

Statistik
und
Wahrscheinlichkeitstheorie
für Informatik | WS 2016
 $\langle \text{StatWth16} \rangle$

Werner Gurker

Copyright © 2016 by Werner Gurker
All rights reserved.

Ass.Prof. Dipl.-Ing. Dr.techn. Werner Gurker
Institut für Stochastik und Wirtschaftsmathematik
Technische Universität Wien
Wiedner Hauptstraße 8–10
Turm A (6. Stock)
A – 1040 Wien

Tel.: (+ 43 1) 58801–10583
E-Mail: W.Gurker@tuwien.ac.at

V o r w o r t

Der nachfolgende Text bildet die Grundlage für die Vorlesung [107.254] und die Übung [107.369] zur **Statistik und Wahrscheinlichkeitstheorie** für Studierende der Bachelorstudien Informatik und Wirtschaftsinformatik im WS 2015 an der TU-Wien. Der mit zahlreichen Beispielen breit angelegte Text geht dabei über den Rahmen einer zweistündigen Vorlesung hinaus und bietet somit interessierten Hörerinnen und Hörern weitere Anregungen und ergänzende Materialien zu den hier behandelten Themenkreisen.

Für die Aufbereitung und Auswertung von Datensätzen, für sonstige Berechnungen und für die Erstellung der Abbildungen wird in diesem Text das unter der **GNU General Public License** frei verfügbare Statistikpaket R verwendet.¹ Neben einer stetig wachsenden Zahl von Lehrbüchern (vgl. Literatur für einige Hinweise) finden sich naturgemäß auch im Internet zahlreiche Hilfestellungen und Manuals zu dieser – speziell im universitären Bereich – weit verbreiteten Statistiksoftware. Zusätzlich empfiehlt sich die Installation einer auf R abgestimmten Entwicklungsumgebung (RStudio, Tinn-R, ...). Zur leichteren Einarbeitung werden die R-Skripts zu den Beispielen im Text sowie zu den mittels R zu bearbeitenden Übungsaufgaben den zur Übung Angemeldeten auf TISS zur Verfügung gestellt.

Wien, September 2016

W. G.

¹<http://www.r-project.org>

Inhaltsverzeichnis

1	Deskriptive und explorative Statistik	1
1.1	Grundgesamtheit	1
1.2	Stichproben	2
1.3	Merkmale	2
1.4	Messniveau	4
1.5	Datenmatrix	6
1.6	Diskrete univariate Merkmale	7
1.6.1	Häufigkeiten	7
1.6.2	Kreisdiagramm	8
1.6.3	Balkendiagramm	9
1.6.4	Mosaikplot	9
1.6.5	Pareto–Diagramm	10
1.7	Stetige univariate Merkmale	13
1.7.1	Ordnungsstatistiken	13
1.7.2	Empirische Verteilungsfunktion	14
1.7.3	Stem-and-Leaf–Plot	15
1.7.4	Klassierung	16
1.7.5	Histogramm	18
1.7.6	Kernschätzung	20
1.7.7	Quantile	23
1.7.8	QQ–Plot	26
1.7.9	Boxplot	27
1.8	Kennzahlen	29
1.8.1	Mittelwert	30
1.8.2	Geometrisches und harmonisches Mittel	31
1.8.3	Getrimmter Mittelwert	33
1.8.4	Median	34

1.8.5	Varianz	35
1.8.6	MAD	37
1.8.7	Datenzusammenfassung	38
1.8.8	Modalwert	39
1.8.9	Momente	41
1.8.10	Schiefe	41
1.8.11	Kurtosis	42
1.8.12	Verteilungsform	43
1.9	Mehrdimensionale Daten	44
1.9.1	Scatterplots	46
1.9.2	Kernschätzung	46
1.9.3	Korrelation	49
1.9.4	Kleinste Quadrate	56
	Aufgaben	60
2	Wahrscheinlichkeit	65
2.1	Gesetz der großen Zahlen	65
2.2	Merkmalraum	67
2.3	Ereignisse	69
2.4	Borelmengen	71
2.5	Wahrscheinlichkeitsmaße	73
2.6	Chancen (Odds)	74
2.7	Endliche W–Räume	76
2.8	Geometrische Wahrscheinlichkeiten	77
2.9	Additionstheorem	78
2.10	Bedingte Wahrscheinlichkeit	81
2.11	Multiplikationstheorem	83
2.12	Vollständige Wahrscheinlichkeit	84
2.13	Bayes'sche Formel	86

2.14	Unabhängigkeit	88
2.15	Mehrstufige Experimente	91
2.16	Beispiele	94
	Aufgaben	102
	Anhang: Abzählende Kombinatorik	107
3	Stochastische Größen und Verteilungen	111
3.1	Stochastische Größen	111
3.2	Verteilungsfunktion	113
3.2.1	Diskrete Verteilungen	118
3.2.2	Stetige Verteilungen	118
3.2.3	Gemischte Verteilungen	122
3.3	Transformationen	125
3.3.1	Transformationen diskreter sGn	126
3.3.2	Transformationen stetiger sGn	127
3.4	Erwartungswert	132
3.5	Varianz	136
3.6	Simulation	139
	Aufgaben	143
4	Spezielle Verteilungen	147
4.1	Diskrete Verteilungen	147
4.1.1	Diskrete uniforme Verteilung	147
4.1.2	Bernoulli–Verteilung	149
4.1.3	Binomialverteilung	150
4.1.4	Negative Binomialverteilung	152
4.1.5	Geometrische Verteilung	154
4.1.6	Hypergeometrische Verteilung	157
4.1.7	Poisson–Verteilung	159

4.2 Stetige Verteilungen	162
4.2.1 Stetige uniforme Verteilung	163
4.2.2 Exponentialverteilung	164
4.2.3 Gamma- und Chiadratverteilung	168
4.2.4 Normalverteilung	170
4.2.5 F–Verteilung	174
4.2.6 t–Verteilung	176
4.2.7 Betaverteilung	177
Aufgaben	179
Anhang: R–Funktionen	182
5 Multivariate Verteilungen	185
5.1 Bivariate Verteilungen	185
5.1.1 Diskrete stochastische Vektoren	186
5.1.2 Stetige stochastische Vektoren	188
5.1.3 Erwartungswert	191
5.1.4 Bedingte Verteilungen	192
5.2 Korrelation	197
5.3 Unabhängigkeit	200
5.4 Mehrdimensionale Erweiterungen	203
5.4.1 Varianz–Kovarianzmatrix	206
5.5 Transformationen	207
5.6 Spezielle multivariate Verteilungen	213
5.6.1 Multinomialverteilung	213
5.6.2 Polyhypergeometrische Verteilung	215
5.6.3 Multivariate Normalverteilung	217
Aufgaben	223
6 Folgen von stochastischen Größen	227

6.1	Lineare Funktionen	227
6.2	Faltung	230
6.2.1	Diskrete Faltung	230
6.2.2	Stetige Faltung	232
6.2.3	Additionstheoreme	233
6.3	Konvergenz	235
6.3.1	Ungleichungen	236
6.3.2	Gesetz der großen Zahlen	238
6.3.3	Zentraler Grenzverteilungssatz	240
6.3.4	Normalapproximation	245
	Aufgaben	247
7	Schließende Statistik	251
7.1	Grundbegriffe	251
7.2	Schätzer	253
7.2.1	Empirische Verteilungsfunktion	253
7.2.2	Momentenschätzer	255
7.2.3	Maximum Likelihood	257
7.2.4	Gütekriterien für Schätzer	263
7.3	Konfidenzintervalle	270
7.3.1	Pivotmethode	270
7.3.2	Approximativer Konfidenzintervall für den Mittelwert	273
7.3.3	Normalverteilung (eine Stichprobe)	274
7.3.4	Normalverteilung (zwei ua. Stichproben)	275
7.3.5	Normalverteilung (verbundene Stichproben)	277
7.3.6	Exponentialverteilung	279
7.3.7	Bernoulli–Verteilung	280
7.3.8	Poisson–Verteilung	282
7.3.9	Resampling und Bootstrapping	284

7.4	Statistische Tests	287
7.4.1	Parametertests	287
7.4.2	p–Wert	292
7.4.3	Beziehung zwischen Tests und Konfidenzintervallen	294
7.4.4	Tests für den Mittelwert einer Normalverteilung (Varianz bekannt)	295
7.4.5	Tests für den Mittelwert einer Normalverteilung (Varianz unbekannt)	296
7.4.6	Tests für die Varianz einer Normalverteilung	299
7.4.7	Tests für einen Anteil	300
7.4.8	Tests für die Mittelwerte von zwei Normalverteilungen	304
7.4.9	Tests für die Varianzen von zwei Normalverteilungen	306
7.4.10	Tests für den Korrelationskoeffizienten	308
7.4.11	Normal-QQ–Plot	309
7.4.12	Chiquadrat–Anpassungstests	312
	Aufgaben	317
	Anhang: Normal-W–Netz	324
8	Bayes–Statistik	325
8.1	A-priori– und A-posteriori–Verteilung	325
8.2	Konjugierte Verteilungsfamilien	328
8.3	Bayes–Schätzer	333
8.4	Bayes'sche Intervallschätzer	335
8.5	Bayes–Tests	336
	Aufgaben	339
9	Regressionsanalyse	341
9.1	Einfache lineare Regression	341
9.1.1	Parameterschätzung	344
9.1.2	Verteilung der Koeffizienten	347
9.1.3	Varianzzerlegung	348

9.1.4	Bestimmtheitsmaß	350
9.1.5	ANOVA–Tafel und F–Test	351
9.1.6	Konfidenzintervalle und t–Tests	354
9.1.7	Residualanalyse	358
9.1.8	Ausreißer und Heelpunkte	360
9.1.9	Matrixschreibweise	364
9.2	Multiple lineare Regression	366
9.2.1	Parameterschätzung	367
9.2.2	ANOVA–Tafel und F–Test	369
9.2.3	Konfidenzintervalle und t–Tests	371
9.2.4	Beispiele	372
	Aufgaben	381
	Tabellen	385
	Literatur	389

1 Deskriptive und explorative Statistik

Die **deskriptive** (beschreibende) **Statistik** beschäftigt sich mit der tabellarischen und grafischen Aufbereitung von Daten sowie mit ihrer zahlenmäßigen Beschreibung (Berechnung von Kenngrößen). In der deskriptiven Statistik verwendet man keine statistischen (stochastischen) Modelle, sodass die aus den Daten gewonnenen Erkenntnisse nicht durch Fehlerwahrscheinlichkeiten abgesichert werden können. Letzteres lässt sich mit Hilfe der **schließenden Statistik** bewerkstelligen, soferne die unterstellten Modellannahmen (zumindest näherungsweise) zutreffen.

Die **explorative Datenanalyse**¹ (oder kurz **EDA**) hat zum Ziel, unbekannte Strukturen und Zusammenhänge in den Daten aufzudecken und Hypothesen über den datengenerierenden Prozess zu formulieren. Neben ihrer Eignung als Einführung in das statistische Denken generell werden Methoden der EDA u. a. im viel diskutierten *Data-Mining* (Verarbeitung sehr großer Datenbestände) eingesetzt.

1.1 Grundgesamtheit

Der erste Schritt jeder Datenanalyse ist die Erhebung der Daten an **statistischen Einheiten**, entweder durch Experimente oder durch Beobachtungsstudien. Im ersten Fall nennt man die statistischen Einheiten auch **Versuchseinheiten**, im zweiten Fall auch **Beobachtungseinheiten**.

Bem: Statistische Untersuchungen werden häufig zur Bestätigung (oder Widerlegung) von *kausalen* Zusammenhängen herangezogen. Dabei ist allerdings Vorsicht geboten. Im strengen Sinn erlauben nur (adäquat durchgeführte) Experimentalstudien Rückschlüsse auf kausale Zusammenhänge, nicht aber Beobachtungsstudien. Letztere können nur Hinweise auf *assoziative* Zusammenhänge liefern. Experimentalstudien sind also zu bevorzugen, aber nicht immer möglich.

Die statistischen Einheiten, über die – deskriptiv und/oder explorativ – Aussagen getroffen werden sollen, bilden die **Grundgesamtheit** oder **Population**. Eine präzise Definition dieser Größen als Basis einer tragfähigen Datenanalyse ist unumgänglich, häufig aber mit Problemen der Ab- bzw. Eingrenzung verbunden. Man betrachte dazu etwa das folgende Beispiel.

Bsp 1.1 Soll beispielsweise die Wirtschaftskraft von kleinen österreichischen IT–Unternehmen untersucht werden, so ist zunächst zu klären, was „kleine IT–Unternehmen“ sind. Als Kriterien bieten sich etwa Mitarbeiterzahl und/oder Umsatz an. Aber auch die Frage, was ein „IT–Unternehmen“ ist, lässt sich nicht eindeutig beantworten. Alle IT–Unternehmen (in diesem Fall sind es Beobachtungseinheiten), die die festgelegten Kriterien erfüllen, bilden dann die Grundgesamtheit. ■

¹Initiiert in den 1970er Jahren vom US-amerikanischen Mathematiker und Statistiker JOHN WILDER TUKEY (1915–2000).

1.2 Stichproben

Eine Untersuchung *aller* Elemente einer Grundgesamtheit (d. h. eine **Gesamterhebung**) ist aus Zeit- und/oder Kostengründen, aber auch aus prinzipiellen Gründen (etwa wenn die Grundgesamtheit – tatsächlich oder potenziell – unendlich ist) nicht immer möglich. In solchen Fällen beschränkt man sich auf eine **Stichprobe**, d. h. auf eine **repräsentative** Teilauswahl aus der Grundgesamtheit.

Um ein getreues Abbild der Grundgesamtheit zu bekommen, sollte die Auswahl rein **zufällig** erfolgen. Besteht die Grundgesamtheit aus N Elementen und soll eine Stichprobe des Umfangs n gezogen werden, so gibt es dafür

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

verschiedene Möglichkeiten, falls die Elemente der Grundgesamtheit unterscheidbar sind. Werden die n Elemente nun so ausgewählt, dass jede der $\binom{N}{n}$ möglichen Stichproben die gleiche Auswahlwahrscheinlichkeit hat, so spricht man von einer (**einfachen**) **Zufallsstichprobe**. In diesem Fall hat jedes Element der Grundgesamtheit die gleiche Chance, in die Stichprobe zu gelangen.

Bei der oben beschriebenen Form der Stichprobenziehung wird jedes Element der Grundgesamtheit *höchstens einmal* ausgewählt. Das nennt man **Ziehen ohne Zurücklegen**. Andererseits ist es aber auch möglich, eine bereits erhobene Einheit ein weiteres Mal zu berücksichtigen. Diese Form der zufälligen Stichprobenentnahme nennt man **Ziehen mit Zurücklegen**.

Eine reine Zufallsauswahl der beschriebenen Art ist in vielen praktisch wichtigen Fällen nicht durchführbar oder auch nicht adäquat. Man betrachte etwa das folgende Beispiel.

Bsp 1.2 Angenommen, ein Industriebetrieb bezieht bestimmte Komponenten von drei verschiedenen Zulieferfirmen, die sich hinsichtlich der Qualität ihrer Produktion unterscheiden. Konkret beziehe der Betrieb $N_1 = 2000$ Komponenten von Firma 1, $N_2 = 1000$ von Firma 2, und $N_3 = 3000$ von Firma 3, insgesamt also $N = 6000$ Stück. Wenn nun der Betrieb eine Qualitätsprüfung durchführen möchte und dafür einen Stichprobenumfang von $n = 300$ festlegt, so liegt es nahe, eine proportionale Schichtung vorzunehmen, d. h., aus den Komponenten von Firma 1 (2, 3) eine Stichprobe der Größe $n_1 = 100$ ($n_2 = 50$, $n_3 = 150$) zu ziehen. Eine Ziehung dieser Art nennt man eine *geschichtete* Stichprobenziehung. ■

1.3 Merkmale

Im nächsten Schritt werden an den ausgewählten Einheiten (der Stichprobe) die interessierenden Größen erhoben, **Merkmale** oder **Variablen** genannt. Die Werte, die von einem Merkmal angenommen werden können (d. h. die möglichen Ausprägungen) nennt

man die **Merkmalsausprägungen**. Die Menge dieser Ausprägungen wird üblicherweise mit M bezeichnet.

Ein Merkmal ist eine Abbildung: Mathematisch ausgedrückt ist ein Merkmal eine Abbildung (Funktion) $X : G \rightarrow M$, die jeder statistischen Einheit $g \in G$ (Grundgesamtheit) eine Ausprägung $X(g) \in M$ zuordnet. Dabei kann es sich auch um Ausprägungsvektoren handeln. Misst man beispielsweise an Personen die Körpergröße und das Körpergewicht, so gilt $X(\text{Person}) = (h, w) \in (\mathbb{R}^+)^2$.

Bsp 1.3 Merkmalsausprägungen können von ganz unterschiedlicher Art sein. Beispielsweise hat das Merkmal „Geschlecht“ nur zwei Ausprägungen (die allein der Unterscheidung dienen); das Merkmal „Mitarbeiterzahl“ (eines Unternehmens) ist eine Zählvariable mit (potenziell) unbeschränkt vielen Ausprägungen. Die Funktionsdauer einer Batterie (in Betriebsstunden) hingegen ist ein auf ein Intervall beschränktes metrisches Merkmal. ■

Studiendesigns: Es gibt eine Reihe von – nicht streng voneinander trennbaren – Formen der Datengewinnung, u. a. die folgenden:

Querschnittsstudien: Bei Querschnittsstudien werden zu einem festen Zeitpunkt die interessierenden Merkmale an den statistischen Einheiten erhoben. Dies führt zu „Momentaufnahmen“. Ein Beispiel sind die alle drei Jahre durchgeführten PISA (*Programme for International Student Assessment*)–Studien der OECD zur Erfassung der Kenntnisse und Fähigkeiten von 15-jährigen Schüler/innen.

Longitudinalstudien: Bei Longitudinalstudien werden an einer unverändert bleibenden Gruppe (*Panel*) von statistischen Einheiten Merkmale zu mehreren Zeitpunkten erhoben. Dadurch sollen zeitliche Entwicklungen erkennbar werden. Ein Beispiel ist das SOEP (Sozio-oekonomisches Panel) des DIW (Deutsches Wirtschaftsforschungsinstitut), eine jährlich wiederholte Befragung ausgewählter privater Haushalte bezüglich Einkommen, Gesundheit, etc., mit teilweiser Anwendung auch auf österreichische Verhältnisse.

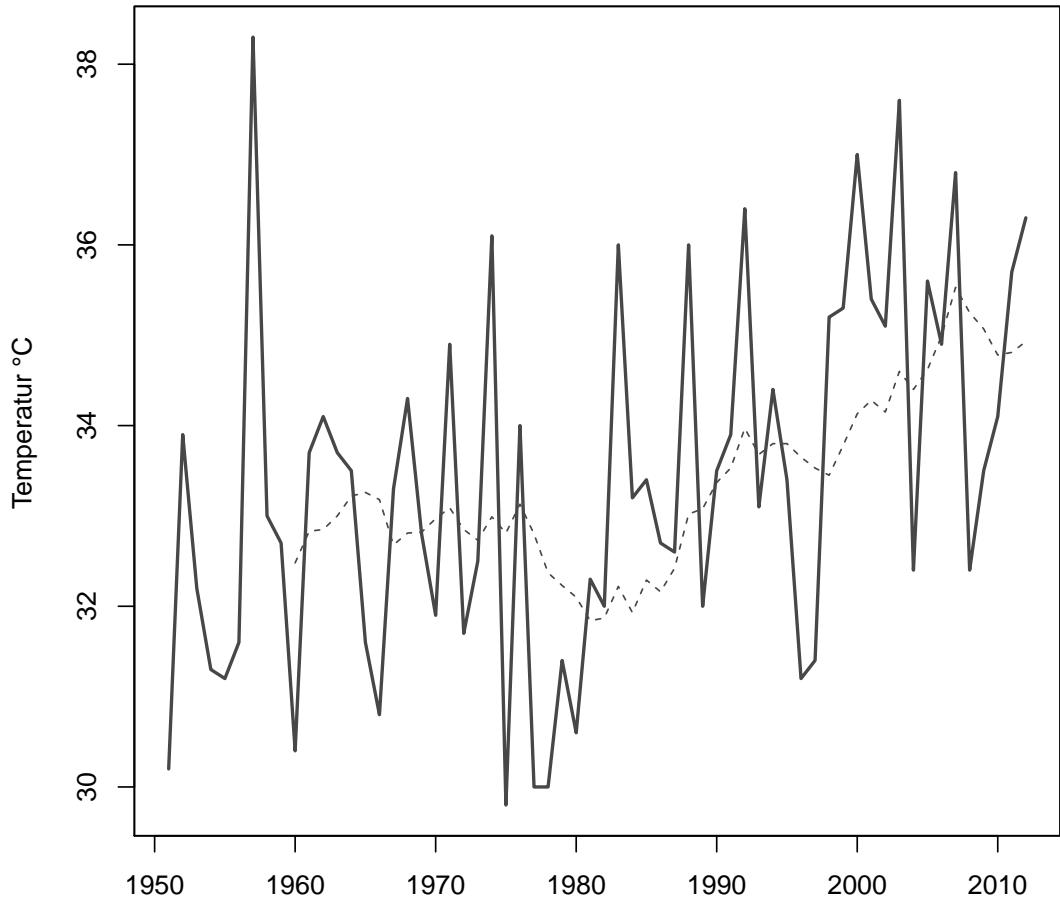
Zeitreihen: Man spricht allgemein von Zeitreihen, wenn die interessierenden Merkmale an einer einzelnen statistischen Einheit zu verschiedenen Zeitpunkten erhoben werden. Beispiele sind etwa Wetterbeobachtungen oder ökonomische Daten.

Bsp 1.4 Als Beispiel für eine Zeireihe betrachten wir die Jahreshöchsttemperaturen in Wien (Messstelle: Hohe Warte) für die Jahre 1951–2012. Überlagert wird der Plot (Abb 1.1) von einem *gleitenden Durchschnitt* der Spannweite $w = 10$. Ist x_t die Beobachtung zum Zeitpunkt t , so berechnet man jeweils den folgenden Durchschnittswert:

$$\hat{x}_t = \frac{x_t + x_{t-1} + \dots + x_{t-w+1}}{w}$$

Gleitende Durchschnitte dienen als *Filter*, um möglicherweise vorhandene Trends klarer zu erkennen. (Bem: Die statistische Analyse von Zeitreihen erfordert spezielle Methoden, die über den Rahmen dieser VO hinausgehen.) ■

Abbildung 1.1: Jahreshöchsttemperaturen in Wien/Hohe Warte von 1951 bis 2012



1.4 Messniveau

Auch wenn Merkmalsausprägungen meist durch Zahlen repräsentiert werden, heißt das nicht automatisch, dass auch alle Rechenoperationen (oder Vergleiche) mit diesen Zahlen durchgeführt werden können bzw. sinnvoll sind. Der Umfang der zulässigen Operationen (oder der zur Verfügung stehenden Methoden der statistischen Analyse) ist abhängig vom **Messniveau** des Merkmals. Man kann zwischen **qualitativen** und **quantitativen** oder zwischen **diskreten** und **stetigen** Merkmalen unterscheiden. Genauer unterscheidet man zwischen den folgenden Messskalen:

Nominalskalen: Hierbei handelt es sich um eine reine Klassifikation, darüberhinaus bestehen keine weiteren Relationen zwischen den Elementen der Grundgesamtheit. Zahlenmäßige Ausprägungen eines solchen Merkmals sind nur eine zweckmäßige Codierung.

Bsp: Geschlecht, Familienstand, Religionsbekenntnis, ...

Ordinalskalen: Kennzeichnend für Rangmerkmale ist eine lineare Ordnungsbeziehung, darüberhinaus sind keine weiteren Beziehungen vorhanden. Zahlenmäßige Ausprägungen eines solchen Merkmals spiegeln diese Ordnung wider.

Bsp: Prüfungsnoten, Güteklassen von Obst, Windstärke (z. B. Beaufort-Skala von 0 bis 12), ...

Bem: Häufig wird ein an sich metrisch skaliertes Merkmal auf ein Rangmerkmal reduziert (ein Beispiel ist die vorhin erwähnte Beaufort-Skala).

Intervallskalen: Die Ausprägungen sind reelle Zahlen (oder Vektoren), wobei der Nullpunkt – sofern vorhanden – keine absolut festgelegte Bedeutung hat (sondern nur zur Definition der Skala dient). Differenzen haben eine sinnvolle Interpretation, Aussagen wie „doppelt so warm“, „halb so spät“, ... hingegen nicht.

Bsp: Zeiteinteilung (0 bis 24 Uhr), Temperatur in Grad Celsius oder Grad Fahrenheit ($F = \frac{9}{5}C + 32$), ...

Verhältnisskalen: Hierbei handelt es sich um Intervallskalen mit ausgeprägtem und interpretierbarem Nullpunkt. Aussagen wie „doppelt so hoch“, „halb so schnell“, ... sind sinnvoll.

Bsp: Körpergröße, Geschwindigkeit, Temperatur in Kelvin, Häufigkeiten, ...

Bemerkungen:

- (a) Manchmal wird eine nominalskalierte Variable durch entsprechende Codierung auf Ordinarniveau „gehoben“, beispielsweise bei einer Befragung nach der Schulbildung (1 = Volksschule, 2 = Berufsschule, 3 = Matura, ...). Bei dieser Vorgangsweise ist allerdings Vorsicht geboten, damit nicht etwa versteckte (eigene) Wertungen in die Erhebung einfließen. Sie ist letztlich nur dort angebracht, wo es sich tatsächlich bereits um ein ordinales Merkmal handelt.
- (b) Diskrete Merkmale werden oft als stetige Merkmale behandelt, wenn die Schrittweite in Bezug auf die beobachtete Größe klein ist.
Bsp: Umsätze eines Betriebes, Schaltvorgänge bis zum Ausfall eines Schalters, ...
- (c) Jede praktische Messung eines stetigen Merkmals ist – bedingt durch die beschränkte Messgenauigkeit – tatsächlich diskret; beträgt die Messgenauigkeit etwa 0.001 mm, ist jede Messung ein Vielfaches von 0.001 mm. Anders ausgedrückt: Ein Messwert x entspricht tatsächlich dem Intervall $\langle x - 0.0005, x + 0.0005 \rangle$ – die Zuordnung der Randpunkte erfolgt entsprechend der Rundungsregel. Allerdings ist die Vorstellung, dass (bei unendlicher Messgenauigkeit) jeder Punkt eines Intervalls prinzipiell als Ausprägung in Frage kommen könnte, für die statistische Modellbildung wichtig.
- (d) Als Folge der durch den Nullpunkt gegebenen (linksseitigen) Beschränkung der Messwerte, weisen verhältnisskalierte Merkmale häufig eine schiefe Verteilung auf.
- (e) Intervall- und Verhältnisskalen werden auch als **metrische** oder **kardinale** Skalen, Nominal- und Ordinalskalen auch als **topologische** Skalen bezeichnet.

1.5 Datenmatrix

Ausgangspunkt für eine tabellarische und/oder grafische Aufbereitung von Datensätzen sind zunächst die **Rohdaten** (oder **Urdaten, Primärdaten**). Die erhobenen Ausprägungen werden in einer **Datenmatrix** (oder einem **Datenframe**) dargestellt. Die Spalten einer Datenmatrix entsprechen den Variablen (Merkmale), die Zeilen den Untersuchungseinheiten.

Bsp 1.5 Der folgende R-Output zeigt einen Ausschnitt aus einem umfangreichen Datensatz (`body.txt`), bestehend aus einer Reihe von anthropometrischen Messwerten.² Hier werden nur 6 der insgesamt 25 Variablen betrachtet: Biacromial diameter (cm), Waist girth (cm), Age (years), Weight (kg), Height (cm), Gender (1/0 = male/female).

	Biacromial	Waist	Age	Weight	Height	Gender
1	42.9	71.5	21	65.6	174.0	1
2	43.7	79.0	23	71.8	175.3	1
3	40.1	83.2	28	80.7	193.5	1
4	44.3	77.8	23	72.6	186.5	1
5	42.5	80.0	22	78.8	187.2	1
.....						
505	34.7	57.9	33	48.6	160.7	0
506	38.5	72.2	33	66.4	174.0	0
507	35.6	80.4	38	67.3	163.8	0

In der i -ten Zeile der Datenmatrix stehen die p (hier ist $p = 6$) an der i -ten statistischen Einheit beobachteten Ausprägungen. In der j -ten Spalte stehen die n (hier ist $n = 507$) beobachteten Werte des j -ten Merkmals; n ist der Stichprobenumfang und p die Dimension der Daten.

Abgesehen vom nominellen Merkmal **Gender** sind hier alle Variablen metrisch skalierte Merkmale auf einer Verhältnisskala. (Bem: Man beachte auch, dass es hier keine *fehlenden* Beobachtungen gibt, bei umfangreichen Datensätzen sonst eher die Regel als die Ausnahme.) ■

Univariate/Multivariate Daten: Für $p = 1$ spricht man von **univariaten** Daten, ansonsten von **multivariaten** Daten. Die n beobachteten Ausprägungen x_1, x_2, \dots, x_n eines univariaten Merkmals werden häufig in einem n -dimensionalen **Datenvektor**³ \mathbf{x} zusammengefasst:

²G. HEINZ, L. J. PETERSON, R. W. JOHNSON, AND C. J. KERK: Exploring Relationships in Body Dimensions, *Journal of Statistics Education*, Vol. 11/2, 2003.

³Vektoren werden meist – so wie hier – als Spalten betrachtet, gelegentlich aber auch als Zeilen.

$$\mathbf{x} = (x_1, x_2, \dots, x_n)' \in \mathbb{R}^n$$

Grafische Darstellung univariater Daten: In den folgenden Abschnitten diskutieren wir die tabellarische und insbesondere grafische Aufbereitung univariater Datensätze. Die Darstellungsmöglichkeiten richten sich dabei nach dem Messniveau; zweckmäßigerweise unterscheidet man zwischen diskreten und stetigen Merkmalen.

1.6 Diskrete univariate Merkmale

Die Darstellung von diskreten (d. h. in erster Linie von nominalen und ordinalen) Daten erfolgt durch Bestimmung von Häufigkeiten und einer geeigneten Visualisierung. Gerade bezüglich des letzteren Punktes trifft man (speziell in den Medien) auf eine Fülle von Umsetzungen, die allerdings manchmal mit einer gewissen Skepsis zu betrachten sind.

1.6.1 Häufigkeiten

Ein diskretes Merkmal, das die Werte $x_1 < x_2 < \dots$ annehmen kann, werde insgesamt n Mal beobachtet. Die **absolute Häufigkeit** mit der x_i beobachtet wird, werde mit n_i bezeichnet. Der größte beobachtete Merkmalswert sei x_k ; dann gilt $\sum_{i=1}^k n_i = n$. Die **relativen Häufigkeiten** seien mit $f_i = n_i/n$ bezeichnet; für sie gilt $\sum_{i=1}^k f_i = 1$.

Nimmt das Merkmal die Werte $0, 1, 2, \dots$ an, handelt es sich um eine **Zählung**. Dabei ist zu beachten, dass die n Beobachtungen an Zählauschnitten (z. B. Zeit-, Längen-, Flächenabschnitten oder Volumen-, Gewichtseinheiten) gleicher Größe durchgeführt werden.

Bei ordinalem Skalenniveau sollten die Kategorien in der tabellarischen/grafischen Darstellung entsprechend angeordnet werden. Bei nominellen Merkmalen wählt man aus Gründen der Übersichtlichkeit meist eine Darstellung nach Häufigkeiten.

Bsp 1.6 Der Datensatz `beginner.txt` umfasst die Zahlen der Studienanfänger/innen an der TU-Wien für die Semester W2010, S2011, ..., W2013, aufgeschlüsselt nach Studienrichtung. In diesem Fall handelt es sich um ein nominelles Merkmal (Studienrichtung), dessen Ausprägungen (nach dem Anfangsbuchstaben) durch die Zahlen 1, 2, ..., 24 repräsentiert werden.

Im Weiteren betrachten wir nur die Wintersemester und zunächst nur das WS 2013. (Bem: Studienrichtungen mit weniger als 10 Neuinskriptionen bleiben unberücksichtigt, ebenso die Kategorie „unbekannt“ mit 177 Hörer/innen.) Gereiht nach der Zahl der Neuinskriptionen ergibt sich die folgende Häufigkeitsverteilung:

	Studienrichtung	Absolut	Relativ	Kumuliert
1	Architektur	1014	21.30	21.30
9	Informatik	690	14.50	35.80
2	Bauingenieurwesen	404	8.49	44.29
13	Maschinenbau	370	7.77	52.06
24	Wirtschaftsingenieurwesen	366	7.69	59.75
19	Technische Physik	356	7.48	67.23
18	Technische Mathematik	342	7.18	74.41
7	Elektrotechnik u Informationstechnik	304	6.39	80.80
17	Technische Chemie	279	5.86	86.66
16	Raumplanung u Raumordnung	230	4.83	91.49
20	Verfahrenstechnik	140	2.94	94.43
23	Wirtschaftsinformatik	125	2.63	97.06
22	Vermessungswesen	62	1.30	98.36
4	Biomedical Engineering	49	1.03	99.39
11	Lehramt	19	0.40	99.79
14	Materialwissenschaften	10	0.21	100.00

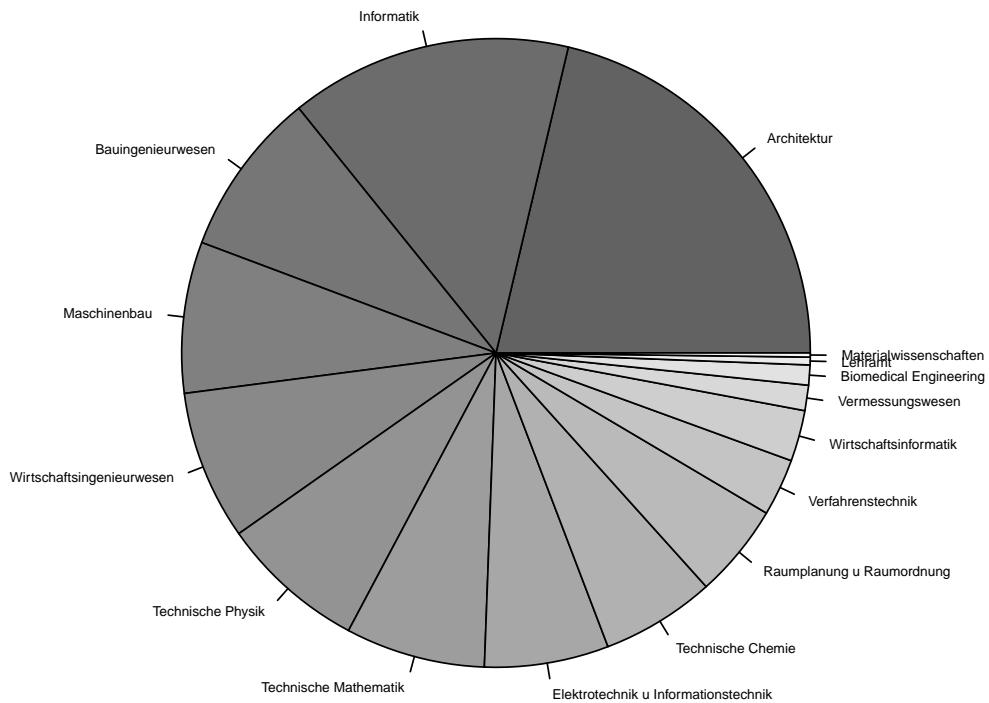
Bereits aus dieser Aufstellung lassen sich einige Einsichten gewinnen (beispielsweise, dass etwas mehr als die Hälfte der Neuinskriptionen auf nur vier Studienrichtungen entfallen), dennoch sind grafische Darstellungen meist aussagekräftiger. ■

1.6.2 Kreisdiagramm

Bei einem **Kreisdiagramm** (auch **Kuchen-** oder **Tortendiagramm** genannt) wird bei einem Kreis der Gesamtwinkel von 360° (bzw. 2π [rad]) entsprechend den absoluten oder relativen Häufigkeiten aufgeteilt. Zur relativen Häufigkeit f_i gehört also der Winkel $\varphi_i = f_i \cdot 360^\circ$ (bzw. $2\pi f_i$ [rad]).

Abb 1.2 zeigt das Kreisdiagramm für die Daten von Bsp 1.6 für das WS 2013. Die großen „Brocken“ Architektur und (in geringerem Ausmaß) Informatik sind augenfällig, hingegen ist eine Unterscheidung zwischen beispielsweise Maschinenbau und Technischer Mathematik nicht so einfach.

Bem: Auch wenn Kreisdiagramme beliebte Darstellungsmittel sind, sollte man Balkendiagramme (s. unten) bevorzugen. Nicht zuletzt auch deshalb, weil Kreisdiagramme durch entsprechende Farbgebung, oder gar durch Herausziehen einzelner Kreissegmente, etc. leicht eine manipulative Wirkung ausüben können. Ein Balkendiagramm hat überdies den Vorteil, dass speziell kleine Unterschiede in den relativen Häufigkeiten leichter erkennbar sind (vgl. Abb 1.3).

Abbildung 1.2: Neuinskriptionen an der TU–Wien im W2013 (Kreisdiagramm)

1.6.3 Balkendiagramm

Das **Balkendiagramm** (auch **Stabdiagramm** oder **Barplot**) ist eine grafische Darstellung der absoluten (oder relativen) Häufigkeiten mit senkrechten (manchmal auch waagrechten) Balken (oder Stäben) der Länge n_i (oder f_i) über den Merkmalswerten x_i .

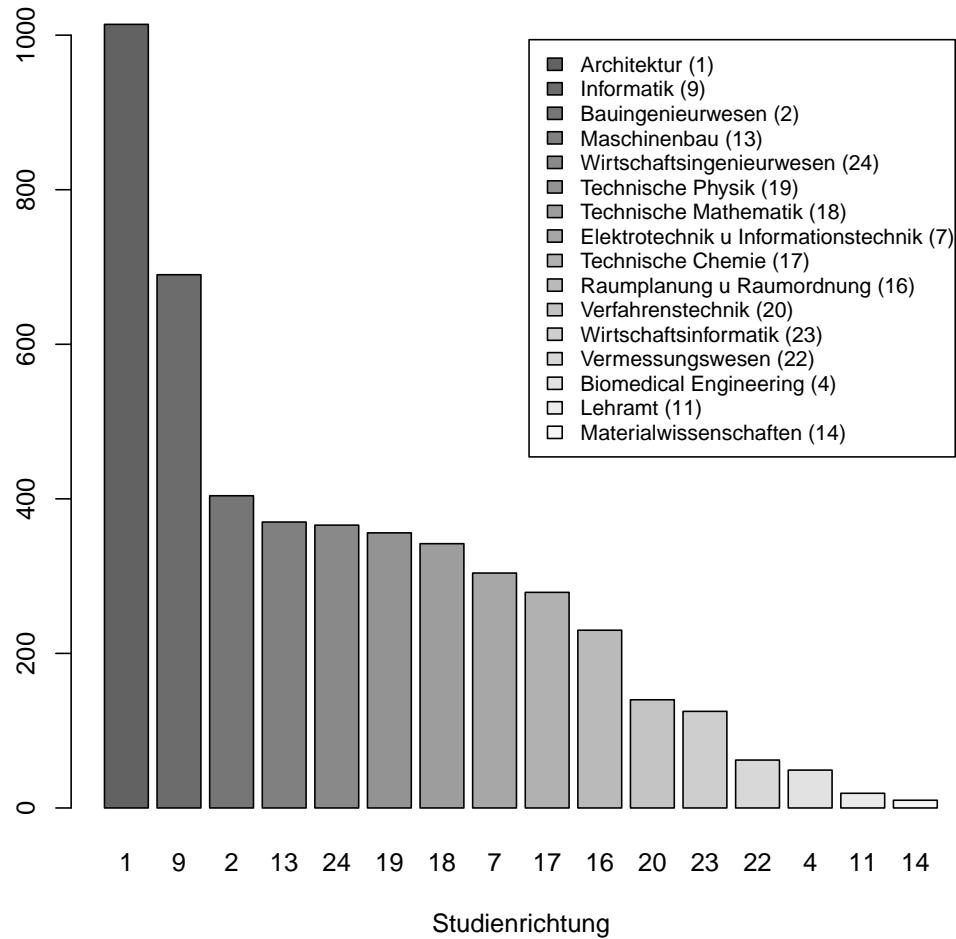
Beim Vergleich mehrerer Häufigkeitsverteilungen können für eine kompaktere Darstellung die Balken auch übereinander gestapelt gezeichnet werden.

Als Beispiel für einen Barplot betrachten wir wieder die Neuinskriptionen im WS 2013 (Abb 1.3), sowie einen Vergleich der Neuinskriptionen für W2010 bis W2013 (Abb 1.4). Für letzteren Vergleich werden die Balken übereinander gestapelt gezeichnet.

1.6.4 Mosaikplot

Der **Mosaikplot** dient zur Visualisierung von Datensätzen mit zwei oder mehreren qualitativen Merkmalen (und ist somit eigentlich eine multivariate Methode). Er gibt einen Überblick über die Daten und ermöglicht gleichzeitig das Erkennen von Zusammenhängen zwischen den verschiedenen Merkmalen. Bei zu vielen gleichzeitig betrachteten Merkmalen wirkt der Mosaikplot allerdings schnell unübersichtlich.⁴

⁴Vgl. <http://de.wikipedia.org/wiki/Mosaikplot> für weitere Details und Beispiele.

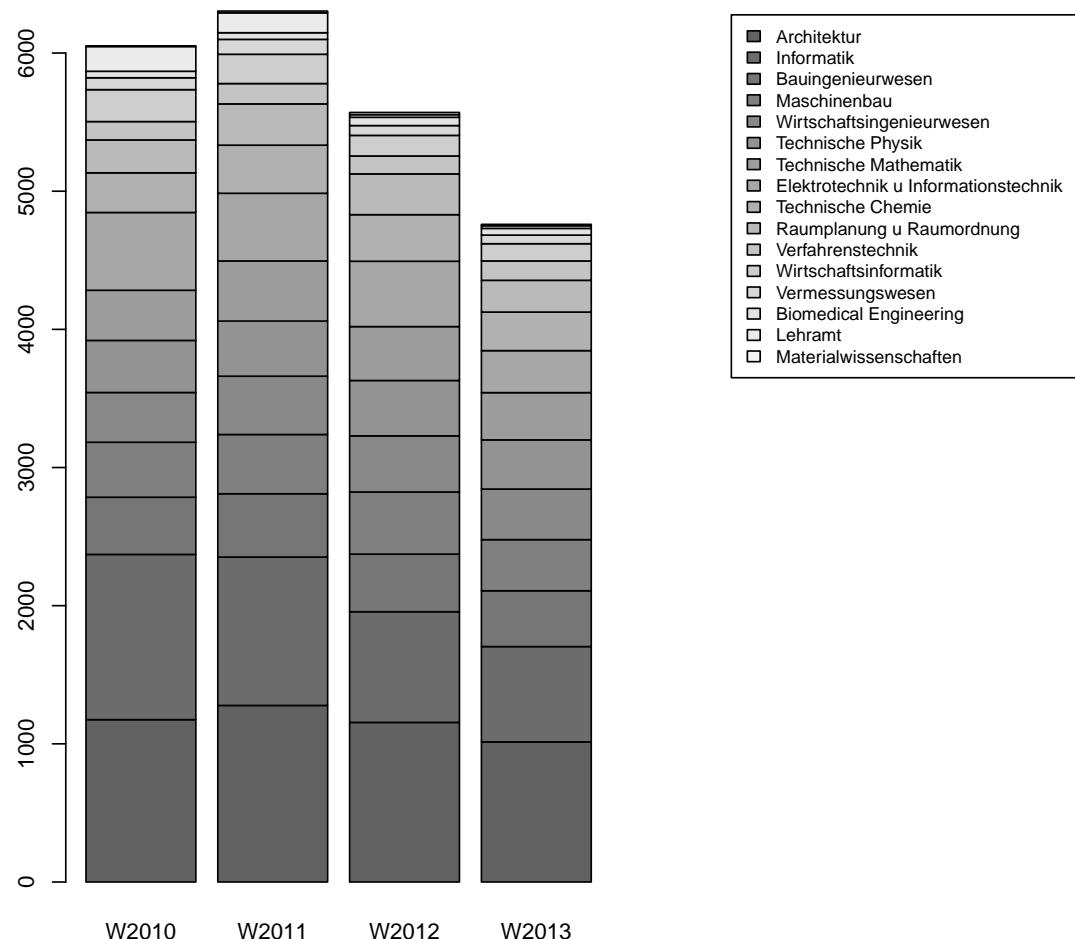
Abbildung 1.3: Neuinskriptionen an der TU–Wien im W2013 (Barplot)

Die Abb 1.5 zeigt den Mosaikplot der Neuinskriptionen für W2010 bis W2013. (Vgl. für die Codierung der Studienrichtungen Abb 1.3 oder den in Bsp 1.6 angegebenen R–Output.)

1.6.5 Pareto–Diagramm

Das **Pareto–Diagramm** ist eine Variante des Balkendiagramms, die vornehmlich im Qualitätsmanagement (aber auch in anderen Bereichen) als Entscheidungshilfe Verwendung findet. Gibt es z. B. mehrere Probleme mit einem (neuen) Produkt, wird man zweckmäßigerweise versuchen, zuerst die häufigsten (und/oder kostspieligsten) Defekte zu eliminieren.

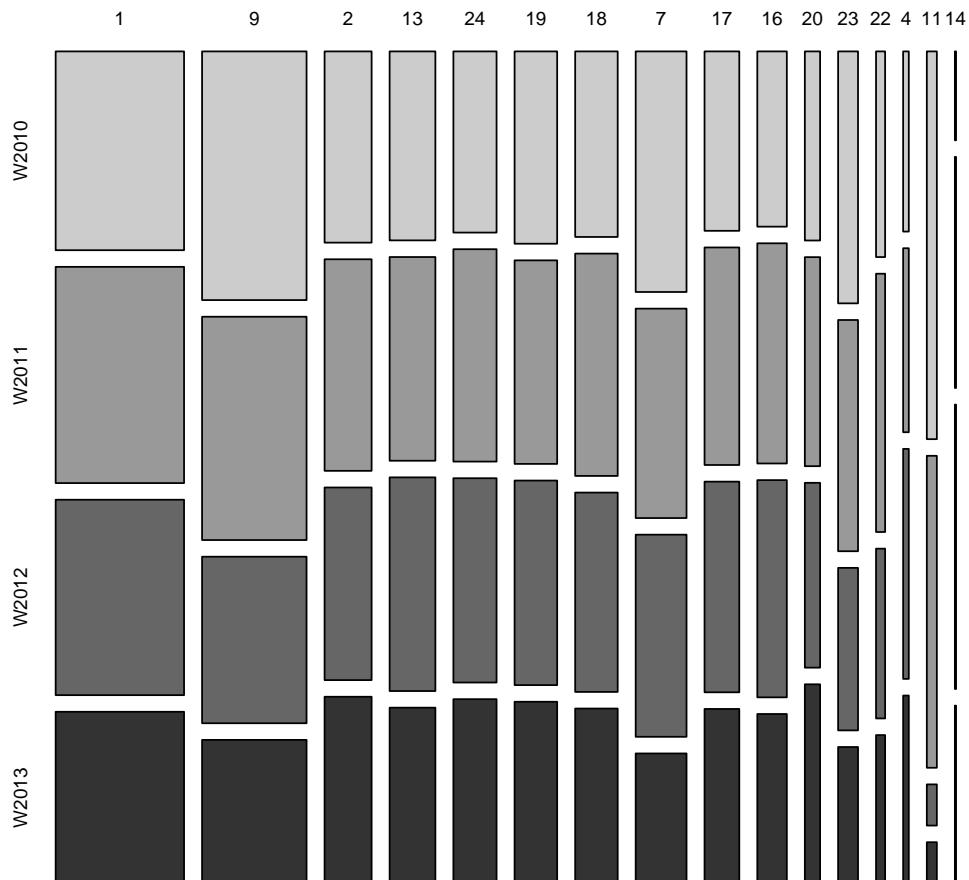
Bem: Benannt nach dem ital.-franz. Ökonomen und Soziologen VILFREDO F.D. PARETO (1848–1923), der erkannte, dass (bezogen auf Märkte) 80% des Geschehens auf 20% der Beteiligten entfällt. Dieses **Pareto–Prinzip** wird daher auch **80/20–Regel** genannt. Im Qualitätsmanagement lässt sich dieses Prinzip wie folgt formulieren: *80% of a problem is*

Abbildung 1.4: Neuinskriptionen an der TU–Wien für W2010 bis W2013 (Barplot)

caused by 20% of the causes, oder: *The rule of the vital few and the trivial (or useful) many.* Das Pareto-Diagramm gehört zu den sogenannten „Sieben Werkzeugen“ zur Verbesserung der Qualität (KAORU ISHIKAWA (1915–1989), japan. Qualitätspionier).

Bsp 1.7 Angenommen, bei 97 elektronischen Einheiten traten die in der 1. Spalte des folgenden R-Outputs angegebenen Defekte auf. Die Häufigkeiten stehen absteigend in der 2. Spalte. (Bem: Man beachte, dass bei einigen Einheiten mehrere Defekte auftraten, und daher die Summe der Häufigkeiten nicht gleich 97 ist.)

Die Abb 1.6 zeigt das zugehörige Pareto-Diagramm. Über den Balken wird das **Summenpolygon** gezeichnet, d. i. eine grafische Darstellung der in der 4. Spalte angegebenen *kumulierten* (relativen) Häufigkeiten.

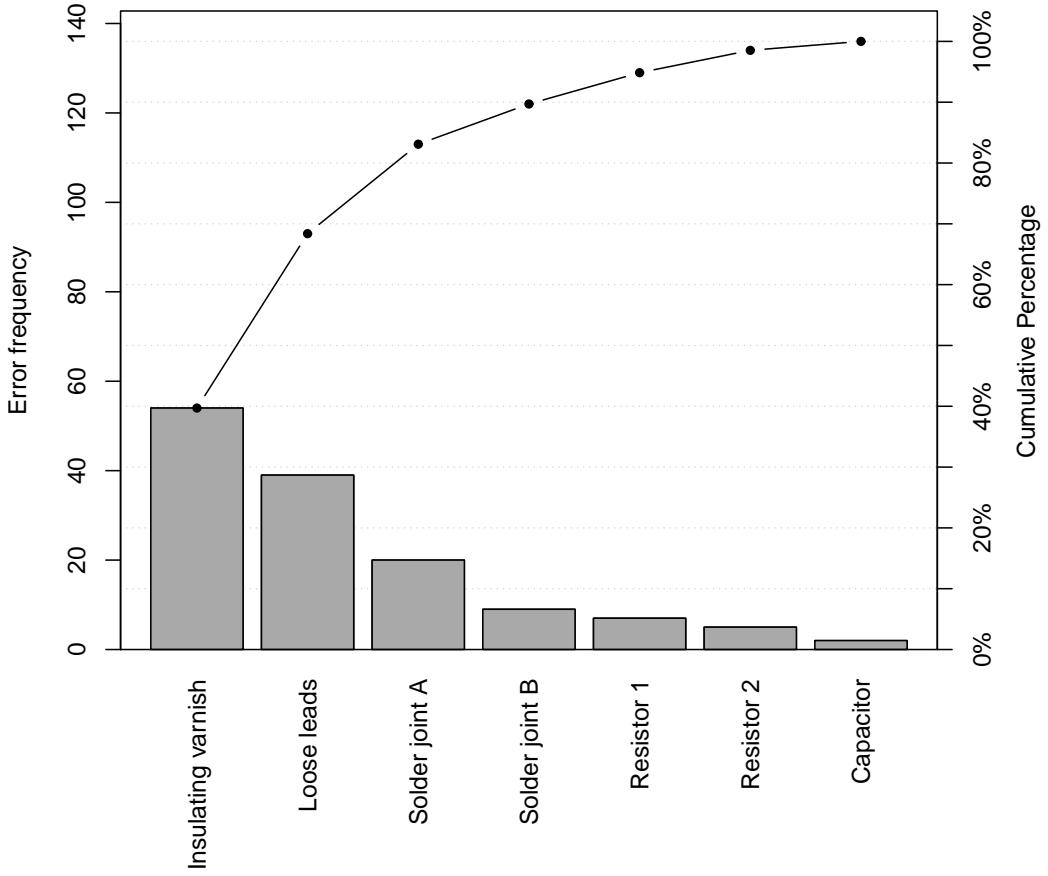
Abbildung 1.5: Neuinskriptionen an der TU–Wien für W2010 bis W2013 (Mosaikplot)

Pareto chart analysis for defect

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
Insulating varnish	54	54	39.71	39.7
Loose leads	39	93	28.68	68.4
Solder joint A	20	113	14.71	83.1
Solder joint B	9	122	6.62	89.7
Resistor 1	7	129	5.15	94.9
Resistor 2	5	134	3.68	98.5
Capacitor	2	136	1.47	100.0

Eine UE–Aufgabe beschäftigt sich mit der Anwendung des Pareto–Diagramms auf die schon mehrfach betrachteten Inskriptionszahlen.

Abbildung 1.6: Pareto-Diagramm (Bsp 1.7)



1.7 Stetige univariate Merkmale

In diesem Abschnitt betrachten wir verschiedene Darstellungsmöglichkeiten für Beobachtungen von stetigen Merkmalen. Da das Messniveau nun höher ist, hat man auch mehr Möglichkeiten als bei qualitativen Merkmalen.

1.7.1 Ordnungsstatistiken

Ein natürlicher erster Schritt in der Aufbereitung von metrischen (oder ordinalen) Merkmalen ist ihre Sortierung nach der Größe. Werden die n Beobachtungswerte eines Merkmals, die in der Reihenfolge ihrer Beobachtung, x_1, x_2, \dots, x_n , als Urliste vorliegen, nach aufsteigender Größe geordnet, entsteht die **Rangfolge**:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

Die $x_{(i)}$ nennt man die **Ordnungsstatistiken**. Sind alle Werte verschieden, bezeichnet

man die Nummer i in der obigen Anordnung als **Rangzahl**. Vielfach (z. B. als Folge einer nur beschränkten Messgenauigkeit) sind mehrere Beobachtungen identisch. Gilt:

$$x_{(i-1)} < x_{(i)} = x_{(i+1)} = \cdots = x_{(i+c)} < x_{(i+c+1)}$$

spricht man von einer **Bindung** vom Ausmaß $c + 1$ und teilt allen Werten von $x_{(i)}$ bis $x_{(i+c)}$ die mittlere Rangzahl $i + c/2$ zu.

Bsp 1.8 Angenommen, die Urliste des Umfangs $n = 10$ ist gegeben wie folgt:

$$0.15 \quad -0.84 \quad -0.83 \quad 0.15 \quad -0.50 \quad -1.62 \quad -0.52 \quad 0.49 \quad 0.08 \quad -0.66$$

Es gibt eine Bindung vom Ausmaß 2 (bei 0.15); die Rangzahlen lauten daher:

$$8.5 \quad 2 \quad 3 \quad 8.5 \quad 6 \quad 1 \quad 5 \quad 10 \quad 7 \quad 4$$

■

Rangtransformation: Wird jede Beobachtung durch ihre Rangzahl (unter Verwendung der obigen Regel bei Bindungen) ersetzt, spricht man von der **Rangtransformation**. Dadurch verzichtet man auf einen Teil der in den ursprünglichen Daten enthaltenen (metrischen) Information und verwendet für weitere Berechnungen nur mehr die relative Position jeder Beobachtung innerhalb des Datensatzes.

Nichtparametrische Statistik: Ordnungsstatistiken (und die Rangtransformation) spielen generell eine große Rolle in der Statistik, insbesondere aber in der sogenannten **nichtparametrischen** Statistik. Bei diesem Zweig der Statistik versucht man mit nur ganz wenigen Voraussetzungen hinsichtlich des zugrunde liegenden statistischen Modells auszukommen.

1.7.2 Empirische Verteilungsfunktion

Eine Funktion von grundlegender Bedeutung in der Statistik ist die **empirische Verteilungsfunktion**, definiert für $x \in \mathbb{R}$ durch:

$$\widehat{F}_n(x) = \begin{cases} 0 & \text{für } x < x_{(1)} \\ \frac{i}{n} & \text{für } x_{(i)} \leq x < x_{(i+1)}, \quad i = 1, 2, \dots, n-1 \\ 1 & \text{für } x_{(n)} \leq x \end{cases}$$

Äquivalente Definition:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(x_i), \quad x \in \mathbb{R}$$

Dabei bezeichnet $I_A(x)$ die **Indikatorfunktion** der Menge A ($A \subseteq \mathbb{R}$):

$$I_A(x) = \begin{cases} 1 & \text{für } x \in A \\ 0 & \text{sonst} \end{cases}$$

$\hat{F}_n(x)$ ist also eine **Treppenfunktion** mit Sprüngen an den Stellen $x_{(i)}$ der Höhe $1/n$ (oder der Höhe c/n , falls es bei $x_{(i)}$ eine Bindung vom Ausmaß c gibt).

Bem: Bei der grafischen Darstellung von \hat{F}_n zeichnet man aus optischen Gründen meist die Stufen aus (vgl. Abb 1.7), gültig sind aber bei Sprüngen jeweils nur die *oberen* Punkte.

Bsp 1.9 Als Beispiel für eine empirische Verteilungsfunktion betrachten wir aus dem Datensatz `body.txt` (vgl. Bsp 1.5) die Variable `Biacromial` (= Schulterbreite), für beide Geschlechter zusammen und getrennt nach Geschlecht, dargestellt in einem Plot (Abb 1.7). Als Folge der beschränkten Messgenauigkeit gibt es hier zahlreiche Bindungen. ■

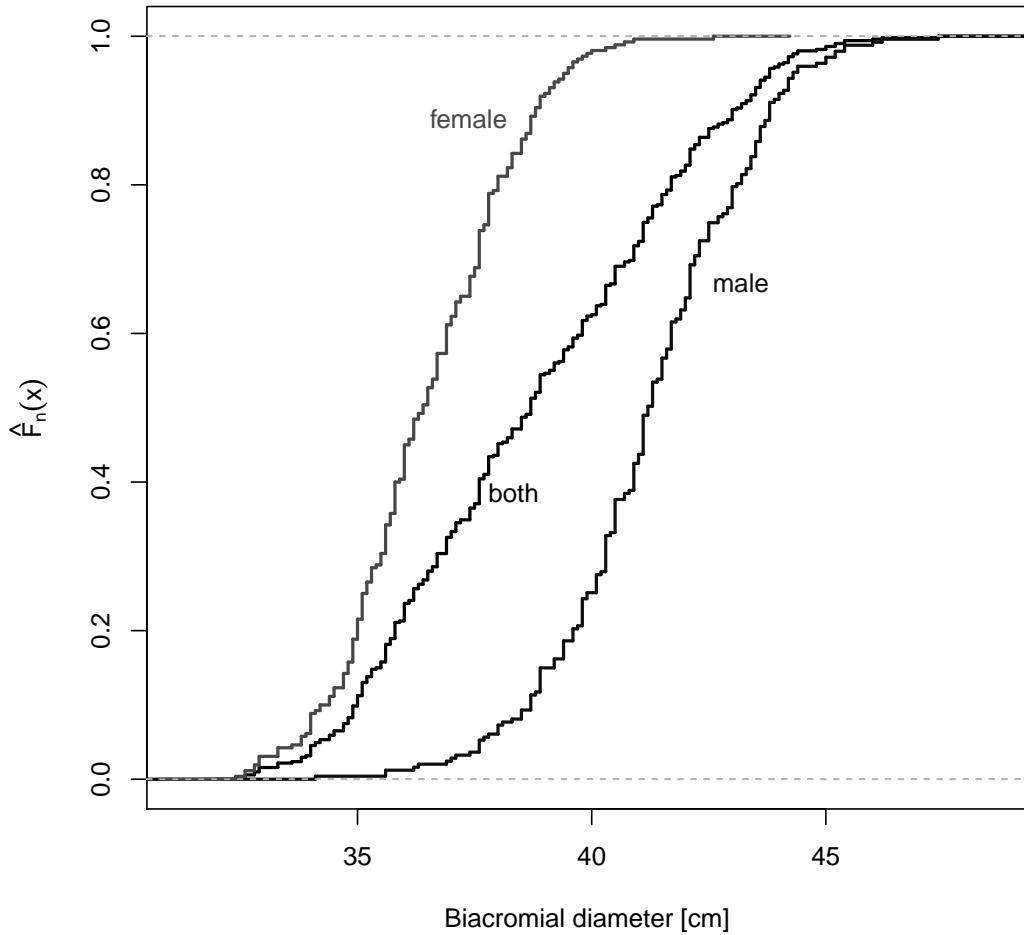
1.7.3 Stem-and-Leaf-Plot

Eine einfache – bei kleineren Datensätzen auch von Hand durchführbare – typografische Darstellung der Daten ist der **Stem-and-Leaf-Plot** (**Stamm-und-Blatt-Darstellung**). Dabei werden die Werte direkt der Größe nach wiedergegeben, wobei die vorderen Dezimalstellen den *Stamm* und die hinteren die *Blätter* bilden. (Vorher werden die Daten auf eine entsprechende Stellenzahl abgeschnitten, nicht gerundet.) Verschiedene Auflösungen (üblich sind 1-, 2- und 5-fache) sind möglich, ihre Sinnhaftigkeit ist aber situationsabhängig.

Bsp 1.10 Die Abb 1.8 zeigt den Stem-and-Leaf-Plot für die Variable `Biacromial` (für `Gender = 1`). In der mittleren Spalte stehen die Stämme, rechts davon die zugehörigen Blätter. Damit der Plot nach rechts hin nicht zu ausladend wird, nehmen wir eine 2-fache Auflösung. Beispielsweise repräsentiert der Eintrag `36|23` die Werte 36.2 und 36.3.

In der ersten – von Hand ergänzten – Spalte stehen die kumulierten Anzahlen der Blätter, von den beiden Enden her betrachtet. Die Bezeichnung (28) bedeutet, dass dieser Stamm 28 Blätter hat. Auf diese Weise lässt sich der Median (s. unten) leichter bestimmen (hier gilt Median = $x_{(124)} = 41.2$). ■

Abbildung 1.7: Empirische Verteilungsfunktion für die Schulterbreite (Biacromial)



1.7.4 Klassierung

Bei größeren Stichprobenumfängen (ab etwa 30) ist eine **Klassenbildung** sinnvoll. Letztere ist naturgemäß nicht eindeutig festgelegt. Hinsichtlich der Anzahl und Breite der **Klassen** (oder **Bins**) haben sich verschiedene Regeln herausgebildet, wobei aber keine in jeder Situation allen anderen überlegen ist. In jedem Fall ist aber darauf zu achten, dass der gesamte Wertebereich (ohne Lücken) überdeckt wird und jede Beobachtung eindeutig einer Klasse zugeordnet werden kann. Üblicherweise nimmt man links offene und rechts abgeschlossene Klassen, also Klassen der Form $(a, b]$. Beispielsweise kann man sich an die folgenden Regeln halten:

- (1) Bestimme zunächst den kleinsten $x_{(1)}$ und größten Wert $x_{(n)}$ der Stichprobe, sowie die **Spannweite** $R = x_{(n)} - x_{(1)}$.
- (2) In der Praxis sind alle Beobachtungen gerundete (oder abgeschnittene) Zahlen. Ist der kleinste Wert beispielsweise 69.6, so steht er für einen Messwert zwischen 69.55 und 69.65. Als unteren Rand der ersten Klasse kann man daher 69.55 nehmen.

Abbildung 1.8: Stem-and-Leaf–Plot für die Schulterbreite (Biacromial)

1 34 1
1 34
1 35
3 35 66
5 36 23
6 36 9
9 37 014
15 37 666678
20 38 00013
37 38 5557777899999999
46 39 222444444
62 39 666678888888899
82 40 1111112333333333334
105 40 5555555555577899999999
(28) 41 0001111111111122233333334
114 41 5555555666777777778999
91 42 0000111111111122233333
68 42 55555577899
57 43 00000001222334444
40 43 5555566666778888889
21 44 00122223344
10 44 8
9 45 002244
3 45
3 46 02
1 46
1 47 4

- (3) Falls die Verteilung nicht sehr schief ist, sind Klassierungen mit **äquidistanten** Klassenbreiten w (gerundet auf die gleiche Genauigkeit wie die Messwerte) zu bevorzugen. Eine grobe Regel besagt:

$$w = \begin{cases} \frac{R}{\sqrt{n}} & \text{falls } 30 < n \leq 400 \\ \frac{R}{20} & \text{falls } n > 400 \end{cases}$$

Eine andere gängige Regel (*Sturges' Rule*) besagt: Nimm a Klassen, wobei $2^{a-1} < n \leq 2^a$. D. h., nimm etwa $\log_2(n)$ (Logarithmus zur Basis 2) Klassen gleicher Breite.

Bem: Man beachte, dass der Stem-and-Leaf–Plot quasi eine auf den Daten selbst basierende Klassierung der Daten vornimmt.

Bsp 1.11 Standardmäßig verwendet R die Sturges–Regel. Für die Variable `Biacromial` (für `Gender = 1`) mit $n = 247$ Beobachtungen, ergibt sich die folgende Klasseneinteilung:

(34,36]	(36,38]	(38,40]	(40,42]	(42,44]	(44,46]	(46,48]	
3	15	44	98	68	17	2	<-- abs. Häufigk.

Nach der Sturges–Regel ($2^7 = 128 < 247 \leq 2^8 = 256$) sind etwa acht Klassen zu nehmen, tatsächlich sind es nur sieben. Hinsichtlich der Klassenbegrenzungen hält sich R nicht an die oben formulierte Regel, sondern versucht möglichst einfache und „glatte“ Zahlen zu finden. ■

1.7.5 Histogramm

Ein **Histogramm** ist eine grafische Darstellung einer (relativen) Häufigkeitsverteilung, basierend auf einer vorherigen Klassierung der Daten. Dabei sollte man sich – zwingend wenn man eine nicht äquidistante Klassierung verwendet oder wenn man mehrere Häufigkeitsverteilungen miteinander vergleichen möchte – an das folgende Prinzip halten:

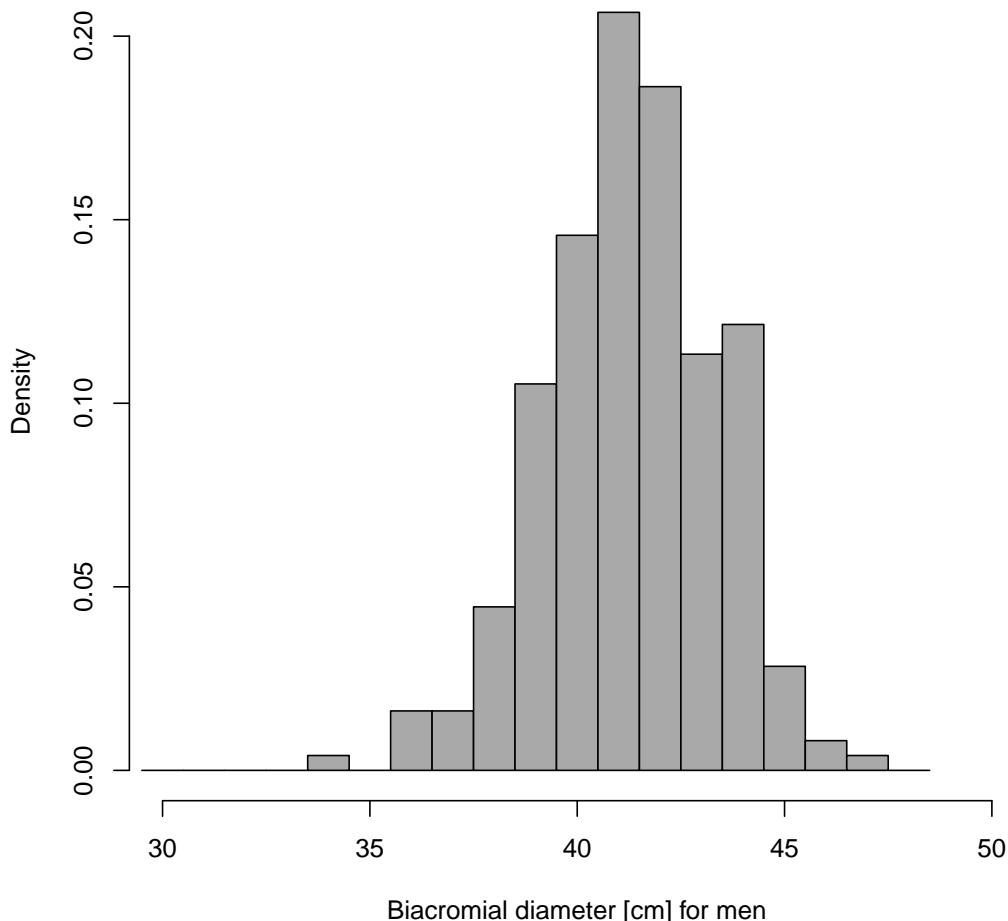
Prinzip der Flächentreue: Zeichne über den k Klassen Rechtecke mit den Höhen:

$$h_i = \frac{f_i}{w_i}, \quad i = 1, 2, \dots, k$$

Dabei bezeichnet f_i die relative Häufigkeit und w_i die Breite der i -ten Klasse. Das so gezeichnete Histogramm nennt man ein **flächentreues Histogramm** (oder ein **Dichtehistogramm**), da unabhängig von der Klasseneinteilung die Summe der Rechtecksflächen genau Eins beträgt:

$$\text{Fläche des Histogramms} = \sum_{i=1}^k h_i w_i = \sum_{i=1}^k f_i = 1$$

Bem: Klasseneinteilungen sind nicht eindeutig bestimmt, daher kann auch das Erscheinungsbild eines Histogramms, abhängig von der verwendeten Klasseneinteilung, u. U. beträchtlich variieren. M. a. W., Histogramme sind nicht „robust“ bezüglich der Klasseneinteilung. Dem trägt man meist dadurch Rechnung, dass man die Klasseneinteilung variiert (Zahl der Klassen, Klassenbreite, Anfangspunkt der Klasseneinteilung) und jenes bevorzugt, das die wenigsten *unechten* „Täler“ und „Gipfel“ aufweist, aber dennoch die Struktur des Datensatzes gut erkennen lässt. Das ist natürlich eine subjektive Entscheidung, bei der man aber auch sonstige Informationen über den Datensatz berücksichtigen sollte.

Abbildung 1.9: Histogramm für die Schulterbreite (Biacromial)

Bsp 1.12 Als Beispiel für ein (Dichte-) Histogramm betrachten wir wieder die Schulterbreite für **Gender** = 1. Wir nehmen eine äquidistante Klasseneinteilung mit links abgeschlossenen und rechts offenen Intervallen wie folgt:

$$\text{Klassen: } [29.5, 30.5), [30.5, 31.5), \dots, [47.5, 48.5)$$

(Bem: Die Klasseneinteilung ist so ausgelegt, dass sie auch für **Gender** = 0 verwendet werden kann und so einen direkten Vergleich der beiden Geschlechter hinsichtlich dieses Merkmals gestattet.) Die Gesamtfläche der grauen Rechtecke in Abb 1.9 ist Eins. Bezeichnet $\hat{f}(x)$ die Funktion, die jedem $x \in \mathbb{R}$ die Höhe des entsprechenden Rechtecks zuordnet, so gilt:

$$\int_{-\infty}^{\infty} \hat{f}(x) dx = 1$$

$\widehat{f}(x)$ ist also eine **Dichte** (vgl. 3.2.2); das erklärt die Bezeichnung *Dichtehistogramm* für ein flächentreues Histogramm. Letzteres lässt sich somit wie folgt interpretieren: Die Rechtecksfläche repräsentiert die relative (Klassen-) Häufigkeit und die Rechteckshöhe repräsentiert die Dichte der Daten.

Ein Vergleich des Histogramms mit Abb 1.8 zeigt, dass der Stem-and-Leaf–Plot quasi ein auf die Seite gelegtes Histogramm ist. ■

1.7.6 Kernschätzung

Die einem Histogramm zugrunde liegende Klasseneinteilung besteht gewissermaßen aus „Fenstern“, durch die man auf die Daten blickt. Nur diejenigen x_i , die im jeweiligen Fenster sichtbar sind, liefern einen Beitrag zur „Dichte“.

Diese Vorstellung lässt sich dahingehend verallgemeinern, dass man als Fenster nicht eine feste Klasseneinteilung nimmt, sondern jedem x_i quasi ein eigenes Fenster zuordnet. Dies führt zum Konzept der **Kerndichteschätzung**.

Dabei versteht man unter einer **Kernfunktion** (oder kurz **Kern**) eine (meist) symmetrische Funktion um Null, deren Fläche Eins ist. Häufig verwendete Kerne (vgl. Abb 1.10 für eine vergleichende grafische Darstellung):

$$\text{Rechteckskern: } K(z) = \frac{1}{2} I_{[-1,1]}(z)$$

$$\text{Dreieckskern: } K(z) = (1 - |z|) I_{[-1,1]}(z)$$

$$\text{Normalkern: } K(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}$$

$$\text{Epanechnikov–Kern: } K(z) = \frac{3}{4} (1 - z^2) I_{[-1,1]}(z)$$

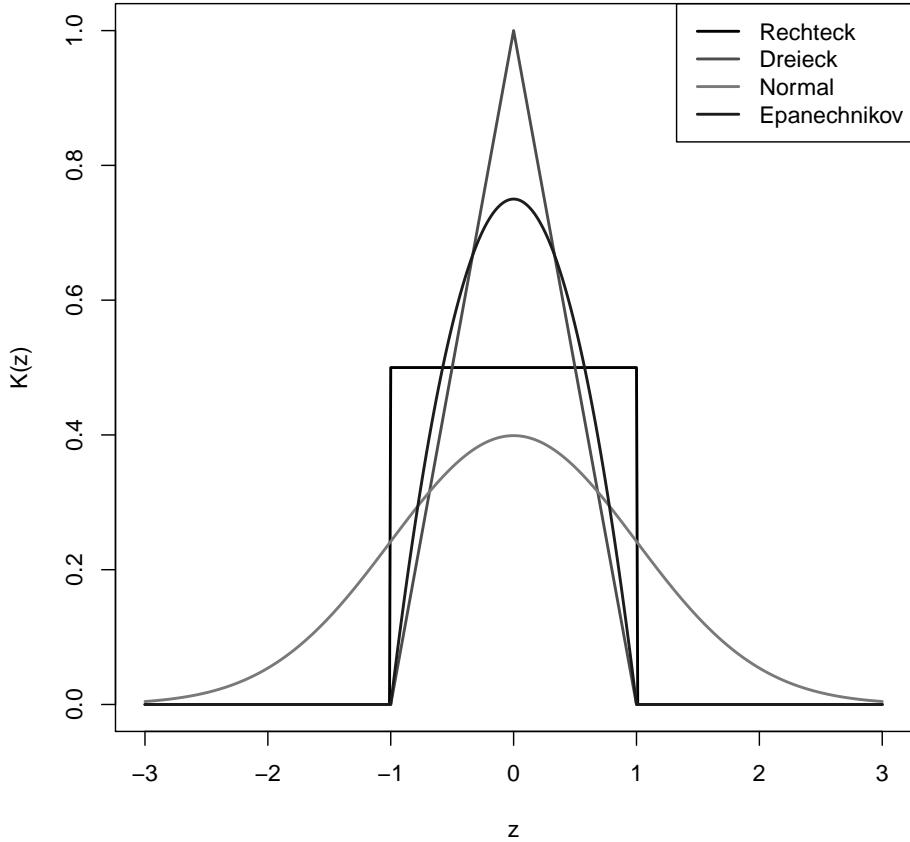
Die **Kerndichteschätzung** ist definiert durch:

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad x \in \mathbb{R}$$

Dabei ist $h > 0$ die sogenannte **Bandbreite**.⁵ Die Kernschätzung „erbt“ die Eigenschaften der Kernfunktion K . Insbesondere gilt, dass $\widehat{f}(x)$ eine stetige und auch hinlänglich „glatte“ Funktion ist. Im Gegensatz dazu sind Histogramme „stufige“ Funktionen. (Letzteres ist aber für *stetige* Merkmale meist nicht erwünscht.)

⁵engl. *bandwidth*

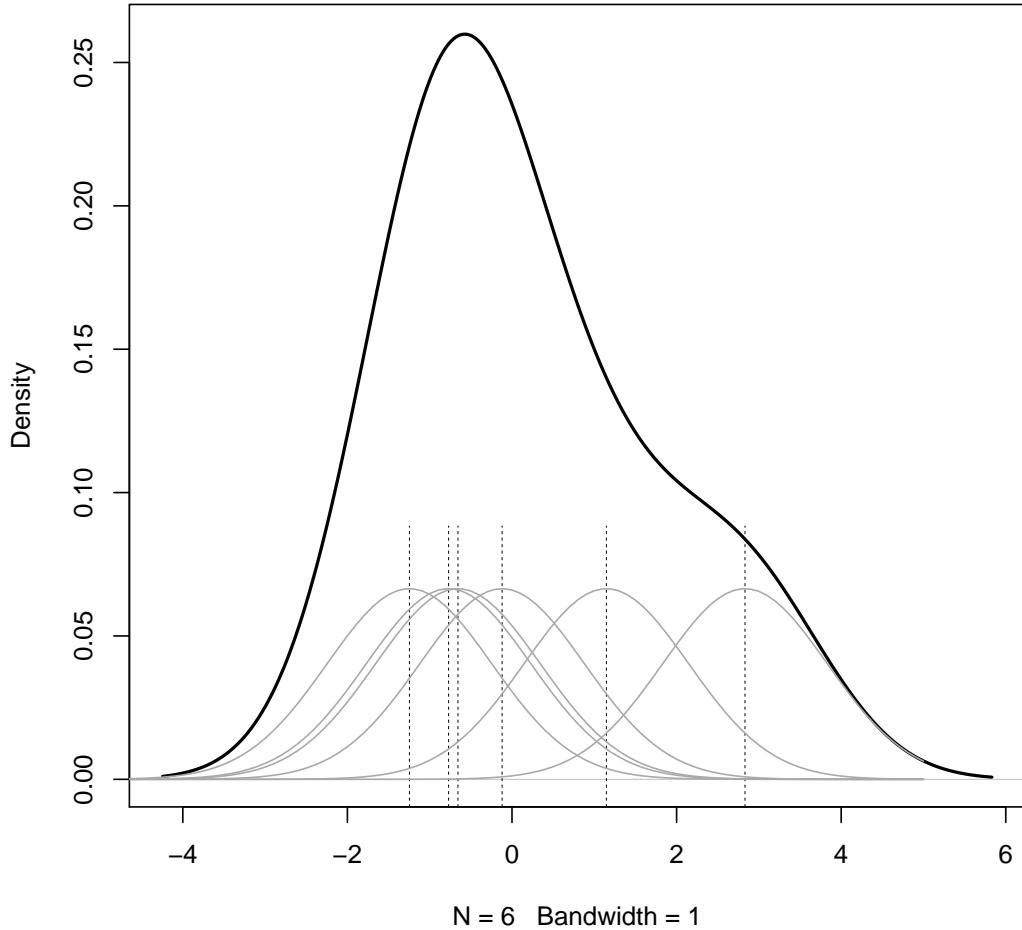
Abbildung 1.10: Gebräuchliche Kernfunktionen



Die Fläche unter $\hat{f}(x)$ ist Eins:

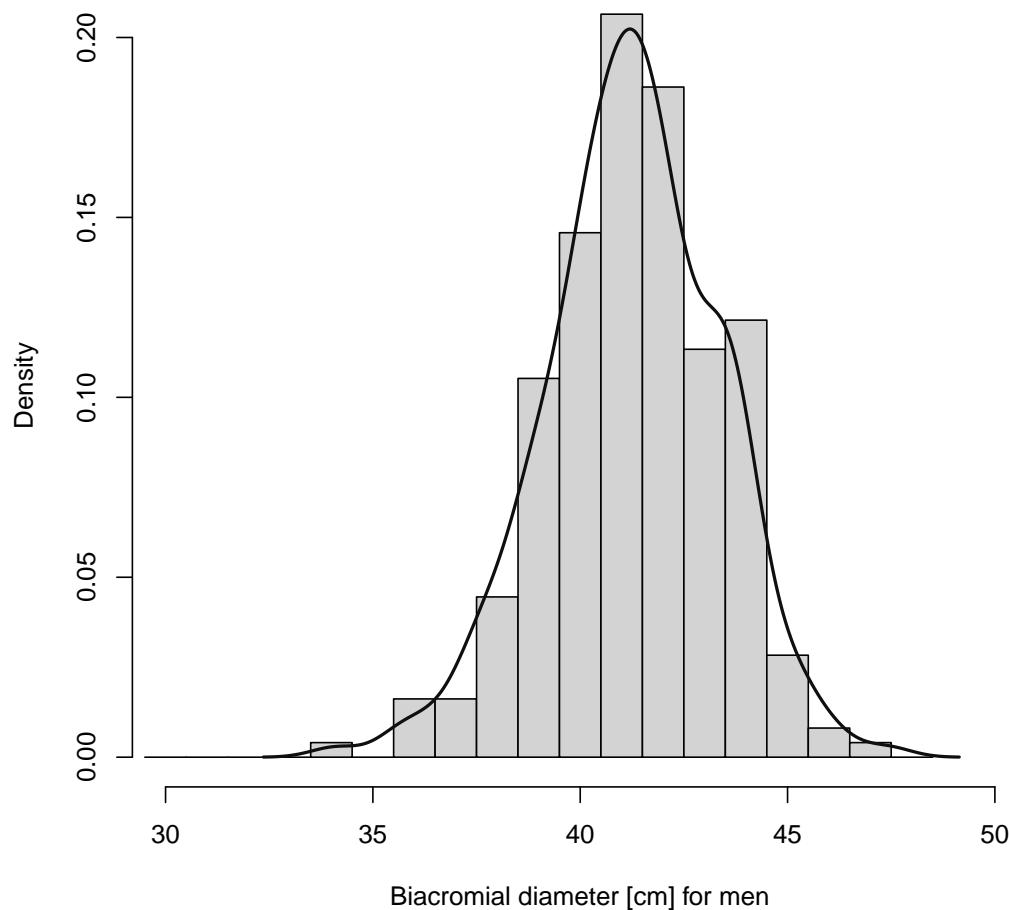
$$\begin{aligned} \int_{-\infty}^{\infty} \hat{f}(x) dx &= \frac{1}{nh} \sum_{i=1}^n \int_{-\infty}^{\infty} K\left(\frac{x-x_i}{h}\right) dx \quad \left[\text{Subst.: } \frac{x-x_i}{h} = z, \, dx = h dz \right] \\ &= \frac{1}{n} \sum_{i=1}^n \underbrace{\int_{-\infty}^{\infty} K(z) dz}_{=1} = 1 \end{aligned}$$

Zur Wahl der Bandbreite: Die Bandbreite h bestimmt die „Glattheit“ der Kernschätzung: Je größer h , umso „träger“ reagiert die Schätzung auf die einzelnen Beobachtungen. Wie sich zeigt, ist die Wahl der Bandbreite kritischer als die Wahl der Kernfunktion. Hier besteht ein ähnliches Dilemma wie beim Histogramm mit seiner Empfindlichkeit gegenüber der Klasseneinteilung. Auch hier gibt es eine Reihe von *Faustregeln* zur Wahl der Bandbreite; als pragmatische Lösung empfiehlt sich das Ausprobieren mehrerer h -Werte.

Abbildung 1.11: Prinzip der Kerndichteschätzung

Bsp 1.13 Die Abb 1.11 zeigt anhand eines einfachen Beispiels, wie die Kerndichteschätzung durch Überlagerung (= Summierung) aus den den einzelnen Beobachtungen – entsprechen den strichlierten vertikalen Linien – zugeordneten Kernfunktionen aufgebaut wird. Als Kernfunktion nehmen wir den Normalkern mit einer Bandbreite von $h = 1$. Man beachte, dass in diesem Fall – auf Basis von nur sechs Beobachtungen – die Konstruktion eines Histogramms nicht möglich wäre. ■

Bsp 1.14 Wir betrachten wieder die Schulterbreite für `Gender = 1` und überlagern das bereits in Abb 1.9 dargestellte Histogramm mit einer Kerndichteschätzung (Abb 1.12). Die R-Funktion `density()` nimmt standardmäßig den Normalkern und eine für *normalverteilte* Beobachtungen optimierte Regel für die Wahl der Bandbreite (*Silverman's Rule*). Im vorliegenden Fall folgen die Daten näherungsweise einer für die Normalverteilung typischen „Glockenkurve“, sodass diese Regel anwendbar ist. Hier ergibt sich eine Bandbreite von $h \approx 0.53$. ■

Abbildung 1.12: Kerndichteschätzung und Histogramm für die Schulterbreite (Biacromial)

1.7.7 Quantile

Empirische (d. h. auf Daten basierende) Quantile werden in der Literatur nicht einheitlich definiert. Grob gesprochen handelt es sich bei einem p -**Quantil** – wobei $0 \leq p \leq 1$ – um einen Wert x_p , der den Datensatz (etwa) im Verhältnis $p : (1 - p)$ teilt. Sind x_1, x_2, \dots, x_n die beobachteten Daten, so gilt:

$$\frac{\text{Anzahl}\{x_i \leq x_p\}}{n} \approx p$$

Die verschiedenen Definitionen lassen sich danach einteilen, ob für x_p nur beobachtete Datenwerte zugelassen sind oder auch Werte dazwischen. In R werden insgesamt neun verschiedene Definitionen (oder Typen) unterschieden. Wir behandeln im Folgenden die Typen 1, 2, 4 und 7 etwas genauer.⁶

⁶Vgl. für eine ausführliche Diskussion ROB J. HYNDMAN and YANAN FAN: Sample Quantiles in Statistical Packages, *The American Statistician*, 50/4, 1996.

Typ 1: Dieser Typ bezieht sich auf die empirische Verteilungsfunktion \widehat{F}_n (vgl. 1.7.2) und ist wie folgt definiert:

$$x_p = \min \{x \in \mathbb{R} : \widehat{F}_n(x) \geq p\}$$

Das so definierte x_p entspricht stets einem Wert aus dem Datensatz.

Bem: Diese Definition entspricht der *verallgemeinerten Inversen* der empirischen Verteilungsfunktion; „verallgemeinert“ deshalb, weil \widehat{F}_n als Treppenfunktion im strengen Sinn nicht invertierbar ist.

Typ 2: Wie Typ 1, allerdings wird bei Unstetigkeiten gemittelt, d. h. auch Werte genau in der Mitte zwischen zwei Datenpunkten sind möglich.

Typ 4: Alle Werte im Intervall $[x_{(1)}, x_{(n)}]$ sind zugelassen und man definiert:

$$x_p = \begin{cases} x_{(1)} & \text{falls } 0 < p \leq \frac{1}{n} \\ x_{(i)} + (np - i)(x_{(i+1)} - x_{(i)}) & \text{falls } \frac{i}{n} < p \leq \frac{i+1}{n}, \quad i = 1, 2, \dots, n-1 \end{cases}$$

Dies entspricht einer *linearen Interpolation* der empirischen Verteilungsfunktion.

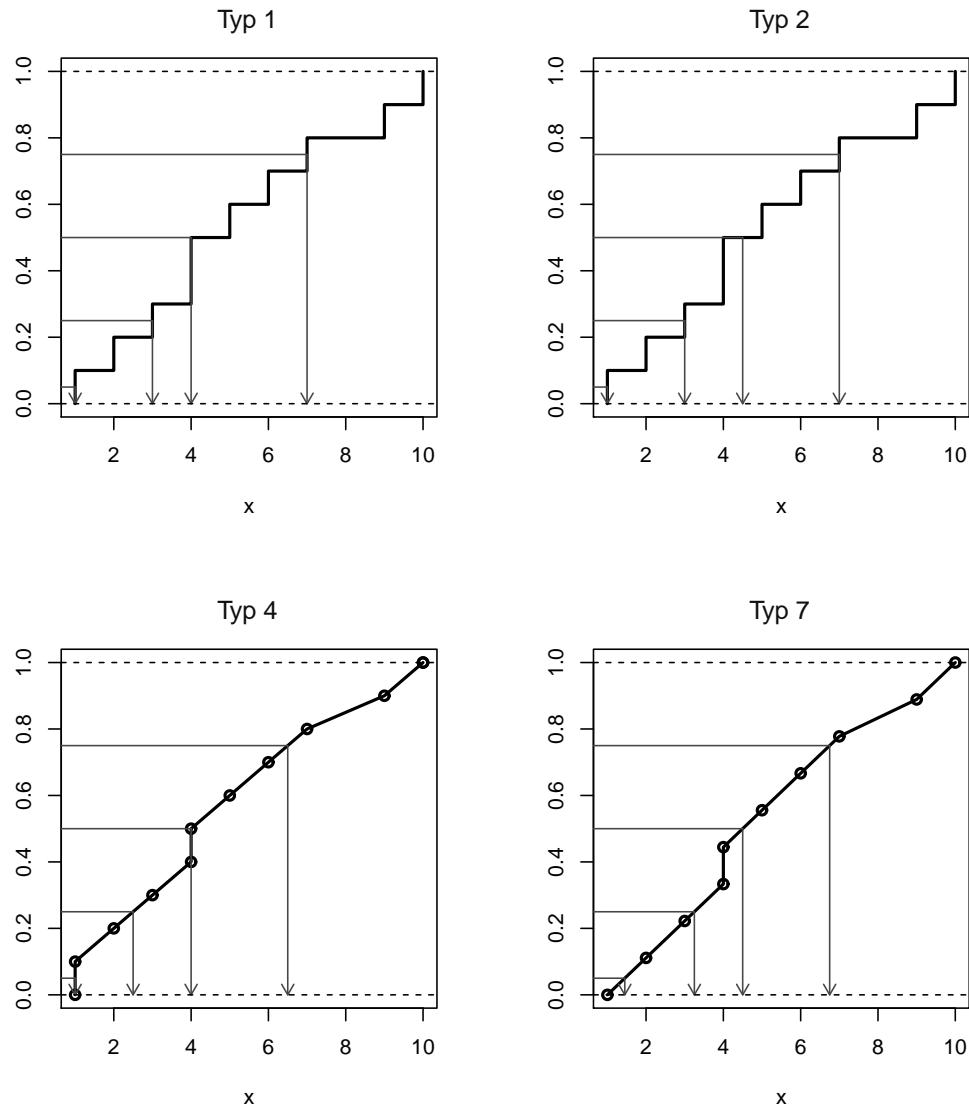
Typ 7: Ähnlich wie Typ 4, allerdings wird bei Typ 7 das Intervall $[0, 1]$ in $n-1$ Teilintervalle (Typ 4: n Teilintervalle) zerlegt (d. h., $x_{(1)}$ entspricht dann dem 0%- und $x_{(n)}$ dem 100%-Quantil). Das ist die von der R-Funktion `quantile()` standardmäßig verwendete Definition.

Einige Quantile sind von besonderer Bedeutung:

Median: Der **Median** ist das 50%-Quantil. Er wird meist mit \tilde{x} bezeichnet und teilt den Datensatz (etwa) in zwei gleich große Hälften. Der Median wird (üblicherweise) einheitlich wie folgt definiert:

$$\tilde{x} = \begin{cases} x_{(k+1)} & \text{falls } n = 2k+1 \quad (\text{d. h. } n \text{ ungerade}) \\ \frac{x_{(k)} + x_{(k+1)}}{2} & \text{falls } n = 2k \quad (\text{d. h. } n \text{ gerade}) \end{cases}$$

Quartile: Die **Quartile** teilen den Datensatz in (etwa) vier gleich große Stücke: $Q_1 = x_{1/4}$ (= 1. Quartil), $Q_2 = x_{1/2}$ (= 2. Quartil = Median), $Q_3 = x_{3/4}$ (= 3. Quartil). Zwischen dem 1. und 3. Quartil liegen die mittleren 50% der Daten.

Abbildung 1.13: Quantilbestimmung mit `quantile()`

Bsp 1.15 In Abb 1.13 ist beispielhaft die Bestimmung von einigen Quantilen (5%, 25%, 50%, 75%) der hier betrachteten Typen für den Datensatz $\{3, 1, 7, 2, 4, 5, 4, 10, 6, 9\}$ dargestellt. Die Unterschiede sind gering und (meist) nur von untergeordneter Bedeutung. Bei großen Stichproben liefern die verschiedenen Definitionen nahezu identische Ergebnisse. ■

Hinges:⁷ Der **untere Hinge** ist der Median der ersten Hälfte der (geordneten) Daten, der **obere Hinge** ist der Median der zweiten Hälfte. Bei ungerader Anzahl von Daten zählt der Median zu beiden Hälften. Die Hinges entsprechen dem 1. und 3. Quartil, sind aber einfacher und schneller zu bestimmen.

⁷hinge engl. = Türangel, Drehachse, Gelenk

Bsp 1.16 Die im vorigen Beispiel betrachteten Daten lauten geordnet 1, 2, 3, 4, 4, 4, 5, 6, 7, 9, 10. Der Median ist $\tilde{x} = 4.5$ und die Hinges werden wie folgt bestimmt:

$$\begin{array}{ccccccccc} 1 & 2 & 3 & \underbrace{4 & 4} & 5 & 6 & 7 & 9 & 10 \\ & & & \text{med} = 3 & & & & \text{med} = 7 & \end{array}$$

Der untere Hinge ist also 3 und der obere Hinge ist 7. Die standardmäßig von R berechneten Quartile (Typ 7) sind $Q_1 = 3.25$, $Q_2 = 4.50$ (Median) und $Q_3 = 6.75$. ■

1.7.8 QQ–Plot

Der **Quantilen-Quantilen–Plot** (oder kurz **QQ–Plot**) ist eine Art Streudiagramm zum grafischen Vergleich zweier Datensätze oder zum Vergleich eines Datensatzes mit einer Referenzverteilung. (Bem: Letztere Anwendung wird in 7.4.11 behandelt.) Ist im ersten Fall die Größe der beiden Datensätze identisch, so zeichnet man einfach die beiden geordneten Stichproben gegeneinander:

$$(x_{(i)}, y_{(i)}), \quad i = 1, 2, \dots, n$$

Sind die Stichprobengrößen unterschiedlich, muss man die Datensätze einander angeleichen. Üblicherweise geht man dabei so vor, dass der größere Datensatz reduziert wird. Man behält Minimum und Maximum und wählt gleichmäßig aufgeteilte (empirische) Quantile dazwischen.

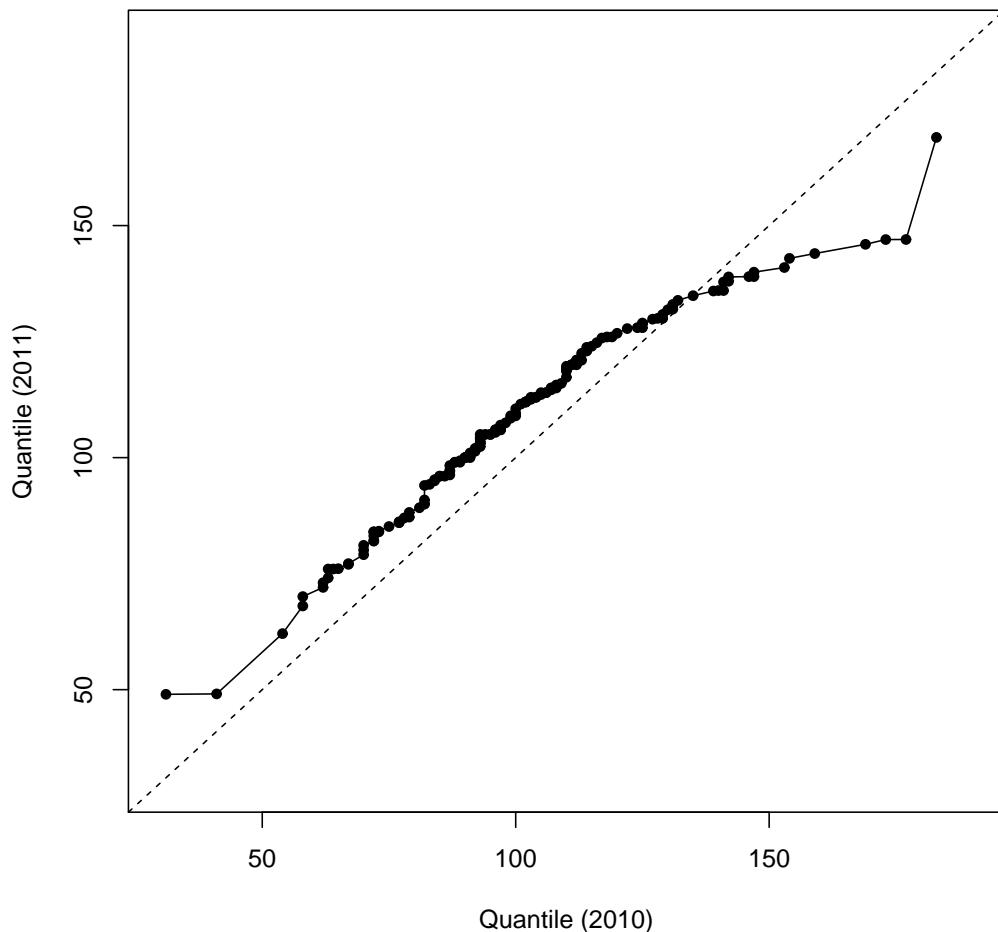
Bsp 1.17 Hat beispielsweise der x –Datensatz 5 Werte, der y –Datensatz aber 20 Werte, so zeichnet man die geordneten x –Werte gegen Minimum, 1. Quartil, Median, 3. Quartil und Maximum der y –Werte. ■

Liegen die Punkte annähernd auf einer Geraden – beispielsweise auf einer „robusten“ Ausgleichsgeraden durch das 1. und 3. Quartil der Punkte – so haben die beiden Verteilungen eine ähnliche *Form* (unterscheiden sich aber möglicherweise hinsichtlich Lage und/oder Streuung). Je nach Anwendung können aber auch andere Geraden sinnvoll sein, beispielsweise eine 45° Gerade durch den Nullpunkt (vgl. das folgende Beispiel). Liegen die Punkte annähernd auf dieser Geraden, besteht kein Unterschied zwischen den Verteilungen.

Bsp 1.18 Der QQ–Plot in Abb 1.14 vergleicht die Ozonwerte (maximale Einstundenmittelwerte) für Illmitz von Mai bis September für 2010 und 2011.⁸ Zusätzlich wurde zum einfacheren Vergleich die 45° Gerade eingezeichnet. Bis auf die hohen Ozonwerte waren die Messwerte im betrachteten Zeitraum 2011 höher als 2010. ■

⁸Die Daten stammen vom UMWELTBUNDESAMT. Die Messstelle Illmitz (Burgenland/Seewinkel) gehört – zusammen mit einigen anderen Messstellen – zu einem europaweiten Messnetz zur Erfassung des großräumigem Luftschadstofftransports.

Abbildung 1.14: QQ-Plot für die Ozonwerte (Messstelle Illmitz)



1.7.9 Boxplot

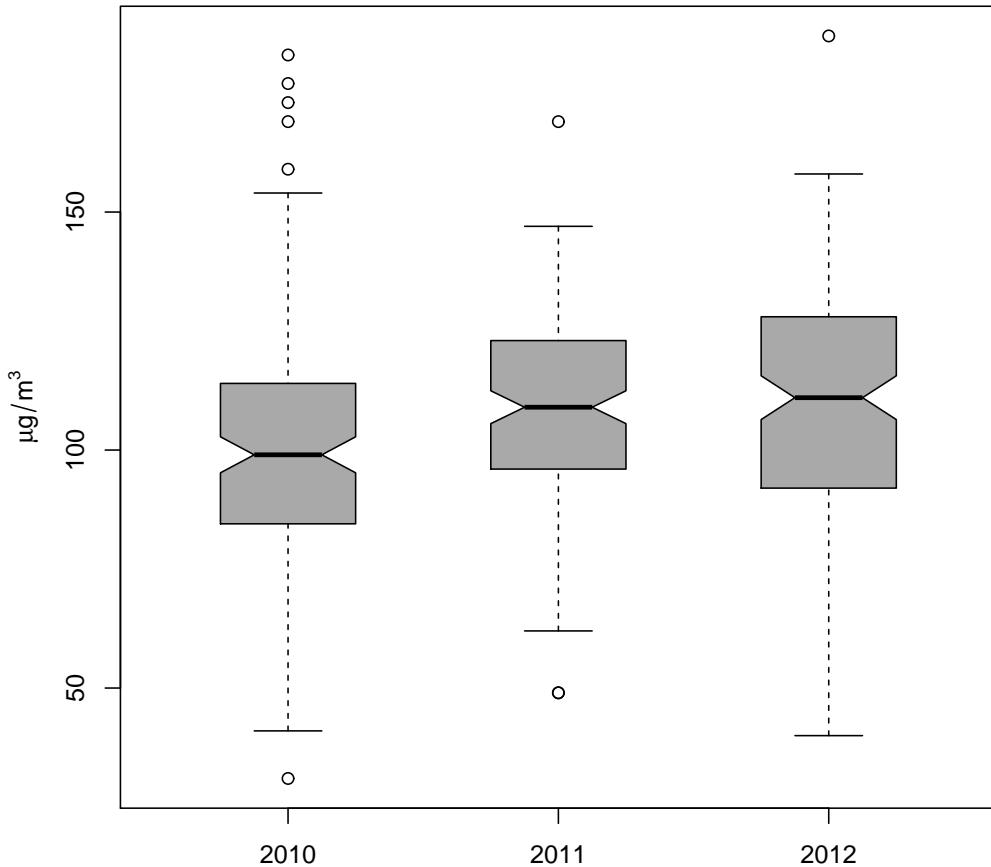
Der **Boxplot** (auch **Box-and-Whisker-Plot**) ist eine grafische Darstellung eines Datensatzes auf Basis der Quartile. Auf diese Weise können auch mehrere Datensätze schnell miteinander verglichen werden. Boxplots sind in der Literatur nicht eindeutig definiert. Die übliche Definition (JOHN W. TUKEY) lautet wie folgt:

Zeichne zunächst die **Box**, d. h. ein Rechteck vom 1. zum 3. Quartil (oder vom unteren zum oberen Hinge). Die Box umfasst also die mittleren 50% der Daten. Der Median (= 2. Quartil) wird durch eine Linie hervorgehoben. Bestimme die **Fences** (= Einzäunungen):

$$\text{Lower Fence: } \text{LF} = Q_1 - \underbrace{1.5(Q_3 - Q_1)}_{=: h}, \quad \text{Upper Fence: } \text{UF} = Q_3 + h$$

Nun zeichnet man die **Whiskers** (= Barthaare), d. h. Linien, die sich vom Rand der Box bis zu den äußersten Datenpunkten, die noch *innerhalb* der Fences liegen, erstrecken.

Abbildung 1.15: Boxplots für die Ozonwerte (Messstelle Illmitz)

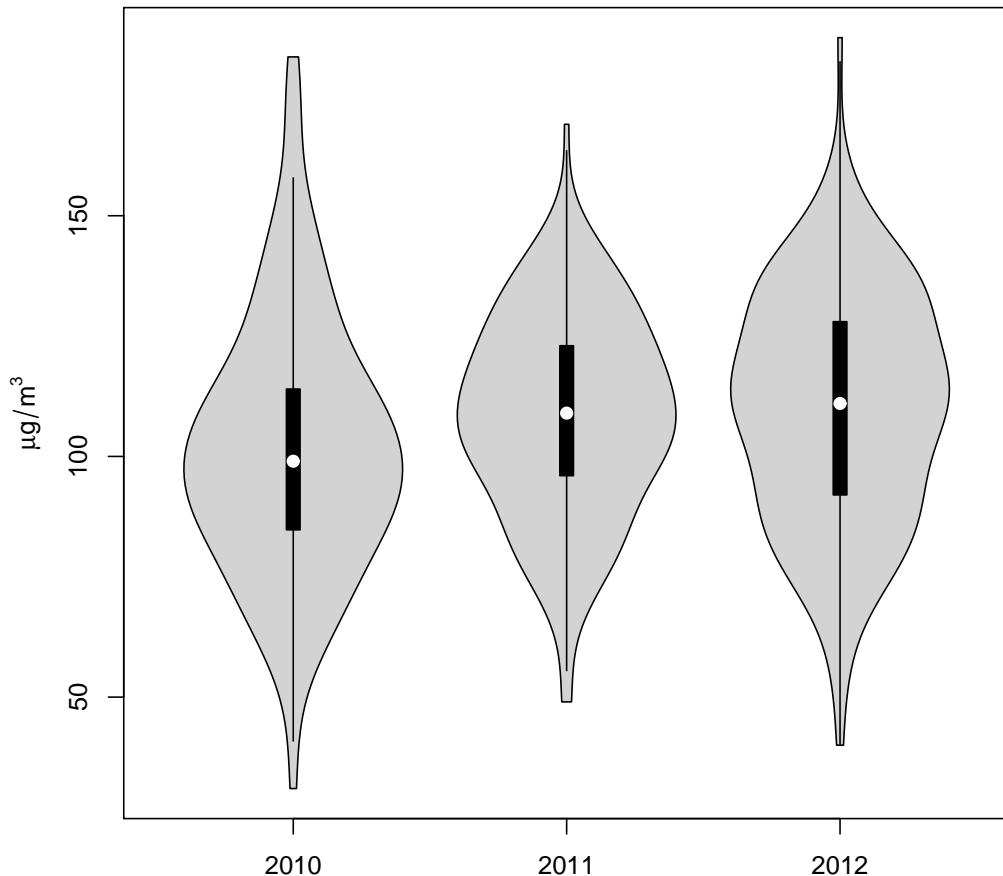


Punkte die außerhalb davon liegen, werden extra gezeichnet. Sie gelten als (potenzielle) „Ausreißer“, d. h. als Punkte, die sich vom Gros der Daten absetzen.

Zusätzlich kann man noch **Notches** (= Einkerbungen) zeichnen. Das sind keilförmige Bereiche, die einem 95%-Konfidenzintervall (vgl. 7.3) für den Median entsprechen.

Bem: Das obige Konzept lässt sich auf verschiedene Weise variieren. Häufig verzichtet man etwa auf die Fences und zeichnet die Whiskers bis zum Maximum bzw. Minimum der Daten. Da letztere Größen aber naturgemäß sehr empfindlich gegenüber Ausreißern sind, kann dadurch der optische Eindruck verfälscht werden. Eine Kombination von Boxplot und Kerndichteschätzung ist der **Violinplot** (vgl. das folgende Beispiel).

Bsp 1.19 Vergleichende Boxplots der Ozonwerte für die Messstelle Illmitz von Mai bis September für 2010, 2011 und 2012 sind in Abb 1.15 dargestellt. Man beachte die zahlreichen Ausreißer speziell für 2010. Zusätzlich sind auch die Notches eingezeichnet. Der bereits vom QQ-Plot (Abb 1.14) gewonne Eindruck (für 2010 und 2011) bestätigt sich. Abb 1.16 zeigt vergleichende Violinplots. Durch die spiegelartige Darstellung bekommt man einen guten Eindruck von der Verteilung der Ozonwerte. ■

Abbildung 1.16: Violinplots für die Ozonwerte (Messstelle Illmitz)

1.8 Kennzahlen

Neben den verschiedenen Möglichkeiten zur grafischen Aufbereitung von Datensätzen ist die Berechnung von **Stichprobenparametern** eine unabdingbare Ergänzung. (Bem: Einige, wie etwa der Median, wurden bereits bei der Erstellung von Grafiken verwendet.) Da **Ausreißer** in der Praxis eher die Regel als die Ausnahme sind, spielt die Frage der **Robustheit** bei der Auswahl der zu berechnenden Parameter keine un wesentliche Rolle.

Legt man (bereits) *klassierte* Daten zugrunde, so werden die jeweiligen Maßzahlen so berechnet, als ob alle Daten einer Klasse in deren Mittelpunkt liegen. (Zum Zwecke einer kürzeren Darstellung werden die entsprechenden Formeln im Folgenden nur gelegentlich angegeben.) Um einen Informationsverlust zu vermeiden, sollten Maßzahlen nach Möglichkeit auf Basis der *unklassierten* Daten (Rohdaten, Urdaten) berechnet werden.

Die Stichprobenparameter lassen sich in solche für die Kennzeichnung der **Lage** und der **Streuung** einteilen. Daneben gibt es auch Kennzahlen für die Beschreibung der **Verteilungsform**.

1.8.1 Mittelwert

Das wichtigste Lagemaß ist der (empirische) **Mittelwert** (oder **Stichprobenmittelwert**), bezeichnet mit \bar{x}_n (oder nur \bar{x} ; sprich: „x quer“). Sind x_1, x_2, \dots, x_n die Daten, so ist \bar{x}_n das arithmetische Mittel:

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$$

Minimumseigenschaft: Für den Mittelwert \bar{x}_n gilt:

$$\sum_{i=1}^n (x_i - \bar{x}_n)^2 \leq \sum_{i=1}^n (x_i - c)^2 \quad \text{für } c \in \mathbb{R}$$

Beweis: Sei $g(c) = (1/n) \sum_{i=1}^n (x_i - c)^2$, so sind die Ableitungen gegeben durch:

$$g'(c) = -\frac{2}{n} \sum_{i=1}^n (x_i - c), \quad g''(c) = 2$$

Aus $g'(c) = 0$ folgt:

$$\sum_{i=1}^n (x_i - c) = 0 \implies c = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} (=: c_0)$$

Wegen $g''(c_0) = 2 > 0$ handelt es sich um ein (relatives) Minimum. Das zeigt die Behauptung. Zur Illustration der Minimumseigenschaft des Mittelwerts zeigt Abb 1.17 die Funktion $g(c)$ auf Basis der Daten $\{3, 1, 7, 2, 4, 5, 4, 10, 6, 9\}$ von Bsp 1.15.

Berechnung aus Teilmittelwerten: Sind m Teilmittelwerte \bar{x}_{n_j} , $j = 1, 2, \dots, m$, gegeben, so gilt für den Gesamtmittelwert $\bar{\bar{x}}$ („x quer quer“):

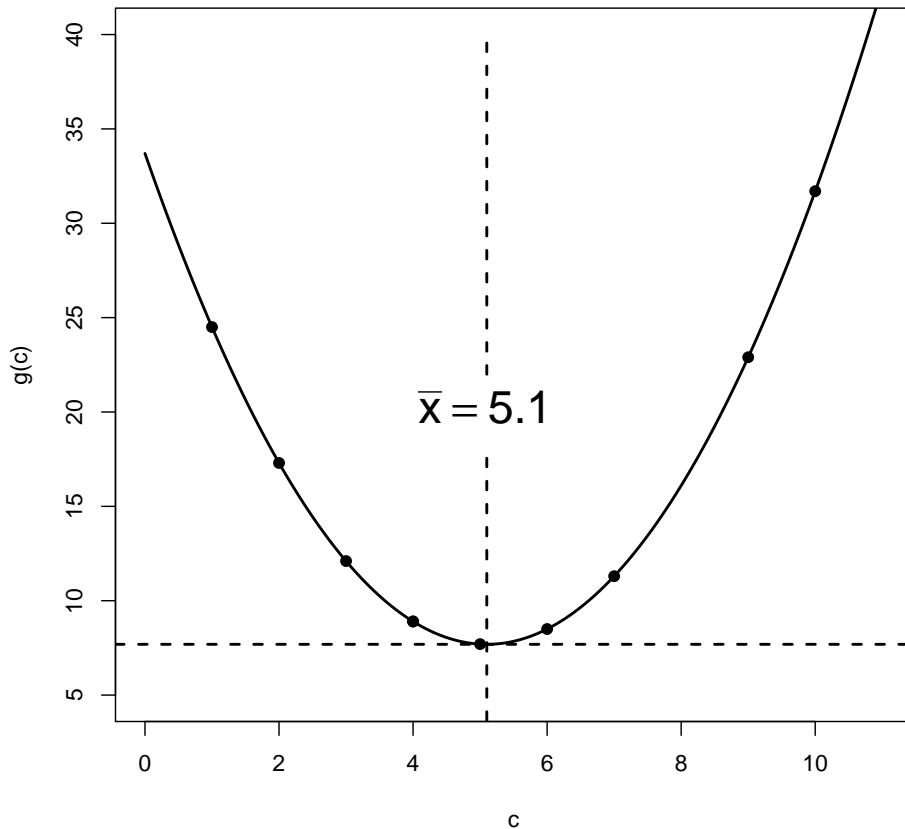
$$\bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^m n_j \bar{x}_{n_j} \quad \text{mit} \quad n = \sum_{j=1}^m n_j$$

Hierbei handelt es sich um ein *gewichtetes* Mittel (mit den Gewichten n_j/n) der Teilmittelwerte. Dieses Konzept lässt sich verallgemeinern.

Gewichteter Mittelwert:

$$\bar{x}_g = \sum_{i=1}^n g_i x_i \quad \text{mit} \quad g_i \geq 0, \quad \sum_{i=1}^n g_i = 1$$

Abbildung 1.17: Minimumseigenschaft des Mittelwerts



Mittelwert aus klassierten Daten: Sind H_j (f_j) die absoluten (relativen) Klassenhäufigkeiten und x_j^* die Mittelpunkte der Klassen K_j , $j = 1, 2, \dots, k$, so berechnet (d. h. approximiert) man den Mittelwert als gewichtetes Mittel der Klassenmitten:

$$\bar{x}_g = \frac{1}{\sum_{j=1}^k H_j} \sum_{j=1}^k H_j x_j^* = \sum_{j=1}^k f_j x_j^*$$

Je nach Verteilung der Daten innerhalb der Klassen kann \bar{x}_g größer oder kleiner als der tatsächliche Mittelwert (berechnet auf Basis der unklassierten Daten) sein. Gleichheit $\bar{x}_g = \bar{x}$ besteht nur dann, wenn die Daten in jeder Klasse symmetrisch um ihren Klassenmittelpunkt verteilt sind.

1.8.2 Geometrisches und harmonisches Mittel

In bestimmten Situationen ist das arithmetische Mittel kein sinnvolles Maß für den Durchschnittswert. Handelt es sich beispielsweise um **relative** Änderungen (z. B. Lohnerhöhung in %), so ist das **geometrische Mittel** geeigneter:

$$\bar{x}_n^{(g)} = \sqrt[n]{x_1 x_2 \cdots x_n} = \left(\prod_{i=1}^n x_i \right)^{1/n}$$

In anderen Fällen wiederum muss man richtigerweise das **harmonische Mittel** bilden:

$$\bar{x}_n^{(h)} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Hat man nur *positive* Beobachtungswerte x_1, x_2, \dots, x_n , so gilt stets die folgende Beziehung zwischen den diversen Mittelwerten:

$$\bar{x}_n^{(h)} \leq \bar{x}_n^{(g)} \leq \bar{x}_n$$

Gleichheit besteht nur für $x_1 = x_2 = \cdots = x_n$.

Sowohl das geometrische als auch das harmonische Mittel können zu **gewichteten** Mittelwerten verallgemeinert werden:

$$\bar{x}_g^{(g)} = \prod_{i=1}^n x_i^{g_i}, \quad \bar{x}_g^{(h)} = \frac{1}{\sum_{i=1}^n \frac{g_i}{x_i}} \quad \text{mit } g_i \geq 0, \quad \sum_{i=1}^n g_i = 1$$

Für $g_i = 1/n$ ergeben sich die gewöhnlichen Mittelwerte.

Bsp 1.20 Ein typisches Beispiel für ein gewichtetes harmonisches Mittel ist die Berechnung von Durchschnittsgeschwindigkeiten. Wird die Strecke W_i [km] mit der (konstanten) Geschwindigkeit V_i [km/h] in der Zeit T_i [h] zurückgelegt, so gilt:

$$W_i = V_i \times T_i, \quad i = 1, 2, \dots, n$$

Ist $T = \sum_{i=1}^n T_i$ die benötigte Zeit für die Gesamtstrecke $W = \sum_{i=1}^n W_i$, so gilt für die Durchschnittsgeschwindigkeit V :

$$V = \frac{W}{T} = \frac{\sum_{i=1}^n W_i}{\sum_{i=1}^n T_i} = \frac{\sum_{i=1}^n W_i}{\sum_{i=1}^n \frac{W_i}{V_i}} = \frac{1}{\sum_{i=1}^n \frac{g_i}{V_i}} \quad \text{mit } g_i = \frac{W_i}{\sum_{j=1}^n W_j}$$

V ist somit das mit g_i gewichtete harmonische Mittel der V_i . ■

1.8.3 Getrimmter Mittelwert

Ungewöhnlich große oder kleine Datenwerte (d. h. **Ausreißer**) können den arithmetischen Mittelwert \bar{x} u. U. stark beeinflussen oder verfälschen. Das ist eine Folge des Umstands, dass $\bar{x} = (1/n) \sum_{i=1}^n x_i$ jeden Datenwert x_i gleich gewichtet (mit $1/n$).

Um den Einfluss von (vermuteten) Ausreißern zu reduzieren, kann man z. B. bei der Berechnung von \bar{x} die kleinsten und größten Datenwerte unberücksichtigt lassen. Für $0 \leq \alpha < 0.5$ und $g = \lfloor \alpha n \rfloor$ ⁹ ist der α -getrimmte Mittelwert definiert durch:

$$\bar{x}_\alpha = \frac{1}{n - 2g} \sum_{i=g+1}^{n-g} x_{(i)}$$

D.h., bei der Berechnung von \bar{x}_α bleiben die g kleinsten und die g größten Werte am Anfang und Ende des (geordneten) Datensatzes unberücksichtigt. Typische Werte für α liegen zwischen 0.05 und 0.2.

Bsp 1.21 Zur Illustration betrachten wir den folgenden (bereits geordneten) Datensatz:

$$\begin{array}{cccccccccc} 77 & 87 & 87 & 114 & 151 & 210 & 219 & 246 & 253 & 262 \\ 296 & 299 & 306 & 376 & 428 & 515 & 666 & 1310 & 2611 \end{array}$$

Der ungetrimmte Mittelwert beträgt $\bar{x} \doteq 448.05$, der Median (= Grenzfall des getrimmten Mittelwerts für $\alpha \rightarrow 0.5$) ist $\tilde{x} = 262$. Für die – häufig empfohlene – 20% Trimming ist $g = \lfloor n\alpha \rfloor = \lfloor (19)(0.2) \rfloor = \lfloor 3.8 \rfloor = 3$ und für die Berechnung von $\bar{x}_{0.2}$ bleiben die drei kleinsten und die drei größten Beobachtungen unberücksichtigt:

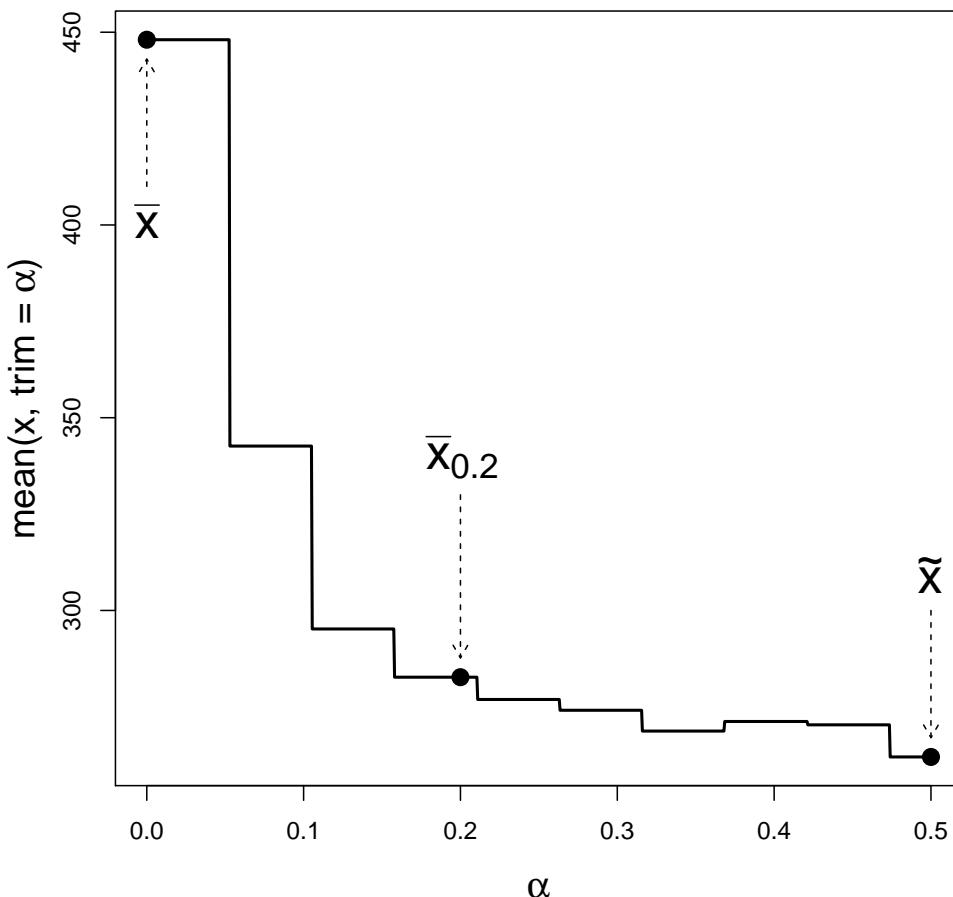
$$\bar{x}_{0.2} = \frac{114 + 151 + \dots + 428 + 515}{13} \doteq 282.69$$

In Abb 1.18 sind alle möglichen getrimmten Mittelwerte in Abhängigkeit von α (für $0 \leq \alpha < 0.5$) dargestellt. Es ergibt sich eine – nicht notwendigerweise monotone – treppenförmige Funktion.

An diesem Datensatz zeigt sich auch eine Problematik der unkritischen Trimming des Mittelwerts. Zeichnet man den Boxplot, so werden nur die zwei größten Beobachtungen (1310, 2611) als (potenzielle) Ausreißer ausgewiesen. Nimmt man aber eine 20% Trimming, so werden – wie oben gesehen – neben den drei größten auch die drei *kleinsten* Werte bei der Mittelwertsberechnung ausgeschlossen. D. h., die Trimming führt in diesem Fall ihrerseits zu einer unerwünschten Verzerrung des Mittelwerts. ■

⁹Für $a > 0$ bezeichnet $\lfloor a \rfloor$ die nächstkleinere ganze Zahl.

Abbildung 1.18: Getrimmte Mittelwerte



Bruchpunkt: Die Robustheit eines Schätzers in Bezug auf Ausreißer lässt sich u. a. durch seinen **Bruchpunkt** bemessen. Man versteht darunter den kleinsten Anteil (in %) der Datenwerte, den man ersetzen müsste, um den Schätzwert beliebig zu verändern. Wie man sich leicht überlegt, beträgt der Bruchpunkt von \bar{x}_α für großes n etwa $100\alpha\%$. Andererseits genügt die Ersetzung *eines* Datenpunkts, um \bar{x} beliebig zu verändern. Für großes n beträgt der Bruchpunkt des ungetrimmten Mittelwerts daher 0%.

1.8.4 Median

Der Median wurde bereits in einem früheren Abschnitt (1.7.7) als 50% Quantil (oder 2. Quartil) eines Datensatzes eingeführt. Die Definition werde hier in leicht abgeänderter Form wiederholt:

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & n \text{ ungerade} \\ \frac{1}{2} [x_{(n/2)} + x_{((n+2)/2)}] & n \text{ gerade} \end{cases}$$

Bruchpunkt: Der Bruchpunkt (vgl. 1.8.3) des Medians beträgt circa 50%. D. h., die Hälfte der Daten müsste ersetzt werden, um den Median beliebig zu verändern. Insofern ist der Median das robusteste Lagemaß.

Minimumseigenschaft: Auch der Median erfüllt eine Minimumseigenschaft:

$$\sum_{i=1}^n |x_i - \tilde{x}| \leq \sum_{i=1}^n |x_i - c| \quad \text{für } c \in \mathbb{R}$$

Beweis: Etwas unorthodox lässt sich das wie folgt zeigen ($\text{sgn}(x)$ bezeichnet die Vorzeichenfunktion: $\text{sign}(x) = -I_{(-\infty, 0)}(x) + I_{(0, \infty)}(x)$):

$$\frac{\partial}{\partial c} \sum_{i=1}^n |x_i - c| = - \sum_{i=1}^n \text{sgn}(x_i - c) = 0$$

Als Lösungen letzterer Gleichung ergeben sich alle c , die in der Summe die gleiche Anzahl von -1 und $+1$ erzeugen. Dies trifft auf den Median zu (für gerades n die einzige Lösung). Der Umstand, dass die Vorzeichenfunktion an der Stelle $x = 0$ nicht differenzierbar ist, spielt bei der obigen Überlegung keine Rolle (die Ableitung wird dort gleich 0 gesetzt). Zur Illustration der Minimumseigenschaft nehmen wir wieder die Daten $\{3, 1, 7, 2, 4, 5, 4, 10, 6, 9\}$. Abb 1.19 zeigt $g(c) := (1/n) \sum_{i=1}^n |x_i - c|$ in Abhängigkeit von c . Man beachte, dass alle c -Werte zwischen 4 und 5 (inklusive) die Funktion $g(c)$ minimieren.

Median aus klassierten Daten: $K_j = \langle u_j, u_j + w_j \rangle$ seien die Klassen und f_j die relativen Klassenhäufigkeiten ($j = 1, 2, \dots, k$). Gilt $\sum_{j=1}^{i-1} f_j < 0.5$ und $\sum_{j=1}^i f_j \geq 0.5$, so liegt der Median in der i -ten Klasse und man definiert:

$$\tilde{x} = u_i + \frac{0.5 - \sum_{j=1}^{i-1} f_j}{f_i} w_i$$

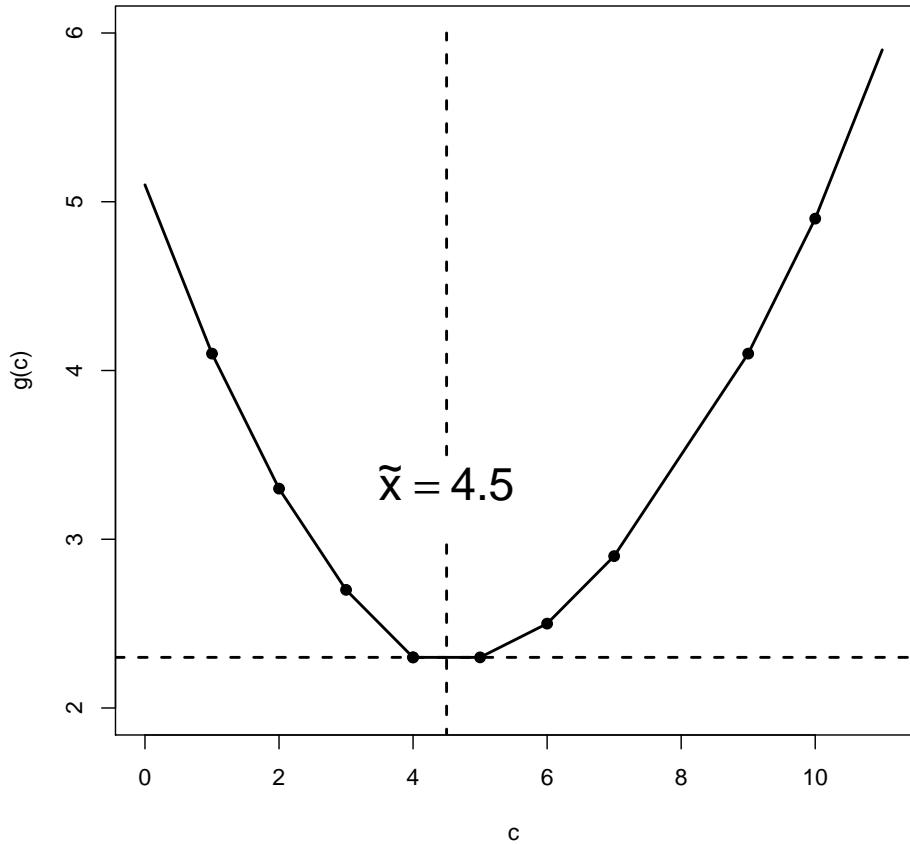
Bem: Nach diesem Muster können auch andere Quantile für klassierte Daten definiert werden.

1.8.5 Varianz

Neben den oben behandelten Kennzahlen für die Lage benötigt man auch Kennzahlen für die Charakterisierung des **Streuungsverhaltens** einer (empirisch gegebenen) Verteilung. Die am häufigsten verwendete Kennzahl dieser Art ist die (empirische) **Varianz** (oder **Stichprobenvarianz**) s_n^2 (kurz s^2), definiert durch:

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

Abbildung 1.19: Minimumseigenschaft des Medians



Die Varianz lässt sich als mittlere quadratische Abweichung der Daten von ihrem Mittelwert interpretieren. Die **Stichprobenstreuung** (oder **Standardabweichung**) ist die (positive) Wurzel aus der Varianz:

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

Bem: Die Bedeutung des auf den ersten Blick wenig einleuchtenden Faktors $1/(n-1)$ wird in 7.2.4 (Bsp 7.9) erklärt. Ist $\{x_1, x_2, \dots, x_n\}$ keine Stichprobe sondern die **Gesamtpopulation**, definiert man:

$$s'_n{}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \quad \text{bzw.} \quad s'_n = \sqrt{s'_n{}^2}$$

Spricht man einfach von der „Varianz“ oder der „Streuung“ eines Datensatzes ist aber stets s_n^2 bzw. s_n gemeint.

Verschiebungssatz: Die Varianz s_n^2 lässt sich auch wie folgt berechnen:

$$s_n^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n(\bar{x}_n)^2 \right] = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right]$$

Bem: Diese Darstellung von s_n^2 ist für numerische Berechnungen gegenüber der ursprünglichen Formel für s_n^2 vorzuziehen, da sich häufig auftretende Rundungsfehler hier weniger stark auswirken. Das gilt insbesondere für große Datensätze.

Berechnung aus Teilvarianzen: Sind m Teilvarianzen und Teilmittelwerte $s_{n_j}^2$, \bar{x}_{n_j} , $j = 1, 2, \dots, m$, gegeben, so gilt für die Gesamtvarianz:

$$s_n^2 = \frac{1}{n-1} \left[\sum_{j=1}^m (n_j - 1)s_{n_j}^2 + \sum_{j=1}^m n_j (\bar{x}_{n_j} - \bar{\bar{x}})^2 \right]$$

Wobei (vgl. 1.8.1):

$$\bar{\bar{x}} = \frac{1}{n} \sum_{j=1}^m n_j \bar{x}_{n_j} \quad \text{und} \quad n = \sum_{j=1}^m n_j$$

1.8.6 MAD

Verwendet man den Median \tilde{x} zur Kennzeichnung der Lage eines Datensatzes, kann man die folgenden Abstände bilden:

$$|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \dots, |x_n - \tilde{x}|$$

Der Mittelwert dieser Abstände, genannt die **mittlere absolute Abweichung** (oder kurz **MAD**), ist ein natürliches Streuungsmaß:

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \tilde{x}|$$

Bem: Manchmal wird auch über die Abstände zum Mittelwert \bar{x} gemittelt: $\sum_{i=1}^n |x_i - \bar{x}| / n$. Auch dieses Streuungsmaß wird als MAD bezeichnet.

Im Gegensatz zum Median ist der MAD nicht robust. Aus diesem Grund verwendet man anstelle des arithmetischen Mittels häufig wiederum den Median der Abstände:

$$\text{Median}\{|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \dots, |x_n - \tilde{x}|\}$$

Auch dieses Streuungsmaß wird als **MAD** bezeichnet (manchmal auch als **MedMed**). Der MAD in diesem Sinn hat wie der Median den maximalen Bruchpunkt 50% (vgl. 1.8.3 für die Definition des Bruchpunkts).

Bsp 1.22 Der MAD wird in R mittels `mad` bestimmt. Standardmäßig wird dabei der MAD mit der Konstanten 1.4826 multipliziert. Das erklärt sich daraus, dass der MAD häufig als (robuster) Schätzer für den Parameter σ einer Normalverteilung Verwendung findet und diese Konstante zur Verzerrungskorrektur benötigt wird. Darüberhinaus gibt es die Möglichkeit, bei der äußeren Medianbildung den *low*- oder *high*-Median zu nehmen, d. h., es wird bei einer geraden Anzahl von Beobachtungen nicht gemittelt. Anhand der Daten $\{1, 2, 3, 4, 5, 6, 7, 8\}$ sollen die verschiedenen Möglichkeiten demonstriert werden.

```
(x <- 1:8)
[1] 1 2 3 4 5 6 7 8
(m <- median(x))
[1] 4.5                                <--- Median (Daten)
sort(abs(x-m))
[1] 0.5 0.5 1.5 1.5 2.5 2.5 3.5 3.5   <--- geordnete Abstände
mad(x, constant=1)
[1] 2                                    <--- Median (Abstände)
mad(x)
[1] 2.9652                             <--- mit Verzerrungskorrektur
mad(x, constant=1, low=TRUE)
[1] 1.5                                 <--- low Median (Abstände)
mad(x, constant=1, high=TRUE)
[1] 2.5                                <--- high Median (Abstände)
```

1.8.7 Datenzusammenfassung

Aus einer übersichtlichen Darstellung von einigen Kennzahlen der Lage und der Streuung lässt sich schon einiges über einen Datensatz erkennen. Bei der **5-Zahlen-Zusammenfassung** werden die folgenden Werte angezeigt:

$$x_{(1)} \text{ (Min)}, \quad Q_1 \text{ (u. Hinge)}, \quad \tilde{x} \text{ (Med)}, \quad Q_3 \text{ (o. Hinge)}, \quad x_{(n)} \text{ (Max)}$$

In einer erweiterten Fassung wird zusätzlich zum Median auch der Mittelwert \bar{x} angezeigt. Man beachte, dass der Boxplot (vgl. 1.7.9) quasi eine grafische Darstellung der 5-Zahlen-Zusammenfassung ist.

Die **Spannweite** (oder **Range**) $R = x_{(n)} - x_{(1)}$ ist ein Maß für die „Spreizung“ des Datensatzes. Der **Interquartilabstand** (kurz **IQA** oder **IQR**) zeigt die Spreizung der mittleren 50% der Daten. Anstelle der Quartilendifferenz kann man auch den **Hingeabstand** (kurz **HA**) nehmen.

Bem: Die von den R-Funktionen angezeigten Kenngrößen kann man einfach um weitere Kenngrößen (z. B. MAD) erweitern. Zu ausladende Darstellungen wirken allerdings schnell unübersichtlich (insbesondere bei mehreren Datensätzen) und sollten vermieden werden.

Einheiten der Kenngrößen: Die meisten Kenngrößen – Ausnahmen sind der Variationskoeffizient (vgl. UE–Aufgabe 1.14) und die in den folgenden Abschnitten diskutierten Maßzahlen der Schiefe und Kurtosis – haben auch **Einheiten**. Der Mittelwert, die Quantile, die Hinges, der MAD, etc. haben jeweils die Dimension [D] der Beobachtungen. Die Einheit der Varianz ist allerdings $[D^2]$; das macht die direkte Interpretation dieser Größe schwierig. Andererseits hat aber die Streuung wiederum die Dimension [D] und lässt sich einfacher interpretieren.

Bsp 1.23 Der folgende R-Output zeigt für die Ozondaten (vgl. Bsp 1.18) eine übersichtliche Darstellung der 5(bzw. 6)-Zahlen–Zusammenfassung getrennt nach Jahr.

