THORSTEN SCHMIDT

STOCHASTIK 2

UNIVERSITÄT FREIBURG

Einführung

In diesem Semester werden wir zum ersten Mal eine auf das Lehramt zugeschnittene Stochastik II hören. Dies ist ein durchaus gewagtes Experiment und Ihre Mitarbeit ist gefragt !

Die einzelnen Vorlesungen werden als 90-Minuten Module gehalten, welche einen Vorlesungsteil und einen Interaktionsteil enthalten. Einiges Material wird in das Skriptum ausgegliedert und ist im Selbststudium zu erarbeiten.

Ziel der Vorlesung ist die folgenden drei Bereiche kennenzulernen. Wenn Sie noch Wünsche haben, bitte gerne in der Vorlesung anbringen oder per Email an mich oder Ihren Tutor.

1. Statistik - eine (kurze) Einführung
2. Machine Learning - Moderne Anwendungen der Statistik
3. Nichtparametrische Statistik - Statistik ohne Modell

Was ist eigentlich Statistik?

Der deutsche Begriff „Statistik“ wurde 1749 von Gottfried Achenwall[1] eingeführt und meint Staatswissenschaft[2]

> „Der Begriff der Sogenannten Statistic, das ist, der Staatswissenschaft einzelner Reiche wird sehr verschiedentlich angegeben . . . “

Solche Statistiken sind mindestens seit dem 5.Jh. v. Chr. bekannt und Mittelwerte wurden sicher bereits in diesem Zusammenhang erhoben, wenn nicht eher.

Die erste *Sterbetafel*, und damit vielleicht der Beginn der demographischen Statistik wurde 1662 von John Graunt und William Petty veröffentlicht. Während des 19. Jahrhunderts wurde das Gebiet auf weitere Felder ausgeweitet unter Verwendung mathematischer Grundlagen mit Hilfe der Wahrscheinlichkeitstheorie. Hiermit kommen Pierre de Fermat, Blaise Pascal, Chrisitiaan Huygens auf den Plan und Jakob Bernoulli (1655 - 1705), dessen prägendes Werk „Ars Conjectandi“ posthum 1713 veröffentlicht wurde[3]. Dieses Werk wird oft als die Geburtsstunde der Stochastik bezeichnet. Die analytischen Methoden wie Maßtheorie und σ-Algebren gehen auf die

[1] 1719 - 1772, studiert in Leipzig, dann Marburg, Göttingen

[2] Abriß der neuesten Staatswissenschaft der vornehmsten Europäischen Reiche und Republicken. Göttingen, 1749

[3] Jakob Bernoulli. *Ars conjectandi: opus posthumum: accedit Tractatus de seriebus infinitis; et Epistola gallice scripta de ludo pilae reticularis*. Impensis Thurnisiorum, 1713

bahnbrechende Arbeit von Andrei Kolmogorov[4] zurück. Wesentliche Fotschritte in der Statistik wurden durch Sir Ronald Fisher[5] und die Einführung der Maximum-Likelihood Methode begründet.

Maschinelles Lernen (ML) ist ein Teilbereich von Künstlicher Intelligenz und bezeichnet Algorithmen, die mit mehr Daten bessere Ergebnisse erzielen - das ist ein Kernaspekt der Statistik. In diesem Sinne ist Statistik stets als ein Verfahren für Maschinelles Lernen zu sehen. Andererseits ist der Unterschied zu ML die meist stringentere Herangehensweise - Exakte Kenntnisse des zugrundeliegenenden Rahmens sind eher die Regel, im Gegensatz zu ML.

Dieses Skript ist *vorläufig* und enthält sicherlich noch viele Fehler. Über Hinweise per Email wäre ich sehr dankbar. Es basiert auf dem Lehrbuch [6] in welchem sich vertiefendes Material und Literaturhinweise finden.

Freiburg, im Sommer 2021 - Thorsten Schmidt

[4] Andrei Kolmogorov. Über die analytischen Methoden in der Wahrscheinlichkeitstheorie. *Math Ann*, 104(1):415–458, 1931

[5] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922

[6] C. Czado and T. Schmidt. *Mathematische Statistik*. Springer Verlag. Berlin Heidelberg New York, 2011

Eine kurzer Streifzug durch die Stochastik

Wir beginnen mit einer sehr kurzen Wiederholung - die vertieften Inhalte finden sich alle im Stochastik I Skript. Zunächst wiederholen wir die zentralen Konzepte: Zufallsvariable, Wahrscheinlichkeitsmaß und Wahrscheinlichkeitstraum.

Der zentrale Begriff *Wahrscheinlichkeit* weisst Mengen Wahrscheinlichkeiten zu, ist also eine Funktion von einem Mengensystem in das Intervall $[0,1]$. Das Mengensystem soll abzählbare Vereinigungen und Komplemente enthalten, was zu dem zentralen Konzept der σ-Algebra führt.

Definition 1. $(\Omega, \mathscr{A}, P)$ heißt *Wahrscheinlichkeitsraum*, falls gilt:

(i) $\mathscr{A}$ ist eine *σ-Algebra*, d. h.

$\Omega \in \mathscr{A}$
$A \in \mathscr{A} \Rightarrow A^c \in \mathscr{A}$
$A_1, A_2, \ldots \in \mathscr{A} \Rightarrow \bigcup_i A_i \in \mathscr{A}$

(ii) P ist ein *Wahrscheinlichkeitsmaß*, d. h.

$P(\Omega) = 1$
Sind $A_1, A_2, \ldots \in \mathscr{A}$ und paarweise disjunkt, so gilt $P\big(\sum_{i=1}^{\infty} A_i\big) = \sum_{i=1}^{\infty} P(A_i)$.

Die wesentlichen Eigenschaften eines Wahrscheinlichkeitsmaßes sind

- Normiertheit und
- σ-Additivität.

Auf einem Wahrscheinlichkeitsraum können wir nun den zentralen Begriff der Zufallsvariablen einführen: Eine Zufallsvariable ist eine Abbildung

$$X : \Omega \to \mathbb{R},$$

so dass $X^{-1}([a,b]) \in \mathscr{A}$ für alle $a \leq b \in \mathbb{R}$. Dieses Konzept lässt sich leicht auf den $\mathbb{R}^d$ erweitern (Wie?).

Eine Zufallsvariable ist eine *meßbare Abbildung*.

Eine Zufallsvariable X beschreibt man mit ihrer Verteilungsfunktion $F_X(x) = P(X \leq x)$. Ist die Zufallsvariable stetig, so ist ihre Dichte f_X die Ableitung der Verteilungsfunktion,

$$P(X \in A) = \int_A dP(\omega) = \int_A f_X(x)dx. \tag{2}$$

Der Erwartungswert ist ein linearer Operator, so dass $E[aX + bY] = aE[X] + bE[Y]$.

Hierauf basierend führt man den Erwartungswert ein,

$$E[X] = \int x f_X(x) dx. \tag{3}$$

Wie war noch einmal die Definition falls X diskret ist? Es ist etwas umständlich für diskrete und stetig Zufallsvariablen unterschiedliche Definition zu benutzen nicht? Mit der Technik der Maße erhält man einen einheitlichen Begriff (Gegenstand der Vorlesung Wahrscheinlichkeitstheorie).

Den Raum aller Zufallsvariabeln, also aller messbaren Abbildungen bezeichen wir mit L^0. Den Raum der p-integrierbaren Zufallsvariablen bezeichnen wir mit $L^p(P) = \{X \in L^0 : E[|X|^p] < \infty\}$.

Der Lesbegue-Raum $L^1(P)$ enthält die integrierbaren Zufallsvariablen.

Ein wichtiges Konzept waren bedingte Verteilungen, und die zentrale Definition hierfür ist

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{4}$$

vorausgesetzt, dass $P(B) > 0$. Arbeitet man mit stetigen Zuvallsvariabeln so erhält man eine ähnliche Formel für Dichten: $f(x|y) = f(x,y)f(y)^{-1}$, wieder unter $f(y) > 0$.

Ebenso wichtig war die Unabhängigkeit: Zwei Ereignisse A und B heißen unabhängig, falls $P(A \cap B) = P(A) \cdot P(B)$. Zwei Zufallsvariablen heißen unabhängig, falls $X^{-1}(A)$ und $Y^{-1}(B)$ unabhängig sind für alle Borel-Mengen A und B.

Die Borel σ-Algebra $\mathscr{B}(\mathbb{R})$ ist die σ-Algebra die von den Halbstrahlen $(-\infty, x]$ erzeugt wird. Sie wird aber auch von allen Intervallen erzeugt. Oft genügt es an Stelle von allen Borel-Mengen nur diejenigen eines Erzeugendensystems anzuschauen.

Erinnern wir uns an ein paar wichtige Verteilungen:

Beispiel 5 (Gleichverteilung)**.** Die Gleichverteilung modelliert eine Zufallsvariable auf einem Intervall $[a,b]$, für welcher jeder Wert in diesem Intervall gleich wahrscheinlich ist. Wir schreiben hierfür $U(a,b)$. Da die Dichte normiert sein muss, erhalten wir direkt (wieso?) $f(x) = \mathbb{1}_{[a,b]} \frac{1}{b-a}$ und als Verteilungsfunktion

$$F(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & x \in (a,b) \\ 1, & x \ge b. \end{cases}$$

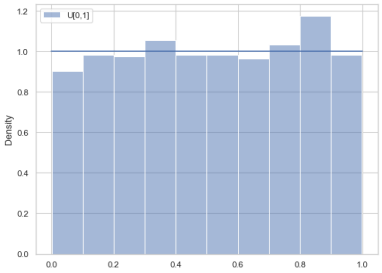


Abbildung 1: Gleichverteilung

Beispiel 6 (Exponentialverteilung)**.** Die Exponentialverteilung modelliert ein überraschendes Auftreten eines Ereignisses mit einer gleichbleibenden Eintrittswahrscheinlichkeit (Bsp: radioaktiver Zerfall). Die Dichte ist proportional zu $e^{-\lambda x}$, $\lambda > 9$ also $f(x) = \mathbb{1}_{\{x \ge 0\}} \lambda e^{-\lambda x}$. Die Verteilungsfunktion der Exponentialverteilung ist

$$F(x) = \begin{cases} 0, & x \le 0 \\ 1 - e^{-\lambda x}, & x > 0. \end{cases}$$

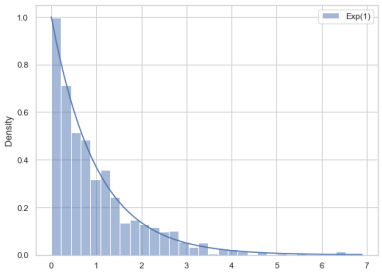


Abbildung 2: Exponential-Verteilung

Charakteristisch für die Exponentialverteilung ist, dass die Verteilung unabhängig von der Wartezeit ist: Zeigen Sie, dass

$$P(X > t + x | X > t) = P(X > x).$$

Beispiel 7 (Normalverteilung)**.** Die wohl wichtigste Verteilung ist die Normalverteilung. Dies begründet sich durch ihre herausragende Position im zentralen Grenzwertsatz. Für die Normalverteilung schreiben wir $\mathcal{N}(\mu, \sigma^2)$. Sei $\mu \in \mathbb{R}$, $\sigma > 0$. Die Dichte der Normalverteilung ist

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

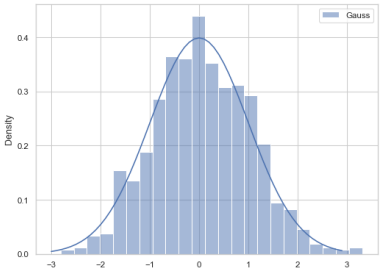


Abbildung 3: Exponential-Verteilung

Hieraus erhalten wir die Verteilungsfunktion:

$$\begin{aligned} F_{\mu,\sigma^2} &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-y)^2}{2\sigma^2}} \, dx \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx = \Phi\Big(\frac{x-\mu}{\sigma}\Big). \end{aligned}$$

Die Verteilung ist symmetrisch um den Ursprung. Überlegen Sie einmal, wie Sie eine symmetrische, positive Funktion erzeugen würden. Sicher landen Sie sehr schnell bei e^{-x^2}, was ja gerade die Basis für die Normalverteilung ist. Welche weiteren Funktionen finden Sie (und zu welchen Verteilungen gehören diese)?

Aufgabe 1 (Zufallsvariablen in Python oder R)**.** Erzeugen Sie die Histogramme und Dichten der obigen Beispiele (und gerne auch für weitere Verteilungen)[7].

[7] Thorsten Schmidt. Google Colab für Stochastik II. `https://colab.research.google.com/drive/1S6uhoYKY87f7XXhCXLSDag8_k0hFaJQA?authuser=1`, 2022

Aufgabe 2 (Erwartungswerte)**.** Rechnen Sie einmal einige Erwartungswerte aus: Den Erwartungswert einer Gleichverteilung, einer Exponentialverteilung, einer Normalverteilung. Probieren Sie auch einmal andere Beispiele. Welche Integrale können Sie noch ausrechnen und welche nicht mehr. Berechnen Sie auch den Erwartungwert einer Binomialverteilung und einer Poissonverteilung.

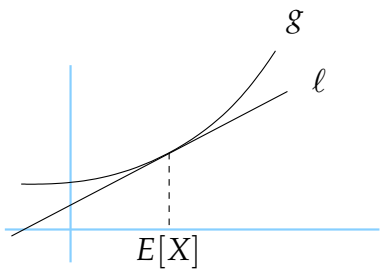
Neben dem Erwartungswert ist die Varianz

$$\mathrm{Var}(X) = E\big[(X - E[X])^2\big] \tag{8}$$

die wichtigste Kennzahl zur Beschreibung einer Verteilung. Hieraus ergibt sich die Standardabweichung $\sigma(X) = \sqrt{\mathrm{Var}(X)}$. Für die Varianz gilt die Regel $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$. Sind die Zufallsvariablen $X_1, \dots, X_n$ unabhängig (oder unkorreliert), so gilt

$$\mathrm{Var}(X_1 + \cdots + X_n) = Var(X_1) + \cdots + \mathrm{Var}(X_n) \tag{9}$$

Aufgabe 3 (Varianzen)**.** Berechnen Sie auch einige Varianzen von Verteilungen.



Ein wichtiges Hilfsmittel, vor allem für Abschätzungen und asymptotische Resultate sind Ungleichungen.

Satz 10 (Jensensche Ungleichung)**.** *Sei $X \in L^1(P)$ und g konvex. Dann ist*

$$E[g(X)] \geq g(E[X]).$$

Schlagen Sie diesen überraschend einfachen Beweis noch einmal in der Stochastik I nach - die Grafik nebenan verdeutlicht bereits die Intuition. Es ist durchaus erstaunlich in wie vielen Situationen die Jensensche Ungleichung als rettende Lösung genommen werden kann. Natürlich gibt es auch eine analoge Version für konkave Funktionen. Der nächste Satz ist die Basis für die Tschebyscheff- und die Markov-Ungleichung.

Satz 11 (Verallgemeinerte Tschebyscheff-Ungleichung)**.** *Sei die Funktion $g \geq 0$ und monoton wachsend. Dann gilt für jedes c mit $g(c) > 0$, dass*

$$P(X \geq c) \leq \frac{E[g(x)]}{g(c)}$$

Auch hier lohnt es sich den Beweis noch einmal anzuschauen. Versuchen Sie es einmal selbst! Wie weit kommen Sie ?

Beweisen Sie damit gleich die Markov-Ungleichung und die Tschebyscheff-Ungleichung: für $c > 0$ gilt

$$P(|X| > c) \leq \frac{E[|X|]}{c},$$
$$P(|X - E[X]| > c) \leq \frac{\mathrm{Var}(X)}{c^2}.$$

Grenzwertsätze

Zwei zentrale Aussagen haben wir in der Stochastik I bewiesen: Das starke Gesetz der großen Zahl und den zentralen Grenzwertsatz. Das starke Gesetz der großen Zahl besagt, dass der arithmetische Mittelwert fast sicher gegen den Erwartungswert konvergiert. Hierfür haben wir zunächst zwei Konvergenzen kennengelernt: Die stochastische Konvergenz,

$$Z_n \xrightarrow{P} Z, \quad \text{falls } P(|Z_n - Z| > \epsilon) \to 0 \quad \forall \epsilon > 0$$

und die stärkere fast sichere Konvergenz,

$$P(\{\omega \in \Omega : \lim Z_n(\omega) = z(\omega)\}) = 1.$$

Fast sichere konvergenz impliziert hierbei stochastische Konvergenz. Und umgekehrt, ist die stochastische Konvergenz schnell genug, und zwar: gilt $\sum_n P(|Z_n - Z| > \epsilon) < \infty$ für alle $\epsilon > 0$, so folgt auch fast sichere Konvergenz.

Aufgabe 4. Beweisen Sie das Schwache Gesetz der großen Zahl: Seien $X_1, X_2, \dots$ unabhängig und identisch verteilt mit $\mathrm{Var}(X_i) < \infty$. Dann gilt

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E[X_1].$$

Mit einigen trickreichen Argumenten waren wir in der Lage, das schwache Gesetz auf die folgende, stärkere Aussage zu erweitern:

Satz 12 (Starkes Gesetz der großen Zahl)**.** *Seien $X_1, X_2, \dots$ unabhängig und identisch verteilt mit* $\mathrm{Var}(X_i) < \infty$*. Dann gilt*

Das SGGZ gilt auch unter der schwächeren Annahme $E[X_1] < \infty$.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow E[X_1] \quad f.s.$$

Für den zentralen Grenzwertsatz müssen wir eine Aussage über die Verteilungen der Zufallsvariablen machen können. Zunächst hatten wir deswegen die schwache Konvergenz einer Folge von Verteilungen (μ_n) auf dem Raum der reellen Zahlen veresehen mit der Borel σ-Algebra $(\mathbb{R}, \mathscr{B})$ eingeführt: Wir sagen μ_n *konvergiert schwach* gegen μ, falls

$$\int f d\mu_n \to \int f d\mu,$$

für alle Funktionen $f : \mathbb{R} \to \mathbb{R}$, welche stetig und beschränkt sind.

Definition 13. Sind $X, X_1, X_2, \dots$ Zufallsvariablen, so sagen wir (X_n) konvergiert in *Verteilung* gegen X, falls F_{X_n} schwach gegen F_X konvergiert.

Wir schreiben $X_n \xrightarrow{\mathscr{L}} X$ für die Konvergenz in Verteilung.

Eine Kombination der beiden obigen Aussagen zeigt, dass dies äquivalent dazu ist, dass

$$E[f(X_n)] \to E[f(X)]$$

für alle $f \in C_b(\mathbb{R})$ (alle stetigen und beschränkten Funktionen auf $\mathbb{R}$).

$$S_n^* = \frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} = \frac{S_n - nE[X_1]}{\sqrt{n\,\mathrm{Var}(X_1)}}, \tag{14}$$

Satz 15. *Seien $(X_i)_{i\geq 1}$ i.i.d. mit $Var(X_i) < \infty$. Dann gilt, dass*

$$S_n^* \xrightarrow{\mathscr{L}} \mathscr{N}(0,1).$$

Wir hatten allgemeiner sogar die berühmten Lindeberg-Bedingungen betrachtet und diese starke Aussage bewiesen.

Aufgabe 5. Simulieren Sie einmal den ZGWS: Starten Sie mit gleichverteilten Zufallsvariablen und testen Sie wie schnell S_n^* gegen eine Normalverteilung konvergiert.

Eine kurzer Streifzug durch die Statistik

Viele Fragen, die uns in unserem Alltag begegnen lassen sich ohne Statistik nicht wissenschaftlich beantworten. Da gehören aktuell sogar eine ganze Menge dazu:

- Ist es gefährlich wenn ich mich mit Corona anstecke ?
- Bin ich durch eine Maske gut geschützt ?
- Ich habe gehört, wenn man sich impfen lässt kann man möglicherweise nicht mehr schwanger werden - stimmt das ?
- Ist es überhaupt sinnvoll sich impfen zu lassen ? Es landen doch auch viele geimpfte im Krankenhaus ?
- Oder etwas allgemeiner: Wie ist die Wirksamkeit von einem Arzneimittel bzw. einer Impfung? Was ist das zugehörige Risko?[8]
- Aktienkäufe sind doch sehr risikoreich - sollte ich mein Geld nicht lieber auf dem Sparbuch lassen ?
- Die Klimakrise ist doch gar nicht durch Menschen verursacht - es gab auch schon früher in der Geschichte der Erde Hitze (und auch Kälteperioden?)

[8] Zu den Nebenwirkungen der Corona Impfungen kursieren eine Reihe von unterschiedlichen Zahlen, siehe Beispiel 16.

Beispiel 16 (Nebenwirkung der Corona Impfung)**.** Es kursieren eine Vielzahl von unterschiedlichen Zahlen, siehe etwa [9]. Ein Grund hierfür ist, dass es ein Meldesystem gibt, wo jeder anonym Fälle melden kann, um eine Früherkennung von Impfnebenwirkungen zu ermöglichen. Hierbei muss kein ursächlicher Zusammenhang vorliegen - etwa durch einen Herzinfarkt in 14/ 30 / 42 Tagen nach der Impfung, denn dieser könnte ja auch so statt gefunden haben. Das Paul-Ehrlich Institut legt eine genaue Analyse dieser Zahlen in ihrem Sicherheitsbericht vor, [10]. Hierin werden die Nebenwirkung der ca 150 Mio Impfungen genauer untersucht. Die Meldungen für Herzinfarkte und Lungenembolie werden mit der durchschnittlich zu erwartenden Anzahl verglichen und liegen sogar darunter, so dass hier kein Risikosignal vermeldet wird. Wohl aber für zwei andere Nebenwirkung (in weniger als 1 von 10.000 bzw. 1 von 100.000 Fällen). Hieraus wurde z.B. abgeleitet, dass für Personen < 30 Jahre ein bestimmter Impfstoff empfohlen wird.

[9] Gabor Paal. Wenn nicht 5.000 Corona-Impftote, wie viele dann?, 2021

[10] Bericht über Verdachtsfälle von Nebenwirkungen und Impfkomplikationen nach Impfung zum Schutz vor COVID-19, 2021

Zwei Dinge sind in diesem Zusammenhang bemerkenswert: Die extrem hohe Zahl an Impfungen und die vergleichweise niedrigen Zahlen an Nebenwirkungen. Es wird geschätzt dass an Todesfällen knapp unter 100 Fälle (in Deutschland) den Impfungen zuzuordnen sind.

Wenn jemand auf der Autobahn auf der Strecke Freiburg - Hamburg ein 1m breites Schild aufstellt und ich während meiner Fahrt auf dieser Strecke blind aus dem Fenster schieße, ist die Wahrscheinlichkeit das Schild zu treffen 100:80.000.000 (und das sind etwas mehr als die Hälfte der 150 Mio Impfungen in Deutschland).

Auf die meisten dieser Fragen kann man ohne Statistik nur mit anekdotischer Evidenz antworten. Sogar im Beginn der Corona-Pandemie, wo nur wenige Daten vorlagen hatte man riesige Probleme mit mangelnden Daten. Mittlerweile können viele der obigen Fragen auf Basis von ausreichend großen Datenmengen statistisch beantwortet werden.

