Mathias Trabs
Moritz Jirak
Konstantin Krenz
Markus Reiß

Statistik und maschinelles Lernen

Eine mathematische Einführung in klassische und moderne Methoden

Vorläufige Arbeitsversion vom 23. Dezember 2025

Springer Nature

Inhaltsverzeichnis

1 Grundlagen der Statistik 1
1.1 Das statistische Modell 1
1.2 Parameterschätzung 5
1.2.1 Konstruktionsprinzipien 5
1.2.2 Minimax- und Bayes-Ansatz 13
1.3 Hypothesentests 20
1.3.1 Statistische Tests und ihre Fehler 20
1.3.2 Das Neyman-Pearson-Lemma 33
1.4 Konfidenzmengen 40
1.5 Aufgaben 44

2 Das lineare Modell 49
2.1 Regression und kleinste Quadrate 49
2.1.1 Lineare Regression 50
2.1.2 Schätzen im linearen Modell 55
2.1.3 Zufälliges Design und Vorhersage 65
2.2 Inferenz unter Normalverteilungsannahme 73
2.3 Varianzanalyse 87
2.4 Aufgaben 95

3 Modellwahl 101
3.1 Informationskriterien 101
3.1.1 Akaike-Informationskriterium (AIC) 106
3.1.2 Das Bayes-Informationskriterium (BIC) 113
3.2 Orakelungleichung für die penalisierte Modellwahl 117
3.3 Aufgaben 122

A Konzepte der Wahrscheinlichkeitstheorie 127
A.1 Grundbegriffe der Maßtheorie und Stochastik 127
A.2 Diskrete Verteilungen 142
A.3 Stetige Verteilungen 144

Literaturverzeichnis 147

Sachverzeichnis 149

Abbkürzungen und Symbole

$\emptyset$	leere Menge
A^C	Komplement der Menge A
$\lvert A\rvert$	Kardinalität der endlichen Menge A
$\mathbb{1}_A(x)$, $\mathbb{1}(x \in A)$	Indikatorfunktion zur Menge A, das heißt falls $x \in A$, dann gilt $\mathbb{1}_A(x) = \mathbb{1}(x \in A) = 1$ und sonst 0
$\mathcal{P}(A)$	Potenzmenge der Menge A
$\mathcal{B}(\Omega)$	Borel-σ-Algebra über Menge Ω
$\mathbb{N}, \mathbb{N}_0$	Menge der natürlichen Zahlen ohne bzw. mit der Null
$\mathbb{R}, \mathbb{R}_+$	Menge der reellen Zahlen bzw. Menge der reellen Zahlen größer oder gleich Null
$a \wedge b, a \vee b$	Minimum bzw. Maximum von $a, b \in \mathbb{R}$
$\arg\min_a f(a)$	Minimalstelle oder Menge der Minimalstellen der Funktion f
$\arg\max_a f(a)$	Maximalstelle oder Menge der Maximalstellen der Funktion f
$(a)_+, (a)_-$	Positiv- bzw. Negativteil, das heißt $(a)_+ = a \vee 0, (a_-) = (-a) \vee 0$
$\propto$	proportional, das heißt Gleichheit bis auf eine multiplikative Konstante
$(\mathcal{X}, \mathcal{F}, \mathbb{P}_\vartheta)$	Wahrscheinlichkeitsraum bezüglich des Parameters $\vartheta \in \Theta$
$\mathbb{E}_\vartheta$, Var_ϑ	Erwarungswert bzw. Varianz bezüglich $\mathbb{P}_\vartheta$
Cov, $\mathbb{C}$ov	Kovarianz und Kovarianzmatrix
$\mathrm{U}(A)$	Gleichverteilung auf der Menge A
$\mathrm{N}(\mu, \sigma^2)$	Normalverteilung mit Mittelwert μ und Varianz σ^2
$\mathrm{Ber}(\vartheta)$	Bernoulli-Verteilung mit Erfolgswahrscheinlichkeit $\vartheta \in [0, 1]$
Φ	Verteilungsfunktion von $\mathrm{N}(0, 1)$
$\sim$	ist verteilt gemäß, zum Beispiel „$Y \sim \mathrm{N}(\mu, \sigma^2)$“
i.i.d.	unabhängig und identisch verteilt (englisch: *independent and identically distributed*)

$\overline{X}_n$	Stichprobenmittel $\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ einer Stichprobe $X_1, \dots, X_n$
q_α	α-Quantil der Standardnormalverteilung
$q_{V,\alpha}$	α-Quantil einer Verteilung $V \in \{\mathrm{U}(A), \mathrm{N}(\mu, \sigma^2), \dots\}$
$\lvert\cdot\rvert, \langle\cdot,\cdot\rangle$	Euklidische Norm und euklidisches Skalarprodukt auf $\mathbb{R}^d$
$\lvert\cdot\rvert_p$	Vektor- oder Folgennorm für $p \in [0, \infty]$
$\lVert\cdot\rVert, \lVert\cdot\rVert_2$	Spektralnorm bzw. Frobeniusnorm einer Matrix
$\mathbb{R}^{m\times n}$	Menge der reellen Matrizen mit $m \in \mathbb{N}$ Zeilen und $n \in \mathbb{N}$ Spalten
E_n	Einheitsmatrix mit $n \in \mathbb{N}$ Spalten und Zeilen
$\mathcal{L}^p$, L^p	Raum der p-fach integrierbaren Funktionen bzw. Äquivalenzklassen
span	Spann, das heißt der lineare Unterraum aufgespannt durch die gegebenen Vektoren
Im	Bildbereich einer Funktion (englisch: *image*)
tr	Spur einer Matrix (englisch: *trace*)
rank	Rang einer Matrix
sgn	Vorzeichen einer reellen Zahl, d. h. $\operatorname{sgn}(x) = 1$, falls $x > 0$ und sonst -1.

Weitere Symbole und Begriffe aus der Wahrscheinlichkeitstheorie sind im Anhang zu finden.

Kapitel 1
Grundlagen der Statistik

In diesem Kapitel werden die grundlegenden Begriffe und Konzepte der Statistik eingeführt. Der Startpunkt ist die mathematische Formulierung eines statistischen Modells. Davon ausgehend werden wir Konstruktionsprinzipien für Parameterschätzer, Hypothesentests und Konfidenzbereiche kennenlernen. Nachdem diese in den einfachen Beispielen dieses Kapitels verstanden sind, können wir aus diesen Grundprinzipien heraus neue Methoden in den komplexeren Modellen späterer Kapitel entwickeln.

1.1 Das statistische Modell

Während die Wahrscheinlichkeitstheorie anhand eines gegebenen Modells die Eigenschaften der (zufälligen) Ereignisse untersucht, ist das Ziel der Statistik entgegengesetzt: Wie kann man aus den gegebenen Beobachtungen Rückschlüsse auf das Modell ziehen?

Beispiel 1.1 (Saskias Umfrage) Versetzen wir uns in die Lage der Studentin Saskia, die für ihre Masterarbeit eine Befragung unter den 20.000 Studentinnen und Studenten ihrer Uni durchführen möchte. Von ihnen möchte Saskia wissen, ob sie mit ihrem Studium zufrieden sind. Dazu führt sie eine Online-Umfrage durch, bei der es nur eine Frage mit zwei Antwortmöglichkeiten gibt – man ist zufrieden oder nicht. Sie lässt den Studierenden die Umfrage über den internen Mailverteiler der Uni zukommen. Alle erhalten die Umfrage, aber nur ein Teil von ihnen macht sich die Mühe, die Umfrage zu beantworten. Welche Schlüsse kann Saskia aus ihrer Umfrage ziehen?

Nehmen wir an, die gewählten Fächer sind zu gleichen Teilen zufriedenstellend und nicht zufriedenstellend, das heisst 50% sind zufrieden, und die Umfrage wird wahrheitsgemäß beantwortet. Würde nur eine Studentin antworten, so entspräche das Ergebnis nicht der Wahrheit, denn es könnte nur eine einhundertprozentige (Un)Zufriedenheit ermittelt werden. Eine sogenannte *Stichprobengröße* von 1 ist

also zu niedrig. Je mehr Leute antworten, desto höher ist die *Wahrscheinlichkeit*, dass der Anteil der positiven Antworten in der Umfrage näher an den wahren 50% liegt. Nur wenn alle 20.000 Studierenden antworten würden, könnte Saskia sicher sein, den wahren Wert gefunden zu haben. Im Allgemeinen wird das Ergebnis von Saskias Umfrage also von Unsicherheit behaftet sein. Um die Aussagekraft des Ergebnisses einschätzen zu können, benötigen wir statistische Methoden. Andersherum wäre es für Saskia gut zu wissen, wie viele Antworten sie benötigt, um ein aussagekräftiges Ergebnis zu erzielen.

Beispiel 1.2 (Wirksamkeit von Impfungen) Ein Corona-Impfstoff wurde in einer ersten Studie bei $n = 43\,500$ Teilnehmenden untersucht. Eine Hälfte bekam den Impfstoff verabreicht, die andere Hälfte als Kontrollgruppe ein Placebo. Innerhalb der darauffolgenden sechs Monate waren 4,03% in der Placebo-Gruppe an Corona erkrankt, aber nur 0,37% in der geimpften Gruppe. Dies wurde mit der Aussage „Das Erkrankungsrisiko wird durch Impfung um etwa 91% gesenkt“ veröffentlicht.

Dies sollten wir kritisch hinterfragen. Könnte es vielleicht nur Zufall gewesen sein, dass weniger Geimpfte erkrankten, ist der Unterschied also eventuell statistisch nicht signifikant? Mathematisch: wie wahrscheinlich ist eine solche Abweichung der Prozentzahlen bei gleicher Erkrankungswahrscheinlichkeit pro Patient in beiden Gruppen? Eine weitergehende Frage wäre, ob wichtige *Kovariablen* wie Geschlecht oder Alter die Erkrankungswahrscheinlichkeit in den beiden Gruppen beeinflussen, sodass sich vielleicht die Effizienz der Impfung detaillierter beschreiben lässt oder das Ergebnis sogar auf eine inhomogene Zusammensetzung der Gruppen zurückzuführen ist.

Beispiel 1.3 (Happiness-Score) Im *World Happiness Report* der Vereinten Nationen (UN) wird jährlich ein Glücksindex bzw. ein *Happiness-Score* zur Lebenszufriedenheit der Bevölkerung aus verschiedenen Ländern bestimmt. Wir wollen die Abhängigkeit des Happiness-Scores aus dem Jahr 2019 vom jeweiligen Pro-Kopf-Bruttoinlandsprodukt untersuchen und betrachten hierzu das *Modell*

$$Y_i = aX_i + b + \varepsilon_i, \quad i = 1, \dots, 156, \tag{1.1}$$

wobei die zufälligen Störgrößen ε_i Unsicherheiten bei den Bevölkerungsumfragen sowie länderspezifische ökonomische/geographische/soziale Einflüsse etc. modellieren. Analog können wir versuchen, die Lebenserwartung bei guter Gesundheit durch das Bruttoinlandsprodukt zu erklären, siehe Abbildung 1.1. Plausible Annahmen an das Modell sind:

1. (ε_i) sind unabhängig (näherungsweise),
2. (ε_i) sind identisch verteilt,
3. $\mathbb{E}[\varepsilon_i] = 0$ (kein systematischer Fehler),
4. ε_i normalverteilt (wegen des zentralen Grenzwertsatzes).

Naheliegende *Ziele* bzw. *Fragestellungen* sind:

1. Für welche Parameter a, b beschreibt das Modell (1.1) die gegebenen Daten am besten? Ein mögliches Schätzverfahren ist der *Kleinste-Quadrate-Schätzer*

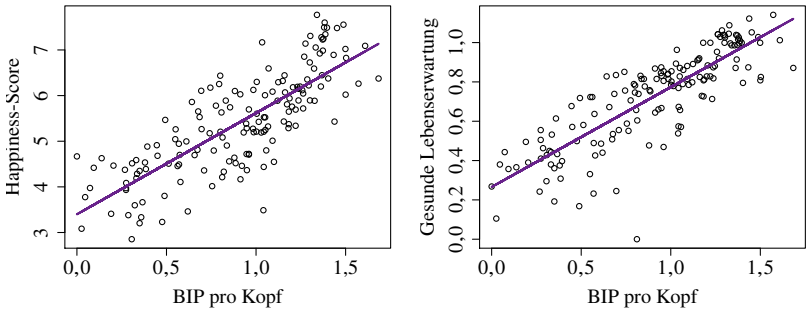

Abb. 1.1 Happiness-Score (*links*) und Lebenserwartung bei guter Gesundheit (*rechts*) in Abhängigkeit von Bruttoinlandsprodukt pro Kopf aus 156 Ländern auf Grundlage des World Happiness Reports von 2019. Die resultierenden Regressionsgeraden aus Beispiel 1.3 sind violett eingezeichnet.

$$(\widehat{a}, \widehat{b}) := \operatorname*{arg\,min}_{a,b\in\mathbb{R}} \sum_{i=1}^{n} (Y_i - aX_i - b)^2$$

(wir minimieren die Summe der quadrierten Residuen), den wir in Kapitel 2 kennenlernen werden. Mit $\widehat{a}$ und $\widehat{b}$ erhalten wir die *Regressionsgerade*

$$y = \widehat{a}x + \widehat{b}.$$

2. Sind die Modellannahmen erfüllt? Hierzu können Histogramme, Boxplots und Quantil-Quantil-Diagramme (QQ-Plots) der Residuen $Y_i - \widehat{a}X_i - \widehat{b}$ verwendet werden.
3. Wenn wir die Verteilung von $\widehat{a}$ kennen (Verteilungsannahme an ε nötig!), können wir Intervalle der Form $I = [\widehat{a} - c, \widehat{a} + c]$ für $c > 0$ konstruieren, sodass der tatsächliche Parameter a mit vorgegebener Wahrscheinlichkeit in I liegt.
4. Wie kann man *testen*, ob das Bruttoinlandsprodukt tatsächlich einen Effekt auf die Lebenszufriedenheit hat, das heißt gilt die Hypothese $H_0 : a = 0$ oder kann sie verworfen werden? Ein mögliches Verfahren ist, die Hypothese zu verwerfen, falls $|\widehat{a}| > c$ für einen kritischen Wert $c > 0$. Um einen sinnvollen Wert zu bestimmen, benötigen wir wieder Verteilungsannahmen an die Fehler (ε_i).

Im Laufe der folgenden Kapitel werden solche und ähnliche Fragen beantwortet, doch vorher müssen wir statistische Modelle formal einführen.

Definition 1.4 Sei $\mathcal{X}$ die Menge aller möglichen Beobachtungen, der sogenannte **Stichprobenraum**. Ein messbarer Raum $(\mathcal{X}, \mathcal{F})$, versehen mit einer Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ von Wahrscheinlichkeitsmaßen mit einer beliebigen Parametermenge $\Theta \neq \emptyset$, heißt **statistisches Experiment** oder **statistisches Modell**.

Durch die Wahl der Familie von Wahrscheinlichkeitsmaßen im statistischen Modell wird zielgerichtet eine Veränderlichkeit oder Unsicherheit des Modells formalisiert. Je nachdem, welche statistische Aufgabe gegeben ist, wählt man statistische

Methoden, die wiederum Rückschlüsse auf das Modell erlauben. Zuletzt werden die im Experiment gewonnenen Daten in die Methoden eingesetzt, sodass man Informationen über den zugrunde liegenden Parameter ϑ gewinnt, unter dem die Daten generiert wurden.

Definition 1.5 Es sei $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $(\mathcal{X}, \mathcal{F})$ ein Messraum. Jede $(\mathcal{A}, \mathcal{F})$-messbare Abbildung $X : \Omega \to \mathcal{X}$ heißt **Beobachtung** oder **Statistik** mit Werten in $(\mathcal{X}, \mathcal{F})$ und induziert das statistische Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}^X_\vartheta)_{\vartheta\in\Theta})$. Sind die Beobachtungen $X_1, ..., X_n, n \in \mathbb{N}$, für jedes $\vartheta \in \Theta$ unabhängig und identisch verteilt (kurz i.i.d. für *independent and identically distributed*), so nennt man $X_1, ..., X_n$ eine **mathematische Stichprobe**.

In der Statistik konzentrieren wir uns auf die Beobachtungen und ihre Verteilung. Entsprechend der vorangegangenen Definition betrachten wir also das statistische Experiment $(\mathcal{X}, \mathcal{F}, (\mathbb{P}^X_\vartheta)_{\vartheta\in\Theta})$, in dem die Werte von $X\colon \Omega \to \mathcal{X}$ liegen. Das statistische Modell $(\Omega, \mathcal{A}, \mathbb{P}_{\vartheta\in\Theta})$, in dem die Urbilder von X liegen, bleibt häufig unspezifiziert. Dies ist vollkommen analog zur Rolle der Zufallsvariablen in der Wahrscheinlichkeitstheorie. Der Einfachheit halber schreiben wir $X \sim \mathbb{P}_\vartheta$ und nicht $X \sim \mathbb{P}^X_\vartheta$.

Bemerkung 1.6 Für $n \in \mathbb{N}$ sei $X_1, \ldots, X_n$ eine mathematische Stichprobe mit Werten in $\mathcal{X}$ und Randverteilung $X_1 \sim \mathbb{P}_\vartheta$ mit Parameter $\vartheta \in \Theta$. Dann ist der Stichprobenvektor $(X_1, \ldots, X_n)$ gemäß dem Produktmaß $\mathbb{P}^{\otimes n}_\vartheta = \bigotimes_{i=1}^n \mathbb{P}_\vartheta$ auf $(\mathcal{X}^n, \mathcal{F}^{\otimes n})$ verteilt, siehe Definition A.20.

Beispiel 1.7 (Statistisches Modell, mathematische Stichprobe) Kommen wir auf Beispiel 1.1 zurück. Die Studentin Saskia bekommt n Anworten auf ihre Umfrage zur Studienzufriedenheit unter $N = 20.000$ Studentinnen und Studenten. Da wir davon ausgehen, dass eine bereits befragte Person nicht noch einmal an der Umfrage teilnimmt, bietet sich die hypergeometrische Verteilung $\mathrm{Hyp}(N, M, n)$ zur Modellierung an, wobei die Anzahl M der zufriedenen Studierenden der unbekannte Parameter ist. Falls deutlich weniger als N Antworten eingehen, können wir die hypergeometrische Verteilung durch die Binomialverteilung beziehungsweise mittels eines Vektors von Bernoulli-verteilten Zufallsvariablen approximieren (siehe Bemerkung A.48).

Folglich kann Saskia die eingegangenen Antworten $X_1, \ldots, X_n$ als unabhängige Bernoulli-verteilte Zufallsvariablen modellieren, wobei 0 für „unzufrieden“ und 1 für „zufrieden“ steht. Der Erfolgs- bzw. Zufriedenheitsparameter $\vartheta = M/N \in [0, 1]$ ist gerade der Anteil der zufriedenen Studentinnen und Studenten. Wir erhalten das statistische Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit

$$\mathcal{X} = \{0, 1\}^n, \quad \mathcal{F} = \mathcal{P}(\mathcal{X}), \qquad \mathbb{P}_\vartheta = \mathrm{Ber}(\vartheta)^{\otimes n} \quad \text{und} \quad \Theta = [0, 1].$$

Beispiel 1.8 (Statistisches Modell II) Wir modellieren nun die Situation aus Beispiel 1.2. Wir notieren den Versuchsausgang als einen 0-1-Vektor x der Länge n, wobei die ersten $n/2$ Einträge die Ergebnisse in Gruppe 1 (Placebo) und die letzten $n/2$ Einträge die Ergebnisse in Gruppe 2 (geimpft) kodieren mit '1'='erkrankt', '0'='nicht erkrankt'. Wir wählen also den Stichprobenraum $\mathcal{X} = \{0, 1\}^n$ mit der

σ-Algebra $\mathcal{F} = \mathcal{P}(\mathcal{X})$. Wenn wir annehmen, dass alle Teilnehmenden in der Studie unabhängig voneinander erkranken oder nicht und die Erkrankungswahrscheinlichkeit $p_1 \in [0,1]$ in Gruppe 1 sowie $p_2 \in [0,1]$ in Gruppe 2 beträgt, so sind die n einzelnen Versuchsausgänge unabhängig Bernoulli-verteilt mit Parametern p_1 bzw. p_2. Unter Verwendung des Produktmaßes und mit $\Theta = [0,1]^2$ ergibt sich daher das statistische Modell $(\mathcal{X}, \mathcal{F}, (\mathrm{Ber}(p_1)^{\otimes n/2} \otimes \mathrm{Ber}(p_2)^{\otimes n/2})_{(p_1,p_2)\in\Theta})$. Einfacher verständlich ist die Beschreibung, dass wir einen $\mathcal{X}$-wertigen Zufallsvektor beobachten, dessen Koordinaten $X_1, \dots, X_n$ unabhängig sind mit $X_i \sim \mathrm{Ber}(p_1)$ für $i = 1, \dots, n/2$ und $X_i \sim \mathrm{Ber}(p_2)$ für $i = n/2+1, \dots, n$, wobei $(p_1, p_2) \in \Theta$ der unbekannte Parameter ist.

Typische Statistiken in dem Modell sind $S_1 = \sum_{j=1}^{n/2} X_j \sim \mathrm{Bin}(n/2, p_1)$ und $S_2 = \sum_{j=n/2+1}^{n} X_j \sim \mathrm{Bin}(n/2, p_2)$. Es ist intuitiv, dass durch diese Zusammenfassung der Erkrankungszahlen in den einzelnen Gruppen keine statistische Information verloren geht.

Wie bereits diese einfachen Beispiele zeigen, bilden Modelle nicht die gesamte Wirklichkeit ab, sondern dienen immer auch der Vereinfachung. Folgendes Zitat sollte man sich daher immer vor Augen halten: *„Alle Modelle sind falsch, doch manche sind nützlich."* (George Box).

Viele statistische Fragestellungen kann man einem der drei Grundprobleme *Parameterschätzung*, *Hypothesentests* und *Konfidenzmengen* zuordnen. Diese werden im Folgenden eingeführt und später weiter vertieft. Weitere wichtige statistische Aufgaben wie Vorhersage oder Klassifikation werden sich durch leichte Modifikationen ergeben.

1.2 Parameterschätzung

Unser erstes Ziel ist, aufgrund von Beobachtungen den unbekannten Parameter $\vartheta \in \Theta$ zu schätzen. Wir sagen „schätzen" und nicht „bestimmen", weil wir in den meisten Fällen mit Wahrscheinlichkeiten, Fehlern und unvollständigen Stichproben konfrontiert sind, wie in Beispiel 1.7 mit Saskia, in dem nur eine zufällige Teilmenge von $n \leqslant 20.000$ Menschen Saskias Umfrage beantwortet. Wir können also nicht erwarten, den „wahren" zugrunde liegenden Parameter, sollte es ihn überhaupt geben, zu finden. Stattdessen soll auf Grundlage der Daten ein Parameterwert gefunden werden, mit dem das resultierende Modell die Daten möglichst gut beschreibt.

1.2.1 Konstruktionsprinzipien

Ein Schätzer ist eine Funktion, manchmal auch *Schätzfunktion* genannt, die jeder realisierten Beobachtung $x \in \mathcal{X}$ einen Parameter $\vartheta \in \Theta$ zuordnet. Wenn wir auch die Schätzung von *abgeleiteten Parametern* $\rho(\vartheta)$ zulassen, wobei $\rho\colon \Theta \to \mathbb{R}^p$, führt dies auf folgende Defintion:

Definition 1.9 Sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\rho\colon \Theta \to \mathbb{R}^p$ ein (abgeleiteter) p-dimensionaler Parameter für $p \in \mathbb{N}$. Ein **Schätzer** ist eine messbare Abbildung $\widehat{\rho}\colon \mathcal{X} \to \mathbb{R}^p$. Gilt $\mathbb{E}_\vartheta[\widehat{\rho}] = \rho(\vartheta)$ für alle $\vartheta \in \Theta$, so heißt $\widehat{\rho}$ **unverzerrt** oder **erwartungstreu** (englisch: *unbiased*). Der Erwartungswert eines Zufallsvektors ist dabei koordinatenweise definiert.

Da für $\Theta \subseteq \mathbb{R}^p$ obige Definition die Identität $\rho(\vartheta) = \vartheta$ einschließt, ist die Betrachtung von abgeleiteten Parametern etwas allgemeiner als nur die Untersuchung von Schätzern des gesamten Parameters.

Beispiel 1.10 (Schätzer, abgeleiteter Parameter – a) Wir setzen Saskias Beispiel 1.7 fort und betrachten die mathematische Stichprobe $X_1, ..., X_n \sim \mathrm{Ber}(\vartheta)$ mit Parameter $\vartheta \in [0, 1]$. Da der Parameterraum nur eindimensional ist, betrachten wir die Identität $\rho(\vartheta) = \vartheta$. Als Schätzer wählen wir den Anteil der zufriedenen Studierenden unter allen Antworten, das heißt

$$\widehat{\rho}_n := \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1.2}$$

Würden wirklich alle 20.000 Studentinnen und Studenten an der Umfrage teilnehmen, würde $\widehat{\rho}_n$ exakt den Anteil der zufriedenen Studierenden ergeben. Unter unserer Modellannahme ist $\widehat{\rho}_n$ aber auch sonst sinnvoll: Es gilt

$$\mathbb{E}_\vartheta[\widehat{\rho}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\vartheta[X_i] = \vartheta \quad \text{und} \quad \mathrm{Var}_\vartheta(\widehat{\rho}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_\vartheta(X_i) = \frac{\vartheta(1-\vartheta)}{n}.$$

Das heißt, der Schätzer $\widehat{\rho}_n$ ist erwartungstreu, und für größer werdenden Stichprobenumfang n sinkt seine Varianz. $\widehat{\rho}_n$ konzentriert sich also für wachsendes n um das wahre ϑ. Beide Eigenschaften sind wünschenswert. Wir werden später Beispiele sehen, bei denen wir jedoch eine Abweichung des Erwartungswerts $\mathbb{E}[\widehat{\rho}_n]$ vom wahren Parameter, man spricht dann von einem *Bias*, in Kauf nehmen, um die Varianz zusätzlich zu reduzieren.

Beispiel 1.11 (Schätzer, abgeleiteter Parameter – b) Es sei $X_1, ..., X_n$ eine normalverteilte mathematische Stichprobe, das heißt $X_1 \sim \mathrm{N}(\mu, \sigma^2)$. Der unbekannte Parameter ist

$$\vartheta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty) =: \Theta.$$

Interessieren wir uns nur für den Erwartungswert μ, dann müssen wir von dem zweidimensionalen Parameter ϑ nur noch die erste Komponente schätzen. Hier kommt nun die den Parameter ableitende Funktion ρ ins Spiel, die ϑ auf das reduziert, was uns interessiert. Wir definieren also $\rho\colon \Theta \to \mathbb{R}, (\mu, \sigma^2) \mapsto \mu$. Als Schätzer für $\rho(\vartheta)$ verwenden wir wieder das arithmetische Mittel $\overline{X}_n$, auch *Stichprobenmittel* genannt, das hier ebenfalls erwartungstreu ist (dies zu zeigen sei den Leserinnen und Lesern überlassen)

$$\widehat{\rho}_n := \overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Um die Güte eines Schätzers zu bestimmen, werden *Verlustfunktionen* und deren zugehöriges *Risiko* verwendet. Diese messen den (erwarteten) Abstand zwischen geschätztem und wahrem Parameter.

Definition 1.12 Eine Funktion $\ell\colon \Theta\times\mathbb{R}^p \to \mathbb{R}_+$ heißt **Verlustfunktion**, falls $\ell(\vartheta,\cdot)$ für jedes $\vartheta\in\Theta$ messbar ist. Der erwartete Verlust $R(\vartheta,\widehat{\rho}) := \mathbb{E}_\vartheta[\ell(\vartheta,\widehat{\rho})]$ eines Schätzers $\widehat{\rho}$ heißt **Risiko**. Typische Verlustfunktionen sind

1. der **0-1-Verlust** $\ell(\vartheta,r) = \mathbb{1}_{\{r\neq\rho(\vartheta)\}}$,
2. der **absolute Verlust** $\ell(\vartheta,r) = |r-\rho(\vartheta)|$ (euklidischer Abstand im $\mathbb{R}^p$) sowie
3. der **quadratische Verlust** $\ell(\vartheta,r) = |r-\rho(\vartheta)|^2$.

Das Risiko bezüglich des quadratischen Verlusts nennt man auch *quadratisches Risiko* bzw. *mittleren quadratischen Fehler* (englisch: *mean squared error*, kurz: MSE). Dieser lässt sich in Bias und Varianz zerlegen. Wir erinnern hierfür daran, dass der Erwartungswert eines Zufallsvektors $X = (X_1,\dots,X_d)^\top \in \mathbb{R}^d$ koeffizientenweise gebildet wird, das heißt $\mathbb{E}[X] := (\mathbb{E}[X_1],\dots,\mathbb{E}[X_d])^\top$, und die Varianz durch

$$\operatorname{Var}(X) := \mathbb{E}\big[|X-\mathbb{E}[X]|^2\big] = \sum_{i=1}^d \mathbb{E}\big[(X_i-\mathbb{E}[X_i])^2\big] = \sum_{i=1}^d \operatorname{Var}(X_i)$$

definiert wird.

Lemma 1.13 (Bias-Varianz-Zerlegung, statistischer Pythagoras) *Es sei $(\mathcal{X},\mathcal{F},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\widehat{\rho}\colon \mathcal{X}\to\mathbb{R}^p$ ein Schätzer des Parameters $\rho(\vartheta)$ mit $\mathbb{E}_\vartheta[|\widehat{\rho}|^2]<\infty$ für alle $\vartheta\in\Theta$. Dann gilt für das quadratische Risiko*

$$\mathbb{E}_\vartheta\big[|\widehat{\rho}-\rho(\vartheta)|^2\big] = \operatorname{Var}_\vartheta(\widehat{\rho}) + |\mathbb{E}_\vartheta[\widehat{\rho}]-\rho(\vartheta)|^2 \quad \textit{für alle } \vartheta\in\Theta.$$

*Die mittlere Abweichung des Schätzers vom wahren Wert $\mathbb{E}_\vartheta[\widehat{\rho}]-\rho(\vartheta)$ wird **Bias** genannt.*

Beweis Für Vektoren $a,b\in\mathbb{R}^p$ gilt $|a+b|^2 = |a|^2+2\langle a,b\rangle+|b|^2$. Aus der Linearität des Erwartungswerts folgt für alle $\vartheta\in\Theta$:

$$\begin{aligned}
\mathbb{E}_\vartheta\big[|\widehat{\rho}-\rho(\vartheta)|^2\big] &= \mathbb{E}_\vartheta\big[|\widehat{\rho}-\mathbb{E}_\vartheta[\widehat{\rho}]+\mathbb{E}_\vartheta[\widehat{\rho}]-\rho(\vartheta)|^2\big]\\
&= \mathbb{E}_\vartheta[|\widehat{\rho}-\mathbb{E}_\vartheta[\widehat{\rho}]|^2] + 2\mathbb{E}_\vartheta[\langle\widehat{\rho}-\mathbb{E}_\vartheta[\widehat{\rho}],\mathbb{E}_\vartheta[\widehat{\rho}]-\rho(\vartheta)\rangle]\\
&\qquad + |\mathbb{E}_\vartheta[\widehat{\rho}]-\rho(\vartheta)|^2\\
&= \operatorname{Var}_\vartheta(\widehat{\rho}) + 2\langle\mathbb{E}_\vartheta[\widehat{\rho}]-\mathbb{E}_\vartheta[\widehat{\rho}],\mathbb{E}_\vartheta[\widehat{\rho}-\rho(\vartheta))]\rangle\\
&\qquad + |\mathbb{E}_\vartheta[\widehat{\rho}]-\rho(\vartheta)|^2\\
&= \operatorname{Var}_\vartheta(\widehat{\rho}) + |\mathbb{E}_\vartheta[\widehat{\rho}]-\rho(\vartheta)|^2
\end{aligned}$$

Damit ist die Behauptung gezeigt. □

Beispiel 1.14 (Quadratisches Risiko) Im Saskia-Beispiel 1.10 betrachten wir nun den Schätzer $\widetilde{\rho}_n := (\sum_{i=1}^n X_i + 1)/(n+2)$. Dieser hat den Bias

$$\mathbb{E}_\vartheta[\widetilde{\rho}_n] - \vartheta = \frac{1-2\vartheta}{n+2}$$

und die Varianz

$$\mathrm{Var}_\vartheta(\widetilde{\rho}_n) = \frac{n\vartheta(1-\vartheta)}{(n+2)^2}.$$

Damit gilt $\mathbb{E}_\vartheta[\widetilde{\rho}_n] = \rho(\vartheta)$ ausschließlich für $\vartheta = 1/2$, sodass $\widetilde{\rho}_n$ kein unverzerrter Schätzer ist. Er besitzt aber einen kleineren quadratischen Fehler als $\widehat{\rho}_n$, wenn $|\vartheta - 1/2| \leqslant 1/\sqrt{8}$.

Obwohl wir nur wenig Asymptotik behandeln, also das Verhalten der Schätzer bei Stichprobenumfängen $n \to \infty$, seien noch zwei weitere wichtige Grundbegriffe erwähnt, die ebenfalls als Gütekriterien eines Schätzers dienen.

Definition 1.15 Für jedes $n \in \mathbb{N}$ sei $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathbb{P}_\vartheta$ eine mathematische Stichprobe. Dann heißt eine Folge von Schätzern $\widehat{\rho}_n = \widehat{\rho}_n(X_1, \ldots, X_n)$ des abgeleiteten Parameters $\rho(\vartheta)$ **konsistent**, falls für alle $\vartheta \in \Theta$

$$\widehat{\rho}_n \xrightarrow{\mathbb{P}_\vartheta} \rho(\vartheta) \quad \text{für} \quad n \to \infty.$$

Falls $\mathbb{E}_\vartheta[|\widehat{\rho}_n|^2] < \infty$ für alle $\vartheta \in \Theta$ und unter jedem $\mathbb{P}_\vartheta$

$$r_n^{-1}(\widehat{\rho}_n - \rho(\vartheta)) \xrightarrow{d} \mathrm{N}(\mu_\vartheta, \sigma_\vartheta^2) \quad \text{für} \quad n \to \infty$$

gilt, nennen wir $\widehat{\rho}_n$ **asymptotisch normalverteilt** mit Konvergenzrate $r_n \to \infty$, asymptotischem Bias $\mu_\vartheta \in \mathbb{R}$ und asymptotischer Varianz σ_ϑ^2.

Warum ist es sinnvoll, einen Schätzer (genauer eine Folge von Schätzern) in diesem Fall „konsistent“ zu nennen? „Konsistent“ bedeutet in der Philosophie, genauer in der Logik, „zusammenhängend in der Gedankenführung“ und tatsächlich stellt obige Definition einen (asymptotischen) Zusammenhang zwischen dem Schätzer $\widehat{\rho}_n$ und dem zu schätzenden Parameter $\rho(\vartheta)$ her.

Ein stärkeres Kriterium beschreibt die asymptotische Verteilung eines Schätzers. Aufgrund des zentralen Grenzwertsatzes sind viele Schätzer asymptotisch normalverteilt, so auch in Beispiel 1.10. Daher kommt der Untersuchung von statistischen Modellen unter Normalverteilungsannahme eine besondere Bedeutung zu.

Bisher haben wir unsere Beispielschätzer eher intuitiv gefunden. Stattdessen wollen wir nun zwei wichtige Konstruktionsprinzipien einführen, die *Momentenmethode* und die *Maximum-Likelihood-Schätzung*, die in einem sehr allgemeinen Rahmen eingesetzt werden können. Beide Ansätze liefern häufig gute Schätzmethoden.

Wir beginnen mit der *Momentenmethode*. Wie der Name suggeriert, wollen wir die Momente von Zufallsvariablen verwenden, um den unbekannten Parameter zu schätzen. Betrachten wir ein statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta \subseteq \mathbb{R}^p$, so

hängen die Momente $m_k(\vartheta) := \mathbb{E}_\vartheta[X^k]$ einer Beobachtung $X \sim \mathbb{P}_\vartheta$ im Allgemeinen vom Parameter ϑ ab. Die Idee ist nun, die p Komponenten von ϑ zu rekonstruieren, indem man ein Gleichungssystem mit p Gleichungen basierend auf den (ersten) p Momenten $m_k(\vartheta)$ löst. Für den Fall, dass dieses Gleichungssystem eine eindeutige Lösung besitzt, muss man nur noch die Momente schätzen. Haben wir n i.i.d. Beobachtungen $X_1, \dots, X_n \sim \mathbb{P}_\vartheta$ zur Verfügung, konvergiert nach dem Gesetz der großen Zahlen das k-te *Stichprobenmoment* $\widehat{m}_k := \frac{1}{n}\sum_{j=1}^n X_j^k$ für $n \to \infty$ gegen $m_k(\vartheta)$.

Methode 1.16 (Momentenmethode, Momentenschätzer) Sei $\vartheta \in \mathbb{R}^p$ der unbekannte Parameter einer mathematischen Stichprobe $X_1, \dots, X_n$ reeller Zufallsvariablen. Für $k \in \mathbb{N}$ bezeichne

$$m_k := m_k(\vartheta) := \mathbb{E}_\vartheta[X_1^k] \qquad \text{bzw.} \qquad \widehat{m}_k := \frac{1}{n}\sum_{j=1}^n X_j^k$$

das k-te Moment von X_1, sofern es existiert ($\mathbb{E}_\vartheta[|X_1|^k] < \infty$), bzw. das k-te Stichprobenmoment. Für gewählte Indizes $1 \leqslant k_1 < k_2 < \cdots < k_p$ mit $\mathbb{E}_\vartheta[|X_1|^{k_p}] < \infty$ ergibt sich der (klassische) **Momentenschätzer** $\widehat{\vartheta}$ von ϑ als die Lösung der p Gleichungen

$$m_{k_1}(\widehat{\vartheta}) = \widehat{m}_{k_1}, \quad m_{k_2}(\widehat{\vartheta}) = \widehat{m}_{k_2}, \quad \dots \quad , m_{k_p}(\widehat{\vartheta}) = \widehat{m}_{k_p}. \tag{1.3}$$

In vielen Fällen ist das Gleichungssystem mit p Gleichungen eindeutig lösbar, aber es gibt Beispiele, in denen keine eindeutig bestimmte Lösung existiert. Dies und die resultierenden Eigenschaften, beispielsweise das quadratische Risiko, bestimmen die Wahl der Exponenten $k_1, \dots, k_p$. Ganz allgemein stellt sich die Frage, wann die Folge der Momente $(m_k)_{k\geqslant 1}$, falls sie existiert, die Verteilung $\mathbb{P}_\vartheta$ eindeutig bestimmt. Eine mögliche Antwort auf dieses sogenannte *Momentenproblem* liefert der Satz A.30. Über die Monome X_1^k hinausgehend, lässt sich die Momentenmethode auch direkt auf Erwartungswerte allgemeinerer Funktionale $f(X_1)$ oder auf zentrierte Momente wie die Varianz $\mathrm{Var}_\vartheta(X_1)$ mit empirischer Varianz $\widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \widehat{m}_1)^2$ verallgemeinern.

Beispiel 1.17 (Momentenschätzer) Sei $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$. Dann ist $m_1 = \mathbb{E}_{\mu,\sigma^2}[X_1] = \mu$ und $m_2 = \mathbb{E}_{\mu,\sigma^2}[X_1^2] = \mathrm{Var}_{\mu,\sigma^2}(X_1) + \mathbb{E}_{\mu,\sigma^2}[X_1]^2 = \sigma^2 + \mu^2$. Wir erhalten das Gleichungssystem

$$\widehat{\mu} = \frac{1}{n}\sum_{j=1}^n X_j \quad \text{und} \quad \widehat{\sigma}^2 + \widehat{\mu}^2 = \frac{1}{n}\sum_{j=1}^n X_j^2.$$

Mithilfe des Stichprobenmittels $\overline{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$ erhalten wir die Lösung

$$\widehat{\mu} = \overline{X}_n, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^n X_j^2 - \Big(\frac{1}{n}\sum_{j=1}^n X_j\Big)^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \overline{X}_n)^2.$$

Kurzbiografie (Karl und Egon Pearson) Karl Pearson wurde 1857 in London geboren. Er studierte in Cambridge, Heidelberg und Berlin Mathematik, Physik, Deutsche Literatur, Biologie, Philosophie, Jura und Geschichte. Den Großteil seiner Laufbahn verbrachte er am Londoner University College. Unter seinen Arbeiten in zahlreichen wissenschaftlichen Disziplinen sind die „Mathematical Contributions to the Theory of Evolution“ sein wertvollster Beitrag zur Statistik. Unter anderem entwickelte er darin die *Momentenmethode*, den Korrelationskoeffizienten (nach einer ähnlichen Idee von Francis Galton) und den χ^2-Test der statistischen Signifikanz. Karl Pearson starb 1936 in Coldharbour, Surrey.

Auch Karl Pearsons Sohn Egon Sharpe Pearson (geboren 1895 in Hampstead, gestorben 1980 in Sussex) wurde einer der führenden britischen Statistiker. Egon Pearson folgte seinem Vater an das University College London, und als Karl Pearson in den Ruhstand ging, wurde sein Lehrstuhl auf seinen Sohn Egon sowie Ronald Aylmer Fisher aufgeteilt. Egon Pearson ist insbesondere für das *Neyman-Pearson-Lemma* bekannt.

Für die zweite Konstruktionsmethode von Schätzern benötigen wir etwas mehr Struktur, die wir auch im weiteren Verlauf immer wieder aufgreifen. Wir fordern Absolutstetigkeit der Wahrscheinlichkeitsmaße (siehe Definition A.32), sodass uns der Satz von Radon-Nikodym (siehe Satz A.33) Dichten liefert.

Definition 1.18 Ein statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt **dominiert**, falls es ein σ-endliches Maß μ gibt, sodass $\mathbb{P}_\vartheta$ absolutstetig bezüglich μ ist für alle $\vartheta \in \Theta$. Für jedes $\vartheta \in \Theta$ bezeichnen wir die Radon-Nikodym-Dichte von $\mathbb{P}_\vartheta$ bezüglich μ mit

$$L(\vartheta, x) := \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x), \quad \vartheta \in \Theta, x \in \mathcal{X},$$

und nennen sie **Likelihood-Funktion** oder kurz **Likelihood**. Äquivalent gilt $L(\vartheta, x) = p_\vartheta(x)$ immer, wenn die μ-Dichte p_ϑ von $\mathbb{P}_\vartheta$ für alle $\vartheta \in \Theta$ existiert.

Damit ist $x \mapsto L(\vartheta, x)$ eindeutig bis auf eine μ-Nullmenge, die dann auch eine $\mathbb{P}_{\vartheta'}$ Nulllmenge für jedes $\vartheta' \in \Theta$ ist. Da wir aufgrund von gegebenen Daten X den zugrunde liegenden Parameter ϑ schätzen wollen, wird in der Statistik die Likelihood-Funktion als durch x parametrisierte Funktion in ϑ aufgefasst. Somit ist $\vartheta \mapsto L(\vartheta) = L(\vartheta, X)$ eine zufällige Funktion, der sogenannte Likelihood-Prozess.

Beispiel 1.19 (Likelihood-Funktionen)

1. Betrachte das statistische Modell $(\mathbb{R}, \mathcal{B}(\mathbb{R}), (\mathrm{N}(\mu, \sigma^2))_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)})$. Als dominierendes Maß kann das Lebesgue-Maß gewählt werden, sodass die Likelihood-Funktion durch die Lebesgue-Dichte der Normalverteilung gegeben ist: $L((\mu, \sigma^2), x) = (2\pi\sigma^2)^{-1/2} e^{(-(x-\mu)^2/(2\sigma^2))}, x \in \mathbb{R}$.
2. Jedes statistische Modell auf dem Stichprobenraum $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ oder allgemeiner auf einem abzählbaren Raum $(\mathcal{X}, \mathcal{P}(\mathcal{X}))$ ist vom Zählmaß dominiert. Damit ist die Likelihood-Funktion durch die Zähldichte gegeben.
3. Ist $\Theta = \{\vartheta_1, \vartheta_2, ...\}$ abzählbar, so ist $\mu = \sum_i c_i \mathbb{P}_{\vartheta_i}$ mit $c_i > 0$ und $\sum_i c_i = 1$ ein dominierendes Maß.

Methode 1.20 (Maximum-Likelihood-Schätzer) Für ein dominiertes statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit Likelihood-Funktion $L(\vartheta, x)$ heißt eine Statistik $\widehat{\vartheta}\colon \mathcal{X} \to \Theta$ (Θ trage eine σ-Algebra) **Maximum-Likelihood-Schätzer** (englisch: *maximum likelihood estimator*, kurz: MLE), falls gilt:

$$L(\widehat{\vartheta}(x), x) = \sup_{\vartheta\in\Theta} L(\vartheta, x) \quad \text{für } \mu\text{-f.a. } x \in \mathcal{X}$$

Die Grundidee der Maximum-Likelihood-Schätzung ist im Fall diskreter Beobachtungen intuitiv: Wir wählen denjenigen Parameter $\widehat{\vartheta}(x)$ aus, unter dem die Beobachtung $X = x$ die größte Wahrscheinlichkeit besitzt. Als einfaches Beispiel betrachte die Beobachtung $X \sim \mathrm{Bin}(2, \vartheta)$ mit $\vartheta \in [0, 1]$ unbekannt. Mit dem Zählmaß als dominierendem Maß gilt dann $L(\vartheta, 0) = (1 - \vartheta)^2$, $L(\vartheta, 1) = 2\vartheta(1 - \vartheta)$, $L(\vartheta, 2) = \vartheta^2$. Der MLE ist also eindeutig gegeben durch $\widehat{\vartheta}(0) = 0$, $\widehat{\vartheta}(1) = 1/2$, $\widehat{\vartheta}(2) = 1$, das heißt $\widehat{\vartheta} = X/2$.

Im allgemeinen Fall, zum Beispiel bei einer Likelihood-Funktion bezüglich des Lebesgue-Maßes, führt das Maximum-Likelihood-Prinzip sehr häufig, aber durchaus nicht immer, zu vernünftigen Schätzern. Es sei der Leserin ans Herz gelegt, nachzuweisen, dass die MLE-Eigenschaft nicht vom dominierenden Maß μ abhängt (Aufgabe 1.3).

Kurzbiografie (Ronald Aylmer Fisher) Sir Ronald Aylmer Fisher wurde 1890 in London geboren. Von 1909 bis 1912 studierte er in Cambridge. Fisher lehrte zunächst an verschiedenen öffentlichen Schulen, bevor er anfing an der Rothamsted Experimental Station zu arbeiten, wo er in den Feldern Genetik, Evolution und Statistik forschte. Nach Pearsons Ausscheiden übernahm er seine Professur am University College London. 1943 besetzte er dann den Lehrstuhl für Genetik in Cambridge. Fisher prägte die theoretische Basis der Statistik und trug maßgeblich zu ihrem heutigen Wesen bei. Unter anderem führte er die Likelihood-Funktionen und den *Maximum-Likelihood-Schätzer*, die *analysis of variance* (ANOVA), das Konzept der Suffizienz, die nach ihm benannte Fisher-Information und vieles mehr ein. Ronald Aylmer Fisher starb 1962 in Adelaide, Australien.

Beispiel 1.21 (Maximum-Likelihood-Schätzer – a) Betrachten wir wieder eine mathematische Stichprobe $X_1, \ldots, X_n$ normalverteilter Zufallsvariablen. Dann ist das statistische Modell $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathbb{P}^{\otimes n}_{\mu,\sigma^2})_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)})$ mit $\mathbb{P}_{\mu,\sigma^2} = \mathrm{N}(\mu, \sigma^2)$ vom Lebesgue-Maß auf $\mathbb{R}^n$ dominiert mit Likelihood-Funktion

$$L(\mu, \sigma^2; x) = (2\pi\sigma^2)^{-n/2} \prod_{j=1}^{n} \exp\Big(- \frac{(x_j - \mu)^2}{2\sigma^2} \Big),\ x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Um den Maximum-Likelihood-Schätzer zu berechnen, nutzen wir, dass Extremstellen unter monotonen Transformationen erhalten bleiben. Die Anwendung des Logarithmus auf die Likelihood-Funktion ist eine solche monotone Transformation, die den Vorteil hat, dass aus dem Produkt eine Summe wird (schon Fisher verwendete diesen Trick). Diese sogenannte *Loglikelihood-Funktion* ist hier

$$l(\mu, \sigma^2; x) := \log L(\mu, \sigma^2; x) = -\frac{n}{2}(\log(2\pi) + \log \sigma^2) - \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{2\sigma^2}.$$

An einer Maximalstelle von ℓ verschwinden die ersten Ableitungen und wir erhalten

$$0 = \sigma^{-2} \sum_{j=1}^{n} (x_j - \mu), \quad \frac{n}{2\sigma^2} = \frac{1}{2\sigma^4} \sum_{j=1}^{n} (x_j - \mu)^2.$$

Umstellen der ersten Gleichung nach μ liefert $\widehat{\mu} = \overline{X}_n$ und Einsetzen in die zweite Gleichung ergibt $\widehat{\sigma}^2 = n^{-1} \sum_j (X_j - \overline{X}_n)^2$. Es ist leicht nachzuprüfen, dass $\widehat{\mu}$ und $\widehat{\sigma}^2$ tatsächlich das Maximierungsproblem lösen (und messbar sind). In diesem Fall stimmt der Maximum-Likelihood-Schätzer also mit dem Momentenschätzer aus Beispiel 1.17 überein.

Beispiel 1.22 (Maximum-Likelihood-Schätzer – b) Es sei $X_1, \dots, X_n \sim \text{Poiss}(\lambda)$ eine Poisson-verteilte mathematische Stichprobe mit Parameter $\lambda > 0$, das heißt $\mathbb{P}_\lambda(X_1 = k) = \lambda^k e^{-k}/k!$. Dann ist die gemeinsame Verteilung gegeben durch

$$\mathbb{P}_\lambda(X_1 = k_1, \dots, X_n = k_n) = \frac{\lambda^{k_1 + \cdots + k_n} e^{-n\lambda}}{k_1! \cdots k_n!}, \quad k_1, \dots, k_n \in \mathbb{N}_0,$$

und wird vom Zählmaß dominiert. Ableiten der Loglikelihood-Funktion nach λ und Nullsetzen führt auf den Maximum-Likelihood-Schätzer $\widehat{\lambda} = \overline{X}_n$. Der zugehörige Nachweis einer hinreichenden Bedingung sei der Leserin überlassen.

Beispiel 1.23 (Maximum-Likelihood-Schätzer – c) Erinnern wir uns an das Saskia-Beispiel 1.7. Im fortführenden Beispiel 1.10 hatten wir im statistischen Modell $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (\text{Ber}(\vartheta)^{\otimes n})_{\vartheta \in [0,1]})$ den Schätzer $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ gewählt. Zu welchem Schätzer führt der Maximum-Likelihood-Ansatz? Die Likelihood-Funktion ist hier gegeben durch

$$L(\vartheta, x) = \prod_{i=1}^{n} \vartheta^{x_i} (1 - \vartheta)^{1 - x_i}, \quad x = (x_1, \dots, x_n) \in \{0,1\}^n,\ \vartheta \in [0,1],$$

sodass wir für $\vartheta \in (0,1)$ die Loglikelihood-Funktion

$$l(\vartheta, x) = \Big(\sum_{i=1}^{n} x_i \Big) \log(\vartheta) + \Big(n - \sum_{i=1}^{n} x_i \Big) \log(1 - \vartheta)$$

erhalten. Dabei fällt auf, dass die Loglikelihood-Funktion nur von der Statistik $s = \sum_{i=1}^n x_i$ abhängt. Ableiten von $l(\cdot, x)$ ergibt $\frac{\partial}{\partial \vartheta} l(\vartheta, x) = \frac{s}{\vartheta} + \frac{s-n}{1-\vartheta}$ und dessen Nullstelle ist gegeben durch $\widehat{\vartheta} = \frac{s}{n} = \frac{1}{n} \sum_{i=1}^n x_i$. Eine weitere Ableitung von $l(\cdot, x)$ liefert die hinreichenden Bedingung und die Fälle $\vartheta \in \{0,1\}$ diskutiert man leicht separat. Insgesamt ist $\widehat{\vartheta} = \frac{1}{n} \sum_{i=1}^n X_i$ tatsächlich der Maximum-Likelihood-Schätzer.

1.2.2 Minimax- und Bayes-Ansatz

Wir haben nun verschiedene Schätzmethoden, wie den Maximum-Likelihood-Schätzer oder die Momentenmethode kennengelernt. Natürlich gibt es weitere Konstruktionen und zahlreiche Variationen. Welche Konstruktionsmethode sollte anhand des gegebenen Schätzproblems ausgewählt werden?

Betrachten wir ein statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit abgeleitetem Parameter $\rho\colon \Theta \to \mathbb{R}^p$. Weiter sei eine Verlustfunktion ℓ gegeben. Als mögliches Vergleichskriterium hatten wir bereits die Risikofunktion $R(\vartheta, \widehat{\rho}) = \mathbb{E}_\vartheta[\ell(\vartheta, \widehat{\rho})]$ eines Schätzers $\widehat{\rho}$ eingeführt. Man beachte jedoch folgendes Beispiel:

Beispiel 1.24 (Risiken von Schätzern – a) Es sei $X \sim \mathrm{N}(\mu, 1)$, $\mu \in \mathbb{R}$, und $\ell(\mu, \widehat{\mu}) := (\widehat{\mu} - \mu)^2$. Wir betrachten die beiden Schätzer $\widehat{\mu}_1 := X$ und $\widehat{\mu}_2 := 5$. Die Risiken sind dann gegeben durch

$$R(\mu, \widehat{\mu}_1) = \mathbb{E}_\vartheta[(X-\mu)^2] = 1 \quad \text{und} \quad R(\mu, \widehat{\mu}_2) = (5-\mu)^2.$$

Damit hat $\widehat{\mu}_1$ genau dann ein kleineres Risiko als $\widehat{\mu}_2$, wenn $\mu \notin [4, 6]$. Welchen Schätzer sollen wir nun wählen in Anbetracht dessen, dass μ unbekannt ist? Ein möglicher Ansatz ist, die maximalen Risiken der Schätzer über alle $\mu \in \Theta$ zu bestimmen und dann den Schätzer zu wählen, der das kleinste maximale Risiko besitzt. Da $R(\mu, \widehat{\mu}_1) = 1$ konstant in $\mu \in \mathbb{R}$ ist, während $\mu \mapsto R(\mu, \widehat{\mu}_2)$ eine nach oben geöffnete Parabel beschreibt und damit beliebig groß werden kann, würden wir den Schätzer $\widehat{\mu}_1$ verwenden.

Definition 1.25 Im statistischen Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit abgeleitetem Parameter $\rho\colon \Theta \to \mathbb{R}^p$ und Verlustfunktion ℓ, heißt ein Schätzer $\widehat{\rho}$ **minimax**, falls für das zugehörige Risiko R

$$\sup_{\vartheta\in\Theta} R(\vartheta, \widehat{\rho}) = \inf_{\widetilde{\rho}} \sup_{\vartheta\in\Theta} R(\vartheta, \widetilde{\rho})$$

gilt, wobei sich das Infimum über alle Schätzer (das heißt messbaren Funktionen) $\widetilde{\rho}\colon \mathcal{X} \to \mathbb{R}^p$ erstreckt.

Anstatt den maximal zu erwartenden Verlust zu minimieren, könnten wir die Risiken abhängig von ϑ gewichten und diese vergleichen. Der Schätzer mit dem geringsten gemittelten Risiko wird dann gewählt. Die nötige Struktur dafür gibt uns folgende Definition.

Definition 1.26 Der Parameterraum Θ trage eine σ-Algebra $\mathcal{F}_\Theta$, $\vartheta \mapsto \mathbb{P}_\vartheta(B)$ sei messbar für alle $B \in \mathcal{F}$, und die Verlustfunktion ℓ sei produktmessbar, das heißt

$$\ell\colon (\Theta \times \mathbb{R}^p, \mathcal{F}_\Theta \otimes \mathcal{B}(\mathbb{R}^p)) \to (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+)) \text{ messbar.}$$

Als **a-priori-Verteilung** π des Parameters ϑ bezeichnen wir ein Wahrscheinlichkeitsmaß auf $(\Theta, \mathcal{F}_\Theta)$. Das zu π assoziierte **Bayes-Risiko** eines Schätzers $\widehat{\rho}$ ist

$$R_\pi(\widehat{\rho}) := \mathbb{E}_\pi[R(\vartheta, \widehat{\rho})] = \int_\Theta \int_\mathcal{X} \ell(\vartheta, \widehat{\rho}(x)) \mathbb{P}_\vartheta(\mathrm{d}x) \pi(\mathrm{d}\vartheta).$$

Der Schätzer $\widehat{\rho}$ heißt **Bayes-Schätzer** oder **Bayes-optimal** (bezüglich π), falls

$$R_\pi(\widehat{\rho}) = \inf_{\widetilde{\rho}} R_\pi(\widetilde{\rho}),$$

wobei sich das Infimum über alle Schätzer (das heißt messbaren Funktionen) $\widetilde{\rho}\colon \mathcal{X} \to \mathbb{R}^p$ erstreckt.

A-priori-Verteilung oder englisch *prior distribution* ist insofern ein passender Name, als dass diese Verteilung gewählt wird, bevor die Daten beobachtet werden.

Bemerkung 1.27 (Bayes-Risiko) Das Bayes-Risiko kann auch als insgesamt zu erwartender Verlust in folgendem Sinne verstanden werden: Definiere $\Omega := \mathcal{X} \times \Theta$ und die gemeinsame Verteilung von Beobachtung und Parameter $\widetilde{\mathbb{P}}$ auf $(\mathcal{X} \times \Theta, \mathcal{F} \otimes \mathcal{F}_\Theta)$ gemäß $\widetilde{\mathbb{P}}(\mathrm{d}x, \mathrm{d}\vartheta) = \mathbb{P}_\vartheta(\mathrm{d}x)\pi(\mathrm{d}\vartheta)$. Bezeichnen X und T die Koordinatenprojektionen von Ω auf $\mathcal{X}$ bzw. Θ, dann gilt $R_\pi(\widehat{\rho}) = \mathbb{E}_{\widetilde{\mathbb{P}}}[\ell(T, \widehat{\rho}(X))]$.

Beispiel 1.28 (Risiken von Schätzern – b) Wir beobachten $X \sim \mathrm{Bin}(n, \vartheta)$ mit $\vartheta \in [0, 1]$ unbekannt und betrachten die Schätzer $\widehat{\vartheta}_1 = X/n$, $\widehat{\vartheta}_2 = 0{,}5$ (unabhängig von den Daten!). Dann gilt für den mittleren quadratischen Fehler $R(\vartheta, \widehat{\vartheta}_1) = \vartheta(1-\vartheta)/n$ und $R(\vartheta, \widehat{\vartheta}_2) = (\vartheta - 0{,}5)^2$. Das maximale Risiko beträgt $1/(4n)$ für $\widehat{\vartheta}_1$ und $1/4$ für $\widehat{\vartheta}_2$, sodass wir unter diesem Kriterium $\widehat{\vartheta}_1$ für $n \geqslant 2$ bevorzugen. Unter der a-priori-Verteilung $\pi = \mathrm{U}([0, 1])$ ist ϑ gleichmäßig auf $[0, 1]$ verteilt, und Integration ergibt die Bayesrisiken $R_\pi(\widehat{\vartheta}_1) = 1/(6n)$ und $R_\pi(\widehat{\vartheta}_2) = 1/12$. Unter diesem Bayesrisiko ziehen wir also nur für $n \geqslant 3$ den Schätzer $\widehat{\vartheta}_1$ vor.

Die a-priori-Verteilung π wird manchmal auch als subjektive Einschätzung der Verteilung des zugrunde liegenden Parameters interpretiert. Nachdem eine Beobachtung gemacht wurde, möchte man die Parameterverteilung mithilfe der zusätzlichen Information aktualisieren („Bayesian updating“). Um diesen Vorgang formal zu beschreiben, erinnern wir uns an die bedingten Dichten aus Definition A.34 und die Bayes-Formel (Satz A.35).

Definition 1.29 Sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein von μ dominiertes statistisches Modell mit Dichten $f^{X|T=\vartheta} := \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}$. Sei π eine a-priori-Verteilung auf $(\Theta, \mathcal{F}_\Theta)$ mit Dichte f^T bezüglich eines Maßes ν. Ist $f^{X|T=\cdot}\colon \mathcal{X} \times \Theta \to \mathbb{R}_+$ $(\mathcal{F} \otimes \mathcal{F}_\Theta)$-messbar, dann ist die **a-posteriori-Verteilung** des Parameters, gegeben die Beobachtung $X = x$, definiert durch die ν-Dichte

$$f^{T|X=x}(\vartheta) = \frac{f^{X|T=\vartheta}(x) f^T(\vartheta)}{\int_\Theta f^{X|T=t}(x) f^T(t)\nu(\mathrm{d}t)}, \quad \vartheta \in \Theta, \tag{1.4}$$

für $\widetilde{\mathbb{P}}^X$-f.a. $x \in \mathcal{X}$ mit dem Wahrscheinlichkeitsmaß $\widetilde{\mathbb{P}}$ aus Bemerkung 1.27. Das **a-posteriori-Risiko** eines Schätzers $\widehat{\rho}$, gegeben $X = x$, ist definiert durch

$$R_\pi(\widehat{\rho}|x) = \int_\Theta \ell(\vartheta, \widehat{\rho}(x)) f^{T|X=x}(\vartheta)\nu(\mathrm{d}\vartheta). \tag{1.5}$$

Der Name der a-posteriori-Verteilung oder englisch *posterior distribution* suggeriert bereits, dass diese Verteilung nach einer gemachten Beobachtung $x \in \mathcal{X}$ berechnet wird.

Bemerkung 1.30 (a-posteriori-Verteilung und -Risiko)

- Beachte, dass im Nenner von (1.4) die Randdichte

$$f^X(x) = \int_\Theta f^{X|T=t}(x) f^T(t) \nu(\mathrm{d}t) \tag{1.6}$$

  von X bezüglich μ in $(\mathcal{X} \times \Theta, \mathcal{F} \otimes \mathcal{F}_\Theta, \widetilde{\mathbb{P}})$ steht, sodass (1.4) für $\widetilde{\mathbb{P}}^X$-f.a. $x \in \mathcal{X}$ wohldefiniert ist.
- Im diskreten Fall wird das Integral in (1.5) und im Nenner von (1.4) zu einer Summe, sodass wir genau die klassische Bayes-Formel erhalten.

Beispiel 1.31 (a-posteriori-Verteilung und -Risiko) Sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein von μ dominiertes statistisches Modell. Setze $\Theta = \{0, 1\}, \mathcal{F}_\Theta := \mathcal{P}(\Theta), \ell(\vartheta, r) = |\vartheta - r|$ (0-1-Verlust) und betrachte eine a-priori-Verteilung π mit $\pi(\{0\}) =: \pi_0$ und $\pi(\{1\}) =: \pi_1 = 1 - \pi_0$. Die Wahrscheinlichkeitsmaße $\mathbb{P}_0$ und $\mathbb{P}_1$ mögen Dichten p_0 und p_1 bezüglich eines Maßes μ besitzen (zum Beispiel $\mu = \mathbb{P}_0 + \mathbb{P}_1$). Dann ist die a-posteriori-Verteilung auf Θ durch die Zähldichte

$$f^{T|X=x}(i) = \frac{\pi_i p_i(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}, \quad i = 0, 1, \quad (\widetilde{\mathbb{P}}^X\text{-f.ü.})$$

gegeben. Damit ist das a-posteriori-Risiko eines Schätzers $\widehat{\vartheta} \colon \mathcal{X} \to \{0, 1\}$ als Erwartungswert bezüglich der a-posteriori-Dichte gegeben durch

$$R_\pi(\widehat{\vartheta}|x) = \frac{\widehat{\vartheta}(x)\pi_0 p_0(x) + (1 - \widehat{\vartheta}(x))\pi_1 p_1(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}.$$

Das a-posteriori-Risiko ist minimal für $\widehat{\vartheta}_\pi(x) := \mathbb{1}(\pi_1 p_1(x) \geqslant \pi_0 p_0(x))$, was genau dem später zu besprechenden *Bayes-Klassifizierer* entspricht (siehe Kapitel **??**).

Kurzbiografie (Thomas Bayes) Thomas Bayes wurde um 1702 in London geboren. Er schrieb sich 1719 an der University of Edinburgh ein und studierte Logik und Theologie. Von etwa 1734 bis 1752 war er Pfarrer der presbyterianischen Kirche in Tunbridge Wells bei London. Erst in seinen späten Jahren vertiefte er sein Interesse an der Beschreibung von Wahrscheinlichkeiten und schrieb seine Ideen und Ergebnisse in Manuskripten nieder, die nach seinem Tod veröffentlicht wurden. Er war damit einer der ersten Statistiker, der sich mit Wahrscheinlichkeit befasste. Nicht nur der *Satz von Bayes* wurde nach ihm benannt, auch das große Gebiet der *Bayes-Statistik*. 1761 starb Thomas Bayes in Tunbridge Wells.

Satz 1.32 (Bayes-Risiko und Bayes-Schätzer) *Es gelten die Bedingungen der vorangegangenen Definition. Für das Bayes-Risiko eines Schätzers $\widehat{\rho}$ mit $R_\pi(\widehat{\rho}) < \infty$ gilt dann mit der Randdichte aus* (1.6)

$$R_\pi(\widehat{\rho}) = \int_{\mathcal{X}} R_\pi(\widehat{\rho}|x) f^X(x) \mu(\mathrm{d}x).$$

Minimiert $\widehat{\rho}(x)$ für $\widetilde{\mathbb{P}}^X$-fast alle x das a-posteriori-Risiko im Sinne von

$$R_\pi(\widehat{\rho}|x) = \min_{r \in \mathbb{R}^p} \int_\Theta \ell(\vartheta, r) f^{T|X=x}(\vartheta) \nu(\mathrm{d}\vartheta),$$

dann ist $\widehat{\rho}$ ein Bayes-Schätzer.

Beweis Umstellen von (1.4) ergibt $f^{T|X=x}(\vartheta) f^X(x) = f^{X|T=\vartheta}(x) f^T(\vartheta)$. Setzen wir dies für die Dichte von $\mathbb{P}_\vartheta(\mathrm{d}x)\pi(\mathrm{d}\vartheta)$ ein, ergibt sich

$$\begin{aligned} R_\pi(\widehat{\rho}) &= \int_\Theta \int_{\mathcal{X}} \ell(\vartheta, \widehat{\rho}(x)) \mathbb{P}_\vartheta(\mathrm{d}x)\pi(\mathrm{d}\vartheta) \\ &= \int_\Theta \int_{\mathcal{X}} \ell(\vartheta, \widehat{\rho}(x)) f^{T|X=x}(\vartheta) f^X(x) \mu(\mathrm{d}x)\nu(\mathrm{d}\vartheta) \\ &= \int_{\mathcal{X}} R_\pi(\widehat{\rho}|x) f^X(x) \mu(\mathrm{d}x), \end{aligned}$$

wobei wir im letzten Schritt den Satz von Fubini anwenden können, da die Integranden nichtnegativ sind. Minimiert $\widehat{\rho}(x)$ punktweise den Integranden, so ist auch das Integral minimal. □

Dieser Satz liefert uns eine neue Methode zur Konstruktion von Schätzern.

Korollar 1.33 *Unter quadratischem Verlust ist der Bayes-Schätzer gegeben durch den a-posteriori-Mittelwert*

$$\widehat{\rho}(x) = \int_\Theta \rho(\vartheta) f^{T|X=x}(\vartheta) \nu(\mathrm{d}\vartheta) = \mathbb{E}[\rho(T)|X = x].$$

Wir verwenden hier den über bedingte Dichten definierten bedingten Erwartungswert $\mathbb{E}[\rho(T)|X = x]$ für $\widetilde{\mathbb{P}}^X$-f.a. $x \in \mathcal{X}$, der sich natürlich in das abstraktere Konzept bedingter Erwartungen einbettet. Den Beweis dieses Korollars überlassen wir der Leserin als Übung. Für den wichtigen Fall, dass $\rho(\vartheta) = \vartheta \in \mathbb{R}^p$ die Identität ist, ergibt sich der *a-posteriori-Mittelwert*

$$\widehat{\vartheta}(x) = \int_\Theta \vartheta f^{T|X=x}(\vartheta) \nu(\mathrm{d}\vartheta) = \mathbb{E}[T|X = x]$$

als Bayes-Schätzer zum quadratischen Verlust. Oft ist ν das Zählmaß oder das Lebesgue-Maß, sodass sich das vorangegangene Integral als Summe bzw. als Riemann-Integral schreiben lässt.