Year:	2010						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
	31.00	84.75	99.00	101.50	114.00	183.00	1.00

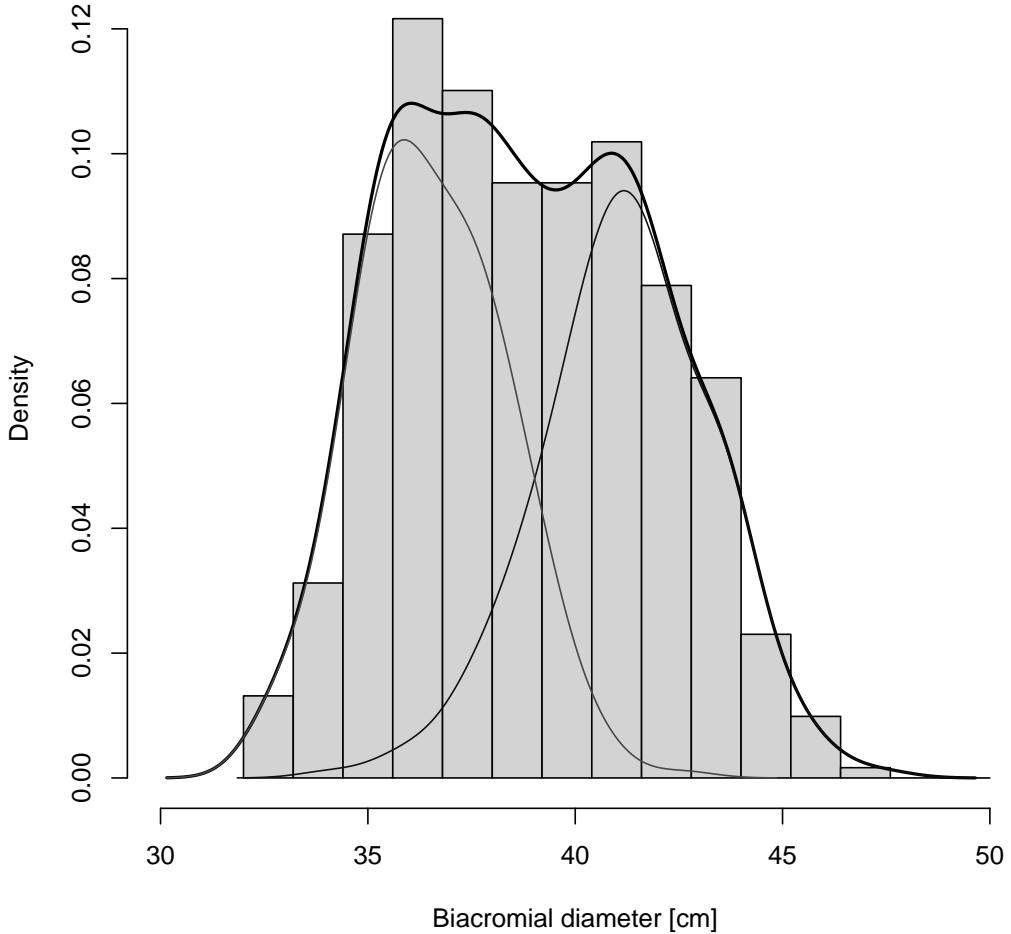
Year:	2011						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	
	49.0	96.0	109.0	108.3	123.0	169.0	

Year:	2012						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	
	40.0	92.0	111.0	110.1	128.0	187.0	

Im Jahr 2010 fehlt eine Beobachtung (NA); die hier verwendete `summary()` Funktion ist aber so voreingestellt, dass fehlende Beobachtungen nicht zu einer Fehlermeldung führen (d. h. `na.rm = TRUE`). Die Einheit der angezeigten Kenngrößen ist [$\mu\text{g}/\text{m}^3$]. ■

1.8.8 Modalwert

Ein Bezugspunkt bei der Beurteilung der Form einer Verteilung ist der **Modalwert** (oder **Modus**). Allgemein versteht man darunter eine Merkmalsausprägung mit höchster „Dichte“. Bei diskreten Merkmalen wäre dies die Ausprägung mit der höchsten Beobachtungshäufigkeit. Bei stetigen Merkmalen bezieht man sich meist auf das Histogramm und

Abbildung 1.20: Beispiel für ein Mischverteilung

betrachtet z. B. den Mittelpunkt der Klasse mit der höchsten beobachteten (relativen) Häufigkeit (d. h. die **Modalklasse**) als Modus. (Bem: Im Falle der Kerndichteschätzung wäre der Modus die Stelle des Maximums der Dichtekurve.)

In vielen Fällen ist der Modalwert mehr oder weniger deutlich ausgeprägt, manchmal gibt es aber auch mehrere (meist zwei) deutlich erkennbare – i. A. nicht gleich hohe – „Gipfel“. Handelt es sich um „echte“ Gipfel, liegt eine mehrgipflige Verteilung vor und man spricht von einer **multimodalen** (im Falle von zwei Gipfeln, von einer **bimodalen**) Verteilung. (Bem: Multimodale Verteilungen sind häufig das Resultat einer *Verteilungsmischung*.)

Bsp 1.24 Als Beispiel für eine bimodale Mischverteilung betrachten wir das Merkmal **Biacromial** aus dem Datensatz **body.txt** (vgl. Bsp 1.5), wobei wir nun nicht nach Geschlecht unterscheiden. Das Histogramm (Abb 1.20) zeigt zwei unterschiedlich stark ausgeprägte Peaks, die sich in der überlagerten Kerndichteschätzung widerspiegeln. Zur Verdeutlichung der Mischung sind auch die Kernschätzungen der beiden Teildatensätze (für **Gender = 0** und **Gender = 1**) eingezeichnet. (Bem: Für alle drei Kernschätzungen wird die gleiche Bandbreite genommen.)

Im vorliegenden Fall sind wir uns des Umstands der (unkorrekteten) Mischung bewusst, in anderen Fällen mag die Situation aber nicht so klar und die „Entmischung“ der Teildatensätze schwierig (d. h. mit großen Unsicherheiten behaftet) sein. ■

1.8.9 Momente

Für einen Datensatz x_1, x_2, \dots, x_n ist das (empirische) **Moment**¹⁰ der **Ordnung** r um den Nullpunkt definiert durch:

$$m'_r = \frac{1}{n} \sum_{i=1}^n x_i^r \quad \text{für } r = 1, 2, \dots$$

Kurz nennt man m'_r einfach das r -te Moment der Daten. Bildet man die Momente um den Mittelwert \bar{x} (= „Schwerpunkt“ der Daten), bekommt man die **zentralen Momente**:

$$m_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r \quad \text{für } r = 1, 2, \dots$$

Bem: Bei Datenmomenten nimmt man stets den Faktor $1/n$. Die Varianz s_n^2 ist in diesem Sinne – bis auf den Faktor $1/(n-1)$ – somit ein zentrales Moment 2. Ordnung, der Mittelwert \bar{x} ($= m'_1$) aber ein Moment 1. Ordnung.

1.8.10 Schiefe

Um die **Schiefe** einer (empirisch gegebenen) Verteilung zu charakterisieren, kann man sich der in 1.8.9 definierten Momente (der Ordnung 2 und 3) bedienen. Mehrere Definitionen (oder Typen) sind gebräuchlich:

$$\begin{aligned} g_1^{(1)} &= \frac{m_3}{m_2^{3/2}} = \frac{\sqrt{n}}{(n-1)\sqrt{n-1}} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3 \\ g_1^{(2)} &= \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3 \\ g_1^{(3)} &= \frac{m_3}{s^3} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3 \end{aligned}$$

Die obigen Schiefekoeffizienten sind i. W. standardisierte (zentrale) Momente 3. Ordnung. Eine auf den Quartilen basierende Definition (wobei die Quartile durch die Hinges ersetzt werden können) lautet:

¹⁰Der Ausdruck kommt aus der Mechanik und meint dort das *Drehmoment* einer Masse um eine Achse.

$$g_1^{(4)} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

Interpretation: Für die Interpretation der obigen Schiefekoeffizienten beziehen wir uns auf den Modalwert (vgl. 1.8.8), im Folgenden mit mod bezeichnet. Typischerweise gilt je nach Vorzeichen von g_1 :

- | | |
|--|--|
| $g_1 > 0$ | linkssteil/rechtsschief: mod < \bar{x} , $\tilde{x} < \bar{x}$, mod < \tilde{x} |
| $g_1 \approx 0$ (annähernd) symmetrisch: | mod $\approx \bar{x}$, $\tilde{x} \approx \bar{x}$, mod $\approx \tilde{x}$ |
| $g_1 < 0$ | rechtssteil/linksschief: mod > \bar{x} , $\tilde{x} > \bar{x}$, mod > \tilde{x} |

Bemerkungen:

- (a) Die Maßzahlen $g_1^{(1)}$, $g_1^{(2)}$ und $g_1^{(3)}$ weisen nur bei kleineren Stichproben größere Unterschiede auf. Infolge der nicht gegebenen Robustheit ist ihre Interpretation aber häufig schwierig.
- (b) $g_1^{(4)}$ ist ein robustes Schiefemaß. Es heißt auch *Quartilenkoeffizient* der Schiefe oder *Bowley-Koeffizient*. Wegen $|g_1^{(4)}| \leq 1$ (-1 : extrem rechtssteil, $+1$: extrem linkssteil) ist dieser Koeffizient auch einfach zu interpretieren.
- (c) Man findet noch andere Maßzahlen für die Schiefe (z. B. auf Basis eines Vergleichs von Mittel- und Modalwert).

1.8.11 Kurtosis

Zur Charakterisierung der **Kurtosis**¹¹ einer (empirisch gegebenen) Verteilung kann man die in 1.8.9 definierten Momente der Ordnung 2 und 4 heranziehen. Mehrere Definitionen (oder Typen) sind gebräuchlich:

$$\begin{aligned} g_2^{(1)} &= \frac{m_4}{m_2^2} = \frac{n}{(n-1)^2} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 \\ g_2^{(2)} &= \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 \\ g_2^{(3)} &= \frac{m_4}{s^4} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 \end{aligned}$$

¹¹*kurtosis* (auch *kyrtosis*) griech. = Krümmung, Wölbung

Die obigen Wölbungskoeffizienten sind i. W. standardisierte (zentrale) Momente 4. Ordnung. Als Referenz fungiert üblicherweise die „Glockenkurve“, deren (theoretische) Wölbung einen Wert von 3 hat. Die Verteilung nennt man daher:

platykurtisch (flach gewölbt), wenn $g_2 < 3$

mesokurtisch (mittel gewölbt), wenn $g_2 \approx 3$

leptokurtisch (steilgipflig), wenn $g_2 > 3$

Eine auf den *Oktilen* (= Achteln; A_i bezeichnet im Folgenden das $i/8$ -Quantil) basierende Definition lautet:

$$g_2^{(4)} = \frac{(A_7 - A_5) + (A_3 - A_1)}{A_6 - A_2}$$

Bemerkungen:

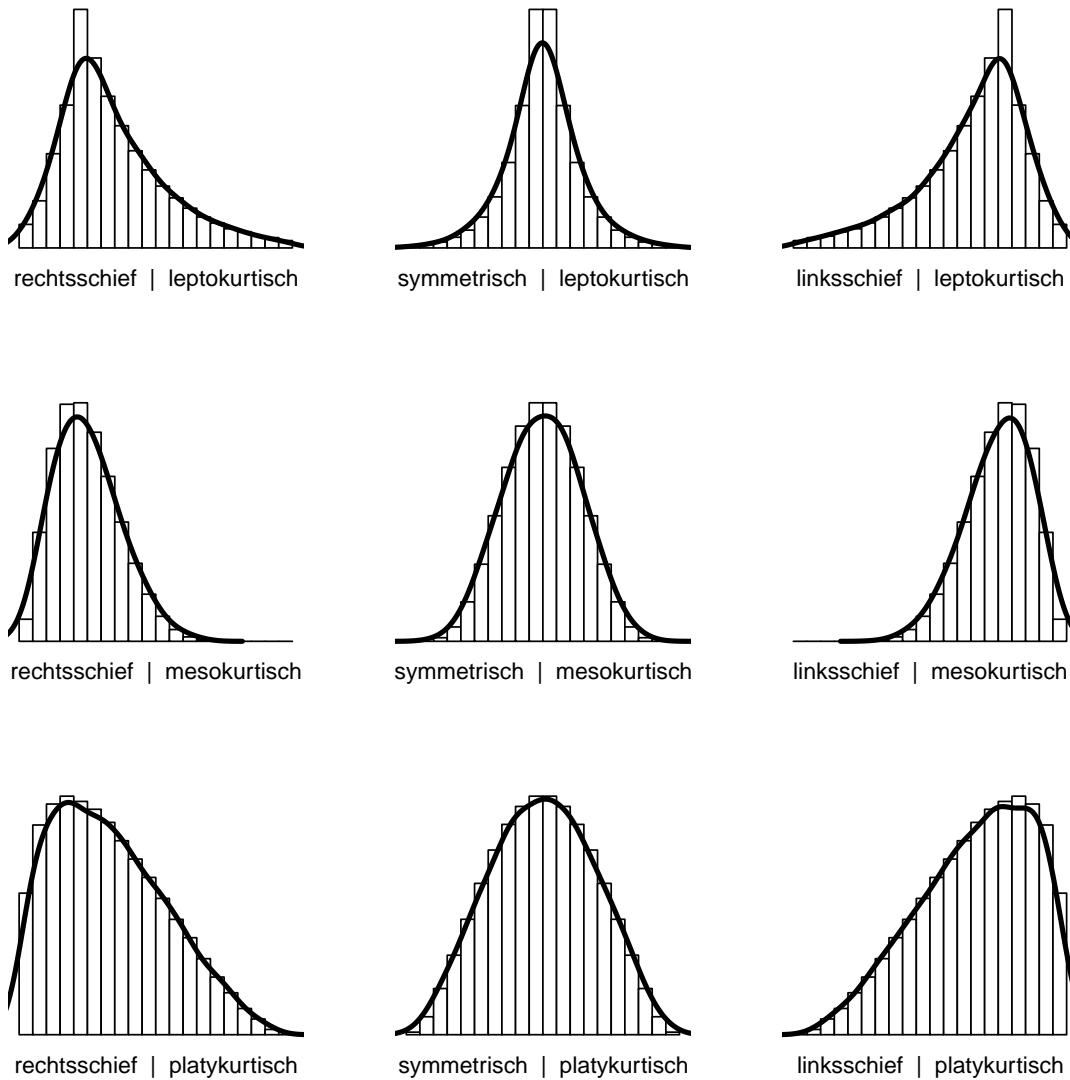
- (a) Die Maßzahlen $g_2^{(1)}$, $g_2^{(2)}$ und $g_2^{(3)}$ weisen nur bei kleineren Stichproben größere Unterschiede auf. Infolge der nicht gegebenen Robustheit ist ihre Interpretation aber häufig schwierig.
- (b) Der *Oktilenkoeffizient* der Wölbung (auch *Moors-Koeffizient* genannt) $g_2^{(4)}$ ist ein robustes Maß. Es liegt zwischen -1 (extrem plati-kurtisch), 1.233 (mesokurtisch; Glockenkurve) und $+\infty$ (extrem leptokurtisch).
- (c) Zieht man von $g_2^{(i)}$, $i = 1, 2, 3$, den Wert 3 ab, spricht man vom **Exzess**.
- (d) Tatsächlich sind $g_2^{(i)}$, $i = 1, 2, 3$, Maßzahlen für die Schwere der Ausläufer relativ zum Mittelteil („Schulter“) der Verteilung. Das hat zur Folge, dass etwa für bimodale Verteilungen der Exzess stark negativ sein kann. (Bsp: Der Exzess der bimodalen Verteilung von Abb 1.20 beträgt ≈ -0.84 .) Ebenso ist für rechtecksförmige Verteilungen (ausgeprägte Schulter, keine Ausläufer) der Exzess negativ. Das gilt sogar für dreiecksförmige Verteilungen (haben relativ zur Glockenkurve eine stärker ausgeprägte Schulter).

1.8.12 Verteilungsform

Mit Hilfe der in den vorigen Abschnitten diskutierten Begriffe „Schiefe“ und „Wölbung“ lassen sich grundsätzliche Formen von unimodalen Verteilungen charakterisieren. In Abb 1.21 sind einige in Anwendungen häufig anzutreffende Verteilungstypen dargestellt.

In der Praxis trifft man aber noch auf eine Reihe von anderen Verteilungsformen, die durch Koeffizienten von der Art g_1 und g_2 nur unzureichend beschreibbar sind. Beispiele sind etwa rechtecksförmige, dreiecksförmige, J-förmige, U-förmige Verteilungen oder Verteilungen mit mehreren Peaks (Abb 1.22).

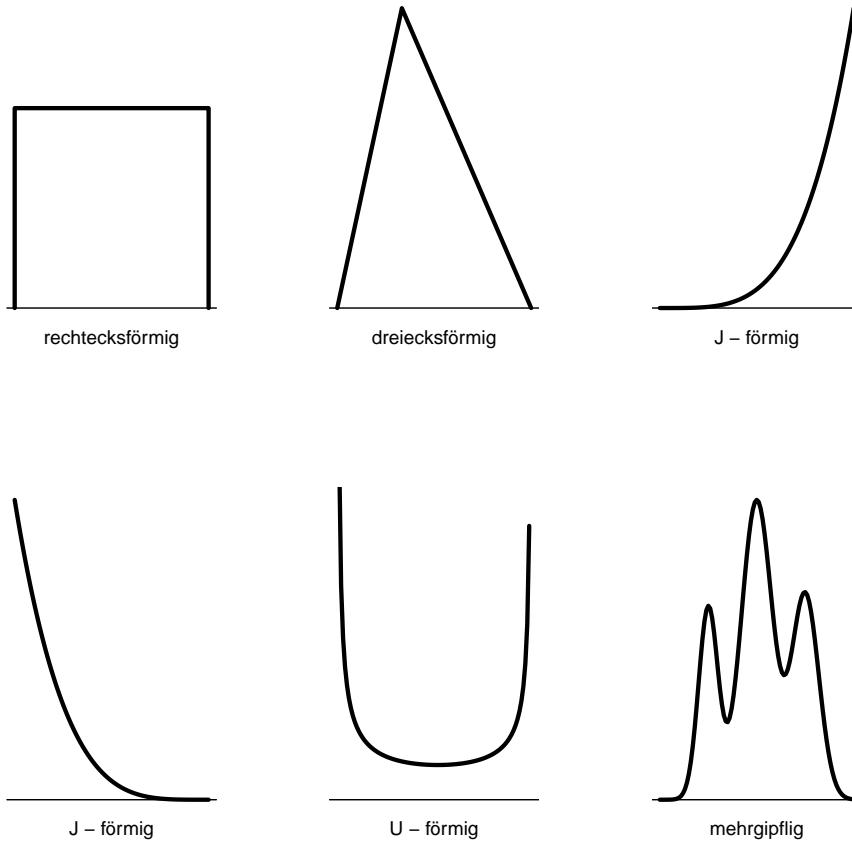
Abbildung 1.21: Typische unimodale Verteilungsformen



1.9 Mehrdimensionale Daten

Werden an beobachteten Einheiten Messungen für mehrere Merkmale vorgenommen, spricht man von **multivariaten** Beobachtungen. Neben der Untersuchung und Charakterisierung der einzelnen Merkmale stehen insbesondere die verschiedenen Beziehungen *zwischen* den Merkmalen im Mittelpunkt des Interesses. Dazu kann man sich – abhängig von Datenstruktur und Zielsetzung – der vielfältigen Methoden der **multivariaten Statistik** bedienen.

In diesem Abschnitt beschränken wir uns allerdings auf einige grafische Methoden und auf Methoden der Korrelations- und Regressionsrechnung für die Analyse von *quantitativen* (metrischen) mehrdimensionalen (speziell zweidimensionalen) Merkmalen.

Abbildung 1.22: Weitere typische Verteilungsformen

Es liege eine Stichprobe von n Beobachtungsvektoren¹² $(x_{i1}, x_{i2}, \dots, x_{ip})$, $i = 1, 2, \dots, n$, zu je p Variablen (Merkmale) vor. Die Beobachtungen lassen sich in Form einer $(n \times p)$ -**Datenmatrix** zusammenfassen (vgl. 1.5):

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Die Zeilen von \mathbf{X} entsprechen den Beobachtungen, die Spalten den Merkmalen. Derartige Schemata – meist ergänzt um eine Zeile mit den Variablennamen und eine Spalte zur Identifizierung der Beobachtungen – werden auch als **Datenframes** bezeichnet. Letztere bilden die grundlegenden Einheiten für statistische Analysen verschiedener Art.

Ein Beispiel ist der schon mehrfach verwendete Datensatz `body.txt`, bestehend aus Beobachtungen zu fünf metrischen Merkmalen (`Biacromial`, `Waist`, ...) und einem nominellen Merkmal (`Gender`).

¹²Als Zeilenvektoren betrachtet.

1.9.1 Scatterplots

Im Falle von zwei (metrischen) Merkmalen kann man die Beobachtungspaare (x_{1i}, x_{i2}) , $i = 1, 2, \dots, n$, als Punkte in einem kartesischen Koordinatensystem interpretieren und in Form eines **Scatterplots** darstellen. Durch „Überladen“ der Punkte eines Scatterplots (Farbe, Größe/Art der Punkte, u. Ä.) können weitere (meist nominelle) Merkmale repräsentiert werden. (Bem: Man sollte derartige Mittel nur sparsam einsetzen, da zu sehr überladene Plots unübersichtlich wirken.)

Bestehen die Daten aus Beobachtungsvektoren (metrischer) Merkmale, $(x_{1i}, x_{2i}, \dots, x_{pi})$, $i = 1, 2, \dots, n$, kann man die einzelnen Merkmale paarweise gegeneinander zeichnen und in Form einer 2-dimensionalen **Scatterplotmatrix** anordnen. Diese Plots können durch zusätzliche grafische (und/oder numerische) Elemente (Histogramme, Boxplots, Trendkurven, etc.) ergänzt werden. Derartige Plots bilden meist den Ausgangspunkt für weitere statistische Analysen.

Bsp 1.25 Einige der Merkmale aus dem Datensatz `body.txt` wurden bereits in früheren Abschnitten auf univariater Basis auf die eine oder andere Art grafisch (und z. T. auch mittels Kenngrößen) aufbereitet. Hier stellen wir zunächst die Merkmale `Weight` und `Height` in Form eines Scatterplots (**Abb 1.23**) dar und überladen den Plot durch Verwendung unterschiedlicher Symbole mit dem nominellen Merkmal `Gender`.

Bem: Man beachte, dass einige Punkte aus dem „Bulk“ der Daten hervorstechen, d. h. ungewöhnliche x - und/oder y -Koordinaten aufweisen. Sollte es sich um echte Datenpunkte (d. h. nicht um Schreibfehler o. Ä.) handeln, sind sie als (potenzielle) Ausreißer zu betrachten und als solche in weiteren Analysen zu berücksichtigen.

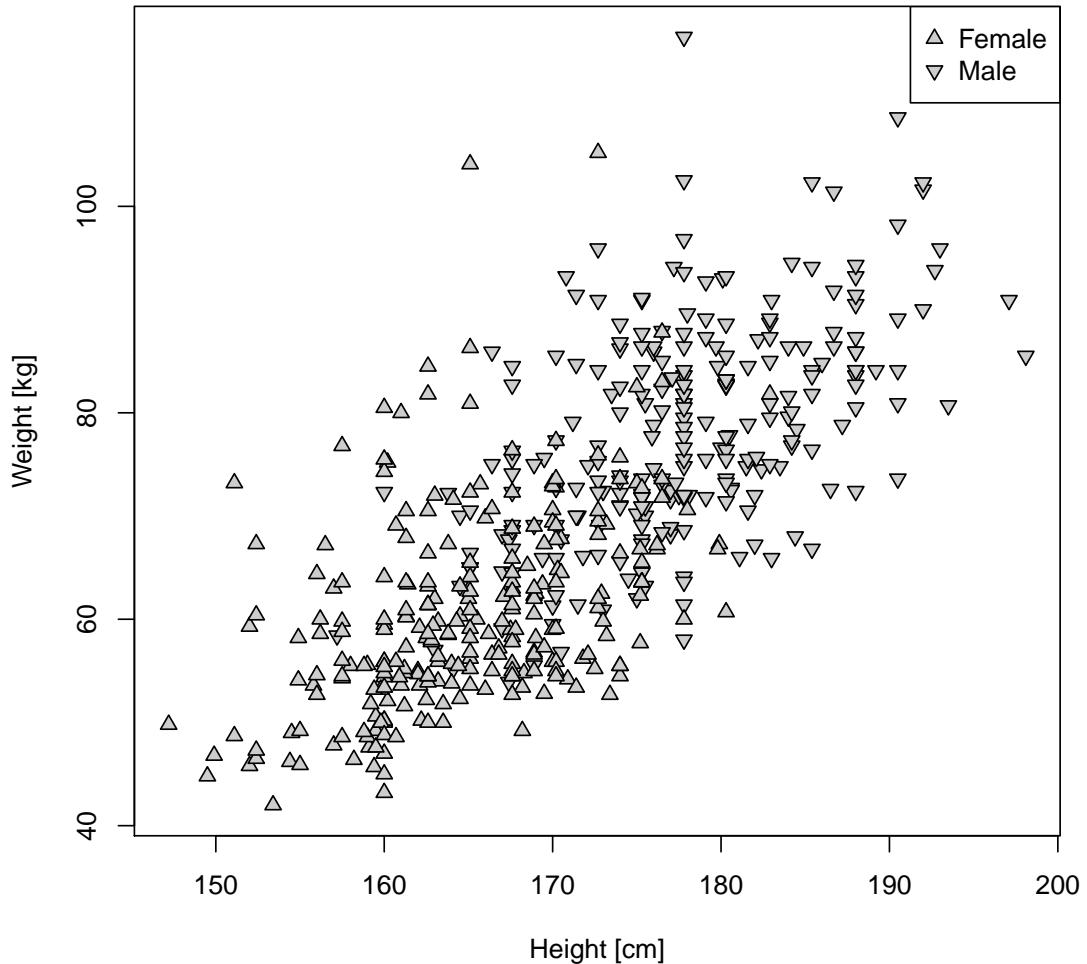
Als nächstes stellen wir alle (metrischen) Merkmale paarweise gegeneinander in Form einer Scatterplotmatrix dar, wobei zusätzlich das Merkmal `Gender` durch unterschiedliche Symbole repräsentiert wird (**Abb 1.24**). Die erkennbaren Zusammenhänge entsprechen weitgehend den Erwartungen, wobei die Abhängigkeit vom Merkmal `Age` nur sehr schwach (wenn überhaupt) ausgeprägt ist (bei erwachsenen Personen ebenfalls zu erwarten). ■

Bem: Dreidimensionale Darstellungen (von je drei Merkmalen) sind nur dann sinnvoll, wenn eine entsprechende Software zur Erzeugung *dynamischer* Grafiken (Drehen, Ändern der Skalierung, etc.) zur Verfügung steht. Eine Alternative besteht darin, durch „Überladen“ zweidimensionaler Scatterplots weitere Merkmale zu repräsentieren (vgl. das obige Bsp 1.25).

1.9.2 Kernschätzung

Das Konzept der **Kerndichteschätzung** (vgl. 1.7.6) lässt sich auf den Fall mehrdimensionaler (stetiger) Beobachtungen erweitern. Mehrere Erweiterungen sind denkbar; der einfachste multivariate Kernschätzer basiert auf dem **Produktkern**:

Abbildung 1.23: Scatterplot von Weight gegen Height



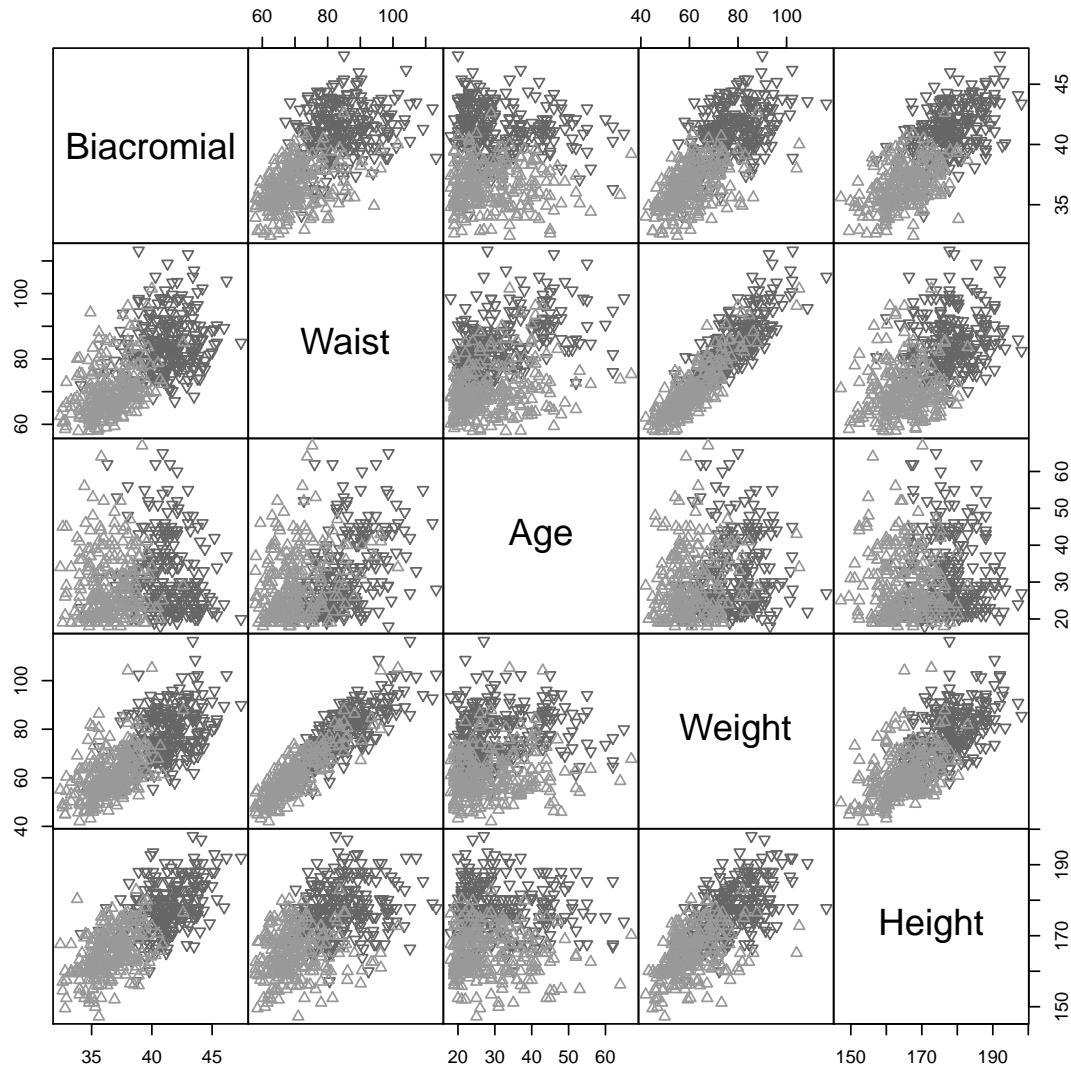
$$\hat{f}_n(\mathbf{x}) = \hat{f}_n(x_1, x_2, \dots, x_p) = \frac{1}{n \prod_{j=1}^p h_j} \sum_{i=1}^n \left\{ \prod_{j=1}^p K \left(\frac{x_j - x_{ij}}{h_j} \right) \right\}$$

für $\mathbf{x} = (x_1, x_2, \dots, x_p) \in \mathbb{R}^p$

Dabei ist p die Dimension der Beobachtungen, x_{ij} die j -te Komponente der i -ten Beobachtung und h_j die Bandbreite der j -ten Kernfunktion ($i = 1, 2, \dots, n$; $j = 1, 2, \dots, p$). In der obigen Form wird für jede Dimension derselbe Kern verwendet (mit möglicherweise verschiedenen Bandbreiten), das ist aber nicht zwingend.

Bem: In der Praxis betrachtet man Dichteschätzungen nur für je zwei Merkmale. Dynamische Grafiken sind zu bevorzugen; in jedem Fall sollte man die Grafiken durch **Contourplots** (= Plots der Höhenschichtlinien) oder dgl. ergänzen.

Abbildung 1.24: Scatterplotmatrix (body.txt)



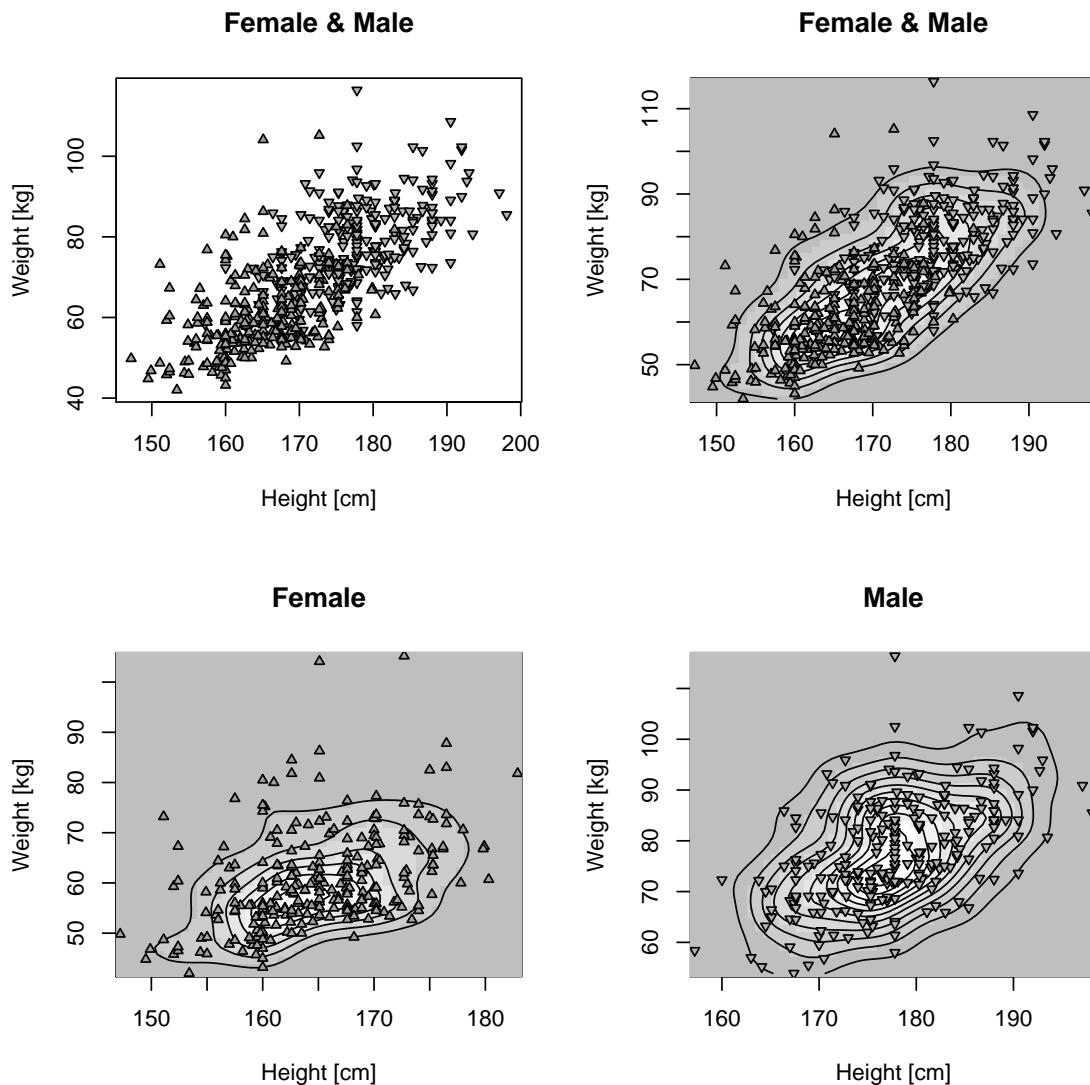
Scott's Rule: Im Falle des Normalkerns (und Beobachtungen nach einer multivariaten „Glockenkurve“) empfiehlt **Scott's Rule** die folgenden Bandbreiten:

$$h_j = \left[\frac{4}{n(p+2)} \right]^{1/(p+4)} \times s_j, \quad j = 1, 2, \dots, p$$

Dabei ist s_j die Streuung der Beobachtungen der j -ten Dimension:

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \quad \text{mit} \quad \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

Abbildung 1.25: Scatterplot und Kernschätzung für Weight gegen Height

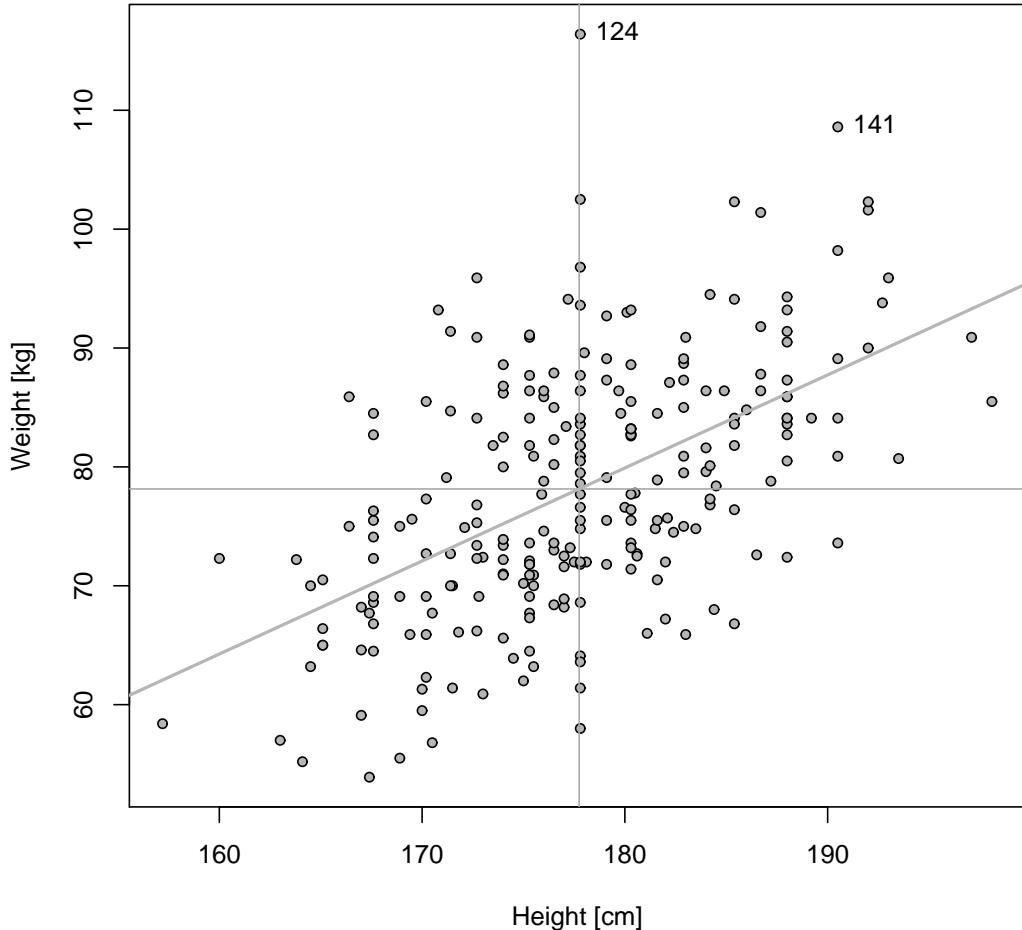


Bsp 1.26 Als Beispiel betrachten wir wieder die Merkmale `Weight` und `Height` aus dem Datenframe `body.txt`. Abb 1.25 zeigt den Scatterplot (wie in Abb 1.23) sowie die von Contourlinien – ermittelt auf Basis einer Kerndichteschätzung unter Verwendung von Scott's Rule – überlagerten Scatterplots für `Gender = 0` und `Gender = 1` gemeinsam und getrennt. ■

1.9.3 Korrelation

Scatterplots geben nicht nur eine grafische Veranschaulichung eines bivariaten Datensatzes, sondern lassen auch Art und Stärke eines eventuell vorhandenen Zusammenhangs zwischen den beiden Merkmalen erkennen. Betrachten wir beispielsweise noch einmal die

Abbildung 1.26: Scatterplot von Weight gegen Height für Gender = 1



Merkmale `Weight` und `Height` (Datenframe: `body.txt`) für `Gender = 1`, ergänzt um vertikale und horizontale Linien an den Stellen der Mittelwerte (`Height`: $\bar{x} = 177.75$ [cm]; `Weight`: $\bar{y} = 78.14$ [kg]), sowie um die „Kleinste-Quadrat-Gerade“ (Abb 1.26). (Bem: Letztere wird in Kapitel 9 ausführlicher behandelt.) Aus dem Plot lassen sich mehrere Einsichten gewinnen:

- (1) Es zeigt sich eine *positive Assoziation* zwischen den Merkmalen. Wie zu erwarten, sind größere Männer tendenziell schwerer als kleinere.
- (2) Der Zusammenhang zwischen den Merkmalen ist grob *linearer* Natur. D.h., jede Einheit an zusätzlicher Körpergröße erhöht das Körpergewicht um etwa den gleichen Betrag. (Hier um ca. 7.8 kg bei Zunahme der Größe um 10 cm.)
- (3) Die Assoziation zwischen den Merkmalen ist nicht sehr stark ausgeprägt. D.h., die Streuung der Punkte um die KQ-Gerade ist vergleichsweise groß. Ein Punkt (Nr. 124) sticht besonders hervor; wie man leicht überprüfen kann, beeinflusst er die KQ-Gerade aber praktisch nicht.

Neben qualitativen Feststellungen der obigen Art ist man aber auch an einer zahlenmäßigen Quantifizierung der Assoziation zwischen Merkmalen interessiert. Letzteres ist insbesondere dann nützlich, wenn man mehrere Datensätze miteinander vergleichen möchte (beispielsweise in der obigen Situation $\text{Gender} = 1$ mit $\text{Gender} = 0$). Die bekannteste Maßzahl dieser Art ist der (Stichproben-) **Korrelationskoeffizient**.¹³ Er misst den Grad der **linearen Assoziation** zwischen zwei Merkmalen.

Zur Motivation betrachte man nochmals Abb 1.26 und die durch die vertikale und horizontale Gerade (durch den jeweiligen Mittelwert) hervorgerufene Aufteilung des Scatterplots in vier Quadranten. Da wir an einem dimensionslosen Assoziationsmaß interessiert sind, betrachten wir die *standardisierten* Abweichungen der einzelnen Beobachtungen von ihrem Mittelwert, d. h. $(x_i - \bar{x})/s_x$ und $(y_i - \bar{y})/s_y$, und die daraus gebildeten Produkte:

$$\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right), \quad i = 1, 2, \dots, n$$

Gibt es eine *positive* Assoziation, werden diese Produkte großteils positiv sein, da y -Werte, die größer (kleiner) als ihr Durchschnitt sind, meist mit x -Werten, die größer (kleiner) als ihr Durchschnitt sind, zusammen auftreten. Im Falle einer *negativen* Assoziation werden die Produkte aus einem analogen Grund großteils negativ sein.

Der **Korrelationskoeffizient** r_{xy} der Stichprobe (x_i, y_i) , $i = 1, 2, \dots, n$, ist nun der „Durchschnitt“ dieser Produkte:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Mit der (empirischen) **Kovarianz**:

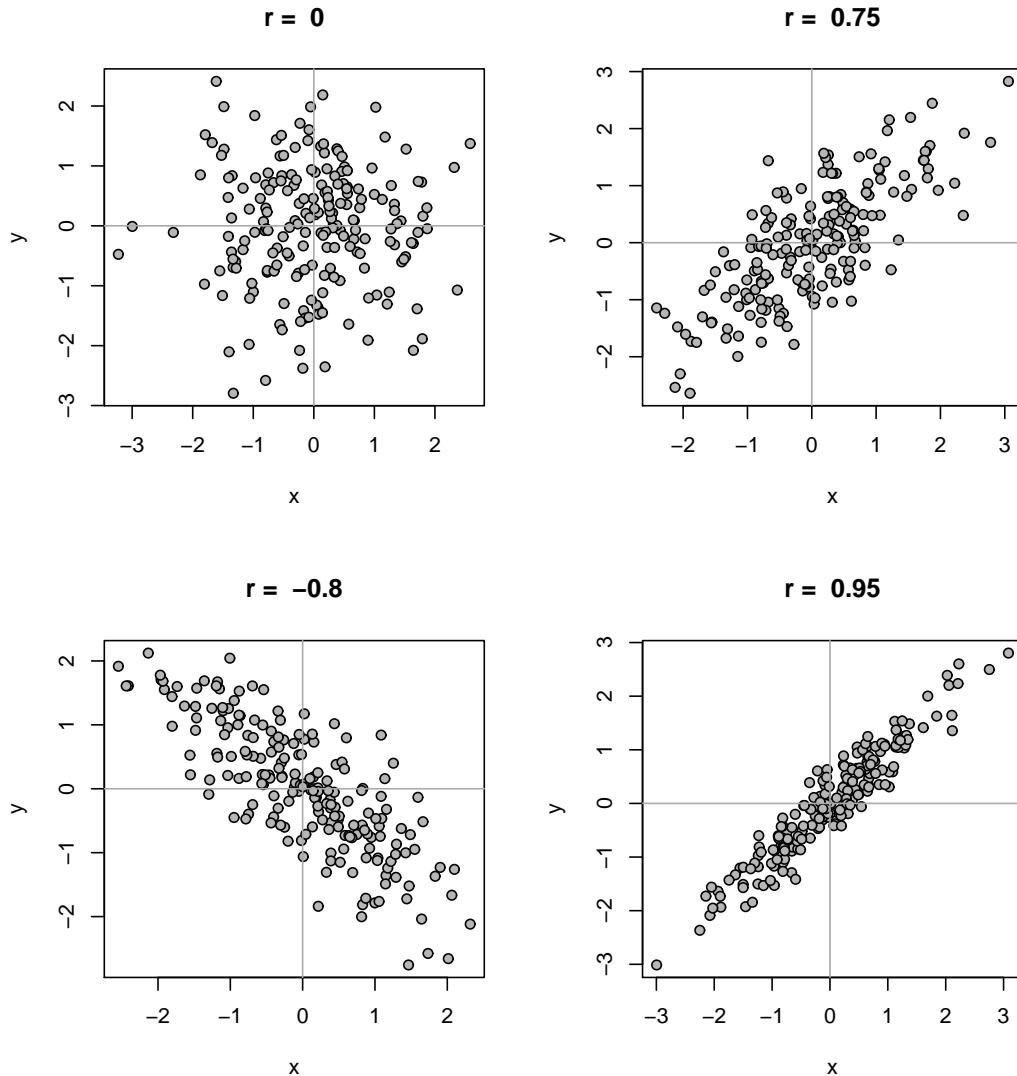
$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

lässt sich der Korrelationskoeffizient auch wie folgt schreiben:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Bem: In den obigen Formeln wird der bereits von der Varianz her bekannte Faktor $1/(n-1)$ (und nicht das vielleicht einleuchtendere $1/n$) verwendet. Das hat zur Folge, dass r stets im Bereich $-1 \leq r \leq 1$ liegt.

¹³Auch *Produkt-Moment-Korrelation* oder (*Bravais-*) *Pearson-Korrelation* genannt.

Abbildung 1.27: Simulierte Beobachtungen mit vorgegebenem r 

Ein numerisch stabilerer Ausdruck für die Berechnung von r_{xy} lautet:

$$r_{xy} = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{\sqrt{\left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] \left[\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right]}}$$

Für die Daten von Abb 1.26 ergibt sich ein Korrelationskoeffizient von $r = 0.5347$. Das bestätigt unseren Eindruck von einer (mittleren) positiven Assoziation zwischen Height und Weight. In Abb 1.27 sind einige weitere typische Situationen dargestellt, wobei in allen Fällen $\bar{x} = \bar{y} = 0$ und $s_x = s_y = 1$.

Der Korrelationskoeffizient ist symmetrisch in den Variablen (d. h. $r_{xy} = r_{yx}$) und es gilt $|r_{xy}| \leq 1$. Letzteres ist eine unmittelbare Folge der *Cauchy-Schwarz'schen Ungleichung*, die im vorliegenden Kontext besagt, dass:

$$\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right]^2 \leq \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]$$

Interpretation:

- (1) Das *Vorzeichen* von r sagt etwas über die Richtung der Assoziation. Ein positiver Wert signalisiert eine positive (oder gleichsinnige) Assoziation: Ist ein Merkmal größer als der Durchschnitt, ist das andere Merkmal tendenziell ebenfalls größer als der Durchschnitt. Ein negativer Wert signalisiert eine negative (oder gegensinnige) Assoziation: Ist ein Merkmal größer als der Durchschnitt, ist das andere Merkmal tendenziell kleiner als der Durchschnitt.
- (2) Der *Absolutwert* von r sagt etwas über die Stärke der Assoziation. Für $r = +1$ liegen alle Punkte (x_i, y_i) exakt auf einer Geraden mit positivem Anstieg (d. h. $y_i = a + bx_i$, $i = 1, 2, \dots, n$, mit $b > 0$). Für $r = -1$ liegen alle Punkte exakt auf einer Geraden mit negativem Anstieg ($b < 0$). Umso näher bei 0 der Wert von r liegt, umso schwächer ist die lineare Assoziation.
- (3) Der Korrelationskoeffizient ist ein Maß für die *lineare Assoziation*. Andere, kompliziertere Formen der Assoziation werden von ihm nicht (ausreichend) erfasst. Zur Veranschaulichung stelle man sich vor, dass alle Punkte äquidistant exakt auf einer Kreislinie liegen. Das bedeutet einen perfekten (deterministischen) Zusammenhang, der allerdings *nichtlinearer* Natur ist. Da aber in allen Quadranten gleich viele Punkte liegen, ist $r = 0$ (d. h. unkorreliert). Eine Lehre aus diesem Beispiel besteht darin, dass man Daten immer grafisch darstellen sollte; ein Scatterplot vermittelt deutlich mehr Information als eine einzelne Zahl.
- (4) *Korrelation* ist nicht gleichbedeutend mit *Kausalität*. Der Umstand, dass zwei Merkmale korrelieren bedeutet nicht notwendigerweise, dass auch eine Ursache-Wirkungsbeziehung zwischen ihnen besteht.¹⁴ Man denke etwa an das obige Beispiel, bei dem es offensichtlich ist, dass das Körpergewicht nicht die Körpergröße (oder umgekehrt) „verursacht“. Beide Merkmale stehen in einer gleichsinnigen Beziehung, nicht mehr und nicht weniger. Auch wenn zwei Merkmale hoch korrelieren, muss es keinen direkten Zusammenhang geben. Möglicherweise ist eine *dritte Variable*¹⁵ im Spiel, die beide Merkmale beeinflusst. Häufige *Confounder* sind etwa „Zeit“ oder „Alter“. (Bem: In diesen Bereich fallen auch die zahlreichen „Nonsense“-Korrelationen, wie etwa zwischen der Zahl der Störche und der Zahl der Geburten.)
- (5) Bei der Beurteilung von Korrelationen ist auch zu beachten, dass die *Beobachtungsbereiche* der beiden Merkmale möglicherweise zu schmal sind, um einen über breiteren

¹⁴Entspricht dem (logischen) *Fehlschluss cum hoc ergo propter hoc* („mit diesem, also wegen diesem“).

¹⁵engl. *confounding factor* (oder *confounder*), *hidden* oder *lurking variable*

Beobachtungsbereichen zutage tretenden Zusammenhang zu erkennen. Hätten wir beispielsweise beim `Weight-Height`-Beispiel dieses Abschnitts nur die Daten von Männern, deren Körpergröße zwischen 175 und 185 cm liegt, zur Verfügung, würde die Korrelation zwischen `Weight` und `Height` von 0.53 auf 0.12 sinken. Der Beobachtungsbereich wäre in diesem Fall zu schmal, um den (ohnehin nicht sehr starken) Zusammenhang zwischen den beiden Merkmalen zu erfassen.

- (6) Auch wenn, wie unter Punkt (4) diskutiert, Korrelation nicht notwendigerweise auch Kausalität bedeutet, so ist es bei Vorliegen einer ausgeprägten Korrelation dennoch möglich, aus Kenntnis der Werte der einen Variablen *Prognosewerte* für die andere Variable zu gewinnen. Das führt zur *Regressionsrechnung*, die in Kapitel 9 noch ausführlicher behandelt wird.

Kovarianz- und Korrelationsmatrix: Bei zwei oder mehr (metrischen) Merkmalen kann man alle *paarweisen* Kovarianzen und Korrelationskoeffizienten bestimmen und in Form einer Matrix anordnen. Ist \mathbf{I} die (entsprechend dimensionierte) Einheitsmatrix, $\mathbf{1} = (1, 1, \dots, 1)'$ der Einsvektor und \mathbf{H} die *Zentriermatrix*:

$$\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}'$$

so gilt mit $\mathbf{D} = \text{diag}(s_1, s_2, \dots, s_p)$ (mit s_i = Streuung des i -ten Merkmals oder der i -ten Spalte der Datenmatrix \mathbf{X}):

$$\text{Kovarianzmatrix: } \mathbf{S} = \frac{1}{n-1} \mathbf{X}' \mathbf{H} \mathbf{X}$$

$$\text{Korrelationsmatrix: } \mathbf{R} = \mathbf{D}^{-1} \mathbf{S} \mathbf{D}^{-1}$$

Beide Matrizen sind symmetrisch und positiv (semi)definit.¹⁶

Bsp 1.27 Der folgende R-Output zeigt die paarweisen (Pearson'schen) Korrelationskoeffizienten für alle metrischen Merkmale des Datensatzes `body.txt` (für `Gender = 1`), gerundet auf vier Stellen.

In der Diagonale stehen überall Einser (da $r_{xx} = 1$); das (4, 5) (oder (5, 4)) -Element (0.53) ist die uns schon bekannte Korrelation zwischen `Weight` und `Height`. Eine noch deutlich höhere positive Korrelation von etwa 0.81 besteht zwischen `Waist` (= Tailenumfang) und `Weight`. Die beiden negativen Korrelationen sind nur ganz schwach ausgeprägt.

¹⁶Allgemein ist eine $(p \times p)$ -Matrix \mathbf{A} *positiv semidefinit*, wenn für alle $\mathbf{x} \in \mathbb{R}^p$ gilt: $\mathbf{x}' \mathbf{A} \mathbf{x} \geq 0$; gilt die strikte Ungleichung für alle $\mathbf{x} \neq \mathbf{0}$, ist die Matrix *positiv definit*. Analog sind *negativ (semi)definite* Matrizen definiert. Ist $\mathbf{x}' \mathbf{A} \mathbf{x}$ sowohl positiv als auch negativ, ist die Matrix *indefinit*.

```
round(cor(datm[,1:5]), 4)
      Biacromial    Waist     Age   Weight   Height
Biacromial  1.0000  0.1757 -0.1010  0.4167  0.4765
Waist        0.1757  1.0000  0.4571  0.8051  0.2059
Age          -0.1010  0.4571  1.0000  0.1444 -0.0374
Weight       0.4167  0.8051  0.1444  1.0000  0.5347
Height       0.4765  0.2059 -0.0374  0.5347  1.0000
```

■

Spearman'sche Rangkorrelation: n Beobachtungspaare (x_i, y_i) , $i = 1, 2, \dots, n$, zu einem 2-dimensionalen *Rangmerkmal* seien gegeben. Gibt es keine Bindungen und sind (k_i, l_i) , $i = 1, 2, \dots, n$, die Rangzahlpaare, so ist der **Spearman'sche Rangkorrelationskoeffizient** definiert durch:

$$r_s = 1 - \frac{6 \sum_{i=1}^n (k_i - l_i)^2}{n(n^2 - 1)}$$

Im Falle von Bindungen verwendet man eine modifizierte Definition: Ist a (bzw. b) die Zahl der Bindungen in den Rangzahlen der x -Werte (bzw. y -Werte) und sind t_1, t_2, \dots, t_a (bzw. w_1, w_2, \dots, w_b) die Ausmaße der Bindungen (d. h. die jeweiligen Anzahlen der gleichen Rangzahlen), so definiert man:

$$r'_s = 1 - \frac{6 \sum_{i=1}^n (k_i - l_i)^2}{n(n^2 - 1) - (T_s + W_s)}$$

$$\text{mit } T_s = \frac{1}{2} \sum_{j=1}^a t_j(t_j^2 - 1) \quad \text{und } W_s = \frac{1}{2} \sum_{j=1}^b w_j(w_j^2 - 1)$$

Bemerkungen:

- (a) In beiden Fällen (d. h. mit und ohne Bindungen) entspricht der Spearman'sche Rangkorrelationskoeffizient dem Pearson'schen Korrelationskoeffizienten der Rangzahlen. Daraus folgt, dass $|r_s| \leq 1$.
- (b) Es gilt $r_s = +1$, wenn beide Rangfolgen exakt übereinstimmen; es gilt $r_s = -1$, wenn die eine Rangfolge die exakte Umkehrung der anderen ist.

- (c) Im Unterschied zum Pearson'schen Koeffizienten (wie oben diskutiert, ein Maß für die Stärke des *linearen* Zusammenhangs) misst der Spearman'sche Koeffizient die Stärke des **monotonen** Zusammenhangs zwischen zwei (Rang-) Merkmalen.
- (d) Der Spearman'sche Korrelationskoeffizient eignet sich als *robustes* Korrelationsmaß. Durch den Verzicht auf die metrische Information in den Daten und die alleinige Verwendung der Ränge wird der Einfluss von (potenziellen) Ausreißern reduziert.

Bsp 1.28 Angenommen, elf Student/inn/en erreichen in zwei Fächern (z. B. Mathematik und Physik) bei einem Test die folgenden Punktzahlen:

Student/in	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
x	41	37	38	39	49	47	42	34	36	48	29
y	36	20	31	24	37	35	42	26	27	29	23

Der folgende R-Output zeigt die Berechnung des Pearson'schen und des Spearman'schen Korrelationskoeffizienten:

```

x <- c(41, 37, 38, 39, 49, 47, 42, 34, 36, 48, 29)
y <- c(36, 20, 31, 24, 37, 35, 42, 26, 27, 29, 23)
cor(x, y)
[1] 0.6295419                                <--- Pearson
cor(x, y, method="spearman")
[1] 0.7181818                                <--- Spearman
(rx <- rank(x))
[1] 7 4 5 6 11 9 8 2 3 10 1      <--- Rangzahlen (x)
(ry <- rank(y))
[1] 9 1 7 3 10 8 11 4 5 6 2      <--- Rangzahlen (y)
cor(rx, ry)
[1] 0.7181818                                <--- Pearson der Rangzahlen

```

In diesem Fall können beide Korrelationskoeffizienten berechnet und sinnvoll interpretiert werden. Wie sich zeigt, ist die Rangkorrelation etwas höher als die Produkt-Moment-Korrelation, d. h., der monotone Zusammenhang ist stärker als der nur lineare. Außerdem wird demonstriert, dass der Spearman'sche Koeffizient tatsächlich der Pearson'sche Koeffizient der Rangzahlen ist. ■

1.9.4 Kleinstes Quadrat

In Abb 1.26 wurde zusätzlich zu den Punkten auch die „Kleinstes-Quadrat-Gerade“ eingezeichnet, die unter allen möglichen Geraden eine bestimmte Optimalitätseigenschaft aufweist. In diesem Abschnitt wollen wir klären, was darunter zu verstehen ist.

Angenommen, an n Punkte (x_i, y_i) , $i = 1, 2 \dots, n$, soll die Kurve $y = h(x; \alpha, \beta)$, die von zwei Parametern α und β abhängt, „bestmöglich“ angepasst werden.¹⁷ Der y -Wert der Kurve an der Stelle x_i ist $h(x_i; \alpha, \beta)$ und der y -Wert des beobachteten Punktes ist y_i . Die Abstand der beiden y -Werte beträgt $d_i = [y_i - h(x_i; \alpha, \beta)]$ und das Quadrat d_i^2 des Abstands ist ein Maß für die Güte der Anpassung an der Stelle x_i . Bei der **Methode der kleinsten Quadrate** werden nun die Parameter α und β so bestimmt, dass die Summe der Abstandsquadrate minimal wird:

$$S(\alpha, \beta) = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n [y_i - h(x_i; \alpha, \beta)]^2 \longrightarrow \text{Min!}$$

Im speziellen Fall einer **Ausgleichsgeraden**, d. h. wenn $h(x; \alpha, \beta) = \alpha + \beta x$, lautet das Minimierungsproblem wie folgt:

$$S(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \longrightarrow \text{Min!}$$

Bildet man die partiellen Ableitungen und setzt sie gleich Null:

$$\frac{\partial S(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-1) = 0$$

$$\frac{\partial S(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-x_i) = 0$$

bekommt man zwei lineare Gleichungen (die **Normalgleichungen**¹⁸):

$$\sum_{i=1}^n y_i = n\alpha + \left(\sum_{i=1}^n x_i \right) \beta$$

$$\sum_{i=1}^n x_i y_i = \left(\sum_{i=1}^n x_i \right) \alpha + \left(\sum_{i=1}^n x_i^2 \right) \beta$$

Als Lösung für β (= Anstieg der Geraden) ergibt sich:

¹⁷Ein anderer Ausdruck dafür lautet, dass die Punkte (x_i, y_i) durch die Kurve $h(x; \alpha, \beta)$ „ausgeglichen“ werden sollen (→ *Ausgleichsrechnung*).

¹⁸Die Bezeichnung verdankt sich dem Umstand, dass die KQ-Lösung aus algebraischer Sicht einer *orthogonalen Projektion* auf einen linearen Unterraum entspricht.

$$\hat{\beta} = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

Nach einfachen Umformungen lässt sich $\hat{\beta}$ auch wie folgt darstellen:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$\hat{\beta}$ steht in enger Beziehung zum Korrelationskoeffizienten r_{xy} :

$$\hat{\beta} = r_{xy} \frac{s_y}{s_x}$$

Dividiert man die erste Normalgleichung durch n , ergibt sich die Lösung für α (= Achsenabschnitt, Interzept):

$$\bar{y} = \hat{\alpha} + \bar{x} \hat{\beta} \implies \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Die im Sinne der kleinsten Abstandsquadrate bestmöglich angepasste Gerade ist also gegeben durch:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x = \bar{y} + \hat{\beta}(x - \bar{x})$$

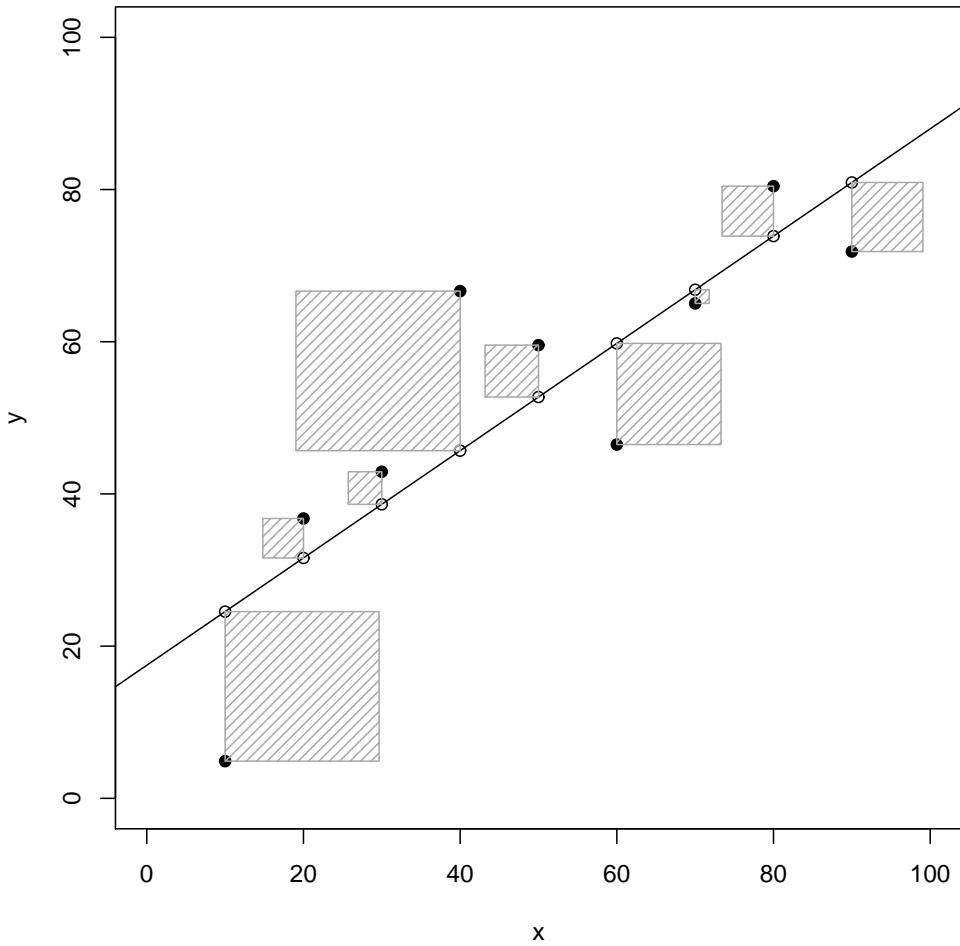
Daraus folgt insbesondere, dass die **KQ–Gerade** (auch **(O)LS–Gerade**¹⁹ durch den Mittelpunkt (\bar{x}, \bar{y}) der Daten verläuft. Für die Daten von Abb 1.26 ergibt sich:

$$\hat{\alpha} = -60.9534 \quad \text{und} \quad \hat{\beta} = 0.7826$$

D.h., jeder zusätzliche Zentimeter an Körpergröße geht mit einer Zunahme von 0.78 kg an Körergewicht einher. Man beachte, dass wir es hier mit Körpergrößen zwischen etwa 155 und 200 cm zu tun haben, der Interzept $\hat{\alpha}$ für sich genommen also keine realistische Interpretation hat (sondern nur der Definition der KQ–Geraden dient).

¹⁹engl. (*ordinary*) least squares

Abbildung 1.28: Prinzip der kleinsten Quadrate



Bsp 1.29 Zur Veranschaulichung des KQ–Prinzips betrachten wir einen simulierten Datensatz aus 9 Punkten (Abb 1.28). Die dick gezeichneten Punkte sind die Beobachtungen und die Gerade entspricht der KQ–Geraden. Die offen gezeichneten Punkte sind die Punkte (x_i, \hat{y}_i) , wobei $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ der auf Basis der KQ–Geraden an der Stelle x_i prognostizierte Wert von y ist. Die von den beiden Punkten aufgespannten Quadratflächen entsprechen den Abstandsquadrate; für die KQ–Gerade ist die Summe dieser Flächen minimal.

Die Differenzen zwischen den tatsächlichen Beobachtungen und den prognostizierten Werten, $e_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, n$, sind die **Residuen**. Aus den Normalgleichungen folgt, dass die Residuen die beiden folgenden Bedingungen erfüllen:

$$\sum_{i=1}^n e_i = 0 \quad \text{und} \quad \sum_{i=1}^n e_i x_i = 0$$

Die Residuen spielen (u. a.) eine wichtige Rolle bei der Beurteilung der Anpassungsgüte der KQ–Geraden. ■

Bem: Die Anpassung einer Geraden (oder von anderen Kurven) an vorgegebene Daten wird in der Statistik unter der Überschrift **Regressionsanalyse** behandelt. Man „regressiert“ eine *Antwortvariable y* (z. B. *Weight*) auf eine *erklärende Variable x* (z. B. *Height*). Die Regressionsanalyse (mit ihren zahlreichen Varianten und Erweiterungen) gehört zu den Kernmethoden der Statistik und wird in Kapitel 9 noch etwas ausführlicher behandelt. Hier betrachten wir die Regressionsanalyse als deskriptive (bzw. explorative) Methode zur Beschreibung eines Datensatzes.²⁰

Aufgaben

- 1.1 Der Datensatz `TempWien1951-2012.txt` (vgl. Bsp 1.4) enthält neben den Höchst- auch die Tiefsttemperaturen und die Jahresdurchschnitte für Wien/Hohe Warte für die Jahre 1951–2012. Stellen Sie die drei Zeitreihen gemeinsam in einem Plot und in einzelnen Plots dar. Überlagern Sie die letzteren Plots mit gleitenden Durchschnitten der Spannweite $w = 10$.
- 1.2 Von der ACEA (*European Automobile Manufacturers' Association; www.acea.be*) werden u. a. Daten über Neuzulassungen von Kraftfahrzeugen gesammelt. Für das Jahr 2011 ergab sich für die PKW–Neuzulassungen das folgende Bild, aufgeschlüsselt nach Herstellergruppen (Zahlen für Westeuropa; Datenfile: `pkw-neuzul11.txt`):

GROUP	TOTAL
ASTON MARTIN	2310
BMW	791658
CHINA	1659
DAIMLER	659268
FIAT	915237
FORD	1033030
GM	1099194
IVECO	704
JAGUAR LAND ROVER	93025
JAPAN	1011765
HYUNDAI	353823
KIA	251334
KOREA	7085
PORSCHE	40714
PSA	1619704
RENAULT	1194752
TOYOTA	520090
VOLKSWAGEN	2939136
OTHER	272904

(Bem: Die Herstellergruppe JAPAN umfasst die Marken Daihatsu, Honda, Mazda, Mitsubishi, Nissan, Subaru, Suzuki und andere.)

²⁰Man beachte, dass durch die Geradenanpassung eine *Dimensionsreduzierung* erfolgt; n Datenpunkte werden durch *zwei* Parameter (α und β) beschrieben.

Fassen Sie Herstellergruppen mit einem Anteil von weniger als 3% mit der Gruppe OTHER zusammen und erstellen Sie ein Kreisdiagramm. Für eine bessere Lesbarkeit des Diagramms empfiehlt sich eine Darstellung nach der Größe der Anteile.

- 1.3 Erstellen Sie Pareto-Diagramme (a) für die Neuinskriptionen an der TU-Wien für das WS 2013 und (b) für die PKW-Neuzulassungen von **Aufgabe 1.2**. Interpretieren Sie die Diagramme.
- 1.4 Ein Hersteller von mikroelektronischen Komponenten benötigt bestimmte keramische Platten. Eine Stichprobe des Umfangs $n = 30$ aus einem größeren Los von derartigen Platten erbrachte die folgenden Fehlerzahlen pro Platte:

0, 2, 0, 0, 1, 3, 0, 3, 1, 1, 0, 0, 1, 2, 0
0, 0, 1, 1, 3, 0, 1, 0, 0, 0, 5, 1, 0, 2, 0

Zeichnen Sie das Balkendiagramm und die **Summentreppe**, d. i. eine treppenförmige Darstellung der kumulierten relativen Häufigkeiten. (Bem: Hier handelt es sich um ein **Zählmerkmal**.)

- 1.5 Bestimmen Sie für das Merkmal **Waist** (Datenframe: `body.txt`) die empirische Verteilungsfunktion für beide Geschlechter zusammen und getrennt.
- 1.6 Bestimmen Sie einen Stem-and-Leaf-Plot für das Merkmal **Biacromial** (Datenframe: `body.txt`) für `Gender = 0`. Zusatz: Erstellen Sie einen *Back-to-Back* Stem-and-Leaf-Plot für `Gender = 0` und `Gender = 1`. (Hinweis: Nehmen Sie die Funktion `stem.leaf.backback()` aus dem Package `aptpack`.)
- 1.7 Der Datensatz **euroweight.txt** umfasst für acht Batches zu jeweils 250 Stück Messwerte des Gewichts von neuen (belgischen) 1€-Münzen.²¹ Zeichnen Sie – angeordnet in einem 4×2 -Array – für alle acht Batches flächentreue Histogramme; nehmen Sie dazu die folgende (gemeinsame) Klasseneinteilung:

$(7.200, 7.210], (7.210, 7.220], \dots, (7.750, 7.760]$

Überlagern Sie die Histogramme mit Kerndichteschätzungen. Kommentieren Sie die Ergebnisse.

- 1.8 Bestimmen Sie – getrennt nach Geschlecht – für das Merkmal **Biacromial** (Datenframe: `body.txt`):
 - (a) den Box- und den Violinplot
 - (b) die 5(6)-Zahlen-Zusammenfassung
 - (c) die Varianz, die Streuung, den MAD

²¹Z. SHKEDY, M. AERTS, AND H. CALLAERT: The Weight of Euro Coins: Its Distribution Might Not Be As Normal As You Would Expect, *Journal of Statistics Education*, Vol. 14/2, 2006.

- 1.9 Bestimmen Sie für großes n den Bruchpunkt der Hinges. (Hinweis: Nehmen Sie als „Datensatz“ beispielsweise die Zahlen von 1 bis 100 und überlegen Sie sich, wieviele Datenpunkte man ändern müsste, um den unteren (oder den oberen) Hinge beliebig zu verändern.)
- 1.10 Laut der Homepage von Eisenstadt/Bgl. entwickelten sich die Einwohnerzahlen von 1951 bis 2011 wie folgt:

Jahr	1951	1961	1971	1981	1991	2001	2011
Bev.	7.568	9.315	10.062	10.102	10.349	11.334	12.995

- (a) Stellen Sie die Zeitreihe grafisch dar (z. B. als Balkendiagramm).
- (b) Wie groß ist die durchschnittliche 10-jährliche Zunahme (in %) und die durchschnittliche jährliche Zunahme (in %)? Wie sind diese Durchschnittswerte zu interpretieren?
- (c) Wenn man die Entwicklung von 2001 auf 2011 zugrunde legt, mit welcher Bevölkerungszahl kann man im Jahr 2030 rechnen?

- 1.11 Ein Handelsbetrieb unterhält in einer Stadt vier Filialen. Bekannt seien für jede Filiale der Anteil am Gesamtumsatz sowie der durchschnittliche Jahresumsatz pro m^2 Verkaufsfläche:

Filiale	Umsatzanteil	Umsatz/ m^2
1	10%	35.000 €
2	20%	42.000 €
3	50%	52.500 €
4	20%	28.000 €

Bestimmen Sie den durchschnittlichen Jahresumsatz pro m^2 Verkaufsfläche für alle Filialen der Stadt zusammen.

- 1.12 Zeigen Sie den Verschiebungssatz für die Stichprobenvarianz:

$$s_n^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n(\bar{x}_n)^2 \right]$$

- 1.13 Eine Stichprobe aus ganzzahligen Werten vom Umfang 72 hat den Modus 54, den Median 54.5 und den Mittelwert 55.7. Eine zusätzliche Beobachtung hat den Wert $x_{73} = 56$. Was lässt sich über den Modus, den Median und den Mittelwert der erweiterten Stichprobe sagen?

- 1.14 Meist bevorzugt man die Streuung s_n eines Datensatzes als Streuungsmaß gegenüber der Varianz s_n^2 , da s_n die gleiche Einheit wie der Mittelwert \bar{x}_n hat. Gelegentlich bevorzugt man aber ein *dimensionsloses* Streuungsmaß. Der **Variationskoeffizient** (kurz **VK**) ist definiert durch $\text{VK}_n = s_n / \bar{x}_n$. Ein Vorteil dieses Streuungsmaßes

besteht darin, dass es unserer (intuitiven) Vorstellung von „Streuung“ meist eher entspricht als etwa s_n . Beispielsweise haben 1, 2, 3, 4 und 1001, 1002, 1003, 1004 zwar die gleiche Standardabweichung (1.291) aber sehr unterschiedliche VK's (0.5164 bzw. 0.0013). Das korrespondiert mit dem Eindruck, dass die zweiten Werte „näher beieinander“ liegen als die ersten.

Bestimmen Sie für die Variable **Height** (Datenfile: **body.txt**) die Standardabweichung und den Variationskoeffizienten für **Gender = 0** und **Gender = 1**. Wie beurteilen Sie das Streuverhalten der beiden Datensätze?

- 1.15 Zeichnen Sie auf Basis der Daten von **body.txt** vergleichende Boxplots sowie Histogramme (plus Kernschätzung) für den BMI (*Body Mass Index* = Gewicht[kg]/(Größe[m])²) für **Gender = 0** und **Gender = 1**. Berechnen Sie Kennzahlen der Lage und der Streuung und beschreiben Sie die Verteilungsform. **Zusatz:** Ein BMI ab 25 kg/m² gilt bereits als gesundheitlich problematisch. Auf Basis des vorliegenden Datensatzes, welcher Anteil bei Männern und Frauen übersteigt diesen Wert? (Bem: Der Mikrozensus²² 2007 erbrachte in der Bevölkerung ab 15 Jahren für Österreich die folgenden Ergebnisse: 54.5% der Männer haben einen BMI von mehr als 25 kg/m², bei den Frauen liegt dieser Anteil bei 41.3%).)
- 1.16 Der Datensatz **brightness** (Package: **UsingR**) umfasst Daten zur Helligkeit von 996 Sternen (Leuchtkraft im sichtbaren Spektralbereich; je kleiner der Wert umso heller der Stern) in einem bestimmten Himmelssektor. Die Daten stammen aus dem sogenannten Hipparcos Katalog.²³ Erstellen Sie ein Histogramm (überlagert mit einer Kerndichteschätzung) und berechnen Sie Koeffizienten der Schiefe und Kurtosis. Kommentieren Sie die Verteilungsform.
- 1.17 Berechnen Sie analog zu Bsp 1.27 alle paarweisen Pearson'schen Korrelationskoeffizienten für die metrischen Merkmale von **body.txt** für **Gender = 0**. Kommentieren Sie die Ergebnisse.
- 1.18 Betrachten Sie die hoch korrelierenden Merkmale **Waist** und **Weight** aus dem Datensatz **body.txt**. Zeichnen Sie den Scatterplot und bestimmen Sie die KQ–Gerade. Zeichnen Sie Letztere in den Scatterplot ein. Unterscheiden Sie dabei nach Geschlecht. Interpretieren Sie die Ergebnisse.
- 1.19 Wie lautet allgemein die KQ–Lösung für den Geradenanstieg β unter der Bedingung, dass die Gerade durch den Nullpunkt verläuft, d. h. für eine Gerade der Form $y = \beta x$? Bestimmen Sie $\hat{\beta}$ konkret für den Datensatz:

x	3	1	5	6	3	4
y	4	2	4	8	6	5

Zeichnen Sie den Scatterplot und die KQ–Gerade durch den Nullpunkt. Bestimmen (und zeichnen) Sie außerdem die uneingeschränkte KQ–Gerade, d. h. die Gerade der Form $y = \hat{\alpha} + \hat{\beta}x$.

²²Stichprobenerhebung, bei der pro Quartal rund 22500 zufällig ausgewählte Haushalte in ganz Österreich befragt werden; jeder Haushalt bleibt für insgesamt fünf Quartale in der Stichprobe.

²³Vgl. z. B. WIKIPEDIA für weitere Informationen (Stichworte: Hipparcos, UVB Photometric System).

- 1.20 Passen Sie nach der KQ–Methode eine Kurve der Form $y = \alpha + \beta x^2$ an die folgenden Daten an:

x	-2	3	-1	0	-3	1	5	-3
y	7	15	3	1	11	6	20	16

Zeichnen Sie den Scatterplot und die KQ–Parabel. (Zusatz: Wie lautet in diesem Fall die allgemeine KQ–Lösung?)

Hinweis: Die folgenden R–Commands führen zum Ziel:

```

x <- c(-2,3,-1,0,-3,1,5,-3)
y <- c(7,15,3,1,11,6,20,16)
plot(y ~ x, type="p", pch=21, bg="lightblue3",
      xlim=c(-5,5), ylim=c(0,25), cex=2, main="KQ - Parabel")
mod <- lm(y ~ I(x^2))
coef(mod) # <<-- KQ-Lösung
xnew <- data.frame(x=seq(-5, 5, by=0.1))
ypred <- predict(mod, newdata=xnew, interval="n")
lines(xnew$x, ypred, lwd=2, col="lightblue3")

```

2 Wahrscheinlichkeit

Wir leben in einer Welt voller zufallsbedingter Unsicherheiten. Dabei bemühen wir den „Zufall“ nicht nur zur Beschreibung vieler Phänomene des Alltags oder wenn wir an einem Glücksspiel teilnehmen, sondern er erweist sich bei genauerem Hinsehen bald als integraler Bestandteil unseres gesamten Naturverständnisses. Beispielsweise können viele Phänomene im Bereich der Elementarteilchen ohne Zuhilfenahme von Modellen aus der Wahrscheinlichkeitstheorie nicht adäquat beschrieben oder interpretiert werden.

Es liegt in der Natur der Sache, dass mehrere „Zufallsmodelle“ vorstellbar sind. Speziell im Bereich der Naturwissenschaften und der Technik ist es aber vorteilhaft, sich bei der Modellentwicklung von Erfahrungen mit „Zufallsexperimenten“ leiten zu lassen. Dabei versteht man unter einem **Zufallsexperiment** allgemein einen zufallsbehafteten Vorgang, dessen Ausgang mehr oder weniger unsicher oder nicht deterministisch bestimmt ist. Ein typisches Beispiel ist etwa das Werfen einer Münze oder eines Würfels.

2.1 Gesetz der großen Zahlen

Die einfachste Form der Beschreibung von Zufallsexperimenten ist das **Zählen**. Man zählt, wie oft ein bestimmtes Ereignis A bei wiederholter Durchführung eines Zufallsexperiments eingetreten ist. Beispielsweise kann man zählen, bei wievielen Patienten ein bestimmtes Medikament eine Besserung bewirkt hat, oder wie oft in den vergangenen 20 Jahren der August verregnelt war, oder wie oft beim Werfen einer 1€–Münze „Zahl“ vorkommt, usw.

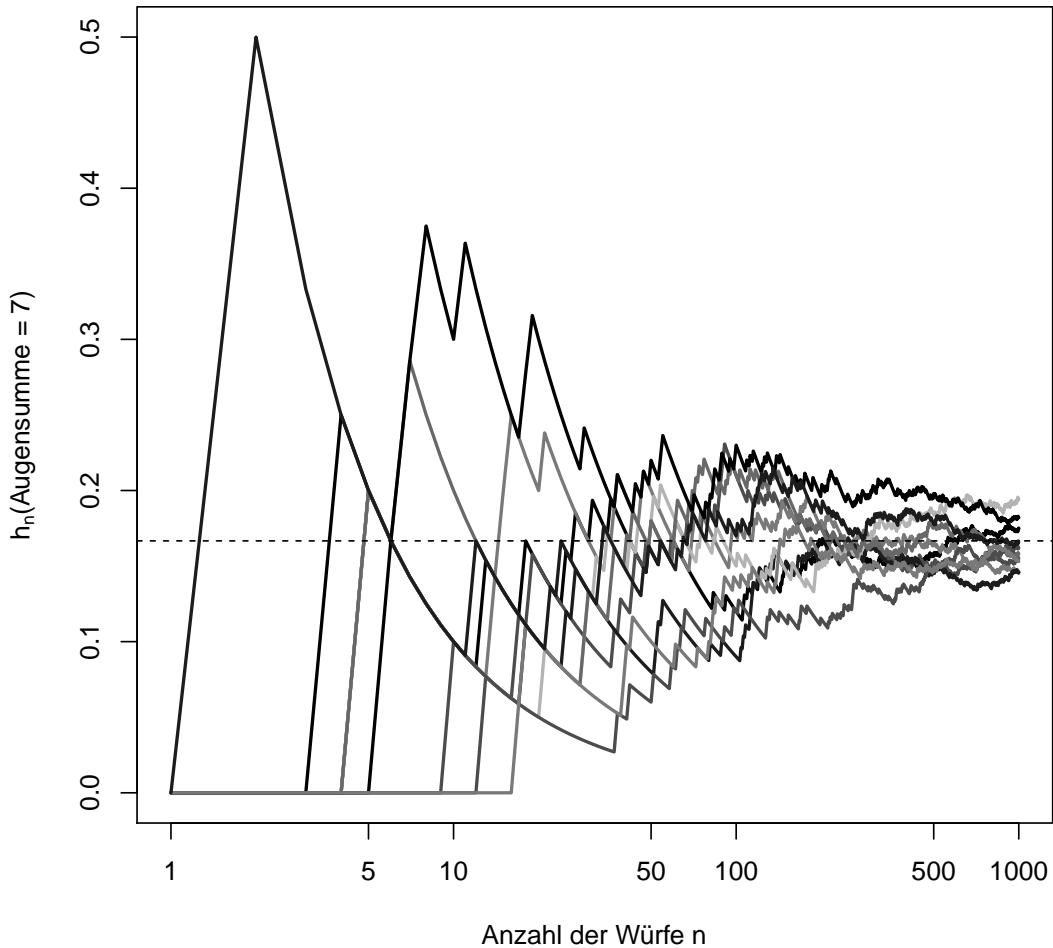
Betrachtet man n Wiederholungen eines Zufallsexperiments und tritt das fragliche Ereignis $H_n(A)$ -mal auf, so hat man häufig den Eindruck, dass sich die **relative Häufigkeit** $h_n(A) := H_n(A)/n$ von A einem Grenzwert nähert:

$$\lim_{n \rightarrow \infty} h_n(A) =: P(A)$$

Es liegt nahe, den Grenzwert $P(A)$ als „Wahrscheinlichkeit“ (des Eintritts) von A zu betrachten. Diese Grenzwertvermutung nennt man das (**empirische**) **Gesetz der großen Zahlen** (eGGZ).

Bsp 2.1 Zur Illustration des eGGZ betrachten wir das Werfen von zwei gleichartigen (ausgewogenen) Würfeln und speziell das Ereignis, dass die Augensumme gleich sieben ist. Abb 2.1 zeigt das Ergebnis von 10 simulierten Wurffolgen zu je 1000 Würfen. (Zur klareren Darstellung nehmen wir eine logarithmische x –Achse.) Am Anfang ist die Fluktuation noch sehr hoch, mit größer werdendem n scheinen sich die relativen Häufigkeiten $h_n(\text{Augensumme} = 7)$ einem Grenzwert zu nähern. (Die strichlierte Linie ist bei der „klassischen“ Wahrscheinlichkeit für das fragliche Ereignis von 1/6.) Man beachte allerdings, dass selbst für $n = 1000$ die Fluktuation noch immer vergleichsweise hoch ist. ■

Abbildung 2.1: Empirisches Gesetz der großen Zahlen



Auch wenn man von der Gültigkeit des eGGZ überzeugt ist, lässt es sich nur unter Inkaufnahme einer Reihe von begrifflichen Schwierigkeiten zur Grundlage eines mathematisch konsistenten Wahrscheinlichkeitsbegriffs machen. Irgendwann muss man die Beobachtungen schließlich abbrechen (oder sie sind von vornherein nur beschränkt verfügbar), sodass man nie ganz sicher sein kann, ob tatsächlich „Konvergenz“ (und in welchem Sinn) vorliegt und ob sich bei einer Wiederholung des gesamten Experiments stets der gleiche Grenzwert einstellen wird.

Man kann allerdings die Eigenschaften von relativen Häufigkeiten zum Vorbild einer **axiomatischen** Definition von Wahrscheinlichkeit nehmen. Letzteren Zugang nennt man die **frequentistische** (oder **objektivistische**) Interpretation des Wahrscheinlichkeitsbegriffs (vgl. 2.5).

Subjektive Wahrscheinlichkeiten: Die frequentistische Interpretation von Wahrscheinlichkeit beruht darauf, dass ein (statistisches) Experiment unter (mehr oder weniger) identischen Bedingungen beliebig oft wiederholbar ist. Das ist keineswegs immer der Fall. Was bedeutet es beispielsweise, wenn jemand behauptet, zu 70% davon überzeugt zu sein, dass

William Shakespeare *Julius Caesar* geschrieben hat, und zu 10% dass es Christopher Marlowe war? Diese Form von **subjektiver** Wahrscheinlichkeit lässt sich nicht frequentistisch interpretieren. Vielmehr handelt es sich um einen – auf persönlicher Expertise basierenden – **Grad des Vertrauens** in eine Behauptung.

Hält man sich bei der Zuschreibung von subjektiven Wahrscheinlichkeiten an bestimmte konsistente und rationale Regeln (verhält sich quasi wie ein rational agierender Spieler), macht es aber – mathematisch gesehen – keinen Unterschied, ob „Wahrscheinlichkeit“ frequentistisch oder subjektivistisch interpretiert wird.

Bem: Der subjektive Wahrscheinlichkeitsbegriff bildet die Grundlage der sog. **Bayes–Statistik**, die in Kapitel 8 noch etwas ausführlicher behandelt wird.

2.2 Merkmalraum

Die möglichen Ergebnisse von statistischen Experimenten lassen sich in einer Grundmenge zusammenfassen. Die Menge aller möglichen Versuchsausgänge nennt man den **Merkmalraum**:¹

$$\Omega = \{\omega \mid \omega \text{ möglicher Versuchsausgang}\}$$

Ein Merkmalraum kann von vielfältiger Gestalt sein: Endlich, unendlich (abzählbar unendlich, überabzählbar), ein-, mehrdimensional, etc. Meist sind die Elemente von Ω mathematische Gebilde (Zahlen, Vektoren, Mengen, …), gelegentlich werden die Versuchsergebnisse aber auch nur verbal beschrieben (beispielsweise beim Münzwurf als „Kopf“ oder „Zahl“).

Man beachte, dass zur Beschreibung eines Experiments durchaus mehrere unterschiedliche Merkmalräume geeignet sein können (vgl. die folgenden Beispiele).

Beispiele:

1. Besteht das statistische Experiment in der Bestimmung des Geschlechts eines neugeborenen Babys, so ist etwa $\Omega = \{g, b\}$ geeignet, wobei $g(\text{irl})$ = Mädchen und $b(\text{oy})$ = Bub bezeichnet.
2. Besteht das statistische Experiment darin, bei einem Pferderennen, an dem sieben Pferde mit den Startnummern 1, 2, …, 7 beteiligt sind, die Reihenfolge des Zieleinlaufs zu bestimmen (oder zu beobachten), so besteht der Merkmalraum aus allen $7! = 5040$ möglichen Permutationen von $(1, 2, 3, 4, 5, 6, 7)$:

$$\Omega = \{(x_1, x_2, \dots, x_7) \mid x_i = 1, 2, \dots, 7; x_i \neq x_j \text{ für } i \neq j\}$$

¹Auch als *Grundraum*, *Grundgesamtheit*, *Stichprobenraum* (engl. *sample space*) o. ä. bezeichnet.

3. Wirft man zwei Münzen (wobei eine die „erste“ und die andere die „zweite“ ist), besteht der Merkmalraum aus vier Elementen:

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$$

Dabei bedeutet z. B. (T, H) , dass die erste Münze auf „Zahl“ und die zweite Münze auf „Kopf“ fällt.

4. Wirft man zwei (übliche) Würfel (wobei einer der „erste“ und der andere der „zweite“ ist) und beobachtet die geworfenen Augenzahlen, besteht der Merkmalraum aus 36 Punkten:

$$\Omega = \{(i, j) \mid i, j = 1, 2, 3, 4, 5, 6\}$$

Dabei bedeutet (i, j) , dass die Augenzahl des ersten Würfels gleich i und die des zweiten gleich j ist.

Bem: Interessiert nur die Augensumme, könnte man auch $\Omega = \{2, 3, \dots, 12\}$ als Merkmalraum nehmen. Ein Nachteil dieses Raumes besteht allerdings darin, dass die Elementarausgänge $\omega \in \Omega$ – in einem intuitiven Sinn – nicht „gleichwahrscheinlich“ sind, die des zuerst betrachteten Raumes aber schon.

5. Besteht das Experiment im Messen der Lebensdauer (in Betriebsstunden) eines Transistors, nimmt man als Merkmalraum ein halbunendliches Intervall:

$$\Omega = \{x \mid 0 \leq x < \infty\} = [0, \infty)$$

Bem: Da Transistoren bereits unmittelbar bei Inbetriebnahme ausfallen können, nimmt man meist $[0, \infty)$ und nicht $(0, \infty)$ als Merkmalraum. Andererseits: Lebensdauern sind endlich, also sollte man eher Intervalle der Form $[0, b)$ (mit $b < \infty$) als Merkmalraum nehmen. Meist ist es aber schwierig, einen Wert für b zu bestimmen, sodass man bei (technischen) Lebensdauerproblemen in der Regel $[0, \infty)$ als Merkmalraum nimmt. (Letztere Wahl vereinfacht auch die statistische Modellierung.)

6. Das statistische Experiment bestehe in der zufälligen Auswahl von k Objekten aus einer Menge von n (unterscheidbaren) Objekten, bezeichnet mit $M = \{1, 2, \dots, n\}$ ($k \leq n$). Die Auswahl erfolge in der Weise, dass ein einmal gewähltes Objekt nicht noch einmal gewählt werden kann. Für die Wahl eines passenden Merkmalraums hat man (zumindest) zwei Möglichkeiten. Spielt die Reihenfolge der Auswahl keine Rolle, nimmt man die Menge aller k -elementigen Teilmengen:

$$\Omega_1 = \{B \mid B \subseteq M, |B| = k\}, \quad |\Omega_1| = \binom{n}{k}$$

Möchte man die Reihenfolge der Auswahl berücksichtigen, nimmt man:

$$\Omega_2 = \{(x_1, x_2, \dots, x_k) \mid x_i \in M; x_i \neq x_j \text{ für } i \neq j\}, \quad |\Omega_2| = \frac{n!}{(n-k)!}$$

Beide Merkmalräume bestehen aus „gleichwahrscheinlichen“ Elementen.

2.3 Ereignisse

Allgemein nennt man eine Teilmenge $A \subseteq \Omega$ eines Merkmalraumes Ω ein **Ereignis**. Gilt für einen Versuchsausgang $\omega \in \Omega$, dass $\omega \in A$, so sagt man, dass A **eingetreten** ist. Alle Ereignisse werden in einem **Ereignissystem** \mathcal{A} zusammengefasst:

$$\mathcal{A} = \{A \mid A \subseteq \Omega \text{ ist ein Ereignis}\}$$

Neben einfachen Ereignissen A, B, \dots möchte man aber auch zusammengesetzte (oder abgeleitete) Ereignisse wie etwa „ A und B treten ein“ oder „ A ist nicht eingetreten“ betrachten. Das hat zur Folge, dass Ereignissysteme eine entsprechende *algebraische Struktur* aufweisen sollten.

Ereignissystem als σ -Algebra: Gegeben sei ein (nichtleerer) Merkmalraum Ω . Ein System \mathcal{A} von Teilmengen aus Ω heißt eine **σ -Algebra** über Ω , wenn es die folgenden Eigenschaften erfüllt:

- (1) $\Omega \in \mathcal{A}$ (d. h., der Merkmalraum selbst ist ein Ereignis)
- (2) Für $A \in \mathcal{A}$ gilt $A^c \in \mathcal{A}$ (d. h., \mathcal{A} ist abgeschlossen unter Komplementbildung²)
- (3) Für eine Folge A_1, A_2, \dots aus \mathcal{A} gilt $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ (d. h., \mathcal{A} ist abgeschlossen unter abzählbaren Vereinigungen)

Aus den Eigenschaften (1) und (2) folgt, dass auch $\emptyset \in \mathcal{A}$. Nach den **De Morgan'schen Regeln**:³

$$\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c \quad \text{und} \quad \left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c$$

folgt aus den Eigenschaften (2) und (3), dass eine σ -Algebra auch abgeschlossen gegenüber abzählbaren Durchschnitten ist:

$$A_1, A_2, \dots \in \mathcal{A} \implies \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$$

Ist Ω ein Merkmalraum und \mathcal{A} eine σ -Algebra über Ω , nennt man das Paar (Ω, \mathcal{A}) einen **Messraum** und die Elemente von \mathcal{A} nennt man **messbare Mengen**.

²Für das zu A komplementäre Ereignis A^c schreibt man auch \overline{A} ; im Folgenden werden beide Schreibweisen verwendet.

³AUGUSTUS DE MORGAN (1806–1871), engl. Mathematiker (zusammen mit GEORGE BOOLE Begründer der formalen Logik).

Bemerkungen:

- (a) Das Präfix „ σ “ bezieht sich auf die in Eigenschaft (3) formulierte *abzählbare* Vereinigung. Wird diese Eigenschaft nur für *endlich* viele Elemente aus \mathcal{A} gefordert, nennt man \mathcal{A} eine **Algebra** (über Ω). Man beachte aber, dass aus der σ -Eigenschaft auch die Abgeschlossenheit gegenüber endlichen Vereinigungen folgt:

$$A_1, A_2, \dots, A_n, \emptyset, \emptyset, \dots \in \mathcal{A} \implies \bigcup_{i=1}^n A_i \in \mathcal{A}$$

- (b) Formal betrachtet ist ein einzelner Versuchsausgang $\omega \in \Omega$ *kein* Ereignis. Das korrespondierende (einelementige) Ereignis lautet korrekt $\{\omega\} \subset \Omega$. Man beachte überdies, dass sich aus (1) bis (3) *nicht* automatisch ergibt, dass einelementige Mengen auch Ereignisse sein müssen. (In der Praxis wird Letzteres aber meist stillschweigend angenommen.)
- (c) In Verallgemeinerung von (b) lässt sich festhalten, dass Ereignisse zwar Teilmengen von Ω sind, aber umgekehrt nicht jede Teilmenge von Ω automatisch auch ein Ereignis sein muss. Kommen Wahrscheinlichkeiten ins Spiel, wäre das aus theoretischen Gründen auch gar nicht wünschenswert (s. unten).
- (d) Die kleinste σ -Algebra über Ω besteht nur aus der leeren Menge und aus dem Merkmalraum: $\mathcal{A} = \{\emptyset, \Omega\}$.
- (e) Die größte σ -Algebra über Ω ist die **Potenzmenge**, d. h. die Menge aller Teilmengen des Merkmalraums:

$$\mathcal{A} = \{A \mid A \subseteq \Omega\} = \mathcal{P}(\Omega)$$

Besteht Ω aus endlich vielen Elementen, gilt $|\mathcal{P}(\Omega)| = 2^{|\Omega|}$.

Bem: Die obige Schreibweise ist auch für unendliche Mengen gebräuchlich. Der *Satz von Cantor* besagt für eine beliebige Menge M , dass die Mächtigkeit (oder Kardinalität) der Potenzmenge $\mathcal{P}(M)$ stets größer ist als die Kardinalität von M , d. h. $|M| < |\mathcal{P}(M)|$. Die sog. *verallgemeinerte Kontinuumshypothese*⁴ besagt für unendliche Mengen M , dass $|\mathcal{P}(M)|$ die nach $|M|$ nächstgrößere Mächtigkeit ist. Insbesondere bedeutet das, dass die Potenzmenge der natürlichen Zahlen \mathbb{N} die Kardinalität von \mathbb{R} hat, d. h. $2^{|\mathbb{N}|} = |\mathbb{R}|$.

- (f) Betrachtet als Ereignisse, nennt man den Merkmalraum $\Omega \in \mathcal{A}$ das **sichere** Ereignis, und die leere Menge $\emptyset \in \mathcal{A}$ das **unmögliche** Ereignis.

⁴Ein zentrales Resultat der Mengentheorie lautet, dass die Kontinuumshypothese im Rahmen der üblichen Axiome der Mengenlehre weder beweis- noch widerlegbar ist, also von den Axiomen unabhängig ist (KURT GÖDEL (1938), PAUL COHEN (1960)).

Festlegung 1: Ist der Merkmalraum Ω endlich oder abzählbar unendlich, wählt man als Ereignissystem stets die Potenzmenge $\mathcal{P}(\Omega)$. Letzteres System ist (trivialerweise) eine σ -Algebra. Man braucht sich also in diesem Fall über die Messbarkeit von Ereignissen keine Gedanken zu machen.

Bsp 2.2 Besteht das statistische Experiment im Werfen eines Würfels und interessiert man sich für die geworfene Augenzahl, ist ein passender Merkmalraum gegeben durch $\Omega = \{1, 2, 3, 4, 5, 6\}$ und das zugehörige Ereignissystem ist die Potenzmenge von Ω :

$$\mathcal{A} = \mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \dots, \{6\}, \{1, 2\}, \{1, 3\}, \dots, \Omega\}$$

Beispielsweise lässt sich das Ereignis, dass die geworfene Augenzahl eine gerade Zahl ist, durch $A = \{2, 4, 6\}$ formulieren. Das Komplement von A , $A^c = \{1, 3, 5\}$, entspricht dem Ereignis, dass die Augenzahl ungerade ist.

Wirft man den Würfel solange, bis zum ersten Mal ein „Sechser“ geworfen wird, wäre $\Omega = \{1, 2, \dots\} = \mathbb{N}$ ein geeigneter Merkmalraum. Als zugehöriges Ereignissystem wählt man wieder die Potenzmenge $\mathcal{A} = \mathcal{P}(\Omega)$. ■

2.4 Borelmengen

Für *überabzählbare* Merkmalräume (z. B. \mathbb{R} , $[0, \infty)$, ...) ist die Potenzmenge $\mathcal{P}(\Omega)$ als Ereignissystem nicht geeignet. Das hat den folgenden Grund: Wie unten ausführlicher dargestellt, möchte man den Ereignissen Wahrscheinlichkeiten zuordnen. Wäre nun *jede* Teilmenge von beispielsweise \mathbb{R} ein Ereignis, hätte man – bildlich gesprochen – „zu viele“ Ereignisse (vgl. dazu Punkt (e) in den Bemerkungen von 2.3) als dass die Zuordnung von Wahrscheinlichkeiten ohne Widerpruch möglich wäre. D. h., man muss sich auf eine *echte* Teilmenge der Potenzmenge beschränken.

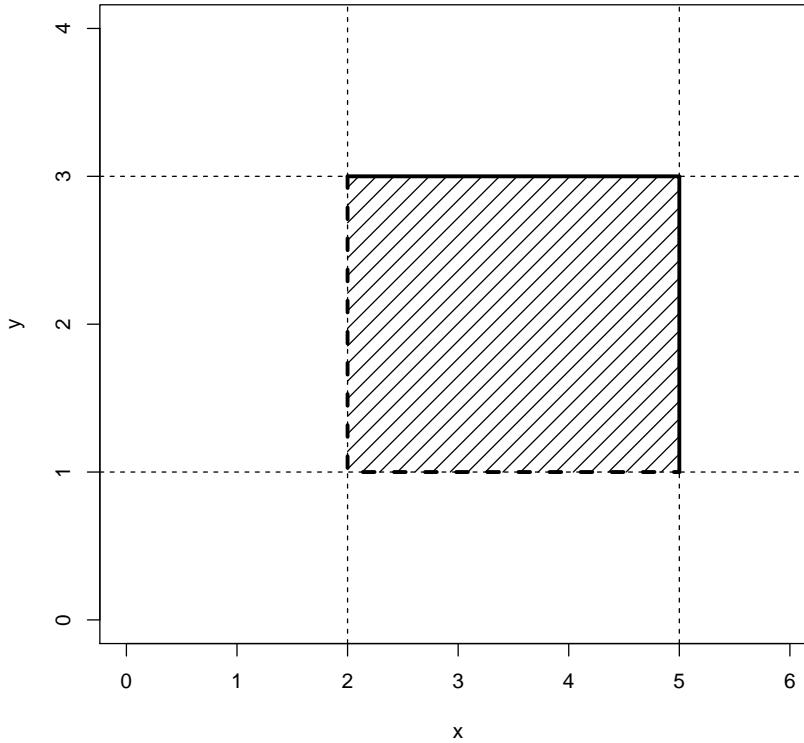
Im Folgenden sei $\Omega = \mathbb{R}$ und für die Konstruktion eines passenden Ereignissystems betrachten wir zunächst das System der links offenen und rechts abgeschlossenen (endlichen) Intervalle:

$$\mathcal{G} = \{(a, b] \mid a, b \in \mathbb{R} \text{ mit } a \leq b\}$$

Klarerweise ist \mathcal{G} noch keine σ -Algebra (z. B. ist das Komplement von $(0, 1]$ kein Element von \mathcal{G}). Aus diesem Grund erzeugt man die *kleinste* σ -Algebra, die alle Elemente von \mathcal{G} umfasst. Diese kleinste σ -Algebra \mathcal{B} nennt man die **Borel σ -Algebra** (oder die **Borelmengen**⁵). Wie man zeigen kann, ist \mathcal{B} eine *echte* Teilmenge der Potenzmenge $\mathcal{P}(\mathbb{R})$.

Bem: Mengen, die nicht zu \mathcal{B} gehören (die also *keine* Borelmengen sind), entziehen sich weitgehend der Anschauung. Anders ausgedrückt: Alle Teilmengen von \mathbb{R} , die man sich vorstellen oder grafisch veranschaulichen kann, *sind* Borelmengen.

⁵ÉMILE BOREL (1871–1956), franz. Mathematiker und Politiker.

Abbildung 2.2: Beispiel für einen halboffenen Quader

Wie in der obigen Bemerkung angedeutet, sind alle Ereignisse von praktischer Bedeutung Borelmengen. Insbesondere gilt, dass alle Typen von **Intervallen** $\langle a, b \rangle$ (offene, abgeschlossene, endliche, unendliche, ...) Borelmengen sind, speziell also auch alle **einpunktigen** Mengen $\{x\}$ ($x \in \mathbb{R}$).

Beweis für letztere Aussage: Eine einpunktige Menge $\{x\}$ lässt sich mit Intervallen aus \mathcal{G} wie folgt darstellen:

$$\{x\} = [x, x] = \bigcap_{n=1}^{\infty} \underbrace{\left(x - \frac{1}{n}, x \right]}_{\in \mathcal{G}} \in \mathcal{B}$$

Die Behauptung ergibt sich nun daraus, dass eine σ -Algebra gegenüber abzählbaren Durchschnitten abgeschlossen ist. Analog argumentiert man für andere Arten von Intervallen.

Mehrdimensionale Borelmengen: Analog zum eindimensionalen Fall lässt sich die Borel σ -Algebra über \mathbb{R}^k ($k \geq 2$) definieren. Beispielsweise ist für $k = 2$ das erzeugende Ereignissystem aus halboffenen *Quadern* (vgl. Abb 2.2) wie folgt gegeben:

$$\mathcal{G}_2 = \{(a, b] \times (c, d] \mid a, b, c, d \in \mathbb{R} \text{ mit } a \leq b, c \leq d\}$$

\mathcal{B}_2 ist nun definiert als die *kleinste* σ -Algebra, die alle Ereignisse aus \mathcal{G}_2 umfasst. Analog sind die k -dimensionalen Borelmengen \mathcal{B}_k definiert.

Festlegung 2: Ist der Merkmalraum Ω eine (überabzählbare) Teilmenge von \mathbb{R}^k ($k \geq 1$), wählt man als Ereignissystem stets die entsprechende Borel σ -Algebra über Ω . Alle praktisch relevanten Ereignisse werden dadurch erfasst.

2.5 Wahrscheinlichkeitsmaße

Die Wahrscheinlichkeitsdefinition nach A. N. KOLMOGOROW⁶ besteht aus drei Axiomen, die – wie bereits in 2.1 erwähnt – von den Eigenschaften der relativen Häufigkeiten motiviert sind. Ist $h_n(A)$ die relative Häufigkeit eines Ereignisses A auf Basis von n wiederholten Versuchen, dann gilt $0 \leq h_n(A) \leq 1$. Sind A_1 und A_2 zwei *disjunkte* Ereignisse (d. h., gilt $A_1 \cap A_2 = \emptyset$), so gilt $h_n(A_1 \cup A_2) = h_n(A_1) + h_n(A_2)$.

Wahrscheinlichkeitsmaß: Gegeben sei ein Messraum (Ω, \mathcal{A}) (d. h. eine σ -Algebra \mathcal{A} über einem Merkmalraum Ω). Eine Abbildung $P : \mathcal{A} \rightarrow \mathbb{R}$ heißt ein **Wahrscheinlichkeitsmaß** (kurz **W-Maß**) auf (Ω, \mathcal{A}) , wenn sie die folgenden Eigenschaften erfüllt:

- (1) $P(A) \geq 0$ für alle $A \in \mathcal{A}$
- (2) $P(\Omega) = 1$
- (3) Für eine Folge A_1, A_2, \dots von (*paarweise*) *disjunkten* Ereignissen (d. h. $A_i \cap A_j = \emptyset$ für $i \neq j$) gilt die σ -**Additivität**:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Auf Basis dieser Definition lässt sich der Messraum (Ω, \mathcal{A}) zu einem **Wahrscheinlichkeitsraum** (kurz **W-Raum**) (Ω, \mathcal{A}, P) erweitern.

Die obigen Axiome haben eine Reihe von Konsequenzen. (Bem: Im Folgenden wird stets angenommen, dass die Ereignisse in der σ -Algebra \mathcal{A} liegen, also messbar sind.)

Behauptung 1: $P(\emptyset) = 0$

Beweis: Man nehme eine Folge von Ereignissen A_1, A_2, \dots aus \mathcal{A} , wobei $A_1 = \Omega$ und $A_i = \emptyset$, $i > 1$. Die Ereignisse sind paarweise disjunkt und $\Omega = \bigcup_{i=1}^{\infty} A_i$; also gilt nach Axiom (2) und (3):

$$\underbrace{P(\Omega)}_{=1} = \sum_{i=1}^{\infty} P(A_i) = \underbrace{P(\Omega)}_{=1} + \sum_{i=2}^{\infty} P(\emptyset)$$

Obige Gleichung lässt sich nicht anders als durch $P(\emptyset) = 0$ erfüllen.

⁶ANDREI NIKOLAJEWITSCH KOLMOGOROW (1903–1987), russ. Mathematiker (bedeutende Beiträge zu mehreren Gebieten der Mathematik).

Behauptung 2: Für eine *endliche* Folge A_1, A_2, \dots, A_n von (paarweise) disjunkten Ereignissen gilt:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

Beweis: Folgt aus (3) und Behauptung 1: Setze $A_i = \emptyset$, $i > n$.

Behauptung 3: $P(A^c) = 1 - P(A)$

Beweis: A und A^c sind disjunkt und $A \cup A^c = \Omega$; nach Axiom (2) und (3) (oder Behauptung 2) gilt:

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$$

Behauptung 4: $A \subseteq B \implies P(A) \leq P(B)$

Beweis: Wegen $A \subseteq B$ lässt sich B darstellen als $B = A \cup (A^c \cap B)$. Die beiden letzteren Ereignisse sind aber disjunkt; also gilt:

$$P(B) = P(A) + P(A^c \cap B)$$

Wegen $P(A^c \cap B) \geq 0$ folgt die Behauptung.

Behauptung 5: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Beweis: $A \cup B$ lässt sich als Vereinigung von disjunkten Ereignissen darstellen: $A \cup B = A \cup (A^c \cap B)$. Nach Axiom (3) gilt:

$$P(A \cup B) = P(A) + P(A^c \cap B)$$

Nun gilt aber auch $B = (A \cap B) \cup (A^c \cap B)$. Da die beiden letzteren Ereignisse disjunkt sind, gilt wieder nach Axiom (3):

$$P(B) = P(A \cap B) + P(A^c \cap B) \implies P(A^c \cap B) = P(B) - P(A \cap B)$$

Daraus folgt die Behauptung.

2.6 Chancen (Odds)

In der Praxis formuliert man Wahrscheinlichkeiten häufig in Form von „Chancen“. Man sagt etwa, „die Chancen stehen 3 zu 1“, ein bestimmtes Spiel zu gewinnen.

Chancen (Odds): Die **Chancen (engl. Odds)** $o(A)$ für den Eintritt eines Ereignisses A sind definiert als der Quotient aus der Wahrscheinlichkeit $p = P(A)$ und der Gegenwahrscheinlichkeit $P(A^c) = 1 - P(A) = 1 - p$:

$$o(A) = \frac{P(A)}{1 - P(A)} = \frac{p}{1 - p}$$

Logarithmiert man $o(A)$, ergeben sich die **logarithmierten Chancen (engl. Log-Odds)**:

$$\log o(A) = \log(p) - \log(1 - p)$$

Die Log-Odds transformieren Wahrscheinlichkeiten (also Zahlen zwischen 0 und 1) in reelle Zahlen zwischen $-\infty$ und ∞ . Sie besitzen eine *Symmetrieeigenschaft*, d. h., die Log-Odds des komplementären Ereignisses A^c sind das Negative der Log-Odds von A :

$$\log o(A^c) = \log \frac{1 - p}{p} = -\log \frac{p}{1 - p} = -\log o(A)$$

Stehen die Chancen 1 zu 1 (d. h., sind A und A^c gleichwahrscheinlich), gilt $o(A) = 1$ und $\log o(A) = 0$.

Bei zwei Ereignissen kann man das Verhältnis der Odds betrachten.

Chancenverhältnis (Odds–Ratio): Die Chancen $o(A)$ und $o(B)$ von zwei Ereignissen A und B werden häufig durch das **Chancenverhältnis (engl. Odds–Ratio)** miteinander verglichen:

$$r(A, B) = \frac{o(A)}{o(B)} = \frac{P(A)/[1 - P(A)]}{P(B)/[1 - P(B)]}$$

Die **Log-Odds–Ratio** ist die Differenz der Log-Odds:

$$\log r(A, B) = \log o(A) - \log o(B)$$

Bsp 2.3 Das Ereignis, ein Spiel zu gewinnen, trete mit Wahrscheinlichkeit $p = 0.75$ ein. Die Chancen, das Spiel zu gewinnen, stehen also 75 zu 25 oder 3 zu 1 (d. h. $o = 3$). Zu gewinnen ist dreimal so wahrscheinlich wie zu verlieren. Ist die Gewinnwahrscheinlichkeit für ein anderes Spiel gleich 0.9, so sind die Odds gleich $0.9/0.1 = 9$, d. h., zu gewinnen ist neunmal wahrscheinlicher als zu verlieren. Die Odds-Ratio beträgt $r = 9/3 = 3$. D. h., die Gewinnchancen sind beim zweiten Spiel um den Faktor 3 günstiger. Auf der logarithmischen Skala erhalten wir $\log(3)$ bzw. $\log(9)$ und die Log-Odds–Ratio beträgt $\log(r) = \log(3)$. ■

2.7 Endliche W–Räume

Ein W–Raum (Ω, \mathcal{A}, P) heißt **endlich**, wenn der Merkmalraum Ω eine endliche Menge ist. Nach **Festlegung 1** (vgl. 2.3) gilt in diesem Fall $\mathcal{A} = \mathcal{P}(\Omega)$, und das W–Maß P ist durch die Angabe der Wahrscheinlichkeiten der **Elementarereignisse** $\{\omega\}$, $\omega \in \Omega$, eindeutig bestimmt:

$$P(\{\omega\}) \geq 0 \quad \forall \omega \in \Omega \quad \text{und} \quad \sum_{\omega \in \Omega} P(\{\omega\}) = 1$$

Die Wahrscheinlichkeit eines beliebigen Ereignisses $A \in \mathcal{P}(\Omega)$ ergibt sich dann durch Addition der Einzelwahrscheinlichkeiten:

$$P(A) = \sum_{\omega \in A} P(\{\omega\})$$

Ein wichtiger Spezialfall ergibt sich, wenn die Elementarereignisse **gleichwahrscheinlich** sind. Statistische Experimente dieser Art nennt man **Laplace–Experimente**.⁷

Laplace–Raum: Ein endlicher W–Raum $(\Omega, \mathcal{P}(\Omega), P)$ heißt **Laplace–Raum**, wenn:

$$P(\{\omega\}) = \frac{1}{|\Omega|} \quad \text{für alle } \omega \in \Omega$$

Die Wahrscheinlichkeit eines beliebigen Ereignisses $A \in \mathcal{P}(\Omega)$ ist dann gegeben durch:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der Elemente von } A}{\text{Anzahl der Elemente von } \Omega}$$

Bemerkungen:

- (a) Die übliche Sprechweise im Laplace–Raum lautet: Die Wahrscheinlichkeit des Ereignisses A ist der Quotient aus der Zahl der für (den Eintritt von) A *günstigen* Fälle und der Zahl der *möglichen* Fälle, oder kurz „ g durch m “. Das nennt man auch die *klassische* Wahrscheinlichkeitsdefinition.
- (b) Vor Anwendung der klassischen W–Definition ist genau zu klären, ob tatsächlich eine *zufällige Entnahme* eines Elements aus Ω vorliegt. Bei Glücksspielen (Lotto, Roulette, ...) mag das hinlänglich gut der Fall sein. Viele Spiele (Poker, Backgammon, ...) sind aber eine Mischung aus Glück und Geschicklichkeit, sodass die klassische W–Definition nur bedingt (oder nur in Teilbereichen) anwendbar ist. Darüberhinaus findet die klassische W–Definition aber auch zahlreiche Anwendungen in anderen Gebieten, etwa in der Genetik oder in der Teilchenphysik.

⁷PIERRE-SIMON DE LAPLACE (1749–1827), franz. Mathematiker, Physiker und Astronom.

- (c) Bei Anwendungen der klassischen W–Definition spielt naturgemäß das *Zählen* eine große Rolle. (Wieviele Elemente hat der Merkmalraum? Wieviele Elemente hat ein bestimmtes Ereignis?) Manchmal können die Elemente direkt abgezählt werden, in den meisten Fällen wird man aber auf *kombinatorische* Methoden zurückgreifen müssen. (Vgl. Anhang: Abzählende Kombinatorik für eine kurze Zusammenfassung der wichtigsten Zähl– und Auswahlprinzipien.)

Bsp 2.4 In einem Behälter seien n gleichartige (aber unterscheidbare) Kugeln, $n - 1$ seien weiß und eine sei rot. Wenn willkürlich k Kugeln hintereinander entnommen werden (Ziehungen ohne Zurücklegen), mit welcher Wahrscheinlichkeit befindet sich darunter die rote Kugel?

Da alle Kugeln auf die gleiche Weise behandelt werden, ist die ausgewählte Menge von k Kugeln mit gleicher Wahrscheinlichkeit eine von den $\binom{n}{k}$ möglichen Auswahlen von k Kugeln. Also gilt:

$$P(\{\text{Die rote Kugel wird ausgewählt}\}) = \frac{\binom{1}{1} \binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}$$

Andere Lösung: Bezeichnet A_i das Ereignis, dass die rote Kugel die i -te gezogene Kugel ist, so gilt auf Grund der Art der Ziehung, dass $P(A_i) = 1/n$, $i = 1, 2, \dots, k$. Die Ereignisse A_i sind paarweise disjunkt, also gilt:

$$P(\{\text{Die rote Kugel wird ausgewählt}\}) = P\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i) = \frac{k}{n}$$

Noch eine andere Lösung: Es gibt $n(n-1)\cdots(n-k+1)$ gleichwahrscheinliche Möglichkeiten, k Kugeln unter Beachtung der Reihenfolge zu ziehen. Die Wahrscheinlichkeit, dass die rote Kugel *nicht* gezogen wird, ist gegeben durch:

$$P(\{\text{Die rote Kugel wird nicht ausgewählt}\}) = \frac{(n-1)(n-2)\cdots(n-k)}{n(n-1)\cdots(n-k+1)} = \frac{n-k}{n}$$

Die gesuchte Wahrscheinlichkeit ist daher $1 - (n-k)/n = k/n$. ■

2.8 Geometrische Wahrscheinlichkeiten

Die **geometrische Definition** von Wahrscheinlichkeit lässt sich anwenden, wenn der Merkmalraum als geometrisches Objekt (Längen-, Flächenstück, Volumen, ...) und Er-

eignisse als Teilbereiche dieses Objekts interpretiert werden können, deren Wahrscheinlichkeit **proportional zur Größe** (d. h. Länge, Fläche, Volumen, ...) des Teilbereichs ist, unabhängig von seiner Position und Form. Insofern stellt die geometrische Interpretation von Wahrscheinlichkeit eine Erweiterung des Laplace-Raums auf *unendlich* viele mögliche Versuchsausgänge dar.

Sind die Voraussetzungen für ihre Anwendung erfüllt, ist die **geometrische Wahrscheinlichkeit** eines Ereignisses $A \subseteq \Omega$ gegeben durch:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Größe von } A}{\text{Größe von } \Omega}$$

Ein einfaches Beispiel soll das Konzept verdeutlichen.

Bsp 2.5 [Rendezvousproblem] Angenommen, zwei Wanderer A und B erreichen, aus unterschiedlichen Richtungen kommend, einen Aussichtspunkt und halten sich dort jeweils 10 (A) bzw. 20 Minuten (B) auf. Ihre Ankunftszeiten am Aussichtspunkt liegen – unabhängig voneinander – zufällig zwischen 10 und 11 Uhr. Mit welcher Wahrscheinlichkeit begegnen sie einander am Aussichtspunkt?

Interpretiert man die Eintreffzeitpunkte der beiden Wanderer als Punkt im (o. B. d. A.) Quadrat $[0, 1] \times [0, 1]$, so entspricht der Begegnungsbereich einem Flächenstück um die Diagonale. In Abb 2.3 ist das die schraffierte Fläche. Rechnet man in der Einheit [h], ist die Wahrscheinlichkeit einer Begegnung gegeben durch:

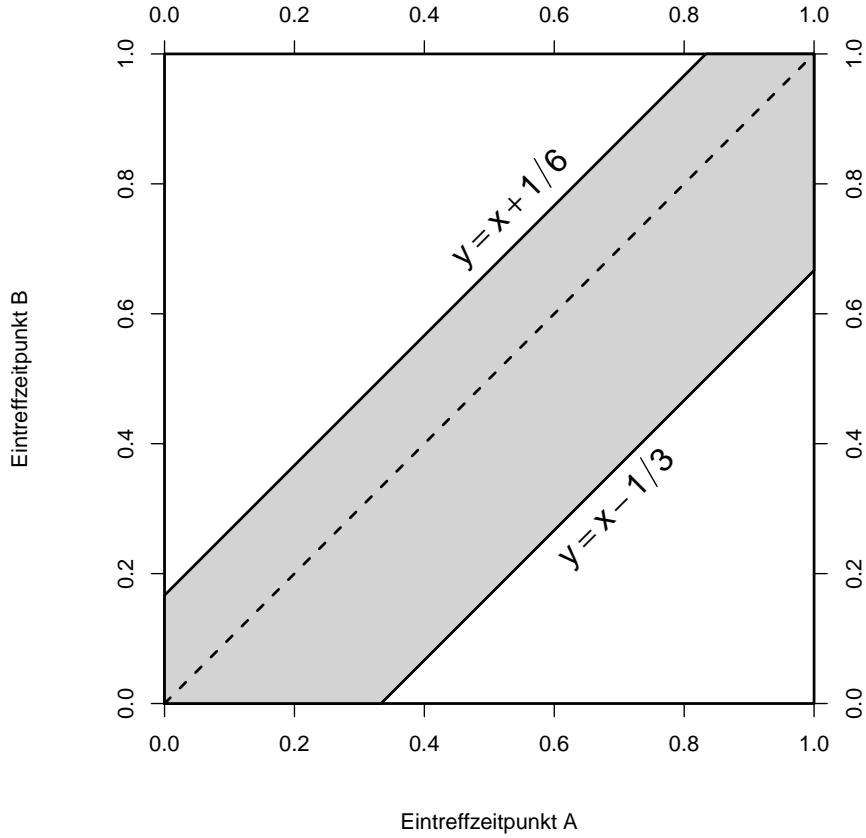
$$P(\{\text{Begegnung}\}) = 1 - \frac{(5/6)^2 + (2/3)^2}{2} = \frac{31}{72} \doteq 0.431$$

Dabei ist es einfacher, zunächst die Fläche des komplementären Ereignisses (d. h. „Keine Begegnung“) zu berechnen.

An diesem Beispiel zeigt sich auch, wie schnell man an die Grenzen der geometrischen Anschaulichkeit gelangt. Wie schaut beispielsweise – unter ähnlichen Bedingungen – der Begegnungsbereich für *drei* Wanderer aus? In komplizierteren Situationen ist eine analytische Lösung meist einfacher. ■

2.9 Additionstheorem

Behauptung 5 aus 2.5 lässt sich auf mehr als zwei Ereignisse verallgemeinern. Da eine Vereinigung von Mengen auch als (mengentheoretische) Addition bezeichnet wird, spricht man vom **Additionstheorem**. Andere Bezeichnungen lauten **Formel der In- und Exklusion** oder **Siebformel**. Letztere beziehen sich auf die „operative“ Interpretation des Additionstheorems (s. unten).

Abbildung 2.3: Rendezvousproblem

Bem: Zur einfacheren Darstellung verwenden wir im Folgenden die übliche Abkürzung AB für $A \cap B$ (analog für mehr als zwei Ereignisse).

(Allgemeines) Additionstheorem: Für Ereignisse A_1, A_2, \dots, A_n aus \mathcal{A} gilt:

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} A_{i_2}) + \dots \\ &\quad + (-1)^{r+1} \sum_{i_1 < i_2 < \dots < i_r} P(A_{i_1} A_{i_2} \dots A_{i_r}) \\ &\quad + \dots + (-1)^{n+1} P(A_1 A_2 \dots A_n) \end{aligned}$$

Beweis: Mittels (mathematischer) Induktion nach n . Induktionsanfang ist das Theorem für zwei Ereignisse (vgl. Behauptung 5 aus 2.5). Für den Schritt von n auf $n + 1$ zeigt man:

$$P\left(\bigcup_{i=1}^{n+1} A_i\right) = P\left(\bigcup_{i=1}^n A_i\right) + P(A_{n+1}) - P\left(\bigcup_{i=1}^n A_i A_{n+1}\right)$$

Letzteres folgt aus der Gültigkeit des Theorems für zwei Ereignisse.

In Worten: Um die Wahrscheinlichkeit der Vereinigung von n Ereignissen zu berechnen, addiere man zunächst die Einzelwahrscheinlichkeiten, subtrahiere davon die Wahrscheinlichkeiten aller paarweisen Durchschnitte, addiere wiederum die Wahrscheinlichkeiten aller dreifachen Durchschnitte, etc.

Kompakte Darstellung der Formel der In- und Exklusion:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n (-1)^{r+1} \sum_{i_1 < i_2 < \dots < i_r} P(A_{i_1} A_{i_2} \cdots A_{i_r})$$

Betrachtet man die Teilsummen auf der rechten Seite des Additionstheorems ergeben sich wechselweise Abschätzungen nach oben und nach unten:

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &\leq \sum_{i=1}^n P(A_i) \\ P\left(\bigcup_{i=1}^n A_i\right) &\geq \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} A_{i_2}) \\ P\left(\bigcup_{i=1}^n A_i\right) &\leq \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} A_{i_2}) + \sum_{i_1 < i_2 < i_3} P(A_{i_1} A_{i_2} A_{i_3}) \\ &\vdots \end{aligned}$$

Die erste der obigen Ungleichungen ist nach G. BOOLE⁸ benannt; sie gilt auch für unendlich viele Ereignisse.

Boole'sche Ungleichung: Für eine Folge von Ereignissen A_1, A_2, \dots aus \mathcal{A} gilt:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$$

Beweis: Das Ereignis $\bigcup_{i=1}^{\infty} A_i$ lässt sich als Vereinigung von (paarweise) disjunkten Ereignissen darstellen:

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_2 A_1^c) \cup (A_3 A_1^c A_2^c) \cup \dots$$

Bezeichnet man die einzelnen Terme der disjunkten Vereinigung mit B_i , so gilt nach Behauptung 4 aus 2.5, dass $P(B_i) \leq P(A_i)$. Die Behauptung folgt dann aus der σ -Additivität:

⁸GEORGE BOOLE (1815–1864), engl. Mathematiker, Logiker und Philosoph.

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i) \leq \sum_{i=1}^{\infty} P(A_i)$$

Bsp 2.6 Angenommen, von den Mitgliedern eines Clubs spielen 36 Tennis, 28 spielen Squash und 18 spielen Badminton. Außerdem spielen 22 Tennis und Squash, 12 spielen Tennis und Badminton, 9 spielen Squash und Badminton und 4 spielen alle drei Sportarten. Wieviele Mitglieder des Clubs betreiben zumindest eine der drei Sportarten?

Für die Beantwortung der Frage stelle man sich vor, dass ein Mitglied des Clubs zufällig ausgewählt wird. Dadurch wird eine W-Verteilung „induziert“: Ist C eine Teilmenge der Clubmitglieder und ist N deren Gesamtzahl, so definiert man:

$$P(C) = \frac{\text{Anzahl der Elemente von } C}{N}$$

Bezeichnet T (S , B) die Menge der Clubmitglieder, die Tennis (Squash, Badminton) spielen, so gilt nach dem Additionstheorem:

$$\begin{aligned} P(T \cup S \cup B) &= P(T) + P(S) + P(B) - P(TS) - P(TB) - P(SB) + P(TSB) \\ &= \frac{36 + 28 + 18 - 22 - 12 - 9 + 4}{N} \\ &= \frac{43}{N} \end{aligned}$$

D. h., 43 Clubmitglieder betreiben zumindest eine der drei Sportarten. ■

2.10 Bedingte Wahrscheinlichkeit

Der Begriff der bedingten Wahrscheinlichkeit gehört zu den wichtigsten Konzepten der W-Theorie. Das erklärt sich daraus, dass der Wahrscheinlichkeitsbegriff eng mit dem Informationsbegriff verknüpft ist. Solange wir nicht wissen, ob ein Ereignis A eingetreten ist oder nicht, bewerten wir das Ereignis mit seiner Wahrscheinlichkeit $P(A)$. Die Kenntnis, dass ein *anderes* Ereignis B eingetreten ist, kann *informativ* für das (mögliche) Eintreten von A sein und die Eintrittswahrscheinlichkeit ändern (d. h. vergrößern oder verkleinern). Man schreibt in diesem Fall $P(A|B)$ und spricht von der **bedingten** Wahrscheinlichkeit von A **gegeben** B .

Die folgende Definition von $P(A|B)$ lässt sich wie folgt motivieren: Wenn B eingetreten ist, sind nur noch diejenigen Versuchsausgänge $\omega \in A$ relevant, die auch in B liegen. Zu betrachten ist also das Ereignis $A \cap B$. Andererseits, wenn B eingetreten ist, wird B zum

neuen (reduzierten) Merkmalraum und die Wahrscheinlichkeit von $A \cap B$ ist relativ zur Wahrscheinlichkeit von B zu bewerten.

Bedingte Wahrscheinlichkeit: Gegeben sei ein W-Raum (Ω, \mathcal{A}, P) und $A, B \in \mathcal{A}$ seien zwei Ereignisse, wobei $P(B) > 0$. Dann ist die **bedingte Wahrscheinlichkeit** von A gegeben B definiert durch:

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

Liegt speziell ein Laplace-Raum vor, dann ist $P(A|B)$ der Anteil der für das Ereignis $A \cap B$ günstigen Fälle, bezogen auf die möglichen Fälle, die dem Ereignis B entsprechen:

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

Man kann in diesem Fall also mit dem reduzierten Merkmalraum B arbeiten. Dazu ein einfaches Beispiel.

Bsp 2.7 Eine Münze wird zweimal geworfen. Unter der Annahme, dass alle vier Punkte des Merkmalraums $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$ (H = Kopf, T = Zahl) gleichwahrscheinlich sind, wie groß ist die bedingte Wahrscheinlichkeit, dass beide Würfe H sind, wenn (a) der erste Wurf H ist? (b) zumindest ein Wurf H ist?

Sei $B = \{(H, H)\}$ ($\hat{=}$ beide Würfe sind H), $F = \{(H, H), (H, T)\}$ ($\hat{=}$ der erste Wurf ist H) und $A = \{(H, H), (H, T), (T, H)\}$ ($\hat{=}$ zumindest ein Wurf ist H). Die Wahrscheinlichkeit für (a) berechnet man wie folgt:

$$P(B|F) = \frac{P(BF)}{P(F)} = \frac{P(\{(H, H)\})}{P(\{(H, H), (H, T)\})} = \frac{1/4}{2/4} = \frac{1}{2}$$

Nun zu (b):

$$P(B|A) = \frac{P(BA)}{P(A)} = \frac{P(\{(H, H)\})}{P(\{(H, H), (H, T), (T, H)\})} = \frac{1/4}{3/4} = \frac{1}{3}$$

Vielfach besteht die Meinung, dass die Wahrscheinlichkeit für (b) ebenfalls $1/2$ ist. Dabei wird wie folgt argumentiert: Wenn zumindest ein Wurf H ist, verbleiben nur zwei Möglichkeiten: Beide Würfe sind H oder nur ein Wurf ist H . Das ist zwar korrekt, der Fehler liegt aber in der Annahme, dass diese beiden Möglichkeiten gleichwahrscheinlich sind (was nicht der Fall ist). Durch Beschränkung auf den durch die Bedingung reduzierten Merkmalraum, lassen sich Fehlschlüsse dieser Art vermeiden. ■

$P(\cdot|B)$ ist ein W-Maß: Bedingte Wahrscheinlichkeiten erfüllen alle Eigenschaften eines (üblichen) W-Maßes. Gegeben sei ein W-Raum (Ω, \mathcal{A}, P) und $B \in \mathcal{A}$ sei ein Ereignis mit $P(B) > 0$. Dann gilt:

- (1) $P(A|B) \geq 0$ für alle $A \in \mathcal{A}$
- (2) $P(\Omega|B) = 1$
- (3) Für eine Folge A_1, A_2, \dots von (paarweise) disjunkten Ereignissen gilt:

$$P\left(\bigcup_{i=1}^{\infty} A_i \mid B\right) = \sum_{i=1}^{\infty} P(A_i|B)$$

2.11 Multiplikationstheorem

Multipliziert man in der Definition der bedingten Wahrscheinlichkeit $P(A|B)$ beide Seiten mit $P(B)$, so ergibt sich:

$$P(A \cap B) = P(A|B)P(B) \quad (\text{Vs.: } P(B) > 0)$$

Vertauschen von A und B ergibt:

$$P(A \cap B) = P(B|A)P(A) \quad (\text{Vs.: } P(A) > 0)$$

Eine Aussage dieser Art nennt man **Multiplikationstheorem** (oder **-regel**); es lässt sich auf mehr als zwei Ereignisse verallgemeinern.

(Allgemeines) Multiplikationstheorem: Für $A_1, A_2, \dots, A_n \in \mathcal{A}$ mit $P(A_1 A_2 \cdots A_n) > 0$ gilt:

$$P\left(\bigcap_{i=1}^n A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 A_2) \cdots P(A_n|A_1 A_2 \cdots A_{n-1})$$

Beweis: Anwendung der Definition der bedingten Wahrscheinlichkeit:

$$P(A_1) \frac{P(A_1 A_2)}{P(A_1)} \frac{P(A_1 A_2 A_3)}{P(A_1 A_2)} \cdots \frac{P(A_1 A_2 \cdots A_n)}{P(A_1 A_2 \cdots A_{n-1})} = P(A_1 A_2 \cdots A_n)$$

Bsp 2.8 Ein übliches Kartenpaket (52 Karten; 4 Farben: Kreuz, Herz, Pik, Karo; 13 Werte: 2–10, Bube (Jack), Dame (Queen), König, Ass) werde zufällig auf 4 Pakete zu je 13 Karten aufgeteilt. Mit welcher Wahrscheinlichkeit enthält jedes Paket ein Ass?

Die gesuchte Wahrscheinlichkeit lässt sich hier mittels einfacher kombinatorischer Überlegungen bestimmen. (Wie?) Ein Vorteil des Multiplikationstheorems besteht jedoch darin, dass komplizierte Probleme in mehrere einfachere Teilprobleme zerlegt werden können. Sei A_i , $i = 1, 2, 3, 4$, das Ereignis, dass das i -te Paket genau ein Ass enthält. Dann gilt:

$$P(A_1) = \frac{\binom{4}{1} \binom{48}{12}}{\binom{52}{13}} = \frac{(4)(13)(37)(38)(39)}{(49)(50)(51)(52)} \doteq 0.4388$$

$$P(A_2|A_1) = \frac{\binom{3}{1} \binom{36}{12}}{\binom{39}{13}} = \frac{(3)(13)(25)(26)}{(37)(38)(39)} \doteq 0.4623$$

$$P(A_3|A_1A_2) = \frac{\binom{2}{1} \binom{24}{12}}{\binom{26}{13}} = \frac{(2)(13)(13)}{(25)(26)} = 0.52$$

$$P(A_4|A_1A_2A_3) = \frac{\binom{1}{1} \binom{12}{12}}{\binom{13}{13}} = 1$$

Die gesuchte Wahrscheinlichkeit ergibt sich mit dem Multiplikationstheorem:

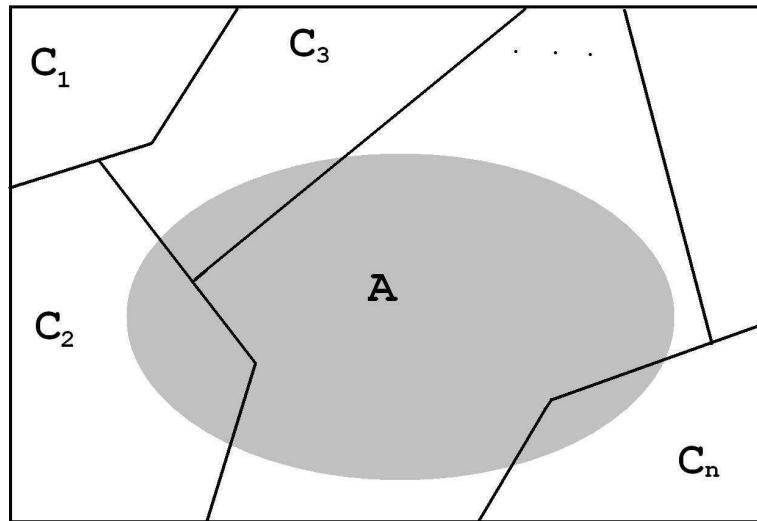
$$P\left(\bigcap_{i=1}^4 A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)P(A_4|A_1A_2A_3) \doteq 0.1055$$

■

2.12 Vollständige Wahrscheinlichkeit

Ähnlich wie das Multiplikationstheorem ermöglicht es auch der **Satz von der vollständigen (oder totalen) Wahrscheinlichkeit**, Wahrscheinlichkeiten von komplizierten Ereignissen aus Wahrscheinlichkeiten von einfacheren Ereignissen zusammenzusetzen. Grundlegend ist dabei die Zerlegung des Merkmalraums in (endlich oder abzählbar unendlich viele) paarweise disjunkte Ereignisse $C_1, C_2, \dots \in \mathcal{A}$, also eine **Partition** von Ω :

$$\Omega = \bigcup_{i=1}^{\infty} C_i \quad \text{mit} \quad C_i \cap C_j = \emptyset \quad \text{für} \quad i \neq j$$

Abbildung 2.4: Illustration zur vollständigen Wahrscheinlichkeit

Bem: Sinnvollerweise sollten die Ereignisse C_i nicht leer sein und eine positive Eintrittswahrscheinlichkeit (d. h. $P(C_i) > 0$) haben.