Wir beginnen mit einer Einführung in die Statistik um die grundlegenden Techniken genauer kennen zu lernen.

Statistik

Die Formulierung von statistischen Modellen bildet die Grundlage der Statistik. Hierbei werden Modelle ausgewählt, welche der Realität zum einen möglichst gut entsprechen sollen, zum anderen die für die statistische Analyse notwendige Handhabbarkeit besitzen. Das statistische Modell beschreibt stets das Ergebnis eines Zufallsexperiments, etwa die Werte einer erhaltenen *Stichprobe* oder gesammelte Messergebnisse eines Experiments. Somit ist die Verteilung der Zufallsvariable das Schlüsselelement. Das statistische Modell ist dann eine geeignete Familie von solchen Verteilungen. Anhand von zwei Beispielen wird im Folgenden die Formulierung von statistischen Modellen illustriert.

Beispiel 17 (Qualitätssicherung)**.** Eine Ladung von N Teilen soll auf ihre Qualität untersucht werden. Die Ladung enthält defekte und nicht defekte Teile. Mit θ sei der Anteil der defekten Teile bezeichnet, von insgesamt N Teilen sind $N\theta$ defekt. Aus Kostengründen wird nur eine Stichprobe von $n \leq N$ Teilen untersucht.

Wir wählen $\Omega = \{0, 1, \dots, n\}$ und $\mathscr{A} = \mathscr{P}(\Omega)$[11]. Die Zufallsvariable $X(\omega) = \omega$ bezeichne die Anzahl der defekten Teile in der Stichprobe. Erfolgt die Auswahl der Stichprobe zufällig, so kann man ein Laplacesches Modell rechtfertigen und erhält eine *hypergeometrische* Verteilung für X:

[11] Die Potenzmenge ist in der Tat eine σ-Algebra (ÜA)! Wir wählen eine Gleichverteilung auf Ω als Modell. Was heißt das noch einmal genau?

$$P(X = k) = \frac{\binom{N\theta}{k}\binom{N-N\theta}{n-k}}{\binom{N}{n}} \tag{18}$$

für $\max\{0, n - N(1-\theta)\} \leq k \leq \min\{N\theta, n\}$. Dies erhält man schlicht durch Abzählen (Günstige durch Mögliche). Wir schreiben kurz

$X \sim \text{Hypergeo}(N, n, \theta)$. Insgesamt kann man dieses Modell wie folgt zusammenfassen:

$$\{\big(\Omega, \mathscr{A}, \text{Hypergeo}(N, n, \theta)\big) : \theta \text{ unbekannt}\}.$$

Dies ist der erste Prototyp eines statistischen Modells. Wir haben immer das gleiche Ω aber jedes Mal ein anderes Wahrscheinlichkeitsmaß - das statistische Modell besteht also aus einer Familie von Wahrscheinlichkeitsräumen.

Der wesentliche Unterschied zu einem festen Wahrscheinlichkeitsraum besteht darin, dass das Wahrscheinlichkeitsmaß $P = P_\theta$ unbekannt ist. Man hat aber Informationen - wir betrachten nicht alle möglichen Wahrscheinlichkeitsmaße auf $(\Omega, \mathscr{A})$ sonder nur die hypergeometrischen. Das Wahrscheinlichkeitsmaß ist sozusagen bis auf den Parameter θ bekann. Ziel ist es natürlich θ möglichst clever zu schätzen.

Messfehler Oft hat man Beobachtungen von einem Modell, die einem Meßfehler unterliegen. Z.B. man misst eine physikalische Größe (Gewicht, etc.) bei einem Experiment. Das Messgerät hat hierbei einen Fehler - mal ist die Messung etwas zu groß, mal zu klein. Hierfür macht man oft die Annahme dass die Meßfehler symmetrisch um 0 verteilt sind.

Definition 19. Eine Zufallsvariable X heißt *symmetrisch um c verteilt*, falls $X - c$ und $-(X - c)$ die gleiche Verteilung besitzen. Dafür schreiben wir

$$X - c \overset{\mathscr{L}}{=} -(X - c). \tag{20}$$

Hat X die Verteilungsfunktion F und Dichte f, so ist (20) äquivalent zu $F(c + x) = 1 - F(c - x)$ für alle $x > 0$. Hieraus folgt, dass X symmetrisch um c genau dann ist, wenn

$$f(c + x) = f(c - x), \qquad x \geq 0.$$

Beispiele von symmetrischen Verteilungen: $\mathcal{N}(\mu, \sigma^2)$ symmetrisch μ und $\text{Bin}(n, \frac{1}{2})$ symmetrisch um $\frac{n}{2}$.

Ist X hingegen diskret mit der Wahrscheinlichkeitsfunktion p, so ist die Symmetrie von X um c äquivalent zu $p(c + x) = p(c - x)$ für alle $x \geq 0$.

Beispiel 21 (Meßmodell)**.** Es werden n Messungen einer physikalischen Konstante μ vorgenommen. Die Messergebnisse seien mit $X_1, \ldots, X_n$ bezeichnet. Man nimmt an, dass die Messungen einem Messfehler unterworfen sind, der *additiv* um μ variiert:

$$X_i \quad = \quad \mu + \varepsilon_i, \quad i = 1, \ldots, n.$$

Hierbei bezeichnet ε_i den Messfehler der i-ten Messung. Wir unterscheiden *typische Annahmen*, welche geringe, oft erfüllte Annahmen an physikalische Messungen beschreiben und *weitere Annahmen*, welche darüber hinaus die Berechnungen erleichtern. Bevor man allerdings die weiteren Annahmen verwendet, sollte man ihre Anwendbarkeit im konkreten Fall unbedingt einer kritischen Überprüfung unterziehen. **Typische Annahmen**:

(i) Die Verteilung von $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$ ist unabhängig von μ (kein systematischer Fehler).

(ii) Der Messfehler der i-ten Messung beeinflusst den Messfehler der j-ten Messung nicht, d.h. $\varepsilon_1, \dots, \varepsilon_n$ sind unabhängig.

(iii) Die Verteilung der einzelnen Messfehler ist gleich, d.h. $\varepsilon_1, \dots, \varepsilon_n$ sind identisch verteilt.

(iv) Die Verteilung von ε_i ist stetig und symmetrisch um 0.

Aus diesen Annahmen folgt, dass $X_i = \mu + \varepsilon_i$ gilt, wobei ε_i nach F und symmetrisch um 0 verteilt ist. Darüber hinaus besitzt X_i eine Dichte und F ist von μ unabhängig. **Weitere Annahmen**:

(v) $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

(vi) σ^2 ist bekannt.

Aus der Annahme (v) folgt, dass $X_i \sim \mathcal{N}(\mu, \sigma^2)$ und $X_1, \dots, X_n$ i.i.d. sind. Unter Annahme (vi) ist μ der einzige unbekannte Parameter, was die Handhabung des Modells wesentlich erleichtert. Bei einem konkreten Messdatensatz ist immer zu diskutieren, welche Annahmen realistisch für das Experiment sind.

Formulierung von statistischen Modellen

Das Ergebnis eines Zufallsexperiments ist eine so genannte *Stichprobe*. Darunter verstehen wir einen Zufallsvektor $\boldsymbol{X} = (X_1, \dots, X_n)^\top$. Falls man konkrete Daten $\boldsymbol{x} = (x_1, \dots, x_n)^\top$ beobachtet, so ist dies gleichbedeutend mit dem Ereignis $\{\boldsymbol{X} = \boldsymbol{x}\}$. Im Folgenden ist der Grundraum Ω wie auch die zugehörige σ-Algebra $\mathscr{A}$ fest.

Definition 22. Unter einem *statistischen Modell* verstehen wir ganz allgemein eine Familie $\mathscr{P}$ von Verteilungen. Für ein statistisches Modell $\mathscr{P}$ verwenden wir stets die Darstellung

$$\mathscr{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\},$$

wobei $P_{\boldsymbol{\theta}}$ für alle $\boldsymbol{\theta} \in \Theta$ ein Wahrscheinlichkeitsmaß ist. Θ heißt *Parameterraum*.

Es gibt zahlreiche Möglichkeiten ein Modell zu parametrisieren. Jede bijektive Funktion $g(\boldsymbol{\theta})$ eignet sich zur Parametrisierung. Es sollten jedoch Parametrisierungen gewählt werden, die eine Interpretation zulassen. Manchmal verlieren solche Parametrisierungen ihre Eindeutigkeit, in diesem Fall spricht man von der *Nichtidentifizierbarkeit* von Parametern.

Definition 23. Ein statistisches Modell $\mathscr{P}$ heißt *identifizierbar*, falls für alle $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ gilt, dass

$$\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \;\Rightarrow\; P_{\boldsymbol{\theta}_1} \neq P_{\boldsymbol{\theta}_2}.$$

Ist $\Theta \subset \mathbb{R}^k$, so spricht man von einem *parametrischen Modell*, ansonsten von einem *nichtparametrischen Modell*. Die Zustandsräume

$$\Theta_1 = \{F : F \text{ ist Verteilungsfunktion symmetrisch um } \mu\} \quad \text{und}$$
$$\Theta_2 = \{(\mu, p) : \mu \in \mathbb{R},\ p \text{ ist Dichte und symmetrisch um } 0\}$$

implizieren zum Beispiel nichtparametrische Modelle.

In dieser Vorlesung beschränken wir uns im Wesentlichen auf parametrische Modelle.

Definition 24. Ein statistisches Modell $\mathscr{P}$ heißt *regulär*, falls eine der folgenden Bedingungen erfüllt ist:

(i) Alle $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$, sind stetig mit Dichte $p_{\boldsymbol{\theta}}(x)$.

(ii) Alle $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$, sind diskret mit Wahrscheinlichkeitsfunktion $p_{\boldsymbol{\theta}}(x)$.

Im Folgenden schreiben wir für ein reguläres Modell oft

$$\mathscr{P} = \{p(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\},$$

wobei durch $p(x, \boldsymbol{\theta}) := p_{\boldsymbol{\theta}}(x)$ die entsprechende Dichte oder Wahrscheinlichkeitsfunktion gegeben ist.

Beispiel 25 (Meßmodell)**.** Reguläre Modelle erhält man etwa durch das Meßmodell aus Beispiel 21. Unter den Annahmen (i)-(iv) und der zusätzlichen Annahme, dass das Modell eine Dichte hat, ist die gemeinsame Dichte durch

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i - \mu)$$

gegeben, wobei f_θ eine von μ unabhängige und um 0 symmetrische Dichte ist. Gilt darüber hinaus die Normalverteilungsannahme (v), so erhält man mit $\boldsymbol{\theta} = (\mu, \sigma)^\top$, dass

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sigma} \phi\left(\frac{x_i - \mu}{\sigma}\right),$$

wobei $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ die Dichte der Standardnormalverteilung ist.

Das Ziel einer statistischen Analyse ist es aus den vorliegenden Daten zu schließen, welche Verteilung P_θ wirklich vorliegt, oder anders ausgedrückt: Welcher Parameter $\boldsymbol{\theta}$ den beobachtetenden Daten zugrunde liegt. Um die vorhandenen Daten bestmöglich auszunutzen, muss die statistische Untersuchung für das Problem speziell angepasst sein, weswegen eine statistische Fragestellung häufig von dem Problem selbst abhängt:

In dem Messmodel aus Beispiel 21 soll der unbekannte Parameter μ geschätzt werden. Ein möglicher Punktschätzer ist durch den arithmetischen Mittelwert gegeben:

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{26}$$

Wie man einen solchen Schätzer bestimmen kann und welche Optimalitätseigenschaften bestimmte Schätzer haben werden im Folgenden untersucht.

Suffizienz

Zur Motiviation betrachten wir den Maximum-Likelihood Schätzer für μ bei einer Normalverteilung,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Hier ist es so, dass der Schätzer nicht alle Informationen über die Daten enthält, sonder nur aggregierte Information: $\sum_{i=1}^{n} X_i$. Kann man besser schätzen wenn man etwa $X_1, \ldots, X_n$ noch zusätzlich betrachtet? Die Antwort hierauf ist nein, und diese Eigenschaft charakterisiert man mit der so genannten *Suffizienz*, die wir nun einführen.

Nach der Wahl des statistischen Modells möchte man irrelevante Informationen aus der Vielzahl der erhobenen Daten herausfiltern, welches zu einer Datenreduktion führt, etwa wie in Gleichung (26) durch den Mittelwert der Daten. Formal gesehen, sind die erhobenen Daten durch den Zufallsvektor $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ charakterisiert. Dies bedeutet, dass die erhobenen Datenwerte als Realisationen

von $\boldsymbol{X}$ angesehen werden. Unter einer *Statistik* versteht[12] man eine (meßbare) Funktion von der Daten, etwa dargestellt durch

$$T := T(\boldsymbol{X}).$$

[12] Eine Statistik $T(\boldsymbol{X})$ ist schlicht eine Funktion der Daten. Wieso sollte sie meßbar sein?

Gilt $T(\boldsymbol{x}_1) = T(\boldsymbol{x}_2)$ für alle Realisierungen $\boldsymbol{x}_1, \boldsymbol{x}_2$ mit gleichen Charakteristika[13] des Experiments, so reicht es aus nur den Wert der Statistik $T(\boldsymbol{x})$ und nicht den ganzen Datenvektor $\boldsymbol{x}$ zu kennen. Das heißt, im Vergleich zur Kenntnis von $\boldsymbol{X}$ geht für die Statistik T keine Information verloren.

[13] Geben Sie hierfür einmal ein Beispiel an.

Ein Schätzer $T(\boldsymbol{X})$ reduziert die in $\boldsymbol{X}$ enthaltene Information auf eine einzelne Größe. Möchte man einen Parameter schätzen, so ist es wesentlich zu wissen, ob durch diese Reduktion wichtige Information verloren geht oder nicht. Ist eine Statistik suffizient für den Parameter $\boldsymbol{\theta}$, so ist das nicht der Fall.

Definition 27. Eine Statistik $T(\boldsymbol{X})$ heißt *suffizient* für $\boldsymbol{\theta}$, falls die bedingte Verteilung von $\boldsymbol{X}$ gegeben $T(\boldsymbol{X}) = t$ nicht von $\boldsymbol{\theta}$ abhängt.

Suffizienz: Falls man den Wert der suffizienten Statistik T kennt, dann enthält $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ keine weiteren Informationen über $\boldsymbol{\theta}$.

Kurz schreiben wir für die Zufallsvariable $\boldsymbol{X}$ bedingt auf $T(\boldsymbol{X}) = t$

$$\boldsymbol{X} \mid T(\boldsymbol{X}) = t.$$

Beispiel 28 (Qualitätssicherung, siehe Beispiel 17)**.** Betrachtet wird die Zufallsvariable $\boldsymbol{X}$ gegeben durch $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$, wobei $X_i \in \{0, 1\}$ ist. X_i hat den Wert 1, falls das i-te Teil defekt ist und sonst 0. Wir nehmen an, dass die X_i unabhängig sind und $P_\theta(X_i = 0) = \theta$, wobei θ der unbekannte Parameter ist. Sei $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \{0, 1\}^n$ der Vektor der beobachteten Werte und $S(\boldsymbol{x}) := \sum_{i=1}^n x_i$. Das zugrundeliegende statistische Modell ist $\{P_\theta : \theta \in [0, 1]\}$ mit

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n) = \theta^{S(\boldsymbol{x})}(1 - \theta)^{n - S(\boldsymbol{x})}.$$

Für die bedingte Verteilung von $\boldsymbol{X}$ gegeben $S(\boldsymbol{X}) = \sum_{i=1}^n X_i$ erhält man

$$P(\boldsymbol{X} = \boldsymbol{x} \,|\, S(\boldsymbol{X}) = t) = \binom{n}{t}^{-1}.$$

Dieser Ausdruck ist unabhängig von θ, also ist $S(\boldsymbol{X})$ eine suffiziente Statistik für den Parameter θ. Damit ist auch der arithmetische Mittelwert $\bar{\boldsymbol{X}} = n^{-1} S(\boldsymbol{X})$ eine suffiziente Statistik für θ.

Suffizienz ist eigentlich recht schwierig nachzuweisen, wenn man es zu Fuß versucht. Mit dem folgenden Satz von Fisher, Neyman, Halmos und Savage kann man Suffizienz oft leichter zeigen. Für diesen Satz nehmen wir an, dass die Werte der Statistik T in Θ liegen.

Satz 29 (Faktorisierungssatz). *Sei $\mathscr{P} = \{p(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ ein reguläres Modell. Dann sind äquivalent:*

(i) $T(\boldsymbol{X})$ *ist suffizient für* $\boldsymbol{\theta}$.
(ii) Es existiert $g : \Theta \times \Theta \to \mathbb{R}$ *und* $h : \mathbb{R}^n \to \mathbb{R}$, *so dass für alle* $\boldsymbol{x} \in \mathbb{R}^n$ *und* $\boldsymbol{\theta} \in \Theta$

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = g(T(\boldsymbol{x}), \boldsymbol{\theta}) \cdot h(\boldsymbol{x}).$$

Beweis. Wir führen den Nachweis nur für den diskreten Fall. $\boldsymbol{X}$ nehme die Werte $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ an. Setze $t_i := T(\boldsymbol{x}_i)$. Dann ist $T = T(\boldsymbol{X})$ eine diskrete Zufallsvariable mit Werten $t_1, t_2, \dots$. Wir zeigen zunächst, dass $(ii) \Rightarrow (i)$. Dazu berechnen wir die bedingte Verteilung von $\boldsymbol{X}$ bedingt auf T. Für $\boldsymbol{\theta} \in \Theta$ mit $P_{\boldsymbol{\theta}}(T = t_i) > 0$ gilt

$$P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}_j | T = t_i) = \frac{P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}_j, T = t_i)}{P_{\boldsymbol{\theta}}(T = t_i)}.$$

Dieser Ausdruck ist 0 und damit unabhängig von $\boldsymbol{\theta}$, falls $T(\boldsymbol{x}_j) \neq t_i$. Gilt hingegen $T(\boldsymbol{x}_j) = t_i$, so ist

$$P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}_j | T = t_i) \ = \frac{g(t_i, \boldsymbol{\theta})\, h(\boldsymbol{x}_j)}{P_{\boldsymbol{\theta}}(T = t_i)}.$$

Aus (ii) folgt, dass

$$P_{\boldsymbol{\theta}}(T = t_i) = \sum_{\{\boldsymbol{x}: T(\boldsymbol{x}) = t_i\}} p(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{\{\boldsymbol{x}: T(\boldsymbol{x}) = t_i\}} g(t_i, \boldsymbol{\theta}) \cdot h(\boldsymbol{x}) \tag{30}$$

und damit

$$P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}_j | T = t_i) \ = \frac{g(t_i, \boldsymbol{\theta}) h(\boldsymbol{x}_j)}{\sum\limits_{\{\boldsymbol{x}: T(\boldsymbol{x}) = t_i\}} g(t_i, \boldsymbol{\theta}) \cdot h(\boldsymbol{x})} = \frac{h(\boldsymbol{x}_j)}{\sum\limits_{\{\boldsymbol{x}: T(\boldsymbol{x}) = t_i\}} h(\boldsymbol{x})}.$$

Da auch dieser Ausdruck unabhängig von $\boldsymbol{\theta}$ ist, ist $T(\boldsymbol{X})$ suffizient für $\boldsymbol{\theta}$.

Es bleibt zu zeigen, dass $(i) \Rightarrow (ii)$. Sei also T eine suffiziente Statistik für $\boldsymbol{\theta}$ und setze

$$g(t_i, \boldsymbol{\theta}) := P_{\boldsymbol{\theta}}(T(\boldsymbol{X}) = t_i) \ , \ h(\boldsymbol{x}) := P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = T(\boldsymbol{x})).$$

Dabei ist h unabhängig von $\boldsymbol{\theta}$, da $T(\boldsymbol{x})$ suffizient ist. Es folgt, dass

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{\theta}) &= P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = T(\boldsymbol{x})) \\ &= P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = T(\boldsymbol{x})) \cdot P_{\boldsymbol{\theta}}(T(\boldsymbol{X}) = T(\boldsymbol{x})) \\ &= h(\boldsymbol{x}) \cdot g(T(\boldsymbol{x}), \boldsymbol{\theta}) \end{aligned}$$

und somit die behauptete Faktorisierung in (ii). □

Beispiel 31 (Warteschlange)**.** Die Ankunft von Kunden an einem Schalter folgt einem *Poisson-Prozess* mit Intensität θ, falls folgende Annahmen erfüllt sind: Bezeichne N_t die zufällige Anzahl der Kunden, welche zum Zeitpunkt $t \geq 0$ angekommen sind.

(i) $N_0 = 0$,
(ii) $N_{t+h} - N_t$ ist unabhängig von N_s für $0 \leq s \leq t$ und $h > 0$,
(iii) $N_{t+h} - N_t \sim \text{Poiss}(\theta h)$ für alle $t \geq 0$ und $h > 0$.

Eine Illustration des Poisson-Prozesses $(N_t)_{t \geq 0}$ findet sich in Abbildung 4.

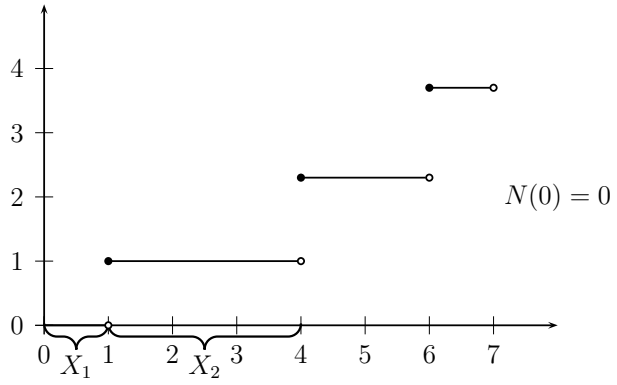


Abbildung 4: Realisation eines Poisson-Prozesses. Die Sprungzeitpunkte stellen Ankünfte von neuen Kunden an einer Warteschlange dar. X_i ist die verstrichene Zeit zwischen der Ankunft des i-ten und des $i-1$-ten Kunden.

Mit X_i sei die verstrichene Zeit zwischen der Ankunft des i-ten und des $i-1$-ten Kunden bezeichnet, X_1 sei die Zeit bis zur Ankunft des ersten Kunden. Dann folgt aus (iii), dass $P(X_1 > t) = P(N(t) = 0) = \exp(-\theta t)$, demzufolge ist X_1 exponentialverteilt mit dem Parameter θ. Zeigen Sie als Übungsaufgabe, dass $X_i \sim \text{Exp}(\theta)$ und $X_1, X_2, \ldots$ sind unabhängig.