Analog ist der Bayes-Schätzer bezüglich absolutem Verlust für $\rho(\vartheta) = \vartheta \in \mathbb{R}$ gegeben durch den *Median* der a-posteriori-Verteilung:

$$\widetilde{\vartheta}(x) = \inf \left\{ q \in \mathbb{R} : \int_{-\infty}^q f^{T|X=x}(\vartheta) \nu(\mathrm{d}\vartheta) \geqslant \frac{1}{2} \right\}$$

Falls ν ein Zählmaß ist und damit die a-posteriori-Verteilung diskret, ist auch der 0-1-Verlust von Interesse. Als Bayes-Schätzer ergibt sich dann die Maximalstelle $\arg\max_{\vartheta\in\Theta} f^{T|X=x}(\vartheta)$ der a-posteriori-Verteilung, die man auch MAP-Schätzer (Maximum-a-posteriori-Schätzer) nennt. Motiviert durch die Maximum-Likelihood-Methode wird der Maximum-a-posteriori-Schätzer auch über den diskreten Fall hinaus eingesetzt. Es handelt sich dann allerdings nicht mehr um einen Bayes-Schätzer im Sinne von Satz 1.32.

Methode 1.34 (Bayes-Schätzer) Durch die Wahl einer a-priori-Verteiltung ergeben sich durch den a-posteriori-Mittelwert, den a-posteriori-Median und die a-posteriori-Maximalstelle Schätzer für den unbekannten Parameter.

Beispiel 1.35 (Bayes-Schätzer – a) Wir betrachten das statistisches Modell $(\{0,...,n\}, \mathcal{P}(\{0,...,n\}), (\mathrm{Bin}(n,\vartheta))_{\vartheta\in[0,1]})$ und wählen wie in Beispiel 1.28 als a-priori-Verteilung $\pi = \mathrm{U}([0,1])$ mit Dichte $f^T = \mathbb{1}_{[0,1]}$. Um die a-posteriori-Verteilung zu berechnen, genügt es, deren Dichte bis auf die Normierungskonstante zu bestimmen, da wir bereits wissen, dass sich wieder eine Wahrscheinlichkeitsdichte in ϑ ergibt. Wir verwenden das Symbol $\propto$ für Gleichheit bis auf eine multiplikative Konstante. Es gilt für $x \in \{0,\dots,n\}$ und $\vartheta \in [0,1]$

$$\begin{aligned} f^{T|X=x}(\vartheta) = \frac{f^{X|T=\vartheta}(x) f^T(\vartheta)}{f^X(x)} &\propto f^{X|T=theta}(x) f^T(\vartheta) \\ &= \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x} \\ &\propto \vartheta^x (1-\vartheta)^{n-x}. \end{aligned}$$

Wir erkennen die Struktur einer Beta-Verteilung $\mathrm{Beta}(\alpha,\beta)$ mit Parametern $\alpha = x+1$ und $\beta = n - x + 1$. Als Bayes-Schätzer unter quadratischem Verlust ergibt sich der Erwartungswert der a-posteriori-Verteilung

$$\widehat{\vartheta} = \int_0^1 \vartheta f^{T|X=x}(\vartheta)\mathrm{d}\vartheta = \frac{\alpha}{\alpha+\beta} = \frac{x+1}{n+2}.$$

Wir erhalten eine Alternative zum Maximum-Likelihood-Schätzer, die uns bereits in Beispiel 1.14 begegnet ist.

Beispiel 1.36 (Bayes-Schätzer – b) Sei $X_1,\dots,X_n \sim \mathrm{N}(\mu,\sigma^2)$ eine mathematische Stichprobe mit bekanntem $\sigma^2 > 0$ und a-priori-Verteilung $\pi = \mathrm{N}(a,b^2)$. Mithilfe der Bayes-Formel kann die a-posteriori-Verteilung für eine Realisierung $x = (x_1,\dots,x_n)$ berechnet werden. Da $f^{T|X=x}$ wieder eine Dichte, also normiert, sein muss, brauchen wir die Konstanten nicht mitzuberechnen und verwenden das Symbol $\propto$ für Gleichheit bis auf eine multiplikative Konstante:

$$\begin{aligned} f^{T|X=x}(\mu) &\propto f^{X|T=\mu}(x) f^T(\mu) \\ &\propto \exp\Big(-\sum_{i=1}^n \tfrac{(x_i-\mu)^2}{2\sigma^2}\Big) \exp\big(-\tfrac{(\mu-a)^2}{2b^2}\big) \end{aligned}$$

$$\begin{aligned}
&\propto \exp\Big(-\tfrac{\mu^2-2\mu\overline{x}_n}{2\sigma^2/n} - \tfrac{\mu^2-2a\mu}{2b^2}\Big)\\
&= \exp\Big(-\tfrac{(b^2+\sigma^2/n)\mu^2-2\mu(b^2\overline{x}_n+a\sigma^2/n)}{2b^2\sigma^2/n}\Big)\\
&\propto \exp\Big(-\tfrac{1}{2}\big(\tfrac{n}{\sigma^2}+\tfrac{1}{b^2}\big)\big(\mu - \tfrac{b^2\overline{x}_n}{b^2+\sigma^2/n} - \tfrac{a\sigma^2/n}{b^2+\sigma^2/n}\big)^2\Big)
\end{aligned}$$

Gegeben die Beobachtung X, ist μ also gemäß

$$\mathrm{N}\Big(\tfrac{b^2}{b^2+\sigma^2/n}\overline{X}_n + \tfrac{\sigma^2/n}{b^2+\sigma^2/n}a, \big(n\sigma^{-2}+b^{-2}\big)^{-1}\Big)$$

a-posteriori-verteilt. Da hier der a-posteriori-Mittelwert mit dem a-posteriori-Median und der Maximalstelle der a-posteriori-Dichte übereinstimmt, ist der Bayes-Schätzer bezüglich quadratischem Verlust und absolutem Verlust sowie der MAP-Schätzer gegeben durch

$$\widehat{\mu}_n = \frac{b^2}{b^2+\sigma^2/n}\overline{X}_n + \frac{\sigma^2/n}{b^2+\sigma^2/n}a.$$

Wir gewichten also den empirischen Mittelwert $\overline{X}_n$ und den a-priori-Mittelwert a entsprechend dem Verhältnis der Varianzen $\frac{\sigma^2}{n}$ und b^2.

In beiden Beispielen konnten wir die a-posteori-Verteilung explizit bestimmen. Hierbei handelt es sich allerdings um Spezialfälle in denen die Berechnung von Bayes-Schätzern besonders einfach ist. Für komplexere Modelle führt die Berechnung der a-posteriori-Dichte mitunter auf hochdimensionale Integrale. Dazu werden häufig numerische Methoden wie MCMC-Methoden (Markov Chain Monte Carlo) verwendet.

Bemerkung 1.37 Eine Familie von Verteilungen $\mathcal{D}$ auf $(\Theta, \mathcal{F}_\Theta)$ heißt durch $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ *konjugiert* (lateinisch für „verbunden"), falls für jede a-priori-Verteilung $\pi \in \mathcal{D}$ und jede Beobachtung $X = x$ die a-posteriori-Verteilung ebenfalls zu $\mathcal{D}$ gehört. Beispiel 1.36 zeigt, dass die Verteilungsklasse der Normalverteilungen durch Normalverteilungen mit unbekanntem Erwartungswert und bekannter Varianz konjugiert werden. Beispiel 1.35 lässt sich auf Beta-verteilte a-priori-Verteilungen verallgemeinern, sodass auch hier eine konjugierte Verteilungsklasse vorliegt (Aufgabe 1.6).

Zum Schluss dieses Kapitels wollen wir noch einen Zusammenhang zwischen Minimax- und Bayes-Schätzer deutlich machen.

Lemma 1.38 *In einem dominierten statistischen Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit messbarem Parameterraum $(\Theta, \mathcal{F}_\Theta)$ gilt für jeden Schätzer $\widehat{\rho}$*

$$\sup_{\vartheta\in\Theta} R(\vartheta, \widehat{\rho}) = \sup_{\pi} R_\pi(\widehat{\rho}),$$

wobei sich das zweite Supremum über alle a-priori-Verteilungen π auf $(\Theta, \mathcal{F}_\vartheta)$ mit $(\mathcal{F}\otimes\mathcal{F}_\Theta)$-messbarer gemeinsamer Dichte $f^{X|T=\cdot}$ erstreckt. Insbesondere ist das Risiko eines Bayes-Schätzer stets kleiner oder gleich dem Minimax-Risiko.

Beweis Natürlich gilt $R_\pi(\widehat{\rho}) = \int_\Theta R(\vartheta, \widehat{\rho})\pi(\mathrm{d}\vartheta) \leqslant \sup_{\vartheta\in\Theta} R(\vartheta, \widehat{\rho})$ für alle π. Andererseits folgt durch Betrachtung der a-priori-Punktverteilungen δ_ϑ, $\vartheta \in \Theta$, dass $\sup_\pi R_\pi(\widehat{\rho}) \geqslant \sup_{\vartheta\in\Theta} R(\vartheta, \widehat{\rho})$ gilt. Dies zeigt die Behauptung. □

Durch dieses Lemma erhalten wir untere Schranken für das Minimax-Risiko über das Risiko von Bayes-Schätzern. Mögliche Anwendungen illustriert der folgende Satz.

Satz 1.39 *Sei $X_1, \ldots, X_n$ eine $\mathrm{N}(\mu, \sigma^2)$-verteilte mathematische Stichprobe mit unbekanntem Erwartungswert $\mu \in \mathbb{R}$ und bekanntem $\sigma^2 > 0$. Bezüglich quadratischem Risiko ist das arithmetische Mittel $\overline{X}_n$ ein Minimax-Schätzer von μ.*

Beweis Wir betrachten a-priori-Verteilungen $\pi = \mathrm{N}(0, b^2)$ für μ. Nach Beispiel 1.36 ist die a-posteriori-Verteilung gegeben durch

$$\mathrm{N}\Big(\tfrac{b^2\overline{X}_n}{b^2+\sigma^2/n}, \big(n\sigma^{-2} + b^{-2}\big)^{-1}\Big).$$

Der Bayes-Schätzer bezüglich quadratischem Risiko ist gegeben durch den a-posteriori-Erwartungswert $\widehat{\mu}_n = b^2\overline{X}_n/(b^2 + \sigma^2 n^{-1})$ und sein a-posteriori-Risiko ist gegeben durch die Varianz der a-posteriori-Verteilung. Ist f^X die Randdichte von X bezüglich $\widetilde{\mathbb{P}}$, folgt aus Satz 1.32:

$$\begin{aligned} R_\pi(\widehat{\mu}_n) &= \int_{\mathbb{R}^n} \mathrm{Var}_{T|X=x}(\mu) f^X(x)\mathrm{d}x \\ &= \int_{\mathbb{R}^n} \big(n\sigma^{-2} + b^{-2}\big)^{-1} f^X(x)\mathrm{d}x = \big(n\sigma^{-2} + b^{-2}\big)^{-1} \end{aligned}$$

Somit können wir das Minimax-Risiko nach unten abschätzen:

$$\begin{aligned} \inf_{\widetilde{\mu}} \sup_{\mu\in\mathbb{R}} R(\mu, \widetilde{\mu}) = \inf_{\widetilde{\mu}} \sup_{\pi} R_\pi(\widetilde{\mu}) &\geqslant \inf_{\widetilde{\mu}} \sup_{b>0} R_{\mathrm{N}(0,b^2)}(\widetilde{\mu}) \\ &\geqslant \sup_{b>0} \inf_{\widetilde{\mu}} R_{\mathrm{N}(0,b^2)}(\widetilde{\mu}) \\ &= \sup_{b>0} \big(n\sigma^{-2} + b^{-2}\big)^{-1} = \frac{\sigma^2}{n}, \end{aligned}$$

wie behauptet, da $R(\mu, \overline{X}_n) = \sigma^2/n$. □

Damit ist $\widehat{\mu}_1$ aus Beispiel 1.24 minimax.

Bemerkung 1.40 Auch in höheren Dimensionen ist das koeffizientenweise Stichprobenmittel $\overline{X}_n$ ein Minimax-Schätzer für den Mittelwertvektor der Normalverteilung, das heißt das maximale Risiko ist kleinstmöglich. Ab Dimension 3 gibt es jedoch Schätzer, inbesondere den sogenannten *James-Stein-Schätzer*, deren Risiko für jedes $\mu \in \mathbb{R}$ stets kleiner ist als das Risiko von $\overline{X}_n$. Dieser Effekt ist als *Stein-Phänomen* bekannt, siehe Lehmann and Casella (1998). In der Sprache der Entscheidungstheorie wird $\overline{X}_n$ als *nicht zulässig* bezeichnet. Hierbei gibt es keinen Widerspruch zur Mimimax-Eigenschaft von $\overline{X}_n$, denn die Verbesserung des James-Stein-Schätzers wird beliebig klein für $|\mu| \to \infty$.

1.3 Hypothesentests

Wir wenden uns nun der zweiten grundlegenden Fragestellung zu. Häufig ist weniger die gesamte zugrunde liegende Verteilung von Interesse als die Frage, ob eine bestimmte Eigenschaft erfüllt ist. Hierfür formuliert man die gefragte Eigenschaft als Hypothese und entscheidet dann, ob die gemachten Beobachtungen für oder gegen diese Hypothese sprechen. Die Entscheidung dafür oder dagegen beruht auf einem statistischen (Hypothesen-)Test.

1.3.1 Statistische Tests und ihre Fehler

Als einführendes Beispiel kann man sich ein Gerichtsverfahren vorstellen, bei dem eine Person beschuldigt wird, eine Straftat begangen zu haben. Das Gericht geht prinzipiell von der Annahme aus, dass der Angeklagte unschuldig ist. Diese Annahme nennen die Juristen Unschuldsvermutung, wir Statistiker nennen sie abstrakter *Nullhypothese* H_0. Die *Alternativhypothese* H_1 lautet, dass der Angeklagte schuldig ist. Es wird nun versucht, die Schuld des Angeklagten zu erörtern.

Das Gerichtsverfahren soll verhindern, den Angeklagten zu Unrecht zu verurteilen. Die Wahrscheinlichkeit für diesen sogenannten *Fehler 1. Art* soll anhand von Indizien und Argumentationen möglichst gering sein. Dies beeinflusst aber auch den sogenannten *Fehler 2. Art*: Ein Schuldiger sollte nicht freigesprochen werden. Beide Fehlerarten werden nicht gleichzeitig minimiert werden können. Die Unschuldsvermutung impliziert eine asymmetrische Gewichtung der Fehler erster und zweiter Art. Ganz ähnlich werden wir bei der Konstruktion statistischer Tests vorgehen.

Definition 1.41 Wir betrachten ein statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$. Die Parametermenge sei in zwei disjunkte Teilmengen Θ_0 und Θ_1 zerlegt, das heißt $\Theta = \Theta_0 \cup \Theta_1$ und $\Theta_0 \cap \Theta_1 = \emptyset$. Das **Testproblem** liest sich dann als

$$H_0 \colon \vartheta \in \Theta_0 \quad \text{gegen} \quad H_1 \colon \vartheta \in \Theta_1.$$

Dabei werden H_0, H_1 als **Hypothesen** bezeichnet, genauer heißt H_0 **Nullhypothese** und H_1 **Alternativhypothese** oder **Alternative**.

Beispiel 1.42 (Hypothesentest) Denken wir wieder an unser Saskia-Beispiel 1.7. Die Studierenden in der Umfrage hatten nur zwei Auswahlmöglichkeiten: Sie stimmten für 1, wenn sie mit ihrem Studium zufrieden waren, und für 0, wenn Sie mit ihrem Studium unzufrieden waren. Die Antworten $X_1, \ldots, X_n$ hatten wir als Bernoulliverteilt modelliert mit Parameter $\vartheta = \mathbb{P}(X_i = 1) \in (0, 1)$.

1. Laut einer Umfrage unter 6000 Studierenden[1] sind 74,9% der Studierendenschaft eher zufrieden bis ganz und gar zufrieden mit ihrem Studium. Dieses Ergebnis kann Saskia anhand ihrer eigenen Erhebung überprüfen. Die entsprechende Nullhypothese lautet dann formal $H_0 : \vartheta \in [\frac{3}{4}, 1) =: \Theta_0$. Da wir die Parametermenge $\Theta = (0, 1)$ nach obiger Definition in disjunkte Teilmengen aufteilen müssen, bleibt für die Alternativhypothese H_1 der Parameterbereich $\Theta_1 = (0, \frac{3}{4})$ übrig. Das Testproblem lautet ausformuliert

$$H_0 : \vartheta \in [\tfrac{3}{4}, 1) \quad \text{gegen} \quad H_1 : \vartheta \in (0, \tfrac{3}{4}).$$

2. Saskia könnte auch die Hypothese aufstellen, dass das Verhältnis zwischen Studierenden, die mit ihrem Studium zufrieden bzw. unzufrieden sind, ausgewogen ist. Dies würde zur Nullhypothese $H_0 : \vartheta = 1/2$ und Alternative $H_1 : \vartheta \neq 1/2$ führen.

Ein statistischer Test entscheidet nun zwischen Nullhypothese und Alternative aufgrund einer Beobachtung $x \in \mathcal{X}$.

Definition 1.43 Ein **nichtrandomisierter statistischer Test** ist eine messbare Abbildung

$$\varphi\colon (\mathcal{X}, \mathcal{F}) \to (\{0, 1\}, \mathcal{P}(\{0, 1\}))$$

mit dem Testentscheid

$$\varphi(x) = \begin{cases} 1: & \text{die Nullhypothese wird verworfen/abgelehnt,} \\ 0: & \text{die Nullhypothese wird nicht verworfen bzw. wird akzeptiert.} \end{cases}$$

Die Menge $\{\varphi = 1\} = \{x \in \mathcal{X} : \varphi(x) = 1\}$ heißt **Ablehnbereich** oder **kritischer Bereich** von φ.

Ein **randomisierter statistischer Test** ist eine messbare Abbildung $\varphi\colon (\mathcal{X}, \mathcal{F}) \to ([0, 1], \mathcal{B}([0, 1]))$. Im Fall $\varphi(x) \in (0, 1)$ entscheidet ein unabhängiges Bernoulli-Zufallsexperiment mit Erfolgswahrscheinlichkeit $p = \varphi(x)$, ob die Nullhypothese verworfen wird.

Testen beinhaltet mögliche Fehlentscheidungen:

(i) **Fehler 1. Art** (auch *α-Fehler*, englisch: *type I error* oder *false positive* genannt): Entscheidung für H_1, obwohl H_0 wahr ist,

(ii) **Fehler 2. Art** (auch *β-Fehler*, englisch: *type II error* oder *false negative* genannt): Entscheidung für H_0, obwohl H_1 wahr ist.

Ein nicht-randomisierter Test lässt sich als randomisierter Test mit Werten nur in $\{0, 1\}$ interpretieren. Wenn wir allgemein von Tests sprechen, meinen wir daher formal randomisierte Tests. In den Beispielen werden wir uns zunächst auf nichtrandomisierte (statistische) Tests konzentrieren. Im folgenden Abschnitt wird sich

[1] HIS (2008). *Studenten, die gegenwärtig mit den folgenden Bereichen ihres Lebens eher zufrieden bis ganz und gar zufrieden sind.* In Statista: `https://de.statista.com/statistik/daten/studie/1440/umfrage/zufriedenheit-von-studenten/` (Zugegriffen am 05. November 2020).

jedoch auch zeigen, dass randomisierte Tests ihre (mathematische) Berechtigung haben.

Im einführenden Gerichtsbeispiel hatten wir die Schwierigkeit angemerkt, eine sinnvolle Schwelle zu finden, ab wann genügend Indizien und Argumente für eine Verurteilung vorliegen. Wir wollen nicht, dass ein Unschuldiger zu Unrecht verurteilt wird (Fehler 1. Art), und wir wollen auch nicht, dass ein Schuldiger frei gesprochen wird (Fehler 2. Art). In dieser Analogie soll also ein statistischer Test entscheiden, ob genügend Indizien und Argumente vorliegen.

Definition 1.44 Sei φ ein Test der Hypothese $H_0\colon \vartheta \in \Theta_0$ gegen die Alternative $H_1\colon \vartheta \in \Theta_1$ im statistischen Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$. Die **Gütefunktion** von φ ist definiert als

$$\beta_\varphi\colon \Theta \to \mathbb{R}_+,\ \vartheta \mapsto \mathbb{E}_\vartheta[\varphi].$$

Ein Test φ erfüllt das **Signifikanzniveau** $\alpha \in [0,1]$ (oder φ ist ein **Test zum Niveau** α), falls $\beta_\varphi(\vartheta) \leqslant \alpha$ für alle $\vartheta \in \Theta_0$ gilt.

Erfüllt ein Test ein *vorher* festgelegtes Niveau α, so nennen wir das Ergebnis des Tests statistisch *signifikant* (zum Niveau α). In Definition 1.44 legen wir also $\alpha \in [0,1]$ vorher fest und konstruieren den Test φ anschließend so, dass für alle Parameter $\vartheta \in \Theta_0$ der Nullhypothese der Erwartungswert $\mathbb{E}_\vartheta[\varphi]$ kleiner gleich α ist.

Ist $\varphi\colon (\mathcal{X}, \mathcal{F}) \to (\{0,1\}, \mathcal{P}(\{0,1\}))$ ein nicht-randomisierter Test, so vereinfacht sich diese Forderung: Da

$$\mathbb{E}_\vartheta[\varphi] = 0 \cdot \mathbb{P}_\vartheta(\varphi = 0) + 1 \cdot \mathbb{P}_\vartheta(\varphi = 1) = \mathbb{P}_\vartheta(\varphi = 1)$$

gilt, erfüllt φ das Signifikanzniveau α, falls

$$\beta_\varphi(\vartheta) = \mathbb{P}_\vartheta(\varphi = 1) \leqslant \alpha \quad \text{für alle } \vartheta \in \Theta_0$$

gilt. Die Wahrscheinlichkeit, dass unser Test die Nullhypothese ablehnt, obwohl H_0 stimmt, soll also höchstens α sein. Das Signifikanzniveau α beschränkt damit die Wahrscheinlichkeit für Fehler 1. Art.

Greifen wir noch einmal die Analogie zum Gerichtswesen auf. Für Juristen ist es schlimmer, Unschuldige zu Unrecht zu verurteilen als Schuldige fälschlicherweise freizusprechen (Unschuldsvermutung). Für sie sind also Fehler 1. Art schwerwiegender als Fehler 2. Art. Deshalb wird versucht, die Fehler 1. Art durch ein geringes Signifikanzniveau α klein zu halten. Typische Werte sind $\alpha = 0{,}05$ und $\alpha = 0{,}01$.

Auf der Alternativmenge Θ_1 sollte die Gütefunktion möglichst groß sein, damit der Fehler 2. Art klein ist. Für einen nichtrandomisierten Test φ ergibt sich direkt der Zusammenhang

$$\mathbb{P}_\vartheta(\varphi = 0) = 1 - \mathbb{P}_\vartheta(\varphi = 1) = 1 - \beta_\varphi(\varphi) \qquad \text{für } \vartheta \in \Theta_1.$$

Entsprechend gibt $1 - \beta_\varphi(\vartheta)$ auch für randomisierte Tests die Wahrscheinlichkeit eines Fehler 2. Art an.

In der Regel lassen sich die Wahrscheinlichkeiten für Fehler 1. und 2. Art nicht gleichzeitig minimieren. Deshalb verfährt man im Allgemeinen so, dass

(i) zuerst die Wahrscheinlichkeit für Fehler 1. Art durch ein vorgegebenes Signifikanzniveau begrenzt wird und dann
(ii) unter der Maßgabe von (i) ein Test φ gesucht wird, der die Wahrscheinlichkeit für Fehler 2. Art minimiert.

Als Nullhypothese H_0 sollte im Allgemeinen der bisherige Standard oder die aufgrund theoretischer Überlegungen erzielte Modellierung gewählt werden. Wird H_0 nicht zugunsten von H_1 abgelehnt, heißt das noch lange nicht, dass H_0 „wahr" ist, sondern nur, dass die Daten nicht im Widerspruch zur Nullhypothese stehen und H_1 die Daten nicht besser als H_0 erklärt. Wird hingegen H_0 abgelehnt, besteht Grund zu der Annahme, dass die Modelle unter H_0 die Daten nicht hinreichend gut beschreiben.

Als Nächstes betrachten wir zwei Formen von Testproblemen, die oft in der Praxis auftauchen.

Definition 1.45 Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, wobei $\Theta \subseteq \mathbb{R}$ ein zusammenhängendes Intervall ist.

1. Testprobleme der Form $H_0 \colon \vartheta \leqslant \vartheta_0$ gegen $H_1 \colon \vartheta > \vartheta_0$ oder $H_0 \colon \vartheta \geqslant \vartheta_0$ gegen $H_1 \colon \vartheta < \vartheta_0$ für ein $\vartheta_0 \in \Theta$ heißen **einseitig**.
2. Testprobleme der Form $H_0 \colon \vartheta = \vartheta_0$ gegen $H_1 \colon \vartheta \neq \vartheta_0$ für ein $\vartheta_0 \in \Theta$ heißen **zweiseitig**.

Hypothesentests für einseitige bzw. zweiseitige Testproblems werden selbst **einseitig** bzw. **zweiseitig** genannt.

Beispiel 1.46 (Einseitiger Binomialtest) Von den 13 Todesfällen unter 55- bis 65-jährigen Arbeitern eines Kernkraftwerks im Jahr 1995 waren 5 auf einen Tumor zurückzuführen. Die Todesursachenstatistik 1995 weist aus, dass Tumore bei etwa 1/5 aller Todesfälle die Ursache in der betreffenden Altersklasse (in der Gesamtbevölkerung) sind. Ist die beobachtete Häufung von tumorbedingten Todesfällen signifikant zum Niveau 5%? Das heißt, ist die Wahrscheinlichkeit dafür, dass 5 von 13 Kernkraftwerksarbeitern an Tumoren gestorben sind, in Anbetracht dessen, dass allgemein jeder 5. an Tumoren stirbt, kleiner oder gleich 5%?

Um diese Frage zu beantworten, beschreiben wir die Anzahl der Tumortoten als Zufallsvariable $X \in \{0, 1, \dots, n\}$ mit $n = 13$. Als statistisches Modell, in das X abbildet, wählen wir $\mathcal{X} = \{0, \dots n\}, \mathcal{F} = \mathcal{P}(\mathcal{X})$ und $\mathbb{P}_p = \mathrm{Bin}(13, p)$. Der Parameter $p \in [0, 1]$ ist die Wahrscheinlichkeit, dass eine Person an einem Tumor gestorben ist. Wir wollen wissen, ob 5 von 13 Todesfällen eine signifikante Häufung zum Niveau $\alpha = 5\%$ sind. Diese Fragestellung führt auf das Testproblem

$$H_0 \colon p \leqslant 1/5 \quad \text{gegen} \quad H_1 \colon p > 1/5.$$

Wir müssen nun einen geeigneten (nichtrandomisierten) Test φ zum Niveau $\alpha = 0{,}05$ konstruieren. Naheliegenderweise wählen wir

$$\varphi(x) := \mathbb{1}(x > c). \tag{1.7}$$

Hierbei wollen wir den kritischen Wert $c > 0$ so wählen, dass $\beta_\varphi(p) \leqslant \alpha$ für alle $p \leqslant 1/5$ gilt. Wegen der zugrunde liegenden Binomialverteilung nennen wir diesen Test *Binomialtest*. Für $p \leqslant 1/5$ gilt

$$\beta_\varphi(p) = \mathbb{P}_p(\varphi = 1) = \mathbb{P}_p(X > c) \leqslant \sup_{p \leqslant 1/5} \mathbb{P}_p(X > c) \overset{!}{\leqslant} \alpha. \tag{1.8}$$

Um eine möglichst große Güte des Tests zu erreichen, sollte c unter dieser Nebenbedingung möglichst klein gewählt werden. Für die Verteilungsfunktion $\mathbb{P}_p(X \leqslant k)$ unserer Binomialverteilung gilt für $k \in \mathcal{X}$

$$\mathbb{P}_p(X \leqslant k) = \sum_{l=0}^{k} \binom{13}{l} p^l (1-p)^{13-l}.$$

Da $p \mapsto \mathbb{P}_p(X \leqslant k)$ für alle $k \in \mathcal{X}$ monoton fallend auf $[0,1]$ ist (dies sieht man durch Ableiten), folgt, dass $p \mapsto \mathbb{P}_p(X > c) = 1 - \mathbb{P}_p(X \leqslant c)$ monoton steigend ist. Folglich gilt mit (1.8), dass

$$\sup_{p \leqslant 1/5} \mathbb{P}_p(X > c) = \mathbb{P}_{1/5}(X > c) \overset{!}{\leqslant} \alpha = 0{,}05 \quad \Leftrightarrow \quad \mathbb{P}_{1/5}(X \leqslant c) \overset{!}{\geqslant} 0{,}95.$$

Da c eine natürliche Zahl sein sollte und wegen

$$\mathbb{P}_{1/5}(X \leqslant 4) \approx 0{,}901 \quad \text{und} \quad \mathbb{P}_{1/5}(X \leqslant 5) \approx 0{,}970,$$

wählen wir $c = 5$, siehe Abbildung 1.4 auf Seite 34. Unser Test lautet also $\varphi(x) = \mathbb{1}_{\{x>5\}}$. Setzen wir nun die Anzahl 5 von Kernkraftwerksarbeitern, die an einem Tumor gestorben sind, in φ ein, so erhalten wir 0. Das heißt, unser Test φ akzeptiert zum Niveau $\alpha = 0{,}05$ die Nullhypothese, dass die beobachteten Todesfälle nicht signifikant sind. Wir erhalten also

$$\sup_{p \leqslant 1/5} \beta_\varphi(p) = \beta_\varphi(1/5) \approx 0{,}03 < 0{,}05 = \alpha,$$

und der Binomialtest kann aufgrund des diskreten Stichprobenraums das Niveau nicht voll ausschöpfen.

Dieses Beispiel führt uns auf ein allgemeines Konstruktionsprinzip von Tests einer Hypothese $H_0 \colon \vartheta \in \Theta_0$ gegen die Alternative $H_1 \colon \vartheta \in \Theta_1$ mit $\Theta_0 \neq \emptyset$ und $\Theta_1 = \Theta \setminus \Theta_0$.

Methode 1.47 (Konstruktion von Tests durch kritische Werte) Für ein gegebenes Testproblem und ein festgelegtes Signifikanzniveau $\alpha \in (0,1)$ wähle eine **Teststatistik** bzw. **Test-** oder **Prüfgröße** $T \colon (\mathcal{X}, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ und konstruiere einen Niveau-α-Test durch

$$\varphi(x) = \mathbb{1}(T(x) > c), \quad x \in \mathcal{X}, \tag{1.9}$$

wobei $c \in \mathbb{R}$ so gewählt wird, dass

$$\sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta(T(X) > c) \leqslant \alpha \tag{1.10}$$

gilt. Das kleinstmögliche $c = c_\alpha$, sodass (1.10) erfüllt ist, heißt **kritischer Wert**. Analog sind auch kritische Bereiche der Formen $\{T(x) \geqslant c\}, \{T(x) < c\}, \{T(x) \leqslant c\}, \{|T(x)| > c\}, \{T(x) < u\} \cup \{T(x) > o\}$ usw. möglich.

Für eine möglichst hohe Güte des Tests möchten wir den Ablehnbereich möglichst groß wählen und minimieren daher c unter der Bedingung (1.10). Um den kritischen Wert zu wählen und somit das Signifikanzniveau des Tests zu gewährleisten, muss die Verteilung der Teststatistik unter der Nullhypothese bekannt sein. Da man häufig Monotonieeigenschaften nutzen kann (wie im Beispiel 1.46), vereinfacht sich die Wahl von c_α im Fall von ein- oder zweiseitigen Testproblemen oft, wenn das Gleichheitszeichen in der Nullhypothese steht (siehe Definition 1.45). Im Spezialfall $\Theta_0 = \{\vartheta_0\}$ ist für einen einseitigen Test der kritische Wert c_α genau das $(1-\alpha)$-*Quantil* der Verteilung von T unter $\mathbb{P}_{\vartheta_0}$.

Natürlich hängt sowohl die Teststatistik als auch der kritische Bereich vom jeweiligen Testproblem ab. Die Herausforderung ist also die Wahl einer Teststatistik, die (gut) geeignet ist, um die jeweilige Hypothese zu überprüfen. Nachdem wir in Beispiel 1.46 bereits einen einseitigen Binomialtest gesehen haben, illustriert das folgende Beispiel den zweiseitigen Fall.

Beispiel 1.48 (Zweiseitiger Binomialtest) Wir wollen die *Hypothese*: „Es werden genauso viele Jungen wie Mädchen geboren." überprüfen.
Sind bei $n \in \mathbb{N}$ Geburten w Mädchen zur Welt gekommen, ist es sinnvoll, als Stichprobenraum $\mathcal{X} = \{0, \dots, n\}$ zu wählen und als statistisches Modell $(\mathcal{X}, \mathcal{P}(\mathcal{X}), (\mathbb{P}_\vartheta)_{\vartheta \in [0,1]})$ mit Binomialverteilungen $\mathbb{P}_\vartheta = \mathrm{Bin}(n, \vartheta)$. Die Hypothese führt auf das zweiseitige Testproblem

$$H_0 : \vartheta = \vartheta_0 \quad \text{gegen} \quad H_1 : \vartheta \neq \vartheta_0$$

für $\vartheta_0 = \frac{1}{2}$, wobei $w \in \mathcal{X}$ beobachtet wird. Wir setzen das Niveau $\alpha = 0{,}05$. Die Teststatistik $T(w) = w$ führt auf den *zweiseitigen Binomialtest*

$$\varphi_\alpha(w) = 1 - \mathbb{1}_{[u_\alpha(\vartheta_0), o_\alpha(\vartheta_0)]}(w)$$

mit kritischen Werten

$$\begin{aligned} u_\alpha(\vartheta_0) &:= \max\{k \in \mathbb{N} : \mathbb{P}_{\vartheta_0}(\{0, \dots, k-1\}) \leqslant \alpha/2\} \quad \text{und} \\ o_\alpha(\vartheta_0) &:= \min\{k \in \mathbb{N} : \mathbb{P}_{\vartheta_0}(\{k+1, \dots, n\}) \leqslant \alpha/2\}. \end{aligned} \tag{1.11}$$

Die symmetrische Verteilung der Fehlerwahrscheinlichkeiten am unteren und am oberen Rand ist dabei eine vernünftige Wahl, wobei auch jede andere Aufteilung des Niveaus α prinzipiell möglich ist.

Beachte, dass dieser zweiseitige Binomialtest nur im Fall $\vartheta_0 = 1/2$ symmetrisch um ϑ_0 ist, denn nur dann ist die Binomialverteilung symmetrisch, siehe

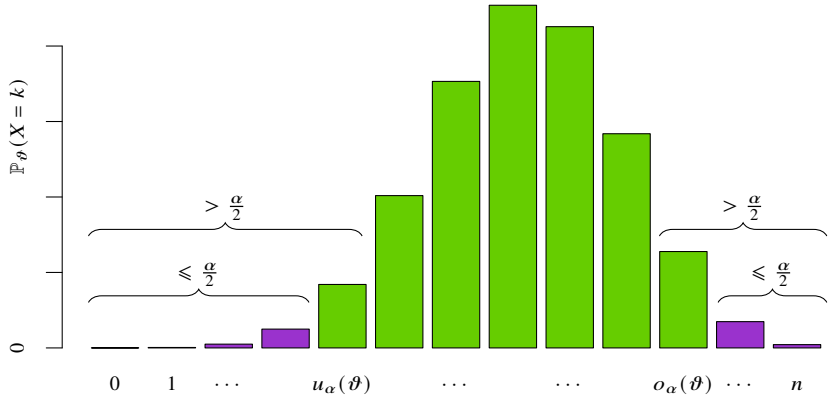

Abb. 1.2 Obere und untere Grenze des Annahmebereichs des zweiseitigen Binomialtests für ein $\vartheta > 1/2$

Abbildung 1.2. Im symmetrischen Fall erhalten wir die einfachere Darstellung $\varphi_\alpha(w) = 1 - \mathbb{1}(|\frac{w}{n} - \frac{1}{2}| \leqslant c_\alpha)$ mit einem geeigneten kritischen Wert c_α.

Laut Statistischem Bundesamt wurden im Jahr 2018 in Hamburg 21.126 Kinder (lebend) geboren. Durch die Berechnung der Quantile der Binomialverteilung führt dies auf

$$u_{0,05}(1/2) = 10.421 \quad \text{und} \quad o_{0,05}(1/2) = 10.705 \quad \text{bzw.} \quad c_{0,05} = 0{,}00672.$$

Von diesen 21.126 Kindern waren 10.215 Mädchen. Setzen wir dies in φ_α ein, erhalten wir

$$\varphi_{0,05}(10.215) = 1 - \mathbb{1}\left(\left|\tfrac{10.215}{21.126} - \tfrac{1}{2}\right| \leqslant c_{0,05}\right) = 1 - 0 = 1,$$

sodass unser Test φ_α zum Niveau $\alpha = 0{,}05$ die Nullhypothese „Es werden genauso viele Jungen wie Mädchen geboren.“ ablehnt.

Bemerkung 1.49 (Normalapproximation) Aufgrund des zentralen Grenzwertsatzes kann die Binomialverteilung $\mathrm{Bin}(n, \vartheta)$ bei hinreichend großen Stichprobenumfängen durch eine Normalverteilung approximiert werden. Es bietet sich in diesem Fall also an, einen *Gauß-Test* zu verwenden, um den Binomialtest zu approximieren: Für $\vartheta \in (0, 1)$ normalisieren wir die Beobachtung $X \sim \mathrm{Bin}(n, \vartheta)$ durch

$$Y := \frac{X - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}.$$

Aus dem zentralen Grenzwertsatz folgt dann, dass die Verteilung von Y für $n \to \infty$ gegen $\mathrm{N}(0, 1)$ konvergiert. Für die Teststatistik $T(X) := |X/n - \vartheta|$ und eine standardnormalverteilte Zufallsvariable $Z \sim \mathrm{N}(0, 1)$ erhalten wir

$$\begin{aligned}
\mathbb{P}_\vartheta(T(X) > c_\alpha) &= \mathbb{P}_\vartheta\Big(\frac{|X - n\vartheta|}{\sqrt{n\vartheta(1-\vartheta)}} > \sqrt{\frac{n}{\vartheta(1-\vartheta)}}c_\alpha\Big) \\
&\overset{n\to\infty}{\approx} \mathbb{P}\Big(|Z| > \sqrt{\frac{n}{\vartheta(1-\vartheta)}}c_\alpha\Big) \\
&= 2\Big(1 - \mathbb{P}\Big(Z \leqslant \sqrt{\frac{n}{\vartheta(1-\vartheta)}}c_\alpha\Big)\Big) \\
&= 2\Big(1 - \Phi\Big(\sqrt{\frac{n}{\vartheta(1-\vartheta)}}c_\alpha\Big)\Big) \overset{!}{=} \alpha,
\end{aligned}$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Ist die Gleichheit für einen kritischen Wert c_α erfüllt, erhalten wir einen nichtrandomisierten Test, der asymptotisch (!) für $n \to \infty$ das Niveau α erreicht. Umformen ergibt

$$c_\alpha = \sqrt{\frac{\vartheta_0(1-\vartheta_0)}{n}}\Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big)$$

mit $\vartheta = \vartheta_0$ unter H_0.

Die Normalapproximation hat den Vorteil, dass im Vergleich zur diskreten Binomialverteilung die Berechnung deutlich vereinfacht wird. Aufgrund der heutigen leistungsfähigen Computer ist dieses Argument in der Praxis allerdings zu vernachlässigen. Zusätzlich vereinfacht die Approximation jedoch die Interpretation der Teststatistik und ermöglicht einen Vergleich zwischen Studien mit verschieden großen Stichprobenumfängen.

Ob die Normalapproximation der Binomialverteilung tatsächlich passend ist, wird in der Praxis mit sogenannten QQ-Plots überprüft.

Bemerkung 1.50 (QQ-Plots) Ein *Quantil-Quantil-Plot*, *QQ-Plot* oder auch *QQ-Diagramm* ist ein exploratives, grafisches Werkzeug, in dem die Quantile zweier Zufallsvariablen gegeneinander aufgetragen werden, um ihre Verteilungen zu vergleichen. Da sie im Allgemeinen von großer Bedeutung ist, erklären wir im Detail, wie man die Verteilung einer Beobachtung mit der Standardnormalverteilung vergleicht: Die *empirische Verteilungsfunktion* einer mathematischen Stichprobe $X_1, \ldots, X_N$ ist definiert als

$$F_N(x) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(X_i \leqslant x).$$

Für große N approximiert F_N die wahre Verteilungsfunktion F, da nach dem starken Gesetz der großen Zahlen $F_N(x) \to \mathbb{E}[\mathbb{1}(X_1 \leqslant x)] = F(x)$ $\mathbb{P}$-f.s. für alle $x \in \mathbb{R}$ gilt (tatsächlich gilt diese Konvergenz sogar gleichmäßig auf $\mathbb{R}$ nach dem Satz von Gliwenko-Cantelli). Im Fall $X_i \sim \mathrm{N}(\mu, \sigma^2)$ gilt $F(x) = \Phi(\frac{x-\mu}{\sigma})$. Für die Quantilfunktion ergibt sich also

$$F^{-1}(\Phi(x)) = \Phi^{-1}(\Phi(x)) \cdot \sigma + \mu = \sigma \cdot x + \mu,$$

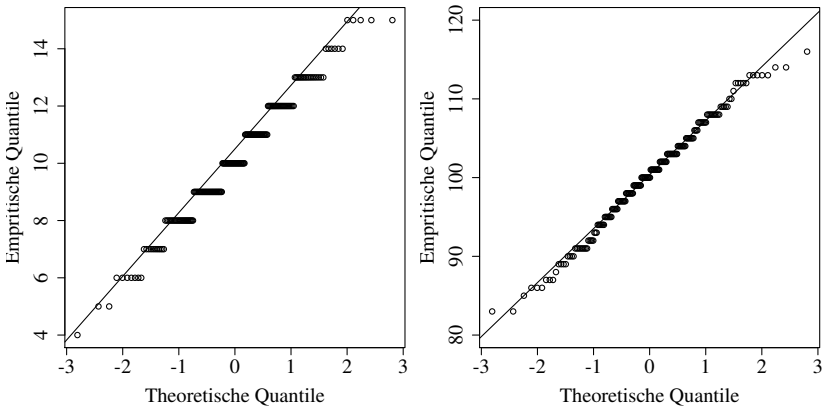

Abb. 1.3 QQ-Plot für eine binomialverteilte Stichprobe mit den Parameter $p = 1/2$, $n = 20$ (links) sowie $n = 200$ (rechts)

und $F^{-1} \circ \Phi$ beschreibt eine Gerade. Im QQ-Plot werden der Größe nach geordnete Werte (x_k) aus dem Intervall $[0, 1]$ in die verallgemeinerte Inverse F_n^{-1} und in Φ^{-1} eingesetzt, sodass man für jedes x_k ein Quantilpaar erhält. Diese Paare werden (als Koordinatenpaare interpretiert) in ein Koordinatensystem eingetragen, und wenn die $X_i \sim \mathrm{N}(\mu, \sigma^2)$ sind, sollten die Quantilpaare in etwa auf einer Geraden liegen.

Zur Illustration betrachten wir eine Stichprobe $X_1, \ldots, X_{200} \sim \mathrm{Bin}(n, 1/2)$ für $n \in \{20, 200\}$ in Abbildung 1.3. Die QQ-Plots zeigen deutlich, dass die Normalapproximation der Binomialverteilung für kleine n problematisch ist, aber für große n gut funktioniert. Genauer sollte $np(1-p)$ hinreichend groß sein, was man aus dem Satz von Berry-Esseen folgern kann. Abbildung 1.3 macht allerdings auch deutlich, dass selbst für $n = 200$ die extremen Quantile der Binomialverteilung noch deutlich von der Normalapproximation abweichen.

Beispiel 1.51 (Planung des Stichprobenumfangs) In Beispiel 1.1 hatten wir uns gefragt, wie viele Studierende an der Umfrage von Saskia teilnehmen sollten, um ein aussagekräftiges Ergebnis zu erhalten. Wie wir im Folgenden sehen werden, kann diese Frage durch eine Betrachtung des Fehlers 2. Art beantwortet werden.

Gemäß Beispiel 1.7 wählen wir das statistische Modell

$$(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (\mathrm{Ber}(\vartheta))_{\vartheta \in [0,1]}^{\otimes n}).$$

Die Summe $Y = X_1 + \cdots + X_n$ der zufriedenen Studentinnen und Studenten ist also wieder binomialverteilt. In Beispiel 1.42 interessierten wir uns für das Testproblem $H_0 : \vartheta \geqslant \vartheta_0$ gegen $H_1 : \vartheta < \vartheta_0$ mit $\vartheta_0 = 3/4$. Wir legen ein Signifikanzniveau $\alpha \in (0, 1)$ fest und betrachten den Test $\varphi_n(x) = \mathbb{1}_{\{x_1 + \ldots + x_n < c_n\}}$. Eine Normalapproximation liefert den kritischen Wert

$$c_n = n\vartheta_0 + \Phi^{-1}(\alpha)\sqrt{n\vartheta_0(1-\vartheta_0)}$$

(dem/der Lesenden sei die Überprüfung überlassen, dass dies tatsächlich ein Test zum asymptotischen Niveau α ist). Wir modifizieren nun die Alternative etwas zu $H_1 : \vartheta < \vartheta_1$ für ein $\vartheta_1 < \vartheta_0$, sodass die Parameter der Hypothese und der Alternative voneinander getrennt sind. Wir wollen die Wahrscheinlichkeit für Fehler 2. Art auf höchstens $1 - \beta$ begrenzen, das heißt, es soll $\inf_{\vartheta \leqslant \vartheta_1} \beta_{\varphi_n}(\vartheta) = \beta_{\varphi_n}(\vartheta_1) \geqslant \beta$ gelten (die Gütefunktion ist monoton fallend). Wir berechnen

$$\begin{aligned}
\beta &\overset{!}{\leqslant} \beta_{\varphi_n}(\vartheta_1) \\
&= \mathbb{P}_{\vartheta_1}\Big(\sum_{j=1}^{n} X_j < c_n\Big) \\
&= \mathbb{P}_{\vartheta_1}\left(\frac{\sum_{j=1}^{n} X_j - n\vartheta_1}{\sqrt{n\vartheta_1(1-\vartheta_1)}} < \frac{c_n - n\vartheta_1}{\sqrt{n\vartheta_1(1-\vartheta_1)}}\right) \\
&= \mathbb{P}_{\vartheta_1}\left(\frac{\sum_{j=1}^{n} X_j - n\vartheta_1}{\sqrt{n\vartheta_1(1-\vartheta_1)}} < \frac{\sqrt{n}(\vartheta_0 - \vartheta_1) + \Phi^{-1}(\alpha)\sqrt{\vartheta_0(1-\vartheta_0)}}{\sqrt{\vartheta_1(1-\vartheta_1)}}\right) \\
&\overset{ZGWS}{\approx} \Phi\left(\sqrt{n}\frac{\vartheta_0 - \vartheta_1}{\sqrt{\vartheta_1(1-\vartheta_1)}} + \Phi^{-1}(\alpha)\sqrt{\frac{\vartheta_0(1-\vartheta_0)}{\vartheta_1(1-\vartheta_1)}}\right).
\end{aligned}$$

Wir erhalten damit die Bedingung

$$n \geqslant \frac{\vartheta_1(1-\vartheta_1)}{(\vartheta_0 - \vartheta_1)^2}\left(\Phi^{-1}(\beta) - \Phi^{-1}(\alpha)\sqrt{\frac{\vartheta_0(1-\vartheta_0)}{\vartheta_1(1-\vartheta_1)}}\right)^2.$$

Insbesondere sehen wir, dass mehr Beobachtungen benötigt werden, um den Fehler 2. Art auch nahe der Nullhypothese, das heißt wenn $\vartheta_0 - \vartheta_1$ klein ist, durch $1 - \beta$ zu beschränken.

Setzen wir $\vartheta_0 = 3/4$, $\vartheta_1 = 0{,}7$, $\alpha = 0{,}1$ und $\beta = 0{,}9$, dann erhalten wir $n \approx 522$, was relativ wenig ist bei einer Grundgesamtheit von 20.000 Studierenden. Wollen wir hingegen die Wahrscheinlichkeiten für Fehler 1. und 2. Art weiter reduzieren, sagen wir mit $\alpha = 0{,}05$ und $\beta = 0{,}95$, erhalten wir $n \approx 860$. $\alpha = 0{,}01$ und $\beta = 0{,}99$ führen zu $n \approx 1720$.

Ein wichtiges Konzept in der Anwendung statistischer Tests sind *p-Werte*. Um diese zu motivieren, bleiben wir beim Saskia-Beispiel. Wir wollen aber das Testproblem etwas umformulieren. Die Nullhypothese ist, dass die Studierenden eher sehr zufrieden mit ihrem Studium sind ($H_0 : \vartheta \in [3/4, 1)$), und wir wollen testen, ob sie mit ihrem Studium eher nicht zufrieden sind ($H_1 : \vartheta \in (0, 3/4)$). Nehmen wir an, wir erhalten durch die Umfrage eine Stichprobe, die im arithmetischen Mittel 0,5 ergibt. Die Hälfte ist also mit dem Studium (un-)zufrieden. Nun könnte man sich fragen, wie stark dieses Ergebnis der Nullhypothese widerspricht, denn immerhin ist die andere Hälfte der Stichprobe zufrieden und es haben (mit sehr hoher Wahr-

scheinlichkeit) nicht alle Studierenden die Umfrage beantwortet. Eine Antwort auf diese Frage liefert der *p-Wert*.

Definition 1.52 Sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\varphi = \mathbb{1}(T > c_\alpha)$ ein Test der Hypothese $H_0\colon \vartheta \in \Theta_0 \neq \emptyset$ mit Teststatistik T und kritischem Wert c_α. Dann ist der **p-Wert** $p_\varphi(x)$ einer Realisierung $x \in \mathcal{X}$ definiert als

$$p_\varphi(x) = \inf_{t<t^*} \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) > t) \qquad \text{für } t^* = T(x).$$

Der p-Wert ist also die Wahrscheinlichkeit, bei Gültigkeit der Hypothese etwas mindestens so Extremes zu beobachten, wie das tatsächlich Beobachtete. Um den Zusammenhang zu Niveau-α-Tests zu erkennen, betrachten wir zunächst den Spezialfall einer einfachen Hypothese $\Theta_0 = \{\vartheta_0\}$ und einer stetigen, streng monoton wachsenden Verteilungsfunktion von T unter $\mathbb{P}_{\vartheta_0}$. Dann vereinfacht sich obige Definition zu

$$p_\varphi(x) = \mathbb{P}_{\vartheta_0}\big(T(X) > T(x)\big).$$

Unter diesen Bedingungen gilt zudem $\mathbb{P}_{\vartheta_0}(T > c_\alpha) = \alpha$. Wir können nun durch einen Vergleich des p-Werts mit dem Niveau ablesen, ob die Hypothese abgelehnt wird oder nicht:

- Falls $p_\varphi(x) < \alpha$, dann folgt aus der Monotonie von $c \mapsto \mathbb{P}_{\vartheta_0}(T > c)$, dass $T(x) > c_\alpha$. Der Test $\varphi(x) = \mathbb{1}(T(x) > c_\alpha)$ lehnt die also Hypothese ab.
- Falls $p_\varphi(x) > \alpha$, ergibt sich analog $T(x) < c_\alpha$, sodass der Test die Hypothese nicht verwirft.

Wir werden diesen Zusammenhang auch unter deutlich allgemeineren Bedingungen in Satz 1.53 nachweisen. Statt nur zu prüfen, ob ein Test eine Hypothese akzeptiert oder ablehnt, gibt der p-Wert (auch „Signifikanzwahrscheinlichkeit“, „Überschreitungswahrscheinlichkeit“ oder „Signifikanzwert“) das kleinste Signifikanzniveau an, zu dem eine Hypothese abgelehnt würde. Damit gibt der p-Wert Aufschluss darüber, „wie stark“ die Daten der Hypothese widersprechen.

Der p-Wert spielt in der Wissenschaft (Biologie, Chemie, Medizin, Physik, Soziologie, ...) eine wichtige Rolle als Qualitätsmaß für die Relevanz beziehungsweise Richtigkeit einer Theorie. Zum Beispiel wurde in der Arbeit Abbott et al. (2016), die erstmals die Existenz der Gravitationswellen nachgewiesen hat, ein p-Wert von $7{,}5 \times 10^{-8}$ angegeben. Zudem hat der p-Wert für die Theorie des *multiplen Testens* eine große Bedeutung und ist als Zwischenresultat für Methoden, die beispielsweise die *false discovery rate* kontrollieren, eine fundamentale Größe, siehe Dickhaus (2014).

Satz 1.53 (Eigenschaften des p-Werts) *Es seien $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\alpha_0 \in (0,1)$ und $\varphi = \mathbb{1}(T > c_{\alpha_0})$ ein Niveau-α_0-Test der Hypothese $H_0\colon \vartheta \in \Theta_0 \neq \emptyset$ mit einer Teststatistik $T\colon \mathcal{X} \to \mathbb{R}$ und kritischen Werten*

$$c_\alpha = \inf\Big\{c \in \mathbb{R} : \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) > c) \leqslant \alpha\Big\}, \quad \alpha \in (0,1). \tag{1.12}$$

Unter der Annahme

$$\sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T > c_\alpha) \leqslant \alpha \qquad \textit{für alle } \alpha \in (0,1)$$

gilt die Darstellung

$$p_\varphi(x) := \inf_{\alpha: T(x) > c_\alpha} \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) > c_\alpha). \tag{1.13}$$

sowie der Testentscheid

$$\varphi(x) = \begin{cases} 1, & \textit{falls } p_\varphi(x) < \alpha_0, \\ 0, & \textit{falls } p_\varphi(x) > \alpha_0. \end{cases}$$

Bemerkung 1.54 (Kritische Werte) Im Allgemeinen muss die Eigenschaft $\sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) > c_\alpha) \leqslant \alpha$ für die kritischen Werte c_α aus (1.12) nicht erfüllt sein. Dieses Defizit kann behoben werden indem man die Gültigkeit der Vertauschungsbedingung

$$\lim_{c_n\downarrow c} \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T > c_n) = \sup_{\vartheta\in\Theta_0} \lim_{c_n\downarrow c} \mathbb{P}_\vartheta(T > c_n) \tag{1.14}$$

fordert. In diesem Fall folgt aus der σ-Stetigkeit von Maßen

$$\sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T > c_\alpha) = \sup_{\vartheta\in\Theta_0} \lim_{c_n\downarrow c_\alpha} \mathbb{P}_\vartheta(T > c_n) = \lim_{c_n\downarrow c_\alpha} \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T > c_n) \leqslant \alpha.$$

Gleichung (1.14) gilt beispielsweise für einfache Hypothesen, das heißt für $\Theta_0 = \{\vartheta_0\}$, da dann $\sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta = \mathbb{P}_{\vartheta_0}$ ein Maß ist, oder wenn Montonie der Verteilungsfunktion von $T(X)$ in ϑ wie in Beispiel 1.56 unten verwendet werden kann.

Beweis Im Folgenden schreiben wir $\mathbb{P}_0 := \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta$. Wir beginnen mit folgenden elementaren Beobachtungen:

(i) Aus der Monotonie der Maße $\mathbb{P}_\vartheta$ folgt sofort, dass die Abbildung $c \mapsto \mathbb{P}_0(T > c)$ monoton fallend ist.

(ii) Sei $c < t^*$, $\alpha = \mathbb{P}_0(T > c)$ und c_α der entsprechende kritische Wert. Aus der Definition der kritischen Werte folgt sofort $c_\alpha \leqslant c$, und laut Voraussetzung gilt $\mathbb{P}_0(T > c_\alpha) \leqslant \alpha$. Aus der Monotonie (i) folgt dann aber

$$\alpha = \mathbb{P}_0(T > c) \leqslant \mathbb{P}_0(T > c_\alpha) \leqslant \alpha,$$

und somit $\mathbb{P}_0(T > c_\alpha) = \alpha$.

Setze $S := \{c_\alpha < t^*, \alpha \in [0,1]\}$ die Menge aller kritischen Werte kleiner als t^*. Dann gilt $S \neq \emptyset$, denn für ein beliebiges $c < t^*$ und den kritischen Wert c_α zu $\alpha = \mathbb{P}_0(T > c)$ gilt $c_\alpha \leqslant c < t^*$ und somit $c_\alpha \in S$. Wir betrachten nun das Supremum $s := \sup\{c_\alpha < t^*\} \leqslant t^*$ der kritischen Werte kleiner t^* und unterscheiden die Fälle $s = t^*$ und $s < t^*$.

Falls $s = t^*$ gilt, gibt es eine Folge kritischer Werte $(c_{\alpha,n})$ mit $\lim_{n\to\infty} c_{\alpha,n} = t^*$ und wir erhalten direkt die Darstellung

$$p_\varphi(x) = \inf_{t<t^*} \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) > t) = \inf_{\alpha:t^*>c_\alpha} \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) > c_\alpha).$$

Wir betrachten nun den Fall $s < t^*$. Wir zeigen zuerst, dass s ein kritischer Wert zum Niveau $\alpha_s = \mathbb{P}_0(T > s)$ ist. Wegen der Monotonie (i) und (ii) gilt für alle $c_\alpha \in S$

$$\mathbb{P}_0(T > c_\alpha) \geqslant \mathbb{P}_0(T > s) = \alpha_s = \mathbb{P}_0(T > c_{\alpha_s}).$$

Aus der Monotonie (i) folgt dann $c_{\alpha_s} \geqslant c_\alpha$ für alle $c_\alpha \in S$. Damit ist c_{α_s} eine obere Schranke für S und muss aufgrund von $c_{\alpha_s} \leqslant s$ mit dem Supremum s übereinstimmen. Aus $c_{\alpha_s} = s < t^*$ und der Monotonie (i) folgt daher

$$\mathbb{P}_0(T > c_{\alpha_s}) \geqslant \inf_{\alpha:t^*>c_\alpha} \mathbb{P}_0(T > c_\alpha) \geqslant \mathbb{P}_0(T > s), \tag{1.15}$$

d.h. $\inf_{\alpha:t^*>c_\alpha} \mathbb{P}_0(T > c_\alpha) = \mathbb{P}_0(T > s)$.

Als Nächstes zeigen wir, dass zwischen s und t^* keine „Masse“ mehr ist. Dazu argumentieren wir per Widerspruch: Angenommen, es gibt ein $s < s' < t^*$, sodass $\mathbb{P}_0(s' < T) < \mathbb{P}_0(s < T) = \alpha_s$. Für $\alpha' = \mathbb{P}_0(s' < T)$ mit kritischem Wert $c_{\alpha'} \leqslant s'$ folgt dann

$$\alpha_s = \mathbb{P}_0(T > s) > \mathbb{P}_0(T > s') = \alpha' \overset{(ii)}{=} \mathbb{P}_0(T > c_{\alpha'}).$$

Insbesondere muss als $c_{\alpha'} > s$ gelten im Widerspruch zu $c_{\alpha'} \leqslant s'$. Damit ergibt sich

$$\inf_{s\leqslant s'<t^*} \mathbb{P}_0(T > s') = \mathbb{P}_0(T > s),$$

und zusammen mit (1.15) folgt die Gültigkeit von (1.13).

Man beachte nun, dass $\varphi(x) = 1$ genau dann gilt, wenn $T(x) = t^* > c_{\alpha_0}$. Die Annahme an die kritischen Werte und die Monotonie (i) von $\mathbb{P}_0$ liefern nun

$$\alpha_0 \geqslant \mathbb{P}_0(T > c_{\alpha_0}) \geqslant \inf_{\alpha:c_\alpha<t^*} \mathbb{P}_0(T > c_\alpha) = p_\varphi(x).$$

Äquivalent impliziert $p_\varphi(x) > \alpha_0$, dass $\varphi(x) = 0$ gelten muss. Im Fall $p_\varphi(x) < \alpha_0$ folgt aus

$$p_\varphi(x) = \inf_{c<t^*} \mathbb{P}_0(T > c),$$

dass es ein $c < t^*$ mit $p_{\varphi(x)} \leqslant \mathbb{P}_0(T > c) \leqslant \alpha_0$ gibt. Die Definition der kritischen Werte impliziert nun, dass $c \geqslant c_{\alpha_0}$ gelten muss, und wegen $t^* > c \geqslant c_{\alpha_0}$ folgt $\varphi(x) = 1$. □

Bemerkung 1.55 (p-Wert)

1. Unter den Voraussetzungen von Satz 1.53 kann man den p-Wert alternativ auch durch

$$p_\varphi(x) = \inf\{\alpha \in [0,1] : t^* > c_\alpha\}$$

darstellen (Übung 1.9).

2. Alle Rahmenbedingungen des Experiments, insbesondere also das Signifikanzniveau, müssen vor seiner Durchführung festgelegt werden! Ein Signifikanzniveau darf nicht a posteriori aufgrund der erzielten p-Werte bestimmt werden. Dies widerspricht korrekter statistischer Praxis! Mathematisch wäre α eine Zufallsvariable (als Funktion in den Beobachtungen), und der vorangegangene Satz kann nicht angewendet werden.
3. Der Vorteil von p-Werten ist, dass sie unabhängig von einem a priori festgesetzten Signifikanzniveau α berechnet werden können. Deshalb werden in allen gängigen Statistik-Softwaresystemen statistische Hypothesentests über die Berechnung von p-Werten implementiert.

Beispiel 1.56 (p-Wert) Auf Grundlage von i.i.d. normalverteilten Beobachtungen $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ mit unbekanntem Mittelwert $\mu \in \mathbb{R}$ und bekanntem $\sigma > 0$ soll das Testproblem

$$H_0 \colon \mu \leqslant \mu_0 \qquad \text{gegen} \qquad H_1 \colon \mu > \mu_0$$

untersucht werden. Da wir bereits wissen, dass im resultierenden statistischen Modell $(\mathbb{R}^n, \mathcal{B}(\mathbb{R})^{\otimes n}, (\mathbb{P}_\mu)_{\mu\in\mathbb{R}})$ mit $\mathbb{P}_\mu = \mathrm{N}(\mu, \sigma^2)^{\otimes n}$ das Stichprobenmittel $\overline{X}_n$ ein guter Schätzer von μ (sogar ein Minimax-Schätzer) ist, wählen wir den Test $\varphi(x) = \mathbb{1}_{\{\overline{x}_n > c_\alpha\}}$ für kritische Werte $(c_\alpha)_{\alpha\in(0,1)}$ und der Teststatistik $T(x) = \overline{x}_n$. Wegen $\overline{X}_n \sim \mathrm{N}(\mu, \sigma^2/n)$ unter $\mathbb{P}_\mu$ und der Monotonie der Verteilungsfunktion Φ der Standardnormalverteilung gilt für jedes $c \in \mathbb{R}$

$$\sup_{\mu\leqslant\mu_0} \mathbb{P}_\mu(\overline{X}_n > c) = \sup_{\mu\leqslant\mu_0} \left\{1 - \Phi\Big(\frac{\sqrt{n}(c-\mu)}{\sigma}\Big)\right\} = 1 - \Phi\Big(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\Big).$$

Zum einen folgt hieraus $c_\alpha = \mu_0 + \frac{\sigma}{\sqrt{n}} q_{1-\alpha}$ für das $(1-\alpha)$-Quantil von $\mathrm{N}(0,1)$ und zum anderen ist die Bedingung $\sup_{\mu\leqslant\mu_0} \mathbb{P}_\mu(\overline{X}_n > c_\alpha) \leqslant \alpha$ aus Satz 1.53 erfüllt. Wir erhalten für den *einseitigen Gauß-Test* $\varphi(x) = \mathbb{1}_{\{x > \mu_0 + \frac{\sigma}{\sqrt{n}} q_{1-\alpha}\}}$ und für Realisierungen $x \in \mathbb{R}^n$ die p-Werte

$$p_\varphi(x) = 1 - \Phi\Big(\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma}\Big) = \inf\big\{\alpha \in (0,1) : T(x) > c_\alpha\big\}.$$

1.3.2 Das Neyman-Pearson-Lemma

Wir werden nun ein Optimalitätskriterium für Tests kennenlernen. Zur Motivation kommen wir auf Beispiel 1.46 zurück.

Beispiel 1.57 (Randomisierte Tests) Wir betrachten erneut ein Binomialmodell $\mathrm{Bin}(13, p)$ auf $\mathcal{X} = \{0, \dots n\}$. Da die Schwelle c des Binomialtests eine ganze

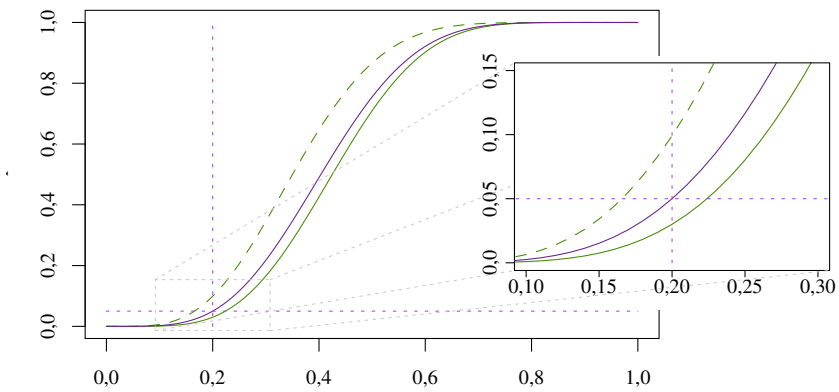

Abb. 1.4 Die Gütefunktionen der Tests $\varphi(x) := \mathbb{1}_{\{x>5\}}$ (grün, durchgezogen), $\overline{\varphi}(x) := \mathbb{1}_{\{x>4\}}$ (grün, gestrichelt) sowie des randomisierten Tests $\widetilde{\varphi}$ (violett) im $\mathrm{Bin}(13, 1/5)$-Modell. Violett markiert sind das Niveau $\alpha = 0{,}05$ und der Parameterwert $p = 1/5$.

Zahl ist, kann der Test $\varphi(x) := \mathbb{1}_{\{x>c\}}$ das Niveau $\alpha = 0{,}05$ insofern nicht voll ausschöpfen, als dass die Wahrscheinlichkeit für Fehler 2. Art nicht perfekt minimiert werden kann. An dieser Stelle hilft eine Randomisierung des Tests. Wir betrachten einen Test der Form

$$\widetilde{\varphi}(x) = \begin{cases} 0, & x < c, \\ \gamma, & x = c, \\ 1, & x > c. \end{cases}$$

Statt die Hypothese wie in Beispiel 1.46 bei $X = c$ immer abzulehnen, fordern wir nun, dass im Fall $X = c$ die Nullhypothese mit einer Wahrscheinlichkeit γ abgelehnt wird, und führen hierzu ein unabhängiges Bernoulli-Experiment $B \sim \mathrm{Ber}(\gamma)$ mit Erfolgswahrscheinlichkeit γ durch. Wir wählen γ so, dass das gesamte Signifikanzniveau ausgeschöpft wird, das heißt

$$0{,}05 \stackrel{!}{=} \mathbb{E}_{1/5}[\widetilde{\varphi}] = \mathbb{P}_{1/5}(X > 5) + \gamma \cdot \mathbb{P}_{1/5}(X = 5).$$

Durch Umstellen nach γ ergibt sich $\gamma \approx 0{,}29$. Der resultierende randomisierte Test $\widetilde{\varphi}$ besitzt somit das Niveau $\alpha = 0{,}05$ und ist zudem unverfälscht, siehe Abbildung 1.4. Da die Gütefunktion des Binomialtests stets unter der Gütefunktion des randomisierten Tests liegt, hat Letzterer einen kleineren Fehler zweiter Art.