Satz von der vollständigen Wahrscheinlichkeit: Ist $C_1, C_2, \dots \in \mathcal{A}$ eine (höchstens) abzählbare Partition von Ω , so lässt sich die Wahrscheinlichkeit für ein Ereignis $A \in \mathcal{A}$ wie folgt berechnen:

$$P(A) = \sum_{i=1}^{\infty} P(A|C_i)P(C_i)$$

Abb 2.4 ist eine Illustration dieses Satzes in Form eines **Venn–Diagramms**.⁹ Der umfassende rechteckige Bereich repräsentiert den Merkmalraum Ω und Ereignisse werden als Teilbereiche dargestellt.

Beweis: Schneidet man A mit allen Mengen der Partition, bekommt man eine disjunkte Vereinigung:

$$A = \bigcup_{i=1}^{\infty} (A \cap C_i) \implies P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$$

Verwendet man für $P(A \cap C_i)$ das Multiplikationstheorem, ergibt sich die Behauptung:

$$P(A) = \sum_{i=1}^{\infty} P(A|C_i)P(C_i)$$

⁹JOHN VENN (1834–1923), engl. Mathematiker.

Bsp 2.9 In einem Behälter befinden sich gut gemischt – und äußerlich nicht unterscheidbar – drei Typen von Batterien im Verhältnis 20 : 30 : 50. Batterien vom Typ 1 arbeiten mit Wahrscheinlichkeit 0.7 länger als 100 Stunden; die entsprechenden Wahrscheinlichkeiten für die beiden anderen Typen sind 0.4 bzw. 0.3. Wenn nun dem Behälter willkürlich eine Batterie entnommen wird, mit welcher Wahrscheinlichkeit wird sie länger als 100 Stunden arbeiten?

Intuitiv sollte die gesuchte Wahrscheinlichkeit ein gewichteter Mittelwert von 0.7, 0.4 und 0.3 sein. Der Satz von der vollständigen Wahrscheinlichkeit sagt uns, wie die Gewichtung vorzunehmen ist. Bezeichnet A das Ereignis, dass die ausgewählte Batterie länger als 100 Stunden arbeiten wird, und B_i , $i = 1, 2, 3$, das Ereignis, dass Typ i gewählt wird, so gilt:

$$P(A) = \sum_{i=1}^3 P(A|B_i)P(B_i) = (0.7)(0.2) + (0.4)(0.3) + (0.3)(0.5) = 0.41$$

Die Wahrscheinlichkeit beträgt also 41%, dass die zufällig ausgewählte Batterie länger als 100 Stunden arbeiten wird. ■

2.13 Bayes'sche Formel

Hat man eine Partition C_1, C_2, \dots des Merkmalraums Ω und kennt für ein Ereignis A die bedingten Wahrscheinlichkeiten $P(A|C_i)$, so stellt sich häufig die Frage, wie daraus die „inversen“ bedingten Wahrscheinlichkeiten $P(C_i|A)$ berechnet werden können. Diese Frage lässt sich durch die **Bayes'sche Formel** (auch **Satz von Bayes**¹⁰ genannt) beantworten.

Bayes'sche Formel: $C_1, C_2, \dots \in \mathcal{A}$ sei eine Partition (d. h. eine disjunkte Zerlegung) von Ω und $P(C_i) > 0$ für alle $i = 1, 2, \dots$. Dann gilt für ein Ereignis A mit $P(A) > 0$:

$$P(C_i|A) = \frac{P(A|C_i)P(C_i)}{\sum_{j=1}^{\infty} P(A|C_j)P(C_j)}$$

Beweis: Nach Definition der bedingten Wahrscheinlichkeit und nach dem Multiplikationssatz gilt:

$$P(C_i|A) = \frac{P(C_i \cap A)}{P(A)} = \frac{P(A|C_i)P(C_i)}{P(A)}$$

Ersetzt man den Nenner durch die Formel für die vollständige Wahrscheinlichkeit, ergibt sich die Behauptung.

¹⁰THOMAS BAYES (1701(?)–1761), engl. (presbyterianischer) Geistlicher, beschäftigt sich auch mit Problemen der Mathematik; sein wichtigster Beitrag, die „Bayes'sche Formel“, wird aber erst posthum veröffentlicht (*An Essay Towards Solving a Problem in the Doctrine of Chances* (1763)).

Sprechweise: $P(C_i)$ nennt man in diesem Zusammenhang die **A-priori–** und $P(C_i|A)$ die **A-posteriori–Wahrscheinlichkeit** von C_i . Diese Ausdrücke beziehen sich auf den „Zeitpunkt“ zu dem die Information, dass A eingetreten ist, bekannt wird. Die Ereignisse C_i aus der Partition von Ω nennt man häufig auch **Hypothesen** (und schreibt H_i).

Bsp 2.10 In Fortsetzung von Bsp 2.9 kann man sich auch fragen, mit welcher Wahrscheinlichkeit es sich um eine Batterie von Typ i handelt, wenn bekannt ist, dass diese Batterie länger als 100 Stunden gearbeitet hat. Nach der Bayes'schen Formel gilt:

$$P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{0.41}$$

Somit:

$$P(B_1|A) = \frac{(0.7)(0.2)}{0.41} = \frac{14}{41} \doteq 0.341$$

$$P(B_2|A) = \frac{(0.4)(0.3)}{0.41} = \frac{12}{41} \doteq 0.293$$

$$P(B_3|A) = \frac{(0.3)(0.5)}{0.41} = \frac{15}{41} \doteq 0.366$$

(Bem: Klärerweise gilt $\sum_{i=1}^3 P(B_i|A) = 1$.) Beispielsweise beträgt a-priori die Wahrscheinlichkeit, dass eine Batterie vom Typ 1 gewählt wird, nur 0.2. Die Information aber, dass die Batterie länger als 100 Stunden gearbeitet hat, erhöht die Wahrscheinlichkeit dieses Ereignisses a-posteriori auf 0.341. ■

Odds–Form der Bayes'schen Formel: Betrachten wir in der obigen Sprechweise nur eine Hypothese H und ihre Gegenhypothese \bar{H} , so lässt sich die Bayes'sche Formel auf Basis der Odds (vgl. 2.6) auch wie folgt schreiben:

$$\underbrace{\frac{P(H|A)}{P(\bar{H}|A)}}_{\text{A-posteriori–Odds}} = \underbrace{\frac{P(H)}{P(\bar{H})}}_{\text{A-priori–Odds}} \times \underbrace{\frac{P(A|H)}{P(A|\bar{H})}}_{\text{Likelihood–Quotient}}$$

Der **Likelihood–Quotient**¹¹ (auch **Likelihood–Ratio**; kurz **LQ** oder **LR**) ist das Verhältnis der Wahrscheinlichkeit von A bedingt durch H und bedingt durch \bar{H} . Um zu den A-posteriori–Odds zu gelangen, muss man also nur die A-priori–Odds mit dem LQ multiplizieren.

¹¹ Likelihood heißt im Englischen zwar auch „Wahrscheinlichkeit“, aber man wählt hier ein anderes Wort, um die Unterschiede zu *probability* zu betonen.

2.14 Unabhängigkeit

Die *bedingte* Wahrscheinlichkeit von A gegeben B , d. h. $P(A|B)$, ist i. A. nicht gleich der *unbedingten* Wahrscheinlichkeit $P(A)$. Wenn bekannt ist, dass B eingetreten ist, verändert sich in der Regel die Wahrscheinlichkeit für den Eintritt von A . Gilt allerdings $P(A|B) = P(A)$, so hat der Eintritt von B quasi keinen „Einfluss“ auf den (möglichen) Eintritt von A . In diesem Fall sagt man, dass A und B **unabhängig** sind. Wegen $P(A|B) = P(AB)/P(B)$ sind A und B unabhängig, falls $P(AB) = P(A)P(B)$.

Unabhängigkeit von zwei Ereignissen: Zwei Ereignisse $A, B \in \mathcal{A}$ sind (stochastisch) **unabhängig** (kurz **ua.**), wenn:

$$P(AB) = P(A)P(B)$$

Andernfalls sind die Ereignisse **abhängig**. Man beachte, dass die Unabhängigkeit eine symmetrische Eigenschaft ist: Aus A, B ua. folgt B, A ua.

Bsp 2.11 Angenommen, wir werfen zwei (ausbalancierte) Würfel. E_1 sei das Ereignis, dass die Summe der Augenzahlen gleich 6 ist, und E_2 sei das Ereignis, dass die Augenzahl des ersten Würfels gleich 4 ist. Verwenden wir den üblichen Merkmalraum $\{(i, j) : i, j = 1, 2, \dots, 6\}$, bestehend aus allen möglichen Paaren von Augenzahlen, so gilt:

$$P(E_1E_2) = P(\{(4, 2)\}) = \frac{1}{36}$$

Andererseits gilt:

$$P(E_1)P(E_2) = \left(\frac{5}{36}\right) \left(\frac{1}{6}\right) = \frac{5}{216}$$

D. h., E_1 und E_2 sind nicht unabhängig. Ist E_3 aber das Ereignis, dass die Augensumme gleich 7 ist, so gilt:

$$P(E_2E_3) = P(\{(4, 3)\}) = \frac{1}{36} \quad \text{und} \quad P(E_2)P(E_3) = \left(\frac{6}{36}\right) \left(\frac{1}{6}\right) = \frac{1}{36}$$

D. h., E_2 und E_3 sind unabhängig! Was ist der Unterschied zum ersten Fall? Wenn die Augensumme gleich 6 ist, werden die Möglichkeiten für den ersten Wurf auf $\{1, 2, 3, 4, 5\}$ eingeschränkt. Ist die Augensumme aber gleich 7, werden die Möglichkeiten für den ersten Wurf nicht beschränkt. ■

Behauptung 1: Sind A, B ua., so sind auch (i) A, B^c , (ii) A^c, B und (iii) A^c, B^c ua.

Beweis für (i): A lässt sich schreiben als $A = AB \cup AB^c$; wegen $AB \cap AB^c = \emptyset$ gilt:

$$P(A) = P(AB) + P(AB^c) = P(A)P(B) + P(AB^c)$$

Daraus folgt:

$$P(AB^c) = P(A) - P(A)P(B) = P(A)[1 - P(B)] = P(A)P(B^c)$$

In den anderen beiden Fällen argumentiert man analog.

Unabhängigkeit von drei Ereignissen: Drei Ereignisse $A, B, C \in \mathcal{A}$ sind (stochastisch) **unabhängig**, wenn:

$$\begin{aligned} P(ABC) &= P(A)P(B)P(C) \\ P(AB) &= P(A)P(B) \\ P(AC) &= P(A)P(C) \\ P(CB) &= P(C)P(B) \end{aligned}$$

Gelten nur die letzten drei Gleichungen, nennt man die Ereignisse **paarweise unabhängig**. Wie man an Beispielen zeigen kann, folgt aus der paarweisen Unabhängigkeit von A, B, C i. A. nicht deren (vollständige) Unabhängigkeit. Umgekehrt kann aus der ersten Gleichung nicht auf die Gültigkeit der paarweisen Gleichungen geschlossen werden.

Behauptung 2: Sind A, B, C ua., dann ist A auch unabhängig von jedem Ereignis, das sich aus B und C bilden lässt.

Beweis: Die Behauptung werde beispielsweise für $B \cup C$ gezeigt:

$$\begin{aligned} P(A(B \cup C)) &= P(AB \cup AC) \\ &= P(AB) + P(AC) - P(ABC) \\ &= P(A)P(B) + P(A)P(C) - P(A)P(B)P(C) \\ &= P(A)[P(B) + P(C) - P(B)P(C)] \\ &= P(A)P(B \cup C) \end{aligned}$$

Dabei wurde zweimal das Additionstheorem verwendet.

Unabhängigkeit von n Ereignissen: Die n Ereignisse $A_1, A_2, \dots, A_n \in \mathcal{A}$ sind (stochastisch) **unabhängig**, wenn für jede Teilmenge $\{i_1, i_2, \dots, i_r\}$ ($r \leq n$) von $\{1, 2, \dots, n\}$ gilt:

$$P(A_{i_1} A_{i_2} \cdots A_{i_r}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_r})$$

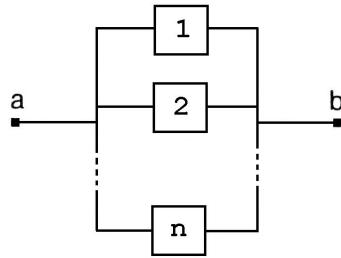
Analog sind *unendlich* viele Ereignisse (stochastisch) unabhängig, wenn jede *endliche* Teilmenge (stochastisch) unabhängig ist.

Unabhängigkeit in der Praxis: Zur Feststellung der Unabhängigkeit von n Ereignissen sind insgesamt $2^n - n - 1$ Bedingungen zu überprüfen. Für beispielsweise $n = 10$ wären das 1013 Bedingungen! Vielfach lässt sich aber bereits aus der Art eines (statistischen) Experiments auf Unabhängigkeit schließen. Wird etwa eine Münze wiederholt – unter gleichen Bedingungen – geworfen, so ist die Annahme nicht unplausibel, dass die (Ergebnisse der) einzelnen Würfe unabhängig sind.

In anderen Fällen ist die Unabhängigkeit aber eine – oft nicht überprüfbare – *Voraussetzung* für weitere Berechnungen. Beispielsweise lässt sich bei komplexen Systemen aus vielen Einzelkomponenten die Abhängigkeit zwischen den Komponenten meist nicht einfach beschreiben, sodass man in erster Näherung von deren Unabhängigkeit hinsichtlich des Ausfallverhaltens ausgeht.

Behauptung 3: Sind $A_1, A_2, \dots, A_n \in \mathcal{A}$ unabhängig, dann sind auch B_1, B_2, \dots, B_k , $k \leq n$, unabhängig, wobei jedes B_i entweder A_i oder A_i^c ist.

Bsp 2.12 Wenn ein System, bestehend aus n einzelnen Komponenten, solange funktioniert, solange zumindest eine Komponente funktioniert, nennt man es ein *Parallelsystem* (vgl. die folgende Abb). Wenn nun Komponente i , unabhängig von den anderen Komponenten, mit Wahrscheinlichkeit p_i , $i = 1, 2, \dots, n$, funktioniert, mit welcher Wahrscheinlichkeit funktioniert dann das System?



Ist C_i das Ereignis, dass Komponente i funktioniert, dann lässt sich das fragliche Ereignis C (= System funktioniert) wie folgt schreiben:

$$C = \bigcup_{i=1}^n C_i$$

Damit folgt unter Verwendung von Behauptung 3:

$$P(C) = 1 - P(C^c) = 1 - P\left(\bigcap_{i=1}^n C_i^c\right) = 1 - \prod_{i=1}^n (1 - p_i)$$

Für $p_1 = p_2 = \dots = p_n = p$ gilt: $P(C) = 1 - (1 - p)^n$. ■

2.15 Mehrstufige Experimente

Bedingte Wahrscheinlichkeiten kommen insbesondere bei **mehrstufigen** Experimenten auf natürliche Weise ins Spiel. Experimente dieser Art laufen in mehreren Stufen (oder Schritten) ab, wobei abhängig von den Ergebnissen einer Stufe verschiedene Ergebnisse auf der folgenden Stufe möglich sind. Dieser Ablauf lässt sich häufig durch einen sog. **Wahrscheinlichkeitsbaum** (vgl. das folgende Beispiel) repräsentieren.

Bsp 2.13 Man betrachte das folgende zweistufige Experiment: Zuerst wird ein regelmäßiger Tetraeder geworfen; beträgt die Augenzahl k ($= 1, 2, 3, 4$) werden anschließend k Münzen geworfen und die Zahl der dabei erzielten „Köpfe“ bestimmt. Der Ablauf des Experiments lässt sich gut durch den in Abb 2.5 wiedergegebenen Wahrscheinlichkeitsbaum verfolgen. Die Rechtecke repräsentieren die auf den einzelnen Stufen möglichen Versuchsausgänge, und in den Kreisen stehen die jeweiligen *bedingten* Wahrscheinlichkeiten (bedingt durch das Ergebnis der vorhergehenden Stufe). Beispielsweise gilt:

$$P(\# \text{ Köpfe} = 2 \mid \text{Augenzahl} = 3) = \frac{3}{8}$$

Die (unbedingte) Wahrscheinlichkeit für beispielsweise $\{\# \text{ Köpfe} = 2\}$ berechnet man mit dem **Satz von der vollständigen Wahrscheinlichkeit**:

$$\begin{aligned} P(\# \text{ Köpfe} = 2) &= \sum_{k=2}^4 P(\# \text{ Köpfe} = 2 \mid \text{Augenzahl} = k) P(\text{Augenzahl} = k) \\ &= \left(\frac{1}{4}\right) \left(\frac{1}{4}\right) + \left(\frac{3}{8}\right) \left(\frac{1}{4}\right) + \left(\frac{6}{16}\right) \left(\frac{1}{4}\right) \\ &= \frac{1}{16} + \frac{3}{32} + \frac{6}{64} = \frac{16}{64} = \frac{1}{4} \end{aligned}$$

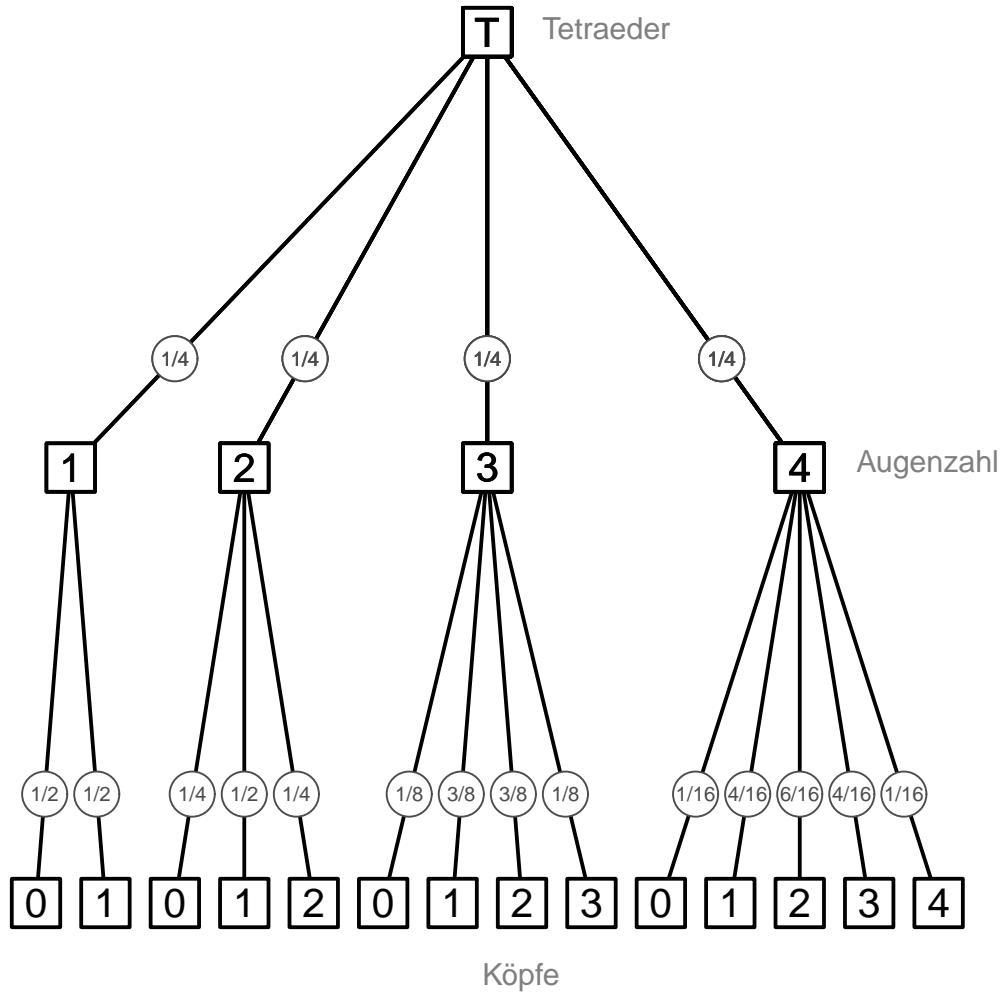
Ebenso berechnet man die anderen Wahrscheinlichkeiten:

k	0	1	2	3	4
$P(\# \text{ Köpfe} = k)$	$\frac{15}{64}$	$\frac{26}{64}$	$\frac{16}{64}$	$\frac{6}{64}$	$\frac{1}{64}$

Wie schon an anderer Stelle erwähnt, handelt es sich beim Satz v. d. vollst. W. um die Bildung eines gewichteten Mittelwerts aus bedingten Wahrscheinlichkeiten. In diesem Fall handelt es sich um die Mittelung zu gleichen Gewichten (jeweils $1/4$) der in Abb 2.6 dargestellten (bedingten) Verteilungen. ■

Formales Modell für mehrstufige Experimente: Besteht ein Zufallsexperiment aus n Stufen mit den Merkmalräumen $\Omega_1, \Omega_2, \dots, \Omega_n$, dann ist das kartesische Produkt:

Abbildung 2.5: Zweistufiges Experiment



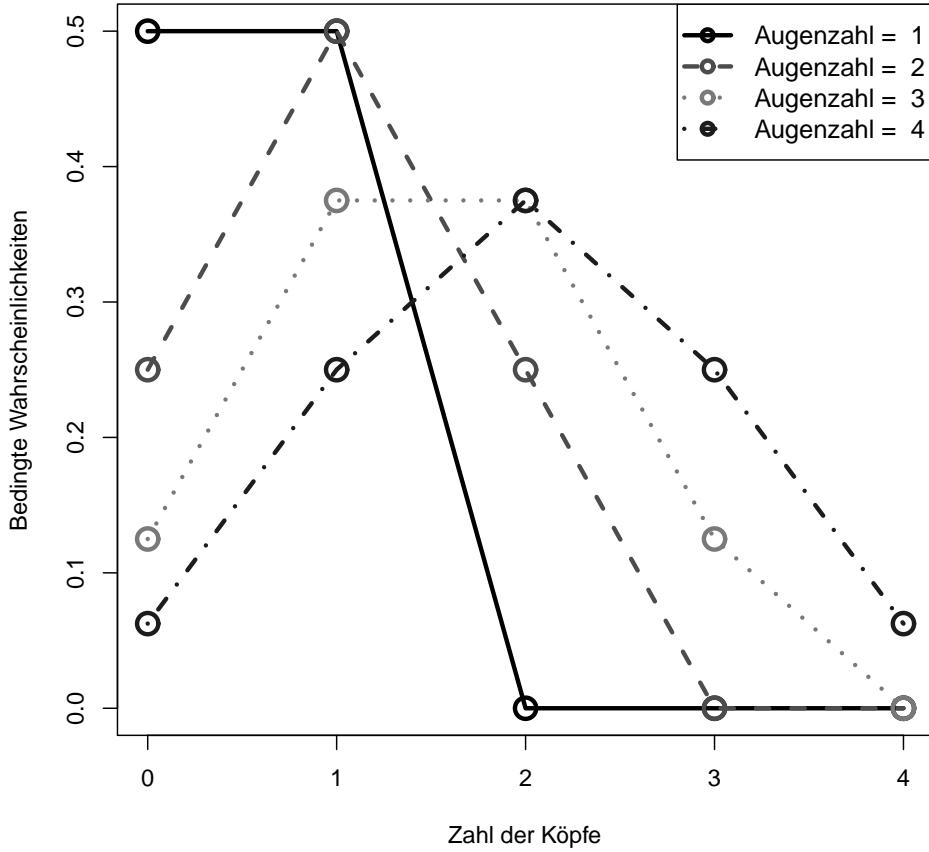
$$\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$$

aller n -Tupel $\omega = (\omega_1, \omega_2, \dots, \omega_n)$ mit $\omega_i \in \Omega_i$, $i = 1, 2, \dots, n$, ein geeigneter Merkmalsraum. Sind alle Ω_i diskret, kann man ein W-Maß auf Ω (genauer auf der zugehörigen σ -Algebra) wie folgt festlegen. Die sog. **Startverteilung** auf Ω_1 :

$$p(\omega_1) \quad \text{für } \omega_1 \in \Omega_1$$

definiert die Wahrscheinlichkeiten von Ereignissen der ersten Stufe. Gegeben den Ausgang ω_1 des ersten Experiments sei $p(\omega_2|\omega_1)$ die bedingte Wahrscheinlichkeit, dass $\omega_2 \in \Omega_2$ eintritt. Auf diese Weise fortlaufend gelangt man zur bedingten Wahrscheinlichkeit, dass ω_j eintritt, wenn auf den Stufen 1 bis $j-1$ die Ausgänge $\omega_1, \omega_2, \dots, \omega_{j-1}$ eingetreten sind:

Abbildung 2.6: Bedingte Verteilungen



$$p(\omega_j | \omega_1, \omega_2, \dots, \omega_{j-1}) \quad \text{für} \quad \omega_j \in \Omega_j$$

Für den Ausgang $\omega = (\omega_1, \omega_2, \dots, \omega_n)$ des Gesamtexperiments gilt nach dem Multiplikationstheorem:

$$p(\omega) = p(\omega_1)p(\omega_2|\omega_1) \cdots p(\omega_n|\omega_1, \omega_2, \dots, \omega_{n-1})$$

Bem: Allgemeine W–Räume dieser Art (sog. **Produkträume**) sind meist sehr komplex, sodass vereinfachende Modellannahmen notwendig sind. Beispielsweise ist bei vielen stochastischen Phänomenen die Annahme gerechtfertigt, dass die bedingten Wahrscheinlichkeiten $p(\omega_j | \omega_1, \omega_2, \dots, \omega_{j-1})$ nur vom vorherigen (letzten) Zustand abhängen:

$$p(\omega_j | \omega_1, \omega_2, \dots, \omega_{j-1}) = p(\omega_j | \omega_{j-1})$$

Ist diese Annahme auf allen Stufen erfüllt, spricht man von einer **Markow–Kette**.

2.16 Beispiele

In diesem Abschnitt betrachten wir einige Anwendungen der in den vorhergehenden Abschnitten diskutierten Konzepte und Sätze. Gelegentlich sind die Resultate etwas überraschend und/oder entsprechen nicht ganz unseren intuitiven Vorstellungen.

1. [Geburtstagsproblem] Wenn sich in einem Raum n Personen befinden, mit welcher Wahrscheinlichkeit haben alle verschiedene Geburtstage?

Da jede Person an einem der 365 möglichen Tage¹² geboren sein kann, gibt es $(365)^n$ mögliche Versuchsausgänge. Ist jeder dieser Ausgänge gleichwahrscheinlich, so ist die gesuchte Wahrscheinlichkeit gegeben durch:

$$\frac{(365)(364) \cdots (365 - n + 1)}{(365)^n}$$

Etwas überraschend ist schon für $n \geq 23$ die obige Wahrscheinlichkeit kleiner als $1/2$. Mit anderen Worten, gibt es 23 oder mehr Personen, so ist die Wahrscheinlichkeit, dass zumindest zwei von ihnen am gleichen Tag geboren sind, größer als $1/2$.

Da 23 im Vergleich zu 365 in der Regel als zu klein empfunden wird, spricht man meist vom „Geburtstagsparadoxon“. Andererseits, jedes Personenpaar hat mit Wahrscheinlichkeit:

$$\frac{365}{(365)^2} = \frac{1}{365}$$

denselben Geburtstag und da es bei 23 Personen $\binom{23}{2} = 253$ verschiedene Personenpaare gibt, erscheint das Resultat nicht mehr ganz so überraschend.

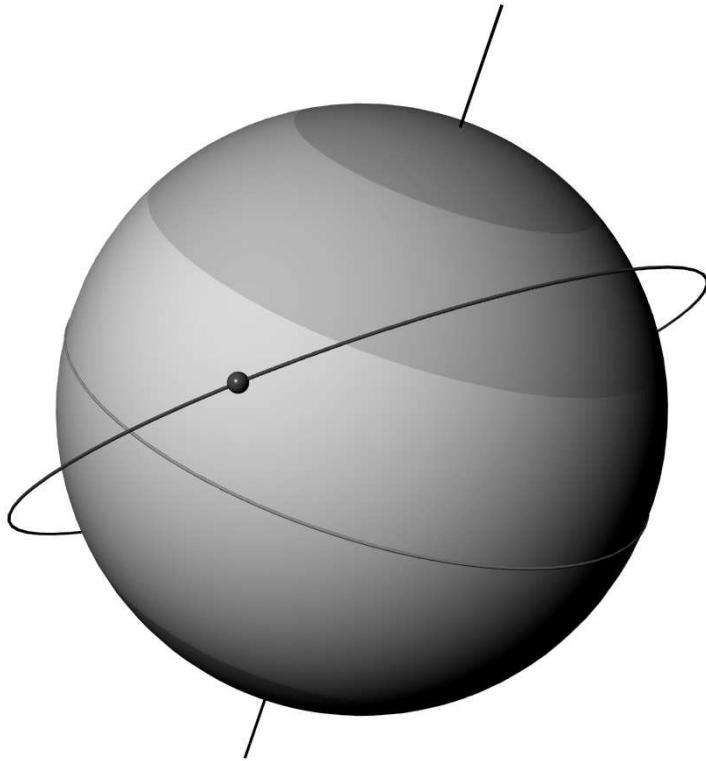
Bei 50 Personen beträgt die Wahrscheinlichkeit, dass zumindest zwei von ihnen am selben Tag geboren sind, näherungsweise 0.970, und bei 100 Personen stehen die Odds besser als 3,000,000 : 1. In letzterem Fall ist die Wahrscheinlichkeit, dass zumindest zwei Personen denselben Geburtstag teilen, also größer als $3 \times 10^6 / (3 \times 10^6 + 1)$.

2. [Satellitenproblem] Ein Satellit, dessen Orbit zwischen 60° nördlicher und 60° südlicher Breite liegt, droht abzustürzen (vgl. Abb 2.7). Wenn jeder Punkt auf dieser Erdkugelzone mit gleicher Wahrscheinlichkeit als Absturzstelle in Frage kommt, mit welcher Wahrscheinlichkeit wird der Satellit zwischen 30° und 60° nördlicher Breite abstürzen?

Abb 2.8 zeigt die (idealisierten) geometrischen Verhältnisse im Längsschnitt durch die Erdkugel. Die Fläche einer *Kugelzone* ist $A = 2\pi rh$, wobei h die Höhe der Zone ist. Nach dieser Formel beträgt die mögliche Fläche (Bem: $\cos(\pi/6) = \sqrt{3}/2$):

$$A_m = 2\pi r h_m = 2\pi r \left[2r \cos\left(\frac{\pi}{6}\right) \right] = 2\pi r^2 \sqrt{3}$$

¹²Schaltjahre bleiben unberücksichtigt; sie beeinflussen die Resultate nur unwesentlich.

Abbildung 2.7: Satellitenproblem (Erdkugel)

Die dem fraglichen Ereignis entsprechende Fläche (repräsentiert durch das dunklere Grau) beträgt (Bem: $\cos(\pi/3) = 1/2$):

$$A_g = 2\pi r h_g = 2\pi r \left[r \cos\left(\frac{\pi}{6}\right) - r \cos\left(\frac{\pi}{3}\right) \right] = \pi r^2 [\sqrt{3} - 1]$$

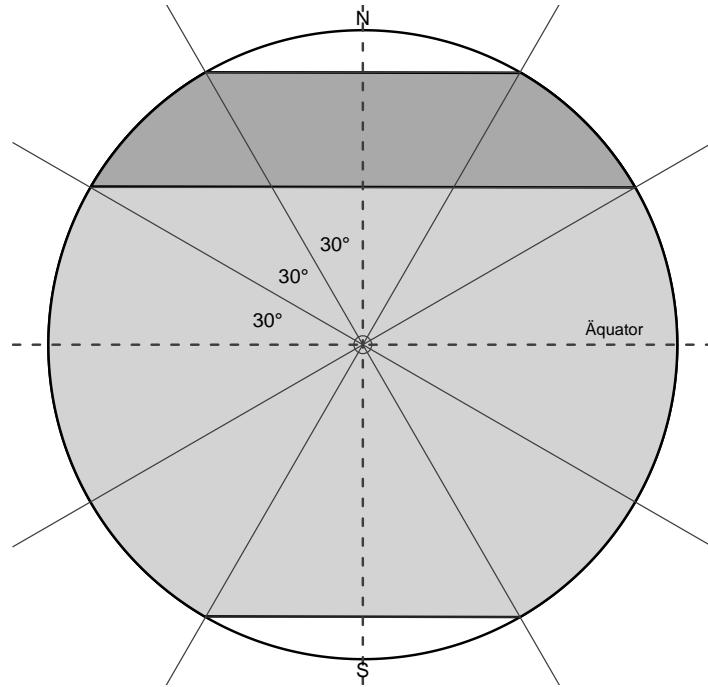
Die gesuchte Wahrscheinlichkeit beträgt also:

$$p = \frac{A_g}{A_m} = \frac{\sqrt{3} - 1}{2\sqrt{3}} \doteq 0.2113$$

3. [Matchingproblem¹³] Das Matchingproblem existiert in zahlreichen Einkleidungen, etwa in der folgenden: Man betrachte zwei zufällige Permutationen der Zahlen $1, 2, \dots, N$ und zähle die Übereinstimmungen. Beispielsweise seien für $N = 10$ die Permutationen gegeben wie folgt:

$$\begin{array}{cccccccccc} 1 & 2 & 4 & 8 & 7 & 5 & 6 & 3 & 10 & 9 \\ 3 & 9 & 8 & 7 & 5 & 10 & 6 & 2 & 1 & 4 \end{array}$$

¹³Auch *Montmort'sches Problem*, benannt nach dem franz. Mathematiker PIERRE-REMOND MONTMORT (1678–1719), der sich als erster mit Problemen dieser Art beschäftigte.

Abbildung 2.8: Satellitenproblem (Längsschnitt durch die Erdkugel)

In diesem Fall gibt es genau eine Übereinstimmung. Äquivalent kann man auch nur eine zufällige Permutation betrachten und die Übereinstimmungen mit der nicht permutierten Folge zählen. Beispielsweise (wieder für $N = 10$):

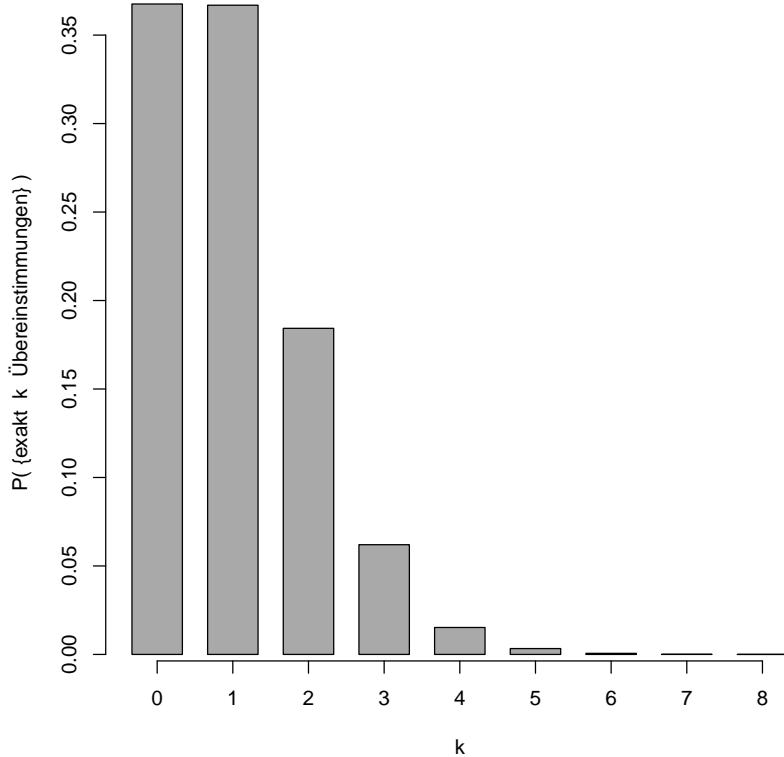
$$\begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 3 & 10 & 2 & 9 & 4 & 5 & 6 & 8 & 7 & 1 \end{array}$$

In diesem Fall ergibt sich ebenfalls genau eine Übereinstimmung.

Eine interessante Frage lautet: Mit welcher Wahrscheinlichkeit gibt es *keine* Übereinstimmung? Dazu berechnen wir zuerst die Wahrscheinlichkeit des komplementären Ereignisses von *zumindest einer* Übereinstimmung. Sei E_i , $i = 1, 2, \dots, N$, das Ereignis, dass die Zahlen auf der i -ten Position übereinstimmen und entsprechend $E_{i_1}E_{i_2}\cdots E_{i_n}$ das Ereignis, dass die Zahlen auf den Positionen i_1, i_2, \dots, i_n übereinstimmen. Da alle $N!$ möglichen Permutationen gleichwahrscheinlich sind (Laplace-Raum), gilt:

$$P(E_i) = \frac{(N-1)!}{N!} = \frac{1}{N} \quad \text{und} \quad P(E_{i_1}E_{i_2}\cdots E_{i_n}) = \frac{(N-n)!}{N!}$$

Es gibt $\binom{N}{n}$ Ereignisse der Form $E_{i_1}E_{i_2}\cdots E_{i_n}$; also gilt:

Abbildung 2.9: Simulation zum Matchingproblem (für $N = 25$)

$$\sum_{i_1 < i_2 < \dots < i_n} P(E_{i_1} E_{i_2} \cdots E_{i_n}) = \frac{N!(N-n)!}{(N-n)!n!N!} = \frac{1}{n!}$$

Nach dem Additionstheorem ist die Wahrscheinlichkeit von $\bigcup_{i=1}^N E_i$ ($\hat{\equiv}$ zumindest eine Übereinstimmung) gegeben durch:

$$P\left(\bigcup_{i=1}^N E_i\right) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{N+1} \frac{1}{N!}$$

Die gesuchte Wahrscheinlichkeit beträgt also:

$$P(\{\text{Keine Übereinstimmung}\}) = \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^N}{N!} = \sum_{j=0}^N \frac{(-1)^j}{j!}$$

Letzterer Ausdruck ist die N -te Partialsumme der Reihe für $e^{-1} = 0.3678\dots$. Die Exponentialreihe konvergiert sehr schnell; bereits für $N \geq 5$ ist der Grenzwert auf zwei Stellen genau erreicht. Bemerkenswerterweise ist also die Wahrscheinlichkeit einer *völligen Unordnung* (d. h. keine Übereinstimmung) nahezu konstant gleich 0.37. (Erwartet hätte man vielleicht $P \rightarrow 1$ für $N \rightarrow \infty$?)

Auf ähnliche Weise zeigt man:

$$P(\{\text{Exakt } k \text{ Übereinstimmungen}\}) = \frac{1}{k!} \sum_{j=0}^{N-k} \frac{(-1)^j}{j!} \quad \rightarrow \quad \frac{e^{-1}}{k!}$$

Beispielsweise sind die Wahrscheinlichkeiten für $N = 5$ gegeben wie folgt:

k	0	1	2	3	4	5
P	0.3667	0.3750	0.1667	0.0833	0	0.0083

(Man beachte, dass exakt $N - 1 = 4$ Übereinstimmungen unmöglich sind.) Das Matchingproblem lässt sich einfach simulieren: Abb 2.9 zeigt den Barplot für die Anzahl der Übereinstimmungen von 100000 simulierten Permutationen von $1, 2, \dots, 25$. Die relativen Häufigkeiten unterscheiden sich praktisch nicht von den Grenzwerten.

4. [Diagnostische Tests] Angenommen, ein Bluttest entdeckt zu 95% eine Krankheit, wenn sie tatsächlich vorhanden ist, liefert aber auch zu 1% ein „falsch positives“ Ergebnis (d. h., reagiert positiv bei einer nicht erkrankten Person). Wenn angenommen 0.5% der Population erkrankt sind, mit welcher Wahrscheinlichkeit ist eine Person, deren Bluttest positiv ist, tatsächlich erkrankt?

Bezeichne $D+/D-$ das Ereignis, dass die getestete Person erkrankt/nicht erkrankt ist, und $T+/T-$ das Ereignis, dass der Test positiv/negativ ist. Dann gilt unter Verwendung der Bayes'schen Formel:

$$\begin{aligned} P(D+|T+) &= \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)} \\ &= \frac{(0.95)(0.005)}{(0.95)(0.005) + (0.01)(0.995)} \\ &= \frac{95}{294} = 0.323 \end{aligned}$$

D. h., nur ca. 32% der Personen, deren Test positiv ist, sind auch tatsächlich erkrankt! Um dieses etwas überraschende Resultat – erwartet hätte man eine deutlich höhere Wahrscheinlichkeit, da der Test augenscheinlich nicht schlecht ist – besser zu verstehen, stellen wir uns vor, dass 10000 (willkürlich herausgegriffene) Personen getestet werden. Unter den obigen Bedingungen erhalten wir (im Durchschnitt) das folgende Bild:

	Test		gesamt
	positiv	negativ	
erkrankt	47.5	2.5	50
nicht erkrankt	99.5	9850.5	9950
gesamt	147.0	9853.0	10000

D.h., nur bei rund 1/3 der Fälle ist ein positives Testergebnis auf die Erkrankung, bei 2/3 der Fälle aber auf andere Effekte zurückzuführen. Andererseits, ist der Test negativ, kann man eine Erkrankung praktisch ausschließen:

$$\begin{aligned} P(D-|T-) &= \frac{P(T-|D-)P(D-)}{P(T-|D-)P(D-) + P(T-|D+)P(D+)} \\ &= \frac{(0.99)(0.995)}{(0.99)(0.995) + (0.05)(0.005)} \\ &= \frac{19701}{19706} \doteq 0.9997 \end{aligned}$$

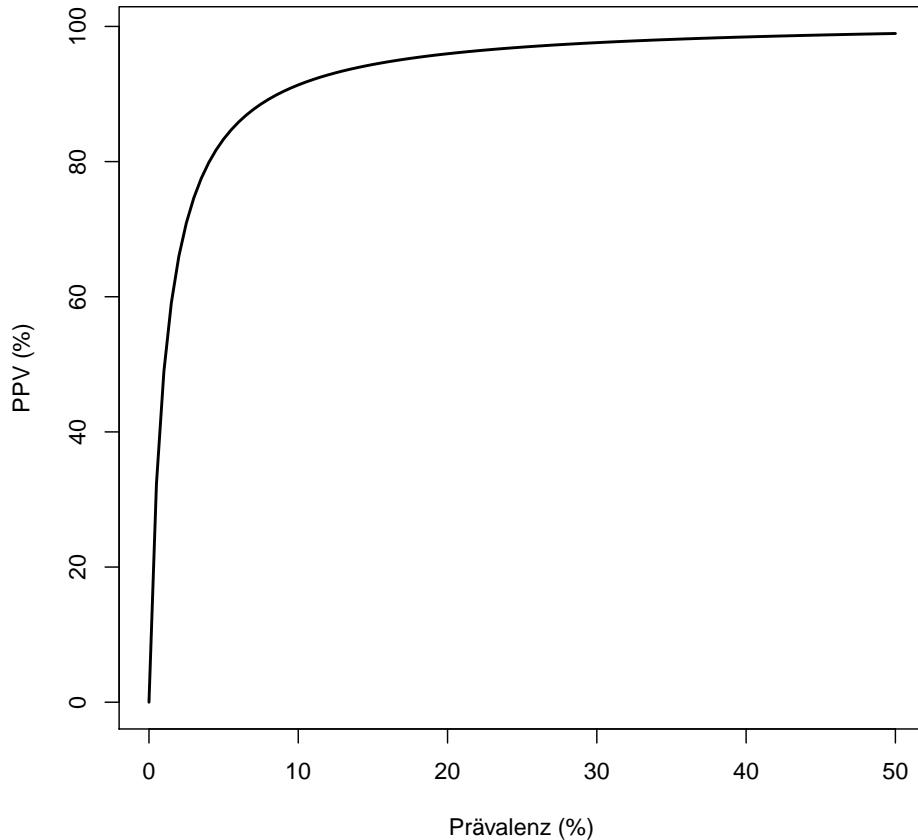
Speziell im medizinischen Kontext sind die folgenden Ausdrücke gebräuchlich:

- (a) Die bedingten Wahrscheinlichkeiten $P(T+|D+)$, genannt die **Sensitivität** (auch **korrekt-positiv Rate**), und $P(T-|D-)$, genannt die **Spezifität** (auch **korrekt-negativ Rate**), bestimmen die Güte des diagnostischen Tests. Um diese Werte bestimmen zu können, muss der tatsächliche Gesundheitszustand der Probanden, der sog. **Goldstandard**, bekannt sein.
- (b) Die A-priori-Wahrscheinlichkeit $P(D+)$ nennt man die **Prävalenz** der fraglichen Krankheit. Sie ist definiert als die relative Häufigkeit der Krankheitsfälle an der gesamten Population. Für die Prävalenz gibt es meist nur mehr oder weniger grobe Schätzwerte, darüberhinaus kann sie für verschiedene Risikogruppen auch beträchtlich variieren (vgl. dazu auch Punkt (c)).
- (c) Die A-posteriori-Wahrscheinlichkeiten $P(D+|T+)$ und $P(D-|T-)$ nennt man den **positiv prädiktiven Wert** (PPV) bzw. den **negativ prädiktiven Wert** (NPV). Da die prädiktiven Werte stark von der Prävalenz abhängen, kommt der möglichst genauen Bestimmung von letzterer eine große Bedeutung zu. (Vgl. Abb 2.10 für die PPV-Kurve unter den Bedingungen des vorliegenden Beispiels.)

Die Odds-Form der Bayes'schen Formel ist hier von besonderer Bedeutung. Abhängig vom Testergebnis gilt:

$$\frac{P(D+|T+)}{P(D-|T+)} = \frac{P(D+)}{P(D-)} \times \underbrace{\frac{P(T+|D+)}{P(T+|D-)}}_{\text{LR}^+}$$

$$\frac{P(D+|T-)}{P(D-|T-)} = \frac{P(D-)}{P(D+)} \times \underbrace{\frac{P(T-|D+)}{P(T-|D-)}}_{\text{LR}^-}$$

Abbildung 2.10: PPV als Funktion der Prävalenz (Sens = 95%, Spez = 99%)

Für das vorliegende Beispiel sind die Likelihood-Ratios gegeben durch:

$$LR^+ = \frac{0.95}{1 - 0.99} = 95 \quad LR^- = \frac{1 - 0.95}{0.99} = 0.051$$

Diese Werte lassen sich wie folgt interpretieren: Die Odds für das Vorliegen der Erkrankung erhöhen sich bei einem positiven Test um das 95-fache, bei negativem Testergebnis reduzieren sie sich um $1 - 0.051 = 94.9\%$.

5. [Monty Hall Problem¹⁴] Bei einer Spielshow gibt es drei Türen. Hinter einer Tür befindet sich ein wertvoller Preis, hinter den beiden anderen aber nur Preise ohne Wert. Ein Spielkandidat wählt zufällig eine Tür und anschließend öffnet der Showmaster eine der beiden anderen Türen, hinter der sich (natürlich) ein wertloser Preis befindet. Nun bekommt der Kandidat die Möglichkeit, seine erste Wahl zu revidieren und zu der vom Showmaster nicht geöffneten Tür zu wechseln. Sollte der Kandidat diese Möglichkeit ergreifen?

¹⁴Populär geworden durch MONTY HALL (*1921 als MAURICE HALPERIN, kanad. Showmaster und TV-Produzent), auch bekannt unter dem Namen *Ziegenproblem*; vgl. unter diesem Stichwort WIKIPEDIA für eine ausführliche Diskussion.

Beharrt der Kandidat auf der ersten Wahl, beträgt seine Gewinnwahrscheinlichkeit $1/3$. Betrachten wir nun die *bedingte* Gewinnwahrscheinlichkeit, wenn der Kandidat wechselt. O. B. d. A. befindet sich der wertvolle Preis hinter Tür 1. Angenommen, der Kandidat hat zunächst Tür 2 gewählt. Der Showmaster kann dann nur Tür 3 öffnen, der Kandidat wechselt zu Tür 1 und gewinnt. Ebenso, wenn der Kandidat zunächst Tür 3 gewählt hat. Hat der Kandidat aber zunächst Tür 1 gewählt, führt ein Wechsel zu einer Niete. Wenn der Kandidat wechselt, beträgt die Gewinnwahrscheinlichkeit also $2/3$.

Bem: Gelegentlich besteht die Meinung, dass auch im Falle, dass der Kandidat nicht wechselt, seine Gewinnwahrscheinlichkeit – ohne sein Zutun – von $1/3$ auf $1/2$ steigt. Nach dem Öffnen einer Tür durch den Showmaster bleiben schließlich nur zwei Türen übrig (und hinter einer befindet sich der Preis). Da der Showmaster aber *immer* eine Tür mit dahinter befindlicher Niete öffnen wird, kann sich diese Aktion nur dann auf die Gewinnwahrscheinlichkeit auswirken, wenn der Kandidat seine erste Wahl revidiert.

Wir betrachten auch eine *formale Lösung*: Sei G_i das Ereignis, dass sich der Preis hinter Tür i befindet, $i = 1, 2, 3$, und S_j das Ereignis, dass der Showmaster Tür j öffnet, $j = 1, 2, 3$. O. B. d. A. werde angenommen, dass der Kandidat Tür 1 wählt und der Showmaster danach Tür 3 öffnet. In Ermangelung anderweitiger Informationen lässt sich in diesem Fall annehmen, dass:

$$\begin{aligned} P(G_1) &= P(G_2) = P(G_3) = \frac{1}{3} \\ P(S_3|G_1) &= \frac{1}{2} \quad P(S_3|G_2) = 1 \quad P(S_3|G_3) = 0 \end{aligned}$$

Interessant ist nun die Frage, mit welcher Wahrscheinlichkeit sich *a-posteriori* (d. h. nach Öffnen von Tür 3) der Preis hinter Tür 2 befindet. Nach der Bayes'schen Formel gilt:

$$\begin{aligned} P(G_2|S_3) &= \frac{P(S_3|G_2)P(G_2)}{P(S_3|G_1)P(G_1) + P(S_3|G_2)P(G_2) + P(S_3|G_3)P(G_3)} \\ &= \frac{(1)(1/3)}{(1/2)(1/3) + (1)(1/3) + (0)(1/3)} = \frac{2}{3} \end{aligned}$$

Durch einen Wechsel lässt sich die Gewinnwahrscheinlichkeit von $1/3$ auf $2/3$ verdoppeln, oder die A-priori-Odds lassen sich von $1 : 2$ auf $2 : 1$ vervierfachen.

(UE-Aufgabe: Bedeutet eine Verdoppelung der Wahrscheinlichkeit stets eine Vervierfachung der Odds?)

6. [Unabhängigkeit/Disjunktheit] Häufig wird – quasi ganz „automatisch“ – von der Disjunktheit zweier Ereignisse A und B (d. h. von $A \cap B = \emptyset$) auf deren Unabhängigkeit

(d. h. auf $P(A \cap B) = P(A)P(B)$) geschlossen. Das ist aber keineswegs der Fall! (Im Gegenteil!) Betrachten wir dazu das Werfen eines (üblichen) Würfels und die Ereignisse $A_1 = \{2, 4, 6\}$ (Augenzahl gerade) und $B_1 = \{1, 3, 5\}$ (Augenzahl ungerade). Zwar gilt $A \cap B = \emptyset$, aber:

$$P(A_1 \cap B_1) = P(\emptyset) = 0 \neq P(A_1)P(B_1) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$$

D.h., A_1 und B_1 sind *nicht* unabhängig. Andererseits sind z. B. $A_2 = \{1, 2, 3\}$ und $B_2 = \{3, 4\}$ zwar nicht disjunkt, aber unabhängig:

$$P(A_2 \cap B_2) = P(\{3\}) = \frac{1}{6} = P(A_2)P(B_2) = \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)$$

Die Erkenntnisse aus dem obigen Beispiel lassen sich wie folgt zusammenfassen (Vs.: $P(A) > 0, P(B) > 0$): Nur Ereignisse, die etwas *gemeinsam* haben (d. h. für die $A \cap B \neq \emptyset$), können auch unabhängig sein! Ob Letzteres tatsächlich der Fall ist, hängt dann von der W-Verteilung ab.

Haben wir beispielsweise einen Würfel, bei dem die Wahrscheinlichkeit, die Augenzahl k zu werfen, proportional zu k ist, d. h. für den $P(\{k\}) = ck$, $k = 1, 2, \dots, 6$, für eine Konstante $c > 0$, so gilt:

$$c \underbrace{\sum_{k=1}^6 k}_{21} = 1 \implies c = \frac{1}{21}$$

Betrachten wir wieder die obigen Ereignisse, so gilt nach wie vor, dass die disjunkten Ereignisse A_1 und B_1 nicht unabhängig sind. Für A_2 und B_2 gilt aber:

$$P(A_2 \cap B_2) = P(\{3\}) = \frac{3}{21} \neq P(A_2)P(B_2) = \left(\frac{6}{21}\right)\left(\frac{7}{21}\right) = \frac{2}{21}$$

D.h., auch A_2 und B_2 sind bei diesem Würfel nicht unabhängig.

Aufgaben

- 2.1 Aus den Anfängen der Wahrscheinlichkeitsrechnung: Der französische Offizier und Schriftsteller CHEVALIER DE MÉRÉ (1607–1684) wandte sich im Jahre 1654 mit der folgenden Frage an BLAISE PASCAL (1623–1662): Was ist vorteilhafter, beim Spiel mit einem Würfel auf das Eintreten mindestens eines Sechzers in vier Würfen oder beim Spiel mit zwei Würfeln auf das Eintreten eines Doppelsechzers in 24 Würfen zu setzen? Als leidenschaftlicher Spieler wusste De Méré, dass die erste

Wette für ihn vorteilhaft ist. Bei der zweiten Wette, von der er annahm, dass sie nur eine Variante der ersten sei, gestalteten sich die Einnahmen aber nicht ganz nach seinen Vorstellungen. Bearbeiten Sie das Problem empirisch unter Verwendung der Funktion `demere()`. (Wie lauten die exakten Wahrscheinlichkeiten?)

- 2.2 Zeigen Sie die Gültigkeit der De Morgan'schen Regeln (vgl. 2.3).
- 2.3 Ermitteln Sie einen möglichst einfachen Ausdruck für das (zusammengesetzte) Ereignis, dass von drei Ereignissen A , B und C :
 - (a) nur A eintritt
 - (b) A und C aber nicht B eintritt
 - (c) zumindest eines eintritt
 - (d) zumindest zwei eintreten
 - (e) alle drei eintreten
 - (f) keines eintritt
 - (g) höchstens eines eintritt
 - (h) höchstens zwei eintreten
 - (i) genau zwei eintreten
 - (j) höchstens drei eintreten
- 2.4 Zeigen Sie, dass (a) alle offenen Intervalle (a, b) , (b) alle abgeschlossenen Intervalle $[a, b]$ und (c) alle Intervalle der Form $(-\infty, a]$ Borelmengen sind.
- 2.5 Jemand behauptet, sechs verschiedene Weinproben den in einer zufälligen Reihenfolge aufgelegten Weinetiketten zuordnen zu können.
 - (a) Wie lautet ein passender Merkmalraum Ω ?
 - (b) Wenn er/sie nur rät, wie lautet eine entsprechende W-Verteilung auf $\mathcal{P}(\Omega)$?
 - (c) Mit welcher Wahrscheinlichkeit werden exakt/zumindest vier Weine richtig zugeordnet, wenn er/sie nur rät?
- 2.6 Zeigen Sie für Ereignisse A_1, A_2, \dots die **Bonferroni-Ungleichung**:¹⁵

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) \geq 1 - \sum_{i=1}^{\infty} P(A_i^c)$$

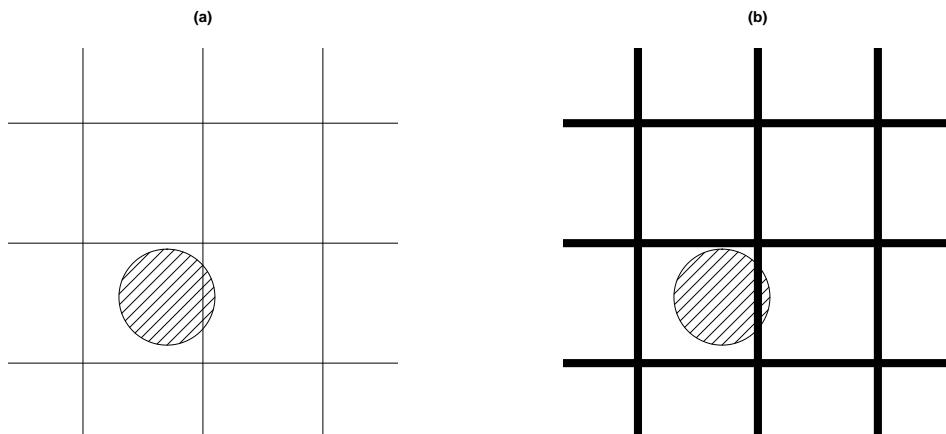
Zeigen Sie, dass die Ungleichung für endlich viele Ereignisse A_1, A_2, \dots, A_n auch wie folgt geschrieben werden kann:

$$P\left(\bigcap_{i=1}^n A_i\right) \geq \sum_{i=1}^n P(A_i) - (n - 1)$$

¹⁵CARLO EMILIO BONFERRONI (1892–1960), ital. Mathematiker.

- 2.7 Um die Dauer einer Meisterschaft abzukürzen, werden die $4n$ teilnehmenden Mannschaften durch Los in 4 gleich große Gruppen aufgeteilt.
- Wieviele verschiedene Aufteilungen gibt es?
 - Wieviele Aufteilungen gibt es, sodass sich die zwei stärksten Mannschaften der Meisterschaft in verschiedenen Gruppen befinden?
 - Mit welcher Wahrscheinlichkeit befinden sich die zwei stärksten Mannschaften in verschiedenen Gruppen? (Ermitteln Sie einen möglichst einfachen Ausdruck.)
- 2.8 Wieviele Personen müssen sich in einem Raum befinden, sodass die Wahrscheinlichkeit größer als $1/2$ ist, dass zumindest zwei von ihnen im selben Monat geboren sind? (Hinweis: Man nehme einfachheitshalber an, dass jeder Monat mit gleicher Wahrscheinlichkeit als Geburtsmonat in Frage kommt.)
- 2.9 Die 8 Titel auf einer CD werden in zufälliger Reihenfolge abgespielt. Mit welcher Wahrscheinlichkeit wird dabei kein/genau ein Titel an der auf der CD angegebenen Stelle wiedergegeben?
- 2.10 S sei eine Menge und S_1, S_2, \dots, S_k seien paarweise disjunkte nichtleere Teilmengen von S , sodass $\bigcup_{i=1}^k S_i = S$. Dann nennt man $\{S_1, S_2, \dots, S_k\}$ eine *Partition* von S . Bezeichnet T_n die Anzahl verschiedener Partitionen von $S = \{1, 2, \dots, n\}$, so gilt $T_1 = 1$ (die einzige Partition von $\{1\}$ ist $\{\{1\}\}$) und $T_2 = 2$ (die zwei Partitionen von $\{1, 2\}$ sind $\{\{1, 2\}\}$ und $\{\{1\}, \{2\}\}$).
- Zeigen Sie direkt, dass $T_3 = 5$ und $T_4 = 15$.
 - Zeigen Sie die Rekursion:
- $$T_{n+1} = 1 + \sum_{k=1}^n \binom{n}{k} T_k$$
- Verwenden Sie diese Beziehung zur Berechnung von T_{10} .
(Hinweis: Eine Möglichkeit, eine Partition von $n+1$ Elementen zu wählen, besteht darin, zunächst ein Element als *speziell* zu kennzeichnen. Anschließend wählt man ein k , $k = 0, 1, \dots, n$, eine Teilmenge der Größe $n-k$ aus den nichtspeziellen Elementen und eine der T_k Partitionen der restlichen k nichtspeziellen Elemente. Gibt man nun das spezielle Element zur vorhin gewählten Teilmenge der Größe $n-k$, bekommt man eine Partition aller $n+1$ Elemente.)
- 2.11 Zehn Studenten und fünf Studentinnen werden zufällig in fünf Arbeitsgruppen zu je drei Personen aufgeteilt. Mit welcher Wahrscheinlichkeit gibt es in jeder Arbeitsgruppe eine Studentin? (Hinweis: Verwenden Sie das Multiplikationstheorem, geben Sie aber auch eine kombinatorische Lösung.)
- 2.12 Geben Sie eine kombinatorische Lösung für das Problem von Bsp 2.8. Wie lautet ein passender Merkmalraum?

- 2.13 Betrachten Sie die folgende Variante des Geburtagsproblems: Angenommen, Sie wollen jemanden finden, der am selben Tag wie Sie geboren ist. Welche Mindestanzahl von Personen müssen Sie befragen, damit die Chancen dafür etwa 50:50 stehen? Wieviele, wenn die Chancen dafür größer sein sollen, etwa 90:10 ? (Zuerst raten!)
- 2.14 Bei einem Spiel wird eine Münze auf eine in quadratische Felder aufgeteilte Tischplatte geworfen. Man gewinnt, falls die Münze zur Gänze innerhalb eines Quadrats zu liegen kommt (Abb (a)). Unter der Voraussetzung, dass die Münze auf dem Tisch landet, wie groß ist die Gewinnwahrscheinlichkeit? (Seitenlänge eines Quadrats = L ; Durchmesser der Münze = $D (< L)$)



Was ändert sich, wenn die Dicke Δ der Begrenzungslinien nicht vernachlässigt werden kann (Abb (b))?

- 2.15 Fortsetzung des Rendezvousproblems (Bsp 2.5): Wie groß ist die Wahrscheinlichkeit, dass sich um 10:30 (i) weder A noch B, (ii) A oder B aber nicht beide, (iii) A oder B, (iv) A und B am Aussichtspunkt befinden? (Hinweis: Argumentieren Sie geometrisch.)
- 2.16 Auf einem (dünnen) Holzstab der Länge $L = 1$ [m] werden willkürlich zwei Stellen markiert; anschließend wird der Stab an diesen Stellen durchgesägt. Mit welcher Wahrscheinlichkeit lässt sich aus den so entstehenden Stücken ein Dreieck bilden? (Hinweis: Argumentieren Sie geometrisch.) Zusatz: Simulieren Sie das Experiment und bestätigen Sie empirisch die gefundene Lösung.
- 2.17 Zeigen Sie, dass die Wahrscheinlichkeit, mit der genau eines der Ereignisse A , B und C eintritt, gegeben ist durch:

$$P(A) + P(B) + P(C) - 2P(AB) - 2P(AC) - 2P(BC) + 3P(ABC)$$

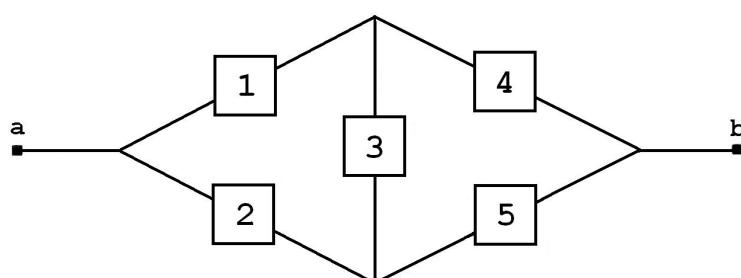
- 2.18 Zeigen Sie, dass für $P(B) > 0$ die bedingten Wahrscheinlichkeiten $P(\cdot | B)$ (vgl. 2.10) alle Eigenschaften eines W-Maßes auf (Ω, \mathcal{A}) erfüllen.

- 2.19 An einem bestimmten Punkt der Ermittlungen ist der Kommissar zu 60% davon überzeugt, dass der Hauptverdächtige der Täter ist. Ein *neues* Beweisstück zeigt, dass der Täter eine bestimmte Eigenart (Linkshänder, braune Haare, o. dgl.) hat. Wenn 20% der Bevölkerung diese Eigenart aufweist, wie überzeugt kann der Kommissar nun sein, wenn sich herausstellt, dass der Verdächtige diese Eigenart hat? (Hinweis: Bayes'sche Formel; verwenden Sie letztere auch in der Odds-Form.)
- 2.20 Ein Würfelpaar wird solange geworfen, bis die Augensumme 5 oder 7 kommt. Mit welcher Wahrscheinlichkeit kommt die Augensumme 5 zuerst? (Hinweis: E_n sei das Ereignis, dass beim n -ten Wurf 5 kommt, aber weder 5 noch 7 bei den ersten $n - 1$ Würfen. Bestimmen Sie $P(E_n)$ und argumentieren Sie, dass $\sum_{n=1}^{\infty} P(E_n)$ die gesuchte Wahrscheinlichkeit ist.)
- 2.21 Man betrachte zwei disjunkte Ereignisse A und B , die bei einem Experiment eintreten können, wobei $P(A) > 0$, $P(B) > 0$ und $P(A) + P(B) \leq 1$. Das Experiment werde solange (unabhängig) wiederholt, bis A oder B eintritt. Mit welcher Wahrscheinlichkeit kommt A vor B ? Zeigen Sie, dass letztere Wahrscheinlichkeit gegeben ist durch:

$$\frac{P(A)}{P(A) + P(B)}$$

(Hinweis: Lässt sich analog zu Aufgabe 2.20 zeigen. Als Alternative kann man aber auch durch das Ergebnis des *ersten* Experiments *bedingen* und den Satz von der vollständigen Wahrscheinlichkeit verwenden. Bearbeiten Sie die Aufgabe mit beiden Methoden.)

- 2.22 Eine Zahl wird zufällig aus der Menge $\{1, \dots, 30\}$ ausgewählt. A sei das Ereignis, dass diese Zahl gerade ist, B das Ereignis, dass sie durch 3 teilbar ist und C das Ereignis, dass sie durch 5 teilbar ist. Diskutieren Sie die stochastische Unabhängigkeit dieser Ereignisse.
- 2.23 Das folgende System ist intakt, wenn es einen Pfad aus intakten Komponenten von a nach b gibt. Dabei nehme man an, dass jede Komponente – unabhängig von den anderen – mit Wahrscheinlichkeit p ($0 \leq p \leq 1$) intakt ist.



- (a) Wie lautet ein passender Merkmalraum Ω ? Wieviele Elemente hat er? Wie lautet und aus wievielen Elementen besteht die zugehörige σ -Algebra? Wie ist $P(\{\omega\})$ für $\omega \in \Omega$ definiert?

- (b) Beschreiben Sie auf Basis des gewählten Merkmalraums die Ereignisse $A_i = \{\text{Komponente } i \text{ ist intakt}\}$ und $A = \{\text{System ist intakt}\}$.
- (c) Berechnen Sie $P(A)$ und stellen Sie letztere Wahrscheinlichkeit in Abhängigkeit von p grafisch dar. (Hinweis: $P(A)$ kann mit Hilfe des Additionstheorems berechnet werden oder mit Hilfe des Satzes v. d. vollst. Wahrscheinlichkeit, indem man nach dem Zustand (defekt/intakt) der „Brücke“ (Komponente 3) bedingt. Versuchen Sie beide Lösungen.)
- 2.24 Jemand fliegt von Los Angeles nach Wien mit Zwischenlandungen in New York, London und Frankfurt. Bei jeder Zwischenlandung wird die Maschine gewechselt, wobei an jedem Flughafen (einschließlich LA) das Gepäck mit gleichbleibender Wahrscheinlichkeit q ($0 < q < 1$) in ein falsches Flugzeug verladen wird. In Wien fehlt das Gepäck; mit welcher Wahrscheinlichkeit ist der Fehler in LA, NY, L, F passiert? Wo ist die Wahrscheinlichkeit am größten? Rechnen Sie allgemein und für $q = 0.05$. Stellen Sie den Weg des Gepäcks in Form eines W-Baums dar.
- 2.25 Betrachten Sie das folgende zweistufige Experiment: Zuerst werfen Sie einen Würfel; anschließend soviele (gleichartige) Münzen wie die zuvor geworfene Augenzahl und zählen dann die Zahl der „Köpfe“. Wenn A_k das Ereignis ist, dass es k Köpfe gibt, bestimmen Sie $P(A_k)$ für $k = 0, 1, \dots, 6$. Welche Zahl von Köpfen ist am wahrscheinlichsten? (Hinweis: Vgl. Bsp 2.13.)

Anhang: Abzählende Kombinatorik

Die **abzählende Kombinatorik**¹⁶ untersucht die Anzahlen möglicher Anordnungen oder Auswahlen von unterscheidbaren oder nicht unterscheidbaren Objekten mit oder ohne Beachtung der Reihenfolge.

- (1) Allgemeines Zählprinzip: Wenn eine Aufgabe durch eine Abfolge von k Schritten beschrieben werden kann, und wenn Schritt 1 auf n_1 verschiedene Arten erledigt werden kann, und wenn Schritt 2 – für jede Art der ersten Stufe – auf n_2 verschiedene Arten erledigt werden kann, usf., dann ist die Zahl der verschiedenen Möglichkeiten, die Aufgabe zu erledigen, gegeben durch:

$$n_1 n_2 \cdots n_k$$

- (2) Permutationen: Anordnungen von n Objekten, wobei alle Objekte vorkommen, mit Beachtung der Reihenfolge.
- (a) Unterscheidbare Objekte: Die Zahl der Permutationen von n verschiedenen Objekten beträgt:

¹⁶Allgemeiner ist die *Kombinatorik* jenes Teilgebiet der *diskreten* Mathematik, das sich mit endlichen oder abzählbar unendlichen diskreten Strukturen beschäftigt.

$$n! = n(n-1)(n-2) \cdots (2)(1)$$

- (b) **Objekte mehrerer Klassen:** Die Zahl der Permutationen von n Objekten, die in k Klassen zu je n_1, n_2, \dots, n_k ($\sum_{i=1}^k n_i = n$) gleichen Objekten vorliegen, beträgt:

$$\frac{n!}{n_1! n_2! \cdots n_k!} = \binom{n}{n_1, n_2, \dots, n_k}$$

Bsp: Wieviele verschiedene Barcodes aus vier dicken, drei mittleren und zwei dünnen Linien gibt es?

$$\frac{9!}{4! 3! 2!} = 1260$$

- (3) **Variationen:** Auswählen von Objekten mit Beachtung der Reihenfolge.

- (a) **Ohne Zurücklegen:** Die Zahl der Möglichkeiten aus n verschiedenen Objekten k ($\leq n$) Objekte ohne Zurücklegen und unter Beachtung der Reihenfolge auszuwählen, beträgt:

$$(n)_k = n(n-1) \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

Bsp: Auf einer Platine gibt es acht verschiedene Stellen, an denen eine Komponente plaziert werden kann. Wenn vier verschiedene Komponenten plaziert werden sollen, wieviele verschiedene Designs gibt es?

$$(8)_4 = (8)(7)(6)(5) = \frac{8!}{4!} = 1680$$

- (b) **Mit Zurücklegen:** Für die Auswahl von k Objekten aus n verschiedenen Objekten mit Zurücklegen und unter Beachtung der Reihenfolge gibt es n^k Möglichkeiten.

- (4) **Kombinationen:** Auswählen von Objekten ohne Beachtung der Reihenfolge.

- (a) **Ohne Zurücklegen:** Die Zahl der Möglichkeiten aus n verschiedenen Objekten k ($\leq n$) Objekte ohne Zurücklegen und ohne Beachtung der Reihenfolge auszuwählen, beträgt:

$$\frac{n!}{(n-k)! k!} = \binom{n}{k} = \binom{n}{n-k}$$

Bsp: Wieviele Möglichkeiten gibt es, aus den Zahlen von 1 bis 45 sechs Zahlen ohne Zurücklegen und ohne Beachtung der Reihenfolge auszuwählen?

$$\binom{45}{6} = 8145060$$

- (b) Mit Zurücklegen: Die Zahl der Möglichkeiten aus n verschiedenen Objekten k Objekte mit Zurücklegen und ohne Beachtung der Reihenfolge auszuwählen, beträgt:

$$\binom{n+k-1}{k}$$

Bsp: Ein gefüllter Getränkeautomat bietet 15 verschiedene Softdrinks an. Wenn Sie drei Flaschen entnehmen möchten, wobei die Marke egal ist, wieviele Möglichkeiten haben Sie?

$$\binom{15+3-1}{3} = \binom{17}{3} = 680$$

3 Stochastische Größen und Verteilungen

Die Modellierung eines Zufallsexperiments mittels einer vollständigen Beschreibung des Merkmalraums Ω und einer W-Verteilung P ist nicht immer notwendig oder auch zweckmäßig. Vielfach interessieren nur Teilespekte in Form numerischer Werte, die den einzelnen Versuchsausgängen $\omega \in \Omega$ zugeordnet werden können. Mathematisch betrachtet handelt es sich dabei um eine Abbildung von Ω nach \mathbb{R} (oder \mathbb{R}^k). Eine Abbildung dieser Art nennt man eine **stochastische Größe** (*Stochastik*, von altgriech. $\sigma\tau\circ\chi\circ\varsigma$ = Kunst des Mutmaßens) oder **Zufallsvariable**.¹ Im Folgenden wird die erste Bezeichnung verwendet.

3.1 Stochastische Größen

In diesem Kapitel beschäftigen wir uns zunächst mit eindimensionalen stochastischen Größen, d.h. mit Abbildungen von Ω nach \mathbb{R} . (Bem: Mehrdimensionale stochastische Größen werden in Kapitel 5 behandelt.) Gegeben sei ein W-Raum (Ω, \mathcal{A}, P) .

Stochastische Größe: Eine Abbildung X von Ω nach \mathbb{R} , die jedem $\omega \in \Omega$ eine reelle Zahl $X(\omega) = x$ zuordnet, nennt man eine **stochastische Größe** (kurz **sG**). Die Zahl $x \in \mathbb{R}$ wird als **Realisation** der sG X bezeichnet. Der **Merkmalraum** (oder **Wertebereich**) von X werde mit M_X (oder kurz M , wenn klar ist, um welche sG es sich handelt) bezeichnet:

$$M_X = \{x \mid x = X(\omega), \omega \in \Omega\}$$

In diesem Text ist $M_X \subseteq \mathbb{R}$ eine abzählbare Menge oder ein Intervall. Im ersten Fall spricht man von einer **diskreten**, im zweiten Fall von einer **stetigen** (oder **kontinuierlichen**) sG. (Bem: In der Praxis spielen allerdings auch Mischtypen eine Rolle; vgl. 3.2.3.)

Bem: Stochastische Größen werden meist mit Großbuchstaben vom Ende des Alphabets bezeichnet: X, Y, Z etc. Man beachte auch genau den Unterschied zwischen X und x . Ersteres bezeichnet die sG (d.h. die Abbildung) und Letzteres eine Realisation der sG (d.h. einen konkreten Funktionswert).

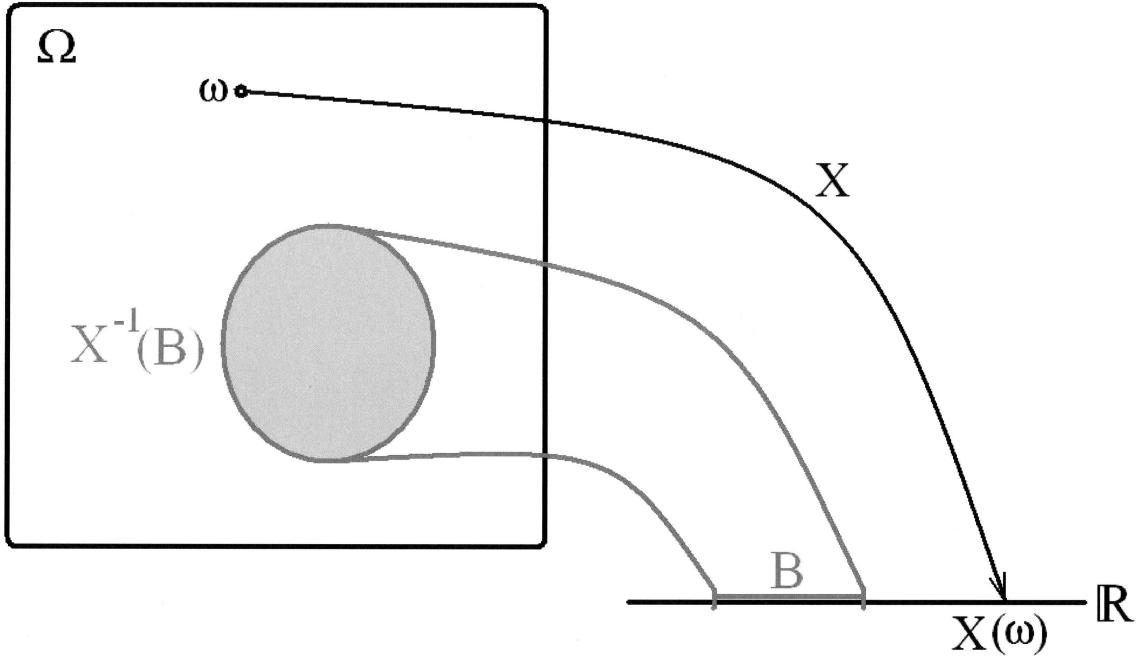
Messbarkeit: Eine Abbildung X von Ω nach \mathbb{R} ist **messbar**, wenn das Urbild jeder Borelmenge $B \in \mathcal{B}$ ein Element von \mathcal{A} ist:

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{A} \quad \text{für alle } B \in \mathcal{B}$$

Im Folgenden werde stets angenommen, dass eine sG X auch messbar² ist. (Bem: Auf Grund von Festlegung 1 (vgl. 2.3) ist diese Eigenschaft trivialerweise erfüllt, wenn Ω

¹engl. *random variable* (abgekürzt rv)

²Die Messbarkeit gehört genaugenommen zu den definierenden Eigenschaften einer sG.

Abbildung 3.1: Symbolische Darstellung einer stochastischen Größe

(höchstens) abzählbar ist.) Vgl. Abb 3.1 für eine symbolische Darstellung einer sG. Um die Messbarkeit von X auch in der Bezeichnung zum Ausdruck zu bringen, schreibt man:

$$X : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}, \mathcal{B})$$

Verteilung von X : Ist X eine sG und A ein Ereignis in \mathbb{R} , so ist auf Grund der Messbarkeit von X das Urbild $X^{-1}(A) = \{\omega \in \Omega \mid X(\omega) \in A\}$ ein Ereignis in Ω . Für letzteres Ereignis schreibt man kurz $X \in A$ und definiert:

$$P_X(A) := P(X \in A) = P(X^{-1}(A))$$

P_X nennt man die (durch P induzierte) **Verteilung** von X .

Bsp 3.1 Als einfache Illustration der obigen Konzepte betrachten wir das Werfen von zwei (symmetrischen) Würfeln. Werden die Würfel hintereinander geworfen, lautet ein passender Merkmalraum $\Omega = \{(i, j) \mid i, j = 1, 2, \dots, 6\}$. Die σ -Algebra ist die Potenzmenge $\mathcal{P}(\Omega)$ und P ist gegeben durch $P(\{(i, j)\}) = 1/36$. Interessiert man sich beispielsweise nur für die geworfene Augensumme, kann man die folgende sG betrachten:

$$X : (i, j) \longmapsto i + j$$

X ist trivialerweise messbar und das Ereignis $X = x$ ist gegeben durch:

$$\{X = x\} = \{(i, j) \mid i + j = x\} = X^{-1}(\{x\})$$

Beispielsweise besteht das Ereignis $X = 8$ aus den folgenden 5 Paaren im ursprünglichen Merkmalraum Ω : $(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)$. Jedes dieser Paare hat die Wahrscheinlichkeit $1/36$ und daher gilt $P(X = 8) = 5/36$. Allgemeiner ist die Verteilung von X gegeben durch:

$$P_X(\{x\}) = P(X = x) = \frac{6 - |7 - x|}{36}, \quad x = 2, 3, \dots, 12$$

■

Intervallereignisse: Neben punktförmigen Ereignissen $\{x\}$, $x \in \mathbb{R}$, spielen in Anwendungen vor allem Ereignisse der Form $A = (a, b]$ (mit $a < b$) eine große Rolle:

$$P_X((a, b]) = P(X \in (a, b]) = P(a < X \leq b)$$

Wegen $(-\infty, b] = (-\infty, a] \cup (a, b]$ (disjunkt) gilt:

$$P(X \leq b) = P(X \leq a) + P(a < X \leq b)$$

Somit:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$

Wahrscheinlichkeiten für Intervalle können also einfach aus Wahrscheinlichkeiten der Form $P(X \leq x)$, $x \in \mathbb{R}$, berechnet werden. Für punktförmige Ereignisse gilt:

$$P_X(\{x\}) = P(X = x) = P(X \leq x) - P(X < x)$$

Letzteres folgt aus $\{x\} = (-\infty, x] \cap \overline{(-\infty, x)}$ und daraus, dass für beliebige Ereignisse A und B gilt:

$$P(A\overline{B}) = P(A) - P(AB)$$

3.2 Verteilungsfunktion

Die im vorhergehenden Abschnitt zuletzt angestellten Überlegungen motivieren die folgende Definition.

Verteilungsfunktion: Die **Verteilungsfunktion**³ (abgekürzt **VF**) F_X einer sG X ist definiert durch:

$$F_X(x) := P(X \leq x), \quad x \in \mathbb{R}$$

Eine Verteilungsfunktion F hat die folgenden Eigenschaften:

- (1) $0 \leq F(x) \leq 1$ für $x \in \mathbb{R}$
- (2) F ist monoton wachsend, d. h., aus $x < y$ folgt $F(x) \leq F(y)$
- (3) $\lim_{x \rightarrow -\infty} F(x) = 0$ und $\lim_{x \rightarrow \infty} F(x) = 1$
- (4) F ist **rechtsstetig**, d. h., $\lim_{h \downarrow 0} F(x+h) = F(x)$ für $x \in \mathbb{R}$

Allgemein nennt man eine Funktion mit den Eigenschaften (1) bis (4) – ohne direkte Bezugnahme auf eine sG – eine **Verteilungsfunktion** (auf \mathbb{R}).

Beweis: Eigenschaft (1) ergibt sich unmittelbar daraus, dass $F(x)$ nach Definition eine Wahrscheinlichkeit ist; Eigenschaft (2) folgt aus Behauptung 5 von Abschnitt 2.5. Zum Beweis der restlichen Eigenschaften benötigt man eine Monotonieeigenschaft des W-Maßes (o. B.):

Lemma: Für eine *wachsende* Folge $\{C_n\}$ von Ereignissen (d. h., wenn $C_n \subseteq C_{n+1}$ für alle n) gilt:

$$\lim_{n \rightarrow \infty} P(C_n) = P\left(\lim_{n \rightarrow \infty} C_n\right) = P\left(\bigcup_{n=1}^{\infty} C_n\right)$$

Für eine *fallende* Folge $\{C_n\}$ von Ereignissen (d. h., wenn $C_n \supseteq C_{n+1}$ für alle n) gilt:

$$\lim_{n \rightarrow \infty} P(C_n) = P\left(\lim_{n \rightarrow \infty} C_n\right) = P\left(\bigcap_{n=1}^{\infty} C_n\right)$$

Sei nun $\{x_n\}$ eine fallende Folge von reellen Zahlen, sodass $x_n \downarrow x$, und sei $C_n = \{X \leq x_n\}$, dann ist $\{C_n\}$ eine monoton fallende Mengenfolge mit $\bigcap_{n=1}^{\infty} C_n = \{X \leq x\}$. Mit dem obigen Lemma folgt:

$$\lim_{n \rightarrow \infty} F(x_n) = P\left(\bigcap_{n=1}^{\infty} C_n\right) = F(x)$$

Das zeigt Eigenschaft (4). (Beweis von Eigenschaft (3) als UE-Aufgabe.)

³engl. *cumulative distribution function* (abgekürzt cdf)

Behauptung 1: Sei X eine sG mit Verteilungsfunktion F_X . Dann gilt für $a < b$:

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

Beweis: Folgt aus der disjunktten Darstellung:

$$\{-\infty < X \leq b\} = \{-\infty < X \leq a\} \cup \{a < X \leq b\}$$

Behauptung 2: Mit $F(x-) := \lim_{h \downarrow 0} F(x-h)$ (= linksseitiger Grenzwert) gilt:

$$P(X = x) = F(x) - F(x-) \quad \text{für } x \in \mathbb{R}$$

Beweis: Die punktförmige Menge $\{x\}$ lässt sich wie folgt darstellen:

$$\{x\} = \bigcap_{n=1}^{\infty} \underbrace{\left(x - \frac{1}{n}, x \right]}_{=: C_n} = \bigcap_{n=1}^{\infty} C_n$$

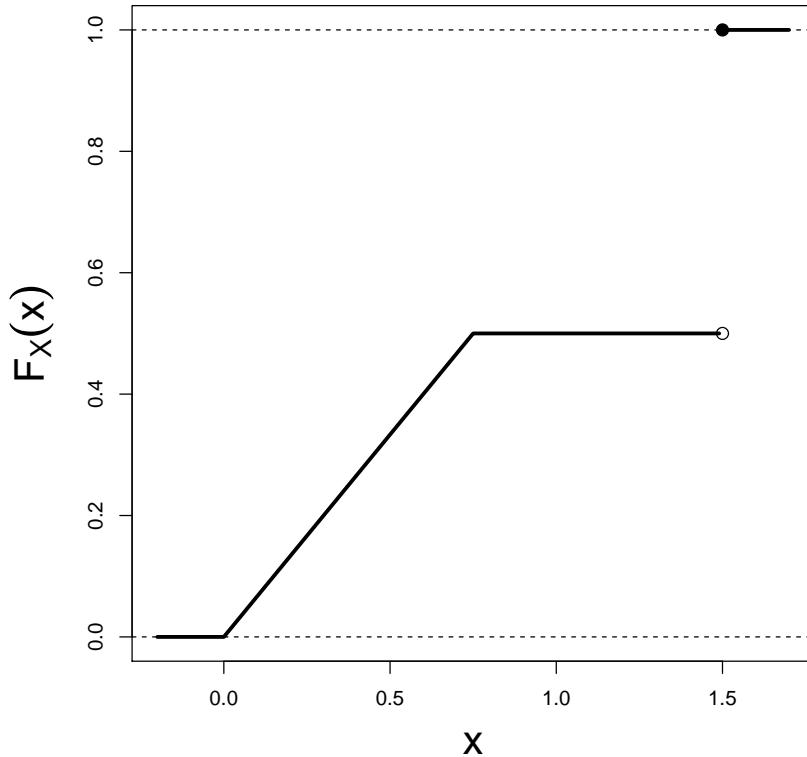
D.h., $\{x\}$ ist der Limes einer fallenden Mengenfolge. Mit dem obigen Lemma (und Behauptung 1) folgt:

$$\begin{aligned} P(X = x) &= P\left(\bigcap_{i=1}^{\infty} \left\{ x - \frac{1}{n} < X \leq x \right\}\right) \\ &= \lim_{n \rightarrow \infty} P\left(x - \frac{1}{n} < X \leq x\right) \\ &= \lim_{n \rightarrow \infty} \left[F_X(x) - F_X\left(x - \frac{1}{n}\right) \right] \\ &= F_X(x) - F_X(x-) \end{aligned}$$

Bsp 3.2 Die sG X habe die folgende Verteilungsfunktion (Abb 3.2):

$$F_X(x) = \begin{cases} 0 & \text{für } x < 0 \\ 2x/3 & \text{für } 0 \leq x < 3/4 \\ 1/2 & \text{für } 3/4 \leq x < 3/2 \\ 1 & \text{für } x \geq 3/2 \end{cases}$$

Die Funktion ist zwar nicht stetig – F_X hat einen Sprung an der Stelle $x = 3/2$ – erfüllt aber alle Eigenschaften einer Verteilungsfunktion. (Die Rechtsstetigkeit wird durch den offen bzw. dick gezeichneten Punkt an der Stelle $x = 3/2$ hervorgehoben.)

Abbildung 3.2: Verteilungsfunktion (Bsp 3.2)

Beispielsweise gilt:

$$P\left(\frac{1}{4} < X \leq 1\right) = F_X(1) - F_X\left(\frac{1}{4}\right) = \frac{1}{2} - \frac{1}{6} = \frac{1}{3}$$

Nach Behauptung 2 gilt an der Stelle $x = 3/2$:

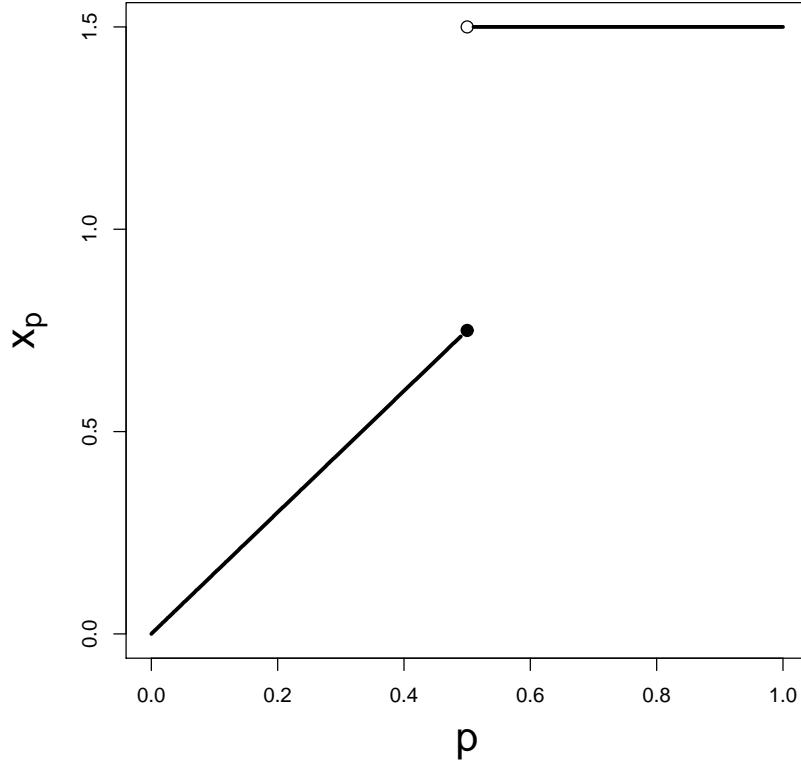
$$P\left(X = \frac{3}{2}\right) = F_X\left(\frac{3}{2}\right) - F_X\left(\frac{3}{2}^-\right) = 1 - \frac{1}{2} = \frac{1}{2}$$

Dieser Wert entspricht der Höhe des Sprungs bei $x = 3/2$. An allen anderen Stellen ist F_X stetig (d. h. rechts- und linksstetig), daher gilt $P(X = x) = 0$ für $x \neq 3/2$. ■

Vielfach benötigt man die Umkehrfunktion einer VF F . Da aber eine VF nicht notwendigerweise *streng* monoton wächst, muss man die Definition der Inversen von F modifizieren.

Verallgemeinerte Inverse: Die **verallgemeinerte Inverse** $F^{-1} : [0, 1] \longrightarrow \mathbb{R}$ einer VF F ist definiert durch:

$$F^{-1}(y) = \inf \{x \mid F(x) \geq y\} \quad \text{für } y \in (0, 1)$$

Abbildung 3.3: Quantilenfunktion (Bsp 3.3)

Bem: Ist F streng monoton wachsend, entspricht F^{-1} der Umkehrfunktion von F .

Aus der deskriptiven Statistik kennen wir das Konzept des Stichprobenquantils (vgl. 1.7.7). Das theoretische Pendant ist die **Quantilenfunktion**:

$$x_p = F^{-1}(p) \quad \text{für } p \in (0, 1)$$

Dabei ist F^{-1} die (verallgemeinerte) Inverse von F . Allgemein nennt man für ein festes $p \in (0, 1)$ den Wert x_p das p -**Quantil** von F (oder von X , wenn $F = F_X$).

Bsp 3.3 Die Quantilenfunktion zur VF von Bsp 3.2 ist gegeben durch (Abb 3.3):

$$x_p = \begin{cases} 3p/2 & \text{für } 0 \leq p \leq 1/2 \\ 3/2 & \text{für } 1/2 < p \leq 1 \end{cases}$$

Die Quantilenfunktion geht aus F_X durch Spiegelung an der 1. Mediane hervor. Man beachte, dass hier auch die Quantile für $p = 0$ und $p = 1$ definiert sind. Bei vielen praktisch wichtigen Verteilungen liegt zumindest eines dieser Quantile im Unendlichen. ■

3.2.1 Diskrete Verteilungen

Eine stochastische Größe X hat eine **diskrete Verteilung** (oder ist **diskret**), wenn ihr Merkmalraum $M_X = \{x_1, x_2, \dots\}$ aus einer endlichen oder abzählbaren Menge von Punkten besteht. Man schreibt in diesem Fall:

$$p_X(x) = P(X = x) \quad \text{für } x \in M_X$$

und nennt $p_X(x)$ die **Wahrscheinlichkeitsfunktion**⁴ (oder die **Punktwahrscheinlichkeiten**, auch die **Zähldichte**) von X .

Eine Wahrscheinlichkeitsfunktion hat die folgenden Eigenschaften:

$$(1) \quad 0 \leq p_X(x) \leq 1, \quad x \in M_X \quad \text{und} \quad (2) \quad \sum_{x \in M_X} p_X(x) = 1$$

Die Wahrscheinlichkeit für eine Teilmenge B von M_X wird wie folgt berechnet:

$$P(X \in B) = \sum_{x \in B} p_X(x)$$

Die Verteilungsfunktion einer diskreten sG ist eine **Treppenfunktion** mit Sprüngen der Höhe $p_X(x)$ an den Stellen $x \in M_X$:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} p_X(x_i), \quad x \in \mathbb{R}$$

Bsp 3.4 Die sG von Bsp 3.1 ist diskret mit Wahrscheinlichkeitsfunktion:

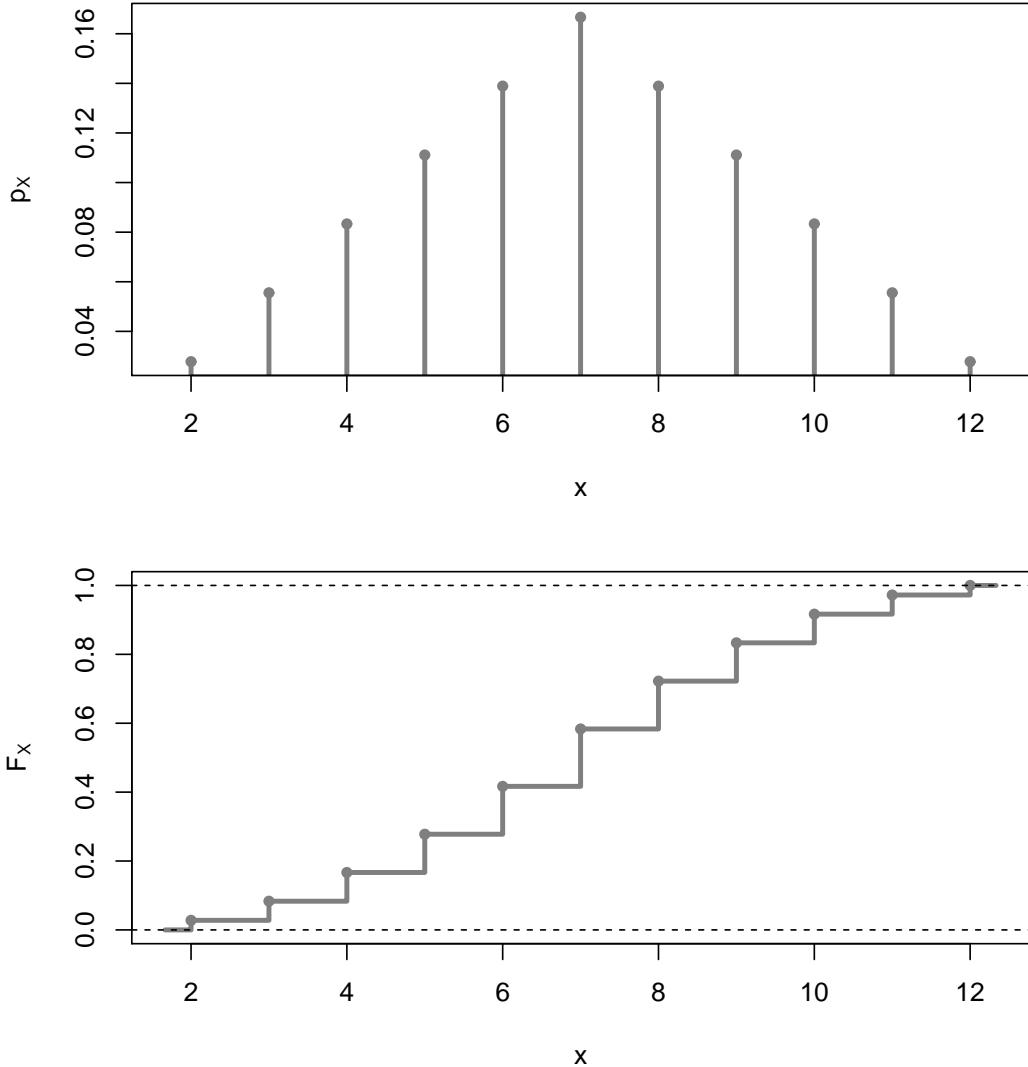
$$p_X(x) = \frac{6 - |7 - x|}{36}, \quad x = 2, 3, \dots, 12$$

Abb 3.4 ist eine graphische Darstellung von p_X und F_X . Man beachte, dass auf Grund der Rechtsstetigkeit von F_X bei Sprüngen jeweils der obere Punkt gültig ist. Letzteres ist insbesondere dann zu beachten, wenn man – so wie hier – die Treppen auszeichnet. ■

3.2.2 Stetige Verteilungen

Eine stochastische Größe X hat eine **stetige Verteilung** (oder ist **stetig**), wenn die Verteilungsfunktion $F_X(x)$ eine stetige Funktion auf \mathbb{R} ist.

⁴engl. *probability mass function* (abgekürzt pmf)

Abbildung 3.4: Wahrscheinlichkeits- und Verteilungsfunktion (Bsp 3.4)

Nach Behauptung 2 gilt allgemein, dass $P(X = x) = F_X(x) - F_X(x-)$. Für eine stetige sG X gibt es also keine Punkte mit positiver Wahrscheinlichkeit, d. h., $P(X = x) = 0$ für alle $x \in \mathbb{R}$. Die meisten stetigen sGn sind **absolut** stetig, d. h., es gibt eine Funktion f_X , sodass:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t) dt$$

Eine Funktion f_X mit dieser Eigenschaft nennt man eine **Dichtefunktion**⁵ (oder kurz **Dichte**) von X . Ist f_X selbst stetig, dann folgt aus dem *Hauptsatz der Differential- und Integralrechnung*, dass:

⁵engl. *probability density function* (abgekürzt pdf)

$$\frac{d}{dx} F_X(x) = F'_X(x) = f_X(x)$$

Eine Dichtefunktion hat die folgenden Eigenschaften:

$$(1) \quad f_X(x) \geq 0, \quad x \in \mathbb{R} \quad \text{und} \quad (2) \quad \int_{-\infty}^{\infty} f_X(t) dt = 1$$

Die Menge S_X aller Punkte $x \in \mathbb{R}$ mit $f_X(x) > 0$ nennt man den **Träger**⁶ von X . Die Wahrscheinlichkeit für eine (Borel-) Menge $B \in \mathcal{B}$ lässt sich wie folgt berechnen:

$$P(X \in B) = \int_B f_X(t) dt$$

Für ein Intervall $B = (a, b]$ ($a < b$) gilt:

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(t) dt$$

Bem: Man beachte, dass für eine stetige sG X gilt:

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b)$$

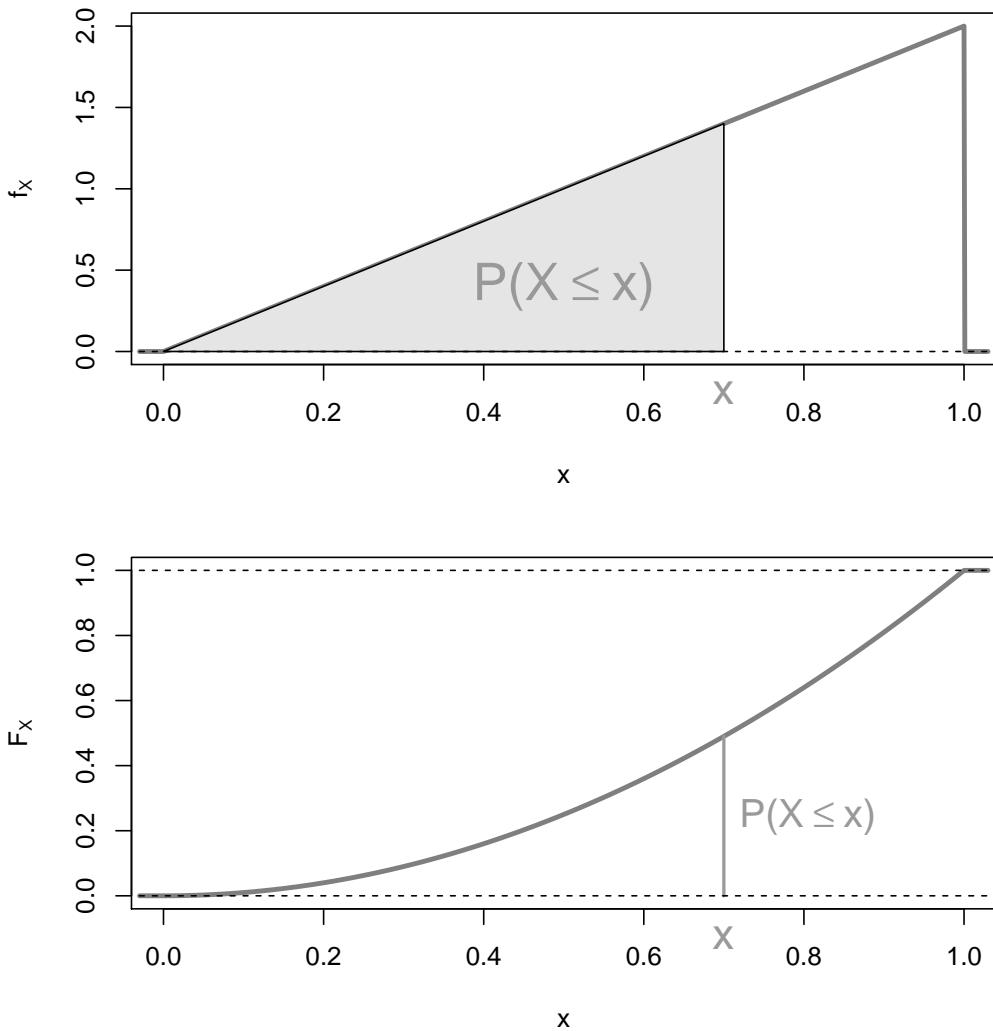
Ist F_X streng monoton wachsend, so ist das p -Quantil von X jener (eindeutig bestimmte) Wert x_p , sodass:

$$F_X(x_p) = P(X \leq x_p) = \int_{-\infty}^{x_p} f_X(t) dt = p \iff x_p = F_X^{-1}(p)$$

Bsp 3.5 Angenommen, wir wählen ganz zufällig einen Punkt im Inneren eines Kreises mit Radius 1. Der Merkmalraum für dieses Experiment ist $\Omega = \{(u, v) \mid u^2 + v^2 < 1\}$. Sei X der Abstand des Punktes vom Ursprung. Da der Punkt zufällig gewählt wird, gilt auf Basis einer einfachen geometrischen Überlegung:

$$P(X \leq x) = \frac{\text{Fläche des Kreises mit Radius } x}{\text{Fläche des Kreises mit Radius } 1} = \frac{x^2 \pi}{\pi} = x^2, \quad 0 \leq x < 1$$

⁶engl. *support*

Abbildung 3.5: Dichte- und Verteilungsfunktion (Bsp 3.5)

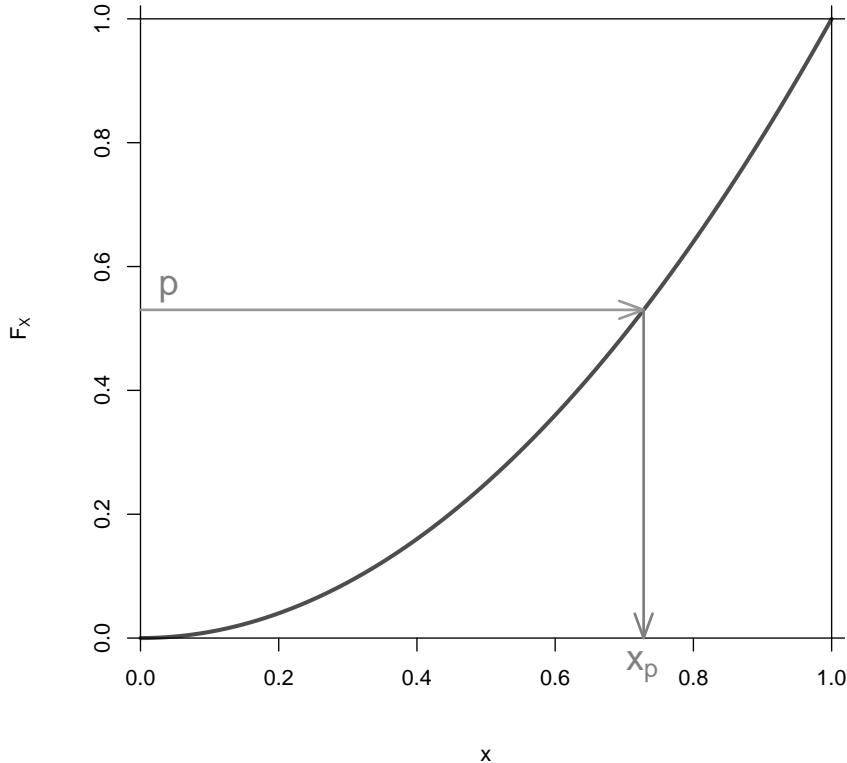
Die Verteilungsfunktion von X ist also gegeben durch:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Dabei handelt es sich um eine stetige Funktion mit Ableitung (= Dichte):

$$f_X(x) = F'_X(x) = \begin{cases} 2x & 0 \leq x < 1 \\ 0 & \text{sonst} \end{cases}$$

In Abb 3.5 sind f_X und F_X grafisch dargestellt, außerdem wird die Beziehung zwischen den beiden Funktionen verdeutlicht.

Abbildung 3.6: Quantilenbestimmung (Bsp 3.5)

Die Wahrscheinlichkeit, dass der Punkt beispielsweise in einen Ring mit den Radien $1/4$ und $1/2$ fällt, lässt sich wie folgt berechnen:

$$P\left(\frac{1}{4} < X \leq \frac{1}{2}\right) = \int_{1/4}^{1/2} 2t \, dt = t^2 \Big|_{t=1/4}^{t=1/2} = \frac{3}{16}$$

Das p -Quantil x_p ist gegeben durch (Abb 3.6):

$$x_p = F_X^{-1}(p) = \sqrt{p}, \quad 0 \leq p \leq 1$$

■

3.2.3 Gemischte Verteilungen

In einigen praktisch wichtigen Situationen ist die Verteilung weder (rein) diskret noch (rein) stetig, die Verteilungsfunktion F also keine (reine) Treppenfunktion aber auch nicht überall stetig. Man denke etwa an ein Produkt (beispielsweise eine Glühlampe), das von Anfang an defekt ist oder unmittelbar bei der ersten Inbetriebnahme ausfällt. Die Lebensdauer dieses Produkts ist also mit positiver Wahrscheinlichkeit gleich Null. Ist das

Produkt aber zu Beginn intakt und/oder überlebt es die erste Inbetriebnahme, ist seine Lebensdauer stetig verteilt. Eine Verteilung (oder sG) dieser Art nennt man **gemischt**, da F – im obigen Beispiel – eine **Mischung** aus einer diskreten (F_d) und einer stetigen (F_s) Verteilungsfunktion ist:

$$F(x) = \alpha F_d(x) + (1 - \alpha) F_s(x), \quad 0 \leq \alpha \leq 1$$

Etwas allgemeiner spricht man von einer **gemischten** Verteilung, wenn sich ihre Verteilungsfunktion F als Mischung von m (≥ 2) Verteilungsfunktionen F_j darstellen lässt:

$$F(x) = \sum_{j=1}^m \alpha_j F_j(x) \quad \text{mit} \quad \alpha_j > 0 \quad \text{und} \quad \sum_{j=1}^m \alpha_j = 1$$

wobei mindestens eine der VFn F_j diskret und mindestens eine stetig ist.

Bem: Generell lässt sich durch Mischen von (endlich oder unendlich vielen) Verteilungen (auch gleicher Art, d. h. alle diskret oder alle stetig) die statistische Modellbildung beträchtlich erweitern. Im vorliegenden Text betrachten wir Mischverteilungen aber nur im obigen enger gefassten Sinn. (Wir werden allerdings dem Mischen von Verteilungen im allgemeineren Sinn wieder in der *Bayes'schen Statistik* (Kapitel 8) begegnen.)

Der Merkmalraum einer gemischten Verteilung hat die Form:

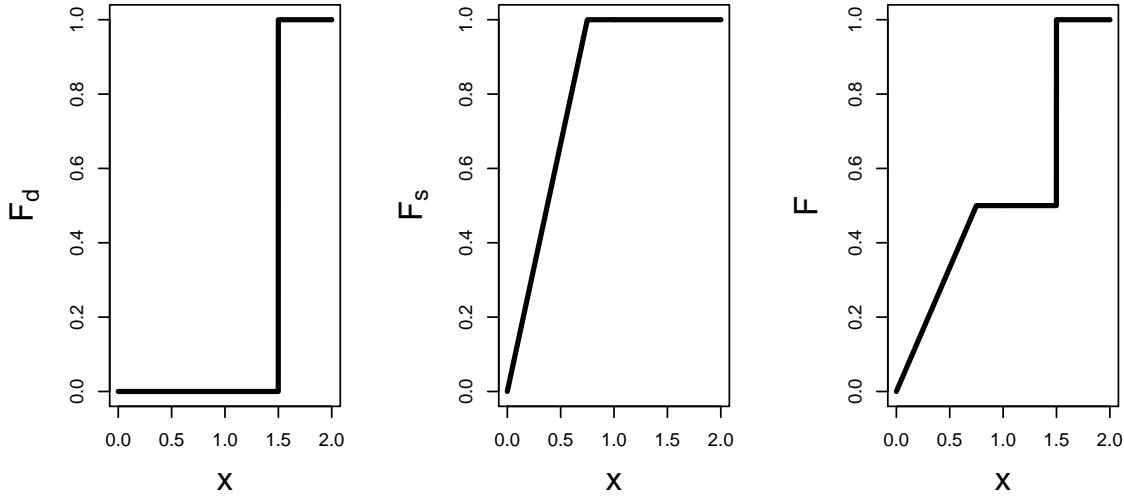
$$M = \{x_1, x_2, \dots, x_m\} \cup \langle a, b \rangle$$

wobei $\langle a, b \rangle$ ein endliches oder unendliches Intervall ist (und die Randpunkte je nach Anwendung dazu gehören oder nicht). Die diskreten Punkte x_i haben positive Wahrscheinlichkeiten $p(x_i) > 0$ und es gibt eine Dichte f^* mit Träger $\langle a, b \rangle$, sodass:

$$\sum_{i=1}^m p(x_i) + \int_a^b f^*(x) dx = 1$$

Man beachte, dass f^* hier keine *vollständige* Dichte ist, da $\int_a^b f^*(x) dx < 1$ ist. (Bem: Aus diesem Grund verwenden wir das $*$ -Symbol.) Die Wahrscheinlichkeit einer (Borel-) Menge $B \in \mathcal{B}$ lässt sich wie folgt berechnen:

$$P(X \in B) = \sum_{x_i \in B} p(x_i) + \int_B f^*(x) dx$$

Abbildung 3.7: F als Mischung von F_d und F_s (Bsp 3.6)

Bsp 3.6 Die VF von Bsp 3.2 ist keine Treppe aber auch nicht überall stetig, d. h., es handelt sich um eine gemischte Verteilung mit Merkmalraum $M = \{3/2\} \cup \langle 0, 3/4 \rangle$. (Bem: Der Bereich zwischen $3/4$ und $3/2$ hat die Wahrscheinlichkeit 0 und gehört nicht zum Träger.)

Betrachten wir den stetigen und den diskreten Teil etwas genauer. Die Dichte f^* bekommt man durch Ableiten von F_X (an den Stellen, an denen F_X differenzierbar ist):

$$f^*(x) = \frac{2}{3} I_{(0,3/4)}(x) \quad \text{mit} \quad \int_0^{3/4} f^*(x) dx = \frac{1}{2} < 1$$

f^* ist konstant auf $(0, 3/4)$; die vollständige Dichte lautet $f(x) = (4/3) I_{(0,3/4)}(x)$. Die (stetige) VF F_s ist also gegeben durch:

$$F_s(x) = \int_0^x \frac{4}{3} dt = \frac{4x}{3}, \quad 0 \leq x \leq \frac{3}{4}$$

Da es nur einen diskreten Punkt gibt ($x_1 = 3/2$), ist die diskrete VF F_d gegeben durch:

$$F_d(x) = I_{[3/2, \infty)}(x)$$

Die VF F ist hier eine Mischung zu gleichen Teilen von F_d und F_s (vgl. Abb 3.7):

$$F(x) = \left(\frac{1}{2}\right) F_d(x) + \left(\frac{1}{2}\right) F_s(x), \quad x \in \mathbb{R}$$

■

Bsp 3.7 Eine wichtige Anwendung von gemischten Verteilungen ergibt sich bei der Analyse von Lebensdauern, wenn Beobachtungen auf die eine oder andere Weise „zensiert“ (d. h. unvollständig) sind.

Angenommen, bestimmte Komponenten sollen hinsichtlich ihrer Lebensdauer („Zuverlässigkeit“) getestet werden. Wenn es sich um sehr zuverlässige Komponenten handelt, kann die Zeitspanne bis zum Ausfall unrealistisch lang sein (u. U. mehrere Jahre). In der Praxis bricht man daher den Versuch nach einer bestimmten Zeitspanne T ab. Ausfälle, die innerhalb von $[0, T]$ auftreten, können beobachtet werden; Ausfälle, die erst *nach* T auftreten, werden aber nicht beobachtet.

Beispielsweise sei die VF der Lebensdauer (Einheit: h) einer bestimmten Komponente gegeben durch:

$$F(x) = 1 - \exp\left(-\frac{x}{1000}\right), \quad 0 \leq x < \infty \quad (= 0 \text{ sonst})$$

Ein Versuch, bei dem derartige Komponenten getestet werden, werde nach $T = 800$ h abgebrochen. Ausfälle, die erst nach 800 h auftreten, werden nicht beobachtet und sind zensiert. Die VF der (*beobachteten*) Lebensdauer springt also bei $x = 800$ von $F(800) \doteq 0.55$ auf Eins:

$$\tilde{F}(x) = \begin{cases} 0 & x < 0 \\ 1 - \exp\left(-\frac{x}{1000}\right) & 0 \leq x < 800 \\ 1 & x \geq 800 \end{cases}$$

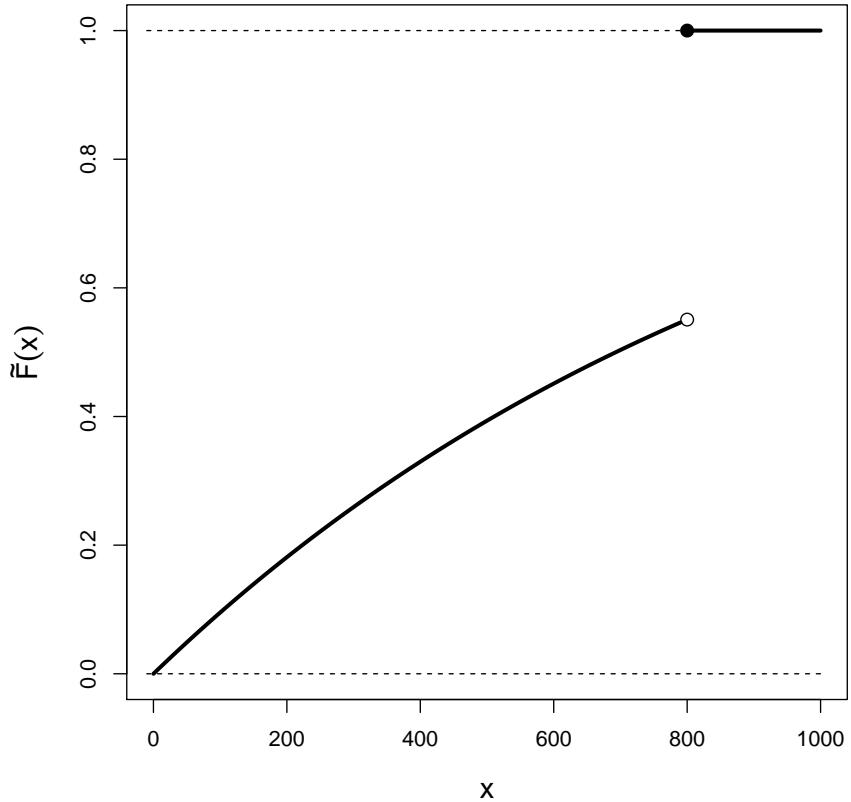
Vgl. Abb 3.8 für eine grafische Darstellung der gemischten Verteilung \tilde{F} . ■

3.3 Transformationen

Häufig ist man mit Problemen der folgenden Art konfrontiert: Man kennt von einer sG X ihre Verteilung (d. h. die Verteilungsfunktion F_X , die Dichte f_X , oder die Wahrscheinlichkeitsfunktion p_X), interessiert sich aber für eine **Transformation** $Y = g(X)$ von X , wobei g eine (messbare⁷) Funktion von \mathbb{R} nach \mathbb{R} ist. Da Y wieder eine sG ist, stellt sich die Frage nach ihrer Verteilung.

Zweckmäßigerweise betrachten wir die Fälle, dass X diskret oder stetig verteilt ist, getrennt voneinander. Weiters unterscheiden wir auch danach, ob g eine umkehrbar eindeutige (d. h. *bijektive*) Funktion ist oder nicht.

⁷Eine Funktion $g : \mathbb{R} \rightarrow \mathbb{R}$ ist *messbar*, wenn das Urbild $g^{-1}(B)$ jeder Borelmenge B wieder eine Borelmenge ist (gilt z. B. für alle stetigen Funktionen).

Abbildung 3.8: Zensierte Beobachtungen (Bsp 3.7)

3.3.1 Transformationen diskreter sGn

Ist g eine umkehrbar eindeutige Funktion, lässt sich die W-Funktion von $Y = g(X)$ einfach wie folgt bestimmen:

$$p_Y(y) = P(Y = y) = P(g(X) = y) = P(X = g^{-1}(y)) = p_X(g^{-1}(y))$$

Bsp 3.8 Eine (symmetrische) Münze wird wiederholt geworfen und X sei die Nummer des Wurfs, bei dem zum ersten Mal „Kopf“ geworfen wird. Der Merkmalraum von X ist $M_X = \{1, 2, \dots\}$ und es gilt:

$$p_X(x) = P(X = x) = \left(\frac{1}{2}\right)^{x-1} \left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^x, \quad x = 1, 2, \dots$$

Y sei nun die Zahl der Würfe *vor* dem ersten Kopf, d. h., $Y = X - 1$. Die Transformation $g(x) = x - 1$ ist umkehrbar eindeutig und $g^{-1}(y) = y + 1$. Der Merkmalraum von Y ist $M_Y = \{0, 1, 2, \dots\}$ und die W-Funktion von Y ist gegeben durch:

$$p_Y(y) = p_X(y+1) = \left(\frac{1}{2}\right)^{y+1}, \quad y = 0, 1, 2, \dots$$

■

Ist die Transformation g nicht umkehrbar eindeutig, lässt sich die Verteilung von $Y = g(X)$ meist durch eine einfache direkte Überlegung bestimmen.

Bsp 3.9 Angenommen, wir spielen das Spiel von Bsp 3.8 gegen die „Bank“. Kommt der erste Kopf bei einem ungeraden Wurf, zahlen wir der Bank 1€, kommt er bei einem geraden Wurf, gewinnen wir 1€. Ist Y unser (Netto-) Gewinn, so gilt $M_Y = \{-1, 1\}$. Die Wahrscheinlichkeit, dass der erste Kopf bei einem ungeraden Wurf kommt, berechnet man wie folgt:

$$P(X \in \{1, 3, 5, \dots\}) = \sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^{2x-1} = \sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^{2x+1} = \frac{1/2}{1 - 1/4} = \frac{2}{3}$$

Die Verteilung von Y ist also gegeben durch:

$$p_Y(-1) = \frac{2}{3}, \quad p_Y(1) = \frac{1}{3}$$

■

3.3.2 Transformationen stetiger sGn

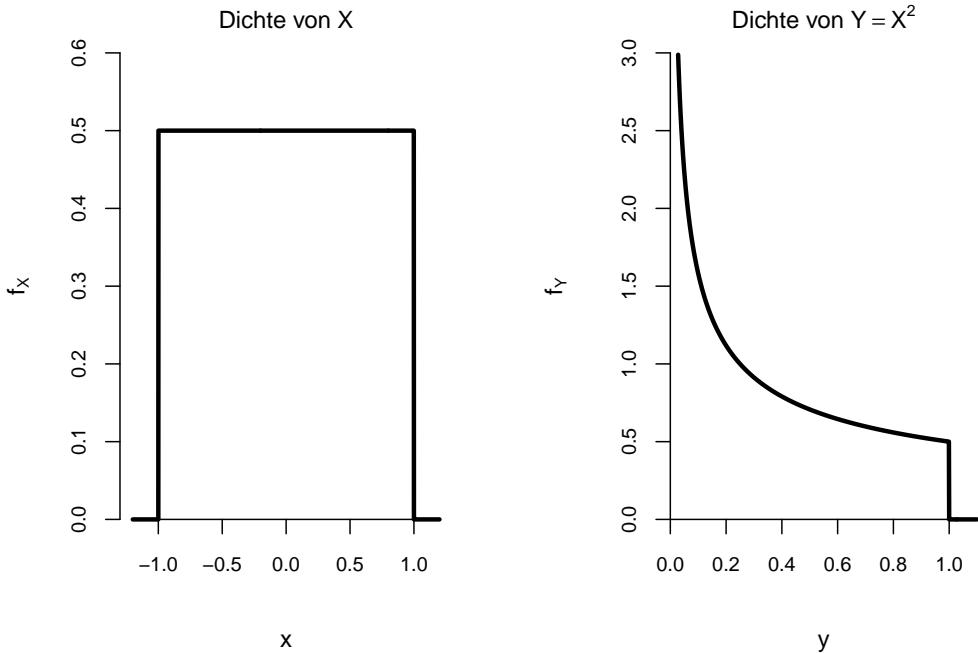
Unabhängig davon, ob g umkehrbar eindeutig ist oder nicht, lässt sich für stetiges X die Verteilung von $Y = g(X)$ mittels der **Methode der Verteilungsfunktion** bestimmen. Die Methode werde anhand eines Beispiels demonstriert.

Bsp 3.10 Die Dichte einer sG X sei gegeben durch:

$$f_X(x) = \begin{cases} \frac{1}{2} & -1 < x < 1 \\ 0 & \text{sonst} \end{cases}$$

Angenommen, wir möchten die Dichte von $Y = X^2$ bestimmen. Dazu bestimmen wir zunächst durch eine direkte Überlegung die Verteilungsfunktion von Y . Für $y \geq 0$ gilt:

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

Abbildung 3.9: Transformation einer stetigen sG (Bsp 3.10)

Die Verteilungsfunktion von Y lässt sich durch die Verteilungsfunktion von X ausdrücken. Letztere ist gegeben durch:

$$F_X(x) = \begin{cases} 0 & x < -1 \\ \frac{x+1}{2} & -1 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Wegen $F_X(\sqrt{y}) - F_X(-\sqrt{y}) = \sqrt{y}$ folgt:

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \sqrt{y} & 0 \leq y < 1 \\ 1 & y \geq 1 \end{cases}$$

Die Dichte von Y bekommt man durch Ableiten:

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} & 0 < y < 1 \\ 0 & \text{sonst} \end{cases}$$

■

In Bsp 3.10 ist die Transformation ($g(x) = x^2$ für $|x| < 1$) nicht umkehrbar eindeutig. Ist aber g umkehrbar eindeutig, kann die Dichte von $Y = g(X)$ mit Hilfe des folgenden Satzes auch direkt bestimmt werden.

Transformationssatz für Dichten: X sei eine stetige sG mit Dichte f_X und Träger S_X , und g sei eine umkehrbar eindeutige differenzierbare Funktion auf S_X . Dann ist die Dichte von $Y = g(X)$ gegeben durch:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|, \quad y \in S_Y$$

Der Träger S_Y von Y ist dabei die Menge $S_Y = \{y = g(x) \mid x \in S_X\}$.

Beweis: Eine umkehrbar eindeutige Funktion ist entweder strikt monoton wachsend oder strikt monoton fallend. Im ersten Fall gilt für die VF von Y :

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

Die Dichte von Y bekommt man durch Ableiten:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(g^{-1}(y)) \frac{dx}{dy}$$

Dabei ist dx/dy die Ableitung der Umkehrfunktion $x = g^{-1}(y)$. Ist g strikt monoton wachsend, gilt $dx/dy > 0$ und $dx/dy = |dx/dy|$. Analog argumentiert man, wenn g strikt monoton fallend ist:

$$F_Y(y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

Somit:

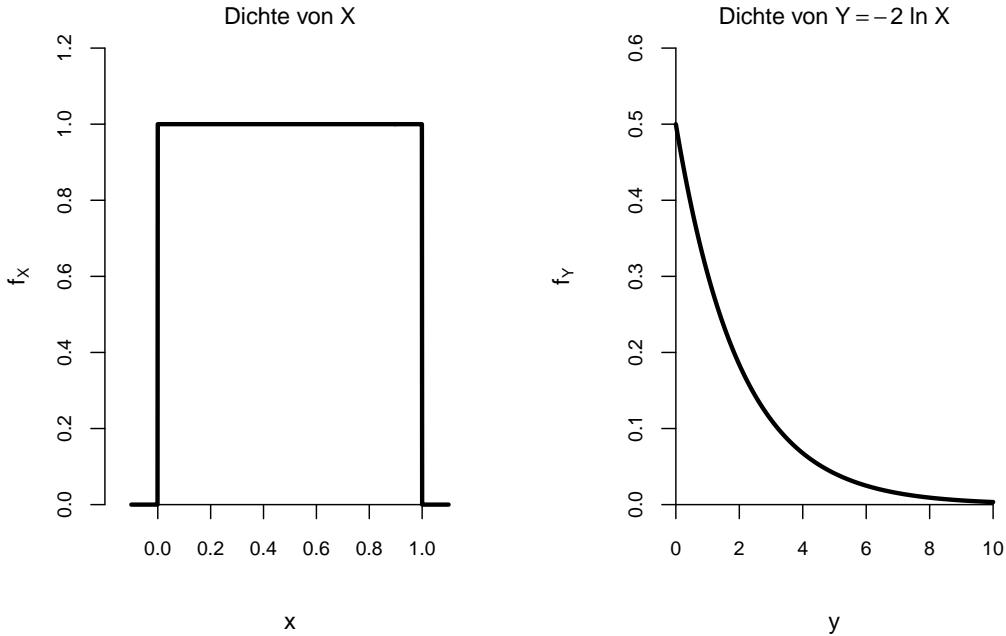
$$f_Y(y) = \frac{dF_Y(y)}{dy} = -f_X(g^{-1}(y)) \frac{dx}{dy}$$

Da in diesem Fall g strikt monoton fallend ist, gilt $dx/dy < 0$ und daher $-dx/dy = |dx/dy|$. Die Behauptung des Satzes ist also für beide Fälle gezeigt.

Jacobian: Die Ableitung $dx/dy = dg^{-1}(y)/dy$ der *Umkehrabbildung* nennt man in englischsprachigen Texten häufig die **Jacobian**⁸ und schreibt:

$$J = \frac{dx}{dy} = \frac{dg^{-1}(y)}{dy} = \frac{1}{\frac{dg(x)}{dx}}$$

⁸Nach CARL GUSTAV JACOB JACOBI (eigentl. JACQUES SIMON; 1804–1851), dt. Mathematiker (bedeutende Beiträge zu mehreren Gebieten der Mathematik und Physik).

Abbildung 3.10: Transformation einer stetigen sG (Bsp 3.11)

Bsp 3.11 Die Dichte der sG X sei gegeben durch:

$$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{sonst} \end{cases}$$

Wie lautet die Dichte von $Y = -2 \ln X$? Die Träger von X und Y sind gegeben durch $S_X = (0, 1)$ bzw. $S_Y = (0, \infty)$. Die Transformation $y = -2 \ln x$ ist umkehrbar eindeutig zwischen S_X und S_Y . Die Umkehrabbildung lautet $x = g^{-1}(y) = e^{-y/2}$ und die Jacobian ist gegeben durch:

$$J = \frac{dx}{dy} = \frac{d(e^{-y/2})}{dy} = -\frac{1}{2}e^{-y/2}$$

Nach dem Transformationssatz lautet die Dichte von Y wie folgt:

$$f_Y(y) = \begin{cases} f_X(e^{-y/2})|J| = \frac{1}{2}e^{-y/2} & 0 < y < \infty \\ 0 & \text{sonst} \end{cases}$$

Vgl. Abb 3.10 für eine grafische Veranschaulichung. ■

Wichtige Spezialfälle sind **affine**⁹ Transformationen der Form $Y = a + bX$.

Dichte einer affinen Transformation: X sei eine stetige sG mit Dichte f_X und $Y = a + bX$, wobei $b \neq 0$. Dann ist die Dichte von Y gegeben durch:

$$f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right)$$

Beweis: Für $b > 0$ (analog für $b < 0$) ist die VF von Y gegeben durch:

$$F_Y(y) = P(Y \leq y) = P\left(X \leq \frac{y-a}{b}\right) = F_X\left(\frac{y-a}{b}\right)$$

Die Dichte ergibt sich durch Ableiten:

$$f_Y(y) = F'_Y(y) = \frac{d}{dy} F_X\left(\frac{y-a}{b}\right) = \frac{1}{b} f_X\left(\frac{y-a}{b}\right)$$

Man kann auch den Transformationssatz verwenden: Die Jacobian der Transformation ist $J = 1/b$ und die Dichte von Y ist gegeben durch:

$$f_Y(y) = f_X(g^{-1}(y)) |J| = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right)$$

Bsp 3.12 Die Dichte der sG X sei gegeben durch:

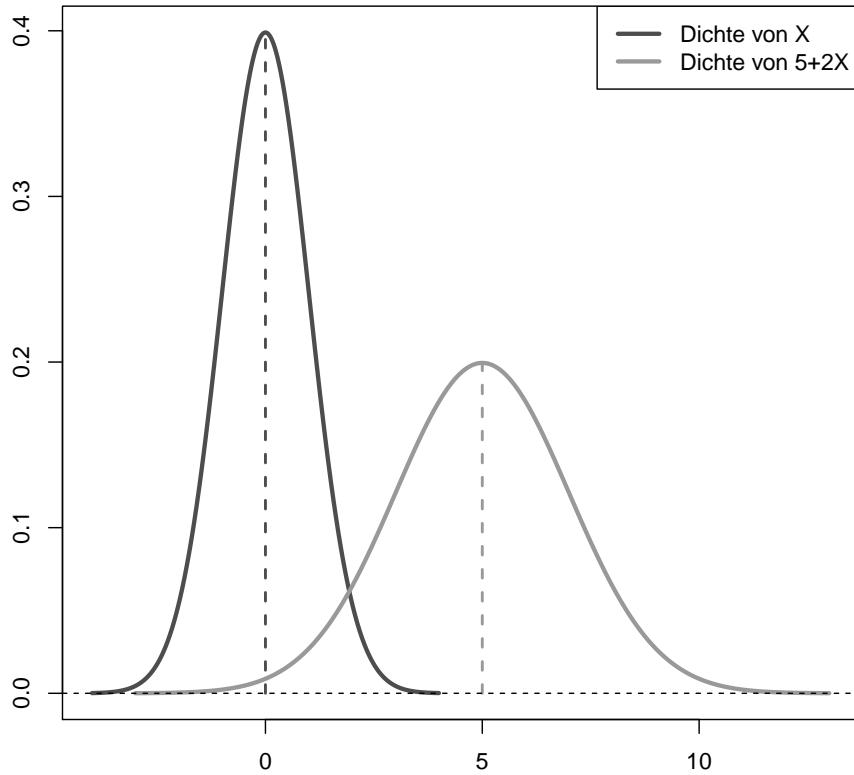
$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty$$

(Bem: Dabei handelt es sich um die (Standard-) Normaldichte („Glockenkurve“), die später noch ausführlicher behandelt wird.) Welche Dichte hat $Y = 5 + 2X$? Nach dem obigen Satz gilt:

$$f_Y(y) = \frac{1}{2\sqrt{2\pi}} \exp\left[-\frac{(y-5)^2}{8}\right], \quad -\infty < y < \infty$$

Abb 3.11 zeigt die ursprüngliche und die transformierte Dichte. Man beachte, dass durch eine affine Transformation wohl die Lage und/oder die Skalierung aber nicht die Form der Dichte geändert wird. ■

⁹Manchmal auch (unkorrekt) als lineare Transformationen bezeichnet. (Lineare Transformationen im strikten Sinn haben die Form $Y = bX$.)

Abbildung 3.11: Affine Transformation (Bsp 3.12)

3.4 Erwartungswert

Die Verteilungsfunktion (W–Funktion, Dichte) enthält die gesamte verfügbare (Wahrscheinlichkeits–) Information über eine sG X . In vielen Situationen genügen allerdings einige wenige charakteristische (numerische) Werte. Einer dieser Werte ist der **Erwartungswert** (auch **Mittelwert** oder kurz **Mittel**) von X , der ein (gewichteter) Durchschnittswert der möglichen Ausprägungen von X ist.

Bem: Aus rein mathematischer Perspektive wäre es im Folgenden nicht notwendig, den diskreten, den stetigen und den gemischten Fall getrennt zu behandeln. Aus praktischer Sicht ist diese Vorgangsweise aber durchaus sinnvoll, wobei Ähnlichkeiten zwischen den einzelnen Fällen offensichtlich sind.

Erwartungswert einer diskreten sG: Ist X eine diskrete sG mit W–Funktion $p(x)$ und gilt $\sum_x |x| p(x) < \infty$, so ist der **Erwartungswert** von X definiert durch:

$$\mathbb{E}(X) = \sum_x x p(x)$$

Der Erwartungswert von X ist also ein gewichteter Mittelwert der möglichen Ausprägungen von X , wobei die Gewichte den Wahrscheinlichkeiten der einzelnen Ausprägungen entsprechen.

Bsp 3.13 Angenommen, bei einem Spiel mit zwei Würfeln ist der Gewinn gleich der größeren der beiden Augenzahlen. Um an diesem Spiel teilzunehmen, ist aber ein Einsatz von d Euro zu entrichten. Wie groß sollte d sein? Handelt es sich um ein *faires* Spiel, sollte der Einsatz dem zu erwartenden Gewinn entsprechen. Bezeichnet X den Gewinn, so gilt:

$$p(x) = P(X = x) = \frac{x^2 - (x-1)^2}{36} = \frac{2x-1}{36}, \quad x = 1, 2, \dots, 6$$

Der Erwartungswert von X ist gegeben durch:

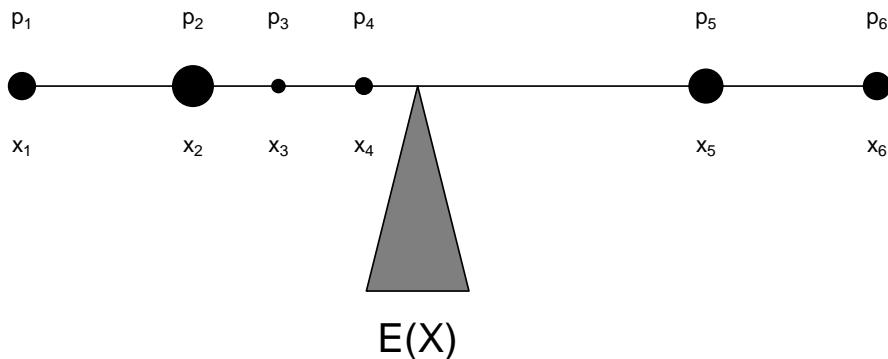
$$\mathbb{E}(X) = \sum_{x=1}^6 xp(x) = \frac{161}{36} \doteq 4.47 \implies d \doteq 4.47$$

Der Erwartungswert lässt sich auch wie folgt interpretieren: Angenommen, man spielt dieses Spiel n Mal, wobei n eine große Zahl sei (beispielsweise $n = 1000$). Ist $H_n(x)$ die (absolute) Häufigkeit eines Gewinns von x Euro, so beträgt der durchschnittliche Gewinn:

$$\frac{1}{n} \sum_{x=1}^6 x H_n(x) = \sum_{x=1}^6 x \frac{H_n(x)}{n} \approx \sum_{x=1}^6 xp(x) = \mathbb{E}(X)$$

Dabei legen wir die frequentistische Interpretation von Wahrscheinlichkeit (vgl. 2.1) zugrunde. ■

$\mathbb{E}(X)$ als Schwerpunkt: Der Erwartungswert lässt sich auch als *Schwerpunkt* von Punktmassen interpretieren. Werden (punktformige) Massen p_1, p_2, \dots, p_n an den Positionen x_1, x_2, \dots, x_n auf der reellen Achse plaziert, entspricht der Schwerpunkt des Systems dem Erwartungswert $\mathbb{E}(X) = \sum_i x_i p_i$.