Beispiel 32 (Warteschlange: Suffizienz)**.** Seien $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ die ersten n Zwischenankunftszeiten eines Poisson-Prozesses. Dann sind $X_1, \ldots, X_n$ unabhängig und $X_i \sim \text{Exp}(\theta)$. Die Dichte von $\boldsymbol{X}$ ist demnach

$$p(\boldsymbol{x}, \theta) = \mathbb{1}_{\{x_1, \ldots, x_n \geq 0\}} \, \theta^n \, e^{-\theta \sum_{i=1}^{n} x_i}.$$

Die Statistik $T(\boldsymbol{X}) := \sum_{i=1}^n X_i$ ist suffizient für θ: In der Tat, wähle $g(t, \theta) = \theta^n \exp\{-\theta t\}$ und $h(\boldsymbol{x}) = \mathbb{1}_{\{x_1, \ldots, x_n \geq 0\}}$. Dann ist die Bedingung (ii) von Satz 29 erfüllt und somit T suffizient für θ. Ebenso ist das arithmetische Mittel eine suffiziente Statistik für θ.

Beispiel 33 (Suffiziente Statistiken für die Normalverteilung)**.** Betrachtet man eine Stichprobe von normalverteilten Daten, so bilden das arithmetische Mittel *und* die Stichprobenvarianz zusammen einen

suffizienten Schätzer: Seien die Zufallsvariablen $X_1, \dots, X_n$ i.i.d. mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Gesucht ist der Parametervektor $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$, d.h. der Erwartungswert μ und die Varianz σ^2 sind unbekannt. Die Dichte von $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ ist

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right).$$

Zunächst betrachten wir $T_1(\boldsymbol{X}) := \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)^\top$. Mit $h(\boldsymbol{x}) := 1$ und

$$g(T_1(\boldsymbol{x}), \boldsymbol{\theta}) := \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{n\mu^2}{2\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i \right) \right)$$

ist $p(\boldsymbol{x}, \boldsymbol{\theta}) = g(T_1(\boldsymbol{x}), \boldsymbol{\theta}) h(\boldsymbol{x})$. Folglich ist $T_1(\boldsymbol{X})$ für $\boldsymbol{\theta}$ suffizient. Der zufällige Vektor T_2, definiert durch

$$T_2(\boldsymbol{X}) := \begin{pmatrix} \bar{X} \\ s^2(\boldsymbol{X}) \end{pmatrix}$$

ist ebenfalls suffizient, denn $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ und $s^2(\boldsymbol{X}) = \frac{1}{n-1} \sum_{i=1}^n (X_i^2 - (\bar{X})^2)$ (ÜA).

Exponentielle Familien

Wir bezeichnen mit $\mathbb{1}_{\{x \in A\}}$ die Indikatorfunktion mit Wert Eins falls $x \in A$ ist und Null sonst. Die folgende Definition führt exponentielle Familien für zunächst einen Parameter ein. K-parametrige exponentielle Familien werden in Definition 47 vorgestellt.

Definition 34. Eine Familie von Verteilungen $\{P_\theta : \theta \in \Theta\}$ mit $\Theta \subset \mathbb{R}$ heißt eine *einparametrige exponentielle Familie*, falls Funktionen $c, d : \Theta \to \mathbb{R}$ und $T, S : \mathbb{R}^n \to \mathbb{R}$ und eine Menge $A \subset \mathbb{R}^n$ existieren, so dass die Dichte oder Wahrscheinlichkeitsfunktion $p(\boldsymbol{x}, \theta),\ \boldsymbol{x} \in \mathbb{R}^n$ von P_θ durch

$$p(\boldsymbol{x}, \theta) = \mathbb{1}_{\{x \in A\}} \cdot \exp\Big(c(\theta) \cdot T(\boldsymbol{x}) + d(\theta) + S(\boldsymbol{x}) \Big) \tag{35}$$

dargestellt werden kann.

Es ist wesentlich, dass A hierbei unabhängig von θ ist. Die Funktion $d(\theta)$ kann als Normierung aufgefasst werden. An dieser Stelle soll betont werden, dass die Verteilung einer mehrdimensionalen Zufallsvariable durchaus zu einer *ein*parametrigen exponentiellen Familie gehören kann. Diese wird allerdings nur von einem eindimensionalen Parameter aufgespannt.

Die Nützlichkeit dieser Darstellung von Verteilungsklassen erschließt sich durch folgende Beobachtung: $T(\boldsymbol{X})$ ist stets suffiziente Statistik[14] für θ; dies folgt aus dem Faktorisierungssatz 29 mit

[14] In exponentiellen Familien ist $T(\boldsymbol{X})$ suffiziente Statistik.

$$g(t,\theta) = \exp\big(c(\theta)t + d(\theta)\big) \quad \text{und} \quad h(\boldsymbol{x}) = \mathbb{1}_{\{\boldsymbol{x}\in A\}} \cdot \exp(S(\boldsymbol{x})).$$

T heißt *natürliche suffiziente Statistik* oder *kanonische Statistik*.

Ist $c(\theta) = \theta$ in Darstellung (35), so spricht man von einer *natürlichen* exponentiellen Familie. Jede exponentielle Familie hat eine Darstellung als natürliche exponentielle Familie, was man stets durch eine Reparametrisierung erreichen kann: Mit $\eta := c(\theta)$ erhält man die Darstellung

$$p_0(\boldsymbol{x},\eta) = \mathbb{1}_{\{x\in A\}} \exp\Big(\eta \cdot T(\boldsymbol{x}) + d_0(\eta) + S(\boldsymbol{x})\Big). \tag{36}$$

Ist p_0 eine Dichte, so ist die zugehörige *Normierungskonstante* gegeben durch

$$d_0(\eta) := -\ln\left(\int_A \exp\big(\eta T(\boldsymbol{x}) + S(\boldsymbol{x})\big)d\boldsymbol{x}\right), \tag{37}$$

was äquivalent ist zu $\int p_0(\boldsymbol{x},\eta)d\boldsymbol{x} = 1$. Ist p_0 hingegen eine Wahrscheinlichkeitsfunktion und nimmt $\boldsymbol{X}$ die Werte $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ an, so gilt

$$d_0(\eta) := -\ln\left(\sum_{\boldsymbol{x}_i\in A} \exp\big(\eta T(\boldsymbol{x}_i) + S(\boldsymbol{x}_i)\big)\right). \tag{38}$$

Bemerkung 39. Ist $c : \Theta \to \mathbb{R}$ eine injektive Funktion, so ist die Normierungskonstante einfacher zu bestimmen, denn in diesem Fall folgt $d_0(\eta) = d(c^{-1}(\eta))$. Gilt weiterhin, dass $\eta = c(\theta)$ für ein $\theta \in \Theta$, so folgt $d_0(\eta) = d(\theta) < \infty$, da $p(\cdot,\theta)$ eine Dichte bzw. eine Wahrscheinlichkeitsfunktion ist.

Beispiel 40 (Normalverteilung mit bekanntem σ)**.** Ausgehend von dem Meßmodell aus Beispiel 21 und den dortigen Annahmen (i)-(vi) betrachten wir ein festes σ_0^2 und das statistische Modell

$$\mathscr{P} = \big\{P_\mu = \mathcal{N}(\mu,\sigma_0^2) : \mu \in \mathbb{R}\big\}.$$

Dann ist $\mathscr{P}$ eine einparametrige exponentielle Familie, denn die zu P_μ zugehörige Dichte lässt sich schreiben als

$$\begin{aligned} p(x,\mu) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(x-\mu)^2\right) \\ &= \exp\left(\frac{\mu}{\sigma_0^2}\cdot x + \frac{-\mu^2}{2\sigma_0^2} - \left(\frac{x^2}{2\sigma_0^2} + \ln\left(\sqrt{2\pi\sigma_0^2}\right)\right)\right). \end{aligned} \tag{41}$$

Mit $c(\mu) := \frac{\mu}{\sigma_0^2}$, $T(x) := x$, $d(\mu) := \frac{-\mu^2}{2\sigma_0^2}$ und $S(x) := -\Big(\frac{x^2}{2\sigma_0^2} + \ln\Big(\sqrt{2\pi\sigma_0^2}\Big)\Big)$ sowie $A := \mathbb{R}$ erhält man die Gestalt (35).

Beispiel 42 (Die $U(0,\theta)$-Verteilung ist keine exponentielle Familie)**.** Als wichtiges Gegenbeispiel für Verteilungen, welche nicht als exponentielle Familie darstellbar sind, betrachte man eine Gleichverteilung auf dem Intervall $(0,\theta)$. Die zugehörige Dichte ist

$$\mathbb{1}_{\{x\in(0,\theta)\}}\,\frac{1}{\theta}$$

und somit handelt es sich nicht um eine exponentielle Familie, da die Menge A in der Darstellung (35) von θ abhängen müsste.

Es sei daran erinnert, dass unabhängige und identisch verteilte Zufallsvariablen als i.i.d. bezeichnet werden.

Zeigen Sie als Übung, dass die i.i.d. Kombination einer exponentiellen Familie wieder eine exponentielle Familie ist.

Beispiel 43 (i.i.d. Normalverteilung mit bekanntem σ)**.** Als Beispiel hierzu betrachten wir $X_1,\dots,X_n$ i.i.d. seien mit $X_i \sim \mathcal{N}(\mu,\sigma_0^2)$ und bekanntem σ_0. Beispiel 40). Dann ist $T(\boldsymbol{X}) := \sum_{i=1}^n X_i$ und somit auch das arithmetische Mittel $\bar{X}$ suffiziente Statistik für μ und die Verteilung von $\boldsymbol{X}$ ist eine einparametrige exponentielle Familie.

Das folgende Resultat beschreibt die Verteilung der natürlichen suffizienten Statistik einer einparametrigen exponentiellen Familie.

Satz 44. *Sei $\{P_\theta : \theta\in\Theta\}$ eine einparametrige exponentielle Familie mit Darstellung (35) und sei T stetig. Hat $\boldsymbol{X}$ die Verteilung P_θ, so hat $T(\boldsymbol{X})$ die Verteilung Q_θ, wobei Q_θ wieder eine einparametrige exponentielle Familie ist mit der Dichte bzw. Wahrscheinlichkeitsfunktion*

$$q(t,\theta) = \mathbb{1}_{\{t\in A^*\}}\exp\Big(c(\theta)\cdot t + d(\theta) + S^*(t)\Big);$$

hierbei ist $A^* := \{T(\boldsymbol{x}) : \boldsymbol{x}\in A\}$. Handelt es sich um eine diskrete Verteilung, so ist

$$S^*(t) = \ln\left(\sum_{\boldsymbol{x}\in A:\, T(\boldsymbol{x})=t}\exp(S(\boldsymbol{x}))\right).$$

Beweis. Wir beweisen den diskreten Fall. Ist $\boldsymbol{X}$ eine diskrete Zufallsvariable mit der Wahrscheinlichkeitsfunktion $p(\boldsymbol{x},\theta)$, so ist $T(\boldsymbol{X})$ ebenfalls eine diskrete Zufallsvariable und besitzt die Wahrscheinlich-

keitsfunktion

$$\begin{aligned} q(t,\theta) := P_\theta(T(\boldsymbol{X}) = t) \; &= \sum_{\boldsymbol{x}\in A\colon T(\boldsymbol{x})=t} p(\boldsymbol{x},\theta) \\ &= \sum_{\boldsymbol{x}\in A\colon T(\boldsymbol{x})=t} \exp\Big(c(\theta)\cdot T(\boldsymbol{x}) + d(\theta) + S(\boldsymbol{x})\Big) \\ &= \mathbb{1}_{A^*}(t)\cdot \exp\big(c(\theta)t + d(\theta)\big)\left(\sum_{\boldsymbol{x}\in A\colon T(\boldsymbol{x})=t} e^{S(\boldsymbol{x})}\right). \end{aligned}$$

Damit ist die Verteilung von T eine exponentielle Familie nach Darstellung (35). □

Wir interessieren uns natürlich für die Verteilung des Schätzers $T(\boldsymbol{X})$. Die kann man in exponentiellen Familien gut berechnen.

Satz 45. *Betrachtet man eine natürliche einparametrige exponentielle Familie mit den Dichten oder Wahrscheinlichkeitsfunktionen $p_0(\boldsymbol{x},\eta)$: $\eta\in\Theta'$ in Darstellung (36) und ist $\boldsymbol{X}\sim p_0$, so gilt*

$$\Psi(s) = E\big[e^{s\cdot T(\boldsymbol{X})}\big] = \exp\big(d_0(\eta) - d_0(\eta+s)\big) < \infty$$

für alle $\eta, \eta+s \in H$ mit $H := \{\eta\in\Theta' : d_0(\eta) < \infty\}$.

Die Verteilung einer Zufallsvariablen kann man in vielen Fällen durch die *momentenerzeugende Funktion*

$$\Psi(s) = E\big[e^{s\cdot T(\boldsymbol{X})}\big], \quad s\in S$$

bestimmen; $S = \{s : \Psi(s) < \infty\}$.

Beweis. Wir führen den Beweis für den Fall, dass p_0 eine Dichte ist. Der diskrete Fall folgt analog. Mit Darstellung (36) erhalten wir

$$\begin{aligned} \Psi(s) &= E\big[\exp(s\cdot T(\boldsymbol{X}))\big] \\ &= \int_A \exp\Big((\eta+s)T(\boldsymbol{x}) + d_0(\eta) + S(\boldsymbol{x})\Big)d\boldsymbol{x} \\ &= \exp\Big(d_0(\eta) - d_0(\eta+s)\Big)\int_A \exp\Big((\eta+s)T(\boldsymbol{x}) + d_0(\eta+s) + S(\boldsymbol{x})\Big)d\boldsymbol{x} \\ &= \exp\Big(d_0(\eta) - d_0(\eta+s)\Big)\int_A p_0(\boldsymbol{x},\eta+s)\,d\boldsymbol{x}. \end{aligned}$$

Nach Voraussetzung ist $\eta+s\in H$, und somit ist $p_0(\cdot,\eta+s)$ eine Dichte und das Integral in der letzten Zeile gleich 1. Weiterhin folgt aus $\eta, \eta+s \in H$, dass $d_0(\eta) - d_0(\eta+s)$ endlich ist und somit $\Psi(s) < \infty$. □

Bemerkung 46. *Erwartungswert und Varianz der suffizienten Statistik in exponentiellen Familien.* Aus der momentenerzeugenden Funktion Ψ kann man folgendermaßen die Momente von $T(\boldsymbol{X})$ bestimmen. Es sei daran erinnert, dass jede exponentielle Familie eine natürliche Darstellung der Form (36) hat. Unter dieser Darstellung ist

$$E[T(\boldsymbol{X})] = \Psi'(0) = \Psi(0)\Big(-d_0'(\eta+s)\Big|_{s=0}\Big) = -d_0'(\eta),$$

da $\Psi(0) = 1$. Weiterhin ist $E[T(\boldsymbol{X})^2] = (d_0'(\eta))^2 - d_0''(\eta)$ und damit

$$\mathrm{Var}(T(\boldsymbol{X})) = -d_0''(\eta).$$

Die Funktion d_0 kann durch (37) bzw. (38) oder mit Hilfe von Bemerkung 39 bestimmt werden. Zusammenfassend erhalten wir:

$$E[T(\boldsymbol{X})] = -d_0'(\eta),$$
$$\mathrm{Var}(T(\boldsymbol{X})) = -d_0''(\eta).$$

Definition 47. Eine Familie von Verteilungen $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ mit $\Theta \subset \mathbb{R}^K$ heißt *K-parametrige exponentielle Familie*, falls Funktionen $c_i, d : \Theta \to \mathbb{R}$, $T_i : \mathbb{R}^n \to \mathbb{R}$ und $S : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \dots, K$ sowie eine Menge $A \subset \mathbb{R}^n$ existieren, so dass die Dichte oder Wahrscheinlichkeitsfunktion $p(\boldsymbol{x}, \boldsymbol{\theta})$ von P_θ für alle $\boldsymbol{x} \in \mathbb{R}^n$ als

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \mathbb{1}_{\{\boldsymbol{x} \in A\}} \exp\left(\sum_{i=1}^{K} c_i(\boldsymbol{\theta}) T_i(\boldsymbol{x}) + d(\boldsymbol{\theta}) + S(\boldsymbol{x}) \right) \tag{48}$$

dargestellt werden kann.

In Analogie zu den einparametrigen Familien ist die Statistik

$$\boldsymbol{T}(\boldsymbol{X}) := \big(T_1(\boldsymbol{X}), \dots, T_K(\boldsymbol{X})\big)^\top$$

suffizient, sie wird als *natürliche suffiziente Statistik* bezeichnet.

Beispiel 49 (Die Normalverteilung ist eine zweiparametrige exponentielle Familie)**.** Die Familie der (eindimensionalen) Normalverteilungen gegeben durch $P_{\boldsymbol{\theta}} = \mathcal{N}(\mu, \sigma^2)$ mit $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ und $\Theta = \{(\mu, \sigma^2)^\top : \mu \in \mathbb{R}, \sigma > 0\}$ ist eine zweiparametrige exponentielle Familie, denn ihre Dichten haben die Gestalt

$$p(x, \boldsymbol{\theta}) = \exp\left(\frac{\mu}{\sigma^2} x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\Big(\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2) \Big) \right).$$

Durch die Wahl von $n = 1$, $c_1(\boldsymbol{\theta}) := \mu/\sigma^2$, $T_1(x) := x$, $c_2(\boldsymbol{\theta}) := -1/2\sigma^2$, $T_2(x) := x^2$, $S(\boldsymbol{x}) := 0$, $A = \mathbb{R}$ und der entsprechenden Normierung $d(\boldsymbol{\theta}) := -1/2(\mu^2\sigma^{-2} + \ln(2\pi\sigma^2))$ erhält man die Darstellung (48).

Beispiel 50 (i.i.d. Normalverteilung als exponentielle Familie)**.** Seien $X_1, \dots, X_n$ i.i.d. und weiterhin $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Dann ist die Verteilung

von $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ darstellbar als zweiparametrige exponentielle Familie: Weiterhin ist

$$T(\boldsymbol{X}) = \left(\sum_{i=1}^n T_1(X_i), \sum_{i=1}^n T_2(X_i) \right)^\top = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)^\top$$

suffizient für $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. Dies wurde in Beispiel 33 bereits auf elementarem Weg gezeigt.

Beispiel 51 (Lineare Regression)**.** Bei der linearen Regression beobachtet man Paare von Daten welche wir mit $(x_1, Y_1), \dots, (x_n, Y_n)$ bezeichnen. Man vermutet einen *linearen* Einfluss der Größen x_i auf Y_i und möchte diesen bestimmen. Die Beobachtungen $x_1, \dots, x_n$ werden als konstant angesehen. Diese Methodik wird in Kapitel 7 wesentlich vertieft und an Beispielen erprobt. Wir gehen von folgendem Modell aus:

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

für $i = 1, \dots, n$. Hierbei sind $\beta_1, \beta_2 \in \mathbb{R}$ unbekannte Konstanten und $\epsilon_1, \dots, \epsilon_n$ i.i.d. mit $\epsilon_1 \sim \mathcal{N}(0, \sigma^2)$ (vergleiche mit dem Meßmodell, Beispiel 21). Setze $\boldsymbol{Y} := (Y_1, \dots, Y_n)^\top$ und $\boldsymbol{\theta} := (\beta_1, \beta_2, \sigma^2)^\top$. Die Dichte von $\boldsymbol{Y}$ ist

$$\begin{aligned}
p(\boldsymbol{y}, \boldsymbol{\theta}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^n \exp\left(- \frac{(y_i - \beta_1 - \beta_2 x_i)^2}{2\sigma^2} \right) \\
&= \exp\Bigg(- \frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2 - \frac{n\beta_1^2}{2\sigma^2} - \frac{\beta_2^2}{2\sigma^2} \sum_{i=1}^n x_i^2 \\
&\qquad + \frac{\beta_1}{\sigma^2} \sum_{i=1}^n y_i + \frac{\beta_2}{\sigma^2} \sum_{i=1}^n x_i y_i - \frac{\beta_1 \beta_2}{\sigma^2} \sum_{i=1}^n x_i - \frac{n}{2} \ln(2\pi\sigma^2) \Bigg) \\
&= \exp\Bigg(- \frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2 + \frac{\beta_1}{\sigma^2} \sum_{i=1}^n y_i + \frac{\beta_2}{\sigma^2} \sum_{i=1}^n x_i y_i \\
&\qquad - \frac{n\beta_1^2}{2\sigma^2} - \frac{\beta_2^2}{2\sigma^2} \sum_{i=1}^n x_i^2 - \frac{\beta_1 \beta_2}{\sigma^2} \sum_{i=1}^n x_i - \frac{n}{2} \ln(2\pi\sigma^2) \Bigg).
\end{aligned}$$

Dies ist eine dreiparametrige exponentielle Familie. In der Tat, setzt man $T_1(\boldsymbol{y}) := \sum_{i=1}^n y_i$, $T_2(\boldsymbol{y}) := \sum_{i=1}^n y_i^2$, $T_3(\boldsymbol{y}) := \sum_{i=1}^n x_i y_i$ sowie $c_1(\boldsymbol{\theta}) := \beta_1/\sigma^2$, $c_2(\boldsymbol{\theta}) := -(2\sigma^2)^{-1}$, $c_3(\boldsymbol{\theta}) := \beta_2/\sigma^2$, so erhält man, mit entsprechender Wahl von d und $S \equiv 0$, $A := \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$ eine Darstellung der Form (48). Damit ist die Statistik

$$T(\boldsymbol{Y}) := \left(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2, \sum_{i=1}^n x_i Y_i \right)^\top$$

suffizient für $\boldsymbol{\theta} = (\beta_1, \beta_2, \sigma^2)^\top$.

Schätzmethoden

In diesem Kapitel gehen wir den wichtigen Schritt von der Theorie zur Praxis: Nun kennen wir den Parameter θ nicht und müssen ihn schätzen. Dazu haben wir folgendes Set-up zur Hand:

- Ein statistisches Modell $\mathscr{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$
- eine Beobachtung $\{\boldsymbol{X} = \boldsymbol{x}\}$ betrachtet
- wir möchten $\boldsymbol{\theta}$ oder etwas allgemeiner $q(\boldsymbol{\theta})$ für eine fest vorgegebene Funktion $q : \Theta \to \mathbb{R}$ schätzen.

In diesem Kapitel stellen wir zwei Methoden für die Auswahl vernünftiger Schätzer vor, das Maximum Likelihood Prinzip und das Kleinste Quadrate Verfahren. Für weitere Methoden sei u.a. auf [15] verwiesen.

Um $q(\boldsymbol{\theta})$ zu schätzen, wählt man eine Statistik T und wertet sie an den beobachteten Datenpunkten $\boldsymbol{x} = (x_1, \dots, x_n)^\top$ aus. Falls der wahre, unbekannte Wert für $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ ist, schätzt man die unbekannte Größe $q(\boldsymbol{\theta}_0)$ durch die bekannte Größe $T(\boldsymbol{x})$, den *Schätzwert*. Oft verwenden wir auch die Notation $T(\boldsymbol{X})$ für den zufälligen *Schätzer* ohne uns auf die beobachteten Daten $\boldsymbol{x}$ festzulegen.