In diesem Beispiel ist es uns gelungen, den Binomialtest zu verbessern. Es stellt sich die Frage, ob es auch einen besten Test gibt und, falls das der Fall ist, wie man diesen konstruieren kann. Die Antwort hierauf wollen wir im einfachen binären Modell studieren.

In diesem Fall besteht die Parametermenge nur aus zwei Parametern, sagen wir $\Theta = \{0, 1\}$. Wir möchten $H_0 \colon \vartheta = 0$ gegen $H_1 \colon \vartheta = 1$ testen. Das folgende grundlegende Resultat beschreibt Tests, die zu vorgegebenem Niveau die Wahrscheinlichkeit von Fehlern zweiter Art minimieren.

Satz 1.58 (Neyman-Pearson-Lemma) *Betrachte ein binäres statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\{0,1\}})$ mit Likelihood-Funktion L bezüglich eines dominierenden Maßes μ (zum Beispiel $\mu = \mathbb{P}_0 + \mathbb{P}_1$) sowie das Testproblem $H_0 \colon \vartheta = 0$ gegen $H_1 \colon \vartheta = 1$. Dann besitzt der (möglicherweise randomisierte) Test*

$$\varphi_\alpha(x) := \begin{cases} 1, & \text{falls } L(1,x) > c_\alpha L(0,x), \\ 0, & \text{falls } L(1,x) < c_\alpha L(0,x), \\ \gamma_\alpha, & \text{falls } L(1,x) = c_\alpha L(0,x) \end{cases} \tag{1.16}$$

mit Konstanten $c_\alpha \geqslant 0$, $\gamma_\alpha \in [0,1]$ unter allen (auch randomisierten) Tests φ mit demselben Niveau $\mathbb{E}_0[\varphi] \leqslant \mathbb{E}_0[\varphi_\alpha]$ die kleinste Wahrscheinlichkeit für Fehler zweiter Art:

$$\mathbb{E}_1[1 - \varphi_\alpha] = \min_\varphi \mathbb{E}_1[1 - \varphi]$$

Beweis Das Resultat folgt, wenn wir $E_1[\varphi - \varphi_\alpha] \leqslant 0$ gezeigt haben. Hierzu verwenden wir ein geschicktes Maßwechselargument sowie $\varphi - \varphi_\alpha = \varphi - 1 \leqslant 0$ auf $\{L(1) > c_\alpha L(0)\}$ und $\varphi - \varphi_\alpha = \varphi \geqslant 0$ auf $\{L(1) < c_\alpha L(0)\}$:

$$\begin{aligned}
\mathbb{E}_1[\varphi - \varphi_\alpha] &= \int_{\{L(1)>c_\alpha L(0)\}} (\varphi - \varphi_\alpha) L(1) \mathrm{d}\mu + \int_{\{L(1)<c_\alpha L(0)\}} (\varphi - \varphi_\alpha) L(1) \mathrm{d}\mu \\
&\quad + \int_{\{L(1)=c_\alpha L(0)\}} (\varphi - \varphi_\alpha) L(1) \mathrm{d}\mu \\
&\leqslant \int_{\{L(1)>c_\alpha L(0)\}} (\varphi - \varphi_\alpha) c_\alpha L(0) \mathrm{d}\mu \\
&\quad + \int_{\{L(1)<c_\alpha L(0)\}} (\varphi - \varphi_\alpha) c_\alpha L(0) \mathrm{d}\mu \\
&\quad + \int_{\{L(1)=c_\alpha L(0)\}} (\varphi - \varphi_\alpha) c_\alpha L(0) \mathrm{d}\mu \\
&= c_\alpha \mathbb{E}_0[\varphi - \varphi_\alpha] \\
&\leqslant 0,
\end{aligned}$$

wobei wir in der letzten Zeile die Niveaubedingung $\mathbb{E}_0[\varphi] \leqslant \mathbb{E}_0[\varphi_\alpha]$ verwendet haben. □

Kurzbiografie (Jerzy Neyman) Jerzy Neyman (bzw. Yuri Czeslawovich) wurde 1894 in Bendery im Russischen Kaiserreich geboren. Er studierte Mathematik und Physik in Charkiw (Ukraine). Nach dem Studium arbeitete er zunächst an den Universtäten in Charkiw und Warschau, bevor er 1934 von Egon Pearson ans University College London geholte wurde. 1938 bekam er einen Ruf an die University of

California in Berkeley, an der er das Statistik-Department aufbaute. 1981 starb Neymann in Kalifornien. Zu den wichtigsten Beiträgen Neymans zählen die Einführung von Konfidenzintervallen und das zusammen mit Egon Pearson bewiesene *Neyman-Pearson-Lemma*.

Definition 1.59 In einem binären statistischen Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\{0,1\}})$ mit Likelihood-Funktion L bezüglich eines dominierenden Maßes heißt ein Test der Form (1.16) **Neyman-Pearson-Test**.

Bemerkung 1.60

1. Es sei der Leserin überlassen, zu zeigen, dass zu jedem Niveau $\alpha \in (0,1)$ durch Wahl von $c_\alpha \geqslant 0$ und $\gamma_\alpha \in [0,1]$ ein Neyman-Pearson-Test φ_α vom Niveau α existiert (Aufgabe 2.10). Im Fall einer einelementigen Hypothese und Alternative, man spricht auch von *einfacher Hypothese* und *einfacher Alternative*, kann somit stets ein optimaler Test angegeben werden. Aus dem vorherigen Beweis lässt sich auch folgern, dass jeder in diesem Sinne optimale Test fast sicher von der Form eines Neyman-Pearson-Tests ist.
2. Die Neyman-Pearson-Tests hängen nicht von der Wahl des dominierenden Maßes μ ab. Mittels Radon-Nikodym-Ableitungen sieht man nämlich mit dem kanonischen dominierenden Maß $\overline{\mu} = \mathbb{P}_0 + \mathbb{P}_1$ und $\overline{L}(\vartheta, x) = \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\overline{\mu}}(x)$:

$$L(\vartheta, x) = \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x) = \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\overline{\mu}}(x)\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu}(x) = \overline{L}(\vartheta, x)\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu}(x).$$

   Die Likelihood-Funktionen L und $\overline{L}$ unterscheiden sich also nur um einen in ϑ konstanten Faktor, der sich bei der Konstruktion des Neyman-Pearson-Tests herauskürzt. Beachte dazu $\mathbb{P}_\vartheta(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu} = 0) = 0$ für $\vartheta \in \{0,1\}$ (Warum gilt das?).

Beispiel 1.61 (Einseitiger Binomialtest) Wir nehmen Beispiel 1.46 wieder auf, testen aber nun für $n = 13$ und $p_0 = 1/5$, $p_1 = 1/4$ die einfache Hypothese „Tod durch Tumorerkrankung mit Wahrscheinlichkeit $p = p_0$“ gegen die einfache Alternative "Tod durch Tumorerkrankung mit Wahrscheinlichkeit $p = p_1$“. Wir erhalten das binäre statistische Modell $(\{0,\dots,n\}, \mathcal{P}(\{0,\dots,n\}), (\mathbb{P}_\vartheta)_{\vartheta\in\{0,1\}})$ mit $\mathbb{P}_0 = \mathrm{Bin}(n; p_0)$ und $\mathbb{P}_1 = \mathrm{Bin}(n, p_1)$. Mit dem Zählmaß als dominierendem Maß μ erhalten wir den *Likelihood-Quotienten*

$$\frac{L(1,k)}{L(0,k)} = \frac{\binom{n}{k}p_1^k(1-p_1)^{n-k}}{\binom{n}{k}p_0^k(1-p_0)^{n-k}} = \frac{(1-p_1)^n}{(1-p_0)^n}\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^k.$$

Wegen $p_1 > p_0$ wächst dieser Quotient in $k \in \{0,\dots,n\}$ streng monoton, und es gibt zu jedem $c_\alpha \geqslant 0$ ein $\widetilde{c}_\alpha \geqslant 0$, sodass

$$\varphi_\alpha(k) := \begin{cases} 1, & \text{falls } L(1,k) > c_\alpha L(0,k) \\ 0, & \text{falls } L(1,k) < c_\alpha L(0,k) \\ \gamma_\alpha, & \text{falls } L(1,k) = c_\alpha L(0,k) \end{cases} = \begin{cases} 1, & \text{falls } k > \widetilde{c}_\alpha, \\ 0, & \text{falls } k < \widetilde{c}_\alpha, \\ \gamma_\alpha, & \text{falls } k = \widetilde{c}_\alpha. \end{cases}$$

Wählen wir nun wie in Beispiel 1.57 $\widetilde{c}_{0,05} = 5$, $\gamma_{0,05} \approx 0{,}29$, so erreicht $\varphi_{0,05}$ genau das Niveau $\alpha = 0{,}05$ und besitzt als Neyman-Pearson-Test unter allen randomisierten Tests zum Niveau $\alpha = 0{,}05$ die kleinste Fehlerwahrscheinlichkeit zweiter Art.

Der aufmerksamen Leserin ist vielleicht aufgefallen, dass die gesamte Testkonstruktion hier nicht vom exakten Wert p_1 abhängt. Wir haben nur $p_1 > p_0$ benutzt. Somit gilt sogar die weit stärkere Eigenschaft, dass es keinen Test φ von $H_0 \colon p = p_0$ gegen $H_1 \colon p > p_0$ vom Niveau $\alpha = 0{,}05$ gibt, der für irgendein $p \in (p_0, 1]$ eine größere Güte $\mathbb{E}_p[\varphi]$ besitzt als $\varphi_{0,05}$. Wenn wir noch das Monotonieargument aus Beispiel 1.46 heranziehen, dass $\varphi_{0,05}$ auch für die zusammengesetzte Hypothese $H_0 \colon p \leqslant p_0$ Niveau α besitzt, so können wir weiter schließen, dass der einseitige Binomialtest auf der Alternative maximale Güte unter allen Niveau-α-Tests für $H_0 \colon p \leqslant p_0$ gegen $H_1 \colon p > p_0$ besitzt. Solche Tests nennt man auch *UMP-Tests* (*uniformly most powerful tests*).

Beispiel 1.62 (Einseitiger Gauß-Test) Wir beobachten eine mathematische Stichprobe normalverteilter Zufallsvariablen $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ mit $\sigma > 0$ bekannt und wollen für feste $\mu_0, \mu_1 \in \mathbb{R}$ die Hypothese $H_0 \colon \mu = \mu_0$ gegen $H_1 \colon \mu = \mu_1$ testen. Mit $\mathbb{P}_0 = \mathrm{N}(\mu_0, \sigma^2)^{\otimes n}$, $\mathbb{P}_1 = \mathrm{N}(\mu_1, \sigma^2)^{\otimes n}$ und dem n-dimensionalen Lebesgue-Maß als dominierendem Maß erhalten wir den Likelihood-Quotienten (als Statistik in den Beobachtungen geschrieben):

$$\begin{aligned}\frac{L(1)}{L(0)} &= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big((X_i-\mu_1)^2-(X_i-\mu_0)^2\Big)\right)\\ &= \exp\left(-\frac{n}{\sigma^2}\left((\mu_0-\mu_1)\overline{X}_n+\frac{\mu_1^2}{2}-\frac{\mu_0^2}{2}\right)\right)\end{aligned}$$

mit dem Stichprobenmittel $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Betrachten wir von nun an den Fall $\mu_1 > \mu_0$, sehen wir, dass der Likelihood-Quotient eine streng monotone Funktion in $\overline{X}_n$ ist und sonst nicht von den Beobachtungen abhängt. Also können wir wie beim Binomialtest durch Modifikation des kritischen Werts c_α jeden Neyman-Pearson-Test schreiben als

$$\varphi_\alpha := \begin{cases} 1, & \text{falls } \overline{X}_n > \widetilde{c}_\alpha,\\ 0, & \text{falls } \overline{X}_n < \widetilde{c}_\alpha,\\ \gamma_\alpha, & \text{falls } \overline{X}_n = \widetilde{c}_\alpha \end{cases}$$

mit $\widetilde{c}_\alpha \geqslant 0$, $\gamma_\alpha \in [0,1]$ geeignet. Unter $\mathbb{P}_0$ und $\mathbb{P}_1$ hat das Ereignis $\{\overline{X}_n = \widetilde{c}_\alpha\}$ die Wahrscheinlichkeit Null, sodass wir auf Randomisierung verzichten und einfach $\gamma_\alpha = 0$ setzen können. Da unter $\mathbb{P}_0$ für das Stichprobenmittel $\overline{X}_n \sim \mathrm{N}(\mu_0, \sigma^2/n)$ gilt, erhalten wir somit einen Neyman-Pearson-Test vom Niveau $\alpha \in (0,1)$ durch

$$\varphi_\alpha = \mathbb{1}(\overline{X}_n > \mu_0 + \sigma n^{-1/2} q_{1-\alpha})$$

mit dem $(1-\alpha)$-Quantil $q_{1-\alpha}$ der Standardnormalverteilung, die gerade mit dem einseitigen Gauß-Test aus Beispiel 1.56 übereinstimmt. Genau wie beim einseitigen

Binomialtest sieht man, dass φ_α sogar ein Test vom Niveau α für $H_0 \colon \mu \leqslant \mu_0$ gegen $H_1 \colon \mu > \mu_0$ mit UMP-Eigenschaft ist.

Wie diese Beispiele demonstrieren, lässt sich die Neyman-Pearson-Theorie manchmal auch vom Fall einfacher Hypothesen auf einseitige Testprobleme übertragen. Die notwendige Strukturvoraussetzung dafür waren immer monotone Likelihood-Quotienten. Für die wichtigen zweiseitigen Testprobleme führt dieser Ansatz jedoch nicht zum Ziel, man kann sogar zeigen, dass für zweiseitige Binomial- oder Gauß-Testprobleme keine UMP-Tests existieren können. Stattdessen sollte für zweiseitige Testprobleme die Klasse der unverfälschten Tests betrachtet werden, für die unter geeigneten Bedingungen die Existenz von *besten, unverzerrten Tests* (englisch: *uniformly most powerfull unbiased*, kurz: UMPU) gezeigt werden kann.

Auch für allgemeine zusammengesetzte Hypothesen führt uns die Neyman-Pearson-Theorie auf einen intuitiven Ansatz zur Wahl der Teststatistik. Nehmen wir dazu ein statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit Likelihood-Funktion L und Partition $\Theta = \Theta_0 \cup \Theta_1$ an, so können wir eine Teststatistik für $H_0 \colon \vartheta \in \Theta_0$ gegen $H_1 \colon \vartheta \in \Theta_1$ konstruieren, indem wir in Θ_0 und Θ_1 jeweils den Parameter wählen, der die Likelihood-Funktion maximiert, und damit, dem Neymann-Pearson-Ansatz folgend, den entsprechenden verallgemeinerten Likelihood-Quotienten bilden:

$$T(x) := \frac{\sup_{\vartheta\in\Theta_1} L(\vartheta, x)}{\sup_{\vartheta\in\Theta_0} L(\vartheta, x)}, \tag{1.17}$$

wobei formal $a/0 := +\infty$ gesetzt werde. Damit erhalten wir eine sehr allgemeine Methode, Tests zu konstruieren.

Methode 1.63 (Likelihood-Quotiententest) Für ein dominiertes statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit Likelihood-Funktion $L(\vartheta, x)$ betrachte das Testproblem $H_0 \colon \vartheta \in \Theta_0$ gegen $H_1 \colon \vartheta \in \Theta_1$. Für $c_\alpha \geqslant 0$ und $\gamma_\alpha \in [0, 1]$ sowie T aus (1.17) ist ein **Likelihood-Quotiententest** (englisch: *likelihood ratio test*, kurz: LR-Test) gegeben durch

$$\varphi_\alpha(x) = \mathbb{1}(T(x) > c_\alpha) + \gamma_\alpha \mathbb{1}(T(x) = c_\alpha), \quad x \in \mathcal{X}.$$

Es ist leicht einzusehen, dass unter H_i, $i = 0, 1$, fast sicher $\sup_{\vartheta\in\Theta_i} L(\vartheta) > 0$ gilt, sodass ein Likelihood-Quotiententest φ_α fast sicher wohldefiniert ist (ist der Nenner in T gleich null, so setze $T = +\infty$, da dies fast sicher nur unter H_1 geschieht, wo der Zähler fast sicher positiv ist).

Eine andere Interpretation des Likelihood-Quotiententests ergibt sich, wenn die Maximum-Likelihood-Schätzer $\widehat{\vartheta}_0$ und $\widehat{\vartheta}_1$ von ϑ über den Parametermengen Θ_0 bzw. Θ_1 existieren. Dann ist der Likelihood-Quotient gerade

$$T(x) = \frac{L(\widehat{\vartheta}_1(x), x)}{L(\widehat{\vartheta}_0(x), x)},$$

und wir können den Likelihood-Quotiententest als einen Neyman-Pearson-Test zwischen den geschätzten Parametern $\widehat{\vartheta}_0$ und $\widehat{\vartheta}_1$ interpretieren. Beachte dazu aber,

dass die Schätzer zufällig sind und das Neyman-Pearson-Lemma keine Anwendung findet. Ähnlich wie bei der Maximum-Likelihood-Methode führt der Likelihood-Quotiententest häufig, aber nicht immer zu guten Tests.

Beispiel 1.64 (Zweiseitiger Binomialtest) In Beispiel 1.48 haben wir den zweiseitigen Binomialtest für $\mathbb{P}_\vartheta = \mathrm{Bin}(n, \vartheta)$ und $H_0 \colon \vartheta = \vartheta_0$ gegen $H_1 \colon \vartheta \neq \vartheta_0$ kennengelernt. Als Likelihood-Quotientenstatistik erhalten wir für $\vartheta \in (0,1)$ fest und $\widehat{\vartheta}(k) = k/n$

$$T(k) = \frac{\sup_{\vartheta \neq \vartheta_0} \vartheta^k (1-\vartheta)^{n-k}}{\vartheta_0^k (1-\vartheta_0)^{n-k}} = \frac{\widehat{\vartheta}(k)^k (1-\widehat{\vartheta}(k))^{n-k}}{\vartheta_0^k (1-\vartheta_0)^{n-k}}, \quad k \in \{0, \dots, n\}.$$

Im Fall $\vartheta_0 = 1/2$ vereinfacht sich dies durch Einsetzen zu $T(k) = (2n)^{-n} k^k (n-k)^{n-k}$. Damit ist T um $n/2$ symmetrisch ($T(k) = T(n-k)$ für $0 \leqslant k \leqslant n/2$) und wachsend ($T(k+1) > T(k)$ für $k \geqslant n/2$). Der Likelihood-Quotiententest lässt sich somit schreiben als $\varphi_\alpha(k) = \mathbb{1}(|k - n/2| > \widetilde{c}_\alpha) + \gamma_\alpha \mathbb{1}(|k - n/2| = \widetilde{c}_\alpha)$, was genau dem zweiseitigen Binomialtest entspricht. Beachte, dass für $\vartheta_0 \neq 1/2$ die Asymmetrie der $\mathrm{Bin}(n, \vartheta_0)$-Verteilung zu einer Teststatistik führt, die nicht symmetrisch um $\vartheta_0 n$ ist.

Beispiel 1.65 (Zweiseitiger Gauß-Test) Aufgrund einer mathematischen Stichprobe $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ mit bekanntem $\sigma > 0$ wollen wir für festes $\mu_0 \in \mathbb{R}$ die Hypothese $H_0 \colon \mu = \mu_0$ gegen $H_1 \colon \mu \neq \mu_0$ testen. Aus Stetigkeitsgründen ist das Supremum der Likelihood-Funktion über $\mu \in \mathbb{R} \setminus \{\mu_0\}$ gleich dem Supremum über ganz $\mathbb{R}$, was am Stichprobenmittel als Maximum-Likelihood-Schätzer $\widehat{\mu} = \overline{X}$ angenommen wird, und die Likelihood-Quotientenstatistik (in den Beobachtungen geschrieben) ist

$$T = \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n \left((X_i - \overline{X})^2 - (X_i - \mu_0)^2\right)\right) = \exp\left(\frac{n}{2\sigma^2} \left(\mu_0 - \overline{X}\right)^2\right).$$

Eine einfache statistische Herleitung der zweiten Identität ergibt sich, wenn man

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2 = (\overline{X} - \mu_0)^2 + \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X})^2$$

als Bias-Varianz-Zerlegung bezüglich der empirischen Verteilung der (X_i) versteht. Wir bemerken, dass T in $|\overline{X} - \mu_0|$ streng monoton wächst und auf Randomisierung verzichtet werden kann, sodass

$$\varphi_\alpha = \mathbb{1}(|\overline{X} - \mu_0| > \sigma n^{-1/2} q_{1-\alpha/2})$$

wegen $\overline{X} \sim \mathrm{N}(\mu_0, \sigma^2/n)$ unter H_0 ein Likelihood-Quotiententest zum Niveau α ist. Man nennt φ_α *zweiseitigen Gauß-Test*.

1.4 Konfidenzmengen

In unserem Saskia-Beispiel 1.7 wird eine Umfrage unter Studierenden durchgeführt, die nur von einem Teil der Studierenden beantwortet wird. Die Antworten aus der Stichprobe ergeben im Mittel einen Wert, von dem wir nicht wissen, ob er der Durchschnittsantwort aller Studierenden (auch jener, die nicht an der Umfrage teilgenommen haben) entspricht. Anhand dieser Stichprobe lässt sich jedoch ein Intervall angeben, das den wahren Mittelwert mit einer gegebenen Wahrscheinlichkeit enthält, und zwar egal, welcher wahre Mittelwert die Wahrscheinlichkeitsverteilung bestimmt.

Statt wie in der Parameterschätzung einen einzelnen Wert zu bestimmen, der möglichst in der Nähe des wahren Parameters liegt, wollen wir also nun Bereiche angeben, von denen wir mit einer gewissen Zuversicht (Konfidenz) sagen können, dass der wahre Parameter in ihnen liegt. Daher werden diese Bereiche auch als „Konfidenzmengen“ oder „Konfidenzbereiche“ bezeichnet. In allen Wissenschaften ist diese Quantifizierung der statistischen Unsicherheit (*uncertainty quantification*) von großer Bedeutung.

Definition 1.66 Sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit abgeleitetem Parameter $\rho\colon \Theta \to \mathbb{R}^d$. Eine mengenwertige Abbildung

$$C\colon \mathcal{X} \to \mathcal{P}(\mathbb{R}^d)$$

heißt **Konfidenzmenge zum Konfidenzniveau** $1-\alpha$ (oder zum Irrtumsniveau α) für $\alpha \in (0,1)$, falls die Messbarkeitsbedingung $\{x \in \mathcal{X} : \rho(\vartheta) \in C(x)\} \in \mathcal{F}$ für alle $\vartheta \in \Theta$ erfüllt ist und

$$\mathbb{P}_\vartheta(\rho(\vartheta) \in C) = \mathbb{P}_\vartheta\big(\{x \in \mathcal{X} : \rho(\vartheta) \in C(x)\}\big) \geqslant 1-\alpha \quad \text{für alle } \vartheta \in \Theta$$

gilt. Im Fall $d = 1$ und falls $C(x)$ für jedes $x \in \mathcal{X}$ ein Intervall ist, heißt C **Konfidenzintervall**.

Bemerkung 1.67 (Konfidenzmengen) Hier ist $\rho(\vartheta)$ fixiert, während C zufällig ist. Konfidenzmengen sind also wie folgt zu interpretieren: Werden in m unabhängigen Experimenten für (verschiedene) Parameter Konfidenzmengen zum Niveau $1-\alpha = 0{,}95$ konstruiert, dann liegt der unbekannte Parameter $\rho(\vartheta)$ in 95% der Fälle in der jeweiligen Konfidenzmenge (für m hinreichend groß; starkes Gesetz der großen Zahlen), unabhängig davon, welches $\vartheta \in \Theta$ vorliegt. Man spricht aber *nicht* davon, dass mit 95%-iger Wahrscheinlichkeit $\rho(\vartheta)$ in $C(x)$ liegt, denn wir betrachten gar kein Wahrscheinlichkeitsmaß auf Θ.

Die Menge $C = \mathbb{R}^d$ ist stets eine (triviale) Konfidenzmenge zu jedem beliebigen Niveau, die uns aber keine Information über $\rho(\vartheta)$ liefert. Je kleiner die Konfidenzmenge ist, desto genauere Aussagen erhalten wir über den unbekannten Parameter. Wir werden uns daher bemühen, Konfidenzmengen möglichst klein zu konstruieren, sodass das Konfidenzniveau gerade noch eingehalten wird.

Wie kann von einer erhobenen Stichprobe ausgehend eine Konfidenzmenge konstruiert werden? Ein häufig verwendetes Konstruktionsprinzip für die Konfidenzintervalle ist die Verwendung eines Schätzers und seiner Verteilung, wie die nächsten Beispiele illustrieren.

Beispiel 1.68 (Konstruktion von Konfidenzintervallen – a) Wir wollen ein Konfidenzintervall für Saskias Umfrage zur Studierendenzufriedenheit konstruieren. In Beispiel 1.7 hatten wir das statistische Modell $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (\mathrm{Ber}(p)^{\otimes n})_{p\in(0,1)})$ betrachtet und als Schätzer $\widehat{\rho}_n$ wählten wir das arithmetische Mittel der Stichprobe. Unsere Grundidee zur Konstruktion des Konfidenzintervalls ist es, um den Schätzer $\widehat{\rho}_n$ ein symmetrisches Intervall

$$C_n := [\widehat{\rho}_n - \varepsilon_n, \widehat{\rho}_n + \varepsilon_n]$$

aufzuspannen, wobei wir ε_n noch näher bestimmen müssen. Damit C_n ein Konfidenzintervall zum Irrtumsniveau $\alpha \in (0,1)$ ist, fordern wir

$$\mathbb{P}_p(p \in C_n) = \mathbb{P}_p(|\widehat{\rho}_n - p| \leqslant \varepsilon_n) = \mathbb{P}_p\Big(\Big|\sum_{i=1}^n (X_i - p)\Big| \leqslant n\varepsilon_n\Big) \overset{!}{\geqslant} 1-\alpha.$$

Wegen $\sum_{i=1}^n X_i \sim \mathrm{Bin}(n,p)$ können wir ε_n numerisch mithilfe der Quantile der Binomialverteilung bestimmen. Alternativ dazu wird im Folgenden eine Normalapproximation verwendet, wobei das resultierende Konfidenzintervall dann nur asymptotisch für große n das Niveau $1-\alpha$ besitzt.

Es gilt für eine Zufallsvariable $Z \sim \mathrm{N}(0,1)$

$$\begin{aligned}
\mathbb{P}_p(p \in C_n) &= \mathbb{P}_p\left(\frac{|\sum_i X_i - np|}{\sqrt{np(1-p)}} \leqslant \frac{n\varepsilon_n}{\sqrt{np(1-p)}}\right) \\
&\approx \mathbb{P}\left(|Z| \leqslant \sqrt{\frac{n}{p(1-p)}}\varepsilon_n\right) \\
&= \Phi\left(\sqrt{\frac{n}{p(1-p)}}\varepsilon_n\right) - \Phi\left(-\sqrt{\frac{n}{p(1-p)}}\varepsilon_n\right) \\
&= 2\Phi\left(\sqrt{\frac{n}{p(1-p)}}\varepsilon_n\right) - 1 \overset{!}{=} 1-\alpha,
\end{aligned}$$

wobei wir die Gleichheit mit $1-\alpha$ fordern, um das Konfidenzniveau voll auszuschöpfen. Mit dem $(1-\alpha/2)$-Quantil $q_{1-\alpha/2}$ erhalten wir also mit gegen $1-\alpha$ konvergierender Wahrscheinlichkeit

$$|\widehat{\rho}_n - p| \leqslant \sqrt{\frac{p(1-p)}{n}} q_{1-\alpha/2}. \tag{1.18}$$

Die rechte Seite ist jedoch noch nicht zur Konstruktion eines Konfidenzintervalls geeignet, da sie vom unbekannten p abhängt. Wir ersetzen p an dieser Stelle daher durch den Schätzer $\widehat{\rho}_n$ und wählen $\varepsilon_n = \sqrt{\frac{\widehat{\rho}_n(1-\widehat{\rho}_n)}{n}} q_{1-\alpha/2}$. Damit ergibt sich das

sogenannte *Wald-Intervall*

$$C_n = \left[\widehat{\rho}_n - \sqrt{\tfrac{\widehat{\rho}_n(1-\widehat{\rho}_n)}{n}}q_{1-\alpha/2}, \widehat{\rho}_n + \sqrt{\tfrac{\widehat{\rho}_n(1-\widehat{\rho}_n)}{n}}q_{1-\alpha/2}\right].$$

In der Tat gilt für jedes $\tau > 0$

$$\begin{aligned}\mathbb{P}_p\big(p \notin C_n\big) &= \mathbb{P}_p\Big(|\widehat{\rho}_n - p| > \sqrt{\tfrac{\widehat{\rho}_n(1-\widehat{\rho}_n)}{n}}q_{1-\alpha/2}\Big)\\ &\leqslant \mathbb{P}_p\Big(|\widehat{\rho}_n - p| > \sqrt{\tfrac{p(1-p)}{n}}q_{1-\alpha/2}(1-\tau)\Big)\\ &\quad + \mathbb{P}_p\Big(\sqrt{p(1-p)} - \sqrt{\widehat{\rho}_n(1-\widehat{\rho}_n)} > \tau\sqrt{p(1-p)}\Big).\end{aligned}$$

Für $n \to \infty$ konvergiert der erste Term wie in obiger Rechnung gegen $2 - 2\Phi(q_{1-\alpha/2}(1-\tau))$, während der zweite Term aufgrund der Konsistenz $\widehat{\rho}_n \xrightarrow{\mathbb{P}_p} p$ (Gesetz der großen Zahlen) für jedes τ gegen 0 konvergiert. Aus $\lim_{\tau\to 0}\Phi(q_{1-\alpha/2}(1-\tau)) = 1-\alpha/2$ ergibt sich damit $\lim_{n\to\infty}\mathbb{P}_p\big(p \notin C_n\big) = \alpha$, sodass C_n tatsächlich ein Konfidenzintervall zum asymptotischen Niveau $1-\alpha$ ist.

Alternativ kann man auch (1.18) durch Quadrieren und Auflösen der quadratischen Gleichung direkt nach p umstellen. Das vermeidet den zusätzlichen Approximationsschritt und führt auf das *Wilson-Intervall* (Aufgabe 1.13).

Beispiel 1.69 (Konstruktion von Konfidenzintervallen – b) Selbst wenn die Verteilung des Schätzers $\widehat{\rho}_n$ nicht explizit bekannt ist, können Konfidenzintervalle konstruiert werden, sofern sich $\widehat{\rho}_n$ um den wahren Wert $\rho(\vartheta)$ konzentriert. In diesem Fall verwenden wir, dass die Wahrscheinlichkeit $\mathbb{P}_\vartheta(|\widehat{\rho}_n - \rho(\vartheta)| > \varepsilon)$ klein wird, wenn ε groß wird. Wir wollen diese Idee im Modell aus Beispiel 1.68 verdeutlichen. Für $C_n := (\widehat{\rho}_n - \varepsilon_n, \widehat{\rho}_n + \varepsilon_n)$ gilt wegen der Tschebyscheff-Ungleichung

$$\begin{aligned}\mathbb{P}_p(p \in C_n) &= \mathbb{P}_p(|\widehat{\rho}_n - p| < \varepsilon_n)\\ &= \mathbb{P}_p\Big(\Big|\sum_{i=1}^n X_i - np\Big| < n\varepsilon_n\Big)\\ &\geqslant 1 - \frac{\operatorname{Var}(\sum_i X_i)}{n^2\varepsilon_n^2}\\ &= 1 - \frac{np(1-p)}{n^2\varepsilon_n^2} \overset{!}{\geqslant} 1-\alpha.\end{aligned}$$

Wir schätzen $p(1-p) \leqslant 1/4$ ab und erhalten $\varepsilon_n = 1/\sqrt{4n\alpha}$. Damit erhalten wir das (nicht asymptotische) Konfidenzintervall $C_n = \Big(\widehat{\rho}_n - \frac{1}{2\sqrt{n\alpha}}, \widehat{\rho}_n + \frac{1}{2\sqrt{n\alpha}}\Big)$. Allerdings ist die Abschätzung mit der Tschebyscheff-Ungleichung sehr grob und führt daher zu einem sehr vorsichtigen bzw. konservativen Konfidenzintervall.

Eine alternative Konstruktion von Konfidenzmengen bietet folgender Korrespondenzsatz:

Satz 1.70 (Korrespondenzsatz) *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\alpha \in (0,1)$. Dann gilt:*

(i) Liegt für jedes $\vartheta_0 \in \Theta$ ein nicht-randomisierter Test φ_{ϑ_0} der Hypothese $H_0\colon \vartheta = \vartheta_0$ zum Signifikanzniveau α vor, so definiert $C(x) := \{\vartheta_0 \in \Theta : \varphi_{\vartheta_0}(x) = 0\}$ für $x \in \mathcal{X}$ eine Konfidenzmenge für ϑ zum Konfidenzniveau $1-\alpha$.

(ii) Ist C eine Konfidenzmenge für ϑ zum Niveau $1-\alpha$, dann ist $\varphi_{\vartheta_0}(x) := 1 - \mathbb{1}_{C(x)}(\vartheta_0)$ ein Niveau-α-Test der Hypothese $H_0\colon \vartheta = \vartheta_0$.

Es gibt also eine Eins-zu-eins-Beziehung zwischen Hypothesentests und Konfidenzmengen. Die Konfidenzmenge in (i) enthält all jene ϑ_0, für die der Test φ_{ϑ_0} die Nullhypothese aufgrund der beobachteten Stichprobe x nicht verwirft.

Beweis Nach Konstruktion erhält man in beiden Fällen

$$\forall \vartheta \in \Theta : \forall x \in \mathcal{X} : \varphi_\vartheta(x) = 0 \quad \Longleftrightarrow \quad \vartheta \in C(x).$$

Damit ist φ_ϑ genau dann ein Test zum Niveau α für alle ϑ, wenn

$$1 - \alpha \leqslant \mathbb{P}_\vartheta(\varphi = 0) = \mathbb{P}_\vartheta(\{x : \vartheta \in C(x)\}),$$

und somit ist C eine Konfidenzmenge zum Niveau α. □

Bezeichnen wir den Annahmebereich der Tests φ_ϑ mit $A(\vartheta)$, liefert uns der Korrespondenzsatz folgende Methode:

Methode 1.71 (Konstruktion von Konfidenzmengen) Sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\alpha \in (0,1)$. Wähle zu jedem $\vartheta \in \Theta$ ein $A(\vartheta) \in \mathcal{F}$ mit $\mathbb{P}_\vartheta(A(\vartheta)) \geqslant 1-\alpha$ und setze $C(x) := \{\vartheta \in \Theta : x \in A(\vartheta)\}$ für $x \in \mathcal{X}$. Dann gilt

$$x \in A(\vartheta) \quad \Leftrightarrow \quad \vartheta \in C(x),$$

woraus mit obigem Satz folgt

$$\mathbb{P}_\vartheta(\{x \in \mathcal{X} : \vartheta \in C(x)\}) \geqslant 1-\alpha \quad \forall \vartheta \in \Theta.$$

Man könnte zunächst meinen, dass durch diese Methode die Schwierigkeit der Konstruktion lediglich von $C(x)$ auf $A(\vartheta)$ verschoben wurde, aber dadurch haben wir einen Vorteil erlangt: $A(\vartheta)$ ist eine Teilmenge von $\mathcal{X}$, die wir mit $\mathbb{P}_\vartheta$ messen können, und aus der Kenntnis von $\mathbb{P}_\vartheta$ ergibt sich meist eine einfache Wahl eines Ereignisses $A(\vartheta)$ mit Wahrscheinlichkeit $1-\alpha$.

Beispiel 1.72 (Konstruktion von Konfidenzmengen) Wir wollen ein Konfidenzintervall für die Geburtswahrscheinlichkeit von Mädchen in Hamburg berechnen. In Beispiel 1.48 hatten wir die Anzahl der Mädchen unter n Geburten mit $X \sim \mathrm{Bin}(n, \vartheta)$ modelliert, wobei $\vartheta \in \Theta := (0,1)$ und der Stichprobenraum $\mathcal{X} = \{0, 1, ..., n\}$ ist. Wir wählen ein Niveau $\alpha \in (0,1)$. Die Annahmebereiche der in Beispiel 1.48 konstruierten Tests sind gegeben durch

$$A(\vartheta) := \{x \in \mathcal{X} : u_\alpha(\vartheta) \leqslant x \leqslant o_\alpha(\vartheta)\}$$

mit $u_\alpha(\vartheta)$ und $o_\alpha(\vartheta)$ aus (1.11). Damit schneiden wir von der gesamten möglichen Wahrscheinlichkeitsmasse an beiden Enden $\frac{\alpha}{2}$ ab, sodass in der Mitte $1-\alpha$ übrig bleibt (siehe Abbildung 1.2). Da wir $A(\vartheta)$ gewählt haben, nutzen wir nun die Beziehung

$$x \in A(\vartheta) \quad \Leftrightarrow \quad \vartheta \in C(x) \quad \text{für alle } x \in \mathcal{X}, \vartheta \in \Theta,$$

um die Konfidenzmenge C zu bestimmen. Die Leserin überprüfe selbst, dass $u_\alpha\colon [0,1] \to \{0,1,...,n\}$ monoton fallend und rechtsseitig stetig ist und $o_\alpha\colon [0,1] \to \{0,1,...,n\}$ monoton wachsend und linksseitig stetig ist. Folglich ist

$$C(x) := \left[\inf\{\vartheta \in \Theta : o_\alpha(\vartheta) = x\}, \sup\{\vartheta \in \Theta : u_\alpha(\vartheta) = x\}\right]$$

ein Konfidenzintervall zum Konfidenzniveau $1-\alpha$. Es wird *Clopper-Pearson-Intervall* genannt. Die Berechnung des Infimums und des Supremums ist eine numerische Aufgabe.

Für $n = 21.126$ Geburten im Jahr 2018, von denen 10.215 weiblich waren, erhalten wir folgendes Konfidenzintervall zum Niveau $1-\alpha = 0{,}95$ für die Wahrscheinlichkeit, dass ein Mädchen geboren wurde:

$$C = [0{,}4768;\ 0{,}4903]$$

Bemerkung 1.73 (Einseitige Konfidenzbereiche) Bisher haben wir nur zweiseitige Konfidenzintervalle gesehen. Eine andere Variante sind *einseitige Konfidenzintervalle*. Das heißt, dass nur eine Seite von den Beobachtungen abhängt und die andere fest ist. Im vorherigen Beispiel könnte man analog das Konfidenzintervall

$$\widetilde{C}(x) := \left[0, \sup\{\vartheta \in \Theta : \widetilde{u}_\alpha(\vartheta) = x\}\right), \quad x \in \mathcal{X},$$

konstruieren mit

$$\widetilde{u}_\alpha(\vartheta) := \max\{k \in \mathcal{X} : \mathbb{P}_\vartheta(X < k) \leqslant \alpha\}.$$

Nach unten verliert diese Konstruktion des Konfidenzintervalls zwar an Aussagekraft, aber nach oben gewinnt es ebenjene, da die Obergrenze schärfer wird.

1.5 Aufgaben

1.1 Wir betrachten eine auf dem Intervall $[a,b]$ gleichverteilte mathematische Stichprobe $X_1,\dots,X_n$ für $n \in \mathbb{N}$ und unbekannte Parameter $-\infty < a < b < \infty$.

(a) Formalisieren Sie das statistische Modell.
(b) Bestimmen Sie Momentenschätzer für a und b.
(c) Bestimmen Sie die Maximum-Likelihood-Schätzer für a und b.
(d) Welches quadratische Risiko hat der Maximum-Likelihood-Schätzer?

1.2 Um die Gesamtanzahl N der insgesamt in Berlin registrierten Taxis zu schätzen, notiert sich ein Tourist die Konzessionsnummern von $n < N$ vorbeifahrenden Taxis (Wiederholungen möglich). Er nimmt an, dass alle Taxis von 1 bis N durchnummeriert sind und die beobachteten Taxinummern unabhängig voneinander und mit gleicher Wahrscheinlichkeit vorbeifahren.

(a) Formalisieren Sie das statistische Modell.
(b) Berechnen Sie aus den notierten Nummern $X_1, \dots, X_n$ einen Maximum-Likelihood-Schätzer $\widehat{N}$ für N. Ist dieser erwartungstreu?
(c) Berechnen Sie approximativ für großes N den relativen Erwartungswert $\mathbb{E}[\widehat{N}]/N$.
Hinweis: Fassen Sie einen geeigneten Ausdruck als Riemann-Summe auf.

1.3 Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein dominiertes statistisches Modell mit zwei dominierenden Maßen μ_1 und μ_2. Die zugehörigen Likelihoodfunktionen seien mit $L_1(\vartheta, x) = \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu_1}(x)$ bzw. $L_2(\vartheta, x) = \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu_2}(x), x \in \mathcal{X}, \vartheta \in \Theta$. Zeigen Sie, dass jeder Maximum-Likelihood-Schätzer unabhängig vom dominierenden Maß ist, d.h. jede Maximimalstelle von L_1 ist auch eine Maximalstelle von L_2.

1.4 Sei $X_1, \dots, X_n$ eine mathematische Stichprobe reeller Zufallsvariablen mit der Verteilung $\mathbb{P}_\vartheta, \vartheta \in \Theta$. Für $k > 0$ betrachten wir die Funktion

$$\psi_k \colon \mathbb{R} \to \mathbb{R}, \quad \psi_k(x) = \begin{cases} -k, & x < -k, \\ x, & |x| \leqslant k, \\ k, & x > k. \end{cases}$$

(a) Beweisen Sie, dass ein $h_k \in \mathbb{R}$ existiert, sodass $\mathbb{E}[\psi_k(X_1 - h_k)] = 0$. Der *Huber-Schätzer* $\widehat{h}_k$ von h_k ist definiert als Nullstelle der Funktion

$$\mathbb{R} \ni h \mapsto \sum_{i=1}^{n} \psi_k(X_i - h).$$

Weisen Sie nach, dass auch $\widehat{h}_k$ stets existiert.
(b) Zeigen Sie, dass h_k für $k \to 0$ gegen einen Median von $\mathbb{P}_\vartheta$ konvergiert. Beweisen Sie im Fall $\mathbb{E}[|X_1|] < \infty$, dass h_k für $k \to \infty$ gegen den Mittelwert von $\mathbb{P}_\vartheta$ konvergiert.
(c) Sei nun $\mathbb{P}_\vartheta$ die Cauchy-Verteilung mit Parameter $\vartheta = (x_0, \gamma) \in \mathbb{R} \times \mathbb{R}_+$ und Lebesgue-Dichte

$$f_{x_0,\gamma}(x) = \frac{1}{\pi} \frac{\gamma}{(x - x_0)^2 + \gamma^2}, \quad x \in \mathbb{R}.$$

Bestimmen Sie h_k in Abhängigkeit von $\vartheta = (x_0, \gamma)$ für alle $k > 0$.
(d) Simulieren Sie $n = 200$ unabhängige Zufallsvariablen mit den Randverteilungen

(i) Lognormalverteilung $\log \mathrm{N}(\mu, \sigma)$ mit Parametern $\mu = 1, \sigma^2 = 1$ sowie $\mu = 1, \sigma^2 = 3$.
(ii) Cauchy-Verteilung mit Parametern $x_0 = 0, \gamma = 1$ und $x_0 = 0, \gamma = 5$.

Bestimmen Sie in allen vier Szenarien den theoretischen Median und den Mittelwert. Berechnen Sie außerdem $\widehat{h}_k$ für $k \in \{\frac{1}{10}, \frac{1}{2}, 1, 2, 5, 10, 20, 50\}$. Stellen Sie Ihre Ergebnisse graphisch dar.

1.5 Sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ mit $\Theta = \mathbb{R}$ ein von μ dominiertes statistisches Modell mit Dichten $f_{X|T=\vartheta} := \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}$ und sei π eine a-priori-Verteilung auf $(\Theta, \mathcal{F}_\Theta)$ mit Dichte f_T bezüglich eines Maßes ν. Nehmen Sie an, dass $f_{X|T=\cdot} : \mathcal{X} \times \Theta \to \mathbb{R}_+$ $(\mathcal{F} \otimes \mathcal{F}_\Theta)$-messbar ist. Zeigen Sie:

(a) Unter quadratischem Verlust ist der Bayes-Schätzer gegeben durch den a-posteriori-Erwartungswert.
(b) Unter absolutem Verlust ist der Bayes-Schätzer gegeben durch den a-posteriori-Median.
(c) Unter 0-1-Verlust ist der Bayes-Schätzer gegeben durch den a-posteriori-Modus (das heißt der Maximalstelle der a-posteriori-Verteilung).

1.6 Es sei $X \sim \mathrm{Bin}(n, p)$ binomialverteilt mit $n \in \mathbb{N}$ und $p \in [0, 1]$, und sei die a-priori-Verteilung π für p auf $[0, 1]$ gegeben durch eine Beta-Verteilung $\mathrm{Beta}(\alpha, \beta)$ mit Parametern $\alpha, \beta > 0$.

(a) Zeigen Sie, dass die Gleichverteilung $\mathrm{U}([0, 1])$ ein Spezialfall der Beta-Verteilung ist.
(b) Beweisen Sie, dass die Beta-Verteilungen zur Binomialverteilung konjugiert sind, das heißt die a-posteriori-Verteilung ist wieder Beta-verteilt. Bestimmen Sie die Parameter der a-posteriori-Beta-Verteilung.
(c) Folgern Sie, dass der Bayes-Schätzer unter quadratischem Verlust gegeben ist durch

$$\widehat{p}_{a,b} = \frac{a + X}{a + b + n}.$$

Bestimmen Sie dessen mittleres quadratisches Risiko $\mathbb{E}_p[|\widehat{\beta}_{a,b} - p|^2]$ in Abhängigkeit von a, b und p.
(d) Wählen Sie $a^*, b^* > 0$ so, dass $\max_{p \in [0,1]} \mathbb{E}_p[|\widehat{\beta}_{a,b} - p|^2]$ minimal ist. Folgern Sie, dass $\widehat{p}_{a^*,b^*}$ minimax-Schätzer ist und der Maximum-Likelihood-Schätzer $\widehat{p} = X/n$ nicht minimax ist.

1.7 Die Vertriebsleiterin einer Getreidemühle geht davon aus, dass ihre angebotenen Mehlpackungen ein mittleres Füllgewicht von 1000 g mit einer Standardabweichung von 12,5 g haben. Sie möchte die Füllmenge mit einem statistischen Test überprüfen. Eine genauere Untersuchung an $n = 40$ Packungen zeigt, dass die Füllmenge im Durchschnitt 1004,32 g beträgt.

(a) Formalisieren Sie das statistische Modell und das Testproblem unter einer Normalverteilungsannahme an die Packungsgewichte.

(b) Konstruieren Sie einen Test φ zum Niveau $\alpha = 0{,}05$ unter Verwendung des Durchschnittsgewichts als Teststatistik. Verwirft dieser Test die Hypothese?
(c) Stellen Sie die Gütefunktion von φ mithilfe der Verteilungsfunktion der Standardnormalverteilung $\Phi\colon \mathbb{R} \to [0,1]$ dar.

1.8 Am Hamburger U-Bahnhof Schlump treffen $n \in \mathbb{N}$ Studierende ein und müssen jeweils X_i Minuten auf die Bahn warten, $i = 1, \dots, n$. Bezeichnet $\vartheta > 0$ die erwartete Wartezeit, soll

$$H_0\colon \vartheta \leqslant \vartheta_0 \quad \text{gegen} \quad H_1\colon \vartheta > \vartheta_0$$

für ein $\vartheta_0 > 0$ getestet werden. Hierzu wird die Teststatistik $\varphi_c(X_1, \dots, X_n) = \mathbb{1}_{(c,\infty)}(\overline{X}_n)$ mit kritischem Wert $c > 0$ und $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ verwendet.

(a) Begründen Sie, warum die Exponentialverteilungsannahme $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathrm{Exp}(\lambda)$ mit Parameter $\lambda > 0$ sinnvoll ist, und formulieren Sie das Testproblem in Abhängigkeit von λ.
(b) Bestimmen Sie die Verteilung von $\overline{X}_n$. *Hinweis:* $\mathrm{Exp}(\lambda) = \Gamma(1, \lambda)$.
(c) Berechnen Sie $c > 0$ so, dass φ_c ein Test zum Niveau $\alpha \in (0,1)$ ist.
(d) Notieren Sie sich die Zeit, die Sie heute auf dem Heimweg auf die U-Bahn warten mussten.

1.9 Betrachten Sie ein statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ und einen Test der Hypothese $H_0\colon \vartheta \in \Theta_0 \neq \emptyset$ zum Niveau $\alpha \in (0,1)$ der Form $\varphi = \mathbb{1}_{\{T > c_\alpha\}}$ mit Teststatistik T und kritischen Werten $c_\alpha, \alpha \in (0,1)$. Falls $\sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T > c_\alpha) \leqslant \alpha$ für alle $\alpha \in (0,1)$, gilt für den p-Wert

$$p_\varphi(x) = \inf\big\{\alpha \in [0,1] : T(x) > c_\alpha\big\}.$$

1.10 Zeigen Sie, dass im statistischen Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\{0,1\}})$ zu jedem $\alpha \in (0,1)$ ein Neyman-Pearson-Test zum Niveau α existiert.

1.11 Die mathematische Stichprobe $X_1, \dots, X_n$ sei gemäß $\mathrm{N}(\mu, \sigma^2)$ verteilt mit unbekanntem $\mu \in \mathbb{R}$ und unbekanntem $\sigma > 0$.

(a) Bestimmen Sie den Maximum-Likelihood-Schätzer $\widehat{\sigma}^2$ für σ^2.
(b) Konstruieren Sie einen Likelihood-Quotiententest für das Testproblem

$$H_0\colon \sigma^2 = \sigma_0^2 \quad \text{gegen} \quad H_1\colon \sigma^2 \neq \sigma_0^2$$

für ein $\sigma_0 > 0$. Zeigen Sie, dass dieser Test für geeignete $a, b > 0$ von der Form $T = 1 - \mathbb{1}(a \leqslant \frac{\widehat{\sigma}^2}{\sigma_0^2} \leqslant b)$ ist.
(c) Verwenden Sie, dass $\frac{n\widehat{\sigma}^2}{\sigma_0}$ unter der Hypothese χ^2-verteilt ist, um a und b so zu wählen, dass T ein Test zum Niveau $\alpha \in (0,1)$ ist.

1.12 Ein Experimentator macht $n \in \mathbb{N}$ unabhängige normalverteilte Messungen mit unbekanntem Erwartungswert $\mu \in \mathbb{R}$. Die Varianz $\sigma^2 > 0$ meint er zu kennen.

(a) Welches Konfidenzintervall C_{α,σ^2} für μ wird er zu einem vorgegeben Irrtumsniveau $\alpha \in (0,1)$ angeben?
(b) Welches Irrtumsniveau hat dieses Konfidenzintervall C_{α,σ^2}, wenn die Varianz in Wirklichkeit den Wert $\widetilde{\sigma}^2 > 0$ annimmt?
(c) Simulieren Sie in 500 Durchgängen jeweils 20 unabhängige $\mathrm{N}(0,1)$-verteilte Zufallsvariablen $X_1, \ldots, X_{20}$ bzw. $\mathrm{N}(0,2)$-verteilte Zufallsvariablen $Y_1, \ldots, Y_{20}$. Bestimmen Sie in jeder Iteration $C_{\alpha,1}$ unter Verwendung von (X_i) bzw. (Y_i). Für wie viele Realisierungen von (X_i) bzw. (Y_i) liegt $\mu = 0$ in $C_{\alpha,1}$ für $\alpha \in \{\frac{1}{100}, \frac{1}{20}, \frac{1}{10}, \frac{1}{4}, \frac{1}{2}\}$. Bestimmen Sie jeweils das Verhältnis zur Anzahl der Simulationsdurchgänge 500.

1.13 Es seien $X_1, \ldots, X_n$ unabhängige $\mathrm{Ber}(p)$ verteilte Zufallsvariablen mit unbekanntem Erfolgsparameter p und dem Schätzer $\widehat{\rho}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Weisen Sie nach, dass für $\alpha \in (0,1)$ das *Wilson-Intervall* $[u_\alpha, o_\alpha]$ mit

$$u_\alpha = \frac{1}{1+\frac{q_{1-\alpha/2}^2}{n}}\Big(\widehat{\rho}_n + \frac{q_{1-\alpha/2}^2}{2n} - \frac{q_{1-\alpha/2}}{\sqrt{n}}\sqrt{\widehat{p}_n(1-\widehat{p}_n) + \frac{q_{1-\alpha/2}^2}{4n}}\Big),$$

$$o_\alpha = \frac{1}{1+\frac{q_{1-\alpha/2}^2}{n}}\Big(\widehat{p}_n + \frac{q_{1-\alpha/2}^2}{2n} + \frac{q_{1-\alpha/2}}{\sqrt{n}}\sqrt{\widehat{p}_n(1-\widehat{p}_n) + \frac{q_{1-\alpha/2}^2}{4n}}\Big).$$

ein Konfidenzintervall zum asymptotischen Konfidenzniveau $1-\alpha$ ist.

Kapitel 2
Das lineare Modell

Mit der einfachen linearen Regression beginnend, werden wir in diesem Kapitel das lineare Modell im Detail studieren. Dieses schließt die multiple sowie die polynomiale Regression ein. Insbesondere befassen wir uns mit der verallgemeinerten Methode der kleinsten Quadrate. Unter einer Normalverteilungsannahme an die Beobachtungsfehler konstruieren wir anschließend Hypothesentests und Konfidenzmengen basierend auf t- und F-Statistiken. Als Spezialfall ergibt sich die Varianzanalyse.

2.1 Regression und kleinste Quadrate

Nachdem wir uns bisher mit einigen Grundbegriffen und Fragestellungen in der Statistik befasst haben, wollen wir nun das lineare Modell studieren. Die beobachteten Daten werden in dieser Modellklasse durch bekannte Einflussvariablen erklärt, wobei eine lineare Abhängigkeit von den unbekannten Parametern angenommen wird. Lineare Modelle finden unzählige Anwendungen und sind mathematisch relativ leicht zu analysieren. Insbesondere die Methode der kleinsten Quadrate, auf die wir noch genauer eingehen werden, hat seit über 200 Jahren nicht an Bedeutung verloren.

Regression bezeichnet die statistische Analyse des (nicht unbedingt linearen) Zusammenhangs zwischen einer *Zielgröße* Y (auch *Regressand*, *Response-Variable* oder *abhängige Variable* genannt) und einem Vektor von *Kovariablen* $X = (X_1, ..., X_n)$ (oder auch *Regressoren*, *erklärenden Variablen*, *unabhängigen Variablen*). Die Kovariablen können zufällig oder deterministisch sein. Wir sprechen in diesen Fällen von *zufälligem Design* bzw. *deterministischem Design*. Im Folgenden betrachten wir zunächst den letzteren Fall und schreiben der Klarheit halber $X = (x_1, ..., x_n)$. Sind die Kovariablen zufällig, aber von den Beobachtungsfehlern unabhängig, so können wir auf X bedingen und die Resultate von deterministischem Design auf zufälliges Design übertragen.

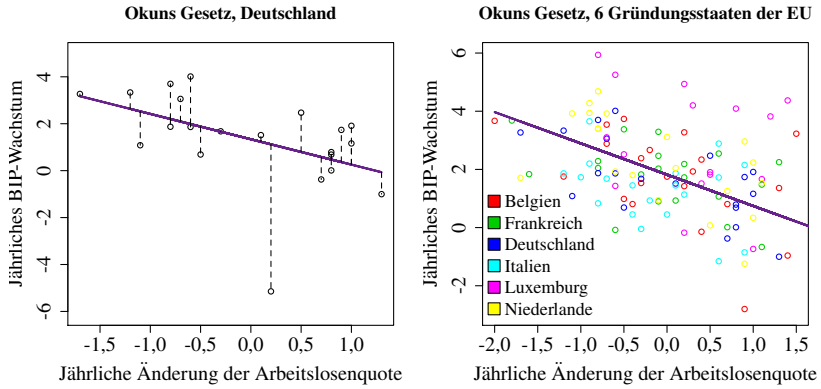

Abb. 2.1 Jährliche prozentuale Veränderung der Arbeitlosenquote und jährliches Wachstum des Bruttoinlandprodukts zwischen 1992 und 2012 für Deutschland (*links*) beziehungsweise für die 6 Gründungsstaaten der EU (*rechts*) sowie jeweilige Regressionsgrade

2.1.1 Lineare Regression

Wir beginnen mit der einfachen linearen Regression:

Definition 2.1 Im Modell der **einfachen linearen Regression** werden

$$Y_i = ax_i + b + \varepsilon_i, \quad i \in \{1, ..., n\},$$

für gegebene Kovariablen $x_1, \ldots, x_n \in \mathbb{R}$ beobachtet. Hierbei sind die Beobachtungsfehler $\varepsilon_1, ..., \varepsilon_n$ zentrierte und unkorrelierte Zufallsvariablen ($\mathbb{E}[\varepsilon_i] = 0$) mit endlicher Varianz $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$. Die Parameter $a, b \in \mathbb{R}$ sind unbekannt und bestimmen die **Regressionsgerade** $y = ax + b$.

Aufgrund der Beobachtungsfehler ε_i ist der Zusammenhang zwischen den Kovariablen x_i und den Regressanden Y_i nicht deterministisch. Stattdessen beobachten wir eine zufällige Punktewolke um die Regressiongerade, die den linearen Zusammenhang möglichst gut beschreibt. Das Ziel ist die Schätzung der Parameter a und b. Der Parameter σ ist typischerweise nicht das Ziel der statistischen Inferenz und somit ein *Störparameter*.

Beispiel 2.2 (Einfache lineare Regression, Okuns Gesetz) Y_i bezeichnet das Wachstum des Bruttoinlandsprodukts von Deutschland im Jahr i. Die Kovariable x_i ist die Veränderung der Arbeitslosenquote im Vergleich zum Vorjahr. Unter Verwendung der Daten von 1992 bis 2012 aus den *World Development Indicators* der Weltbank erhalten wir als Regressionsgrade $y = -1{,}080x + 1{,}338$. Betrachten wir alle sechs Gründungsmitglieder der EU im gleichen Zeitraum, ergibt sich ganz ähnlich $y = -1{,}075x + 1{,}819$, siehe Abbildung 2.1. Der lineare Zusammenhang beider Größen ist als *Okuns Gesetz* bekannt.

Um die den Daten zugrunde liegenden Parameter a, b und damit die Regressionsgerade zu schätzen, werden wir die uns schon bekannte Maximum-Likelihood-Methode anwenden. Dafür nehmen wir an, dass $\varepsilon_1, ..., \varepsilon_n$ unabhängig und $\mathrm{N}(0, \sigma^2)$-verteilt sind. Weil $ax_i + b$ für alle i deterministisch in $\mathbb{R}$ ist, gilt für die Beobachtungen in der einfachen linearen Regression

$$Y_i = ax_i + b + \varepsilon_i \sim \mathrm{N}(ax_i + b, \sigma^2).$$

Das statistische Modell ist somit durch

$$\Big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \Big(\bigotimes_{i=1}^{n} \mathrm{N}(ax_i + b, \sigma^2)\Big)_{a,b\in\mathbb{R},\,\sigma>0}\Big)$$

gegeben. Es ergibt sich für $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ die Likelihood-Funktion

$$\begin{aligned} L(a, b, \sigma; y) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\Big(-\frac{(y_i - ax_i - b)^2}{2\sigma^2}\Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - ax_i - b)^2\Big). \end{aligned}$$

Die Terme $y_i - ax_i - b$ nennt man auch *Residuen*. Das Maximieren der Likelihood über a, b ist also äquivalent zum Minimieren der Summe der quadrierten Residuen (englisch: *residual sum of squares*, kurz: RSS). Auch wenn die Fehler nicht normalverteilt sind, kann diese Methode gute Ergebnisse erzielen.

Methode 2.3 (Kleinste Quadrate) In der einfachen linearen Regression sind die Kleinste-Quadrate-Schätzer (englisch: *least squares estimator*, kurz: LSE) $\widehat{a}, \widehat{b}$ durch Minimierung der Summe der quadrierten Residuen gegeben:

$$(\widehat{a}, \widehat{b}) := \operatorname*{arg\,min}_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (Y_i - ax_i - b)^2$$

Da die Kleinste-Quadrate-Schätzer nicht von σ^2 abhängen, sind sie auch im Fall von unbekannter Fehlervarianz anwendbar. Zudem kann die Lösung dieses Minimierungsproblems explizit berechnet werden.

Lemma 2.4 *In der einfachen linearen Regression mit unabhängigen und* $\mathrm{N}(0, \sigma^2)$*-verteilten Fehlern ist der Maximum-Likelihood-Schätzer gleich dem Kleinste-Quadrate-Schätzer. Zudem gilt*

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2} \quad \textit{und} \quad \widehat{b} = \overline{Y}_n - \widehat{a}\,\overline{x}_n$$

mit $\overline{Y}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i$ *und* $\overline{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$*, falls es* $i, j \in \{1, \ldots, n\}$ *gibt mit* $x_i \neq x_j$*.*

Beweis Es bleibt festzustellen, dass wir durch Differenzieren in a und b folgende Normalengleichungen erhalten:

$$0 = \sum_{i=1}^{n} x_i(Y_i - ax_i - b) \quad \text{und} \quad 0 = \sum_{i=1}^{n} (Y_i - ax_i - b)$$

Man prüft leicht nach, dass diese Gleichungen durch $\widehat{a}$ und $\widehat{b}$ gelöst werden, sofern die Stichprobenvarianz der (x_i) nicht null ist, das heißt falls $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 > 0$ gilt. Dies ist genau dann der Fall, wenn es x_i, x_j mit $x_i \neq x_j$ gibt. Offensichtlich liegt bei $(\widehat{a}, \widehat{b})$ ein Minimum des streng konvexen Kleinste-Quadrate-Kriteriums vor. □

Der Maximum-Likelihood-Ansatz liefert auch unter anderen Verteilungsannahmen an die (ε_i) sinnvolle Schätzer, siehe Aufgabe 2.1. Die Kleinste-Quadrate-Methode ist jedoch aufgrund ihrer guten Eigenschaften mit Abstand die populärste.

Kurzbiografie (Carl Friedrich Gauß) Carl Friedrich Gauß wurde 1777 in Braunschweig geboren. Früh galt er als Wunderknabe, was ihm die Gönnerschaft des Herzogs von Braunschweig einbrachte. Er unterstützte ihn vor allem finanziell, sodass er ab 1795 in Göttingen studieren konnte. Gauß beschäftigte sich mit Philologie, Mathematik, Physik und insbesondere mit Astronomie. An der Universität Helmstedt reichte er 1799 seine Doktorarbeit ein. Danach arbeitete er intensiv an seinem einflussreichen Lehrbuch zu höherer Arithmetik, der *Disquisitiones Arithmeticae*. 1801 gelang es ihm, den Zwergplaneten Ceres mithilfe seiner *Methode der kleinsten Quadrate* wiederaufzufinden. Wenig später wurde er Universitätsprofessor und Direktor der Sternwarte in Göttingen. Zahlreiche Methoden und Ideen sind nach ihm benannt, unter anderem das gaußsche Eliminationsverfahren zur Diagonalisierung und Invertierung von Matrizen, die Gauß-Verteilung (trotz reichlicher Vorarbeit von de Moivre, Laplace und Poisson) und die Methode der kleinsten Quadrate. 1855 starb Gauß in Göttingen.

Auch wenn kein linearer Zusammenhang vorliegt, kann die lineare Regression zur Untersuchung aller möglichen Zusammenhänge herangezogen werden: Der Abstand zur nächsten Autobahn (Regressor) als Einfluss auf die Anzahl der nächtlich geschlafenen Stunden (Regressand), der Betrag des Geldes auf dem Konto (Regressor) als Einfluss auf das Wohlbefinden (Regressand), die Anzahl der Sonnenstunden (Regressor) als Einfluss auf den Vitamin-D-Gehalt im Körper (Regressand) etc. Dabei muss man jedoch beachten, dass die Kalbrierung eines linearen Modells anhand von Daten, die keinem linearen Zusammenhang folgen, zu einem möglicherweise großen Modellfehler führen und gegebenfalls falsche Schlussfolgerungen suggerieren, siehe Beispiel 2.9 sowie Aufgabe 2.3.

Bei der einfachen linearen Regression wird der Regressand durch eine Kovariable erklärt. Allgemeiner kann man den Einfluss mehrerer Kovariablen untersuchen.

Definition 2.5 Bei $k \geqslant 2$ Kovariablen $x_j = (x_{1,j}, \ldots, x_{n,j})^\top \in \mathbb{R}^n$, $j = 1, \ldots, k$, und n Beobachtungen Y_i erhalten wir das **multiple lineare Regressionsmodell**

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \varepsilon_i, \quad i = 1, \ldots, n,$$

wobei die Fehlerterme $(\varepsilon_i)_{i=1,\ldots,n}$ zentriert und unkorreliert sind mit $0 < \mathrm{Var}(\varepsilon_i) =: \sigma^2 < \infty$. In Vektorschreibweise erhalten wir die lineare Gleichung

$$Y = X\beta + \varepsilon$$

mit

$$\begin{aligned}
\textbf{Responsevektor} \qquad & Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n, \\
\textbf{Designmatrix} \qquad & X := \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix} \in \mathbb{R}^{n\times(k+1)}, \\
\textbf{Fehlervektor} \qquad & \varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n, \\
\textbf{Parametervektor} \qquad & \beta := (\beta_0, \ldots, \beta_k)^\top \in \mathbb{R}^{k+1}.
\end{aligned}$$

Bemerkung 2.6 Wechselwirkungen zwischen zwei Kovariablen x_i und x_j werden in der Praxis oft durch Interaktionsterme $x_{i,k} \cdot x_{i,j}$ in der Designmatrix modelliert.

Kategorielle Kovariablen können durch eine Menge von sogenannten *Dummy-Variablen*, das heisst $\{0, 1\}$-wertigen Variablen, kodiert werden, um nicht implizit eine (inadäquate) Metrisierung auf dem diskreten Wertebereich solcher Kovariablen zu erzeugen. Eine kategorielle Kovariable mit ℓ möglichen Ausprägungen wird dabei durch $(\ell - 1)$ viele $\{0, 1\}$-wertige Variablen repräsentiert. Die j-te Dummy-Variable kodiert das Ereignis, dass die Kategorie $(j + 1)$ bei der zugehörigen Kovariablen vorliegt. Sind alle $(\ell - 1)$ Indikatoren gleich null, so entspricht dies der (Referenz-) Kategorie 1 der zugehörigen kategoriellen Kovariablen. Wir werden dieses Vorgehen insbesondere bei der Varianzanalyse in Kapitel 2.3 verwenden, wo die *Faktoren* kategorielle Kovariablen sind.

Die Vektorschreibweise führt uns auf die allgemeine Form des linearen Modells. Dabei muss insbesondere die erste Spalte der Designmatrix nicht nur aus Einsen bestehen. Zudem werden Korrelationen zwischen den Fehlertermen ε_i zugelassen. Zunächst betrachten wir den klassischen Fall, in dem die Parameterdimension p kleiner als die Stichprobengröße n ist. Für den Fall $p \geqslant n$ werden wir in Kapitel **??** eine Schätzmethode kennenlernen.

Definition 2.7 Ein **lineares Modell** mit n reellwertigen Beobachtungen $Y = (Y_1, \ldots, Y_n)^\top$ und p-dimensionalem Parameter $\beta \in \mathbb{R}^p$, $p \leqslant n$, besteht aus einer reellen Matrix $X \in \mathbb{R}^{n\times p}$ von vollem Rang p, der **Designmatrix**, und einem Zufallsvektor $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, den **Fehler- oder Störgrößen**, mit $\mathbb{E}[\varepsilon_i] = 0$ und positiv definiter Kovarianzmatrix $\Sigma := \mathbb{C}\mathrm{ov}(\varepsilon) = (\mathrm{Cov}(\varepsilon_i, \varepsilon_j))_{i,j=1,\ldots,n}$. Beobachtet wird eine Realisierung von

$$Y = X\beta + \varepsilon.$$

Wir sprechen vom **gewöhnlichen linearen Modell**, falls $\Sigma = \sigma^2 E_n$ für ein Fehlerniveau $\sigma > 0$ gilt.