Bem: Wie an den obigen Beispielen zu sehen, ist der Erwartungswert einer (diskreten) sG X *nicht* notwendigerweise ein Element des Merkmalraums M_X . Weiteres Beispiel: Die mittlere Augenzahl eines (balancierten) Würfels ist $\mathbb{E}(X) = 7/2 \notin M_X = \{1, 2, 3, 4, 5, 6\}$.

Erwartungswert einer stetigen sG: Ist X eine stetige sG mit der Dichtefunktion $f(x)$ und gilt $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$, so ist der **Erwartungswert** von X definiert durch:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Bsp 3.14 Die Dichte der sG X sei gegeben durch:

$$f(x) = \begin{cases} 4x^3 & 0 < x < 1 \\ 0 & \text{sonst} \end{cases}$$

Dann gilt:

$$\mathbb{E}(X) = \int_0^1 x (4x^3) dx = \int_0^1 4x^4 dx = \frac{4x^5}{5} \Big|_0^1 = \frac{4}{5}$$

■

Erwartungswert einer gemischten Verteilung: Den Erwartungswert einer gemischten Verteilung (vgl. 3.2.3) berechnet man wie folgt:

$$\mathbb{E}(X) = \sum_{i=1}^m x_i p(x_i) + \int_a^b x f^*(x) dx$$

(Dabei wird vorausgesetzt, dass die Summe und das Integral absolut konvergieren.)

Erwartungswert einer Funktion von X : Die sG $Y = g(X)$ sei eine Funktion der sG X .

- (a) Ist X diskret mit W-Funktion $p_X(x)$ und gilt $\sum_{x \in S_X} |g(x)| p_X(x) < \infty$, dann existiert der Erwartungswert von Y und ist gegeben durch:

$$\mathbb{E}(Y) = \sum_{x \in S_X} g(x) p_X(x)$$

- (b) Ist X stetig mit Dichte $f_X(x)$ und gilt $\int_{-\infty}^{\infty} |g(x)|f_X(x) dx < \infty$, dann existiert der Erwartungswert von Y und ist gegeben durch:

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} g(x)f_X(x) dx$$

Bem: Auch wenn die obigen Aussagen in manchen Lehrbüchern als *Definition* von $\mathbb{E}[g(X)]$ Verwendung finden, sollte man sich dessen bewusst sein, dass es sich tatsächlich um einen (mathematischen) *Satz* handelt. In der englischsprachigen Literatur wird er manchmal *Law of the Unconscious Statistician*¹⁰ (kurz LotUS) genannt. Diese – auf den ersten Blick seltsam anmutende – Bezeichnung soll darauf hinweisen, dass jemand, der den Erwartungswert von $g(X)$ nach den obigen Regeln berechnet (und glaubt, dass es sich dabei um eine Definition handelt), sich unbewusst wie ein/e Statistiker/in verhält (der/die weiß, dass es ein Satz ist).

Beweis(skizze): Nur für (a): Die W-Funktion $p_Y(y)$ von Y lässt sich wie folgt durch $p_X(x)$ ausdrücken:

$$p_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x:g(x)=y} P(X = x) = \sum_{x:g(x)=y} p_X(x)$$

Damit folgt:

$$\mathbb{E}(Y) = \sum_y y p_Y(y) = \sum_y \sum_{x:g(x)=y} y p_X(x) = \sum_x g(x) p_X(x)$$

(Punkt (b) lässt sich auf ähnliche Weise zeigen; dabei benötigt man aber etwas tieferliegende Resultate aus der Analysis.)

Bsp 3.15 Die Kernaussage des obigen Satzes besteht darin, dass man zur Berechnung des Erwartungswerts einer Funktion $Y = g(X)$ von X nicht zuerst die Verteilung von Y bestimmen muss, sondern auf die Verteilung von X zurückgreifen kann. Dazu ein einfaches Beispiel. Die Dichte von X sei gegeben durch:

$$f_X(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{sonst} \end{cases}$$

Dann lässt sich der Erwartungswert von beispielsweise $Y = g(X) = 1/X$ einfach wie folgt berechnen:

$$\mathbb{E}\left(\frac{1}{X}\right) = \int_{-\infty}^{\infty} g(x)f_X(x) dx = \int_0^1 \frac{1}{x} (2x) dx = \int_0^1 2 dx = 2$$

¹⁰Geprägt von SHELDON M. Ross, Prof. em. University of California/Berkeley.

Man kann aber auch zuerst die Dichte von Y bestimmen. Die Jacobian der Transformation $g(x) = 1/x$ ist $J = -1/y^2$ und mit dem Transformationssatz (vgl. 3.3.2) bekommt man die Dichte von Y :

$$f_Y(y) = f_X\left(\frac{1}{y}\right) \left| -\frac{1}{y^2} \right| = \frac{2}{y^3}, \quad 1 < y < \infty$$

Der Erwartungswert von Y lässt sich dann wie folgt berechnen:

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_1^{\infty} \frac{2}{y^2} dy = -\frac{2}{y} \Big|_1^{\infty} = 2$$

Klarerweise stimmen die Erwartungswerte bei beiden Berechnungen überein. Die erste Berechnung war aber deutlich einfacher. (Allerdings haben wir bei der zweiten Berechnung mehr an Information gewonnen, nicht nur den Erwartungswert sondern auch die Verteilung (Dichte) von Y .) ■

Eigenschaften des Erwartungswerts: Für Konstanten a, b, k_1, k_2 und Funktionen g, h gilt:

- (1) $\mathbb{E}(a) = a$
- (2) $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
- (3) $\mathbb{E}[k_1g(X) + k_2h(X)] = k_1\mathbb{E}[g(X)] + k_2\mathbb{E}[h(X)]$

(Beweis als UE–Aufgabe.)

3.5 Varianz

Im vorigen Abschnitt wurde der Erwartungswert $\mathbb{E}(X)$ einer sG X als die wichtigste Maßzahl für die **Lage** einer Verteilung (oder einer sG) definiert. Als Standardbezeichnung hat sich μ_X (oder kurz μ) etabliert. Weitere wichtige Maßzahlen der Lage sind die p –Quantile x_p (vgl. 3.2), insbesondere der **Median** (= 0.5–Quantil):

$$x_{0.5} = F_X^{-1}\left(\frac{1}{2}\right)$$

(F^{-1} ist die verallgemeinerte Inverse von F ; vgl. 3.2.) Für den Median sind mehrere Bezeichnungen gebräuchlich. Neben $x_{0.5}$ schreibt man auch \tilde{x} , $\text{Median}(X)$, $\text{med}(X)$, o. Ä.

Neben Maßzahlen der Lage benötigt man aber auch Maßzahlen für das **Streuungsverhalten** einer Verteilung (oder sG). Die wichtigste Maßzahl dieser Art ist die **Varianz**.

Varianz/Streuung einer sG: X sei eine sG mit endlichem Mittelwert μ_X und derart, dass $\mathbb{E}[(X - \mu_X)^2]$ endlich ist, dann ist die **Varianz** von X definiert durch:

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2]$$

Die Standardbezeichnung für die Varianz ist σ_X^2 (oder kurz σ^2) Die (positive) Wurzel aus der Varianz nennt man die **Streuung** (oder die **Standardabweichung**¹¹) von X :

$$\text{Streuung } (X) = +\sqrt{\text{Var}(X)}$$

Die Standardbezeichnung für die Streuung ist σ_X (oder kurz σ).

Die Varianz ist also die **mittlere quadratische Abweichung** einer sG von ihrem Mittelwert, somit der Erwartungswert von $Y = g(X) = (X - \mu)^2$. Aus dem vorigen Abschnitt wissen wir, wie ein derartiger Erwartungswert zu berechnen ist:

$$\text{diskret: } \text{Var}(X) = \sum_x [x - \mathbb{E}(X)]^2 p_X(x)$$

$$\text{stetig: } \text{Var}(X) = \int_{-\infty}^{\infty} [x - \mathbb{E}(X)]^2 f_X(x) dx$$

$$\text{gemischt: } \text{Var}(X) = \sum_{i=1}^m [x_i - \mathbb{E}(X)]^2 p(x_i) + \int_{-\infty}^{\infty} [x - \mathbb{E}(X)]^2 f^*(x) dx$$

Meist ist die Varianzberechnung mit Hilfe des folgenden Satzes einfacher.

Verschiebungssatz für die Varianz: Die Varianz σ_X^2 einer sG X lässt sich auch wie folgt berechnen:

$$\sigma_X^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \mathbb{E}(X^2) - \mu_X^2$$

Beweis: Unter Verwendung der Rechenregeln für den Erwartungswert gilt:

$$\begin{aligned} \sigma^2 &= \mathbb{E}[(X - \mu)^2] = \mathbb{E}(X^2 - 2\mu X + \mu^2) \\ &= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 \\ &= \mathbb{E}(X^2) - \mu^2 \end{aligned}$$

¹¹engl. *standard deviation* (abgekürzt sd)

Da Varianzen nichtnegative Größen sind, folgt aus dem Verschiebungssatz die wichtige Ungleichung:

$$\mathbb{E}(X^2) \geq [\mathbb{E}(X)]^2$$

(Bem: Für den Ausdruck auf der rechten Seite schreibt man meist kürzer $\mathbb{E}^2(X)$.)

Bsp 3.16 Die Varianz der (stetigen) sG von Bsp 3.14 lässt sich nach Definition berechnen (UE-Aufgabe):

$$\text{Var}(X) = \int_0^1 \left(x - \frac{4}{5}\right)^2 (4x^3) dx$$

Einfacher ist die Berechnung mittels Verschiebungssatz:

$$\mathbb{E}(X^2) = \int_0^1 x^2 (4x^3) dx = 4 \int_0^1 x^5 dx = \frac{2x^6}{3} \Big|_0^1 = \frac{2}{3}$$

Somit:

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75} \implies \sigma_X = \sqrt{\frac{2}{75}}$$

■

Eigenschaften der Varianz/Streuung: Für Konstanten a, b gilt:

- (1) $\text{Var}(a) = 0$
- (2) $\text{Var}(aX + b) = a^2 \text{Var}(X)$
- (3) $\sigma_{aX+b} = a\sigma_X$

(Beweis als UE-Aufgabe.)

Einheiten der Kenngrößen: Bei konkreten Anwendungen ist zu beachten, dass die hier behandelten Kenngrößen μ_X (Mittelwert), σ_X^2 (Varianz) und σ_X (Streuung) **Einheiten** haben. Ist beispielsweise die sG X ein Gewicht in der Einheit [kg], hat μ_X die Einheit [kg], σ_X^2 die Einheit [kg²] und σ_X die Einheit [kg]. Auch der Median (oder ein anderes Quantil) hat in diesem Beispiel die Einheit [kg].

MAD: Ein weiteres wichtiges Streuungsmaß ist die **mittlere absolute Abweichung** vom Median. Für eine stetige sG X mit Dichte f_X ist der **MAD** definiert durch:

$$\text{MAD}(X) = \int_{-\infty}^{\infty} |x - \text{med}(X)| f_X(x) dx$$

Bem: Der MAD ist zwar in gewisser Weise ein „natürlicheres“ Streuungsmaß als die Streuung σ_X , wegen des Absolutbetrages aber meist schwieriger zu berechnen.

3.6 Simulation

Die Simulation von stochastischen Vorgängen verschiedenster Art ist mittlerweile ein unverzichtbares Werkzeug der modernen (Computer-) Statistik. Der erste (und – aus statistischer Sicht – schwierigste) Schritt dabei ist die Erzeugung von (unbeschränkt vielen unabhängigen) Realisationen einer sG U mit der folgenden **uniformen** Dichte:

$$f_U(u) = \begin{cases} 1 & 0 < u < 1 \\ 0 & \text{sonst} \end{cases}$$

Es existieren zahlreiche Tabellen mit **echten** Zufallszahlen.¹² In der Praxis verwendet man aber computergenerierte **Pseudozufallszahlen**, d. h., algorithmisch erzeugte Zahlen, die den *Anschein* von echten Zufallszahlen erwecken.

Einfachere (und ältere) **Generatoren** sind meist von folgender Bauart: Starte mit einem Anfangswert x_0 (genannt *Seed*) und berechne rekursiv Werte wie folgt:

$$x_{n+1} = (ax_n + c)(\text{mod } m), \quad n \geq 0$$

Dabei sind a , c und m natürliche Zahlen. Durch die Rechnung modulo m ist jedes x_n eine Zahl aus $0, 1, \dots, m-1$ und als Näherung für eine Realisation von U nimmt man x_n/m . Durch entsprechende Wahl von a , c und m lassen sich auf diese Weise Zahlen erzeugen, die den Anschein von echten U -Realisationen erwecken.

Bsp 3.17 Wir betrachten als Illustration einen Generator mit $m = 23^4 = 279841$, $a = 7200$ und $c = 1$:

$$x_{n+1} = (7200x_n + 1)(\text{mod } 279841) \quad \text{und} \quad u_n = \frac{x_n}{279841}$$

¹²Die bekannteste Tabelle dieser Art wurde von der RAND (*Research AND Development*) Corporation herausgegeben: *A Million Random Digits with 100,000 Normal Deviates* (1955).

Wählt man als Startwert beispielsweise $x_0 = 1$, werden die folgenden Zahlen erzeugt:

	x	u
1	7201	0.025732
2	76616	0.273784
3	68590	0.245103
4	208477	0.744984
5	247118	0.883066
6	20523	0.073338
7	9553	0.034137
8	220556	0.788148
9	185367	0.662401
10	80672	0.288278
.	.	.
.	.	.
.	.	.
279838	227376	0.812519
279839	37351	0.133472
279840	0	0.000000
279841	1	0.000004
279842	7201	0.025732
.	.	.
.	.	.
.	.	.

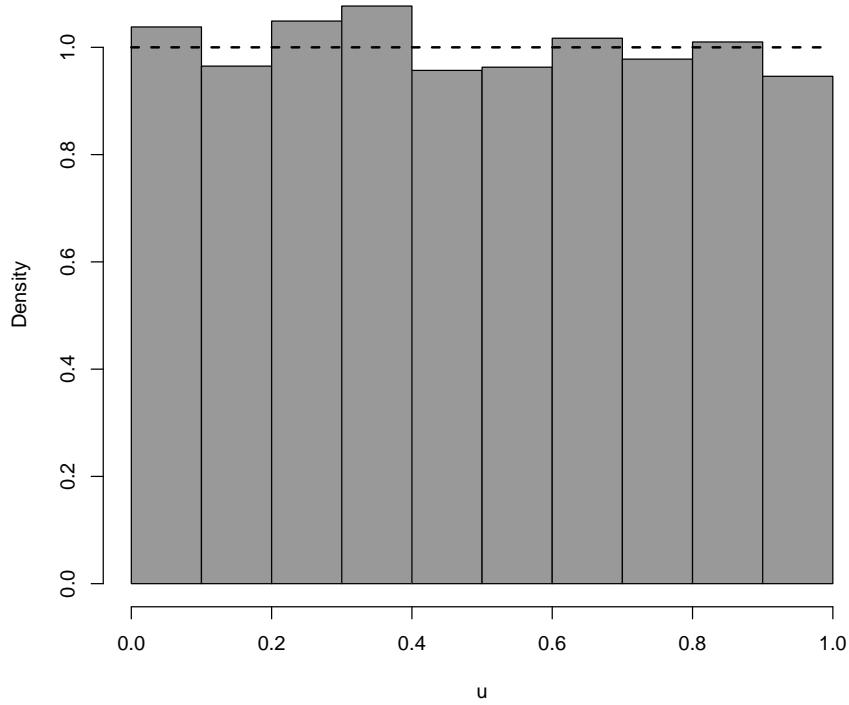
Man beachte, dass für einen Generator dieser Art, unabhängig vom Startwert x_0 , die **Periode** gleich m ist, d.h., ab der m -ten Zufallszahl wiederholt sich exakt die gleiche Folge. (Abb 3.12 zeigt ein Histogramm der ersten 10000 Zufallszahlen.) ■

Bem: Neuere Generatoren sind komplexer (und versuchen, einige der Probleme mit den obigen *linearen Kongruenzgeneratoren* zu vermeiden). Ein häufig verwendeter Generator neueren Typs ist der sogenannte *Mersenne-Twister*¹³, entwickelt 1997 von MAKOTO MATSUMOTO und TAKUJI NISHIMURA. Das ist auch der standardmäßig von der R-Funktion `runif()` verwendete Generator.

Hat man eine zuverlässige Methode zur Erzeugung von auf $(0, 1)$ uniform verteilten (Pseudo-) Zufallszahlen zur Verfügung, stellt sich im nächsten Schritt die Frage, wie Realisationen für eine beliebige sG X erzeugt werden können. Zur Beantwortung dieser Frage gibt es eine ganze Reihe von (z.T. sehr speziellen) Methoden. Im Folgenden soll nur *eine* allgemein anwendbare (in vielen Fällen aber nicht sehr effiziente) Methode vorgestellt werden.¹⁴

¹³<http://de.wikipedia.org/wiki/Mersenne-Twister>

¹⁴Vgl. für einen Überblick über die verschiedenen Methoden ROBERT & CASELLA (2010).

Abbildung 3.12: Histogramm von 10000 Zufallszahlen (Generator von Bsp 3.17)

Bezeichnung: Hat die sG X die Verteilungsfunktion F_X , so schreibt man kurz $X \sim F_X$. Ebenso bedeutet $X \sim f_X$, dass X die Dichte f_X hat, oder $X \sim p_X$, dass X die W-Funktion p_X hat. (Meist lässt man den Index X auch weg.)

Behauptung: Ist F eine (beliebige) Verteilungsfunktion und U eine sG mit uniformer Dichte $f_U(u) = I_{(0,1)}(u)$, so gilt:

$$X := F^{-1}(U) \sim F$$

Dabei ist F^{-1} die verallgemeinerte Inverse von F .

Beweis (für streng monoton wachsendes F): Für die VF von X gilt:

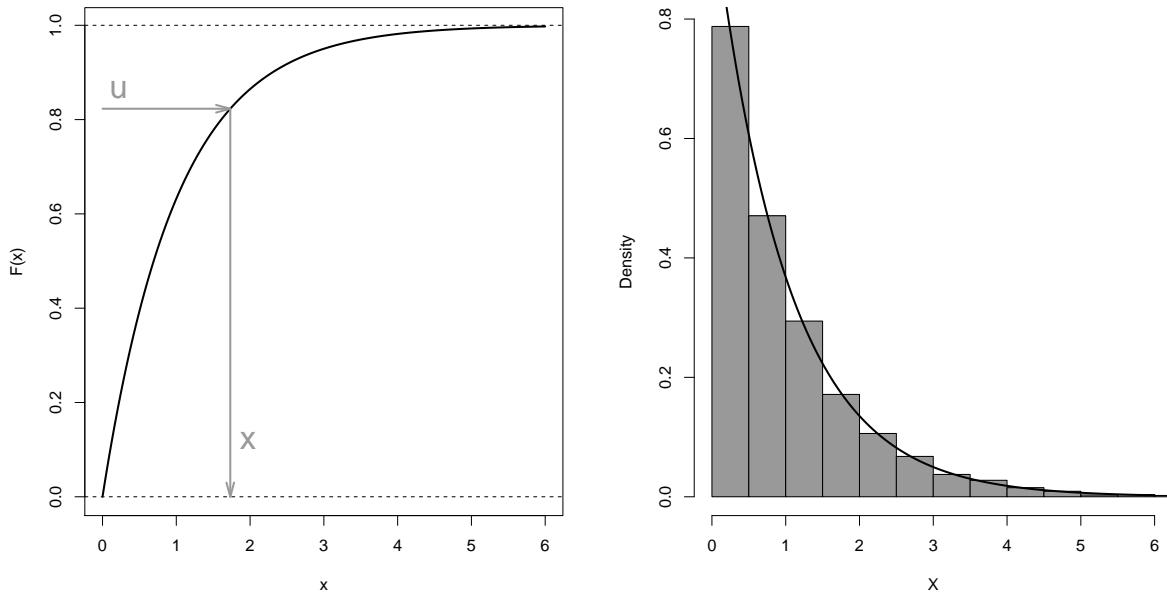
$$F_X(x) = P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F_U(F(x))$$

Die VF von U ist gegeben durch:

$$F_U(u) = P(U \leq u) = \begin{cases} 0 & u < 0 \\ u & 0 \leq u < 1 \\ 1 & u \geq 1 \end{cases}$$

Damit folgt: $F_X(x) = F_U(F(x)) = F(x)$. Das war zu zeigen.

Abbildung 3.13: Inversionsmethode (stetige Verteilung)



Inversionsmethode: Für die Erzeugung einer Realisation x einer sG $X \sim F$ genügen die beiden folgenden Schritte:

- (1) Erzeuge eine Realisation u von $U \sim f_U(u) = I_{(0,1)}(u)$.
- (2) Bilde $x = F^{-1}(u)$.

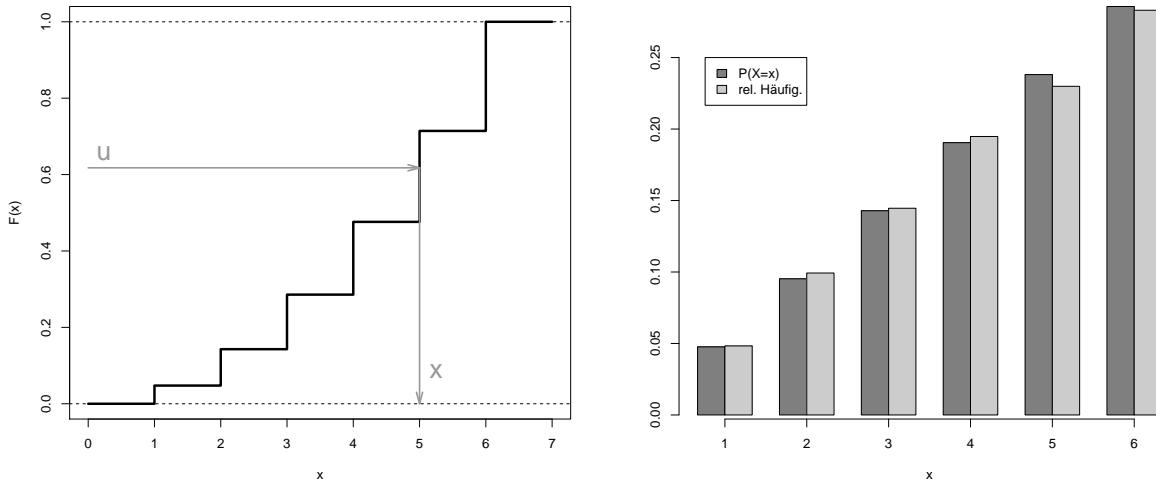
Bsp 3.18 Zur Generierung von Realisationen einer sG X mit beispielsweise der Dichte $f_X(x) = e^{-x} I_{(0,\infty)}(x)$ bestimmt man zuerst die Verteilungsfunktion von X :

$$F_X(x) = \int_0^x e^{-t} dt = 1 - e^{-x}, \quad 0 < x < \infty$$

Nach der Inversionsmethode ist die VF zu invertieren:

$$1 - e^{-x} = u \implies x = -\ln(1 - u)$$

Der linke Teil von Abb 3.13 ist eine grafische Veranschaulichung der Inversionsmethode. Der rechte Teil zeigt ein Histogramm von 10000 auf diese Weise generierten Zufallszahlen. (Die darüber gezeichnete Linie entspricht der Dichte von X .) ■

Abbildung 3.14: Inversionsmethode (diskrete Verteilung)

Bsp 3.19 Die Inversionsmethode lässt sich auch im diskreten (oder gemischten) Fall anwenden. Angenommen, wir möchten einen Würfel simulieren, bei dem die Wahrscheinlichkeit, die Augenzahl k zu werfen, proportional zu k ist (vgl. auch Bsp 6 von 2.16). Ist X die geworfene Augenzahl, so gilt in diesem Fall:

$$p_X(x) = P(X = x) = \frac{x}{21} \quad \text{für } x = 1, 2, 3, 4, 5, 6$$

Der linke Teil von Abb 3.14 zeigt die Verteilungsfunktion sowie das Prinzip der Generierung von nach p_X verteilten Zufallszahlen. Nach der Inversionsmethode ist die VF zu invertieren, wobei in diesem Fall auf die verallgemeinerte Inverse zurückgegriffen werden muss. Der rechte Teil von Abb 3.14 stellt die tatsächlichen Wahrscheinlichkeiten p_X den relativen Häufigkeiten von 10000 auf diese Weise generierten Realisationen von X gegenüber. ■

Aufgaben

- 3.1 Zwei (symmetrische) Würfel werden geworfen und X_{\max} sei die größere der beiden Augenzahlen. Bestimmen Sie die W-Funktion und die Verteilungsfunktion von X_{\max} . Wiederholen Sie die Aufgabe für X_{\min} (= kleinere der beiden Augenzahlen). Stellen Sie die beiden Verteilungsfunktionen in einem Plot dar.
- 3.2 In einem Behälter befinden sich (gut gemischt) N weiße und M schwarze Kugeln. Die Kugeln werden eine nach der anderen zufällig mit Zurücklegen solange gezogen, bis man die erste schwarze Kugel bekommt. Wenn X die Nummer der Ziehung der ersten schwarzen Kugel ist, bestimmen Sie (a) $P(X = x)$ und (b) $F_X(x) = P(X \leq x)$. (Hinweis: Bestimmen Sie zunächst $P(X > x)$.) Stellen Sie die Verteilungsfunktion von X grafisch dar (beispielsweise für $N = 20$ und $M = 10$).

- 3.3 Ein W–Raum (Ω, \mathcal{A}, P) sei gegeben durch $\Omega = \{\omega \mid 0 < \omega < 10\}$ und für $A \in \mathcal{A}$ sei $P(A) = \int_A \frac{1}{10} dx$. Begründen Sie, warum $X(\omega) = \omega^2$ eine stochastische Größe ist. Bestimmen Sie M_X und $F_X(x) = P_X(X \leq x)$ für $x \in \mathbb{R}$.

- 3.4 Die Verteilungsfunktion einer stetigen sG X sei gegeben durch:

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2/5 & 0 \leq x \leq 1 \\ (-x^2 + 6x - 4)/5 & 1 < x \leq 3 \\ 1 & x > 3 \end{cases}$$

- (a) Stellen Sie die Verteilungsfunktion grafisch dar.
 (b) Bestimmen Sie die Dichte und stellen Sie auch Letztere grafisch dar.
 (c) Berechnen Sie die folgenden Wahrscheinlichkeiten: $P(X \leq 2)$, $P(1 < X \leq 2)$, $P(1 \leq X \leq 2)$ und $P(X > 1/2)$.

- 3.5 Die Grünphase (einschließlich Blinkphase) bei einer Fußgängerampel beträgt 25 Sekunden, die Rotphase 65 Sekunden. Sie kommen zu einem zufälligen Zeitpunkt zu dieser Ampel und X sei die Wartezeit. Bestimmen Sie:

- (a) die Verteilungsfunktion von X (plus Zeichnung). Um welchen Verteilungstyp (diskret, stetig, gemischt) handelt es sich?
 (b) die Wahrscheinlichkeit, dass Sie länger als 20 Sekunden warten.
 (c) die (bedingte) Wahrscheinlichkeit, dass Sie noch mindestens weitere 20 Sekunden warten, wenn Sie bereits 20 Sekunden gewartet haben.
 (d) das 10%-, 25%-, 50%- und das 90%-Quantil von X .

- 3.6 Die Verteilungsfunktion einer stetigen sG X ist gegeben durch:

$$F(x) = \frac{e^x}{1 + e^x}, \quad -\infty < x < \infty$$

- (a) Stellen Sie die Funktion grafisch dar und überzeugen Sie sich davon, dass F alle Eigenschaften einer (stetigen) VF erfüllt.
 (b) Ermitteln Sie allgemein einen Ausdruck für das p -Quantil x_p und bestimmen Sie konkret die drei Quartile (d. h., 25%, 50%, 75%) der Verteilung.
 (c) Bestimmen Sie die zugehörige Dichte f und stellen Sie sie grafisch dar.

- 3.7 Die sG X habe die Dichte $f_X(x) = x^2/9$, $0 < x < 3$, gleich Null sonst. Bestimmen Sie mittels Transformationssatz die Dichte von $Y = X^3$.

- 3.8 X sei uniform verteilt mit Dichte $f_X(x) = 1/\pi$, $-\pi/2 < x < \pi/2$. Bestimmen Sie mittels Transformationssatz die Dichte von $Y = \tan(X)$.

- 3.9 Die sG X habe eine Verteilung mit der Dichte $f(x) = e^{-x}$, $x > 0$. Bestimmen Sie die Dichte von $Y = \sqrt{X}$. Verwenden Sie dazu (a) die Methode der VF und (b) den Transformationssatz. Erstellen Sie eine Abbildung der Dichte.
- 3.10 Eine Übung wird in vier Gruppen zu 20, 25, 35 bzw. 40 Student/inn/en abgehalten. Wenn von den insgesamt 120 Personen, die an der Übung teilnehmen, eine Person zufällig ausgewählt wird und X die Größe der Gruppe ist, aus der die Person stammt, berechnen Sie $\mathbb{E}(X)$. Geben Sie eine anschauliche Erklärung dafür, warum $\mathbb{E}(X)$ größer als die durchschnittliche Gruppengröße $(20 + 25 + 35 + 40)/4 = 30$ ist.
- 3.11 Bestimmen Sie für Aufgabe 3.1 den Erwartungswert von X_{\max} und X_{\min} .
- 3.12 Berechnen Sie den Erwartungswert einer sG X mit der Dichte $f(x) = e^{-x} I_{(0,\infty)}(x)$.
- 3.13 Berechnen Sie den Erwartungswert der sG von Aufgabe 3.4.
- 3.14 Berechnen Sie den Erwartungswert der Wartezeit bei der Fußgängerampel von Aufgabe 3.5.
- 3.15 Für eine *positive* sG X mit der Verteilungsfunktion F kann der Erwartungswert auch wie folgt berechnet werden:

$$\mathbb{E}(X) = \int_0^{\infty} [1 - F(x)] dx$$

Berechnen Sie auf diese Weise die Erwartungswerte von Aufgabe 3.12 und 3.14.

- 3.16 Zeigen Sie, dass der Erwartungswert der sG Y von Aufgabe 3.8 nicht existiert.
- 3.17 X habe die Dichte $f(x) = 3x^2$, $0 < x < 1$, gleich Null sonst. Betrachten Sie ein Rechteck mit den Seiten X und $1 - X$. Bestimmen Sie den Erwartungswert der Fläche.
- 3.18 Sei $f(x) = 3x^2$, $0 < x < 1$, gleich Null sonst, die Dichte von X .
- (a) Berechnen Sie $\mathbb{E}(X^3)$.
 - (b) Bestimmen Sie die Dichte von $Y = X^3$.
 - (c) Berechnen Sie $\mathbb{E}(Y)$ mit Hilfe von (b) und vergleichen Sie mit (a).
- 3.19 Betrachten Sie eine diskrete sG X mit Merkmalraum $M = \{1, 2, \dots, k\}$ und W-Funktion $p(x) = 1/k$ für $x \in M$. Bestimmen Sie (a) den Mittelwert $\mu = \mathbb{E}(X)$ und (b) die Varianz $\sigma^2 = \text{Var}(X)$. Betrachten Sie speziell den Fall $k = 6$ (\cong Augenzahl eines üblichen Würfels). (Hinweis zu (b): Verwenden Sie den Verschiebungssatz.)
- 3.20 Berechnen Sie die Varianz einer sG X mit der Dichte $f(x) = e^{-x} I_{(0,\infty)}(x)$. (Hinweis: Verwenden Sie den Verschiebungssatz; vgl. auch Aufgabe 3.12.)
- 3.21 Berechnen Sie die Varianz und die Streuung der Wartezeit bei der Fußgängerampel von Aufgabe 3.5. (Hinweis: Verwenden Sie den Verschiebungssatz; vgl. auch Aufgabe 3.14.)

- 3.22 Wie kann man Realisationen einer sG X mit der Dichte $f(x) = 3x^2 I_{(0,1)}(x)$ erzeugen? Schreiben Sie eine R-Funktion und erzeugen Sie damit $N = 1000$ Zufallszahlen. Stellen Sie das Ergebnis in Form eines Histogramms dar.
- 3.23 Wie kann man Realisationen einer sG X mit der (logistischen) Verteilung von Aufgabe 3.6 erzeugen? Schreiben Sie eine R-Funktion und erzeugen Sie damit $N = 10000$ Zufallszahlen. Stellen Sie das Ergebnis in Form eines Histogramms dar.
- 3.24 Wie kann man Wartezeiten bei der Fußgängerampel von Aufgabe 3.5 simulieren? Schreiben Sie eine R-Funktion und erzeugen Sie damit $N = 100$ Wartezeiten.
- 3.25 Schreiben Sie eine R-Funktion für die Simulation einer sG X mit der Dichte:

$$f(x) = 30(x^2 - 2x^3 + x^4) \quad \text{für } 0 < x < 1$$

Erzeugen Sie $N = 10000$ Realisationen von X und stellen Sie das Ergebnis in Form eines Histogramms dar. (Hinweis: Verwenden Sie zur Invertierung der VF die Funktion `uniroot()`.)

4 Spezielle Verteilungen

4.1 Diskrete Verteilungen

4.1.1 Diskrete uniforme Verteilung

Eine stochastische Größe X hat eine (diskrete) **uniforme Verteilung** (oder (diskrete) **Gleichverteilung**) auf der Menge $M = \{x_1, x_2, \dots, x_n\}$ (mit $x_i \neq x_j$ für $i \neq j$), wenn jedes Element von M die gleiche Wahrscheinlichkeit hat:

$$p(x_i) = \frac{1}{n}, \quad i = 1, 2, \dots, n$$

Der Erwartungswert von X ist gegeben durch:

$$\mathbb{E}(X) = \sum_{x \in M} xp(x) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

Für die Varianz von X gilt:

$$\text{Var}(X) = \sum_{x \in M} [x - \mathbb{E}(X)]^2 p(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Letzterer Ausdruck lässt sich auch wie folgt schreiben:

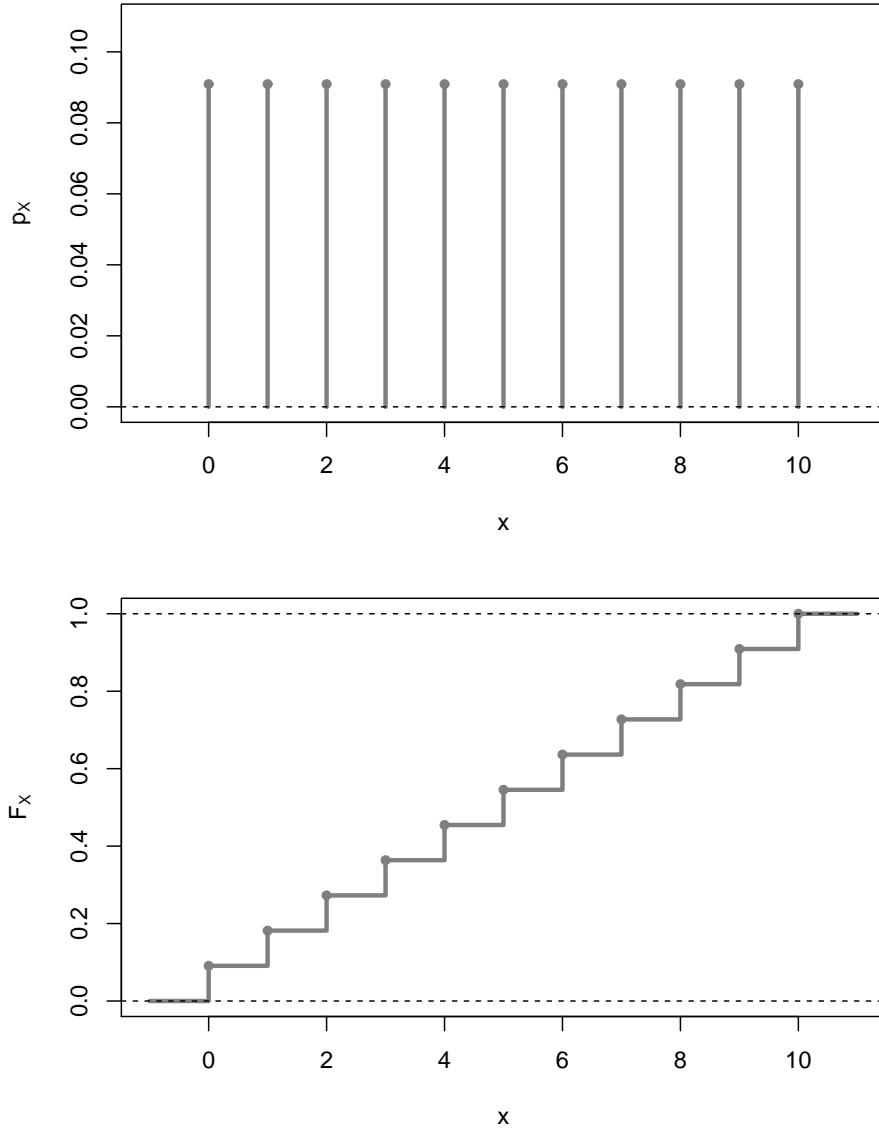
$$\text{Var}(X) = \frac{1}{n} \left[\sum_{i=1}^n x_i^2 - n(\bar{x})^2 \right]$$

Spezialfall: Besteht der Merkmalraum M aus aufeinanderfolgenden ganzen Zahlen, d. h., gilt $M = \{a, a+1, a+2, \dots, b\}$ mit $a \leq b$, $a \in \mathbb{Z}$, so ist der Mittelwert und die Varianz gegeben durch:

$$\mathbb{E}(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a+1)^2 - 1}{12}$$

Speziell für $M = \{1, 2, \dots, k\}$ gilt:

$$\mathbb{E}(X) = \frac{1+k}{2}, \quad \text{Var}(X) = \frac{k^2 - 1}{12}$$

Abbildung 4.1: Diskrete uniforme Verteilung auf $M = \{0, 1, 2, \dots, 10\}$ 

Bsp 4.1 Hat X eine diskrete uniforme Verteilung auf $M = \{0, 1, 2, \dots, 10\}$, so gilt:

$$\mathbb{E}(X) = \frac{0 + 10}{2} = 5, \quad \text{Var}(X) = \frac{(10 - 0 + 1)^2 - 1}{12} = \frac{120}{12} = 10$$

Abb 4.1 zeigt die W-Funktion (p_X) und die Verteilungsfunktion (F_X). ■

Zufallszahlen: Um Realisationen einer diskreten uniformen Verteilung zu generieren, kann man sich der R-Funktion `sample()` bedienen. Beispielsweise lassen sich 100 Realisationen der sG von Bsp 4.1 wie folgt erzeugen:

```
(x <- sample(0:10, size=100, replace=TRUE))
[1] 8 6 9 4 9 0 3 1 4 9 6 3 7 0 9 6 1 8 7 9
[21] 3 2 8 9 10 6 0 0 0 3 4 2 0 8 7 1 10 5 2 7
[41] 3 10 8 8 0 7 0 8 1 10 6 0 4 1 3 0 5 7 4 1
[61] 6 10 7 7 7 7 10 10 9 10 10 7 5 10 10 7 4 8 0 3
[81] 3 5 9 1 1 2 9 3 6 7 0 8 10 0 3 4 5 8 5 1
table(x)
x
0 1 2 3 4 5 6 7 8 9 10
13 9 4 10 7 6 7 13 10 9 12
```

4.1.2 Bernoulli–Verteilung

Man spricht von einem **Bernoulli–Experiment**¹, wenn man nur beobachtet, ob ein bestimmtes Ereignis A eintritt oder nicht. Die zugehörige sG ist nur ein **Indikator** für den Eintritt von A :

$$X = \begin{cases} 1 & A \text{ tritt ein („Erfolg“)} \\ 0 & A \text{ tritt nicht ein („Misserfolg“)} \end{cases}$$

Gilt $p = P(X = 1)$ und $q = 1 - p = P(X = 0)$, so hat X eine **Bernoulli–Verteilung** (oder **Alternativverteilung**) $A(p)$ und die W–Funktion von X ist gegeben durch:

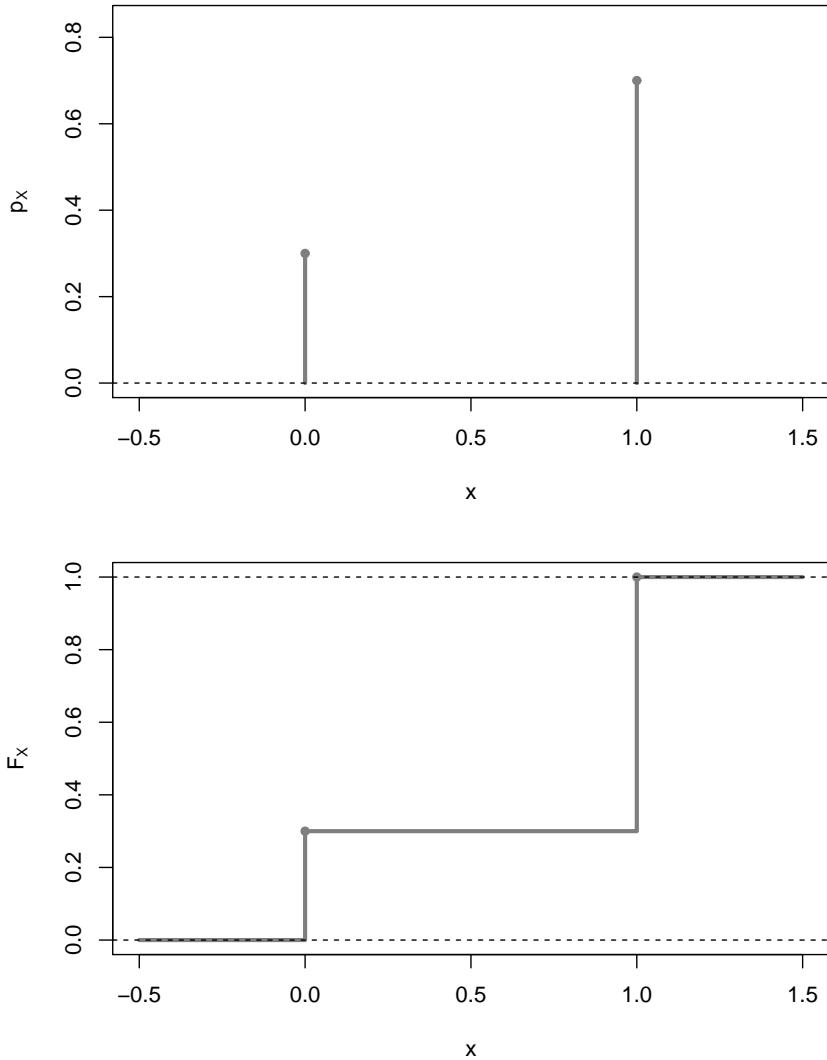
$$p(x) = p^x(1 - p)^{1-x} \quad \text{für } x \in \{0, 1\}$$

Den Erwartungswert und die Varianz berechnet man wie folgt:

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=0}^1 xp^x(1 - p)^{1-x} = (0)(1 - p) + (1)(p) = p \\ \mathbb{E}(X^2) &= \sum_{x=0}^1 x^2 p^x(1 - p)^{1-x} = (0^2)(1 - p) + (1^2)(p) = p\end{aligned}$$

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = p - p^2 = p(1 - p)$$

¹JAKOB (I.) BERNOULLI (1655–1705), Schweizer Mathematiker (wesentliche Beiträge zur Wahrscheinlichkeitstheorie) und Physiker.

Abbildung 4.2: Bernoulli–Verteilung A(0.7)

Bsp 4.2 Die W–Funktion (p_X) und die Verteilungsfunktion (F_X) einer A(0.7)–Verteilung ist in Abb 4.2 grafisch dargestellt. Der Erwartungswert ist $\mu = 0.7$ und die Varianz beträgt $\sigma^2 = (0.7)(0.3) = 0.21$. ■

4.1.3 Binomialverteilung

Werden n unabhängige und identische Bernoulli–Experimente durchgeführt, so ist das Ergebnis ein n –Tupel aus Nullen und Einsen, beispielsweise:

$$\underbrace{(0, 0, 1, 0, 1, \dots, 1)}_{n \text{ Elemente}}$$

Häufig ist man aber nur an der **Anzahl** der Erfolge interessiert und nicht an der Reihenfolge ihres Auftretens. Bezeichnet die sG X die Zahl der Erfolge bei n Bernoulli-Experimenten, so kann X die Werte $0, 1, \dots, n$ annehmen. Gibt es x Erfolge (und daher $n - x$ Misserfolge), so hat man:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

verschiedene Möglichkeiten, die Positionen für die x Erfolge zu wählen. Die Wahrscheinlichkeit für jede dieser Möglichkeiten beträgt $p^x(1-p)^{n-x}$. Die W-Funktion von X ist daher gegeben durch:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{für } x \in \{0, 1, 2, \dots, n\}$$

Eine sG mit der obigen W-Funktion hat eine **Binomialverteilung** und man schreibt $X \sim \mathcal{B}(n, p)$. Mit Hilfe des *Binomischen Lehrsatzes* zeigt man, dass die Summe der Punkt-wahrscheinlichkeiten $p(x) = P(X = x)$ gleich Eins ist:

$$\sum_{x=0}^n p(x) = \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = [p + (1-p)]^n = 1$$

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim \mathcal{B}(n, p)$ sind gegeben durch:

$$\mathbb{E}(X) = np, \quad \text{Var}(X) = np(1-p)$$

Beweis: Nur für den Erwartungswert (Herleitung der Varianz als UE-Aufgabe):

$$\begin{aligned} \mu &= \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^n x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= np \sum_{x=1}^n \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \\ &= np \sum_{x=0}^{n-1} \frac{(n-1)!}{x!(n-1-x)!} p^x (1-p)^{n-1-x} \\ &= np \underbrace{\sum_{x=0}^{n-1} \binom{n-1}{x} p^x (1-p)^{n-1-x}}_{= 1 \dots \mathcal{B}(n-1, p)} \\ &= np \end{aligned}$$

Spezialfall: Der Fall $n = 1$ entspricht der Bernoulli–Verteilung: $B(1, p) \equiv A(p)$.

Modalwert(e): Die Modalwerte (d. h. die x –Werte mit $p(x) = \max$) sind gegeben durch:

$$x_{\text{mod}} = \begin{cases} \lfloor (n+1)p \rfloor & \text{falls } (n+1)p \notin \mathbb{N} \\ (n+1)p - 1, (n+1)p & \text{falls } (n+1)p \in \mathbb{N} \end{cases}$$

(Man beachte, dass es für $(n+1)p \in \mathbb{N}$ zwei Modalwerte gibt.)

Beweis: Man betrachte den Quotienten zweier aufeinanderfolgender Wahrscheinlichkeiten:

$$\frac{p(x+1)}{p(x)} = \frac{\binom{n}{x+1} p^{x+1} (1-p)^{n-x-1}}{\binom{n}{x} p^x (1-p)^{n-x}} = \left(\frac{n-x}{x+1}\right) \left(\frac{p}{1-p}\right) \geq 1 \iff (n+1)p \geq x+1$$

Das kleinste x , das die Bedingung *nicht* erfüllt, ist der Modalwert.

Bsp 4.3 Die W–Funktion (p_X) und die Verteilungsfunktion (F_X) einer $B(10, 0.7)$ –Verteilung ist in Abb 4.3 grafisch dargestellt. Der Erwartungswert ist $\mu = (10)(0.7) = 7$ und die Varianz beträgt $\sigma^2 = (10)(0.7)(0.3) = 2.1$. Der Modalwert ist in diesem Fall eindeutig bestimmt und gegeben durch $x_{\text{mod}} = \lfloor (11)(0.7) \rfloor = \lfloor 7.7 \rfloor = 7$. ■

4.1.4 Negative Binomialverteilung

Man betrachte eine Folge von unabhängigen Wiederholungen eines Bernoulli–Experiments mit konstanter Erfolgswahrscheinlichkeit p . Ist X die Gesamtzahl der Versuche, die notwendig sind, um exakt r Erfolge zu bekommen, so gilt:

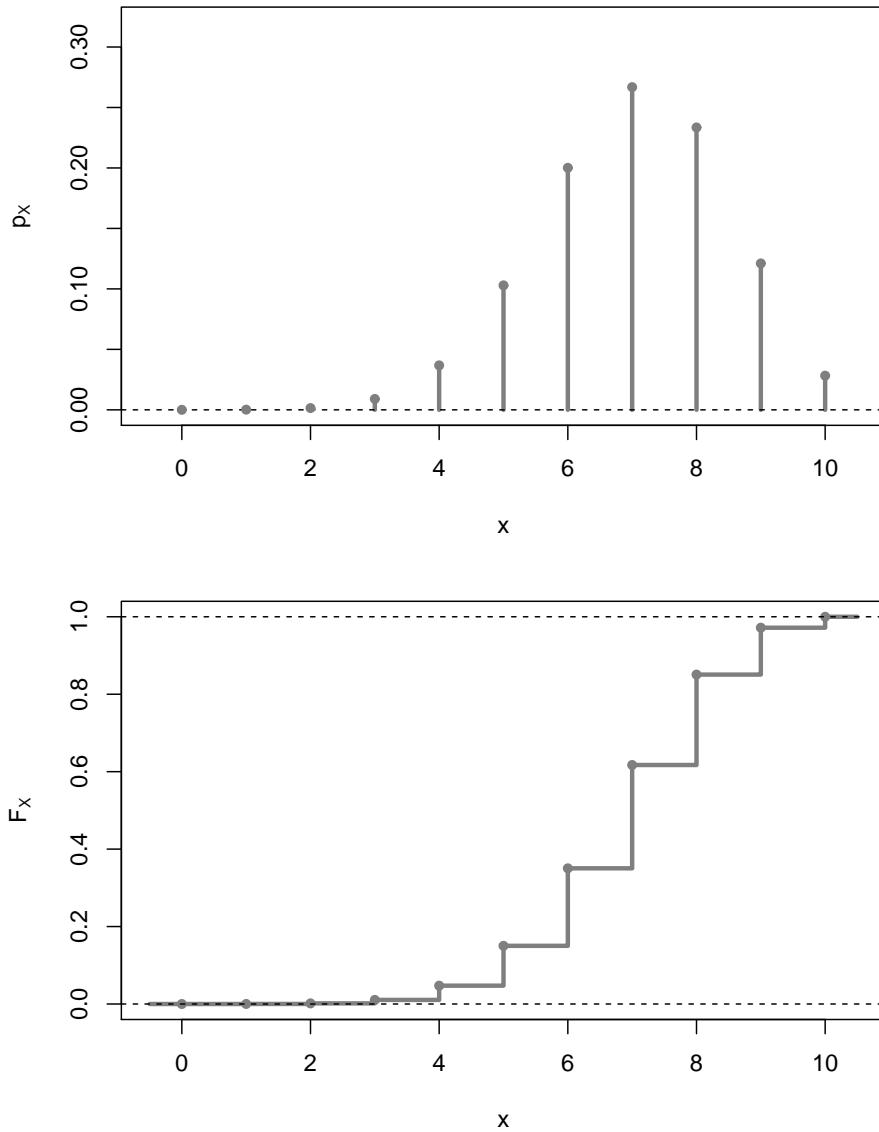
$$p(x) = P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad \text{für } x \in \{r, r+1, \dots\}$$

Eine Verteilung mit dieser W–Funktion nennt man eine **Negative Binomialverteilung**² und schreibt $X \sim NB(r, p)$.

Bem: In Lehrbüchern wird häufig auch die Verteilung von $Y = X - r$ (= Zahl der *Miss-Erfolge* vor dem r –ten Erfolg) als $NB(r, p)$ –Verteilung bezeichnet. Die W–Funktion von Y lässt sich einfach wie folgt bestimmen:

$$p(y) = P(Y = y) = P(X = y + r) = \binom{y+r-1}{r-1} p^r (1-p)^y, \quad y \in \{0, 1, 2, \dots\}$$

²Manchmal auch als *Pascal–Verteilung* bezeichnet.

Abbildung 4.3: Binomialverteilung $B(10, 0.7)$ 

Ein Vorteil von letzterer Definition der NB–Verteilung besteht darin, dass – unabhängig vom Wert von r – der Merkmalraum der Verteilung stets gleich $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ ist. Wir verwenden im Folgenden aber die zuerst gegebene Definition.

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim NB(r, p)$ sind gegeben durch:

$$\mathbb{E}(X) = \frac{r}{p}, \quad \text{Var}(X) = \frac{r(1-p)}{p^2}$$

(Der Beweis ist nicht ganz einfach und wird hier nicht gegeben; vgl. für den Spezialfall $r = 1$ den folgenden Abschnitt.)

4.1.5 Geometrische Verteilung

Die Negative Binomialverteilung für $r = 1$ nennt man **Geometrische Verteilung** und schreibt $G(p)$ ($\equiv \text{NB}(1, p)$). Ist X die Gesamtzahl der Versuche, die notwendig sind, um exakt *einen* Erfolg zu bekommen, so gilt:

$$p(x) = P(X = x) = p(1 - p)^{x-1} \quad \text{für } x \in \{1, 2, \dots\}$$

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim G(p)$ sind gegeben durch:

$$\mathbb{E}(X) = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}$$

Beweis: Nur für den Erwartungswert:

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x=1}^{\infty} xp(1-p)^{x-1} \\ &= \sum_{x=1}^{\infty} (x-1+1)p(1-p)^{x-1} \\ &= \sum_{x=1}^{\infty} (x-1)p(1-p)^{x-1} + \underbrace{\sum_{x=1}^{\infty} p(1-p)^{x-1}}_{=1} \\ &= \sum_{x=0}^{\infty} xp(1-p)^x + 1 \\ &= (1-p) \underbrace{\sum_{x=1}^{\infty} xp(1-p)^{x-1}}_{=\mathbb{E}(X)} + 1 \\ &= (1-p)\mathbb{E}(X) + 1 \end{aligned}$$

Daraus folgt, dass $\mathbb{E}(X) = 1/p$. (Bem: Die Herleitung der Varianz erfolgt auf ähnliche Weise mittels Verschiebungssatz.)

Die Geometrische Verteilung hat eine besondere Eigenschaft (vgl. das folgende Bsp 4.4 für eine anschauliche Interpretation).

Gedächtnislosigkeit: Für $X \sim G(p)$ gilt:

$$P(X > a+b \mid X > a) = P(X > b) \quad \text{für } a, b \in \{1, 2, \dots\}$$

Beweis: Nach Definition der bedingten Wahrscheinlichkeit gilt:

$$P(X > a + b \mid X > a) = \frac{P(\{X > a + b\} \cap \{X > a\})}{P(X > a)} = \frac{P(X > a + b)}{P(X > a)}$$

Nun gilt $P(X > x) = (1 - p)^x$ für $x \in \{1, 2, \dots\}$; somit:

$$P(X > a + b \mid X > a) = \frac{(1 - p)^{a+b}}{(1 - p)^a} = (1 - p)^b = P(X > b)$$

Bsp 4.4 Angenommen, man nimmt wiederholt an einem (reinen) Glücksspiel teil, bei dem man mit Wahrscheinlichkeit $p = 1/10$ gewinnt. (Bem: Letzteres ist beispielsweise die Gewinnwahrscheinlichkeit beim *Joker*.) Ist X die Nummer der Runde des *ersten* Gewinns, so hat X eine $G(p)$ -Verteilung; daher gilt:

$$\mathbb{E}(X) = \frac{1}{1/10} = 10, \quad \text{Var}(X) = \frac{1 - 1/10}{(1/10)^2} = 90, \quad \sqrt{\text{Var}(X)} \approx 9.5$$

(Vgl. Abb 4.4 für eine grafische Darstellung der Verteilung.) Hat man nun bereits a Runden erfolglos gespielt, so besagt die „Gedächtnislosigkeit“ der $G(p)$ -Verteilung, dass die Wahrscheinlichkeit, bei der $(a + 1)$ -ten Runde zu gewinnen, genau gleich groß ist wie zu Beginn, d. h. unverändert gleich $1/10$. Anders ausgedrückt, es gibt keine „Prämie“ für erfolglose Spiele (etwa in Form einer höheren Gewinnwahrscheinlichkeit), das Spiel startet quasi nach jeder (erfolglosen) Runde von vorne.

Bem: Diese – eigentlich selbstverständliche – Eigenschaft von (reinen) Glücksspielen wird von einigen Spielern nicht verstanden, die der festen Überzeugung sind, dass mit der Zahl erfolgloser Spiele im Gegenzug die Gewinnwahrscheinlichkeit steigen *muss*. Das ist aber nicht der Fall, da die zugrunde liegende $G(p)$ -Verteilung eben kein „Gedächtnis“ hat.

■

Die Eigenschaft der Gedächtnislosigkeit charakterisiert die $G(p)$ -Verteilung.

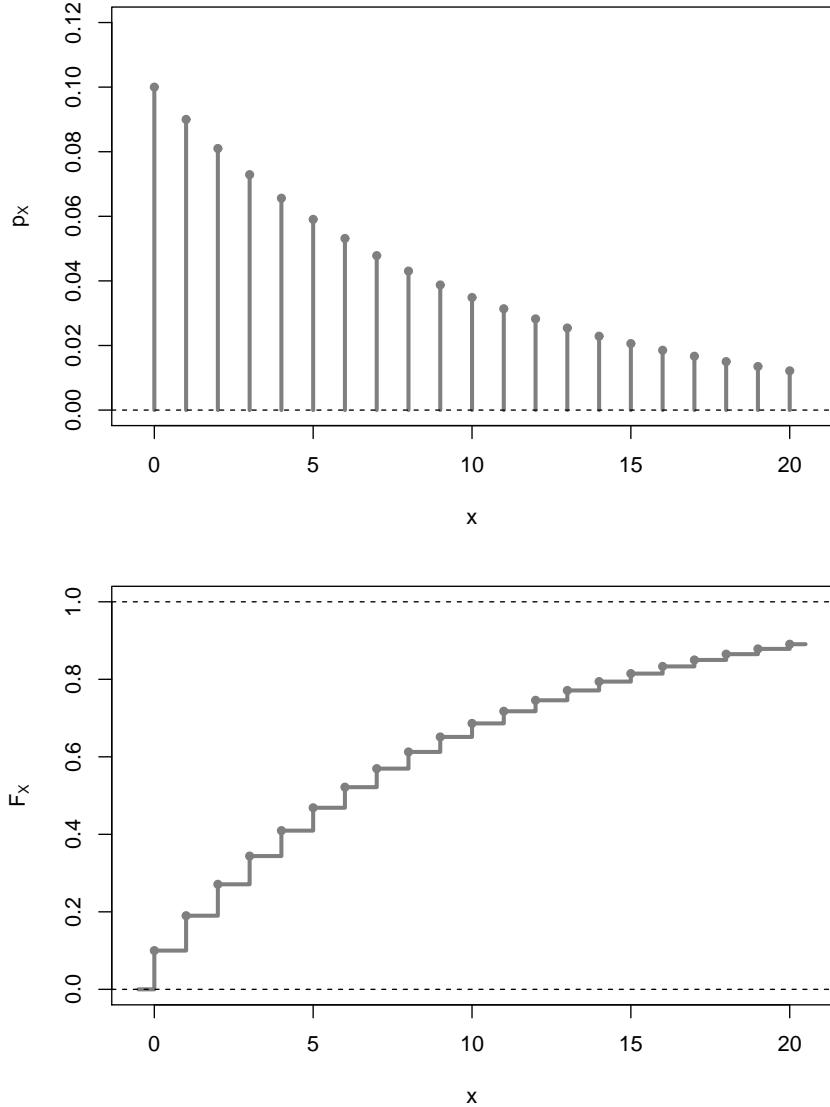
Behauptung: Die $G(p)$ -Verteilung ist die *einzige* diskrete Verteilung auf $\{1, 2, \dots\}$ ohne Gedächtnis.

Beweis: Die Gedächtnislosigkeit lässt sich auch wie folgt ausdrücken:

$$P(X > a + b) = P(X > a)P(X > b) \quad \text{für } a, b \in \mathbb{N}$$

Setzt man $p := P(X = 1)$, so gilt für eine gedächtnislose Verteilung:

$$\begin{aligned} P(X > 1) &= 1 - p \\ P(X > 2) &= P(X > 1)P(X > 1) = (1 - p)^2 \\ P(X > 3) &= P(X > 2)P(X > 1) = (1 - p)^2(1 - p) = (1 - p)^3 \\ &\vdots \end{aligned}$$

Abbildung 4.4: Geometrische Verteilung $G(0.1)$ 

Allgemein für $x \in \mathbb{N}$:

$$P(X > x) = (1 - p)^x$$

Damit folgt:

$$\begin{aligned} P(X = x) &= P(X > x - 1) - P(X > x) = (1 - p)^{x-1} - (1 - p)^x \\ &= (1 - p)^{x-1} [1 - (1 - p)] \\ &= p(1 - p)^{x-1} \end{aligned}$$

D. h., eine gedächtnislose Verteilung auf $\mathbb{N} = \{1, 2, \dots\}$ ist notwendigerweise eine $G(p)$ -Verteilung. Das war zu zeigen.

4.1.6 Hypergeometrische Verteilung

Viele praktische Situationen lassen sich durch ein Modell der folgenden Art beschreiben: In einem Behälter befinden sich (gut gemischt) N (gleichartige) Objekte; davon haben $A \leq N$ eine bestimmte Eigenschaft („Erfolge“) und entsprechend $N - A$ haben diese Eigenschaft nicht („Misserfolge“). Nun werden auf zufällige Weise $n \leq N$ Objekte **ohne Zurücklegen** entnommen (d. h., ein gezogenes Objekt wird nicht mehr in den Behälter zurückgelegt und kann kein weiteres Mal entnommen werden). Letztere Objekte bilden eine (einfache) **Stichprobe** der Größe (oder des Umfangs) n .

Bem: Üblicherweise stellt man sich vor, dass die Ziehungen *hintereinander* erfolgen. Man kann sich aber auch vorstellen, dass alle n Objekte der Stichprobe *zugleich* entnommen werden. In beiden Fällen ist aber die völlige Zufälligkeit der Stichprobenentnahme sicherzustellen.

Die sG X sei nun die Zahl der Erfolge in der Stichprobe. Mittels einer kombinatorischen Überlegung sieht man, dass die W-Funktion von X gegeben ist durch:

$$p(x) = P(X = x) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}}, \quad x \in \{ \max\{0, n+A-N\}, \dots, \min\{A, n\} \}$$

Eine Verteilung mit der obigen W-Funktion nennt man eine **Hypergeometrische Verteilung** und schreibt $X \sim H(N, A, n)$.

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim H(N, A, n)$ sind gegeben durch:

$$\mathbb{E}(X) = n \frac{A}{N}, \quad \text{Var}(X) = n \frac{A}{N} \left(1 - \frac{A}{N}\right) \frac{N-n}{N-1}$$

(Der Beweis ist etwas aufwendiger und wird hier nicht gegeben.)

Bem: Setzt man $p = A/N$, so ähneln die obigen Ausdrücke den entsprechenden Ausdrücken für die $B(n, p)$ -Verteilung. Das ist kein Zufall (s. unten). Den bei der Varianz hinzukommenden Faktor $(N-n)/(N-1)$ nennt man den **Korrekturfaktor für endliche Grundgesamtheiten**. Ist N sehr viel größer als n , ist der Korrekturfaktor annähernd gleich Eins.

Binomialapproximation: Ist N sehr viel größer als n , macht es keinen großen Unterschied, ob die Stichprobe auf Basis von Ziehungen **ohne** oder **mit Zurücklegen** zustande kommt.

In letzterem Fall handelt es sich aber um n unabhängige und identische Bernoulli-Experimente und die Zahl der Erfolge in der Stichprobe folgt einer $B(n, p)$ -Verteilung mit $p = A/N$. Unter bestimmten Umständen lässt sich also die $H(N, A, n)$ -Verteilung durch die (einfachere) $B(n, p = A/N)$ -Verteilung approximieren. Für die **Zulässigkeit** dieser Approximation gibt es mehrere „Faustregeln“; eine typische Regel lautet wie folgt:

Faustregel: Sind A und $N - n$ beide nicht zu klein und ist $n/N \leq 0.05$, so gilt in guter Näherung:

$$P(X = x) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}} \approx \binom{n}{x} \left(\frac{A}{N}\right)^x \left(1 - \frac{A}{N}\right)^{n-x}$$

Bsp 4.5 Anwendungen der Hypergeometrischen Verteilung finden sich beispielsweise in der *Qualitätskontrolle*. Angenommen, ein Los³ bestehend aus $N = 100$ (gleichartigen) Elementen (z. B. Glühlampen), soll auf seine Qualität geprüft werden. Dazu werden dem Los willkürlich $n = 22$ Elemente ohne Zurücklegen entnommen. Befinden sich darunter mehr als 2 defekte Elemente, wird das Los zurückgewiesen (an den Hersteller), andernfalls wird es akzeptiert.

Befinden sich im Los A defekte Einheiten, beträgt der *Defektanteil* $p = A/100$. Für Letzteren kommen die folgenden diskreten Werte in Frage:

$$p = 0, \frac{1}{100}, \frac{2}{100}, \dots, \frac{99}{100}, 1$$

Ist $X \sim H(N = 100, A = 100p, n = 22)$ die Zahl der defekten Elemente in der Stichprobe, so ist die Wahrscheinlichkeit, mit der das Los bei einem Defektanteil von p akzeptiert wird, gegeben durch:

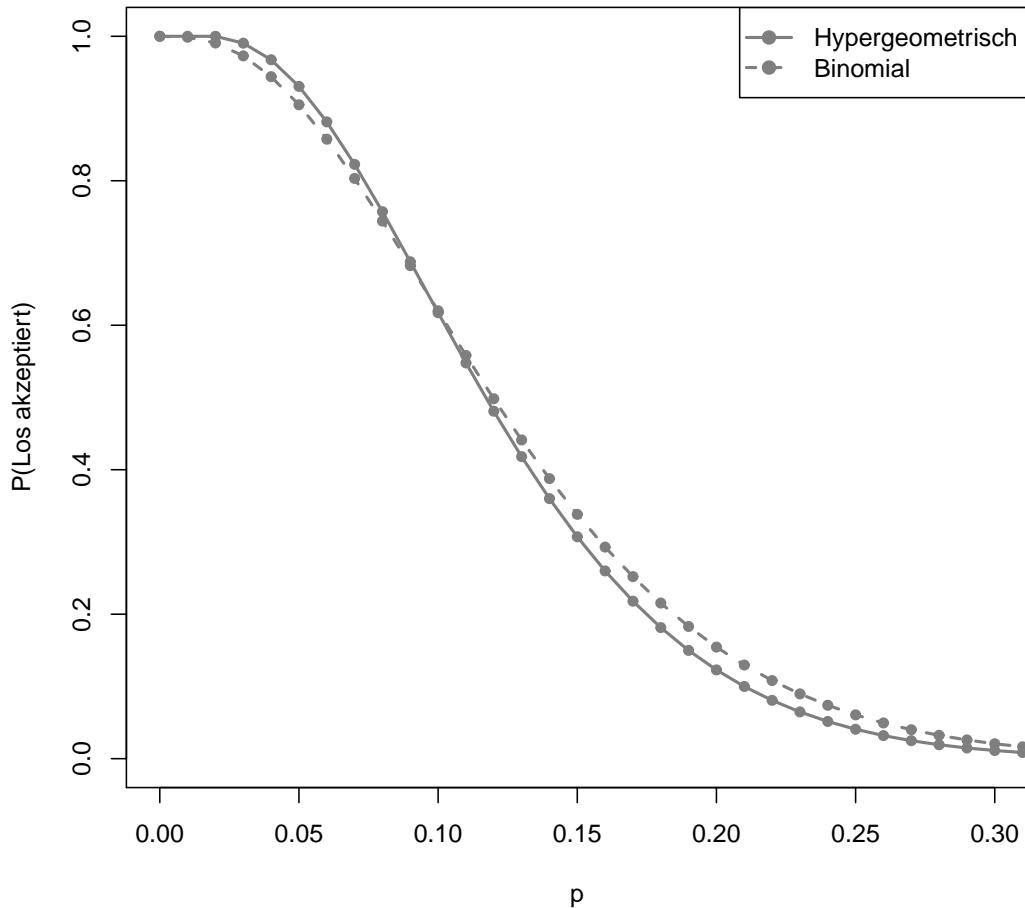
$$P(\text{Los akzeptiert}) = P(X \leq 2) = \sum_{x=0}^2 \frac{\binom{100p}{x} \binom{100-100p}{22-x}}{\binom{100}{22}}$$

Zum Vergleich betrachten wir auch die Berechnung über die Binomialapproximation der Hypergeometrischen Verteilung. Mit $Y \sim B(n = 22, p)$ gilt:

$$P(\text{Los akzeptiert}) \approx P(Y \leq 2) = \sum_{x=0}^2 \binom{22}{x} p^x (1-p)^{22-x}$$

³engl. *lot* oder *batch*

Abbildung 4.5: Annahmewahrscheinlichkeit (Bsp 4.5)



In Abb 4.5 ist die Annahmewahrscheinlichkeit des Loses in Abhängigkeit vom Defektanteil grafisch dargestellt. Auch wenn nach der Faustregel die Approximation hier nicht zulässig ist (die *Auswahlquote* $n/N = 22/100 = 0.22$ ist zu hoch), so ist dennoch die Binomialapproximation für praktische Zwecke ausreichend genau. (Bem: Bei Problemen der Qualitätskontrolle rechnet man meist mit der einfacheren Binomialverteilung.) ■

4.1.7 Poisson–Verteilung

Bekanntlich konvergiert die *Exponentialreihe* für alle $\lambda \in \mathbb{R}$:

$$1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$$

Für $\lambda > 0$ lässt sich also die W–Funktion einer sG X wie folgt definieren:

$$p(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{für } x \in \{0, 1, 2, \dots\}$$

Eine Verteilung mit der obigen W-Funktion nennt man eine **Poisson-Verteilung**⁴ und schreibt $X \sim P(\lambda)$.

Bem: Wie sich in der Praxis zeigt, lässt sich die Poisson-Verteilung in vielen Situationen mit befriedigenden Resultaten anwenden. So folgt beispielsweise die Anzahl X der von einer radioaktiven Substanz während einer bestimmten Zeitspanne emittierten α -Teilchen in guter Näherung einer $P(\lambda)$ -Verteilung. Weitere Anwendungen: Zahl der Lackierungsfehler auf einem Autoblech; Zahl der Verkehrsunfälle während einer bestimmten Zeitspanne; Zahl der Kunden, die im Laufe eines Tages ein Geschäft betreten; Zahl der Blitze während einer Minute bei einem Gewitter, etc.

Lässt sich ein (zufälliger) Prozess durch eine Poisson-Verteilung beschreiben, spricht man von einem **Poisson-Prozess**.

Bedingungen für einen Poisson-Prozess: Bezeichnet $p(x, w)$ die Wahrscheinlichkeit von x „Vorkommnissen“ in einem Intervall⁵ der Länge w , so lauten (hinreichende) Bedingungen für das Vorliegen eines Poisson-Prozesses wie folgt:

(1) **Proportionalität im Kleinen:** $p(1, h) = \lambda h + o(h)$ für $\lambda > 0$ (Konstante) und $h > 0$.

Bem: Das aus der Mathematik bekannte *Landau-Symbol* $o(h)$ (lies: „klein o von h “) bedeutet hier eine Funktion mit $\lim_{h \rightarrow 0} [o(h)/h] = 0$.

(2) **Nachwirkungsfreiheit:** $\sum_{x=2}^{\infty} p(x, h) = o(h)$.

(D. h., für kleine Intervalle kann die Wahrscheinlichkeit des Auftretens von zwei oder mehr Vorkommnissen vernachlässigt werden.)

(3) **Unabhängigkeit:** Die Anzahlen von Vorkommnissen in nicht überlappenden Intervallen sind unabhängig.

Sind die obigen Bedingungen erfüllt, so kann man zeigen, dass die Zahl X der Vorkommnisse in einem Intervall der Länge w einer $P(\lambda w)$ -Verteilung folgt.

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim P(\lambda)$ sind gegeben durch:

$$\mathbb{E}(X) = \lambda, \quad \text{Var}(X) = \lambda$$

⁴SIMÉON DENIS POISSON (1781–1840), franz. Physiker und Mathematiker.

⁵Kann ein zeitliches oder räumliches Intervall sein.

Beweis: Nur für den Erwartungswert (Herleitung der Varianz als UE–Aufgabe):

$$\mathbb{E}(X) = \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} \underbrace{\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}}_{=e^{\lambda}} = \lambda$$

Modalwert(e): Die Modalwerte der $P(\lambda)$ –Verteilung sind gegeben durch:

$$x_{\text{mod}} = \begin{cases} \lfloor \lambda \rfloor & \text{falls } \lambda \notin \mathbb{N} \\ \lambda - 1, \lambda & \text{falls } \lambda \in \mathbb{N} \end{cases}$$

Man beachte, dass es für $\lambda \in \mathbb{N}$ zwei Modalwerte gibt.

(Beweis als UE–Aufgabe.)

Poisson–Verteilung als Grenzfall der Binomialverteilung: Für $X_n \sim B(n, \lambda/n)$ gilt:

$$\lim_{n \rightarrow \infty} P(X_n = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Beweis: Für festes $x \in \{0, 1, 2, \dots, n\}$ gilt:

$$\begin{aligned} P(X_n = x) &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!} \underbrace{\frac{n \times (n-1) \times \cdots \times (n-x+1)}{n \times n \times \cdots \times n}}_{\rightarrow 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\rightarrow e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\rightarrow 1} \\ &\longrightarrow \frac{\lambda^x e^{-\lambda}}{x!} \end{aligned}$$

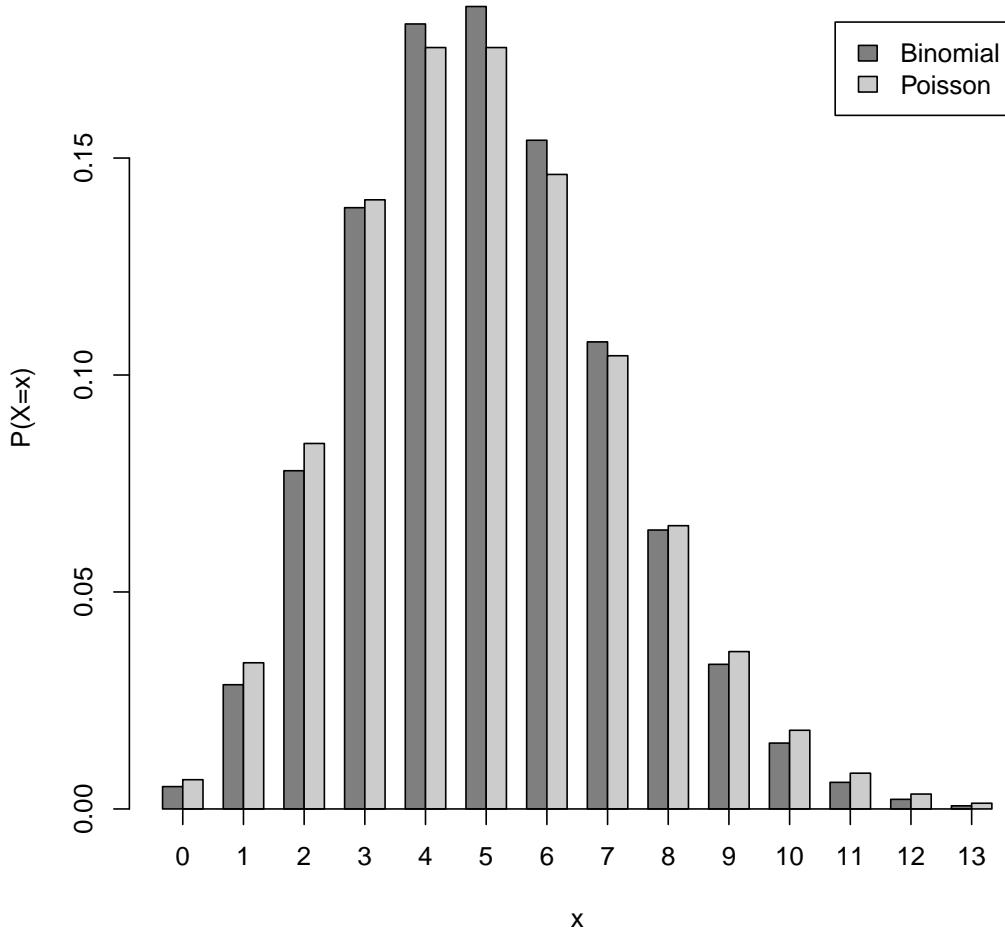
Bem: Die Poisson–Verteilung lässt sich also wie folgt interpretieren: Gibt es *viele* unabhängige und identische Bernoulli–Experimente mit *kleiner* Erfolgswahrscheinlichkeit, so folgt die Zahl der Erfolge in guter Näherung einer Poisson–Verteilung. Aus diesem Grund nennt man die Poisson–Verteilung manchmal auch die „Verteilung der seltenen Ereignisse“.

Eine gängige Regel für die **Zulässigkeit** der Approximation einer $B(n, p)$ –Verteilung durch eine $P(\lambda = np)$ –Verteilung lautet wie folgt:

Faustregel: Für $n \geq 50$, $p \leq 1/10$ und $np \leq 10$ gilt in guter Näherung für $X \sim B(n, p)$:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{(np)^x e^{-np}}{x!}$$

Abbildung 4.6: Binomial– und approximierende Poissonverteilung (Bsp 4.6)



Bsp 4.6 Die sG X habe eine $B(50, 1/10)$ -Verteilung und Y habe die approximierende $P(5)$ -Verteilung. Abb 4.6 zeigt einen grafischen Vergleich der beiden W-Funktionen. (Bem: Nur Wahrscheinlichkeiten größer als 0.001 werden dargestellt.) Die Bedingungen der obigen Faustregel sind hier (gerade noch) erfüllt. Wegen $(n+1)p = 5.1$ hat die Binomialverteilung einen eindeutig bestimmten Modalwert (bei $x = 5$) und wegen $\lambda = np = 5$ hat die Poissonverteilung zwei Modalwerte (bei $x = 4$ und $x = 5$). ■

4.2 Stetige Verteilungen

Im vorliegenden Abschnitt werden eine Reihe von wichtigen *stetigen* Verteilungen definiert und ihre Eigenschaften diskutiert. Wie auch im diskreten Fall hängen diese Verteilungen von (einem oder mehreren) **Parametern** ab, sodass man genauer von **Verteilungsfamilien** sprechen kann. Beispielsweise lässt sich die Familie der Normalverteilungen (vgl. 4.2.4) wie folgt schreiben:

$$\mathcal{F}_N = \{\mathsf{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 > 0\}$$

μ und σ^2 (bzw. σ) sind die Parameter der Verteilungsfamilie.

4.2.1 Stetige uniforme Verteilung

Eine sG X hat eine (stetige) **uniforme Verteilung** (oder eine (stetige) **Gleichverteilung**) auf dem Intervall (a, b) ($a < b, a, b \in \mathbb{R}$), wenn die Dichte von X wie folgt lautet:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } x \in (a, b) \\ 0 & \text{für } x \notin (a, b) \end{cases}$$

Man schreibt $X \sim \mathsf{U}(a, b)$ oder $X \sim \mathsf{U}[a, b]$ (falls die Randpunkte zum Träger gehören).

Bem: Man beachte, dass es für (Wahrscheinlichkeits-) Berechnungen keine Rolle spielt, ob man das offene (a, b) oder das abgeschlossene Intervall $[a, b]$ (oder ein halboffenes Intervall) zugrunde legt.

Die Verteilungsfunktion von $X \sim \mathsf{U}(a, b)$ ist gegeben durch:

$$F(x) = \begin{cases} 0 & \text{für } x \leq a \\ \frac{x-a}{b-a} & \text{für } a < x < b \\ 1 & \text{für } x \geq b \end{cases}$$

Abb 4.7 zeigt eine grafische Darstellung von VF und Dichte.

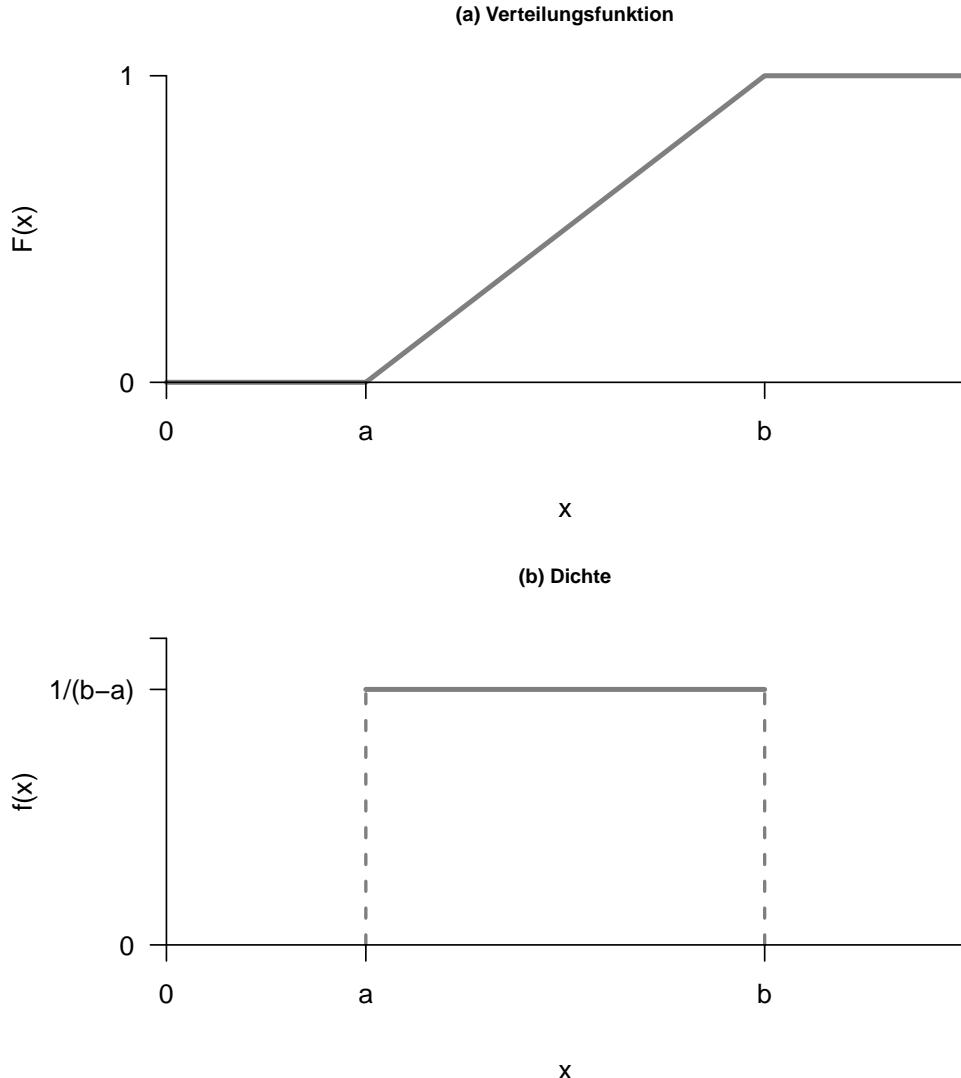
Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim \mathsf{U}(a, b)$ sind gegeben durch:

$$\mathbb{E}(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

Beweis: Aus Gründen der Symmetrie ist der Erwartungswert der Mittelpunkt des Intervalls:

$$\mathbb{E}(X) = \int_a^b \frac{x}{b-a} dx = \frac{x^2}{2(b-a)} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

Abbildung 4.7: Stetige uniforme Verteilung



$$\begin{aligned}\text{Var}(X) &= \mathbb{E}(X^2) - \mathbb{E}^2(X) = \int_a^b \frac{x^2}{b-a} dx - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{b^3 - a^3}{3(b-a)} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}\end{aligned}$$

4.2.2 Exponentialverteilung

Die Geometrische Verteilung $G(p)$ (vgl. 4.1.5) lässt sich als (diskrete) *Wartezeitverteilung* (= Zahl der Versuche bis zum *ersten* Erfolg) interpretieren. Eine stetige Version der $G(p)$ -Verteilung ist die Exponentialverteilung.

Eine sG X hat eine **Exponentialverteilung** mit dem **Skalierungsparameter** $\tau > 0$, wenn ihre Dichte gegeben ist durch:

$$f(x) = \frac{1}{\tau} e^{-x/\tau} \quad \text{für } x \geq 0$$

Setzt man $\lambda = 1/\tau$, lautet die Dichte wie folgt:

$$f(x) = \lambda e^{-\lambda x} \quad \text{für } x \geq 0$$

Man schreibt $X \sim \text{Exp}(\lambda)$ (oder $X \sim \text{Exp}(\tau)$).

Verteilungsfunktion: Die Verteilungsfunktion von $X \sim \text{Exp}(\lambda)$ ist gegeben durch:

$$F(x) = 1 - e^{-\lambda x} \quad \text{für } x \geq 0$$

Beweis:

$$F(x) = P(X \leq x) = \int_0^x \lambda e^{-\lambda u} du = -e^{-\lambda u} \Big|_0^x = 1 - e^{-\lambda x}, \quad x \geq 0$$

Abb 4.8 zeigt die Verteilungsfunktion und die Dichte für $\tau = 1/2, 1, 2$ (bzw. $\lambda = 2, 1, 1/2$). Man beachte, dass die Dichte die y -Achse bei $\lambda = 1/\tau$ schneidet.

Erwartungswert/Varianz/Streuung: Der Erwartungswert, die Varianz und die Streuung von $X \sim \text{Exp}(\lambda)$ sind gegeben durch:

$$\mathbb{E}(X) = \frac{1}{\lambda} = \tau, \quad \text{Var}(X) = \frac{1}{\lambda^2} = \tau^2, \quad \sqrt{\text{Var}(X)} = \frac{1}{\lambda} = \tau$$

Beweis: Partielle Integration:

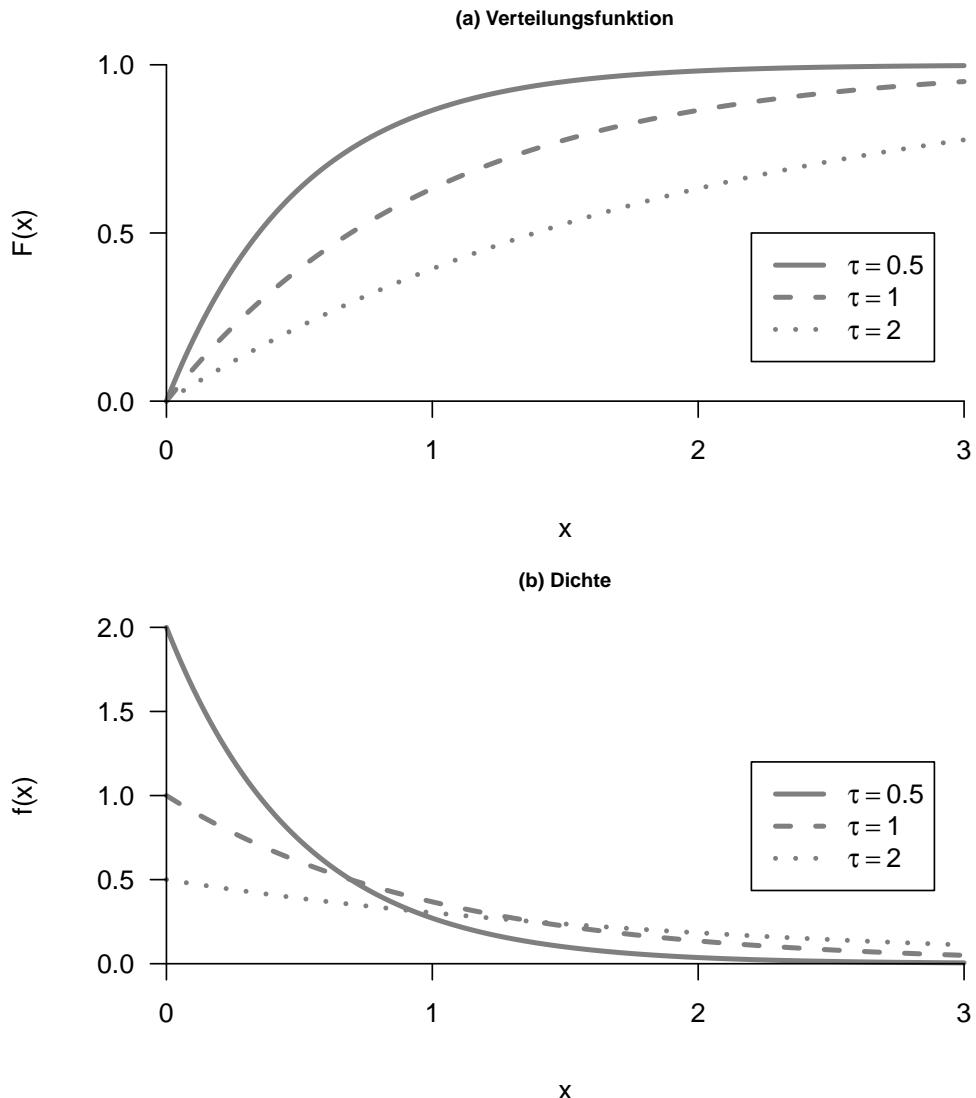
$$\mathbb{E}(X) = \int_0^\infty x \lambda e^{-\lambda x} dx = \underbrace{-x e^{-\lambda x}}_{=0} \Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda} \underbrace{\int_0^\infty \lambda e^{-\lambda x} dx}_{=1} = \frac{1}{\lambda}$$

Analog bestimmt man die Varianz (UE-Aufgabe).

Wie die $G(p)$ -Verteilung hat auch die $\text{Exp}(\lambda)$ -Verteilung kein „Gedächtnis“.

Gedächtnislosigkeit: Für $X \sim \text{Exp}(\lambda)$ gilt:

$$P(X > s + t \mid X > s) = P(X > t) \quad \text{für } s, t > 0$$

Abbildung 4.8: Exponentialverteilung

Beweis: Nach Definition der bedingten Wahrscheinlichkeit gilt:

$$P(X > s + t \mid X > s) = \frac{P(X > s + t, X > s)}{P(X > s)} = \frac{P(X > s + t)}{P(X > s)}$$

Mit:

$$P(X > x) = 1 - F(x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}, \quad x \geq 0$$

folgt:

$$\frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t)$$

Interpretation: Die *Memoryless Property* lässt sich als „Nicht–Alterung“ interpretieren. Ist beispielsweise X die exponentialverteilte Lebensdauer einer Komponente, und ist die Komponente zum Zeitpunkt s noch intakt, so hat die *restliche Lebensdauer* der Komponente, d. h. $X - s$, die gleiche Exponentialverteilung wie eine komplett neue Komponente. M. a. W., die Komponente „erinnert“ sich nicht an ihr Alter und ist zu jedem Zeitpunkt, zu dem sie noch intakt ist, so gut wie neu.

Die Eigenschaft der Gedächtnislosigkeit charakterisiert die Exponentialverteilung.

Behauptung: Die $\text{Exp}(\lambda)$ –Verteilung ist die *einzige* stetige Verteilung auf $[0, \infty)$ ohne Gedächtnis.

Beweis: Sei X eine sG auf $[0, \infty)$ mit dieser Eigenschaft. Mit $G(x) := P(X > x)$ gilt dann:

$$G(x+y) = G(x)G(y), \quad x, y \in [0, \infty)$$

Für $a \in \mathbb{N}$ folgt daraus:

$$G(a) = G\left(\sum_{i=1}^a 1\right) = G(1)^a$$

Weiter gilt für $b \in \mathbb{N}$:

$$G(1) = G\left(\sum_{i=1}^b \frac{1}{b}\right) = G\left(\frac{1}{b}\right)^b \implies G\left(\frac{1}{b}\right) = G(1)^{1/b}$$

Für rationale Zahlen $q = a/b$ gilt daher:

$$G(q) = G\left(\frac{a}{b}\right) = G\left(\sum_{i=1}^a \frac{1}{b}\right) = G\left(\frac{1}{b}\right)^a = G(1)^{a/b} = G(1)^q$$

Jede reelle Zahl $x > 0$ kann aber von rechts durch rationale Zahlen $q_n > 0$ angenähert werden: $q_n \rightarrow x$. Wegen der Rechtsstetigkeit von $G(x) = 1 - F(x)$ folgt daher:

$$G(x) = \lim_{n \rightarrow \infty} G(q_n) = \lim_{n \rightarrow \infty} G(1)^{q_n} = G(1)^x$$

Setzt man $\lambda := -\ln G(1)$ (d. h. $G(1) = e^{-\lambda}$), so gilt:

$$G(x) = P(X > x) = e^{-\lambda x}$$

D. h., X ist exponentialverteilt:

$$F(x) = 1 - G(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

Das war zu zeigen.

4.2.3 Gamma- und Chiadratverteilung

Die Gammaverteilung ist eine Verallgemeinerung der Exponentialverteilung. Eine SG X hat eine **Gammaverteilung** mit dem **Formparameter** $\alpha > 0$ und dem **Skalierungsparameter** $\beta > 0$, wenn ihre Dichte gegeben ist durch:

$$f(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} \quad \text{für } x > 0$$

Setzt man $\lambda = 1/\beta$, lautet die Dichte wie folgt:

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \quad \text{für } x > 0$$

Man schreibt $X \sim \text{Gam}(\alpha, \lambda)$ (oder $X \sim \text{Gam}(\alpha, \beta)$).

Die **Gammafunktion** Γ ist eine wichtige Funktion in der Mathematik, definiert durch:

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du \quad \text{für } \alpha > 0$$

Sie hat die folgenden Eigenschaften:

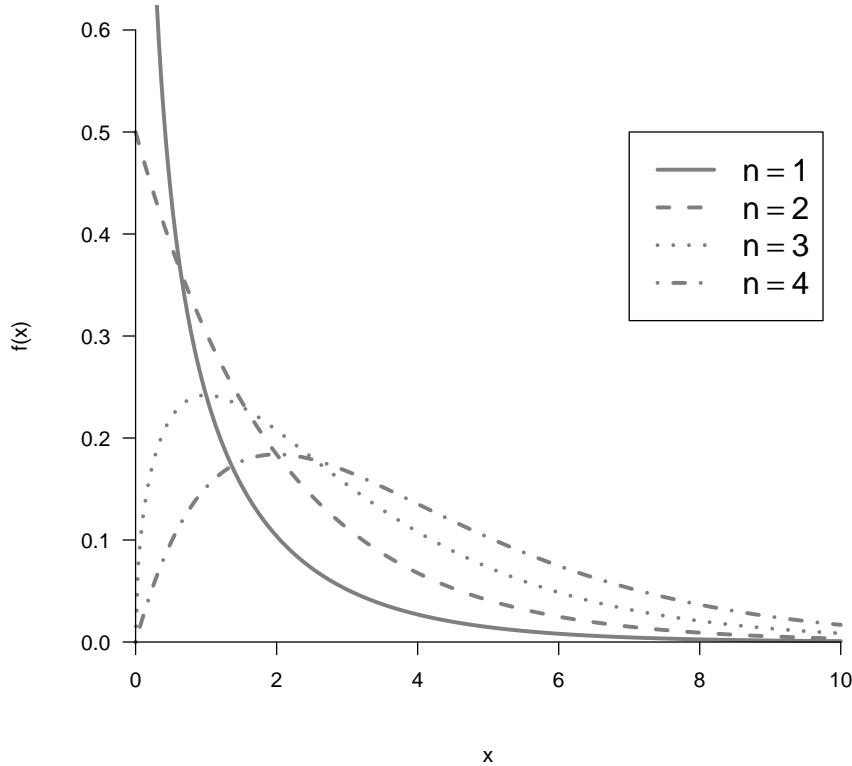
- (1) $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$ für $\alpha \in (0, \infty)$
- (2) $\Gamma(n) = (n - 1)!$ für $n = 1, 2, \dots$
- (3) $\Gamma(1/2) = \sqrt{\pi}$

An den Eigenschaften (1) und (2) erkennt man, dass die Gammafunktion eine Verallgemeinerung der Fakultät auf positive reelle Zahlen ist.⁶

Spezialfälle: Zwei Spezialfälle der Gammaverteilung sind von besonderer Bedeutung:

- (1) Für $\alpha = 1$ ergibt sich die Exponentialverteilung $\text{Exp}(\lambda)$ (oder $\text{Exp}(\beta)$).
- (2) Für $\alpha = n/2$ (mit $n \in \mathbb{N}$) und $\beta = 2$ (oder $\lambda = 1/2$) ergibt sich die **Chiadratverteilung** mit n **Freiheitsgraden** (vgl. Abb 4.9). Man schreibt $X \sim \chi^2(n)$ (oder $X \sim \chi_n^2$). Die Quantile der $\chi^2(n)$ -Verteilung werden (meist) mit $\chi_{n,p}^2$ bezeichnet und sind ausführlich tabelliert (vgl. Anhang: Tabellen).

⁶Aus diesem Grund schreibt man manchmal auch $\Gamma(x) = (x - 1)!$ für $x \in \mathbb{R}^+$.

Abbildung 4.9: Dichte der $\chi^2(n)$ -Verteilung

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim \text{Gam}(\alpha, \lambda)$ sind gegeben durch:

$$\mathbb{E}(X) = \frac{\alpha}{\lambda} = \alpha\beta, \quad \text{Var}(X) = \frac{\alpha}{\lambda^2} = \alpha\beta^2$$

Insbesondere gilt für $X \sim \chi^2(n)$:

$$\mathbb{E}(X) = n, \quad \text{Var}(X) = 2n$$

Beweis: Den Erwartungswert berechnet man wie folgt:

$$\mathbb{E}(X) = \int_0^\infty x \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} dx = \int_0^\infty \frac{\lambda^\alpha x^\alpha e^{-\lambda x}}{\Gamma(\alpha)} dx = \underbrace{\frac{\Gamma(\alpha+1)}{\lambda \Gamma(\alpha)}}_{=\alpha/\lambda} \underbrace{\int_0^\infty \frac{\lambda^{\alpha+1} x^\alpha e^{-\lambda x}}{\Gamma(\alpha+1)} dx}_{=1 \dots \text{Gam}(\alpha+1, \lambda)} = \frac{\alpha}{\lambda}$$

Auf analoge Weise zeigt man $\mathbb{E}(X^2) = (\alpha + 1)\alpha/\lambda^2$ (UE-Aufgabe) und mittels Verschiebungssatz:

$$\text{Var}(X) = \frac{(\alpha + 1)\alpha}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2}$$

4.2.4 Normalverteilung

Die in diesem Abschnitt behandelte Verteilung gehört zu den wichtigsten Verteilungen in Statistik und Wahrscheinlichkeitstheorie.

Eine sG X hat eine **Normalverteilung** (auch **Gauß-Verteilung**⁷) mit dem **Lageparameter** $\mu \in \mathbb{R}$ und dem **Skalierungsparameter** $\sigma > 0$, wenn ihre Dichte („Glockenkurve“) gegeben ist durch:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \quad \text{für } -\infty < x < \infty$$

Man schreibt $X \sim N(\mu, \sigma^2)$.⁸ Die Verteilung mit $\mu = 0$ und $\sigma = 1$ nennt man die **Standardnormalverteilung** $N(0, 1)$. Die Dichte von Letzterer wird üblicherweise mit φ bezeichnet und ist gegeben durch:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{für } -\infty < x < \infty$$

Für die Verteilungsfunktion der $N(\mu, \sigma^2)$ gibt es keinen expliziten Ausdruck; sie lässt sich allerdings mittels **Standardisierung** auf die VF der $N(0, 1)$ zurückführen. Letztere wird üblicherweise mit Φ bezeichnet und ist ausführlich tabelliert (vgl. Anhang: Tabellen).

Behauptung: Für die VF Φ der Standardnormalverteilung $N(0, 1)$ gilt:

$$\Phi(-x) = 1 - \Phi(x) \quad \text{für } -\infty < x < \infty$$

Beweis: Folgt unmittelbar aus der Symmetrie der Standardnormalverteilung um Null.

Standardisierung: Gilt $X \sim N(\mu, \sigma^2)$ und ist $Z = (X - \mu)/\sigma$ die **standardisierte** sG, so hat Z eine Standardnormalverteilung: $Z \sim N(0, 1)$.

Beweis: Die VF von Z ist gegeben durch:

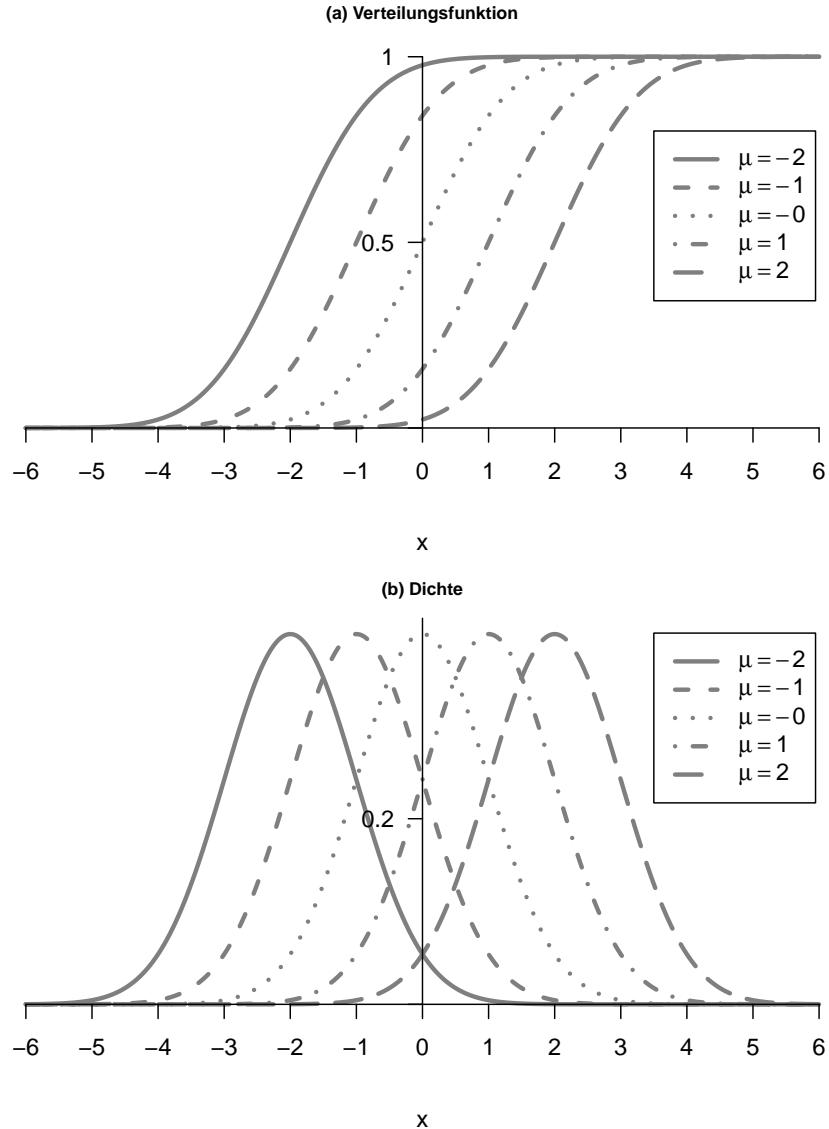
$$F_Z(z) = P(Z \leq z) = P \left(\frac{X - \mu}{\sigma} \leq z \right) = P(X \leq \mu + \sigma z)$$

Mittels der Variablensubstitution $y = (x - \mu)/\sigma$ ($\rightarrow dx = \sigma dy$) bekommt man:

$$P(X \leq \mu + \sigma z) = \int_{-\infty}^{\mu + \sigma z} \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] dx = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \Phi(z)$$

⁷JOHANN CARL FRIEDRICH GAUSS (1777–1855), dt. Mathematiker, Astronom, Geodät und Physiker; genannt *Princeps mathematicorum*.

⁸Manchmal schreibt man auch $X \sim N(\mu, \sigma)$; hier wird aber stets die erste Schreibweise verwendet.

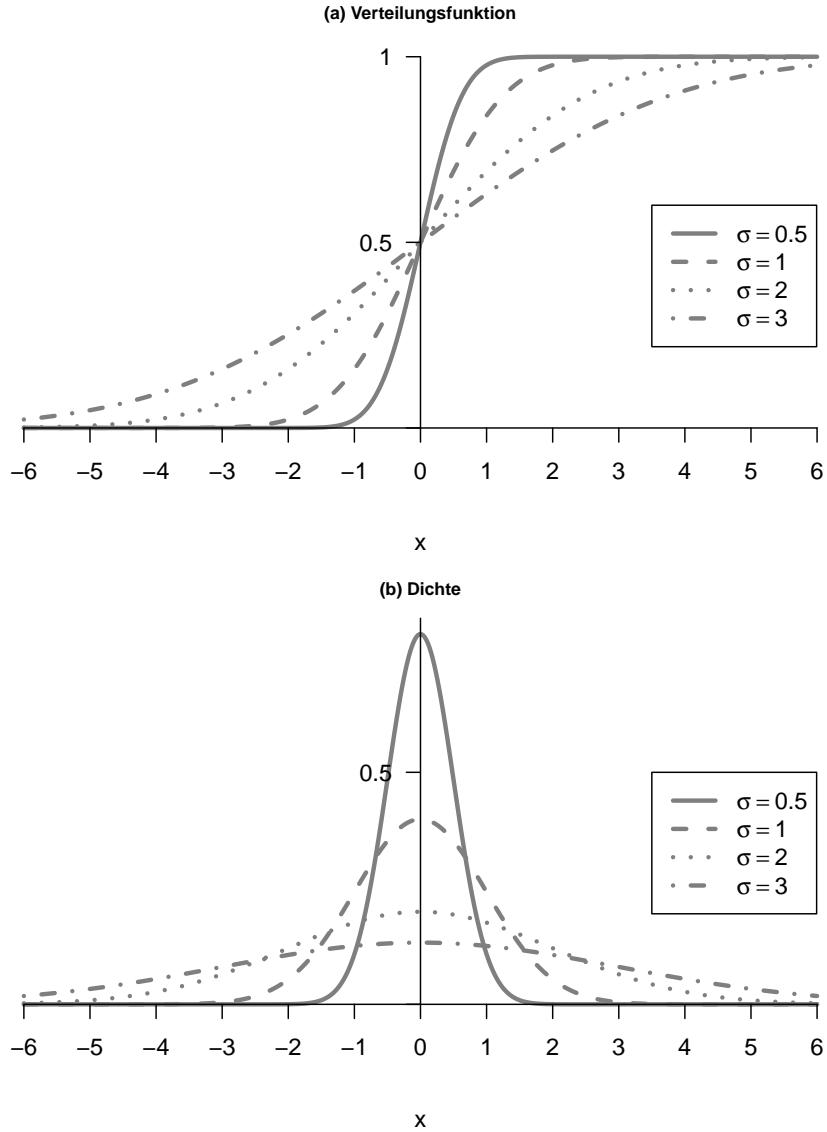
Abbildung 4.10: Normalverteilung ($\sigma = 1$)

Daraus folgt, dass $X \sim N(\mu, \sigma^2)$ als **affine Transformation** (vgl. 3.3.2) von $Z \sim N(0, 1)$ dargestellt werden kann:

$$X = \mu + \sigma Z \quad \text{mit} \quad Z \sim N(0, 1)$$

Für die VF F und die Dichte f von $X \sim N(\mu, \sigma^2)$ gilt:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad f(x) = F'(x) = \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}\right)$$

Abbildung 4.11: Normalverteilung ($\mu = 0$)

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim N(\mu, \sigma^2)$ sind gegeben durch:

$$\mathbb{E}(X) = \mu, \quad \text{Var}(X) = \sigma^2$$

Beweis: Da die Dichte symmetrisch um μ ist, und – wie man zeigen kann – der Erwartungswert von X auch existiert, gilt:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] dx = \mu$$

Um zu zeigen, dass die Varianz von X gleich σ^2 ist, schreiben wir zunächst $X = \mu + \sigma Z$, wobei $Z \sim N(0, 1)$. Daraus folgt, dass $\text{Var}(X) = \sigma^2 \text{Var}(Z)$. D. h., es genügt zu zeigen, dass $\text{Var}(Z) = 1$. Wegen $E(Z) = 0$ gilt:

$$\text{Var}(Z) = E(Z^2) = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \int_{-\infty}^{\infty} \underbrace{z}_u \times \underbrace{\frac{z}{\sqrt{2\pi}} e^{-z^2/2}}_{v'} dz$$

Partielle Integration ergibt:

$$E(Z^2) = \underbrace{-\frac{z}{\sqrt{2\pi}} e^{-z^2/2}}_{=0} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \int_{-\infty}^{\infty} \varphi(z) dz = 1$$

Quantile: Zwischen dem p -Quantil x_p von $N(\mu, \sigma^2)$ und dem p -Quantil z_p (manchmal auch mit u_p bezeichnet) von $N(0, 1)$ besteht die folgende Beziehung:

$$x_p = \mu + \sigma z_p \quad \text{für } 0 < p < 1$$

Die Quantile der $N(0, 1)$ -Verteilung sind für $p \geq 0.5$ ausführlich tabelliert (vgl. Anhang: Tabellen). Für $p < 0.5$ benutzt man die für alle $p \in (0, 1)$ gültige Beziehung:

$$z_p = -z_{1-p}$$

Beweis: Der erste Teil folgt aus der Standardisierung von X :

$$p = P(Z \leq z_p) = P\left(\frac{X - \mu}{\sigma} \leq z_p\right) = P(X \leq \underbrace{\mu + \sigma z_p}_{= x_p})$$

Der zweite Teil folgt aus der Symmetrie der $N(0, 1)$ -Verteilung um Null.

$$\underline{\text{Beziehung zur Chiquadratverteilung}}: X \sim N(\mu, \sigma^2) \implies \left(\frac{X - \mu}{\sigma}\right)^2 \sim \chi^2(1)$$

Beweis: Zunächst gilt $Z = (X - \mu)/\sigma \sim N(0, 1)$ (Standardisierung). Da die Funktion $y = z^2$ auf \mathbb{R} nicht umkehrbar eindeutig ist, benutzen wir für den weiteren Beweis die Methode der Verteilungsfunktion (vgl. 3.3.2). Die VF von $Y = Z^2$ ist gegeben durch:

$$F_Y(y) = P(Z^2 \leq y) = P(-\sqrt{y} \leq Z \leq \sqrt{y}) = 2\Phi(\sqrt{y}) - 1$$

Letzteres gilt wegen $\Phi(-\sqrt{y}) = 1 - \Phi(\sqrt{y})$. Die Dichte von Y bekommt man durch Ableiten:

$$f_Y(y) = F'_Y(y) = 2\varphi(\sqrt{y}) \frac{1}{2\sqrt{y}} = \frac{y^{1/2-1} e^{-y/2}}{\sqrt{2\pi}}, \quad y > 0$$

Wegen $\Gamma(1/2) = \sqrt{\pi}$ handelt es sich um die Dichte einer $\chi^2(1)$ -Verteilung.

Vgl. Abb 4.10 und 4.11 für grafische Darstellungen der Verteilungsfunktion und Dichte der $N(\mu, \sigma^2)$ -Verteilung für einige Werte von μ und σ^2 .

4.2.5 F-Verteilung

Die Verteilungen dieses und des folgenden Abschnitts spielen eine große Rolle in der (klassischen) Statistik.

Eine sG X hat eine **F-Verteilung**⁹ mit m und n **Freiheitsgraden**, wenn ihre Dichte gegeben ist durch:

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right) \left(\frac{m}{n}\right)^{m/2} x^{(m-2)/2}}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) \left[1 + \left(\frac{m}{n}\right)x\right]^{(m+n)/2}} \quad \text{für } x \geq 0$$

Dabei ist Γ die in 4.2.3 definierte Gammafunktion. Man schreibt $X \sim F(m, n)$ (oder $X \sim F_{m,n}$). Vgl. Abb 4.12 für grafische Darstellungen der Dichte für einige (m, n) -Kombinationen.

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim F(m, n)$ sind gegeben durch:

$$\mathbb{E}(X) = \frac{n}{n-2} \quad (n > 2), \quad \text{Var}(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad (n > 4)$$

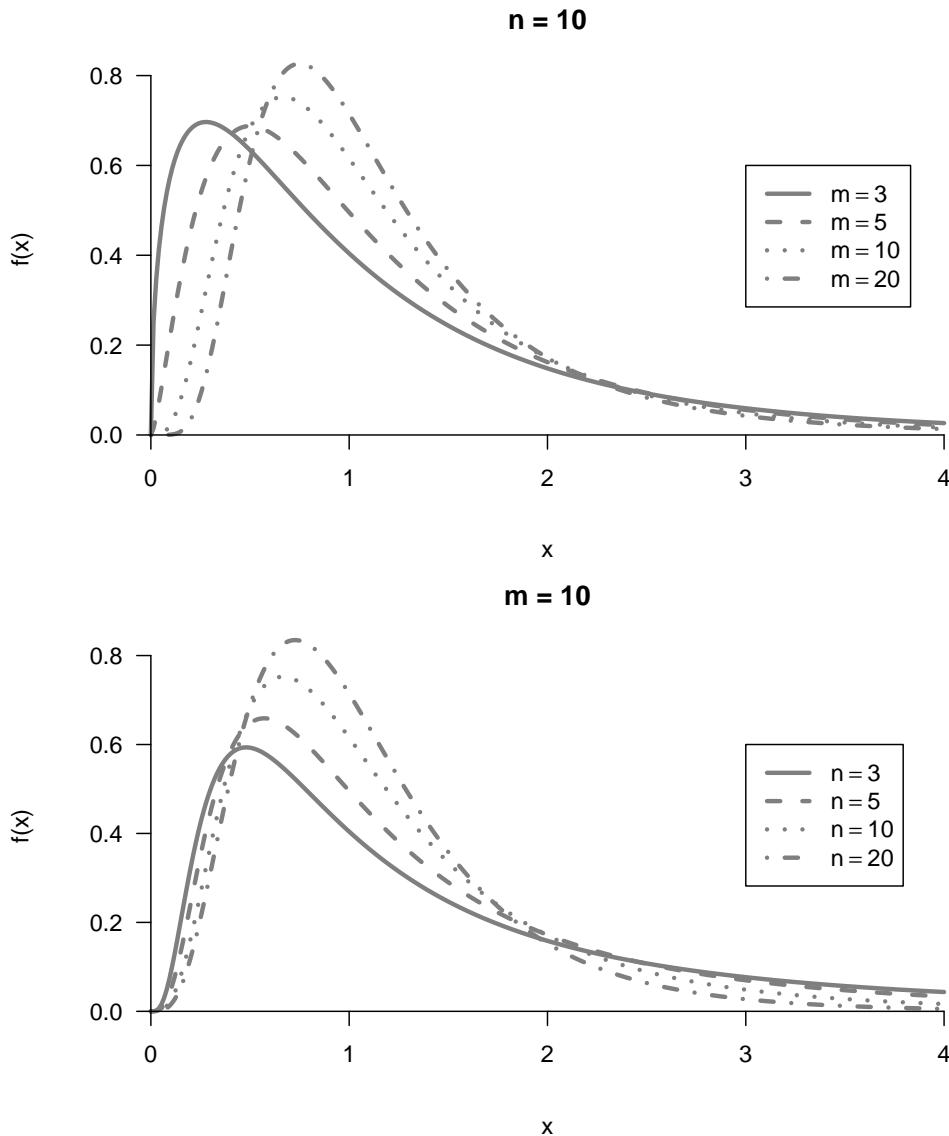
(Der Beweis ist etwas aufwendiger und wird hier nicht gegeben.)

Symmetrie: Es gilt die folgende Symmetriebeziehung:

$$X \sim F(m, n) \iff \frac{1}{X} \sim F(n, m)$$

Beweis: Mittels Transformationssatz für Dichten (vgl. 3.3.2).

⁹Benannt nach (SIR) RONALD AYLMER FISHER (1890–1962), engl. Biologe, Genetiker und einer der bedeutendsten Statistiker des 20. Jh.

Abbildung 4.12: Dichte der $F(m, n)$ -Verteilung

Quantile: Die Quantile der $F(m, n)$ -Verteilung werden (meist) mit $F_{m,n; p}$ bezeichnet und sind für $p \geq 0.5$ ausführlich tabelliert (vgl. Anhang: Tabellen). Für $p < 0.5$ benützt man die aus der obigen Symmetrie folgende für alle $p \in (0, 1)$ gültige Beziehung:

$$F_{m,n; p} = \frac{1}{F_{n,m; 1-p}}$$

Bem: Die F -Verteilung spielt speziell in der Regressionsanalyse (vgl. Kapitel 9) und in der Varianzanalyse¹⁰ (wird in diesem Text nicht behandelt) eine große Rolle.

¹⁰Vgl. DALGAARD (2008) oder VERZANI (2014).

4.2.6 t–Verteilung

Eine sG X hat eine **t–Verteilung** (oder **Student–Verteilung**¹¹) mit n **Freiheitsgraden**, wenn ihre Dichte gegeben ist durch:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} \quad \text{für } -\infty < x < \infty$$

Dabei ist Γ die in 4.2.3 definierte Gammafunktion. Man schreibt $X \sim t(n)$ (oder $X \sim t_n$). Die Dichte ist symmetrisch (um Null) und konvergiert für wachsende Freiheitsgrade gegen die Dichte der Standardnormalverteilung:

$$\lim_{n \rightarrow \infty} f(x) = \varphi(x) \quad \text{für } x \in \mathbb{R}$$

Man beachte aber, dass alle t–Dichten **schwerere Ausläufer** als die Normaldichte haben (Abb 4.13). Die $t(1)$ –Verteilung nennt man auch **Cauchy–Verteilung**¹² und schreibt $C(0, 1)$. Die Dichte der Cauchy–Verteilung lautet wie folgt:

$$f(x) = \frac{1}{\pi(1+x^2)} \quad \text{für } -\infty < x < \infty$$

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim t(n)$ sind gegeben durch:

$$\mathbb{E}(X) = 0 \quad (n > 1), \quad \text{Var}(X) = \frac{n}{n-2} \quad (n > 2)$$

(Der Beweis ist etwas aufwendiger und wird hier nicht gegeben.)

Quantile: Die Quantile der $t(n)$ –Verteilung werden (meist) mit $t_{n;p}$ bezeichnet und sind für $p \geq 0.5$ ausführlich tabelliert (vgl. Anhang: Tabellen). Für $p < 0.5$ benutzt man die aus der Symmetrie der Verteilung folgende für alle $p \in (0, 1)$ gültige Beziehung:

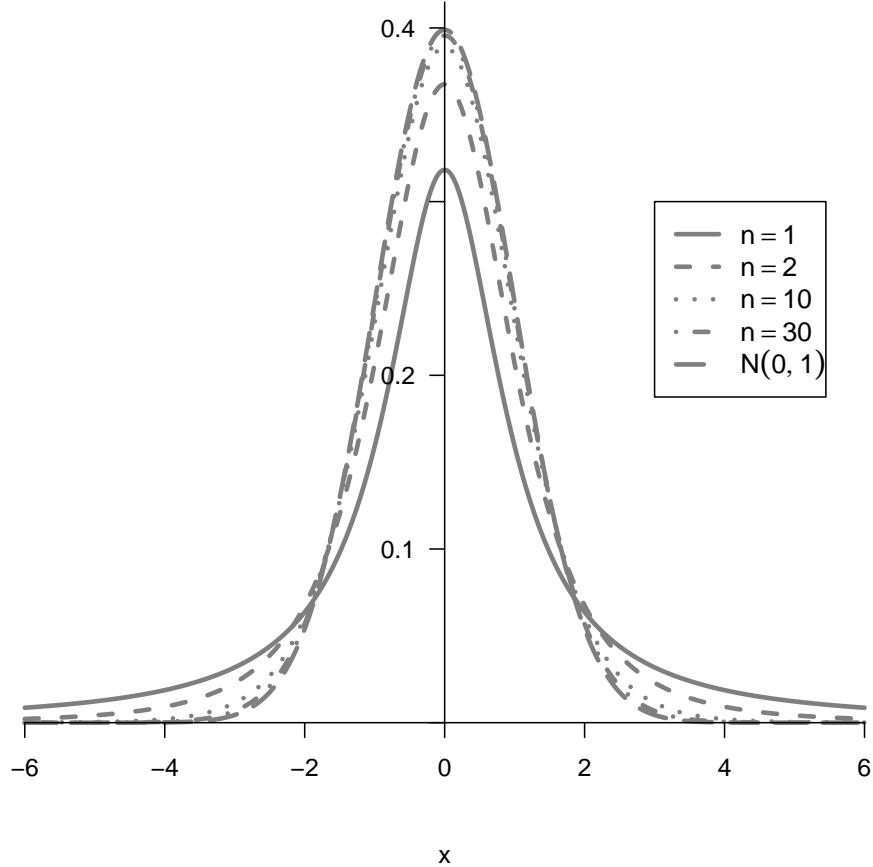
$$t_{n;p} = -t_{n;1-p}$$

Beziehung zur F–Verteilung: $X \sim t(n) \implies X^2 \sim F(1, n)$

Beweis: Da die Funktion $y = x^2$ auf \mathbb{R} nicht umkehrbar eindeutig ist, benutzt man zum Beweis die Methode der Verteilungsfunktion (vgl. 3.3.2).

¹¹Benannt nach dem engl. Statistiker WILLIAM SEALY GOSSET (1876–1937); angestellt bei der Dubliner Brauerei *Arthur Guinness & Son*; publiziert unter dem Pseudonym *Student*.

¹²AUGUSTIN-LOUIS CAUCHY (1789–1857), franz. Mathematiker (bedeutende Beiträge zur Analysis).

Abbildung 4.13: Dichte der $t(n)$ -Verteilung

4.2.7 Betaverteilung

Eine sG X hat eine **Betaverteilung** $\text{Be}(a, b)$ mit den **Formparametern** $a > 0$ und $b > 0$, wenn ihre Dichte gegeben ist durch:

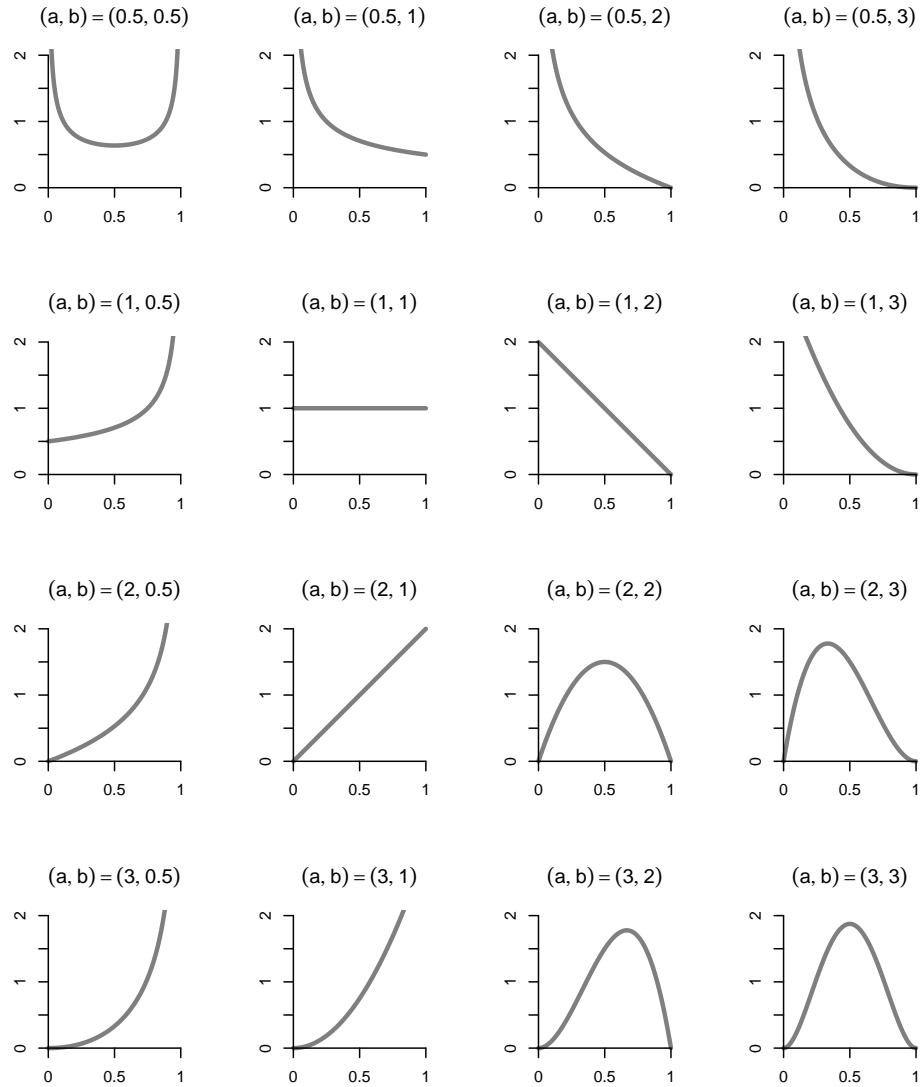
$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \quad \text{für } 0 < x < 1$$

Dabei ist Γ die in 4.2.3 definierte Gammafunktion. Die **Betafunktion** ist definiert durch:

$$B(a, b) = \int_0^1 u^{a-1} (1-u)^{b-1} du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Die Dichte der Betaverteilung lässt sich also auch wie folgt schreiben:

$$f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} \quad \text{für } 0 < x < 1$$

Abbildung 4.14: Dichte der $\text{Be}(a, b)$ -Verteilung

Durch die beiden Formparameter zeigt die $\text{Be}(a, b)$ -Verteilung eine große Formenvielfalt (vgl. Abb 4.14). Man beachte auch die Symmetrie um 0.5 bei vertauschten Parametern.

Erwartungswert/Varianz: Der Erwartungswert und die Varianz von $X \sim \text{Be}(a, b)$ sind gegeben durch:

$$\mathbb{E}(X) = \frac{a}{a+b}, \quad \text{Var}(X) = \frac{ab}{(a+b+1)(a+b)^2}$$

Beweis: Den Erwartungswert berechnet man wie folgt:

$$\mathbb{E}(X) = \int_0^1 x \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \underbrace{\int_0^1 x^a (1-x)^{b-1} dx}_{= B(a+1,b)}$$

Mit der Fakultätseigenschaft der Gammafunktion (vgl. 4.2.3) gilt:

$$B(a+1, b) = \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)}$$

Einsetzen in den ersten Ausdruck ergibt den Erwartungswert. Analog bestimmt man die Varianz mittels Verschiebungssatz.

Die beiden folgenden Eigenschaften zeigt man mit dem Transformationssatz für Dichten (vgl. 3.3.2):

Symmetrie: $X \sim \text{Be}(a, b) \implies 1 - X \sim \text{Be}(b, a)$

Beziehung zur F-Verteilung: Für $m, n \in \mathbb{N}$ gilt:

$$X \sim \text{Be}(m, n) \implies \frac{X/m}{(1-X)/n} \sim F(2m, 2n)$$

Aufgaben

- 4.1 Zeigen Sie für eine diskrete uniforme Verteilung auf $M = \{1, 2, \dots, k\}$, dass der Erwartungswert gleich $\mu = (1+k)/2$ und die Varianz gleich $\sigma^2 = (k^2 - 1)/12$ ist.
- 4.2 (a) Stellen Sie die $B(15, 0.2)$ -Verteilung grafisch dar.
 (b) Zeichnen Sie die $B(n, p)$ -Verteilung für $n = 15$ und $p = 0.10, 0.20, \dots, 0.90$.
 (c) Zeichnen Sie die $B(n, p)$ -Verteilung für $p = 0.05$ und $n = 10, 20, 50, 200$.
- 4.3 Zeigen Sie für eine $B(n, p)$ -Verteilung, dass $\sigma^2 = np(1-p)$. (Hinweis: Verwenden Sie den Verschiebungssatz.)
- 4.4 Ein Kommunikationssystem bestehe aus n (unabhängigen) Komponenten, wobei jede Komponente mit Wahrscheinlichkeit p funktioniert. Das System funktioniert nur, wenn zumindest die Hälfte der Komponenten funktioniert. Für welche Werte von p ist ein 5-Komponentensystem einem 3-Komponentensystem vorzuziehen? (Hinweis: Die Lösung führt auf eine Gleichung 3. Grades. Falls Sie diese Gleichung nicht explizit lösen können, lösen Sie sie numerisch unter Verwendung der R-Funktion `polyroot()`.)

- 4.5 Zwei Personen A und B werfen je zehn Freiwürfe mit einem Basketball. A ist bei jedem Wurf mit Wahrscheinlichkeit 0.80 erfolgreich, B mit Wahrscheinlichkeit 0.85. Mit welcher Wahrscheinlichkeit gewinnt A, B, keiner von beiden? Welche (Unabhängigkeits-) Voraussetzungen liegen den Berechnungen zugrunde?
- 4.6 Zeigen Sie die folgende Beziehung zwischen der Negativen Binomialverteilung und der Binomialverteilung: Für $X \sim \text{NB}(r, p)$ und $Y \sim \text{B}(x, p)$, $x \in \mathbb{N}$, gilt:

$$P(X > x) = P(Y < r)$$

- 4.7 Aus einer Gruppe, bestehend aus 6 Männern und 9 Frauen, soll ein Gremium aus 5 Personen gebildet werden. Das Gremium werde ganz zufällig gebildet und X sei die Zahl der Männer im Gremium. Wie ist X verteilt? Bestimmen Sie den Erwartungswert und die Varianz von X .

Zusatz1: Wie ließe sich die zufällige Zusammenstellung des Gremiums mit Hilfe von R praktisch realisieren? (Hinweis: `sample()`.)

Zusatz2: Angenommen, im Gremium gibt es 4 Männer. Erfolgte die Auswahl rein zufällig? (Hinweis: Wie groß ist die Wahrscheinlichkeit, dass bei zufälliger Auswahl $X \geq 4$?)

- 4.8 Bestimmen Sie die Varianz einer $P(\lambda)$ -Verteilung. (Hinweis: Berechnen Sie zuerst $\mathbb{E}[X(X - 1)]$ und verwenden Sie dann den Verschiebungssatz.)
- 4.9 Bestimmen Sie die Modalwerte der $P(\lambda)$ -Verteilung. (Hinweis: Betrachten Sie den Quotienten zweier aufeinanderfolgender Wahrscheinlichkeiten, d. h. $p(x + 1)/p(x)$.)
- 4.10 Anfragen erreichen einen Server gemäß einer Poissonverteilung mit einem Mittelwert von 10 pro Stunde. Bestimmen Sie die Länge eines Zeitintervalls (in Sekunden), sodass mit einer Wahrscheinlichkeit von 0.90 während dieses Intervalls keine Anfrage eintrifft.
- 4.11 Angenommen, bei der Herstellung von optischen Speichermedien (CDs) treten Verunreinigungen durch Staubteilchen gemäß einer Poissonverteilung mit einem Mittelwert von 0.0002 Teilchen pro cm^2 auf. Die CDs haben eine Fläche von 100 cm^2 .
- (a) Wenn 50 CDs untersucht werden, wie groß ist die Wahrscheinlichkeit, dass keine Teilchen entdeckt werden?
 - (b) Wieviele CDs müssen im Mittel untersucht werden, bevor ein Teilchen entdeckt wird?
 - (c) Wenn 50 CDs untersucht werden, wie groß ist die Wahrscheinlichkeit, dass es darunter höchstens 2 CDs mit einem oder mehr Teilchen gibt?
- 4.12 In Ö gibt es etwa 35000 Eheschließungen im Jahr. Berechnen Sie (approximativ) die Wahrscheinlichkeit dafür, dass bei zumindest einem der Paare:
- (a) beide Partner am 1. Oktober geboren sind.
 - (b) beide Partner am selben Tag geboren sind.

Welche Voraussetzungen liegen den Berechnungen zugrunde?

- 4.13 Ein Produkt wird in Losen der Größe $N = 500$ geliefert. Für eine Qualitätsprüfung werden dem Los willkürlich $n = 50$ Elemente ohne Zurücklegen entnommen und geprüft. Gibt es unter den geprüften Elementen mehr als ein defektes Element, wird das Los zurückgewiesen. Angenommen, das Los enthält (i) 0.8%, (ii) 9% defekte Elemente. Mit welcher Wahrscheinlichkeit wird das Los zurückgewiesen? Rechnen Sie (a) mit der (exakten) Hypergeometrischen Verteilung, (b) einer passenden Binomialapproximation und (c) einer passenden Poissonapproximation. (Sind die Approximationen hier zulässig?)

Zusatz: Der Ausschussanteil betrage allgemein $100p\%$. Bestimmen Sie unter Verwendung aller drei Verteilungen die Wahrscheinlichkeit mit der das Los angenommen wird und stellen Sie die Wahrscheinlichkeiten als Funktion von p grafisch dar.

- 4.14 Zeigen Sie, dass die Varianz einer Exponentialverteilung gleich $1/\lambda^2 = \tau^2$ ist. (Hinweis: Verwenden Sie den Verschiebungssatz.)
- 4.15 Die Kilometerleistung einer Autobatterie sei exponentialverteilt mit einem Mittelwert von 10000 km.
- (a) Mit welcher Wahrscheinlichkeit lässt sich eine 5000 km lange Reise ohne Ersetzung der Batterie absolvieren?
 - (b) Wie lang darf eine Reise höchstens sein, dass sie mit 90% Wahrscheinlichkeit ohne Ersetzung der Batterie beendet werden kann?
 - (c) Bestimmen Sie den Median, den Mittelwert und die Streuung der Kilometerleistung der Batterie.
- 4.16 Die Anzahl N_t von bestimmten Ereignissen (z. B. Telefonanrufe, Aufträge an einen Netzwerkdrucker, etc.) im Zeitintervall $(0, t]$ sei eine nach $\text{P}(\lambda t)$ verteilte sG und T sei die Zeitspanne bis zum Auftreten des *ersten* Ereignisses. Bestimmen Sie die Verteilung von T . (Hinweis: Bestimmen Sie zunächst $P(T > x)$.)

- 4.17 Zeigen Sie für $X \sim \text{Gam}(\alpha, \lambda)$:

$$\mathbb{E}(X^2) = \frac{(\alpha + 1)\alpha}{\lambda^2} = (\alpha + 1)\alpha\beta^2$$

- 4.18 Die Lebensdauer eines Bildschirms sei eine normalverteilte sG mit Mittelwert 8.2 Jahre und Streuung 1.4 Jahre.
- (a) Welcher Anteil solcher Bildschirme funktioniert länger als 10 Jahre, nicht länger als 5 Jahre, zwischen 5 und 10 Jahren?
 - (b) Bestimmen Sie das 10% und das 90% Quantile der Lebensdauer. Wie sind diese Werte zu interpretieren?
 - (c) Sie kaufen einen 3 Jahre alten gebrauchten Bildschirm. Mit welcher Wahrscheinlichkeit funktioniert er noch länger als 5 Jahre? (Hat die Normalverteilung ein „Gedächtnis“?)

- 4.19 Angenommen, die Wegzeit von zu Hause zur TU ist normalverteilt mit Mittelwert 40 Minuten und Standardabweichung 7 Minuten. Wenn man um 13 Uhr eine Prüfung hat und mit Wahrscheinlichkeit 0.95 nicht zu spät kommen möchte, wann spätestens müsste man aufbrechen? Wann, wenn man mit Wahrscheinlichkeit 0.99 nicht zu spät kommen möchte?
- 4.20 Die sG X sei standardnormalverteilt, d. h. $X \sim N(0, 1)$.
- Bestimmen Sie $E(|X|)$. (Hinweis: LotUS)
 - Bestimmen (und zeichnen) Sie die Verteilungsfunktion und (durch Ableiten) die Dichte von $Y = |X|$. (Bem: Die Verteilung von $|X|$ heißt auch **Halbnormalverteilung**. Warum?)
 - Bestimmen Sie einen Ausdruck für das p -Quantil von Y .
- 4.21 Berechnen Sie $\int_2^3 \exp[-2(x-3)^2] dx$.
- 4.22 Bestimmen Sie für $X \sim N(\mu, \sigma^2)$ die Verteilungsfunktion und (durch Ableiten) die Dichte von $Y = e^X$. Stellen Sie die VF und die Dichte von Y für $\mu = 0$ und $\sigma^2 = 1$ grafisch dar. Bestimmen Sie einen allgemeinen Ausdruck für das p -Quantil von Y . (Bem: Die Verteilung von Y nennt man **Log(arithmische)-Normalverteilung** und schreibt $Y \sim L(\mu, \sigma^2)$.)
- 4.23 Wenn X nach $\chi^2(5)$ verteilt ist, bestimmen Sie c und d so, dass $P(c < X < d) = 0.95$ und $P(X < c) = 0.025$.
- 4.24 Bestimmen Sie für $X \sim F(5, 10)$ zwei Werte a und b so, dass $P(X \leq a) = 0.05$ und $P(X \leq b) = 0.95$, und daher $P(a < X < b) = 0.90$.
- 4.25 Bestimmen Sie für $X \sim t(14)$ einen Wert b so, dass $P(-b < X < b) = 0.90$.

Anhang: R-Funktionen

Bernoulli-/Binomialverteilung:

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

Negative Binomialverteilung: $x \hat{=} \text{Zahl der Misserfolge vor dem } r\text{-ten Erfolg}$

```
dnbinom(x, size, prob, mu, log = FALSE)
pnbinom(q, size, prob, mu, lower.tail = TRUE, log.p = FALSE)
qnbinom(p, size, prob, mu, lower.tail = TRUE, log.p = FALSE)
rnbinom(n, size, prob, mu)
```

Geometrische Verteilung: $x \hat{=} \text{Zahl der Misserfolge vor dem ersten Erfolg}$

```
dgeom(x, prob, log = FALSE)
pgeom(q, prob, lower.tail = TRUE, log.p = FALSE)
qgeom(p, prob, lower.tail = TRUE, log.p = FALSE)
rgeom(n, prob)
```

Hypergeometrische Verteilung: $m \hat{=} A, n \hat{=} N - A, k \hat{=} n$

```
dhyper(x, m, n, k, log = FALSE)
phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)
qhyper(p, m, n, k, lower.tail = TRUE, log.p = FALSE)
rhyper(nn, m, n, k)
```

Poisson-Verteilung:

```
dpois(x, lambda, log = FALSE)
ppois(q, lambda, lower.tail = TRUE, log.p = FALSE)
qpois(p, lambda, lower.tail = TRUE, log.p = FALSE)
rpois(n, lambda)
```

Stetige uniforme Verteilung:

```
dunif(x, min = 0, max = 1, log = FALSE)
punif(q, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE)
qunif(p, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE)
runif(n, min = 0, max = 1)
```

Exponentialverteilung: $\text{rate} \hat{=} \lambda = 1/\tau$

```
dexp(x, rate = 1, log = FALSE)
pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)
qexp(p, rate = 1, lower.tail = TRUE, log.p = FALSE)
rexp(n, rate = 1)
```

Gammaverteilung: $\text{shape} \hat{=} \alpha, \text{rate} \hat{=} \lambda = 1/\beta$

```
dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)
pgamma(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE,
       log.p = FALSE)
qgamma(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE,
```

```
log.p = FALSE)
rgamma(n, shape, rate = 1, scale = 1/rate)
```

```
# Gammafunktion
gamma(x)
```

Normalverteilung: $\text{mean} \hat{=} \mu$, $\text{sd} \hat{=} \sigma$

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

χ^2 -Verteilung: $\text{df} \hat{=} n$

```
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)
```

F-Verteilung: $\text{df1} \hat{=} m$, $\text{df2} \hat{=} n$

```
df(x, df1, df2, ncp, log = FALSE)
pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
rf(n, df1, df2, ncp)
```

t-Verteilung: $\text{df} \hat{=} n$

```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

Betaverteilung: $\text{shape1} \hat{=} a$, $\text{shape2} \hat{=} b$

```
dbeta(x, shape1, shape2, ncp = 0, log = FALSE)
pbeta(q, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qbeta(p, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rbeta(n, shape1, shape2, ncp = 0)
```

```
# Betafunktion
beta(a, b)
```

5 Multivariate Verteilungen

Häufig benötigt man zur Beschreibung von Zufallsexperimenten mehrere stochastische Größen. Man betrachte etwa die folgenden Beispiele:

1. Wir wählen zufällig $n = 10$ Personen und beobachten ihre Körpergrößen. Die einzelnen Beobachtungen seien X_1, X_2, \dots, X_n .
2. Wir werfen wiederholt eine Münze. Sei $X_i = 1$, wenn der i -te Wurf ein „Kopf“ ist, und $X_i = 0$ im anderen Fall. Das Experiment lässt sich durch eine Folge X_1, X_2, \dots von Bernoulli-Größen beschreiben.
3. Wir wählen zufällig eine Person aus einer großen Population und messen ihr Körpergewicht X und ihre Körpergröße Y .

Wie lässt sich das Verhalten der obigen sGn beschreiben? Die Spezifikation der einzelnen Verteilungen allein genügt nicht, wir müssen auch den **Zusammenhang** (oder das *Fehlen* desselben) zwischen den einzelnen Größen beschreiben. Wenn beispielsweise im dritten Experiment Y groß ist, dann ist sehr wahrscheinlich auch X groß. Andererseits, in den ersten beiden Experimenten kann man davon ausgehen, dass die einzelnen Größen **unabhängig** sind. D. h., wissen wir etwas über eine dieser Größen, so haben wir dadurch keine *zusätzliche* Information über die anderen. M. a. W., wir benötigen die **gemeinsame Verteilung** der sGn.

5.1 Bivariate Verteilungen

Man betrachte ein Zufallsexperiment mit Merkmalraum Ω und zwei stochastische Größen X_1 und X_2 , die jedem Element $\omega \in \Omega$ eine reelle Zahl zuordnen:

$$X_1(\omega) = x_1 \quad \text{und} \quad X_2(\omega) = x_2$$

Dann nennt man (X_1, X_2) einen (2-dimensionalen) **stochastischen Vektor** (kurz **sV**) mit dem **Merkmalraum**:

$$M = \{(x_1, x_2) \mid x_1 = X_1(\omega), x_2 = X_2(\omega), \omega \in \Omega\}$$

Häufig bezeichnet man den stochastischen Vektor mit $\mathbf{X} = (X_1, X_2)'$ (= transponierter Zeilenvektor).

Wie im eindimensionalen Fall nennt man Teilmengen $B \subseteq M$ **Ereignisse** und die Wahrscheinlichkeit $P(\mathbf{X} \in B)$ für den Eintritt von B lässt sich durch die (2-dimensionale) Verteilungsfunktion charakterisieren.

Verteilungsfunktion: Die (gemeinsame) **Verteilungsfunktion** des stochastischen Vektors $\mathbf{X} = (X_1, X_2)'$ ist definiert durch:

$$F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$$

Dabei handelt es sich um eine Funktion von \mathbb{R}^2 nach $[0, 1]$.

Bem: Der Ausdruck $P(X_1 \leq x_1, X_2 \leq x_2)$ ist eine Kurzschreibweise für $P(\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\})$.

Behauptung: Die Wahrscheinlichkeit von Ereignissen der Form $(a_1, b_1] \times (a_2, b_2]$ (= halboffener Quader) lässt sich wie folgt mittels F bestimmen:

$$P(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$$

Beweis: Man veranschauliche sich das fragliche Ereignis in der Ebene (vgl. Abb 2.2). Die Behauptung folgt dann durch Anwendung elementarer Regeln der Wahrscheinlichkeitsrechnung (vgl. Kapitel 2).

5.1.1 Diskrete stochastische Vektoren

Ist der Merkmalraum M ($\subseteq \mathbb{R}^2$) eines stochastischen Vektors $\mathbf{X} = (X_1, X_2)'$ endlich oder abzählbar, handelt es sich um einen **diskreten sV**. Die (gemeinsame) **W–Funktion** ist gegeben durch:

$$p(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) \quad \text{für alle } (x_1, x_2) \in M$$

Die W–Funktion hat die folgenden Eigenschaften:

$$(1) \quad 0 \leq p(x_1, x_2) \leq 1, \quad (x_1, x_2) \in M \quad \text{und} \quad (2) \quad \sum_{(x_1, x_2) \in M} p(x_1, x_2) = 1$$

Ist die W–Funktion bekannt, lässt sich die Wahrscheinlichkeit für ein beliebiges Ereignis $\{(X_1, X_2) \in B\}$ ($B \subseteq \mathbb{R}^2$) wie folgt bestimmen:

$$P((X_1, X_2) \in B) = \sum_{(x_1, x_2) \in B} p(x_1, x_2)$$

Bsp 5.1 In einem Behälter befinden sich drei Würfel: Würfel 1 ist ein üblicher Würfel, Würfel 2 hat keine Augenzahl 6, dafür zwei Seiten mit der Augenzahl 5, und Würfel 3 hat keine Augenzahl 5, dafür zwei Seiten mit der Augenzahl 6. Das Zufallsexperiment besteht in der zufälligen Auswahl eines Würfels und dem anschließenden Werfen des gewählten

Würfels. Sei X_1 die Nummer des Würfels und X_2 die geworfene Augenzahl. Wie lautet die W-Funktion $p(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$? Die folgende Tabelle zeigt die gemeinsame Verteilung:

		X_2					
		1	2	3	4	5	6
1		$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$
X_1	2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{9}$	0
	3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	$\frac{1}{9}$

1

Träger: Wie im eindimensionalen Fall besteht der **Träger** eines diskreten stochastischen Vektors (X_1, X_2) aus allen Punkten (x_1, x_2) mit $p(x_1, x_2) > 0$. Im obigen Beispiel ist der Träger von (X_1, X_2) gegeben durch $\{(x_1, x_2) \mid x_1 = 1, 2, 3; x_2 = 1, 2, 3, 4, 5, 6\}$, ohne die Punkte $(3, 5)$ und $(2, 6)$.

Randverteilungen: Die Elemente X_1 und X_2 eines stochastischen Vektors (X_1, X_2) sind selbst (1-dimensionale) sGn. Wie bestimmt man ihre Verteilungen? Die W-Funktionen sind gegeben durch:

$$X_1 : \ p_1(x_1) = \sum_{x_2} p(x_1, x_2) \quad \quad X_2 : \ p_2(x_2) = \sum_{x_1} p(x_1, x_2)$$

Um beispielsweise die Wahrscheinlichkeit von $\{X_1 = x_1\}$ zu bestimmen, hält man x_1 fest und summiert $p(x_1, x_2)$ über alle möglichen Werte von x_2 . Die auf diese Weise bestimmten Verteilungen nennt man die **Randverteilungen** von (X_1, X_2) .

Bsp 5.2 Die Randverteilungen von (X_1, X_2) aus Bsp 5.1 ergeben sich durch Summation der Zeilen bzw. Spalten in der gemeinsamen Verteilung:

Man beachte, dass hier aus der Kenntnis der beiden Randverteilungen von X_1 und X_2 allein die gemeinsame Verteilung von (X_1, X_2) *nicht* rekonstruiert werden kann. Der Grund dafür liegt darin, dass X_1 und X_2 nicht *unabhängig* sind. Dieses Konzept wird später noch ausführlicher diskutiert. ■

5.1.2 Stetige stochastische Vektoren

Ist die Verteilungsfunktion $F(x_1, x_2)$ eines stochastischen Vektors (X_1, X_2) eine stetige Funktion, spricht man von einem **stetigen stochastischen Vektor**. In den meisten Fällen lässt sich die VF eines stetigen sVs wie folgt darstellen:

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(w_1, w_2) dw_1 dw_2 \quad \text{für } (x_1, x_2) \in \mathbb{R}^2$$

Den Integranden f nennt man die (gemeinsame) **Dichtefunktion** (kurz **Dichte**) von (X_1, X_2) . Analog zum eindimensionalen Fall gilt an den Stetigkeitspunkten von $f(x_1, x_2)$:

$$\frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} = f(x_1, x_2)$$

Eine Dichtefunktion hat die folgenden Eigenschaften:

$$(1) \quad f(x_1, x_2) \geq 0, \quad (x_1, x_2) \in M \quad \text{und} \quad (2) \quad \iint_M f(x_1, x_2) dx_1 dx_2 = 1$$

Ist die Dichte bekannt, lässt sich die Wahrscheinlichkeit für ein Ereignis $\{(X_1, X_2) \in B\}$ ($B \subseteq \mathbb{R}^2$) wie folgt bestimmen:

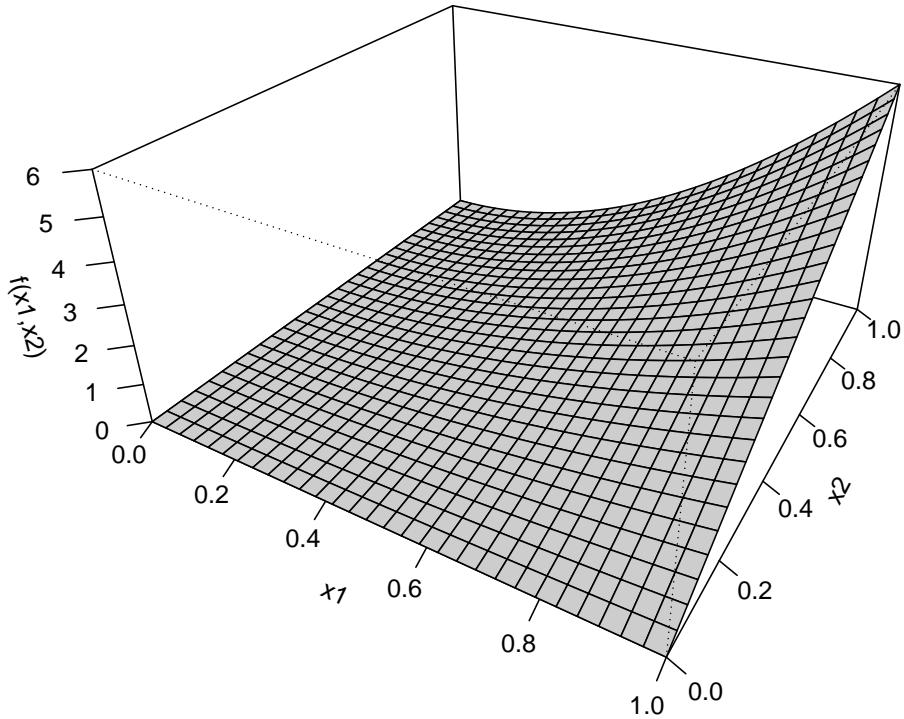
$$P((X_1, X_2) \in B) = \iint_B f(x_1, x_2) dx_1 dx_2$$

Man beachte, dass $P((X_1, X_2) \in B)$ dem **Volumen** unter der Fläche $z = f(x_1, x_2)$ über der Menge B entspricht.

Bsp 5.3 Die Dichte eines sVs (X_1, X_2) sei gegeben wie folgt (vgl. Abb 5.1):

$$f(x_1, x_2) = \begin{cases} 6x_1^2 x_2 & 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{sonst} \end{cases}$$

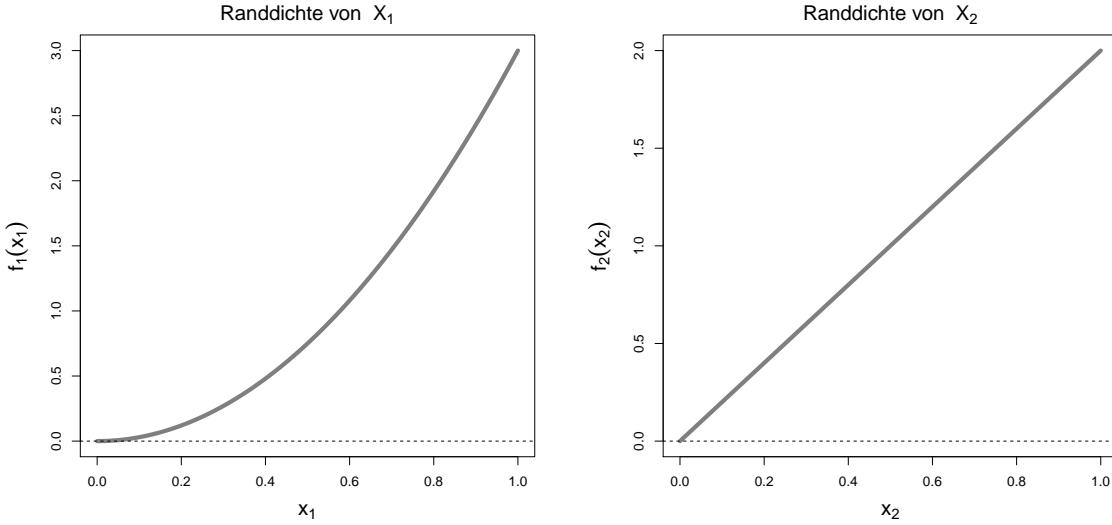
Abbildung 5.1: Gemeinsame Dichte (Bsp 5.3)



Beispielsweise lässt sich die Wahrscheinlichkeit des Ereignisses $\{0 < X_1 < 3/4\} \cap \{1/3 < X_2 < 2\}$ wie folgt berechnen:

$$\begin{aligned}
 P\left(0 < X_1 < \frac{3}{4}, \frac{1}{3} < X_2 < 2\right) &= \int_{1/3}^2 \int_0^{3/4} f(x_1, x_2) dx_1 dx_2 \\
 &= \int_{1/3}^1 \int_0^{3/4} 6x_1^2 x_2 dx_1 dx_2 + \underbrace{\int_1^2 \int_0^{3/4} (0) dx_1 dx_2}_{=0} \\
 &= \int_{1/3}^1 \left[2x_1^3\right]_0^{3/4} x_2 dx_2 = 2 \left(\frac{3}{4}\right)^3 \left[\frac{x_2^2}{2}\right]_{1/3}^1 \\
 &= 2 \left(\frac{3}{4}\right)^3 \left[\frac{1}{2} - \frac{1}{18}\right] = \frac{3}{8}
 \end{aligned}$$

Diese Wahrscheinlichkeit entspricht dem Volumen unter der Fläche $f(x_1, x_2) = 6x_1^2 x_2$ über dem rechteckförmigen Bereich $\{(x_1, x_2) | 0 < x_1 < 3/4, 1/3 < x_2 < 1\}$. ■

Abbildung 5.2: Randdichten (Bsp 5.4)

Träger: Der **Träger** eines stetigen stochastischen Vektors (X₁, X₂) besteht aus allen Punkten (x₁, x₂) mit f(x₁, x₂) > 0. Im obigen Beispiel ist der Träger von (X₁, X₂) gegeben durch (0, 1) × (0, 1).

Randdichten: Analog zum diskreten Fall bestimmt man die **Randdichten** von X₁ bzw. X₂ aus der gemeinsamen Dichte f(x₁, x₂) von (X₁, X₂) wie folgt:

$$X_1 : f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \quad X_2 : f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1$$

Um die Randdichte von X₁ zu bestimmen, ist die gemeinsame Dichte f(x₁, x₂) über x₂ zu integrieren; zur Bestimmung der Randdichte von X₂ ist über x₁ zu integrieren.

Bsp 5.4 Die Randdichten der gemeinsamen Dichte von Bsp 5.3 bestimmt man wie folgt:

$$f_1(x_1) = \int_0^1 6x_1^2 x_2 dx_2 = 6x_1^2 \left[\frac{x_2^2}{2} \right]_0^1 = 3x_1^2 \quad \text{für } 0 < x_1 < 1$$

$$f_2(x_2) = \int_0^1 6x_1^2 x_2 dx_1 = 6x_2 \left[\frac{x_1^3}{3} \right]_0^1 = 2x_2 \quad \text{für } 0 < x_2 < 1$$

Vgl. Abb 5.2 für eine grafische Darstellung. ■

5.1.3 Erwartungswert

Das Konzept des Erwartungswerts lässt sich direkt auf den 2-dimensionalen Fall übertragen. Sei (X_1, X_2) ein stochastischer Vektor und $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ eine reellwertige Funktion. Dann ist $Y = g(X_1, X_2)$ eine (1-dimensionale) sG und existiert ihr **Erwartungswert**, so ist er gegeben durch (LotUS; vgl. 3.4):

$$\text{diskret: } \mathbb{E}(Y) = \sum_{x_1} \sum_{x_2} g(x_1, x_2) p(x_1, x_2)$$

$$\text{stetig: } \mathbb{E}(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2) f(x_1, x_2) dx_1 dx_2$$

Behauptung: (X_1, X_2) sei ein sV und $Y_1 = g_1(X_1, X_2)$ und $Y_2 = g_2(X_1, X_2)$ seien zwei sGn, deren Erwartungswerte existieren. Dann gilt für Konstanten k_1 und k_2 :

$$\mathbb{E}(k_1 Y_1 + k_2 Y_2) = k_1 \mathbb{E}(Y_1) + k_2 \mathbb{E}(Y_2)$$

(Beweis als UE-Aufgabe.)

Bsp 5.5 In der Situation von Bsp 5.3 berechnet man den Erwartungswert von beispielsweise $Y = X_1/X_2$ wie folgt:

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}\left(\frac{X_1}{X_2}\right) = \int_0^1 \int_0^1 \left(\frac{x_1}{x_2}\right) 6x_1^2 x_2 dx_1 dx_2 \\ &= \int_0^1 \int_0^1 6x_1^3 dx_1 dx_2 = \int_0^1 6 \left[\frac{x_1^4}{4}\right]_0^1 dx_2 \\ &= \int_0^1 \frac{3}{2} dx_2 = \frac{3}{2} \end{aligned}$$

■

Bem: Der Erwartungswert des stochastischen Vektors $\mathbf{X} = (X_1, X_2)'$ ist gegeben durch:

$$\mathbb{E}(\mathbf{X}) = \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \end{bmatrix}$$

Dabei wird vorausgesetzt, dass die Erwartungswerte von X_1 und X_2 existieren.

Bsp 5.6 Ist (X_1, X_2) ein stochastischer Vektor, so hat man für die Berechnung des Erwartungswerts von beispielsweise X_1 zwei Möglichkeiten. Man bestimmt zuerst die Randdichte von X_1 und berechnet nach Definition $\mathbb{E}(X_1)$. Die zweite Möglichkeit besteht darin, den Erwartungswert von $Y = g(X_1, X_2) = X_1$ über die gemeinsame Verteilung von (X_1, X_2) zu berechnen.

In der Situation von Bsp 5.3 gilt $f_1(x_1) = 3x_1^2 I_{(0,1)}(x_1)$ (vgl. Bsp 5.4) und der Erwartungswert von X_1 ist nach Definition gegeben durch:

$$\mathbb{E}(X_1) = \int_{-\infty}^{\infty} x_1 f_1(x_1) dx_1 = \int_0^1 3x_1^3 dx_1 = \frac{3}{4}$$

Zweite Möglichkeit:

$$\mathbb{E}(X_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f(x_1, x_2) dx_1 dx_2 = \int_0^1 \int_0^1 6x_1^3 x_2 dx_1 dx_2 = \int_0^1 \frac{3x_2}{2} dx_2 = \frac{3}{4}$$

■

5.1.4 Bedingte Verteilungen

In den vorhergehenden Abschnitten haben wir uns mit der gemeinsamen Verteilung von (X_1, X_2) und den Randverteilungen von X_1 und X_2 beschäftigt. Häufig kennt man aber den Wert einer Variablen (d. h. von X_1 oder X_2) und es stellt sich die Frage, welche Auswirkungen sich dadurch für die Verteilung der anderen Variablen ergeben. Dies führt zum Konzept der **bedingten Verteilung**.

Diskreter Fall: (X_1, X_2) sei ein diskreter sV mit der (gemeinsamen) W–Funktion $p(x_1, x_2)$ (positiv auf dem Träger S) und $p_1(x_1)$ bzw. $p_2(x_2)$ seien die Randverteilungen von X_1 und X_2 . Sei $x_1 \in S_1$ ein Punkt aus dem Träger S_1 von X_1 (d. h. $p_1(x_1) > 0$). Dann gilt nach Definition der bedingten Wahrscheinlichkeit (vgl. 2.10):

$$P(X_2 = x_2 | X_1 = x_1) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_1 = x_1)} = \frac{p(x_1, x_2)}{p_1(x_1)}$$

Für alle x_2 aus dem Träger S_2 von X_2 . Für jedes feste $x_1 \in S_1$ nennt man:

$$p(x_2|x_1) = \frac{p(x_1, x_2)}{p_1(x_1)} \quad \text{für } x_2 \in S_2 \quad (x_1 \in S_1 \text{ fest})$$

die durch $X_1 = x_1$ **bedingte W–Funktion** von X_2 . Letztere hat alle Eigenschaften einer W–Funktion: Es gilt $0 \leq p(x_2|x_1) \leq 1$ und für die Summe gilt:

$$\sum_{x_2} p(x_2|x_1) = \frac{1}{p_1(x_1)} \underbrace{\sum_{x_2} p(x_1, x_2)}_{=p_1(x_1)} = 1$$

Der durch $X_1 = x_1$ bedingte **Erwartungswert** von X_2 ist gegeben durch:

$$\mathbb{E}(X_2|x_1) = \sum_{x_2} x_2 p(x_2|x_1)$$

Ist $u(X_2)$ eine **Funktion** von X_2 , so ist der durch $X_1 = x_1$ bedingte Erwartungswert von $u(X_2)$ gegeben durch:

$$\mathbb{E}[u(X_2)|x_1] = \sum_{x_2} u(x_2) p(x_2|x_1)$$

Die durch $X_1 = x_1$ bedingte **Varianz** von X_2 lässt sich wie folgt berechnen:

$$\text{Var}(X_2|x_1) = \mathbb{E}(X_2^2|x_1) - [\mathbb{E}(X_2|x_1)]^2$$

Analoge Formeln gelten für die durch $X_2 = x_2$ bedingte Verteilung von X_1 :

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p_2(x_2)} \quad \text{für } x_1 \in S_1 \quad (x_2 \in S_2 \text{ fest})$$

$$\mathbb{E}(X_1|x_2) = \sum_{x_1} x_1 p(x_1|x_2)$$

$$\mathbb{E}[u(X_1)|x_2] = \sum_{x_1} u(x_1) p(x_1|x_2)$$

$$\text{Var}(X_1|x_2) = \mathbb{E}(X_1^2|x_2) - [\mathbb{E}(X_1|x_2)]^2$$

Bsp 5.7 In der Situation von Bsp 5.1 (bzw. 5.2) ist beispielsweise die durch $x_1 = 2$ bedingte Verteilung von X_2 gegeben durch:

x_2	1	2	3	4	5
$p(x_2 x_1 = 2)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$

Bedingter Erwartungswert von $X_2|x_1 = 2$:

$$\mathbb{E}(X_2|x_1 = 2) = \sum_{k=1}^4 \frac{k}{6} + \frac{5}{3} = \frac{10}{3}$$

Bedingter Erwartungswert von $X_2^2|x_1 = 2$:

$$\mathbb{E}(X_2^2|x_1 = 2) = \sum_{k=1}^4 \frac{k^2}{6} + \frac{25}{3} = \frac{40}{3}$$

Bedingte Varianz von $X_2|x_1 = 2$:

$$\text{Var}(X_2|x_1 = 2) = \frac{40}{3} - \left(\frac{10}{3}\right)^2 = \frac{20}{9}$$

■

Stetiger Fall: (X_1, X_2) sei ein stetiger sV mit der (gemeinsamen) Dichte $f(x_1, x_2)$ (positiv auf dem Träger S) und $f_1(x_1)$ bzw. $f_2(x_2)$ seien die Randdichten von X_1 und X_2 . Sei $x_1 \in S_1$ ein Punkt aus dem Träger S_1 von X_1 (d. h. $f_1(x_1) > 0$). Dann definiert man die durch $X_1 = x_1$ **bedingte Dichte** von X_2 wie folgt:

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} \quad \text{für } x_2 \in S_2 \quad (x_1 \in S_1 \text{ fest})$$

Die bedingte Dichte hat alle Eigenschaften einer Dichtefunktion: Es gilt $f(x_2|x_1) \geq 0$ und für das Integral gilt:

$$\int_{-\infty}^{\infty} f(x_2|x_1) dx_2 = \frac{1}{f_1(x_1)} \underbrace{\int_{-\infty}^{\infty} f(x_1, x_2) dx_2}_{= f_1(x_1)} = 1$$

Den durch $X_1 = x_1$ **bedingten Erwartungswert** von X_2 berechnet man wie folgt:

$$\mathbb{E}(X_2|x_1) = \int_{-\infty}^{\infty} x_2 f(x_2|x_1) dx_2$$

Ist $u(X_2)$ eine **Funktion** von X_2 , so ist der durch $X_1 = x_1$ bedingte Erwartungswert von $u(X_2)$ gegeben durch:

$$\mathbb{E}[u(X_2)|x_1] = \int_{-\infty}^{\infty} u(x_2) f(x_2|x_1) dx_2$$

Die durch $X_1 = x_1$ bedingte Varianz von X_2 lässt sich wie folgt berechnen:

$$\text{Var}(X_2|x_1) = \mathbb{E}(X_2^2|x_1) - [\mathbb{E}(X_2|x_1)]^2$$

Analoge Formeln gelten für die durch $X_2 = x_2$ bedingte Verteilung von X_1 :

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \quad \text{für } x_1 \in S_1 \quad (x_2 \in S_2 \text{ fest})$$

$$\mathbb{E}(X_1|x_2) = \int_{-\infty}^{\infty} x_1 f(x_1|x_2) dx_1$$

$$\mathbb{E}[u(X_1)|x_2] = \int_{-\infty}^{\infty} u(x_1) f(x_1|x_2) dx_1$$

$$\text{Var}(X_1|x_2) = \mathbb{E}(X_1^2|x_2) - [\mathbb{E}(X_1|x_2)]^2$$

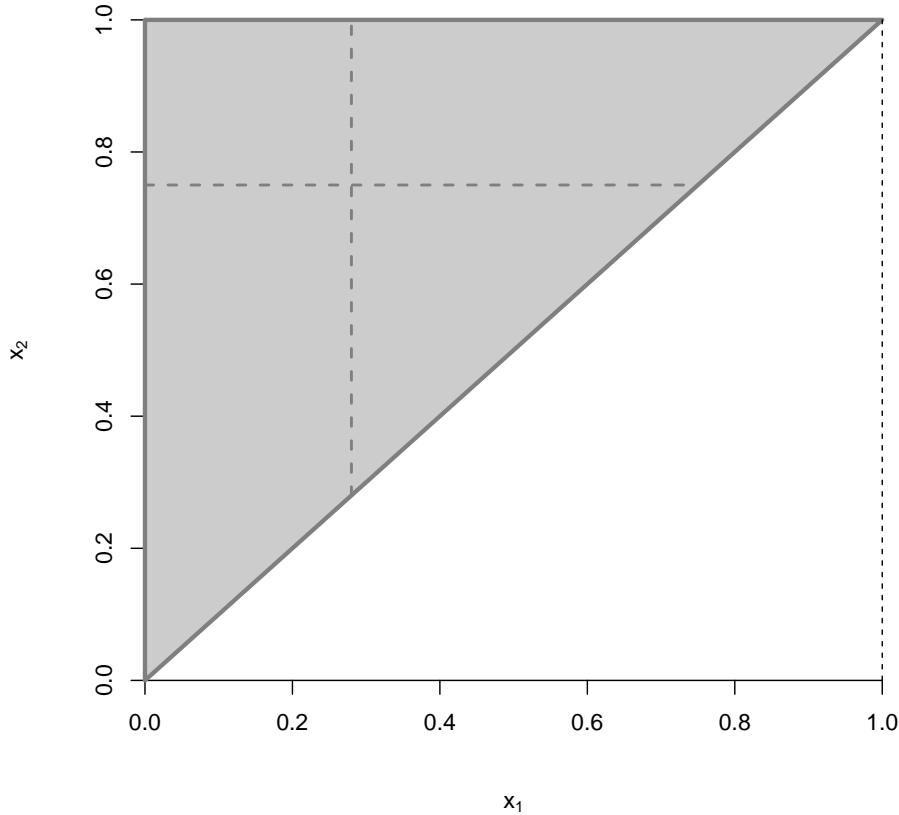
Bsp 5.8 Die gemeinsame Dichte von (X_1, X_2) sei gegeben wie folgt:

$$f(x_1, x_2) = \begin{cases} 2 & 0 < x_1 < x_2 < 1 \\ 0 & \text{sonst} \end{cases}$$

Bei der Bestimmung der Randdichten ist zu beachten, dass der Träger von (X_1, X_2) nicht rechtecksförmig ist (vgl. Abb 5.3; die strichlierten Linien sind zwei exemplarische Integrationswege):

$$f_1(x_1) = \int_{x_1}^1 2 dx_2 = 2(1 - x_1) \quad \text{für } 0 < x_1 < 1$$

$$f_2(x_2) = \int_0^{x_2} 2 dx_1 = 2x_2 \quad \text{für } 0 < x_2 < 1$$

Abbildung 5.3: Träger von (X_1, X_2) (Bsp 5.8)

Die durch $X_2 = x_2$ bedingte Dichte von X_1 lautet wie folgt:

$$f(x_1|x_2) = \frac{2}{2x_2} = \frac{1}{x_2} \quad \text{für } 0 < x_1 < x_2 \quad (x_2 \text{ fest})$$

Es handelt sich also um eine (stetige) uniforme Verteilung (vgl. 4.2.1) auf dem Intervall $(0, x_2)$. Der durch $X_2 = x_2$ bedingte Erwartungswert von X_1 ist daher gegeben durch:

$$\mathbb{E}(X_1|x_2) = \int_{-\infty}^{\infty} x_1 f(x_1|x_2) dx_1 = \frac{x_2}{2}, \quad 0 < x_2 < 1$$

Bedingte Varianz:

$$\text{Var}(X_1|x_2) = \int_0^{x_2} x_1^2 \frac{1}{x_2} dx_1 - \left(\frac{x_2}{2}\right)^2 = \frac{x_2^2}{3} - \frac{x_2^2}{4} = \frac{x_2^2}{12}, \quad 0 < x_2 < 1$$

■

5.2 Korrelation

Wir kennen bereits das Konzept der *empirischen* Korrelation von 1.9.3. Nun betrachten wir das entsprechende (bivariate) Verteilungskonzept. Da man üblicherweise in diesem Zusammenhang statt X_1 und X_2 die Bezeichnungen X und Y verwendet, folgen wir in diesem Abschnitt ebenfalls dieser Konvention.

Die (gemeinsame) W-Funktion bzw. Dichte des stochastischen Vektors (X, Y) sei $p(x, y)$ bzw. $f(x, y)$. Die Mittelwerte von X und Y seien mit μ_1 bzw. μ_2 bezeichnet, die Varianzen mit σ_1^2 bzw. σ_2^2 . Weiters setzen wir voraus, dass im Folgenden alle betrachteten Erwartungswerte auch existieren.

Kovarianz: Die **Kovarianz** $\text{Cov}(X, Y)$ (auch σ_{12}) von X und Y ist definiert durch:

$$\sigma_{12} = \text{Cov}(X, Y) = \mathbb{E}[(X - \mu_1)(Y - \mu_2)]$$

Die Kovarianz lässt sich auch wie folgt berechnen:

$$\begin{aligned}\mathbb{E}[(X - \mu_1)(Y - \mu_2)] &= \mathbb{E}(XY - \mu_2 X - \mu_1 Y + \mu_1 \mu_2) \\ &= \mathbb{E}(XY) - \mu_2 \mathbb{E}(X) - \mu_1 \mathbb{E}(Y) + \mu_1 \mu_2 \\ &= \mathbb{E}(XY) - \mu_1 \mu_2\end{aligned}$$

Letzteren Ausdruck nennt man den **Verschiebungssatz** (für die Kovarianz).

Bem: Die Kovarianz von X und X ist die Varianz von X :

$$\text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \text{Var}(X)$$

Der Verschiebungssatz für die Kovarianz reduziert sich in diesem Fall auf den Verschiebungssatz für die Varianz:

$$\text{Cov}(X, X) = \text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$

Korrelation: Sind σ_1 und σ_2 beide positiv, ist der **Korrelationskoeffizient** ρ_{XY} (kurz ρ) von X und Y gegeben durch:

$$\rho = \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

Bem: Ersetzt man den Zähler von ρ durch den Verschiebungssatz, so folgt für den Erwartungswert des Produkts XY :

$$\mathbb{E}(XY) = \mu_1\mu_2 + \rho\sigma_1\sigma_2 = \mu_1\mu_2 + \text{Cov}(X, Y)$$

Eigenschaften von ρ : Der Korrelationskoeffizient hat die folgenden Eigenschaften:

- (1) Es gilt $-1 \leq \rho \leq 1$ (oder $|\rho| \leq 1$).
- (2) Im Grenzfall $\rho = \pm 1$ (oder $|\rho| = 1$) gilt:

$$P(Y = a + bX) = 1$$

D. h., die gesamte Wahrscheinlichkeitsverteilung von (X, Y) konzentriert sich für $|\rho| = 1$ auf einer Geraden. Für $\rho = 1$ gilt $b > 0$, für $\rho = -1$ gilt $b < 0$.

Bem: Für $\rho = 1$ gilt $b = \sigma_2/\sigma_1 > 0$; für $\rho = -1$ gilt $b = -\sigma_2/\sigma_1 < 0$.

Beweis für (1): Man betrachte die folgende nichtnegative quadratische Funktion:

$$h(z) := \mathbb{E} \left\{ [(X - \mu_1) + z(Y - \mu_2)]^2 \right\}$$

Ausquadriert lautet $h(z)$ wie folgt:

$$h(z) = \sigma_1^2 + 2z\sigma_{12} + z^2\sigma_2^2 \geq 0$$

Damit $h(z) \geq 0$ für alle z , muss die quadratische Gleichung $h(z) = 0$ zwei (konjugiert) komplexe Lösungen haben. Das ist genau dann der Fall, wenn die Diskriminante $\sigma_{12}^2 - \sigma_1^2\sigma_2^2$ kleiner oder gleich Null ist:

$$\sigma_{12}^2 - \sigma_1^2\sigma_2^2 \leq 0 \iff \rho^2 = \left(\frac{\sigma_{12}}{\sigma_1\sigma_2} \right)^2 \leq 1$$

Das war zu zeigen.

Bsp 5.9 In Bsp 5.8 haben wir die Randdichten von $X (= X_1)$ und $Y (= X_2)$ bestimmt. Nach einfachen Rechnungen (UE-Aufgabe) ergibt sich:

$$\mathbb{E}(X) = \frac{1}{3}, \quad \mathbb{E}(Y) = \frac{2}{3}, \quad \text{Var}(X) = \text{Var}(Y) = \frac{1}{18}$$

$$\mathbb{E}(XY) = \int_0^1 \int_0^y 2xy \, dx \, dy = \frac{1}{4} \implies \text{Cov}(X, Y) = \frac{1}{4} - \left(\frac{1}{3} \right) \left(\frac{2}{3} \right) = \frac{1}{36}$$

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{1/36}{1/18} = \frac{1}{2}$$

■

Interpretation: Der Korrelationskoeffizient ρ lässt sich als Maß für den **linearen Zusammenhang** zwischen X und Y interpretieren. Das ergibt sich einerseits aus der obigen Eigenschaft (2) und andererseits aus dem folgenden Sachverhalt:

Ist die (gemeinsame) Verteilung des stochastischen Vektors (X, Y) derart, dass der bedingte Erwartungswert $\mathbb{E}(Y|x) = a + bx$ eine *Gerade* ist, dann kann man zeigen, dass diese Gerade die folgende Gestalt hat:

$$\mathbb{E}(Y|x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1)$$

(Analog, falls $\mathbb{E}(X|y)$ eine Gerade ist.) Beispielsweise gilt im Kontext von Bsp 5.8, dass (UE-Aufgabe):

$$\mathbb{E}(Y|x) = \frac{1+x}{2}$$

Der durch $X = x$ bedingte Erwartungswert von Y ist eine Gerade; also gilt (vgl. auch Bsp 5.9):

$$\mathbb{E}(Y|x) = \frac{1+x}{2} = \underbrace{\frac{2}{3}}_{\rho} + \underbrace{\left(\frac{1}{2}\right) \frac{\sqrt{1/18}}{\sqrt{1/18}}}_{\sigma_2/\sigma_1} \left(x - \frac{1}{3}\right) = \frac{2}{3} + \left(\frac{1}{2}\right) \left(x - \frac{1}{3}\right)$$

Bsp 5.10 X sei eine auf dem Intervall $(-1, 1)$ (stetig) uniform verteilte sG. Die Dichte von X ist $f(x) = (1/2)I_{(-1,1)}(x)$ und es gilt:

$$\mathbb{E}(X) = \int_{-1}^1 \frac{x}{2} dx = 0, \quad \mathbb{E}(X^2) = \int_{-1}^1 \frac{x^2}{2} dx = \frac{1}{3}, \quad \mathbb{E}(X^3) = \int_{-1}^1 \frac{x^3}{2} dx = 0$$

Definiert man $Y = X^2$, so gilt:

$$\text{Cov}(X, Y) = \text{Cov}(X, X^2) = \mathbb{E}(X^3) - \mathbb{E}(X) \mathbb{E}(X^2) = 0 \implies \rho_{XY} = 0$$

D.h., auch wenn die Korrelation gleich Null ist, so gibt es hier dennoch einen *perfekten* (deterministischen) Zusammenhang zwischen X und Y (nämlich $Y = X^2$). Letzterer ist allerdings nichtlinearer Natur. ■

5.3 Unabhängigkeit

Die gemeinsame Dichte des (stetigen) stochastischen Vektors (X_1, X_2) sei $f(x_1, x_2)$, und die beiden Randdichten seien $f_1(x_1)$ bzw. $f_2(x_2)$. Aus der Definition der bedingten Dichte $f(x_2|x_1)$ folgt, dass die gemeinsame Dichte wie folgt geschrieben werden kann:

$$f(x_1, x_2) = f(x_2|x_1)f_1(x_1)$$

Wenn nun die bedingte Dichte $f(x_2|x_1)$ nicht von x_1 abhängt, so gilt für die Randdichte von X_2 :

$$\begin{aligned} f_2(x_2) &= \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 = \int_{-\infty}^{\infty} f(x_2|x_1)f_1(x_1) dx_1 \\ &= f(x_2|x_1) \underbrace{\int_{-\infty}^{\infty} f_1(x_1) dx_1}_{=1} \\ &= f(x_2|x_1) \end{aligned}$$

D.h., im Falle, dass $f(x_2|x_1)$ nicht von x_1 abhängt, gilt:

$$f_2(x_2) = f(x_2|x_1) \quad \text{und} \quad f(x_1, x_2) = f_1(x_1)f_2(x_1)$$

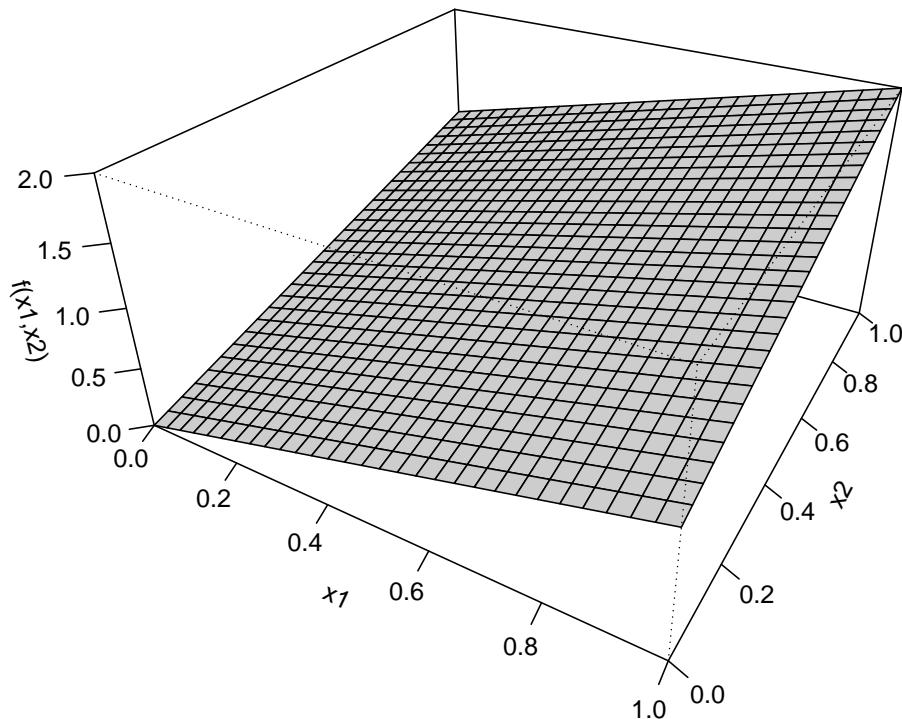
(Analoge Überlegungen gelten auch im diskreten Fall.) Die obigen Überlegungen motivieren die folgende Definition.

Unabhängigkeit: Die Dichte (W-Funktion) von (X_1, X_2) sei $f(x_1, x_2)$ ($p(x_1, x_2)$) und die Randdichten (Randverteilungen) seien $f_1(x_1)$ ($p_1(x_1)$) bzw. $f_2(x_2)$ ($p_2(x_2)$). Die sGn X_1 und X_2 sind (stochastisch) **unabhängig** (kurz **ua.**), wenn:

$$\text{stetig: } f(x_1, x_2) \equiv f_1(x_1)f_2(x_2) \quad \text{diskret: } p(x_1, x_2) \equiv p_1(x_1)p_2(x_2)$$

Nicht unabhängige sGn nennt man (stochastisch) **abhängig**.

Bem: Die Äquivalenz (\equiv) in der obigen Definition ist wie folgt zu verstehen: Es mag auch für unabhängige sGn Punkte $(x_1, x_2) \in S$ (= Träger von (X_1, X_2)) geben, für die $f(x_1, x_2) \neq f_1(x_1)f_2(x_2)$. Ist aber A die Menge aller derartigen Punkte, so gilt $P(A) = 0$. Allerdings, im *diskreten* Fall muss für die Unabhängigkeit von X_1 und X_2 die Gleichung $p(x_1, x_2) = p_1(x_1)p_2(x_2)$ für alle Punkte aus dem Träger von (X_1, X_2) gelten.

Abbildung 5.4: Dichte von (X_1, X_2) (Bsp 5.11)

Bsp 5.11 Die gemeinsame Dichte von X_1 und X_2 sei gegeben durch (Abb 5.4):

$$f(x_1, x_2) = \begin{cases} x_1 + x_2 & 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{sonst} \end{cases}$$

Randdichten:

$$f_1(x_1) = \int_0^1 (x_1 + x_2) dx_2 = x_1 + \frac{1}{2} \quad \text{für } 0 < x_1 < 1$$

$$f_2(x_2) = \int_0^1 (x_1 + x_2) dx_1 = x_2 + \frac{1}{2} \quad \text{für } 0 < x_2 < 1$$

Da $f(x_1, x_2) \neq f_1(x_1)f_2(x_2)$, sind X_1 und X_2 abhängig. ■

Die Unabhängigkeit von sGn lässt sich auch über die Verteilungsfunktionen formulieren. Dabei muss man nicht zwischen stetigen und diskreten sVn unterscheiden.

Behauptung 1: Die gemeinsame Verteilungsfunktion von (X_1, X_2) sei $F(x_1, x_2)$ und die Verteilungsfunktionen von X_1 und X_2 seien $F_1(x_1)$ bzw. $F_2(x_2)$. Dann gilt: X_1 und X_2 sind genau dann unabhängig, wenn:

$$F(x_1, x_2) = F_1(x_1)F_2(x_2) \quad \text{für alle } (x_1, x_2) \in \mathbb{R}^2$$

Behauptung 2: Existieren $\mathbb{E}[u(X_1)]$ und $\mathbb{E}[v(X_2)]$ für zwei Funktionen u und v und sind X_1 und X_2 unabhängig, dann gilt:

$$\mathbb{E}[u(X_1)v(X_2)] = \mathbb{E}[u(X_1)]\mathbb{E}[v(X_2)]$$

Beweis: Im stetigen Fall gilt:

$$\begin{aligned} \mathbb{E}[u(X_1)v(X_2)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(x_1)v(x_2)f(x_1, x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(x_1)v(x_2)f_1(x_1)f_2(x_2) dx_1 dx_2 \\ &= \left[\int_{-\infty}^{\infty} u(x_1)f_1(x_1) dx_1 \right] \left[\int_{-\infty}^{\infty} v(x_2)f_2(x_2) dx_2 \right] \\ &= \mathbb{E}[u(X_1)]\mathbb{E}[v(X_2)] \end{aligned}$$

Im diskreten Fall argumentiert man analog.

Folgerung: X und Y seien zwei sGn mit den Mittelwerten μ_1 und μ_2 und den Varianzen $\sigma_1^2 > 0$ und $\sigma_2^2 > 0$. Dann folgt aus der Unabhängigkeit von X und Y auch die Unkorreliertheit:

$$X, Y \text{ ua.} \implies \rho_{XY} = 0$$

Bem: Die Umkehrung gilt nicht, d. h., aus der Unkorreliertheit folgt i. A. nicht die Unabhängigkeit. (Vgl. Bsp 5.10 für ein Gegenbeispiel.)

Beweis: Es genügt zu zeigen, dass die Kovarianz gleich Null ist:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_1)(Y - \mu_2)] = \mathbb{E}(X - \mu_1)\mathbb{E}(Y - \mu_2) = 0$$

Dabei haben wir Behauptung 2 verwendet.

5.4 Mehrdimensionale Erweiterungen

Die für zwei sGn entwickelten Konzepte lassen sich unschwer auf mehrere sGn erweitern. Man betrachte ein Zufallsexperiment mit Merkmalraum Ω und n sGn X_1, X_2, \dots, X_n , die jedem Element $\omega \in \Omega$ eine reelle Zahl zuordnen:

$$X_i(\omega) = x_i \quad \text{für } i = 1, 2, \dots, n$$

Dann nennt man (X_1, X_2, \dots, X_n) einen (n -dimensionalen) **stochastischen Vektor** und schreibt $\mathbf{X} = (X_1, X_2, \dots, X_n)'$. Der **Merkmalraum** von \mathbf{X} ist gegeben durch:

$$M = \{(x_1, x_2, \dots, x_n) \mid x_i = X_i(\omega), \omega \in \Omega, i = 1, 2, \dots, n\}$$

Die (gemeinsame) **Verteilungsfunktion** des stochastischen Vektors \mathbf{X} ist definiert durch:

$$F(\mathbf{x}) = F(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

Dabei handelt es sich um eine Funktion von \mathbb{R}^n nach $[0, 1]$.

Sind alle n sGn X_1, X_2, \dots, X_n vom diskreten Typ, spricht man von einem **diskreten** stochastischen Vektor. Die (gemeinsame) **W-Funktion** $p(x_1, x_2, \dots, x_n)$ ist gegeben durch:

$$p(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \quad \text{für } (x_1, x_2, \dots, x_n) \in M$$

Eine W-Funktion hat die folgenden Eigenschaften:

$$(1) \quad 0 \leq p(x_1, x_2, \dots, x_n) \leq 1 \quad \text{und} \quad (2) \quad \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} p(x_1, x_2, \dots, x_n) = 1$$

Ist die Verteilungsfunktion $F(x_1, x_2, \dots, x_n)$ eine stetige Funktion, spricht man von einem **stetigen** stochastischen Vektor. In den meisten Fällen lässt sich die VF eines stetigen sVs wie folgt darstellen:

$$F(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(w_1, w_2, \dots, w_n) dw_1 dw_2 \cdots dw_n$$

Den Integranden f nennt man die (gemeinsame) **Dichtefunktion** (kurz **Dichte**) von \mathbf{X} . An den Stetigkeitspunkten von $f(x_1, x_2, \dots, x_n)$ gilt:

$$\frac{\partial^n F(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n} = f(x_1, x_2, \dots, x_n)$$

Eine Dichtefunktion hat die folgenden Eigenschaften:

$$(1) \quad f(x_1, x_2, \dots, x_n) \geq 0 \quad \text{und} \quad (2) \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n = 1$$

Ist $Y = u(X_1, X_2, \dots, X_n)$ eine **Funktion** des stochastischen Vektors, lässt sich der **Erwartungswert** von Y wie folgt berechnen:

$$\text{diskret: } \mathbb{E}(Y) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} u(x_1, x_2, \dots, x_n) p(x_1, x_2, \dots, x_n)$$

$$\text{stetig: } \mathbb{E}(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n$$

Für $Y_j = u_j(X_1, X_2, \dots, X_n)$ und Konstanten k_j , $j = 1, 2, \dots, m$, gilt:

$$\mathbb{E} \left(\sum_{j=1}^m k_j Y_j \right) = \sum_{j=1}^m k_j \mathbb{E}(Y_j)$$

Das Konzept der **Randverteilung** lässt sich ebenfalls einfach auf mehrere Dimensionen erweitern. Im stetigen Fall ist beispielsweise die **Randdichte** von X_1 gegeben durch:

$$f_1(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_2 \cdots dx_n$$

Die durch $X_1 = x_1$ **bedingte Dichte** von (X_2, \dots, X_n) ist definiert durch:

$$f(x_2, \dots, x_n | x_1) = \frac{f(x_1, x_2, \dots, x_n)}{f_1(x_1)}$$

Der durch $X_1 = x_1$ **bedingte Erwartungswert** von $u(X_2, \dots, X_n)$ ist gegeben durch:

$$\mathbb{E}[u(X_2, \dots, X_n) | x_1] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_2, \dots, x_n) f(x_2, \dots, x_n | x_1) dx_2 \cdots dx_n$$

Analoge Ausdrücke gelten für andere Randdichten, bedingte Dichten oder bedingte Erwartungswerte (oder für diskrete sVn).

Die sGn X_1, X_2, \dots, X_n sind (stochastisch) **unabhängig**, wenn:

$$\text{diskret: } p(x_1, x_2, \dots, x_n) \equiv p_1(x_1)p_2(x_2) \cdots p_n(x_n)$$

$$\text{stetig: } f(x_1, x_2, \dots, x_n) \equiv f_1(x_1)f_2(x_2) \cdots f_n(x_n)$$

Bem: Sind die Größen X_1, X_2, \dots, X_n unabhängig, so sind sie auch **paarweise unabhängig**, d.h., X_i und X_j sind unabhängig für alle $i \neq j$. Die Umkehrung gilt aber nicht, d.h., aus der paarweisen Unabhängigkeit folgt nicht die (vollständige) Unabhängigkeit von X_1, X_2, \dots, X_n .

Existieren die Erwartungswerte $\mathbb{E}[u_i(X_i)]$ für Funktionen u_i , $i = 1, 2, \dots, n$, so gilt für unabhängige Größen X_1, X_2, \dots, X_n :

$$\mathbb{E}\left[\prod_{i=1}^n u_i(X_i)\right] = \prod_{i=1}^n \mathbb{E}[u_i(X_i)]$$

Bsp 5.12 Die sGn X_1, X_2 und X_3 seien unabhängig mit identischer Dichte:

$$f(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{sonst} \end{cases}$$

Die gemeinsame Dichte von (X_1, X_2, X_3) ist gegeben durch:

$$f(x_1, x_2, x_3) = f(x_1)f(x_2)f(x_3) = 8x_1x_2x_3 \quad \text{für } 0 < x_i < 1, i = 1, 2, 3$$

Der Erwartungswert von beispielsweise $Y = 5X_1X_2^3 + 3X_2X_3^4$ lässt sich wie folgt berechnen:

$$\mathbb{E}(Y) = \int_0^1 \int_0^1 \int_0^1 (5x_1x_2^3 + 3x_2x_3^4) 8x_1x_2x_3 dx_1 dx_2 dx_3 \quad (= 2)$$

Wegen der Unabhängigkeit von X_1, X_2, X_3 aber auch wie folgt:

$$\mathbb{E}(Y) = 5\mathbb{E}(X_1X_2^3) + 3\mathbb{E}(X_2X_3^4) = 5\mathbb{E}(X_1)\mathbb{E}(X_2^3) + 3\mathbb{E}(X_2)\mathbb{E}(X_3^4)$$

Berechnen Sie $\mathbb{E}(Y)$ als UE-Aufgabe nach beiden Methoden. ■

Sind die sGn X_1, X_2, \dots, X_n unabhängig mit identischer Verteilung, nennt man sie **uiv** oder **iid**.¹ So sind in Bsp 5.12 die Größen X_1, X_2, X_3 iid mit (identischer) Dichte $f(x)$.

¹independent and identically distributed

5.4.1 Varianz–Kovarianzmatrix

Die in 5.2 diskutierte Kovarianz zwischen zwei sGn lässt sich auf den mehrdimensionalen Fall erweitern. Sei $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ ein stochastischer Vektor. Der Erwartungswert von \mathbf{X} ist der Vektor der Erwartungswerte:

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \mathbb{E}(X_2), \dots, \mathbb{E}(X_n))'$$

Ist $\mathbf{W} = [W_{ij}]$ eine $(m \times m)$ –Matrix aus sGn, so ist der Erwartungswert von \mathbf{W} die Matrix der Erwartungswerte:

$$\mathbb{E}(\mathbf{W}) = [\mathbb{E}(W_{ij})]$$

Für einen stochastischen Vektor $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ mit Mittelwert $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$ ist die **Varianz–Kovarianzmatrix** definiert durch:

$$\text{Cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] = [\sigma_{ij}]$$

Dabei ist $\sigma_{ii} = \sigma_i^2 = \text{Var}(X_i)$ die Varianz von X_i und $\sigma_{ij} = \text{Cov}(X_i, X_j)$ die Kovarianz von X_i und X_j . Ausführlich geschrieben lautet $\text{Cov}(\mathbf{X})$ wie folgt:

$$\text{Cov}(\mathbf{X}) = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Var}(X_n) \end{bmatrix}$$

Wegen $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$ handelt es sich um eine **symmetrische** Matrix. Außerdem ist $\text{Cov}(\mathbf{X})$ **positiv semidefinit**, d. h.:

$$\mathbf{a}' \text{Cov}(\mathbf{X}) \mathbf{a} \geq 0 \quad \text{für alle Vektoren } \mathbf{a} \in \mathbb{R}^n$$

Im **bivariaten** Fall (d. h. für $n = 2$) hat die Varianz–Kovarianzmatrix die folgende Form:

$$\text{Cov}(\mathbf{X}) = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

Dabei ist ρ der Korrelationskoeffizient von X_1 und X_2 .

Bsp 5.13 Für die bivariate Größe von Bsp 5.8 gilt (vgl. auch Bsp 5.9):

$$\mathbb{E}(\mathbf{X}) = \frac{1}{3} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{und} \quad \text{Cov}(\mathbf{X}) = \frac{1}{36} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

■

Behauptung: Sei $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ ein stochastischer Vektor mit dem Erwartungswert $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$ und \mathbf{A} eine $(m \times n)$ -Matrix aus Konstanten. Dann gilt:

- (1) $\mathbb{E}(\mathbf{AX}) = \mathbf{A}\boldsymbol{\mu}$
- (2) $\text{Cov}(\mathbf{X}) = \mathbb{E}(\mathbf{XX}') - \boldsymbol{\mu}\boldsymbol{\mu}'$
- (3) $\text{Cov}(\mathbf{AX}) = \mathbf{ACov}(\mathbf{X})\mathbf{A}'$

Beweis: (1) folgt aus der Linearität des Erwartungswerts; (2) zeigt man wie folgt:

$$\begin{aligned} \text{Cov}(\mathbf{X}) &= \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] \\ &= \mathbb{E}[\mathbf{XX}' - \boldsymbol{\mu}\mathbf{X}' - \mathbf{X}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}'] \\ &= \mathbb{E}[\mathbf{XX}'] - \underbrace{\boldsymbol{\mu}\mathbb{E}[\mathbf{X}']}_{\boldsymbol{\mu}'} - \underbrace{\mathbb{E}(\mathbf{X})\boldsymbol{\mu}'}_{\boldsymbol{\mu}} + \boldsymbol{\mu}\boldsymbol{\mu}' \\ &= \mathbb{E}(\mathbf{XX}') - \boldsymbol{\mu}\boldsymbol{\mu}' \end{aligned}$$

Nach Definition gilt:

$$\begin{aligned} \text{Cov}(\mathbf{AX}) &= \mathbb{E}[(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu})(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu})'] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'\mathbf{A}'] \\ &= \mathbf{A} \underbrace{\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})']}_{\text{Cov}(\mathbf{X})} \mathbf{A}' \\ &= \mathbf{ACov}(\mathbf{X})\mathbf{A}' \end{aligned}$$

Das zeigt (3).

5.5 Transformationen

Der Transformationssatz für Dichten (vgl. 3.3.2) lässt sich auf (stetige) stochastische Vektoren verallgemeinern.² Um die Verteilung einer reellwertigen Funktion von \mathbf{X} zu bestimmen, kann man aber auch die **Methode der Verteilungsfunktion** anwenden. Im diskreten Fall führt häufig eine direkte Überlegung zum Ziel. Dazu einige Beispiele.

²Vgl. HOGG ET AL. (2005) für eine detaillierte Darstellung des mehrdimensionalen Falls.

Bsp 5.14 [Diskreter Fall] Die gemeinsame W-Funktion von X_1 und X_2 sei gegeben durch:

$$p(x_1, x_2) = \frac{x_1 x_2}{36} \quad \text{für } x_1, x_2 = 1, 2, 3$$

Wie lautet die W-Funktion von $Y = X_1 X_2$? Der Merkmalraum von Y ist gegeben durch $M_Y = \{1, 2, 3, 4, 6, 9\}$ und die Wahrscheinlichkeit von $Y = y$ lässt sich wie folgt berechnen:

$$p_Y(y) = P(Y = y) = \sum_{(x_1, x_2): x_1 x_2 = y} p(x_1, x_2)$$

Beispielsweise gilt für $y = 6$:

$$p_Y(6) = \frac{(2)(3)}{36} + \frac{(3)(2)}{36} = \frac{12}{36}$$

Ebenso behandelt man die anderen Fälle:

y	1	2	3	4	6	9
$p_Y(y)$	$\frac{1}{36}$	$\frac{4}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{12}{36}$	$\frac{9}{36}$

■

Bsp 5.15 [Stetiger Fall] Die gemeinsame Dichte von X_1 und X_2 sei gegeben durch:

$$f(x_1, x_2) = \begin{cases} e^{-(x_1+x_2)} & 0 < x_i < \infty, i = 1, 2 \\ 0 & \text{sonst} \end{cases}$$

Wie lautet die Dichte von $Y = X_1 + X_2$? Dazu bestimmt man zunächst die Verteilungsfunktion von Y :

$$\begin{aligned} F_Y(y) &= P(X_1 + X_2 \leq y) = \iint_{x_1+x_2 \leq y} e^{-(x_1+x_2)} dx_1 dx_2 = \int_0^y \int_0^{y-x_2} e^{-(x_1+x_2)} dx_1 dx_2 \\ &= \int_0^y e^{-x_2} [1 - e^{-(y-x_2)}] dx_2 = \int_0^y (e^{-x_2} - e^{-y}) dx_2 \\ &= 1 - e^{-y} - ye^{-y} \quad \text{für } y > 0 \end{aligned}$$

Durch Ableiten bekommt man die Dichte von Y :

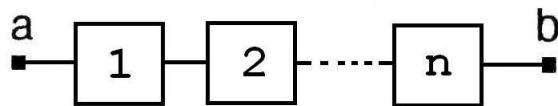
$$f_Y(y) = F'_Y(y) = e^{-y} - (e^{-y} - ye^{-y}) = ye^{-y} \quad \text{für } y > 0$$

Das entspricht einer **Gam(2, 1)**-Verteilung. Man beachte, dass hier X_1 und X_2 unabhängig nach **Exp(1)** verteilt sind.

Bem: Die Bestimmung der Verteilung der Summe $Y = X_1 + X_2$ von zwei (unabhängigen) sGn X_1 und X_2 nennt man **Faltung**. (Vgl. Kapitel 6 für eine ausführlichere Diskussion.)

■

Bsp 5.16 [Minimum] Die sGn X_1, X_2, \dots, X_n seien Lebensdauern von Komponenten in einem **Seriensystem**:



Das Seriensystem fällt aus, sobald die *erste* Komponente ausfällt. Die Lebensdauer Y_1 des Seriensystems ist also das **Minimum** der Lebensdauern der Komponenten:

$$Y_1 = \min\{X_1, X_2, \dots, X_n\}$$

Sind die Größen X_i , $i = 1, 2, \dots, n$, **unabhängig**, so ist die Verteilungsfunktion von Y_1 gegeben durch:

$$\begin{aligned} F_{\min}(y) &= P(Y_1 \leq y) = 1 - P(Y_1 > y) \\ &= 1 - P(X_1 > y, X_2 > y, \dots, X_n > y) \\ &= 1 - \prod_{i=1}^n P(X_i > y) \\ &= 1 - \prod_{i=1}^n [1 - P(X_i \leq y)] \end{aligned}$$

Bezeichnet F_i die Verteilungsfunktion von X_i , so gilt:

$$F_{\min}(y) = 1 - \prod_{i=1}^n [1 - F_i(y)]$$

Speziell: Für X_1, X_2, \dots, X_n iid $\text{Exp}(\lambda)$ (oder $\text{Exp}(\tau)$) ist F_{\min} gegeben durch:

$$F_{\min}(y) = 1 - \prod_{i=1}^n [1 - (1 - e^{-\lambda y})] = 1 - e^{-n\lambda y} \quad \text{für } y > 0$$

D. h. $Y_1 \sim \text{Exp}(n\lambda)$ (oder $\text{Exp}(\tau/n)$). Die Dichte von Y_1 lautet:

$$f_{\min}(y) = F'_{\min}(y) = n\lambda e^{-n\lambda y} \quad \text{für } y > 0$$

Die mittlere Lebensdauer einer Komponente beträgt $1/\lambda = \tau$ und die mittlere Lebensdauer des Seriensystems ist gegeben durch:

$$\mathbb{E}(Y_1) = \frac{1}{n\lambda} = \frac{\tau}{n}$$

Für X_i id³ $\text{Exp}(\lambda_i)$, $i = 1, 2, \dots, n$, ist F_{\min} gegeben durch:

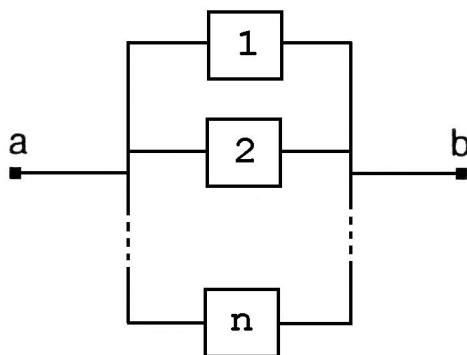
$$F_{\min}(y) = 1 - e^{-\tilde{\lambda}y} \quad \text{mit} \quad \tilde{\lambda} = \sum_{i=1}^n \lambda_i$$

D. h., $Y_1 \sim \text{Exp}(\tilde{\lambda})$ und für die mittlere Lebensdauer des Seriensystems gilt:

$$\mathbb{E}(Y_1) = \frac{1}{\tilde{\lambda}} = \frac{1}{\lambda_1 + \lambda_2 + \dots + \lambda_n}$$

■

Bsp 5.17 [Maximum] Die sGn X_1, X_2, \dots, X_n seien Lebensdauern von Komponenten in einem **Parallelsystem**:



³unabhängig (aber nicht notwendigerweise identisch verteilt)

Das Parallelsystem fällt erst aus, wenn *alle* Komponenten ausgefallen sind.⁴ Die Lebensdauer Y_n des Parallelsystems ist also das **Maximum** der Lebensdauern der Komponenten:

$$Y_n = \max\{X_1, X_2, \dots, X_n\}$$

Sind die Größen X_i , $i = 1, 2, \dots, n$, **unabhängig**, so ist die Verteilungsfunktion von Y_n gegeben durch:

$$F_{\max}(y) = P(Y_n \leq y) = P(X_1 \leq y, X_2 \leq y, \dots, X_n \leq y) = \prod_{i=1}^n F_i(y)$$

Dabei bezeichnet F_i wieder die Verteilungsfunktion von X_i .

Speziell: Für X_1, X_2, \dots, X_n iid $\text{Exp}(\lambda)$ (oder $\text{Exp}(\tau)$) ist F_{\max} gegeben durch:

$$F_{\max}(y) = \prod_{i=1}^n (1 - e^{-\lambda y}) = (1 - e^{-\lambda y})^n \quad \text{für } y > 0$$

Die Dichte von Y_n bekommt man durch Ableiten:

$$f_{\max}(y) = F'_{\max}(y) = n\lambda e^{-\lambda y} (1 - e^{-\lambda y})^{n-1} \quad \text{für } y > 0$$

Man beachte, dass die Lebensdauer des Parallelsystems *nicht* wieder exponentialverteilt ist. Mit Hilfe der *Gedächtnislosigkeit* der Exponentialverteilung (vgl. 4.2.2) lässt sich aber zeigen, dass die mittlere Lebensdauer von Y_n gegeben ist durch:

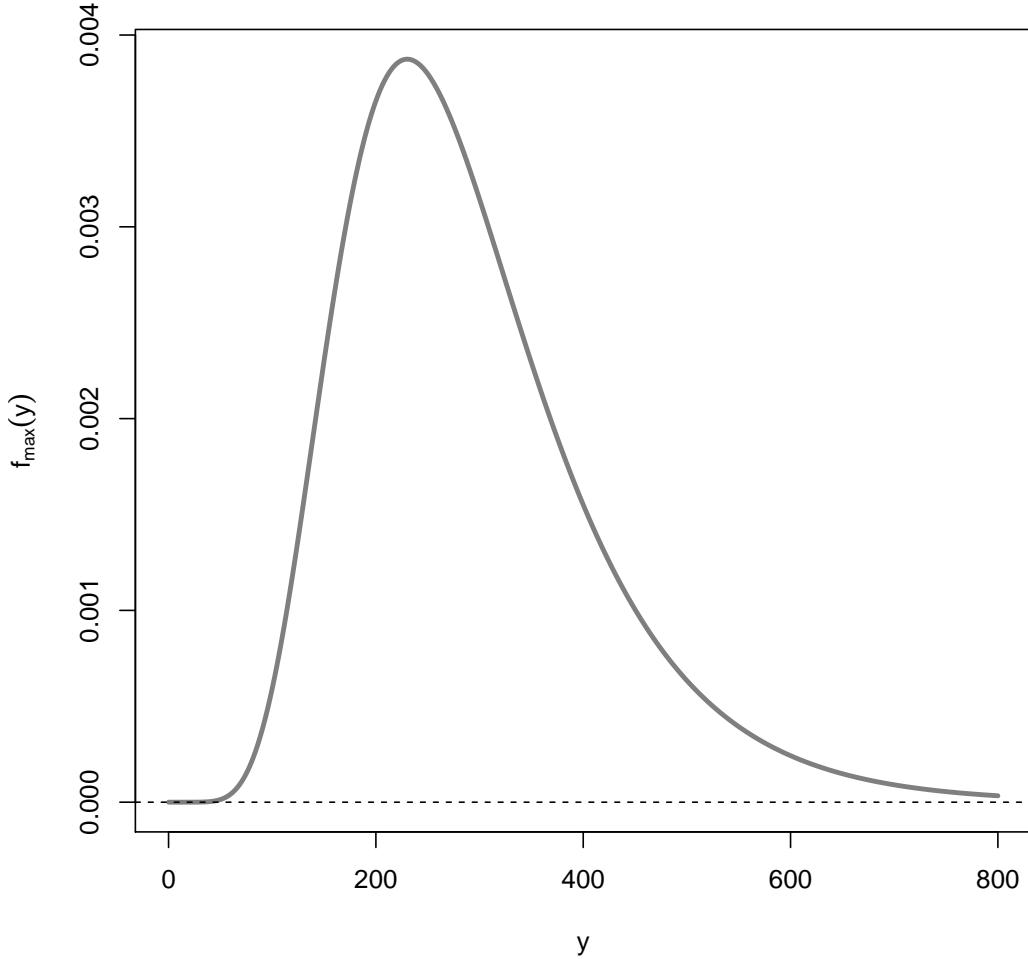
$$\mathbb{E}(Y_n) = \frac{1}{\lambda} \left(\frac{1}{n} + \frac{1}{n-1} + \dots + \frac{1}{2} + 1 \right)$$

Beispielsweise gilt für $n = 10$:

$$\mathbb{E}(Y_{10}) = \frac{1}{\lambda} \left(\frac{1}{10} + \frac{1}{9} + \dots + \frac{1}{2} + 1 \right) \approx \frac{2.93}{\lambda}$$

Abb 5.5 zeigt die Dichte von Y_n für $n = 10$ und $\lambda = 1/100$; die mittlere Lebensdauer des Parallelsystems beträgt in diesem Fall etwa 293 [ZE]. ■

⁴Ein Parallelsystem nennt man auch ein *vollständig redundantes* System.

Abbildung 5.5: Dichte des Maximums von $n = 10$ ua. $\text{Exp}(\lambda = 1/100)$ -Größen

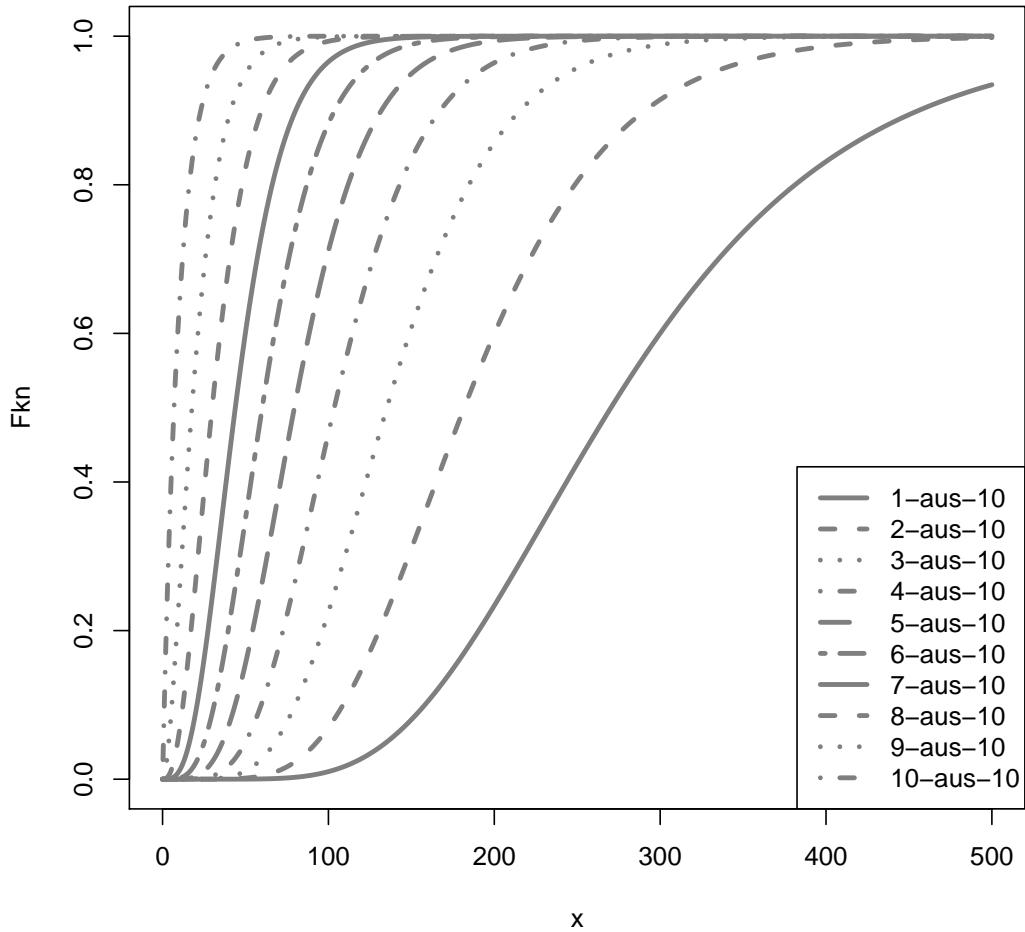
Bsp 5.18 [k-aus-n] Ein **k -aus- n -System** ist intakt, wenn *zumindest k* der insgesamt n Komponenten intakt sind. Die beiden vorhin betrachteten Systeme sind Spezialfälle: Ein Seriensystem ist ein n -aus- n -System und ein Parallelsystem ist ein 1-aus- n -System.

Bezeichnet $Y_{k|n}$ die Lebensdauer eines k -aus- n -Systems, so gilt für X_1, X_2, \dots, X_n iid F :

$$F_{k|n}(y) = P(Y_{k|n} \leq y) = \sum_{j=0}^{k-1} \binom{n}{j} [1 - F(y)]^j [F(y)]^{n-j}$$

Beweis: Die Wahrscheinlichkeit, dass eine Komponente nach (vor) y ausfällt, ist $1 - F(y)$ ($F(y)$). Da die Komponenten nach Voraussetzung unabhängig sind, ergibt sich der obige Ausdruck durch Anwendung der Binomialverteilung.

Abb 5.6 ist für $k = 1, 2, \dots, n$ eine vergleichende Darstellung der Verteilungsfunktionen von $Y_{k|n}$ für $n = 10$ ua. $\text{Exp}(\lambda = 1/100)$ -Komponenten. Die flachste Kurve entspricht dem Parallelsystem (1-aus-10), die steilste Kurve dem Seriensystem (10-aus-10). ■

Abbildung 5.6: VF von $Y_{k|n}$ für $n = 10$ ua. $\text{Exp}(\lambda = 1/100)$ -Größen

5.6 Spezielle multivariate Verteilungen

5.6.1 Multinomialverteilung

Ein Experiment bestehe aus n identischen und unabhängigen Versuchen, wobei jeder Versuch mit den (konstanten) Wahrscheinlichkeiten p_1, \dots, p_k , wobei $\sum_{i=1}^k p_i = 1$, auf eine von k Arten ausgehen kann.

Ist X_i , $i = 1, \dots, k$, die Anzahl der Versuche, die auf die i -te Art ausgehen, so hat der stochastische Vektor (X_1, X_2, \dots, X_k) eine **Multinomialverteilung** $M(n, p_1, p_2, \dots, p_k)$ mit der W-Funktion:

$$p(x_1, \dots, x_k) = \binom{n}{x_1, x_2, \dots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \quad \text{mit} \quad \sum_{i=1}^k x_i = n$$

Dabei ist der **Multinomialkoeffizient** gegeben durch:

$$\binom{n}{x_1, x_2, \dots, x_k} = \frac{n!}{x_1! x_2! \cdots x_k!} \quad \text{mit} \quad \sum_{i=1}^k x_i = n$$

Bem: Man beachte, dass es sich – entgegen der Schreibweise – wegen $\sum_{i=1}^k x_i = n$ nur um eine $(k - 1)$ -dimensionale Verteilung handelt.

Für $k = 2$ ergibt sich die **Binomialverteilung** $M(n, p_1, p_2) \equiv B(n, p_1)$. Man schreibt $p_1 = p$ und $p_2 = 1 - p$ und die W-Funktion lautet (vgl. 4.1.3)

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \text{für } x = 0, 1, 2, \dots, n$$

Für $k = 3$ ergibt sich die **Trinomialverteilung** $M(n, p_1, p_2, p_3)$ (mit $p_1 + p_2 + p_3 = 1$). Man schreibt $X = X_1$, $Y = X_2$ (und $Z = n - X_1 - X_2$) und die W-Funktion von (X, Y) ist gegeben durch:

$$p(x, y) = \frac{n!}{x!y!(n-x-y)!} p_1^x p_2^y (1-p_1-p_2)^{n-x-y}$$

Dabei sind x und y Zahlen aus $\{0, 1, 2, \dots, n\}$ mit $x + y \leq n$.

Randverteilungen: Die Randverteilungen der Multinomialverteilung sind wieder Multinomialverteilungen. Insbesondere ergibt sich für $(X_1, X_2, \dots, X_k) \sim M(n, p_1, p_2, \dots, p_k)$:

- (1) $X_i \sim M(n, p_i, 1 - p_i) \equiv B(n, p_i)$
- (2) $(X_i, X_j) \sim M(n, p_i, p_j, 1 - p_i - p_j) \quad (\text{für } i < j)$

Beweis: Ergibt sich aus der zu Beginn dieses Abschnitts gegebenen *inhaltlichen* Interpretation der Multinomialverteilung.

Erwartungswert/Varianz: Wegen $X_i \sim B(n, p_i)$ sind Erwartungswert und Varianz von X_i gegeben durch (vgl. 4.1.3):

$$\mathbb{E}(X_i) = np_i, \quad \text{Var}(X_i) = np_i(1 - p_i)$$

Kovarianz/Korrelation: Für $i \neq j$ gilt:

$$\text{Cov}(X_i, X_j) = -np_i p_j, \quad \rho_{X_i X_j} = -\sqrt{\frac{p_i}{1-p_i}} \sqrt{\frac{p_j}{1-p_j}}$$

Man beachte, dass die Größen X_i nicht unabhängig sind und negativ korrelieren.

Bsp 5.19 Drei Kugeln werden zufällig und *mit* Zurücklegen aus einem Behälter, bestehend aus 3 roten, 4 weißen und 5 blauen Kugeln, entnommen, und X bzw. Y sei die Zahl der roten bzw. weißen Kugeln in der Stichprobe.

Die gemeinsame Verteilung von X , Y (und $Z = 3 - X - Y$) ist eine Trinomialverteilung, $M(n = 3, p_1, p_2, p_3)$ wobei:

$$p_1 = \frac{3}{12}, \quad p_2 = \frac{4}{12}, \quad p_3 = \frac{5}{12}$$

Die Randverteilungen von X , Y und Z sind Binomialverteilungen:

$$X \sim B(3, p_1), \quad Y \sim B(3, p_2), \quad Z \sim B(3, p_3)$$

Der Korrelationskoeffizient von X und Y ist gegeben durch:

$$\rho_{XY} = -\sqrt{\frac{3/12}{1 - 3/12}} \sqrt{\frac{4/12}{1 - 4/12}} = -\frac{1}{\sqrt{6}}$$

■

5.6.2 Polyhypergeometrische Verteilung

Unter N (gleichartigen) Objekten gebe es A_i Objekte der i -ten Art, $i = 1, \dots, k$, wobei $\sum_{i=1}^k A_i = N$. Werden zufällig n Objekte *ohne* Zurücklegen gezogen und ist X_i , $i = 1, \dots, k$, die Zahl der dabei erhaltenen Objekte der i -ten Art, so hat der stochastische Vektor (X_1, X_2, \dots, X_k) eine **Polyhypergeometrische Verteilung** $H(N, A_1, A_2, \dots, A_k, n)$ mit der W-Funktion:

$$p(x_1, \dots, x_k) = \frac{\binom{A_1}{x_1} \binom{A_2}{x_2} \cdots \binom{A_k}{x_k}}{\binom{N}{n}} \quad \text{mit} \quad \sum_{i=1}^k x_i = n$$

Bem: Man beachte, dass es sich – entgegen der Schreibweise – wegen $\sum_{i=1}^k x_i = n$ nur um eine $(k - 1)$ -dimensionale Verteilung handelt.

Für $k = 2$ ergibt sich die (übliche) **Hypergeometrische Verteilung** $H(N, A_1, A_2, n) \equiv H(N, A_1, n)$. Man schreibt $A_1 = A$ und $A_2 = N - A$ und die W-Funktion lautet (vgl. 4.1.6):

$$p(x) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}} \quad \text{für } x \in \{ \max\{0, n+A-N\}, \dots, \min\{A, n\} \}$$

Randverteilungen: Die Randverteilungen der Polyhypergeometrischen Verteilung sind wieder Polyhypergeometrische Verteilungen. Insbesondere ergibt sich für $(X_1, X_2, \dots, X_k) \sim \mathsf{H}(N, A_1, A_2, \dots, A_k, n)$:

- (1) $X_i \sim \mathsf{H}(N, A_i, N - N_i, n) \equiv \mathsf{H}(N, A_i, n)$
- (2) $(X_i, X_j) \sim \mathsf{H}(N, A_i, A_j, N - A_i - A_j, n) \quad (\text{für } i < j)$

Beweis: Ergibt sich aus der zu Beginn dieses Abschnitts gegebenen *inhaltlichen* Interpretation der Polyhypergeometrischen Verteilung.

Erwartungswert/Varianz: Wegen $X_i \sim \mathsf{H}(N, A_i, n)$ sind Erwartungswert und Varianz von X_i gegeben durch (vgl. 4.1.6):

$$\mathbb{E}(X_i) = n \frac{A_i}{N}, \quad \text{Var}(X_i) = n \frac{A_i}{N} \left(1 - \frac{A_i}{N}\right) \frac{N-n}{N-1}$$

Kovarianz/Korrelation: Für $i \neq j$ gilt:

$$\text{Cov}(X_i, X_j) = -n \frac{A_i}{N} \frac{A_j}{N} \frac{N-n}{N-1}, \quad \rho_{X_i X_j} = -\sqrt{\frac{A_i}{N-A_i}} \sqrt{\frac{A_j}{N-A_j}}$$

Man beachte, dass die Größen X_i nicht unabhängig sind und negativ korrelieren. (Bem: Der Korrekturfaktor für endliche Grundgesamtheiten $(N-n)/(N-1)$ kürzt sich im Ausdruck für ρ heraus.)

Bsp 5.20 Erfolgen die Ziehungen in der Situation von Bsp 5.19 *ohne* Zurücklegen, sind X, Y und $Z = 3 - X - Y$ gemeinsam polyhypergeometrisch $\mathsf{H}(12, 3, 4, 5, 3)$ verteilt. Der Korrelationskoeffizient von X und Y ist gegeben durch:

$$\rho_{XY} = -\sqrt{\frac{3}{12-3}} \sqrt{\frac{4}{12-4}} = -\frac{1}{\sqrt{6}}$$

Man beachte, dass sich der gleiche Korrelationskoeffizient wie bei Ziehungen *mit* Zurücklegen ergibt. Das zeigt sich auch daran, dass man $\rho_{X_i X_j}$ für die Polyhypergeometrische Verteilung auch wie folgt schreiben kann:

$$\rho_{X_i X_j} = -\sqrt{\frac{A_i/N}{1-A_i/N}} \sqrt{\frac{A_j/N}{1-A_j/N}}$$

Setzt man $p_i = A_i/N$ und $p_j = A_j/N$, entspricht der obige Ausdruck dem Korrelationskoeffizienten für die Multinomialverteilung. ■

5.6.3 Multivariate Normalverteilung

Neben der (univariaten) Normalverteilung (vgl. 4.2.4) ist auch ihre multivariate Verallgemeinerung von zentraler Bedeutung in Wahrscheinlichkeitstheorie und Statistik. (Bem: Beispielsweise basiert die klassische Regressionsanalyse auf der multivariaten Normalverteilung.)

Ein stochastischer Vektor $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ hat eine (n -dimensionale) **multivariate Normalverteilung**, $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, wenn seine Dichte gegeben ist durch:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \quad \text{für } \mathbf{x} = (x_1, x_2, \dots, x_n)' \in \mathbb{R}^n$$

Dabei ist $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)'$ ein Vektor aus \mathbb{R}^n und $\boldsymbol{\Sigma}$ ist eine symmetrische und positiv definite ($n \times n$)-Matrix.⁵

Im Spezialfall $\boldsymbol{\mu} = \mathbf{0}$ und $\boldsymbol{\Sigma} = \mathbf{I}_n$ (= n -dimensionale Einheitsmatrix) spricht man von einer (n -dimensionalen) **Standardnormalverteilung**. Wegen $|\mathbf{I}_n| = 1$ und $\mathbf{I}_n^{-1} = \mathbf{I}_n$ ist die Dichte von $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ ⁶ gegeben durch:

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} \exp \left(-\frac{1}{2} \mathbf{z}' \mathbf{z} \right) \quad \text{für } \mathbf{z} = (z_1, z_2, \dots, z_n)' \in \mathbb{R}^n$$

Erwartungswert/Varianz-Kovarianzmatrix: Für $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ gilt:

$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \quad \text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Mit $\mu_i = \mathbb{E}(X_i)$, $\sigma_i^2 = \text{Var}(X_i)$ und $\sigma_{ij} = \text{Cov}(X_i, X_j)$ ($= \sigma_{ji}$) für $i, j = 1, 2, \dots, n$.

⁵ $|\boldsymbol{\Sigma}|$ bezeichnet die Determinante und $\boldsymbol{\Sigma}^{-1}$ die Inverse von $\boldsymbol{\Sigma}$.

⁶ Üblicherweise bezeichnet man einen standardnormalverteilten sV mit \mathbf{Z} .

Die folgenden (ohne Beweis angegebenen) Behauptungen unterstreichen die Bedeutung der multivariaten Normalverteilung.

Affine Transformationen: \mathbf{X} habe eine $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ -Verteilung und $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$, mit \mathbf{A} eine $(m \times n)$ -Matrix und $\mathbf{b} \in \mathbb{R}^m$, sei eine affine Transformation von \mathbf{X} . Dann gilt:

$$\mathbf{Y} \sim N_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$$

Randverteilungen: Speziell folgt aus der obigen Behauptung, dass die Randverteilungen einer multivariaten Normalverteilung wieder multivariate Normalverteilungen sind. Sei beispielsweise \mathbf{X}_1 der Untervektor der ersten m Elemente von \mathbf{X} und \mathbf{X}_2 der Untervektor der restlichen $n - m$ Elemente:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

Partitioniert man $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ auf die gleiche Weise:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

so gilt:

$$\mathbf{X}_1 \sim N_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \quad \text{und} \quad \mathbf{X}_2 \sim N_{n-m}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

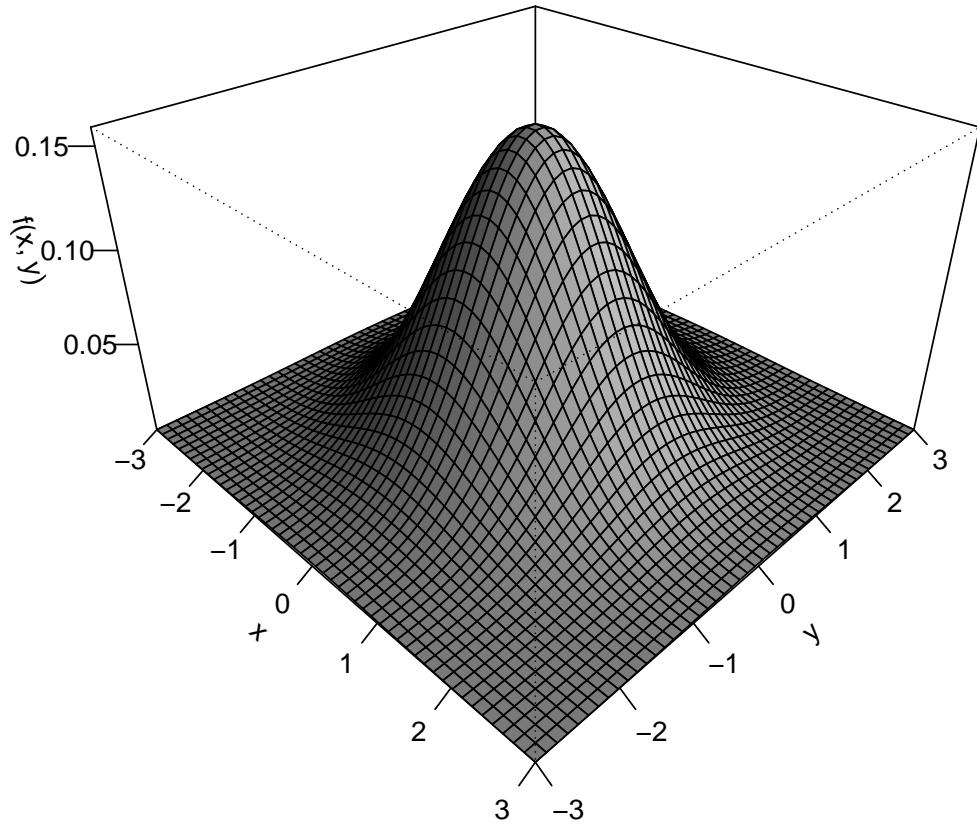
Bem: $\boldsymbol{\Sigma}_{11}$ ($\boldsymbol{\Sigma}_{22}$) ist die Varianz–Kovarianzmatrix von \mathbf{X}_1 (\mathbf{X}_2) und $\boldsymbol{\Sigma}_{12}$ ($= \boldsymbol{\Sigma}'_{21}$) umfasst alle (paarweisen) Kovarianzen der Komponenten von \mathbf{X}_1 und \mathbf{X}_2 .

Bsp 5.21 [Bivariate Normalverteilung] In diesem und im folgenden Beispiel spezialisieren wir die allgemeinen Überlegungen auf den Fall $n = 2$, d. h. auf die **bivariate Normalverteilung**. Schreibt man – wie üblich – (X, Y) statt (X_1, X_2) , so gilt für $(X, Y) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

Dabei ist μ_1 (σ_1^2) der Mittelwert (die Varianz) von X , μ_2 (σ_2^2) ist der Mittelwert (die Varianz) von Y und $\sigma_{12} = \sigma_{21} = \text{Cov}(X, Y)$ ist die Kovarianz von X und Y . Es gilt $\sigma_{12} = \rho\sigma_1\sigma_2$, wobei ρ der Korrelationskoeffizient von X und Y ist. Allgemein gilt $\rho^2 \leq 1$, im Folgenden nehmen wir aber an, dass $\rho^2 < 1$. In letzterem Fall ist $\boldsymbol{\Sigma}$ invertierbar (und positiv definit), und es gilt:

Abbildung 5.7: Bivariate Standardnormaldichte



$$|\Sigma| = \sigma_1^2 \sigma_2^2 (1 - \rho^2), \quad \Sigma^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho \sigma_1 \sigma_2 \\ -\rho \sigma_1 \sigma_2 & \sigma_1^2 \end{bmatrix}$$

Substituiert man diese Ausdrücke in der allgemeinen Formel für die multivariate Normaldichte, so bekommt man die Dichte der bivariaten Normalverteilung:

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2 \sqrt{1-\rho^2}} \exp\left(-\frac{q}{2}\right) \quad \text{für } (x, y) \in \mathbb{R}^2$$

wobei:

$$q = \frac{1}{1-\rho^2} \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x-\mu_1}{\sigma_1} \right) \left(\frac{y-\mu_2}{\sigma_2} \right) + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right]$$

Man schreibt im bivariaten Fall auch $(X, Y) \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Vgl. Abb 5.7 für eine grafische Darstellung der bivariaten Standardnormaldichte, d. h. von $N_2(0, 0, 1, 1, 0)$.

Die Randverteilungen sind gegeben durch $X \sim N(\mu_1, \sigma_1^2)$ und $Y \sim N(\mu_2, \sigma_2^2)$. (Bem: Man beachte, dass die Randverteilungen *nicht* von ρ abhängen.)

Die *Höhenschichtlinien* der bivariaten Normaldichte sind **Ellipsen**. (Bem: Für $\sigma_1 = \sigma_2$ und $\rho = 0$ handelt es sich um Kreise.) Abb 5.8 zeigt einige *Contourplots* der Normaldichte für verschiedene Werte von σ_1 , σ_2 und ρ . (In allen Fällen ist $\mu_1 = \mu_2 = 0$.)

Allgemein gilt, dass zwei unabhängige sGn X und Y auch unkorreliert sind (vgl. 5.3). Im Falle $(X, Y) \sim N_2$ gilt auch die **Umkehrung**: Aus der Unkorreliertheit folgt die Unabhängigkeit. Für $\rho = 0$ lässt sich $f(x, y)$ nämlich wie folgt schreiben:

$$\begin{aligned} f(x, y) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ -\frac{1}{2} \left[\left(\frac{x - \mu_1}{\sigma_1} \right)^2 + \left(\frac{y - \mu_2}{\sigma_2} \right)^2 \right] \right\} \\ &= \underbrace{\frac{1}{\sigma_1\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x - \mu_1}{\sigma_1} \right)^2 \right]}_{N(\mu_1, \sigma_1^2)} \times \underbrace{\frac{1}{\sigma_2\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{y - \mu_2}{\sigma_2} \right)^2 \right]}_{N(\mu_2, \sigma_2^2)} \\ &= f_1(x) \times f_2(y) \quad \text{für } (x, y) \in \mathbb{R}^2 \end{aligned}$$

Da sich die gemeinsame Dichte als Produkt der beiden Randdichten darstellen lässt, folgt die Unabhängigkeit von X und Y . ■

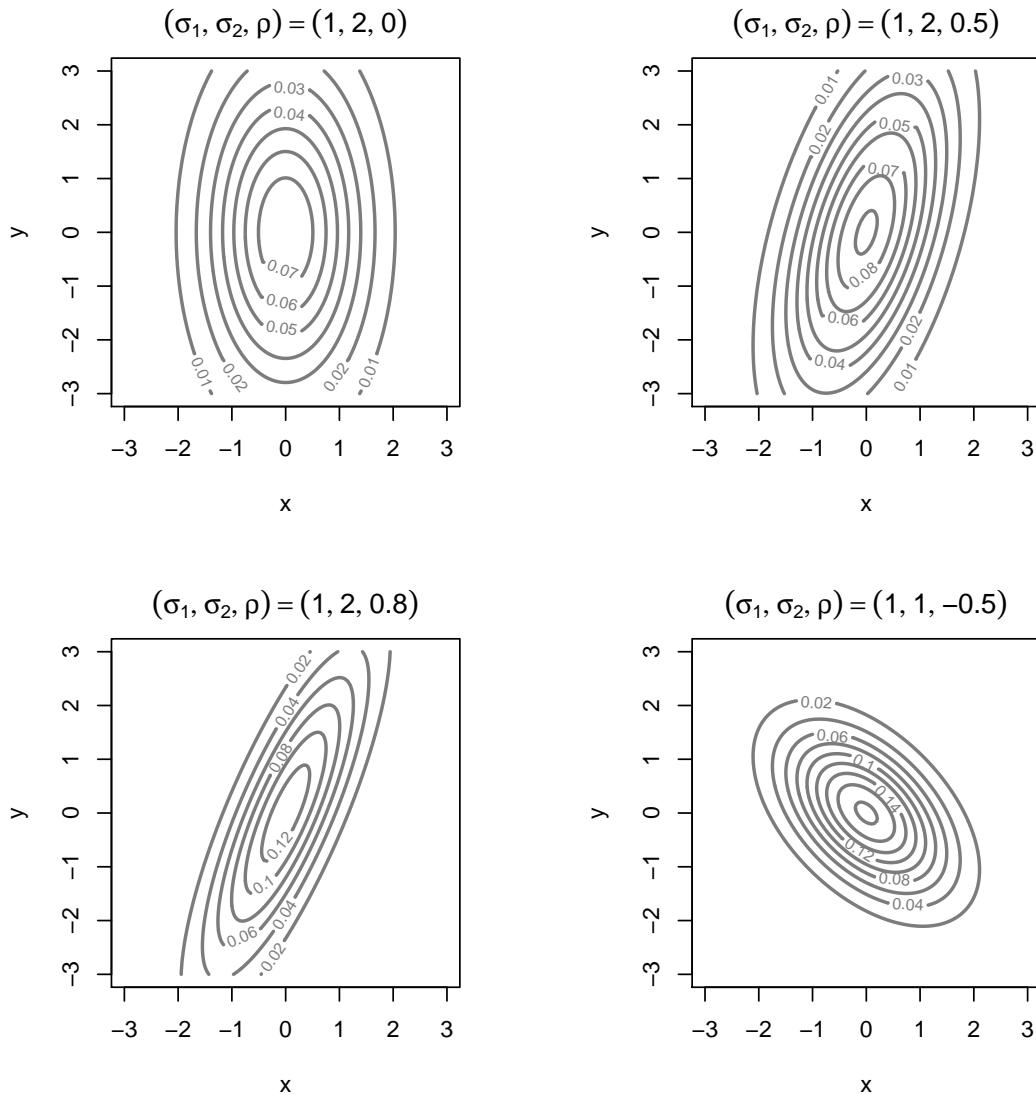
In Bsp 5.21 wurde gezeigt, dass für $(X, Y) \sim N_2$ aus $\rho_{XY} = 0$ auch die Unabhängigkeit von X und Y folgt. (Bem: Die Bedingung, dass X und Y *gemeinsam* normalverteilt sind, ist hier wesentlich; es genügt nicht, dass X und Y jeweils für sich normalverteilt sind.) Letzteres gilt allgemeiner; dazu legen wir wieder die für die **Randverteilungen** vorgenommene Partitionierung $\mathbf{X} = [\mathbf{X}'_1 \ \mathbf{X}'_2]'$ zugrunde.

Unabhängigkeit/Unkorreliertheit: Für $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ gilt, dass \mathbf{X}_1 und \mathbf{X}_2 genau dann unabhängig sind, wenn $\boldsymbol{\Sigma}_{12} = \mathbf{O}$ (d. h., wenn alle paarweisen Kovarianzen/Korrelationen zwischen den Komponenten von \mathbf{X}_1 und \mathbf{X}_2 gleich Null sind).

Nicht nur die Randverteilungen einer multivariaten Normalverteilung sind selbst wieder multivariate Normalverteilungen, auch die bedingten Verteilungen.

Bedingte Verteilungen: Legt man wieder die Partitionierung $\mathbf{X} = [\mathbf{X}'_1 \ \mathbf{X}'_2]'$ zugrunde, so gilt für $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2 \sim N_m(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

Abbildung 5.8: Contourplots von bivariaten Normaldichten

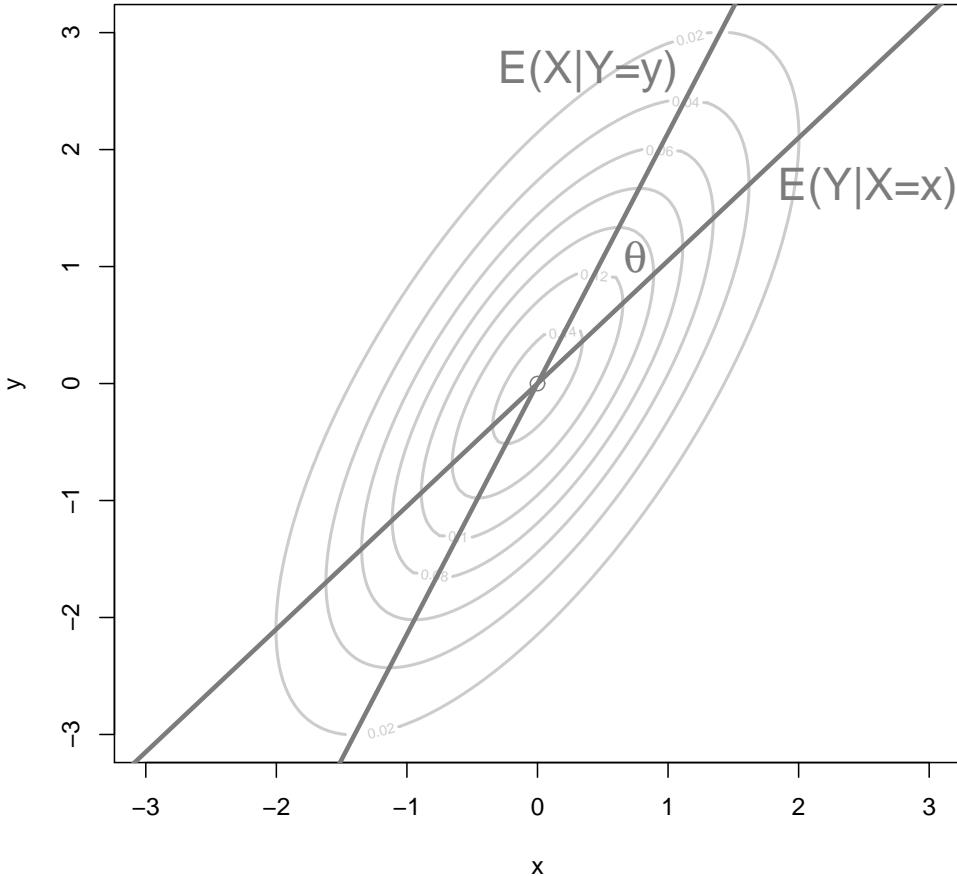
Bsp 5.22 [Bivariate Normalverteilung] Spezialisiert auf den Fall $n = 2$ bedeutet die obige Aussage über die bedingten Verteilungen, dass:

$$Y|X = x \sim N\left(\mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1), \sigma_2^2(1 - \rho^2)\right)$$

D. h., der durch $X = x$ bedingte Erwartungswert von Y ist eine Gerade:

$$\mathbb{E}(Y|x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1)$$

Diese Gerade nennt man auch die **Regressionsgerade** von Y auf X . Analog gilt für die durch $Y = y$ bedingte Verteilung von X :

Abbildung 5.9: Regressionsgeraden („Regressionsschere“)

$$X|Y = y \sim N\left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2), \sigma_1^2(1 - \rho^2)\right)$$

Die **Regressionsgerade** von X auf Y ist gegeben durch:

$$\mathbb{E}(X|y) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2)$$

Die beiden Regressionsgeraden („Regressionsschere“) schneiden sich im Punkt (μ_1, μ_2) . Vgl. Abb 5.9 für eine grafische Veranschaulichung für den Fall $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1.5$ und $\rho = 0.7$. (Bem: Die Schere klappt zusammen, wenn $|\rho| = 1$; sie hat maximale Öffnung (d. h. rechtwinkelig), wenn $\rho = 0$. In letzterem Fall verlaufen die Regressionsgeraden parallel zu den Koordinatenachsen.) Für den Tangens des Öffnungswinkels θ gilt:

$$\tan \theta = \frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2} \frac{1 - \rho^2}{\rho}$$

■

Aufgaben

5.1 Der Merkmalraum des stochastischen Vektors (X, Y) sei \mathbb{R}^2 . Betrachten Sie die folgenden Ereignisse und ihre Wahrscheinlichkeiten:

$$A_1 = \{(x, y) \mid x \leq 2, y \leq 4\}, \quad P(A_1) = 7/8$$

$$A_2 = \{(x, y) \mid x \leq 2, y \leq 1\}, \quad P(A_2) = 4/8$$

$$A_3 = \{(x, y) \mid x \leq 0, y \leq 4\}, \quad P(A_3) = 3/8$$

$$A_4 = \{(x, y) \mid x \leq 0, y \leq 1\}, \quad P(A_4) = 2/8$$

Bestimmen Sie die Wahrscheinlichkeit von $A_5 = \{(x, y) \mid 0 < x \leq 2, 1 < y \leq 4\}$.

5.2 Die gemeinsame W–Funktion von X und Y sei gegeben wie folgt:

$$p(1, 1) = 1/8, \quad p(1, 2) = 1/4$$

$$p(2, 1) = 1/8, \quad p(2, 2) = 1/2$$

- (a) Bestimmen Sie die Randverteilungen von X und Y .
- (b) Bestimmen Sie die durch $Y = i$, $i = 1, 2$, bedingte Verteilung von X .
- (c) Berechnen Sie $P(XY \leq 3)$, $P(X + Y > 2)$, $P(X/Y > 1)$.

5.3 Drei Kugeln werden zufällig und *ohne* Zurücklegen aus einem Behälter, bestehend aus 3 roten, 4 weißen und 5 blauen Kugeln, entnommen. X bzw. Y sei die Zahl der roten bzw. weißen Kugeln in der Stichprobe. Bestimmen Sie:

- (a) die gemeinsame Verteilung (2–dimensionale Tabelle) von X und Y .
- (b) die Randverteilungen von X und Y .

Wiederholen Sie (a) und (b) für Ziehungen *mit* Zurücklegen.

5.4 Die gemeinsame W–Funktion von X , Y und Z sei gegeben wie folgt:

$$p(1, 2, 3) = p(2, 1, 1) = p(2, 2, 1) = p(2, 3, 2) = 1/4$$

Berechnen Sie (a) $\mathbb{E}(XYZ)$ und (b) $\mathbb{E}(XY + XZ + YZ)$.

5.5 Die gemeinsame Dichte von X und Y sei gegeben durch:

$$f(x, y) = \frac{6}{7} \left(x^2 + \frac{xy}{2} \right) \quad \text{für } 0 < x < 1, \quad 0 < y < 2$$

- (a) Bestätigen Sie, dass es sich um eine Dichtefunktion handelt und geben Sie eine grafische Darstellung.
- (b) Bestimmen Sie die Randdichten von X und Y .
- (c) Berechnen Sie $P(X > Y)$.
- (d) Bestimmen Sie $\mathbb{E}(X)$ und $\mathbb{E}(Y)$.

5.6 Bestimmen Sie für Bsp 5.1:

- (a) die Kovarianz von X_1 und X_2 .
- (b) den Korrelationskoeffizienten von X_1 und X_2 .

5.7 Bestimmen Sie für Bsp 5.11:

- (a) die Kovarianz von $X (= X_1)$ und $Y (= X_2)$.
- (b) den Korrelationskoeffizienten von X und Y .
- (c) die Dichten von $X|Y = y$ und $Y|X = x$.
- (d) $\mathbb{E}(X|y)$ und $\mathbb{E}(Y|x)$.

5.8 Ein Punkt (X, Y) wird zufällig im Einheitskreis um den Nullpunkt gewählt.

- (a) Wie lautet die gemeinsame Dichte von (X, Y) ?
- (b) Bestimmen Sie die Randdichten von X und Y .
- (c) Sind X und Y unabhängig?
- (d) Zeigen Sie, dass die Kovarianz (und daher auch der Korrelationskoeffizient) von X und Y gleich Null ist.
- (e) $D = \sqrt{X^2 + Y^2}$ sei der Abstand des Punktes (X, Y) von $(0, 0)$. Bestimmen Sie die Verteilungsfunktion und die Dichte von D und berechnen Sie $\mathbb{E}(D)$.
(Hinweis: Bestimmen Sie die Verteilungsfunktion mit Hilfe einer geometrischen Überlegung.)

5.9 Angenommen, A macht sich zwischen 8:00 und 8:30 auf den Weg ins Büro und benötigt dazu zwischen 40 und 50 Minuten. X sei der Zeitpunkt des Aufbruchs und Y die benötigte Zeitspanne. Wenn diese sGn unabhängig und uniform verteilt sind, bestimmen Sie die Wahrscheinlichkeit, dass A vor 9:00 im Büro eintrifft.

5.10 Der Input eines Programms sei eine sG X mit Dichte $f_X(x) = e^{-x} I_{(0,\infty)}(x)$ (d. h. eine $\text{Exp}(1)$ -Verteilung). Bedingt durch $X = x$ sei die Ausführungszeit des Programms eine exponentialverteilte sG mit Mittelwert $1/x$. Bestimmen Sie die Dichte der Ausführungszeit Y des Programms. (**Hinweis:** Bestimmen Sie zuerst die gemeinsame Dichte von (X, Y) und anschließend die Randdichte von Y .)

5.11 Die Kantenlängen X, Y, Z eines Quaders seien unabhängige $U(0, 1)$ verteilte sGn. Bestimmen Sie den Erwartungswert und die Varianz des Volumens $V = XYZ$.
(Hinweis: Nehmen Sie für die Varianzberechnung den Verschiebungssatz.)

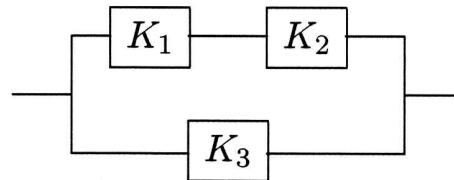
5.12 Die gemeinsame Dichte von X und Y sei gegeben durch:

$$f(x, y) = \begin{cases} Ce^{-(x+2y)} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{sonst} \end{cases}$$

- (a) Bestimmen Sie die Konstante C und stellen Sie die Dichte grafisch dar.
- (b) Bestimmen Sie die Randdichten von X und Y .
- (c) Sind X und Y unabhängig?
- (d) Bestimmen Sie $\mathbb{E}(Y|x)$ und $\mathbb{E}(X|y)$.
- (e) Bestimmen Sie die Dichte von $Z = X/Y$. (Hinweis: Bestimmen Sie zuerst die Verteilungsfunktion von Z .)

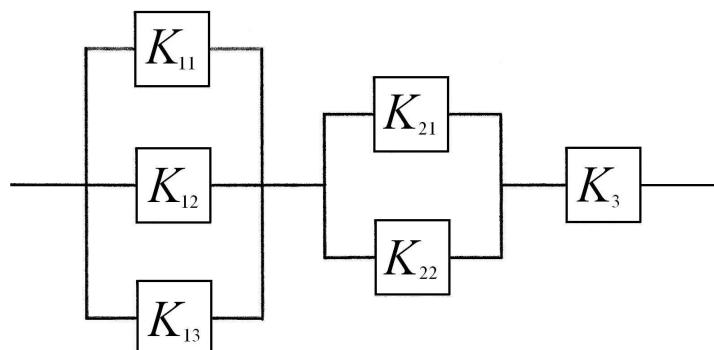
5.13 Ein Seriensystem bestehe aus drei Komponenten mit unabhängigen exponentialverteilten Lebensdauern mit den Mittelwerten 100, 200 bzw. 300 Stunden. Bestimmen Sie die Verteilungsfunktion und die Dichte der Lebensdauer des Systems sowie den Mittelwert und die Streuung.

5.14 Die logische Struktur eines Systems sei gegeben wie folgt:



Die Lebensdauern der Komponenten seien unabhängig und identisch verteilt mit Dichte $f(x) = e^{-x} I_{(0,\infty)}(x)$. Bestimmen Sie die Verteilungsfunktion und die Dichte der Lebensdauer des Systems sowie den Mittelwert.

5.15 Für die Komponenten des folgenden Systems gelte: Die Lebensdauern der Komponenten der ersten Parallelgruppe sind exponentialverteilt mit Mittelwert 1000 Stunden, die der zweiten Parallelgruppe sind exponentialverteilt mit Mittelwert 3000 Stunden, und die Lebensdauer der letzten Serienkomponente ist exponentialverteilt mit Mittelwert 5000 Stunden. Alle Lebensdauern seien unabhängig.



- (a) Bestimmen Sie einen Ausdruck für die Verteilungsfunktion der Lebensdauer des Systems.
- (b) Simulieren Sie die Systemlebensdauer mehrere tausend Mal und stellen Sie das Ergebnis in Form eines (Dichte-) Histogramms dar.

Hinweis: Eine simulierte Lebensdauer für beispielsweise die erste Parallelgruppe lässt sich mittels `max(rexp(3, rate=1/1000))` erzeugen. Nehmen Sie eine `for` Schleife.

- 5.16 Der stochastische Vektor $\mathbf{X} = (X_1, X_2, X_3)'$ sei normalverteilt $N_3(\mathbf{0}, \Sigma)$, wobei:

$$\Sigma = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 3 \end{bmatrix}$$

Bestimmen Sie (a) die Verteilung von $Y = X_1 - 2X_2 + X_3$ und berechnen Sie (b) $P(Y^2 > 15.36)$.

- 5.17 In einem (amerikanischen) Lehrbuch findet sich die folgende Aufgabe: Angenommen, der Korrelationskoeffizient zwischen der Körpergröße des Mannes (X) und der Frau (Y) von verheirateten Paaren beträgt 0.70, und die mittlere Körpergröße des Mannes beträgt 5 ft. 10 in. mit der Standardabweichung 2 in., und die mittlere Körpergröße der Frau beträgt 5 ft. 4 in. mit der Standardabweichung $1\frac{1}{2}$ in. Wenn man von einer bivariaten Normalverteilung ausgeht:

- (a) Wie lautet die gemeinsame Verteilung der Körpergrößen in der Einheit cm?
 (Hinweis: 1 ft. = 12 in. = 30.48 cm, 1 in. = 2.54 cm)
- (b) Welche Größe würden Sie für die Frau prognostizieren, wenn der Mann 6 ft. groß ist? (Hinweis: Betrachten Sie $E(Y|x)$.)
- (c) Bestimmen und zeichnen Sie die beiden Regressionsgeraden. (Wie sind diese Geraden zu interpretieren?)

- 5.18 X und Y seien bivariat normalverteilt mit den Parametern $\mu_1 = 5$, $\mu_2 = 10$, $\sigma_1^2 = 1$, $\sigma_2^2 = 25$ und $\rho > 0$. Wenn $P(4 < Y < 16 | X = 5) = 0.954$, bestimmen Sie ρ .

- 5.19 Mit Hilfe der (eigenen) Funktion `biv.rnorm()` lassen sich bivariat normalverteilte Beobachtungen simulieren. Erzeugen Sie mit dieser Funktion $n = 500$ Beobachtungen einer (a) $N_2(0, 0, 1, 1, 0)$, einer (b) $N_2(100, 200, 25, 36, 0.8)$ und einer (c) $N_2(100, 200, 25, 36, -0.6)$ Verteilung und stellen Sie die Ergebnisse mittels Scatterplot (vgl. 1.9.1) grafisch dar. Verwenden Sie für eine erweiterte grafische Darstellung auch die Funktionen `scatter.with.hist()` und `scatter.with.box()` (vgl. den R-Code zu Kapitel 5).

- 5.20 Erzeugen Sie $n = 500$ Beobachtungen des (3-dim.) normalverteilten stochastischen Vektors $\mathbf{X} = (X_1, X_2, X_3)'$ von Aufgabe 5.16 und stellen Sie das Ergebnis grafisch dar. (Hinweis: Nehmen Sie für die Simulation die Funktion `mvrnorm()` aus dem Package MASS, und für die grafische Darstellung `pairs()`.)

6 Folgen von stochastischen Größen

In diesem Kapitel betrachten wir Folgen von (unabhängigen) stochastischen Größen, X_1, X_2, \dots, X_n , und speziell die wichtige Klasse der **linearen Funktionen**, d. h. Funktionen der Form $T = \sum_{i=1}^n a_i X_i$, etwas genauer. Dabei untersuchen wir nicht nur die Eigenschaften dieser Funktionen für festes n (d. h. Erwartungswert, Varianz, Verteilung), sondern auch das **Konvergenzverhalten** für $n \rightarrow \infty$. Da es sich dabei aber um **stochastische Folgen** handelt, sind die aus der Analysis bekannten Konvergenzbegriffe entsprechend zu adaptieren bzw. zu erweitern.

6.1 Lineare Funktionen

Für stochastische Größen X_1, X_2, \dots, X_n und (reelle) Konstanten a_1, a_2, \dots, a_n nennt man eine Funktion der Form:

$$T = \sum_{i=1}^n a_i X_i$$

eine **lineare Funktion** (oder eine **Linearkombination**) von X_1, X_2, \dots, X_n .

Bem: Im Folgenden wird generell vorausgesetzt, daß alle betrachteten Erwartungswerte, Varianzen und Kovarianzen existieren und endlich sind.

Behauptung 1: Sei $T = \sum_{i=1}^n a_i X_i$ eine lineare Funktion von X_1, X_2, \dots, X_n , dann gilt:

$$\mathbb{E}(T) = \sum_{i=1}^n a_i \mathbb{E}(X_i)$$

Beweis: Ergibt sich unmittelbar aus der Linearität des Erwartungswerts.

Um die Varianz einer linearen Funktion zu bestimmen, betrachten wir zunächst ein allgemeines Resultat über die Kovarianz von *zwei* Linearkombinationen.

Behauptung 2: Für $T = \sum_{i=1}^n a_i X_i$ und $W = \sum_{j=1}^m b_j Y_j$ gilt:

$$\text{Cov}(T, W) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(X_i, Y_j)$$

Beweis: Nach Definition der Kovarianz von stochastischen Größen (vgl. 5.2) gilt:

$$\begin{aligned}\text{Cov}(T, W) &= \mathbb{E} \left[\sum_{i=1}^n \sum_{j=1}^m (a_i X_i - a_i \mathbb{E}(X_i)) (b_j Y_j - b_j \mathbb{E}(Y_j)) \right] \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \underbrace{\mathbb{E}[(X_i - \mathbb{E}(X_i))(Y_j - \mathbb{E}(Y_j))]}_{=\text{Cov}(X_i, Y_j)}\end{aligned}$$

In der zweiten Gleichung wird Behauptung 1 verwendet.

Um die Varianz von T zu bestimmen, setzen wir in Behauptung 2 einfach $W = T$.

Folgerung 1: Für $T = \sum_{i=1}^n a_i X_i$ gilt:

$$\text{Var}(T) = \text{Cov}(T, T) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i < j} a_i a_j \text{Cov}(X_i, X_j)$$

Sind X_1, X_2, \dots, X_n unabhängig, so gilt $\text{Cov}(X_i, X_j) = 0$ für $i \neq j$. Das führt uns zur zweiten Folgerung.

Folgerung 2: Sind X_1, X_2, \dots, X_n unabhängig, so gilt:

$$\text{Var}(T) = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$$

Bem: Für die Gültigkeit von Folgerung 2 genügt die (paarweise) Unkorreliertheit von X_i und X_j für alle $i \neq j$.

Bsp 6.1 [Stichprobenmittelwert] Ist X_1, X_2, \dots, X_n eine iid-Folge von stochastischen Größen (d. h. eine „Stichprobe“; vgl. 7.1) mit dem Mittelwert μ und der Varianz σ^2 , so nennt man die folgende lineare Funktion:

$$\overline{X}_n := \sum_{i=1}^n \frac{1}{n} X_i = \frac{1}{n} \sum_{i=1}^n X_i$$

den **Stichprobenmittelwert** von X_1, X_2, \dots, X_n . Nach Behauptung 1 gilt:

$$\mathbb{E}(\overline{X}_n) = \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}(X_i)}_{=\mu} = \frac{n\mu}{n} = \mu$$

Nach Folgerung 2 gilt:

$$\text{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \underbrace{\text{Var}(X_i)}_{=\sigma^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Für wachsendes n strebt die Varianz von \bar{X}_n gegen Null. Da der Mittelwert aber konstant bleibt, hat es den Anschein, dass die Verteilung von \bar{X}_n für $n \rightarrow \infty$ gegen μ konvergiert. Diese Form der „Konvergenz“ wird später noch ausführlicher diskutiert. ■

Bsp 6.2 [Matchingproblem] Wir betrachten noch einmal das Matchingproblem (Bsp 3 von 2.16). Eine interessante Frage betrifft die zu *erwartende* Anzahl von Übereinstimmungen bei zwei zufälligen Permutationen von $1, 2, \dots, N$. Dazu definieren wir **Indikatorvariablen** für die Übereinstimmung an der i -ten Position:

$$X_i = \begin{cases} 1 & \text{Übereinstimmung an der } i\text{-ten Position} \\ 0 & \text{sonst} \end{cases} \quad \text{für } i = 1, 2, \dots, N$$

Da die Permutationen ganz zufällig erfolgen, gilt:

$$P(X_i = 1) = \frac{(N-1)!}{N!} = \frac{1}{N} \quad \text{und} \quad P(X_i = 0) = \frac{N-1}{N}$$

Der Erwartungswert von X_i ist also gegeben durch:

$$\mathbb{E}(X_i) = (1)P(X_i = 1) + (0)P(X_i = 0) = \frac{1}{N}$$

Nach Behauptung 1 gilt für die Zahl $X = \sum_{i=1}^N X_i$ der Übereinstimmungen:

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^N X_i\right) = \sum_{i=1}^N \mathbb{E}(X_i) = \frac{N}{N} = 1$$

Man kann also – unabhängig von N – genau *eine* Übereinstimmung erwarten.

Varianz von X : Die Berechnung der Varianz von X wird dadurch erschwert, dass die X_i nicht unabhängig sind. Zunächst gilt:

$$\text{Var}(X_i) = \frac{1}{N} \frac{N-1}{N} = \frac{N-1}{N^2}$$

Für die paarweisen Kovarianzen gilt nach dem Verschiebungssatz:

$$i < j : \quad \text{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i) \mathbb{E}(X_j)$$

Nun gilt $X_i X_j = 1$ (gleich 0 sonst) genau dann, wenn es an der i -ten und j -ten Position eine Übereinstimmung gibt:

$$\begin{aligned} P(X_i X_j = 1) &= \frac{(N-2)!}{N!} = \frac{1}{N(N-1)} \implies \mathbb{E}(X_i X_j) = \frac{1}{N(N-1)} \\ \implies \text{Cov}(X_i, X_j) &= \frac{1}{N(N-1)} - \left(\frac{1}{N}\right)^2 = \frac{1}{N^2(N-1)} \end{aligned}$$

Nach **Folgerung 1** gilt:

$$\begin{aligned} \text{Var}(X) &= \sum_{i=1}^N \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j) \\ &= \frac{N(N-1)}{N^2} + \frac{N(N-1)}{N^2(N-1)} \\ &= \frac{N-1}{N} + \frac{1}{N} = 1 \end{aligned}$$

Wir bekommen also das bemerkenswerte Resultat, dass – unabhängig von N – nicht nur der Erwartungswert sondern auch die Varianz der Anzahl der Überstimmungen bei zwei zufälligen Permutationen exakt gleich 1 ist. ■

6.2 Faltung

Möchte man die Verteilung von $X + Y$ aus den Verteilungen von zwei unabhängigen sGn X und Y bestimmen, spricht man von **Faltung**.¹ (Bem: Der Name bezieht sich auf den geometrischen Aspekt der Art und Weise, wie die Verteilung von $X + Y$ bestimmt wird.²)

6.2.1 Diskrete Faltung

X und Y seien zwei ua. diskret verteilte sGn mit den W-Funktionen p_X bzw. p_Y . Das Ereignis $\{X + Y = a\}$ lässt sich als Vereinigung von disjunkten Ereignissen darstellen:

$$\{X + Y = a\} = \bigcup_{(x,y): x+y=a} \{X = x, Y = y\}$$

¹engl. convolution

²Vgl. WIKIPEDIA für animierte Grafiken.

Aus der Additivität der W–Verteilung folgt:

$$p_{X+Y}(a) = P(X + Y = a) = \sum_{(x,y): x+y=a} P(X = x, Y = y)$$

Als Folge der vorausgesetzten Unabhängigkeit von X und Y gilt:

$$P(X = x, Y = y) = P(X = x)P(Y = y) = p_X(x)p_Y(y)$$

Bezeichnet M_X , M_Y bzw. M_{X+Y} den Merkmalraum von X , Y bzw. $X + Y$, so ist die W–Funktion von $X + Y$ gegeben durch:

$$p_{X+Y}(a) = \sum_{x \in M_X} p_X(x)p_Y(a - x) = \sum_{y \in M_Y} p_X(a - y)p_Y(y) \quad \text{für } a \in M_{X+Y}$$

Die obigen Summendarstellungen für p_{X+Y} nennt man auch das **Faltprodukt** von p_X und p_Y und man schreibt:

$$p_{X+Y} = p_X * p_Y = p_Y * p_X$$

Bsp 6.3 Als Beispiel für eine diskrete Faltung bestimmen wir die Verteilung der Summe von zwei ua. poissonverteilten sGn $X \sim \mathsf{P}(\lambda_1)$ und $Y \sim \mathsf{P}(\lambda_2)$. Das Faltprodukt von:

$$p_X(x) = \frac{\lambda_1^x e^{-\lambda_1}}{x!}, \quad x \in \mathbb{N}_0 \quad \text{und} \quad p_Y(y) = \frac{\lambda_2^y e^{-\lambda_2}}{y!}, \quad y \in \mathbb{N}_0$$

ist gegeben durch:

$$\begin{aligned} p_{X+Y}(a) &= \sum_{x=0}^a \frac{\lambda_1^x e^{-\lambda_1}}{x!} \frac{\lambda_2^{a-x} e^{-\lambda_2}}{(a-x)!} \\ &= e^{-(\lambda_1 + \lambda_2)} \sum_{x=0}^a \frac{\lambda_1^x \lambda_2^{a-x}}{x!(a-x)!} \\ &= \frac{e^{-(\lambda_1 + \lambda_2)}}{a!} \underbrace{\sum_{x=0}^a \frac{a!}{x!(a-x)!} \lambda_1^x \lambda_2^{a-x}}_{=(\lambda_1 + \lambda_2)^a} \\ &= \frac{(\lambda_1 + \lambda_2)^a e^{-(\lambda_1 + \lambda_2)}}{a!}, \quad a \in \mathbb{N}_0 \end{aligned}$$

D.h., $X + Y$ hat wieder eine Poissonverteilung mit Mittelwert $\lambda_1 + \lambda_2$. ■

6.2.2 Stetige Faltung

X und Y seien zwei unabhängige, stetig verteilte sGn mit den Dichten f_X bzw. f_Y . Zunächst bestimmen wir die Verteilungsfunktion von $X + Y$:

$$\begin{aligned} F_{X+Y}(a) &= P(X + Y \leq a) = \iint_{x+y \leq a} f_X(x)f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{a-y} f_X(x) dx \right] f_Y(y) dy \\ &= \int_{-\infty}^{\infty} F_X(a-y)f_Y(y) dy \end{aligned}$$

Durch Ableiten bekommt man die Dichte:

$$\begin{aligned} f_{X+Y}(a) &= \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y)f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \frac{d}{da} [F_X(a-y)] f_Y(y) dy \\ &= \int_{-\infty}^{\infty} f_X(a-y)f_Y(y) dy \end{aligned}$$

Aus Symmetriegründen gilt auch:

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(x)f_Y(a-x) dx$$

Die letzteren beiden Integraldarstellungen für f_{X+Y} nennt man auch das **Faltprodukt** von f_X und f_Y und man schreibt:

$$f_{X+Y} = f_X * f_Y = f_Y * f_X$$

Bsp 6.4 Als Beispiel bestimmen wir die Verteilung der Summe von zwei ua. stetig uniform verteilten sGn $X \sim U(0, 1)$ und $Y \sim U(0, 1)$. Das Faltprodukt von:

$$f_X(x) = I_{(0,1)}(x) \quad \text{und} \quad f_Y(y) = I_{(0,1)}(y)$$

ist gegeben durch:

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} I_{(0,1)}(a-y) I_{(0,1)}(y) dy \quad \text{für } a \in (0, 2)$$

Der Integrand ist genau dann gleich Eins (sonst gleich Null), wenn:

$$\left\{ 0 < a - y < 1 \quad \text{und} \quad 0 < y < 1 \right\} \iff \left\{ a - 1 < y < a \quad \text{und} \quad 0 < y < 1 \right\}$$

Die zweite Form der Bedingung legt die folgende Fallunterscheidung nahe:

$$(1) \quad 0 < a \leq 1: \quad f_{X+Y}(a) = \int_0^a dy = a$$

$$(2) \quad 1 < a < 2: \quad f_{X+Y}(a) = \int_{a-1}^1 dy = 2 - a$$

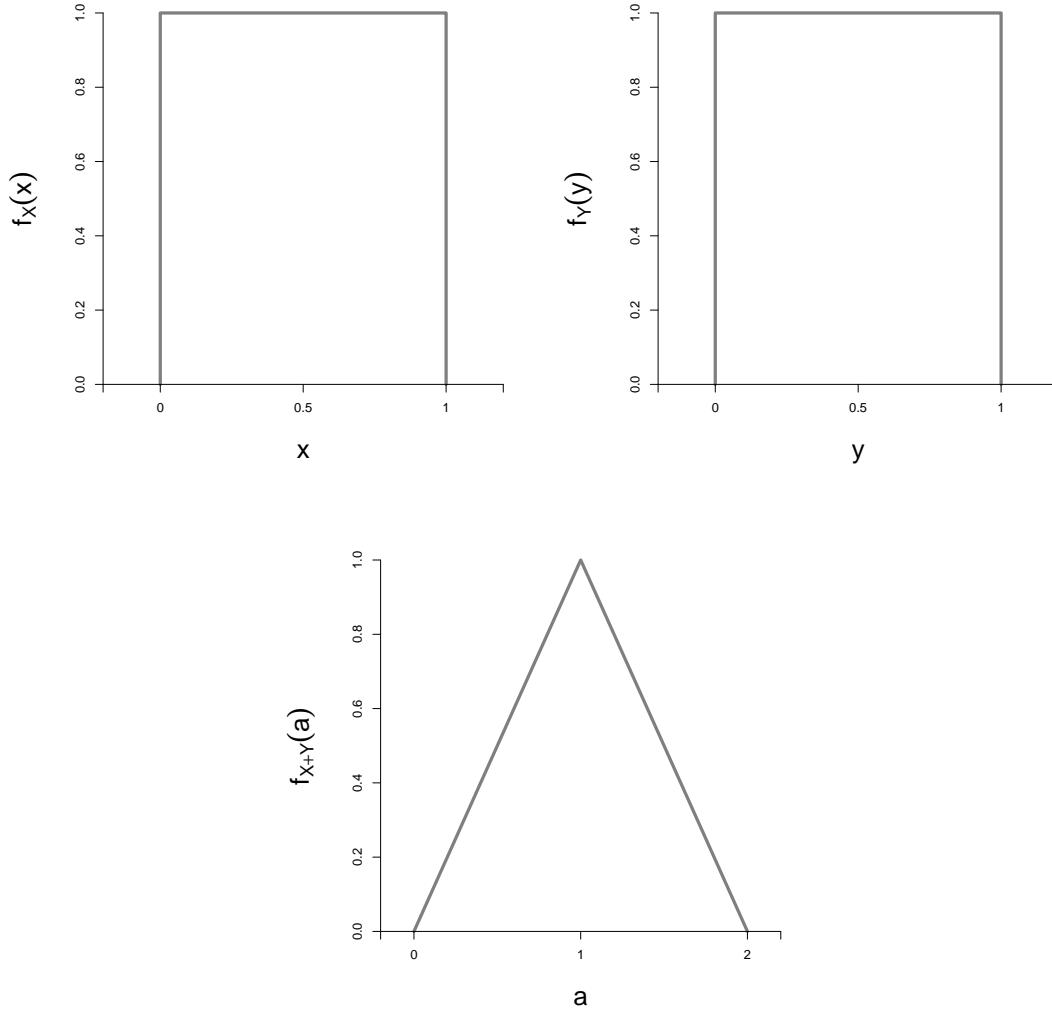
Zusammenfassung:

$$f_{X+Y}(a) = \begin{cases} a & 0 < a \leq 1 \\ 2 - a & 1 < a < 2 \\ 0 & \text{sonst} \end{cases}$$

Abb 6.1 ist eine grafische Darstellung der Faltung von X und Y . ■

6.2.3 Additionstheoreme

Die Faltung lässt sich unschwer von zwei auf mehrere stochastische Größen erweitern. Im Folgenden ein Überblick über die wichtigsten derartigen **Additionstheoreme**. Man beachte, dass vielfach bestimmte Einschränkungen notwendig sind, und dass in allen Fällen die Größen X_1, X_2, \dots, X_n als unabhängig vorausgesetzt werden.

Abbildung 6.1: Faltprodukt $U(0, 1) * U(0, 1)$ 

(1) Bernoulli-Verteilung:

$$X_i \sim A(p), \quad i = 1, 2, \dots, n, \text{ ua.} \quad \Rightarrow \quad \sum_{i=1}^n X_i \sim B(n, p)$$

(2) Binomialverteilung:

$$X_i \sim B(n_i, p), \quad i = 1, 2, \dots, n, \text{ ua.} \quad \Rightarrow \quad \sum_{i=1}^n X_i \sim B\left(\sum_{i=1}^n n_i, p\right)$$

(3) Poissonverteilung:

$$X_i \sim P(\lambda_i), \quad i = 1, 2, \dots, n, \text{ ua.} \quad \Rightarrow \quad \sum_{i=1}^n X_i \sim P\left(\sum_{i=1}^n \lambda_i\right)$$

(4) Geometrische Verteilung:

$$X_i \sim G(p), i = 1, 2, \dots, r, \text{ ua.} \implies \sum_{i=1}^r X_i \sim NB(r, p)$$

(5) Exponentialverteilung:

$$X_i \sim Exp(\lambda), i = 1, 2, \dots, n, \text{ ua.} \implies \sum_{i=1}^n X_i \sim Gam(n, \lambda)$$

Bem: Eine Gammaverteilung $Gam(\alpha, \lambda)$, deren Formparameter α aus \mathbb{N} ist, nennt man auch eine **Erlang-Verteilung**³ und schreibt $Er(n, \lambda)$ ($\equiv Gam(n, \lambda)$).

(6) Gammaverteilung:

$$X_i \sim Gam(\alpha_i, \lambda), i = 1, 2, \dots, n, \text{ ua.} \implies \sum_{i=1}^n X_i \sim Gam\left(\sum_{i=1}^n \alpha_i, \lambda\right)$$

(7) Chiquadratverteilung:

$$X_i \sim \chi^2(n_i), i = 1, 2, \dots, n, \text{ ua.} \implies \sum_{i=1}^n X_i \sim \chi^2\left(\sum_{i=1}^n n_i\right)$$

(8) Normalverteilung:

$$X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2, \dots, n, \text{ ua.} \implies \sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

Mit Konstanten a_1, a_2, \dots, a_n gilt etwas allgemeiner:

$$X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2, \dots, n, \text{ ua.} \implies \sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

6.3 Konvergenz

In diesem Abschnitt erweitern wir den aus der Analysis bekannten Konvergenzbegriff auf Folgen von stochastischen Größen und formulieren zwei klassische Theoreme der Wahrscheinlichkeitstheorie: das **Gesetz der großen Zahlen** und den **Zentralen Grenzwertungssatz**. Für beide Theoreme gibt es mehrere Versionen mit unterschiedlich starken Voraussetzungen. Wir betrachten jeweils nur die einfachste Version.

³AGNER KRARUP ERLANG (1878–1929), dänischer Mathematiker und Ingenieur; Beiträge zur Warteschlangentheorie (*Erlang-C-Formel*).

6.3.1 Ungleichungen

Die folgenden (klassischen) Ungleichungen sind nicht nur für sich von Interesse, sondern spielen auch eine große Rolle beim Beweis von Konvergenzaussagen.

Markow'sche Ungleichung:⁴ X sei eine nichtnegative sG (d. h. $X \geq 0$), deren Erwartungswert $\mathbb{E}(X)$ existiert. Dann gilt für $a > 0$:

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

Beweis: Wir betrachten nur den stetigen Fall:

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty xf(x) dx \\ &= \int_0^a xf(x) dx + \int_a^\infty xf(x) dx \\ &\geq \int_a^\infty xf(x) dx \\ &\geq \int_a^\infty af(x) dx \\ &= a \int_a^\infty f(x) dx \\ &= a P(X \geq a) \end{aligned}$$

Allgemeinere Form der Markow'schen Ungleichung: $u(X)$ sei eine nichtnegative Funktion der sG X (d. h., $u(X) \geq 0$). Existiert $\mathbb{E}[u(X)]$, dann gilt für $a > 0$:

$$P(u(X) \geq a) \leq \frac{\mathbb{E}[u(X)]}{a}$$

Tschebyschew'sche Ungleichung:⁵ Ist X eine sG mit Mittelwert μ und Varianz σ^2 , dann gilt für $k > 0$:

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

⁴ ANDREI ANDREJEWITSCH MARKOW (1856–1922), russ. Mathematiker (bedeutende Beiträge zur Wahrscheinlichkeitstheorie und Analysis).

⁵ PAFNUTI LWOWITSCH TSCHEBYSCHEW (richtiger: TSCHEBYSCHEW; 1821–1894), russ. Mathematiker (bedeutende Beiträge zu mehreren Gebieten der Mathematik und Physik).

Beweis: Da $(X - \mu)^2$ eine nichtnegative sG ist, lässt sich die Markow'sche Ungleichung anwenden:

$$P((X - \mu)^2 \geq k^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2}$$

Da $(X - \mu)^2 \geq k^2$ genau dann, wenn $|X - \mu| \geq k$, kann die obige Ungleichung auch wie folgt geschrieben werden:

$$P(|X - \mu| \geq k) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

Das war zu zeigen.

Äquivalente Formen der Tschebyschew'schen Ungleichung:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \text{oder} \quad P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Bsp 6.5 Angenommen, die Zahl der in einer Fabrik während einer Woche produzierten Einheiten ist eine sG mit Mittelwert 500. Was lässt sich über die Wahrscheinlichkeit sagen, mit der die Wochenproduktion zumindest 1000 Einheiten beträgt? Diese Frage lässt sich mit der Markow'schen Ungleichung beantworten:

$$P(X \geq 1000) \leq \frac{\mathbb{E}(X)}{1000} = \frac{500}{1000} = \frac{1}{2}$$

Wenn bekannt ist, dass die Streuung der wöchentlichen Produktionszahlen gleich 10 ist, was lässt sich über die Wahrscheinlichkeit sagen, mit der die wöchentliche Produktion zwischen 400 und 600 Einheiten liegt? Diese Frage lässt sich mit der Tschebyschew'schen Ungleichung beantworten:

$$P(|X - 500| \geq 100) \leq \frac{\sigma^2}{(100)^2} = \frac{100}{(100)^2} = \frac{1}{100}$$

Somit:

$$P(|X - 500| < 100) \geq 1 - \frac{1}{100} = \frac{99}{100}$$

D.h., die Wahrscheinlichkeit, mit der die wöchentliche Produktion zwischen 400 und 600 Einheiten liegt, beträgt mindestens 0.99. ■

Bsp 6.6 X sei eine diskrete sG mit dem Merkmalraum $M = \{-1, 0, 1\}$ und der W-Funktion:

$$p(-1) = p(1) = \frac{1}{8} \quad \text{und} \quad p(0) = \frac{6}{8}$$

In diesem Fall gilt:

$$\mathbb{E}(X) = 0 \quad \text{und} \quad \text{Var}(X) = \mathbb{E}(X^2) = \frac{2}{8} = \frac{1}{4}$$

Für $k = 2$ gilt:

$$P(|X - \mu| \geq k\sigma) = P(|X| \geq 1) = \frac{2}{8} = \frac{1}{4}$$

Andererseits gilt nach der Tschebyschew'schen Ungleichung:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} = \frac{1}{4}$$

D.h., $P(|X - \mu| \geq k\sigma)$ erreicht hier die obere Grenze der Ungleichung. Dieses Beispiel zeigt, dass die Tschebyschew'sche Ungleichung *scharf* ist, d.h., ohne zusätzliche Voraussetzungen nicht verbessert („verschärft“) werden kann. ■

6.3.2 Gesetz der großen Zahlen

Die Vorstellung, dass sich eine Folge von stochastischen Größen einer anderen stochastischen Größe „nähert“, lässt sich wie folgt formalisieren.

Stochastische Konvergenz: $\{X_n\}$ sei eine Folge von stochastischen Größen und X sei eine andere stochastische Größe. Dann **konvergiert X_n stochastisch** (oder **in der Wahrscheinlichkeit**⁶) gegen X , wenn für alle $\epsilon > 0$:

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \epsilon) = 0$$

Oder äquivalent:

$$\lim_{n \rightarrow \infty} P(|X_n - X| < \epsilon) = 1$$

Man schreibt in diesem Fall:

$$X_n \xrightarrow{P} X$$

⁶engl. *convergence in probability*

Vielfach ist X eine Konstante, d. h., die W–Verteilung von X konzentriert sich in einem Punkt a (d. h. $p_X(a) = 1$). In diesem Fall schreibt man:

$$X_n \xrightarrow{P} a$$

Behauptung: Angenommen, die Folge $\{X_n\}$ konvergiert in der Wahrscheinlichkeit gegen eine Konstante a , d. h. $X_n \xrightarrow{P} a$. Dann gilt für eine an der Stelle a stetige Funktion g :

$$g(X_n) \xrightarrow{P} g(a)$$

Die obige Behauptung hat viele nützliche Anwendungen. Beispielsweise ergeben sich aus $X_n \xrightarrow{P} a$ auch die folgenden Aussagen:

$$X_n^2 \xrightarrow{P} a^2, \quad 1/X_n \xrightarrow{P} 1/a \quad (\text{falls } a \neq 0), \quad \dots$$

Schwaches Gesetz der großen Zahlen (schGGZ): $\{X_n\}$ sei eine iid–Folge mit dem Mittelwert μ und der Varianz $\sigma^2 < \infty$. Sei $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ der Stichprobenmittelwert der ersten n Elemente der Folge. Dann gilt:

$$\bar{X}_n \xrightarrow{P} \mu$$

Beweis: Von Bsp 6.1 wissen wir, dass:

$$\mathbb{E}(\bar{X}_n) = \mu \quad \text{und} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

Nach der Tschebyschew'schen Ungleichung gilt für alle $\epsilon > 0$:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0 \quad \text{für } n \longrightarrow \infty$$

Das war zu zeigen.

Bemerkungen:

- (a) Anschaulich besagt das schwache GGZ, dass für großes n der Stichprobenmittelwert \bar{X}_n mit hoher Wahrscheinlichkeit in der Nähe von μ liegt. (Aber wie nahe? Mit dieser Frage beschäftigen wir uns im folgenden Abschnitt.)

- (b) Das erste GGZ (für Bernoulli–Größen) wurde von JAKOB (I.) BERNOULLI formuliert und bewiesen (*Ars Conjectandi*, posthum 1713). Da aber die Tschebyschew'sche Ungleichung zu seiner Zeit noch nicht bekannt war, beruht der Beweis auf einer Reihe von spitzfindigen Überlegungen.
- (c) Man kann im schwachen GGZ auf die Existenz der Varianz verzichten. In dieser Form wurde es vom russ. Mathematiker A. J. CHINTSCHIN (auch: KHINTCHINE; 1894–1959) bewiesen.
- (d) Es gibt auch ein *starkes* Gesetz der großen Zahlen: Ist $\{X_n\}$ eine iid–Folge mit dem Mittelwert μ , dann gilt:

$$P \left(\lim_{n \rightarrow \infty} \overline{X}_n = \mu \right) = 1$$

Diese Art der (stochastischen) Konvergenz nennt man *fast sichere Konvergenz*.

Bsp 6.7 Als Beispiel für das schwache GGZ betrachten wir eine diskrete sG X auf dem Merkmalraum $M = \{0, 1, 2, 3, 4\}$ mit der folgenden W–Funktion:

x	0	1	2	3	4
$p(x)$	0.1	0.2	0.3	0.35	0.05

Der Erwartungswert von X ist gegeben durch:

$$\mathbb{E}(X) = \sum_{x=0}^4 xp(x) = 2.05$$

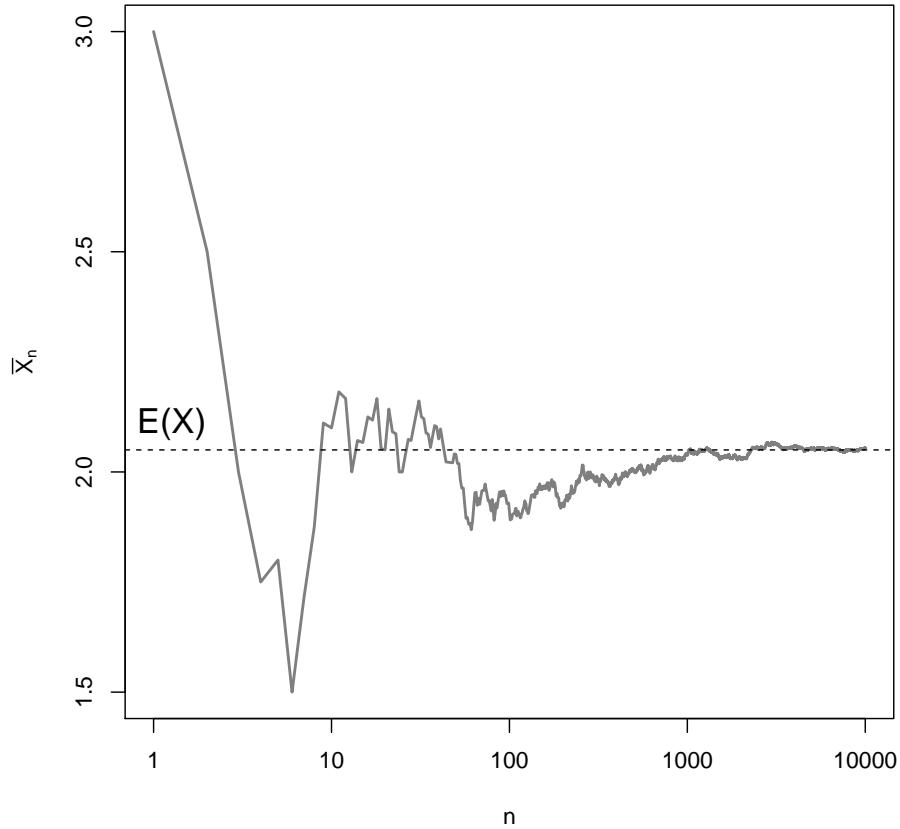
Da auch die Varianz von X existiert, besagt das schwache GGZ, dass:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}(X) = 2.05$$

Vgl. Abb 6.2 für die grafische Darstellung einer simulierten Folge von Stichprobenmittelwerten \overline{X}_n für $n = 1, 2, \dots, 10000$. ■

6.3.3 Zentraler Grenzverteilungssatz

Das schwache GGZ besagt, dass sich der Stichprobenmittelwert \overline{X}_n für wachsendes n dem Erwartungswert $\mu = \mathbb{E}(X)$ nähert. Lässt sich etwas über die Güte dieser Näherung aussagen? Zur Beantwortung dieser Frage benötigen wir einen weiteren Konvergenzbegriff.

Abbildung 6.2: Simulation zum schwachen GGZ

Konvergenz in der Verteilung: $\{X_n\}$ sei eine Folge von stochastischen Größen und X sei eine andere stochastische Größe. Sind F_{X_n} und F_X die Verteilungsfunktionen von X_n bzw. X und ist $C(F_X)$ die Menge aller Stetigkeitspunkte von F_X , so **konvergiert X_n in der Verteilung⁷** gegen X , wenn:

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \quad \text{für alle } x \in C(F_X)$$

Man schreibt in diesem Fall:

$$X_n \xrightarrow{D} X$$

Behauptung:⁸ Konvergiert X_n in der Wahrscheinlichkeit gegen X , so konvergiert X_n auch in der Verteilung gegen X :

$$X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$$

⁷engl. *convergence in distribution* (oder *in law*)

⁸Gilt $X_n \xrightarrow{D} a$ für eine Konstante a , so gilt auch die Umkehrung, d. h. $X_n \xrightarrow{P} a$.

Bem: Die obige Behauptung besagt, dass die Konvergenz in der Verteilung schwächer als die Konvergenz in der Wahrscheinlichkeit ist. Aus diesem Grund nennt man (in mathematischen Texten) die Konvergenz in der Verteilung auch die *schwache Konvergenz*.

Zentraler Grenzverteilungssatz (ZGVS): $\{X_n\}$ sei eine iid-Folge mit dem Mittelwert μ und der Varianz $\sigma^2 < \infty$. Dann konvergieren die Größen Y_n , definiert durch:

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

in der Verteilung gegen eine standardnormalverteilte stochastische Größe $Z \sim N(0, 1)$, d. h., für $n \rightarrow \infty$ gilt:

$$P(Y_n \leq z) \rightarrow \Phi(z) \quad \text{für alle } z \in \mathbb{R}$$

Historische Bemerkung: Die erste Version des ZGVS wurde vom franz. Mathematiker ABRAHAM DE MOIVRE (1667–1754) für $A(p = 1/2)$ -Größen bewiesen (1733) und später von PIERRE-SIMON DE LAPLACE auf allgemeines p erweitert. Laplace bewies auch den allgemeineren ZGVS in der obigen Form. (Sein Beweis hatte allerdings eine Lücke, die erst vom russ. Mathematiker und Physiker A. M. LJAPUNOW (1857–1918) um 1902 geschlossen wurde.)

Bsp 6.8 Wenn 10 symmetrische Würfel geworfen werden, mit welcher Wahrscheinlichkeit liegt dann die Augensumme zwischen 30 und 40 (inklusive)? Ist X_i die geworfene Augenzahl des i -ten Würfels, $i = 1, 2, \dots, 10$, so gilt:

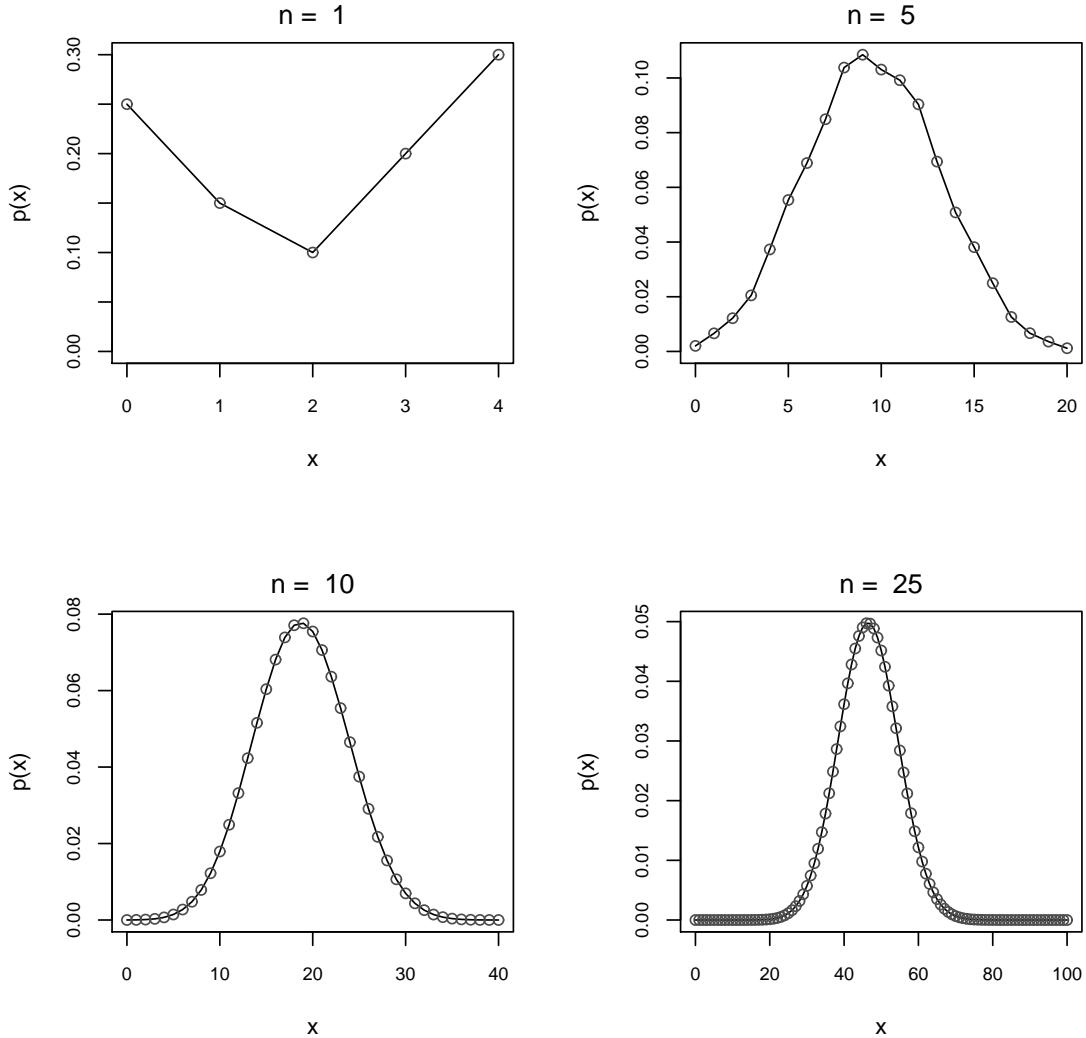
$$\mathbb{E}(X_i) = \frac{7}{2} \quad \text{und} \quad \text{Var}(X_i) = \frac{6^2 - 1}{12} = \frac{35}{12}$$

Nach dem ZGVS gilt für $X = \sum_{i=1}^{10} X_i$:

$$\begin{aligned} P(29.5 \leq X \leq 40.5) &= P\left(\frac{29.5 - 35}{\sqrt{350/12}} \leq \frac{X - 35}{\sqrt{350/12}} \leq \frac{40.5 - 35}{\sqrt{350/12}}\right) \\ &\approx 2\Phi\left(\frac{5.5}{\sqrt{350/12}}\right) - 1 \\ &= 2\Phi(1.0184) - 1 \\ &= 0.6915 \end{aligned}$$

■

Abbildung 6.3: Illustration zum ZGVS (Bsp 6.9)

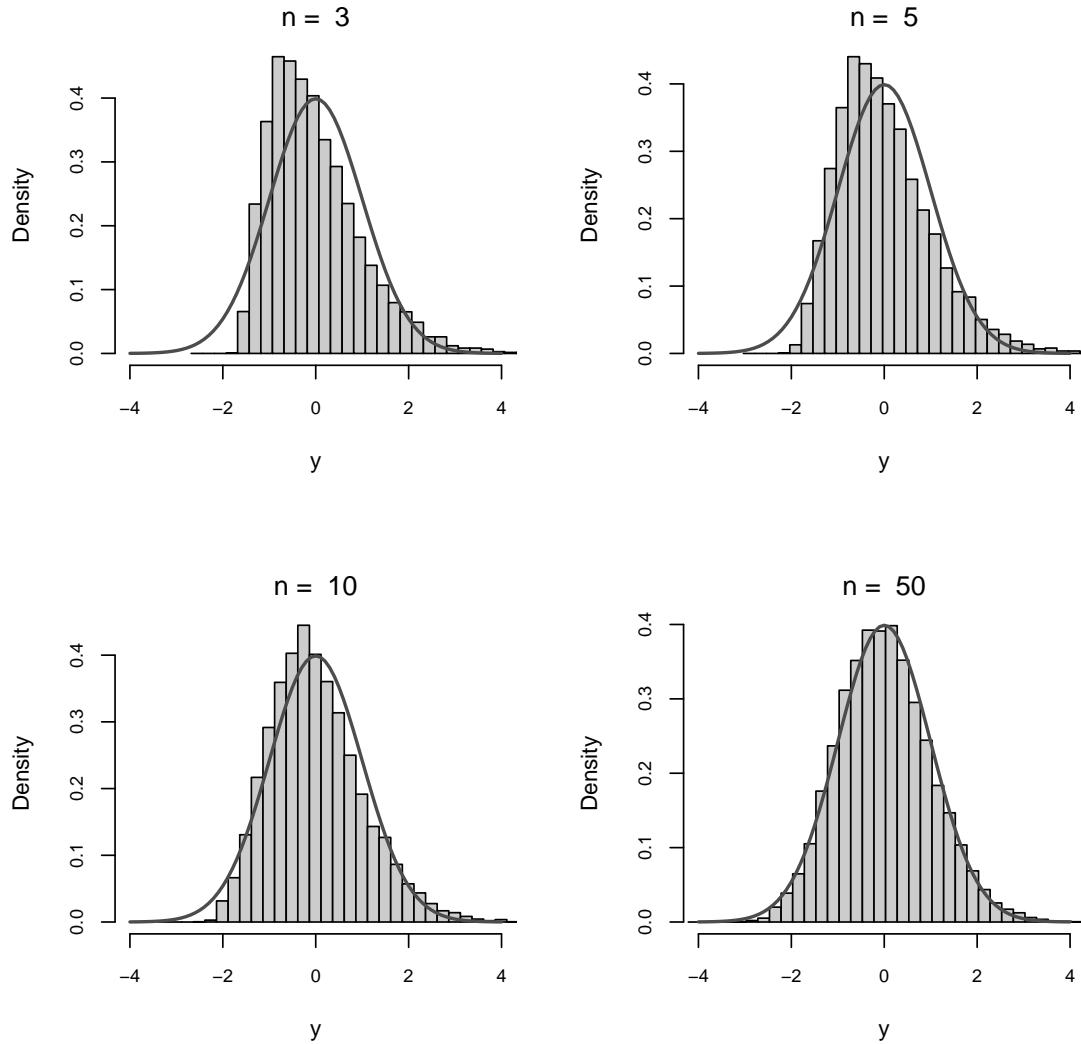


Bsp 6.9 Als Illustration des ZGVS für den diskreten Fall betrachten wir die W-Funktion der Summe $\sum_{i=1}^n X_i$ von n iid-Größen X_i mit dem Merkmalraum $M = \{0, 1, 2, 3, 4\}$ und der folgenden W-Funktion:

x	0	1	2	3	4
$p(x)$	0.25	0.15	0.1	0.2	0.3

Abb 6.3 zeigt das Ergebnis für $n = 1, 5, 10, 25$. Die Ausgangsverteilung ist noch weit von einer Normalverteilung entfernt, aber bereits für $n = 5$ ist die Form der Glockenkurve deutlich erkennbar, und für höhere n -Werte ist mit freiem Auge kein Unterschied mehr feststellbar. Das illustriert den durch den ZGVS ausgedrückten Sachverhalt, dass durch wiederholte Faltung die Ausgangsverteilung quasi „abgestreift“ wird. ■

Abbildung 6.4: Illustration zum ZGVS (Bsp 6.10)



Bsp 6.10 Als Illustration des ZGVS für den stetigen Fall betrachten wir die Dichte der standardisierten Summe Y_n von n iid–Größen mit $X_i \sim \text{Exp}(1)$:

$$Y_n = \frac{\sum_{i=1}^n X_i - n}{\sqrt{n}} = \sqrt{n}(\bar{X}_n - 1)$$

Abb 6.4 zeigt das Ergebnis für $n = 3, 5, 10, 50$ auf Basis von jeweils $N = 10000$ simulierten Werten für Y_n . Die darüber gezeichneten Kurven entsprechen der Dichte $\varphi(x)$ der $N(0, 1)$ –Verteilung. Da die Ausgangsverteilung hier sehr schief ist (vgl. Abb 4.8(b)), braucht es vergleichsweise große n –Werte, um ihren Einfluss auf die Faltung „abzustreifen“ (vgl. die Schlussbemerkung zu Bsp 6.9). Selbst für $n = 50$ macht sich die Schiefe der Exp–Verteilung noch bemerkbar. ■

6.3.4 Normalapproximation

Nach dem ZGVS lässt sich die Verteilung der Summe $\sum_{i=1}^n X_i$ von iid-Größen X_i (diskret oder stetig) mit Mittelwert μ und Varianz σ^2 für nicht zu kleines n in guter Näherung wie folgt durch eine Normalverteilung approximieren:

$$\sum_{i=1}^n X_i \approx N(n\mu, n\sigma^2)$$

Im diskreten Fall, d. h., wenn die X_i – und daher auch die Summe – diskrete sGn sind, lässt sich die obige Approximation häufig auf einfache Weise verbessern.

Stetigkeitskorrektur: Die Approximation einer diskreten Verteilung durch eine stetige Verteilung (insbesondere Normalverteilung) lässt sich häufig durch die **Stetigkeitskorrektur** verbessern. Ist X die diskrete und Y die stetige Größe, und besteht der Merkmalraum von X aus aufeinanderfolgenden ganzen Zahlen, lautet die Approximation unter Verwendung der Stetigkeitskorrektur wie folgt:

$$P(a \leq X \leq b) \approx P\left(a - \frac{1}{2} \leq Y \leq b + \frac{1}{2}\right)$$

D. h., am unteren Randpunkt von $[a, b]$ wird $1/2$ abgezogen, am oberen Rand addiert. (Bem: Diese Korrektur wurde bereits in **Bsp 6.8** verwendet.)

Normalapproximation der $B(n, p)$ -Verteilung: Nach Additionstheorem (1) von 6.2.3 hat die Summe $X_n = \sum_{i=1}^n X_i$ von n unabhängigen $A(p)$ -Größen X_i eine $B(n, p)$ -Verteilung. Unter Verwendung der Stetigkeitskorrektur gilt für $a \leq b$ ($a, b \in \{0, 1, \dots, n\}$):

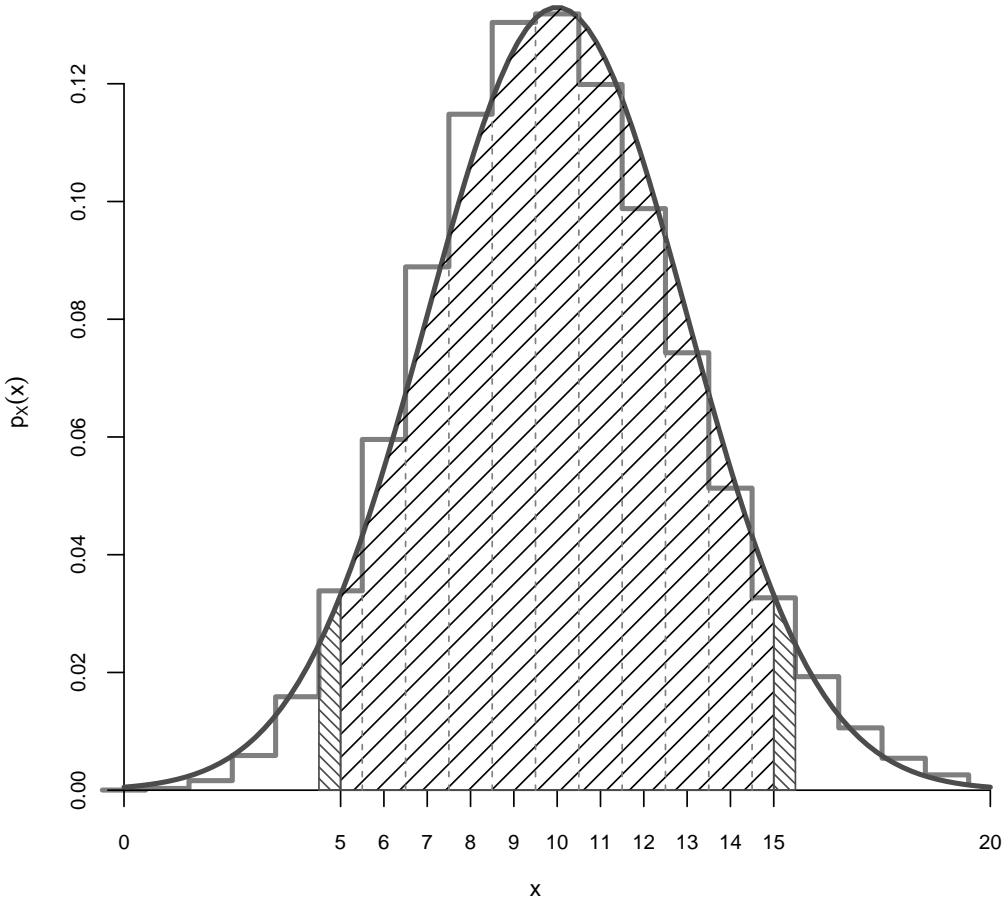
$$P(a \leq X_n \leq b) \approx \Phi\left(\frac{b + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 1/2 - np}{\sqrt{np(1-p)}}\right)$$

Approximation der Verteilungsfunktion von X_n :

$$P(X_n \leq x) \approx \Phi\left(\frac{x + 1/2 - np}{\sqrt{np(1-p)}}\right)$$

Nach einer gängigen Regel ist die Approximation ausreichend gut, wenn $np(1-p) \geq 10$.

Bsp 6.11 Angenommen, man möchte für $X \sim B(100, 0.1)$ die Wahrscheinlichkeit von $\{5 \leq X \leq 15\}$ mit Hilfe der Normalapproximation $X \approx N(np, np(1-p)) = N(10, 9)$ berechnen.

Abbildung 6.5: Normalapproximation der $B(100, 0.1)$ -Verteilung

ohne Korrektur: $P(5 \leq X \leq 15) \approx \Phi\left(\frac{15 - 10}{3}\right) - \Phi\left(\frac{5 - 10}{3}\right) \doteq 0.9044$

mit Korrektur: $P(5 \leq X \leq 15) \approx \Phi\left(\frac{15.5 - 10}{3}\right) - \Phi\left(\frac{4.5 - 10}{3}\right) \doteq 0.9332$

exakt: $P(5 \leq X \leq 15) = \sum_{x=5}^{15} \binom{100}{x} (0.1)^x (0.9)^{100-x} \doteq 0.9364$

Abb 6.5 zeigt die geometrischen Verhältnisse. Die dick gezeichnete Linie ist die Dichte der approximierenden Normalverteilung, der stärker schraffierte Bereich entspricht der Stetigkeitskorrektur. Wie auch an den numerischen Werten erkennbar, wird durch die Korrektur die Approximation deutlich verbessert. (Man beachte, dass hier die oben erwähnte Regel knapp nicht erfüllt ist.) ■

Normalapproximation der $P(\lambda)$ -Verteilung: Hat X eine $P(\lambda)$ -Verteilung, so gilt unter Verwendung der Stetigkeitskorrektur für $a \leq b$ ($a, b \in \mathbb{N}_0$):

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b + 1/2 - \lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{a - 1/2 - \lambda}{\sqrt{\lambda}}\right)$$

Approximation der Verteilungsfunktion von X :

$$P(X \leq x) \approx \Phi\left(\frac{x + 1/2 - \lambda}{\sqrt{\lambda}}\right)$$

Nach einer gängigen Regel ist die Approximation ausreichend gut, wenn $\lambda > 9$.

Bsp 6.12 Angenommen, man möchte für $X \sim P(1000)$ den Wert von $P(X \leq 950)$ mit Hilfe der Normalapproximation $X \approx N(\lambda, \lambda) = N(1000, 1000)$ berechnen. Die exakte Wahrscheinlichkeit ist gegeben durch:

$$P(X \leq 950) = \sum_{x=0}^{950} \frac{1000^x e^{-1000}}{x!}$$

Die rechnerischen Schwierigkeiten mit diesem Ausdruck sind offensichtlich. (Bem: Die R-Funktion `ppois()` liefert einen Wert von 0.0578.) Die Wahrscheinlichkeit lässt sich aber wie folgt approximieren:

$$P(X \leq 950) \approx \Phi\left(\frac{950.5 - 1000}{\sqrt{1000}}\right) \doteq 0.0588$$

Ohne Stetigkeitskorrektur ergibt sich:

$$P(X \leq 950) \approx \Phi\left(\frac{950 - 1000}{\sqrt{1000}}\right) \doteq 0.0569$$

In diesem Fall sogar ein leicht besserer Wert (absolut und relativ). ■

Aufgaben

- 6.1 Ein Hersteller von Cornflakes legt den Packungen Figuren aus einem aktuellen Film bei. Insgesamt gibt es m verschiedene Figuren und jede Packung enthält mit gleicher Wahrscheinlichkeit eine dieser Figuren. Natürlich möchte man die komplette Serie haben. Die erste Packung setzt den Beginn; die zweite Packung enthält eine neue

Figur oder dieselbe wie in der ersten Packung, usf. Berechnen Sie den Erwartungswert der Anzahl X_n von verschiedenen Figuren, die Sie mit n Packungen bekommen. Berechnen Sie $\mathbb{E}(X_n)$ konkret für $m = 20$ und $n = 10, 20, 100$.

Hinweis: Nummerieren Sie die verschiedenen Figuren mit $1, 2, \dots, m$ und stellen Sie X_n als Summe dar:

$$X_n = \sum_{i=1}^m Y_i \quad \text{mit} \quad Y_i = \begin{cases} 1 & \text{wenn (mindestens) eine } i\text{-Figur dabei ist} \\ 0 & \text{sonst} \end{cases}$$

6.2 X_1, X_2, \dots, X_n seien identisch verteilte Größen mit Mittelwert μ und Varianz σ^2 . Ermitteln Sie für ihre Summe $S_n = \sum_{i=1}^n X_i$ den Mittelwert $\mathbb{E}(S_n)$ und die Varianz $\text{Var}(S_n)$, wenn:

- (a) die Größen stochastisch unabhängig sind.
- (b) je zwei Größen eine Korrelation von $0 \leq \rho \leq 1$ aufweisen. (Was ergibt sich speziell für $\rho = 0$ und $\rho = 1$?)

6.3 Ein regelmäßiger Tetraeder mit den Seiten 1, 2, 3, 4 wird mehrfach geworfen. Wenn X_i die beim i -ten Wurf unten liegende Seite ist, bestimmen Sie die Verteilung von (a) $X_1 + X_2$ und von (b) $X_1 + X_2 + X_3$. Bestimmen Sie jeweils auch den Mittelwert und die Varianz.

6.4 X und Y seien unabhängige stochastische Größen mit den Dichten:

$$f_X(x) = I_{(0,1)}(x) \quad \text{und} \quad f_Y(y) = \frac{1}{2} I_{(0,2)}(y)$$

Bestimmen Sie mittels Faltformel die Dichte von $X + Y$. (**Zusatz:** Simulieren Sie die Faltung mehrere tausend Mal und stellen Sie das Ergebnis in Form eines Histogramms grafisch dar.)

6.5 Ein System bestehe aus einer Arbeits- und einer Reservekomponente. Fällt die Arbeitskomponente aus, wird sie unverzüglich durch die Reservekomponente ersetzt. Wenn die Lebensdauern der Komponenten unabhängig exponentialverteilt mit Mittelwert 4 bzw. 3 sind, bestimmen Sie für die Zeitspanne bis zum Ausfall des Systems (a) die Dichte und (b) den Mittelwert und die Streuung. (**Zusatz:** Simulieren Sie das System mehrere tausend Mal und beantworten Sie die Fragen empirisch.)

6.6 An einem Schalter folgen die Servicezeiten einer Exponentialverteilung mit Mittelwert 10 Minuten. Wie ist Ihre Wartezeit verteilt, wenn beim Eintreffen drei Personen vor dem Schalter warten und eine Person bedient wird? Mittelwert? Streuung? (**Hinweis:** Nützen Sie die Gedächtnislosigkeit der Exponentialverteilung; vgl. 4.2.2.)

6.7 Wenn man keinen $10\text{ M}\Omega$ Widerstand hat, einen solchen aber durch Hintereinanderschalten von (1) zehn $1\text{ M}\Omega$, oder (2) fünf $2\text{ M}\Omega$ Widerständen herstellen kann, welche der beiden Möglichkeiten sollte man wählen, wenn der $10\text{ M}\Omega$ Widerstand

möglichst genau sein sollte und die Widerstände aus einer Normalverteilung stammen, deren Mittelwert gleich dem Nominalwert und deren Streuung 1.5% des Nominalwerts beträgt?

- 6.8 Wenn X eine stochastische Größe mit Mittelwert = Varianz = 20 ist, was lässt sich über $P(0 < X < 40)$ sagen?
- 6.9 Angenommen, die Punktezahl pro Student/in bei einem Abschlusstest ist eine sG mit dem Mittelwert 75 und der Varianz 25.
- (a) Geben Sie eine obere Schranke für die Wahrscheinlichkeit, dass die Punktezahl 85 übersteigt.
 - (b) Was lässt sich über die Wahrscheinlichkeit sagen, dass die Punktezahl zwischen 65 und 85 liegt?
 - (c) Wieviele Student/inn/en müssten bei der Prüfung antreten, sodass mit einer Wahrscheinlichkeit von mindestens 0.9 der Punktedurchschnitt um weniger als 5 vom Mittelwert 75 abweicht?
- 6.10 Betrachten Sie ein Quadrat der Seitenlänge 2 (in Nullpunktslage) und den eingeschriebenen Kreis. Wählt man zufällig einen Punkt (V_1, V_2) im Quadrat, so ist die Wahrscheinlichkeit, dass der Punkt innerhalb des Kreises liegt, gleich $\pi/4$. (Warum?) Simuliert man eine Folge von Punkten und definiert:

$$X_i = \begin{cases} 1 & \text{wenn der } i\text{-te Punkt innerhalb des Kreises liegt} \\ 0 & \text{sonst} \end{cases}$$

so folgt, dass $\{X_i\}$ eine iid-Folge mit $\mathbb{E}(X_i) = \pi/4$ ist. Nach dem schGGZ gilt:

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{P} \frac{\pi}{4}$$

D. h., durch Simulation einer großen Zahl von Punkten (V_1, V_2) lässt sich der Wert von π approximieren.⁹ Erzeugen Sie auf diese Weise einige tausend Punkte und ermitteln Sie einen Näherungswert für π . (Streuung des Näherungswerts?)

- 6.11 Ein symmetrischer Würfel wird 1000 Mal geworfen. Berechnen Sie approximativ die Wahrscheinlichkeit, dass die Augenzahl 6 zwischen 150 und 200 Mal inklusive geworfen wird. Wenn die Augenzahl 6 exakt 200 Mal geworfen wird, berechnen Sie approximativ die Wahrscheinlichkeit, dass die Augenzahl 5 weniger als 150 Mal geworfen wird. (Rechnen Sie mit Stetigkeitskorrektur.)
- 6.12 Beim (französischen) Roulette gibt es 37 Felder, nummeriert mit 0, 1, 2, ..., 36. Wenn Sie 1€ auf eine bestimmte Zahl setzen, so gewinnen Sie entweder 35€, wenn diese Zahl kommt, oder Sie verlieren den Einsatz, wenn die Zahl nicht kommt. Wenn

⁹Diese Idee zur Bestimmung von π geht zurück auf den franz. Naturforscher GEORGES-LOUIS LECLERC DE BUFFON (1707–1788), bekannt v. a. durch seine Nadelexperimente (*Buffon'sche Nadel*).

Sie kontinuierlich auf diese Weise spielen, mit welcher approximativen Wahrscheinlichkeit sind Sie (a) nach 35 Spielen, (b) nach 1000 Spielen, (c) nach 100000 Spielen im Plus? (Rechnen Sie mit Stetigkeitskorrektur.)

- 6.13 Die Zahl X der Zugriffe auf eine Webseite folge einer Poissonverteilung mit einem Mittelwert von 10000 pro Tag. Bestimmen Sie approximativ:
- (a) Die Wahrscheinlichkeit von mehr als 20000 Zugriffen pro Tag.
 - (b) Die Wahrscheinlichkeit von weniger als 9900 Zugriffen pro Tag.
 - (c) Einen Wert c so, dass $P(X > c) \approx 0.01$.
 - (d) Die zu erwartende Anzahl von Tagen in einem Jahr (365 Tage), an denen es mehr als 10200 Zugriffe gibt.
 - (e) Die Wahrscheinlichkeit, dass es in einem Jahr (365 Tage) mehr als 15 Tage mit jeweils mehr als 10200 Zugriffen gibt.
- 6.14 Angenommen, eine bestimmte Komponente ist kritisch für die Funktionsfähigkeit eines Systems, und muss nach Ausfall sofort ausgetauscht werden. Wenn die mittlere Lebensdauer dieser Komponente 100 [h] und die Standardabweichung 30 [h] beträgt, wieviele derartige Komponenten müssen vorrätig sein, sodass die Funktion des Systems für die nächsten 2000 Stunden mit einer Mindestwahrscheinlichkeit von 0.95 gewährleistet ist?
- 6.15 A hat 20 Jobs zu erledigen, wobei die für die Erledigung der Jobs benötigten Zeitspannen unabhängige sGn mit Mittelwert 50 [min] und Standardabweichung 10 [min] sind. B hat ebenfalls 20 Jobs zu erledigen, wobei die für die Erledigung der Jobs benötigten Zeitspannen unabhängige sGn mit Mittelwert 52 [min] und Standardabweichung 15 [min] sind. Mit welcher (approximativen) Wahrscheinlichkeit ist A vor B fertig?

7 Schließende Statistik

Allgemein formuliert besteht die Grundaufgabe der **schließenden Statistik**¹ darin, basierend auf **Stichproben** (d. h. Daten oder Beobachtungen) Rückschlüsse auf das zu Grunde liegende („datengenerierende“) **statistische Modell** zu ziehen. Häufig sind statistische Modelle durch **Parameter** charakterisiert und die Aufgabe besteht konkreter darin, diese Parameter zu **schätzen**, Aussagen über die **Genauigkeit** der Schätzungen zu treffen und Hypothesen über die Parameter zu **testen**. Naturgemäß ist das nur unter Inkaufnahme von mehr oder weniger großen Unsicherheiten möglich.

7.1 Grundbegriffe

Man unterscheidet zwischen **parametrischen** und **nichtparametrischen** statistischen Modellen. Erstere sind dadurch charakterisiert, dass sie durch einen ein- oder mehrdimensionalen **Parameter** $\theta \in \Theta \subseteq \mathbb{R}^k$ beschrieben werden können. Die Menge Θ aller möglichen Parameter nennt man den **Parameterraum**.

Bsp 7.1 Ein Beispiel für ein diskretes parametrisches Modell ist etwa die Klasse aller $B(n, p)$ -Verteilungen (mit festem $n \in \mathbb{N}$):

$$\mathcal{P} = \{B(n, p) \mid p \in (0, 1)\}$$

Ein Beispiel für ein stetiges parametrisches Modell ist etwa die Klasse aller $N(\mu, \sigma^2)$ -Verteilungen:

$$\mathcal{P} = \{N(\mu, \sigma^2) \mid \mu \in \mathbb{R}, 0 < \sigma^2 < \infty\}$$

Im ersten Fall handelt es sich um einen eindimensionalen, im zweiten Fall um einen zweidimensionalen Parameter. Hingegen lässt sich ein statistisches Modell der folgenden Art:

$$\mathcal{P} = \{F \mid F \text{ eine stetige Verteilungsfunktion}\}$$

nicht durch einen *endlichdimensionalen* Parameter charakterisieren. In diesem Fall handelt es sich um ein nichtparametrisches Modell. ■

Stichprobe: Man nennt die stochastischen Größen X_1, X_2, \dots, X_n eine **Stichprobe** (oder auch **Zufallsstichprobe**²) einer stochastischen Größe X , wenn die Größen X_i unabhängig und so wie X verteilt sind (d. h., wenn es sich um iid-Größen handelt). Häufig schreibt man $\mathbf{X} = (X_1, X_2, \dots, X_n)'$.

¹Auch *inferentielle Statistik* genannt.

²engl. *random sample*

Bem: Man unterscheide genau zwischen den stochastischen Größen X_1, X_2, \dots, X_n und ihren **Realisationen** x_1, x_2, \dots, x_n (d. h. der *konkreten* Stichprobe).

Ist $p(x)$ (bzw. $f(x)$) die W-Funktion (bzw. Dichte) von X , ist die gemeinsame Verteilung der Stichprobe gegeben durch:

$$\text{diskret: } p(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i)$$

$$\text{stetig: } f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i)$$

Statistik: Eine Funktion $T = T(X_1, X_2, \dots, X_n)$ einer Stichprobe X_1, X_2, \dots, X_n nennt man allgemein eine **Statistik**. Handelt es sich bei T um eine Abbildung in den Parameterraum Θ , d. h., gilt $T : \mathbb{R}^n \rightarrow \Theta$, nennt man die Statistik eine **Schätzfunktion** (kurz einen **Schätzer**) für den Parameter $\theta \in \Theta$. In diesem Fall schreibt man:

$$\hat{\theta}_n = T(X_1, X_2, \dots, X_n)$$

Ebenso verfährt man bei anderen unbekannten Größen. So bezeichnet beispielsweise $\hat{F}_n(x)$ einen Schätzer (auf Basis von n Beobachtungen) für die Verteilungsfunktion $F(x)$.

Bsp 7.2 Statistiken sind uns schon an mehreren Stellen begegnet. So sind beispielsweise die in Kapitel 1 diskutierten grafischen Darstellungen (Histogramm, Boxplot, ...) von Daten x_1, x_2, \dots, x_n Statistiken im obigen Sinn. Ebenso handelt es sich bei den diversen Kennzahlen (Mittelwert, Median, ...) um Beispiele für Schätzfunktionen. ■

Die bei weitem wichtigsten Schätzfunktionen in der Statistik sind der **Stichprobenmittelwert** \bar{X}_n und die **Stichprobenvarianz** S_n^2 :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Allgemein gilt: Der Stichprobenmittelwert ist ein Schätzer für den Mittelwert μ und die Stichprobenvarianz ist ein Schätzer für die Varianz σ^2 einer Verteilung. Ein Schätzer für die Streuung σ ist die **Stichprobenstreuung** S_n :

$$S_n = \sqrt{S_n^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

Die Eigenschaften dieser (und anderer) Schätzer werden im Folgenden noch ausführlicher diskutiert.

7.2 Schätzer

7.2.1 Empirische Verteilungsfunktion

Die Verteilung einer sG X ist durch ihre Verteilungsfunktion spezifiziert:

$$F(x) = P(X \leq x) \quad \text{für } x \in \mathbb{R}$$

Hat man X mehrfach beobachtet, d. h., hat man eine Stichprobe X_1, X_2, \dots, X_n von X , so stellt sich die Frage, wie F geschätzt werden kann. Dazu nehmen wir die bereits in 1.7.2 diskutierte **empirische Verteilungsfunktion**:³

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X_i) \quad \text{für } x \in \mathbb{R}$$

Für ein festes $x \in \mathbb{R}$ ist $\hat{F}_n(x)$ der Anteil der Beobachtungen X_1, X_2, \dots, X_n , die kleiner oder gleich x sind.

Eigenschaften der empirischen VF: Für festes $x \in \mathbb{R}$ gilt:

$$(1) \quad \mathbb{E}[\hat{F}_n(x)] = F(x)$$

$$(2) \quad \text{Var}[\hat{F}_n(x)] = \frac{F(x)[1 - F(x)]}{n}$$

$$(3) \quad \hat{F}_n(x) \xrightarrow{P} F(x) \quad \text{für } n \rightarrow \infty$$

Beweis: Aus der Definition von $\hat{F}_n(x)$ folgt, dass $Y_n = n\hat{F}_n(x)$ (= Zahl der Beobachtungen kleiner oder gleich x) binomialverteilt $B(n, p)$ ist, wobei $p = F(x)$ (= Wahrscheinlichkeit, dass eine Beobachtung kleiner oder gleich x ist). Damit gilt:

$$\mathbb{E}(Y_n) = np = nF(x) \implies \mathbb{E}[\hat{F}_n(x)] = \mathbb{E}\left(\frac{Y_n}{n}\right) = F(x)$$

$$\text{Var}(Y_n) = np(1-p) = nF(x)[1 - F(x)] \implies \text{Var}[\hat{F}_n(x)] = \text{Var}\left(\frac{Y_n}{n}\right) = \frac{F(x)[1 - F(x)]}{n}$$

Das zeigt (1) und (2); (3) folgt aus dem schGGZ (UE-Aufgabe).

Eigenschaft (3) besagt für *festes* $x \in \mathbb{R}$, dass $\hat{F}_n(x)$ in Wahrscheinlichkeit gegen $F(x)$ konvergiert. Es gilt aber noch mehr:

³engl. *empirical (cumulative) distribution function* (abgekürzt e(c)df)

Satz von Gliwenko–Cantelli:⁴ Für eine Stichprobe X_1, X_2, \dots, X_n von $X \sim F(x)$ gilt:

$$P\left(\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = 0\right) = 1$$

D. h., mit Wahrscheinlichkeit 1 konvergiert $\hat{F}_n(x)$ *gleichmäßig* gegen die zugrunde liegende Verteilungsfunktion $F(x)$.

Bem: Wegen seiner großen Bedeutung heißt dieser Satz auch **Fundamentalsatz** – oder **Hauptsatz – der Statistik**.

Bsp 7.3 Zur Illustration des Satzes von Gliwenko–Cantelli simulieren wir Beobachtungen einer $\text{Exp}(\tau = 2)$ -Verteilung, zeichnen die empirische Verteilungsfunktion $\hat{F}_n(x)$ und bestimmen Stelle und Wert des größten Abstands D_n zur (theoretischen) Verteilungsfunktion $F(x) = 1 - e^{-x/2}$:

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - (1 - e^{-x/2})|$$

Abb 7.1 zeigt das Ergebnis für eine kleine Stichprobe ($n = 10$) und Abb 7.2 das Ergebnis für eine große Stichprobe ($n = 100$). Deutlich zeigt sich der über ganz \mathbb{R}^+ gleichmäßig kleinere Abstand von \hat{F}_n und F für die größere Stichprobe. ■

Bem: Insbesondere ist die Schätzung der VF mittels empirischer VF für *stetige* Verteilungen von Bedeutung, also für das – nichtparametrische – Verteilungsmodell:

$$\mathcal{P} = \{F \mid F \text{ eine stetige Verteilungsfunktion}\}$$

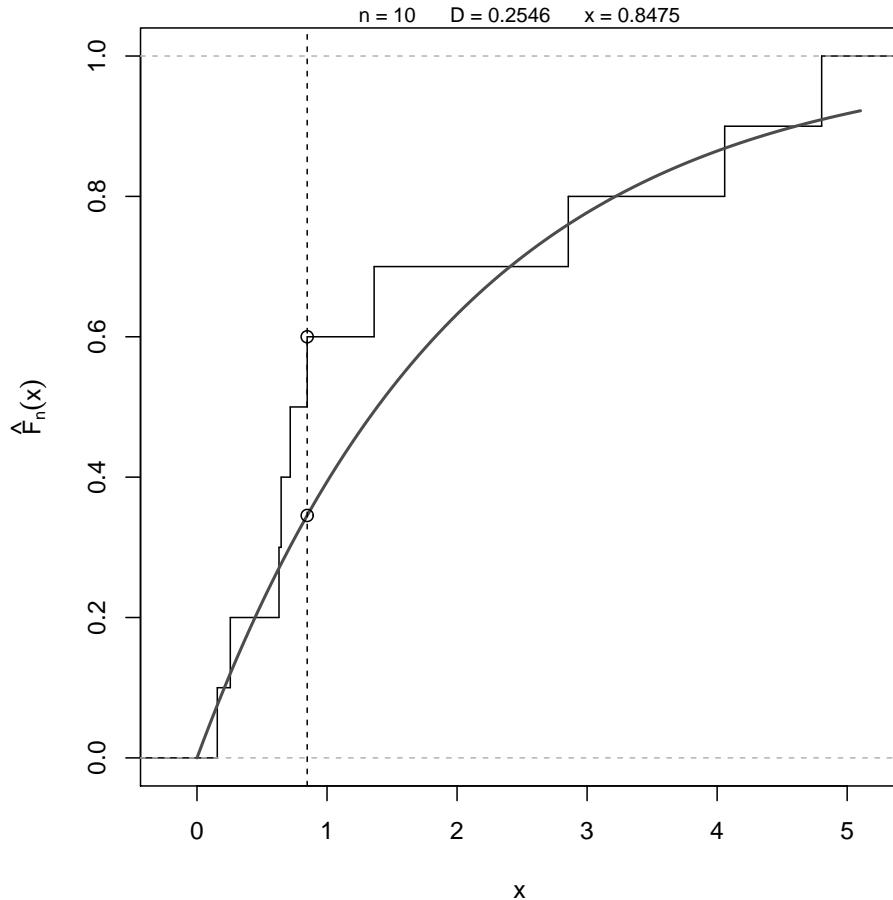
Häufig interessiert man sich aber auch für die Schätzung der Dichte, legt also das folgende – nichtparametrische – Verteilungsmodell zugrunde:

$$\mathcal{P} = \{f \mid f \text{ eine Dichtefunktion}\}$$

In Kapitel 1 haben wir zwei diesbezügliche **Dichteschätzer** kennengelernt, das Histogramm (vgl. 1.7.5) und die Kernschätzung (vgl. 1.7.6), und dabei auch einige kritische Punkte angesprochen (Wahl der Bins, der Bandbreite, ...). Generell lässt sich sagen, dass die Schätzung der Dichte ein statistisch schwierigeres Problem darstellt als die Schätzung der Verteilungsfunktion. (Eine weitere Diskussion geht aber über den Rahmen dieser VO hinaus.)

⁴WALERI IWANOWITSCH GLIWENKO (1897–1940), russ. Mathematiker; FRANCESCO PAOLO CANTELLI (1875–1966), ital. Mathematiker.

Abbildung 7.1: Illustration zum Fundamentalsatz (kleine Stichprobe)



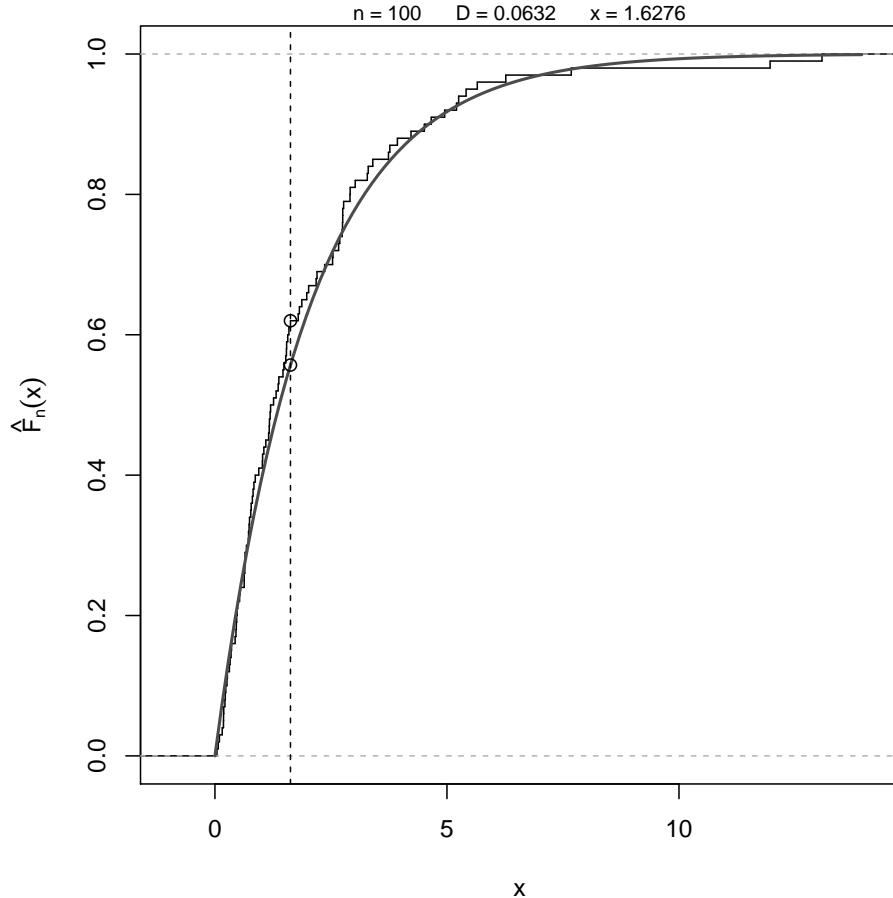
7.2.2 Momentenschätzer

Die Idee hinter der **Momentenmethode** zur Schätzung von Parametern besteht darin, die theoretischen Momente der Verteilung den entsprechenden Stichprobenmomenten gleichzusetzen. Da Erstere Funktionen der unbekannten Parameter sind, lassen sich durch Auflösen dieser Gleichungen Schätzer für die Parameter gewinnen.

Das k -te **Moment** einer stochastischen Größe X ist definiert durch $\mathbb{E}(X^k)$. (Speziell ist etwa der Mittelwert $\mathbb{E}(X)$ das erste Moment.) Ist X_1, X_2, \dots, X_n eine Stichprobe von X , so ist das k -te **Stichprobenmoment** definiert durch $(1/n) \sum_{i=1}^n X_i^k$. (Speziell ist etwa der Stichprobenmittelwert \bar{X}_n das erste Stichprobenmoment.)

Gibt es im Verteilungsmodell m unbekannte Parameter, $\theta_1, \theta_2, \dots, \theta_m$, so lassen sich durch Auflösen des folgenden Gleichungssystems:

$$\mathbb{E}(X^k) = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad k = 1, 2, \dots, m$$

Abbildung 7.2: Illustration zum Fundamentalsatz (große Stichprobe)

Schätzer für die Parameter gewinnen:

$$\hat{\theta}_k = T_k(X_1, X_2, \dots, X_n), \quad k = 1, 2, \dots, m$$

Bem: Meist handelt es sich um ein nichtlineares Gleichungssystem, sodass – speziell bei mehreren Parametern – die explizite Auslösung nach θ_k schwierig sein kann. In diesem Fall müssen numerische Methoden (z. B. Iterationsverfahren) angewendet werden.

Bsp 7.4 Sei X_1, X_2, \dots, X_n eine Stichprobe von $X \sim N(\mu, \sigma^2)$. Wie lauten die Momentenschätzer von μ und σ^2 ? Die ersten beiden Momente von X sind gegeben durch:

$$\mathbb{E}(X) = \mu, \quad \mathbb{E}(X^2) = \mu^2 + \sigma^2$$

(Letzteres folgt aus dem Verschiebungssatz.) Das Gleichungssystem lautet also wie folgt:

$$\mu = \bar{X}_n, \quad \mu^2 + \sigma^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$$

In diesem Fall ist die Auflösung einfach:

$$\hat{\mu} = \bar{X}_n, \quad \hat{\sigma}^2 = \frac{1}{n} \left(\sum_{i=1}^n X_i^2 - n \bar{X}_n^2 \right) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Die Schätzer haben eine plausible Form; man beachte allerdings, dass der Schätzer für σ^2 nicht identisch mit dem üblichen Varianzschätzer S_n^2 ist. Für nicht zu kleine Stichproben sind die Unterschiede aber gering. ■

7.2.3 Maximum Likelihood

Neben der Momentenschätzung ist auch die **Maximum-Likelihood-Schätzung** eine konstruktive Methode zur Gewinnung von Schätzfunktionen. (Bem: Entwickelt von R. A. FISHER in den 1920er Jahren.) Wie unten noch ausführlicher diskutiert, bekommt man mit dieser Methode – unter bestimmten Bedingungen – „optimale“ Schätzer (zumindest für große Stichproben).

Maximum-Likelihood-Schätzer (diskreter Fall): Ist X eine diskrete sG mit der W-Funktion $p(x; \theta)$, wobei $\theta \in \Theta$ ein einzelner unbekannter Parameter ist, und sind x_1, x_2, \dots, x_n (konkrete) Beobachtungen einer Stichprobe X_1, X_2, \dots, X_n von X , so ist die **Likelihood-Funktion** (kurz **Likelihood**⁵) der Stichprobe definiert durch:

$$L(\theta) = \prod_{i=1}^n p(x_i; \theta) \quad \text{für } \theta \in \Theta$$

Der **Maximum-Likelihood-Schätzer** (kurz **ML-Schätzer**) von θ ist nun jener Wert aus Θ , der $L(\theta)$ maximiert.

Bemerkungen:

- (a) Im diskreten Fall lässt sich die Likelihood der Stichprobe wie folgt interpretieren: Für $\theta \in \Theta$ entspricht $L(\theta)$ gerade der Wahrscheinlichkeit, die Stichprobenwerte x_1, x_2, \dots, x_n zu beobachten:

$$L(\theta) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)$$

⁵likelihood = Wahrscheinlichkeit, Plausibilität

Dabei wird für die Berechnung der Wahrscheinlichkeit(en) der Parameterwert θ zugrunde gelegt.

- (b) Der ML–Schätzer ist jener θ –Wert, der die Beobachtung der (konkreten) Stichprobe x_1, x_2, \dots, x_n am wahrscheinlichsten (oder plausibelsten) macht.
- (c) Das der ML–Schätzung zugrunde liegende **Likelihood–Prinzip** lässt sich wie folgt formulieren: Entscheide dich für das plausibelste Verteilungsmodell. Oder: Entscheide dich für jenes Modell, das die Daten mit höchster Wahrscheinlichkeit (oder Plausibilität) erzeugt (hat).

Bsp 7.5 Für eine Bernoulli–Größe X lautet die W–Funktion wie folgt:

$$p(x; \theta) = \begin{cases} \theta^x (1 - \theta)^{1-x} & x = 0, 1 \\ 0 & \text{sonst} \end{cases}$$

Dabei ist $0 \leq \theta \leq 1$ der zu schätzende Parameter (= Erfolgswahrscheinlichkeit). (Bem: Es ist nicht unüblich, in der schließenden Statistik den in Frage stehenden Parameter allgemein mit θ zu bezeichnen.) Die Likelihood–Funktion für eine (konkrete) Stichprobe x_1, x_2, \dots, x_n der Größe n ist gegeben durch:

$$L(\theta) = \prod_{i=1}^n p(x_i; \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$$

Letzterer Ausdruck ist nach θ zu maximieren. Das ist zwar nicht schwierig, einfacher ist es jedoch, anstelle von $L(\theta)$ die logarithmierte Likelihood–Funktion $\ln L(\theta)$ zu maximieren:⁶

$$\ln L(\theta) = \left(\sum_{i=1}^n x_i \right) \ln \theta + \left(n - \sum_{i=1}^n x_i \right) \ln(1 - \theta)$$

Letzteres nennt man die **Log-Likelihood (–Funktion)**. Die Stelle des Maximums bestimmt man auf die übliche Weise:

$$\frac{d \ln L(\theta)}{d\theta} = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1 - \theta}$$

Setzt man die Ableitung gleich Null und löst nach θ auf, ergibt sich:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$$

⁶Der Logarithmus als strikt monoton wachsende Funktion verändert die Stelle des Maximums nicht.

Man überzeugt sich leicht davon (2. Ableitung), dass es sich um die Stelle eines Maximums handelt. Der ML–Schätzer von θ ist also gegeben durch:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$$

Der Erwartungswert von X ist θ ; der Momentenschätzer ist in diesem Fall also identisch mit dem ML–Schätzer. ■

Bem: Man unterscheide allgemein zwischen dem **Schätzer** (oder der **Schätzfunktion**) und dem **Schätzwert**. Beispielsweise ist der ML–Schätzer von θ im obigen Beispiel gegeben durch \bar{X}_n (eine sG), der ML–Schätzwert von θ aber ist \bar{x}_n (eine konkrete Zahl).

Auch wenn sich die in der obigen **Bemerkung (a)** gegebene Interpretation der Likelihood einer Stichprobe genaugenommen auf diskrete sGn beschränkt, lässt sich die ML–Methode sinngemäß auf den stetigen Fall übertragen.

Maximum-Likelihood–Schätzer (stetiger Fall): Ist X eine stetige sG mit der Dichte $f(x; \theta)$, wobei $\theta \in \Theta$ ein einzelner unbekannter Parameter ist, und sind x_1, x_2, \dots, x_n (konkrete) Beobachtungen einer Stichprobe X_1, X_2, \dots, X_n von X , so ist die **Likelihood–Funktion** (kurz **Likelihood**) der Stichprobe definiert durch:

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta) \quad \text{für } \theta \in \Theta$$

Der **ML–Schätzer** von θ ist nun jener Wert aus Θ , der $L(\theta)$ maximiert.

Bsp 7.6 Die sG X sei exponentialverteilt $X \sim \text{Exp}(\lambda)$. Die Likelihood einer (konkreten) Stichprobe x_1, x_2, \dots, x_n von X ist gegeben durch:

$$L(\lambda) = \prod_{i=1}^n f(x_i; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

Die Likelihood ist bezüglich λ zu maximieren. Wieder ist es in diesem Fall einfacher, dafür die **Log-Likelihood** heranzuziehen:

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^n x_i$$

Ableiten und Nullsetzen:

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \implies \hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}_n}$$

Der ML–Schätzer von λ ist also gegeben durch:

$$\hat{\lambda} = \frac{1}{\bar{X}_n}$$

Bem: Der Erwartungswert von X ist $1/\lambda$; der Momentenschätzer ist in diesem Fall also identisch mit dem ML–Schätzer.

Als konkretes Beispiel seien etwa die Ausfallzeiten (Einheit [h]) von $n = 8$ gleichartigen Komponenten wie folgt:

Ausfallzeiten [h]							
11.96	5.03	67.40	16.07	31.50	7.73	11.10	22.38

Handelt es sich – in guter Näherung – um Beobachtungen einer $\text{Exp}(\lambda)$ –Verteilung, so ist der ML–Schätzwert von λ gegeben durch:

$$\hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{21.65} = 0.0462$$

Die durchgezogene Linie in Abb 7.3 zeigt die Log-Likelihood⁷ für ein Intervall um den ML–Schätzwert $\hat{\lambda}$. Diese Kurve ist als Folge der nur kleinen Stichprobe vergleichsweise flach um das Maximum, d. h., die Präzision der Schätzung ist nicht sehr hoch. Hätten wir – für einen unveränderten Wert von \bar{x} – die Schätzung auf $n = 20$ (strichliert) oder sogar $n = 40$ (punktliert) Beobachtungen stützen können, wären die Kurven um das Maximum stärker gewölbt und die Schätzungen daher präziser. ■

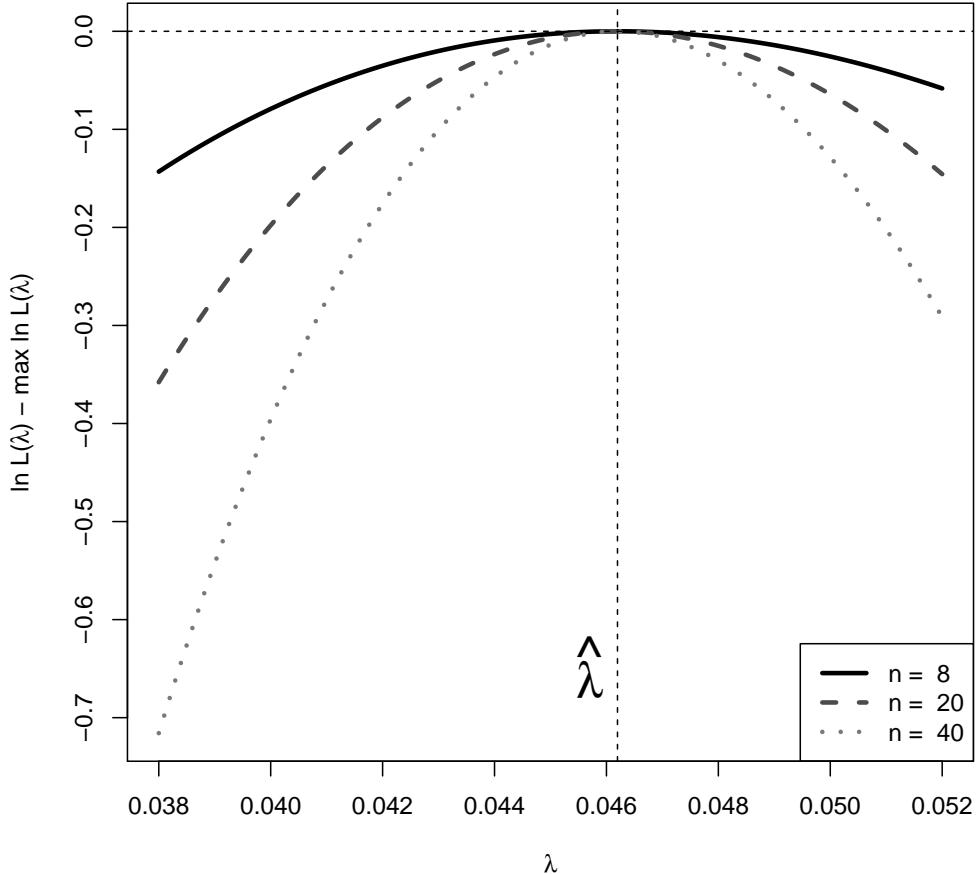
Die ML–Schätzmethode hat eine sehr nützliche Eigenschaft unter Transformationen.

Invarianz der ML–Schätzung: X_1, X_2, \dots, X_n sei eine Stichprobe von $X \sim f(x; \theta)$ (oder $X \sim p(x; \theta)$) und $\eta = g(\theta)$ sei eine Funktion des Parameters. Ist $\hat{\theta}$ der ML–Schätzer von θ , so ist der ML–Schätzer von η gegeben durch:

$$\hat{\eta} = \widehat{g(\theta)} = g(\hat{\theta})$$

⁷Genauer die Funktion $\ln L(\lambda) - \max_{\lambda} \ln L(\lambda)$.

Abbildung 7.3: Log-Likelihood (Bsp 7.6)



Beispielsweise ist im Kontext von Bsp 7.6 der ML-Schätzer von $\tau = 1/\lambda$ (= Erwartungswert von X) ohne weitere Rechnung gegeben durch:

$$\hat{\tau} = \frac{1}{\bar{\lambda}} = \bar{X}_n$$

Die ML-Methode lässt sich auch für die Schätzung von mehreren Parametern anwenden, im Folgenden formuliert nur für den stetigen Fall (analog für den diskreten Fall).

Mehrere Parameter: Ist X eine stetige sG mit der Dichte $f(x; \boldsymbol{\theta})$, wobei $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)'$ ein k -dimensionaler Parameter aus $\Theta \subseteq \mathbb{R}^k$ ist, so ist für eine (konkrete) Stichprobe x_1, x_2, \dots, x_n von X die **Likelihood (-Funktion)** gegeben durch:

$$L(\boldsymbol{\theta}) = L(\theta_1, \theta_2, \dots, \theta_k) = \prod_{i=1}^n f(x_i; \boldsymbol{\theta}) \quad \text{für } \boldsymbol{\theta} \in \Theta$$

Der **ML-Schätzer** von $\boldsymbol{\theta}$ ist nun jener Wert aus Θ , der $L(\boldsymbol{\theta})$ maximiert.

Manchmal lässt sich der ML-Schätzer durch Lösen der folgenden Gleichungen bestimmen:

$$\frac{\partial L(\theta_1, \theta_2, \dots, \theta_k)}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, k$$

Oder – meist einfacher – durch Lösen der folgenden Gleichungen auf Basis der **Log-Likelihood**:

$$\frac{\partial \ln L(\theta_1, \theta_2, \dots, \theta_k)}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, k$$

Bem: Die obigen sog. **ML-Gleichungen** haben vielfach keine explizite Lösung, sodass man – ausgehend von Startwerten – iterative Lösungsmethoden anwenden muss. Man kann aber auch versuchen, die Stelle des Maximums von $L(\boldsymbol{\theta})$ direkt numerisch zu bestimmen (beispielsweise mittels der R-Funktion `optim()`).

Bsp 7.7 [Normalverteilung] Für eine (konkrete) Stichprobe x_1, x_2, \dots, x_n von $X \sim N(\mu, \sigma^2)$ ist die Likelihood gegeben durch:

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{(x_i - \mu)^2}{2\sigma^2} \right] = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right]$$

Die Log-Likelihood lautet wie folgt:

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

Partiell ableiten und Nullsetzen:

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

Dieses Gleichungssystem lässt sich einfach nach μ und σ^2 auflösen:

$$\hat{\mu} = \bar{X}_n \quad \text{und} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Die oben erwähnte **Invarianzeigenschaft** der ML–Methode gilt auch für mehrdimensionale Parameter, sodass der ML–Schätzer für σ gegeben ist durch:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

Man beachte, dass der ML–Schätzer für σ^2 nicht mit dem üblichen Varianzschätzer S_n^2 übereinstimmt. (Für große Stichproben ist der Unterschied aber gering.) ■

Bsp 7.8 Wir betrachten noch einmal die $n = 8$ konkreten Ausfallzeiten von **Bsp 7.6**, passen diesmal aber die allgemeinere $\text{Gam}(\alpha, \lambda)$ –Verteilung an. Die ML–Gleichungen haben in diesem Fall eine komplizierte Form und es gibt keine explizite Lösung. Der folgende R–Output zeigt die Lösung mittels `fitdistr()` (Package: MASS).

```
require(MASS)
x <- c(11.96, 5.03, 67.40, 16.07, 31.50, 7.73, 11.10, 22.38)

# avoid spurious accuracy
op <- options(digits = 3)

fitdistr(x, "gamma")
  shape      rate
  1.7457    0.0806
(0.8032) (0.0429)

# now do this with more control
fitdistr(x, dgamma, start=list(shape=1, rate=0.1), lower=0.001)
  shape      rate
  1.7459    0.0807
(0.8032) (0.0429)
```

Die ML–Schätzwerte für α (= `shape`) und λ (= `rate`) sind also gegeben durch:

$$\hat{\alpha} = 1.746 \quad \text{und} \quad \hat{\lambda} = 0.081$$

Die eingeklammerten Werte sind Schätzwerte für die Streuungen der ML–Schätzer. ■

7.2.4 Gütekriterien für Schätzer

Für die Schätzung von Parametern auf Basis einer Stichprobe hat man vielfach mehrere Schätzer zur Auswahl und es stellt sich die Frage, welche(n) man bevorzugen sollte. Man

kann sich auch fragen, welche Schätzer in einer bestimmten Situation „optimal“ sind. Zur Beantwortung dieser Fragen benötigt man entsprechende **Gütekriterien** für Schätzer.

Erwartungstreue: Ein Schätzer $\hat{\theta}_n = T(X_1, X_2, \dots, X_n)$ für einen Parameter $\theta \in \Theta$ heißt **erwartungstreu** (oder **unverzerrt**⁸), wenn:

$$\mathbb{E}_{\theta}(\hat{\theta}_n) = \theta \quad \text{für alle } \theta \in \Theta$$

Gilt für $n \rightarrow \infty$, dass:

$$\mathbb{E}_{\theta}(\hat{\theta}_n) \rightarrow \theta \quad \text{für alle } \theta \in \Theta$$

nennt man $\hat{\theta}_n$ **asymptotisch erwartungstreu** (oder **unverzerrt**).

Bemerkungen:

- (a) Anschaulich bedeutet die obige Definition, dass man bei Verwendung eines erwartungstreuen Schätzers keinen *systematischen* Fehler macht, sondern *im Mittel* (oder *im Durchschnitt*) an der gewünschten Stelle ist.
- (b) Die Schreibweise $\mathbb{E}_{\theta}(\hat{\theta}_n)$ soll darauf hinweisen, dass der Erwartungswert von $\hat{\theta}_n$ mit dem Parameterwert θ zu berechnen ist.
- (c) Ein wesentlicher Punkt bei der obigen Definition besteht darin, dass die Bedingung für *alle* $\theta \in \Theta$ erfüllt sein muss, und nicht etwa nur für den „wahren“ Wert des Parameters.
- (d) Für einen verzerrten Schätzer $\hat{\theta}_n$ definiert man die **Verzerrung** (engl. *Bias*) durch:

$$\text{Bias}(\hat{\theta}_n; \theta) = \mathbb{E}_{\theta}(\hat{\theta}_n) - \theta, \quad \theta \in \Theta$$

Meist ist die Verzerrung eine Funktion von θ .

Bsp 7.9 [Stichprobenmittelwert/Stichprobenvarianz] In diesem Beispiel zeigen wir, dass für eine Stichprobe X_1, X_2, \dots, X_n von X (deren Mittelwert und Varianz existieren) der **Stichprobenmittelwert** \bar{X}_n und die **Stichprobenvarianz** S_n^2 erwartungstreue Schätzer für $\mu = \mathbb{E}(X)$ bzw. $\sigma^2 = \text{Var}(X)$ sind. Der Nachweis der Erwartungstreue von \bar{X}_n ist einfach:

$$\mathbb{E}(\bar{X}_n) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}(X_i)}_{=\mu} = \frac{n\mu}{n} = \mu$$

⁸engl. *unbiased*

Für den Nachweis der Erwartungstreue von S_n^2 müssen wir etwas weiter ausholen. Zunächst gilt nach dem empirischen Verschiebungssatz:

$$(n-1)S_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n X_i^2 - n\bar{X}_n^2$$

Nach dem Verschiebungssatz für sGn wiederum gilt:

$$\mathbb{E}(X_i^2) = \text{Var}(X_i) + \mathbb{E}^2(X_i) = \sigma^2 + \mu^2$$

Die Varianz von \bar{X}_n ist gegeben durch:

$$\begin{aligned} \text{Var}(\bar{X}_n) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \underbrace{\text{Var}(X_i)}_{=\sigma^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \\ \implies \mathbb{E}(\bar{X}_n^2) &= \text{Var}(\bar{X}_n) + \mathbb{E}^2(\bar{X}_n) = \frac{\sigma^2}{n} + \mu^2 \end{aligned}$$

Damit folgt:

$$\begin{aligned} \mathbb{E}[(n-1)S_n^2] &= \sum_{i=1}^n \mathbb{E}(X_i^2) - n\mathbb{E}(\bar{X}_n^2) \\ &= n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= (n-1)\sigma^2 \end{aligned}$$

D.h., $\mathbb{E}(S_n^2) = \sigma^2$ (für alle Werte von μ und σ^2). Das erklärt den Faktor $1/(n-1)$ im Ausdruck für S_n^2 . Hätten wir die Stichprobenvarianz wie folgt definiert:

$$S_n'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n-1}{n} S_n^2$$

würden wir σ^2 systematisch *unterschätzen* ($\text{Bias} = -\sigma^2/n$). ■

Effizienz: Neben dem Erwartungswert spielt auch die Varianz eine wesentliche Rolle bei der Beurteilung von Schätzern. Man sagt, dass ein erwartungstreuer Schätzer $\hat{\theta}_1$ des Parameters θ **effizienter** als ein anderer erwartungstreuer Schätzer $\hat{\theta}_2$ desselben Parameters ist, wenn:

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$$

Ist ein erwartungstreuer Schätzer des Parameters θ effizienter als jeder andere erwartungstreue Schätzer desselben Parameters, nennt man ihn **effizient**.

Bsp 7.10 [Linear effiziente Schätzer] Der Nachweis der Effizienz eines Schätzers erfordert in der Regel weitergehende Konzepte der Statistik. Beschränkt man sich allerdings auf **lineare** Schätzer, d. h. auf Schätzer der Form $T_n = \sum_{i=1}^n a_i X_i$, genügen meist einfachere Überlegungen. Im Folgenden zeigen wir, dass der Stichprobenmittelwert \bar{X}_n der **linear effiziente Schätzer** des Mittelwerts $\mu = \mathbb{E}(X)$ ist.

Sei $T_n = \sum_{i=1}^n a_i X_i$ ein beliebiger linearer erwartungstreuer Schätzer für μ auf Basis einer Stichprobe X_1, X_2, \dots, X_n von X ; dann gilt für alle μ :

$$\mu = \mathbb{E}(T_n) = \mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \underbrace{\mathbb{E}(X_i)}_{=\mu} = \mu \sum_{i=1}^n a_i \implies \sum_{i=1}^n a_i = 1$$

Für die Varianz von T_n gilt:

$$\text{Var}(T_n) = \text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \underbrace{\text{Var}(X_i)}_{=\sigma^2} = \sigma^2 \sum_{i=1}^n a_i^2 \longrightarrow \text{Min!}$$

Nun gilt:

$$\begin{aligned} \sum_{i=1}^n a_i^2 &= \sum_{i=1}^n \left(a_i - \frac{1}{n} + \frac{1}{n}\right)^2 \\ &= \sum_{i=1}^n \left[\left(a_i - \frac{1}{n}\right)^2 + 2\left(a_i - \frac{1}{n}\right)\frac{1}{n} + \left(\frac{1}{n}\right)^2\right] \\ &= \sum_{i=1}^n \left(a_i - \frac{1}{n}\right)^2 + \frac{2}{n} \underbrace{\left(\sum_{i=1}^n a_i - 1\right)}_{=0} + \frac{1}{n} \\ &= \sum_{i=1}^n \left(a_i - \frac{1}{n}\right)^2 + \frac{1}{n} \geq 0 \end{aligned}$$

Der letztere Ausdruck ist minimal ($= 1/n$), wenn die erste Summe gleich Null ist, d. h., wenn $a_i = 1/n$ für alle $i = 1, 2, \dots, n$. Der linear effiziente Schätzer für μ lautet also wie folgt:

$$T_n = \sum_{i=1}^n \frac{1}{n} X_i = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$$

■

Konsistenz: Ein Schätzer $\hat{\theta}_n = T(X_1, X_2, \dots, X_n)$, basierend auf einer Stichprobe der Größe n , heißt (schwach) **konsistent** für θ , wenn:

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{für } n \rightarrow \infty$$

Bemerkungen:

- (a) Anschaulich bedeutet Konsistenz, dass sich ein Schätzer mit dieser Eigenschaft für wachsendes n mit hoher Wahrscheinlichkeit in der Nähe des zu schätzenden Parameters aufhält. Letzteres ist eine sehr wünschenswerte Eigenschaft von „guten“ Schätzern.
- (b) Aus den Eigenschaften der stochastischen Konvergenz (vgl. 6.3.2) folgt: Ist $\hat{\theta}_n$ konsistent für θ und ist g eine stetige Funktion, so ist auch $g(\hat{\theta}_n)$ konsistent für $g(\theta)$.
- (c) Ist $\hat{\theta}_n$ ein asymptotisch erwartungstreuer Schätzer (d. h. $\lim_{n \rightarrow \infty} \mathbb{E}(\hat{\theta}_n) = \theta$) und gilt $\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0$, dann ist $\hat{\theta}_n$ auch ein konsistenter Schätzer von θ .

Bsp 7.11 [Stichprobenmittelwert/Stichprobenvarianz] In diesem Beispiel zeigen wir, dass für eine Stichprobe X_1, X_2, \dots, X_n von X (deren Mittelwert und Varianz existieren) der **Stichprobenmittelwert** \bar{X}_n und die **Stichprobenvarianz** S_n^2 konsistente Schätzer für $\mu = \mathbb{E}(X)$ bzw. $\sigma^2 = \text{Var}(X)$ sind. Ersteres wissen wir schon: Die Konsistenz von \bar{X}_n ist äquivalent zum schGGZ (vgl. 6.3.2):

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}(X) = \mu$$

Daraus folgt nach der obigen **Bemerkung (b)**, dass $\bar{X}_n^2 \xrightarrow{P} \mu^2$. Das schGGZ lässt sich aber auch auf die iid-Folge $\{X_n^2\}$ anwenden:

$$\frac{1}{n} \sum_{i=1}^n X_i^2 \xrightarrow{P} \mathbb{E}(X^2) = \sigma^2 + \mu^2$$

Damit folgt:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \underbrace{\frac{n}{n-1}}_{\rightarrow 1} \left[\underbrace{\frac{1}{n} \sum_{i=1}^n X_i^2}_{\rightarrow \sigma^2 + \mu^2} - \underbrace{\bar{X}_n^2}_{\rightarrow \mu^2} \right] \xrightarrow{P} \sigma^2$$

Das zeigt die Konsistenz von S_n^2 . Auf ähnliche Weise zeigt man, dass $S_n = +\sqrt{S_n^2}$ ein asymptotisch erwartungstreuer (d. h. $\lim_{n \rightarrow \infty} \mathbb{E}(S_n) = \sigma$) und konsistenter Schätzer von σ ist (Letzteres folgt wieder aus der obigen **Bemerkung (b)**):

$$S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2} \xrightarrow{P} \sigma$$

■

Asymptotische Normalverteilung: Ein Schätzer $\hat{\theta}_n = T(X_1, X_2, \dots, X_n)$ ist **asymptotisch normalverteilt**, wenn er in Verteilung (vgl. 6.3.3) gegen eine normalverteilte sG konvergiert, d. h., wenn für alle $z \in \mathbb{R}$:

$$\lim_{n \rightarrow \infty} P\left(\frac{\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)}{\sqrt{\text{Var}(\hat{\theta}_n)}} \leq z\right) = \Phi(z)$$

Das lässt sich auch wie folgt ausdrücken:

$$\hat{\theta}_n \stackrel{\text{asympt.}}{\sim} N(\mathbb{E}(\hat{\theta}_n), \text{Var}(\hat{\theta}_n))$$

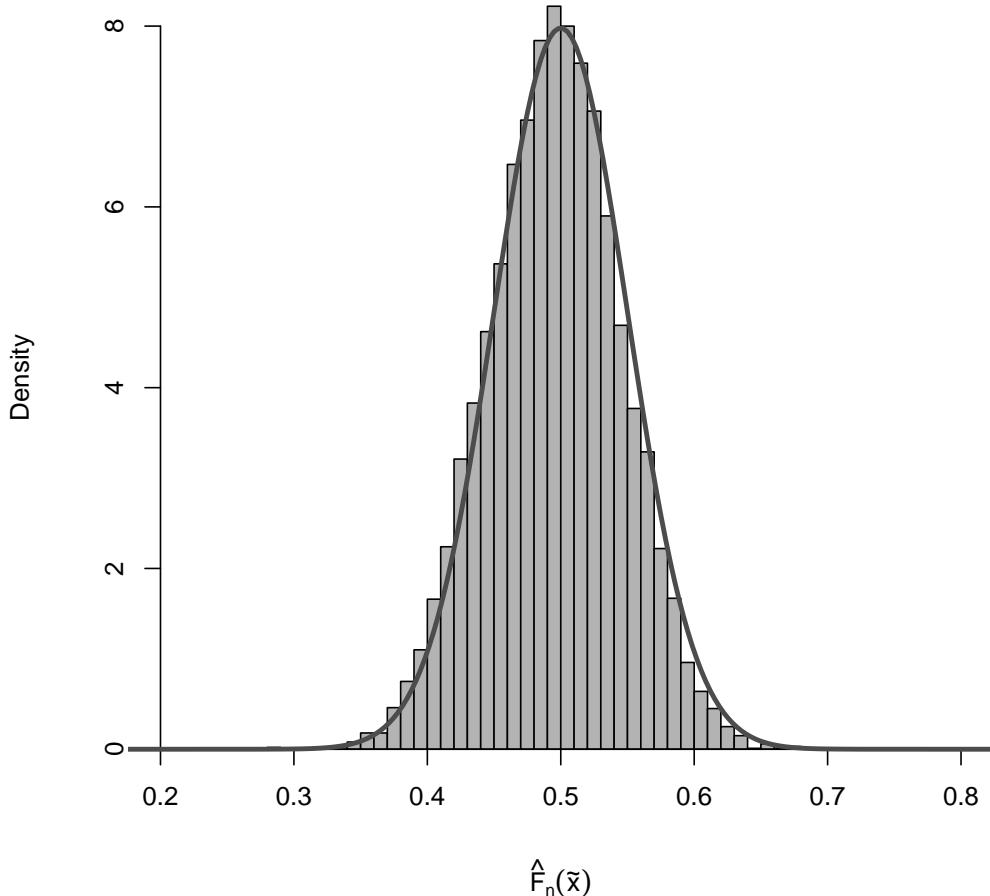
Bsp 7.12 [Empirische Verteilungsfunktion] Die in 7.2.1 diskutierten Eigenschaften der **empirischen Verteilungsfunktion** können auch so formuliert werden, dass $\hat{F}_n(x)$ für festes $x \in \mathbb{R}$ ein erwartungstreuer und konsistenter Schätzer von $F(x)$ ist. Darüberhinaus gilt, dass $\hat{F}_n(x)$ nach dem ZGVS (vgl. 6.3.3) auch asymptotisch normalverteilt ist:

$$\hat{F}_n(x) \stackrel{\text{asympt.}}{\sim} N\left(F(x), \frac{F(x)[1-F(x)]}{n}\right)$$

Zur Illustration dieses Sachverhalts betrachten wir ein konkretes Beispiel: X sei nach $\text{Exp}(1)$ verteilt und $\tilde{x} = -\ln(1-1/2) = \ln 2$ sei der Median von X . Dann gilt $F(\tilde{x}) = 1/2$ und $F(\tilde{x})[1-F(\tilde{x})] = 1/4$. D. h., an der Stelle $x = \tilde{x}$ gilt:

$$\hat{F}_n(\tilde{x}) \stackrel{\text{asympt.}}{\sim} N\left(\frac{1}{2}, \frac{1}{4n}\right)$$

Abb 7.4 zeigt in Form eines Histogramms der $\hat{F}_n(\tilde{x})$ -Werte das Ergebnis von $N = 10000$ Simulationen zu je $n = 100$ Beobachtungen einer $\text{Exp}(1)$ -Verteilung. Die Linie entspricht der Dichte der $N(1/2, 1/(4n))$ -Verteilung. Man beachte, dass in der „Mitte“ der Verteilung die Normalapproximation besonders gut ist; an den „Rändern“ der Verteilung ist sie weniger gut. (Man überprüfe das als UE-Aufgabe.) ■

Abbildung 7.4: Asymptotische Normalverteilung der empVF (Bsp 7.12)

Im Folgenden ein Überblick über die wichtigsten Eigenschaften der ML–Schätzmethoden. Dadurch wird deutlich, dass ML–Schätzer **asymptotisch** (d. h. für $n \rightarrow \infty$) **optimale Schätzer** sind. (Bem: Bis auf (1) sind alle Eigenschaften asymptotischer Natur.)

Eigenschaften von Maximum-Likelihood-Schätzern: Unter bestimmten *Regularitätsvoraussetzungen*⁹ sind ML–Schätzer:

- (1) invariant
- (2) asymptotisch erwartungstreu
- (3) asymptotisch effizient
- (4) konsistent
- (5) asymptotisch normalverteilt

⁹Erfüllt für eine große Klasse von Verteilungen (Normal, Gamma, Poisson, ...); vgl. HOGG ET AL. (2005) für eine ausführliche Diskussion der Bedingungen.

7.3 Konfidenzintervalle

Ein wesentlicher Teil jeder Schätzprozedur sind Aussagen betreffend die **Genauigkeit** (oder **Präzision**) der Schätzer. Ohne derartige Aussagen wäre die bloße Angabe von Schätzwerten für z. B. Verteilungsparameter nur von geringer Bedeutung. Allgemein bieten sog. **Intervallschätzer** eine präzise Möglichkeit zur Beschreibung der Ungenauigkeit in den Schätzwerten.

Ist $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ eine Stichprobe der sG X , deren Verteilung von einem unbekannten Parameter $\theta \in \Theta$ abhängt und sind $T_1(\mathbf{X}) < T_2(\mathbf{X})$ zwei Funktionen der Stichprobe, so nennt man das Zufallsintervall $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ ein **Konfidenzintervall** (kurz **KI**) für θ mit **Konfidenzkoeffizient** $1 - \alpha$, wenn:

$$P_\theta(T_1(\mathbf{X}) < \theta < T_2(\mathbf{X})) \geq 1 - \alpha \quad \text{für alle } \theta \in \Theta$$

Gilt die obige Aussage nur approximativ, spricht man von einem **approximativen Konfidenzintervall**. Die Wahrscheinlichkeit $P_\theta(T_1(\mathbf{X}) < \theta < T_2(\mathbf{X}))$ nennt man die **Überdeckungswahrscheinlichkeit**¹⁰ (kurz **ÜW**). Für ein exaktes $(1 - \alpha)$ -Konfidenzintervall beträgt die ÜW mindestens $1 - \alpha$ für alle $\theta \in \Theta$; für approximative KIe kann die tatsächliche ÜW für bestimmte θ auch kleiner als $1 - \alpha$ sein.

Es gibt mehrere Methoden zur Konstruktion von Konfidenzintervallen. Wir behandeln im Folgenden eine klassische – auf R. A. FISHER zurückgehende – Methode und eine auf Simulation basierende Methode etwas genauer.

7.3.1 Pivotmethode

Unter einer **Pivotgröße** (kurz **Pivot**¹¹) versteht man eine sG $T = T(\mathbf{X}, \theta)$, die eine Funktion der Stichprobe \mathbf{X} und des Parameters θ ist, deren Verteilung aber bekannt ist und *nicht* von θ abhängt.

Bsp 7.13 X_1, X_2, \dots, X_n sei eine Stichprobe von $X \sim N(\mu, 1)$ (d. h., σ^2 sei bekannt und gleich 1). Wie schon mehrfach diskutiert, lässt sich μ durch den Stichprobenmittelwert \bar{X}_n schätzen. Für Letzteren gilt in diesem Fall:

$$\bar{X}_n \sim N\left(\mu, \frac{1}{n}\right) \implies T = \frac{\bar{X}_n - \mu}{1/\sqrt{n}} = \sqrt{n}(\bar{X}_n - \mu) \sim N(0, 1)$$

T ist eine Funktion der Stichprobe und des Parameters μ , die Verteilung von T ist aber bekannt und hängt nicht von μ ab. D. h., T ist eine Pivotgröße. Um nun ein KI für μ

¹⁰engl. *coverage probability*

¹¹pivot engl./franz. = Dreh-, Angelpunkt

mit Konfidenzkoeffizient $1 - \alpha$ für μ zu konstruieren, nehmen wir das $(\alpha/2)$ - und das $(1 - \alpha/2)$ -Quantil der **Pivotverteilung** (hier $N(0, 1)$):

$$P_\mu(z_{\alpha/2} < T < z_{1-\alpha/2}) = 1 - \alpha \quad \text{für alle } \mu \in \mathbb{R}$$

Der Ausdruck in Klammern lässt sich äquivalent wie folgt schreiben:

$$\underbrace{\bar{X}_n - z_{1-\alpha/2} \frac{1}{\sqrt{n}}}_{T_1} < \mu < \underbrace{\bar{X}_n + z_{1-\alpha/2} \frac{1}{\sqrt{n}}}_{T_2}$$

(Man beachte, dass $z_{\alpha/2} = -z_{1-\alpha/2}$.) Das Zufallsintervall (T_1, T_2) ist symmetrisch um \bar{X}_n ; daher schreibt man manchmal auch kürzer:

$$\bar{X}_n \pm z_{1-\alpha/2} \frac{1}{\sqrt{n}}$$

Wie lässt sich dieses KI interpretieren? zieht man wiederholt Stichproben der Größe n aus $X \sim N(\mu, 1)$ und bestimmt jeweils das obige KI, so werden etwa $100(1 - \alpha)\%$ dieser Intervalle das wahre μ überdecken. Das lässt sich mittels einer Simulation empirisch überprüfen.

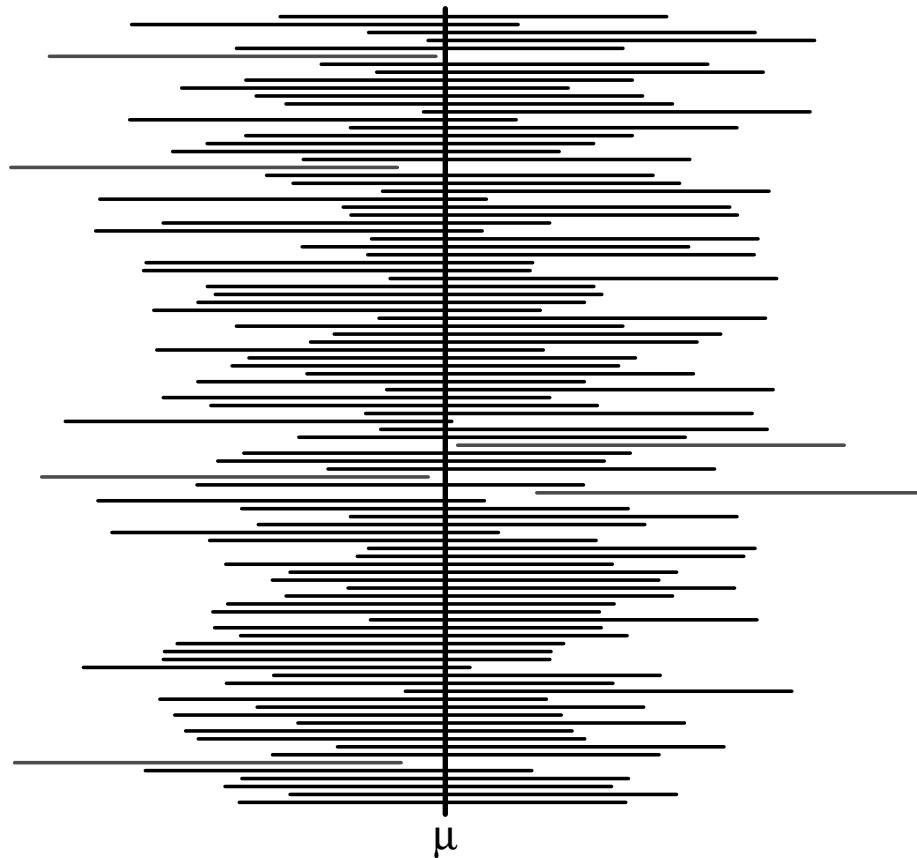
In Abb 7.5 ist das Ergebnis einer Simulation von 100 Stichproben der Größe $n = 10$ aus $X \sim N(\mu, 1)$ für $\mu = 0$ und $\alpha = 0.05$ grafisch dargestellt. Beim abgebildeten Durchlauf überdecken 6 der 100 Intervalle den wahren Wert von μ nicht. Die geschätzte ÜW des Konfidenzintervalls beträgt also 94%. ■

In Verallgemeinerung des obigen Beispiels besteht die Konstruktion von Konfidenzintervallen mittels Pivotmethode aus den folgenden Schritten:

- (1) Formuliere ein statistisches Modell für die Stichprobe \mathbf{X} .
- (2) Wähle eine geeignete Pivotgröße $T(\mathbf{X}, \theta)$.
- (3) Bestimme die Verteilung des Pivots.
- (4) Bestimme zwei Quantile q_1 und q_2 der Pivotverteilung, sodass:

$$P(q_1 < T(\mathbf{X}, \theta) < q_2) = 1 - \alpha$$

- (5) Bringe das Ereignis $\{q_1 < T(\mathbf{X}, \theta) < q_2\}$ in die Form $\{T_1(\mathbf{X}) < \theta < T_2(\mathbf{X})\}$.
- (6) $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ ist ein $100(1 - \alpha)\%$ -Konfidenzintervall für θ .

Abbildung 7.5: 95%-Konfidenzintervalle für μ Bemerkungen:

- (a) Üblicherweise wählt man für q_1 das $(\alpha/2)$ - und für q_2 das $(1 - \alpha/2)$ -Quantil der Pivotverteilung. In diesem Fall spricht man von **Equal-Tails-Konfidenzintervallen**.
- (b) Je kleiner α umso breiter das KI; sehr breite KIe sind aber nur von geringer praktischer Relevanz. Übliche Werte für α sind 0.01, 0.05 oder 0.1.
- (c) Vielfach findet man keine exakten sondern nur **approximative Pivots** (etwa auf Basis des ZGVS). Dabei ist zu beachten, dass bei kleinen (oder auch mittleren) Stichprobengrößen die tatsächliche ÜW von mit approximativen Pivots konstruierten KIn u. U. erheblich vom nominellen $1 - \alpha$ abweichen kann.

7.3.2 Approximativer Konfidenzintervall für den Mittelwert

X_1, X_2, \dots, X_n sei eine Stichprobe von einer sG X mit Mittelwert μ und Varianz $\sigma^2 < \infty$, wobei beide Parameter unbekannt seien. Vom ZGVS (vgl. 6.3.3) wissen wir, dass:

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) \implies \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

Letztere Größe ist noch keine (approximative) Pivotgröße für μ , da sie noch von der (unbekannten) Streuung σ abhängt. Die Stichprobenstreuung S_n ist aber ein konsistenter Schätzer für σ (vgl. Bsp 7.11). Ersetzt man σ durch S_n , bekommt man einen approximativen Pivot für μ :

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \approx N(0, 1)$$

Damit folgt für großes n :

$$P\left(\bar{X}_n - z_{1-\alpha/2} \frac{S_n}{\sqrt{n}} < \mu < \bar{X}_n + z_{1-\alpha/2} \frac{S_n}{\sqrt{n}}\right) \approx 1 - \alpha$$

Ein approximativer $(1 - \alpha)$ -Konfidenzintervall für μ ist also gegeben durch:

$$\bar{X}_n \pm z_{1-\alpha/2} \frac{S_n}{\sqrt{n}}$$

Bsp 7.14 [Monte Carlo Integration] Angenommen, wir möchten $I = \int_0^\infty \sqrt{x} e^{-x} dx$ berechnen. Das Integral lässt sich auf die Gammafunktion (vgl. 4.2.3) zurückführen:

$$I = \int_0^\infty x^{3/2-1} e^{-x} dx = \Gamma\left(\frac{3}{2}\right) = \left(\frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2} \doteq 0.8862$$

Das Integral lässt sich aber auch als Erwartungswert von $Y = \sqrt{X}$, wobei $X \sim \text{Exp}(1)$, interpretieren. Diesen Erwartungswert kann man auf Basis einer Stichprobe X_1, X_2, \dots, X_n von X wie folgt konsistent schätzen:

$$\hat{I} = \frac{1}{n} \sum_{i=1}^n \sqrt{X_i}$$

Außerdem lässt sich mittels eines (beispielsweise) 95%-Konfidenzintervalls für $\mathbb{E}(\sqrt{X})$ eine Aussage über die Genauigkeit der Schätzung machen.

```

n <- 10^6
x <- rexp(n, rate=1)
y <- sqrt(x)
alph <- 0.05
options(digits=5)
(Ihat <- mean(y))
[1] 0.88617
(Ihat + c(-1,1)*qnorm(1-alph/2)*sd(y)/sqrt(n))
[1] 0.88526 0.88708

```

Man beachte, dass der wahre Wert von I im 95%-KI enthalten ist. ■

7.3.3 Normalverteilung (eine Stichprobe)

Auf Basis einer Stichprobe X_1, X_2, \dots, X_n von $X \sim N(\mu, \sigma^2)$ können exakte Konfidenzintervalle für μ und σ^2 konstruiert werden. Dazu benötigen wir das folgende fundamentale Resultat.¹²

Behauptung: Für eine Stichprobe X_1, X_2, \dots, X_n von $X \sim N(\mu, \sigma^2)$ gilt:

$$(1) \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$(2) \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

(3) \bar{X}_n und S_n^2 sind (stochastisch) unabhängig.

$$(4) \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

Man beachte, dass die Größen von (2) und (4) Pivotgrößen für μ bzw. σ^2 sind. Auf Basis dieser Pivots lassen sich exakte Konfidenzintervalle konstruieren.

Konfidenzintervall für μ : Auf Basis von Behauptung (4) gilt wegen $t_{n-1; \alpha/2} = -t_{n-1; 1-\alpha/2}$:

$$P \left(-t_{n-1; 1-\alpha/2} < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < t_{n-1; 1-\alpha/2} \right) = 1 - \alpha$$

¹²Häufig als *Hauptsatz der mathematischen Statistik* oder als *Satz von Student* bezeichnet.

Ein $(1 - \alpha)$ -Konfidenzintervall für μ ist also gegeben durch:

$$\left(\bar{X}_n - t_{n-1; 1-\alpha/2} \frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{n-1; 1-\alpha/2} \frac{S_n}{\sqrt{n}} \right)$$

In der Kurzform:

$$\bar{X}_n \pm t_{n-1; 1-\alpha/2} \frac{S_n}{\sqrt{n}}$$

Konfidenzintervall für σ^2 : Auf Basis von Behauptung (2) gilt:

$$P \left(\chi^2_{n-1; \alpha/2} < \frac{(n-1)S_n^2}{\sigma^2} < \chi^2_{n-1; 1-\alpha/2} \right) = 1 - \alpha$$

Ein $(1 - \alpha)$ -Konfidenzintervall für σ^2 ist also gegeben durch:

$$\left(\frac{(n-1)S_n^2}{\chi^2_{n-1; 1-\alpha/2}}, \frac{(n-1)S_n^2}{\chi^2_{n-1; \alpha/2}} \right)$$

Konfidenzintervall für σ : Zieht man im $(1 - \alpha)$ -KI für σ^2 auf beiden Seiten die Wurzel, bekommt man ein $(1 - \alpha)$ -KI für σ :

$$\left(\sqrt{\frac{(n-1)S_n^2}{\chi^2_{n-1; 1-\alpha/2}}}, \sqrt{\frac{(n-1)S_n^2}{\chi^2_{n-1; \alpha/2}}} \right)$$

7.3.4 Normalverteilung (zwei ua. Stichproben)

Hat man Stichproben X_1, X_2, \dots, X_m und Y_1, Y_2, \dots, Y_n von zwei ua. sGn $X \sim N(\mu_X, \sigma_X^2)$ bzw. $Y \sim N(\mu_Y, \sigma_Y^2)$, lassen sich Konfidenzintervalle für die Differenz der Mittelwerte $\mu_X - \mu_Y$ bzw. für den Quotienten der Varianzen σ_X^2 / σ_Y^2 konstruieren.

Die Konstruktion eines KIs für $\mu_X - \mu_Y$ verläuft ähnlich wie im Falle einer Stichprobe, vorausgesetzt man trifft die zusätzliche **Annahme**, dass die beiden **Varianzen gleich** sind, d. h., dass $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (unbekannt). In diesem Fall gilt zunächst:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{1/m + 1/n}} \sim N(0, 1)$$

Die gemeinsame Varianz σ^2 lässt sich durch einen **gepoolten Varianzschätzer**¹³, d. h. durch einen **gewichteten Mittelwert** der beiden Stichprobenvarianzen S_X^2 und S_Y^2 , erwartungstreu schätzen:

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

Ersetzt man σ^2 durch S_p^2 , bekommt man einen Pivot für $\mu_X - \mu_Y$:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{1/m + 1/n}} \sim t(m+n-2)$$

Konfidenzintervall für $\mu_X - \mu_Y$: Unter der Voraussetzung $\sigma_X^2 = \sigma_Y^2$ ist ein $(1-\alpha)$ -KI für $\mu_X - \mu_Y$ gegeben durch:

$$\bar{X} - \bar{Y} \pm t_{m+n-2; 1-\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

Bem: Lässt man die Voraussetzung $\sigma_X^2 = \sigma_Y^2$ fallen, gibt es keinen *exakten* Pivot für $\mu_X - \mu_Y$, wohl aber *approximative* Pivots. Sind beide Stichprobengrößen m und n nicht zu klein, kann man etwa das folgende approximative $(1-\alpha)$ -KI für $\mu_X - \mu_Y$ nehmen:

$$\bar{X} - \bar{Y} \pm z_{1-\alpha/2} \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}$$

Für die Konstruktion von KIn für σ_X^2/σ_Y^2 benötigen wir das folgende Resultat.

Behauptung: Für zwei unabhängige Stichproben X_1, X_2, \dots, X_m und Y_1, Y_2, \dots, Y_n von $X \sim N(\mu_X, \sigma_X^2)$ bzw. $Y \sim N(\mu_Y, \sigma_Y^2)$ gilt:

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1, n-1)$$

Konfidenzintervall für σ_X^2/σ_Y^2 : Auf Basis der obigen Behauptung gilt:

$$P\left(F_{m-1, n-1; \alpha/2} < \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} < F_{m-1, n-1; 1-\alpha/2}\right) = 1 - \alpha$$

¹³engl. *pooled sample variance*

Ein $(1 - \alpha)$ -Konfidenzintervall für σ_X^2 / σ_Y^2 ist also gegeben durch:

$$\left(\frac{1}{F_{m-1,n-1;1-\alpha/2}} \frac{S_X^2}{S_Y^2}, \frac{1}{F_{m-1,n-1;\alpha/2}} \frac{S_X^2}{S_Y^2} \right)$$

Bem: Bei der Verwendung von Tabellen für die Bestimmung der F-Quantile ist zu beachten, dass meist nur $F_{m-1,n-1;1-\alpha/2}$ tabelliert ist; für $F_{m-1,n-1;\alpha/2}$ verwendet man die folgende Beziehung:

$$F_{m-1,n-1;\alpha/2} = \frac{1}{F_{n-1,m-1;1-\alpha/2}}$$

7.3.5 Normalverteilung (verbundene Stichproben)

Ein praktisch wichtiges Problem ist die Entwicklung von Konfidenzintervallen für die Differenz der Mittelwerte, wenn die Stichproben **abhängig** (oder **verbunden**) sind. Das betrifft in erster Linie Vorher/Nachher-Situationen (u. Ä.) an *denselben* Untersuchungseinheiten. In derartigen Situationen würde das t-Intervall für unabhängige Stichproben einen falschen Eindruck vermitteln. Die korrekte Vorgangsweise in solchen Fällen ist die Bildung der Differenzen $D_i = X_i - Y_i$ der Beobachtungen.

Nach Folgerung 1 von 6.1 gilt für $D = X - Y$:

$$D \sim N\left(\underbrace{\mu_X - \mu_Y}_{\mu_D}, \underbrace{\sigma_X^2 + \sigma_Y^2 - 2 \operatorname{Cov}(X, Y)}_{\sigma_D^2}\right) = N(\mu_D, \sigma_D^2)$$

Nun kann man auf Basis der Stichprobe D_1, D_2, \dots, D_n von $D \sim N(\mu_D, \sigma_D^2)$ ein Konfidenzintervall für $\mu_D = \mu_X - \mu_Y$ bestimmen.

Bsp 7.15 Belastend bei gleichförmiger Bildschirmarbeit wirken u. a. die vielen *kleinen* Bewegungen des Oberarms (Hebungen um weniger als 30°). In einer Studie an 16 Personen wurde der Zeitanteil der Arbeitszeit mit Bewegungen des Oberarms um weniger als 30° erhoben. Einige Monate später wurde diese Untersuchung an denselben Personen wiederholt, wobei in der Zwischenzeit eine Umstellung der Arbeit vorgenommen wurde. Hat sich der Anteil der Arbeitszeit mit Bewegungen des Oberarms um weniger als 30° signifikant verändert?

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Vorher	81	87	86	82	90	86	96	73	74	75	72	80	66	72	56	82
Nachher	78	91	78	78	84	67	92	70	58	62	70	58	66	60	65	73
Differenz	3	-4	8	4	6	19	4	3	16	13	2	22	0	12	-9	9

Abbildung 7.6: Boxplot für die Differenzen $d_i = x_i - y_i$

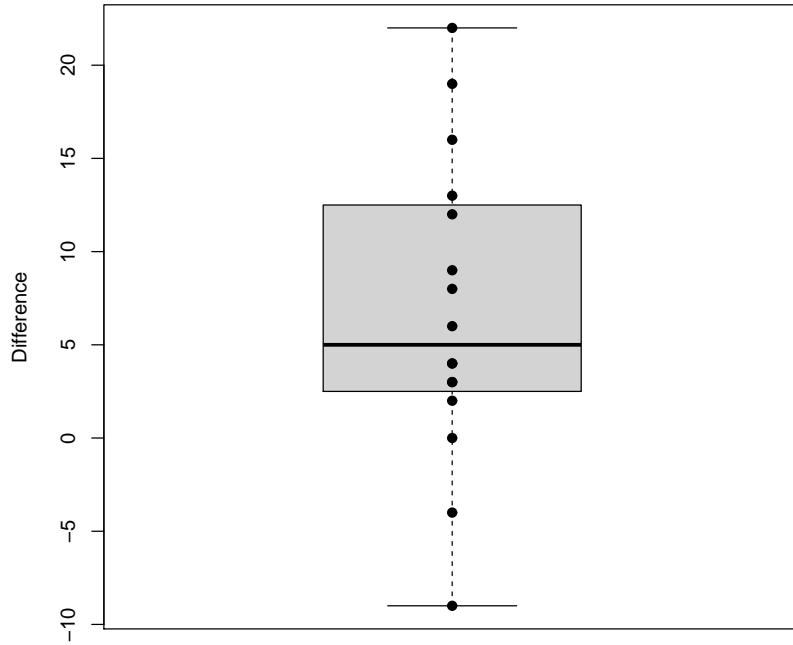


Abb 7.6 zeigt den Boxplot für die Differenzen $d_i = x_i - y_i$, $i = 1, 2, \dots, 16$. Die Box liegt zur Gänze im positiven Bereich, sodass bereits hier eine signifikante Abnahme des Zeitanteils mit kleinen Bewegungen behauptet werden kann.

Der folgende R–Ouput zeigt die Berechnung eines 95%–Konfidenzintervalls für μ_D mit Hilfe der Funktion `t.test()`. Als Kontrast wird auch ein 95%–Konfidenzintervall für $\mu_X - \mu_Y$ unter der Annahme unabhängiger Stichproben berechnet (Varianzen gleich).

```
x <- c(81,87,86,82,90,86,96,73,74,75,72,80,66,72,56,82)
y <- c(78,91,78,78,84,67,92,70,58,62,70,58,66,60,65,73)
d <- x-y

t.test(x, y, paired=TRUE)$conf.int
[1] 2.3624 11.1376
attr(),"conf.level")
[1] 0.95

t.test(x, y, var.equal=TRUE)$conf.int
[1] -0.74626 14.24626
attr(),"conf.level")
[1] 0.95
```

Die Schlussfolgerungen sind ganz verschieden; im ersten (korrekten) Fall zeigt sich eine deutliche Signifikanz (Null ist kein Element des Intervalls), im zweiten Fall aber nicht. ■

7.3.6 Exponentialverteilung

Auf Basis einer Stichprobe X_1, X_2, \dots, X_n von $X \sim \text{Exp}(\tau)$ (τ = Erwartungswert von X) ist der ML-Schätzer von τ gegeben durch:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$$

Mit dem Additionstheorem für Exp-Verteilungen (vgl. 6.2.3) – und mit dem Transformationssatz (vgl. 3.3.2) – zeigt man, dass:

$$\frac{2n\bar{X}_n}{\tau} \sim \chi^2(2n)$$

eine (exakte) Pivotgröße für τ ist.

Exaktes Konfidenzintervall für τ : Auf Basis des obigen Pivots ist ein exaktes $(1 - \alpha)$ -Konfidenzintervall für τ gegeben durch:

$$\left(\frac{2n\bar{X}_n}{\chi^2_{2n; 1-\alpha/2}}, \frac{2n\bar{X}_n}{\chi^2_{2n; \alpha/2}} \right)$$

Wir betrachten noch zwei *approximative* KIe. Die Streuung von X ist τ ; nach dem ZGVS (vgl. 6.3.3) gilt daher für großes n :

$$\frac{\hat{\tau} - \tau}{\tau/\sqrt{n}} \approx N(0, 1) \quad (*)$$

Ersetzt man im Nenner τ durch den (konsistenten) Schätzer $\hat{\tau}$, bekommt man einen approximativen Pivot¹⁴ für τ :

$$\frac{\hat{\tau} - \tau}{\hat{\tau}/\sqrt{n}} \approx N(0, 1) \quad (**)$$

Auf Basis dieses Pivots lässt sich nun einfach ein (approximatives) KI für τ konstruieren.

¹⁴Tatsächlich sind auch die Größen $(**)$ und $(*)$ *exakte* Pivots, deren Verteilungen – für großes n – durch $N(0, 1)$ approximiert werden können.

Bem: Nach diesem Prinzip konstruierte KIe werden in der Literatur meist **Wald–Intervalle**¹⁵ genannt.

Wald–Intervall für τ : Auf Basis des Pivots (***) ist ein approximatives $(1 - \alpha)$ –Konfidenzintervall für τ gegeben durch:

$$\left(\hat{\tau} - z_{1-\alpha/2} \frac{\hat{\tau}}{\sqrt{n}}, \hat{\tau} + z_{1-\alpha/2} \frac{\hat{\tau}}{\sqrt{n}} \right) = \hat{\tau} \pm z_{1-\alpha/2} \frac{\hat{\tau}}{\sqrt{n}}$$

Auf Basis des Pivots (*) gilt:

$$P\left(-z_{1-\alpha/2} < \frac{\hat{\tau} - \tau}{\tau/\sqrt{n}} < z_{1-\alpha/2}\right) \approx 1 - \alpha$$

Die obige Doppelungleichung lässt sich einfach nach τ auflösen.

Bem: Nach diesem Prinzip konstruierte KIe werden in der Literatur meist **Scoreintervalle** genannt. (Bem: Der Ausdruck stammt aus der Maximum-Likelihood-Theorie.)

Scoreintervall für τ : Auf Basis des Pivots (*) ist ein approximatives $(1 - \alpha)$ –Konfidenzintervall für τ gegeben durch:

$$\left(\frac{\hat{\tau}}{1 + z_{1-\alpha/2}/\sqrt{n}}, \frac{\hat{\tau}}{1 - z_{1-\alpha/2}/\sqrt{n}} \right)$$

7.3.7 Bernoulli–Verteilung

Auf Basis einer Stichprobe X_1, X_2, \dots, X_n von $X \sim A(p)$ (Bernoulli–Verteilung) ist der ML–Schätzer von p gegeben durch:

$$\hat{p} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = \text{Anteil der Einser in der Stichprobe}$$

In diesem Fall ist es schwierig, einen *exakten* Pivot für p zu finden. Ist n nicht zu klein, kann man sich aber auf den ZGVS berufen (vgl. 6.3.4):

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right) \implies \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0, 1)$$

¹⁵Nach ABRAHAM WALD (1902–1950), geb. in Siebenbürgen (damals Ungarn); gehört zu den bedeutendsten Statistikern des 20. Jh.

Auf Basis dieses *approximativen* Pivots gilt:

$$P\left(-z_{1-\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{1-\alpha/2}\right) \approx 1 - \alpha \quad (*)$$

Damit folgt:

$$P\left(\hat{p} - z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha$$

Ersetzt man im Wurzausdruck das unbekannte p durch den (konsistenten) Schätzer \hat{p} , lautet ein approximatives $(1 - \alpha)$ -Konfidenzintervall für p wie folgt:

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Das ist das **Standardintervall** (oder **Wald–Intervall**) für p . Als Alternative kann man auch die Doppelungleichung (*) nach p auflösen. Setzt man $a = z_{1-\alpha/2}$, lautet die zu lösende quadratische Gleichung wie folgt:

$$n(\hat{p} - p)^2 = a^2 p(1 - p)$$

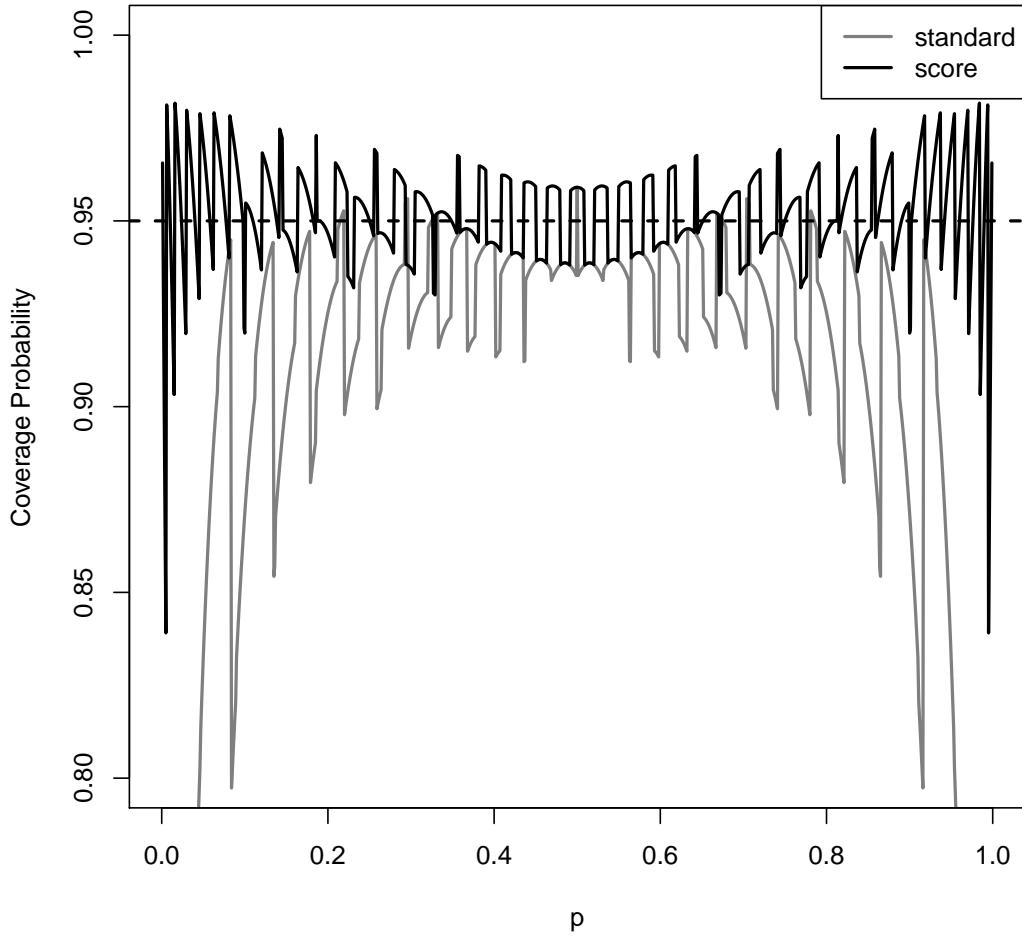
Die quadratische Gleichung hat zwei reelle Lösungen:

$$T_{1,2} = \frac{a^2 + 2n\hat{p} \pm a \sqrt{a^2 + 4n\hat{p}(1-\hat{p})}}{2(a^2 + n)}$$

Das Intervall (T_1, T_2) ist das **Scoreintervall** für p . Wie sich zeigt (vgl. das folgende Beispiel) hat es hinsichtlich ÜW – speziell an den Rändern, d. h. für kleine oder große Werte von p – deutlich bessere Eigenschaften als das Standardintervall.

Bsp 7.16 Es ist interessant, die tatsächliche ÜW des Standard- und des Scoreintervalls zu bestimmen und miteinander zu vergleichen. (Bem: Die ÜW lässt sich direkt durch Abzählen bestimmen; vgl. den R-Code.) Abb 7.7 zeigt die ÜW als Funktion von p für $n = 35$ und $\alpha = 0.05$. Die ÜW des Scoreintervalls stimmt über den ganzen Bereich $0 < p < 1$ – insbesondere aber an den Rändern – deutlich besser mit der nominellen ÜW von $1 - \alpha = 0.95$ überein als die ÜW des Standardintervalls. (Bem: Der gezackte Verlauf der beiden Kurven ist typisch für diskrete sGn.) ■

Bem: Es gibt auch „exakte“ Konfidenzintervalle für p , die sog. **Pearson–Clopper–Intervalle**. Ihre Herleitung ist allerdings schwieriger und wird hier nicht weiter diskutiert. (Die R-Funktion `binom.test()` ist eine Implementierung der exakten Prozedur.)

Abbildung 7.7: Vergleich der ÜW des Standard– und das Scoreintervalls für p 

7.3.8 Poisson–Verteilung

Auf Basis einer Stichprobe X_1, X_2, \dots, X_n von $X \sim P(\lambda)$ ist der ML–Schätzer von λ gegeben durch:

$$\hat{\lambda} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Auch in diesem Fall ist es schwierig, einen *exakten* Pivot für λ zu finden. Ist n nicht zu klein, kann man sich aber auf den ZGVS berufen (vgl. 6.3.4):

$$\hat{\lambda} \approx N\left(\lambda, \frac{\lambda}{n}\right) \implies \frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}} \approx N(0, 1)$$

Auf Basis dieses *approximativen* Pivots gilt:

$$P\left(-z_{1-\alpha/2} < \frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}} < z_{1-\alpha/2}\right) \approx 1 - \alpha \quad (*)$$

Damit folgt:

$$P\left(\hat{\lambda} - z_{1-\alpha/2} \sqrt{\frac{\lambda}{n}} < p < \hat{\lambda} + z_{1-\alpha/2} \sqrt{\frac{\lambda}{n}}\right) \approx 1 - \alpha$$

Ersetzt man im Wurzelausdruck das unbekannte λ durch den (konsistenten) Schätzer $\hat{\lambda}$, lautet das **Standardintervall** (oder **Wald–Intervall**) für λ wie folgt:

$$\hat{\lambda} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\lambda}}{n}}$$

Als Alternative kann man auch die Doppelungleichung (*) nach λ auflösen. Setzt man $a = z_{1-\alpha/2}$, lautet die zu lösende quadratische Gleichung wie folgt:

$$n(\hat{\lambda} - \lambda)^2 = a^2 \lambda$$

Die quadratische Gleichung hat zwei reelle Lösungen:

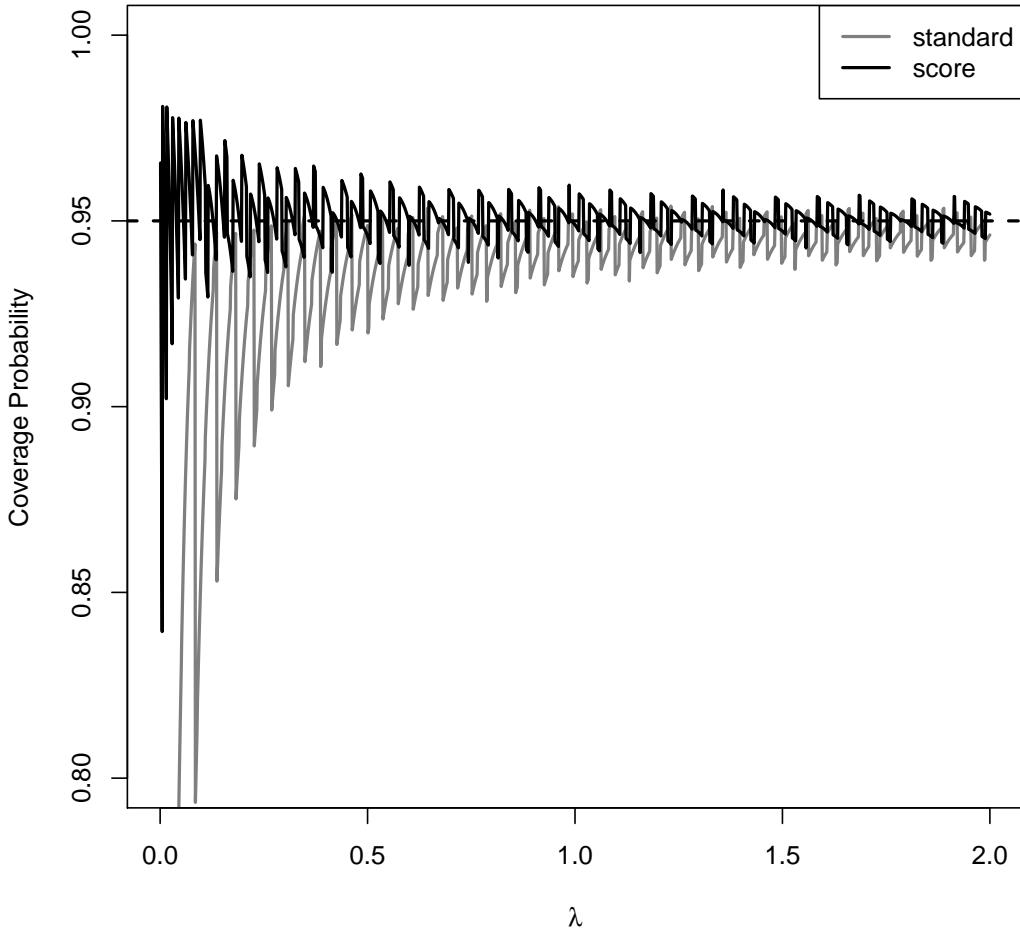
$$T_{1,2} = \frac{a^2 + 2n\hat{\lambda} \pm a \sqrt{a^2 + 4n\hat{\lambda}}}{2n}$$

Das Intervall (T_1, T_2) ist das **Scoreintervall** für λ . Wie sich zeigt (vgl. das folgende Beispiel) hat es hinsichtlich ÜW – speziell für kleine λ – deutlich bessere Eigenschaften als das Standardintervall.

Bsp 7.17 Ahnlich wie im Falle der Bernoulli–Verteilung ist es auch hier interessant, die tatsächliche ÜW des Standard– und des Scoreintervalls zu bestimmen und miteinander zu vergleichen. Abb 7.8 zeigt die ÜW als Funktion von λ für $n = 35$ und $\alpha = 0.05$. Die ÜW des Scoreintervalls stimmt über den hier betrachteten Bereich für λ – insbesondere aber für kleine λ –Werte – deutlich besser mit der nominellen ÜW von $1 - \alpha = 0.95$ überein als die ÜW des Standardintervalls. Für große λ –Werte werden die Unterschiede aber immer geringer. (Bem: Auch hier ist der stark gezackte Verlauf der Diskretheit der Poisson–Verteilung geschuldet.) ■

Bem: Es gibt auch für den Poissonparameter „exakte“ Konfidenzintervalle. Ihre Herleitung ist allerdings schwieriger und wird hier nicht weiter diskutiert.

Abbildung 7.8: Vergleich der ÜW des Standard– und das Scoreintervalls für λ



7.3.9 Resampling und Bootstrapping

Die Idee hinter dem **Resampling** ist sehr einfach: Hat man eine (konkrete) Stichprobe $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ aus einer stetigen sG X mit unbekannter Dichte f , so ist – ohne weitere Annahmen über f – die beste Information über f die Stichprobe \mathbf{x} selbst. Die beste Möglichkeit, das Experiment, das zu \mathbf{x} geführt hat, zu „wiederholen“, besteht nun darin, die (ursprüngliche) Stichprobe \mathbf{x} zur (neuen) Grundgesamtheit zu erklären und aus ihr (neue) Stichproben zu ziehen. Das entspricht einem Ziehen mit Zurücklegen aus einem Behälter, der die Elemente $\{x_i\}$ der ursprünglichen Stichprobe enthält.

Anders ausgedrückt: Das Resampling entspricht dem Ziehen von Stichproben von einer sG, deren Verteilungsfunktion gleich \widehat{F}_n (= empirische Verteilungsfunktion auf Basis der ursprünglichen Stichprobe) ist.

Durch Resampling kann man quasi das Experiment, das zur Originalstichprobe geführt hat, beliebig oft wiederholen. Das ist etwa dann sehr nützlich, wenn man etwas über die Eigenschaften einer auf Basis von \mathbf{x} bestimmten Statistik erfahren möchte. Hat man

beispielsweise auf Basis der Originaldaten die Statistik $T(\mathbf{x})$ bestimmt, kann man durch Resampling von \mathbf{x} Informationen über die Verteilung von $T(\mathbf{X})$ (also über die zugehörige sG) gewinnen.

Das **Bootstrapping** ist eine Formalisierung der Idee hinter dem Resampling. Wir betrachten hier nur die Bestimmung eines **Quantilen-Bootstrap-Konfidenzintervalls** für einen Parameter θ etwas detaillierter. Der Parameter werde auf eine bestimmte Weise auf Basis der Originalstichprobe $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ geschätzt, d. h. $\hat{\theta} = \hat{\theta}(\mathbf{x})$. Im folgenden Algorithmus bezeichnet B die Zahl der durch Resampling bestimmten Stichproben (üblich sind etwa $B = 3000$ oder mehr Resamples).

Algorithmus zur Bestimmung eines Bootstrap-Konfidenzintervalls:

- (1) Setze $j = 1$.
- (2) Solange $j \leq B$, gehe zu (3) – (5).
- (3) Ermittle durch Resampling eine Stichprobe \mathbf{x}_j^* der Größe n aus \hat{F}_n .
- (4) Bestimme $\hat{\theta}_j^* = \hat{\theta}(\mathbf{x}_j^*)$.
- (5) Setze $j = j + 1$.
- (6) Sind $\hat{\theta}_{(1)}^* \leq \hat{\theta}_{(2)}^* \leq \dots \leq \hat{\theta}_{(B)}^*$ die der Größe nach geordneten Werte von $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$, setze $m = \lfloor (\alpha/2)B \rfloor$ und bilde das Intervall:

$$(\hat{\theta}_{(m)}^*, \hat{\theta}_{(B-m+1)}^*)$$

Bem: Die offensichtlich selbstreferenzielle Vorgangsweise beim Bootstrapping¹⁶ wirkt auf den ersten Blick nicht sehr vielversprechend. Aber die einzige Information über die Variabilität der Stichprobe ist die (konkrete) Stichprobe selbst, und durch Resampling kann man diese Variabilität simulieren. Klarerweise gibt es Situationen, in denen diese Methode nicht funktioniert (jedenfalls nicht besser als herkömmliche Methoden). Andererseits zeigt aber die Praxis, dass es in zahlreichen Fällen funktioniert (vgl. das folgende Beispiel) und vielfach bessere Ergebnisse liefert als Methoden, die sich auf die Theorie der großen Stichproben berufen.

Bsp 7.18 Zur Illustration des Bootstrapping und zum Vergleich mit klassischen Methoden betrachten wir eine simulierte Stichprobe der Größe $n = 25$ aus einer $\text{Exp}(\tau = 3)$ -Verteilung. Von 7.3.6 kennen wir ein exaktes und zwei approximative Konfidenzintervalle für den Parameter τ . Wir berechnen für $\alpha = 5\%$ alle drei Intervalle und zusätzlich nach dem obigen Algorithmus das Bootstrapintervall auf Basis von $B = 3000$ Resamples der (simulierten) Stichprobe. Als Schätzer für τ (= Mittelwert der Verteilung) nehmen wir den Stichprobenmittelwert \bar{x} .

¹⁶bootstrap engl. = Stiefellasche, -riemen; Redewendung: *pull yourself up by your bootstraps* (= sich am eigenen Schopf aus dem Sumpf ziehen)

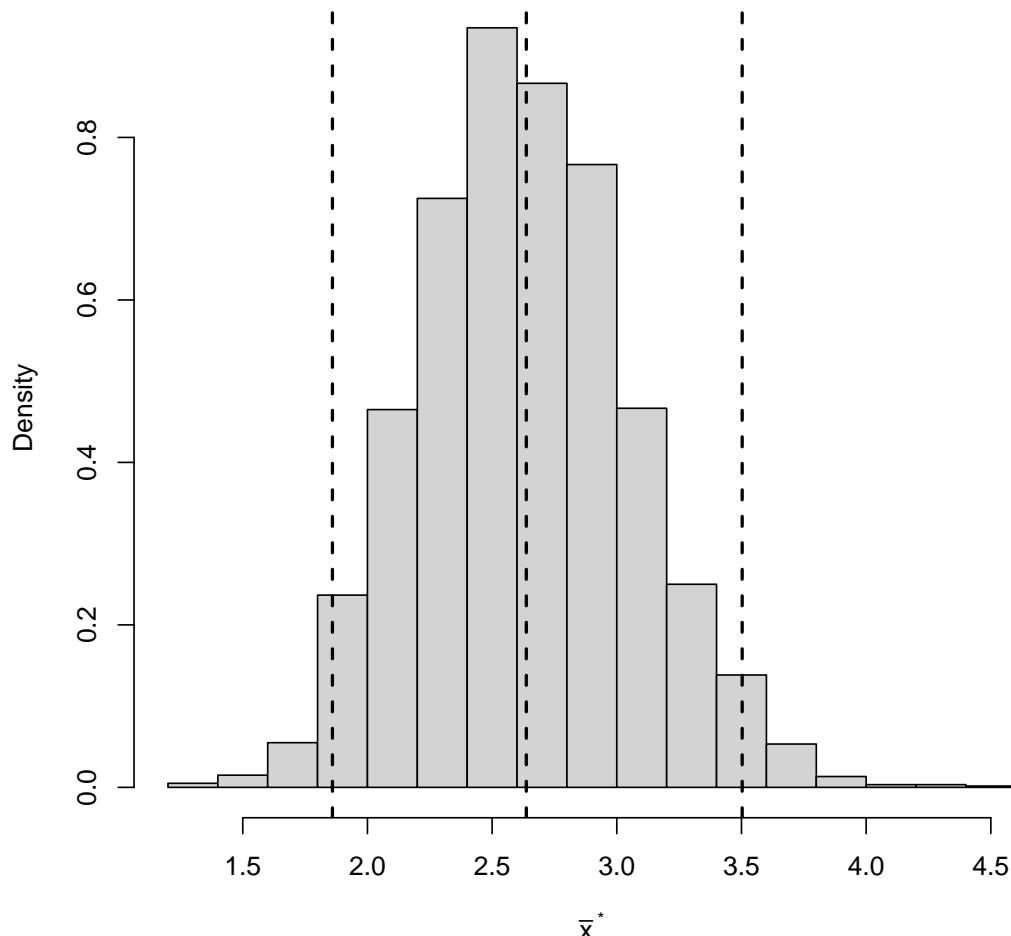
Abbildung 7.9: Histogramm der Bootstrapmittelwerte \bar{x}^* 

Abb 7.9 zeigt das Histogramm der Bootstrapmittelwerte; die äußeren strichlierten Linien markieren die Grenzen des 95%-Bootstrapintervalls, die mittlere Linie ist an der Stelle des Mittelwerts \bar{x} der Originalstichprobe. Der folgende R-Output zeigt eine Zusammenfassung aller vier Intervalle (vgl. `chap7.r` für den R-Code).

	Lower2.5%	Upper97.5%	Length
Exact	1.846	4.075	2.229
Wald	1.603	3.671	2.068
Score	1.895	4.337	2.443
Boot	1.860	3.503	1.643

Alle vier Intervalle überdecken den (hier bekannten) wahren Wert von τ ; das Bootstrapintervall ist aber mit Abstand am kürzesten. ■

7.4 Statistische Tests

Neben Punktschätzungen und Konfidenzintervallen betrachtet man in der schließenden Statistik auch das **Testen von statistischen Hypothesen**. Dabei unterscheidet man grundsätzlich zwischen Parameter- und Verteilungshypothesen. Unter einer **Parameterhypothese** versteht man eine Behauptung über den (oder die) Parameter von einer (oder mehreren) Verteilung(en). Bei dieser Art von Hypothesen wird angenommen, dass der Verteilungstyp bekannt ist (beispielsweise, dass es sich um eine Normalverteilung handelt). Ist der Verteilungstyp aber nicht bekannt und möchte man testen, ob eine bestimmte Verteilung oder eine bestimmte Verteilungsfamilie (beispielsweise, die Familie der Normalverteilungen) ein zufriedenstellendes Modell für die vorliegenden Beobachtungen darstellt, spricht man von einer **Verteilungshypothese**.

7.4.1 Parametertests

Wie in den vorigen Abschnitten nehmen wir an, dass unser Interesse einer sG $X \sim f(x; \theta)$ (oder $X \sim p(x; \theta)$) mit unbekanntem Parameter $\theta \in \Theta$ gilt. Auf Grund einer Theorie (oder einer Vermutung, einem früheren Experiment, ...) gelte $\theta \in \Theta_0$ oder $\theta \in \Theta_1$ mit $\Theta_0 \cap \Theta_1 = \emptyset$ und $\Theta_0 \cup \Theta_1 \subseteq \Theta$. Die erste Behauptung nennt man die **Nullhypothese**, die zweite die **Alternativ-** oder **Gegenhypothese** und schreibt das Testproblem wie folgt:

$$\mathcal{H}_0 : \theta \in \Theta_0 \quad \text{gegen} \quad \mathcal{H}_1 : \theta \in \Theta_1$$

Bem: Es ist nicht gleichgültig, welche Behauptung die Null- und welche die Gegenhypothese ist; das ist eine Folge der Asymmetrie des Testens. Als Nullhypothese wählt man in der Regel diejenige Behauptung, die die bisherige Situation oder den „Normalfall“ (oder „Status quo“) repräsentiert. Die Alternativhypothese ist häufig einfach das Komplement zur Nullhypothese, oder diejenige Behauptung, deren Zutreffen ein bestimmtes Handeln erfordert oder die gravierenderen Konsequenzen (positive oder negative) nach sich zieht.

Ein-/Zweiseitige Alternativhypothesen: Im Folgenden betrachten wir nur **einfache** Nullhypothesen der Form $\mathcal{H}_0 : \theta = \theta_0$. Lautet die Alternativhypothese $\theta \neq \theta_0$ nennt man sie **zweiseitig**:

$$\mathcal{H}_0 : \theta = \theta_0 \quad \text{gegen} \quad \mathcal{H}_1 : \theta \neq \theta_0$$

In den beiden folgenden Fällen nennt man die Alternativhypothese **einseitig**:

$$\mathcal{H}_0 : \theta = \theta_0 \quad \text{gegen} \quad \mathcal{H}_1 : \theta < \theta_0 \quad \text{oder} \quad \mathcal{H}_1 : \theta > \theta_0$$

Testentscheidung: Eine auf einer Stichprobe $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ von X basierende Entscheidungsregel über Hypothesen nennt man einen (statistischen) **Test**. Ein Test wird

durch seinen **kritischen Bereich** C charakterisiert. Dabei handelt es sich um eine Teilmenge des Stichprobenraumes M_X^n (= Menge aller möglichen Stichproben von X) mit:

Verwerfe \mathcal{H}_0 (Akzeptiere \mathcal{H}_1) falls $\mathbf{X} \in C$

Akzeptiere \mathcal{H}_0 (Verwerfe \mathcal{H}_1) falls $\mathbf{X} \in C^c$

Typ I/Typ II–Fehler: Allgemein unterscheidet man **Typ I–** und **Typ II–Fehler** (oder auch **Fehler 1. und 2. Art**). Der erste tritt auf, wenn die \mathcal{H}_0 verworfen wird, obwohl sie richtig ist; der zweite tritt auf, wenn die \mathcal{H}_0 nicht verworfen wird, obwohl sie falsch ist. Die folgende Tabelle zeigt die möglichen (Fehl–) Entscheidungen:

		Wahrer Zustand	
Entscheidung	\mathcal{H}_0 trifft zu	\mathcal{H}_1 trifft zu	
	Verwerfe \mathcal{H}_0	Typ I–Fehler	Korrekte Entscheidung
Akzeptiere \mathcal{H}_0	Korrekte Entscheidung		Typ II–Fehler

Die Wahrscheinlichkeit eines Typ I–Fehlers bezeichnet man mit α :

$$\alpha = P(\text{Typ I–Fehler}) = P_{\theta_0}(\mathbf{X} \in C)$$

Die Wahrscheinlichkeit eines Typ II–Fehlers bezeichnet man mit β :

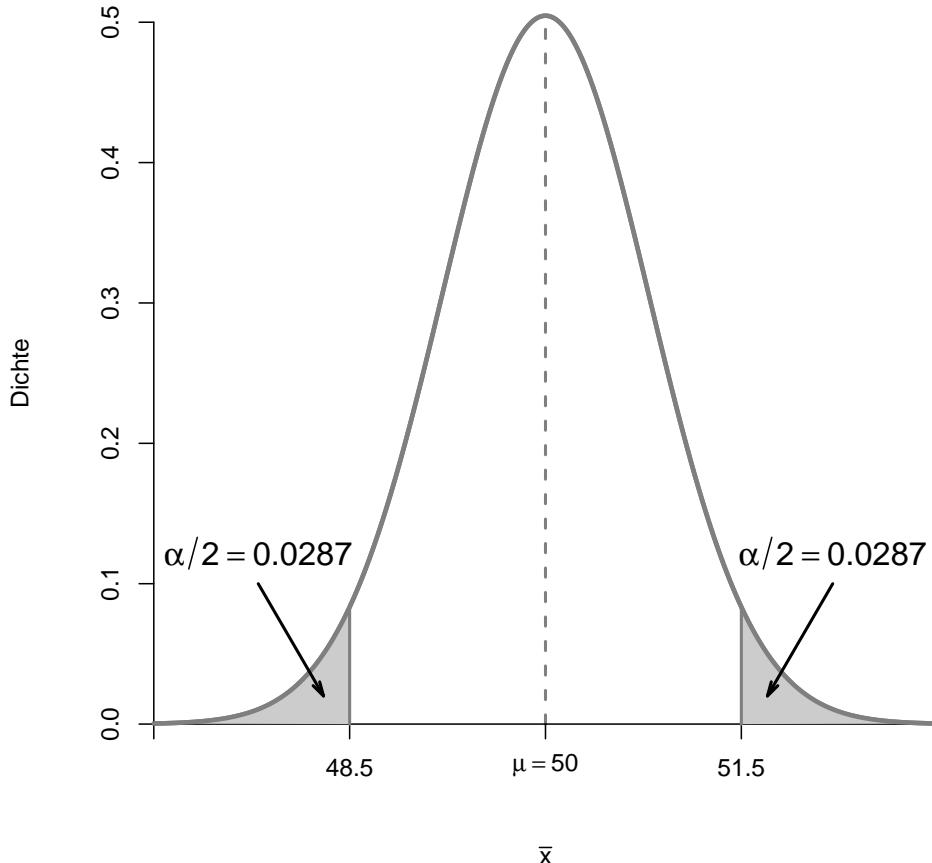
$$\beta = P(\text{Typ II–Fehler}) = P_{\theta}(\mathbf{X} \in C^c)$$

Um β berechnen zu können, brauchen wir einen spezifischen Wert θ aus der Alternativhypothese Θ_1 , d. h., $\beta = \beta(\theta)$ ist keine Konstante, sondern hängt vom wahren Wert des Parameters ab.

Bem: Die Wahrscheinlichkeit α eines Typ I–Fehlers nennt man auch das (**Signifikanz-**) **Niveau** des Tests.

Bsp 7.19 Als Illustration betrachten wir für $X \sim N(\mu, (2.5)^2)$ das folgende zweiseitige Testproblem:

$$\mathcal{H}_0 : \mu = 50 (= \mu_0) \quad \text{gegen} \quad \mathcal{H}_1 : \mu \neq 50$$

Abbildung 7.10: Dichte von \bar{X} unter \mathcal{H}_0 und kritischer Bereich

Angenommen, wir ziehen eine Stichprobe der Größe $n = 10$ und die Entscheidungsregel lautet: Verwerfe \mathcal{H}_0 , wenn $\bar{x} < 48.5$ oder $\bar{x} > 51.5$, andernfalls verwerfe \mathcal{H}_0 nicht. Der kritische Bereich des Tests ist also gegeben durch:

$$C = \{(x_1, x_2, \dots, x_n) \mid \bar{x} < 48.5 \text{ oder } \bar{x} > 51.5\}$$

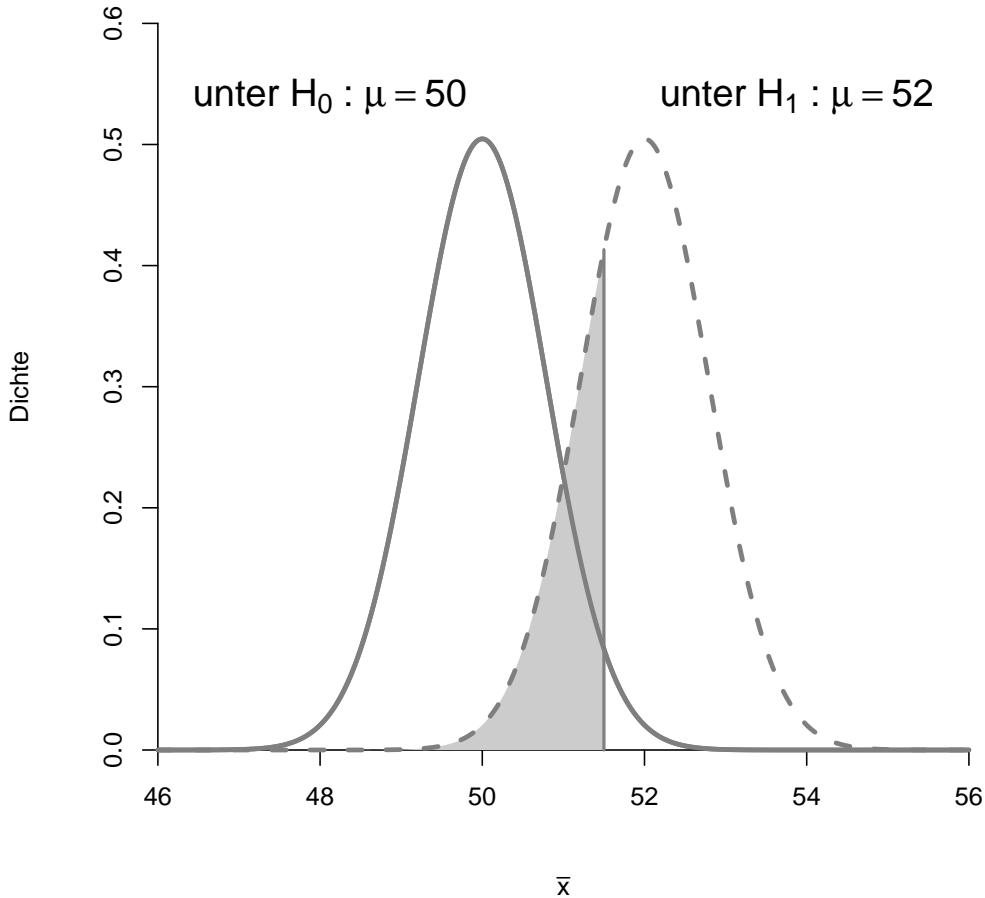
Wie groß ist bei diesem Test die Fehlerwahrscheinlichkeit 1. Art? Unter \mathcal{H}_0 gilt:

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right) = N\left(50, \frac{(2.5)^2}{10}\right)$$

Die Wahrscheinlichkeit für einen Typ I–Fehler ist daher gegeben durch:

$$\alpha = \Phi\left(\frac{48.5 - 50}{2.5/\sqrt{10}}\right) + \left[1 - \Phi\left(\frac{51.5 - 50}{2.5/\sqrt{10}}\right)\right] = 0.0287 + 0.0287 = 0.0574$$

(Vgl. Abb 7.10 für eine grafische Veranschaulichung.) Wie groß ist bei diesem Test die

Abbildung 7.11: Wahrscheinlichkeit eines Typ II–Fehlers für $\mu = 52$ 

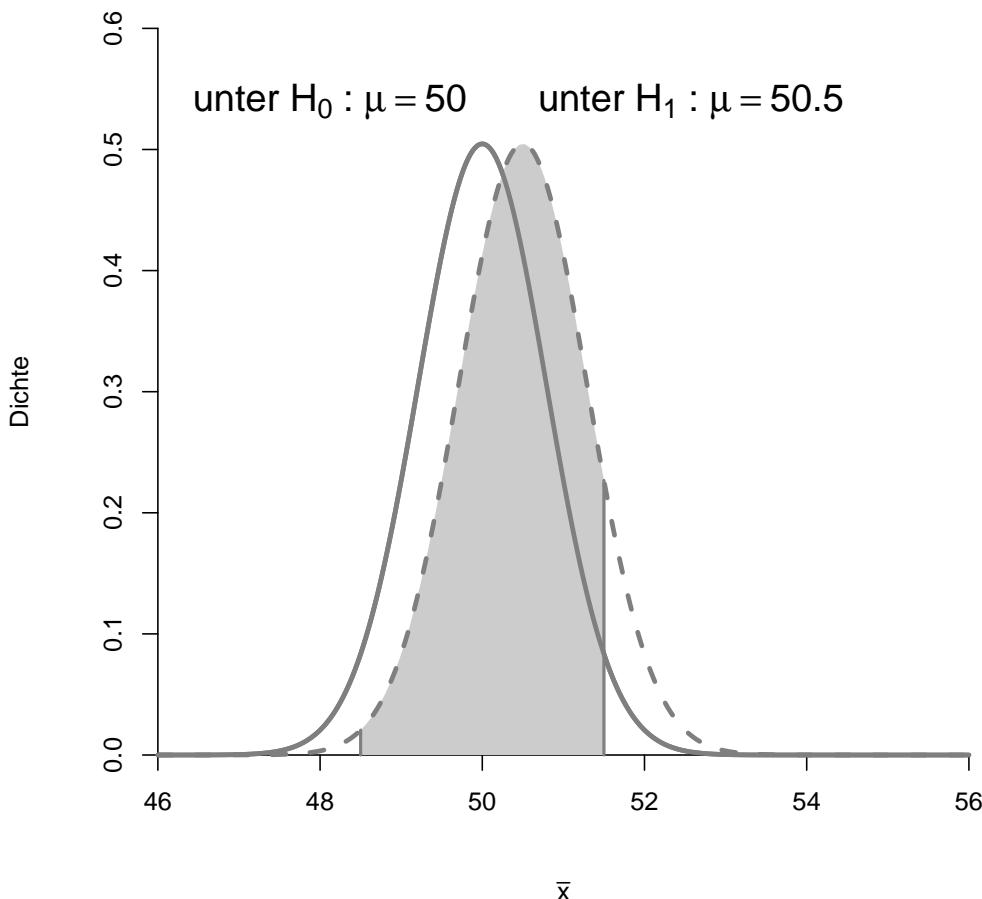
Fehlerwahrscheinlichkeit 2. Art? Um diese Frage beantworten zu können, müssen wir einen Wert aus der Alternativhypothese spezifizieren. Für beispielsweise $\mu = 52$ gilt:

$$\bar{X} \sim N\left(52, \frac{(2.5)^2}{10}\right)$$

Die Wahrscheinlichkeit für einen Typ II–Fehler ist daher gegeben durch:

$$\beta = \Phi\left(\frac{51.5 - 52}{2.5/\sqrt{10}}\right) - \Phi\left(\frac{48.5 - 52}{2.5/\sqrt{10}}\right) = 0.2643$$

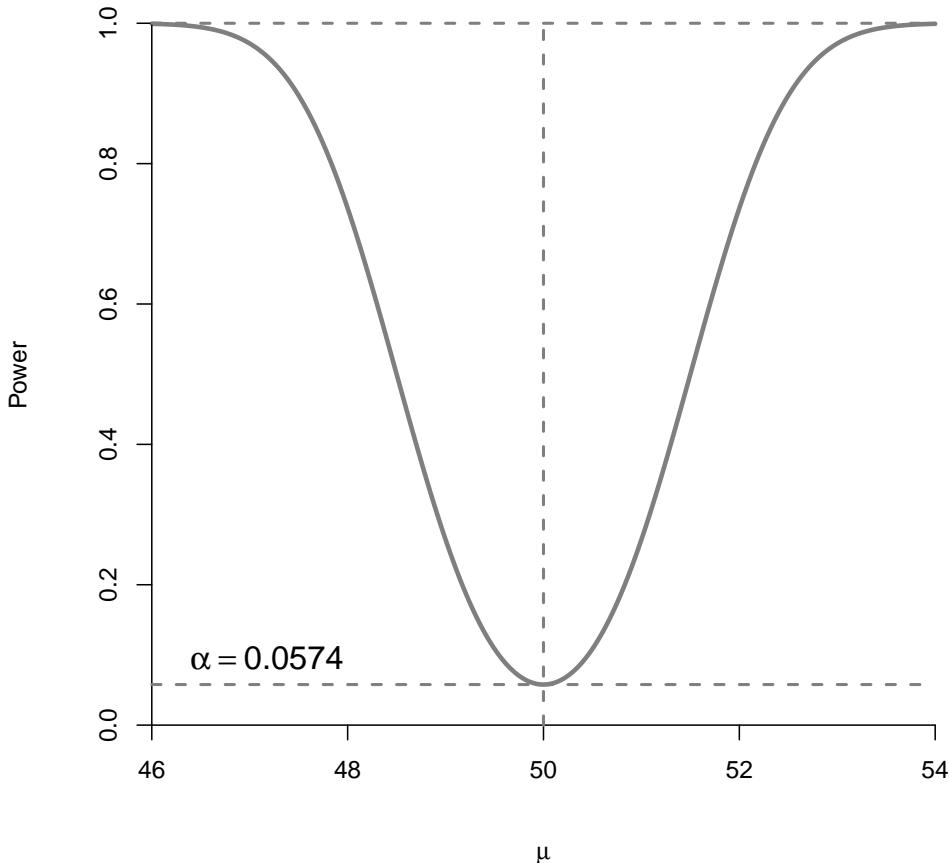
(Vgl. Abb 7.11 für eine grafische Veranschaulichung.) Aus Symmetriegründen ergibt sich der gleiche Wert für β , wenn $\mu = 48$. Liegt der wahre Wert von μ sehr nahe bei $\mu_0 = 50$, erhöht sich die Wahrscheinlichkeit für einen Typ II–Fehler drastisch. Für beispielsweise $\mu = 50.5$ ergibt sich $\beta = 0.8923$ (vgl. Abb 7.12). ■

Abbildung 7.12: Wahrscheinlichkeit eines Typ II-Fehlers für $\mu = 50.5$ 

Starke/Schwache Schlussfolgerungen: In der Praxis verwendet man Tests, die eine vorgegebene (kleine) Wahrscheinlichkeit für einen Typ I-Fehler nicht überschreiten. Wird \mathcal{H}_0 verworfen, spricht man daher von einer **starken** Schlussfolgerung. Wird \mathcal{H}_0 nicht verworfen, hat man möglicherweise einen Typ II-Fehler begangen, und über seine Größe weiß man meist nur wenig (β ist eine Funktion des wahren Parameterwerts, und der ist eben nicht bekannt). In letzterem Fall spricht man daher von einer **schwachen** Schlussfolgerung und sagt meist vorsichtiger, dass man \mathcal{H}_0 *nicht verwerfen kann* (und nicht, dass man \mathcal{H}_0 „akzeptiert“).

Schärfe: Die **Schärfe** (oder **Power**) eines Tests ist die Wahrscheinlichkeit der Verwerfung der Nullhypothese \mathcal{H}_0 , wenn die Alternativhypothese zutrifft (d. h., die richtige Entscheidung zu treffen, wenn \mathcal{H}_0 falsch ist). Betrachtet man die Schärfe als Funktion von θ , spricht man von der **Schärfefunktion** (oder **Powerfunktion**):

$$\gamma_C(\theta) = 1 - \beta(\theta) \quad \text{für } \theta \in \Theta_1$$

Abbildung 7.13: Powerfunktion für den Test von Bsp 7.19

Bem: Die Power eines Tests hängt eng mit seiner **Sensitivität** zusammen, d. h., mit seiner Fähigkeit, Abweichungen von der Nullhypothese \mathcal{H}_0 als solche zu erkennen. Ist die Power eines Tests zu gering, kann man entweder α oder – nach Möglichkeit – die Stichprobengröße n erhöhen.

Bsp 7.20 Die Powerfunktion für den Test von Bsp 7.19 ist gegeben durch:

$$\gamma_C(\mu) = 1 - \beta(\mu) = \Phi\left(\frac{51.5 - \mu}{2.5/\sqrt{10}}\right) - \Phi\left(\frac{48.5 - \mu}{2.5/\sqrt{10}}\right), \quad \mu \in \mathbb{R}$$

Für $\mu = \mu_0 = 50$ entspricht die Powerfunktion der Wahrscheinlichkeit eines Typ I–Fehlers, d. h. $\gamma_C(50) = \alpha = 0.0574$. (Vgl. Abb 7.13 für eine grafische Darstellung.) ■

7.4.2 p-Wert

Die allermeisten Statistikpakete (auch R) verfolgen beim Testen von Hypothesen nicht die im vorigen Abschnitt beschriebene „klassische“ Vorgangsweise, sondern berechnen statt

dessen einen Wahrscheinlichkeitswert. Der p -Wert¹⁷ (oder das **beobachtete Signifikanzniveau**) der \mathcal{H}_0 entspricht der Wahrscheinlichkeit – bei Zutreffen von \mathcal{H}_0 – den beobachteten Wert der Teststatistik oder einen extremeren zu bekommen. Was konkret unter „extremer“ zu verstehen ist, hängt von der Gegenhypothese (oder vom kritischen Bereich) ab.

Bsp 7.21 Angenommen, im Kontext von Bsp 7.19 ergibt sich ein Stichprobenmittelwert von $\bar{x} = 51.8$ (= Wert der Teststatistik). Beim vorliegenden zweiseitigen Testproblem bedeutet „extremer“, dass sich \bar{X} unter \mathcal{H}_0 um mehr als 1.8 von $\mu_0 = 50$ unterscheidet. D.h., der p -Wert ist wie folgt zu berechnen:

$$\begin{aligned} p\text{-Wert} &= P_{\mu_0}(|\bar{X} - \mu_0| \geq 1.8) \\ &= 1 - P_{\mu_0}(48.2 < \bar{X} < 51.8) \\ &= 1 - \left[\Phi\left(\frac{51.8 - 50}{2.5/\sqrt{10}}\right) - \Phi\left(\frac{48.2 - 50}{2.5/\sqrt{10}}\right) \right] \\ &= 0.0228 \end{aligned}$$

Nach dem unten angegebenen Beurteilungsschema bedeutet dieser Wert, dass für einen beobachteten Wert von $\bar{x} = 51.8$ starke Einwände gegen die Gültigkeit der $\mathcal{H}_0 : \mu = 50$ vorliegen. ■

Bezug zum klassischen Testen: Ein klassischer Test ergibt sich dadurch, dass eine \mathcal{H}_0 , deren p -Wert kleiner als α ist, auf dem Niveau α verworfen wird. Anders ausgedrückt:

Der p -Wert der \mathcal{H}_0 ist der *größte* Wert von α , für den die \mathcal{H}_0 *nicht* verworfen wird.

Die Beurteilung von Hypothesen mittels p -Wert hat u.a. den Vorteil, dass man auf Basis einer Zahl für alle Werte von α die Testentscheidung unmittelbar ablesen kann.

Interpretation des p -Werts: Bei der Interpretation des p -Werts hält man sich meist an das folgende Beurteilungsschema:

p -Wert	Signifikanz	
< 0.01	sehr hoch	(sehr starke Einwände gegen \mathcal{H}_0)
0.01 – 0.05	hoch	(starke Einwände gegen \mathcal{H}_0)
0.05 – 0.10	schwach	(schwache Einwände gegen \mathcal{H}_0)
> 0.10	keine	(sehr schwache/keine Einwände gegen \mathcal{H}_0)

¹⁷engl. *p-value*

Bemerkungen:

- (a) Die oben verwendete Sprechweise von der „Signifikanz“ eines Tests ist zwar weit verbreitet aber mit einer gewissen Vorsicht zu gebrauchen. Ein Test ist **signifikant**, wenn er die Nullhypothese verwirft. Das ist eine formale Aussage, die von den Hypothesen, vom verwendeten Test, von der Stichprobengröße und von α abhängt. Diese *statistische* Signifikanz sollte nicht mit der *praktischen* (oder *wissenschaftlichen*) Signifikanz verwechselt werden. Möglicherweise ist ein formal signifikantes Ergebnis nur von geringer praktischer Bedeutung.
- (b) Bei der Beurteilung des p -Werts nach dem obigen Schema ist eine gewisse Vorsicht angebracht. Ein „großer“ p -Wert (beispielsweise größer als 0.10) bedeutet *nicht* automatisch eine Unterstützung für \mathcal{H}_0 . Ein möglicher anderer Grund dafür könnte auch sein, dass die \mathcal{H}_0 falsch ist, aber der Test eine zu geringe Power hat, um das zu erkennen.
- (c) Man verwechsle den p -Wert einer Nullhypothese nicht mit $P(\mathcal{H}_0|\text{Daten})$. Derartige Aussagen sind nur im Rahmen der *Bayes'schen Statistik* (vgl. Kapitel 8) möglich und sinnvoll. Der p -Wert ist *nicht* die Wahrscheinlichkeit für die Gültigkeit der \mathcal{H}_0 !

7.4.3 Beziehung zwischen Tests und Konfidenzintervallen

Es gibt eine enge Beziehung zwischen Parametertests und Konfidenzintervallen. Angenommen, $(T_1(\mathbf{x}), T_2(\mathbf{x}))$ ist ein $(1 - \alpha)$ -Konfidenzintervall für einen Parameter $\theta \in \Theta$ auf Basis einer (konkreten) Stichprobe $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ von X . Dann ist ein Test zum Niveau α für die Hypothesen:

$$\mathcal{H}_0 : \theta = \theta_0 \quad \text{gegen} \quad \mathcal{H}_1 : \theta \neq \theta_0$$

gegeben durch:

$$\theta_0 \in (T_1(\mathbf{x}), T_2(\mathbf{x})) \quad \longrightarrow \quad \mathcal{H}_0 \text{ nicht verwerfen}$$

$$\theta_0 \notin (T_1(\mathbf{x}), T_2(\mathbf{x})) \quad \longrightarrow \quad \mathcal{H}_0 \text{ verwerfen}$$

Bsp 7.22 Ebenso wie in **Bsp 7.21** nehmen wir an, dass sich im Kontext von **Bsp 7.19** ein Stichprobenmittelwert von $\bar{x} = 51.8$ ergibt. Ein 95%-Konfidenzintervall für μ ist dann gegeben durch:

$$\bar{x} \pm z_{0.975} \frac{2.5}{\sqrt{10}} = (50.251, 53.349)$$

Da $\mu_0 = 50$ kein Element dieses Intervalls ist, wird $\mathcal{H}_0 : \mu = 50$ zum Niveau 5% verworfen. (Man beachte auch, dass der in **Bsp 7.21** zu $\bar{x} = 51.8$ berechnete p -Wert der \mathcal{H}_0 kleiner als $\alpha = 0.05$ ist.) ■

Bem: Auch wenn Parametertests und Konfidenzintervalle äquivalente Konzepte darstellen, so vermitteln sie dennoch unterschiedliche Einsichten. Durch ein Konfidenzintervall bekommt man – zu einem bestimmten Konfidenzlevel – einen Bereich von „plausiblen“ Werten für den in Frage stehenden Parameter. Auf Basis von Tests andererseits gewinnt man – beispielsweise durch Berechnung von p -Werten – Einsichten hinsichtlich der mit bestimmten Entscheidungen verbundenen *Risiken*.

7.4.4 Tests für den Mittelwert einer Normalverteilung (Varianz bekannt)

Gegeben sei eine Stichprobe X_1, X_2, \dots, X_n von $X \sim N(\mu, \sigma_0^2)$, wobei die Schreibweise σ_0^2 für die Varianz andeuten soll, dass sie als bekannt vorausgesetzt wird. Wie schon früher diskutiert, ist der Stichprobenmittelwert \bar{X}_n allgemein ein unverzerrter Schätzer für μ mit Varianz σ_0^2/n . Für Stichproben aus einer Normalverteilung ist die **Stichprobenverteilung**¹⁸ von \bar{X}_n gegeben durch:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma_0^2}{n}\right)$$

Für die Entwicklung von Tests für $\mathcal{H}_0 : \mu = \mu_0$ (gegen ein- oder zweiseitige Alternativen) ist es vorteilhaft, \bar{X}_n zu *standardisieren* und den kritischen Bereich durch die folgende **Teststatistik** zu definieren:

$$Z_0 = \frac{\bar{X}_n - \mu_0}{\sigma_0/\sqrt{n}}$$

Unter $\mathcal{H}_0 : \mu = \mu_0$ ist Z_0 standardnormalverteilt, $Z_0 \sim N(0, 1)$, und exakte Tests für den Mittelwert μ zum Niveau α sind gegeben wie folgt:

Nullhypothese: $\mathcal{H}_0 : \mu = \mu_0$

Teststatistik: $Z_0 = \frac{\bar{X}_n - \mu_0}{\sigma_0/\sqrt{n}}$

Alternativhypothese \mathcal{H}_1	\mathcal{H}_0 verwerfen, falls
$\mu \neq \mu_0$	$ Z_0 > z_{1-\alpha/2}$
$\mu > \mu_0$	$Z_0 > z_{1-\alpha}$
$\mu < \mu_0$	$Z_0 < z_\alpha (= -z_{1-\alpha})$

¹⁸engl. *sampling distribution*

p-Wert: Ist z_0 der beobachtete Wert der Teststatistik Z_0 , so ist der p-Wert der \mathcal{H}_0 – abhängig von der Alternativhypothese – wie folgt zu berechnen:

Alternativhypothese \mathcal{H}_1	p-Wert
$\mu \neq \mu_0$	$2[1 - \Phi(z_0)]$
$\mu > \mu_0$	$1 - \Phi(z_0)$
$\mu < \mu_0$	$\Phi(z_0)$

Bem: Tests werden häufig nach den vorkommenden **Schwellenwerten** (= Quantile der Teststatistik unter \mathcal{H}_0) benannt. Die Tests dieses Abschnitts werden dementsprechend meist als z -Tests bezeichnet.

Powerfunktion: Bei bekannter Varianz σ_0^2 ist es nicht schwierig, explizite Ausdrücke für die Testpower zu finden. Instruktiver und nützlicher für Anwendungen sind aber grafische Darstellungen der Powerfunktion für ein (grobes) Raster von Stichprobengrößen.

Abb 7.14 zeigt einige Powerfunktionen für den z -Test für zweiseitige Alternativen in Abhängigkeit von $\delta = |\mu - \mu_0|/\sigma_0$ (d. h. in standardisierten Abweichungen von μ_0). Möchte man beispielsweise die Power des z -Tests von $\mathcal{H}_0 : \mu = 50$ (gegen $\mathcal{H}_1 : \mu \neq 50$) an der Stelle $\mu = 51$ bestimmen, wenn $n = 25$, $\sigma_0 = 2$ und $\alpha = 5\%$, so findet man für $\delta = |51 - 50|/2 = 1/2$ einen Wert von 70%. D.h., in etwa 30% der Fälle wird der Test eine Abweichung von der Größe einer halben Standardabweichung nicht entdecken.

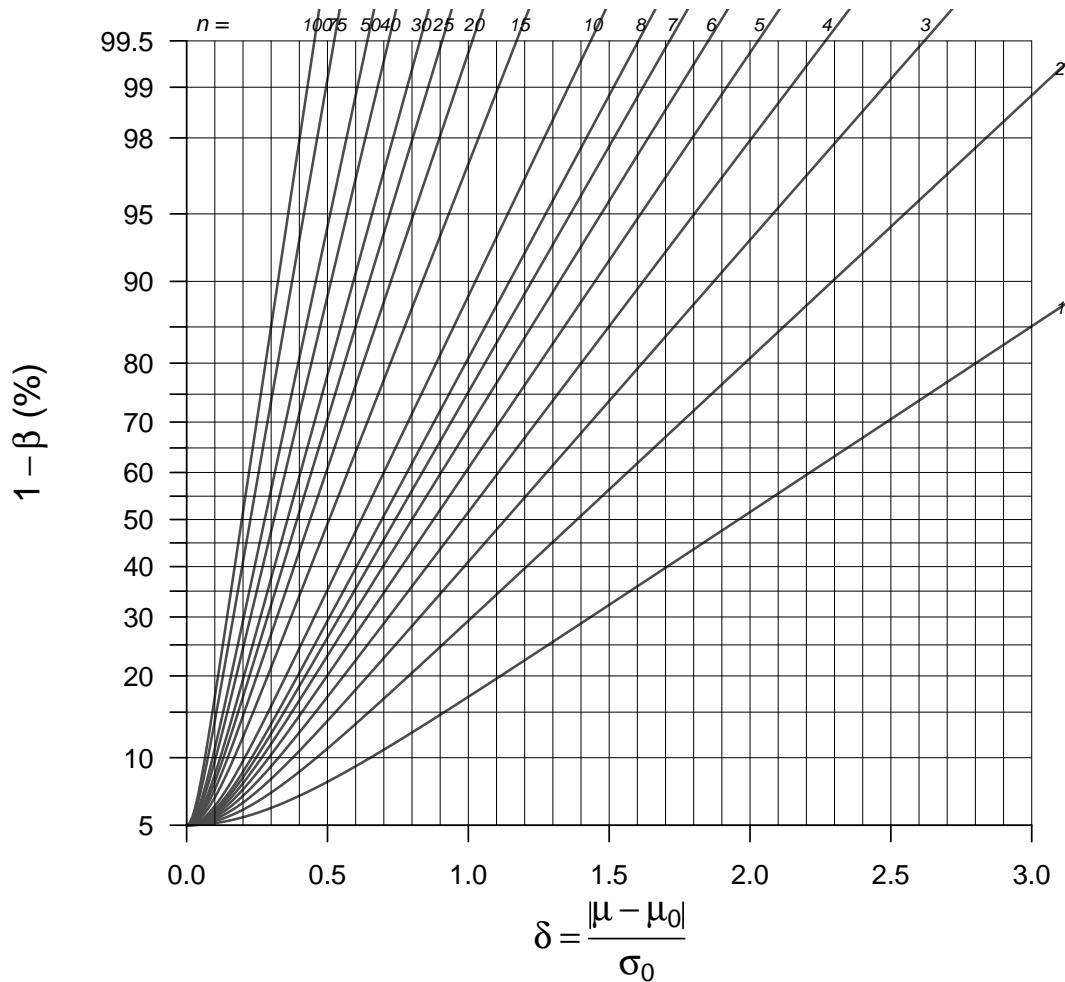
Umgekehrt lässt sich aus Diagrammen dieser Art auch abschätzen, wie groß die Stichprobe sein müsste, um eine bestimmte Power zu erzielen. Wenn man beispielsweise für eine standardisierte Abweichung von $\delta = 1/2$ eine hohe Power von (mindestens) 90% haben möchte, so findet man aus Abb 7.14 eine Stichprobengröße von etwa $n = 40$.

Tests für große Stichproben: Ist die Stichprobe nicht zu klein (etwa $n > 40$), können die Tests dieses Abschnitts in guter Näherung auch dann verwendet werden, wenn σ^2 nicht bekannt ist (und durch s_n^2 ersetzt wird), ungeachtet der tatsächlichen Form der zugrunde liegenden Verteilung. Dabei beruft man sich auf den ZGVS und auf die Konsistenz von S_n^2 zur Schätzung von σ^2 .

7.4.5 Tests für den Mittelwert einer Normalverteilung (Varianz unbekannt)

Gegeben sei eine Stichprobe X_1, X_2, \dots, X_n von $X \sim N(\mu, \sigma^2)$, wobei wir nun davon ausgehen, dass auch σ^2 nicht bekannt ist und durch die Stichprobenvarianz S_n^2 erwartungstreu und konsistent geschätzt werden kann. Nach der Behauptung (4) in 7.3.3 gilt in diesem Fall, dass:

$$T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \sim t(n-1)$$

Abbildung 7.14: Powerfunktionen für den z -Test bei zweiseitigen Alternativen ($\alpha = 5\%$)

Exakte t-Tests für den Mittelwert μ zum Niveau α sind dann gegeben wie folgt:

Nullhypothese: $\mathcal{H}_0 : \mu = \mu_0$

Teststatistik: $T_0 = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}$

Alternativhypothese \mathcal{H}_1	\mathcal{H}_0 verwerfen, falls
$\mu \neq \mu_0$	$ T_0 > t_{n-1; 1-\alpha/2}$
$\mu > \mu_0$	$T_0 > t_{n-1; 1-\alpha}$
$\mu < \mu_0$	$T_0 < t_{n-1; \alpha} (= -t_{n-1; 1-\alpha})$

p-Wert: Ist t_0 der beobachtete Wert der Teststatistik T_0 , so ist der p-Wert der \mathcal{H}_0 – abhängig von der Alternativhypothese – wie folgt zu berechnen:

Alternativhypothese \mathcal{H}_1	p-Wert
$\mu \neq \mu_0$	$2[1 - F(t_0)]$
$\mu > \mu_0$	$1 - F(t_0)$
$\mu < \mu_0$	$F(t_0)$

Dabei bezeichnet F die Verteilungsfunktion einer $t(n - 1)$ -Verteilung.

Bsp 7.23 Zehn Beobachtungen aus einer Normalverteilung seien gegeben wie folgt:

52.1 49.0 51.4 50.0 50.3 49.6 50.6 50.8 51.0 51.7

Sind die Beobachtungen zum Niveau $\alpha = 5\%$ kompatibel mit der Behauptung $\mu = 50$? Der folgende R-Output zeigt das Ergebnis des t-Tests gegen die zweiseitige Alternative $\mu \neq 50$.

```
x <- c(52.1,49.0,51.4,50.0,50.3,49.6,50.6,50.8,51.0,51.7)
t.test(x, mu=50)

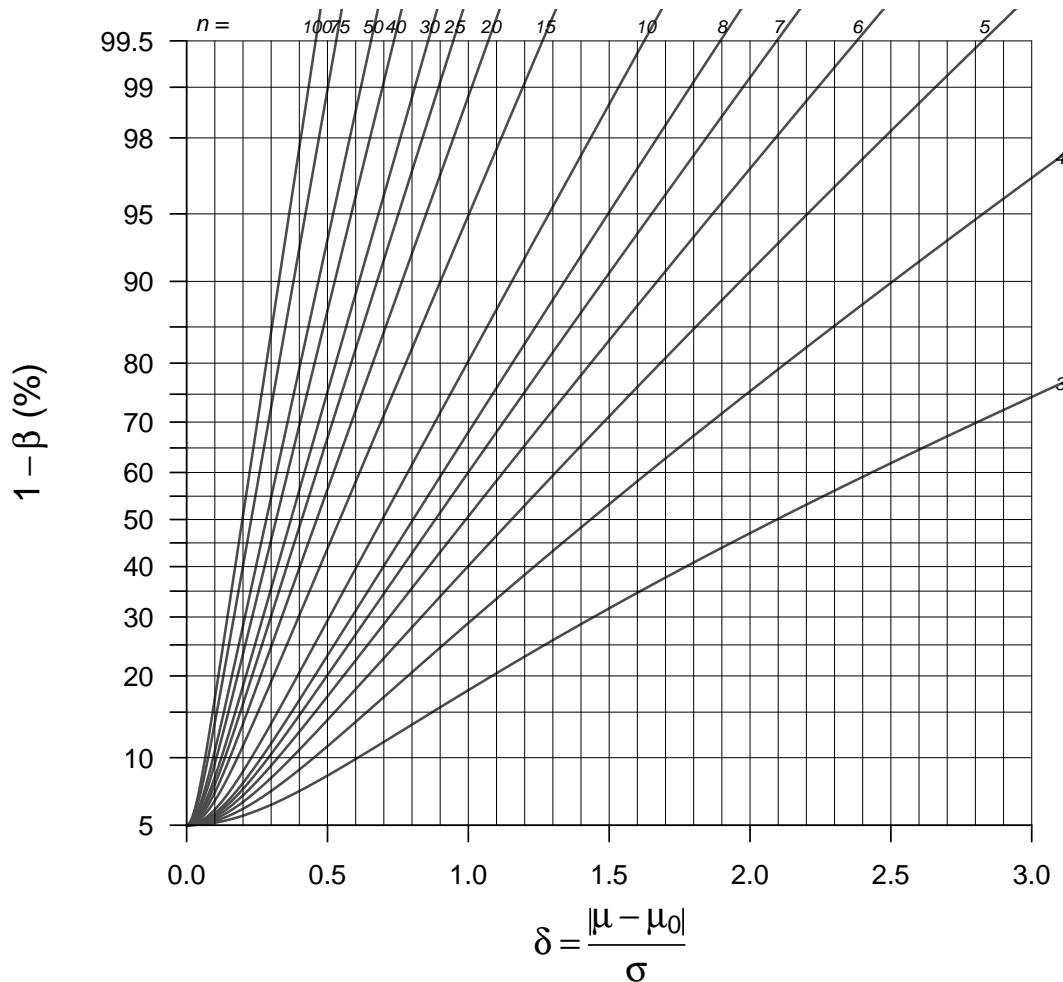
One Sample t-test

data: x
t = 2.1423, df = 9, p-value = 0.06079
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
49.964 51.336
sample estimates:
mean of x
50.65
```

Da der p-Wert größer als 0.05 ist, wird die Nullhypothese auf dem Niveau 5% nicht verworfen. Äquivalent dazu kann man auch das 95%-Konfidenzintervall für μ heranziehen. Da $\mu = 50$ Element des Intervalls ist, wird $\mathcal{H}_0 : \mu = 50$ nicht verworfen. ■

Powerfunktion: Bei unbekannter Varianz σ^2 ist die Berechnung der Testpower schwieriger als bei bekannter Varianz.¹⁹ Instruktiver und nützlicher für Anwendungen sind aber

¹⁹Dabei kommt die sog. *nichtzentrale t*-Verteilung ins Spiel.

Abbildung 7.15: Powerfunktionen für den t–Test bei zweiseitigen Alternativen ($\alpha = 5\%$)

grafische Darstellungen der Powerfunktion für ein (grobes) Raster von Stichprobengrößen (Abb 7.15). Auch in diesem Fall ist es sinnvoll, die Powerfunktion in Abhängigkeit von der standardisierten Abweichung $\delta = |\mu - \mu_0|/\sigma$ darzustellen.

Möchte man beispielsweise die Power des (zweiseitigen) t–Tests für $\delta = 1$, $n = 10$ und $\alpha = 5\%$ bestimmen, so findet man aus dem Diagramm einen Wert von 80%. Für dieselben Vorgaben findet man für den z –Test eine Power von etwa 88% (Abb 7.14). Die höhere Power verdankt sich dem Umstand, dass die Varianz beim z –Test als bekannt vorausgesetzt wird und nicht aus den Beobachtungen geschätzt werden muss.

7.4.6 Tests für die Varianz einer Normalverteilung

Für die Entwicklung von Tests für die Varianz einer Normalverteilung beziehen wir uns auf Behauptung (2) von 7.3.3. Ist X_1, X_2, \dots, X_n eine Stichprobe von $X \sim N(\mu, \sigma^2)$ und ist S_n^2 die Stichprobenvarianz, so gilt:

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

Exakte Tests für die Varianz σ^2 zum Niveau α sind dann gegeben wie folgt:

Nullhypothese: $\mathcal{H}_0 : \sigma^2 = \sigma_0^2$

Teststatistik: $\chi_0^2 = \frac{(n-1)S_n^2}{\sigma_0^2}$

Alternativhypothese \mathcal{H}_1	\mathcal{H}_0 verwerfen, falls
$\sigma^2 \neq \sigma_0^2$	$\chi_0^2 < \chi_{n-1; \alpha/2}^2$ oder $\chi_0^2 > \chi_{n-1; 1-\alpha/2}^2$
$\sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi_{n-1; 1-\alpha}^2$
$\sigma^2 < \sigma_0^2$	$\chi_0^2 < \chi_{n-1; \alpha}^2$

Bsp 7.24 Angenommen, wir testen $\mathcal{H}_0 : \sigma^2 = 5$ gegen $\mathcal{H}_1 : \sigma^2 < 5$ und der Wert der Teststatistik für eine Stichprobengröße von $n = 15$ beträgt $\chi_0^2 = 4.2$. Wie groß ist der p -Wert? Der p -Wert ist der größte Wert von α , für den die \mathcal{H}_0 nicht verworfen wird:

$$p\text{-Wert} = P(\chi^2(14) \leq 4.2) \doteq 0.0059$$

Testet man zweiseitig (d. h. gegen $\mathcal{H}_1 : \sigma^2 \neq 5$), ist der p -Wert wie folgt zu berechnen:

$$p\text{-Wert} = 2 \min \left\{ \underbrace{P(\chi^2(14) \leq 4.2)}_{= 0.0059}, \underbrace{P(\chi^2(14) \geq 4.2)}_{= 1 - 0.0059} \right\} \doteq 0.0117$$

In beiden Fällen ist zum (üblichen) Niveau von $\alpha = 5\%$ die Nullhypothese $\mathcal{H}_0 : \sigma^2 = 5$ zugunsten der Alternativhypothese zu verwirfen. Zum „vorsichtigeren“ Niveau $\alpha = 1\%$ allerdings nur im ersten Fall. ■

7.4.7 Tests für einen Anteil

Ist X_1, X_2, \dots, X_n eine Stichprobe von $X \sim A(p)$ (Bernoulli-Verteilung), so hat $Y = \sum_{i=1}^n X_i$ (= Anzahl der Einser in der Stichprobe) eine Binomialverteilung $B(n, p)$. Betrachten wir zunächst einen Test von:

$$(1) \quad \mathcal{H}_0 : p = p_0 \quad \text{gegen} \quad \mathcal{H}_1 : p > p_0$$

Intuitiv wird man \mathcal{H}_0 verwerfen, wenn Y einen bestimmten Schwellenwert überschreitet, d. h., wenn $Y \geq k$. Unter \mathcal{H}_0 gilt:

$$P(Y \geq k) = \sum_{i=k}^n \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

Für einen Test zum (vorgegebenen) Niveau α müsste man k so wählen, dass die Summe auf der rechten Seite gleich α ist. Da es aber aufgrund der Diskrettheit der Binomialverteilung ein derartiges k in der Regel nicht gibt, wählt man den Schwellenwert k^* wie folgt:

$$k^* = \min \left\{ k \mid \sum_{i=k}^n \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha \right\}$$

Als Alternative zur obigen Vorgangsweise kann man auch den p -Wert der \mathcal{H}_0 bestimmen. Ist y der beobachtete Wert von Y , so gilt:

$$p\text{-Wert} = \sum_{i=y}^n \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

Bsp 7.25 Ein Hersteller von Computerchips behauptet, dass nicht mehr als 2% seiner Chips defekt sind. Ein Abnehmer testet 300 Chips und findet darunter 10 defekte Chips. Lässt sich damit die Behauptung des Herstellers widerlegen? Der Abnehmer testet die folgenden Hypothesen:

$$\mathcal{H}_0 : p = p_0 = 0.02 \quad \text{gegen} \quad \mathcal{H}_1 : p > p_0$$

Eine einfache Suchprozedur ergibt für $\alpha = 0.05$ einen Schwellenwert von $k^* = 11$. Die Nullhypothese kann also zum Niveau 5% nicht verworfen werden. Man kann auch den p -Wert berechnen:

$$p\text{-Wert} = P_{p_0}(Y \geq 10) = 1 - \sum_{i=0}^9 \binom{300}{i} (0.02)^i (0.98)^{300-i} = 0.0818$$

Auch am p -Wert zeigt sich, dass zum Niveau 5% die Nullhypothese nicht verworfen werden kann, wohl aber zum weniger vorsichtigen Niveau 10%. ■

Analoge Überlegungen gelten für einen Test von:

$$(2) \quad \mathcal{H}_0 : p = p_0 \quad \text{gegen} \quad \mathcal{H}_1 : p < p_0$$

Nun ist \mathcal{H}_0 zu verwerfen, wenn $Y \leq k^*$, wobei k^* wie folgt gewählt wird:

$$k^* = \max \left\{ k \mid \sum_{i=0}^k \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha \right\}$$

Ist y der beobachtete Wert von Y , so ist der p -Wert von \mathcal{H}_0 gegeben durch:

$$p\text{-Wert} = \sum_{i=0}^y \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

Im Falle einer zweiseitigen Alternative:

$$(3) \quad \mathcal{H}_0 : p = p_0 \quad \text{gegen} \quad \mathcal{H}_1 : p \neq p_0$$

wird man \mathcal{H}_0 verwerfen, wenn der beobachtete Wert y von $Y = \sum_{i=1}^n X_i$ entweder deutlich größer oder kleiner als der Wert ist, den man für $p = p_0$ erwarten würde, d. h., wenn:

$$P_{p_0}(Y \geq y) \leq \frac{\alpha}{2} \quad \text{oder} \quad P_{p_0}(Y \leq y) \leq \frac{\alpha}{2}$$

D. h., der p -Wert für $Y = y$ ist gegeben durch:

$$p\text{-Wert} = 2 \min \left\{ P_{p_0}(Y \geq y), P_{p_0}(Y \leq y) \right\}$$

Bsp 7.26 Angenommen, die Ausschussquote eines Prozesses liegt schon seit längerer Zeit bei 4%. Nach einer Umstellung der Arbeitsabläufe möchte man herausfinden, ob sich die Ausschussquote verändert hat. In einer Stichprobe der Größe $n = 500$ gibt es 16 defekte Teile (entspricht einer Ausschussquote von 3.2%). Da die ursprüngliche Fragestellung auf eine *Veränderung* der Ausschussquote abzielt (und nicht auf eine Verbesserung), sind die folgenden Hypothesen zu testen:

$$\mathcal{H}_0 : p = p_0 = 0.04 \quad \text{gegen} \quad \mathcal{H}_1 : p \neq p_0$$

Für $p = 0.04$ gilt $P(Y \leq 16) = 0.2158$ und $P(Y \geq 16) = 0.8487$; der p -Wert der \mathcal{H}_0 ist also gegeben durch:

$$p\text{-Wert} = 2P(Y \leq 16) = 0.4316$$

Der vergleichsweise große p -Wert zeigt, dass der Stichprobenbefund nicht ausreicht, um auf eine Veränderung der Ausschussquote schließen zu können.

Der folgende R-Output zeigt die Verwendung der Funktion `binom.test()`:

```
binom.test(16, 500, p=0.04)

Exact binomial test

data: 16 and 500
number of successes = 16, number of trials = 500, p-value = 0.4242
alternative hypothesis: true probability of success is not equal to 0.04
95 percent confidence interval:
 0.018399 0.051447
sample estimates:
probability of success
                0.032
```

Der kleine Unterschied zum zuerst berechneten p -Wert erklärt sich aus der – im zweiseitigen Fall – etwas anderen Berechnung, und zwar als Summe aller Binomialwahrscheinlichkeiten, deren Wert kleiner oder gleich $P_{p_0}(Y = 16)$ ist. Praktisch ist die unterschiedliche Berechnung des p -Werts aber nur von geringer Bedeutung. ■

Approximative Tests für große Stichproben: Für große Stichproben gilt unter Verwendung des ZGVS (vgl. 6.3.3):

$$Y = \sum_{i=1}^n X_i \approx N(np, np(1-p))$$

Approximative Tests zum Niveau α sind also gegeben durch:

Nullhypothese: $\mathcal{H}_0 : p = p_0$

Teststatistik: $Z_0 = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}}$

Alternativhypothese \mathcal{H}_1	\mathcal{H}_0 verwerfen, falls
$p \neq p_0$	$ Z_0 > z_{1-\alpha/2}$
$p > p_0$	$Z_0 > z_{1-\alpha}$
$p < p_0$	$Z_0 < z_\alpha (= -z_{1-\alpha})$

Bsp 7.27 Wir betrachten noch einmal die Situation von Bsp 7.26. Die Stichprobengröße von $n = 500$ ist ausreichend groß, sodass die Normalapproximation der Binomialverteilung auch für $p = 0.04$ zulässig ist (Faustregel: $(500)(0.04)(0.96) = 19.2 \geq 10$). Für den Wert der Teststatistik ergibt sich:

$$z_0 = \frac{16 - (500)(0.04)}{\sqrt{(500)(0.04)(0.96)}} = -0.9129$$

Der (approximative) p -Wert ist also gegeben durch:

$$p\text{-Wert} = 2[1 - \Phi(|z_0|)] = 0.3613$$

Auch dieser Wert zeigt, dass nicht auf eine Veränderung der Ausschussquote geschlossen werden kann. ■

7.4.8 Tests für die Mittelwerte von zwei Normalverteilungen

Für Stichproben X_1, X_2, \dots, X_m und Y_1, Y_2, \dots, Y_n von zwei ua. sGn $X \sim N(\mu_X, \sigma_X^2)$ bzw. $Y \sim N(\mu_Y, \sigma_Y^2)$, betrachten wir nun Tests für die Differenz $\mu_X - \mu_Y$ der Mittelwerte. Dabei nehmen wir zunächst an, dass die beiden Varianzen σ_X^2 und σ_Y^2 unbekannt aber **gleich** sind, d. h. $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (unbekannt). Ebenso wie für die Konstruktion von Konfidenzintervallen für $\mu_X - \mu_Y$ (vgl. 7.3.4) ist es in diesem Fall sinnvoll, die Varianzschätzungen S_X^2 und S_Y^2 zu kombinieren. Der **gepoolte Varianzschätzer** von σ^2 ist gegeben durch:

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

Mit diesem Varianzschätzer gilt:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{1/m + 1/n}} \sim t(m+n-2)$$

Exakte (**gepoolte**) **t-Tests** für die Differenz $\mu_X - \mu_Y$ der Mittelwerte zum Niveau α sind dann gegeben wie folgt:

Nullhypothese: $\mathcal{H}_0 : \mu_X - \mu_Y = \Delta_0$

Teststatistik: $T_0 = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{S_p \sqrt{1/m + 1/n}}$

Alternativhypothese \mathcal{H}_1	\mathcal{H}_0 verwerfen, falls
$\mu_X - \mu_Y \neq \Delta_0$	$ T_0 > t_{m+n-2; 1-\alpha/2}$
$\mu_X - \mu_Y > \Delta_0$	$T_0 > t_{m+n-2; 1-\alpha}$
$\mu_X - \mu_Y < \Delta_0$	$T_0 < t_{m+n-2; \alpha} (= -t_{m+n-2; 1-\alpha})$

In vielen Fällen kann man nicht davon ausgehen, dass die beiden Varianzen σ_X^2 und σ_Y^2 gleich sind. Im Falle $\sigma_X^2 \neq \sigma_Y^2$ (beide unbekannt) gibt es keinen exakten Test für die Differenz der Mittelwerte. Sind *beide* Stichprobengrößen m und n nicht zu klein, kann man den folgenden approximativen Test verwenden:

Nullhypothese: $\mathcal{H}_0 : \mu_X - \mu_Y = \Delta_0$

Teststatistik: $Z_0 = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{S_X^2/m + S_Y^2/n}}$

Alternativhypothese \mathcal{H}_1	\mathcal{H}_0 verwerfen, falls
$\mu_X - \mu_Y \neq \Delta_0$	$ Z_0 > z_{1-\alpha/2}$
$\mu_X - \mu_Y > \Delta_0$	$Z_0 > z_{1-\alpha}$
$\mu_X - \mu_Y < \Delta_0$	$Z_0 < z_\alpha (= -z_{1-\alpha})$

Bsp 7.28 Zwei Stichproben aus unabhängigen Normalverteilungen seien gegeben wie folgt:

Stichprobe 1	3, 7, 25, 10, 15, 6, 12, 25, 15, 7
Stichprobe 2	48, 44, 40, 38, 33, 21, 20, 12, 1, 18

Die Stichprobenstreuung der ersten Stichprobe ist $s_1 = 7.63$, der zweiten $s_2 = 15.3$. Die Annahme, dass die beiden Streuungen gleich sind (d.h., $\sigma_1 = \sigma_2$) erscheint hier wenig plausibel und das Poolen der beiden Stichprobenvarianzen daher wenig sinnvoll. Der folgende R-Output zeigt die Verwendung der Funktion `t.test()` für den Test von:

$$\mathcal{H}_0 : \mu_1 = \mu_2 \quad \text{gegen} \quad \mathcal{H}_1 : \mu_1 \neq \mu_2$$

```

x <- c(3,7,25,10,15,6,12,25,15,7)
y <- c(48,44,40,38,33,21,20,12,1,18)
t.test(x, y, var.equal=TRUE)

Two Sample t-test

data: x and y
t = -2.7669, df = 18, p-value = 0.0127
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-26.3894 -3.6106
sample estimates:
mean of x mean of y
12.5      27.5

t.test(x, y, var.equal=FALSE)

Welch Two Sample t-test

data: x and y
t = -2.7669, df = 13.196, p-value = 0.01583
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-26.6941 -3.3059
sample estimates:
mean of x mean of y
12.5      27.5

```

Der erste **t**-Test wird unter der (hier fragwürdigen) Voraussetzung $\sigma_1^2 = \sigma_2^2$ durchgeführt, der zweite ohne diese Voraussetzung. Die Ergebnisse unterscheiden sich allerdings nur wenig, die p -Werte sind ähnlich, und in beiden Fällen wird die Gleichheit der Mittelwerte zum Niveau 5% verworfen (nicht aber zum vorsichtigeren Niveau 1%). Man beachte auch, dass für `var.equal=FALSE` ein anderer als der oben angegebene Näherungstest verwendet wird, nämlich der sog. **Welch-Test**. Letzterer ist im Fall ungleicher Varianzen der bevorzugte Näherungstest.²⁰ ■

7.4.9 Tests für die Varianzen von zwei Normalverteilungen

Für Stichproben X_1, X_2, \dots, X_m und Y_1, Y_2, \dots, Y_n von zwei ua. sGn $X \sim N(\mu_X, \sigma_X^2)$ bzw. $Y \sim N(\mu_Y, \sigma_Y^2)$, betrachten wir nun Tests für die Varianzen von X und Y . Dabei greifen wir auf die in 7.3.4 angegebene Behauptung zurück, dass unter den gegebenen Bedingungen:

²⁰Vgl. z. B. WIKIPEDIA für Details zum Welch-Test.

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1, n-1)$$

Exakte **F–Tests** für den Quotienten σ_X^2/σ_Y^2 der Varianzen zum Niveau α sind dann gegeben wie folgt:

Nullhypothese: $\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2$

Teststatistik: $F_0 = \frac{S_X^2}{S_Y^2}$

Alternativhypothese \mathcal{H}_1	\mathcal{H}_0 verwerfen, falls
$\sigma_X^2 \neq \sigma_Y^2$	$F_0 < F_{m-1, n-1; \alpha/2}$ oder $F_0 > F_{m-1, n-1; 1-\alpha/2}$
$\sigma_X^2 > \sigma_Y^2$	$F_0 > F_{m-1, n-1; 1-\alpha}$
$\sigma_X^2 < \sigma_Y^2$	$F_0 < F_{m-1, n-1; \alpha}$

Bsp 7.29 Für die beiden Stichproben von **Bsp 7.28** sind die Stichprobenstreuungen gegeben durch $s_1 = 7.63$ bzw. $s_2 = 15.3$. Da s_2 etwa doppelt so groß wie s_1 ist, erscheint die Annahme $\sigma_1 = \sigma_2$ nur wenig plausibel zu sein. Der folgende R–Output zeigt die Verwendung der Funktion `var.test()` für einen formalen Test von:

$$\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2 \quad \text{gegen} \quad \mathcal{H}_1 : \sigma_1^2 \neq \sigma_2^2$$

```
var.test(x, y)

  F test to compare two variances

data: x and y
F = 0.2473, num df = 9, denom df = 9, p-value = 0.04936
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.061438 0.995819
sample estimates:
ratio of variances
0.24735
```

Wie der p -Wert zeigt, wird – etwas überraschend – die \mathcal{H}_0 zum Niveau 5% nur ganz knapp verworfen. Man beachte allerdings, dass aufgrund der nur kleinen Stichproben der F-Test hier nicht sehr scharf ist. Testet man einseitig gegen $\mathcal{H}_1 : \sigma_1^2 < \sigma_2^2$ ist der p -Wert nur halb so groß (0.02468), die Verwerfung der \mathcal{H}_0 also deutlicher. (Man beachte allerdings, dass es i. A. *nicht* korrekt ist, die zu testenden Hypothesen erst *nach* Ansicht der Stichprobenwerte zu formulieren!)

Bem: Da es keine Rolle spielt, welche Stichprobe die „erste“ und welche die „zweite“ ist, wählt man – bei händischer Rechnung – beim zweiseitigen Test die Reihenfolge zweckmäßigerweise so, dass F_0 größer als 1 ist. In diesem Fall genügt der Vergleich mit dem oberen Schwellenwert (der untere Schwellenwert ist kleiner als 1 und muss nicht überprüft werden). Im vorliegenden Fall wäre es also zweckmäßiger, die Reihenfolge zu vertauschen:

$$F_0 = \frac{s_2^2}{s_1^2} = 4.0429 > F_{9,9; 0.975} = 4.026 \quad \longrightarrow \quad H_0 \text{ verwerfen } (\alpha = 5\%)$$

(Klarerweise sind bei Vertauschung von Zähler und Nenner von F_0 auch die Freiheitsgrade der F-Verteilung zu vertauschen!) ■

7.4.10 Tests für den Korrelationskoeffizienten

Der Korrelationskoeffizient ρ einer bivariaten Normalverteilung (vgl. Bsp 5.21):

$$(X, Y) \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$$

kann mittels des (**Stichproben-**) **Korrelationskoeffizienten** R geschätzt werden (vgl. auch 1.9.3). Ist $(X_1, Y_1), \dots, (X_n, Y_n)$ (mit $n > 2$) eine Stichprobe von (X, Y) , so ist R definiert wie folgt:

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}$$

Bem: R wird allgemein zur Schätzung der Korrelation von zwei gemeinsam stetig verteilten SGn X und Y verwendet. Im Fall einer bivariaten Normalverteilung kann man aber zeigen, dass R der ML-Schätzer von ρ ist.

Die Verteilung von R hängt vom tatsächlichen Wert von ρ ab und hat eine komplizierte Form. Für $\rho = 0$ (d. h., wenn X und Y unabhängig sind) gilt:

Behauptung: Ist $(X_1, Y_1), \dots, (X_n, Y_n)$ ($n > 2$) eine Stichprobe aus einer bivariaten Normalverteilung, so gilt für $\rho = 0$:

$$T = \frac{R \sqrt{n-2}}{\sqrt{1-R^2}} \sim t(n-2)$$

Exakte **Unabhängigkeitstests** zum Niveau α sind dann gegeben wie folgt:

Nullhypothese: $\mathcal{H}_0 : \rho = 0$

Teststatistik: $T = \frac{R \sqrt{n-2}}{\sqrt{1-R^2}}$

Alternativhypothese \mathcal{H}_1	\mathcal{H}_0 verwerfen, falls
$\rho \neq 0$	$ T > t_{n-2; 1-\alpha/2}$
$\rho > 0$	$T > t_{n-2; 1-\alpha}$
$\rho < 0$	$T < t_{n-2; \alpha} (= -t_{n-2; 1-\alpha})$

Bsp 7.30 Angenommen, für eine Stichprobe der Größe $n = 35$ aus einer bivariaten Normalverteilung $(X, Y) \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ ergibt sich ein Korrelationskoeffizient von $R = 0.30$. Der Wert der Teststatistik für einen Test von:

$$\mathcal{H}_0 : \rho = 0 \quad \text{gegen} \quad \mathcal{H}_1 : \rho > 0$$

ist gegeben durch:

$$T = \frac{0.3 \sqrt{33}}{\sqrt{1-0.09}} = 1.8066 \implies p\text{-Wert} = P(t(33) \geq 1.8066) \doteq 0.0400$$

Wie der p -Wert zeigt, wird die Nullhypothese ($\hat{=} X, Y$ unabhängig) auf dem Niveau 5% zugunsten von $\rho > 0$ verworfen, nicht aber auf dem Niveau 1%. ■

7.4.11 Normal-QQ-Plot

Häufig möchte man überprüfen, ob ein Datensatz aus einer bestimmten Verteilungsfamilie stammt. Für derartige **Verteilungshypothesen** gibt es zahlreiche formale statistische

Tests (vgl. den folgenden Abschnitt) aber auch grafische Methoden, wie den **Q(uantilen)-Q(uantilen)-Plot**. Der QQ-Plot ist insbesondere dann eine gute Methode zur Überprüfung von Verteilungshypothesen, wenn die Verteilungsfunktion wie folgt darstellbar ist:

$$F(x) = P(X \leq x) = F_0\left(\frac{x - c}{d}\right), \quad x \in \mathbb{R}, \quad c \in \mathbb{R}, \quad d > 0$$

(F_0 ist eine nicht von unbekannten Parametern abhängige VF.) Dann sagt man, dass die Verteilung zu einer **L(age)S(kalen)-Familie** gehört. In diesem Fall lässt sich der Graph von F durch „Strecken“ der Ordinatenachse in eine Gerade transformieren.

Im Folgenden betrachten wir speziell die LS-Familie der **Normalverteilungen** etwas genauer. Für die VF einer normalverteilten sG $X \sim N(\mu, \sigma^2)$ gilt:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad x \in \mathbb{R}$$

(F_0 entspricht der VF Φ der Standardnormalverteilung $N(0, 1)$.) Zwischen den Quantilen von F und den Quantilen von Φ besteht die folgende **lineare** Beziehung:

$$x_p = \mu + \sigma z_p$$

Die Quantile $z_p = \Phi^{-1}(p)$ sind bekannt, die Quantile $x_p = F^{-1}(p)$ sind nicht bekannt (da sie von den unbekannten Parametern μ und σ abhängen), können aber aus den gegebenen Beobachtungen x_1, x_2, \dots, x_n geschätzt werden. Ist $x_{(i)}$ die i -te Ordnungsstatistik (vgl. 1.7.1), so gilt:

$$x_{(i)} \approx F^{-1}\left(\frac{i}{n}\right), \quad i = 1, 2, \dots, n$$

Nun gilt für $i = n$ ungünstigerweise $F^{-1}(n/n) = F^{-1}(1) = \infty$. Aus diesem Grund betrachtet man (meist) die folgende modifizierte Beziehung:

$$x_{(i)} \approx F^{-1}\left(\frac{i - 0.5}{n}\right), \quad i = 1, 2, \dots, n$$

Der Normal-QQ-Plot wird nun wie folgt erstellt:

- (1) Daten der Größe nach ordnen: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$
- (2) Der kumulierte Anteil der Daten links von $x_{(i)}$ ist gegeben durch:

$$p_i = \frac{i - 0.5}{n}, \quad i = 1, 2, \dots, n$$

(3) Bestimme die p_i -Quantile der Standardnormalverteilung:

$$z_i = \Phi^{-1}(p_i), \quad i = 1, 2, \dots, n$$

(4) Zeichne die folgenden Punkte in ein übliches Koordinatensystem:

$$(x_{(i)}, z_i), \quad i = 1, 2, \dots, n$$

Für eine Normalverteilung liegen die Punkte annähernd auf einer Geraden:

$$z_i \approx \frac{x_{(i)} - \mu}{\sigma}$$

Bem: Häufig zeichnet man auch die Punkte mit vertauschten Koordinaten, d. h. die Punkte $(z_i, x_{(i)}), i = 1, 2, \dots, n$.

(5) Lege eine Vergleichsgerade durch die Punkte, beispielsweise eine „robuste“ Gerade durch das 1. und 3. Quartil (oder durch die Hinges). Nicht unüblich ist es auch, die ML-Gerade durchzulegen, d. h. die Gerade:

$$z = \frac{x - \hat{\mu}}{\hat{\sigma}} \quad \text{mit} \quad \hat{\mu} = \bar{x}_n, \quad \hat{\sigma} = \sqrt{\frac{n-1}{n}} s_n \quad (\text{oder } \hat{\sigma} = s_n)$$

Bem: Üblicherweise verwendet man entsprechende Software zur Erstellung von QQ-Plots. Es gibt aber auch vorgefertigte **W-Netze**, meist versehen mit Skalen und diversen Hilfslinien, die das Eintragen der Punkte und das Ablesen von Schätzwerten für die Parameter erleichtern. (Vgl. den Anhang: Normal-W-Netz für ein typisches Normalnetz.)

Bsp 7.31 Zehn Beobachtungen der effektiven Nutzungsdauer [min] von in Notebooks verwendeten Akkus waren wie folgt:

176 183 185 190 191 192 201 205 214 220

Es besteht die Vermutung, dass die Nutzungsdauer der Akkus eine normalverteilte sG ist. Der folgende R-Output zeigt die einzelnen Schritte zur Erstellung eines entsprechenden QQ-Plots (vgl. Abb 7.16).

```
x <- c(176,183,185,190,191,192,201,205,214,220)
n <- length(x)
x <- sort(x)
p <- (1:n-0.5)/n
z <- qnorm((1:n-0.5)/n)
```

```
nnetz <- data.frame(x=x, p=p, z=z)
round(nnetz, 3)
    x      p      z
1 176  0.05 -1.645
2 183  0.15 -1.036
3 185  0.25 -0.674
4 190  0.35 -0.385
5 191  0.45 -0.126
6 192  0.55  0.126
7 201  0.65  0.385
8 205  0.75  0.674
9 214  0.85  1.036
10 220  0.95  1.645
```

Die eingezeichnete Linie entspricht der ML–Geraden. Wie sich zeigt, lässt sich die Normalverteilung als Verteilungsmodell nicht ausschließen. (Das ist kein „Beweis“ für die Gültigkeit des Normalmodells, andere Modelle mögen ebenso adäquat sein.) Die ML–Schätzwerte von μ und σ lassen sich hier einfach berechnen:

$$\hat{\mu} = \bar{x} = 195.7 \quad \text{und} \quad \hat{\sigma} = \sqrt{\frac{9}{10}} \underbrace{s_n}_{=14.032} = 13.312$$

Grobe Schätzwerte für μ und σ können im positiven Fall (d. h., wenn die Punkte ausreichend gut auf einer Geraden liegen) aber auch wie in Abb 7.16 angedeutet dem QQ–Plot entnommen werden. ■

7.4.12 Chiquadrat–Anpassungstests

Die diversen **Chiquadrat–Tests**²¹ gehören zu den ältesten Methoden der schließenden Statistik. Wie in 4.2.4 gezeigt, gilt für $X \sim N(\mu, \sigma^2)$:

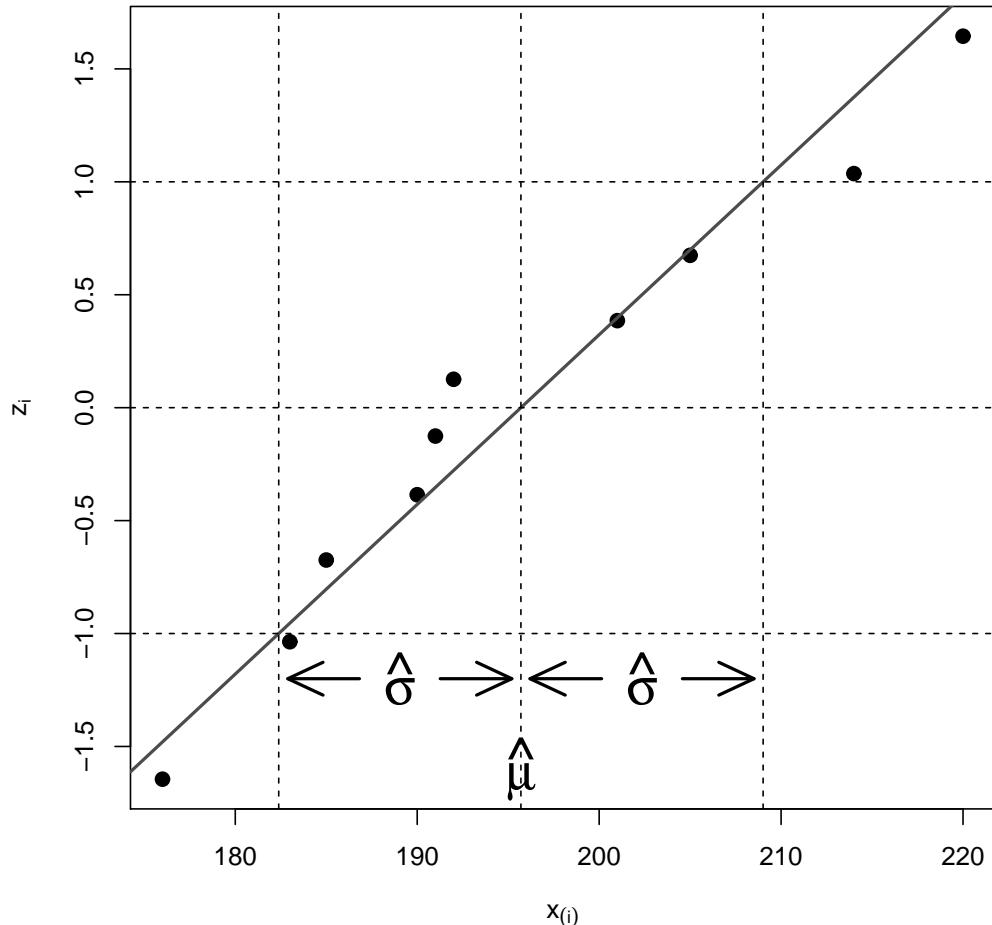
$$\left(\frac{X - \mu}{\sigma} \right)^2 \sim \chi^2(1)$$

Hat man n unabhängige sGn $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, n$, so gilt nach dem Additionstheorem für Chiquadratverteilungen (vgl. 6.2.3):

$$\sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2(n)$$

²¹Entwickelt um 1900 vom engl. Statistiker KARL PEARSON (1857–1938).

Abbildung 7.16: Normal-QQ-Plot für die Daten von Bsp 7.30



Neben diesen exakten Resultaten gibt es aber auch sGn, deren Summe ihrer Quadrate *approximativ* einer Chiquadratverteilung folgt. So gilt für einen multinomial verteilten stochastischen Vektor $\mathbf{X} = (X_1, X_2, \dots, X_k)' \sim M(n, p_1, p_2, \dots, p_k)$, dass:

$$Q_{k-1} = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} \xrightarrow{D} \chi^2(k-1)$$

Diese Überlegungen bilden die Basis für die folgenden Tests.

Einfacher Chiquadrat-Anpassungstest: Der Merkmalraum M eines statistischen Experiments zerfalle in k paarweise disjunkte Teilmengen A_1, \dots, A_k , wobei $p_i = P(A_i)$ für $i = 1, \dots, k$. Das Experiment werde n Mal (unabhängig) wiederholt und X_i sei die Anzahl der Versuchsausgänge in A_i . Dann ist die Teststatistik eines Tests von:

$$\mathcal{H}_0 : p_i = p_{i0}, \quad i = 1, \dots, k \quad \text{gegen} \quad \mathcal{H}_1 : \exists i \text{ mit } p_i \neq p_{i0}$$

gegeben durch:

$$Q_{k-1} = \sum_{i=1}^k \frac{(X_i - np_{i0})^2}{np_{i0}}$$

\mathcal{H}_0 wird verworfen, falls:

$$Q_{k-1} > \chi_{k-1; 1-\alpha}^2$$

Bem: Der obige Test hat nur approximativ das Niveau α . Verschiedene Regeln haben sich etabliert, um die Zulässigkeit der χ^2 -Approximation zu gewährleisten. Die übliche Regel besagt, dass man den Test nur dann verwenden soll, wenn $np_{i0} \geq 5$, $i = 1, \dots, k$; andernfalls müssen benachbarte Klassen zusammengefasst werden. Eine andere Regel verlangt nur, dass $np_{i0} \geq 1$, $i = 1, \dots, k$, und dass 80% der np_{i0} größer oder gleich 5 sind.

Bsp 7.32 Ein Würfel wird 60 Mal geworfen, mit dem folgenden Ergebnis:

Augenzahl	1	2	3	4	5	6
Häufigkeit	13	19	11	8	5	4

Handelt es sich um einen „fairen“ Würfel? Bezeichnet p_i die Wahrscheinlichkeit, die Augenzahl i zu werfen, so sind die folgenden Hypothesen zu testen:

$$\mathcal{H}_0 : p_i = \frac{1}{6}, \quad i = 1, 2, 3, 4, 5, 6 \quad \text{gegen} \quad \mathcal{H}_1 : \exists i \text{ mit } p_i \neq \frac{1}{6}$$

Die folgende Tabelle zeigt die einzelnen Rechenschritte für den χ^2 -Anpassungstest:

Klasse	X_i	p_{i0}	np_{i0}	$(X_i - np_{i0})^2 / np_{i0}$
1	13	1/6	10	0.9
2	19	1/6	10	8.1
3	11	1/6	10	0.1
4	8	1/6	10	0.4
5	5	1/6	10	2.5
6	4	1/6	10	3.6
Summe	60	1	60	15.6

Testet man auf dem Niveau $\alpha = 5\%$, so ist der Wert der Teststatistik $Q_5 = 15.6$ mit $\chi_{5; 0.95}^2 = 11.07$ zu vergleichen. Wegen $15.6 > 11.07$ wird die Nullhypothese verworfen. Den (approximativen) p -Wert der \mathcal{H}_0 berechnet man wie folgt:

$$p\text{-Wert} = P(\chi^2(5) \geq 15.6) \doteq 0.0081$$

Die Nullhypothese wird also auch auf dem vorsichtigeren Niveau 1% verworfen. (Man beachte, dass hier wegen $np_{i0} \geq 5$ die Chi-Quadrat-Approximation der Teststatistik nach der strengeren Faustregel ausreichend gut ist.) ■

Zusammengesetzter Chi-Quadrat-Anpassungstest: Sind die Wahrscheinlichkeiten p_i für die Klassen A_i durch die \mathcal{H}_0 nicht vollständig spezifiziert, gilt also $p_i = p_i(\theta)$ für einen (unbekannten) s -dimensionalen Parametervektor $\theta \in \Theta$, so ist der Anpassungstest zu modifizieren. Die Teststatistik eines Tests von

$$\mathcal{H}_0 : p_i = p_i(\theta), \quad i = 1, 2, \dots, k \quad \text{gegen} \quad \mathcal{H}_1 : \exists i \text{ mit } p_i \neq p_i(\theta)$$

ist gegeben durch:

$$Q_{k-s-1} = \sum_{i=1}^k \frac{[X_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})}$$

Dabei ist der Schätzwert $\hat{\theta}$ für den Parameter θ so zu wählen, dass Q_{k-s-1} minimal wird. Die Nullhypothese ist zu verwerfen, falls:

$$Q_{k-s-1} > \chi^2_{k-s-1; 1-\alpha}$$

Bem: Den obigen Schätzer von θ nennt man den *Minimum-Chi-Quadrat-Schätzer*. Das ist die korrekte Vorgangsweise. Praktisch geht man aber fast nie auf diese Weise vor. Der Grund dafür liegt darin, dass der Minimum-Chi-Quadrat-Schätzwert nur in Ausnahmefällen einfach ermittelt werden kann. Üblich ist die folgende Vorgangsweise: Man bestimmt zunächst auf Basis der *ursprünglichen* Beobachtungen (also nicht auf Basis der Zählvariablen X_i) den ML-Schätzwert von θ und setzt diesen Schätzwert in Q_{k-s-1} ein. Da sich die ML-Schätzwerte meist aber von den Minimum-Chi-Quadrat-Schätzwerten unterscheiden, ist dabei zu bedenken, dass sich dadurch das Niveau des Tests erhöht, der χ^2 -Test also „verwerfungsfreudiger“ wird.

Bsp 7.33 Wir demonstrieren die in der obigen **Bemerkung** angesprochene Vorgangsweise an einem simulierten Datensatz. Mittels (zusammengesetztem) Chi-Quadrat-Anpassungstest (mit $\alpha = 0.05$) soll überprüft werden, ob die folgenden Werte (Datenfile: `rn01.txt`) aus einer Normalverteilung stammen.

-0.0895	-1.0233	0.9375	-1.1317	-0.7107
-1.1695	1.0654	-0.6804	-1.7258	0.8132
1.4419	0.6723	0.1387	-0.8595	-0.7523
1.2296	1.1508	-0.6080	0.8062	0.2171
-0.3735	-0.8320	0.2869	-1.8189	-1.5731
2.0157	-0.0720	2.6289	-0.2433	0.1733
0.9232	-0.1786	-0.5217	1.4320	-0.8701
0.8075	-0.5106	0.7435	0.8479	-0.8299

Für den (zusammengesetzten) Chiquadrat–Anpassungstest von:

$$\mathcal{H}_0 : X \sim N(\mu, \sigma^2) \quad \text{gegen} \quad \mathcal{H}_1 : X \not\sim N(\mu, \sigma^2)$$

werden zunächst die ML–Schätzwerte von μ und σ^2 bestimmt:

$$\hat{\mu} = \bar{x}_n = 0.0439, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = 1.0652$$

Außerdem muss eine Klasseneinteilung vorgenommen werden; dafür gibt es zahlreiche Möglichkeiten. Eine komfortable Einteilung ergibt sich, wenn man unter \mathcal{H}_0 *gleichwahrscheinliche* Klassen nimmt. Hält man sich dabei an die Regel $np_{i0} \geq 5$, so ist bei 40 Beobachtungen die maximale Klassenzahl gleich 8. (Die minimale Klassenzahl ist 4, da $k - s - 1 = k - 3 \geq 1$ sein muss.) Bei 8 Klassen lautet diese Einteilung mit $\hat{x}_p = \hat{\mu} + z_p \hat{\sigma}$ wie folgt:

$$(-\infty, \hat{x}_{1/8}], (\hat{x}_{1/8}, \hat{x}_{2/8}], \dots, (\hat{x}_{7/8}, \infty)$$

Klassen	X	\hat{p}	$n\hat{p}$	$(X - n\hat{p})^2/n\hat{p}$
$(-\infty, -1.1433]$	4	0.125	5	0.2
$(-1.1433, -0.6522]$	9	0.125	5	3.2
$(-0.6522, -0.2849]$	4	0.125	5	0.2
$(-0.2849, 0.0439]$	4	0.125	5	0.2
$(0.0439, 0.3728]$	4	0.125	5	0.2
$(0.3728, 0.7401]$	1	0.125	5	3.2
$(0.7401, 1.2312]$	10	0.125	5	5.0
$(1.2312, \infty)$	4	0.125	5	0.2
Summe	40	1.000	40	12.4

Hier gilt $k - s - 1 = 8 - 2 - 1 = 5$ und wegen $Q_5 = 12.4 > \chi^2_{5;0.95} = 11.071$ wird die Nullhypothese auf dem Niveau 5% verworfen.

Der obige Test lässt sich mit den folgenden R-Zeilen durchführen:

```

dat <- scan("rn01.txt")
m <- mean(dat)
s2 <- var(dat)*(length(dat)-1)/length(dat)
class.2 <- cut(dat, breaks=c(-Inf,m+sqrt(s2)*qnorm((1:7)/8),Inf))
result <- chisq.test(table(class.2), p=rep(1/8,8))
result

Chi-squared test for given probabilities

data: table(class.2)
X-squared = 12.4, df = 7, p-value = 0.08815

```

Die Zahl der Freiheitsgrade $df = 7$ ist hier *nicht* korrekt (und daher auch der p -Wert nicht), da `chisq.test()` stets von einem einfachen Anpassungstest ausgeht. Die korrekte Anzahl²² ist $df = 5$ und der p -Wert ist:

$$p\text{-Wert} = P(\chi^2(5) \geq 12.4) \doteq 0.0297$$

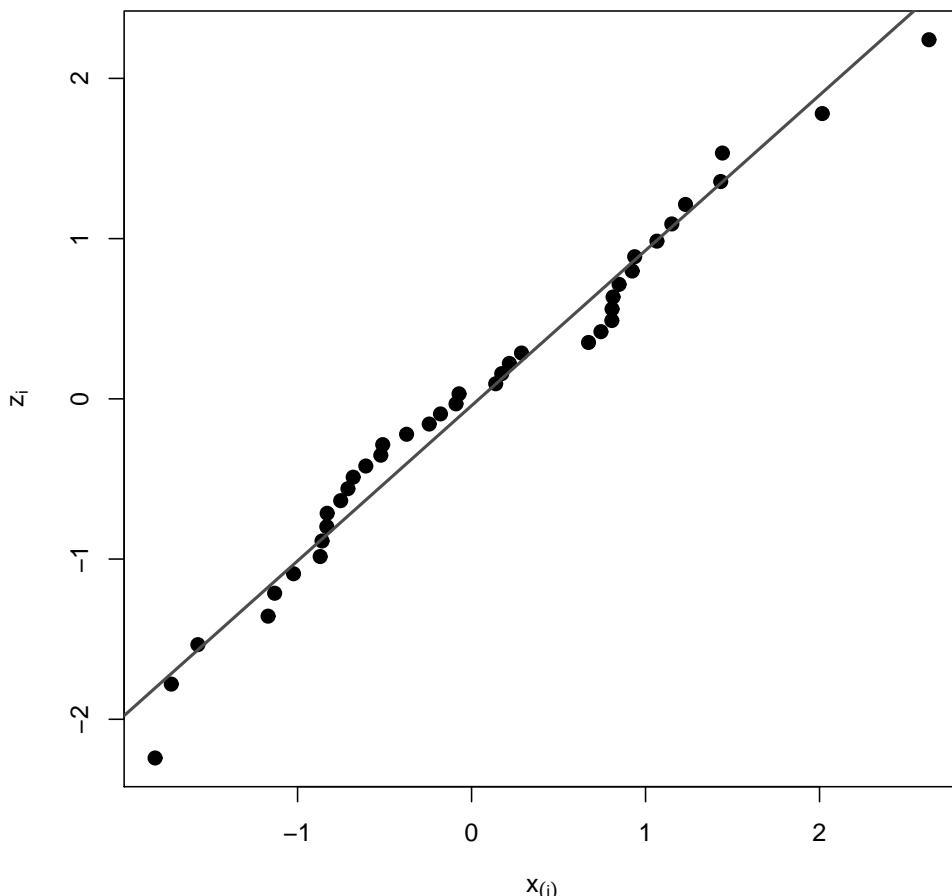
Am p -Wert zeigt sich, dass die Nullhypothese zumindest zweifelhaft ist; für $\alpha = 5\%$ wird sie verworfen, für das „vorsichtigere“ $\alpha = 1\%$ aber nicht.

Interessant ist hier auch der graphische Test mittels Normal-QQ-Plot (Abb 7.17). Die Geradenanpassung ist augenscheinlich nicht schlecht, das Normalmodell daher nicht auszuschließen. (Bem: Tatsächlich handelt es sich bei diesem Datensatz um simulierte Werte aus einer Standardnormalverteilung.) ■

Aufgaben

- 7.1 Simulieren Sie (a) $n = 10$ bzw. (b) $n = 100$ standardnormalverteilte Beobachtungen, zeichnen Sie die empirische Verteilungsfunktion \hat{F}_n und bestimmen Sie grafisch den größten Abstand von Φ (= VF der Standardnormalverteilung), d. h., bestimmen Sie $\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - \Phi(x)|$. Führen Sie die Simulationen mehrfach durch und kommentieren Sie die Ergebnisse. (Hinweis: Sie können dazu auch die (eigene) Funktion `dist.norm()` verwenden.)

²²Wie Simulationsstudien zeigen, liegt der korrekte Wert meist zwischen $k - 1$ und $k - s - 1$.

Abbildung 7.17: Normal-QQ-Plot für rn01.txt

7.2 X_1, X_2, \dots, X_n sei eine Stichprobe von $X \sim U(0, \theta)$ ($\theta > 0$).

- (a) Bestimmen Sie den Momentenschätzer von θ .
- (b) Ist der Momentenschätzer unverzerrt? Konsistent?
- (c) Wie lautet in diesem Fall der linear effiziente Schätzer von θ ?

7.3 Die folgende Tabelle ist die Zusammenfassung einer Stichprobe der Größe $n = 55$ von $X \sim P(\lambda)$:

x	0	1	2	3	4	5
Häufigkeit	7	14	12	13	6	3

- (a) Bestimmen Sie den ML-Schätzer von λ .
- (b) Ist der ML-Schätzer unverzerrt? Konsistent?
- (c) Bestimmen Sie den ML-Schätzwert von λ .
- (d) Bestimmen Sie den ML-Schätzwert von $P(X = 2)$.

- 7.4 Bei einem Glücksspiel benötigten zehn Spieler die folgenden Anzahlen von Runden bis zum ersten Gewinn:

9 10 1 4 3 12 13 13 8 22

Bestimmen Sie (a) den Momentenschätzer und (b) den ML–Schätzer für die Wahrscheinlichkeit p ($0 < p < 1$) mit der man bei diesem Spiel gewinnt.

- 7.5 X_1, X_2, \dots, X_n sei eine Stichprobe von $X \sim N(0, \sigma^2)$. Bestimmen Sie den ML–Schätzer von σ^2 und von σ .

- 7.6 Argumentieren Sie, dass für eine Stichprobe X_1, X_2, \dots, X_n von $X \sim U(0, \theta)$ ($\theta > 0$) der ML–Schätzer von θ gegeben ist durch:

$$\hat{\theta} = \max\{X_1, X_2, \dots, X_n\} = X_{(n)}$$

- 7.7 Zeigen Sie für eine Stichprobe X_1, X_2, \dots, X_n von $X \sim U(0, \theta)$ ($\theta > 0$), dass:

$$T = \frac{\max\{X_1, X_2, \dots, X_n\}}{\theta} = \frac{X_{(n)}}{\theta}$$

eine Pivotgröße ist.

- 7.8 Berechnen Sie $I = \int_0^1 4\sqrt{1-x^2} dx$ mittels Monte Carlo Integration und bestimmen Sie ein (approximativer) 95%–Konfidenzintervall für I (vgl. Bsp 7.14). Enthält das Intervall den exakten Wert von I ?

- 7.9 Fünfzehn unabhängige Beobachtungen von $X \sim N(\mu, \sigma^2)$ waren wie folgt:

492 512 502 487 500 483 490 498 489 503
497 494 508 506 497

Bestimmen Sie die ML–Schätzer/Schätzwerte (a) von μ , (b) von σ^2 und (c) von σ .

- 7.10 Bestimmen Sie (a) ein 90%–, (b) ein 95%– und (c) ein 99%–Konfidenzintervall für den Mittelwert μ auf Basis der Daten von Aufgabe 7.9.
- 7.11 Bestimmen Sie ein 95%–Konfidenzintervall (a) für die Varianz σ^2 und (b) für die Streuung σ auf Basis der Daten von Aufgabe 7.9.

- 7.12 Zusammengefasst ergab sich für zwei Stichproben aus unabhängigen Normalverteilungen $X \sim N(\mu_X, \sigma_X^2)$ bzw. $Y \sim N(\mu_Y, \sigma_Y^2)$:

$$\begin{array}{lll} n_1 = 10 & \bar{x} = 104 & s_X^2 = 290 \\ n_2 = 20 & \bar{y} = 114 & s_Y^2 = 510 \end{array}$$

Ermitteln Sie unter der Voraussetzung $\sigma_X^2 = \sigma_Y^2$ ein 95%–Konfidenzintervall für die Differenz $\mu_Y - \mu_X$ der Mittelwerte.

- 7.13 Werden (herkömmliche) Glühlampen (60 W) unter normalen Bedingungen (230 V, Glühfadentemperatur 2700 K) vom Einschalten bis zum Ausfall beobachtet, so folgt die Brenndauer näherungsweise einer Exponentialverteilung. Angenommen, bei 25 Glühlampen ergibt sich eine mittlere Brenndauer von 976 Stunden. Bestimmen Sie – inklusive Herleitung – den ML–Schätzer/Schätzwert (a) für den Mittelwert und (b) für den Median der Brenndauer, sowie (c) für die Wahrscheinlichkeit, dass eine Glühlampe länger als 2000 Stunden brennt.
- 7.14 Bestimmen Sie (a) ein exaktes 95%–Konfidenzintervall, (b) das 95%–Wald–Intervall und (c) das 95%–Scoreintervall für die mittlere Brenndauer der Glühlampen von Aufgabe 7.13.
- 7.15 Angenommen, bei der Herstellung von ICs mittels Photolithographie stellt sich heraus, dass von 300 zufällig ausgewählten ICs 13 defekt sind. Bestimmen Sie (a) den ML–Schätzwert, (b) das 95%–Standardintervall und (c) das 95%–Scoreintervall für den Defektanteil p bei dieser Produktionsmethode.
- 7.16 Bei der optischen Prüfung von 20 zufällig herausgegriffenen Autoblechen wurden die folgenden Anzahlen von Lackierungsfehlern pro Blech gefunden:

1 7 1 3 2 5 2 8 5 4 6 5 4 6 2 4 5 2 3 6

Bestimmen Sie (a) den ML–Schätzwert, (b) das 95%–Standardintervall und (c) das 95%–Scoreintervall für die mittlere Fehlerzahl λ pro Blech.

- 7.17 Bestimmen Sie ein 95%–Bootstrapintervall für den Mittelwert μ auf Basis der folgenden zehn Beobachtungen:

79 88 39 17 40 27 45 100 50 71

(Hinweis: Sie können dazu die (eigene) Funktion `percentciboot()` verwenden.)

- 7.18 Ein Produzent behauptet, dass (höchstens) 1% seiner Produkte fehlerhaft sind. Zur Prüfung dieser Behauptung entnehmen Sie – ohne Zurücklegen – aus einem Los der Größe $N = 1000$ zufällig 55 Einheiten, und beschließen, das Los nur dann zu akzeptieren, wenn die Stichprobe nicht mehr als 1 fehlerhafte Einheit enthält.
- Formulieren Sie die Null– und Alternativhypothese.
 - Wie groß ist bei diesem Test die Wahrscheinlichkeit eines Fehlers 1. Art?
 - Wie groß ist die Wahrscheinlichkeit eines Fehlers 2. Art, wenn der Defektanteil tatsächlich 5% (10%) beträgt?

(Hinweis: Rechnen Sie mit der Binomialverteilung.)

- 7.19 Für eine normalverteilte stochastische Größe X , deren Varianz mit $\sigma^2 = 4$ bekannt ist, möchten wir $\mathcal{H}_0 : \mu = 100$ gegen $\mathcal{H}_1 : \mu \neq 100$ auf Basis einer Stichprobe der Größe $n = 9$ testen.

- (a) Wenn der kritische Bereich durch $\bar{x} < 98.5$ oder $\bar{x} > 101.5$ gegeben ist, wie groß ist die Wahrscheinlichkeit eines Typ I–Fehlers?
- (b) Wenn der tatsächliche Mittelwert gleich 103 ist, wie groß ist die Wahrscheinlichkeit eines Typ II–Fehlers?

7.20 Testen Sie auf Basis der Daten von **Aufgabe 7.9** die Hypothese $\mathcal{H}_0 : \mu = 500$ gegen $\mathcal{H}_1 : \mu \neq 500$. Die Wahrscheinlichkeit eines Typ I–Fehlers soll $\alpha = 0.05$ betragen. (Gibt es einen Zusammenhang mit den in **Aufgabe 7.10** bestimmten Konfidenzintervallen?) Wie groß ist der p –Wert?

7.21 Bei 15 unabhängigen Messungen des Gewichts von einem Blatt Papier ergibt sich eine Stichprobenstreuung von $s = 0.0083$ g. Wenn die Messwerte normalverteilt sind, testen Sie $\mathcal{H}_0 : \sigma = 0.01$ gegen $\mathcal{H}_1 : \sigma \neq 0.01$ mit $\alpha = 5\%$. Wie groß ist der p –Wert?

7.22 Der Natriumgehalt [mg] von dreißig 300g Packungen Cornflakes war wie folgt (Datenfile: `sodium.txt`):

131.15	130.69	130.91	129.54	129.64	128.77	130.72	128.33	128.24	129.65
130.14	129.29	128.71	129.00	129.39	130.42	129.53	130.12	129.78	130.92
131.15	130.69	130.91	129.54	129.64	128.77	130.72	128.33	128.24	129.65

Wenn es sich um normalverteilte Beobachtungen handelt:

- (a) Unterscheidet sich der mittlere Natriumgehalt signifikant von 130 mg? (Nehmen Sie $\alpha = 5\%$.) Wie groß ist der p –Wert?
- (b) Wie groß ist die Power des Tests, wenn der wahre Natriumgehalt 130.5 mg beträgt? (Hinweis: Verwenden Sie Abb 7.15.)
- (c) Lässt sich (a) auch mit einem Konfidenzintervall für den mittleren Natriumgehalt beantworten?

7.23 Angenommen, Sie finden bei der Losprüfung von **Aufgabe 7.18** in der Stichprobe 2 fehlerhafte Einheiten. Wie groß ist in diesem Fall der p –Wert?

7.24 Fortsetzung von **Aufgabe 7.12**:

- (a) Testen Sie unter der Voraussetzung $\sigma_X^2 = \sigma_Y^2$, ob die beiden Mittelwerte gleich sind, d. h., testen Sie $\mathcal{H}_0 : \mu_X = \mu_Y$ gegen $\mathcal{H}_1 : \mu_X \neq \mu_Y$. ($\alpha = 5\%$)
- (b) Testen Sie, ob die beiden Varianzen als gleich angesehen werden können, d. h., testen Sie $\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2$ gegen $\mathcal{H}_1 : \sigma_X^2 \neq \sigma_Y^2$. ($\alpha = 10\%$)

7.25 Bei 15 männlichen Erwachsenen wurde die Cholesterinkonzentration im Blut vor und nach einem 3–monatigen Diät– und Bewegungsprogramm gemessen, mit dem folgenden Ergebnis (Datenfile: `cholesterol.txt`):

	vor	nach		vor	nach		vor	nach
1	265	229	6	245	241	11	283	246
2	240	231	7	287	234	12	240	218
3	258	227	8	314	256	13	238	219
4	295	240	9	260	247	14	225	226
5	251	238	10	279	239	15	247	233

Bewirkt die Therapie eine signifikante Reduktion der mittleren Cholesterinkonzentration im Blut? ($\alpha = 5\%$) Wie groß ist der p -Wert?

- 7.26 Die folgenden acht Beobachtungspaare stammen von einer bivariaten Normalverteilung $(X, Y) \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$:

	1	2	3	4	5	6	7	8
x	10.33	9.53	9.82	10.11	8.99	10.37	9.99	9.01
y	9.75	8.44	10.03	10.67	9.30	10.68	11.14	7.87

Bestimmen Sie den Korrelationskoeffizienten R und testen Sie mit $\alpha = 5\%$, ob X und Y unabhängig (gegen eine positive Abhängigkeit) sind, d. h., testen Sie:

$$\mathcal{H}_0 : \rho = 0 \quad \text{gegen} \quad \mathcal{H}_1 : \rho > 0$$

Wie groß ist der p -Wert?

- 7.27 Erstellen Sie den Normal-QQ-Plot für die Daten von Bsp 7.31 mit der Hand unter Verwendung eines vorgefertigten W-Netzes (vgl. Anhang: Normal-W-Netz).
- 7.28 Prüfen Sie mittels QQ-Plot, ob die Daten von Aufgabe 7.22 als normalverteilt angesehen werden können.
- 7.29 Erstellen Sie für die folgenden vier Datenvektoren einen Normal-QQ-Plot und kommentieren Sie die Ergebnisse:

```
x <- rnorm(50)
y <- x + 2
z <- x/2
u <- 2*x
```

- 7.30 Erstellen Sie für die acht Batches von `euroweight.txt` – angeordnet in einem 4×2 -Array – einen Normal-QQ-Plot und kommentieren Sie die Ergebnisse. (Vgl. auch Aufgabe 1.7.)

- 7.31 Ein Würfel wird 100 Mal geworfen, mit dem Ergebnis:

Augenzahl	1	2	3	4	5	6
Häufigkeit	13	17	9	17	18	26

Ist der Würfel ausbalanciert? ($\alpha = 5\%$)

- 7.32 Die folgenden Daten wurden mittels `round(sort(runif(30)), 4)` erzeugt:

0.0920	0.1469	0.1696	0.1903	0.2304	0.2415	0.2550	0.2917	0.2949	0.3201
0.3300	0.3474	0.3690	0.4259	0.4725	0.4749	0.5155	0.5820	0.5959	0.6509
0.6829	0.6950	0.7144	0.7415	0.8392	0.8459	0.8678	0.8853	0.9005	0.9640

Prüfen Sie mit $\alpha = 5\%$, ob die Daten als Stichprobe von $X \sim U(0, 1)$ angesehen werden können. Nehmen Sie dazu die Klasseneinteilung $[0, 0.2), [0.2, 0.4), \dots, [0.8, 1]$.

- 7.33 Prüfen Sie mit $\alpha = 5\%$, ob die Daten von `rn01.txt` aus einer Standardnormalverteilung stammen, d. h., testen Sie $H_0 : X \sim N(0, 1)$ gegen $H_1 : X \not\sim N(0, 1)$. Nehmen Sie dazu 8 unter H_0 gleichwahrscheinliche Klassen. Vergleichen Sie das Ergebnis mit dem zusammengesetzten χ^2 -Anpassungstest von Bsp 7.33. (Hinweis: Nehmen Sie die Funktion `chisq.test()`; die Klasseneinteilung lässt sich mittels `cut()` und `table()` durchführen.)

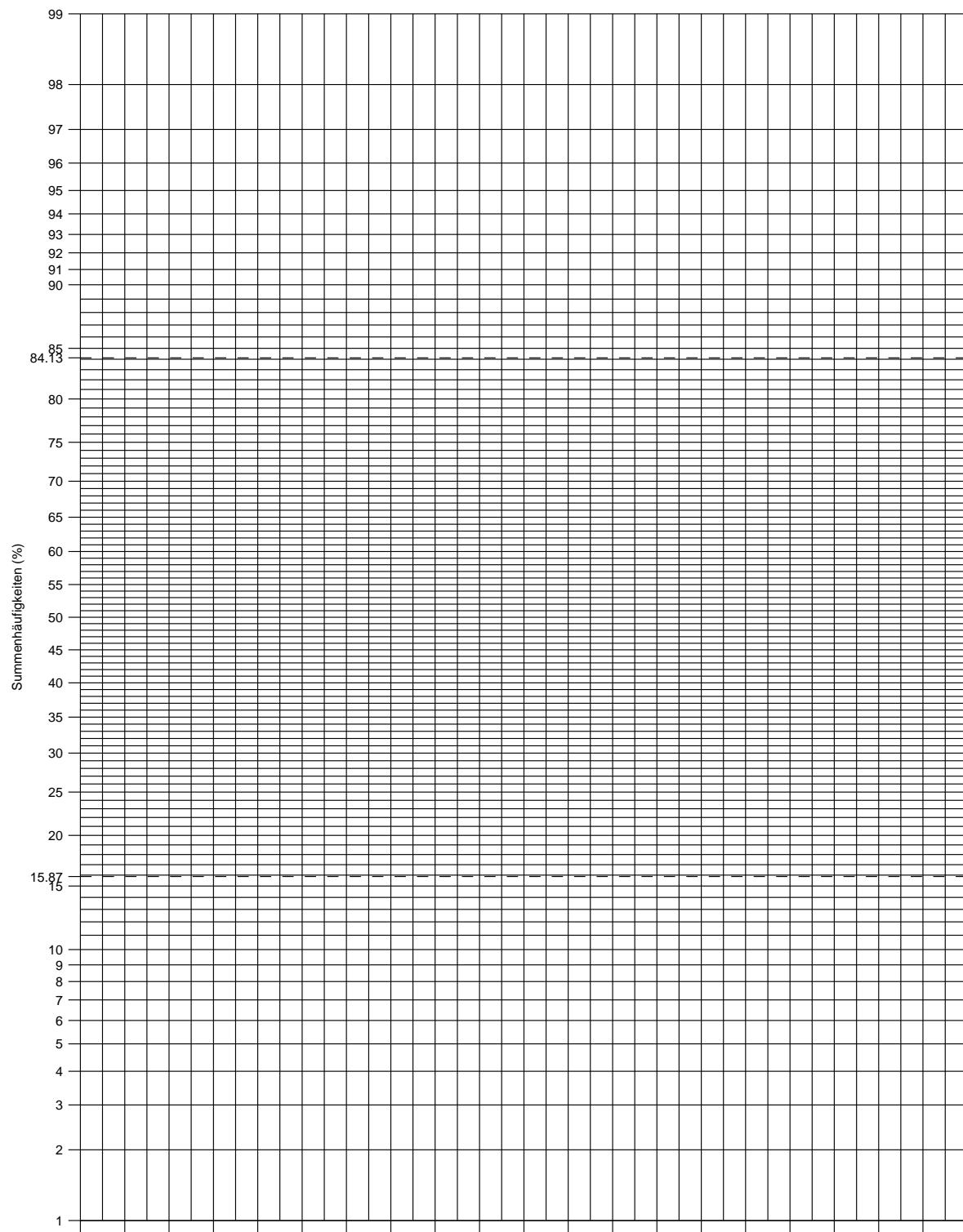
- 7.34 Eine sG X wurde 100 Mal beobachtet, mit dem folgenden Ergebnis:

Wert	0	1	2	3	4
Häufigkeit	24	30	31	11	4

Ist die Poissonverteilung ein geeignetes Modell? ($\alpha = 5\%$) Wie groß ist der p -Wert? (Hinweis: Bestimmen Sie zuerst den ML-Schätzwert für den Parameter λ der Poissonverteilung. Achten Sie auf die Einhaltung der Faustregel $n\hat{p} \geq 5$.)

- 7.35 Der Datensatz `lifetimes.txt` umfasst 24 Beobachtungen der Ausfallzeit einer elektronischen Komponente. Prüfen Sie mit $\alpha = 5\%$, ob die Exponentialverteilung ein geeignetes Modell ist. Wie groß ist der p -Wert? (Hinweis: Bestimmen Sie zuerst den ML-Schätzwert für den Parameter τ der Exponentialverteilung. Klassieren Sie die Daten z. B. wie folgt: $(0, 44], (44, 106], (106, 212], (212, \infty)$. Für die konkrete Durchführung des Tests können Sie auch die (eigene) Funktion `chi2.exp()` nehmen.)

Anhang: Normal-W–Netz



8 Bayes–Statistik

In der (klassischen) **frequentistischen Statistik** bilden Stichproben X_1, X_2, \dots, X_n von $X \sim f(x; \theta)$ die alleinige Quelle von Unsicherheit, wobei (unbekannte) Parameter $\theta \in \Theta$ als **fest** betrachtet werden. Statistische Prozeduren (Schätzer, Konfidenzintervalle, Tests) basieren auf der gemeinsamen Verteilung der Daten:

$$f(\mathbf{x}; \theta) = f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

In der **Bayes–Statistik**¹ andererseits betrachtet man den (unbekannten) Parameter θ als weitere Quelle von Unsicherheit und modelliert das **Vorwissen** (d. h. das Wissen *vor* Ziehung der Stichprobe) über θ in Form einer W–Verteilung (d. h. durch eine W–Funktion $p(\theta)$ oder eine Dichte $f(\theta)$).

Bsp 8.1 Wirft man beispielsweise eine (neue) 1€ Münze fünfmal und bekommt einmal „Kopf“, so würde man aus frequentistischer Perspektive die Wahrscheinlichkeit p für Kopf mit $1/5$ schätzen. Wir wissen aber, dass dieser Wert viel zu niedrig ist, da p in der Nähe von $1/2$ liegen wird. Aus Bayes’scher Perspektive wird man daher dieses Vorwissen in Form einer *A-priori–Verteilung* für p , etwa durch eine Dichte wie in Abb 8.1 dargestellt, modellieren. ■

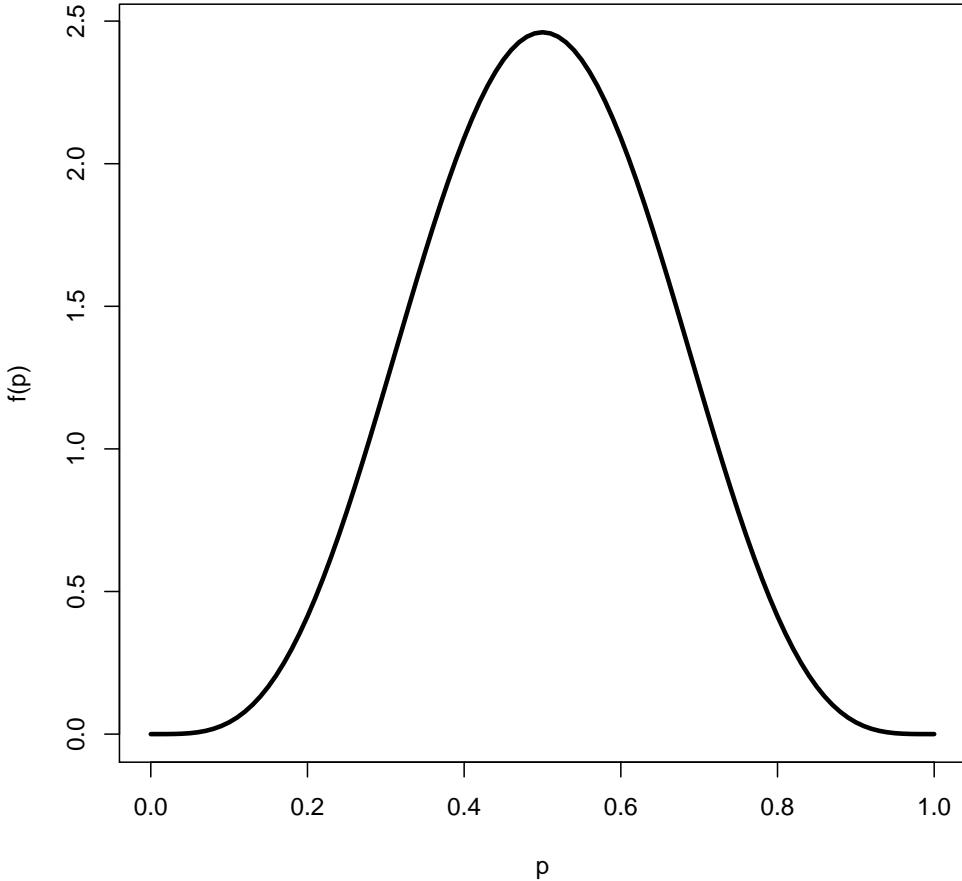
Bem: Ein Vorteil dieser Betrachtungsweise besteht u. a. darin, dass man sich für die Interpretation von Resultaten nicht auf die „Auf-lange-Sicht“–Perspektive berufen muss. In vielen Fällen hat man nur eine beschränkte Zahl von Beobachtungen zur Verfügung und die Vorstellung, dass sich das Experiment beliebig oft wiederholen lässt, ist der Problemstellung häufig nicht adäquat. Man denke auch an folgende Situation: Eine unbekannte Klaviersonate taucht auf und es besteht die Vermutung, dass sie von Mozart stammen könnte. Hier gibt es kein (wiederholbares) „Experiment“; wohl aber lässt sich auf Basis bestimmter Eigenheiten (des Auffindungsorts, der Komposition, etc.) eine A-priori–Wahrscheinlichkeit für die Vermutung angeben.

8.1 A-priori– und A-posteriori–Verteilung

Das Vorwissen über den Parameter $\theta \in \Theta$ wird durch die **A-priori–Verteilung**² modelliert. Unabhängig davon, ob letztere Verteilung diskret oder stetig ist (d. h., durch eine W–Funktion oder durch eine Dichte beschrieben wird), wird sie üblicherweise mit $\pi(\theta)$ bezeichnet. Das ist die eine Informationsquelle über den Parameter θ ; als Zweites gibt es aber auch die Dateninformation $\mathbf{X} = (X_1, X_2, \dots, X_n)'$, im stetigen Fall gegeben durch die folgende **bedingte** Dichte:

¹Vgl. zum Namensgeber die Fußnote zu 2.13.

²engl. meist kurz *prior*

Abbildung 8.1: A-priori-Verteilung für $p = P(\text{Kopf})$ 

$$f(\mathbf{x}|\theta) = f(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$$

Bem: Handelt es sich – wie in der klassischen Statistik – um einen *festen* (unbekannten) Parameter, schreibt man meist $f(x; \theta)$ oder $p(x; \theta)$. Möchte man aber hervorstreichen, dass es sich – wie in der Bayes-Statistik – um eine *bedingte* Dichte handelt, schreibt man $f(x|\theta)$ oder $p(x|\theta)$. (Vgl. dazu auch 5.1.4.)

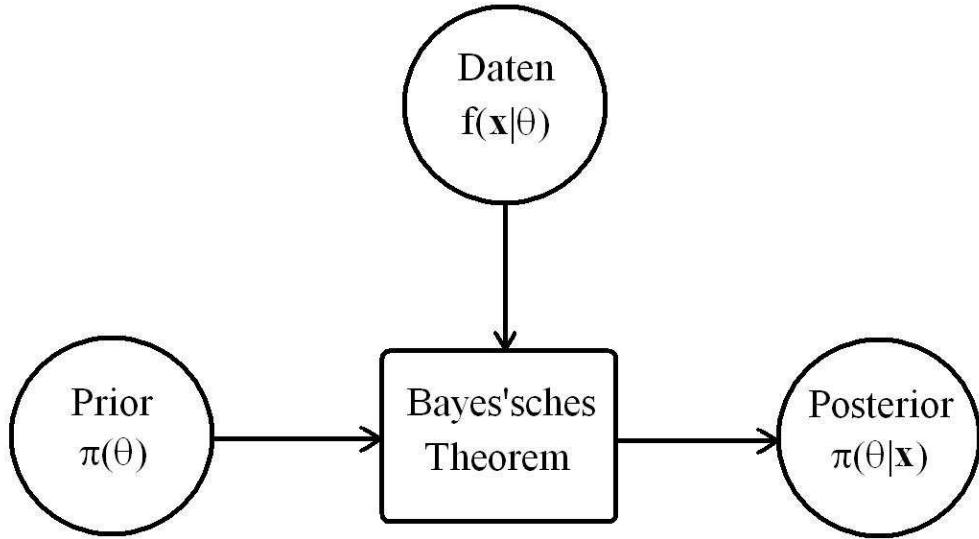
A-priori-Verteilung $\pi(\theta)$ und Dateninformation $f(\mathbf{x}|\theta)$ werden nun über das **Bayes'sche Theorem**³ zur **A-posteriori-Verteilung**⁴ von θ verknüpft:

$$\pi(\theta|\mathbf{x}) = \pi(\theta|\mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}$$

³Eine Verallgemeinerung der Bayes'schen Formel von 2.13.

⁴engl. meist kurz *posterior*

Abbildung 8.2: Verknüpfung von A-priori– und Dateninformation



Der Nenner $m(\mathbf{x})$ repräsentiert die – nicht von θ abhängige – **Randverteilung**⁵ von \mathbf{X} , berechnet über den Satz von der vollständigen Wahrscheinlichkeit (vgl. 2.12):

$$\pi(\theta) \text{ diskret: } m(\mathbf{x}) = \sum_{\theta} f(\mathbf{x}|\theta)\pi(\theta)$$

$$\pi(\theta) \text{ stetig: } m(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}|\theta)\pi(\theta) d\theta$$

Vgl. Abb 8.2 für eine schematische Darstellung der Verknüpfung von A-priori– und Dateninformation; als Schnittpunkt fungiert das Bayes'sche Theorem.

Bsp 8.2 Ein Hersteller behauptet, dass der Defektanteil seiner Produkte nur 5% beträgt, der Abnehmer ist aber der Meinung, dass er bei 10% liegt. Bevor das Ergebnis einer Stichprobenprüfung bekannt wird, geben wir beiden Anteilen eine 50–50 Chance:

$$\pi(0.05) = \pi(0.10) = 0.5$$

Angenommen, in einer Stichprobe der Größe 20 gibt es 3 defekte Einheiten. Legen wir die Binomialverteilung $B(20, \theta)$ zugrunde, ist die Dateninformation gegeben wie folgt:

⁵Auch *A-priori-Prädiktivverteilung* genannt.

$$p(3|\theta = 0.05) = \binom{20}{3} (0.05)^3 (0.95)^{17} = 0.0596$$

$$p(3|\theta = 0.10) = \binom{20}{3} (0.10)^3 (0.90)^{17} = 0.1901$$

Die Randverteilung von X (= Zahl der defekten Einheiten in der Stichprobe) lautet für $x = 3$ wie folgt:

$$m(3) = p(3|0.05)\pi(0.05) + p(3|0.10)\pi(0.10) = 0.1249$$

Die A-posteriori-Wahrscheinlichkeiten sind nun gegeben wie folgt:

$$\pi(0.05|X = 3) = \frac{p(3|0.05)\pi(0.05)}{m(3)} = 0.2387$$

$$\pi(0.10|X = 3) = \frac{p(3|0.10)\pi(0.10)}{m(3)} = 0.7613$$

A-priori hatten wir keine Präferenz für einen der beiden Defektanteile. Nach Beobachtung eines vergleichsweise hohen Defektanteils von $3/20 = 15\%$ in der Stichprobe ist a-posteriori $\theta = 0.10$ aber etwa dreimal so wahrscheinlich wie $\theta = 0.05$. ■

8.2 Konjugierte Verteilungsfamilien

Durch geeignete Wahl der A-priori-Verteilung lässt sich die Bestimmung der A-posteriori-Verteilung vereinfachen.

Konjugierte A-priori-Verteilung: Man nennt eine Familie von A-priori-Verteilungen **konjugiert** zum Modell $f(\mathbf{x}|\theta)$, wenn die A-posteriori-Verteilung zur selben Familie gehört.

Wir betrachten dazu drei Standardsituationen. (Bem: Den im Mittelpunkt des Interesses stehenden Parameter bezeichnen wir im Folgenden – abweichend von früher verwendeten Bezeichnungen – stets mit θ .)

Poisson-Modell: $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ sei eine Stichprobe von $X \sim P(\theta)$. Dann gilt für die durch θ bedingte W-Funktion:

$$p(\mathbf{x}|\theta) = \prod_{i=1}^n p(x_i|\theta) = \prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!}$$

Da für die Berechnung von $\pi(\theta|\mathbf{x})$ nur die von θ abhängigen Terme relevant sind, schreiben wir kürzer:⁶

$$p(\mathbf{x}|\theta) \propto \theta^{\sum_{i=1}^n x_i} e^{-n\theta}$$

An der Form von $p(\mathbf{x}|\theta)$ sieht man, dass die Familie der Gammaverteilungen $\{\text{Gam}(\alpha, \lambda)\}$ (vgl. 4.2.3) konjugiert ist:

$$\pi(\theta) = \frac{\lambda^\alpha \theta^{\alpha-1} e^{-\lambda\theta}}{\Gamma(\alpha)} \propto \theta^{\alpha-1} e^{-\lambda\theta}, \quad \theta > 0$$

(Man beachte, dass auch von $\pi(\theta)$ nur die von θ abhängigen Terme relevant sind.) Da die Randverteilung $m(\mathbf{x})$ (definitionsgemäß) nicht von θ abhängt, ergibt sich die A-posteriori-Dichte von θ wie folgt:

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto p(\mathbf{x}|\theta)\pi(\theta) \\ &\propto \left(\theta^{\sum_{i=1}^n x_i} e^{-n\theta}\right) \left(\theta^{\alpha-1} e^{-\lambda\theta}\right) \\ &\propto \theta^{\alpha+\sum_{i=1}^n x_i-1} e^{-(\lambda+n)\theta}, \quad \theta > 0 \end{aligned}$$

Vergleicht man den zuletzt erhaltenen Ausdruck mit der allgemeinen Gammadichte, so erkennt man, dass $\pi(\theta|\mathbf{x})$ wieder einer $\text{Gam}(\alpha^*, \lambda^*)$ -Verteilung entspricht, wobei:

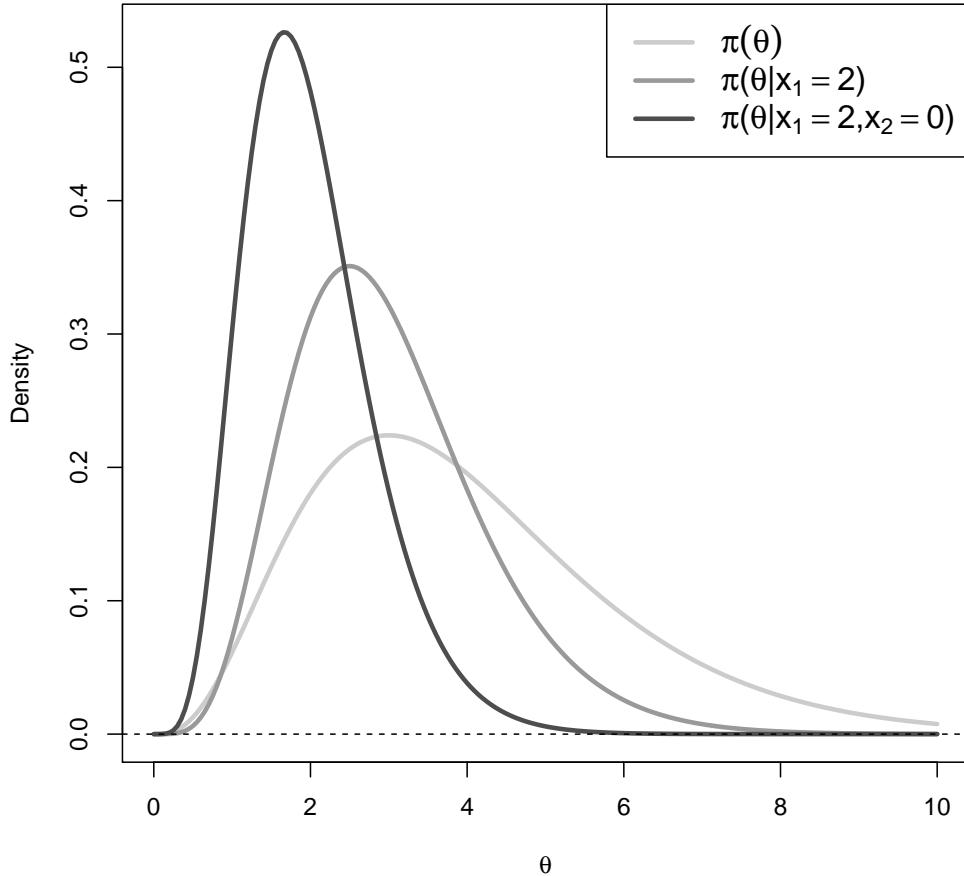
$$\alpha^* = \alpha + \sum_{i=1}^n x_i \quad \text{und} \quad \lambda^* = \lambda + n$$

Wählt man also die A-priori-Verteilung aus der Familie der Gammaverteilungen und beobachtet $\mathbf{X} = \mathbf{x}$, so ergibt sich die A-posteriori-Verteilung durch **Updating** der A-priori-Parameter:

$$\pi(\theta) \cong \text{Gam}(\alpha, \lambda) \implies \pi(\theta|\mathbf{x}) \cong \text{Gam}\left(\alpha + \sum_{i=1}^n x_i, \lambda + n\right)$$

Bsp 8.3 Die Zahl der wöchentlichen Blackouts eines Netzwerks folge einer $P(\theta)$ -Verteilung. Der Parameter θ ist nicht genau bekannt, aber aus der Vergangenheit weiß man, dass es pro Woche durchschnittlich 4 Blackouts gibt, mit einer Streuung von 2. Man findet leicht eine Verteilung aus der konjuguierten $\text{Gam}(\alpha, \lambda)$ -Familie, deren Erwartungswert gleich 4 und deren Streuung gleich 2 ist:

⁶Das Zeichen \propto bedeutet „proportional zu“.

Abbildung 8.3: A-priori-Dichte und A-posteriori-Dichten (Bsp 8.3)

$$\frac{\alpha}{\lambda} = 4, \quad \frac{\sqrt{\alpha}}{\lambda} = 2 \quad \Rightarrow \quad \alpha = 4, \quad \lambda = 1$$

Gibt es in der laufenden Woche beispielsweise $x_1 = 2$ Blackouts und wählt man als A-priori für θ eine $\text{Gam}(4, 1)$, so ist die A-posteriori eine $\text{Gam}(\alpha^*, \lambda^*)$, wobei:

$$\alpha^* = \alpha + 2 = 6, \quad \lambda^* = \lambda + 1 = 2$$

Gibt es in der nächsten Woche $x_2 = 0$ Blackouts, ergeben sich die folgenden Parameter:

$$\alpha^* = \alpha + 2 + 0 = 6, \quad \lambda^* = \lambda + 2 = 3$$

Vgl. Abb 8.3 für ein grafische Darstellung der A-priori-Dichte und der beiden A-posteriori-Dichten von θ . ■

Bernoulli–Modell: $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ sei eine Stichprobe von $X \sim \text{A}(\theta)$. Dann gilt für die durch θ bedingte W–Funktion:

$$p(\mathbf{x}|\theta) = \prod_{i=1}^n p(x_i|\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}$$

An der Form von $p(\mathbf{x}|\theta)$ sieht man, dass die Familie der Betaverteilungen $\{\text{Be}(a, b)\}$ (vgl. 4.2.7) konjugiert ist:

$$\pi(\theta) = \frac{1}{B(a, b)} \theta^{a-1} (1-\theta)^{b-1} \propto \theta^{a-1} (1-\theta)^{b-1}, \quad 0 < \theta < 1$$

Die A-posteriori–Dichte von θ ergibt sich dann wie folgt:

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto p(\mathbf{x}|\theta) \pi(\theta) \\ &\propto \left[\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \right] \left[\theta^{a-1} (1-\theta)^{b-1} \right] \\ &\propto \theta^{a+\sum_{i=1}^n x_i - 1} (1-\theta)^{b+n-\sum_{i=1}^n x_i - 1}, \quad 0 < \theta < 1 \end{aligned}$$

Vergleicht man den zuletzt erhaltenen Ausdruck mit der allgemeinen Betadichte, so erkennt man, dass $\pi(\theta|\mathbf{x})$ wieder einer $\text{Be}(a^*, b^*)$ –Verteilung entspricht, wobei:

$$a^* = a + \sum_{i=1}^n x_i \quad \text{und} \quad b^* = b + n - \sum_{i=1}^n x_i$$

Wählt man also die A-priori–Verteilung aus der Familie der Betaverteilungen und beobachtet $\mathbf{X} = \mathbf{x}$, so ergibt sich die A-posteriori–Verteilung durch **Updating** der A-priori–Parameter:

$$\pi(\theta) \stackrel{\text{def}}{=} \text{Be}(a, b) \implies \pi(\theta|\mathbf{x}) \stackrel{\text{def}}{=} \text{Be}\left(a + \sum_{i=1}^n x_i, b + n - \sum_{i=1}^n x_i\right)$$

Normalmodell: Für eine Stichprobe $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ von $X \sim \mathcal{N}(\theta, \sigma^2)$, wobei die Varianz σ^2 als *bekannt* vorausgesetzt werde, gilt:

$$f(\mathbf{x}|\theta) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - \theta)^2}{2\sigma^2}\right] \propto \exp\left[-\sum_{i=1}^n \frac{(x_i - \theta)^2}{2\sigma^2}\right], \quad \theta \in \mathbb{R}$$

In diesem Fall ist die Familie der Normalverteilungen $\{\mathcal{N}(\mu, \tau^2)\}$ konjugiert:

$$\pi(\theta) \propto \exp\left[-\frac{(\theta - \mu)^2}{2\tau^2}\right], \quad \theta \in \mathbb{R}$$

Nach längerer Rechnung zeigt sich, dass die A-posteriori-Dichte von θ gegeben ist durch:

$$\pi(\theta|\mathbf{x}) \propto \exp\left[-\frac{(\theta - \mu^*)^2}{2\tau^{*2}}\right], \quad \theta \in \mathbb{R}$$

wobei:

$$\mu^* = \frac{\mu/\tau^2 + n\bar{x}/\sigma^2}{1/\tau^2 + n/\sigma^2} \quad \text{und} \quad \tau^{*2} = \frac{1}{1/\tau^2 + n/\sigma^2}$$

Wählt man also die A-priori-Verteilung aus der Familie der Normalverteilungen und beobachtet $\mathbf{X} = \mathbf{x}$, so ergibt sich die A-posteriori-Verteilung durch **Updating** der A-priori-Parameter:

$$\pi(\theta) \stackrel{!}{=} N(\mu, \tau^2) \implies \pi(\theta|\mathbf{x}) \stackrel{!}{=} N\left(\frac{\mu/\tau^2 + n\bar{x}/\sigma^2}{1/\tau^2 + n/\sigma^2}, \frac{1}{1/\tau^2 + n/\sigma^2}\right)$$

Drei Grenzfälle:

- (1) Für $n \rightarrow \infty$ nähert sich μ^* dem Stichprobenmittelwert \bar{x} (= klassischer Schätzwert für θ) und τ^{*2} konvergiert gegen Null. In diesem Fall *dominiert* die Stichprobeninformation die A-priori-Information und die Bayes'sche Analyse nähert sich der frequentistischen Analyse.
- (2) Für $\tau \rightarrow \infty$ nähert sich μ^* ebenfalls dem Stichprobenmittelwert \bar{x} und τ^{*2} nähert sich dem Wert σ^2/n :

$$\pi(\theta|\mathbf{x}) \approx N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

Für großes τ wird die A-priori-Verteilung sehr *flach* und daher die A-priori-Information sehr vage. In diesem Fall spricht man von einer **nichtinformativen** A-priori-Verteilung. (Bem: Nichtinformative A-priori-Verteilungen spielen generell eine wichtige Rolle in der Bayes-Statistik; sie kommen zur Anwendung, wenn man sich möglichst „objektiv“ verhalten möchte, oder wenn über den interessierenden Parameter nur wenig bekannt ist.)

- (3) Für $\sigma \rightarrow \infty$ steckt sehr viel Unsicherheit in der Stichprobe. In diesem Fall bekommt die A-priori-Information ein großes Gewicht und es gilt:

$$\pi(\theta|\mathbf{x}) \approx N(\mu, \tau^2)$$

8.3 Bayes–Schätzer

Die gesamte Information über den (unbekannten) Parameter, *nach* Beobachtung von \mathbf{X} , steckt in der A-posteriori–Verteilung. Naheliegenderweise verwenden wir daher Letztere für weitere statistische Analysen. Um den Parameter θ zu schätzen, nehmen wir den **A-posteriori–Erwartungswert**:

$$\widehat{\theta}_B = \mathbb{E}(\theta|\mathbf{X} = \mathbf{x}) = \begin{cases} \sum_{\theta} \theta \pi(\theta|\mathbf{x}) = \frac{\sum_{\theta} \theta p(\mathbf{x}|\theta)\pi(\theta)}{\sum_{\theta} p(\mathbf{x}|\theta)\pi(\theta)} & \theta \text{ diskret} \\ \int_{\Theta} \theta \pi(\theta|\mathbf{x}) d\theta = \frac{\int \theta f(\mathbf{x}|\theta)\pi(\theta) d\theta}{\int f(\mathbf{x}|\theta)\pi(\theta) d\theta} & \theta \text{ stetig} \end{cases}$$

Den bedingten Erwartungswert $\widehat{\theta}_B$, gegeben die Beobachtungen $\mathbf{X} = \mathbf{x}$, nennt man den **Bayes–Schätzer** von θ .

Wie genau ist der Bayes–Schätzer? Unter allen Schätzern $\widehat{\theta}$ (von θ) hat $\widehat{\theta}_B = \mathbb{E}(\theta|\mathbf{x})$ die kleinste **A-posteriori–Varianz** (vgl. dazu auch 1.8.1):

$$\mathbb{E}[(\theta - \widehat{\theta})^2 | \mathbf{X} = \mathbf{x}]$$

Diese Varianz nennt man auch das **A-posteriori–Risiko** (bezüglich eines quadratischen Fehlers $(\theta - \widehat{\theta})^2$). Der Bayes–Schätzer minimiert auch das **Bayes–Risiko**, für stetiges X und stetiges θ gegeben durch:

$$R(\pi, \widehat{\theta}) = \int \left\{ \int (\theta - \widehat{\theta})^2 f(\mathbf{x}|\theta) d\mathbf{x} \right\} \pi(\theta) d\theta$$

Bsp 8.4 Für die Situation von Bsp 8.3, nach zwei Wochen mit zwei bzw. keinem Ausfall, ist die A-posteriori–Verteilung eine $\text{Gam}(6, 3)$ –Verteilung. Der Bayes–Schätzer von θ ist also gegeben durch:

$$\widehat{\theta}_B = \mathbb{E}(\theta|\mathbf{x}) = \frac{\alpha^*}{\lambda^*} = \frac{6}{3} = 2 \quad [\text{Blackouts/Woche}]$$

Das A-posteriori–Risiko beträgt:

$$\text{Var}(\theta|\mathbf{x}) = \frac{\alpha^*}{\lambda^{*2}} = \frac{2}{3}$$

■

Im Folgenden ein Überblick bezüglich Bayes-Schätzer und A-posteriori-Risiko für die drei in Abschnitt 8.2 betrachteten Standardmodelle:

Poisson-Modell: Der Bayes-Schätzer lautet wie folgt:

$$\hat{\theta}_B = \mathbb{E}(\theta|\mathbf{x}) = \frac{\alpha + \sum_{i=1}^n x_i}{\lambda + n} = \frac{\alpha + n\bar{x}}{\lambda + n}$$

Es ist instruktiv, den Bayes-Schätzer als gewichteten Mittelwert aus A-priori-Mittelwert ($= \alpha/\lambda$) und Stichprobenmittelwert ($= \bar{x}$) darzustellen:

$$\hat{\theta}_B = \left(\frac{\lambda}{\lambda + n} \right) \left(\frac{\alpha}{\lambda} \right) + \left(\frac{n}{\lambda + n} \right) (\bar{x})$$

Das A-posteriori-Risiko des Bayes-Schätzers ist gegeben durch:

$$\text{Var}(\theta|\mathbf{x}) = \frac{\alpha + n\bar{x}}{(\lambda + n)^2}$$

Bernoulli-Modell: Der Bayes-Schätzer lautet wie folgt:

$$\hat{\theta}_B = \mathbb{E}(\theta|\mathbf{x}) = \frac{a + \sum_{i=1}^n x_i}{a + b + n} = \frac{a + n\bar{x}}{a + b + n}$$

Darstellung als gewichteter Mittelwert aus A-priori-Mittelwert ($= a/(a+b)$) und Stichprobenmittelwert ($= \bar{x}$):

$$\hat{\theta}_B = \left(\frac{a + b}{a + b + n} \right) \left(\frac{a}{a + b} \right) + \left(\frac{n}{a + b + n} \right) (\bar{x})$$

Das A-posteriori-Risiko des Bayes-Schätzers ist gegeben durch (mit $a^* = a + n\bar{x}$ und $b^* = b + n(1 - \bar{x})$):

$$\text{Var}(\theta|\mathbf{x}) = \frac{a^* b^*}{(a^* + b^*)^2 (a^* + b^* + 1)}$$

Normalmodell: Der Bayes-Schätzer lautet wie folgt:

$$\hat{\theta}_B = \mathbb{E}(\theta|\mathbf{x}) = \frac{\mu/\tau^2 + n\bar{x}/\sigma^2}{1/\tau^2 + n/\sigma^2}$$

Darstellung als gewichteter Mittelwert aus A-priori–Mittelwert ($= \mu$) und Stichprobenmittelwert ($= \bar{x}$):

$$\hat{\theta}_B = \left(\frac{1/\tau^2}{1/\tau^2 + n/\sigma^2} \right) (\mu) + \left(\frac{n/\sigma^2}{1/\tau^2 + n/\sigma^2} \right) (\bar{x})$$

Das A-posteriori–Risiko des Bayes–Schätzers ist gegeben durch:

$$\text{Var}(\theta|\mathbf{x}) = \frac{1}{1/\tau^2 + n/\sigma^2}$$

8.4 Bayes'sche Intervallschätzer

Konfidenzintervalle haben in der Bayes–Statistik eine vollkommen andere Bedeutung als in der klassischen Statistik. Da wir eine A-posteriori–Verteilung für θ haben, müssen wir nicht mehr auf die „Auf-lange-Sicht“–Interpretation der klassischen Konfidenzintervalle zurückgreifen, sondern können von der (a-posteriori) Wahrscheinlichkeit sprechen, mit der θ von einem Intervall überdeckt wird. Derartige Aussagen sind in der frequentistischen Statistik unmöglich.

Vertrauensintervalle: Ist $C = (u(\mathbf{x}), v(\mathbf{x}))$ ein Intervall, sodass:

$$P(\theta \in C | \mathbf{X} = \mathbf{x}) = \int_C \pi(\theta|\mathbf{x}) d\theta = 1 - \alpha$$

so ist C ein Intervallschätzer für θ in dem Sinn, dass die bedingte Wahrscheinlichkeit, dass θ zu diesem Intervall gehört, gleich $1 - \alpha$ ist. Zur Unterscheidung von den klassischen Konfidenzintervallen spricht man hier von (Bayes'schen) **Vertrauensintervallen** (oder allgemeiner von **Vertrauensbereichen**) mit **Sicherheit** $1 - \alpha$.

HPD–Intervalle: Wünschenswert sind möglichst kurze Intervalle (oder möglichst kleine Bereiche); das führt zum Begriff des **HPD–Intervalls**⁷ (oder **HPD–Bereichs**). Darunter versteht man Vertrauensbereiche C der folgenden Form:

$$C = \{ \theta \mid \pi(\theta|\mathbf{x}) \geq c \}$$

In einigen Fällen können derartige Bereiche explizit bestimmt werden, meist ist man aber auf numerische Methoden angewiesen.

Bsp 8.5 [Normalmodell] Für das Normalmodell (mit bekannter Varianz) ist das $(1 - \alpha)$ –HPD–Intervall für θ aufgrund der Symmetrie der A-posteriori–Verteilung gegeben durch:

⁷engl. *highest posterior density credible set*

$$\mu^* \pm z_{1-\alpha/2} \tau^* = (\mu^* - z_{1-\alpha/2} \tau^*, \mu^* + z_{1-\alpha/2} \tau^*)$$

Für die *nichtinformative* A-priori-Verteilung (d. h. für $\tau \rightarrow \infty$ oder für $\pi(\theta) \propto c$) lautet das HPD-Intervall wie folgt:

$$\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Dieses Intervall stimmt *formal* mit dem klassischen Konfidenzintervall für θ (bei bekannter Varianz) überein (vgl. Bsp 7.13). Die Interpretation der Intervalle ist aber gänzlich verschieden. Für das HPD-Intervall gilt auch *a-posteriori* (d. h. *nach* den Beobachtungen):

$$P \left(\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \theta < \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \mid \mathbf{X} = \mathbf{x} \right) = 1 - \alpha$$

Eine analoge Aussage gilt für das klassische Intervall nur *a-priori* (d. h. *vor* den Beobachtungen); dann bezieht sie sich allerdings nicht auf θ sondern auf die stochastische Größe \bar{X}_n (θ ist im klassischen Fall eine – unbekannte – Konstante):

$$P \left(\bar{X}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \theta < \bar{X}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

■

Bsp 8.6 Für die Situation von Bsp 8.3, nach zwei Wochen mit zwei bzw. keinem Ausfall, ist die A-posteriori-Verteilung eine $\text{Gam}(6, 3)$ -Verteilung. Bezeichnet $q_{\alpha/2}$ bzw. $q_{1-\alpha/2}$ das $(\alpha/2)$ - bzw. das $(1 - \alpha/2)$ -Quantil dieser Verteilung, so ist das $(1 - \alpha)$ -Equal-Tails-Intervall für θ gegeben durch:

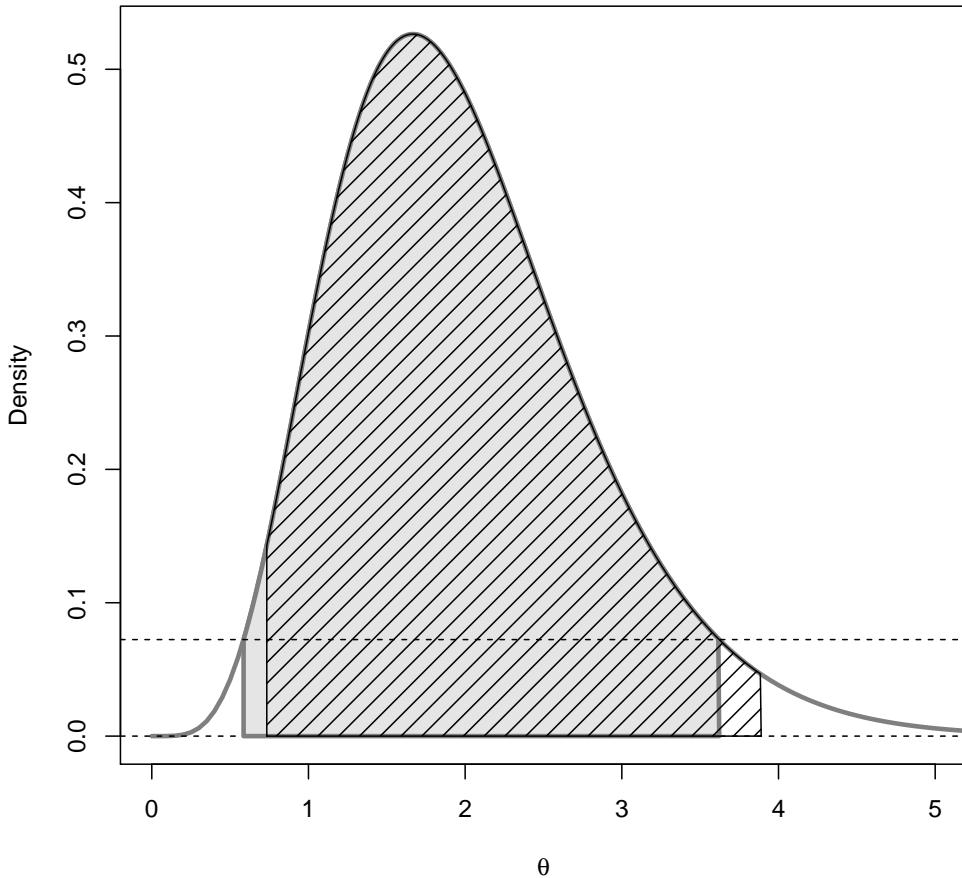
$$(q_{\alpha/2}, q_{1-\alpha/2})$$

Wegen der Schiefe der Gammaverteilung entspricht dieses Intervall nicht dem kürzesten $(1 - \alpha)$ -HPD-Intervall für θ . Letzteres muss numerisch bestimmt werden (vgl. die (eigene) Funktion `gam.hpd()`). Abb 8.4 zeigt einen Vergleich beider Intervalle für $\alpha = 5\%$ (Equal-Tails: schraffiert; HPD: grau unterlegt).

■

8.5 Bayes-Tests

Die Grundidee beim Bayes'schen Testen von (Parameter-) Hypothesen ist sehr einfach: Berechne die A-posteriori-Wahrscheinlichkeiten von \mathcal{H}_0 und \mathcal{H}_1 und wähle die wahrscheinlichere Hypothese. Derartiges ist beim klassischen Testen nicht möglich, da Parameter feste (unbekannte) Größen sind und keine Verteilung haben.

Abbildung 8.4: 95% Equal-Tails– und HPD–Intervall

Beim Bayes'schen Testen nimmt man für die Null– und Alternativhypothese meist Teilmengen $\Theta_0, \Theta_1 \subset \Theta$ des Parameterraums Θ (mit $\Theta_0 \cap \Theta_1 = \emptyset$, aber nicht notwendigerweise auch $\Theta_0 \cup \Theta_1 = \Theta$):

$$\mathcal{H}_0 : \theta \in \Theta_0 \quad \text{gegen} \quad \mathcal{H}_1 : \theta \in \Theta_1$$

Bem: Die beim klassischen Testen häufig genommenen *einfachen* Nullhypotesen der Form $\theta = \theta_0$ sind im Bayes'schen Kontext meist nicht sinnvoll, da ihr Zutreffen in der Regel unwahrscheinlich ist und in der Praxis ohnehin ein (kleines) *Intervall* um θ_0 gemeint ist. (Vgl. allerdings Bsp 8.7 für eine sinnvolle Verwendung von einfachen Hypothesen.)

Wir verwenden die A-posteriori–Verteilung um die folgenden bedingten Wahrscheinlichkeiten zu berechnen:

$$\alpha_0 = P(\theta \in \Theta_0 | \mathbf{x}) \quad \text{und} \quad \alpha_1 = P(\theta \in \Theta_1 | \mathbf{x})$$

Eine einfache Entscheidungsregel besagt nun: Akzeptiere \mathcal{H}_0 , falls $\alpha_0 \geq \alpha_1$.

Bsp 8.7 In Bsp 8.2 haben wir eigentlich ein Testproblem für den Defektanteil θ :

$$\mathcal{H}_0 : \theta = 0.05 \quad \text{gegen} \quad \mathcal{H}_1 : \theta = 0.10$$

Auf Basis des Stichprobenbefundes (3 defekte Einheiten in einer Stichprobe der Größe 20) ergaben sich die folgenden A-posteriori-Wahrscheinlichkeiten:

$$\pi(0.05|\mathbf{x}) = 0.2387 \quad \text{und} \quad \pi(0.10|\mathbf{x}) = 0.7613$$

Nach der obigen Entscheidungsregel würde man \mathcal{H}_0 verwerfen. Eine Wahrscheinlichkeit von 23.87% mag allerdings nicht klein genug sein, um die gesamte Lieferung zurückzuweisen. Jedenfalls bestehen starke Zweifel an der Behauptung des Herstellers; eine weitere Stichprobenziehung oder andere genauere Überprüfungen sind angezeigt.

Allgemein verwirft der Bayes-Test die Nullhypothese, wenn $\pi(0.05|\mathbf{x}) < 1/2$. Wie die folgende Rechnung zeigt, ist das genau dann der Fall, wenn es in der Stichprobe 3 oder mehr defekte Einheiten gibt:

```

x <- 0:20
p0 <- 0.5; p1 <- 0.5
b0 <- dbinom(x, 20, 0.05)
b1 <- dbinom(x, 20, 0.10)
post0 <- b0*p0/(b0*p0+b1*p1)
post1 <- b1*p1/(b0*p0+b1*p1)

round(data.frame(post0, post1), 4)
      post0    post1
1  0.7467  0.2533
2  0.5828  0.4172
3  0.3982  0.6018
4  0.2386  0.7614
5  0.1293  0.8707

.....
21 0.0000  1.0000

# Fehler 1. Art
sum(dbinom(3:20, 20, 0.05))
[1] 0.07548367

```

Interpretiert man den Bayes-Test als klassischen Test, entspricht das einer Fehlerwahrscheinlichkeit 1. Art von ca. 7.5%. ■

Aufgaben

8.1 Der Defektanteil θ in einem großen Los sei entweder 0.1 oder 0.2 und a-priori gelte:

$$\pi(0.1) = 0.7, \quad \pi(0.2) = 0.3$$

Wenn 8 Einheiten zufällig aus dem Los entnommen werden und davon genau 2 defekt sind, wie lautet die A-posteriori-Verteilung von θ ? Bayes-Schätzer? A-posteriori-Risiko?

8.2 Die Zahl der Bläschen auf einer Glasscheibe folge einer Poissonverteilung, deren Mittelwert θ entweder 1.0 oder 1.5 ist. Wenn a-priori gilt:

$$\pi(1.0) = 0.4, \quad \pi(1.5) = 0.6$$

und bei einer zufällig ausgewählten Glasscheibe 3 Bläschen gefunden werden, wie lautet die A-posteriori-Verteilung von θ ? Bayes-Schätzer? A-posteriori-Risiko?

8.3 Der Defektanteil θ in einem großen Los sei unbekannt. A-priori gelte:

$$(a) \pi(\theta) = I_{(0,1)}(\theta) \quad (b) \pi(\theta) = 2(1 - \theta)I_{(0,1)}(\theta)$$

Wenn von 8 zufällig ausgewählten Einheiten genau 3 defekt sind, wie lautet die A-posteriori-Verteilung? Bayes-Schätzer?

8.4 Fortsetzung von Aufgabe 8.3: Bestimmen Sie (a) 95%-Equal-Tails-Intervalle und (b) 95%-HPD-Intervalle für θ . (Hinweis zu (b): Nehmen Sie die (eigene) Funktion `beta.hpd()`.)

8.5 Die Zeit [min], die eine Person in der Früh auf den Bus warten muss, sei auf dem Intervall $(0, \theta)$ uniform verteilt, wobei $\theta > 0$ unbekannt ist. Die A-priori-Verteilung sei gegeben wie folgt:

$$\pi(\theta) = \begin{cases} \frac{192}{\theta^4} & \text{für } \theta \geq 4 \\ 0 & \text{sonst} \end{cases}$$

Wenn an drei aufeinanderfolgenden Tagen die Wartezeiten 5, 3 und 8 min betragen, bestimmen Sie (a) die A-posteriori-Verteilung, (b) den Bayes-Schätzer und (c) das 95%-HPD-Intervall für θ .

8.6 Die folgende Stichprobe stammt aus einer Poissonverteilung mit Mittelwert θ :

$$\begin{array}{cccccccccc} 11 & 7 & 11 & 6 & 5 & 9 & 14 & 10 & 9 & 5 \\ 8 & 10 & 8 & 10 & 12 & 9 & 3 & 12 & 14 & 4 \end{array}$$

Wir vermuten, dass θ etwa 12 ist, aber wir sind nicht sicher. Daher wählen wir eine $\text{Gam}(\alpha = 10, \lambda = 5/6)$ als A-priori-Verteilung für θ .

- (a) Bestimmen Sie die A-posteriori–Verteilung von θ .
- (b) Wie lautet der Bayes–Schätzer von θ ?
- (c) Bestimmen Sie das 95%–Equal-Tails/HPD–Intervall für θ .
- (d) Testen Sie die folgenden Hypothesen:

$$\mathcal{H}_0 : \theta \leq 10 \quad \text{gegen} \quad \mathcal{H}_1 : \theta > 10$$

8.7 X sei eine Beobachtung einer $G(\theta)$ –Verteilung. Wenn θ a-priori nach $U(0, 1)$ verteilt ist, bestimmen Sie:

- (a) die A-posteriori–Verteilung von θ .
- (b) die Randverteilung von X .
- (c) den Modus der A-posteriori–Verteilung.
- (d) den Bayes–Schätzer von θ .

8.8 X sei normalverteilt mit unbekanntem Mittelwert θ und bekannter Varianz $\sigma^2 = 9$. A-priori sei θ normalverteilt mit $\mu = 4$ und $\tau^2 = 1$. Eine Stichprobe des Umfangs $n = 25$ ergibt einen Stichprobenmittelwert von $\bar{x} = 4.85$.

- (a) Bestimmen Sie die A-posteriori–Verteilung von θ .
- (b) Wie lautet der Bayes–Schätzer von θ ? (ML–Schätzer?)
- (c) Wie groß ist das A-posteriori–Risiko des Bayes–Schätzers?
- (d) Bestimmen Sie das 95%–HPD–Intervall für θ .
- (e) Beantworten Sie die vorhergehenden Fragen, wenn für θ eine nichtinformative A-priori–Verteilung der Form $\pi(\theta) \propto c$ gewählt wird.

8.9 Eine Normalverteilung mit unbekanntem Mittelwert θ und bekannter Varianz $\sigma^2 = 2$ wird n Mal beobachtet. A-priori sei θ normalverteilt mit $\tau^2 = 4$. Wie groß muss n mindestens sein, sodass das A-posteriori–Risiko nicht größer als 0.01 ist?

8.10 Angenommen, $x_1 = 1.1065$, $x_2 = 0.5343$, $x_3 = 11.1438$, $x_4 = 0.4893$, $x_5 = 2.4748$ sind die beobachteten Werte einer Stichprobe von einer $Exp(\lambda)$ –Verteilung. Wenn für λ eine nichtinformative A-priori–Verteilung der Form $\pi(\lambda) \propto 1/\lambda$ gewählt wird, bestimmen Sie:

- (a) die A-posteriori–Verteilung von λ .
- (b) den Bayes–Schätzer von λ .

9 Regressionsanalyse

Das Ziel vieler wissenschaftlicher Untersuchungen besteht darin, Zusammenhänge zwischen mehreren Variablen zu erkennen und zu modellieren. Häufig interessiert die Stärke des Zusammenhangs zwischen einer **Antwortvariablen**¹ und einer oder mehreren **erklärenden Variablen**² (auch **Prädiktorvariablen** oder **Prädiktoren** genannt). Als Beispiel denke man etwa an die Beziehung zwischen Benzinverbrauch und verschiedenen Charakteristiken (Gewicht, Hubraum, Antrieb, etc.) eines Fahrzeugs.

Nur in Ausnahmefällen kennt man einen *exakten* funktionalen Zusammenhang zwischen der Antwortvariablen und den erklärenden Variablen. Beziehungen dieser Art nennt man **deterministisch**, da wiederholte Experimente unter identischen Einstellungen der erklärenden Variablen zur gleichen Antwort führen. Ein Beispiel für einen deterministischen Zusammenhang ist etwa das *Ohm'sche Gesetz* (Spannung = Widerstand \times Stromstärke).

In der überwiegenden Zahl der Fälle sind die Beziehungen zwischen den Variablen aber nicht bekannt oder zu kompliziert, als dass sie durch einige wenige erklärende Variablen beschrieben werden könnten. In solchen Fällen muss man auf – die reale Situation nur approximierende – **statistische Modelle** zurückgreifen. Die Antwortvariable ist nun eine stochastische Größe, die um einen – von den Werten der erklärenden Variablen abhängigen – Mittelwert streut.

Die **Regressionsanalyse** beschäftigt sich mit der Entwicklung von derartigen statistischen Modellen, die trotz ihrer nur approximativen Natur ein äußerst nützliches Instrument der Datenanalyse darstellen. Häufig bekommt man auf diese Weise einfache – aber in vielen Fällen dennoch „leistungsfähige“ – Modelle, die das Wesen des Zusammenhangs zwischen mehreren Variablen erfassen und beschreiben.

9.1 Einfache lineare Regression

Im einfachsten Fall hat man für eine Antwortvariable (Y) nur eine erklärende Variable (x) und die Beziehung zwischen den beiden Größen wird durch ein lineares Modell beschrieben:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Die üblichen Annahmen über die Parameter und stochastischen Größen in diesem Modell lauten wie folgt:

1. Die Größen x_i sind die beobachteten Werte der erklärenden Variablen. Bei *geplanten* Experimenten handelt es sich um feste vorgegebene Werte aus dem interessierenden Versuchsbereich.

¹engl. *response variable*

²engl. *explanatory variables*

2. Die sGn Y_i sind die zu x_i gehörigen Werte der Antwortvariablen. (Bem: Aus diesem Grund schreibt man manchmal auch Y_{x_i} anstelle von Y_i .)
3. Die Größen β_0 (Interzept) und β_1 (Anstieg) sind die Koeffizienten in der linearen Beziehung. Eine Veränderung um eine Einheit in der erklärenden Variablen x geht mit einer Veränderung um β_1 Einheiten in der Antwortvariablen einher.
4. Die sGn ε_i modellieren die Fehler, die das Streuen der Beobachtungspaare (Y_i, x_i) um die Gerade $\beta_0 + \beta_1 x_i$ bewirken. Wir nehmen an, dass diese Fehler unabhängig und normalverteilt sind, mit Mittelwert Null und konstanter Varianz σ^2 :

$$\varepsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \dots, n; \text{ ua.}$$

Die Größen ε_i subsummieren einerseits nicht berücksichtigte erklärende Variablen, andererseits aber auch Fehler, die beim Messen (oder Beobachten) von Y auftreten.

Aus den obigen Annahmen folgt, dass sich die Antwortvariable Y_i aus einem *deterministischen* Teil (**Signal**) und einem normalverteilten *zufälligen* Teil (**Noise**) zusammensetzt:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Signal}} + \underbrace{\varepsilon_i}_{\text{Noise}}$$

D. h., Y_i ist eine normalverteilte sG mit Erwartungswert:

$$\mathbb{E}(Y_i) = \mathbb{E}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i + \underbrace{\mathbb{E}(\varepsilon_i)}_{=0} = \beta_0 + \beta_1 x_i$$

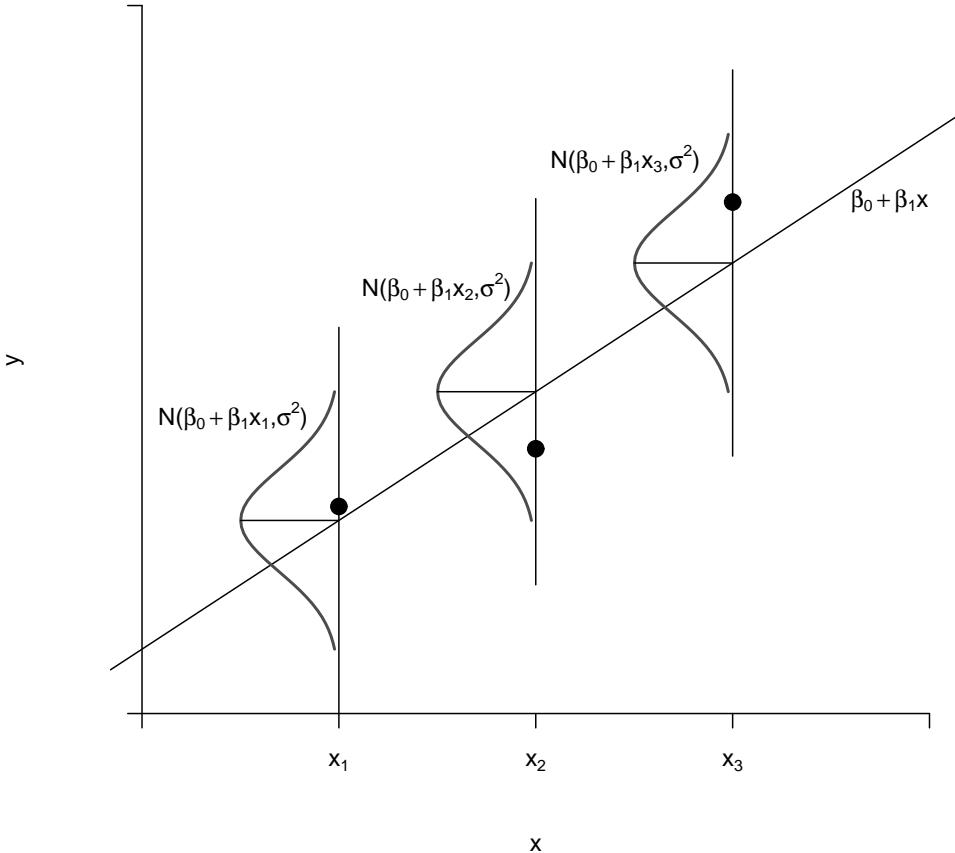
und Varianz:

$$\text{Var}(Y_i) = \text{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \underbrace{\text{Var}(\varepsilon_i)}_{=\sigma^2} = \sigma^2$$

Aus der Unabhängigkeit der ε_i folgt, dass auch die Y_i unabhängig sind. D. h., es gilt:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, 2, \dots, n; \text{ ua.}$$

Bem zur Notation: Die beobachteten Werte der erklärenden Variablen x werden mit Kleinbuchstaben bezeichnet, x_1, x_2, \dots, x_n . Damit soll zum Ausdruck gebracht werden, dass diese Werte als fest (oder gegeben) betrachtet werden und *nicht* als sGn. Die Werte der Antwortvariablen Y andererseits werden *vor* ihrer Beobachtung (oder *vor* Durchführung des Experiments) mit Großbuchstaben bezeichnet, Y_1, Y_2, \dots, Y_n . Das entspricht der üblichen Konvention zur Bezeichnung von sGn. Konkrete Beobachtungen der Antwortvariablen wiederum werden mit Kleinbuchstaben bezeichnet, y_1, y_2, \dots, y_n .

Abbildung 9.1: Beobachtungen und einfache lineare Regression

Bsp 9.1 Man betrachte die Beziehung zwischen dem Gewicht (x) eines Fahrzeugs und dem Benzinverbrauch (Y). Nun ist der Benzinverbrauch annähernd proportional zum Kraftaufwand, der notwendig ist, um das Fahrzeug zu bewegen. Kraft wiederum ist proportional zum Gewicht, sodass man davon ausgehen kann, dass der Benzinverbrauch annähernd proportional zum Gewicht ist. Es ist also sinnvoll, näherungsweise ein lineares Modell anzusetzen:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Dabei ist ε ein Fehlerterm, der verschiedene Abweichungen von einem strikten Geradenmodell $\beta_0 + \beta_1 x$ subsummiert. Einerseits ist Gewicht sicher nicht die einzige Größe, die den Benzinverbrauch bestimmt (andere Faktoren sind etwa Bauart, Motortyp, etc.). Andererseits lässt sich der Benzinverbrauch aber auch nicht ohne Fehler messen, sodass auch *Messfehler* in Rechnung zu stellen sind. Die weiteren Voraussetzungen (normalverteilte Fehler, konstante Varianz, Unabhängigkeit) sind zunächst nur Annahmen, die überprüft werden müssen. Vgl. Abb 9.1 für eine grafische Veranschaulichung des einfachen linearen Regressionsmodells. ■

9.1.1 Parameterschätzung

Die drei Parameter des einfachen linearen Modells, die Koeffizienten β_0 , β_1 und die Varianz σ^2 , sind auf Basis einer Stichprobe $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ zu schätzen. Dazu gibt es mehrere Möglichkeiten. Die am häufigsten verwendete – und mathematisch einfachste – Methode besteht darin, die Koeffizienten β_0 und β_1 so zu wählen, dass die folgende Quadratsumme minimal wird:

$$S(\beta_0, \beta_1) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Dieses **Prinzip der kleinsten Quadrate** wurde bereits in 1.9.4 aus deskriptiver Perspektive diskutiert. (Bem: Die Koeffizienten werden dort mit α und β bezeichnet.) Als Lösung ergeben sich die **KQ– (oder (O)LS–) Schätzwerte**:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Die **Prognosewerte**³ sind gegeben durch:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \dots, n$$

Die Differenzen zwischen den (tatsächlichen) Beobachtungen (y_i) und den Prognosewerten (\hat{y}_i) sind die **Residuen**:⁴

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n$$

Die Residuen unterliegen den beiden folgenden Bedingungen:

$$\sum_{i=1}^n e_i = 0 \quad \text{und} \quad \sum_{i=1}^n e_i x_i = 0$$

Eine wichtige Größe ist die **Residuenquadratsumme**:

$$\text{SSE} = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = \sum_{i=1}^n e_i^2$$

³engl. *fitted values*

⁴engl. *residuals*

Auf Basis von SSE bekommt man einen (erwartungstreuen) Schätzer für σ^2 :

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - 2} = \frac{\sum_{i=1}^n e_i^2}{n - 2}$$

Bsp 9.2 In Fortsetzung von Bsp 9.1 betrachten wir den Benzinverbrauch (in gpm⁵) von einundvierzig 2007er Modellen in Abhängigkeit vom Gewicht (in 1000 lb⁶) des Fahrzeugs. (Datensatz: `carsnew.txt`) Die R-Funktion für die Anpassung von linearen Modellen ist `lm()`. Aus dem sich ergebenden `lm`-Objekt können Detailergebnisse (Koeffizienten, Prognosewerte, Residuen, etc.) ausgelesen werden.

```

carsn <- read.table("carsnew.txt", header=TRUE)
attach(carsn)
plot(100/MPGHwy ~ CurbWeight, type="p", pch=19,
     xlab="weight", ylab="gpm (Highway)", col="grey50")
mod <- lm(100/MPGHwy ~ CurbWeight)
abline(mod, lwd=2)

coef(mod)
(Intercept)  CurbWeight
 0.8462777   0.7455568

fit <- data.frame(y=100/MPGHwy, yhat=fitted(mod), e=resid(mod))
round(fit, 4)
      y      yhat       e
 1 2.7027  2.8116 -0.1089
 2 3.5714  3.2917  0.2797
 3 3.4483  3.5623 -0.1141
 4 4.3478  4.0335  0.3143
 5 3.4483  3.1963  0.2520
 6 3.1250  3.2969 -0.1719
 7 3.5714  3.6138 -0.0424
 8 3.8462  4.0067 -0.1605
 9 2.7778  2.9010 -0.1233
10 2.9412  3.2619 -0.3207

      .....

40 4.0000  3.8375  0.1625
41 4.0000  4.1170 -0.1170
detach(carsn)

```

⁵ 1 gpm (gallons per 100 miles) = 100/mpg = 2,352 l/100km

⁶ 1 lb (pound) = 0,45359 kg

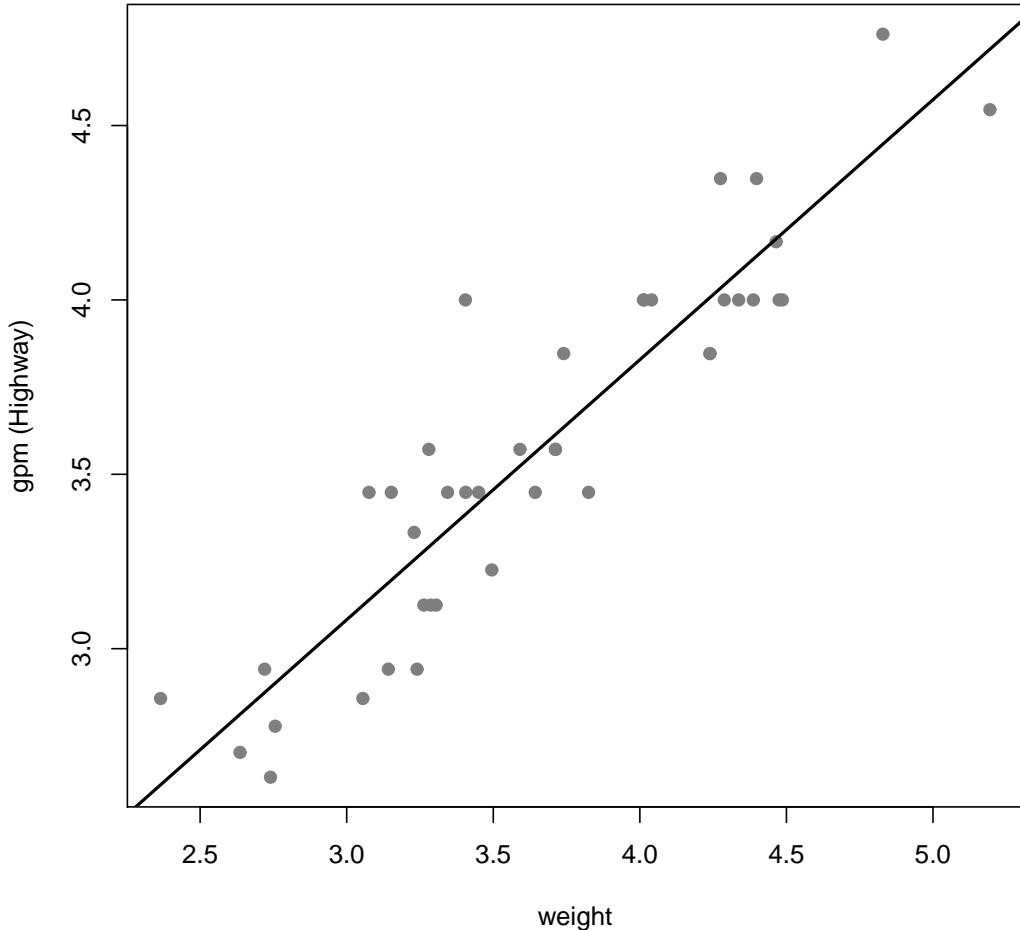
Abbildung 9.2: Benzinverbrauch in Abhängigkeit vom Fahrzeuggewicht

Abb 9.2 zeigt eine grafische Darstellung der Beobachtungspaare (x_i, y_i) sowie die angepasste KQ- (oder LS-) Gerade:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 0.846 + 0.746 x$$

Der Anstieg $\hat{\beta}_1 = 0.746$ lässt sich wie folgt interpretieren: Ein zusätzliches Gewicht von 1000 lb geht mit einer (mittleren) Erhöhung des Verbrauchs von 0.746 gpm einher.

Über den Beobachtungsbereich von etwa 2300 bis 5200 lb ist der Zusammenhang von Verbrauch und Gewicht annähernd linear; außerhalb dieses Bereichs mag die Beziehung von anderer Form sein (aber Daten dazu sind nicht verfügbar).

Man beachte auch, dass dem Interzept $\hat{\beta}_0 = 0.846$ keine allzu große Bedeutung zukommt (dient nur der Definition der Geraden); er ist jedenfalls *nicht* als Verbrauch von „Fahrzeugen“ mit Gewicht = 0 lb zu interpretieren! ■

9.1.2 Verteilung der Koeffizienten

Auf Basis der Beobachtungspaare $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ lässt sich der KQ-Schätzer des Anstiegs β_1 (= Regressionskoeffizient) wie folgt darstellen:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n c_i Y_i \quad \text{mit} \quad c_i = \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

D.h., $\hat{\beta}_1$ ist eine lineare Funktion von Y_1, Y_2, \dots, Y_n . Für die Koeffizienten c_i der Linear-kombination gelten die folgenden Aussagen:

$$\begin{aligned} \sum_{i=1}^n c_i &= \frac{1}{\sum_{j=1}^n (x_j - \bar{x})^2} \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{=0} = 0 \\ \sum_{i=1}^n c_i x_i &= \frac{1}{\sum_{j=1}^n (x_j - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) x_i = \frac{1}{\sum_{j=1}^n (x_j - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x})^2 = 1 \\ \sum_{i=1}^n c_i^2 &= \frac{1}{\left[\sum_{j=1}^n (x_j - \bar{x})^2 \right]^2} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Damit folgt, dass $\hat{\beta}_1$ ein erwartungstreuer (unverzerrter) Schätzer von β_1 ist:

$$\mathbb{E}(\hat{\beta}_1) = \sum_{i=1}^n c_i \mathbb{E}(Y_i) = \sum_{i=1}^n c_i (\beta_0 + \beta_1 x_i) = \underbrace{\beta_0 \sum_{i=1}^n c_i}_{=0} + \underbrace{\beta_1 \sum_{i=1}^n c_i x_i}_{=1} = \beta_1$$

Für die Varianz von $\hat{\beta}_1$ gilt:

$$\text{Var}(\hat{\beta}_1) = \sum_{i=1}^n c_i^2 \underbrace{\text{Var}(Y_i)}_{=\sigma^2} = \sigma^2 \sum_{i=1}^n c_i^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Die Größen Y_i sind ua. mit $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$; somit folgt nach dem Additionstheorem für Normalverteilungen (vgl. 6.2.3):

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

Analoge Überlegungen gelten für den Schätzer $\hat{\beta}_0$:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

Die Schätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ sind nicht unabhängig; ihre Kovarianz ist gegeben durch:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = -\bar{x} \text{Var}(\hat{\beta}_1)$$

Gilt $\bar{x} = 0$ (d. h., ist die erklärende Variable *zentriert*), sind $\hat{\beta}_0$ und $\hat{\beta}_1$ unkorreliert bzw. unabhängig.

Bem: Wie man leicht nachweist, sind unter den gegebenen Voraussetzungen die KQ–Schätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ auch die ML–Schätzer von β_0 bzw. β_1 . Der ML–Schätzer der Modellvarianz σ^2 ist allerdings gegeben durch:

$$\tilde{\sigma}^2 = \frac{\text{SSE}}{n} = \frac{\sum_{i=1}^n e_i^2}{n}$$

Der ML–Schätzer $\tilde{\sigma}^2 = (n-2)\hat{\sigma}^2/n$ ist also (leicht) verzerrt. In der Praxis verwendet man aber ausschließlich den (unverzerrten) Schätzer $\hat{\sigma}^2$.

9.1.3 Varianzzerlegung

Das Ziel der Regressionsanalyse besteht darin, die in der erklärenden Variablen (x) enthaltene Information dazu zu benutzen, um (zumindest einen Teil) der Variation in der

Anwortvariablen (Y) zu erklären. Ignoriert man die in x_1, x_2, \dots, x_n enthaltene Information, kann man die Variation in Y_1, Y_2, \dots, Y_n durch die folgende **totale Quadratsumme** beschreiben:

$$\text{SST} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Bis auf den Faktor $1/(n-1)$ ist SST identisch mit der (üblichen) Stichprobenvarianz S_Y^2 :

$$S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Ein Teil der Variation in Y_1, Y_2, \dots, Y_n lässt sich aber (möglicherweise) auf die verschiedenen Werte x_1, x_2, \dots, x_n der erklärenden Variablen zurückführen. In Bsp 9.2 liegt das Gewicht der Fahrzeuge zwischen etwa 2300 und 5200 lb, und diese Gewichtsunterschiede tragen sicher ihren Teil zur Variation im Benzinverbrauch bei. Die Variation der auf Basis des Regressionsmodells prognostizierten Werte $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, 2, \dots, n$, um den Mittelwert \bar{Y} , beschrieben durch:

$$\text{SSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

entspricht jenem Teil von SST, der durch das Regressionsmodell „erklärbar“ ist. Aus diesem Grund nennt man SSR auch die **Regressionsquadratsumme**. Der durch das Regressionsmodell nicht erklärbare Rest lässt sich durch die in 9.1.1 definierte **Residuenquadratsumme** (oder **Fehlerquadratsumme**) SSE beschreiben.

Mittels einfacher algebraischer Umformungen zeigt man, dass die oben intuitiv hergeleitete **Varianzzerlegung** auch formal gültig ist:

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{\text{SSE}}$$

Zum besseren Verständnis der Varianzzerlegung zwei Grenzfälle:

- (1) Gilt $\hat{\beta}_1 = 0$, so gibt es keine (lineare) Assoziation zwischen erklärender Variable und Antwortvariable. Unabhängig vom Wert von x gilt stets $\hat{Y}_i = \bar{Y}$, sodass SSR = 0 und SSE = SST. In diesem Fall trägt das Regressionsmodell *nichts* zur Erklärung der Variation in Y_1, Y_2, \dots, Y_n bei.

- (2) Verläuft die Regressionsgerade exakt durch alle Punkte (x_i, Y_i) , $i = 1, 2, \dots, n$, so gilt stets $\hat{Y}_i = Y_i$, sodass $SSE = 0$ und $SSR = SST$. In diesem Fall wird durch das Regressionsmodell die *gesamte* Variation in Y_1, Y_2, \dots, Y_n erklärt.

9.1.4 Bestimmtheitsmaß

Auf Basis der Varianzzerlegung von 9.1.3 lässt sich eine Maßzahl für die Güte der Anpassung des Regressionsmodells an die Daten definieren. Das **Bestimmtheitsmaß**⁷ ist gegeben durch:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Das Bestimmtheitsmaß misst den Anteil der Variation in Y , der durch das Regressionsmodell erklärt wird. Wegen $0 \leq SSR \leq SST$ gilt:

$$0 \leq R^2 \leq 1$$

Die Fälle $R^2 = 0$ bzw. 1 entsprechen den in 9.1.3 diskutierten Grenzfällen (1) bzw. (2).

Im Falle des einfachen linearen Regressionsmodells besteht eine direkte Beziehung zwischen R^2 und dem in 1.9.3 definierten Korrelationskoeffizienten r . Dazu bringen wir SSR in eine etwas andere Form:

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n [\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x})]^2 = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

$\hat{\beta}_1$ ist gegeben durch (vgl. 9.1.1):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Substitution in den obigen Ausdruck ergibt:

⁷engl. coefficient of determination

$$\text{SSR} = \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) \right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Damit lässt sich R^2 auch wie folgt schreiben:

$$R^2 = \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) \right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

D. h., R^2 ist das Quadrat des Korrelationskoeffizienten der Punkte (x_i, Y_i) , $i = 1, 2, \dots, n$:

$$R^2 = r_{xY}^2$$

Bsp 9.3 Für die Daten von Bsp 9.2 bekommt man für die einzelnen Quadratsummen die folgenden Werte:

$$\text{SST} = 11.1635, \quad \text{SSR} = 9.3724 \quad \text{SSE} = 1.7911$$

Damit ergibt sich das Bestimmtheitsmaß zu:

$$R^2 = \frac{9.3724}{11.1635} = 1 - \frac{1.7911}{11.1635} = 0.8396$$

D. h., etwa 84% der Variation im Benzinverbrauch wird durch das Fahrzeuggewicht erklärt. Anders ausgedrückt, das einfache lineare Regressionsmodell reduziert die Variation in der Antwortvariablen um 84%. ■

9.1.5 ANOVA-Tafel und F-Test

Die Zerlegung von SST in SSR und SSE wird meist in Form einer sog. **Varianzanalysetafel**⁸ dargestellt. Dabei spielen die den Quadratsummen zugeordneten **Freiheitsgrade**⁹ eine zentrale Rolle. Da wegen $\sum(Y_i - \bar{Y}) = 0$ für die Berechnung von $\text{SST} = \sum(Y_i - \bar{Y})^2$ nur $n - 1$ Komponenten benötigt werden, sind der totalen Quadratsumme $n - 1$ Freiheitsgrade zugeordnet. Wegen $\sum e_i = \sum e_i x_i = 0$ sind der Fehlerquadratsumme $\text{SSE} = \sum e_i^2$

⁸engl. *analysis-of-variance (ANOVA) table*

⁹engl. *degrees of freedom* (abgekürzt df)

$n - 2$ Freiheitsgrade zugeordnet, und da es im einfachen linearen Modell nur eine Prädiktorvariable gibt, hat die Regressionsquadratsumme nur 1 Freiheitsgrad.

ANOVA–Tafel für das einfache lineare Regressionsmodell

	df	SS	MS	F
Regression	1	SSR	MSR = SSR/1	MSR/MSE
Fehler	$n - 2$	SSE	MSE = SSR/($n - 2$)	
Total	$n - 1$	SST		

Die MS–Spalte beinhaltet die **mittleren Quadratsummen**, d. h. die Quadratsummen geteilt durch die Freiheitsgrade. In der F–Spalte steht eine Teststatistik für einen Test der Hypothesen:

$$\mathcal{H}_0 : \beta_1 = 0 \quad \text{gegen} \quad \mathcal{H}_1 : \beta_1 \neq 0$$

Wie man zeigen kann, gilt unter \mathcal{H}_0 :¹⁰

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F(1, n - 2)$$

D. h., \mathcal{H}_0 wird zum Niveau α verworfen, falls:

$$F > F_{1, n-2; 1-\alpha}$$

Den p –Wert berechnet man wie folgt:

$$p\text{--Wert} = P(F(1, n - 2) \geq F)$$

Bemerkungen zur Form der Teststatistik: In 9.1.4 haben wir gezeigt, dass SSR wie folgt dargestellt werden kann:

$$\text{SSR} = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

Unter \mathcal{H}_0 (d. h. für $\beta_1 = 0$) gilt (vgl. 9.1.2):

¹⁰Vgl. HOGG ET AL.(2005) oder GURKER (2015).

$$\hat{\beta}_1 \sim N\left(0, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

Daraus folgt (vgl. 4.2.4):

$$\frac{SSR}{\sigma^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} \sim \chi^2(1) \implies \mathbb{E}\left(\frac{SSR}{\sigma^2}\right) = 1$$

D.h., unter \mathcal{H}_0 ist SSR ein unverzerrter Schätzer von σ^2 . Nun haben wir festgestellt, dass $\hat{\sigma}^2 = SSE/(n - 2)$ in jedem Fall (d.h., auch wenn \mathcal{H}_0 nicht zutrifft) ein unverzerrter Schätzer von σ^2 ist. Unter \mathcal{H}_0 ist die F -Statistik also ein Quotient aus zwei unverzerrten Schätzern von σ^2 :

$$F = \frac{MSR}{MSE} = \frac{SSR/(1)}{SSE/(n - 2)} \approx 1$$

Trifft \mathcal{H}_0 nicht zu (d.h., gilt $\beta_1 \neq 0$) ist der Zähler von F tendenziell größer als σ^2 und $F = MSR/MSE > 1$.

Bsp 9.4 Für die carsnew-Daten (Bsp 9.2) ergibt sich die ANOVA-Tafel wie folgt:

```
anova(mod)
Analysis of Variance Table

Response: 100/MPGHwy
          Df Sum Sq Mean Sq F value    Pr(>F)
CurbWeight   1 9.3724  9.3724 204.07 < 2.2e-16 ***
Residuals   39 1.7911  0.0459
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Die mittlere Fehlerquadratsumme von $MSE = 0.0459$ ist der Schätzwert von σ^2 . Der Wert der F -Statistik (204.07) ist sehr viel größer als $F_{1,39;0.99} = 7.33$, sodass wir mit hoher Sicherheit davon ausgehen können, dass $\beta_1 \neq 0$. Das zeigt sich auch am kleinen p -Wert von nahezu Null. ■

9.1.6 Konfidenzintervalle und t–Tests

Unter den gegebenen Voraussetzungen können auf Basis der t–Verteilung einfache Konfidenzintervalle und Tests für die beiden Koeffizienten β_0 und β_1 hergeleitet werden (vgl. dazu auch 7.3.3). Im Folgenden betrachten wir in erster Linie Intervalle und Tests für den meist im Mittelpunkt des Interesses stehenden Koeffizienten β_1 . In 9.1.2 haben wir gezeigt, dass:

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \implies \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2}} \sim N(0, 1)$$

Ersetzt man das unbekannte σ^2 durch $\widehat{\sigma}^2 = \text{MSE}$, so gilt:

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\text{MSE} / \sum_{i=1}^n (x_i - \bar{x})^2}} \sim t(n-2)$$

Den Ausdruck im Nenner von T nennt man auch den (geschätzten) **Standardfehler** des Schätzers $\widehat{\beta}_1$:

$$s(\widehat{\beta}_1) := \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Konfidenzintervall für β_1 : T ist eine Pivotgröße und ein $(1 - \alpha)$ –Konfidenzintervall für β_1 ist gegeben durch:

$$\widehat{\beta}_1 \pm t_{n-2; 1-\alpha/2} s(\widehat{\beta}_1) = \widehat{\beta}_1 \pm t_{n-2; 1-\alpha/2} \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

t–Test für β_1 : Ein Test zum Niveau α für die Hypothesen:

$$\mathcal{H}_0 : \beta_1 = \beta_{10} \quad \text{gegen} \quad \mathcal{H}_1 : \beta_1 \neq \beta_{10}$$

ist gegeben durch:

$$\text{Verwerfe } \mathcal{H}_0, \text{ falls: } \frac{|\hat{\beta}_1 - \beta_{10}|}{s(\hat{\beta}_1)} > t_{n-2; 1-\alpha/2}$$

Häufig interessiert man sich für die folgenden Hypothesen:

$$\mathcal{H}_0 : \beta_1 = 0 \quad \text{gegen} \quad \mathcal{H}_1 : \beta_1 \neq 0$$

Ein Test zum Niveau α lautet wie folgt:

$$\text{Verwerfe } \mathcal{H}_0, \text{ falls: } |t(\hat{\beta}_1)| = \frac{|\hat{\beta}_1|}{s(\hat{\beta}_1)} > t_{n-2; 1-\alpha/2}$$

Letzteren Test nennt man kurz den **t–Test** für β_1 . (Er gehört auch zum Standardoutput von R; vgl. dazu das folgende Beispiel.) Der p –Wert ist gegeben durch:

$$p\text{–Wert} = 2P(t(n-2) \geq |t(\hat{\beta}_1)|)$$

Bem: Wir sind diesem Testproblem bereits im Zuge der ANOVA–Tafel von 9.1.5 begegnet, haben dort aber einen **F–Test** dafür angegeben. Beide Tests sind aber äquivalent, denn:

$$[t(\hat{\beta}_1)]^2 = \frac{\hat{\beta}_1^2}{[s(\hat{\beta}_1)]^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\text{MSE}} = \frac{\text{MSR}}{\text{MSE}} = F$$

Überdies gilt (vgl. dazu auch 4.2.6):

$$(t_{n-2; 1-\alpha/2})^2 = F_{1, n-2; 1-\alpha}$$

Bsp 9.5 Der folgende R–Output zeigt für die **carsnew**–Daten (Bsp 9.2) die Schätzwerte, die (geschätzten) Standardfehler und die t–Tests für β_0 und β_1 . Beide p –Werte sind nahezu Null, die Tests daher hoch signifikant.

Darunter steht der Wert von $\hat{\sigma} = \sqrt{\text{MSE}}$, das Bestimmtheitsmaß R^2 und die uns schon von der ANOVA–Tafel bekannte F–Statistik. (Bem: Die Bedeutung des *adjusted R*² wird in 9.2.2 diskutiert.)

```

summary(mod)

.....
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.84628   0.19416   4.359 9.24e-05 ***
CurbWeight  0.74556   0.05219  14.285 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2143 on 39 degrees of freedom
Multiple R-squared:  0.8396,    Adjusted R-squared:  0.8354
F-statistic: 204.1 on 1 and 39 DF,  p-value: < 2.2e-16

```

■

Konfidenzintervall für $\mathbb{E}(Y_0)$: Sind $\hat{\beta}_0$ und $\hat{\beta}_1$ die unverzerrten Schätzer für β_0 und β_1 , so ist ein unverzerrter Schätzer für die mittlere Antwort $\mathbb{E}(Y_0) = \beta_0 + \beta_1 x_0$ an der (neuen) Stelle $x = x_0$ gegeben durch:

$$\tilde{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Unter Verwendung der in 9.1.2 angegebenen Ausdrücke für die Varianzen und die Kovarianz von $\hat{\beta}_0$ und $\hat{\beta}_1$, ergibt sich nach einfacher Rechnung:

$$\begin{aligned} \text{Var}(\tilde{Y}_0) &= \text{Var}(\hat{\beta}_0) + x_0^2 \text{Var}(\hat{\beta}_1) + 2x_0 \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \end{aligned}$$

(Man beachte, dass diese Varianz für die Stelle $x_0 = \bar{x}$, also im Zentrum der Daten, am kleinsten ist.) Ersetzt man die (unbekannte) Varianz σ^2 durch MSE, so ergibt sich ein $(1 - \alpha)$ -Konfidenzintervall für $\mathbb{E}(Y_0)$ wie folgt:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2; 1-\alpha/2} \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

Prognoseintervall für Y_0 : Sind $\hat{\beta}_0$ und $\hat{\beta}_1$ die unverzerrten Schätzer für β_0 und β_1 , so ist ein Prognosewert für die Antwort $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ an der (neuen) Stelle $x = x_0$ gegeben durch:

$$\tilde{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

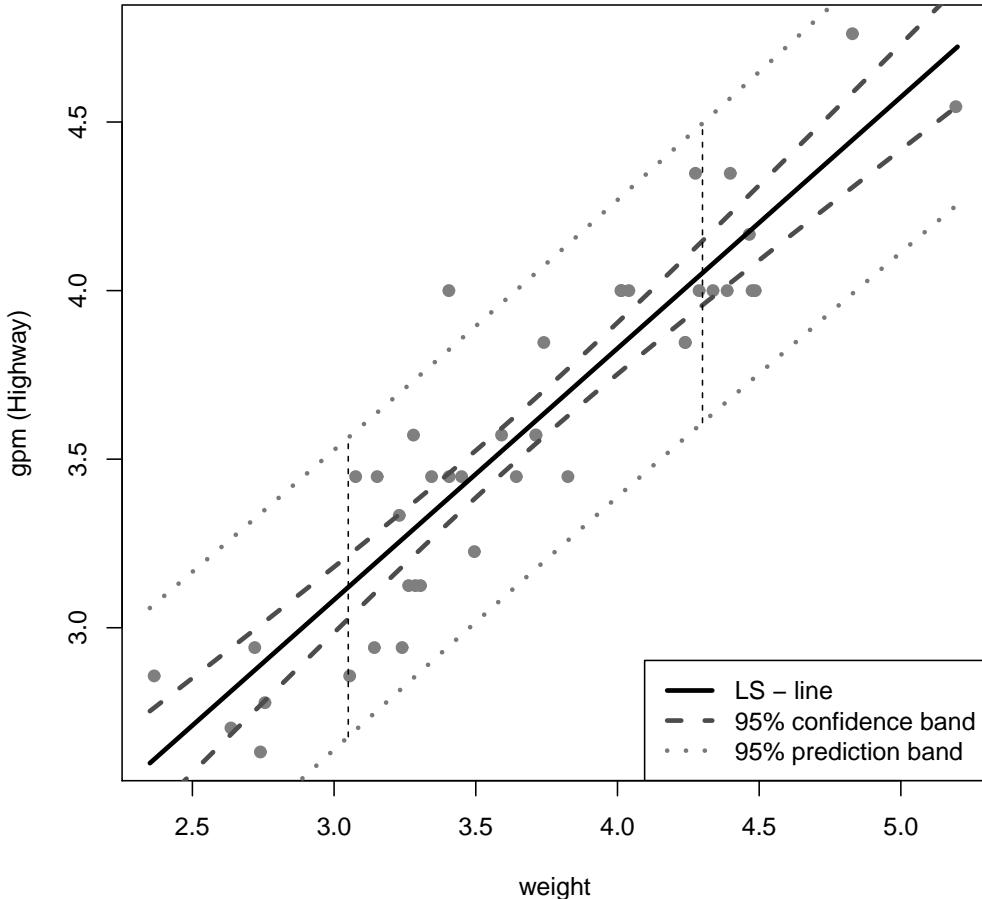
(Man beachte, dass hier die *Prognose* der sG Y_0 von Interesse ist, und nicht die *Schätzung* von $\mathbb{E}(Y_0)$.) Ein $(1 - \alpha)$ -Prognoseintervall für Y_0 ist gegeben wie folgt:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2; 1-\alpha/2} \sqrt{\text{MSE} \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

Bsp 9.6 Mit Hilfe der folgenden R-Commands werden auf einem gleichmäßigen Gitter von Punkten (erzeugt mit `pretty()`) aus dem Beobachtungsbereich der erklärenden Variablen `CurbWeight` jeweils 95%-Konfidenz- bzw. Prognoseintervalle für $\mathbb{E}(Y_x)$ bzw. Y_x bestimmt und grafisch dargestellt (Abb 9.3).

```
attach(carsn)
plot(100/MPGHwy ~ CurbWeight, type="p", pch=19,
     xlab="weight", ylab="gpm (Highway)", col="grey50")
mod <- lm(100/MPGHwy ~ CurbWeight)
x <- CurbWeight
CuWe.new <- data.frame(CurbWeight=pretty(range(x), 100))
pred.c <- predict(mod, CuWe.new, interval="confidence")
pred.p <- predict(mod, CuWe.new, interval="prediction")
matplot(CuWe.new, cbind(pred.c, pred.p[, -1]), type="l",
        lty=c(1, 2, 2, 3, 3), lwd=3, col=c(1, 2, 2, 3, 3), add=TRUE)
legend("bottomright", c("LS - line", "95% confidence band",
                        "95% prediction band"), lty=1:3, lwd=3, col=1:3)
detach(carsn)
```

Werden die Endpunkte der Intervalle durch Geradenstücke verbunden, ergeben sich Konfidenz- bzw. Prognosebänder. (Man beachte allerdings, dass diese Bänder – wie in Abb 9.3 für zwei Stellen angedeutet – *punktweise* für jedes einzelne x und nicht „*simultan*“ auf dem Beobachtungsbereich zu verstehen sind.) ■

Abbildung 9.3: 95%-Konfidenz– und Prognosebänder

9.1.7 Residualanalyse

Ein einfaches lineares Modell (d. h. ein Geradenmodell) sollte nur dann angepasst werden, wenn ein (grob) linearer Zusammenhang zwischen x und Y vorliegt. Gibt es im Scatterplot erkennbare quadratische oder andere nichtlineare Patterns, muss das einfache Modell modifiziert werden. Ebenso sollten auch die anderen Voraussetzungen – zumindest annähernd – erfüllt sein, d. h. eine über dem Beobachtungsbereich konstante Varianz und unabhängige (unkorrelierte) Fehler. Letzterer Punkt ist möglicherweise dann verletzt, wenn die Beobachtungen (x_i, Y_i) , $i = 1, 2, \dots, n$, zeitlich (oder räumlich) hintereinander erhoben werden und i die zeitliche (oder räumliche) Ordnung repräsentiert.

Zur Überprüfung der Modellvoraussetzungen verwendet man hauptsächlich auf den Residuen $e_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, n$, basierende grafische Methoden. Insbesondere zeichnet man die folgenden **Residualplots**:

- (1) Einen Plot der Residuen e_i gegen die Prognosewerte \hat{y}_i .
- (2) Einen Plot der Residuen e_i gegen x_i .

- (3) Plots der Residuen e_i gegen andere erklärende Variablen, die *nicht* im ursprünglichen Modell enthalten sind (z. B. gegen die Zeit oder gegen die Ordnung, wenn die Daten sequenziell erhoben wurden).
- (4) Einen Plot der Residuen e_i gegen die Lag-1-Residuen e_{i-1} , wenn die Daten sequenziell erhoben wurden.

Sind die Voraussetzungen erfüllt, sollten sich in den Plots keine Patterns zeigen und die Residuen sollten großteils innerhalb eines horizontalen 2σ -Bandes um Null liegen. (Als Schätzwerte von unabhängigen $N(0, \sigma^2)$ -Fehlern sollten etwa 95% der Residuen innerhalb von $\pm 2\sqrt{\text{MSE}}$ liegen.) Betrachtet man *standardisierte* Residuen der Form $s_i = e_i / \sqrt{\text{MSE}}$, so sollten sie großteils innerhalb von ± 2 um Null liegen.

Bem: Die in R mittels `rstandard()` bestimmten Residuen werden aber wie folgt berechnet:

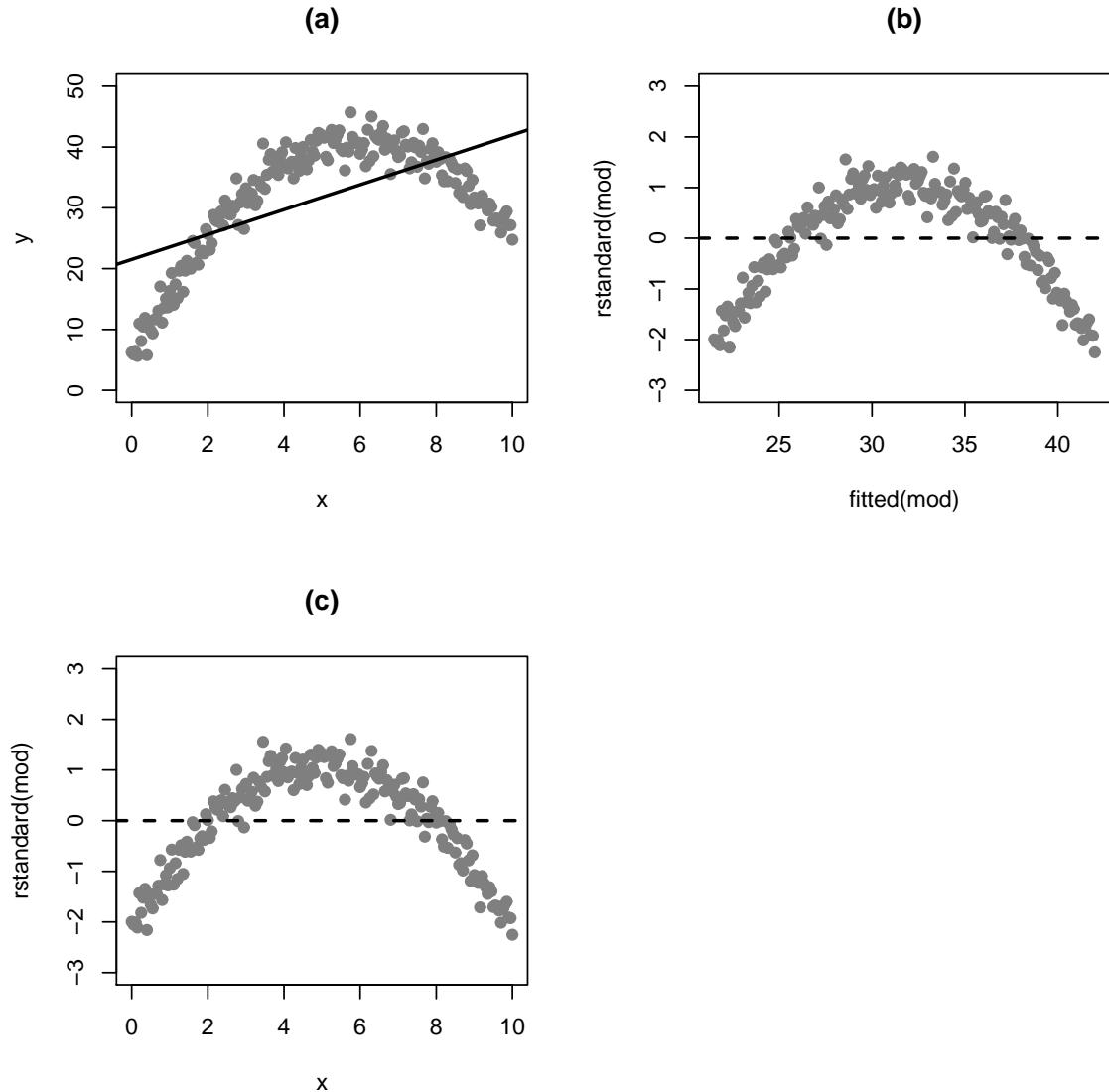
$$r_i = \frac{e_i}{\sqrt{\text{MSE}} \sqrt{1 - h_i}}$$

Man kann nämlich zeigen, dass $\text{Var}(e_i) = \sigma^2(1 - h_i)$, wobei h_i die sog. *Hatwerte* sind. (Letztere werden mittels `hatvalues()` bestimmt.)

Bsp 9.7 Zur Illustration betrachten wir zwei Fälle, bei denen einige der Modellvoraussetzungen nicht erfüllt sind. Im 1. Fall wird die Regressionsfunktion (d. h. $\mathbb{E}(Y|x)$) falsch spezifiziert. Der Zusammenhang ist quadratischer Natur, wir passen aber ein einfaches lineares Modell an. (Vgl. Abb 9.4.) Im 2. Fall ist die Varianz nicht konstant, sondern wächst mit der Größe von Y . (Vgl. Abb 9.5.) In beiden Fällen zeigen sich charakteristische Patterns in den Residualplots, die leicht zu deuten sind.

Die Feststellung allein, dass es Probleme mit dem Regressionsmodell gibt, ist allerdings zu wenig. Man muss auch versuchen, die Modelldefizite zu beheben. Im 1. Fall würde man im nächsten Schritt ein Modell der Form $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ anpassen (vgl. dazu den folgenden Abschnitt 9.2). Im 2. Fall würde man nach einer *varianzstabilisierenden* Transformation von Y suchen. Wächst die *Streuung* von Y proportional zu Y , ist die logarithmische Transformation ($\ln Y$) geeignet. Wächst die *Varianz* von Y proportional zu Y , sollte man die Wurzeltransformation (\sqrt{Y}) versuchen. In der Praxis versucht man mehrere Transformationen ($1/Y$, $\ln Y$, \sqrt{Y} , ...) und nimmt diejenige, die die beste varianzstabilisierende Wirkung zeigt. (Man beachte allerdings, dass mit jeder Transformation von Y auch die Fehlerverteilung verändert wird!) ■

Bsp 9.8 Für die `carsnew`-Daten von Bsp 9.2 sind die Residualplots in Abb 9.6 dargestellt. Es zeigen sich keine besonderen Auffälligkeiten, sodass das einfache lineare Modell als durchaus adäquat betrachtet werden kann. Lediglich ein Punkt (Nr. 27 $\hat{=}$ Mercedes C230 Sport Sedan) sticht etwas hervor. (Vgl. 9.1.8 für eine ausführlichere Diskussion derartiger Punkte.) Man beachte allerdings, dass man durch Hinzunahme von weiteren erklärenden Variablen (wie etwa Anzahl der Zylinder oder Hubraum) die „Erklärungskraft“ des Modells deutlich erhöhen könnte. ■

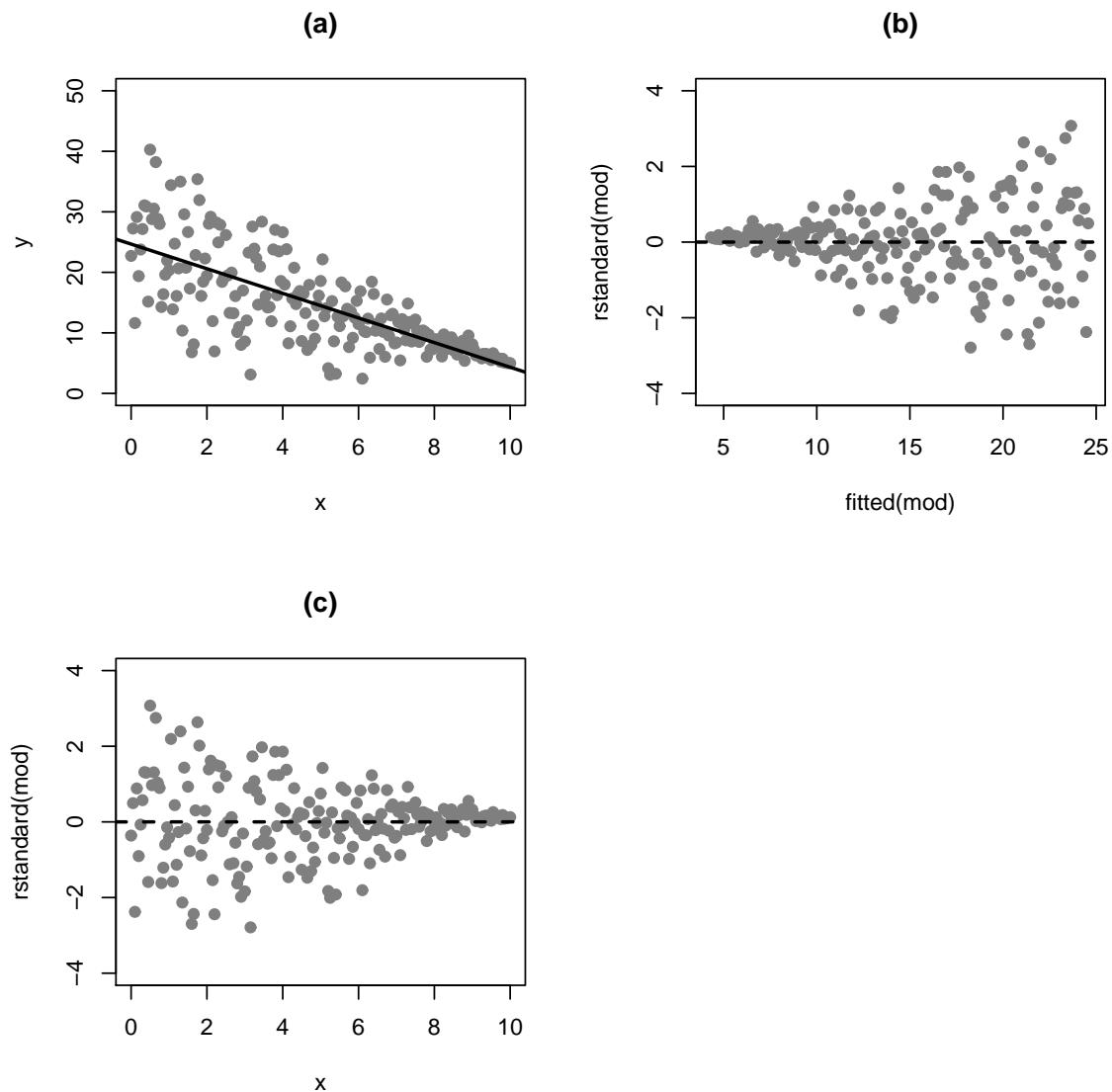
Abbildung 9.4: Scatterplot und Residualplots für Fall 1 (Bsp 9.7)

9.1.8 Ausreißer und Heelpunkte

Ein Scatterplot der Datenpunkte ist ein unabdingbarer erster Schritt in jeder Regressionsanalyse. Dadurch lassen sich bereits im Vorfeld grobe Irrtümer vermeiden. Man betrachte dazu etwa den aus vier Unterdatensätzen gleicher Größe bestehenden Datensatz `anscombe.txt`.¹¹ Für alle vier Datensätze stimmen die KQ-Schätzwerte, die Standardfehler, die ANOVA-Tafel und R^2 exakt überein, aber nur im Fall oben/rechts ist ein einfaches lineares Modell adäquat (Abb 9.7). Der Scatterplot oben/rechts zeigt einen quadratischen Zusammenhang; die Plots in der unteren Reihe zeigen die Auswirkungen von ungewöhnlichen einzelnen Datenpunkten. Im Plot unten-links liegen 10 Punkte exakt auf einer Geraden, aber ein Punkt weicht davon ab. Punkte mit ungewöhnlichen Y -Werten nennt

¹¹F. J. ANSCOMBE: Graphs in Statistical Analysis, *The American Statistician*, Vol. 27, 1973.

Abbildung 9.5: Scatterplot und Residualplots für Fall 2 (Bsp 9.7)

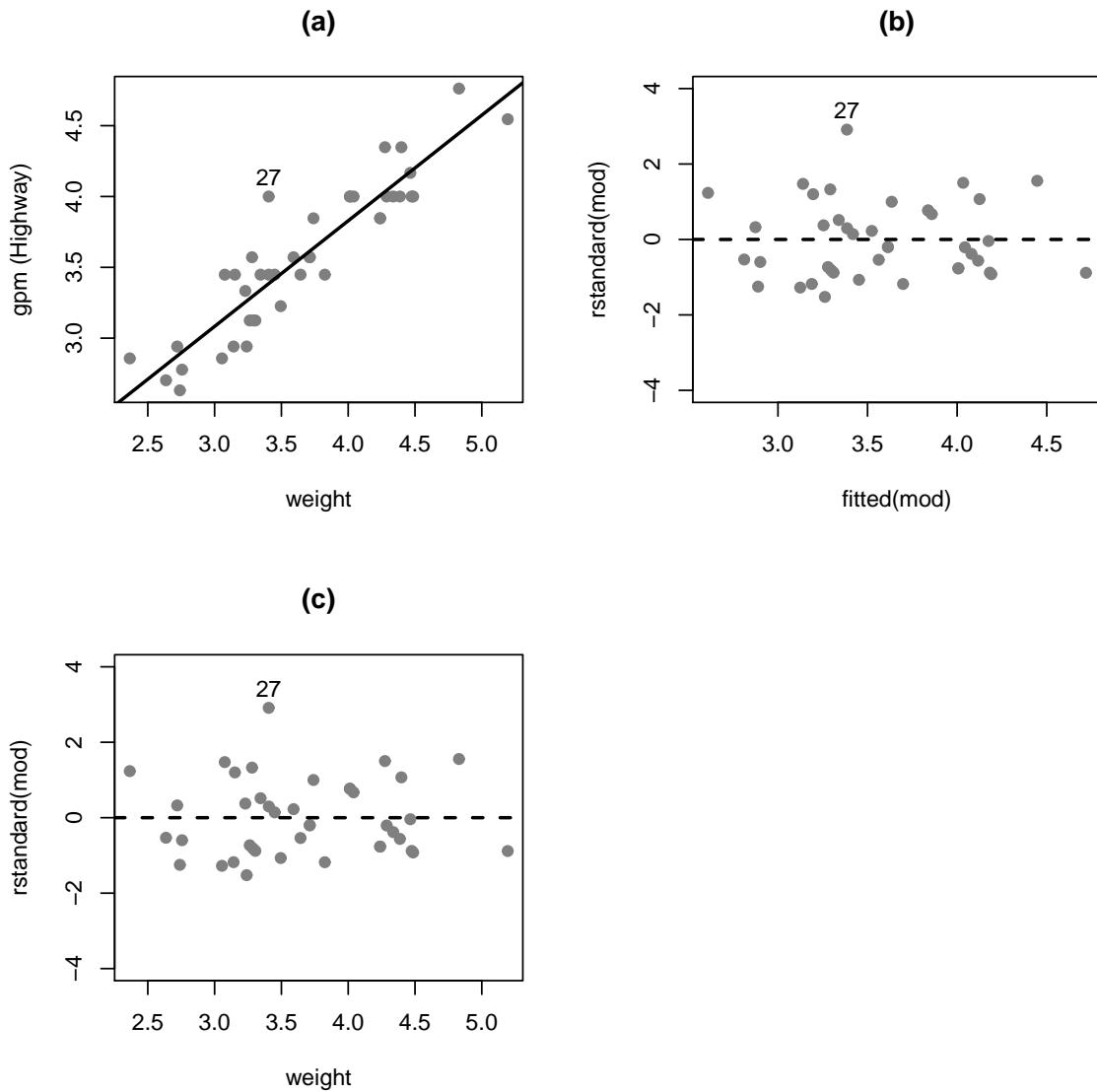


man **Ausreißer**.¹² Ausreißer verändern (mehr oder weniger stark) den Anstieg der KQ-Geraden. Der Plot unten/rechts zeigt den extremen Fall, wie ein einzelner Punkt mit ungewöhnlichem x -Wert die KQ-Gerade komplett an sich zieht. Derartige Punkte – auch mit weniger drastischen Auswirkungen – nennt man **Heelpunkte**.¹³

Was soll man mit ungewöhnlichen Datenpunkten tun? Das lässt sich nicht einfach generell beantworten. Kann man die Punkte auf bestimmte (außerstatistische) Ursachen zurückführen (Schreib-, Ablesefehler, fehlerhaftes Messinstrument, ...), sollten diese Werte klarerweise korrigiert, oder – falls das nicht möglich ist – nicht für die Modellanpassung verwendet werden.

¹²engl. *outlier*

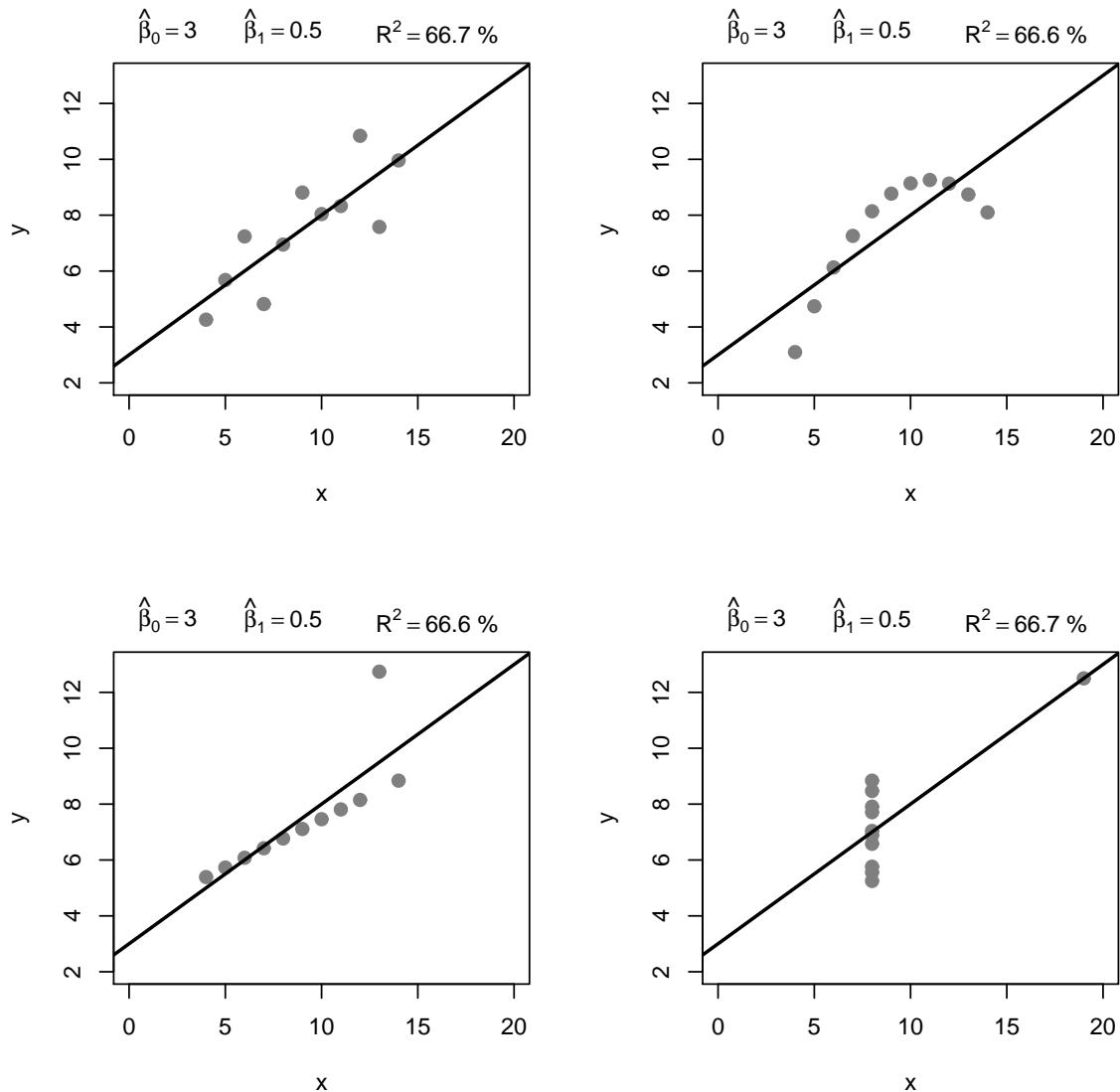
¹³engl. *leverage points*

Abbildung 9.6: Scatterplot und Residualplots für die carsnew–Daten

Häufig zeigt sich jedoch, dass bestimmte Beobachtungen zwar ungewöhnlich sind aber ansonsten korrekt erhoben wurden. Das Weglassen derartiger Punkte (um die Anpassung zu „verbessern“) ist problematisch, da auf diese Weise ein falscher Eindruck von der Präzision der Anpassung entstehen kann. Nicht selten sind die „Ausreißer“ die interessantesten Punkte des Datensatzes, weil sie die Aufmerksamkeit auf entscheidende Regionen des Modells lenken, Modelldefizite aufzeigen, oder auf andere Weise für die Untersuchung von Bedeutung sind. Eine genauere Analyse dieser Punkte (und der Bedingungen, unter denen sie erhoben wurden) ist unumgänglich und führt nicht selten zur Identifizierung von bisher nicht beachteten Prädiktoren.

Ungewöhnliche Punkte einfach wegzulassen ist keine gute Strategie. Besser ist es, die Auswirkungen zu analysieren, indem man die Anpassung einmal mit und einmal ohne diese Punkte durchführt. (Untersuchen Sie als UE–Aufgabe, welchen Einfluss der als Ausreißer

Abbildung 9.7: Die Anscombe–Datensätze



identifizierte Datenpunkt Nr. 27 im Datensatz `carsnew.txt` auf die KQ–Gerade hat.¹⁴⁾ Klärerweise möchte man nicht mit einem Modell arbeiten, das stark von einigen wenigen Datenpunkten abhängt. Falls möglich, wird man in so einem Fall danach trachten, an den kritischen Regionen zusätzliche Daten zu erheben und so zu einem stabileren Modell zu kommen.

Als Alternative zu der gegenüber Ausreißern sehr empfindlichen KQ–Methode gibt es mittlerweile eine ganze Reihe von *robusten* Schätzmethoden.¹⁵⁾ Allerdings sollte man auch bei ihrer Anwendung nicht auf eine genaue Erforschung der Ursachen für die Instabilität verzichten.

¹⁴⁾ Antwort: Praktisch keinen.

¹⁵⁾ Vgl. GURKER (2015).

9.1.9 Matrixschreibweise

In Matrixform können die meisten (praktischen und theoretischen) Berechnungen einfacher und übersichtlicher ausgeführt werden. Das gilt insbesondere für die multiple Regression, aber auch schon bei nur *einer* erklärenden Variablen ergeben sich dadurch Vorteile. Mit den Bezeichnungen:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

lässt sich das einfache lineare Regressionsmodell wie folgt schreiben:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Die **Normalgleichungen** sind gegeben durch:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$$

Dabei ist $\mathbf{X}'\mathbf{X}$ eine (2×2) -Matrix der Gestalt:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

Ist diese Matrix invertierbar, so gilt:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

Der **KQ-Schätzer** ist gegeben durch:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Bem: $\mathbf{X}'\mathbf{X}$ ist genau dann singulär, wenn alle x_i identisch sind, d. h. $x_1 = x_2 = \dots = x_n$; das bedeutet nicht, dass die Normalgleichungen nicht lösbar sind, sondern nur, dass sie nicht *eindeutig* lösbar sind. Anders ausgedrückt, β ist in diesem Fall nicht *identifizierbar*.

Für die **Prognosewerte** gilt:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Der **Residuenvektor** lässt sich wie folgt schreiben:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Dabei ist \mathbf{I} die $(n \times n)$ -Einheitsmatrix und \mathbf{H} die sog. *Hatmatrix*:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Bem: Die Bezeichnung kommt daher, dass der Prognosevektor $\hat{\mathbf{Y}}$ auch wie folgt geschrieben werden kann: $\hat{\mathbf{Y}} = \mathbf{HY}$. Die Diagonalelemente von \mathbf{H} nennt man die *Hatwerte* (vgl. dazu auch die Bem in 9.1.7).

Sind die Fehler ε_i unabhängig (unkorreliert) mit gleicher Varianz σ^2 , so ist die Varianz-Kovarianzmatrix von $\hat{\beta}$ gegeben durch:

$$\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Beweis: Unter Verwendung der Rechenregeln für die Varianz-Kovarianzmatrix (vgl. 5.4.1) gilt:

$$\text{Cov}(\hat{\beta}) = \text{Cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Letzteres gilt wegen:

$$\text{Cov}(\mathbf{Y}) = \text{Cov}(\varepsilon) = \sigma^2\mathbf{I}$$

Der unverzerrte Schätzer für die Varianz σ^2 ist gegeben durch:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-2}$$

9.2 Multiple lineare Regression

Bisher gab es nur *eine* erklärende Variable (x), die in einer linearen Beziehung zur Antwortvariablen (Y) stand. Derartige Modelle sind häufig gute Approximationen für kompliziertere funktionale Beziehungen, insbesondere über nicht zu großen x -Bereichen. Vielfach sind die Beziehungen aber nichtlinearer Natur, etwa wenn x und Y in einer **quadratischen** Beziehung stehen:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Oder allgemeiner in einer **polynomialen** Beziehung:

$$Y_i = \beta_0 + \beta_1 x_i + \dots + \beta_k x_i^k + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Modelle dieser Art sind zwar nichtlinear in x , aber immer noch linear in den Koeffizienten $\beta_0, \beta_1, \dots, \beta_k$; aus diesem Grund spricht man nach wie vor von **linearen** Regressionsmodellen. Neben polynomialen gibt es auch andere nichtlineare Beziehungen zwischen x und Y , beispielsweise Beziehungen der folgenden Art:

$$Y_i = (\beta_0 + \beta_1 x_i) e^{\beta_2 x_i} + \varepsilon_i, \quad Y_i = \frac{\beta_0 x_i}{1 + \beta_1 x_i} + \varepsilon_i, \quad \dots$$

Modelle dieser Art sind nicht nur nichtlinear in x sondern auch in den Koeffizienten. Aus diesem Grund spricht man von **nichtlinearen** Regressionsmodellen.

Die obigen Modelle beinhalten nach wie vor nur *eine* erklärende Variable. Vielfach gibt es aber mehrere Größen, die eine Antwortvariable beeinflussen. Betrachten wir zunächst den Fall zweier Einflussgrößen x_1 und x_2 und ein lineares Modell der Form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Die Fehler $\{\varepsilon_i\}$ seien unabhängige, nach $N(0, \sigma^2)$ verteilte, stochastische Größen. Der Erwartungswert $\mathbb{E}(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ definiert eine Ebene im 3-dimensionalen Raum. Häufig ist die Beziehung zwischen x_1, x_2 und Y aber etwas komplexer. Beispielsweise lautet ein volles **polynomiales** Modell (2. Ordnung) in x_1 und x_2 wie folgt:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \beta_2 x_{i2} + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

Wegen des Terms $\beta_{12} x_{i1} x_{i2}$ hängt der Effekt einer Änderung von x_1 um eine Einheit vom Wert der anderen erklärenden Variablen x_2 ab (und umgekehrt). Man sagt in diesem Fall, dass x_1 und x_2 *interagieren*.

Ein **multiples lineares Regressionsmodell** mit p erklärenden Variablen (**Prädiktoren** oder **Regressoren**) lautet wie folgt:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Die Fehler $\{\varepsilon_i\}$ seien unabhängige, nach $N(0, \sigma^2)$ verteilte, stochastische Größen. Bei den Regressoren kann es sich um p verschiedene Einflussgrößen handeln oder um Funktionen einer kleineren Menge von Variablen. Beispielsweise gibt es beim obigen vollen polynomialen Modell 2. Ordnung in x_1 und x_2 nur zwei erklärende Variablen aber insgesamt $p = 5$ Regressoren.

Matrixschreibweise: Mit den Bezeichnungen:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

lässt sich das Regressionsmodell wie folgt darstellen:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$$

Dabei bezeichnet $\mathbf{0} = (0, 0, \dots, 0)'$ den n -dimensionalen Nullvektor und \mathbf{I} die $(n \times n)$ -Einheitsmatrix. Damit folgt:

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$$

9.2.1 Parameterschätzung

Die Koeffizienten $\beta_0, \beta_1, \dots, \beta_p$ und die Varianz $\text{Var}(\varepsilon_i) = \sigma^2$ sind üblicherweise unbekannt und müssen auf Basis von Beobachtungen $(y_i, x_{i1}, \dots, x_{ip})$, $i = 1, 2, \dots, n$, geschätzt werden. Nehmen wir dazu wieder das **KQ-Prinzip**, so sind die Koeffizienten $\beta_0, \beta_1, \dots, \beta_p$ so zu bestimmen, dass die folgende Quadratsumme minimal wird:

$$S(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2$$

Dieses Minimierungsproblem lässt sich auf die übliche Weise lösen, indem wir die $(p+1)$ partiellen Ableitungen bilden und gleich Null setzen:

$$\frac{\partial S(\beta_0, \beta_1, \dots, \beta_p)}{\partial \beta_j} = 0, \quad j = 0, 1, \dots, p$$

Für die Lösung des auf diese Weise entstehenden linearen Gleichungssystems (**Normalgleichungen**) verwendet man in der Praxis ein entsprechendes Computerprogramm. In Matrixschreibweise lauten die Normalgleichungen wie folgt:

$$\mathbf{X}' \mathbf{X} \boldsymbol{\beta} = \mathbf{X} \mathbf{Y}$$

Üblicherweise kann man davon ausgehen, dass $\mathbf{X}' \mathbf{X}$ invertierbar ist; in diesem Fall ist der **KQ-Schätzer** gegeben durch:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}$$

Der KQ-Schätzer ist erwartungstreu:

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbb{E}(\mathbf{Y}) = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}$$

Die Varianz-Kovarianzmatrix ist gegeben durch:

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \underbrace{\text{Cov}(\mathbf{Y})}_{= \sigma^2 \mathbf{I}} \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$$

Der KQ-Schätzer ist eine lineare Funktion von \mathbf{Y} ; damit folgt:

$$\hat{\boldsymbol{\beta}} \sim N_{p+1} \left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} \right)$$

Der Vektor der **Prognosewerte** ist gegeben durch:

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{X} \hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'}_{= \mathbf{H}} \mathbf{Y} = \mathbf{H} \mathbf{Y}$$

(\mathbf{H} ist die *Hatmatrix*; vgl. 9.1.9.) Für den Vektor der **Residuen** $e_i = Y_i - \hat{Y}_i$ gilt:

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Bem: Die Residuen unterliegen $(p + 1)$ linearen Bedingungen:

$$\sum_{i=1}^n e_i = 0, \quad \sum_{i=1}^n e_i x_{ij} = 0, \quad j = 1, 2, \dots, p$$

Auf Basis der Residuen ist ein erwartungstreuer Schätzer von σ^2 gegeben durch:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - p - 1} = \frac{\mathbf{e}' \mathbf{e}}{n - p - 1}$$

9.2.2 ANOVA-Tafel und F-Test

Ebenso wie für das einfache lineare Regressionsmodell gilt auch im multiplen Fall eine **Varianzzerlegung** der folgenden Form:

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}_{\text{SSE}}$$

SST (= totale Quadratsumme) hat nach wie vor $n - 1$ Freiheitsgrade, SSR (= Regressionsquadratsumme) sind p und SSE (= Fehlerquadratsumme) sind $n - p - 1$ Freiheitsgrade zugeordnet. Die verschiedenen (mittleren) Quadratsummen werden üblicherweise in Form einer **ANOVA-Tafel** angeordnet. Für das multiple lineare Modell ist diese Tafel gegeben wie folgt:

ANOVA-Tafel für das multiple lineare Regressionsmodell

	df	SS	MS	F
Regression	p	SSR	$\text{MSR} = \text{SSR}/p$	MSR/MSE
Fehler	$n - p - 1$	SSE	$\text{MSE} = \text{SSR}/(n - p - 1)$	
Total	$n - 1$	SST		

Zunächst stellt sich die Frage, ob die Regressoren in ihrer Gesamtheit überhaupt etwas zur Erklärung der Variation in der Antwortvariablen beitragen, d. h., man interessiert sich zunächst für einen Test von:

$$\mathcal{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{gegen} \quad \mathcal{H}_1 : \exists i \text{ mit } \beta_i \neq 0$$

Eine geeignete Teststatistik steht in der F -Spalte der ANOVA-Tafel; wie man zeigen kann, gilt unter \mathcal{H}_0 :¹⁶

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F(p, n - p - 1)$$

D. h., \mathcal{H}_0 wird zum Niveau α verworfen, falls:

$$F > F_{p, n-p-1; 1-\alpha}$$

Den p -Wert berechnet man wie folgt:

$$p\text{-Wert} = P(F(p, n - p - 1) \geq F)$$

Multiples Bestimmtheitsmaß: Auf Basis der Varianzzerlegung lässt sich auch für das multiple lineare Modell das **Bestimmtheitsmaß** wie folgt definieren:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

R^2 liegt zwischen 0 und 1 und repräsentiert den Anteil an der Variation von Y , der durch das Regressionsmodell erklärt wird. Allerdings ist R^2 in mehrfacher Hinsicht kein ideales Maß für die „Qualität“ eines Regressionsmodells. So kann man zeigen, dass R^2 bei Hinzunahme eines weiteren (möglicherweise irrelevanten) Regressors nicht kleiner werden kann. Ein besseres Maß bekommt man dadurch, dass man nicht die „rohen“ Quadratsummen miteinander vergleicht, sondern die auf diesen Quadratsummen basierenden *Varianzschätzungen*. Ignoriert man die erklärenden Variablen ist $\text{SST}/(n - 1)$ ein Schätzer für die Varianz; auf Basis des Modells ist $\text{SSE}/(n - p - 1)$ ein unverzerrter Varianzschätzer. Das **modifizierte**¹⁷ **Bestimmtheitsmaß** ist dann definiert wie folgt:

$$R_a^2 = 1 - \frac{\text{SSE}/(n - p - 1)}{\text{SST}/(n - 1)}$$

Wie man zeigen kann, gilt $R_a^2 \leq R^2$ und das modifizierte R^2 kann bei Hinzunahme eines weiteren Regressors auch kleiner werden.

¹⁶Unter der – hier getroffenen – Voraussetzung, dass $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$.

¹⁷engl. *adjusted*

9.2.3 Konfidenzintervalle und t–Tests

Mit Hilfe des F–Tests des vorigen Abschnitts kann man testen, ob die Prädiktoren (oder Regressoren) in ihrer Gesamtheit einen Beitrag zur Erklärung der Variation in der Antwortvariablen leisten. Das ist aber nur ein erster Schritt. Ist der F–Test signifikant (d. h., wird $\mathcal{H}_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ verworfen), stellt sich als nächstes die Frage nach der Signifikanz der einzelnen Koeffizienten β_j , $j = 1, 2, \dots, p$. Schreibt man:

$$\mathbf{C} = (\mathbf{X}' \mathbf{X})^{-1} = \begin{bmatrix} C^{00} & C^{01} & \dots & C^{0p} \\ C^{10} & C^{11} & \dots & C^{1p} \\ \vdots & \vdots & \ddots & \vdots \\ C^{p0} & C^{p1} & \dots & C^{pp} \end{bmatrix}$$

so folgt wegen $\widehat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}' \mathbf{X})^{-1})$, dass (vgl. 5.6.3):

$$\widehat{\beta}_j \sim N(\beta_j, \sigma^2 C^{jj}), \quad j = 0, 1, \dots, p$$

Ersetzt man das unbekannte σ^2 durch den unverzerrten Schätzer MSE, so ist der (geschätzte) Standardfehler von $\widehat{\beta}_j$ gegeben durch:

$$s(\widehat{\beta}_j) = \sqrt{\text{MSE } C^{jj}}$$

und es gilt:

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)} \sim t(n - p - 1)$$

Konfidenzintervall für β_j : T_j ist eine Pivotgröße und ein $(1 - \alpha)$ –Konfidenzintervall für β_j ist gegeben durch:

$$\widehat{\beta}_j \pm t_{n-p-1; 1-\alpha/2} s(\widehat{\beta}_j)$$

t–Test für β_j : Ein Test zum Niveau α für die Hypothesen:

$$\mathcal{H}_0 : \beta_j = \beta_{j0} \quad \text{gegen} \quad \mathcal{H}_1 : \beta_j \neq \beta_{j0}$$

ist gegeben durch:

$$\text{Verwerfe } \mathcal{H}_0, \text{ falls: } \frac{|\hat{\beta}_j - \beta_{j0}|}{s(\hat{\beta}_j)} > t_{n-p-1; 1-\alpha/2}$$

Häufig interessiert man sich für die folgenden Hypothesen:

$$\mathcal{H}_0 : \beta_j = 0 \quad \text{gegen} \quad \mathcal{H}_1 : \beta_j \neq 0$$

Ein Test zum Niveau α lautet wie folgt:

$$\text{Verwerfe } \mathcal{H}_0, \text{ falls: } |t(\hat{\beta}_j)| = \frac{|\hat{\beta}_j|}{s(\hat{\beta}_j)} > t_{n-p-1; 1-\alpha/2}$$

Letzteren Test nennt man den **partiellen t–Test** für β_j . (Er gehört zum Standardoutput von R.) Der p –Wert ist gegeben durch:

$$p\text{–Wert} = 2P(t(n-p-1) \geq |t(\hat{\beta}_j)|)$$

Interpretation des partiellen t–Tests: Da der Schätzer $\hat{\beta}_j$ von β_j nicht nur vom Regressor x_j sondern i. A. auch von den anderen Regressoren x_i ($i \neq j$) des Modells abhängt, handelt es sich um einen **partiellen** Test, d. h., der Regressor x_j wird so interpretiert, als ob er der *letzte* war, der in das Modell aufgenommen wurde (alle anderen Regressoren im Modell). Ein nichtsignifikanter partieller t–Test besagt also *nicht*, dass der entsprechende Regressor keinen Einfluss hat, sondern lediglich, dass der über alle anderen Regressoren hinausgehende *zusätzliche* Einfluss gering ist.

Bem: Folgende (paradoxe) Situation kann vorkommen: Der F–Test verwirft (ist signifikant), jedoch *keiner* der partiellen t–Tests. Das ist ein Hinweis darauf, dass man auf (zumindest) einen Regressor verzichten kann. Auch der umgekehrte Fall kann vorkommen (allerdings wesentlich seltener): Der F–Test ist nicht signifikant, wohl aber ein oder mehrere partielle t–Tests. Letzteres ist ein Hinweis auf ein sehr ungünstiges „Modelldesign“.

9.2.4 Beispiele

1. [Polynomiales Modell] Der Scatterplot (Abb 9.8) der folgenden Daten deutet auf einen nichtlinearen Zusammenhang hin:

x	0	1	2	3	4	5	6	7	8
y	1.2	16.5	28.9	23.1	81.7	120.3	132.5	197.6	283.8

Das einfachste Modell für einen nichtlinearen Zusammenhang ist ein quadratisches Modell. Zum Vergleich betrachten wir auch ein einfaches lineares und ein kubisches Modell:

$$\text{Modell (1)} \quad Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\text{Modell (2)} \quad Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \varepsilon$$

$$\text{Modell (3)} \quad Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \beta_{111} x^3 + \varepsilon$$

Modell (1): Der F–Test und der (äquivalente) t–Test (für β_1) sind beide (hoch) signifikant. An diesem Beispiel kann man auch sehen, dass ein signifikantes Modell keineswegs auch ein adäquates Modell sein muss. (Signifikanz und Adäquatheit sind verschiedene Dinge!) $R^2 \approx 90\%$ ist zwar akzeptabel, lässt sich aber noch verbessern.

```
mod1 <- lm(y ~ x)
summary(mod1)

....
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-33.473	20.221	-1.655	0.14182
x	32.968	4.247	7.762	0.00011 ***

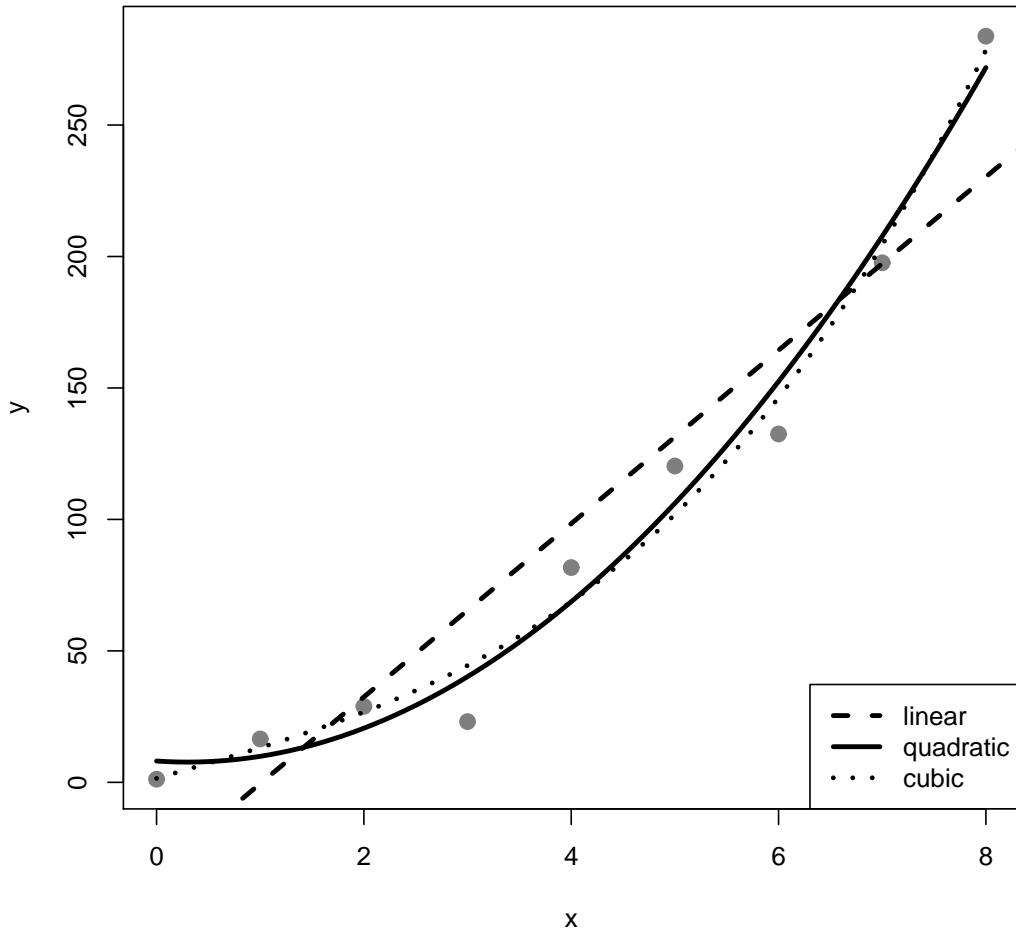
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.9 on 7 degrees of freedom
Multiple R-squared: 0.8959, Adjusted R-squared: 0.881
F-statistic: 60.25 on 1 and 7 DF, p-value: 0.0001104

Modell (2): Der F–Test ist (hoch) signifikant; von den partiellen t–Tests ist nur der Test für β_{11} (hoch) signifikant, nicht aber der für β_1 . Man sollte allerdings daraus nicht den Schluss ziehen, dass man auf den linearen Term ($\beta_1 x$) verzichten kann! Aus folgendem Grund: Verändert man die Skalierung von x , ersetzt beispielsweise x durch $x + 5$, ist der lineare Term signifikant. (Generell: Gibt es einen Term höherer Ordnung im Modell, sollten auch alle Terme niedrigerer Ordnung enthalten sein.) R^2 und R_a^2 haben gegenüber Modell (1) deutlich zugelegt.

```
mod2 <- lm(y ~ x + I(x^2))
summary(mod2)
```

Abbildung 9.8: Polynomiale Regression



```
.....
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.0994   12.7025   0.638  0.54726
x          -2.6654    7.4043  -0.360  0.73119
I(x^2)      4.4542    0.8905   5.002  0.00245 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.63 on 6 degrees of freedom
Multiple R-squared:  0.9799,    Adjusted R-squared:  0.9732
F-statistic: 146 on 2 and 6 DF,  p-value: 8.161e-06
```

Modell (3): Der F–Test ist (hoch) signifikant, aber keiner der partiellen t–Tests! Das illustriert die Bem am Schluss des vorigen Abschnitts. Man kann also auf den kubischen

Term verzichten. R^2 hat praktisch nicht mehr zugelegt, R_a^2 ist aber gegenüber Modell (2) leicht gesunken.

```

mod3 <- lm(y ~ x + I(x^2) + I(x^3))
summary(mod3)

.....

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.4939    14.6221   0.102   0.923
x           11.5677   16.8760   0.685   0.524
I(x^2)      -0.2640    5.0953  -0.052   0.961
I(x^3)       0.3932    0.4179   0.941   0.390

Residual standard error: 15.78 on 5 degrees of freedom
Multiple R-squared:  0.9829,    Adjusted R-squared:  0.9726
F-statistic: 95.77 on 3 and 5 DF,  p-value: 7.748e-05

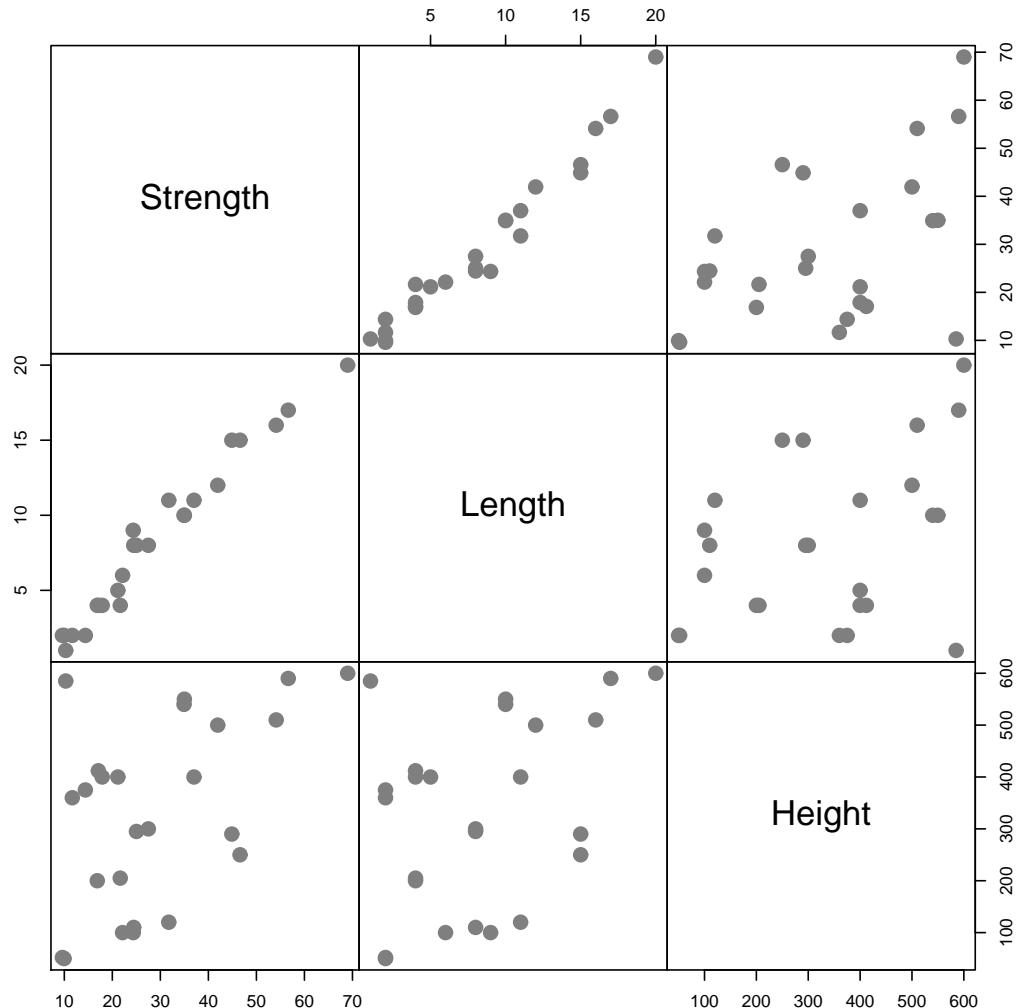
```

2. [Multiples Modell] Bei einer Untersuchung zum Drahtbonding von Chips wurde neben der Reißfestigkeit (y) auch die Drahtlänge (x_1) und die Chiphöhe (x_2) erhoben (wirebond.txt):

n	y	x_1	x_2	n	y	x_1	x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

Gesucht ist ein empirisches Modell für den Zusammenhang von y und x_1 und x_2 . Ein erster Schritt sind paarweise Scatterplots (Abb 9.9); auffällig ist der starke lineare Zusammenhang zwischen Strength und Length.

Abbildung 9.9: Scatterplotmatrix der wirebond–Daten



In Ermangelung einer physikalischen (oder sonstigen) Theorie nehmen wir ein multiples lineares Modell der Form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Der folgende R–Output zeigt die Detailergebnisse der Anpassung dieses Modells:

```
bond <- read.table("wirebond.txt", header=TRUE) [,-1]
mod <- lm(Strength ~ Length + Height, data=bond)
summary(mod)

.....
```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.263791   1.060066   2.136 0.044099 *
Length       2.744270   0.093524  29.343 < 2e-16 ***
Height      0.012528   0.002798   4.477 0.000188 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.288 on 22 degrees of freedom
Multiple R-squared: 0.9811,    Adjusted R-squared: 0.9794
F-statistic: 572.2 on 2 and 22 DF,  p-value: < 2.2e-16

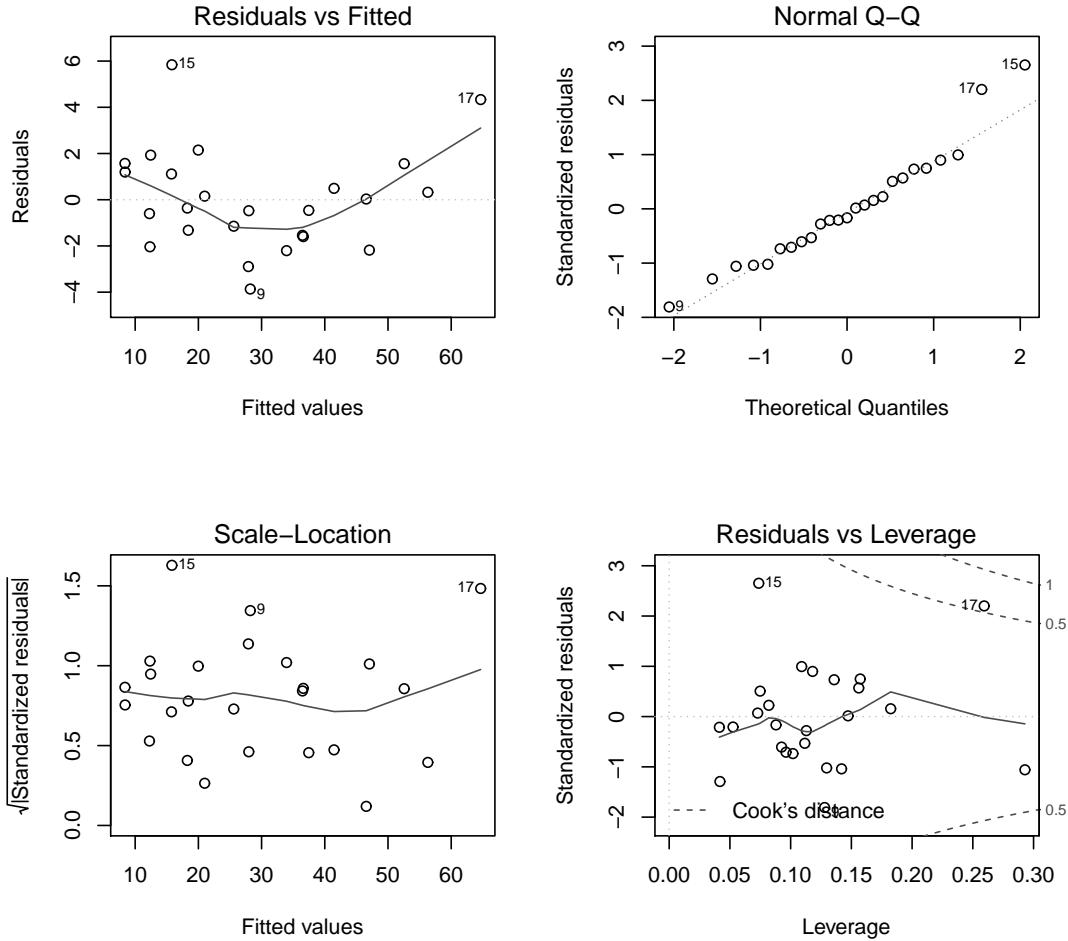
```

Der F–Test und die partiellen t–Tests für Length und Height sind alle (hoch) signifikant. Letzteres mag etwas überraschen, da im Scatterplot der Zusammenhang zwischen Strength und Height nur schwach ausgeprägt ist. Über Length hinausgehend leistet im *multiplen* Modell aber auch Height einen (hoch) signifikanten Beitrag zur Erklärung der Variation in Strength. R^2 und R_a^2 sind beide sehr hoch, sodass sich das Modell von dieser Seite für Prognosezwecke eignen sollte. Bevor das Modell verwendet werden kann, muss aber eine Residualanalyse durchgeführt werden, um etwaige Modelldefizite zu erkennen. Mittels `plot(mod)` bekommt man die Plots von Abb 9.10.

Interpretation: Der Plot oben/rechts ist der Normal-QQ–Plot der (standardisierten) Residuen. Keine groben Abweichungen sind erkennbar, allerdings liegen die zwei größten Residuen (möglicherweise Ausreißer) etwas abseits der eingezeichneten Geraden. Der Plot oben/links ist der Residualplot von e_i gegen \hat{y}_i . Drei (bezeichnete) Punkte stechen etwas hervor; auffällig ist aber vor allem eine näherungsweise quadratische Struktur: Das Modell unterschätzt niedrige und hohe, überschätzt aber mittlere Reißfestigkeiten. Möglicherweise lässt sich das Modell durch Hinzunahme von quadratischen Termen ($\beta_{11}x_1^2$ oder $\beta_{22}x_2^2$) verbessern (UE–Aufgabe 9.8) oder andere (nicht im Modell befindliche) Regressoren beeinflussen die Antwortvariable. Der Plot unten/links ist der sog. *Spread-Location–Plot*. Er zeigt, ob die Annahme einer konstanten Varianz σ^2 möglicherweise verletzt ist. Das könnte der Fall sein, wenn ein Trend erkennbar ist (hier nicht der Fall). Schließlich zeigt der Plot unten/rechts, ob es einflussreiche Punkte gibt.¹⁸ Nach diesem Plot hat Punkt Nr. 17 einen mittelstarken Einfluss auf die Anpassung der Regressionsebene.

3. [Dummy–Variablen] Die bisher betrachteten Regressionsmodelle basieren auf *quantitativen* Variablen (Spannung, Länge, etc.). Gelegentlich muss man aber auch *kategoriale* oder *qualitative* Variablen (Geschlecht, Schulabschluss, etc.) einbeziehen. Die übliche Methode, um derartige Variablen zu berücksichtigen, besteht in der Verwendung von **Indikator–** oder **Dummy–Variablen**. Hat eine qualitative Variable k Levels, so definiert man $k - 1$ Dummys. Für $k = 3$ beispielsweise wie folgt:

¹⁸Punkte mit *Cook–Distanz* größer als 1 gelten als einflussreich (vgl. GURKER (2015)).

Abbildung 9.10: Residualanalyse mittels plot(mod)

	x_1	x_2	
	0	0	falls Level = 1
	1	0	falls Level = 2
	0	1	falls Level = 3

(Bem: Die Codierung mit 0 und 1 ist nicht zwingend, auch andere Codierungen sind zulässig; 0/1 ist aber am besten.) Im Folgenden befassen wir uns mit einem Datensatz¹⁹, bestehend aus Messungen der Körpertemperatur (`temp`) in Abhängigkeit von der Herzfrequenz (`hr`) und dem Geschlecht (`gender`). (Datenfile: `normtemp.txt`) Ist x die Indikatorvariable für die qualitative Variable `gender`, so lautet ein passendes Modell etwa wie folgt:

$$(1) \quad \text{temp} = \beta_0 + \beta_1 \text{hr} + \beta_2 x + \varepsilon$$

¹⁹A. L. SHOEMAKER: What's Normal? – Temperature, Gender, and Heart Rate, *Journal of Statistics Education*, Vol. 4/2, 1996.

Für $x = 0$ und $x = 1$ unterscheiden sich die Modelle bezüglich Interzept:

$$x = 0: \text{ temp} = \beta_0 + \beta_1 \text{hr} + \varepsilon$$

$$x = 1: \text{ temp} = (\beta_0 + \beta_2) + \beta_1 \text{hr} + \varepsilon$$

Man kann aber auch unterschiedliche Anstiege modellieren:

$$(2) \quad \text{temp} = \beta_0 + \beta_1 \text{hr} + \beta_2 x + \beta_3 (\text{hr} \times x) + \varepsilon$$

Nun lauten die einzelnen Modelle wie folgt:

$$x = 0: \text{ temp} = \beta_0 + \beta_1 \text{hr} + \varepsilon$$

$$x = 1: \text{ temp} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{hr} + \varepsilon$$

Der folgende R–Output zeigt die Ergebnisse der Anpassung von Modell (1). Man beachte, dass die Variable **gender** als Faktor zu deklarieren ist.

```
mod1 <- lm(temp ~ hr + factor(gender), data=normtemp)
summary(mod1)

....
```

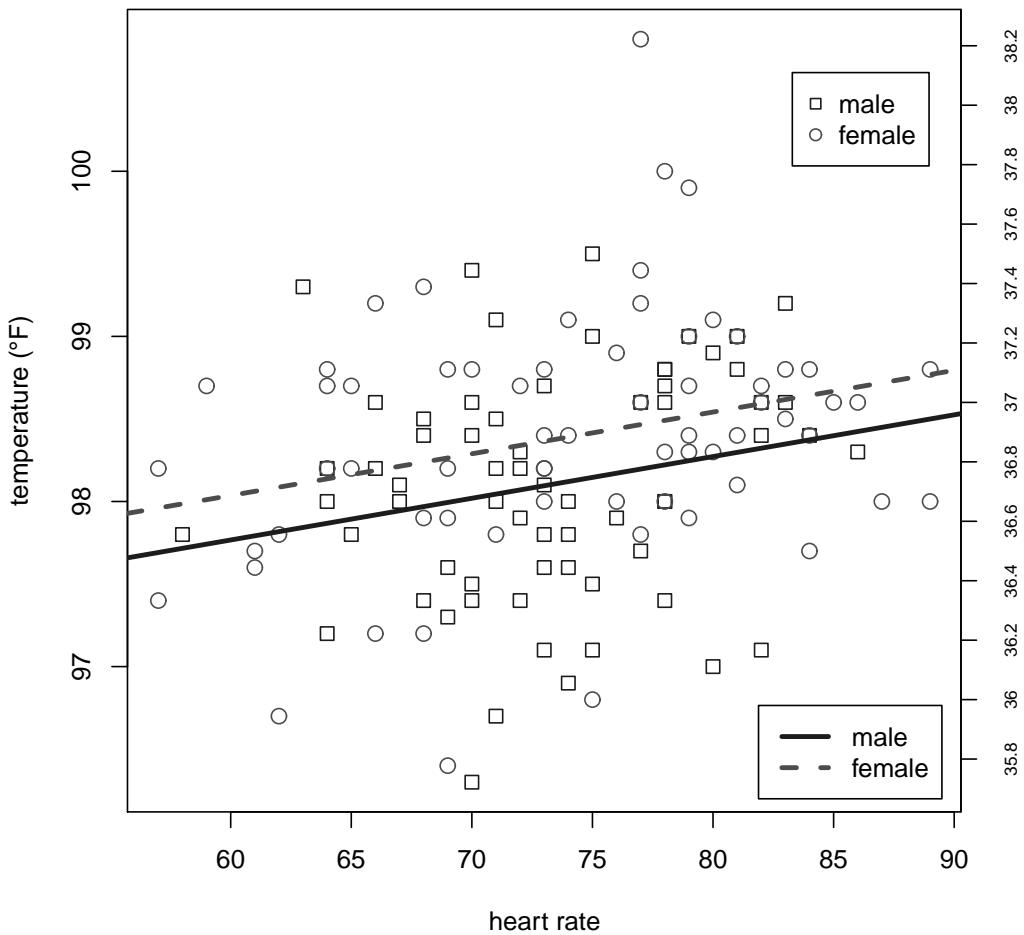
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	96.250814	0.648717	148.371	< 2e-16	***
hr	0.025267	0.008762	2.884	0.00462	**
factor(gender)2	0.269406	0.123277	2.185	0.03070	*

```
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7017 on 127 degrees of freedom
Multiple R-squared: 0.09825, Adjusted R-squared: 0.08405
F-statistic: 6.919 on 2 and 127 DF, p-value: 0.001406
```

Alle Koeffizienten sind signifikant; der Koeffizient von **factor(gender)2** besagt, dass – auf dem Niveau 5% – die mittlere Körpertemperatur von Frauen um etwa 0.27°F ($\cong 0.15^\circ\text{C}$) höher ist als diejenige der Männer. Abb 9.11 zeigt die Daten mit verschiedenen aber parallelen Regressionsgeraden. Die letzte Zeile im R–Output besagt, dass der F–Test hoch signifikant ist. Das ist ein Test für die Hypothese, dass alle Koeffizienten mit Ausnahme des Interzepts gleich Null sind. Besser wäre es allerdings zu testen, ob **gender** signifikant ist, wenn **hr** bereits im Modell ist. Dieser Test entspricht dem partiellen t–Test für

Abbildung 9.11: Körpertemperatur in Abhängigkeit von Herzfrequenz und Geschlecht



`factor(gender)2`, der auf dem Niveau 5% signifikant ist. Eine Alternative ist die Verwendung der Funktion `anova()`:

```

mod0 <- lm(temp ~ hr, data=normtemp)
anova(mod0, mod1)
Analysis of Variance Table

Model 1: temp ~ hr
Model 2: temp ~ hr + factor(gender)
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     128 64.883
2     127 62.532  1   2.3515 4.7758 0.0307 *

```

Die p -Werte stimmen für beide Tests überein.

Aufgaben

9.1 Im Zuge der Anpassung eines einfachen linearen Regressionsmodells ergeben sich auf Basis von $n = 14$ Beobachtungen (x_i, y_i) die folgenden Werte:

$$\begin{aligned}\sum_{i=1}^n x_i &= 43, & \sum_{i=1}^n x_i^2 &= 157.42 \\ \sum_{i=1}^n y_i &= 572, & \sum_{i=1}^n y_i^2 &= 23530 \\ \sum_{i=1}^n x_i y_i &= 1697.80\end{aligned}$$

- (a) Bestimmen Sie die KQ–Schätzwerte von β_0 und β_1 und schätzen Sie σ^2 .
- (b) Wenn $x = 3.7$, welcher Prognosewert ergibt sich für Y ?
- (c) Erstellen Sie die ANOVA–Tafel und bestimmen Sie den p –Wert.
- (d) Wie groß ist R^2 ?

9.2 Zeigen Sie für ein einfaches lineares Regressionsmodell:

- (a) $\sum_{i=1}^n e_i = \sum_{i=1}^n x_i e_i = 0$
- (b) $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$
- (c) Der Punkt (\bar{x}, \bar{y}) liegt exakt auf der KQ–Geraden.

9.3 [Regression durch den Nullpunkt] Betrachten Sie ein lineares Modell der Form:

$$Y_i = \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

und bestimmen Sie den KQ–Schätzer von β_1 . Wie lautet ein unverzerrter Schätzer für σ^2 ? (Zusatz: Bestimmen Sie ein $(1 - \alpha)$ –Konfidenzintervall für β_1 .)

9.4 In einer Studie wurde der Zusammenhang zwischen Lärmpegel x (in db) und Blutdrucksteigerung y (in mmHg) untersucht. Die folgenden Daten sind repräsentativ für die an der Studie beteiligten Personen:

y	1	0	1	2	5	1	4	6	2	3
x	60	63	65	70	70	70	80	90	80	80

y	5	4	6	8	4	5	7	9	7	6
x	85	89	90	90	90	90	94	100	100	100

(a) Zeichnen Sie einen Scatterplot von y gegen x . Ist ein einfaches lineares Modell der Form $Y = \beta_0 + \beta_1 x + \varepsilon$ ein geeignetes Modell?

(b) Bestimmen Sie die KQ-Schätzwerte von β_0 und β_1 und zeichnen Sie die KQ-Gerade über den Scatterplot. Schätzwert für σ^2 ? F-Test? R^2 ?

(c) Führen Sie eine Residualanalyse durch. (Gibt es einflussreiche Punkte?)

9.5 Fortsetzung von Aufgabe 9.4: Ermitteln und zeichnen Sie 95%-Konfidenz- bzw. Prognosebänder. (Hinweis: Vgl. Bsp 9.6.)

9.6 [Gewichtete Kleinste Quadrate] Angenommen, wir möchten ein einfaches lineares Modell der Form $Y = \beta_0 + \beta_1 x + \varepsilon$ anpassen, aber die Varianz von Y hängt vom x -Level ab, d. h., es gelte:

$$\text{Var}(Y_i|x_i) = \sigma_i^2 = \frac{\sigma^2}{w_i}, \quad i = 1, 2, \dots, n$$

Dabei seien $w_i > 0$ (bekannte) *Gewichte*. In diesem Fall liegt es nahe, in der zu minimierenden Quadratsumme, Beobachtungen mit einer kleineren Varianz ein höheres Gewicht zu geben, d. h., statt der üblichen Quadratsumme $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ die folgende Quadratsumme zu minimieren:

$$S_w(\beta_0, \beta_1) = \sum_{i=1}^n w_i (y_i - \beta_0 - \beta_1 x_i)^2$$

Wie lauten in diesem Fall die Normalgleichungen und ihre Lösungen? Die auf diese Weise bestimmten Schätzer von β_0 und β_1 nennt man die **gewichteten KQ-Schätzer**.

9.7 Zeichnen Sie für die folgenden Daten einen Scatterplot:

y	1.81	1.70	1.65	1.55	1.48	1.40	1.30	1.26	1.24	1.21	1.20	1.18
x	20	25	30	35	40	50	60	65	70	75	80	90

- (a) Passen Sie ein quadratisches Modell der Form $Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \varepsilon$ an.
- (b) Ist die Regression insgesamt signifikant? (p -Wert?)
- (c) Benötigt man den quadratischen Term? D. h., testen Sie $\mathcal{H}_0 : \beta_{11} = 0$ gegen $\mathcal{H}_1 : \beta_{11} \neq 0$.
- (d) Zeichnen Sie 95%-Konfidenz- und Prognosebänder.

9.8 Lässt sich das Modell für die `wirebond`-Daten (Bsp 2 von 9.2.4) durch Hinzunahme von quadratischen Termen verbessern? Zeichnen Sie zur Beantwortung dieser Frage die Residuen e_i gegen x_1 (= `Length`) und x_2 (= `Height`). Bei welchem Plot zeigt sich näherungsweise eine quadratische Abhängigkeit? Erweitern Sie das Modell um den entsprechenden quadratischen Term ($\beta_{11} x_1^2$ oder $\beta_{22} x_2^2$) und wiederholen Sie die Residualanalyse.

- 9.9 Traditionellerweise wird die Qualität eines neuen Jahrgangs von Weinen aus Bordeaux im März des folgenden Jahres durch Experten beurteilt. Diese Beurteilungen sind aber meist recht unzuverlässig, sodass möglicherweise ein Regressionsmodell zur Prognose des letztlich erzielbaren Preises besser geeignet sein könnte. Der (historische) Datensatz `wine.txt` umfasst den Preis (relativ zum Jahrgang 1961²⁰) für die Jahrgänge 1952 bis 1980 zusammen mit vier weiteren Variablen (vgl. Datenfile für Details). Zu den Jahrgängen 1954 und 1956, die unter Weinkennern als schwache Jahrgänge gelten, gibt es keine Angaben. Außerdem gibt es Angaben zu den Jahrgängen 1987 bis 1991, die zur Validierung des Modells verwendet werden können.

Year	Temp	Rain	PrevRain	Age	Price
1952	17.12	160	600	31	0.368
1953	16.73	80	690	30	0.635
1954	–	–	–	29	–
1955	17.15	130	502	28	0.446
1956	–	–	–	27	–
1957	16.13	110	420	26	0.221
1958	16.42	187	582	25	0.180
1959	17.48	187	485	24	0.658
1960	16.42	290	763	23	0.139
1961	17.33	38	830	22	1.000
1962	16.30	52	697	21	0.331
1963	15.72	155	608	20	0.168
1964	17.27	96	402	19	0.306
1965	15.37	267	602	18	0.106
1966	16.53	86	819	17	0.473
1967	16.23	118	714	16	0.191
1968	16.20	292	610	15	0.105
1969	16.55	244	575	14	0.117
1970	16.67	89	622	13	0.404
1971	16.77	112	551	12	0.272
1972	14.98	158	536	11	0.101
1973	17.07	123	376	10	0.156
1974	16.30	184	574	9	0.111
1975	16.95	171	572	8	0.301
1976	17.65	247	418	7	0.253
1977	15.58	87	821	6	0.107
1978	15.82	51	763	5	0.270
1979	16.17	122	717	4	0.214
1980	16.00	74	578	3	0.136
1987	16.98	115	452	-4	0.135
1988	17.10	59	808	-5	0.271
1989	18.60	82	443	-6	0.432
1990	18.70	80	468	-7	0.568
1991	17.70	183	570	-8	0.142

²⁰Gilt als einer der besten Nachkriegsjahrgänge.

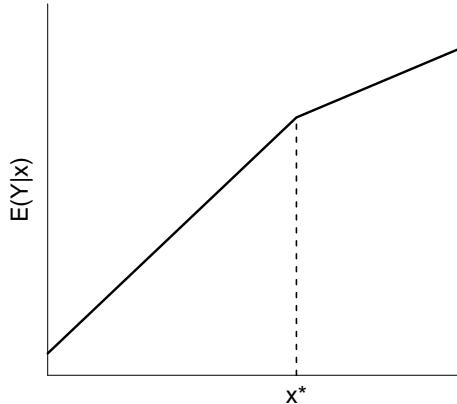
- (a) Zeichnen Sie für die Jahrgänge 1952 bis 1980 paarweise Scatterplots von `Price` gegen die anderen Variablen. (Hinweis: `pairs()`)
- (b) Passen Sie für die Jahrgänge 1952 bis 1980 ein Modell der folgenden Form an:

$$\log(\text{Price}) = \beta_0 + \beta_1 \text{Temp} + \beta_2 \text{Rain} + \beta_3 \text{PrevRain} + \beta_4 \text{Age} + \varepsilon$$

Interpretieren Sie die Ergebnisse der Anpassung (F -Test, partielle t -Tests, ...) und führen Sie eine Residualanalyse durch.

- (c) Prognostizieren Sie auf Basis des Regressionsmodells die Preise für die Jahrgänge 1987 bis 1991 und vergleichen Sie mit den tatsächlichen Preisen.

- 9.10 [Broken Stick Regression] Angenommen, die Antwortvariable Y steht in einer linearen Beziehung zu einer erklärenden Variablen x . Wie in der folgenden Abbildung dargestellt, gibt es allerdings an einer bestimmten (bekannten) Stelle x^* eine abrupte Änderung im Anstieg.



Wie lautet ein Regressionsmodell, mit dem man die Signifikanz der Änderung im Anstieg testen könnte? (Hinweis: Definieren Sie eine entsprechende Dummy-Variable.)

Tabellen

Tabelle 1: Verteilungsfunktion $\Phi(x)$ der Standardnormalverteilung $N(0, 1)$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Tabelle 2: Quantile z_p der $N(0, 1)$

p	0.60	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995	0.999
z_p	0.2533	0.6745	0.8416	1.0364	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902

Tabelle 3: Quantile $t_{n;p}$ der t–Verteilung

n	p								
	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995	0.999
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	318.309
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385
31	0.682	0.853	1.054	1.309	1.696	2.040	2.453	2.744	3.375
32	0.682	0.853	1.054	1.309	1.694	2.037	2.449	2.738	3.365
33	0.682	0.853	1.053	1.308	1.692	2.035	2.445	2.733	3.356
34	0.682	0.852	1.052	1.307	1.691	2.032	2.441	2.728	3.348
35	0.682	0.852	1.052	1.306	1.690	2.030	2.438	2.724	3.340
36	0.681	0.852	1.052	1.306	1.688	2.028	2.434	2.719	3.333
37	0.681	0.851	1.051	1.305	1.687	2.026	2.431	2.715	3.326
38	0.681	0.851	1.051	1.304	1.686	2.024	2.429	2.712	3.319
39	0.681	0.851	1.050	1.304	1.685	2.023	2.426	2.708	3.313
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.261
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232
70	0.678	0.847	1.044	1.294	1.667	1.994	2.381	2.648	3.211
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195
90	0.677	0.846	1.042	1.291	1.662	1.987	2.368	2.632	3.183
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090

Bem: In der letzten Zeile stehen die entsprechenden Quantile der $N(0, 1)$ (vgl. Tabelle 2).

Tabelle 4: Quantile $\chi^2_{n,p}$ der Chiquadratverteilung

n	<i>p</i>									
	0.005	0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.960	22.106	24.075	26.492	48.363	52.192	55.668	59.892	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	52.949	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	30.765	54.090	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	31.625	55.230	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	32.487	56.369	60.481	64.201	68.710	71.893
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
46	25.041	26.657	29.160	31.439	34.215	58.641	62.830	66.617	71.201	74.437
47	25.775	27.416	29.956	32.268	35.081	59.774	64.001	67.821	72.443	75.704
48	26.511	28.177	30.755	33.098	35.949	60.907	65.171	69.023	73.683	76.969
49	27.249	28.941	31.555	33.930	36.818	62.038	66.339	70.222	74.919	78.231
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490

Tabelle 5: Quantile $F_{m,n; p}$ der F-Verteilung

p	n	m											
		1	2	3	4	5	6	7	8	9	10	12	15
0.95	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9
0.975		647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9
0.99		4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157
0.95	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43
0.975		38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.42	39.43
0.99		98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43
0.95	3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.703
0.975		17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25
0.99		34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87
0.95	4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.858
0.975		12.22	10.65	9.979	9.605	9.364	9.197	9.074	8.980	8.905	8.844	8.751	8.657
0.99		21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20
0.95	5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.619
0.975		10.01	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619	6.525	6.428
0.99		16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.888	9.722
0.95	6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.938
0.975		8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600	5.523	5.461	5.366	5.269
0.99		13.75	10.92	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.559
0.95	7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.511
0.975		8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823	4.761	4.666	4.568
0.99		12.25	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.314
0.95	8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.284	3.218
0.975		7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295	4.200	4.101
0.99		11.26	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.515
0.95	9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	3.006
0.975		7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026	3.964	3.868	3.769
0.99		10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.962
0.95	10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.845
0.975		6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717	3.621	3.522
0.99		10.04	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.558
0.95	12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.617
0.975		6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374	3.277	3.177
0.99		9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	4.010
0.95	15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.403
0.975		6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.060	2.963	2.862
0.99		8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.522

$$F_{m,n; p} = \frac{1}{F_{n,m; 1-p}}$$

Literatur

- Baron, M. (2013): *Probability and Statistics for Computer Scientists*, 2nd Ed., Chapman & Hall/CRC.
- Bosch, K. (2010): *Elementare Einführung in die angewandte Statistik*, 9. Aufl., Vieweg/Teubner.
- Bosch, K. (2011): *Elementare Einführung in die Wahrscheinlichkeitsrechnung*, 11. Aufl., Vieweg/Teubner.
- Dalgaard, P. (2008): *Introductory Statistics with R*, 2nd Ed., Springer.
- DeGroot, M. H. and Schervish, M. J. (2014): *Probability and Statistics*, 4th Ed., Pearson.
- Gurker, W. (2016): *Angewandte Mathematische Statistik* [Skriptum zur VO].
- Gurker, W. (2016): *Introduction to Regression Modeling* [Skriptum zur VO].
- Hogg, R. V., McKean, J. W., and Craig A. T. (2005): *Introduction to Mathematical Statistics*, 6th Ed., Pearson/Prentice Hall.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013): *An Introduction to Statistical Learning – with Applications in R*, Springer.
- Kroese, D. P. and Chan, J. C. C. (2014): *Statistical Modeling and Computation*, Springer.
- Pruim, R. (2011): *Foundations and Applications of Statistics — An Introduction Using R*, American Mathematical Society (AMS).
- Robert, C. P. and Casella, G. (2010): *Introducing Monte Carlo Methods with R*, Springer.
- Ross, S. M. (2014): *Introduction to Probability and Statistics for Engineers and Scientists*, 5th Ed., Academic Press.
- Ross, S. M. (2014): *Introduction to Probability Models*, 11th Ed., Academic Press.
- Steland, A. (2013): *Basiswissen Statistik*, 3. Aufl., Springer.
- Trivedi, K. S. (2002): *Probability and Statistics with Reliability, Queuing and Computer Science Applications*, 2nd Ed., Wiley.
- Venables, W. N. and Ripley, B. D. (2003): *Modern Applied Statistics with S*, 4th Ed., Springer.
- Verzani, J. (2014): *Using R for Introductory Statistics*, 2nd Ed., Chapman & Hall/CRC.
- Viertl, R. (2003): *Einführung in die Stochastik – mit Elementen der Bayes-Statistik und der Analyse unscharfer Information*, 3. Aufl., Springer.
- Wollschläger, D. (2016): *R kompakt – Der schnelle Einstieg in die Datenanalyse*, 2. Aufl., Springer.