[15] C. Czado and T. Schmidt. *Mathematische Statistik*. Springer Verlag. Berlin Heidelberg New York, 2011

Maximum-Likelihood-Schätzung

Die wichtigste und flexibelste Methode zur Bestimmung von Schätzern ist die Maximum-Likelihood-Methode. Wir betrachten ein reguläres statistisches Modell $\mathscr{P}$ gegeben durch eine Familie von Dichten oder Wahrscheinlichkeitsfunktionen $\left\{p(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\right\}$ mit $\Theta \subset \mathbb{R}^k$.

Die Funktion $L : \Theta \times \mathbb{R}^n \to \mathbb{R}^+$, gegeben durch

$$L(\boldsymbol{\theta}, \boldsymbol{x}) := p(\boldsymbol{x}, \boldsymbol{\theta})$$

mit $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{x} \in \mathbb{R}^n$ heißt *Likelihood-Funktion* des Parameters $\boldsymbol{\theta}$ für die Beobachtung $\boldsymbol{x}$.

Sind $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(\mu, 1)$, so ist

$$L(\mu, \boldsymbol{x}) = C \prod_{i=1}^{n} e^{-\frac{(x_i - \mu)^2}{2}}$$

mit einer von μ unabhängigen Konstante C.

Falls $\boldsymbol{X}$ eine diskrete Zufallsvariable ist, dann gibt $L(\boldsymbol{\theta}, \boldsymbol{x})$ die Wahrscheinlichkeit an, die Beobachtung $\{\boldsymbol{X} = \boldsymbol{x}\}$ unter dem Parameter $\boldsymbol{\theta}$ zu erhalten. Aus diesem Grund kann man $L(\boldsymbol{\theta}, \boldsymbol{x})$ als Maß dafür interpretieren, wie wahrscheinlich (likely) der Parameter $\boldsymbol{\theta}$ ist, falls $\boldsymbol{x}$ beobachtet wird. Im stetigen Fall kann diese Interpretation

ebenfalls erlangt werden, indem man das Ereignis $\{\boldsymbol{X}$ liegt in einer ϵ-Umgebung von $\boldsymbol{x}\}$ betrachtet und ϵ gegen Null gehen lässt.

Die *Maximum-Likelihood-Methode* besteht darin, den Schätzwert $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ zu finden, unter dem die beobachteten Daten die höchste Wahrscheinlichkeit erlangen.

Definition 52. Gibt es in dem regulären statistischen Modell $\mathscr{P}$ eine meßbare Funktion $\widehat{\boldsymbol{\theta}} : \mathbb{R}^n \mapsto \Theta$, so dass

$$L(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{x}) \quad = \quad \max\left\{L(\boldsymbol{\theta}, \boldsymbol{x}) : \boldsymbol{\theta} \in \Theta\right\} \qquad \text{für alle } \boldsymbol{x} \in \mathbb{R}^n,$$

so heißt $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ *Maximum-Likelihood-Schätzer* (MLS) von $\boldsymbol{\theta}$.

Falls der MLS $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ existiert, dann schätzen wir $q(\boldsymbol{\theta})$ durch $q(\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))$. In diesem Fall heißt

$$q(\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))$$

der *Maximum-Likelihood-Schätzer* von $q(\boldsymbol{\theta})$. Dieser wird auch als MLE oder Maximum-Likelihood-Estimate von $q(\boldsymbol{\theta})$ bezeichnet.

Ist die Likelihood-Funktion differenzierbar in $\boldsymbol{\theta}$, so sind *mögliche* Kandidaten für den Maximum-Likelihood-Schätzwert durch die Bedingung

$$\frac{\partial}{\partial \theta_i} L(\boldsymbol{\theta}, \boldsymbol{x}) = 0, \quad i = 1, \dots, k$$

gegeben. Darüber hinaus ist die zweite Ableitung zu überprüfen, um festzustellen, ob es sich tatsächlich um ein Maximum handelt. Weitere Maxima könnten auch auf dem Rand des Parameterraums angenommen werden.

Für die praktische Anwendung ist es äußerst nützlich den Logarithmus der Likelihood-Funktion zu betrachten. Da der Logarithmus eine streng monoton wachsende Funktion ist, bleibt die Maximalität unter dieser Transformation erhalten.

Die *Log-Likelihood-Funktion* $l : \Theta \times \mathbb{R}^n \to \mathbb{R}$ ist definiert durch

$$l(\boldsymbol{\theta}, \boldsymbol{x}) := \ln L(\boldsymbol{\theta}, \boldsymbol{x}).$$

In der Tat ist natürlich

$$l(\mu, \boldsymbol{x}) = C' - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2}$$

viel leichter abzuleiten. (Bestimmen Sie das Maximum von l).

Falls Θ offen, l differenzierbar in $\boldsymbol{\theta}$ für festes $\boldsymbol{x}$ und $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ existiert, so muß der Maximum-Likelihood-Schätzerwert $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ die *Log-Likelihood-Gleichung* erfüllen:

$$\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{x})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(\boldsymbol{x})} = \mathbf{0}. \tag{53}$$

Des Weiteren sind hinreichende Bedingungen, etwa an die zweite Ableitung, zu überprüfen um zu verifizieren, dass $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ auch tatsächlich eine Maximalstelle ist.

Bemerkung 54. *Konkavität der Likelihood-Funktion.* Nicht immer muss man die zweite Ableitung bemühen, um Maximalität zu zeigen: Falls L konkav ist, so ist eine Lösung von $\frac{\partial}{\partial\theta}L(\theta,\boldsymbol{x}) = 0$ für $\theta \in \mathbb{R}$ stets Maximum-Likelihood-Schätzwert für θ. Gleiches gilt ebenso für l. In Abbildung 5 wird dies an einer konkaven Funktion illustriert. Hierbei ist eine Funktion $f : \mathbb{R} \to \mathbb{R}$ konkav, falls $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y)$ für alle $\lambda \in (0,1)$. Angewendet etwa auf die Log-Likelihood-Funktion l heißt das: Ist l zweimal differenzierbar in θ, so ist l konkav in θ genau dann, wenn $\frac{\partial^2}{\partial\theta^2}l(\theta,\boldsymbol{x}) \leq 0$.

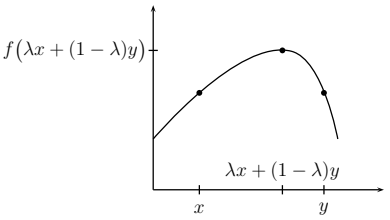


Abbildung 5: Ist die Funktion L konkav, so ist das Verschwinden der ersten Ableitung auch hinreichend für ein Maximum von L.

Beispiel 55 (Log-Likelihood-Funktion unter Unabhängigkeit)**.** Sind die $X_1, \ldots, X_n$ unabhängig und hat X_i die Dichte oder Wahrscheinlichkeitsfunktion $p_i(\cdot, \boldsymbol{\theta})$, so ist die Log-Likelihood-Funktion gegeben durch

$$l(\boldsymbol{\theta}, \boldsymbol{x}) = \ln\left(\prod_{i=1}^{n} p_i(x_i, \boldsymbol{\theta})\right) = \sum_{i=1}^{n} \ln p_i(x_i, \boldsymbol{\theta}).$$

Maximum-Likelihood in eindimensionalen Modellen

In diesem Abschnitt nehmen wir an, dass $\theta \in \mathbb{R}$ ein eindimensionaler Parameter ist. Wir beginnen mit zwei Beispielen.

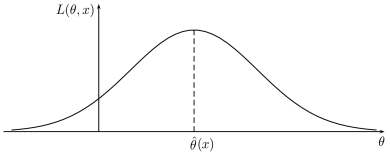


Abbildung 6: Die Likelihood-Funktion L als Funktion von θ. Der Maximum-Likelihood-Schätzwert $\widehat{\theta}(x)$ maximiert die Likelihood-Funktion $L(\theta, x)$ für ein festes x.

Beispiel 56 (Normalverteilungsfall, σ bekannt)**.** (Siehe Beispiel 40). Sei X normalverteilt, $X \sim \mathcal{N}(\theta, \sigma^2)$ und die Varianz σ^2 sei bekannt. Man erhält die Likelihood-Funktion

$$L(\theta, x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(\theta - x)^2\Big).$$

Diese ist in der Abbildung 6 dargestellt. Nach Beispiel 55 kann man dies leicht auf die i.i.d.-Situation übertragen: Seien $X_1, \ldots, X_n$ i.i.d. mit $X_1 \sim \mathcal{N}(\theta, \sigma^2)$. Die Varianz σ^2 sei bekannt. Dann gilt für die Likelihood-Funktion[16]

$$L(\theta, \mathbf{x}) \propto \exp\Big(-\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2}\Big).$$

Daraus erhält man die Log-Likelihood-Funktion mit einer geeigneten Konstanten $c \in \mathbb{R}$

$$l(\theta, \mathbf{x}) = c - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2}.$$

Die Log-Likelihood-Gleichung (53) ergibt direkt, dass

$$\hat{\theta}(\boldsymbol{x}) = \bar{x}.$$

[16] In dieser Gleichung ist L nur bis auf multiplikative Konstanten angegeben. $L(\theta) \propto f(\theta)$ bedeutet, es existiert eine von θ unabhängige Konstante C, so dass $L(\theta) = C \cdot f(\theta)$.

Die zweite Ableitung von l nach θ ist negativ und somit ist das gefundene $\widehat{\theta}$ Maximalstelle.

Die verschiedenen Schätzmethoden für den Normalverteilungsfall, etwa die Kleinste-Quadrate-Methode in Beispiel 75, ergeben folglich den gleichen Schätzer wie die Maximum-Likelihood-Methode.

Beispiel 57 (Gleichverteilung)**.** Es werde eine Population mit θ Mitgliedern betrachtet. Die Mitglieder seien nummeriert mit $1, \dots, \theta$. Von dieser Population werde n-mal mit Wiederholung gezogen. Mit X_i werde die gezogene Nummer des i-ten Zuges bezeichnet und das Maximum der Beobachtungen durch $x_{(n)} := \max\{x_1, \dots, x_n\}$. Es gilt, dass $P(X_i = r) = \theta^{-1} \mathbb{1}_{\{r \in \{1,\dots,\theta\}\}}$.

Nach Beispiel 55 ist die Likelihood-Funktion gegeben durch

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} \theta^{-1} \mathbb{1}_{\{x_i \in \{1,\dots,\theta\}\}} = \theta^{-n} \mathbb{1}_{\{x_{(n)} \le \theta, x_1,\dots,x_n \in \mathbb{N}\}} \tag{58}$$

$$= \begin{cases} 0 & \text{für } \theta \in \{1, \dots, x_{(n)} - 1\} \\ \max\{x_1, \dots, x_n\}^{-n} & \text{für } \theta = x_{(n)} \\ \theta^{-n} & \text{für } \theta > x_{(n)}. \end{cases}$$

Damit ergibt sich $\widehat{\theta} = X_{(n)}$ als Maximum-Likelihood-Schätzer. Die Likelihood-Funktion ist in Abbildung 7 dargestellt.

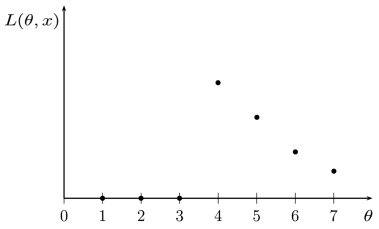


Abbildung 7: Die Likelihood-Funktion als Funktion von θ für eine Population mit θ Mitgliedern. Die Darstellung ist für $x_{(n)} = 4$.

Beispiel 59 (Warteschlange)**.** (Siehe Beispiel 31) Sei X die Anzahl der Kunden, welche an einem Schalter in n Stunden ankommen. Wir nehmen an, dass die Anzahl der ankommenden Kunden einem Poisson-Prozess folgt und bezeichnen die Intensität (beziehungsweise die erwartete Anzahl von Kunden pro Stunde) mit λ. Dann gilt $X \sim \text{Poiss}(n\lambda)$. Mit der Wahrscheinlichkeitsfunktion einer Poisson-Verteilung erhält man die Likelihood-Funktion

$$L(\lambda, x) = \frac{e^{-\lambda n}(\lambda n)^x}{x!}$$

für $x = 0, 1, \dots$. Damit ist die Log-Likelihood-Funktion

$$l(x, \lambda) = -\lambda n + x \ln(\lambda n) - \ln x!$$

und die Log-Likelihood-Gleichung (53) egibt

$$0 = \left.\frac{\partial l(\lambda, x)}{\partial \lambda}\right|_{\lambda = \widehat{\lambda}} = -n + \frac{x \cdot n}{\widehat{\lambda} \cdot n} = 0.$$

Somit ist $\widehat{\lambda} = \widehat{\lambda}(x) = x/n$. Die zweite Ableitung ist $-x/\lambda^2$, welche für $x > 0$ negativ ist. Somit erhält man für $x > 0$ das arithmetische Mittel

$$\widehat{\lambda}(x) = \frac{x}{n}$$

als den Maximum-Likelihood-Schätzwert für λ. Gilt allerdings $x = 0$, so existiert kein MLS für λ.

In dem regulären statistischen Modell $\mathscr{P} = \{p(\cdot,\theta) : \theta \in \Theta\}$ sei $E_\theta[T(\boldsymbol{X})]$ der Erwartungswert von $T(\boldsymbol{X})$ bezüglich der Dichte oder Wahrscheinlichkeitsfunktion $p(\cdot,\theta)$. Weiterhin sei das Bild von c durch $c(\Theta) := \{c(\theta) : \theta \in \Theta\}$ bezeichnet.

Satz 60 (MLS für eindimensionale exponentielle Familien)**.** *Betrachtet werde das reguläre statistische Modell $\mathscr{P} = \{p(\cdot,\theta) : \theta \in \Theta\}$ mit $\Theta \subset \mathbb{R}$ und*

$$p(\boldsymbol{x},\theta) = \mathbb{1}_{\{\boldsymbol{x}\in A\}} \exp\Big(c(\theta)T(\boldsymbol{x}) + d(\theta) + S(\boldsymbol{x})\Big), \quad \boldsymbol{x} \in \mathbb{R}^n.$$

Sei C das Innere von $c(\Theta)$, c injektiv und $\boldsymbol{x} \in \mathbb{R}^n$. Falls

$$E_\theta[T(\boldsymbol{X})] = T(\boldsymbol{x})$$

eine Lösung $\widehat{\theta}(\boldsymbol{x})$ besitzt mit $c(\widehat{\theta}(\boldsymbol{x})) \in C$, dann ist $\widehat{\theta}(\boldsymbol{x})$ der eindeutige Maximum-Likelihood-Schätzwert von θ.

Der Beweis geht in 3 Schritten:
1. - Wir betrachten die natürliche Familie
2. - Dann ist direkt $0 = l' = T(\boldsymbol{x}) + d_0'$ wobei wir schon wissen, dass $-d_0' = E[T(\boldsymbol{X})]$
3. - Jetzt ist $c(\Theta) \subset \{\eta \in H\}$, (Hierfür wichtig: Bemerkung 39) das Maximum wird also höchstens kleiner wenn wir nicht die natürliche Familie betrachten. Aber wir haben ja ein Maximum ! □.

Beweis. Betrachte zunächst die zugehörige natürliche exponentielle Familie in Darstellung (36). Sie ist gegeben durch $\{p_0(\cdot,\eta) : \eta \in H\}$ wobei $H := \{\eta \in \mathbb{R} : d_0(\eta) < \infty\}$ und

$$p(\boldsymbol{x},\eta) \quad = \quad \mathbb{1}_{\{\boldsymbol{x}\in A\}} \exp\Big(\eta \cdot T(\boldsymbol{x}) + d_0(\eta) + S(\boldsymbol{x})\Big).$$

Somit ist für einen inneren Punkt $\eta \in H$

$$\frac{\partial}{\partial\eta} l(\eta,\boldsymbol{x}) \quad = \quad T(\boldsymbol{x}) + d_0'(\eta) \quad \text{und} \quad \frac{\partial^2}{\partial\eta^2} l(\eta,\boldsymbol{x}) = d_0''(\eta).$$

Dann gilt nach Bemerkung 46 auch, dass

$$\begin{aligned} E_\eta[T(\boldsymbol{X})] &= -d_0'(\eta), \\ \mathrm{Var}_\eta(T(\boldsymbol{X})) &= -d_0''(\eta) > 0 \end{aligned}$$

und $d_0''(\eta) < 0$. Daraus folgt, dass die Log-Likelihood-Funktion l strikt konkav ist und somit ist die Log-Likelihood-Gleichung (53) äquivalent zu $E_\eta(T(\boldsymbol{X})) = T(\boldsymbol{x})$. Existiert eine Lösung $\boldsymbol{x}$ für $E_\eta(T(\boldsymbol{X})) = T(\boldsymbol{x})$, so muß diese Lösung der MLS sein. Eindeutigkeit folgt aus der strikten Konkavität von l.

Den allgemeinen Fall behandeln wir wie folgt. Sei $\boldsymbol{x} \in \mathbb{R}^n$ beliebig. Für die möglichen Werte der Log-Likelihood-Funktion gilt, dass

$$\big\{l(\theta,\boldsymbol{x}) = c(\theta)T(\boldsymbol{x}) + d(\theta) + S(\boldsymbol{x}) : \theta \in \Theta\big\} \subset \big\{\eta \cdot T(\boldsymbol{x}) + d_0(\eta) + S(\boldsymbol{x}) : \eta \in H\big\}, \tag{61}$$

denn für $\theta \in \Theta$ folgt aus der Injektivität von c, dass $d_0(c^{-1}(\theta)) < \infty$ nach Bemerkung 39. Falls $\widehat{\theta}(\boldsymbol{x})$ Lösung von $E_\theta(T(\boldsymbol{X})) = T(\boldsymbol{x})$

ist, dann maximiert $c(\widehat{\theta}(\boldsymbol{x}))$ die Gleichung $\eta \cdot T(\boldsymbol{x}) + d_0(\eta) + S(\boldsymbol{x})$ für alle $\eta \in H$ und weiterhin ist $\widehat{\eta}(\boldsymbol{x}) = c(\widehat{\theta}(\boldsymbol{x}))$. Dies folgt aus der Eindeutigkeit von $\widehat{\eta}(\boldsymbol{x})$ und der Injektivität von $c : \Theta \to \mathbb{R}$. Vergleichen wir mit (61), so erhält man das Maximum der Menge $\{\eta \cdot T(\boldsymbol{x}) + d_0(\eta) + S(\boldsymbol{x}) : \eta \in H\}$ mit $l(\widehat{\theta}(\boldsymbol{x}), \boldsymbol{x})$. Hierbei ist $\widehat{\theta}(\boldsymbol{x}) \in \Theta$ und somit maximiert $\widehat{\theta}(\boldsymbol{x})$ die Log-Likelihood-Funktion $l(\cdot, \boldsymbol{x})$. □

Beispiel 62 (Normalverteilungsfall, σ bekannt)**.** (Siehe Beispiel 56) Seien $X_1, \dots, X_n$ i.i.d. mit $X_1 \sim \mathcal{N}(\theta, \sigma^2)$ und die Varianz σ^2 sei bekannt. Nach Beispiel 50 ist die Verteilung von $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ eine exponentielle Familie mit natürlicher suffizienter Statistik $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$. Da

$$E_\theta[T(\boldsymbol{X})] = n\theta,$$

ist die Bedingung $E_\theta(T(\boldsymbol{X})) = T(\boldsymbol{x})$ äquivalent zu

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

Da $c(\theta) = \theta/\sigma^2$ nach Beispiel 40 gilt, ist c injektiv und das Bild von c ist $\mathbb{R}$. Damit liegt $\widehat{\theta}(\boldsymbol{X}) := \bar{X}$ im Inneren des Bildes von c. Mit Satz 60 folgt somit, dass $\widehat{\theta}(\boldsymbol{X}) = \bar{X}$ ein eindeutiger MLS ist.

Maximum-Likelihood in mehrdimensionalen Modellen

In diesem Abschnitt wird die Verallgemeinerung der Maximum-Likelihood-Methode vorgestellt, in welcher der Parameterraum Θ k-dimensional ist. Hierzu betrachten wir das reguläre statistische Modell $\mathscr{P}$ gegeben durch eine Familie von Dichten oder Wahrscheinlichkeitsfunktionen $\{p(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ mit $\Theta \subset \mathbb{R}^k$. Das zu $p(\cdot, \boldsymbol{\theta})$ gehörige Wahrscheinlichkeitsmaß sei mit $P_{\boldsymbol{\theta}}$ bezeichnet. Wir nehmen an, dass Θ offen ist. Falls die partiellen Ableitungen der Log-Likelihood-Funktion existieren und der MLS $\widehat{\boldsymbol{\theta}}$ existiert, so löst $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ die Log-Likelihood-Gleichung (53),

$$\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{x})\Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x})} = \mathbf{0}.$$

Wieder bezeichnen wir mit $E_{\boldsymbol{\theta}}(T(\boldsymbol{X}))$ den Erwartungswert von $T(\boldsymbol{X})$ bezüglich der Verteilung $P_{\boldsymbol{\theta}}$ und das Bild von c mit $c(\Theta) := \{c(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. Der folgende Satz gibt Kriterien für einen eindeutigen Maximum-Likelihood-Schätzer in K-parametrigen exponentiellen Familien.

Satz 63. *Betrachtet werde das reguläre statistische Modell* $\mathscr{P} = \{p(\cdot,\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta\}$ *aus einer K-parametrigen exponentiellen Familie, so dass für alle* $\boldsymbol{x} \in \mathbb{R}^n$ *und* $\boldsymbol{\theta} \in \Theta$

$$p(\boldsymbol{x},\boldsymbol{\theta}) = \mathbb{1}_{\{\boldsymbol{x}\in A\}} \exp\left(\sum_{i=1}^{K} c_i(\boldsymbol{\theta})T_i(\boldsymbol{x}) + d(\boldsymbol{\theta}) + S(\boldsymbol{x})\right), \quad \boldsymbol{\theta} \in \Theta. \tag{64}$$

Sei C das Innere von $c(\Theta)$ *und* $c_1,\dots,c_K$ *injektiv. Falls*

$$E_{\boldsymbol{\theta}}[T_i(\boldsymbol{X})] = T_i(\boldsymbol{x}), \quad i = 1,\dots,K$$

eine Lösung $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ *besitzt mit* $(c_1(\widehat{\boldsymbol{\theta}}(\boldsymbol{x})),\dots,c_K(\widehat{\boldsymbol{\theta}}(\boldsymbol{x})))^\top \in C$*, dann ist* $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ *der eindeutige Maximum-Likelihood-Schätzwert von* $\boldsymbol{\theta}$*.*

Der Beweis des Satzes ist dem eindimensionalen Fall ähnlich. In Verallgemeinerung von Beispiel 56 betrachten wir nun die Situation der MLS von normalverteilten Beobachtungen.

Beispiel 65 (MLS für Normalverteilung, μ und σ unbekannt)**.** Seien $X_1,\dots,X_n$ i.i.d. mit $X_i \sim \mathcal{N}(\mu,\sigma^2)$ und sowohl μ als auch σ^2 unbekannt. Setze $\boldsymbol{\theta} := (\mu,\sigma^2)^\top$ und $\Theta := \mathbb{R}\times\mathbb{R}^+$. Nach Beispiel 49 führt die Darstellung der Normalverteilung als exponentielle Familie gemäß Gleichung (64) zu $c_1(\boldsymbol{\theta}) = \mu/\sigma^2$ und $c_2(\boldsymbol{\theta}) = -1/2\sigma^2$. Damit ist $C = \mathbb{R}\times\mathbb{R}^-$ mit $\mathbb{R}^- := \{x \in \mathbb{R} : x < 0\}$. Weiterhin sind

$$T_1(\boldsymbol{x}) = \sum_{i=1}^{n} x_i, \quad T_2(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2.$$

Daraus ergeben sich die folgenden beiden Gleichungen. Zunächst ist $E_{\boldsymbol{\theta}}\big[T_1(\boldsymbol{X})\big] = n\mu$. Damit ist $E_{\boldsymbol{\theta}}[T_1(\boldsymbol{X})] = T_1(\boldsymbol{x})$ äquivalent zu

$$n\mu = \sum_{i=1}^{n} x_i,$$

woraus $\widehat{\mu} = \widehat{\theta}_1(\boldsymbol{X}) = \bar{X}$ folgt. Weiterhin ist

$$E_{\boldsymbol{\theta}}\big[T_2(\boldsymbol{X})\big] = \sum_{i=1}^{n} E_{\boldsymbol{\theta}}(X_i^2) = n\big(\sigma^2+\mu^2\big).$$

Damit ist $E_{\boldsymbol{\theta}}[T_2(\boldsymbol{X})] = T_2(\boldsymbol{x})$ äquivalent zu $n\big(\sigma^2+\mu^2\big) = \sum_{i=1}^n x_i^2$. Wir erhalten

$$\widehat{\sigma}^2 = \widehat{\theta}_2(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}\big(X_i-\bar{X}\big)^2, \tag{66}$$

falls $n \geq 2$. Damit erhalten wir den MLS für die Normalverteilung mit unbekanntem Mittelwert und unbekannter Varianz:

Mit Satz 63 folgt, dass für $X_1, \dots, X_n$ i.i.d. und $X_1 \sim \mathcal{N}(\mu, \sigma^2)$

$$\widehat{\boldsymbol{\theta}} = \left(\bar{X}, \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \right)^\top$$

der eindeutige Maximum-Likelihood-Schätzer für $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ ist.

Methode der kleinsten Quadrate

Die lineare Regression und in diesem Zusammenhang die Methode der kleinsten Quadrate ist eine Methode, die bereits Gauß für astronomische Messungen verwendete.

Das zur Anpassung der Regressionsgeraden an die Daten verwendete Prinzip der Minimierung eines quadratischen Abstandes findet in vielen unterschiedlichen Bereichen Anwendung. Die erhaltenen Formeln werden in der Numerik oft auch als verallgemeinerte Inverse verwendet.

In der Tat gab es sogar einen erheblichen Disput darum, wer die Methode als erster erfunden hatte. Legendre hat zumindest die „Mſthode des moindres carrś“als erster veröffentlicht - und zwar 1805. Gauß hatte diese Methoder aber bereits vorher angewendet (als 24 jähriger zur Vorhersage der Bahn des Zwergplaneten Ceres). Eine interessante Darstellung der Hintergründe findet sich unter `https://de.wikipedia.org/wiki/Methode_der_kleinsten_Quadrate`, von wo auch der Ausschnitt aus einem Bild von Gottlieb Biermann entnommen wurde, Foto: A. Wittmann.

Allgemeine und lineare Regressionsmodelle

Regressionsprobleme untersuchen die Abhängigkeit der *Zielvariablen* (Response, endogene Variable) von anderen Variablen (Kovariablen, unabhägige Variablen, exogene Variablen). Der Begriff Regression geht hierbei auf Experimente zur Schätzung der Köpergröße von Söhnen basierend auf der Körpergröße ihrer Väter zurück.

Definition 67. Eine *allgemeine Regression* ist gegeben durch einen zu bestimmenden r-dimensionalen Parametervektor $\boldsymbol{\theta} \in \Theta$ und bekannte, parametrische Funktionen $g_1, \dots, g_n : \Theta \to \mathbb{R}$. Das zugehörige *Modell* ist

$$Y_i = g_i(\boldsymbol{\theta}) + \epsilon_i \qquad i = 1, \dots, n.$$

Fehler, welche die folgende Annahme (WN) erfüllen, werden als *weißes Rauschen* (white noise) bezeichnet.

(WN) Für die Zufallsvariablen $\epsilon_1, \dots, \epsilon_n$ gilt:

(i) $E[\epsilon_i] = 0$ für alle $i = 1, \dots, n$.
(ii) $\mathrm{Var}(\epsilon_i) = \sigma^2 > 0$ für alle $i = 1, \dots, n$. σ^2 ist unbekannt.
(iii) $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ für alle $1 \leq i \neq j \leq n$.

Die Zufallsvariablen $\epsilon_1, \dots, \epsilon_n$ stellen wie in Beispiel 21 Abweichungen von der systematischen Beziehung $Y_i = g_i(\boldsymbol{\theta})$ dar. Die Bedingung (i) veranschaulicht, dass die Regression keinen systematischen Fehler macht. Die Bedingung (ii) verlangt eine homogene Fehlervarianz, was man als *homoskedastisch* bezeichnet.

Die Bedingungen (i)-(iii) gelten, falls $\epsilon_1, \dots, \epsilon_n$ i.i.d. mit Erwartungswert 0 und $\mathrm{Var}(\epsilon_i) > 0$. Ein wichtiger Spezialfall ist durch die zusätzliche Normalverteilungsannahme $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ gegeben. An

dieser Stelle sei noch einmal auf die Analogie zu den Annahmen des Meßmodells aus Beispiel 21 verwiesen.

Beispiel 68 (Meßmodell aus Beispiel 21)**.** Es werden n Messungen einer physikalischen Konstante θ vorgenommen. Variiert der Messfehler additiv um θ, so erhält man

$$Y_i = \theta + \epsilon_i, \quad i = 1, \dots, n.$$

In diesem Fall ist $r = 1$ und $g_i(\theta) = \theta$. Die Messergebnisse werden stets mit $y_1, \dots, y_n$ bezeichnet.

Beispiel 69 (Einfache lineare Regression)**.** Die einfache lineare Regression wurde bereits in Beispiel 51 im Kontext von exponentiellen Familien betrachtet, welches wir an dieser Stelle wieder aufgreifen. Man beobachtet Paare von Daten $(x_1, y_1), \dots, (x_n, y_n)$. Die Größen $x_1, \dots, x_n$ werden als deterministisch und bekannt betrachtet und es wird folgendes statistisches Modell angenommen:

$$Y_i = \theta_1 + \theta_2 x_i + \epsilon_i.$$

Y_i heißt *Zielvariable* mit Beobachtungen y_i und x_i heißt *Kovariable*. Wir verwenden $g_i(\theta_1, \theta_2) = \theta_1 + \theta_2 x_i$ als parametrische Funktion. In Abbildung 8 werden die Beobachtungen zusammen mit der geschätzten Regressionsgeraden $x \mapsto \widehat{\theta}_1 + \widehat{\theta}_2 x$ bei einer einfachen linearen Regression gezeigt.

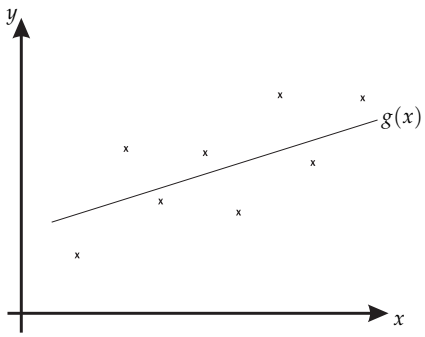


Abbildung 8: Eine einfache lineare Regression. Beobachtet werden Paare (x_i, y_i), $i = 1, \dots, n$, welche in der Abbildung durch Kreuze gekennzeichnet sind. Die den Daten angepasste Regressionsgerade $g : x \to \hat{\theta}_1 + \hat{\theta}_2 x$ mit geschätzten Paramtern $\widehat{\theta}_1$ und $\widehat{\theta}_2$ ist ebenfalls dargestellt.

Methode der kleinsten Quadrate

Bei dieser Methode schätzt man den unbekannten Parameter $\boldsymbol{\theta}$ durch den Schätzwert $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{y})$, welcher den Abstand von $E_{\boldsymbol{\theta}}(\boldsymbol{Y})$ und den beobachteten Daten $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ unter allen $\boldsymbol{\theta} \in \Theta$ minimiert. Der Abstand wird hierbei durch einen *quadratischen* Abstand Q gemessen. Das allgemeine Regressionsmodell wurde bereits in Definition 67 definiert.

Definition 70. Der quadratische Abstand $Q : \Theta \times \mathbb{R}^n \to \mathbb{R}^+$ sei definiert durch

$$Q(\boldsymbol{\theta}, \boldsymbol{y}) := \sum_{i=1}^{n} \big(y_i - g_i(\boldsymbol{\theta})\big)^2, \quad \boldsymbol{y} \in \mathbb{R}^n. \tag{71}$$

Gilt für eine meßbare Funktion $\widehat{\boldsymbol{\theta}} : \mathbb{R}^n \to \Theta$, dass

$$Q(\widehat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{y}) \leq Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{y}) \qquad \text{für alle } \tilde{\boldsymbol{\theta}} \in \Theta \text{ und } \boldsymbol{y} \in \mathbb{R}^n,$$

so heißt $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ *Kleinste-Quadrate-Schätzer* (KQS) von $\boldsymbol{g}(\boldsymbol{\theta})$.

Ein KQS wird auch als *Least Squares Estimator* (LSE) bezeichnet. Sind die Funktionen g_i differenzierbar, und ist das Bild von $(g_1, \dots, g_n)$ abgeschlossen, so ist dies eine hinreichende Bedingung dafür, dass $\widehat{\boldsymbol{\theta}}$ wohldefiniert ist. Ist darüber hinaus $\Theta \subset \mathbb{R}^r$ offen, so muss $\widehat{\boldsymbol{\theta}}$ notwendigerweise die *Normalengleichungen*

$$\frac{\partial}{\partial \theta_j} Q(\boldsymbol{\theta}, \boldsymbol{y})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(\boldsymbol{y})} = 0, \quad j = 1, \dots, r$$

erfüllen. Mit der Definition von Q aus (71) sind die Normalengleichungen äquivalent zu folgender Gleichung:

$$\sum_{i=1}^{n} \left(\big(y_i - g_i(\boldsymbol{\theta})\big) \cdot \frac{\partial}{\partial \theta_j} g_i(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(\boldsymbol{y})} \right) = 0, \qquad j = 1, \dots, r. \tag{72}$$

Bemerkung 73. In der *linearen Regression* sind die Funktionen $g_i(\theta_1, \dots, \theta_r)$ linear in $\theta_1, \dots, \theta_r$. In diesem Fall erhält man ein lineares Gleichungssystem, welches man explizit lösen kann.

Die Kleinste-Quadrate-Methode soll nun an den obigen Beispielen illustriert werden.

Beispiel 74 (Meßmodell)**.** Gegeben sei wie in Beispiel 68 ein lineares Modell

$$Y_i = \theta + \epsilon_i, \quad i = 1, \dots, n.$$

Dann ist $g_i(\theta) = \theta$ und somit $\frac{\partial}{\partial \theta} g_i(\theta) = 1$ für alle $i = 1, \dots, n$. Die Normalengleichungen (72) ergeben

$$\sum_{i=1}^{n} (y_i - \widehat{\theta}(\boldsymbol{y})) = 0.$$

Hieraus folgt unmittelbar, dass $\widehat{\theta}(\boldsymbol{y}) = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ ist, das arithmetische Mittel der Beobachtungen. Nach Beispiel 50 ist $\bar{Y}$ darüber hinaus eine suffiziente Statistik für θ.

Beispiel 75 (Einfache lineare Regression)**.** In Fortsetzung von Beispiel 69 betrachten wir ein lineares Modell gegeben durch

$$Y_i = \theta_1 + \theta_2 x_i + \epsilon_i, \quad i = 1, \dots, n.$$

In diesem Fall ist $g_i(\boldsymbol{\theta}) = \theta_1 + \theta_2 x_i$ und $\frac{\partial g_i}{\partial \theta_1}(\boldsymbol{\theta}) = 1$, $\frac{\partial g_i}{\partial \theta_2}(\boldsymbol{\theta}) = x_i$. Schreiben wir kurz $\widehat{\theta}_i = \widehat{\theta}_i(\boldsymbol{y})$, $i = 1, 2$ so erhalten die Normalengleichungen (72) folgende Gestalt:

$$\sum_{i=1}^{n} \big(y_i - \widehat{\theta}_1 - \widehat{\theta}_2 \, x_i\big) \cdot 1 = 0 \tag{76}$$

$$\sum_{i=1}^{n} \big(y_i - \widehat{\theta}_1 - \widehat{\theta}_2 \, x_i\big) \cdot x_i = 0. \tag{77}$$

Aus Gleichung (76) erhält man mit $\bar{y} := \frac{\sum_{i=1}^n}{n} y_i$ und $\bar{x} := \frac{\sum_{i=1}^n}{n} x_i$, dass

$$\widehat{\theta}_1 = \bar{y} - \widehat{\theta}_2 \, \bar{x}.$$

Setzt man dies in (77) ein, so ergibt sich

$$\begin{aligned} &\sum_{i=1}^n x_i y_i - \left(\bar{y} - \widehat{\theta}_2 \bar{x}\right) \sum_{i=1}^n x_i - \widehat{\theta}_2 \sum_{i=1}^n x_i^2 = 0 \\ \Leftrightarrow \qquad & \frac{\sum_{i=1}^n}{n} x_i y_i - \bar{y}\,\bar{x} = \widehat{\theta}_2 \Big(\frac{\sum_{i=1}^n}{n} x_i^2 - \bar{x}^2 \Big). \end{aligned}$$

Da weiterhin $\sum_{i=1}^n x_i^2 - n(\bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ und $\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ gilt, erhält man folgende Aussage.

In der *einfachen linearen Regression* ist

$$\widehat{\theta}_2(\boldsymbol{y}) = \frac{\sum_{i=1}^n \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}$$

$$\widehat{\theta}_1(\boldsymbol{y}) = \bar{y} - \widehat{\theta}_2 \bar{x}.$$

Die Gerade $x \mapsto \widehat{\theta}_1(\boldsymbol{y}) + \widehat{\theta}_2(\boldsymbol{y}) x$ heißt *Regressionsgerade*. Sie minimiert die Summe der quadratischen Abstände zwischen (x_i, y_i) und $(x_i, \theta_1 + \theta_2 x_i)$. Der Erwartungswert von Y_i, gegeben durch $E(Y_i) = \theta_1 + \theta_2 x_i$ wird durch

$$\widehat{y}_i := \widehat{\theta}_1(\boldsymbol{y}) + \widehat{\theta}_2(\boldsymbol{y})\, x_i, \quad i = 1, \ldots, n$$

geschätzt. Die Regressionsgerade zusammen mit y_i und $\widehat{y}_i$ werden in Abbildung 9 illustriert.

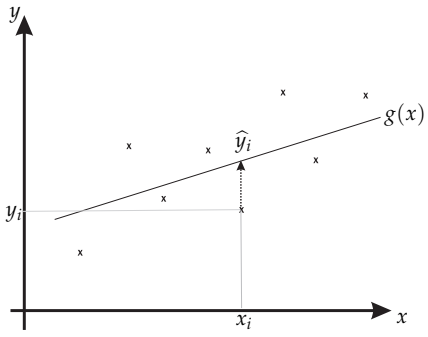


Abbildung 9: Illustration der Regressionsgeraden $g : x \mapsto \widehat{\theta}_1(\boldsymbol{y}) + \widehat{\theta}_2(\boldsymbol{y}) x$ und der Erwartung eines Datenpunktes $\widehat{y}_i = \widehat{\theta}_1(\boldsymbol{y}) + \widehat{\theta}_2(\boldsymbol{y}) x_i$.

Vergleich der Maximum-Likelihood-Methode mit anderen Schätzverfahren

In diesem Abschnitt halten wir einige Beobachtungen fest, die den MLS in andere Schätzmethoden einordnen.

(i) Das Maximum-Likelihood-Verfahren für diskrete Zufallsvariablen entspricht dem Substitutionsprinzip.

(ii) Der Kleinste-Quadrate-Schätzer einer allgemeinen Regression unter Normalverteilungsannahme aus Abschnitt kann als Maximum-Likelihood-Schätzer betrachtet werden: Für $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ und

$$Y_i = g_i(\boldsymbol{\theta}) + \epsilon_i, \quad i = 1, \ldots, n$$

mit i.i.d. $\epsilon_1, \ldots, \epsilon_n$ und $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ ist die Likelihood-Funktion gegeben durch

$$L(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - g_i(\boldsymbol{\theta}))^2 \right). \tag{78}$$

Für alle $\sigma^2 > 0$ ist (78) genau dann maximal, wenn

$$\sum_{i=1}^{n} \left(x_i - g_i(\theta_1, \ldots, \theta_r)\right)^2$$

minimal ist. Damit entspricht der Kleinste-Quadrate-Schätzer in diesem Fall dem Maximum-Likelihood-Schätzer.

Anpassungstests

In diesem Buch gehen wir stets von einem parametrischen Modell von der Form $\mathscr{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ aus. Wie wir in diesem Abschnitt gesehen haben, kann man unter dieser Annahme verschiedene Schätzern herleiten und in den folgenden Kapiteln werden wir deren Optimalitätseigenschaften analysieren. In der praktischen Anwendung muss man Annahme dass die Daten dem Modell $\mathscr{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ entstammen mit einem geeigneten Test überprüfen. Dies führt auf natürliche Weise zu so genannten nichtparametrischen Tests, wie z.B. den χ^2-Anpassungstest oder eine der vielen Varianten des Kolmogorov-Smirnov-Anpassungstests.

Konfidenzintervalle und Hypothesentests

Dieses Kapitel stellt zunächst Konfidenzintervalle im ein- und mehrdimensionalen Fall vor und behandelt danach Hypothesentests nach dem Ansatz von Neyman und Pearson. Abschließend wird die Dualität zwischen den beiden Begriffen erläutert.

Konfidenzintervalle

Schätzt man einen Parameter aus Daten, so erhält man als Ergebnis eines Schätzverfahrens einen Schätzwert. Es ist allerdings unerläßlich, neben einem Schätzwert stets eine Angabe über seine Qualität oder seine Präzision zu machen. Ein zuverläßliches und allgemeines Merkmal für die Qualtität eines Schätzers ist ein *Konfidenzintervall*. Als Ergebnis einer Schätzung sollte stets Schätzwert und Konfidenzintervall mit zugehörigem Konfidenzniveau angegeben werden.

Der eindimensionale Fall

Sei $T(\boldsymbol{X})$ ein Schätzer von $q(\theta) \in \mathbb{R}$. Wir suchen zufällige Grenzen $\underline{T}(\boldsymbol{X}) \leq q(\theta) \leq \overline{T}(\boldsymbol{X})$, so dass die Wahrscheinlichkeit, dass $q(\theta)$ von $[\underline{T}(\boldsymbol{X}), \overline{T}(\boldsymbol{X})]$ überdeckt wird, ausreichend hoch ist. Ein solches zufälliges Intervall nennen wir Zufallsintervall. Fixiert man ein kleines Toleranzniveau α, so interessiert man sich für Statistiken $\underline{T}$ und $\overline{T}$ mit der folgenden Eigenschaft.

Definition 79. Ein durch $\underline{T}(\boldsymbol{X}) \leq \overline{T}(\boldsymbol{X})$ gegebenes Zufallsintervall $[\underline{T}(\boldsymbol{X}), \overline{T}(\boldsymbol{X})]$ für welches für alle $\theta \in \Theta$ gilt, dass

$$P_\theta \left(q(\theta) \in [\underline{T}(\boldsymbol{X}), \overline{T}(\boldsymbol{X})] \right) \geq 1 - \alpha, \tag{80}$$

heißt $(1-\alpha)$-*Konfidenzintervall* für $q(\theta)$ zum *Konfidenzniveau* $1 - \alpha \in [0, 1]$.

Hierbei verwenden wir folgenden Sprachgebrauch: Ein $(1-\alpha)$-Konfidenzintervall bedeutet ein $(1-\alpha) \cdot 100$ %-Konfidenzintervall;

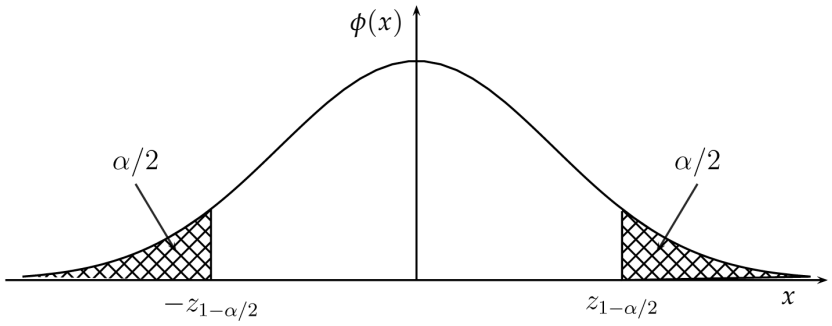


Abbildung 10: Dichte der Standardnormalverteilung mit den $\alpha/2$ und $1-\alpha/2$-Quantilen.

ist etwa $\alpha = 0.05$, so verwenden wir synonym die Bezeichnung 0.95-Konfidenzintervall und 95%-Konfidenzintervall. Für ein gegebenes Konfidenzintervall ist ein Intervall, welches dieses einschließt wieder ein Konfidenzintervall (auch zum gleichen Konfidenzniveau). Allerdings sind wir typischerweise daran interessiert, für ein vorgegebenes Konfidenzniveau das kleinste Intervall zu finden, welches die Überdeckungseigenschaft (80) erfüllt. Ist dies der Fall, so erwartet man approximativ, dass in n Beobachtungen $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ von i.i.d. Zufallsvariablen mit der gleichen Verteilung wie $\boldsymbol{X}$ in $(1-\alpha)n$ Fällen $[\underline{T}(\boldsymbol{x}_i), \overline{T}(\boldsymbol{x}_i)]$ den wahren Parameter $q(\theta)$ enthält.

Handelt es sich um ein symmetrisches Intervall, so nutzen wir die Schreibweise

$$a \pm b := [a-b, a+b].$$

Beispiel 81 (Normalverteilung, σ bekannt: Konfidenzintervall)**.** Seien $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\theta, \sigma^2)$ und σ^2 sei bekannt. Als Schätzer für θ verwenden wir $\bar{X}$. Da die $\mathcal{N}(\theta, \sigma^2)$-Verteilung symmetrisch um θ ist, liegt es nahe als Konfidenzintervall ein symmetrisches Intervall um $\bar{X}$ zu betrachten. Für $c > 0$ gilt

$$P_\theta\left(\bar{X} - c\frac{\sigma}{\sqrt{n}} \le \theta \le \bar{X} + c\frac{\sigma}{\sqrt{n}}\right) = P_\theta\left(\left|\frac{\bar{X}-\theta}{\sigma/\sqrt{n}}\right| \le c\right).$$

Da $\frac{\bar{X}-\theta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, folgt

$$P_\theta\left(\left|\frac{\bar{X}-\theta}{\sigma/\sqrt{n}}\right| \le c\right) = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1.$$

Wir suchen das kleinste Konfidenzintervall, welches die Überdeckungseigenschaft (80) erfüllt, suchen wird ein $c > 0$ so, dass $2\Phi(c) - 1 = 1 - \alpha$ gilt. Mit

$$z_a := \Phi^{-1}(a)$$

sei das *a-Quantil* der Standardnormalverteilung bezeichnet. Dann ist das symmetrische Intervall

$$\bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

ein $(1-\alpha)$-Konfidenzintervall für θ; siehe Abbildung 10. Da $z_{0.975} = 1.96$ gilt, ist in einer Stichprobe mit $\bar{x} = 5, \sigma = 1, n = 100$ das 95%-Konfidenzintervall für θ gegeben durch 5 ± 1.96.

Die Größe $z_{0.975} = 1.96$ taucht somit immer in solchen Kontexten aus. Das Konfidenzintervall ist fast doppelt so groß wie das Interval $\pm\sigma$ um den Mittelwert (was man auch gerne die 2σ-Regel nennt). Das mit einem σ überdeckte Intervall hat die Überdeckungswahrscheinlichkeit von 0,683, das mit 3σ überdeckte Intervall diejenige von 0,997 - Entnommen aus Lambacher / Schweizer: Kursstufe.

Das Testen von Hypothesen

Bisher haben wir Schätzverfahren betrachtet und entwickelt, welche man beispielsweise nutzen kann, um aus den Daten die Wirksamkeit einer Therapie zu schätzen. Allerdings ist man oft nicht direkt an dem Schätzwert interessiert, sondern man möchte entscheiden, ob diese Therapie hilft oder nicht. Hierfür wird man wegen der Zufälligkeit des Problems keine absolute Entscheidung treffen können, sondern zu jeder Zeit muss man eine gewisse Wahrscheinlichkeit für eine Fehlentscheidung akzeptieren, ähnlich wie bei den Konfidenzintervallen.

Im Folgenden führen wir das Konzept des statistischen Tests zur Überprüfung von Hypothesen auf Basis einer Stichprobe ein. Stets gehen wir von einem statistischen Modell $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ mit $\boldsymbol{X} \sim P_{\boldsymbol{\theta}}$ aus. Allerdings zerlegt die betrachtete Fragestellung den Parameterraum disjunkt in die zwei Hypothesen Θ_0 und Θ_1 mit $\Theta = \Theta_0 \oplus \Theta_1$, was gleichbedeutend ist mit $\Theta_0 \cap \Theta_1 = \emptyset$ und $\Theta_0 \cup \Theta_1 = \Theta$. Die beiden Parameterbereiche Θ_0 und Θ_1 stehen für unterschiedliche Hypothesen. Im obigen Beispiel würde man Θ_0 als den Bereich wählen, in welchem die Therapie nicht hilft; in dem Bereich Θ_1 hilft hingegen die Therapie. Wir verwenden die folgenden Bezeichnungen:

$H_0 = \{\boldsymbol{\theta} \in \Theta_0\}$ heißt *Null-Hypothese* und

$H_1 = \{\boldsymbol{\theta} \in \Theta_1\}$ heißt *Alternative*.

Oft schreiben wir hierfür $H_0 : \boldsymbol{\theta} \in \Theta_0$ gegen $H_1 : \boldsymbol{\theta} \in \Theta_1$. Die Bezeichnung Null-Hypothese stammt vom englischen Begriff *to nullify* = entkräften, widerlegen. Wie wir später sehen werden, ist die Hypothese, die widerlegt werden soll, stets als Null-Hypothese zu wählen.

Besteht Θ_0 aus einem einzigen Element, $\Theta_0 = \{\boldsymbol{\theta}_0\}$, so spricht man von einer *einfachen* Hypothese, ansonsten handelt es sich um eine

zusammengesetzte Hypothese. Ist $\Theta \subset \mathbb{R}$ und die Alternative von der Form $\Theta_1 = \{\theta : \theta \neq \theta_0\}$, so nennt man sie *zweiseitig*; ist sie von der Form $\Theta_1 = \{\theta : \theta > \theta_0\}$, so heißt sie *einseitig*.

Um eine Entscheidung zwischen den beiden Hypothesen H_0 und H_1 treffen zu können, stellt man eine Entscheidungsregel auf, welche wir *Test* nennen.

Definition 82. Ein *Test* δ ist eine messbare Funktion der Daten $\boldsymbol{X}$ mit Werten in $[0,1]$. Dabei bedeutet

- $\delta(\boldsymbol{X}) = 0$: Die Null-Hypothese wird akzeptiert.
- $\delta(\boldsymbol{X}) = 1$: Die Null-Hypothese wird verworfen.

Der Bereich $\{\boldsymbol{x} : \delta(\boldsymbol{x}) = 1\}$ heißt der *kritische Bereich* oder *Verwerfungsbereich* des Tests. Ist $T(\boldsymbol{X})$ eine Statistik und gilt $\delta(\boldsymbol{X}) = \mathbb{1}_{\{T(\boldsymbol{X}) \geq c\}}$, so heißt c *kritischer Wert* des Tests.

Beispiel 83 (Test für Bernoulli-Experiment)**.** Ein neues Medikament soll getestet werden, welches die Gesundungsrate einer Krankheit erhöhen soll. Die *Null-Hypothese* ist, dass das Medikament keine Wirkung hat. Aus Erfahrung weiß man, dass ein Anteil $\theta_0 = 0.2$ von Probanden ohne Behandlung gesundet. Es werden n Patienten getestet und deren Gesundungsrate beobachtet. Als statistisches Modell betrachten wir $X_1, \dots, X_n$ i.i.d. mit $X_1 \sim \text{Bernoulli}(\theta)$. Interessiert sind wir an der Entscheidung, ob $H_0 : \theta = \theta_0$ oder $H_1 : \theta > \theta_0$ vorliegt. Letztere, einseitige Hypothese verdeutlicht, dass wir nachweisen wollen, dass das Medikament nicht schädlich ist, sondern eine Verbesserung der Gesundungsrate bewirkt. Als Teststatistik verwenden wir den MLS-Schätzer $\bar{X}$. Ist $\bar{X}$ deutlich größer als θ_0, so spricht dies für H_1 und gegen H_0. Für ein noch zu bestimmendes Niveau wird man sich für H_1 entscheiden, falls $\bar{X}$ über diesem Niveau liegt, und sonst für H_0. Die Verteilung von $n\bar{X} = \sum_{i=1}^n X_i$ lässt sich leichter handhaben als die von $\bar{X}$. Folglich verwenden wir die Tests δ_k mit

$$\delta_k(\boldsymbol{X}) := \begin{cases} 1 & \sum_{i=1}^n X_i \geq k \\ 0 & \text{sonst.} \end{cases} \tag{84}$$

Die Wahl eines geeigneten k hängt von einer Fehlerwahrscheinlichkeit ab, die wir im folgenden Abschnitt einführen.

Fehlerwahrscheinlichkeiten und Güte

In unseren statistischen Tests betrachten wir stets zwei Hypothesen. Bei der Entscheidung für eine jede kann man einen Fehler machen.

Diese beiden Fehler können eine unterschiedliche Wahrscheinlichkeit haben und aus diesem Grund müssen wir stets beide Fehlerquellen im Auge behalten. Man erhält folgende Fälle: Ist H_0 wahr und ergibt der Test „H_0 wird akzeptiert", so macht man keinen Fehler; ebenso falls H_1 wahr ist und der Test ergibt „H_0 wird verworfen". Ist allerdings H_0 wahr und der Test ergibt „H_0 wird verworfen", so macht man den so genannten *Fehler 1. Art*. Andererseits, ist H_1 wahr, und ergibt der Test „H_0 wird akzeptiert", so macht man den *Fehler 2. Art*. Wir fassen dies in der folgenden Tabelle zusammen.

	H_0 wahr	H_1 wahr
H_0 wird akzeptiert	kein Fehler	Fehler 2.Art
H_0 wird verworfen	Fehler 1. Art	kein Fehler

Man geht wie folgt vor: Die Hypothese H_0 ist so gewählt, dass man sie ablehnen will. Somit ist der Fehler 1. Art für die Fragestellung wichtiger als der Fehler 2. Art. Man gibt sich ein Niveau α vor und wählt den Test so, dass der Fehler 1. Art höchstens α ist. Unterschiedliche Tests werden anhand ihres Fehlers 2. Art (Güte) verglichen.

Definition 85. Für einen Test δ ist die *Gütefunktion* $G_\delta : \Theta \to [0,1]$ definiert durch

$$G_\delta(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\delta(\boldsymbol{X})).$$

Ist $\delta \in \{0,1\}$, so ist die Güte eines Tests für vorgegebenes $\boldsymbol{\theta}$ gerade die Wahrscheinlichkeit, sich für die Alternative H_1 zu entscheiden. Ist $\boldsymbol{\theta} \in \Theta_0$, so ist das gerade die Wahrscheinlichkeit für einen Fehler 1. Art. Damit erhält man folgende Interpretation von $G_\delta(\boldsymbol{\theta})$:

$$\begin{cases} \text{Güte des Tests gegen die Alternative,} & \boldsymbol{\theta} \in \Theta_1 \\ \text{Wahrscheinlichkeit des Fehlers 1. Art für den wahren Wert } \boldsymbol{\theta}, & \boldsymbol{\theta} \in \Theta_0. \end{cases}$$

Gilt für einen Test δ, dass

$$\sup_{\boldsymbol{\theta} \in \Theta_0} G_\delta(\boldsymbol{\theta}) \leq \alpha$$

sagt man, der Test hat das *Signifikanzniveau* α. Gilt für δ

$$\sup_{\boldsymbol{\theta} \in \Theta_0} G_\delta(\boldsymbol{\theta}) = \alpha,$$

so nennen wir den Test δ einen *Level-α-Test*. Bei einem Test mit Signifikanzniveau α könnte man möglicherweise auch ein kleineres

Niveau α wählen; bei einem Level-α-Test ist das nicht der Fall, siehe Beispiel 86.

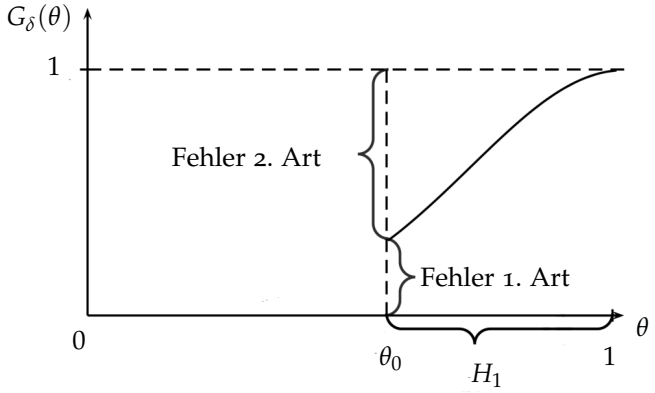


Abbildung 11: Illustration der Fehlerwahrscheinlichkeiten und der Gütefunktion eines Tests δ für das Testproblem $H_0 : \theta = \theta_0$ gegen $H_1 : \theta > \theta_0$ im Parameterraum $\Theta = \{\theta : 0 \leq \theta_0 \leq \theta \leq 1\}$.

Beispiel 86 (Test mit Signifikanzniveau α und Level-α-Test)**.** Ist $X \sim \mathcal{N}(\mu, 1)$, so ist $\delta(X) = \mathbb{1}_{\{X>c\}}$ ein Test für $H_0 : \mu = 0$ gegen $H_1 : \mu > 0$. Für ein vorgegebenes $\alpha \in (0,1)$ erhält man für jedes $c \geq \Phi^{-1}(1-\alpha)$ einen Fehler 1. Art mit einer Wahrscheinlichkeit kleiner als α. Diese Tests sind somit alle Tests mit Signifikanzniveau α. Aber nur für $c = \Phi^{-1}(1-\alpha)$ erhält man einen Level-α-Test.

Beispiel 87 (Fortführung von Beispiel 83)**.** Für das Testproblem $H_0 : \theta = \theta_0$ gegen $H_1 : \theta > \theta_0$ sollen die Tests δ_k aus Gleichung (84) verwendet werden. Wir setzen $S := n\bar{X} = \sum_{i=1}^n X_i$ und erinnern daran, dass S nach Aufgabe **??** gerade $\mathrm{Bin}(n,\theta)$-verteilt ist. Die Wahrscheinlichkeit, einen Fehler 1. Art zu begehen ist demnach

$$P_{\theta_0}\left(\delta_k(\boldsymbol{X}) = 1\right) = P_{\theta_0}(S \geq k) = \sum_{j=k}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j}.$$

Die Wahrscheinlichkeit einen Fehler 2. Art zu begehen hingegen hängt von dem *unbekannten* Wert $\theta \in \Theta_1$ ab. Für den Fehler 2. Art gilt $\theta \in \Theta_1$ und wir erhalten folgende Wahrscheinlichkeit für einen Fehler 2. Art:

$$P_{\theta}\left(\delta_k(\boldsymbol{X}) = 0\right) = P_{\theta}(S < k) = \sum_{j=0}^{k-1} \binom{n}{j} \theta^j (1-\theta)^{n-j}.$$

Schließlich ergibt sich folgende Gütefunktion

$$G_{\delta_k}(\theta) = P_\theta(S \geq k) = \sum_{j=k}^{n} \binom{n}{j} \theta^j (1-\theta)^{n-j}, \quad \theta \in \Theta.$$

Die zugehörigen Fehlerwahrscheinlichkeiten und die Gütefunktion sind in Abbildung 11 illustriert.

Beispiel 88 (Tests: Anwendungsbeispiele)**.** Zur Illustration von statistischen Tests stellen wir zwei Beispiele aus der Anwendung vor.

1. Eine Medizinerin möchte die Wirkung eines neuen Medikaments testen. Dabei erwartet sie, dass das neue Medikament wirksam ist. Aus diesem Grund verwendet sie die Hypothesen H_0: Medikament hat keine Wirkung gegen H_1: Medikament hat Wirkung. Ihr Ziel ist es, H_0 abzulehnen; falls H_0 aber nicht abgelehnt werden kann, dann wird sie nichts vermelden und an Verbesserungen arbeiten.

2. Ein Verbraucherberater untersucht Kindersitze für Autos. Er möchte nachweisen, dass die mittlere Kraft μ, welche benötigt wird bis der Kindersitz zerbricht, bei einer bestimmten Marke niedriger ist als die entsprechende Kraft μ_0 für andere Marken. Das heißt, er möchte H_0: $\mu \geq \mu_0$ gegen H_1: $\mu < \mu_0$ testen. Falls H_0 nicht abgelehnt werden kann, dann wird er nichts vermelden, da in diesem Fall eine Warnung vor diesem Typ von Kindersitzen nicht berechtigt wäre.

Generell kann man Folgendes formulieren: Falls die Null-Hypothese H_0 abgelehnt wird, dann wird ein Fehler (Fehler 1. Art) höchstens mit der Wahrscheinlichkeit α gemacht. Falls H_0 jedoch nicht abgelehnt werden kann, dann ist der Fehler (in diesem Fall der Fehler 2. Art) nicht kontrolliert, d.h. die Wahrscheinlichkeit für einen Fehler 2. Art kann in bestimmten Situationen beliebig nahe an 1 sein. Daher sagt man, dass „H_0 nicht verworfen werden kann“ oder „es gibt nicht genügend Evidenz für einen signifikanten Effekt“.

Beispiel 89 (Fortsetzung von Beispiel 83)**.** Für das Testproblem $H_0 : \theta = \theta_0$ gegen $H_1 : \theta > \theta_0$ sollen die Tests δ_k aus Gleichung (84) verwendet werden. Hierbei ist wieder

$$S = S(\boldsymbol{X}) = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, \theta).$$

Man wählt $k_0 = k(\theta_0, \alpha)$ so, dass die Wahrscheinlichkeit für einen Fehler 1. Art kleiner oder gleich α ist, also

$$P_{\theta_0}(S \geq k_0) \leq \alpha \tag{90}$$

gilt. Ein solches k_0 existiert, da

$$P_{\theta_0}(S \geq k) = \sum_{j=k}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j}$$

monoton fallend in k ist. Für genügend großes n mit $\min(n\theta_0, n(1-\theta_0)) \geq 5$ kann man auch folgende Approximation durch die Normal-

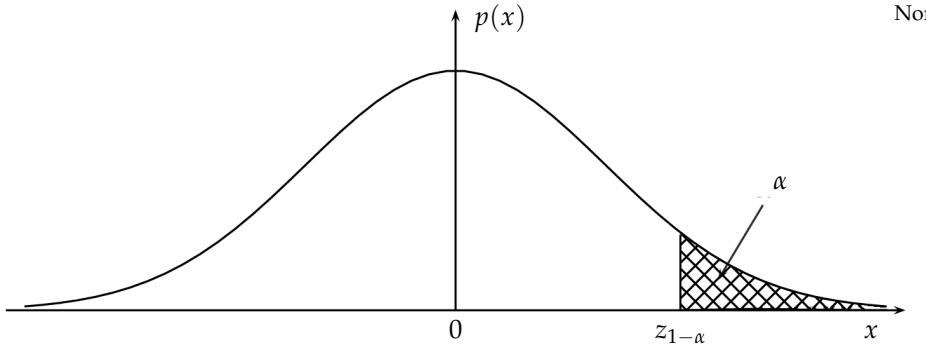


Abbildung 12: Das $(1-\alpha)$-Quantil der Normalverteilung $z_{1-\alpha}$.

verteilung verwenden:

$$P_\theta(S \geq k) \approx P_\theta\left(\frac{\sqrt{n}(\bar{X}-\theta)}{\sqrt{\theta(1-\theta)}} \geq \frac{k-n\theta-0.5}{\sqrt{n\theta(1-\theta)}}\right) \approx 1-\Phi\left(\frac{k-n\theta-0.5}{\sqrt{n\theta(1-\theta)}}\right).$$

Hierbei ist der Term 0.5 im Zähler die so genannte *Stetigkeitskorrektur*, die die Approximation verbessert. Dann gilt

$$P_{\theta_0}(S \geq k) \approx 1-\Phi\Big(\frac{k-n\theta_0-0.5}{\sqrt{n\theta_0(1-\theta_0)}}\Big) \leq \alpha.$$

Demnach ist (90) (approximativ) gleichbedeutend mit

$$k_0 \geq x_0 \text{ mit } x_0 = n\theta_0 + 0.5 + z_{1-\alpha}\sqrt{n\theta_0(1-\theta_0)}, \tag{91}$$

wobei $z_{1-\alpha}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung ist (siehe Abbildung 12). Somit ist der Test

$$\delta_{k_0}(\boldsymbol{X}) := \mathbb{1}_{\{S(\boldsymbol{X})>k_0\}} = \mathbb{1}_{\{n\bar{X}>k_0\}}$$

ein Test mit (approximativem) Signifikanzniveau α für H_0 gegen H_1, falls (91) (und damit (90), ebenfalls approximativ) gilt.

Beispiel 92 (Normalverteilung: Einseitiger Gauß-Test für μ)**.** In diesem Beispiel wird ein einseitiger Test für den Erwartungswert einer Normalverteilung mit bekannter Varianz vorgestellt. Seien dazu $X_1, \ldots, X_n$ i.i.d. mit $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ und σ^2 sei bekannt. Für das Testproblem $H_0 : \mu \leq 0$ gegen $H_1 : \mu > 0$ verwenden wir den UMVUE-Schätzer $T(X) := \bar{X}$. Ist $\bar{X}$ zu groß, so spricht das für H_1 und gegen H_0. Somit erhalten wir einen sinnvollen Test durch $\delta_c(\boldsymbol{X}) := \mathbb{1}_{\{\bar{X}\geq c\}}$. Dieser Test wird auch als *einseitiger Gauß-Test* bezeichnet. Er hat die Gütefunktion

$$\begin{aligned} G_c(\mu) &= P_\mu(\delta_c(\boldsymbol{X}) = 1) = P_\mu\Big(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq \frac{c-\mu}{\sigma/\sqrt{n}}\Big) \\ &= 1-\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right). \end{aligned} \tag{93}$$

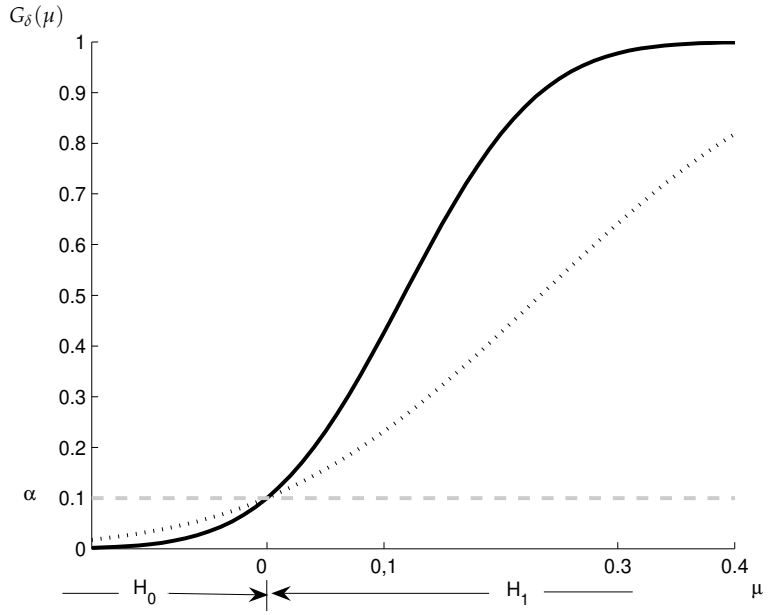


Abbildung 13: Gütefunktion des Tests $\delta(\boldsymbol{X}) = \mathbb{1}_{\{\bar{X} \geq \sigma z_{1-\alpha}/\sqrt{n}\}}$ für $H_0 : \mu \leq 0$ gegen $H_1 : \mu > 0$. Hierbei ist $\bar{X}$ normalverteilt mit bekannter Varianz σ^2. In der Darstellung wurde $\sigma = 0.5$ (gestrichelt) und $\sigma = 0.1$ (durchgezogene Linie) gewählt.

Demnach ist G_c monoton wachsend in μ. Da

$$\sup_{\mu \in \Theta_0} G_c(\mu) = 1 - \Phi\Big(\frac{c}{\sigma/\sqrt{n}}\Big) \leq \alpha$$

gelten muss, erhält man das kleinste c, welches das Signifikanzniveau α einhält durch $c_\alpha := \sigma/\sqrt{n} \cdot z_{1-\alpha}$. Der Test

$$\delta(\boldsymbol{X}) := \mathbb{1}_{\{\bar{X} \geq \frac{\sigma z_{1-\alpha}}{\sqrt{n}}\}} \tag{94}$$

ist somit der gesuchte Level-α-Test für das betrachtete Testproblem. Die entsprechende Gütefunktion ist in Abbildung 13 illustriert.

Der p-Wert: Die Teststatistik als Evidenz

Zur Durchführung eines Tests gehört immer die Wahl eines Signifikanzniveaus α. Diese Wahl hängt jedoch von der Problemstellung ab. Beim Testen eines Präzisionsinstrumentes wird man α sehr klein wählen, während bei statistischen Testproblemen die etwa auf einer Umfrage basieren ein größeres α sinnvoll ist. Um diese problemspezifische Wahl dem Anwender zu überlassen, führt man den p-Wert ein. Für die feste Beobachtung $\{\boldsymbol{X} = \boldsymbol{x}\}$ definiert man den *p-Wert* als kleinstes Signifikanzniveau, an welchem der Test die Null-Hypothese H_0 verwirft. Damit kann man H_0 stets verwerfen, falls man α gleich dem p-Wert wählt. Ein kleiner p-Wert kann als starke Evidenz gegen die Null-Hypothese interpretiert werden.

Beispiel 95 (Fortsetzung von Beispiel 92: p-Wert)**.** Das kleinste α, an welchem der Test unter der Beobachtung $\{\boldsymbol{X} = \boldsymbol{x}\}$ verwirft, erhält man wie folgt: Zunächst ist $\delta(\boldsymbol{x}) = 1$ nach Gleichung (94) äquivalent zu

$$\bar{x} \geq \frac{\sigma}{\sqrt{n}} z_{1-\alpha} = \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1-\alpha).$$

Löst man diese Gleichung nach α auf, so erhält man

$$p\text{-Wert}(\boldsymbol{x}) = 1 - \Phi\Big(\frac{\bar{x}}{\sigma/\sqrt{n}}\Big).$$

Offensichtlich übernimmt hier $\bar{x}$ die Rolle des vorherigen c.

Allgemeiner als in diesem Beispiel gilt falls $\boldsymbol{X}$ eine stetige Zufallsvariable ist:

Ist der Test von der Form $\delta_c(\boldsymbol{X}) = \mathbb{1}_{\{T(\boldsymbol{X}) \geq c\}}$, so ist

$$\gamma(c) := \sup_{\boldsymbol{\theta} \in \Theta_0} P_{\boldsymbol{\theta}}\big(T(\boldsymbol{X}) \geq c\big)$$

die Wahrscheinlichkeit für einen Fehler 1. Art. Der größte Wert c, für welchen man H_0 verwerfen kann, wenn $\{\boldsymbol{X} = \boldsymbol{x}\}$ beobachtet wurde, ist $T(\boldsymbol{x})$ und somit

$$p\text{-Wert}(\boldsymbol{x}) = \gamma(T(\boldsymbol{x})).$$

Dualität zwischen Konfidenzintervallen und Tests

Ein Konfidenzintervall ist ein zufälliger Bereich, der mit mindestens einer vorgegebenen Wahrscheinlichkeit den wahren Parameter überdeckt. Bei einem Test hingegen wird überprüft ob ein Wert von Interesse unter Einbezug einer gewissen Fehlerwahrscheinlichkeit mit den Daten in Einklang gebracht werden kann. Liegt etwa der Wert von Interesse in einem Konfidenzintervall, so würde man dies bejahen und man erhält aus einem Konfidenzintervall einen Test. Dies funktioniert auch umgekehrt und führt zu einer nützlichen Dualität zwischen Konfidenzintervallen und Tests. Wir beginnen mit einem Beispiel.

Beispiel 96 (Normalverteilung: Zweiseitiger Gauß-Test über den Erwartungswert)**.** Wir betrachten den Fall, dass eine Wissenschaftlerin eine physikalische Theorie untersucht. Bisher wurde angenommen, dass eine physikalische Konstante den Wert θ_0 hat. Die Wissenschaftlerin glaubt, dass diese These falsch ist und möchte

sie widerlegen. Dazu untersucht sie das zweiseitige Testproblem $H_0 : \theta = \theta_0$ gegen $H_1 : \theta \neq \theta_0$. Sie macht die (zu überprüfende) Annahme, dass $X_1, \ldots, X_n$ i.i.d. sind mit $X_1 \sim \mathcal{N}(\theta, \sigma^2)$. Weiterhin sei σ^2 bekannt. Ein Konfidenzintervall für θ wurde in Beispiel 81 bestimmt: $\bar{X} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$. Einen Test mit Signifikanzniveau α erhält man folgendermaßen aus diesem Konfidenzintervall: Die Annahme der Null-Hypothese $\theta = \theta_0$ sei gleichbedeutend damit, dass θ_0 in dem Konfidenzintervall liegt, also

$$\theta_0 \in \left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\,,\, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{97}$$

Mit $T(\boldsymbol{X}) := \sqrt{n}(\bar{X}-\theta_0)/\sigma$ ist (97) gleichbedeutend mit $|T(\boldsymbol{X})| \geq z_{1-\alpha/2}$, man erhält folgenden Test für $H_0 : \theta = \theta_0$ gegen $H_1 : \theta \neq \theta_0$:

$$\delta(\boldsymbol{X}, \theta_0) = \mathbb{1}_{\left\{\frac{|\sqrt{n}(\bar{X}-\theta_0)|}{\sigma} \geq z_{1-\alpha/2}\right\}}.$$

Dies ist in der Tat ein Test mit Signifikanzniveau α für jedes $\theta_0 \in \Theta$, denn

$$\begin{aligned} P_{\theta_0}(\delta(\boldsymbol{X}) = 1) &= 1 - P_{\theta_0}\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \theta_0 \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &\leq 1 - (1-\alpha) = \alpha \end{aligned}$$

da (97) ein $(1-\alpha)$-Konfidenzintervall war. Der durch δ gegebene Test ist ein *zweiseitiger Test*, weil er sowohl für kleine (und negative) als auch für große (und positive) Werte von T verwirft.

Dies kann man natürlich auch umgekehrt macchen und aus Tests Konfidenzintervalle konstruieren.

Kleiner Einschub: die χ^2-Verteilung und Tests über σ

Wir wiederholen zwei Verteilungen, die wir bereits aus der Stochastik 1 kennen: die χ^2- und die t-Verteilung. Ziel ist zweierlei: Zunächst möchten wir ein Konfidenzintervale für den Schätzer der Varianz angeben. Den MLS von σ haben wir bereits in Gleichung (**??**) bestimmt. Er besteht aus einer Summe von Quadraten. Die zu quadrierenden Zufallsvariablen selbst sind normalverteilt, was die folgende Verteilung motiviert.

Die χ^2-Verteilung ist eine Summe von unabhängigen, Standardnormalverteilten Zufallsvariablen.

Die χ^2-Verteilung entsteht als Summe von quadrierten, normalverteilten Zufallsvariablen.

Definition 98. Sind $X_1, \dots, X_n$ unabhängig und standardnormalverteilt, heißt

$$V := \sum_{i=1}^{n} X_i^2$$

χ^2-verteilt mit n Freiheitsgraden, kurz χ^2_n-verteilt. D

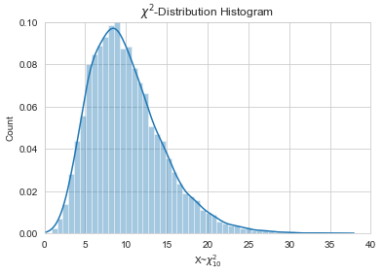


Abbildung 14: χ^2-Verteilung.

Hierbei verwenden wir die *Gamma-Funktion*, definiert durch

$$\Gamma(a) := \int_0^\infty t^{a-1} e^{-t} dt, \qquad a > 0.$$

Dann ist $\Gamma(n) = (n-1)!$, $n \in \mathbb{N}$ und $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Weiterhin gilt $E(V) = n$ und $\mathrm{Var}(V) = 2n$.

Die Dichte von V kann man mit Hilfe von Induktion ausrechnen und erhält

$$p_{\chi^2_n}(x) = \mathbb{1}_{\{x>0\}} \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}. \tag{99}$$

Als Übung hierzu kann mam einmal die Dichte von X_1^2 ausrechnen.

Bemerkung 100. Die Darstellung der Dichte in (99) zeigt, dass die χ^2_n-verteilte Zufallsvariable V für $n = 2$ exponentialverteilt ist mit Parameter $\frac{1}{2}$. Aus dem zentralen Grenzwertsatz folgt, dass

$$\frac{\chi^2_n - n}{\sqrt{2n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

Möchte man ein Konfidenzintervall für den Mittelwert einer Normalverteilung mit unbekannter Varianz bilden, so muss man diese schätzen. Dabei taucht die Wurzel einer Summe von Normalverteilungsquadraten (mit Faktor $\frac{1}{n}$) im Nenner auf. Hierüber gelangt man zur t-Verteilung, welche oft auch als Student-Verteilung oder Studentsche t-Verteilung bezeichnet wird.

Definition 101. Ist X standardnormalverteilt und V χ^2_n-verteilt und unabhängig von X, so heißt die Verteilung von

$$T := \frac{X}{\sqrt{\frac{1}{n}V}} \tag{102}$$

die *t-Verteilung* mit n Freiheitsgraden, kurz t_n-Verteilung.

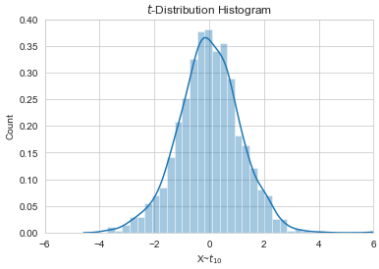


Abbildung 15: t-Verteilung.

Auch die Dichte der t_n-Verteilung kann man berechnen:

$$p_{t_n}(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(n/2)\Gamma(1/2)\sqrt{n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

für alle $x \in \mathbb{R}$.

Unbekannte Varianz

Das letzte Kapitel in diesem Abschnitt soll sich mit dem Fall beschäftigen in welchem σ unbekannt ist. Dies ist natürlich der für die Praxis wichtigste Fall! Allerdings ist er auch deutlich komplizierter.

Beispiel 103 (Normalverteilung: Konfidenzintervall für σ^2)**.** Wieder seien $X_1, \ldots, X_n$ i.i.d. mit $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. Der Mittelwert μ sei nun bekannt. In diesem Fall ist

$$\tilde{\sigma}^2(\boldsymbol{X}) := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

der Maximum-Likelihood- und UMVUE-Schätzer für σ^2. Ein Pivot ist leicht gefunden, da

$$\frac{n\tilde{\sigma}^2(\boldsymbol{X})}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n.$$

Sei $\chi^2_{n,a}$ das a-Quantil der χ^2_n-Verteilung. Durch die Beobachtung, dass

$$P\left(\chi^2_{n,\alpha/2} \leq \frac{n\tilde{\sigma}^2(\boldsymbol{X})}{\sigma^2} \leq \chi^2_{n,1-\alpha/2}\right) = 1 - \alpha$$

erhält man ein $(1-\alpha)$-Konfidenzintervall für σ^2 gegeben durch

$$\left[\frac{n\tilde{\sigma}^2(\boldsymbol{X})}{\chi^2_{n,1-\alpha/2}}, \frac{n\tilde{\sigma}^2(\boldsymbol{X})}{\chi^2_{n,\alpha/2}}\right].$$

Allerdings handelt es sich hier nicht um ein unverzerrtes Konfidenzintervall. Weiterhin ist es nicht symmetrisch um $\tilde{\sigma}^2(\boldsymbol{X})$.

Beispiel 104 (Normalverteilung — Konfidenzintervall)**.** Die Zufallsvariablen $X_1, \dots, X_n$ seien i.i.d. mit $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. Gesucht ist ein Konfidenzintervall für den Mittelwert μ, aber auch σ ist unbekannt. Wie bisher bezeichne

$$s_n^2 = s_n^2(\boldsymbol{X}) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

die Stichprobenvarianz und weiterhin sei $c := t_{n-1,1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden. Man erhält mit $\boldsymbol{\theta} := (\mu, \sigma^2)^\top$, dass

$$P_{\boldsymbol{\theta}}\left(\bar{X} - \frac{cs_n}{\sqrt{n}} \le \mu \le \bar{X} + \frac{cs_n}{\sqrt{n}}\right) = P_{\boldsymbol{\theta}}\left(\left|\frac{\bar{X}-\mu}{s_n/\sqrt{n}}\right| \le c\right).$$

Nach Satz 107 folgt, dass $\bar{X}$ von $s_n^2(\boldsymbol{X})$ unabhängig ist und $(n-1)\frac{s_n^2(\boldsymbol{X})}{\sigma^2} \sim \chi^2_{n-1}$. Wir erhalten nach Definition 101, dass

$$T_{n-1}(\boldsymbol{X}) := \frac{\sqrt{n}(\bar{X}-\mu)}{s_n(\boldsymbol{X})} = \frac{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{(n-1)s_n^2(\boldsymbol{X})}{\sigma^2}}}$$

t_{n-1}-verteilt ist. Da diese Verteilung unabhängig von $\boldsymbol{\theta}$ ist, ist T_{n-1} ein Pivot. Somit ergibt sich folgendes Konfidenzintervall für μ:

$$\bar{X} \pm \frac{s_n}{\sqrt{n}}\, t_{n-1,1-\alpha/2}.$$

Nun betrachten wir noch kurz den mehrdimensionalen Fall, in welchem ein Konfidenzintervall für die vektorwertige Transformation $\boldsymbol{q}(\boldsymbol{\theta}) = (q_1(\boldsymbol{\theta}), \dots, q_n(\boldsymbol{\theta}))^\top$ bestimmt werden soll. Analog zum eindimensionalen Fall definieren wir:

Definition 105. Das durch $\underline{T_j}(\boldsymbol{X}) \le \overline{T_j}(\boldsymbol{X}), 1 \le j \le n$ gegebene Zufallsrechteck

$$I(\boldsymbol{X}) := \left\{\boldsymbol{x} \in \mathbb{R}^n : \underline{T_j}(\boldsymbol{X}) \le x_j \le \overline{T}_j(\boldsymbol{X}), j = 1, \dots, n\right\}$$

heißt $(1-\alpha)$-*Konfidenzbereich* für $q(\boldsymbol{\theta})$, falls für alle $\boldsymbol{\theta} \in \Theta$

$$P_{\boldsymbol{\theta}}\big(\boldsymbol{q}(\boldsymbol{\theta}) \in I(\boldsymbol{X})\big) \ge 1 - \alpha.$$

Man kann die für den eindimensionalen Fall erhaltenen Konfidenzintervalle unter gewissen Umständen auf den n-dimensionalen Fall übertragen. Allerdings erhält man dann ein anderes, deutlich schlechteres Konfidenzniveau.

(i) Falls $I_j(\boldsymbol{X}) := \left[\underline{T}_j(\boldsymbol{X}), \overline{T}_j(\boldsymbol{X})\right]$ jeweils $(1-\alpha_j)$-Konfidenzintervall für $q_j(\boldsymbol{\theta})$ ist und falls $(\underline{T}_1, \overline{T}_1), \ldots, (\underline{T}_n, \overline{T}_n)$ unabhängig sind, so ist

$$I(\boldsymbol{X}) := I_1(\boldsymbol{X}) \times \cdots \times \ I_r(\boldsymbol{X})$$

ein $\prod_{j=1}^{n}(1-\alpha_j)$-Konfidenzbereich für $\boldsymbol{q}(\boldsymbol{\theta})$. Mit $\alpha_j = \sqrt[n]{1-\alpha}$ erhält man so einen $(1-\alpha)$-Konfidenzbereich.

(ii) Falls die I_j nicht unabhängig sind, so kann man die *Bonferroni Ungleichung*[17] verwenden, und erhält daraus, dass jedes Intervall I_j das Konfidenzniveau α_j einhält, dass

$$P_{\boldsymbol{\theta}}(\boldsymbol{q}(\boldsymbol{\theta}) \in I(\boldsymbol{X})) \quad \geq \quad 1 - \sum_{j=1}^{n} P_\theta(q_j(\boldsymbol{\theta}) \notin I_j(\boldsymbol{X})) \geq 1 - \sum_{j=1}^{n} \alpha_j.$$

Dann ist $I(\boldsymbol{X})$ ein $(1-\alpha)$-Konfidenzbereich, falls man $\alpha_j = \alpha/n$ wählt.

[17] Die Bonferroni Ungleichung lautet $P(A \cap B) \geq 1 - (P(\bar{A}) + P(\bar{B}))$ für alle $A, B \in \mathcal{A}$.

Beispiel 106 (Normalverteilungsfall: Konfidenzbereich für (μ, σ^2))**.** Wir übertragen die eindimensionalen Konfidenzintervalle wobei wir das Konfidenzintervall für σ^2 mit dem Faktor $n-1$ statt n multiplizieren um Unverzerrtheit zu erhalten: Seien $X_1, \ldots, X_n$ i.i.d. mit $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. Dann ist

$$I_1(\boldsymbol{X}) := \bar{X} \pm \frac{s(\boldsymbol{X})}{\sqrt{n}} t_{n-1,1-\alpha/4}$$

ein $(1-\alpha/2)$-Konfidenzintervall für μ, wenn σ^2 unbekannt ist und

$$I_2(\boldsymbol{X}) := \left[\frac{(n-1)s^2(\boldsymbol{X})}{\chi^2_{n-1,1-\alpha/4}}, \frac{(n-1)s^2(\boldsymbol{X})}{\chi^2_{n-1,\alpha/4}}\right]$$

ein $(1-\alpha/2)$-Konfidenzintervall für σ^2, wenn μ unbekannt ist. Man erhält den *gemeinsamen Konfidenzbereich* für (μ, σ^2) durch $I_1(\boldsymbol{X}) \times I_2(\boldsymbol{X})$ mit Konfidenzniveau $1 - \left(\frac{\alpha}{2} + \frac{\alpha}{2}\right) = 1 - \alpha$.

Mit einigem Aufwand lässt sich zeigen, dass $\bar{X}$ und $s^2(\boldsymbol{X})$ unabhängig sind. Etwas überraschend lässt sich mit Hilfen der linearen Regression dieser Beweis durchaus einfacher führen (siehe Czado & Schmidt[18], Satz 7.14 - der Satz wird hier nur zur Illustration angegeben).

[18] C. Czado and T. Schmidt. *Mathematische Statistik*. Springer Verlag. Berlin Heidelberg New York, 2011

Satz 107. *Sei* $\widehat{\boldsymbol{\zeta}}(\boldsymbol{Y}) := X\widehat{\boldsymbol{\beta}}(\boldsymbol{Y})$ *und* $s^2(\boldsymbol{Y}) := \frac{1}{n-r} \parallel \boldsymbol{Y} - \widehat{\boldsymbol{\zeta}} \parallel^2$. *Dann gilt im allgemeinen linearen Modell:*

(i) $\widehat{\boldsymbol{\zeta}}$ *und* $\boldsymbol{Y} - \widehat{\boldsymbol{\zeta}}$ *sind unabhängig.*

(ii) $(n-r)\frac{s^2(\boldsymbol{Y})}{\sigma^2} \sim \chi^2_{n-r}$ *und ist unabhängig von* $\widehat{\boldsymbol{\zeta}}$.

Beweis. Zunächst ist nach Definition $\widehat{\zeta} = \sum_{i=1}^{r} Z_i v_i$. Es folgt, dass

$$Y - \widehat{\zeta} = \sum_{i=r+1}^{n} Z_i v_i.$$

Da $Z_1, \dots, Z_n$ unabhängig sind (!) folgt Behauptung (i).

Somit ist auch $(n-r)s^2 = \sum_{i=r+1}^{n} Z_i^2$ unabhängig von $\widehat{\zeta}$. Die Zufallsvariablen $Z_{r+1}, \dots, Z_n$ sind i.i.d. mit $Z_i \sim \mathcal{N}(0, \sigma^2)$ für $i = r+1, \dots, n$ und somit gilt, dass

$$\frac{(n-r)s^2}{\sigma^2} = \sum_{i=r+1}^{n} \left(\frac{Z_i}{\sigma}\right)^2 \sim \chi^2_{n-r}.$$

□

Ein kurzer Ausflug in die Finanzmathematik

Am Markt für Aktien werden nicht nur Aktion gehandelt, sondern auch derivative Instrumente die eine zentrale Rolle in der Absicherung von Investitionsstrategien bilden.

Definition 1. Eine *europäische Call (Put) Option* ist das *Recht*, das Basisgut an $T \geq t$ zum Preis K zu kaufen (zu verkaufen). Eine *amerikanische Call (Put) Option* ist das Recht, das Basisgut an jedem Zeitpunkt bis T zum Preis K zu kaufen (zu verkaufen).

Wir nennen K den *Ausübungspreis* (Strike) und $T - t$ *Restlaufzeit*. Man beachte, dass im Gegensatz zu einem Termingeschäft nicht Verpflichtung besteht, das Basisgut zu kaufen oder zu verkaufen. Den Preisprozess des Basisgutes bezeichnen wir mit $S = (S_t)_{t \geq 0}$.

Wir betrachten zunächst die Call Option. Ein rationaler Investor wird das Optionsrecht nur ausüben, falls $S_T > K$ (anderenfalls kann er die Aktie billiger am Markt kaufen); in diesem Fall erzielt er einen Gewinn in Höhe von $S_T - K$. So ergibt sich der Wert der Call Option an T als

$$C(T) = \max\{S_T - K, 0\} =: (S_T - K)^+. \tag{2}$$

Ganz analog ergibt sich für den Endwert der Put Option

$$P(T) = \max\{K - S_T, 0\} =: (K - S_T)^+.$$

Beide Funktionen werden in Abbildung 16 illustriert.

Beispiel 3 (Absichern eines Aktiendepots mit Put Optionen)**.** Eine Anlegerin hält (an $t = 0$) 10 Akien im Depot mit Kurs S_0. Sie möchte vermeiden, dass der Wert der Aktienposition an $T = 1$ unter den Ausganswert $A := 10\, S_0$ fällt. Hierzu geht sie wie folgt vor: Sie kauft 10 europäische Put Optionen auf die Aktie mit Ausübungspreis

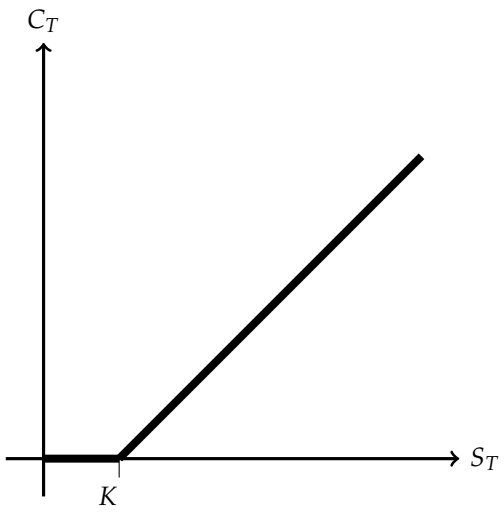


Abbildung 16: Auszahlungsschema eins Calls - wie sieht das eines Puts aus?

$K = S_0$. Dies ergibt an $T = 1$ den Wert

$$\begin{aligned} &10\,(S_1 + 0) && \text{falls } S_1 > S_0 \\ &10\,(S_1 + (S_0 - S_1)) = 10\,S_0 && \text{falls } S_1 < S_0. \end{aligned}$$

Somit hat das Portfolio mindestens den Wert A. Allerdings ist zu Beginn die Zahlung der Optionsprämie erforderlich.

Das Modell mit einer Periode

Für unser Modell betrachten wir einen Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, P)$. Wir gehen zunächst von zwei Zeitpunkten aus, welche wir mit 0 und T bezeichnen. Der Handel von Wertpapieren findet zu Beginn der Periode, also an $t = 0$ statt. Die Wertpapiere werden bis zum Zeipunkt T gehalten und dann verkauft. Die entstehenden Gewinne bzw. Verluste sind zufällig.

Wir werden vektorwertige Zufallsvariablen an den Zeitpunkten 0 und T betrachten und schreiben deswegen $S = (S^1, \dots, S^d)$ für einen Vektor S; S_0 und S_T sind die Werte eines stochastischen Prozesses an 0 und T, so dass jeweils

$$S_t = (S_t^1, \dots, S_t^d)$$

gilt. Für Vektoren, die nicht die Werte von stochastischen Prozessen sind schreiben klassisch $(x_1, \dots, x_k)$.

Marktteilnehmer können zur Zeit 0 Wertpapiere kaufen und an T verkaufen. Neben den Wertpapieren gibt es ein risikoloses Bankkonto, welches wir mit S^0 bezeichnen. Wir nehmen an, dass $S_0^0 = 1$, also dass das Bankkonto mit dem Wert 1 startet, und dass

$S_1^0 > 1$ ist. Typischerweise haben wir einen Zins r, so dass

$$S_1^0 = 1 + r$$

ist. Allerdings müssen wir nicht annehmen, dass $r > 0$ ist, Inflation oder negative Zinsen sind demnach zugelassen.

Wir haben nun immer den d-dimensionalen Vektor $S = (S^1, \dots, S^d)$ und den $d+1$-dimensionalen Vektor $\bar{S} = (S^0, \dots, S^d)$. Dies werden wir stets in der Notation kenntlich machen (also ohne Strich für die Wertpapiere $S^1, \dots, S^d$ und mit Strich für das Bankkonto und die Wertpapiere $S^0, \dots, S^d$.)

Zur Zeit 0 entscheidet der Investor, welche Wertpapiere sie kaufen will. Auch negative Positionen sind erlaubt, sogenannte Leerverkäufe (short-selling), d.h. ein Marktteilnehmer hat die Möglichkeit in Wertpapieren negative Positionen zu beziehen. Ebenso kann sie beliebig stückeln und beispielsweise $1/3$ Aktien kaufen. Formal beschreiben wir die Position eines Investors durch einen Vektor

$$\bar{H} = (H^0, \dots, H^d) \in \mathbb{R}^{d+1}.$$

Dabei ist H^i die Anzahl der i-ten Wertpapiere im Portfolio. Ist $H^i < 0$ spricht man von einer *Short Position* in Wertpapier i, im Fall $H^i > 0$ entsprechend von einer *Long Position*. Bei $i = 0$, also dem Bankkonto entspricht H^0 der Einlage in das Bankkonto (falls $H^0 > 0$) oder dem aufgenommenen Geld (falls $H^0 < 0$).

Die *Auszahlung* oder der Wert des Portfolios H ist wieder eine Zufallsvariable, welche wir mit $W_T = W_T^H$ bezeichnen. Der Wert des Portfolios an T ist gegeben durch die Summe der einzelnen Auszahlungen, also

$$W_T^H = \bar{H} \cdot \bar{S}_T = \sum_{i=0}^{d} H^i S_T^i.$$

Definition 4. Ein *bedingter Anspruch* ist eine Zufallsvariable ξ. Sie heißt *erreichbar*, falls ein Portfolio $\bar{H}$ existiert, so dass

$$\xi = W_T^H.$$

Oft sagen wir auch *Europäische Option* für die Zufallsvariable ξ. Das Portfolio H heißt *Replikationsportfolio* von ξ. Im Englischen spricht man bei einer bedingten Auszahlung von einem *Contingent Claim*. Ein erreichbarer bedingter Anspruch lässt sich also exakt durch ein Portfolio von Wertpapieren nachbilden. Das ist zum Beispiel für jedes gehandelte Wertpapier selbst der Fall.

Beispiel 5 (Binomialmodell)**.** Wir betrachten ein Modell mit zwei Wertpapieren, Nullkuponanleihe und Aktie. Es gebe *zwei* Möglichkeiten für den Aktienkurs in T, und zwar 120 und 180, was wir durch $\Omega = \{\omega_1, \omega_2\}$ modellieren. Das Bankkonto hat stets den Wert 1.

Es soll eine Call Option auf die Aktie mit Ausübungspreis $K = 150$ und Fälligkeit T bewertet werden. Dies ist der bedingte Anspruch ξ mit

$$\xi = \begin{cases} 30 & \text{für } \omega_1 \\ 0 & \text{sonst.} \end{cases}$$

Die Auszahlung des Calls ist erreichbar, falls das lineare Gleichungssystem

$$\begin{array}{rcrcl} H^0 & + & 180\, H^1 & = & 30 \\ H^0 & + & 120\, H^1 & = & 0 \end{array}$$

eine Lösung hat. Dies ist der Fall für $H^0 = -60,\ H^1 = 1/2$. Wir erhalten in der Tat ein Replikationsportfolio durch eine Short Position von 60 Nullkuponanleihen und eine Long Position von 0.5 Aktien.

Beispiel 6 (Trinomialmodell)**.** Nun betrachten wir *drei* Möglichkeiten für den Wert der Aktie an T: 120, 150 und 180. Wir müssen nun das Gleichungssystem

$$H^0 + 180\, H^1 = 30 \tag{7}$$
$$H^0 + 150\, H^1 = 0 \tag{8}$$
$$H^0 + 120\, H^1 = 0. \tag{9}$$

lösen. Aus dem vorigen Beispiel wissen wir, dass (7) und (9) auf $H^0 = -60, H^1 = 1/2$ führen. Setzen wir diese Werte in (8) ein, so erhalten wir $-60 + 1/2 \cdot 150 = -60 + 75 \neq 0$. Der Call ist also *nicht* erreichbar.

Arbitragefreiheit

In diesem Abschnitt wird der Begriff Arbitrage präzise definiert und Kriterien bestimmt, die einen arbitragefreien Markt charakterisieren.

Arbitragefreiheit und Martingalmaße

Für den Kauf von Wertpapieren zur Zeit 0 ist ein jeweils ein Preis zu zahlen. Wie bereits dargestellt, ist für das i-te Wertpapier der Preis S_0^i zu zahlen, so dass der Preis eines Portfolios H an 0

$$W_0^H = \bar{H} \cdot \bar{S}_0 \tag{10}$$

ist.

Definition 11. Eine *Arbitragemöglichkeit* ist ein Portfolio $\bar{H}$, so dass

(i) $W_0^H \leq 0$,
(ii) $W_T^H \geq 0$,
(iii) entweder ist $W_0^H < 0$, oder $P(W_T^H > 0) > 0$.

Ein Markt heißt *arbitragefrei*, wenn es keine Arbitragemöglichkeit gibt.

Eine Arbitragemöglichkeit ist demnach ein Portfolio, für welches zunächst der Preis Null oder negativ und die Auszahlung Null oder positiv ist. Allerdings ist entweder der Kaufpreis echt verschieden von Null oder der Wert an T mit positiver Wahrscheinlichkeit echt positiv.

Denken Sie hierbei an das Binomialmodell: $\Omega = \{\omega_1, \omega_2\}$ mit der Voraussetzung $P(\omega_i) > 0$, $i = 1, 2$. Dann ist Q äquivalent zu P genau dann wenn, $Q(\omega_i) > 0$. Damit ist das Modell auch ein Binomialmodell unter Q (denn etwa $Q(\omega_1) = 0$ ist nicht erlaubt).

Ein Wahrscheinlichkeitsmaß Q heißt *äquivalent* zu P falls $Q(A) = 0$ genau dann gilt, wenn $P(A) = 0$ für alle $A \in \mathscr{F}$. Mit $E_Q[\cdot] = \int \cdot dQ$ bezeichen wir den Erwartungswert unter dem Wahrscheinlichkeitsmaß Q.

Definition 12. Ein zu P äquivalentes Wahrscheinlichkeitsmaß Q heißt *risikoneutral*, falls für jede erreichbare Auszahlung $X = W_T^H$ gilt, dass

$$W_0^H = \frac{1}{S_T^0} E_Q[W_T^H]. \tag{13}$$

Die Bewertungsregel (13) heißt *risikoneutrale Bewertungsregel*. Das risikoneutrale Wahrscheinlichkeitsmaß wird oft auch als *Martingalmaß* bezeichnet, da die abdiskontierten Auszahlungen der gehandelten Wertpapiere unter Q Martingale sind, was wir im Mehrperiodenmodell noch zeigen werden. Die Wahrscheinlichkeiten $Q(\cdot)$ sind durch die Marktstruktur (S_0 und S_T) festgelegt und sind typischerweise verschieden von den realen Eintrittswahrscheinlichkeiten der Zustände, $P(\cdot)$. Der Name risikoneutral kommt daher, dass diese Wahrscheinlichkeiten gerade den Erwartungen eines risikoneutralen Investors entsprechen, die mit dem Preissystem S in Einklang stehen.

Lemma 14. *Ein zu P äquivalentes Maß Q ist genau dann risikoneutral, falls*

$$S_0^i = (S_T^0)^{-1} E_Q[S_T^i], \qquad 1 \leq i \leq d. \tag{15}$$

Beweis. Zunächst folgt aus (15), dass

$$W_0^H = \bar{H} \cdot \bar{S}_0 = (S_T^0)^{-1} \bar{H} \cdot E_Q[\bar{S}_T] = (S_T^0)^{-1} E_Q[W_T^H]$$

und Q ist risikoneutrales Maß. Umgekehrt erhalten wir aus (13) direkt (15) wenn für H die Einheitsvektoren auf dem $\mathbb{R}^d$ gewählt werden. □

Das folgende Resultat ist der *1. Hauptsatz der Wertpapierbewertung*. Wie zuvor betrachten wir einen Markt mit risikolosem Wertpapier (S, r).

Theorem 16. *Der Markt ist genau dann arbitragefrei, falls ein risikoneutrales Maß Q existiert.*

Wir betrachten einen allgemeineren Satz in mehreren Perioden, wo wir zumindest eine Richtung dieses Satzes beweisen.

Mehrere Perioden

Wir betrachten nun die Perioden $t = 0, \dots, T$. In mehreren Perioden gibt es einen *Informationsfluss*: an $t_2 > t_1$ steht mehr Information zur Verfügung, etwa neue Aktienkurse, etc. Dies modelliert man mit einer Filtration: Eine Filtration ist eine wachsende Folge von Sub-σ-Algebren $\mathbb{F} = (\mathscr{F}_t)_{t=0,\dots,T}$, d.h.

$$\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}_T \subseteq \mathscr{F}.$$

Der Finanzmarkt wird modelliert durch einen $d+1$-dimensionalen adaptierten stochastischen Prozess $S = (S_t)_{t=,\dots,T}$. *Adaptiert* heißt hierbei, dass S_t $\mathscr{F}_t$-messbar ist für $t = 0, \dots, T$. Diese Bedingung bedeutet, dass an Zeitpunkt t alle Preise S_t^i bekannt sind (und nicht etwa von nicht vorhandenen, zukünftigen Infomrationen abhängen).

Die *Handelsstrategie* $\bar{H}$ ist ein besonderer Prozess, der nicht nur adaptiert, sonderen sogar *vorhersehbar* ist. Hierzu stellen wir uns vor, dass H_t^i die Anzahl der i-ten Aktie ist, die von Beginn der Periode $(t-1, t]$ bis zum Ende der Periode gehalten werden. Diese Anzahl muss zum Zeitpunkt $t-1$ bekannt sein, also $\mathscr{F}_{t-1}$-messbar. Genau da nennt man vorhersehbar: Ein stochastischer Process $(X_t)_{t=1,\dots,T}$ heißt *vorhersehbar*, falls X_t $\mathscr{F}_{t-1}$-messbar ist für $t = 1, \dots, T$.

In mehreren Perioden gibt es noch einen weiteren wichtigen Begriff für die Handelsstrategie: An t wird die Handelsstrategie $\bar{H}_t$ liquidiert und direkt in die neue Handelsstrategie $\bar{H}_{t+1}$ übergeführt. Dabei darf kein Geld verlorengehen, was in der folgenden Bedingung festgesetzt wird.

Definition 17. Eine Handelsstrategie $\bar{H}$ heißt *selbstfinanzierend*, falls

$$\bar{H}_t \cdot \bar{S}_t = \bar{H}_{t+1} \cdot \bar{S}_t.$$

Wie bereits in einer Periode angedeutet, gibt es ein risikoloses Wertpapier (Bankkonto) S^0, welches die Wertenwicklung des Geldes beinhaltet. Insbesondere ist ein Euro in einem Jahr natürlich weniger wert als ein Euro heute (vorausgesetzt, die Zinsen sind positiv) – dies bildet man durch *diskontieren* an: $(S_t^0)^{-1}$ ist der Wert eines Euros zur Zeit t betrachtet von der heutigen Warte $t = 0$ aus. Denn, durch investieren von $(S_t^0)^{-1}$ in S^0 erreiche ich an t genau

$$(S_t^0)^{-1} S_t^0 = 1.$$

Wir führen demzufolge den *diskontierten Wertprozess* ein. Dieser nimmt also die Wertentwicklung des Bargeldes aus der Entwicklung des Aktienkurses heraus, betrachtet also die Wertentwicklung relative zu S^0. Insbesondere lässt sich somit leicht feststellen, ob sich eine Aktie besser oder schlecher entwickelt hat als S^0. (Wie ?)

Wir definieren den diskontierten Wertprozess X durch

$$X_t^i = \frac{S_t^i}{S_t^0}, \qquad t = 0, \dots, T,\ i = 0, \dots, d.$$

Hierbei ist das diskontierte Bankkonto X^0 ist immer genau 1.

Wir führen ebenso den diskontierten Wertprozess $V = V^H$ einer Handelsstrategie H ein durch $V_0 := \bar{H}_1 \cdot X_0$ und

$$V_t := \bar{H}_t \cdot \bar{X}_t.$$

Ein besonders nützliches Hilfsmittel ist der zugehörige Gewinnprozess $G = G^H$ einer selbstfinanzierenden Handelsstrategie, gegeben durch $G_0 = 0$ und

$$G_t = \sum_{s=1}^{t} H_s \cdot (X_s - X_{s-1}) =: \sum_{s=1}^{t} H_s \cdot \Delta X_s.$$

Mit ΔX_s bezeichnen wir den (diskontierten) Gewinn/Verlust in dem (diskontierten) Wertpapierportfolio X.

Man beachte, dass man bei G die Striche nun weglassen kann, denn $\Delta X^0 = 1 - 1 = 0$! Somit dürfen wir ab diesem Zeitpunkt mit dem d-dimensionalen Prozess X rechnen.

Nun können wir auch in mehreren Perioden den Begriff Arbitrage ganz präzise einführen. Beachten Sie bitte die Unterschiede zu der Definition in einer Periode (und zeigen Sie, dass die Begriffe übereinstimmen).

Definition 18. Eine *Arbitrage* ist eine selbstfinanzierende Handelsstrategie H, so dass

(i) $V_0^H \leq 0$,
(ii) $V_T^H \geq 0$,
(iii) $P(V_T^H > 0) > 0$.

Besonders nützlich ist folgendes Kriterium: Genau dann gibt es eine Arbitrage, wenn es eine einzelne Zeitperiode $(t-1, t]$ gibt, in der es eine Arbitrage gibt. In unserem Setting ist es also nicht möglich eine Arbitrage über zwei Perioden spannen. Dies liegt an der Adaptiertheit der Preisprozesse. Wäre S nicht adaptiert, kann man auch eine Zwei-Perioden Arbitrage konstruieren. Dieses Setting ist allerdings bis heute Gegenstand aktueller Forschung und noch nicht komplett verstanden.

Satz 19. *Der Markt ist frei von Arbitrage, genau dann, falls ein $t \in \{1, \dots, T\}$ und ein $\eta \in \mathscr{F}_{t-1}$ existiert, so dass*

$$\eta \cdot \Delta X_t \geq 0 \qquad P(\eta \cdot \Delta X_t > 0) > 0.$$

Martingale haben wir schon in der Stochastik 1 kennengelernt - in der Vorlesung gab es einen kleinen Exkurs zu den Rechenregeln hierzu, der im Skriptum erst etwas später hinzugefügt werden kann.

Definition 20. Ein *Martingal* ist ein adaptierter Prozess M, so dass

(i) $E[|M_t|] < \infty$, für $t = 0, \dots, T$, und
(ii) $E[M_t|\mathscr{F}_{t-1}] = M_{t-1}$ für $t = 1, \dots, T$.

Wir nennen ein Maß Q nun *Martingalmaß*, falls der diskontierte Preisprozess X ein Martingal unter Q ist, d.h. $E_Q[X_t^i|\mathscr{F}_{t-1}] = X_{t-1}^i$ für alle t und i.

Theorem 21. *Der Markt ist frei von Arbitrage genau dann, wenn ein äquivalentes Martingalmaß existiert.*

Beweis. Nach Satz 19 dürfen wir uns auf eine Periode beschränken. Sei also η so wie in diesem Satz angegeben. Dann ist $P(\eta \Delta X_t > 0) > 0$. Da Q zu P äquivalent ist, folgt aber sofort

$$Q(\eta \Delta X_t > 0) > 0.$$

Nun ist aber Q ein Martingalmaß, d.h.

$$\begin{aligned} E_Q[\eta \Delta X_t | \mathscr{F}_{t-1}] &= \eta E_Q[\Delta X_t | \mathscr{F}_{t-1}] \\ &= \eta \cdot \Big(E_Q[X_t | \mathscr{F}_{t-1}] - X_{t-1} \Big) = 0, \end{aligned}$$

da ja $E_Q[X_t|\mathscr{F}_{t-1}] = X_{t-1}$. Dies ist aber ein Widerspruch zu $\eta \Delta X_t \geq 0$ und $Q(\eta \Delta X_t > 0) > 0$.

Die Rückrichtung ist deutlich schwieriger zu beweisen, siehe etwa [19]. □

[19] H. Föllmer and A. Schied. *Stochastic Finance*. Walter de Gruyter, Berlin, 2011

Wir erhalten ebenfalls: Füge ich einem Markt ein Wertpapier X^{d+1} hinzu, so dass der diskontierte Preisprozess X^{d+1} ein Q-Martingal ist, so bleibt der Markt arbitragefrei. Dies begründet die wichtige *risikoneutrale Bewertungsformel*

$$X_t^{d+1} = E_Q\left[\frac{X_T}{S_T^0} | \mathscr{F}_t \right],$$

$t = 0, \ldots, T$, womit ein Derivat arbitragefrei bewertet werden kann.

Hedging und Replikation

Das Bewerten eines Derivats ist allerdings erst der erste Schritt - die Bank muss die Auszahlung durch geschicktes Handeln erreichen. Dies nennt man *Replikation*, *Absichern* oder *Hedgen*.

Definition 22. Eine europäische Option ξ heißt *erreichbar*, falls eine selbstfinanzierende Handelsstrategie $\bar{H}$ existiert, so dass

$$\xi = \bar{W}_T^H \cdot \bar{S}_T.$$

Man kann relativ leicht zeigen, dass eine europäische Option genau dann erreichbar ist, falls für ihre diskontierte Auszahlung gilt:

$$\frac{\xi}{S_T^0} = V_T^H \cdot X_t,$$

so dass wir auch hier wieder direkt mit diskontierten Größen arbeiten können.

Für eine erreichbare europäische Option gibt es einen eindeutigen Preis (ansonsten gibt es Arbitrage) - für eine nicht erreichbare Option muss das nicht gelten.

Satz 23. *Für eine erreichbare europäische Option ξ gilt $E[|\xi(S_T^0)^{-1}|] < \infty$ und für alle äquivalenten Martingalmaße Q ist*

$$V_t = E_Q[\frac{\xi}{S_T^0}|\mathscr{F}_{t-1}].$$

Beweis. Den ersten Teil findet man in Theorem 5.14 und 5.25 in [20].

Weiterhin gilt

$$E_Q[V_T - V_t|\mathscr{F}_t] = V_t + \sum_{s=t+1}^{T} E_Q\Big[H_s \cdot \Delta X_s|\mathscr{F}_t] = V_t.$$

□

[20] H. Föllmer and A. Schied. *Stochastic Finance*. Walter de Gruyter, Berlin, 2011

Beispiel 24 (Hedging Strategie im Binomialmodell)**.** Wir betrachten das ein-Perioden-Binomialmodell, also $T = 1$ und $\Omega = \{\omega_1, \omega_2\}$. Schauen wir uns ein beliebiges Derivat mit der Auszahlung $\xi_1(\omega_1)$ bzw. $\xi_1(\omega_2)$ an. Dann muss für eine Repliation gelten, dass

$$\xi_1 = \xi_0 + H_1 \cdot (S_1 - S_0).$$

Achten Sie darauf, welche Terme zufällig sind und welche nicht. Dies führt uns zu dem Gleichungssystem

$$\begin{aligned}\xi_1(\omega_1) &= \xi_0 + H_1 \cdot (S_1(\omega_1) - S_0)\\ \xi_1(\omega_2) &= \xi_0 + H_1 \cdot (S_1(\omega_) - S_0),\end{aligned}$$

und wir errechnen leicht den *Hedge*

$$H_0 = \frac{\xi_1(\omega_1) - \xi_1(\omega_2)}{S_1(\omega_1) - S_1(\omega_2)}. \tag{25}$$

Literaturverzeichnis

[1] Bericht über Verdachtsfälle von Nebenwirkungen und Impfkomplikationen nach Impfung zum Schutz vor COVID-19, 2021.

[2] Jakob Bernoulli. *Ars conjectandi: opus posthumum: accedit Tractatus de seriebus infinitis; et Epistola gallice scripta de ludo pilae reticularis*. Impensis Thurnisiorum, 1713.

[3] C. Czado and T. Schmidt. *Mathematische Statistik*. Springer Verlag. Berlin Heidelberg New York, 2011.

[4] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

[5] H. Föllmer and A. Schied. *Stochastic Finance*. Walter de Gruyter, Berlin, 2011.

[6] Andrei Kolmogorov. Über die analytischen Methoden in der Wahrscheinlichkeitstheorie. *Math Ann*, 104(1):415–458, 1931.

[7] Gabor Paal. Wenn nicht 5.000 Corona-Impftote, wie viele dann?, 2021.

[8] Thorsten Schmidt. Google Colab für Stochastik II. `https://colab.research.google.com/drive/1S6uhoYKY87f7XXhCXLSDag8_k0hFaJQA?authuser=1`, 2022.