Wahrscheinlichkeitstheorie, Statistik und Datenanalyse

Prof. Dr. Wolfgang von der Linden
DI Alexander Prüll

10. Dezember 2002

Inhaltsverzeichnis

WAHRSCHEINLICHKEITSTHEORIE, STATISTIK UND DATENANALYSE

THERE ARE THREE LIES: LIES, DAMNED LIES AND STATISTICS. (DISRAELI, MARK TWAIN)

Teil I

Einführung

Kapitel 1

Statistische und klassische Definition von „Wahrscheinlichkeit"

1.1 Klassische Definition

Die erste quantitative Definition des Begriffes „Wahrscheinlichkeit" geht auf Blaise Pascal (1623 - 1662) und Pierre de Fermat (1601 - 1665) zurück. Sie wurden von Antoine Gombauld Chevalier de Méré, Sieur de Baussay (1607-1685) ermahnt, dass „die Mathematik nicht im Einklang mit dem praktischen Leben sei". Das „praktische Leben" war für den Edelmann das Glücksspiel. Ihn interessierte besonders ein damals übliches Würfelspiel, bei dem die Bank gewinnt, wenn bei viermaligem Würfeln mindestens einmal die Augenzahl 6 erscheint. Pascal und Fermat haben sich mit dieser und ähnlichen Fragen auseinandergesetzt und damit die klassische Wahrscheinlichkeitstheorie begründet.

Def. 1.1 (Klassische Definition von Wahrscheinlichkeit) *Ein Ereignis trete zufällig auf. Die Wahrscheinlichkeit für das Auftreten des Ereignisses E ist definiert durch den Quotienten aus der Anzahl g der für das Ereignis günstigen und der Zahl m der möglichen Fälle*

$$P = \frac{g}{m} \qquad .$$

Beispiel: Wahrscheinlichkeit für Herz in einem Skat-Spiel ist $P = \frac{8}{32} = \frac{1}{4}$.

Für die klassische Wahrscheinlichkeit lassen sich für zwei beliebige Ereignisse A und B folgende Regeln ableiten

$$P(A \vee B) = \frac{n_A + n_B - n_{A \wedge B}}{N} = P(A) + P(B) - P(A \wedge B) \tag{1.1a}$$

$$P(N) = \frac{0}{N} = 0 \qquad \text{unmögliches Ereignis, N: Null-Element} \tag{1.1b}$$

$$P(E) = \frac{N}{N} = 1 \qquad \text{sicheres Ereignis, E: Eins-Element} \tag{1.1c}$$

$$0 \leq P(A) \leq 1 \qquad \text{folgt aus der Definition} \tag{1.1d}$$

$$P(A|B) = \frac{n_{A \wedge B}}{n_B} = \frac{P(A \wedge B)}{P(B)} \qquad . \tag{1.1e}$$

Das Symbol $P(A|B)$ steht für die BEDINGTE WAHRSCHEINLICHKEIT für A, vorausgesetzt das Ereignis B liegt vor. In der klassischen Wahrscheinlichkeitstheorie heißt das, dass wir die möglichen Ereignisse vorsortieren. Wir berücksichtigen nur noch solche, die die Bedingung B erfüllen, das seien n_B Ereignisse. Von diesen interessieren uns nun gerade solche, die auch gleichzeitig A erfüllen. Damit ist die Zahl der günstigen Fälle gleich der Zahl der Fälle, die sowohl A als auch B erfüllen, also $n_{A \wedge B}$.

Def. 1.2 (Exklusive Ereignisse) *Sich gegenseitig ausschließende Ereignisse, d.h. $A \wedge B = N$, wollen wir in Anlehnung an die englische Nomenklatur kurz* EXKLUSIV *nennen.*

Def. 1.3 (Komplementäre Ereignisse) *Ein Ereignis $\overline{A}$ ist komplementär zu A, wenn gilt*

$$\overline{A} \vee A = E; \qquad \textit{und} \qquad \overline{A} \wedge A = N \qquad .$$

Aus der Summenregel folgt für exklusive Ereignisse eine vereinfachte Summenregel und daraus wiederum der Zusammenhang zwischen den Wahrscheinlichkeitenkomplementärer Ereignisse

$$P(A \vee B) = P(A) + P(B) \tag{1.2}$$

$$P(\overline{A}) = 1 - P(A) \tag{1.3}$$

Diese klassische Definition der Wahrscheinlichkeit wurde später von Jacob Bernoulli (1654-1705) in seinem Buch Ars Conjectandi (1713 posthum publiziert) weiterentwickelt. Dieses Buch enthält wegweisende Beiträge zur Wahrscheinlichkeitstheorie, unter anderem eine ausführliche Diskussion der wahren Bedeutung des Wahrscheinlichkeitsbegriffes: *Wahrscheinlichkeit ist ein messbares Maß der Gewissheit. Bernoulli unterschied bereits klar zwischen Prior- und Posterior-Wahrscheinlichkeit*. Neben Pascal, Fermat und Bernoulli war Pierre-Simon Laplace (1749-1827) einer der führenden Begründer der modernen Wahrscheinlichkeitstheorie. Laplace benutzte die Wahrscheinlichkeitstheorie u.a. für inverse Schlüsse (z.B. die Straße ist nass, wie groß ist die Wahrscheinlichkeit, dass es geregnet hat?). Die Formel, die er hierzu verwendet hat, geht auf Reverend Thomas Bayes (1702-1761) zurück. Die nach ihm benannte Formel (Bayes Theorem, sogenannt nach Pointcaré viele Jahre später) wurde erst

posthum (1764) publiziert. Eingereicht wurde sie von einem seiner Freunde (Richard Price). Die Formel war allerdings in einer wenig transparenten Form und wurde erst durch Laplace in die heute bekannte Gestalt gebracht. Im Zusammenhang mit der Himmelsmechanik beruhen wesentliche seiner Ergebnisse auf Wahrscheinlichkeitsüberlegungen. Weitere Anwendungen seiner Wahrscheinlichkeitstheorie sind: Buffonsches Nadel-Problem, Glücksspiele, naturwissenschaftliche Probleme, juristische Probleme, Kindersterblichkeit, etc. Für die Ableitung seiner Ergebnisse führte er das Prinzip der Ununterscheidbarkeit (principle of indifference) ein. Auf Laplace gehen auch die Anfänge der Fehlerrechnung zurück.

Es war bereits Bernoulli bekannt, dass die Zahlen m aller Ereignisse und die der Zahl g der günstigen Ereignisse nicht immer eindeutig festgelegt werden können. Zum Beispiel: Es werden zwei Würfel geworfen und wir suchen die Wahrscheinlichkeit, dass die Summe der Augenzahlen 7 ist.

a) Wir betrachten als mögliche Ergebnisse die 11 möglichen Summen der Augenzahlen $(2, 3, \ldots, 12)$. Von diesen Summen ist nur ein Ergebnis (die Summe 7) das positive Ergebnis $\Rightarrow P = 1/11$.

b) Wir betrachten als mögliche Ergebnisse alle der Größe nach sortierten Zahlen-Paare, ohne zwischen dem ersten und zweiten Würfel zu unterscheiden, $(1,1),(1,2),\ldots(1,6),\ (2,2),(2,3),\ldots(6,6)$. Davon gibt es $6+5+4+3+2+1 = 21$. Es gibt 3 positive Ereignisse $(1,6),(2,5),(3,4)$. Daraus folgt $P = 1/7$.

c) Wir unterscheiden zusätzlich, welcher Würfel welche Augenzahl liefert. Der Unterschied besteht darin, dass es nun 2 Möglichkeiten für ungleiche Augenzahlen und weiterhin nur eine für gleiche Augenzahlen gibt. Nun ist $N = 36$ und $g = 6$, mit $(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)$. Das heißt, $P = 1/6$.

In diesem Beispiel ist offensichtlich die letzte Variante die korrekte. Man ist überzeugt, dass das a-priori so sein muss. Die Quanten-Physik lehrt uns aber etwas anderes. Es gibt hier Situationen, in denen die zweite Variante korrekt ist (Bosonen). Hierauf gehen wir später näher ein. Das obige Beispiel zeigt, dass die Definition der Wahrscheinlichkeit präzisiert werden muss:

Def. 1.4 (Präzisierte Definition der klassischen Wahrscheinlichkeit) *Die Wahrscheinlichkeit eines Ereignisses ist gegeben durch das Verhältnis der Zahl der günstigen Ergebnisse zu der aller möglichen; vorausgesetzt, alle Ergebnisse sind gleich-wahrscheinlich.*

Diese Definition eliminiert aber nicht wirklich die Probleme:

- Konzeptionell unbefriedigend, Wahrscheinlichkeit darüber definiert, dass Ereignisse dieselbe Wahrscheinlichkeit haben (Ringschluss).
- Das obige Beispiel hat gezeigt, dass die Gleich-Wahrscheinlichkeit nicht eindeutig ist.

- Kann nur in den Fällen angewandt werden, in denen alle Elementar-Ereignisse dieselbe Wahrscheinlichkeit haben. Kann in vielen Fällen leicht geklärt werden. (Glücksspiel, auch in der statistischen Physik)

Es wurden zur Zuweisung von Wahrscheinlichkeiten zwei Prinzipien vorgeschlagen: das „Principle of Insufficient Reasoning" von Jacob Bernoulli und das „Principle of Indifference" von Pierre-Simon Laplace. Sie besagen: wenn es kein Vorwissen gibt, das die einzelnen Elementar-Ereignisse unterscheidet, kann man sie neu durchnummerieren, ohne die Situation zu verändern. Daraus muss man schließen, dass alle Ereignisse, zwischen denen a-priori nicht unterschieden werden kann, „gleichwahrscheinlich" sind. Es gibt aber Gegenbeispiele für diskrete Probleme aus der Physik. Besonders augenfällig sind die Schwierigkeiten bei kontinuierlichen Problemen, da hier das Maß in's Spiel kommt. Ein prominentes Beispiel ist das sogenannte BERTRAND-PARADOXON (1888), das erst 1968 von E.T.Jaynes (1922-1998) zufriedenstellend gelöst wurde.

1.2 Bertrand Paradoxon

Über einen Kreis hinweg werden zufällig Geraden gezeichnet. Wie groß ist die Wahrscheinlichkeit, dass der Abstand vom Zentrum kleiner ist als die Hälfte des Radius r? Ohne Beschränkung der Allgemeinheit setzen wir $r = 1$ in passend gewählten Einheiten. Es bieten sich mehrere Lösungen an. Wir wollen hier zunächst nur drei diskutieren.

a) Wir nehmen an, dass – wie in Abbildung (1.1) dargestellt – der Abstand vom Zentrum mit gleicher Wahrscheinlichkeit alle Werte zwischen 0 und 1 annehmen kann. Unter dieser Voraussetzung ist das Intervall der günstigen Ereignisse $1/2$ und somit die gesuchte Wahrscheinlichkeit $P = 1/2$.

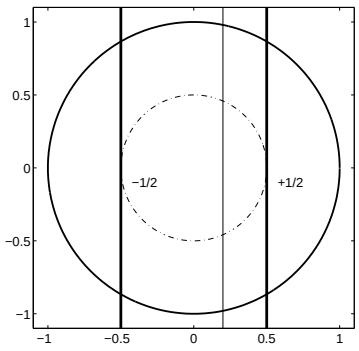

Abbildung 1.1: *Skizze 1 zum Bertrand-Paradoxon.*

b) Alternativ können wir, wie in Abbildung (1.2) skizziert, den Winkel zwischen der Geraden und der Tangente an den Kreis als Größe betrachten, die gleich-

wahrscheinlich alle Werte zwischen 0 und π annehmen kann. Der „günstige" Winkelbereich ist $\pi/3$ und die gesuchte Wahrscheinlichkeit ist demnach in diesem Fall $P = 1/3$.

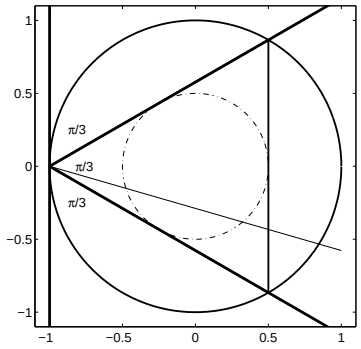

Abbildung 1.2: *Skizze 2 zum Bertrand-Paradoxon.*

c) Eine weitere Möglichkeit ist, die Fläche A des konzentrischen Kreises, der die Gerade – wie in Abbildung (1.3) – berührt, als gleich-verteilt zwischen 0 und π anzunehmen. Der günstige Bereich ist $A \in [0, \pi/4]$, woraus sich eine Wahrscheinlichkeit $P = 1/4$ ergibt.

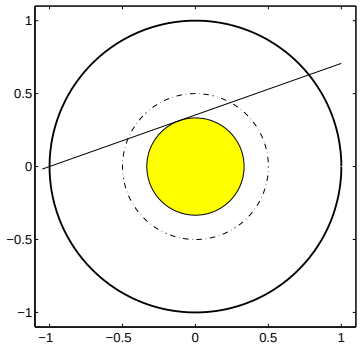

Abbildung 1.3: *Skizze 3 zum Bertrand-Paradoxon.*

Es gibt noch andere plausibel erscheinende Darstellungen, die zu wieder anderen Ergebnissen führen. Wie kann es sein, dass ein eindeutig definiertes Problem keine eindeutige Antwort liefert? Hinter dem Bertrand-Paradoxon verbirgt sich allgemein das Problem, wie man Unwissenheit bei kontinuierlichen Freiheitsgraden zu beschreiben hat. Nehmen wir an, wir wollen etwas über eine Größe x aussagen, die die reellen Zahlen zwischen 0 und 1 annehmen kann. Es ist naheliegend das klassische Konzept zu verwenden, dass, wenn wir nichts weiter wissen, die Wahrscheinlichkeit für

$x \in (x, x+dx)$ als

$$P(x \in (x, x+dx)) = \frac{dx}{1} = p_x(x)dx$$

anzusetzen ist. Hierbei ist $p_x(x) = 1$ die Wahrscheinlichkeitsdichte, die später genauer behandelt wird. Unwissenheit wird hier also durch eine konstante Wahrscheinlichkeitsdichte beschrieben. Die Wahrscheinlichkeitsdichte für x^n ist dann aber nicht mehr konstant sondern, wie wir später zeigen werden,

$$p_z(z = x^n) = p_x(x) \left| \frac{dx}{dz} \right| = \frac{1}{n} z^{\frac{1}{n}-1} \quad .$$

Die Wahrscheinlichkeitsdichte für die Variable z ist demnach scharf bei $z = 0$ gepeakt. Das heißt unsere flache Verteilung in x impliziert, dass es eine sehr große Wahrscheinlichkeit für kleine Werte von z gibt.
Ganz allgemein erkennen wir, dass nichtlineare Transformationen offensichtlich Probleme machen. In welcher Darstellung sind die Ereignisse gleicher Größe (Länge, Fläche, ...) gleich-wahrscheinlich, und wie beschreibt man völliges Unwissen? Auf das Problem PRIOR-WAHRSCHEINLICHKEITEN zuzuweisen kommen wir im Teil III zu sprechen. Hierzu gibt es einen Zugang über Transformationsgruppen[1] und einen differentialgeometrischen, welche beide äquivalent sind. In diesem Zusammenhang werden wir auch das wichtige Prinzip der maximalen Entropie behandeln.
Die klassische Definition der Wahrscheinlichkeit war trotz dieser Probleme mehre Jahrhunderte in Gebrauch und wird auch heute noch auf viele Probleme angewandt; insbesondere solche kombinatorischer Natur.
Bereits Bernoulli hatte sich die Frage gestellt, wie die Wahrscheinlichkeit mit der relativen Häufigkeit zusammenhängt. Beispiel: Die Wahrscheinlichkeit für eine gerade Augenzahl beim Würfeln ist 1/2. Wenn man, das Würfel-Experiment N-mal wiederholt, wie häufig wird man dann eine gerade Augenzahl vorfinden. Wir werden zeigen, dass diese relative Häufigkeit im Limes $N \to \infty$ gegen die „intrinsische" Wahrscheinlichkeit konvergiert. Bernoulli hat aber eigentlich das umgekehrte Problem beschäftigt. Gegeben eine „Stichprobe" vom endlichen Umfang N, was kann man dann über die zugrundeliegende (intrinsische) Wahrscheinlichkeit aussagen?

1.3 Statistische Definition

Um die Prior-Probleme der klassischen Wahrscheinlichkeitstheorie zu vermeiden, verfolgten Ellis (1842), Boole (1854), Venn (1866) und von Mises (1928) eine andere Definition. Da, wie Bernoulli gezeigt hatte, die relative Häufigkeit im Limes $N \to \infty$ gegen die „intrinsische" Wahrscheinlichkeit konvergiert, führten sie folgende Definition der Wahrscheinlichkeit ein

[1] Dieser Zugang wurde von E.T. Jaynes verwendet, um das Bertand-Paradoxon zu klären.

Def. 1.5 (Statistische Definition von Wahrscheinlichkeit) *Ein „Ereignis" A trete zufällig auf. Die Wahrscheinlichkeit ist definiert durch die relative Häufigkeit*

$$P(A) = \lim_{N\to\infty} \frac{n}{N} \qquad , \tag{1.4}$$

dass das Ereignis bei N Versuchen n-mal auftritt, im Limes $N \to \infty$.

Diese Definition soll vermeiden, dass man je in die Verlegenheit kommt, „Prior-Wahrscheinlichkeit" angeben zu müssen, sondern die Wahrscheinlichkeit eine dem zu untersuchenden Objekt anhaftende intrinsische Eigenschaft ist, die nur experimentell bestimmt werden kann, nämlich indem man ein Stichprobe von unendlichem Umfang untersucht. Da eine unendlich Stichprobe nie untersucht werden kann, ist diese Definition von rein hypothetischem Charakter. Diese Definition des Wahrscheinlichkeitsbegriffes ist allerdings weit verbreitet und bildet die Grundlage der „orthodoxen Statistik".

Nachteile des statistischen Wahrscheinlichkeitsbegriffes

- Für viele Probleme gibt es keine Häufigkeitsverteilung
  - War Herr X der Täter?
  - Liegt die Temperatur des Fusionsplasmas $T \in (1.0, 2.0)10^8 K$?
  - War Julius Cäsar Linkshänder?
- Nur in den wenigsten Fällen ist $N \gg 1$
  - Großexperimente, z.B. Fusionsforschung (Kostenfrage)
  - Extraterrestische Physik, Stichprobenumfang von der Natur vorgegeben.
- der Limes $N \to \infty$ ist in der Praxis nicht möglich, und Gl. (1.4) ist keine Definition für $P(A)$ sondern muss als Hypothese gesehen werden. Es handelt sich nicht um eine experimentell bestimmbare Größe.
- Interpretationsprobleme
  - Was bedeutet die Aussage, die Wahrscheinlichkeit ist 0.1, dass Herr X einen Virus hat?
    (100 mal klonen, 10 Klone haben dann den Virus?!)
  - Die Wahrscheinlichkeit im Lotto zu gewinnen ist $7 \cdot 10^{-8}$
    (10^8 mal einzahlen, um einmal zu gewinnen?!)

Man erkennt an den beiden Beispielen auch, wie unterschiedlich Wahrscheinlichkeiten, je nach Kontext, bewertet werden. Im ersten Fall wird man geneigt sein, die Möglichkeit des Virus nahezu auszuschließen, während der Zuspruch des Lottospiels zeigt, dass man ein Wahrscheinlichkeit von $7 \cdot 10^{-8}$ als hinreichend groß bewertet. Wie Wahrscheinlichkeiten tatsächlich zu bewerten sind, besagt die ENTSCHEIDUNGSTHEORIE. Hierbei wird zusätzlich eine Bewertungsfunktion (Kostenfunktion) einfließen, die den unterschiedlichen Ausgängen eines „Experiments" Werte (Kosten) zuweist.

Mit der statistischen Definition kann nur eine stark eingeschränkte Klasse von Problemen behandelt werden, nämlich solche, bei denen ein „Versuch" zumindest theoretisch beliebig oft wiederholt werden kann.
Die statistische Definition der Wahrscheinlichkeit führt allerdings zu denselben Rechenregeln für Wahrscheinlichkeiten, wie die klassische Definition.

Wir haben bei der obigen Diskussion gesehen, dass einige Begriffe und Ideen im Zusammenhang mit der Wahrscheinlichkeitstheorie auftreten, die genauer definiert bzw. analysiert und genauer hinterfragt werden müssen. Hierauf werden wir in einem späteren Kapitel eingehen. Zunächst benötigen wir ein paar wichtige Begriffe zur Charakterisierung von Wahrscheinlichkeitsverteilungen, die im nächsten Kapitel dargestellt werden.

Kapitel 2

Definition von Mittelwert, Momenten und marginaler Verteilung

Im letzten Kapitel haben wir uns damit beschäftigt, Ereignissen Wahrscheinlichkeiten für ihr Auftreten zuzuweisen. Jetzt wollen wir die Ereignisse selbst mit Zahlen in Verbindung bringen.

2.1 Verteilungen einer diskreten Zufallsvariablen

Für die folgenden Definitionen sei eine abzählbare Menge $\mathcal{G}$ von Elementarereignissen gegeben. Jedes Elementarereignis $\omega \in \mathcal{G}$ trete mit Wahrscheinlichkeit P_ω auf.

Def. 2.1 (Zufallsvariable) *Eine Zufallsvariable ist ein Funktional, das jedem Ergebnis $\omega \in \mathcal{G}$ eine reelle Zahl $x = X(\omega)$ zuordnet. x heißt Realisierung von X. Die Menge R der möglichen Realisierungen x heißt Wertebereich von X. $\mathcal{G}$ wird hierbei auf R abgebildet.*

Beispiel:

- Beim Münzwerfen kann man z.B. dem Ereignis „Kopf" die Zahl 0 und dem Ereignis „Zahl" eine 1 zuweisen.
- Beim Würfeln ist es naheliegend, dem Ereignis „n Augen" die Zahl n zuzuweisen.
- Wir wiederholen das Münz-Experiment dreimal. Die Menge der möglichen Ergebnisse ist

$$\begin{aligned}\mathcal{G} =\{&(K,K,K),(K,K,Z),(K,Z,K),(Z,K,K),\\ &(K,Z,Z),(Z,K,Z),(Z,Z,K),(Z,Z,Z)\}\end{aligned} \quad .$$

Wir können diesen 8 Ergebnissen z.B. die Zufallsvariablen

$$3, 2, 2, 2, 1, 1, 1, 0$$

zuweisen, also die Anzahl der Köpfe.

Nun kann man auch Funktionen $Y = f(X)$ der Zufallsvariablen X untersuchen. Natürlich ist Y wieder eine Zufallsvariable. Jedem Ereignis ω wird jetzt er Wert $f(x)$ zugeordnet.

Ein grundlegendes Werkzeug, um Aussagen über Wahrscheinlichkeitsverteilungen treffen zu können, ist die Mittelwertbildung.

Def. 2.2 (Mittelwert einer Zufallsvariablen) *Man definiert als*

MITTELWERT EINER DISKRETEN ZUFALLSVARIABLEN

$$\langle X \rangle := \sum_{\omega \in \mathcal{G}} X(\omega)\ P_\omega \qquad . \tag{2.1}$$

Bemerkungen:

a) Der MITTELWERT wird auch oft ERWARTUNGSWERT genannt.

b) In vielen Fällen ist der Wertebereich R der Zufallsvariablen X gleich der typischen Größen der Elementarereignisse. Z.B. wird dem Ereignis ω_n: *Der Würfel zeigt n Augen* oft der Wert $X(\omega_n) = n$ zugewiesen. Bezeichnet man nun das Ereignis ω_n kurz mit n, dann wird aus Gl. (2.1)

$$\langle n \rangle = \sum_{n \in \mathcal{G}} n\ P_n \qquad . \tag{2.2}$$

Wir werden im Folgenden diese Notation bevorzugen.

c) Der Mittelwert ist keine Zufallsvariable, sondern ein exakt definierter Wert.

d) Die Mittelwertbildung ist eine lineare Operation, d.h.

$$\langle \alpha\ f + \beta\ g \rangle = \alpha\ \langle f \rangle + \beta\ \langle g \rangle$$

für Funktionen $f(n)$, $g(n)$ und Konstanten α, β.

Beweis:

$$\begin{aligned}\langle \alpha\ f + \beta\ g \rangle &= \sum_{n \in M} (\alpha\ f(n) + \beta\ g(n))\ P_n \\ &= \alpha \sum_{n \in M} f(n)\ P_n + \beta \sum_{n \in M} g(n)\ P_n \\ &= \alpha\ \langle f \rangle + \beta\ \langle g \rangle \qquad .\end{aligned} \tag{2.3}$$

e) Für eine Funktion $Y = f(X)$ einer Zufallsvariablen ergibt sich

$$\langle f(X) \rangle = \sum_{n \in M} f(n)\ P_n \qquad . \tag{2.4}$$

Hierfür schreiben wir kurz $\langle f \rangle$ oder $\langle f(n) \rangle$.

f) Kontinuierliche Zufallsvariablen werden im Kapitel 9 vorgestellt.

Beispiel: Beim einmaligen Würfeln sind die Elemantarereignisse die Augenzahlen $n \in \{1, 2, 3, 4, 5, 6\}$, die mit den Wahrscheinlichkeiten $P_n = \frac{1}{6}$ auftreten. Die Zufallsvariable X habe die jeweilige Augenzahl als Wert. Der Mittelwert ist dann $\langle n \rangle = 3.5$. Die Funktion $Y = f(X)$ weise jeder Augenzahl n einen Gewinn $f(n)$ zu. Der Mittelwert $\langle f \rangle$ ist dann der mittlere Gewinn.

Wahrscheinlichkeitsverteilungen kann man nicht nur durch die Angabe der Wahrscheinlichkeiten sondern auch durch andere Größen, wie z.B. die Momente charakterisieren. Momente sind als Mittelwerte bestimmter Zufallsvariablen definiert.

Def. 2.3 (i-tes Moment einer Zufallsvariablen) *Den Mittelwert der Funktion $f(n) = n^i$ bezeichnet man als*

i-TES MOMENT EINER ZUFALLSVARIABLEN

$$m_i := \langle n^i \rangle \qquad . \tag{2.5}$$

Bemerkung: Das erste Moment m_1 einer Verteilung ist offensichtlich der Mittelwert $\langle n \rangle$ der Verteilung, das nullte Moment wegen der Normierung $m_0 = 1$.

Def. 2.4 (i-tes zentrales Moment einer Zufallsvariablen) *Den Mittelwert der Funktion $f(n) = (n - \langle n \rangle)^i$ bezeichnet man als*

i-TES ZENTRALES MOMENT EINER ZUFALLSVARIABLEN
$$\mu_i := \left\langle (\Delta n)^i \right\rangle = \left\langle (n - \langle n \rangle)^i \right\rangle \quad . \tag{2.6}$$

Das zweite Momente erhält einen eigenen Namen: VARIANZ. Aus Gleichung (2.6) erhält man mit der Eigenschaft (2.3)

$$\begin{aligned} \left\langle (n - \langle n \rangle)^2 \right\rangle &= \left\langle n^2 - 2\, n\, \langle n \rangle + \langle n \rangle^2 \right\rangle \\ &= \left\langle n^2 \right\rangle - 2\, \langle n \rangle\, \langle n \rangle + \langle n \rangle^2 \\ &= \left\langle n^2 \right\rangle - \langle n \rangle^2 \quad . \end{aligned} \tag{2.7}$$

Def. 2.5 (Varianz) *Das zweite zentrale Moment heißt auch*

VARIANZ EINER ZUFALLSVARIABLEN
$$\mathrm{var}(n) := \sigma^2 := \left\langle (\Delta n)^2 \right\rangle = \left\langle (n - \langle n \rangle)^2 \right\rangle = \left\langle n^2 \right\rangle - \langle n \rangle^2 \quad . \tag{2.8}$$

Die Varianz ist also die mittlere quadratische Abweichung der Zufallsvariablen vom Mittelwert. Um ein lineares Maß zu bekommen, definiert man die

Def. 2.6 (Standardabweichung) *Die Wurzel der Varianz nennt man*

STANDARDABWEICHUNG
$$\mathrm{std}(x) := \sigma := \sqrt{\mathrm{var}(x)} \tag{2.9}$$

Nehmen wir an, wir wollen durch wiederholtes Würfeln, den Mittelwert, der bei einem symmetrischen Würfel 3.5 beträgt, ermitteln. Dazu würfeln wir N mal und erhalten eine Sequenz (Stichprobe) von Augenzahlen x_i. Daraus bilden wir das arithmetische Mittel (Stichprobenmittelwert)

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

Wie wir später noch eingehend besprechen werden, ist die Abweichung des Stichprobenmittelwertes vom wahren Mittelwert

STANDARDFEHLER EINER STICHPROBE VOM UMFANG N

$$\text{Standardfehler} = \frac{\sigma}{\sqrt{N}} \quad . \tag{2.10}$$

2.2 Verteilungen mehrerer diskreter Zufallsvariablen

Für die nächsten Definitionen soll uns folgendes Beispiel leiten: In einer Firma ist das Gewicht und die Größe der Angestellten bestimmt worden. Folgende Tabelle zeigt die Anzahl der Personen mit entsprechendem Gewicht und Größe.

kg\m	1.6	1.7	1.8	#
60	3	3	1	7
70	0	5	6	11
80	1	4	7	12
#	4	12	14	30

Die Grössenangaben sind als Intervalle der Länge 0.1, die bei den angegebenen Werten zentriert sind, zu verstehend. Entsprechend stehen die Gewichtsangaben für zentrierte Intervalle der Länge 10. Es gibt also z.B. 5 Personen, deren Grösse zwischen 1.65m und 1.75m liegt und deren Gewicht aus dem Intervall $(65kg, 75kg]$ ist. Ganz rechts bzw. unten steht die Summe der entsprechenden Zeile bzw. Spalte. Insgesamt sind also 30 Leute vermessen worden. Den Zeilen kann man eine Zufallsvariable Masse m mit dem Wertebereich $\{60, 70, 80\}$ den Spalten eine Zufallsvariable Höhe h mit dem Wertebereich $\{1.6, 1.7, 1.8\}$ zuweisen. Die Wahrscheinlichkeit, dass ein zufällig ausgewählter Angestellter dieser Firma z.B. 61 kg groß und 1.73 m schwer ist, ist $P = \frac{3}{30}$. Man kann also jedem Eintrag die relative Häufigkeit als Wahrscheinlichkeit P_{mh} zuweisen.

Def. 2.7 (marginale Verteilung) *Seien $P_{n_1,\ldots,n_N}$ die Wahrscheinlichkeiten von Elementarereignissen, die durch die Angabe des N-Tupels von Eigenschaften $n_1, \ldots, n_N$ eindeutig bestimmt sind. Man nennt folgende Verteilung*

MARGINALE VERTEILUNG ODER RANDVERTEILUNG

$$P_{n_1,\ldots,n_j} := \sum_{n_{j+1}} \cdots \sum_{n_N} P_{n_1,\ldots,n_N} \tag{2.11}$$

der ersten j Eigenschaften. Entsprechend definiert man Randverteilungen bezüglich einer anderen Auswahl von Eigenschaften.

Beispiel: In unserem Beispiel kann man sich die Randverteilungen $P_h = \sum_m P_{mh}$ und $P_m = \sum_h P_{mh}$ anschauen. Für erstere erhält man $P_{h=1.6} = \frac{2}{15}$, $P_{h=1.7} = \frac{2}{5}$ und $P_{h=1.8} = \frac{7}{15}$.

Analog zu Gleichungen (2.4), (2.5) bzw (2.6) definieren wir den Mittelwert, die Momente bzw. die zentralen Momente für Funktionen mehrerer Zufallsvariablen.

Def. 2.8 (Mittelwert mehrerer diskreter Zufallsvariablen) *Gegeben seien N Zufallsvariablen $X_1, \ldots, X_N$ mit Wertebereichen $n_1 \in M_1, \ldots, n_N \in M_N$ und eine Funktion $Y = f(X_1, \ldots, X_N)$. Dann ist der*

MITTELWERT EINER FUNKTION MEHRERER DISKRETER ZUFALLSVARIABLEN

$$\langle f(X_1, \ldots, X_N) \rangle := \sum_{n_1} \cdots \sum_{n_N} f(n_1, \ldots, n_N)\ P_{n_1,\ldots,n_N} \quad . \tag{2.12}$$

Beispiel: Der Body Mass Index ist definiert durch $BMI = \frac{\text{Masse in kg}}{(\text{Größe in m})^2}$. Der mittlere Body Mass Index der Firmenangestellten ist durch den Mittelwert der Funktion $f(m,h) = \frac{m}{h^2}$ gegeben. Wir erhalten $\langle f \rangle = 23.9$.

Def. 2.9 (Momente mehrerer Zufallsvariablen) *Den Mittelwert der Funktion $f(n_1, \ldots, n_N) = n_1^{i_1}\ n_2^{i_2} \cdots n_N^{i_N}$ bezeichnet man als*

MOMENT DER ORDNUNG $i_1, i_2, \ldots, i_N$

$$m_{i_1,i_2,\ldots,i_N} := \left\langle n_1^{i_1}\ n_2^{i_2} \cdots n_N^{i_N} \right\rangle \quad . \tag{2.13}$$

Def. 2.10 (zentrale Momente mehrerer Zufallsvariablen) *Den Mittelwert der Funktion* $f(n_1, \ldots, n_N) = (n_1 - \langle n_1 \rangle)^{i_1} \, (n_2 - \langle n_2 \rangle)^{i_2} \, \cdots \, (n_N - \langle n_N \rangle)^{i_N}$ *bezeichnet man als*

ZENTRALES MOMENT DER ORDNUNG $i_1, i_2, \ldots, i_N$

$$\mu_{i_1,i_2,\ldots,i_N} := \left\langle (n_1 - \langle n_1 \rangle)^{i_1} \, (n_2 - \langle n_2 \rangle)^{i_2} \, \cdots \, (n_N - \langle n_N \rangle)^{i_N} \right\rangle \quad . \tag{2.14}$$

Die Mittelwerte der n_i sind durch Momente gegeben

$$\begin{aligned} m_{100\ldots0} &= \langle n_1 \rangle \\ m_{010\ldots0} &= \langle n_2 \rangle \\ &\vdots \\ m_{000\ldots1} &= \langle n_N \rangle \quad , \end{aligned} \tag{2.15}$$

die entsprechenden Varianzen durch zentrale Momente

$$\begin{aligned} \mu_{200\ldots0} &= \left\langle (n_1 - \langle n_1 \rangle)^2 \right\rangle = \mathrm{var}(n_1) \\ \mu_{020\ldots0} &= \left\langle (n_2 - \langle n_2 \rangle)^2 \right\rangle = \mathrm{var}(n_2) \\ &\vdots \\ \mu_{000\ldots2} &= \left\langle (n_N - \langle n_N \rangle)^2 \right\rangle = \mathrm{var}(n_N) \quad . \end{aligned} \tag{2.16}$$

Beispiel: In unserem Beispiel erhalten wir für die mittlere Masse $\langle m \rangle = 71.7\,\mathrm{kg}$, für die mittlere Größe $\langle h \rangle = 1.73\,\mathrm{m}$ und für die entsprechenden Varianzen $\mathrm{var}(m) = 60.5\,\mathrm{kg}^2$ und $\mathrm{var}(h) = 0.0049\,\mathrm{m}^2$.

Def. 2.11 (Kovarianz) *Ein zentrales Moment der Ordnung* $i_1, i_2, \ldots, i_N$ *mit* $i_k = i_{l \neq k} = 1$ *und alle anderen* $i_j = 0$ *heißt*

KOVARIANZ ZWEIER ZUFALLSVARIABLEN

$$\begin{aligned} \mathrm{cov}(n_k, n_l) := \mu_{0\ldots010\ldots010\ldots0} &= \langle (n_{i_k} - \langle n_{i_k} \rangle) \, (n_{i_l} - \langle n_{i_l} \rangle) \rangle \\ &= \langle n_{i_k} \, n_{i_l} \rangle - \langle n_{i_k} \rangle \langle n_{i_l} \rangle \quad . \end{aligned} \tag{2.17}$$

Paaren sich Werte $n_k > \langle n_k \rangle$ hauptsächlich mit Werten $n_l > \langle n_l \rangle$, dann ist die Kovarianz positiv. Treten andererseits Werte $n_k > \langle n_k \rangle$ vorzüglich mit Werten $n_l < \langle n_l \rangle$, auf ist sie negativ. Lässt sich aus der relativen Lage von n_k bezüglich $\langle n_k \rangle$ nicht auf die Lage von n_l bezüglich $\langle n_l \rangle$ schließen, ist die Kovarianz Null. Somit ist die Kovarianz ein Maß für die gegenseitige Abhängigkeit der Zufallsvariablen.

Beispiel: Die Kovarianz zwischen Masse und Größe der Firmenangestellten ergibt $\mathrm{cov}(m, h) = 0.21\,\mathrm{kg\,m}$. Das heißt, eine größere Masse bedeutet im Mittel, dass die entsprechende Person größer ist.

Häufig „spüren" die einzelnen Zufallsvariablen nichts von einander. Solche Verteilungen zeichnen sich durch eine einfache Struktur aus.

Def. 2.12 (unabhängige Zufallsvariablen) *Lässt sich die Verteilung mehrerer Zufallsvariablen als Produkt ihrer marginalen Verteilungen schreiben, nennt man sie*

UNABHÄNGIGE ZUFALLSVARIABLEN

$$P_{n_1,n_2,\ldots,n_N} = \prod_{i=1}^{N} P_{n_i} \quad . \tag{2.18}$$

Bemerkung: Die Kovarianz unabhängiger Zufallsvariablen ist Null.
Beweis:

$$\begin{aligned}
\mathrm{cov}(n_i, n_j) &= \langle (n_i - \langle n_i \rangle)(n_j - \langle n_j \rangle) \rangle \\
&= \sum_{n_1} \cdots \sum_{n_N} \left((n_i - \langle n_i \rangle)(n_j - \langle n_j \rangle) \prod_l P_{n_l} \right) \\
&= \underbrace{\sum_{n_i} (n_i - \langle n_i \rangle) P_{n_i}}_{\langle (n_i - \langle n_i \rangle) \rangle = 0} \underbrace{\sum_{n_j} (n_j - \langle n_j \rangle) P_{n_j}}_{\langle (n_j - \langle n_j \rangle) \rangle = 0} \underbrace{\sum_{\substack{n_1,\ldots,n_N \\ \neq n_i \\ \neq n_j}} \prod_{\substack{l \\ i \neq l \neq j}} P_{n_l}}_{=1} \\
&= 0 \quad .
\end{aligned}$$

Kapitel 3

Einführung in die Kombinatorik

Damit wir auch konkret etwas rechnen können, benötigen wir einige Grundlagen der Kombinatorik.

3.1 Vorbemerkungen

Def. 3.1 (Paare) *Zwischen m Elementen $a_1, a_2, \ldots, a_m$ vom Typ a und n Elementen $b_1, b_2, \ldots, b_n$ vom Typ b sollen alle möglichen unterschiedlichen Paare (a_i, b_k) gebildet werden.*

ANZAHL DER PAARE

$$N_P = n * m \tag{3.1}$$

1	(a_1, b_1)
2	(a_1, b_2)
3	(a_1, b_3)
4	(a_2, b_1)
5	(a_2, b_2)
6	(a_2, b_3)

Tabelle 3.1: Paare der Elemente a_1, a_2 und b_1, b_2, b_3

Beweis: Es gibt m Möglichkeiten ein Element vom Typ a auszuwählen und zu jeder Wahl a_i gibt es n Möglichkeiten für b.
Beispiel: Bridge Karten. Als Menge von Elementen betrachten wir die 4 Farben und die 13 Bilder. Jede Karte ist durch das Paar (Farbe, Bild) definiert. Demnach gibt es $4 * 13 = 52$ Karten.

Def. 3.2 (Multiplets) *Gegeben seien n_1 Elemente $a_1, a_2, \ldots, a_{n_1}$ vom Typ a, n_2 Elemente $b_1, b_2, \ldots, b_{n_2}$ vom Typ b, und so weiter bis zu n_r Elementen $x_1, x_2, \ldots, x_{n_r}$ vom Typ x. Es sollen alle Multiplets (r-Tupel, die von jedem Typ je ein Element enthalten) $(a_{j_1}, b_{j_2}, \ldots, x_{j_r})$ gebildet werden.*

ANZAHL DER MULTIPLETS

$$N_M = n_1 * n_2 * \ldots * n_r = \prod_{i=1}^{r} n_i \tag{3.2}$$

Beweis: Für $r = 2$ reduziert sich die Behauptung auf Paarungen und stimmt somit. Für $r = 3$ betrachten wir die $n_1 * n_2$ Elemente der Paare (a_i, b_j) als Elemente $u_l,\ l = 1, 2, \ldots, n_1 * n_2$ einer neuen Menge. Jedes Tripel (a_i, b_j, c_k) kann dann als Paar (u_l, c_k) aufgefasst werden. Dafür gibt es nun $n_1 * n_2 * n_3$ Möglichkeiten. Wenn man durch Induktion weiter macht, erhält man die Behauptung für beliebiges r.

1	(a_1, b_1, c_1)
2	(a_1, b_1, c_2)
3	(a_1, b_2, c_1)
4	(a_1, b_2, c_2)
5	(a_1, b_3, c_1)
6	(a_1, b_3, c_2)
7	(a_2, b_1, c_1)
8	(a_2, b_1, c_2)
9	(a_2, b_2, c_1)
10	(a_2, b_2, c_2)
11	(a_2, b_3, c_1)
12	(a_2, b_3, c_2)

Tabelle 3.2: Multiplets der Elemente a_1, a_2, b_1, b_2, b_3 und c_1, c_2.

Beispiel: TEILCHEN AUF ZELLEN VERTEILEN läuft darauf hinaus, für jedes Teilchen eine Zelle auszusuchen. Es gebe r Teilchen und n Zellen. Für jedes Teilchen gibt es n unabhängige Möglichkeiten. Somit gibt es insgesamt $n * n * \ldots * n = n^r$ Multiplets (Verteilungen der Teilchen auf die Zellen).
Das Beispiel wird später in der statistischen Physik wichtig. Man kann das Beispiel auch direkt auf Würfel übertragen. Wir identifizieren hierzu die $n = 6$ Augenzahlen mit den Zellen. Der Würfel werde r-mal geworfen (entspricht der Zahl der Teilchen). Es gibt somit 6^r mögliche Ergebnisse (Sequenzen). Hiervon erfüllen 5^r die Bedingung, dass keine 6 vorkommt. Wenn alle Ergebnisse gleich-wahrscheinlich sind, ist

die klassische Wahrscheinlichkeit, bei 6 Würfen nicht die Augenzahl 6 zu erhalten $P(\text{keine 6}) = \frac{5^6}{6^6}$, bzw. das komplementäre Ereignis, bei 6 Würfen mindestens eine 6 zu erhalten, hat die Wahrscheinlichkeit $P(\text{Augenzahl=6 bei 6 Würfen}) = 1 - \frac{5^6}{6^6} = 0.6651 < 2/3$.
Entgegen der naiven Vorstellung reicht es offensichtlich bei 6 Möglichkeiten nicht aus 6-mal zu würfeln, um jedes Ergebnis zu finden.
Wir stellen nun die Frage, wie oft muss man dann würfeln, damit diese Wahrscheinlichkeit größer als 0.9 ist?

$$\begin{aligned} 1 - (\frac{5}{6})^R &> 0.9 \\ (\frac{5}{6})^R &< 0.1 \\ \ln(5/6)\ R &< \ln(0.1) \\ R &> \ln(0.1)/\ln(5/6) \\ R &\geq 13 \quad . \end{aligned}$$

Man benötigt also mindestens doppelt soviele Versuche, wie es Möglichkeiten gibt. Das gilt qualitativ für eine beliebige Zahl von Alternativen $n \geq 3$.
Die Zahl der Zellen n und die der Teilchen r seinen beliebig. Die Wahrscheinlichkeit, dass eine ausgewählte Zelle i leer bleibt, lässt sich annähern durch

$$P(\text{Zelle i leer}) = \left(\frac{n-1}{n}\right)^r = \left(1 - \frac{1}{n}\right)^r = e^{r\ln(1-\frac{1}{n})} = e^{-r/n+O(n^{-2})} \approx e^{-r/n} \ .$$

3.2 Geordnete Stichproben

Wir betrachten eine Menge von n Elementen $a_1, a_2, \ldots, a_n$, die wir in diesem Zusammenhang POPULATION nennen. Aus dieser Population werden Stichproben vom Umfang r ausgewählt $a_{j_1}, a_{j_2}, \ldots, a_{j_r}$. Da es uns auf die Reihenfolge der Elemente ankommt, nennen wir sie GEORDNETE STICHPROBE[1].
Es gibt zwei Möglichkeiten.

a) Die Elemente werden aus der Population kopiert. Da diese Überlegungen auf Urnen-Experimente zurückgehen, nennt man das AUSWÄHLEN MIT ZURÜCKLEGEN.

b) Die Elemente werden aus der Population herausgenommen. Diese Alternative nennt man entsprechend AUSWÄHLEN OHNE ZURÜCKLEGEN. In diesem Fall kann der Umfang der Stichproben r natürlich nicht größer sein als die Größe der Population. Man nennt diese Variante auch VARIATION VON n ELEMENTEN ZUR r-TEN KLASSE.

[1] Das soll gerade nicht bedeuten, dass die Elemente nach irgendeinem Kriterium geordnet werden!

Bei der Auswahl mit Zurücklegen gibt es für jedes Element der Stichprobe n Möglichkeiten aus der Population auszuwählen und somit insgesamt n^r. Ohne Zurücklegen gibt es für das erste Element n Möglichkeiten, danach für das zweite Element nur noch $n-1$, da bereits ein Element in der Population fehlt, usw. bis zum r-ten Element, für das es nur noch $n-r+1$ Möglichkeiten gibt.

ZAHL DER GEORDNETEN STICHPROBEN VOM UMFANG r AUS EINER POPULATION DER GRÖSSE n

$$N_{\text{op}}^{\text{mz}} = n^r \qquad \text{mit Zurücklegen} \tag{3.3a}$$

$$N_{\text{op}}^{\text{oz}} = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!} \qquad \text{ohne Zurücklegen} \tag{3.3b}$$

Ein Spezialfall einer Stichprobe ohne Zurücklegen stellt $r = n$ dar. In diesem Fall stellt die Stichprobe eine mögliche Anordnung (PERMUTATION) der Elemente der Population dar.

ZAHL DER PERMUTATIONEN

$$N_{\text{perm}} = n! \tag{3.4}$$

3.2.1 Beispiele

Wir nehmen eine Stichprobe vom Umfang r mit Zurücklegen aus einer Population der Größe n. Wie groß ist Wahrscheinlichkeit, dass kein Element in der Stichprobe doppelt vorkommt. Das heißt, die Stichprobe hätte auch ohne Zurücklegen erzeugt worden sein können. Gemäß Gl. (3.3a) gibt es n^r geordnete Stichproben mit Zurücklegen. Die Zahl der Variationen von n Elementen zur Klasse r ist $n!/(n-r)!$, das ist genau die Zahl der Arrangements von r Elementen aus einer Population der Größe n, bei denen keine Elemente doppelt vorkommen, bzw. die Zahl der geordneten Stichproben ohne Zurücklegen.
Somit ist die Wahrscheinlichkeit, dass in der Stichprobe kein Element doppelt vor-

kommt

$$P = \frac{n!}{(n-r)!}\left(\frac{1}{n}\right)^r \quad . \tag{3.5}$$

Es gibt verschiedene interessante Interpretationen dieses Ergebnisses:

a) Es gibt $n = 10$ Ziffern $(0, 1, \ldots, 9)$. Wir nehmen einmal an, dass die letzten Ziffern von Zahlen in großen mathematischen Tabellen „zufälligen" Charakter haben. Wir nehmen eine sehr lange Dezimalzahl und analysieren die Nachkommastellen. Wir betrachten die letzten 5 Ziffern von Dezimalzahlen, d.h. die Stichprobe hat den Umfang $r = 5$. Gemäß Gl. (3.5) ist die Wahrscheinlichkeit, dass in den 5-er Blöcken keine Ziffer doppelt vorkommt, $P = \frac{10!}{5!}10^{-5} = 0.3024$. Es wurde ein Experiment durchgeführt, bei dem in Tabellenwerken mit 16-stelligen Zahlenangaben die letzten 5 Ziffern der Einträge auf doppelte Ziffern untersucht wurden. Dazu wurden 12 Pakete mit je 100 Zahlen untersucht und ausgewertet in wievielen keine Ziffern doppelt sind. Es ergaben die 12 Pakete die Anzahlen: $30, 27, 30, 34, 26, 32, 37, 36, 26, 31, 36, 32$. Das arithmetische Mittel der Häufigkeit ist 0.31 in verblüffender Übereinstimmung mit dem obigen Ergebnisse. Allerdings ist es eine Aufgabe der induktiven Logik festzustellen, ob aus dieser „Vorwärtsrechnung" geschlossen werden kann, dass die Ziffern in Tabellen wirklich zufällig sind.

b) Wenn n Kugeln zufällig auf n Zellen verteilt werden, ist die Wahrscheinlichkeit, dass jede Zelle genau eine Kugel enthält $P = n!/n^n$ (entspricht $r = n$ im obigen Beispiel). Für $n = 6$ ist die Wahrscheinlichkeit 0.015. Das ist z.B. die Wahrscheinlichkeit, dass beim Würfeln nach 6 Würfen jede Ziffer genau einmal vorkommt. Der sehr kleine Zahlenwert von 0.015 offenbart einen unerwarteten Aspekt „zufälliger" Ereignisse.

3.3 Unterpopulationen und Partitionierungen

Wie zuvor betrachten wir Populationen der Größe n. Wie nennen zwei Populationen ungleich, wenn es mindestens ein Element einer Population in der anderen nicht gibt. Wir betrachten eine Unterpopulation der Größe r einer gegebenen Population. Da es auf die Reihenfolge nicht ankommt ist

ZAHL DER UNTERPOPULATIONEN DER GRÖSSE r
EINER POPULATION DER GRÖSSE n

$$N(r|n) = \binom{n}{r} = \frac{n!}{r!(n-r)!} \qquad \text{ohne Zurücklegen} \qquad (3.6a)$$

$$N^{\mathrm{mz}} = \binom{n+r-1}{r} \qquad \text{mit Zurücklegen} \qquad (3.6b)$$

Man nennt das auch die Zahl der Kombinationen von n Elementen zu Gruppen der Größe r. Der **Beweis** für Gl. (3.6a) (der für Gl. (3.6b) wird im Abschnitt 3.4 nachgereicht) folgt aus der Zahl der geordneten Stichproben ohne Zurücklegen vom Umfang r aus Populationen der Größe n. Die Zahl der geordnete Stichproben ist nach Gl. (3.3b) $N_{\mathrm{op}}^{\mathrm{oz}} = \frac{n!}{(n-r)!}$. Bei den r Elementen der Stichprobe kommt es uns aber nun nicht mehr auf die Reihenfolge an. Da es $r!$ mögliche Permutationen gibt, folgt somit Gl. (3.6a). Mit anderen Worten, es gibt $\binom{n}{r}$ Teilmengen mit r Elementen einer Menge mit n Elementen. Es gibt natürlich genauso viele Unterpopulationen der Größe r wie solche der Größe $n-r$

$$\binom{n}{r} = \binom{n}{n-r} \qquad .$$

Es ist sinnvoll zu definieren

$$\binom{n}{0} = 1, \qquad 0! = 1 \qquad \text{und} \quad \binom{n}{r} = 0 \text{ für } r > n \qquad .$$

Def. 3.3 (Binomial-Koeffizienten) *Die Größen*

$$\binom{n}{r}$$

werden Binomial-Koeffizienten genannt.

Gegeben seinen n Kugeln, davon seien n_1 blau und $n_2 = n - n_1$ rot. Wieviele unterschiedliche Farbsequenzen gibt es? Die Antwort lautet

$$N_{\mathrm{Sequenzen}} = \binom{n}{n_1} \qquad .$$

Der Beweis ist wie folgt: Eine Farbsequenz ist dadurch gekennzeichnet, dass man angibt, an welcher Position sich die blauen Kugeln befinden. Aus den n Positionen (Population) werden n_1 ausgewählt.

BINOMISCHER SATZ

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r} \qquad . \tag{3.7}$$

Beweis: Wenn man die linke Seite explizit ausmultipliziert erhält man Summen von Produkten von Faktoren a und b in allen möglichen Sequenzen, die insgesamt immer n Faktoren enthalten, z.B.

$$\underbrace{a\ a\ b\ a\ b\ b\ b\ a\ b\ \ldots\ a}_{\text{n Faktoren}}$$

Zu fester Anzahl r der Faktoren a gibt es $\binom{n}{r}$ unterschiedliche Sequenzen, die – da es auf die Reihenfolge nicht ankommt – alle denselben Wert $a^r b^{n-r}$ beitragen.
Eine Anwendung des Binomischen Satzes ist die BINOMIALVERTEILUNG. Wir führen ein Zufalls-Experiment durch, bei dem es nur zwei mögliche Ausgänge gibt, z.B: rote/blaue Kugel, Kopf/Zahl defekt/intakt, etc. Wird ein solches Experiment wiederholt durchgeführt, spricht man von einem BERNOULLI-VERSUCH. Die Wahrscheinlichkeit für die erste Alternative sei p, und die Wahrscheinlichkeit für die zweite Alternative ist dann $q = 1 - p$. Die Wahrscheinlichkeit, dass bei einer Stichprobe vom Umfang n die erste Alternative r-mal erscheint, ist unter Berücksichtigung der Reihenfolge $p^r q^{n-r}$. Die Wahrscheinlichkeit ohne Berücksichtigung der Reihenfolge ist, da es $\binom{n}{r}$ möglicher Sequenzen zu festem r gibt,

BINOMIAL-VERTEILUNG

$$P(r|n,p) = \binom{n}{r} p^r (1-p)^{n-r} \tag{3.8a}$$

$$\langle r \rangle = n\ p \tag{3.8b}$$

$$\mathrm{var}(r) = n\ p\ (1-p) \qquad . \tag{3.8c}$$

Die richtige Normierung der Wahrscheinlichkeit auf eins folgt aus dem Binomischen Satz

$$\sum_{r=0}^{n} \binom{n}{r} p^r q^{n-r} = (p+q)^n = 1 \qquad .$$

Es wurde ausgenutzt, dass $p + q = 1$.

Beweis:

$$\begin{aligned}
\langle r \rangle &= \sum_{r=0}^{n} r \binom{n}{r} p^r (1-p)^{n-r} = \sum_{r=0}^{n} r \frac{n!}{r!\,(n-r)!} p^r (1-p)^{n-r} \\
&= \sum_{r=1}^{n} n \frac{(n-1)!}{(r-1)!\,((n-1)-(r-1))!} p\, p^{r-1} (1-p)^{(n-1)-(r-1)} \\
&= n\, p \underbrace{\sum_{k=0}^{n-1} \frac{(n-1)!}{k!\,((n-1)-k)!} p^k (1-p)^{(n-1)-k}}_{=1} \\
&= n\, p
\end{aligned}$$

$$\begin{aligned}
\left\langle r^2 \right\rangle &= \sum_{r=0}^{n} r^2 \binom{n}{r} p^r (1-p)^{n-r} = \sum_{r=0}^{n} r^2 \frac{n!}{r!\,(n-r)!} p^r (1-p)^{n-r} \\
&= \sum_{r=1}^{n} n\, r \frac{(n-1)!}{(r-1)!\,((n-1)-(r-1))!} p\, p^{r-1} (1-p)^{(n-1)-(r-1)} \\
&= n\, p \sum_{k=0}^{n-1} (k+1) \frac{(n-1)!}{k!\,((n-1)-k)!} p^k (1-p)^{(n-1)-k} \\
&= n\, p \left(1 + \underbrace{\sum_{k=0}^{n-1} k \frac{(n-1)!}{k!\,((n-1)-k)!} p^k (1-p)^{(n-1)-k}}_{=(n-1)\,p} \right) \\
&= n\, p\, (1 + (n-1)\, p)
\end{aligned}$$

$$\begin{aligned}
\mathrm{var}(r) &= \left\langle r^2 \right\rangle - \langle r \rangle^2 \\
&= n\, p\, (1-p)
\end{aligned}$$

Wir erweitern diese Überlegungen auf mehr als zwei alternative Ausgänge des Experiments. Es seien ganze Zahlen $n_1, n_2, \ldots, n_k$ mit $\sum_{i=1}^{k} = n$ gegeben. Die Zahl der möglichen Aufteilung (PARTITIONIERUNG) der Population auf k Unterpopulationen von denen die erste n_1 Elemente, die zweite n_2 Elemente, etc. hat, ist

MULTINOMIAL-KOEFFIZIENTEN

$$N(\{n_i\}|n,k) = \binom{n}{\{n_i\}} = \frac{n!}{\prod_{i=1}^{k} n_i!} \quad . \tag{3.9}$$

Beweis: Wir nummerieren die Elemente so durch, dass die Elemente der Unterpopulationen aufeinanderfolgende Nummern haben. Elemente der Unterpopulationen

eins haben die Nummern $1, 2, \ldots n_1$, die zweite Unterpopulation hat die Nummern $n_1 + 1, n_1 + 2, \ldots n_1 + n_2$, etc. bis zur k-ten Unterpopulation mit den Nummern $n_{k-1}+1, n_{k-1}+2, \ldots, n$. Wir erzeugen alle Partitionierungen, indem wir die Elemente auf die Positionen $1 \ldots n$ verteilen (permutieren). Es gibt $n!$ Permutationen. Permutationen, bei denen nur die Elemente in den Unterpopulationen permutiert werden, beschreiben dieselbe Partitionierung. Daraus folgt, dass bei den $n!$ Permutationen von jeder Unterpopulation ein Faktor $n_\alpha!$ ($\alpha = 1, 2, \ldots k$) zu viel vorkommt. Damit haben wir den Beweis für Gl. (3.9) erbracht.
Die Binomial-Koeffizienten stellen den Spezialfall $k = 2$ dar, $N(r|n) = N(\{r, n - r\}|n, 2)$.
Die Verallgemeinerung des Binomischen Satzes ergibt

$$(a_1 + a_2 + \ldots + a_k)^n = \widetilde{\sum_{n_1,n_2,\ldots,n_k=0}^{n}} \binom{n}{\{n_i\}} a_1^{n_1} a_2^{n_2} \ldots a_k^{n_k} \quad . \tag{3.10}$$

Hierbei soll die Tilde andeuten, dass nur solche Zahlen $\{n_i\}$ erlaubt sind, die in der Summe n ergeben. Der Beweis ist analog zu dem des Binomischen Satzes.
Genauso können wir die Binomial-Verteilung verallgemeinern

MULTINOMIAL-VERTEILUNG

$$P(\{n_i\}|n, k) = \binom{n}{\{n_i\}} \prod_{i=1}^{k} p_i^{n_i} \tag{3.11a}$$

$$\langle n_i \rangle = n\, p_i \tag{3.11b}$$

$$\mathrm{var}(n_i) = n\, p_i\, (1 - p_i) \tag{3.11c}$$

$$\mathrm{cov}(n_i, n_j) = n\, p_i\, (\delta_{ij} - p_j) \quad . \tag{3.11d}$$

Die Multinomial-Verteilung gibt die Wahrscheinlichkeit an, von n Teilchen n_α in Zelle α anzutreffen, wenn p_α die Prior-Wahrscheinlichkeit ist, ein Teilchen in Zelle α anzutreffen. Wenn alle Zellen gleich-wahrscheinlich sind, ist $p_\alpha = 1/k$.
Aus Gl. (3.10) folgt unmittelbar, dass die Multinomial-Verteilung auf eins normiert ist.

$$\widetilde{\sum_{n_1,n_2,\ldots,n_k=0}^{n}} P(\{n_i\}|n, k) = \widetilde{\sum_{n_1,n_2,\ldots,n_k=0}^{n}} \binom{n}{\{n_i\}} \prod_{i=1}^{k} p_i^{n_i} = (p_1 + p_2 + \ldots + p_k)^n = 1 \; .$$

Beweis: Für den Beweis der oben angegebenen Gleichungen verwenden wir einen

Trick, der in der Wahrscheinlichkeitsrechnung oft verwendet wird.

$$\begin{aligned}
\langle n_i \rangle &= \widetilde{\sum_{n_1,n_2,\ldots,n_k=0}^{n}} n_i \binom{n}{\{n_m\}} \prod_{l=1}^{k} p_l^{n_l} \\
&= p_i \frac{\partial}{\partial p_i} \widetilde{\sum_{n_1,n_2,\ldots,n_k=0}^{n}} \binom{n}{\{n_m\}} \prod_{l=1}^{k} p_l^{n_l} \\
&= p_i \frac{\partial}{\partial p_i} (p_1 + p_2 + \ldots + p_k)^n \\
&= p_i \, n \underbrace{(p_1 + p_2 + \ldots + p_k)}_{=1}^{n-1} \\
&= n \, p_i
\end{aligned}$$

Analog verfährt man für $\langle n_i \, n_j \rangle$:

$$\begin{aligned}
\langle n_i \, n_j \rangle &= \widetilde{\sum_{n_1,n_2,\ldots,n_k=0}^{n}} n_i \, n_j \binom{n}{\{n_m\}} \prod_{l=1}^{k} p_l^{n_l} \\
&= p_i \frac{\partial}{\partial p_i} \, p_j \frac{\partial}{\partial p_j} \widetilde{\sum_{n_1,n_2,\ldots,n_k=0}^{n}} \binom{n}{\{n_m\}} \prod_{l=1}^{k} p_l^{n_l} \\
&= p_i \frac{\partial}{\partial p_i} \, p_j \frac{\partial}{\partial p_j} (p_1 + p_2 + \ldots + p_k)^n \\
&= p_i \frac{\partial}{\partial p_i} \, p_j \, n \, (p_1 + p_2 + \ldots + p_k)^{n-1} \\
&= p_i \Big[\delta_{ij} \, n \, (p_1 + p_2 + \ldots + p_k)^{n-1} \\
&\quad + p_j \, n \, (n-1) \, (p_1 + p_2 + \ldots + p_k)^{n-2} \Big] \\
&= p_i \, (n \, \delta_{ij} + n \, (n-1) \, p_j)
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}(n_i, n_j) &= \langle n_i \, n_j \rangle - \langle n_i \rangle \langle n_j \rangle \\
&= n \, p_i \, \delta_{ij} + n^2 \, p_i \, p_j - n \, p_i \, p_j - n^2 \, p_i \, p_j \\
&= n \, p_i \, (\delta_{ij} - p_j)
\end{aligned}$$

$$\mathrm{var}(n_i) = \mathrm{cov}(n_i, n_i) = n \, p_i \, (1 - p_i)$$

3.3.1 Vollständige Paarungen einer Population

Ausgehend von einer Population mit $N = 2m$ Elementen fragen wir nach der Zahl der vollständigen Paarungen (Kontraktionen).
Beispiel: Population a_1, a_2, a_3, a_4 hat die 3 vollständigen Paarungen

1	$(a_1, a_2)(a_3, a_4)$
2	$(a_1, a_3)(a_2, a_4)$
3	$(a_1, a_4)(a_2, a_3)$

Zur Erzeugung aller vollständigen Paarungen (Kontraktionen) reihen wir die Paare hintereinander. An Positionen 1 und 2 steht das erste Paar, an Position 3 und 4 das zweite Paar, und so weiter bis zu den Positionen $2m-1$ und $2m$ für das m-te Paar. Es gibt $(2m)!$ Permutationen der Elemente der Population auf die Positionen. Da es auf die Reihenfolge innerhalb der Paare nicht ankommt, wird hierbei ein Faktor 2 bei jedem Paar zu viel gezählt. Das ergibt insgesamt einen Faktor 2^m. Darüber hinaus kommt es uns auch nicht auf die Reihenfolge der Paare an, d.h. $(a_1, a_2)(a_3, a_4)$ beschreibt dieselben Kontraktionen wie $(a_3, a_4)(a_1, a_2)$. Bei m Paaren ist in allen Permutationen ein Faktor $m!$ zu viel enthalten. Damit ist die Anzahl der Kontraktionen

$$N_{\mathrm{K}} = \frac{(2m)!}{2^m\, m!} = \frac{(2m)!}{[2m][2(m-1)][2(m-2)]\cdots[2]} = \frac{(2m)!}{[2m][2m-2][2m-4]\cdots[2]}$$

$$= \frac{[2m](2m-1)[2m-2](2m-3)\cdots(3)[2](1)}{[2m][2m-2][2m-4]\cdots[2]}$$

$$= 1\cdot 3\cdot 5\cdot \ldots \cdot (2m-1) = (2m-1)!!$$

ZAHL DER VOLLSTÄNDIGEN PAARUNGEN VON $N = 2m$ ELEMENTEN

$$N_{\mathrm{K}} = (N-1)!! \tag{3.12}$$

3.3.2 Beispiel: der Random-Walk

In einem schief gestellten Brett befinden sich Nägel an regelmäßigen Positionen, wie in Abbildung 3.1 dargestellt. Kugeln fallen durch den eingezeichneten Schacht auf den oberen Nagel. Mit gleicher Wahrscheinlichkeit sollen die Kugeln entweder nach links oder rechts abgelenkt werden. Daraufhin treffen sie auf einen der Nägel in der zweiten Reihe. Auch hier werden sie mit gleicher Wahrscheinlichkeit nach links oder rechts abgelenkt. Uns interessiert nun, mit welcher Häufigkeit die Kugeln in die Zellen am unteren Rand des *Pinball-Brettes* fallen. Wir können das Problem wie folgt

analysieren. Vor dem ersten Schritt ist die Kugel an der Position $x = 0$. Nach dem ersten Schritt wird die Kugel eine der beiden Positionen $x = \pm 1$ annehmen. Allgemein befindet sich die Kugel nach dem i-ten Schritt an der Position x_i und kann diese im nächsten Schritt um $\Delta x = \pm 1$ ändern. Es handelt sich hier also um einen wiederholten Versuch, bei dem mit gleicher Wahrscheinlichkeit eines von zwei Ereignisse ($\Delta x = \pm 1$) eintritt. Nach n Schritten sei $\Delta x = +1$ mit der Häufigkeit k vorgekommen[2]. Dementsprechend ist $\Delta x = -1$ mit der Häufigkeit $n - k$ aufgetreten und die Kugel hat die Position

$$x_n(k) = (+1) \cdot k + (-1) \cdot (n - k) = 2k - n$$

erreicht. Im folgenden werden wir zunächst eine beliebige Wahrscheinlichkeit p für die Ablenkung nach rechts und $q = 1 - p$ für die Ablenkung nach links annehmen. Das heißt, die Wahrscheinlichkeit dafür, dass sich die Kugel nach n Schritten am Platz $x_n(k)$ befindet, entspricht der Wahrscheinlichkeit für $k = (x_n(k) + n)/2$ (wenn n gerade/ungerade muss auch $x_n(k)$ gerade/ungerade sein)

$$\begin{aligned} P(x_n|n,p) &= P(k = \frac{x_n + n}{2}|n,p) \\ &= \binom{n}{k} p^k q^{n-k}\Big|_{k=\frac{x_n+n}{2}} \quad . \end{aligned} \tag{3.13}$$

Der Binomial-Koeffizient $\binom{n}{k}$ gibt an, wieviele Möglichkeiten es gibt, dass eine Kugel im n-ten Schritt den Nagel am Platz $i = x_n(k)$ erreicht. Wir erkennen an der Formel oder anhand der Abbildung (3.1), dass die Plätze im n-ten Schritt die Werte $i = -n, -n + 2, \ldots, n$ annehmen können.
Um im n-ten Schritt zum Platz i zu gelangen, muss die Kugel zuvor am Platz $i - 1$ oder $i + 1$ gewesen sein. Die Zahl der Möglichkeiten $N_{i,n}$, mit n Schritten an den Platz i zu gelangen, ist also die Summe der Möglichkeiten in $(n - 1)$ Schritten an die Plätze $i - 1$ oder $i + 1$ zu gelangen. Am Rand fehlt jeweils ein Nachbar und die zugehörige Zahl der Möglichkeiten ist dort Null

$$N_{i,n} = 0 \qquad \text{für } |i| > n \qquad .$$

Beginnt man mit dem obersten Nadel ($N_{0,0} = 1$) und bestimmt iterativ die Zahl der Möglichkeit $N_{i,n}$ aus der Formel

$$N_{i,n} = N_{i-1,n-1} + N_{i+1,n-1} \; ; \qquad \text{für } i = -n, -n + 2, \ldots, n$$

so erhält man die in in Abb. (3.1-b) dargestellten Binomial-Koeffizienten. Man nennt das Gebilde auch PASCALSCHES DREIECK. Die Häufigkeitsverteilung der Kugeln in der n-ten Reihe des *Pinball-Brettes* entspricht demnach den Binomial-Koeffizienten.
Als Verallgemeinerung könnten wir die Anordnung dadurch verkomplizieren, dass wir die Wahrscheinlichkeit p/q für eine Ablenkung nach rechts/links vom Ort (n, i)

[2] Natürlich gilt $0 \leq k \leq n$.

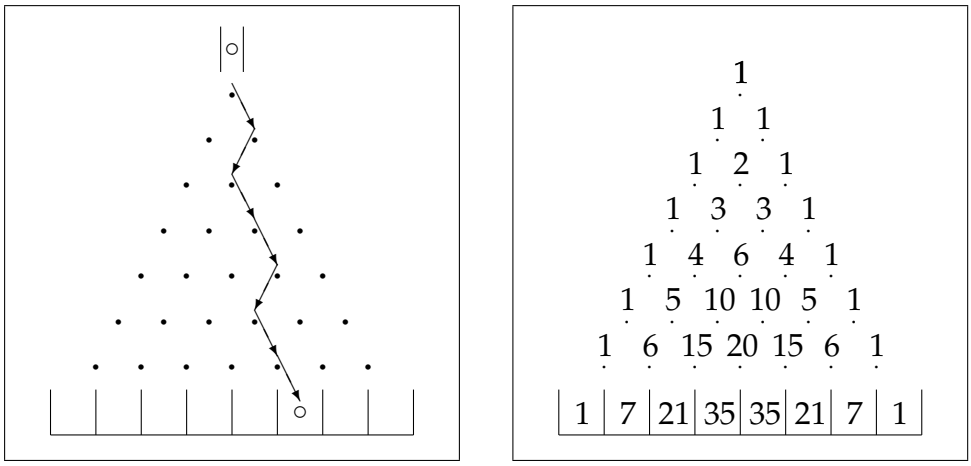

Abbildung 3.1: a) Pinball, b) Pascalsches Dreieck

abhängig machen. Außerdem könnten wir noch einführen, dass es Nägel gibt, die nicht nur in die nächsten-Nachbar-Plätze „streuen", sondern auch weiter reichende Steuprozesse beschreiben. Schließlich können wir auch mehrere Kugeln gleichzeitig auf die Reise schicken, die sich gegenseitig behindern. Dieses Modell ähnelt dann den Problemen, die bei der Pfadintegral-Behandlung von Vielteilchen-Problem in der Festkörperphysik auftreten.

Wir wissen nun, dass die Wahrscheinlichkeit dafür, dass sich die Kugel nach n Schritten an der Position x_n befindet, durch Gl. (3.13) gegeben ist. Hieraus können wir ermitteln, wo sich die Kugeln im Mittel befinden werden und wie weit die Positionen der einzelnen Kugeln von diesem Wert abweichen können. Zur Berechnung, wo sich die Kugeln im Mittel befinden werden und welche Abweichungen hiervon zu erwarten sind, benötigen wir $\langle x^\nu \rangle$ für $\nu = 1, 2$

$$\begin{aligned} \langle x^\nu \rangle &= \sum_{k=0}^{n} \big(x_n(k)\big)^\nu \binom{n}{k} p^k q^{n-k} \\ &= \sum_{k=0}^{n} \big(2k-n\big)^\nu \binom{n}{k} p^k q^{n-k} \qquad . \end{aligned}$$

Das heißt

$$\begin{aligned} \langle x \rangle &= 2 \underbrace{\sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k}}_{\langle k \rangle} - n \underbrace{\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}}_{1} \\ &= 2\langle k \rangle - n \\ \big\langle x^2 \big\rangle &= 4\big\langle k^2 \big\rangle - 4\, n\, \langle k \rangle + n^2 \qquad . \end{aligned}$$

$\langle k \rangle$ und $\langle k^2 \rangle$ sind nach Gl. (3.8b) und Gl. (3.8c) $\langle k \rangle = n\ p$ und $\langle k^2 \rangle = \mathrm{var}(k) + \langle k \rangle^2 = n\ p\ (1-p) + (n\ p)^2$. Damit lautet der gesuchte Erwartungswert der Position

$$\langle x \rangle = 2\ p\ n - n = (2\ p - 1)\ n \qquad .$$

Es ist sinnvoll die Definition

$$p = \frac{1}{2}(1+v) \qquad \Rightarrow \qquad q = (1-p) = \frac{1}{2}(1-v)$$

einzuführen. Damit lautet

$$\langle x \rangle = v\ n \qquad .$$

Weiter erhalten wir

$$\begin{aligned} \left\langle x^2 \right\rangle &= 4npq + 4(np)^2 - 4nnp + n^2 \\ &= n(1-v^2) + n^2\ (4p^2 - 4p + 1) \\ &= n(1-v^2) + n^2\ (2p-1)^2 \\ &= n(1-v^2) + n^2\ v^2 \\ &\Rightarrow \\ \mathrm{var}(x) &= \left\langle x^2 \right\rangle - \langle x \rangle^2 = n(1-v^2) \qquad . \end{aligned} \tag{3.14}$$

Das heißt, der Schwerpunkt der Wahrscheinlichkeitsverteilung „driftet" mit $v\ n$. Identifiziert man n mit der Zeit t, kann man diesen Vorgang physikalisch interpretieren: der Schwerpunkt bewegt sich gleichförmig $x_S = v\ t$. Die Schwerpunktsgeschwindigkeit (Drift-Geschwindigkeit) ist durch

$$v = p - q$$

gegeben. Die Breite B der Verteilung wird durch $\mathrm{std}(x) = \sqrt{\mathrm{var}(x)}$ wiedergegeben und ist

$$B = \sqrt{n\ (1-v^2)} \quad \left(= \sqrt{t\ (1-v^2)} \right) \qquad .$$

Die Breite nimmt durch Diffusion proportional zu $\sqrt{t}$ zu. Die Diffusionsgeschwindigkeit nimmt mit abnehmender Driftgeschwindigkeit zu und ist maximal für $v = 0$.

3.3.3 Beispiel: Korrektur bei der Informationsübertragung

Die Übertragung von Daten (Bits) sei aufgrund von zufälligen Ereignissen fehlerbehaftet. Dies ist insbesondere in Verbindung mit Quantencomputern interessant. Die Wahrscheinlichkeit, dass ein Bit (0/1) korrekt übertragen wird, sei $p > 1/2$. Um Unsicherheiten zu eliminieren, wird die Bit-Übertragung n-mal wiederholt und die sogenannte MAJORITÄTSREGEL verwendet. Sie besagt, dass das zu bestimmende Bit den

Wert 0 oder 1 erhält, je nachdem, welcher Wert häufiger vorkommt. Um eine Patt-Situation auszuschließen, betrachten wir nur ungeradzahlige n.
Wir wollen die Wahrscheinlichkeit $P(\text{Bit korrekt}|p,n)$ bestimmen, dass der so ermittelte Bit-Wert korrekt ist. Um sagen zu können, ob die Majoritätsregel den richtigen Wert liefert, müssen wir die Zahl m der korrekt übertragenen Bits über die Marginalisierungsregel einführen

$$\begin{aligned} P(\text{Bit korrekt}|p,n) &= \sum_{m=0}^{n} P(\text{Bit korrekt}, m|p,n) \\ &= \sum_{m=0}^{n} P(\text{Bit korrekt}|m,p,n)\ P(m|p,n) \end{aligned} \quad .$$

Die beiden Faktoren sind nun bekannt. Der zweite Faktor ist die Binomial-Verteilung und der erste Faktor ist gemäß der Majoritätsregel

$$P(\text{Bit korrekt}|m,p,n) = \begin{cases} 1 & \text{für } m \geq \frac{n+1}{2} \\ 0 & \text{sonst} \end{cases} \quad .$$

Somit gilt

$$\sum_{m=0}^{n} P(\text{Bit korrekt}, m|p,n) = \sum_{m=\frac{n+1}{2}}^{n} \binom{n}{m} p^m\ (1-p)^{n-m}$$

$$\backsimeq \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-x^2/2}\, dx \quad = \frac{1}{2}\big(1 - \text{erf}(\frac{a}{\sqrt{2}})\big)$$

$$\text{mit} \qquad a = -\frac{n(p-1/2)}{\sqrt{np(1-p)}}$$

Im letzten Schritt wurde die de-Moivresche Integralformel verwendet, die im nächsten Kapitel (Abschnitt 4.3) vorgestellt wird.
In Abbildung 3.2 ist diese Wahrscheinlichkeit über der Zahl n der für die Majoritätsregel verwendeten Bits für verschiedene Werte der Wahrscheinlichkeit p aufgetragen. Je größer p ist, desto schneller konvergiert die Treffer-Wahrscheinlichkeit der Majoritätsregel gegen Eins. Es ist umgekehrt interessant sich zu überlegen, welchen Wert n annehmen muss, damit die Treffer-Wahrscheinlichkeit eine vorgegebene Schwelle überschreitet. In Abbildung 3.2 sind die Ergebnisse für die Schwellwerte $P \geq 0.8, P \geq 0.9$ und $P \geq 0.99$ als Funktion von p aufgetragen.

3.4 Anwendung auf Besetzungszahl-Probleme

a) In vielen Anwendungen, insbesondere in der Statistischen Physik, hat man es mit einer Art von Problemen zu tun, bei denen identische Objekte (Teilchen/Kugeln) auf

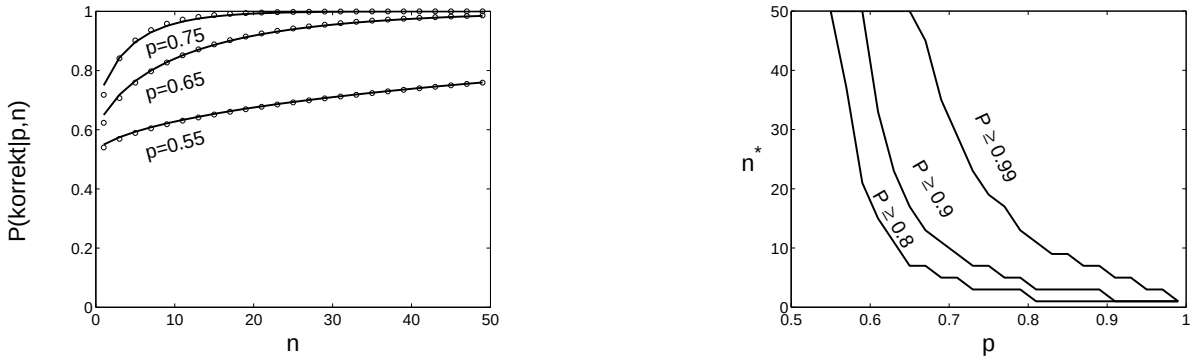

Abbildung 3.2: *Bit-Übertragung mit dem Majoritätsverfahren.*
Linke Abbildung: Wahrscheinlichkeit für korrekte Übertragung als Funktion der Wiederholungen n für verschiedene Einzelwahrscheinlichkeiten p. Die Kreise stehen für die Ergebnisse mit der de-Moivreschen Integralformel.
Rechte Abbildung: Benötigte Zahl der Wiederholungen als Funktion von p für unterschiedliche Zuverlässigkeiten $P \geq P^*$, die beim Majoritätsverfahren erzielt werden sollen.

Zellen verteilt werden, wobei es nur auf die Besetzungszahlen der Zellen ankommt und nicht auf die beteiligten 'Individuen'. Diese Situation liegt auch bei statistischen Untersuchungen von Verkehrsunfällen, Geburtstagen, etc. vor. Die Aufgabe impliziert wieder, dass Kugeln/Teilchen auf Zellen verteilt werden. Die Reihenfolge, in der die Kugeln in die Zellen eintreffen ist hierbei jedoch irrelevant. Wir führen hierzu die BESETZUNGSZAHLEN $\{n_i\} = \{n_1, n_2, \ldots, n_k\}$ ein, wobei n_i angibt, wieviele Teilchen sich in der Zelle i befinden. Die Gesamtzahl der Teilchen ist fixiert

$$\sum_i n_i = N \qquad . \tag{3.15}$$

Mit identischen Teilchen sind zwei Verteilungen nur dann unterschiedlich, wenn sich die k-Tupel $\{n_i\}$ der Besetzungszahlen unterscheiden. Die Zahl der unterscheidbaren Verteilungen von N identischen Teilchen auf k Zellen[3] ist

$$A_{N,k} = \binom{N+k-1}{N} = \binom{N+k-1}{k-1} \tag{3.16}$$

Beweis: Wir stellen die Teilchen durch Kreise ◯ und die Begrenzungen der Zellen durch senkrechte Striche $|$ dar. Zum Beispiel besagt die Darstellung

$$\| \bigcirc\bigcirc\bigcirc | \quad | \bigcirc\bigcirc | \bigcirc | \bigcirc\bigcirc\bigcirc\bigcirc \| \qquad ,$$

[3]Das ist gleichzeitig die Zahl der unterschiedlichen Lösungen von Gl. (3.15) für natürliche n_i.

dass wir insgesamt $N = 10$ Teilchen auf $k = 5$ Zellen mit den Besetzungszahlen $3, 0, 2, 1, 4$ verteilt haben. Die äußeren Begrenzungen $\|$ können hierbei nicht verändert werden. Wir erhalten alle möglichen Besetzungszahlverteilungen der Teilchen auf die Zellen, wenn wir die N Kreise und $k-1$ inneren Zellbegrenzungen ($|$) auf die dazu zur Verfügung stehenden $N+k-1$ Plätze verteilen. Auf diese Weise wird klar, dass wir das obige Ergebnis für $A_{N,k}$ erhalten.
Beweis von Gl. (3.6b): Das Auswählen einer Unterpopulation der Größe r einer Population der Größe n mit Zurücklegen kann man sich auch als das Aufteilen von r Teilchen auf n Zellen vorstellen. Die Anzahl der Teilchen in einer Zelle entspricht der Häufigkeit des Auftretens des entsprechenden Elements. Jedes Element kommt zwischen 0- und r-mal vor. Insgesamt werden genau r Elemente gewählt. Nach Gl. (3.16) erhalten wir damit Gl. (3.6b).

b) Die Multinomial-Verteilung vereinfacht sich für gleich-wahrscheinliche Ereignisse $p_\alpha = 1/k$ zu

BOLTZMANN-VERTEILUNG

$$P_B(\{n_i\}|N,k) = \binom{N}{\{n_i\}} k^{-N} \tag{3.17a}$$

$$\langle n_i \rangle = \frac{N}{k} \tag{3.17b}$$

$$\mathrm{var}(n_i) = N\,\frac{1}{k}(1-\frac{1}{k}) \tag{3.17c}$$

$$\mathrm{cov}(n_i,n_j) = N\,\frac{1}{k}(\delta_{ij}-\frac{1}{k}) \quad . \tag{3.17d}$$

Das ist ein Ergebnis, das in der Statistischen Physik Boltzmann-Verteilung genannt wird. Dieses Ergebnis ist eine Verallgemeinerung des Beispiels, bei dem zwei Würfel geworfen werden und nach der Wahrscheinlichkeit für die Summe 7 gefragt wird.
Es war eine große Überraschung als man in der Physik erkennen musste, dass die Boltzmann-Verteilung für identische Teilchen nicht korrekt ist; die k^N möglichen Verteilungen der Teilchen auf die Zellen haben nicht dieselbe a-priori Wahrscheinlichkeit. Es gibt zwei Arten von Teilchen, Bosonen und Fermionen, die sich in der sogenannten TEILCHEN-STATISTIK unterscheiden. Von Bosonen kann man beliebig viele in eine Zelle packen, von Fermionen hingegen maximal eins.
Beide Teilchen-Arten haben gemeinsam, dass sie absolut ununterscheidbar sind. Was bedeutet es, wenn man N UNUNTERSCHEIDBARE Teilchen auf k Zellen verteilen will? Die Anzahl k^N kam dadurch zustande, dass wir zunächst das erste Teilchen auf eine der k Zellen verteilt und anschließend das zweite Teilchen und so weiter. Das heißt

z.B., dass die Konfiguration mit dem ersten Teilchen in Zelle i und dem zweiten Teilchen in Zelle j eine andere Konfiguration darstellt als die, bei der das erste Teilchen in Zelle j und das zweite in Zelle i ist. Diesen Unterschied gibt es bei UNUNTERSCHEIDBAREN TEILCHEN nicht mehr. Es kommt nun nur noch auf die Besetzungszahlen an. Für Bosonen, bei denen beliebige Besetzungszahlen erlaubt sind, ist die Gesamtzahl der unterschiedlichen Verteilungen von N Teilchen auf k Zellen durch $A_{N,k}$ (Gl. (3.16)) gegeben. All diese Konfigurationen haben dieselbe Wahrscheinlichkeit $1/A_{N,k}$.
Bei Fermionen sind nun nur Konfigurationen mit maximal einem Teilchen pro Zelle erlaubt. Bei N Teilchen und k Zellen gibt es N Zellen mit einem und $k-N$ Zellen mit Null Teilchen. Es gibt $\binom{k}{N}$ unterschiedliche Verteilungen der N Teilchen auf die k Zellen. Diese Verteilungen haben alle die Wahrscheinlichkeit $P = 1/\binom{k}{N}$.
Der Sachverhalt ist in Tabelle (3.6) für drei Zellen und zwei Teilchen illustriert.
Zusammenfassend: Die Wahrscheinlichkeit, dass sich in Zelle i ($i = 1, 2, \ldots, k$) n_i Teilchen befinden (wobei $\sum_i n_i = N$) ist

$$
\begin{aligned}
P &= \binom{N}{\{n_i\}} k^{-N} && \text{Boltzmann} \\
P &= \frac{N!(k-1)!}{(N+k-1)!} && \text{Bose-Einstein} \\
P &= \frac{N!(k-N)!}{k!} && \text{Fermi} \qquad ,
\end{aligned}
$$

wobei im letzten Fall vorausgesetzt wurde, dass $n_\alpha \in \{0, 1\}$.
Das soll an einem Zahlenbeispiel verdeutlicht werden. Wir nehmen zwei Würfel, d.h. $N = 2$, $k = 6$, wobei die Augenzahlen als Zellen interpretiert werden und die Würfel als Teilchen. Was ist die Wahrscheinlichkeit dafür, dass die Augenzahl in der Summe 7 ergibt? Die günstigen Ereignisse hierfür sind $(1,6), (6,1), (2,5), (5,2), (3,4), (4,3)$. Die zugehörigen Besetzungszahlen sind $\{1,0,0,0,0,1\}$ für $(1,6)$ bzw. $(6,1)$ und analog für die anderen Wertepaare. Gemäß der Boltzmann-Statistik ist die Wahrscheinlichkeit

$$
P = \frac{2!}{1!0!0!0!0!1!} 6^{-2} + \frac{2!}{0!1!0!0!1!0!} 6^{-2} + \frac{2!}{0!0!1!1!0!0!} 6^{-2} = 3\frac{2}{36} = \frac{1}{6} \qquad .
$$

Für Bosonen hingegen ist die Wahrscheinlichkeit

$$
P = 3\frac{2!5!}{7!} = \frac{3*2}{6*7} = \frac{1}{7} \qquad .
$$

Für Fermionen erhalten wir schließlich

$$
P = 3\frac{2!4!}{6!} = \frac{2*3}{5*6} = \frac{1}{5} \qquad .
$$

Dieses Beispiele verdeutlichen noch einmal die Problematik der Gleichwahrscheinlichkeit.

3.5 Geometrische und hypergeometrische Verteilung

Viele kombinatorische Probleme haben mit einer Population zu tun, in der sich zwei Typen von Elementen befinden. Z.B. Kugeln unterschiedlicher Farben (rot, schwarz), Teilchen unterschiedlichen Spins (spin-up, spin-down) oder, wie im Fall der Qualitätskontrolle, defekte und intakte Produkte. In der Population der Größe n gebe es n_I Elemente vom ersten und $n_{II} = n - n_I$ Elemente vom zweiten Typ.
Es gibt hierzu zwei interessierende Fragen:

1. Es werden Elemente nacheinander (mit oder ohne Zurücklegen) herausgenommen. Wie groß ist die Wahrscheinlichkeit, dass erst beim k-ten Zug ein Element vom Typ II beobachtet wird?

2. Es wird eine geordnete Stichprobe (mit oder ohne Zurücklegen) vom Umfang k entnommen. Wie groß ist die Wahrscheinlichkeit, dass in der Stichprobe k_I Elemente vom ersten und $k_{II} = k - k_I$ Elemente vom zweiten Typ vorkommen?

Beide Fragestellungen sind an sich bereits interessant und kommen in unterschiedlichen Anwendungen vor. Insbesondere benötigt man diese Wahrscheinlichkeiten aber auch, um aus der Stichprobe im inversen Schluss auf die Population zu schließen.

3.5.1 Fragestellung 1 ohne Zurücklegen

Die Zahl der günstigen Ereignisse ergibt sich aus der Zahl der geordneten Stichproben (o.Z.) vom Umfang $k-1$ aus den n_I Elementen vom Typ I. Diese Zahl ist gemäß Gl. (3.3b) $\frac{n_I!}{(n_I-(k-1))!}$. Für den k-ten Zug gibt es n_{II} günstige Ereignisse. Die Gesamtzahl der Ereignisse ist die Zahl der geordneten Stichproben (o.Z.) vom Umfang k aus einer Population mit n Elementen, also $\frac{n!}{(n-k)!}$. Die gesuchte Wahrscheinlichkeit ist also

$$\frac{n_I!\,(n-k)!\,n_{II}}{n!(n_I-(k-1))!} = \frac{\binom{n_I}{k-1}\,n_{II}}{\binom{n}{k}\,k} \quad .$$

3.5.2 Fragestellung 1 mit Zurücklegen

Hier ist die Berechnung analog. Wir müssen lediglich den Ausdruck Gl. (3.3a) für geordnete Stichproben mit Zurücklegen verwenden. Für die ersten $k-1$ Elemente der Stichprobe gibt es demnach ${n_I}^{k-1}$ günstige Ereignisse. Für den k-ten Zug ist die Zahl weiterhin n_{II}. Die Gesamtzahl der Ereignisse beträgt nun n^k. Die gesuchte Wahrscheinlichkeit ist also

$$\frac{{n_I}^{k-1}\,n_{II}}{n^k} = \left(\frac{n_I}{n}\right)^{k-1}\frac{n_{II}}{n} \quad .$$

Dieses Ergebnis kann auch anders hergeleitet werden. Die Wahrscheinlichkeit, ein Element vom Typ I/II aus der Population zu ziehen ist $p_I = \frac{n_I}{n}$, bzw. $p_{II} = \frac{n_{II}}{n}$. Es gilt

natürlich $p_{II} = 1 - p_I$. Wie wir später zeigen werden, ist die Wahrscheinlichkeit, beim unabhängigen Ziehen ein Element vom Typ I $(k_I = k - 1)$-mal zu entnehmen durch $p_I^{k_I}$ gegeben, und wir erhalten ebenfalls für die gesuchte Wahrscheinlichkeit

GEOMETRISCHE VERTEILUNG

$$P(k_I|p_I) = p_I^{k_I}(1 - p_I) \tag{3.19a}$$

$$\langle k_I \rangle = \frac{p_I}{1 - p_I} \tag{3.19b}$$

$$\mathrm{var}(k_I) = \frac{p_I}{(1 - p_I)^2} \quad . \tag{3.19c}$$

Diese Verteilung nennt man GEOMETRISCHE VERTEILUNG. Man vergewissert sich leicht, dass diese Verteilung auf eins normiert ist, da $\sum_{k=0}^{\infty} p^k = 1/(1-p)$.

Die Formeln (3.19b) und (3.19c) für Mittelwert und Varianz kann man leicht zeigen:

$$\begin{aligned}
\langle k_I \rangle &= \sum_{k_I=0}^{\infty} k_I \; p_I^{k_I} \; (1 - p_I) \\
&= \sum_{k_I=0}^{\infty} k_I \; p_I^{k_I} - \sum_{k_I=0}^{\infty} k_I \; p_I^{k_I+1} \\
&= p_I \; \frac{d}{dp_I} \sum_{k_I=0}^{\infty} p_I^{k_I} - p_I^2 \; \frac{d}{dp_I} \sum_{k_I=0}^{\infty} p_I^{k_I} \\
&= (p_I - p_I^2) \frac{d}{dp_I} \frac{1}{1 - p_I} \\
&= \frac{p_I}{1 - p_I}
\end{aligned}$$

$$\begin{aligned}
\langle k_I^2 \rangle &= \sum_{k_I=0}^{\infty} k_I^2 \; p_I^{k_I} \; (1 - p_I) \\
&= p_I \; \frac{d}{dp_i} p_I \; \frac{d}{dp_i} \sum_{k_I=0}^{\infty} p_I^{k_I} - p_I^2 \; \frac{d}{dp_i} p_I \; \frac{d}{dp_i} \sum_{k_I=0}^{\infty} p_I^{k_I} \\
&= \frac{p_I(1 + p_I)}{(1 - p_I)^2}
\end{aligned}$$

$$\begin{aligned}\mathrm{var}(k_I) &= \left\langle k_I^2 \right\rangle - \langle k_I \rangle^2 \\ &= \frac{p_I}{(1-p_I)^2}\end{aligned}$$

3.5.3 Fragestellung 2 ohne Zurücklegen

Wir suchen die Wahrscheinlichkeit, dass die Stichprobe k_I Elemente vom Typ I und $k_{II} = k - k_I$ Elemente vom Typ II enthält. Hierbei kann k_I irgendeine ganze Zahl zwischen Null und $\min(n_I, k)$ sein. Die k_I Elemente vom Typ I können auf $\binom{n_I}{k_I}$ Arten gezogen werden. Für die Elemente vom Typ II gibt es $\binom{n_{II}}{k_{II}}$ Möglichkeiten. Die Gesamtzahl der Möglichkeiten k Elemente aus der Population der Größe $n = n_I + n_{II}$ auszuwählen ist $\binom{n}{k}$. Die gesuchte Wahrscheinlichkeit ist somit

HYPERGEOMETRISCHE VERTEILUNG

$$P(k_I|k = k_I + k_{II}, n_I, n_{II}) = \frac{\binom{n_I}{k_I}\binom{n_{II}}{k_{II}}}{\binom{n_I+n_{II}}{k_I+k_{II}}} \tag{3.20a}$$

$$\langle k_I \rangle = \frac{n_I\,(k_I + k_{II})}{n_I + n_{II}} \tag{3.20b}$$

$$\mathrm{var}(k_I) = \frac{(k_I + k_{II})(n_I + n_{II} - k_I - k_{II})\,n_I\,n_{II}}{(n_I + n_{II} - 1)(n_I + n_{II})^2} \quad . \tag{3.20c}$$

Beweis: Laut Bronstein gilt

$$\binom{\alpha}{0}\binom{\beta}{k} + \binom{\alpha}{1}\binom{\beta}{k-1} + \ldots + \binom{\alpha}{k}\binom{\beta}{0} = \binom{\alpha+\beta}{k} \quad ,$$

womit mit $\alpha = n_I$ und $\beta = n_{II}$ sofort die Normierung von Gl. (3.20a) folgt. Die Beweise für den Erwartungswert und die Varianz lassen sich leicht führen.

$$\begin{aligned}\langle k_I \rangle &= \sum_{k_I=0}^{k} k_I \frac{\binom{n_I}{k_I}\binom{n_{II}}{k-k_I}}{\binom{n_I+n_{II}}{k}} \\ &= \sum_{k_I=0}^{k} k_I \frac{n_I!}{k_I!\,(n_I-k_I)!} \frac{n_{II}!}{(k-k_I)!\,(n_{II}-k+k_I)!} \frac{k!\,(n_I+n_{II}-k)!}{(n_I+n_{II})!}\end{aligned}$$

$$= \sum_{k_I=1}^{k} n_I \, \frac{(n_I-1)!}{(k_I-1)!\,[(n_I-1)-(k_I-1)]!} \, \frac{n_{II}!}{(k-k_I)!\,(n_{II}-k+k_I)!}$$

$$\times \frac{k}{n_I+n_{II}} \, \frac{(k-1)!\,[(n_I+n_{II}-1)-(k-1)]!}{(n_I+n_{II}-1)!}$$

$$= \frac{n_I\,k}{n_I+n_{II}} \sum_{k_I=1}^{k} \frac{\binom{n_I-1}{k_I-1}\binom{n_{II}}{k-k_I}}{\binom{n_I+n_{II}-1}{k-1}}$$

ersetze den Summationsindex $k_I \to k_I + 1$

$$= \frac{n_I\,k}{n_I+n_{II}} \sum_{k_I=0}^{k-1} \frac{\binom{n_I-1}{k_I}\binom{n_{II}}{k-k_I-1}}{\binom{n_I+n_{II}-1}{k-1}}$$

definiere: $\tilde{k} := k-1$, $\tilde{n}_I := n_I - 1$

$$= \frac{n_I\,k}{n_I+n_{II}} \underbrace{\sum_{k_I=0}^{\tilde{k}} \frac{\binom{\tilde{n}_I}{k_I}\binom{n_{II}}{\tilde{k}-k_I}}{\binom{\tilde{n}_I+n_{II}}{\tilde{k}}}}_{=1}$$

$$= \frac{n_I\,(k_I+k_{II})}{n_I+n_{II}}$$

Für $\langle k_I^2 \rangle$ verfährt man analog:

$$\langle k_I^2 \rangle = \sum_{k_I=0}^{k} k_I^2 \, \frac{\binom{n_I}{k_I}\binom{n_{II}}{k-k_I}}{\binom{n_I+n_{II}}{k}}$$

$$= \sum_{k_I=0}^{k} k_I^2 \, \frac{n_I!}{k_I!\,(n_I-k_I)!} \, \frac{n_{II}!}{(k-k_I)!\,(n_{II}-k+k_I)!} \, \frac{k!\,(n_I+n_{II}-k)!}{(n_I+n_{II})!}$$

$$= \sum_{k_I=1}^{k} k_I\,n_I \, \frac{(n_I-1)!}{(k_I-1)!\,[(n_I-1)-(k_I-1)]!} \, \frac{n_{II}!}{(k-k_I)!\,(n_{II}-k+k_I)!}$$

$$\times \frac{k}{n_I+n_{II}} \, \frac{(k-1)!\,[(n_I+n_{II}-1)-(k-1)]!}{(n_I+n_{II}-1)!}$$

$$= \frac{n_I\,k}{n_I+n_{II}} \sum_{k_I=1}^{k} k_I \, \frac{\binom{n_I-1}{k_I-1}\binom{n_{II}}{k-k_I}}{\binom{n_I+n_{II}-1}{k-1}}$$

ersetze den Summationsindex $k_I \to k_I + 1$

$$= \frac{n_I\,k}{n_I+n_{II}} \sum_{k_I=0}^{k-1} (k_I+1) \, \frac{\binom{n_I-1}{k_I}\binom{n_{II}}{k-k_I-1}}{\binom{n_I+n_{II}-1}{k-1}}$$

definiere: $\tilde{k} := k-1$, $\tilde{n}_I := n_I - 1$

$$
\begin{aligned}
&= \frac{n_I\, k}{n_I + n_{II}} \sum_{k_I=0}^{\tilde{k}} (k_I+1)\, \frac{\binom{\tilde{n}_I}{k_I}\binom{n_{II}}{\tilde{k}-k_I}}{\binom{\tilde{n}_I+n_{II}}{\tilde{k}}} \\
&= \frac{n_I\, k}{n_I + n_{II}} \left(1 + \underbrace{\sum_{k_I=0}^{\tilde{k}} k_I\, \frac{\binom{\tilde{n}_I}{k_I}\binom{n_{II}}{\tilde{k}-k_I}}{\binom{\tilde{n}_I+n_{II}}{\tilde{k}}}}_{=\frac{\tilde{k}\,\tilde{n}_I}{\tilde{n}_I+n_{II}}} \right) \\
&= \frac{n_I\, k}{n_I + n_{II}} \left(1 + \frac{(k-1)\,(n_I-1)}{n_1 + n_{II} - 1} \right)
\end{aligned}
$$

Damit können wir leicht die Varianz berechnen:

$$
\begin{aligned}
\mathrm{var}(k_I) &= \left\langle k_I^2 \right\rangle - \left\langle k_I \right\rangle^2 \\
&= \frac{n_I\, k}{n_I + n_{II}} + \frac{n_I\, k\,(k-1)\,(n_I-1)}{(n_I+n_{II})\,(n_1+n_{II}-1)} - \frac{n_I^2\, k^2}{(n_I+n_{II})^2} \\
&= \frac{k\,(n_I+n_{II}-k)\, n_I\, n_{II}}{(n_I+n_{II})^2\,(n_1+n_{II}-1)} \\
&= \frac{(k_I+k_{II})\,(n_I+n_{II}-k_I-k_{II})\, n_I\, n_{II}}{(n_I+n_{II})^2\,(n_1+n_{II}-1)}
\end{aligned}
$$

Beispiele:

a) Qualitätskontrolle:
Wenn in einer Produktion von n Teilen n_I defekt und $n_{II} = n - n_I$ intakt sind, wie groß ist die Wahrscheinlichkeit, dass bei einer Stichprobe vom Umfang k die Zahl der defekten Teile k_I beträgt. Die gesuchte Wahrscheinlichkeit ist die hypergeometrische Verteilung $P(k_I|k, n_I, n_{II})$. Die wirklich interessante Frage lautet jedoch, wie groß ist die Zahl n_I der defekten Teile der gesamte Produktion vom Umfang n, wenn bei Stichprobe vom Umfang k die Zahl der defekten Teilen k_I beträgt. Diese Frage werden wir mit dem Bayesschen Theorem beantworten können.

b) Abschätzprobleme:
Es werden n_I Fische gefangen, rot markiert und wieder freigelassen. Nach einiger Zeit werden erneut Fische gefangen, diesmal k Stück, davon sind k_I Fische rot markiert. Wie groß ist die Zahl der Fische im See? Dieses inverse Problem können wir ebenfalls noch nicht beantworten. Wir können aber die umgekehrte Frage beantworten, die bei der Lösung des inversen Problems eine wichtige Rolle spielt: Wie groß ist die Wahrscheinlichkeit, dass der zweite Fang k_I markierte Fische enthält, vorausgesetzt, die Zahl der Fische im See ist n und davon sind n_I rot markiert. Diese Wahrscheinlichkeit ist ebenfalls durch die hypergeometrische Verteilung $P(k_I|k, n_I, n-n_I)$ gegeben.

3.5.4 Fragestellung 2 mit Zurücklegen

Diese Frage ist leicht zu beantworten, wenn wir von den Wahrscheinlichkeiten $p_I = \frac{n_I}{n}$ bzw. $p_{II} = \frac{n_{II}}{n}$ ausgehen, Elemente vom Typ I oder II unabhängig voneinander zu ziehen. Die Wahrscheinlichkeit, dass die Stichprobe k_I Elemente vom Type I und k_{II} Elemente vom Typ II in einer ganz bestimmten Reihenfolge enthält, ist demnach

$$ {p_I}^{k_I}\ {p_{II}}^{k_{II}} = {p_I}^{k_I}\ (1-p_I)^{k-k_I} \qquad . $$

Da es in der Fragestellung nicht auf die Reihenfolge ankommt, ist die gesuchte Wahrscheinlichkeit

$$ \binom{k}{k_I} {p_I}^{k_I}\ (1-p_I)^{k-k_I} \qquad . $$

Das ist nichts anderes als die BINOMIAL-VERTEILUNG.

	Mit Zurücklegen		Ohne Zurücklegen
1	(a_1, a_1)	1	(a_1, a_2)
2	(a_1, a_2)	2	(a_1, a_3)
3	(a_1, a_3)	3	(a_2, a_1)
4	(a_2, a_1)	4	(a_2, a_3)
5	(a_2, a_2)	5	(a_3, a_1)
6	(a_2, a_3)	6	(a_3, a_2)
7	(a_3, a_1)		
8	(a_3, a_2)		
9	(a_3, a_3)		

Tabelle 3.3: Geordnete Stichproben vom Umfang 2 aus der Population a_1, a_2, a_3.

1	(a_1, a_2, a_3)
2	(a_2, a_3, a_1)
3	(a_3, a_1, a_2)
4	(a_3, a_2, a_1)
5	(a_2, a_1, a_3)
6	(a_1, a_3, a_2)

Tabelle 3.4: Permutationen der Elemente der Population a_1, a_2, a_3.

1	(a_1, a_2)
2	(a_1, a_3)
3	(a_2, a_3)

Tabelle 3.5: Unterpopulationen der Größe 2 aus der Population a_1, a_2, a_3.

BOLTZMANN STATISTIK

	1	2	3	4	5	6	7	8	9
Z_1	a,b			a	b	a	b		
Z_2		a,b		b	a			a	b
Z_3			a,b			b	a	b	a

BOSE-EINSTEIN-STATISTIK

	1	2	3	4	5	6
Z_1	a,a			a	a	
Z_2		a,a		a		a
Z_3			a,a		a	a

FERMI-STATISTIK

	1	2	3
Z_1	a	a	
Z_2	a		a
Z_3		a	a

Tabelle 3.6: Vergleich der Teilchen-Statistiken.

Kapitel 4

Grenzwertsätze

4.1 Stirlingsche Formel

Wichtig im Zusammenhang mit kombinatorischen Problemen ist die Stirlingsche Formel zur approximativen Bestimmung von $n!$, da in vielen Fällen große Zahlen n auftreten. Wir werden später sehen, dass wir neben der Fakultät auch die Gamma-Funktion benötigen,

$$\Gamma(x) = \int_0^\infty t^{x-1}\, e^{-t}\, dt \qquad , \tag{4.1}$$

die für ganzzahlige Argumente mit der Fakultät über

$$\Gamma(n+1) = n! \tag{4.2}$$

zusammenhängt. Ebenso wie die Fakultät erfüllt auch die Gamma-Funktion (für beliebige reelle Argumente) die Rekursionsformel

$$\Gamma(x+1) = x\,\Gamma(x) \qquad . \tag{4.3}$$

Die asymptotische Darstellung der Gamma-Funktion für große Werte von $|x|$ lautet

$$\Gamma(x) = x^{x-\frac{1}{2}} e^{-x} \sqrt{2\pi} \left\{ 1 + \frac{1}{12x} + O(x^{-2}) \right\} = x^{x-\frac{1}{2}} e^{-x} \sqrt{2\pi} \left\{ 1 + O(x^{-1}) \right\} . \tag{4.4}$$

Für die Fakultät erhält man daraus für große n

$$\begin{aligned} n! = \Gamma(n+1) &= (n+1)^{n+\frac{1}{2}} e^{-(n+1)} \sqrt{2\pi} \left\{ 1 + O(n^{-1}) \right\} \\ &= e^{(n+\frac{1}{2})\ln(n+1) - n - 1} \sqrt{2\pi} \left\{ 1 + O(n^{-1}) \right\} \\ &= e^{(n+\frac{1}{2})\left(\ln(n) + \ln(1+1/n) \right) - n - 1} \sqrt{2\pi} \left\{ 1 + O(n^{-1}) \right\} \\ &= e^{(n+\frac{1}{2})\left(\ln(n) + 1/n + O(1/n^2) \right) - n - 1} \sqrt{2\pi} \left\{ 1 + O(n^{-1}) \right\} \end{aligned}$$

$$= \quad e^{(n+\frac{1}{2})\ln(n)}\; e^{1+O(1/n)-n-1}\; \sqrt{2\pi}\; \left\{1+O(n^{-1})\right\} \qquad .$$

Das bringt uns zur gesuchten Formel

STIRLINGSCHE FORMEL

$$n! = n^{(n+\frac{1}{2})}\; e^{-n}\; \sqrt{2\pi}\; \left\{1+O(n^{-1})\right\} \tag{4.5a}$$

$$\ln(n!) = (n+\frac{1}{2})\ln(n) - n + \ln(\sqrt{2\pi}) + O(n^{-1}) \qquad . \tag{4.5b}$$

Der Term $O(n^{-1})$ kann für große n vernachlässigt werden. Das führt in der linearen Darstellung (Gl. (4.5a)) zu einem zwar über alle Grenzen anwachsenden Absolut-Fehler, aber der relativer Fehler verschwindet wie $1/n$. In den Anwendungen werden i.d.R. Verhältnisse von Fakultäten benötigt, so dass hier auch der absolute Fehler mit $1/n$ verschwindet. In der logarithmischen Darstellung verschwindet bereits der Absolut-Fehler. In Tabelle (4.1) ist der relative Fehler für einige Werte von n angegeben. Man erkennt in der Tat, dass der relative Fehler sehr schnell klein wird und die Näherung bereits für $n = 1$ plausible Werte liefert[1].

n	n!	ε
1	1	0.0779
2	2	0.0405
3	6	0.0273
4	24	0.0206
5	120	0.0165
10	3628800	0.0083
20	2432902008176640000	0.0042
30	265252859812191058636308480000000	0.0028
40	815915283247897734345611269596115894272000000000	0.0021

Tabelle 4.1: Relativer Fehler ε der Stirlingschen Formel.

4.2 Lokaler Grenzwertsatz (de Moivre)

Es kann gezeigt werden (siehe Lehrbuch von H.Meschkowski), dass die Binomial-Verteilung (Newtonsche Formel) für $np(1-p) \gg 1$ durch eine Gauß-Funktion appro-

[1] Die Stirlingsche Formel liefert für $n = 1$ den Wert 0.922

ximiert werden kann

DE-MOIVRE-LAPLACE THEOREM
(LOKALER GRENZWERTSATZ)

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k} \backsimeq g(k|k_0,\sigma) \tag{4.6}$$

mit

$$g(k|k_0,\sigma) := \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(k-k_0)^2}{2\sigma^2}} \tag{4.7}$$

$$k_0 = np$$

$$\sigma = \sqrt{np(1-p)}$$

für $\sigma \gg 1$.

Dieser Satz besagt präzise formuliert:
Zu jeder reellen Zahl ε und zu jedem Paar positiver reeller Zahl x_1 und x_2 gibt es eine natürliche Zahl $N(\varepsilon, x_1, x_2)$, so dass

$$\left| \frac{P(k|n,p)}{g(k|np, \sqrt{np(1-p)})} - 1 \right| < \varepsilon$$

wenn $n > N(\varepsilon, x_1, x_2)$ und wenn zudem k so gewählt wird, dass

$$\frac{|k-np|}{\sqrt{np(1-p)}} \in [x_1, x_2] \quad .$$

Die Beschränkung von $\frac{|k-np|}{\sqrt{np(1-p)}}$ auf $[x_1, x_2]$ bedeutet, dass die Approximation für extrem große oder kleine k unbrauchbar ist. Für festes $k = c$ gilt ja offensichtlich $\lim_{n\to\infty} \frac{|k-np|}{\sqrt{np(1-p)}} = \infty$. Man sieht, dass selbst in der Nähe des Maximums große Abweichungen (im Prozentbereich) auftreten. Die Grenzwertsätze haben für die praktische Anwendung an Bedeutung verloren, da man die Binomial-Verteilung numerisch genauso bequem auswerten kann, wie die Gauß-Funktion. Wir werden allerdings Situation antreffen, in denen wir diese Näherung benötigen, um analytische Auswertungen zu ermöglichen.

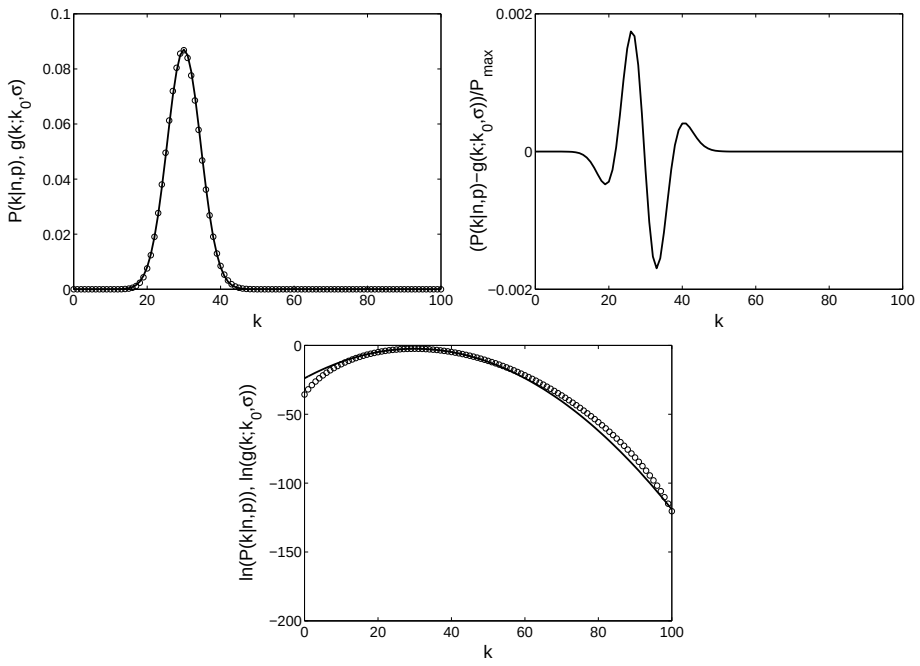

Abbildung 4.1: *Links: Vergleich der Binomial Verteilung (Kreise) mit der Gaußschen Näherung (de-Moivre-Laplace) (durchgezogene Kurve) für* $n = 100$, $p = 0.3$. *Für diese Parameter ist* $k_0 = 30$ *und* $\sigma = 4.58$.
Rechts: Differenz „Binomial-Gauß" normiert auf den Maximalwert der Binomial-Verteilung. Unten: Vergleich der Binomial-Verteilung (Kreise) mit der Gaußschen Näherung (de-Moivre-Laplace) (durchgezogene Kurve) auf logarithmischer Skala.

4.3 Integralsatz von de-Moivre

In der Regel interessiert man sich weniger für die Wahrscheinlichkeit, dass bei n Versuchen genau k-mal ein bestimmtes Ereignis eintritt, sondern eher für Fragen folgender Art:
In einer Produktion von Bauelementen sei $p = 0.002$ *die Wahrscheinlichkeit dafür, dass ein Teil fehlerhaft ist. Wie groß ist die Wahrscheinlichkeit, dass in einer Produktion von 5000 Stück höchstens 20 unbrauchbar sind?*
Hier fragt man nach der Wahrscheinlichkeit, dass die exklusiven Ereignisse $k = 0, 1, 2, \ldots, 20$ vorkommen. Nach der Summenregel ist diese Wahrscheinlichkeit

$$P(k \in \{0, 1, 2, \ldots, 20\}|n, p) = \sum_{k=0}^{20} P(k|n, p) \qquad . \tag{4.8}$$

Auch diese Summe ist eine Trivialität für heutige Rechner. Dennoch kann es in manchen Fällen, insbesondere für analytische Rechnungen, sinnvoll sein, die de-Moivre-Laplace-Näherung zu verwenden. Das führt zum

INTEGRALSATZ VON DE-MOIVRE

Es sei p die Wahrscheinlichkeit für das Eintreten eines Ereignisses E in einem Versuch. Dieses Ereignis soll bei n unabhängigen Wiederholungen des Versuches k-mal auftreten. Wir definieren die Menge

$$I_{a,b} = \{k | a \leq \frac{k - np}{\sqrt{np(1-p)}} < b\} \qquad .$$

Es gilt die Grenzwertbeziehung

$$\lim_{n\to\infty} P(k \in I_{a,b}|n,p) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx \qquad . \tag{4.9}$$

Weiter gilt

$$P(k \in I_{a,b}|n,p) = \sum_{k \in I_{a,b}} P(k|n,p) \qquad .$$

In diesem Zusammenhang ist die Funktion

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \tag{4.10}$$

wichtig. Diese Funktion steigt monoton an und hat folgende speziellen Funktionswerte

$$\begin{aligned} \Phi(-\infty) &= 0 \\ \Phi(0) &= 1/2 \\ \Phi(\infty) &= 1 \qquad . \end{aligned} \tag{4.11}$$

Mit der Funktion $\Phi(x)$ gilt

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx = \Phi(b) - \Phi(a) \qquad .$$

Eine weitere wichtige Beziehungen ist

$$\Phi(-x) = 1 - \Phi(x) \qquad . \tag{4.12}$$

Wir können nun die eingangs gestellte Aufgabe mit dem Integralsatz näherungsweise lösen. Hierbei ist $np = 10$ und $np(1-p) = 5000 * 0.002 * .998 = 9.98$, bzw. $\sqrt{np(1-p)} = 3.159$. Damit

$$\frac{k - np}{\sqrt{npq}} < b$$

ist, kann für die Obergrenze

$$\frac{k_{\max} - np}{\sqrt{npq}} < b < \frac{k_{\max} + 1 - np}{\sqrt{npq}}$$

gewählt werden. Die Summe über die 21 k-Werte $k = 0, 1, \ldots, 20$ wird durch ein Integral ersetzt. Damit die Intervall-Länge ebenfalls 21 beträgt, sollte man $b = (k_{\max} + 1 - np)/\sqrt{npq}$ wählen. Es erweist sich aber als besser, $b = (k_{\max} + 0.5 - np)/\sqrt{npq}$ zu verwenden. Für die Intervallgrenzen verwenden wir deshalb $a = (0-10)/3.159 = -3.165$ und $b = (20.5 - 10)/3.159 = 3.3237$ und erhalten für die gesuchte Wahrscheinlichkeit

$$P(k \in I_{a,b}) = \Phi(3.324) - \Phi(-3.165) = 0.9988.$$

Der exakte Wert ist 0.9984. Die Übereinstimmung ist gut. In der Abbildung ist der Vergleich für unterschiedliche $k_{\max}$ aufgetragen.

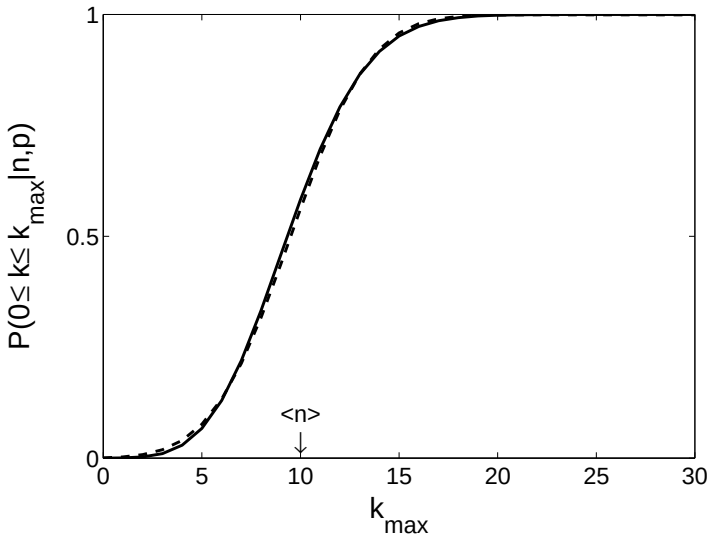

Abbildung 4.2: *Vergleich der exakten Werte (durchgezogene Kurve) für die Wahrscheinlichkeit $P(0 \le k \le k_{\max}|n,p)$ mit denen des Integralsatzes (gestrichelt) als Funktion von $k_{\max}$ für $n = 5000$, $p = 0.002$.*

In diesem Zusammenhang ist auch die FEHLERFUNKTION $\mathrm{erf}(x)$ zu erwähnen

$$\begin{aligned} \mathrm{erf}(x) &= \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{2}x} e^{-t^2/2}\, dt \\ \frac{1}{2}\mathrm{erf}(\frac{x}{\sqrt{2}}) &= \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\, dt \qquad . \end{aligned} \tag{4.13}$$

Demnach können wir den Integralsatz in der Fehlerfunktion wie folgt ausdrücken

$$P(k \in I_{a,b}|n,p) = \frac{1}{2}\left(\mathrm{erf}(\frac{b}{\sqrt{2}}) - \mathrm{erf}(\frac{a}{\sqrt{2}})\right) \quad . \tag{4.14}$$

4.4 Bernoullis Gesetz der großen Zahlen

Wir betrachten Bernoulli-Experimente, die durch eine Binomial-Verteilung beschrieben werden. Wir wissen bereits aus Gl. (3.8b), dass der Mittelwert bei n Versuchen durch $\langle x\rangle = p\ n$ gegeben ist. Wie groß ist die Wahrscheinlichkeit, dass bei n Versuchen ein Ereignis, das die Wahrscheinlichkeit p besitzt, $k = n\ p$ mal auftritt? Die Antwort ist offensichtlich in der de-Moivre-Laplace-Näherung

$$P(k = n\ p|n,p) = \frac{1}{\sqrt{2\pi np(1-p)}} \underset{n\to\infty}{\longrightarrow} 0 \quad . \tag{4.15}$$

Das heißt, der Mittelwert $\langle k\rangle = np$, der hier auch gleichzeitig der wahrscheinlichste Wert ist (siehe Abbildung 4.1), hat eine für $n \to \infty$ verschwindende Wahrscheinlichkeit. Das liegt natürlich daran, dass es sehr viele k-Werte in der Nähe von $k = \langle k\rangle$ gibt, die vergleichbare Wahrscheinlichkeit besitzen. Der de-Moivresche Integralsatz besagt, dass die Wahrscheinlichkeit, dass ein k-Wert aus dem Bereich

$$\begin{aligned} I_\sigma &= \{k \mid np - \sqrt{np(1-p)} \le k < np + \sqrt{np(1-p)}\} \\ &= \{k \mid \langle k\rangle - \sigma \le k < \langle k\rangle + \sigma\} \end{aligned}$$

auftritt (σ-Bereich), durch

$$P(k \in I_\sigma|n,p) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-\frac{x^2}{2}}\, dx = \mathrm{erf}(\frac{1}{\sqrt{2}}) = 0.6827 \tag{4.16}$$

gegeben ist. Die Wahrscheinlichkeit ein k aus dem „2-σ"-Bereich zu finden, ist entsprechend

$$P(k \in I_{2\sigma}|n,p) = \frac{1}{\sqrt{2\pi}} \int_{-2}^{2} e^{-\frac{x^2}{2}}\, dx = \mathrm{erf}(\sqrt{2}) = 0.9545 \quad .$$

Das bedeutet, wenn wir ein beliebig kleines $\varepsilon > 0$ vorgeben, ist die Wahrscheinlichkeit für $|k/n - p| < \varepsilon$

$$P(|k/n - p| < \varepsilon|n,p) = P(pn - \varepsilon n \le k < pn + \varepsilon n) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx$$

mit

$$\begin{aligned} a &= -\frac{\varepsilon n}{\sqrt{np(1-p)}} = -\sqrt{n}\,\frac{\varepsilon}{\sqrt{p(1-p)}} \underset{n\to\infty}{\longrightarrow} -\infty \\ b &= +\sqrt{n}\,\frac{\varepsilon}{\sqrt{p(1-p)}} \underset{n\to\infty}{\longrightarrow} \infty \quad . \end{aligned}$$

Daraus folgt

BERNOULLIS GESETZ DER GROSSEN ZAHLEN

$$P(|k/n - p| < \varepsilon | n, p) \underset{n\to\infty}{\longrightarrow} 1 \qquad \text{für bel. } \varepsilon > 0 \qquad . \tag{4.17}$$

Wenn wir also die intrinsische Wahrscheinlichkeit p aus der relativen Häufigkeit $\frac{k}{n}$ abschätzen, geht die Abweichung (Ungenauigkeit) mit $n \to \infty$ gegen Null.

4.5 Der Satz von Poisson

Die Näherung von de-Moivre-Laplace liefert brauchbare Werte in der Nähe des Maximums der Verteilung. Für sehr kleine oder sehr große Werte $O(n)$ von k wird der Fehler groß. Gerade in diesem Fall hilft ein Satz von Poisson.

SATZ VON POISSON

Es werden n Versuche durchgeführt, bei denen ein Ereignisse eintritt oder nicht. Die Wahrscheinlichkeit p für das Ereignis erfülle die Bedingung

$$n \cdot p = \mu = const \qquad . \tag{4.18}$$

Das heißt, das Ereignis tritt im Mittel μ-mal auf, unabhängig von der Zahl n der Versuche. Dann gilt

$$\lim_{n\to\infty} P(k|n, p = \frac{\mu}{n}) = e^{-\mu}\frac{\mu^k}{k!} =: P(k|\mu) \qquad . \tag{4.19}$$

Die resultierende Verteilung $P(k|\mu)$ heißt POISSON-VERTEILUNG. Offensichtlich ist die Poisson-Verteilung korrekt auf Eins normiert

$$\sum_{k=0}^{\infty} e^{-\mu}\frac{\mu^k}{k!} = e^{-\mu}e^{\mu} = 1 \qquad .$$

Die Bedingung Gl. (4.18) kann auf zwei Arten verstanden werden

1. $p \ll 1$ mit $np \not\gg 1$. Die Poisson-Verteilung wird als Näherung für die Binomial-Verteilung verwendet.

2. Wir betrachten ein Zeitintervall t, in dem im Mittel μ Ereignisse stattfinden. Wir unterteilen das Intervall in n gleich große Teile $\Delta t = t/n$. Die Wahrscheinlichkeit, dass in einem der Teilintervalle ein Ereignis stattfindet, sei p. Wenn die Ereignisse unabhängig voneinander sind, haben wir es wieder mit n Wiederholungen identischer Versuche zu tun, und wir können die Binomial-Verteilung verwenden, von der wir wissen, dass im Mittel pn Ereignisse auftreten werden. Damit diese Zahl unabhängig von der Zahl der Unterteilungen des gesamten Zeitintervalls ist, muss $p = \mu/n$ sein.

In der Abbildung 4.3 sind die Ergebnisse der Poissonschen Näherung mit denen der exakten Binomial-Verteilung für verschiedene Parameter verglichen. Die Übereinstimmung ist generell sehr gut. Zusätzlich enthalten die Abbildungen auch die Kurven der Gauß-Verteilungen mit denselben Werten für Mittelwert und Varianz. Für $\mu \gg 1$ geht die Poisson-Verteilung ebenfalls in eine Gauß-Verteilung über.

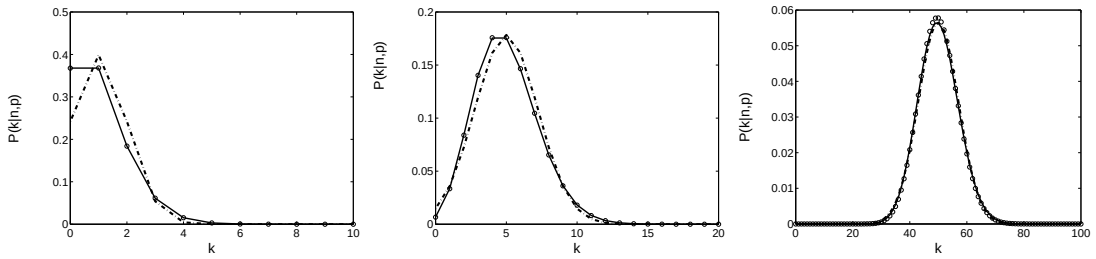

Abbildung 4.3: *Vergleich der Binomial Verteilung (Kreise) mit der Poissonschen Näherung (durchgezogene Kurve) für* $n = 1000$.
Links: $p = 0.001,\ \mu = 1,\ \sigma = 1.00$.
Mitte: $p = 0.005,\ \mu = 5,\ \sigma = 2.23$.
Rechts: $p = 0.05,\ \mu = 50,\ \sigma = 6.89$.
Zusätzlich wurde die Gauß-Funktion $g(x|\mu, \sqrt{\mu})$ *gestrichelt eingezeichnet.*

Der Erwartungswert der Poisson-Verteilung ist

$$\begin{aligned} \langle k \rangle &= \sum_{k=0}^{\infty} k\, e^{-\mu} \frac{\mu^k}{k!} & &= e^{-\mu}\, \mu\, \frac{\partial}{\partial \mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} \\ &= e^{-\mu}\, \mu\, \frac{\partial}{\partial \mu}\, e^{\mu} & &= \mu \qquad . \end{aligned}$$

Analog erhalten wir für das zweite Moment der Poisson-Verteilung

$$\begin{aligned} \left\langle k^2 \right\rangle &= \sum_{k=0}^{\infty} k^2\, e^{-\mu} \frac{\mu^k}{k!} = e^{-\mu}\, \mu\, \frac{\partial}{\partial \mu}\, \mu\, \frac{\partial}{\partial \mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} \\ &= e^{-\mu}\, \mu\, \frac{\partial}{\partial \mu}\, \mu\, \frac{\partial}{\partial \mu} e^{\mu} = e^{-\mu}\, \mu\, \left(\mu\, e^{\mu} + e^{\mu}\right) \\ &= \mu^2 + \mu \qquad . \end{aligned}$$

Somit haben wir

POISSON-VERTEILUNG

$$P(k|\mu) = e^{-\mu}\,\frac{\mu^k}{k!} \tag{4.20a}$$

$$\langle k\rangle = \mu \tag{4.20b}$$

$$\mathrm{var}(k) = \mu \quad . \tag{4.20c}$$

Erwartungswert und Varianz der Poisson-Verteilung stimmen mit den entsprechenden Werten der Binomial-Verteilung (Gl. (3.8b) und (4.20c)) überein, wenn man berücksichtigt, dass $\mu = p\,n$ und $p \to 0$.
Man erkennt in Abbildung 4.3 auch, dass bereits bei einem Mittelwert von $\mu = 5$ die Gaußsche Näherung gut mit der Poisson-Verteilung übereinstimmt.
Wichtig ist festzuhalten, dass die Gauß-Verteilung bereits bei kleinen Werten $\mu > 10$ sehr gut mit der Poisson-Verteilung übereinstimmt. Aber selbst bei $\mu = 50$ ist noch eine Diskrepanz zur Binomial-Verteilung zu erkennen, da in diesem Fall n nicht groß genug ist.
Viele Experimente sind Zählexperimente, z.B. inverse Photoemission, bei denen Teilchen gezählt werden. Zählexperimente unterliegen generell der Poisson-Statistik. Die Zählraten sind aber in der Regel sehr groß ($\mu > 100$), und es ist daher gerechtfertigt, die Streuung der experimentellen Werte um den wahren Wert mit einer Gauß-Verteilung zu beschreiben, bei der die Varianz gleich dem Messwert ist. Die Fehlerstatistik eines Zählexperimentes lautet demnach

FEHLERSTATISTIK VON ZÄHLEXPERIMENTEN

$$P(N|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\; e^{-\frac{(N-\mu)^2}{2\sigma^2}} \tag{4.21}$$
$$\sigma = \sqrt{\mu} \approx \sqrt{N}$$

Das bedeutet, die experimentellen Zählraten werden gemäß einer Gauß-Verteilung um den wahren Wert μ mit einer Breite $\sqrt{\mu}$ streuen. Da der wahre Wert μ nicht bekannt ist, sonst würde man das Experiment erst gar nicht durchführen, verwendet man diese Formel i.d.R. um von den beobachteten Zählraten auf den wahren Wert zu schließen. Solange die Zählrate $N > 10$, kann die Varianz durch die experimentelle

Zählrate approximiert werden, und man liest diese Verteilung umgekehrt. Man gibt

$$\mu = N \pm \sqrt{N} \tag{4.22}$$

als Schätzwert für den wahren Wert μ an.

Kapitel 5

Begriffsdefinitionen und Diskussion

Wir haben mittlerweile einige Erfahrung mit typischen Fragestellungen der Wahrscheinlichkeitstheorie und auch das nötige mathematische Rüstzeug, um die Begriffe, die wir in der Wahrscheinlichkeitstheorie verwenden werden, präzisieren zu können. Das soll anhand eines speziell konstruierten Beispiels geschehen.

5.1 Das Schätzexperiment mit drei Urnen

Gegeben sind drei Behälter (Urnen) U_1, U_2, U_3, in denen sich jeweils 100 Kugeln befinden. Die Kugeln sind entweder rot oder gelb. Die Urnen unterscheiden sich im Anteil q_α der gelben Kugeln

Urne	Anteil gelb	Anteil rot
U_1	$q_1 = 0.2$	0.8
U_2	$q_2 = 0.4$	0.6
U_3	$q_3 = 0.7$	0.3

AUFGABENSTELLUNG

- Es wird zufällig eine der drei Urnen U_α ausgewählt.
- Wir wissen nicht welche.
- Aus U_α werden 40 Kugeln mit Zurücklegen gezogen.
  Davon sind n_g Kugeln gelb.

Def. 5.1 (Bedingungskomplex) *Wir nennen diese Information und sämtliche Annahmen, die die Aufgabenstellung eindeutig definieren, den Bedingungskomplex $\mathcal{B}$.*

Der Bedingungskomplex $\mathcal{B}$ enthält im Falle des Drei-Urnen-Experiments die Information: es gibt drei Urnen, es wird gezogen mit Zurücklegen, alle Kugeln sind gleich groß und schwer[1], die Kugeln werden „zufällig" entnommen (ohne hinzusehen oder abzutasten), etc... . Erst die präzise Angabe des Bedingungskomplexes macht eine eindeutige Antwort möglich! Es ist nicht immer klar, was für den Bedingungskomplex relevant ist. Z.B. ist es wichtig, dass es sich um drei Urnen handelt, aber nicht welche Farbe diese Urnen haben, außer die Farben werden kommuniziert. Die Größe der Urnen könnte relevant sein, wenn davon die Wahrscheinlichkeit abhängt, dass diese Urne ausgewählt wurde.

FRAGESTELLUNGEN

Wie groß ist die WAHRSCHEINLICHKEIT, dass

- $\alpha = 1, 2, 3$.
- der nächste Zug wieder gelb liefert.
- sich eine Urne unbekannten Typs eingeschlichen hat.

Def. 5.2 (Versuch, Experiment) *Unter einem* VERSUCH *oder* EXPERIMENT *verstehen wir die Verwirklichung eines Bedingungskomplexes $\mathcal{B}$ und die Beobachtung des Ergebnisses.*

Ein Versuch muss nicht notwendig von Menschen durchgeführt werden. So ist eine durch $\mathcal{B}$ spezifizierte Naturbeobachtung auch ein Versuch.

Def. 5.3 (Zufallsversuch) *Wird bei Wiederholung des Versuchs, stets dasselbe Ergebnis erwartet, dann spricht man von einem* DETERMINISTISCHEN *Versuch. Demgegenüber heißt ein Versuch* ZUFÄLLIG, *wenn im einzelnen nicht vorhersagbare Ergebnisse zu erwarten sind.*

Def. 5.4 (Grundgesamtheit) *Die Menge aller unter $\mathcal{B}$ durchführbaren Zufallsversuche (kurz Versuche oder Experiment) heißt Grundgesamtheit $\mathcal{G}$.*

Die Grundgesamtheit kann endlich oder unendlich sein. Beim Urnen-Experiment ist sie unendlich, da immer wieder zurückgelegt wird. Ohne Zurücklegen, hätte die Grundgesamtheit die MÄCHTIGKEIT (Zahl der Kugeln) 100.

Def. 5.5 (Elementarereignisse) *Ereignisse, die sich nicht weiter zerlegen lassen, heißen Elementarereignisse oder Ergebnisse. Alle anderen heißen zusammengesetzte Ereignisse.*

[1] Ansonsten modifiziert das die Verteilung der Kugeln, die an der Oberfläche anzutreffen sind.

Beim Roulette kann rot auf 18 Arten zustandekommen: 18 Felder sind rot. Unter diesen ist auch z.B. das mit 12 markierte Feld. 12 ist in diesem Beispiel ein Elementarereignis.
Allgemein sind Ereignisse E Teilmengen der Grundgesamtheit $\mathcal{G}$, $E \subseteq \mathcal{G}$. Ein Ereignis E tritt ein, wenn ein Ergebnis $\omega \in E$ beobachtet wird. Demnach ist ein Ereignis entweder ein Elementarereignis oder eine Vereinigung von Elementarereignissen.
Auch $\mathcal{G}$ und die leere Menge $\emptyset$ sind Ereignisse, nämlich das sichere bzw. unmögliche Ereignis. Denn $\omega \in \mathcal{G}$ tritt mit Sicherheit ein; und $E = \emptyset$ bedeutet, dass es kein Ergebnis ω mit $\omega \in E$ gibt.
$\mathcal{G}$ heißt diskret, wenn die möglichen Ergebnisse abzählbar sind, sonst heißt $\mathcal{G}$ kontinuierlich. Es kann auch ein gemischter Fall vorliegen. Oft haben Versuche nur zwei Ausgänge: sogenannte DICHOTOME Versuche (z.B. Münzwurf).
Mit $|\mathcal{G}|$ bezeichnen wir die Mächtigkeit. Bei dichotomen Versuchen ist $|\mathcal{G}| = 2$. Beim Würfel-Experiment ist $|\mathcal{G}| = 6$. Bei Stichproben vom Umfang n bildet die Menge aller Stichproben-Ergebnisse den Ergebnis-Raum $\mathcal{G}_n$, auch Stichproben-Raum genannt. Bei dichotomen Experimenten ist offensichtlich $|\mathcal{G}_n| = 2^n$.

Def. 5.6 (Bernoulli-Versuche) *Unter einem* BERNOULLI-VERSUCH *versteht man, dass ein Versuch mit zwei Ausgängen wiederholt durchgeführt wird. Hierbei sollen die Ausgänge der einzelnen Versuche voneinander unab-hängig sein.*

Das eben besprochene Urnenexperiment kann als Bernoulli-Experiment aufgefasst werden, bei dem nacheinander 40 Kugeln gezogen werden, die entweder rot oder gelb sind.

Def. 5.7 (Stichprobe) *Man kann Versuche bündeln. Eine n-malige Wiederholung eines einzelnen Versuchs heißt ebenfalls Versuch: Man sagt dann, man habe eine* STICHPROBE VOM UMFANG n *gezogen.*

In diesem Sinne handelt es sich in obigem Beispiel um eine Stichprobe vom Umfang 40.
Duale Bedeutung wiederholter Versuche: In der Wahrscheinlichkeitstheorie hat der Begriff „wiederholte Versuche" zwei unterschiedliche Bedeutungen. Der erste ist die approximative Beziehung zwischen intrinsischer Wahrscheinlichkeit und relativer Häufigkeit. Die zweite ist die Durchführung eines zusammengesetzten Experiments. Z.B. kann das Urnen-Experiment zweifach interpretiert werden:
1) Das eigentliche Experiment (Versuch) ist der einzelne Zug einer Kugel. Der Versuch wird nun n mal wiederholt, um die intrinsische Wahrscheinlichkeit q_α des Einzelexperimentes zu ermitteln.
2) Das Experiment besteht darin, die n Kugeln zu ziehen. Der zugrunde liegende WAHRSCHEINLICHKEITSRAUM hat nun 2^n Elemente. Um die zugehörigen Wahrscheinlichkeiten über relative Häufigkeiten zu ermitteln, muss das zusammengesetzte Experiment mehrfach wiederholt werden.

Def. 5.8 (Ereignis) *Den Ausgang eines Versuchs nennt man* EREIGNIS. **Ereignisse treten ein oder nicht***! Ein mehr oder wenig gibt es nicht.*

Ein mögliches EREIGNIS ist: *gelb im i-ten Zug*. Aber auch die Auswahl der Urne α ist ein Ereignis. Im Urnen-Experiment wird von bekannten Ereignissen auf unbekannte geschlossen. Abgesehen von Ereignissen spielen Propositionen eine zentrale Rolle in der Wahrscheinlichkeitstheorie.

Def. 5.9 (Propositionen) *Aussagen, die entweder wahr oder falsch sind, nennt man* PROPOSITIONEN.

- *Die ausgewählt Urne hat den Index α.*
- *Die nächste Kugel wird gelb sein.*
- *Es gibt eine weitere Urne U_4 mit unbekanntem q_4.*
- *Julius Cäsar war ein Mann.*
- *Morgen wird es regnen.*
- *Die Länge des Stabes liegt zwischen $1.0m$ und $1.1m$.*
  Auf diese Weise können auch überabzählbare Propositionen für kontinuierliche Freiheitsgrade eingeführt werden.

Def. 5.10 (Hypothesen) *Hypothesen sind Aussagen über unbekannte Ereignisse.*

- *Die gewählte Urne hat die Nummer drei.*

Hypothesen sind wichtig in der orthodoxen Statistik. Streng genommen gibt es keinen Unterschied zwischen Hypothesen und Propositionen.

Def. 5.11 (Vorwärts-/Rückwärtsrechnung) *Es liege U_1 vor. Die Proposition* DIE NÄCHSTE KUGEL WIRD GELB SEIN *erhält man aus einer* VORWÄRTSRECHNUNG. *Im Urnen-Experiment interessiert der Umkehrschluss, der von den beobachteten Folgen der Farben auf die mögliche, hypothetische Ursache rückschließt (*RÜCKWÄRTSRECHNUNG*). Man spricht auch von induktivem Schluss* INDUKTIVER LOGIK.

Im Drei-Urnen-Experiment scheint es zwei Arten von Wahrscheinlichkeiten zu geben:

W_1 : Wahrscheinlichkeit, aus der Urne U_α eine gelbe Kugel zu ziehen.

W_2: Wahrscheinlichkeit, dass die Urne α vorliegt, wenn die Stichprobe vom Umfang N n_g gelbe Kugeln enthält.

Die erste Wahrscheinlichkeit hat einen "objektiven" Charakter, da sie von der Person losgelöst ist, die die Auswertung des Experimentes vornimmt. Sie beschreibt eine Eigenschaft der ausgewählten Urne. Präziser formuliert ist W_1 die *Wahrscheinlichkeit, im nächsten Zug gelb zu ziehen, wenn U_α vorliegt* eine bedingte Wahrscheinlichkeit, die wir in Zukunft kompakt

$$W_1 = P(\text{gelb}|U_\alpha, \mathcal{B}) = q_\alpha \tag{5.1}$$

schreiben werden. Gemäß der klassischen Definition weisen wir dieser Wahrscheinlichkeit den Wert q_α zu, da es unter Berücksichtigung des Bedingungskomplexes $\mathcal{B}$ bei jedem Zug insgesamt 100 Möglichkeiten (Kugeln) gibt, wovon $100 * q_\alpha$ günstige Ereignisse darstellen. Andererseits wissen wir vom Bernoulli-Gesetz der großen Zahlen, dass $n_g/N \underset{N\to\infty}{\longrightarrow} q_\alpha$. Also ist auch im Rahmen der statistischen Definition der Wahrscheinlichkeit W_1 der Wert q_α zuzuweisen. Bei der Wahrscheinlichkeit W_1 handelt es sich um eine sogenannte VORWÄRTS-WAHRSCHEINLICHKEIT, deren zufälliges Verhalten intrinsisch vom Experiment ($\mathcal{B}$) vorgegeben ist.
Anders verhält es sich bei der Wahrscheinlichkeit W_2

$$W_2 = P(U_\alpha|N, n_g, \mathcal{B}) \qquad .$$

Hierbei handelt es sich um eine RÜCKWÄRTS-WAHRSCHEINLICHKEIT, denn wenn wir die Argumente vor und hinter dem Bedingungsstrich umarrangieren zu $P(n_g|U_\alpha, N, \mathcal{B})$, erhalten wir die Vorwärts-Wahrscheinlichkeit, dass bei einer Stichprobe vom Umfang N aus der Urne U_α die Anzahl der gelben Kugeln n_g sein wird. Das ist eine Wahrscheinlichkeit, die wieder objektiven Charakter hat. Diese Wahrscheinlichkeit kennen wir bereits. Es handelt sich um die Bernoulli-Verteilung $P(n_g|N, q_\alpha)$.
Auf die gesuchte Rückwärtswahrscheinlichkeit kann nur "induktiv" geschlossen werden: Nehmen wir an, $n_g/N = 0.25$. Da q_1 am nächsten an dem beobachten Verhältnis n_g/N liegt, wird man schließen, dass es sich um die Urne 1 handelt. Aber wie kann man diese Wahrscheinlichkeit quantifizieren?

5.2 Orthodoxe Statistik versus Bayessche Wahrscheinlichkeitstheorie

Es gibt zwei kontroverse Sichtweisen bei Problemen der induktiven Logik, die Sicht der orthodoxen Statistik und die der Bayesschen Wahrscheinlichkeitstheorie.

5.2.1 Orthodoxe Statistik

In der ORTHODOXEN STATISTIK beschreibt man Vorwärts-Wahrscheinlichkeiten durch Zufallsvariablen und Rückwärts-Wahrscheinlichkeiten über Hypothesen-Tests.
Die Zufälligkeit wird in der orthodoxen Statistik als intrinsische (physikalische) Eigenschaft des Experiments betrachtet. Man unterscheidet streng zwischen folgenden Situationen.
a) Eine Münze ist völlig symmetrisch. Die Wahrscheinlichkeit, beim nächsten Wurf Kopf zu erhalten, ist $w_{\text{Kopf}} = 1/2$.
b) Die Münze ist manipuliert und fällt immer auf eine (mir aber unbekannte) Seite. Nun ist $X(\text{Kopf/Zahl})$ keine Zufallsvariable mehr, da sie immer denselben Wert annehmen wird. Die Wahrscheinlichkeit lässt sich nun nicht mehr über n_{Kopf}/n bestimmen, da dieses Verhältnis immer 0 oder 1 sein wird. Aber das wissen wir erst,

nachdem wir das Experiment durchgeführt haben. Man kann vorher keine Aussagen machen.
Vor dem Auswählen der Urne ist der Index α ebenfalls eine Zufallsvariable mit der Prior-Wahrscheinlichkeit $w(\alpha) = 1/3$.
Wenn die Urne bereits ausgewählt wurde, α also bereits vorliegt, wir nur den Wert nicht wissen, betrachtet die Orthodoxe Statistik α nun nicht mehr als Zufallsvariable.[2] Man kann im Rahmen der Orthodoxen Statistik z.B. nicht mehr nach der Wahrscheinlichkeit, dass $\alpha = 1$ ist, fragen. Um dennoch Aussagen über solche Fragestellungen machen zu können, wurden die Signifikanz-Tests erfunden.

5.2.2 Signifikanz-Test

Wir wollen wissen, ob die Daten mit der Hypothese $\alpha = 1$ konsistent sind, wenn NICHT, nennt man das Experiment SIGNIFIKANT. Das heißt, die Daten sind zu weit vom Erwartungswert entfernt, dass die Abweichungen nicht mehr zufälligen Charakter haben, sondern SIGNIFIKANT sind. Um zu zeigen, dass Daten signifikant sind, greift man in der orthodoxen Statistik auf einen TRICK zurück. Man könnte das den *indirekten Beweis der Statistik* nennen. Man geht davon aus, dass die Daten einer Wahrscheinlichkeitsverteilung genügen, die durch die Hypothese vorgegeben ist. Liegen die beobachteten Daten in den Ausläufern der Wahrscheinlichkeitsverteilung, dann ist es unwahrscheinlich, dass die Hypothese korrekt ist. Die Daten sind dann signifikant, und die Hypothese muss verworfen werden.
Im vorliegenden Drei-Urnen-Problem wollen wir z.B. die Hypothese testen, dass die Urne U_1 vorliegt. Wenn die Hypothese korrekt ist, erwarten wir im Mittel $\mu = N\ q_1$ gelbe Kugeln. Die beobachtete Anzahl $n_g^{\exp}$ weicht hiervon um $\Delta n^* = \left|\mu - n_g^{\exp}\right|$ ab. Nun reicht es nicht die Wahrscheinlichkeit $P(n_g^{\exp}|N, q_\alpha)$ für den beobachteten Wert anzuschauen, da diese Wahrscheinlichkeit mit zunehmendem N gegen Null geht. Man berechnet stattdessen die Wahrscheinlichkeit dafür, dass eine Abweichung Δn vom Mittelwert auftritt, die so groß wie oder größer als die beobachtete Abweichung Δn^* ist. Das heißt, man berechnet die Wahrscheinlichkeit

$$P(n_g \leq \mu - \Delta n^* \vee n_g \geq \mu + \Delta n^* | N, q_1, \mathcal{B}) = \sum_{n_g=0}^{\mu-\Delta n^*} P(n_g|N, q_1) \ + \sum_{n_g=\mu+\Delta n^*}^{N} P(n_g|N, q_1) \quad . \tag{5.2}$$

Wenn diese Wahrscheinlichkeit kleiner als ein vorgegebenes Signifikanz-Niveau p_S (z.B. 5%) ist, verwirft man die Hypothese.

Def. 5.12 (Statistischer Fehler erster Art, Irrtumswahrscheinlichkeit) *Ein Fehler erster Art liegt vor, wenn man eine Hypothese verwirft, obwohl sie richtig ist. Die Wahrschein-*

[2] Strenggenommen dürfte man mit (Pseudo-)Zufallszahlen keine Wahrscheinlichkeitsrechnungen machen, da die Sequenz deterministisch vorgegeben ist und jeder Wert mit Wahrscheinlichkeit Eins vorhergesagt werden kann. Es liegt also genau die Situation vor, wie im Fall der bereits gewählten Urne.

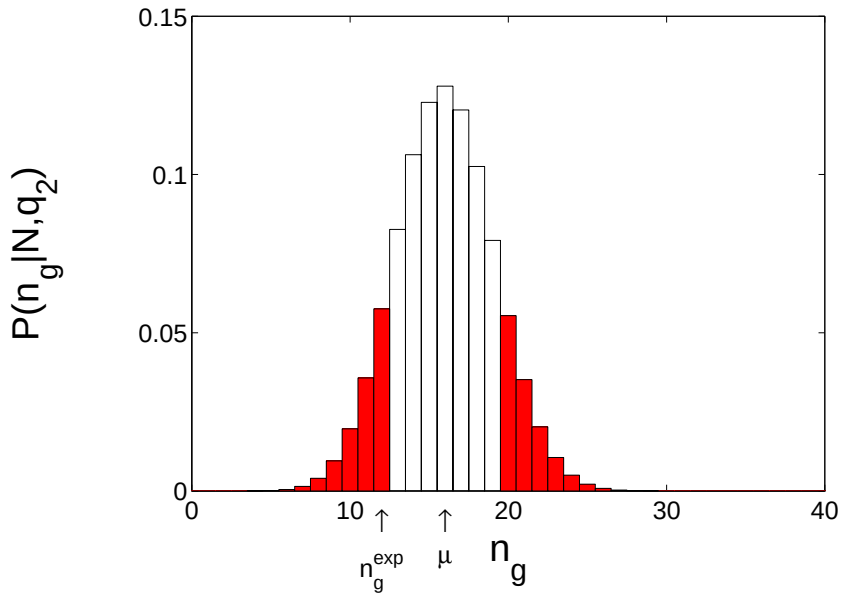

Abbildung 5.1: *Binomial-Verteilung zu* $q_2 = 0.4$ *und* $N = 40$ *Die Fläche unter dem gefüllten Bereich gibt die Wahrscheinlichkeit* $P(\Delta n \geq \Delta n^* | U_\alpha, \mathcal{B})$ *für* $g^{\exp} = 12$ *an, auf der der Signifikanz-Test beruht.*

lichkeit, dass ein Fehler erster Art bei gegebenem statistischen Test auftritt, nennt man IRRTUMSWAHRSCHEINLICHKEIT.

Def. 5.13 (Statistischer Fehler zweiter Art) *Ein Fehler zweiter Art liegt vor, wenn man eine Hypothese akzeptiert, obwohl sie falsch ist.*

Wie groß ist die Irrtumswahrscheinlichkeit beim Verwerfen der Hypothese, wenn $P(\Delta n \geq \Delta n^* | U_\alpha, \mathcal{B}) < p_S$? Vorausgesetzt, die Hypothese ist korrekt, dann kennen wir die Wahrscheinlichkeitsverteilung, mit der die Realisierungen $n_g = n_g^{\exp}$ auftreten. Es gibt Grenzwerte $n_g^{(1)}$ und $n_g^{(2)}$, die vom Signifikanz-Niveau abhängen, mit der Eigenschaft, dass die Hypothese im Test verworfen wird, wenn immer der beobachtete Wert $n_g^{\exp}$ kleiner als $n_g^{(1)}$ oder größer als $n_g^{(2)}$ ist. Die Wahrscheinlichkeit, dass solche Werte $n_g^{(\exp)}$ im Rahmen der Hypothese durch „Zufallsschwankungen" auftreten, ist aber gerade

$$P(n_g < n_g^{(1)} \vee n_g > n_g^{(2)}) = p_S \qquad .$$

Das bedeutet, dass wir bei dem Signifikanz-Test mit Wahrscheinlichkeit p_S einen Fehler erster Art machen werden.
Man könnte nun auf die Idee kommen diesem Fehler dadurch zu minimieren, dass man das Signifikanz-Niveau kleiner wählt. Dann wird man seltener versehentlich die Hypothese verwerfen, obwohl sie richtig ist. Andererseits steigt damit der Fehler zweiter Art drastisch an, denn man wird nun zunehmend die Hypothese akzeptieren, obwohl sie nicht korrekt ist. Umso kleiner die Rand-Wahrscheinlichkeit

$P(\Delta n \geq \Delta n^* | U_\alpha, \mathcal{B})$ ist, umso größer ist die Wahrscheinlichkeit, dass es sich um signifikante Daten handelt. Das wiederum bedeutet, dass die Hypothese falsch ist.
Die Rand-Wahrscheinlichkeit ist für das Drei-Urnen-Problem in Abbildung 5.2 über den möglichen experimentellen Zählern $n_g^{\exp}$ aufgetragen.

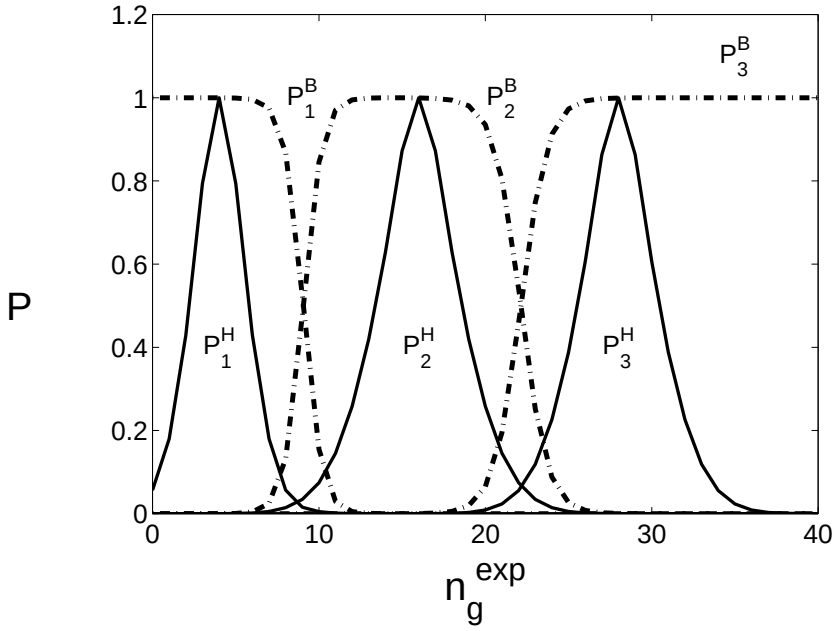

Abbildung 5.2: *Drei-Urnen-Problem für $N = 40$ und q-Werte wie im Text. Signifikanz-Test: (durchgezogene Kurve) Rand-Wahrscheinlichkeit P_α^H unter der Hypothese, dass U_α vorliegt. Bayesscher Hypothesen-Vergleich: (gestrichelte Kurve) $P_\alpha^B = P(U_\alpha | N, n_g^{\exp}, \mathcal{B})$.*

Nachteile dieser Methode

- Ad hoc.
- Nur sinnvoll, wenn unimodale Wahrscheinlichkeitsverteilung vorliegt.
- Es werden Hypothesen isoliert betrachtet.
  Es geht nicht ein, welche Alternativen es gibt. Es kann sein, dass
  - es nur 2 Alternativen gibt und separate Tests beide Alternativen verwerfen.
  - die betrachtete Hypothese völlig falsch ist und dennoch den Test besteht und die richtige Hypothese gar nicht untersucht wird.
  - Bsp: Hypothese *es hat geregnet*. Daten: Der Boden ist nass. Die Hypothese besteht jeden Test, da

    $$P(\text{Der Boden ist nass} | \text{Es hat geregnet}) = 1 \quad .$$

    Es fehlen alle anderen Erklärungen.

- Wie soll man das Signifikanz-Niveau wählen? Üblich ist 5% oder 1%.

5.2.3 Bayessche Wahrscheinlichkeitstheorie

Die Bayessche Wahrscheinlichkeitstheorie (BWT) unterscheidet sich in zwei wesentlichen Punkten von der orthodoxen Statistik, nämlich in der Interpretation des Begriffes Wahrscheinlichkeit und in einer konträren Sicht von Zufallsereignissen.
a) In der Baysschen Wahrscheinlichkeitstheorie versteht man Wahrscheinlichkeiten als Maß dafür, dass eine Proposition wahr ist. Insofern stellt die BWT eine Erweiterung der Aussagenlogik auf Teilwahrheiten dar.
b) Die Zufälligkeit ist keine intrinsische (physikalische) Eigenschaft des untersuchten Objektes, sondern resultiert aus mangelnder Information. Diese Sicht ist natürlich wesentlich sinnvoller und weitreichender als die orthodoxe. Sie trifft zu bei

- Münzwurf (klassisch, deterministisch).
  Es fehlt nur die nötige Information, um alles zu berechnen.
- Würfel
- (Pseudo-)Zufallszahlen. Man hat sogar die gesamte Information um die Zahlen vorherzusagen.
- Drei-Urnen-Problem. Die Urne wird nicht wirklich zufällig ausgewählt. Die Person, die sie auswählt hat i.d.R ein Konzept, wie sie die Urne auswählt.
- Wählerverhalten. Es gibt sicher bei den meisten klare Vorstellungen warum sie eine bestimmte Partei wählen.
- Klassische Vielteilchen-Systeme.

Was den Zufall charakterisiert, ist die Unberechenbarkeit aufgrund mangelnder Information.
Im Rahmen der Bayesschen Wahrscheinlichkeitstheorie gibt es auch keinen prinzipiellen Unterschied zwischen der Behandlung des oben beschriebenen Münz-Problems, bei dem einmal die Münze symmetrisch und einmal manipuliert ist. In beiden Fällen wird man vor dem ersten Experiment dieselbe Wahrscheinlichkeit erhalten

$$P(\text{Kopf}|\text{fair}, \mathcal{B}) = 1/2$$

sowie

$$P(\text{Kopf}|\text{manipuliert}, \mathcal{B}) = 1/2 \quad .$$

Der Unterschied manifestiert sich erst beim zweiten Versuch

$$P(\text{Kopf}|\text{erster Versuch: Zahl}, \text{fair}, \mathcal{B}) = 1/2$$

aber

$$P(\text{Kopf}|\text{erster Zug Zahl}, \text{manipuliert}, \mathcal{B}) = 0 \quad .$$

Wir werden im übernächsten Kapitel zeigen, dass die BWT die einzige konsistente Theorie ist, um Teilwahrheiten quantitativ zu beschreiben. Die BWT stellt einen Kalkül für Propositionen dar. Die Rechenregeln, die sich aus der Konsistenzforderung ergeben, sind die bereits bekannte Summen- und Produktregel und daraus abgeleitet das Bayessche Theorem. Sie wird Bayessche Wahrscheinlichkeitstheorie genannt, da sie ausgiebig vom Bayesschen Theorem gebraucht macht.
Das Bayessche Theorem gibt es auch in der Orthodoxen Statistik. Jedoch dürfen die Argumente nur Zufallsvariablen sein, da dann die bedingten Wahrscheinlichkeiten als Quotient von Wahrscheinlichkeiten definiert sind, zu denen man relative Häufigkeiten angeben kann. Deshalb ist es nicht möglich gewesen, dass Bayessche Theorem heranzuziehen, um zu ermitteln, welche Urne U_α vorliegt. Im Rahmen der BWT gibt es diese Probleme nicht. Wir benötigen folgende Propositionen

- U_α: *Die Urne α wurde ausgewählt.*
- $q = q_\alpha$: *Der Anteil q der gelben Kugeln ist q_α.*
- N: *Die Stichprobe hat den Umfang N.*
- n_g: *In der Stichprobe sind n_g gelbe Kugeln.*
- $\mathcal{B}$: Die Propositionen, die den Bedingungskomplex definieren.

Wir bestimmen die gesuchte Wahrscheinlichkeit $P(U_\alpha|N, n_g, \mathcal{B})$ aus dem Bayesschen Theorem (siehe Kapitel 7.2 bzw. Gl. (7.5))

$$p_\alpha := P(U_\alpha|n_g, N, \mathcal{B}) = \frac{P(n_g|U_\alpha, N, \mathcal{B})\ P(U_\alpha|N, \mathcal{B})}{P(n_g|N, \mathcal{B})} \quad . \tag{5.3}$$

Der erste Term im Zähler $P(n_g|U_\alpha, N, \mathcal{B})$ ist die Vorwärts-Wahrscheinlichkeit, dass eine Stichprobe vom Umfang N aus der Urne U_α n_g gelbe Kugeln enthält. Das ist natürlich die Binomial-Verteilung $P(n_g|N, q_\alpha)$ (Gl. (3.8a)).
Der zweite Term im Zähler $P(U_\alpha|N, \mathcal{B})$ ist die sogenannte PRIOR-WAHRSCHEINLICHKEIT, dass die Urne U_α vorliegt. Für die Bestimmung dieser Wahrscheinlichkeit ist die Kenntnis des Stichprobenumfangs irrelevant. Es gilt deshalb $P(U_\alpha|N, \mathcal{B}) = P(U_\alpha|\mathcal{B})$. Diese Wahrscheinlichkeit ist $P(U_\alpha|N, \mathcal{B}) = 1/3$, da es 3 Urnen gibt und nur eines der Ereignisse günstig ist.
Schließlich kommt in Gl. (5.3) noch ein Normierungsnenner vor, der unabhängig von α ist. Er sorgt dafür, dass $\sum_{\alpha=1}^{3}\ p_\alpha = 1$. Daraus folgt

$$p_\alpha := \frac{P(n_g|N, q_\alpha)}{\sum_{\beta=1}^{3}\ P(n_g|N, q_\beta)} \quad . \tag{5.4}$$

Diese Wahrscheinlichkeit ist in Abbildung 5.2 als Funktion von $n_g = n_g^{\text{exp}}$ aufgetragen. Wir erkennen nun klarer, was der Unterschied zwischen den beiden Arten von Wahrscheinlichkeiten ist.

Def. 5.14 (Prior-Wahrscheinlichkeit) *Wir bezeichnen hier mit* PRIOR-WAHRSCHEINLICHKEIT *die Wahrscheinlichkeit $P(X|\mathcal{B})$ für eine Proposition X, wenn nur der Bedingungskomplex gegeben ist und keine Daten vorliegen. $\mathcal{B}$ kann aber durchaus Vorwissen anderer Form (z.B. Summenregeln, etc) enthalten.*[3]

Wir werden später erfahren, wie man konsistent Prior-Wahrscheinlichkeiten zuweist. Im Drei-Urnen-Experiment konnte die Prior-Wahrscheinlichkeit bereits aus der klassischen Definition der Wahrscheinlichkeit bestimmt werden.
Im Gegensatz zur Prior-Wahrscheinlichkeit gibt es die

Def. 5.15 (Posterior-Wahrscheinlichkeit) *Unter der* POSTERIOR-WAHRSCHEINLICHKEIT *verstehen wir die Wahrscheinlichkeit $P(X|D,\mathcal{B})$ für X, wenn neben dem Bedingungskomplex auch Daten D vorliegen.*

Schließlich hatten wir neben der Posterior- (bzw. Rückwärts-) Wahrscheinlichkeit die Vorwärts-Wahrscheinlichkeit, die man auch Likelihood-Funktion nennt

Def. 5.16 (Likelihood-Funktion) *Unter der* LIKELIHOOD-FUNKTION *verstehen wir die Wahrscheinlichkeit $P(D|X,\mathcal{B})$, die Daten D zu messen, wenn die Proposition X wahr ist und weiterhin der Bedingungskomplex vorliegt. Man nennt sie Likelihood, da man an der Abhängigkeit von X interessiert ist. In dieser Größe ist sie keine Wahrscheinlichkeit, da sie hierin nicht normiert ist.*

Wichtig: $P(D|X,\mathcal{B}) + P(D|\overline{X},\mathcal{B}) \neq 1$.

Im Urnen-Beispiel war die Likelihood die Binomial-Verteilung. Die Likelihood-Funktion ist i.d.R. die einzige Größe, die auch in der Orthodoxen Statistik verwendet wird.

[3] Man findet auch häufig den Begriff „a-priori-Wahrscheinlichkeit". Der Begriff ist allerdings bereits in der Philosophie mit anderer Bedeutung in Gebrauch.

Kapitel 6

Boolsche Algebren und Borel-Körper

6.1 Halbordnung

Def. 6.1 (Halbordnung) *Eine Menge M (mit Elementen $a, b, c, \ldots$) heißt* HALBGEORDNET, *wenn in M eine Relation $\preceq$ definiert ist mit folgenden Eigenschaften:*

$$a \preceq a \qquad \forall a \in M \qquad , \tag{6.1a}$$

$$[a \preceq b \wedge b \preceq c] \Rightarrow a \preceq c \qquad , \tag{6.1b}$$

$$[a \preceq b \wedge b \preceq a] \Rightarrow a = b \qquad . \tag{6.1c}$$

Beispiele sind:

- Die Menge der reellen Zahlen mit der Relation $\leq$.
- Die Potenzmenge einer beliebigen Menge mit der Mengen-Relation $\subseteq$.

(*) Die Menge $M = \{2, 3, 5, 6, 10, 15, 30\}$ ist halbgeordnet durch $a|b$, das heißt, a ist Teiler von b.

Dieses Beispiel zeigt auch, dass nicht zwischen allen Elementen eine Halbordnung existieren muss. Die Menge der halbgeordneten Paare ist

$$\begin{aligned}\{&(2,2),(2,6),(2,10),(2,30), \quad (3,3),(3,6),(3,15),(3,30),\\ &(5,5),(5,10),(5,15),(5,30), \quad (6,6),(6,30),\\ &(10,10),(10,30), \quad (15,15),(15,30) \quad (30,30)\} \qquad .\end{aligned}$$

Def. 6.2 (Nullelement, Einselement) *Eine halbgeordnete Menge M besitzt ein Nullelement $\mathcal{N}_\preceq$, wenn gilt*

$$\mathcal{N}_\preceq \preceq a \qquad \forall a \in M \qquad .$$

Eine halbgeordnete Menge M besitzt ein Einselement $\mathcal{E}_\preceq$, wenn gilt

$$a \preceq \mathcal{E}_\preceq \qquad \forall a \in M \qquad .$$

Im Zusammenhang mit der Wahrscheinlichkeitstheorie ist das Nullelement das unmögliche Ereignis und das Einselement das sichere Ereignis.

Bemerkung 6.1 *Es existiert nicht immer ein Null- oder Einselement, wie das Beispiel (*) zeigt.*

Satz 6.1 (Eindeutigkeit des Null- oder Einselement) *Existiert ein Null- oder Einselement, so ist es eindeutig.*

Beweis: Seien $\mathcal{N}_{\preceq,1}$ und $\mathcal{N}_{\preceq,2}$ Nullelemente. Dann gilt $\mathcal{N}_{\preceq,1} \preceq a$ für alle a, insbesondere auch $\mathcal{N}_{\preceq,1} \preceq \mathcal{N}_{\preceq,2}$. Mit der gleichen Argumentation gilt $\mathcal{N}_{\preceq,2} \preceq \mathcal{N}_{\preceq,1}$, also nach Gl. (6.1c) $\mathcal{N}_{\preceq,2} = \mathcal{N}_{\preceq,1}$. Der Beweis für das Einselement läuft völlig analog.

6.2 Boolsche Algebra

Def. 6.3 (distributiver Verband) *Eine Menge V (mit Elementen $a, b, c, \ldots$) bildet einen distributiven Verband, wenn zweistellige Verknüpfungen $\sqcap$ und $\sqcup$ definiert sind, die folgende Eigenschaften besitzen:*

$$a \sqcap b = b \sqcap a \qquad a \sqcup b = b \sqcup a \tag{6.2a}$$
$$(a \sqcap b) \sqcap c = a \sqcap (b \sqcap c) \qquad (a \sqcup b) \sqcup c = a \sqcup (b \sqcup c) \tag{6.2b}$$
$$a \sqcap (a \sqcup b) = a \qquad a \sqcup (a \sqcap b) = a \tag{6.2c}$$
$$a \sqcap (b \sqcup c) = (a \sqcap b) \sqcup (a \sqcap c) \qquad a \sqcup (b \sqcap c) = (a \sqcup b) \sqcap (a \sqcup c) \quad . \tag{6.2d}$$

Bemerkung 6.2 *Gl. (6.2a) spiegelt die Kommutativität und Gl. (6.2b) die Assoziativität wieder. Die Axiome Gl. (6.2a),Gl. (6.2b),Gl. (6.2c) nennt man auch Verbandsaxiome, wenn nur sie erfüllt sind, liegt ein Verband vor. Erst die Axiome Gl. (6.2d) machen den Verband zum distributiven Verband.*

Bemerkung 6.3 *In den Axiomen treten die Operationen $\sqcap$ und $\sqcup$ gleichberechtigt auf. Hat man irgendeine Ausage bewiesen, gilt die entsprechende* DUALE AUSAGE, *d.h. jene in der $\sqcap$ durch $\sqcup$ ersetzt wird und umgekeht, genauso.*

Satz 6.2 (Idempotenz) *In einem Verband gilt für alle a:*

$$a \sqcap a = a \tag{6.3a}$$
$$a \sqcup a = a \quad . \tag{6.3b}$$

Beweis: Für Gl. (6.3a):

$$a \sqcap a \overset{Gl.(6.2c),\ \text{rechts}}{=} a \sqcap (a \sqcup (a \sqcap b)) \overset{Gl.(6.2c),\ \text{links}}{=} a \quad .$$

Gl. (6.3b) ist der duale Ausdruck zu Gl. (6.3a)

Def. 6.4 (Null-, Einselement) *In einem Verband gelte:*
Ein Objekt $\mathcal{N}$ mit $a \sqcup \mathcal{N} = a$ und $a \sqcap \mathcal{N} = \mathcal{N}$ für alle a heißt Nullelement.
Ein Objekt $\mathcal{E}$ mit $a \sqcap \mathcal{E} = a$ und $a \sqcup \mathcal{E} = \mathcal{E}$ für alle a heißt Einselement.

Satz 6.3 (Eindeutigkeit des Null-, Einselements) *Existiert in einem Verband ein Null- oder Einselement, so ist es eindeutig.*

Beweis: Sind $\mathcal{N}_1$ und $\mathcal{N}_2$ beide Nullelemente, so gilt $\mathcal{N}_2 \sqcap \mathcal{N}_1 = \mathcal{N}_1$ (da $\mathcal{N}_1$ Nullelement) und $\mathcal{N}_1 \sqcap \mathcal{N}_2 = \mathcal{N}_2$ (da $\mathcal{N}_2$ Nullelement), also ist $\mathcal{N}_1 = \mathcal{N}_2 \sqcap \mathcal{N}_1 \overset{Gl.(6.2a),\text{ links}}{=} \mathcal{N}_1 \sqcap \mathcal{N}_2 = \mathcal{N}_2$. Der Beweis fürs Einselement geht mit den dualen Ausdrücken.

Bemerkung 6.4 *Die Elemente $\mathcal{N}$ und $\mathcal{E}$ sind zueinander dual (falls sie existieren).*

Def. 6.5 (Komplement) *In einem Verband mit Null- und Einselement heißt ein Element $\overline{a}$ Komplement von a, wenn*

$$a \sqcap \overline{a} = \mathcal{N}, \quad \textit{und} \quad a \sqcup \overline{a} = \mathcal{E} \qquad .$$

Satz 6.4 (Eindeutigkeit des Komplements) *In einem distributiven Verband gibt es zu jedem Objekt höchstens ein Komplement.*

Beweis: Sind $\overline{a}_1$ und $\overline{a}_2$ Komplemente von a, so gilt nach Definition $a \sqcup \overline{a}_1 = \mathcal{E} = a \sqcup \overline{a}_2$, sowie $a \sqcap \overline{a}_1 = \mathcal{N} = a \sqcap \overline{a}_2$.
Dann ist
$\overline{a}_1 \overset{Gl.(6.2c),\text{ rechts}}{=} \overline{a}_1 \sqcup (\overline{a}_1 \sqcap a) \overset{Gl.(6.2a),\text{ links}}{=} \overline{a}_1 \sqcup (a \sqcap \overline{a}_1) \overset{Vor.}{=}$
$\overline{a}_1 \sqcup (a \sqcap \overline{a}_2) \overset{Gl.(6.2d),\text{ rechts}}{=} (\overline{a}_1 \sqcup a) \sqcap (\overline{a}_1 \sqcup \overline{a}_2) \overset{Vor.,Gl.(6.2a),\text{ rechts}}{=}$
$(\overline{a}_2 \sqcup a) \sqcap (\overline{a}_2 \sqcup \overline{a}_1) \overset{Gl.(6.2d),\text{ rechts}}{=} \overline{a}_2 \sqcup (a \sqcap \overline{a}_1) \overset{Vor.}{=} \overline{a}_2 \sqcup (a \sqcap \overline{a}_2) \overset{Gl.(6.2a),\text{ links}}{=} \overline{a}_2 \sqcup (\overline{a}_2 \sqcap a) \overset{Gl.(6.2c),\text{ rechts}}{=} \overline{a}_2$.

Satz 6.5 *In einem Verband wird durch $a \preceq b :\Leftrightarrow a \sqcap b = a$ (oder alternativ $a \preceq b :\Leftrightarrow a \sqcup b = b$) eine Halbordnungsstruktur definiert.*

Beweis: Wir müssen zeigen, dass die 3 Halbordnungsaxiome gelten.

Gl. (6.1a): Aus dem Satz 6.2 (Idempotenz) folgt nach der Definition in Satz 6.5: $a \preceq a$.

Gl. (6.1b): Aus $a \preceq b$ und $b \preceq c$ folgt zunächst nach Definition $a \sqcap b = a$ und $b \sqcap c = b$. Dann ist durch Einsetzen $a \sqcap c \overset{Vor.}{=} (a \sqcap b) \sqcap c \overset{Gl.(6.2b),\text{ links}}{=} a \sqcap (b \sqcap c) \overset{Vor.}{=} a \sqcap b \overset{Vor.}{=} a$, also $a \preceq c$.

Gl. (6.1c): Aus $a \preceq b$ und $b \preceq a$ folgt zunächst nach Definition $a \sqcap b = a$ und $b \sqcap a = b$, daher gilt: $a \overset{Vor.}{=} a \sqcap b \overset{Gl.(6.2a),\text{ links}}{=} b \sqcap a \overset{Vor.}{=} b$.

Satz 6.6 *In einem Verband sei durch obige Vorschrift eine Halbordnung definiert. Falls Verband wie Halbordnungsstuktur über $\mathcal{N}/\mathcal{E}$ bzw. $\mathcal{N}_\preceq/\mathcal{E}_\preceq$ verfügen, gilt:*

$$\mathcal{N} = \mathcal{N}_\preceq \qquad \textit{und} \qquad \mathcal{E} = \mathcal{E}_\preceq \qquad .$$

Beweis: Nach Definition 6.4 gilt im Verband für das Nullelement für alle a: $a \sqcap \mathcal{N} = \mathcal{N}$. Nach Satz 6.5 gilt daher $a \preceq \mathcal{N}$ für alle a, daher ist $\mathcal{N}$ ein Nullelement $\mathcal{N}_{\preceq}$ im Sinne der Definition 6.2. Der Beweis fürs Einselement läuft dual.

Satz 6.7 (Kürzungsregel) *In einem disrtributiven Verband gilt für alle a, b, c:*

$$a \sqcup b = a \sqcup c\,, \quad a \sqcap b = a \sqcap c \quad \Rightarrow \quad b = c \qquad .$$

Beweis: Es ist: $b \overset{Gl.(6.2c),\text{ rechts}}{=} b \sqcup (a \sqcap b) \overset{Vor.}{=} b \sqcup (a \sqcap c) \overset{Gl.(6.2d),\text{ rechts}}{=} (b \sqcup a) \sqcap (b \sqcup c) \overset{Gl.(6.2a),\text{ links}}{=} (a \sqcup b) \sqcap (b \sqcup c) \overset{Vor.}{=} (a \sqcup c) \sqcap (b \sqcup c) \overset{Gl.(6.2a),\text{ rechts}}{=} (c \sqcup a) \sqcap (c \sqcup b) \overset{Gl.(6.2d),\text{ rechts}}{=} c \sqcup (a \sqcap b) \overset{Vor.}{=} c \sqcup (a \sqcap c) \overset{Gl.(6.2a),\text{ links}}{=} c \sqcup (c \sqcap a) \overset{Gl.(6.2c),\text{ rechts}}{=} c.$

Satz 6.8 (Doppelte Negation) *In einem distributiven Verband gilt:*

$$\overline{\overline{a}} = a \qquad .$$

Beweis: Da $\overline{a}$ das Komplement von a ist, gilt nach Definition 6.5: $a \sqcap \overline{a} = \mathcal{N}$ und $a \sqcup \overline{a} = \mathcal{E}$. Da $\overline{\overline{a}}$ das Komplement von $\overline{a}$ ist, gilt ebenso: $\overline{a} \sqcap \overline{\overline{a}} = \mathcal{N}$ und $\overline{a} \sqcup \overline{\overline{a}} = \mathcal{E}$. Also gilt auch $\overline{a} \sqcap a \overset{Gl.(6.2a),\text{ links}}{=} a \sqcap \overline{a} \overset{s.o.}{=} \mathcal{N} \overset{s.o.}{=} \overline{a} \sqcap \overline{\overline{a}}$ und $\overline{a} \sqcup a \overset{Gl.(6.2a),\text{ rechts}}{=} a \sqcup \overline{a} \overset{s.o.}{=} \mathcal{E} \overset{s.o.}{=} \overline{a} \sqcup \overline{\overline{a}}$. Nach der Kürzungsregel 6.7 ergibt sich $a = \overline{\overline{a}}$.

Def. 6.6 (Boolsche Algebra) *Eine distributiver Verband bildet eine Boolsche Algebra, wenn der Verband sowohl ein Null- und Einselement besitzt und zusätzlich komplementär ist, das heißt, dass zu jedem Element der Menge das Komplement in der Menge existiert.*

Satz 6.9 *In einer Boolschen Algebra gilt:*

$$\overline{\mathcal{E}} = \mathcal{N} \qquad \textit{und} \qquad \overline{\mathcal{N}} = \mathcal{E} \qquad .$$

Beweis: Für das Komplement $\overline{\mathcal{N}}$ von $\mathcal{N}$ muss gelten: $\mathcal{N} \sqcap \overline{\mathcal{N}} = \mathcal{N}$ und $\mathcal{N} \sqcup \overline{\mathcal{N}} = \mathcal{E}$. $\overline{\mathcal{N}} = \mathcal{E}$ erfüllt diese Bedingungen, denn es ist $\mathcal{N} \sqcap \mathcal{E} = \mathcal{N}$ und $\mathcal{N} \sqcup \mathcal{E} = \mathcal{E}$. Der Beweis für $\overline{\mathcal{E}} = \mathcal{N}$ geht analog.

Satz 6.10 *In einer Boolschen Algebra gilt für alle Elemente a*

$$a \sqcup \mathcal{N} = a \tag{6.4a}$$
$$a \sqcap \mathcal{E} = a \qquad . \tag{6.4b}$$

Beweis: Das folgt aus Axiom Gl. (6.2c) und der Definition des Komplements

$$a \sqcup \underbrace{(a \sqcap \overline{a})}_{\mathcal{N}} = a$$
$$a \sqcap \underbrace{(a \sqcup \overline{a})}_{\mathcal{E}} = a \qquad .$$

Satz 6.11 *Für zwei beliebige Elemente a, b einer Boolschen Algebra gilt:*

$$[a = b] \Leftrightarrow [\overline{a} = \overline{b}] \tag{6.5a}$$

$$\overline{a \sqcup b} = \overline{a} \sqcap \overline{b} \tag{6.5b}$$

$$\overline{a \sqcap b} = \overline{a} \sqcup \overline{b} \quad . \tag{6.5c}$$

Beweis:

Gl. (6.5a): Wir gehen von $a = b$ aus und verknüpfen beide Seiten einmal mit $\overline{a} \sqcap$, dann mit $\overline{a} \sqcup$. Das liefert

$$\underbrace{\overline{a} \sqcap a}_{\mathcal{N}} = \overline{a} \sqcap b\,, \qquad \underbrace{\overline{a} \sqcup a}_{\mathcal{E}} = \overline{a} \sqcup b \quad .$$

Diese Gleichungen stellen also die Definition des Komplements von b dar, d.h. $\overline{a} = \overline{b}$.

Gl. (6.5b): Wenn $\overline{a} \sqcap \overline{b}$ das Komplement $a \sqcup b$ sein soll, dann muss gelten

$$(\overline{a} \sqcap \overline{b}) \sqcup (a \sqcup b) = \mathcal{E} \quad ,$$

$$(\overline{a} \sqcap \overline{b}) \sqcap (a \sqcup b) = \mathcal{N} \quad .$$

Wir nutzen die Distributivität aus und erhalten

$$(\overline{a} \sqcup (a \sqcup b)) \sqcap (\overline{b} \sqcup (a \sqcup b)) = \underbrace{(\underbrace{(\overline{a} \sqcup a)}_{\mathcal{E}} \sqcup b)}_{\mathcal{E}} \sqcap \underbrace{(\underbrace{(\overline{b} \sqcup b)}_{\mathcal{E}} \sqcup a)}_{\mathcal{E}} = \mathcal{E} \quad .$$

Analog erhalten wir

$$((\overline{a} \sqcap \overline{b}) \sqcap a) \sqcup ((\overline{a} \sqcap \overline{b}) \sqcap b) = \underbrace{(\underbrace{(a \sqcap \overline{a})}_{\mathcal{N}} \sqcap \overline{b})}_{\mathcal{N}} \sqcup \underbrace{(\overline{a} \sqcap \underbrace{(\overline{b} \sqcap b)}_{\mathcal{N}})}_{\mathcal{N}} = \mathcal{N} \quad .$$

Somit ist der Beweis erbracht.

Gl. (6.5c): Dual zu oben.

Wir erkennen sofort, dass die bekannten Mengenverknüpfungen $\cap$ und $\cup$, aber ebenso das logische Und bzw. Oder der Aussagenlogik, die geforderten Eigenschaften erfüllen. Diese beiden stellen auch die in der Anwendung wichtigsten Verknüpfungen dar.
Es gibt einen engen Bezug zwischen den beiden bekannten Typen von Verknüpfungen, der logischen und der Mengen-Verknüpfung. Das sei am Beispiel der Ereignisse eines Würfels erläutert. Die Elementarereignisse sind die Aussagen (Propositionen)

"ich würfle die Augenzahl i", kurz e_i. Die Elementarereignisse entsprechen den "Atomen" der Potenzmenge $P(\Omega)$, wobei $\Omega = \{1, 2, 3, 4, 5, 6\}$ (Gesamtmenge), mit der Zuordnung

$$e_i \longleftrightarrow \{i\}, \quad i = 1, 2, 3, 4, 5, 6 \qquad .$$

Weiters gelten die Zuordnungen

$$\{i\} \cup \{j\} \longleftrightarrow e_i \vee e_j \qquad .$$

Weitere Propositionen sind z.B. e_a :"ich würfle eine Augenzahl kleiner 5" oder e_b :"ich würfle eine gerade Augenzahl" Die Zuordnung zur Potenzmenge ist dann

$$e_a \longleftrightarrow \{1, 2, 3, 4\} \; ; \qquad e_b \longleftrightarrow \{2, 4, 6\} \qquad .$$

Die Verknüpfung $\sqcap$ lautet in den beiden Darstellungen

$$e_a \wedge e_b = \text{"ich würfle 2 oder 4"} \longleftrightarrow \{2, 4\} \qquad .$$

Interessant ist noch das Null- und das Einselement. In der Darstellung mit Propositionen ist das Nullelements das unmögliche Ereignis und das Einselement das sichere Ereignis, d.h. irgendeine Augenzahl. In Mengen mit den Mengen-Relationen $\cap$ und $\cup$ ist $\mathcal{N} = \{\emptyset\}$ und $\mathcal{E} = \Omega$ die Gesamtmenge.

6.2.1 Beispiele:

- Mengen mit Mengenverknüpfungen: $\mathcal{N} = \emptyset$, $\mathcal{E} = \Omega$: Gesamtmenge.
- Menge der N-stelligen Bit-Muster mit den elementweisen logischen Verknüpfungen $\wedge$ und $\vee$: $\mathcal{N} = \{0, 0, \ldots, 0\}$, $\mathcal{E} = \{1, 1, \ldots, 1\}$[1]
  Das Komplement ist das elementweise Vertauschen $0 \leftrightarrow 1$.

Def. 6.7 (Potenzmengen) *Als Potenzmenge $P(\Omega)$ bezeichnet man die Menge aller Teilmengen einer Menge Ω.*

Def. 6.8 (Partition) *Die Partition $\mathcal{A} = [\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n]$ einer Menge Ω ist eine Kollektion sich gegenseitig ausschließender* DISJUNKTER *Untermengen $\mathcal{A}_i$ $(i = 1, 2, \ldots, N)$ von Ω,*

$$\mathcal{A}_i \cap \mathcal{A}_j = 0 \qquad \forall i \neq j \qquad ,$$

deren Vereinigung die Gesamtmenge Ω aufspannt

$$\bigcup_{i=1}^{N} \mathcal{A}_i = \Omega \qquad .$$

Theorem 6.1 *Jede endliche Boolsche Algebra ist isomorph zu einer geeignet gewählten Potenzmenge.*

[1] 0 kann als logische falsch und 1 als wahr interpretiert werden.

6.2.2 Normierung

Def. 6.9 (Norm) *Ein in einer Boolschen Algebra $\mathcal{B}$ definiertes Funktional heißt Norm, wenn dieses Funktional*

$$P : a \mapsto P(a)$$

für alle $a, b \in \mathcal{B}$ folgende Eigenschaften hat:

$$P(a) \geq 0 \tag{6.6a}$$

$$P(\mathcal{E}) = 1 \tag{6.6b}$$

$$[a \sqcap b = \mathcal{N}] \Rightarrow \qquad P(a \sqcup b) = P(a) + P(b) \qquad . \tag{6.6c}$$

Eine Boolsche Algebra, in der eine Norm definiert ist, heißt normiert.

Es ist zu erkennen, dass die Eigenschaft der Norm stark an die Definition der klassischen Wahrscheinlichkeit angelehnt sind.
Es kann nicht immer allen Untermengen von Ω widerspruchsfrei eine Norm (Maß) zugeordnet werden. Da wir die Norm mit der Wahrscheinlichkeit gleichsetzen werden, kann eine Wahrscheinlichkeitstheorie nur auf solchen Untermengen widerspruchsfrei aufgebaut werden, bei denen die Normzuweisung widerspruchsfrei möglich ist. Diese Untermengen bilden eine σ-Algebra, auch σ-Körper genannt.

Def. 6.10 (σ-Körper) *Ein Körper ist in diesem Zusammenhang eine Klasse $\mathcal{F}$ von Mengen, so dass für jedes Paar $A, B \in \mathcal{F}$ auch $A \sqcap B \in \mathcal{F}$ und $A \sqcup B \in \mathcal{F}$. Darüber hinaus verlangt man, dass zu jedem $A \in \mathcal{F}$ auch das Komplement $\overline{A}$ zu $\mathcal{F}$ gehört.*
Ein σ-Körper besitzt zusätzlich die Eigenschaft, dass die Verknüpfung von abzählbar vielen Elementen A_i, $\quad (i = 1, 2, \ldots, N)$[2] aus $\mathcal{F}$ auch immer zu $\mathcal{F}$ gehören

$$\bigsqcap_{i=1}^{N} A_i \quad \in \mathcal{F}$$

$$\bigsqcup_{i=1}^{N} A_i \quad \in \mathcal{F} \qquad .$$

Im Fall beliebiger Verknüpfungen spricht man auch von der BOOLSCHEN ALGEBRA. Im Spezialfall der Mengen-Verknüpfungen nennt man den σ-Körper auch BOREL-KÖRPER.[3]

Beispiel:

- Menge der ganzen Zahlen $\{1, 2, 3, 4, 5, 6\}$. Der σ-Körper sind alle Untermengen, die aus den Elementarmengen $\{i\}$ erzeugt werden können.
- Ω sei die Menge der Punkte auf der reellen Achse. Untermengen sind dann alle Mengen von reellen Zahlen. Ein σ-Körper sind die Intervalle $\{x_1 \leq x \leq x_2\}$ und die Vereinigungen und Durchschnitte abzählbar vieler davon. Dieser Körper enthält alle offenen und geschlossenen Intervalle und alle Punkte, insbesondere auch $(-\infty, \lambda], \ \lambda \in \mathbb{R}$.

[2] $N = \infty$ ist auch möglich!
[3] In der englischen Literatur spricht man vom Borel-Field, wenn gleichzeitig eine Norm definiert ist.

Kapitel 7

Axiomatische Wahrscheinlichkeitstheorie

Die axiomatische Definition des Wahrscheinlichkeitsbegriffes geht auf Kolmogoroff (1933) zurück. Der Zugang hat einen deduktiven Charakter, vermeidet konzeptionelle Willkür und liefert zumindest einen ersten Zugang zu einem tieferen Verständnis. Allerdings ist immer noch nicht klar, wie die Norm (Wahrscheinlichkeit) zu interpretieren ist und ob sie in allen Fällen mit dem gesunden Menschenverstand übereinstimmt. Ein noch größeres Problem ist, dass die Zuweisung der Wahrscheinlichkeit zu den Ergebnissen von der Theorie nicht vorgegeben wird und bei weitem nicht eindeutig ist.

Def. 7.1 (Axiomatische Definition der Wahrscheinlichkeit) *Die Wahrscheinlichkeitstheorie ist die Theorie der normierten Boolschen Algebren. Der Wahrscheinlichkeitsraum ist durch drei Bestandteile ausgezeichnet*

- *Die Gesamtmenge Ω aller interessierenden "Ereignisse", bzw. experimentellen Resultate.*
- *Der σ-Körper der "Ereignisse", eine Klasse von Untermengen von Ω.*
- *Die Normen der Elemente des σ-Körpers.*

Die Wahrscheinlichkeit eines Ereignisses (Ereignisse sind Untermengen von Ω), das durch das Element des σ-Körpers repräsentiert wird, ist die Norm des Elementes.

Die axiomatische Definition der Wahrscheinlichkeit ist eine Verallgemeinerung der klassischen bzw. der statistischen Definition. Natürlich, löst die axiomatische Definition nicht das Problem der Zuweisung von Prior-Wahrscheinlichkeiten. Aber sie bildet ein mathematisch konsistentes Gerüst zum Rechnen mit Wahrscheinlichkeiten.

7.1 Regeln der Wahrscheinlichkeitsrechnung

Für die Wahrscheinlichkeit einer normierten Ereignis-Algebra B gilt für zwei beliebige Elemente $a, b \in B$ und für indizierte, sich gegenseitig ausschließende Ereignisse $a_i \in B$

$$P(\overline{a}) = 1 - P(a) \tag{7.1a}$$
$$P(a) \leq 1 \tag{7.1b}$$
$$P(\mathcal{N}) = 0 \tag{7.1c}$$
$$[a \preceq b] \Rightarrow \quad P(a) \leq P(b) \tag{7.1d}$$
$$P(a \sqcup b) = P(a) + P(b) - P(a \sqcap b) \tag{7.1e}$$
$$\text{(Summenregel)} \qquad P(\sqcup_{i=1}^{n} a_i) = \sum_{i=1}^{n} P(a_i) \quad . \tag{7.1f}$$

Falls $[a_1, a_2, \ldots]$ eine Partition von B ist, gilt weiter

$$P(\sqcup_i a_i) = \sum_i P(a_i) = 1 \tag{7.2a}$$
$$\text{(Marginalisierungsregel)} \qquad P(b) = P((\sqcup_i a_i) \sqcap b) = \sum_i P(a_i \sqcap b) \quad . \tag{7.2b}$$

Beweise:

- Gl. (7.1a): $1 = P(\mathcal{E}) = P(a \sqcup \overline{a}) \overset{Gl.(6.6c)}{=} P(a) + P(\overline{a})$.
- Gl. (7.1b): sonst wäre $P(\overline{a}) < 0$ im Widerspruch zu Gl. (6.6a).
- Gl. (7.1c): $P(\mathcal{N}) = P(\overline{\mathcal{E}}) = 1 - P(\mathcal{E}) = 1 - 1 = 0$.
- Gl. (7.1d): Gemäß Definition 6.5 bedeutet die Halbordnung $a \preceq b$, dass $a \sqcup b = b$.

  $$b = \mathcal{E} \sqcap b = (a \sqcup \overline{a}) \sqcap b \overset{a \preceq b}{=} (a \sqcup \overline{a}) \sqcap (a \sqcup b) \overset{Gl.(6.2d)}{=} a \sqcup \underbrace{(\overline{a} \sqcap b)}_{b \backslash a}$$

  Die Teilmenge $b \backslash a$ ist disjunkt zu a

  $$a \sqcap b \backslash a = a \sqcap (\overline{a} \sqcap b) \overset{Gl.(6.2b)}{=} \underbrace{(a \sqcap \overline{a})}_{\mathcal{N}} \sqcap b = \mathcal{N}$$

  Damit gilt gemäß Gl. (6.6c)

  $$P(b) = P(a \sqcup b \backslash a) = P(a) + P(b \backslash a) \overset{Gl.(6.6a)}{\geq} P(a)$$

- Gl. (7.1e): Um diese Form der Summenregel zu beweisen, schreiben wir die Ereignisse $a \sqcup b$ und b als Vereinigung ($\sqcup$) von sich gegenseitig ausschließenden Ereignissen

$$a \sqcup b = a \sqcup (\overline{a} \sqcap b); \qquad b = (a \sqcap b) \sqcup (\overline{a} \sqcap b) \qquad ,$$

  auf die nun die Norm-Eigenschaft Gl. (6.6c) anwendbar ist

$$P(a \sqcup b) = P(a) + P(\overline{a} \sqcap b); \qquad P(b) = P(a \sqcap b) + P(\overline{a} \sqcap b) \qquad .$$

  Wir lösen die zweite Gleichung nach $P(\overline{a} \sqcap b)$ auf und setzen das Ergebnis in die erste Gleichung ein und erhalten somit die gesuchte Summenregel für den Fall "überlappender" Ereignisse.

- Gl. (7.1f): Folgt leicht durch Induktion. Es gelte für $P(\sqcup_i^n a_i)$. Dann sind $\sqcup_i^n a_i$ und a_{n+1} sich gegenseitig ausschließende Ereignisse auf die Gl. (6.6c) anwendbar ist und somit gilt Gl. (7.1f) auch für $n + 1$.

- Gl. (7.2a): Folgt sofort aus der Tatsache, dass $\mathcal{E} = \sqcup_i a_i$ und $P(\mathcal{E}) = 1$.

- Gl. (7.2b): Folgt ebenso aus der Tatsache, dass $\mathcal{E} = \sqcup_i a_i$ zusammen mit $b = b \sqcap \mathcal{E}$.

7.2 Bedingte Wahrscheinlichkeiten

Wir hatten bereits früher darauf hingewiesen, dass es nur bedingte Wahrscheinlichkeiten gibt. Im Rahmen der axiomatischen Wahrscheinlichkeitstheorie verbirgt sich der Bedingungskomplex hinter der Bedeutung und Struktur des Wahrscheinlichkeitsraumes, also der Angabe aller Experimente (Ω), der interessierenden Ereignisse (σ-Algebra) und der Prior-Wahrscheinlichkeiten. Insofern handelt es sich zwar auch hier um bedingte Wahrscheinlichkeiten, der Bedingungskomplex wird aber nicht explizit in die Notation aufgenommen, da er sich während der Analyse eines Problems nicht ändert. Das geht in den meisten Fällen gut, hat aber in der Literatur zu Paradoxa geführt, die Jahrzehnte lang unverstanden waren.
Neben der Abhängigkeit der Wahrscheinlichkeit vom Bedingungskomplex, kann man natürlich auch nach der Wahrscheinlichkeit $P(a|b)$ für ein Ereignis a fragen, unter der Annahme, dass das Ereignis b vorliegt. In Anlehnung an die Regeln der klassischen Wahrscheinlichkeitstheorie definiert man in der axiomatischen Theorie die

BEDINGTE WAHRSCHEINLICHKEIT

$$P(a|b) = \frac{P(a \sqcap b)}{P(b)} \qquad . \tag{7.3}$$

Man überzeugt sich leicht davon, dass bedingte Wahrscheinlichkeiten $P(a|x)$ ebenfalls alle Regeln der Wahrscheinlichkeitstheorie erfüllen; schließlich sind sie nichts anderes als Wahrscheinlichkeiten.
Aus der Definition der bedingten Wahrscheinlichkeit ergibt sich sofort die

PRODUKTREGEL

$$P(a \sqcap b) = P(a|b) * P(b) = P(b|a) * P(a) \quad , \tag{7.4}$$

und durch Auflösen der beiden möglichen Darstellungen der Produktregel erhalten wir das Bayessche Theorem

BAYESSCHES THEOREM

$$P(a|b) = \frac{P(b|a) * P(a)}{P(b)} \quad , \tag{7.5}$$

das es erlaubt, inverse Probleme zu lösen. Man interessiert sich in sehr vielen Fällen für die Posterior-Wahrscheinlichkeit $P(x|d)$ für ein Ereignis x, gegeben experimentelle Daten d und der Bedingungskomplex. Die Posterior-Wahrscheinlichkeit ist gemäß des Bayesschen Theorems verknüpft mit der Prior-Wahrscheinlichkeit $P(x)$, der sogenannte Daten-Evidenz $P(d)$[1] und der Likelihood-Funktion $P(d|x)$.
Man kann die Produkt-Regel verallgemeinern und in eine für die Anwendung sehr nützliche Gestalt bringen. Es sei $[a_1, \ldots, a_n]$ eine Partitionierung der Gesamtmenge Ω bzw. des sicheren Ereignisses und b ein beliebiges Element der σ-Algebra. Dann gilt wegen der Marginalisierungsregel Gl. (7.2b)

MARGINALISIERUNGSREGEL

$$P(b) = \sum_{i=1}^{n} P(a_i \sqcap b) = \sum_{i=1}^{n} P(b|a_i)\ P(a_i) \quad . \tag{7.6}$$

Damit lässt sich das Bayessche Theorem auch in folgender Form schreiben:

[1] Die Daten-Evidenz spielt i.d.R. nur die Rolle des Normierungsfaktors.

$$P(a_i|b) = \frac{P(b|a_i)P(a_i)}{\sum_j P(b|a_j)\ P(a_j)} \tag{7.7}$$

Def. 7.2 (Unabhängigkeit) *Man nennt zwei Ereignisse der σ-Algebra logisch unabhängig, wenn*

$$P(a \sqcap b) = P(a)P(b) \quad \Longleftrightarrow \quad P(a|b) = P(a)$$

Kapitel 8

Bayessche Wahrscheinlichkeitstheorie

8.1 Was ist Wahrscheinlichkeit?

Im Lexikon findet man zwei Erklärungen:
1. Wahrscheinlichkeit ist ein Begriff, der die Einstufung von Aussagen oder Urteilen nach dem Grad ihres Geltungsanspruchs bzw. Möglichkeit und Gewissheit bezeichnet, wobei die Gründe für den Geltungsanspruch nicht oder noch nicht ausreichen, um die Annahme des Gegenteils auszuschließen.
2. speziell in Mathematik, Naturwissenschaft und Statistik: Wahrscheinlichkeit ist der Grad der Möglichkeit bzw. Voraussagbarkeit (Prognostizierbarkeit) des Eintretens eines Ereignisses.
Wahrscheinlichkeit ist somit ein Maß für den Wahrheitsgehalt bzw. Wahrheitsgrad einer Proposition.

Es gibt keine unbedingten (absoluten) Wahrscheinlichkeiten!

Z.B. die Wahrscheinlichkeit $P(Z|\mathcal{B})$, beim Wurf einer Münze „Zahl" vorzufinden, ist nur dann $1/2$, wenn vorausgesetzt wird, dass die Münze nur zwei Möglichkeiten hat, zur Ruhe zu kommen (auf der Kante ist ausgeschlossen) und wenn sichergestellt ist, dass keine der beiden Seiten ausgezeichnet ist.
Auch wenn dieser Bedingungskomplex nahezu selbsterklärend ist, ist nur durch diese zumeist implizite Angabe die Fragestellung eindeutig definiert.

Die bedingte Wahrscheinlichkeit $P(X|Y)$ kann somit auch als Maß dafür verstanden werden, wie stark die Proposition Y die Proposition X impliziert. Man kann Wahrscheinlichkeit deshalb auch als IMPLIKATIONSMASS betrachten. Eigentlich ist es genau diese Interpretation, die uns in echten Problemen interessiert. Das Ziel ist immer auszuarbeiten, wie stark die vorliegenden Informationen inklusive experimenteller Daten implizieren, dass ein Parameter bestimmte Werte annimmt, oder dass eine Hypothese wahr ist, oder dass ein Modell die Daten beschreibt.
Wir werden sehen, dass dieses Maß eindeutig definiert ist und dass aus den Regeln der elementaren Logik der Kalkül für dieses Maß folgt.

Hier werden wir die Summen- und Produktregel der Wahrscheinlichkeitstheorie von einfachen Konsistenzforderungen der elementaren Aussagenlogik ableiten.

8.2 Das Universalgesetz der Wahrscheinlichkeitstheorie

Die Wahrheitswerte von Propositionen gehorchen der Boolschen Algebra. Diese implizieren korrespondierende Regeln für die Wahrscheinlichkeiten. Wir werden diese Regeln aus denen der Aussagenlogik ableiten. Diese Strategie geht auf R.T. Cox (1946) zurück, der damit, 2 Jahrhunderte nachdem die Summen- und Produktregel eingeführt worden sind, die Rechtfertigung geliefert hat, dass man die bekannten Regeln der Wahrscheinlichkeitsrechnung auch auf das Maß des Wahrheitsgehalts einer Proposition, bzw. das Implikationsmaß, anwenden kann. Damit ist es gelungen, die Axiome der axiomatischen Wahrscheinlichkeitstheorie aus Konsistenzforderungen herzuleiten.
Im Fall, dass eindeutige Aussagen möglich sind, reduziert sich alles auf die gewöhnliche Aussagenlogik.
Wir benötigen im Folgenden die logische Verknüpfung „NAND"

$$A \uparrow B := \overline{A \wedge B} \quad . \tag{8.1}$$

Die Regeln der Wahrscheinlichkeitsrechnung lassen sich aus einer einzigen Regel, der für das logische „NAND",

$$P(A \uparrow B|\mathcal{B})$$

ableiten.[1] Das ist möglich, da aus dem logischen „NAND" alle logischen Verknüpfungen aufgebaut werden können.

8.3 Aussagenlogik

Es sollen hier die wichtigsten Eigenschaften der Aussagenlogik zusammengefasst werden[2].

Proposition: Aussage, die entweder wahr („TRUE", T) oder falsch („FALSE", F) ist.

Negation: Zu jeder Proposition existiert das Inverse $\overline{A}$, mit $\overline{\overline{A}} = A$.

Logisches UND: $A \wedge B$ ist dann und nur dann wahr, wenn beide Propositionen A und B wahr sind. Als Argument von Wahrscheinlichkeiten schreiben wir die UND-Verknüpfung zweier Propositionen in der Form A, B.

[1] Dieser Beweis geht ursprünglich auf R.T.Cox (1946) zurück und wurde 1998 von A.Garrett vereinfacht.

[2] Eine gute Darstellung der Boolschen Algebra findet man in Kuntzmann, J. 1967. FUNDAMENTAL BOOLEAN ALGEBRA, London, UK.

Logisches ODER: $A \vee B$ ist dann und nur dann falsch, wenn beide Proposition falsch sind.

WICHTIGE EIGENSCHAFTEN DER BOOLSCHEN ALGEBRA

$$A \wedge A = A \tag{8.2a}$$
$$A \wedge B = B \wedge A \quad \text{Kommutativität} \tag{8.2b}$$
$$(A \wedge B) \wedge C = A \wedge (B \wedge C) = A \wedge B \wedge C \quad \text{Assoziativität} \tag{8.2c}$$
$$\overline{A \wedge B} = \overline{A} \vee \overline{B} \tag{8.2d}$$
$$A \wedge T = A \tag{8.2e}$$
$$A \wedge F = F \quad . \tag{8.2f}$$

Wir erkennen, dass eine Dualität zwischen UND und ODER-Verknüpfung besteht,

$$A \vee B = \overline{\overline{A} \wedge \overline{B}} \quad .$$

Es genügt deshalb, das logische „NAND" um alle logischen Funktionen aufzubauen[3].

$$\overline{A} = A \uparrow A \tag{8.3a}$$
$$A \wedge B = (A \uparrow B) \uparrow (A \uparrow B) \tag{8.3b}$$
$$A \vee B = (A \uparrow A) \uparrow (B \uparrow B) \quad . \tag{8.3c}$$

8.4 Herleitung der Wahrscheinlichkeitsrechnung

Wir wollen die Herleitung der Regeln der Wahrscheinlichkeitsrechnung hier explizit vorführen, da sie ungewohnte und interessante Elemente enthält und weil es hierbei klar wird, dass die Regeln der Wahrscheinlichkeitsrechnung eine logische Konsequenz elementarer Konsistenzforderungen sind.
Wir drücken die Wahrscheinlichkeit $P(A \uparrow B|\mathcal{B})$ als Funktion einfacherer bedingter Wahrscheinlichkeiten aus. Welche Möglichkeiten gibt es hierzu. Zunächst kommen keine logischen Ausdrücke, die komplexer sind als die Ausgangsverknüpfung $A \uparrow B$, in Frage. Da der Bedingungskomplex $\mathcal{B}$ vorgegeben ist, bleibt er in allen Wahrscheinlichkeiten hinter dem Bedingungsstrich erhalten. Es gibt vor dem Bedingungsstrich

[3] Jeder logische Schaltkreis kann allein aus „NAND"-Gattern aufgebaut werden.

nur die Möglichkeiten A bzw. B. Alle anderen Ausdrücke können durch NAND-Verknüpfungen hierdurch ausgedrückt werden. Hinter dem Bedingungsstrich wäre zusätzlich $A \uparrow B$ als Proposition denkbar, aber auch das liefert nichts Neues

$$P(A|A \uparrow B, \mathcal{B}) = \begin{cases} P(A|\overline{A}, \mathcal{B}) = P_F & \text{für } B = T, \\ P(A|T, \mathcal{B}) = P(A|\mathcal{B}) & \text{für } B = F \end{cases} .$$

Hier beschreibt P_T die Wahrscheinlichkeit für die wahre (T) und P_F die Wahrscheinlichkeit für die falsche (F) Proposition.
Analoges gilt für $P(B|A \uparrow B, \mathcal{B})$. Somit lautet das allgemeinste Funktional

$$P(A \uparrow B|\mathcal{B}) = \mathcal{F}\big(\, P(A|B, \mathcal{B}), P(B|\mathcal{B}), P(B|A, \mathcal{B}), P(A|\mathcal{B}) \,\big) \quad . \tag{8.4}$$

Der Bedingungskomplex hängt nicht von A, $\overline{A}$, B or $\overline{B}$ ab, außer wenn es explizit erwähnt wird.
Wir werden nun Gl. (8.4) weiter vereinfachen. Dazu betrachten wir den Spezialfall $B = A \wedge C$. B kann zwar nicht mehr unabhängig von A beliebige Werte annehmen, aber es gibt keine Einschränkung von A oder C. Auf der linken Seite wird $A \uparrow B$ zu

$$A \uparrow B = A \uparrow (A \wedge C) = \overline{A \wedge A \wedge C} = \overline{A \wedge C} = A \uparrow C \quad . \tag{8.5}$$

Die rechte Seite vereinfacht sich, da

$$P(A|A, C, \mathcal{B}) = P_T, \qquad P(A, C|A, \mathcal{B}) = P(C|A, \mathcal{B}) \, . \tag{8.6}$$

Wir haben somit

$$P(A \uparrow C|\mathcal{B}) = \mathcal{F}\big(P_T, P(A, C|\mathcal{B}), P(C|A, \mathcal{B}), P(A|\mathcal{B})\big) \tag{8.7}$$

oder durch Umbenennen $C \to B$,

$$P(A \uparrow B|\mathcal{B}) = \mathcal{F}\big(P_T, P(A, B|\mathcal{B}), P(B|A, \mathcal{B}), P(A|\mathcal{B})\big) \quad . \tag{8.8}$$

Diese Beziehung muss mit Gl. (8.4) verglichen werden. Wir haben also bereits erreicht, dass $P(A \uparrow B|\mathcal{B})$ nur von drei Wahrscheinlichkeiten, $P(A, B|\mathcal{B})$, $P(B|A, \mathcal{B})$ und $P(A|\mathcal{B})$, abhängt.
Nun betrachten wir den Spezialfall $B = A$ in Gl. (8.8). Wegen $A \uparrow A = \overline{A}$ erhalten wir

$$P(\overline{A}|\mathcal{B}) = \mathcal{F}\big(P_T, P(A|\mathcal{B}), P_T, P(A|\mathcal{B})\big) \quad . \tag{8.9}$$

Diese Beziehung besagt, dass es einen direkten Bezug zwischen der Wahrscheinlichkeit einer Proposition und ihrer Negation gibt

$$P(\overline{A}|\mathcal{B}) = \Xi\big(P(A|\mathcal{B})\big) \tag{8.10}$$

mit

$$\Xi(z) = \mathcal{F}(P_T, z, P_T, z) \quad . \tag{8.11}$$

Um $P(A,B|\mathcal{B})$ auf der rechten Seite von Gl. (8.8) durch $P(A\uparrow B|\mathcal{B})$ auszudrücken, ersetzen wir die Proposition A in Gl. (8.10) durch $A\uparrow B$ und erhalten zusammen mit Gl. (8.3b)

$$P(A,B|\mathcal{B}) = P(\overline{A\uparrow B}|\mathcal{B}) = \Xi\big(P(A\uparrow B|\mathcal{B})\big) \quad . \tag{8.12}$$

Wenn wir diese Relation auf der rechten Seite von Gl. (8.8) einsetzen, erhalten wir

$$P(A\uparrow B|\mathcal{B}) = \mathcal{F}\big(P_T, \Xi\big(P(A\uparrow B|\mathcal{B})\big), P(B|A,\mathcal{B}), P(A|\mathcal{B})\big) \quad . \tag{8.13}$$

Nun gibt es zwei Möglichkeiten:
a) Die rechte Seite von Gl. (8.13) ist identisch zu $P(A\uparrow B|\mathcal{B})$ und hängt gar nicht von $P(B|A,\mathcal{B})$ oder $P(A|\mathcal{B})$ ab. In diesem Fall ist Gl. (8.13) eine Tautologie und bringt nichts Neues.
b) (8.13) stellt eine implizite Beziehung zwischen $P(A\uparrow B|\mathcal{B})$, $P(B|A,\mathcal{B})$ und $P(A|\mathcal{B})$ dar. In diesem Fall ist es uns gelungen, Gl. (8.4) weiter zu vereinfachen.
Wir gehen davon aus, dass es tatsächlich eine nicht-triviale Lösung gibt

$$P(A\uparrow B|\mathcal{B}) = \mathcal{G}(P(B|A,\mathcal{B}), P(A|\mathcal{B})) \quad . \tag{8.14}$$

Hierbei stellt $\mathcal{G}$ ein neues unbekanntes Funktional dar. Wenn der Fall a) zutreffen sollte, würden wir feststellen, dass Gl. (8.14) keine Lösung besitzt. Wir werden nun das Funktional $\mathcal{G}$ bestimmen. Dazu setzen wir noch einmal $B = A$ an und erhalten

$$P(\overline{A}|\mathcal{B}) = \mathcal{G}\big(P_T, P(A|\mathcal{B})\big) \quad . \tag{8.15}$$

Es ist nützlich, die Definition

$$\chi(v) = \mathcal{G}(P_T, v) \tag{8.16}$$

einzuführen, mit der wir

$$P(\overline{A}|\mathcal{B}) = \chi\big(P(A|\mathcal{B})\big) \tag{8.17}$$

erhalten. Wenn wir in Gl. (8.17) A durch $\overline{A}$ ersetzen und Gl. (8.17) erneut verwenden, folgt

$$P(A|\mathcal{B}) = \chi\big(P(\overline{A}|\mathcal{B})\big) = \chi\big(\chi(P(A|\mathcal{B}))\big)$$

für alle A und somit für alle Werte von $P(A|\mathcal{B})$. Demnach gilt

$$\chi\circ\chi(z) = z \quad . \tag{8.18}$$

Das heißt, $\chi\circ\chi = \mathcal{I}$, und somit ist χ gleich seinem Inversen.
Als nächstes wenden wir uns dem logischen „UND" zu.

$$\begin{aligned} P(A,B|\mathcal{B}) &= P(\overline{A\uparrow B}|\mathcal{B}) = \chi(P(A\uparrow B|\mathcal{B})) = \chi(\mathcal{G}(P(B|A,\mathcal{B}), P(A|\mathcal{B}))) \\ &= \psi\big(P(B|A,\mathcal{B}), P(A|\mathcal{B})\big) \qquad (8.19) \\ \psi(a,b) &:= \chi(\mathcal{G}(a,b)) \quad . \qquad (8.20) \end{aligned}$$

Hierbei wurde Gl. (8.14) verwendet. Wir können das logische Produkt weiter spezifizieren, indem wir Gl. (8.19) verwenden, um das logische Produkt von drei Propositionen (A, B, C) zu berechnen. Das Produkt kann auf mehrere Arten ausgewertet

werden, da die Kommutativität und Assoziativität gilt. Hierbei geht nun erneut die Konsistenz-Forderungen ein.
Wir beginnen mit

$$\begin{aligned} P(A,B,C|\mathcal{B}) &= P\big((A,B),C|\mathcal{B}\big) \\ &= \psi\big(P(A,B|C,\mathcal{B}),P(C|\mathcal{B})\big) \\ &= \psi\Big(\psi\big(P(A|B,C,\mathcal{B}),P(B|C,\mathcal{B})\big),P(C|\mathcal{B})\Big) \quad . \end{aligned}$$

Alternativ gilt auch

$$\begin{aligned} P(A,B,C|\mathcal{B}) &= P\big(A,(B,C)|\mathcal{B}\big) \\ &= \psi\big(P(A|B,C,\mathcal{B}),P(B,C|\mathcal{B})\big) \\ &= \psi\Big(P(A|B,C,\mathcal{B}),\psi\big(P(B|C,\mathcal{B}),P(C|\mathcal{B})\big)\Big) \quad . \end{aligned}$$

Die rechten Seiten beider Alternativen müssen gleich sein. Mit den Abkürzungen

$$a = P(A|B,C,\mathcal{B}), \qquad b = P(B|C,\mathcal{B}), \qquad c = P(C|\mathcal{B}), \tag{8.21}$$

erhalten wir für die Funktionalgleichung

$$\psi\big(a,\psi(b,c)\big) = \psi\big(\psi(a,b),c\big) \quad . \tag{8.22}$$

Die Variablen a, b und c sind voneinander unabhängig.
Man nennt Gl. (8.22) Assoziativitätsgleichung. Sie wurde von Abel bereits im 19. Jahrhundert eingeführt und untersucht. Sie hat die allgemeine Lösung

$$\psi(u,v) = \varphi^{-1}\big(\varphi(u)\varphi(v)\big) \quad , \tag{8.23}$$

wobei φ eine stetige Funktion mit dem Inversen φ^{-1} ist. Beide Funktion müssen eindeutig sein. Ansonsten ist φ beliebig. Die Herleitung der Lösung ist langwierig. Es ist jedoch leicht, sich davon zu überzeugen, dass sie Gl. (8.22) erfüllt. Einsetzen in Gl. (8.22) liefert für die linke Seite

$$\begin{aligned} \psi\big(a,\psi(b,c)\big) &= \varphi^{-1}\big(\varphi(a)\varphi(\psi(b,c))\big) \\ &= \varphi^{-1}\big(\varphi(a)\varphi(\varphi^{-1}\big(\varphi(b)\varphi(c)\big))\big) \\ &= \varphi^{-1}\big(\varphi(a)\varphi(b)\varphi(c)\big) \quad . \end{aligned}$$

Dasselbe Ergebnis erhalten wir für die rechte Seite

$$\begin{aligned} \psi\big(\psi(a,b),c\big) &= \varphi^{-1}\big(\varphi(\psi(a,b))\varphi(c)\big) \\ &= \varphi^{-1}\big(\varphi(\varphi^{-1}\big(\varphi(a)\varphi(b)\big))\varphi(c)\big) \\ &= \varphi^{-1}\big(\varphi(a)\varphi(b)\varphi(c)\big) \quad . \end{aligned}$$

Wir setzen nun die Lösung, Gl. (8.23), in Gl. (8.19) für das logische „UND" ein und wenden auf beide Seiten die Funktion φ an

$$\varphi(P(A,B|\mathcal{B})) = \varphi(P(B|A,\mathcal{B}))\ \varphi(P(A|\mathcal{B})) \quad . \tag{8.24}$$

Um diese Relation in die bekannte Form der Produktregel zu bringen, tauschen wir das ohnehin noch nicht genauer spezifizierte Funktional $P(\ldots)$ gegen $P'(\ldots)$ ein

$$P'(\ldots) = \varphi(P(\ldots))$$

und erhalten damit

$$P'(A, B|\mathcal{B}) = P'(B|A, \mathcal{B})\ P'(A|\mathcal{B}) \qquad . \tag{8.25}$$

Hieraus können wir nun die Grenzwerte ermitteln. Z.B. liefert $A = T$

$$P'(B|\mathcal{B}) = P'(T, B|\mathcal{B}) = P'(B|T, \mathcal{B})\ \underbrace{P'(T|\mathcal{B})}_{P'_T} = P'(B|\mathcal{B})\ P'_T \qquad .$$

Daraus folgt zwingend für die Darstellung P', in der die Produktregel in der einfachen Form gilt,

$$P(T|\mathcal{B}) = 1 \qquad . \tag{8.26}$$

Weiter erhalten wir für $B = F$

$$P'(F|\mathcal{B}) = P'(A, F|\mathcal{B}) = P'(F|A, \mathcal{B})\ P'(A|\mathcal{B}) = P'(F|\mathcal{B})\ P'(A|\mathcal{B}) \qquad .$$

Eine Lösung hiervon ist

$$P'(F|\mathcal{B}) = 0 \qquad . \tag{8.27}$$

Als weitere Lösung ist $P'(F|\mathcal{B}) = \infty$ ebenso möglich. Es ergeben sich hieraus zwei mögliche Abbildungen von Propositionen auf die reellen Zahlen. Entweder wie gewohnt auf das Intervall $[0, 1]$ oder auf das Intervall $[1, \infty)$. Wir werden später sehen, dass zwischen diesen beiden Darstellungen die Beziehung $P' \leftrightarrow 1/P'$ besteht. Wir werden hier aber zunächst die Standard-Darstellung weiter verfolgen.
Schließlich bleibt noch das logische „NOT" zu bestimmen. In Gl. (8.17) hatten wir bereits

$$P(\overline{A}|\mathcal{B}) = \chi(P(A|\mathcal{B}))$$

abgeleitet. Anwenden von φ auf beide Seiten liefert

$$\begin{aligned} \varphi(P(\overline{A}|\mathcal{B})) &= \varphi \circ \chi \circ \varphi^{-1} \circ \varphi(P(A|\mathcal{B})) \\ P'(\overline{A}|\mathcal{B}) &= \underbrace{\varphi \circ \chi \circ \varphi^{-1}}_{\Theta}(P'(A|\mathcal{B})) \\ P'(\overline{A}|\mathcal{B}) &= \Theta(P'(A|\mathcal{B})) \qquad . \end{aligned}$$

Wenn wir, wie zuvor, diese Gleichung auf die doppelte Verneinung zweimal anwenden erhalten wir $\Theta \circ \Theta = \mathcal{I}$. Das heißt $\Theta^{-1} = \Theta$.
Um das Funktional Θ zu bestimmen, betrachten wir speziell

$$P'(\overline{A}, \overline{B}|\mathcal{B}) = P'(\overline{A}|\overline{B}, \mathcal{B})\ P'(\overline{B}|\mathcal{B})$$

und ersetzen alle Negationen durch das Funktional Θ.

$$P'(\overline{A}, \overline{B}|\mathcal{B}) = P'(\overline{A}|\overline{B}, \mathcal{B})\ P'(\overline{B}|\mathcal{B}) = \Theta\left[P'(A|\overline{B}, \mathcal{B})\right]\ \Theta\left[P'(B|\mathcal{B})\right]$$

Die Negation hinter dem Bedingungsstrich werden wir dadurch los, dass wir ausnutzen, dass das logische „UND" kommutativ ist und wir die Produktregel auf unterschiedliche Weise anwenden können

$$\begin{aligned} P'(A, \overline{B}|\mathcal{B}) &= P'(A|\overline{B}, \mathcal{B})\ P'(\overline{B}|\mathcal{B}) \\ P'(A, \overline{B}|\mathcal{B}) &= P'(\overline{B}|A, \mathcal{B})\ P'(A|\mathcal{B}) \qquad . \end{aligned}$$

Gleichsetzen der rechten Seiten liefert

$$P'(A|\overline{B}, \mathcal{B}) = \frac{P'(\overline{B}|A, \mathcal{B})\ P'(A|\mathcal{B})}{P'(\overline{B}|\mathcal{B})} = \frac{\Theta\left[P'(B|A, \mathcal{B})\right]\ P'(A|\mathcal{B})}{\Theta\left[P'(B|\mathcal{B})\right]} \qquad . \tag{8.28}$$

Somit haben wir schließlich

$$\begin{aligned} P'(\overline{A}, \overline{B}|\mathcal{B}) &= \Theta\left[\frac{\Theta\left[P'(B|A, \mathcal{B})\right]\ P'(A|\mathcal{B})}{\Theta\left[P'(B|\mathcal{B})\right]}\right]\ \Theta\left[P'(B|\mathcal{B})\right] \\ &= \Theta\left[\frac{\Theta\left[\frac{P'(A,B|\mathcal{B})}{P'(A|\mathcal{B})}\right]\ P'(A|\mathcal{B})}{\Theta\left[P'(B|\mathcal{B})\right]}\right]\ \Theta\left[P'(B|\mathcal{B})\right] \qquad . \end{aligned}$$

Wir definieren nun

$$a \equiv P'(A|\mathcal{B})\ , \quad b \equiv P'(B|\mathcal{B})\ , \quad c \equiv P'(A, B|\mathcal{B}). \tag{8.29}$$

Damit erhalten wir die etwas handlichere Form

$$P'(\overline{A}, \overline{B}|\mathcal{B}) = \Theta\left[b\right]\ \Theta\left[\frac{a\ \Theta\left[\frac{c}{a}\right]}{\Theta\left[b\right]}\right] \qquad . \tag{8.30}$$

Nun ist das logische „UND" kommutativ, und wir sollten dasselbe erhalten, wenn wir A und B vertauschen. Das entspricht dem Vertauschen von a und b auf der rechten Seite. Da die linken Seiten dabei invariant sind, muss also gelten

$$\Theta\left[a\right]\ \Theta\left[\frac{b\ \Theta\left[\frac{c}{b}\right]}{\Theta\left[a\right]}\right] = \Theta\left[b\right]\ \Theta\left[\frac{a\ \Theta\left[\frac{c}{a}\right]}{\Theta\left[b\right]}\right] \ . \tag{8.31}$$

Aus dieser Funktionalgleichung soll nun Θ, unter Berücksichtigung der Nebenbedingungen $\Theta \circ \Theta = \mathcal{I}$ und

$$\Theta\left[P'(T|\mathcal{B})\right] = \Theta(1) = P'(F|\mathcal{B}) = 0$$
$$\Theta\left[P'(F|\mathcal{B})\right] = \Theta(0) = P'(T|\mathcal{B}) = 1$$

bestimmt werden. Die Lösung lautet

$$\Theta(u) = \left(1 - u^k\right)^{1/k} \quad , \tag{8.32}$$

wobei k eine beliebige positive reelle Zahl ist. Für die Randbedingung $P'(F|\mathcal{B}) = \infty$ erhält man dieselbe Lösung aber mit negativem k. Man überzeugt sich leicht, dass dies tatsächlich die Lösung der Funktional-Gleichung ist. Aus der Relation für die Negation folgt dann

$$P'(\overline{A}|\mathcal{B}) = \Theta\left[P'(A|\mathcal{B})\right] = \left(1 - P'(A|\mathcal{B})^k\right)^{1/k}$$
$$P'(\overline{A}|\mathcal{B})^k = 1 - P'(A|\mathcal{B})^k \quad .$$

Wir definieren Wahrscheinlichkeit noch einmal um in

$$P'(\ldots)^k \to P''(\ldots) \quad .$$

Hierbei bleiben die Form der Produktregel sowie die Grenzwerte $P''(T|\mathcal{B}) = 1$ und $P''(F|\mathcal{B}) = 0$ erhalten. Die Summenregel vereinfacht sich in dieser Darstellung zu

$$P''(\overline{A}|\mathcal{B}) = 1 - P''(A|\mathcal{B}) \quad .$$

Die andere Darstellung liefert identische Ergebnisse, ist aber wesentlich unhandlicher. Ab sofort benennen wir P'' wieder in P um!
Abschließend wollen wir noch die allgemeine Form der Summenregel ableiten.

$$\begin{aligned}
P(A \vee B|\mathcal{B}) &= P(\overline{\overline{A}, \overline{B}}|\mathcal{B}) \\
&= 1 - P(\overline{A}, \overline{B}|\mathcal{B}) \\
&= 1 - P(\overline{A}|\overline{B}, \mathcal{B})\ P(\overline{B}|\mathcal{B}) \\
&= 1 - \left(1 - P(A|\overline{B}, \mathcal{B})\right)\ P(\overline{B}|\mathcal{B}) \\
&= 1 - \left(1 - \frac{P(A, \overline{B}|\mathcal{B})}{P(\overline{B}|\mathcal{B})}\right)\ P(\overline{B}|\mathcal{B}) \\
&= 1 - P(\overline{B}|\mathcal{B}) + P(A, \overline{B}|\mathcal{B}) \\
&= \underbrace{1 - P(\overline{B}|\mathcal{B})}_{P(B|\mathcal{B})} + P(\overline{B}|A, \mathcal{B})\ P(A|\mathcal{B}) \\
&= P(B|\mathcal{B}) + (1 - P(B|A, \mathcal{B}))\ P(A|\mathcal{B}) \\
&= P(B|\mathcal{B}) + P(A|\mathcal{B}) - P(B|A, \mathcal{B})\ P(A|\mathcal{B}) \\
&= P(B|\mathcal{B}) + P(A|\mathcal{B}) - P(A, B|\mathcal{B})
\end{aligned}$$

Somit haben wir bewiesen, dass zur konsistenten Zuweisung eines Maßes für den Wahrheitsgehalt einer Proposition die Regeln der Wahrscheinlichkeitsrechnung gelten müssen. **Jede andere quantitative Behandlung von Problemen der induktiven Logik ist entweder hierzu äquivalent oder falsch.**

8.5 Spezielle Propositionen

8.5.1 Indizierte Propositionen

In vielen Problemen ist es sinnvoll, Propositionen zu indizieren $A_i, \quad (i = 1, 2 \ldots, N)$, wobei die Anzahl der indizierten Propositionen endlich ($N < \infty$) oder unendlich ($N = \infty$) sein kann.
Beispiel: A_i: *Die Augenzahl beim Würfeln ist* i.

Von besonderer Bedeutung sind hierbei paarweise DISJUNKTE Propositionen

PAARWEISE DISJUNKTE PROPOSITIONEN

$$A_i \wedge A_j = \begin{cases} A_i & \text{für } i = j \\ F & \text{sonst.} \end{cases} \qquad \forall\, i, j \qquad . \tag{8.33}$$

Die A_i bilden eine PARTITIONIERUNG wenn sie disjunkt und VOLLSTÄNDIG sind

PARTITIONIERUNG

$$\vee_i\, A_i = T \qquad . \tag{8.34}$$

Die Produktregel vereinfacht sich für disjunkte Propositionen zu

$$P(A_i, A_j|\mathcal{B}) = \delta_{ij} P(A_i|\mathcal{B}) \qquad .$$

Die Summenregel lautet für disjunkte, vollständige Propositionen

$$1 = P(\vee_i\, A_i|\mathcal{B}) = \sum_i\, P(A_i|\mathcal{B}) \qquad . \tag{8.35}$$

SUMMENREGEL FÜR DISKRETE FREIHEITSGRADE

$$\sum_i P(A_i|\mathcal{B}) = 1 \qquad \text{Normierung} \tag{8.36a}$$

$$P(B|\mathcal{B}) = \sum_i P(B|A_i,\mathcal{B})\ P(A_i|\mathcal{B}) \qquad \text{Marginalisierungsregel} \tag{8.36b}$$

8.5.2 Kontinuierliche Propositionen

Zur Behandlung von Problemen mit kontinuierlichen Freiheitsgraden, führen wir Propositionen vom Typ

$A_{<a}$: *Die Variable A hat einen Wert kleiner als a*

ein. Die Wahrscheinlichkeit dieser Proposition $P(A_{<a}|\mathcal{B}) = P(A < a|\mathcal{B})$ ist eine VERTEILUNGSFUNKTION.

Daraus erhalten wir durch Differenzieren die Wahrscheinlichkeitsdichte

$$p(A_a|\mathcal{B}) = \lim_{da\to 0} \frac{P(A_{<a+da}|\mathcal{B}) - P(A_{<a}|\mathcal{B})}{da} \qquad .$$

Der Ausdruck im Zähler ist nichts anderes als die Wahrscheinlichkeit

$$P(dA_a|\mathcal{B})$$

für die „infinitesimale" Proposition

dA_a: *A liegt im Intervall* $[a, a+da)$.

Die Wahrscheinlichkeit $P(dA_a)$ ist demnach ein Produkt aus der infinitesimalen Intervallgröße und der WAHRSCHEINLICHKEITSDICHTE $p(A_a|\mathcal{B})$. Diese Form gilt allgemein für Probleme beliebiger Dimension $a \in \mathbb{R}^n$.

Wahrscheinlichkeitsdichten werden wir immer durch kleingeschriebene Buchstaben kennzeichnen. Die Summenregel geht in diesem Fall über in

SUMMENREGEL FÜR KONTINUIERLICHE FREIHEITSGRADE

$$\int p(x|\mathcal{B})\, dx = 1 \qquad \text{Normierung} \tag{8.37a}$$

$$P(B|\mathcal{B}) = \int P(B|x,\mathcal{B})\, p(x|\mathcal{B})\, dx \qquad \text{Marginalisierungsregel} \tag{8.37b}$$

Die Bezeichnung „Bayessche Wahrscheinlichkeitstheorie" rührt daher, dass das Bayessche Theorem

$$P(A|B,\mathcal{B}) = \frac{P(B|A,\mathcal{B})\; P(A|\mathcal{B})}{P(B|\mathcal{B})} \tag{8.38}$$

intensiv zur Lösung inverser Probleme verwendet wird. Hierbei enthält B die experimentellen Daten und A sind die Parameter, Hypothesen oder Modelle, die untersucht werden sollen.

8.6 Einfache Beispiele

8.6.1 Propagatoren

(R. D. Mattuck, *A Guide to Feynman Diagrams in the Many-Body Problem*, Dover Publications, Inc., New York, (92))

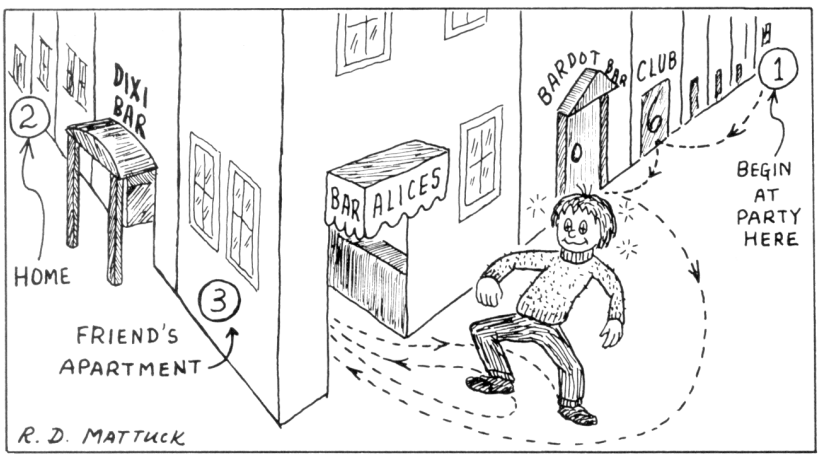

Fig. 1.1 *Propagation of Drunken Man*

(Reproduced with the kind permission of *The Encyclopedia of Physics*)

Ein Party-Gast (PG) propagiert nach der Party nach Hause. Dieses Modell wird im obenerwähnten Buch herangezogen um Greensche Funktionen in der Vielteilchen-Theorie zu erklären. Wir betrachten hier ein ganz einfaches (relativ unrealistisches) Modell. Bei jeder Bar wird erneut überlegt, ob er einkehren soll (Wahrscheinlichkeit P_B) oder nicht ($1-P_B$). Wenn er einkehrt, besteht ein Wahrscheinlichkeit $P_R < 1$, dass er wieder herauskommt und weiter propagiert.
Die disjunkten Ereignisse (Propositionen) sind hier
E_n: *Er kehrt in n Bars ein.*
Die zu untersuchende Proposition lautet:
H: *Er kehrt am Abend heim.*
Mit der Marginalisierungsregel erhalten wir

$$P(H|N,\mathcal{B}) = \sum_{n=0}^{N} P(H|E_n,N,\mathcal{B})\ P(E_n|N,\mathcal{B}) \qquad .$$

Die Proposition N besagt, dass es N Bars gibt, und der Bedingungskomplex enthält u.a. die Information, dass jede Bar höchstens einmal aufgesucht wird. Nun sollen die Entscheidungen, in die individuellen Bars einzukehren, unkorreliert (unabhängig) sein. Es handelt sich also um einen Bernoulli-Versuch

$$P(E_n|N,\mathcal{B}) = \binom{N}{n}\ {P_B}^n\ (1-P_B)^{N-n} \qquad .$$

Die Wahrscheinlichkeit, dass nach einem Bar-Besuch die Heimreise fortgesetzt werden kann, sei P_R. Damit er noch in dieser Nacht zu Hause ankommt, muss er in allen n Fällen die Heimreise antreten. Das geschieht mit der Wahrscheinlichkeit ${P_R}^n$. Somit haben wir

$$\begin{aligned} P(H|N,\mathcal{B}) &= \sum_{n=0}^{N}\ P(H|E_n,N,\mathcal{B})\ P(E_n|N,\mathcal{B}) \\ &= \sum_{n=0}^{N}\ \binom{N}{n}\ {P_B}^n\ (1-P_B)^{N-n}\ {P_R}^n \\ &= \sum_{n=0}^{N}\ \binom{N}{n}\ (P_B P_R)^n\ (1-P_B)^{N-n} \\ &= (P_B P_R + 1 - P_B)^N = (1-P_B(1-P_R))^N \qquad . \end{aligned}$$

Das Ergebnis macht Sinn, denn $P_B\ (1-P_R)$ ist die Wahrscheinlichkeit, dass er von einer Bar absorbiert wird. Demnach ist $1-P_B\ (1-P_R)$ die Wahrscheinlichkeit, dass er eine Bar „überwindet". Damit er nach Hause kommt, muss dieses Ereignis N-mal eintreten.
Wenn $P_B(1-P_R) \ll 1$, dann gilt

$$P(H|N,\mathcal{B}) = e^{N\ \ln(1-P_B(1-P_R))} \simeq e^{-P_B(1-P_R)\ \ N} \triangleq e^{-\alpha x} \qquad .$$

Wir erkennen sofort den Bezug zu physikalischen Problemen, z.B. Dämpfung von Wellen, Teilchen in Materie. Die Anzahl N der Bars entspricht dem zurückgelegten Weg x, und die Wahrscheinlichkeit $P_B(1-P_R)$, von einer Bar „absorbiert" zu werden, entspricht der Dämpfungkonstante α pro Wegeinheit.

8.6.2 Das 3 Türen Problem

Aufgabenstellung eines Glücksspiels:
Es gibt drei Türen. Hinter einer Tür befindet sich der Hauptgewinn. Der Kandidat wählt eine Tür aus. Der Quizmaster öffnet eine der beiden anderen Türen, allerdings niemals die Tür mit dem Hauptgewinn. Danach kann der Kandidat entscheiden, ob er bei seiner Wahl bleibt, oder die noch verschlossene Tür wählt.
Ist es besser, bei der ersten Wahl zu bleiben, zu wechseln, oder spielt es keine Rolle?
Wir wollen die Wahrscheinlichkeit berechnen, dass wir mit der Strategie S gewinnen.
Die benötigten Propositionen sind

- S: *Die verfolgte Strategie ist, bei der ersten Wahl zu bleiben.*
- G: *Wir gewinnen.*

Damit ist die gesuchte Wahrscheinlichkeit $P(G|S,\mathcal{B})$. Wann sind alle Wahrscheinlichkeiten bekannt? Genau dann, wenn wir wissen, ob unsere erste Wahl richtig war! Also verwenden wir die Propositionen
$\sigma = 1/0$: *Die erste Wahl war richtig/falsch*
Die Marginalisierungsregel liefert das gesuchte Ergebnis

$$P(G|S,\mathcal{B})) = \sum_{\sigma=0}^{1} \underbrace{P(G|\sigma,S,\mathcal{B})}_{\delta_{\sigma,1}} \; P(\sigma|S,\mathcal{B}) = P(\sigma=1|S,\mathcal{B}) = 1/3.$$

Somit ist es offensichtlich wesentlich besser, die Tür zu wechseln. Wie kann man das anschaulich verstehen? Es ist gut, bei der ersten Entscheidung zu bleiben, wenn die erste Wahl richtig war. Hierfür gibt es zu Beginn nur eine Möglichkeit. Es ist hingegen gut zu wechseln, wenn die erste Wahl falsch war. Hierfür gibt es zu Beginn zwei Möglichkeiten.

8.6.3 Detektor für seltene Teilchen

Wie gut muss ein Detektor sein, um seltene Teilchen nachzuweisen? Benötigte Propositionen

- T / $\overline{T}$: *Teilchen vorhanden / nicht vorhanden.*
- D: *Detektor spricht an.*

Wir wollen die Wahrscheinlichkeit dafür berechnen, dass ein Teilchen vorhanden ist, wenn der Detektor anspricht? Wir verwenden hierzu das Bayessche Theorem

$$P(T|D,\mathcal{B}) = \frac{P(D|T,\mathcal{B})\ P(T|\mathcal{B})}{P(D|T,\mathcal{B})\ P(T|\mathcal{B}) + P(D|\overline{T},\mathcal{B})\ P(\overline{T}|\mathcal{B})}$$

Zahlenbeispiel:

$$\begin{aligned} P(T|\mathcal{B}) &= 10^{-8} \\ P(D|T,\mathcal{B}) &= 0.9999 \\ P(D|\overline{T},\mathcal{B}) &= 0.0001 \qquad . \end{aligned}$$

Einsetzen der Zahlenwerte liefert

$$P(T|D,\mathcal{B}) = 10^{-4} \qquad .$$

Obwohl der Detektor scheinbar sehr gut ist, ist die Wahrscheinlichkeit verschwindend klein, dass tatsächlich ein Teilchen vorhanden ist, wenn der Detektor reagiert. Wie gut muss dann der Detektor überhaupt sein, damit $P(T|D,\mathcal{B})$ wenigstens $1/2$ ist? Offensichtlich muss gelten

$$\begin{aligned} P(D|\overline{T},\mathcal{B})\ P(\overline{T}|\mathcal{B}) &\leq P(D|T,\mathcal{B})\ P(T|\mathcal{B}) \\ P(D|\overline{T},\mathcal{B}) &\leq P(D|T,\mathcal{B})\ \frac{P(T|\mathcal{B})}{1-P(T|\mathcal{B})} \simeq P(T|\mathcal{B}) \qquad . \end{aligned}$$

Das heißt, die Fehlerrate $P(D|\overline{T},\mathcal{B})$ muss kleiner als 10^{-8} sein.

Medizinische Untersuchungen

Diese Überlegungen lassen sich sofort auf medizinische Reihen-Untersuchungen übertragen. Die Propositionen sind dann

- T: *Der Patient hat einen Virus*.
- D: *Der Befund ist positiv*.

Gesucht wird die Wahrscheinlichkeit, dass der Patient tatsächlich einen Virus hat, wenn der Befund positiv ausfällt. Typische Werte sind

$$\begin{aligned} P(T|\mathcal{B}) &= 0.001 \\ P(D|T,\mathcal{B}) &= 0.9 \\ P(D|\overline{T},\mathcal{B}) &= 0.01 \qquad . \end{aligned}$$

Das liefert

$$\begin{aligned} P(D|T,\mathcal{B})\ P(T|\mathcal{B}) &= 0.9 \cdot 0.001 = 0.0009 \\ P(D|\overline{T},\mathcal{B})\ P(\overline{T}|\mathcal{B}) &= 0.01 \cdot (1 - 0.001) = 0.00999 \\ P(T|D,\mathcal{B}) &= \frac{1}{1 + \frac{0.00999}{0.0009}} = \frac{1}{12.1} \simeq 0.08 \qquad . \end{aligned}$$

Das heißt die Wahrscheinlichkeit ist lediglich 0.08, dass der Patient den Virus hat. Allerdings ist dadurch die Wahrscheinlichkeit von $P(T|\mathcal{B}) = 0.001$ vor der Untersuchung um einen Faktor 80 angestiegen.
Der anschauliche Grund in beiden Fällen ist der, dass es für das Reagieren des Detektors/Tests sehr viel mehr Möglichkeiten $P(D|\overline{T},\mathcal{B})\ P(\overline{T}|\mathcal{B})$ der Fehldiagnose als der korrekten Diagnose $P(D|T,\mathcal{B})\ P(T|\mathcal{B})$ gibt, da es sich um seltene Teilchen/Fälle $P(\overline{T}|\mathcal{B}) \gg P(T|\mathcal{B})$ handelt.

8.6.4 Ist die Münze symmetrisch?

Wenn man Hypothesen testen will, betrachtet man üblicherweise das sogenannte Odds-Ratio

$$o = \frac{P(H|D,\mathcal{B})}{P(\overline{H}|D,\mathcal{B})} \qquad ,$$

wobei H die zu analysierende Hypothese und D die Daten sind. Der Vorteil des Odds-Ratios ist, dass man vermeidet, den Normierungsnenner berechnen zu müssen.
Hier interessieren wir uns für

$$o = \frac{P(H|n_K,N,\mathcal{B})}{P(\overline{H}|n_K,N,\mathcal{B})} = \frac{P(n_K|H,N,\mathcal{B})\ P(H|\mathcal{B})}{P(n_K|\overline{H},N,\mathcal{B})\ P(\overline{H}|\mathcal{B})} = \underbrace{\frac{P(n_K|H,N,\mathcal{B})}{P(n_K|\overline{H},N,\mathcal{B})}}_{Bayes-Faktor}\ \underbrace{\frac{P(H|\mathcal{B})}{P(\overline{H}|\mathcal{B})}}_{Prior-Odds} .$$

Die verwendeten Propositionen lauten

- H: *Die Münze ist symmetrisch*
- n_K: *Die Anzahl „Kopf" ist n_K.*
- N: *Die Münze wird N-mal geworfen.*

Wir gehen davon aus, dass wir vorab nicht wissen, ob die Münze symmetrisch ist oder nicht

$$\frac{P(H|\mathcal{B})}{P(\overline{H}|\mathcal{B})} = 1 \qquad .$$

Die MARGINALE LIKELIHOOD-FUNKTION $P(n_K|N,H,\mathcal{B})$ und $P(n_K|N,\overline{H},\mathcal{B})$ kann unter Ausnutzung der Marginalisierungsregel aus der Likelihood-Funktion berechnet werden

$$P(n_K|N,A,\mathcal{B}) = \int_0^1 P(n_K|N,q,A,\mathcal{B})\; p(q|N,A,\mathcal{B})\; dq \qquad , \tag{8.39}$$

mit $A = H$ oder $A = \overline{H}$. Die Likelihood-Funktion ist unabhängig von A, denn sie gibt die Wahrscheinlichkeit dafür, dass bei N-maligem Münzwurf n_k-mal Kopf erscheint, wenn die Wahrscheinlichkeit für Kopf im Einzelwurf q ist. Durch die Bedingung, dass q bekannt ist, wird die Hypothese H oder deren Negation $\overline{H}$ in der Likelihood-Funktion überflüssig. Es handelt sich wieder um die Binomial-Verteilung

$$P(n_K|N,q,\mathcal{B}) = \binom{N}{n_K}\; q^{n_K}\; (1-q)^{N-n_K} \qquad .$$

Die Abhängigkeit von A steckt in $P(q|N,A,\mathcal{B})$. Diese Wahrscheinlichkeit ist logisch unabhängig von der Zahl der Würfe N

$$P(q|A,\mathcal{B}) = \begin{cases} \delta(q-1/2) & \text{für} A = H \\ \theta(0 \le q \le 1) & \text{für} A = \overline{H} \end{cases} \qquad .$$

Im Fall, dass die Hypothese zutrifft, ist $q = 1/2$. Strenggenommen ist in diesem Fall $P(n_K|N,q,H,\mathcal{B})$ nur für $q = 1/2$ erlaubt, da für andere Werte von q hinter dem Bedingungsstrich ein Widerspruch steht. Die Wahrscheinlichkeit $P(n_K|N,q,H,\mathcal{B})$ ist für $q \neq 1/2$ nicht definiert, aber diese Terme tragen wegen der δ-Funktion ohnehin nicht zum Integral Gl. (8.39) bei.
Die Hypothese $\overline{H}$ bedeutet, dass wir nicht wissen welchen Wert q annimmt. Er kann somit jeden Wert zwischen 0 und 1 gleich-wahrscheinlich annehmen.
Damit haben wir

$$P(n_K|N,H,\mathcal{B}) = \binom{N}{n_K}\; (\frac{1}{2})^{n_K}\; (1-\frac{1}{2})^{N-n_K} = \binom{N}{n_K}\; 2^{-N}$$

$$P(n_K|N,\overline{H},\mathcal{B}) = \binom{N}{n_K}\; \int_0^1 q^{n_K}\; (1-q)^{N-n_K}\; dq = \binom{N}{n_K}\; \frac{n_K!\;(N-n_K)!}{(N+1)!} \qquad .$$

Das Odds-Ratio ist somit

$$o = \frac{2^{-N}\;(N+1)!}{n_K!\;(N-n_K)!}$$

Wir betrachten die beiden Extremfälle

- Die Münze ist manipuliert und fällt immer auf Kopf.
- Die Münze verhält sich perfekt symmetrisch $q = 1/2$.

Im ersten Fall ist $n_K = N$. Damit vereinfacht sich der Bayes-Faktor zu

$$o = 2^{-N}\ (N+1) \qquad .$$

Aus dem Odds-Ratio erhalten wir die Wahrscheinlichkeit

$$P(H|n_K = N, N, \mathcal{B}) = \frac{o}{1+o} = \frac{1}{1+1/o} = \frac{1}{1+2^N/(N+1)} \qquad .$$

Die Ergebnisse hierzu sind in der Tabelle 8.1 aufgelistet.

N	$P(H\|n_K = N, N, \mathcal{B})$
0	0.500
1	0.500
2	0.429
5	0.158
10	0.011
15	0.0005
20	0.00002

Tabelle 8.1:

Für $N = 0$ ist alles noch unentschieden $P(H|n_K, N, \mathcal{B}) = 1/2$. Nach dem ersten Wurf ist die Wahrscheinlichkeit immer noch $P(H|n_K, N, \mathcal{B}) = 1/2$. Weitere Werte sind in der Tabelle aufgelistet. Man erkennt, dass die Wahrscheinlichkeit sehr schnell klein wird und man bereits bei zehn Würfen sicher sein kann, dass die Münze manipuliert ist.
Im Fall einer symmetrische Münze erhalten wir im Idealfall $N = 2n_K$. Damit wird aus dem Odds-Ratio

$$o = 2^{-2\,n_K}\,\frac{(2\,n_K+1)!}{n_K!\;n_K!} \qquad .$$

Für diesen Datensatz sind die Ergebnisse in Tabelle 8.2 aufgeführt.

N	$P(H\|n_K = N/2, N, \mathcal{B})$
0	0.500
2	0.600
10	0.730
20	0.787
40	0.837
100	0.889
200	0.919
2000	0.973

Tabelle 8.2:

Es ist natürlich wesentlich schwerer in diesem Fall zu entscheiden, ob die Münze symmetrisch ist oder nicht.

8.6.5 Produktionsrate des Wettbewerbers

Eine Firma hat von einem Produkt N Stück hergestellt. Die einzelnen Teile haben fortlaufende Seriennummern. Es wird eine Stichprobe vom Umfang L ohne Zurücklegen genommen, in der sich die laufenden Nummern $\{n_1, n_2, \ldots, n_L\}$ befinden. Was können wir daraus über die Gesamtproduktion aussagen?
Die Wahrscheinlichkeit $P(N|n_1, n_2, \ldots, n_L, \mathcal{B})$ können wir wieder über das Bayessche Theorem bestimmen

$$P(N|n_1, n_2, \ldots, n_L, L, \mathcal{B}) = \frac{1}{Z}\, P(n_1, n_2, \ldots, n_L|N, L, \mathcal{B})\; P(N|L, \mathcal{B}) \qquad .$$

Hierbei ist die Bedeutung der Propositionen

- N: *Die Stückzahl beträgt N.*
- L: *Der Umfang der Stichprobe ist L.*
- n_j: *Das j-te Element der Stichprobe hat die Seriennummer n_j.*

Die Wahrscheinlichkeit $P(n_1, n_2, \ldots, n_L|N, L, \mathcal{B})$ kann mit der Produktregel weiter vereinfacht werden

$$\begin{aligned} P(n_1, \ldots, n_L|N, L, \mathcal{B}) &= P(n_L|n_1, \ldots, n_{L-1}, N, L, \mathcal{B}) \\ &\quad \cdot P(n_1, \ldots, n_{L-1}|N, L, \mathcal{B}) \\ &= P(n_L|n_1, \ldots, n_{L-1}, N, L, \mathcal{B}) \\ &\quad \cdot P(n_{L-1}|n_1, \ldots, n_{L-2}, N, L, \mathcal{B}) \\ &\quad \cdots P(n_2|n_1, N, L, \mathcal{B})\; P(n_1|N, L, \mathcal{B}) \qquad . \end{aligned}$$

Die Stichprobe soll „zufällig" entnommen sein. Das bedeutet, dass $\{1, 2, \ldots, N\}$ mit gleicher Wahrscheinlichkeit in der Stichprobe vorkommen kann

$$P(n_j|n_1, n_2, \ldots, n_{j-1}, N, L, \mathcal{B}) = \frac{1}{N - j + 1}\, \theta(n_j \leq N) \qquad ,$$

da es für die Seriennummer n_j noch $N - j + 1$ mögliche Werte gibt, da durch die Seriennummern $n_1, n_2, \ldots, n_{j-1}$ bereits $j - 1$ Nummern von N vergeben sind. Die $N - j + 1$ Möglichkeiten haben alle dieselbe Wahrscheinlichkeit.
Damit ist die Likelihood-Funktion

$$\begin{aligned} P(n_1, n_2, \ldots, n_L|N, L, \mathcal{B}) &= \prod_{j=1}^{L} \left(\frac{1}{N + 1 - j}\, \theta(n_j \leq N) \right) \\ &= \theta(N \geq n_{\max})\, \frac{(N - L)!}{N!} \end{aligned}$$

Hierbei ist $n_{\max} = \max_j\; n_j$ die größte Seriennummer in der Stichprobe. Die Prior-Wahrscheinlichkeit $P(N|L, \mathcal{B})$ enthält nur die Bedingung, dass die Stichprobe den

Umfang L hat. Das bedeutet, dass $N \geq L$ sein muss. Wir werden einen Maximalwert $N_{\max}$ einführen, den wir am Ende der Rechnung nach Unendlich gehen lassen

$$P(N|L,\mathcal{B}) = \frac{1}{N_{\max} - L}\, \theta(L \leq N < N_{\max}) \qquad .$$

Damit ist die gesuchte Wahrscheinlichkeit

$$P(N|n_1, n_2, \ldots, n_L, L, \mathcal{B}) = \frac{1}{Z'}\, \theta(n_{\max} \leq N < N_{\max})\, \frac{(N-L)!}{N!}$$
$$Z' = \sum_{N=n_{\max}}^{N_{\max}} \frac{(N-L)!}{N!} \qquad .$$

Um ohne Computer weiterzukommen, gehen wir davon aus, dass $n_{\max} \gg L$. Dann sind die Werte für N ebenfalls sehr viel größer als L und wir können

$$\frac{(N-L)!}{N!} \simeq N^{-L}$$

verwenden. Die Normierungskonstante Z' lässt sich dann näherungsweise berechnen, indem wir die Summe durch ein Integral ersetzen

$$Z' \simeq \int_{N=n_{\max}}^{N_{\max}} N^{-L}\, dN = \frac{N^{-L+1}}{L-1}\Big|_{N_{\max}}^{n_{\max}} \underset{N_{\max}\to\infty}{\longrightarrow} \frac{n_{\max}^{-L+1}}{L-1} \qquad .$$

Das führt zu

$$P(N|n_1, n_2, \ldots, n_L, L, \mathcal{B}) = \frac{\theta(n_{\max} \leq N < \infty)\, N^{-L}}{n_{\max}^{-L+1}/(L-1)} \qquad .$$

Die Wahrscheinlichkeit für N im Lichte der Stichproben-Information hat ihr Maximum beim kleinsten erlaubten Wert $N = n_{\max}$ und fällt von da wie N^{-L} ab. Der Mittelwert liegt bei

$$\begin{aligned}
\langle N \rangle &= \frac{1}{Z'} \sum_{N=n_{\max}}^{N_{\max}} N\, N^{-L} \\
&\simeq \frac{1}{Z'} \int_{N=n_{\max}}^{N_{\max}} N^{-L+1}\, dN \\
&\underset{N_{\max}\to\infty}{\longrightarrow} \frac{L-1}{n_{\max}^{-L+1}}\, \frac{n_{\max}^{-L+2}}{L-2} \\
&= n_{\max}\, \frac{L-1}{L-2} = n_{\max}\, \left(1 + \frac{1}{L-2}\right) \qquad .
\end{aligned}$$

Das Ergebnis macht erst Sinn für $L > 2$. Für kleinere Stichproben hängt die Posterior-Wahrscheinlichkeit vom Cutoff $N_{\max}$ ab, und für $N_{\max} \to \infty$ existiert weder die Norm noch das erste Moment.
Wir wollen die Verteilungsfunktion untersuchen

$$\begin{aligned} F(N) &:= \sum_{N'=n_{\max}}^{N} P(N'|n_1, n_2, \ldots, n_L, L, \mathcal{B}) \\ &\simeq \theta(N \geq n_{\max}) \, \frac{L-1}{n_{\max}^{-L+1}} \int_{n_{\max}}^{N} N'^{-L} \, dN' \\ &= \theta(N \geq n_{\max}) \, \frac{L-1}{n_{\max}^{-L+1}} \, \frac{n_{\max}^{-L+1} - N^{-L+1}}{L-1} \\ F(N) &= \theta(N \geq n_{\max}) \left(1 - \left(\frac{n_{\max}}{N}\right)^{L-1}\right) \end{aligned}$$

Damit z.B. 90% der Wahrscheinlichkeitsmasse abgedeckt ist, muss $\left(\frac{n_{\max}}{N}\right)^{L-1} = 0.1$ sein. Daraus folgt

$$N = n_{\max}(0.1)^{-1/(L-1)} = n_{\max} \cdot (10)^{1/(L-1)} \qquad .$$

Für eine Stichprobe vom Umfang $L = 10$ bedeutet das, $N = 1.29 \; n_{\max}$. Mit einer Wahrscheinlichkeit von 0.9 liegt N im Intervall $N \in [n_{\max}, 1.29 \; n_{\max}]$.

8.6.6 Anzahl der Fische

Wir hatten in einem früheren Kapitel folgendes Problem diskutiert. In einem See befinden sich N Fische. Davon werden N_r Fische gefangen, rot gefärbt und wieder in den See zurückgegeben. Nach einiger Zeit werden wieder $n = N_r$ Fische gefangen. Davon sind n_r rot. Wir hatten damals die Wahrscheinlichkeit berechnet, dass sich in der Stichprobe n_r rote Fische befinden, wenn das Verhältnis der roten Fische im See $q = N_r/N$ bekannt ist.
Wir wollen nun der wesentlich interessanteren Frage nachgehen, wie wir aus der Stichprobe auf die Zahl N der Fische im See schließen können. Die gesuchte Wahrscheinlichkeit ist also $P(N|n_r, n = N_r, N_r, \mathcal{B})$ zu den Propositionen

- N: *Die Zahl der Fische im See ist* N.
- N_r: *Die Zahl der roten Fische im See ist* N_r.
- n: *Die Stichprobe hat den Umfang* $n = N_r$.
- n_r: *In der Stichprobe befinden sich* n_r *rote Fische.*

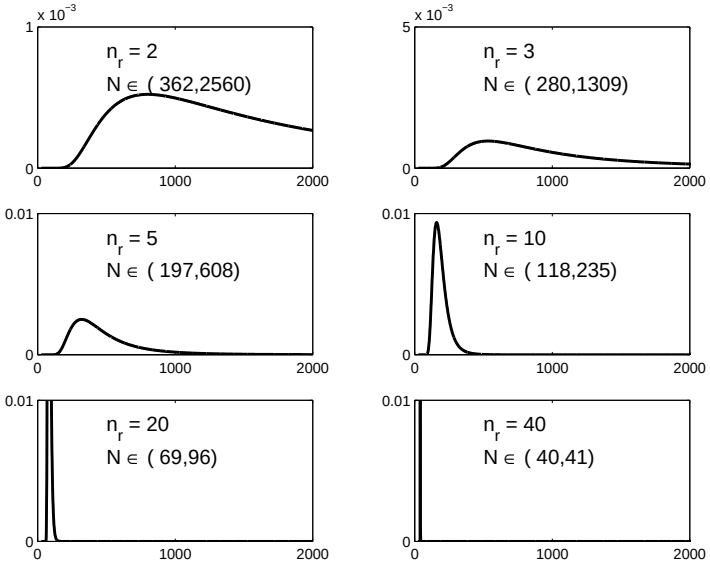

Abbildung 8.1: *Wahrscheinlichkeit $P(N|n_r, n = N_r, N_r, \mathcal{B})$ als Funktion von N für den Fall $N_r = 40$ und $n = N_r = 40$. Zusätzlich angegeben ist das Intervall, in dem die Wahrscheinlichkeit größer ist als ein e-tel des Maximalwertes.*

Nach dem Bayesschen Theorem gilt

$$P(N|n_r, n = N_r, N_r, \mathcal{B}) = \frac{1}{Z} P(n_r|N, n = N_r, N_r, \mathcal{B})\ P(N|n = N_r, N_r, \mathcal{B}) \qquad .$$

Der Normierungsnenner Z hängt nicht von N ab und wird nachträglich bestimmt. Der erste Faktor $P(n_r|N, n = N_r, N_r, \mathcal{B})$ ist die bereits bestimmte Vorwärts-Wahrscheinlichkeit, in einer Stichprobe vom Umfang $n = N_r$, n_r rote Fische zu haben, wenn $q = N_r/N$ bekannt ist. Das ist die hypergeometrische Verteilung

$$P(n_r|N, n = N_r, N_r, \mathcal{B}) = \left.\frac{\binom{N_r}{n_r}\ \binom{N-N_r}{n-n_r}}{\binom{N}{n}}\right|_{n=N_r} \qquad .$$

Die Prior-Wahrscheinlichkeit $P(N|n = N_r, N_r, \mathcal{B})$ hängt nun davon ab, was wir über die Zahl der Fische im See wissen. Sie ist zumindest größer als N_r. Wenn wir sonst nichts wissen, setzen wir

$$P(N|n, N_r, \mathcal{B}) = \frac{1}{N_{\max} - N_r}\ \theta(N_r \leq N < N_{\max})$$

an. $N_{\max}$ ist die Maximalzahl, die wir entweder abschätzen können oder am Ende der Analyse nach Unendlich gehen lassen. Später werden wir bessere Methoden, uninformative Wahrscheinlichkeitsverteilungen zu bestimmen, kennenlernen. Damit ist die gesuchte Wahrscheinlichkeit

$$P(N|n, n_r, N_r, \mathcal{B}) = \frac{1}{Z'}\ \theta(0 \leq N < N_{\max})\ \frac{\binom{N-N_r}{N_r-n_r}}{\binom{N}{N_r}} \qquad .$$

Das erste Argument der θ-Funktion ist hier auf Null gesetzt worden, da der Likelihood-Anteil ohnehin verschwindet, wenn $N < 2N_r - n_r$. Alle N-unabhängigen Anteile sind in Z' zusammengefasst. Wenn in der Stichprobe kein roter Fisch enthalten ist ($n_r = 0$), dann geht die Wahrscheinlichkeit für $N \to \infty$ gegen 1. Das heißt, dieses Ergebnis wird vom Cutoff $N_{\max}$ bestimmt und ist somit nicht sehr aussagekräftig. Das ist natürlich verständlich, denn $n_r = 0$ ist kompatibel mit $N = \infty$. Für $n_r = 1$ geht die Wahrscheinlichkeit für $N \to \infty$ wie $1/N$. Hierbei spielt der Cutoff schon keine so große Rolle mehr. Er ist zwar für die Normierung wichtig, aber wir können bereits das Maximum der Verteilung bestimmen und die Punkte, an denen die Verteilung auf ein e-tel abgefallen ist. Für $n_r > 1$ ist der Cutoff irrelevant und kann auf unendlich gesetzt werden.
In Abbildung 8.1 sind die Wahrscheinlichkeiten zu $n = N_r = 40$ für verschiedene Werte von n_r aufgetragen. Man erkennt, dass die experimentellen Daten mit zunehmendem n_r immer aussagekräftiger werden. Man kann diese Analyse auch verwenden, um das Experiment zu planen. Das heißt, den Umfang der Stichprobe so zu wählen, dass belastbare Ergebnisse herauskommen, sprich, dass die Varianz möglichst klein ist.

8.6.7 Beste Auswahl aus N Vorschlägen

Wir betrachten nun folgendes Problem. Es wird nacheinander N Vorschläge geben. Maximal einen davon können wir akzeptieren. Bei jedem Ereignis müssen wir uns entscheiden, ob wir akzeptieren oder verwerfen. Diese Entscheidung ist endgültig. Die Ereignisse sind alle durch eine Zufallszahl x_i $(i = 1, \ldots, N)$ derselben Verteilung $p(x)$ charakterisiert. Welches ist die beste Strategie, um einen möglichst hohen Wert zu akzeptieren? Sollen wir den erstbesten Vorschlag akzeptieren?
Eine mögliche Strategie ist, von den N Vorschlägen die ersten L als „Lernphase" aufzufassen. Die ersten L Vorschläge sollen den Maximalwert ξ haben. Von den nächsten $K = N - L - 1$ Vorschlägen akzeptieren wir den ersten, der größer-gleich ξ ist. Wenn davon keiner das Kriterium erfüllt, nehmen wir den letzten Vorschlag x_N.
Wir wollen uns zunächst überlegen, wie die Wahrscheinlichkeitsdichte

$$p(R|N, L, \mathcal{B})$$

der so erzielten Resultate R aussieht

$$R : \textit{Der Wert des akzeptierten Vorschlags ist } R.$$

Welche Propositionen müssen wir einführen, um die Wahrscheinlichkeit eindeutig angeben zu können? Wir benötigen den Maximalwert ξ der ersten L Vorschläge, die Werte der Zufallszahlen $x_{L+1}, \ldots, x_{L+K}$ und den der letzten Zufallszahl x_N. Um die Indizes übersichtlich zu halten, nennen wir

$$\begin{aligned} y_i &= x_{L+i}, \qquad i = 1, 2, \ldots, K \\ z &= x_N \qquad . \end{aligned}$$

Die Marginalisierungsregel liefert demnach

$$p(R|N,L,\mathcal{B}) = \int d\xi \int dy^K \int dz\; p(R|N,L,\xi,y_1,\ldots,y_K,z,\mathcal{B}) \\ \times\; p(\xi|N,L,\mathcal{B})\; p(y_1,y_2,\ldots,y_K|N,L,\mathcal{B})\; p(z|N,L,\mathcal{B}) \quad . \tag{8.40}$$

Es wurde bereits ausgenutzt, dass die Zufallszahlen $x_1, x_2, \ldots, x_N$ unkorreliert sind. Es gilt weiter

$$p(y_1,y_2,\ldots,y_K|N,L,\mathcal{B}) = p(y_1)\;\ldots\;p(y_K) \\ p(z|N,L,\mathcal{B}) = p(z) \quad .$$

Für die Dichte $p(\xi|N,L,\mathcal{B})$ ist N irrelevant, d.h $p(\xi|N,L,\mathcal{B}) = p(\xi|L,\mathcal{B})$. Nun müssen wir noch $p(R|N,L,\xi,y_1,\ldots,y_K,z,\mathcal{B})$ ermitteln. Der Wert R ist gleich y_1, wenn $y_1 \geq \xi$, sonst ist er gleich y_2, wenn $y_2 \geq \xi$ etc. Wenn kein $y_i \geq \xi$ ist $R = z$. Auf eine mathematische Formel gebracht, heißt das

$$p(R|N,L,\xi,y_1,\ldots,y_K,z,\mathcal{B}) = \sum_{j=1}^{K} \left(\delta(R-y_j)\;\theta(y_j \geq \xi) \prod_{i=1}^{j-1} \theta(y_i < \xi) \right) \\ + \delta(R-z) \prod_{j=1}^{K} \theta(y_j < \xi) \quad .$$

Einsetzen in Gl. (8.40) liefert

$$p(R|N,L,\mathcal{B}) = \sum_{j=1}^{K} \left[\int d\xi\; p(\xi|L,\mathcal{B}) \left(\int dy_j\; \delta(R-y_j)\;\theta(y_j \geq \xi)\; p(y_j) \right) \right. \\ \left. \times \prod_{i=1}^{j-1} \left(\int dy_i\; p(y_i)\;\theta(y_i < \xi) \right) \right] \left(\int dz\; p(z) \right) \\ + \int d\xi\; p(\xi|L,\mathcal{B}) \left(\int dz\; \delta(R-z)\; p(z) \right) \prod_{j=1}^{K} \left(\int dy_j\; p(y_j)\;\theta(y_j < \xi) \right) \quad . \tag{8.41}$$

Die Integrale sind leicht auszuwerten

$$p(R|N,L,\mathcal{B}) = \sum_{j=1}^{K} \int d\xi\; p(\xi|L,\mathcal{B})\;\theta(R \geq \xi)\; p(R) \prod_{i=1}^{j-1} \Big(F(\xi) \Big) \\ + \int d\xi\; p(\xi|L,\mathcal{B})\; p(R) \prod_{j=1}^{K} \Big(F(\xi) \Big) \\ = p(R) \sum_{j=1}^{K} \int_{-\infty}^{R} d\xi\; p(\xi|L,\mathcal{B})\; F(\xi)^{j-1} \\ + p(R) \int d\xi\; p(\xi|L,\mathcal{B})\; F(\xi)^K \quad .$$

Die Wahrscheinlichkeitsdichte des Maximalwertes von L Zufallszahlen werden wir im Abschnitt 9.3 Gl. (9.5) kennen lernen. Damit vereinfacht sich das Ergebnis zu

$$\begin{aligned} p(R|N,L,\mathcal{B}) &= L\, p(R) \sum_{j=0}^{K-1} \int_{-\infty}^{R} d\xi\, p(\xi)\, F(\xi)^{L-1}\, F(\xi)^{j} \\ &\quad + L\, p(R) \int d\xi\, p(\xi)\, F(\xi)^{L-1}\, F(\xi)^{K} \\ &= L\, p(R) \left(\sum_{j=0}^{K-1} \int_{-\infty}^{R} d\xi\, p(\xi)\, F(\xi)^{L+j-1} \right. \\ &\quad \left. + \int_{-\infty}^{\infty} d\xi\, p(\xi)\, F(\xi)^{L+K-1} \right) \quad . \end{aligned}$$

Wegen $L + K = N - 1$ ist $K = N - L - 1$, und somit gilt

$$\begin{aligned} p(R|N,L,\mathcal{B}) &= L\, p(R) \left(\sum_{j=0}^{N-L-2} \int_{-\infty}^{R} d\xi\, p(\xi)\, F(\xi)^{L+j-1} \right. \\ &\quad \left. + \int_{-\infty}^{\infty} d\xi\, p(\xi)\, F(\xi)^{N-2} \right) \quad . \end{aligned} \tag{8.42}$$

Wir betrachten nun den Spezialfall gleich-verteilter Zufallszahlen auf dem Einheitsintervall

$$p(x) = \theta(0 \le x < 1) \quad .$$

Die kumulative Wahrscheinlichkeit ist

$$F(x) = \theta(0 \le x < 1) \int_{0}^{x} dx' = x\, \theta(0 \le x < 1) \quad .$$

Damit wird aus Gl. (8.42)

$$\begin{aligned} p(R|N,L,\mathcal{B}) &= L\, \theta(0 \le R < 1) \left(\sum_{j=0}^{N-L-2} \int_{0}^{R} d\xi\, \xi^{L+j-1} + \int_{0}^{1} d\xi\, \xi^{N-2} \right) \\ &= L\, \theta(0 \le R < 1) \left(\sum_{j=0}^{N-L-2} \frac{R^{L+j}}{L+j} + \frac{1}{N-1} \right) \quad . \end{aligned}$$

Nach dieser längeren Rechnung ist es ratsam zu testen, ob die Normierung noch

stimmt.

$$\begin{aligned}\int_0^1 p(R|N,L,\mathcal{B})\,dR &= L\left(\sum_{j=0}^{N-L-2}\int_0^1 \frac{R^{L+j}}{L+j}\,dR + \frac{1}{N-1}\int_0^1 dR\right)\\ &= L\left(\sum_{j=0}^{N-L-2}\frac{1}{(L+j)(L+j+1)} + \frac{1}{N-1}\right)\\ &= L\left(\sum_{j=0}^{N-L-2}\left(\frac{1}{L+j} - \frac{1}{L+j+1}\right) + \frac{1}{N-1}\right)\\ &= L\left(\frac{1}{L} - \frac{1}{L+N-L-2+1} + \frac{1}{N-1}\right) = 1 \qquad .\end{aligned}$$

Das ist schon einmal beruhigend. Als nächstes bestimmen wir den Erwartungswert

$$\begin{aligned}\langle R\rangle &= \int_0^1 R\,p(R|N,L,\mathcal{B})\,dR\\ &= L\left(\sum_{j=0}^{N-L-2}\int_0^1 \frac{R^{L+j+1}}{L+j}\,dR + \frac{1}{N-1}\int_0^1 R\,dR\right)\\ &= L\left(\sum_{j=0}^{N-L-2}\frac{1}{(L+j)(L+j+2)} + \frac{1}{2(N-1)}\right)\\ &= \frac{L}{2}\left(\sum_{j=0}^{N-L-2}\left(\frac{1}{L+j} - \frac{1}{L+j+2}\right) + \frac{1}{N-1}\right)\\ &= \frac{L}{2}\left(\frac{1}{L} + \frac{1}{L+1} - \frac{1}{L+N-L-1} - \frac{1}{L+N-L} + \frac{1}{N-1}\right)\\ &= \frac{1}{2}\left(2 - \frac{1}{L+1} - \frac{L}{N}\right) \qquad .\end{aligned}$$

Das Ergebnis ist korrekt für $L = 0$. In diesem Fall ist das Maximum der Lernphase $\xi = 0$, da die Lernphase die Länge Null hat. Das bedeutet, der erste Wert wird akzeptiert und der Mittelwert hiervon ist $1/2$. Ebenso muss bei $L = N-1$ der nächste und letzte Vorschlag akzeptiert werden. Deshalb ist auch hierbei der Mittelwert $1/2$.
Das Maximum des mittleren Gewinns als Funktion von L ergibt sich aus

$$\frac{\partial}{\partial L}\langle R(L)\rangle = \frac{1}{(L+1)^2} - \frac{1}{N} = 0 \qquad .$$

Das Maximum liegt also bei

$$L^* = \sqrt{N} - 1 \qquad .$$

Der maximale mittlere Gewinn ist demnach

$$\langle R(L^*)\rangle = \frac{1}{2}\big(2 - \frac{1}{\sqrt{N}} - \frac{\sqrt{N}-1}{N}\big) = 1 - \frac{1}{\sqrt{N}} + \frac{1}{2N} \qquad .$$

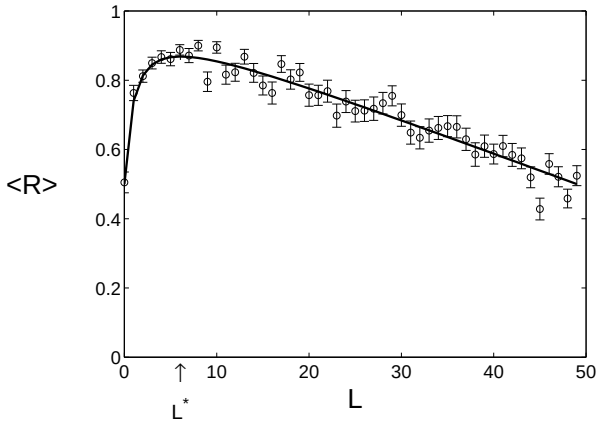

Abbildung 8.2: *Erwarteter Gewinn mit der im Text beschriebenen Strategie und einer Lernphase der Länge L als Funktion von L mit $N = 50$. Die durchgezogene Kurve stellt die berechneten Ergebnisse dar, die mit denen der Computersimulation (Kreise+Fehlerbalken) zu vergleichen sind.*

Kapitel 9

Kontinuierliche Variablen

Bisher haben wir uns fast ausschließlich mit diskreten Problemen beschäftigt, wie z.B. Würfelprobleme, Münzwurf, Urnenprobleme etc. Viele Probleme sind jedoch kontinuierlicher Natur.

9.1 Verteilungsfunktion und Dichtefunktion

Die einfachste Möglichkeit, kontinuierlichen Variablen in die Wahrscheinlichkeitstheorie einzuführen, bietet die Verteilungsfunktion.

Def. 9.1 (Verteilungsfunktion) *Unter der Verteilungsfunktion versteht man die Wahrscheinlichkeit*

$$F(x) = P(\mathbf{x} \leq x|\mathcal{B}) \qquad ,$$

dass eine Zufalls-Variable $\mathbf{x}$ *Werte kleiner oder gleich* x *annimmt. Man nennt die Verteilungsfunktion auch kumulative Wahrscheinlichkeit.*

Daneben definiert man die Dichtefunktion (Wahrscheinlichkeitsdichte)

Def. 9.2 (Wahrscheinlichkeitsdichte) *Das ist die Ableitung der Verteilungsfunktion nach der Zufalls-Variablen*

$$p(x) = \frac{d}{dx}F(x) \qquad .$$

Nach der Definition der Ableitung ist

$$p(x)\,dx = \Big(F(x+dx) - F(x)\Big) = P\big(x < \mathbf{x} \leq x + dx|\mathcal{B}\big) \qquad .$$

Das heißt die Wahrscheinlichkeitsdichte ist die Wahrscheinlichkeit, dass die Zufalls-Variable $\mathbf{x}$ Werte aus dem infinitesimalen Intervall dx bei x annimmt, dividiert durch die Intervall-Größe dx. Die Wahrscheinlichkeitsdichte lässt sich sofort auf höhere Dimensionen $x \in \mathbb{R}^n$ verallgemeinern. Sie ist dann die Wahrscheinlichkeit, dass die

Zufalls-Variable Werte aus dem infinitesimalen Volumen dV_x bei x annimmt, dividiert durch das Volumen.

Man kann in einer Dimension umgekehrt aus der Wahrscheinlichkeitsdichte leicht die Verteilungsfunktion berechnen

$$F(x) = \int_{-\infty}^{x} p(x')\, dx' \qquad .$$

Für mehrere Dimensionen ist dieser Ausdruck entsprechend zu erweitern. Wir werden diesen Fall im Weiteren aber nicht benötigen.

Die Verteilungsfunktion und die Dichtefunktion lassen sich auch für diskrete Probleme definieren. Das bietet eine Möglichkeit, alle Probleme als kontinuierliche Probleme zu behandeln. Die Werte der Zufalls-Variablen seien x_k $(k = 0, 1, \ldots, N)$, mit den Wahrscheinlichkeiten P_k. Die Verteilungsfunktion $F(x)$ ist dann die Wahrscheinlichkeit, dass der Wert der Zufalls-Variablen x_k kleiner oder gleich als x ist

$$F(x) = P(x_k \leq x|\mathcal{B}) = \sum_{\substack{k=0 \\ x_k \leq x}}^{N} P_k \qquad , \tag{9.1}$$

Die Wahrscheinlichkeitsdichte $p(x)$ ist

$$p(x) = \sum_{k=0}^{N} P_k\, \delta(x - x_k) \qquad , \tag{9.2}$$

wie man durch Einsetzen verifizieren kann

$$F(x) = \int_{-\infty}^{x} p(x)\, dx = \sum_{k=0}^{N} P_k \underbrace{\int_{-\infty}^{x} \delta(x' - x_k)\, dx'}_{\theta(x_k \leq x)} = \sum_{\substack{k=0 \\ x_k \leq x}}^{N} P_k \qquad .$$

9.1.1 Beispiel eines kontinuierlichen Problems

Es werden zufällig Teilchen auf ein kreisförmiges Target geschossen. Es ist sichergestellt, dass alle Teilchen innerhalb des Radius R landen. Wir suchen die Wahrscheinlichkeit, dass die Teilchen innerhalb des Kreises vom Radius $r \leq R$ angetroffen werden. Die günstigen Ereignisse haben die Fläche πr^2. Die gesamte Fläche beträgt πR^2. Somit ist die gesuchte Wahrscheinlichkeit

$$F(x) = P(\mathbf{x} < x|\mathcal{B}) = \begin{cases} 0 & x \leq 0 \\ \frac{x^2}{R^2} & 0 \leq x \leq R \\ 1 & R < x \end{cases} \qquad .$$

Es soll hier noch einmal die Bedeutung des Bedingungskomplex erklärt werden. $\mathcal{B}$ enthält hierbei z.B. die Information, dass der Abstand vom Radius nicht größer als R werden kann, und dass alle Punkte auf dem Kreis vom Radius R gleich wahrscheinlich sind.
Die Wahrscheinlichkeitsdichte erhalten wir aus der Ableitung der Verteilungsfunktion

$$p(x|\mathcal{B}) = \frac{d}{dx}F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{2\,x}{R^2} & 0 < x \leq R \\ 0 & R < x \end{cases} \quad .$$

Die Wahrscheinlichkeit, in den infinitesimale Kreisring mit Radius $x \in [x, x + dx)$ zu treffen, ist Null für nicht erlaubte Radien, also für $x < 0$ oder $x > R$. Für erlaubte Radien steigt diese Wahrscheinlichkeit mit der Fläche der Kreisrings, also mit x linear an. In Abbildung 9.1 sind die Verteilungsfunktion und die Wahrscheinlichkeitsdichte

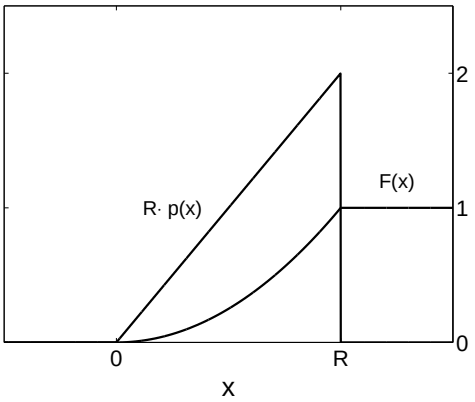

Abbildung 9.1: *Verteilungsfunktion und Wahrscheinlichkeitsdichte des Target-Problems.*

dargestellt.

9.1.2 Beispiel eines diskreten Problems

Wir betrachten ein Bernoulli-Experiment mit $n = 6$ Wiederholungen und einer Einzel-Wahrscheinlichkeit $p = 1/2$ für das Auftreten eines Ereignisses E. Die Wahrscheinlichkeit, dass das Ereignis k-mal auftritt, ist durch die Binomial-Verteilung gegeben

$$P(k|n = 6, p = 1/2) = 2^{-6}\binom{6}{k} \quad .$$

Die Verteilungsfunktion lautet dann gemäß Gl. (9.1)

$$F(x) = P(k \leq x|\mathcal{B}) = 2^{-6} \sum_{\substack{k=0 \\ k \leq x}}^{N} \binom{6}{k} \quad .$$

Die Wahrscheinlichkeitsdichte erhalten wir aus Gl. (9.2)

$$p(x|\mathcal{B}) = 2^{-6} \sum_{k=0}^{6} \binom{6}{k} \delta(x-k) \qquad .$$

Dichte- und Verteilungsfunktion der Bernoulli-Verteilung sind in Abbildung 9.2 dargestellt.

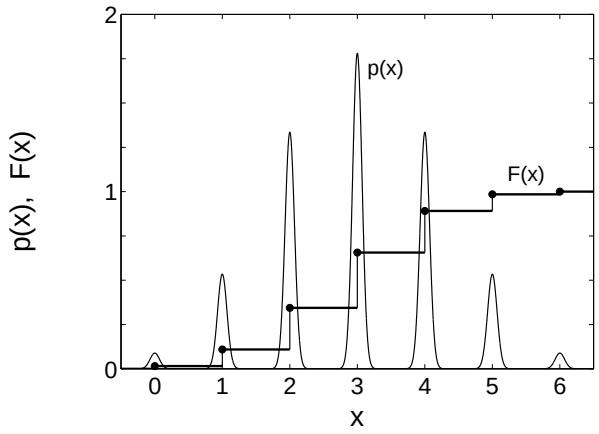

Abbildung 9.2: *Dichte- und Verteilungsfunktion der Bernoulli-Verteilung für* $n = 6$ *und* $p = 1/2$. *Die* δ*-Funktionen in der Wahrscheinlichkeitsdichte wurden durch normierte Gauß-Funktionen endlicher Breite approximiert, um eine graphische Darstellung zu ermöglichen. Die Punkte sollen andeuten, dass die Intervalle links abgeschlossen und die rechts offen sind.*

9.2 Weitere Definitionen

9.2.1 Definition von Mittelwert, Momenten und marginaler Verteilung

In Kapitel 2 haben wir für diskrete Zufallsvariablen den Mittelwert und die Momente definiert. Analog definieren wir

Def. 9.3 (Mittelwert einer kontinuierlichen Zufallsvariable) *Gegeben sei eine Zufallsvariable* X *und die dazugehörige Wahrscheinlichkeitsdichte* $p(X = x) = p(x)$. *Dann definiert man als*

MITTELWERT EINER KONTINUIERLICHEN ZUFALLSVARIABLE

$$\langle X \rangle := \int_{-\infty}^{\infty} x\, p(x)\, dx \qquad . \tag{9.3}$$

Wieder schreiben wir in Zukunft $\langle x \rangle$ anstatt $\langle X \rangle$. MITTELWERTE VON FUNKTIONEN von Zufallsvariablen und MARGINALE VERTEILUNG definiert man analog zu Gl. (2.12) und (2.11), Summen sind durch Integrale zu ersetzen. Die Definitionen der MOMENTE können wir wörtlich aus Kapitel 2 übernehmen, es muss nur auf die korrekte Mittelwertbildung geachtet werden.

9.2.2 Definition einer Stichprobe

In der Praxis kennt man oft die theoretische Verteilung von gewünschten (Mess-)Größen nicht. Einzelne Realisierungen davon sind aber durchaus (experimentell) zugänglich. Sie spielen daher eine zentrale Rolle beim Analysieren unbekannter Verteilungen.

Def. 9.4 (Stichprobe) *X sei eine Zufallsvariable mit einer Verteilungsfunktion $x \to F(x)$. L unabhängige Feststellungen (z.B. Ergebnisse von Experimenten) $x_1, x_2, \ldots, x_L$ über die Zufallsvariable X heißen eine Stichprobe vom Umfang L.*

9.3 Ordnungs-Statistik

Wir betrachten folgende Aufgabenstellung. Gegeben sei eine Stichprobe vom Umfang L von identisch unabhängig verteilten (i.u.v.) (englisch i.i.d.) kontinuierlichen Zufallszahlen der Wahrscheinlichkeitsdichte $\rho(x)$. Die zugehörige Verteilungsfunktion sei $F(x)$. Die Elemente der Stichprobe werden nach ansteigendem Wert sortiert

$$s_1 \leq s_2 \leq \ldots \leq s_k \leq s_{k+1} \leq \ldots \leq s_L \qquad .$$

Wir suchen die Wahrscheinlichkeit $P(s_k \in (x, x+dx)|L, \rho, \mathcal{B})$, dass das k-te Element s_k der sortierten Liste im Intervall $(x, x+dx)$ liegt. Dazu müssen drei Propositionen gleichzeitig erfüllt sein

- A: *$k-1$ Elemente sind kleiner-gleich x.*
- B: *$L-k$ Elemente sind größer-gleich x.*
- C: *Ein Elemente liegt im Intervall $(x, x+dx)$*

Die Wahrscheinlichkeit, dass ein Element kleiner ist als x, ist durch die Verteilungsfunktion $p_1 = F(x)$ gegeben. Entsprechend ist die Wahrscheinlichkeit, dass ein Element größer-gleich x ist, $p_2 = (1 - F(x))$. Schließlich ist die Wahrscheinlichkeit, ein Element in $(x, x + dx)$ anzutreffen, $p_3 = \rho(x)\ dx$. Es handelt sich im Prinzip um das Problem, L Teilchen auf drei Boxen zu verteilen, von denen die Einzel-Wahrscheinlichkeiten p_α sind. Wie groß ist die Wahrscheinlichkeit, in der ersten Box $k-1$, in der zweiten $L-k$ und in der dritten Box ein Teilchen anzutreffen. Die Lösung ist die Multinomial-Verteilung (siehe Gl. (3.11a))

ORDNUNGS-STATISTIK, ORDER STATISTICS

$$P(s_k \in (x, x+dx)|L, \rho, \mathcal{B}) = \frac{L!}{(k-1)!(L-k)!}\ F(x)^{k-1}\ (1 - F(x))^{L-k}\ \underbrace{\rho(x)\ dx}_{dF(x)} \quad . \tag{9.4}$$

9.3.1 Wahrscheinlichkeitsverteilung von Maximalwerten

Wir betrachten L Zufallszahlen $x_1, x_2, \ldots, x_L$ einer Wahrscheinlichkeitsdichte $p(x)$. Wir fragen nach der Wahrscheinlichkeitsdichte der dabei auftretenden Maximalwerte

$$p(\xi|L, \mathcal{B}) \quad .$$

Diese Wahrscheinlichkeit ist nichts anderes als die Ordnungs-Statistik für $k = L$. Die gesuchte Wahrscheinlichkeit ist demnach

MAXIMA-STATISTIK

$$p(\xi|L, \mathcal{B}) = L\ p(\xi)\ F(\xi)^{L-1} \quad . \tag{9.5}$$

$F(x)$ ist die Verteilungsfunktion zur Wahrscheinlichkeitsdichte $p(x)$ ist.

Wir illustrieren das am Spezialfall gleich-verteilter Zufallszahlen. In diesem Fall ist die Wahrscheinlichkeitsdichte $p(x) = \theta(0 \le x \le 1)$ und die Verteilungsfunktion $F(x) = \theta(0 \le x \le 1)\ x$. Die Maximalwert-Wahrscheinlichkeit ist somit

$$p(\xi|L, \mathcal{B}) = L\ \theta(0 \le x \le 1)\ \xi^{L-1} \quad .$$

Der Erwartungswert der Maxima ist dann

$$\langle \xi \rangle = \int_0^1 \xi \; p(\xi|L,\mathcal{B}) \; d\xi = L \int_0^1 \xi^L \; d\xi = \frac{L}{L+1} = 1 - \frac{1}{L+1} \qquad . \tag{9.6}$$

Diese Überlegungen können leicht auf dem Computer simuliert und getestet werden. Es werden L Zufallszahlen erzeugt, daraus das Maximum bestimmt und das Ganze M mal wiederholt um Mittelwert und Standardfehler zu ermitteln. Die Ergebnisse sind in Abbildung 9.3 als Funktion von L dargestellt.

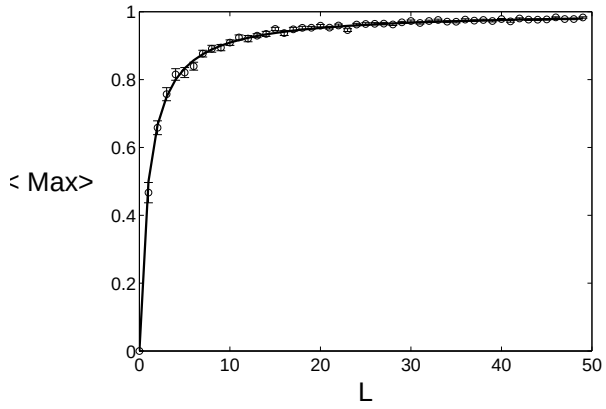

Abbildung 9.3: *Mittelwert der Maxima von L gleich-verteilten Zufallszahlen als Funktion von L. Die durchgezogene Kurve stellt die berechneten Ergebnisse dar, die mit denen der Computersimulation (Kreise+Fehlerbalken) innerhalb der Fehlerbalken übereinstimmen.*

9.4 Gängige Wahrscheinlichkeitsverteilungen

9.4.1 Gleich-Verteilung im Intervall $[a, b]$

Für die Gleich-Verteilung im Intervall $[a, b]$ führen wir die erweiterte Stufenfunktion ein

$$p(x|a,b) = \frac{1}{b-a} \; \theta(a \le x \le b) \qquad .$$

Der Mittelwert ist offensichtlich $(b+a)/2$, und die Varianz erhalten wir aus

$$\left\langle x^2 \right\rangle = \frac{1}{b-a} \int_a^b x^2 \; dx = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3} \qquad .$$

Damit ist die Varianz

$$\mathrm{var}(x) = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12} \qquad .$$

GLEICH-VERTEILUNG

Def.bereich: $x \in [a, b]$

$$p_g(x|a,b) = \frac{1}{b-a} \tag{9.7a}$$

$$F_g(x|a,b) := \frac{x-a}{b-a} \tag{9.7b}$$

$$\langle x \rangle = \frac{a+b}{2} \tag{9.7c}$$

$$\mathrm{var}(x) = \frac{(b-a)^2}{12} \tag{9.7d}$$

Besonders prominent ist die Gleich-Verteilung auf dem Einheitsintervall $p_g(x|0,1)$ mit Mittelwert $1/2$ und Varianz $1/12$. Hierfür gibt es in jeder vernünftigen Programmiersprache einen PSEUDO-ZUFALLSZAHLEN-Generator, der i.d.R. RND oder RAND heißt.

9.4.2 β-Verteilung

Im Zusammenhang mit Bernoulli-Problemen spielt die β-Verteilung eine wichtige Rolle.

β-VERTEILUNG

Def.bereich: $x \in [0, 1]$

$$p_\beta(x|\alpha, \rho) = \frac{1}{B(\alpha, \rho)}\, x^{\alpha-1}\, (1-x)^{\rho-1} \tag{9.8a}$$

$$F_\beta(x|\alpha, \rho) := \frac{1}{B(\alpha, \rho)} \int_0^x p^{\alpha-1}\, (1-p)^{\rho-1}\, dp \tag{9.8b}$$

$$\langle x \rangle = \frac{\alpha}{\alpha + \rho} \tag{9.8c}$$

$$\mathrm{var}(x) = \frac{\langle x \rangle\, (1 - \langle x \rangle)}{\alpha + \rho + 1} \tag{9.8d}$$

Der Normierungsfaktor $B(\alpha, \rho)$ ist das β-Integral bzw. die β-Funktion.

β-FUNKTION UND UNVOLLSTÄNDIGE β-FUNKTION

$$B(\alpha, \rho) := \int_0^1 p^{\alpha-1}\, (1-p)^{\rho-1}\, dp = \frac{\Gamma(\alpha)\, \Gamma(\rho)}{\Gamma(\alpha + \rho)} \tag{9.9a}$$

$$B(x; \alpha, \rho) := \int_0^x p^{\alpha-1}\, (1-p)^{\rho-1}\, dp \tag{9.9b}$$

Die unvollständige β-Funktion ist bis auf die Normierung gleich der Verteilungsfunktion der β-Verteilung.

Die Kenngrößen Mittelwert und Varianz der β-Verteilung lassen sich leicht mit Gl.

(9.9a) bestimmen

$$\begin{aligned}\langle x\rangle &= \int_0^1 x\; p_\beta(x|\alpha,\rho)\; dx \\ &= \frac{B(\alpha+1,\rho)}{B(\alpha,\rho)} = \frac{\Gamma(\alpha+1)\;\Gamma(\rho)\;\Gamma(\alpha+\rho)}{\Gamma(\alpha)\;\Gamma(\rho)\;\Gamma(\alpha+\rho+1)} = \frac{\alpha}{\alpha+\rho}\end{aligned}$$

$$\begin{aligned}\left\langle x^2\right\rangle &= \frac{B(\alpha+2,\rho)}{B(\alpha,\rho)} = \frac{\Gamma(\alpha+2)\;\Gamma(\rho)\;\Gamma(\alpha+\rho)}{\Gamma(\alpha)\;\Gamma(\rho)\;\Gamma(\alpha+\rho+2)} \\ &= \frac{\alpha\;(\alpha+1)}{(\alpha+\rho)\;(\alpha+\rho+1)} \qquad .\end{aligned}$$

Hieraus erhält man die Varianz

$$\mathrm{var}(x) = \left\langle x^2\right\rangle - \langle x\rangle^{\,2} = \frac{\langle x\rangle(1-\langle x\rangle)}{\alpha+\rho+1} \qquad .$$

In Abbildung 9.4 ist die Verteilungsfunktion und die Dichte der β-Verteilung abgebildet. Die Parameter α, ρ wurden so gewählt, dass $\langle x\rangle = .5$ und $\mathrm{var}(x) = 0.01, 1/12, 0.2$. Der Wert $\mathrm{var}(x) = 1/12$ ist insofern ausgezeichnet, als dass die Varianz der Gleich-Verteilung auf dem Einheits-Intervall ist. Für kleinere Werte der Varianz besitzt die Beta-Dichte ein Maximum bei $x = 1/2$. Um eine größere Varianz als die der Gleich-Verteilung zu erreichen, muss die Dichte an den Rändern divergieren. Für $\langle x\rangle = 1/2$ ist

$$\alpha = \rho = \frac{1}{2}\left(\frac{1}{4\mathrm{var}(x)} - 1\right) \qquad .$$

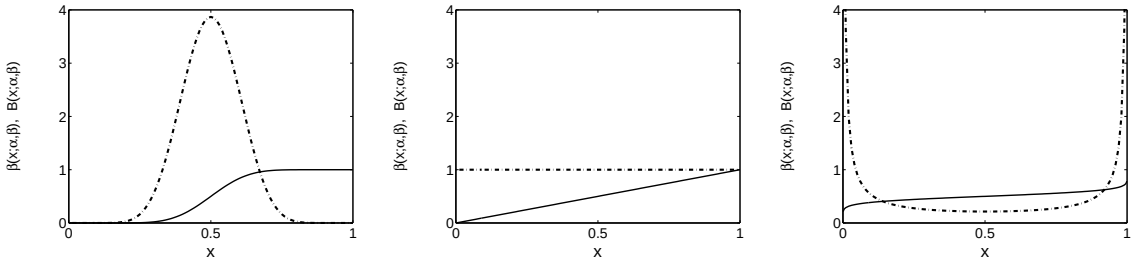

Abbildung 9.4: *Verteilungsfunktion (durchgezogene Linie) und Dichte (gestrichelte Linie) der β-Verteilung. Die Parameter α, ρ wurden so gewählt, dass $\langle x\rangle = 1/2$ und $\mathrm{var}(x) = 0.01, 1/12, 0.2$.*

Wir können die Beta-Funktionen verwenden, um die kumulative Wahrscheinlichkeit der Bernoulli-Verteilung anzugeben[1]

[1] Den Beweis findet man z.B. im Buch von Meschkowski (S.109ff).

KUMULATIVE WAHRSCHEINLICHKEIT DER BERNOULLI-VERTEILUNG

$$
\begin{aligned}
F_\beta(x|n,p) &= \sum_{\substack{r=0 \\ r<x}}^{n} \binom{n}{r} p^r \, (1-p)^{n-r} \\
&= 1 - \frac{B(p; r+1, n-r)}{B(r+1, n-r)} \,, \\
&\qquad \text{mit } r < x \leq r+1, \quad p < 1 \qquad .
\end{aligned}
\tag{9.10}
$$

Die β-Funktionen werden später bei der Analyse von Urnen-Problemen wiederholt auftreten.

9.4.3 Γ-Verteilung, χ^2-Verteilung

Eine weitere sehr wichtige Verteilung ist die

Γ-VERTEILUNG

Def.bereich: $x \in [0, \infty)$

$$p_\Gamma(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \, x^{\alpha-1} \, e^{-\beta\, x} \tag{9.11a}$$

$$F_\Gamma(x|\alpha,\beta) := \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} \, e^{-\beta\, t} \, dt \tag{9.11b}$$

$$\langle x \rangle = \frac{\alpha}{\beta} \tag{9.11c}$$

$$\mathrm{var}(x) = \frac{\alpha}{\beta^2} \tag{9.11d}$$

Mittelwert und Varianz der Γ-Verteilung lassen sich leicht berechnen

$$\begin{aligned}
\langle x \rangle &= \int_0^\infty x\, p_\Gamma(x|\alpha,\beta)\, dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{(\alpha+1)-1}\, e^{-\beta\, x}\, dx \\
&= \frac{1}{\beta\, \Gamma(\alpha)} \int_0^\infty (\beta x)^{(\alpha+1)-1}\, e^{-\beta\, x}\, d(\beta x) \\
&= \frac{1}{\beta\, \Gamma(\alpha)} \int_0^\infty z^{(\alpha+1)-1}\, e^{-z}\, dz \\
&= \frac{\Gamma(\alpha+1)}{\beta\, \Gamma(\alpha)} = \frac{\alpha}{\beta}
\end{aligned}$$

$$\begin{aligned}
\mathrm{var}(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{(\alpha+2)-1}\, e^{-\beta\, x}\, dx \;-\; \left(\frac{\alpha}{\beta}\right)^2 \\
&= \frac{\Gamma(\alpha+2)}{\beta^2\, \Gamma(\alpha)} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha\,(\alpha+1)}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}
\end{aligned}$$

In Abbildung 9.5 ist die Verteilungsfunktion und die Dichte der Γ-Verteilung für $\alpha = 4$ abgebildet. Der Parameter β geht nur in der Skalierung der x-Achse in der Kombination βx ein. Für $0 < \alpha < 1$ besitzt die Γ-Dichte eine Divergenz bei $x = 0$, für $\alpha = 1$ ist sie identisch der Exponential-Funktion und für $\alpha > 1$ besitzt sie ein Maximum bei $\beta x = \alpha - 1$. Die kumulative Wahrscheinlichkeit der Γ-Verteilung ist die auf

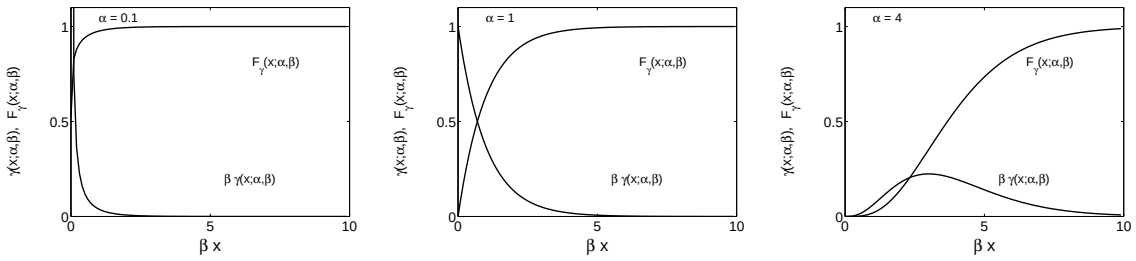

Abbildung 9.5: *Verteilungsfunktion und Dichte der* Γ*-Verteilung für* $\alpha = 0.1, 1, 4$.

Eins normierte unvollständige Γ-Funktion

$$F_\Gamma(x) = \frac{\Gamma(\beta\, x; \alpha}{\Gamma(\alpha)} \qquad .$$

Γ-Funktion und unvollständige Γ-Funktion

$$\Gamma(\alpha) := \int_0^\infty t^{\alpha-1}\, e^{-t}\, dt \tag{9.12a}$$

$$\Gamma(x;\alpha) := \int_0^x t^{\alpha-1}\, e^{-t}\, dt \tag{9.12b}$$

$$\Gamma(\alpha) = (\alpha-1)\,\Gamma(\alpha-1) \tag{9.12c}$$

$$\Gamma(\alpha) = (n-1)! \quad \text{für } n \in \mathbb{N} \tag{9.12d}$$

$$\Gamma(1/2) = \sqrt{\pi} \tag{9.12e}$$

Einen Spezialfall ($\alpha = \frac{n}{2}, \beta = \frac{1}{2}$) hiervon stellt die χ^2-Verteilung dar

χ^2-Verteilung

Def.bereich: $x \in [0, \infty)$

$$p_{\chi^2}(x|n) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})}\, x^{\frac{n}{2}-1}\, e^{-\frac{1}{2}x} \tag{9.13a}$$

$$= p_\Gamma(x|\alpha = \tfrac{n}{2}, \beta = \tfrac{1}{2})$$

$$F_{\chi^2}(x|n) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})} \int_0^\infty t^{\frac{n}{2}-1}\, e^{-\frac{1}{2}t}\, dt \tag{9.13b}$$

$$\langle x \rangle = n \tag{9.13c}$$

$$\mathrm{var}(x) = 2n \tag{9.13d}$$

n : Zahl der Freiheitsgrade.

Wir wollen hier die kumulative Wahrscheinlichkeit der Poisson-Verteilung $w(k|\lambda)$ untersuchen

$$F(x) = \sum_{\substack{k=0 \\ k<x}}^{\infty} e^{-\lambda}\, \frac{\lambda^k}{k!} \qquad .$$

Ähnlich wie im Falle der Bernoulli-Verteilung lässt sich die Summation über die

Poisson-Verteilung ausführen, und das führt zu einem ähnlich aussehenden Ergebnis

KUMULATIVE WAHRSCHEINLICHKEIT DER POISSON-VERTEILUNG

$$F(x) = 1 - \frac{\Gamma(\lambda; r+1)}{\Gamma(r+1)}\,, \qquad r < x \le r+1 \qquad . \tag{9.14}$$

Da der Beweis an sich interessant ist, soll er hier vorgeführt werden. Die Beweisidee ist dieselbe im Fall von Gl. (9.10).
Wir gehen aus von der unvollständigen Γ-Funktion

$$\Gamma(\lambda; \alpha)$$

und erhalten durch partielle Integration daraus die Rekursionsformel

$$\begin{aligned}
\Gamma(\lambda; \alpha+1) &= \int_0^\lambda t^\alpha\, e^{-t}\, dt = -\int_0^\lambda t^\alpha \left(\frac{d}{dt}\, e^{-t}\right) dt \\
&= -t^\alpha\, e^{-t}\Big|_0^\lambda + \int_0^\lambda \left(\frac{d}{dt} t^\alpha\right) e^{-t}\, dt \\
&= \begin{cases} -\lambda^\alpha\, e^{-\lambda} + \alpha\Gamma(\lambda;\alpha) & \alpha > 0, \\ 1 - e^{-\lambda} & \alpha = 0 \end{cases} \qquad .
\end{aligned}$$

Für $\alpha > 0$ erhalten wir daraus

$$\frac{\lambda^\alpha\, e^{-\lambda}}{\alpha!} = \frac{\Gamma(\lambda, \alpha)}{(\alpha-1)!} - \frac{\Gamma(\lambda, \alpha+1)}{\alpha!}$$

und somit

$$\begin{aligned}
\sum_{\alpha=0}^{r} \frac{\lambda^\alpha\, e^{-\lambda}}{\alpha!} &= e^{-\lambda} + \sum_{\alpha=1}^{r} \frac{\lambda^\alpha\, e^{-\lambda}}{\alpha!} \\
&= e^{-\lambda} + \sum_{\alpha=1}^{r} \frac{\Gamma(\lambda, \alpha)}{(\alpha-1)!} - \sum_{\alpha=1}^{r} \frac{\Gamma(\lambda, \alpha+1)}{\alpha!} \\
&= e^{-\lambda} + \sum_{\alpha=0}^{r-1} \frac{\Gamma(\lambda, \alpha+1)}{\alpha!} - \sum_{\alpha=1}^{r} \frac{\Gamma(\lambda, \alpha+1)}{\alpha!} \\
&= e^{-\lambda} + \frac{\Gamma(\lambda, 1)}{0!} - \frac{\Gamma(\lambda, r+1)}{r!} \\
&= e^{-\lambda} + 1 - e^{-\lambda} - \frac{\Gamma(\lambda, r+1)}{\Gamma(\alpha+1)} \qquad .
\end{aligned}$$

9.4.4 Exponential-Verteilung

Ein wichtiger Spezialfall der Γ-Verteilung ist die EXPONENTIAL-VERTEILUNG. Die Dichte und Verteilungsfunktion der EXPONENTIAL-VERTEILUNG sind

EXPONENTIAL-VERTEILUNG

Def.bereich: $x \in [0, \infty)$

$$p_e(x|\lambda) = \lambda \; e^{-\lambda\, x} = p_\Gamma(x|\alpha = 1, \beta = \lambda) \tag{9.15a}$$

$$F_e(x|\lambda) = 1 - e^{-\lambda\, x} \tag{9.15b}$$

$$\langle x \rangle = \frac{1}{\lambda} \tag{9.15c}$$

$$\mathrm{var}(x) = \frac{1}{\lambda^2} \quad . \tag{9.15d}$$

9.4.5 Normal-Verteilung

Def. 9.5 (Normal-Verteilung, Gauß-Verteilung) *Eine Zufalls-Variable heißt* NORMAL VERTEILT, *wenn die Wahrscheinlichkeitsdichte eine Gauß-Funktion ist*

$$p(x) = \mathcal{N}(x|x_0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-x_0)^2/2\sigma^2} \quad . \tag{9.16}$$

Der Definitionsbereich ist die gesamte reelle Achse $x \in \mathbb{R}$. Die Verteilungsfunktion der Normal-Verteilung ist

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-(x'-x_0)^2/2\sigma^2}\, dx' \\ &= \Phi\left(\frac{x - x_0}{\sigma}\right) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{x - x_0}{\sqrt{2}\sigma}\right) \quad . \end{aligned} \tag{9.17}$$

Die Definition von Φ ist in Gl. (4.10) gegeben, die von erf *in Gl. (4.13).*

Wahrscheinlichkeitsdichte und kumulative Wahrscheinlichkeit der Normal-Verteilung sind in Abbildung 9.6 abgebildet.
Die Berechnung der Momente ist hierbei etwas aufwendiger. Der Mittelwert ist wegen der Symmetrie der Gauß-Funktion $\langle x \rangle = x_0$. Wir berechnen nun die geraden

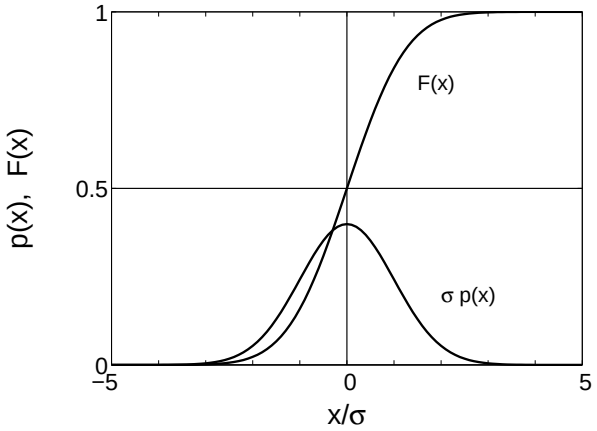

Abbildung 9.6: *Wahrscheinlichkeitsdichte und kumulative Wahrscheinlichkeit der bei Null zentrierten Normal-Verteilung.*

zentrierten Momente. Die ungeraden Momente sind Null wegen der Symmetrie der Gaußfunktion.

$$\begin{aligned}\langle (x-x_0)^n \rangle &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-x_0)^n \; e^{-(x-x_0)^2/2\sigma^2} \; dx \\ &= \frac{2}{\sqrt{2\pi\sigma^2}} \int_{0}^{\infty} z^{n+1} \; e^{-z^2/2\sigma^2} \; \frac{dz}{z}\end{aligned}$$

Nun substituieren wir $t = z^2/2\sigma^2$. D.h. $\frac{dz}{z} = \frac{dt}{2t}$

$$\begin{aligned}\langle (x-x_0)^n \rangle &= \frac{2}{2\sqrt{2\pi\sigma^2}} (2\sigma^2)^{\frac{n+1}{2}} \int_{0}^{\infty} t^{\frac{n+1}{2}} \; e^{-t} \; \frac{dt}{t} \\ &= \frac{1}{\sqrt{\pi}} (2\sigma^2)^{n/2} \; \Gamma\left(\frac{n+1}{2}\right)\end{aligned}$$

Hieraus erhalten wir speziell die Varianz (n=2)

$$\mathrm{var}(x) = \frac{1}{\sqrt{\pi}} 2\sigma^2 \; \Gamma\left(\frac{3}{2}\right) = \frac{1}{\sqrt{\pi}} 2\sigma^2 \; \frac{\sqrt{\pi}}{2} = \sigma^2$$

NORMAL-VERTEILUNG $\mathcal{N}(x|x_0,\sigma)$

Def.bereich: $x \in \mathbb{R}$

$$p(x|x_0,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-x_0)^2/2\sigma^2} \tag{9.18a}$$

$$F(x|x_0,\sigma) = \Phi\left(\frac{x-x_0}{\sigma}\right) \tag{9.18b}$$

$$\langle x\rangle = x_0 \tag{9.18c}$$

$$\mathrm{var}(x) = \sigma^2 \tag{9.18d}$$

Varianz-Marginalisierung der Normalverteilung

Es kommt in Anwendungen häufig vor, dass man zwar weiß, dass die Fehler der Messungen normal verteilt sind, aber man kennt die Varianz nicht. Angenommen, man habe N solcher Datenpunkte, die zur Bestimmung einer physikalischen Größe μ dienen und die alle einen gemeinsamen Messfehler σ aufweisen. Die Likelihood lautet dann

$$\begin{aligned} p(\vec{d}|\mu,\sigma,\mathcal{B}) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu-d_i)^2}{2\sigma^2}\right) \\ &= (2\pi)^{-N/2}\, \sigma^{-N}\, \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}(\mu-d_i)^2\right)\right\} \quad . \end{aligned}$$

Die Summe in der Exponentialfunktion vereinfachen wir, indem wir Stichprobenmittelwerte $\overline{d^n} := \frac{1}{N}\sum_{i=1}^{N} d^n$ einführen

$$\begin{aligned} \sum_{i=1}^{N}(\mu-d_i)^2 &= \sum_{i=1}^{N}(\mu^2 - 2\,\mu\, d_i + d_i^2) \\ &= N\mu^2 - 2\,N\,\mu\,\overline{d} + N\,\overline{d^2} \\ &= N(\mu-\overline{d})^2 - N;\overline{d}^2 + N\,\overline{d^2} \\ &= N(\mu-\overline{d})^2 + \overline{(\Delta d)^2} \quad . \end{aligned}$$

Damit wird die Likelihood zu

$$p(\vec{d}|\mu,\sigma,\mathcal{B}) = (2\pi)^{-N/2}\, \sigma^{-N}\, \exp\left\{-\frac{N}{2\sigma^2}\left(\left(\mu-\overline{d}\right)^2 + \overline{(\Delta d)^2}\right)\right\} \quad . \tag{9.19}$$

Wenn nun die Varianz σ nicht bekannt ist, benötigen wir stattdessen die marginale Likelihood $p(\vec{d}|\mu, \mathcal{B})$, die wir über die Marginalisierungsregel bestimmen

$$p(\vec{d}|\mu, \mathcal{B}) = \int_0^\infty p(\vec{d}|\mu, \sigma, \mathcal{B})\; p(\sigma|\mathcal{B})\; d\sigma \qquad . \tag{9.20}$$

Bei σ handelt es sich um eine Skalenvariable. Falls keine weiter Information vorliegt, ist $p(\sigma|\mathcal{B})$ durch Jeffreys' Prior (siehe Kapitel 14) zu quantifizieren. Wir wollen hier diesen uninformative Fall annehmen. Da Jeffreys' Prior nicht normiert ist, können wir die Normierung der marginalen Likelihood nur nachträglich vornehmen. Das zu bestimmende Integral hat die Form

$$p(\vec{d}|\mu, \mathcal{B}) \propto I(\alpha; N) := \int_0^\infty \sigma^{-N-1}\; e^{-\frac{\alpha}{\sigma^2}}\;\; d\sigma\;. \tag{9.21}$$

Hierbei ist $\alpha = \frac{N}{2}\left(\left(\mu - \overline{d}\right)^2 + \overline{(\Delta d)^2}\right)$. Durch die Substitution $t = \frac{\alpha}{\sigma^2}$ lässt sich das Integral auf eine Gamma-Funktion abbilden und liefert schließlich

$$I(\alpha; N) := \int_0^\infty \sigma^{-N-1}\; e^{-\frac{\alpha}{\sigma^2}}\; d\sigma \;= \frac{1}{2}\;\frac{\Gamma(N/2)}{\alpha^{N/2}}\;. \tag{9.22}$$

Somit ist die gesuchte marginale Likelihood

$$p(\vec{d}|\mu, \mathcal{B}) = \frac{1}{Z}\left(\left(\mu - \overline{d}\right)^2 + \overline{(\Delta d)^2}\right)^{-N/2} = \frac{1}{Z'}\left(1 + \frac{\left(\mu - \overline{d}\right)^2}{\overline{(\Delta d)^2}}\right)^{-N/2}$$

Die Normierungskonstante erhalten wir aus

$$Z' = \int \left(1 + \frac{\left(\mu - \overline{d}\right)^2}{\overline{(\Delta d)^2}}\right)^{-N/2} d\mu$$

Hierzu benötigen wir das Integral

$$\int \left(1 + \frac{\mu^2}{s^2}\right)^{-N/2} = \sqrt{\pi s^2}\;\frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N}{2})}\;. \tag{9.23}$$

Die Normierung Z' lautet also

$$Z' = \sqrt{\pi\overline{(\Delta d)^2}}\frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N}{2})}\;.$$

Damit haben wir das gesuchte Ergebnis

MARGINALE LIKELIHOOD

$$p(\vec{d}|\mu,\mathcal{B}) = \frac{\Gamma(\frac{N}{2})}{\sqrt{\pi\overline{(\Delta d)^2}\,\Gamma(\frac{N-1}{2})}}\left(1+\frac{\left(\mu-\overline{d}\right)^2}{\overline{(\Delta d)^2}}\right)^{-N/2} \tag{9.24}$$

Wie wir gleich sehen werden ist dies eine Student-t-Verteilung. Interessant ist noch das Verhalten für $N \to \infty$

$$p(\vec{d}|\mu,\mathcal{B}) \underset{N\gg 1}{\longrightarrow} \frac{1}{\sqrt{2\pi\sigma_{se}^2}}\exp\left\{-\frac{-\left(\mu-\overline{d}\right)^2}{2\sigma_{se}^2}\right\}, \text{with} \quad \sigma_{se} = \overline{(\Delta d)^2}/N\ . \tag{9.25}$$

9.4.6 Student-t-Verteilung, Cauchy-Verteilung

STUDENT-t-VERTEILUNG

Def.bereich: $t \in \mathbb{R}$

$$p_t(t|\nu) := \frac{1}{\sqrt{\nu}\ B(\frac{1}{2},\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)} \tag{9.26a}$$

$$F_t(\tau|\nu) = 1\ -\ \frac{1}{2}\ \frac{B\left((1+\frac{\tau^2}{\nu})^{-1};\frac{1}{2},\frac{\nu}{2}\right)}{B(\frac{1}{2},\frac{\nu}{2})} \tag{9.26b}$$

$$\langle t\rangle = 0 \tag{9.26c}$$

$$\mathrm{var}(t) = \frac{\nu}{\nu-2}, \quad \text{für } \nu > 2 \tag{9.26d}$$

ν : Zahl der FREIHEITSGRADE.

Sie entsteht, wenn man aus der Dichte der Normal-Verteilungen $p_{\mathcal{N}}(x|x_0,\sigma)$ die Varianz σ^2 als unbekannt ausintegriert. Sie wird wichtig werden, wenn wir bestimmen wollen, ob zwei Stichproben, die beide einer Normal-Verteilung unbekannter Varianz genügen, denselben Mittelwert haben.

Die Verteilungsfunktion berechnet sich wie folgt

$$\begin{aligned}
F_t(\tau|\nu) &= \frac{1}{B(\frac{1}{2},\frac{\nu}{2})} \int_{-\infty}^{\tau} \left(1+\frac{t^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)} \frac{dt}{\sqrt{\nu}} \\
&= \frac{1}{B(\frac{1}{2},\frac{\nu}{2})} \int_{-\infty}^{\tau/\sqrt{\nu}} \left(1+x^2\right)^{-\frac{1}{2}(\nu+1)} dx \\
&= \frac{1}{B(\frac{1}{2},\frac{\nu}{2})} \int_{-\infty}^{0} \left(1+x^2\right)^{-\frac{1}{2}(\nu+1)} dx \\
&\quad + \frac{1}{B(\frac{1}{2},\frac{\nu}{2})} \int_{0}^{\tau/\sqrt{\nu}} \left(1+x^2\right)^{-\frac{1}{2}(\nu+1)} dx \qquad .
\end{aligned}$$

Wir gehen davon aus, dass $\tau > 0$ ist. Wir substituieren

$$\begin{aligned}
\xi &= (1+x^2)^{-1} \\
x &= \sqrt{\frac{1-\xi}{\xi}} \\
dx &= \frac{1}{2}\, \xi^{-\frac{3}{2}}\, (1-\xi)^{-\frac{1}{2}}\, d\xi
\end{aligned}$$

Somit erhalten wir

$$\begin{aligned}
F_t(\tau|\nu) &= \frac{1}{2}\, \frac{1}{B(\frac{1}{2},\frac{\nu}{2})} \int_0^1 \xi^{\frac{1}{2}(\nu+1)}\, \xi^{-\frac{3}{2}}\, (1-\xi)^{-\frac{1}{2}}\, d\xi \\
&\quad + \frac{1}{2}\, \frac{1}{B(\frac{1}{2},\frac{\nu}{2})} \int_{1/(1+\frac{\tau^2}{\nu})}^1 \xi^{\frac{1}{2}(\nu+1)}\, \xi^{-\frac{3}{2}}\, (1-\xi)^{-\frac{1}{2}}\, d\xi \\
&= \frac{1}{2} \;+\; \frac{1}{2}\, \frac{1}{B(\frac{1}{2},\frac{\nu}{2})} \int_{1/(1+\frac{\tau^2}{\nu})}^1 \xi^{\frac{\nu}{2}-1}\, (1-\xi)^{\frac{1}{2}-1}\, d\xi \\
&= \frac{1}{2} \;+\; \frac{1}{2} \left(1 - \frac{1}{B(\frac{1}{2},\frac{\nu}{2})} \int_0^{1/(1+\frac{\tau^2}{\nu})} \xi^{\frac{\nu}{2}-1}\, (1-\xi)^{\frac{1}{2}-1}\, d\xi\right) \\
&= 1 \;-\; \frac{1}{2}\, \frac{B\left((1+\frac{\tau^2}{\nu})^{-1}; \frac{1}{2}, \frac{\nu}{2}\right)}{B(\frac{1}{2},\frac{\nu}{2})} \qquad .
\end{aligned}$$

Die Cauchy-Verteilung ist ein Spezialfall der Student-t-Verteilung mit $\nu = 1$.

CAUCHY-VERTEILUNG

Def.bereich: $t \in \mathbb{R}$

$$p_C(x) = \frac{1}{\pi\,(1+x^2)} \tag{9.27a}$$

$$F_C(x) = \frac{1}{2} + \frac{\arctan(x)}{\pi} \tag{9.27b}$$

$$\langle x \rangle = 0 \tag{9.27c}$$

$$\mathrm{var}(x) = \infty \tag{9.27d}$$

Man nennt diese Wahrscheinlichkeitsdichte auch LORENTZ-FUNKTION.
Die Dichte- und Verteilungsfunktion der Cauchy-Verteilung sind in Abbildung 9.7 wiedergegeben.

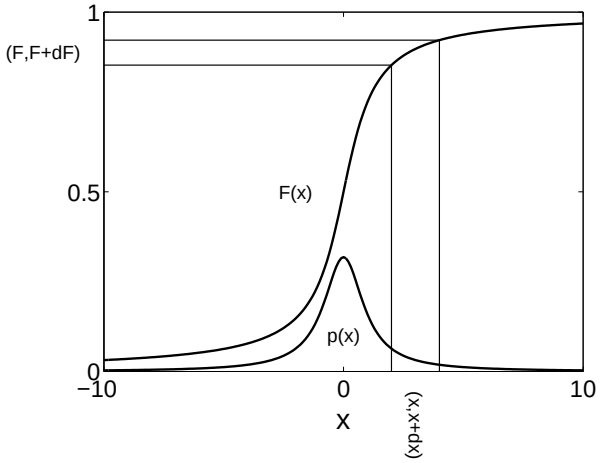

Abbildung 9.7: *Dichte- und Verteilungsfunktion der Cauchy-Verteilung.*

9.4.7 Multivariate Normal-Verteilung

In sehr vielen Anwendungen spielt die Normalverteilung von mehreren Zufallsvariablen (MULTIVARIATE NORMALVERTEILUNG) eine zentrale Rolle. Die Wahrscheinlichkeitsdichte dieser Verteilung lautet

MULTIVARIATE NORMAL-VERTEILUNG

Def.bereich: $x \in \mathbb{R}^n$

$$p_N(x|C, x^0) := \frac{1}{\sqrt{(2\pi)^n \ |C|}} \ e^{-\frac{1}{2}(x-x^0)^T \ C^{-1} \ (x-x^0)} \tag{9.28a}$$

$$\langle x_i \rangle = x_i^0 \tag{9.28b}$$

$$\mathrm{cov}(x_i, x_j) = C_{ij} \tag{9.28c}$$

$$|C| = \det(C) \qquad . \tag{9.28d}$$

Wir werden in diesem Zusammenhang sehr häufig Gauß-Integrale benötigen.

GAUSS-INTEGRALE

$$\int e^{-\frac{1}{2}((x-x^0)^T A(x-x^0)+2x^T b)} \ d^n x = \frac{(2\pi)^{\frac{n}{2}}}{\sqrt{|A|}} \ e^{\frac{1}{2}b^T A^{-1} b - b^T x^0} \quad . \tag{9.29}$$

Zur Berechnung des Mittelwertes $\langle x_i \rangle$ benötigen wir die marginale Dichte. Wenn die Zufallsvariablen unabhängig sind ($p(x_1, \ldots, x_n) = p_0(x_1) \ldots p_0(x_n)$) ist die marginale Dichte $p_i(x_i) = p_0(x_i)$. In der Regel sind die marginalen Dichten nicht alle gleich. Die marginale Dichte der multivariaten Normalverteilung lautet

MARGINALE DICHTE DER MULTIVARIATEN NORMALVERTEILUNG

$$p_i(x) = \frac{1}{\sqrt{2 \ \pi \ C_{ii}}} \ e^{-\frac{(x-x_i^0)^2}{2 \ C_{ii}}} \qquad . \tag{9.30}$$

9.5 Transformationseigenschaften

Wenn man eine kontinuierliche Variable transformiert

$$x \to y = f(x), \qquad x, y \in \mathbb{R}^n \qquad ,$$

transformiert sich auch das Volumenelement. Wenn wir dasselbe Ereignis in zwei unterschiedlichen Koordinatensystemen x bzw. y beschreiben, sollte gelten

$$p_x(x)\ dV_x = p_y(y)\ dV_y \qquad , \tag{9.31}$$

da in beiden Fällen die Wahrscheinlichkeit desselben „infinitesimalen" Ereignisse angegeben ist. Somit transformiert sich die Wahrscheinlichkeitsdichte

VARIABLEN TRANSFORMATION

$$p_y(y) = p_x(x)\ \left|\frac{\partial x_i}{\partial y_j}\right| \qquad . \tag{9.32}$$

Die Änderung der Volumina wird durch die JAKOBI-DETERMINANTE

$$\left|\frac{\partial x_i}{\partial y_j}\right| = \left|\begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}\right| \tag{9.33}$$

berücksichtigt.

9.5.1 Beispiele mit einer Variablen

Wir gehen von der Gleich-Verteilung auf dem Einheits-Intervall $x \in (0, 1]$ aus

$$p_x(x) = \theta(0 < x \leq 1) \qquad .$$

Nun gehen wir zu einer neuen Variablen

$$y = -\ln(x)\ \in [0, \infty)$$

über. Die Umkehr-Transformation lautet $x = e^{-y}$. Wir berechnen die zugehörige Wahrscheinlichkeitsdichte $p_y(y)$

$$p_y(y) = p_x(x) \left|\frac{\partial x}{\partial y}\right| = e^{-y} \qquad .$$

Als weiteres Beispiel betrachten wir eine UNEIGENTLICHE Wahrscheinlichkeitsdichte

$$p_x(x) = 1 \ , \qquad x \in (-\infty, \infty) \qquad .$$

Es handelt sich wieder um eine Gleich-Verteilung, nun aber auf der gesamten reellen Achse. Diese Wahrscheinlichkeitsdichte lässt sich nicht normieren. Sie spielt aber, wie wir später sehen werden, eine wichtige Rolle in Wahrscheinlichkeitstheorie. Um damit rechnen zu können muss das Intervall zunächst auf ein endliches Intervall $[-a, a]$ eingeschränkt werden. Am Ende der Rechnung lassen wir dann a gegen unendlich gehen. Wir betrachten nun die Transformation $x = \ln(\sigma)$ (bzw. $\sigma = e^x$) für $\sigma \in (0, \infty)$. Die Wahrscheinlichkeitsdichte für σ ist dann

$$p_\sigma(\sigma) = p_x(x) \frac{dx}{d\sigma} = \frac{1}{\sigma} \qquad . \tag{9.34}$$

Das heißt, eine Gleich-Verteilung im Logarithmus wird zu einem $1/\sigma$-Verhalten

$$p_\sigma(\sigma)\, d\sigma = \frac{d\sigma}{\sigma} \qquad .$$

Diese Wahrscheinlichkeitsdichte ist ebenfalls uneigentlich, aber sie besitzt die wichtige Eigenschaft, dass sie SKALEN-INVARIANT ist. Diese Verteilung ist bis auf den unbestimmten Vorfaktor invariant gegen die Transformationen

$$\begin{aligned} \sigma &\to \alpha\, \sigma \\ \sigma &\to \sigma^n \end{aligned} \qquad ,$$

da diese Transformationen im Logarithmus einer Verschiebung oder Streckung entsprechen. Es wird sich zeigen, dass diese sogenannte JEFFREYS-VERTEILUNG die korrekte Prior-Wahrscheinlichkeit zur Beschreibung von SKALEN-VARIABLEN ist.

9.5.2 Beispiel mit zwei Variablen

Bei einem zweidimensionalen Problem liege eine Wahrscheinlichkeitsdichte $p_{xy}(x, y)$ in kartesischen Koordinaten x, y vor
Der Übergang auf Kreiskoordinaten lautet

$$\begin{aligned} x &= r\ \cos(\varphi) \\ y &= r\ \sin(\varphi) \end{aligned} \qquad .$$

Die Jakobi-Determinante ist demnach

$$\left|\frac{\partial\{x, y\}}{\partial\{r, \varphi\}}\right| = \left|\begin{pmatrix} \frac{dx}{dr} & \frac{dx}{d\varphi} \\ \frac{dy}{dr} & \frac{dy}{d\varphi} \end{pmatrix}\right| = \left|\begin{pmatrix} \cos(\varphi) & -r\ \sin(\varphi) \\ \sin(\varphi) & r\ \cos(\varphi) \end{pmatrix}\right| = r$$

Das heißt

$$p_{r\varphi}(r, \varphi) = p_{xy}(r\cos(\varphi), r\sin(\varphi))\ \cdot\ r \qquad .$$

9.6 Aufenthaltswahrscheinlichkeit des harmonischen Oszillators

Wir betrachten hier ein Beispiel, bei dem die Transformation nicht ein-eindeutig ist und die obige Formel nicht direkt angewendet werden kann. Es soll die Wahrscheinlichkeitsdichte untersucht werden, ein Teilchen, das sich in einem harmonischen Potential bewegt, am Ort x anzutreffen. Das Koordinatensystem liege so, dass die Gleichgewichtslage des harmonischen Oszillators bei $x = 0$ ist. Die Auslenkung ist bekanntlich

$$x(t) = A\ \cos(\omega t + \varphi) \qquad , \tag{9.35}$$

wobei A die Amplitude, ω die Frequenz und φ die Phase zur Zeit $t = 0$ darstellt. Uns interessiert die Wahrscheinlichkeitsdichte $p(x|A, \omega, \varphi, \mathcal{B})$, das Teilchen bei x anzutreffen. Die Amplitude, die Frequenz und die Phase φ seien bekannt. Wenn wir ohne Kenntnis des Zeitpunktes wiederholt den Ort messen, erhalten wir die gesuchte Wahrscheinlichkeitsdichte. Wenn zusätzlich der Zeitpunkt bekannt wäre, hätten wir

$$p(x|t, \varphi, A, \omega, \mathcal{B}) = \delta(x - A\cos(\omega t + \varphi)) \qquad .$$

Mit Hilfe der Marginalisierungsregel schreiben wir

$$p(x|A, \omega, \varphi, \mathcal{B}) = \int\ p(x|t, A, \omega, \varphi, \mathcal{B})\ p(t|A, \omega, \varphi, \mathcal{B})\ dt \tag{9.36}$$

$$= \int\ \delta(x - A\cos(\omega t + \varphi))\ p(t|A, \omega, \varphi, \mathcal{B})\ dt \qquad . \tag{9.37}$$

Die Information über A, ω, φ sagt nichts über den Zeitpunkt der Messung aus.

$$p(t|A, \omega, \varphi, \mathcal{B}) = p(t|\mathcal{B}) \qquad .$$

Wir transformieren nun die Zufalls-Variable t in

$$\xi = A\cos(\omega t + \varphi)$$

um. Diese Abbildung ist eindeutig, aber die Umkehr-Abbildung ist es nicht. Wir können die Jakobi-Determinante nicht ohne weiteres anwenden. Stattdessen können wir aber die δ-Funktion transformieren. Es gilt bekanntlich

$$\delta(\Psi(t)) = \sum_{l=1}^{L} \frac{\delta(t - t_l)}{\left|\frac{d\Psi(t)}{dt}\right|_{t=t_l}} \qquad ,$$

wobei $t_l, \quad (l = 1, \ldots, L)$ die Nullstellen von $\Psi(t)$ sind. Die Nullstellen erfüllen im vorliegenden Fall

$$\cos(\omega t_l + \varphi) = \frac{x}{A} \tag{9.38}$$

$$\text{bzw.} \qquad t_l^{\pm} = (\pm \arccos(x/A) - \varphi + l\ 2\pi)\ /\omega, \quad l \in \mathbb{Z} \qquad . \tag{9.39}$$

Hierbei ist $\arccos(x/A)$ auf das Intervall $(0, \pi]$ beschränkt. Die Ableitung von $\Psi(t) = x - A\cos(\omega t + \varphi)$ liefert

$$\left|\frac{d\Psi(t)}{dt}\right| = A\,\omega\;\;|\sin(\omega t + \varphi)| = A\,\omega\;\sqrt{1 - \cos^2(\omega t + \varphi)} \qquad .$$

Die Ableitung an der Stelle t_i liefert wegen Gl. (9.38)

$$\left|\frac{d\Psi(t)}{dt}\right|_{t=t_l} = A\,\omega\,\sqrt{1 - \left(\frac{x}{A}\right)^2} \qquad .$$

Damit ist die gesuchte Wahrscheinlichkeit

$$p(x|A,\omega,\varphi,\mathcal{B}) = \theta(|x| \le A)\;\frac{1}{\omega\sqrt{A^2 - x^2}}\sum_l\;\left(p(t_l^+|\mathcal{B}) + p(t_l^-|\mathcal{B})\right) \qquad . \tag{9.40}$$

Die Theta-Funktion resultiert daher, dass die Nullstellen in Gl. (9.38) nur existieren, wenn $|x| \le A$. Über den Zeitpunkt wissen wir nichts, deshalb müssen wir eine konstante Wahrscheinlichkeitsdichte $p(t|\mathcal{B}) = c$ ansetzen. Dieser UNEIGENTLICHE PRIOR hat den Nachteil, dass er nicht normierbar ist. Wir sehen die daraus resultierende Problematik sofort. Wenn wir ein endliches c ansetzen, liefert das Normierungs-Integral $Unendlich$. Setzen wir $c = 0$ so ist auch die Normierung Null. Außerdem liefert die Summe in Gl. (9.40) $2 * \infty * c$. Dieses triviale Beispiel ist sehr gut geeignet, die korrekte Vorgehensweise zu demonstrieren. Wir sollten immer einen flachen Prior als Grenzwert eines eigentlichen Priors ansetzen und den Grenzübergang ganz am Ende der Rechnung durchführen. Wenn die gesuchte Wahrscheinlichkeitsdichte existiert und normierbar ist, ist alles in Ordnung. Wenn nicht, ist das Ergebnis so stark vom Prior abhängig, dass auch das Ergebnis eine uneigentliche Wahrscheinlichkeitsdichte liefert. Wir setzen hier

$$p(t|\mathcal{B}) = p(t|\sigma,\mathcal{B}) \qquad .$$

Der Parameter σ^2 gibt die Varianz der Verteilung an. Am Ende lassen wir σ gegen Unendlich gehen. Die Nullstelle $t_i^{\pm}$ liegen gemäß Gl. (9.39) äquidistant im Abstand $\Delta t = 2\pi/\omega$. Wir können die Summe in Gl. (9.40) auch schreiben als

$$p(x|A,\omega,\varphi,\sigma,\mathcal{B}) = \theta(|x| \le A)\;\frac{1}{2\pi\sqrt{A^2 - x^2}}\sum_l\;\left(p(t_l^+|\sigma,\mathcal{B}) + p(t_l^-|\sigma,\mathcal{B})\right)\;\frac{2\pi}{\omega} \qquad .$$

Für $\sigma \to \infty$ werden die Variationen von $p(t_i^{\pm}|\sigma,\mathcal{B})$ und somit auch die Ableitungen immer kleiner, und die Summe geht in ein Integral über

$$\begin{aligned}\lim_{\sigma\to\infty} p(x|A,\omega,\varphi,\sigma,\mathcal{B}) &= \theta(|x| \le A)\;\frac{1}{2\pi\sqrt{A^2 - x^2}}\;2\;\underbrace{\int_{-\infty}^{\infty} p(t|\sigma,\mathcal{B})\;dt}_{=1} \\ &= \theta(|x| \le A)\;\frac{1}{\pi\sqrt{A^2 - x^2}} \qquad .\end{aligned}$$

Das gesuchte Ergebnis ist somit

AUFENTHALTS-WAHRSCHEINLICHKEIT DES HARMONISCHEN OSZILLATORS

$$p(x|A,\omega,\varphi,\mathcal{B}) = \theta(|x| \leq A)\ \frac{1}{\pi\sqrt{A^2-x^2}} \qquad . \tag{9.41}$$

Wie zu erwarten war, hängt das Ergebnis nicht von der Frequenz und von der Phase φ ab. Ein anderes Ergebnis wäre im Widerspruch dazu gewesen, dass wir den Prior für die Zeit konstant gewählt haben.
Das Ergebnis kann man mit physikalischen Argumenten wesentlich kürzer herleiten. Allerdings werden wir auf die ausführliche Ableitung im Zusammenhang mit Laser-Speckle-Phänomenen noch zurückgreifen. Die Rechnung hat auch wesentlichen Elemente einer Rechnung aufgezeigt, die bei komplexen Datenanalyse-Problemen auftreten.
Bei der physikalisch begründeten Herleitung geht man davon aus, dass die gesuchte Wahrscheinlichkeit proportional zur Verweildauer Δt des Teilchens im betrachteten Intervall $(x, x+\Delta x)$ ist

$$P(x' \in (x, x+\Delta x)) \propto \Delta t = \frac{\Delta x}{|v(x)|} \qquad .$$

Aus der Oszillatorbewegung folgt

$$\begin{aligned} x(t) &= A \cdot \cos(\omega t + \varphi) \\ |v(x)| = |\dot{x}| &= |\omega \cdot A| \cdot |\sin(\omega t + \varphi)| = |\omega A|\sqrt{1-\cos^2(\omega t+\varphi)} \\ &= \omega A \sqrt{1-\left(\frac{x}{A}\right)^2} \qquad . \end{aligned}$$

Nach der Normierung auf 1 erhalten wir

$$P(x' \in (x, x+dx)) = \frac{1}{\pi A}\frac{1}{\sqrt{1-(\frac{x}{A})^2}}dx \qquad .$$

Kapitel 10

Der zentrale Grenzwertsatz

Wir gehen von i.u.v. Zufalls-Variablen $x_n, n = 1, \ldots, N$ einer Wahrscheinlichkeitsdichte $\rho(x)$ aus. Mittelwert und Varianz dieser Verteilung seien

$$\mu := \langle x \rangle = \int x \, \rho(x) \, dx$$
$$\mathrm{var}(x) = \int (\Delta x)^2 \, \rho(x) \, dx \qquad .$$

Wir interessieren uns für die Summe

$$S = \sum_{i=1}^{N} c_i \, x_i \qquad , \tag{10.1}$$

wobei die c_i fest vorgegebene Gewichte sind, mit denen die Zufallszahlen x_i in die Summe eingehen. Ziel dieses Kapitels ist es, die Wahrscheinlichkeitsdichte $p(S|N, \rho, \mathcal{B})$ zu bestimmen. Es bietet sich an, in diesem Zusammenhang die Fouriertransformation einer Wahrscheinlichkeitsdichte (CHARAKTERISTISCHE FUNKTION) einzuführen. Die charakteristische Funktion spielt in vielen Anwendungsgebieten eine wichtige Rolle.

10.1 Charakteristische Funktion

10.1.1 Alternative Beschreibung einer Zufalls-Variablen

Def. 10.1 (Charakteristische Funktion) *Wir wollen die charakteristische Funktion gleich für m-dimensionale Zufalls-Variablen $x \in \mathbb{R}^m$ formulieren.*

CHARAKTERISTISCHE FUNKTION

$$\Phi(\omega) = \int\limits_{\mathbb{R}^m} d^m x \; \rho(x) \; e^{i\omega^T x} \; , \quad x, \omega \in \mathbb{R}^m \qquad . \tag{10.2}$$

Das ^{T} bezeichnet die Transponierte. Das innere Produkt zweier Spatenvektoren ω und x ist somit $\omega^T x$. Man kann die charakteristische Funktion auch als Erwartungswert

$$\Phi(\omega) = \langle e^{i\omega^T x} \rangle \tag{10.3}$$

der Zufallsvariablen $e^{i\omega^T x}$ betrachten.

Die charakteristische Funktion enthält dieselbe Information wie die Wahrscheinlichkeitsdichte, da aus ihr die Wahrscheinlichkeitsdichte über eine inverse Fourier-Transformation berechnet werden kann

$$p(x) = \left(\frac{1}{2\pi} \right)^m \int\limits_{\mathbb{R}^m} d^m \omega \; \Phi(\omega) \; e^{-i \, \omega^T \, x} \qquad . \tag{10.4}$$

10.1.2 Das Shift-Theorem

Angenommen, wir nehmen an der Zufalls-Variablen eine lineare Transformation vor

$$y = Ax + b \qquad ,$$

wobei A eine $(m \times m)$-Matrix und $y, b \in \mathbb{R}^m$ ist. Die charakteristische Funktion hierzu ist

$$\Phi_y(\omega) = \langle e^{i\omega^T y} \rangle_y = \langle e^{i\omega^T (Ax+b)} \rangle_y = e^{i\omega^T b} \; \langle e^{i(A^T \omega)^T x} \rangle_x$$

Das heißt

$$\Phi_y(\omega) = e^{i\omega^T b} \; \Phi_x(A^T \omega) \quad \text{für } y = Ax + b \qquad . \tag{10.5}$$

10.1.3 Erzeugung von Momenten

Die charakteristische Funktion ist sehr nützlich als erzeugendes Funktional der Momente der Zufalls-Variablen x. Wir beschränken uns auf die wichtigsten Spezialfälle

A) Eindimensionale Zufalls-Variablen $x \in \mathbb{R}$

Die n-te Ableitung der charakteristischen Funktion nach ω liefert

$$\Phi^{(n)}(\omega) = i^n \int\limits_{-\infty}^{\infty} dx\ x^n\ \rho(x)\ e^{i\,\omega\,x} \qquad .$$

Daraus folgt

$$\langle x^n \rangle = i^{-n}\ \Phi^{(n)}(0) \qquad . \tag{10.6}$$

Insbesondere gilt

$$1 = \Phi(0) \tag{10.7a}$$
$$\langle x \rangle = -i\ \Phi^{(1)}(0) \tag{10.7b}$$
$$\left\langle x^2 \right\rangle = -\ \Phi^{(2)}(0) \tag{10.7c}$$

Man kann $\Phi(0)$ also auch verwenden, um die Normierung zu bestimmen.

Beispiel: Γ-Verteilung

Die charakteristische Funktion der Γ-Verteilung lautet

$$\begin{aligned}
\Phi(\omega) &= \int\limits_{-\infty}^{\infty} dx\ p_\Gamma(x|\alpha,\beta)\ e^{i\,\omega\,x} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int\limits_{0}^{\infty} dx\ x^{\alpha-1}\ e^{-(\beta-i\omega)\,x} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\ (\beta - i\omega)^{-(\alpha)} \int\limits_{0}^{(\beta-i\omega)\cdot\infty} dx\ x^{\alpha-1}\ e^{-x} \qquad .
\end{aligned}$$

Innerhalb des Kreissektors, der durch die Wegstücke $C_1 = (0 \to (\beta - i\omega)\infty)$, $C_2 = ((\beta - i\omega) \cdot \infty \to \infty)$ und $C_3 = (\infty \to 0)$ aufgespannt wird, liegen keine Pole des Integranden, wenn $\alpha \geq 1$. Demnach gilt

$$\int\limits_{C_1+C_2+C_3} dx\ x^{\alpha-1}\ e^{-x} = 0 \qquad .$$

Außerdem verschwindet der Integrand auf C_2, so dass

$$\int_0^{(\beta-i\omega)\cdot\infty} dx\; x^{\alpha-1}\; e^{-x} = \int_0^{\infty} dx\; x^{\alpha-1}\; e^{-x} = \Gamma(\alpha) \qquad .$$

Damit lautet die

CHARAKTERISTISCHE FUNKTION DER Γ-VERTEILUNG

$$\Phi(\omega) = \left(\frac{\beta}{\beta - i\,\omega}\right)^{\alpha} = \left(1 - i\,\frac{\omega}{\beta}\right)^{-\alpha} \qquad . \tag{10.8}$$

Es sei daran erinnert, dass die Exponential-Funktion einen Spezialfall der Γ-Verteilung darstellt. Offensichtlich ist die Normierung korrekt $\Phi(0) = 1$. Die ersten beiden Ableitungen sind

$$\begin{aligned}
\Phi^{(1)}(\omega) &= (-\alpha)\; \left(1 - i\,\frac{\omega}{\beta}\right)^{-\alpha-1}\; (-i\frac{1}{\beta}) \\
&= i\,\frac{\alpha}{\beta}\; \left(1 - i\,\frac{\omega}{\beta}\right)^{-\alpha-1} \\
\Phi^{(2)}(\omega) &= i\,\frac{\alpha}{\beta}\,(-\alpha-1)\; \left(1 - i\,\frac{\omega}{\beta}\right)^{-\alpha-2}\; (-i\frac{1}{\beta}) \\
&= -\,\frac{\alpha(\alpha+1)}{\beta^2}\; \left(1 - i\,\frac{\omega}{\beta}\right)^{-\alpha-2} \qquad .
\end{aligned}$$

Mit Gl. (10.7b) und Gl. (10.7c) erhalten wir die korrekten Ergebnisse für Mittelwert und Varianz, in Übereinstimmung mit Gl. (9.11c) und Gl. (9.11d).

Beispiel: Normal-Verteilung

Die charakteristische Funktion der Normal-Verteilung berechnet sich aus

$$\begin{aligned}
\Phi(\omega) &= \int_{-\infty}^{\infty} dx\; \mathcal{N}(x|x_0,\sigma)\; e^{i\,\omega\,x} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} dx\; e^{-\frac{1}{2\sigma^2}((x-x_0)^2 - 2\,i\,\sigma^2\omega\,x)} \qquad .
\end{aligned}$$

Das Argument der Exponential-Funktion lässt sich quadratisch ergänzen

$$\begin{aligned}(x-x_0)^2 - 2\,i\,\sigma^2\omega\,x &= x^2 - 2\,x\,(x_0 + i\,\sigma^2\,\omega) + x_0^2 \\ &= (x - (x_0 + i\,\sigma^2\,\omega))^2 + (\sigma^2\,\omega)^2 - 2\,i\,\sigma^2\,\omega\,x_0 \qquad .\end{aligned}$$

Wir führen eine neue Integrationsvariable $z = x - (x_0 + i\,\sigma^2\,\omega)$ ein. Der Integrationsweg wird dann zu $C = (-\infty - i\,\sigma^2\,\omega \to \infty - i\,\sigma^2\,\omega)$

$$\begin{aligned}\Phi(\omega) &= \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-((\sigma^2\,\omega)^2 - 2\,i\,\sigma^2\,\omega)/2\sigma^2} \underbrace{\int\limits_C dx\; e^{-\frac{z^2}{2\sigma^2}}}_{\sqrt{2\pi\sigma^2}} \\ &= e^{-\frac{1}{2}\,\sigma^2\,\omega^2 + i\,\omega\,x_0} \qquad .\end{aligned}$$

CHARAKTERISTISCHE FUNKTION DER NORMAL-VERTEILUNG
$$\Phi(\omega) = e^{i\omega x_0}\; e^{-\frac{1}{2}\,\sigma^2\,\omega^2} \qquad . \tag{10.9}$$

Auch hierbei ist die Normierung korrekt $\Phi(0) = 1$. Nach dem Shift-Theorem ist die charakteristische Funktion der Zufalls-Variablen $z = x - x_0$ gegeben durch

$$\Phi_z(\omega) = \Phi_x(\omega)\; e^{-i\omega x_0} = e^{-\frac{1}{2}\,\sigma^2\,\omega^2} \qquad .$$

Die Momente dieser verschobenen Zufalls-Variablen sind gleichzeitig die zentrierten Momente

$$\langle (x - x_0)^n \rangle \qquad .$$

Die n-te Ableitung von $\Phi_z(\omega)$ lautet

$$\Phi_z^{(n)}(0) = \begin{cases} (-1)^{\frac{n}{2}}\;\sigma^n\;(n-1)!! & \text{für gerades } n, \\ 0 & \text{sonst.} \end{cases}$$

Gemäß Gl. (10.6) gilt dann

ZENTRALE MOMENTE DER NORMAL-VERTEILUNG
$$\langle (x - x_0)^n \rangle = \begin{cases} \sigma^n\;(n-1)!! & n \text{ gerade}, \\ 0 & \text{sonst.} \end{cases} \tag{10.10}$$

Man kann die charakteristische Funktion auch für diskrete Probleme berechnen. Die Wahrscheinlichkeitsdichte einer diskreten Verteilung P_n lautet

$$p(x) = \sum_n \; P_n \; \delta(x - x_n) \qquad .$$

Die charakteristische Funktion ergibt dann

$$\Phi(\omega) = \int \; dx \; p(x) \; e^{i\omega x} = \sum_n \; P_n \; \left(\int \; dx \; \delta(x - x_n) \; e^{i\omega x} \right) = \sum_n \; P_n \; e^{i\omega x_n} \quad .$$

CHARAKTERISTISCHE FUNKTION DISKRETER PROBLEME

$$\Phi(\omega) = \sum_n \; P_n \; e^{i\omega x_n} \qquad . \tag{10.11}$$

Beispiel: Poisson-Verteilung

Im Fall der Poisson-Verteilung erhalten wir

$$\begin{aligned} \Phi(\omega) &= e^{-\mu} \; \sum_{n=0}^{\infty} \; \frac{\mu^n}{n!} \; e^{i\omega n} = e^{-\mu} \; \sum_{n=0}^{\infty} \; \frac{\left(e^{i\omega}\mu\right)^n}{n!} = e^{-\mu} \; e^{\mu \, e^{i\omega}} \\ &= e^{-\mu \, (1 - e^{i\omega})} \qquad . \end{aligned}$$

Wir haben hier den sehr seltenen Fall der Funktion $\exp(\exp(.))$ vorliegen. Die Normierung ist korrekt. Die ersten beiden Ableitungen lauten

$$\begin{aligned} \Phi^{(1)}(\omega) &= e^{-\mu \, (1 - e^{i\omega})} \; \left(\mu \; e^{i\omega} \right) \; (i) && = i\mu \; e^{i\omega} \; e^{-\mu \, (1 - e^{i\omega})} \\ \Phi^{(2)}(\omega) &= i\mu \; \left(i\mu e^{i\omega} \; e^{i\omega} + i \; e^{i\omega} \right) \; e^{-\mu \, (1 - e^{i\omega})} && = -\mu \; \left(\mu e^{i2\omega} + \; e^{i\omega} \right) \; e^{-\mu \, (1 - e^{i\omega})} \; . \end{aligned}$$

An der Stelle $\omega = 0$ liefern die Ableitungen

$$\begin{aligned} \Phi^{(1)}(0) &= i\mu \\ \Phi^{(2)}(0) &= -\mu \; (\mu + 1) \qquad . \end{aligned}$$

Das Ergebnis ist wieder in Übereinstimmung mit dem, das wir früher abgeleitet hatten.

Beispiel: Binomial-Verteilung

Die charakteristische Funktion der Binomial-Verteilung lässt sich ebenfalls elementar berechnen.

$$\Phi(\omega) = \sum_{n=0}^{\infty} \binom{N}{n} p^n \, q^{(N-n)} \, e^{i\omega n} = \sum_{n=0}^{N} \binom{N}{n} (p \, e^{i\omega})^n \, q^{(N-n)}$$
$$= (p \, e^{i\omega} + q)^N \qquad .$$

Auch in diesem Fall kann leicht gezeigt werden, dass Mittelwert und Varianz mit den alten Ergebnissen übereinstimmen.

B) Multivariater Fall $x \in \mathbb{R}^m$

Es soll hier zumindest angegeben werden, dass die charakteristische Funktion auch im multivariaten Fall sehr hilfreich sein kann, ohne dass wir sie später wirklich verwenden werden. Der Mittelwert der i-ten Komponente folgt wieder aus

$$\langle x_i \rangle = -i \, \frac{\partial}{\partial \omega_i} \Phi(\omega) \Big|_{\omega=0} \qquad .$$

Die Kovarianz-Matrix erhalten wir analog aus

$$\langle \Delta x_i \Delta x_j \rangle = - \frac{\partial^2}{\partial \omega_i \partial \omega_j} \Phi(\omega) \Big|_{\omega=0} - \langle x_i \rangle \langle x_j \rangle \qquad .$$

Weitere Details hierzu können in den Büchern von R. Frieden oder A. Papoulis nachgelesen werden.
Neben der charakteristischen Funktion ist auch

$$p(x) = \sum_n p_n \, x^n \tag{10.12}$$

ein nützliches und weit verbreitetes erzeugendes Funktional, das wir im Zusammenhang mit Poisson-Prozessen noch besprechen werden (siehe Kapitel 12).

10.2 Summe von Zufalls-Variablen

Besonders nützlich ist die charakteristische Funktion bei der Behandlung von Summen von unabhängigen Zufalls-Variablen

$$S = \sum_{n=1}^{N} x_n \qquad . \tag{10.13}$$

Die Zufalls-Variablen x_n haben die Dichten $p_n(x)$ und die charakteristischen Funktionen $\Phi_n(\omega)$. Die Wahrscheinlichkeitsdichte von S ist

$$\begin{aligned} p(S|N,\mathcal{B}) &= \int d^N x \; p(S|x_1,\ldots,x_N,N,\mathcal{B}) \; p(x_1,\ldots,x_N|N,\mathcal{B}) \\ &= \int d^N x \; \delta(S - \sum_n x_n) \prod_n p_n(x_n) \qquad . \end{aligned}$$

Die zugehörige charakteristische Funktion lautet

$$\begin{aligned} \Phi(\omega) &= \int dS \; p(S|N,\mathcal{B}) \; e^{iS\omega} \\ &= \int d^N x \; \left(\int dS \; \delta(S - \sum_n x_n) \; e^{iS\omega} \right) \prod_n p_n(x_n) \\ &= \int d^N x \;\; e^{i\sum_n x_n \omega} \prod_n p_n(x_n) = \prod_n \int dx_n \;\; e^{i\, x_n \omega} \; p_n(x_n) \\ &= \prod_n \Phi_n(\omega) \qquad . \end{aligned}$$

CHARAKTERISTISCHE FUNKTION
EINER SUMME VON ZUFALLS-VARIABLEN

$$\begin{aligned} S &= \sum_{n=1}^{N} x_n \\ \Phi(\omega) &= \prod_{n=1}^{N} \Phi_n(\omega) \qquad . \end{aligned} \tag{10.14}$$

Die charakteristische Funktion der Zufalls-Variablen S ist also das Produkt der charakteristischen Funktionen der beteiligten Zufalls-Variablen. Die Wahrscheinlichkeitsdichte von S erhalten wir aus der inversen Fourier-Transformation. Bekanntlich ist die inverse FT eines Produktes die FALTUNG.
Wie verschiebt sich der Mittelwert der Summe, wenn der der individuellen Verteilungen um x_0 verschoben wurde, so dass $\langle x \rangle = 0$? Wenn wir die Zufalls-Variablen zum Schwerpunkt verschieben, ändert sich die charakteristischen Funktion nach dem Shift-Theorem zu

$$\Phi_n(\omega) = \Phi_n^0(\omega) \; e^{i\omega x_n^0} \qquad .$$

Der obere Index 0 soll andeuten, dass die charakteristischen Funktion zur Wahrscheinlichkeitsdichte mit Mittelwert Null gehört. Für die charakteristischen Funktion

von S bedeutet das

$$\Phi_S(\omega) = e^{i\sum_n \omega x_n^0} \prod_{n=1}^{N} \Phi_n^0(\omega).$$

Wir kommen nun zum zentralen Grenzwertsatz zurück. Hierbei handelt es sich um die Summe

$$S = \sum_{n=1}^{N} c_n \, x_n$$

von unabhängigen Zufalls-Variablen. Da die x_n unabhängige Zufalls-Variablen sind, sind auch die Größen $c_n x_n$ mit festen Koeffizienten c_n unabhängige Zufalls-Variablen. Mittelwert und Varianz der Zufalls-Variablen S kennen wir bereits aus dem Abschnitt über Fehler-Fortpflanzung

$$\begin{aligned} \langle S \rangle &= \mu \sum_n c_n \\ \mathrm{var}(S) &= \sigma_x^2 \sum_n c_n^2 \qquad . \end{aligned}$$

Nach Gl. (10.14) hat die Summe demnach die charakteristische Funktion

$$\Phi_S(\omega) = \prod_n \Phi_n(\omega) \qquad .$$

Die Zufalls-Variablen x_n sind i.u.v. gemäß der Dichte $p(x)$. Die charakteristischen Funktion zu dieser Dichte sei $\Phi_x(\omega)$ mit Mittelwert μ. Die neuen Zufalls-Variablen $c_n x_n$ haben Mittelwert $c_n \mu$. Sie gehen aus den zentrierten i.u.v. Größen x_n^0 hervor über

$$c_n x_n = c_n x_n^0 + c_n \mu \qquad .$$

Nach dem Shift-Theorem gilt für die charakteristische Funktion der Zufalls-Variablen $c_n x_n$

$$\Phi_x(c_n \omega) = e^{i\omega c_n \mu} \, \Phi_x^0(c_n \omega) \qquad .$$

Damit wird aus der charakteristischen Funktion von S

$$\Phi_S(\omega) = e^{i\omega \, \mu \, \sum_n \, c_n} \prod_n \Phi_x^0(c_n \omega) \qquad .$$

Wir erkennen in der Phase den Mittelwert $\langle S \rangle$ wieder

$$\Phi_S(\omega) = e^{i\omega \langle S \rangle} \prod_n \Phi_x^0(c_n \omega) \qquad . \tag{10.15}$$

Wir können auch ausnutzen, dass die Standardabweichung $\sqrt{\mathrm{var}(S)}$ eine typische Skala des Problems definiert. Für die folgenden Überlegungen ist es sinnvoll, zu skalierten Größen

$$\tilde{S} = S/\sqrt{\mathrm{var}(S)}$$

überzugehen. Für die charakteristischen Funktion dieser Größen gilt

$$\Phi_{\tilde{S}}(\omega) = \Phi_S(\underbrace{\omega/\sqrt{\mathrm{var}(S)}}_{\tilde{\omega}}) = \Phi_S(\tilde{\omega}) \tag{10.16}$$

Wir werden nun das Produkt weiterverarbeiten. Der Logarithmus hiervon ist

$$\sum_n \ln(\Phi_x^0(c_n\tilde{\omega})) \qquad . \tag{10.17}$$

Als nächstes entwickeln wir $\Phi_x^0(c_n\tilde{\omega})$ um den Punkt $\tilde{\omega} = 0$

$$\Phi_x^0(c_n\tilde{\omega}) = \Phi_x^0(0) + \Phi_x^{0\,(1)}(0)\; c_n\tilde{\omega} + \tfrac{1}{2}\Phi_x^{0\,(2)}(0)\; (c_n\tilde{\omega})^2 + \tfrac{1}{6}\Phi_x^{0\,(3)}(0)\; (c_n\tilde{\omega})^3 + \ldots \qquad .$$

Nun ist $\Phi_x^0(0) = 1$ die Normierung. Der Mittelwert ist proportional zur ersten Ableitung. Da der Mittelwert Null ist, verschwindet $\Phi_x^{0\,(1)}(0) = 0$. Die zweite Ableitung liefert, da der Mittelwert Null ist, bis auf das Vorzeichen die Varianz

$$\Phi_x^{0\,(2)}(0) = -\sigma_x^2 \qquad .$$

Für das dritte Moment führen wir eine Abkürzung ein

$$\tfrac{1}{6}\Phi_x^{0\,(3)}(0) =: \kappa \qquad .$$

Damit haben wir

$$\Phi_x^0(c_n\tilde{\omega}) = 1 - \tfrac{1}{2}\sigma_x^2\; (c_n\tilde{\omega})^2 + \kappa\; (c_n\tilde{\omega})^3 + \ldots \qquad ,$$

und Gl. (10.17) wird zu

$$\begin{aligned}\sum_n \ln(\Phi_x^0(c_n\tilde{\omega})) &= \sum_n \ln\left(1 - \tfrac{1}{2}\; c_n^2\; \sigma_x^2\; \tilde{\omega}^2 + \; c_n^3\; \kappa\; \tilde{\omega}^3 + \ldots\right) \\ &= \sum_n \ln\left(1 - \tfrac{1}{2}\; \sigma_x^2\; \tilde{\omega}^2\; c_n^2 \left(1 - c_n\; \frac{2\kappa}{\sigma_x^2}\; \tilde{\omega}\right) + \ldots\right) \qquad .\end{aligned}$$

Wir betrachten nur solche Werte von $\tilde{\omega}$, für die die Terme der Reihenentwicklung sehr viel kleiner als Eins sind, damit die Reihenentwicklung gerechtfertigt ist. Dann können wir auch den Logarithmus entwickeln

$$\begin{aligned}\sum_n \ln(\Phi_x^0(c_n\tilde{\omega})) = &-\tfrac{1}{2}\; \sigma_x^2\; \tilde{\omega}^2 \sum_n c_n^2 \left(1 - c_n\; \frac{2\kappa}{\sigma_x^2}\; \tilde{\omega}\right) \\ &- \tfrac{1}{4}\; \sigma_x^4\; \tilde{\omega}^4 \sum_n c_n^4 \left(1 - c_n\; \frac{2\kappa}{\sigma_x^2}\; \tilde{\omega}\right)^2 + \ldots \qquad .\end{aligned} \tag{10.18}$$

Nun kommt der Grund, warum wir zu den skalierten Größen übergegangen sind, zum Tragen. Die modifizierte Frequenz ist

$$\begin{aligned}\tilde{\omega} &= \frac{\omega}{\sqrt{\mathrm{var}(S)}} \\ &= \frac{\omega}{\sigma_x \sqrt{\sum_n c_n^2}} \qquad .\end{aligned}$$

Von den Gewichtsfaktoren c_n verlangen wir, dass die Summen $\sum_{n=1}^{N} c_n^2$ und $\sum_{n=1}^{N} c_n^3$ für große N linear mit N anwachsen

$$\sum_{n=1}^{N} c_n^\nu = \alpha_\nu \, N \qquad ,$$

wobei die Größen α_ν nicht mehr von N abhängen. Damit haben wir

$$\tilde{\omega} = \frac{\omega}{\sigma_x \sqrt{N \, \alpha_2}} \qquad .$$

Dann liefern die einzelnen Terme

$$\begin{aligned}-\tfrac{1}{2} \, \sigma_x^2 \, \tilde{\omega}^2 \sum_n c_n^2 \left(1 - c_n \frac{2\kappa}{\sigma_x^2} \, \tilde{\omega}\right) &= -\tfrac{1}{2} \, \sigma_x^2 \, \tilde{\omega}^2 \, N\alpha_2 \left(1 - \frac{2\kappa\alpha_3}{\alpha_2\sigma_x^2} \, \tilde{\omega}\right) \\ &= -\tfrac{1}{2} \, \sigma_x^2 \, \frac{\omega^2 \, N\alpha_2}{\sigma_x^2 \, N \, \alpha_2} \left(1 - \frac{2\kappa\alpha_3}{\alpha_2\sigma_x^2} \, \frac{\omega}{\sigma_x \sqrt{N \, \alpha_2}}\right) \\ &= -\tfrac{1}{2} \, \omega^2 \left(1 - \frac{2\kappa\alpha_3}{\alpha_2\sigma_x^2} \, \frac{\omega}{\sigma_x \sqrt{N \, \alpha_2}}\right)\end{aligned}$$

Der zweite Term geht mit $1/\sqrt{N}$ gegen Null und kann deshalb vernachlässigt werden. Die nächsten Terme in Gl. (10.18) lauten

$$\begin{aligned}\tfrac{1}{4} \, \sigma_x^4 \, \tilde{\omega}^4 \sum_n c_n^4 \left(1 - c_n \frac{2\kappa}{\sigma_x^2} \, \tilde{\omega}\right)^2 &= \tfrac{1}{4} \, \sigma_x^4 \, \tilde{\omega}^4 \sum_n c_n^4 \; + \; \ldots \\ &= \tfrac{1}{4} \, \sigma_x^4 \, \tilde{\omega}^4 \, N \, \alpha_4 \; + \; \ldots \\ &= \tfrac{1}{4} \, \sigma_x^4 \, \frac{\omega^4}{\sigma_x^4 \, \alpha_2^2 \, N^2} \, N \, \alpha_4 \; + \; \ldots \\ &= O(1/N) \qquad .\end{aligned}$$

Damit ist der Logarithmus Gl. (10.17) für festes ω und großes N

$$\sum_n \ln(\Phi_x^0(c_n\tilde{\omega})) = -\tfrac{1}{2} \, \omega^2 + O(1/\sqrt{N}) \qquad .$$

Mit Gl. (10.15) und Gl. (10.16) erhalten wir somit

$$\Phi_{\tilde{S}}(\omega) = e^{i\frac{\omega}{\sqrt{\mathrm{var}(S)}}\langle S\rangle}\; e^{-\frac{1}{2}\omega^2} = \Phi_S(\frac{\omega}{\sqrt{\mathrm{var}(S)}}) \qquad ,$$

bzw.

$$\Phi_S(\omega') = e^{i\omega'\langle S\rangle}\; e^{-\frac{1}{2}\,\mathrm{var}(S)\,\omega'^2} \qquad .$$

Der Vergleich mit der charakteristischen Funktion der Normal-Verteilung (Gl. (10.9)) zeigt, dass S einer Normal-Verteilung mit Mittelwert $\langle S\rangle$ und Varianz $\mathrm{var}(S)$ genügt. Das bedeutet, die Summe von gewichteten Zufallszahlen ist für große N normal-verteilt

ZENTRALER GRENZWERT-SATZ

Es sei die Zufalls-Variable

$$S = \sum_{n=1}^{N} c_n\, x_n$$

definiert, wobei die x_n i.u.v. Zufallsgrößen mit Mittelwert μ und Varianz σ_x^2 sind. Die c_n sind feste Gewichtsfaktoren, mit der Eigenschaft

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} c_n^\nu = \alpha_\nu = konst., \qquad \nu \in \mathbb{Z}$$

Dann gilt

$$\lim_{N\to\infty} p(S|N,\mathcal{B}) = \mathcal{N}(S|\langle S\rangle, \mathrm{var}(S))$$

$$\langle S\rangle = \mu \sum_{n=1}^{N} c_n \tag{10.19}$$

$$\mathrm{var}(S) = \sigma_x^2 \sum_{n=1}^{N} c_n^2$$

10.2.1 Beispiel: Summe von exponentiell verteilten Zufallszahlen

Wir betrachten hier als Beispiel eine Stichprobe vom Umfang N, deren Elemente exponentiell verteilt sind. In Abbildung 10.1 ist die Wahrscheinlichkeitsdichte $p(S|N, \mathcal{B})$ für verschiedene N aufgetragen. Man erkennt, wie das arithmetische Mittel $S = \frac{1}{N} \sum_i^N x_i$ sich sehr schnell einer Gauß-Verteilung annähert. Wir können dieses

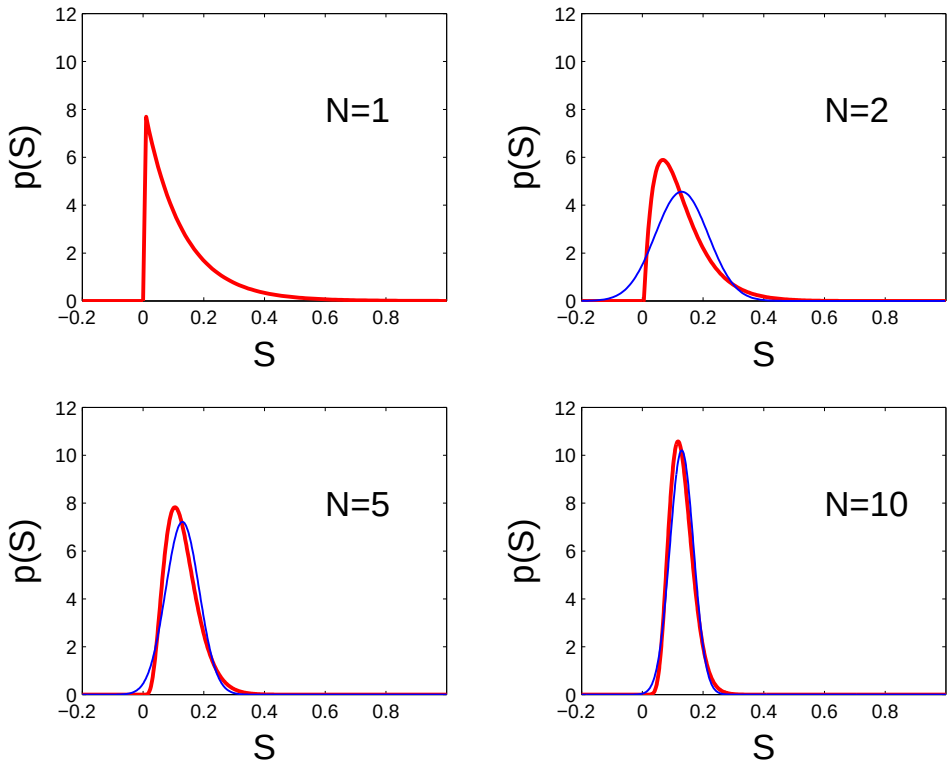

Abbildung 10.1: *Zentraler Grenzwert-Satz für* $S = \frac{1}{N} \sum_{i=1}^N x_i$ *für verschiedene* N *verglichen mit den entsprechenden Gauß-Funktionen. Alle* x_i *sind exponentiell verteilt. Man erkennt, dass bereits bei* $N = 5$ *eine gute Übereinstimmung mit der Gauß-Funktion besteht.*

Beispiel auch exakt durchrechnen. Die charakteristische Funktion der Exponential-Funktion

$$p(x) = \beta \, e^{-\beta x}$$

ist gemäß Gl. (10.8)

$$\Phi(\omega) = \frac{\beta}{\beta - i\omega} \qquad .$$

Die charakteristische Funktion von S ist dann

$$\Phi_S(\omega) = \Phi(\omega)^N = \left(\frac{\beta}{\beta - i\omega} \right)^N \qquad .$$

Das ist wiederum nach Gl. (10.8) die charakteristische Funktion der Γ-Verteilung. Das heißt

$$p(S|N,\mathcal{B}) = \Gamma(S|N,\beta) = \frac{\beta^N}{\Gamma(N)}\, S^{N-1}\, e^{-\beta S} \qquad . \tag{10.20}$$

Diese Funktion kann, wie wir wissen, durch eine Gauß-Funktion approximiert werden.

10.2.2 Beispiel: Summe gleichverteilter Zufallszahlen

Nun betrachten wir das arithmetische Mittel von gleichverteilten Zufallszahlen aus dem Intervall $[0,1]$. In Abbildung 10.2 ist das Mittel $S = \frac{1}{N}\sum_i^N x_i$ für verschiedene N dargestellt. Im Vergleich mit dem vorigen Beispiel (Abbildung 10.1) wird hier die Gauß-Verteilung noch schneller angenommen. Auch die hier auftretenden Wahr-

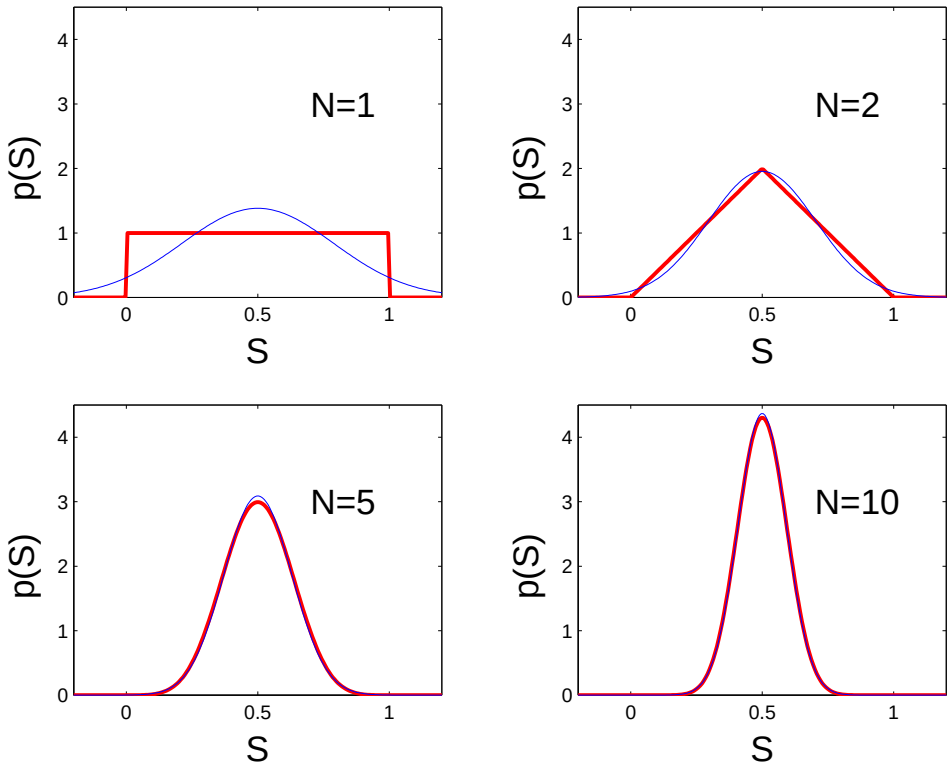

Abbildung 10.2: *Zentraler Grenzwert-Satz für* $S = \frac{1}{N}\sum_{i=1}^N x_i$ *für verschiedene* N *verglichen mit den entsprechenden Gauß-Funktionen. Alle* x_i *sind gleichverteilt. Man erkennt, dass bereits bei* $N = 2$ *eine gewisse Übereinstimmung mit der Gauß-Funktion besteht.*

scheinlichkeitsdichten kann man exakt berechnen. Als erstes bestimmen wir die charakteristische Funktion der Gleich-Verteilung

$$p(x) = \theta(0 \le x \le 1) \qquad .$$

Mit Gl. (10.2) erhalten wir

$$\Phi(\omega) = \int_0^1 e^{i\,\omega\,x}\,dx \;=\; \frac{e^{i\,\omega}-1}{i\,\omega} \qquad . \tag{10.21}$$

Die charakteristische Funktion der Summe $S = \sum_{i=1}^{N} x_i$ ist nach Gl. (10.14) die N-te Potenz von Gl. (10.21). Die Rücktransformation ergibt

$$\begin{aligned} p(S|N,\mathcal{B}) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\frac{e^{i\omega}-1}{i\omega}\right)^N e^{-i\omega S}\,d\omega \\ &= \frac{(-i)^N}{2\pi} \int_{-\infty}^{\infty} \omega^{-N} \left(\sum_{k=0}^{N} \binom{N}{k} e^{ik\omega}\,(-1)^{N-k}\right) e^{-i\omega S}\,d\omega \\ &= \frac{i^N}{2\pi} \sum_{k=0}^{N} \binom{N}{k} (-1)^k \underbrace{\int_{-\infty}^{\infty} \omega^{-N}\, e^{i\omega(k-S)}\,d\omega}_{=:I} \qquad . \end{aligned}$$

Das Integral I lässt sich mit komplexen Methoden berechnen. Wir erhalten

$$I = i^N\,\pi\,\frac{(k-S)^{N-1}}{(N-1)!}\,\mathrm{sign}(k-S) \qquad .$$

Wir können die Wahrscheinlichkeitsdichte für S als

$$p(S|N,\mathcal{B}) = \frac{(-1)^N}{2} \sum_{k=0}^{N} \binom{N}{k} (-1)^k\,\frac{(k-S)^{N-1}}{(N-1)!}\,\mathrm{sign}(k-S)$$

schreiben. Die Dichte $p(S|N,\mathcal{B})$ besteht also stückweise aus Polynomen. Für die ersten paar N erhält man durch Einsetzen

$$p(S|N=1,\mathcal{B}) = \begin{cases} 0 & \text{wenn } S \le 0 \\ 1 & \text{wenn } 0 < S \le 1 \\ 0 & \text{wenn } 1 < S \end{cases}$$

$$p(S|N=2,\mathcal{B}) = \begin{cases} 0 & \text{wenn } S \le 0 \\ S & \text{wenn } 0 < S \le 1 \\ -S+2 & \text{wenn } 1 < S \le 2 \\ 0 & \text{wenn } 2 < S \end{cases}$$

$$p(S|N=3,\mathcal{B}) = \begin{cases} 0 & \text{wenn } S \leq 0 \\ \frac{1}{2}S^2 & \text{wenn } 0 < S \leq 1 \\ -S^2 + 3S - \frac{3}{2} & \text{wenn } 1 < S \leq 2 \\ \frac{1}{2}S^2 - 3S + \frac{9}{2} & \text{wenn } 2 < S \leq 3 \\ 0 & \text{wenn } 3 < S \end{cases} \quad .$$

10.2.3 Monte-Carlo-Integration

Der Zentrale Grenzwertsatz bildet die Grundlage für eine der wichtigsten Methoden der Computersimulation, dem Monte-Carlo-Verfahren. Hiermit kann man eine Vielzahl von Problemen der klassischen Statistik und der Quantenmechanik numerisch simulieren. In der klassischen statistischen Physik ist man an thermodynamischen Erwartungswerten

$$\langle O \rangle_T = \int d^m x \; O(x) \; p(x) \tag{10.22}$$

interessiert. Falls diskreten Freiheitsgrade vorliegen, ist das Integral durch eine Summe zu ersetzen. Die Wahrscheinlichkeitsdichte $p(x)$ hängt nur von der Energie $E(x)$ ab und ist der normierte Boltzmann-Faktor

$$p(x) = \frac{1}{Z} \, e^{-E(x)/kT} \quad .$$

Die Problematik der Berechnung liegt darin, dass es sich um extrem hochdimensionale Integrale oder Summen handelt, die nur in den seltensten Fällen exakt berechnet werden können. Wir werden nun den Zentralen Grenzwertsatz etwas verallgemeinern. Wir gehen von m-dimensionalen Zufalls-Variablen $x \in \mathbb{R}^m$ aus und definieren zu jedem Vektor x eine ein-dimensionale Zufalls-Variable $O(x)$, die eine eindeutige Funktion von x darstellt. Die Wahrscheinlichkeitsdichte von O ist

$$p(O) = \int d^m x \; \delta(O - O(x)) \; p(x) \quad .$$

Mittelwert und Varianz von O lassen sich leicht angeben

$$\langle O \rangle = \int dO \; O \; p(O) = \int d^m x \; O(x) \; p(x) \tag{10.23a}$$

$$\mathrm{var}(O) = \int dO \; (O - \langle O \rangle)^2 \; p(O) = \int d^m x \; (O(x) - \langle O \rangle)^2 \; p(x) \quad . \tag{10.23b}$$

Das heißt, der Mittelwert von O ist genau das gesuchte Integral. Nun besagt der Zentrale Grenzwertsatz, dass das arithmetische Mittel

$$\mathcal{O} = \frac{1}{N} \sum_{n=1}^{N} O(x_n)$$

der Zufallszahlen $O(x_n)$ für hinreichend große N einer Normalverteilung genügt

$$p(\mathcal{O}) \underset{N\gg 1}{=} \mathcal{N}(\mathcal{O}|\langle O\rangle, \mathrm{var}(O)/N) \qquad .$$

Somit kann man umgekehrt hoch-dimensionale Integrale/Summen vom Typ

$$\langle O\rangle = \int O(x)\; p(x)\; dx$$

aus einer Stichprobe $\{x_1, \ldots, x_N\}$ von Zufallszahlen der Verteilung $p(x)$ durch das arithmetische Mittel approximieren, da ja gilt

$$\langle O\rangle = \frac{1}{N}\sum_{n=1}^{N} O(x_n) \pm \sqrt{\frac{\mathrm{var}(O)}{N}} \qquad .$$

Das arithmetische Mittel ist demnach Gauß-verteilt um den Mittelwert der Einzel-Beiträge mit einer Varianz, die um den Faktor N reduziert ist. Das erklärt, warum es Sinn macht, zur experimentellen Bestimmung einer Größe, die statistischen Schwankungen unterliegt, das Experiment oft zu wiederholen und daraus das arithmetische Mittel zu bilden. Die Ungenauigkeit des arithmetischen Mittels ist der

STANDARDFEHLER

$$\mathrm{SF} = \frac{\sigma}{\sqrt{N}} \qquad . \tag{10.24}$$

10.3 Zentraler Grenzwertsatz: multivariater Fall

ZENTRALER GRENZWERT-SATZ IN M DIMENSIONEN

Es sei die Zufalls-Variable $S \in \mathbb{R}^m$

$$S = \frac{1}{N} \sum_{n=1}^{N} x_n$$

definiert, wobei die x_n *i.u.v. Zufallsgrößen mit Mittelwert* $\mu \in \mathbb{R}^m$ *und Kovarianzmatrix* C *sind. Dann gilt*

$$\lim_{N \to \infty} p(S|N, \mathcal{B}) = \mathcal{N}(S|\mu, C/N) \tag{10.25}$$

Kapitel 11

Laser-Speckle

Als anspruchsvollere Anwendung des bisher Gelernten wollen wir Laser-Speckle untersuchen. Laser-Speckle kann man z.B. beobachten, wenn Laser-Licht von einer Wand reflektiert wird. Es bilden sich Erscheinungen, in denen das Laser-Licht in zufällige granulare Muster zerbrochen scheint. Es bietet eine interessante Anwendung des Zentralen Grenzwertsatzes und der Transformationstheorie von Zufalls-Variablen.
Wir gehen davon aus, dass kohärentes Licht auf einen sogenannten Diffuser fällt, der hierbei überall einheitlich und phasengleich ausgeleuchtet werden soll. Der Diffuser ist eine dünne transparente Schicht, deren Dicke zufällig variiert. Das Speckle-Muster ist in allen Ebenen hinter dem Diffuser vorhanden. Wir greifen eine Ebene „B" heraus und analysieren die komplexe Amplitude $A(x_B, y_B)$ des Lichts an einem Punkt $\vec{x}_B = (x_B, y_B)$ in dieser Ebene.
Wir sind insbesondere an der Intensität (I) und der Phase Φ der Speckle interessiert. Diese Größen erhalten wir aus der Amplitude gemäß

$$I = |A|^2 = A_{\mathrm{Re}}^2 + A_{\mathrm{Im}}^2 \tag{11.1}$$

$$\Phi = \arctan(\frac{A_{\mathrm{Im}}}{A_{\mathrm{Re}}}) \qquad . \tag{11.2}$$

Wir werden zunächst die gemeinsame Wahrscheinlichkeitsdichte $p(A_{\mathrm{Re}}, A_{\mathrm{Im}}|\mathcal{B})$ bestimmen und daraus die gesuchten Größen berechnen.

11.1 Das Statistische Modell

Um die Laser-Speckle zu analysieren, müssen wir Modell-Annahmen machen. Das ist bereits teilweise durch den Aufbau mit dem Diffuser geschehen. Der Diffuser habe am Ort $\vec{x}_D = (x_D, y_D)$ in der Diffuser-Ebene die Dicke $\Delta(x_D, y_D)$. Der Abstand R zwischen Diffuser und der Ebene B soll groß sein im Vergleich zur Diffuser-Dicke

$$R \gg \Delta(x_D, y_D), \qquad \forall (x_D, x_D) \qquad .$$

Von jedem Punkt $\vec{x}_D$ des Diffusers geht nach dem Huygensschen Prinzip eine Kugelwelle aus, die an einem beliebigen Punkt $\vec{x}_B$ der Bild-Ebene die Amplitude

$$\frac{e^{ik|\vec{x}_B - \vec{x}_D|}}{|\vec{x}_B - \vec{x}_D|}$$

beiträgt. Für die Wellenzahl gilt

$$k = \frac{2\pi}{\lambda} \qquad .$$

Amplituden aller Punkte des Diffuser tragen additiv zur Gesamtamplitude im Punkt $\vec{x}_B$ bei. Um die Analyse einfach zu halten, nehmen wir an, dass der Diffuser aus vielen (N) Bereichen („Scattering spots") der Größe Δa bestehen. Innerhalb jeden Spots ist die Dicke konstant und liefert perfekt korrelierte Beiträge zur Amplitude. Die Dicken-Variation von Spot zu Spot ist unkorreliert. Das ist Teil der Modell-Annahmen, die sich sehr gut bei der Beschreibung von Speckle-Phänomenen bewährt haben. Das heißt, die Gesamtamplitude im Punkt $\vec{x}_B$ ist somit

$$A(\vec{x}_B) = A_0 \, \Delta a \sum_{n=1}^{N} \frac{e^{ik|\vec{x}_B - \vec{x}_n|}}{|\vec{x}_B - \vec{x}_n|} \qquad . \tag{11.3}$$

Hierbei sind $\vec{x}_n$ die Positionen der Scattering-Spots in der Diffuser-Ebene. Wir hatten angenommen, dass $R \gg \Delta$, das bedeutet, dass wir denn Nenner $|\vec{x}_B - \vec{x}_n|$ durch R approximieren können. Zusätzlich können wir das Argument der Exponential-Funktion in eine Reihe um $\Delta_n = 0$ entwickeln

$$\begin{aligned}
r_n = |\vec{x}_B - \vec{x}_n| &= \sqrt{(x_n - x_B)^2 + (y_n - y_B)^2 + (R - \Delta_n)^2} \\
&= \sqrt{(x_n - x_B)^2 + (y_n - y_B)^2 + R^2} \\
&\quad - \frac{R}{\sqrt{(x_n - x_b)^2 + (y_n - y_B)^2 + R^2}} \, \Delta_n + \dots \\
&= R + \frac{(x_n - x_B)^2 + (y_n - y_B)^2}{2R} - \Delta_n + O(\Delta^2) \\
&= D_n - \Delta_n + O(\Delta^2) \qquad .
\end{aligned}$$

Hier wurde ausgenutzt, dass auch der Strahldurchmesser, also der Bereich der Diffuser-Ebene der zum Integral beiträgt, sehr viel kleiner als R ist. Die Größen D_n sind deterministisch und hängen nicht von den Zufalls-Variablen ab. Damit erhalten wir für die Amplitude im Punkt $\vec{x}_B$

$$A(\vec{x}_B) = \kappa \sum_{n=1}^{N} e^{ikD_n - k\Delta_n} \qquad . \tag{11.4}$$

$$A_{\mathrm{Re}} = \kappa \sum_{n=1}^{N} \cos(kD_n - k\Delta_n)$$
$$A_{\mathrm{Im}} = \kappa \sum_{n=1}^{N} \sin(kD_n - k\Delta_n)$$

Die Terme kD_n sind deterministisch und hängen von den Koordinaten des n-ten Spots ab. Die Größen $r_n := k\Delta_n$ sind unkorrelierte Zufalls-Variablen. Es soll außerdem kein Punkt auf dem Diffuser ausgezeichnet sein. Deshalb sind diese Zufalls-Variablen i.u.v. Die Wahrscheinlichkeitsdichte von $C = \cos(kD_n - k\Delta_n)$ ist

$$p(C|k, D_n, \mathcal{B}) = \int \delta(C - \cos(kD_n - r_n))\; p(r_n|\mathcal{B})\; dr_n \qquad .$$

Wir gehen davon aus, dass die Zufalls-Variablen $r_n = k\delta_n$ eine flache Wahrscheinlichkeitsdichte besitzen, d.h. es soll kein Wert ausgezeichnet sein. Physikalisch bedeutet das, dass die Verteilung der Diffuser-Dicken sehr viel größer sind als die Wellenlänge λ und somit die Verteilung von $k\Delta_n = 2\pi\Delta_n/\lambda$ sehr breit ist. Damit ist die Wahrscheinlichkeitsdichte $p(C|k, D_n, \mathcal{B})$ identisch zu der des harmonischen Oszillators Gl. (9.37) für $A = 1$

$$p(C|k, D_n, \mathcal{B}) = \frac{1}{\pi\sqrt{1 - C^2}} \qquad .$$

Das heißt, die Größen $\cos(kD_n - k\Delta_n)$ sind i.u.v. Zufallsvariablen mit der Wahrscheinlichkeitsdichte $p(C|k, D_n, \mathcal{B})$. Da $\sin(kD_n - k\Delta_n) = \cos(kD_n - k\Delta_n + \pi/2)$, sind auch die Größen $\sin(kD_n - k\Delta_n)$ i.u.v. Zufalls-Variablen mit derselben Wahrscheinlichkeitsdichte.
Das heißt schließlich, dass sowohl A_{Re} als auch A_{Im} Summen von i.u.v. Zufallsvariablen mit derselben Wahrscheinlichkeitsdichte sind. Für den Zentralen Grenzwertsatz benötigen wir noch den Mittelwert

$$\langle C \rangle = \frac{1}{\pi} \int_{-1}^{1} \frac{1}{\sqrt{1 - x^2}}\, dx = 0$$

und die Varianz

$$\mathrm{var}(C) = \frac{1}{\pi} \int_{-1}^{1} \frac{x^2}{\sqrt{1 - x^2}}\, dx \qquad .$$

Die Substitution $x = \cos(\varphi)$ führt zu

$$\begin{aligned} \mathrm{var}(C) &= \frac{1}{\pi} \int_0^{\pi} \frac{\cos^2(\varphi)}{\sqrt{1 - \cos^2(\varphi)}}\, \sin(\varphi)\, d\varphi \\ &= \frac{1}{\pi} \int_0^{\pi} \frac{\cos^2(\varphi)}{\sin(\varphi)}\, \sin(\varphi)\, d\varphi = \frac{1}{\pi} \int_0^{\pi} \cos^2(\varphi)\, d\varphi = \frac{1}{\pi}\frac{\pi}{2} \\ &= \frac{1}{2} \qquad . \end{aligned}$$

Die Varianz der Variablen $\kappa \cos(kD_n - r_n)$ ist demnach $\kappa^2/2$. Der Zentrale Grenzwertsatz besagt deshalb, dass

$$p(A_{\mathrm{Re}}|\kappa, N, \mathcal{B}) = \frac{1}{\sqrt{N\pi\kappa^2}}\, e^{-A_{\mathrm{Re}}^2/N\kappa^2}$$
$$p(A_{\mathrm{Im}}|\kappa, N, \mathcal{B}) = \frac{1}{\sqrt{N\pi\kappa^2}}\, e^{-A_{\mathrm{Im}}^2/N\kappa^2} \qquad .$$

Damit kennen wir die marginalen Wahrscheinlichkeitsdichte. Wir benötigen aber die joint probability $p(A_{\mathrm{Re}}, A_{\mathrm{Im}}|\kappa, N, \mathcal{B})$. Wir werden beweisen, dass die Zufalls-Variablen A_{Re} und A_{Im} unkorreliert sind.

$$\begin{aligned}\langle A_{\mathrm{Re}}\, A_{\mathrm{Im}}\rangle &= \kappa^2 \sum_n \sum_m \langle \cos(kD_n - k\Delta_n)\sin(kD_m - k\Delta_m)\rangle \\ &= \kappa^2 \sum_n \langle \cos(kD_n - k\Delta_n)\sin(kD_n - k\Delta_n)\rangle \\ &+ \kappa^2 \sum_{\substack{n,m \\ n\neq m}} \langle \cos(kD_n - k\Delta_n)\rangle\, \langle \sin(kD_m - k\Delta_m)\rangle \qquad .\end{aligned}$$

Es wurde ausgenutzt, dass die Δ_n und Δ_m für $n \neq m$ unabhängig sind. Zudem ist der Mittelwert $\langle \sin(kD_m - k\Delta_m)\rangle = 0$. Somit bleibt nur noch

$$\begin{aligned}\langle A_{\mathrm{Re}}\, A_{\mathrm{Im}}\rangle &= \kappa^2 \sum_n \langle \cos(kD_n - k\Delta_n)\sin(kD_n - k\Delta_n)\rangle \\ &= \kappa^2 \sum_n \frac{1}{2}\langle \sin(2kD_n - 2k\Delta_n)\rangle = 0 \qquad .\end{aligned}$$

Der letzte Mittelwert ist ebenfalls null, da es keinen Unterschied macht, ob wir k oder $2k$ einsetzen. Somit haben wir gezeigt, dass die Kovarianz von A_{Re} und A_{Im} null ist und deshalb gilt

$$p(A_{\mathrm{Re}}, A_{\mathrm{Im}}|\kappa, N, \mathcal{B}) = \frac{1}{\pi\sigma^2}\, e^{-\frac{1}{\sigma^2}(A_{\mathrm{Re}}^2 + A_{\mathrm{Im}}^2)} \qquad , \tag{11.5}$$

mit $\sigma^2 = N\kappa^2$.
Nun können wir daraus die gesuchte Wahrscheinlichkeitsdichte der Intensität und der Phase berechnen

$$p(I, \Phi|\kappa, N, \mathcal{B}) = p(A_{\mathrm{Re}}, A_{\mathrm{Im}}|\kappa, N, \mathcal{B}) \left|\frac{\partial A_{\mathrm{Re}}, A_{\mathrm{Im}}}{\partial I, \Phi}\right| \qquad .$$

Wegen

$$A_{\mathrm{Re}} = \sqrt{I}\, \cos(\Phi)$$
$$A_{\mathrm{Im}} = \sqrt{I}\, \sin(\Phi)$$

gilt

$$\left|\frac{\partial A_{\mathrm{Re}}, A_{\mathrm{Im}}}{\partial I, \Phi}\right| = \left|\begin{pmatrix} \frac{\partial A_{\mathrm{Re}}}{I} & \frac{\partial A_{\mathrm{Re}}}{\Phi} \\ \frac{\partial A_{\mathrm{Im}}}{I} & \frac{\partial A_{\mathrm{Im}}}{\Phi} \end{pmatrix}\right| = \left|\begin{pmatrix} \frac{1}{2\sqrt{I}}\cos(\Phi) & -\sqrt{I}\sin(\Phi) \\ \frac{1}{2\sqrt{I}}\sin(\Phi) & \sqrt{I}\cos(\Phi) \end{pmatrix}\right| = \frac{1}{2} \quad .$$

Schließlich ist also

$$p(I, \Phi|\kappa, N, \mathcal{B}) = \underbrace{\frac{1}{\sigma^2}\, e^{-I/\sigma^2}\, \theta(I \geq 0)}_{p(I|\kappa,N,\mathcal{B})} \underbrace{\frac{1}{2\pi}\theta(0 \leq \Phi < 2\pi)}_{p(\Phi|\kappa,N,\mathcal{B})} \quad . \tag{11.6}$$

Die Phase ist also gleich-verteilt. Die Intensität genügt einer Exponential-Verteilung mit Mittelwert $\langle I \rangle = \sigma^2$ (vgl. Gl. (9.15)). Diese sehr breite Verteilung erklärt die granulare Erscheinung der Speckle-Muster. Die Fluktuationen sind vergleichbar zum Mittelwert.

11.2 Signal-zu-Rauschen (S/R) Verhältnis

Das Signal-zu-Rauschen (S/R) Verhältnis gibt an, in welchem Ausmaß ein Signal vom Rauschen gestört ist. Es ist definiert als das Verhältnis von mittlerer Intensität zur Standardabweichung der Intensität. Das mittlere Signal ist bereits bekannt

$$\langle I \rangle = \sigma^2 \quad .$$

Die Standard-Abweichung der Exponential-Verteilung kennen wir bereits aus Gl. (9.15d)

$$\sqrt{\mathrm{var}(I)} = \langle I \rangle \quad .$$

Somit ist das Signal-zu-Rauschen Verhältnis

$$S/R = \frac{\langle I \rangle}{\sqrt{\mathrm{var}(I)}} = 1 \quad .$$

Das Ergebnis zeigt, wie stark verrauscht der Speckle-Prozess ist. Das erklärt den granularen Charakter der Laser-Speckle. I.d.R. ist die Granularität unerwünscht, und man ist bestrebt sie loszuwerden. Das geht offensichtlich nicht, indem man die Eingangsamplitude A_0 (Intensität) vergrößert, da S/R unabhängig hiervon ist.

11.3 Verbesserung des Signal-zu-Rauschen Verhältnisses

Eine Möglichkeit besteht darin, das Bild in der Ebene B über einen Bereich ΔA, der viele Speckle-Körner enthält, zu mitteln. Man verliert hierdurch zwar an Auflösung, der Kontrast wird aber deutlich verbessert.

Das Bild besteht aus vielen zufälligen „Körnern". Im Mittel befinden sich in einem Bereich der Größe ΔA M Körner, über deren Intensität wir mitteln werden. Die Intensitäten innerhalb des Bereiches, über den gemittelt wird, seien alle i.u.v. gemäß einer Exponential-Verteilung. Wir nummerieren die Körner fortlaufend durch. Die Intensität im Bereich ΔA ist dann

$$\mathcal{I} = \sum_{m=1}^{M} I_m \qquad .$$

Es handelt sich wieder um eine Summe von i.u.v. Zufalls-Variablen. Da M aber nicht unbedingt sehr viel größer als eins sein soll, werden wir hier den Zentralen Grenzwertsatz nicht verwenden. Wir kennen ohnehin bereits die Wahrscheinlichkeitsdichte $p(\mathcal{I}|M, \sigma, \mathcal{B})$ von $\mathcal{I}$ aus Gl. (10.20). Sie lautet

$$p(\mathcal{I}|M, \sigma, \mathcal{B}) = p_\Gamma(\mathcal{I}|\alpha = M, \beta = \sigma^{-2}) = \frac{\sigma^{-2M}}{\Gamma(M)}\, \mathcal{I}^{M-1}\, e^{-\mathcal{I}/\sigma^2} \qquad .$$

Diese Wahrscheinlichkeitsdichte wird auch χ^2-VERTEILUNG genannt. Wir kennen aus Gl. (9.11c) und Gl. (9.11d) den Mittelwert und Varianz der Γ-Verteilung. Das Signal-zu-Rauschen Verhältnis ist demnach

$$S/R = \frac{\langle \mathcal{I} \rangle}{\sqrt{\mathrm{var}(\mathcal{I})}} = \frac{M\sigma^2}{\sqrt{M\sigma^4}} = \sqrt{M} \qquad .$$

Das zeigt den Vorteil der Mittelwert-Bildung. Das Signal-zu-Rauschen Verhältnis verbessert sich mit $\sqrt{M}$, der Wurzel aus der Anzahl der Körner im Bereich ΔA, über den gemittelt wird. Das Verfahren hat allerdings auch große Nachteile, da es wegen der Faltung mit einer stufenförmigen Fenster-Funktion hohe Fourierkomponenten beimischt und außerdem die räumliche Auflösung verringert. Die Speckle-Reduktion gehört zu den aktuellen Forschungsaktivitäten. Verbesserte Ansätze verwenden die Wavelet-Transformation zur optimalen lokale Mittelwertbildung in Kombination mit einer wahrscheinlichkeitstheoretische Auswertung.

11.4 Die Standard-Form der χ^2-Verteilung

Die χ^2-Verteilung ist hier in einem besonderen physikalischen Kontext entstanden. Von der Herleitung erkennen wir, dass diese Verteilung immer entsteht, wenn eine Zufalls-Variable z

$$z = \sum_{n=1}^{N} x_n^2$$

aus einer Summe von N i.u.v. Zufalls-Variablen x_n aufgebaut ist, die einer zentrierten Normal-Verteilung $\mathcal{N}(x_n|0, \sigma)$ genügen. Die x_n nennt man in diesem Zusammenhang FREIHEITSGRADE und N die ZAHL DER FREIHEITSGRADE.

Die χ^2-Verteilung entsteht im Zusammenhang mit vielen physikalischen Problemen, z.B. Maxwellsche Geschwindigkeitsverteilung. Sie charakterisiert auch das Gesetz der Stichproben-Varianz und wird in Signifikanz-Tests verwendet, die wir in einem späteren Kapitel besprechen werden.

Teil II

Poisson-Prozeß, Poisson-Punkte und Wartezeiten

Kapitel 12

Poisson-Prozess, Poisson-Punkte und Wartezeiten

12.1 Stochastische Prozesse

Stochastische Prozesse sind in der Physik weit verbreitet.
Wir gehen von einer parametrisierten Funktion

$$f(x|\lambda)$$

aus, bei der $x \in \mathbb{R}$ kontinuierliche Werte annimmt und $\lambda \in \mathbb{R}^{N_p}$ die Parameter der Funktion darstellen. Wenn die Parameter λ Zufalls-Variablen sind, nennt man $f(x|\lambda)$ einen stochastischen Prozess. Die Unsicherheit des stochastischen Prozesses liegt in der stochastischen Natur der Parameter. Ein Beispiel stellt die Energie-Verlust-Kurve $E(x)$ beim Durchgang von Strahlung durch Materie dar, die angibt mit welcher Energie die Strahlung am Ort x ankommt, wenn sie mit einer Anfangsenergie E_0 bei $x = 0$ erzeugt wird. Der Energieverlust hängt vom lokalen Aufbau der Materie ab, der stochastische Charakterzüge haben kann, wie z.B. beim Durchgang von Licht durch Wolken, Emulsion, etc. Wenn die Parameter λ, hier die räumliche Dichte-Verteilung der Materie, gegeben sind, ist die Funktion $E(x)$ i.d.R eine stetige Funktion von x. Ein weiteres Beispiel sind Radarsignale. Ein Radarsignal hat die Form $f(t|d) = \cos(\omega t - 2kd)$, wobei ω die Frequenz, $k = 2\pi/\lambda$ das Inverse der Wellenlänge darstellt. Die Zufalls-Variable in diesem Zusammenhang ist der Abstand d zur unbekannten Quelle. Die Radarsignale haben daher zufällige Verzögerungen. Der zeitliche Ablauf des Signals ist jedoch eine glatte Cosinus-Kurve.
Es kann jedoch auch vorkommen, dass sich das stochastische Verhalten auch in der Abhängigkeit von der unabhängigen Variablen x in $f(x|\lambda)$ äußert. Das ist z.B. der Fall bei additivem Rauschen. Ein Signal $s(t)$ sei durch unkorreliertes Rauschen $\eta(t)$ gestört

$$\tilde{s}(t) = s(t) + \eta(t) \qquad .$$

Das Rauschen hat hier die Rolle des Parameters λ übernommen. Eine Realisierung

liefert eine feste zeitliche Abfolge von Rauschbeiträgen $\eta(t)$, die unkorreliert sein sollen. Daraus resultiert, dass $\tilde{s}(t)$ nicht mehr stetig ist.
Stochastische Prozesse unterscheiden sich von anderen Zufalls-Variablen lediglich darin, dass die Zufalls-Variable nun einen kontinuierlichen Index erhält. Ansonsten gelten alle Regeln der Wahrscheinlichkeitstheorie unverändert. Z.B. bedeutet die Mittelungen von stochastischen Prozessen, dass über die stochastischen Parameter gemittelt wird

$$\langle f(x) \rangle = \int f(x|\lambda)\, p(\lambda|\mathcal{B})\, d\lambda \qquad .$$

Die Mittelwerte hängen dann auch von der unabhängigen Variablen (hier x) ab. Das ist eigentlich nichts anderes als die Marginalisierungsregel.
Ein weiteres Beispiel stochastischer Prozesse stellt das epitaktische Oberflächenwachstum dar, bei dem Atome stochastisch auf die Oberfläche auftreffen. Ein verwandtes Phänomen sind Warteschlangen an Kassen etc.. Wir werden in diesem Kapitel noch den Poisson-Prozess im Detail kennenlernen.

12.2 Poisson Punkte

Es werden zufällig N Punkte in einem Intervall $\Omega = (0, L)$ der Länge L erzeugt. Die so erzeugten Punkte nennt man Poisson-Punkte (PP). Die Punktdichte ist dabei

$$\rho := \frac{N}{L} \qquad . \tag{12.1}$$

Wir greifen nun ein beliebiges Teilintervall $I \subset \Omega$ der Länge x heraus und fragen nach der Wahrscheinlichkeit, dass bei den N Versuchen n Teilchen in das Teilintervall x treffen. Die Wahrscheinlichkeit, dass bei einem dieser Versuche das Teilchen in das Intervall gelangt, ist nach der klassischen Definition von Wahrscheinlichkeit, das Verhältnis der „Zahl der günstigen Ereignisse" zur „Zahl der gesamten Möglichkeiten", also

$$p = \frac{x}{L} \qquad .$$

Die gesuchte Wahrscheinlichkeit ist demnach die Binomial-Verteilung $P(n|N, p = x/L)$. Die mittlere Zahl der Teilchen, die in I landen, ist

$$\mu = p\, N = \frac{x}{L}\, N \; = x\, \rho \qquad . \tag{12.2}$$

Die Größe des Intervalls I behalten wir nun bei, verdoppeln aber sowohl die Länge des Grundintervalls $L \to 2L$ als auch die Zahl der Teilchen $N \to 2N$. Hierbei bleibt die Teilchendichte $\rho := \frac{N}{L}$ und demnach auch die mittlere Zahl der Teilchen in L konstant. Wir wiederholen die Intervall-Verdopplung unendlich oft, bis das Intervall L schließlich die gesamte reelle Achse von $-\infty$ bis ∞ abdeckt. Diese Konstruktion erfüllt die Voraussetzungen für den Satz von Poisson ($\mu = p\, N = \text{const}$ und $N \to \infty$)

und demnach geht die gesuchte Wahrscheinlichkeit im Limes $N \to \infty$[1] in die Poisson-Verteilung über

$$P(n|x,\rho,\mathcal{B}) := e^{-\rho x}\,\frac{(\rho x)^n}{n!} \qquad . \tag{12.3}$$

12.3 Intervall-Verteilung der Poisson-Punkte

Wir wollen nun untersuchen, wie die Abstände Δ_x zwischen PPen verteilt sind. D.h. wir wollen $p(\Delta_x|\rho, I)$ berechnen. Wir verwenden hier die Proposition
Δ_x: *Der Wert des Abstands zum nächsten PP ist aus* $[x, x+dx)$.
Damit die Proposition Δ_x zutrifft, darf im Intervall der Länge x kein PP liegen und anschließend muss in dx ein PP vorhanden sein. Diese beiden Ereignisse sind unabhängig voneinander. Der erste Faktor ist die Wahrscheinlichkeit $P(n=0|x,\rho,\mathcal{B})$ aus Gl. (12.3). Die Wahrscheinlichkeit, in dx einen PP anzutreffen, ist, da die Dichte der Punkte homogen gleich ρ ist, durch $\rho\ dx$ gegeben. Somit lautet die gesuchte Wahrscheinlichkeit

$$\begin{aligned} P(\Delta_x|\rho,\mathcal{B}) &= P(n=0|x,\rho,\mathcal{B})\ P(n=1|dx,\rho,\mathcal{B}) \\ &= e^{-\rho\,x}\ \rho\ dx \qquad . \end{aligned}$$

Die zugehörige Wahrscheinlichkeitsdichte ist demnach

INTERVALL-LÄNGEN-VERTEILUNG IM POISSON-PROZESS

$$p(\Delta_x|\rho,\mathcal{B}) = \rho\ e^{-\rho x} \tag{12.4}$$

die Exponential-Verteilung Gl. (9.15). Der Mittelwert dieser Verteilung ist $1/\rho$ in Einklang mit der mittleren Dichte ρ.

12.3.1 Alternative Sicht der Poisson Punkte

Man kann die Entstehung der Poisson-Punkte auch anders interpretieren. Bisher wurden die Punkte mit vorgegebener Dichte ρ auf ein Intervall zufällig verteilt. Auf diese Weise ist gleichzeitig gewährleistet, dass die Abstände zwischen benachbarten Punkten unabhängig voneinander sind und alle derselben Wahrscheinlichkeitsdichte genügen. Die Intervall-Längen sind i.u.v. (identisch unabhängig verteilt). Wir können die Poisson-Punkte also auch sequentiell erzeugen. Wir beginnen bei $x=0$ und bestimmen den nächsten Poisson-Punkt x_1 im Abstand Δx_1 gemäß der Exponential-Verteilung, die in Gl. (12.4) gegeben ist. Danach erzeugen wir x_2 ausgehend von x_1

[1] Siehe auch Gl. (4.19).

im Abstand Δx_2, der wiederum aus $p(\Delta|\rho, \mathcal{B})$ ermittelt wird, etc. Die Poisson-Punkte haben dann die Werte

$$x_n = \sum_{i=1}^{n} \Delta x_i \qquad .$$

12.4 Wartezeiten-Paradoxon

Wir wissen, dass die Abstände Δx zwischen zwei Poisson-Punkten exponentiell verteilt sind.
Wir interpretieren die x-Achse nun als Zeitachse. Hieraus folgt, dass die Abstände zeitlicher Ereignisse, die einem Poisson-Prozess genügen, ebenfalls exponentiell verteilt sind

$$p(\Delta t|\lambda, \text{exp.}) = \lambda e^{-\lambda\, \Delta t} \qquad . \tag{12.5}$$

Die zeitliche Dichte ist hier mit λ bezeichnet.
Nehmen wir an, die Ankunft von Bussen an einer Haltestelle sei Poisson-verteilt. Wir kommen zufällig an der Haltestelle an. Wie lange müssen wir im Mittel auf den nächsten Bus warten?
Da der Abstand der Ereignisse im Mittel λ beträgt und wir zufällig zwischen den Ereignissen angekommen sind, sollten alle Zeiten zwischen 0 und λ zu erwarten sein. Von daher erwarten wir eine mittlere Wartezeit $\lambda/2$.
Dieses Ergebnis ist aber falsch!
Wir wollen nun die korrekte Antwort mit den elementaren Regeln der Wahrscheinlichkeitstheorie ableiten. Gesucht ist die Wahrscheinlichkeitsdichte $p(\Delta t|t \in I, \mathcal{B})$, dass die Wartezeit Δt beträgt, wenn wir zufällig in einem der Intervall, nennen wir es I, ankommen. Hierbei gibt t den Zeitpunkt unserer Ankunft an. Wir führen über die Marginalisierungsregel die Länge L des Intervalls ein, in dem wir ankommen.

$$\begin{aligned}
p(\Delta t|t \in I, \mathcal{B}) &= \int_0^\infty p(\Delta t, L_I = L|t \in I, \mathcal{B})\; dL \\
&= \int_0^\infty \underbrace{p(\Delta t|L_I = L, t \in I, \mathcal{B})}_{\frac{1}{L}\theta(0<\Delta t\leq L)}\; P(L_I = L|t \in I, \mathcal{B})\; dL \\
&= \int_{\Delta t}^\infty \frac{1}{L}\; P(L_I = L|t \in I, \mathcal{B})\; dL \\
&= \int_{\Delta t}^\infty \frac{1}{L}\; \frac{P(t \in I|L_I = L, \mathcal{B})\; P(L_I = L|\mathcal{B})}{P(t \in I|\mathcal{B})}\; dL
\end{aligned}$$

Die Wahrscheinlichkeit, dass wir in einem Intervall der Länge L ankommen ist proportional zu L

$$P(t \in I|L_I = L, \mathcal{B}) = \kappa\, L \qquad .$$

Die Wahrscheinlichkeit, dass die Intervall-Länge L beträgt, ist durch die exponentielle Verteilung der Intervall-Längen vorgegeben. Der Ausdruck im Nenner ist unabhängig von L und Δt und dient hier nur der Normierung. Wir können die beiden unbekannten Faktoren κ und $P(t \in I|\mathcal{B})$ zu einem Normierungsfaktor Z zusammenfassen, den wir anschließende über die Normierung bestimmen.

$$\begin{aligned} p(\Delta t|t \in I, \mathcal{B}) &= \frac{1}{Z} \int_{\Delta t}^{\infty} \frac{1}{L}\, L\, \lambda\, e^{-\lambda L}\, dL \\ &= \frac{1}{Z}\, e^{-\lambda \Delta t} \qquad . \end{aligned}$$

$$Z = \int_0^{\infty} e^{-\lambda \Delta t}\, dt = \frac{1}{\lambda} \qquad .$$

Somit lautet das gesuchte Ergebnis

$$p(\Delta t|t \in I, \mathcal{B}) = \lambda\, e^{-\lambda\, \Delta t} \qquad . \tag{12.6}$$

Das heißt, die Wartezeiten sind genauso verteilt wie die Abstände zwischen den Ereignissen und somit ist die mittlere Wartezeit gleich dem mittleren Abstand der Ereignisse $\langle \Delta t \rangle = \lambda$. Der Grund ist, dass die Wahrscheinlichkeit größer ist, zwischen zwei Ereignissen einzutreffen, die einen großen zeitlichen Abstand voneinander haben.
Das erklärt auch, warum auch das erste Intervall vom willkürlich gewählten zeitlichen Nullpunkt zum ersten Ereignis, eine mittlere Länge λ hat.

12.4.1 Verteilung der Intervall-Längen eines zufällig ausgewählten Intervalls

Wir wählen zufällig einen Punkt x aus und fragen nach der Wahrscheinlichkeit, dass das so ausgewählte Intervall I die Größe L besitzt

$$p(L|I \ni x, \mathcal{B}) \qquad .$$

Nach dem Bayesschen Theorem gilt hierfür

$$p(L|I \ni x, \rho, \mathcal{B}) = \frac{p(x \in I|L, \rho, \mathcal{B})\; p(L|\rho, \mathcal{B})}{p(x \in I|\mathcal{B})} \qquad .$$

Wie zuvor ist $p(x \in I|L, \mathcal{B}) = \kappa\, L$ proportional zur Intervall-Länge und $p(x \in I|\mathcal{B})$ ein von L-unabhängiger Normierungsfaktor

$$
\begin{aligned}
p(L|I \ni x, \rho, \mathcal{B}) &= \frac{1}{\tilde{Z}} L \, p(L|\mathcal{B}) \\
&= \frac{1}{\tilde{Z}} \, L \, \rho \, e^{-\rho L} \\
\tilde{Z} &= \int_0^\infty L \, \rho \, e^{-\rho L} \, dL = \frac{1}{\rho} \\
p(L|I \ni x, \rho, \mathcal{B}) &= \rho^2 \, L \, e^{\rho L} \qquad .
\end{aligned}
$$

Die mittlere Intervall-Länge ist demnach

$$
\langle L \rangle = \int_0^\infty (\rho L)^2 \, e^{-\rho L} dL = \frac{2}{\rho} \qquad .
$$

Der Grund ist natürlich wie zuvor, dass mit größerer Wahrscheinlichkeit der Aufpunkt in einem Intervall liegt, das größer ist.

12.5 Poisson-Prozess

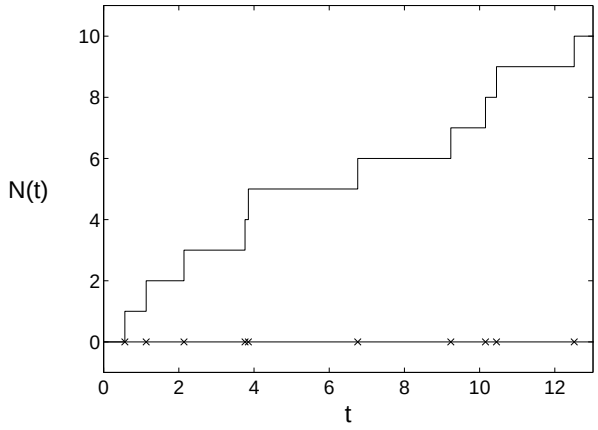

Abbildung 12.1: *Poisson-Punkte (Kreuze) und Poisson-Prozess $N(t)$ für $\lambda = 1$.*

Die Punkte $t_1, t_2, \ldots, t_n$ sind PP. Mit $N(t)$ bezeichnen wir den stochastischen Prozess, der die ganzzahligen Werte $0, 1, \ldots$ annimmt, die angeben, wieviele Poisson-Punkte bis zur Zeit t aufgetreten sind.

$$
N(t) = \sum_{n=0}^{\infty} \theta(t \leq t_n) \qquad .
$$

Die Wahrscheinlichkeit, dass $N(t) = n$ ist natürlich die Wahrscheinlichkeit der Poisson-Verteilung.

12.6 Ordnungsstatistik des Poisson-Prozesses

Wir wollen die Wahrscheinlichkeitsdichte bestimmen, dass der n-te PP die Koordinate x hat. Wir können die allgemeinen Überlegungen über Ordungsstatistiken übernehmen. Damit der n-te PP die Koordinate x hat müssen im Intervall $[0, x)$ $n-1$ PPe liegen. Die Wahrscheinlichkeit hierfür ist die Poisson-Verteilung

$$P(n-1|x,\rho) = e^{-\rho\, x}\, \frac{(\rho\, x)^{n-1}}{(n-1)!} \qquad .$$

Die Wahrscheinlichkeit, den nächsten Punkt in $(x, x+dx)$ anzutreffen, ist $\rho\, dx$. Da die PPe unkorreliert sind, ist die gesuchte Wahrscheinlichkeitsdichte das Produkt aus beiden Faktoren. Das führt zur

ERLANG-VERTEILUNG

$$p(x_n = x|\rho, \mathcal{B}) = \rho\, e^{-\rho\, x}\, \frac{(x\, \rho)^{n-1}}{(n-1)!} \tag{12.7}$$

Der Erwartungswert dieser Verteilung liefert

$$\begin{aligned}
\langle x_n \rangle &= \frac{\rho}{(n-1)!} \int_0^\infty e^{-\rho\, x} (\rho\, x)^{n-1}\, x\, dx \\
&= \frac{1}{\rho\, (n-1)!} \int_0^\infty e^{-t} (t)^n\, dt \\
&= \frac{n!}{\rho\, (n-1)!} \\
&= \frac{n}{\rho} \qquad .
\end{aligned}$$

Das Ergebnis ist in Einklang mit der Erwartung, da der mittlere Abstand der Punkte $1/\rho$ beträgt.

12.7 Alternative Herleitung des Poisson-Prozesses

Wir messen den radioaktiven Zerfall mit einem Geiger Zähler. Hiermit wird die Zahl der Zerfälle $X(t)$ bis zum Zeitpunkt t ermittelt.
Der Poisson-Prozess ist folgendermaßen spezifiziert:

- Die Wahrscheinlichkeit, dass in einem infinitesimalen Zeitintervall dt ein Zerfall stattfindet, ist unabhängig von der Vorgeschichte und vom gegenwärtigen Zustand.

- Die Wahrscheinlichkeit ist analytisch in der Zeit $P(\text{Zerfall in dt}|\lambda, \mathcal{B}) = \lambda\ dt$.
- Da dt infinitesimal ist, ist die Wahrscheinlichkeit für zwei Zerfälle $O(dt^2)$ vernachlässigbar.
- Die Wahrscheinlichkeit, dass in dt kein Zerfall stattfindet, ist demnach $1 - \lambda\ dt + O(dt^2)$.

Wir definieren die Proposition

$$A(n,t) = \textit{Zahl der Zerfälle bis zur Zeit t beträgt n.}$$

Gesucht ist die Wahrscheinlichkeit $P_n(t) := P(A(n,t)|t,\mathcal{B})$, dass bis zur Zeit t die Zahl der Poisson-Punkte n beträgt. Wir wollen hierfür eine Differentialgleichung in der Zeit ableiten und betrachten dazu die Wahrscheinlichkeit zur Zeit $t + dt$. Über die Marginalisierungsregel summieren wir über die Propositionen $A(m,t)$ zur Zeit t

$$\begin{aligned} P(A(n,t+dt)|\mathcal{B}) &= \sum_{m=0}^{\infty} P(A(n,t+dt), A(m,t)|\mathcal{B}) \\ &= \sum_{m=0}^{\infty} P(A(n,t+dt)|A(m,t),\mathcal{B})\ P(A(m,t)|\mathcal{B}) \end{aligned}$$

Da in dt höchstens ein Zerfall stattfinden kann, muß die Zahl der Poisson-Punkte zur Zeit t entweder ebenfalls $m = n$ oder $m = n-1$ betragen. Die vernachlässigten Terme sind von der Ordnung dt^2 und verschwinden für $dt \to 0$. Wir haben somit

$$\begin{aligned} P(A(n,t+dt)|\mathcal{B}) &= P(A(n,t+dt)|A(n,t),\mathcal{B})\ P(A(n,t)|\mathcal{B}) \\ &\quad + P(A(n,t+dt)|A(n-1,t),\mathcal{B})\ P(A(n-1,t)|\mathcal{B}) \\ &= (1-\lambda\ dt)\ P(A(n,t)|\mathcal{B}) + \lambda\ dt\ P(A(n-1,t)|\mathcal{B}) \\ \frac{d}{dt}\ P(A(n,t)|\mathcal{B}) &= \lambda \left(P(A(n-1,t)|\mathcal{B}) - P(A(n,t)|\mathcal{B}) \right) \quad . \end{aligned}$$

Es wurde $P_{-1}(t) = 0$ definiert.

$$\frac{d}{dt}\ P_n(t) = \lambda \left(P_{n-1}(t) - P_n(t) \right) \tag{12.8}$$

Die Anfangsbedingungen sind

$$P_n(0) = \delta_{n,0} \quad , \tag{12.9}$$

da zur Zeit $t = 0$ noch kein Zerfall stattgefunden hat. Diese Differenzen-Differential-Gleichung kann mit elementaren Methoden gelöst werden. Es soll hier aber ein allgemeineres Lösungsverfahren über ERZEUGENDE FUNKTIONEN vorgestellt werden, da sie in der Wahrscheinlichkeitstheorie eine wichtige Rolle spielen.

Wir definieren die ERZEUGENDE FUNKTION $\varphi(x,t)$ über die folgende Reihenentwicklung

ERZEUGENDE FUNKTION

$$\varphi(x,t) = \sum_{n=0}^{\infty} P_n(t)\, x^n \qquad . \tag{12.10}$$

Zu beliebiger aber fester Zeit t stellen die $P_n(t)$ die Entwicklungskoeffizienten einer Taylorentwicklung dar. Da $\varphi(x,t)$ nach wie vor die gesamte Information der $P_n(t)$ enthält, können alle wahrscheinlichkeitstheoretischen Berechnungen auch über $\varphi(x,t)$ vorgenommen werden. Dieser Zugang ist in der Regel wesentlich effizienter. Z.B. ist die Norm von $P_n(t)$ über

$$\sum_n P_n(t) = \varphi(1,t)$$

gegeben. Die Momente der Verteilung

$$\mu_\nu(t) := \sum_n n^\nu\, P_n(t)$$

erhält man aus

$$\left.\frac{d^\nu}{dx^\nu}\, \varphi(x,t)\right|_{x=1} \qquad .$$

Zum Beispiel gilt

$$\begin{aligned} \left.\frac{d}{dx}\, \varphi(x,t)\right|_{x=1} &= \mu_1 \\ \left.\frac{d^2}{dx^2}\, \varphi(x,t)\right|_{x=1} &= \mu_2 - \mu_1 \qquad . \end{aligned}$$

Wenn $\varphi(x,t)$ bekannt ist, erhält man daraus die Wahrscheinlichkeiten $P_n(t)$

$$P_\nu(t) = \frac{1}{\nu!}\, \left.\frac{d^\nu}{dx^\nu}\, \varphi(x,t)\right|_{x=0} \tag{12.11}$$

Beweis

$$\begin{aligned}
\frac{1}{n!}\,\frac{d^n}{dx^n}\,\varphi(x,t)\Big|_{x=0} &= \frac{1}{n!}\,\frac{d^n}{dx^n}\,\sum_{m=0}^{\infty} P_m(t)\,x^m\,\Big|_{x=0} \\
&= \frac{1}{n!}\,\sum_{m=n}^{\infty} P_m(t)\,m(m-1)\cdots(m-n+1)\,\underbrace{x^{m-n}\,\Big|_{x=0}}_{\delta_{m,n}} \\
&= \frac{1}{n!}\,P_n(t)\,n(n-1)\cdots(n-n+1) \\
&= P_n(t) \qquad .
\end{aligned}$$

Wir multiplizieren Gl. (12.8) mit x^n und summieren über n und erhalten

$$\begin{aligned}
\frac{d}{dt}\underbrace{\sum_{n=0}^{\infty} P_n(t)\,x^n}_{\varphi(x,t)} &= \lambda\left(\underbrace{\sum_{n=0}^{\infty} P_{n-1}(t)\,x^n}_{x\,\varphi(x,t)} - \underbrace{\sum_{n=0}^{\infty} P_n(t)\,x^n}_{\varphi(x,t)}\right) \\
\frac{d}{dt}\varphi(x,t) &= \lambda\,(x-1)\,\varphi(x,t) \qquad .
\end{aligned}$$

Es wurde $P_{-1}(t) = 0$ ausgenutzt. Die Anfangsbedingung Gl. (12.9) überträgt sich auf

$$\varphi(x,0) = 1 \qquad .$$

Die Lösung der Differentialgleichung zu dieser Anfangsbedingung ist leicht zu ermitteln

$$\varphi(x,t) = e^{\lambda\,(x-1)\,t} \qquad .$$

Mit Gl. (12.11) erhalten wir hieraus die gesuchten Wahrscheinlichkeiten

$$\begin{aligned}
P_n(t) &= \frac{1}{n!}\,\frac{d^n}{dx^n} e^{\lambda\,(x-1)\,t}\Big|_{x=0} \\
&= \frac{1}{n!}\,(\lambda\,t)^n\,e^{-\lambda\,t} \qquad .
\end{aligned}$$

Das ist genau die Poisson-Verteilung.

12.8 Shot-Noise

Wir betrachten folgendes physikalische Problem. Eine Lichtquelle emittiert zufällig Photonen mit einer mittleren Rate von λ Photonen pro Zeiteinheit. Die Photonen werden unabhängig von einander ausgesandt, und in jedem infinitesimalen Zeitintervall dt ist die Zahl der Photonen entweder 0 oder 1. Die Wahrscheinlichkeit, dass im Intervall dt ein Photon emittiert wird ist demnach

$$p = \lambda\,dt \qquad .$$

Wie groß ist die Wahrscheinlichkeit, dass in einem endlichen Intervall der Länge t die Zahl der emittierten Photonen k beträgt? Offensichtlich handelt es sich wieder um ein Bernoulli-Experiment. Die Zahl der Versuche ist $n = t/dt$ und die Wahrscheinlichkeit, dass in einem Versuch ein Photon emittiert wird ist $p = \lambda\ dt$. Die mittlere Zahl der Photonen ist $\mu = p\ n = \lambda\ t$. Da dt infinitesimal sein soll, geht $n \to \infty$ und $p \to 0$, wobei aber der Mittelwert μ konstant ist. Es sind also die Bedingungen für den Satz von Poisson erfüllt, und die gesuchte Wahrscheinlichkeit ist die Poisson-Verteilung

$$P(n|N = t/dt, p = \lambda dt) \underset{dt\to 0}{\longrightarrow} e^{-\lambda t}\frac{(\lambda t)^n}{n!} \qquad .$$

Das Shot-Rauschen ist also ebenfalls Poisson-verteilt.

12.9 Die Hartnäckigkeit des Pechs

Man hat oft das Gefühl, dass gerade die Warteschlange, in der man sich befindet, am langsamsten abgefertigt wird. Was kann die Wahrscheinlichkeitstheorie hierzu aussagen.

Wir nehmen an, x charakterisiert die Wartezeit. x_0 sei unsere Wartezeit, und x_i ($i = 1, \ldots$) die anderer Personen. Wir gehen davon aus, dass in Wirklichkeit niemand bevorzugt wird und die Variationen der Werte x_i alle denselben statistischen Fluktuationen unterworfen sind und sich nicht gegenseitig beeinflussen. D.h. sie sind i.u.v. gemäß einer Wahrscheinlichkeitsdichte $p(x|\mathcal{B})$, wobei der Bedingungskomplex die Ursache der Fluktuationen genauer spezifiziert. Wie groß ist nun die Wahrscheinlichkeit, dass erst die n-te Wartezeit (x_n) größer ist als x_0? Zu systematischen Behandlung definieren wir die Proposition

$$A_n : \text{erst der n-te Wert } x_n \text{ ist größer als } x_0 \qquad ,$$

Zur Berechnung der Wahrscheinlichkeit $P(A_n|\mathcal{B})$ verwenden wir die Marginalisierungsregel, um den Wert x_0 einzuführen

$$\begin{aligned} P(A_n|\mathcal{B}) &= \int p(A_n, x_0|\mathcal{B})\ dx_0 \\ &= \int P(A_n|x_0, \mathcal{B})\ p(x_0|\mathcal{B})\ dx_0 \qquad . \end{aligned}$$

Die Aussage A_n ist äquivalent dazu, dass die Wartezeit der nächsten $n - 1$ Personen

kleiner ist als x_0 und die der n-ten Person größer

$$\begin{aligned}
P(A_n|\mathcal{B}) &= \int P(x_i < x_0 (i = 1, \ldots, n-1),\ x_n > x_0|\mathcal{B})\ p(x_0|\mathcal{B})\ dx_0 \\
&= \int \prod_{i=1}^{n-1} \underbrace{P(x_i < x_0|\mathcal{B})}_{=:q(x_0)\text{ unabh. von i}}\ P(x_n > x_0|\mathcal{B})\ p(x_0|\mathcal{B})\ dx_0 \\
&= \int q(x_0)^{n-1}\ (1 - q(x_0))\ p(x_0|\mathcal{B})\ dx_0 \\
&= \int \left(\int \delta(r - q(x_0))\ dr \right) q(x_0)^{n-1}\ (1 - q(x_0))\ p(x_0|\mathcal{B})\ dx_0 \\
&= \int \left(\int \delta(r - q(x_0))\ \ p(x_0|\mathcal{B})\ dx_0 \right) r^{n-1}\ (1 - r)\ dr \\
&= \int \left(\int \frac{\delta(x_0 - q^{-1}(r))}{|q'(x_0)|}\ \ p(x_0|\mathcal{B})\ dx_0 \right) r^{n-1}\ (1 - r)\ dr
\end{aligned}$$

Nun ist $q(x_0) = P(x_i < x_0|\mathcal{B}) = \int_0^{x_0}\ p(x'|\mathcal{B})\ dx'$. Das heißt, $q'(x_0) = p(x_0|\mathcal{B})$ und das innere Integral vereinfacht sich zu

$$\int_0^1 \frac{\delta(x_0 - q^{-1}(r))}{|q'(x_0)|}\ p(x_0|\mathcal{B})\ dx_0 = \int_0^1 \delta(x_0 - q^{-1}(r))\ dx_0 = 1 \qquad .$$

Damit erhalten wir schließlich

$$\begin{aligned}
P(A_n|\mathcal{B}) &= \int_0^1 r^{n-1}\ (1 - r)\ dr \\
&= \frac{\Gamma(n)\Gamma(2)}{\Gamma(n+2)} = \frac{(n-1)!}{(n+1)!} \\
&= \frac{1}{n(n+1)} \qquad .
\end{aligned}$$

Man kann leicht überprüfen, dass die Wahrscheinlichkeit auf Eins normiert ist

$$\begin{aligned}
\sum_{n=1}^{\infty} P(A_n|\mathcal{B}) &= \sum_{n=1}^{\infty} \frac{1}{n(n+1)} \\
&= \lim_{L\to\infty} \sum_{n=1}^{L} \left(\frac{1}{n} - \frac{1}{n+1} \right) \\
&= \lim_{L\to\infty} \left(\sum_{n=1}^{L} \frac{1}{n} - \sum_{n=2}^{L+1} \frac{1}{n} \right) \\
&= \lim_{L\to\infty} \left(1 - \frac{1}{L+1} \right) \\
&= 1 \qquad .
\end{aligned}$$

Das verblüffende an dem Ergebnis ist

- es ist unabhängig von der Verteilung $p(x|\mathcal{B})$.
- Es existiert kein Mittelwert, $\langle n \rangle = \infty$.

Das heißt, im Mittel muss man unendlich lang warten, bis noch eine Warteschlange so lang ist wie unsere. Allerdings ist $P(A_1|\mathcal{B}) = \frac{1}{2}$, d.h. der Median ist $n = 1.5$.
Man kann dieses Ergebnis auch ganz anders begründen. Die Wartezeiten x_i $(i = 0, 1, \ldots, n)$ seinen zufällig i.u.v. erzeugt worden. Es gibt ein größtes Elemente x_g und ein zweit-größtes Element x_{zg}. Da die Zahlen unabhängig voneinander erzeugt wurden, sind alle Reihenfolgen der Elemente gleich-wahrscheinlich. Insgesamt gibt es $(n+1)!$ Ereignisse (Permutationen). Die günstigen Ereignisse sind solche, bei denen x_{zg} an der ersten und x_g an der letzten Stelle vorkommt. Auf die Reihenfolge der mittleren $(n-1)$ Elemente kommt es hierbei nicht an. Es gibt demnach $(n-1)!$ günstige Ereignisse und die Wahrscheinlichkeit ist nach der klassischen Definition

$$P = \frac{(n-1)!}{(n+1)!} = \frac{1}{n\,(n+1)} \qquad .$$

Dieses Ergebnis kann auf dem Computer leicht simuliert werden. Wir verwenden hierbei zwischen 0 und 1 gleich-verteilte Zufallszahlen.

- Zunächst wird die Referenz-Zufallszahl x_0 erzeugt.
- Danach werden solange Zufallszahlen erzeugt, bis eine größer ist als x_0.
- Das Experiment wird L-mal wiederholt.
- Hieraus ermitteln wir das arithmetische Mittel und die Standardabweichung.

Die Ergebnisse von $L = 1000$ solchen Simulationen sind in Abbildung 12.2 wiedergegeben. Man erkennt, dass weder der Mittelwert noch die Standardabweichung konvergieren, da ja beide für $L \to \infty$ divergieren.
Obgleich die Wahrscheinlichkeit dafür, dass bereits das nächste oder übernächste Element größer als das Referenzelement ist, bereits $2/3$ beträgt, gibt es doch nicht vernachlässigbare Wahrscheinlichkeit, dass sehr große Wartezeiten vorkommen können, wie man im Bild gut erkennen kann. Man beachte, dass die Wartezeiten logarithmisch aufgetragen sind. In diesem Beispiel ist das arithmetische Mittel 548 in die Standardabweichung beträgt 17151 und ist somit 30-mal größer als der Mittelwert.

12.10 Schätzen der Halbwertszeit aus einer Stichprobe

Gegeben sei eine Stichprobe von Zerfallszeiten $\{t_1, \ldots, t_L\}$. Die Wahrscheinlichkeit für einen Zerfall in $(t, t+dt)$ ist durch die Exponential-Verteilung

$$p(t|\tau) = \frac{1}{\tau}\, e^{-t/\tau}$$

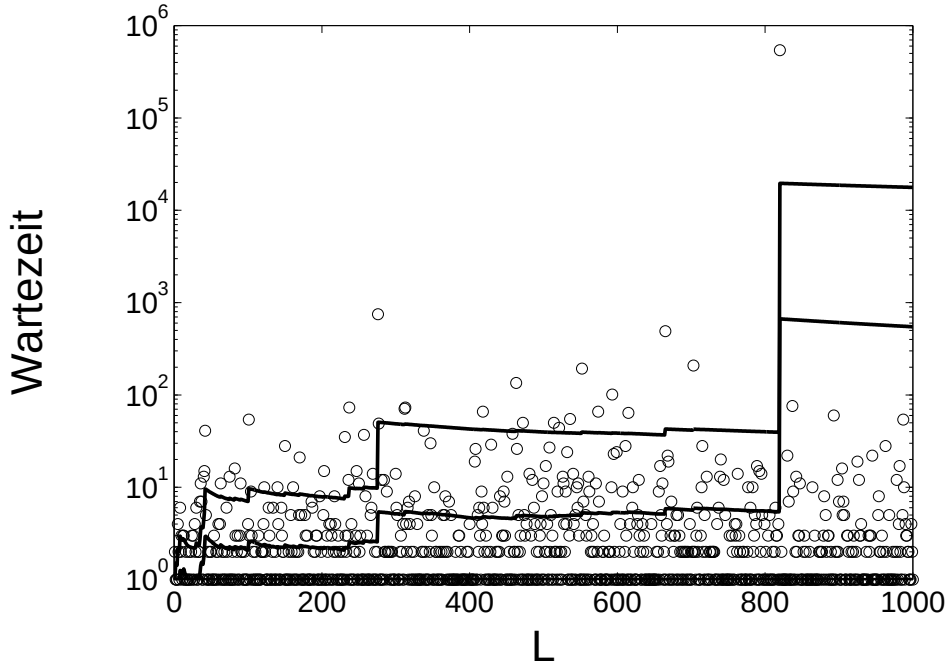

Abbildung 12.2: *Computersimulation der Wartezeiten, wie im Text beschrieben. Eingezeichnet ist zusätzlich das arithmetische Mittel und das Fehlerband.*

gegeben. Wir wollen die Wahrscheinlichkeitsdichte $p(\tau|t_1,\ldots,t_L,\mathcal{B})$ für die Relaxationszeit τ berechnen. Das Bayessche Theorem liefert den Bezug zur Likelihood

$$p(\tau|t_1,\ldots,t_L,\mathcal{B}) \propto p(t_1,\ldots,t_L|\tau,\mathcal{B})\; p(\tau|\mathcal{B}) \qquad .$$

Die Likelihood lautet, da die Zerfälle unkorreliert sind,

$$p(t_1,\ldots,t_L|\tau,\mathcal{B}) = \prod_{i=1}^{L} \left(\frac{1}{\tau}\; e^{-t_i/\tau}\right) = \tau^{-L}\; e^{-L\bar{t}/\tau} \qquad .$$

Hierbei ist $\bar{t}$ der Stichproben-Mittelwert der Zerfallszeiten. Der Prior ist, da es sich um eine Skalen-Variable handelt, Jeffreys' Prior. Somit lautet die Posterior-Verteilung

$$p(\tau|t_1,\ldots,t_L,\mathcal{B}) = \frac{1}{Z}\tau^{-L-1}\; e^{-L\bar{t}/\tau} \qquad .$$

Die Normierung lässt sich leicht berechnen

$$\begin{aligned} Z &= \int_0^\infty \tau^{-L}\; e^{-L\bar{t}/\tau}\; \frac{d\tau}{\tau} \\ &= \int_0^\infty \left(\frac{x}{L\bar{t}}\right)^L\; e^{-x}\; \frac{dx}{x} \\ &= \left(L\bar{t}\right)^{-L}\; \Gamma(L) \qquad . \end{aligned}$$

Der Erwartungswert dieser Verteilung lautet

$$\begin{aligned}\langle\tau\rangle &= \frac{1}{Z}\int_0^\infty \tau^{-L+1}\, e^{-L\bar{t}/\tau}\, \frac{d\tau}{\tau} \\ &= \frac{\left(L\bar{t}\right)^{-L+1}\,\Gamma(L-1)}{\left(L\bar{t}\right)^{-L}\Gamma(L)} \\ &= \frac{L}{L-1}\,\bar{t} \qquad .\end{aligned}$$

Die Varianz liefert analog

$$\begin{aligned}\mathrm{var}(\tau) &= \frac{\left(L\bar{t}\right)^{-L+2}\,\Gamma(L-2)}{\left(L\bar{t}\right)^{-L}\Gamma(L)} - \overline{\tau}^2 \\ &= \frac{L^2}{(L-1)(L-2)}\,\bar{t}^2 - \frac{L^2}{(L-1)(L-1)}\,\bar{t}^2 \\ &= \langle\tau\rangle^2\,\left(\frac{L-1}{L-2}-1\right) \\ &= \langle\tau\rangle^2\,\frac{1}{L-2} \qquad .\end{aligned}$$

Somit gilt

$$\tau = \bar{t}\,\frac{L}{L-1}\,\left(1 \pm \frac{1}{\sqrt{L-2}}\right) \qquad .$$

Teil III

Zuweisen von Wahrscheinlichkeiten

Kapitel 13

Vorbemerkungen

Die Regeln der Wahrscheinlichkeitstheorie besagen, wie man Wahrscheinlichkeiten umformen kann. Insbesondere sagt das Bayessche Theorem, dass man Posterior-Wahrscheinlichkeiten durch Likelihood und Prior

$$P(a|d,\mathcal{B}) = \frac{1}{Z}\, P(d|a,\mathcal{B})\; P(a|\mathcal{B}) \qquad .$$

ausdrücken kann. Die Likelihood-Funktionen sind trivialerweise quantitativ bekannt, da die Aussage des Bedingungskomplexes $\mathcal{B}$ beinhaltet, nach welcher Wahrscheinlichkeitsverteilung die experimentellen Daten verteilt sind. Die Proposition a legt zudem fest, welche Werte die Parameter der Wahrscheinlichkeitsverteilung besitzen.
Die Prior-Wahrscheinlichkeiten hingegen können nicht aus den bisher abgeleiteten Regeln der Wahrscheinlichkeitstheorie alleine berechnet werden.
Unbekannte Prior-Wahrscheinlichkeiten können auch auf komplexere Weise ins Spiel kommen, wenn nur ein Teil b der Parameter der Likelihood-Funktion bekannt ist und weitere Parameter a unbekannt sind. Die Marginalisierungsregel liefert dann

$$p(d|b,\mathcal{B}) = \int p(d|a,b,\mathcal{B})\; p(a|b,\mathcal{B})\; da \qquad . \tag{13.1}$$

Hierbei taucht wieder eine Prior-Wahrscheinlichkeit auf, die mit den Regeln der Wahrscheinlichkeitstheorie nicht weiter vereinfacht oder quantifiziert werden kann. Ein typisches Beispiel ist die Situation, dass die Fehler der Daten einer Gauß-Verteilung mit Mittelwert Null genügen, deren Varianz jedoch nicht bekannt ist. Die MARGINALE LIKELIHOOD $p(d|\mu = 0,\mathcal{B})$ kann dann über die Marginalisierungsregel bestimmt werden

$$p(\underline{d}|\mu = 0,\mathcal{B}) = \int_0^\infty p(\underline{d}|\sigma,\mu = 0,\mathcal{B})\; p(\sigma|\mu = 0,\mathcal{B})\; d\sigma \qquad . \tag{13.2}$$

Es bleibt aber zunächst der Prior für die Varianz zu spezifizieren. Generell bleibt ein Problem übrig, die Prior-Wahrscheinlichkeit von Parametern oder gar ganzen Wahrscheinlichkeitsverteilungen P_i oder Wahrscheinlichkeitsdichte $p(x)$ zu quantifizieren.

Es gibt drei Fälle zu unterscheiden

- **uninformative Prioren** (IGNORANT PRIORS) für Parameter, die außer der Definition des Problems keine Information enthalten.
- Exakte überprüfbare Information **testable information**. Z.B. Momente der Verteilung, $\Phi\{p(\mathbf{x}|I)\} = 0$
- Fehlerbehaftete überprüfbare Information

Wir werden uns hier auf die ersten beiden Fälle beschränken. Der letzte Fall führt auf die sogenannte QUANTIFIED MAXIMUM ENTROPY METHOD, die in den letzten Jahren sehr erfolgreich auf unterschiedlichste Arten von formfreier Rekonstruktion, z.B. in der Astro- und Plasmaphysik und in der Bilderkennung eingesetzt worden ist.
Wir hatten für Wahrscheinlichkeiten P_i diskreter Ereignisse bereits das Laplacesche Prinzip der Indifferenz kennengelernt, dass besagt, dass wenn die einzelnen Ereignisse AUSTAUSCHBAR sind, d.h. kein Index ausgezeichnet ist, dann sollten alle P_i gleich sein. Dieses Prinzip muss in zweierlei Hinsicht erweitert werden

1. Konsistente Berücksichtigung von exaktem Vorwissen, z.B.

$$\sum_i i^\nu \; P_i = \mu_\nu$$

2. Erweiterung auf kontinuierliche Freiheitsgrade $p(x)$.

Kapitel 14

Uninformative Prioren für Parameter

Wenn wir nichts über die Wahrscheinlichkeiten der verschiedenen Ausgänge eines Versuchs, sagen wir, $A_1, A_2, \ldots A_M$ wissen, dann können wir bei diskreten Problemen das Laplacesche Prinzip heranziehen.
Wie bereits mehrfach erwähnt, ist dieses Prinzip jedoch unbrauchbar im Fall kontinuierlicher Freiheitsgrade, wie wir am Beispiel des 1889 von Bertrand formulierten Paradoxons gesehen haben. Gesucht war hierbei die Wahrscheinlichkeit, eine „zufällige" Gerade im Abstand kleiner als der halbe Radius vom Kreiszentrum anzutreffen. Wir hatte hierfür unterschiedliche Werte $1/2, 1/3$ und $1/4$ erhalten, je nachdem, in welcher Variablen ein flacher Prior gewählt wurde.
Eine zufriedenstellende Antwort hierzu gelang erst 1973. E.T. Jaynes führte hierzu das Transformations-Invarianz-Prinzip (TIP) ein. Wenn eine Transformation T einer Zufalls-Variable x auf eine Variable x' existiert, die die Aufgabenstellung nicht ändert, und die zu neuen Variablen führt über die wir gleichermaßen wenig wissen, so darf diese Transformation die Wahrscheinlichkeitsdichte nicht ändern. Das heißt,

$$p_{\mathbf{x}}(\xi) = p_{\mathbf{x}'}(\xi) = p(\xi)$$

mit einer gemeinsamen Wahrscheinlichkeitsdichte $p(x)$. Andererseits gilt bei einer Transformation grundsätzlich

$$p_{\mathbf{x}}(\mathbf{x})d\mathbf{x} = p_{\mathbf{x}'}(\mathbf{x}')d\mathbf{x}' \Rightarrow p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}'}(T(\mathbf{x}))\left|\frac{\partial T}{\partial \mathbf{x}}\right| \quad . \tag{14.1}$$

Zusammen mit der Transformationsinvarianz folgt eine Bestimmungsgleichung für die Wahrscheinlichkeitsdichte

$$p(\mathbf{x}) = p(T(\mathbf{x}))\left|\frac{\partial T}{\partial \mathbf{x}}\right| \quad . \tag{14.2}$$

Diese Bedingung lässt sich besonders einfach über infinitesimale Transformationen T_ϵ erfüllen, wobei ϵ ein Vektor sein kann. Es muss zusätzlich zur Invarianzbedingung gelten,

$$\lim_{\epsilon \to 0} T_\varepsilon(\mathbf{x}) = x \qquad . \tag{14.3}$$

Wenn wir die infinitesimale Transformation in Gl. (14.2) einsetzen und nach ϵ_i ableiten, erhalten wir die Differentialgleichung

$$\frac{\partial}{\partial \epsilon_i} p(\mathbf{x})\Big|_{\epsilon=0} = \frac{\partial}{\partial \epsilon_i} \left(p(T_\epsilon(\mathbf{x})) \left| \frac{\partial T_\epsilon}{\partial \mathbf{x}} \right| \right) \Bigg|_{\epsilon=0} \quad . \tag{14.4}$$

Die Wahrscheinlichkeit $p(\mathbf{x})$ auf der linken Seite hängt nicht von ϵ ab, und somit verschwindet die Ableitung. Die rechte Seite ist für $\epsilon = 0$ auszuwerten

TIP-GLEICHUNG

$$\frac{\partial}{\partial \epsilon_i} \left[p(T_\epsilon(\mathbf{x})) \left| \frac{\partial T_\epsilon(x)}{\partial \mathbf{x}} \right| \right]_{\epsilon=0} = 0 \quad . \tag{14.5}$$

Man beachte, dass $\left|\frac{\partial T_\epsilon(x)}{\partial \mathbf{x}}\right|$ bei mehreren Variablen die Bedeutung der Jakobi-Determinante hat. Im Fall des Bertrand-Paradoxons gibt es drei Transformationen gegen die die Wahrscheinlichkeitsdichte invariant sein muss

- Rotation um das Kreiszentrum
- Verschiebung des gesamten Kreises
- Reskalierung des Radius R.

Diese Invarianz-Forderungen reichen aus, den Prior eindeutig festzulegen. Man findet

$$p(r) = \frac{1}{R}\, \theta(0 \leq r \leq R) \quad , \tag{14.6}$$

und dementsprechend ist die gesuchte Wahrscheinlichkeit $P = 1/2$.

14.1 Jeffreys' Prior für Skalen-Variablen

Wir wollen an einem einfacheren Beispiel das TIP detailliert vorführen. Wir betrachten eine Skalen-Variable, wie zum Beispiel die Standard-Abweichung. Hierbei ist a-priori z.B. nicht vorgeschrieben, in welchen Einheiten diese Größe anzugeben ist. Wenn es sich z.B. um Längen handelt, können wir sie in Meter, Angström, etc. angeben. Von dieser Wahl darf die Wahrscheinlichkeitsdichte demnach nicht abhängen. Darüber hinaus ist in diesem Beispiel auch nicht klar, ob der Prior für den Standardfehler oder für die Varianz σ^2 zu wählen ist. Wenn einer z.B. flach gewählt wird, ist es der andere nicht. Ein flacher Prior beschreibt in diesem Beispiel also nicht vollständiges Unwissen. Die Transformationen, gegen die der Prior einer Skalen-Variablen invariant sein muss, sind

$$x' = T_\varepsilon^{(1)}(x) = (1+\varepsilon)\, x \qquad x' = T_\varepsilon^{(2)}(x) = x^{(1+\varepsilon)} \quad .$$

Es wurde diese spezielle Form gewählt, damit die Einheitstransformation dem Fall $\varepsilon = 0$ entspricht. Für Gl. (14.5) benötigen wir im Fall der ersten Transformation

$$\left| \frac{\partial T_\epsilon^{(1)}}{\partial x} \right| = (1 + \varepsilon) \qquad .$$

Damit lautet die TIP-Gleichung

$$\begin{aligned} \frac{\partial}{\partial \epsilon} \left[p\left(T_\epsilon^{(1)}(x)\right) \; (1+\varepsilon) \right]_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} \left[p\left((1+\varepsilon)\, x\right) \; (1+\varepsilon) \right]_{\epsilon=0} \\ &= p'(x)\, x + p(x) \stackrel{!}{=} 0 \qquad . \end{aligned}$$

Die Lösung dieser Dgl. ist

$$p(x) = \frac{c}{x} \qquad .$$

Damit ist der Prior bereits festgelegt. Es bleibt also nur zu prüfen, ob die zweite Transformation zum selben Ergebnis führt, bzw. ob der Prior invariant gegen die Transformation ist. Strenggenommen sollte

$$p(x)\, dx = p(x')\, dx'$$

gelten, wobei $x' = x^\nu$ ist[1]. Wir müssen aber berücksichtigen, dass der eben abgeleitete Prior nicht normierbar ist. Also wird die Gleichung nur bis auf einen unbestimmten Normierungsfaktor gelten. Wir betrachten

$$\begin{aligned} p(x')dx' = p(x^\nu)d(x^\nu) &= \frac{c}{x^\nu}\, \nu\, x^{\nu-1}\, dx \\ &= \frac{c\nu}{x}\, dx \qquad . \end{aligned}$$

Damit hat der Prior in der Tat (bis auf die unbestimmte Normierung) die geforderte Invarianz. Somit ist der Prior für eine Skalen-Variable

JEFFREYS' PRIOR FÜR SKALEN-VARIABLEN	
$p(x) = \frac{1}{x} \qquad .$	(14.7)

Jeffreys' Prior hat, wie bereits erwähnt, die unschöne Eigenschaft, dass er nicht normierbar ist. Solche Prioren nennt man UNEIGENTLICH (IMPROPER). In praktischen Rechnungen wird man Cutoffs einführen und den eigentlichen Prior

[1] In der obigen Notation war $\nu = 1 + \varepsilon$. Hier ist diese Darstellung weniger geeignet.

EIGENTLICHER JEFFREYS' PRIOR

$$p_J(x) = \frac{1}{V_\sigma}\,\theta(x_u \le x \le x_o)\,\frac{1}{x} \qquad V_\sigma = \ln x_o/x_u \tag{14.8}$$

verwenden und erst im Endergebnis die Cutoffs eliminieren. Dadurch erhält man i.d.R. normierbare Posterior-Verteilung. Ist das nicht der Fall, sind die Daten in der Likelihood-Funktion nicht informativ genug.

Benford's Law

Benford hatte anhand von Tabellen und Zahlentafeln festgestellt, dass die führende Ziffer der Zahlen nicht gleich häufig vorkommen, sondern dass die Eins z.B. am häufigsten vorkommt. Das daraus abgeleitet Gesetz (BENFORD'S LAW) kann mit Jeffreys' Prior sofort hergeleitet werde. Wir berechnen die Wahrscheinlichkeit, dass die führende Ziffer einer beliebigen Zahl x den Wert $i \in (1, 2, 3, \ldots, 9)$[2] besitzt. Hierzu schreiben wir die Zahl als

$$x = d_0, d_1 d_2 \ldots \times 10^y \qquad \text{mit } d_0 > 0 \text{ und } y \in \mathbb{Z} \tag{14.9}$$

Das kann kann auch als $x = (d_0 + R) \times 10^y$ mit $R \in [0, 1)$ geschrieben werden. Die Wahrscheinlichkeit, dass die führende Ziffer den Wert d_0 hat, ist

$$P(d_0|I) = P(x \in [d_0, d_0 + 1)) = \frac{\int_{d_0}^{d_0+1} p(x)\,dx}{\int_1^{10} p(x)\,dx} = \frac{\int_{d_0}^{d_0+1} 1/x\,dx}{\int_1^{10} 1/x\,dx} = \log_{10}(\frac{d_0+1}{d_0}) \quad . \tag{14.10}$$

Da x eine Skalen-Variable ist, wurde Jeffreys' Prior verwendet. Hierbei kommt es nicht auf die Normierung an, und wir konnten ohne Cutoffs das Ergebnis direkt ermitteln. Wie in Abbildung 14.1 zu sehen, finden wir in der Tat, dass die niedrigen Ziffern wahrscheinlicher sind als die höheren.

14.2 Prior für die Parameter einer Geraden

Wir gehen von dem Problem aus, dass eine Gerade durch Punkte in der Ebene gelegt werden soll, wobei die Ungenauigkeit bei der Messung der Punkt in alle Raumrichtungen gleich sind $\sigma_x = \sigma_y = \sigma$, oder darüber nichts bekannt ist. Diese Information stecke im Bedingungskomplex $\mathcal{B}$. Wir suchen die Prior-Wahrscheinlichkeit

$$p(a, b|\mathcal{B})$$

[2] Führende Nullen werden gegebenenfalls eliminiert.

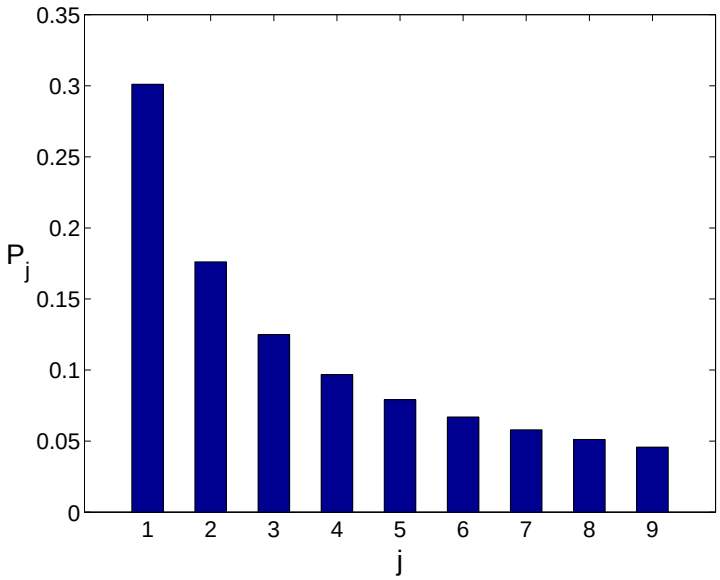

Abbildung 14.1: Benford's Law der Wahrscheinlichkeit der führenden Ziffern 1 to 9.

für die Parameter der Geradengleichung

$$y = a\ x + b \qquad .$$

Die Punkte $\mathbf{x} \in \mathbb{R}^2$ der Geraden erfüllen die Normalgleichung

$$\hat{n}^T \mathbf{x} = d \qquad .$$

Hierbei ist $\hat{n}$ der Einheitsvektor senkrecht zur Geraden und d ist der vorzeichenbehaftete und in Richtung $\hat{n}$ gemessene Abstand der Geraden vom Koordinatenursprung. In Abbildung 14.2 ist die Gerade skizziert. Der Einheitsvektor kann daraus abgelesen werden

$$\hat{n} = \begin{pmatrix} \sin(\Phi) \\ \cos(\Phi) \end{pmatrix} \qquad . \tag{14.11}$$

Auch liest man aus 14.2 leicht ab, dass

$$a = -\tan(\Phi) \tag{14.12}$$

$$b = \frac{d}{\cos(\Phi)} \qquad . \tag{14.13}$$

Das Problem ist invariant gegen Drehung und Verschiebung des Koordinatensystems. Hierbei geht ein beliebiger Punkt $\mathbf{x}$ über in

$$\mathbf{x}' = U\mathbf{x} + s \qquad ,$$

wobei die Matrix

$$U = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \tag{14.14}$$

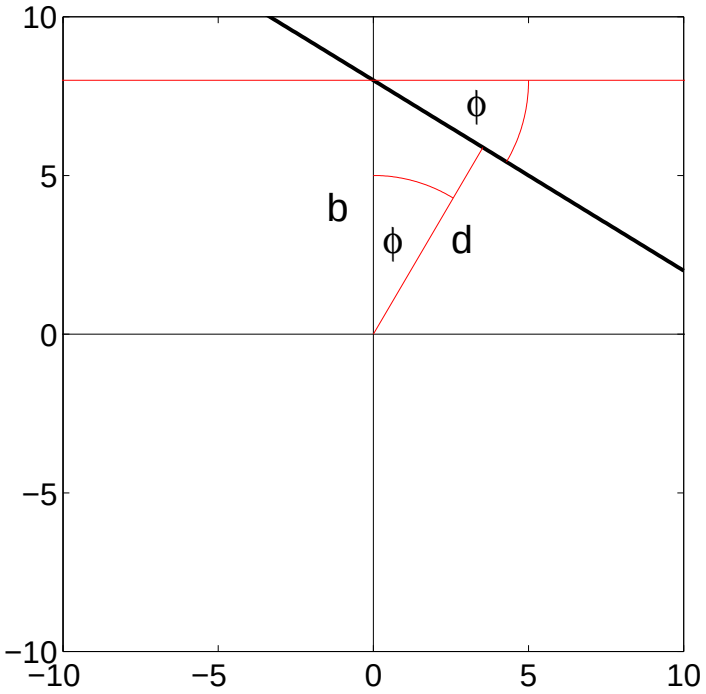

Abbildung 14.2: *Geraden-Problem*

eine Drehung am Koordinatenursprung bewirkt und $s \in \mathbb{R}^2$ den Verschiebungsvektor des Koordinatenursprungs darstellt. Die Geradengleichung im transformierten Koordinatensystem lautet

$$\begin{aligned} \hat{n}'^T \mathbf{x}' &= d' \\ \hat{n}'^T (U\mathbf{x} + s) &= d' \\ (\underbrace{U^T \hat{n}'}_{\hat{n}})^T \mathbf{x} &= \underbrace{d' - \hat{n}'^T s}_{d} \end{aligned} \quad .$$

Daraus folgt

$$\begin{aligned} \hat{n}' &= U\ \hat{n} \\ d' &= d + \hat{n}'^T s \end{aligned} \quad . \tag{14.15}$$

Mit Gl. (14.14) und Gl. (14.11) erhalten wir

$$\hat{n}' = \begin{pmatrix} \cos(\theta)\sin(\Phi) - \sin(\theta)\cos(\Phi) \\ \cos(\theta)\cos(\Phi) + \sin(\theta)\sin(\Phi) \end{pmatrix} = \begin{pmatrix} \sin(\Phi - \theta) \\ \cos(\Phi - \theta) \end{pmatrix} \quad .$$

Bei dieser Transformation gilt offenbar

$$\Phi \rightarrow \Phi' = \Phi - \theta \quad ,$$

wie man es bei einer Drehung auch erwarten würde. Die Transformation von d gemäß Gl. (14.15) ergibt

$$d \to d' = d + \Delta \qquad ,$$

hierbei kann Δ je nach Richtung und Länge von s beliebige Werte annehmen. Da das Problem invariant gegen beliebige Verschiebungen s ist, muss es auch invariant gegen beliebige Änderungen Δ sein.
Es ist offensichtlich sinnvoll zunächst die Gerade als Funktion von Φ und d zu formulieren und anschließend zur Standarddarstellung a, b überzugehen. Wir werden also zunächst $p(\Phi, d)$ ermitteln. Die infinitesimale Transformation ist somit

$$T_\varepsilon(\Phi, d) = \begin{pmatrix} \Phi' \\ d' \end{pmatrix} = \begin{pmatrix} \Phi - \varepsilon_\Phi \\ d + \varepsilon_d \end{pmatrix} \qquad .$$

Demnach ist

$$\frac{\partial \Phi'}{\partial \Phi} = 1 \qquad \frac{\partial \Phi'}{\partial d} = 0 \qquad \frac{\partial d'}{\partial \Phi} = 0 \qquad \frac{\partial d'}{\partial d} = 1 \qquad .$$

Das bedeutet, dass die Jakobi-Determinante Eins ist. Die TIP-Gleichungen werden also zu

$$\begin{aligned} \frac{\partial}{\partial \varepsilon_\Phi} p(\Phi - \varepsilon_\Phi, d + \varepsilon_d)\Big|_{\varepsilon=0} &= -\frac{\partial}{\partial \Phi} p(\Phi, d) = 0 \\ \frac{\partial}{\partial \varepsilon_d} p(\Phi - \varepsilon_\Phi, d + \varepsilon_d)\Big|_{\varepsilon=0} &= \frac{\partial}{\partial d} p(\Phi, d) = 0 \qquad . \end{aligned}$$

Daraus folgt

$$p_{\Phi,d}(\Phi, d) = const \qquad .$$

Die Rücktransformation auf die Größen a, b ist in Gl. (14.12) und Gl. (14.13) angegeben. Dazu benötigen wir gemäß

$$p_{a,b}(a, b) \left| \frac{\partial(a, b)}{\partial(\Phi, d)} \right| = p_{\Phi,d}(\Phi, d)$$

Die zugehörige Jakobi-Determinante lautet

$$\left| \frac{\partial(a, b)}{\partial(\Phi, d)} \right| = \left| \begin{pmatrix} \frac{\partial a}{\partial \Phi} & \frac{\partial a}{\partial d} \\ \frac{\partial b}{\partial \Phi} & \frac{\partial b}{\partial d} \end{pmatrix} \right| = \left| \begin{pmatrix} -\frac{1}{\cos^2(\Phi)} & 0 \\ \star & \frac{1}{\cos(\Phi)} \end{pmatrix} \right| = \frac{1}{\cos^3(\Phi)} \qquad .$$

Wegen $\cos^2 = 1/(1 + \tan^2)$ folgt somit

$$p(a, b) = c \left| \frac{\partial(a, b)}{\partial(\Phi, d)} \right|^{-1} = c \, (1 + a^2)^{-\frac{3}{2}} \qquad .$$

Dieser Prior ist wieder uneigentlich, da er in b nicht normierbar ist.

PRIOR DES GERADEN-PROBLEMS

$$p(a,b) = (1+a^2)^{-\frac{3}{2}} \tag{14.16}$$

Kapitel 15

Der entropische Prior für diskrete Probleme

Wir beschäftigen uns nun mit Wahrscheinlichkeitsverteilung P_i diskreter Ereignisse. Es soll exakte, überprüfbare Information (TESTABLE INFORMATION) vorliegen. Man nennt Nebenbedingungen überprüfbar, wenn es möglich ist festzustellen, ob gegebene P_i diese Nebenbedingung erfüllen oder nicht[1]. Es gibt mehrere Zugänge unterschiedlicher Strenge zum ENTROPISCHEN PRIOR

- Anschaulich, qualitativ
- Axiomatisch
- Über Konsistenzforderung.

Wir werden hier nur die ersten beiden Zugänge im Detail diskutieren.
Wir suchen die „uninformativste" Verteilung, die mit dem Vorwissen verträglich ist. „Uninformativ" soll bedeuten, dass wir uns damit am wenigsten festlegen, bzw. dass wir damit in den meisten Fällen richtig liegen. Nehmen wir z.B. einen Würfel mit den Nebenbedingungen

$$\begin{aligned} \mu_0 &= \sum_i q_i & &= 1 & &\text{Normierung} \\ \mu_1 &= \sum_i i\, q_i & &= 3.5 & &\text{1. Moment} \end{aligned} \quad .$$

Diese Nebenbedingungen sind durch

$$\{q_i\} = (\tfrac{1}{2}, 0, 0, 0, 0, \tfrac{1}{2})$$

erfüllt. Damit legen wir uns allerdings maximal fest, da nun definitiv immer eine Eins oder eine Sechs erscheinen sollte und niemals die anderen Seiten des Würfels. Das erscheint unsinnig, da auch die Wahrscheinlichkeiten

$$\{q_i\} = (\tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6})$$

[1] Das ist ganz immer Sinne von Karl Popper's Forderung, dass Theorien falsifizierbar sein müssen.

mit den Nebenbedingungen verträglich sind. In diesem Fall würden wir sicherlich diese Wahl im Sinne des Laplaceschen „principle of indifference" vorziehen. Mit dieser Wahl legen wir uns am wenigsten fest. Das bedeutet auch, dass der INFORMATIONSGEHALT DER WAHRSCHEINLICHKEITSVERTEILUNG minimal ist. Um Nebenbedingungen berücksichtigen zu können, die nicht mit der Gleichverteilung verträglich sind, benötigen wir ein Maß für die Information. Wir werden dann diejenige Wahrscheinlichkeitsverteilung $\{q_i\}$ wählen, die unter Berücksichtigung der Nebenbedingungen das Informationsmaß minimiert.

15.1 Shannon-Entropie: Informationsgehalt bei binären Fragen

Wieviele Fragen benötigt man, um einen Gegenstand aus $N = 2^L$ möglichen zu identifizieren[2]? Die Überlegungen hier haben motivierenden Charakter. Von daher ist die Einschränkung auf Zweierpotenzen unbedeutend.

- Nummerieren der Gegenstände durch $1, \ldots, N$.
- n sei die Nummer der gesuchten Gegenstandes.
- optimale Strategie, wenn alle Möglichkeiten gleich-wahrscheinlich sind:
  - Definiere die Menge $\mathcal{M} = \{1, \ldots, N\}$.

  a) Wenn $|\mathcal{M}| = 1 \Rightarrow$ es gibt nur eine Möglichkeit $\Rightarrow$ fertig!

  b) Teile die beiden Mengen in (annähernd) gleich großen Teilmengen $\mathcal{M}_1 = \{1, \ldots, \frac{N}{2}\}$ und $\mathcal{M}_2 = \{\frac{N}{2} + 1, \ldots, N\}$

  c) Frage ob der gesuchte Gegenstand in der ersten Teilmenge ist.
  wenn ja: $\mathcal{M} = \mathcal{M}_1$;
  wenn nein: $\mathcal{M} = \mathcal{M}_2$.
  gehe zu (a).
  - Man benötigt maximal $L = \log_2 N$ Fragen[3].

Alternative Begründung: Wähle binäre Darstellung von

$$N = \sum_{i=0}^{L} 2^i \, N_i \quad N_i \in \{0, 1\} \qquad .$$

Die n-te Frage lautet, *ist* $N_n = 1$? Somit benötigt man ebenfalls $L = \log_2 N$ Fragen. Damit kann $\log_2 N$ im Fall von N gleich wahrscheinlichen Ereignissen als Maß für die Ungewissheit interpretieren.

[2] Ableitung nach R.T.Cox: „The algebra of probable inference", Johns Hopkins univ. press, 1961
[3] Um Komplikationen zu vermeiden betrachten wir nur Zahlen die Zweierpotenzen sind.

Wir verallgemeinern nun diese Überlegungen auf Ereignisse mit unterschiedlichen Wahrscheinlichkeiten. Es gilt wieder, einen von N Gegenständen zu erraten. Die Gegenstände sind nun zu Gruppen zusammengefasst.
Beispiel: Es gebe N Lotterielose, die auf m Eimer verteilt sind. In jedem Eimer befinden sich $\frac{N}{m}$ Lose. Sowohl N als auch m soll eine Zweierpotenz sein.
Wir könnten zunächst erfragen, in welchem Eimer sich der Gewinn befindet. Dazu benötigen wir $L_1 = \log_2 m$ Fragen. Anschließend eruieren wir, welches Los in diesem Eimer der Gewinn ist, dazu benötigen wir zusätzlich $L_2 = \log_2 \frac{N}{m}$ Fragen.
Die Gesamtzahl ist somit

$$L_1 + L_2 = \log_2 m + \log_2 \frac{N}{m} = \log_2 N \qquad .$$

Das entspricht der Zahl, die wir ohne Unterteilung auch erhalten hätten. Diese ADDITIVITÄT werden wir auch im allgemeinen Fall fordern, wenn die Ereignisse nicht mehr alle gleich wahrscheinlich sind.
Wir definieren nun als Maß der Ungewissheit gerade die Anzahl der Binär-Fragen, die nötig ist, um die spezielle Aufgabenstellung Q bei einem gegebenen Bedingungskomplex $\mathcal{B}$ zu lösen.

$$U(Q|\mathcal{B}) : \text{Ungewißheit bzgl. } Q \text{ gegeben } \mathcal{B} \qquad .$$

Beispiel:

- $\mathcal{B}$: Es gebe N sich gegenseitig ausschließende Propositionen
- Q : Welche Proposition ist wahr?

Wir betrachten nun den Fall, dass die einzelnen Propositionen nicht gleichwahrscheinlich sind.
Wir wollen das an einem Beispiel diskutieren. Es gebe m Eimer mit Lotterielosen, in denen sich insgesamt N Lose befinden. Der Hauptgewinn ist durch Binär-Fragen zu ermitteln. Im Gegensatz zu vorher enthalten die Eimer unterschiedlich viele Lose. Im Eimer i seien n_i Lose (Zweierpotenz!). Es gilt natürlich

$$\sum_{i=1}^{m} n_i = N \qquad .$$

Die verwendeten Abkürzungen sind

- $\mathcal{B}$: *beschreibt diese Situation.*
- A_i : *der Gewinn befindet sich im Eimer i.*
- Q_1 : *Welches Los enthält den Gewinn?*

Es gilt

$$U(Q_1|\mathcal{B}) = \log_2 N$$
$$U(Q_1|A_i, \mathcal{B}) = \log_2 n_i \qquad .$$

Bei der optimalen Strategie benötigt man im ungünstigsten Fall $\log_2 N$ Binär-Fragen. Es gibt keine Strategie, die diese WORST-CASE Situation unterbietet.
Angenommen, wir wollen zunächst den Eimer ermitteln, in dem sich das richtige Los befindet und anschließend das Los. Der Gewinn soll sich im Eimer j befinden. Wir führen hierzu ein

- Q_2 : *Welcher Eimer enthält das richtige Los?*
- C : *Zuerst muss Q_2 beantwortet werden.*

Da die Eimer nun unterschiedlich viele Lose enthalten, ändert sich bei dieser Vorgehensweise die Zahl der Binär-Fragen.

$$U(Q_1|C, A_j, \mathcal{B}) = U(Q_2|\mathcal{B}) + U(Q_1|A_j, \mathcal{B}) \qquad .$$

A_j gibt an, in welchem Eimer sich der Gewinn befindet.
Beispiel: Die Anzahl der Eimer sei $m = 8$. Die Zahl der Lose in den Eimern sei

$$\{n_i\} = (2, 2, 2, 2, 2, 2, 4, 16) \qquad .$$

Insgesamt gibt es $N = 32$ Lose. Somit ist

$$U(Q_1|\mathcal{B}) = \log_2 32 = 5 \qquad .$$

Für $U(Q_1|C, \mathcal{B})$ erhalten wir

$$U(Q_1|C, A_j, \mathcal{B}) = U(Q_2|\mathcal{B}) + U(Q_1|A_j, \mathcal{B}) = \log_2 8 + \log_2 n_j = 3 + \log_2 n_j \qquad .$$

Wenn der Gewinn in den Eimern 1-6 ist ($j \in \{1, \ldots, 6\}$), dann gilt

$$U(Q_1|C, A_j, \mathcal{B}) = 3 + \log_2 2 = 3 + 1 = 4 \qquad ,$$

für $j = 7$ gilt

$$U(Q_1|C, A_j, \mathcal{B}) = 3 + \log_2 4 = 3 + 2 = 5$$

und für $j = 8$ haben wir

$$U(Q_1|C, A_j, \mathcal{B}) = 3 + \log_2 16 = 3 + 4 = 7 \qquad .$$

Im ungünstigsten Fall $j = 8$ benötigt man also 7 statt wie vorher 5 Fragen. Das bedeutet, die Additivität gilt nur bei Vorgabe von j. Allgemein gilt aber

$$U(Q_1|C, \mathcal{B}) \neq U(Q_2|\mathcal{B}) + U(Q_1|\mathcal{B}) \qquad .$$

Stattdessen fordern wir diese Additivität im Mittel.

$$\langle U(Q_1|C,\mathcal{B})\rangle = U(Q_2|\mathcal{B}) + \sum_{j=1}^{m} P(A_j|\mathcal{B})\, U(Q_1|A_j,\mathcal{B}) \qquad .$$

Der erste Summand liefert die Zahl der Binär-Fragen, die nötig sind, den richtigen Eimer herauszufinden. Der zweite Summand liefert die mittlere Zahl der Fragen, um dann noch das richtige Los zu finden. Gemittelt wird über die möglichen Eimer, in denen sich der Gewinn befinden kann.
Es gibt andererseits keine bessere Strategie[4], als alle Lose zu nummerieren und simultan zu bearbeiten. In diesem Fall ist die Ungewissheit $U(Q_1|\mathcal{B})$. Damit gilt also

$$\begin{aligned} \langle U(Q_1|C,\mathcal{B})\rangle \geq U(Q_1|\mathcal{B}) \quad &\Rightarrow \\ U(Q_2|\mathcal{B}) + \sum_{j=1}^{m} P(A_j|\mathcal{B})\, U(Q_1|A_j,\mathcal{B}) \geq U(Q_1|\mathcal{B}) \quad &\Rightarrow \\ U(Q_2|\mathcal{B}) \geq U(Q_1|\mathcal{B}) - \sum_{j=1}^{m} P(A_j|\mathcal{B})\, U(Q_1|A_j,\mathcal{B}) \qquad . \end{aligned}$$

Daraus folgt

$$\begin{aligned} U(Q_2|\mathcal{B}) &\geq \log_2 N - \sum_{j=1}^{m} P(A_j|\mathcal{B})\ \log_2 n_j \\ &\geq \sum_{j=1}^{m} P(A_j|\mathcal{B})\ \log_2 N - \sum_{j=1}^{m} P(A_j|\mathcal{B})\ \log_2 n_j \\ &\geq -\sum_{j=1}^{m} P(A_j|\mathcal{B})\ \log_2 \frac{n_j}{N} \qquad . \end{aligned}$$

Nun ist aber die Wahrscheinlichkeit, dass sich der Hauptgewinn im Eimer j befindet, $P(A_j|\mathcal{B}) = \frac{n_j}{N} =: p_j$, da alle Lose gleich-wahrscheinlich sind. Das liefert die Ungleichung

$$U(Q_2|\mathcal{B}) \geq -\sum_{j=1}^{m} p_j\ \log_2(p_j) \qquad .$$

Wegen $\log_2 x = \frac{\ln x}{\ln 2}$ steht somit auf der rechten Seite ein Ausdruck der proportional ist zur

SHANNON-ENTROPIE

$$S(\{p_i\}) = -\sum_{i=1}^{m} p_i \ln(p_i) \qquad . \tag{15.1}$$

[4]Wir hatten im Beispiel bereits gesehen, dass im Einzelfall, die Zahl der Fragen auf 7 ansteigt

Die Shannon-Entropie ist also ein Maß für die Ungewissheit. Man bezeichnet S auch als Informationsgehalt.

15.2 Eigenschaften der Shannon-Entropie

Da $0 \leq p_j \leq 1$ gilt für jeden Summanden $p_j \ln(p_j) \leq 0$. Daraus folgt

$$S \geq 0 \qquad .$$

Wir untersuchen für welche $\{p_j\}$ die Entropie (das Maß an Unsicherheit) unter der Nebenbedingung $\sum_j p_j = 1$ maximal ist

$$\frac{\partial}{\partial p_i}\left(S - \lambda \sum_{j=1}^{m} p_j\right) = -\ln(p_i) - 1 - \lambda \stackrel{!}{=} 0 \quad \Rightarrow$$
$$p_i = \frac{1}{m} \qquad .$$

Die zweite Ableitung liefert

$$\frac{\partial^2}{\partial p_i^2} S = -\frac{1}{p_i} < 0 \qquad .$$

Es handelt sich also in der Tat um das Maximum. Die Maximum-Entropie-Lösung (ME) ist also für den Fall, dass nur die Normierungsbedingung vorliegt,

$$p_i^{\mathrm{ME}} = \frac{1}{m} \qquad .$$

Somit hat die Entropie tatsächlich die gewünschte Eigenschaft, die wir im Würfel-Beispiel gefordert hatten. Der Wert der Entropie der Maxent-Lösung ist

$$S_{\mathrm{Max}} = -\sum_{j=1}^{m} \frac{1}{m} \ln(\frac{1}{m}) = \ln(m) \qquad .$$

Wenn als weitere Nebenbedingung der Mittelwert

$$\mu_1 = \sum_{j=1}^{m} j \; p_j$$

vorgegeben ist, so erhalten wir die Maxent-Lösung aus

$$\frac{\partial}{\partial p_l}\left(S + \lambda_0 \sum_{j=1}^{m} p_j + \lambda_1 \sum_{j=1}^{m} j\ p_j\right) = -\ln(p_l) - 1 + \lambda_0 + \lambda_1 l \quad \Rightarrow$$

$$p_l = e^{1+\lambda_0+\lambda_1 l}$$

$$p_l = \frac{1}{Z}\ q^l \qquad .$$

Wir haben neue Lagrange-Parameter Z und q eingeführt, die über die beiden Nebenbedingungen bestimmt werden müssen. Die Normierung liefert

$$Z = \sum_{i=1}^{m} q^i = q\ \frac{1-q^m}{1-q} \qquad .$$

Das erste Moment ist

$$\mu_1 = \frac{\sum_{i=1}^{m} i\ q^i}{Z} = \frac{q\frac{d}{dq}\sum_{i=1}^{m} q^i}{Z} = q\frac{d}{dq}\ \ln(Z)$$

$$= q\frac{d}{dq}\left(\ln(q) + \ln(1-q^m) - \ln(1-q)\right)$$

$$= \left(1 - m\frac{q^m}{1-q^m} + \frac{q}{1-q}\right) \qquad .$$

Für $q \to 1$ erhält man

$$\mu_1 = \frac{m+1}{2} \qquad .$$

Das entspricht genau dem ersten Moment, wenn alle Wahrscheinlichkeiten gleich sind, also für

$$p_i = \frac{q^i}{Z} = \frac{1}{m} \qquad .$$

Die Werte q zu gegebenem μ_1 müssen numerisch, z.B. mit dem Newton-Verfahren, ermittelt werden. Für einen Würfel ($L = 6$) und den Mittelwerten $\mu_1 = \{1.5, 2.5, 3.5, 4.5, 5.5\}$ sind die ME-Lösungen der Wahrscheinlichkeiten P_j in Abbildung 15.1 dargestellt. Man erkennt das exponentielle Verhalten zu einem der beiden Ränder hin, sowie die flache Verteilung für $\mu_1 = 3.5$.

15.3 Axiomatische Ableitung der Shannon-Entropie

Die folgende axiomatische Ableitung geht auf C. Shannon (1948) zurück. Auf diese Weise kommen wir von der unbefriedigenden Ungleichung weg. Das bedeutet aber auch, dass das Maß nicht exakt mit der Zahl der binären Fragen übereinstimmt.
Das Maß der Ungewissheit einer Wahrscheinlichkeitsverteilung $\underline{P} = \{P_1, \ldots, P_L\}$ soll folgende Axiome erfüllen:

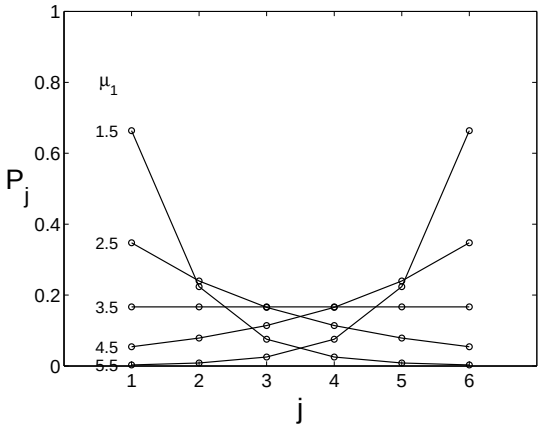

Abbildung 15.1: *Maxent-Lösungen der Wahrscheinlichkeiten P_j eines Würfels zu vorgegebenem Mittelwert μ_1.*

Eindeutigkeit Es existiert eine eindeutige Abbildung $H(\underline{P})$ der Wahrscheinlichkeitsverteilung auf die reellen Zahlen.

Stetigkeit Die Abbildung $H(\underline{P})$ ist stetig in allen P_j.
Kleine Änderungen in den P_j sollen nicht zu großen Änderungen in der Ungewissheit führen.

Monotonie $H(\{\frac{1}{L}, \ldots, \frac{1}{L}\})$ soll monoton mit L ansteigen.
Common sense: Wenn es mehr gleich-wahrscheinliche Propositionen gibt, steigt die Ungewissheit über den Ausgang des Experimentes.

Additivität

$$H(\{P_1, P_2, P_3, \ldots, P_L\}) = H(\{\overbrace{P_1 + P_2}, P_3, \ldots, P_L\}) + (P_1 + P_2)\ H(\{\frac{P_1}{P_1 + P_2}, \frac{P_2}{P_1 + P_2}\}) \quad .$$

Das entspricht der Aufteilung von Losen auf Eimer. Es soll für die Ungewissheit egal sein, ob man zuerst den Eimer und dann das Los in diesem Eimer bestimmt oder direkt das Los.
Hier können die $P_j \in \mathbb{R}$ sein, nicht mehr nur rationale Zahlen $\frac{n_i}{N}$.

Wir nutzen zunächst die Additivität aus. Dazu gehen wir von einer Grobunterteilung (Eimer) mit Wahrscheinlichkeiten P_j aus. Zu jedem Index j gibt es eine Feinunterteilung der Wahrscheinlichkeiten in

$$P_j \to \{P_{j,1}, \ldots, P_{j,L_j}\} \quad .$$

Es gelten die Summenregeln

$$\sum_{j=1}^{L} P_j = 1 \tag{15.2a}$$

$$\sum_{\nu=1}^{L_j} P_{j,\nu} = 1 \qquad . \tag{15.2b}$$

D.h., die Wahrscheinlichkeiten sind auch in jeder „Zelle" auf Eins normiert.
Aus der Additivität folgt

$$H(P_{1,1},\ldots,P_{1,n_1},\ldots,P_{L,1},\ldots,P_{L,n_L}) = \\ H(P_1,\ldots,P_L) + \sum_{j=1}^{L} P_j\, H(P_{j,1},\ldots,P_{j,n_j}) \qquad .$$

Der erste Term beschreibt die Ungewissheit der Grobunterteilung. Der zweite Term ist so zu verstehen; P_j ist die Wahrscheinlichkeit, dass das gesuchte Los sich im Eimer j befindet und $H(P_{j,1},\ldots,P_{j,L_j})$ ist dann die Ungewissheit dieses Eimers (Fragen die für diesen Eimer nötig sind).
Da $H(\underline{P})$ stetig sein soll, genügt es, nur rationale Zahlen

$$P_j = \frac{n_j}{\sum_j\ n_j}$$

zu betrachten, da sie jeder reellen Zahl beliebig nahe kommen. Mit der Abkürzung $N = \sum_j\ n_j$ haben wir dann

$$H(\underline{P}) = H(\tfrac{n_1}{N},\ldots,\tfrac{n_L}{N}) \qquad .$$

Wir unterteilen nun die Zelle j in n_j gleich-wahrscheinliche Teile, das heißt

$$P_j \to \{P_{j,1},\ldots,P_{j,n_j}\} = \{\underbrace{\tfrac{1}{n_j},\ldots\tfrac{1}{n_j}}_{n_j}\} \qquad .$$

Es wurde hierbei die Normierung Gl. (15.2b) berücksichtigt. Aufgrund der Additivität gilt

$$H(\underbrace{\tfrac{1}{N},\ldots\tfrac{1}{N}}_{n_1},\ldots,\underbrace{\tfrac{1}{N},\ldots\tfrac{1}{N}}_{n_L}) = H(P_1,\ldots,P_L) + \sum_{j=1}^{L} P_j\, H(\underbrace{\tfrac{1}{n_j},\ldots,\tfrac{1}{n_j}}_{n_j}) \qquad .$$

Die Einträge $1/n_j$ im Argument im letzten Ausdruck stehen deshalb, da dort normierte Wahrscheinlichkeiten einzutragen sind. Mit der Definition

$$h(m) = H(\underbrace{\tfrac{1}{m},\ldots,\tfrac{1}{m}}_{m}) \tag{15.3}$$

wird hieraus

$$h(N) = H(\underline{P}) + \sum_{j=1}^{L} P_j \; h(n_j) \qquad . \tag{15.4}$$

Nun wählen wir speziell $n_j = n$ mit

$$\frac{n}{N} = \frac{1}{M} \qquad \text{bzw.} \quad N = n\; M \qquad \text{und } P_j = \frac{1}{M} \qquad .$$

Damit erhalten wir

$$H(\underline{P}) = H(\{\tfrac{1}{M}, \ldots, \tfrac{1}{M}\}) = h(M)$$

und somit gilt

$$h(nM) = h(N) = h(M) + \underbrace{\sum_{j=1}^{L} P_j}_{=1} \; h(n)$$

$$h(nM) = h(M) + h(n) \qquad . \tag{15.5}$$

Hieraus folgt unmittelbar

$$h(1 \cdot M) = h(M) + h(1) \qquad .$$

Das heißt

$$h(1) = 0 \qquad . \tag{15.6}$$

Das ist sehr sinnvoll, denn wenn es nur eine Möglichkeit $n = 1$ gibt, ist die Ungewissheit Null. Wir betrachten noch die spezielle Wahl $n = M^l$, für die wir

$$\begin{aligned} h(M^{l+1}) &= h(M) + h(M^l) \quad \Rightarrow \\ h(M^l) &= l\; h(M) \end{aligned} \tag{15.7}$$

erhalten.
Zur Lösung der Funktional-Gleichung Gl. (15.5), die eigentlich für $n \in \mathbb{N}$ gelten soll, betrachten wir speziell $n = 1 + \varepsilon$ mit infinitesimalem ε. Wir werden anschließend prüfen, ob die so erhaltene Lösung auch für alle $n \in \mathbb{N}$ gilt und ob sie eindeutig ist. Im Rahmen dieser versuchsweisen Lösung nehmen wir auch an, dass die Ableitung von $h(x)$ existiert, dann erhalten wir aus Gl. (15.5)

$$h(M + M\varepsilon) = h(M) + h'(M)\varepsilon M \stackrel{!}{=}$$
$$h(M) + h(1 + \varepsilon) = h(M) + h(1) + h'(1)\varepsilon$$

$$\Rightarrow$$

$$\begin{aligned} h'(M)\varepsilon M &= h(1) + h'(1)\varepsilon \overset{Gl.(15.6)}{=} h'(1)\varepsilon \\ h'(M)M &= h'(1) =: \kappa \\ dh(M) &= \kappa \frac{dM}{M} \qquad . \end{aligned}$$

Die Lösung dieser Differentialgleichung liefert mit Gl. (15.6) ($h(1) = 0$)

$$h(M) = \kappa \ln(M) \qquad . \tag{15.8}$$

Es bleibt zu prüfen, ob die so gewonnene Gleichung tatsächlich die Funktionalgleichung Gl. (15.5) für natürliche Zahlen n, M erfüllt. Einsetzen der rechten Seite von Gl. (15.8) in Gl. (15.5) liefert

$$\kappa \ln(nM) \overset{?}{=} \kappa \left(\ln(M) + h(n)\right) \qquad .$$

Diese Gleichung wird offensichtlich von der Lösung Gl. (15.8) erfüllt.
Die Untersuchung der Eindeutigkeit der Lösung ist für natürliche n, M nicht so einfach wie für reelle. Wir betrachten hierzu die Primfaktorenzerlegung

$$M = \prod_{\nu=1}^{\infty} q_\nu^{m_\nu} \qquad ,$$

wobei q_ν alle möglichen Primzahlen sind und m_ν angibt, wie häufig die Primzahl q_ν in M vorkommt. Für die Primzahlen, die in der Zerlegung von M nicht vorkommen ist $m_\nu = 0$. Die Primfaktorenzerlegung ist eindeutig! Einsetzen in Gl. (15.5) und verwenden von Gl. (15.7) liefert

$$h(M) = h(\prod_\nu q_\nu^{m_\nu}) \overset{Gl.(15.5)}{=} \sum_\nu h(q_\nu^{m_\nu}) \overset{Gl.(15.7)}{=} \sum_\nu m_\nu \, h(q_\nu)$$

Wir führen die Abkürzung $h_\nu = h(q_\nu)$ ein. Für n schreiben wir analog die Primfaktorenzerlegung

$$n = \prod_{\nu=1}^{\infty} q_\nu^{n_\nu} \qquad .$$

Daraus folgt für Mn

$$Mn = \prod_\nu q_\nu^{m_\nu + n_\nu}$$

bzw.

$$h(Mn) = \sum_\nu (m_\nu + n_\nu) \, h_\nu$$

Das können wir weiter umformen

$$h(Mn) = \underbrace{\sum_\nu m_\nu \, h_\nu}_{h(M)} + \underbrace{\sum_\nu n_\nu \, h_\nu}_{h(n)} = h(M) + h(n)$$

Damit ist Gl. (15.5) für beliebige Zuweisungen $h_\nu = h(q_\nu)$ für die Primzahlen q_ν erfüllt.
Für die Eindeutigkeit benötigen wir Monotonie (Axiom 3). Es seien

$$t, s \in \mathbb{N}, \quad t, s > 1 \quad \Rightarrow \ln(t), \ln(s) > 0 \qquad .$$

Für hinreichend große n existiert ein M, so dass

$$\frac{n}{M} \leq \frac{\ln(t)}{\ln(s)} \leq \frac{n+1}{M} \qquad , \tag{15.9}$$

da sich jede reelle Zahl (hier $\frac{\ln(t)}{\ln(s)}$) durch rationale Zahlen der angegebenen Form einschachteln lässt. Wir multiplizieren obige Gleichung mit $M \ln(s)$ und erhalten unter Ausnutzung der Monotonie des Logarithmus

$$\begin{aligned} n \ln(s) \leq M \ln(t) &\leq (n+1) \ln(s) \\ s^n \leq t^M &\leq s^{n+1} \qquad . \end{aligned} \tag{15.10}$$

Nun soll nach Axiom 3 auch $h(n)$ eine monoton steigende Funktion sein. Das heißt für Gl. (15.10)

$$\begin{aligned} h(s^n) \leq h(t^M) \leq h(s^{n+1}) &\overset{Gl.(15.7)}{\Rightarrow} \\ nh(s) \leq Mh(t) \leq (n+1)h(s) &\quad \Rightarrow \\ \frac{n}{M} \leq \frac{h(t)}{h(s)} \leq \frac{n+1}{M} &\qquad . \end{aligned} \tag{15.11}$$

Aus Gl. (15.9) und Gl. (15.11) folgt

$$\begin{aligned} \frac{\ln(t)}{\ln(s)} &= \frac{n}{M} + \frac{\xi}{M} && \xi \in [0,1] \\ \frac{h(t)}{h(s)} &= \frac{n}{M} + \frac{\xi'}{M} && \xi' \in [0,1] \\ \Rightarrow \qquad \left| \frac{\ln(t)}{\ln(s)} - \frac{h(t)}{h(s)} \right| &= \frac{|\xi - \xi'|}{M} \leq \frac{1}{M} \qquad . \end{aligned}$$

Da die letzte Ungleichung für beliebige M gilt und M beliebig groß gewählt werden kann, folgt daraus

$$\begin{aligned} \frac{h(t)}{h(s)} = \frac{\ln(t)}{\ln(s)} &\quad \Rightarrow \\ \frac{h(t)}{\ln(t)} = \frac{h(s)}{\ln(s)} = \text{konst} &\quad \Rightarrow \\ h(s) = \text{konst } \ln(s) &\qquad . \end{aligned}$$

Aus Gl. (15.4) folgt dann

$$\begin{aligned} H(\underline{P}) &= h(N) - \sum_{j=1}^{L} P_j \; h(n_j) \\ &= \kappa \left(\ln(N) - \sum_{j=1}^{L} P_j \; \ln(n_j) \right) \\ &= -\kappa \sum_{j=1}^{L} P_j \; (\ln(n_j) - \ln(N)) \\ &= -\kappa \sum_{j=1}^{L} P_j \; \ln(\frac{n_j}{N}) = -\kappa \sum_{j} P_j \; \ln(P_j) \qquad \text{q.e.d.} \end{aligned}$$

Das Maß der Ungewissheit, das obige Axiome erfüllt, ist demnach proportional zur Shannon-Entropie S.
Wir benötigen nur die Wahrscheinlichkeitsverteilung, die dieses Funktional maximiert. Hierbei ist die Konstante κ irrelevant und statt H wird die Entropie maximiert. Deswegen heißt dieses Prinzip auch das MAXIMUM ENTROPIE (MAXENT) Prinzip.

15.4 Eigenschaften der Entropie

$$S = -\sum_{j=1}^{N} P_j \; \ln(P_j) \qquad .$$

- Invariant gegen Permutation der Indizes.
  Das war ja auch gewollt, da ohne Nebenbedingungen kein Index ausgezeichnet sein soll.
- Es besteht zwischen den Wahrscheinlichkeiten P_i und P_j keine Korrelation. Die Hesse-Matrix ist diagonal.
- $-p \ln(p)$ ist eine stetige Funktion im Intervall $0 \leq p \leq 1$
- $S \geq 0$, da alle Summanden nicht negativ sind.
- $S = 0 \;\Leftrightarrow\; P_j = \delta_{jj_0}$ Entspricht Gewissheit.
- S mit der Normierungsbedingung ist maximal für $P_j = \frac{1}{N}$
- S ist unabhängig von unmöglichen Propositionen $P_j = 0$.
- S ist konvex
  $$\frac{\partial^2 S}{\partial P_j \partial P_j} = -\delta_{ij} \; \frac{1}{P_i} \qquad .$$
  Das heißt es gibt nur ein Maximum.

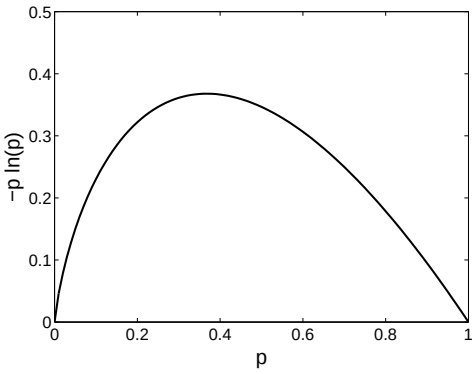

Abbildung 15.2: *Summand zur Shannon-Entropie*

15.5 MaxentPrinzip

E.T.Jaynes schlug 1963[5] vor, Prior-Wahrscheinlichkeiten bei gegebenen überprüfbaren Informationen durch Maximieren der Entropie unter Berücksichtigung der Nebenbedingungen zu bestimmen. 1984 gelang es Shore und Johnson[6] aus Konsistenz-Überlegungen sowohl das Entropie-Funktional als auch das Maxent-Prinzip herzuleiten.
Wir werden uns hier nur mit solchen Nebenbedingungen beschäftigen, die sich in Gleichungen ausdrücken lassen. Es ist auch möglich, Ungleichungen zu inkorporieren.
Wir betrachten überprüfbare Information in der Form von Gleichungen

$$\varphi_\mu\{P_j\} = 0, \quad \mu = 1, \ldots, L \qquad , \tag{15.12}$$

wobei L die Zahl der Nebenbedingungen angibt. Neben diesen Bedingungen ist in allen Fällen die Normierung

$$\sum_{j=1}^{N} P_j - 1 = 0 \tag{15.13}$$

zu erfüllen. Das Maxent-Prinzip besagt, dass die Entropie unter Berücksichtigung der Nebenbedingungen zu maximieren ist. Die Nebenbedingungen können am bequemsten über den Formalismus der Lagrange-Parameter berücksichtigt werden. Wir definieren die Lagrange-Funktion

$$\mathcal{L} = -\sum_j P_j \ln(P_j) + \lambda_0 \,(\sum_j P_j - 1) + \sum_\mu \lambda_\mu \varphi_\mu\{P_j\} \qquad . \tag{15.14}$$

Die Maxent-Lösung erhält man aus der Nullstelle der Ableitungen nach P_j und die

[5] E.T. Jaynes, papers on prob., stat. and stat.phys., academic press (83)
[6] Shore and Johnson, IEEE trans.inf.theory,26,1,84

Nebenbedingungen aus den Ableitungen nach den Lagrange-Parametern.

$$\begin{aligned}
\Psi_i &= \frac{\partial}{\partial P_i}\,\mathcal{L} \\
&= \frac{\partial}{\partial P_i}\left(-\sum_j P_j \ln(P_j) + \lambda_0\,(\sum_j P_j - 1) + \sum_\mu \lambda_\mu \varphi_\mu\{P_j\}\right) \\
\Psi_i &= -\ln(P_i) - 1 + \lambda_0 + \sum_\mu \lambda_\mu \frac{\partial}{\partial P_i}\,\varphi_\mu\{P_j\} \stackrel{!}{=} 0 \qquad (15.15) \\
\ln(P_i) &= -1 + \lambda_0 + \sum_\mu \lambda_\mu \frac{\partial}{\partial P_i}\,\varphi_\mu\{P_j\} \qquad .
\end{aligned}$$

Die Lösung lautet also allgemein

MAXIMUM-ENTROPIE-LÖSUNG

$$P_i = \frac{1}{Z}\, e^{\sum_\mu \lambda_\mu \frac{\partial}{\partial P_i} \varphi_\mu\{P_j\}} \qquad (15.16a)$$

$$Z = \sum_{j=1}^{N} e^{\sum_\mu \lambda_\mu \frac{\partial}{\partial P_i} \varphi_\mu\{P_j\}} \qquad . \qquad (15.16b)$$

Die Normierung wurde bereits explizit über die „Zustandssumme" Z sichergestellt. Die übrigen Lagrange-Parameter λ_ν sind so zu bestimmen, dass die Nebenbedingungen Gl. (15.12) erfüllt sind.

Von besonderer Bedeutung sind lineare Nebenbedingungen

$$\varphi_\mu\{P_j\} = \sum_{j=1}^{N} K_{\mu j}\, P_j \;-\; \kappa_\mu \qquad . \qquad (15.17)$$

Die in Gl. (15.16a) und Gl. (15.16b) benötigten Ableitungen lassen sich dann leicht berechnen

$$\frac{\partial}{\partial P_i}\,\varphi_\mu\{P_j\} = K_{\mu i} \qquad (15.18)$$

und wir erhalten

MAXIMUM-ENTROPIE LÖSUNG BEI LINEAREN NEBENBEDINGUNGEN

$$P_j = \frac{1}{Z}\, e^{\sum_\mu \lambda_\mu K_{\mu j}} \tag{15.19a}$$

$$Z = \sum_{j=1}^{N} e^{\sum_\mu \lambda_\mu K_{\mu j}} \quad . \tag{15.19b}$$

In diesem Fall lassen sich die zusätzlichen Nebenbedingungen ebenfalls elegant formulieren

$$\begin{aligned} \kappa_\mu &= \sum_{j=1}^{N} K_{\mu j}\, P_j \\ &= \frac{1}{Z} \sum_{j=1}^{N} K_{\mu j}\, e^{\sum_\mu \lambda_\mu K_{\mu j}} \\ &= \frac{\partial}{\partial \lambda_\mu} \ln(Z) \quad . \end{aligned}$$

- Formale Ähnlichkeit zur stat. Mechanik
- erweiterbar auf (Nichtgleichgewichts)-Thermodynamik[7]
- weitere publizierte Anwendungen[8]:
  Logistische Probleme, Volkswirtschaft, queuing theory, nichtlin. Spektralanalyse, Mustererkennung, Bildverarbeitung, Tomographie, Flüssigkeitsdynamik, Wachstumsdynamik, ...

Es stellt sich die Frage, ob es mehrere Lösung geben kann. Die Maxent-Lösung folgt aus der Nullstelle von Ψ_i, der partiellen Ableitung der Lagrange-Funktion nach P_i. Wir bestimmen die zweite partielle Ableitung der Lagrange-Funktion, bzw. die partiellen Ableitung von Ψ_i Gl. (15.15) mit Gl. (15.18)

$$\frac{\partial^2 \mathcal{L}}{\partial P_i \partial P_j} = \frac{\partial \Psi_i}{\partial P_j} = \frac{\partial^2 S}{\partial P_i \partial P_j} = -\delta_{ij}\, \frac{1}{P_i} \quad .$$

Die Lagrange-Funktion ist also global konvex und besitzt von daher nur ein (globales) Maximum. Alternativ kann man auch sagen, Ψ_i hängt nur von P_i ab und ist in dieser Variablen monoton fallend; also gibt es nur eine Nullstelle.
Wir wollen nun einige Beispiele diskutieren.

[7] W.T.Grandy, Foundations of stat.mech. I/II D.Reidel Publ., Kluwer (87)
[8] J.Kapur, Entropy Optimization Principle, academic Press (92)

15.6 Maxwell-Boltzmann-Verteilung

Die einzige Nebenbedingung neben der Normierung sei

$$\langle E \rangle = \sum_j P_j \, E_j \qquad . \tag{15.20}$$

Das heißt, jedem Ereignis wird ein Kostenwert (Energie) zugeordnet und als überprüfbare Information sind die mittleren Kosten bekannt. Das sind genau die Voraussetzungen, die der Maxwell-Boltzmann-Verteilung in der statistischen Mechanik zugrunde liegen. Die Größe $K_{\mu j}$ in Gl. (15.17) ist

$$K_{\mu j} = E_j, \quad \mu \equiv 1 \qquad .$$

Damit lautet die Maxent-Lösung nach Gl. (15.19a) und Gl. (15.19b)

MAXWELL-BOLTZMANN-VERTEILUNG

$$P_i = \frac{1}{Z} \, e^{-\beta \, E_i}$$
$$Z = \sum_j \, e^{-\beta \, E_j} \qquad .$$

Der Lagrange-Parameter β folgt aus der Nebenbedingung

$$-\frac{\partial \ln(Z)}{\partial \beta} \stackrel{!}{=} \langle E \rangle \qquad .$$

Die Lösung ist, wie oben allgemein gezeigt, eindeutig. Es existiert jedoch nur dann eine Lösung, wenn

$$E_{\min} \leq \langle E \rangle \leq E_{\max} \qquad .$$

Andere Mittelwerte sind nicht mit dem Modell kompatibel.
Wenn man experimentell Abweichungen hiervon findet,

$$S^{Maxent} > S^{\exp} \qquad ,$$

dann heißt das, dass weitere Nebenbedingungen existieren.

Anwendungsbeispiel aus der Flüssigkeitsdynamik:

P_i sei das Verhältnis von Tropfen der Masse m_i an der Gesamtmasse m, die pro Zeiteinheit erzeugt werden, wenn Flüssigkeit mit einer Rate $\dot{m}$ und einer konstanten

Tropfenrate $\dot{n}$ aus einer Düse heraustritt. Die Nebenbedingung (neben der Normierung) erhalten wir folgendermaßen. Die Flüssigkeitsmasse Δm, die in der Zeiteinheit Δt austritt, enthält Δn_j Tropfen der Masse m_j und beträgt demnach

$$\sum_j \Delta n_j \; m_j = \Delta m \qquad .$$

Die Teilchenzahl Δn_j erhalten wir aus der gesamten Teilchenzahl Δn, die pro Zeiteinheit austritt, über die Beziehung

$$\Delta n_j = \Delta n \; P_j \qquad .$$

Damit haben wir

$$\sum_j P_j \; m_j \; \Delta n = \Delta m \qquad .$$

Division durch Δt liefert schließlich

$$\dot{n} \sum_j P_j \; m_j = \dot{m} \qquad .$$

Die Maxent-Lösung lautet dann

$$P_i = \frac{1}{Z} \; e^{-\lambda \dot{n} m_i}$$

15.7 Bose-Einstein-Verteilung

Wir nehmen nun an, dass wir nicht genau wissen, wieviele Teilchen sich in einem betrachteten Volumen befinden. Als überprüfbare Information sei weiterhin die mittlere Energie bekannt. Hinzukommt als Nebenbedingung die mittlere Teilchenzahl.
P_{in} sei die Wahrscheinlichkeit, n Teichen im Zustand i mit der Energie E_i anzutreffen. Die Normierungs-Nebenbedingungen lauten

$$\sum_{n=0}^{\infty} P_{in} = 1 \quad \forall i \qquad . \tag{15.21}$$

Die mittlere Zahl (BESETZUNGSZAHL) an Teilchen im Zustand i ist

$$n_i = \sum_n n \; P_{in} \qquad . \tag{15.22}$$

Die mittlere Teilchenzahl ist damit

$$\sum_j \sum_{n=0}^{\infty} n \; P_{jn} = \langle N \rangle \qquad . \tag{15.23}$$

Schließlich lautet die letzte Nebenbedingung

$$\sum_j E_j \; n_j = \sum_j E_j \sum_{n=0}^{\infty} n \; P_{jn} = \langle E \rangle \qquad . \tag{15.24}$$

Die Entropie sieht nun leicht anders aus, da die Ereignisse durch ein Index-Paar (in) gekennzeichnet sind

$$S = -\sum_{jn} P_{jn} \; \ln(P_{jn}) \qquad . \tag{15.25}$$

Die Lagrange-Funktion lautet

$$\mathcal{L} = S - \sum_j \lambda_{0,j} (\sum_m P_{jm} - 1) - \lambda_1 \; (\sum_{jm} m \; P_{jm} - \langle N \rangle) - \lambda_2 \; (\sum_{jm} E_j \; m \; P_{jm} - \langle E \rangle) \qquad .$$

Die Ableitung nach P_{in} liefert

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial P_{in}} &= -\ln P_{in} - 1 - \lambda_{0,i} - \lambda_1 \; n - \lambda_2 \; n \; E_i \stackrel{!}{=} 0 \Rightarrow \\ P_{in} &= \frac{1}{Z_i} \; e^{-n(\lambda_1 + \lambda_2 E_i)} \qquad . \end{aligned} \tag{15.26}$$

Die Normierung Gl. (15.21) ergibt

$$Z_i = \sum_{n=0}^{\infty} e^{-n(\lambda_1 + \lambda_2 E_i)} = \sum_{n=1}^{\infty} \left(e^{-(\lambda_1 + \lambda_2 E_i)} \right)^n = \frac{1}{1 - e^{-(\lambda_1 + \lambda_2 E_i)}} \qquad .$$

Daraus folgt

$$P_{in} = \left(1 - e^{-(\lambda_1 + \lambda_2 E_i)} \right) \; e^{-n \; (\lambda_1 + \lambda_2 E_i)} \qquad . \tag{15.27}$$

Daraus berechnen wir die Besetzungszahl

$$\begin{aligned} n_i &= \sum_{n=0}^{\infty} n \; P_{in} \\ &= \left(1 - e^{-(\lambda_1 + \lambda_2 E_i)} \right) \; \sum_{n=0}^{\infty} n \; e^{-n \; (\lambda_1 + \lambda_2 E_i)} \\ &= -\left(1 - e^{-\lambda} \right) \; \frac{\partial}{\partial \lambda} \sum_{n=0}^{\infty} e^{-n \; \lambda} \Big|_{\lambda = \lambda_1 + \lambda_2 E_i} \\ &= -\left(1 - e^{-\lambda} \right) \; \frac{\partial}{\partial \lambda} \frac{1}{1 - e^{-\lambda}} \Big|_{\lambda = \lambda_1 + \lambda_2 E_i} \\ &= \left(1 - e^{-\lambda} \right) \; \frac{e^{-\lambda}}{(1 - e^{-\lambda})^2} \Big|_{\lambda = \lambda_1 + \lambda_2 E_i} \\ &= \frac{1}{e^{\lambda_1 + \lambda_2 E_i} - 1} \qquad . \end{aligned}$$

Mit Symbolen für die Lagrange-Parameter, die in der statistischen Physik üblich sind, erhalten wir

BESETZUNGSZAHLEN DER BOSE-EINSTEIN-VERTEILUNG

$$n_i = \frac{1}{e^{\beta(E_i-\mu)} - 1} \quad . \tag{15.28}$$

Die Parameter β und μ folgen aus den Nebenbedingungen für $\langle E \rangle$ und $\langle N \rangle$.

Anwendung in der BWL

Es werden drei Produkte mit den Kosten (in beliebigen Einheiten)

$$K_i = \{50, 7.5, 1\}$$

hergestellt. Die Kosten entsprechen den Energien E_i. Pro Monat sei die mittlere Zahl der Produkte

$$\langle N \rangle = 275$$

und die mittleren Kosten betragen

$$\langle K \rangle = 829 \quad .$$

Wenn das die einzige Information ist, wie ist dann die mittlere Zahl der einzelnen Produkte, d.h. die Besetzungszahl? Die Maxent-Lösung liefert

$$n_i = \{6, 40, 229\} \quad .$$

15.8 Fermi-Dirac-Verteilung

Der einzige Unterschied zu den Voraussetzungen der BE-Verteilung ist die Bedingung, dass in jedem Zustand i nur 0 oder 1 Teilchen sein kann. Da ansonsten alles gleich bleibt, erhalten wir dieselbe allgemeine Lösung wie in Gl. (15.26). Die Normierung ist nun aber anders zu berechnen

$$Z_i = \sum_{n=0}^{1} e^{-n(\lambda_1+\lambda_2 E_i)} \qquad = 1 + e^{-(\lambda_1+\lambda_2 E_i)} \quad .$$

$$P_{in} = \frac{1}{1 + e^{-(\lambda_1+\lambda_2 E_i)}} \, e^{-n\,(\lambda_1+\lambda_2 E_i)} \quad . \tag{15.29}$$

Die Besetzungszahl liefert

$$\begin{aligned} n_i &= \frac{\sum_{n=0}^{1} n\, e^{-n\,(\lambda_1+\lambda_2 E_i)}}{1+e^{-(\lambda_1+\lambda_2 E_i)}} \\ &= \frac{e^{-\,(\lambda_1+\lambda_2 E_i)}}{1+e^{-(\lambda_1+\lambda_2 E_i)}} \\ &= \frac{1}{e^{(\lambda_1+\lambda_2 E_i)}+1} \qquad . \end{aligned}$$

Mit den in der statistischen Physik üblichen Symbolen liefert das

BESETZUNGSZAHLEN DER FERMI-DIRAC-VERTEILUNG

$$n_i = \frac{1}{e^{\beta(E_i-\mu)}+1} \qquad . \tag{15.30}$$

15.9 Vergleich mit Zufallsexperiment

Die Anwendung des Maxent-Prinzips ist nicht darauf angewiesen, dass die gesuchte Verteilung das Ergebnis eines Zufallsexperiments ist. Falls es aber ein Zufallsexperiment gibt, sollte es eine enge Beziehung zum Maxent-Ergebnis geben.

Maxent

Eine Zufalls-Variable x soll die Werte $\{\xi_1, \ldots, \xi_N\}$ annehmen können. Die überprüfbare Information für die Wahrscheinlichkeiten P_i, mit denen diese Werte angenommen werden, möge der Einfachheit halber lineare Nebenbedingungen liefern

$$\kappa_\mu = \sum_i P_i\, K_\mu(\xi_i), \quad \mu = 1, \ldots, L \qquad .$$

Die Wahrscheinlichkeitsverteilung, die diese Nebenbedingungen erfüllt, sonst aber frei von weiteren Bedingungen ist, ist die Maxent-Verteilung.

Zufallsexperiment

Die Werte von x werden in einem Zufallsexperiment ermittelt. Das Zufallsexperiment liefere die Stichprobe

$$\zeta_1, \ldots, \zeta_M, \quad \text{mit } \zeta_l \in \{\xi_1, \ldots, \xi_N\} \qquad .$$

Die Nebenbedingung fordert

$$\kappa_\mu = \frac{1}{M} \sum_{l=1}^{M} K_\mu(\zeta_l), \quad \mu = 1, \ldots, L \qquad .$$

Wir ermitteln die Häufigkeit m_i, mit der die einzelnen Werte ξ_i in der Stichprobe vorkommen und sortieren die Summe nach den angenommenen Werten ξ_i um

$$\kappa_\mu = \frac{1}{M} \sum_{i=1}^{N} m_i \, K_\mu(\xi_i) \qquad .$$

Die Nebenbedingungen lauten dann

$$\begin{aligned} \sum_{i=1}^{N} m_i &= M \\ \sum_{i=1}^{N} m_i \, K_\mu(\xi_i) &= M \, \kappa_\mu, \quad \mu = 1, \ldots, L \qquad . \end{aligned} \tag{15.31}$$

Wie wird die Häufigkeitsverteilung m_i aussehen, die in zukünftigen Zufallsexperimenten am häufigsten vorkommen wird? Dazu müssen wir berücksichtigen, dass die Häufigkeitsverteilung aus den Stichproben $\zeta_1, \ldots, \zeta_M$ ermittelt wird. Die Elemente der Stichprobe sollen aber unkorreliert sein, da nichts anderes bekannt ist. Das ist eine wichtige Annahme! Wenn etwas über die Korrelation bekannt ist, muss das berücksichtigt werden. Erlaubte Stichproben sind solche, deren Häufigkeitsverteilung die Nebenbedingungen Gl. (15.31) erfüllt. Die Anzahl der Sequenzen ist durch den Multinomialkoeffizienten gegeben

$$N(\underline{m}, M) = \frac{M!}{\prod_i \, m_i} \qquad .$$

Unter allen Häufigkeitsverteilungen $\underline{m}$, die mit dem Vorwissen verträglich sind, kommen diejenigen bei wiederholten Zufallsexperimenten am häufigsten vor, für die $N(\underline{m}, M)$ am größten ist. Das bedeutet, dass die wahrscheinlichste Häufigkeitsverteilungen $\underline{m}$ aus der Variationsrechnung

$$\max_{\underline{m}} \, \ln\left(N(\underline{m}, M)\right) \quad \& \quad \text{Nebenbedingungen} \qquad . \tag{15.32}$$

Wenn

$$M, m_i \gg 1 \qquad ,$$

dann kann die Stirling-Formel Gl. (4.5b) auf die Fakultäten angewandt werden

$$\ln(m_i!) \simeq m_i \ln(m_i) - m_i + \ln(\sqrt{2\pi}) \qquad .$$

Daraus ergibt sich

$$\begin{aligned}
\ln(N(\underline{m}, M)) &\simeq M\ln(M) - M - \sum_{i=1}^{N} (m_i \ln(m_i) - m_i) \\
&= M\ln(M) - M - \sum_{i=1}^{N} m_i \ln(m_i) + \underbrace{\sum_{i=1}^{N} m_i}_{=M} \\
&= M\ln(M) - \sum_{i=1}^{N} m_i \ln(m_i) \\
&= \sum_{i=1}^{N} m_i \ln(M) - \sum_{i=1}^{N} m_i \ln(m_i) \\
&= -M \sum_{i=1}^{N} \frac{m_i}{M} \ln(\frac{m_i}{M}) \\
&= -M \sum_{i=1}^{N} \tilde{P}_i \ln(\tilde{P}_i) \qquad . \qquad \triangleq M\,S
\end{aligned}$$

Damit entspricht die wahrscheinlichste Verteilung des Zufallsexperiments, die über Gl. (15.32) ermittelt wird, genau der Maxent-Lösung. Es stellt sich die Frage, wie relevant die wahrscheinlichste (Maxent) Lösung ist, bzw. wie breit die Verteilung der $\underline{m}$ ist. Es wurde von E.T.Jaynes im sogenannten ENTROPIE-KONZENTRATIONSTHEOREM gezeigt, dass für großen Stichprobenumfang M die mit den Nebenbedingungen verträglichen Häufigkeiten $\underline{m}$ zu Werten $x = 2M(S_{\max} - S)$ führen, die einer χ^2-Verteilung mit $\nu = N - L - 1$ Freiheitsgraden genügen. Das heißt, für $M \gg N$ ist das Intervall, in dem mit der Wahrscheinlichkeit α die Entropie-Werte liegen werden,

$$\left[S_{\max} - \frac{\chi^2_\nu(\alpha)}{2M}, S_{\max}\right] \qquad .$$

Nun ist $\chi^2_\nu(\alpha) \propto \nu$, das heißt das Entropie-Intervall geht mit zunehmendem M gegen Null. Das bezeichnet man als Entropie-Konzentrationstheorem, da die Häufigkeitsverteilungen $\underline{m}$ mit zunehmendem Stichprobenumfang Entropie-Werte in der Nähe des Maximalwertes haben und somit der Maxent-Verteilung entsprechen.

Kapitel 16

Maxent bei kontinuierlichen Variablen

Für den diskreten Fall ist die Entropie durch

$$S^D = -\sum_{j=1}^{N} P_j \ \ln(P_j)$$

gegeben. Die Variable i soll nun zu einer kontinuierlichen Größe werden. Hierbei geht P_j über in

$$P(\Delta x_j) = p(x_j)\ \Delta x_j \qquad .$$

Der Abstand Δx_j der Punkte x_j geht mit $N \to \infty$ wie $1/N$ gegen Null. Wir definieren hierzu

$$\Delta x_j = \frac{1}{m(x_j)} \frac{1}{N} \qquad .$$

Die Bedeutung von $m(x_j)$ ist

$$m(x_j) = \lim_{N\to\infty} \frac{1/N}{\Delta x_j} = \frac{\text{Zahl der Zustände in } \Delta x_j}{\text{Intervall-Länge}} \triangleq \text{Maß} \qquad .$$

Es gilt

$$\int m(x)\ dx = \lim_{N\to\infty} \sum_{j=1}^{N} m(x_j)\ \Delta x_j = \lim_{N\to\infty} \sum_{j=1}^{N} \frac{1}{N} = 1 \qquad .$$

Damit wird aus der Entropie

$$\begin{aligned} S^D &= -\sum_{j=1}^{N} \Delta x_j\ p(x_j)\ \ln\left(p(x_j)\ \Delta x_j\right) \\ &= -\sum_{j=1}^{N} \Delta x_j\ p(x_j)\ \ln\left(\frac{p(x_j)}{m(x_j)N}\right) \\ &= -\sum_{j=1}^{N} \Delta x_j\ p(x_j)\ \ln\left(\frac{p(x_j)}{m(x_j)}\right) + \ln(N) \underbrace{\sum_{j=1}^{N} \Delta x_j\ p(x_j)}_{=1} \qquad . \end{aligned}$$

Man definiert die Entropie für kontinuierliche Freiheitsgrade als Grenzwert

$$S^C := \lim_{N\to\infty} (S^D - \ln(N)) \qquad . \tag{16.1}$$

Die Konstante hat keinen Einfluss auf die Maxent-Lösung. Die Entropie ist somit

ENTROPIE FÜR KONTINUIERLICHE FREIHEITSGRADE

$$S^C = -\int p(x)\ \ln\left(\frac{p(x)}{m(x)}\right)\ dx, \quad m(x) : \text{invariantes Maß}$$

Nur mit der Normierungsbedingung ist die Maxent-Lösung, wie wir gleich sehen werden, identisch zu $m(x)$. Da keine weitere Information bzgl. $p(x)$ vorliegt, muss diese Lösung mit der uninformativen Prior-Verteilung übereinstimmen. Deshalb nennt man sie INVARIANTES MASS. Die Entropie wird auch RELATIVE ENTROPIE, CROSS-ENTROPY und KULLBACK-LEIBLER-ENTROPIE bezeichnet. Sie hat die Bedeutung des Abstands

$$D(p : m)$$

im Raum der Wahrscheinlichkeitsdichten zwischen der Wahrscheinlichkeitsdichte $p(x)$ und dem „Aufpunkt" $m(x)$. Hierfür gilt allerdings nicht die Dreiecksungleichung.

Wie im Fall diskrete Freiheitsgrade besteht die Maxent-Lösung darin, die Entropie unter Berücksichtigung der Nebenbedingungen

$$\varphi_\mu\{p(x)\} = 0, \quad \mu = 1, \ldots, L \qquad ,$$

zu maximieren. Dazu definieren wir wieder die Lagrange-Funktion

$$\mathcal{L} = -\int p(x)\ \ln\left(\frac{p(x)}{m(x)}\right)\ dx - \lambda_0\ \left(\int p(x)\ dx - 1\right) - \sum_\mu \lambda_\mu \varphi_\mu\{p(x)\} \qquad . \tag{16.2}$$

Die Maxent-Lösung erhält man aus der Nullstelle der Funktionalsableitung

$$\begin{aligned}
0 &= \frac{\delta}{\delta p(x)}\,\mathcal{L} \\
&= \frac{\delta}{\delta p(x)}\,\left(-\int\, p(x')\,\ln\left(\frac{p(x')}{m(x')}\right)\,dx'\right. \\
&\qquad\qquad \left. -\,\lambda_0\,\left(\int\, p(x')\,dx' - 1\right) - \sum_\mu \lambda_\mu \varphi_\mu\{p(x)\}\right) \\
&= -\ln\left(\frac{p(x)}{m(x)}\right) - 1 - \lambda_0 - \sum_\mu \lambda_\mu\,\frac{\delta}{\delta p(x)}\,\varphi_\mu\{p(x)\} \\
\ln\left(\frac{p(x)}{m(x)}\right) &= -1 - \lambda_0 - \sum_\mu\,\lambda_\mu\,\frac{\delta}{\delta p(x)}\,\varphi_\mu\{p(x)\} \qquad .
\end{aligned}$$

Wir beschränken uns wieder auf lineare Nebenbedingungen

$$\varphi_\mu\{p(x)\} = \int\, p(x)\,K_\mu(x)\,dx - \kappa_\mu, \quad \mu = 1, \ldots, L$$

Die Funktionalableitung ist in diesem Fall

$$\frac{\delta}{\delta p(x)}\,\varphi_\mu\{p(x)\} = K_\mu(x) \qquad .$$

Die Maxent-Lösung lautet also allgemein

MAXENT-LÖSUNG BEI LINEAREN NEBENBEDINGUNGEN FÜR KONTINUIERLICHE FREIHEITSGRADE

$$p(x) = \frac{1}{Z}\,m(x)\,e^{-\sum_\mu\,\lambda_\mu K_\mu(x)} \tag{16.3a}$$

$$Z = \int\, m(x)\,e^{-\sum_\mu\,\lambda_\mu\,K_\mu(x)}\,dx \qquad . \tag{16.3b}$$

Auch in in diesem Fall lassen sich die zusätzlichen Nebenbedingungen elegant formulieren

$$\begin{aligned}
\kappa_\mu &= \int\, K_\mu(x)\,p(x) \\
&= \frac{1}{Z}\,\int\, m(x)\,K_\mu(x)\,e^{-\sum_\mu\,\lambda_\mu\,K_\mu(x)} \\
&= -\frac{\partial}{\partial\lambda_\mu}\,\ln(Z) \qquad .
\end{aligned}$$

Es seinen z.B. die unteren Momente der Wahrscheinlichkeitsdichte $p(x)$ vorgegeben.

$$\kappa_\mu = \langle x^n \rangle = \int x^n \rho(x) dx$$

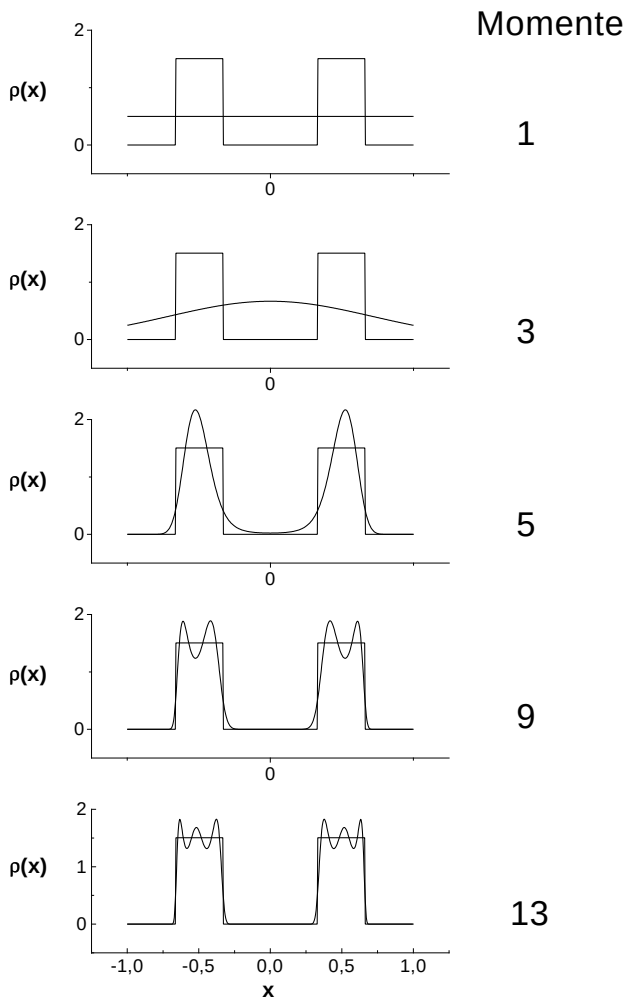

Abbildung 16.1: *Maximum-Entropie Lösung des Momenten-Problems.*

Weitere einfache Beispiele sind

Barometrische Höhenformel

Gesucht sei die Wahrscheinlichkeit $p(z)$, Teilchen der Masse m in der Höhe z (mit $z \in [0, \infty)$) im Schwerefeld der Erde anzutreffen. Neben der Normierung sei die mittlere potentielle Energie bekannt

$$\kappa = \frac{\langle E_p \rangle}{m\ g} = \int\ p(z)\ z\ dz = \frac{1}{\beta\ m\ g} \qquad .$$

Das die mittlere Energie $kT = 1/\beta$ ist, folgt natürlich aus physikalischen Überlegungen. Als invariantes Maß verwenden wir die flache Verteilung und erhalten als normierte Maxent-Lösung

$$p(z) = \frac{1}{Z}\, e^{-\lambda\, z}$$
$$Z = \frac{1}{\lambda} \qquad .$$

Die Nebenbedingung verlangt

$$-\frac{d}{d\lambda} \ln(Z) = \frac{d}{d\lambda} \ln(\lambda) = +\frac{1}{\lambda} \stackrel{!}{=} \kappa$$
$$\Rightarrow \qquad \lambda = \frac{1}{\kappa} \qquad .$$

Damit haben wir das aus der statistischen Mechanik bekannte Ergebnis

BAROMETRISCHE HÖHENFORMEL
$$p(z) = \beta\ m\ g\ e^{-\beta\ m\ g\ z} \qquad .$$ (16.4)

Gauß-Verteilung

Wir suchen die Wahrscheinlichkeitsdichte $p(x)$, mit $x \in (-\infty, \infty)$. Es seien die unteren zwei Momente, Mittelwert μ und Varianz σ^2 einer Verteilung gegeben. Die Maxent-Lösung ist

$$p(x) = \frac{1}{Z}\ e^{-\lambda_1\ x - \lambda_2\ x^2} \qquad .$$

Das lässt sich immer quadratisch ergänzen und in die bereits korrekt normierte Form

$$p(x) = \frac{1}{\sqrt{2\ \pi\ a}}\ e^{-\frac{(x-b)^2}{2\ a}}$$

bringen. Damit liegt bereits eine Normal-Verteilung vor, von der wir wissen, dass der Mittelwert b und die Varianz a beträgt. Also ist die zu den ersten zwei Momenten gehörige Maxent-Lösung die Normalverteilung mit diesen Momenten.
Ein Spezialfall hiervon ist die Maxwellsche Geschwindigkeitsverteilung, bei der die kinetische Energie der Teilchen, also das zweite Moment, bekannt ist.
Die Maxent-Ableitung liefert eine weitere Erklärung, warum die Gauß-Funktion allgegenwärtig ist. Diese Erklärung setzt nicht voraus, anders als der zentrale Grenzwertsatz, dass die Größe x eine Summe i.u.nv. Zufallszahlen ist und dass sehr viele Summanden beitragen. Die einzige Voraussetzung ist, dass nur die unteren Momente bekannt oder überhaupt mit hinreichender Genauigkeit bestimmbar sind.

Multivariate-Normal-Verteilung

Für die Variable $x \in \mathbb{R}^N$ wird die Wahrscheinlichkeitsdichte $p(x)$ gesucht. Als exakte Information seien die Mittelwerte

$$\mu_i = \int x_i \; p(x) \; d^N x, \qquad i = 1, \ldots, N \tag{16.5}$$

und die Kovarianzen

$$C_{ij} = \int \Delta x_i \Delta x_j \; p(x) \; d^N x \tag{16.6}$$

mit $\Delta x_i = x_i - \mu_i$ gegeben. Die Lagrange-Funktion ist in diesem Fall

$$\begin{aligned} \mathcal{L} = -\int p(x) \; \ln\left(\frac{p(x)}{m(x)}\right) \; d^N x - \lambda_0 \int p(x) \; d^N x \\ - \sum_{i=1}^{N} \lambda_i \left(\int x_i \; p(x) \; d^N x - \mu_i \right) \\ - \sum_{i,j=1}^{N} \Lambda_{ij} \; \left(\int \Delta x_i \Delta x_j \; p(x) \; d^N x - C_{ij} \right) \quad . \end{aligned}$$

Die Funktionalableitung nach $p(x)$ liefert

$$\frac{\delta}{\delta p(x)} \; \mathcal{L} = -\ln(\frac{p(x)}{m(x)}) - 1 - \lambda_0 - \sum_{i=1}^{N} \lambda_i \; x_i - \sum_{ij} \; \Lambda_{ij} \; \Delta x_i \Delta x_j \quad .$$

Aus der Nullstelle folgt

$$p(x) = \frac{m(x)}{Z} \; e^{-\sum_i \lambda_i \; x_i - \sum_{ij} \; \Lambda_{ij} \; \Delta x_i \Delta x_j} \quad .$$

Wir gehen wieder von einer flachen Verteilung $m(x)$ aus und modifizieren den ersten Term zu

$$p(x) = \frac{1}{Z'} \; e^{-\sum_i \lambda_i \; \Delta x_i - \sum_{ij} \; \Lambda_{ij} \; \Delta x_i \Delta x_j} \quad .$$

Damit die Nebenbedingung Gl. (16.5) erfüllt ist, muss gelten

$$\int \Delta x_i \; p(x) \; d^N x = 0 \qquad .$$

Das wiederum gilt nur, wenn $\lambda_i = 0$ ist. Damit lautet die vorläufige Maxent-Lösung

$$p(x) = \frac{1}{Z''} \; e^{-\sum_{ij} \; \Lambda_{ij} \; \Delta x_i \Delta x_j} \qquad . \tag{16.7}$$

Das ist eine multivariaten Normalverteilung von der wir die Kovarianz aus Gl. (9.28a) kennen. Damit die Kovarianz-Nebenbedingung Gl. (16.6) erfüllt ist, muss also gelten

MAXENT-LÖSUNG: MITTELWERTE UND KOVARIANZ GEGEBEN

$$p(x) = (2\pi \; |C|)^{-\frac{N}{2}} \; e^{-\frac{1}{2} \Delta x^T C^{-1} \Delta x} \qquad . \tag{16.8}$$

Kapitel 17

Das invariante Riemann-Maß

Wir hatten bereits das Transformations-Invarianz-Prinzip besprochen, um uninformative, invariante Prior-Wahrscheinlichkeit zu bestimmen.

Die uninformative Prior-Wahrscheinlichkeit von Parametern a hängt natürlich von ihrer Bedeutung ab. Ein und derselbe Buchstabe, z.B. σ kann je nach Problem unterschiedliche Bedeutungen und demnach unterschiedliche Invarianz-Eigenschaften haben. Diese werden eindeutig durch die Likelihood-Funktion festgelegt. Es macht daher wenig Sinn zu fordern, dass eine Prior-Wahrscheinlichkeit in willkürlichen Parametern flach sein soll. Aber es macht durchaus Sinn, zu verlangen, dass die Wahrscheinlichkeitsdichte so sein muss, dass alle Likelihood-Verteilungen

$$p(d|a,\mathcal{B})$$

im Raum der Wahrscheinlichkeitsdichten $p(d)$, die man als Riemannsche Mannigfaltigkeiten auffassen kann, gleich-wahrscheinlich sind. Mit einigen Überlegungen aus der Differential-Geometrie kann man daraus ableiten, dass

$$\begin{aligned} p(a|\mathcal{B}) &= |g|^{\frac{1}{2}} \\ g_{ij} &= \int p(d|a,\mathcal{B}) \, \frac{\partial^2}{\partial a_i \partial a_j} \, \ln\left(p(d|a,\mathcal{B})\right) \, d^N d \\ &= \left\langle \frac{\partial^2}{\partial a_i \partial a_j} \, \ln\left(p(d|a,\mathcal{B})\right) \right\rangle \end{aligned}$$

Die Riemann-Metrik g ist auch gleichzeitig die Fisher-Informations-Matrix. Der Vorteil dieses Zuganges ist, dass man sich nicht überlegen muss, gegen welche Transformationen das Problem invariant ist. Diese Information wird von der Likelihood

implizit geliefert und im Riemann-Maß automatisch implementiert.

$$
\begin{aligned}
p(a|\mathcal{B})\ da = |g(a)|^{\frac{1}{2}}\ da &= \left| \left\langle \frac{\partial^2}{\partial a_i \partial a_j}\ \ln\left(p(d|a,\mathcal{B})\right) \right\rangle \right|^{\frac{1}{2}}\ da \\
&= \left| \frac{\partial a'_m}{\partial a_i}\ \frac{\partial a'_n}{\partial a_j}\ \left\langle \frac{\partial^2}{\partial a'_m \partial a'_n}\ \ln\left(p(d|a',\mathcal{B})\right) \right\rangle \right|^{\frac{1}{2}}\ da \\
&= \left| J^{-1}\ g(a')\ J^{-1} \right|^{\frac{1}{2}}\ da \\
&= |g(a')|^{\frac{1}{2}}\ J^{-1}\ da \\
&= |g(a')|^{\frac{1}{2}}\ da' \qquad = p(a'|\mathcal{B})\ da' \qquad .
\end{aligned}
$$

Die so definierte Wahrscheinlichkeitsdichte $p(a|\mathcal{B})$ ist somit in der Tat invariant gegen Re-Parametrisierung.

Kapitel 18

Fehlerbehaftete überprüfbare Information

Der häufigste Fall ist, dass die Daten, die zur Verfügung stehen, verrauscht sind, egal ob sie von Computersimulationen oder „echten" Experimenten stammen. Wir wollen nun eine formlose Rekonstruktion von Verteilungsfunktionen ableiten.

Zu Beginn wollen wir uns die Problematik vor Augen führen. Angenommen wir sind am Vektor $\underline{\rho}$ interessert, der mit dem Datenvektor $\underline{d}$ über eine Matrix M verknüpft ist:

$$M \cdot \underline{\rho} = \underline{d}$$

In der Regel sind die Daten $\underline{d}$ von einem Rauschen $\underline{\eta}$ überlagert. In vielen Fällen ist zusätzlich die Matrix M schlecht konditioniert, was bedeutet, dass das Verhältnis vom höchsten zum niedrigsten Eigenwert sehr groß ist. Damit kann man das Gleichungssystem

$$M \cdot \underline{\rho} = \underline{d} + \underline{\eta} \tag{18.1}$$

durch direktes Invertieren nicht lösen, da kleine Eigenwerte von M das Rauschen überproportional verstärken. Das kann man leicht einsehen, wenn man annimmt, dass M eine quadratische Matrix ist, die man in der Form

$$M = \sum_i a_i \, \underline{u}_i \, \underline{u}_i^{\mathrm{T}}$$

schreiben kann, wobei a_i die Eigenwerte und $\underline{u}_i$ die Eigenvektoren sind. Wendet man die Inverse von M auf Gl. (18.1) an, erhält man

$$M^{-1}(\underline{d} + \underline{\eta}) = \sum_i \frac{1}{a_i} \, \underline{u}_i \, \underline{u}_i^{\mathrm{T}}(\underline{d} + \underline{\eta}) = \underline{\rho} + \sum_i \underline{u}_i \, \frac{\underline{u}_i^{\mathrm{T}} \underline{\eta}}{a_i} \qquad ,$$

wobei $\underline{\rho}$ die exakte Lösung ist. Gibt es einen großen Überlapp $\underline{u}_i^{\mathrm{T}} \underline{\eta}$ zwischen einem Eigenvektor $\underline{u}_i^{\mathrm{T}}$ mit kleinem Eigenwert a_i und dem Rauschvektor $\underline{\eta}$, wird der Fehler stark verstärkt. Stammt M von einer experimentellen Aparatefunktion, (??? general point spread function ???), so oszillieren die Eigenfunktionen kleiner Eigenwerte stark. Damit ist der Überlapp $\underline{u}_i^{\mathrm{T}} \underline{\eta}$ mit dem Rauschen groß.

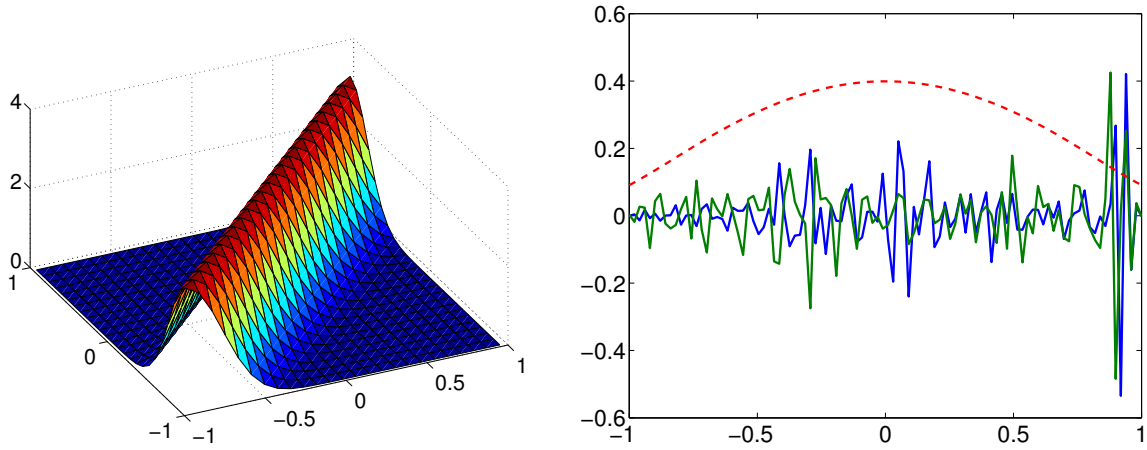

Abbildung 18.1: Gaußförmige Apparatefunktion (links) und die Eigenvektoren zum größten (strichliert) und den beiden kleinsten Eigenwerten. Man erkennt deutlich die starken Oszillationen der Eigenvektoren zu den kleinsten Eigenwerten.

Das sieht man schon am typischen Beispiel für eine Apparatefunktion, nämlich an der Gaußschen Glockenkurve. Wir nehmen an, die gewünschten physikalischen Größen $\rho(x)$ werden durch eine Messapparatur $M(x, x')$ auf die Daten $d(x)$ durch

$$d(x) = \int M(x, x')\, \rho(x)\, dx \quad \text{mit} \quad M(x, x') = \sqrt{\beta}\, e^{-\beta\, |x-x'|^2}$$

($\beta > 0$) abgebildet (Faltung). Eine einfache Diskretisierung der Funktionen ($d_i = d(x_i)$) und des Integrals liefert den linearen Zusammenhang

$$d_i = \sum_j M_{i,j}\, \rho_j \quad \text{mit der Matrix} \quad M_{i,j} = \sqrt{\beta}\, e^{-\beta\, |x_i - x_j|^2}\ .$$

Nimmt man an, dass $x \in [-1, 1]$ und diskretisiert das Intervall in 80 Punkte, so ergibt sich für $\beta = 10$ die in Abb. 18.1 gezeichnete Apparatefunktion. Der kleinste Eigenwert von M ist $\simeq 10^{-14}$ also beinahe 0, der größte ist ≈ 84. Man erkennt in Abb. 18.1, dass die Eigenvektoren zu den kleinen Eigenwerten stark oszillieren und deshalb stark ans Rauschen ankoppeln.

Ein Beispiel eines solch schlecht konditionierten Problems in der Vielteilchenphysik ist die Rekonstruktion der Spektralfunktion $A(\omega)$ aus der gemessenen (berechneten) Greenschen Funktion $g(\tau)$, die über eine Laplace-Transformation

$$g(\tau_i) = \int_0^\infty A(\omega)\, e^{-\omega\tau_i}\, d\omega \tag{18.2}$$

miteinander verknüpft sind. Hier entspricht die Greensche Funktion $g(\tau)$ den Daten $d(.)$, die Spektralfunktion $A(\omega)$ ist die Funktion $\underline{\rho}(.)$ und die Abbildung M ist durch Integration über die Exponentialfunktion $e^{-\omega\tau_i}$ gegeben.

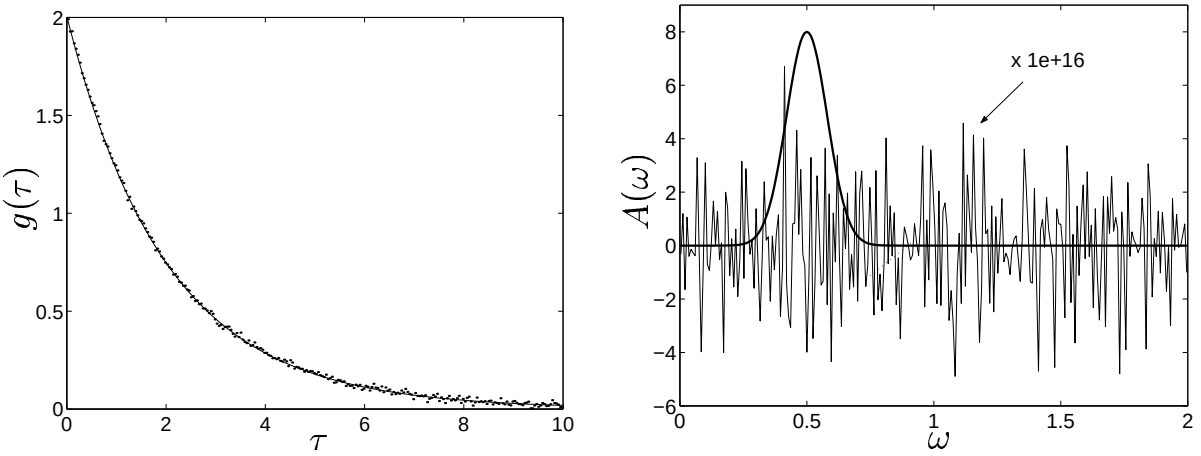

Abbildung 18.2: *Die glatte Kurve im linken Bild wird vom glatten Peak im rechten Bild erzeugt. Beim Invertieren reicht minimales Rauschen (meist genügt schon das Rechnen mit endlicher Genauigkeit) von $g(\tau)$ aus, um das Rekonstruierte $A(\omega)$ unbrauchbar zu machen.*

Der exponentielle Anteil des Integranden fällt mit wachsendem ω stark ab, wodurch der entsprechende Anteil von $A(\omega)$ kaum zum Ergebnis $g(\tau_i)$ beiträgt. Umgekehrt kann man bei gegebenem $g(\tau_i)$ kaum etwas über diesen Anteil von $A(\omega)$ sagen. Noch schlimmer, da die Abbildung schlecht konditioniert ist — die größte Eigenwert im Beispiel Abbildung 18.2 ist rund 74, der kleinste 3×10^{-19} — ergibt eine direkte Invertierung von Gl. (18.2) ein völlig unbrauchbares, stark oszillierendes Ergebnis.

Da der direkte Zugang nicht zum Erfolg führt, wählt man einen wahrscheinlichkeitstheoretischen. Man fragt nach der wahrscheinlichsten Funktion $\rho(.)$ gegeben die Daten $d(.)$. Die Anwendung des Bayesschen Theorems liefert

$$p(\rho(.)|d(.), \mathcal{B}) = \frac{1}{Z}\, p(d(.)|\rho(.), \mathcal{B})\; p(\rho(.)|\mathcal{B}) \qquad . \tag{18.3}$$

Die Schwierigkeit besteht nun darin, den Prior $p(\rho(.)|\mathcal{B})$ zu bestimmen.

Bestimmung des Priors $p(\rho(x)|\mathcal{B})$

Die folgende Vorgangsweise wird QUANTIFIED MAXIMUM ENTROPY (QME) genannt. Als Vorraussetzung geht ein, dass man die gesuchte Funktion $\rho(x)$ als Wahrscheinlichkeitsdichte interpretieren kann, das heißt, dass $\rho(x)$ POSITIV ist. Unser Ziel ist es, die Wahrscheinlichkeitsdichte $p(\rho(x)|\underline{d}, \mathcal{B})$ zu berechnen, dass die Verteilungsfunktion $\rho(x)$ die echte ist, gegeben die N_d Datenpunkte d_ν. Die Prozedur kann in vier Schritte unterteilt werden.

Schritt eins Wir DISKRETISIEREN den Raum x, indem wir ihn in N Untermengen aufteilen. Oft sind die aus den Messdaten zu berechnenden Größen bereits diskret,

dann muss nichts gemacht werden. Ist der x-Raum jedoch wirklich kontinuierlich, integrieren wir über bestimmte Volumselemente ΔV_i, um ein diskrete Menge zu erhalten

$$\rho(x) \to \rho_i = \rho(x_i)\,\Delta V_i \qquad .$$

Diese diskrete Menge wird PIXELMENGE genannt.

Schritt zwei Im sogenannten QUANTISIERUNGSSCHRITT suchen wir eine kleinste Einheit $\Delta\rho$, sodass, gemessen in dieser Einheit, sich die ρ_i durch ganze Zahlen ausdrücken lassen:

$$\rho_i \to n_i \qquad \text{mit} \quad \rho_i \approx n_i\,\Delta\rho \qquad . \tag{18.4}$$

Manchmal sind die Ergebnisse aus physikalischen Gründen schon Vielfache einer natürlichen Einheit $\Delta\rho$.

Schritt drei Als nächstes müssen wir eine PRIOR-WAHRSCHEINLICHKEIT $P(\underline{n}|\mathcal{B})$ zuweisen, ohne die Daten $\underline{d}$ zu kennen. Dazu führen wir die mittlere Zahl μ_i in jedem Pixel i als DEFAULT-MODELL ein. Wir nehmen nun an, dass A PRIORI die Dichte $\rho(x)$ dadurch entsteht, dass eine gewisse Anzahl von Einheiten $\Delta\rho$ zufällig auf der reellen Achse verteilt wird. Als Nebenbedingung dieses Verteilungsprozesses soll nur gelten, dass im Mittel in jedem Pixel i genau μ_i dieser Einheiten sind. Dann ist die Wahrscheinlichkeit, dass A PRIORI die Zahlen n_i auftreten durch eine Poisson-Verteilung gegeben:

$$P(n_i|\mu_i,\mathcal{B}) = \frac{\mu_i^{n_i}}{n_i!}\,e^{-\mu_i}\;, \qquad n_i = 0, 1, 2, \ldots$$

Die Wahrscheinlichkeit, dass ein Pixel eine weitere Einheiten $\Delta\rho$ dazubekommt, ist unabhängig von den schon verteilten $\Delta\rho$'s. Aus dem Bild ergibt sich, dass die einzelnen Pixel unkorreliert sind. Die kombinierte Wahrscheinlichkeit $P(\underline{n}|\underline{\mu},\mathcal{B})$ für alle N Pixels ist daher durch das Produkt

$$P(\underline{n}|\underline{\mu},\mathcal{B}) = \prod_{i=1}^{N} P(n_i|\mu_i,\mathcal{B}) = e^{-\sum_i \mu_i} \prod_{i=1}^{N} \frac{\mu_i^{n_i}}{n_i!}$$

gegeben. Da wir uns in allen weitern Schritten auf das Default-Modell beziehen, nehmen wir es in den Bedingungskomplex $\mathcal{B}$ auf und schreiben in Zukunft statt $P(\underline{n}|\underline{\mu},\mathcal{B})$ kurz $P(\underline{n}|\mathcal{B})$. Wir nähern die Fakultät mittels der Stirlingschen Formel $n! \approx \sqrt{2\pi\,n}\,n^n\,e^{-n}$ und führen für die Normalisierungskonstante die Abkürzung Z ein. Damit erhalten wir

$$P(\underline{n}|\mathcal{B}) = \frac{1}{Z}\,\frac{1}{\prod_i \sqrt{n_i}}\,e^{\sum_{i=1}^{N} n_i - \mu_i - n_i \log(n_i/\mu_i)} \qquad . \tag{18.5}$$

Der Exponent in diesem Ausdruck ist eine Verallgemeinerung der SHANNON ENTROPY (Gl. (15.1)) und wird deshalb mit $\mathcal{S}$ bezeichnet. Diese Verallgemeinerung kommt dadurch zustande, weil wir uns auf ein Default-Modell beziehen (ähnlich

dem invariantem Maß bei kontinuierlichen Variablen). $\mathcal{S}$ ist eine Summe über die Entropien $\mathcal{S}_i$ jedes einzelnen Pixels

$$\mathcal{S} = \sum_{i=1}^{N} \mathcal{S}_i = \sum_{i=1}^{N} n_i - \mu_i - n_i \log(n_i/\mu_i) \qquad .$$

Schritt vier Im letzten Schritt machen wir die Quantisierung rückgängig und gehen damit auf kontinuierliche Größen ρ_i und m_i für jedes Pixel $i = 1, \ldots, N$ über.

$$n_i = \frac{\rho_i}{\Delta\rho} \;\Rightarrow\; \mu_i = \frac{m_i}{\Delta\rho} \tag{18.6}$$

Das Default-Modell wird jetzt von den Größen m_i getragen. Der Prior für n_i, Gl. (18.5), und damit der Prior für ρ_i kann durch ρ_i und m_i ausgedrückt werden. Einsetzen von Gl. (18.6) in Gl. (18.5) liefert

$$\begin{aligned} p(\underline{\rho}|\mathcal{B}) &= \frac{1}{Z} \frac{1}{\prod_i \sqrt{\rho_i}} e^{\frac{1}{\Delta\rho}\sum_i \rho_i - m_i - \rho_i \log(\rho_i/m_i)} \\ &= \frac{1}{Z(\alpha)} \frac{1}{\prod_i \sqrt{\rho_i}} e^{\alpha S} \qquad , \end{aligned} \tag{18.7}$$

mit der Entropie

$$S = \sum_{i=1}^{N} S_i = \sum_{i=1}^{N} \rho_i - m_i - \rho_i \log(\rho_i/m_i) \qquad . \tag{18.8}$$

Wir haben den Parameter $\alpha = \frac{1}{\Delta\rho}$ eingeführt. Da er ein zusätzlicher, nicht bekannter und damit unerwünschter Parameter ist, wird er auch NUISANCE PARAMETER oder HYPER-PARAMETER genannt. Er wird später näher bestimmt (noch besser ausintegriert) werden müssen. Einstweilen notieren wir seine Existenz, indem wir ihn hinter dem Bedingungsstrich aufnehmen.

Mit dem Auffinden des Priors $p(\underline{\rho}|\mathcal{B})$ ist die Grundlage des QME gelegt worden. Der Rest ist nur noch reines Anwenden der Bayesschen Wahrscheinlichkeitstheorie, wie es schon viele Male gezeigt worden ist. Da jedoch einige technische Details zu beachten sind, sollen die einzelnen Schritte vorgezeigt werden. Zuerst jedenfalls muss der Prior normiert werden.

Normierung des Priors in Steepest-Descent-Näherung

Die Normierung von $p(\underline{\rho}|\alpha, \mathcal{B})$ kann man numerisch oder analytisch durchführen. Letzteres wird nur über Näherungen möglich sein, wobei wir die STEEPEST-DESCENT-Näherung verwenden wollen.

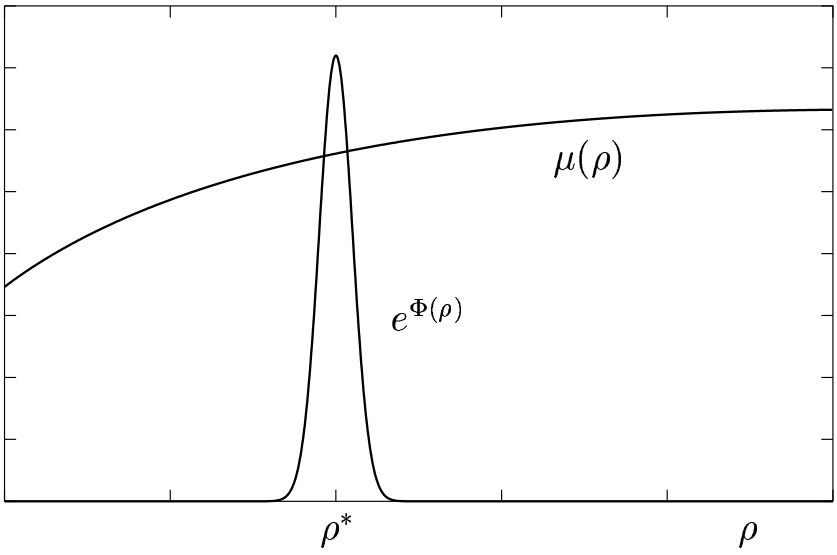

Abbildung 18.3: *Illustration der Steepest-Descent-Methode: Das Maß $\mu(\rho)$ ist im Vergleich zum Exponentialfaktor sehr flach.*

Steepest-Descent-Näherung: Diese Methode wendet man an, wenn der Integrand ein Produkt eines schwach veränderlichen Maßes $\mu(\underline{\rho})$ mit einem scharf gepeakten Faktor $\exp(\Phi(\underline{\rho}))$ ist. Man bestimmt das Maximum $\underline{\rho}^*$ des Exponenten $\Phi(\underline{\rho})$ und entwickelt ihn um $\underline{\rho}^*$ bis zur zweiten Ordnung in ein Taylor-Polynom. Das ist in Abbildung 18.3 schematisch dargestellt. Die Entwicklung ist durch

$$\int e^{\Phi(\underline{\rho})}\,\mu(\underline{\rho})\,d\underline{\rho} \approx \mu(\underline{\rho}^*)\int_{\mathbb{R}^N} e^{\Phi(\underline{\rho}^*)+\frac{1}{2}\sum_{i,j}\Delta\rho_i\frac{\partial^2}{\partial\rho_i\partial\rho_j}\Phi\Big|_{\underline{\rho}^*}\Delta\rho_j}\,d\underline{\rho}$$

gegeben. Diese Approximation ist immer dann gerechtfertigt, wenn das Maximum scharf gepeakt ist und alle Eigenwerte der Matrix

$$\frac{\partial^2}{\partial\rho_i\partial\rho_j}\Phi\Big|_{\underline{\rho}^*}$$

kleiner als Null sind (sonst konvergiert das Integral nicht). Da in unserem Fall die Funktion $\Phi(\underline{\rho})$ konvex ist, ist diese Forderung erfüllt. Wir definieren die positive definite Hessematrix H_{ij}

$$H_{ij} = -\frac{\partial^2}{\partial\rho_i\partial\rho_j}\Phi\Big|_{\underline{\rho}^*} \tag{18.9}$$

und führen die Integration über $\mathbb{R}^N$ aus. Das führt zu folgender Näherungsformel

$$\int_0^\infty e^{\Phi(\underline{\rho})}\,\mu(\underline{\rho})\,d\underline{\rho} \approx \mu(\underline{\rho}^*)\,e^{\Phi(\underline{\rho}^*)}\,(2\pi)^{N/2}\,|H|^{-1/2} \quad . \tag{18.10}$$

Der Normierungsfaktor $Z(\alpha)$: Wir wollen Gl. (18.10) nun benutzen, um den Normierungsfaktor $Z(\alpha)$ aus Gl. (18.7) zu bestimmen. Dazu schauen wir uns die Entropie S näher an. Sie besteht aus einer Summe

$$S = \sum_{i=1}^{N} \rho_i - m_i - \rho_i \log \frac{\rho_i}{m_i} \qquad ,$$

in der jeder Summand nur aus Beiträgen eines Pixels besteht, d.h. es werden keine Korrelationen zwischen den Pixels berücksichtigt. Permutationen des Index i haben keinen Einfluss, somit wird Stetigkeit von der rekonstruierten Verteilung ρ nicht gefordert. Diese spezielle Form bewirkt auch, dass die Norm $Z(\alpha)$ zu

$$Z(\alpha) = \int e^{\alpha S} \frac{1}{\prod_i \sqrt{\rho_i}} d\underline{\rho} = \prod_{i=1}^{N} \int_0^\infty e^{\alpha(\rho_i - m_i - \rho_i \log \frac{\rho_i}{m_i})} \frac{d\rho_i}{\sqrt{\rho_i}}$$

faktorisiert. Für die Steepest-Descent-Näherung benötigen wir das Maximum jedes Terms S_i:

$$S_i = \rho_i - m_i - \rho_i \log \frac{\rho_i}{m_i} \;\Rightarrow\; \frac{\partial S_i}{\partial \rho_i} = 1 - 1 - \log \frac{\rho_i}{m_i} = 0$$

Die Lösung $\rho_i^* = m_i$ ist unser Default-Modell. Die Matrix der zweiten Ableitungen ist diagonal und ergibt sich zu

$$\frac{\partial^2 S}{\partial \rho_i^2}\Big|_{\rho_i = \rho_i^* = m_i} = -\frac{1}{\rho_i}\Big|_{\rho_i = \rho_i^* = m_i} = -\frac{1}{m_i} \qquad .$$

Nun können wir das Taylor-Polynom zweiter Ordnung eines Entropie-Terms S_i um das Maximum $\rho_i^* = m_i$ angeben:

$$S_i(\rho_i) = 0 - \frac{1}{2}\frac{\Delta\rho_i^2}{m_i} \;, \qquad \text{mit } \Delta\rho_i = \rho_i - m_i$$

Der Faktor $\frac{1}{\sqrt{\rho_i}}$ entspricht dem Maß $\mu(\rho)$ in der Steepest-Descent-Näherung Gl. (18.10) und wird am Maximum vor das Integral gezogen. Damit erhalten wir

$$Z(\alpha) = \prod_{i=1}^{N} \frac{1}{\sqrt{m_i}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{\Delta\rho_i^2}{m_i/\alpha}} d\rho_i = \prod_{i=1}^{N} \frac{1}{\sqrt{m_i}} \sqrt{2\pi\, m_i/\alpha} = (2\pi)^{N/2}\, \alpha^{-N/2} \qquad .$$

Insgesamt erhalten wir für den Prior bei gegebenem α in der Steepest-Descent-Näherung

ENTROPISCHER PRIOR

$$p(\underline{\rho}|\alpha,\mathcal{B}) = (2\pi)^{-N/2}\,\frac{\alpha^{N/2}}{\prod_i \sqrt{\rho_i}}\,e^{\alpha S} \tag{18.11}$$

mit

$$S = \sum_{i=1}^{N} \rho_i - m_i - \rho_i \log\frac{\rho_i}{m_i} \quad .$$

Die Posterior-Wahrscheinlichkeitsdichte $p(\underline{\rho}|\underline{d},\mathcal{B})$

Nachdem wir den Prior für $\underline{\rho}$ bestimmt haben, wollen wir die Wahrscheinlichkeitsdichte $p(\underline{\rho}|\underline{d},\mathcal{B})$ für $\underline{\rho}$ im Lichte der Daten $\underline{d}$ berechnen. Nach Marginalisierung und Anwenden der Produktregel erhalten wir

$$\begin{aligned} p(\underline{\rho}|\underline{d},\mathcal{B}) &= \int p(\alpha,\underline{\rho}|\underline{d},\mathcal{B})\,d\alpha \\ &= \int p(\underline{\rho}|\alpha,\underline{d},\mathcal{B})\;p(\alpha|\underline{d},\mathcal{B})\,d\alpha \quad . \end{aligned} \tag{18.12}$$

Als Funktion von α ist der erste Faktor $p(\underline{\rho}|\alpha\,\underline{d},\mathcal{B})$ flach im Vergleich zum Faktor $p(\alpha|\underline{d},\mathcal{B})$, der schon bei wenigen Daten scharf gepeakt ist. Deshalb nähern wir das Integral durch

$$p(\underline{\rho}|\underline{d},\mathcal{B}) \approx p(\underline{\rho}|\alpha^*,\underline{d},\mathcal{B}) \underbrace{\int p(\alpha|\underline{d},\mathcal{B})\,d\alpha}_{=1} = p(\underline{\rho}|\alpha^*,\underline{d},\mathcal{B}) \quad , \tag{18.13}$$

wobei α^* jenes α ist, das $p(\alpha|\underline{d},\mathcal{B})$ maximiert. Diese Näherung wird EVIDENZ-NÄHERUNG genannt. Setzen wir das in Gl. 18.3 ein, so erhalten wir in dieser Näherung

$$p(\underline{\rho}|\underline{d},\mathcal{B}) = p(\underline{\rho}|\underline{d},\alpha^*,\mathcal{B}) = \frac{p(\underline{d}|\underline{\rho},\mathcal{B})\;p(\underline{\rho}|\alpha^*,\mathcal{B})}{p(\underline{d}|\mathcal{B})} \quad . \tag{18.14}$$

Nun müssen wir noch α^* berechnen, wozu wir $p(\alpha|\underline{d},\mathcal{B})$ kennen müssen. Diese Dichte bestimmen wir abermals durch Marginalisierung und Anwenden der Produktregel:

$$\begin{aligned} p(\alpha|\underline{d},\mathcal{B}) &= \int p(\alpha,\underline{\rho}|\underline{d},\mathcal{B})\,d\underline{\rho} = \frac{1}{p(\underline{d}|\mathcal{B})}\int p(\alpha,\underline{\rho},\underline{d}|\mathcal{B})\,d\underline{\rho} \\ &= \frac{1}{p(\underline{d}|\mathcal{B})}\int p(\underline{d}|\underline{\rho},\alpha,\mathcal{B})\,p(\underline{\rho}|\alpha,\mathcal{B})\,p(\alpha|\mathcal{B})\,d\underline{\rho} \end{aligned}$$

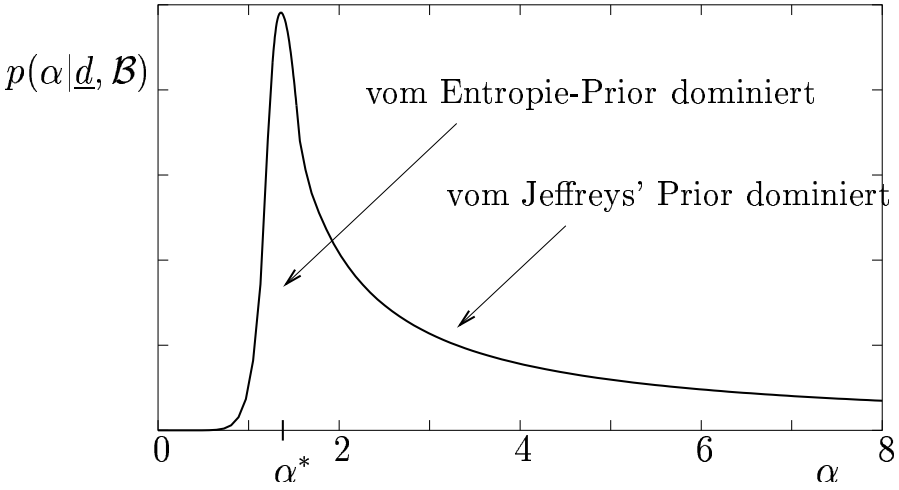

Abbildung 18.4: *Qualitatives Verhalten der Funkton* $p(\alpha|\underline{d},\mathcal{B})$.

Die Wahrscheinlichkeitsdichte $p(\underline{d}|\underline{\rho},\alpha,\mathcal{B})$ hängt nicht von α ab, da bei gegebenen Dichten $\underline{\rho}$ und Bedingungskompex $\mathcal{B}$ die Daten $\underline{d}$ bestimmt sind. Daher ergibt sich

$$p(\alpha|\underline{d},\mathcal{B}) = \frac{1}{p(\underline{d}|\mathcal{B})} \int p(\underline{d}|\underline{\rho},\mathcal{B})\, p(\underline{\rho}|\alpha,\mathcal{B})\, p(\alpha|\mathcal{B})\, d\underline{\rho} \tag{18.15}$$

Der Prior $p(\alpha|\mathcal{B})$ ist durch JEFFREYS' PRIOR $p(\alpha|\mathcal{B}) \propto \frac{1}{\alpha}$ gegeben, da es sich bei α um eine Skalen-Variable handelt. Die Dichte $p(\underline{\rho}|\alpha,\mathcal{B})$ ist durch den ENTROPISCHEN PRIOR (Gl. (18.11)) gegeben. Für die LIKELIHOOD $p(\underline{d}|\underline{\rho},\mathcal{B})$ benötigen wir ein theoretisches Modell, welches die Verteilung der Messdaten $\underline{d}$ bei gegebenen Dichten $\rho(x)$ liefert.

Für $p(\alpha|\underline{d},\mathcal{B})$ können zwei extreme Fälle unterscheiden (vgl. Abbildung 18.4):

- Keine Regularisierung, $\alpha \to 0$: Das bewirkt, dass der Exponent im entropischen Prior verschwindet und daher $p(\alpha|\underline{d},\mathcal{B}) \sim \alpha^{N/2-1}$. Das Verhalten wird vom entropieschen Prior bestimmt.

- Starke Regularisierung, $\alpha \to \infty$: Der Einfluss der Daten wird immer geringer, da kleine Abweichungen vom Default-Modell stark exponentiell unterdrückt werden, was dazu führt, dass das Default-Modell angenommen wird.

  $$p(\underline{\rho}|\alpha,\mathcal{B}) \to \delta(\underline{\rho} - \underline{m}) \;\Rightarrow\; p(\underline{d}|\underline{m},\mathcal{B})\, p(\alpha|\mathcal{B})$$

  Da $p(\underline{d}|\underline{m},\mathcal{B})$ nicht von α abhängt, wird das Verhalten von Jeffreys' Prior dominiert, also $p(\alpha|\underline{d},\mathcal{B}) \sim 1/\alpha$.

Jetzt müssen wir noch die Likelihood-Funktion $p(\underline{d}|\underline{\rho},\mathcal{B})$ festlegen. Dazu benötigen wir ein THEORETISCHES MODELL, das die Daten $\underline{d}$ mit der Verteilungsfunktion $\underline{\rho}$ verknüpft. Prinzipiell kann dieses Modell so komplex wie erforderlich sein. In vielen

Fällen ist das Modell aber linear, d.h. der Vektor der theoretisch erwarteten Daten $\underline{d}^{\mathrm{th}}$ wird durch Anwendung einer Matrix M auf die Verteilung $\underline{\rho}$ gewonnen. Die allgemeine Form eines linearen Modells lässt sich als

$$\underline{d}^{\mathrm{th}}(\underline{\rho}) \;=\; M \cdot \underline{\rho} \tag{18.16}$$

schreiben. Die Abweichung $\Delta\underline{d}$ der echten von den theoretisch erwarteten Daten ist durch

$$\Delta\underline{d} = \underline{d} - \underline{d}^{\mathrm{th}}(\underline{\rho}) = \underline{d} - M \cdot \underline{\rho} \tag{18.17}$$

gegeben. Nimmt man an, dass das Rauschen additiv und normalverteilt ist, ist die Likelihood-Funktion eine multivariate Normalverteilung

$$p(\underline{d}|\underline{\rho}, \mathcal{B}) = (2\pi)^{-N_d/2}\, |C|^{-1/2}\, e^{-\frac{1}{2}\Delta\underline{d}^{\mathrm{T}} C^{-1} \Delta\underline{d}} \qquad , \tag{18.18}$$

wobei N_d die Anzahl der Daten ist. Die Kovaranzmatrix C sollte aus dem Experiment abgeschätzt werden können. Setzen wir diesen Ansatz in Gl. (18.15) ein, so ergibt sich

$$p(\alpha|\underline{d}, \mathcal{B}) = (2\pi)^{-(N_d+N)/2}\, |C|^{-1/2}\, \alpha^{N/2-1} \underbrace{\int_0^\infty \cdots \int_0^\infty}_{N\text{-mal}} e^{-\frac{1}{2}\Delta\underline{d}^{\mathrm{T}} C^{-1} \Delta\underline{d} + \alpha S} \prod_i \frac{d\rho_i}{\sqrt{\rho_i}} \tag{18.19}$$

Der Exponent

$$\Phi = -\frac{1}{2} \underbrace{\Delta\underline{d}^{\mathrm{T}} C^{-1} \Delta\underline{d}}_{\chi^2} + \alpha\, S\,. \tag{18.20}$$

setzt sich aus dem MISS-FIT χ^2 der Daten und dem entropischen Anteil αS zusammen. Das Integral wollen wir wieder in der Steepest-Descent-Näherung auswerten. Dazu benötigen wir das Maximum $\underline{\rho}^*$ des Exponenten. Die Position $\underline{\rho}^*$ des Maximums des Exponenten hängt wieder von α ab, wodurch die Bestimmung von $p(\alpha|\underline{d}, \mathcal{B})$ im Prinzip für jedes α getrennt durchgeführt werden muss. In der Steepest-Descent-Näherung erhalten wir

$$p(\alpha|\underline{d}, \mathcal{B}) \simeq (2\pi)^{-N_d/2}\, |C|^{-1/2}\, \alpha^{N/2-1}\, |H|^{-1/2}\, \frac{1}{\prod_i \sqrt{\rho_i^*}} \qquad . \tag{18.21}$$

Die Hessematrix am Maximum $\underline{\rho}^*$ lautet

$$H_{ij} = -\frac{\partial^2 \Phi}{\partial\rho_i \partial\rho_j} = (M^{\mathrm{T}} C^{-1} M)_{ij} + \frac{\alpha}{\rho_i^*}\, \delta_{ij} \qquad . \tag{18.22}$$

Setzen wir das nun in Gl. (18.14) ein, so erhalten wir die POSTERIOR-WAHRSCHEINLICHKEITSDICHTE in Evidenz-Näherung

POSTERIORI WAHRSCHEINLICHKEIT
$$p(\underline{\rho}\vert\underline{d},\mathcal{B}) = \frac{1}{Z}\, e^{-\frac{1}{2}\chi^2+\alpha^* S} \quad , \tag{18.23}$$

wobei wir in der letzten Zeile der Faktor $\prod_i \rho_i^{-1/2}$ aus Gl. (18.11) vernachlässigen, da wir annehmen, dass er flach im Vergleich zum exponentiellen Ausdruck ist. Weiters haben wir den Nenner durch Z abgekürzt haben.

Lösung mittels Legendre-Transformation

Das Maximum des Exponenten $\Phi = -\frac{1}{2}\chi^2 + \alpha S$ führt auf einen Schätzer von $\underline{\rho}^*$. Direkte Maximierung ist äußerst schlecht konditioniert. Deshalb führen wir eine LEGENDRE-TRANSFORMATION der Funktion

$$\Phi(\underline{\rho},\underline{d}) = -\frac{1}{2}\chi^2(\underline{d}) + \alpha S(\underline{\rho}) \tag{18.24}$$

durch. Dies ist möglich, da die Abbildung $\underline{\rho} \mapsto \Phi(\underline{\rho},\underline{d})$ konvex ist. Durch die Transformation wird die quadratische Abhängigkeit $\underline{\rho}$ von zu einer linearen. Wir ersetzen die Daten $\underline{d}$ durch die partiellen Ableitungen $\underline{\lambda}$ und Φ durch die Transformierte $\hat{\Phi}$

$$\lambda_\nu = \frac{1}{\alpha}\frac{\partial\Phi}{\partial d_\nu}\,, \qquad \hat{\Phi}(\underline{\rho},\underline{\lambda}) = \Phi(\underline{\rho},\underline{d}(\underline{\lambda})) - \alpha\sum_{\nu=1}^{N_d}\lambda_\nu d_\nu \quad . \tag{18.25}$$

Die neuen unabhängigen Variablen λ_ν sind durch

$$\lambda_\nu = \frac{1}{\alpha}\frac{\partial\Phi}{\partial d_\nu} = -\frac{1}{2\alpha}\frac{\partial\chi^2}{\partial d_\nu} = -\frac{1}{\alpha}\sum_{\nu'=1}^{N_d} C^{-1}_{\nu\nu'}\Big(d_{\nu'} - d^{\rm th}_{\nu'}(\underline{\rho})\Big) \tag{18.26}$$

gegeben, wodurch der Vektor der theoretisch erwarteten Daten durch

$$\underline{d}^{\rm th}(\underline{\rho}) = \underline{d} + \alpha\, C\, \underline{\lambda} \tag{18.27}$$

ausgedrückt werden kann. Der letzte Ausdruck spiegelt wider, dass für $\alpha \to 0$ die theoretischen mit den experimentellen Daten übereinstimmen. Die Transformierte von Φ ist also insgesamt durch

$$\begin{aligned}\hat{\Phi}(\underline{\rho},\underline{\lambda}) &= -\frac{1}{2}\,\alpha^2\,\underline{\lambda}^{\rm T}\,C\,C^{-1}\,C\,\underline{\lambda} - \alpha\underline{\lambda}^{\rm T}\underline{d}^{\rm th}(\underline{\rho}) + \alpha^2\underline{\lambda}^{\rm T}C\underline{\lambda} + \alpha S(\underline{\rho}) \\ &= \frac{\alpha^2}{2}\underline{\lambda}^{\rm T}\,C\,\underline{\lambda} - \alpha\underline{\lambda}^{\rm T}\,\underline{d}^{\rm th}(\underline{\rho}) + \alpha S(\underline{\rho})\end{aligned}$$

gegeben. Verwenden wir das lineare Modell, Gl. (18.16), erhalten wir

$$\hat{\Phi}(\underline{\rho}, \underline{\lambda}) = \frac{\alpha^2}{2} \underline{\lambda}^{\mathrm{T}} C \, \underline{\lambda} - \alpha \underline{\lambda}^{\mathrm{T}} M \, \underline{\rho} + \alpha S(\underline{\rho}) \qquad .$$

Dieser Ausdruck muss bezüglich $\underline{\rho}$ maximiert werden. Dazu berechnen wir den Gradienten und setzen ihn null:

$$\frac{\partial \hat{\Phi}}{\partial \rho_i} = -\alpha \log \frac{\rho_i}{m_i} - \alpha (\underline{\lambda}^{\mathrm{T}} M)_i \stackrel{!}{=} 0$$

Als Ergebnis erhalten wir

$$\rho_i = m_i \, e^{-(\underline{\lambda}^{\mathrm{T}} M)_i} \qquad . \tag{18.28}$$

Diese spezielle Gestalt erzwingt die POSITIVITÄT der Verteilung ρ_i. Für $\underline{\lambda} = 0$ ergibt sich das Default-Modell. Wir setzen Gl. (18.28) in Gl. (18.27) ein und führen den Fehlervektor $\underline{\psi}$ ein:

$$\begin{aligned} \psi_\nu(\underline{\lambda}) &= d_\nu^{\mathrm{th}} - d_\nu - \alpha (C \, \underline{\lambda})_\nu = (M \underline{\rho})_\nu - d_\nu - \alpha (C \, \underline{\lambda})_\nu \\ &= \sum_{i=1}^{N} M_{\nu i} \, m_i \, e^{-\sum_\mu \lambda_\mu M_{\mu i}} - d_\nu - \alpha \sum_{\nu'=1}^{N_d} C_{\nu\nu'} \lambda_{\nu'} \quad \stackrel{!}{=} 0 \end{aligned} \tag{18.29}$$

Numerische Lösung mittels Newton-Verfahren

Gl. (18.29) muss man numerisch lösen. Eine einfache Methode dazu ist das Newton-Verfahren. Das Taylor-Polynom erster Ordnung von $\psi_\nu(\underline{\lambda} + \Delta\underline{\lambda})$ ist

$$\psi_\nu(\underline{\lambda} + \Delta\underline{\lambda}) \approx \psi_\nu(\underline{\lambda}) + \underbrace{\frac{\partial}{\partial \lambda_\mu} \psi_\nu(\underline{\lambda})}_{D_{\nu\mu}} \Delta\lambda_\mu \qquad .$$

Das Iterarionsschema erhält man, indem man das Taylor-Polynom null setzt. Nach Inversion der Matrix D erhalten wir folgende Rekursionsformel

$$\underline{\lambda}^{(n+1)} = \underline{\lambda}^{(n)} - D^{-1} \, \underline{\psi}(\underline{\lambda}^{(n)}) \qquad . \tag{18.30}$$

Die Praxis zeigt, dass es oft vorteilhaft ist, am Beginn der Iteration nicht den gesamten Newton-Schritt durchzuführen, da zu diesem Zeitpunkt ein Taylor-Polynom erster Ordnung eine zu grobe Näherung ist, und die Routine eventuell nicht konvergiert. Dieses Problem wird in Abbildung 18.5 veranschaulicht. Besser ist es am Anfang nur einen Bruchteil $\gamma < 1$ eines Newton-Schritts durchzuführen

$$\underline{\lambda}^{(n+1)} = \underline{\lambda}^{(n)} - \gamma \, D^{-1} \, \underline{\psi}(\underline{\lambda}^{(n)}) \qquad . \tag{18.31}$$

Im Laufe der Iterationen kann man dann denn Bruchteil langsam erhöhen, $\gamma \to 1$. Die Matrix D hat folgende Form

$$D_{\nu\mu} = -\sum_{i=1}^{N} M_{\nu i} \, M_{\mu i} \underbrace{m_i \, e^{-\sum_\mu \lambda_\mu M_{\mu i}}}_{\rho_i(\underline{\lambda})} - \alpha \, C_{\nu\mu} \qquad . \tag{18.32}$$

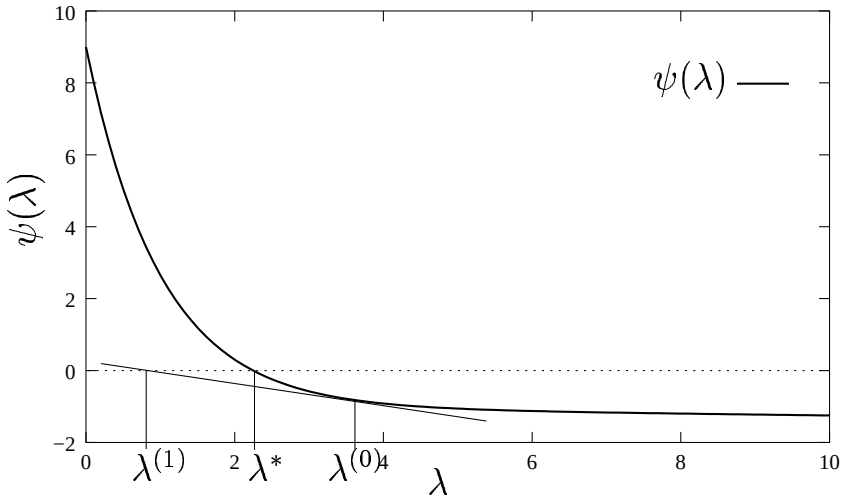

Abbildung 18.5: *Eindimensionales Beispiel zum Erleutern des Newton-Verfahrens. In diesem Fall überschätzt das Verfaren den Abstand zwischen $\lambda^{(0)}$ und der Wurzel λ^*.*

Optimale Größe von N

Wie groß ist die optimale Anzahl N der rekonstruierten Punkte von ρ gegeben die Daten $\underline{d}$. Die Wahrscheinlichkeit für diese Größe ist

$$P(N|\underline{d},\mathcal{B}) = \frac{p(N,\underline{d}|\mathcal{B})}{p(\underline{d}|\mathcal{B})} \quad . \tag{18.33}$$

Über die Marginalisierungsregel führen wir die Verteilung $\underline{\rho}$ ein:

$$\begin{aligned} P(N|\underline{d},\mathcal{B}) &= \frac{1}{Z}\int_{\mathbb{R}^N} p(N,\underline{d},\underline{\rho}|\mathcal{B})\, d\underline{\rho} \\ &= \frac{1}{Z}\int_{\mathbb{R}^N} p(\underline{d}|N,\underline{\rho},\mathcal{B})\, p(\underline{\rho}|N,\mathcal{B})\, p(N|\mathcal{B})\, d\underline{\rho}\,. \end{aligned} \tag{18.34}$$

Der erste Faktor des Integranden ist die Likelihood-Funktion, Gl. (18.18), die beiden hinteren Faktoren sind Prioren, die wir flach ansetzen

$$\begin{aligned} p(\underline{\rho}|N,\mathcal{B}) &= \prod_{i=1}^{N} \frac{\theta(0\le\rho_i\le a)}{a} \\ p(N|\mathcal{B}) &= \frac{\theta(N_0\le N\le N_1)}{N_1-N_0+1} \quad . \end{aligned}$$

Das bedeutet, dass wir, ohne Messungen durchzuführen, nur wissen, dass $\rho_i \in [0,a]$, und dass $N \in \{N_0,\ldots,N_1\}$. Setzen wir das in Gl. (18.34) ein, erhalten wir

$$P(N|\underline{d},\mathcal{B}) = \frac{1}{Z'}|C|^{-1/2}\int_{\mathbb{R}^N} e^{-\frac{1}{2}\Delta\underline{d}^{\mathrm{T}}C^{-1}\Delta\underline{d}} \prod_{i=1}^{N}\frac{\theta(0\le\rho_i\le a)}{a}\, d\underline{\rho} \quad ,$$

mit $Z' = Z(N_1 - N_0 + 1)(2\pi)^{N_d/2}$ und $\Delta\underline{d} = \underline{d} - M \cdot \underline{\rho}$. Wählt man die Konstante a genügend groß, sodass es keine Einschränkungen für die statistisch möglichen $\underline{\rho}$ gibt, kann man die Funktionen $\theta(0 \leq \rho_i \leq a)$ vernachlässigen. Damit vereinfacht sich obiger Ausdruck zu

$$P(N|\underline{d}, \mathcal{B}) = \frac{1}{Z'} \frac{|C|^{-1/2}}{a^N} \int_{\mathbb{R}^N} e^{-\frac{1}{2}\Delta\underline{d}^{\mathrm{T}} C^{-1} \Delta\underline{d}} \, d\underline{\rho} \qquad . \tag{18.35}$$

Wir schreiben den Exponenten in der Form

$$\begin{aligned} \varphi(\underline{\rho}) &\equiv \frac{1}{2}\Delta\underline{d}^{\mathrm{T}} C^{-1} \Delta\underline{d} \\ &= \varphi(\underline{\rho}^*) + \Delta\underline{\rho}^{\mathrm{T}} M^{\mathrm{T}} C^{-1} M \Delta\underline{\rho} \qquad , \end{aligned}$$

wobei $\Delta\underline{\rho}$ die Differenz zwischen $\underline{\rho}$ und der Maximum-Likelihood-Lösung $\underline{\rho}^*$ ist. Letztere erhält man aus

$$\frac{\partial\varphi}{\partial\underline{\rho}} = -M^{\mathrm{T}} C^{-1} \left(\underline{d} - M \cdot \underline{\rho}\right) \stackrel{!}{=} \underline{0} \qquad ,$$

woraus

$$\underline{\rho}^* = \left(M^{\mathrm{T}} C^{-1} M\right)^{-1} M^{\mathrm{T}} C^{-1} \underline{d} \tag{18.36}$$

folgt. Setzen wir das in Gl. (18.35) ein, erhalten wir

$$\begin{aligned} P(N|\underline{d}, \mathcal{B}) &= \frac{1}{Z'} \frac{|C|^{-1/2}}{a^N} e^{-\varphi(\underline{\rho}^*)} \int_{\mathbb{R}^N} e^{-\Delta\underline{\rho}^{\mathrm{T}} M^{\mathrm{T}} C^{-1} M \Delta\underline{\rho}} \, d\underline{\rho} \\ &= \frac{1}{Z'} \frac{|C|^{-1/2}}{a^N} e^{-\varphi(\underline{\rho}^*)} (2\pi)^{N/2} \left|M^{\mathrm{T}} C^{-1} M\right|^{-1/2} \qquad . \end{aligned} \tag{18.37}$$

Im Falle unkorrelierter Daten, $C = \sigma^2 \, \mathbb{I}$, vereinfacht sich diese Formel zu

$$P(N|\underline{d}, \mathcal{B}) = \frac{1}{Z''} \left(\frac{\sqrt{2\pi}}{a}\right)^N \left|M^{\mathrm{T}} M\right|^{-1/2} e^{-\varphi(\underline{\rho}^*)} \qquad , \tag{18.38}$$

mit $Z'' = Z'/\sigma^{2N_d}$.

18.1 Beispiele

In diesem Abschnitt wollen wir die Maximum-Entropie-Methode an zwei Beispielen demonstrieren. Im ersten Beispiel wird die Inversion der Laplace-Transformation unter die Lupe genommen. Das ist ein Problem, das z.B. bei Quanten-Monte-Carlo-Methoden auftritt. Das zweite Beispiel ist eine optische Anwendung, die Abel-Inversion.

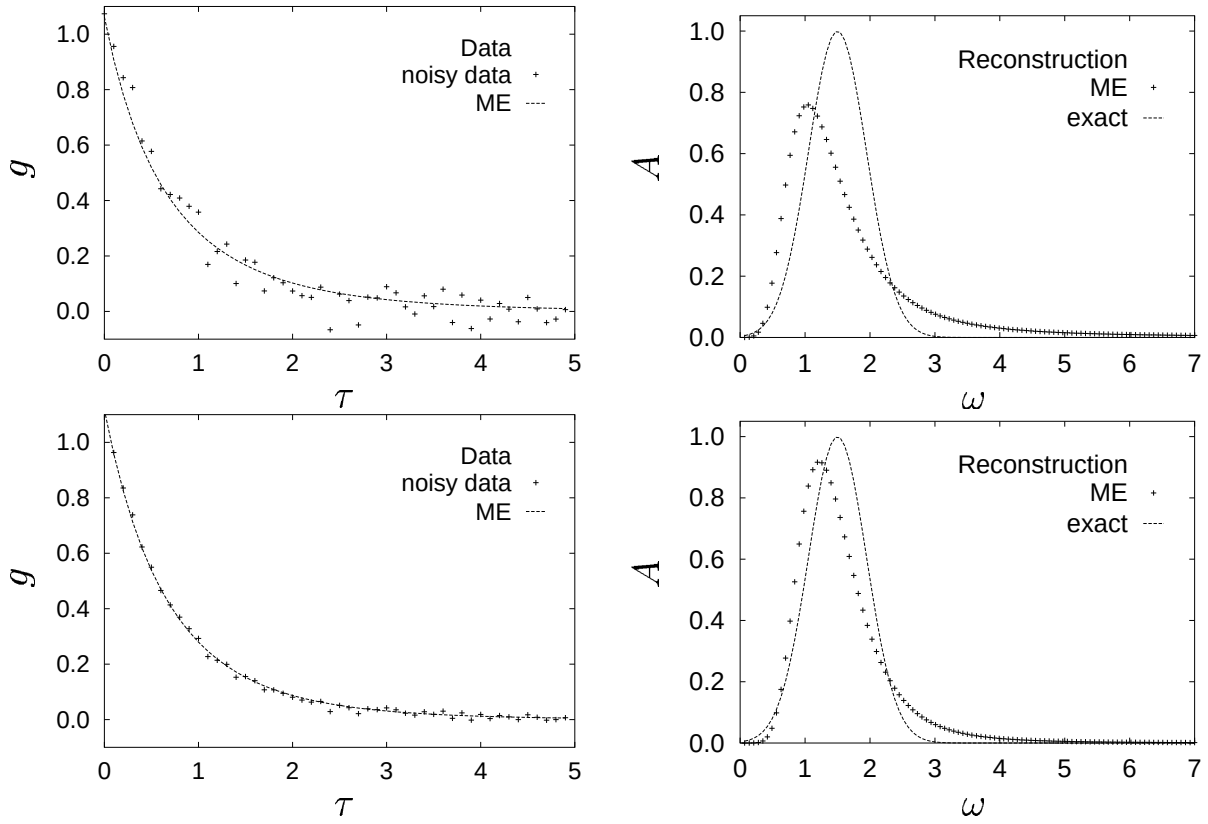

Abbildung 18.6: *Abhängigkeit der Güte der Rekonstruktion vom Rauschen. Die linken Abbildungen zeigen die Daten g, die rechten deren Maximum-Entropie-Rekonstruktionen. Die strichlierte Linie in den linken Abbildungen zeigt die g, die aus den rekonstruierten A berechnet worden sind.*

18.1.1 Invertieren der Laplace-Transformation

Wir wollen nun das eingangs vorgestellte Beispiel mit der QME-Methode bearbeiten. Die experimentellen Daten g seien über eine Laplace-Transformation mit der gewünschten physikalischen Größe A gekoppelt

$$g(\tau) = \int_0^\infty e^{-\tau\omega}\, A(\omega)\, d\omega \qquad . \tag{18.39}$$

Die Daten sind auf einer Menge $\tau_1, \ldots, \tau_{N_d}$ gegeben. Die Aufgabenstellung ist, die Funktion $A(\omega)$ zu rekonstruieren.

Als erstes diskretisieren wir die Beziehung (18.39). Das erreicht man am einfachsten, indem man die obere Integrationsgrenze durch eine endliche Integrationsgrenze $\omega = \omega^f$ ersetzt und das Integral durch eine Summe ersetzt. Führen wir N ω-Werte, $\omega_i =$

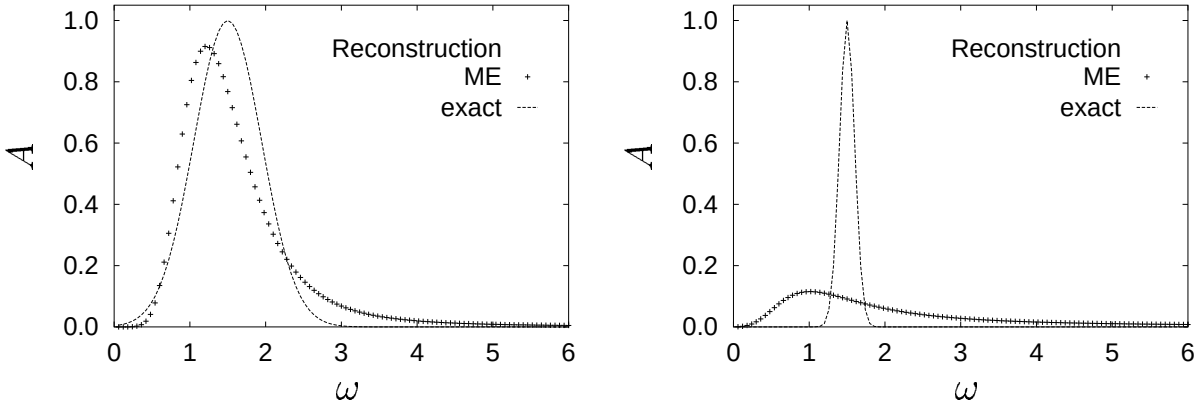

Abbildung 18.7: *Rekonstruktion von verrauschten Daten. Die Qualität der Rekonstruktion in Abhängigkeit von der Peakbreite.*

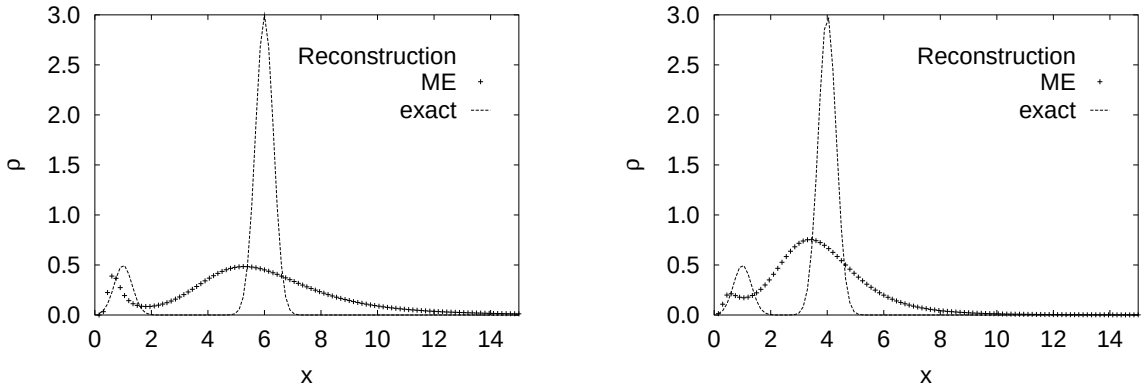

Abbildung 18.8: *Auflösungsvermögen zweier Peaks. Je weiter die Peaks von einander entfernt sind, desto leichter können sie vom Maximum-Entopie-Verfahren aufgelöst werden.*

$(i - 1/2)\,\omega^f/N$, ein, erhalten wir

$$\begin{aligned} g_\nu \equiv g(\tau_\nu) &\approx \sum_{i=1}^{N} e^{-\tau_\nu \omega_i}\, A(\omega_i)\, \Delta\omega \\ &= \sum_{i=1}^{N} e^{-\tau_\nu \omega_i}\, A_i\, \Delta\omega \qquad , \end{aligned} \tag{18.40}$$

mit $A_i \equiv A(\omega_i)$ und $\Delta\omega = \omega^f/N$. Somit ist die Matrix des Modells durch

$$M_{\nu i} = e^{-\tau_\nu \omega_i}\, \Delta\omega$$

gegeben.
Einige Resultate einer Maximum-Entropie-Rekonstruktion von $A(\omega)$ sind in den Abbildungen 18.6–18.8 zu sehen. Abbildung 18.6 zeigt den Einfluss vom Rauschen auf die Qualität der Rekonstruktion: Wegen des stärkeren Rauschens, das man in den

Abbildungen auf der linken Seite sehen kann, ist der rekonstruierte Peak im ersten Plot breiter und niedriger als der im zweiten. Über Gl. (18.40) kann man aus den rekonstruierten $\underline{A}$ die daraus theoretisch erwarteten Daten berechnen. Diese sind in der linken Spalte der Abbildung 18.6 als strichlierte Linie eingezeichnet.

In Abbildung 18.7 wird der Einfluss der Peakbreite auf die Güte der Rekonstruktion demonstriert. Um so schmäler der Peak, desto schlechter fällt der Vergleich mit der exakten Verteilung aus.

Das Auflösungsvermögens wird in Abbildung 18.8 gezeigt. Zwei Fälle sind geplottet: Die Rekonstruktion auf der linken Seite kann die beiden Peaks auflösen, während in der rechten Abbildung der kleinere Peak zu einer Schulter verkommen ist.

18.1.2 Abel-Inversion

In diesem Abschnitt wenden wir die QME-Methode auf ein weiteres physikalisches Inversionsproblem an, die Abel-Inversion.

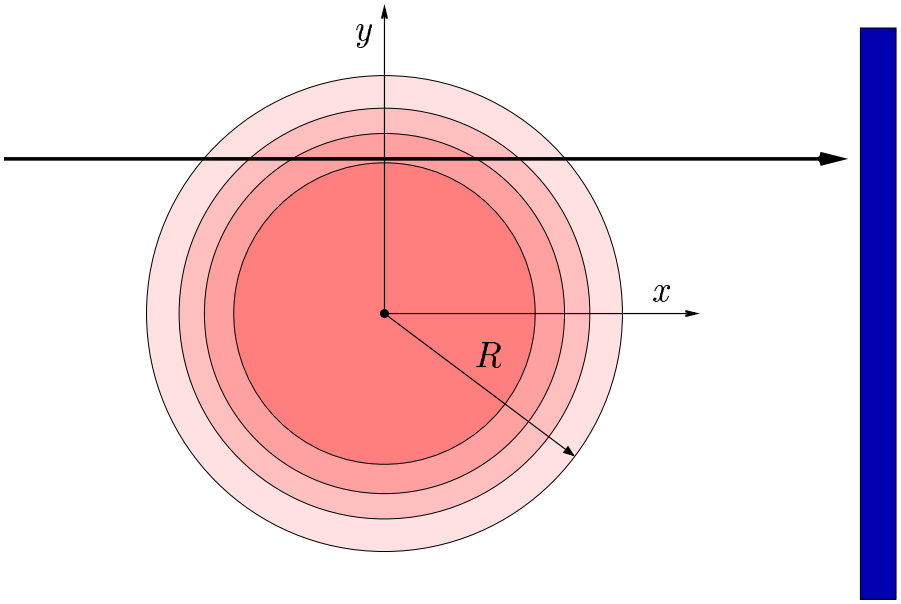

Abbildung 18.9: *Geometrie der Abel-Inversion. Die Dichte des Mediums hängt nur vom Radius* $r \in [0, R]$ *ab.*

Bei der Abel-Inversion geht man von einer zylindrisch symmetrischen Verteilung des absorbierenden Mediums aus. Die Daten werden entlang der Sekanten (vgl. Abbildung 18.9), wo der Laser das Material durchstoßt, gemessen. Der Absorptionsindex ist proportional zur Dichte $\rho(r)$. Weiters nehmen wir an, dass das Ausmaß der Absorption klein gegenüber der Laser-Intensität ist, wodurch das Problem linear wird.

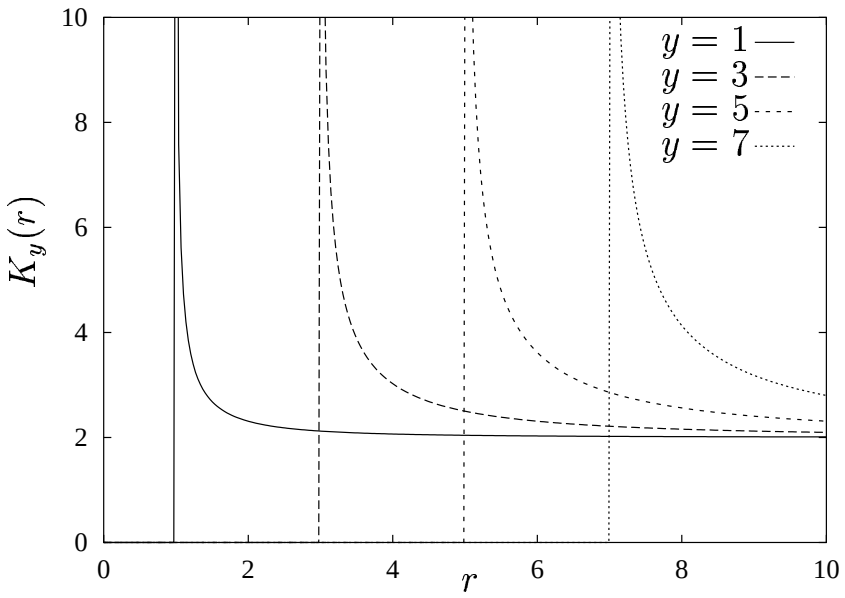

Abbildung 18.10: *Der Integrationskern $K_y(r)$ der Abel-Inversion für unterschiedliche Werte von y. Die Punkte $r = y$ sind Singularitäten, wo der Kern gegen Unendlich geht.*

Die Absorption ist dann

$$A(y) = \int_{-\sqrt{R^2-y^2}}^{\sqrt{R^2-y^2}} \rho(\sqrt{x^2+y^2})\, dx \qquad .$$

Nach der Substitution $r = \sqrt{x^2+y^2}$ erhalten wir

$$A(y) = \int_y^R \frac{2r}{\sqrt{r^2-y^2}}\, \rho(r)\, dr = \int_0^R \Theta(r-y)\, \frac{2r}{\sqrt{r^2-y^2}}\, \rho(r)\, dr \qquad ,$$

wobei $\Theta(.)$ die Stufenfunktion ist. Damit haben wir unser Problem in eine Standardform für lineare Probleme gebracht. Mit dem Integralkern $K_y(r)$ können wir schreiben

$$A(y) = \int_0^R K_y(r)\, \rho(r)\, dr\,, \qquad \text{mit } K_y(r) = \Theta(r-y)\, \frac{2r}{\sqrt{r^2-y^2}} \qquad . \tag{18.41}$$

Wegen der Singularität des Kerns $K_y(r)$ an der Stelle $r = y$ ist die Inversion des Integrals sehr schlecht konditioniert. Abbildung 18.10 zeigt eine Skizze des Kerns für verschiedene Werte von y.
Wir benötigen ein Modell für die Dichte des absorbierenden Mediums. Der Einfachheit halber nehmen wir eine stückweise konstante Funktion der Form

$$\rho(r) = \sum_{j=1}^{N_0} \rho_j\, \chi_j(r) \tag{18.42}$$

an, wobei χ_j charakteristische Funktionen der Intervalle $[(j-1)R/N_0, jR/N_0)$ sind. Stetigkeit der Dichte wird dann über eine Spline-Interpolation erreicht. Setzt man das Modell (18.42) in die Gl. (18.41), erhält man

$$A(y) = \int_0^R K_y(r)\,\rho(r)\,dr = \int_y^R \frac{2r}{\sqrt{r^2-y^2}} \sum_{j=1}^{N_0} \rho_j\,\chi_j(r) \qquad .$$

Die Integrale sind leicht auszuwerten, nur bei den Integrationsgrenzen muss man ein wenig aufpassen. Vertauscht man die Summe mit dem Integral, wird obiger Ausdruck zu

$$A(y) = \sum_{j=1}^{N_0} \rho_j \int_y^R \frac{2r}{\sqrt{r^2-y^2}}\,\chi_j(r) = \sum_{j=1}^{N_0} \rho_j\,A_j(y) \qquad ,$$

mit

$$A_j(y) = \begin{cases} 2\sqrt{r^2-y^2}\Big|_{r=(j-1)N_0/R}^{jN_0/R} & \dots\ y \le (j-1)R/N_0 \\ 2\sqrt{(jR/N_0)^2-y^2} & \dots\ (j-1)R/N_0 < y \le jR/N_0 \\ 0 & \dots\ y > jR/N_0 \end{cases} \qquad .$$

Gemessen wird nur an diskreten Punkten entlang der y-Achse. Wir nehmen vereinfachend an, dass der Laserstrahl unendlich dünn ist. Messungen sollen an folgenden Punkten vorgenommen werden:

$$y_\nu = \frac{(\nu-1/2)R}{N_d}\,, \qquad \nu = 1,\dots,N_d$$

Damit ergibt sich für die Modellmatrix, die die Daten mit den Dichten verknüpft

$$A_{\nu j} = A_j(y_\nu) = \begin{cases} 2\sqrt{r^2-\left(\frac{\nu-1/2}{N_d}\right)^2}\Big|_{r=(j-1)N_0/R}^{jN_0/R} & \dots\ \frac{\nu-1/2}{N_d} \le \frac{j-1}{N_0} \\ 2\sqrt{\left(\frac{jR}{N_0}\right)^2-\left(\frac{\nu-1/2}{N_d}\right)^2} & \dots\ \frac{j-1}{N_0} < \frac{\nu-1/2}{N_d} \le \frac{j}{N_0} \\ 0 & \dots\ \frac{\nu-1/2}{N_d} > \frac{j}{N_0} \end{cases} \qquad .$$

Um glatte Kurven für die rekonstruierten Dichten zu bekommen, wird zusätzlich noch eine Spline-Interpolation für die Dichten durchgeführt. Die N_0 Werte der Dichte werden durch eine glatte Interpolation mit $N < N_0$ Knoten erzeugt.
Da die Spline-Interpotation eine lineare Abbildung ist, kann man sie durch eine Matrix $S_{ji}, j = 1,\dots,N_0,\ i = 1,\dots,N$ ausdrücken, die die N_0 interpolierten Dichten mit den N Dichteknoten verknüpft. Die explizite Form von S kann man in diversen Numerikbüchern nachschlagen. Die Matrix M des Modells entsteht also durch eine Matrixmultiplikation

$$M_{\nu i} = \sum_{j=1}^{N_0} A_{\nu j} S_{ji} \qquad .$$

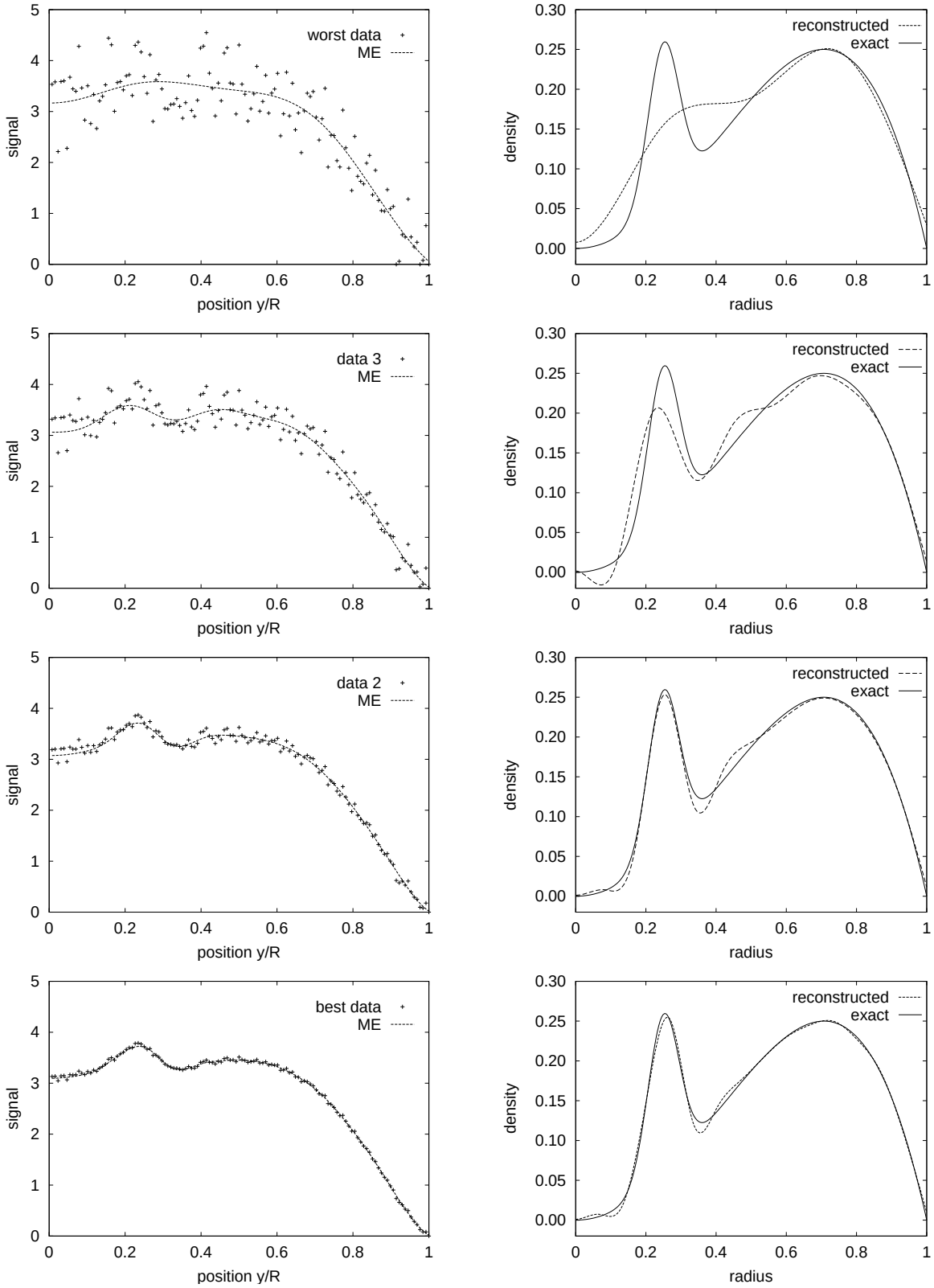

Abbildung 18.11: *Daten und rekonstruierte Dichten für vier verschieden stark verrauschte Datensätze.*

Abbildung (18.11) zeigt einige Resultate einer Abel-Inversion. Die Daten (linke Spalte) sind durch eine direkte Abel-Transformation der Funktion

$$\rho(r) = r^2\,(1 - r^2) + \frac{1}{5}\,e^{-250(r-1/4)^2}$$

entstanden. Ein normalverteiltes Rauschen mit Standard-Abweichung $\sigma = 0.65, 0.46, 0.29, 0.17$ (von oben nach unten) und Mittelwert null ist addiert worden. In allen Plots ist die Zahl der Daten $N_d = 128$ und $N_0 = 256$.
Für eine gegebene Zahl N von Dichteknoten ist der Hyper-Parameter α derart iterativ bestimmt worden, dass das Kriterium $\chi^2/N_d \approx 1$ erfüllt ist. Das ist eine alternative Vorgehensweise, anstatt das Maximum der Wahrscheinlichkeit für α zu bestimmen. Obiges Kriterium garantiert, dass der mittlere Fehler der Interpolation dem Standard-Fehler der Daten entspricht.
Die Anzahl der Dichteknoten N wird optimiert, indem man das Maximum der Wahrscheinlichkeit $P(N|\underline{d}, \mathcal{B})$ sucht. In den gezeigten Fällen haben sich die Werte $N = 5, 10, 13, 16$ (von oben nach unten) ergeben. Für zu kleine N gibt es kein α, das das Kriterium $\chi^2/N_d \approx 1$ erfüllt, da die Spline-Interpolation zu steif ist und $\alpha \to 0$. In der Nähe des optimalen N nimmt α sein Maximum an. Erhöht man N zu weit, fängt die rekonstruierte Dichte zu oszillieren an, wenn α verkleinert wird.

Teil IV

Parameterschätzen

Kapitel 19

Entscheidungstheorie

19.1 Elemente der Entscheidungstheorie

Eine Entscheidung ist eine Auswahl aus alternativen AKTIONEN. Als Grundlage für die Entscheidung stehen Daten oder Informationen zur Verfügung. Das Ergebnis wird durch den ZUSTAND charakterisiert. Die Aktionen müssen nicht unbedingt einen Einfluss auf den Endzustand haben. Sie können die Bedeutung haben, dass man auf den Zustand optimal vorbereitet sein will.
Wir interessieren uns nur für solche Situationen, in denen es keine deterministische Beziehung zwischen den Handlungen und den Resultaten gibt. Für die Resultate existiert eine Bewertungsfunktion. Die Entscheidungstheorie beschäftigt sich mit der optimalen Auswahl von Handlungen angesichts von Unwägbarkeiten. Die Unsicherheit kann in der Beziehung zwischen Handlung und Resultat oder in der Zuverlässigkeit der zugrundeliegenden Information (Daten) liegen. Es gibt im Wesentlichen drei Elemente der Entscheidungstheorie:

Information in Form von Daten $D_i, \qquad i = 1, \ldots, N_D$ und Vorwissen $\mathcal{B}$.

Handlungen $A_j, j = 1, \ldots, N_A$, die die möglichen Alternativen aufzeigen

Zustände $Z_l, l = 1, \ldots, N_Z$, die die Konsequenzen repräsentieren.

BEISPIEL Ein Mann liest einen Wetterbericht. Die möglichen Informationen sind

D_1: Der Wetterbericht sagt Sonnenschein voraus.

D_2: Der Wetterbericht sagt Regen voraus.

D_3: Der Wetterbericht sagt wechselhaftes Wetter voraus.

D_4: Der Wetterbericht sagt Schnee voraus.

Es gibt mehrere Aktionen, aus denen der Mann wählen kann

A_1: Er bleibt den ganzen Tag im Haus

A_2: Er zieht seine Winterkleidung an

A_3: Er zieht Sommerkleidung an

A_4: Er nimmt einen Regenschirm mit

Schließlich gibt es mehrere Endzustände

Z_1: Die Sonne scheint

Z_2: Es regnet

Z_3: Es schneit

Z_4: Es gibt einen Sturm

Für jedes Tripel (D_i, A_j, Z_l) benötigen wir eine Bewertungsfunktion, die wir als

$$K_{i,j,l} = K(D_i, A_j, Z_l)$$

darstellten. Zum Beispiel könnte man im obigen Beispiel die Situation (D_1, A_3, Z_2) mit $K_{1,3,2} = -10 \in$ bewerten, da der Mann ein Taxi nehmen muss.
Es gibt drei gängige Unterteilungen

pragmatisch: $K = K(A_j, Z_l)$
rituell: $K = K(A_j, D_i)$
gemischt: $K = K(A_j, Z_l, D_i)$

Der Bedingungskomplex enthält alle Parameter, die zusätzlich benötigt werden.
Die erwarteten Kosten sind

$$\begin{aligned} E(K|\mathcal{B}) &= \sum_K K\ P(K|\mathcal{B}) \\ &= \sum_{A_j,D_i,Z_l} \sum_K K\ \underbrace{P(K|A_j, D_i, Z_l, \mathcal{B})}_{\delta_{K,K(A_j,D_i,Z_l,\mathcal{B})}}\ P(A_j, D_i, Z_l|\mathcal{B}) \\ &= \sum_{A_j,D_i,Z_l} K(A_j, D_i, Z_l, \mathcal{B})\ P(A_j, D_i, Z_l|\mathcal{B}) \qquad . \end{aligned}$$

Es hängt nun vom Problem ab, in welcher Form die Wahrscheinlichkeit $P(A_j, D_i, Z_l|\mathcal{B})$ weiter zerlegt werden kann. Wenn die Handlung keinen Einfluss auf den Zustand hat und der Zustand zum Zeitpunkt der Handlung noch nicht vorliegt oder zumindest nicht bekannt ist, ist folgende Zerlegung sinnvoll

$$\begin{aligned} E(K|\mathcal{B}) &= \sum_{A_j,D_i,Z_l} K(A_j, D_i, Z_l, \mathcal{B})\ P(A_j|D_i, Z_l, \mathcal{B})\ P(D_i|Z_l, \mathcal{B})\ P(Z_l|\mathcal{B}) \\ &= \sum_{A_j,D_i,Z_l} K(A_j, D_i, Z_l, \mathcal{B})\ P(A_j|D_i, \mathcal{B})\ P(D_i|Z_l, \mathcal{B})\ P(Z_l|\mathcal{B}) \qquad . \end{aligned} \tag{19.1}$$

Im letzten Schritt wurde ausgenutzt, dass die Aktion logisch unabhängig vom Zustand ist, da für die Entscheidung nur die Daten zur Verfügung stehen. Wenn die Aktion hingegen den Endzustand beeinflusst, ist die Zerlegung

$$E(K|\mathcal{B}) = \sum_{A_j,D_i,Z_l} K(A_j, D_i, Z_l, \mathcal{B})\ P(Z_l|A_j, D_i, \mathcal{B})\ P(A_j|D_i, \mathcal{B})\ P(D_i|\mathcal{B}) \quad . \tag{19.2}$$

vorzuziehen.

19.1.1 Beispiel: Qualitätskontrolle

Eine Firma stelle elektronische Bauelemente her, die zu 1000 Stück in Schachteln verpackt werden. Es stellt sich im Nachhinein heraus, dass die Hälfte der Produktion fehlerhaft ist. Aufgrund des Verpackungsprozesses hat das dazu geführt, dass in der Hälfte der Pakete 10% der Teile defekt sind und in der anderen Hälfte 90%. Der Käufer teilt mit, dass für ihn eine Defektrate von 10% akzeptabel ist und er hierfür einen Preis von $25 \in$ bezahlt. Wenn er jedoch die zu 90% defekten Teile einbaut entstehen im hierdurch Mehrkosten in Höhe von $40 \in$, die er der Zulieferfirma in Rechnung stellen wird.
Die Firma hat nun die Möglichkeit, alle Teile in allen Kartons zu testen. Die Überprüfung eines Teils koste $0.2 \in$. Eine Kartonladung (1000 Stück) neu zu produzieren kostet hingegen $20 \in$. Es lohnt sich deshalb nicht, alle Teile zu prüfen. Die Firma beschließt, zu jedem Karton eine Stichprobe vom Umfang $N \leq 20$ durchzuführen und erarbeitet ein Entscheidungskriterium, bei welcher Zahl von defekten Teilen in der Stichprobe der Karton weggeschmissen und der Inhalt neu produziert werden soll. Die Kostenfunktion ist pragmatisch und ist in der nachstehenden Tabelle zusammengefasst.

	Zustand	Aktion	Kosten
K_{11}	gut	versenden	-25
K_{21}	schlecht	versenden	40
K_{12}	gut	neu produzieren	20
K_{22}	schlecht	neu produzieren	20

Negative Kosten bedeuten Gewinn. Es vereinfacht die nachfolgende Rechnung, wenn wir den „Kosten-Nullpunkt" auf $K_0 = 20 \in$ legen. Dadurch verändert sich die Kostenfunktion zu

	Zustand	Aktion	Kosten
$\tilde{K}_{11}$	gut	versenden	-45
$\tilde{K}_{21}$	schlecht	versenden	20
$\tilde{K}_{12}$	gut	neu produzieren	0
$\tilde{K}_{22}$	schlecht	neu produzieren	0

Zu diesen Karton-bezogenen Kosten kommen noch die Kosten für die Prüfung $K_P N$ ($K_P = 0.2$ €) hinzu. Die Gesamtkosten sind somit

$$K(Z_l, A_j, D_i) = K_0 + K_P \cdot N_i + \tilde{K}_{l,j} \qquad .$$

Die Daten D_i bestehen aus dem Werte-Paar (N_i, n_i), dem Stichproben-Umfang N_i und der Zahl der defekten Bauelemente n_i in der Stichprobe. Die Strategie wird sein, immer denselben Stichproben-Umfang $N_i = N$ zu verwenden, das heißt

$$P(D_i|Z_l, \mathcal{B}) = P(n|Z_l, N, \mathcal{B}) = \binom{N}{n} q_l^n \, (1 - q_l)^{N-n} \qquad .$$

Der Zustand gibt in diesem Beispiel an, ob der Karton gut ($q = 0.1$) oder schlecht ($q = .9$) ist. Dieser Zustand wird durch die Aktion, versenden oder neu produzieren, nicht beeinflusst. Der Zustand liegt zwar zum Zeitpunkt der Entscheidung bereits vor, ist aber nicht bekannt. Aus diesem Grund werden die Stichproben entnommen. Das heißt, $P(A_j|D_i, Z_l, \mathcal{B}) = P(A_j|D_i, \mathcal{B})$, die Handlung A_j ist logisch unabhängig vom Zustand Z_l. Die mittleren Kosten sind somit gemäß Gl. (19.1)

$$\begin{aligned}
E(K|\mathcal{B}) &= K_0 + K_P \cdot N \; + \sum_{n=0}^{N} \sum_{A_j, Z_l} \tilde{K}_{l,j} \; P(A_j|D_i, \mathcal{B}) \; P(D_i|Z_l, \mathcal{B}) \; P(Z_l|\mathcal{B}) \\
&= K_0 + K_P \cdot N \; + \tfrac{1}{2} \sum_{n=0}^{N} \sum_{l=1}^{2} \tilde{K}_{l,1} \; P(A_1|n, N, \mathcal{B}) \; P(n|Z_l, N, \mathcal{B}) \\
&= K_0 + K_P \cdot N \\
&+ \tfrac{1}{2} \sum_{n=0}^{N} \binom{N}{n} P(A_1|n, N, \mathcal{B}) \underbrace{\left(\sum_{l=1}^{2} \tilde{K}_{l,1} \; q_l^n \; (1 - q_l)^{N-n} \right)}_{f(n)}
\end{aligned}$$

Es wurde ausgenutzt, dass $P(Z_l|\mathcal{B}) = 1/2$ und $\tilde{K}_{Z_l, A_j=2} = 0$. Im Zustand Z_1 ist $q = 0.1$ und somit ist

$$\tilde{K}_{11} q_1^n (1 - q_1)^{N-n} = - \left| \tilde{K}_{11} \right| q_1^n (1 - q_1)^{N-n}$$

als Funktion von n monoton steigend. Im anderen Fall ($q = .9$) ist diese Funktion ebenfalls monoton steigend, da gleich $\tilde{K}_{21} > 0$ ist. Das heißt, $f(n)$ ist am kleinsten für $n = 0$ und maximal für $n = N$. Der Minimalwert ist

$$f(0) = - \left| \tilde{K}_{11} \right| (1 - q_1)^N + \tilde{K}_{21} (1 - q_2)^N = -0.1^N \left(9^N \; \left| \tilde{K}_{11} \right| - \tilde{K}_{21} \right) < 0 \qquad \forall N$$

und der Maximalwert lautet

$$f(N) = - \left| \tilde{K}_{11} \right| q_1^N + \tilde{K}_{21} q_2^N = 0.1^N \left(- \left| \tilde{K}_{11} \right| + 9^N \; \tilde{K}_{21} \right) > 0 \qquad \forall N > 0 \qquad .$$

Die Funktion ist in Abbildung 19.1 für $N = 5$ dargestellt. Die Kosten sollen minimiert werden. Die beste Strategie ist demnach

$$P(A_1|n, N, \mathcal{B}) = \begin{cases} 1 & \text{wenn } f(n) < 0 \\ 0 & \text{sonst.} \end{cases}$$

Wir erhalten somit

$$E(K|\mathcal{B}) = K_0 + K_P \cdot N + \tfrac{1}{2} \sum_{\substack{n=0 \\ f(n)<0}}^{N} \binom{N}{n} f(n) \qquad .$$

Die Kosten sind in Abbildung 19.1 als Funktion des Stichprobenumfangs N aufgetragen. Das Minimum (Optimum) liegt bei $N = 5$ mit einem Wert von -1.2218 €. Die

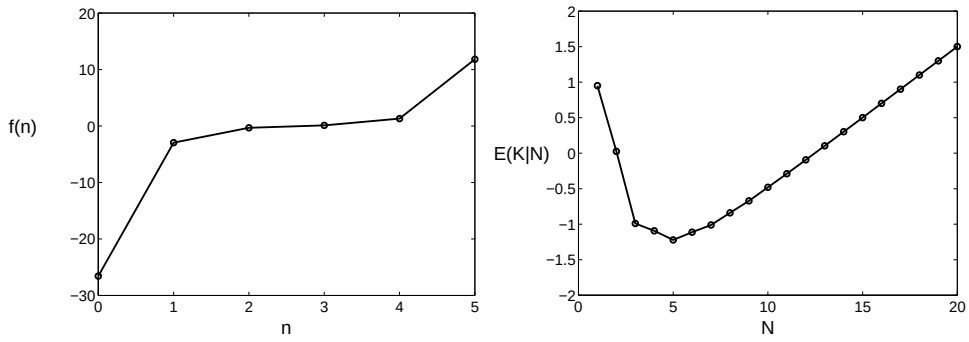

Abbildung 19.1: Links: $f(n)$ mit den im Text angegebenen Parametern. Rechts: mittlere Kosten als Funktion des Stichprobenumfangs.

Tatsache, dass die Hälfte der Bauelemente defekt ist, hat also den Gewinn pro Karton von 25 € auf 1.22 € reduziert.

19.1.2 Beispiel: Optimale Produktionsrate

Die Herstellungskosten eines Produkt seinen pro Stück H. Es werde zu einem Preis $\mathcal{P}$ verkauft. Die Fixkosten der Firma betragen pro Tag F. Die Produktionsrate pro Tag sei N. Wenn die Nachfrage pro Tag n beträgt, ist die Zahl der verkauften Teile $\min(n, N)$. Die Kosten pro Tag sind somit

$$K(n, N) = -\min(n, N)\, \mathcal{P} + NH + F \qquad .$$

Wir identifizieren in diesem Problem den Zustand mit der Nachfrage

$$Z = n$$

und die möglichen Aktionen entsprechen der Produktionsrate

$$A = N \qquad .$$

Die vorliegenden Daten enthalten die Werte der Nachfrage der letzten Wochen. Es habe sich hierbei ergeben, dass die Nachfrage einer Wahrscheinlichkeitsverteilung $P(n|\mathcal{B})$ genügt. Die Strategie soll sein, über längere Zeit denselben Wert N zu verwenden.
Die mittleren Kosten erhalten wir aus

$$\begin{aligned} E(K|N,\mathcal{B}) &= \sum_n \; K(n,N)\; P(n|\mathcal{B}) \\ &= \sum_n \; (-\min(n,N)\;\mathcal{P} + NH + F)\;\; P(n|\mathcal{B}) \\ &= F + NH - \mathcal{P}\;\left(\sum_{n=0}^{N} n\; P(n|\mathcal{B}) + N \sum_{n=N+1}^{\infty} P(n|\mathcal{B})\right) \qquad . \end{aligned}$$

Um analytisch rechnen zu können, verwenden wir zunächst eine flache Wahrscheinlichkeitsverteilung für die Nachfrage pro Tag

$$P(n|\mu,\mathcal{B}) = \frac{1}{\mu+1}\;\theta(n \le \mu) \qquad ,$$

mit einer Obergrenze μ. Damit erhalten wir

$$\begin{aligned} E(K|N,\mu,\mathcal{B}) &= F + NH - \frac{\mathcal{P}}{\mu+1}\;\left(\sum_{n=0}^{N} n\; - N \sum_{n=N+1}^{\mu} 1\right) \\ &= F + NH - \frac{\mathcal{P}}{\mu+1}\;\left(\frac{(N+1)N}{2}\; + N\;(\mu - N)\right) \\ &= F + NH - \frac{\mathcal{P}}{\mu+1}\;\left(\frac{N}{2}\; + N\;\mu - \frac{N^2}{2}\right) \qquad . \end{aligned}$$

Das Minimum erhalten wir aus der Ableitung nach N[1]

$$\begin{aligned} \frac{d}{dN}E(K|N,\mu,\mathcal{B}) &= H - \frac{\mathcal{P}}{\mu+1}\;\left(\frac{1}{2}\; + \mu - N\right) \stackrel{!}{=} 0 \\ &\Rightarrow \\ N^* &= \frac{1}{2} + \mu - \frac{(\mu+1)H}{\mathcal{P}} = -\frac{1}{2} + (\mu+1)\frac{\mathcal{P}-H}{\mathcal{P}} \qquad . \end{aligned}$$

Wir schreiben den Preis als Vielfaches der Herstellungskosten

$$\mathcal{P} = \nu\; H$$

[1]Wir betrachten hierzu N zunächst als kontinuierliche Variable. Der optimale Wert N ist dann einer der beiden ganzzahligen Nachbarn.

und erhalten

$$N^* = -\frac{1}{2} + (\mu + 1)\ \frac{\nu - 1}{\nu} \qquad .$$

Das heißt z.B., wenn $\nu = \frac{P}{H} = 2$, dass die optimale Produktionsrate

$$N^* = -\frac{1}{2} + \frac{\mu + 1}{2} = \frac{\mu}{2}$$

der Hälfte der maximalen Kundenzahl entspricht.
Für $\mu \gg 1$ ist

$$N^* = \mu\ \frac{\nu - 1}{\nu} \qquad ,$$

und die mittleren Kosten für $N = N^*$ ergeben

$$\begin{aligned}
E(K|\mathcal{B}) &= F + N^* H\ \left(1 - \frac{\nu}{\mu + 1}\ \left(\frac{1}{2}\ + \mu - \frac{N^*}{2}\right)\right) \\
&\simeq F + \mu\ \frac{\nu - 1}{\nu}\ H\ \left(1 - \frac{\nu}{\mu}\ \left(\mu - \frac{\mu(\nu - 1)}{2\nu}\right)\right) \\
&= F + \mu\ \frac{\nu - 1}{\nu}\ H\ \left(1 - \nu\ + \frac{\nu - 1}{2}\right) \\
&= F - \frac{\mu}{2}\ \frac{(\nu - 1)^2}{\nu}\ H \\
&= F - \frac{\mu}{2}\ \frac{(P - H)^2}{P} \qquad .
\end{aligned}$$

Dieses Ergebnis könnte zu der Annahme verleiten, dass der Gewinn durch Anheben des Preises beliebig erhöht werden kann. Das ist nicht richtig, da die maximale Kundenzahl μ mit zunehmendem Preis sinken wird. Mit der Modell-Annahme

$$\mu = \frac{C}{P^\kappa}$$

erhält man

$$E(K|\mathcal{B}) = F - \frac{C}{2}\ \frac{(P - H)^2}{P^{\kappa+1}} \qquad .$$

Die Ableitung nach P ergibt

$$\begin{aligned}
E(K|\mathcal{B}) &= -\frac{C}{2}\ \frac{(P - H)}{P^{\kappa+2}}\ (2P - (\kappa + 1)(P - H)) \\
&= -\frac{C}{2}\ \frac{(P - H)}{P^{\kappa+2}}\ ((\kappa + 1)H - (\kappa - 1)P) \stackrel{!}{=} 0 \qquad\qquad \Rightarrow \\
P &= \frac{\kappa + 1}{\kappa - 1}\ H \qquad .
\end{aligned}$$

Man kann dieses Beispiel leicht realistischer gestalten. Man wird dann die Optimierung jedoch numerisch durchführen müssen.

Kapitel 20

Parameter-Schätzen

20.1 Unverzerrte Schätzwerte

Es sei ein Datenanalyse-Problem gegeben, das von Parametern $a \in \mathbb{R}^{N_p}$ abhängt. Es werden Messwerte (Stichprobe) $\underline{y} = \{y_1, \ldots, y_N\}$ ermittelt, die dazu dienen sollen, die Parameter a zu bestimmen. Die zugrundeliegenden Parameter definieren die Wahrscheinlichkeitsdichte der Stichprobe, das heißt der Likelihood-Funktion

$$p(\underline{y}|a) \qquad .$$

Ein Beispiel stellen additive Gaußsche Fehler dar

$$p(\underline{y}|a) = (2\pi\sigma^2)^{-\frac{N}{2}} \; e^{-\frac{1}{2}\sum_i \frac{(y_i - f_i(a))^2}{\sigma_i^2}} \qquad . \tag{20.1}$$

Hierbei ist $f_i(a) = f(s_i, a)$ das theoretische Modell, das von Steuergrößen s_i und den zu bestimmenden Parametern a abhängt. Betrachten wir z.B. eine gedämpfte Schwingung, die zur Zeit t die Auslenkung

$$f(t|A_0, \omega, \lambda, \varphi) = A(t) = A_0 \; \cos(\omega \; t + \varphi) \; e^{-\lambda t}$$

aufweist. Zu verschiedenen Zeiten t_i wird die Amplitude experimentell bestimmt. Der Messwert zur Zeit t_i sei y_i. In diesem Beispiel sind die Steuergrößen die Zeiten t_i und die unbekannten Parameter sind $a = \{A_0, \omega, \lambda, \varphi\}$. Die Funktion $f(s, a)$ entspricht der Auslenkung $A(t)$. Die Messwerte streuen gemäß einer Gauß-Verteilung mit einer Standardabweichung σ um die theoretischen Werte. Aus der Stichprobe sollen nun die vier unbekannten Parameter bestimmt werden.
Der Einfachheit halber betrachten wir zunächst nur Probleme mit einem unbekannten Parameter a. Im Rahmen der orthodoxen Statistik sucht man ein Funktional $\hat{a}(\underline{y})$, das einen Schätzwert für den Parameter bei gegebener Stichprobe liefert. Von besonderer Bedeutung sind solche Funktionale die UNVERZERRTE SCHÄTZWERTE (UNBIASED ESTIMATOR) liefern. Als unverzerrt werden solche Schätzwerte bezeichnet, die

bei ein und demselben Problem, d.h. auch bei denselben Parameterwerten, gemittelt über alle denkbaren Stichproben den exakten Wert liefern. Das heißt

$$\langle \hat{a}(\underline{y}) \rangle = \int d^N y \; \hat{a}(\underline{y}) \; p(\underline{y}|a) \stackrel{!}{=} a \qquad . \tag{20.2}$$

20.2 Maximum-Likelihood Schätzwert

Unter dem Maximum-Likelihood (ML) Schätzwert versteht man die Lösung der Maximierungsaufgabe

$$\max_a \; p(\underline{y}|a)$$

oder äquivalent

$$\left. \frac{\partial \ln(p(\underline{y}|a))}{\partial a_i} \right|_{a=\hat{a}^{\mathrm{ML}}} = 0 \qquad .$$

20.2.1 Beispiel: univariate Normal-Verteilung

Als Beispiel betrachten wir die einfachste Variante des obigen Problems

$$p(\underline{y}|a) = (2\pi\sigma^2)^{-\frac{N}{2}} \; e^{-\frac{1}{2\sigma^2} \; \sum_i (y_i - a)^2} \qquad ,$$

bei der die Messungen direkt den gesuchten Parameter liefert, z.B. Längenmessung. Die LOG-LIKELIHOOD lautet

$$\ln(p(\underline{y}|a)) = C - \frac{1}{2\sigma^2} \; \sum_i (y_i - a)^2 \qquad .$$

Die Nullstelle der Ableitung nach a liefert

$$\hat{a}^{\mathrm{ML}} = \langle y \rangle \qquad .$$

In diesem Fall stimmt der ML-Schätzwert mit dem arithmetischen Mittel überein. Wir wissen bereits, dass dieser Schätzwert unverzerrt ist.
Die ML-Lösung soll noch an einem weiteren Beispiel illustriert werden.

20.2.2 Beispiel: Halbwertszeit eines Poissonprozesses

Gegeben sei eine Stichprobe von Zerfallszeiten $\{t_1, \ldots, t_L\}$. Die Wahrscheinlichkeit für einen Zerfall in $(t, t + dt)$ ist durch die Exponential-Verteilung

$$p_P(t|\tau) = \frac{1}{\tau} \; e^{-t/\tau}$$

gegeben. Die Log-Likelihood lautet, da die Zerfälle unkorreliert sind,

$$\begin{aligned}\ln(p(t_1,\ldots,t_L|\tau,\mathcal{B})) &= \sum_{i=1}^{L} \ln\left(\frac{1}{\tau}\, e^{-t_i/\tau}\right) \\ &= \sum_{i=1}^{L} \left(-\frac{t_i}{\tau} - \ln(\tau)\right) \\ &= -L\, \left(\frac{\langle t\rangle}{\tau} + \ln(\tau)\right) \qquad .\end{aligned}$$

Hierbei ist $\langle t\rangle$ der Stichproben-Mittelwert der Zerfallszeiten. Die Ableitung verschwindet an der Stelle

$$\tau^{\mathrm{ML}} = \langle t\rangle \qquad .$$

Somit ist auch in diesem Beispiel der ML-Schätzwert gleich dem arithmetischen Mittel. Das ist nicht immer so, wie das folgende Beispiel zeigt

20.2.3 Cauchy-Verteilung

Gegeben sei eine Stichprobe $\{x_1,\ldots,x_L\}$ der Cauchy-Verteilung mit der Wahrscheinlichkeitsdichte

$$p_C(x|a,b) = \frac{1}{\pi}\, \frac{b}{(x-a)^2+b^2} \qquad .$$

Die Log-Likelihood lautet, da die Elemente der Stichprobe unabhängig sind,

$$\begin{aligned}\ln(p(x_1,\ldots,x_L|a,b,\mathcal{B}) &= \sum_{i=1}^{L} \ln\left(\frac{1}{\pi}\, \frac{b}{(x_i-a)^2+b^2}\right) \\ &= C + L\ln(b) - \sum_{i=1}^{L} \ln\left((x_i-a)^2+b^2\right) \qquad .\end{aligned}$$

Die Ableitung nach b ergibt

$$\begin{aligned}\frac{\partial}{\partial b}\ln(p(x_1,\ldots,x_L|a,b,\mathcal{B})) &= \frac{L}{b} - b\, \sum_{i=1}^{L}\, \frac{2}{(x_i-a)^2+b^2} \stackrel{!}{=} 0 \\ &\Rightarrow \\ b^2 &= \left(\frac{1}{L}\sum_{i=1}^{L}\, \frac{2}{(x_i-a)^2+b^2}\right)^{-1} \qquad .\end{aligned}$$

Die Ableitung nach a liefert

$$\frac{\partial}{\partial a} \ln(p(x_1, \ldots, x_L|a, b, \mathcal{B})) = 2 \sum_{i=1}^{L} \frac{x_i - a}{(x_i - a)^2 + b^2} \stackrel{!}{=} 0 \qquad .$$

Damit haben wir als ML-Lösung

$$a = \frac{\frac{1}{L}\sum_{i=1}^{L} \frac{x_i}{(x_i-a)^2+b^2}}{\frac{1}{L}\sum_{i=1}^{L} \frac{1}{(x_i-a)^2+b^2}}$$
$$b^2 = \frac{1}{\frac{2}{L}\sum_{i=1}^{L} \frac{1}{(x_i-a)^2+b^2}} \qquad .$$

ein gekoppeltes System transzendenter Gleichungen vor.

20.2.4 Bernoulli-Problem

Wir betrachten noch einmal das binäre Bernoulli-Problem mit gelben und roten Kugeln. Es werde eine Stichprobe vom Umfang N mit Zurücklegen gezogen. Davon seien n_g Kugeln gelb. Wir wollen hieraus den Wert des Verhältnisses $q = n_g/N_p$ der Population abschätzen. Die Likelihood-Funktion ist in diesem Fall

$$P(n_g|q, N, \mathcal{B}) = \binom{N}{n_g} q^{n_g} (1-q)^{N-n_g} \qquad .$$

Die ML-Lösung erhalten wir aus Nullstelle der Ableitung des Logarithmus

$$\frac{d}{dq}\left(n_g \ln(q) + (N - n_g)\ln(1-q)\right) = \frac{n_g}{q} - \frac{N-n_g}{1-q} \tag{20.3}$$
$$\stackrel{!}{=} 0$$
$$\Rightarrow \qquad q_{\mathrm{ML}} = \frac{n_g}{N} \qquad . \tag{20.4}$$

Wir wollen noch untersuchen, ob es sich um einen unverzerrten Schätzwert handelt. Dazu benötigen wir die Schätzwert-Verteilung

$$p(q_{\mathrm{ML}}|N, q, \mathcal{B}) = \sum_{k=0}^{N} p(q_{\mathrm{ML}}|k, q, N, \mathcal{B})\ P(k|q, N, \mathcal{B})$$
$$= \sum_{k=0}^{N} \delta(q_{\mathrm{ML}} - \tfrac{k}{N})\ P(k|q, N, \mathcal{B})$$

Die Momente dieser Verteilung sind

$$\begin{aligned}\langle q_{\mathrm{ML}}^{\nu}\rangle &= \int_0^1 q_{\mathrm{ML}}^{\nu}\; p(q_{\mathrm{ML}}|q,N,\mathcal{B})\; dq_{\mathrm{ML}} \\ &= \int_0^1 \sum_{k=0}^{N} q_{\mathrm{ML}}^{\nu}\; \delta(q_{\mathrm{ML}} - \tfrac{k}{N})\; P(k|q,N,\mathcal{B})\; dq_{\mathrm{ML}} \\ &= \frac{1}{N^{\nu}} \sum_{k=0}^{N} k^{\nu}\; P(k|q,N,\mathcal{B}) = \frac{\langle k^{\nu}\rangle}{N^{\nu}} \qquad .\end{aligned}$$

Daraus ergibt sich

$$\begin{aligned}\langle q_{\mathrm{ML}}\rangle &= \frac{\langle k\rangle}{N} \\ \mathrm{var}(q_{\mathrm{ML}}) &= \frac{\mathrm{var}(k)}{N^2}\end{aligned}$$

und mit dem Mittelwert und der Varianz der Binomial-Verteilung erhalten wir schließlich

LÖSUNG DES BERNOULLI-PROBLEMS

$$\langle q_{\mathrm{ML}}\rangle = \frac{N\,q}{N} = q \tag{20.5}$$

$$\mathrm{var}(q_{\mathrm{ML}}) = \frac{N\,q\,(1-q)}{N^2} = \frac{q\,(1-q)}{N} \tag{20.6}$$

Der Schätzwert ist in der Tat unverzerrt und hat eine Varianz $q(1-q)/N$, die mit zunehmendem Stichprobenumfang abnimmt.

20.2.5 Abbruchskriterien bei Experimenten

Wir wollen untersuchen, welchen Einfluss die Abbruchskriterien eines Experiments auf das Ergebnis haben können. Dazu betrachten wir als einfaches Beispiel erneut das eben behandelte binäre Bernoulli-Problem aus Abschnitt 20.2.4. Da man vor dem Experiment nicht weiß, wie groß q ist, ist es auch schwer zu sagen, welchen Umfang N die Stichproben haben sollte, um ein zuverlässiges Ergebnis zu erhalten. Wenn N zu klein gewählt wird, kann es sein, dass wir fast ausschließlich $n_g = 0$ finden. Wäre es da nicht sinnvoller, den Stichprobenumfang vom Ausgang des Experimentes

abhängen zu lassen? Man könnte z.B. das Experiment solange durchzuführen, bis die Zahl der gelben Kugeln einen bestimmten Wert n_g^* erreicht hat?
Zur Untersuchung dieser Fragestellung wollen wir zunächst Ergebnisse einer Computersimulation diskutieren. Es wurden Stichproben vom Umfang $N = 10$ erzeugt. Zu jeder Stichprobe wurde die Anzahl n_g der gelben Kugeln ermittelt und daraus der Schätzwert $q_{\mathrm{ML}} = \frac{n_g}{N}$ berechnet. Dieses Experiment wurde N_{exp}-mal wiederholt und daraus der Mittelwert und die Standardabweichung berechnet. In Abbildung 20.1 ist das Ergebnis als Funktion von N_{exp} aufgetragen.

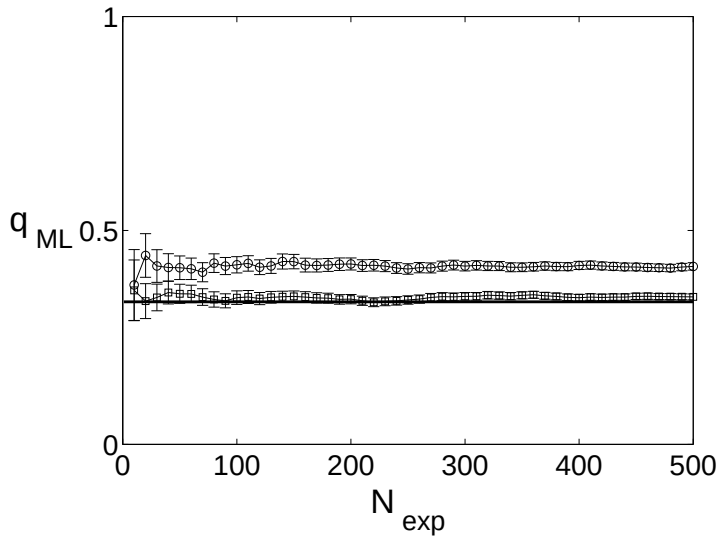

Abbildung 20.1: *Schätzwert für q aus*
a) Stichproben vom Umfang $N = 10$ (untere Kurve).
b) abgebrochene Stichproben, mit $n_g^* = 3$ (obere Kurve).
Die gestrichelte Linie kennzeichnet den intrinsische Wert $q = 1/3$.

Man erkennt, dass das arithmetische Mittel mit zunehmender Zahl der Experimente gegen den wahren Wert konvergiert und der Standardfehler monoton abnimmt. Aufgrund des ZENTRALEN GRENZWERTSATZES geht der Standardfehler mit $1/\sqrt{N_{\mathrm{exp}}}$ gegen Null.
Nun führen wir das Experiment durch, bei dem wir die Messung abbrechen, sobald genau 3 gelbe Kugeln vorliegen. Wieder verwenden wir $q_{\mathrm{ML}} = \frac{n_g}{N}$ als Schätzwert für q und bilden das arithmetische Mittel über viele Wiederholungen des Experiments. Wie zuvor nehmen die statistischen Schwankungen (Standardfehler) wie $1/\sqrt{N_{\mathrm{exp}}}$ ab. Nur konvergiert das Ergebnis gegen einen falschen Wert. Dieses Experiment hat einen BIAS!
Dieses Ergebnis soll nun analytisch, mit den Regeln der Wahrscheinlichkeitstheorie berechnet werden. Das heißt, wir werden die Verteilung $p(q_{\mathrm{ML}}|q, n_g^*, \mathcal{B})$ der ML-Schätzwerte bestimmen, die sich ergibt, wenn das Experiment immer bei $n_g = n_g^*$ abgebrochen wird. Diese Information ist Teil des Bedingungskomplexes $\mathcal{B}$.

Es ist an der Zeit, noch einmal darauf hinzuweisen, dass im Bezug auf den Bedingungskomplex große Vorsicht geboten ist. Er ist gewissem Sinne der wichtigste Teil der bedingten Wahrscheinlichkeit und wird aber in der Regel am sorglosesten behandelt. Man verwendet z.B. immer dasselbe Symbol für den Bedingungskomplex, obwohl sich seine Bedeutung von Problem zu Problem ändert. Es gilt generell

$$P(A|C,\mathcal{B}) \neq P(A|C,\mathcal{B}') \qquad .$$

Man muss also aufpassen, dass innerhalb einer Rechnung tatsächlich immer derselbe Bedingungskomplex vorliegt.
Um $p(q_{\mathrm{ML}}|q,n_g^*,\mathcal{B})$ berechnen zu können, benötigen wir noch die Information über den Stichprobenumfang, den wir hier mit N^* kennzeichnen wollen. Die Marginalisierungsregel liefert

$$p(q_{\mathrm{ML}}|q,n_g^*,\mathcal{B}) = \sum_{N^*=1}^{\infty} p(q_{\mathrm{ML}}|N^*,q,n_g^*,\mathcal{B})\; P(N^*|q,n_g^*,\mathcal{B}) \qquad . \tag{20.7}$$

Der erste Faktor ist nun bekannt, da der Bedingungskomplex besagt, dass das Experiment bei $n_g = n_g^*$ abgebrochen wird und gleichzeitig der Stichproben-Umfang N^* vorliegt. Damit liegt der Wert des ML-Schätzwertes fest $q_{\mathrm{ML}} = \frac{n_g^*}{N^*}$. Das heißt,

$$p(q_{\mathrm{ML}}|N^*,q,n_g^*,\mathcal{B}) = \delta(q_{\mathrm{ML}} - \frac{n_g^*}{N^*}) \qquad .$$

Somit liefert Gl. (20.7)

$$p(q_{\mathrm{ML}}|q,n_g^*,\mathcal{B}) = \sum_{N^*=1}^{\infty} \delta(q_{\mathrm{ML}} - \frac{n_g^*}{N^*})\; P(N^*|q,n_g^*,\mathcal{B}) \qquad . \tag{20.8}$$

Statistik der Stichprobenumfänge

Um Gl. (20.8) weiter auswerten zu können, benötigen wir die Verteilung $P(N^*|q,n_g^*,\mathcal{B})$ der Stichprobenumfänge N^*, die sich einstellen, wenn das Experiment bei $n_g = n_g^*$ abgebrochen wird. Das Ergebnis der Computersimulation für $P(N^*|q,n_g^*,\mathcal{B})$ ist in Abbildung 20.2 als Histogramm aufgetragen. Wir wollen dieses Ergebnis auch mit den Regeln der Bayessche Wahrscheinlichkeitstheorie berechnen. Damit die N^*-te Kugel die n_g^*-te gelbe ist, müssen unter den ersten N^*-1 Kugeln in beliebiger Reihenfolge n_g^*-1 gelbe Kugeln vorkommen und die N^*-te ebenfalls gelb sein. Da die einzelnen Kugeln unabhängig voneinander gezogen werden, gilt

$$\begin{aligned} P(N^*|n_g^*,q,\mathcal{B}) &= P(n_g = n_g^*-1|q,N^*-1,\mathcal{B})\; P(n_g=1|q,N=1,\mathcal{B}) \\ &= \binom{N^*-1}{n_g^*-1}\, q^{n_g^*}\,(1-q)^{N^*-n_g^*} \end{aligned}$$

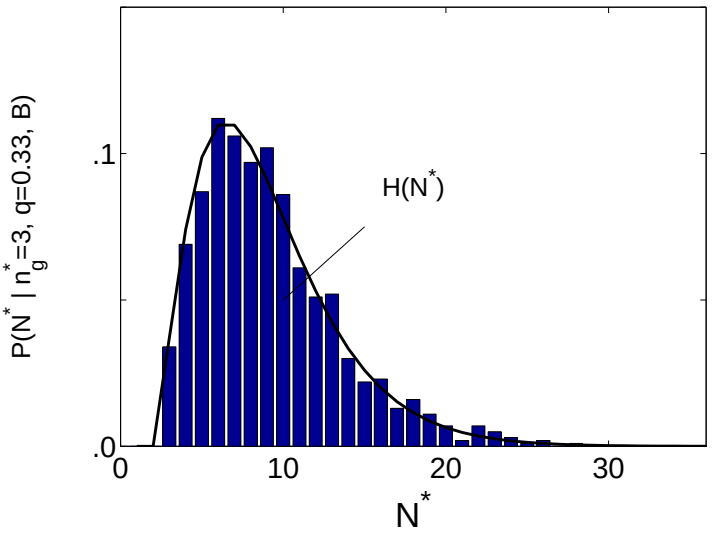

Abbildung 20.2: *Verteilung der Stichproben-Umfänge, die sich ergeben, wenn die Experimente bei $n_g = 3$ abgebrochen werden. Der Anteil der gelben Kugeln beträgt $q = 1/3$.*

Diese Wahrscheinlichkeit ist in Abbildung 20.2 mit dem Histogramm verglichen. Wir sehen, dass das theoretische Ergebnis mit dem der Computersimulation übereinstimmt.
Nachdem wir nun die Verteilung der Stichprobenumfänge kennen, können wir daraus die Verteilung der Schätzwerte $q_{\mathrm{ML}} = \frac{n_g^*}{N^*}$ mit Hilfe von Gl. (20.8) berechnen. Der Mittelwert dieser Verteilung ist

$$\begin{aligned}
\langle q_{\mathrm{ML}} \rangle &= \int_0^1 q_{\mathrm{ML}}\; p(q_{\mathrm{ML}}|n_g^*, q, \mathcal{B})\; dq_{\mathrm{ML}} \\
&= \sum_{N^*=n_g^*}^{\infty} \frac{n_g^*}{N^*}\; p(N^*|n_g^*, q, \mathcal{B}) \\
&= n_g^* \left(\frac{q}{1-q}\right)^{n_g^*} \sum_{N^*=n_g^*}^{\infty} \frac{1}{N^*} \binom{N^*-1}{n_g^*-1} (1-q)^{N^*} \qquad . \qquad (20.9)
\end{aligned}$$

Für $n_g^* = 1$ lässt sich das leicht auswerten

$$\begin{aligned}
\langle q_{\mathrm{ML}} \rangle &= \frac{q}{1-q} \sum_{N^*=1}^{\infty} \frac{1}{N^*} (1-q)^{N^*} = \frac{q}{1-q} \sum_{N^*=1}^{\infty} \int_{p=0}^{1-q} p^{N^*-1}\, dp \\
&= \frac{q}{1-q} \int_{p=0}^{1-q} \frac{1}{1-p}\, dp = -\frac{q}{1-q}\, \ln(q) \qquad .
\end{aligned}$$

Für $q = 1/3$ und $n_g^* = 1$ ist $\langle q_{\mathrm{ML}} \rangle = \frac{1}{2}\ln(3) = 0.549$, also weit entfernt vom korrekten Wert $q = 0.334$. Eine genauere Auswertung von Gl. (20.9) für $q = 1/3$ und $n_g^* = 3$ liefert $\langle q_{\mathrm{ML}} \rangle = \frac{3}{8}\ln(3) = 0.412$. Das ist immer noch weit vom korrekten Wert entfernt und erklärt den systematischen Fehler in Abbildung 20.1. Wir erkennen aber auch, dass das Abbruchskriterium für zunehmende Werte n_g^* an Einfluss verliert.

20.2.6 Least-Squares-Fit

Wir betrachten den sehr häufig vorliegenden Fall unabhängiger, normal-verteilter, additiver Fehler, mit der in Gl. (20.1) angegebenen Likelihood-Funktion. Maximieren der Log-Likelihood ist äquivalent zum Minimieren der gewichteten mittleren quadratischen Abweichung

LEAST-SQUARES-FIT

$$\min_a \sum_i \frac{(y_i - f(s_i, a))^2}{\sigma_i^2} \quad . \tag{20.10}$$

Der einfachste Spezialfall liegt vor, wenn $\sigma_i^2 = \sigma^2$.
Wir wollen nun noch korrelierte Fehler betrachten. Der Logarithmus der multivariaten Normalverteilung mit Kovarianz-Matrix C lässt sich wie folgt umformen

$$\begin{aligned}\log(\mathcal{L}) &= -\frac{1}{2}\ (y - f(a))^T\ C^{-1}\,(y - f(a)) \\ &= -\frac{1}{2}\ \left(C^{-\frac{1}{2}}\,(y - f(a))\right)^T\ \left(C^{-\frac{1}{2}}\,(y - f(a))\right) = -\frac{1}{2}\ \left|\tilde{y} - \tilde{f}(a)\right|^2 \quad .\end{aligned}$$

Es wurde ausgenutzt, dass die Kovarianz-Matrix symmetrisch ist. Die Transformation mit $C^{-\frac{1}{2}}$ bewirkt, dass in dieser Darstellung die Fehler der Daten unkorreliert sind und alle die Varianz Eins besitzen.

TRANSFORMIERTE MULTIVARIATE NORMALVERTEILUNG

$$\begin{aligned} p(d|C, a, \mathcal{B}) &= (2\pi)^{-\frac{N}{2}}\ |C|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\left|\tilde{y} - \tilde{f}(a)\right|^2} \\ &\text{mit} \\ \tilde{y} &= C^{-\frac{1}{2}}\ y \\ \tilde{f}(a) &= C^{-\frac{1}{2}}\ f(a) \quad . \end{aligned} \tag{20.11}$$

Das bedeutet, dass auch in diesem allgemeinen Fall wieder ein least-squares-fit durchzuführen ist

LEAST-SQUARES-FIT MIT KORRELIERTEN FEHLERN

$$\min_a \sum_i \left(\tilde{y}_i - \tilde{f}_i(a)\right)^2 \quad . \tag{20.12}$$

20.3 Cramer-Rao Untergrenze des Schätzfehlers

Der unverzerrte Schätzwert ist nicht eindeutig. Wenn z.B. das arithmetische Mittel einen unverzerrten Schätzwert darstellt, dann ist auch jedes Element x_i der Stichprobe ein unverzerrter Schätzwert. Nur wird in diesem Fall der Schätzwert weiter um den wahren Wert des Parameters streuen als es das arithmetische Mittel tut. Der beste Schätzwert ist sicherlich derjenige, der am wenigsten vom wahren Wert abweicht (streut). Es stellt sich die Frage, ob es möglich ist diese Streuung beliebig klein zu machen? Die Antwort liefert der Cramer-Rao-Grenzwert.
Der CR-Grenzwert stellt ein fundamentales Ergebnis der Statistik dar.
Wir gehen von einem unverzerrten Schätzwert $\hat{a}(\underline{x})$ für den Parameter a aus. Dann gilt

$$\int p(\underline{x}|a)\ (\hat{a}(\underline{x}) - a)\ d^L x = 0 \quad .$$

Wir leiten diese Gleichung nach a ab und erhalten

$$\int \left(\frac{\partial}{\partial a}\, p(\underline{x}|a)\right)\ (\hat{a}(\underline{x}) - a)\ d^L x - \underbrace{\int p(\underline{x}|a)\ d^L x}_{=1} = 0 \quad .$$

Mit $\frac{\partial}{\partial a}\ p = p\frac{\partial}{\partial a}\ \ln(p)$ und der Schwarzschen Ungleichung folgt daraus

$$\begin{aligned} 1 &= \int p(\underline{x}|a)\ \left(\frac{\partial}{\partial a}\ \ln(p(\underline{x}|a))\right)\ (\hat{a}(\underline{x}) - a)\ d^L x \\ &= \left\langle \left(\frac{\partial}{\partial a}\ \ln(p(\underline{x}|a))\right)\ (\hat{a}(\underline{x}) - a) \right\rangle \\ &\leq \sqrt{\left\langle \left(\frac{\partial}{\partial a}\ \ln(p(\underline{x}|a))\right)^2 \right\rangle\ \left\langle (\hat{a}(\underline{x}) - a)^2 \right\rangle} \quad . \end{aligned} \tag{20.13}$$

Da wir von einem unverzerrten Schätzwert ausgegangen sind, ist $\langle \hat{a}(\underline{x}) \rangle = a$ und somit ist

$$\left\langle (\hat{a}(\underline{x}) - a)^2 \right\rangle = \mathrm{var}(\hat{a}(\underline{x})) \qquad .$$

Somit haben wir die gesuchte Ungleichung

CRAMER-RAO UNGLEICHUNG

$$\mathrm{var}(\hat{a}(\underline{x})) \geq \frac{1}{I} \tag{20.14}$$

$$I = \int p(\underline{x}|a) \left(\frac{\partial}{\partial a} \ln(p(\underline{x}|a)) \right)^2 d^L x$$

$$= \int \frac{\left(\frac{\partial}{\partial a} p(\underline{x}|a) \right)^2}{p(\underline{x}|a)} d^L x \qquad . \tag{20.15}$$

I ist die sogenannte FISHER-INFORMATION.

Beispiel: Stichprobe Normal-Verteilung

Die Log-Likelihood ist in diesem Fall

$$\ln(p(\underline{x}|a)) = C - \frac{1}{2\sigma^2} \sum_{i=1}^{L} (x_i - a)^2 \qquad .$$

Die benötigte Ableitung liefert

$$\frac{d}{da} \ln(p(\underline{x}|a)) = \frac{1}{\sigma^2} \sum_{i=1}^{L} (x_i - a) \qquad . \tag{20.16}$$

$$= \frac{L}{\sigma^2} (\overline{x} - a) \qquad . \tag{20.17}$$

Hierbei ist $\overline{x}$ der Stichproben-Mittelwert. Wir wissen bereits aus dem Kapitel über Stichproben-Verteilungen, dass der Stichproben-Mittelwert im Mittel den wahren Mittelwert a liefert (unbiased estimator). Somit ist die Fisher-Information

$$I = \frac{L^2}{\sigma^4} \mathrm{var}(\overline{x})$$
$$= \frac{L^2}{\sigma^4} \frac{\sigma^2}{L}$$
$$= \frac{L}{\sigma^2} \qquad .$$

Die Varianz eines beliebigen Schätzwertes erfüllt somit die Ungleichung

$$\mathrm{var}(\hat{a}(\underline{x})) \geq \frac{\sigma^2}{L} = \mathrm{var}(\overline{x}) \qquad .$$

Der Stichproben-Mittelwert $\hat{a}(\underline{x}) = \overline{x}$ stellt in diesem Fall einen effizienten Schätzwert dar, da

$$\mathrm{var}(\hat{a}(\underline{x})) = \mathrm{var}(\overline{x}) = \frac{\sigma^2}{L} \qquad .$$

Beispiel: Stichprobe einer Cauchy-Verteilung

Wir betrachten hier die Cauchy-Verteilung für $b = 1$

$$p(x|a) = \frac{1}{\pi} \, \frac{1}{(x-a)^2 + 1} \qquad .$$

Die Log-Likelihood lautet dann

$$\ln(p(\underline{x}|a)) = C - \sum_{i=1}^{L} \ln((x_i - a)^2 + 1) \qquad .$$

Die benötigte Ableitung liefert

$$\frac{d}{da} \ln(p(\underline{x}|a)) = 2 \sum_{i=1}^{L} \frac{x_i - a}{(x_i - a)^2 + 1} \qquad . \tag{20.18}$$

Die Fisher-Information ist dann

$$\begin{aligned}
I &= 4 \left\langle \left(\sum_{i=1}^{L} \frac{x_i - a}{(x_i - a)^2 + 1} \right)^2 \right\rangle \\
&= 4 \sum_{i,j} \left\langle \frac{x_i - a}{(x_i - a)^2 + 1} \, \frac{x_j - a}{(x_j - a)^2 + 1} \right\rangle \\
&= 4 \sum_{i \neq j} \left\langle \frac{x_i - a}{(x_i - a)^2 + 1} \right\rangle \left\langle \frac{x_j - a}{(x_j - a)^2 + 1} \right\rangle + 4 \sum_{i} \left\langle \frac{(x_i - a)^2}{((x_i - a)^2 + 1)^2} \right\rangle \\
&= 4L(L-1) \left(\left\langle \frac{x - a}{(x - a)^2 + 1} \right\rangle \right)^2 + 4L \left\langle \frac{(x - a)^2}{((x - a)^2 + 1)^2} \right\rangle \qquad .
\end{aligned}$$

Die verbleibenden Erwartungswerte sind

$$\left\langle \frac{x - a}{(x - a)^2 + 1} \right\rangle = \int \frac{x - a}{(x - a)^2 + 1} \, p(x|a) \, dx = \frac{1}{\pi} \int \frac{x - a}{((x - a)^2 + 1)^2} \, dx = 0$$

und

$$\left\langle \frac{(x-a)^2}{((x-a)^2+1)^2} \right\rangle = \frac{1}{\pi} \int \frac{(x-a)^2}{((x-a)^2+1)^3}\, dx = \frac{1}{\pi} \int \frac{z^2}{(z^2+1)^3}\, dz = \frac{1}{8} \quad .$$

Somit ist die Fisher-Information $I = \frac{L}{2}$ und die Ungleichung für die Varianz lautet

$$\mathrm{var}(\hat{a}(\underline{x})) \geq \frac{2}{L} \quad .$$

Wir hatten im Kapitel über Stichproben-Verteilungen den Median von Stichproben der Cauchy-Verteilung untersucht und gefunden, dass dieser Schätzwert unverzerrt ist und für $L \gg 1$ die Varianz

$$\mathrm{var}(\hat{a}(\underline{x})) = \frac{\pi^2}{4L} = \frac{2.47}{L}$$

besitzt. Diese Varianz ist etwas größer als die Cramer-Rao Untergrenze. Der Median ist somit kein effizienter Schätzwert. Es stellt sich die Frage, ob es immer einen Schätzwert gibt, der die Untergrenze erfüllt und wie dieser Schätzwert konstruiert werden kann.

20.3.1 Wann wird die CR-Grenze erreicht?

Die Schwarzsche Ungleichung liefert dann die Gleichheit, wenn die beiden beteiligten Vektoren $\vec{a}$ und $\vec{b}$ parallel sind

$$\vec{a} = \alpha\, \vec{b} \quad .$$

Das bedeutet für die CR-Untergrenze aus Gl. (20.13)

$$\frac{\partial}{\partial a}\, \ln(p(\underline{x}|a)) = \alpha\ (\hat{a}(\underline{x}) - a) \quad , \tag{20.19}$$

wobei die Proportionalitätskonstante von a, aber nicht von der Stichprobe $\underline{x}$, abhängen kann.
Wenn die Ableitung der Log-Likelihood in die Form der rechten Seite von Gl. (20.19) gebracht werden kann, dann existiert ein unverzerrter effizienter Schätzwert. Der Schätzwert kann unmittelbar hieraus abgelesen werden. Umgekehrt bedeutet das aber auch, wenn die Ableitung nicht in die Form der rechten Seite gebracht werden kann, existiert kein unverzerrter effizienter Schätzwert.
Wir kommen nun zu einem wichtigen Satz: Wenn die Gleichung Gl. (20.19) erfüllt ist, dann ist der ML-Schätzwert ein unverzerrter effizienter Schätzwert.
Wir gehen davon aus, dass Gl. (20.19) erfüllt ist. Das heißt die Ableitung der Log-Likelihood kann in die Form

$$\frac{\partial}{\partial a}\, \ln(p(\underline{x}|a)) = \alpha(a)\ (f(\underline{x}) - a)$$

gebracht werden. Hierbei darf $\alpha(a)$ nicht von $\underline{x}$ abhängen. $f(\underline{x})$ ist eine beliebige Funktion der Stichprobe, die jedoch nicht von vom Parameter a abhängt. Diese Funktion stellt dann den Schätzwert dar.
Die ML-Lösung erhalten wir definionsgemäß aus der Nullstelle der Ableitung der Log-Likelihood

$$\frac{\partial}{\partial a} \ln(p(\underline{x}|a))\Big|_{a^{\mathrm{ML}}} = 0 \qquad .$$

Nun ist aber gemäß Gl. (20.19) die benötigte Ableitung von der Form $\alpha(a)\ (f(\underline{x}) - a)$ und hieraus folgt

$$\frac{\partial}{\partial a} \ln(p(\underline{x}|a))\Big|_{a^{\mathrm{ML}}} = \alpha(a^{\mathrm{ML}})\ \left(f(\underline{x}) - a^{\mathrm{ML}}\right) \stackrel{!}{=} 0 \qquad .$$

Damit ist zu gegebener Stichprobe, die ML-Lösung

$$a^{\mathrm{ML}} = f(\underline{x}) \qquad .$$

Wie behauptet, ist die ML-Lösung gleich dem unverzerrten effizienten Schätzwert $f(\underline{x})$.

20.3.2 Beispiele

Normal-Verteilung

Aus Gl. (20.16) kennen wir bereits die Ableitung der Log-Likelihood einer Stichprobe normal-verteilter Zufalls-Variablen. Sie ist unmittelbar in der geforderten Gestalt, und der effiziente Schätzwert ist das arithmetische Mittel und, er stimmt somit in der Tat mit der ML-Lösung überein.

Cauchy-Verteilung

Die Ableitung der Log-Likelihood einer Stichprobe Cauchy-verteilter Zufalls-Variablen wurde bereits in Gl. (20.18) ermittelt und liefert

$$\frac{d}{da} \ln(p(\underline{x}|a)) = 2 \sum_{i=1}^{L} \frac{x_i - a}{(x_i - a)^2 + 1} \qquad .$$

Es ist offensichtlich, dass sich dieser Ausdruck nicht in die gewünschte Form bringen lässt. Es existiert in diesem Fall also kein (effizienter) Schätzwert, der die CR-Untergrenze erreicht.

Bernoulli-Problem

Wir betrachten den Schätzwert $\frac{n_g}{N}$ des Bernoulli-Problems (siehe Abschnitt 20.2.4) für die intrinsische Wahrscheinlichkeit q. Die Ableitung der Log-Likelihood ist nach Gl. (20.3)

$$\begin{aligned}\frac{d}{dq}\log\left(P(n_g|q,N,\mathcal{B})\right) &= \frac{n_g}{q} - \frac{N-n_g}{1-q} \\ &= \frac{N}{q(1-q)}\left(\frac{n_g}{N} - q\right) \quad .\end{aligned}$$

Damit ist die Bedingung (Gl. (20.19)) für einen effizienten Schätzwert erfüllt.

20.4 Parameter-Schätzen im Rahmen der Wahrscheinlichkeitstheorie

Die Wahrscheinlichkeitstheorie erlaubt es, aus Stichproben Parameter des zugrundeliegenden (physikalischen) Problems zu eruieren. Die Vorgehensweise ist straightforward. Gegeben ist die tatsächlich vorliegende Stichprobe[1] $\underline{y}$. Daraus ermittelt man die Wahrscheinlichkeitsdichte für die gesuchten Parameter

$$p(a|\underline{y}, \mathcal{B}) \qquad .$$

Das Bayessche Theorem liefert hierfür

$$p(a|\underline{y}, \mathcal{B}) = \frac{p(\underline{y}|a, \mathcal{B})\ p(a|\mathcal{B})}{p(\underline{y}|\mathcal{B})} \propto p(\underline{y}|a, \mathcal{B})\ p(a|\mathcal{B}) \qquad .$$

Es hängt nun von den Anforderungen an das Ergebnis ab, wie man aus dieser Wahrscheinlichkeitsdichte einen einzigen Wert für die Parameter a auswählt. Eine Möglichkeit ist die sogenannte MODE, das ist das Maximum der a-posteriori Verteilung $p(a|\underline{y}, \mathcal{B})$. Wenn der Prior-Anteil vernachlässigbar ist, weil z.B. keine detailliertes Vorwissen vorliegt, oder die Zahl der Daten sehr groß ist, ist diese Lösung identisch mit der ML-Lösung. Neben der Mode gibt es weitere Möglichkeiten, die Wahrscheinlichkeitsdichte zu charakterisieren.
Es muss zunächst eine Kostenfunktion definiert werden. Diese Funktion hängt vom jeweiligen Problem ab.

$$K = \int d^{N_p}a\ K(a, \hat{a})\ p(a|\underline{y}, \mathcal{B}) \qquad .$$

In der Regel verwendet man Kostenfunktionen $K(a-\hat{a})$, die nur von der Abweichung $a - \hat{a}$ abhängen. Gängige Kostenfunktionen

- Quadratische Kosten $K(a - \hat{a}) = \sum_i (a_i - \hat{a}_i)^2$
  Die quadratische Kostenfunktion „bestraft" große Abweichungen stärker als kleine bis mittlere Abweichungen. Der daraus resultierende Schätzwert wird somit zu große Abweichungen zu vermeiden.

- Absolute Kosten $K(a - \hat{a}) = \sum_i |a_i - \hat{a}_i|$
  Es kann Situationen geben, in denen große Abweichungen nicht so viel negativer zu bewerten sind als kleine oder mittlere. In diesem Fall wird man eine absolute Kostenfunktion verwenden.

- Die stufenförmige Kosten-Funktion kommt zum Einsatz, wenn Abweichungen innerhalb einer kleinen Toleranz akzeptabel sind und ansonsten das Ergebnis

[1] Man geht hierbei nicht von fiktiven Stichproben-Replika aus.

unbrauchbar wird. Z.B. bei der Herstellung von Geräten. Innerhalb der Toleranz können die Geräte verkauft werden, ansonsten müssen sie neu produziert werden, und es fallen hierbei feste Kosten an.

$$K(a-\hat{a}) = -\prod_i \Theta\big(\,|a_i - \hat{a}_i| \leq \varepsilon\big)$$

Der beste Schätzwert ist nun der, der die zu erwartenden Kosten minimiert. Wir verlangen deshalb, dass die Ableitungen nach $\hat{a}_i$ verschwinden.

A) Quadratische Kosten

Das führt im Fall der quadratischen Kosten zur Bedingung

$$\begin{aligned}
\frac{\partial}{\partial \hat{a}_i} K &= 2 \int (a_i - \hat{a}_i)\; p(a|\underline{y}, \mathcal{B})\; d^{N_p}a \\
&= 2 \left(\int a_i\; p(a|\underline{y}, \mathcal{B})\; d^{N_p}a - \hat{a}_i \right) \stackrel{!}{=} 0 \\
&\Rightarrow \\
\hat{a}_i &= \int a_i\; p(a|\underline{y}, \mathcal{B})\; d^{N_p}a = \langle a_i \rangle_{\underline{y}} \\
\hat{a}_i &= \hat{a}_{\mathrm{PM},i}
\end{aligned}$$

Bei quadratischen Kosten ist das arithmetische Mittel der Posterior-Verteilung (Posterior Mittelwert (PM)) der beste Schätzwert.

B) Absolute Kosten

$$
\begin{aligned}
\frac{\partial}{\partial \hat{a}_i} K &= \int \sum_j \frac{\partial}{\partial \hat{a}_i} \; |\hat{a}_j - a_j| \; p(a|\underline{y}, \mathcal{B}) \; d^{N_p} a \\
&= \int \operatorname{sgn}(\hat{a}_i - a_i) \; p(a|\underline{y}, \mathcal{B}) \; d^{N_p} a \\
&= \int_{-\infty}^{\hat{a}_i} da_i \underbrace{\int \prod_{j \neq i} da_j \; p(a|\underline{y}, \mathcal{B})}_{=: p(a_i|\underline{y}, \mathcal{B})} \\
&\quad - \int_{\hat{a}_i}^{\infty} da_i \underbrace{\int \prod_{j \neq i} da_j \; p(a|\underline{y}, \mathcal{B})}_{=: p(a_i|\underline{y}, \mathcal{B})} \stackrel{!}{=} 0 \\
&\Rightarrow \\
\int_{-\infty}^{\hat{a}_i} p(a_i|\underline{y}, \mathcal{B}) \; da_i &= \int_{\hat{a}_i}^{\infty} p(a_i|\underline{y}, \mathcal{B}) \; da_i \\
\hat{a}_i &= \hat{a}_{\mathrm{MED},i}
\end{aligned}
$$

Bei absoluten Kosten ist der Median der beste Schätzwert.

C) Kastenförmige Kostenfunktion

In der folgenden Ableitung wird angenommen, dass die Intervalllänge ε klein ist.

$$
\begin{aligned}
\frac{\partial}{\partial \hat{a}_i} K &= -\frac{\partial}{\partial \hat{a}_i} \int \left(\prod_i \theta(|\hat{a} - a| \geq \varepsilon) \right) p(a|\underline{y}, \mathcal{B}) \; d^{N_p} a \\
&= -\frac{\partial}{\partial \hat{a}_i} \int_{\hat{a}_i - \varepsilon}^{\hat{a}_i + \varepsilon} da_i \underbrace{\left(\prod_{j \neq i} \int_{\hat{a}_j - \varepsilon}^{\hat{a}_j + \varepsilon} da_j \right) p(a|\underline{y}, \mathcal{B})}_{= (2\varepsilon)^{N_p - 1} p(a_i, \hat{a}_{j \neq i}|\underline{y}, \mathcal{B})} \\
&= -(2\varepsilon)^{N_p - 1} \frac{\partial}{\partial \hat{a}_i} \; \big(F(\hat{a}_i + \varepsilon|\underline{y}, \mathcal{B}) - F(\hat{a}_i - \varepsilon|\underline{y}, \mathcal{B}) \big) \\
&= -(2\varepsilon)^{N_p - 1} \big(p(\hat{a}_i + \varepsilon, \hat{a}_{j \neq i}|\underline{y}, \mathcal{B}) - p(\hat{a}_i - \varepsilon, \hat{a}_{j \neq i}|\underline{y}, \mathcal{B}) \big) \\
&= -(2\varepsilon)^{N_p} \frac{\partial}{\partial a_i} p(a|\underline{y}, \mathcal{B}) \Big|_{a = \hat{a}} \stackrel{!}{=} 0 \\
&\Rightarrow \\
\hat{a} &= \hat{a}_{\mathrm{MAP}}
\end{aligned}
$$

In diesem Fall erhalten wir die maximale a-posteriori Lösung (MAP).
Wir wollen diesen Fall noch etwas genauer untersuchen. Die Posterior-Wahrscheinlichkeit ist das Produkt aus Likelihood-Funktion und Prior-Wahrscheinlichkeit. Im Fall eines flachen Priors ist die MAP-Lösung somit gleichbedeutend mit der ML-Lösung. Das gibt der ML-Lösungen somit eine alternative Bedeutung.

20.4.1 Nächste-Nachbar-Abstände von d-dimensionalen Poisson-Punkten

Wir fragen nach der Verteilung der Abstände zu den nächsten Nachbarn von Poisson-Punkten (PP) der Dichte ρ in beliebiger Dimension d. Wir gehen zunächst von einer festen Anzahl N von PPen aus, die wir anschließend gegen Unendlich gehen lassen. Wir greifen nun einen der PP als Bezugspunkt heraus und fragen nach der Wahrscheinlichkeit $p(r)dr$, den ersten weiteren Punkt im Abstand r anzutreffen. Das heißt, von den verbleibenden $N-1$ Punkten sind $N-2$ in einem Abstand größer-gleich r und einer im Intervall $r, r+dr$. Es ist einfacher, von der komplementären Verteilungsfunktion

$$\tilde{F}(r) = \int_r^\infty f(r)\, dr$$

auszugehen, die gleichbedeutend ist mit der Wahrscheinlichkeit, den nächsten Nachbar im Abstand größer-gleich r anzutreffen. Das heißt, alle $N-1$ Punkte haben einen Abstand größer-gleich r vom Bezugspunkt. Da die Punkte alle unkorreliert sind, ist diese Wahrscheinlichkeit

$$\tilde{F}(r) = (P(\text{Abstand eines PPs ist größer-gleich } r)^{N-1} \qquad . \tag{20.20}$$

Die Wahrscheinlichkeit, dass ein PP im Abstand größer-gleich r anzutreffen ist, ist nach der klassischen Definition der Wahrscheinlichkeit gleich

$$P(\text{Abstand eines PPs ist größer-gleich } r) = \frac{V^>(r)}{V} \qquad ,$$

wobei $V^>(r)$ das „Volumen" des Bereiches ist, in dem die Punkte einen Abstand größer-gleich r vom Aufpunkt haben und V das Gesamtvolumen darstellt. Dafür kann man auch

$$P(\text{Abstand eines PPs ist größer-gleich } r) = \frac{V - V^<(r)}{V} = 1 - \frac{V^<(r)}{V}$$

schreiben. Hierbei ist $V^<(r)$ entsprechend das Volumen des Bereichs der Punkte, die einen kleineren Abstand als r aufweisen. Dieses Volumen ist aber nichts anderes als das Volumen der „Kugel" in d Dimensionen mit Radius r, also

$$V^<(r) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}\, r^d = c_d\, r^d \qquad . \tag{20.21}$$

Gemäß Gl. (20.20) ist die gesuchte komplementäre Verteilungsfunktion, wenn wir zusätzlich $\frac{N}{V} = \rho$ ausnutzen,

$$\begin{aligned}\tilde{F}(r) &= \left(1 - \frac{\rho}{N}\, V^<(r)\right)^{N-1} \\ &= e^{(N-1)\,\ln\left(1-\frac{\rho}{N}\, V^<(r)\right)} \\ &= e^{(N-1)\,\left(-\frac{\rho}{N}\, V^<(r) + O(\frac{1}{N^2})\right)} \\ &= e^{-\rho\, V^<(r) + O(\frac{1}{N})} \quad .\end{aligned}$$

Im Limes $N \to \infty$ verschwinden die Terme der Ordnung $O(\frac{1}{N})$. Aus der komplementären Verteilungsfunktion erhalten wir die eigentlich interessierende Wahrscheinlichkeitsdichte über

$$p(r) = -\frac{d}{dr}\, \tilde{F}(r) = \rho\, \frac{d}{dr}\, V^<(r)\, e^{-\rho\, V^<(r)} \quad .$$

Mit Gl. (20.21) liefert das

$$p(r) = \rho\, d\, c_d\, r^{d-1}\, e^{-\rho\, c_d\, r^d} \tag{20.22}$$

Wir wollen nun aus einer Stichprobe $\underline{r} = \{r_1, \ldots, r_L\}$ von L nächsten-Nachbar-Abstände die Dichte ρ ermitteln. Die Likelihood-Funktion ist in diesem Fall

$$\begin{aligned}p(\underline{r}|\rho, \mathcal{B}) &= \underbrace{\left(\prod_{i=1}^{L} r_i^{d-1}\right)\, (d\, c_d)^L}_{c}\; \rho^L\, e^{-\rho\, c_d\, \sum_i\, r_i^d} \\ &= c\, e^{L\left(\ln(\rho) - \rho\, c_d\, \overline{r^d}\right)} \quad .\end{aligned}$$

Die ML-Lösung ist somit gegeben durch

$$\begin{aligned}&\frac{d}{d\rho}\, \left(\ln(\rho) - \rho\, c_d\, \overline{r^d}\right) = 0 \\ &\Rightarrow \\ &\rho = \frac{1}{c_d\, \overline{r^d}} = \frac{1}{\overline{V}} \quad .\end{aligned}$$

Der ML-Schätzwert ist durch das inverse mittlere Stichproben-Volumen

$$\overline{V} = \frac{1}{L}\, \sum_{i=1}^{L}\, V^<(r_i)$$

gegeben.

20.4.2 Beispiel: Bernoulli-Problem

Wir wollen nun das binäre Bernoulli-Problem aus Abschnitt 20.2.4 wahrscheinlichkeitstheoretisch behandeln. Die Wahrscheinlichkeitsdichte $p(q|N, n_g, \mathcal{B})$ für den Wert des Verhältnisses $q = n_g/N_p$ der Population folgt aus dem Bayesschen Theorem

$$p(q|N, n_g, \mathcal{B}) = \frac{1}{Z}\, q^{n_g}\, (1-q)^{N-n_g}\, p(q|\mathcal{B}) \qquad .$$

Wenn wir für $p(q|\mathcal{B})$ einen flachen Prior $p(q|\mathcal{B}) = \theta(0 \leq q \leq 1)$ verwenden, so erhalten wir mit der richtigen Normierung (β-Verteilung Gl. (9.8a))

$$p(q|N, n_g, \mathcal{B}) = \frac{(N+1)!}{n_g!\,(N-n_g)!}\, q^{n_g}\, (1-q)^{N-n_g} \qquad .$$

Die MAP-Lösung ist bei einem flachen Prior wieder gleich der ML-Lösung

$$q_{\mathrm{MAP}} = q_{ML} = \frac{n_g}{N} \qquad .$$

Der Posterior-Mittelwert (PM) und das Vertrauensintervall folgen aus dem Mittelwert und der Standardabweichung der β-Verteilung (Gl. (9.8c), Gl. (9.8d))

$$\begin{aligned} q_{\mathrm{PM}} &= \langle q \rangle = \frac{n_g + 1}{N+2} \\ \mathrm{VI} &= \sqrt{\mathrm{var}(q)} = \sqrt{\frac{\langle q \rangle\, (1 - \langle q \rangle)}{N+3}} \end{aligned} \qquad .$$

Der Unterschied zwischen MAP-Lösung und Posterior-Mittelwert ist von der Ordnung $O(1/N)$, wohingegen das Vertrauensintervall von der Ordnung $O(1/\sqrt{N})$, also wesentlich größer ist. Für $n_g = 6$ und $N = 20$ ist die Wahrscheinlichkeit in Abbildung 20.3 dargestellt.

20.4.3 Risiko

Die mittleren Kosten, als Mittelwert über alle möglichen Datensätze und Parameter

$$R = \int\int K(a - \hat{a}(\underline{y}))\, p(a, \underline{y}|\mathcal{B})\, d^N y\, d^{N_p} a$$

nennt man Risiko. Die Produkt-Regel liefert

$$R = \int p(y|\mathcal{B}) \left(\int K(a - \hat{a}(\underline{y}))\, p(a|\underline{y}, \mathcal{B})\, d^{N_p} a \right) d^N y \qquad .$$

Wir wollen dasjenige Funktional $\hat{a}(\underline{y})$ bestimmen, dass das Risiko minimiert. Wir benötigen hierzu die Nullstelle der Funktional-Ableitung

$$0 = \frac{\delta}{\delta \hat{a}(\underline{y})} R = \int p(y|\mathcal{B}) \left(\frac{\delta}{\delta \hat{a}(\underline{y})} \int K(a - \hat{a}(\underline{y}))\, p(a|\underline{y}, \mathcal{B})\, d^{N_p} a \right) d^N y \qquad .$$

Da $p(y|\mathcal{B}) \geq 0$ muss die innere Klammer verschwinden. Das heißt, die eben hergeleiteten Schätzwerte optimieren auch gleichzeitig das Risiko.

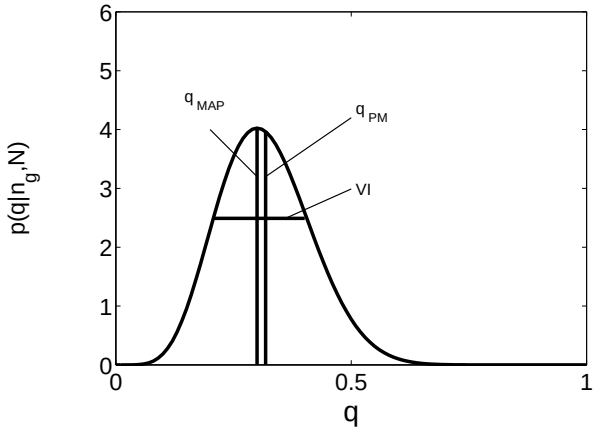

Abbildung 20.3: *Wahrscheinlichkeit* $p(q|N, n_g, \mathcal{B})$ *als Funktion von* q *für* $n_g = 6$ *und* $N = 20$.

20.4.4 Vertrauensintervall

Im Rahmen der Bayesschen Wahrscheinlichkeitstheorie ist das Vertrauensintervall eines Parameters a_i durch die Standardabweichung der Posterior-Wahrscheinlichkeit

$$\int (\Delta a_i)^2 \; p(a|d, \mathcal{B}) \; d^{N_p} a$$

gegeben.

20.5 Lineare Regression

Viele Anwendungen sind von Hause aus linear in den Parametern oder können im interessierenden Bereich linearisiert werden. Das Modell lautet in diesem Fall

$$y = \sum_i \; x_i \; a_i \qquad , \tag{20.23}$$

wobei die x_i Funktionen der experimentellen Steuergrößen und eventuell zusätzlicher Parameter sind. Zum Beispiel

$$y = a_0 + a_1 \; x + a_2 \; x^2 \qquad .$$

Das Experiment werde nun für L unterschiedliche Sätze von Steuergrößen $\{x_{\nu i}\}$ wiederholt, wobei der Index $\nu = 1, \ldots, L$ die Experimente durchnummeriert

$$y_\nu = \sum_{n=1}^{N} \; x_{\nu n} \; a_n \qquad .$$

Man kann diesen Ausdruck auch als die Entwicklung eines Vektors y nach Basisvektoren auffassen, wobei $x_{\nu n}$ die n-te Komponente des ν-ten Basisvektors ist. Die a_n sind die Entwicklungskoeffizienten. Das lässt sich kompakt in Vektor-Notation schreiben

$$y = X\ a \qquad .$$

Hierbei ist $y \in \mathbb{R}^L$ der Vektor der experimentellen Ergebnisse, $a \in \mathbb{R}^N$ der Vektor der Entwicklungsparameter und X die $(L \times N)$-Matrix mit den Elementen $x_{\nu n}$, die von den unterschiedlichen Steuergrößen und zusätzlichen vorgegebenen Parametern in bekannter Weise abhängen. Wir gehen davon aus, dass das Modell stimmt. Allerdings sind experimentelle Daten i.d.R. fehlerbehaftet, so dass die gemessenen Größen y von den Vorhersagen des Modells abweichen werden. Wir machen folgende Annahmen

- Lineares Modell mit additiven Fehlern

$$d = Xa + \eta \qquad .$$

- die Fehler η_ν der einzelnen Experimente sind unabhängig vom Messwert d_ν
- Die Zufalls-Variablen η_ν sind multivariat normalverteilt mit der Kovarianz-Matrix C. Die Mittelwert seien alle Null. Wenn wir wissen, dass ein Mittelwert nicht Null ist, heißt das, dass das Experiment systematische Fehler produziert und es wäre sinnvoll diese zunächst zu eliminieren, bzw. wenn der Bias bekannt ist, ihn von den Daten abzuziehen.

Diese Annahmen definieren das LINEARE REGRESSIONSMODELL[2] mit der Likelihood gemäß Gl. (20.11)

LINEARES REGRESSIONSMODELL

$$\begin{aligned} p(d|a,N,L,X,\sigma,\mathcal{B}) &= (2\pi)^{-\frac{L}{2}}\ |C|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\left|\tilde{d}-\tilde{X}a\right|^2} \\ &\text{mit} \\ \tilde{d} &= C^{-\frac{1}{2}}\ d \\ \tilde{X} &= C^{-\frac{1}{2}}\ X \qquad . \end{aligned} \tag{20.24}$$

Die ML-Lösung bzw. bei konstantem Prior die MAP-Lösung erhalten wir durch Minimieren der FEHLANPASSUNG

$$\chi^2 := \left|\tilde{d} - \widetilde{X}a\right|^2 = \left|\tilde{d}\right|^2 - 2a^T\widetilde{X}^T\tilde{d} + a^T\widetilde{X}^T\widetilde{X}a \qquad ,$$

[2] Häufig findet man die Definition für den Spezialfall i.u.nv. Fehler.

d.h. an der Nullstelle des Gradienten

$$0 = \vec{\nabla}\chi^2 = \left(-2\widetilde{X}^T\tilde{d} + 2\widetilde{X}^T\widetilde{X}a\right) \qquad .$$

Das lineare Regressionsmodell hat die Lösung

ML-LÖSUNG DES LINEAREN REGRESSIONSMODELLS
$$a^{\mathrm{ML}} = (\widetilde{X}^T\widetilde{X})^{-1}\ \widetilde{X}^T\tilde{d} \qquad . \tag{20.25}$$

Voraussetzung ist, dass von der Matrix

$$H = \widetilde{X}^T\widetilde{X}$$

das Inverse existiert. Aus den ML-Parametern folgen die Modell-Werte

$$y^{\mathrm{ML}} = \widetilde{X}a^{\mathrm{ML}} \qquad .$$

Man kann die multivariate Normal-Verteilung Gl. (20.24) ausdrücken in

$$\Delta a = a - a^{\mathrm{ML}} \qquad ,$$

in der die Log-Likelihood quadratisch ist, in der Form

$$p(d|a, N, L, X, \sigma, \mathcal{B}) = \frac{1}{Z}\, e^{-\frac{1}{2}\,\Delta a^T H \Delta a} \qquad , \tag{20.26}$$

mit der

HESSE-MATRIX
$$H = \tfrac{1}{2}\,\nabla\nabla^T\,\chi^2 = \widetilde{X}^T\widetilde{X} \qquad . \tag{20.27}$$

Im Rahmen der Bayessche Wahrscheinlichkeitstheorie gilt

$$p(a|d, N, L, X, \sigma, \mathcal{B}) = \frac{1}{Z'}\, p(d|a, N, L, X, \sigma, \mathcal{B})\ p(a|d, N, L, X, \sigma, \mathcal{B}) \qquad .$$

Immer dann, wenn das Experiment wesentliche, neue Erkenntnisse über die Parameter a liefert – davon sollte man bei einem guten Experiment eigentlich ausgehen

können – dominiert die Likelihood-Funktion die funktionelle Form der Posterior-Verteilung und man kann einen flachen Prior annehmen. Unter Berücksichtigung der Normierung der multivariaten Normalverteilung (Gl. (9.28a)) erhalten wir dann

$$p(a|d, N, L, X, \sigma, \mathcal{B}) = (2\pi\sigma^2)^{-\frac{L}{2}} \, |H|^{\frac{1}{2}} \;\; e^{-\frac{1}{2}\,\Delta a^T H \Delta a}$$

Aus Gl. (9.28c) wissen wir, dass die Kovarianz der Parameter

$$\langle \Delta a_n \Delta a_m \rangle = H^{-1}_{m,n} \tag{20.28}$$

durch die Matrix-Elemente der inversen Hesse-Matrix gegeben sind. Diese Matrix wird in der Regel nicht diagonal sein. Das bedeutet, dass die Fehler korreliert sind. Das ist einleuchtend, wenn wir z.B. an einen Geraden-Fit denken: Wenn hierbei, z.B. die Steigung verändert wird muss man auch gleichzeitig den Achsen-Abschnitt verändern, um zur veränderten Steigung den optimalen Fit an die Daten zu erhalten. Wir wollen zwei wichtige Spezialfälle betrachten

20.5.1 Schätzen einer Konstanten

Das einfachste lineare Modell ist die Konstante

$$y = a \qquad .$$

Weiter wollen wir unkorrelierte Fehler mit individuellen Varianzen annehmen

$$C_{ij} = \delta_{ij} \, \sigma_i^2 \qquad .$$

Wir benötigen

$$C_{ij}^{-\frac{1}{2}} = \delta_{ij} \, \sigma_i^{-1} \qquad .$$

Die Matrix X hat die einfache Gestalt

$$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

und weiter gilt

$$\widetilde{X} = C^{-\frac{1}{2}} X = \begin{pmatrix} 1/\sigma_1 \\ \vdots \\ 1/\sigma_i \\ \vdots \\ 1/\sigma_L \end{pmatrix} \qquad .$$

Die transformierten Daten sind

$$\tilde{d} = C^{-\frac{1}{2}} d = \begin{pmatrix} d_1/\sigma_1 \\ \vdots \\ d_i/\sigma_i \\ \vdots \\ d_L/\sigma_L \end{pmatrix} \quad .$$

Die Hessematrix ist dann

$$H = \widetilde{X}^T \widetilde{X} = \sum_{i=1}^{L} \frac{1}{\sigma_i^2} \stackrel{\text{def}}{=} \Sigma^{-2} \quad .$$

Schließlich ist die ML-Lösung

$$a^{\text{ML}} = H^{-1} \widetilde{X}^T \tilde{d} = \Sigma^2 \sum_{i=1}^{L} \frac{d_i}{\sigma_i^2} \quad .$$

Die Lösung des Konstanten-Problems lautet also

ML-/MAP-LÖSUNG DER KONSTANTEN

$$\begin{aligned} a^{\text{ML}} &= \frac{\sum_{i=1}^{L} \frac{d_i}{\sigma_i^2}}{\sum_{i=1}^{L} \frac{1}{\sigma_i^2}} =: \langle d \rangle_\sigma \\ \text{var}(a) &= 1 / \sum_{i=1}^{L} \frac{1}{\sigma_i^2} \quad . \end{aligned} \tag{20.29}$$

Die ML-/MAP-Lösung ist also das mit den inversen Varianzen gewichtete Mittel. Im Fall $\sigma_i = \sigma$ geht das Ergebnis in das alt-vertraute Ergebnis $a^{\text{ML}} = \overline{d}$ und $\text{var}(a) = \sigma^2/L$ über.

20.5.2 Schätzen der Parameter einer Geraden

Ein weiteres weit verbreitetes Problem ist die Bestimmung der Parameter einer Geraden

$$y = a_1 + x\, a_2 \quad .$$

Die zu untersuchenden Parameter sind der Achsen-Abschnitt a_1 und die Steigung a_2. Wir setzen dieselbe Kovarianz-Matrix vom letzten Beispiel voraus. Die Matrix X ist

in diesem Fall

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_L \end{pmatrix} \qquad .$$

Daraus ergibt sich

$$\widetilde{X} = \begin{pmatrix} 1/\sigma_1 & x_1/\sigma_1 \\ \vdots & \vdots \\ 1/\sigma_i & x_i/\sigma_i \\ \vdots & \vdots \\ 1/\sigma_L & x_L/\sigma_L \end{pmatrix} \qquad .$$

Die Hesse-Matrix lautet nun

$$\widetilde{X}^T \widetilde{X} = \begin{pmatrix} 1/\sigma_1 & \dots & 1/\sigma_i & \dots 1/\sigma_l \\ x_1/\sigma_1 & \dots & x_i/\sigma_i & \dots x_L/\sigma_L \end{pmatrix} \begin{pmatrix} 1/\sigma_1 & x_1/\sigma_1 \\ \vdots & \vdots \\ 1/\sigma_i & x_i/\sigma_i \\ \vdots & \vdots \\ 1/\sigma_L & x_L/\sigma_L \end{pmatrix} = \begin{pmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{pmatrix}$$

$$= \Sigma^{-2} \begin{pmatrix} 1 & \langle x \rangle_\sigma \\ \langle x \rangle_\sigma & \langle x^2 \rangle_\sigma \end{pmatrix} \qquad .$$

Die Determinante hiervon ist

$$|H| = \Sigma^{-4} \left(\langle x^2 \rangle_\sigma - \langle x \rangle_\sigma^2 \right) = \Sigma^{-4} \left\langle (\Delta x)^2 \right\rangle_\sigma \stackrel{def}{=} \Sigma^{-4} \, C_{xx} \qquad .$$

Für das Inverse einer (2×2) Matrix gilt

$$A^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = \frac{1}{|A|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} \qquad .$$

Für die ML-Lösung benötigen wir

$$\widetilde{X}^T \tilde{d} = \begin{pmatrix} 1/\sigma_1 & \dots & 1/\sigma_i & \dots 1/\sigma_L \\ x_1/\sigma_1 & \dots & x_i/\sigma_i & \dots x_L/\sigma_L \end{pmatrix} \begin{pmatrix} d_1/\sigma_1 \\ \vdots \\ d_i/\sigma_i \\ \vdots \\ d_L/\sigma_L \end{pmatrix} = \Sigma^{-2} \begin{pmatrix} \langle d \rangle_\sigma \\ \langle xd \rangle_\sigma \end{pmatrix} \qquad .$$

Damit ist die ML-Lösung

$$\begin{aligned} a^{\mathrm{ML}} &= \frac{\Sigma^{-2}\Sigma^{-2}}{\Sigma^{-4}\, C_{xx}} \begin{pmatrix} \langle x^2\rangle_\sigma & -\langle x\rangle_\sigma \\ -\langle x\rangle_\sigma & 1 \end{pmatrix} \begin{pmatrix} \langle d\rangle_\sigma \\ \langle xd\rangle_\sigma \end{pmatrix} \qquad . \\ &= \frac{1}{C_{xx}} \begin{pmatrix} \langle x^2\rangle_\sigma \langle d\rangle_\sigma - \langle x\rangle_\sigma \langle xd\rangle_\sigma \\ \langle xd\rangle_\sigma - \langle x\rangle_\sigma \langle d\rangle_\sigma \end{pmatrix} \\ &= \frac{1}{C_{xx}} \begin{pmatrix} (C_{xx} + \langle x\rangle_\sigma^2)\langle d\rangle_\sigma - \langle x\rangle_\sigma (C_{xd+}\langle x\rangle_\sigma \langle d\rangle_\sigma) \\ C_{xd} \end{pmatrix} \\ &= \begin{pmatrix} \langle d\rangle_\sigma - \langle x\rangle_\sigma \frac{C_{xd}}{C_{xx}} \\ \frac{C_{xd}}{C_{xx}} \end{pmatrix} \qquad . \end{aligned} \tag{20.30}$$

Das heißt

PARAMETER DES GERADEN-FITS

$$\begin{aligned} a_1 &= \langle d\rangle_\sigma - \langle x\rangle_\sigma \, a_2 \\ a_2 &= \frac{C_{xd}}{C_{xx}} \\ \mathrm{var}(a_1) &= H_{11}^{-1} = \frac{\Sigma^2 \langle x^2\rangle_\sigma}{C_{xx}} \\ \mathrm{var}(a_2) &= H_{22}^{-1} = \frac{\Sigma^2}{C_{xx}} \\ \mathrm{cov}(a_1, a_2) &= H_{12}^{-1} = -\frac{\Sigma^2 \langle x\rangle_\sigma}{C_{xx}} \qquad . \end{aligned} \tag{20.31}$$

20.5.3 Vorhersagen bei einem linearen Modell

Man wird in vielen Fällen nicht allein an den Parametern interessiert sein, sondern Vorhersagen für neue, experimentell schwer zugängliche Steuergrößen x machen wollen. In dem einfachen Geraden-Problem wollen wir also y_x zu einem beliebigen Wert der unabhängigen Variablen ξ vorhersagen. Diese Überlegungen wollen wir nur für den Fall gleicher Varianzen $\sigma_i^2 = \sigma^2$ vorrechnen. Gesucht ist demnach

$$p(y_\xi | \xi, \underline{x}, \underline{y}, N, \sigma, \mathcal{B}) \qquad .$$

Diese Größe erhalten wir aus der bereits bekannten Wahrscheinlichkeitsdichte der Parameter über die Marginalisierungsregel

$$\begin{aligned} p(y_\xi|\xi,\underline{x},\underline{y},N,L,\sigma,\mathcal{B}) &= \int da_1\ da_2\ p(y_\xi|a_1,a_2,\xi,\underline{x},\underline{y},N,L,\sigma,\mathcal{B})\ \times \\ &\qquad p(a_1,a_2|\underline{x},\underline{y},N,L,\sigma,\mathcal{B}) \\ &= \int da_1\ da_2\ \delta(y_\xi - a_1 - a_2\ \xi)\ p(a_1,a_2|\underline{x},\underline{y},N,L,\sigma,\mathcal{B}) \\ &= \int da_2\ p(a_1 = y_\xi - a_2\ \xi, a_2|\underline{x},\underline{y},N,L,\sigma,\mathcal{B}) \quad . \end{aligned}$$

Gemäß Gl. (20.24) führt das zu

$$p(y_\xi|\xi,\underline{x},\underline{y},N,L,\sigma,\mathcal{B}) \propto \int da_2\ e^{-\frac{1}{2\sigma^2}\sum_\nu (y_\nu - (y_\xi - a_2\ \xi) - a_2 x_\nu)^2} \quad . \tag{20.32}$$

Das Argument der Exponentialfunktion kann vereinfacht werden

$$\begin{aligned} &\sum_{\nu=1}^{L} (y_\nu - (y_\xi - a_2\ \xi) - a_2 x_\nu)^2 = \\ &\qquad = \sum_{\nu=1}^{L} ((y_\nu - y_\xi) - a_2(x_\nu - \xi))^2 \\ &\qquad = L\left(\overline{(y-y_\xi)^2} - 2\ a_2\ \overline{(y-y_\xi)(x-\xi)} + a_2^2 \overline{(x-\xi)^2}\right) \\ &\qquad = L\overline{(x-\xi)^2}\left(a_2^2 - 2\ a_2\ \underbrace{\frac{\overline{(y-y_\xi)(x-\xi)}}{\overline{(x-\xi)^2}}}_{a_2^0}\right) + L\ \overline{(y-y_\xi)^2} \\ &\qquad = L\overline{(x-\xi)^2}\left(a_2 - a_2^0\right)^2 + L\left(\overline{(y-y_\xi)^2} - \frac{\overline{(y-y_\xi)(x-\xi)}^2}{\overline{(x-\xi)^2}}\right) \quad . \end{aligned}$$

Wir setzen dieses Ergebnis in Gl. (20.32) ein und führen die Integration über a_2 aus. Uns interessieren nur die Terme, die von y_ξ abhängen. Alle anderen Faktoren erhalten wir einfacher nachträglich über die Normierung

$$p(y_\xi|\xi,\underline{x},\underline{y},N,L,\sigma,\mathcal{B}) \propto \exp\left(-\frac{L\ (\overline{(y-y_\xi)^2}\ \overline{(x-\xi)^2} - \overline{(y-y_\xi)(x-\xi)}^2)}{2\sigma^2\overline{(x-\xi)^2}}\right). \tag{20.33}$$

Wir drücken nun die Vektoren x und y über die Abweichungen Δx und Δy von den Mittelwerten aus

$$\begin{aligned} x &= \overline{x} + \Delta x \\ \xi &= \overline{x} + \Delta\xi \\ y_\xi &= \overline{y} + \Delta y_\xi \quad . \end{aligned}$$

Damit vereinfachen sich die Terme in Gl. (20.33)

$$\begin{aligned}
\overline{(x-\xi)^2} &= \mathrm{var}(x) + (\overline{x}-\xi)^2 &&= \mathrm{var}(x) + \Delta\xi^2 \\
\overline{(y-y_\xi)^2} &= \mathrm{var}(y) + (\overline{y}-y_\xi)^2 &&= \mathrm{var}(y) + {\Delta y_\xi}^2 \\
\overline{(x-\xi)(y-y_\xi)} &= &&= \mathrm{cov}(xy) + \Delta\xi\ \Delta y_\xi \qquad .
\end{aligned}$$

Damit vereinfachen wir den Nenner in Gl. (20.33)

$$\overline{(x-\xi)^2}\ \overline{(y-y_\xi)^2} - \overline{(y-y_\xi)(x-\xi)}^2 \ =$$

$$\begin{aligned}
&= (\mathrm{var}(x) + \Delta\xi^2)\ (\mathrm{var}(y) + {\Delta y_\xi}^2) - (\mathrm{cov}(xy) + \Delta\xi\ \Delta y_\xi)^2 \\
&= \mathrm{var}(x)\mathrm{var}(y) + \Delta\xi^2\ \mathrm{var}(y) + {\Delta y_\xi}^2\ \mathrm{var}(x) + {\Delta y_\xi}^2\ \Delta\xi^2 \\
&\qquad - \mathrm{cov}(xy)^2 - 2\ \mathrm{cov}(xy)\ \Delta\xi\ \Delta y_\xi - {\Delta y_\xi}^2\ \Delta\xi^2 \\
&= A + {\Delta y_\xi}^2\ \mathrm{var}(x) - 2\ \mathrm{cov}(xy)\ \Delta\xi\ \Delta y_\xi \qquad .
\end{aligned}$$

Hierbei enthält A alle Terme, die nicht von y_ξ abhängen. Diese Terme können bequemer nachträglich mit der Normierung berücksichtigt werden. Damit ist die gesuchte Wahrscheinlichkeitsdichte in Gl. (20.33)

$$\begin{aligned}
p(y_\xi|\xi,\underline{x},\underline{y},N,L,\sigma,\mathcal{B}) \propto\ & \exp\left(-\frac{L}{2\sigma^2}\ \frac{\mathrm{var}(x)\left({\Delta y_\xi}^2 - 2\ \overbrace{\frac{\mathrm{cov}(xy)}{\mathrm{var}(x)}}^{a_2^{\mathrm{ML}}}\ \Delta\xi\ \Delta y_\xi\right)}{\mathrm{var}(x)+\Delta\xi^2}\right) \\
\propto\ & \exp\left(-\frac{L}{2\sigma^2}\ \frac{({\Delta y_\xi}^2 - 2\ \Delta y_\xi\ a_2^{\mathrm{ML}}\ \Delta\xi)}{1+\frac{\Delta\xi^2}{\mathrm{var}(x)}}\right) \\
\propto\ & \exp\left(-\frac{L}{2\sigma^2}\ \frac{(\Delta y_\xi - a_2^{\mathrm{ML}}\ \Delta\xi)^2}{1+\frac{\Delta\xi^2}{\mathrm{var}(x)}}\right) \\
\propto\ & \exp\left(-\frac{L}{2\sigma^2}\ \frac{(y_\xi - \overline{y} + a_2^{\mathrm{ML}}\ \overline{x} - a_2^{\mathrm{ML}}\ \xi)^2}{1+\frac{\Delta\xi^2}{\mathrm{var}(x)}}\right) \\
\propto\ & \exp\left(-\frac{L}{2\sigma^2}\ \frac{(y_\xi - a_1^{\mathrm{ML}} - a_2^{\mathrm{ML}}\ \xi)^2}{1+\frac{\Delta\xi^2}{\mathrm{var}(x)}}\right) \qquad .
\end{aligned}$$

GERADEN-FIT VORHERSAGE

$$p(y_\xi|\xi,\underline{x},\underline{y},N,L,\sigma,\mathcal{B}) = \frac{1}{\sqrt{2\pi\sigma_\xi^2}}\, e^{-\frac{\left(y_\xi - a_1^{\mathrm{ML}} - a_2^{\mathrm{ML}}\,\xi\right)^2}{2\,\sigma_\xi^2}} \tag{20.34}$$

$$\text{mit} \qquad \sigma_\xi^2 = \frac{\sigma^2}{L}\left(1 + \frac{(\xi-\overline{x})^2}{\mathrm{var}(x)}\right) \qquad .$$

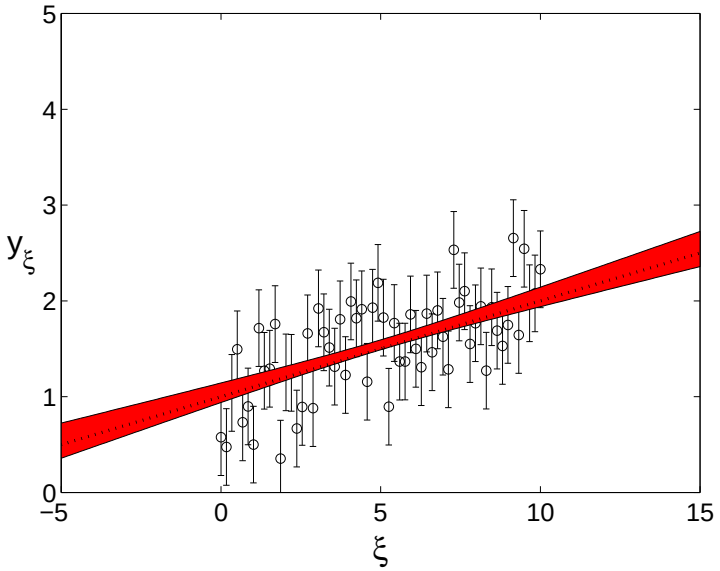

Abbildung 20.4: *Geradenfit mit Fehlerband. Die Gerade, mit der die Daten erzeugt wurden, ist gestrichelt eingetragen.*

Das ist ein einfaches und einleuchtendes Ergebnis. Der Wert y_ξ zur Steuergröße $x = \xi$ ist normal-verteilt. Der Mittelwert ist durch das lineare Modell mit Maximum-Likelihood-Werten für die Modell-Parameter a_1, a_2 gegeben. Die Varianz der Normal-Verteilung ist minimal im Zentrum der Daten bei $\xi = \langle x \rangle$. Dort ist

$$\sigma_\xi^2 = \frac{\sigma^2}{L} \qquad .$$

Es erscheint die bekannte $1/L$ Abhängigkeit. Weg vom Daten-Zentrum steigt die Varianz quadratisch an. Für ξ weit außerhalb des Datenbereichs, d.h. für

$$\frac{(\xi - \langle x \rangle)^2}{\mathrm{var}(x)} \gg 1 \qquad ,$$

wächst die Unsicherheit der Vorhersage (das Fehlerband)

$$\sigma_\xi = \frac{\sigma}{\sqrt{L}} \frac{|\xi - \langle x \rangle|}{\sqrt{\mathrm{var}(x)}}$$

linear mit dem Abstand vom Daten-Zentrum. Für äquidistante Daten gilt am Daten-Rand $\xi = x_1$ oder $\xi = x_N$ und für $L \gg 1$

$$\sigma_\xi = \frac{2\,\sigma}{\sqrt{L}} \quad .$$

Der Fehler ist hier also doppelt so groß wie im Zentrum der Daten, da dieser Bereich durch weniger Daten abgedeckt wird.

20.5.4 Zahl der Datenpunkte innerhalb des Fehlerband

Wir wollen einmal überlegen, wieviele Datenpunkte mit ihrem Fehlerbalken von einer Standardabweichung den Geradenfit berühren. In Gl. (4.16) hatten wir gezeigt, dass bei normal-verteilten experimentellen Fehlern die Wahrscheinlichkeit, einen Datenpunkt innerhalb einer Standard-Abweichung zu finden

$$q_{sd} = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-x^2/2}\, dx = 0.6827$$

ist. Die Wahrscheinlichkeit, für 2 Standardabweichungen war $q_{2sd} = 0.9545$. Bei L Datenpunkten, deren Fehler i.u.nv. sind, ist die Wahrscheinlichkeit, dass l Punkte innerhalb des Fehlerbandes liegen, die Binomial-Verteilung

$$P(l|L, q_{sd}, \mathcal{B}) = \binom{L}{l} q_{sd}^l \, (1 - q_{sd})^{L-l} \quad .$$

Somit haben wir ein wichtiges Ergebnis abgeleitet. Die mittlere Zahl der Daten im Fehlerband ist $L\ q_{sd}$ und die Varianz lautet $L\ q_{sd}(1 - q_{sd})$

ZAHL DER DATEN IM FEHLERBAND

$$\frac{L_{SD}}{L} = 0.6827 \pm \frac{0.4654}{\sqrt{L}} \quad .$$

Bei 10 Daten sollten (7 ± 2) innerhalb des Fehlerbandes liegen. Bei 100 Daten sollten es (68 ± 5) sein.
In Abbildung 20.4 liegen 65% der Daten innerhalb einer Standardabweichung von der ermittelten Geraden.

20.6 Parameter-Schätzen von nichtlinearen Modellen

Auch bei nichtlinearen Modellen ist der Ausgangspunkt die Posterior-Wahrscheinlichkeit

$$p(a|\underline{x}, \underline{y}, N, \mathcal{B}) \qquad .$$

In der Regel ist es bei diesen Problemen nicht möglich analytisch fortzufahren. Eine weitverbreitete Auswertungsmethode sind die numerischen Monte-Carlo-Verfahren, genauer gesagt die Markov-Chain-Monte-Carlo (MCMC) Verfahren, die im Rahmen der Computersimulationen besprochen werden.
Eine weitverbreitete Näherungsmethode stellt die sogenannte GAUSSSCHE NÄHERUNG oder STEEPEST-DESCENT-NÄHERUNG dar. Hierbei wird die Posterior-Wahrscheinlichkeit in eine exponentielle Form gebracht

$$p(a|\underline{x}, \underline{y}, N, \mathcal{B}) = e^{\Phi(a)} \qquad ,$$

was natürlich immer möglich ist, da die Wahrscheinlichkeitsdichte positiv ist. Die eigentlich Näherung besteht nun darin, das Argument der Exponential-Funktion in eine Taylorreihe um das Maximum a_{MAP} zu entwickeln bis zu Termen zweiter Ordnung

$$\begin{aligned} \Phi(a) &\simeq \Phi(a_{\mathrm{MAP}}) + \frac{1}{2}\, \Delta a^T\, H\, \Delta a \\ \Delta a &= a - a_{\mathrm{MAP}} \\ H_{ij} &= \left.\frac{\partial^2 \Phi(a)}{\partial a_i \partial a_j}\right|_{a=a_{\mathrm{MAP}}} \qquad . \end{aligned} \tag{20.35}$$

Die Matrix H ist die Hesse-Matrix. Die Prior-Wahrscheinlichkeit beinhaltet i.d.R. Schranken für die erlaubten Parameter. In der Steepest-Descent-Näherung setzt man voraus, dass die Posterior-Wahrscheinlichkeit an den Schranken bereits vernachlässigbar klein ist, so dass man auf die Schranken verzichten kann, und Integrale über den gesamten $\mathbb{R}^{n_\alpha}$ erstrecken kann. Somit lautet die Posterior-Wahrscheinlichkeit

$$p(a|\underline{x}, \underline{y}, N, \mathcal{B}) \simeq (2\pi)^{-\frac{N}{2}}\ |H|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\,\Delta a^T H \Delta a} \qquad . \tag{20.36}$$

Nach Konstruktion ist sie multivariat normal. Das hat zur Folge, dass $a_{\mathrm{PM}} = a_{\mathrm{MAP}}$, und die Kovarianz-Matrix ist gleich der inversen Hesse-Matrix

$$\mathrm{cov}(a_i a_j) = H^{-1}_{ij} \qquad .$$

20.7 Fehler in Abszisse und Ordinate

Wir wollen hier der Vollständigkeit halber angeben, wie der Geraden-Fit aussieht, wenn ein konstanter Fehler in Abszisse und Ordinate vorliegt σ_x und σ_y. Das ist der Fall, wenn man beim Ablesen der Steuergröße einen vergleichbaren Fehler macht, wie bei der Bestimmung der Messgröße. Es gibt aber auch den Fall, dass beide Koordinaten Messgrößen sind, z.B. $x-$, und $y-$Koordinate von 2-d Punkten. Gegeben ist eine Stichprobe die aus x-Werten $\underline{x} = \{x_1, \ldots, x_n\}$ und entsprechenden y-Werten besteht. Die Likelihood-Funktion lautet in diesem Fall

$$\begin{aligned} p(b|a, \underline{x}, \underline{y}, \underline{y}, \sigma_x, \sigma_y, \mathcal{B}) &= \tfrac{1}{\sqrt{2\pi\tilde{\sigma}^2}}\ e^{-\frac{(b-(\langle y\rangle - a\langle x\rangle))^2}{2\tilde{\sigma}^2}} \\ p(a|\underline{x}, \underline{y}, \sigma_x, \sigma_y, \mathcal{B}) &= \left(2\pi\tilde{\sigma}^2\right)^{-\frac{N}{2}}\ \ e^{-\frac{\frac{1}{N}\sum_i(\Delta y_i - a\Delta x_i)^2}{2\tilde{\sigma}^2}} \\ \tilde{\sigma}^2 &= \frac{\sigma_y^2 + a^2\sigma_x^2}{N} \qquad . \end{aligned}$$

20.7.1 Leuchtturm Problem

Wir betrachten folgendes Problem. Ein Leuchtturm befinde sich im Abstand a von der Küste und die Projektion auf die Küste liege an der Stelle b. Der Leuchtturm sendet zufällige Lichtblitze in alle Richtungen parallel zur Erdoberfläche. An der Küste befinden sich Detektoren, die die Lichtblitze nachweisen, aber nicht woher sie kommen. Wir führen einen Winkel Φ ein, der beim Lot auf die Küste den Wert Null annimmt. Die Lichtblitze sollen in Φ gleich-wahrscheinlich sein

$$p(\Phi|\mathcal{B}) = \frac{1}{\pi}\ \theta(-\frac{\pi}{2}, \frac{\pi}{2}) \qquad .$$

Der Bedingungskomplex beinhaltet die Einschränkung, dass nur Lichtblitze, die irgendwo die Küste treffen, berücksichtigt werden. Es werden nun L Lichtblitze an den Positionen $x_1, x_2, \ldots, x_L$ detektiert. Wie groß ist nun die Wahrscheinlichkeit $p(a, b|x_1, x_2, \ldots, x_L, \mathcal{B})$. Das Bayessche Theorem liefert

$$p(a, b|x_1, x_2, \ldots, x_L, \mathcal{B}) = \frac{1}{Z} p(x_1, x_2, \ldots, x_L|a, b, \mathcal{B})\ p(a, b|\mathcal{B}) \qquad .$$

Da die Lichtblitze zufällig sein sollen, gilt weiter

$$p(x_1, x_2, \ldots, x_L|a, b, \mathcal{B}) = \prod_{i=1}^{L}\ p(x_i|a, b, \mathcal{B}) \qquad .$$

Die Wahrscheinlichkeit $p(x|a, b, \mathcal{B})$, einen Lichtblitz im Intervall $(x, x + dx)$ anzutreffen, wenn wir die Leuchtturm-Position kennen, erhalten wir aus der Winkel-

Information

$$
\begin{aligned}
p(x|a,b,\mathcal{B}) &= \int_{-\pi/2}^{\pi/2} p(x|\Phi,a,b,\mathcal{B})\; p(\Phi|a,b,\mathcal{B}) \\
&= \int_{-\pi/2}^{\pi/2} \delta(x-(b+a\tan(\Phi))\;\frac{d\Phi}{\pi} \\
&= \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \delta(\Phi-\Phi^*)\frac{\cos^2(\Phi^*)}{a}\;d\Phi = \frac{\cos^2(\Phi^*)}{a\pi} \\
&= \frac{1}{a\;\pi\;(1+\tan^2(\Phi^*))} \\
\Phi^* &= \arctan(\frac{x-b}{a})
\end{aligned}
$$

Damit ergibt sich für die Wahrscheinlichkeit eines Blitzes am Ort x die Cauchy-Verteilung

$$p(x|a,b,\mathcal{B}) = \frac{a}{\pi\;(a^2+(x-b)^2)} \qquad .$$

Die gesuchte Wahrscheinlichkeit lautet schließlich

$$p(a,b|x_1,x_2,\ldots,x_L,\mathcal{B}) = \frac{1}{Z}\left(\prod_{i=1}^{L}\frac{a}{\pi\;(a^2+(x_i-b)^2)}\right)p(a,b|\mathcal{B}) \qquad .$$

Die Prior-Wahrscheinlichkeit $p(a,b|\mathcal{B})$ wollen wir hier nicht näher spezifizieren und setzen sie daher konstant an.

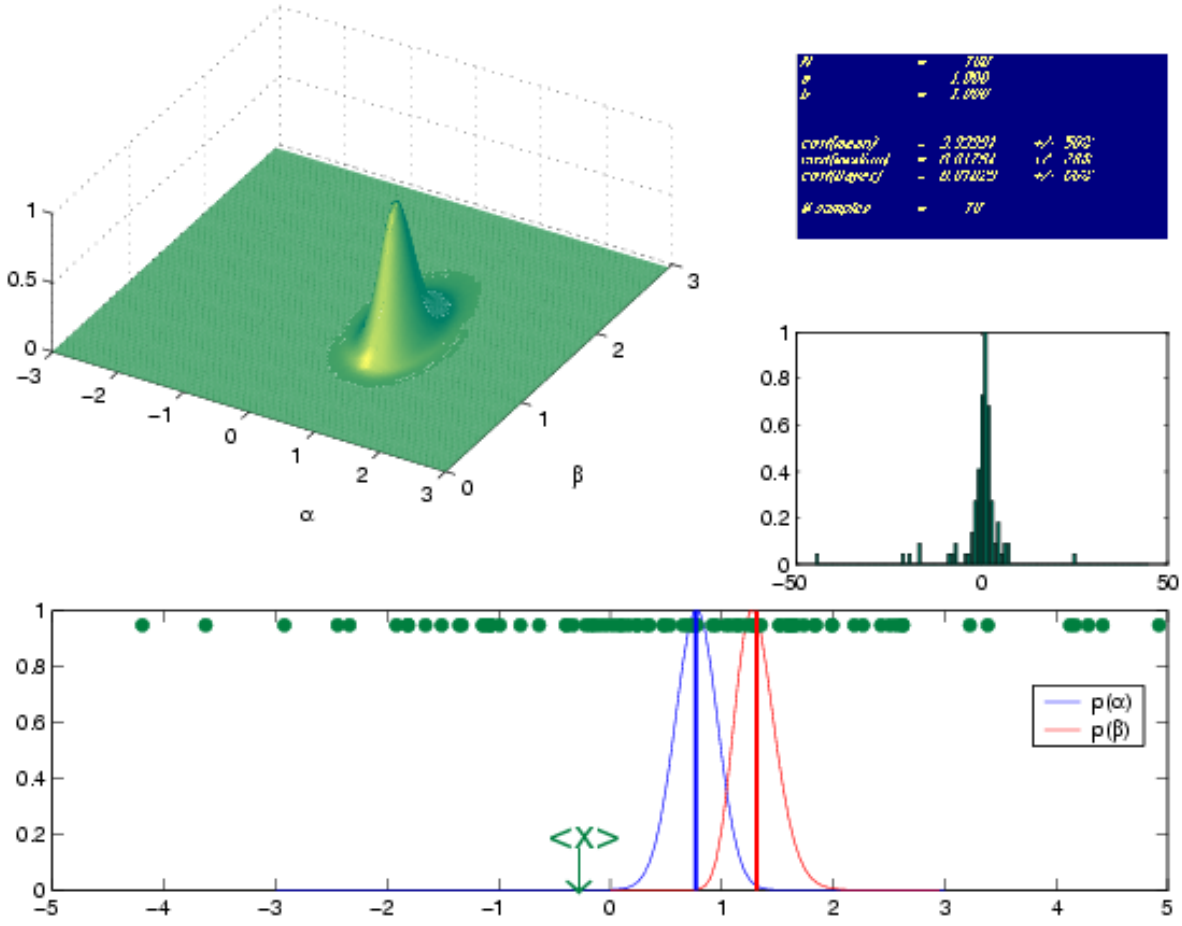

Abbildung 20.5: *Das Leuchtturm Problem. Beschreibung im Text.*

20.8 Ausreißer-tolerante Parameter-Schätzung

Es kommt häufig vor, dass einzelne Datenpunkte nicht dem angegebenen Fehlergesetz genügen. Das kann z.B. daran liegen, dass eine momentane externe Störung die Messung beeinflusst hat oder, dass die Daten aus unterschiedlichen Experimenten stammen, von denen eins oder mehrere inkorrekte Fehlerangaben machen oder gar systematische Fehler aufweisen.
Es ist gängige Praxis, offensichtliche Ausreißer wegzulassen. Das ist allerdings aus mehreren Gründen nicht befriedigend

- Bei höher-dimensionalen Problemen sind Ausreißer nicht mehr graphisch zu erkennen.
- Wenn viele Ausreißer vorliegen.
- Die Ausreißer entsprechen seltenen Ereignissen der Wahrscheinlichkeitsverteilung.

Das Eliminieren von vermeintlichen Ausreißern kann zu systematischen Fehlern beim Parameter-Schätzen führen. Schließlich kann es auch sein, dass die Ausreißer in Wirklichkeit physikalische Effekte wiederspiegeln.
Wir wollen hier besprechen, wie man trotz der möglichen Anwesenheit von Ausreißern zuverlässig und konsistent Modell-Parameter bestimmen kann.
Wir gehen von gemessenen Daten

$$d_i \quad , \qquad i = 1, \ldots, N$$

aus, die wie zuvor durch ein Modell

$$y_i = f(s_i, a)$$

beschrieben werden sollen, das eine Funktion der Steuergrößen s ist und von den Modell-Parametern a abhängt. Das Ziel der Experimente und der Datenanalyse ist die Bestimmung der Parameter. Die Daten sollen additive Gaußsche Fehler η_i aufweisen

$$p(\eta_i|\sigma_i, \mathcal{B}) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\, e^{-\frac{\eta_i^2}{2\sigma_i^2}} \qquad .$$

Wir gehen davon aus, dass die Fehler der einzelnen Messungen unkorreliert sind. Das ist sicherlich in den meisten Fällen richtig und trifft insbesondere auf Daten unterschiedlicher Experimente zu. Für diejenigen Datenpunkte, die keine Ausreißer sind, ist die Likelihood

$$p(d_i|\overline{A}, s_i, \sigma_i, a, \mathcal{B}) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\, e^{-\frac{(d_i - f(s_i|a))^2}{2\sigma_i^2}} \qquad .$$

Die Proposition $\overline{A}$ besagt, dass der Datenpunkt kein Ausreißer ist. Ausreißer behandeln wir dadurch, dass wir einen modifizierten Fehler $\tilde{\sigma}_i = \kappa\,\sigma_i > \sigma_i$ annehmen (d.h. $\kappa > 1$). Die Likelihood-Funktion lautet im Fall von Ausreißern

$$p(d_i|A, s_i, \kappa, a, \mathcal{B}) = \frac{1}{\sqrt{2\pi(\kappa\sigma_i)^2}}\, e^{-\frac{(d_i - f(s_i,a))^2}{2(\kappa\sigma_i)^2}} \quad .$$

Die Proposition A besagt, dass ein Ausreißer vorliegt. Uns interessiert die Wahrscheinlichkeit

$$p(a|\underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B}) = \frac{1}{Z}\, p(\underline{d}|a, \underline{s}, \underline{\sigma}, \mathcal{B})\; p(a|\underline{s}, \underline{\sigma}, \mathcal{B}) \quad .$$

Wir gehen hier davon aus, dass wir über die Parameter vor dem Experiment nichts aussagen können. Daher nehmen Wahrscheinlichkeitsrechnung hier eine konstante Wahrscheinlichkeit $p(a|\underline{s}, \underline{\sigma}, \mathcal{B})$ an[3]. Damit erhalten wir

$$p(a|\underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B}) = \frac{1}{Z'}\, p(\underline{d}|a, \underline{s}, \underline{\sigma}, \mathcal{B}) \quad .$$

Da die Daten unkorreliert sind, gilt

$$p(\underline{d}|a, \underline{s}, \underline{\sigma}, \mathcal{B}) = \prod_i\; p(d_i|a, s_i, \sigma_i, \mathcal{B}) \quad .$$

Über die Marginalisierungsregel führen wir die Proposition A ein, ob es sich um einen Ausreißer handelt oder nicht

$$p(d_i|a, s_i, \sigma_i, \mathcal{B}) = p(d_i|A, a, s_i, \kappa\,\sigma_i, \mathcal{B})\; P(A|\mathcal{B}) + p(d_i|\overline{A}, a, s_i, \sigma_i, \mathcal{B})\; P(\overline{A}|\mathcal{B}) \quad .$$

Es wurde davon ausgegangen, dass die Wahrscheinlichkeit $P(A|a, s_i, \sigma_i, \mathcal{B})$ nicht von den Parametern und den Steuergrößen abhängt, sondern allein eine Eigenschaft der Messapparatur ist. Wir führen die Abkürzung $\beta = P(A|\mathcal{B})$ ein. Daraus folgt $P(\overline{A}|\mathcal{B}) = 1 - \beta$. Die Likelihood ist demnach, unter Vorgabe der Werte β und κ,

$$p(\underline{d}|\beta, \kappa, a, \underline{s}, \underline{\sigma}, \mathcal{B}) = \prod_i \left((1-\beta)\, \mathcal{N}(d_i|y_i, \sigma_i) \;+\; \beta\, \mathcal{N}(d_i|y_i, \kappa\,\sigma_i) \right) \quad .$$

Nun sind diese HYPER-PARAMETER aber nicht bekannt und müssen über die Marginalisierungsregel eliminiert werden. Der Wert von β muss, da es sich um eine Wahrscheinlichkeit handelt, zwischen Null und Eins liegen. Wir gehen hier davon aus, dass alle Werte gleich-wahrscheinlich sind[4]. Da es sich bei κ um einen Skalen-Parameter

[3] Wenn detaillierteres Vorwissen vorliegt, lässt sich das leicht ergänzen.

[4] Es wird in realistischen Problemen eher so sein, dass man die Ausreißer-Wahrscheinlichkeit β aufgrund der Problemstellung genauer abschätzen kann.

handelt, ist JEFFREYS' PRIOR, also $P(\kappa|\mathcal{B}) \propto \frac{1}{\kappa}$, zu verwenden. Die Likelihood ist somit

$$\begin{aligned} p(\underline{d}|a, \underline{s}, \underline{\sigma}, \mathcal{B}) &= \int_0^1 d\beta \int_1^\infty d\kappa \prod_i p(d_i, \beta, \kappa|a, s_i, \sigma_i, \mathcal{B}) \\ &= \int_0^1 d\beta \int_1^\infty d\kappa \prod_i p(d_i|\beta, \kappa, a, s_i, \sigma_i, \mathcal{B}) \times \end{aligned}$$

$$\underbrace{p(\beta|\kappa, a, s_i, \sigma_i, \mathcal{B})}_{=p(\beta|\mathcal{B})=1} \underbrace{p(\kappa|a, s_i, \sigma_i, \mathcal{B})}_{=p(\kappa|\mathcal{B})\propto\frac{1}{\kappa}}$$

$$= \frac{1}{Z''} \int_0^1 d\beta \int_1^\infty \frac{d\kappa}{\kappa} \prod_i \Big((1-\beta)\, \mathcal{N}(d_i|y_i, \sigma_i) + \beta\, \mathcal{N}(d_i|y_i, \kappa\, \sigma_i) \Big) \quad .$$

Die weitere Analyse geht wie bisher. Es ist z.B. der Posterior-Erwartungswert $\langle a \rangle_i$ der Parameter zu bestimmen und die Vertrauensintervalle erhält man aus der Kovarianzmatrix

$$\langle \Delta a_i \Delta a_j \rangle = \int da\ (a_i - \langle a \rangle_i)(a_j - \langle a \rangle_j)\ p(a|\underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B})\ da \quad .$$

20.8.1 Vorhersagen

In der Regel ist man nicht an den Werten der Parameter selbst interessiert, sondern verwendet sie, um Vorhersagen für andere Werte der Steuergröße s zu machen. Der theoretische Wert der Messgröße y zur Steuergröße s bei gegebenen Parameter-Werten a ist

$$y_s = f(s, a) \quad .$$

Da die Parameter nicht exakt bekannt sind, sondern aus den gegebenen Daten $\underline{d}$ abgeschätzt werden, müssen wir nach der Wahrscheinlichkeitsdichte

$$p(y_s|s, \underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B})$$

für y_s fragen gegeben die Daten und das restliche Vorwissen. Mit der Marginalisierungsregel führen wir die Parameter ein

$$\begin{aligned} p(y_s|s, \underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B}) &= \int \underbrace{p(y_s|a, s, \cancel{\underline{d}}, \cancel{\underline{s}}, \cancel{\underline{\sigma}}, \mathcal{B})}_{\delta(y_s - f(s,a))}\ p(a|\cancel{s}, \underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B})\ da \\ &= \int \delta(y_s - f(s, a))\ p(a|\underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B})\ da \quad . \end{aligned}$$

Uns interessiert nicht die komplette Wahrscheinlichkeitsverteilung, sondern nur der Mittelwert und die Varianz

$$\begin{aligned} \langle y_s \rangle = E(y_s|s,\underline{d},\underline{s},\underline{\sigma}) &= \int f(s,a)\; p(a|\underline{d},\underline{s},\underline{\sigma},\mathcal{B})\; da \\ \left\langle (\Delta y_s)^2 \right\rangle = E((\Delta y_s)^2|s,\underline{d},\underline{s},\underline{\sigma}) &= \int (f(s,a) - \langle y_s \rangle)^2\; p(a|\underline{d},\underline{s},\underline{\sigma},\mathcal{B})\; da \end{aligned} \quad .$$

Für den wichtigen Spezialfall des lineare Modells $y_s = as + b$ gilt

$$\begin{aligned} \langle y_s \rangle &= \langle a \rangle s + \langle b \rangle \\ \left\langle (\Delta y_s)^2 \right\rangle &= \left\langle ((a - \langle a \rangle)s + (b - \langle b \rangle))^2 \right\rangle \\ &= \left\langle (\Delta a)^2 \right\rangle s^2 + 2 \left\langle \Delta a\, \Delta b \right\rangle s + \left\langle (\Delta b)^2 \right\rangle \end{aligned}$$

20.8.2 Beispiel: Schätzen des Mittelwerts

Den vorgestellten Formalismus kann man verwenden, um ausreißer-tolerant den Mittelwert von Daten

$$d_i, \qquad i = 1, \ldots, N$$

zu schätzen, indem man ein konstante Funktion als Modell wählt

$$y_i = f(s_i, a) = \mu \quad .$$

Ohne auf Ausreißer Rücksicht zu nehmen würde man den Parameter und dessen Varianz aus den Formeln (20.29)

$$\mu_0 = \frac{\sum_{i=1}^N d_i/\sigma_i^2}{\sum_{i=1}^N 1/\sigma_i^2} \qquad \text{und}$$
$$\sigma_0 = \frac{1}{\sum_{i=1}^N 1/\sigma_i^2}$$

berechnen. Diese Standard-Lösungen erhält man aus den obigen Formeln für den Prior $p(\beta|\mathcal{B}) = \delta(\beta)$. Abbildung 20.6 zeigt eine Studie, in der verschiedene Datensätze mit dieser Methode und mit der Standard-Lösung ausgewertet worden sind. Im Bild links oben sind alle Daten und Fehler mit dem Mittelwert kompatibel. Beide Methoden liefern dasselbe Ergebnis. Das ändert sich, wenn man Ausreißer hinzufügt. Nicht nur, dass bei der herkömmlichen Methode der Mittelwert durch die Ausreißer merklich verschoben wird, es wird auch der Fehler des Mittelwerts stark unterschätzt. Sind

zu viele Ausreißer vorhanden, sieht man, dass sich in der Wahrscheinlichkeitsdichte mehrere (in unserem Fall zwei) Moden ausbilden. Das ist ein Zeichen dafür, dass man nicht mehr sicher sagen kann, welche Daten die Ausreißer und welche die „wahren" Daten sind. In diesem Fall muss man sich zwangsläufig um mehr Daten bemühen, die dann (hoffentlich) eine Entscheidung erlauben.

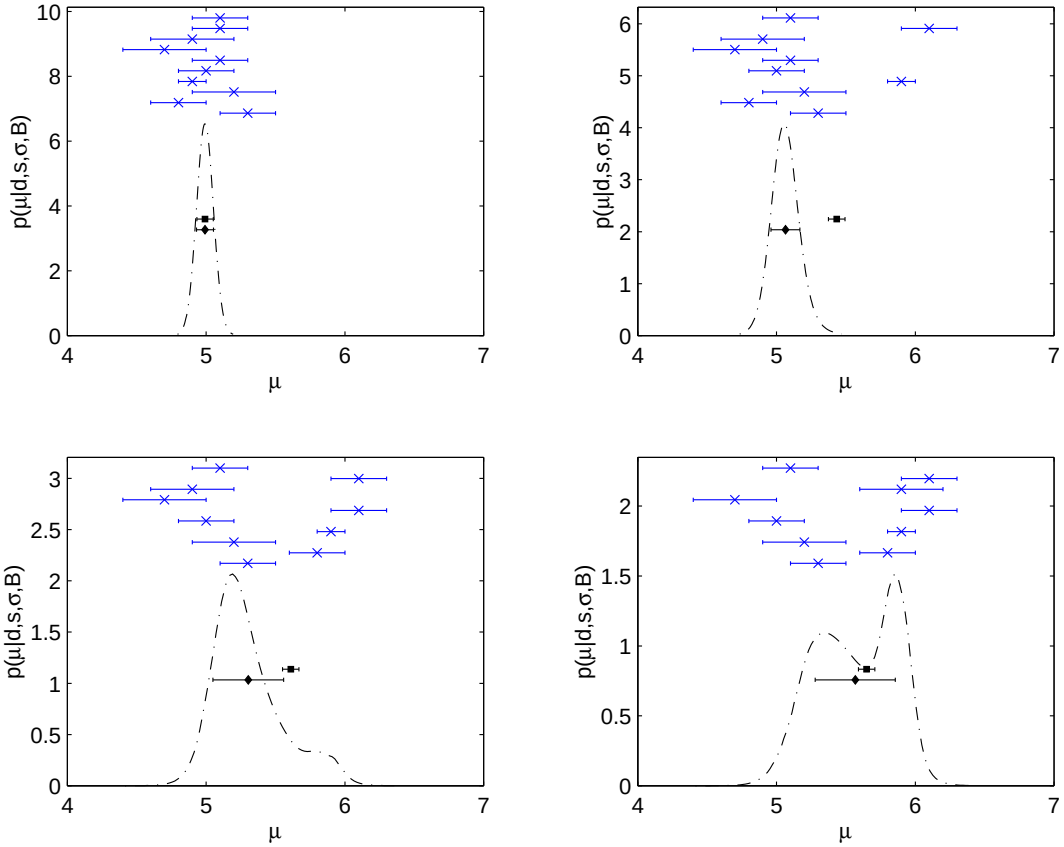

Abbildung 20.6: In diesen Abbildungen ist für verschiedene Datensätze (oberes Bilddrittel zeigt die Daten mit Fehlerbalken) die Wahrscheinlichkeitsdichte $p(\mu|\underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B})$ (strichpunktierte Linie), der Mittelwert mit Fehler (Raute) und der Mittelwert mit Fehler nach der Stanardmethode (Quadrat) dargestellt. Links oben sind die Daten kompatibel. In den folgenden Abbildungen wurden sukzessive Ausreißer hinzugefügt.

20.8.3 Beispiel: Geradenfit

Als nächstes wollen wir eine Gerade durch eine Punktmenge fitten. Unser Modellfunktion ist daher

$$y_i = a\, s_i + b$$

Abbildungen 20.7 und 20.8 zeigen die Ergebnisse zweier Testdatensätze. Das erste Beispiel zeigt, dass die Methode robust gegen Ausreißer ist, während ein herkömmlicher Fit natürlich Ausreißer nicht selbständig erkennen kann. An der Dichteverteilung des zweiten Beispiels kann man erkennen, dass in diesem Fall eine klare Trennung zwischen Ausreißern und „wahren" Daten nicht möglich ist.

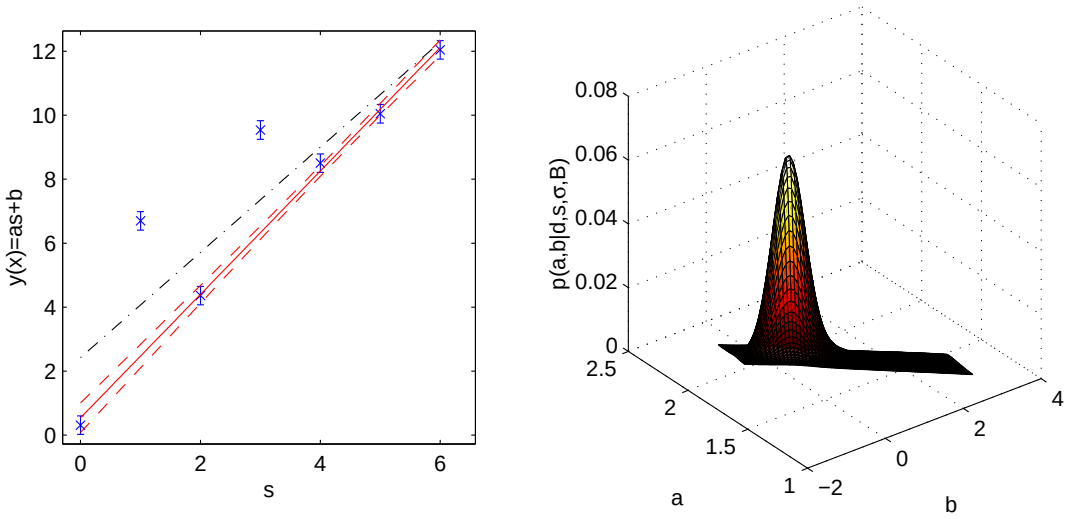

Abbildung 20.7: Ausreißer-toleranter Geradenfit. Im linken Bild sieht man die robust gerechnete Ausgleichsgerade (durchgezogen) mit dem Fehlerbereich (strichliert). Die strich-punktierte Linie zeigt einen herkömmlichen Fit. Das rechte Bild zeigt die Wahrscheinlichkeitsdichte $p(a, b|\underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B})$. In diesem Fall sind genügend Daten vorhanden, um die Ausreißer zu erkennen.

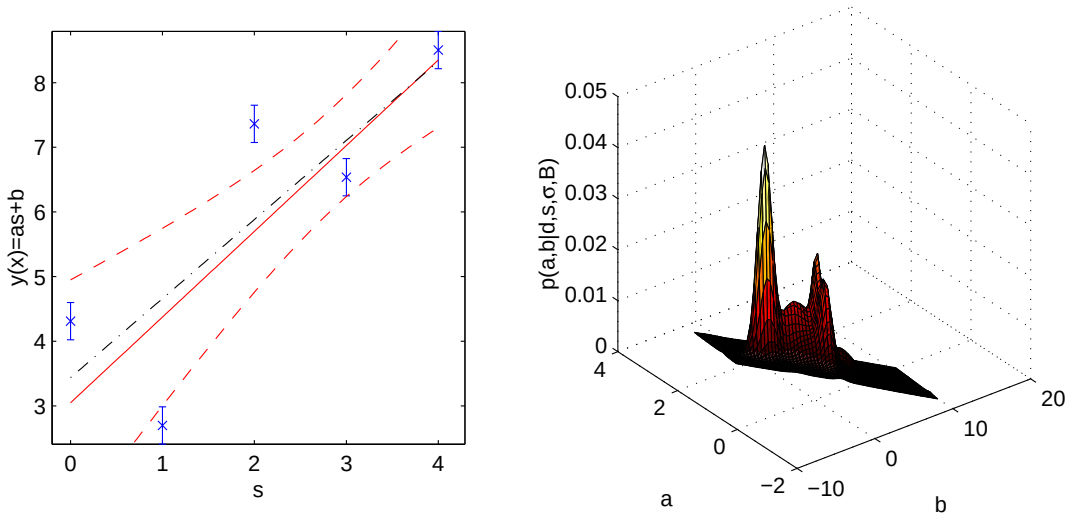

Abbildung 20.8: Ausreißer-toleranter Geradenfit. Im linken Bild sieht man die robust gerechnete Ausgleichsgerade (durchgezogen) mit dem Fehlerbereich (strichliert). Die strich-punktierte Linie zeigt einen herkömmlichen Fit. Das rechte Bild zeigt die Wahrscheinlichkeitsdichte $p(a, b|\underline{d}, \underline{s}, \underline{\sigma}, \mathcal{B})$. Die geringe Datenmenge mit offensichtlich falschen Fehlerangaben macht ein eindeutiges Erkennen der Ausreißer unmöglich.

Teil V

Hypothesentests

Kapitel 21

Stichproben-Verteilungen

21.1 Mittelwert und Median

Wir gehen von einer Stichprobe vom Umfang L i.u.v. Zufallszahlen einer Wahrscheinlichkeitsdichte $\rho(x)$ aus. Mittelwert und Varianz dieser Verteilung sind

$$\begin{aligned} \langle x \rangle &= \int x \; \rho(x) \; dx \\ \mathrm{var}(x) &= \int (x - \langle x \rangle)^2 \; \rho(x) \; dx \qquad . \end{aligned}$$

21.1.1 Verteilung des Stichproben-Mittelwertes

Def. 21.1 (Stichproben-Mittelwert) *Man definiert als*

STICHPROBEN-MITTELWERT (SAMPLE MEAN)
$$\overline{x} := \frac{1}{L} \sum_{i=1}^{L} x_i \qquad . \tag{21.1}$$

Der aktuelle Wert, den $\overline{x}$ annimmt, hängt von der jeweiligen Stichprobe ab. Es gibt also eine Wahrscheinlichkeitsverteilung $p(\overline{x}|L, \rho, \mathcal{B})$. Diese Wahrscheinlichkeit erhalten wir aus der Marginalisierungsregel

$$p(\overline{x}|L, \rho, \mathcal{B}) = \int \underbrace{p(\overline{x}|x_1, \ldots, x_L, L, \rho, \mathcal{B})}_{\delta\left(\overline{x} - \frac{1}{L}\sum_{i=1}^{L} x_i\right)} \; p(x_1, \ldots, x_L|L, \rho, \mathcal{B}) \; dx_1 \ldots dx_L \quad .$$

indem wir über die möglichen Stichproben integrieren. Da die Elemente der Stichprobe i.u.v. sind, gilt

$$p(\overline{x}|L,\rho,\mathcal{B}) = \int \delta\big(\overline{x} - \frac{1}{L}\sum_{i=1}^{L} x_i\big)\ \rho(x_1)\ldots\rho(x_L)\ dx_1\ldots dx_L \qquad .$$

Wir wollen hier diese Verteilung nicht weiter berechnen sondern lediglich den daraus resultierenden Mittelwert

$$\begin{aligned}
\langle\overline{x}\rangle &= \int \overline{x}\ p(\overline{x}|L,\rho,\mathcal{B})\ d\overline{x} \\
&= \int \left(\frac{1}{L}\sum_{i=1}^{L} x_i\right)\ \rho(x_1)\ldots\rho(x_L)\ dx_1\ldots dx_L \\
&= \frac{1}{L}\sum_{i=1}^{L} \int x_i\ \rho(x_1)\ldots\rho(x_L)\ dx_1\ldots dx_L \\
&= \frac{1}{L}\sum_{i=1}^{L} \left(\underbrace{\int x_i\ \rho(x_i)\ dx_i}_{=\langle x\rangle} \prod_{j\neq i} \underbrace{\int \rho(x_j)\ dx_j}_{=1} \right) \\
\langle\overline{x}\rangle &= \langle x\rangle \qquad .
\end{aligned}$$

Das heißt, der Stichproben-Mittelwert liefert im Mittel den wahren Mittelwert der zugrunde liegenden Wahrscheinlichkeitsdichte $\rho(x)$.
Man nennt deshalb das arithmetische Mittel der Stichprobe einen unverzerrten bzw. erwartungstreuen Schätzwert (UNBIASED ESTIMATOR) für den intrinsischen Mittelwert der Verteilung. Inwieweit das relevant ist, werden wir noch diskutieren.

21.1.2 Varianz der Stichproben-Mittelwerte

Als nächstes wollen wir untersuchen, wie weit der Stichproben-Mittelwert $\overline{x}$ um den Mittelwert streut. Dazu benötigen wir $\langle\overline{x}^2\rangle$

$$\begin{aligned}
\langle\overline{x}^2\rangle &= \int \overline{x}^2\ p(\overline{x}|L,\rho,\mathcal{B})\ d\overline{x} \\
&= \int \left(\frac{1}{L}\sum_{i=1}^{L} x_i\right)^2\ \rho(x_1)\ldots\rho(x_L)\ dx_1\ldots dx_L \\
&= \frac{1}{L^2}\sum_{i,j=1}^{L} \int x_i\ x_j\ \rho(x_1)\ldots\rho(x_L)\ dx_1\ldots dx_L
\end{aligned}$$

Hier müssen wir nun zwischen $i = j$ und $i \neq j$ unterscheiden

$$\begin{aligned}
\left\langle \overline{x}^2 \right\rangle &= \frac{1}{L^2} \sum_{i=1}^{L} \int x_i^2 \, \rho(x_1) \ldots \rho(x_L) \, dx_1 \ldots dx_L \\
&+ \frac{1}{L^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{L} \int x_i \, x_j \, \rho(x_1) \ldots \rho(x_L) \, dx_1 \ldots dx_L \\
&= \frac{1}{L^2} \sum_{i=1}^{L} \left\langle x^2 \right\rangle + \frac{1}{L^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{L} \langle x \rangle \, \langle x \rangle \\
&= \frac{1}{L} \left\langle x^2 \right\rangle + \frac{L \, (L-1)}{L^2} \langle x \rangle^2 = \frac{1}{L} \left(\left\langle x^2 \right\rangle - \langle x \rangle^2 \right) + \langle x \rangle^2 \\
&= \frac{\langle (\Delta x)^2 \rangle}{L} + \langle x \rangle^2 = \frac{\mathrm{var}(x)}{L} + \langle x \rangle^2
\end{aligned}$$

Daraus folgt wegen $\langle \overline{x} \rangle = \langle x \rangle$ die

VARIANZ DER STICHPROBEN-MITTELWERTE

$$\mathrm{var}(\overline{x}) = \frac{\mathrm{var}(x)}{L} \quad . \tag{21.2}$$

Die mittlere quadratischen Abweichung (Varianz) besagt, dass die Streuung von $\overline{x}$ um den wahren Mittelwert $\langle x \rangle$ im Mittel $\sqrt{\mathrm{var}(\overline{x})}$ beträgt. Diese Größe nennt man statistischen Fehler des Mittelwertes bzw.

STANDARDFEHLER

$$\Delta \overline{x} := \mathrm{SF} := \frac{\sqrt{\mathrm{var}(x)}}{\sqrt{L}} \tag{21.3}$$

Er unterschiedet sich von der STANDARD-ABWEICHUNG $\sqrt{\mathrm{var}(x)}$ durch den Faktor $1/\sqrt{L}$. Das ist die Grundlage fast aller experimenteller Bestimmungen von unbekannten Größen, die im Experiment mit einem statistischen Fehler behaftet sind. Dasselbe gilt für (Quanten-)Monte-Carlo-Verfahren. Der Standardfehler besagt, dass man mit zunehmendem Stichproben-Umfang den statistischen Fehler immer stärker unterdrücken kann.
Eine kleine Unzulänglichkeit besteht noch. Wir verwenden die Stichprobe, um den wahren Mittelwert der Wahrscheinlichkeitsverteilung abzuschätzen. Der Fehler den

wir dabei machen, hängt von $\mathrm{var}(x)$ ab, das wir aber ebensowenig kennen wie den Mittelwert. Woher bekommen wir also $\mathrm{var}(x)$? Es ist naheliegend, auch die Varianz aus der Stichprobe abzuschätzen

$$v = \frac{1}{L} \sum_{i=1}^{L} \left(x_i - \overline{x}\right)^2 \qquad .$$

Den wahren Mittelwert haben wir hier durch den Schätzwert $\overline{x}$ ersetzt. Der Schätzwert für die Varianz kann umgeschrieben werden in

$$\begin{aligned} v &= \frac{1}{L} \sum_{i=1}^{L} x_i^2 - 2\,\overline{x}\,\underbrace{\frac{1}{L}\sum_{i=1}^{L} x_i}_{\overline{x}} + \frac{1}{L}\sum_{i=1}^{L} \overline{x}^2 \\ &= \frac{1}{L} \sum_{i=1}^{L} x_i^2 - \overline{x}^2 \qquad . \end{aligned}$$

Auch dieser Schätzwert ist eine Zufalls-Variable, die von der jeweiligen Stichprobe abhängt. Der Mittelwert liefert

$$\begin{aligned} \langle v \rangle &= \underbrace{\frac{1}{L} \sum_{i=1}^{L} \left\langle x_i^2 \right\rangle}_{\langle x^2 \rangle} - \left\langle \overline{x}^2 \right\rangle \\ &= \left\langle x^2 \right\rangle - \frac{\mathrm{var}(x)}{L} - \langle x \rangle^2 \\ &= \frac{L-1}{L}\,\mathrm{var}(x) \qquad . \end{aligned}$$

Offensichtlich weist dieser Schätzwert im Mittel eine Abweichung (Bias) vom wahren Wert der Varianz auf. Um diese Abweichung zu beheben führt man den Schätzwert

UNBIASED ESTIMATOR FