  $X_* = x_*$  durchzuführen.

Hierfür gehen weiter davon aus, dass die  $Y_n$  nicht unabhängig von den zugehörigen  $X_n$  sind. Im Gegenteil, wir möchten einen funktionalen Zusammenhang herauskehren. Bei der **einfachen linearen Regression**<sup>2</sup> liegt einem solchen Zusammenhang der folgende Ansatz zugrunde: Für jede Realisierung  $x_n$  von  $X_n$  wird angenommen, dass

$$Y_n = mx_n + c + \varepsilon_n$$

mit Konstanten  $m, c \in \mathbb{R}$  und normalverteilten, unabhängigen Zufallsvariablen  $\varepsilon_n \sim \mathcal{N}(\cdot | 0, \sigma^2)$  gilt. Wir gehen also von folgendem Szenario aus:

- Die Realisierungen  $y_*$  von  $Y_*$  streuen normalverteilt um Mittelwerte herum,
- und diese Mittelwerte liegen auf einer Geraden  $f(x_*) = mx_* + c$ , in Abhängigkeit der Realisierungen  $x_*$  von  $X_*$ .

Bei Regressionsmodellen wird die Variable  $Y_*$  die **Zielgröße** oder abhängige Variable genannt, während  $X_*$  eine unabhängige Variable oder **Einflussgröße** darstellt. Die Zufallsvariable  $\varepsilon_*$  wird **Störgröße** genannt.

Dieses Modell führt auf die folgende bedingte Wahrscheinlichkeitsdichte, um die Beobachtungen zu erklären:

$$p(y_n|x_n; m, c, \sigma) = \mathcal{N}(y_n|mx_n + c, \sigma^2)$$

für alle  $n \in \{1, \dots, N, *, \dots\}$ .

Die aus dem Modell abgeleitete Log-Likelihood-Funktion ergibt sich somit wie folgt:

$$\begin{aligned}\ell(m, c, \sigma) &= \sum_{n=1}^N \ln(\mathcal{N}(y_n|mx_n + c, \sigma^2)) \\ &= -\frac{1}{2\sigma^2} \sum_{n=1}^N (y_n - mx_n - c)^2 - N \ln(\sigma) - \frac{N}{2} \ln(2\pi)\end{aligned}$$

Eine Berechnung der Maximalstelle  $(\hat{m}, \hat{c}, \hat{\sigma})$  führt schließlich auf die gesuchten Modellparameter:

Die Steigung und der  $y$ -Achsenabschnitt der **Regressionsgeraden** oder **Ausgleichsgeraden**

$$\hat{f}: \mathbb{R} \rightarrow \mathbb{R}, \hat{f}(x_*) = \hat{m}x_* + \hat{c}$$

zweier gepaarter Stichproben  $x_1, \dots, x_N$  und  $y_1, \dots, y_N$  metrischer Werte bestimmen sich wie folgt:

---

<sup>2</sup> Der Zusatz „einfach“ bezieht sich darauf, dass nur eine unabhängige Größe  $X$  herangezogen wird. Daher können wir auch von univariater linearer Regression sprechen.

$$\hat{m} = \frac{\sum_{n=1}^N (x_n - \bar{x}) \cdot (y_n - \bar{y})}{\sum_{n=1}^N (x_n - \bar{x})^2} = \frac{s(x, y)}{s^2(x)},$$

$$\hat{c} = \frac{1}{N} \sum_{n=1}^N (y_n - \hat{m}x_n) = \bar{y} - \hat{m}\bar{x}$$

Der mittlere quadrierte Abstand der  $y$ -Werte von der Regressionsgeraden ist wie folgt gegeben:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (y_n - \hat{m}x_n - \hat{c})^2$$

Die Kennzahl  $\hat{\sigma}$  kann als ein Maß für die Güte des Modells herangezogen werden. Sie gibt an, wie breit die Datenpunkte um die Ausgleichsgerade herum streuen: Im Grenzfall  $\hat{\sigma} = 0$  würden sie alle genau auf der Geraden liegen. Die Größe

$$N\hat{\sigma}^2 = \sum_{n=1}^N (y_n - \hat{m}x_n - \hat{c})^2 = \sum_{n=1}^N (y_n - \hat{y}_n)^2$$

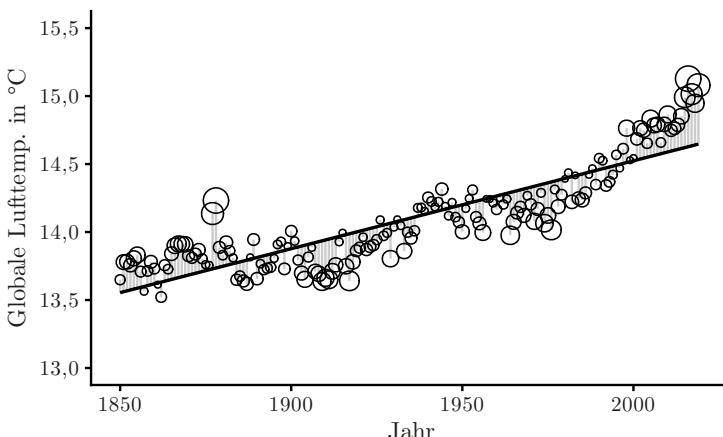
wird auch **Residuenquadratsumme** genannt. Bezuglich der Parameter  $m, c$  ist eine Maximierung der Likelihood äquivalent zur *Minimierung* der Residuenquadratsumme, aufgefasst als Funktion der Parameter:

$$R(m, c) = \sum_{n=1}^N (y_n - mx_n - c)^2.$$

Diese Sichtweise auf die lineare Regression wird **Methode der kleinsten Quadrate** genannt. Viele Verfahren der Statistik oder des maschinellen Lernens basieren wesentlich auf der Minimierung oder Maximierung einer solchen geeigneten **Zielfunktion**.

**Anwendungsbeispiel.** Das Projekt Berkeley Earth Surface Temperature stellt Messdaten zur Quantifizierung der globalen Erderwärmung zur Verfügung [10]. Die im Folgenden dargestellte Abbildung zeigt den zeitlichen Verlauf der mittleren globalen Lufttemperatur seit 1850 zusammen mit dem Ergebnis einer linearen Regression.

Die Steigung der Ausgleichsgeraden beträgt  $\hat{m} = 0,006 \text{ K/a}$ , das entspricht einer Erwärmung um 0,6 Kelvin im Jahrhundert. Die Größe der Beobachtungspunkte zeigt zudem das jeweilige **Residuum** an: den  $y$ -parallelen Abstand des Datenpunkts zur ermittelten Ausgleichsgeraden. Auf diese Weise werden Ausreißer visuell hervorgehoben. Die mittlere Abweichung von der Regressionsgeraden liegt bei  $\hat{\sigma} = 0,17 \text{ K}$ .



**Abb. 4.10.** Globale Temperaturentwicklung mit Ausgleichsgerade

Aus der Abbildung geht auch hervor: Messdaten im 20. Jahrhundert liegen unterhalb der Regressionsgeraden, während Messdaten vor 1900 bzw. nach 2000 oberhalb der Geraden liegen. Dieser Umstand deutet darauf hin, dass der wahre funktionale Zusammenhang eine konvexe Funktion darstellt, die Erwärmung also nicht gleichmäßig, sondern *beschleunigt* ist. Darauf kommen wir später noch einmal zurück, vgl. Abb. 6.2.

Eine alternative Schreibweise für die Gleichung der Regressionsgeraden ist die folgende:

$$\hat{f}: \mathbb{R} \rightarrow \mathbb{R}, \hat{f}(x_*) = \hat{\beta}(x_* - \bar{x}) + \hat{\alpha}$$

mit  $\hat{\beta} = \hat{m}$  und  $\hat{\alpha} = \bar{y}$ . Wir erkennen unter anderem, dass jede Regressionsgerade stets durch den geometrischen Schwerpunkt  $(\bar{x}, \bar{y})$  der Daten verläuft.

Mit Mitteln der statistischen Theorie kann anstelle eines bloßen Punktschätzers  $\hat{f}(x_*)$  auch ein  $\gamma$ -Vertrauensintervall berechnet werden, auf dessen Herleitung wir hier verzichten. Für eine Stelle  $x_* \in \mathbb{R}$  und hinreichend große Stichproben gilt

$$[\hat{f}(x_*)]_\gamma = \left[ \hat{f}(x_*) \pm z(\gamma) \cdot \frac{\hat{\sigma}}{\sqrt{N}} \cdot \sqrt{1 + \frac{(x_* - \bar{x})^2}{s^2(x)}} \right]$$

mit z. B.  $z(0,95) = 1,96$ . Mit variablem  $x_*$  definiert obige Formel einen Streifen um die Ausgleichsgerade. Dieser **Vertrauensbereich** spiegelt unsere Unsicherheit gegenüber der Position und Orientierung der Geraden wider. Mit wachsendem Abstand zum Mittelwert  $\bar{x}$  wächst auch die Breite dieses Bereichs und das Vertrauen in die Richtigkeit der Schätzung sinkt.

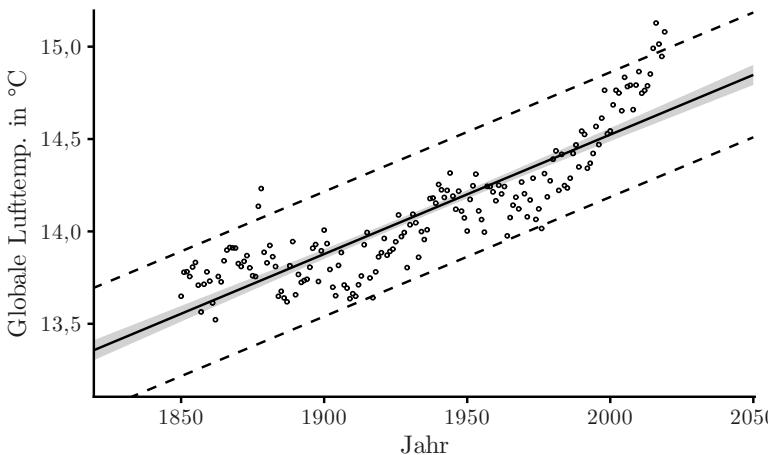
Der Vertrauensbereich ist nicht mit einem **Vorhersage-** bzw. **Prognosebereich** zu verwechseln. In einem solchen Bereich kann ein bestimmter Anteil

$0 < \delta \leq 1$  der Datenpunkte erwartet werden. Für hinreichend große Stichproben ist ein  $\delta$ -Vorhersagebereich für die einfache lineare Regresson von folgender Form:

$$\left[ \hat{f}(x_*) \pm z(\delta) \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{N} \cdot \left( 1 + \frac{(x_* - \bar{x})^2}{s^2(x)} \right)} \right]$$

Ist der Abstand der betrachteten Stelle  $x_*$  zum Mittelwert  $\bar{x}$  nicht zu groß, ist der Vorhersagebereich näherungsweise durch  $[\hat{f}(x_*) \pm z(\delta) \cdot \hat{\sigma}]$  gegeben.

Im Folgenden ist das obige Anwendungsbeispiels erneut illustriert, anstelle der empirischen Residuen ist der 95 %-Vertrauensbereich als schmaler grauer Streifen um die Ausgleichsgerade hervorgehoben. Die gestrichelten Kurven dienen hingegen der Angabe des wesentlich breiteren 95 %-Vorhersagebereichs.



**Abb. 4.11.** Vertrauens- und Vorhersagebereich der linearen Regression

Schließlich wollen wir anhand des soeben Gelernten die Bedeutung des Korrelationkoeffizienten nach Bravais-Pearson etwas genauer beleuchten. Die Steigung der mittels linearer Regression ermittelten Ausgleichsgeraden beträgt

$$\hat{m}(y, x) = \frac{s(x, y)}{s^2(x)},$$

wenn wir  $x$  als Ausprägungen der Einflussgröße und  $y$  als Ausprägungen der Zielgröße auffassen. Vertauschen wir die Rollen der Merkmale, so erhalten wir eine Steigung von

$$\hat{m}(x, y) = \frac{s(y, x)}{s^2(y)} = \frac{s(x, y)}{s^2(y)}.$$

Zunächst einmal fällt auf, dass  $\text{sgn}(\hat{m}(y, x)) = \text{sgn}(\hat{m}(x, y))$  gilt: Die Steigung beider Regressionsgeraden haben das gleiche Vorzeichen. Hiernach lässt sich die folgende Formel für den Korrelationskoeffizienten ableiten:

$$r(x, y) = \frac{s(x, y)}{s(x) \cdot s(y)} = \operatorname{sgn}(\hat{m}(y, x)) \cdot \sqrt{\hat{m}(y, x) \cdot \hat{m}(x, y)}$$

Betragsmäßig ist der Korrelationskoeffizient also gerade das geometrische Mittel der Geradensteigungen. Dies unterstreicht noch einmal die Interpretation dieser Kennzahl als ein Maß für linearen Zusammenhang.

#### 4.5.2 Theil-Sen-Verfahren

Eine Alternative zur einfachen linearen Regression stellt das folgende Verfahren dar.

Seien  $x_1, \dots, x_N$  und  $y_1, \dots, y_N$  gepaarte Stichproben metrischer Werte. Beim **Theil-Sen-Verfahren** werden Steigung und  $y$ -Achsenabschnitt einer Ausgleichsgeraden wie folgt berechnet:

$$\begin{aligned}\hat{m}_{\text{TS}} &= \underset{\substack{k, l \in \{1, \dots, N\} \\ x_k \neq x_l}}{\operatorname{median}} \left( \frac{y_l - y_k}{x_l - x_k} \right), \\ \hat{c}_{\text{TS}} &= \underset{k \in \{1, \dots, N\}}{\operatorname{median}} (y_k - \hat{m}_{\text{TS}} \cdot x_k)\end{aligned}$$

Das Theil-Sen-Verfahren ist eine robustere Methode als die lineare Regression: Durch die Verwendung des Medians haben Ausreißer einen geringeren Einfluss auf die ermittelte Regressionsgerade. Bei der linearen Regression können hingegen bereits wenige Ausreißer in den Daten zu ganz anderen Ergebnissen führen.

**Anwendungsbeispiel.** Neben dem monatlichen Nettoeinkommen erfassst die ALLBUS-Studie auch die tägliche Fernsehgesamtdauer der Befragten. Wir wollen herausfinden, ob zwischen beiden Merkmalen ein Zusammenhang besteht. Die im Folgenden dargestellte Abbildung zeigt für die Kohorte der alleinwohnenden und ganztägig berufstätigen Männer die Fernsehdauer dem Einkommen gegenübergestellt.

Die mithilfe linearer Regression ermittelte Ausgleichsgerade hat eine leicht negative Steigung, woraus gefolgert werden könnte, dass Personen in der gewählten Kohorte mit höherem Einkommen weniger fernsehen. Genauer gilt  $\hat{m} = -0,012 \frac{\text{min.}}{\text{EUR}}$ , also eine Verminderung der Fernsehdauer von 12 Minuten pro 1000 EUR zusätzlichem Einkommen.

Allerdings gilt  $\hat{\sigma} = 68 \text{ min.}$ , sodass die prognostische Stärke dieser Aussage zweifelhaft ist. Dariüber hinaus hat die über den robusten Theil-Sen-Schätzer ermittelte Ausgleichsgerade eine *verschwindende* Steigung  $\hat{m}_{\text{TS}} = 0$ . Diese Analyse kommt somit zu einem anderen Schluss: Im Schnitt sieht jede Person täglich 120 Minuten fern, unabhängig vom Einkommen.

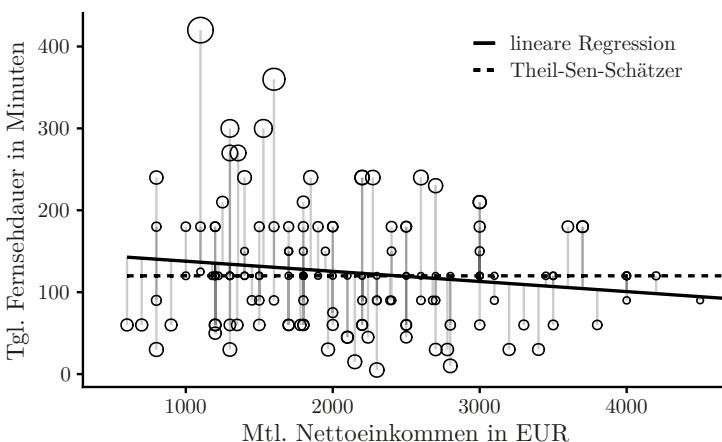


Abb. 4.12. Theil-Sen-Regression

### 4.5.3 Einfache logistische Regression

Bei der logistischen Regression ist die Zielgröße  $Y_*$  keine stetige Zufallsvariable wie bei der linearen Regression, sondern eine diskrete Bernoulli-verteilte Zufallsvariable: Es sind nur die Realisierungen  $Y_* = 0$  und  $Y_* = 1$  möglich. Der Ansatz der logistischen Regression ist der folgende, wobei die  $x_n$  wieder Realisierungen der Einflussgröße darstellen:

$$Y_n = \begin{cases} 1 & \text{falls } mx_n + c + \varepsilon_n > 0 \\ 0 & \text{sonst} \end{cases}$$

Dabei wird davon ausgegangen, dass die voneinander unabhängigen Störgrößen  $\varepsilon_n$  einer sogenannten **logistischen Verteilung** folgen:  $\varepsilon_n \sim \text{Logist}(\cdot | 0, 1)$ . Die logistische Verteilung ist wie folgt definiert:

$$\text{Logist}(u | \mu, s) = \frac{1}{4s} \text{sech}^2 \left( \frac{u - \mu}{2s} \right)$$

mit dem Sekans hyperbolicus

$$\text{sech}(u) = \frac{2}{e^u + e^{-u}}.$$

Wir können uns den Ansatz wie folgt vorstellen: Um die Realisierungen der diskreten Zufallsvariablen  $Y_n$  zu erklären, führen wir die stetige Zufallsvariable  $Y_n^* = mx_n + c + \varepsilon_n$  ein, welche über die Bedingung  $Y_n^* > 0$  die Realisierung  $Y_n = 1$  bewirkt. Die Zufallsvariable  $Y_n^*$  wird in diesem Zusammenhang eine **latente Variable** genannt, da sie sozusagen „im Hintergrund“ besteht: Ihre Realisierungen entsprechen keiner tatsächlichen Beobachtung.

Mit diesem Ansatz ergeben sich die folgenden bedingten Wahrscheinlichkeiten für die möglichen Realisierungen von  $Y_n$ :

$$\begin{aligned}\Pr(Y_n = 1|X_n = x_n) &= \Pr(Y_n^* > 0|X_n = x_n) \\ &= \int_0^\infty \text{Logist}(\xi|mx_n + c, 1) d\xi \\ &= \frac{1}{1 + e^{-mx_n - c}}\end{aligned}$$

bzw.

$$\begin{aligned}\Pr(Y_n = 0|X_n = x_n) &= 1 - \Pr(Y_n = 1|X_n = x_n) \\ &= 1 - \frac{1}{1 + e^{-mx_n - c}} \\ &= \frac{1}{1 + e^{mx_n + c}}\end{aligned}$$

Zusammengefasst können wir die Situation durch das folgende statistische Modell beschreiben:

$$p(y_n|x_n; m, c) = \frac{1}{1 + \exp((-1)^{y_n} \cdot (mx_n + c))}$$

für alle  $n \in \{1, \dots, N, \dots\}$ .

Damit ergibt sich die Log-Likelihood-Funktion der einfachen logistischen Regression:

$$\ell(m, c) = - \sum_{n=1}^N \ln(1 + \exp((-1)^{y_n} \cdot (mx_n + c)))$$

Die Maximalstellen  $\hat{m}, \hat{c}$  dieser Funktion können nicht in geschlossener Form berechnet werden; es werden numerische Methoden zuhilfe genommen.

Die geschätzten Parameter führen auf eine sogenannte **Entscheidungsgrenze**  $\hat{x} = -\frac{\hat{c}}{\hat{m}}$ . An dieser Stelle gilt gerade  $\hat{m}\hat{x} + \hat{c} = 0$ . Für beobachtete Werte von  $x_*$  mit  $\hat{m}x_* + \hat{c} > 0$  ist es gemäß dem logistischen Modell wahrscheinlicher,  $Y_* = 1$  vorzufinden:  $\Pr(Y_* = 1|X_* = x_*) > \Pr(Y_* = 0|X_* = x_*)$ .

In diesem Fall besteht die Entscheidungsgrenze nur aus einem Punkt, da wir nur eine unabhängige Variable betrachten. In der Praxis ist die univariate logistische Regression von geringer Bedeutung, aufbauend auf diesen Überlegungen behandeln wir jedoch die multivariate logistische Regression im Abschn. 6.3.1.

## Quellen

- [1] CDC Population Health Surveillance Branch. *Behavioral Risk Factor Surveillance System (BRFSS) Survey Data 2018*. Aufgerufen am 01. Feb. 2020. URL: <https://www.cdc.gov/brfss/>.
- [2] UN Department of Economic and Social Affairs. *World Population Prospects 2019*. Aufgerufen am 10. Juli 2020. URL: <https://population.un.org/wpp/>.
- [3] Klaus D. Schmidt. *Maß und Wahrscheinlichkeit*. 2. Aufl. Springer, Berlin, Heidelberg, 2011. ISBN: 978-3-642-21026-6. DOI: [10.1007/978-3-642-21026-6](https://doi.org/10.1007/978-3-642-21026-6).
- [4] GESIS-Leibniz-Institut für Sozialwissenschaften. *Allgemeine Bevölkerungs umfrage der Sozialwissenschaften ALLBUS 2018*. 2019. DOI: [10.4232/1.13250](https://doi.org/10.4232/1.13250).
- [5] Eugene Lukacs. „A Characterization of the Normal Distribution“. In: *Ann. Math. Statist.* 13.1 (März 1942), S. 91–93. DOI: [10.1214/aoms/1177731647](https://doi.org/10.1214/aoms/1177731647).
- [6] Radha G. Laha. „On an extension of Geary’s theorem“. In: *Biometrika* 40.1-2 (1953), S. 228–229. DOI: [10.1093/biomet/40.1-2.228](https://doi.org/10.1093/biomet/40.1-2.228).
- [7] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. 2. Aufl. New Jersey, USA: Lawrence Earlbaum Associates, 1988. ISBN: 0-8058-0283-5.
- [8] Shlomo S. Sawilowsky. „New Effect Size Rules of Thumb“. In: *Journal of Modern Applied Statistical Methods* 8.2 (Nov. 2009), S. 597–599. DOI: [10.22237/jmasm/1257035100](https://doi.org/10.22237/jmasm/1257035100).
- [9] Richard O. Duda, Peter E. Hart und David G. Stork. *Pattern Classification*. 2. Aufl. Wiley, 2000. ISBN: 978-0-471-05669-0.
- [10] Berkeley Earth. *Time Series Data – Monthly Global Average Temperature (Annual Summary)*. Aufgerufen am 01. Feb. 2020. URL: <http://berkeleyearth.org/data/>.



## Multivariate Statistik

Mit der Vorstellung von Assoziationsmaßen und Regressionsverfahren haben wir bereits einen ersten Einblick in multivariate Verfahren gewonnen. Multivariate Methoden ermöglichen die gemeinsame Untersuchung aller relevanten Merkmale und ihrer Beziehungen untereinander – mit dem Ziel ein möglichst vollständiges Bild der Daten zu erfassen.

### 5.1 Datenmatrizen

Betrachten wir eine Stichprobe vom Umfang  $N$ , bei der Werte für insgesamt  $D$  Merkmale erhoben wurden. Wenn nichts anderes gesagt wird, gehen wir im Folgenden von metrischen Merkmalen aus. Zunächst einmal können wir die  $n$ -te Beobachtung,  $n \in \{1, \dots, N\}$ , als einen  $D$ -dimensionalen Merkmalsvektor  $x_n \in \mathbb{R}^D$  auffassen. Um mit Werkzeugen der linearen Algebra wie etwa der Matrixmultiplikation arbeiten zu können, müssen wir uns auf Konventionen für das Format der involvierten Vektoren und Matrizen einigen. Zum einen können wir die Merkmalsvektoren  $x_1, \dots, x_N$  als Spalten einer Matrix hintereinander schreiben,  $x = (x_1, \dots, x_N)$ . Im Allgemeinen hat diese Matrix das Format  $D \times N$ . Am Beispiel der zwei Merkmale Körpergröße (in Meter) und Körpergewicht (in Kilogramm) könnte eine solche Matrix wie folgt aussehen:

$$x = \begin{pmatrix} 1,63 & 1,66 & 1,47 & 1,79 \\ 59 & 91 & 64 & 86 \end{pmatrix}$$

Diese Konvention ist insbesondere dann hilfreich, wenn wir Hilfsmittel der mehrdimensionalen Analysis anwenden wollen. Andernfalls müssten wir dort gebräuchliche Konventionen brechen, sodass z. B. der Gradient zu einem Zeilenvektor würde. In der Statistik ist es allerdings auch oder sogar öfter üblich, die Merkmalsvektoren als Zeilen untereinander zu schreiben, in den Spalten stehen nun die Stichprobenvektoren für die jeweiligen Merkmale:

$$x^T = \begin{pmatrix} 1,63 & 59 \\ 1,66 & 91 \\ 1,47 & 64 \\ 1,79 & 86 \end{pmatrix}$$

Die so konstruierte Matrix heißt **Datenmatrix** oder **Modellmatrix**. Sie hat dasselbe Format wie die Datentabellen, von denen wir bereits einige in vorherigen Kapiteln gesehen haben. Wir bezeichnen sie auch mit einem eigenen Symbol:  $\mathbf{X} = x^T$ . Trotz dieser verbreiteten Konvention, auf die wir ebenfalls zurückgreifen, werden wir stets Spaltenvektoren meinen, wenn wir von Merkmalsvektoren oder Datenpunkten sprechen. Ebenso sind mit Zufallsvektoren, also einem Tupel von Zufallsvariablen, stets Spaltenvektoren gemeint. Im Fließtext schreiben wir solche Spaltenvektoren auch als transponierte Zeilen, z. B.  $u = (u_1, u_2, u_3)^T$ .

Stichprobenvektoren sind hingegen in der Regel als Zeilenvektoren aufzufassen. Wird der Übersichtlichkeit der Formeln halber an einzelnen Stellen eine abweichende Konvention verwendet, so zeigen wir dies ähnlich wie bei der Datenmatrix mit Fettdruck an, also z. B.  $\mathbf{y} = y^T = (y_1, \dots, y_N)^T$ .

Wir werden auch z. B. Modellparameter zu Vektoren und Matrizen zusammenfassen. Wir weisen dann explizit auf das Format hin, sofern erforderlich.

Allgemein hat eine Datenmatrix mit  $N$  Zeilen bzw. Beobachtungen und  $D$  Spalten bzw. Merkmalen die folgende Form:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{pmatrix} = (x_{nd})_{\substack{n \in \{1, \dots, N\} \\ d \in \{1, \dots, D\}}}$$

Die Zahl  $D$  wird auch die **Dimensionalität** der Daten genannt. Die Zeilen  $\mathbf{X}_{1\bullet}, \dots, \mathbf{X}_{N\bullet}$  einer Datenmatrix  $\mathbf{X}$  sind die transponierten Merkmalsvektoren:

$$\mathbf{X}_{1\bullet} = x_1^T, \mathbf{X}_{2\bullet} = x_2^T, \dots, \mathbf{X}_{N\bullet} = x_N^T$$

Die Spalten (also die transponierten Stichprobenvektoren) können wir mit  $\mathbf{X}_{\bullet 1}, \dots, \mathbf{X}_{\bullet D}$  bezeichnen. Ziehen wir von diesen deren arithmetische Mittelwerte ab, so erhalten wir die **mittelwertzentrierte Datenmatrix**:

$$\mathbf{X}_{\bullet d} \mapsto \mathbf{X}_{\bullet d} - \overline{\mathbf{X}_{\bullet d}}$$

für alle  $d \in \{1, \dots, D\}$ .

Teilen wir auch noch durch die jeweilige Standardabweichung, so erhalten wir die **standardisierte Datenmatrix** bestehend aus den sogenannten  **$z$ -Werten**:

$$\mathbf{X}_{\bullet d} \mapsto \mathbf{Z}_{\bullet d} = \frac{\mathbf{X}_{\bullet d} - \overline{\mathbf{X}_{\bullet d}}}{s(\mathbf{X}_{\bullet d})}$$

Falls  $s(\mathbf{X}_{\bullet d}) = 0$  gilt, kann z. B.  $\mathbf{Z}_{\bullet d} = 0$  gesetzt werden. Nach Konstruktion sind arithmetischer Mittelwert eines mittelwertzentrierten oder standardisierten Stichprobenvektors stets gleich null. Die empirische Varianz eines  $z$ -Werts ist zudem stets gleich eins (wenn nicht gerade  $s(\mathbf{X}_{\bullet d}) = 0$  gilt).

Eine Standardisierung ist insbesondere dann angezeigt, wenn die Merkmale verschiedene Einheiten oder Skalen haben. Beispielsweise wird die obige Datenmatrix aus Körpergröße und Gewicht (unter Verwendung der korrigierten Varianz) auf diese Weise zu:

$$\mathbf{Z} = \begin{pmatrix} -0,05 & -1,00 \\ 0,17 & 1,00 \\ -1,27 & -0,69 \\ 1,16 & 0,69 \end{pmatrix}$$

Für einige Anwendungen ist es praktisch, eine „nullte“ Spalte hinzuzufügen, deren Einträge alle eins sind:

$$\mathbf{X}_{\bullet 0} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \left. \right\} N\text{-mal}$$

Wir sprechen in diesem Fall von der **erweiterten Datenmatrix**.

## 5.2 Abstands- und Ähnlichkeitsmaße

Für den Vergleich zweier Ausprägungen  $u, v \in \mathbb{R}$  eines univariaten metrischen Merkmals ist der Absolutbetrag das wesentliche Maß, um einen numerischen Abstand zu ermitteln:  $\delta(u, v) = |u - v|$ . Bei einem ordinalen Merkmal kann die Rangdifferenz herangezogen werden. Bei einem kategorialen Merkmal mit den möglichen Ausprägungen  $m_1, \dots, m_K$  bleibt für einen Vergleich letztlich nur die diskrete Metrik:  $\delta(m_k, m_l) = 1$  falls  $k \neq l$  und  $\delta(m_k, m_l) = 0$  falls  $k = l$ .

Wollen wir im multivariaten Fall zwei Beobachtungen anhand ihrer Merkmalslisten bzw. -vektoren zahlenmäßig miteinander vergleichen, so gibt es eine ganze Reihe sinnvoller Maßzahlen, die als Abstand oder Ähnlichkeit interpretiert werden können.

### 5.2.1 Metrische Abstands- und Ähnlichkeitsmaße

Zunächst definieren wir verschiedene Maße für die Länge eines einzelnen Merkmalsvektors. Ein solches Maß für Vektorlänge wird im Allgemeinen als **Norm** bezeichnet.

Für einen metrischen Merkmalsvektor  $u = (u_1, \dots, u_D)^T \in \mathbb{R}^D$  und ein  $p \in \mathbb{N}$ ,  $p \geq 1$ , definieren wir dessen **Minkowski-Norm der Ordnung  $p$**  (kurz:  **$p$ -Norm**):

$$\|u\|_p = \left( \sum_{d=1}^D |u_d|^p \right)^{\frac{1}{p}},$$

weiterhin die **Maximumsnorm**:

$$\|u\|_\infty = \max_{d \in \{1, \dots, D\}} |u_d|$$

Die Maximumsnorm ist der Grenzfall der  $p$ -Norm für große Werte von  $p$ , für alle  $u \in \mathbb{R}^D$  gilt nämlich:

$$\lim_{p \rightarrow \infty} \|u\|_p = \|u\|_\infty$$

Die 2-Norm ist die gewöhnliche euklidische Norm; wenn kein spezifischer  $p$ -Index angegeben und sonst keine andere Angabe gemacht wird, meinen wir diese:  $\|u\| = \|u\|_2$ . Die euklidische Norm hängt wiederum mit dem **euklidischen Skalarprodukt**

$$\langle u, v \rangle = \sum_{d=1}^D u_d \cdot v_d$$

wie folgt zusammen:

$$\|u\| = \sqrt{\langle u, u \rangle}$$

Für zwei Spaltenvektoren  $u, v \in \mathbb{R}^D$  kann das euklidische Skalarprodukt wie folgt als eine Matrixmultiplikation geschrieben werden:

$$\langle u, v \rangle = u^T \cdot v$$

Der Abstand zweier Vektoren kann als die Länge von deren Differenz erklärt werden.

Für zwei metrische Merkmalsvektoren  $u, v \in \mathbb{R}^D$  definieren wir deren **euklidischen Abstand**

$$\delta_2(u, v) = \|u - v\|_2,$$

den **Manhattan-Abstand**

$$\delta_1(u, v) = \|u - v\|_1$$

sowie den **Tschebyscheff-Abstand**

$$\delta_\infty(u, v) = \|u - v\|_\infty.$$

Abstandsmaße können für jede Minkowski-Norm auf diese Weise definiert werden, die obige Liste gibt lediglich die in der Datenanalyse am häufigsten verwendeten wieder. Wenn in diesem Band keine andere Angabe gemacht wird,

kann davon ausgegangen werden, dass metrische Merkmalsvektoren über den euklidischen Abstand miteinander verglichen werden.

Als Rechenbeispiel betrachten wir die Vektoren:

$$u = \begin{pmatrix} -0,5 \\ 0,5 \end{pmatrix}, v = \begin{pmatrix} 1,0 \\ 1,0 \end{pmatrix}$$

Für die verschiedenen Abstandsmaße ergeben sich die folgenden Werte:

$$\delta_1(u, v) = |-0,5 - 1,0| + |0,5 - 1,0| = 2,0$$

$$\delta_2(u, v) = \sqrt{(-0,5 - 1,0)^2 + (0,5 - 1,0)^2} \approx 1,6$$

$$\delta_\infty(u, v) = \max\{|-0,5 - 1,0|, |0,5 - 1,0|\} = 1,5$$

**Anwendungsbeispiel.** Als ein konkretes Beispiel betrachten wir drei Pflanzen der Gattung der Schwertlilien aus einem größeren Datensatz von insgesamt 150 Exemplaren [1, 2]:

<b><i>n</i></b>	<b>Art</b>	<b>Länge Kelchblatt</b>	<b>Breite Kelchblatt</b>	<b>Länge Kronblatt</b>	<b>Breite Kronblatt</b>
1	<i>Iris setosa</i>	5,1 cm	3,5 cm	1,4 cm	0,2 cm
2	<i>Iris setosa</i>	4,9 cm	3,0 cm	1,4 cm	0,2 cm
101	<i>Iris virginica</i>	6,3 cm	3,3 cm	6,0 cm	2,5 cm

**Tabelle 5.1.** Maße von Kelch- und Kronblatt von Schwertlilien

Die Abmessungen von Kelch- und Kronblatt können wir zu Merkmalsvektoren  $x_1, x_2, x_{101} \in \mathbb{R}^4$  zusammenfassen. Der Tschebyscheff-Abstand zwischen den Exemplaren der Art *Iris setosa* beträgt  $\delta_\infty(x_1, x_2) = 0,5$  cm. Der Abstand dieser Exemplare zum Exemplar der Art *Iris virginica* ist wesentlich größer und beträgt jeweils  $\delta_\infty(x_1, x_{101}) = \delta_\infty(x_2, x_{101}) = 4,6$  cm aufgrund des deutlich längeren Kronblatts.

Sind die Komponenten der verglichenen Vektoren mit Einheiten behaftet, so müssen diese identisch sein. In obigem Beispiel wurden beispielsweise Zentimeter gewählt. In vielen Fällen ist eine vorherige Standardisierung der Datenmatrix bzw. Merkmale anzuraten, sodass die Ausprägungen einheitlos sind und dieselbe Streuung aufweisen. Bei gänzlich verschiedenen Dimensionen wie z. B. Länge gegenüber Gewicht ist eine Skalierung in jedem Fall für einen sinnvollen Vergleich erforderlich.

Der zum Abstand duale Begriff ist die Ähnlichkeit: Wir können uns vorstellen, dass Entitäten mit Merkmalsvektoren von geringem Abstand eine große Ähnlichkeit aufweisen.

Für zwei metrische Merkmalsvektoren mit nichtnegativen Werten  $u, v \in [0, \infty[^D$  definieren wir die **Kosinusähnlichkeit**

$$\sigma_{\cos}(u, v) = \frac{\langle u, v \rangle}{\|u\| \cdot \|v\|} = \cos \sphericalangle(u, v),$$

falls nicht gerade  $u = 0$  oder  $v = 0$ ; in diesem Fall kann  $\sigma_{\cos}(u, v) = 0$  gesetzt werden.

Die **Tanimoto-Ähnlichkeit** ist wie folgt gegeben:

$$\sigma_{\text{Tanim}}(u, v) = \frac{\langle u, v \rangle}{\|u\|^2 + \|v\|^2 - \langle u, v \rangle},$$

sofern nicht gerade  $u = v = 0$  gilt; in diesem Fall ist die Größe nicht definiert.

Die Konventionen, mit denen die Lücken im Definitionsbereich geschlossen werden, sind nicht in Stein gemeißelt. Beispielsweise mag die Auffassung, dass  $\sigma_{\text{Tanim}}(0, 0) = 0$  gilt, auch vertreten werden.

Sind die miteinander verglichenen Vektoren orthogonal, verschwinden sowohl die Kosinus- als auch die Tanimoto-Ähnlichkeit. Sind sie kolinear, nimmt die Kosinus-Ähnlichkeit ihren maximal möglichen Wert von eins an: Es gilt genau dann  $\sigma_{\cos}(u, v) = 1$ , wenn  $v = \lambda u$  für ein  $\lambda > 0$  vorliegt. Maximale Tanimoto-Ähnlichkeit  $\sigma_{\text{Tanim}}(u, v) = 1$  ist hingegen dann und nur dann gegeben, wenn  $u = v$  gilt.

**Anwendungsbeispiel.** Die Abmessungen von Kelch- und Kronblatt von Schwerlilien aus obigem Beispiel können wir auch über Ähnlichkeitsmaße miteinander vergleichen. Zwischen den Exemplaren der Art *Iris setosa* beträgt die Tanimoto-Ähnlichkeit  $\sigma_{\text{Tanim}}(x_1, x_2) \approx 0,99$ . Die Ähnlichkeit mit dem Exemplar der Art *Iris virginica* ist deutlich geringer und beträgt  $\sigma_{\text{Tanim}}(x_1, x_{101}) \approx 0,65$  bzw.  $\sigma_{\text{Tanim}}(x_2, x_{101}) \approx 0,64$ .

## 5.2.2 Kategoriale und binäre Abstands- und Ähnlichkeitsmaße

Angenommen, wir haben zwei gleich lange, geordnete Listen von Symbolen oder anderer Objekte gegeben. Ein naheliegender Gedanke, ein Maß für deren Abstand oder Ähnlichkeit zu ermitteln, besteht in einem paarweisen Vergleich und Zählung der verschiedenen bzw. identischen Symbole bzw. Objekte. So hätten die beiden Zeichenketten ABCDE und ABXDY etwa einen Abstand von zwei, da zwei Zeichen nicht miteinander übereinstimmen. Siehe hierzu auch die im Abschn. 1.4.5 für Zwecke der Datendeduplikation vorgestellten Editierabstandsmaße.

Im Kontext der Statistik ergibt sich aus dieser Überlegung die folgende formale Definition.

Für zwei Listen  $u, v$  der Länge  $D$  von Ausprägungen derselben kategorialen Merkmale definieren wir den **Hamming-Abstand**:

$$\delta_{\text{Hamm}}(u, v) = |\{d \in \{1, \dots, D\} | u_d \neq v_d\}|$$

Der **normierte Hamming-Abstand** ist durch  $\frac{1}{D} \cdot \delta_{\text{Hamm}}(u, v)$  gegeben.

Der Hamming-Abstand gibt einfach die Anzahl der Merkmale wieder, bei denen unterschiedliche Ausprägungen vorliegen. Ein Alltagsbeispiel würde der Vergleich zweier Pflanzen anhand kategorialer Bestimmungsmerkmale darstellen. Zwei Pflanzen mit identischen Merkmalsausprägungen (z. B. Wuchsform, Blattform, Zahl der Blütenblätter usw.) haben einen Hamming-Abstand von null und würden als derselben Art zugehörig vermutet werden.

Im Falle von binären Merkmalen gibt der Hamming-Abstand die Anzahl unterschiedlicher „Bits“ wieder: Seien  $x_1 = (1, 1, 0, 1)^T$  und  $x_2 = (0, 1, 0, 1)^T$  zwei binäre Merkmalsvektoren. Für einen leichteren Vergleich können wir diese auch übereinander schreiben und zu einer Datenmatrix zusammenfassen:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

Nur die ersten Einträge sind verschieden, also gilt  $\delta_{\text{Hamm}}(x_1, x_2) = 1$ .

Weitere Kennzahlen zum Vergleich von Listen binärer Merkmale sind die folgenden.

Für zwei Listen binärer Merkmalsausprägungen  $u, v \in \{0, 1\}^D$  definieren wir den **Jaccard-Koeffizienten**

$$\sigma_{\text{Jacc}}(u, v) = \frac{|\{d | u_d = 1 \text{ und } v_d = 1\}|}{|\{d | u_d = 1 \text{ oder } v_d = 1\}|}$$

sowie den **Szymkiewicz-Simpson-Koeffizienten** oder **Überlappungskoeffizienten**

$$\sigma_{\text{overlap}}(u, v) = \frac{|\{d | u_d = 1 \text{ und } v_d = 1\}|}{\min\{|\{d | u_d = 1\}|, |\{d | v_d = 1\}|\}}.$$

Dabei setzen wir voraus, dass nicht gerade  $u = v = 0$  gilt; in diesem Fall sind die obigen Ähnlichkeitsmaße nicht definiert.

Den Jaccard-Koeffizient haben wir bereits auf binäre Stichprobenlisten angewendet und als Assoziationsparameter zwischen zwei Merkmalen oder Teilsammlungen interpretiert (siehe Abschn. 2.5.3). Hier wenden wir die Definition auf Merkmalslisten an und interpretieren sie als Maß für die Ähnlichkeit zweier Informationsobjekte. Auch in diesem Fall gibt es eine mengentheoretische Interpretation. Anstelle der binären Vektoren  $u, v$  können wir auch die Mengen  $U$  und  $V$  der Merkmale mit jeweils positiven Einträgen angeben. Dann bemessen

$$\sigma_{\text{Jacc}}(U, V) = \frac{|U \cap V|}{|U \cup V|}$$

und

$$\sigma_{\text{overlap}}(U, V) = \frac{|U \cap V|}{\min\{|U|, |V|\}}$$

deren Grad an Überschneidung. Beide Maße nehmen Werte zwischen null und eins an, wobei  $\sigma_{\text{Jacc}}(U, V) = \sigma_{\text{overlap}}(U, V) = 0$  einer verschwindenden Überschneidung entspricht:  $U \cap V = \emptyset$ . Die Maximalwerte sind wie folgt charakterisiert:

$$\begin{aligned}\sigma_{\text{Jacc}}(U, V) &= 1 \Leftrightarrow U = V, \\ \sigma_{\text{overlap}}(U, V) &= 1 \Leftrightarrow U \subseteq V \text{ oder } U \supseteq V\end{aligned}$$

Eine in der Praxis wichtige alternative Schreibweise für den Jaccard-Koeffizienten ist die folgende:

$$\sigma_{\text{Jacc}}(U, V) = \frac{|U \cap V|}{|U| + |V| - |U \cap V|}$$

Als Rechenbeispiel betrachten wir die zwei binären Merkmalsvektoren  $u = (1, 1, 0, 1)^T$  und  $v = (0, 1, 0, 1)^T$ . Bezeichnen wir die zu jedem Eintrag von  $u$  und  $v$  gehörigen Merkmale mit  $a$  bis  $d$ , so entsprechen diese den Mengen  $U = \{a, b, d\}$  und  $V = \{b, d\}$ . Die Ähnlichkeitsmaße berechnen sich wie folgt:

$$\begin{aligned}\sigma_{\text{Jacc}}(u, v) &= \sigma_{\text{Jacc}}(U, V) = \frac{2}{3} \approx 0,67, \\ \sigma_{\text{overlap}}(u, v) &= \sigma_{\text{overlap}}(U, V) = \frac{2}{\min\{3, 2\}} = 1,00\end{aligned}$$

### 5.2.3 Abstands- und Ähnlichkeitsmatrizen

Nachdem wir im letzten Abschnitt einige Beispiele für Abstands- und Ähnlichkeitsmaße kennengelernt haben, wollen wir uns einer allgemeinen Definition nähern.

Sei  $\Omega$  eine beliebige Menge, z. B. ein metrischer oder binärer Merkmalsraum  $\Omega \subseteq \mathbb{R}^D$  oder  $\Omega = \{0, 1\}^D$ . Für eine Abbildung  $\delta: \Omega \times \Omega \rightarrow [0, \infty[$  betrachten wir die folgenden Bedingungen, welche für alle  $u, v, w \in \Omega$  gelten sollen:

$$(1) \quad \delta(u, u) = 0$$

$$(2) \quad \delta(u, v) = 0 \Rightarrow u = v$$

$$\text{Symmetrie:} \quad \delta(u, v) = \delta(v, u)$$

$$\text{Dreiecksungleichung:} \quad \delta(u, w) \leq \delta(u, v) + \delta(v, w)$$

Ist wenigstens Bedingung (1) erfüllt, so wollen wir  $\delta(\cdot, \cdot)$  eine **Prämetrik**<sup>1</sup> nennen. Sind alle Bedingungen erfüllt, so wird die Abbildung eine **Metrik** genannt.

Sämtliche durch die Minkowski-Normen  $\|\cdot\|_p$  induzierten Abstandsmaße sind Metriken, ebenso der Hamming-Abstand.

Aus der Jaccard-Ähnlichkeit kann der **Jaccard-Abstand** wie folgt konstruiert werden:

$$\delta_{\text{Jacc}}(u, v) := 1 - \sigma_{\text{Jacc}}(u, v)$$

für alle  $u, v \in \{0, 1\}^D$ . Es kann gezeigt werden, dass dieses Abstandsmaß eine Metrik ist.

Ein Beispiel für eine symmetrische Prämetrik, die jedoch keine Metrik ist (weder die Dreiecksungleichung noch Bedingung (2) sind erfüllt):

$$\delta_{\text{overlap}}(u, v) := 1 - \sigma_{\text{overlap}}(u, v)$$

für alle  $u, v \in \{0, 1\}^D$ .

Viele in der Datenwissenschaft gebräuchlichen Abstandsmaße sind wenigstens symmetrische Prämetriken, und Ähnlichkeitsmaße können oft auf solche zurückgeführt werden. Daraus ergeben sich analog die folgenden Bedingungen, die alle oder zum Teil sinnvollerweise an Ähnlichkeitsmaße gestellt werden können:

$$(1) \quad \sigma(u, u) = \sigma(v, v)$$

$$(2) \quad \sigma(u, v) < \sigma(v, v) \Leftrightarrow u \neq v$$

$$\text{Symmetrie:} \quad \sigma(u, v) = \sigma(v, u)$$

für alle  $u, v \in \Omega$ . Ist  $\delta(\cdot, \cdot)$  eine Metrik, so werden alle obigen Eigenschaften insbesondere durch das folgende mithilfe dieser Metrik konstruierte Ähnlichkeitsmaß erfüllt:

$$\sigma(u, v) := a^2 \cdot e^{-\frac{1}{2} \left( \frac{\delta(u, v)}{h} \right)^q}$$

mit reellen Zahlen  $a, h, q > 0$ ; oft wird  $q = 1$  oder  $q = 2$  gewählt.

Mitunter wird auch noch die folgende multiplikative Dreiecksungleichung als wünschenswerte Eigenschaft genannt:

$$\sigma(u, w) \geq \sigma(u, v) \cdot \sigma(v, w)$$

Sei  $x_1, \dots, x_N$  eine Stichprobe von Merkmalslisten oder -vektoren. Für ein fest gewähltes Abstandsmaß  $\delta(\cdot, \cdot)$  definieren wir die **Abstandsmatrix**:

---

<sup>1</sup> Das ist keine standardisierte Definition, manche Autoren mögen etwas anderes unter einer „Prämetrik“ verstehen.

$$\Delta(x) = \begin{pmatrix} \delta(x_1, x_1) & \delta(x_1, x_2) & \cdots & \delta(x_1, x_N) \\ \delta(x_2, x_1) & \delta(x_2, x_2) & \cdots & \delta(x_2, x_N) \\ \vdots & \vdots & & \vdots \\ \delta(x_N, x_1) & \delta(x_N, x_2) & \cdots & \delta(x_N, x_N) \end{pmatrix}$$

Für ein Ähnlichkeitsmaß  $\sigma(\cdot, \cdot)$  wird analog die Ähnlichkeitsmatrix  $\Sigma(x)$  definiert.

Die Abstands- bzw. Ähnlichkeitsmatrix besteht also gerade aus den paarweise berechneten Abständen bzw. Ähnlichkeiten zwischen den Beobachtungen.

Typischerweise ist das Abstandsmaß wenigstens eine symmetrische Prämetrik, in diesem Fall ist die Abstandsmatrix symmetrisch und hat verschwindende Diagonaleinträge.

Als Rechenbeispiel ziehen wir die paarweisen Distanzen (in Kilometern gemessen) der deutschen Städte Berlin, Hamburg, München, Köln und Frankfurt heran. Diese können in einer Abstandsmatrix zusammengefasst werden:

$$\Delta((\text{Berlin}, \text{Hamburg}, \text{München}, \text{Köln}, \text{Frankfurt})) = \begin{pmatrix} 0 & 282 & 508 & 534 & 459 \\ 282 & 0 & 615 & 377 & 396 \\ 508 & 615 & 0 & 470 & 312 \\ 534 & 377 & 470 & 0 & 163 \\ 459 & 396 & 312 & 163 & 0 \end{pmatrix}$$

Als weiteres Beispiel betrachten wir die standardisierte Datenmatrix aus Körpergröße und -gewicht aus Abschn. 5.1, die sich aus vier Merkmalsvektoren zusammensetzt:

$$z_1 = \begin{pmatrix} -0,05 \\ -1,00 \end{pmatrix}, z_2 = \begin{pmatrix} 0,17 \\ 1,00 \end{pmatrix}, z_3 = \begin{pmatrix} -1,27 \\ -0,69 \end{pmatrix}, z_4 = \begin{pmatrix} 1,16 \\ 0,69 \end{pmatrix}$$

Die euklidische Abstandsmatrix ist wie folgt gegeben:

$$\Delta(z) = \begin{pmatrix} 0 & 2,01 & 1,26 & 2,08 \\ 2,01 & 0 & 2,22 & 1,04 \\ 1,26 & 2,22 & 0 & 2,79 \\ 2,08 & 1,04 & 2,79 & 0 \end{pmatrix}$$

Schließlich können wir mit Ähnlichkeitsmaßen etwa auch Kochrezepte durch das Vorkommen an Zutaten charakterisieren. Jede Zutat definiert ein binäres Merkmal (die Zutat wird benötigt, oder sie wird nicht benötigt). Beispiele sind die folgenden:

- $U_1 = \text{Spaghetti aglio e olio} = \{\text{Chili, Knoblauch, Olivenöl, Spaghetti}\},$
- $U_2 = \text{Curry} = \{\text{Brokkoli, Chili, Currysauce, Knoblauch, Kokosmilch, Tofu}\},$

$U_3 = \text{Spaghetti Napoli} = \{\text{Basilikum, Olivenöl, Spaghetti, Tomaten, Zwiebeln}\},$

$U_4 = \text{Ratatouille} = \{\text{Auberginen, Basilikum, Knoblauch, Paprika, Olivenöl, Rosmarin, Tomaten, Zucchini, Zwiebeln}\}$

Die paarweise Berechnung von Jaccard- bzw. Überlappungs-Koeffizienten für diese Listen führt auf folgende Ähnlichkeitsmatrizen:

$$\Sigma_{\text{Jacc}}(U) = \begin{pmatrix} 1 & 0,25 & 0,29 & 0,18 \\ 0,25 & 1 & 0 & 0,07 \\ 0,29 & 0 & 1 & 0,40 \\ 0,18 & 0,07 & 0,40 & 1 \end{pmatrix},$$

$$\Sigma_{\text{overlap}}(U) = \begin{pmatrix} 1 & 0,5 & 0,5 & 0,5 \\ 0,5 & 1 & 0 & 0,17 \\ 0,5 & 0 & 1 & 0,8 \\ 0,5 & 0,17 & 0,8 & 1 \end{pmatrix}$$

### 5.3 Multivariate Lage- und Streuungsparameter

Wir betrachten eine Stichprobe univariater Werte  $x_1, \dots, x_N \in \mathbb{R}$ . Wir wollen jene Stelle  $\hat{u} \in \mathbb{R}$  finden, welche die Summe der quadrierten Abstände zu ihr minimiert, also die Funktion:

$$\ell_2: \mathbb{R} \rightarrow [0, \infty[, \ell_2(u) = \sum_{n=1}^N (u - x_n)^2$$

Die Ableitung dieser Funktion ist wie folgt gegeben:

$$\frac{d}{du} \ell_2(u) = \sum_{n=1}^N 2(u - x_n) = 2N \cdot u - 2 \sum_{n=1}^N x_n$$

Die zweite Ableitung ist positiv, daher liegt die eindeutig bestimmte Minimalstelle bei der Nullstelle der Ableitung:  $\hat{u} = \frac{1}{N} \sum_{n=1}^N x_n$ . Das ist gerade der arithmetische Mittelwert. Ebenso kann gezeigt werden, dass der empirische Median eine Minimalstelle der Summe der Abstände ist, ohne dass diese quadriert werden, also der folgenden Funktion:

$$\ell_1: \mathbb{R} \rightarrow [0, \infty[, \ell_1(u) = \sum_{n=1}^N |u - x_n|$$

Der arithmetische Mittelwert und der Median können zu multivariaten Lageparametern verallgemeinert werden, indem statt des Abstands entlang der Zahlengeraden andere Abstandsmaße verwendet werden, etwa der euklidische Abstand.

Bei der multivariaten Betrachtung der Streuung von Datenpunkten ist darauf zu achten, dass diese entlang verschiedener Richtungen des Merkmalsraums unterschiedlich stark ausgeprägt sein kann.

### 5.3.1 Geometrischer Schwerpunkt und Median, Medoid

Der geometrische Schwerpunkt ist das multivariate Analogon zum arithmetischen Mittel.

Für eine Reihe von Merkmalsvektoren  $x_1, \dots, x_N$  mit  $x_n \in \mathbb{R}^D$  für alle  $n \in \{1, \dots, N\}$  ist der **geometrische Schwerpunkt** wie folgt definiert:

$$\bar{x}_{\text{centroid}} = \frac{1}{N} \sum_{n=1}^N x_n$$

Es ist nicht schwer zu zeigen, dass der geometrische Schwerpunkt gerade die Summe der quadrierten euklidischen Abstände zu den Merkmalsvektoren minimiert; er ist die eindeutig bestimmte Minimalstelle der Funktion:

$$\ell_2: \mathbb{R}^D \rightarrow [0, \infty[, \ell_2(u) = \sum_{n=1}^N \|u - x_n\|^2$$

Weiterhin kann der Schwerpunkt als das spaltenweise berechnete arithmetische Mittel der zugehörigen Datenmatrix charakterisiert werden:

$$\bar{x}_{\text{centroid}} = \begin{pmatrix} \overline{\mathbf{X}_{\bullet 1}} \\ \overline{\mathbf{X}_{\bullet 2}} \\ \vdots \\ \overline{\mathbf{X}_{\bullet D}} \end{pmatrix}$$

Anders ausgedrückt: Der Schwerpunkt ist der Vektor, der sich aus den arithmetischen Mittelwerten der einzelnen univariaten Merkmale zusammensetzt. Insbesondere sieht man, dass eine Mittelwertzentrierung gleichbedeutend mit der Operation ist, von jedem Datenpunkt den Schwerpunkt abzuziehen. Der Schwerpunkt der mittelwertzentrierten Datenpunkte ist daher stets der Koordinatenursprung.

Wie beim arithmetischen Mittel werden im Folgenden auch alternative Schreibweisen gebraucht:  $\mu(x) = \bar{x} = \bar{x}_{\text{centroid}}$ .

Für eine Reihe von Merkmalsvektoren  $x_1, \dots, x_N$  mit  $x_n \in \mathbb{R}^D$  für alle  $n \in \{1, \dots, N\}$  ist ein **geometrischer Median** jeder Punkt  $\bar{x}_{\text{median}} \in \mathbb{R}^D$ , der eine Minimalstelle der folgenden Funktion ist:

$$\ell_1: \mathbb{R}^D \rightarrow [0, \infty[, \ell_1(u) = \sum_{n=1}^N \|u - x_n\|$$

Ein geometrischer Median minimiert also gerade die Summe der Abstände zu den Datenpunkten in der Stichprobe. Es kann gezeigt werden: Sind die Merk-

maßvektoren  $x_1, \dots, x_N$  linear unabhängig, so ist deren geometrischer Median eindeutig bestimmt.

Im Gegensatz zum Schwerpunkt gibt es für den geometrischen Median keine geschlossene Formel für dessen Berechnung; er ist mithin auch *nicht* durch den komponentenweise ermittelten einfachen Median gegeben. Numerisch kann der geometrische Median mithilfe des **Weiszfeld'schen Algorithmus** berechnet werden [3, 4]. Dieser wendet die folgende Iteration an:

$$y_{k+1} = \left( \sum_{n=1}^N \frac{1}{\|x_n - y_k\|} \right)^{-1} \cdot \sum_{n=1}^N \frac{x_n}{\|x_n - y_k\|}$$

mit einem geeigneten Startwert  $y_0$ , beispielsweise  $y_0 = \bar{x}_{\text{centroid}}$ . Wenn der geometrische Median eindeutig bestimmbar ist, und wenn eines der Zwischenergebnisse  $y_k$  nicht gerade auf einem der Datenpunkte  $x_n$  zu liegen kommt, dann konvergiert die Folge der  $y_k$  gegen  $\bar{x}_{\text{median}}$ .

**Anwendungsbeispiel.** Abb. 5.1 zeigt oben das Streudiagramm der Breite und Länge des Kronblattes von Schwertliliengewächsen, zusammen mit geometrischem Median und Schwerpunkt der Datenpunkte.

Der geometrische Median bzw. Schwerpunkt minimiert die Summe der euklidischen Abstände bzw. der quadrierten euklidischen Abstände zu den Datenpunkten  $x_1, \dots, x_N$ . Denken wir diese Charakterisierung für allgemeine Metriken  $\delta(\cdot, \cdot)$  weiter, so führt dies auf den verallgemeinerten geometrischen Median bzw. den **Fréchet'schen Mittelwert**. Diese Kennzahlen sind bei gegebenem Merkmalsraum  $\Omega$  durch Minimalstellen der Funktionen

$$\ell_\alpha: \Omega \rightarrow [0, \infty[, \ell(u) = \sum_{n=1}^N (\delta(u, x_n))^\alpha$$

mit  $\alpha = 1$  bzw.  $\alpha = 2$  gegeben [5, 6].

Eine Berechnung dieser Minimalstellen ist im Allgemeinen nicht einfach. Eine ähnliche Idee liegt dem folgenden Lagemaß zugrunde, das einfacher zu berechnen ist.

Seien Datenpunkte oder Merkmalslisten  $x_1, \dots, x_N$  gegeben, für die eine symmetrische Prämetrik  $\delta(\cdot, \cdot)$  definiert ist. Ein **Medoid** ist ein Datenpunkt  $\bar{x}_{\text{medoid}} \in \{x_1, \dots, x_N\}$ , der die Summe der Abstände zu den übrigen Datenpunkten minimiert. Er ist also ein Minimum der folgenden Funktion:

$$f: \{x_1, \dots, x_N\} \rightarrow [0, \infty[, f(u) = \sum_{n=1}^N \delta(u, x_n)$$

Im Unterschied zu den zuvor vorgestellten Lagemaßen wird der Medoid stets aus der Menge der vorliegenden Datenpunkte ausgewählt.

Über die Distanzmatrix  $\Delta$  ausgedrückt entspricht ein Medoid  $x_i$  gerade einem Index  $i \in \{1, \dots, N\}$  mit minimaler Zeilen- oder Spaltensumme:

$$\Delta_{\bullet i} = \sum_{n=1}^N \Delta_{ni} = \sum_{n=1}^N \Delta_{in} = \Delta_{i\bullet}$$

### 5.3.2 Empirische Kovarianz- und Korrelationsmatrix

Eine weitere naheliegende Operation besteht darin, Assoziationsmaße zwischen allen Merkmalen zu berechnen und in einer Matrix zusammenzufassen.

Seien  $x_1, \dots, x_N \in \mathbb{R}^D$  Merkmalsvektoren und  $\mathbf{X}$  die zugehörige Datenmatrix. Die **empirische Kovarianzmatrix** ist durch die paarweise berechneten Kovarianzen von deren Spalten gegeben:

$$S(x) = \begin{pmatrix} s(\mathbf{X}_{\bullet 1}, \mathbf{X}_{\bullet 1}) & s(\mathbf{X}_{\bullet 1}, \mathbf{X}_{\bullet 2}) & \cdots & s(\mathbf{X}_{\bullet 1}, \mathbf{X}_{\bullet D}) \\ s(\mathbf{X}_{\bullet 2}, \mathbf{X}_{\bullet 1}) & s(\mathbf{X}_{\bullet 2}, \mathbf{X}_{\bullet 2}) & \cdots & s(\mathbf{X}_{\bullet 2}, \mathbf{X}_{\bullet D}) \\ \vdots & \vdots & & \vdots \\ s(\mathbf{X}_{\bullet D}, \mathbf{X}_{\bullet 1}) & s(\mathbf{X}_{\bullet D}, \mathbf{X}_{\bullet 2}) & \cdots & s(\mathbf{X}_{\bullet D}, \mathbf{X}_{\bullet D}) \end{pmatrix}$$

Analog besteht die **empirische Korrelationsmatrix**  $R(x)$  aus den paarweise berechneten Korrelationskoeffizienten nach Bravais-Pearson.

Sowohl Kovarianz- als auch Korrelationsmatrix sind symmetrische Matrizen. Ist die Datenmatrix  $\mathbf{X}$  mittelwertzentriert, so kann die Kovarianzmatrix recht kompakt als ein Matrixprodukt ausgedrückt werden:

$$S(x) = \frac{1}{N} \mathbf{X}^T \cdot \mathbf{X}$$

bzw.  $S_{\text{kor}}(x) = \frac{1}{N-1} \mathbf{X}^T \cdot \mathbf{X}$ , falls eine Bessel-korrigierte Varianzschätzung zugrundegelegt werden soll. Ist die Datenmatrix zudem noch standardisiert, so dass alle Stichprobenvektoren die Varianz eins haben, so stimmen Kovarianzmatrix und Korrelationsmatrix überein.

**Anwendungsbeispiel.** Wir betrachten nochmals den Datensatz mit den Maßen des Kronblatts von Schwertlilien, diesmal nach vorheriger Standardisierung der Länge bzw. Breite auf die entsprechenden  $z$ -Werte. Die Varianz der so standardisierten Kronblattmaße ist eins, die Kovarianzmatrix bzw. Korrelationsmatrix sieht wie folgt aus:

$$R(x) = \begin{pmatrix} 1 & 0,96 \\ 0,96 & 1 \end{pmatrix}$$

Länge und Breite sind demnach stark korreliert. Abb. 5.1 zeigt unten das Streudiagramm zusammen mit der **Kovarianzellipse**. Diese hat als Mittelpunkt den Schwerpunkt, und die Halbachsen zeigen jeweils in die Richtungen kleinster bzw. größter Varianz. Die Längen der Halbachsen geben die entsprechende Standardabweichung wieder. Die Richtungen sind durch die Eigenvektoren der Kovarianzmatrix gegeben – weitere Details können im Abschn. 7.2.1 zum Thema „Hauptachsentransformation“ in Erfahrung gebracht werden.

## 5.4 Zufallsvektoren und -matrizen

So wie wir uns die Einträge eines Stichprobenvektors als Realisierungen einer Zufallsvariablen vorstellen können, so können die Merkmalsvektoren als Realisierungen eines **Zufallsvektors** aufgefasst werden. Zufallsvektoren sind Zufallsvariablen mit Werten in  $\mathbb{R}^D$ ; eine alternative Sichtweise von Zufallsvektoren ist als ein  $D$ -Tupel univariater Zufallsvariablen  $X_1, \dots, X_D$ . Analog können auch **Zufallsmatrizen** definiert werden: Dabei handelt es sich um Matrizen, deren Einträge Zufallsgrößen sind.

### 5.4.1 Erwartungswertvektor und Kovarianzmatrix

Den Begriff von Erwartungswert und Varianz einer Zufallsvariablen können wir auf Zufallsvektoren verallgemeinern, indem wir die entsprechenden Operationen komponentenweise anwenden.

Sei  $X = (X_1, \dots, X_D)^T$  ein Zufallsvektor. Dann ist der zugehörige **Erwartungswertvektor** wie folgt gegeben:

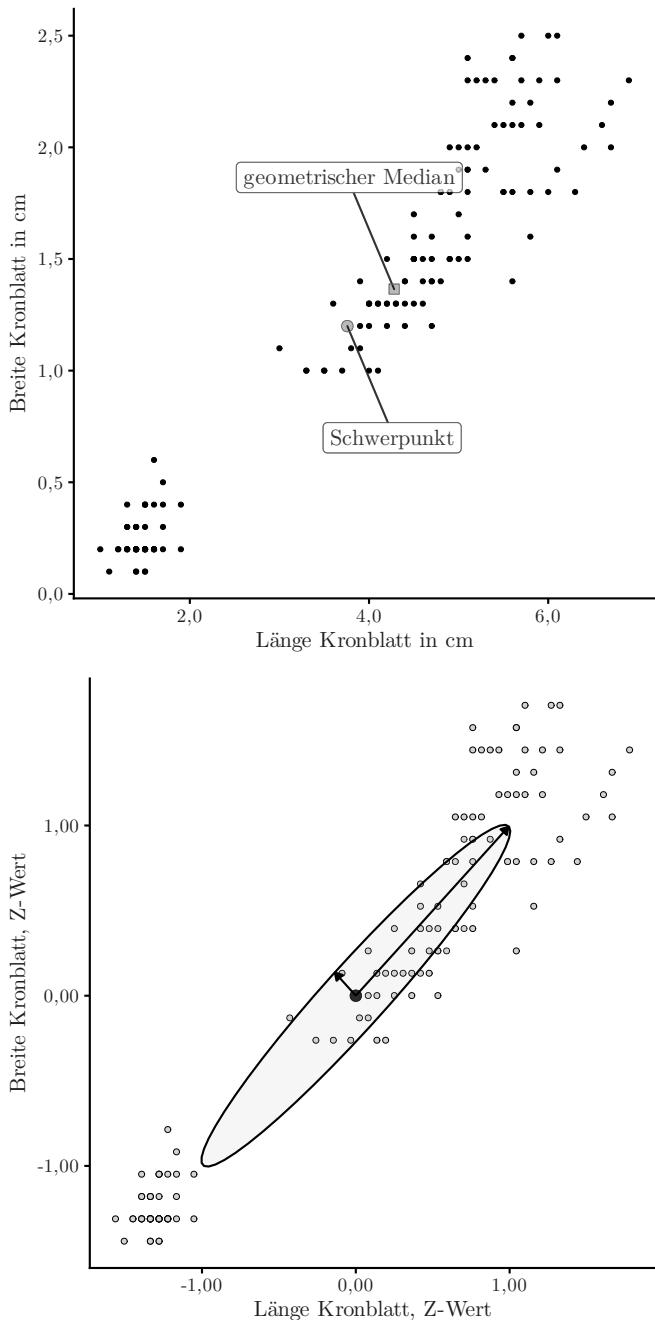
$$E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_D] \end{pmatrix}$$

Der Erwartungswert einer Zufallsmatrix wird analog definiert.

Die **Kovarianzmatrix** ist die symmetrische Matrix der paarweise berechneten Kovarianzen:

$$\Sigma[X] = \begin{pmatrix} \sigma[X_1, X_1] & \sigma[X_1, X_2] & \cdots & \sigma[X_1, X_D] \\ \sigma[X_2, X_1] & \sigma[X_2, X_2] & \cdots & \sigma[X_2, X_D] \\ \vdots & & & \vdots \\ \sigma[X_D, X_1] & \sigma[X_D, X_2] & \cdots & \sigma[X_D, X_D] \end{pmatrix}$$

Die Linearitätseigenschaft des Erwartungswerts überträgt sich sinngemäß, es gilt stets:



**Abb. 5.1.** Geometrischer Median und Schwerpunkt; Kovarianzellipse

$$E[A \cdot X + B \cdot Y + C] = A \cdot E[X] + B \cdot E[Y] + C$$

Dabei sind  $X, Y$  Zufallsvektoren mit  $D$  Einträgen,  $A, B$  Matrizen vom Format  $K \times D$  und  $C$  ein Spaltenvektor der Länge  $K$ .

Zur Erinnerung, zwischen Kovarianz und Varianz besteht der folgende Zusammenhang:  $\sigma[X_i, X_i] = \sigma^2[X_i]$ . Auf der Diagonalen der Kovarianzmatrix stehen daher gerade die Varianzen der Komponenten des Zufallsvektors.

Alternativ lässt sich die Kovarianzmatrix auch als Erwartungswert einer Zufallsmatrix schreiben, wobei wir uns den Zufallsvektor  $X$  unser Konvention gemäß als Spaltenvektor vorstellen müssen:

$$\Sigma[X] = E[(X - E[X]) \cdot (X - E[X])^T]$$

Eine quadratische  $D \times D$ -Matrix  $A$  ist positiv semidefinit, wenn sie – als lineare Abbildung aufgefasst – bei keinem Vektor dessen Richtung umkehrt, also für alle Vektoren  $v \in \mathbb{R}^D$  und Skalare  $\lambda \in \mathbb{R}$  gilt:

$$A \cdot v = \lambda v \Rightarrow \lambda \geq 0$$

Anders ausgedrückt: Die Matrix besitzt keine negativen Eigenwerte. Eine weitere, äquivalente Formulierung:

$$\langle v, A \cdot v \rangle \geq 0$$

für alle  $v \in \mathbb{R}^D$ .

Die Kovarianzmatrix ist stets positiv semidefinit. Es gilt nämlich mit der Abkürzung  $Y = X - E[X]$ :

$$\begin{aligned} \langle v, \Sigma[X] \cdot v \rangle &= v^T \cdot \Sigma[X] \cdot v = v^T \cdot E[Y \cdot Y^T] \cdot v \\ &= E[v^T \cdot Y \cdot Y^T \cdot v] = E[(v^T \cdot Y)^2] \geq 0 \end{aligned}$$

Dabei haben wir die Linearität und Monotonie des Erwartungswerts genutzt.

### 5.4.2 Multivariate Normalverteilung

Die **multivariate Dichtefunktion** eines Zufallsvektors  $X = (X_1, \dots, X_D)^T$ , dessen Komponenten stetige Zufallsvariablen darstellen, ist einfach durch die gemeinsame Dichtefunktion dieser Komponenten gegeben:

$$p_X : \mathbb{R}^D \rightarrow [0, \infty[, p_X(u) = p_{X_1, \dots, X_D}(u_1, \dots, u_D)$$

Eine der wichtigsten Familien von multivariaten Dichtefunktionen stellt die Verallgemeinerung der univariaten Normalverteilung dar.

Die  **$D$ -dimensionale Normalverteilung** ist die folgende (multivariate) Wahrscheinlichkeitsdichtefunktion:

$$\begin{aligned}\mathcal{N}(u|\mu, \Sigma) &= \mathcal{N}(u_1, \dots, u_D|\mu, \Sigma) \\ &= \frac{1}{\sqrt{(2\pi)^D \cdot \det(\Sigma)}} \cdot e^{-\frac{1}{2}(u-\mu)^T \cdot \Sigma^{-1} \cdot (u-\mu)}\end{aligned}$$

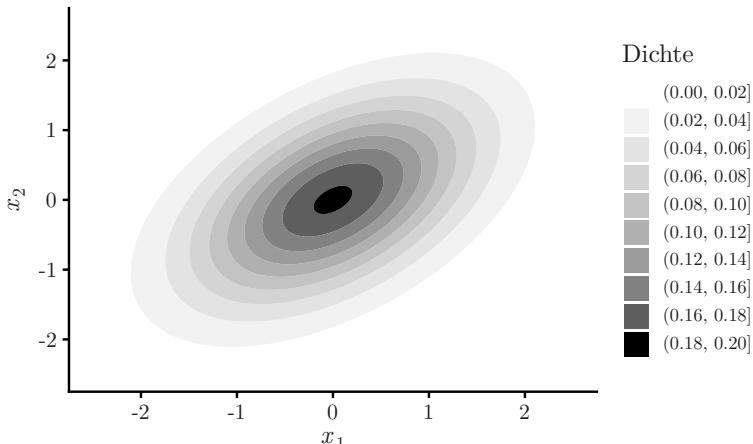
Dabei sind  $\mu \in \mathbb{R}^D$  ein Lage- bzw. Schwerpunktsvektor und  $\Sigma$  eine symmetrische, positiv definite  $D \times D$ -Matrix.

Für den Erwartungswertvektor und die Kovarianzmatrix eines normalverteilten Zufallsvektors  $X \sim \mathcal{N}(\cdot | \mu, \Sigma)$  gilt:

$$E[X] = \mu, \quad \Sigma[X] = \Sigma$$

Die folgende Abbildung zeigt zum Beispiel die Dichte einer bivariaten Normalverteilung mit dem Lagevektor bzw. der Kovarianzmatrix

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$



**Abb. 5.2.** Dichtefunktion einer bivariaten Normalverteilung

Die quadrierte euklidische Norm eines im Ursprung zentrierten  $D$ -dimensionalen normalverteilten Zufallsvektors  $X$ , dessen Kovarianzmatrix durch die Einheitsmatrix  $\Sigma = \text{diag}(1, \dots, 1)$  gegeben ist, folgt einer Chi-Quadrat-Verteilung mit  $D$  Freiheitsgraden (siehe Abschn. 3.5.1):

$$\|X\|^2 \sim \chi_D^2$$

Die folgenden Lehrsätze über Randverteilungen und bedingte Verteilungen können durch etwas längere Rechnungen gezeigt werden, siehe etwa [7, Kap. VIII, Abschn. 9]. Sei im Folgenden  $X = (X_1, \dots, X_D)^T$  stets ein normalverteilter Zufallsvektor mit  $X \sim \mathcal{N}(\cdot | \mu, \Sigma)$ .

**Verteilung der linearen Transformation eines normalverteilten Zufallsvektors.** Sei  $A$  eine Matrix vom Format  $K \times D$ ,  $K \leq D$ , von maximalem Rang  $K$ .

Dann ist der Zufallsvektor  $Y := A \cdot X$  ebenfalls normalverteilt, mit Varianz  $A \cdot \mu$  und Kovarianz  $A \cdot \Sigma \cdot A^T$ :

$$Y \sim \mathcal{N}(\cdot | A \cdot \mu, A \cdot \Sigma \cdot A^T)$$

Insbesondere kann die Formel für eine Linearkombination  $Y = \sum_{n=1}^N a_n X_n$  unabhängiger univariat normalverteilter Zufallsvariablen  $X_1, \dots, X_N$  mit Erwartungswerten  $\mu_1, \dots, \mu_N$  und Varianzen  $\sigma_1^2, \dots, \sigma_N^2$  abgeleitet werden. Dazu setzen wir

$$\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_N^2) \text{ und } A = (a_1, \dots, a_N)$$

und erhalten:

$$Y \sim \mathcal{N}(\cdot | \mu_Y, \sigma_Y^2) \text{ mit } \mu_Y = \sum_{n=1}^N a_n \mu_n \text{ und } \sigma_Y^2 = \sum_{n=1}^N (a_n \sigma_n)^2$$

Wir wollen nun die Randdichten betrachten. Hierzu stellen wir uns  $X$  in zwei Zufallsvektoren der Längen  $1 \leq K < D$  bzw.  $D - K$  aufgeteilt vor:

$$X^{(0)} = \begin{pmatrix} X_1 \\ \vdots \\ X_K \end{pmatrix} \text{ und } X^{(1)} = \begin{pmatrix} X_{K+1} \\ \vdots \\ X_D \end{pmatrix}$$

Ebenso verfahren wir mit dem Lagevektor:

$$\mu^{(0)} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix} \text{ und } \mu^{(1)} = \begin{pmatrix} \mu_{K+1} \\ \vdots \\ \mu_D \end{pmatrix}$$

Entsprechend bringen wir die Varianzmatrix in die folgende Blockform:

$$\Sigma = \begin{pmatrix} \Sigma^{(00)} & \Sigma^{(01)} \\ \Sigma^{(10)} & \Sigma^{(11)} \end{pmatrix}$$

Dabei entsteht die  $K \times K$ -Matrix  $\Sigma^{(00)}$  durch Streichen der letzten  $D - K$  Spalten und Zeilen von  $\Sigma$ , und  $\Sigma^{(11)}$  entsteht durch Streichen der ersten  $K$  Spalten und Zeilen.

**Randverteilungen eines normalverteilten Zufallsvektors.** Der Zufallsvektor  $X^{(0)}$  ist ebenfalls normalverteilt:

$$X^{(0)} \sim \mathcal{N}(\cdot | \mu^{(0)}, \Sigma^{(00)})$$

Entsprechendes gilt natürlich sinngemäß für  $X^{(1)}$  oder jede andere beliebige Auswahl an Komponenten. Insbesondere folgt jede Komponente  $X_d$  eines normalverteilten Zufallsvektors einer eindimensionalen Normalverteilung mit Mittelwert  $\mu_d$  und Varianz  $\Sigma_{dd}$ ,  $d \in \{1, \dots, D\}$ .

**Bedingte Verteilungen von Komponenten eines normalverteilten Zufallsvektors.** Sei  $x^{(0)} \in \mathbb{R}^K$ . Die bedingte Verteilung von  $X^{(1)}$  unter der Bedingung  $X^{(0)} = x^{(0)}$  ist eine Normalverteilung:

$$p_{X^{(1)}|X^{(0)}}(\cdot | x^{(0)}) = \mathcal{N}(\cdot | \mu^{(1|0)}, \Sigma^{(1|0)})$$

mit

$$\begin{aligned}\mu^{(1|0)} &= \mu^{(1)} + \Sigma^{(10)} \cdot (\Sigma^{(00)})^{-1} \cdot (x^{(0)} - \mu^{(0)}), \\ \Sigma^{(1|0)} &= \Sigma^{(11)} - \Sigma^{(10)} \cdot (\Sigma^{(00)})^{-1} \cdot \Sigma^{(01)}\end{aligned}$$

Wir kommen zurück auf das Beispiel einer bivariaten Normalverteilung und betrachten einen Zufallsvektor  $X = (X_1, X_2)^T \sim \mathcal{N}(\cdot | 0, \Sigma)$  mit

$$\Sigma = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

Die Randdichten sind Standardnormalverteilungen:  $X_1, X_2 \sim \mathcal{N}(\cdot | 0, 1)$ .

Die Verteilung von  $X_2$  unter der Bedingung  $X_1 = x_*$ ,  $x_* \in \mathbb{R}$ , ist eine Normalverteilung mit Lageparameter  $\mu_{2|1} = \frac{1}{2}x_*$  und Streuungsparameter  $\sigma_{2|1}^2 = \frac{3}{4}$ .

Wir können dieses Ergebnis wie folgt interpretieren: Die Zufallsvariablen  $X_1$  und  $X_2$  sind positiv korreliert, es gilt  $\sigma[X_1, X_2] = \Sigma_{12} = \frac{1}{2}$ . Bei Kenntnis der Bedingung  $X_1 = x_*$  führt dies zu einer Aktualisierung unserer Kenntnis vom Wert von  $X_2$ : Aufgrund der positiven Korrelation erwarten wir nun, dass der Wert nun nicht mehr in der Nähe von  $\mu_2 = 0$ , sondern in der Nähe von  $\mu_{2|1} = \frac{1}{2}x_*$  zu finden ist. Zugleich sinkt mit der Kenntnis dieser neuen Datenlage unsere Unsicherheit über die Lageinformation:  $\sigma_{2|1} \approx 0,87 < 1 = \sigma_2$ .

### 5.4.3 Multinomialverteilung

Die **multivariate Massenfunktion** eines Zufallsvektors  $X = (X_1, \dots, X_D)^T$ , dessen Komponenten sämtlich diskrete Zufallsvariablen darstellen, ist durch die gemeinsame Massenfunktion dieser Komponenten gegeben:

$$p_X: \text{supp}(X_1) \times \cdots \times \text{supp}(X_D) \rightarrow [0, \infty[, p_X(u) = p_{X_1, \dots, X_D}(u_1, \dots, u_D)$$

Die folgende Verteilung kann als multivariate Verallgemeinerung der Binomialverteilung (Abschn. 4.1.1) erachtet werden.

Die **Multinomialverteilung** ist die folgende (multivariate) Massenfunktion:

$$\begin{aligned} \mathcal{M}(\cdot | p, N) &: \mathbb{N} \rightarrow [0, 1], \\ \mathcal{M}(k|p, N) &= \mathcal{M}(k_1, \dots, k_D | p_1, \dots, p_D, N) \\ &= \begin{cases} \frac{N!}{\prod_{d=1}^D k_d!} \cdot \prod_{d=1}^D p_d^{k_d} & \text{falls } \sum_{d=1}^D k_d = N \\ 0 & \text{sonst} \end{cases} \end{aligned}$$

Dabei sind  $p_1, \dots, p_D \in [0, 1]$  mit  $\sum_{d=1}^D p_d = 1$  und  $N \in \mathbb{N}$  Parameter. Es gilt hier die Konvention „ $0^0 = 0$ “.

Erwartungswertvektor und Kovarianzmatrix eines multinomialverteilten Zufallsvektors  $X \sim \mathcal{M}(\cdot | p, N)$  sind wie folgt gegeben:

$$E[X] = N \cdot p, \quad \Sigma[X] = N \cdot (\Sigma_p - p \cdot p^T)$$

Dabei fassen wir  $p$  als Spaltenvektor auf, und  $\Sigma_p = \text{diag}(p_1, \dots, p_D)$  ist die Matrix mit den Diagonaleinträgen  $p_1, \dots, p_D$  und ansonsten verschwindenden Einträgen.

## Quellen

- [1] Ronald Aylmer Fisher. „The use of multiple measurements in taxonomic problems“. In: *Annals of Eugenics* 7.2 (Sep. 1936), S. 179–188. DOI: [10.1111/j.1469-1809.1936.tb02137.x](https://doi.org/10.1111/j.1469-1809.1936.tb02137.x).
- [2] Center for Machine Learning and Intelligent Systems, University of California, Irvine. *Iris Data Set*. Aufgerufen am 20. Dez. 2020. URL: <https://archive.ics.uci.edu/ml/datasets/iris>.
- [3] Endre Weiszfeld. „Sur le point pour lequel la somme des distances de  $n$  points donnés est minimum“. In: *Tohoku Mathematical Journal* 43 (1937), S. 355–386.
- [4] Amir Beck und Shoham Sabach. „Weiszfeld’s Method: Old and New Results“. In: *Journal of Optimization Theory and Applications* 164.1 (Mai 2014), S. 1–40. DOI: [10.1007/s10957-014-0586-7](https://doi.org/10.1007/s10957-014-0586-7).
- [5] Maurice René Fréchet. „Les éléments aléatoires de nature quelconque dans un espace distancié“. In: *Annales de l’Institut Henri Poincaré* 10.4 (1948), S. 215–310.
- [6] Miroslav Bacák. „Computing Medians and Means in Hadamard Spaces“. In: *SIAM J. Optim.* 24 (2014), S. 1542–1566. [arXiv:1210.2145](https://arxiv.org/abs/1210.2145).
- [7] Richard von Mises. *Mathematical Theory of Probability and Statistics*. Hrsg. von Hilda Geiringer. 1. Aufl. Academic Press, 1964.

## **Teil III**

---

Maschinelles Lernen



## Überwachtes maschinelles Lernen

Gemäß internationalem Technologiestandard ist ein **Algorithmus** [1] eine „endliche geordnete Menge wohldefinierter Regeln für die Lösung eines Problems<sup>1</sup>“.

Ein Beispiel für eine Aufgabe, die sich mithilfe eines Algorithmus gut lösen lässt, ist das alphabetische Sortieren einer Anzahl von beliebigen Zeichenketten, z. B. von Wörtern oder Namen. Die Eingabe eines solchen Algorithmus könnte z. B. das Tupel (David, Robert, Anna, Karl) sein. Die korrekte Ausgabe wäre dann das Tupel (Anna, David, Karl, Robert). Ein einfacher Algorithmus zur allgemeinen Lösung des Problems ist **Bubblesort**:

```
N := Länge des Eingabetupels x
solange vertauscht = falsch tue
    vertauscht := falsch
    für i = 1 bis N - 1 tue
        wenn xi > xi+1 dann
            vertausche xi mit xi+1
            vertauscht := wahr
        Ende
    Ende
Ende
```

Ausgabe: x

Dabei steht „ $x_i > x_{i+1}$ “ für „ $x_i$  steht in der alphabetischen Reihenfolge nach  $x_{i+1}$ “. Bei Bubblesort werden solange direkt benachbarte Zeichenketten in die korrekte Reihenfolge gebracht, bis keine Vertauschungen mehr notwendig sind.

Computer können in kürzester Zeit sehr viele der nötigen Verarbeitungsschritte durchführen. Die Mächtigkeit von Algorithmen besteht darin, dass das Ergebnis für jede beliebige Eingabe stets korrekt ist. Im obigen Beispiel wird dies durch Regeln ermöglicht, die vom Programmierer explizit vorgegeben werden, damit der Computer zum gewünschten Ergebnis gelangt.

---

<sup>1</sup> finite ordered set of well-defined rules for the solution of a problem

Eine solche rein regelbasierte Vorgehensweise ist jedoch nicht immer praktikabel. Eine wichtige Aufgabe von Algorithmen ist z. B. die Klassifikation von Bildern oder Texten. Wie aber sollten explizite Regeln aussehen, welche einen Algorithmus zur Lösung der folgenden Aufgabe befähigt: „Ordne Fotos in Kategorien, je nachdem, ob es sich um eine Landschaftsaufnahme, eine Porträtaufnahme oder eine andere Art von Bild handelt“. Um diese und ähnliche Aufgaben zu lösen, wird sich in der Regel dem **maschinellen Lernen** bedient.

In diesem Kapitel werden wir uns mit **überwachten Verfahren** des maschinellen Lernens befassen. Überwachte Verfahren der Klassifikation basieren auf der statistischen Auswertung einer Stichprobe mit bereits bekannter Kategoriezuordnung, die in diesem Zusammenhang als **Trainingsdatensatz** bezeichnet wird. In obigem Beispiel bestünde ein solcher Trainingsdatensatz etwa aus einer Anzahl von Fotos, von denen jedes einer der Kategorien „Landschaft“, „Porträt“ usw. händisch zugeordnet wurde. Das Verfahren ist dann idealerweise in der Lage, Muster in den Fotos zu erkennen, die Landschafts- und Porträtaufnahmen jeweils charakterisieren und voneinander abgrenzen. Anhand dieser Muster werden dann Regeln erzeugt, mit dem auch neue, nicht im Trainingsdatensatz enthaltene Fotos den Kategorien zugeordnet werden können. Diese Regeln müssen vom Programmierer jedoch nicht explizit vorgegeben werden, sondern werden durch die Maschine anhand des Trainingsdatensatzes „erlernt“.

Es gibt einige Charakteristiken, die Probleme des maschinellen Lernens für gewöhnlich auszeichnen:

- Nicht immer ist es offensichtlich, wie Merkmale aus den zu verarbeitenden Daten extrahiert werden können, und welche Merkmale für die Modellbildung nützlich sind: Es ist ein **Feature-Engineering** erforderlich. Beispielsweise werden Textdaten der statistischen Analyse erst zugänglich gemacht, indem diese in geeigneter Weise numerisch repräsentiert werden.
- Die verarbeiteten Datensätze sind oft **hochdimensional**, d. h., es gibt sehr viele Merkmale. Werden z. B. digitale Fotos verarbeitet, so entsprechen die konkreten Farbwerte jedes Pixels im Bild einer Merkmalsausprägung. Bei einer Auflösung von beispielweise  $1280 \times 720$  Pixeln und drei Farbkanälen sind das  $D = 1280 \cdot 720 \cdot 3 = 2.764.800$  Werte pro Bild. Daher kann es sinnvoll oder gar erforderlich sein, eine **Merkmalsauswahl** oder andere Verfahren der **Dimensionsreduktion** anzuwenden.
- Die für Aufgaben des maschinellen Lernens verwendeten Datensätze bestehen nicht selten auch aus einer großen Anzahl von Merkmalsträgern. ImageNet ist beispielsweise ein Datensatz von über 14 Millionen digitalen Bildern, die insgesamt 1000 verschiedenen Kategorien zugeordnet sind [2]. Datensätze für die maschinelle Verarbeitung von Text können ebenfalls sehr schnell einen Umfang von Millionen von Sätzen oder Absätzen erreichen. Die skalierbare Umsetzung von Algorithmen des maschinellen Lernens für große Datenmengen ist ein wichtiges Thema, auf das wir in diesem Buch jedoch nicht näher eingehen.

- Maschinelles Lernen dient in erster Linie der **Prognose** von Merkmalen von Informationsobjekten, die *nicht* im Trainingsdatensatz enthalten sind. Beispielsweise sollte ein entsprechender Klassifikationsalgorithmus in der Lage sein, Landschaften, Porträts usw. beliebiger Art auf neuen Eingabebildern zu erkennen. Mithin ist die Erwartung an einen solchen Algorithmus nicht, dass dieser allein ein gutes Modell für den Trainingsdatensatz darstellt: Er sollte auf neue Situationen verallgemeinerbar sein.
- Einige Probleme, die ein Computer mit maschinellem Lernen zu lösen in der Lage ist, werden zwar von Menschen (teils mühelos) bewältigt, verschlossen sich aber lange Zeit der Lösung durch den Computer. Dazu gehören z. B. die Bilderkennung, speziell Gesichtserkennung, Spracherkennung, die Übersetzung natürlicher Sprachen [3] oder das Spielen des japanischen Brettspiels Go oberhalb von Amateurniveau [4]. Aus diesem Grund stellt das maschinelle Lernen wichtige Werkzeuge im Bereich der **künstlichen Intelligenz** zur Verfügung.

## 6.1 Elemente des überwachten Lernens

Eine wichtige Aufgabe von intelligenten Computersystemen besteht in der Vorhersage der Ausprägungen einer Variable auf der Grundlage neuer Beobachtungen. Beruht diese Vorhersage auf der Verallgemeinerung von zuvor anhand eines sogenannten Trainingsdatensatzes beobachteten Zusammenhängen, so wird von überwachtem Lernen gesprochen.

**Klassifikationsverfahren** dienen der Vorhersage der Ausprägungen einer kategorialen Variable. Ist die Zielgröße eine metrischen Variable, so wird von einem **Regressionsverfahren** im engeren Sinne gesprochen.<sup>2</sup> Beispielsweise fiele die Vorhersage von Börsenkursen aufgrund historischer Verläufe in den Bereich der Regression. Die automatische Einordnung von digitalen Fotos in Kategorien (z. B. Landschaftsaufnahmen, Porträtaufnahmen usw.) anhand von zuvor händisch sortierten Fotos gehört in den Bereich der Klassifikation. Ein weiteres Beispiel für eine Klassifikationsaufgabe ist die Filterung von Spam-E-Mails anhand von zuvor durch den Nutzer identifizierten Beispielen für unerwünschte E-Mail-Nachrichten.

Formal kann die Situation wie folgt beschrieben werden. Gegeben ist zum einen ein Merkmalsraum  $\mathcal{X}$ , welcher den Wertebereich der Einflussgrößen darstellt, anhand derer eine Vorhersage vorgenommen werden soll. Beispielsweise könnte  $\mathcal{X} = \mathbb{R}^D$  gelten, es würden also  $D$  metrische Merkmale für die Vorhersage herangezogen. Diese könnten etwa die Farbwerte eines digitalen Fotos sein.

Zum anderen ist der Wertebereich  $\mathcal{Y}$  der vorherzusagenden Zielgröße anzugeben. Für Regressionsverfahren gilt  $\mathcal{Y} = \mathbb{R}$ . Im Falle der Klassifikation

---

<sup>2</sup> Sowohl Klassifikations- als auch Regressionsverfahren der statistischen Lerntheorie basieren wesentlich auf der Idee der Regressionsanalyse; hier hat sich keine einheitliche Sprechweise durchgesetzt.

besteht  $\mathcal{Y}$  aus einer endlichen Anzahl von Kategorien oder Klassen, z. B.  $\mathcal{Y} = \{\text{Landschaft, Porträt, andere}\}$ . Diese werden auch mit dem englischen Fachbegriff **Labels** bezeichnet. Im Folgenden können wir durch geeignete Numerierung der Kategorien stets  $\mathcal{Y} \subset \mathbb{N}$  annehmen.

Die binäre Klassifikation,  $\mathcal{Y} = \{0, 1\}$ , ist der Anschauung besonders dienlich. Außerdem ist es mitunter sinnvoll, eine allgemeine Klassifikationsaufgabe in eine Reihe von binären Merkmalen der Form „gehört zur Klasse“/„gehört nicht zur Klasse“ zu überführen. Eine solche Übersetzung wird auch **One-Hot-Kodierung** genannt.

In den folgenden Abschnitten 6.2 und 6.3 werden eine Reihe konkreter Regressions- bzw. Klassifikationsverfahren vorgestellt. Ergebnis eines solchen Verfahrens ist eine anhand des Trainingsdatensatzes erlernte Abbildungsvorschrift  $\hat{f}: \mathcal{X} \rightarrow \mathcal{Y}$ , die als **Entscheidungsregel** oder **Hypothese** bezeichnet wird. Eine solche Entscheidungsregel ordnet etwa ein Foto, charakterisiert durch die Farbwerte der Pixel, einer Bildkategorie zu.

Im Falle eines Klassifikationsverfahrens nennen wir eine Entscheidungsregel auch Klassifikationsregel oder **Klassifikator**, bei Regressionsverfahren können wir von einer **Regressionsfunktion** sprechen. Der Klassifikator bzw. die Regressionsfunktion wird dabei aus einem begrenzten **Hypothesenraum**  $\mathcal{F} \subset \{f: \mathcal{X} \rightarrow \mathcal{Y}\}$  ausgewählt. In vielen Fällen kann der Hypothesenraum als eine parametrisierte Familie von Funktionen von der Form

$$\mathcal{F} = \{f(\cdot; \theta): \mathcal{X} \rightarrow \mathcal{Y} | \theta \in \mathcal{P}\}$$

beschrieben werden. Dabei ist  $\mathcal{P}$  der Raum der vom Verfahren zu erlernenden **Modellparameter**, typischerweise  $\mathcal{P} \subseteq \mathbb{R}^K$ . Die optimale Parameterbelegung  $\hat{\theta} \in \mathcal{P}$  wird während des Trainingsvorgangs durch statistische Schätzverfahren mithilfe des Trainingsdatensatzes berechnet.<sup>3</sup>

Oft wird nicht bloß ein Verfahren isoliert betrachtet, sondern eine Reihe von Verfahren werden herangezogen, die durch **Hyperparameter** voneinander unterschieden werden. Formal zusammengefasst:

Seien  $\mathcal{X}, \mathcal{Y}$  die Wertebereiche von Einfluss- bzw. Zielgröße und  $\mathcal{S}_N = (\mathcal{X} \times \mathcal{Y})^N$  der Raum der Trainingsdatensätze vom Umfang  $N$ .

Eine durch die Hyperparameter  $\alpha \in \mathcal{H}$  parametrisierte Familie von **Verfahren des überwachten Lernens** besteht zum einen aus einer Familie von Hypothesenräumen

$$\mathcal{F}_\alpha = \{f_\alpha(\cdot; \theta): \mathcal{X} \rightarrow \mathcal{Y} | \theta \in \mathcal{P}_\alpha\}_{\alpha \in \mathcal{H}},$$

<sup>3</sup> Wir übergehen hier und im Folgenden ein paar mathematische Details. Beispielweise sollten Entscheidungsregeln  $\hat{f}$  wenigstens messbare Abbildungen sein, also z. B. Erwartungswerte wie  $E[\hat{f}(X)]$  für Zufallsvariablen  $X: \Omega \rightarrow \mathcal{X}$  berechnet werden können.

zum anderen aus einer Berechnungsvorschrift, die jedem Trainingsdatensatz (von in der Regel beliebigem Umfang  $N$ ) eine Belegung von Modellparametern zuordnet:

$$\hat{\theta}_\alpha: \mathcal{S}_N \rightarrow \mathcal{P}_\alpha$$

Das Verfahren verarbeitet also einen Trainingsdatensatz  $(x, y) \cong ((x_1, y_1), \dots, (x_N, y_N))$  mit dem Ziel des Erlernens einer Entscheidungsregel

$$\hat{f}(\cdot) = f_\alpha(\cdot; \hat{\theta}_\alpha(x, y))$$

Diese Entscheidungsregel dient dann der Zuordnung noch „ungesehener“ Daten  $x_*$  einer Klasse oder einem Zielwert  $\hat{y}_* = \hat{f}(x_*)$ .

Die Hyperparameter  $\alpha$  werden anhand von A-priori-Annahmen festgestellt oder durch Validierungsverfahren ermittelt, auf die wir in Abschn. 6.1.3 näher eingehen.

Eine binäre Entscheidungs- bzw. Klassifikationsregel  $f: \mathcal{X} \rightarrow \{0, 1\}$  ist gleichbedeutend mit einer Einteilung des Merkmalsraums in zwei disjunkte Klassenbereiche:

$$\mathcal{X} = f^{-1}(\{0\}) \cup f^{-1}(\{1\}) \text{ mit } f^{-1}(\{0\}) \cap f^{-1}(\{1\}) = \emptyset$$

Die beiden Klassenbereiche werden durch eine **Entscheidungsgrenze** voneinander getrennt, in einem gewissem Sinn ist die Regel entlang dieser Grenze eigentlich „unentschieden“. Eine wichtige Familie von Hypothesen über  $\mathcal{X} = \mathbb{R}^D$  ist die der **linearen Klassifikatoren**; bei diesen ist die Entscheidungsgrenze durch eine (affine) Hyperebene gegeben:

$$\mathcal{F} = \left\{ f(u; w) = \begin{cases} 1 & \text{falls } \sum_{d=1}^D w_d u_d < w_0 \\ 0 & \text{sonst} \end{cases} \mid w = (w_0, \dots, w_D) \in \mathbb{R}^{D+1} \right\}$$

Die im Abschn. 6.3.1 beschriebene logistische Regression ist ein Beispiel für ein Verfahren, das als Entscheidungsregel einen linearen Klassifikator lernt.

Um zu illustrieren, wie sich ein uns bereits bekanntes Regressionsverfahren in obige Charakterisierung einfindet, betrachten wir die einfache lineare Regression. Es handelt sich dabei um ein Regressionsverfahren mit  $\mathcal{X} = \mathbb{R}$  und  $\mathcal{Y} = \mathbb{R}$ . Der Hypothesenraum ist durch den Raum aller Ausgleichsgeraden gegeben:

$$\mathcal{F}_1 = \{f: \mathbb{R} \rightarrow \mathbb{R}, f(u; m, c) = mu + c | m, c \in \mathbb{R}\}$$

Der Modellparameterraum setzt sich aus allen grundsätzlich möglichen Werten für die Steigung und den  $y$ -Achsenabschnitt zusammen:  $\mathcal{P} = \{(m, c) | m, c \in \mathbb{R}\} = \mathbb{R}^2$ . Eine weitere Abhängigkeit von Hyperparametern besteht nicht.

Der Trainingsvorgang besteht in der Berechnung der Minimalstelle der Residuenquadratsumme; hier im Ergebnis für einen Trainingsdatensatz  $(x, y) \cong ((x_1, y_1), \dots, (x_N, y_N))$  noch einmal ausgeschrieben:

$$\hat{\theta}(x, y) = \left( \frac{s(x, y)}{s^2(x)}, \bar{y} - \frac{s(x, y)}{s^2(x)} \cdot \bar{x} \right) = (\hat{m}, \hat{c})$$

Die gelernte Entscheidungsregel  $\hat{f}: \mathbb{R} \rightarrow \mathbb{R}$  ist dann die aus den Daten geschätzte Ausgleichsgerade  $\hat{f}(u) = \hat{m}u + \hat{c}$  für alle  $u \in \mathbb{R}$ .

Im Abschn. 6.2.1 werden wir sehen, wie nicht nur Ausgleichsgeraden, sondern auch Polynome höheren Grades an eine Stichprobe von Datenpunkten optimal angepasst werden können. Für jeden Polynomgrad  $K$ , der als Hyperparameter gedacht werden kann, wird auf diese Weise ein Regressionsverfahren mit dem folgenden Hypothesenraum festgelegt:

$$\mathcal{F}_K = \left\{ f_K(u; w) = \sum_{k=0}^K w_k u^k \mid w = (w_0, \dots, w_K) \in \mathbb{R}^{K+1} \right\}_{K \in \mathbb{N}}$$

### 6.1.1 Verlustfunktionen und empirisches Risiko

Um die Güte der Vorhersage einer Entscheidungsregel quantitativ bewerten zu können, muss eine **Verlustfunktion**

$$\lambda: \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty[, (y, \hat{y}) \mapsto \lambda(y, \hat{y})$$

festgelegt werden. Die Verlustfunktion soll die „Kosten“ oder den „Schaden“ bemessen, der bei einer Prognose von  $\hat{y}$  hervorgerufen wird, wenn der wahre Wert der Zielgröße  $y$  beträgt.

Eine oft für Regressionsverfahren herangezogene Verlustfunktion ist die quadratische Verlustfunktion:

$$\lambda_2(y, \hat{y}) = (y - \hat{y})^2$$

für alle  $y, \hat{y} \in \mathbb{R}$ . Eine weitere mögliche Verlustfunktion ist  $\lambda_1(y, \hat{y}) = |y - \hat{y}|$ .

Im Fall einer Klassifikation ist die sogenannte **Null-Eins-Verlustfunktion** besonders einfach: Eine korrekte Kategoriezuordnung führt zum Verlust  $\lambda = 0$ , eine inkorrekte zum Verlust  $\lambda = 1$ . Im Allgemeinen kann die Verlustfunktion für eine Klassifikation stets durch eine **Kostenmatrix** beschrieben werden, in dieser wird der Verlust für jede mögliche Kombination von wahrer Zuordnung und Vorhersage tabellarisch aufgeführt.

Eine Kostenmatrix für die Identifizierung von Spam-E-Mails könnte etwa wie folgt aussehen ( $a, b \in \mathbb{R}$  mit  $a, b > 0$ ):

		Prognose	
		kein Spam	Spam
wahre Klasse	kein Spam	0	$a$
	Spam	$b$	0

**Tabelle 6.1.** Kostenmatrix, Beispiel „Spamfilter“

Wird eine Nachricht korrekt klassifiziert, so ist der Verlust stets null.

Wird eine Nachricht fälschlich als unerwünscht klassifiziert, so werden die durch diese Fehlklassifikation entstandenen Kosten mit  $\lambda(\text{kein Spam}, \text{Spam}) = a$  bewertet. Lässt der Spam-Filter eine unerwünschte Nachricht durch, so sind die entstandenen Kosten  $\lambda(\text{Spam}, \text{kein Spam}) = b$ .

Wenn es also als ein größerer Schaden angesehen wird, wenn eine erwünschte Nachricht versehentlich im Spamordner landet, als dass eine unerwünschte Nachricht im Posteingang verbleibt, so sollte  $a > b$  gewählt werden.

Das **empirische Risiko** bzw. der **Trainingsfehler** einer Entscheidungsregel  $f: \mathcal{X} \rightarrow \mathcal{Y}$  bezüglich eines Trainingsdatensatzes  $((x_1, y_1), \dots, (x_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$  und einer Verlustfunktion  $\lambda: \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty[$  ist durch den durchschnittlichen Verlust über die Trainingsbeispiele gegeben:

$$\hat{R}[f] = \frac{1}{N} \sum_{n=1}^N \lambda(y_n, f(x_n))$$

Wird die Entscheidungsregel aus einer Familie von Hypothesenräumen  $\mathcal{F}_\alpha = \{f_\alpha(\cdot; \theta)\}$  ausgewählt, so kann der Trainingsfehler als Funktion  $R_\alpha(\cdot)$  in den Modellparametern aufgefasst werden:

$$R_\alpha(\theta) = \hat{R}[f_\alpha(\cdot; \theta)] = \frac{1}{N} \sum_{n=1}^N \lambda(y_n, f_\alpha(x_n; \theta))$$

Zahlreiche Verfahren des überwachten maschinellen Lernens bestehen im Kern in der Definition eines Hypothesenraums und der Wahl einer Verlustfunktion. Die Modellparameter werden bei gegebenem Trainingsdatensatz durch Minimierung des empirischen Risikos ermittelt.

Die einfache lineare Regression entspricht zum Beispiel der Minimierung des folgenden empirischen Risikos über dem Hypothesenraum der Ausgleichsgeraden auf Basis der quadratischen Verlustfunktion:

$$R(m, c) = \frac{1}{N} \sum_{n=1}^N (y_n - mx_n - c)^2$$

mit  $m, c \in \mathbb{R}$ . Bis auf den Vorfaktor „ $1/N$ “ ist das gerade die Residuenquadratsumme.

Allein die Minimierung des empirischen Risikos genügt jedoch nicht, um die **Verallgemeinerbarkeit** der gelernten Entscheidungsregel zu gewährleisten. Damit ist gemeint, dass die Güte der Prognose anhand *neuer* Daten stark eingeschränkt sein kann, obwohl der Klassifikator bzw. die Regressionsfunktion sehr gut an die Trainingsdaten angepasst ist. Dieser Umstand kann durch folgendes extreme Beispiel illustriert werden. Aus einem Trainingsdatensatz  $((x_1, y_1), \dots, (x_N, y_N)) \in (\mathbb{R} \times \mathbb{R})^N$  sei eine Entscheidungsregel wie folgt abgeleitet:

$$\hat{f}(x_*) = \begin{cases} y_i & \text{falls } x_* = x_i \text{ für ein } i \in \{1, \dots, N\} \\ 10^{42} & \text{sonst} \end{cases}$$

Dabei setzen wir voraus, dass die  $x_i$  paarweise verschieden sind oder wenigstens die  $y_i$  eindeutig festlegen. In Worten: Nimmt die Einflussgröße einen im Trainingsdatensatz vorhandenen Wert an, so wird der in diesem Datensatz entsprechende Wert der Zielgröße zugeordnet. Andernfalls wird angenommen, dass die Zielgröße den willkürlichen Wert  $\hat{y} = 10^{42}$  annimmt. Für jede „vernünftige“ Verlustfunktion, z. B. die quadratische Verlustfunktion, verschwindet das empirische Risiko bei diesem „Verfahren“: Die Entscheidungsregel ist stets ideal an die Trainingsdaten angepasst, alle Wertepaare werden mit verschwindender Abweichung reproduziert. Dennoch ist der Algorithmus nicht besser als zufälliges Raten, da er für *fast alle* Werte der Einflussgröße ein Ergebnis mit potenziell beliebig hohem Verlust liefert.

Dieses Beispiel zeigt: Eine Minimierung des empirischen Risikos bzw. Trainingsfehlers  $\hat{R}[f]$  entspricht nicht in jedem Fall einer Minimierung des **erwarteten Risikos** bzw. **Testfehlers**  $R[f]$ . Für eine *fest gewählte* Entscheidungsregel  $f: \mathcal{X} \rightarrow \mathcal{Y}$  und alle  $\varepsilon > 0$  gilt zwar die für den arithmetischen Mittelwert übliche Konsistenzbedingung:

$$\lim_{N \rightarrow \infty} \Pr \left( \left| \hat{R}_N[f] - R[f] \right| < \varepsilon \right) = 0$$

mit

$$\hat{R}_N[f] = \frac{1}{N} \sum_{n=1}^N \lambda(Y_n, f(X_n)), \quad R[f] = E[\lambda(Y_*, f(X_*))]$$

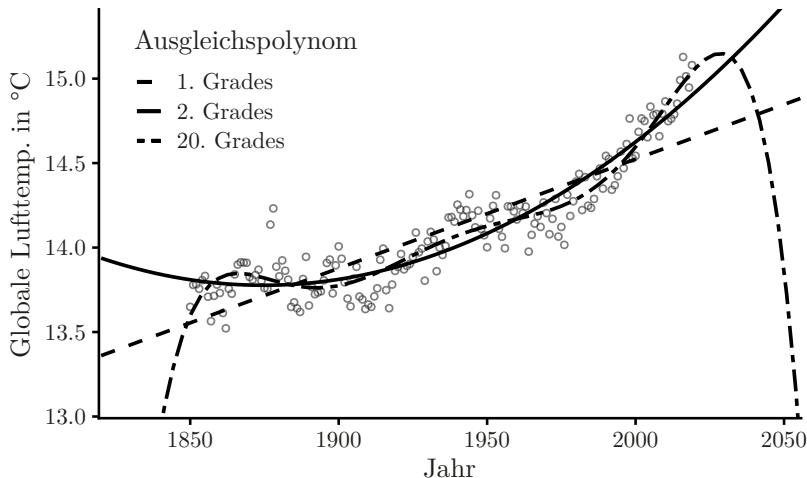
für unabhängige und identisch verteilte Stichprobenvariablen  $X_1, \dots, X_N, X_*$  bzw.  $Y_1, \dots, Y_N, Y_*$  mit Werten in  $\mathcal{X}$  bzw.  $\mathcal{Y}$ . Bei einem Verfahren des maschinellen Lernens wird die betrachtete Entscheidungsregel  $\hat{f}$  jedoch aus einem Hypothesenraum  $\mathcal{F}$  in Abhängigkeit der Trainingsdaten ausgewählt. Eine zur obigen Bedingung analoge Forderung der *gleichmäßigen* Konvergenz wäre dann etwa:

$$\lim_{N \rightarrow \infty} \Pr \left( \sup_{f \in \mathcal{F}} \left| \hat{R}_N[f] - R[f] \right| < \varepsilon \right) = 0$$

für alle  $\varepsilon > 0$ . In diesem Fall kann das empirische Risiko nicht in beliebigem Maße vom erwarteten Risiko abweichen. Eine wesentliche Aufgabe der Theorie des statistischen Lernens besteht in der Untersuchung, in welchen Fällen und inwieweit derartige Anforderungen an die Verallgemeinerbarkeit der Verfahren gewährleistet werden können.

### 6.1.2 Überanpassung und Unteranpassung

Vergleichen wir ein paar Ausgleichspolynome verschiedenen Grades miteinander, ermittelt für den schon in einem vorigen Kapitel als Beispiel betrachteten Datensatz [5] von Messungen der globalen Durchschnittstemperatur der Erde:



**Abb. 6.1.** Globale Temperaturentwicklung mit Ausgleichspolynomen verschiedenen Grades

Je größer der Grad des Polynoms ist, desto besser passt es sich der Stichprobe bzw. den Trainingsdaten an: Das empirische Risiko, also die mittlere quadrierte Abweichung zwischen Ausgleichskurve und Datenpunkten, wird minimiert.

Dies kann als ein allgemeines Phänomen angesehen werden: Je mehr Parameter ein Modell hat (in diesem Fall die Koeffizienten des Ausgleichspolynoms), desto mehr Spielraum besteht, es an die Trainingsdaten anzupassen. Es scheint also zunächst sinnvoll zu sein, möglichst komplexe Modelle zu verwenden, da dann das empirische Risiko minimiert wird. Innerhalb des Intervalls, wo ausreichend Trainingsdaten vorliegen, ist das Polynom 20. Grades am besten an die Trainingsdaten angepasst. Wir sehen aber auch deutlich, dass dieses Modell für Vorhersagen wenig geeignet ist: Für frühe und späte Zeiten jenseits der Trainingsdaten fällt die Ausgleichskurve sehr schnell ab. Die Entscheidungsregel sagt für Vergangenheit und Zukunft der Erde eine „plötzliche Eiszeit“ voraus.

Um genauer zu untersuchen, welchen Einfluss die Modellkomplexität auf den Testfehler hat, machen wir zunächst die folgenden, noch immer vergleichsweise allgemeinen Annahmen in Hinblick auf Regressionsverfahren:

- Die Stichprobenvariablen der Zielgröße sind gemäß  $Y_n = f(x_n) + \varepsilon_n$  verteilt, wobei die unabhängigen und identisch verteilten Störgrößen  $\varepsilon_n$  einen verschwindenden Erwartungswert und endliche Varianz  $\sigma^2 > 0$  haben.
- Trainingsdatensatz  $((x_1, y_1), \dots, (x_N, y_N))$  und ein Punkt im Testdatensatz  $(x_*, y_*)$  sind Realisierungen von unabhängigen und identisch verteilten Zufallsvariablen  $(X_1, \dots, X_N, X_*)$  bzw.  $(Y_1, \dots, Y_N, Y_*)$ .
- Die Verlustfunktion ist die quadratische Verlustfunktion  $\lambda(y, \hat{y}) = (y - \hat{y})^2$  für alle  $y, \hat{y} \in \mathbb{R}$ .

Sei  $x_* \in \mathcal{X}$  eine konkrete Realisierung der Einflussgröße im Testdatensatz. Dann ist der Testfehler (unter der Bedingung  $X_* = x_*$ ) wie folgt gegeben:

$$R[\hat{f}_\alpha | X_* = x_*] = E\left[\left(f(x_*) + \varepsilon_* - \hat{f}_\alpha(x_*)\right)^2\right]$$

mit der Modellvorhersage  $\hat{f}_\alpha(x_*) = f_\alpha(x_*; \hat{\theta}_\alpha)$ . Dabei fassen wir die Vorhersage nun als Schätzfunktion in den Stichprobenvariablen auf:

$$\hat{\theta}_\alpha = \hat{\theta}_\alpha((X_1, Y_1), \dots, (X_N, Y_N))$$

Der Übersichtlichkeit halber führen wir die Abkürzungen  $\hat{f} = \hat{f}_\alpha(x_*)$ ,  $f = f(x_*)$  und  $\varepsilon = \varepsilon_*$  ein. Die Störgröße  $\varepsilon$  mit  $E[\varepsilon] = 0$  und  $E[\varepsilon^2] = \sigma^2$  erzeugt die Schwankungen im Testdatensatz. Da der Schätzer  $\hat{f}$  mittels des Trainingsdatensatzes gelernt wurde, ist dieser mit der Störgröße unkorreliert. Die Größe  $f$  ist bei gegebenem  $x_*$  eine deterministische Konstante. Insgesamt haben wir also:

$$E[\varepsilon \cdot \hat{f}] = E[\varepsilon \cdot f] = 0, E[f \cdot \hat{f}] = E[f] \cdot E[\hat{f}]$$

Das erwartete Risiko kann daher wie folgt umgeformt werden:

$$\begin{aligned} E\left[\left(f - \hat{f} + \varepsilon\right)^2\right] &= E\left[\left(f - \hat{f}\right)^2\right] + E\left[2\varepsilon \cdot \left(f - \hat{f}\right)\right] + E\left[\varepsilon^2\right] \\ &= E\left[\left(f - \hat{f}\right)^2\right] + \sigma^2 \\ &= E\left[f^2\right] - 2E\left[f \cdot \hat{f}\right] + E\left[\hat{f}^2\right] + \sigma^2 \\ &= \left(\left(E\left[\hat{f}\right]\right)^2 - 2E\left[f\right] \cdot E\left[\hat{f}\right] + E\left[f^2\right]\right) \\ &\quad + \left(E\left[\hat{f}^2\right] - \left(E\left[\hat{f}\right]\right)^2\right) + \sigma^2 \\ &= \left(E\left[\hat{f} - f\right]\right)^2 + E\left[\left(\hat{f} - E[\hat{f}]\right)^2\right] + \sigma^2 \end{aligned}$$

**Verzerrung-Varianz-Zerlegung.** Das erwartete Risiko eines Regressionsverfahrens mit quadratischer Verlustfunktion kann wie folgt zerlegt werden:

$$\begin{aligned} R[\hat{f}_\alpha | X_* = x_*] &= \left(E\left[\hat{f}_\alpha(x_*) - f(x_*)\right]\right)^2 + E\left[\left(\hat{f}_\alpha(x_*) - E\left[\hat{f}_\alpha(x_*)\right]\right)^2\right] + \sigma^2 \\ &= \left(\text{Verzerrung}\left[\hat{f}_\alpha | X_* = x_*\right]\right)^2 + \text{Varianz}\left[\hat{f}_\alpha | X_* = x_*\right] + \sigma^2 \end{aligned}$$

Der Testfehler setzt sich also aus den folgenden Anteilen zusammen:

- Aufgrund der zufälligen Schwankungen der Zielgröße  $Y_*$  um den wahren Wert  $f(x_*)$  ist eine Vorhersage mit verschwindendem Fehler auf Grundlage der Daten nicht möglich. Es ist wenigstens der **irreduzible Fehler**  $\sigma^2$  zu erwarten.
- Die **Verzerrung**  $E[\hat{f}_\alpha(x_*) - f(x_*)]$  (engl. *bias*) ist die erwartete Abweichung der Schätzung vom wahren Wert.
- Die **Varianz**  $E[(\hat{f}_\alpha(x_*) - E[\hat{f}_\alpha(x_*)])^2]$  gibt die Schwankung bei wiederholter Anwendung des Verfahrens auf verschiedene Trainingsdatensätze an, auch wenn deren Erhebung dieselbe Verteilung zugrundeliegt.

Das Problem der **Modellauswahl** – der optimalen Wahl der Hyperparameter  $\alpha$  – ist im Kern durch das **Verzerrung-Varianz-Dilemma** gekennzeichnet: Ein gut verallgemeinerbares Modell minimiert nicht nur das empirische, sondern auch das erwartete Risiko. Komplexe Modelle, die an Trainingsdaten gut angepasst werden können und daher ein geringes empirisches Risiko bzw. eine geringe Verzerrung aufweisen, zeigen oft zugleich aber auch eine hohe Varianz: Die Modelle sind möglicherweise zu sensibel gegenüber zufälligen Schwankungen in den Trainingsdaten. In diesem Fall wird von einer **Überanpassung** (im Englischen: *overfitting*) gesprochen. Bei zu einfachen Modellen, welche eine hohe Verzerrung aufweisen, liegt hingegen eine **Unteranpassung** vor.

Die folgende Zerlegung des erwarteten Risikos ist unabhängig von einer spezifischen Verlustfunktion und kann auch auf Klassifikationsverfahren angewendet werden.

**Approximations- und Schätzfehler.** Das erwartete Risiko eines überwachten Verfahrens für maschinelles Lernen kann wie folgt zerlegt werden:

$$\begin{aligned} R[\hat{f}_\alpha] &= \left( \inf_{f \in \mathcal{F}_\alpha} R[f] - R[f_{\text{Bayes}}] \right) + \left( R[\hat{f}_\alpha] - \inf_{f \in \mathcal{F}_\alpha} R[f] \right) + R[f_{\text{Bayes}}] \\ &= \text{Approxationsfehler} + \text{Schätzfehler} + \text{Bayes-Fehler} \end{aligned}$$

Hierbei ist  $R[f_{\text{Bayes}}]$  das minimale erwartete Risiko unter allen grundsätzlich zulässigen Entscheidungsregeln, ohne Einschränkung auf den Hypothesenraum.

Im Detail haben die Terme die folgende Bedeutung:

- Der **Bayes-Fehler** spielt die gleiche Rolle wie der irreduzible Fehler bei der Verzerrung-Varianz-Zerlegung. Er entspricht dem erwarteten Risiko einer optimalen Entscheidungsregel  $f_{\text{Bayes}}$ .
- Der **Approxationsfehler**  $\inf R[f] - R[f_{\text{Bayes}}]$  ist analog zur Verzerrung und gibt das überschüssige Risiko an, welches durch die Auswahl der optimalen Entscheidungsregel aus einem begrenzten Hypothesenraum erzeugt wird.

- Der **Schätzfehler**  $R[\hat{f}_\alpha] - \inf R[f]$  ergibt sich aus der Schätzung der Modellparameter anhand eines begrenzten Trainingsdatensatzes: Selbst wenn durch den Algorithmus das empirische Risiko minimiert wird, minimiert  $\hat{f}_\alpha$  nicht zwangsläufig auch das erwartete Risiko. Er ist analog zur Varianz.

Ein Klassifikationsmodell mit großem Schätzfehler ist überangepasst, ein Modell mit hohem Approximationsfehler unterangepasst.

### 6.1.3 Training, Modellauswahl und Test

Eine wesentliche Herausforderung in der Modellauswahl besteht darin, dass eine Minimierung des empirischen Risikos für eine Verallgemeinerbarkeit nicht garantieren kann und das erwartete Risiko einer Schätzung nicht direkt zugänglich ist. Ein Ausweg besteht in der Verwendung eines zweiten Datensatzes mit bekannter Belegung der Zielgröße, welcher Validierungsdatensatz genannt wird. Die Idee dabei ist, eine Verallgemeinerbarkeit zu gewährleisten, indem für die Validierung bzw. den Test des Verfahrens Daten verwendet werden, die vom Algorithmus noch nicht verarbeitet wurden. Auf diese Weise kann insbesondere eine Überanpassung erkannt und vermieden werden.

**Trainingsdatensatz.** Stichprobe von vorklassifizierten bzw. vorbewerteten Trainingsbeispielen, anhand dessen ein Verfahren des maschinellen Lernens die Parameter eines Modells (durch Minimierung des Trainingsfehlers) bestimmt.

**Validierungsdatensatz.** Stichprobe von vorklassifizierten bzw. vorbewerteten Testbeispielen, die der Modellauswahl dient: Der Testfehler wird anhand dieses Datensatzes geschätzt, um die optimalen Hyperparameter zu ermitteln.

**Testdatensatz.** Stichprobe von Testbeispielen, die ebenfalls der Schätzung des Testfehlers dienen, um die Güte des finalen Klassifikators bzw. des Regressionsmodells zu bewerten.

Ein typischer Arbeitsablauf für die Implementierung und den Test eines Verfahrens des überwachten Lernens ist der folgende:

- Der gesamte Trainingsdatensatz wird partitioniert, in der Regel durch zufällige Auswahl: in einen Trainingsdatensatz im engeren Sinne sowie einen Validierungsdatensatz und einen Testdatensatz. Eine typische Aufteilung im Verhältnis der Datenmengen besteht zu 70 % aus Trainingsdaten und 15 % aus Validierungs- bzw. Testdatensatz.
- Die Modellparameter werden für eine Auswahl von Hyperparametern bestimmt. Die Auswahl der Hyperparameter kann zum Beispiel händisch erfolgen (**Rastersuche**) oder (teilweise) zufällig (**Zufallssuche**).
- Für jede Belegung der Hyperparameter wird der durchschnittliche Verlust anhand des Validierungsdatensatzes ermittelt. Gegebenenfalls werden wei-

tere Kennzahlen für die Bewertung herangezogen. In der Praxis kann z. B. auch die **Effizienz** des implementierten Verfahrens, etwa im Sinne der Datenverarbeitungsgeschwindigkeit oder erforderlicher Speicherressourcen, eine wichtige Rolle spielen.

4. Das Modell bzw. die Hyperparameter, welches bzw. die im Validierungsschritt auf das beste Verfahren führen, wird bzw. werden ausgewählt und mithilfe des Testdatensatzes final bewertet.

In der Praxis wird unter Umständen nicht strikt zwischen Validierungs- und Testdatensatz unterschieden. Entscheidend ist jedoch, dass der Trainingsdatensatz keine Überschneidung mit dem Validierungs-/Testdatensatz hat!

Eine besondere Form der Modellbewertung ist die  **$K$ -fache Kreuzvalidierung**. Bei dieser Methode wird der Ausgangsdatensatz (ggf. nachdem ein Testdatensatz zurückgehalten wurde) in insgesamt  $K$  gleich große und sich nicht überschneidende Validierungsdatensätze aufgeteilt. Zu jedem dieser Validierungsdatensätze gehört ein Trainingsdatensatz: Dies sind einfach alle übrigen Beobachtungen, die nicht im jeweiligen Validierungsdatensatz enthalten sind. Das Training und die Validierung wird mithilfe dieser Sequenz an Datensätzen insgesamt  $K$ -mal ausgeführt, die Modellbewertungen über die Ausführungen gemittelt. Auf diese Weise wird vermieden, dass die Hyperparameter auf Grundlage eines einzelnen, jedoch unglücklich gewählten Validierungsdatensatzes bestimmt werden und daher nicht optimal sind.

## Gütemaße für binäre Klassifikatoren

Bewerten wir das Ergebnis einer binären Entscheidungsregel  $f: \mathcal{X} \rightarrow \{0, 1\}$  mithilfe eines Testdatensatzes vom Umfang  $M$ , so kann für jede der  $M$  Beobachtungen  $(x_m, y_m)$  einer der folgenden Fälle auftreten:

	$f(x_m) = 0$	$f(x_m) = 1$
$y_m = 0$	richtig negativ	falsch positiv
$y_m = 1$	falsch negativ	richtig positiv

**Tabelle 6.2.** Richtig/falsch positive/negative Ergebnisse

In der Literatur ist auch die Konvention  $y \in \{-1, +1\}$  für die Zielgröße binärer Klassifikationsprobleme durchaus üblich, diese machte die Sprechweise vom „negativen“ bzw. „positiven“ Ergebnis etwas einleuchtender.

Durch Auszählen der (absoluten) Häufigkeiten von richtiger oder falscher Klassifikation erhalten wir eine entsprechende Kontingenztafel, welche in diesem Zusammenhang auch **Wahrheitsmatrix** genannt wird:

		$f(x) = 0$	$f(x) = 1$	$\sum$
		$M_{--}$	$M_{-+}$	$M_{-\bullet}$
$y = 0$	$M_{+-}$	$M_{++}$	$M_{+\bullet}$	
	$M_{\bullet-}$	$M_{\bullet+}$	$M$	
$\sum$				

Im Allgemeinen sollten die Raten der Fehlklassifikation  $M_{-+}$  und  $M_{+-}$  möglichst gering gehalten werden. Bei gegebener Kostenmatrix der Form

		$f(x) = 0$	$f(x) = 1$	
		0	a	
$y = 0$	$b$		0	

ist das empirische Risiko durch  $\hat{R}[f] = a \cdot \frac{M_{-+}}{M} + b \cdot \frac{M_{+-}}{M}$  gegeben.

Anstelle des empirischen Risikos werden oft auch die folgenden Kennzahlen für die Beurteilung der gelernten Klassifikationsregel herangezogen, welche im Gegensatz zum Risiko jedoch *maximiert* werden sollten.

Für eine binäre Klassifikationsregel können die folgenden Gütekennzahlen aus den Einträgen der Wahrheitsmatrix abgeleitet werden:

$$\begin{aligned}\text{Korrektklassifikationsrate} &= \frac{M_{--} + M_{++}}{M} \\ \text{Spezifität} &= \frac{M_{--}}{M_{-\bullet}} \\ \text{Trefferquote, Sensitivität} &= \frac{M_{++}}{M_{+\bullet}} \\ \text{Genauigkeit} &= \frac{M_{++}}{M_{\bullet+}}\end{aligned}$$

Für einen vorgegebenen Gewichtungsparameter  $\beta > 0$  wird außerdem das  $F_\beta$ -Maß wie folgt erklärt:

$$\begin{aligned}F_\beta &= \frac{(1 + \beta^2) \cdot M_{++}}{(1 + \beta^2) \cdot M_{++} + \beta^2 \cdot M_{+-} + M_{-+}} \\ &= (1 + \beta^2) \cdot \frac{\text{Genauigkeit} \cdot \text{Trefferquote}}{\beta^2 \cdot \text{Genauigkeit} + \text{Trefferquote}}\end{aligned}$$

Die Korrektklassifikationsrate (im Englischen: *accuracy*) ist gerade gleich  $1 - \hat{R}[f]$ , wenn eine einfache Null-Eins-Verlustfunktion zugrunde gelegt wird. Folgendes sollte beachtet werden: Liegt der Zielgröße eine schiefe Verteilung zugrunde, so kann die Korrektklassifikationsrate ein falsches Bild von der Güte der Klassifikation erzeugen. Haben beispielweise a priori nur 5 % der Beobachtungen eine positive Klassenzuordnung  $y = 1$ , so hätte eine Entscheidungsregel,

die pauschal jeder Beobachtung  $y = 0$  zuweisen würde, bereits eine Korrektklassifikationsrate von 95 %.

Das  $F_\beta$ -Maß ist für  $\beta = 1$  gerade das harmonische Mittel von Genauigkeit (engl. *precision*) und Trefferquote (engl. *recall*). Ist der Wert für  $\beta$  klein, so wird der Genauigkeit eine höhere Bedeutung beigemessen, ein größerer Wert für  $\beta$  entspricht einer höheren Gewichtung der Trefferquote.

Als illustratives Beispiel stellen wir uns vor, wir wollten Spam-E-Mail automatisiert identifizieren und filtern. Das Testergebnis eines trainierten Klassifikators führe auf die folgende Wahrheitsmatrix:

		Prognose	kein Spam	Spam	$\Sigma$
		wahre Klasse			
		kein Spam	400	20	420
		Spam	200	1000	1200
		$\Sigma$	600	1020	1620

**Tabelle 6.3.** Wahrheitsmatrix, Beispiel „Spamfilter“

Damit ergibt sich:

$$\text{Genauigkeit der Spam-Erkennung} = \frac{1000}{1020} \approx 98\%$$

$$\text{Trefferquote der Spam-Erkennung} = \frac{1000}{1200} \approx 83\%$$

In Fehlerquoten ausgedrückt: Von den im Spam-Ordner vorhandenen E-Mails sind 2 % in Wahrheit erwünscht, und 17 % der unerwünschten E-Mails verbleiben trotz des Filters im Posteingang.

Anstelle der Zuordnung von Spam-E-Mails kann auch die Identifikation *erwünschter* Nachrichten als positive Klassenzuordnung erklärt werden. In diesem Fall ergibt sich:

$$\text{Genauigkeit der Erkennung erwünschter E-Mails} = \frac{400}{600} \approx 67\%$$

$$\text{Trefferquote der Erkennung erwünschter E-Mails} = \frac{400}{420} \approx 95\%$$

Von den im Posteingang vorhandenen E-Mails sind also 33 % in Wahrheit unerwünscht; 5 % der erwünschten E-Mails werden in den Spam-Ordner verschoben.

Es ergibt sich die folgende Auswahl an Werten für das  $F_\beta$ -Maß für die Erkennung erwünschter E-Mails:

$$F_{0,5} \approx 70\%, F_1 \approx 78\%, F_2 \approx 88\%$$

Genauigkeit und Trefferquote werden auch für die Beurteilung von Systemen für **Information Retrieval** wie z. B. Suchmaschinen herangezogen. Aufgabe

einer Suchmaschine ist die Ausgabe einer für die Suche relevanten Ergebnismenge von Informationsobjekten durch geeigneten Abgleich der Eingabe mit dem Datenbestand. Die Kennzahlen geben in diesem Fall die folgenden Verhältnisse an:

$$\text{Genauigkeit} = \frac{\text{Anzahl relevanter Ergebnisse}}{\text{Anzahl Ergebnisse}}$$

$$\text{Trefferquote} = \frac{\text{Anzahl relevanter Ergebnisse}}{\text{Anzahl relevanter Entitäten im Datenbestand}}$$

Die Genauigkeit gibt also an, wie viele der Entitäten in der Ergebnismenge für den Nutzer des Systems auf Grundlage seiner Sucheingabe tatsächlich relevant sind. Die Trefferquote ist ein Maß dafür, inwieweit auch alle relevanten Entitäten aufgefunden und ausgegeben wurden.

#### 6.1.4 Numerische Optimierung

Das empirische Risiko stellt sich bei gegebenem Trainingsdatensatz und festgelegten Hyperparametern als eine reellwertige Funktion in den Parametern  $\theta_1, \dots, \theta_K$  dar. Das Training besteht dann im Auffinden der Minimalstelle. In der Regel kann diese Minimalstelle nicht durch eine geschlossene Formel angegeben, sondern muss mittels numerischer Verfahren errechnet werden.

Tatsächlich führen viele Verfahren des statistischen Lernens, auch des unüberwachten Lernens, auf die numerische Berechnung von Extremalstellen einer geeigneten Zielfunktion  $R: \mathbb{R}^K \rightarrow \mathbb{R}$ . Der prototypische Algorithmus für die Lösung dieser Aufgabe ist das **Gradientenverfahren**, auch **Verfahren des steilsten Abstiegs** genannt.

Die Idee ist dabei die folgende. Die Ableitung von  $R$  in Richtung  $h \in \mathbb{R}^K$ ,  $\|h\| = 1$ , kann durch den Gradienten von  $R$  ausgedrückt werden:

$$\frac{dR}{dh}(\theta) = \lim_{\alpha \rightarrow 0} \frac{R(\theta + \alpha h) - R(\theta)}{\alpha} = \langle h, \text{grad } R(\theta) \rangle$$

Ist  $\theta$  nicht gerade eine stationäre Stelle (also eine Stelle mit verschwindendem Gradienten) und zeigt  $h$  in Richtung des negativen Gradienten, so wird die Richtungsableitung negativ sein:

$$\frac{dR}{dh}(\theta) = -\|\text{grad } R(\theta)\|^{-1} \cdot \langle \text{grad } R(\theta), \text{grad } R(\theta) \rangle < 0,$$

falls  $h = -\|\text{grad } R(\theta)\|^{-1} \cdot \text{grad } R(\theta)$  und  $\text{grad } R(\theta) \neq 0$ . Wählen wir ein  $\alpha > 0$  nicht zu groß, dann können wir die Richtungsableitung durch den Differenzenquotienten annähern:

$$\frac{dR}{dh}(\theta) \approx \frac{R(\theta + \alpha h) - R(\theta)}{\alpha}$$

In ausreichend weiter Entfernung von einer stationären Stelle wird diese Näherung ebenfalls einen negativen Wert haben. Folglich gilt in diesen Fällen  $R(\theta) > R(\theta + \alpha h)$ . Diesen Umstand können wir nutzen, um eine Iteration

$$\theta^{(j+1)} = \theta^{(j)} + \alpha h^{(j)}$$

mit absteigenden Funktionswerten zu ermitteln:

$$R(\theta^{(0)}) > R(\theta^{(1)}) > \dots$$

Wenn der zugehörige Gradient bzw. seine Norm  $\|\text{grad } R(\theta^{(j)})\|$  nur noch einen kleinen Wert annimmt, dann befindet sich das Argument  $\theta^{(j)}$  in der Nähe eines lokalen Minimums und die Iteration kann beendet werden.

Sei eine (stetig differenzierbare) Funktion  $R: \mathbb{R}^K \rightarrow \mathbb{R}$ ,  $\theta \mapsto R(\theta)$  gegeben. Ziel des **einfachen Gradientenverfahrens** ist die numerische Berechnung einer lokalen Minimalstelle  $\hat{\theta}$  von  $R(\cdot)$ . Das Verfahren besteht in der folgenden Iteration:

```

Initialisiere  $\theta \in \mathbb{R}^K$ ,
konvergiert := falsch,  $j := 0$ 
solange konvergiert = falsch und  $j < j_{\max}$  tue
    aktualisiere  $j \leftarrow j + 1$ 
    aktualisiere  $\theta \leftarrow \theta + \alpha \cdot h$ , wobei
         $h := -\|\text{grad } R(\theta)\|^{-1} \cdot \text{grad } R(\theta)$ 
    wenn  $\|\text{grad } R(\theta)\| \leq \tau$  dann
        | konvergiert := wahr
    Ende
Ende
Ausgabe: konvergiert,  $\hat{\theta} = \theta$ 
```

Dabei sind  $\alpha > 0$  und  $\tau > 0$  Schrittweite- bzw. Toleranzparameter, die klein gewählt werden. Die Zahl  $j_{\max}$  gibt an, wie viele Iterationsschritte maximal durchgeführt werden sollen.

Um statt einer Minimalstelle eine Maximalstelle von  $R$  aufzufinden, wird der Algorithmus einfach auf  $-R$  angewendet. Im Kontext des maschinellen Lernens wird die Schrittweite  $\alpha$  auch **Lerngeschwindigkeit** (engl. *learning rate*) genannt.

Eine Verwendung des nichtnormierten Gradienten, „ $h = -\text{grad } R(\theta)$ “, ist auch üblich, wenn nicht sogar üblicher. Beharren wir auf der Konvention  $\|h\| = 1$ , so kann ein solches Vorgehen auch als fortlaufend angepasste Schrittweite interpretiert werden:

$$\alpha^{(j)} = \alpha^{(0)} \cdot \|\text{grad } R(\theta^{(j)})\|$$

Eine solche Steuerung der Schrittweite bzw. Lerngeschwindigkeit ist durchaus sinnvoll: Auf diese Weise wird vermieden, „über das Ziel hinauszuschießen“. Ein ausgeklügelteres Verfahren für eine Anpassung der Schrittweite ist die **Methode nach Barzilai und Borwein** [6], das wir hier ohne weitere Herleitung angeben:

$$\alpha^{(j)} = \alpha^{(j-1)} \cdot \frac{\|g^{(j)}\|}{\|g^{(j-1)}\|} \cdot \frac{\langle g^{(j-1)}, g^{(j)} - g^{(j-1)} \rangle}{\|g^{(j)} - g^{(j-1)}\|^2}$$

mit  $g^{(j)} := \text{grad } R(\theta^{(j)})$ .

Abb. 6.3 illustriert oben einen Gradientenabstieg einer Funktion in zwei Variablen,  $R(\theta) = R(\theta_1, \theta_2)$ . Es wurde der nichtnormierte Gradient verwendet, die Pfeile geben die Differenzvektoren aufeinanderfolgender Wertepaare an:

$$\alpha^{(j)} h^{(j)} = -\alpha^{(0)} \cdot \text{grad } R(\theta^{(j)}) = \theta^{(j+1)} - \theta^{(j)}$$

Die Vektoren stehen stets senkrecht auf den ebenfalls eingezeichneten Niveaulinien der Funktion und zeigen in Richtung von deren Minimalstelle.

Im Kontext des statistischen Lernens werden oft Extrempunkte von Funktionen aufgesucht, die in der folgenden Form vorliegen:

$$R(\theta) = \frac{1}{N} \sum_{n=1}^N R_n(\theta)$$

Das empirische Risiko liegt zum Beispiel in dieser Form vor, in diesem Fall wird der Verlust über die Beobachtungen im Trainingsdatensatz summiert. Häufig liegen sehr viele Beobachtungen bzw. Summanden vor. Ein gewöhnlicher Gradientenabstieg bestünde in der wiederholten Berechnung des negativen Gradienten als Abstiegsrichtung:

$$h \propto -\text{grad } R(\theta) = -\frac{1}{N} \sum_{n=1}^N \text{grad } R_n(\theta)$$

Eine solche Berechnung kann bei umfangreichem Trainingsdatensatz „teuer“ werden und würde unter Umständen Systemressourcen in Beschlag nehmen, die nicht zur Verfügung stehen. Beim **stochastischen Gradientenabstieg** wird stattdessen mit jedem Iterationsschritt nur ein einzelner Summand bzw. eine Beobachtung mit Index  $n_1$  für die Berechnung ausgewählt und nur für diese(n) der Gradient bestimmt:

$$h_{\text{stoch}} \propto -\text{grad } R_{n_1}(\theta)$$

Die Auswahl von  $n_1 \in \{1, \dots, N\}$  erfolgt gewöhnlich rein zufällig. Eine weitere Variante des Gradientenabstiegs besteht in der zufälligen Auswahl von bis zu  $1 < M \ll N$  Trainingsbeispielen, einem **Mini-Batch**:

$$h_{\text{batch}} \propto -\frac{1}{M} \sum_{k=1}^M \text{grad } R_{n_k}(\theta)$$

Die Zufallsauswahl erfolgt „ohne Zurücklegen“, sodass sämtliche Trainingsdaten nach insgesamt  $\lceil N/M \rceil$  Iterationsschritten, die eine sogenannte **Epoche** ausmachen, verarbeitet wurden. Für die folgende Epoche werden dann wieder alle

Trainingsbeispiele zur Verfügung gestellt und erneut in zufällig ausgewählten Mini-Batches durchlaufen.

Abb. 6.3 zeigt unten einen stochastischen Gradientenabstieg: Die Abstiegsrichtung steht nicht mehr zwangsläufig orthogonal auf den Niveaulinien der zu minimierenden Funktion. Die generelle Laufrichtung bleibt – über viele Iterationsschritte gemittelt – dennoch mit hoher Wahrscheinlichkeit die richtige.

Ein weiteres Verfahren ist das Broyden-Fletcher-Goldfarb-Shanno-Verfahren (kurz: BFGS-Verfahren) [7, 8, 9, 10]. Mithilfe dieses Verfahrens wurden Ergebnisse für eine Reihe von Beispielen in diesem Band berechnet, daher stellen wir kurz seine wesentlichen Charakteristiken vor.

Das Verfahren basiert im Kern auf der Betrachtung des Taylor-Polynoms zweiter Ordnung anstelle einer nur linearen Näherung der Funktion wie beim Gradientenverfahren:

$$R(\theta + \alpha h) \approx R(\theta) + \alpha \cdot \langle h, \text{grad } R(\theta) \rangle + \frac{\alpha^2}{2} \langle h, \text{Hess } R(\theta) \cdot h \rangle$$

Die Suchrichtung  $h$  soll wieder so gewählt sein, dass der Funktionswert mit dem nächsten Iterationsschritt kleiner wird, der Algorithmus sich also in Richtung eines Minimums bewegt. Auf der anderen Seite wollen wir nicht zu weit in diese Richtung laufen, da andernfalls die Näherung keine Gültigkeit mehr hat. Unter der Annahme, dass die Hesse-Matrix positiv definit ist, liegt das Minimum des Taylor-Polynoms in der folgenden Richtung:

$$h \propto -(\text{Hess } R(\theta))^{-1} \cdot \text{grad } R(\theta)$$

Die Schrittweite  $\alpha$  wird adaptiv gesteuert, und zwar durch näherungsweise Minimierung der eindimensionalen Funktion  $g(\alpha) = R(\theta + \alpha h)$ , etwa vermöge einer **Backtracking-Liniensuche**:

```

Lege Parameter  $c_1 \in ]0, 1[$ ,  $c_2 \in ]0, 1[$  fest,
initialisiere  $\alpha > 0$ ,
berechne  $m := \langle \text{grad } R(\theta), h \rangle$ 
solange  $R(\theta + \alpha h) \geq R(\theta) + c_1 m \alpha$  true
| aktualisiere  $\alpha \leftarrow c_2 \alpha$ 
Ende
Ausgabe:  $\alpha$ 
```

Die Abbruchbedingung wird auch **Armijo-Bedingung** genannt [11]; gewöhnlich wird  $c_1$  sehr klein gewählt.

Schließlich ist eine wesentliche Komponente des BFGS-Algorithmus, dass die Hesse-Matrix nicht exakt berechnet, sondern eine Näherung  $H$  verwendet wird. Diese entspringt der folgenden **Quasi-Newton-Bedingung** im  $j$ -ten Iterationsschritt [12, Abschnitt 3.2]; der Übersichtlichkeit halber ist der Iterationsindex hier als Subskript notiert:

$$H_{j+1} \cdot h_j = \frac{\text{grad } R(\theta_j + \alpha_j h_j) - \text{grad } R(\theta_j)}{\alpha_j} =: \frac{v_j}{\alpha_j}$$

Ist  $H_j$  eine beliebige symmetrische und positiv definite Matrix, so ist  $H_{j+1}$  mit folgender Definition ebenfalls symmetrisch und positiv definit und erfüllt zudem die obige Bedingung:

$$H_{j+1} = H_j + (\alpha_j \cdot \langle v_j, h_j \rangle)^{-1} \cdot v_j \cdot v_j^T - (\langle h_j, H_j \cdot h_j \rangle)^{-1} \cdot (H_j \cdot h_j) \cdot (H_j \cdot h_j)^T$$

Die finale, im folgenden Pseudocode verwendete Formel für die Inverse von  $H$  folgt schließlich aus der sogenannten Sherman-Morrison-Formel [12, Übungsaufgabe 3.13].

Sei eine (zweimal stetig differenzierbare) Funktion  $R: \mathbb{R}^K \rightarrow \mathbb{R}$ ,  $\theta \mapsto R(\theta)$  gegeben. Ziel des **BFGS-Verfahrens** ist die numerische Berechnung einer Minimalstelle  $\hat{\theta}$  von  $R(\cdot)$ . Es besteht aus der folgenden Iteration:

```

Initialisiere:  $j := 0$ , konvergiert := falsch,  $\theta \in \mathbb{R}^K$ ,
 $H^{-1} :=$  Einheitsmatrix vom Format  $K \times K$ 
solange konvergiert = falsch und  $j < j_{\max}$  tue
    aktualisiere  $j \leftarrow j + 1$ 
    # Ermittle Schrittweite und Suchrichtung:
    bestimme (näherungsweise) eine Minimalstelle  $\alpha \in \mathbb{R}$  der
        Funktion  $g(\alpha) := R(\theta + \alpha \cdot h)$  mit  $h := -H^{-1} \operatorname{grad} R(\theta)$ 
    # Aktualisiere Näherung für Funktionswert und Hesse-Matrix:
         $\theta \leftarrow \theta + \alpha \cdot h$ ,
         $H^{-1} \leftarrow H^{-1} + \left(1 + \alpha \frac{v^T H^{-1} v}{h^T v}\right) \cdot \frac{h h^T}{h^T v} - \frac{h v^T H^{-1} + H^{-1} v h^T}{h^T v}$ 
    wobei  $v = \operatorname{grad} R(\theta + \alpha \cdot h) - \operatorname{grad} R(\theta)$ 
    wenn  $\|\operatorname{grad} R(\theta)\| \leq \tau$  dann
        | konvergiert := wahr
    Ende
Ende
Ausgabe: konvergiert,  $\hat{\theta} = \theta$ 
```

Dabei ist  $\tau > 0$  und  $j_{\max}$  ein Toleranzparameter bzw. die maximale Anzahl von Iterationsschritten. Alle Vektoren sind hier als Spaltenvektoren aufgefasst.

## 6.2 Regressionsverfahren

Im den folgenden Abschnitten stellen wir konkrete Regressionsverfahren vor: Gegeben ist dabei stets ein Trainingsdatensatz gepaarter Stichproben  $(x, y) \cong ((x_1, y_1), \dots, (x_N, y_N))$  mit den Realisierungen einer im Allgemeinen multivariaten Einflussgröße  $x_1, \dots, x_N \in \mathbb{R}^D$  und den Realisierungen der Zielgröße  $y_1, \dots, y_N \in \mathbb{R}$ . Ziel der Verfahren ist das Erlernen einer Regressionsfunktion

$$\hat{f}: \mathbb{R}^D \rightarrow \mathbb{R}.$$

Diese dient dazu, bei Eingabe eines neuen, noch nicht verarbeiteten Testbeispiels  $x_* \in \mathbb{R}^D$  eine geeignete Vorhersage  $\hat{y}_* = \hat{f}(x_*)$  machen zu können.

### 6.2.1 Lineare Regression

Im Abschn. 4.5.1 hatten wir die einfache lineare Regression kennengelernt, bei der eine affin-lineare funktionale Abhangigkeit einer Zielgroe von einer univariaten Einflussgroe untersucht wird. Das Ergebnis ist eine Ausgleichsgerade durch die Datenpunkte mit minimaler Residuenquadratsumme.

Ist die Einflussgroe ein Zufallsvektor, so stellen ihre Realisierungen Merkmalsvektoren  $x_1, \dots, x_N \in \mathbb{R}^D$  dar. Diese konnen wir uns zeilenweise untereinander geschrieben als eine Datenmatrix mit den Eintragen  $x_{nd}$ ,  $n \in \{1, \dots, N\}$ ,  $d \in \{1, \dots, D\}$ , zusammengefasst vorstellen. Mithin hat jeder der Merkmalsvektoren  $x_n$  die Koordinaten  $x_{n1}, \dots, x_{nD}$ . Bei der multivariaten linearen Regression nimmt die Zielgroe dann die folgende Form an:

$$Y_n = w_0 + \sum_{d=1}^D w_d \cdot x_{nd} + \varepsilon_n$$

mit Konstanten  $w_0, w_1, \dots, w_D$  und normalverteilten, unabhangigen Storgroen  $\varepsilon_n \sim \mathcal{N}(\cdot | 0, \sigma^2)$ ,  $\sigma > 0$ .

Die Methode der kleinsten Quadrate, also die Minimierung des empirischen Risikos bezuglich der quadratischen Verlustfunktion, fuhrt auf das folgende Kriterium.

Seien Datenpunkte  $x_1, \dots, x_N \in \mathbb{R}^D$  und damit gepaarte Realisierungen der Zielgroe  $y_1, \dots, y_N \in \mathbb{R}$  gegeben. Die Zielfunktion der **linearen Regression** ist die Residuenquadratsumme  $R: \mathbb{R}^{D+1} \rightarrow [0, \infty[$ :

$$R(w_0, \dots, w_D) = \sum_{n=1}^N \left( y_n - w_0 - \sum_{d=1}^D w_d x_{nd} \right)^2$$

Die geschatzten Modellparameter sind durch die Minimalstelle  $\hat{w} = (\hat{w}_0, \dots, \hat{w}_D)$  von  $R(\cdot)$  bestimmt, und die Regressionsfunktion ist:

$$\hat{f}(x_*) = \hat{w}_0 + \sum_{d=1}^D \hat{w}_d x_{*d}$$

fur alle  $x_* \in \mathbb{R}^D$  mit Koordinaten  $x_{*1}, \dots, x_{*D}$ .

Im nachsten Abschnitt werden wir ein allgemeines Verfahren vorstellen, mit dem die Minimalstelle in einer geschlossenen Form angegeben werden kann.

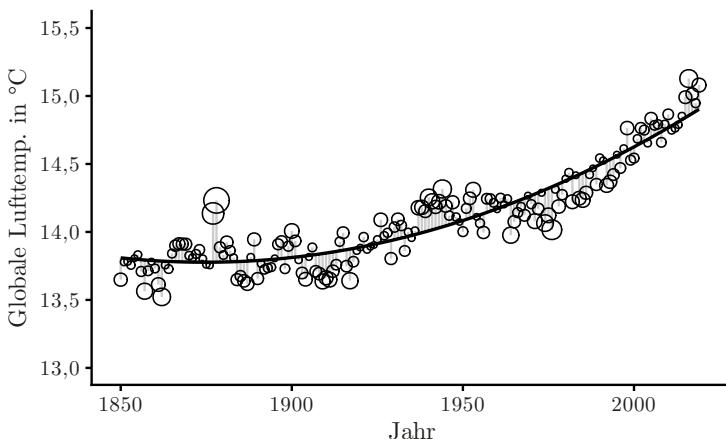
Ergebnis der einfachen linearen Regression ist eine Ausgleichsgerade durch Datenpunkte  $(x_n, y_n) \in \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ . Anstelle dieser Geraden tritt bei der allgemeinen linearen Regression eine Ausgleichshyperebene durch Datenpunkte  $(x_n, y_n) \in \mathbb{R}^D \times \mathbb{R} \cong \mathbb{R}^{D+1}$ .

Zum einen ist eine wichtige Anwendung eines solchen Modells mit mehreren Einflussgrößen natürlich die Berücksichtigung weiterer Merkmale. Eine weitere wichtige Anwendung der multivariaten linearen Regression besteht in der Modellierung nichtlinearer Einflüsse. Beispielsweise können wir quadratische Terme einer univariaten Einflussgröße wie folgt in das Modell einbeziehen:

$$R(w_0, w_1, w_2) = \sum_{n=1}^N (y_n - w_0 - w_1 \cdot x_n - w_2 \cdot x_n^2)^2$$

Die Modellparameter sind dann gerade die Koeffizienten des quadratischen Polynoms, mit dem wir die Verteilung der Beobachtungspaare approximieren wollen. Trotz der nichtlinearen Form der Regressionsfunktion können wir weiterhin von linearer Regression sprechen: Wir haben lediglich den Katalog an Merkmalen erweitert.

**Anwendungsbeispiel.** In Abschn. 4.5.1 hatten wir eine Ausgleichsgerade durch Messdaten der mittleren globalen Lufttemperatur [5] ermittelt (siehe Abb. 4.10). Wird das Modell um einen quadratischen Term erweitert, so führt dies auf ein Polynom als Ausgleichskurve:



**Abb. 6.2.** Globale Temperaturentwicklung und quadratisches Ausgleichspolynom

Das ermittelte Ausgleichspolynom genügt der folgenden Gleichung:

$$\begin{aligned}\hat{f}(x_*) &= 203 \text{ } ^\circ\text{C} - 0,202 \frac{\text{K}}{\text{a}} \cdot x_* + 5,39 \cdot 10^{-5} \frac{\text{K}}{\text{a}^2} \cdot x_*^2 \\ &= 13,8 \text{ } ^\circ\text{C} + 5,39 \cdot 10^{-5} \frac{\text{K}}{\text{a}^2} \cdot (x_* - 1875 \text{ a})^2\end{aligned}$$

Gemäß diesem Modell zeigt sich also seit dem Jahr 1875 eine beschleunigter Anstieg der mittleren globalen Lufttemperatur auf der Erde; die mittlere Abweichung von der Regressionskurve beträgt  $\hat{\sigma} = 0,13 \text{ K}$ .

### Moore-Penrose-Inverse

Die Parameter der linearen Regression sind durch die Minimalstelle der Residuenquadratsumme gegeben. Wie im Folgenden gezeigt werden soll, kann diese Minimalstelle in einer geschlossenen Form angegeben werden. Ein wesentliches Werkzeug ist dabei die sogenannte Moore-Penrose-Inverse einer Matrix. Zunächst erinnern wir an die Definition der *erweiterten* Datenmatrix, bei der die Merkmalsvektoren um eine Eins im „nullten“ Eintrag ergänzt werden:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1D} \\ 1 & x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{ND} \end{pmatrix} = (x_{nd})_{\substack{n \in \{1, \dots, N\} \\ d \in \{0, 1, \dots, D\}}}$$

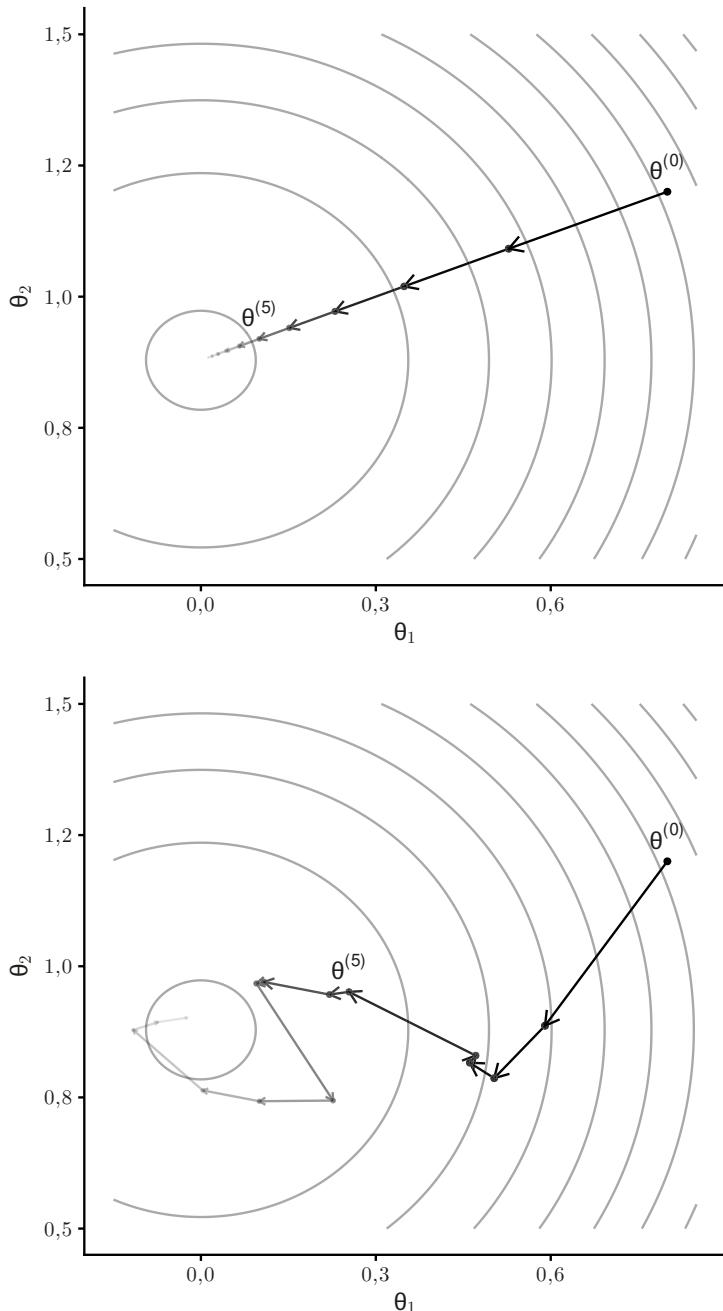
Weiterhin fassen wir die Modellparameter und Realisierungen der Zielgröße zu Spaltenvektoren der Länge  $D + 1$  bzw.  $N$  zusammen:

$$w = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix} \quad \text{und} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

Damit kann die Residuenquadratsumme  $R(w_0, \dots, w_D) = R(w)$  nun wesentlich effizienter notiert werden:

$$\begin{aligned} R(w) &= \sum_{n=1}^N \left( y_n - w_0 - \sum_{d=1}^D w_d x_{nd} \right)^2 \\ &= \sum_{n=1}^N \left( y_n - \sum_{d=0}^D x_{nd} w_d \right)^2 \\ &= \|\mathbf{y} - \mathbf{X} \cdot w\|^2 \end{aligned}$$

Gäbe es eine Lösung  $\hat{w}$  des linearen Gleichungssystems  $\mathbf{X} \cdot w = \mathbf{y}$ , so würde  $R(\hat{w}) = 0$  gelten und die Residuenquadratsumme offensichtlich minimiert werden. Formal könnten wir dann „ $\hat{w} = \mathbf{X}^{-1} \cdot \mathbf{y}$ “ schreiben. Allerdings wird in der Praxis eine solche Lösung im Allgemeinen nicht existieren: Das Gleichungssystem besteht aus  $D + 1$  Unbekannten mit  $N$  Gleichungen, und in der Regel wird die Anzahl der Modellparameter deutlich kleiner als die Stichprobengröße sein: Das Gleichungssystem ist daher überbestimmt. Im Rahmen der einfachen linearen Regression entspricht dies dem Umstand, dass in der Regel wesentlich mehr als nur zwei Datenpunkte gegeben sind, welche die Ausgleichsgerade bestimmen.



**Abb. 6.3.** Gewöhnlicher Gradientenabstieg (oben) und stochastischer Gradientenabstieg (unten)

Dennoch existiert stets eine Minimalstelle der Residuenquadratsumme, und diese kann als  $\hat{w} = \mathbf{X}^\dagger \cdot \mathbf{y}$  geschrieben werden, wobei  $\mathbf{X}^\dagger$  die **verallgemeinerte Matrixinverse** oder **Moore-Penrose-Inverse** der Datenmatrix  $\mathbf{X}$  ist.

Um  $\mathbf{X}^\dagger$  zu bestimmen, berechnen wir zunächst den Gradienten der zu minimierenden Funktion  $R(\cdot)$ :

$$\begin{aligned}\text{grad } R(w) &= (\text{D}R(w))^T = (\text{D}_w \|\mathbf{y} - \mathbf{X} \cdot w\|^2)^T \\ &= (2(\mathbf{y} - \mathbf{X} \cdot w)^T \cdot (-\mathbf{X}))^T = (-2\mathbf{y}^T \mathbf{X} + 2w^T \mathbf{X}^T \mathbf{X})^T \\ &= -2\mathbf{X}^T \cdot \mathbf{y} + 2\mathbf{X}^T \cdot \mathbf{X} \cdot w\end{aligned}$$

Dabei wurden die Kettenregel sowie die grundlegenden Ableitungsregeln

$$\text{D}_z(\|z\|^2) = 2z^T, \quad \text{D}_z(\mathbf{X} \cdot z) = \mathbf{X}$$

verwendet. Außerdem erinnern wir hier an die Rechenregel für Matrizen  $(A \cdot B)^T = B^T \cdot A^T$  und natürlich  $(A^T)^T = A$ .

Der Gradient verschwindet folglich für  $\hat{w} \in \mathbb{R}^{D+1}$  mit

$$\mathbf{X}^T \cdot \mathbf{X} \cdot \hat{w} = \mathbf{X}^T \cdot \mathbf{y}.$$

Diese Gleichung kann direkt nach  $\hat{w}$  aufgelöst werden, sofern  $\mathbf{X}^T \cdot \mathbf{X}$  invertierbar ist:

$$\hat{w} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y},$$

Dies ist dann auch die eindeutig bestimmte Minimalstelle, da  $R(\cdot)$  nichtnegativ und somit nach unten beschränkt ist. In diesem Fall ist  $\mathbf{X}^\dagger = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T$  auch die Moore-Penrose-Inverse von  $\mathbf{X}$ .

Wenn  $\mathbf{X}^T \cdot \mathbf{X}$  *nicht* invertierbar ist, so können wir wie folgt vorgehen. Zunächst bemerken wir, dass die Matrix  $\Sigma := \mathbf{X}^T \cdot \mathbf{X}$  nur nichtnegative Eigenwerte hat, ansonsten könnten wir Vektoren  $w \in \mathbb{R}^{D+1}$  mit  $R(w) < 0$  finden, was nicht möglich ist. Außerdem ist  $\Sigma$  eine symmetrische Matrix und daher wie folgt diagonalisierbar (siehe z. B. [13, Abschnitt 5.6.2]): Es gibt eine orthogonale Matrix  $V$  und eine Diagonalmatrix  $\Lambda$ , sodass gilt:

$$\Sigma = V \cdot \Lambda \cdot V^T$$

In den Spalten von  $V$  stehen die normierten und paarweise orthogonalen Eigenvektoren  $v_0, \dots, v_D \in \mathbb{R}^{D+1}$  von  $\Sigma$ . Auf der Diagonalen von  $\Lambda$  stehen die zugehörigen reellen Eigenwerte  $\lambda_0, \dots, \lambda_D$ . Wir können die Diagonalzerlegung aber auch wie folgt darstellen:

$$\Sigma = \sum_{d=0}^D \lambda_d v_d \cdot v_d^T$$

Dass diese Darstellung korrekt ist, sehen wir daran, dass die Wirkung von  $\Sigma$  auf die Eigenvektorbasis die erwartete ist:

$$\Sigma \cdot v_i = \sum_{d=0}^D \lambda_d v_d \cdot v_d^T \cdot v_i = \sum_{d=0}^D \lambda_d \langle v_d, v_i \rangle v_d = \lambda_i v_i$$

für alle  $i \in \{0, \dots, D\}$ .

Wir legen fest:

$$\mathbf{X}^\dagger = \sum_{d=0}^D \lambda_d^\dagger v_d \cdot v_d^T \cdot \mathbf{X}^T$$

mit

$$\lambda_d^\dagger = \begin{cases} \frac{1}{\lambda_d} & \text{falls } \lambda_d > 0 \\ 0 & \text{falls } \lambda_d = 0 \end{cases}$$

Diese Festlegung löst unser Problem, denn wie wir im Folgenden zeigen werden, erfüllt sie die obige Bedingung eines verschwindenden Gradienten  $\mathbf{X}^T \mathbf{X} \hat{w} = \Sigma \mathbf{X}^\dagger \mathbf{y} = \mathbf{X}^T \mathbf{y}$ .

Hierzu stellen wir zunächst fest, dass für jeden Eigenvektor  $v$  zum Eigenwert Null der Vektor  $v^T \mathbf{X}^T \mathbf{y}$  verschwinden muss:

$$\begin{aligned} \Sigma v = 0 &\Rightarrow \mathbf{X}^T \mathbf{X} v = 0 \Rightarrow v^T \mathbf{X}^T \mathbf{X} v = 0 \\ &\Rightarrow \langle \mathbf{X} v, \mathbf{X} v \rangle = 0 \Rightarrow \mathbf{X} v = 0 \Rightarrow v^T \mathbf{X}^T \mathbf{y} = 0 \end{aligned}$$

Weiterhin gilt  $\sum_{d=0}^D v_d v_d^T v_i = v_i$  zunächst für jeden Eigenvektor  $v_i$  und vermöge Linearität für jeden beliebigen Vektor, folglich ist  $\sum_{d=0}^D v_d \cdot v_d^T$  einfach die Einheitsmatrix.

Daraus ergibt sich schließlich:

$$\begin{aligned} \Sigma \mathbf{X}^\dagger \mathbf{y} &= \Sigma \cdot \left( \sum_{d=0}^D \lambda_d^\dagger v_d v_d^T \mathbf{X}^T \right) \cdot \mathbf{y} = \sum_{d=0}^D \lambda_d^\dagger \Sigma v_d v_d^T \mathbf{X}^T \mathbf{y} \\ &= \sum_{d=0}^D \lambda_d^\dagger \lambda_d v_d v_d^T \mathbf{X}^T \mathbf{y} = \sum_{d=0}^D v_d v_d^T \mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{y} \end{aligned}$$

Wir fassen zusammen:

Die Residuenquadratsumme

$$R(w) = \|y^T - \mathbf{X} \cdot w\|^2$$

mit der erweiterten Datenmatrix  $\mathbf{X}$  vom Format  $N \times (D + 1)$  und den gepaarten Realisierungen der Zielgröße  $y = (y_1, \dots, y_N)$  hat die Minimalstelle

$$\hat{w} = \mathbf{X}^\dagger \cdot y^T,$$

wobei  $\mathbf{X}^\dagger$  die weiter oben beschriebene **Moore-Penrose-Inverse** von  $\mathbf{X}$  ist.

### 6.2.2 Gauß-Prozess-Regression

Für die Herleitung eines weiteren Regressionsverfahrens betrachten wir erneut einen Trainingsdatensatz  $(x, y) \cong ((x_1, y_1), \dots, (x_N, y_N))$  mit den Merkmalsvektoren  $x_1, \dots, x_N \in \mathbb{R}^D$  und Realisierungen der Zielgröße  $y_1, \dots, y_N \in \mathbb{R}$ . Wir gehen im Folgenden der Einfachheit halber davon aus, dass  $\bar{y} = 0$  gilt: Dies kann stets durch Mittelwertzentrierung erreicht werden, von der Zielgröße also deren arithmetischer Mittelwert subtrahiert wird. Anstelle einer linearen Abhängigkeit betrachten wir nun zunächst ein Regressionsmodell von allgemeinerer Form:

$$Y_n = f(x_n) + \varepsilon_n$$

mit normalverteilten Störgrößen  $\varepsilon_n \sim \mathcal{N}(\cdot | 0, \delta_n^2)$  für alle  $n \in \{1, \dots, N\}$ . Wir wollen zulassen, dass die Varianz  $\delta_n > 0$  der Störgrößen im Allgemeinen variiert, also *nicht* zwangsläufig  $\delta_1 = \delta_2 = \dots = \delta_N$  gilt. In diesem Fall wird gesagt, dass **Heteroskedastie** vorliegt.

Betrachten wir die Wertepaare von Einflussgröße und dem zugehörigen (unbekannten, nicht direkt beobachtbaren) Wert der Regressionsfunktion:

$$((x_1, f_1), \dots, (x_N, f_N)) = ((x_1, f(x_1)), \dots, (x_N, f(x_N)))$$

Die Gauß-Prozess-Regression ist eine Bayes'sche Methode: Ihr liegt die Annahme zugrunde, dass unsere Unkenntnis über die Werte  $f_1, \dots, f_N$  durch eine multivariate Normalverteilung modelliert werden kann, mit einem Lagevektor  $m = m(x_1, \dots, x_N) \in \mathbb{R}^N$  und einer  $N \times N$ -Kovarianzmatrix  $\tilde{\Sigma} = \tilde{\Sigma}(x_1, \dots, x_N)$ . Es gilt also:

$$p(f_1, \dots, f_N | x_1, \dots, x_N) = p(f|x) = \mathcal{N}(f|m(x), \tilde{\Sigma}(x))$$

Diese Annahme soll für jede beliebige endliche Auswahl von Wertpaaren Gültigkeit haben, nicht bloß jene im Trainingsdatensatz. Eine Familie  $x_* \mapsto f_*$  von derart verteilten Zufallsgrößen wird als **Gauß-Prozess** bezeichnet. Da  $\bar{y} = 0$  gilt, treffen wir die weitere Annahme  $m(x) = 0$ . Als Ansatz für die Kovarianzmatrix dient eine Ähnlichkeitsmatrix

$$\tilde{\Sigma}(x) = \begin{pmatrix} \sigma(x_1, x_1) & \cdots & \sigma(x_1, x_N) \\ \vdots & & \vdots \\ \sigma(x_N, x_1) & \cdots & \sigma(x_N, x_N) \end{pmatrix}$$

mit einem geeigneten Ähnlichkeitsmaß  $\sigma: \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ .

Die Grundidee eines solchen Ansatzes: Weisen zwei Datenpunkte  $x_m$  und  $x_n$  einen geringen Abstand auf, so sollten die zugehörigen Werte der Zielgröße nicht unabhängig voneinander sein, sondern im Gegenteil eine große Korrelation aufweisen. Sind sie hingegen weit voneinander entfernt, ist davon auszugehen, dass wenig gemeinsame Einflüsse die entsprechenden Werte der Zielgröße bestimmen: Über große Distanzen hinweg ist die Zielgröße dekorreliert. Wie

grundsätzlich alle Modellannahmen muss auch diese kritisch und auf den Anwendungsfall bezogen hinterfragt werden. So blieben bei einer naiven Anwendung des Verfahrens auf eine Zeitreihe periodische Einflüsse unberücksichtigt, welche Korrelationen über weite Zeitspannen hinweg bedingen.

Weiterhin nehmen wir im Folgenden an, dass die Kovarianzterme explizit von einer Gauß'schen Form sind:

$$\sigma(u, v) = a^2 \exp\left(-\frac{1}{2} u^T \cdot \Sigma_h^{-1} \cdot v\right)$$

für alle Spaltenvektoren  $u, v \in \mathbb{R}^D$ . Die Zahlen  $h_1, \dots, h_D, a > 0$  sind weitere Modellparameter und  $\Sigma_h^{-1}$  ist die Matrix mit den Diagonaleinträgen  $h_1^{-2}, \dots, h_D^{-2}$  (alle anderen Einträge sind gleich null).

Um die Verteilung der tatsächlich gemachten Beobachtungen  $y_1, \dots, y_N$  angeben zu können, müssen noch die Störterme berücksichtigt werden:

$$p(y_1, \dots, y_N | x_1, \dots, x_N) = p(y|x) = \mathcal{N}(y|0, K(x))$$

mit  $K(x) = \tilde{\Sigma}(x) + \Sigma_\delta$ . Dabei ist  $\Sigma_\delta = \text{diag}(\delta_1^2, \dots, \delta_N^2)$  die Diagonalmatrix mit den Diagonaleinträgen  $\delta_1^2, \dots, \delta_N^2$ .

Hielten wir alle Parameter in der Notation fest, so hätten wir  $\sigma(\cdot, \cdot) = \sigma_{h,a}(\cdot, \cdot)$ ,  $\tilde{\Sigma}(x) = \tilde{\Sigma}_{h,a}(x)$  und  $K(x) = K_\delta(x; h, a)$ . Der Übersichtlichkeit halber werden wir die Parameter im Folgenden an den meisten Stellen nicht explizit aufführen.

Wir wollen zu gegebenem neuen Wert  $x_*$  den zugehörigen Wert der Regressionsfunktion  $f_* = f(x_*)$  vorhersagen. Zu diesem Zweck ermitteln wir eine A-posteriori-Verteilung, die unsere Kenntnis von diesem Wert beschreibt. Die gemeinsame Dichtefunktion von den Beobachtungen und der Vorhersage ist ebenfalls eine Normalverteilung:

$$p(y_1, \dots, y_N, f_* | x_1, \dots, x_N, x_*) = p(y, f_* | x, x_*) = \mathcal{N}((y, f_*)|0, K_*)$$

Die Kovarianzmatrix  $K_*$  ist dabei durch die folgende Blockmatrix vom Format  $(N+1) \times (N+1)$  gegeben:

$$K_* = \begin{pmatrix} K(x) & K(x_*, x) \\ K(x_*, x)^T & a^2 \end{pmatrix}$$

mit dem Spaltenvektor  $K(x_*, x) := (\sigma(x_*, x_1), \dots, \sigma(x_*, x_N))^T$ . Wir können nun die Formel für die bedingte Wahrscheinlichkeitsdichte von Komponenten eines normalverteilten Zufallsvektors verwenden (siehe Abschn. 5.4.2), um die A-posteriori-Verteilung der Prognose abzuleiten:

$$p(f_* | x, x_*, y) = \mathcal{N}(f_* | K(x_*, x)^T K(x)^{-1} \mathbf{y}, a^2 - K(x_*, x)^T K(x)^{-1} K(x_*, x))$$

mit dem Spaltenvektor der Ausprägungen der Zielgröße  $\mathbf{y} = (y_1, \dots, y_N)^T$ .

Insgesamt verfügt das Modell über die folgenden Parameter: Die Varianzen der Störgröße  $\delta_1^2, \dots, \delta_N^2$ , die Korrelationsreichweiten  $h_1, \dots, h_D$  und die Skalenvariable  $a$ . Die Korrelationsreichweiten und die Skalenvariable sollten als Modellparameter behandelt und aus den Daten geschätzt werden. Würden alle Varianzen der Störgröße ebenfalls aus den Daten geschätzt, kann es leicht zu einer Überanpassung kommen. Alternativ gibt es die folgenden Möglichkeiten:

- Die Varianz der Störgröße wird als gleichbleibend angenommen, es liegt **Homoskedastie** vor – eine Annahme, welche die Anzahl der Modellparameter drastisch reduziert:  $\delta_1 = \dots = \delta_N$ .
- Die Standardabweichungen  $\delta_1, \dots, \delta_N$  werden als Hyperparameter behandelt und aus anderen Quellen erschlossen. Dabei kann es sich z. B. um Angaben von Messgenauigkeit handeln.

In jedem Fall kann aus obiger A-posteriori-Verteilung ein Maximum-a-posteriori-Schätzer und ein Glaubwürdigkeitsintervall gewonnen werden. Wir fassen unsere Überlegungen zusammen, wobei wir die  $\delta_1, \dots, \delta_N$  als Hyperparameter notieren:

Die Log-Likelihood-Funktion der **Gauß-Prozess-Regression** für Datenpunkte  $x_1, \dots, x_N \in \mathbb{R}^D$  und mittelwertzentrierten Realisierungen der Zielgröße  $\mathbf{y} = (y_1, \dots, y_N)^T$  ist wie folgt gegeben:

$$\begin{aligned}\ell_\delta(h, a) &= \ln(p(y|x)) = \ln(\mathcal{N}(y|0, K_\delta(x; h, a))) \\ &= -\frac{1}{2}\mathbf{y}^T(K_\delta(x; h, a))^{-1}\mathbf{y} - \frac{1}{2}\ln(\det(K_\delta(x; h, a))) - \frac{N}{2}\ln(2\pi),\end{aligned}$$

wobei  $K_\delta(x; h, a) = \tilde{\Sigma}_{h,a}(x) + \Sigma_\delta$  wie in obiger Herleitung definiert ist.

Eine Maximalstelle  $(\hat{h}_1, \dots, \hat{h}_D, \hat{a})$  der Log-Likelihood-Funktion führt auf die folgende Regressionsfunktion:

$$\hat{f}: \mathbb{R}^D \rightarrow \mathbb{R}, \quad \hat{f}(x_*) = \hat{K}(x_*, x)^T \hat{K}(x)^{-1} \mathbf{y}$$

mit

$$\hat{K}(x) = K_\delta(x; \hat{h}, \hat{a}), \quad \hat{K}(x_*, x) = (\sigma_{\hat{h}, \hat{a}}(x_*, x_1), \dots, \sigma_{\hat{h}, \hat{a}}(x_*, x_N))^T$$

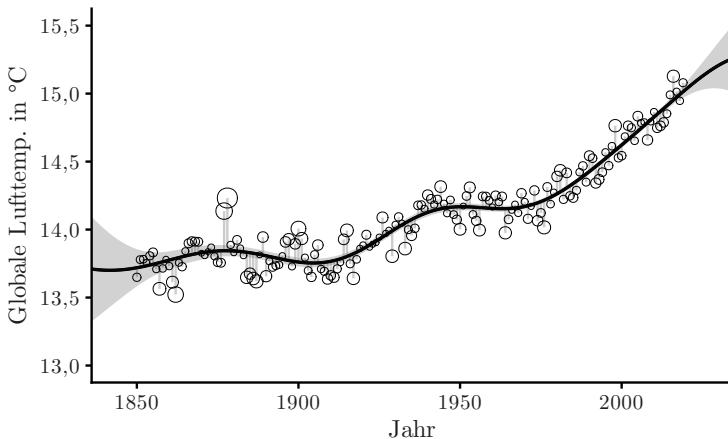
Ein  $\gamma$ -Glaubwürdigkeitsbereich der Schätzung ist durch

$$\left[ \hat{f}(x_*) \pm z(\gamma) \cdot \sqrt{a^2 - \hat{K}(x_*, x)^T \hat{K}(x)^{-1} \hat{K}(x_*, x)} \right]$$

gegeben, wobei z. B.  $z(0,95) = 1,96$ .

Im Falle einer univariaten Einflussgröße ( $D = 1$ ) besteht die gesuchte Ausgleichskurve aus den Wertpaaren  $(x_*, f_*)$ , im multivariaten Fall ( $D > 1$ ) wird eine Hyperfläche im  $(D + 1)$ -dimensionalen Raum beschrieben.

**Anwendungsbeispiel.** Wir wenden das Verfahren der Gauß-Prozess-Regression auf Messdaten der mittleren globalen Lufttemperatur an; der graue Streifen um die Kurve markiert den 95 %-Glaubwürdigkeitsbereich:



**Abb. 6.4.** Globale Temperaturrentwicklung und Ausgleichskurve einer Gauß-Prozess-Regression

Die mit den Rohdaten gelieferten Unsicherheiten in der Messung (je nach Jahr zwischen 0,03 K und 0,21 K) sind direkt als Hyperparameter  $\delta_1, \dots, \delta_N$  in das Modell eingeflossen. Mithilfe des BFGS-Verfahrens können eine Maximalstelle der Likelihood-Funktion und somit die übrigen Parameter ermittelt werden:

$$\bar{y} = 14,10 \text{ } ^\circ\text{C}, h = 30 \text{ y}, a = 0,52 \text{ K}$$

## 6.3 Klassifikationsverfahren

Im den folgenden Abschnitten werden überwachte Verfahren für die Klassifikation vorgestellt. Gegeben ist dabei stets ein Trainingsdatensatz  $(x, y) \cong ((x_1, y_1), \dots, (x_N, y_N))$  mit den Merkmalsvektoren  $x_1, \dots, x_N \in \mathbb{R}^D$  sowie den Klassenlabels  $y_1, \dots, y_N \in \{0, 1, \dots, K - 1\}$ . Eine binäre Klassifikation entspricht dem Fall  $K = 2$ . Ziel der Verfahren ist das Erlernen einer Entscheidungsregel

$$\hat{f}: \mathbb{R}^D \rightarrow \{0, 1, \dots, K - 1\}.$$

Mit dieser kann dann ein noch nicht verarbeitetes Testbeispiel  $x_* \in \mathbb{R}^D$  einer Klasse  $\hat{y}_* = \hat{f}(x_*)$  zugeordnet werden.

### 6.3.1 Logistische Regression

Das Modell der einfachen logistischen Regression lässt sich in ähnlicher Weise wie die einfache lineare Regression auf den multivariaten Fall verallgemeinern.

Der Trainingsdatensatz besteht dann aus einer metrischen erweiterten Datenmatrix  $\mathbf{X} = (x_{nd})$  mit  $N$  Zeilen und  $D + 1$  Spalten sowie den mit diesen Beobachtungen gepaarten, binären Klassenzugehörigkeiten  $y_1, \dots, y_N \in \{0, 1\}$ .

Die Log-Likelihood-Funktion der **logistischen Regression** in den Modellparametern  $w_0, \dots, w_D \in \mathbb{R}$  ist wie folgt gegeben:

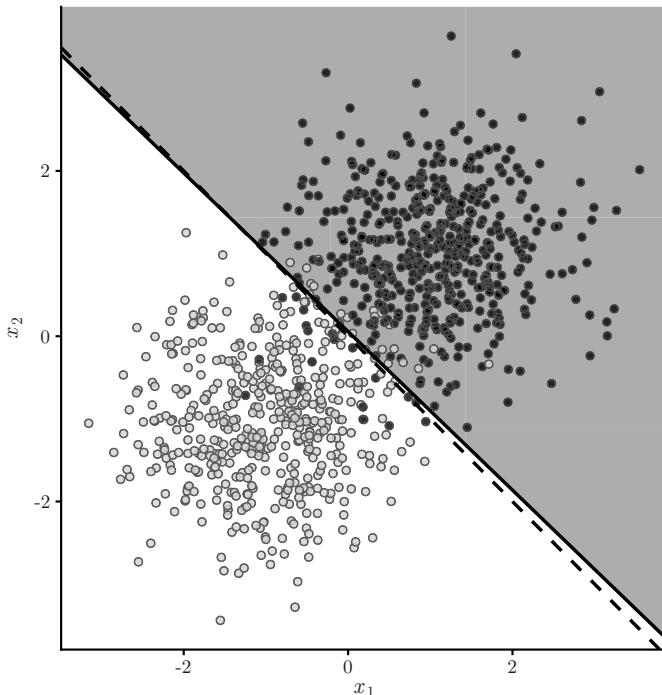
$$\ell(w_0, \dots, w_D) = - \sum_{n=1}^N \ln \left( 1 + \exp \left( (-1)^{y_n} \cdot \sum_{d=0}^D w_d x_{nd} \right) \right)$$

Eine Maximalstelle  $\hat{w} = (\hat{w}_0, \dots, \hat{w}_D)$  von  $\ell(\cdot)$  führt auf die folgende Klassifikationsregel:

$$\hat{f}: \mathbb{R}^{D+1} \rightarrow \{0, 1\}, \quad \hat{f}(x_*) = \begin{cases} 1 & \text{falls } \sum_{d=0}^D \hat{w}_d x_{*d} > 0 \\ 0 & \text{sonst} \end{cases}$$

Die Maximalstelle der Log-Likelihood-Funktion wird mittels numerischer Verfahren ermittelt, etwa dem BFGS-Verfahren.

Um einen ersten Eindruck vom Verhalten des Algorithmus zu gewinnen, wenden wir ihn auf einen synthetischen Trainingsdatensatz niedriger Dimensionalität an:



**Abb. 6.5.** Entscheidungsgrenze der logistischen Regression

Die Abbildung zeigt das Streudiagramm zweier Klassen von insgesamt  $N = 1000$  Datenpunkten, die mittels bivariater Gauß-Verteilungen erzeugt wurden. Die anhand des Datensatzes gelernte Entscheidungsgrenze ist im Allgemeinen eine Hyperebene mit dem Normalenvektor  $w^\perp = (\hat{w}_1, \dots, \hat{w}_D)^T$  und dem Abstand  $\hat{w}_0/\|w^\perp\|$  zum Ursprung. Hier gilt  $D = 2$ , daher handelt es sich um eine Gerade, die als durchgezogene Linie eingezeichnet ist und den Merkmalsraum in die Bereiche einteilt, in denen die Prognosen  $\hat{y} = 0$  (heller Bereich) bzw.  $\hat{y} = 1$  (dunkler Bereich) gemacht werden.

Wir wollen anhand des Beispiels auch noch einmal die Konzepte von empirischem und erwartetem Risiko diskutieren. Das empirische Risiko verschwindet nicht vollständig, denn einige der Datenpunkte im Trainingsdatensatz würden augenscheinlich nicht korrekt klassifiziert werden: Sie befinden sich auf der „falschen“ Seite der Entscheidungsgrenze.

Die gestrichelte Linie zeigt die optimale Entscheidungsgrenze an. In der Praxis ist diese natürlich unbekannt, und hier können wir sie nur deshalb angeben, weil der Datensatz synthetisch erzeugt wurde. Für einen Klassifikator mit dieser Entscheidungsgrenze entspricht das erwartete Risiko dem minimalen Bayes-Fehler, dennoch verschwindet auch für diesen das empirische Risiko nicht.

Für dieses einfache Beispiel stellt die optimale Entscheidungsgrenze ebenfalls einen linearen Klassifikator dar und ist im Hypothesenraums des angewendeten Verfahrens enthalten. Daher verschwindet der Approximationsfehler: Für größere Trainingsdatensätze wird es immer wahrscheinlicher, dass die gelernte Klassifikationsregel der optimalen entspricht; gestrichelte und durchgezogene Gerade fallen dann zusammen. Die in der Abbildung sichtbare Abweichung zwischen gelernter und optimaler Entscheidungsgrenze ist also allein auf den Schätzfehler zurückzuführen.

In Abb. 6.6 ist ein Datensatz zu sehen, bei dem eine logistische Regression zu einer Unteranpassung führen würde, denn die spiralförmig verteilten Klassen können nicht durch eine Entscheidungsgrenze in Form einer einzelnen Geraden voneinander getrennt werden. In diesem Fall sagen wir, dass die Klassen nicht **linear separierbar** sind.

Eine Möglichkeit, die spiralförmigen Bereiche dennoch mithilfe eines linearen Klassifikators zu trennen, besteht im Hinzufügen höherer Potenzen der Einflussgrößen. Dies ist eine Vorgehensweise, die mit der Verallgemeinerung der linearen Regression von Ausgleichsgeraden auf Ausgleichspolynome vergleichbar ist. Im vorliegenden Beispiel können etwa Terme bis maximal zur dritten Potenz hinzugefügt werden. Anstelle einer Geraden  $w_0 + w_1x_1 + w_2x_2 = 0$  ist die Entscheidungsgrenze dann eine algebraische Kurve vom Grad drei:

$$\begin{aligned} w_0 + w_1x_1 + w_2x_2 + w_{11}x_1^2 + w_{12}x_1x_2 + w_{22}x_2^2 \\ + w_{112}x_1^2x_2 + w_{122}x_1x_2^2 + w_{111}x_1^3 + w_{222}x_2^3 = 0 \end{aligned}$$

Das Ergebnis ist in Abb. 6.6 oben zu sehen; eine signifikante Unteranpassung liegt augenscheinlich nicht vor.

Eine andere Sichtweise auf das Verfahren ist die folgende. Die Datenpunkte sind im zweidimensionalen Merkmalsraum nicht linear separiert und werden durch folgende Abbildung in einen höherdimensionalen Raum überführt:

$$\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^9, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \\ x_2^2 \\ x_1^2 x_2 \\ x_1 x_2^2 \\ x_1^3 \\ x_2^3 \end{pmatrix}$$

In diesem höherdimensionalen Raum sind die Klassen nun linear separiert, und die logistische Regression kann erfolgreich angewendet werden. Diese Idee macht sich auch die sogenannte Kernel-Methode zunutze, welche im nächsten Abschnitt beschrieben wird.

### Logistische Regression mit Kern

Als Ausgangspunkt für die folgenden Überlegungen dient uns wieder ein Trainingsdatensatz, der aus den Merkmalsvektoren  $x_1, \dots, x_N \in \mathbb{R}^{D+1}$  sowie den mit diesen Datenpunkten gepaarten Klassenlabels  $y_1, \dots, y_N \in \{0, 1\}$  besteht. Die Merkmalsvektoren wurden um eine Eins im nullten Eintrag erweitert. Weiterhin sei ein symmetrisches Ähnlichkeitsmaß  $\sigma: \mathbb{R}^{D+1} \times \mathbb{R}^{D+1} \rightarrow \mathbb{R}$  gegeben, in diesem Zusammenhang **Kern** (engl. *kernel*) genannt. Eine beliebte Wahl ist einmal mehr der Gauß'sche Kern

$$\sigma_h(u, v) = e^{-\frac{\|u-v\|^2}{2h^2}}$$

für alle  $u, v \in \mathbb{R}^{D+1}$  mit Bandbreite  $h > 0$ . Eine weitere Möglichkeit ist ein polynomialer Kern  $\sigma_k(u, v) = (\langle u, v \rangle)^k$  der Ordnung  $k \in \mathbb{N}$ ,  $k \geq 1$ .

Bei der **logistischen Regression mit Kern**  $\sigma(\cdot, \cdot)$  wird die folgende Zielfunktion betrachtet:

$$\ell(\alpha_1, \dots, \alpha_N) = - \sum_{n=1}^N \ln \left( 1 + \exp \left( (-1)^{y_n+1} \cdot \sum_{m=1}^N \alpha_m \sigma(x_m, x_n) \right) \right)$$

Eine Maximalstelle  $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_N)$  von  $\ell(\cdot)$  führt auf die folgende Klassifikationsregel:

$$\hat{f}: \mathbb{R}^{D+1} \rightarrow \{0, 1\}, \hat{f}(x_*) = \begin{cases} 1 & \text{falls } \sum_{m=1}^N \hat{\alpha}_m \sigma(x_m, x_*) > 0 \\ 0 & \text{sonst} \end{cases}$$

Abb. 6.6 zeigt unten das Ergebnis der Anwendung des Verfahrens mit Gauß'schem Kern (Bandbreite:  $h = 0,05$ ) auf den schon bekannten synthetischen Datensatz spiralförmiger Verteilungen. Bei der praktischen Anwendung des Kernel-Verfahrens gilt es zu beachten, dass für eine Auswertung der gelernten Klassifikationsregel das Ähnlichkeitsmaß zwischen dem zu klassifizierenden Datenpunkt  $x_*$  im Testdatensatz und den Datenpunkten  $x_1, \dots, x_N$  im Trainingsdatensatz berechnet werden muss. Verfahren dieser Art werden als **instanzbasierte Verfahren** bezeichnet. Die Komplexität der Hypothese steigt also mit der Größe des Trainingsdatensatzes, was viele Systemressourcen in Anspruch nehmen kann. Eine Reduktion dieser Komplexität kann daher erforderlich sein, siehe [14] für einen Überblick über entsprechende Methoden.

Eine Begründung für die Funktionsweise des Verfahrens bzw. eine Interpretation auf Grundlage der gewöhnlichen logistischen Regression kann wie folgt gegeben werden. Wir setzen zunächst voraus, dass der Kern die Eigenschaft hat, dass die Matrix mit den Einträgen  $\sigma(x_m, x_n)$  für jede endliche Auswahl von Merkmalsvektoren  $x_1, x_2, \dots$  symmetrisch und positiv semidefinit ist. Es kann nachgewiesen werden, dass diese Bedingung für die oben vorgestellten Beispiele von Gauß'schen und polynomialem Kernen erfüllt ist.

Sei nun neben den Trainingsdaten auch ein Testdatensatz mit den Merkmalsvektoren  $x_{N+1}, \dots, x_{N+M}$  gegeben, und wir betrachten die Matrix

$$\Sigma = (\sigma(x_m, x_n))_{m,n \in \{1, \dots, N+M\}}.$$

Zunächst ziehen wir die Wurzel aus  $\Sigma$ : Wir suchen eine quadratische Matrix  $\Phi$  mit  $\Sigma = \Phi \cdot \Phi^T$ . Ein Weg eine solche Matrix zu konstruieren, ist der folgende.

Da das Ähnlichkeitsmaß als symmetrisch vorausgesetzt wurde, ist  $\Sigma$  eine symmetrische Matrix. Daher existiert eine orthogonale Matrix  $V$  und eine Diagonalmatrix  $\Lambda$ , sodass  $\Sigma = V \cdot \Lambda \cdot V^T$  gilt. Außerdem ist  $\Sigma$  – ebenfalls nach Voraussetzung – positiv semidefinit, daher stehen auf der Diagonalen von  $\Lambda$  nur nichtnegative Werte. Daher können wir die Quadratwurzel aus den Diagonaleinträgen ziehen, wir schreiben für die dadurch entstehende Matrix  $\sqrt{\Lambda}$ . Definieren wir  $\Phi = V \cdot \sqrt{\Lambda}$ , so gilt:

$$\Phi \cdot \Phi^T = V \cdot \sqrt{\Lambda} \cdot \left( V \cdot \sqrt{\Lambda} \right)^T = V \cdot \sqrt{\Lambda} \cdot \sqrt{\Lambda}^T \cdot V^T = V \cdot \Lambda \cdot V^T = \Sigma$$

Weiterhin gilt für alle  $m, n \in \{1, \dots, N, \dots, N+M\}$ , wobei  $\phi_m$  bzw.  $\phi_n$  die  $m$ -te bzw.  $n$ -te Zeile von  $\Phi$  ist:

$$\langle \phi_m, \phi_n \rangle = \phi_m \cdot \phi_n^T = \Sigma_{mn} = \sigma(x_m, x_n)$$

Die Entscheidungsregel der logistischen Regression mit Kern macht also für all jene Merkmalsvektoren  $x_*$  im Testdatensatz eine positive Vorhersage, wenn die folgende Bedingung erfüllt ist:

$$\sum_{m=1}^N \hat{\alpha}_m \sigma(x_m, x_*) = \sum_{m=1}^N \hat{\alpha}_m \langle \phi_m, \phi_* \rangle = \langle w^\perp, \phi_* \rangle > 0$$

mit  $w^\perp = \sum_{m=1}^N \hat{\alpha}_m \phi_m$ . Die Punkte positiver Vorhersage finden sich also auf einer Seite einer Hyperebene mit Normalenvektor  $w^\perp$  wieder.

Somit kann die logistische Regression mit Kern als ein linearer Klassifikator aufgefasst werden – jedoch in einem (in der Regel höherdimensionalen) Raum, der durch die **Merkmalsabbildung** (engl. *feature map*)

$$\Phi: \{x_1, \dots, x_{M+N}\} \rightarrow \mathbb{R}^{N+M}, x_n \mapsto \phi_n = \Phi(x_n)$$

mit der charakteristischen Eigenschaft  $\sigma(x_m, x_n) = \langle \phi_m, \phi_n \rangle$  vermittelt wird.

Unsere Konstruktion einer Merkmalsabbildung hängt explizit vom Testdatensatz ab, auch wenn dieser beliebigen Umfang haben kann. Für ein solides theoretisches Fundament wäre es wünschenswert, wenn diese Einschränkung nicht bestünde und die Merkmalsabbildung universell und von der Form

$$\Phi: \mathbb{R}^{D+1} \rightarrow \mathcal{X}, u \mapsto \phi(u)$$

wäre. Dabei ist  $\mathcal{X}$  ein geeigneter Vektorraum mit Skalarprodukt  $\langle \cdot, \cdot \rangle$ , der eine vollständige Darstellung des Kerns im Bildraum ermöglichte:

$$\sigma(u, v) = \langle \Phi(u), \Phi(v) \rangle$$

für alle  $u, v \in \mathbb{R}^{D+1}$ . Wir gehen auf diese Möglichkeit hier nicht im Detail ein, weisen jedoch darauf hin, dass die Existenz einer solchen Konstruktion unter recht allgemeinen Umständen gezeigt werden kann, siehe z. B. [15, Theorem 6.8]. Der Zielraum  $\mathcal{X}$  ist allerdings nicht mehr von endlicher Dimension, er wird in der englischen Fachliteratur ein *reproducing kernel Hilbert space* genannt.

### 6.3.2 Nächste-Nachbarn-Klassifikation

Die logistische Regression erzeugt einen linearen Klassifikator, der eine einfache geometrische Interpretation zulässt: Die positiv ( $\hat{y} = 1$ ) bzw. negativ ( $\hat{y} = 0$ ) klassifizierten Entitäten werden durch eine Hyperebene im Merkmalsraum voneinander getrennt. Durch nichtlineare Transformation der Einflussgrößen, beispielsweise über die Kernel-Methode, können allgemeinere Hyperflächen als Entscheidungsgrenzen gelernt werden.

Die sogenannte Nächste-Nachbarn-Klassifikation kann unmittelbar über die Anschauung des Merkmalsraums als einen geometrischen Raum motiviert und beschrieben werden: Bei diesem Verfahren wird die Klasse für einen Punkt im Testdatensatz anhand der nächsten Nachbarn durch eine **Mehrheitsentscheidung** vorhergesagt.

Sei  $(x, y)$  mit  $x = (x_1, \dots, x_N) \in \mathcal{X}^N$  und Klassenlabels  $y = (y_1, \dots, y_N)$  ein Trainingsdatensatz, wobei über  $\mathcal{X}$  ein Abstands- oder ein Ähnlichkeitsmaß definiert sei.

Sei ferner der Hyperparameter  $K \in \mathbb{N}$ ,  $K \geq 1$  gegeben. Die  $K$  nächsten Nachbarn eines Merkmalsvektors bzw. einer Merkmalsliste  $x_* \in \mathcal{X}$  sind jene Beobachtungen  $x_{\ell(1)}, \dots, x_{\ell(K)}$ , welche den geringsten Abstand zu bzw. die größte Ähnlichkeit mit  $x_*$  haben.

Der  **$K$ -nächste-Nachbarn-Klassifikator** (kurz: KNN-Klassifikator) ist durch die Entscheidungsregel gegeben, welche  $x_*$  den Modus der benachbarten Klassen  $y_{\ell(1)}, \dots, y_{\ell(K)}$  zuordnet.

Wir setzen hierbei voraus, dass die  $K$  nächsten Nachbarn und der Modus eindeutig bestimmt sind.

Es gibt verschiedene Möglichkeiten, die Fälle zu behandeln, in denen die  $K$  nächsten Nachbarn oder der Modus nicht eindeutig bestimmt sind. Wir stellen hier jeweils eine vor:

- Nach Auswahl von  $K$  Beobachtungen  $x_{\ell(1)}, \dots, x_{\ell(K)}$  mit minimalem Abstand bzw. größter Ähnlichkeit zu  $x_*$  werden alle Beobachtungen mit einem Abstand von höchstens

$$\delta_{\max} = \max_{k \in \{1, \dots, K\}} \{\delta(x_*, x_{\ell(k)})\}$$

um  $x_*$  für die Konstruktion herangezogen, auch wenn dies mehr als  $K$  Beobachtungen sein mögen. Dabei bezeichnet  $\delta(\cdot, \cdot)$  ein Abstandsmaß, für ein Ähnlichkeitsmaß  $\sigma(\cdot, \cdot)$  würden analog alle Punkte mit einem Ähnlichkeitsmaß von wenigstens

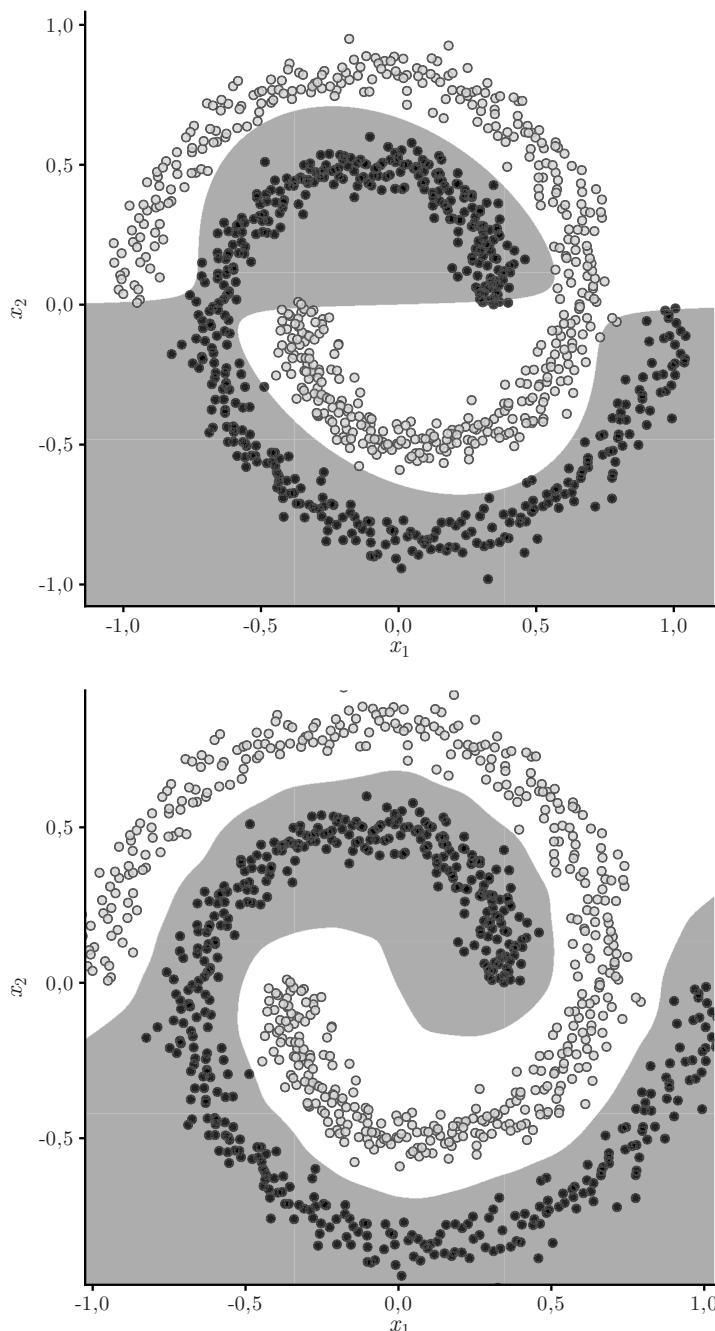
$$\sigma_{\min} = \min_{k \in \{1, \dots, K\}} \{\sigma(x_*, x_{\ell(k)})\}$$

herangezogen.

- Liegt unter den  $K$  nächsten Nachbarn ein Stimmgleichgewicht vor, sind also zwei verschiedene Merkmalsausprägungen mit gleicher maximaler Häufigkeit vorhanden, so werden stattdessen die  $K - 1$  nächsten Nachbarn herangezogen. Die Anzahl der betrachteten Nachbarn wird so lange reduziert, bis eine eindeutige Mehrheitsentscheidung erzielt werden kann.

Die KNN-Klassifikation ist ein instanzbasiertes Verfahren: In seiner grundlegenden Form müssen in der Testphase die Abstände oder Ähnlichkeiten des geprüften Datenpunkts zu allen Datenpunkten im Trainingsdatensatz berechnet werden.

Im Fall  $K = 1$  wird einfach der nächste bzw. ähnlicheste Punkt im Trainingsdatensatz herangezogen und dessen Klasse zugeordnet. Ein so geringer Wert für  $K$  kann jedoch leicht zu einer Überanpassung führen. Der Hyperparameter  $K$  kann durch Validierung ermittelt werden, in der Praxis finden aber auch Faustregeln wie  $K \approx \sqrt{N}$  Verwendung. In Abb. 6.7 wird das Verfahren für zwei Werte von  $K$  anhand eines synthetischen Datensatzes demonstriert, es wird die gewöhnliche euklidische Metrik als Abstandsmaß verwendet.



**Abb. 6.6.** Klassifikation mit logistischer Regression: mittels polynomialer Merkmale (oben) und Kernel-Methode (unten)

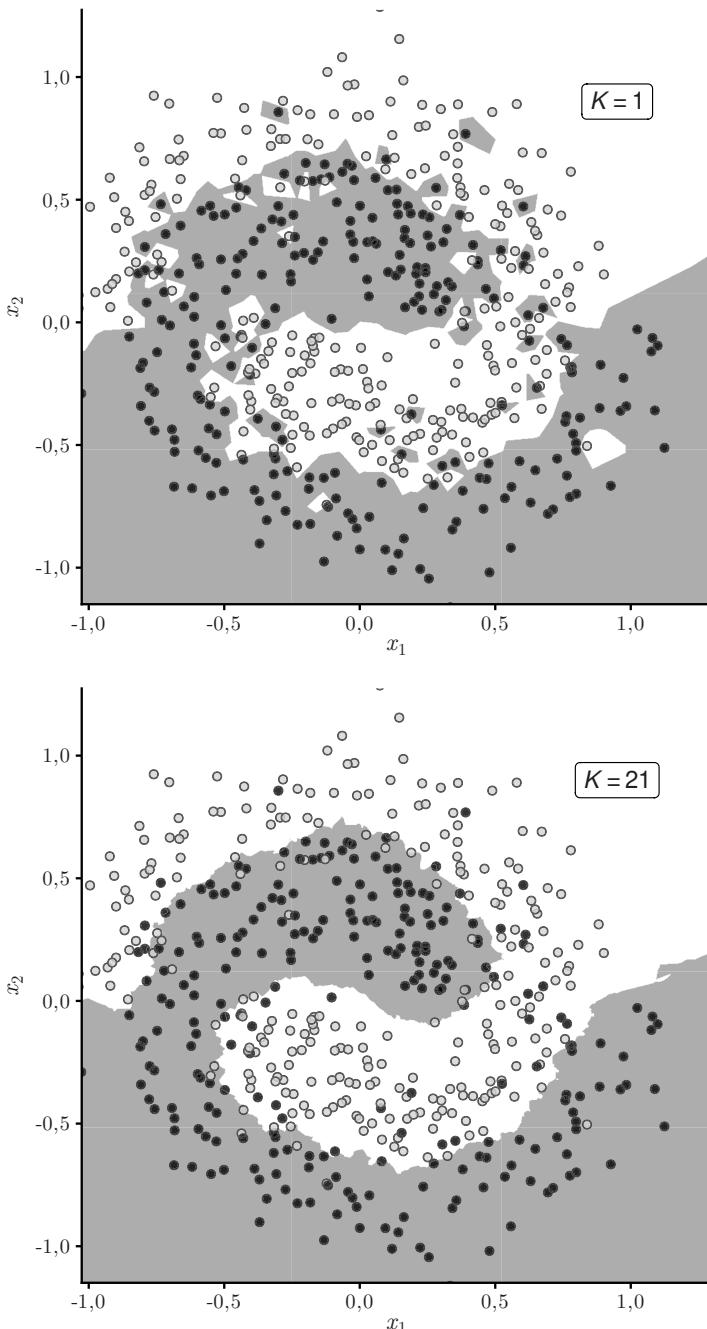


Abb. 6.7.  $K$ -nächste-Nachbarn-Klassifikation

### 6.3.3 Bayes'sche Klassifikationsverfahren

Der Bayes-Klassifikator ist die optimale Entscheidungsregel bei vollständiger Kenntnis der Wahrscheinlichkeitsverteilungen, die den Beobachtungen zugrundeliegen.

Seien  $\mathcal{X}$  der Merkmalsraum der Einflussgröße, z. B.  $\mathcal{X} \subseteq \mathbb{R}^D$ , und  $\mathcal{Y} = \{0, 1, \dots, K\}$  die möglichen Klassenzuordnungen. Auf Grundlage der Null-Eins-Verlustfunktion ordnet der **Bayes-Klassifikator** einem Testbeispiel  $x_* \in \mathcal{X}$  jene Klasse  $y_* \in \mathcal{Y}$  zu, für welche die A-posteriori-Wahrscheinlichkeit

$$\Pr(Y = y_* | X = x_*)$$

maximiert wird.

Um die Notation nicht unnötig zu überladen, haben wir den  $*$ -Index fortgelassen, der die zu den Testdaten gehörige Stichprobenvariable bezeichnet:  $Y = Y_*$  bzw.  $X = X_*$ . Der Einfachheit halber betrachten wir im Folgenden das binäre Klassifikationsproblem mit  $\mathcal{Y} = \{0, 1\}$ . Der Bayes-Klassifikator schreibt sich dann wie folgt:

$$f: \mathcal{X} \rightarrow \mathcal{Y}, f_{\text{Bayes}}(x_*) = \begin{cases} 1 & \text{falls } \Pr(Y = 1 | X = x_*) > \Pr(Y = 0 | X = x_*) \\ 0 & \text{sonst} \end{cases}$$

Handelt es sich bei den Einflussgrößen um kategoriale Merkmale, also Ausprägungen diskreter Zufallsvariablen, so können wir die Bedingung für eine positive Klassenzuordnung mithilfe des Satzes von Bayes umschreiben:

$$\frac{\Pr(X = x_* | Y = 1)}{\Pr(X = x_*)} \cdot \Pr(Y = 1) > \frac{\Pr(X = x_* | Y = 0)}{\Pr(X = x_*)} \cdot \Pr(Y = 0)$$

für alle Testbeispiele  $x_*$  mit  $\Pr(X = x_*) > 0$ . Oder äquivalent dazu:

$$\Pr(X = x_* | Y = 1) \cdot \Pr(Y = 1) > \Pr(X = x_* | Y = 0) \cdot \Pr(Y = 0)$$

Diese Form der Bayes'schen Entscheidungsregel bleibt auch für stetige Einflussgrößen sinnvoll, wenn wir eine bedingte Wahrscheinlichkeitsdichtefunktion zugrundelegen:

$$p_{X|Y}(x_* | 1) \cdot p_Y(1) > p_{X|Y}(x_* | 0) \cdot p_Y(0)$$

In Worten: Jene Klasse wird zugeordnet, bei der die Likelihood multipliziert mit der A-priori-Wahrscheinlichkeit den größeren Wert hat. Bislang haben wir so im Grunde noch nichts erreicht, denn die Verteilungen sind unbekannt. Allerdings können wir nun auf bekannte Methoden der Parameter- und Dichteschätzung zurückgreifen: Wir nehmen an, dass Likelihood und A-priori-Wahrscheinlichkeit durch geeignete statistische Modelle beschrieben werden können:

$$p(x_* | \theta_1) \cdot p_{\text{prior}}(1 | \alpha) > p(x_* | \theta_0) \cdot p_{\text{prior}}(0 | \alpha)$$

Wir machen also die folgenden Annahmen:

- Die Massen- oder Dichtefunktion der Einflussgröße  $X$  unter der Bedingung  $Y = 0$  bzw.  $Y = 1$ , also die Likelihood-Funktion, kann durch ein statistisches Modell  $p(\cdot | \theta_0)$  bzw.  $p(\cdot | \theta_1)$  abgebildet werden. Wird ein solches Modell durch eine Normalverteilung bestimmt, so wird auch von einer **linearen Diskriminanzanalyse** gesprochen.
- A priori sind die Klassen gemäß  $p_{\text{prior}}(\cdot | \alpha)$  verteilt.

Wir erkennen hier das Prinzip der Maximum-a-posteriori-Schätzung wieder (siehe Abschn. 4.4.2). Die Modellparameter  $\theta_0, \theta_1$  werden aus den Daten geschätzt. Der „reinen Lehre“ nach sollten die A-priori-Wahrscheinlichkeiten nur von den Hyperparametern  $\alpha$  abhängig sein, welche eigentlich nicht anhand der vorliegenden Daten erlernt werden. Dennoch ist es üblich, die A-priori-Wahrscheinlichkeiten ebenfalls aus den Trainingsdaten zu schätzen; eine solche Vorgehensweise wird dann als **empirische Bayes-Methode** bezeichnet.

**Anwendungsbeispiel.** Wir bedienen uns eines Beispiels, das von geringer praktischer Bedeutung, aber illustrativ ist. Wir wollen das biologische Geschlecht einer Person allein anhand der Kenntnis der Körpergröße vorhersagen und dabei den CDC-Datensatz [16] als Trainingsdatensatz heranziehen. In Abschn. 4.4.1 hatten wir bereits festgestellt, dass die Häufigkeitsverteilung der Körpergröße  $x$  unter der Bedingung einer bestimmten Geschlechtszugehörigkeit gut durch Normalverteilungen beschrieben werden können:

$$p(x|\hat{\mu}_0, \hat{\sigma}_0^2) = \mathcal{N}(x|\hat{\mu}_0, \hat{\sigma}_0^2), \quad p(x|\hat{\mu}_1, \hat{\sigma}_1^2) = \mathcal{N}(x|\hat{\mu}_1, \hat{\sigma}_1^2)$$

mit den aus den Daten geschätzten Parametern  $\hat{\mu}_0 = 178 \text{ cm}$ ,  $\hat{\sigma}_0 = 7,8 \text{ cm}$  für männliche Studienteilnehmer und  $\hat{\mu}_1 = 163 \text{ cm}$ ,  $\hat{\sigma}_1 = 7,3 \text{ cm}$  für Teilnehmerinnen. Diese Dichtefunktionen können als Likelihood-Funktionen verwendet werden.

Die A-priori-Verteilung ist eine Bernoulli-Verteilung, die etwa wie folgt bestimmt werden kann:

- Die nichtinformative A-priori-Verteilung, also eine Gleichverteilung mit  $p_{\text{prior}}(0) = p_{\text{prior}}(1) = \frac{1}{2}$ .
- Die empirische A-priori-Verteilung, die aus den Daten anhand der relativen Häufigkeit über alle Beobachtungen ermittelt wird:  $p_{\text{prior}}(0) = 0,45$  und  $p_{\text{prior}}(1) = 0,55$ . In diesem speziellen Fall stimmen die Häufigkeiten näherungsweise mit einer Gleichverteilung überein.

Wir gehen im folgenden von einer nichtinformativen A-priori-Verteilung aus. Die Bayes'sche Entscheidungsregel für die Klassifikation einer Person der Größe  $x$  als „weiblich“ wird unter diesen Modellannahmen zu:

$$\begin{aligned}
 p(x|\hat{\mu}_1, \hat{\sigma}_1) \cdot p_{\text{prior}}(1) &> p(x|\hat{\mu}_0, \hat{\sigma}_0) \cdot p_{\text{prior}}(0) \Leftrightarrow \\
 \frac{1}{\sqrt{2\pi}\hat{\sigma}_1} \cdot \exp\left(-\frac{(x-\hat{\mu}_1)}{2\hat{\sigma}_1^2}\right) \cdot \frac{1}{2} &> \frac{1}{\sqrt{2\pi}\hat{\sigma}_0} \cdot \exp\left(-\frac{(x-\hat{\mu}_0)}{2\hat{\sigma}_0^2}\right) \cdot \frac{1}{2} \Leftrightarrow \\
 \hat{\sigma}_0|x - \hat{\mu}_1| < \hat{\sigma}_1|x - \hat{\mu}_0|
 \end{aligned}$$

Die Entscheidungsgrenze liegt an der Stelle

$$x = \frac{\hat{\sigma}_0\hat{\mu}_1 + \hat{\sigma}_1\hat{\mu}_0}{\hat{\sigma}_0 + \hat{\sigma}_1} \approx 1,70 \text{ cm}$$

Eine Person mit einer Körpergröße unterhalb dieses Wertes würde als „weiblich“ klassifiziert. Diese Entscheidungsgrenze ist in Abb. 4.6 eingezeichnet.

## Naive Bayes-Klassifikation

Eine stark vereinfachende Annahme besteht darin, dass die Merkmalsausprägungen unter der Bedingung der Klassenzugehörigkeit voneinander unabhängige Ereignisse sind. Die Likelihood zerfällt dann in ein Produkt über den Merkmalen.

Der binäre **naive Bayes-Klassifikator** auf Grundlage von  $K$  Merkmalen und der Null-Eins-Verlustfunktion ist durch folgende Bedingung an eine positive Klassifikation für das Trainingsbeispiel  $x_* = (x_{*1}, \dots, x_{*K})$  gegeben:

$$\Pr(Y=1) \cdot \prod_{k=1}^K \Pr(X_k = x_{*k}|Y=1) > \Pr(Y=0) \cdot \prod_{k=1}^K \Pr(X_k = x_{*k}|Y=0)$$

Ein wichtiger Anwendungsfall besteht in der Klassifikation von Sequenzen  $t = (t_{x_1}, \dots, t_{x_K})$ , wobei die  $t_{x_k}$  aus einem Inventar von  $D$  Symbolen oder Zeichenketten  $\{t_1, \dots, t_D\}$  stammen, dem **Vokabular**. Ein konkreter Anwendungsbereich ist dabei in der Computerlinguistik zu finden, also der Verarbeitung natürlichsprachlicher Texte. Durch einen Vorverarbeitungsschritt der **Tokenisierung** werden die Texte in Einheiten zerlegt; für gewöhnlich sind dies die einzelnen Wörter, aus denen sich der Text zusammensetzt. Das Vokabular stellt dann die Gesamtheit dieser **Tokens** dar.

Beispielsweise könnte der Text „Das Vokabular stellt die Gesamtheit der Tokens dar.“ nach einer Normierung der Groß- und Kleinschreibung und einer Tokenisierung in diese Folge von Zeichenketten überführt werden:

(das, vokabular, stellt, die, gesamtheit, der, tokens, dar, .)

Eine oft gemachte Annahme besteht darin, dass die Likelihood der Klassenzuordnung unabhängig von der Reihenfolge der Tokens ist: In diesem Fall wird

von einem **Bag-of-Tokens-Modell** oder **Bag-of-Words-Modell** gesprochen. Der obige Satz würde dann nicht z. B. von dieser Folge unterschieden werden:

(., dar, das, der, die, gesamtheit, stellt, tokens, vokabular)

### Multinomiales Ereignismodell

Es gibt zwei wesentliche Möglichkeiten der statistischen Modellierung im Rahmen des Bag-of-Tokens-Ansatzes. Eine besteht darin, die Tokenfolge bzw. die zugehörigen Indizes als Liste kategorialer Variablen  $x = (x_1, \dots, x_K)$  mit  $D$  möglichen Ausprägungen zu betrachten, wobei  $D$  der Umfang des Vokabulars ist:  $x_k \in \{1, \dots, D\}$ .

Die Entscheidungsregel für eine Sequenz  $x_*$  im Testdatensatz ist dann von folgender Form:

$$\Pr(Y = 1) \cdot \prod_{k=1}^K \Pr(X_k = x_{*k} | Y = 1) > \Pr(Y = 0) \cdot \prod_{k=1}^K \Pr(X_k = x_{*k} | Y = 0)$$

mit  $x_{*k} \in \{1, \dots, D\}$ . Der Bag-of-Tokens-Ansatz führt auf die vereinfachende Annahme, dass ein Token mit derselben Wahrscheinlichkeit an jeder Position auftreten kann:  $\Pr(X_k = d | Y = y) = \Pr(X_l = d | Y = y)$  für alle  $k, l \in \{1, \dots, K\}$ ,  $d \in \{1, \dots, D\}$ ,  $y \in \{0, 1\}$ . Wir führen die Abkürzungen  $q = \Pr(Y = 1)$  sowie  $p_{d|1} = \Pr(X_k = d | Y = 1)$  und  $p_{d|0} = \Pr(X_k = d | Y = 0)$  ein. Damit schreibt sich die Entscheidungsregel wie folgt:

$$\frac{q}{1-q} \cdot \prod_{k=1}^K \frac{p_{x_{*k}|1}}{p_{x_{*k}|0}} > 1$$

Eine alternative Schreibweise ist die folgende:

$$\frac{q}{1-q} \cdot \prod_{d=1}^D \left( \frac{p_{d|1}}{p_{d|0}} \right)^{n_{*d}} > 1$$

Dabei ist  $n_{*d}$  die absolute Häufigkeit, mit der das Token  $t_d$  in der zu klassifizierenden Sequenz  $t^{(*)}$  vorkommt. Dies ist das **multinomiale Ereignismodell**: Es wird angenommen, dass das Vorkommen der Tokens innerhalb einer Klasse einer multinomialen Verteilung folgt. Logarithmieren ergibt:

$$\text{logit } q + \sum_{d=1}^D n_{*d} \cdot (\ln p_{d|1} - \ln p_{d|0}) > 0$$

Dabei ist

$$\text{logit: } ]0, 1[ \rightarrow \mathbb{R}, \text{logit}(q) = \ln \left( \frac{q}{1-q} \right)$$

die sogenannte **Logit-Funktion**.

Das Training besteht in der Schätzung der auftretenden Wahrscheinlichkeiten anhand eines Trainingsdatensatzes  $((t^{(1)}, y_1), \dots, (t^{(N)}, y_N))$ . Die A-priori-Wahrscheinlichkeit  $q$  könnte etwa durch die empirische gegeben sein:

$$\hat{q} = \frac{N_+}{N}$$

mit  $N_+ = \sum_{n=1}^N y_n$ , der Anzahl von Sequenzen mit positiver Klassenzuordnung.

Um die Likelihood zu schätzen, konstruieren wir eine Datenmatrix  $\mathbf{N} = (n_{nd})$ , deren Einträge die absolute Häufigkeit des  $d$ -ten Tokens in der  $n$ -ten Sequenz sind. Die relative Häufigkeit des Vorkommens des  $d$ -ten Tokens in jeder Klasse ergibt die folgenden Schätzungen:

$$\hat{p}_{d|1} = \frac{\sum_{n=1}^{N_+} n_{nd}^{(+)}}{\sum_{d=1}^D \sum_{n=1}^{N_+} n_{nd}^{(+)}}, \quad \hat{p}_{d|0} = \frac{\sum_{n=1}^{N_-} n_{nd}^{(-)}}{\sum_{d=1}^D \sum_{n=1}^{N_-} n_{nd}^{(-)}}$$

Dabei sind  $n^{(+)}$  bzw.  $n^{(-)}$  jene Zeilen der Datenmatrix  $\mathbf{N}$ , die mit positiver bzw. negativer Klasse gepaart sind.

In der Praxis besteht eine Herausforderung darin, dass seltene Tokens mitunter überhaupt nicht mit einer bestimmten Klasse vorkommen, sodass für diese nach obiger Rechnung etwa  $\hat{p}_{d|1} = 0$  gelten könnte. Abhilfe schafft eine **Lidstone-Glättung**, mit der die Schätzungen wie folgt korrigiert werden:

$$\hat{p}_{d|1,s} = \frac{\sum_{n=1}^{N_+} n_{nd}^{(+)} + s}{\sum_{d=1}^D \sum_{n=1}^{N_+} n_{nd}^{(+)} + s \cdot D}, \quad \hat{p}_{d|0,s} = \frac{\sum_{n=1}^{N_-} n_{nd}^{(-)} + s}{\sum_{d=1}^D \sum_{n=1}^{N_-} n_{nd}^{(-)} + s \cdot D}$$

mit dem Glättungsparameter  $s \geq 0$ . Der Glättungsparameter ist ein Hyperparameter, dessen optimaler Wert beispielsweise über eine Kreuzvalidierung ermittelt werden kann. Andernfalls ist z.B. der Wert  $s = 1$  üblich, bei diesem Parameterwert wird auch von einer **Laplace-Glättung** gesprochen.

## Bernoulli'sches Ereignismodell

Ein alternative Ansatz der Modellierung besteht darin, jeder Trainingssequenz  $t^{(n)}$  eine Liste binärer Variablen  $x_{n1}, \dots, x_{nD}$  zuzuordnen. Dabei ist  $x_{nd} = 1$ , wenn das Token  $t_d$  in der Sequenz vorkommt, andernfalls gilt  $x_{nd} = 0$ . Die Trainingssequenzen werden auf diese Weise in eine Datenmatrix  $\mathbf{X} = (x_{nd})$  vom Format  $N \times D$  mit binären Einträgen übersetzt. Diese Art der Darstellung wird **One-Hot-Kodierung** genannt und führt auf das **Bernoulli'sche Ereignismodell**.

Die Entscheidungsregel ist dann von folgender Form:

$$\Pr(Y = 1) \cdot \prod_{d=1}^D \Pr(X_d = x_{*d} | Y = 1) > \Pr(Y = 0) \cdot \prod_{d=1}^D \Pr(X_d = x_{*d} | Y = 0)$$

mit  $x_{*d} \in \{0, 1\}$ .

Mit den Abkürzungen  $q = \Pr(Y = 1)$  und  $p_{d|1} = \Pr(X_d = 1|Y = 1)$  bzw.  $p_{d|0} = \Pr(X_d = 1|Y = 0)$  führt dies auf die folgende Entscheidungsregel:

$$\begin{aligned} \text{logit } q + \sum_{d=1}^D \ln \left\{ \begin{array}{ll} \frac{p_{d|1}}{p_{d|0}} & \text{falls } x_{*d} = 1 \\ \frac{1-p_{d|1}}{1-p_{d|0}} & \text{falls } x_{*d} = 0 \end{array} \right\} = \\ \text{logit } q + \sum_{d=1}^D \ln \left( \frac{1 - p_{d|1} + x_{*d} \cdot (2p_{d|1} - 1)}{1 - p_{d|0} + x_{*d} \cdot (2p_{d|0} - 1)} \right) > 0 \end{aligned}$$

Die auftretenden Likelihood-Werte lassen sich wieder aus dem Trainingsdatensatz über die relativen Häufigkeiten der Merkmalsausprägungen schätzen und glätten:

$$\hat{p}_{d|1,s} = \frac{\sum_{n=1}^{N_+} x_{nd}^{(+)} + s}{N_+ + sD}, \quad \hat{p}_{d|0,s} = \frac{\sum_{n=1}^{N_-} x_{nd}^{(-)} + s}{N_- + sD}$$

Dabei sind  $x^{(+)}$  bzw.  $x^{(-)}$  jene Zeilen der Datenmatrix  $\mathbf{X}$ , die positive bzw. negative Klassenzuordnung aufweisen.

## 6.4 Künstliche neuronale Netzwerke

**Künstliche neuronale Netzwerke** sind Verfahren des maschinellen Lernens, die sowohl für Regressions- als auch für Klassifikationsaufgaben verwendet werden. Wie der Name schon andeutet, sind diese Verfahren in der Funktion biologischen Nervensystemen entlehnt. In Abb. 6.10 ist zur Illustration das neuronale Netzwerk des Fadenwurms *Caenorhabditis elegans* dargestellt [17, 18]: Jeder Knoten entspricht einem Neuron, jede Kante einer interneuronalen Synapse (vgl. [19]).

Das wohl erste künstliche neuronale Netzwerk, das „Perzepton Mark I“, wurde bereits 1960 unter der Federführung des Psychologen und Informatikers Frank Rosenblatt entwickelt [20, 21, 22]. Seit etwa 2010 haben neuronale Netzwerke im Bereich des maschinellen Lernens immens an Bedeutung gewonnen und werden für zahlreiche Aufgaben eingesetzt, für die große oder enorme Mengen an Trainingsdaten vorliegen. Sogenannte tiefe neuronale Netzwerke können eine große oder immens große Anzahl von Modellparametern in Größenordnungen von  $10^4$ - $10^{11}$  aufweisen, um aus diesen Trainingsdaten hoch angepasste Klassifikatoren und Regressionsfunktionen zu lernen, die dennoch einen ebenso hohen Grad an Verallgemeinerbarkeit aufweisen.

Die einfachste Form eines künstlichen neuronalen Netzwerks kann wie folgt beschrieben werden.

Einem **vorwärtsgerichteten neuronalen Netzwerk** oder **Feedforward-Netzwerk** liegt ein Hypothesenraum zugrunde, der aus Funktionen der folgenden Form besteht:

$$f: \mathbb{R}^{D_0} \rightarrow \mathbb{R}^{D_L}, f(u) = (f_L \circ f_{L-1} \circ \dots \circ f_1)(u),$$

wobei jede **Schicht** (engl. *layer*)  $f_l$  für alle  $l \in \{1, \dots, L\}$  eine Funktion der folgenden Form ist:

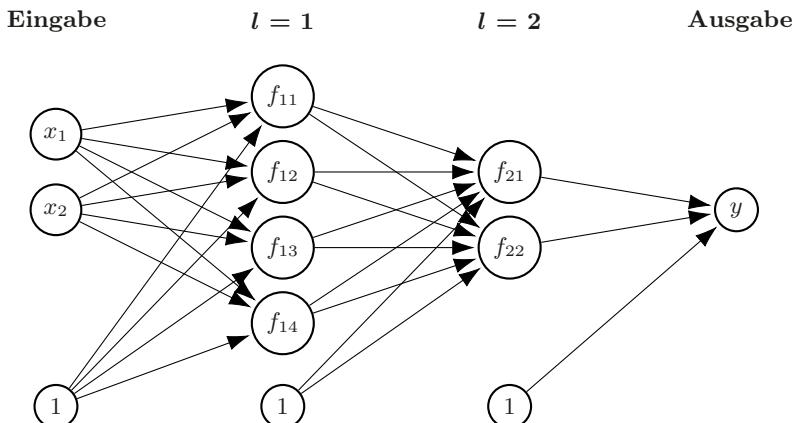
$$f_l: \mathbb{R}^{D_{l-1}} \rightarrow \mathbb{R}^{D_l}, f_l(u) = \phi_l \left( w^{(l)} \cdot u + b^{(l)} \right)$$

mit sogenannten **Aktivierungsfunktionen**  $\phi_l: \mathbb{R}^{D_l} \rightarrow \mathbb{R}^{D_l}$  und Matrizen von **Gewichten**  $w^{(l)}$  vom Format  $D_l \times D_{l-1}$  und den **Verzerrungsvektoren**  $b^{(l)} \in \mathbb{R}^{D_l}$ .

Die Komponentenfunktionen  $f_{l1}, \dots, f_{lD_l}$  stellen die **Neuronen** der jeweiligen  $l$ -ten Schicht dar, ein bestimmter Funktionswert repräsentiert eine **Aktivierung** des Neurons.

Wir stellen uns vor, dass dem Netzwerk eine 0-te Schicht  $f_0: u \mapsto u$  hinzugefügt wird, die **Eingabeschicht**. Die letzte Schicht  $f_L$  stellt die **Ausgabeschicht** dar. Die dazwischenliegenden Schichten mit  $1 \leq l < L$  werden als **verborgen** bezeichnet. Die Anzahl der Schichten bestimmt die **Tiefe** des Netzwerks, die Anzahl der Neuronen in jeder Schicht dessen **Breite**. Hat das Netzwerk mehr als nur eine verborgene Schicht, gilt also  $L > 2$ , so wird von **mehrschichtigem Lernen** gesprochen bzw. der englische Fachbegriff **Deep Learning** verwendet.

Ein vorwärtsgerichtetes neuronales Netzwerk kann als azyklischer gerichteter Graph aufgefasst werden. Die folgende Abbildung zeigt das Knoten-Kanten-Diagramm eines Feedforward-Netzwerks mit zwei verborgenen Schichten:



**Abb. 6.8.** Feedforward-Netzwerk

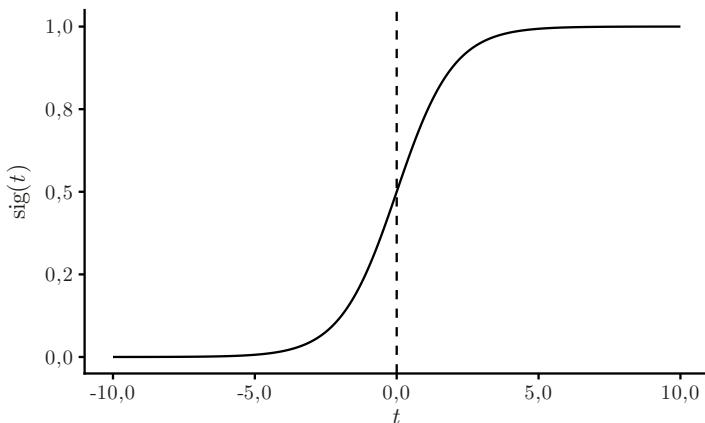
Dabei entspricht jeder Knoten einem Neuron. Die Neuronen einer gegebenen Schicht sind über die Gewichte mit jedem Neuron der darauffolgenden Schicht verknüpft. Diese Verknüpfungen werden als gerichtete Kanten dargestellt und könnten auch als **Synapsen** bezeichnet werden. Die zusätzlichen, mit „1“ bezeichneten Neuronen stellen den Einfluss der Verzerrungsvektoren dar. Die Anzahl der Modellparameter eines neuronalen Netzwerks entspricht somit genau der Anzahl der Kanten im zugehörigen Graphen, im Falle des oben skizzierten Netzwerks sind dies 25 Parameter.

Grundsätzlich kann jeder azyklische Graph als Architektur eines neuronalen Netzwerks verwendet werden. Bei sogenannten **residualen neuronalen Netzwerken** verbinden Synapsen auch Schichten miteinander, die nicht unmittelbar aufeinander folgen.

Eine mögliche Wahl für die Aktivierungsfunktion besteht in einer **Sigmoidfunktion** oder **Fermi-Funktion**:

$$\text{sig}: \mathbb{R} \rightarrow \mathbb{R}, \text{sig}(t) = \frac{1}{1 + e^{-t}} = \frac{1}{2} \cdot \left( 1 + \tanh \frac{t}{2} \right)$$

In der folgenden Abbildung ist der Funktionsgraph dargestellt:



**Abb. 6.9.** Sigmoidfunktion

Dabei ist die Anwendung der Sigmoidfunktion in einer Schicht des neuronalen Netzwerks komponentenweise zu verstehen:

$$\phi_l: \mathbb{R}^{D_l} \rightarrow \mathbb{R}^{D_l}, \phi_l(u) = (\text{sig}(u_1), \text{sig}(u_2), \dots, \text{sig}(u_{D_l}))$$

Die „biologische“ Interpretation: Ein Neuron in der folgenden Schicht wird erst dann aktiviert, wenn das durch die Synapse übertragene Signal die „Reizschwelle“  $t = 0$  überwunden hat.

Eine weitere Möglichkeit besteht in der komponentenweisen Anwendung einer **Gleichrichterfunktion** (engl. *rectifier*) der folgenden Form:

$$\text{rect}_\alpha: \mathbb{R} \rightarrow \mathbb{R}, \text{rect}_\alpha(t) = \begin{cases} t & \text{falls } t > 0 \\ \alpha \cdot t & \text{falls } t \leq 0 \end{cases}$$

Dabei ist  $\alpha \geq 0$  ein Parameter, der wesentlich kleiner als eins gewählt wird;  $\alpha = 0$  und  $\alpha = 0,01$  sind übliche Werte. Im Falle  $\alpha = 0$  handelt es sich um die gewöhnliche Gleichrichterfunktion, bei  $\alpha > 0$  um eine **durchlässige Gleichrichterfunktion** (engl. *leaky rectifier* [23]).

Die Gleichrichterfunktion ist an der Stelle  $t = 0$  nicht differenzierbar. Da für viele Optimierungsverfahren eine Ableitung auch von den Aktivierungsfunktionen berechnet werden muss, ist die folgende geglättete Variante der Gleichrichterfunktion in Gebrauch, die insbesondere für  $\alpha = 0$  auch **Softplus-Funktion** genannt wird:

$$\text{splus}_\alpha(t) = \alpha t + (1 - \alpha) \cdot \ln(1 + e^t)$$

Die **Softmax-Funktion** ist schließlich ein weiteres wichtiges Beispiel für eine Aktivierungsfunktion. Diese wird in der Ausgabeschicht eines für Klassifikationsaufgaben verwendeten neuronalen Netzes verwendet und ist wie folgt definiert:

$$\text{smax}: \mathbb{R}^K \rightarrow \mathbb{R}^K, \text{smax}(u) = \left( \sum_{k=1}^K e^{u_k} \right)^{-1} \cdot (e^{u_1}, e^{u_2}, \dots, e^{u_K})$$

Die Softmax-Funktion hat die Eigenschaft, dass ihre Werte stets als eine Massenfunktion über den Ausgabeneuronen interpretiert werden können:

$$(\text{smax}(u))_k > 0, \sum_{l=1}^K (\text{smax}(u))_l = 1$$

für alle  $k \in \{1, \dots, K\}$  und  $u \in \mathbb{R}^K$ .

#### 6.4.1 Regression und Klassifikation mittels neuronaler Netzwerke

Die Modellparameter eines neuronalen Netzwerks sind die Einträge der Gewichtsmatrizen  $w^{(1)}, \dots, w^{(L)}$  bzw. der Verzerrungsvektoren  $b^{(1)}, \dots, b^{(L)}$  aller Schichten. Die Einträge der Gewichtsmatrix  $w^{(l)}$  bzw. des Verzerrungsvektors  $b^{(l)}$  in der  $l$ -ten Schicht,  $1 \leq l \leq L$ , bezeichnen wir mit  $w^{(l)}_{ji}$  bzw.  $b^{(l)}_j$ , wobei  $i \in \{1, \dots, D_{l-1}\}$  und  $j \in \{1, \dots, D_l\}$ . Der Übersichtlichkeit der Formeln halber fassen wir all diese Parameter auch mit dem Buchstaben „ $\theta$ “ zusammen.

Das Training besteht in der Minimierung der Zielfunktion, die durch den durchschnittlichen Verlust über die Trainingsbeispiele gegeben ist:

$$R(\theta) = \frac{1}{N} \sum_{n=1}^N \lambda(y_n, f(x_n; \theta))$$

Dabei sind  $(x_1, y_1), \dots, (x_N, y_N)$  die vorklassifizierten bzw. vorbewerteten Trainingsbeispiele und  $(y, \hat{y}) \mapsto \lambda(y, \hat{y})$  ist die Verlustfunktion.

Wird das neuronale Netzwerk für Regressionsaufgaben verwendet, so sind auch hier wieder  $\lambda_2(y, \hat{y}) = (y - \hat{y})^2$  oder  $\lambda_1(y, \hat{y}) = |y - \hat{y}|$  geeignete Verlustfunktionen. Das denkbar einfachste neuronale Netzwerk besteht aus genau einem Ausgabeneuron, weist sonst keine verborgenen Schichten auf, und als Aktivierungsfunktion wird die identische Abbildung  $\phi: u \mapsto u$  verwendet. Unter Verwendung der quadratischen Verlustfunktion ist die zu minimierende Zielfunktion dann wie folgt gegeben:

$$\begin{aligned} R(w_0, \dots, w_D) &= \frac{1}{N} \sum_{n=1}^N (y_n - w_0 - w \cdot x_n)^2 \\ &= \frac{1}{N} \sum_{n=1}^N \left( y_n - w_0 - \sum_{d=1}^D w_d x_{nd} \right)^2 \end{aligned}$$

Da es nur eine Schicht gibt, haben wir die Gewichtsmatrix  $w^{(1)} = (w^{(1)}_1, \dots, w^{(1)}_D)$  als Zeilenvektor von Gewichten  $w = (w_1, \dots, w_D)$  abgekürzt, und  $w_0 \in \mathbb{R}$  ist der einzige Eintrag im Verzerrungsvektor  $b^{(1)}$ . Bis auf den Faktor „ $1/N$ “ ist das gerade die Residuenquadratsumme der linearen Regression (siehe Abschn. 6.2.1). Auch die gelernte Regressionsfunktion ist dieselbe. Daher sind beide Verfahren äquivalent und wir können die lineare Regression als ein spezielles neuronales Netzwerk auffassen.

Für Klassifizierungsaufgaben entspricht jedes Neuron der Ausgabeschicht einer der insgesamt  $K$  Klassen, sodass die Breite der Ausgabeschicht  $D_L = K$  beträgt. Eine Softmax-Funktion sorgt dafür, dass die Aktivierung eines Ausgabeneurons als Wahrscheinlichkeit für die jeweilige Klassenzugehörigkeit interpretiert werden kann. Als Verlustfunktion eignet sich in diesem Fall die **Kreuzentropie**:

$$\lambda(y, \hat{y}) = \lambda(y_1, \dots, y_K, \hat{y}_1, \dots, \hat{y}_K) = - \sum_{k=1}^K y_k \ln(\hat{y}_k)$$

Dabei gilt  $y_k = 1$ , wenn das Trainingsbeispiel der  $k$ -ten Klasse zugehörig ist, andernfalls  $y_k = 0$ : Dies ist wieder die One-Hot-Kodierung der Klassenbelegung. Alle Summanden bis auf jenen mit der wahren Klassenzugehörigkeit sind dann gleich null. Ein Training mit unscharf vorklassifizierten Daten ist aber auch möglich; dann ist die Klassenzuordnung nicht binär, sondern abgestuft:  $0 \leq y_k \leq 1$  für alle  $k \in \{1, \dots, K\}$ . Ein Verfahren der Datenvorverarbeitung, das auf eine abgestufte Vorklassifikation führt, ist die **Labelglättung** (engl. *label smoothing*) [24]. Dabei wird den übrigen Klassen eine kleine Wahrscheinlichkeit der Zugehörigkeit zugestanden, dies kann zu einer verbesserten Verallgemeinerbarkeit des Modells führen [25]:

$$y_{k,\varepsilon} = (1 - \varepsilon) \cdot y_k + \frac{\varepsilon}{K}$$

Dabei ist  $\varepsilon > 0$  ein klein gewählter Glättungsparameter.

Unabhängig davon, ob die Klassenlabels der Trainingsdaten unscharf sind oder nicht, ist die Ausgabe des neuronalen Netwerks in aller Regel keine „harte“ Klassenzuordnung, sondern eine Verteilung von Klassenzugehörigkeitswahrscheinlichkeiten. Die finale Entscheidungsregel kann durch Auswahl der Klasse mit größter Zugehörigkeitswahrscheinlichkeit erklärt werden.

Für eine binäre Klassifikation kann als Ausgabeschicht auch ein einzelnes Sigmoid-aktiviertes Neuron verwendet werden, um eine Klassenzugehörigkeitswahrscheinlichkeit zwischen null und eins zu garantieren. In dem Fall nimmt die Kreuzentropie die folgende Form an:

$$\lambda(y, \hat{y}) = -y \ln \hat{y} - (1 - y) \ln(1 - \hat{y})$$

Es ist instruktiv, ein solches Netzwerk mit dieser Ausgabeschicht bzw. Verlustfunktion zu betrachten, das ansonsten keine verborgene Schichten aufweist. Die Ausgabe dieses sehr flachen Netzwerks nimmt, bei Eingabe  $u \in \mathbb{R}^D$ , die folgende Form an:

$$\begin{aligned} f(u; w_0, \dots, w_D) &= \text{sig}(w \cdot u + w_0) \\ &= \frac{1}{1 + e^{-(w \cdot u + w_0)}} \end{aligned}$$

Dabei ist  $w = (w_1, \dots, w_D)$  ein Zeilenvektor von Gewichten, und  $w_0 \in \mathbb{R}$  ist der einzige Eintrag im Verzerrungsvektor.

Das empirische Risiko, gegeben ein Trainingsdatensatz  $x_1, \dots, x_N \in \mathbb{R}^D$ , ist daher die folgende Funktion in den Modellparametern:

$$\begin{aligned} N \cdot R(w_0, \dots, w_D) &= \sum_{k=1}^N \lambda(y_n, f(x_n; w_0, \dots, w_D)) \\ &= \sum_{k=1}^N \lambda(y_n, \text{sig}(w \cdot x_n + w_0)) \\ &= - \sum_{k=1}^N y_n \ln(\text{sig}(w \cdot x_n + w_0)) - \\ &\quad \sum_{k=1}^N (1 - y_n) \ln(\text{sig}(-(w \cdot x_n + w_0))) \\ &= - \sum_{k=1}^N \ln(\text{sig}((-1)^{y_n} (w \cdot x_n + w_0))) \\ &= \sum_{k=1}^N \ln(1 + \exp((-1)^{y_n} (w \cdot x_n + w_0))) \\ &= \sum_{n=1}^N \ln \left( 1 + \exp \left( (-1)^{y_n} \cdot \sum_{d=0}^D w_d x_{nd} \right) \right) \end{aligned}$$

Dabei haben wir im letzten Schritt die Eingabedaten als Zeilen einer erweiterten Datenmatrix  $\mathbf{X} = (x_{nd})$  aufgefasst und davor die folgende Eigenschaft der Sigmoidfunktion ausgenutzt:

$$1 - \text{sig}(t) = 1 - \frac{1}{1 + e^{-t}} = \frac{e^{-t}}{1 + e^{-t}} = \frac{1}{1 + e^t} = \text{sig}(-t)$$

Ein Vergleich mit der Log-Likelihood-Funktion  $\ell(\cdot)$  der logistischen Regression (Abschn. 6.3.1) ergibt:  $\ell(w_0, \dots, w_D) = -NR(w_0, \dots, w_D)$ . Eine Maximierung dieser Log-Likelihood-Funktion entspricht also genau der Minimierung des Trainingsfehlers dieses einfachen neuronalen Netzwerks. Schließlich sind auch die Entscheidungsregeln identisch, denn aufgrund von  $\text{sig}(t) > 1/2 \Leftrightarrow t > 0$  gilt:

$$f(x_*; \hat{w}_0, \dots, \hat{w}_D) > \frac{1}{2} \Leftrightarrow \sum_{d=0}^D \hat{w}_d x_{*d} > 0$$

Es lohnt sich, obige Ergebnisse hervorzuheben.

**Lineare und logistische Regression als spezielle neuronale Netzwerke.** Das Verfahren der **linearen Regression** ist äquivalent zu einem Feedforward-Netzwerk ohne verborgene Schichten mit einem einzelnen trivial aktivierten Ausgabeneuron unter Verwendung der quadratischen Verlustfunktion.

Das Verfahren der **logistischen Regression** (ohne Einsatz der Kernel-Methode) ist äquivalent zu einem Feedforward-Netzwerk ohne verborgene Schichten mit einem einzelnen Sigmoid-aktivierten Ausgabeneuron unter Verwendung der Kreuzentropie als Verlustfunktion.

Abb. 6.11 zeigt die Ergebnisse einer Klassifikation auf zwei ähnlichen synthetischen Datensätzen mithilfe eines neuronalen Netzwerks. Die hierfür verwendete Architektur ist die in Abb. 6.8 dargestellte, also zwei verborgene Schichten mit vier bzw. zwei Neuronen; als Aktivierungsfunktionen wurden durchweg Sigmoidfunktionen verwendet. Die Entscheidungsgrenze ist augenscheinlich keine Gerade: Durch das Hinzufügen verborgener Schichten ist das Netzwerk in der Lage nichtlineare Klassifikatoren zu lernen – im Gegensatz zur logistischen Regression, deren Ergebnis ohne Verwendung der Kernel-Methode nur ein linearer Klassifikator sein kann.

Die grundlegenden Modelle der linearen und logistischen Regression sind Teil des Hypothesenraums der neuronalen Netzwerke. Welche weiteren Modelle können als ein neuronales Netzwerk dargestellt werden? Es zeigt sich, dass dies in einem gewissen Sinne *alle* Modelle sind: Jede beliebige stetige Funktion kann durch ein geeignetes neuronales Netzwerk näherungsweise dargestellt werden. Ohne Beweis führen wir die folgenden Lehrsätze an, die diese Aussage mathematisch präzisieren.

**Universelle Approximationseigenschaft neuronaler Netzwerke beliebiger Breite [26, Theorem 3.1].** Sei  $\phi: \mathbb{R} \rightarrow \mathbb{R}$  eine stetige Funktion, und sei  $\mathcal{W}(\phi)$  die Menge aller Funktionen der folgenden Form:

$$f: \mathbb{R}^D \rightarrow \mathbb{R}, f(u) = \sum_{k=1}^K w^{(2)}_k \cdot \phi \left( \sum_{d=1}^D w^{(1)}_{kd} \cdot u_d + b_k \right)$$

mit Parametern  $K \in \mathbb{N}$  und  $w^{(1)}_k, w^{(2)}_k, b_k \in \mathbb{R}$ . Der Hypothesenraum  $\mathcal{W}(\phi)$  besteht aus neuronalen Netzwerken mit genau einer verborgenen,  $\phi$ -aktivierten Schicht von potenziell unbegrenzter Breite.

Falls  $\phi$  kein Polynom ist, dann gibt es für jedes  $\varepsilon > 0$  und jede auf einer kompakten Menge  $K \subset \mathbb{R}^D$  (zum Beispiel  $K = [0, 1]^D$ ) definierten stetigen Funktion  $g: K \rightarrow \mathbb{R}$  eine Funktion  $f \in \mathcal{W}(\phi)$  mit der folgenden Eigenschaft:

$$\sup_{u \in K} |f(u) - g(u)| < \varepsilon$$

Jede auf einem Kompaktum definierte stetige Funktion kann also durch neuronale Netzwerke aus  $\mathcal{W}(\phi)$  gleichmäßig approximiert werden. Umgekehrt gilt: Besitzt  $\mathcal{W}(\phi)$  diese **universelle Approximationseigenschaft**, so kann  $\phi$  kein Polynom sein.

Ein entsprechender Lehrsatz kann auch für tiefe Netzwerke mit begrenzter Breite bewiesen werden.

**Universelle Approximationseigenschaft neuronaler Netzwerke beliebiger Tiefe und begrenzter Breite [27, Theorem 3.2].** Sei  $\phi: \mathbb{R} \rightarrow \mathbb{R}$  eine stetige Funktion, die in wenigstens einem Punkt stetig differenzierbar ist und dort nichtverschwindende Ableitung hat.

Sei weiterhin  $\mathcal{D}(\phi)$  die Menge aller Feedforward-Netzwerke  $f: \mathbb{R}^D \rightarrow \mathbb{R}$  beliebiger Tiefe mit  $\phi$ -aktivierten verborgenen Schichten, wobei keine Schicht aus mehr als  $D + 3$  Neuronen besteht.

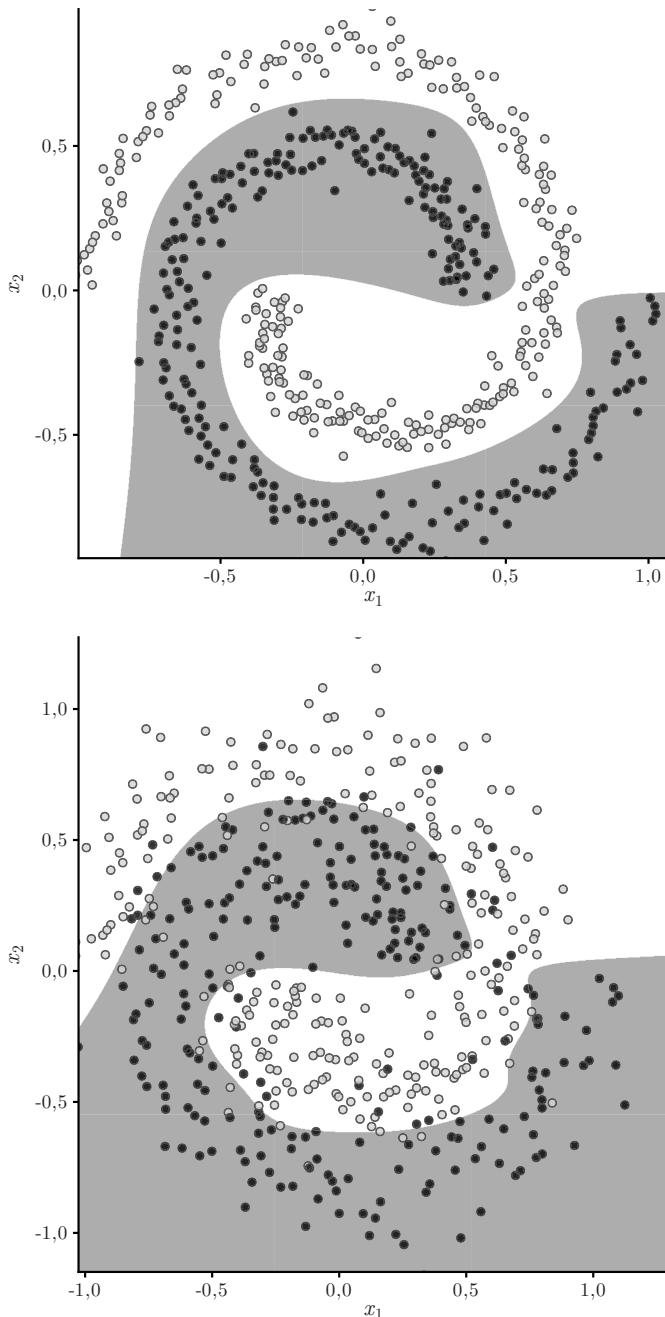
Wenn  $\phi$  keine affin-lineare Funktion darstellt – also nicht gerade von der Form  $\phi(u) = m \cdot u + c$  ist – dann besitzt  $\mathcal{D}(\phi)$  die universelle Approximationseigenschaft: Jede auf einem Kompaktum definierte stetige Funktion  $g: \mathbb{R}^D \supset K \rightarrow \mathbb{R}$  kann durch Funktionen aus  $\mathcal{D}(\phi)$  gleichmäßig approximiert werden.

#### 6.4.2 Training neuronaler Netzwerke durch Fehlerrückführung

Um den Trainingsfehler mithilfe von Gradientenabstiegsverfahren zu minimieren und so die optimalen Modellparameter zu ermitteln, muss mit jedem Iterationsschritt der Gradient einer Funktion der Form



**Abb. 6.10.** Neuronen und interneuronale Synapsen des Fadenwurms *Caenorhabditis elegans* (*Hermaphrodite*)



**Abb. 6.11.** Klassifikation mit neuronalem Netzwerk

$$R(\theta) = \frac{1}{M} \sum_{k=1}^M \lambda(y_{n_k}, f(x_{n_k}; \theta))$$

berechnet werden. Je nachdem, ob ein Gradientenabstieg, ein stochastischer Gradientenabstieg oder eine Mini-Batch-Optimierung vorgenommen wird, gilt  $M = N$ ,  $M = 1$  oder  $1 < M < N$ , wobei  $N$  die Gesamtzahl der Trainingsbeispiele darstellt.

Die jeweilige Zielfunktion ist also im Wesentlichen durch eine Summe über den Verlust gegeben, sodass wir den Gradienten ebenfalls als Summe der Gradienten jedes Summanden schreiben können. Mithin betrachten wir im Folgenden nur den Beitrag eines einzelnen Trainingsbeispiels  $(x_n, y_n)$ , wir sind also an der Berechnung des Gradienten der Funktion

$$R_n(\theta) = \lambda(y_n, f(x_n; \theta))$$

interessiert. Der Übersichtlichkeit halber schreiben wir im Folgenden statt obigem Ausdruck  $r(\theta) = \lambda(y, f(x; \theta))$ .

Der Gradient von  $r(\cdot)$  wird durch ein effizientes Rechenschema ermittelt, das als **Fehlerrückführung** oder mit dem englischen Fachbegriff **Backpropagation** bezeichnet wird.

Bevor wir dieses beschreiben, führen wir zunächst ein paar Abkürzungen ein. Den Vektor der Aktivierungen in der  $l$ -ten Schicht bezeichnen wir mit  $a^{(l)}$ , es gilt also

$$a^{(l)} = (f_l \circ f_{l-1} \circ \dots \circ f_1)(x)$$

für  $1 \leq l \leq L$  sowie  $a^{(0)} = x$ . Weiterhin führen wir für jede Schicht den Vektor der gewichteten Eingaben

$$z^{(l)} = w^{(l)} \cdot a^{(l-1)} + b^{(l)}$$

ein, es gilt also insbesondere  $a^{(l)} = \phi_l(z^{(l)})$ , wenn  $\phi_l$  die Aktivierungsfunktion der  $l$ -ten Schicht ist. Die Verlustfunktion wird wie üblich mit  $(y, \hat{y}) \mapsto \lambda(y, \hat{y})$  notiert.

Eine **Fehlerrückführung** für die Berechnung des Gradienten der Zielfunktion eines Feedforward-Netzwerks besteht in der Durchführung folgender Rechenschritte (vgl. [28, Kap. 2]):

1. **Eingabe.** Belege die Aktivierung der Eingabeschicht,  $a^{(0)} = x$ , und setze die aktuellen Werte der Modellparameter  $w_{ij}^l$  und  $b_j^l$ .
2. **Feedforward.** Für  $l = 1, \dots, L$ , berechne nacheinander die gewichteten Eingaben und Aktivierungen:  $z^{(l)} = w^{(l)} \cdot a^{(l-1)} + b^{(l)}$  und  $a^{(l)} = \phi_l(z^{(l)})$ .
3. **Ausgabefehler.** Berechne  $\Delta^{(L)} := (\mathbf{D}\phi_L(z^{(L)}))^T \cdot \nabla_{\hat{y}} \lambda(y, a^{(L)})$ .

4. **Fehlerrückführung.** Für  $l = L - 1, L - 2, \dots, 1$ , berechne nacheinander  $\Delta^{(l)} = (\mathbf{D}\phi_l(z^{(l)}))^T \cdot (w^{(l+1)})^T \cdot \Delta^{(l+1)}$ .

5. **Ausgabe.** Der Gradient der Zielfunktion ist schließlich durch die folgenden partiellen Ableitungen gegeben:

$$\frac{\partial r}{\partial w^{(l)}_{ji}} = \Delta^{(l)}_j \cdot a^{(l-1)}_i, \quad \frac{\partial r}{\partial b^{(l)}_j} = \Delta^{(l)}_j$$

für alle  $l \in \{1, \dots, L\}$ ,  $i \in \{1, \dots, D_{l-1}\}$  und  $j \in \{1, \dots, D_l\}$ .

Das Verfahren der Fehlerrückführung basiert zum einen auf der Erkenntnis, dass die Änderung der Verlustfunktion mit den Gewichten und den Einträgen des Verzerrungsvektors in der  $l$ -ten Schicht durch die Änderung der gewichteten Eingaben allein in dieser Schicht ausgedrückt werden kann. Aus der Kettenregel für partielle Ableitungen folgt:

$$\begin{aligned} \frac{\partial r}{\partial w^{(l)}_{ji}} &= \sum_{k=1}^{D_l} \frac{\partial r}{\partial z^{(l)}_k} \cdot \frac{\partial z^{(l)}_k}{\partial w^{(l)}_{ji}} \\ &= \sum_{k=1}^{D_l} \frac{\partial r}{\partial z^{(l)}_k} \cdot \frac{\partial}{\partial w^{(l)}_{ji}} \left( \sum_{m=1}^{D_{l-1}} w^{(l)}_{km} a^{(l-1)}_m + b^{(l)}_k \right) \\ &= \frac{\partial r}{\partial z^{(l)}_j} \cdot a^{(l-1)}_i, \\ \frac{\partial r}{\partial b^{(l)}_j} &= \sum_{k=1}^{D_l} \frac{\partial r}{\partial z^{(l)}_k} \cdot \frac{\partial z^{(l)}_k}{\partial b^{(l)}_j} = \frac{\partial r}{\partial z^{(l)}_j} \end{aligned}$$

Die Aktivierungen  $a^{(l-1)}_i$  werden während des Feedforward-Schritts berechnet. Die partiellen Ableitungen  $\frac{\partial r}{\partial z^{(l)}_j}$  fassen wir zu einem Gradientenvektor  $\Delta^{(l)} := \nabla_{z^{(l)}} r$  zusammen. Für die Jacobi-Matrix, also den transponierten Gradientenvektor, gilt vermöge der Kettenregel:

$$\begin{aligned} (\Delta^{(L)})^T &= \mathbf{D}_{\hat{y}} \lambda(y, a^{(L)}) \cdot \mathbf{D}\phi_L(z^{(L)}), \\ (\Delta^{(L-1)})^T &= \mathbf{D}_{\hat{y}} \lambda(y, a^{(L)}) \cdot \mathbf{D}\phi_L(z^{(L)}) \cdot w^{(L)} \cdot \mathbf{D}\phi_{L-1}(z^{(L-1)}), \\ &\vdots \\ (\Delta^{(l)})^T &= \mathbf{D}_{\hat{y}} \lambda(y, a^{(L)}) \cdot \mathbf{D}\phi_L(z^{(L)}) \cdot w^{(L)} \cdot \mathbf{D}\phi_{L-1}(z^{(L-1)}) \cdots \\ &\quad \cdot w^{(l+2)} \cdot \mathbf{D}\phi_{l+1}(z^{(l+1)}) \cdot w^{(l+1)} \cdot \mathbf{D}\phi_l(z^{(l)}) \\ &\quad \vdots \\ (\Delta^{(1)})^T &= \mathbf{D}_{\hat{y}} \lambda(y, a^{(L)}) \cdot \mathbf{D}\phi_L(z^{(L)}) \cdot w^{(L)} \cdot \mathbf{D}\phi_{L-1}(z^{(L-1)}) \cdots \\ &\quad \cdot w^{(3)} \cdot \mathbf{D}\phi_2(z^{(2)}) \cdot w^{(2)} \cdot \mathbf{D}\phi_1(z^{(1)}) \end{aligned}$$

Es werden also sukzessive die inneren Ableitungen von rechts heranmultipliziert, um die entscheidenden Fehlerterme  $\Delta^{(l)}$  zu produzieren: Der Fehler wird von der Ausgabeschicht hin zur Eingabeschicht zurückgeführt. Eine Multiplikation von links ist die üblichere Darstellung der Fehlerrückführung, dies kann durch Transponieren erreicht werden:

$$\begin{aligned}\Delta^{(L)} &= (\mathbf{D}\phi_L(z^{(L)}))^T \cdot \nabla_{\hat{y}} \lambda(y, a^{(L)}), \\ \Delta^{(L-1)} &= (\mathbf{D}\phi_{L-1}(z^{(L-1)}))^T \cdot (w^{(L)})^T \cdot \Delta^{(L)}, \\ &\vdots \\ \Delta^{(l)} &= (\mathbf{D}\phi_l(z^{(l)}))^T \cdot (w^{(l+1)})^T \cdot \Delta^{(l+1)} \\ &\vdots \\ \Delta^{(1)} &= (\mathbf{D}\phi_1(z^{(1)}))^T \cdot (w^{(2)})^T \cdot \Delta^{(2)}\end{aligned}$$

Ein **Dropout** bzw. eine „Ausdünnung“ (engl. *dilution*) des neuronalen Netzwerks ist eine stochastische Anpassung der Fehlerrückführung und stellt eine effektive und einfache Methode dar, mit der eine Überanpassung vermieden werden soll. Die Idee besteht darin, während des Trainings nur einen Teil des Netzwerks zu verwenden, der mit jedem Iterationsschritt zufällig ausgewählt wird [29].

Bei einem **Dropout** mit Ausfallwahrscheinlichkeiten  $0 \leq q_l < 1$  für jede verborgene Schicht  $l \in \{1, \dots, L-1\}$  wird der Algorithmus der Fehlerrückführung wie folgt angepasst:

1. Deaktiviere mit jeder Iteration im Feedforward-Schritt eine zufällige Auswahl von Neuronen:
  - a) Für jeden Index  $d \in \{1, \dots, D_l\}$  treffe eine zufällige Auswahl von  $\rho^{(l)}_d \in \{0, 1\}$  mit unabhängiger Wahrscheinlichkeit von  $q_l$  für den Ausgang  $\rho^{(l)}_d = 0$ .
  - b) Setze  $z^{(l)} = w^{(l)} \cdot a^{(l-1)} + b^{(l)}$  und  $(a^{(l)})_d = \rho^{(l)}_d \cdot (\phi_l(z^{(l)}))_d$  für alle  $d \in \{1, \dots, D_l\}$ .
2. Bei der Berechnung der rückgeführten Fehler bleiben deaktivierte Neuronen unberücksichtigt. Entsprechend werden nur Gewichte  $w^{(l)}_{ji}$  bzw. Verzerrungen  $b^{(l)}_j$  aktualisiert, für die  $\rho^{(l-1)}_i = \rho^{(l)}_j = 1$  bzw.  $\rho^{(l)}_j = 1$  gilt.
3. Nach Abschluss des Trainings werden alle Gewichte in der  $l$ -ten Schicht um dem Faktor  $1 - q_l$  skaliert.

Die Ausfallwahrscheinlichkeiten  $q_1, \dots, q_{L-1}$  können als neue Hyperparameter des Modells aufgefasst werden. Bei der Anwendung des trainierten Modells auf die Testdaten bleiben alle Neuronen aktiviert.

### 6.4.3 Convolutional Neural Networks

**Convolutional Neural Networks** (zu deutsch etwa: „faltende neuronale Netzwerke“) sind eine besondere Variante von neuronalen Netzwerken. Die automatisierte Klassifikation digitaler Bilder stellt ein hauptsächliches Einsatzgebiet solcher Netzwerke dar. Wir stellen in diesem Abschnitt Grundidee und Kernkomponenten vor.

Ein digitales Bild kann als ein dreidimensionales Raster oder Gitter von Bildpunkten  $M$  der Breite  $B$  und der Höhe  $H$  sowie einer „Tiefe“ von  $T$  Farbkanälen angesehen werden. Jedem Pixel mit Position  $(i, j)$  wird im Farbkanal  $t$  ein Farbwert  $M(i, j, t)$  zugeordnet,  $j \in \{1, \dots, B\}$ ,  $i \in \{1, \dots, H\}$ ,  $t \in \{1, \dots, T\}$ . Ein hochauflösendes Farbbild könnte z. B. eine Breite von  $B = 1920$  Pixeln, eine Höhe von  $H = 1080$  Pixeln und eine „Tiefe“ von  $T = 3$  Farbkanälen aufweisen. Gewöhnlich werden digitale Bilder im Format „Breite mal Höhe“ angegeben. Um in Übereinstimmung mit der mathematischen Konvention für die Angabe von Matrizenformaten zu bleiben, nennen wir im Folgenden jedoch die Anzahl der Bildzeilen, also die Höhe, zuerst.

Jedes abstrakte Zahlenraster mit drei Dimensionen – unabhängig davon, ob es ein digitales Bild darstellt – wollen wir eine **Merkmalskarte** nennen. Bei einem Convolutional Neural Network wird das Quellbild mit jeder Schicht in neue Merkmalskarten umgewandelt. Diese Umwandlung findet vermöge zweier zentraler Operationen statt, die wir im Folgenden beschreiben wollen: Die **Faltung** und das **Pooling**.

Eine **Filtermaske** oder **Faltungskern** der Bandbreite  $2F + 1$ ,  $F \in \mathbb{N}$ , ist eine quadratische Merkmalskarte vom Format  $(2F + 1) \times (2F + 1) \times T$ . Typische Werte für in der Praxis verwendete Filtermasken sind  $F = 1$  oder  $F = 2$ . Bei Filtermasken wollen wir Breite und Höhe symmetrisch um Null herum mit  $-F, \dots, -1, 0, 1, \dots, F$  indizieren, das macht die Definition der folgenden Operation übersichtlicher.

Die **zweidimensionale Faltung** (kurz: **2D-Faltung**) einer Merkmalskarte oder Bildes  $M$  vom Format  $H \times B \times T$  und einem Filter  $\kappa$  vom Format  $(2F + 1) \times (2F + 1) \times T$  führt auf eine Karte  $M \star \kappa$  vom Format  $H \times B \times 1$  und besteht in der folgenden Operation:

$$(M \star \kappa)(i, j, 1) = \sum_{t=1}^T \sum_{f_1=-F}^F \sum_{f_2=-F}^F M(i + f_1, j + f_2, t) \cdot \kappa(f_1, f_2, t)$$

für alle  $j \in \{1, \dots, B\}$ ,  $i \in \{1, \dots, H\}$ ,  $t \in \{1, \dots, T\}$ . Wo der Summenindex  $(i + f_1, j + f_2)$  aus der Merkmalskarte bzw. Bildfläche herausläuft, nehmen wir an, dass diese dort den Wert Null hat.

Die am Ende erwähnte Rechenvorschrift kann auch so verstanden werden, dass wir das zu filternde Bild zuvor an den Rändern durch schwarze bzw. mit einem Farbwert von null belegte Pixel fortsetzen (**Zero-Padding**).

Jede Faltung ist eine lineare Abbildung über dem Vektorraum der Merkmalskarten eines festen Formats; es gilt für eine Filtermaske  $\kappa$ , beliebige Karten  $M_1, M_2$  und Skalare  $\lambda \in \mathbb{R}$ :

$$(M_1 + M_2) * \kappa = M_1 * \kappa + M_2 * \kappa, (\lambda \cdot M_1) * \kappa = \lambda \cdot (M_1 * \kappa)$$

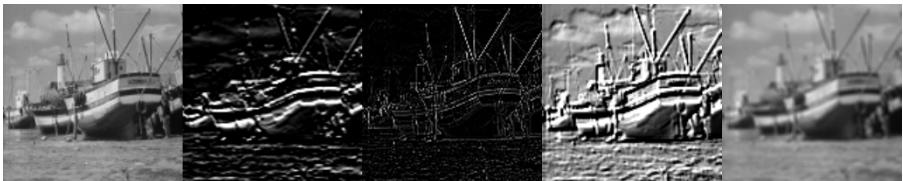
Die Addition und Skalarmultiplikation ist dabei „pixelweise“ mit jedem Eintrag der Karte zu verstehen.

Obige Definition der Faltungsoperation ist in der Literatur zu Deep Learning verbreitet. In der allgemeinen Signal- oder Bildverarbeitung wird sie allerdings oft auch als **Kreuzkorrelation** bezeichnet, während stattdessen die folgende Operation für Bilder vom Format  $H \times B$  gewöhnlich als Faltung bekannt ist:

$$(M * k)(i, j) = \sum_{f_1=-F}^F \sum_{f_2=-F}^F M(i - f_1, j - f_2) \cdot \kappa(f_1, f_2)$$

Faltung und Kreuzkorrelation gehen durch Spiegelung der Filtermaske an der zentralen Zeile und Spalte ineinander über. Diese Feinheit spielt bei der Implementierung in einem Convolutional Neural Network jedoch eine untergeordnete Rolle, da die Einträge der Filtermaske ohnehin während des Trainings angepasst werden. Nichtsdestotrotz können die verschiedenen Konventionen beim Studium des Themas für Verwirrung sorgen.

Die folgenden Bilder gehen aus dem Originalfoto [30] ganz links durch Anwendung verschiedener, in der Bildverarbeitung gebräuchlicher Filtermasken hervor:



**Abb. 6.12.** Faltungsfilter in der Bildverarbeitung

Im Einzelnen sind die angewendeten Filtermasken (in der Abbildung von links nach rechts) die folgenden:

$$\begin{aligned} \kappa_1 &= \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}, \kappa_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix} \\ \kappa_3 &= \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \\ 0 & -1 & -2 \end{pmatrix}, \kappa_4 = \frac{1}{9} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{aligned}$$

Die ersten beiden Filtermasken sind der **horizontale Sobel-Operator** und der **diskrete Laplace-Operator**. Beide dienen in der Bildverarbeitung der

**Kantendetektion.** Darauf folgen der **Relieffilter** und der **Mittelwertfilter**. Der Mittelwertfilter berechnet den durchschnittlichen Farb-/Grauwert in der Umgebung jedes Pixels, was zu einer Weichzeichnung des Bildes führt.

Bei allen Bildfiltrern, die auf einer Faltung beruhen, bestimmt lediglich die Nachbarschaft eines Pixels den Wert des gefilterten Bildes an dieser Stelle. Das Ergebnis sind Merkmalskarten, die *lokale* Bildmerkmale wie z. B. Bildkanten widerspiegeln. Die Grundidee eines Convolutional Neural Networks besteht darin, es selbstständig erlernen zu lassen, welche Filter bzw. Merkmale am besten für eine Charakterisierung und Klassifikation der Bildinformationen geeignet sind.

Eine Reduktion der Auflösung stellt eine weitere wichtige Komponente eines Convolutional Neural Networks dar. Eine Möglichkeit ist die Einführung einer größeren **Schrittweite**  $s > 1$  der Faltung (engl. *stride*), bei der Pixel übersprungen werden:

$$(M \star_s \kappa)(i, j) = \sum_{f_1=-F}^F \sum_{f_2=-F}^F M(1 + s \cdot (i - 1) + f_1, 1 + s \cdot (j - 1) + f_2) \cdot \kappa(f_1, f_2)$$

mit  $j \in \{1, \dots, \lfloor B/s \rfloor\}$ ,  $i \in \{1, \dots, \lfloor H/s \rfloor\}$ .

Eine weitere Methode der Datenreduktion besteht in die Auswahl der größten Aktivierung in Ausschnitten der Merkmalskarte.

Ein **Maximums-Pooling** einer Merkmalskarte  $M$  vom Format  $B \times H \times T$ , mit geraden Zahlen  $B$  und  $H$ , besteht in der folgenden Operation:

$$\text{max-pool}(M)(i, j, t) = \max_{k, l \in \{0, 1\}} \{M(2i - 1 + k, 2j - 1 + l, t)\}$$

für alle  $j \in \{1, \dots, \lfloor B/2 \rfloor\}$ ,  $i \in \{1, \dots, \lfloor H/2 \rfloor\}$ ,  $t \in \{1, \dots, T\}$ .

Die oben definierten Operationen auf Bildern/Merkmalskarten stellen wesentliche Komponenten der Architektur eines **Convolutional Neural Networks** dar. Ein solches Netzwerk besteht aus einer Anordnung der folgenden Typen von Schichten:

- Eine **Convolution-Schicht** besteht in der Faltung der Eingabekarte  $M$  vom Format  $H \times B \times T$  mit Filtermasken  $\kappa_1, \dots, \kappa_K$  vom Format  $(2F + 1) \times (2F + 1) \times T$ , wobei mit jeder Faltung noch eine globale Verzerrung  $b_1, \dots, b_K \in \mathbb{R}$  zu allen Einträgen addiert wird. Anschließend wird, in der Regel ebenfalls „pixelweise“ mit jedem Eintrag der Merkmalskarte, eine Aktivierungsfunktion  $\phi$  angewendet. Hierdurch entstehen  $K$  neue Bilder bzw. Merkmalskarten vom Format  $H \times B \times 1$ :

$$\phi(M \star \kappa_1 + b_1), \dots, \phi(M \star \kappa_K + b_K)$$

Diese werden in der Tiefe zu einer neuen Merkmalskarte vom Format  $H \times B \times K$  zusammengeführt. Die Einträge der Filtermasken und die Verzerrungen stellen die zu lernenden Modellparameter dar.

- Eine **Pooling-Schicht**, in der zwecks Datenreduktion ein Maximums-Pooling durchgeführt wird.
- Eine **voll vernetzte Schicht** besteht zunächst einmal in einer **Flattening-Operation**: Die Eingabekarte  $M$  wird zu einem Vektor  $\text{Flatten}(M)$  der Länge  $D = H \cdot B \cdot T$  „aufgerollt“. Hiernach folgt eine affine Transformation und Anwendung einer Aktivierungsfunktion, wie wir sie schon von gewöhnlichen Feedforward-Netzwerken kennen:

$$\phi(w \cdot \text{Flatten}(M) + b)$$

mit einer Matrix von Gewichten  $w$  vom Format  $K \times D$  und einem Verzerrungsvektor  $b$  der Länge  $K$ .

In folgender Tabelle sind – bei einem Format der Eingabekarte von  $H \times B \times T$  – für jede dieser Schichten die Anzahl der Modellparameter und das Format der Ausgabekarte aufgeführt:

Schichttyp	Anzahl Modellparameter	Ausgabeformat
Convolution-Schicht, $K$ Filtermasken	$K \cdot (T \cdot (2F + 1)^2 + 1)$	$H \times B \times K$
Pooling-Schicht	0	$\lfloor \frac{B}{2} \rfloor \times \lfloor \frac{H}{2} \rfloor \times T$
voll vernetzte Schicht, Ausgabedimension $K$	$K \cdot (H \cdot B \cdot T + 1)$	$1 \times 1 \times K$

**Tabelle 6.4.** Schichttypen eines Convolutional Neural Networks

Eine voll vernetzte Schicht verknüpft Ein- und Ausgabe über eine im Prinzip beliebige affine Abbildung miteinander, wie bei einem gewöhnlichen Feedforward-Netzwerk:

$$M \mapsto w \cdot \text{Flatten}(M) + b$$

Bei einer Convolution-Schicht sind die möglichen Operationen zwar auch affine Abbildungen, jedoch auf die Familie der Faltungen (plus Verzerrung) eingeschränkt:

$$M \mapsto M \star \kappa + b$$

Die Anzahl der zu trainierenden Modellparameter einer Convolution-Schicht ist daher in der Regel wesentlich kleiner als bei einer voll vernetzten Schicht mit gleicher Eingabe.

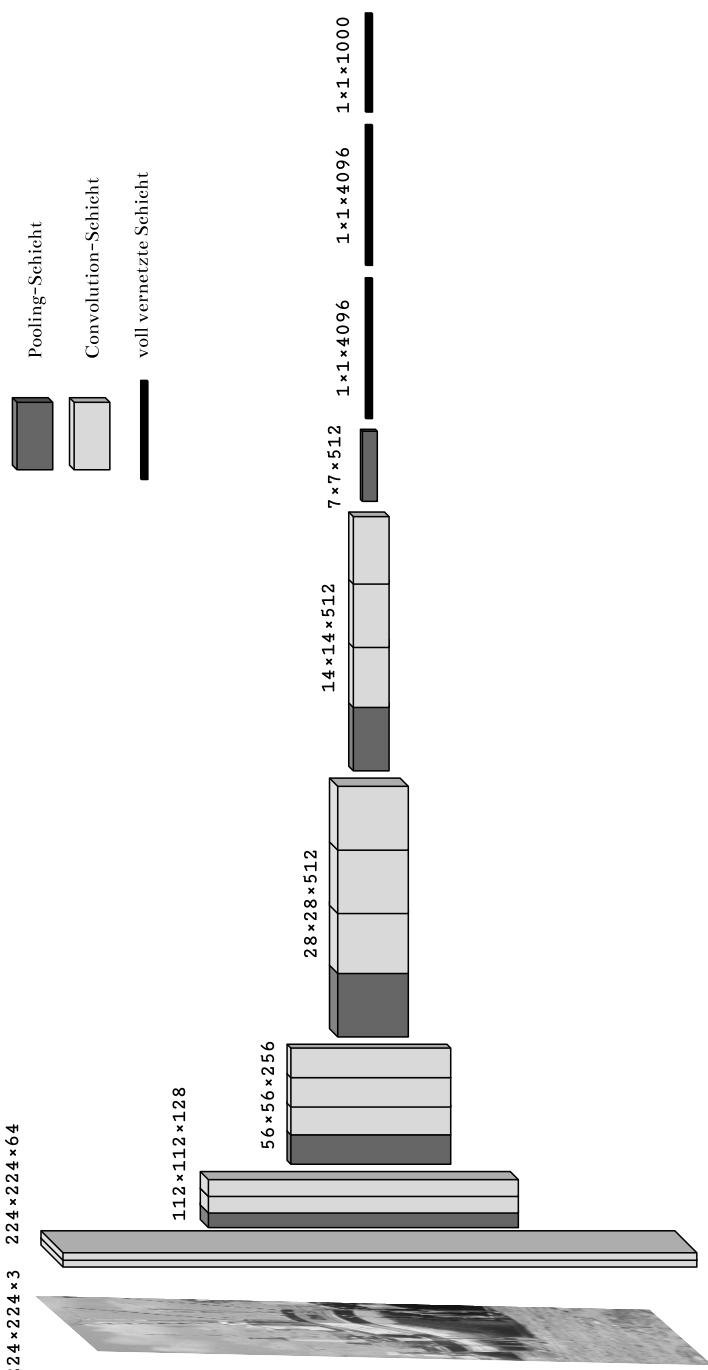
Die typische Architektur eines Convolutional Neural Networks besteht in der abwechselnden Hintereinanderausführung von Convolution- und Pooling-Schichten, deren Ausgabe schließlich durch eine oder mehrere voll vernetzte

Schichten verarbeitet wird. Abb. 6.13 ist eine Darstellung der Architektur VGG-16<sup>4</sup> [31], die noch 2014 im Vergleich mit anderen Verfahren auf dem Datensatz ImageNet [2, 32] hohe Korrektklassifikationsraten von 74,4 % erzielte [33]. Architekturen auf dem aktuellen Stand der Technik (2020) erzielen Korrektklassifikationsraten jenseits der 85 % und basieren auf residualen Architekturen sowie effektiver Hyperparameteroptimierung (Breite, Tiefe und Auflösung) und Datenaugmentation [34, 35, 36, 37].

---

<sup>4</sup> „VGG-16“ steht für Visual Geometric Group der University of Oxford, 16 Schichten.

Abb. 6.13. Architektur des Convolutional Neural Networks VGG-16



## Quellen

- [1] ISO Central Secretary. *Information technology – Vocabulary*. Standard ISO/IEC 2382:2015. Genf, Schweiz: International Organization for Standardization, 2015, S. 2121376.
- [2] Jia Deng u. a. „ImageNet: A large-scale hierarchical image database“. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2009, S. 248–255. DOI: [10.1109/CVPR.2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848).
- [3] Martin Popel u. a. „Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals“. In: *Nature Communications* 11.1 (Sep. 2020), S. 4381. DOI: [10.1038/s41467-020-18073-9](https://doi.org/10.1038/s41467-020-18073-9).
- [4] David Silver u. a. „Mastering the game of Go without human knowledge“. In: *Nature* 550.7676 (Okt. 2017), S. 354–359. DOI: [10.1038/nature24270](https://doi.org/10.1038/nature24270).
- [5] Berkeley Earth. *Time Series Data – Monthly Global Average Temperature (Annual Summary)*. Aufgerufen am 01. Feb. 2020. URL: <http://berkeleyearth.org/data/>.
- [6] Jonathan Barzilai und Jonathan M. Borwein. „Two-Point Step Size Gradient Methods“. In: *IMA Journal of Numerical Analysis* 8.1 (1988), S. 141–148. DOI: [10.1093/imanum/8.1.141](https://doi.org/10.1093/imanum/8.1.141).
- [7] Charles George Broyden. „The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations“. In: *IMA Journal of Applied Mathematics* 6.1 (1970), S. 76–90. DOI: [10.1093/imamat/6.1.76](https://doi.org/10.1093/imamat/6.1.76).
- [8] Roger Fletcher. „A new approach to variable metric algorithms“. In: *The Computer Journal* 13.3 (März 1970), S. 317–322. DOI: [10.1093/comjnl/13.3.317](https://doi.org/10.1093/comjnl/13.3.317).
- [9] Donald Goldfarb. „A family of variable-metric methods derived by variational means“. In: *Mathematics of Computation* 24.109 (Jan. 1970), S. 23–23. DOI: [10.1090/s0025-5718-1970-0258249-6](https://doi.org/10.1090/s0025-5718-1970-0258249-6).
- [10] David F. Shanno. „Conditioning of quasi-Newton methods for function minimization“. In: *Mathematics of Computation* 24.111 (Sep. 1970), S. 647–647. DOI: [10.1090/s0025-5718-1970-0274029-x](https://doi.org/10.1090/s0025-5718-1970-0274029-x).
- [11] Larry Armijo. „Minimization of functions having Lipschitz continuous first partial derivatives“. In: *Pacific Journal of Mathematics* 16.1 (Jan. 1966), S. 1–3. DOI: [10.2140/pjm.1966.16.1](https://doi.org/10.2140/pjm.1966.16.1).
- [12] Roger Fletcher. *Practical methods of optimization*. 2. Aufl. Wiley, 1987. ISBN: 978-0-471-91547-8.
- [13] Gerd Fischer. *Lineare Algebra*. 18. Aufl. Springer Spektrum, Wiesbaden, 2014. DOI: [10.1007/978-3-658-03945-5](https://doi.org/10.1007/978-3-658-03945-5).
- [14] D. Randall Wilson und Tony R. Martinez. „Reduction Techniques for Instance-Based Learning Algorithms“. In: *Machine Learning* 38 (2000), S. 257–286. DOI: [10.1023/a:1007626913721](https://doi.org/10.1023/a:1007626913721).
- [15] Mehryar Mohri, Afshin Rostamizadeh und Ameet Talwalkar. *Foundations of Machine Learning*. 2. Aufl. MIT Press, 2018. ISBN: 978-0-262-03940-6.

- [16] CDC Population Health Surveillance Branch. *Behavioral Risk Factor Surveillance System (BRFSS) Survey Data 2018*. Aufgerufen am 01. Feb. 2020. URL: <https://www.cdc.gov/brfss/>.
- [17] David H. Hall, Zeynep F. Altun und Laura A. Herndon. *Wormatlas. Neuronal Wiring*. Aufgerufen am 30. Dez. 2020. New York, USA. URL: <https://www.wormatlas.org/neuronalwiring.html>.
- [18] Lav R. Varshney u. a. „Structural Properties of the Caenorhabditis elegans Neuronal Network“. In: *PLoS Computational Biology* 7.2 (Feb. 2011). Hrsg. von Olaf Sporns, e1001066. DOI: [10.1371/journal.pcbi.1001066](https://doi.org/10.1371/journal.pcbi.1001066).
- [19] Gang Yan u. a. „Network control principles predict neuron function in the Caenorhabditis elegans connectome“. In: *Nature* 550.7677 (Okt. 2017), S. 519–523. DOI: [10.1038/nature24056](https://doi.org/10.1038/nature24056).
- [20] Frank Rosenblatt. „The perceptron: A probabilistic model for information storage and organization in the brain.“ In: *Psychological Review* 65.6 (1958), S. 386–408. DOI: [10.1037/h0042519](https://doi.org/10.1037/h0042519).
- [21] Frank Rosenblatt. *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms*. Washington, D.C., USA: Spartan Books, 1962.
- [22] Melanie Lefkowitz. „Professor’s perceptron paved the way for AI – 60 years too soon“. In: *Cornell Chronicle* (Sep. 2019). URL: <https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon>.
- [23] Andrew L. Maas, Awni Y. Hannun und Andrew Y. Ng. „Rectifier nonlinearities improve neural network acoustic models“. In: *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. 2013.
- [24] Christian Szegedy u. a. „Rethinking the Inception Architecture for Computer Vision“. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Juni 2016. DOI: [10.1109/cvpr.2016.308](https://doi.org/10.1109/cvpr.2016.308). arXiv:1512.00567.
- [25] Rafael Müller, Simon Kornblith und Geoffrey E Hinton. „When does label smoothing help?“ In: *Advances in Neural Information Processing Systems*. Hrsg. von H. Wallach u. a. Bd. 32. Curran Associates, Inc., 2019, S. 4694–4703. arXiv:1906.02629.
- [26] Allan Pinkus. „Approximation theory of the MLP model in neural networks“. In: *Acta Numerica* 8 (Jan. 1999), S. 143–195. DOI: [10.1017/s0962492900002919](https://doi.org/10.1017/s0962492900002919).
- [27] Patrick Kidger und Terry Lyons. „Universal Approximation with Deep Narrow Networks“. In: *33rd Conference on Learning Theory*. Hrsg. von Jacob Abernethy und Shivani Agarwal. Bd. 125. Proceedings of Machine Learning Research. PMLR, Juli 2020, S. 2306–2327. arXiv:1905.08539.
- [28] Michael A. Nielsen. *Neural networks and deep learning*. Determination Press, 2015. URL: <http://neuralnetworksanddeeplearning.com/>.
- [29] Nitish Srivastava u. a. „Dropout: A Simple Way to Prevent Neural Networks from Overfitting“. In: *J. Mach. Learn. Res.* 15.1 (Jan. 2014), S. 1929–1958.

- [30] Allan G. Weber. *The USC-SIPI Image Database: Version 6*. Techn. Ber. Los Angeles, USA: Signal und Image Processing Institute, University of Southern California, Feb. 2018. URL: <http://sipi.usc.edu/database>.
- [31] Karen Simonyan und Andrew Zisserman. „Very Deep Convolutional Networks for Large-Scale Image Recognition“. In: *3rd International Conference on Learning Representations, San Diego, USA*. Hrsg. von Yoshua Bengio und Yann LeCun. Mai 2015. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- [32] Olga Russakovsky u. a. „ImageNet Large Scale Visual Recognition Challenge“. In: *International Journal of Computer Vision (IJCV) 115.3* (2015), S. 211–252. DOI: [10.1007/s11263-015-0816-y](https://doi.org/10.1007/s11263-015-0816-y). [arXiv:1409.0575](https://arxiv.org/abs/1409.0575).
- [33] Papers with Code Community. *ImageNet Benchmark (Image Classification)*. Hrsg. von Robert Stojnic u. a. Aufgerufen am 28. Dez. 2020. URL: <https://paperswithcode.com/sota/image-classification-on-imagenet>.
- [34] Mingxing Tan und Quoc V. Le. „EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks“. In: *36th International Conference on Machine Learning, Long Beach, California, USA*. Hrsg. von Kamalika Chaudhuri und Ruslan Salakhutdinov. Bd. 97. Proceedings of Machine Learning Research. PMLR, Juni 2019, S. 6105–6114. [arXiv:1905.11946](https://arxiv.org/abs/1905.11946).
- [35] Longhui Wei u. a. „Circumventing Outliers of AutoAugment with Knowledge Distillation“. In: *Computer Vision – ECCV 2020*. Hrsg. von A. Vedaldi u. a. Bd. 12348. Lecture Notes in Computer Science. Springer, Cham, 2020, S. 608–625. DOI: [10.1007/978-3-030-58580-8\\_36](https://doi.org/10.1007/978-3-030-58580-8_36). [arXiv:2003.11342](https://arxiv.org/abs/2003.11342).
- [36] Chengyue Gong u. a. *MaxUp: A Simple Way to Improve Generalization of Neural Network Training*. Feb. 2020. [arXiv:2002.09024v1](https://arxiv.org/abs/2002.09024v1).
- [37] Cihang Xie u. a. „Adversarial Examples Improve Image Recognition“. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Juni 2020. DOI: [10.1109/cvpr42600.2020.00090](https://doi.org/10.1109/cvpr42600.2020.00090). [arXiv:1911.09665](https://arxiv.org/abs/1911.09665).



# Unüberwachtes maschinelles Lernen

Ziel von Verfahren des überwachten maschinellen Lernens ist die Ableitung einer Entscheidungsregel  $f: \mathcal{X} \rightarrow \mathcal{Y}$  anhand eines Trainingsdatensatzes  $(\mathcal{X} \times \mathcal{Y})^N$ . Im Gegensatz dazu ist bei Verfahren des **unüberwachten Lernens** keines der Merkmale des zu analysierenden Datensatzes a priori als Zielgröße oder Klassenlabel ausgezeichnet.

Zwei wesentliche Typen von unüberwachten Verfahren sind die folgenden:

- Ziel einer **Dimensionsreduktion** ist es, die Anzahl der Merkmale zu verringern, ohne bestimmte Charakteristiken – etwa den paarweisen Abstand der Datenpunkte voneinander – wesentlich zu verändern. Verfahren der Dimensionsreduktion können als unüberwachtes Analogon zu Regressionsverfahren verstanden werden – hierzu mehr im nächsten Abschnitt.
- Verfahren der **Clusteranalyse** haben zum Ziel, den Datensatz in Gruppen einander ähnlicher Informationsobjekte einzuteilen. Die Clusteranalyse ist damit ähnlich einer Klassifikationsaufgabe – nur werden die Clusterlabels nicht durch einen Trainingsdatensatz vorgegeben, sondern vom Algorithmus allein aus der Struktur der Eingangsdaten erlernt.

In diesem Kapitel befassen wir uns zunächst mit der geometrischen und der topologischen Sicht auf Daten. Dieser Blickwinkel dient der einleitenden Motivation von Verfahren der Dimensionsreduktion und der Clusteranalyse, die in den darauf folgenden Abschnitten erläutert werden.

## 7.1 Elemente des unüberwachten Lernens

Gegeben ein Datensatz von Punkten  $x_1, x_2, \dots, x_N \in \mathbb{R}^D$  galt unser Hauptaugenmerk bislang deren statistischer Beschreibung; etwa welcher Verteilung die Punkte gehorchen. Bereits im Kapitel über das überwachte maschinelles Lernen haben wir gesehen, dass darüber hinaus eine geometrische Sichtweise einem besseren Verständnis der Verfahren oft dienlich ist. Um ein paar Beispiele zu nennen: Der  $K$ -nächste-Nachbarn-Klassifikator bedient sich direkt der Anschauung

von Datenpunkten als Objekte, die in einem Raum in gewissen Abständen zueinander angeordnet sind. Affine Unterräume spielen als geometrische Objekte ebenso eine wesentliche Rolle: Das Ergebnis einer linearen Regression ist eine Hyperebene im Merkmalsraum, und durch einen linearen Klassifikator werden Klassen durch ebensolche Hyperebenen voneinander getrennt.

### 7.1.1 Intrinsische Dimension von Daten

Für die folgenden Ausführungen sollten nochmals die Streudiagramme betrachtet werden, die zuvor als Beispiele für Regressionsverfahren oder stark korrelierte Merkmale angeführt wurden, etwa Abb. 2.3 oder Abb. 6.2. Die Datenpunkte streuen in diesen Fällen nicht beliebig in alle Richtungen in der Ebene, sondern befinden sich in der Nähe einer Regressionskurve. Abgesehen von einer geringen, durch Störgrößen hervorgerufenen Streuung kann die Position eines Datenpunkts also als ein Punkt entlang dieser Kurve beschrieben werden. Auf diese Weise wurde die Dimensionalität des Datensatzes effektiv von zwei (der von Eingangs- und Zielgröße aufgespannte Merkmalsraum) auf eins (der Position entlang der Regressionskurve) verringert.

Der folgende Lehrsatz gibt einen Hinweis darauf, inwieweit die Hoffnung bestehen darf, dass sich eine Anzahl von Datenpunkten *generell* in der Nähe eines Unterraumes von geringerer Dimension befindet.

**Lemma von Johnson und Lindenstrauss [1].** Seien eine Reihe von Datenpunkten  $x_1, \dots, x_N \in \mathbb{R}^D$  sowie  $\varepsilon \in ]0, 1[$  gegeben. Sei weiterhin  $K \in \mathbb{N}$  mit

$$K \geq 4 \cdot \left( \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} \right)^{-1} \cdot \ln N.$$

Dann existiert eine lineare Abbildung  $f: \mathbb{R}^D \rightarrow \mathbb{R}^K$ , sodass für alle  $m, n \in \{1, \dots, N\}$  gilt:

$$(1 - \varepsilon) \cdot \|x_m - x_n\|^2 \leq \|f(x_m) - f(x_n)\|^2 \leq (1 + \varepsilon) \cdot \|x_m - x_n\|^2$$

Für kleine Werte von  $\varepsilon$  unterscheiden sich die paarweisen Abstände zwischen den Datenpunkten unter der Abbildung  $f$  also nur unwesentlich. Eine weniger strikte, aber vielleicht einfacher zu merkende Schranke für die Zieldimension ist  $K \geq 20 \cdot \varepsilon^{-2} \cdot \ln N$ , falls  $0 < \varepsilon < 0,9$  (vgl. [2, Lemma 15.4]).

Die Beweisidee besteht in der zufälligen Auswahl eines  $K$ -dimensionalen Unterraums von  $\mathbb{R}^D$ . Bezeichnet  $\text{proj}(\cdot)$  die orthogonale Projektion auf besagten Unterraum, so kann gezeigt werden, dass für die Abbildung  $f(v) := \sqrt{D/K} \cdot \text{proj}(v)$ ,  $v \in \mathbb{R}^D$ , gilt: Die Wahrscheinlichkeit, dass die gewünschte Schranke für alle Datenpunktpaare  $x_m, x_n$  Gültigkeit hat, beträgt *mindestens*  $1/N$ . Die Wahrscheinlichkeit ist insbesondere also echt größer als null, daher muss es einen Unterraum bzw. eine Projektion geben, welche die gewünschte Bedingung mit Sicherheit erfüllt.

Die geschätzte Ausgangsdimension  $K$  hängt nicht von der ursprünglichen Dimensionalität  $D$  ab, und nur logarithmisch von der Anzahl der Datenpunkte  $N$ . Wir illustrieren die Tragweite dieses Umstands anhand eines Rechenbeispiels: Sei ein Datensatz von  $N = 10.000$  digitalen Farbfotos im Format von  $1920 \times 1080$  Pixeln gegeben. Unter Berücksichtigung von drei Farbkanälen müssen für jedes unkomprimierte Foto  $D = 1920 \cdot 1080 \cdot 3 = 6.220.800$  Farbwerte gespeichert werden. Aus Sicht der Datenanalyse kann somit jedes Bild als ein Vektor in einem Raum mit einer Dimension von etwa 6 Millionen aufgefasst werden.

Das Lemma von Johnson und Lindenstrauss impliziert jedoch, dass die **intrinsische Dimensionalität** deutlich geringer ist: Es existiert eine Abbildung  $f: \mathbb{R}^D \rightarrow \mathbb{R}^K$  mit  $K = 7895$ , sodass die relativen quadrierten Distanzen zwischen den projizierten Datenpunkten in keinem Fall um mehr als  $\varepsilon = 10\%$  von den relativen quadrierten Distanzen im Eingangsdatensatz abweichen. Diese potenzielle Reduktion in der Dimensionalität ändert sich nur wenig mit der Größe des Datensatzes;  $N = 1.000.000$  Bilder führten auf eine maximale intrinsische Dimension des Datensatzes von  $K = 11.842$ .

Dabei wurden noch keinerlei Anforderungen an die Verteilung der ursprünglichen Datenpunkte gestellt: Die Inhalte der Fotos könnten beliebig sein und gar nur aus Bildrauschen bestehen. Darüber hinaus macht der Satz nur eine Aussage über die Möglichkeit der *linearen* Abbildung der Daten.

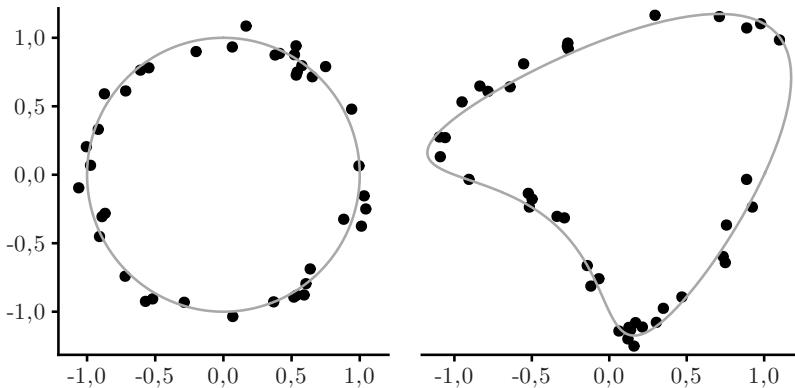
Die Konstruktion einer niederdimensionalen Darstellung von Daten mit anfangs hoher extrinsischer Dimensionalität ist Aufgabe von Verfahren zur **Dimensionsreduktion**, das Ergebnis einer solchen Reduktion wird mitunter auch als **Einbettung** bezeichnet.

In der Bildanalyse [3] oder der Textanalyse [4] gebräuchliche Einbettungen weisen mitunter eine Dimensionalität von wenigen hundert Komponenten auf.

### 7.1.2 Topologische Merkmale von Daten

Gemäß dem Lemma von Johnson und Lindenstrauss liegen Datenpunkte der Dimensionalität  $D$  in der Nähe eines  $K$ -dimensionalen linearen Unterraums. In der Praxis zeigt sich oft, dass die intrinsische Dimensionalität  $K$  als wesentlich kleiner als die extrinsische  $D$  angenommen werden kann.

Diese Erkenntnis beschränkt sich nicht bloß auf lineare Unterräume: Sie kann auch auf allgemeinere Kurven, Flächen oder deren höherdimensionale Analoga erweitert werden. Die folgenden Streudiagramme zeigen zum Beispiel zwei (synthetische) Datensätze, die sich beide um eine geschlossene Kurve herum verteilen:



**Abb. 7.1.** Datenpunkte entlang einer Kurve

Geometrisch sind beide Kurven verschieden, so ist die linke eine Kreislinie mit konstantem Radius, während die Krümmung der rechten Kurve variabel ist. Die qualitative Gestalt als zusammenhängende, geschlossene Kurven ist jedoch dieselbe: Wir können uns vorstellen, dass die Kurven durch stetige Verformung ineinander überführt werden können. Die mathematische Disziplin, welche sich dem Studium der qualitativen Gestalt verschrieben hat, ist die **Topologie** – scherhaft auch „Gummituchgeometrie“ genannt.

Eine wichtige topologische Eigenschaft ist der Zusammenhang bzw. die Möglichkeit der Zerlegung in zusammenhängende Teile.

Sei  $U$  eine Teilmenge von  $\mathbb{R}^D$ . Eine Menge  $V \subseteq U$  nennt sich **Wegzusammenhangskomponente** von  $U$ , falls gilt:

1. Für alle  $x, y \in V$  existiert eine stetige Kurve, die ganz in  $V$  liegt und  $x$  mit  $y$  verbindet.
2. Es gibt keine von  $V$  verschiedene Menge  $W$  mit  $V \subset W \subseteq U$ , die ebenfalls obige Eigenschaft hat.

Es kann nachgewiesen werden, dass es ausreicht, anstelle beliebiger stetiger Kurven bloß Polygonzüge zu betrachten. Die Wegzusammenhangskomponenten bilden eine Partition von  $U$ , d. h., sie sind paarweise disjunkt und überdecken  $U$ . Die in der Praxis der Datenanalyse untersuchten Mengen haben in der Regel nur endlich viele Wegzusammenhangskomponenten  $S_1, \dots, S_K$ :

$$U = S_1 \cup \dots \cup S_K, S_k \cap S_l = \emptyset$$

für alle  $k, l \in \{1, \dots, K\}$ .

Eine endliche Menge von  $N$  paarweise verschiedenen Datenpunkten  $U_0 = \{x_1, \dots, x_N\} \subset \mathbb{R}^D$  hat immer genau  $N$  Wegzusammenhangskomponenten:  $S_n = \{x_n\}$ ,  $n \in \{1, \dots, N\}$ . Jeder Weg zwischen zwei verschiedenen Datenpunkten führt nämlich zwangsläufig aus der Menge heraus. Dies stellt also

keinen besonderen Fortschritt in der Analyse der topologischen Struktur der Daten dar. Betrachten wir stattdessen für ein gegebenes  $\varepsilon > 0$  die folgende Überdeckung durch Kugeln:  $U_\varepsilon = B_\varepsilon(x_1) \cup \dots \cup B_\varepsilon(x_N)$  mit

$$B_\varepsilon(x_n) = \{u \in \mathbb{R}^D \mid \|u - x_n\| \leq \varepsilon\}$$

für alle  $n \in \{1, \dots, N\}$ .

In Abb. 7.2 sind für einen synthetischen Datensatz eine Reihe von solchen Überdeckungen durch Kugeln (im Fall  $D = 2$  sind dies Kreis Scheiben) dargestellt. Die Anzahl der Wegzusammenhangskomponenten von  $U_\varepsilon$  fällt mit steigendem Überdeckungsradius  $\varepsilon$ . Jede dieser Komponenten steht für eine Gruppierung von Datenpunkten, die sich im Merkmalsraum nahe sind, und welche **Cluster** genannt werden. Das Ziel einer **Clusteranalyse** besteht in der Identifikation solcher Cluster.

Cluster sind die einfachsten, aber nicht die einzigen topologischen Merkmale, die in Daten identifiziert werden können. Die Überdeckung mit größtem Radius in Abb. 7.2 offenbart etwa eine ring- oder kreisförmige Struktur. Die vergleichsweise junge Disziplin der **topologischen Datenanalyse** befasst sich mit der Identifikation solcher Strukturen und ihrer höherdimensionalen Analoga.

## 7.2 Dimensionsreduktion

Unter Dimensionsreduktion verstehen wir die Reduktion der Anzahl der Merkmale. Zwei Ansätze können unterschieden werden:

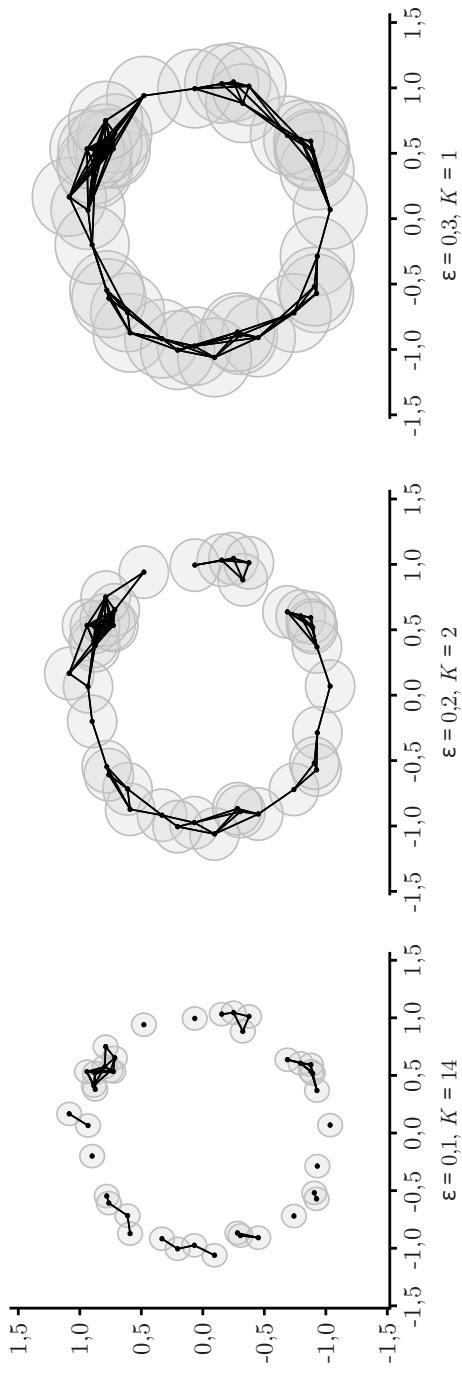
- Bei Verfahren der **Merkmalsauswahl** werden von den  $D$  Eingangsmerkmalen  $K < D$  Merkmale nach bestimmten Kriterien ausgewählt.
- Bei einer **Merkmalsextraktion** werden die  $D$  Eingangsmerkmale im Allgemeinen verändert: Eine geeignete Transformation führt diese in  $K < D$  neue Ausgangsmerkmale über.

In beiden Fällen sollen die Ausgangsmerkmale, obwohl weniger an der Zahl, wesentliche Charakteristiken des Datensatzes widerspiegeln.

Im Rahmen des überwachten Lernens kann die Dimensionsreduktion ein Verarbeitungsschritt sein, welcher der eigentlichen Aufgabe, etwa dem Training eines Klassifikators, vorangestellt wird. Hintergrund der Erfordernis einer Dimensionsreduktion kann zum einen die Optimierung der Laufzeit des Algorithmus sein: Je weniger Merkmale verarbeitet werden müssen, desto geringer ist diese in der Regel. Ein weiterer wichtiger Faktor ist die Beobachtung, dass die Verwendung einer großen Anzahl von Merkmalen zu einer Überanpassung führen kann.

Herausforderungen, die mit der Verarbeitung hochdimensionaler Daten verbunden sind, werden unter dem Schlagwort „**Fluch der Dimensionalität**“ zusammengefasst. Für eine Dimensionsreduktion zum „Brechen“ des Fluchs kann

Abb. 7.2. Überdeckung durch Kreisscheiben mit variablem Radius  $\varepsilon$  und Clusteranzahl  $K$



die sogenannte Hauptkomponentenanalyse (engl. *principal component analysis, PCA*) eingesetzt werden, die wir im Folgenden beschreiben.

Im Kontext der Datenvisualisierung besteht das Ziel einer Dimensionsreduktion hingegen darin, wesentliche Aspekte der hochdimensionalen Eingangsdaten grafisch sichtbar machen zu können. Das im Abschn. 7.2.4 vorgestellte t-SNE-Verfahren ist dazu geeignet.

### 7.2.1 Hauptkomponentenanalyse

Seien  $X = (X_1, \dots, X_D)^T$  ein Zufallsvektor und  $\Sigma[X]$  dessen Kovarianzmatrix. Da  $\Sigma[X]$  eine symmetrische Matrix ist, kann für diese die folgende Diagonalisierung gefunden werden (siehe z. B. [5, Abschnitt 5.6.2]): Es gibt eine orthogonale Matrix  $V$  und eine Diagonalmatrix  $\Lambda$ , sodass gilt:

$$\Sigma[X] = V \cdot \Lambda \cdot V^T$$

In den Spalten von  $V$  stehen die normierten und paarweise orthogonalen Eigenvektoren  $v_1, \dots, v_D \in \mathbb{R}^D$  von  $\Sigma[X]$ . Auf der Diagonalen von  $\Lambda$  stehen die zugehörigen reellen Eigenwerte  $\lambda_1, \dots, \lambda_D$ . Da  $\Sigma[X]$  positiv semidefinit ist, ist von diesen keiner negativ.

Definieren wir den neuen Zufallsvektor  $Z := V^T \cdot (X - E[X])$ , so gilt  $E[Z] = 0$  und außerdem für dessen Varianz:

$$\Sigma[Z] = E[ZZ^T] = V^T E[(X - E[X]) \cdot (X - E[X])^T] V = V^T \Sigma[X] V = \Lambda$$

Die Transformation führt also einen beliebigen Zufallsvektor  $X$  in einen neuen Zufallsvektor  $Z$  mit diagonaler Kovarianzmatrix über.

Ein entsprechendes Resultat gilt auch für eine Stichprobe von Merkmalsvektoren  $x_1, \dots, x_N \in \mathbb{R}^D$ . Da die empirische Kovarianzmatrix  $S(x)$  ebenfalls symmetrisch und positiv semidefinit ist, kann für diese eine Diagonaldarstellung mit obigen Eigenschaften ermittelt werden:

$$S(x) = V \cdot \Lambda \cdot V^T$$

In diesem Fall besitzt die Kovarianzmatrix des transformierten Datensatzes  $V^T x_1, V^T x_2, \dots, V^T x_n$  Diagonalgestalt. Dies können wir ganz ähnlich wie oben einsehen: Zunächst gehen wir davon aus, dass die entsprechende Datenmatrix  $\mathbf{X}$  mittelwertzentriert ist und schreiben für die empirische Kovarianz  $S(x) = \frac{1}{N} \mathbf{X}^T \mathbf{X}$ . Die Datenmatrix der transformierten Stichprobe  $z_1, \dots, z_N$  ist durch  $\mathbf{Z} = \mathbf{X} \cdot V$  gegeben, und es folgt:

$$NS(z) = \mathbf{Z}^T \mathbf{Z} = (\mathbf{X}V)^T \mathbf{X}V = V^T \mathbf{X}^T \mathbf{X}V = NV^T S(x)V = N\Lambda$$

Stellen wir die Methoden der Datentransformation gegenüber, die wir bislang kennengelernten:

- Eine Mittelwertzentrierung verschiebt die Datenpunkte so, dass der Schwerpunkt im Ursprung zum Liegen kommt.

- Eine Standardisierung skaliert die Daten nach Mittelwertzentrierung so, dass die Varianz der einzelnen Merkmale identisch (gleich eins) ist. Die neuen Merkmale können nach wie vor korreliert sein.
- Obige **Hauptachsentransformation** führt die Daten in Merkmale mit verschwindender Kovarianz bzw. Korrelation über.

Alle diese Transformationen stellen eine affine Abbildung dar. Allen geht eine Verschiebung um den Schwerpunkt voraus, gefolgt von einer linearen Abbildung: Im Fall der Standardisierung eine Skalierung, im Fall der Hauptachsentransformation eine lineare Abbildung mit der darstellenden Matrix  $V^T$ .

Wir wollen nun annehmen, dass die Eigenwerte nach absteigender Größe sortiert wurden:  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D \geq 0$ . Dann hat der Stichprobenvektor  $Z_{\bullet 1}$  die größte Varianz  $\sigma_1^2 = \lambda_1$ , während  $Z_{\bullet D}$  die kleinste Varianz aufweist. Die Grundidee bei der Dimensionsreduktion durch Haupkomponentenanalyse ist es, Richtungen kleiner Varianz zu vernachlässigen.

**Hauptkomponentenanalyse.** Seien Datenpunkte  $x_1, \dots, x_N \in \mathbb{R}^D$  mit Schwerpunkt  $\bar{x}$  gegeben. Eine **Karhunen-Loève-Transformation** mit  $1 \leq K \leq D$  ist die affine Abbildung

$$\text{pca}_K: \mathbb{R}^D \rightarrow \mathbb{R}^K, \text{pca}_K(u) = (V_K)^T \cdot (u - \bar{x}).$$

Die Spalten der  $D \times K$ -Matrix  $V_K$  sind die ersten  $K$  der zu den größten Eigenwerten gehörigen normierten Eigenvektoren der Kovarianzmatrix  $S(x)$ .

In der Literatur sind verschiedene Konventionen<sup>1</sup> für die Benennung der beteiligten Objekte geläufig; wir wollen die folgende verwenden. Die nach Größe der zugehörigen Varianz absteigend sortierten Eigenvektoren  $v_1, \dots, v_K, \dots, v_D$  der Kovarianzmatrix werden die **Hauptrichtungen** genannt. Die von der  $k$ -ten Hauptrichtung  $v_k$  aufgespannte affine Gerade mit dem Schwerpunkt als Fußpunkt hat die Parametergleichung

$$\mathbb{R} \rightarrow \mathbb{R}^D, \lambda \mapsto \bar{x} + \lambda \cdot v_k$$

und stellt eine **Hauptachse** dar. Alle Hauptrichtungen bzw. Hauptachsen stehen paarweise senkrecht aufeinander. Projizieren wir einen beliebigen Vektor  $u \in \mathbb{R}^D$  auf die  $k$ -te Hauptachse, so ist der orientierte Abstand der Projektion vom Schwerpunkt durch  $\langle v_k, u - \bar{x} \rangle$  gegeben, der  $k$ -ten **Hauptkomponente** oder **Hauptkoordinate**. Die ersten  $K$  Hauptkomponenten bilden den Karhunen-Loève-transformierten Vektor:

$$\text{pca}_K: \mathbb{R}^D \rightarrow \mathbb{R}^K, \text{pca}_K: u \mapsto \begin{pmatrix} \langle v_1, u - \bar{x} \rangle \\ \vdots \\ \langle v_K, u - \bar{x} \rangle \end{pmatrix}$$

<sup>1</sup> Insbesondere der Begriff „Hauptkomponente“ wird sehr frei und austauschbar für Hauptachsen, -richtungen oder -koordinaten verwendet. Zu allem Überfluss gibt es auch noch den „Hauptvektor“ als Verallgemeinerung des Begriffs des Eigenvektors.

Anders ausgedrückt sind dies gerade die Koordinaten der Projektion des Vektors auf den von den ersten  $K$  Hauptachsen aufgespannten  $K$ -dimensionalen affinen Unterraum bzgl. der Basis von Hauptrichtungen.

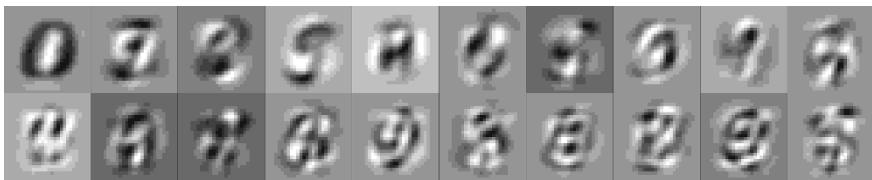
Wurde der Eingabevektor durch eine ähnliche Verteilung erzeugt, wie sie dem Hauptachsen-transformierten Datensatz zugrundeliegt, so kann der projizierte Vektor als eine Näherung des Eingabevektors verstanden werden: In Richtungen, die auf den ersten Hauptachsen senkrecht stehen, ist die Varianz gering. Daher ist es sehr wahrscheinlich (wir können dabei wieder an die Tschebyscheff'sche Ungleichung denken, Abschn. 3.4.3), dass die Komponenten in diesen Richtungen kleinen Betrag haben und vernachlässigt werden können. Entsprechend können wir unter diesen Umständen auch erwarten, den Vektor näherungsweise anhand von dessen Hauptkomponenten rekonstruieren zu können: Es gilt  $u \approx ((\text{pca}_K)^\dagger \circ \text{pca}_K)(u)$  mit

$$(\text{pca}_K)^\dagger: \mathbb{R}^K \rightarrow \mathbb{R}^D, (\text{pca}_K)^\dagger: y \mapsto \bar{x} + \sum_{k=1}^K y_k \cdot v_k$$

**Anwendungsbeispiel.** Der MNIST-Datensatz [6], den wir auch in Abschn. 8.1.1 als Anwendungsbeispiel für Klassifikationsverfahren heranziehen, besteht aus digitalen Bildern handgeschriebener Ziffern im Format von  $28 \times 28$  Pixeln, siehe auch Abb. 8.6 oben. Für die Eingabe in die Hauptkomponentenanalyse ignorieren wir die Anordnung der Pixel als Raster zunächst und fassen die Grauwerte als Komponenten eines Vektors der Länge  $D = 28 \cdot 28 = 784$  auf. In Abb. 7.5 wurde oben eine kleine Auswahl von Ziffern auf die Position der jeweiligen zwei ersten Hauptkoordinaten gesetzt. Es ist eine gewisse Tendenz zu erkennen, mit der gleiche Ziffern gruppiert werden, besonders deutlich wird dies bei den Ziffern Null und Eins. Dies ist ein erster Hinweis darauf, dass die ersten Hauptkomponenten bereits wesentliche Charakteristiken der Merkmalsträger widerspiegeln: Wir können eine Ziffer anhand weniger Hauptkomponenten anstelle der Grauwerte aller Pixel identifizieren.

Die ersten zwei Hauptkomponenten sind jedoch für eine gute Rekonstruktion noch nicht ausreichend: In Abb. 7.5 sind unten Linearkombinationen der ersten zwei Hauptrichtungen dargestellt, zurück auf Bildformat gebracht. Diese zeigen eine gewisse Ähnlichkeit mit den Ziffern Null, Eins und Neun, spannen jedoch augenscheinlich noch keinen hinreichend großen Teil des Merkmalsraums auf.

Zwanzig Hauptkomponenten sind für eine Rekonstruktion akzeptabler Güte jedoch bereits ausreichend. Die ersten zwanzig Eigenvektoren  $v_1, \dots, v_{20}$  der Kovarianzmatrix sehen, zurück auf Bildformat gebracht, wie folgt aus:



**Abb. 7.3.** „Eigenziffern“ des MNIST-Datensatzes

Diese Bilder können quasi als Erzeugendensystem für den MNIST-Datensatz angesehen werden: Jede Ziffer lässt sich näherungsweise durch Linearkombination dieser Bilder und Addition zu dem aus den Mittelwerten aller Grauwerte bestehenden Bild rekonstruieren. Am Beispiel zeigt dies die folgende Abbildung; oben ist das Original und unten die Rekonstruktion:



**Abb. 7.4.** Bildrekonstruktion anhand von Hauptkomponenten

## 7.2.2 Autoencoder

Neuronale Netzwerke können auch für Aufgaben des unüberwachten Lernens eingesetzt werden. Sogenannte **Autoencoder** basieren wesentlich auf zwei Ideen oder Annahmen:

- Ein neuronales Netzwerk kann auch ohne ausgezeichnete Zielgröße als Regressionsverfahren eingesetzt werden, indem jedes Trainingsbeispiel sich selbst als Zielgröße zugeordnet wird. Das Netzwerk lernt auf diese Weise eine Näherung der identischen Abbildung id:  $\mathbb{R}^D \rightarrow \mathbb{R}^D$ ,  $u \mapsto u$ .
- Die Aktivierungen der Neuronen verborgener Schichten spiegeln wesentliche Charakteristiken der Ein- bzw. Ausgabedaten wider.

Ein **Autoencoder** ist ein neuronales Netzwerk mit wenigstens einer verborgenen Schicht, bei dem Ein- und Ausgabeschicht dieselbe Breite haben:

$$f: \mathbb{R}^D \rightarrow \mathbb{R}^D, f(u) = (f_L \circ f_{L-1} \circ \dots \circ f_1)(u), L \geq 2$$

Eine der verborgenen Schichten des Autoencoders, etwa  $f_m: \mathbb{R}^{D_{m-1}} \rightarrow \mathbb{R}^{D_m}$ ,  $1 \leq m < L$ , wird als **latente Schicht** ausgezeichnet.

Gegeben ein Datensatz  $x_1, \dots, x_N \in \mathbb{R}^D$  wird der Autoencoder mit identischer Ein- und Ausgabe trainiert; das zu minimierende empirische Risiko hat also die folgende Form:

$$\hat{R}[f] = \frac{1}{N} \sum_{n=1}^N \lambda(x_n, f(x_n))$$

Die so gelernten Abbildungen

$$\text{code}: \mathbb{R}^D \rightarrow \mathbb{R}^{D_m}, \text{code}(u) = (\hat{f}_m \circ \hat{f}_{m-1} \circ \cdots \circ \hat{f}_1)(u)$$

und

$$\text{code}^\dagger: \mathbb{R}^{D_m} \rightarrow \mathbb{R}^D, \text{code}^\dagger(y) = (\hat{f}_L \circ \hat{f}_{L-1} \circ \cdots \circ \hat{f}_{m+1})(y)$$

stellen die **Kodierungs-** bzw. **Dekodierungsabbildung** dar.

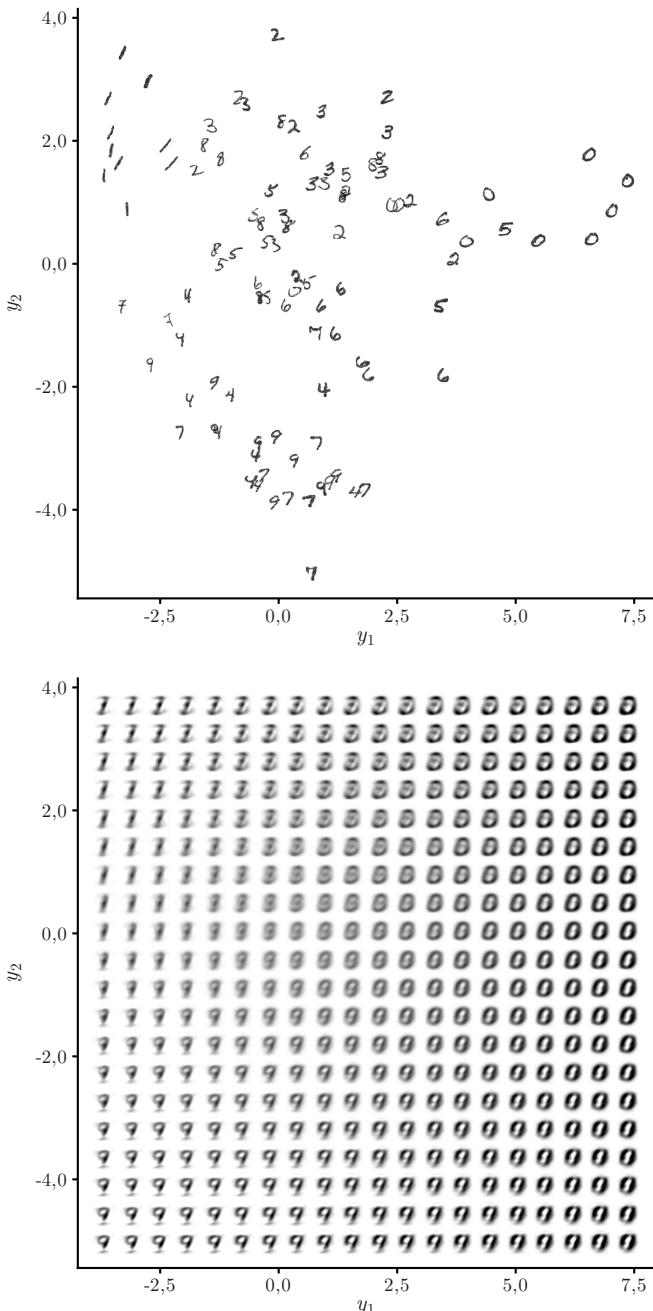
Wird die Breite der latenten Schicht (deutlich) geringer als die Dimensionalität der Eingabedaten gewählt,  $D_m \ll D$ , so definiert die Kodierungsabbildung eine entsprechende Dimensionsreduktion.

**Anwendungsbeispiel.** Abb. 7.6 zeigt oben ein Streudiagramm der Kodierung eines Autoencoders einer Auswahl von Ziffern aus dem MNIST-Datensatz; für das Training wurde der gesamte Datensatz verwendet. Der Autoencoder selbst ist ein gewöhnliches Feedforward-Netzwerk mit fünf verborgenen Schichten und der folgenden Anzahl an Neuronen in den verborgenen Schichten: 32, 64, 2, 64, 32. Die mittlere Schicht mit zwei Neuronen stellt die latente Schicht dar. Als Aktivierungsfunktionen wurden durchlässige Gleichrichter verwendet. Die Minimierung des Trainingsfehlers auf Grundlage der quadratischen Verlustfunktion wurde mithilfe eines stochastischen Gradientenabstiegs durchgeführt. Die Implementierung erfolgte mittels der Programmbibliothek Keras [7, 8].

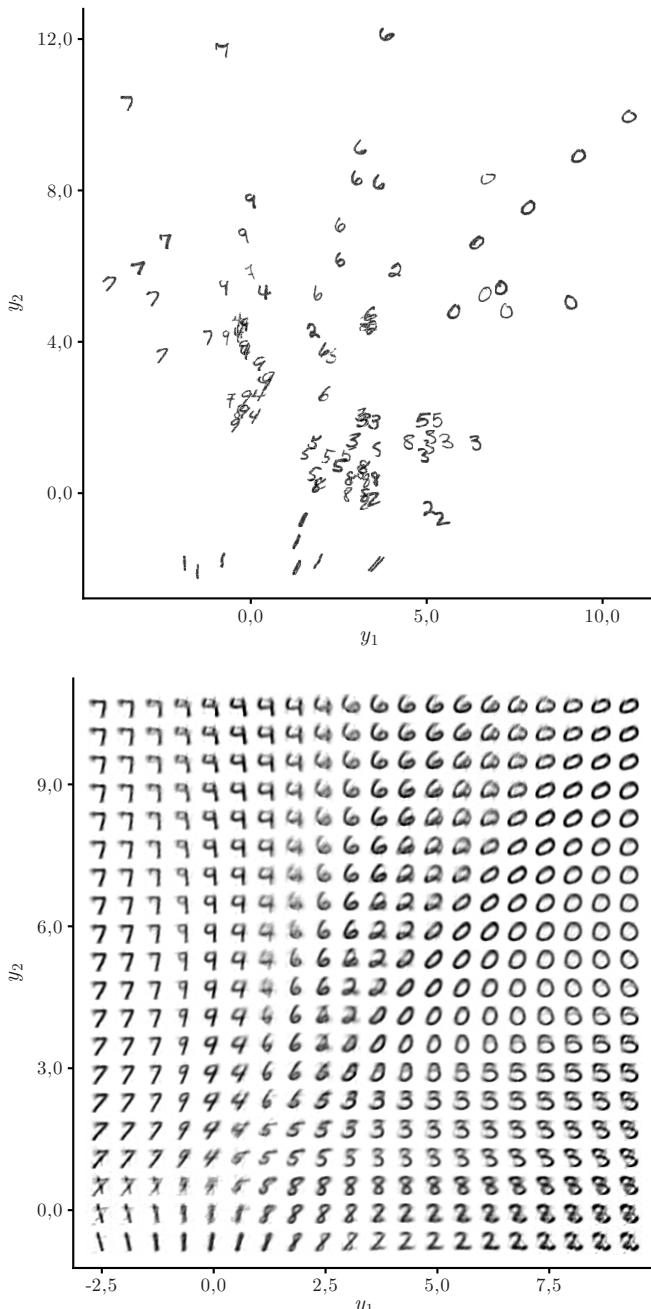
Die untere Abbildung ist eine Darstellung der Dekodierungsabbildung. Im Vergleich mit der Hauptkomponentenanalyse (Abb. 7.5 unten) zeigt sich, dass das neuronale Netzwerk augenscheinlich schon sehr viel besser in der Lage ist, den gesamten Merkmalsraum relativ verlustfrei auf eine niedrigdimensionale Darstellung zu komprimieren.

### 7.2.3 Multidimensionale Skalierung

Seien Datenpunkte  $x_1, \dots, x_N$  mit einer symmetrischen Prämetrik  $\delta(\cdot, \cdot)$  gegeben. Dabei kann es sich um Punkte in einem hochdimensionalen euklidischen Raum handeln, oder auch um andere Arten von Daten wie z. B. Listen binärer Merkmale, die mittels der Jaccard-Distanz verglichen werden. Wir möchten diese Datenpunkte auf eine Konfiguration von Zielpunkten  $y_1, \dots, y_N \in \mathbb{R}^K$  in einem euklidischen Raum mit vorgeschriebener, in der Regel vergleichsweise niedriger Dimension  $K$  überführen. Diese Zuordnung soll derart sein, dass sich die paarweisen Abstände zwischen den Punkten möglichst wenig ändern



**Abb. 7.5.** Streudiagramm der ersten zwei Hauptkoordinaten einer Auswahl von MNIST-Bildern (oben); Linearkombinationen der ersten zwei Hauptrichtungen (unten)



**Abb. 7.6.** Dimensionsreduktion durch einen Autoencoder: Streudiagramm der Kodierung einer Auswahl von MNIST-Bildern (oben); Dekodierungsabbildung (unten)

– in einem geeigneten Sinn soll für alle oder wenigstens die meisten Paare von Punkten gelten:

$$\Delta_{mn} = \delta(x_m, x_n) \approx \|y_m - y_n\|$$

Eine Idee besteht darin, eine solche Konfiguration durch die Optimierung geeigneter Zielfunktionen zu finden.

Sei eine Abstandsmatrix  $\Delta$  vom Format  $N \times N$  gegeben. Eine **metrische multidimensionale Skalierung** (MDS) besteht in der Minimierung der folgenden Zielfunktion:

$$R_{\text{mMDS}}(y_1, \dots, y_N) = \sum_{k=1}^N \sum_{l=1}^N (\Delta_{kl} - \|y_k - y_l\|)^2$$

Die **multidimensionale Skalierung nach Sammon** besteht in der Minimierung der folgenden Zielfunktion:

$$R_{\text{Samm}}(y_1, \dots, y_N) = \sum_{k=1}^N \sum_{l=1}^N \frac{(\Delta_{kl} - \|y_k - y_l\|)^2}{\Delta_{kl}}$$

Dabei werden nicht definierte Summanden mit verschwindendem Nenner gleich null gesetzt.

Die metrische multidimensionale Skalierung besteht in der Minimierung einer quadratischen Verlustfunktion mit dem Ziel, die paarweisen Abstände zwischen den Datenpunkten im Zielraum möglichst zu erhalten.

Die Variante nach Sammon unterscheidet sich von der grundlegenden metrischen MDS um den Faktor „ $1/\Delta_{kl}$ “ unter der Summe: Nah beieinander liegende Datenpunkte erhalten ein größeres Gewicht als weiter voneinander entfernt liegende. Auf diese Weise ist die Rekonstruktion der Abstände im Zielraum weniger global, in verstärktem Maße sollen lokale Nachbarschaftsbeziehungen erhalten bleiben. Dieses Ziel verfolgt auch das im nächsten Abschnitt besprochene t-SNE-Verfahren.

In der Literatur sind auch die folgenden normierten Zielfunktionen gebräuchlich, welche auf dieselben Minimalstellen führen:

$$S_{\text{mMDS}}(y_1, \dots, y_N) = \left( \sum_{k=1}^N \sum_{l=1}^N (\Delta_{kl})^2 \right)^{-1} \cdot \sum_{k=1}^N \sum_{l=1}^N (\Delta_{kl} - \|y_k - y_l\|)^2$$

$$S_{\text{Samm}}(y_1, \dots, y_N) = \left( \sum_{k=1}^N \sum_{l=1}^N \Delta_{kl} \right)^{-1} \cdot \sum_{k=1}^N \sum_{l=1}^N \frac{(\Delta_{kl} - \|y_k - y_l\|)^2}{\Delta_{kl}}$$

**Anwendungsbeispiel.** In Abb. 7.7 sind oben die Koordinaten einer multidimensionalen Skalierung nach Sammon dargestellt, angewandt auf einen Teil des MNIST-Datensatzes.

### 7.2.4 T-distributed Stochastic Neighbor Embedding (t-SNE)

Seien Datenpunkte  $x_1, \dots, x_N$  sowie eine symmetrische Prämetrik  $\delta(\cdot, \cdot)$  gegeben, mit der Abstände zwischen den Datenpunkten ermittelt werden können. Wir möchten die Datenpunkte auf Bildpunkte in einem Raum mit vorgeschriebener Dimension  $K$  überführen:  $y = (y_1, \dots, y_N)$  mit  $y_n \in \mathbb{R}^K$  für alle  $n \in \{1, \dots, N\}$ . Diese Zuordnung soll so konstruiert werden, dass sich Abstände zwischen *benachbarten* Punkten wenig ändern, in einem geeigneten Sinn soll also  $\Delta_{mn} = \delta(x_m, x_n) \approx \|y_m - y_n\|$  für *kleine*  $\Delta_{mn}$  gelten.

Die Grundidee des sogenannten **t-SNE-Verfahrens** ist wie folgt. Wir stellen uns eine **zufällige Schrittfolge** (engl. *random walk*) von Datenpunkt zu Datenpunkt vor: Befindet sich der Läufer an einem Punkt mit dem Index  $m$ , so begibt er sich beim folgenden Schritt mit der Wahrscheinlichkeit  $p(n|m)$  zum Datenpunkt mit dem Index  $n$ . Je geringer der Abstand von  $x_m$  und  $x_n$  sind, um so größer sei diese Wahrscheinlichkeit. Für alle  $m \in \{1, \dots, N\}$  gilt  $\sum_{n=1}^N p(n|m) = 1$ . Damit erhalten wir eine Familie von Massenfunktionen, welche die lokale, intrinsische Geometrie widerspiegelt. Die Verteilung der Datenpunkte im Zielraum soll möglichst große Ähnlichkeit mit dieser Ursprungsverteilung haben.

Um diese Idee weiter zu konkretisieren, definieren wir folgende Massenfunktionen; dabei sind die Bandbreiten  $\sigma = (\sigma_1, \dots, \sigma_N)$  noch zu bestimmende Parameter:

$$p_{x,\sigma}(n|m) = e^{-\frac{1}{2}\left(\frac{\Delta_{mn}}{\sigma_m}\right)^2} \cdot \left( -1 + \sum_{k=1}^N e^{-\frac{1}{2}\left(\frac{\Delta_{mk}}{\sigma_m}\right)^2} \right)^{-1}$$

sowie

$$q_y(m, n) = \frac{1}{1 + \|y_m - y_n\|^2} \cdot \left( -N + \sum_{k=1}^N \sum_{l=1}^N \frac{1}{1 + \|y_k - y_l\|^2} \right)^{-1}$$

falls  $m, n \in \{1, \dots, N\}$  und  $m \neq n$ , ansonsten ist  $p_{x,\sigma}(n|m) = q_y(m, n) = 0$ .

Für die Originalverteilung wird also eine Form von Kerndichteschätzer mit Gauß'schem Kern verwendet, während für die Zielverteilung eine endlastige Cauchy-Verteilung als Kern angesetzt wird. Hintergrund ist dabei, dass auf diese Weise ein „Zusammenklumpen“ der Bildpunkte vermieden wird. Die Cauchy-Verteilung  $\mathcal{L}(u|0, 1) \propto (1 + u^2)^{-1}$  stimmt für positive Argumente mit der *t*-Verteilung mit einem Freiheitsgrad überein, daher auch der Name des Verfahrens.

Sei  $e^H > 0$  ein vorgegebener Parameter, die **Perplexität**. Mit den obigen Definitionen besteht das ***t*-SNE-Verfahren** aus folgenden Berechnungen [9]:

1. Die Bandbreiten  $\sigma_1, \dots, \sigma_N$  werden anhand der vorgegebenen Perplexität so bestimmt, dass für alle  $m \in \{1, \dots, N\}$  gilt:

$$-\sum_{n=1}^N p_{x,\sigma_m}(n|m) \cdot \ln(p_x(n|m)) = H$$

2. Es wird festgesetzt:

$$p_x(n, m) = \frac{1}{2N} (p_{x,\sigma_m}(n|m) + p_{x,\sigma_n}(m|n))$$

3. Die Bildpunkte werden durch Minimierung der folgenden Zielfunktion bestimmt, der **Kreuzentropie** zwischen den Verteilungen  $p_x(\cdot, \cdot)$  und  $q_y(\cdot, \cdot)$ :

$$R(y_1, \dots, y_N) = -\sum_{k=1}^N \sum_{l=1}^N p_x(k, l) \ln(q_y(k, l))$$

Auch hier gilt die Vereinbarung „ $0 \cdot \ln 0 = 0$ “.

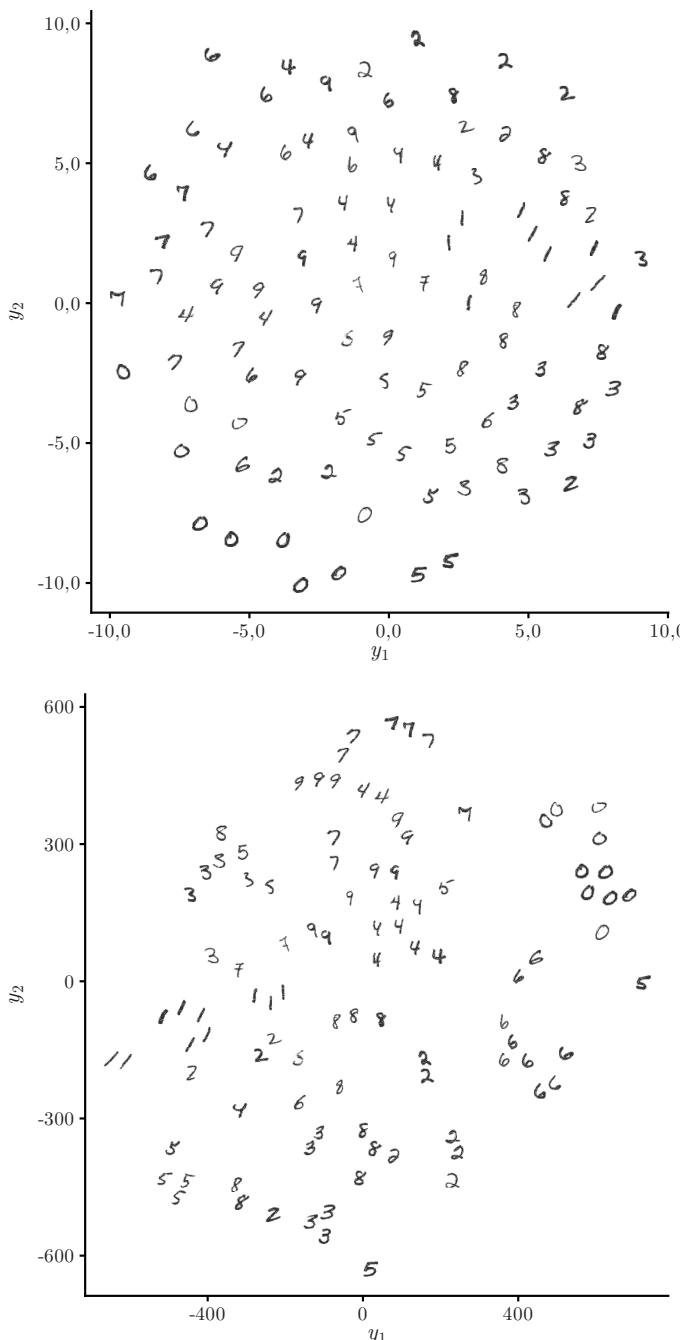
Typische Werte für die Wahl der Perplexität liegen bei  $5 \leq e^H \leq 50$ . Vorrangiges Anwendungsgebiet des *t*-SNE-Verfahrens ist die Visualisierung, daher gilt in der Regel  $K = 2$  oder  $K = 3$ . Eine algorithmisch ähnliche, aber neuere Methode ist das UMAP-Verfahren<sup>2</sup> [10].

Die numerische Implementierung wird dadurch erleichtert, dass der Gradient der obigen Zielfunktion explizit berechnet werden kann. Für alle  $n \in \{1, \dots, N\}$  gilt:

$$\nabla_{y_n} R(y_1, \dots, y_N) = 4 \sum_{k=1}^N \frac{p_x(n, k) - q_y(n, k)}{1 + \|y_n - y_k\|^2} \cdot (y_n - y_k)$$

**Anwendungsbeispiel.** Abb. 7.7 zeigt unten das Ergebnis des *t*-SNE-Verfahrens, angewandt auf einen Teil des MNIST-Datensatzes. Es wurde eine Perplexität von  $e^H = 12$  gewählt.

<sup>2</sup> „UMAP“ steht für *Uniform Manifold Approximation and Projection*.



**Abb. 7.7.** Multidimensionale Skalierung nach Sammon (oben);  $t$ -distributed Stochastic Neighbor Embedding ( $t$ -SNE) (unten)

## 7.3 Clusteranalyse

Vorrangiges Ziel einer Clusteranalyse ist die Partition eines Datensatzes  $x = (x_1, \dots, x_N) \in \mathcal{X}^N$  in Teilmengen  $S_1, \dots, S_K \subseteq \{x_1, \dots, x_N\}$ , den **Clustern**, wobei jede Beobachtung  $x_n$  in genau einem Cluster enthalten ist: Für alle  $n \in \{1, \dots, N\}$  gibt es genau ein  $k \in \{1, \dots, K\}$  mit  $x_n \in S_k$ .

Die oben beschriebene Zielstellung ist die einfachste Form der Clusteranalyse; es gibt auch Varianten, bei denen jede Beobachtung in *höchstens* einem Cluster enthalten ist, manche Datenpunkte also als Ausreißer betrachtet und *keinem* Cluster zugeordnet werden.

Ebenso gibt es Algorithmen, welche einen Datenpunkt mehreren Clustern zuordnen. In der Regel wird dann die Zuordnung eines Datenpunktes  $x_n$  zu einem Cluster  $S_k$  noch mit einer gewichtenden Kennzahl  $w_{nk} \in ]0, 1]$  charakterisiert, welche den Grad der Zugehörigkeit widerspiegeln soll. Die Cluster können in diesem Fall als **unscharfe Mengen** (auch: Fuzzy-Mengen) aufgefasst werden.

Über eine Partition des vorliegenden Datensatzes hinaus gehend kann das Ziel einer Clusteranalyse im Erlernen einer Entscheidungsregel  $f: \mathcal{X} \rightarrow \{1, \dots, K\}$  bestehen. Diese ermöglichte eine Clusterzuordnung von Datenpunkten, die nicht notwendigerweise im ursprünglichen Datensatz enthalten waren. Liefert das Verfahren keine solche Entscheidungsregel mit, so kann diese mithilfe eines Klassifikationsverfahrens anhand des nun mit Klassenlabels versehenen Datensatzes im Nachhinein erzeugt werden.

Die angestrebte Zuordnung von Clustern ist natürlich nicht beliebig: In der Regel sollen die Datenpunkte innerhalb eines Clusters einen geringen Abstand oder eine hohe Ähnlichkeit bzgl. eines geeigneten Maßes haben, während der Abstand bzw. die Ähnlichkeit von Datenpunkten zweier verschiedener Cluster wiederum verhältnismäßig groß sein soll.

### 7.3.1 K-Means-Verfahren

Das *K*-Means-Verfahren kann auf Grundlage der Methoden der Dichteschätzung wie folgt motiviert werden. Betrachten wir hierzu eine Dichtefunktion, welche eine spezielle Form eines **Gauß'schen Mischmodells**  $p_{\text{GMM}}: \mathbb{R}^D \rightarrow \mathbb{R}$  darstellt:

$$p_{\text{GMM}}(u|\mu_1, \dots, \mu_K, h^2) = \frac{1}{K} \cdot \frac{1}{(2\pi)^{\frac{D}{2}}} \sum_{k=1}^K \exp\left(-\frac{\|u - \mu_k\|^2}{2h^2}\right)$$

mit den Lagevektoren  $\mu_1, \dots, \mu_K \in \mathbb{R}^D$  und einer als universell angenommenen Bandbreite  $h > 0$ . Jeder der Lagevektoren bestimmt den Schwerpunkt einer multivariaten Normalverteilung, die bei Realisierung einen Cluster von Datenpunkten erzeugt. Wir nehmen an, dass die Gauß'schen Verteilungen in der obigen Summe eine geringe Überlappung aufweisen. Anders ausgedrückt: Die Clusterschwerpunkte haben – in Vielfachen von  $h$  gemessen – paarweise

hinreichend großen Abstand voneinander. Mit dieser Annahme können wir die Dichtefunktion näherungsweise wie folgt darstellen:

$$p_{\text{GMM}}(u|\mu_1, \dots, \mu_K, h^2) \approx \frac{1}{K} \cdot \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{\|u - \mu_{f(u)}\|^2}{2h^2}\right)$$

Dabei ist  $\mu_{f(u)} \in \{\mu_1, \dots, \mu_k\}$  der Clusterschwerpunkt, der zu  $u \in \mathbb{R}^D$  den geringsten Abstand hat. Wir nehmen an, dass dieser den entscheidenden Beitrag zur Dichte liefert. Die Entscheidungsregel  $f: \mathbb{R}^D \rightarrow \{1, \dots, K\}$  ordnet jeden Punkt seinem Clusterlabel zu und ist bereits im Wesentlichen – mit Ausnahme von Entscheidungsgrenzen, die gleichen Abstand zu mehr als einem Schwerpunkt haben – durch die Clusterschwerpunkte bestimmt:  $f(\cdot) = f(\cdot; \mu_1, \dots, \mu_K)$ .

Eine Partition  $S_1, \dots, S_K$  eines Datensatzes  $x = (x_1, \dots, x_N)$  mit  $x_n \in \mathbb{R}^D$  für alle  $n \in \{1, \dots, N\}$  auf Basis einer solchen Entscheidungsregel lässt sich wie folgt charakterisieren:

$$S_k = \{x_n | n \in \{1, \dots, N\}, f(x_n) = k\}$$

für alle  $k \in \{1, \dots, K\}$ . Ebensogut können wir aber auch die Indizes der Datenpunkte zu Clustern zusammenfassen:

$$I_k = \{n \in \{1, \dots, N\} | f(x_n) = k\} = \{n \in \{1, \dots, N\} | x_n \in S_k\}$$

Die zu dem Modell und dem Datensatz gehörige Log-Likelihood-Funktion berechnet sich näherungsweise wie folgt:

$$\begin{aligned} \ell(\mu_1, \dots, \mu_K, h^2) &= \sum_{n=1}^N \ln(p_{\text{GMM}}(x_n|\mu_1, \dots, \mu_K, h)) \\ &\approx -\frac{1}{2h^2} \cdot \sum_{n=1}^N \|x_n - \mu_{f(x_n; \mu_1, \dots, \mu_K)}\|^2 \\ &\quad - N \cdot \ln(K) - \frac{ND}{2} \ln(2\pi) \end{aligned}$$

Die Maximierung der Log-Likelihood-Funktion in Näherung führt auf die Minimierung der Summe der quadrierten Abstände zu den jeweiligen Clusterschwerpunkten:

$$R = \sum_{n=1}^N \|x_n - \mu_{f(x_n; \mu_1, \dots, \mu_K)}\|^2 = \sum_{k=1}^K \sum_{n \in I_k} \|x_n - \mu_k\|^2$$

In Worten: Es werden innerhalb jeden Clusters die quadrierten Abstände der Datenpunkte zum Clusterschwerpunkt summiert, und diese Werte werden wiederum addiert.

Für eine optimale Belegung der Clusterlabels  $I_1, \dots, I_K$  müssen  $\mu_1, \dots, \mu_K$  notwendigerweise die jeweiligen empirischen Schwerpunkte der Cluster sein, da auf diese Weise jeder der nichtnegativen Summanden minimiert wird. Diese Überlegungen führen auf das folgende Verfahren.

Seien  $x_1, \dots, x_N \in \mathbb{R}^D$  Datenpunkte. Das  **$K$ -Means-Verfahren** der Clusteranalyse besteht in der Minimierung der folgenden Zielfunktion über den möglichen Belegungen von Clusterlabels  $I_1, \dots, I_K$ :

$$R(I_1, \dots, I_K) = \sum_{k=1}^K \sum_{n \in I_k} \|x_n - \mu_k\|^2 \text{ mit } \mu_k = \frac{1}{|I_k|} \sum_{n \in I_k} x_n$$

Der folgende **Lloyd'sche Algorithmus** ist geeignet, ein lokales Minimum der Zielfunktion aufzufinden:

```

initialisiere Schwerpunkte  $\mu_1, \dots, \mu_K$ 
solange der Wert von  $R$  verändert sich mit jeder Iteration true
  für  $i = 1$  bis  $N$  true
    |  $f(x_i) :=$  Index des nächstgelegenen Schwerpunkts
  Ende
  für  $j = 1$  bis  $K$  true
    |  $\mu_j :=$  Schwerpunkt aller Datenpunkte im  $j$ -ten Cluster
  Ende
  berechne neu:  $R = \sum_{n=1}^N \|x_n - \mu_{f(x_n)}\|^2$ 
Ende
Ausgabe:
Clusterschwerpunkte  $\mu_1, \dots, \mu_K$ 
Clusterzugehörigkeiten  $f(x_1), \dots, f(x_N)$ 
```

Die Initialisierung der Schwerpunkte kann z.B. durch zufällige Auswahl von  $K$  Punkten aus  $x_1, \dots, x_N$  durchgeführt werden. Der Lloyd'sche Algorithmus findet in jedem Fall ein *lokales* Minimum auf – eine ungünstige Initialisierung kann jedoch zu einem vom globalen Optimum weit entfernten und daher minderwertigen Ergebnis führen. Daher empfiehlt es sich, den Algorithmus mit verschiedenen initialen Schwerpunkten mehrmals auszuführen und die Ergebnisse, etwa anhand des finalen Gütemaßes  $R$ , miteinander zu vergleichen.

Abb. 7.8 zeigt die Anwendung des Verfahrens auf einen synthetischen Datensatz, der durch ein Gauß'sches Mischmodell mit drei Schwerpunkten erzeugt wurde. Die Qualität des Ergebnisses hängt mit entscheidend vom Hyperparameter  $K$  ab, der die Anzahl der aufzufindenden Cluster festlegt.

Varianten des  $K$ -Means-Verfahrens stellen  **$K$ -Medoids-Verfahren** dar [11]. Bei diesen werden anstelle der Schwerpunkte der Cluster deren Medoiden als Lagemaß herangezogen. Ein Vorteil besteht in der Möglichkeit, ein allgemeines Abstandsmaß  $\delta(\cdot, \cdot)$  zugrundezulegen.

## K-Means-Verfahren mit Kern

Wie aus dessen Herleitung ersichtlich wird, macht das *K-Means-Verfahrens* einige recht restriktive Annahmen: Die Cluster haben gleiche Größe und jeder Cluster ist radialsymmetrisch bezüglich seines jeweiligen Schwerpunkts. Sind diese Annahmen nicht gerechtfertigt, kann die Clusteranalyse ein minderwertiges Ergebnis liefern, wie etwa das Beispiel in Abb. 7.9 oben zeigt.

Den Beschränkungen des gewöhnlichen *K-Means-Verfahrens* kann mit der Anwendung der Kernel-Methode begegnet werden. Wir legen eine Merkmalsabbildung  $\phi: \mathbb{R}^D \rightarrow \mathcal{X}$  zugrunde, wobei die Dimension des Zielraums  $\mathcal{X}$  in der Regel wesentlich größer als  $D$  oder gar unendlich ist. Das Skalarprodukt  $\langle \cdot, \cdot \rangle$  im Zielraum wird durch einen Kern  $\sigma: \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$  vermittelt, sodass für alle  $u, v \in \mathbb{R}^D$  gilt:

$$\langle \phi(u), \phi(v) \rangle = \sigma(u, v)$$

Es kann z. B. wieder ein Gauß'scher Kern verwendet werden:  $\sigma_h(u, v) = \exp(-\frac{1}{2}h^{-2}\|u - v\|^2)$ .

Weiterhin schreiben wir für die Clusterschwerpunkte unter der Merkmalsabbildung:

$$\phi^* \mu_k := \frac{1}{|I_k|} \sum_{n \in I_k} \phi(x_n)$$

Für den quadrierten Abstand eines transformierten Datenpunktes  $u \in \mathbb{R}^D$  zum  $k$ -ten Clusterschwerpunkt ergibt sich:

$$\begin{aligned} r_k^2(u) &= \|\phi(u) - \phi^* \mu_k\|^2 \\ &= \langle \phi(u), \phi(u) \rangle - 2\langle \phi(u), \phi^* \mu_k \rangle + \langle \phi^* \mu_k, \phi^* \mu_k \rangle \\ &= \sigma(u, u) - \frac{2}{|I_k|} \sum_{m \in I_k} \sigma(u, x_m) + \frac{1}{|I_k|^2} \sum_{m \in I_k} \sum_{n \in I_k} \sigma(x_m, x_n) \end{aligned}$$

Seien  $x_1, \dots, x_N \in \mathbb{R}^D$  Datenpunkte. Das ***K-Means-Verfahren mit Kern***  $\sigma: \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$  besteht in der Minimierung der folgenden Zielfunktion über den möglichen Belegungen von Clusterlabels  $I_1, \dots, I_K$ :

$$R(I_1, \dots, I_K) = \sum_{k=1}^K \sum_{n \in I_k} r_k^2(x_n),$$

wobei sich  $r_k^2(x_n)$  wie aus der Herleitung oben ersichtlich berechnet.

Der Lloyd-Algorithmus kann an die neue Situation wie folgt angepasst werden:

```

initialisiere Clusterlabels für  $x_1, \dots, x_N$ 
solange der Wert von  $R$  verändert sich mit jeder Iteration true
  für  $i = 1$  bis  $N$  true
    neues Clusterlabel von  $x_i$  := Index  $k$  mit kleinstem Wert für
       $r_k^2(x_i)$ 
  Ende
  berechne neu:  $R = \sum_{k=1}^K \sum_{n \in I_k} r_k^2(x_n)$ 
Ende
Ausgabe: Clusterlabels für  $x_1, \dots, x_N$ 
```

In Abb. 7.9 ist unten das Ergebnis einer  $K$ -Means-Clusteranalyse mit Kern für synthetische Daten dargestellt ( $K = 4$ , Gauß'scher Kern mit  $h = 0,5$ ).

### 7.3.2 Hierarchische Clusteranalyse

Die Grundidee des Algorithmus einer **agglomerativen hierarchischen Clusteranalyse** ist die folgende. Zunächst wird mit der feinsten möglichen Partition  $S^{(0)}$  des Datensatzes initialisiert: Jeder Cluster enthält genau einen Datenpunkt. Mit jedem folgenden Verfahrensschritt werden dann jeweils die Cluster fusioniert, in denen die jeweiligen Datenpunkte insgesamt minimalen Abstand zueinander haben. Die so konstruierte neue Partition besteht aus weniger Clustern, ist also größer. Das Verfahren wird nach  $N - 1$  Schritten abgebrochen, wenn die konstruierte Partition  $S^{(N-1)}$  nur noch aus einem einzelnen Cluster besteht, der alle Datenpunkte enthält.

Eine **agglomerative hierarchische Clusteranalyse** erzeugt eine Reihe von Partitionen eines Datensatzes  $x_1, \dots, x_N$  mit Abstandsmaß  $\delta(\cdot, \cdot)$  wie folgt:

```

initialisiere  $t := 0$ ,  $S^{(0)} := \{S_1^{(0)}, \dots, S_N^{(0)}\} = \{\{x_1\}, \dots, \{x_N\}\}$ 
solange  $|S^{(t)}| > 1$  true
  aktualisiere  $t \leftarrow t + 1$ , initialisiere  $S^{(t)} := S^{(t-1)}$ 
  # Bestimme Cluster mit kleinster Distanz:
  für  $i = 2$  bis  $|S^{(t)}|$  true
    für  $j = 1$  bis  $i - 1$  true
      berechne  $\Delta_{ij} := D(S_i^{(t)}, S_j^{(t)})$  (siehe unten)
    Ende
  Ende
  bestimme  $i, j$  mit kleinster Distanz  $\Delta_{ij}$ 
  # Fusioniere Cluster mit kleinster Distanz:
  aktualisiere  $S_i^{(t)} \leftarrow S_i^{(t)} \cup S_j^{(t)}$ , lösche  $S_j^{(t)}$ 
Ende
Ausgabe: Partitionen  $S^{(0)}, S^{(1)}, \dots, S^{(T)}$ 
```

Dabei ordnet  $D(\cdot, \cdot)$  zwei Clustern deren Distanz zu. Üblich sind folgende Definitionen:

$$D_{\min}(A, B) = \min_{u \in A, v \in B} \delta(u, v),$$

$$D_{\text{avg}}(A, B) = \langle \delta(u, v) \rangle_{u \in A, v \in B},$$

$$D_{\max}(A, B) = \max_{u \in A, v \in B} \delta(u, v)$$

Je nach Verwendung von  $D_{\min}$ ,  $D_{\text{avg}}$  oder  $D_{\max}$  als Abstandsmaß zwischen den Clustern wird von einem **Single-Linkage-**, **Average-Linkage-** oder **Complete-Linkage-Verfahren** gesprochen.

Im Falle einer **divisiven hierarchischen Clusteranalyse** werden die Partitionen, ausgehend von einem einzelnen Cluster, fortlaufend verfeinert. Auf diese Verfahren gehen wir hier nicht näher ein.

Das Ergebnis einer hierarchischen Clusteranalyse kann grafisch mithilfe eines **Baumdiagramms** dargestellt werden. Abb. 7.10 zeigt ein solches Baumdiagramm. Es entstand durch eine hierarchische Clusteranalyse anhand der geografischen Distanzen von vierzig deutschen Städten im Complete-Linkage-Verfahren. Die Blätter des Baums entsprechen den zu gruppierenden Entitäten, und jede Gabelung repräsentiert die Fusion zweier Cluster. Die Position der Gabelung ist ebenfalls von Bedeutung, an ihr kann die Wert der Clusterdistanzfunktion  $D(\cdot, \cdot)$  abgelesen werden, bei dem die Fusion stattfand.

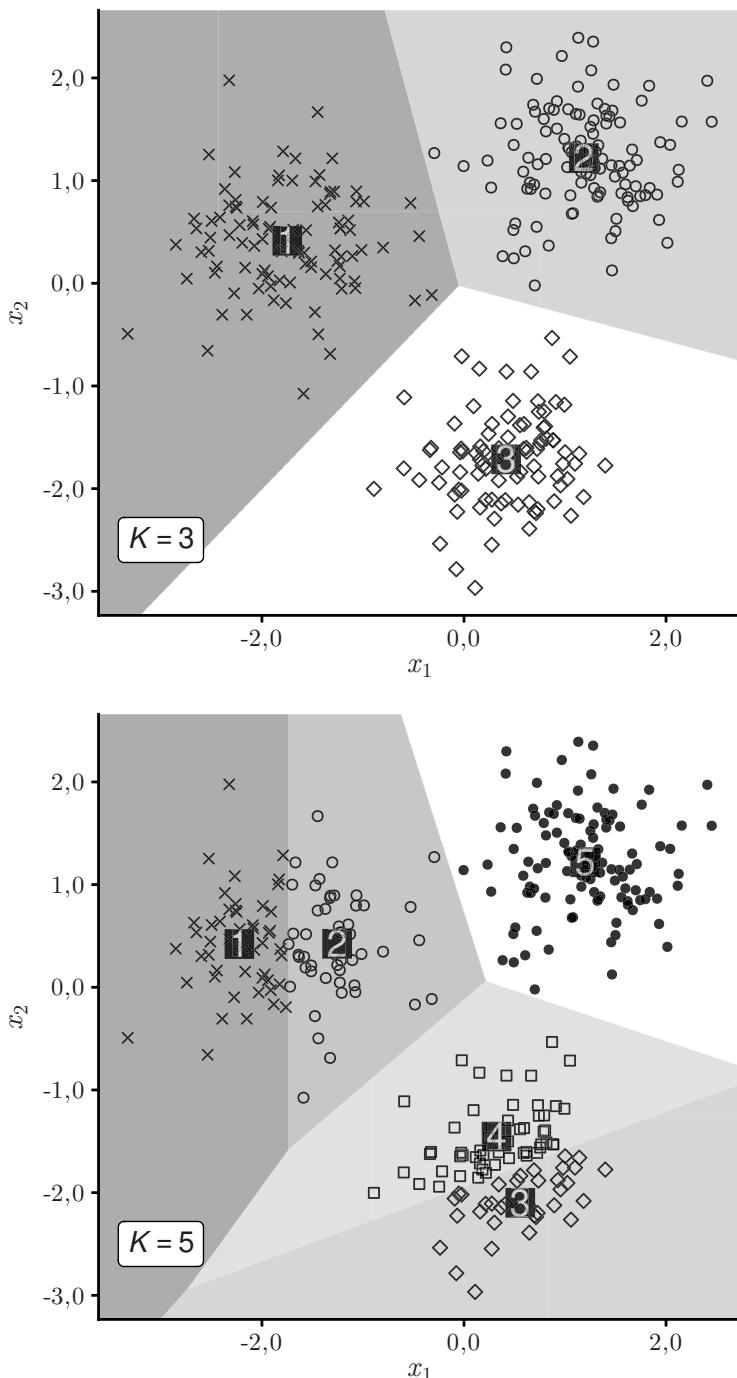


Abb. 7.8. Clusteranalyse mit dem  $K$ -Means-Verfahren

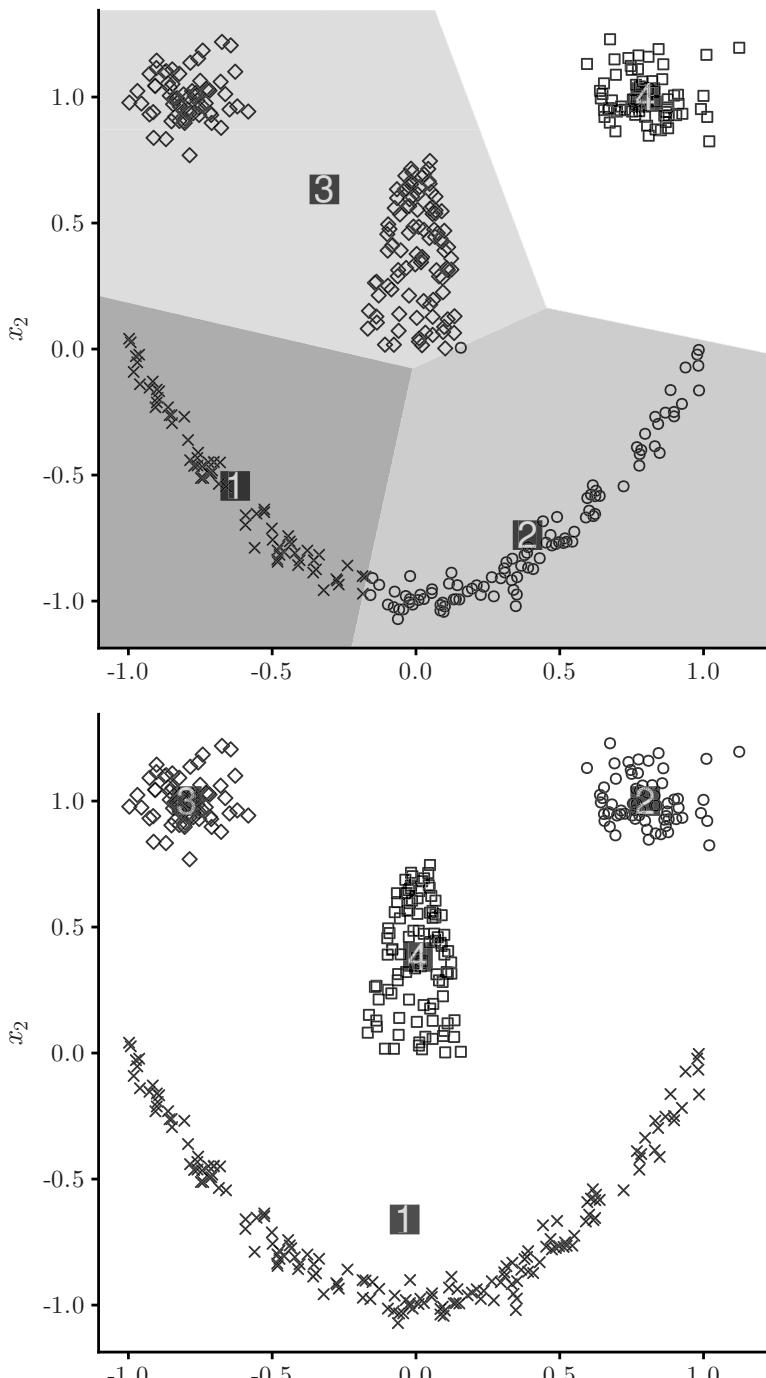
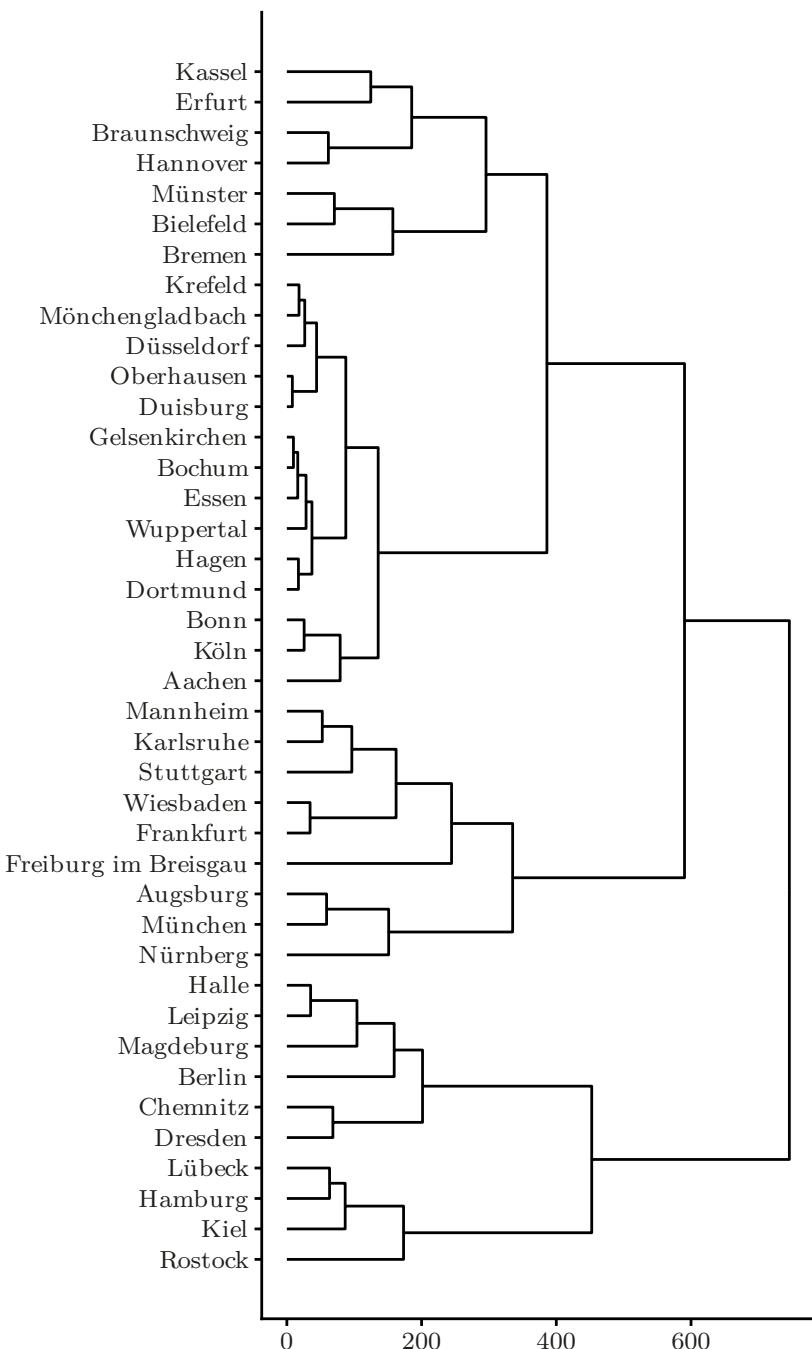


Abb. 7.9. K-Means-Clusteranalyse: gewöhnlich (oben) und mit Kern (unten)



**Abb. 7.10.** Baumdiagramm einer hierarchischen Clusteranalyse deutscher Städte

## Quellen

- [1] Sanjoy Dasgupta und Anupam Gupta. „An elementary proof of a theorem of Johnson and Lindenstrauss“. In: *Random Structures & Algorithms* 22.1 (2003), S. 60–65. DOI: [10.1002/rsa.10073](https://doi.org/10.1002/rsa.10073).
- [2] Mehryar Mohri, Afshin Rostamizadeh und Ameet Talwalkar. *Foundations of Machine Learning*. 2. Aufl. MIT Press, 2018. ISBN: 978-0-262-03940-6.
- [3] Sixue Gong, Vishnu Naresh Boddeti und Anil K. Jain. „On the Intrinsic Dimensionality of Image Representations“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Juni 2019. [arXiv:1803.09672](https://arxiv.org/abs/1803.09672).
- [4] Christian S. Perone, Roberto Silveira und Thomas S. Paula. *Evaluation of sentence embeddings in downstream and linguistic probing tasks*. Juni 2018. [arXiv:1806.06259v1](https://arxiv.org/abs/1806.06259v1).
- [5] Gerd Fischer. *Lineare Algebra*. 18. Aufl. Springer Spektrum, Wiesbaden, 2014. DOI: [10.1007/978-3-658-03945-5](https://doi.org/10.1007/978-3-658-03945-5).
- [6] Yann LeCun, Corinna Cortes und Christopher J. C. Burges. *The MNIST database of handwritten digits*. 2010. URL: <http://yann.lecun.com/exdb/mnist/>.
- [7] François Collet u. a. *Keras*. URL: <https://keras.io>.
- [8] J. J. Allaire und François Chollet. *keras: R Interface to 'Keras'*. R-Paket, Version 2.3.0.0. 2020. URL: <https://CRAN.R-project.org/package=keras>.
- [9] Laurens J. P. van der Maaten und Geoffrey E. Hinton. „Visualizing High-Dimensional Data Using t-SNE“. In: *Journal of Machine Learning Research* 9 (Nov. 2008), S. 2579–2605.
- [10] Leland McInnes, John Healy und James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Sep. 2020. [arXiv:1802.03426v3](https://arxiv.org/abs/1802.03426v3).
- [11] Erich Schubert und Peter J. Rousseeuw. *Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms*. Okt. 2019. [arXiv:1810.05691v4](https://arxiv.org/abs/1810.05691v4).



## Maschinelles Lernen in der Anwendung

Datenwissenschaftliche und statistische Verfahren im Allgemeinen und Methoden des maschinellen Lernens im Besonderen finden breite Anwendung in vielfältigen Bereichen von Wissenschaft und Technik. Um nur ein paar Beispiele solcher Bereiche zu nennen:

- Für die Steuerung autonomer Fahrzeuge [1],
- in der medizinischen Bildverarbeitung und -analyse [2], beispielsweise für die Diagnose von Erkrankungen wie COVID-19 auf Grundlage computertomographischer Thoraxaufnahmen [3, 4],
- bei der Verarbeitung astronomischer Daten [5], etwa für die morphologische Klassifikation von Galaxien [6, 7] oder die Entdeckung neuer Exoplaneten [8, 9],
- zur Betrugserkennung und -prävention im Kapitalverkehr [10],
- bei der Implementierung von Gesten- und Spracherkennung für Mensch-Maschine-Kommunikation [11, 12].

Mit der Verbreitung immer leistungsfähigerer Verfahren auf Grundlage stetig wachsender Datenbestände stellen sich vermehrt auch ethische Fragen beim **Datenschutz** [13] und dem Einsatz künstlicher Intelligenz [14, 15].

### 8.1 Anwendungsbeispiele für überwachtes Lernen

In den folgenden Abschnitten demonstrieren wir die Anwendung von Verfahren für die automatisierte Kategorisierung von Daten: Zum einen die Klassifikation von digitalen Bildern, zum anderen von Textdokumenten.

### 8.1.1 MNIST: Handschrifterkennung

Der MNIST-Datensatz [16, 17] besteht aus 70.000 digitalen Bildern<sup>1</sup> in Graustufen mit einer Auflösung von  $28 \times 28$  Pixeln. Jedes der Bilder zeigt eine handschriftlich notierte Ziffer, siehe Abb. 8.6 oben. Jedes Bild ist mit der notierten Ziffer vorklassifiziert. Die Klassifikationsaufgabe besteht im automatisierten Erkennen der Ziffer anhand der handschriftlichen Notierung. Eine rein regelbasierte Klassifikation erscheint impraktikabel oder zumindest sehr aufwändig, sodass Methoden des maschinellen Lernens zum Einsatz kommen.

Wir wollen das folgende vereinfachte, binäre Klassifikationsproblem lösen: Das Erkennen der Ziffer Eins im Unterschied zu allen anderen Ziffern. Zunächst wandeln wir die Bilder um, indem die Grauwerte der Pixel jedes Bilds einfach in einer Folge aneinander gereiht werden. Diese Transformation wird auch **Flattening** genannt. Dadurch entstehen aus den Grauwertmatrizen einfache Vektoren aus  $\mathbb{R}^D$  mit  $D = 28 \cdot 28 = 784$  Einträgen.

Der Datensatz wird dann in einen Trainingsdatensatz von 60.000 Bildern und einen Testdatensatz von 10.000 aufgeteilt. Mit dem Trainingsdatensatz können die in den vorigen Abschnitten vorgestellten Algorithmen trainiert werden. Die Güte der trainierten Modelle wird anhand von Kennzahlen wie Genauigkeit, Trefferquote und  $F_1$ -Maß bewertet, die über die Klassifikation der Bilder im Testdatensatz ermittelt werden.

Die folgende Tabelle zeigt eine Gegenüberstellung der Verfahren von logistischer Regression und linearer Diskriminanzanalyse, dem Ein-nächste-Nachbarn-Verfahren und einem Feedforward-Netzwerk:

	Genauigkeit	Trefferquote	$F_1$ -Maß
logist. Regression	93,1 %	97,1 %	95,1 %
LDA	89,6 %	95,9 %	92,6 %
1-NN	96,7 %	99,5 %	98,1 %
neuronales Netz	98,7 %	99,0 %	98,9 %
„Münzwurf“	11,4 %	50,0 %	18,5 %

**Tabelle 8.1.** Güte verschiedener Klassifikatoren für den MNIST-Datensatz

Insgesamt schneiden für diesen Datensatz alle Verfahren vergleichsweise gut ab. Zu Vergleichszwecken stehen in der letzten Zeile die Kennzahlen, die von einem Algorithmus erwartet werden können, der mit einer Wahrscheinlichkeit von 50 % eine Klassenzugehörigkeit zufällig auswählte.

Um die lineare Diskriminanzanalyse durchzuführen, wurden die empirischen Kovarianzmatrizen  $\Sigma_0$  und  $\Sigma_1$  sowie die geometrischen Schwerpunkte  $\mu_0$  und  $\mu_1$  aus den Trainingsdaten mit der Klassenzugehörigkeit  $y_n = 0$  bzw.  $y_n = 1$  ermittelt, um die jeweilige Likelihood durch eine multivariate Normalverteilung zu modellieren:

<sup>1</sup> „MNIST“ steht für Modified National Institute of Standards and Technology.

$$p(x|y=k) = \mathcal{N}(x|\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^D \cdot \det(\Sigma_k)}} \cdot e^{-\frac{1}{2}(x-\mu_k)^T \cdot (\Sigma_k)^{-1} \cdot (x-\mu_k)}$$

mit  $k \in \{0, 1\}$ . Da die  $\Sigma_k$  jedoch singulär und somit nicht invertierbar sind bzw. verschwindende Determinante haben, wurden stattdessen **regularisierte Kovarianzmatrizen** verwendet:  $\Sigma_{k,\varepsilon} = \Sigma_k + \text{diag}(\varepsilon, \varepsilon, \dots, \varepsilon)$  mit einem kleinen gewählten Glättungsparameter  $\varepsilon > 0$ .

Der optimale Hyperparameter  $K = 1$  für die  $K$ -nächste-Nachbarn-Klassifikation kann durch Kreuzvalidierung ermittelt werden. Das folgende Diagramm zeigt die Spannweite und den arithmetischen Mittelwert des  $F_1$ -Maßes über  $K$  bei 6-facher Kreuzvalidierung:

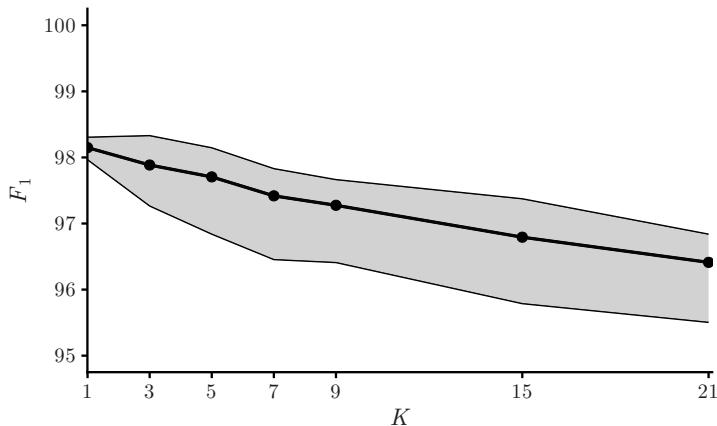


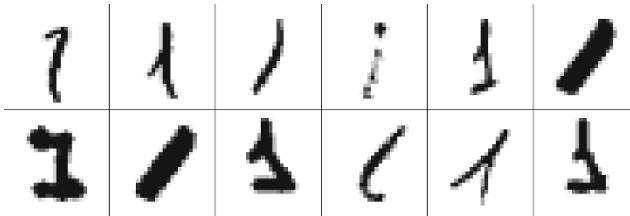
Abb. 8.1. Kreuzvalidierung eines KNN-Klassifikators

Das durch das neuronale Netzwerk trainierte Modell zeigt die größte Güte im Sinne des  $F_1$ -Maßes. Dennoch identifizierte das Verfahren fälschlich die folgenden Abbildungen als solche der Ziffer Eins:



Abb. 8.2. Falsch positive Ergebnisse der Klassifikation von MNIST-Ziffern

Umgekehrt wurden folgende handschriftliche Varianten der Ziffer Eins nicht als solche erkannt:



**Abb. 8.3.** Falsch negative Ergebnisse der Klassifikation von MNIST-Ziffern

Das für obige Klassifikation verwendete neuronale Netzwerk besteht aus drei verborgenen Schichten mit 128, 64 und 32 Neuronen. Für die verborgenen Schichten wird ein durchlässiger Gleichrichter als Aktivierungsfunktion verwendet. Die Ausgabeschicht besteht aus nur einem mithilfe einer Sigmoidfunktion aktvierten Neuron und als Verlustfunktion wurde die Kreuzentropie gewählt. Der Trainingsfehler auf Basis dieser Verlustfunktion wurde mittels eines Gradientenabstiegs bei einer Batch-Größe von jeweils zwanzig Trainingsbeispielen minimiert.

### 8.1.2 CIFAR-10: Objekterkennung

Der CIFAR-10-Datensatz [18] besteht aus 60.000 digitalen RGB-Farbbildern<sup>2</sup> mit einer Auflösung von  $32 \times 32$  Pixeln. Dies entspricht bei drei Farbkanälen einer extrinsischen Dimension der Eingangsdaten von  $D = 32 \cdot 32 \cdot 3 = 3072$ . Auf jedem Bild ist ein Objekt abgebildet, das einer der folgenden zehn Objektklassen zugeordnet und entsprechend vorklassifiziert ist – siehe Abb. 8.6 unten: Flugzeug, Pkw, Vogel, Katze, Hirsch, Hund, Frosch, Pferd, Schiff, Lkw.

Es dienen 10.000 der Bilder als Testdatensatz. Wir wollen Klassifikationsverfahren einsetzen, um in diesem Datensatz automatisiert Bilder von Katzen zu identifizieren. Eine Gegenüberstellung der Verfahren, die schon für die MNIST-Aufgabe herangezogen wurden, ergibt die folgende Tabelle:

	Genauigkeit	Trefferquote	$F_1$ -Maß
logist. Regression	22 %	27 %	24 %
LDA	16 %	29 %	21 %
1-NN	29 %	24 %	26 %
neuronales Netz	41 %	9 %	15 %
„Münzwurf“	10 %	50 %	17 %

**Tabelle 8.2.** Güte verschiedener Klassifikatoren für den CIFAR-10-Datensatz

Die Güte der Verfahren bewegt sich großenordnungsmäßig im Bereich zufälligen Ratens!

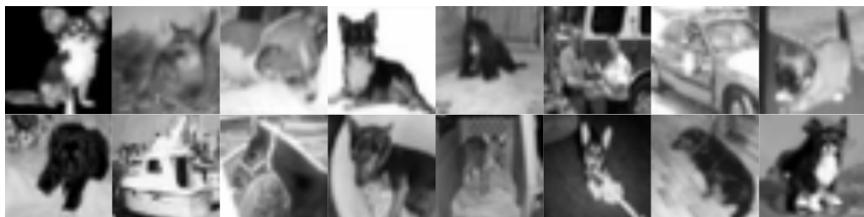
Mithilfe eines tiefen Convolutional Neural Networks kann jedoch ein Klassifikator trainiert werden, der ein  $F_1$ -Maß von 60 % erzielt, bei einer Genauigkeit von

<sup>2</sup> „CIFAR-10“ steht für Canadian Institute for Advanced Research, 10 Objektklassen.

70 % und einer Trefferquote von 53 %. Der wesentliche Teil des Programmcodes in R (vgl. [19]) ist mit Abb. 8.7 aufgeführt, für die Implementierung wurden die Programmhbibliotheken Keras [20, 21] und TensorFlow [22, 23] verwendet. Das Training kann auf gebrauchsüblicher Hardware durchgeführt werden. Der mittlere Verlust wurde mithilfe von Adam<sup>3</sup> minimiert [24], einem effizienten Algorithmus für numerische Optimierung auf dem aktuellen Stand der Technik (2020). Darüber hinaus wurde für die letzte verborgene, voll vernetzte Schicht ein Dropout mit einer Ausfallwahrscheinlichkeit von  $q = 0,5$  eingesetzt.

Eine weitere Komponenten der Architektur besteht in einer sogenannten Batch-Standardisierung [25]. Bei diesem Verfahren werden die Aktivierungen mit jeder Iteration über den Mini-Batch hinweg standardisiert, d. h., mittelwertzentriert und auf eine Varianz von eins gebracht. Dies kann zu einer verbesserten numerischen Stabilität führen.

Folgende Testbeispiele klassifizierte das Verfahren fälschlich als Abbildungen einer Katze – in der Mehrzahl handelt es sich um Fotos von Hunden:



**Abb. 8.4.** Falsch positive Ergebnisse der Klassifikation von CIFAR-10-Bildern

Umgekehrt wurden folgende Fotos einer Katze nicht als solche erkannt:



**Abb. 8.5.** Falsch negative Ergebnisse der Klassifikation von CIFAR-10-Bildern

Das relativ neue Forschungsfeld der **Explainable AI** (xAI) befasst sich mit der Aufgabe, für den Menschen und im einzelnen Anwendungsfall nachvollziehbar zu machen, auf welche Weise komplexe Verfahren wie Deep Learning zu bestimmten Ergebnissen gelangen.

Wird das Ausgabeneuron durch eine Schicht mit zehn Softmax-aktivierten Neuronen ersetzt, so kann mit der Architektur auch das vollständige Problem der Einordnung der Fotos in alle Bildklassen in Angriff genommen werden. Die

---

<sup>3</sup> „Adam“ steht für *Adaptive Moment Estimation*.

Korrektklassifikationsrate beträgt dann 80 %, mithilfe einer Datenaugmentation kann diese auf 90 % gesteigert werden [19].

Nach aktuellem Stand der Technik können auf dem CIFAR-10-Datensatz Korrektklassifikationsraten jenseits der 99 % erzielt werden [26, 27, 28, 29]. Zum Vergleich: Ein Mensch kann etwa 94 % der Bilder korrekt zuordnen [30].

### 8.1.3 Large Movie Review Dataset: Sentimentanalyse

Aufgabe einer **Sentimentanalyse** ist im weitesten Sinne die systematische Identifikation, Extraktion und Bemessung von Informationen, die durch ein subjektives Empfinden gekennzeichnet sind. Ein biometrisches Bild- oder Videoerfassungssystem könnte etwa darauf trainiert werden zu erkennen, in welchem Gemütszustand sich eine erfasste Person befindet: Zeigt sich die Person fröhlich, traurig, wütend usw.

Im engeren Sinne ist die Sentimentanalyse eine Aufgabe der Computerlinguistik und besteht in der Analyse von Textdaten hinsichtlich Inhalten wie Gefühls- oder Meinungsäußerungen. Im einfachsten Falle besteht eine Sentimentanalyse in der Bewertung der **Polarität** eines Textdokuments oder eines Teils davon: Spricht der Text insgesamt positiv oder negativ über ein Thema?

Eine einfache Möglichkeit, die Polarität zu bestimmen, besteht in der Verwendung eines Wörterbuchs mit Begriffen, die positiv bzw. negativ konnotiert sind, oder die positives bzw. negatives Sentiment ausdrücken. Der untersuchte Text würde dann mit dem Wörterbuch abgeglichen, z. B. würden im Satz „Ich liebe diesen großartigen Film“ die positiven Begriffe „lieben“ und „großartig“ identifiziert werden, welche positives Sentiment nahelegen.

Zwei solche Wörterbücher, die für eine Sentimentanalyse zur Verfügung stehen, sind Lexicoder 2015 [31, 32] und VADER<sup>4</sup> [33]. Eine weitere Möglichkeit besteht in der Verwendung von Methoden des überwachten Lernens, wie wir im Folgenden demonstrieren wollen.

Wir verwenden den Datensatz „Large Movie Review Dataset v1.0“ [34, 35]. Auf der Data-Science-Plattform Kaggle steht eine im Format vereinfachte Variante zur Verfügung [36]. Der Datensatz enthält insgesamt 50.000 englischsprachige Rezensionen zu Spielfilmen und Fernsehserien, verfasst von Nutzern der Internet Movie Database. Je nachdem, ob der Film dem Nutzer gefallen oder dieser den Film verrissen hat, sind die Rezensionen als „positiv“ oder „negativ“ gekennzeichnet.

Hier zwei besonders prägnante, in hohem Maße polarisierte Beispieldokumente:

---

<sup>4</sup> „VADER“ steht für Valence Aware Dictionary and sEntiment Reasoner.



Abb. 8.6. MNIST-Datensatz für Handschrifterkennung (oben) und CIFAR-10-Datensatz für Objekterkennung (unten; Originalbilder in Farbe)

```

model <- keras_model_sequential() %>%
  layer_conv_2d(filters = 32, kernel_size = c(3,3)
    , input_shape = c(32, 32, 3), padding = 'same') %>%
  layer_activation_leaky_relu() %>%
  layer_batch_normalization(axis = -1) %>%
  layer_conv_2d(filters = 32, kernel_size = c(3,3)
    , padding = 'same') %>%
  layer_activation_leaky_relu() %>%
  layer_batch_normalization(axis = -1) %>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  layer_conv_2d(filters = 64, kernel_size = c(3,3)
    , padding = 'same') %>%
  layer_activation_leaky_relu() %>%
  layer_batch_normalization(axis = -1) %>%
  layer_conv_2d(filters = 64, kernel_size = c(3,3)
    , padding = 'same') %>%
  layer_activation_leaky_relu() %>%
  layer_batch_normalization(axis = -1) %>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  layer_conv_2d(filters = 128, kernel_size = c(3,3)
    , padding = 'same') %>%
  layer_activation_leaky_relu() %>%
  layer_batch_normalization(axis = -1) %>%
  layer_conv_2d(filters = 128, kernel_size = c(3,3)
    , padding = 'same') %>%
  layer_activation_leaky_relu() %>%
  layer_batch_normalization(axis = -1) %>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  layer_flatten() %>%
  layer_dense(units = 512) %>%
  layer_activation_leaky_relu() %>%
  layer_batch_normalization(axis = -1) %>%
  layer_dropout(rate = 0.5) %>%
  layer_dense(units = 1, activation = "sigmoid");

model %>% compile(
  optimizer = 'adam',
  loss = 'binary_crossentropy', metrics = 'accuracy'
);

set.seed(1234)
history <- model %>%
  fit(
    x = X_train, y = Y_train, epochs = 20, batch_size = 20,
    validation_data = unname(list(x = X_val, y = Y_val))
  );

```

**Abb. 8.7.** Beispielcode für die Definition und das Training eines Convolutional Neural Networks

Polarität	Rezension
positiv	If you like original gut wrenching laughter you will like this movie. If you are young or old then you will love this movie, hell even my mom liked it. Great Camp!!!
negativ	Hated it with all my being. Worst movie ever. Mentally scarred. Help me. It was that bad. TRUST ME!!!

**Tabelle 8.3.** Positive/negative Filmrezensionen

Wir wollen das Verfahren der naiven Bayes-Klassifikation einsetzen, um eine Filmrezension automatisiert als „positiv“ oder „negativ“ einzuführen. Als Tokens verwenden wir zunächst einzelne Wörter, Groß- und Kleinschreibung bleiben dabei unberücksichtigt. Trotz der bereits recht fortgeschrittenen Theorie (Multinomial- gegenüber Bernoulli-Modell, Lidstone- bzw. Laplace-Glättung usw.) sollte nicht vergessen werden, dass das Grundprinzip hinter dem Klassifikationsverfahren immer noch ein sehr einleuchtendes ist: Mittels einer **Worthäufigkeitsanalyse** soll herausgefunden werden, welche Wörter auf eine positive oder negative Rezension hindeuten. Dass ein solches Vorgehen praktikabel ist, kann anhand der Häufigkeiten gesehen werden, mit denen gewisse Wörter mit positiven bzw. negativen Rezensionen assoziiert sind. So sind die Häufigkeiten, mit denen folgende Wörter in positiven Rezensionen vorkommen, sehr hoch:

Token	flawless	superbly	perfection	wonderfully	must-see
Likelihood (pos.)	90 %	89 %	89 %	88 %	88 %

**Tabelle 8.4.** Positiv konnotierte Texteinheiten, aufwertende Begriffe

Hingegen deutet ein Vorkommen der folgenden Texteinheiten<sup>5</sup> auf eine negative Rezension hin:

Token	stinker	mst3k	waste	unwatchable	0	unfunny
Likelihood (neg.)	96 %	96 %	94 %	93 %	92 %	92 %

**Tabelle 8.5.** Negativ konnotierte Texteinheiten, abwertende Begriffe

Die so identifizierten Wörter dienen dann als Merkmale, um jede vom Algorithmus „noch nicht gesehene“ Filmbewertung als „positiv“ oder „negativ“ einzuführen.

Der naive Bayes-Klassifikator wurde mithilfe von 40.000 der Rezensionen trainiert. Insgesamt können aus den Texten  $D = 149.653$  Wörter extrahiert werden. Für alle Berechnungen wurden die relativen Häufigkeiten des Auftretens der Wörter über eine Laplace-Glättung regularisiert. Angewandt auf den Test-

<sup>5</sup> Mystery Science Theater 3000, kurz MST3K, ist eine Fernsehshow, in der ausgewählte B-Movies veralbert werden.

datensatz der übrigen 10.000 Rezensionen bewirkt der Einsatz maschinellen Lernens bereits eine deutliche Verbesserung gegenüber den regelbasierten Verfahren:

	Genauigkeit	Trefferquote	$F_1$ -Maß
Lexicoder 2015	70,9 %	74,6 %	72,7 %
VADER	65,0 %	79,0 %	71,3 %
Multinomial	87,2 %	81,1 %	84,1 %
Bernoulli	88,6 %	81,1 %	84,7 %
„Münzwurf“	50,0 %	50,0 %	50,0 %

**Tabelle 8.6.** Güte verschiedener Verfahren der Sentimentanalyse

Eine weitere Verbesserung kann erreicht werden, indem nicht nur Wörter, sondern auch  **$N$ -Gramme** als Merkmale verwendet werden. Dabei handelt es sich um Kombinationen von  $N$  direkt im Text aufeinander folgender Tokens. Es wurden Monogramme ( $N = 1$ ) und Bigramme ( $N = 2$ ) verwendet:

	Genauigkeit	Trefferquote	$F_1$ -Maß
Multinomial, $N$ -Gramme	88,7 %	87,1 %	87,9 %
Bernoulli, $N$ -Gramme	87,0 %	89,5 %	88,2 %

**Tabelle 8.7.** Güte naiver Bayes-Klassifikatoren auf Basis von  $N$ -Grammen

Für die Klassifikation mittels  $N$ -Grammen wurde vor dem Training außerdem eine Merkmalsauswahl vorgenommen: Zum einen wurden  $N$ -Gramme, die in weniger als zehn Dokumenten vorkommen, entfernt. Von diesen wurden wiederum die 10 % an Merkmalen entfernt, welche die kleinste Transinformation mit der Verteilung von positivem/negativem Sentiment aufweisen. Insgesamt wurden  $D = 137.806$   $N$ -Gramme für die Sentimentanalyse herangezogen.

Hier zwei Beispiele einer Fehlklassifikation mit händisch hervorgehobenen Schlüsselelementen:

Polarität	Prognose	Rezension
positiv	negativ	In Black Mask, Jet Li plays a bio-engineered super-killer turned pacifist, who has to fight against other super-killers. <b>Bad plot, bad sfx</b> (60 million dollar budget), <b>but the fighting scenes were excellent!</b> Jet Li is the greatest martial-arts star alive!
negativ	positiv	<b>The first part</b> of Grease with John Travolta and Olivia Newton John <b>is one of the best movie</b> for teens, <b>This one is a very bad copy</b> . The change is only in the sex. In the first one the good one was Sandy, here it's Michael. I prefer to watch the first Grease.

**Tabelle 8.8.** Fehlklassifikationen einer Sentimentanalyse

Im ersten Fall handelt es sich um eine differenzierte Rezension, bei der auch negative Aspekte des Films aufgeführt werden. Bei der zweiten Fehlklassifikation besteht der positive Aspekt in der Bewertung des Vorgängerfilms – der Film, der den eigentlichen Gegenstand der Rezension darstellt, wurde jedoch negativ bewertet. Für diese Fälle wäre es also u. U. von Vorteil gewesen, die Sentimentanalyse nicht auf Dokumentebene, sondern auf einer kleinteiligeren **Aspektebene** durchzuführen [37].

## 8.2 Anwendungsbeispiele für unüberwachtes Lernen

Für die Anwendungsbeispiele in den folgenden Abschnitten verwenden wir wie schon im vorigen Kapitel einen Auszug aus der IMDb, der Internet Movie Database [38]. Der Datensatz enthält insgesamt 85.855 Filme mit Attributen wie Titel, englischsprachige Beschreibung, durchschnittliche Nutzerbewertung, Anzahl der Bewertungen, Genre usw. Den Filmen sind 297.705 Personen verknüpft, die bei der Produktion der Filme als Schauspieler/-innen, Regisseur/-in oder Kameramann/-frau usw. mitgewirkt haben.

### 8.2.1 Textanalyse: Themenmodellierung

Zunächst wollen wir die Filmbeschreibungen untersuchen. Um die Domäne der Texte bereits im Vorfeld etwas einzugrenzen, beschränken wir uns auf zwei Genres und eine zeitliche Schaffensspanne: Familienfilme und Science-Fiction-Filme, jeweils aus den Jahren 1980-2020. Insgesamt enthält der Datensatz jeweils 3141 bzw. 2835 solcher Filme und deren Beschreibungen. Wir würden gern einen schnellen Überblick über typische Themenfelder gewinnen, die die Handlung der Filme ausmachen. Damit wir hierzu nicht alle etwa 6000 Filmbeschreibungen lesen und händisch zusammenfassen wollen, bedienen wir uns Methoden der automatisierten **Textanalyse**.

Damit die unstrukturierten Textdaten einer solchen Analyse zugänglich gemacht werden können, müssen diese zunächst vorverarbeitet werden. Ähnlich wie bei der Sentimentanalyse von Filmrezensionen (Abschn. 8.1.3) nehmen wir hierzu eine Tokenisierung vor. Die Tokens sollen in diesem Fall nicht nur im Hintergrund dem Algorithmus zugänglich gemacht werden, sondern sind Teil des Endergebnisses einer sogenannten Themenkarte. Daher müssen wir bei der Merkmalsauswahl etwas sorgfältiger vorgehen und wollen ausschließlich Textbestandteile extrahieren, die eine gewisse Relevanz für die Textinhalte besitzen. Beispielsweise sollten wenigstens sogenannte **Stoppwörter** von der Analyse ausgeschlossen werden, die in den meisten Textkorpora sehr häufig auftreten wie z. B. die englischen Artikel „the“, „a“ oder häufig vorkommende Präpositionen wie „with“ oder „from“.

Hier werden wir ausschließlich Nominalphrasen verwenden: Eine Nominalphrase kann ein einzelnes Nomen wie „astronaut“ sein, aber auch Wortzusammensetzungen wie „virtual reality“ oder „time machine“. Für die Extraktion von

Nominalphrasen und andere Aufgaben der automatischen Verarbeitung natürlicher Sprache gibt es Softwarebibliotheken wie etwa quanteda [39] oder spaCy [40, 41]. Außerdem beschränken wir uns auf Phrasen, die aus maximal zwei Wörtern zusammengesetzt sind.

Insgesamt wurden aus dem Korpus 105.292 solcher Nominalphrasen extrahiert. Diese sind von variabler Relevanz für die Inhalte. Hier eine kleine Auswahl von extrahierten Phrasen und deren Vorkommen in den jeweiligen Genres in Prozent:

Genre Phrase \	Crime	Family	Horror	Romance	Sci-Fi	Western
Genre	Crime	Family	Horror	Romance	Sci-Fi	Western
undercover cop	65	0	4	0	2	0
fairy tale	2	39	6	16	0	2
occult	7	2	80	5	2	0
young lovers	8	0	8	43	5	0
outer space	2	13	22	4	73	0
cavalry	1	4	0	18	0	70

**Tabelle 8.9.** Filmgenretypische Nominalphrasen

An 100 % fehlende Anteile werden von den übrigen Genres abgedeckt. Obige Phrasen können als „Genre-typisch“ angesehen werden, z. B. handelt es sich bei 73 % aller Filme, deren Beschreibung die Phrase „outer space“ enthält, um Science-Fiction-Filme. Wir wollen uns in der Analyse eher auf solche Genretypischen Texteinheiten konzentrieren und berechnen daher für die spätere Merkmalsauswahl für jede der Phrasen die Transinformation mit der Verteilung über die Genres.

Werden von den 10.522 der in Familienfilmen vorkommenden Phrasen zunächst jene entfernt, die entweder nur in weniger als zehn Texten oder öfter als in jedem vierten Text vorkommen, dann reduziert sich die Anzahl der Tokens bereits auf 674. Schließlich werden von diesen verbleibenden Nominalphrasen jene 60 ausgesucht, die den größten Wert für die vorberechnete Transinformation aufweisen.

Die Merkmale, mit denen wir die Science-Fiction-Filme charakterisieren wollen, werden mit derselben Methode ausgewählt. Schließlich wenden wir den t-SNE-Algorithmus auf die Ähnlichkeitsmatrix zwischen den Phrasen an, die mittels des Überlappungsindex berechnet wird. Dadurch entstehen zwei **Themenkarten**, die in Abb. 8.8 dargestellt sind. Außerdem wurde die Größe der Phrasen mit der Häufigkeit von deren Auftreten skaliert.

## 8.2.2 Netzwerkanalyse: Gemeinschaftsstrukturen

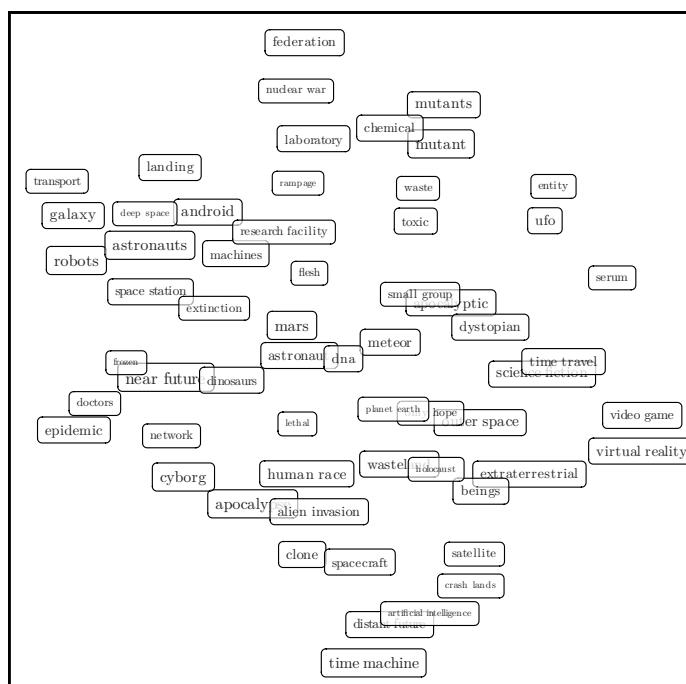
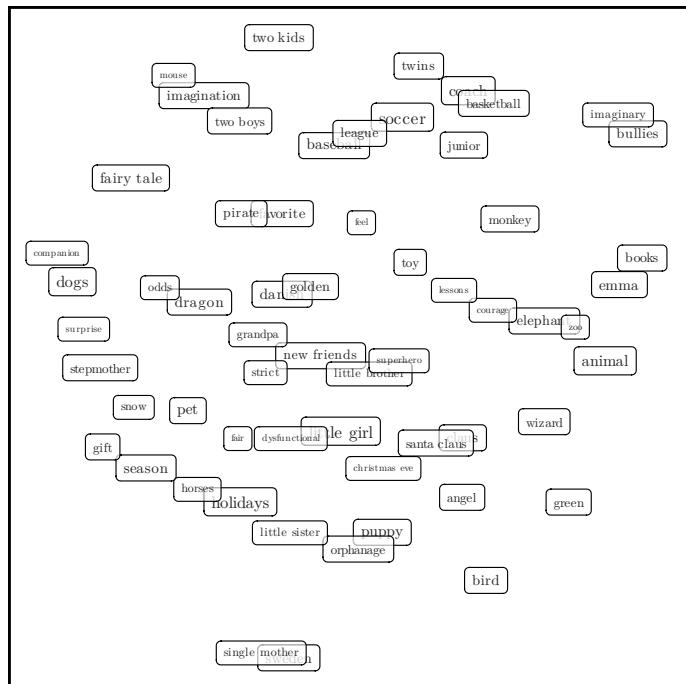
Der IMDb-Datensatz enthält neben der Beschreibung für jeden Film auch eine Liste von Schauspielern/Schauspielerinnen, die an dem Film mitgewirkt haben. Aus diesen Daten lässt sich ein **Kooperationsnetzwerk** konstruieren: Treten

zwei Schauspieler/-innen in wenigstens einem Film gemeinsam vor die Kamera, so stellen wir sie uns durch diese Kooperation verbunden vor. Mathematisch ist ein solches Netzwerk ein ungerichteter Graph: Jeder Knoten repräsentiert einen/eine Darsteller/-in und jede Kooperation eine Kante. Mithin können wir auch ein Knoten-Kanten-Diagramm für einen solchen Graphen aufzeigen, in Abb. 8.9 ist ein solches dargestellt. Für das Beispiel wurden nur Schauspieler/-innen berücksichtigt, die in wenigstens einem Film mitspielten, bei dem Martin Scorsese Regie führte. Obwohl es sich dabei um Personen des öffentlichen Lebens handelt, führt die Analyse auf umfassende personenbezogene Informationen. Daher wurden alle Namen bis auf ein paar wenige Ausnahmen pseudonymisiert, damit der Datenschutz gewährleistet bleibt.

Für die Netzwerkdarstellung wurden die Knoten mit dem *t*-SNE-Verfahren automatisch positioniert, die Visualisierung wurde mithilfe spezialisierter Programmbibliotheken für R erzeugt [42, 43, 44]. Beispiele für Softwareprogramme, die ganz der Netzwerkanalyse und -visualisierung gewidmet sind, sind beispielsweise Gephi [45] und Cytoscape [46].

Obwohl die Grafik nur einen sehr kleinen Ausschnitt aus dem Kooperationsnetzwerk zeigt, ist sie bereits recht unübersichtlich. Gerade in der Nähe stark vernetzter Knoten (sogenannte **Hubs**) ist es nicht leicht, die Kooperationen zwischen den Akteuren alle nachzuvollziehen.

Eine Möglichkeit, die Komplexität zu reduzieren, besteht in der Identifikation von Clustern im Netzwerk, den **Gemeinschaften**: In Abb. 8.10 ist das Baumdiagramm einer hierarchischen Clusteranalyse im Average-Linkage-Verfahren dargestellt. Dabei wurde die Jaccard-Ähnlichkeit verwendet: Die Anzahl der Filme, in denen beide Darsteller/-innen mitwirkten, geteilt durch die Anzahl der Filme, in denen wenigstens eine/-r von ihnen mitspielte.



**Abb. 8.8.** Themenkarten für Familien- bzw. Science-Fiction-Filme (1980-2020)

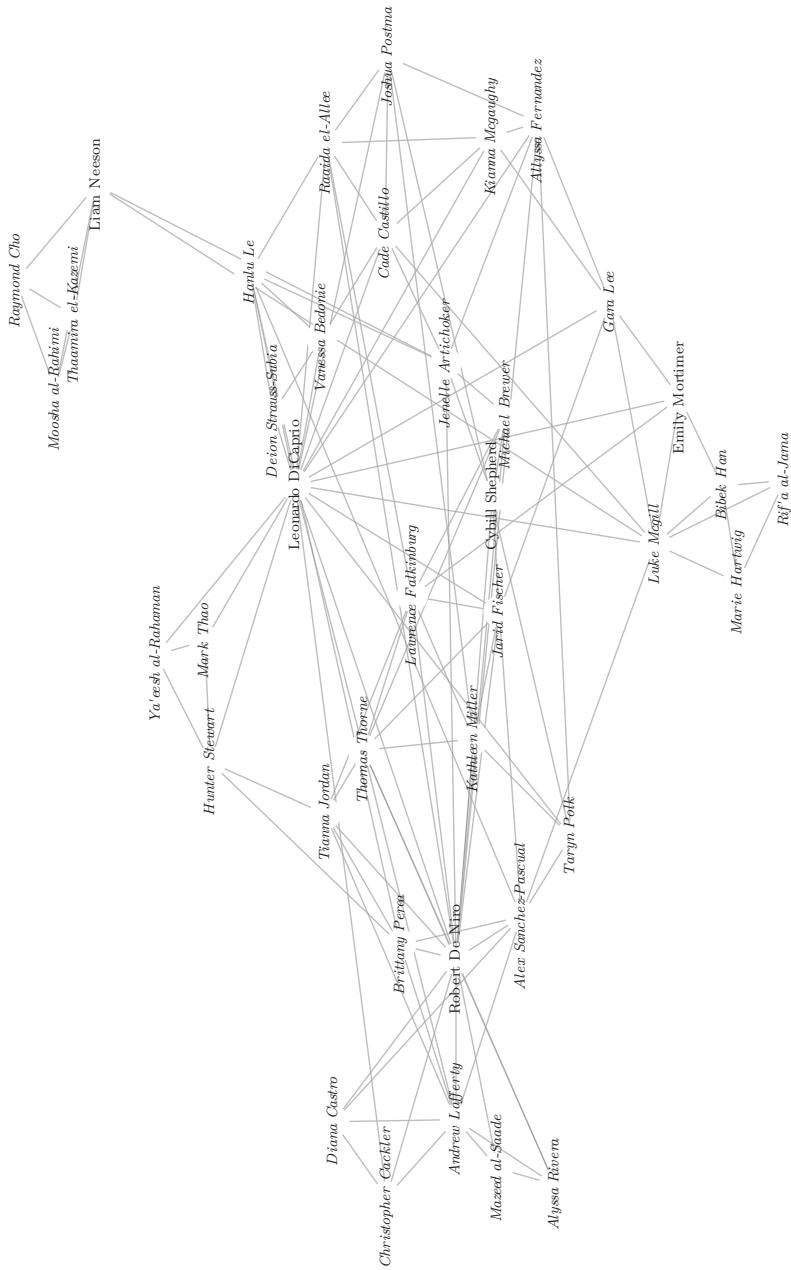
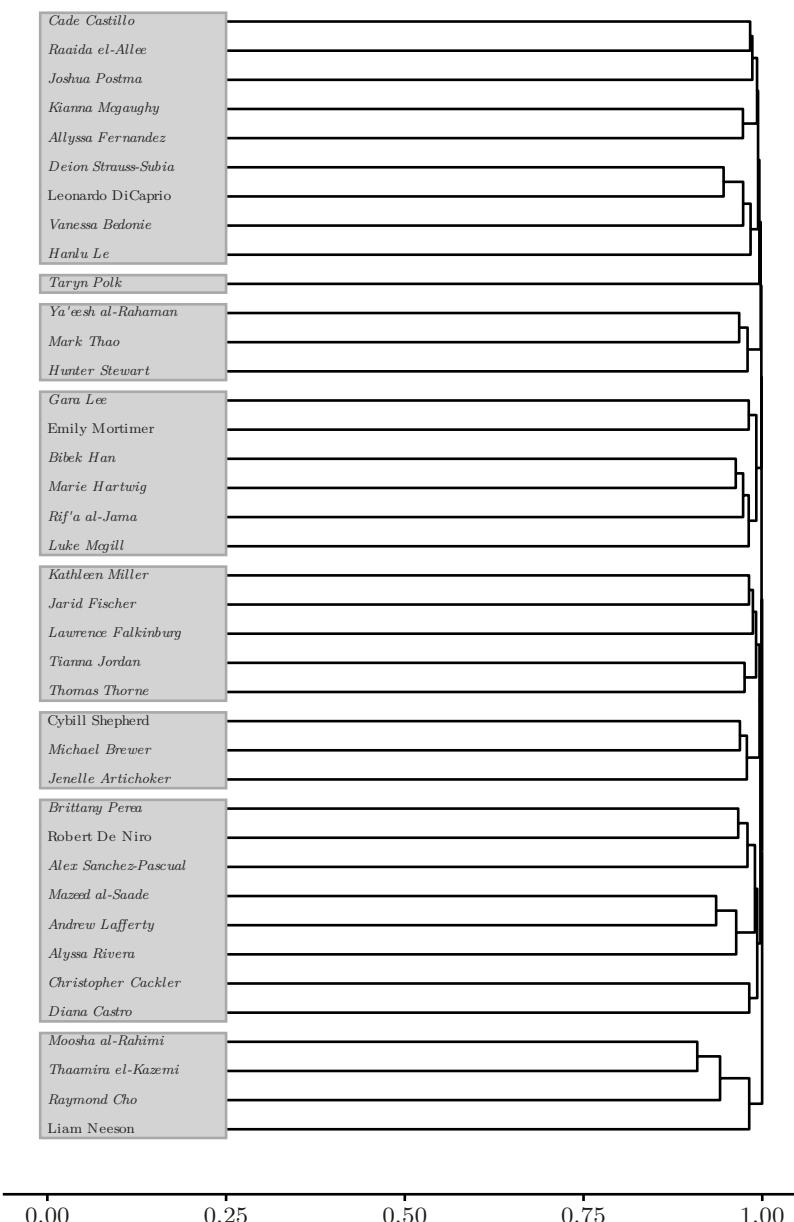


Abb. 8.9. Kooperationsnetzwerk von Schauspielern/Schauspielerinnen (teilweise pseudonymisiert)



**Abb. 8.10.** Hierarchische Clusteranalyse eines Kooperationsnetzwerks von Schauspielern/Schauspielerinnen (teilweise pseudonymisiert)

## Quellen

- [1] Sampo Kuutti u. a. „A Survey of Deep Learning Applications to Autonomous Vehicle Control“. In: *IEEE Transactions on Intelligent Transportation Systems* (2020), S. 1–22. DOI: [10.1109/tits.2019.2962338](https://doi.org/10.1109/tits.2019.2962338). arXiv:[1912.10773](https://arxiv.org/abs/1912.10773).
- [2] S. Kevin Zhou, Hayit Greenspan und Dinggang Shen, Hrsg. *Deep Learning for Medical Image Analysis*. Academic Press, Jan. 2017. ISBN: 978-0128104088.
- [3] Edward H. Lee u. a. „Deep COVID DeteCT: an international experience on COVID-19 lung detection and prognosis using chest CT“. In: *npj Digital Medicine* 4.1 (Jan. 2021). DOI: [10.1038/s41746-020-00369-1](https://doi.org/10.1038/s41746-020-00369-1).
- [4] Nikolas Lessmann u. a. „Automated Assessment of COVID-19 Reporting and Data System and Chest CT Severity Scores in Patients Suspected of Having COVID-19 Using Artificial Intelligence“. In: *Radiology* 298.1 (Jan. 2021), E18–E28. DOI: [10.1148/radiol.2020202439](https://doi.org/10.1148/radiol.2020202439).
- [5] Dalya Baron. *Machine Learning in Astronomy: a practical overview*. Apr. 2019. arXiv:[1904.07248v1](https://arxiv.org/abs/1904.07248v1).
- [6] Sander Dieleman, Kyle W. Willett und Joni Dambre. „Rotation-invariant convolutional neural networks for galaxy morphology prediction“. In: *Monthly Notices of the Royal Astronomical Society* 450.2 (Apr. 2015), S. 1441–1459. DOI: [10.1093/mnras/stv632](https://doi.org/10.1093/mnras/stv632). arXiv:[1503.07077](https://arxiv.org/abs/1503.07077).
- [7] Helena Domínguez Sánchez u. a. „Improving galaxy morphologies for SDSS with Deep Learning“. In: *Monthly Notices of the Royal Astronomical Society* 476.3 (Feb. 2018), S. 3661–3676. DOI: [10.1093/mnras/sty338](https://doi.org/10.1093/mnras/sty338). arXiv:[1711.05744](https://arxiv.org/abs/1711.05744).
- [8] Carlos Alberto Gomez Gonzalez, Olivier Absil und Marc van Droogenbroeck. „Supervised detection of exoplanets in high-contrast imaging sequences“. In: *Astronomy & Astrophysics* 613 (Mai 2018), A71. DOI: [10.1051/0004-6361/201731961](https://doi.org/10.1051/0004-6361/201731961). arXiv:[1712.02841](https://arxiv.org/abs/1712.02841).
- [9] Faustine Cantalloube u. a. „Exoplanet imaging data challenge: benchmarking the various image processing methods for exoplanet detection“. In: *Adaptive Optics Systems VII*. Hrsg. von Dirk Schmidt, Laura Schreiber und Elise Vernet. Bd. 11448. International Society for Optics and Photonics. SPIE, Dez. 2020, S. 1027–1062. DOI: [10.1117/12.2574803](https://doi.org/10.1117/12.2574803).
- [10] Fabrizio Carcillo u. a. „Combining unsupervised and supervised learning in credit card fraud detection“. In: *Information Sciences* (Mai 2019). DOI: [10.1016/j.ins.2019.05.042](https://doi.org/10.1016/j.ins.2019.05.042).
- [11] Fan Zhang u. a. *MediaPipe Hands: On-device Real-time Hand Tracking*. Juni 2020. arXiv:[2006.10214v1](https://arxiv.org/abs/2006.10214v1).
- [12] Dong Yu und Li Deng. *Automatic Speech Recognition*. Springer, London, 2015. DOI: [10.1007/978-1-4471-5779-3](https://doi.org/10.1007/978-1-4471-5779-3).
- [13] Bernhard C. Witt. *Datenschutz kompakt und verständlich*. Wiesbaden: Vieweg+Teubner, 2010. DOI: [10.1007/978-3-8348-9653-7](https://doi.org/10.1007/978-3-8348-9653-7).

- [14] Matthias Plaue. „Rise of the Mindless Machines“. In: *towards data science* (Nov. 2018). URL: <https://towardsdatascience.com/rise-of-the-mindless-machines-c0e578061e65>.
- [15] Anna Jobin, Marcello Ienca und Effy Vayena. „The global landscape of AI ethics guidelines“. In: *Nature Machine Intelligence* 1.9 (Sep. 2019), S. 389–399. DOI: [10.1038/s42256-019-0088-2](https://doi.org/10.1038/s42256-019-0088-2). arXiv:1906.11668.
- [16] Yann LeCun, Corinna Cortes und Christopher J. C. Burges. *The MNIST database of handwritten digits*. 2010. URL: <http://yann.lecun.com/exdb/mnist/>.
- [17] Jiang Junfeng. *readmnist: Read MNIST Dataset*. R-Paket, Version 1.0.6. 2018. URL: <https://CRAN.R-project.org/package=readmnist>.
- [18] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. Techn. Ber. 2009.
- [19] Moritz Hambach. *Image Augmentation in Keras (CIFAR-10)*. Jan. 2018. URL: <https://github.com/moritzhambach/Image-Augmentation-in-Keras-CIFAR-10->.
- [20] François Collet u. a. *Keras*. URL: <https://keras.io>.
- [21] J. J. Allaire und François Chollet. *keras: R Interface to 'Keras'*. R-Paket, Version 2.3.0.0. 2020. URL: <https://CRAN.R-project.org/package=keras>.
- [22] Martín Abadi u. a. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. URL: <https://www.tensorflow.org/>.
- [23] J. J. Allaire und Yuan Tang. *tensorflow: R Interface to 'TensorFlow'*. R-Paket, Version 2.2.0. 2020. URL: <https://CRAN.R-project.org/package=tensorflow>.
- [24] Diederik P. Kingma und Jimmy Ba. „Adam: A Method for Stochastic Optimization“. In: *3rd International Conference on Learning Representations, San Diego, USA*. Hrsg. von Yoshua Bengio und Yann LeCun. Mai 2015. arXiv:1412.6980.
- [25] Sergey Ioffe und Christian Szegedy. „Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift“. In: *32nd International Conference on Machine Learning, Lille, France*. Hrsg. von Francis Bach und David Blei. Bd. 37. Proceedings of Machine Learning Research. PMLR, Juli 2015, S. 448–456. arXiv:1502.03167.
- [26] Papers with Code Community. *CIFAR-10 Benchmark (Image Classification)*. Hrsg. von Robert Stojnic u. a. Aufgerufen am 28. Dez. 2020. URL: <https://paperswithcode.com/sota/image-classification-on-cifar-10>.
- [27] Pierre Foret u. a. *Sharpness-Aware Minimization for Efficiently Improving Generalization*. Okt. 2020. arXiv:2010.01412v1.
- [28] Alexey Dosovitskiy u. a. *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale*. Okt. 2020. arXiv:2010.11929v1.
- [29] Alexander Kolesnikov u. a. „Big Transfer (BiT): General Visual Representation Learning“. In: *Computer Vision – ECCV 2020*. Bd. 12350. Lecture Notes in Computer Science. Springer, Cham, 2020, S. 491–507. DOI: [10.1007/978-3-030-58558-7\\_29](https://doi.org/10.1007/978-3-030-58558-7_29). arXiv:1912.11370.

- [30] Andrej Karpathy. *Lessons learned from manually classifying CIFAR-10*. Apr. 2011. URL: <http://karpathy.github.io/2011/04/27/manually-classifying-cifar10/>.
- [31] Lori Young und Stuart Soroka. *Lexicoder Sentiment Dictionary*. 2012.
- [32] Lori Young und Stuart Soroka. „Affective News: The Automated Coding of Sentiment in Political Texts“. In: *Political Communication* 29.2 (2012), S. 205–231. DOI: [10.1080/10584609.2012.671234](https://doi.org/10.1080/10584609.2012.671234).
- [33] Clayton J. Hutto und Eric Gilbert. „VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.“ In: *ICWSM*. Hrsg. von Eytan Adar u. a. AAAI Press, 2014.
- [34] Andrew L. Maas u. a. „Learning Word Vectors for Sentiment Analysis“. In: *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA*. Association for Computational Linguistics, Juni 2011, S. 142–150.
- [35] Andrew L. Maas. *Large Movie Review Dataset v1.0*. Aufgerufen am 15. Nov. 2020. URL: <http://ai.stanford.edu/~amaas/data/sentiment/>.
- [36] N. Lakshmi pathi. *IMDb dataset of 50k movie reviews. Large Movie Review Dataset*. Aufgerufen am 15. Nov. 2020. URL: <https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>.
- [37] Ambreen Nazir u. a. „Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey“. In: *IEEE Transactions on Affective Computing* (2020). DOI: [10.1109/taffc.2020.2970399](https://doi.org/10.1109/taffc.2020.2970399).
- [38] Stefano Leone. *IMDb movies extensive dataset. 81k+ movies and 175k+ cast members scraped from IMDb*. Aufgerufen am 17. Nov. 2020. URL: <https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>.
- [39] Kenneth Benoit u. a. „quanteda: An R package for the quantitative analysis of textual data“. In: *Journal of Open Source Software* 3.30 (2018), S. 774. DOI: [10.21105/joss.00774](https://doi.org/10.21105/joss.00774). URL: <https://quanteda.io>.
- [40] Matthew Honnibal und Ines Montani. *spaCy. Industrial-Strength Natural Language Processing*. Aufgerufen am 09. Dez. 2020. URL: <https://spacy.io/>.
- [41] Kenneth Benoit und Akitaka Matsuo. *spacyr: Wrapper to the 'spaCy' NLP Library*. R-Paket, Version 1.2.1. 2020. URL: <https://CRAN.R-project.org/package=spacyr>.
- [42] Barret Schloerke u. a. *GGally: Extension to 'ggplot2'*. R-Paket, Version 2.0.0. 2020. URL: <https://CRAN.R-project.org/package=GGally>.
- [43] Carter T. Butts. *network: Classes for Relational Data*. R-Paket, Version 1.16.1. The Statnet Project (<http://www.statnet.org>). 2020. URL: <https://CRAN.R-project.org/package=network>.
- [44] Carter T. Butts. „network: a Package for Managing Relational Data in R“. In: *Journal of Statistical Software* 24.2 (2008). URL: <https://www.jstatsoft.org/v24/i02/paper>.
- [45] Mathieu Bastian, Sébastien Heymann und Mathieu Jacomy. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. URL: <https://gephi.org/>.

- [46] Paul Shannon u.a. „Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks“. In: *Genome Research* 13.11 (Nov. 2003), S. 2498–2504. DOI: [10.1101/gr.1239303](https://doi.org/10.1101/gr.1239303). URL: <https://cytoscape.org/>.



---

## Ergänzende Literatur

Die zum Verständnis des vorliegenden Buches notwendigen mathematischen Grundlagen werden beispielsweise in den ersten zwei Bänden der Reihe „Mathematik für das Bachelorstudium“, welche von Mike Scherfner und mir verfasst sind, vermittelt [1, 2]. Die dortige Notation von Formeln stimmt mit der hier verwendeten in weiten Teilen überein. Natürlich kann auch auf bewährte Klassiker zurückgegriffen werden, etwa auf „den Fischer“ [3] für lineare Algebra oder „den Forster“ für Analysis [4, 5].

Gerade Quereinsteigern aus anderen Fachrichtungen als der Informatik empfiehlt sich ein Blick in das „Taschenbuch der Informatik“, um sich mit wesentlicher Terminologie vertraut zu machen [6].

Für weiterführende Literatur verweise ich gerne auf die folgenden Texte.

**Datenorganisation.** Einen allgemeinen Überblick über den Einsatz von Informationstechnik in der Wirtschaft, insbesondere das Daten- und Wissensmanagement, liefert das – inzwischen in 17. Auflage erschienene – Buch von Mertens u. a. [7]. Auf die Themen Datenintegration, Datenqualität und Datenbereinigung geht der Band von Leser und Naumann umfassend ein [8]. Mehr über Entity-Relationship-Modelle erfährt die Leserin oder der Leser in der kompakten Darstellung von Andreas Gadatsch [9]. Über graphenbasierte Ansätze im Allgemeinen und die Verwendung der Graphendatenbank neo4j im Besonderen klärt ein Buch von Ian Robinson u. a. auf [10]. Wer sich eingehender mit der mathematischen Theorie der Graphen auseinandersetzen möchte, dem sei das klassische Buch von Harary ans Herz gelegt [11].

**Wahrscheinlichkeitstheorie und Statistik.** Einen kompakten und anwendungsorientierten Überblick verschafft das Buch „Basiswissen Statistik“ von Ansgar Steland [12]. Mit dem Band „Maß und Wahrscheinlichkeit“ stellt Klaus Schmidt einen klaren Überblick über die mathematischen Grundlagen der Wahrscheinlichkeitstheorie bereit [13]. Ebenfalls mathematisch geneigten Leser/-innen sei darüber hinaus der Klassiker von Fisz [14] und der etwas neuere Band von Casella und Berger empfohlen [15]. Roman Vershynin zeigt in sei-

nem Buch „High-Dimensional Probability“ wesentliche mathematische Konzepte und Resultate auf, die der Analyse hochdimensionaler Daten zugrundeliegen [16]. Ein Artikel von Gunnar Carlsson beleuchtet die topologische Perspektive [17]. Die „Encyclopedia of Distances“ hat sich schließlich ganz dem Thema der Abstandsmaße verschrieben [18].

**Maschinelles Lernen.** Empfohlene Klassiker über Verfahren des maschinellen Lernens und der Mustererkennung sind die Werke von Duda, Hart und Stork bzw. Hastie, Tibshirani und Friedman [19, 20]. Auch der Band von Theodoridis und Koutroumbas teilt mit den Leserinnen und Lesern einen reichen Erfahrungsschatz [21]. Der Band von Ian Goodfellow u. a. enthält eine umfassende Darstellung von künstlichen neuronalen Netzwerken [22]. François Chollet, Initiator der Deep-Learning-Programmbibliothek Keras, ist Autor von Praxisbüchern zum Thema [23, 24]. Eine kompakte und klare Darstellung der Kernaspekte der Theorie des statistischen Lernens liefert ein Übersichtsartikel von Ulrike von Luxburg und Bernhard Schölkopf [25]. Ein empfehlenswertes Buch zum Thema ist der Band von Mehryar Mohri u. a. [26].

**Weitere Themen.** Ein Standardwerk über digitale Bildverarbeitung ist das Lehrbuch von Bernd Jähne [27]. Eine moderne Einführung in die Computerlinguistik bietet das Buch von Jacob Eisenstein [28]. Wer sich im Detail zum Thema Sentimentanalyse informieren möchte, kann den Band von Bing Liu zu Rate ziehen [29].

Weiterhin sollte noch das Buch „Foundations of Data Science“ von Avrim Blum, John Hopcroft und Ravindran Kannan nicht unerwähnt bleiben, das einen Einblick in verschiedenste Themen liefert – ein hervorragendes Buch für den Nachtisch interessierter Leserinnen und Leser [30].

Schließlich möchte ich noch auf den YouTube-Kanal 3Blue1Brown von Grant Sanderson aufmerksam machen, auf dem – unter anderem – eine unterhaltsame und aufschlussreiche Einführung in neuronale Netzwerke zu finden ist [31].

## Quellen

- [1] Matthias Plaue und Mike Scherfner. *Mathematik für das Bachelorstudium I.* 2. Aufl. Springer Spektrum, Berlin, Heidelberg, 2019. ISBN: 978-3-662-58351-7. DOI: [10.1007/978-3-662-58352-4](https://doi.org/10.1007/978-3-662-58352-4).
- [2] Matthias Plaue und Mike Scherfner. *Mathematik für das Bachelorstudium II.* Springer Spektrum, Berlin, Heidelberg, 2019. ISBN: 978-3-8274-2068-8. DOI: [10.1007/978-3-8274-2557-7](https://doi.org/10.1007/978-3-8274-2557-7).
- [3] Gerd Fischer und Boris Springborn. *Lineare Algebra. Eine Einführung für Studienanfänger.* 19. Aufl. Springer Spektrum, Berlin, Heidelberg, 2020. DOI: [10.1007/978-3-662-61645-1](https://doi.org/10.1007/978-3-662-61645-1).
- [4] Otto Forster. *Analysis 1.* 12. Aufl. Springer Spektrum, Wiesbaden, 2016. DOI: [10.1007/978-3-658-11545-6](https://doi.org/10.1007/978-3-658-11545-6).
- [5] Otto Forster. *Analysis 2.* 11. Aufl. Springer Spektrum, Wiesbaden, 2017. DOI: [10.1007/978-3-658-19411-6](https://doi.org/10.1007/978-3-658-19411-6).
- [6] Georg Disterer u. a. *Taschenbuch der Informatik.* Hrsg. von Uwe Schneider. 7. Aufl. München: Carl Hanser, 2017. ISBN: 978-3-446-42638-2.
- [7] Peter Mertens u. a. *Grundzüge der Wirtschaftsinformatik.* 17. Aufl. Springer, Berlin, Heidelberg, 2017. DOI: [10.1007/978-3-662-53362-8](https://doi.org/10.1007/978-3-662-53362-8).
- [8] Ulf Leser und Felix Naumann. *Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen.* 1. Aufl. OCLC: 180130256. Heidelberg: dpunkt.verlag, 2007. ISBN: 978-3-898-64400-6.
- [9] Andreas Gadatsch. *Datenmodellierung.* Springer Vieweg, Wiesbaden, 2019. DOI: [10.1007/978-3-658-25730-9](https://doi.org/10.1007/978-3-658-25730-9).
- [10] Ian Robinson, Jim Webber und Emil Eifrem. *Graph Databases.* 2. Aufl. Sebastopol, USA: O'Reilly, 2015.
- [11] Frank Harary. *Graph Theory.* Reading, USA: Addison Wesley, 1969.
- [12] Ansgar Steland. *Basiswissen Statistik.* Springer Spektrum, Berlin, Heidelberg, 2016. DOI: [10.1007/978-3-662-49948-1](https://doi.org/10.1007/978-3-662-49948-1).
- [13] Klaus D. Schmidt. *Maß und Wahrscheinlichkeit.* 2. Aufl. Springer, Berlin, Heidelberg, 2011. ISBN: 978-3-642-21026-6. DOI: [10.1007/978-3-642-21026-6](https://doi.org/10.1007/978-3-642-21026-6).
- [14] Marek Fisz. *Wahrscheinlichkeitsrechnung und Mathematische Statistik.* Berlin: VEB Deutscher Verlag der Wissenschaften, 1962.
- [15] George Casella und Roger L. Berger. *Statistical Inference.* Pacific Grove, USA: Duxbury Thomson Learning, 2001.
- [16] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge University Press, Sep. 2018. DOI: [10.1017/9781108231596](https://doi.org/10.1017/9781108231596).
- [17] Gunnar Carlsson. „Topological pattern recognition for point cloud data“. In: *Acta Numerica* 23 (Mai 2014), S. 289–368. DOI: [10.1017/s0962492914000051](https://doi.org/10.1017/s0962492914000051).
- [18] Michel Marie Deza und Elena Deza. *Encyclopedia of Distances.* Springer, Berlin, Heidelberg, 2009. DOI: [10.1007/978-3-642-00234-2](https://doi.org/10.1007/978-3-642-00234-2).
- [19] Richard O. Duda, Peter E. Hart und David G. Stork. *Pattern Classification.* 2. Aufl. Wiley, 2000. ISBN: 978-0-471-05669-0.

- [20] Trevor Hastie, Robert Tibshirani und Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. Aufl. Springer, New York, 2009. ISBN: 978-0-387-84857-0. DOI: [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7).
- [21] Sergios Theodoridis und Konstantinos Koutroumbas. *Pattern Recognition*. 4. Aufl. Academic Press, 2008. ISBN: 1597492728.
- [22] Ian Goodfellow, Yoshua Bengio und Aaron Courville. *Deep Learning*. MIT Press, Nov. 2016. ISBN: 978-0-262-03561-3. URL: <http://www.deeplearningbook.org/>.
- [23] François Chollet. *Deep Learning mit Python und Keras*. Frechen: mitp-Verlag, Mai 2018. ISBN: 978-3-958-45838-3.
- [24] François Chollet und J. J. Allaire. *Deep Learning mit R und Keras*. Frechen: mitp-Verlag, Okt. 2018. ISBN: 978-3-958-45893-2.
- [25] Ulrike von Luxburg und Bernhard Schölkopf. „Statistical Learning Theory: Models, Concepts, and Results“. In: *Handbook of the History of Logic*. Bd. 10. Amsterdam, Niederlande: Elsevier North Holland, Mai 2011, S. 651–706. DOI: [10.1016/b978-0-444-52936-7.50016-1](https://doi.org/10.1016/b978-0-444-52936-7.50016-1). arXiv:0810.4752.
- [26] Mehryar Mohri, Afshin Rostamizadeh und Ameet Talwalkar. *Foundations of Machine Learning*. 2. Aufl. MIT Press, 2018. ISBN: 978-0-262-03940-6.
- [27] Bernd Jähne. *Digitale Bildverarbeitung und Bildgewinnung*. 8. Aufl. Springer Vieweg, Wiesbaden, 2021. ISBN: 978-3-662-59510-7.
- [28] Jacob Eisenstein. *Introduction to natural language processing*. MIT Press, 2019. ISBN: 978-0-262-04284-0.
- [29] Bing Liu. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge University Press, 2015. ISBN: 978-1-107-01789-4.
- [30] Avrim Blum, John Hopcroft und Ravi Kannan. *Foundations of Data Science*. Cambridge University Press, Jan. 2020. DOI: [10.1017/9781108755528](https://doi.org/10.1017/9781108755528).
- [31] Grant Sanderson. *3Blue1Brown, Season 3: Neural Networks*. Aug. 2018. URL: [https://youtube.com/playlist?list=PLZHQBObOWTQDNU6R1\\_67000Dx\\_ZCJB-3pi](https://youtube.com/playlist?list=PLZHQBObOWTQDNU6R1_67000Dx_ZCJB-3pi).

---

# Sachverzeichnis

- 2D-Faltung, 245
- A**
- Adam, 287
  - Aggregation, 17
  - Aktivierungsfunktion, 233
  - Algorithmus, 189
  - ALLBUS-Studie, 44, 137, 139, 147, 161
  - Alternativhypothese, 137
  - Anscombe-Quartett, 60
  - A-posteriori-Verteilung, 151
  - A-posteriori-Wahrscheinlichkeit, 79
  - Approximationsfehler, 199
  - A-priori-Verteilung, 151
    - nichtinformative, 152
    - uneigentliche, 151
  - A-priori-Wahrscheinlichkeit, 79
  - Armijo-Bedingung, 207
  - Assoziationsparameter, 56
  - Attribut, 13
  - Ausgleichsgerade, 157
  - Ausreißer, 49, 123
  - Autoencoder, 264
  - Average-Linkage-Verfahren, 277
- B**
- Backpropagation, 242
  - Backtracking-Liniensuche, 207
  - Bag-of-Words-Modell, 230
  - Barzilai-Borwein-Methode, 205
  - Batch-Standardisierung, 287
- Baum, 21
- Baumdiagramm, 277
- Bayes-Fehler, 199
- Bayes-Klassifikator, 227
  - naiver, 229
- Beobachtung, 39
- Bernoulli-Experiment, 117
- Bernoulli-Verteilung, 117
- Bessel-Korrektur, 131
- Beziehungstyp, 13
- BFGS-Verfahren, 207
- Big Data, 11
- Bildverarbeitung, 246
- Binomialkoeffizient, 118
- Binomialverteilung, 118
- Body-Mass-Index, 40, 42
- Broca-Index, 43
- Bubblesort, 189
- Business Intelligence, 1
- C**
- Cauchy-Lorentz-Verteilung, 122
  - CDC-Studie, 37, 41, 42, 58, 128, 136, 142, 148, 153, 156, 228
  - Chen-Notation, 13
  - Chi-Quadrat-Verteilung, 110
    - mit einem Freiheitsgrad, 91
  - Choroplethenkarte, 44
  - CIFAR-10, 286
  - Clusteranalyse, 259, 272
    - agglomerative hierarchische, 276

- divisive hierarchische, 277
- Codeplan, 23
- Complete-Case-Analyse, 26
- Complete-Linkage-Verfahren, 277
- Convolutional Neural Network, 247, 287
- Convolution-Schicht, 248
- Cromwell'sche Regel, 81
  
- D**
- Damerau-Levenshtein-Distanz, 30
- Data Engineering, 12
- Data-Mining, 3
- Data-Profilierung, 23
- Daten, 1
  - empirische, 78
  - hochdimensionale, 190
- Datenanalyse, 3
  - topologische, 258
- Datenaugmentation, 28, 288
- Datenbestand, 11
- Datendeduplikation, 29
- Datendokumentation, 12, 23
- Datenelement, 14
- Datenfeld, 14
- Datenherkunft, 24
- Datenimputation, 26
- Datenintegration, 11
- Datenmatrix, 166
  - erweiterte, 167
  - mittelwertzentrierte, 166
  - standardisierte, 166
- Datenmodell
  - graphbasiertes, 19
  - hierarchisches, 21
  - konzeptionelles, 12
  - logisches, 14
  - physisches, 12
  - relationales, 15
- Datennormierung, 26
- Datenpunkt, 39
- Datenqualität, 23
- Datenquelle, 11
- Datensatz, 14
- Datenschutz, 24, 283
- Datentupel, 14
  
- Datenvalidierung, 25
- Datenversionierung, 24
- Datenvisualisierung, 2
- Datenvorverarbeitung, 24
- Deep Learning, 233
- Dichtefunktion, 86
  - bedingte, 94
  - empirische, 145
  - gemeinsame, 92
  - multivariate, 181
- differenzielle Entropie, 101
- Dimensionalität, 166
  - intrinsische, 257
- Dimensionsreduktion, 257
- Dirac-Maß, 72
- Dropout, 244
  
- E**
- Editierabstand, 30
- Effektstärke
  - nach Cohen, 143
- Effizienz, 201
- Einbettung, 257
- Einflussgröße, 157
- Elementarereignis, 71
- Entität, 12, 39
- Entitätstyp, 12
- Entity-Relationship-Diagramm, 13
- Entity-Relationship-Modell, 12
- Entity-Resolution, 29
- Entscheidungsgrenze, 149, 193
- Entscheidungsregel, 192
  - Bayes'sche, 227
  - verallgemeinerbare, 195
- Epoche, 206
- Ereignis, 71
  - komplementäres, 73
- Ereignismodell
  - Bernoulli'sches, 231
  - multinomiales, 230
- Ereignisse
  - unabhängige, 76
- Ergebnisraum, 70
- Erwartungswert, 97
  - empirischer, 48, 128
- Erwartungswertvektor, 179

euklidische Norm, 168  
Explainable AI, 287

**F**

Fakultät, 110, 118  
Faltung  
  von Bildern, 245  
  von Funktionen, 107  
fast sicher, 127  
Feature, 38  
Feature-Engineering, 190  
Feedforward-Netzwerk, 233  
Fehlerrückführung, 242  
Fellegi-Sunter-Modell, 33  
Fermi-Funktion, 234  
Flattening, 284  
Fluch der Dimensionalität, 259  
Fréchet'scher Mittelwert, 177  
Fuzzy-Menge, 272  
Fuzzy-Suche, 29

**G**

Gauß-Glocke, 123  
Gauß-Prozess, 215  
Gauß-Prozess-Regression, 217  
Gauß'sche Glockenkurve, 100, 154  
Gauß'sches Fehlerintegral, 88, 102  
Gauß'sches Mischmodell, 272  
Gauß-Test, 139  
Gauß-Verteilung, 121  
Gemeinschaftserkennung, 295  
Genauigkeit, 202  
Geometrische Verteilung, 119  
gepaarte Stichproben, 57  
Gesetz der großen Zahlen  
  Bernoulli'sches, 128  
  Tschebyscheff'sches, 130  
Gesetze der großen Zahlen, 123  
Glaubwürdigkeitsintervall, 153  
Gleichrichterfunktion, 234  
  durchlässige, 235  
Gleichverteilung, 56, 72  
  einer diskreten Zufallsvariablen,  
    117  
  einer stetigen Zufallsvariablen,  
    120  
globale Erwärmung, 158, 210, 218

Gradientenverfahren, 204  
  einfaches, 205  
  stochastisches, 206

**Graph**

  azyklischer, 20  
  gerichteter, 17  
  Multi-, 18  
  ungerichteter, 18

Grundgesamtheit, 39

**H**

Hamming-Abstand, 171  
Häufigkeit  
  absolute, 55  
  gemeinsame, 62  
  relative, 54  
Häufigkeitsverteilung, 37  
  multimodale, 45, 54  
  schiefe, 49  
  symmetrische, 49  
  unimodale, 45  
Hauptachse, 262  
Hauptachsentransformation, 262  
Hauptkomponente, 262  
Hauptkomponentenanalyse, 262  
Hauptkoordinate, 262  
Hauptrichtung, 262  
Heatmap, 44  
Heteroskedastie, 215  
hierarchische Clusteranalyse, 276  
Histogramm, 41, 143  
Homoskedastie, 217  
Hub, 295  
Hyperparameter, 151, 192, 201  
Hypothese, 78, 137, 192  
Hypothesenraum, 192

**I**

ImageNet, 249  
IMDb, 288, 293  
Indifferenzprinzip, 72, 117  
Indikatorfunktion, 105  
Inferenzstatistik, 115  
Information Retrieval, 203  
Informationsobjekt, 12  
instanzbasierte Verfahren, 222, 224  
Intervallschätzer, 135

irreduzibler Fehler, 199

Irrtumswahrscheinlichkeit, 136

## J

Jaccard-Abstand, 173

Jaccard-Koeffizient, 64, 171

Jaro-Ähnlichkeit, 32

Jaro-Winkler-Ähnlichkeit, 32

Join, 16

## K

Kaggle, 288

Kantendetektion, 246

Kantenzug, 20

Karhunen-Loëve-Transformation,  
262

Keras, 287

Kerndichteschätzung, 154

Kernglättung, 155

Klassenbreite, 41

Klasseneinteilung, 41

Klassifikation, 191

binäre, 192

Klassifikator, 192

linearer, 193

*K*-Means-Verfahren, 274

mit Kern, 275

*K*-Medoids-Verfahren, 274

*K*-nächste-Nachbarn-Klassifikator,  
224

Knoten-Kanten-Diagramm, 18

Kohorte, 39

Kolmogoroff'sche Axiome, 71

Konfidenzintervall, 133

Kontingenztafel, 62

Konvergenz in Verteilung, 132

Kooperationsnetzwerk, 294

Korrektklassifikationsrate, 202

Korrelation, 101

Korrelationskoeffizient

nach Bravais-Pearson, 57, 160

nach Kendall, 60

nach Spearman, 60

Kosinusähnlichkeit, 170

Kostenmatrix, 194

Kovarianz, 101

empirische, 57

Kovarianzellipse, 179

Kovarianzmatrix, 179

regularisierte, 285

Kredibilitätsintervall, 153

Kreisdiagramm, 44

Kreuzentropie, 236, 270

Kreuzkorrelation, 245

Kreuzvalidierung, 201, 285

kritischer Wert, 139

künstliche Intelligenz, 191

## L

Label, 192

Labelglättung, 236

Lageparameter, 45

Laplace-Glättung, 231

Laplace-Operator, 246

Laplace'sche Formel, 72

Lebesgue-Zerlegung, 83

Lemma von Johnson und

Lindenstrauss, 256

Levenshtein-Distanz, 30

normierte, 31

Lidstone-Glättung, 231

Likelihood, 79

Likelihood-Funktion, 145

lineare Diskriminanzanalyse, 228

lineare Separierbarkeit, 220

Liniendiagramm, 40

Lloyd-Algorithmus, 274

logistische Verteilung, 162

Logit-Funktion, 230

Log-Likelihood-Funktion, 145

## M

Manhattan-Abstand, 168

maschinelles Lernen, 190

Massenfunktion

bedingte, 93

gemeinsame, 92

multivariate, 184

Matrix

positiv semidefinite, 181

Maximalkardinalität, 13

Maximum-a-posteriori-Schätzer,

152, 228

- Maximum-Likelihood-Schätzer, 146  
 Maximumsnorm, 168  
 Maximums-Pooling, 247  
 Median, 96  
   empirischer, 48  
   geometrischer, 176  
   mittlere Abweichung vom, 53  
 Medoid, 33, 177  
 mehrschichtiges Lernen, 233  
 Merkmal, 38  
   binäres, 39  
   bivariate, 39  
   kategoriales, 38  
   metrisches, 38  
   multivariate, 39  
   nominales, 38  
   ordinale, 38  
   qualitative, 38  
   quantitative, 38  
   univariates, 39  
 Merkmalsausprägung, 38  
 Merkmalsauswahl, 259  
 Merkmalsextraktion, 259  
 Merkmalsliste, 39  
 Merkmalsraum, 38  
 Merkmalsträger, 39  
 Merkmalsvektor, 39  
 messbare Teilmenge, 71  
 Metadaten, 12  
 Methode der kleinsten Quadrate, 158  
 Metrik, 173  
   diskrete, 29  
 Mini-Batch, 206  
 Minkowski-Norm, 168  
 Mittel  
   arithmetisches, 48  
   geometrisches, 51  
   harmonisches, 51  
   quadratisches, 51  
 Mittelwertfilter, 246  
 Mittelwertimputation, 26  
 Mittelwertzentrierung, 166, 176  
 MNIST, 266, 267, 271, 284  
 Modellauswahl, 199  
 Modellmatrix, 166  
 Modellparameter, 192  
 Modus, 45  
 Moore-Penrose-Inverse, 213  
 multidimensionale Skalierung  
   metrische, 268  
   nach Sammon, 268  
 Multigraph, 18  
 Multinomialverteilung, 185  
  
**N**  
 Netzwerkdiagramm, 233, 295  
 neuronales Netzwerk, 233  
   residuale, 234  
   vorwärtsgerichtetes, 233  
 N-Gramme, 292  
 Normalverteilung, 100, 121, 147  
   multivariate, 182  
 Null-Eins-Verlustfunktion, 194  
 Nullhypothese, 116, 137  
  
**O**  
 Objektidentifikation, 29  
 One-Hot-Kodierung, 192  
  
**P**  
 Pareto-Verteilung, 123, 146  
 Pearson-Korrelation, 57, 160  
 Perplexität, 270  
 PMI, 76  
 $p$ -Norm, 168  
 Poisson-Verteilung, 119  
 Polarität, 288  
 Pooling, 247  
 Prämetrik, 173  
 Prävalenz, 80  
 Prinzip der maximalen Entropie, 122  
 Prognosebereich, 160  
 Prognoseintervall, 137  
 Projektion, 16  
 Property-Graph-Modell, 19  
 Prozentrang, 59  
 Punktschätzung, 133  
  
**Q**  
 quanteda, 294

- Quantil, 96  
  empirisches, 50
- Quantilfunktion, 96
- Quartil, 51
- Quartilsabstand, 53
- Quasi-Newton-Bedingung, 207
- R**
- Randdichte, 92  
  der Normalverteilung, 184
- Randhäufigkeit, 62
- Rang, 59
- Rangkorrelationskoeffizient, 60
- Rastersuche, 200
- Realisierung, 69, 82
- Record-Linkage, 29
- Regression, 191  
  lineare, 157, 209  
  logistische, 162, 219  
  logistische mit Kern, 221
- Regressionsfunktion, 192
- Regressionsgerade, 157
- reguläre Ausdrucke, 25
- Relation, 15
- relationale Algebra, 15
- Relieffilter, 246
- Residuenquadratsumme, 211
- Ringdiagramm, 44
- Risiko  
  empirisches, 195  
  erwartetes, 196
- Robustheit, 49, 161
- Rohdaten, 3
- S**
- Sankt-Petersburg-Paradoxon, 97
- Satz von  
  Bayes, 79  
  Gliwenko-Cantelli, 145  
  Greary, 140  
  Slutsky, 136
- Säulendiagramm, 40
- Schätzer  
  asymptotisch erwartungstreuer, 131  
  erwartungstreuer, 127  
  konsistenter, 127, 130
- Schätzfehler, 199
- Schicht  
  Pooling-, 248  
  voll vernetzte, 248
- Schlüsselattribut, 15
- Schwerpunkt, 159, 176
- Schwertlilien, 54, 169, 177, 178
- Selektion, 16
- Sensitivität, 79, 202
- Sentimentanalyse, 288  
  auf Aspektbene, 293
- Shannon-Entropie, 101  
  empirische, 55  
  empirische gemeinsame, 62
- Sigmoidfunktion, 234
- Signifikanzniveau, 136
- Single-Linkage-Verfahren, 277
- Sobel-Operator, 246
- Softmax-Funktion, 235
- Softplus-Funktion, 235
- Sonifikation, 2
- spaCy, 294
- Spamfilter, 80, 194, 203
- Spannweite, 53
- Spezifität, 79, 202
- SQL, 17
- Standardabweichung, 99  
  empirische, 53
- Standardnormalverteilung, 87, 121
- Statistik  
  Bayes'sche, 150  
  beschreibende, 37  
  deskriptive, 37  
  explorative, 40  
  inferenzielle, 115  
  schließende, 115  
  statistische Einheit, 39  
  statistisches Modell, 116
- Stichprobe, 38
- Stichprobengröße, 39
- Stichprobenliste, 39
- Stichprobenvariablen, 129
- Stichprobenvektor, 39
- Stochastik, 69
- Stoppwörter, 293
- Störgröße, 157

Streudiagramm, 42  
 Streuungsparameter, 53  
 Student'sche  $t$ -Verteilung, 111  
 Student'scher  $t$ -Test, 138  
 Synapse, 234  
 Szymkiewicz-Simpson-Koeffizient, 171

**T**

Tanimoto-Ähnlichkeit, 170

Teilgesamtheit, 39

TensorFlow, 287

Testdatensatz, 200

Testfehler, 196

Textanalyse, 293

Theil-Sen-Verfahren, 161

Themenkarte, 294

Tokenisierung, 229

Träger, 83

Trainingsdatensatz, 200

Trainingsfehler, 195

Transinformation

  empirische, 62

Transinformationsgehalt, 76

Trefferquote, 202

Tschebyscheff-Abstand, 168

Tschebyscheff'sche Ungleichung, 104

$t$ -SNE-Verfahren, 270

$t$ -Verteilung, 111

Typkompatibilität, 16

**U**

Überanpassung, 199

Überlappungskoeffizient, 171

überwachtes Lernen, 192

UMAP-Verfahren, 270

universelle

  Approximationseigenschaft, 239

unscharfe Menge, 272

unscharfe Suche, 29

Unteranpassung, 199

unüberwachtes Lernen, 255

**V**

Validierungsdatensatz, 200

Varianz, 99, 199  
  empirische, 53  
  gepoolte, 143  
  korrigierte empirische, 54, 131

Varianzschätzung, 131

Vektornorm, 167

Verbund, 16

Verfahren des steilsten Abstiegs, 204

Verlustfunktion, 194

Verteilungsfunktion, 82

  empirische, 143

  gemeinsame, 92

  logistische, 162

Vertrauensbereich, 159

Vertrauensintervall, 133

Vertrauensniveau, 137

Verzerrung, 199

Verzerrung-Varianz-Dilemma, 199

Vokabular, 229

Vollerhebung, 128

Vorhersagebereich, 160

Vorhersageintervall, 137

**W**

Wahlforschung, 44, 65, 69, 78

Wahrheitsmatrix, 201

Wahrscheinlichkeit

  Bayes'sche, 70

  bedingte, 75

  empirische, 69

  frequentistische, 70

Wahrscheinlichkeitsdichte

  endlastige, 123

Wahrscheinlichkeitsdichtefunktion, 86

  bedingte, 94

  multivariate, 181

Wahrscheinlichkeitsfunktion

  gemischte, 93

Wahrscheinlichkeitsmaß, 71

Wahrscheinlichkeitsmassenfunktion  
  bedingte, 93

Wegzusammenhangskomponente, 258

Weiszfeld-Algorithmus, 177

Worthäufigkeitsanalyse, 291

**X**

XML, 22

**Z**

Zähldichte, 85

Zeitreihe, 40

Zentraler Grenzwertsatz von Lindeberg-Lévy, 132

Zero-Padding, 245

Zielfunktion, 158

Zielgröße, 157

Zufallsexperiment, 70

Zufallsmatrix, 179

Zufallssuche, 200

Zufallsvariable, 69, 82

absolutstetige, 83

diskrete, 83

latente, 162

stetige, 83

transformierte, 88

Zufallsvariablen

unabhängige, 95

unkorrelierte, 101

Zufallsvektor, 179

Zusammenhangskomponente, 258

z-Wert, 166