Bemerkung 2.8 (symmetrisch, positiv-definit) Wir schreiben $\Sigma > 0$, falls Σ eine symmetrische, positiv-definite Matrix ist. Dann ist $\Sigma = TDT^\top$ diagonalisierbar mit einer Diagonalmatrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, Eigenwerten $\lambda_i > 0$, $i = 1, \ldots, n$, und einer Orthogonalmatrix T. Wir setzen $\Sigma^{-1/2} := TD^{-1/2}T^\top$ mit $D^{-1/2} := \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_n^{-1/2})$ und erhalten

$$(\Sigma^{-1/2})^2 = \Sigma^{-1} \quad \text{und} \quad |\Sigma^{-1/2} v|^2 = \langle \Sigma^{-1} v, v \rangle.$$

Zusätzlich zur einfachen und multiplen Regression umfasst das lineare Modell weitere wichtige Beispiele.

Beispiel 2.9 (Polynomiale Regression) Wir beobachten für ein $p \in \mathbb{N}$

$$Y_i = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_{p-1} x_i^{p-1} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Die Regressionsfunktion ist damit keine Gerade mehr, sondern ein Polynom vom Grad $p-1$. Die Koeffizienten des Polynoms bilden den unbekannten Parametervektor $\beta = (a_0, \ldots, a_{p-1})^\top$. Es ergibt sich eine Designmatrix vom Vandermonde-Typ

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix}.$$

Die Matrix hat vollen Rang, sofern p der Designpunkte (x_i) verschieden sind, was über die sogenannte Vandermonde-Determinante leicht nachzuweisen ist. Abbildung 2.2 zeigt $n = 100$ Beobachtungen, die durch ein Regressionspolynom vom Grad vier, äquidistante Designpunkte $x_i = (i-1)/(n-1)$ und i.i.d. Beobachtungsfehler $\varepsilon_i \sim \mathrm{N}(0; 0{,}1)$ erzeugt wurden.

Beispiel 2.10 (Orthogonales Design) Beobachten wir $(x_i, Y_i)_{i=1,\ldots,n}$ für reellwertige Kovariablen $x_i \in \mathbb{R}$ und Regressanden Y_i, können wir das Regressionsmodell in der Form

$$Y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

mit einer unbekannten Regressionsfunktion f schreiben. In der polynomiellen Regression postulieren wir, dass f ein Polynom vom Grad $p-1$ ist, und beschreiben f als Linearkombination der ersten p Monome $(x^{k-1})_{k=1,\ldots,p}$. Analog können andere Basisfunktionen $(\varphi_k)_{k=1,\ldots,p}$ verwendet werden, um $f(x) = \sum_{k=1}^p \beta_k \varphi_k(x)$ zu modellieren. Im Hinblick auf das Beobachtungsschema ist es nützlich, wenn die $(\varphi_i)_{i=1,\ldots,p}$ ein Orthonormalsystem bezüglich des *empirischen Skalarproduktes*

$$\langle f, g \rangle_n := \frac{1}{n} \sum_{i=1}^n f(x_i) g(x_i)$$

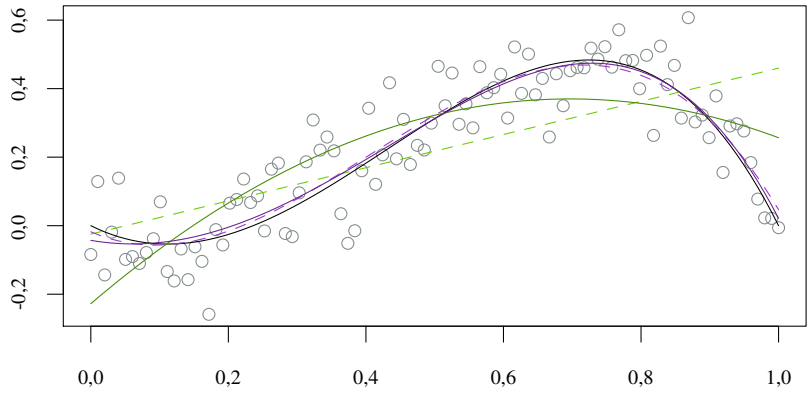

Abb. 2.2 Beobachtungen $(X_i, Y_i)_{i=1,\dots,n}$ aus einem polynomialen Regressionsmodell vom Grad 4 sowie die wahre Regressionsfunktion (schwarz) und geschätzte Regressionspolynome vom Grad eins (grün, gestrichelt), zwei (grün, durchgezogen), drei (violett, gestrichelt), vier (violett, durchgezogen)

bilden. In diesem Fall besitzt die Designmatrix $X = (\varphi_j(x_i))_{i=1,\dots,n, j=1,\dots,p}$ die Eigenschaft $X^\top X = nE_p$, weshalb wir von *orthogonalem Design* sprechen. Beispielsweise können die Monome mittels Gram-Schmidt-Verfahren bezüglich $\langle \cdot, \cdot \rangle_n$ orthogonalisiert werden. Es sei bemerkt, dass für äquidistantes Design das empirische Skalarprodukt für $n \to \infty$ gegen das L^2-Skalarprodukt konvergiert, siehe hierzu Beispiel 3.13.

2.1.2 Schätzen im linearen Modell

Übertragen wir die Kleinste-Quadrate-Methode (siehe Methode 2.3) von der einfachen linearen Regression auf den allgemeinen Fall, erhalten wir folgendes Verfahren zur Schätzung des Parametervektors β im linearen Modell:

Methode 2.11 (Kleinste-Quadrate-Schätzer) Der **gewichtete Kleinste-Quadrate-Schätzer** $\widehat{\beta}$ von β minimiert den gewichteten euklidischen Abstand zwischen Beobachtungen und Modellvorhersage:

$$|\Sigma^{-1/2}(Y - X\widehat{\beta})|^2 = \inf_{b \in \mathbb{R}^p} |\Sigma^{-1/2}(Y - Xb)|^2$$

Im gewöhnlichen Fall $\Sigma = \sigma^2 E_n$ ergibt sich der **gewöhnliche Kleinste-Quadrate-Schätzer** (englisch: *ordinary least squares*, kurz: OLS) $\widehat{\beta}$ mit

$$|Y - X\widehat{\beta}|^2 = \inf_{b \in \mathbb{R}^p} |Y - Xb|^2,$$

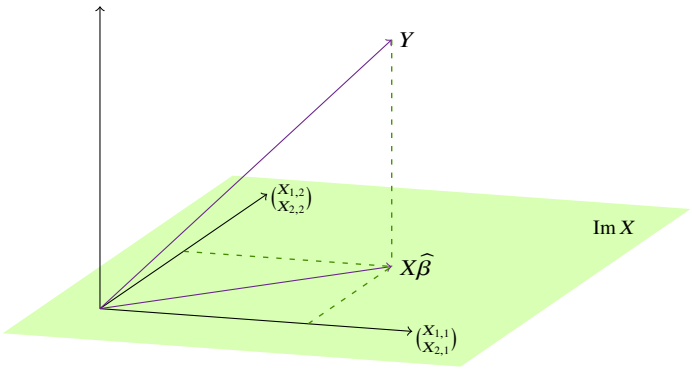

Abb. 2.3 Geometrische Interpretation des Kleinste-Quadrate-Schätzers mit $n = 3$ und $p = 2$

der unabhängig von der Kenntnis von σ^2 ist.

Bemerkung 2.12 (Gewichteter Kleinster-Quadrate-Schätzer) In der einfachen linearen Regression hatten wir den Kleinste-Quadrate-Schätzer mit normalverteilten, unabhängigen und identisch verteilten Fehlern hergeleitet. Das allgemeine lineare Modell können wir hierauf zurückführen, indem wir die beobachteten Daten entsprechend der Kovarianzmatrix Σ gewichten, genauer betrachten wir $\Sigma^{-1/2}Y$. Für die entsprechend gewichteten Fehler $\Sigma^{-1/2}\varepsilon$ gilt nämlich

$$\mathbb{C}\mathrm{ov}(\Sigma^{-1/2}\varepsilon) = \Sigma^{-1/2}\,\mathbb{C}\mathrm{ov}(\varepsilon)(\Sigma^{-1/2})^\top = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = E_n.$$

Aus $\Sigma^{-1/2}\varepsilon = \Sigma^{-1/2}(Y - X\beta)$ folgt der Ansatz des gewichteten Kleinste-Quadrate-Schätzers.

Beispiel 2.13 (Polynomiale Regression) Im Modell aus Beispiel 2.9 und Abbildung 2.2 wenden wir nun die Kleinste-Quadrate-Methode zur Schätzung der unbekannten Koeffizienten des zugrunde liegenden Regressionspolynoms an. Anhand der Beobachtungen aus Abbildung 2.2 kalibrieren wir polynomiale Regressionsmodelle der Grade 1, 2, 3 und 4. Da in den ersten drei Fällen das zur Schätzung verwendete Modell nicht mit dem wahren Modell, das die Daten erzeugt hat, übereinstimmt, weisen diese Schätzungen einen Modellfehler auf. Dieser zeigt sich insbesondere für die Grade eins und zwei in einer deutlichen Abweichung der geschätzten von der wahren Regressionsfunktion. Andererseits scheint bereits ein Polynom vom Grad drei die Daten gut zu beschreiben, da der zusätzliche vierte Grad kaum zu Änderungen führt.

Wir betrachten das gewöhnliche lineare Modell mit $\Sigma = E_n$ und einem zweidimensionalen Parameter $\beta \in \mathbb{R}^2$, das heißt $p = 2$. Die Designmatrix X ist dann eine $(n \times 2)$-Matrix und wir beobachten einen Punkt $Y \in \mathbb{R}^n$. Der Kleinste-Quadrate-Schätzer gibt nun den Wert $b = \widehat{\beta}$ an, an dem der euklidische Abstand zwischen Y und der Ebene $\mathrm{Im}\, X = \{Xb : b \in \mathbb{R}^2\}$ minimal ist. Folglich ist $X\widehat{\beta}$ die *Orthogonalprojektion* von Y auf $\mathrm{Im}\, X$, siehe Abbildung 2.3.

Diese geometrische Anschauung lässt sich auch ganz allgemein formulieren und führt zu einer expliziten Lösung des Minimierungsproblems.

Lemma 2.14 *Es seien* $\Sigma > 0$ *und* X *von vollem Rang. Setze*

$$X_\Sigma := \Sigma^{-1/2} X \qquad \textit{und} \qquad \Pi_{X_\Sigma} := X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top.$$

Dann ist Π_{X_Σ} *die Orthogonalprojektion von* $\mathbb{R}^n$ *auf den Bildraum* $\mathrm{Im}(X_\Sigma) = \{X_\Sigma b : b \in \mathbb{R}^p\}$ *und für den Kleinste-Quadrate-Schätzer gilt*

$$\widehat{\beta} = X_\Sigma^{-1}(\Pi_{X_\Sigma} \Sigma^{-1/2} Y) = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} Y.$$

Dabei ist $X_\Sigma^{-1}(v)$ *für* $v \in \mathrm{Im}(X_\Sigma)$ *das eindeutig bestimmte Urbild von* v *unter* X_Σ *(für* $p < n$ *ist* X_Σ *keine invertierbare Matrix). Insbesondere existiert der Kleinste-Quadrate-Schätzer, ist eindeutig und erwartungstreu, und es gilt* $X_\Sigma \widehat{\beta} = \Pi_{X_\Sigma}(\Sigma^{-1/2} Y)$.

Beweis Die Positivität und Symmetrie von Σ liefert

$$\Sigma^{-1/2} = (\Sigma^{-1/2})^\top \quad \text{und} \quad X_\Sigma^\top = \big(\Sigma^{-1/2} X\big)^\top = X^\top \big(\Sigma^{-1/2}\big)^\top = X^\top \Sigma^{-1/2}. \tag{2.1}$$

Wir zeigen zuerst, dass $X_\Sigma^\top X_\Sigma = X^\top \Sigma^{-1} X$ invertierbar und somit Π_{X_Σ} wohldefiniert ist: Für jedes $v \in \mathbb{R}^p$ mit $X^\top \Sigma^{-1} X v = 0$ folgt aus (2.1)

$$0 = v^\top X^\top \Sigma^{-1} X v = |\Sigma^{-1/2} X v|^2.$$

Da $\Sigma^{-1/2}$ vollen Rang hat, muss dann $|Xv| = 0$ gelten. Aus dem vollen Rang von X folgt wiederum $v = 0$. Also besteht der Kern von $X_\Sigma^\top X_\Sigma$ nur aus dem Nullvektor, und $X_\Sigma^\top X_\Sigma$ ist invertierbar.

Wir setzen nun $\Pi_{X_\Sigma} := X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top$ und $w = \Pi_{X_\Sigma} v$ für ein $v \in \mathbb{R}^n$. Dann folgt $w \in \mathrm{Im}(X_\Sigma)$ und im Fall $v = X_\Sigma u$ durch Einsetzen $w = \Pi_{X_\Sigma} X_\Sigma u = v$, sodass Π_{X_Σ} eine Projektion auf $\mathrm{Im}(X_\Sigma)$ ist. Zudem ist Π_{X_Σ} selbstadjungiert (symmetrisch) wegen

$$\begin{aligned} \big((X_\Sigma^\top X_\Sigma)^{-1}\big)^\top &= \big((X^\top \Sigma^{-1} X)^{-1}\big)^\top = \big((X^\top \Sigma^{-1} X)^\top\big)^{-1} \\ &= \big(X^\top (\Sigma^{-1})^\top X\big)^{-1} = \big(X^\top \Sigma^{-1} X\big)^{-1} = \big(X_\Sigma^\top X_\Sigma\big)^{-1}. \end{aligned} \tag{2.2}$$

Somit ist Π_{X_Σ} sogar eine Orthogonalprojektion:

$$\forall u \in \mathbb{R}^n, \forall w \in \mathrm{Im}(X_\Sigma) : \langle u - \Pi_{X_\Sigma} u, w \rangle = \langle u, w \rangle - \langle u, \Pi_{X_\Sigma} w \rangle = 0$$

Aus der Eigenschaft $\widehat{\beta} = \arg\min_b |\Sigma^{-1/2}(Y - Xb)|^2$ folgt, dass $\widehat{\beta}$ die beste Approximation von $\Sigma^{-1/2} Y$ durch $X_\Sigma b$ liefert. Diese ist durch die Orthogonalprojektionseigenschaft

$$\Pi_{X_\Sigma} \Sigma^{-1/2} Y = X_\Sigma \widehat{\beta}$$

bestimmt. Es folgt

$$X_\Sigma^\top \Pi_{X_\Sigma} \Sigma^{-1/2} Y = (X_\Sigma^\top X_\Sigma)\widehat{\beta} \Rightarrow (X_\Sigma^\top X_\Sigma)^{-1} X^\top \Sigma^{-1} Y = \widehat{\beta}.$$

Schließlich folgt aus der Linearität des Erwartungswerts und $\mathbb{E}[\varepsilon] = 0$

$$\mathbb{E}[\widehat{\beta}] = \mathbb{E}[(X_\Sigma^\top X_\Sigma)^{-1} X^\top \Sigma^{-1}(X\beta + \varepsilon)] = \beta + 0 = \beta.$$

Damit ist $\widehat{\beta}$ erwartungstreu. □

Bemerkung 2.15

- Im gewöhnlichen linearen Modell gilt $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$, sodass der Kleinste-Quadrate-Schätzer unabhängig vom unbekannten Parameter $\sigma > 0$ ist.
- $X_\Sigma^\dagger := (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top$ heißt auch *Moore-Penrose-(Pseudo-)Inverse* von X_Σ. Die Bezeichnung Pseudoinverse ist motiviert durch die Eigenschaften $X_\Sigma^\dagger X_\Sigma = E_p$ und $X_\Sigma X_\Sigma^\dagger|_{\mathrm{Im}(X_\Sigma)} = E_n|_{\mathrm{Im}(X_\Sigma)}$. Insbesondere erhalten wir $\widehat{\beta} = X_\Sigma^\dagger \Sigma^{-1/2} Y$ bzw. die Vereinfachung $\widehat{\beta} = X^\dagger Y$ im gewöhnlichen linearen Modell.

Der folgende zentrale Satz der Regressionsanalyse zeigt, dass der Kleinste-Quadrate-Schätzer optimal ist, wenn wir uns auf Schätzer beschränken, die linear in den Daten Y und erwartungstreu sind.

Satz 2.16 (Gauß-Markov) *Im linearen Modell gilt:*

(i) Der Parameter $\rho = \langle \beta, v\rangle$ für ein $v \in \mathbb{R}^p$ wird von $\widehat{\rho} = \langle \widehat{\beta}, v\rangle$ erwartungstreu geschätzt, und $\widehat{\rho}$ besitzt unter allen linearen erwartungstreuen Schätzern die minimale Varianz $\mathrm{Var}(\widehat{\rho}) = |X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} v|^2$.

(ii) Der Kleinste-Quadrate-Schätzer $\widehat{\beta}$ besitzt unter allen linearen erwartungstreuen Schätzern von β minimale Kovarianzmatrix, nämlich $\mathbb{C}\mathrm{ov}(\widehat{\beta}) = (X_\Sigma^T X_\Sigma)^{-1}$, das heißt für alle linearen, erwartungstreuen Schätzer $\widetilde{\beta}$ ist $\mathbb{C}\mathrm{ov}(\widetilde{\beta}) - \mathbb{C}\mathrm{ov}(\widehat{\beta}) \geqslant 0$ eine positiv semi-definite Matrix.

Beweis *(i)* Die Linearität ist klar und aus dem vorangegangenen Lemma folgt, dass $\widehat{\rho}$ erwartungstreu ist. Für die Varianz ergibt sich

$$\begin{aligned}
\mathrm{Var}(\widehat{\rho}) &= \mathbb{E}\big[\langle \widehat{\beta} - \beta, v\rangle^2\big] \\
&= \mathbb{E}\big[\langle (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \varepsilon, v\rangle^2\big] \\
&= \mathbb{E}\big[\langle \varepsilon, \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} v\rangle^2\big] \\
&= v^\top (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \Sigma \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} v \\
&= |X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} v|^2.
\end{aligned}$$

Sei nun $\widetilde{\rho}$ ein beliebiger linearer Schätzer von ρ. Dann gibt es ein $w \in \mathbb{R}^n$, sodass $\widetilde{\rho} = \langle Y, w\rangle$ (Satz von Riesz). Dies impliziert, dass für alle $\beta \in \mathbb{R}^p$

$$\mathbb{E}[\langle Y, w\rangle] = \rho \Rightarrow \mathbb{E}[\langle Y, w\rangle] = \langle X\beta, w\rangle = \langle \beta, v\rangle \Rightarrow \langle X^\top w - v, \beta\rangle = 0$$

und somit $v = X^\top w = X_\Sigma^\top \Sigma^{1/2} w$. Da Π_{X_Σ} eine Orthogonalprojektion ist, erhalten wir mit dem Satz von Pythagoras

$$\begin{aligned}\mathrm{Var}(\widetilde{\rho}) &= \mathbb{E}[\langle \varepsilon, w\rangle^2] = \mathbb{E}[w^\top \varepsilon \varepsilon^\top w] \\ &= w^\top \Sigma w = |\Sigma^{1/2} w|^2 = |\Pi_{X_\Sigma}(\Sigma^{1/2} w)|^2 + |(E_n - \Pi_{X_\Sigma})(\Sigma^{1/2} w)|^2.\end{aligned}$$

Damit gilt

$$\begin{aligned}\mathrm{Var}(\widetilde{\rho}) \geqslant |\Pi_{X_\Sigma}(\Sigma^{1/2} w)|^2 &= |X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} X^\top w|^2 \\ &= |X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} v|^2 = \mathrm{Var}(\widehat{\rho}).\end{aligned}$$

(ii) Sei $\widetilde{\beta}$ ein linearer, erwartungstreuer Schätzer. Dann ist $\mathbb{C}\mathrm{ov}(\widetilde{\beta}) - \mathbb{C}\mathrm{ov}(\widehat{\beta})$ genau dann eine positiv semi-definite Matrix, wenn

$$\forall v \in \mathbb{R}^p : v^\top (\mathbb{C}\mathrm{ov}(\widetilde{\beta}) - \mathbb{C}\mathrm{ov}(\widehat{\beta}))v \geqslant 0.$$

Sei $v \in \mathbb{R}^p$. Nach Annahme sind $\langle v, \widetilde{\beta}\rangle$ und $\langle v, \widehat{\beta}\rangle$ lineare, erwartungstreue Schätzer und aus *(i)* folgt daher

$$0 \leqslant \mathrm{Var}(\langle v, \widetilde{\beta}\rangle) - \mathrm{Var}(\langle v, \widehat{\beta}\rangle) = v^\top (\mathbb{C}\mathrm{ov}(\widetilde{\beta}) - \mathbb{C}\mathrm{ov}(\widehat{\beta}))v.$$

Weiter gilt wegen *(i)* für beliebiges $v \in \mathbb{R}^p$

$$\begin{aligned}v^\top \mathbb{C}\mathrm{ov}(\widehat{\beta}) v &= \mathrm{Var}(\langle \widehat{\beta}, v\rangle) \\ &= |X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} v|^2 \\ &= v^\top (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} v \\ &= v^\top (X_\Sigma^\top X_\Sigma)^{-1} v,\end{aligned}$$

woraus $\mathbb{C}\mathrm{ov}(\widehat{\beta}) = (X_\Sigma^\top X_\Sigma)^{-1}$ folgt. □

Kurzbiografie (Andrey Andreyevich Markov) Andrey Andreyevich Markov wurde 1856 in Rjasan, rund 200 km südöstlich von Moskau, geboren und wuchs in St. Petersburg auf. Er studierte an der St. Petersburger Universität Mathematik und Physik, wobei er unter anderem Vorlesungen von Tschebyscheff besuchte. Kurz nach seiner Promotion wurde er Professor und in die russische Akademie der Wissenschaften gewählt. Markov leistete wichtige Beiträge zur Stochastik und Analysis und insbesondere zu stochastischen Prozessen, wobei etliche Resultate nach ihm benannt wurden, beispielsweise die Markov-Eigenschaft, die Markov-Ungleichung, der Markov-Prozess und der *Satz von Gauß-Markov*. Markov starb 1922.

Da wir das Funktional $\rho := \rho(\beta) := \langle v, \beta\rangle$ mit $\widehat{\rho} = \rho(\widehat{\beta})$ durch Einsetzen von $\widehat{\beta}$ schätzen, nennt man $\widehat{\rho}$ *plugin-Schätzer*. Eine typische Anwendung sind Vorhersagen. Aufgrund des Schätzers $\widehat{\beta}$ von β können wir für neue Kovariablen $v := (x_{n+1,1}, ..., x_{n+1,p})^\top$ die zugehörige Beobachtung $Y_{n+1} = v^\top \beta + \varepsilon_{n+1}$ vorhersagen, indem wir $\langle v, \widehat{\beta}\rangle$ berechnen. Eine weitere wichtige Anwendung ist die Schätzung einzelner Koeffizienten des Vektors β.

Bemerkung 2.17 (BLUE, GLS)

1. Man sagt, dass der Schätzer $\widehat{\rho}$ im Satz von Gauß-Markov **bester linearer erwartungstreuer Schätzer** (englisch: *best linear unbiased estimator*, kurz: BLUE) ist. Eingeschränkt auf lineare, erwartungstreue Schätzer ist der Kleinste-Quadrate-Schätzer damit minimax bezüglich quadratischem Verlust, siehe Definition 1.25. Ob es einen besseren nichtlinearen Schätzer geben kann, werden wir in Kapitel **??** beantworten.
2. Der Kleinste-Quadrate-Schätzer mit allgemeiner Kovarianzmatrix Σ wird im Englischen zur Abhebung vom gewöhnlichen Kleinste-Quadrate-Schätzer (OLS) auch verallgemeinerter Kleinste-Quadrate-Schätzer (englisch: *generalized least squares*, kurz: GLS) genannt und wurde erstmals von Alexander Aitken im Jahre 1934 beschrieben. Aitken hatte den Satz von Gauß-Markov, der sich eigentlich nur mit dem gewöhnlichen Fall beschäftigte, auf das allgemeine Modell übertragen.
3. Wegen $\mathbb{E}[|Z|^2] = \sum_{i=1,\dots,p} \mathbb{E}[Z_i^2] = \operatorname{tr}(\mathbb{E}[ZZ^\top])$ für Zufallsvektoren $Z \in \mathbb{R}^p$ und mit der Spur $\operatorname{tr}(\cdot)$ einer Matrix ist der mittlere quadratische Fehler von $\widehat{\beta}$ gegeben durch $\mathbb{E}[|\widehat{\beta} - \beta|^2] = \operatorname{tr}((X_\Sigma^T X_\Sigma)^{-1})$. Für ein gewöhnliches lineares Modell mit orthogonalem Design (Beispiel 2.10) gilt $X_\Sigma^T X_\Sigma = X^\top \Sigma^{-1} X = \frac{n}{\sigma^2} E_p$ und daher

$$\mathbb{E}[|\widehat{\beta} - \beta|^2] = \frac{\sigma^2 p}{n}. \tag{2.3}$$

   Die Abhängigkeit des Fehlers vom Stichprobenumfang n, der Parameterdimension p und dem Rauschniveau σ ist hier besonders offenkundig.

Im Spezialfall des gewöhnlichen linearen Modells ist es von großem Interesse, das *Rauschniveau* σ^2 zu schätzen. Dieses ist der einzige unbekannte Wert in der Fehlerformel (2.3) und wird uns die Konstruktion von Tests und Konfidenzbereichen ermöglichen.

Lemma 2.18 *Im gewöhnlichen linearen Modell mit $\sigma > 0$ und Kleinste-Quadrate-Schätzer $\widehat{\beta}$ gilt $X\widehat{\beta} = \Pi_X Y$ (mit $\Pi_X := \Pi_{X_{E_n}}$). $R := Y - X\widehat{\beta}$ bezeichne den Vektor der **Residuen**. Dann ist für $p < n$ die Stichprobenvarianz*

$$\widehat{\sigma}^2 := \frac{|R|^2}{n-p} = \frac{|(E_n - \Pi_X)Y|^2}{n-p}$$

ein erwartungstreuer Schätzer von σ^2.

Beweis $X\widehat{\beta} = \Pi_X Y$ folgt aus Lemma 2.14. Einsetzen zeigt $\mathbb{E}[|Y - X\widehat{\beta}|^2] = \mathbb{E}[|Y - \Pi_X Y|^2] = \mathbb{E}[|(E_n - \Pi_X)\varepsilon|^2]$. Mithilfe der Spur und Eigenschaften der Orthogonalprojektion $E_n - \Pi_X$ vom Rang $n - p$ berechnen wir

$$\mathbb{E}\big[|(E_n - \Pi_X)\varepsilon|^2\big] = \operatorname{tr}\big((E_n - \Pi_X)\mathbb{E}[\varepsilon\varepsilon^\top](E_n - \Pi_X)\big) = \sigma^2 \operatorname{tr}(E_n - \Pi_X) = \sigma^2(n-p),$$

was die Behauptung impliziert. □

Eine alternative Normierung der Stichprobenvarianz ergibt sich aus dem Maximum-Likelihood-Ansatz im normalverteilten gewöhnlichen linearen Modell. Dieser

liefert $\widehat{\sigma}^2_{ML} = |R|^2/n$ als Schätzer für σ^2 (siehe Aufgabe 2.6). In der Praxis wird der erwartungstreue Schätzer $\widehat{\sigma}^2$ häufig bevorzugt, wobei dieser jedoch eine größere Varianz als $\widehat{\sigma}^2_{ML}$ aufweist.

Nachdem uns die Maximum-Likelihood-Methode auf den Kleinste-Quadrate-Schätzer geführt hat, soll nun der Bayes-Ansatz für das lineare Modell untersucht werden.

Satz 2.19 (Bayes-Schätzer im linearen Modell) *Im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ und bekanntem $\sigma > 0$ folge $\beta \in \mathbb{R}^p$ der a-priori-Verteilung*

$$\beta \sim \mathrm{N}(m, \sigma^2 M)$$

mit $m \in \mathbb{R}^p$ und symmetrischer, positiv definiter Matrix $M \in \mathbb{R}^{p\times p}$. Dann ist die a-posteriori-Verteilung von β, gegeben eine realisierte Beobachtung $y \in \mathbb{R}^n$, wiederum normalverteilt:

$$\beta|Y = y \sim \mathrm{N}(\mu_y, \sigma^2\Sigma_y) \quad \textit{mit} \quad \Sigma_y = \left(X^\top X + M^{-1}\right)^{-1},\ \mu_y = \Sigma_y(X^\top y + M^{-1}m)$$

Insbesondere ist der Bayes-Schätzer bezüglich quadratischem Verlust gegeben durch

$$\widehat{\beta}_{\mathrm{Bayes}} = \left(X^\top X + M^{-1}\right)^{-1}(X^\top Y + M^{-1}m).$$

Beweis Für die a-posteriori-Dichte an der Stelle $t \in \mathbb{R}^p$ gilt

$$\begin{aligned}
f^{\beta|Y=y}(t) &\propto f^{Y|\beta=t}(y) f_\beta(t) \\
&\propto \exp\Big(-\frac{1}{2\sigma^2}(y - Xt)^\top(y - Xt)\Big)\exp\Big(-\frac{1}{2\sigma^2}(t-m)^\top M^{-1}(t-m)\Big) \\
&\propto \exp\Big(\frac{1}{\sigma^2}t^\top X^\top y - \frac{1}{2\sigma^2}t^\top X^\top X t - \frac{1}{2\sigma^2}t^\top M^{-1}t + \frac{1}{\sigma^2}t^\top M^{-1}m\Big) \\
&= \exp\Big(\frac{1}{\sigma^2}t^\top \underbrace{\left(X^\top y + M^{-1}m\right)}_{=\Sigma_y^{-1}\mu_y} - \frac{1}{2\sigma^2}t^\top \underbrace{\left(X^\top X + M^{-1}\right)}_{=\Sigma_y^{-1}} t\Big) \\
&\propto \exp\Big(-\frac{1}{2\sigma^2}\left(t^\top\Sigma_y^{-1}t - 2t^\top\Sigma_y^{-1}\mu_y + \mu_y^\top\Sigma_y^{-1}\mu_y\right)\Big) \\
&= \exp\Big(-\frac{1}{2\sigma^2}(t-\mu_y)^\top\Sigma_y^{-1}(t-\mu_y)\Big).
\end{aligned}$$

Daher ist β, gegeben $Y = y$, normalverteilt mit der Kovarianzmatrix $\sigma^2\Sigma_y$ und dem Erwartungswert μ_y. Korollar 1.33 liefert den Rest der Behauptung. □

Bemerkung 2.20 (Mehrstufiges Bayes-Modell) Indem wir den Parameter σ^2 mit einer a-priori-Verteilung versehen und β gemäß einer von σ abhängigen a-priori-Verteilung wählen, erhalten wir ein mehrstufiges Bayes-Modell. Da besonders konjugierte Verteilungsklassen von Interesse sind, wird hierzu oft die *inverse Gammaverteilung* verwendet: Ist $Z \sim \Gamma(a, b)$, so ist $1/Z \sim \mathrm{IG}(a, b)$ invers gammaverteilt

mit Parametern $a, b > 0$ und Lebesgue-Dichte

$$f_{a,b}(x) = \frac{b^a}{\Gamma(a)} x^{-(a-1)} e^{-b/x} \mathbb{1}_{(0,\infty)}(x), \quad x \in \mathbb{R}.$$

Skalieren wir die Varianz des normalverteilten β mit σ^2, erhalten wir das Bayes-Modell

$$Y|\beta, \sigma^2 \sim \mathrm{N}(X\beta, \sigma^2 E), \quad \beta|\sigma^2 \sim \mathrm{N}(m, \sigma^2 M), \quad \sigma^2 \sim \mathrm{IG}(a, b).$$

Die gemeinsame Verteilung von $(\beta, \sigma^2) \sim \mathrm{NIG}(m, M, a, b)$ wird *normal-inverse Gammaverteilung* genannt und besitzt die Dichte

$$f(\beta, \sigma^2) = \frac{C}{(\sigma^2)^{p/2+a+1}} \exp\Big(-\frac{1}{2\sigma^2}\big((\beta - m)^\top M^{-1}(\beta - m) + 2b\big)\Big) \qquad \text{mit}$$
$$C = \frac{b^a}{(2\pi)^{p/2}|M|^{1/2}\Gamma(a)}, \qquad \beta \in \mathbb{R}^p, \sigma^2 > 0.$$

In diesem Modell ist die a-posteriori-Verteilung von σ^2, gegeben β und Y, gegeben durch $\sigma^2|\beta, Y \sim \mathrm{IG}(a', b')$ mit $a' = a + \frac{n}{2} + \frac{p}{2}$ und

$$b' = b + \frac{1}{2}(Y - X\beta)^\top (Y - X\beta) + \frac{1}{2}(\beta - m)^\top M^{-1}(\beta - m).$$

Die a-posteriori-Verteilung von (β, σ^2), gegeben Y, ist $(\beta, \sigma^2)|Y \sim \mathrm{NIG}(\widetilde{m}, \widetilde{M}, \widetilde{a}, \widetilde{b})$ mit den Parametern

$$\widetilde{M} = (X^\top X + M^{-1})^{-1}, \quad \widetilde{m} = \widetilde{M}(M^{-1}m + X^\top y),$$
$$\widetilde{a} = a + \frac{n}{2}, \quad \widetilde{b} = b + \frac{1}{2}\Big(Y^\top Y + m^\top M^{-1} m - \widetilde{m}^\top \widetilde{M}^{-1} \widetilde{m}\Big),$$

siehe Fahrmeir et al. (2009, Kapitel 3.5).

In einem Spezialfall von Satz 2.19 erhalten wir eine weitere Darstellung des Bayes-Schätzers.

Korollar 2.21 *Unter den Voraussetzungen von Satz 2.19 mit $m = 0$ und $M = \lambda^{-1} E_p$ für ein $\lambda > 0$ gilt für den Bayes-Schätzer unter quadratischem Verlust*

$$\widehat{\beta}_{\mathrm{Bayes}} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \Big(|Y - X\beta|^2 + \lambda |\beta|^2 \Big).$$

Beweis Im Spezialfall $m = 0$ und $M = \lambda^{-1} E_p$ folgt aus Satz 2.19

$$\widehat{\beta}_{\mathrm{Bayes}} = \big(X^\top X + \lambda E_p\big)^{-1} X^\top Y.$$

Andererseits gilt

$$\arg\min_{\beta} \Big(|Y - X\beta|^2 + \lambda\beta^\top\beta \Big) = \arg\min_{\beta} \Big(-2Y^\top X\beta + \beta^\top (X^\top X + \lambda E_p)\beta \Big).$$

Null setzen des Differenzials in β liefert $0 = -2Y^\top X + 2\beta^\top (X^\top X + \lambda E_p)$, sodass aus der Positivität und Symmetrie von $X^\top X + \lambda E_p$ die Behauptung folgt. □

Der Bayes-Ansatz führt uns also zu einer neuen Schätzmethode im linearen Modell:

Methode 2.22 (Ridge-Regression) Im linearen Modell $Y = X\beta + \varepsilon$ ist der **Ridge-Regressionsschätzer** oder **Shrinkage-Schätzer** mit Penalisierung $\lambda > 0$ definiert als

$$\widehat{\beta}_{\text{ridge}} = \arg\min_{\beta \in \mathbb{R}^p} \big(|Y - X\beta|^2 + \lambda |\beta|^2 \big).$$

Der Strafterm $\lambda|\beta|^2$ führt zu Lösungen des Minimierungsproblems, die kleine Parametervektoren β bevorzugen, und daher zur Bezeichnung *shrinkage* („Schrumpfung“) führt. Diesen Effekt sieht man auch direkt am Bayes-Ansatz in Satz 2.19, da dort eine a-priori-Verteilung verwendet wird, die um $\beta = 0$ zentriert ist. Um die Bezeichnung *ridge* zu verstehen, erinnern wir uns daran, dass die ursprüngliche Motivation des Kleinste-Quadrate-Schätzers das Maximum-Likelihood-Prinzip war. Führen nun mehrere Parameterwahlen β zu vergleichbar großen Likelihoods, ähnelt die Kontur der Likelihood-Funktion einem Bergkamm (englisch: *ridge*). Das Finden des Maximums oder äquivalent des Minimums der quadrierten Residuen (in einem langen Tal) ist numerisch schwierig. Wenn wir durch Hinzufügen des strikt konvexen Strafterms das Tal an den Seiten anheben, entsteht ein leichter zu findendes globales Minimum.

Abbildung 2.4 illustriert diesen Sachverhalt, wobei wir für die Simulation folgendes Modell benutzt haben: $X \in \mathbb{R}^{10\times 2}$, $X_{i,1} = 1$, $X_{j,2} \sim \mathrm{U}([0,1])$ für alle $i, j = 1, ..., 10$, $\beta = (2, 1/2)^\top$, $\varepsilon_i \sim \mathrm{N}(0, 1/10)$ mit $\Sigma = 0,1 \cdot E_{10}$ und $\lambda = 15$.

Den Einfluss des Strafterms wollen wir im Spezialfall von orthogonalem Design, das heißt für $X^\top X = nE_p$ und $p \leqslant n$, genauer untersuchen, siehe auch Aufgabe 2.9 zum Vergleich zwischen Ridge-Regressionsschätzer und dem Kleinste-Quadrate-Schätzer.

Lemma 2.23 *Im gewöhnlichen linearen Modell $Y = X\beta+\varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$, $\sigma > 0$ und $X^\top X = nE_p$ gilt für den Ridge-Regressionsschätzer mit Penalisierungsparameter $\lambda > 0$*

$$\mathbb{E}\big[|\widehat{\beta}_{\text{ridge}} - \beta|^2\big] = \frac{|\beta|^2}{(1 + n/\lambda)^2} + \frac{1}{(1 + \lambda/n)^2}\frac{\sigma^2 p}{n}.$$

Beweis Aus Korollar 2.21 erhalten wir $\widehat{\beta}_{\text{ridge}} = (X^\top X + \lambda E_p)^{-1}(X^\top Y)$, sodass

$$\begin{aligned}\widehat{\beta}_{\text{ridge}} - \beta &= (X^\top X + \lambda E_p)^{-1} X^\top X\beta - \beta + (X^\top X + \lambda E_p)^{-1} X^\top \varepsilon \\ &= -(X^\top X + \lambda E_p)^{-1}\lambda\beta + (X^\top X + \lambda E_p)^{-1} X^\top \varepsilon.\end{aligned}$$

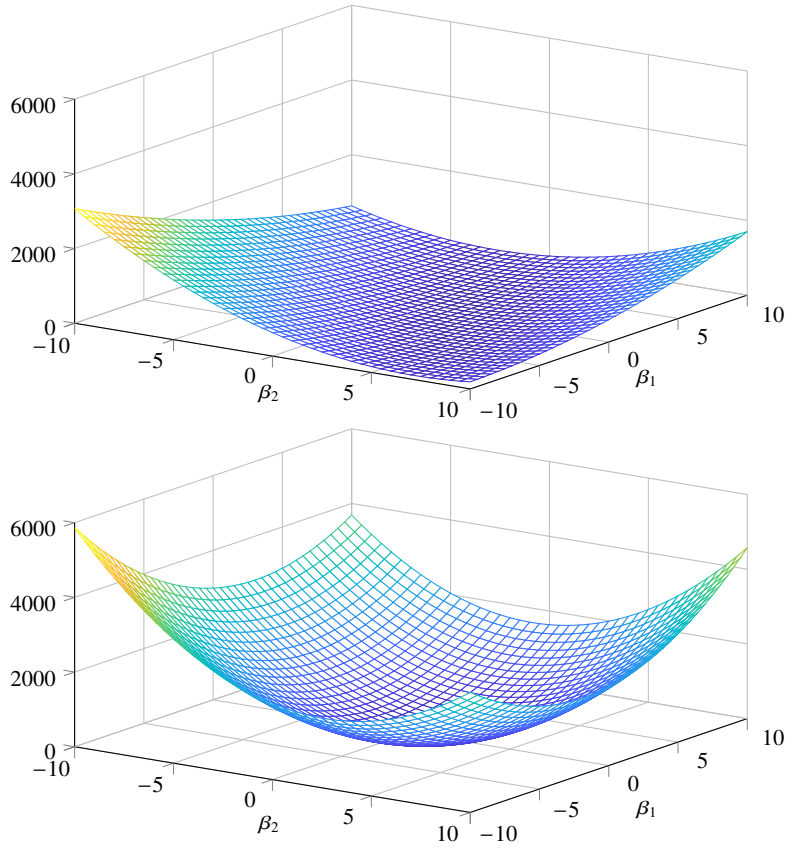

Abb. 2.4 Quadrierte Residuen in Abhängigkeit von $\beta = (\beta_1, \beta_2)^\top \in \mathbb{R}^2$ ohne Strafterm (*oben*) und mit ℓ^2-Strafterm (*unten*)

Aus der Bias-Varianz-Zerlegung, $\mathbb{E}[|Z|^2] = \operatorname{tr}(\mathbb{E}[ZZ^\top])$ für beliebige Zufallsvektoren $Z \in \mathbb{R}^p$ und dem Einsetzen von $X^\top X = nE_p$ folgt

$$\begin{aligned}\mathbb{E}[|\widehat{\beta}_{\text{ridge}} - \beta|^2] &= \lambda^2 \left|(X^\top X + \lambda E_p)^{-1}\beta\right|^2 \\ &\quad + \sigma^2 \operatorname{tr}\left((X^\top X + \lambda E_p)^{-1} X^\top X (X^\top X + \lambda E_p)^{-1}\right) \\ &= \frac{|\beta|^2}{(1 + n\lambda^{-1})^2} + \frac{\sigma^2 pn}{(n+\lambda)^2}.\end{aligned}$$

Damit ergibt sich die behauptete Darstellung des mittleren quadratischen Fehlers von $\widehat{\beta}_{\text{ridge}}$. □

Im Vergleich zum mittleren quadratischen Fehler $\mathbb{E}[|\widehat{\beta} - \beta|^2] = \sigma^2 p/n$ des Kleinste-Quadrate-Schätzers $\widehat{\beta}$ wird die Varianz des Ridge-Regressionsschätzers auf Kosten eines zusätzlichen Bias verringert. Da der Bias proportional zu $|\beta|^2$ wächst, ist

der Shrinkageansatz insbesondere sinnvoll, wenn $|\beta|$ klein ist. Die richtige Wahl des Penalisierungsparameters λ ist hierbei allerdings entscheidend. Eine optimale Wahl von λ hängt vom unbekannten Vektor β ab und ist dem Statistiker nicht zugänglich (Aufgabe 2.9).

Der Ridge-Regressionschätzer liefert bei kleinen $|\beta|^2$ auch noch gute Schätzergebnisse, wenn die Parameterdimension nicht viel kleiner als die Anzahl der Beobachtungen ist: Die in der Dimension linear wachsende Varianz $\sigma^2 p$ des Kleinste-Quadrate-Schätzers wird durch den Penalisierungsparameter reduziert.

Die Ridge-Regression wird insbesondere bei schlecht konditionierter Designmatrix X, also wenn das Verhältnis von größtem zu kleinstem Eigenwert von $X^\top X$ groß ist, verwendet. Dann ist einerseits die Lösung des Minimierungsproblems $\widehat{\beta}_{\text{ridge}} = (X^\top X + \lambda E_p)^{-1} X^\top Y$ numerisch stabiler, da die Matrix $X^\top X + \lambda E_p$ besser konditioniert ist. Andererseits kann man analog zum Lemma beweisen, dass die Varianz des Schätzers insbesondere in Richtung der Eigenvektoren zu den kleinen Eigenwerten von $X^\top X$ stark reduziert wird. Der Strafterm führt also zu einer numerischen wie statistischen Regularisierung.

Möchte man β in einem hochdimensionalen linearen Modell schätzen, also im Fall $p \gg n$, ist auch der Ridge-Regressionsschätzer nicht immer zielführend. Man kann aber den ℓ^2-Strafterm durch einen ℓ^1-Strafterm ersetzen, was typischerweise zu spärlich besetzten (englisch: *sparse*) Lösungen des Minimierungsproblems führt und auch bei sehr großen Parameterdimensionen p oft gut funktioniert. Die resultierende Schätzmethode ist der sogenannte Lasso-Schätzer, den wir in Kapitel **??** studieren werden.

2.1.3 Zufälliges Design und Vorhersage

Oft sind die Kovariablen im linearen Modell selbst zufällig und damit auch die Designmatrix X. Da wir neben der Zielgröße Y auch X beobachten, gelten die bisherigen Resultate zur Schätzung von β und σ weiter, sofern wir nur auf die Realisierung von X bedingen. Ein neuer wesentlicher Aspekt ergibt sich aber, wenn wir die Vorhersage der Zielgröße, gegeben die zugehörige Kovariable, betrachten.

Definition 2.24 Im linearen Modell mit **zufälligem Design** beobachten wir $(X_i, Y_i)_{i=1,\dots,n}$ mit

$$Y_i = \langle X_i, \beta \rangle + \varepsilon_i = (X\beta)_i + \varepsilon_i, \quad i = 1, \dots, n,$$

und der Designmatrix $X = (X_1, \dots, X_n)^\top$ ($X_i^\top$ ist die i-te Zeile in X). Dabei seien $\beta \in \mathbb{R}^p$ der unbekannte Parameter und $(X_i, \varepsilon_i)_{i=1,\dots,n}$ unabhängige Zufallsvariablen mit identisch verteilten Zufallsvektoren X_i im $\mathbb{R}^p$ und identisch verteilten reellwertigen Zufallsvariablen ε_i, für die $\mathbb{E}[\varepsilon_i] = 0$, $\sigma^2 := \mathbb{E}[\varepsilon_i^2] < \infty$ gilt und die unabhängig von X_i sind. Zusätzlich gelte $\mathbb{E}[|X_i|^2] < \infty$, sodass die **Design-Kovarianzmatrix**

$$\Sigma_X = \mathbb{E}[X_1 X_1^\top]$$

wohldefiniert ist.

Wegen der Unabhängigkeit der ε_i liegt ein gewöhnliches lineares Modell $Y = X\beta + \varepsilon$ mit $\Sigma = \sigma^2 E_n$ vor. Allerdings fordern wir hier nicht a priori, dass X vollen Rang p besitzt. Die Beobachtungen $(X_i, Y_i)_{i=1,\dots,n}$ sind also i.i.d. mit $\mathbb{E}[Y_i - \langle X_i, \beta\rangle] = 0$. Man beachte, dass oft allgemeiner nur mit Hilfe der bedingten Erwartung $\mathbb{E}[Y_i \mid X_i] = \langle X_i, \beta\rangle$, also $\mathbb{E}[\varepsilon_i \mid X_i] = 0$ gefordert wird, was bei uns aus der Unabhängigkeit von ε_i und X_i sowie $\mathbb{E}[\varepsilon_i] = 0$ folgt. Falls X_1 nicht zentriert ist, ist Σ_X nicht mehr die Kovarianzmatrix des Designs und man müsste eigentlich von der Matrix der zweiten Momente sprechen, was aber nicht gängig ist.

In der Sprache des maschinellen Lernens wollen wir nun auf der Grundlage der Trainingsdaten $(X_i, Y_i)_{i=1,\dots,n}$ bei einer neuen Kovariablen (oder einem neuen Feature) X_{n+1} die Zielgröße Y_{n+1} vorhersagen. Im Folgenden schreiben wir $\mathbb{E}^{(X_{n+1}, Y_{n+1})}$ für den Erwartungswert nur bezüglich der Zufallsvariablen X_{n+1}, Y_{n+1}.

Definition 2.25 Für $(X_i, Y_i)_{i=1,\dots,n+1}$ gelte das lineare Modell mit zufälligem Design. Für einen Schätzer $\widehat{\beta}_n$, basierend auf $(X_i, Y_i)_{i=1,\dots,n}$, bezeichnet

$$\ell^{\text{pred}}(\widehat{\beta}_n, \beta) = \mathbb{E}_\beta^{(X_{n+1}, Y_{n+1})}\left[\left(Y_{n+1} - \langle X_{n+1}, \widehat{\beta}_n\rangle\right)^2\right]$$

den **Vorhersageverlust** (engl. *out-of-sample prediction loss*) und

$$R^{\text{pred}}(\widehat{\beta}_n, \beta) = \mathbb{E}_\beta[\ell^{\text{pred}}(\widehat{\beta}_n, \beta)] = \mathbb{E}_\beta\left[\left(Y_{n+1} - \langle X_{n+1}, \widehat{\beta}_n\rangle\right)^2\right]$$

den **Vorhersagefehler** (auch Vorhersagerisiko, *out-of-sample prediction error*).

Der Vorhersageverlust ist also bezüglich der Realisierung von $\widehat{\beta}_n$ weiterhin zufällig, während der Vorhersagefehler als dessen Erwartungswert deterministisch ist. Im weiteren Verlauf werden wir einen kleinen Vorhersageverlust mit großer Wahrscheinlichkeit nachweisen können, während sich die Abschätzung seines Erwartungswerts als technisch schwierig herausstellt und wenig zusätzliche Aussagekraft besitzt.

Lemma 2.26 *Für den Vorhersageverlust gilt*

$$\ell^{\text{pred}}(\widehat{\beta}_n, \beta) = \langle \Sigma_X(\widehat{\beta}_n - \beta), \widehat{\beta}_n - \beta\rangle + \sigma^2.$$

Beweis Mit $Y_{n+1} = \langle X_{n+1}, \beta\rangle + \varepsilon_{n+1}$ und $\mathbb{E}[X_{n+1}\varepsilon_{n+1}] = \mathbb{E}[X_{n+1}]\mathbb{E}[\varepsilon_{n+1}] = 0$ folgt

$$\begin{aligned}
\ell^{\text{pred}}(\widehat{\beta}_n, \beta) &= \mathbb{E}^{(X_{n+1}, \varepsilon_{n+1})}\left[\left(\langle X_{n+1}, \beta - \widehat{\beta}_n\rangle + \varepsilon_{n+1}\right)^2\right] \\
&= (\beta - \widehat{\beta}_n)^\top \mathbb{E}[X_{n+1} X_{n+1}^\top](\beta - \widehat{\beta}_n) + \mathbb{E}[\varepsilon_{n+1}^2] \\
&= (\beta - \widehat{\beta}_n)^\top \Sigma_X(\beta - \widehat{\beta}_n) + \sigma^2
\end{aligned}$$

und durch Umformulieren mit Skalarprodukt das Ergebnis. □

Da die Matrix Σ_X positiv semi-definit ist, ist der Vorhersageverlust immer mindestens σ^2, was für $\widehat{\beta}_n = \beta$ auch erreicht wird. Also selbst wenn wir β kennen,

machen wir noch einen Fehler bei der Vorhersage von Y_{n+1} auf Grund des stochastischen Fehlers in ε_{n+1}. Ein wichtiges Konzept des statistischen Lernens ist es, den nicht zugänglichen Vorhersagefehler, bei dem β und Σ_X nicht bekannt sind, durch einen empirischen Vorhersagefehler zu ersetzen und dann zu minimieren.

Definition 2.27 Im linearen Modell mit zufälligem Design bezeichnet

$$R_n(\widehat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle X_i, \widehat{\beta}_n\rangle\right)^2$$

das **empirische Risiko** (engl. *empirical risk, in-sample risk*) eines Schätzers $\widehat{\beta}_n$. Ein Schätzer $\widehat{\beta}_n$ mit

$$R_n(\widehat{\beta}_n) = \inf_{\widetilde{\beta}_n} R_n(\widetilde{\beta}_n),$$

wobei das Infimum über alle Schätzer $\widetilde{\beta}_n$, basierend auf $(X_i, Y_i)_{i=1,\dots,n}$, genommen wird, heißt **empirischer Risiko-Minimierer** (ERM, *empirical risk minimiser*).

Für deterministische $b \in \mathbb{R}^p$ sehen wir sofort, dass $\mathbb{E}_\beta[R_n(b)] = \ell^{\text{pred}}(b, \beta)$ gilt. Für den ERM $\widehat{\beta}_n$ statt b ist dies jedoch falsch.

Lemma 2.28 *Kleinste Quadrate-Schätzer $\widehat{\beta}_n$ sind gerade die empirischen Risiko-Minimierer, und es gilt*

$$\mathbb{E}_\beta[R_n(\widehat{\beta}_n)] = \tfrac{n-p}{n}\sigma^2,$$

sofern X fast sicher Rang p besitzt.

Beweis Beachte, dass $R_n(\widehat{\beta}_n) = \frac{1}{n}|Y - X\widehat{\beta}_n|^2$ gilt. Die rechte Seite wird vom Kleinste-Quadrate-Schätzer minimiert, die linke Seite vom empirischen Risikominimierer, so dass beide Konzepte übereinstimmen. Aus Lemma 2.18 folgt dann

$$\mathbb{E}[R_n(\widehat{\beta}_n)] = \tfrac{1}{n}\mathbb{E}\left[|Y - X\widehat{\beta}_n|^2\right] = \tfrac{n-p}{n}\sigma^2.$$

Wegen der Unabhängigkeit von (X_i) und (ε_i) nimmt man dafür zunächst den Erwartungswert über (ε_i) gemäß Lemma 2.18 und erhält $\frac{n-p}{n}\sigma^2$ fast sicher. Der Erwartungswert über (X_i) liefert dann den deterministischen Wert $\frac{n-p}{n}\sigma^2$. □

Eine einfache hinreichende Bedingung dafür, dass X vollen Rang p besitzt, ist, dass X_i gemäß einer Lebesguedichte im $\mathbb{R}^p$ verteilt ist und $n \geqslant p$ gilt. Das Argument beruht darauf, dass die Menge der $p \times p$-Matrizen A mit Determinante Null (A hat nicht vollen Rang) eine Lebesgue-Nullmenge im $\mathbb{R}^{p\times p}$ ist.

Lemmata 2.26 und 2.28 zeigen sofort einen wichtigen Unterschied zwischen empirischem Risiko und Vorhersagefehler des empirischen Risikominimierers auf:

$$\mathbb{E}_\beta[R_n(\widehat{\beta}_n)] = \tfrac{n-p}{n}\sigma^2 < \sigma^2 \leqslant R^{\text{pred}}(\widehat{\beta}_n, \beta)$$

Da wir beim empirischen Risiko dieselben Daten für die Fehlerquantifizierung benutzen, die auch für den Schätzer verwendet wurden, ist der Fehler innerhalb der

Stichprobe stets kleiner als bei der Vorhersage außerhalb der Stichprobe. Wir wollen nun verstehen, inwiefern trotzdem die empirische Risikominimierung zu guten Schätzern bezüglich Vorhersagefehler führen, sofern der Stichprobenumfang n hinreichend groß ist. Man sagt auch, dass der Schätzer einen kleinen *Verallgemeinerungsfehler* (engl. *generalisation error*) besitzt, eine wichtige Frage auch später bei den Methoden des maschinellen Lernens.

Satz 2.29 *Mit der* ***empirischen Design-Kovarianzmatrix***

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top = \frac{1}{n} X^\top X$$

gilt für den Vorhersageverlust und den Vorhersagefehler des Kleinste-Quadrate-Schätzers $\widehat{\beta}_n$

$$\begin{aligned}\mathbb{E}^\varepsilon\left[\ell^{\mathrm{pred}}(\widehat{\beta}_n, \beta)\right] &= \tfrac{\sigma^2}{n}\operatorname{tr}(\Sigma_X \Sigma_n^{-1}) + \sigma^2,\\ R^{\mathrm{pred}}(\widehat{\beta}_n, \beta) &= \tfrac{\sigma^2}{n}\mathbb{E}[\operatorname{tr}(\Sigma_X \Sigma_n^{-1})] + \sigma^2,\end{aligned}$$

wobei $\mathbb{E}^\varepsilon$ *bedeutet, dass der Erwartungswert nur bezüglich* $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ *genommen wird.*

Beweis Es gilt $\widehat{\beta}_n - \beta = (X^\top X)^{-1} X^\top \varepsilon = \Sigma_n^{-1} \frac{1}{n} X^\top \varepsilon$, und wir erhalten mit Lemma 2.26 und $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 E_n$

$$\begin{aligned}\mathbb{E}^\varepsilon\left[\ell^{\mathrm{pred}}(\widehat{\beta}_n, \beta)\right] &= \tfrac{1}{n^2}\mathbb{E}^\varepsilon\left[\langle \Sigma_X \Sigma_n^{-1} X^\top \varepsilon, \Sigma_n^{-1} X^\top \varepsilon\rangle\right] + \sigma^2\\ &= \tfrac{1}{n^2}\operatorname{tr}\left(\Sigma_X \Sigma_n^{-1} X^\top \mathbb{E}[\varepsilon\varepsilon^\top] X \Sigma_n^{-1}\right) + \sigma^2\\ &= \tfrac{\sigma^2}{n}\operatorname{tr}\left(\Sigma_X \Sigma_n^{-1} \Sigma_n \Sigma_n^{-1}\right) + \sigma^2,\end{aligned}$$

woraus die erste Behauptung folgt. Für den Vorhersagefehler verwenden wir dann $R^{\mathrm{pred}}(\widehat{\beta}_n, \beta) = \mathbb{E}^X[\mathbb{E}^\varepsilon[\ell^{\mathrm{pred}}(\widehat{\beta}_n, \beta)]]$ wegen der Unabhängigkeit der X_i und ε_i. □

Im einfachen Fall eines Gaußschen Designs $X_i \sim N(0, \Sigma_X)$ kann man mit Hilfe der sogenannten inversen Wishart-Verteilung zeigen, dass $\mathbb{E}[\Sigma_n^{-1}] = \frac{n}{n-p-1}\Sigma_X^{-1}$ für $n \geqslant p+2$ gilt, so dass $R^{\mathrm{pred}}(\widehat{\beta}_n, \beta) = \frac{\sigma^2 p}{n-p-1} + \sigma^2$ folgt. Im Allgemeinen ist nicht einmal klar, dass Σ_n^{-1} einen Erwartungswert besitzt. Die Theorie der zufälligen Matrizen liefert immerhin, dass Σ_n nahe an Σ_X ist, wenn n groß ist und große Werte der Kovariablen X_i ähnlich unwahrscheinlich sind wie unter einer Normalverteilung.

Definition 2.30 Ein Zufallsvektor X im $\mathbb{R}^p$ heißt (Σ, C)**-subgaußsch**, falls $\mathbb{P}(|\langle X, u\rangle| > t) \leqslant C e^{-t^2/(2\langle \Sigma u, u\rangle)}$ für alle $t > 0$, $u \in \mathbb{R}^p \setminus \{0\}$ gilt mit einer Konstanten $C > 0$ und einer positiv-definiten Matrix $\Sigma \in \mathbb{R}^{p\times p}$.

Insbesondere ist jeder normalverteilte Zufallsvektor $X \sim \mathrm{N}(\mu, \Sigma)$ (Σ, C)-subgaußsch mit geeignetem $C > 0$, aber auch jeder Zufallsvektor, dessen Verteilung einen kompakten Träger besitzt. Ist mit der positiv-definiten Kovarianzmatrix Σ_X

von X der standardisierte Zufallsvektor $\Sigma_X^{-1/2} X$ (cE_p, C)-subgaußsch für ein $c > 0$, so ist X $(c\Sigma_X, C)$-subgaußsch, und umgekehrt. In diesem Sinn interpretieren wir im Folgenden Σ häufig als Vielfaches der Design-Kovarianzmatrix Σ_X. Ebenso wie normalverteilte Zufallsvektoren besitzen $(c\Sigma_X, C)$-subgaußsche Zufallsvektoren X endliche Momente, genauer gilt

$$\mathbb{E}[|\Sigma_X^{-1/2} X|^m]^{1/m} \leqslant \widetilde{C}_m p^{1/2}, \quad \mathbb{E}[|\langle \Sigma_X^{-1/2} X, v\rangle|^m]^{1/m} \leqslant \widetilde{C}_m |v|, \quad m \geqslant 1, \quad (2.4)$$

für deterministische $v \in \mathbb{R}^p$, vorausgesetzt Σ_X ist invertierbar. Die Konstanten $\widetilde{C}_m$ hängen nur von c, C und m ab.

Um die Abweichung der empirischen Design-Kovarianzmatrix Σ_n von der Design-Kovarianzmatrix $\Sigma_x = \mathbb{E}[\Sigma_n]$ zu messen, stellt sich die *Spektralnorm* als geeignete Metrik heraus. Für eine Matrix M ist sie definiert als

$$\|M\| = \sup_{|v|=1} |Mv|.$$

Insbesondere gilt dann $|Mu| \leqslant \|M\||u|$ für alle Vektoren u. Die Spektralnorm ist beschränkt durch die Frobeniusnorm, für die wir in Übung 2.11 $\mathbb{E}[\|\Sigma_n - \Sigma_X\|_2^2]^{1/2} = O(p/\sqrt{n})$ nachweisen. Ein erstaunliches Ergebnis der Theorie zufälliger Matrizen mit vielen Anwendungen in der hochdimensionalen Statistik ist, dass $\|\Sigma_n - \Sigma_X\|$ unter milden Voraussetzungen von der stochastischen Ordnung $\sqrt{p/n}$ ist. Insbesondere ist also $\|\Sigma_n - \Sigma_X\|$ bereits klein, sobald die Dimension p von kleinerer Größenordnung als der Stichprobenumfang n ist. Wir passen Satz 4.7.1 von Vershynin (2018) auf unsere Situation an, siehe auch Satz 6.5 in Wainwright (2019).

Satz 2.31 (Fehler der empirischen Kovarianzmatrix) *Sind $X_1, \dots, X_n$ i.i.d. $(c\Sigma_X, C)$-subgaußsche Zufallsvektoren in $\mathbb{R}^p$ mit $C, c > 0$ und invertierbarer Kovarianzmatrix $\Sigma_X \in \mathbb{R}^{p\times p}$, so gilt für $p \leqslant n$*

$$\mathbb{E}\big[\|\Sigma_X^{-1/2} \Sigma_n \Sigma_X^{-1/2} - E_p\|\big] \leqslant \widetilde{C}\sqrt{p/n}$$

mit einer Konstanten $\widetilde{C}$ (abhängig von c, C).

Verwenden wir den $O_{\mathbb{P}}$-Kalkül aus Definition A.41, so gilt also

$$\|\Sigma_X^{-1/2} \Sigma_n \Sigma_X^{-1/2} - E_p\| = O_{\mathbb{P}}\big(\sqrt{p/n}\big).$$

Schreiben wir $A := \Sigma_n^{1/2} \Sigma_X^{-1/2}$, so erlaubt es dieses Resultat, weitere durch A definierte Matrizen abzuschätzen. Dazu nehmen wir $\|A^\top A - E_p\| \leqslant \varepsilon$ mit $\varepsilon \in (0, 1)$ an. Die variationelle Charakterisierung von Eigenwerten symmetrischer Matrizen zeigt dann

$$\sup_{|v|=1} ||Av|^2 - 1| = \sup_{|v|=1} |\langle (A^\top A - E_p)v, v\rangle| = \|A^\top A - E_p\| \leqslant \varepsilon.$$

Direkt folgt also $\|A\| \leqslant \sqrt{1+\varepsilon} \leqslant 1 + \varepsilon/2$. Darüberhinaus ist $\inf_{|v|=1}|Av| \geqslant \sqrt{1-\varepsilon} > 0$, so dass A invertierbar ist mit (setze $w = A^{-1}v$)

$$\|A^{-1}\| = \sup_{v\neq 0} \frac{|A^{-1}v|}{|v|} = \sup_{w\neq 0} \frac{|w|}{|Aw|} \leqslant \frac{1}{\sqrt{1-\varepsilon}} \leqslant 1 + \frac{\varepsilon}{1-\varepsilon}.$$

Mit denselben Argumenten schließen wir auf

$$\|(A^{-1})^\top A^{-1} - E_p\| = \sup_{|v|=1} ||A^{-1}v|^2 - 1| \leqslant \max\Big(\frac{1}{1-\varepsilon} - 1, 1 - \frac{1}{1+\varepsilon}\Big) = \frac{\varepsilon}{1-\varepsilon}.$$

Aus Satz 2.31 folgen daher durch Einsetzen und unter Verwendung von $\|A^\top\| = \|A\|$ die weiteren asymptotischen Schranken unter der Asymptotik $p/n \to 0$

$$\|\Sigma_n^{1/2}\Sigma_X^{-1/2}\| \leqslant 1 + O_{\mathbb{P}}\big(\sqrt{p/n}\big),\ \|\Sigma_X^{1/2}\Sigma_n^{-1/2}\| \leqslant 1 + O_{\mathbb{P}}\big(\sqrt{p/n}\big), \tag{2.5}$$

$$\|\Sigma_n^{-1/2}\Sigma_X\Sigma_n^{-1/2} - E_p\| = O_{\mathbb{P}}\big(\sqrt{p/n}\big),\ \|\Sigma_X^{1/2}\Sigma_n^{-1}\Sigma_X^{1/2} - E_p\| = O_{\mathbb{P}}\big(\sqrt{p/n}\big). \tag{2.6}$$

Insbesondere heißt das, dass $\mathbb{P}(\Sigma_n^{-1}$ existiert$) \to 1$ gilt. Wir können nun immerhin den Vorhersageverlust des Kleinste-Quadrate-Schätzers in stochastischer Ordnung unter allgemeinen Bedingungen abschätzen. Dies ist für viele Zwecke vollkommen ausreichend und vermeidet zusätzliche Bedingungen, um die Existenz von $\mathbb{E}[\Sigma_n^{-1}]$ sicherzustellen.

Korollar 2.32 *Die Kovariablen X_i seien $(c\Sigma_X, C)$-subgaußsche Zufallsvektoren im $\mathbb{R}^p$ für Konstanten $c, C > 0$. Für den Vorhersageverlust des Kleinste-Quadrate-Schätzers $\widehat{\beta}_n$ gilt*

$$\ell^{\text{pred}}(\widehat{\beta}_n, \beta) = \frac{\sigma^2 p}{n}\Big(1 + O_{\mathbb{P}}\big(\sqrt{p/n}\big)\Big) + \sigma^2,$$

sofern $p \leqslant n$ und Σ_X invertierbar ist.

Beweis Aus Satz 2.29 folgt mit $\operatorname{tr}(E_p) = p$ und Kommutativität unter der Spur

$$\mathbb{E}^\varepsilon\big[\ell^{\text{pred}}(\widehat{\beta}_n, \beta)\big] = \tfrac{\sigma^2}{n}\Big(p + \operatorname{tr}(\Sigma_n^{-1/2}\Sigma_X\Sigma_n^{-1/2} - E_p)\Big) + \sigma^2.$$

Wir verwenden $|\operatorname{tr}(M)| \leqslant p\|M\|$ für $p \times p$-Matrizen M sowie (2.6) und schließen

$$|\operatorname{tr}(\Sigma_n^{-1/2}\Sigma_X\Sigma_n^{-1/2} - E_p)| = O_{\mathbb{P}}\big(p\sqrt{p/n}\big).$$

Also gilt

$$\mathbb{E}^\varepsilon\big[\ell^{\text{pred}}(\widehat{\beta}_n, \beta)\big] = \tfrac{\sigma^2 p}{n}\Big(1 + O_{\mathbb{P}}\big(\sqrt{p/n}\big)\Big) + \sigma^2.$$

Da ℓ^{pred} nicht-negativ ist, impliziert die Konvergenzrate des Erwartungswerts bezüglich ε dieselbe $O_{\mathbb{P}}$-Konvergenzrate und damit die Behauptung. □

Wir sehen, dass der zusätzliche Vorhersageverlust, neben dem unvermeidlichen Fehler σ^2, mit Rate $\sigma^2 p/n$ gegen Null geht selbst bei wachsender Dimension p, falls nur $p/n \to 0$ gilt.

Schließlich wollen wir noch den allgemeineren Zugang des statistischen Lernens kennenlernen. Der Hauptunterschied ist, dass die Daten mit einem sehr allgemeinen Modell beschrieben werden, für die aber nur eine Teilklasse von statistischen Methoden untersucht wird. Im vorliegenden Vorhersageproblem wird nurmehr angenommen, dass $(X_i, Y_i)_{i=1,\dots,n+1}$ i.i.d. sind mit $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$. Wir betrachten die Klasse aller linearen Vorhersagen $f_b(X_{n+1}) = \langle X_{n+1}, b\rangle$ von Y_{n+1} mit $b \in \mathbb{R}^p$. Der Vorhersageverlust ist dann

$$\ell^{\text{pred}}(b) = \mathbb{E}^{(X_{n+1},Y_{n+1})}[(Y_{n+1} - f_b(X_{n+1}))^2],$$

was mit der vorherigen Definition von $\ell^{\text{pred}}(b, \beta)$ übereinstimmt in dem Fall, dass wirklich das lineare Modell $Y_{n+1} = \langle X_{n+1}, \beta\rangle + \varepsilon_{n+1}$ gilt. Im Allgemeinen gilt dies jedoch nicht, und wir vergleichen die Qualität eines Schätzers mit der besten Vorhersage innerhalb der linearen Klasse. Wir setzen daher

$$\beta^* := \operatorname*{arg\,min}_{b\in\mathbb{R}^p} \ell^{\text{pred}}(b)$$

und nennen β^* das *Orakel* oder den *Orakelparameter*. Da wir mit linearen Vorhersagen keinen geringeren Verlust als $\ell^{\text{pred}}(f_{\beta^*})$ erzielen können, betrachten wir für einen Schätzer $\widehat{\beta}$ den *Exzessverlust*

$$\mathcal{E}^{\text{pred}}(\widehat{\beta}) := \ell^{\text{pred}}(\widehat{\beta}) - \ell^{\text{pred}}(\beta^*) = \ell^{\text{pred}}(\widehat{\beta}) - \inf_{b\in\mathbb{R}^p} \ell^{\text{pred}}(b).$$

Mit dem empirischen Risiko von oben gilt weiterhin $\mathbb{E}[R_n(b)] = \ell^{\text{pred}}(b)$ für deterministische $b \in \mathbb{R}^p$, und der Kleinste-Quadrate-Schätzer $\widehat{\beta}_n$ ist empirischer Risikominimierer, da die Berechnung ja unabhängig vom datenerzeugenden Modell ist. Ziel ist es nun, das Orakel β^* und die Größenordnung des Exzessrisikos von $\widehat{\beta}_n$ im allgemeinen Modell zu verstehen.

Satz 2.33 *Mit dem Kovarianzvektor $C_{XY} := \mathbb{E}[X_1 Y_1]$ gilt für den Orakelparameter $\beta^* = \Sigma_X^{-1} C_{XY}$. Der Exzessverlust eines Schätzers $\widehat{\beta}$ ist gegeben durch*

$$\mathcal{E}^{\text{pred}}(\widehat{\beta}) = \langle \Sigma_X(\widehat{\beta} - \beta^*), \widehat{\beta} - \beta^*\rangle, \tag{2.7}$$

wobei Σ_X^{-1} existieren möge. Sind ferner die Kovariablen X_i $(c\Sigma_X, C)$-subgaußsch, gilt $\mathbb{E}[Y_1^4] < \infty$ und ist $n \geqslant p$, so gilt für den Kleinste-Quadrate-Schätzer $\widehat{\beta}_n$

$$\mathcal{E}^{\text{pred}}(\widehat{\beta}_n) = O_{\mathbb{P}}(p/n),$$

wobei die Konstante nur von c, C und $\mathbb{E}[Y_1^4]$ abhängt.

Beweis Durch Einsetzen und quadratische Ergänzung erhalten wir

$$\ell^{\text{pred}}(b) = \mathbb{E}[Y_{n+1}^2] - 2\mathbb{E}[Y_{n+1}\langle X_{n+1}, b\rangle] + \mathbb{E}[\langle X_{n+1}, b\rangle^2]$$

$$= \mathbb{E}[Y_1^2] - \langle C_{XY}, b\rangle + \langle \Sigma_X b, b\rangle$$
$$= \langle \Sigma_X(b - \Sigma_X^{-1}C_{XY}), b - \Sigma_X^{-1}C_{XY}\rangle + \mathbb{E}[Y_1^2] - \langle \Sigma_X^{-1}C_{XY}, C_{XY}\rangle.$$

Dies wird offensichtlich durch $\beta^* = \Sigma_X^{-1}C_{XY}$ in b minimiert. Weiterhin folgt aus der Darstellung sofort (2.7) für den Exzessverlust.

Zur asymptotischen Analyse von $\widehat{\beta}_n = \Sigma_n^{-1}C_n$ mit

$$C_n = \frac{1}{n}X^\top Y = \frac{1}{n}\sum_{i=1}^n X_i Y_i$$

wollen wir zunächst Erwartungswert und Varianz von $\Sigma_X^{-1/2}C_n$ abschätzen. Es gilt $\mathbb{E}[C_n] = C_{XY}$ sowie mit der Cauchy-Schwarz-Ungleichung

$$\begin{aligned}
|\Sigma_X^{-1/2}C_{XY}|^2 &= \sup_{|v|=1} \langle \mathbb{E}[\Sigma_X^{-1/2}X_1Y_1], v\rangle^2 \\
&= \sup_{|v|=1} \mathbb{E}[\langle \Sigma_X^{-1/2}X_1, v\rangle Y_1]^2 \\
&\leqslant \sup_{|v|=1} \mathbb{E}[\langle \Sigma_X^{-1/2}X_1, v\rangle^2]\mathbb{E}[Y_1^2] \\
&\leqslant \widetilde{C}_2^2\mathbb{E}[Y_1^2]
\end{aligned}$$

mit $\widetilde{C}_2$ aus (2.4). Wegen X_iY_i i.i.d. folgt ferner für die Varianz des Zufallsvektors $\Sigma_X^{-1/2}C_n$ aus der Cauchy-Schwarz-Ungleichung

$$\begin{aligned}
\operatorname{Var}(\Sigma_X^{-1/2}C_n) &= \frac{1}{n^2}\sum_{i=1}^n \operatorname{Var}(\Sigma_X^{-1/2}X_iY_i) \leqslant \frac{1}{n}\mathbb{E}[|\Sigma_X^{-1/2}X_1Y_1|^2] \\
&\leqslant \frac{1}{n}\mathbb{E}[|\Sigma_X^{-1/2}X_1|^4]^{1/2}\mathbb{E}[Y_1^4]^{1/2}.
\end{aligned}$$

Unter Verwendung von $\widetilde{C}_4$ aus (2.4) und $\mathbb{E}[Y_1^2]^2 \leqslant \mathbb{E}[Y_1^4]$ implizieren beide Abschätzungen die stochastischen Ordnungen

$$|\Sigma_X^{-1/2}(C_n - C_{XY})| = O_{\mathbb{P}}(\sqrt{p/n}), \quad |\Sigma_X^{-1/2}C_n| = O_{\mathbb{P}}(1)$$

mit einem Faktor in Abängigkeit von $c, C, \mathbb{E}[Y_1^4]$. Einsetzen in (2.7) und Verwenden von (2.6) ergibt damit

$$\begin{aligned}
\mathcal{E}^{\text{pred}}(\widehat{\beta}_n) &= |\Sigma_X^{1/2}(\widehat{\beta}_n - \beta^*)|^2 \\
&= |\Sigma_X^{1/2}\Sigma_n^{-1}C_n - \Sigma_X^{-1/2}C_{XY}|^2 \\
&\leqslant \left(\|\Sigma_X^{1/2}\Sigma_n^{-1}\Sigma_X^{1/2} - E_p\|\,|\Sigma_X^{-1/2}C_n| + |\Sigma_X^{-1/2}(C_n - C_{XY})|\right)^2 \\
&= \left(O_{\mathbb{P}}(\sqrt{p/n})O_{\mathbb{P}}(1) + O_{\mathbb{P}}(\sqrt{p/n})\right)^2 = O_{\mathbb{P}}(p/n),
\end{aligned}$$

wie behauptet. □

Wenn wir mit Korollar 2.32 im korrekt spezifizierten linearen Modell vergleichen, so sehen wir, dass in beiden Fällen die Rate p/n für den Exzessverlust vorliegt (beachte $\inf_b \ell^{\text{pred}}(b, \beta) = \sigma^2$). Allerdings ist dort $\sigma^2 p/n$ die exakte Form des Terms größter Ordnung, während der Faktor im allgemeinen Fall unbestimmt bleibt. Trotzdem garantiert dieses Resultat eine Robustheit gegenüber Modellmisspezifikation, also falls die Annahme eines linearen Modellzusammenhangs verletzt ist.

2.2 Inferenz unter Normalverteilungsannahme

Statistische Inferenz umfasst die Konstruktion von Tests und Konfidenzintervallen. Im Gegensatz zur Schätztheorie aus dem letzten Abschnitt, benötigen wir hier eine explizite Annahme an die Verteilung des Fehlervektors ε im linearen Modell. Im Folgenden werden wir stets das gewöhnliche lineare Modell unter der Normalverteilungsannahme $(\varepsilon_i) \sim \mathrm{N}(0, \sigma^2 E_n)$ betrachten, da die Abweichungen der (Mess-)Werte vieler natur-, wirtschafts- und ingenieurswissenschaftlicher Vorgänge durch die Normalverteilung in guter Näherung beschrieben werden können. Zusätzlich beschränken wir uns auf den Fall, dass die Designmatrix X deterministisch ist und vom Rang $p < n$.

Ist die Fehlervarianz σ^2 bekannt, lassen sich unter der Normalverteilungsannahme leicht Konfidenzintervalle konstruieren:

Beispiel 2.34 Sind die Messfehler $(\varepsilon_i) \sim \mathrm{N}(0, \sigma^2 E_n)$ gemeinsam normalverteilt und $\rho = \langle v, \beta \rangle$ für $v \in \mathbb{R}^k$, so gilt

$$\widehat{\beta} \sim \mathrm{N}(\beta, \sigma^2 (X^\top X)^{-1}) \quad \text{und} \quad \widehat{\rho} = \langle v, \widehat{\beta} \rangle \sim \mathrm{N}(\rho, \sigma^2 \langle (X^\top X)^{-1} v, v \rangle).$$

Ist $\sigma > 0$ bekannt, so ist ein Konfidenzintervall zum Niveau $1 - \alpha$ für ρ gegeben durch

$$\left[\widehat{\rho} - q_{1-\alpha/2} \sigma \sqrt{\langle (X^\top X)^{-1} v, v \rangle}, \widehat{\rho} + q_{1-\alpha/2} \sigma \sqrt{\langle (X^\top X)^{-1} v, v \rangle}\right],$$

mit dem $(1 - \alpha/2)$-Quantil $q_{1-\alpha/2}$ der Standardnormalverteilung. Beachte, dass dieses Konfidenzintervall über den Korrespondenzsatz 1.70 dem zweiseitigen Gauß-Test aus Beispiel 1.65 entspricht. Insbesondere ergibt sich für $\alpha = 0{,}05$ der Wert $q_{0{,}975} \approx 1{,}96$, der in Anwendungen häufig auftaucht.

Die Annahme in diesem Beispiel, dass σ bekannt sei, ist in den wenigstens Fällen erfüllt. Bei unbekanntem Rauschniveau können wir σ durch den Schätzer $\widehat{\sigma}$ ersetzen. Ist $\widehat{\sigma}$ konsistent, dann folgt aus Slutzkys Lemma, dass das resultierende Konfidenzintervall beziehungsweise der entsprechende Test das vorgegebene Niveau zumindest asymptotisch erreicht. Im nächsten Abschnitt werden wir sehen, dass man sogar die Verteilung der normalisierten Statistiken explizit bestimmen und so Konfidenzbereiche und Tests konstruieren kann, die auch für endliche Stichprobengrößen ein gegebenes Niveau genau erreichen.

Folgende Verteilungen sind essentiell für die angestrebte Inferenz und bilden die Grundlage für die vielfach genutzten t- und F-Tests.

Definition 2.35 Die **t-Verteilung** $\mathrm{t}(n)$ (oder Student-t-Verteilung) mit $n \in \mathbb{N}$ Freiheitsgraden auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ist gegeben durch die Lebesgue-Dichte

$$t_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi n}}\Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \quad x \in \mathbb{R}.$$

Dabei bezeichnet $\Gamma(p) = \int_0^\infty t^{p-1}e^{-t}\mathrm{d}t$ die Gammafunktion.

Definition 2.36 Die **F-Verteilung** $\mathrm{F}(m, n)$ (oder Fisher-Verteilung) mit $(m, n) \in \mathbb{N}^2$ Freiheitsgraden auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ist gegeben durch die Lebesgue-Dichte

$$f_{m,n}(x) = \frac{m^{m/2}n^{n/2}}{B(\frac{m}{2}, \frac{n}{2})}\frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}}\mathbb{1}_{\mathbb{R}^+}(x), \quad x \in \mathbb{R}.$$

Dabei bezeichnet $B(p, q) = \dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ die Betafunktion.

Lemma 2.37 *Es seien $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ unabhängige $\mathrm{N}(0,1)$-verteilte Zufallsvariablen. Dann gilt*

$$T_n := \frac{X_1}{\sqrt{\frac{1}{n}\sum_{j=1}^n Y_j^2}} \sim \mathrm{t}(n) \quad \textit{und} \quad F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^m X_i^2}{\frac{1}{n}\sum_{j=1}^n Y_j^2} \sim \mathrm{F}(m, n).$$

Beweis Es gilt $T_n^2 = F_{1,n}$, sodass mittels Dichtetransformation $f^{|T_n|}(x) = f^{F_{1,n}}(x^2)2x$, $x \geqslant 0$, folgt. Da T_n symmetrisch (wie $-T_n$) verteilt ist, folgt $f^{T_n}(x) = f^{F_{1,n}}(x^2)|x|, x \in \mathbb{R}$, und Einsetzen zeigt die Behauptung für T_n, sofern $F_{1,n}$ $\mathrm{F}(1, n)$-verteilt ist.

Um die Behauptung für $F_{m,n}$ nachzuweisen, benutzen wir, dass $X := \sum_{i=1}^m X_i^2$ bzw. $Y := \sum_{j=1}^n Y_j^2$ gemäß $\chi^2(m)$ bzw. $\chi^2(n)$ verteilt sind. Wegen der Unabhängigkeit von X und Y und des Satzes von Fubini gilt für $z > 0$ (setze $w = x/y$)

$$\begin{aligned}\mathbb{P}(X/Y \leqslant z) &= \int_0^\infty \int_0^\infty \mathbb{1}_{\{x/y \leqslant z\}} f^X(x) f^Y(y)\mathrm{d}x\mathrm{d}y \\ &= \int_0^z \Big(\int_0^\infty f^X(wy) f^Y(y) y \mathrm{d}y \Big)\mathrm{d}w.\end{aligned}$$

Setzen wir die χ^2-Dichten ein und substituieren $w = (z+1)y$ und $t = w/2$, ergibt sich die Dichte von X/Y:

$$\begin{aligned}f^{X/Y}(z) &= \int_0^\infty f^X(zy) f^Y(y) y \mathrm{d}y \\ &= \frac{2^{-(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty (zy)^{m/2-1} y^{n/2} e^{-(zy+y)/2}\mathrm{d}y\end{aligned}$$

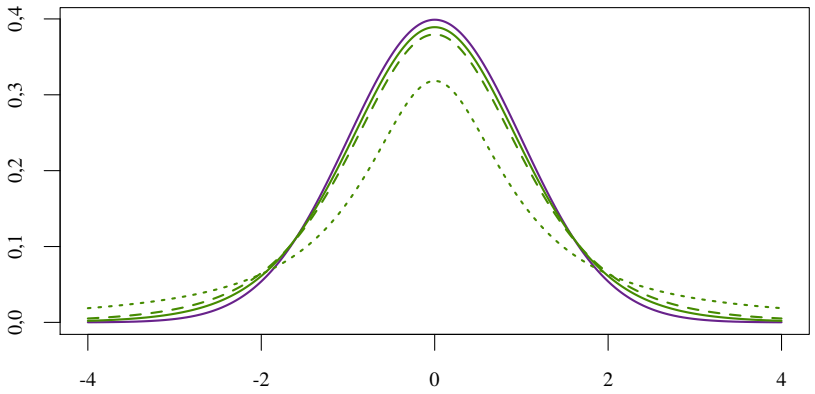

Abb. 2.5 Für $n \to \infty$ konvergiert die Dichte der $\mathrm{t}(n)$-Verteilung (grün) gegen jene der Standardnormalverteilung (violett). Gepunktet entspricht $\mathrm{t}(1)$, gestrichelt $\mathrm{t}(5)$ und durchgezogen $\mathrm{t}(10)$.

$$
\begin{aligned}
&= \frac{2^{-(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty (zw/(z+1))^{m/2-1} (w/(z+1))^{n/2} e^{-w/2} (z+1)^{-1} \mathrm{d}w \\
&= \frac{z^{m/2-1}(z+1)^{-(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty 2^{-((m+n)/2-1)} 2^{-1} w^{(m+n)/2-1} e^{-w/2} \mathrm{d}w \\
&= \frac{z^{m/2-1}(z+1)^{-(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty t^{((m+n)/2-1)} e^t \mathrm{d}t \\
&= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} z^{m/2-1}(z+1)^{-(m+n)/2}, \quad z > 0.
\end{aligned}
$$

Dichtetransformation ergibt damit für $F_{m,n} = \frac{n}{m}\frac{X}{Y}$ die Dichte $\frac{m}{n} f^{X/Y}(\frac{m}{n}x) = f_{m,n}(x)$. □

Bemerkung 2.38 (t- und F-Verteilung) Die t-Verteilung ist symmetrisch zur Symmetrieachse $x = 0$ und glockenförmig. Die Dichte fällt jedoch nur polynomiell statt exponentiell schnell ab. Insbesondere besitzt die $\mathrm{t}(n)$-Verteilung für jedes $n \in \mathbb{N}$ nur Momente bis zur Ordnung $p < n$. Man spricht von *heavy tails*, was man mit „schweren Flanken“ übersetzen könnte. Für $n = 1$ ist die $\mathrm{t}(n)$-Verteilung gerade die Cauchy-Verteilung, und für $n \to \infty$ konvergiert sie schwach gegen die Standardnormalverteilung, was man leicht aus dem letzten Lemma folgern kann, siehe auch Abbildung 2.5. Die F-Verteilung hat ebenfalls heavy tails. Es gilt $F_{1,n} = T_n^2$. Für $n \to \infty$ konvergiert $mF_{m,n}$ gegen die $\chi^2(m)$-Verteilung.

Folgendes Hilfsresultat zur Verteilung von quadratischen Formen wird uns auch bei der Konstruktion von Tests und Konfidenzbändern im linearen Modell helfen:

Lemma 2.39 *Sei* $Z \sim \mathrm{N}(0, E_n)$ *und sei* $R \in \mathbb{R}^{n\times n}$ *eine Orthogonalprojektion vom Rang* $\mathrm{rank}(R) = r \leqslant n$. *Dann gilt*

(i) $Z^\top RZ \sim \chi^2(r)$,
(ii) $Z^\top RZ$ *ist unabhängig von* BZ *für jede Matrix* $B \in \mathbb{R}^{p\times n}$ *mit* $BR = 0$,
(iii) für jede weitere Orthogonalprojektion $S \in \mathbb{R}^{n\times n}$ *mit* $\operatorname{rank}(S) = s \leqslant n$ *und* $RS = 0$ *sind* $Z^\top RZ$ *und* $Z^\top SZ$ *unabhängig, und es gilt*

$$\frac{s}{r}\frac{Z^\top RZ}{Z^\top SZ} \sim \mathrm{F}(r,s).$$

Beweis *(i)* Als Orthogonalprojektion ist R symmetrisch und idempotent ($R = R^\top$ und $R^2 = R$). Daher existiert eine Orthogonalmatrix T mit

$$R = TD_rT^\top \text{ und } D_r = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Da T orthogonal ist und Z standardnormalverteilt, folgt $W := T^\top Z \sim \mathrm{N}(0, E_n)$. Wegen

$$Z^\top RZ = Z^\top(TD_rT^\top)Z = (T^\top Z)^\top D_r(T^\top Z) = W^\top D_r W = \sum_{i=1}^{r} W_i^2$$

ist $Z^\top RZ$ $\chi^2(r)$-verteilt.

(ii) Wir setzen $Y := BZ \sim \mathrm{N}(0, B^\top B)$ und $V := RZ \sim \mathrm{N}(0, R)$. Dann gilt

$$\mathbb{C}\mathrm{ov}(Y,V) = B\,\mathbb{C}\mathrm{ov}(Z)R^\top = BR = 0.$$

Weiter haben wir $Z^\top RZ = Z^\top R^2 Z = (RZ)^\top(RZ) = V^\top V$. Da (Y,V) als Lineartransformation von Z gemeinsam normalverteilt ist, folgt aus der Unkorreliertheit bereits die Unabhängigkeit von Y und V und somit auch die von $Y = BZ$ und $V^\top V = Z^\top RZ$.

(iii) Genau wie in (ii) folgt die Unabhängigkeit von $Y := SZ$ und $V := RZ$ und somit auch die Unabhängigkeit von $Y^\top Y = Z^\top SZ$ und $V^\top V = Z^\top RZ$. Zusammen mit (i) und dem vorangegangenen Lemma folgt die Behauptung. □

Als Folgerung erhalten wir Tests und Konfidenzbereiche für die Schätzung von β und linearen Funktionalen im gewöhnlichen linearen Modell unter der Normalverteilungsannahme:

Satz 2.40 (F-Test) *Betrachte im gewöhnlichen linearen Modell unter der Normalverteilungsannahme* $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ *mit unbekanntem* $\sigma > 0$ *das Testproblem* $H_0\colon \beta = \beta_0, \sigma > 0$ *gegen* $H_1\colon \beta \neq \beta_0, \sigma > 0$. *Dann ist der zweiseitige* ***F-Test*** *(auch* ***Fisher-Test****)*

$$\varphi_\alpha = \mathbb{1}(F > q_{\mathrm{F}(p,n-p);1-\alpha}) \quad \textit{mit} \qquad F = \frac{\frac{1}{p}|X(\widehat{\beta} - \beta_0)|^2}{\widehat{\sigma}^2},$$

dem Kleinste-Quadrate-Schätzer $\widehat{\beta}$, der empirischen Stichprobenvarianz $\widehat{\sigma}^2 = \frac{1}{n-p}|Y - X\widehat{\beta}|^2$ und dem $(1-\alpha)$-Quantil $q_{\mathrm{F}(p,n-p);1-\alpha}$ der $\mathrm{F}(p, n-p)$-Verteilung ein Likelihood-Quotiententest zum Niveau $\alpha \in (0,1)$.

Beweis Da der Kleinste-Quadrate-Schätzer gerade der Maximum-Likelihood-Schätzer unter Normalverteilung ist, können wir die Likelihood für alle $\sigma^2 > 0$ über

$$\sup_{\beta \in \mathbb{R}^p \setminus \{\beta_0\}} L(\sigma^2, \beta) = L(\sigma^2, \widehat{\beta})$$

maximieren. Wir erhalten für den Likelihood-Quotienten (Methode 1.63)

$$T = \frac{\sup_{\sigma^2>0} L(\widehat{\beta}, \sigma^2)}{\sup_{\sigma^2>0} L(\beta_0, \sigma^2)} = \frac{\sup_{\sigma^2>0} \sigma^{-n} \exp(-\frac{1}{2\sigma^2}|Y - X\widehat{\beta}|^2)}{\sup_{\sigma^2>0} \sigma^{-n} \exp(-\frac{1}{2\sigma^2}|Y - X\beta_0|^2)}.$$

Wegen

$$\frac{d}{d\sigma^2}\Big(-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}|Y - X\widehat{\beta}|^2\Big) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}|Y - X\widehat{\beta}|^2,$$

wird das Supremum im Zähler von T bei $\widehat{\sigma}^2_{\mathrm{MLE},1} = \frac{1}{n}|Y - X\widehat{\beta}|^2$ angenommen. Analog wird das Supremum im Nenner von T bei $\widehat{\sigma}^2_{\mathrm{MLE},0} = \frac{1}{n}|Y - X\beta_0|^2$ angenommen. Mit $X\widehat{\beta} = \Pi_X Y$ und der Orthogonalität $(E_n - \Pi_X)\Pi_X = 0$ erhalten wir

$$\begin{aligned} T = \frac{\widehat{\sigma}^{-n}_{\mathrm{MLE},1}}{\widehat{\sigma}^{-n}_{\mathrm{MLE},0}} &= \Big(\frac{|Y - X\beta_0|^2}{|Y - \Pi_X Y|^2}\Big)^{n/2} \\ &= \Big(\frac{|(E_n - \Pi_X)Y|^2 + |\Pi_X(Y - X\beta_0)|^2}{|Y - \Pi_X Y|^2}\Big)^{n/2} = \Big(1 + \frac{|X(\widehat{\beta} - \beta_0)|^2}{(n-p)\widehat{\sigma}^2}\Big)^{n/2} \end{aligned}$$

mit der empirischen Stichprobenvarianz $\widehat{\sigma}^2 = |Y - \Pi_X Y|^2/(n-p)$. Durch monotone Transformation hat also ein Likelihood-Quotiententest die Form

$$\varphi_\alpha = \mathbb{1}\Big(\frac{\frac{1}{p}|X(\widehat{\beta} - \beta_0)|^2}{\widehat{\sigma}^2} > \widetilde{c}_\alpha\Big),$$

wobei wir wegen der stetigen Verteilung auf die Randomisierung verzichten können. Da Π_X und $E_n - \Pi_X$ als Projektionen auf $\mathrm{Im}(X)$ bzw. $(\mathrm{Im}(X))^\perp$ symmetrische, idempotente Matrizen mit Rang p bzw. $(n-p)$ sind und $(E_n - \Pi_X)\Pi_X = 0$ gilt, folgt aus Lemma 2.39 unter H_0:

$$\frac{\frac{1}{p}|X(\widehat{\beta} - \beta_0)|^2}{\widehat{\sigma}^2} = \frac{(n-p)}{p}\frac{\varepsilon^\top \Pi_X \varepsilon}{\varepsilon^\top (E_n - \Pi_X)\varepsilon} \sim \mathrm{F}(p, n-p).$$

Man beachte, dass sich der wahre Wert von σ^2 im Bruch herauskürzt. Wählen wir also $\widetilde{c}_\alpha = q_{\mathrm{F}(p,n-p);1-\alpha}$, so besitzt φ_α Niveau α unter H_0. □

Im Allgemeinen bezeichnen wir einen Hypothesentest als F-Test, wenn seine Teststatistik unter der Nullhypothese einer F-Verteilung folgt. Zwei Hauptanwendungen von F-Tests sind die Überprüfung, ob die Regressionskoeffizienten in der linearen Regression signifikant von vorgegebenen Koeffizienten abweichen, und der Nachweis, ob sich die Mittelwerte aus zwei oder mehr Stichproben aus unterschiedlichen, normalverteilten Populationen signifikant unterscheiden (Varianzanalyse, Methode 2.62).

Beispiel 2.41 (F-Test, Happiness-Score) Erinnern wir uns an den World Happiness Report und speziell an den Zusammenhang zwischen pro-Kopf-Bruttoinlandsprodukt und Happiness-Score (siehe Beispiel 1.3). Ein Glücksforscher gibt als Faustregel die Formel $y = 2x + 4$ für den Happiness-Score y und das pro-Kopf-Bruttoinlandsprodukt x an. Mit $p = 2$ und $n = 156$ ist die Hypothese damit $H_0 \colon \beta = \beta_0$ für $\beta_0 = (4, 2)^\top$. Wir legen das Signifikanzniveau auf $\alpha = 0{,}05$ fest. Unter Zuhilfenahme einer Statistik-Software bestimmen wir den Kleinste-Quadrate-Schätzer, die Stichprobenvarianz, das entsprechende $(1 - \alpha)$-Quantil der $F_{2;198}$-Verteilung sowie den Wert der F-Statistik:

$$\widehat{\beta} \approx \begin{pmatrix} 3{,}40 \\ 2{,}22 \end{pmatrix}, \quad \widehat{\sigma}^2 \approx 0{,}46, \quad q_{\mathrm{F}(p,n-p);1-\alpha} \approx 3{,}05 \quad \text{und} \quad F \approx 28{,}77$$

Wegen $F > q_{\mathrm{F}(p,n-p);1-\alpha}$ wird die Nullhypothese verworfen. Die Faustregel ist demnach eine zu starke Vereinfachung der tatsächlichen Parameterwerte.

Für das Testen des reellen Parameters $\rho = \langle v, \beta \rangle$ zu gegebenem $v \in \mathbb{R}^p$ erhalten wir:

Satz 2.42 (t-Test) *Im gewöhnlichen linearen Modell unter Normalverteilungsannahme $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ mit unbekanntem $\sigma > 0$ betrachten wir den abgeleiteten Parameter $\rho = \langle v, \beta \rangle$ für ein $v \in \mathbb{R}^p \setminus \{0\}$. Dann ist der Likelihood-Quotiententest der Hypothese $H_0 \colon \rho = \rho_0, \sigma > 0$ gegen die Alternative $H_1 \colon \rho \neq \rho_0, \sigma > 0$ für ein $\rho_0 \in \mathbb{R}$ zum Niveau $\alpha \in (0, 1)$ gegeben durch den zweiseitigen **t-Test** (auch **Student-t-Test**)*

$$\varphi_\alpha = \mathbb{1}\left(|T| > q_{\mathrm{t}(n-p);1-\alpha/2}\right) \quad \textit{mit} \quad T := \frac{\widehat{\rho} - \rho_0}{\widehat{\sigma}\sqrt{\langle (X^\top X)^{-1} v, v \rangle}}$$

und $\widehat{\rho} = \langle v, \widehat{\beta} \rangle$, dem Kleinste-Quadrate-Schätzer $\widehat{\beta}$, der empirischen Stichprobenvarianz $\widehat{\sigma}^2$ und dem $(1 - \alpha/2)$-Quantil $q_{\mathrm{t}(n-p);1-\alpha/2}$ der $\mathrm{t}(n - p)$-Verteilung.

Im Allgemeinen bezeichnen wir Hypothesentests, deren Teststatistik unter der Nullhypothese t-verteilt sind, als t-Tests. Neben dem Einstichproben-t-Test gibt es auch den *Zweistichproben-t-Test*, der die Mittelwerte zweier unabhängiger Stichproben auf Gleichheit testet (siehe Korollar 2.65).

Beweis Der direkte Nachweis, dass der angegebene t-Test ein Likelihood-Quotiententest ist, verbleibt als Übung, siehe Aufgabe 2.13. Später wird diese Aussage in Beispiel 2.51 auch aus der allgemeinen Theorie folgen.

Aus dem Satz von Gauß-Markov und der Normalverteilungsannahme folgt $\widehat{\rho} \sim \mathrm{N}(\rho_0, \sigma^2 v^\top (X^\top X)^{-1} v)$ unter $H_0 \colon (\beta, \sigma) \in \{b \in \mathbb{R}^p : \langle v, b\rangle = \rho_0\} \times (0, \infty)$. Daraus folgt

$$\frac{\widehat{\rho} - \rho_0}{\sigma\sqrt{\langle (X^\top X)^{-1} v, v\rangle}} \sim \mathrm{N}(0,1) \text{ unter } H_0.$$

Andererseits sind $\widehat{\rho}$ und $\widehat{\sigma}^2$ unabhängig (wegen $(E_n - \Pi_X)\Pi_X = 0$ und Lemma 2.39), und es gilt $\widehat{\sigma}^2 = \sigma^2 Y/(n-p)$ für eine Zufallsvariable $Y \sim \chi^2(n-p)$. Damit ist

$$\frac{\widehat{\rho} - \rho_0}{\sqrt{\widehat{\sigma}^2\langle (X^\top X)^{-1} v, v\rangle}} \sim \mathrm{t}(n-p) \text{ unter } H_0,$$

und die Behauptung folgt durch Wahl des richtigen Quantils. □

Beispiel 2.43 (Einstichproben-t-Test) Mit dem Einstichproben-t-Test wird auf den Mittelwert einer i.i.d. normalverteilten Stichprobe getestet. Mögliche Anwendungsfälle sind Nachweise, ob Sollwerte, zum Beispiel das Füllgewicht von Zuckerpackungen, eingehalten werden oder die Istwerte signifikant davon abweichen.

Wir beobachten also $Y_i = \beta + \varepsilon_i$ mit $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ und unbekanntem Mittelwert β. Wir erhalten ein lineares Modell mit $X = (1, \ldots, 1)^\top \in \mathbb{R}^n$ und dem empirischen Mittelwert als Schätzer für β:

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Wir wenden nun Satz 2.42 an, wobei $p = 1$ und $v = 1$, sodass $\rho = \beta$ und $\widehat{\rho} = \widehat{\beta}$. Wir erhalten die Teststatistik

$$T = \frac{\sqrt{n}(\overline{Y} - \rho_0)}{\widehat{\sigma}}.$$

Im Gegensatz zum Gauß-Test aus Beispiel 2.34 wird also im t-Test die Varianz σ^2 durch die Stichprobenvarianz $\widehat{\sigma}^2$ ersetzt und ein Quantil der t-Verteilung benutzt.

Kurzbiografie (William Sealy Gosset) William Sealy Gosset wurde 1876 in Canterbury (Südostengland) geboren und studierte Chemie und Mathematik am New College in Oxford. Im Anschluss begann er bei der Dubliner Brauerei Arthur Guiness & Son zu arbeiten. Gossets Augenmerk war auf statistische Tests mit kleinen Stichprobengrößen gerichtet, was ein typisches Problem von Brauereien war. Da die Guiness-Brauerei ihren Mitarbeitern verbot, Arbeiten zu veröffentlichen, publizierte Gosset seine Erkenntnisse unter dem Pseudonym *Student* – daher auch der Name *Student-t-Verteilung*. Insbesondere Ronald Aylmer Fisher erkannte die Bedeutung von Gossets Arbeiten und entwickelte sie weiter, woraus die *t-Teststatistik* und die Anwendung der t-Verteilung in der Regressionsanalyse entstand. 1935 nahm Gosset eine Führungsposition in der neuen Guiness-Brauerei in London an, starb jedoch schon zwei Jahre später.

Der Korrespondenzsatz 1.70 zwischen Tests und Konfidenzbereichen liefert uns sofort folgende wichtige Konstruktionen von Konfidenzmengen:

Satz 2.44 ***(Konfidenzmengen im linearen Modell)*** *Im gewöhnlichen linearen Modell unter der Normalverteilungsannahme $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ für $\sigma > 0$ gelten für den Kleinste-Quadrate-Schätzer $\widehat{\beta}$ und die empirische Stichprobenvarianz $\widehat{\sigma}^2$ folgende Konfidenzaussagen für gegebenes Niveau $\alpha \in (0,1)$:*

(i) Ist $q_{\mathrm{F}(p,n-p);1-\alpha}$ das $(1-\alpha)$-Quantil der $\mathrm{F}(p, n-p)$-Verteilung, so ist

$$C := \left\{\beta \in \mathbb{R}^p \middle| |X(\beta - \widehat{\beta})|^2 \leqslant p\widehat{\sigma}^2 q_{\mathrm{F}(p,n-p);1-\alpha}\right\}$$

ein Konfidenzellipsoid zum Konfidenzniveau $1-\alpha$ für β.

(ii) Ist $q_{\mathrm{t}(n-p);1-\alpha/2}$ das $(1-\frac{\alpha}{2})$-Quantil der $\mathrm{t}(n-p)$-Verteilung, so ist

$$I := \left[\widehat{\rho} - \widehat{\sigma}\sqrt{\langle (X^\top X)^{-1} v, v\rangle} q_{\mathrm{t}(n-p);1-\alpha/2}, \widehat{\rho} + \widehat{\sigma}\sqrt{\langle (X^\top X)^{-1} v, v\rangle} q_{\mathrm{t}(n-p);1-\alpha/2}\right]$$

ein Konfidenzintervall zum Konfidenzniveau $1-\alpha$ für $\rho = \langle v, \beta\rangle$.

Eine allgemeinere Klasse statistischer Fragestellungen im linearen Modell sind lineare (bzw. affine) Testprobleme. Sie bieten eine Vielzahl von Anwendungsmöglichkeiten und sind aufgrund ihrer hohen Relevanz standardmäßig in Statistik-Software, wie zum Beispiel R (siehe linear.hypothesis im car-Paket), implementiert.

Definition 2.45 Im gewöhnlichen linearen Modell ist ein (zweiseitiges) **lineares Testproblem** gegeben durch

$$H_0 \colon K\beta = c, \sigma > 0 \quad \text{gegen} \quad H_1 \colon K\beta \neq c, \sigma > 0$$

für eine (deterministische) Matrix $K \in \mathbb{R}^{r\times p}$ mit vollem Rang $\operatorname{rank}(K) = r \leqslant p$ und einem Vektor $c \in \mathbb{R}^r$. K wird **Kontrastmatrix** genannt.

Für eine bessere Übersichtlichkeit verzichten wir im Folgenden auf die Bedingung $\sigma > 0$ in der Angabe von Hypothese und Alternative. Unter der Hypothese H_0 werden also insgesamt $r \leqslant p$ linear unabhängige Bedingungen an die Parameter des linearen Modells gestellt. Neben den Testproblemen die mit dem F-Test aus Satz 2.40 und dem t-Test aus Satz 2.42 behandelt wurden, umfassen lineare Testprobleme weitere wichtige Beispiele.

Beispiel 2.46 (Lineare Testprobleme)

1. Ein Test auf Gleichheit zweier Regressionskoeffizienten ist für $j, l \in \{1, \ldots, p\}$, $j \neq l$, gegeben durch

   $$H_0 \colon \beta_j = \beta_l \quad \text{gegen} \quad H_1 \colon \beta_j \neq \beta_l.$$

   Dies wird durch die Kontrastmatrix $K = (a_{1,i}) \in \mathbb{R}^{1\times p}$ mit $a_{1,i} = \mathbb{1}_{\{i=j\}} - \mathbb{1}_{\{i=l\}}$ und $c = 0$ modelliert.
2. Der Globaltest

   $$H_0 \colon \forall j \in \{1, \ldots, d\} : \beta_j = 0 \quad \text{gegen} \quad H_1 \colon \exists j \in \{1, \ldots, d\} : \beta_j \neq 0$$

   wird mit der Kontrastmatrix $K = E_p$ und $c = (0, \ldots, 0)^\top$ beschrieben.

3. Der Test eines Subvektors $\beta^* = (\beta_1^*, \dots, \beta_r^*)^\top$ mit $r \leqslant p$, das heisst

$$H_0 \colon \forall j \in \{1, \dots, r\} : \beta_j = \beta_j^* \quad \text{gegen} \quad H_1 \colon \exists j \in \{1, \dots, r\} : \beta_j \neq \beta_j^*,$$

führt auf die Kontrastmatrix $K = (\mathbb{1}_{i=j})_{i,j} \in \mathbb{R}^{r \times p}$ und $c = \beta^*$.

Die Grundidee für das Testen linearer Hypothesen ist es, die Residuen des auf $H_0 \colon K\beta = c$ eingeschränkten Kleinste-Quadrate-Schätzers mit denen des uneingeschränkten Kleinste-Quadrate-Schätzers $\widehat{\beta}$ zu vergleichen.

Definition 2.47 Der auf die Hypothese $H_0 \colon K\beta = c$ **eingeschränkte Kleinste-Quadrate-Schätzer** $\widehat{\beta}_{H_0}$ ist gegeben durch

$$\widehat{\beta}_{H_0} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p : K\beta = c} |Y - X\beta|^2. \tag{2.8}$$

Wir bezeichnen die **Summe der quadrierten Residuen** von $\widehat{\beta}$ und $\widehat{\beta}_{H_0}$ mit

$$RSS = |Y - X\widehat{\beta}|^2 \quad \text{bzw.} \quad RSS_{H_0} := |Y - X\widehat{\beta}_{H_0}|^2.$$

Per Definitionem ist $RSS_{H_0} \geqslant RSS$. Ist die Abweichung zu groß, spricht dies gegen die Hypothese H_0. In der Tat führt auch die Methode des Likelihood-Quotiententests zu diesem Ansatz.

Lemma 2.48 *Im gewöhnlichen linearen Modell unter Normalverteilungsannahme $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ besitzt jeder nichtrandomisierte Likelihood-Quotiententest für*

$$H_0 \colon K\beta = c \quad \textit{gegen} \quad H_1 \colon K\beta \neq c$$

mit Kontrastmatrix $K \in \mathbb{R}^{r \times p}$ und $c \in \mathbb{R}^r$ die Form

$$\varphi_\alpha = \mathbb{1}(F > \widetilde{c}_\alpha) \textit{ mit } F := \frac{n-p}{r} \frac{RSS_{H_0} - RSS}{RSS}, \quad \widetilde{c}_\alpha \geqslant 0.$$

Beweis Beachtet man, dass $\widehat{\beta}_{H_0}$ der Maximum-Likelihood-Schätzer unter der Nullhypothese ist, so ergibt sich genau wie im Beweis von Satz 2.40 für die Likelihood-Quotientenstatistik

$$\begin{aligned} T = \frac{\sup_{\beta \in \mathbb{R}^p, \sigma > 0} L(\beta, \sigma)}{\sup_{\beta \in \mathbb{R}^p \text{ mit } K\beta = c, \sigma > 0} L(\beta, \sigma)} &= \Big(\frac{|Y - X\widehat{\beta}_{H_0}|^2}{|Y - X\widehat{\beta}|^2} \Big)^{n/2} \\ &= \Big(1 + \frac{RSS_{H_0} - RSS}{RSS} \Big)^{n/2}. \end{aligned}$$

Durch monotone Transformation lässt sich daher ein nichtrandomisierter Likelihood-Quotiententest als $\varphi_\alpha = \mathbb{1}(F > \widetilde{c}_\alpha)$ mit $\widetilde{c}_\alpha \geqslant 0$ schreiben. □

Um den kritischen Wert $\widetilde{c}_\alpha$ von φ_α zu bestimmen, müssen wir die Teststatistik F, auch *Fisher-Statistik* genannt, genauer analysieren. Wir erhalten folgende umfassende Aussage:

Satz 2.49 (F-Test für lineare Hypothesen) *Im gewöhnlichen linearen Modell unter Normalverteilungsannahme $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ ist die lineare Hypothese*

$$H_0 \colon K\beta = c \quad \textit{gegen} \quad H_1 \colon K\beta \neq c$$

mit Kontrastmatrix $K \in \mathbb{R}^{r\times p}$ und $c \in \mathbb{R}^r$ zu testen.

(i) Für den Schätzer $\widehat{\beta}_{H_0}$ aus (2.8) *gilt*

$$\widehat{\beta}_{H_0} = \widehat{\beta} - (X^\top X)^{-1} K^\top \big(K(X^\top X)^{-1} K^\top\big)^{-1} (K\widehat{\beta} - c). \tag{2.9}$$

(ii) Es gilt

$$RSS_{H_0} - RSS = |X(\widehat{\beta} - \widehat{\beta}_{H_0})|^2 = \langle (K(X^\top X)^{-1}K^\top)^{-1}(K\widehat{\beta} - c), K\widehat{\beta} - c\rangle$$

und unter H_0 ist $(RSS_{H_0} - RSS)/\sigma^2$ $\chi^2(r)$-verteilt.

*(iii) Der **F-Test***

$$\varphi_\alpha = \mathbb{1}(F > q_{\mathrm{F}(r,n-p),1-\alpha}) \textit{ mit } F := \frac{n-p}{r}\frac{RSS_{H_0} - RSS}{RSS}$$

ist ein Likelihood-Quotiententest zum Niveau $\alpha \in (0,1)$.

Beweis *(i)* Wir bezeichnen die rechte Seite von (2.9) mit $\widetilde{\beta}_{H_0}$ und müssen zeigen, dass $\widetilde{\beta}_{H_0}$ die eindeutige Lösung der Optimierung in (2.8) ist. Zunächst weisen wir die Nebenbedingung nach. In der Tat gilt:

$$K\widetilde{\beta}_{H_0} = K\widetilde{\beta} - K(X^\top X)^{-1}K^\top (K(X^\top X)^{-1}K^\top)^{-1}(K\widetilde{\beta} - c) = c,$$

Aus Lemma 2.14 wissen wir, dass $Y - X\widehat{\beta} = (E_n - \Pi_X)Y \perp \mathrm{Im}(X)$, sodass für $\gamma \in \mathbb{R}^p$ nach dem Satz von Pythagoras

$$|Y - X\gamma|^2 = |Y - X\widehat{\beta} + X(\widehat{\beta} - \gamma)|^2 = |Y - X\widehat{\beta}|^2 + |X(\widehat{\beta} - \gamma)|^2$$

gilt. Außerdem ist

$$|X(\widehat{\beta} - \gamma)|^2 = |X(\widehat{\beta} - \widetilde{\beta}_{H_0})|^2 + |X(\widetilde{\beta}_{H_0} - \gamma)|^2 + 2\langle X(\widehat{\beta} - \widetilde{\beta}_{H_0}), X(\widetilde{\beta}_{H_0} - \gamma)\rangle.$$

Die Wahl von $\widetilde{\beta}_{H_0}$ impliziert jedoch für γ mit $K\gamma = c$

$$\begin{aligned}
&\langle X(\widehat{\beta} - \widetilde{\beta}_{H_0}), X(\widetilde{\beta}_{H_0} - \gamma)\rangle \\
&= \big((X^\top X)^{-1}K^\top (K(X^\top X)^{-1}K^\top)^{-1}(K\widehat{\beta} - c)\big)^\top X^\top X(\widetilde{\beta}_{H_0} - \gamma) \\
&= (K\widehat{\beta} - c)^\top (K(X^\top X)^{-1}K^\top)^{-1}(K\widetilde{\beta}_{H_0} - K\gamma) = 0,
\end{aligned}$$

wobei die letzte Gleichheit aus $K\widetilde{\beta}_{H_0} = K\gamma = c$ folgt. Insgesamt erhalten wir

$$|Y - X\gamma|^2 = |Y - X\widehat{\beta}|^2 + |X(\widehat{\beta} - \widetilde{\beta}_{H_0})|^2 + |X(\widetilde{\beta}_{H_0} - \gamma)|^2, \tag{2.10}$$

was offensichtlich für $\gamma = \widetilde{\beta}_{H_0}$ minimal ist, sodass $\widehat{\beta}_{H_0} = \widetilde{\beta}_{H_0}$ gilt.

(ii) Aus (2.10) mit $\gamma = \widehat{\beta}_{H_0}$ folgt durch Einsetzen von $\widehat{\beta}_{H_0}$

$$\begin{aligned} RSS_{H_0} - RSS =& |Y - X\widehat{\beta}_{H_0}|^2 - |Y - X\widehat{\beta}|^2 = |X(\widehat{\beta} - \widehat{\beta}_{H_0})|^2 \\ =& (\widehat{\beta} - \widehat{\beta}_{H_0})^\top X^\top X(\widehat{\beta} - \widehat{\beta}_{H_0}) \\ =& (K\widehat{\beta} - c)^\top (K(X^\top X)^{-1}K^\top)^{-1}(K\widehat{\beta} - c). \end{aligned}$$

Nach dem Satz von Gauß-Markov ist $\widehat{\beta}$ unter H_0 ein erwartungstreuer Schätzer von β mit $\mathbb{C}\mathrm{ov}_0(\widehat{\beta}) = \sigma^2(X^\top X)^{-1}$. Es ergibt H_0

$$\mathbb{E}_0[K\widehat{\beta} - c] = K\mathbb{E}_0[\widehat{\beta}] - c = K\beta - c = 0,$$

und

$$\mathrm{Var}_0(K\widehat{\beta} - c) = K\,\mathbb{C}\mathrm{ov}_0(\widehat{\beta})K^\top = \sigma^2 K(X^\top X)^{-1}K^\top.$$

Aus der Normalverteilung von $\widehat{\beta}$ folgt daher $(RSS_{H_0} - RSS)/\sigma^2 \sim \chi^2(r)$.

(iii) Da $RSS_{H_0} - RSS$ eine messbare Funktion von $\widehat{\beta}$ und somit auch von $X\widehat{\beta} = \Pi_X Y$ ist, ist $RSS = |(E_n - \Pi_X)Y|^2$ unabhängig von $RSS_{H_0} - RSS$. $F \sim \mathrm{F}(r, n-p)$ folgt daher aus der Charakterisierung der $\mathrm{F}(r, n-p)$-Verteilung aus Lemma 2.37. Wir schließen mit Lemma 2.48. □

Bemerkung 2.50 $W := rF$ heißt auch *Wald-Statistik*. Unter Verwendung von Satz 2.49(ii) und Lemma 2.18 können wir die Fisher-Statistik auch als

$$F = \frac{\frac{1}{r}|X\widehat{\beta} - X\widehat{\beta}_{H_0}|^2}{\widehat{\sigma}^2}$$

schreiben.

Beispiel 2.51 (t-Test als Spezialfall des F-Tests) Wir betrachten den Spezialfall einer eindimensionalen Nebenbedingung. Mit $r = 1$, $K = v^\top$, $\rho = \langle v, \beta\rangle$, $c = \rho_0$ testen wir also $H_0 \colon \rho = \rho_0$ gegen $H_1 \colon \rho \neq \rho_0$. Satz 2.49(ii) zeigt mit $\widehat{\rho} = \langle v, \widehat{\beta}\rangle$

$$F = \frac{|X\widehat{\beta} - X\widehat{\beta}_{H_0}|^2}{\widehat{\sigma}^2} = \frac{(\widehat{\rho} - \rho_0)^2}{\widehat{\sigma}^2\langle (X^\top X)^{-1}v, v\rangle} \sim \mathrm{F}(1, n-p) \text{ unter } H_0.$$

Damit ist der F-Test äquivalent zum zweiseitigen t-Test mit der Teststatistik

$$T = \frac{\widehat{\rho} - \rho_0}{\widehat{\sigma}\sqrt{\langle (X^\top X)^{-1}v, v\rangle}} \sim \mathrm{t}(n-p) \text{ unter } H_0.$$

Wir erhalten in diesem Fall also genau Satz 2.42.

Wir wollen nun die entwickelte Theorie auf reale Daten anwenden.

Beispiel 2.52 (Klimadaten) Wir betrachten die mittleren Julitemperaturen zwischen 1719 und 2020, die an einer Wetterstation in Berlin-Dahlem gemessen wurden. Bis

auf wenige Ausnahmen haben wir Messungen aus jedem Jahr. Insgesamt liegen $n = 291$ Beobachtungen vor, die über das Data Climate Center des Deutschen Wetterdienstes[1] frei verfügbar sind. Eine polynomiale Regression in der Zeit t (in Jahrhunderten beginnend bei 1719, siehe Abbildung 2.6) mit Polynomgraden $d = 1, \dots, 4$ liefert (mit gerundeten Werten)

$$\begin{aligned} p_1(t) &= 18{,}72 - 0{,}002\,t, \\ p_2(t) &= 18{,}63 + 0{,}16\,t - 0{,}05\,t^2, \\ p_3(t) &= 17{,}43 + 4{,}49\,t - 3{,}52\,t^2 + 0{,}75\,t^3, \\ p_4(t) &= 17{,}39 + 4{,}71\,t - 3{,}84\,t^2 + 0{,}92\,t^3 - 0{,}03\,t^4. \end{aligned}$$

Wir sehen hier sogar einen leichten negativen Anstieg in p_1, was den allgemeinen Erkenntnissen zur Klimaentwicklung zu widersprechen scheint. Andererseits ist der Faktor $-0{,}002$ nur sehr klein. Es fällt zudem auf, dass auch der jeweils letzte Grad in p_2 und p_4 einen nur sehr geringen Einfluss auf die Regressionskurve hat. Welches Polynom ist nun am besten geeignet, um die Daten zu beschreiben?

Um diese Frage beantworten zu können, verwenden wir die vorangegangene Testtheorie. Zunächst ist es plausibel, dass die zufälligen Schwankungen zwischen den Jahren unabhängig voneinander sind und als näherungsweise normalverteilt angenommen werden können (QQ-Plot). Um statistisch verwertbare Aussagen zu treffen, setzen wir noch das Niveau $\alpha = 0{,}05$ fest. Der Parametervektor ist $\beta = (\beta_0, \dots, \beta_d)^\top$.

Frage 1: Ist der negative Trend von p_1 signifikant, wenn wir das lineare Modell mit $d = 1$ annehmen? $H_0 \colon \beta_1 \geqslant 0$ gegen $H_1 \colon \beta_1 < 0$. Die zugehörige t-Statistik $T = -\frac{\widehat{\beta}_1}{\widehat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v}} \approx 0{,}01$ mit $v = (0, 1)^\top \in \mathbb{R}^2$ liegt deutlich unter dem kritischen Wert $q_{\mathrm{t}(n-2),1-\alpha} \approx 1{,}65$ (einseitiger T-Test), sodass die Hypothese nicht verworfen werden kann. Es gibt also keinen signifikant negativen Trend.

Frage 2: Liegt den Beobachtungen (im Modell mit $d = 4$) ein linearer Zusammenhang zugrunde? $H_0 \colon \beta_2 = \beta_3 = \beta_4 = 0$. Mittels Bemerkung 2.50 (oder direkt über die quadrierten Residuen) berechnen wir die Fisher-Statistik

$$F = \frac{\sum_{i=1}^n (p_4(t_i) - p_1(t_i))^2}{3\widehat{\sigma}^2} \approx 5{,}25 > 2{,}63 \approx q_{\mathrm{F}(3,n-5),1-\alpha}.$$

Folglich kann die Hypothese abgelehnt werden, und wir schlussfolgern, dass eine Regressionsgerade unzureichend ist.

Frage 3: Benötigen wir ein Polynom vierten Grades? $H_0 \colon \beta_4 = 0$. Die zugehörige t-Statistik hat den Wert $-0{,}11$, dessen Absolutbetrag kleiner als das Quantil $q_{\mathrm{t}(n-5);0{,}975} \approx 1{,}97$ ist (zweiseitiger t-Test). Diese Nullhypothese kann also akzeptiert werden.

Frage 4: Benötigen wir ein Polynom dritten Grades ? $H_0 \colon \beta_3 = 0$ (im Modell mit $d = 3$). Die zugehörige t-Statistik hat den Wert 3,96, dessen Absolutbetrag größer

[1] Siehe `https://cdc.dwd.de/portal/`

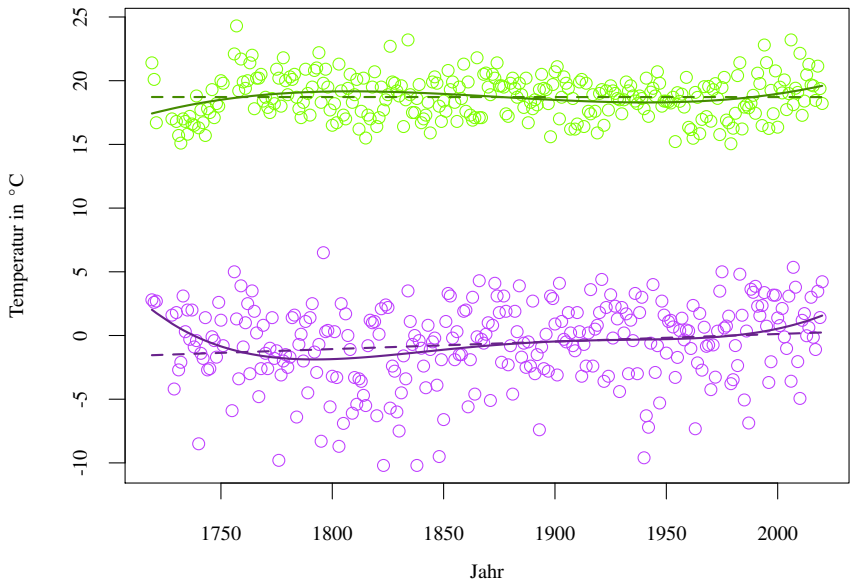

Abb. 2.6 Polynomielle Regression über die mittleren Julitemperaturen (grün) und Januartemperaturen (violett) in Berlin-Dahlem von 1719 bis 2020: Gestrichelte Linien zeigen die jeweiligen Regressionsgraden. Durchgezogene Linien zeigen das Regressionpolynom dritten Grades für den Juli und vierten Grades für den Januar. Datenbasis: Deutscher Wetterdienst

als das Quantil $q_{t(n-4);0,975} \approx 1{,}97$ ist. Die Hypothese kann also abgelehnt werden, und der kubische Anteil im Regressionspolynom ist signifikant.

Es sei darauf hingewiesen, dass wir die jeweiligen Testprobleme einzeln betrachten. Wenn wir alle Fragen simultan zu einem Niveau α beantworten wollen, so liegt ein multiples Testproblem vor, und die kritischen Werte müssen korrigiert werden (sogenannte *Multiplizitätskorrektur*), beispielsweise indem man bei m simultanen Tests jeweils das Niveau α/m verwendet (*Bonferroni-Korrektur*).

p_3 zeigt einen deutlichen Anstieg der Temperaturen in der zweiten Hälfte des 20. Jahrhunderts. Da wir hier nur eine einzelne Zeitreihe betrachtet haben, kann daraus aber kein allgemeiner Zusammenhang geschlossen werden. Das muss Ergebnis einer Kooperation mit Klimatologen sein. Eine analoge Analyse der Januar-Mitteltemperaturen zeigt übrigens einen signifikanten Koeffizienten vierten Grades (Aufgabe 2.17) und ebenfalls einen deutlichen Anstieg in den letzten 100 Jahren.

Zum Abschluss dieses Kapitels sei nochmal betont, dass vor der Anwendung der Inferenztheorie aus diesem Kapitel in Praxis zunächst geprüft werden muss, ob ein gewöhnliches lineares Modell unter Normalverteilungsannahme tatsächlich geeignet ist, um die beobachteten Daten $(x_i, Y_i)_{i=1,\dots,n}$ zu beschreiben.

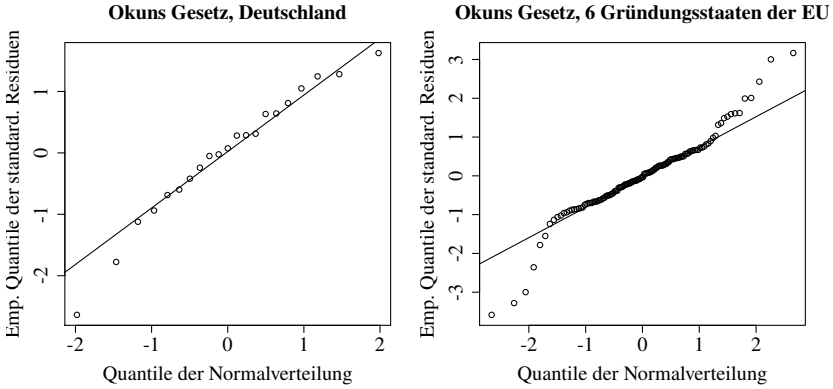

Abb. 2.7 QQ-Plots der standardisierten Residuen im einfachen linearen Modell für das jährliche Wachstum des BIP in Abhängigkeit von der jährlichen Veränderung der Arbeitslosenquote für Deutschland (links) und die 6 Gründungsstaaten der EU (rechts) in den Jahren 1992 bis 2012

Um dies zu prüfen, wird unter anderem der aus Lemma 2.18 bekannte Residuenvektor

$$R = Y - X\widehat{\beta} = (E_n - \Pi_X)\varepsilon$$

herangezogen, wobei $\widehat{\beta}$ der Kleinste-Quadrate-Schätzer und $\Pi_X = X(X^\top X)^{-1}X^\top$ die Projektionsmatrix auf Im X sind. Unter den Modellannahmen sind die Residuen R_i zentriert und normalverteilt, aber nicht mehr unabhängig voneinander, und sie besitzen verschiedene Varianzen $(1 - h_i)\sigma^2$ mit $h_i = (X(X^\top X)^{-1}X^\top)_{ii}$. Ein Plot der Residuen gegen die vorhergesagten Werte $X\widehat{\beta} = \Pi_X Y$ gibt Aufschluss, ob ein systematischer Modellfehler vorliegt. Hier sollte keine Abhängigkeit erkennbar sein, was aus Aufgabe 2.8 folgt.

Um die Normalverteilungsannahme zu überprüfen, bietet sich ein QQ-Plot der standardisierten Residuen

$$T_i = \frac{R_i}{\sqrt{(1 - h_i)\widehat{\sigma}^2}}$$

an. Man beachte, dass wir die unbekannte Varianz σ^2 durch ihren Schätzer aus Lemma 2.18 ersetzt haben, sodass T_i nur noch approximativ normalverteilt ist.

Beispiel 2.53 (QQ-Plots für Okuns Gesetz) Wir greifen den empirischen Zusammenhang zwischen der änderung der Arbeitslosenquote x_i und Wachstum des Bruttoinlandsproduktes Y_i aus Beispiel 2.2 auf. Abbildung 2.7 zeigt die aus dem einfachen linearen Modell $Y_i = ax_i + b + \varepsilon_i$ für die Jahre $i = 1992, \ldots, 2012$ resultierenden QQ-Plots, wobei Deutschland einzeln und zusammen mit 5 weiteren EU-Staaten betrachtet wird. In beiden Fällen sehen wir, dass die Normalverteilungsannahme im Zentrum gut zutrifft, jedoch sehr große und sehr kleine Quantile zum Teil zu deutlichen Abweichungen führen.

Statistische Tests zum überprüfen der Verteilungsannahme beruhen beispielsweise auf dem χ^2-Test oder dem Kolmogorov-Smirnov-Test, auf die wir hier jedoch nicht eingehen werden. Die interessierte Leserin sei auf Lehmann and Romano (2005) verwiesen. Die gesamte Modellverifikation für das lineare Modell wird von Fahrmeir et al. (2009) ausführlich diskutiert.

2.3 Varianzanalyse

In der Varianzanalyse werden Varianzen verschiedener Gruppen benutzt, um mögliche Unterschiede zwischen diesen Gruppen nachzuweisen. Die Varianzanalyse beruht auf dem Testen von Spezialfällen linearer Hypothesen im linearen Modell. Wir werden also unsere bisherigen Resultate anwenden, aber gewisse Strukturen zusätzlich ausnutzen.

Beispiel 2.54 (Düngemittel – a) Um den Einfluss von $p \in \mathbb{N}$ verschiedenen Düngemitteln auf den Ernteertrag zu vergleichen, wird jedes Düngemittel $i \in \{1, \dots, p\}$ auf n_i verschiedenen Agrarflächen ausgebracht. Der durch Witterungseinflüsse etc. zufällige Ernteertrag kann mittels $Y_{ij} = \mu_i + \varepsilon_{ij}$ für $j = 1, \dots, n_i$ und $i = 1, \dots, p$ modelliert werden, wobei μ_i der mittlere Ernteertrag von Düngemittel i ist und ε_{ij} unabhängige, zentrierte Störgrößen sind. Die zu untersuchende Frage ist, ob die einzelnen Düngemittel einen unterschiedlichen Einfluss auf den mittleren Ernteertrag haben.

Definition 2.55 Das Modell der **einfaktoriellen Varianzanalyse** (englisch: *(one-way) analysis of variance*, kurz: ANOVA1) ist gegeben durch Beobachtungen

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, p, j = 1, \dots, n_i,$$

mit i.i.d.-verteilten Störgrößen $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$. Wir bezeichnen den ersten Index als den **Faktor** und den Wert $i = 1, \dots, p$ als die **Faktorstufe**. Folglich geben $(n_i)_{i=1,\dots,p}$ die Anzahl der unabhängigen Versuchswiederholungen pro Faktor an und $n := \sum_{i=1}^{p} n_i$ ist der Gesamtstichprobenumfang. Gilt $n_1 = \cdots = n_p$, so sprechen wir von **balanciertem Design**.

Beispiel 2.56 (Düngemittel – b) Der Faktor sind die Düngemittel, und weil wir zwei Düngemittel testen, haben wir zwei Faktorstufen $i \in \{1, 2\}$. Der Gesamtstichprobenumfang beträgt $n_1 + n_2 = 2 + 3 = 5$, und es liegt wegen $n_1 \neq n_2$ kein balanciertes Design vor.

Bemerkung 2.57 (ANOVA1) Das ANOVA1-Modell ist ein Spezialfall des gewöhnlichen linearen Modells der Form

$$\mathbb{R}^n \ni Y := \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_p} \end{pmatrix} = \underbrace{\begin{pmatrix} 1\,0\cdots 0 \\ \vdots\,\vdots\quad\ \vdots \\ 1\,0\cdots 0 \\ \vdots\,\vdots\quad\ \vdots \\ 0\,0\cdots 1 \\ \vdots\quad\ \vdots \\ 0\,0\cdots 1 \end{pmatrix}}_{=:X\in\mathbb{R}^{n\times p}} \cdot \underbrace{\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}}_{=:\mu\in\mathbb{R}^p} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{p1} \\ \vdots \\ \varepsilon_{pn_p} \end{pmatrix}.$$

Hierbei gilt $\mathrm{Im}(X) = p$.

Die klassische Fragestellung der Varianzanalyse lautet: „Existieren Unterschiede in den faktorstufenspezifischen Mittelwerten μ_i?“ oder anders formuliert „Hat der Faktor einen Einfluss auf die Response-Variable oder nicht?“. Dies führt auf das *Testproblem*

$$H_0 \colon \mu_1 = \cdots = \mu_p \quad \text{gegen} \quad H_1 \colon \exists i, l \in \{1, \ldots, p\} : \mu_i \neq \mu_l. \tag{2.11}$$

Eine erste Idee, um diese Hypothese zu überprüfen, wäre die Mittelwerte jeder Faktorstufe zu berechnen und diese zu vergleichen. Eine große Abweichung würde gegen die Hypothese sprechen. Dieser Vergleich muss jedoch die Streuung der Beobachtungen um den jeweiligen Mittelwert berücksichtigen, da andernfalls nicht klar ist, was eine große Abweichung ist (man vergleiche beispielsweise $\mathrm{N}(0; 1)$ mit $\mathrm{N}(0{,}1; 1)$ und $\mathrm{N}(0; 0{,}01)$ mit $\mathrm{N}(0{,}1; 0{,}01)$). Wir sollten also *die Nullhypothese ablehnen, wenn die Streuung zwischen den Gruppen größer ist als die Streuung innerhalb der Gruppen*. Dieses Vorgehen motiviert die Bezeichnung Varianzanalyse.

Beispiel 2.58 (Düngemittel – c) Wir bleiben bei zwei Faktorstufen (zwei Düngemittel), erhöhen aber zur besseren Illustration die Anzahl der Felder pro Düngemittel $n_1 = n_2 = 50$. Abbildung 2.8 zeigt die Histogramme von drei verschiedenen Ernteszenarien, bei denen sich die Mittelwerte annähern, während die Varianzen gleich bleiben. Man sieht deutlich, dass im ersten Fall beide Faktorstufen leicht voneinander zu unterscheiden sind, während dies im dritten Fall kaum noch möglich ist.

Lemma 2.59 (Streuungszerlegung) *Im Modell der einfaktoriellen Varianzanalyse definieren wir das i-te Gruppenmittel, $i = 1, \ldots, p$, bzw. das Gesamtmittel als*

$$\overline{Y}_{i\bullet} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad \textit{bzw.} \quad \overline{Y}_{\bullet\bullet} := \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}$$

sowie die Streuungsmaße

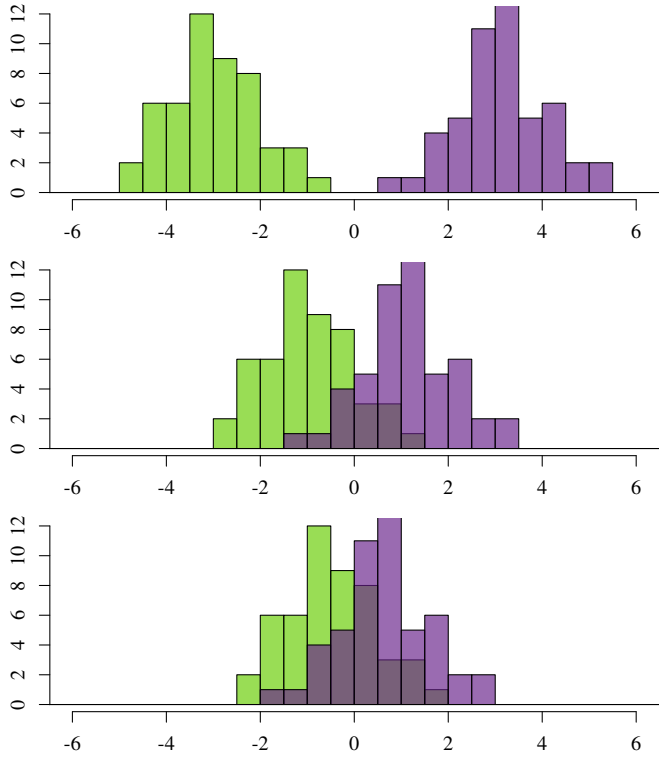

Abb. 2.8 Teilweise überlappende Histogramme von drei verschiedenen, normalverteilten Beobachtungen von zwei Gruppen (grün und violett). Von oben nach unten rücken die Erwartungswerte der Gruppen näher aneinander (erst liegen sie bei ±3, dann bei ±1 und zuletzt bei ±0,5) mit gleichbleibenden Varianzen ($\sigma^2 = 1$).

$$SSB := \sum_{i=1}^{p} n_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \quad \textit{und} \quad SSW := \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2.$$

Dann gilt

$$SST := \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\bullet\bullet})^2 = SSB + SSW.$$

Beweis Es gilt

$$\begin{aligned} SST &= \sum_i \sum_j (Y_{ij} - \overline{Y}_{\bullet\bullet})^2 = \sum_i \sum_j (Y_{ij} - \overline{Y}_{i\bullet} + \overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \\ &= \sum_i \sum_j \left((Y_{ij} - \overline{Y}_{i\bullet})^2 + 2(Y_{ij} - \overline{Y}_{i\bullet})(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) + (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \right), \end{aligned}$$

wobei

$$\sum_i \sum_j (Y_{ij} - \overline{Y}_{i\bullet})(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) = \sum_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) \sum_j (Y_{ij} - \overline{Y}_{i\bullet})$$
$$= \sum_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})(n_i \overline{Y}_{i\bullet} - n_i \overline{Y}_{i\bullet}) = 0.$$

Damit ist die Darstellung von *SST* gezeigt. □

Bemerkung 2.60 Nach Normierung mit $1/n$ handelt es sich bei den Größen *SSB*, *SSW* bzw. *SST* um die gewichteten empirischen Varianzen der Mittelwerte der Gruppen (englisch: *sum of squares between groups*), der Summe der empirischen Varianz innerhalb der Gruppen (englisch: *sum of squares within groups*) bzw. der empirischen Varianz der gesamten Stichprobe (englisch: *total sum of squares*).

Satz 2.61 *Im Modell der einfaktoriellen Varianzanalyse mit $n > p \geqslant 2$ gilt:*

(i) Der Kleinste-Quadrate-Schätzer von $\mu = (\mu_1, \ldots, \mu_p)^\top$ ist gegeben durch

$$\widehat{\mu} = (\overline{Y}_{1\bullet}, \ldots, \overline{Y}_{p,\bullet})^\top.$$

(ii) $SSW/\sigma^2 \sim \chi^2(n-p)$ und unter $H_0\colon \mu_1 = \cdots = \mu_p$ gilt $SSB/\sigma^2 \sim \chi^2(p-1)$.

(iii) SSW und SSB sind unabhängig und somit $F := \frac{n-p}{p-1} \frac{SSB}{SSW} \overset{H_0}{\sim} \mathrm{F}(p-1, n-p)$.

Beweis *(i)* Nachrechnen zeigt

$$\widehat{\mu} = (X^\top X)^{-1} X^\top Y = \begin{pmatrix} 1/n_1 & & 0 \\ & \ddots & \\ 0 & & 1/n_p \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1j} \\ \vdots \\ \sum_{j=1}^{n_p} Y_{kj} \end{pmatrix} = \begin{pmatrix} \overline{Y}_{1\bullet} \\ \vdots \\ \overline{Y}_{p\bullet} \end{pmatrix}.$$

(ii) und (iii) Wegen $SSW = |Y - X\widehat{\mu}|^2 = |R|^2$ für die Residuen R aus Lemma 2.18 folgt $SSW/\sigma^2 \sim \chi^2(n-p)$ und die Unabhängigkeit von *SSW* und $\widehat{\mu}$ aus Lemma 2.39. Nach dem vorangegangen Satz gilt weiterhin $SSB = SST - SSW$. Somit folgt die Behauptung aus Satz 2.49, falls $SST = |Y - X\widehat{\mu}_{H_0}|^2$, wobei $\widehat{\mu}_{H_0}$ gegeben ist durch

$$|Y - X\widehat{\mu}_{H_0}|^2 = \min_{\mu \in \mathbb{R}} \Big| Y - X \underbrace{\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}}_{\in \mathbb{R}^p} \Big|^2 = \min_{\mu \in \mathbb{R}} \Big| Y - \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{=: X_0 \in \mathbb{R}^{n_\bullet \times 1}} \mu \Big|^2.$$

Dieses Minimierungsproblem wird durch $\widehat{\mu}_{H_0} = (X_0^\top X_0)^{-1} X_0^\top Y = n^{-1} \sum_{i,j} Y_{ij} = \overline{Y}_{\bullet\bullet}$ gelöst. Damit folgt die Behauptung. □

Folglich können wir die Hypothese aus (2.11) mit dem Likelihood-Quotiententest aus Satz 2.49, das heißt einem F-Test, überprüfen.

	FG	Quadratsummen	Quadratmittel	F-Statistik
zwischen	$p-1$	SSB= $\sum_{i=1}^{p} n_i(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2$	SSB$/(p-1)$	$\frac{n-p}{p-1}\frac{\text{SSB}}{\text{SSW}}$
innerhalb	$n-p$	SSW= $\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2$	SSW$/(n-p)$	
total	$n-1$	SST= $\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\bullet\bullet})^2$	SST$/(n-1)$	

Tabelle 2.1 ANOVA1-Tafel zur Darstellung von Freiheitsgraden (FG), Quadratsummen, Quadratmittel und der resultierenden F-Statistik in der einfaktoriellen Varianzanalyse

	FG	Quadratsummen	Quadratmittel	F-Statistik
zwischen	1	$SSB = 199{,}314$	$SSB/1 = 199{,}314$	
innerhalb	98	$SSW = 2{,}402$	$SSW/98 = 0{,}025$	$F = 8131{,}686$
total	99	$SST = 201{,}716$	$SST/99 = 2{,}037$	

	FG	Quadratsummen	Quadratmittel	F-Statistik
zwischen	1	$SSB = 7{,}471$	$SSB/1 = 7{,}471$	
innerhalb	98	$SSW = 8{,}445$	$SSW/98 = 0{,}086$	$F = 86{,}700$
total	99	$SST = 15{,}916$	$SST/99 = 0{,}161$	

	FG	Quadratsummen	Quadratmittel	F-Statistik
zwischen	1	$SSB = 12{,}171$	$SSB/1 = 12{,}171$	
innerhalb	98	$SSW = 116{,}014$	$SSW/98 = 1{,}184$	$F = 10{,}281$
total	99	$SST = 128{,}185$	$SST/99 = 1{,}295$	

Tabelle 2.2 ANOVA1-Tafeln für den Ernteertrag unter dem Einsatz von zwei verschiedenen Düngemitteln in drei verschiedenen Szenarien

Methode 2.62 (Einfaktorielle Varianzanalyse) Im Modell der einfaktoriellen Varianzanalyse testen wir

$$H_0 \colon \mu_1 = \cdots = \mu_p \quad \text{versus} \quad H_1 \colon \exists i, l \in \{1, \dots, p\} : \mu_i \neq \mu_l$$

zum Niveau $\alpha \in (0, 1)$ durch den F-Test

$$\varphi_\alpha = \mathbb{1}(F > q_{\mathrm{F}(p-1,n-p);1-\alpha}) \quad \text{mit} \quad F := \frac{n-p}{p-1}\frac{\text{SSB}}{\text{SSW}}.$$

Es ist übersichtlich, die einzelnen Zwischenergebnisse der ANOVA1 in eine Tabelle zu schreiben (siehe Tabelle 2.1 und das folgende Beispiel).

Beispiel 2.63 (Düngemittel – d) Wie zuvor betrachten wir zwei Düngemittel, die jeweils auf 50 Feldern eingesetzt werden und wenden den F-Test auf die drei Beobachtungssätze aus Abbildung 2.8 an. Für die drei Szenarien erhalten wir ANOVA1-Tafeln, siehe Tabelle 2.2.

Es gilt $q_{\mathrm{F}(1,98);0,95} = 3{,}938$, $q_{\mathrm{F}(1,98);0,99} = 6{,}901$ und $q_{\mathrm{F}(1,98);0,999} = 11{,}510$. Bis auf den Test zum Niveau $\alpha = 0.001$ der dritten Beobachtung wird die Nullhypothese also immer abgelehnt.

Bemerkung 2.64 (Effektdarstellung) Das einfaktorielle Varianzanalysemodell lässt sich zu

$$Y_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, p, j = 1, \dots, n_i,$$

umformen, wobei $\mu_0 := \frac{1}{n} \sum_{i=1}^{p} n_i \mu_i = \mathbb{E}[\overline{Y}_{\bullet\bullet}]$ das Gesamtmittel ist und $\alpha_i := \mu_i - \mu_0$, $i = 1, \dots, p$, den **Effekt der Faktorstufe** beschreibt. Diese Form heißt **Effektdarstellung**, und sie verlangt die Nebenbedingung

$$0 = \sum_{i=1}^{p} n_i \alpha_i \quad \text{oder äquivalent} \quad \alpha_p = -\frac{1}{n_p} \sum_{i=1}^{p-1} n_i \alpha_i,$$

damit die Designmatrix weiter vollen Rang hat. Der Parametervektor ist also gegeben durch $(\mu_0, \alpha_1, \dots, \alpha_{p-1})^\top$. Die F-Statistik, um die Globalhypothese

$$H_0 \colon \alpha_1 = \cdots = \alpha_{p-1} = 0$$

zu überprüfen, ist identisch zur Statistik aus Satz 2.61. Per Konstruktion lässt sich somit anhand der Schätzungen des Parametervektors ablesen, wie stark der Effekt, zum Beispiel von Düngemitteln, im Vergleich zum Gesamtdurchschnitt ist (negativ wie auch positiv).

Im Fall $p = 2$ führt die Varianzanalyse auf den Zweistichproben-t-Test.

Korollar 2.65 (Zweistichproben-t-Test) *Im Modell der einfaktoriellen Varianzanalyse mit $p = 2$ und dem Testproblem $H_0 \colon \mu_1 = \mu_2, \sigma > 0$ gegen $H_1 \colon \mu_1 \neq \mu_2, \sigma > 0$ ist*

$$\varphi_\alpha = \mathbb{1}(|T| > q_{\mathrm{t}(n-2),1-\alpha/2}) \quad \textit{mit} \quad T := \frac{\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) SSW/(n-2)}}$$

*mit dem $(1-\alpha/2)$-Quantil $q_{\mathrm{t}(n-2),1-\alpha/2}$ der $\mathrm{t}(n-2)$-Verteilung der sogenannte **Zweistichproben-t-Test** der Hypothese H_0 zum Niveau $\alpha \in (0,1)$.*

Beweis Wegen $p = 2$ und Satz 2.61 gilt unter H_0

$$\frac{n-p}{p-1} \cdot \frac{SSB}{SSW} = \frac{SSB}{SSW/(n-2)} \sim F(1, n-2).$$

Zur Erinnerung gilt $T_n^2 = F_{1,n}$. Wegen $n\overline{Y}_{\bullet\bullet} = n_1 \overline{Y}_{1\bullet} + n_2 \overline{Y}_{2\bullet}$ gilt

$$\begin{aligned} SSB &= n_1(\overline{Y}_{1\bullet} - \overline{Y}_{\bullet\bullet})^2 + n_2(\overline{Y}_{2\bullet} - \overline{Y}_{\bullet\bullet})^2 \\ &= n_1 \overline{Y}_{1\bullet}^2 + n_2 \overline{Y}_{2\bullet}^2 + n\overline{Y}_{\bullet\bullet}^2 - 2(n_1 \overline{Y}_{1\bullet} + n_2 \overline{Y}_{2\bullet})\overline{Y}_{\bullet\bullet} \\ &= n_1 \overline{Y}_{1\bullet}^2 + n_2 \overline{Y}_{2\bullet}^2 - \frac{1}{n}\big(n_1 \overline{Y}_{1\bullet} + n_2 \overline{Y}_{2\bullet}\big)^2 = \frac{n_1 n_2}{n}\big(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}\big)^2. \end{aligned}$$

Folglich gilt unter H_0

$$T := \frac{\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})SSW/(n-2)}} \sim \mathrm{t}(n-2).$$

Bemerkung 2.66 (Zweistichproben-t-Test) Alternativ kann die Teststatistik des Zweistichproben-t-Tests auch als

$$T = \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frac{\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}}{\sqrt{\frac{n_1-1}{n-2}\widehat{\sigma}_1^2 + \frac{n_2-1}{n-2}\widehat{\sigma}_2^2}} \quad \text{mit} \quad \widehat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$$

geschrieben werden. Dabei wird die Varianz σ^2 durch das gewichtete Mittel der gruppenweisen empirischen Varianzen $\widehat{\sigma}_i^2$ geschätzt.

Sind die Varianzen in den Grundgesamtheiten (zum Beispiel die Felder mit den verschiedenen Düngemitteln) ungleich, dann kann obiges Resultat nicht angewendet werden. Eine Modifikation des Zweistichproben-t-Test führt auf den Welch-Test, der den verschiedenen Varianzen Rechnung trägt, siehe Lehmann and Romano (2005).

Nach dem Studium der einfaktoriellen Varianzanalyse liegt es nahe, den Einfluss von mehreren Faktoren zu berücksichtigen. Wir beschränken uns hier auf zwei Faktoren. Eine Varianzanalyse mit mehr als zwei Faktoren führt zu ähnlichen Verteilungsaussagen und resultierenden Tests.

Definition 2.67 Das Modell der **zweifaktoriellen Varianzanalyse** (kurz: ANOVA2) mit balanciertem Design ist gegeben durch Beobachtungen

$$\begin{aligned} Y_{ijk} =& \mu_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K \\ =& \mu_0 + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \end{aligned}$$

mit $I, J, K \geqslant 2$, i.i.d.-verteilten Störgrößen $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$ und Nebenbedingungen (der Effektdarstellung)

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0.$$

Wir haben also zwei Faktoren mit Faktorstufen $i = 1, \ldots, I$ und $j = 1, \ldots, J$. (α_i) und (β_j) heißen **Haupteffekte** des ersten beziehungsweise zweiten Faktors. (γ_{ij}) heißen **Interaktions-** oder **Wechselwirkungseffekte**.

Das ANOVA2-Modell ist also ein lineares Modell mit zwei kategoriellen Kovariablen. Die Gesamtanzahl an Beobachtungen ist gegeben durch $n = I \cdot J \cdot K$ und die Parameterdimension ist $p = I \cdot J$. Die typische Testprobleme sind

$$H_0: \quad \forall i : \alpha_i = 0 \quad \text{gegen} \quad H_1 : \exists i \in \{1, \ldots, I\} : \alpha_i \neq 0, \tag{2.12}$$

$$H_0: \quad \forall j : \beta_j = 0 \quad \text{gegen} \quad H_1 : \exists j \in \{1, \ldots, J\} : \beta_j \neq 0, \tag{2.13}$$

$$H_0: \forall i, j : \gamma_{ij} = 0 \quad \text{gegen} \quad H_1 : \exists i \in \{1, \ldots, I\}, j \in \{1, \ldots, J\} : \gamma_{ij} \neq 0. \tag{2.14}$$

Satz 2.68 *Im Modell der zweifaktoriellen Varianzanalyse mit balanciertem Design gilt:*

(i) Die Kleinsten-Quadrate-Schätzer für μ_0, α_i, β_j *und* γ_{ij}, $i = 1, \dots, I-1, j = 1, \dots, J-1$, *sind gegeben durch*

$$\widehat{\mu}_0 = \overline{Y}_{\bullet\bullet\bullet}, \quad \widehat{\alpha}_i = \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}, \quad \widehat{\beta}_j = \overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet},$$
$$\widehat{\gamma}_{ij} = (\overline{Y}_{ij\bullet} - \overline{Y}_{\bullet\bullet\bullet}) - \widehat{\alpha}_i - \widehat{\beta}_j = \overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet},$$

wobei wir wieder in den mit $\bullet$ *gekennzeichneten Koordinaten Mittelwerte bilden.*

(ii) Definieren wir

$$SSW := \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{ij\bullet})^2,$$
$$SSB_1 := JK\sum_{i=1}^{I}(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2, \quad SSB_2 := IK\sum_{j=1}^{J}(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2,$$
$$SSB_{12} := K\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2,$$

dann können die Hypothesen (2.12), (2.13) bzw. (2.14) mit den F-Statistiken

$$\frac{IJ(K-1)}{I-1}\frac{SSB_1}{SSW} \sim \mathrm{F}(I-1, IJ(K-1)),$$
$$\frac{IJ(K-1)}{J-1}\frac{SSB_2}{SSW} \sim \mathrm{F}(J-1, IJ(K-1)) \quad \textit{bzw.}$$
$$\frac{IJ(K-1)}{(I-1)(J-1)}\frac{SSB_{12}}{SSW} \sim \mathrm{F}\big((I-1)(J-1), IJ(K-1)\big)$$

getestet werden, wobei die Statistiken jeweils unter der Nullhypothese F-verteilt sind.

Den Beweis überlassen wir den Leserinnen und Lesern als Übung (Aufgabe 2.18).

Beispiel 2.69 (ANOVA2) Eine Bäuerin interessiert sich neben dem Effekt ihrer Düngemittel auch für den Einfluss der genutzten Samenarten. Sie verwendet $I = 3$ verschiedene Samenarten und $J = 5$ verschiedene Düngemittel. Das ergibt $3 \cdot 5 = 15$ verschiedene Feldergruppen, auf denen je genau eine Samenart und ein Düngemittel ausgebracht wird. Jede Samen-Düngemittel-Kombination wird auf $K = 2$ Feldern eingesetzt. Die Erträge der Bäuerin sind in Tabelle 2.3 zusammengefasst.

Für $\alpha = 0{,}05$ erhalten wir die ANOVA-Tafel in Tabelle 2.4. Die Hypothesen, dass die Haupteffekte null seien, werden also verworfen, aber die Hypothese, dass die Interaktionseffekte null sind, wird bestätigt.

	DM1	DM2	DM3	DM4	DM5
SA1	111;116	100;106	99;113	108;110	105;108
SA2	115;118	103;105	105;107	113;118	110;113
SA3	99;103	91;93	103;105	104;107	99;104

Tabelle 2.3 Messung der Ernteerträge von je zwei Feldern pro Samenart-Düngemittel- Kombination. DM steht für Düngemittel und SA für Samenart.

	FG	Quadratsummen	Quadratmittel	F-Statistik	F-Quantil
Zwischen	2	$SSB_1 = 512{,}867$	256,433	20,298	3,682
	4	$SSB_2 = 449{,}467$	112,367	8,894	3,056
	8	$SSB_{12} = 143{,}133$	17,892	1,416	2,641
Innerhalb	15	$SSW = 189{,}500$	12,633		
Total	29	$SST = 1294{,}967$	44,654		

Tabelle 2.4 ANOVA2-Tafel, wobei in der ersten Spalte von rechts das $(1-\alpha)$-Quantil der F-Verteilung mit entsprechenden Freiheitsgraden steht

2.4 Aufgaben

2.1 Bestimmen Sie im einfachen linearen Modell

$$Y_i = ax_i + b + \varepsilon_i \quad \text{für} \quad i = 1, \ldots, n, n \in \mathbb{N}$$

die Maximum-Likelihood-Schätzer für die unbekannten Parameter $a, b \in \mathbb{R}$ aufgrund der Beobachtungen $Y_1, \ldots, Y_n$, wenn

(a) $(\varepsilon_1, \ldots, \varepsilon_n)$ ein Vektor unabhängiger (zentrierter) Laplace-verteilter Zufallsvariablen mit Skalierungsparameter $\beta > 0$ ist, das heißt die Verteilung von ε_i hat die Lebesgue-Dichte $f_\beta(x) = \frac{2}{\beta} e^{-|x|/\beta}, x \in \mathbb{R}$,

(b) $(\varepsilon_1, \ldots, \varepsilon_n)$ ein Vektor unabhängiger $\mathrm{Exp}(\lambda)$-verteilter Zufallsvariablen mit Parameter $\lambda > 0$ ist, das heißt die Verteilung von ε_i hat die Lebesgue-Dichte $g_\lambda(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,\infty)}(x), x \in \mathbb{R}$.

2.2 Untersuchen Sie den Zusammenhang zwischen Bruttoinlandsprodukt und Happiness-Score aus Beispiel 1.3. Bestimmen Sie anhand der Daten aus dem Jahr 2019 die Regressionsgrade. Recherchieren Sie das Bruttoinlandsprodukt von Deutschland im Jahr 2020. Welchen Happiness-Score würden Sie vorhersagen?

2.3 Zehn erkrankte Patienten nehmen dasselbe Medikament, jedoch in verschiedenen Dosen. Folgende Tabelle zeigt die Anzahl an Tagen bis zur Genesung:

(a) Verwenden Sie ein einfaches lineares Modell, um festzustellen, ob eine höhere Dosis zu einer schnelleren Genesung führt.

(b) Die ersten fünf Patienten in der Tabelle waren Frauen, während die hinteren fünf Patienten Männer waren. Schätzen Sie für jede der beiden Gruppen eine eigene Regressiongrade. ändert dies die Schlussfolgerung aus (a)?

Dosis	45	55	70	60	75	80	100	90	110	125
Genesungszeit	5	6	8	8	9	3	4	6	5	7

2.4 Im linearen Modell $Y = X\beta + \varepsilon$ mit dem Mittelwert der Beobachtung $\overline{Y}$ und den geschätzten Zielgrößen $\widehat{Y} := X\widehat{\beta}$ heißt

$$R^2 := \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

Bestimmtheitsmaß. Zeigen Sie, dass R^2 in $[0, 1]$ liegt. Weisen Sie nach, dass $R^2 = 0$ einen linearen Zusammenhang ausschließt und dass $R^2 = 1$ einen perfekt linearen Zusammenhang impliziert.

2.5 Leiten Sie aus dem Satz von Gauß-Markov ab, dass die optimale Varianz im gewöhnlichen linearen Modell

(a) von $\widehat{\beta}$ gleich $\sigma^2(X^TX)^{-1}$ ist,
(b) von $\langle v, \widehat{\beta}\rangle$ gleich $\sigma^2|X(X^TX)^{-1}v|^2$ ist.

2.6 Bestimmen Sie im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n), \sigma > 0$, den Maximum-Likelihood-Schätzer $\widehat{\sigma}^2_{ML}$ von σ^2.Vergleichen Sie Bias und Varianz von $\widehat{\sigma}^2_{ML}$ und dem Schätzer $\widehat{\sigma}^2 = |Y - X\widehat{\beta}|^2/(n-k)$ aus Lemma 2.18.

2.7 Beweisen Sie im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n), \sigma > 0$, dass für den Kleinste-Quadrate-Schätzer $\widehat{\beta}$ gilt:

$$\frac{1}{n}\mathbb{E}\big[|X\beta - X\widehat{\beta}|^2\big] = \sigma^2\frac{p}{n}.$$

2.8 Im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n), \sigma > 0$, und Kleinste-Quadrate-Schätzer $\widehat{\beta} = X^\dagger Y$ sind $\widehat{Y} := (\widehat{Y}_1, \ldots, \widehat{Y}_n)^\top := X\widehat{\beta}$ die geschätzten Erwartungswerte und $R = (R_1, \ldots, R_n)^\top = Y - \widehat{Y}$ die Residuen. Beweisen Sie folgende geometrische Eigenschaften:

(a) $\widehat{Y}$ ist orthogonal zu $\widehat{\varepsilon}$, das heiÃt $\langle \widehat{Y}, R\rangle = 0$.
(b) Die Spalten von X sind orthogonal zu den Residuen, d. h. $X^\top R = 0$.

Gilt zusätzlich, dass wir ein Regressionsmodell mit Absolutglied haben, das heißt in der ersten Spalte der Designmatrix X stehen nur Einsen, gilt weiterhin:

(c) Die Residuen sind im Mittel gleich null, das heißt $\sum_{i=1}^n R_i = 0$.
(d) Der arithmetische Mittelwert der $\widehat{Y}_i$ ist gleich dem Mittelwert der Beobachtungen Y_i selbst, das heißt $\sum_{i=1}^n \widehat{Y}_i = \sum_{i=1}^n Y_i$.

2.9 Im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ mit $\beta \in \mathbb{R}^p, \varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ und $\sigma > 0$ bezeichne $\widehat{\beta}$ den Kleinste-Quadrate-Schätzer und $\widehat{\beta}_{\text{ridge}}$ den Ridge-Regressionsschätzer.

(a) Finden Sie für den Ridge-Regressionsschätzer den optimalen Tuning-Parameter λ in Abhängigkeit vom unbekannten $\beta \in \mathbb{R}^p$ und bestimmen Sie das resultierende minimale Risiko $\mathbb{E}\big[|\widehat{\beta}_{\text{ridge}} - \beta|^2\big]$.
(b) Folgern Sie, dass es immer ein $\lambda > 0$ gibt, sodass gilt

$$\mathbb{E}\big[|\widehat{\beta}_{\text{ridge}} - \beta|^2\big] < \mathbb{E}\big[|\widehat{\beta} - \beta|^2\big].$$

2.10 Untersuchen Sie das Verhalten des Ridge-Regressionsschätzers im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, E_n)$ bei $n = 50$ Beobachtungen und $p = 30$ unbekannten Parametern, wenn davon 10 groß und 20 klein sind. Gehen Sie wie folgt vor:

(a) Simulieren Sie die Einträge der Designmatrix $X \in \mathbb{R}^{n\times p}$ als unabhängige $\mathrm{N}(0,1)$-verteilte Zufallsvariablen und erzeugen Sie den Vektor $\beta \in \mathbb{R}^p$ aus 10 $\mathrm{U}([\frac{1}{2}, 1])$-verteilten und 20 $\mathrm{U}([0, \frac{3}{10}])$-verteilten Zufallsvariablen.
(b) Erzeugen Sie in 200 Durchgängen jeweils den Fehlervektor

$$\varepsilon^{(i)} \in \mathbb{R}^{50}, i = 1, \ldots, 200,$$

und berechnen Sie aus den resultierenden Beobachtungen den Ridge-Regressionsschätzer $\widehat{\beta}_\lambda^{(i)}$ für $\lambda \in \{\frac{k}{2} : k = 0, \ldots, 50\}$.
(c) Bestimmen Sie für jedes λ den (empirischen) mittleren quadratischen Fehler $R_\lambda := \frac{1}{200}\sum_{i=1}^{200} |\widehat{\beta}_\lambda^{(i)} - \beta|^2$. Stellen Sie die Abbildung $\lambda \mapsto R_\lambda$ graphisch dar und vergleichen Sie die Fehler der Ridge-Regressionsschätzer mit dem des gewöhnlichen Kleinste-Quadrate-Schätzers.

2.11 Beweisen Sie für die empirische Kovarianzmatrix $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ im Fall unabhängiger p-dimensionaler Zufallsvektoren $X_i \sim \mathrm{N}(0, \Sigma)$:

(a) Es gilt $\mathbb{E}[\Sigma_n] = \Sigma$ (der Erwartungswert einer Matrix ist die Matrix der Erwartungswerte).
(b) Mit der *Frobenius-Norm* $\|M\|_2 = (\sum_{i,j=1}^p M_{i,j}^2)^{1/2}$ und Spur $\mathrm{tr}(M) = \sum_{i=1}^p M_{i,i}$ einer $p \times p$-Matrix M gilt

$$\mathbb{E}[\|\Sigma_n - \Sigma\|_2^2] = n^{-1}(\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_2^2),\ \mathbb{E}[\|\Sigma^{-1/2}(\Sigma_n - \Sigma)\Sigma^{-1/2}\|_2^2] = n^{-1}(p^2 + p),$$

wobei in der zweiten Gleichung Σ als invertierbar angenommen wird.

2.12 Betrachten Sie das lineare Modell $Y_i = X_i\beta + \varepsilon_i, i = 1, \dots, n$, mit i.i.d. $(X_i, Y_i)_{i=1,\dots,n} \subseteq [-R, R]^p \times \mathbb{R}$ für ein $R > 0$ und $\mathbb{E}[\varepsilon_i^2] = \sigma^2 > 0$. Zudem seien X_i und ε_i für alle i unabhängig und $\Sigma_X = \mathbb{E}[X_1 X_1^\top] \in \mathbb{R}^{p\times p}$ sei wohldefiniert und positiv definit. Beweisen Sie für den Kleinste-Quadrate-Schätzer $\widehat{\beta}$ die asymptotische Normalität:

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \mathrm{N}(0, \sigma^2 \Sigma_X^{-1}).$$

2.13 Betrachten Sie das gewöhnliche lineare Modell unter Normalverteilungsannahme $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ mit unbekanntem $\sigma > 0$ und wahrem Parameter $\beta \in \mathbb{R}^p$. Für $v, \beta_0 \in \mathbb{R}^p$ seien $\rho = \langle v, \beta\rangle$ und $\rho_0 = \langle v, \beta_0\rangle$ gegeben. Zeigen Sie, dass der Likelihood-Quotiententest der Hypothese $H_0 \colon \rho = \rho_0, \sigma > 0$ gegen die Alternative $H_1 \colon \rho \neq \rho_0, \sigma > 0$ zum Niveau $\alpha \in (0, 1)$ von der Form

$$\varphi_\alpha = \mathbb{1}\left(|T_{n-p}(Y)| > c_\alpha\right) \quad \text{mit} \quad T_{n-p}(Y) := \frac{\widehat{\rho} - \rho_0}{\widehat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v}}$$

und einem geeigneten kritischen Wert $c_\alpha > 0$ ist.

2.14 Im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ unter der Normalverteilungsannahme $(\varepsilon_1, \dots, \varepsilon_n)^\top \sim \mathrm{N}(0, \sigma^2 E_n)$ mit unbekanntem $\beta \in \mathbb{R}^p$ und $\sigma > 0$ bestimmt der Kovariablenvektor $x_{n+1} \in \mathbb{R}^p$ die zukünftige Beobachtung $Y_{n+1} = \langle x_{n+1}, \beta\rangle + \varepsilon_{n+1}$ mit $\varepsilon_{n+1} \sim \mathrm{N}(0, \sigma^2)$ unabhängig von $(\varepsilon_1, \dots, \varepsilon_n)$. Sei $\alpha \in (0, 1)$.

(a) Konstruieren Sie ein $(1 - \alpha)$-Konfidenzintervall für den zu erwartenden Wert $\langle x_{n+1}, \beta\rangle$.

(b) Konstruieren Sie ein $(1 - \alpha)$-Prognoseintervall für die zu beobachtende Realisierung von Y_{n+1}.

Geben Sie im Modell aus Aufgabe 2 beide Intervalle zum Niveau 0,95 für den Happiness-Score in Deutschland 2020 explizit an.

2.15 Zeigen Sie im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ unter der Normalverteilungsannahme $(\varepsilon_1, \dots, \varepsilon_n)^\top \sim \mathrm{N}(0, \sigma^2 E_n)$ mit unbekanntem $\beta \in \mathbb{R}^k$ und $\sigma > 0$, dass

$$C := \left[(n-k)\widehat{\sigma}^2 / q_{\chi^2(n-k),1-\alpha/2}, (n-k)\widehat{\sigma}^2 / q_{\chi^2(n-k),\alpha/2}\right]$$

ein Konfidenzintervall für σ^2 zum Konfidenzniveau $1 - \alpha$ ist, wobei $\widehat{\sigma}^2 = |Y - X\widehat{\beta}|^2/(n-k)$ die Stichprobenvarianz und $q_{\chi^2(n-k),\tau}$ für $\tau \in (0, 1)$ das τ-Quantil der $\chi^2(n-k)$-Verteilung sind. Nutzen Sie dieses Resultat, um einen χ^2-Test für die Varianz σ^2 zum Niveau $\alpha \in (0, 1)$ zu konstruieren.

2.16 Im gewöhnlichen linearen Modell unter Normalverteilungsannahme $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ mit $\sigma^2 > 0$ soll die lineare Hypothese

$$H_0 : \beta_j = \beta_l \quad \text{gegen} \quad H_1 : \beta_j \neq \beta_l.$$

getestet werden. Zeigen Sie, dass die zugehörige Fisher-Statistik von der Form

$$F = \frac{(\widehat{\beta}_j - \widehat{\beta}_l)^2}{\widehat{\mathrm{Var}}(\widehat{\beta}_j - \widehat{\beta}_l)}$$

und unter H_0 $F(1, n-p)$-verteilt ist, wobei $\widehat{\mathrm{Var}}(\widehat{\beta}_j - \widehat{\beta}_l)$ ein geeigneter Schätzer der Varianz $\mathrm{Var}(\widehat{\beta}_j - \widehat{\beta}_l)$ ist. Warum ist dieser F-Test äquivalent zum (zweiseitigen) t-Test mit der Teststatistik

$$T = \frac{\widehat{\beta}_j - \widehat{\beta}_l}{(\widehat{\mathrm{Var}}(\widehat{\beta}_j - \widehat{\beta}_l))^{1/2}} \overset{H_0}{\sim} t(n-p)?$$

2.17 Führen Sie eine lineare Regressionsanalyse analog zu Beispiel 2.52 für die mittleren Januartemperaturen in Berlin-Dahlem zwischen 1719 und 2020 basierend auf den Messungen des Deutschen Wetterdienstes aus.

2.18 Beweisen Sie im Modell der zweifaktoriellen Varianzanalyse den Satz 2.68.

Kapitel 3
Modellwahl

Am Ende des vorangegangenen Kapitels haben wir eine mögliche Modellmissspezifikation betrachtet, die Daten werden also durch eine Verteilung erzeugt, die gar nicht in dem Modell enthalten ist, das unserem statistischen Verfahren zugrunde liegt. Dies wirft die Frage auf, was überhaupt ein geeignetes Modell ist. In diesem Kapitel werden wir verschiedene Ansätze kennenlernen, die eine versierte Wahl des Modells ermöglichen, nämlich die Informationskritierien AIC und BIC, die Idee der Kreuzvalidierung und die Lasso-Methode. Um diese Methoden zu analysieren, werden wir Orakelungleichungen herleiten.

3.1 Informationskriterien

Im Allgemeinen wird ein statistisches Modell die zugrunde liegenden Zusammenhänge nicht exakt darstellen. Unabhängig davon, ob wir die möglicherweise sehr komplexen Abhängigkeiten kennen oder nicht, ist eine exakte Beschreibung oft gar nicht das Ziel. Stattdessen möchten wir ein möglichst einfaches Modell, das sowohl die gegebenen Beobachtungen als auch neue Daten aus demselben Experiment möglichst gut beschreibt. Bei der Wahl eines Modells müssen daher zwei entgegengesetzt wirkende Probleme ausbalanciert werden:

1. Ist das Modell zu stark vereinfacht, werden Haupteffekte nicht mehr ausreichend erklärt. Dieses sogenannte *underfitting* führt zu einer systematischen Verzerrung/einem Bias.
2. Ist das Modell sehr komplex mit vielen Parametern, dann können die gegebenen Beobachtungen sehr genau darstellt werden, bishin zur Interpolation der Daten. Viele Modellparameter führen jedoch zu einem großen statistischen Fehler und starken Schwankungen, sodass man für neue Beobachtungen keine vernünftige Modellvorhersage mehr erwarten kann. Man spricht von *overfitting*.

Folgendes Beispiel verdeutlicht diese beiden Effekte:

Beispiel 3.1 Für i.i.d. zentrierte Fehler $(\varepsilon_i)_{i=1,\dots,n}$ beobachten wir

$$Y_i = f(x_i) + \varepsilon_i, \quad 1 \leqslant i \leqslant n, \tag{3.1}$$

mit einer unbekannten Funktion $f\colon \mathbb{R}^d \to \mathbb{R}$ und $x_i \in \mathbb{R}^d$, $i = 1, \dots, n$. Wir betrachten nun die empirische Seminorm

$$\|f\|_n := \sqrt{\langle f, f\rangle_n}, \qquad \text{wobei} \qquad \langle f, g\rangle_n := \frac{1}{n}\sum_{i=1}^n f(x_i)g(x_i),$$

auf dem Funktionenraum $\mathscr{F} := \{f\colon \mathbb{R}^d \to \mathbb{R}\}$. Für eine Orthonormalbasis $(e_j)_{j=1,\dots,n}$ dieses Funktionenraumes bezüglich $\langle\cdot,\cdot\rangle_n$ können wir die Regressionsfunktion durch

$$f(x_i) = \sum_{j=1}^n \langle f, e_j\rangle_n e_j(x_i), \qquad i = 1, \dots, n,$$

darstellen. Statt ein allgemeines $f \in \mathscr{F}$ zu betrachten, können wir hoffen, dass die unbekannte Regressionsfunktion f für ein $p \in \{1, \dots, n\}$ bereits durch ein Element aus dem linearen Unterraum $\operatorname{span}\{e_1, \dots, e_p\}$ gut approximiert werden kann. Diese Annahme führt auf das vereinfachte lineare Modell

$$Y_i = \sum_{j=1}^p \beta_j e_j(x_i) + \varepsilon_i, \quad 1 \leqslant i \leqslant n,\ p \in \mathbb{N},$$

sodass die Modellparameter $(\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ durch ein $\widehat{\beta}^{(p)} = (\widehat{\beta}_1, \dots, \widehat{\beta}_p)$ geschätzt werden müssen. Da die Daten jedoch gemäß (3.1) erzeugt werden, liegt hier eine *Modellmissspezifikation* vor (siehe auch Kapitel **??**). Dennoch kann ein resultierender Schätzer $\widehat{f}_{n,p} = \sum_{j=1}^p \widehat{\beta}_j e_j$ eine gute Näherung von f sein.

Wir haben somit verschiedene Modelle mit Parameterdimensionen $p = 1, \dots, n$ zur Auswahl. Wie groß sollen wir p wählen? Ein großes p bedeutet, dass der (quadrierte) Bias $\|f - \sum_{j=1}^p \beta_j e_j\|_n^2 = \sum_{j=p+1}^n \langle f, e_j\rangle^2$ klein ist. Gleichzeitig führt die große Parameterdimension zu großen stochastischen Fehlern, denn wir müssen p Koeffizienten schätzen, siehe (2.3) und Aufgabe 2.7. Wir sollten also p so wählen, dass Bias und stochastischer Fehler annähernd gleich sind.

In diesem Kapitel wollen wir Methoden entwickeln, die automatisch und anhand der gegeben Daten ein geeignetes Modell wählen. Hierzu untersuchen wir Daten aus einem Wahrscheinlichkeitsraum $(\mathcal{X}, \mathcal{F}, \mathbb{P})$, wobei wir das zugrunde liegende Wahrscheinlichkeitsmaß $\mathbb{P}$ nicht kennen. Wie in Beispiel 3.1 illustriert, modellieren wir die Beobachtungen durch $P \in \mathbb{N}$ verschiedene parametrische Modelle

$$(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta_p}),$$

wobei die Dimension $p \in \{1, ..., P\}$ des Parameterraums Θ_p wächst und $P \in \mathbb{N}$ die maximal betrachtete Parameterdimension ist. Innerhalb des p-ten Modells liefert

uns die Maximum-Likelihood-Methode einen Schätzer $\widehat{\vartheta}_p$. Unser Ziel ist, ein $\widehat{p}$ zu finden, sodass $\mathbb{P}_{\widehat{\vartheta}_{\widehat{p}}}$ die wahre Verteilung $\mathbb{P}$ gut approximiert. Man beachte dabei, dass $\mathbb{P}$ in keiner der Familien $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta_p}$ liegen muss.

Um zu entscheiden, ob ein Kandidat $\mathbb{P}_{\widehat{\vartheta}_{\widehat{p}}}$ die unbekannte Verteilung $\mathbb{P}$ gut approximiert, benötigen wir ein Abstandsmaß zwischen den Wahrscheinlichkeitsverteilungen.

Definition 3.2 Für Wahrscheinlichkeitsmaße $\mathbb{P}$ und $\mathbb{Q}$ auf $(\mathcal{X}, \mathcal{F})$ heißt

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \begin{cases} \int_{\mathcal{X}} \log(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x))\mathbb{P}(\mathrm{d}x), & \text{falls } \mathbb{P} \ll \mathbb{Q}, \\ \infty, & \text{sonst,} \end{cases}$$

Kullback-Leibler-Divergenz von $\mathbb{P}$ bezüglich $\mathbb{Q}$.

Besitzen $\mathbb{P}$ und $\mathbb{Q}$ μ-Dichten p und q, so berechnet man $\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \int_{\{q>0\}} \log(p/q)p\mathrm{d}\mu$ im Fall $\mathbb{P} \ll \mathbb{Q}$. Die Kullback-Leibler-Divergenz ist keine Metrik auf dem Raum der Wahrscheinlichkeitsmaße (es gilt weder Symmetrie noch Dreiecksungleichung, vergleiche Aufgabe 4.1), jedoch besitzt sie nützliche Eigenschaften, um verschiedene Wahrscheinlichkeitsmaße zu vergleichen.

Lemma 3.3 (Eigenschaften der Kullback-Leibler-Divergenz) *Für Wahrscheinlichkeitsmaße $\mathbb{P}$ und $\mathbb{Q}$ auf $(\mathcal{X}, \mathcal{F})$ gilt:*

(i) $\mathrm{KL}(\mathbb{P}|\mathbb{Q}) \geqslant 0$ *und* $\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = 0$ *gilt genau dann, wenn* $\mathbb{P} = \mathbb{Q}$.

(ii) $\mathrm{KL}(\mathbb{P}^{\otimes n}|\mathbb{Q}^{\otimes n}) = n\,\mathrm{KL}(\mathbb{P}|\mathbb{Q})$ *für alle* $n \in \mathbb{N}$.

(iii) Bildet $(\mathbb{P}_\eta)_{\eta\in\Xi}$ *eine natürliche p-parametrische Exponentialfamilie der Form*

$$\frac{\mathrm{d}\mathbb{P}_\eta}{\mathrm{d}\mu}(x) = c(x)\exp\big(\langle\eta, T(x)\rangle - \zeta(\eta)\big)$$

und liegt η_0 im Inneren von Ξ, so ist

$$\mathrm{KL}(\mathbb{P}_{\eta_0}|\mathbb{P}_\eta) = \zeta(\eta) - \zeta(\eta_0) - \langle\nabla\zeta(\eta_0), \eta - \eta_0\rangle, \qquad \eta \in \Xi.$$

(iv) Für unabhängige und gemäß $\mathbb{P}$ identisch verteilte Beobachtungen $(X_i)_{i\in\mathbb{N}}$ mit Werten in $(\mathcal{X}, \mathcal{F})$ und Loglikelihood-Funktion $l(\vartheta) := \log\frac{\mathrm{d}\mathbb{P}_\vartheta^{X_1}}{\mathrm{d}\mu}$ in einem statistischen Modell bezüglich einem dominierenden Maß μ auf $(\mathcal{X}, \mathcal{F})$ gilt

$$-\frac{1}{n}\sum_{i=1}^{n} l(\vartheta, X_i) \overset{\mathbb{P}-f.s.}{\longrightarrow} -\mathbb{E}_{\mathbb{P}}[l(\vartheta, X_1)] = \mathrm{KL}(\mathbb{P}|\mathbb{P}_\vartheta^{X_1}) - \mathrm{KL}(\mathbb{P}|\mu) \quad \text{für } n \to \infty,$$

sofern die Kullback-Leibler-Divergenzen (für allgemeine Maße μ analog definiert) endlich sind.

Beweis

(i) Ohne Einschränkung können wir $\mathbb{P} \ll \mathbb{Q}$ annehmen, da andernfalls $\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \infty > 0$ stets erfüllt ist. Dann gilt

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \int_{\mathcal{X}} \underbrace{\log\Big(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}}_{=:f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)} \mathrm{d}\mathbb{Q}$$

für $f(x) = x\log(x)$ mit $f(0) = 0$. Wegen $f''(x) = \frac{1}{x} > 0$ ist f strikt konvex. Die Jensen-Ungleichung liefert somit

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \int_{\mathcal{X}} f\Big(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big)\mathrm{d}\mathbb{Q} \geqslant f\Big(\int_{\mathcal{X}} \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\mathrm{d}\mathbb{Q}\Big) = f(1) = 0.$$

Hierbei gilt genau dann Gleichheit, wenn $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$ $\mathbb{Q}$-f.s. konstant ist.

(ii) Wir können wieder $\mathbb{P} \ll \mathbb{Q}$ annehmen. Da die Radon-Nikodym-Dichte von Produktmaßen gleich dem Produkt der Randdichten ist, folgt

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}^{\otimes n}\middle|\mathbb{Q}^{\otimes n}\right) &= \int_{\mathcal{X}^n} \log\Big(\frac{\mathrm{d}\mathbb{P}^{\otimes n}}{\mathrm{d}\mathbb{Q}^{\otimes n}}\Big)\mathrm{d}\mathbb{P}^{\otimes n} \\ &= \sum_{i=1}^{n} \int_{\mathcal{X}} \log\Big(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big)\mathrm{d}\mathbb{P} = n\,\mathrm{KL}(\mathbb{P}|\mathbb{Q}).\end{aligned}$$

(iii) Einsetzen der μ-Dichten liefert

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\eta_0}|\mathbb{P}_\eta) &= \mathbb{E}_{\eta_0}\Big[\log\Big(\frac{\mathrm{d}\mathbb{P}_{\eta_0}/\mathrm{d}\mu}{\mathrm{d}\mathbb{P}_\eta/\mathrm{d}\mu}\Big)\Big] \\ &= \mathbb{E}_{\eta_0}\Big[\log\Big(\frac{\mathrm{d}\mathbb{P}_{\eta_0}}{\mathrm{d}\mu}\Big) - \log\Big(\frac{\mathrm{d}\mathbb{P}_\eta}{\mathrm{d}\mu}\Big)\Big] \\ &= \mathbb{E}_{\eta_0}\big[\langle\eta_0, T\rangle - \zeta(\eta_0) - \big(\langle\eta, T\rangle - \zeta(\eta)\big)\big] \\ &= \zeta(\eta) - \zeta(\eta_0) - \langle\eta - \eta_0, \mathbb{E}_{\eta_0}[T]\rangle.\end{aligned}$$

Schließlich gilt $\nabla\zeta(\eta_0) = \mathbb{E}_{\eta_0}[T]$ nach Satz **??**(iii).

(iv) Unter den Annahmen $\mathrm{KL}(\mathbb{P}|\mathbb{P}_\vartheta) < \infty$ und $\mathrm{KL}(\mathbb{P}|\mu) < \infty$ gilt

$$\begin{aligned}\mathrm{KL}(\mathbb{P}|\mathbb{P}_\vartheta) - \mathrm{KL}(\mathbb{P}|\mu) &= \int_{\mathcal{X}} \Big(\log\Big(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}_\vartheta}\Big) - \log\Big(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}\Big)\Big)\mathrm{d}\mathbb{P} \\ &= -\int_{\mathcal{X}} \log\Big(\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}\Big)\mathrm{d}\mathbb{P}.\end{aligned}$$

Damit ist $l(\vartheta) = \log(\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}) \in \mathcal{L}^1(\mathbb{P})$, und das starke Gesetz der großen Zahlen liefert für $\mathbb{P}$-verteilte, unabhängige (X_i)

$$-\frac{1}{n}\sum_{i=1}^{n} l(\vartheta, X_i) \overset{\mathbb{P}-f.s.}{\longrightarrow} -\mathbb{E}_{\mathbb{P}}\big[l(\vartheta, X_1)\big] = \mathrm{KL}(\mathbb{P}|\mathbb{P}_\vartheta) - \mathrm{KL}(\mathbb{P}|\mu)$$

für $n \to \infty$. □

Die Kullback-Leibler-Divergenz von Normalverteilungen ergibt sich als Spezialfall von (iii):

Korollar 3.4 *Für $\mu_0, \mu \in \mathbb{R}^p$ und eine symmetrische, positiv-definite Matrix $\Sigma \in \mathbb{R}^{p\times p}$ gilt*

$$\mathrm{KL}(\mathrm{N}(\mu_0, \Sigma)|\mathrm{N}(\mu, \Sigma)) = \frac{1}{2}\left|\Sigma^{-1/2}(\mu - \mu_0)\right|^2 = \frac{1}{2}\langle\Sigma^{-1}(\mu - \mu_0),\, \mu - \mu_0\rangle.$$

Beweis Wir verwenden (iii) aus dem vorangegangenen Lemma, wobei $\zeta(\mu) = \frac{1}{2}\langle\Sigma^{-1}\mu, \mu\rangle$. Daraus folgt

$$\begin{aligned}\mathrm{KL}(\mathrm{N}(\mu_0, \Sigma)|\mathrm{N}(\mu, \Sigma)) &= \frac{1}{2}\Big(\langle\Sigma^{-1}\mu, \mu\rangle - \langle\Sigma^{-1}\mu_0, \mu_0\rangle - 2\langle\Sigma^{-1}(\mu - \mu_0), \mu_0\rangle\Big)\\ &= \frac{1}{2}\Big(\langle\Sigma^{-1}\mu, \mu\rangle + \langle\Sigma^{-1}\mu_0, \mu_0\rangle - 2\langle\Sigma^{-1}\mu, \mu_0\rangle\Big)\\ &= \frac{1}{2}\langle\Sigma^{-1}(\mu - \mu_0), \mu - \mu_0\rangle.\end{aligned}$$

Der letzte Ausdruck ist gerade gleich $\frac{1}{2}|\Sigma^{-1/2}(\mu - \mu_0)|^2$. □

Beispiel 3.5 Wir setzen Beispiel 3.1 unter der Annahme von i.i.d. Beobachtungsfehlern $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ fort. Der Vektor der Beobachtungen $Y_i = f(x_i) + \varepsilon_i, i = 1, \dots n$, mit unbekannter Regressionsfunktion $f\colon \mathbb{R}^d \to \mathbb{R}$ ist dann gemäß $\mathbb{P} = \mathrm{N}(\mu, \sigma^2 E_n)$ verteilt mit Mittelwertvektor $\mu = (f(x_1), \dots, f(x_n))^\top \in \mathbb{R}^n$. In Modell p ist die Verteilungsfamilie gegeben durch $(\mathbb{P}_{\beta^{(p)}})_{\beta^{(p)}\in\mathbb{R}^p}$ mit $\mathbb{P}_{\beta^{(p)}} = \mathrm{N}(X^{(p)}\beta^{(p)}, \sigma^2 E_n)$ und der Designmatrix $X^{(p)} = (e_j(x_i))_{i=1,\dots,n;j=1,\dots,p} \in \mathbb{R}^{n\times p}$. Die Kullback-Leibler-Divergenz zwischen $\mathbb{P}$ und $\mathbb{P}_{\beta^{(p)}}$ ist damit gegeben durch

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}|\mathbb{P}_{\beta^{(p)}}\right) &= \frac{1}{2\sigma^2}\left|\mu - X^{(p)}\beta^{(p)}\right|^2\\ &= \frac{1}{2\sigma^2}\sum_{i=1}^n\Big(f(x_i) - \sum_{j=1}^p \beta_j^{(p)} e_j(x_i)\Big)^2 = \frac{n}{2\sigma^2}\Big\|f - \sum_{j=1}^p \beta_j^{(p)} e_j\Big\|_n^2.\end{aligned}$$

Der „Abstand“ zwischen $\mathbb{P}$ und der bestmöglichen Wahl in $(\mathbb{P}_{\beta^{(p)}})_{\beta^{(p)}\in\mathbb{R}^p}$ ist also klein, wenn f gut durch die ersten p Basisfunktionen approximiert werden kann. Als Kriterium zur Wahl von p können wir dies allerdings nicht direkt einsetzen, da f und damit auch $\mathrm{KL}(\mathbb{P}|\mathbb{P}_{\beta^{(p)}})$ unbekannt sind.

Für die Konvergenz in Lemma 3.3(iv) müssen wir nicht annehmen, dass das Wahrscheinlichkeitsmaß $\mathbb{P}$, unter dem die Beobachtungen erzeugt werden, in der Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ enthalten ist. Wir benötigen lediglich $\mathrm{KL}(\mathbb{P}|\mathbb{P}_\vartheta) < \infty$ für $\vartheta \in \Theta$. Daher erhalten wir eine interessante Interpretation des *Maximum-Likelihood-Schätzers für missspezifizierte Modelle*: Während die linke Seite (die negative log-likelihood, skaliert mit $1/n$)

$$-\frac{1}{n}\sum_{i=1}^{n} l(\vartheta, X_i)$$

im Modell $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ durch den Maximum-Likelihood-Schätzer $\widehat{\vartheta}$ minimiert wird, ist der Grenzwert

$$\mathrm{KL}(\mathbb{P}|\mathbb{P}_\vartheta) - \mathrm{KL}(\mathbb{P}|\mu)$$

unter demjenigen Parameter ϑ kleinstmöglich, der die Kullback-Leibler-Divergenz $\mathrm{KL}(\mathbb{P}|\mathbb{P}_\vartheta)$ zum wahren $\mathbb{P}$ minimiert. Dieser Wert ist damit ein natürlicher Kandidat für einen Grenzwert des Maximum-Likelihood-Schätzers. Zumindest für große Stichprobenumfänge können wir hoffen, dass $\widehat{\vartheta}_p$ nahe am Minimierer von $\mathrm{KL}(\mathbb{P}|\mathbb{P}_\vartheta)$ liegt.

3.1.1 Akaike-Informationskriterium (AIC)

Es sei $X \in \mathcal{X}$ eine Beobachtung aus dem Wahrscheinlichkeitsraum $(\mathcal{X}, \mathcal{F}, \mathbb{P})$. Wir betrachten dominierte statistische Modelle $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_{\vartheta^{(p)}})_{\vartheta^{(p)}\in\Theta_p})$ mit $\Theta_p \subseteq \mathbb{R}^p$ mit nichtleerem Inneren (die Parametermenge Θ_p ist p-dimensional).

Ohne Einschränkung können wir annehmen, dass alle Modelle durch dasselbe Maß μ dominiert werden und auch $\mathbb{P}$ absolutstetig bezüglich μ ist. Im Modell $(\mathbb{P}_{\vartheta^{(p)}})_{\vartheta^{(p)}\in\Theta_p}$ bezeichnen wir die Likelihood-Funktionen bezüglich μ mit $L_p(\vartheta^{(p)}, x)$. Der Maximum-Likelihood-Schätzer $\widehat{\vartheta}^{(p)}$ im p-ten Modell wird unter der Annahme $X \sim \mathbb{P}_{\vartheta^{(p)}}$ für ein $\vartheta^{(p)} \in \Theta_p$ bestimmt:

$$\widehat{\vartheta}^{(p)} \in \operatorname*{arg\,max}_{\vartheta^{(p)}\in\Theta_p} L_p(\vartheta^{(p)}, X).$$

Nachdem $\mathbb{P}_{\widehat{\vartheta}^{(p)}}$ ein guter Kandidat innerhalb des p-ten Modells ist, wollen wir ein geeignetes Modell auswählen, indem wir die Kullback-Leibler-Divergenz $\mathrm{KL}(\mathbb{P}|\mathbb{P}_{\widehat{\vartheta}^{(p)}})$ über p minimieren. Da wir $\mathbb{P}$ nicht kennen, ist jedoch $\mathrm{KL}(\mathbb{P}|\mathbb{P}_{\widehat{\vartheta}^{(p)}})$ nicht zugänglich und muss geschätzt werden. Als reelles Funktional wird $\mathrm{KL}(\mathbb{P}|\mathbb{P}_{\widehat{\vartheta}^{(p)}})$ dabei mit kleinerem Fehler schätzbar sein als $\widehat{\vartheta}^{(p)}$ selbst. Hierfür verwenden wir im p-ten Modell für $\vartheta^{(p)} \in \Theta_p$ die Darstellung (siehe Lemma 3.3)

$$\mathrm{KL}(\mathbb{P}|\mathbb{P}_{\vartheta^{(p)}}) = \mathrm{KL}(\mathbb{P}|\mu) - \mathbb{E}_{\mathbb{P}}\Big[\log L_p(\vartheta^{(p)}, X)\Big],$$

sofern der letzte Erwartungswert existiert. In diesem Fall bezeichnen wir mit

$$d_p(\vartheta^{(p)}) := -2\mathbb{E}_{\mathbb{P}}[\log L_p(\vartheta^{(p)}, X)], \qquad \vartheta^{(p)} \in \Theta_p,$$

die *Kullback-Leibler-Diskrepanz*.

Da $\mathrm{KL}(\mathbb{P}|\mu)$ unabhängig von ϑ ist, ist die Minimierung von $p \mapsto \mathrm{KL}(\mathbb{P}|\mathbb{P}_{\widehat{\vartheta}^{(p)}})$ äquivalent zur Minimierung von $p \mapsto d_p(\widehat{\vartheta}^{(p)})$. Auch $d_p(\vartheta)$ ist jedoch unbekannt.

Dessen empirische Version ist gegeben durch $-2\log L_p(\vartheta^{(p)}, X)$, da X gemäß $\mathbb{P}$ verteilt ist. Setzen wir in $-\log L_p(\vartheta^{(p)}, X)$ den Maximum-Likelihood-Schätzer $\widehat{\vartheta}^{(p)} = \widehat{\vartheta}^{(p)}(X)$ ein, erhalten wir aufgrund der Abhängigkeit zwischen $\widehat{\vartheta}^{(p)}$ und X in der Regel keinen erwartungstreuen Schätzer von $d_p(\widehat{\vartheta}^{(p)})$. Stattdessen ist zu vermuten, dass $d_p(\widehat{\vartheta}^{(p)})$ unterschätzt wird, denn $\widehat{\vartheta}^{(p)}$ minimiert gerade die Funktion $\vartheta^{(p)} \mapsto -\log L_p(\vartheta^{(p)}, X)$.

Beispiel 3.6 (Signalerkennung) Für eine P-dimensionale mathematische Stichprobe $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\vartheta,\, E_P)$ betrachten wir die Modelle

$$\Theta_p := \big\{\vartheta \in \mathbb{R}^P \,\big|\, \forall j > p : \vartheta_j = 0\big\}, \qquad p = 1, \dots, P,$$

für den unbekannten Mittelwertvektor. Im p-ten Modell ist der Maximum-Likelihood-Schätzer durch das Stichprobenmittel in den ersten p Koordinaten gegeben, während die letzten $P - p$ Koordinaten null sind:

$$\widehat{\vartheta}^{(p)} = (\widehat{\vartheta}_1, \dots, \widehat{\vartheta}_p, 0, \dots, 0) \quad \text{wobei} \quad \widehat{\vartheta}_j := (\overline{X}_n)_j = \frac{1}{n}\sum_{i=1}^n X_{i,j}$$

Schreiben wir $|t|^2_{\leqslant p} := \sum_{j=1}^p t_j^2$ und $|t|^2_{>p} := \sum_{j=p+1}^P t_j^2$ für $t \in \mathbb{R}^P$, dann gilt im p-ten Modell für $\vartheta^{(p)} \in \Theta_p$

$$\begin{aligned} -2\log L_p(\vartheta^{(p)},\, X) &= -2\sum_{i=1}^n \Big(\log((2\pi)^{-\frac{P}{2}}) - \frac{|X_i - \vartheta^{(p)}|^2_{\leqslant p} + |X_i|^2_{>p}}{2} \Big) \\ &= \sum_{i=1}^n \Big(|X_i - \vartheta^{(p)}|^2_{\leqslant p} + |X_i|^2_{>p} + P\log(2\pi) \Big). \end{aligned}$$

Unter der Annahme $X \sim \mathbb{P} = \mathrm{N}(\vartheta_0,\, E_P)$ für ein $\vartheta_0 \in \mathbb{R}^P$ erhalten wir mit der Bias-Varianz-Zerlegung

$$\begin{aligned} d_p(\vartheta^{(p)}) = -2\mathbb{E}_{\mathbb{P}}[\log L_p(\vartheta^{(p)}, X)] &= \sum_{i=1}^n \Big(\mathbb{E}_{\mathbb{P}}\big[|X_i - \vartheta^{(p)}|^2_{\leqslant p} + |X_i|^2_{>p} \big] + P\log(2\pi) \Big) \\ &= n\Big(|\vartheta_0 - \vartheta^{(p)}|^2_{\leqslant p} + |\vartheta_0|^2_{>p} + P + P\log(2\pi) \Big). \end{aligned}$$

Die Kullback-Leibler-Diskrepanz d_p spiegelt also wider, wie gut ϑ_0 durch einen Vektor aus Θ_p approximiert werden kann: Auch wenn die Modellannahme $\vartheta_0 \in \Theta_p$ nicht erfüllt ist, kann Θ_p ein gutes Modell sein, wenn alle Einträge von ϑ_0 für große Indizes klein sind, das heißt wenn $|\vartheta_0|_{>p}$ nahe 0 ist.

Setzen wir den Maximum-Likelihood-Schätzer $\widehat{\vartheta}^{(p)}$ ein, erhalten wir

$$-2\log L_p(\widehat{\vartheta}^{(p)},\, X) = \sum_{i=1}^n \Big(|X_i - \overline{X}_n|^2_{\leqslant p} + |X_i|^2_{>p} + P\log(2\pi) \Big),$$

und aufgrund von $\mathbb{E}[|X_i - \overline{X}|^2_{\leqslant p}] = p\frac{n-1}{n}$ gilt

$$\begin{aligned} -2\mathbb{E}_{\mathbb{P}}\big[\log L_p(\widehat{\vartheta}^{(p)}, X)\big] &= n\Big(p\frac{n-1}{n} + |\vartheta_0|^2_{>p} + (P-p) + P\log(2\pi)\Big) \\ &= n\Big(-\frac{p}{n} + |\vartheta_0|^2_{>p} + P + P\log(2\pi)\Big). \end{aligned}$$

Anderseits gilt

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}[d_p(\widehat{\vartheta}^{(p)})] &= n\Big(\mathbb{E}_{\mathbb{P}}\big[|\vartheta^0 - \overline{X}_n|^2_{\leqslant p}\big] + |\vartheta_0|^2_{>p} + P + P\log(2\pi)\Big) \\ &= n\Big(\frac{p}{n} + |\vartheta_0|^2_{>p} + P + P\log(2\pi)\Big). \end{aligned}$$

Daraus folgt

$$\mathbb{E}_{\mathbb{P}}\big[-2\log L_p(\widehat{\vartheta}^{(p)}, X)\big] - \mathbb{E}_{\mathbb{P}}[d(\widehat{\vartheta}^{(p)})] = -2p,$$

sodass $-2\log L_p(\widehat{\vartheta}^{(p)}, X)$ die zu schätzende Größe $d(\widehat{\vartheta}^{(p)})$ systematisch um $-2p$ unterschätzt.

Deutlich allgemeiner weisen wir den Zusammenhang $\mathbb{E}_{\mathbb{P}}[-2\log L(\widehat{\vartheta}_p(X), X)] = \mathbb{E}_{\mathbb{P}}[d(\widehat{\vartheta}_p)] - 2p$ in Satz 3.8 für das lineare Modell nach. Eine Biaskorrektur führt auf Akaikes Informationskriterium.

Methode 3.7 (Modellwahl durch AIC) Für die Modelle $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta_p})$ mit $\Theta_p \subseteq \mathbb{R}^p$ und $p = 1, \dots, P$ ist das **Akaike-Informationskriterium** (kurz: AIC) definiert als

$$\mathrm{AIC}(p) := -2\log L_p(\widehat{\vartheta}^{(p)}, X) + 2p$$

für die Maximum-Likelihood-Schätzer $\widehat{\vartheta}^{(p)}$ im Modell $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta_p}$. Das Modell $\widehat{p}$ wird als Minimierer $\widehat{p} \in \arg\min_{p=1,\dots,P} \mathrm{AIC}(p)$ gewählt und der resultierende, gemäß AIC ausgewählte Schätzer ist $\widehat{\vartheta}^{(\widehat{p})}$.

Kurzbiografie (Hirotugu Akaike) Hirotugu Akaike wurde 1927 in der Präfektur Shizuoka in Japan geboren. Er studierte an der Universität Tokio und promovierte dort 1961 am Institut für Statistische Mathematik. Akaike blieb auch in seiner weiteren wissenschaftlichen Laufbahn an diesem Institut, dessen Direktor er später wurde. Er schrieb wesentliche Arbeiten zur Theorie der Zeitreihen. Sein bekanntester Beitrag ist die Einführung des *Informationskriteriums* AIC im Jahr 1973, das später nach ihm benannt wurde. Akaike starb 2009.

Im Folgenden untersuchen wir, wie sich das Akaike-Informationskriterium auf das lineare Modell anwenden lässt.

Satz 3.8 (AIC im Gaußschen linearen Modell) *Gegeben sei das wahre lineare Modell*

$$Y = \mu + \varepsilon \qquad \textit{mit} \qquad \mu \in \mathbb{R}^n, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$$

unter $\mathbb{P}$ mit bekannter Varianz $\sigma^2 > 0$ sowie für $p = 1, \dots, P$ mit $P \leqslant n$ die Modelle

$$Y = X^{(p)}\beta^{(p)} + \varepsilon \qquad \text{mit} \qquad \beta^{(p)} \in \mathbb{R}^p, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$$

und die Designmatrix $X^{(p)} \in \mathbb{R}^{n\times p}$ mit vollem Rang p. Dann gilt:

(i) Der Maximum-Likelihood-Schätzer in Modell p ist der Kleinste-Quadrate-Schätzer

$$\widehat{\beta}^{(p)} = (X^{(p)\top}X^{(p)})^{-1}X^{(p)\top}Y.$$

(ii) Das Akaike-Informationskriterium ist gegeben durch

$$\mathrm{AIC}(p) = n\log(2\pi\sigma^2) + \frac{|Y - X^{(p)}\widehat{\beta}^{(p)}|^2}{\sigma^2} + 2p.$$

(iii) $\mathrm{AIC}(p)$ *ist ein erwartungstreuer Schätzer der Kullback-Leibler-Diskrepanz:*

$$\mathbb{E}_{\mathbb{P}}[\mathrm{AIC}(p)] = \mathbb{E}_{\mathbb{P}}[d_p(\widehat{\beta}^{(p)})].$$

Beweis

(i) Da die Loglikelihood-Funktion von der Form $\log L_p(\beta^{(p)}, Y) = \log((2\pi\sigma^2)^{-\frac{n}{2}}) - \frac{1}{2\sigma^2}|Y - X^{(p)}\beta^{(p)}|^2$ ist, entspricht der Maximum-Likelihood-Schätzer dem Kleinste-Quadrate-Schätzer, der in Lemma 2.14 explizit berechnet wurde.

(ii) Einsetzen in die Definition von 3.7 liefert unmittelbar

$$\mathrm{AIC}(p) = n\log(2\pi\sigma^2) + \sigma^{-2}|Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + 2p.$$

(iii) Wegen $d_p(\beta^{(p)}) = \mathbb{E}[-2\log L_p(\beta^{(p)}, Y)]$ folgt aus der Bias-Varianz-Zerlegung

$$\begin{aligned} d_p(\beta^{(p)}) &= n\log(2\pi\sigma^2) + \sigma^{-2}\mathbb{E}[|Y - X^{(p)}\beta^{(p)}|^2] \\ &= n\log(2\pi\sigma^2) + \sigma^{-2}\big(|\mu - X^{(p)}\beta^{(p)}|^2 + \underbrace{\mathbb{E}[|\varepsilon|^2]}_{=\sigma^2 n}\big) \\ &= n\big(\log(2\pi\sigma^2) + 1\big) + \sigma^{-2}|\mu - X^{(p)}\beta^{(p)}|^2. \end{aligned}$$

Einsetzen von $X^{(p)}\widehat{\beta}^{(p)} = \Pi^{(p)}Y = \Pi^{(p)}\mu + \Pi^{(p)}\varepsilon$ mit der Orthogonalprojektion $\Pi^{(p)} := \Pi_{X^{(p)}}$ liefert

$$\begin{aligned} \mathbb{E}[d_p(\widehat{\beta}^{(p)})] &= n(\log(2\pi\sigma^2) + 1) + \sigma^{-2}\mathbb{E}[|\mu - X^{(p)}\widehat{\beta}^{(p)}|^2] \\ &= n(\log(2\pi\sigma^2) + 1) + \sigma^{-2}\Big(\big|\mu - \Pi^{(p)}\mu\big|^2 + \mathbb{E}\big[\big|\Pi^{(p)}\varepsilon\big|^2\big]\Big) \\ &= n(\log(2\pi\sigma^2) + 1) + \sigma^{-2}|(E_n - \Pi^{(p)})\mu|^2 + p, \end{aligned}$$

da $\Pi^{(p)}$ Rang p hat und damit $\mathbb{E}[|\Pi^{(p)}\varepsilon|^2] = \sigma^2 p$ gilt. Analog ist $E_n - \Pi^{(p)}$ die orthogonale Projektion auf einen $(n-p)$-dimensionalen Unterraum und $\mathbb{E}[|(E_n - \Pi^{(p)})\varepsilon|^2] = (n-p)\sigma^2$. Wir erhalten

$$\begin{aligned}\mathbb{E}[\mathrm{AIC}(p)] &= n\log(2\pi\sigma^2) + \sigma^{-2}\mathbb{E}\big[|Y - X^{(p)}\widehat{\beta}^{(p)}|^2\big] + 2p \\ &= n\log(2\pi\sigma^2) + \sigma^{-2}\big(|\mu - \Pi^{(p)}\mu|^2 + \mathbb{E}\big[|\varepsilon - \Pi^{(p)}\varepsilon|^2\big]\big) + 2p \\ &= \mathbb{E}[d_p(\widehat{\beta}^{(p)})].\end{aligned}$$

Damit ist die Erwartungstreue von $\mathrm{AIC}(p)$ gezeigt. □

Die Darstellung von $\mathrm{AIC}(p)$ erinnert an die Bias-Varianz-Zerlegung. Durch Multiplikation mit σ^2 (was für die Minimierung von $\mathrm{AIC}(p)$ nichts ändert) erhalten wir die quadrierten Residuen $|Y - X^{(p)}\widehat{\beta}^{(p)}|^2$ als Maß für die Modellgüte sowie den doppelten Varianzterm $2p\sigma^2$. Die Modellwahl ist gegeben durch

$$\begin{aligned}\widehat{p} &\in \arg\min_p \left(n\log(2\pi\sigma^2) + \frac{|Y - X^{(p)}\widehat{\beta}^{(p)}|^2}{\sigma^2} + 2p\right) \\ &= \arg\min_p \underbrace{\left(|Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + 2p\sigma^2\right)}_{\mathrm{AIC}_{\mathrm{LM}}(p)}.\end{aligned}$$

Hierbei fällt der empirische Verlust $|Y - X^{(p)}\widehat{\beta}^{(p)}|^2$ in $\mathrm{AIC}_{\mathrm{LM}}(p)$ mit wachsender Dimension p, während der Strafterm $2p\sigma^2$ in p wächst. $\widehat{p}$ balanciert also die Güte der Datenanpassung mit der Komplexität des Modells aus. Man beachte, dass die Erwartungstreue von $\mathrm{AIC}(p)$ allein nicht sicherstellt, dass $\widehat{p}$ tatsächlich ein gutes Modell bzw. $X^{(\widehat{p})}\widehat{\beta}^{(\widehat{p})}$ eine gute Vorhersage für μ ist. Die Qualität von $X^{(\widehat{p})}\widehat{\beta}^{(\widehat{p})}$ beurteilen wir in Kapitel 3.2 anhand einer Orakelungleichung.

Korollar 3.9 (Unverzerrte Risikoschätzung) *In der Situation von Satz 3.8 ist* $|Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + 2p\sigma^2 - n\sigma^2$ *eine erwartungstreue Schätzung des quadratischen Vorhersagefehlers* $|\mu - X^{(p)}\widehat{\beta}^{(p)}|^2$*:*

$$\mathbb{E}_{\mathbb{P}}\big[|Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + 2p\sigma^2 - n\sigma^2\big] = \mathbb{E}_{\mathbb{P}}\big[|\mu - X^{(p)}\widehat{\beta}^{(p)}|^2\big]$$

Insbesondere kann das Akaike-Informationskriterium als „unbiased risk estimation"-Kriterium interpretiert werden.

Beweis Es gilt für die quadrierten Residuen wie oben gesehen, dass

$$\mathbb{E}_{\mathbb{P}}\big[|Y - X^{(p)}\widehat{\beta}^{(p)}|^2\big] = |\mu - \Pi^{(p)}\mu|^2 + (n-p)\sigma^2.$$

Mit der Bias-Varianz-Zerlegung folgt

$$\mathbb{E}_{\mathbb{P}}[|\mu - X^{(p)}\widehat{\beta}^{(p)}|^2] = |\mu - \Pi^{(p)}\mu|^2 + p\sigma^2.$$

Ein Vergleich der beiden Erwartungswerte liefert die Behauptung. □

Bemerkung 3.10 (Mallows' C_p-Kriterium) Der Beweis von Korollar 3.9 zeigt, dass für die erwartungstreue Fehlerschätzung die Normalverteilungsannahme unerheblich ist. Es genügt, $\mathbb{E}[\varepsilon] = 0$ und $\mathbb{C}\mathrm{ov}(\varepsilon) = \sigma^2 E_n$ vorauszusetzen. Das Modell allgemein

aufgrund einer erwartungstreuen Risikoschätzung auszuwählen, ist der Ansatz von *Mallows' C_p-Kriterium*, wobei

$$C_p := |Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + 2p\sigma^2 - n\sigma^2$$

gilt. Bis auf einen von p unabhängigen Summanden stimmt Mallows' C_p-Kriterium im linearen Modell mit AIC überein.

Wir wollen nun Satz 3.8 auf den Fall einer unbekannten Varianz erweitern:

Satz 3.11 (AIC im linearen Modell mit unbekannter Varianz) *Gegeben sei das wahre lineare Modell*

$$Y = \mu + \varepsilon \qquad \textit{mit} \qquad \mu \in \mathbb{R}^n, \quad \varepsilon \sim \mathrm{N}(0,\, \sigma_0^2 E_n)$$

unter $\mathbb{P}$ sowie für $p = 1, \dots, P$ mit $P \leqslant n-3$ die Modelle

$$Y = X^{(p)}\beta^{(p)} + \varepsilon \qquad \textit{mit} \qquad \beta^{(p)} \in \mathbb{R}^p, \quad \varepsilon \sim \mathrm{N}(0,\, \sigma^2 E_n)$$

und der Designmatrix $X^{(p)} \in \mathbb{R}^{n\times p}$ mit vollem Rang p. Für $k = p+1$ ist $\vartheta^{(k)} = (\beta^{(p)}, \sigma^2)$ der unbekannte Parameter in $\mathbb{R}^k$. Dann gilt:

(i) Der Maximum-Likelihood-Schätzer ist $\widehat{\vartheta}^{(k)} = (\widehat{\beta}^{(p)}, \widehat{\sigma}_p^2)$ mit dem Kleinste-Quadrate-Schätzer $\widehat{\beta}^{(p)} \in \mathbb{R}^p$ und $\widehat{\sigma}_k^2 := \frac{1}{n}|Y - X^{(p)}\widehat{\beta}^{(p)}|^2$. Es gilt

$$\mathrm{AIC}(k) = n\log(2\pi\widehat{\sigma}_k^2) + n + 2k.$$

(ii) Im Fall, dass $\mu = X^{(p)}\beta_0^{(p)}$ für ein p und ein $\beta_0^{(p)} \in \mathbb{R}^p$ erfüllt ist, gilt mit $k = p+1$

$$\mathbb{E}\big[\,\mathrm{AIC}(k)\big] = \mathbb{E}\big[d_k(\widehat{\vartheta}_k)\big] - 2\frac{k(k+1)}{n-k-1}.$$

Für eine unbekannte Varianz liefert $\mathrm{AIC}(k)$ also keine unverzerrte Schätzung der Kullback-Leibler-Diskrepanz. Für allgemeines $\mu \in \mathbb{R}^n$ in (ii) kann man $\mathbb{E}[d_k(\widehat{\vartheta}_k)] - \mathbb{E}[\mathrm{AIC}(k)] \in [0, 2k(k+1)/(n-k-1)]$ zeigen. Die Abweichung ist also von der Größenordnung $O(k^2/n)$ in der Parameterdimension k und der Stichprobengröße n, was für große n bei beschränkter maximaler Modellgröße P vernachlässigbar ist. Ähnliches gilt für AIC in allgemeinen Likelihood-Modellen.

Beweis

(i) Die Loglikelihood-Funktion erfüllt

$$-2\log L_k(\vartheta^{(k)}) = n\log((2\pi\sigma^2)) + \frac{1}{\sigma^2}|Y - X^{(p)}\beta^{(p)}|^2,$$

sodass die Form des Maximum-Likelihood-Schätzers klar ist, siehe auch Aufgabe 2.6. Einsetzen liefert

$$\begin{aligned}\mathrm{AIC}(k) &= -2\log L_k(\widehat{\vartheta}^{(k)}) + 2k = n\log\left(2\pi\widehat{\sigma}_k^2\right) + \frac{1}{\widehat{\sigma}_k^2}|Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + 2k \\ &= n\log\left(2\pi\widehat{\sigma}_k^2\right) + n + 2k.\end{aligned}$$

(ii) Für $\vartheta^{(k)} = (\beta^{(p)}, \sigma^2)$ gilt

$$d_k(\vartheta^{(k)}) = \mathbb{E}_{\mathbb{P}}[-2\log L_k(\beta^{(p)}, \sigma^2)] = n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}(|\mu - X^{(p)}\beta^{(p)}|^2 + n\sigma_0^2),$$

sodass

$$\mathbb{E}_{\mathbb{P}}[d_k(\widehat{\vartheta}^{(k)})] = n\mathbb{E}[\log(2\pi\widehat{\sigma}_k^2)] + \mathbb{E}\Big[\frac{1}{\widehat{\sigma}_k^2}\left(|\mu - X^{(p)}\widehat{\beta}^{(p)}|^2 + n\sigma_0^2\right)\Big].$$

Wegen $\mathbb{E}_{\mathbb{P}}[\mathrm{AIC}(k)] = n\mathbb{E}[\log(2\pi\widehat{\sigma}_k^2)] + n + 2k$ folgt

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[\mathrm{AIC}(k)] - \mathbb{E}_{\mathbb{P}}[d_k(\widehat{\vartheta}^{(k)})] &= n + 2k - \mathbb{E}\Big[\frac{1}{\widehat{\sigma}_k^2}\left(|\mu - X^{(p)}\widehat{\beta}^{(p)}|^2 + n\sigma_0^2\right)\Big] \\ &= n + 2k - \mathbb{E}\Big[\frac{|\mu - \Pi^{(p)}Y|^2 + n\sigma_0^2}{\frac{1}{n}|(E_n - \Pi^{(p)})Y|^2}\Big].\end{aligned}$$

Da $\Pi^{(p)}$ und $E_n - \Pi^{(p)}$ Orthogonalprojektion auf $\mathrm{Im}\, X^{(p)}$ bzw. $\mathrm{Im}(X^{(p)})^\perp$ sind, sind $\Pi^{(p)}Y$ und $(E_n - \Pi^{(p)})Y$ unkorreliert und damit unabhängig. Es folgt

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}}[\mathrm{AIC}(k)] - \mathbb{E}_{\mathbb{P}}[d(\widehat{\vartheta}^{(k)})] \\ &\quad= n + 2k - \mathbb{E}\big[|\mu - \Pi^{(p)}Y|^2 + n\sigma_0^2\big]\mathbb{E}\Big[\frac{n}{|(E_n - \Pi^{(p)})Y|^2}\Big] \\ &\quad= n + 2k - \left(|\mu - \Pi^{(p)}\mu|^2 + p\sigma_0^2 + n\sigma_0^2\right)\frac{n}{\sigma_0^2}\mathbb{E}\left[Z^{-1}\right],\end{aligned}$$

wobei $Z := \frac{1}{\sigma_0^2}|(E_n - \Pi^{(p)})Y|^2 = \frac{1}{\sigma_0^2}|(E_n - \Pi^{(p)})\varepsilon|^2 \sim \chi^2(n-p)$ und $|\mu - \Pi^{(p)}\mu|^2 = 0$ im korrekt spezifizierten Modell gilt. Wir berechnen

$$\mathbb{E}\big[Z^{-1}\big] = \int_0^\infty \frac{1}{z}\frac{2^{-\frac{n-p}{2}}}{\Gamma(\frac{n-p}{2})}z^{\frac{n-p}{2}-1}e^{-\frac{z}{2}}\mathrm{d}z = \frac{1}{n-p-2} = \frac{1}{n-k-1}$$

für $n \geqslant p + 3$ und erhalten

$$\mathbb{E}_{\mathbb{P}}[\mathrm{AIC}(k)] - \mathbb{E}_{\mathbb{P}}[d_k(\widehat{\vartheta}^{(k)})] = n + 2k - \frac{(k-1+n)n}{n-k-1} = -2\frac{k(k+1)}{n-k-1}.$$

Beispiel 3.12 (AIC für Klimadaten) Wir erinnern uns an die Klimadatenzeitreihe aus Beispiel 2.52, in der wir die mittleren Julitemperaturen in Berlin-Dahlem zwischen 1719 und 2020 betrachtet haben. Wir hatten hier Regressionspolynome mit den Graden $d \in \{1, \ldots, 4\}$ verwendet. $\mathrm{AIC}(d)$ für $d \in \{1, \ldots, 30\}$ ist in Abbildung 3.1

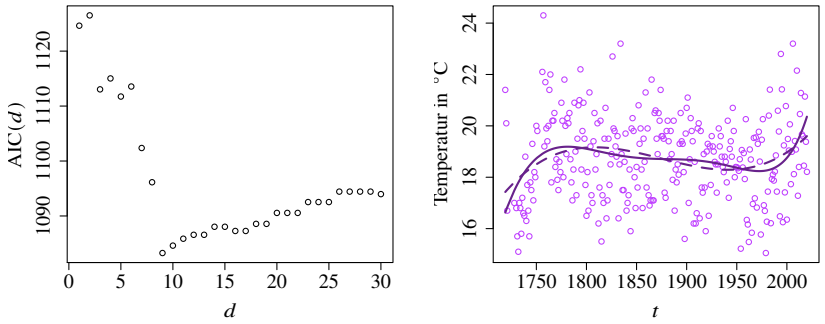

Abb. 3.1 *Links:* AIC für polynomielle Regression in Abhängigkeit vom Polynomgrad für die Temperaturdaten aus Beispiel 2.52. *Rechts:* Das Regressionpolynom des Grades 9 (durchgezogene Linie) sowie des Grades 3 (gestrichtelte Linie). Datenquelle: Deutscher Wetterdienst

dargestellt. Bis $d = 9$ fällt das Informationskriterium im Wesentlichen, hier profitieren wir also durch die Erhöhung der Parameterdimension noch deutlich in den Residuen. Ab $d = 10$ wächst $\mathrm{AIC}(d)$, sodass ab hier der Einfluss des Strafterms dominiert. Ebenfalls ist in Abbildung 3.1 das durch AIC gewählte Regressionspolynom zusammen mit dem Polynom vom Grad 3, für das wir uns in Beispiel 2.52 entschieden hatten, zu sehen.

3.1.2 Das Bayes-Informationskriterium (BIC)

Anstelle des Ansatzes, die Kullback-Leibler-Diskrepanz oder das Risiko erwartungstreu zu schätzen, ist der Bayes-Ansatz (der auf Gideon Schwarz zurückgeht), eine gleichmäßige a-priori-Verteilung für die Modelle $p \in \{1, \ldots, P\}$ anzunehmen. Die a-posteriori-Verteilung liefert uns dann ein Kriterium zur Auswahl des Modells p.

Kurzbiografie (Gideon Schwarz) Gideon E. Schwarz wurde 1933 in Salzburg geboren. Er flüchtete nach dem Anschluss Øßterreichs 1938 in das damaligen Palästina, heute Israel. Gideon E. Schwarz studierte an der Hebrew University Mathematik. Seinen Ph.D. erwarb er an der Columbia University 1961 im Bereich der mathematischen Statistik. Später wurde er Professor an der Hebrew University mit zahlreichen Gastaufenthalten in Stanford, Tel Aviv und Berkeley. Er verfasste vor allem wichtige Beiträge zur Bayes-Statistik. Als Reaktion auf das *Informationskriterium* AIC entwickelte er 1976 das *Bayes-Informationskriterium* BIC, manchmal auch Schwarz Information Criterion (SIC) genannt. Schon damals entwickelte sich eine große Diskussion über die verschiedenen Sichtweisen, die teilweise bis heute anhält. Schwarz starb im Jahr 2007.

Um das Bayes-Informationskriterium (englisch: *Bayesian information criterion*, kurz: BIC) herzuleiten, betrachten wir wieder das wahre lineare Modell

$$Y = \mu + \varepsilon \qquad \text{mit} \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$$

und dem unbekannten Mittelwertvektor $\mu \in \mathbb{R}^n$, den es zu schätzen gilt. Das Rauschniveau $\sigma > 0$ nehmen wir als bekannt an. Unser Ziel ist, aus den Modellen

$$Y = X^{(p)}\beta^{(p)} + \varepsilon$$

mit $\beta^{(p)} \in \mathbb{R}^p$, $X^{(p)} \in \mathbb{R}^{n\times p}$, $\operatorname{rank} X = p$ und $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ für $p \in \{1, \dots, P\}$ ein geeignetes auszuwählen.

Für jedes p sei die a-priori-Verteilung des Parameters $\beta^{(p)}$ durch eine Lebesgue-Dichte π_p auf $\mathbb{R}^p$ gegeben. Für das Modell p wählen wir die uninformative a-priori-Verteilung $U := \mathrm{U}(\{1, \dots, P\})$, sodass jedes Modell p mit der gleichen Wahrscheinlichkeit $\frac{1}{P}$ ausgewählt wird. Das gemäß U ausgewählte Modell bezeichnen wir mit κ. Die a-priori-Verteilungen von p und $\beta^{(p)}$ seien unabhängig.

Die gemeinsame Verteilung von Modell κ, Parameter β und Beobachtung Y ist dann gegeben durch

$$\mathbb{P}^{(\kappa,\beta,Y)}(A) = \int \mathbb{1}_A(p, \beta^{(p)}, y) \frac{e^{-|y-X^{(p)}\beta^{(p)}|^2/(2\sigma^2)}}{(2\pi\sigma^2)^{n/2}} \mathrm{d}y \pi_p(\beta^{(p)}) \mathrm{d}\beta^{(p)} U(\mathrm{d}p).$$

Da die Dimension von $\beta^{(p)}$ von p abhängt, ist in dieser Formel Vorsicht geboten. Um einen gemeinsamen Grundraum zu haben, müssen wir $A \subseteq \{1, \dots, P\} \times \mathbb{R}^P \times \mathbb{R}^n$ wählen und in der Indikatorfunktion $\beta^{(p)}$ als Vektor in $\mathbb{R}^P$ interpretieren, dessen letzte $P - p$ Koordinaten null sind. Da wir nur an der gemeinsamen Verteilung von κ und Y interessiert sind, entfällt dieses technische Problem ohnehin: Die Verteilung von (κ, Y) besitzt die Dichte

$$f^{(\kappa,Y)}(p, y) = \int_{\mathbb{R}^p} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(- \frac{|y - X^{(p)}\beta^{(p)}|^2}{2\sigma^2} \Big) \pi_k(\beta^{(p)}) \mathrm{d}\beta^{(p)}$$

bezüglich des Produktmaßes $U \otimes \lambda^{\otimes n}$. Der BIC-Ansatz besteht nun darin, das Modell mit der größten a-posteriori-Wahrscheinlichkeit auszuwählen, siehe *MAP*-Schätzer aus Korollar 1.33:

$$\widehat{p} \in \operatorname*{arg\,min}_{p=1,\dots,P} \mathbb{P}(\kappa = p|Y)$$

Für dieses Kriterium leiten wir nun eine deutlich zugänglichere (approximative) Darstellung her. Für die a-posteriori-Verteilung von κ gilt

$$\mathbb{P}(\kappa = p|Y) \propto \int_{\mathbb{R}^k} \exp\Big(- \frac{|Y - X^{(p)}\beta^{(p)}|^2}{2\sigma^2} \Big) \pi_p(\beta^{(p)}) \mathrm{d}\beta^{(p)}.$$

Es bezeichne wieder $\widehat{\beta}^{(p)}$ den Kleinste-Quadrate-Schätzer in Modell p. Aufgrund der Orthogonalität der Residuen $Y - X^{(p)}\widehat{\beta}^{(p)} = (E_n - \Pi_{X^{(p)}})Y$ zum Bild von $X^{(p)}$ können wir den Exponenten in der a-posteriori-Verteilung gemäß

$$|Y - X^{(p)}\beta^{(p)}|^2 = |Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + |X^{(p)}\widehat{\beta}^{(p)} - X^{(p)}\beta^{(p)}|^2$$

$$
\begin{aligned}
&\quad - 2\langle Y - X^{(p)}\beta^{(p)}, X^{(p)}(\widehat{\beta}^{(p)} - \beta^{(p)})\rangle \\
&= |Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + |X^{(p)}\widehat{\beta}^{(p)} - X^{(p)}\beta^{(p)}|^2
\end{aligned}
$$

zerlegen. Da $\widehat{\beta}^{(p)}$ nur von Y (und p), aber nicht von β abhängt, erhalten wir

$$
\begin{aligned}
\mathbb{P}(\kappa = p|Y) \propto \exp\Big(-\frac{1}{2\sigma^2}|Y - X^{(p)}\widehat{\beta}^{(p)}|^2\Big) \\
\times \int_{\mathbb{R}^p} \exp\Big(-\frac{1}{2\sigma^2}|X^{(p)}(\widehat{\beta}^{(p)} - \beta^{(p)})|^2\Big)\pi_p(\beta^{(p)})\mathrm{d}\beta^{(p)}.
\end{aligned}
$$

Durch eine Substitution $h = \sqrt{n}(\widehat{\beta}^{(p)} - \beta^{(p)})$ erhalten wir für $\Sigma_n^{(p)} := \frac{1}{n}(X^{(p)})^\top X^{(p)} \in \mathbb{R}^{p\times p}$

$$
\begin{aligned}
&\int_{\mathbb{R}^p} \exp\Big(-\frac{1}{2\sigma^2}|X^{(p)}(\widehat{\beta}^{(p)} - \beta^{(p)})|^2\Big)\pi_p(\beta^{(p)})\mathrm{d}\beta^{(p)} \\
&\quad = n^{-p/2}\int_{\mathbb{R}^p} \exp\Big(-\frac{1}{2\sigma^2}\langle\Sigma_n^{(p)}h, h\rangle\Big)\pi_p\Big(\widehat{\beta}^{(p)} - \frac{h}{\sqrt{n}}\Big)\mathrm{d}h =: n^{-p/2}I_{n,p}.
\end{aligned}
$$

Bezeichnen wir die Normierungskonstante der Dichte von $\mathbb{P}(\kappa = p|Y)$ mit $C > 0$, so ergibt sich

$$
\log\mathbb{P}(\kappa = p|Y) = -\log C - \frac{p}{2}\log n - \frac{1}{2\sigma^2}\big|Y - X^{(p)}\widehat{\beta}^{(p)}\big|^2 + \log I_{n,p} \tag{3.2}
$$

Um den letzten Term $\log I_{n,p}$ abzuschätzen, [RD]betrachten wir die Asymptotik $n \to \infty$ und nehmen an, dass

$$
\Sigma_n^{(p)} \to \Sigma^{(p)} \qquad \text{für} \qquad n \to \infty
$$

konvergiert.

Beispiel 3.13 (Asymptotik der Designmatrix) Es seien $\varphi_1, \dots, \varphi_n \colon [0,1] \to \mathbb{R}$ stetige Funktionen, beispielsweise die Monome $\varphi_k(x) = x^{k-1}$. Unter äquidistantem Design ist die Designmatrix durch

$$
X^{(p)} = \begin{pmatrix}
\varphi_1(1/n) & \cdots & \varphi_p(1/n) \\
\varphi_1(2/n) & \cdots & \varphi_p(2/n) \\
\vdots & & \vdots \\
\varphi_1(n/n) & \cdots & \varphi_p(n/n)
\end{pmatrix}
$$

gegeben, und wir erhalten für alle $i, j \in \{1, \dots, p\}$

$$
(\Sigma_n^{(p)})_{i,j} = \frac{1}{n}\sum_{k=1}^{n}\varphi_i\Big(\frac{k}{n}\Big)\varphi_j\Big(\frac{k}{n}\Big) \to \int_0^1 \varphi_i(x)\varphi_j(x)dx =: (\Sigma^{(p)})_{i,j}
$$

für $n \to \infty$.

Sind nun noch die Dichten π_p gleichmäßig beschränkt, finden wir mit dominierter Konvergenz eine obere Schranke für $I_{n,p}$, die gleichmäßig in n und p gilt. Aus (3.2) folgt dann für eine Nullfolge $o(1) \to 0$ für $n \to \infty$

$$-2 \log \mathbb{P}(\kappa = p|Y) = p \log n(1 + o(1)) + \frac{1}{\sigma^2}|Y - X^{(p)}\widehat{\beta}^{(p)}|^2.$$

Vernachlässigt man den $o(1)$-Term, entspricht das Bayes-Informationskriterium gerade

$$\widehat{p} \in \operatorname*{arg\,min}_{p=1,\dots,P} \left\{ \frac{1}{\sigma^2}|Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + p \log n \right\}.$$

Die quadrierten Residuen entsprechen wieder der negativen Loglikelihood-Funktion im linearen Modell (mit Faktor $2\sigma^2$). Hinzu kommt eine Penalisierung mit $p \log n$. Analog zu Akaikes Informationskriterium definieren wir daher ganz allgemein:

Methode 3.14 (Modellwahl durch BIC) Für $p = 1, \dots, P$ und dominierte statistische Modelle $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_{\vartheta^{(p)}})_{\vartheta^{(p)} \in \Theta_p})$ mit $\mathcal{X} = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$ und $\Theta_p \subseteq \mathbb{R}^p$ ist das **Bayes-Informationskriterium** (kurz: BIC) definiert als

$$\mathrm{BIC}(p) := -2 \log L_p(\widehat{\vartheta}^{(p)}, X) + p \log n$$

für die Likelihood-Funktion L_p und den Maximum-Likelihood-Schätzer $\widehat{\vartheta}_p$ im Modell p. Das Modell $\widehat{p}$ wird als Minimierer $\widehat{p} \in \arg\min_{p=1,\dots,P} \mathrm{BIC}(p)$ gewählt.

Entsprechend der vorangegangenen Herleitung erhalten wir:

Lemma 3.15 *Für $p = 1, \dots, P$ mit $P \leqslant n$ und die linearen Modelle*

$$Y = X^{(p)}\beta^{(p)} + \varepsilon \qquad \textit{mit} \qquad \beta^{(p)} \in \mathbb{R}^p, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2 E_n),$$

Designmatrix $X^{(p)} \in \mathbb{R}^{n \times p}$ von vollem Rang p und mit bekannter Varianz $\sigma > 0$ gilt

$$\mathrm{BIC}(p) = n \log(2\pi\sigma^2) + \frac{1}{\sigma^2}|Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + p \log n.$$

Im Vergleich zum AIC wird im BIC die Modelldimension mit $p \log n$ statt mit $2p$ bestraft. BIC wählt also für $n \geqslant 8$ tendenziell ein kleineres Modell aus. Andererseits führt die AIC-Wahl typischerweise auf einen kleineren Vorhersagefehler $|Y - X^{(p)}\widehat{\beta}^{(p)}|^2$.

Man beachte, dass $|Y - X^{(p)}\widehat{\beta}^{(p)}|^2$ in n wächst und ohne weitere Annahmen von der Größenordnung n ist. Reskalieren wir BIC durch $\frac{1}{n}|Y - X^{(p)}\widehat{\beta}^{(p)}|^2 + p\frac{\log n}{n}$, wird deutlich, dass die Erhöhung der Modelldimension mit steigendem n geringer bestraft wird.

Beispiel 3.16 (BIC für Klimadaten) Wir wenden das Bayes-Informationskriterium auf die Daten aus den Beispielen 2.52 bzw. 3.12 über die langfristige Entwicklung

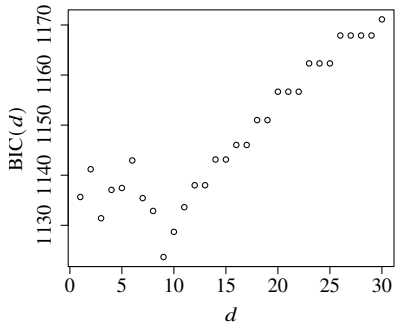

Abb. 3.2 BIC für polynomielle Regression in Abhängigkeit vom Polynomgrad für die Temperaturdaten aus Beispiel 2.52

der mittleren Julitemperaturen an. Die resultierenden Werte $\mathrm{BIC}(d)$ für die Polynomgrade $d \in \{1, \dots, 30\}$ sind in Abbildung 3.2 dargestellt. Im Vergleich zu den AIC-Werten steigt BIC wesentlich schneller in d an. Zwar wählt auch das Bayes-Informationskriterium den Grad 9, aber der Abstand zum Grad 3 ist hier viel kleiner als bei Akaikes Informationskriterium.

3.2 Orakelungleichung für die penalisierte Modellwahl

Im vorangegangenen Abschnitt haben wir die Informationskriterien AIC und BIC hergeleitet und Beispiele gesehen, in denen diese gute Resultate erzielen. Die offene Frage ist, ob wir im Allgemeinen nachweisen können, dass die Modellwahlkriterien den Trade-off zwischen Datenanpassung und zu hoher Modellkomplexität lösen.

Wie zuvor seien die Beobachtungen durch

$$Y = \mu + \varepsilon, \qquad \mu \in \mathbb{R}^n, \varepsilon \sim \mathrm{N}(0,\, \sigma^2 E_n),$$

gegeben. Für die Modelle $Y = X^{(p)}\beta^{(p)} + \varepsilon$ mit den Designmatrizen $X^{(p)} \in \mathbb{R}^{n\times p}$ von vollem Rang bezeichne wieder $\widehat{\beta}^{(p)}$ den jeweiligen Kleinste-Quadrate-Schätzer, wobei $p \in \{1, \dots, P\}$. Die Modellwahlkriterien AIC und BIC zur Wahl des finalen Schätzers $\widehat{\beta}^{(\widehat{p})}$ waren beide von der Form

$$\widehat{p} \in \operatorname*{arg\,min}_{p=1,\dots,P} \left(\left| Y - X^{(p)}\widehat{\beta}^{(p)} \right|^2 + \mathrm{Pen}(p) \right)$$

mit dem Penalisierungsterm $\mathrm{Pen}(p)$. Für AIC ist Letzterer durch $\mathrm{Pen}(p) = 2\sigma^2 p$ gegeben, während BIC dem Strafterm $\mathrm{Pen}(p) = \log(n)\sigma^2 p$ entspricht. In beiden Fällen hängt der Strafterm also von der Dimension des Parameterraums ab. Betrachtet man

$$S_p := \Im X^{(p)} \quad \text{und} \quad \widehat{\mu}^{(p)} := X^{(p)}\widehat{\beta}^{(p)} = \Pi_{S_p} Y$$

mit der Orthogonalprojektion Π_{S_p} auf das Bild von $X^{(p)}$, betten sich beide Verfahren in die folgende allgemeine Modellwahl ein.

Definition 3.17 Für eine Beobachtung $Y \in \mathbb{R}^n$ seien lineare Unterräume $S_m \subseteq \mathbb{R}^n$ der Dimension p_m gegeben, wobei $m = 1, \dots, M$. Für eine monoton wachsende Funktion $\mathrm{Pen}\colon \mathbb{N} \to [0, \infty)$ betrachten wir die **penalisierte Modellwahl**

$$\widehat{m} \in \operatorname*{arg\,min}_{m=1,\dots,M} \left(\left| Y - \widehat{\mu}^{(m)} \right|^2 + \mathrm{Pen}(p_m) \right)$$

mit den Kleinste-Quadrate-Schätzern $\widehat{\mu}^{(m)} := \Pi_{S_m} Y \in S_m$. Wir setzen $\mu^{(m)} = \Pi_{S_m} \mu$.

Wir werden nun eine Orakelungleichung für die penalisierte Modellwahl beweisen. Eine erste Orakelungleichung ist uns bereits in Kapitel **??** begegnet. Mit ihrer Hilfe konnten wir das Verhalten des Kleinste-Quadrate-Schätzers im missspezifierten Modell analysieren, siehe Korollar **??**. Dieses Resultat müssen wir um den Penalisierungsterm erweitern. Die größte Schwierigkeit besteht darin, die Abhängigkeit zwischen $\widehat{m}$ und $\widehat{\mu}^{(m)}$ im gewählten Parameter $\widehat{\mu}^{(\widehat{m})}$ zu kontrollieren.

Ausgangspunkt für den Beweis der Orakelungleichung war die Fundamentalungleichung aus Lemma (**??**). Im penalisierten Fall (und für quadratischen Verlust) erhalten wir folgendes Analogon:

Lemma 3.18 (Fundamentalungleichung) *Im Modell $Y = \mu + \varepsilon$ gilt für den penalisierten empirischen Risikominimierer*

$$\widehat{m} := \operatorname*{arg\,min}_{m=1,\dots,M} \left\{ |Y - \widehat{\mu}^{(m)}|^2 + \mathrm{Pen}(\overline{p}_m) \right\}$$

die Ungleichung

$$|\mu - \widehat{\mu}^{(\widehat{m})}|^2 + \mathrm{Pen}(p_{\widehat{m}}) \leqslant |\mu - \mu^{(m)}|^2 + \mathrm{Pen}(p_m) + 2\langle \varepsilon,\, \widehat{\mu}^{(\widehat{m})} - \mu^{(m)} \rangle$$

für jedes $m = 1, \dots, M$.

Beweis Nach Definition von $\widehat{m}$ gilt für jedes m

$$|Y - \widehat{\mu}^{(\widehat{m})}|^2 + \mathrm{Pen}(p_{\widehat{m}}) \leqslant |Y - \widehat{\mu}^{(m)}|^2 + \mathrm{Pen}(p_m) \leqslant |Y - \mu^{(m)}|^2 + \mathrm{Pen}(p_m),$$

wobei wir $|Y - \widehat{\mu}^{(m)}|^2 = |Y - \Pi_{S_m} Y|^2 \leqslant |Y - \mu^{(m)}|^2$ aus der Eigenschaft der Orthogonalprojektion Π_{S_m} und $\mu^{(m)} \in S_m$ folgern. Aus $|Y - v|^2 = |v - \mu|^2 + |\varepsilon|^2 - 2\langle \varepsilon,\, v - \mu \rangle$ für jedes $v \in \mathbb{R}^n$ folgt durch Einsetzen

$$|\mu - \widehat{\mu}^{(\widehat{m})}|^2 - 2\langle \varepsilon,\, \widehat{\mu}^{(\widehat{m})} - \mu \rangle + \mathrm{Pen}(p_{\widehat{m}}) \leqslant |\mu - \mu^{(m)}|^2 - 2\langle \varepsilon,\, \mu^{(m)} - \mu \rangle + \mathrm{Pen}(p_m).$$

Umstellen liefert die Behauptung. □

Satz 3.19 (Orakelungleichung zur Modellwahl) *Betrachte das datenerzeugende Modell $Y = \mu + \varepsilon$, $\mu \in \mathbb{R}^n$, $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ und für lineare Unterräume $S_m \subseteq \mathbb{R}^n$,*

$m = 1, \dots, M$, mit $\dim(S_m) = p_m$ die penalisierte Modellwahl $\widehat{m}$, wobei $\mathrm{Pen}(p_m) \geqslant K\sigma^2(p_m + 1)$ für ein $K > 1$ gelte. Weiter seien $\widehat{\mu}^{(m)} = \Pi_{S_m} Y$, $\mu^{(m)} = \Pi_{S_m}\mu$ und $\kappa \in (0, \sqrt{K} - 1)$.

(i) Für jedes $\tau > 0$ gilt mit der Wahrscheinlichkeit $1 - e^{-\tau/2} \sum_{m=1}^{M} e^{-p_m\kappa^2/2}$ die Orakelungleichung:

$$|\widehat{\mu}^{(\widehat{m})} - \mu|^2 \leqslant C(K,\kappa)\Big(\min_{m=1,\dots,M} \big(|\mu^{(m)} - \mu|^2 + \mathrm{Pen}(p_m)\big) + \sigma^2\tau \Big) \tag{3.3}$$

mit einer Konstanten $C(K, \kappa) > 0$, die nur von K und κ abhängt, wobei $C(K,\kappa) \to \infty$ für $\kappa \to \sqrt{K} - 1$.

(ii) Es gilt

$$\mathbb{E}\big[|\widehat{\mu}^{(\widehat{m})} - \mu|^2\big] \leqslant \widetilde{C}(K,\kappa)\Big(\min_{m=1,\dots,M} \big(|\mu^{(m)} - \mu|^2 + \mathrm{Pen}(p_m)\big) + \sigma^2 \sum_{m=1}^{M} e^{-p_m\kappa^2/2} \Big)$$

mit einer Konstanten $\widetilde{C}(K, \kappa) > 0$, die nur von K und κ abhängt, wobei $C(K,\kappa) \to \infty$ für $\kappa \to \sqrt{K} - 1$.

Beweis Der Beweis von *(i)* erfolgt in vier Schritten:

Schritt 1: Anwenden der Fundamentalungleichung. Aus Lemma 3.18 folgt für beliebiges $m \in \{1, \dots, M\}$

$$|\widehat{\mu}^{(\widehat{m})} - \mu|^2 \leqslant |\mu^{(m)} - \mu|^2 + \mathrm{Pen}(p_m) - \mathrm{Pen}(p_{\widehat{m}}) + 2\langle \varepsilon, \widehat{\mu}^{(\widehat{m})} - \mu^{(m)} \rangle.$$

Wegen $\widehat{\mu}^{(\widehat{m})} - \mu^{(m)} \in \mathrm{span}(S_{\widehat{m}}, \mu^{(m)}) =: S^*_{\widehat{m}}$ mit $\dim S^*_{\widehat{m}} \leqslant p_{\widehat{m}} + 1$ folgt

$$\langle \varepsilon, \widehat{\mu}^{(\widehat{m})} - \mu^{(m)} \rangle \leqslant |\widehat{\mu}^{(\widehat{m})} - \mu^{(m)}| \sup_{s \in S^*_{\widehat{m}}} \frac{|\langle \varepsilon,\, s\rangle|}{|s|}.$$

Bezeichnen wir die Orthogonalprojektion auf $S^*_{\widehat{m}}$ mit $\Pi_{S^*_{\widehat{m}}}$, dann wird das Supremum bei $s = \Pi_{S^*_{\widehat{m}}}\varepsilon$ mit dem maximalen Wert $|\Pi_{S^*_{\widehat{m}}}\varepsilon|$ angenommen. Wir erhalten

$$|\widehat{\mu}^{(\widehat{m})} - \mu|^2 \leqslant |\mu^{(m)} - \mu|^2 + \mathrm{Pen}(p_m) - \mathrm{Pen}(p_{\widehat{m}}) + 2|\widehat{\mu}^{(\widehat{m})} - \mu^{(m)}|\,|\Pi_{S^*_{\widehat{m}}}\varepsilon|.$$

Schritt 2: Konzentrationsungleichung für $\chi^2(d)$. Sei $Z_d \sim \chi^2(d)$, das heißt, $Z_d \overset{d}{=} \sum_{i=1}^{d} X_i^2$ für i.i.d. $X_i \sim \mathrm{N}(0,1)$. Wegen $\mathbb{E}[e^{uX_i^2}] = (1-2u)^{-1/2}$ für jedes $u \in (0, 1/2)$ erhalten wir für $\rho > 1$

$$\mathbb{P}(Z_d \geqslant \rho d) \leqslant \mathbb{E}\big[e^{uZ_d} e^{-u\rho d}\big] = (1-2u)^{-\frac{d}{2}} e^{-u\rho d} = e^{-\frac{d}{2}(\rho - 1 - \log\rho)},$$

wobei für den letzten Schritt $u = (\rho - 1)/(2\rho)$ gewählt wurde. Aufgrund von $\log(1+t) \leqslant t$ für $t \geqslant 0$ folgt daraus

$$\mathbb{P}\Big(Z_d \geqslant \big(1 + \kappa + \sqrt{\tau/d}\big)^2 d\Big)$$

$$\leqslant \exp\Big(-\frac{d}{2}\Big[\big(\kappa+\sqrt{\tau/d}\big)^2+2\big(\kappa+\sqrt{\tau/d}\big)-2\log\big(1+\kappa+\sqrt{\tau/d}\big)\Big]\Big)$$
$$\leqslant \exp\Big(-\frac{d}{2}\Big(\kappa+\sqrt{\tau/d}\Big)^2\Big) \leqslant \exp\Big(-\frac{d}{2}\kappa^2-\frac{\tau}{2}\Big).$$

Schritt 3: Kontrolle des stochastischen Fehlers. Es gilt $Z_{d_m} := \sigma^{-2}|\Pi_{S_m^*}\varepsilon|^2 \sim \chi^2(d_m)$ für ein $d_m \in \{p_m, p_m+1\}$ und jedes m. Insbesondere folgt $\mathbb{E}[|\Pi_{S_m^*}\varepsilon|^2] = \sigma^2 d_m$ für jedes deterministische m. Wir normieren und korrigieren daher $|\Pi_{S_{\widehat{m}}^*}\varepsilon|$ um einen Term dieser Größe. Anschließend schätzen wir $\widehat{m}$ durch das Maximum in $\{1,\dots,M\}$ ab, sodass wir die Abhängigkeit zwischen $\widehat{m}$ und ε im Term $\Pi_{S_{\widehat{m}}^*}\varepsilon$ auflösen:

$$|\Pi_{S_{\widehat{m}}^*}\varepsilon| \leqslant \sigma\Big((\kappa+1)\sqrt{p_{\widehat{m}}+1}+\frac{1}{\sigma}|\Pi_{S_{\widehat{m}}^*}\varepsilon|-(\kappa+1)\sqrt{p_{\widehat{m}}+1}\Big)$$
$$\leqslant \sigma\Big((\kappa+1)\sqrt{p_{\widehat{m}}+1}+\max_{m=1,\dots,M}\Big(\frac{1}{\sigma}|\Pi_{S_m^*}\varepsilon|-(\kappa+1)\sqrt{d_m}\Big)\Big)$$

Wir definieren nun das gute Ereignis

$$\mathcal{G} := \Big\{\max_{1\leqslant m\leqslant M}\Big(\frac{1}{\sigma}|\Pi_{S_m^*}\varepsilon|-(\kappa+1)\sqrt{d_m}\Big) \leqslant \sqrt{\tau}\Big\}.$$

Aus der Subadditivität von $\mathbb{P}$ und Schritt 2 angewendet auf $Z_{d_m} = \sigma^{-2}|\Pi_{S_m^*}\varepsilon|^2$ folgt

$$\mathbb{P}(\mathcal{G}^c) = \mathbb{P}\Big(\max_{1\leqslant m\leqslant M}\Big(\frac{1}{\sigma}|\Pi_{S_m^*}\varepsilon|-(\kappa+1)\sqrt{d_m}\Big) \geqslant \sqrt{\tau}\Big)$$
$$\leqslant \sum_{m=1}^{M}\mathbb{P}\Big(\frac{1}{\sigma}|\Pi_{S_m^*}\varepsilon|-(\kappa+1)\sqrt{d_m} \geqslant \sqrt{\tau}\Big)$$
$$\leqslant \sum_{m=1}^{M}\mathbb{P}\Big(\sqrt{Z_{d_m}} \geqslant \Big(1+\kappa+\sqrt{\frac{\tau}{d_m}}\Big)\sqrt{d_m}\Big)$$
$$\leqslant \sum_{m=1}^{M}\exp\Big(-\frac{\kappa^2 d_m}{2}-\frac{\tau}{2}\Big) \leqslant e^{-\tau/2}\sum_{m=1}^{M}e^{-p_m\kappa^2/2}.$$

Damit gilt auf $\mathcal{G}$ mit der gesuchten Wahrscheinlichkeit

$$|\Pi_{S_{\widehat{m}}^*}\varepsilon| \leqslant \sigma\big((\kappa+1)\sqrt{p_{\widehat{m}}+1}+\sqrt{\tau}\big) \leqslant \frac{\kappa+1}{\sqrt{K}}\sqrt{\mathrm{Pen}(p_{\widehat{m}})}+\sigma\sqrt{\tau}.$$

Schritt 4: Umordnung der Terme. Auf dem Ereignis $\mathcal{G}$ erhalten wir durch Kombination von Schritt 1 und Schritt 3:

$$|\widehat{\mu}^{(\widehat{m})}-\mu|^2 \leqslant |\mu^{(m)}-\mu|^2+\mathrm{Pen}(p_m)-\mathrm{Pen}(p_{\widehat{m}})$$
$$+2|\widehat{\mu}^{(\widehat{m})}-\mu^{(m)}|\Big(\frac{\kappa+1}{\sqrt{K}}\sqrt{\mathrm{Pen}(p_{\widehat{m}})}+\sigma\sqrt{\tau}\Big).$$

Wir zerlegen nun $|\widehat{\mu}^{(\widehat{m})} - \mu^{(m)}| \leqslant |\widehat{\mu}^{(\widehat{m})} - \mu| + |\mu^{(m)} - \mu|$ und verwenden zweimal $2AB \leqslant \eta A^2 + \eta^{-1}B^2$, $\eta > 0$, sodass wir

$$|\widehat{\mu}^{(\widehat{m})} - \mu|\Big(\frac{\kappa+1}{\sqrt{K}}\sqrt{\mathrm{Pen}(p_{\widehat{m}})} + \sigma\sqrt{\tau}\Big) \leqslant (\eta_1^{-1} + \eta_2^{-1})|\widehat{\mu}^{(\widehat{m})} - \mu|^2 + \eta_1 \frac{(1+\kappa)^2}{K}\mathrm{Pen}(p_{\widehat{m}}) + \eta_2\sigma^2\tau$$

für $\eta_1, \eta_2 > 0$ erhalten. Behandelt man den Summanden mit $|\mu^{(m)} - \mu|$ analog, dann folgt für $\eta_3, \eta_4 > 0$

$$|\widehat{\mu}^{(\widehat{m})} - \mu|^2 \leqslant (1 + \eta_3^{-1} + \eta_4^{-1})|\mu^{(m)} - \mu|^2 + \mathrm{Pen}(p_m) + (\eta_2 + \eta_4)\tau\sigma^2 + \Big((\eta_1 + \eta_3)\frac{(1+\kappa)^2}{K} - 1\Big)\mathrm{Pen}(p_{\widehat{m}}) + (\eta_1^{-1} + \eta_2^{-1})|\widehat{\mu}^{(\widehat{m})} - \mu|^2.$$

Aufgrund der Annahme $\frac{K}{(1+\kappa)^2} > 1$ finden wir nun $\eta_1, \eta_2, \eta_3 > 0$, sodass $\eta_1^{-1} + \eta_2^{-1} < 1$ und $\eta_1 + \eta_3 = \frac{K}{(1+\kappa)^2}$ erfüllt sind. Außerdem können wir $\eta_4 = 1$ wählen. Umstellen liefert auf $\mathcal{G}$, da m beliebig war,

$$|\widehat{\mu}^{(\widehat{m})} - \mu|^2 \leqslant C(\eta_1, \eta_2, \eta_3) \min_{m=1,\dots,M} \big(|\mu^{(m)} - \mu|^2 + \mathrm{Pen}(p_m) + \sigma^2\tau\big).$$

Wir folgern abschließend *(ii)*. Setze $t^* := C(K,\kappa)\min_{m=1,\dots,M}(|\mu - \mu^{(m)}|^2 + \mathrm{Pen}(p_m))$. Dann impliziert *(i)*

$$\begin{aligned}
\mathbb{E}\big[|\widehat{\mu}^{(\widehat{m})} - \mu|^2\big] &= \int_0^\infty \mathbb{P}\big(|\widehat{\mu}^{(\widehat{m})} - \mu|^2 > t\big)\mathrm{d}t \\
&\leqslant \int_0^{t^*} \mathbb{P}\big(|\widehat{\mu}^{(\widehat{m})} - \mu|^2 > t\big)\mathrm{d}t \\
&\quad + \int_0^\infty \mathbb{P}\big(|\widehat{\mu}^{(\widehat{m})} - \mu|^2 > t^* + C(K,\kappa)\sigma^2\tau\big)C(K,\kappa)\sigma^2\mathrm{d}\tau \\
&\leqslant t^* + C(K,\kappa)\sigma^2 \sum_{m=1}^M e^{-p_m\kappa^2/2}\int_0^\infty e^{-\tau/2}\mathrm{d}\tau.
\end{aligned}$$

Die Behauptung folgt, da das letzte Integral endlich ist. □

Bemerkung 3.20 (Interpretation der Orakelungleichung)

1. Das m^*, mit dem das Minimum auf der rechten Seite in (3.3) angenommen wird, nennt man auch *Orakelmodell*. Aus der Bias-Varianz-Zerlegung folgt

$$\mathbb{E}[|\widehat{\mu}^{(m)} - \mu|^2] = |\mu^{(m)} - \mu|^2 + \sigma^2 p_m \approx |\mu^{(m)} - \mu|^2 + \mathrm{Pen}(p_m)/K$$

für $\mathrm{Pen}(d_m) = K\sigma^2(d_m + 1)$. Damit liegt $|\mu^{(m^*)} - \mu|^2 + \mathrm{Pen}(d_{m^*})$ nahe am *Orakelfehler* $\min_m \mathbb{E}[|\widehat{\mu}^{(m)} - \mu|^2]$. Mit $\tau = d_{m^*}$ ist der Restterm $\sigma^2\tau$ von der Ordnung $\mathrm{Pen}(d_{m^*})$ (oder kleiner) und kann damit durch das Minimum in (3.3) abgeschätzt

werden. Für dieses τ ist die Wahrscheinlichkeit, mit der die Orakelungleichung gilt, gegeben durch

$$1 - \sum_{m=1}^{M} e^{-p_m \kappa^2/2} e^{-d_{m^*}/2}.$$

Falls asymptotisch $n \to \infty$, $d_{m^*} \to \infty$ (also μ in keinem S_m exakt liegt) und diese Summe gleichmäßig in M beschränkt ist, dann konvergiert diese Wahrscheinlichkeit exponentiell schnell in d_{m^*} gegen eins.

2. In Teil *(ii)* kann man für festes M und $n \to \infty$ im Fall des BIC mit $\mathrm{Pen}(p_m) = \log(n) p_m \sigma^2$ auch κ zunehmend größer wählen, sodass der hintere Term sehr klein wird.

Im Satz 3.19 wird nicht gefordert, dass die Modelle geordnet sind. In der multiplen Regression können daher alle 2^p Untermodelle betrachtet werden. Allerdings muss dann κ hinreichend groß gewählt werden, um eine große Wahrscheinlichkeit sicher zu stellen. In der Tat zeigt sich bei dieser „full subset"-Variablenselektion, dass in der Praxis AIC und BIC nicht so gut funktionieren wie größere $\mathrm{Pen}(p_m)$-Penalisierungen. Eine weitergehende Analyse von penalisierten Modellwahlverfahren ist in Massart (2007) zu finden.

3.3 Aufgaben

4.1 Zeigen Sie, dass die Kullback-Leibler-Divergenz nicht symmetrisch und somit keine Metrik ist. Zeigen Sie, dass auch die symmetrisierte Version $\mathrm{KL}(\mathbb{P}|\mathbb{Q}) + \mathrm{KL}(\mathbb{Q}|\mathbb{P})$ keine Metrik ist.

4.2 Bestimmen Sie die Kullback-Leibler-Divergenz $\mathrm{KL}(\mathrm{N}(\mu_1, \Sigma_1)|\mathrm{N}(\mu_2, \Sigma_2))$ zwischen zwei d-dimensionalen Normalverteilungen mit $\mu_1, \mu_2 \in \mathbb{R}^d$ und symmetrischen, positiv-semidefiniten Matrizen $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d \times d}$.

4.3 Der *Totalvariationsabstand* zweier Wahrscheinlichkeitsmaße $\mathbb{P}$ und $\mathbb{Q}$ auf einem messbaren Raum $(\mathcal{X}, \mathcal{F})$ ist definiert als

$$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

(a) Zeigen Sie

$$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \int |p - q| \mathrm{d}\nu$$

für ein dominierendes Maß ν und die Radon-Nikodym-Dichten $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu}$, $q = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu}$.

(b) Zeigen Sie $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \leqslant \sqrt{\mathrm{KL}(\mathbb{P}|\mathbb{Q})/2}$.
Hinweis: Betrachten Sie die Funktion $h(z) := z \log(z) - z + 1, z > 0$, mit stetiger Fortsetzung in Null. Zeigen Sie für alle $z \geqslant 0$

$$\left(\tfrac{4}{3} + \tfrac{2}{3}z\right)h(z) \geqslant (z-1)^2.$$

Verwenden Sie anschließend die Cauchy-Schwarz-Ungleichung.

(c) Prüfen Sie, ob folgende Folgen $(\mathbb{P}_n)_{n\in\mathbb{N}}$ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ schwach, im Totalvariationsabstand oder in der Kullback-Leibler-Divergenz gegen einen geeigneten Grenzwert $\mathbb{P}$ konvergieren (betrachten Sie sowohl $\mathrm{KL}(\mathbb{P}_n|\mathbb{P})$ als auch $\mathrm{KL}(\mathbb{P}|\mathbb{P}_n)$):
   (i) $\mathbb{P}_n = \delta_{1/n}$, Dirac-Maß in $1/n$;
   (ii) $\mathbb{P}_n = (1 - \frac{1}{n})\nu + \frac{1}{n}\delta_1$, wobei ν das Maß der $\mathrm{N}(0,1)$-Verteilung ist.

4.4 Wir verwenden polynomielle Untermodelle im nicht parametrischen Regressionsmodell

$$Y_i = f\left(\tfrac{i}{n}\right) + \varepsilon_i, \quad i = 1, \ldots, 100,$$

mit $\varepsilon_1, \ldots, \varepsilon_n \overset{i.i.d.}{\sim} \mathrm{N}(0,1)$ und für die (wahren) Funktionen

(a) $f\colon [0,1] \to \mathbb{R}$, $f(x) = \sin(\frac{3}{2}\pi x)$,
(b) $f\colon [0,1] \to \mathbb{R}$, $f(x) = x^2$.

Verwenden Sie jeweils in 10.000 unabhängigen Simulationen AIC und BIC zur Wahl des Grades $p-1$ der linearen Modelle $f(x) = f_k(x) = \beta_0 + \cdots + \beta_{p-1}x^{p-1}$. Bestimmen Sie in allen vier Fällen die Monte-Carlo-Approximation des Vorhersagefehlers $\mathbb{E}[|Y - X\widehat{\beta}|^2]$ und stellen Sie die Wahl von p jeweils in einem Boxplot dar.

4.5 Die reellwertige Zufallsvariable X_p sei $\chi^2(p)$-verteilt mit $p \in \mathbb{N}$ Freiheitsgraden. Bestimmen Sie eine möglichst scharfe Abschätzung der Wahrscheinlichkeit

$$\mathbb{P}\big(X_p - \mathbb{E}[X_p] \geqslant \sqrt{\mathrm{Var}(X_p)}\kappa\big) \quad \text{für} \quad \kappa > 0.$$

Gehen Sie wie folgt vor:

(i) Berechnen Sie $\mathbb{E}[X_p]$, $\mathrm{Var}(X_p)$ und $\mathbb{E}[\exp(\lambda X_p)]$ für $\lambda > 0$.
(ii) Verwenden Sie die Markov-Ungleichung, um eine Abschätzung der gewünschten Wahrscheinlichkeit zu erhalten.
(iii) Wählen Sie λ optimal.

4.6 Das *empirische Skalarprodukt* sei definiert als $\langle f, g\rangle_n := \frac{1}{n}\sum_{i=1}^n f(\frac{i}{n})g(\frac{i}{n})$ bzw. $\langle x, g\rangle_n := \frac{1}{n}\sum_{i=1}^n x_i g(\frac{i}{n})$ für Funktionen f, g auf $[0,1]$ und einen Vektor $x \in \mathbb{R}^n$. Die empirische Norm ist gegeben durch $\|f\|_n^2 := \langle f, f\rangle_n$. Sei $(\varphi_k)_{k=1}^n \subseteq L^2([0,1])$ eine Orthonormalbasis bezüglich $\langle\cdot,\cdot\rangle_n$.
Für $f(x) = \sum_{k=1}^n \alpha_l \varphi_k(x), x \in [0,1]$, mit Koeffizienten $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ betrachten wir das Regressionsmodell

$$Y_i = f\left(\tfrac{i}{n}\right) + \varepsilon_i, \quad i = 1, \ldots, n,$$

mit unabängig und identisch $\mathrm{N}(0, \sigma^2)$-verteilten Fehlern $\varepsilon_i, i = 1, \ldots, n$.

(a) Weisen Sie nach, dass $\widehat{\alpha}_k := \langle Y, \varphi_k \rangle_n$ der Maximum-Likelihood-Schätzer von α_k für alle $k = 1, \ldots, n$ ist.
(b) Bestimmen Sie für $m \in \{1, \ldots, n\}$ den Fehler

$$\mathbb{E}\big[\|f - \widehat{f}_m\|_n^2\big] \quad \text{für} \quad \widehat{f}_m(x) := \sum_{k=1}^{m} \widehat{\alpha}_k \varphi_k(x).$$

Auf welches Minimierungsproblem führt eine optimale Wahl von m?
(c) Nehmen Sie an, dass es ein $s > 0$ und $0 < c < C$ gibt, sodass $ck^{-s} \leqslant \alpha_k \leqslant Ck^{-s}$ gilt. Wie wächst das optimale m in n?

4.7 Betrachten Sie die linearen Modelle $Y = X^{(p)}\beta^{(p)} + \varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$, $\beta^{(p)} \in \mathbb{R}^p$ und den Designmatrizen $X^{(p)} \in \mathbb{R}^{n \times p}$. Leiten Sie die explizite Darstellung des Leave-one-out-Kleinste-Quadrate-Schätzers her.

4.8 Betrachten Sie die mittleren Jannuartemperaturen in Berlin-Dahlem in den letzten 300 Jahren (siehe Beispiel 2.52). Verwenden Sie sowohl die polynomiale Regression mit AIC- und BIC-Wahl für den Polynomgrad als auch die Lasso-Methode. Untersuchen Sie die Abhängigkeit vom Tuning-Parameter. Liefern alle Verfahren ähnliche Resultate?

Die beiden folgenden Aufgaben beleuchten das Optimierungsproblem des Lasso-Schätzers genauer. Zu ihrer Lösung sind der Subdifferentialkalkül und die Karush-Kuhn-Tucker-Bedingungen aus der nichtlinearen Optimierung hilfreich.

4.9 Betrachten Sie das Regressionsmodel $Y = \mu + \varepsilon$ mit $\mu \in \mathbb{R}^n$, Beobachtungsfehlern $\varepsilon \in \mathbb{R}^n$ und eine Designmatrix $X \in \mathbb{R}^{n \times p}$. Zeigen Sie, dass das Lasso-Optimierungsproblem äquivalent durch

$$\widehat{\beta}' = \operatorname*{arg\,min}_{\beta' : |\beta'| \leqslant s} \frac{1}{n} |Y - X\beta'|^2$$

für einen Sparsity-Parameter $s > 0$ beschrieben werden kann. Welche Beziehung besteht zwischen s und dem Penalisierungsparameter λ des Lasso-Schätzers?

4.10 Betrachten Sie das lineare Modell $Y = X\beta + \varepsilon$ mit $X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p$ und $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$. Es sei $G(\beta) := \frac{1}{n}|Y - X\beta|^2$.

(a) Kann es mehr als eine Lösung des Optimierungsproblems

$$\min_{\beta \in \mathbb{R}^p} \{G(\beta) + \lambda |\beta|_1\} \tag{$*$}$$

geben? Begründen Sie Ihre Antwort.
(b) Zeigen Sie, dass eine notwendige und hinreichende Bedingung an eine Lösung $\widehat{\beta}$ von $(*)$ gegeben ist durch

$$\nabla_j G(\widehat{\beta}) = -\operatorname{sign}(\widehat{\beta}_j)\lambda, \qquad \text{falls } \widehat{\beta}_j \neq 0,$$

$$|\nabla_j G(\widehat{\beta})| \leqslant \lambda, \qquad \text{falls } \widehat{\beta}_j = 0.$$

(c) Beweisen Sie im Fall, dass $(*)$ keine eindeutige Lösung besitzt: Gilt $\nabla G_j(\widehat{\beta}) < \lambda$ für eine Lösung $\widehat{\beta}$ von $(*)$, dann gilt $\widehat{\beta}_j = 0$ für alle Lösungen. Inbesondere hängt die aktive Menge $\{j : \widehat{\beta}_j \neq 0, j = 1, \ldots, p\}$ nicht von der Lösung von $(*)$ ab.

Anhang A
Konzepte der Wahrscheinlichkeitstheorie

A.1 Grundbegriffe der Maßtheorie und Stochastik

Die Konzepte der Maß- und Wahrscheinlichkeitstheorie bilden die Grundlage der Stochastik und damit insbesondere der Statistik. Der Vollständigkeit halber wollen wir hier die grundlegenden Begriffe und Konzepte zusammenfassen, ohne zu sehr ins Detail zu gehen. Wir verzichten daher in diesem Kapitel auf Beweise.

Für eine grundlegende Einführung sei beispielsweise auf Küchler (2016) verwiesen. Eine tiefgehende und umfangreiche Darstellung der Maßtheorie findet man in Elstrodt (2005). Als Einführung in die Stochastik kann man Georgii (2007) empfehlen. Ein darüber hinausgehendes und weitreichendes Lehrbuch zur Stochastik ist Klenke (2008).

Wir modellieren die Gesamtheit aller möglichen Ausgänge eines Zufallsexperiments durch eine nichtleere Menge Ω. Teilmengen von Ω beschreiben Ereignisse. Ein Ereignis $A \subseteq \Omega$ tritt ein, wenn das Ergebnis $\omega \in \Omega$ des Zufallsexperiments in A liegt. Die Potenzmenge $\mathcal{P}(\Omega)$ ist definiert als die Menge aller Teilmengen von Ω. Um später jedem Ereignis in konsistenter Art und Weise eine Wahrscheinlichkeit zuordnen zu können, ist im Allgemeinen die gesamte Potzenmenge zu groß. Alle uns interessierenden Ereignisse fassen wir daher in einem Mengensystem $\mathcal{A}$ aus Ω zusammen. Je größer $\mathcal{A}$ ist, desto genauere Aussagen können wir über den Ausgang des Zufallsexperiments treffen.

Definition A.1 Für $\Omega \neq \emptyset$ heißt ein Mengensystem $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ **σ-Algebra über Ω**, falls folgende Eigenschaften erfüllt sind:

1. $\Omega \in \mathcal{A}$,
2. für alle $A \in \mathcal{A}$ gilt auch $A^c \in \mathcal{A}$ und
3. für alle $A_1, A_2, A_3, ... \in \mathcal{A}$ gilt auch $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{A}$.

Das Paar $(\Omega, \mathcal{A})$ heißt **messbarer Raum**.

Beispiel A.2 Es sei Ω eine nichtleere Menge.

(a) $\{\emptyset, \Omega\}$ und die Potenzmenge $\mathcal{P}(\Omega)$ sind σ-Algebren.

(b) Ist $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ ein System von Teilmengen von Ω, dann ist

$$\sigma(\mathcal{E}) := \bigcap \{\mathcal{A} : \mathcal{A} \text{ ist } \sigma\text{-Algebra aus } \Omega \text{ mit } \mathcal{E} \subseteq \mathcal{A}\}$$

die kleinste σ-Algebra, die $\mathcal{E}$ umfasst. $\sigma(\mathcal{E})$ heißt die von $\mathcal{E}$ **erzeugte σ-Algebra**.

(c) Ist Ω mit einer Metrik d versehen, dann ist die **Borel-σ-Algebra** $\mathcal{B}(\Omega)$ definiert als die kleinste σ-Algebra, die alle offenen Teilmengen von Ω enthält:

$$\mathcal{B}(\Omega) = \sigma\big(\{O \mid O \subseteq \Omega \text{ offen}\}\big)$$

Die Elemente dieser σ-Algebra heißen **Borel-Mengen**.

Die Borel-Mengen sind das (für uns) wichtigste Beispiel einer σ-Algebra. Wenn nichts anderes gefordert oder angenommen wurde, dann verwenden wir für metrische Räume stets die Borel-σ-Algebra.

Maße sind Abbildungen von einer σ-Algebra nach $[0, \infty]$, die jeder Menge aus der σ-Algebra eine „Größe“ zuordnen, beispielsweise Flächen oder Volumina. Es ist intuitiv, dass diese Abbildungen nichtnegativ und additiv sein sollten.

Definition A.3 Auf einem messbaren Raum $(\Omega, \mathcal{A})$ heißt eine nichtnegative Abbildung $\mu\colon \mathcal{A} \to [0, \infty]$ **Maß**, falls

1. $\mu(\emptyset) = 0$, die leere Menge also das Maß null hat, und
2. μ σ-additiv ist, das heißt für paarweise disjunkte Mengen $A_1, A_2, A_3, \ldots \in \mathcal{A}$ mit $A_i \cap A_j = \emptyset$ für alle $i \neq j$ gilt

$$\mu\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

Das Tripel $(\Omega, \mathcal{A}, \mu)$, bestehend aus einer nichtleeren Menge Ω, einer σ-Algebra $\mathcal{A}$ und einem Maß μ auf $\mathcal{A}$, heißt **Maßraum**. Die Mengen $A \in \mathcal{A}$ mit $\mu(A) = 0$ heißen μ-**Nullmengen**.

Aus der σ-Additivität von Maßen folgt insbesondere die σ-Subadditivität für beliebige Mengen $A_i \in \mathcal{A}, i \in \mathbb{N}$:

$$\mu\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) \leqslant \sum_{i\in\mathbb{N}} \mu(A_i)$$

In der Stochastik werden Maße verwendet, um den Ereignissen Wahrscheinlichkeiten zuzuordnen. Das beschränkt Wertebereich von Wahrscheinlichkeitsmaßen in natürlicher Weise auf $[0, 1]$.

Definition A.4 Auf einem messbaren Raum $(\Omega, \mathcal{A})$ heißt ein Maß μ **σ-endlich**, falls eine abzählbare Folge $(A_n)_{n\in\mathbb{N}}$ von Mengen aus $\mathcal{A}$ existiert, sodass $\bigcup_{n\in\mathbb{N}} A_n = \Omega$ sowie $\mu(A_n) < \infty$ für alle $n \in \mathbb{N}$ gilt. Ist sogar $\mu(\Omega) < \infty$, heißt μ **endlich**. Im Spezialfall $\mu(\Omega) = 1$ wird μ **Wahrscheinlichkeitsmaß** und $(\Omega, \mathcal{A}, \mu)$ **Wahrscheinlichkeitsraum** genannt.

Beispiel A.5 Wir betrachten einen beliebigen messbaren Raum $(\Omega, \mathcal{A})$ für eine nichtleere Menge Ω.

1. Das **Zählmaß** auf $\mathcal{A}$ ist definiert als

$$\mu\colon \mathcal{A} \to [0,\infty], \qquad \mu(A) = \begin{cases} |A|, & \text{falls } A \text{ endlich ist,} \\ +\infty, & \text{falls } A \text{ unendlich ist.} \end{cases}$$

   Jeder messbaren Menge $A \in \mathcal{A}$ wird also die Anzahl der Elemente in A zugeordnet. μ ist genau dann σ-endlich, wenn Ω abzählbar ist. Ist Ω endlich, dann ist $\mathbb{P}(A) := \frac{|A|}{|\Omega|}$, $A \subseteq \Omega$, ein Wahrscheinlichkeitsmaß, nämlich die Gleichverteilung.
2. Für ein fixiertes Element $\omega \in \Omega$ ist das **Dirac-Maß** in ω definiert via

$$\delta_\omega\colon \mathcal{A} \to [0,\infty], \qquad \delta_\omega(A) = \begin{cases} 1, & \text{falls } \omega \in A, \\ 0, & \text{sonst.} \end{cases}$$

   Man sieht leicht, dass δ_ω ein Wahrscheinlichkeitsmaß ist.
3. Ist $\Omega = \mathbb{R}$ und $\mathcal{A} = \mathcal{B}(\mathbb{R})$, dann ist das (eindeutig bestimmte) Maß λ, für das

$$\lambda\big((a,b]\big) = b - a, \qquad \text{für alle } a, b \in \mathbb{R}, a < b,$$

   gilt, das **Lebesgue-Maß** auf $\mathbb{R}$. Analog kann das Lebesgue-Maß auf $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ über Volumina von Quadern eindeutig bestimmt werden.

Meist betrachten wir Wahrscheinlichkeitsmaße aus einer der beiden folgenden Klassen.

Definition A.6 Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

1. Ist Ω eine endliche oder abzählbar unendliche Menge und $\mathcal{A} = \mathcal{P}(\Omega)$, so nennen wir $(\Omega, \mathcal{A}, \mathbb{P})$ einen **diskreten Wahrscheinlichkeitsraum**. In diesem Fall existiert stets eine **Zähldichte** von $\mathbb{P}$ gegeben durch $p\colon \Omega \to [0,1]$ mit $\sum_{\omega\in\Omega} p(\omega) = 1$ und $p(\omega) = \mathbb{P}(\{\omega\})$. Es gilt

$$\mathbb{P}(A) = \sum_{\omega \in A} p(\omega) \qquad \text{für alle } A \in \mathcal{A}.$$

2. Ist $\Omega = \mathbb{R}$ versehen mit der σ-Algebra $\mathcal{A} = \mathcal{B}(\mathbb{R})$ und existiert eine Funktion $f\colon \mathbb{R} \to [0,\infty)$ mit $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$ und

$$\mathbb{P}((a,b]) = \int_a^b f(x)\mathrm{d}x \qquad \text{für alle } a, b \in \mathbb{R}, a < b,$$

   so heißt f **Wahrscheinlichkeitsdichte** von $\mathbb{P}$ oder kurz **Dichte** von $\mathbb{P}$. Wir sprechen in diesem Fall von einer **stetigen Wahrscheinlichkeitsverteilung**.

Man beachte, dass durch $\mathbb{P}((a,b]) = \int_a^b f(x)\mathrm{d}x$ das Maß $\mathbb{P}$ bereits eindeutig durch seine Dichte festgelegt wird (was aus dem Eindeutigkeitssatz aus der Maßtheorie

folgt). Zudem ist der Fall von stetigen Verteilungen leicht auf den mehrdimensionalen Fall $\mathbb{R}^d$ zu übertragen. Wichtige Beispiele für diskrete und stetige Verteilungen werden in den Abschnitten A.2 und A.3 beschrieben.

Definition A.7 In einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ gilt ein Ereignis $A \in \mathcal{A}$ **$\mathbb{P}$-fast sicher** (kurz: $\mathbb{P}$-f.s.), wenn $\mathbb{P}(A) = 1$ gilt. Die Eigenschaft $E(\omega)$ sei für die Elemente $\omega \in \Omega$ sinnvoll. Dann sagt man, die Eigenschaft E gilt **$\mathbb{P}$-fast überall** auf Ω (kurz: $\mathbb{P}$-f.ü.), wenn es eine $\mathbb{P}$-Nullmenge $N \in \mathcal{A}$ gibt, sodass alle $\omega \in N^c$ die Eigenschaft E haben. Man sagt auch, dass E für **$\mathbb{P}$-fast alle** ($\mathbb{P}$-f.a.) $\omega \in \Omega$ gilt.

Nachdem wir einen Wahrscheinlichkeitsraum definiert haben, wollen wir als Nächstes Zufallsvariablen einführen.

Definition A.8 Seien $(\Omega, \mathcal{A})$ und $(\mathcal{X}, \mathcal{F})$ zwei messbare Räume. Eine Abbildung $X\colon \Omega \to \mathcal{X}$ heißt $(\mathcal{A}, \mathcal{F})$**-messbar**, falls

$$\forall A \in \mathcal{F}\colon\ X^{-1}(A) := \{\omega \in \Omega \mid X(\omega) \in A\} \in \mathcal{A}.$$

Hierbei bezeichnet $X^{-1}(A)$ das Urbild von A unter X. Gehen die σ-Algebren $\mathcal{A}$ und $\mathcal{F}$ aus dem Kontext eindeutig hervor, nennen wir X kurz **messbar**. Messbare Abbildungen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ werden **Zufallsvariablen** genannt.

Ist eine Zufallsvariable $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$-wertig, so sprechen wir auch kurz von einer reellen Zufallsvariable.

Beispiel A.9 Gegeben sei der messbare Raum $(\Omega, \mathcal{A})$.

1. Gilt $n \in \mathbb{N}$, $A_1, \ldots, A_n \in \mathcal{A}$ und $a_1, \ldots, a_n \in \mathbb{R} \setminus \{0\}$, so ist

   $$X(\omega) = \sum_{k=1}^{n} a_k \mathbb{1}_{A_k}(\omega), \quad \omega \in \Omega,$$

   $\mathcal{A}$-messbar. Funktionen dieser Gestalt nennen wir **einfache Funktionen**. Man kann zeigen, dass jede messbare nichtnegative Funktion durch eine monotone Folge einfacher Funktionen approximiert werden kann.
2. Jede stetige Funktion $X\colon \mathbb{R} \to \mathbb{R}$ ist Borel-messbar, das heißt $(\mathcal{B}(\mathbb{R}), \mathcal{B}(\mathbb{R}))$-messbar. Die umgekehrte Implikation gilt nicht, da die Dirichlet-Funktion $X(y) = \mathbb{1}_{\mathbb{Q}}(y)$, $y \in \mathbb{R}$, messbar, aber nirgendwo stetig ist.
3. Es seien Ω eine Menge, $(\mathcal{X}, \mathcal{F})$ ein messbarer Raum und $X\colon \Omega \to \mathcal{X}$ eine Abbildung. Dann ist $\sigma(X) := X^{-1}(\mathcal{F}) = \{X^{-1}(A) : A \in \mathcal{F}\}$ die kleinste σ-Algebra $\mathcal{A}$ auf Ω, sodass X eine $(\mathcal{A}, \mathcal{F})$-messbare Abbildung ist. Wir nennen $\sigma(X)$ die von X **erzeugte σ-Algebra**.

Die folgende Eigenschaft erlaubt es Stochastikern, sich auf Wahrscheinlichkeitsmaße auf dem Ergebnisraum eines Zufallexperiments zu beschränken, statt den gesamten zugrunde liegenden Wirkmechanismus beschreiben zu müssen.

Lemma A.10 *Es seien $(\Omega, \mathcal{A}, \mu)$ ein Maßraum, $(\mathcal{X}, \mathcal{F})$ ein messbarer Raum und $X\colon \Omega \to \mathcal{X}$ eine $(\mathcal{A}, \mathcal{F})$-messbare Abbildung. Durch*

$$\mu^X(A) := \mu(X^{-1}(A)), \quad A \in \mathcal{F},$$

*ist auf $\mathcal{F}$ ein Maß μ^X definiert, das als von X **induziertes Maß** oder als **Bildmaß** von X bezeichnet wird. Ist μ ein Wahrscheinlichkeitsmaß, so ist auch μ^X ein Wahrscheinlichkeitsmaß.*

Definition A.11 Für eine Zufallsvariable X auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ wird das induzierte Maß $\mathbb{P}^X$ auch **Verteilung der Zufallsvariable** X genannt, und wir schreiben dann $X \sim \mathbb{P}^X$.

Analog zu Definition A.6 sprechen wir von diskret bzw. stetig verteilten Zufallsvariablen X, falls ihre Bildmaße diskret mit Zähldichte p^X bzw. stetig mit Dichte f^X sind.

Bemerkung A.12 Wir betrachten einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$, einen messbaren Raum $(\mathcal{X}, \mathcal{F})$ und eine Zufallsvariable $X\colon (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{F})$. Für $A \in \mathcal{F}$ sind folgende Schreibweisen in der Stochastik üblich und zweckmäßig:

$$\{X \in A\} := \{\omega \in \Omega \mid X(\omega) \in A\} = X^{-1}(A) \qquad \text{sowie}$$
$$\mathbb{P}(X \in A) := \mathbb{P}(\{X \in A\}) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\}) = \mathbb{P}^X(A).$$

Die Verteilung einer reellen Zufallsvariable kann (eindeutig) durch ihre Verteilungsfunktion beschrieben werden. Letztere ist wie folgt definiert.

Definition A.13 Sei X eine reelle Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Die **Verteilungsfunktion** von X ist definiert als

$$F^X\colon \mathbb{R} \to [0,1], \quad F^X(x) := \mathbb{P}(X \leqslant x) = \mathbb{P}^X((-\infty, x]).$$

Für eine diskret auf einer abzählbaren Teilmenge S von $\mathbb{R}$ verteilten Zufallsvariable ist ihre Verteilungsfunktion eine stückweise konstante Treppenfunktion mit Sprüngen auf S. Für stetig verteilte Zufallsvariablen mit Wahrscheinlichkeitsdichte f^X ist die Verteilungsfunktion stetig und es gilt $F^X(x) = \int_{-\infty}^{x} f^X(y)\mathrm{d}y$.

Verteilungsfunktionen sind stets monoton wachsend, rechtsstetig, das heißt $F^X(x) = \lim_{y \downarrow x} F^X(y)$ für alle $x \in \mathbb{R}$, und es existieren die linken Grenzwerte $F^X(x-) := \lim_{y \uparrow x} F^X(y)$. Zudem gilt $\lim_{x \to -\infty} F^X(x) = 0$ sowie $\lim_{x \to \infty} F^X(x) = 1$. Diese Eigenschaften implizieren insbesondere, dass eine (verallgemeinerte) Inverse von F^X existiert.

Definition A.14 Sei $F\colon \mathbb{R} \to \mathbb{R}$ eine monoton wachsende Funktion, wobei wir den Definitionsbereich um $\pm\infty$ erweitern:

$$F(-\infty) := \lim_{x \to -\infty} F(x), \quad F(\infty) = \lim_{x \to \infty} F(x)$$

Die **verallgemeinerte Inverse** $F^-\colon \mathbb{R} \to \overline{\mathbb{R}} = [-\infty, \infty]$ von F ist definiert als

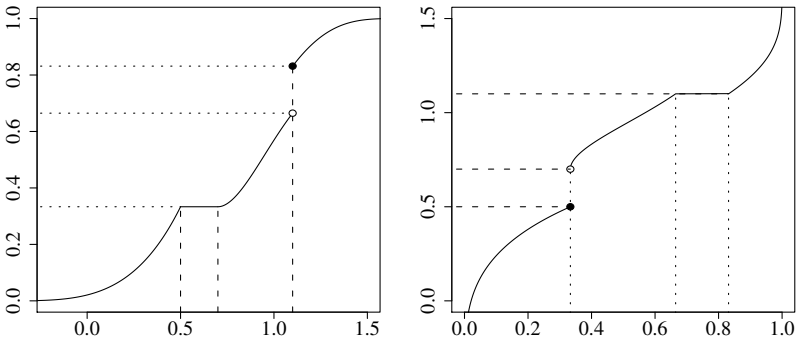

Abb. A.1 Eine Verteilungsfunktion (*links*) und die zugehörige Quantilfunktion (*rechts*)

$$F^-(y) := \inf\{x \in \mathbb{R} : F(x) \geqslant y\}, \quad y \in \mathbb{R},$$

mit der Konvention $\inf \emptyset := \infty$. Falls $F\colon \mathbb{R} \to [0,1]$ eine Verteilungsfunktion ist, heißt $F^-\colon [0,1] \to \overline{\mathbb{R}}$ **Quantilfunktion** von F.

Definition A.15 Es sei X eine Zufallsvariable auf dem Wahrscheinlichkeitsraum $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ mit zugehöriger Verteilungsfunktion F^X. Für jedes $p \in (0,1)$ ist die Menge aller p-**Quantile** von F^X gegeben durch

$$\big\{x \in \mathbb{R} :\ \mathbb{P}(X \leqslant x) \geqslant p \text{ und } \mathbb{P}(X \geqslant x) \geqslant 1-p\big\}.$$

Die Menge der p-Quantile ist ein abgeschlossenes Intervall mit der Obergrenze

$$\sup\{x \in \mathbb{R} : \mathbb{P}(X \geqslant x) \geqslant 1-p\}$$

und der Untergrenze $(F^X)^-(p)$. Falls F^X invertierbar ist, fallen Ober- und Untergrenze zusammen und das p-Quantil ist eindeutig. Im Verlauf des Buches wird klar, dass Quantile für die Konstruktion und Kalibrierung vieler statistischer Methoden von grundlegender Bedeutung sind.

Wir kommen nun zu zwei zentralen Begriffen der Wahrscheinlichkeitstheorie, nämlich bedingten Wahrscheinlichkeiten und Unabhängigkeit. Erstere formalisieren die Intuition, dass sich Wahrscheinlichkeiten von Ereignissen ändern, wenn über den Ausgang eines Zufallsexperiments bereits eine Teilinformation vorhanden ist.

Definition A.16 Es seien $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ Ereignisse mit $\mathbb{P}(B) > 0$. Dann ist die **bedingte Wahrscheinlichkeit** von A gegeben B definiert als

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

Es ist leicht nachzuprüfen, dass $\mathbb{P}(\cdot|B)$ tatsächlich ein Wahrscheinlichkeitsmaß ist, falls $\mathbb{P}(B) > 0$. Falls $\mathbb{P}(A|B) = \mathbb{P}(A)$ für zwei Ereignisse $A, B \in \mathcal{A}$ gilt, so enthält

B keinerlei Information über A. Die Ereignisse A und B sind also unabhängig. Umstellen zeigt, dass $\mathbb{P}(A|B) = \mathbb{P}(A)$ äquivalent zu $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ ist.

Definition A.17 Auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ heißen zwei Ereignisse $A, B \in \mathcal{A}$ **(stochastisch) unabhängig**, falls $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ gilt. Eine Familie von Ereignissen $(A_i)_{i\in I}$ mit nichtleerer Indexmenge $I \neq \emptyset$ heißt **(stochastisch) unabhängig**, falls $\mathbb{P}(\bigcap_{i\in J} A_i) = \prod_{i\in J} \mathbb{P}(A_i)$ für jede endliche Teilmenge $J \subseteq I$.

Der folgende Satz enthält zwei elementare, aber wichtige Eigenschaften von bedingten Wahrscheinlichkeiten.

Satz A.18 (Gesetz der totalen Wahrscheinlichkeit, Bayes-Formel) *Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $(B_i)_{i\in I}$ eine höchstens abzählbare Folge paarweiser disjunkter Mengen aus $\mathcal{A}$ mit $\bigcup_{i\in I} B_i = \Omega$ und $\mathbb{P}(B_i) > 0$ für alle $i \in I$.*

(i) Für jedes Ereignis $A \in \mathcal{A}$ gilt das Gesetz der totalen Wahrscheinlichkeit

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(B_i)\mathbb{P}(A|B_i).$$

(ii) Für jedes $A \in \mathcal{A}$ mit $\mathbb{P}(A) > 0$ und alle $i \in I$ gilt die Bayes-Formel

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(B_i)\mathbb{P}(A|B_i)}{\sum_{k\in I} \mathbb{P}(B_k)\mathbb{P}(A|B_k)} = \frac{\mathbb{P}(B_i)\mathbb{P}(A|B_i)}{\mathbb{P}(A)}.$$

Definition A.19 Eine Familie $(X_i)_{i\in I}$ von $(S_i, \mathcal{S}_i)$-wertigen Zufallsvariablen heißt **unabhängig**, falls für jede beliebige Wahl $A_i \in \mathcal{S}_i$ die Familie von Ereignissen $(X_i \in A_i)_{i\in I}$ unabhängig ist.

Ein Vektor oder eine Folge $X = (X_i)_{i\in I}$ von S_i-wertigen Zufallsvariablen X_i, $i \in I$, nimmt Werte im Produktraum $\prod_{i\in I} S_i$ an. Wollen wir die Verteilung X auf diesem beschreiben, müssen wir den Produktraum zunächst mit einer σ-Algebra versehen.

Definition A.20 Es seien $(\Omega_i, \mathcal{A}_i, \mathbb{P}_i)$, $i \in I$, Wahrscheinlichkeitsräume für eine beliebige nichtleere Indexmenge $I \neq \emptyset$. Die **Produkt-σ-Algebra** über dem kartesischen Produkt $\Omega := \prod_{i\in I} \Omega_i$ ist definiert als

$$\mathcal{A} := \bigotimes_{i\in I} \mathcal{A}_i := \sigma\big(\{\pi_i^{-1}(A_i) | i \in I, A_i \in \mathcal{A}_i\}\big),$$

wobei $\pi_i \colon \Omega \to \Omega_i$ die i-te Koordinatenprojektion bezeichnet. Gilt für ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf $\mathcal{A}$

$$\mathbb{P}\Big(\bigcap_{i\in J} \pi_i^{-1}(A_i)\Big) = \prod_{i\in J} \mathbb{P}_i(A_i) \qquad \text{für alle } J \subseteq I \text{ endlich}, A_i \in \mathcal{A}_i,$$

so heißt $\mathbb{P}$ **Produktmaß**. Wir schreiben $\mathbb{P} = \bigotimes_{i\in I} \mathbb{P}_i$.

Sind $X_i \colon \Omega \to \mathcal{X}_i$ für $i = 1, \dots, n$ Zufallsvariablen auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit Werten in $(\mathcal{X}_i, \mathcal{F}_i)$, dann ist auch der Vektor $(X_1, \dots, X_n)^\top$ eine messbare Abbildung, das heißt eine Zufallsvariable:

$$(X_1, \dots, X_n)^\top \colon \quad (\Omega, \mathcal{A}) \to \Big(\prod_{i=1}^n \mathcal{X}_i, \bigotimes_{i=1}^n \mathcal{F}_i \Big)$$

Die Verteilung $\mathbb{P}^{(X_1,\dots,X_n)}$ von $(X_1, \dots, X_n)^\top$ wird auch die **gemeinsame Verteilung** der Zufallsvariablen $X_1, \dots, X_n$ genannt.

Satz A.21 *Es seien $X_i \colon \Omega \to \mathcal{X}_i$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit Werten in den messbaren Räumen $(\mathcal{X}_i, \mathcal{F}_i)$ für $i \in I \neq \emptyset$. Dann ist die Familie $(X_i)_{i \in I}$ genau dann unabhängig, wenn für alle endlichen Teilmengen $J \subseteq I$ die gemeinsame Verteilung von $(X_i)_{i \in J}$ auf $(\prod_{i \in J} \mathcal{X}_i, \bigotimes_{i \in J} \mathcal{F}_i)$ das Produktmaß der Marginalverteilungen ist, das heißt wenn $\mathbb{P}^{(X_i)_{i \in J}} = \bigotimes_{i \in J} \mathbb{P}^{X_i}$.*

Man kann zeigen, dass für gegebene $(\Omega_i, \mathcal{A}_i, \mathbb{P}_i)_{i \in I}$ mit beliebiger Indexmenge $I \neq \emptyset$, stets ein Produktmaß $\mathbb{P} = \bigotimes_{i \in I} \mathbb{P}_i$ auf dem Produktraum $(\prod_{i \in I} \Omega_i, \bigotimes_{i \in I} \mathcal{A}_i)$ existiert. Insbesondere findet man also einen Wahrscheinlichkeitsraum, auf dem beliebig viele unabhängige Zufallsvariablen X_i mit vorgegeben Randverteilungen $\mathbb{P}^{X_i} = \mathbb{P}_i$ existieren. Sind (X_i) unabhängig und identisch verteilt, das heißt $\mathbb{P}^{X_i} = \mathbb{P}^{X_j}$ für alle $i, j \in I$ beschreiben wir dies häufig mit **i.i.d.**, was das englische *independent and identically distributed* abkürzt.

Um Momente, wie den Erwartungswert oder die Varianz, von reellwertigen Zufallsvariablen zu definieren, benötigen wir einen geeigneten Integralbegriff auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Die Integrationstheorie von Lebesgue erlaubt eine Integration bezüglich allgemeiner (Wahrscheinlichkeits-)Maße. Die Integralkonstruktion beruht auf drei Schritten. Für einfache Zufallsvariablen

$$X = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i} \quad \text{mit} \quad n \in \mathbb{N}, \alpha_1, \dots, \alpha_n \in \mathbb{R}, A_1, \dots, A_n \in \mathcal{A},$$

siehe Beispiel A.9, ist der Erwartungswert bzw. das Integral von X bezüglich $\mathbb{P}$ definiert als

$$\mathbb{E}[X] := \int_\Omega X \mathrm{d}\mathbb{P} := \sum_{i=1}^n \alpha_i \mathbb{P}(A_i).$$

Dieser Erwartungswert hängt nicht von der Darstellung von X ab, ist linear in X und monoton. Für jede nichtnegative Zufallsvariable X finden wir im zweiten Schritt stets eine Folge von einfachen Zufallsvariablen X_n, sodass $X_n(\omega) \uparrow X(\omega)$ für alle $\omega \in \Omega$ gilt, siehe Beispiel A.9. Dann definieren wir

$$\mathbb{E}[X] := \lim_{n \to \infty} \mathbb{E}[X_n].$$

Aufgrund der Monotonie des Erwartungswerts und der Folge (X_n) ist dieser Grenzwert wohldefiniert und man kann zeigen, dass diese Definition nicht von der Wahl der approximierenden Folge abhängt.

Definition A.22 Auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ ist die Menge aller endlich integrierbaren Zufallsvariablen gegeben durch

$$\mathcal{L}^1 := \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P}) := \left\{X\colon \Omega \to \mathbb{R} \text{ messbar} \,\middle|\, \mathbb{E}[|X|] < \infty\right\}.$$

Für $X \in \mathcal{L}^1$ definieren wir mit $X_+ := \max(X, 0)$ und $X_- := \max(-X, 0)$ den **Erwartungswert** als

$$\mathbb{E}[X] := \mathbb{E}[X_+] - \mathbb{E}[X_-] \in \mathbb{R}.$$

Wir schreiben $\mathbb{E}[X] = \int_\Omega X \,\mathrm{d}\mathbb{P} = \int_\Omega X(\omega)\,\mathbb{P}(\mathrm{d}\omega)$ und $\int_A X \,\mathrm{d}\mathbb{P} = \int_\Omega X(\omega)\mathbb{1}_A(\omega)\,\mathbb{P}(\mathrm{d}\omega)$ für $A \in \mathcal{A}$.

Analog wird das Integral bezüglich allgemeiner Maße definiert. Folgender Satz fasst die elementarsten Eigenschaften des Erwartungswerts zusammen.

Satz A.23 *Für $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ gilt:*

(i) $\mathbb{E}[X] = \int_{\mathbb{R}} x\mathbb{P}^X(\mathrm{d}x)$*, insbesondere hängt der Erwartungswert nur von der Verteilung $\mathbb{P}^X$ von X ab.*

(ii) Der Erwartungswert ist linear und monoton: Ist $Y \in \mathcal{L}^1$ eine weitere Zufallsvariable und sind $\alpha, \beta \in \mathbb{R}$, so gilt $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$. Aus $X \leqslant Y$ folgt $\mathbb{E}[X] \leqslant \mathbb{E}[Y]$.

(iii) Falls $X, Y \in \mathcal{L}^1$ unabhängig sind, so gilt $X \cdot Y \in \mathcal{L}^1$ und $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

In den Spezialfällen diskret bzw. stetig verteilter Zufallsvariablen ergibt sich:

Korollar A.24 *Es sei X eine Zufallsvariable auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$.*

(i) Besitzt X einen abzählbaren Wertebereich $X(\Omega) \subseteq \mathbb{R}$, so gilt $X \in \mathcal{L}^1$ genau dann, wenn $\sum_{x\in X(\Omega)} |x|\mathbb{P}(X = x)$ endlich ist. In diesem Fall gilt für den Erwartungswert

$$\mathbb{E}[X] = \sum_{x\in X(\Omega)} x\mathbb{P}(X = x).$$

(ii) Ist X eine Zufallsvariable mit Dichte $f^X\colon \mathbb{R} \to [0, \infty)$, so gilt $X \in \mathcal{L}^1$ genau dann, wenn $\int_{\mathbb{R}} |x| f^X(x)\mathrm{d}x$ endlich ist. In diesem Fall gilt für den Erwartungswert

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f^X(x)\mathrm{d}x.$$

Die Darstellung des Erwartungswerts in Abhängigkeit von der Verteilung der Zufallsvariable gilt auch allgemeiner. Ist $X\colon \Omega \to \mathbb{R}^d$ ein Zufallsvektor und $h\colon \mathbb{R}^d \to \mathbb{R}$ Borel-messbar, dann ist die Zufallsvariable $h(X)$ genau dann in $\mathcal{L}^1$, wenn $\int_{\mathbb{R}^d} |h(x)|\mathbb{P}^X(\mathrm{d}x) < \infty$. In diesem Fall ist der Erwartungswert gegeben durch

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}^d} h(x)\mathbb{P}^X(\mathrm{d}x).$$

Bemerkung A.25 Für eine $\mathbb{R}^d$-wertige Zufallsvariable X auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ ist die Familie der Funktionen $h_u(x) = \exp(\mathrm{i}\langle x, u\rangle)$ für $u \in \mathbb{R}^d$ und mit der imaginären Einheit $\mathrm{i} = \sqrt{-1}$ ein wichtiger Spezialfall. Der Erwartungswert von $h_u(X)$ wird durch

$$\begin{aligned}\varphi(u) :=& \mathbb{E}[e^{\mathrm{i}\langle u,X\rangle}] := \mathbb{E}[\mathrm{Re}\, e^{\mathrm{i}\langle u,X\rangle}] + \mathrm{i}\mathbb{E}[\Im e^{\mathrm{i}\langle u,X\rangle}] \\ =& \mathbb{E}[\cos(\langle u, X\rangle)] + \mathrm{i}\mathbb{E}[\sin(\langle u, X\rangle)]\end{aligned}$$

definiert. Da Kosinus und Sinus durch eins beschränkt sind, sind letztere Erwartungswerte stets wohldefiniert. Die Funktion $u \mapsto \varphi(u)$ heißt **charakteristische Funktion**, und man kann beweisen, dass die Verteilung von X eindeutig durch φ bestimmt wird.

Die Monome $h(x) = x^p$ führen uns auf folgende Definition:

Definition A.26 Eine Zufallsvariable X liegt in $\mathcal{L}^p$ für $p > 0$, falls $|X|^p \in \mathcal{L}^1$, also falls $\mathbb{E}[|X|^p] < \infty$ gilt. In diesem Fall heißt $\mathbb{E}[|X|^p]$ das **p-te absolute Moment** von X. Für $X \in \mathcal{L}^p$ und $p \in \mathbb{N}$ heißt $\mathbb{E}[X^p]$ das **p-te Moment** von X.

Man beachte, dass $\mathcal{L}^q \subseteq \mathcal{L}^p$ für $0 < p \leqslant q$ gilt, wobei hierfür essentiell ist, dass $\mathbb{P}$ ein endliches Maß ist. Von herausragender Bedeutung ist das zentrierte zweite Moment, also die Varianz, da sie die Streuung einer Zufallsvariable um ihren Erwartungswert angibt. Die lineare Abhängigkeit zwischen zwei Zufallsvariablen wird über ihre Korrelation quantifiziert.

Definition A.27 Für eine Zufallsvariable $X \in \mathcal{L}^2$ bezeichnet

$$\mathrm{Var}(X) := \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

die **Varianz** von X. Ihre Wurzel $\sqrt{\mathrm{Var}(X)}$ heißt **Standardabweichung** von X. Für $X, Y \in \mathcal{L}^2$ definiert

$$\mathrm{Cov}(X, Y) := \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

die **Kovarianz** zwischen X und Y. Gilt $\mathrm{Var}(X), \mathrm{Var}(Y) > 0$, ist die **Korrelation** gegeben durch

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Im Fall $\mathrm{Cov}(X, Y) = 0$ heißen X und Y **unkorreliert**.

Bemerkung A.28 Diese Definitionen lassen sich leicht auf d-dimensionale Zufallsvariablen übertragen $X = (X_1, ..., X_d)^\top$: Den Erwartungwert verstehen wir komponentenweise, das heißt

$$\mathbb{E}[X] := (\mathbb{E}[X_1], ..., \mathbb{E}[X_d])^\top.$$

Die Varianz des Zufallsvektors X definieren wir als erwarteten euklidischen Abstand vom Erwartungswert:

$$\mathrm{Var}(X) := \mathbb{E}[|X - \mathbb{E}[X]|^2]$$

Die Kovarianzen zwischen allen Komponenten von X werden in der **Kovarianzmatrix** zusammengefasst:

$$\mathbb{C}\mathrm{ov}(X) := \mathbb{E}\big[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\big] = \big(\mathrm{Cov}(X_i, X_j)\big)_{i,j=1,\dots,d} \in \mathbb{R}^{d\times d}$$

Satz A.29 *Für reellwertige Zufallsvariablen X, Y auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ gelten folgende Ungleichungen:*

1. *Für $X \in \mathcal{L}^1(\mathbb{P})$ gilt die **Markov-Ungleichung***

$$\mathbb{P}(|X| > \kappa) \leqslant \frac{\mathbb{E}[|X|]}{\kappa} \qquad \text{für alle } \kappa > 0.$$

*Gilt $X \in \mathcal{L}^2(\mathbb{P})$, folgt die **Tschebyscheff-Ungleichung***

$$\mathbb{P}(|X - \mathbb{E}[X]| > \kappa) \leqslant \frac{\mathrm{Var}(X)}{\kappa^2} \qquad \text{für alle } \kappa > 0.$$

2. *Für $X, Y \in \mathcal{L}^2(\mathbb{R})$ ist $XY \in \mathcal{L}^1(\mathbb{P})$, und es gilt die **Cauchy-Schwarz-Ungleichung***

$$\mathbb{E}[|XY|] \leqslant \mathbb{E}[X^2]^{1/2}\mathbb{E}[Y^2]^{1/2}.$$

*Sind allgemeiner $p, q \geqslant 1$ derart, dass $1 = \frac{1}{p} + \frac{1}{q}$ und $X \in \mathcal{L}^p(\mathbb{P}), Y \in \mathcal{L}^q(\mathbb{P})$, dann folgt $XY \in \mathcal{L}^1(\mathbb{P})$, und es gilt die **Hölder-Ungleichung***

$$\mathbb{E}[|XY|] \leqslant \mathbb{E}[|X|^p]^{1/p}\mathbb{E}[|Y|^q]^{1/q}.$$

3. *Ist $\varphi\colon \mathbb{R} \to \mathbb{R}$ konvex und $\varphi(X) \in \mathcal{L}^1(\mathbb{P})$, dann gilt die **Jensen-Ungleichung***

$$\varphi\big(\mathbb{E}[X]\big) \leqslant \mathbb{E}\big[\varphi(X)\big].$$

Die Momente einer Verteilung beschreiben Eigenschaften wie den Mittelwert, die Streuung, die Schiefe (zentriertes drittes Moment, englisch: *skewness*) oder die Wölbung (zentriertes viertes Moment, englisch: *kurtosis*). Darüber hinaus kann man sich fragen, ob die Folge der Momente $m_n = \mathbb{E}[X^n]$, $n \in \mathbb{N}$, für $X \sim \mathbb{P}$ ein Wahrscheinlichkeitsmaß $\mathbb{P}$ eindeutig bestimmt, falls alle Momente wohldefiniert und endlich sind. Diese Frage ist als das *Momentenproblem* bekannt und auch von statistischer Relevanz, siehe Momentenmethode in Abschnitt 1.2. Basierend auf der charakteristischen Funktion aus Bemerkung A.25 geben wir exemplarisch folgende Charakterisierung an:

Satz A.30 *Seien $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf dem messbaren Raum $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ und $X \sim \mathbb{P}$ mit existierenden Momenten $m_n = \mathbb{E}\big[X^n\big], n \in \mathbb{N}$, und charakteristischer Funktion $\varphi(u) = \mathbb{E}[e^{iuX}]$, $u \in \mathbb{R}$. Dann sind folgende Eigenschaften äquivalent:*

(i) φ *ist analytisch auf einer Umgebung um die Null.*
(ii) φ *ist analytisch auf* $\mathbb{R}$.
(iii) $\limsup_{n\to\infty}\left(|m_n|/n!\right)^{1/n} < \infty$.

Jede der drei Eigenschaften impliziert, dass die Momentenfolge $(m_n)_{n\in\mathbb{N}}$ *das Maß* $\mathbb{P}$ *eindeutig bestimmt.*

Aus Eigenschaft (iii) erhalten wir das folgende Korollar.

Korollar A.31 *Sei* $\mathbb{P}$ *ein Wahrscheinlichkeitsmaß auf* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *mit kompaktem Träger, das heißt, die kleinste abgeschlossene Menge mit Maß eins ist kompakt. Dann ist* $\mathbb{P}$ *eindeutig durch seine Momente bestimmt.*

Um den Einfluss verschiedener Parameterwahlen auf die Wahrscheinlichkeitsverteilung zu beschreiben, gibt es in statistischen Modellen nicht nur ein Wahrscheinlichkeitsmaß auf dem zugrunde liegenden Raum $(\Omega, \mathcal{A})$, sondern mehrere. Um diese zueinander in Beziehung zu setzen, spielt der Satz von Radon-Nikodym eine zentrale Rolle.

Definition A.32 Es seien μ und ν Maße auf einem messbaren Raum $(\Omega, \mathcal{A})$. ν heißt **absolutstetig** bezüglich μ, falls jede μ-Nullmenge eine ν-Nullmenge ist, das heißt, für jedes $A \in \mathcal{A}$ mit $\mu(A) = 0$ gilt auch $\nu(A) = 0$. Wir schreiben $\nu \ll \mu$.

Satz A.33 (Radon-Nikodym) *Es sei* $(\mathcal{X}, \mathcal{A})$ *ein messbarer Raum mit einem* σ*-endlichen Maß* μ *und einem Maß* ν*, das absolutstetig bezüglich* μ *ist. Dann existiert eine messbare Funktion* $f\colon X \to [0, \infty]$*, sodass*

$$\nu(A) = \int_A f \,\mathrm{d}\mu \qquad \textit{für alle} \qquad A \in \mathcal{A}.$$

f *ist* μ*-fast überall eindeutig bestimmt und heißt* ***Radon-Nikodym-Dichte*** *von* ν *bezüglich* μ *oder* μ***-Dichte von*** ν*. Wir schreiben* $\dfrac{\mathrm{d}\nu}{\mathrm{d}\mu} := f$.

Mit Blick auf Definition A.6 stellen wir fest, dass uns der Satz von Radon-Nikodym einen einheitlichen Rahmen für diskrete und stetige Verteilungen liefert: Die Zähldichte einer diskreten Verteilung ist gerade die Radon-Nikodym-Dichte bezüglich des Zählmaßes, während die Wahrscheinlichkeitsdichte einer stetigen Verteilung als Radon-Nikodym-Dichte bezüglich des Lebesgue-Maßes aufgefasst werden kann.

Bedingte Wahrscheinlichkeiten $\mathbb{P}(A \mid B)$ sind nur für Ereignisse B mit $\mathbb{P}(B) > 0$ wohldefiniert. Mithilfe von Radon-Nikodym-Dichten lässt sich dies verallgemeinern.

Definition A.34 Es seien $(\mathcal{X}, \mathcal{F}, \mu)$ und $(\mathcal{Y}, \mathcal{G}, \nu)$ Maßräume und X, Y Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit Werten in $\mathcal{X}$ bzw. $\mathcal{Y}$, deren gemeinsame Verteilung $\mathbb{P}^{X,Y}$ auf $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G})$ eine Dichte $f^{X,Y}\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ bezüglich dem Produktmaß $\mu \otimes \nu$ besitzt. Dann heißt

$$f^{Y|X=x}(y) := \frac{f^{X,Y}(x,y)}{f^X(x)} \quad \text{mit} \quad f^X(x) := \int_{\mathcal{Y}} f(x,z)\nu(\mathrm{d}z)$$

für alle $x \in \mathcal{X}$ mit positiver Randdichte $f^X(x) > 0$ **bedingte Dichte** von Y gegeben $X = x$. Für alle $x \in \mathcal{X}$ mit $f^X(x) = 0$ wird $f^{Y|X=x}$ beliebig gesetzt, zum Beispiel null. Für messbare Mengen $A \in \mathcal{G}$ und messbare Funktionen $\varphi\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ bezeichnet

$$\mathbb{P}(Y \in A \mid X = x) := \int_A f^{Y|X=x}(y)\, \nu(\mathrm{d}y)$$

die bedingte Wahrscheinlichkeit für $Y \in A$, gegeben $X = x$, und

$$\mathbb{E}[\varphi(X,Y) \mid X = x] := \int_{\mathcal{Y}} \varphi(x,y) f^{Y|X=x}(y)\, \nu(\mathrm{d}y)$$

den **bedingten Erwartungswert** von $\varphi(X,Y)$, gegeben $X = x$, sofern das Integral wohldefiniert ist.

In Analogie zu $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$ gilt also $f^{X,Y}(x,y) = f^{Y|X=x}(y) f^X(x)$. Die Formel von der totalen Wahrscheinlichkeit $\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$ für eine Partition $\bigcup_i B_i = \Omega$ findet ihre Entsprechung in

$$\mathbb{P}(Y \in A) = \int_{\mathcal{X}} \mathbb{P}(Y \in A \mid X = x) f^X(x)\, \mu(\mathrm{d}x).$$

Allgemeiner gilt auch

$$\mathbb{E}[\varphi(X,Y)] = \int_{\mathcal{X}} \mathbb{E}[\varphi(X,Y) \mid X = x] f^X(x)\, \mu(\mathrm{d}x).$$

Die Bayes-Formel aus Satz A.18 verallgemeinert sich für bedingte Dichten zu folgendem Zusammenhang:

Satz A.35 (Bayes-Formel) *In der Situation von Definition A.34 gilt*

$$f^{Y|X=x}(y) = \frac{f^{X|Y=y}(x) f^Y(y)}{\int_{\mathcal{Y}} f^{X|Y=z}(x) f^Y(z) \mathrm{d}z} = \frac{f^{X|Y=y}(x) f^Y(y)}{f^X(x)} \quad \textit{für } \mathbb{P}^X\textit{-f.a. } x \in \mathcal{X}.$$

Mit dem Satz von Radon-Nikodym lässt sich eine abstrakte Definition von $\mathbb{E}[\varphi(X,Y) \mid X = x]$ für beliebige Zufallsvariablen X, Y mit $\varphi(X,Y) \in \mathcal{L}^1$ geben, die im Fall einer gemeinsamen Dichte $f^{X,Y}$ der obigen entspricht. Mit Indikatorfunktionen φ ergeben sich dann auch entsprechende bedingte Wahrscheinlichkeiten. Für die Zwecke dieses Buches reicht obige Definition aus.

Wir schließen dieses Grundlagenkapitel mit den zentralen Sätzen aus der Wahrscheinlichkeitstheorie ab: dem Gesetz der großen Zahlen und dem zentralen Grenzwertsatz. Im Gegensatz zur klassischen Analysis unterscheiden wir in der Stochastik zwischen verschiedenen Konvergenzarten.

Definition A.36 Es sei $(\mathcal{X}, d)$ ein metrischer Raum. Eine Folge von Wahrscheinlichkeitsmaßen $(\mathbb{P}_n)_{n\in\mathbb{N}}$ auf $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ **konvergiert schwach** gegen ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, wenn für alle stetigen und beschränkten Funktion

$\varphi\colon \mathcal{X} \to \mathbb{R}$

$$\lim_{n\to\infty} \int_{\mathcal{X}} \varphi \, \mathrm{d}\mathbb{P}_n = \int_{\mathcal{X}} \varphi \, \mathrm{d}\mathbb{P}$$

gilt. Wir schreiben $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$, wobei das w für *weak* steht, oder $\mathbb{P}_n \xrightarrow{d} \mathbb{P}$ mit d für *distribution*. Eine Folge von Zufallsvariablen $(X_n)_{n\in\mathbb{N}}$ mit Werten in $(\mathcal{X}, d)$ **konvergiert in Verteilung** bzw. **schwach** gegen eine Zufallsvariable X in $\mathcal{X}$, falls $\mathbb{P}^{X_n} \xrightarrow{w} \mathbb{P}^X$. Wir schreiben kurz $X_n \xrightarrow{w} X$ oder $X_n \xrightarrow{d} X$.

Die schwache Konvergenz der Folge $(X_n)_{n\in\mathbb{N}}$ gegen die Zufallsvariable X ist also äquivalent zu

$$\lim_{n\to\infty} \mathbb{E}[\varphi(X_n)] = \mathbb{E}[\varphi(X)]$$

für alle stetigen und beschränkten Funktionen $\varphi\colon \mathcal{X} \to \mathbb{R}$. Ist die Folge $(X_n)_{n\in\mathbb{N}}$ reellwertig, so konvergiert sie genau dann in Verteilung gegen die Zufallsvariable X, wenn die Folge ihrer Verteilungsfunktionen $(F^{X_n})_{n\in\mathbb{N}}$ an jeder Stetigkeitsstelle der Verteilungsfunktion F^X punktweise gegen F^X konvergiert, das heißt, für alle Stetigkeitsstellen x von F^X gilt

$$\lim_{n\to\infty} F^{X_n}(x) = F^X(x).$$

Definition A.37 Eine Folge $(X_n)_{n\in\mathbb{N}}$ von Zufallsvariablen auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit Werten in einem normierten Raum $(\mathcal{X}, |\cdot|)$ konvergiert **stochastisch** bzw. **in Wahrscheinlichkeit** gegen die Zufallsvariable X, wenn für alle $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$$

gilt. Wir schreiben kurz $X_n \xrightarrow{\mathbb{P}} X$. Die Folge $(X_n)_{n\in\mathbb{N}}$ konvergiert **fast sicher** gegen die Zufallsvariable X, wenn

$$\mathbb{P}\big(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\big) = 1$$

ist. In diesem Fall schreiben wir $X_n \xrightarrow{\text{f.s.}} X$.

Man beachte, dass fast sichere Konvergenz stochastische Konvergenz impliziert und dass aus stochastischer Konvergenz schwache Konvergenz folgt.

Satz A.38 ***(Starkes Gesetz der großen Zahlen)*** *Es sei $(X_i)_{i\in\mathbb{N}}$ eine Folge von Zufallsvariablen in $\mathcal{L}^1$ mit demselben Erwartungswert $\mu = \mathbb{E}[X_i]$. Weiter sei eine der beiden folgenden Bedingungen erfüllt:*

(i) $(X_i)_{i\geqslant 1}$ sind identisch verteilt und paarweise unabhängig.
(ii) $(X_i)_{i\geqslant 1}$ liegen in $\mathcal{L}^2$, sind paarweise unkorreliert, und es gilt $\sup_{i\in\mathbb{N}} \mathrm{Var}(X_i) < \infty$.

Dann gilt

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{f.s.}} \mu.$$

Satz A.39 ***(Zentraler Grenzwertsatz)*** *Es sei $(X_i)_{i\geqslant 1}$ eine Folge unabhängiger und identisch verteilter Zufallsvariablen in $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mu = \mathbb{E}[X_1]$ und $\sigma^2 = \mathrm{Var}(X_1)$. Dann erfüllt ihre standardisierte Summe*

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \xrightarrow{d} \mathrm{N}(0,1),$$

wobei $\mathrm{N}(0,1)$ *die Standardnormalverteilung bezeichnet. Insbesondere gilt für $a < b$ also $\lim_{n\to\infty} \mathbb{P}(a < Z_n \leqslant b) = \Phi(b) - \Phi(a)$ mit der Verteilungsfunktion Φ der Standardnormalverteilung.*

Ein weiterer, wichtiger Satz ist das *continuous mapping*-Theorem, mit dem man Konsistenz und Grenzwertverteilungseigenschaften übertragen kann.

Satz A.40 (Continuous Mapping) *Seien die Abbildung $f\colon \mathbb{R}^d \mapsto \mathbb{R}^k$ stetig und $(X_n)_{n\in\mathbb{N}}$ eine Folge von d-dimensionalen Zufallsvariablen, die schwach, fast sicher bzw. stochastisch gegen $X \in \mathbb{R}^d$ konvergiert. Dann konvergiert auch $f(X_n)$ schwach, fast sicher bzw. stochastisch gegen $f(X)$.*

Folgender Kalkül der stochastischen Ordnung verallgemeinert Landaus O-Symbol der Analysis und ist oft hilfreich.

Definition A.41 Für eine Folge reellwertiger Zufallsvariablen X_n und positive Zahlen a_n schreiben wir $X_n = O_{\mathbb{P}}(a_n)$, falls

$$\lim_{R\to\infty} \limsup_{n\to\infty} \mathbb{P}(|X_n| > Ra_n) = 0,$$

und sagen, dass X_n die **stochastische Ordnung** a_n besitzt. Im Fall $X_n = O_{\mathbb{P}}(1)$ nennen wir die Folge (X_n) **stochastisch beschränkt**.

Ist (X_n) deterministisch, so gilt $X_n = O_{\mathbb{P}}(a_n)$ genau dann, wenn $X_n \leqslant Ca_n$ für eine Konstante $C > 0$ gilt, das heißt $X_n = O(a_n)$. Für eine Nullfolge (a_n) folgt aus $X_n = O_{\mathbb{P}}(a_n)$, dass $\mathbb{P}(|X_n| > \varepsilon)$ für jedes $\varepsilon > 0$ gegen null konvergiert, also $X_n \xrightarrow{\mathbb{P}} 0$. Stochastische Beschränktheit einer Folge (X_n) ist auch als Straffheit von Verteilungen in der Wahrscheinlichkeitstheorie bekannt. Weitere wichtige Eigenschaften werden in dem folgenden Satz zusammengefasst.

Satz A.42 (Eigenschaften des $O_{\mathbb{P}}$-Kalküls) *Für reellwertige Zufallsvariablen X_n, Y_n und positive Zahlen a_n, b_n gilt:*

1. *Aus $\mathbb{E}[|X_n|^p]^{1/p} \leqslant Ca_n$ für Konstanten $C > 0$ und $p > 0$ folgt $X_n = O_{\mathbb{P}}(a_n)$ (aus L^p-Beschränktheit folgt stochastische Beschränktheit).*
2. *Aus $a_n^{-1} X_n \xrightarrow{d} Y$ für eine Zufallsvariable Y folgt $X_n = O_{\mathbb{P}}(a_n)$ (Konvergenz in Verteilung impliziert stochastische Beschränktheit).*
3. *Aus $X_n = O_{\mathbb{P}}(a_n)$, $Y_n = O_{\mathbb{P}}(b_n)$ folgt $X_n + Y_n = O_{\mathbb{P}}(a_n + b_n)$ und $X_n Y_n = O_{\mathbb{P}}(a_n b_n)$, symbolisch:*

$$O_{\mathbb{P}}(a_n) + O_{\mathbb{P}}(b_n) = O_{\mathbb{P}}(a_n + b_n), \quad O_{\mathbb{P}}(a_n) O_{\mathbb{P}}(b_n) = O_{\mathbb{P}}(a_n b_n).$$

Wir verwenden die letzte Eigenschaft auch für zufällige Matrizen M_n und Vektoren v_n in der Form, dass $\|M_n\| = O_{\mathbb{P}}(a_n)$ (mit Spektralnorm $\|\cdot\|$) und $|v_n| = O_{\mathbb{P}}(b_n)$ implizieren $|M_n v_n| = O_{\mathbb{P}}(a_n b_n)$, was wegen $|M_n v_n| \leqslant \|M_n\||v_n|$ offensichtlich ist.

A.2 Diskrete Verteilungen

Im Folgenden sollen häufig auftretende diskrete Verteilungen eingeführt werden. Wir beginnen mit dem einfachsten Fall, bei dem es nur zwei mögliche Versuchsausgänge gibt.

Definition A.43 Eine Zufallsvariable $X \in \{0, 1\}$ heißt **Bernoulli-verteilt**, falls $\mathbb{P}(X = 1) = p$ und $\mathbb{P}(X = 0) = 1 - p$ für eine Erfolgswahrscheinlichkeit $p \in [0, 1]$ gilt. Man schreibt $X \sim \mathrm{Ber}(p)$.

Für $X \sim \mathrm{Ber}(p)$ gilt $\mathbb{E}[X] = p$ und $\mathrm{Var}(X) = p(1 - p)$.

Auf endlichen Grundräumen ist die Gleichverteilung die wohl wichtigste, insbesondere mit Blick auf Urnenmodelle.

Definition A.44 Ist $\Omega \neq \emptyset$ ein endlicher Grundraum mit Kardinatilität $|\Omega| \in \mathbb{N}$, dann ist die diskrete **Gleichverteilung** $\mathrm{U}(\Omega)$ gegeben durch die Zähldichte

$$\mathrm{U}(\Omega)(\{\omega\}) = \frac{1}{|\Omega|} \qquad \text{für alle } \omega \in \Omega.$$

Zur Beschreibung der Anzahl der Erfolge in einer Serie von $n \in \mathbb{N}$ gleichartigen und unabhängigen Bernoulli-Versuchen nutzt man die Binomialverteilung.

Definition A.45 Für $n \in \mathbb{N}$ und $p \in [0, 1]$ ist die **Binomialverteilung** $\mathrm{Bin}(n, p)$ durch die Zähldichte

$$\mathrm{Bin}(n, p)(\{k\}) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad k \in \{0, 1, \ldots, n\},$$

gegeben.

Der Erwartungswert einer binomialverteilten Zufallsvariable $X \sim \mathrm{Bin}(n, p)$ ist np. Die Varianz beträgt $np(1 - p)$.

Wir verallgemeinern ein Binomialexperiment auf mehrere mögliche Versuchsausgänge. Betrachten wir also wieder $n \in \mathbb{N}$ unabhängige und gleich verteilten Durchläufe, wobei jeweils $s \in \mathbb{N}$ verschiedene Versuchsausgänge $\{1, ..., s\}$ möglich sind und $j \in \{1, ..., s\}$ mit Wahrscheinlichkeit p_j eintritt. Bezeichne X_j die Anzahl der Durchgänge aus n Versuchen, bei denen j aufgetreten ist, dann ist der Vektor $X := (X_1, ..., X_s)$ multinomialverteilt.

Definition A.46 Für $n, s \in \mathbb{N}$ und $p_1, \ldots, p_s \in [0, 1]$ mit $\sum_{j=1}^{s} p_j = 1$ ist eine Zufallvariable $X := (X_1, ..., X_s) \sim \mathrm{Mult}(n, p_1, ..., p_s)$ **multinomial-verteilt**, falls für $k = (k_1, ..., k_s) \in \mathcal{X}$ mit

$$\mathcal{X} := \{k = (k_1, ..., k_s) : k_1, ..., k_s \in \mathbb{N}_0 \text{ mit } k_1 + ... + k_s = n\}$$

gilt:

$$\mathbb{P}(X = k) = \frac{n!}{k_1! \cdot k_2! \cdot ... \cdot k_s!} \cdot p_1^{k_1} \cdot p_2^{k_2} \cdot ... \cdot p_s^{k_s}$$

Wir nennen $\frac{n!}{k_1! \cdot k_2! \cdot ... \cdot k_s!}$ **Multinomialkoeffizient**.

Das Binomial- und das Multinomialmodell können wir zur Beschreibung von Urnenmodellen nutzen, wobei die Erfolgswahrscheinlichkeiten dem relativen Anteil der Kugeln in einer bestimmten Farbe entsprechen. Da diese Wahrscheinlichkeiten in jedem Versuchsdurchgang, also in jeder Ziehung, gleich sind, handelt es sich um eine Ziehung mit Zurücklegen. Was passiert ohne Zurücklegen? In einer Grundgesamtheit gäbe es $N \in \mathbb{N}$ Elemente, wobei jedes Element nur eine von zwei möglichen Ausprägungen hat (zum Beispiel Erfolg/Misserfolg oder rot/blau). Sind $M \leqslant N$ Elemente mit der gewünschten Eigenschaft (Erfolg oder rot) in der Grundgesamtheit, so gibt die hypergeometrische Verteilung die Wahrscheinlichkeit an, beim Ziehen ohne Zurücklegen in der Stichprobe $k \leqslant M$ Elemente mit der gewünschten Eigenschaft zu finden.

Definition A.47 Für $N, M, n \in \mathbb{N}$ mit $M \leqslant N$ besitzt eine Zufallsvariable $X \sim \mathrm{Hyp}(N, M, n)$ die **hypergeometrische Verteilung**, falls

$$\mathbb{P}(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \qquad k \in \{0, 1, \ldots, n\}.$$

Der Erwartungswert einer $\mathrm{Hyp}(N, M, n)$-verteilten Zufallsvariable X ist $\mathbb{E}[X] = n\frac{M}{N}$.

Bemerkung A.48 Der Unterschied zwischen der hypergeometrischen und der Binomialverteilung ist das Zurücklegen. Dieser Effekt ist bei einem relativ kleinen Stichprobenumfang n im Vergleich zu M, das heißt wenn $n/M \to 0$, vernachlässigbar gering. Tatsächlich gilt für alle $0 \leqslant k \leqslant n$

$$\begin{aligned}\binom{M}{k} &= \frac{M^k}{k!}\frac{M(M-1)\cdots(M-k+1)}{M^k} \\ &= \frac{M^k}{k!}\Big(1 - \frac{(M-1)}{M}\Big)\cdots\Big(1 - \frac{(M-k+1)}{M}\Big) = \frac{M^k}{k!}(1 + o(1)) \quad \text{für } \frac{n}{M} \to 0.\end{aligned}$$

Für $X \sim \mathrm{Hyp}(N, M, n)$ gilt also punktweise für jedes $k \in \{0, 1, \ldots, n\}$:

$$\begin{aligned}\mathbb{P}(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} &= \binom{n}{k}\frac{M^k(N-M)^{n-k}}{N^n}(1 + o(1)) \\ &= \binom{n}{k}\Big(\frac{M}{N}\Big)^k\Big(1 - \frac{M}{N}\Big)^{n-k}(1 + o(1)).\end{aligned}$$

Folglich kann man für $n/M \to 0$ die hypergeometrische Verteilung $\mathrm{Hyp}(N, M, n)$ durch die einfacher zu handhabende Binomialverteilung $\mathrm{Bin}(n, \frac{M}{N})$ approximieren.

Abschließend führen wir noch eine Verteilung ein, die auch als Verteilung seltener Ereignisse bezeichnet wird (wegen des Grenzwertsatzes von Poisson).

Definition A.49 Die **Poisson-Verteilung** $\mathrm{Poiss}(\lambda)$ mit dem Intensitätsparameter $\lambda > 0$ ist durch die Zähldichte

$$\mathrm{Poiss}(\lambda)(\{k\}) = \frac{\lambda^k}{k!} \exp(-\lambda), \quad k \in \mathbb{N}_0,$$

gegeben.

Für eine Zufallsvariable $X \sim \mathrm{Poiss}(\lambda)$ gilt $\mathbb{E}[X] = \mathrm{Var}(X) = \lambda$.

A.3 Stetige Verteilungen

In diesem Kapitel wollen wir einige wichtige stetige Wahrscheinlichkeitsverteilungen zusammenfassend einführen. Diese sind insbesondere durch ihre Dichte charakterisiert. Wir betrachten dabei stets die Borel-σ-Algebra auf dem jeweiligen (Teil-)Raum der reellen Zahlen. Im einfachsten Fall ist die Dichte konstant.

Definition A.50 Sei $\Omega = [a, b]$ für $a < b \in \mathbb{R}$. Mit $\mathrm{U}([a, b])$ bezeichnet man die **Gleichverteilung** auf $[a, b]$. Die Dichte f einer Zufallsvariable $X \sim \mathrm{U}([a, b])$ ist für alle $x \in \mathbb{R}$ gegeben durch

$$f(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x).$$

Der Erwartungswert und die Varianz einer $\mathrm{U}([a, b])$-verteilten Zufallsvaribale X sind $\mathbb{E}[X] = \frac{a+b}{2}$ und $\mathrm{Var}(X) = \frac{1}{12}(b-a)^2$.

Im zentralen Grenzwertsatz ist uns bereits die Normalverteilung begegnet, die von fundamentaler Bedeutung ist.

Definition A.51 Eine Zufallsvariable X auf $\mathbb{R}$ mit der Wahrscheinlichkeitsdichte $f : \mathbb{R} \to \mathbb{R}$,

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

heißt **normalverteilt** (oder auch **Gauß-verteilt**) mit den Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$. Wir schreiben $X \sim \mathrm{N}(\mu, \sigma^2)$. Die Verteilungsfunktion der Standardnormalverteilung $\mathrm{N}(0, 1)$ wird mit

$$\Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t, \qquad x \in \mathbb{R},$$

bezeichnet.

Für eine Zufallsvaribale $X \sim \mathrm{N}(\mu, \sigma)$ geben die beiden Parameter gerade den Erwartungswert $\mathbb{E}[X] = \mu$ und die Varianz $\mathrm{Var}(X) = \sigma^2$ an. Die Normalverteilung konzentriert sich sehr stark um ihren Mittelwert: Für $X \sim \mathrm{N}(0,1)$ gilt nämlich

$$\mathbb{P}(|X| \geqslant t) = 2\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\mathrm{d}x \leqslant \frac{2}{\sqrt{2\pi}}\int_t^\infty \frac{x}{t} e^{-x^2/2}\mathrm{d}x = \frac{\sqrt{2}}{\sqrt{\pi}t} e^{-t^2/2}. \quad \text{(A.1)}$$

Definition A.52 Für $d \in \mathbb{N}$ ist ein d-dimensionaler reeller Zufallsvektor $X = (X_1, \ldots, X_d)$ **mehrdimensional** (oder auch **multivariat**) **normalverteilt** mit Erwartungsvektor $\mu \in \mathbb{R}^d$ und positiv definiter Kovarianzmatrix $\Sigma \in \mathbb{R}^{d\times d}$, wenn sie eine Dichtefunktion der Form

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big), \quad x \in \mathbb{R}^d,$$

besitzt. Man schreibt $X \sim \mathrm{N}(\mu, \Sigma)$.

Für einen Zufallsvektor $X \sim \mathrm{N}(\mu, \Sigma)$ erhalten wir $\mathbb{E}[X_i] = \mu_i$ und $\mathrm{Cov}(X_i, X_j) = \Sigma_{i,j}$ für alle $i, j \in \{1, \ldots, d\}$ bzw. $\mathbb{E}[X] = \mu$ und $\mathbb{C}\mathrm{ov}(X) = \Sigma$ in Vektor- bzw. Matrixnotation.

Beispiel A.53 Für einen Zufallsvektor $X \sim \mathrm{N}(\mu, \Sigma)$ mit $\mu \in \mathbb{R}^d$ und positiv definiter Kovarianzmatrix $\Sigma \in \mathbb{R}^{d\times d}$ gilt für die Lineartransformation $AX + b \in \mathbb{R}^p$ mit $A \in \mathbb{R}^{p\times d}$, $b \in \mathbb{R}^p$, dass

$$AX + b \sim \mathrm{N}(A\mu + b, A\Sigma A^\top).$$

Sind $X_1, ..., X_m \sim \mathrm{N}(0,1)$ unabhängig, dann kann man auch die Dichte von $X := \sum_{i=1}^m X_i^2$ explizit bestimmen und erhält die folgende Verteilung:

Definition A.54 Eine reellwertige Zufallsvariable X ist **χ^2-verteilt** mit m Freiheitsgraden, geschrieben $X \sim \chi^2(m)$, wenn ihre Verteilung durch die Dichte

$$f_X(x) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{m/2-1} e^{-x/2} \mathbb{1}_{\{x>0\}}$$

gegeben ist, wobei $\Gamma(x) := \int_0^\infty t^{x-1}\exp(-t)dt$ die Gammafunktion bezeichnet.

Es folgen weitere stetige Verteilungen.

Definition A.55 Mit $\mathrm{Exp}(\lambda)$, $\lambda > 0$, bezeichnet man die **Exponentialverteilung**. Die Dichte f einer Zufallsvariable $X \sim \mathrm{Exp}(\lambda)$ ist für alle $x \in \mathbb{R}$ gegeben durch

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{\{x\geqslant 0\}}.$$

Der Erwartungswert und die Varianz der $\mathrm{Exp}(\lambda)$-Verteilung sind $\mathbb{E}[X] = \frac{1}{\lambda}$ und $\mathrm{Var}(X) = \frac{1}{\lambda^2}$. Eine Verallgemeinerung von Exponential- und χ^2-Verteilung ist durch die Gammaverteilung gegeben.

Definition A.56 Eine Zufallsvariable X mit der Wahrscheinlichkeitsdichte $f\colon \mathbb{R} \to \mathbb{R}$,

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}_{\{x>0\}},$$

heißt **gammaverteilt** mit den positiven, reellen Parametern $\alpha, \beta > 0$, geschrieben $X \sim \Gamma(\alpha, \beta)$.

Wir erhalten $\mathrm{Exp}(\lambda) = \Gamma(1, \lambda)$ und $\chi^2(m) = \Gamma(m/2, 1/2)$. Für eine gammaverteilte Zufallsvariable X gilt $\mathbb{E}[X] = \alpha/\beta$ und $\mathrm{Var}(X) = \alpha/\beta^2$.

Definition A.57 Die **Betaverteilung** $\mathrm{Beta}(\alpha, \beta)$ ist eine stetige Wahrscheinlichkeitsverteilung auf $([0,1], \mathcal{B}([0,1]))$ mit den Parametern $\alpha, \beta > 0$. Ihre Wahrscheinlichkeitsdichte ist

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbb{1}_{\{x\in[0,1]\}}$$

mit der **Betafunktion**

$$B(\alpha, \beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \mathrm{d}x.$$

Der Erwartungswert und die Varianz einer Zufallsvariable $X \sim \mathrm{Beta}(\alpha, \beta)$ sind

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \quad \text{und} \quad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

Beispiel A.58 Sortieren wir $n \in \mathbb{N}$ unabhängige und $\mathrm{U}([0,1])$-verteilte Zufallsvariablen $U_1, \ldots, U_n$ der Größe nach, erhalten wir die Ordnungsstatistiken $U_{(1)} \leqslant \ldots \leqslant U_{(n)}$. Für die j-te Ordnungstatistik gilt $U_{(j)} \sim \mathrm{Beta}(j, n-j+1)$ und insbesondere $\mathbb{E}[U_{(j)}] = \frac{j}{n+1}$. Der Abstand zweier Ordnungsstatistiken ist $U_{(j)} - U_{(j-1)}$ ist $\mathrm{Beta}(1, n)$-verteilt.

Literaturverzeichnis

Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., Acernese, F., Ackley, K., et al. (2016). Binary black hole mergers in the first advanced ligo observing run. *Phys. Rev. X*, 6:041015.

Dickhaus, T. (2014). *Simultaneous Statistical Inference with Applications in the Life Sciences*. Springer-Verlag, Berlin Heidelberg.

Elstrodt, J. (2005). *Maß- und Integrationstheorie*. Springer-Lehrbuch. Grundwissen Mathematik. Springer-Verlag, Berlin, 4. edition.

Fahrmeir, L., Kneib, T., and Lang, S. (2009). *Regression. Modelle, Methoden und Anwendungen*. Springer-Verlag Berlin Heidelberg.

Georgii, H.-O. (2007). *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik*. de Gruyter Lehrbuch. Walter de Gruyter & Co., Berlin, 3. edition.

Klenke, A. (2008). *Wahrscheinlichkeitstheorie*. Springer-Verlag Berlin Heidelberg, 2. edition.

Küchler, U. (2016). *Maßtheorie für Statistiker. Grundlagen der Stochastik*. Springer-Verlag Berlin Heidelberg.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, 2. edition.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Vershynin, R. (2018). *High-dimensional probability. An introduction with applications in data science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. With a foreword by Sara van de Geer.

Wainwright, M. J. (2019). *High-dimensional statistics. A non-asymptotic viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Sachverzeichnis

A

a-posteriori-Risiko 14
a-posteriori-Verteilung 14
a-priori-Verteilung 13
Ablehnbereich 21
absolutstetig 138
AIC 108
Akaike-Informationskriterium 108
Alternativhypothese 20
asymptotisch normalverteilt 8

B

balanciertes Design 87
Bayes-Formel 139
Bayes-Informationskriterium 116
Bayes-optimal 14
Bayes-Risiko 13, 15
bedingte Dichte 139
bedingte Wahrscheinlichkeit 132
bedingter Erwartungswert 139
Bestimmtheitsmaß 96
Bias-Varianz-Zerlegung 7
BIC 116
Binomialtest
  einseitiger 23, 36
  zweiseitiger 25, 39
BLUE 60
Borel-σ-Algebra 128
Borel-Mengen 128

C

charakteristische Funktion 136

D

Design 49
  zufällig 65
Design-Kovarianzmatrix 65
Designmatrix 53
Dichte 129, 138
dominiert 10
Dummy-Variablen 53

E

Effektdarstellung 92
empirische Risiko 67
empirischen Design-Kovarianzmatrix 68
empirischer Risiko-Minimierer 67
empirischer Risikominimierer
  penalisierter 118
empirisches Skalarprodukt 54, 123
endlich 128
erwartungstreu 6
Erwartungswert 135

F

F-Test 76, 82
Faktor 87
Fehler 1. und 2. Art 21
Fehler- oder Störgrößen 53
Fehlervektor 53
Fisher-Test 76
Fundamentalungleichung 118

G

Gütefunktion 22
Gauß-Test
  einseitiger 33, 37
  zweiseitiger 26, 39

H

Hypothesen 20

I

i.i.d. 134

K

Konfidenzintervall 40
Konfidenzmenge 40
konjugierte Verteilungsklassen 18
konsistent 8
Kontrastmatrix 80
Korrelation 136
Korrespondenzsatz 43
Kovarianz 136
Kovarianzmatrix 137
kritischer Bereich 21
kritischer Wert 25
Kullback-Leibler-Diskrepanz 106
Kullback-Leibler-Divergenz 103

L

Likelihood-Funktion 10
Likelihood-Quotiententest 38
lineares Modell 53
  gewöhnliches 54

M

Maß 128
  Dirac- 129
  induziert 131
  Lebesgue- 129
  Zähl- 129
Maßraum 128
Mallows' C_p-Kriterium 111
mathematische Stichprobe 4
messbar 130
messbarer Raum 127
minimax 13
MLE 11
Modellmissspezifikation 102
  Maximum-Likelihood-Schätzer unter 105
Momente 136
MSE 7

N

Neyman-Pearson-Lemma 35
Neyman-Pearson-Test 36
Nullhypothese 20

O

Orakelungleichung
  Modellwahl 118
orthogonales Design 54
overfitting 101

P

p-Wert 30
Parametervektor 53
penalisierte Modellwahl 118

Q

Quantil-Quantil-Plot 27
Quantile 132
Quantilfunktion 132

R

Radon-Nikodym-Dichte 138
Regression 53
  einfache 50
  multiple 52
  polynomiale 54
Regressionsgerade 50
Residuen 51, 60, 81
Responsevektor 53
Ridge-Regression 63
Risiko 7

S

σ-Algebra 127
σ-endlich 128
Satz
  Gauß-Markov 58
  Radon-Nikodym 138
  von Bayes 133
Schätzer 6
  Bayes- 14, 17
  bester linearer erwartungstreuer 60
  Kleinste-Quadrate- 55
  Maximum-Likelihood- 11
  Momenten- 9
  Shrinkage- 63
signifikant 22
Signifikanzniveau 22
Standardabweichung 136
Statistik 4
statistischer Test
  nichtrandomisierter 21
  randomisierter 21
statistisches Experiment 3

statistisches Modell 3
Stichprobenmittel 6
Stichprobenraum 3
stochastisch beschränkt 141
stochastische Ordnung 141
Student-t-Test 78
subgaußsch 68

T

t-Test 78
  Zweistichproben- 92
Test zum Niveau α 22
Testproblem 20
  einseitiges 23
  lineares 80
  zweiseitig 23
Teststatistik 24
Totalvariationsabstand 122

U

UMP-Tests 37
Unabhängigkeit
  von Ereignissen 133
  von Zufallsvariablen 133
underfitting 101
Ungleichung
  Cauchy-Schwarz 137
  Hölder 137
  Jensen 137
  Markov 137
  Tschebyscheff 137
unkorreliert 136
unverzerrt 6
unverzerrte Risikoschätzung 110

V

Varianz 136
Varianzanalyse
  einfaktorielle 87, 91
  zweifaktorielle 93
verallgemeinerte Inverse 131
Verlust 7
  0-1- 7
  absoluter 7
  quadratischer 7
Verteilung
  χ^2- 145
  Bernoulli- 142
  Beta- 146
  Binomial- 142
  einer Zufallsvariable 131
  F- 74
  Gamma- 146
  Gleich- 142, 144
  hypergeometrische 143
  inverse Gamma- 61
  Multinomial- 142
  Normal- bzw. Gauß- 144
  normal-inverse Gamma- 62
  Poisson- 144
  t- 74
Verteilungsfunktion 131
  empirische 27
Vorhersagefehler 66
Vorhersageverlust 66

W

Wahrscheinlichkeitsdichte 129
Wahrscheinlichkeitsmaß 128
Wahrscheinlichkeitsraum 128
Wahrscheinlichkeitsverteilung
  diskrete 129
  stetige 129
Wald-Intervall 42
Wilson-Intervall 42, 48

Z

Zähldichte 129
zufälligem Design 65
Zufallsvariablen 130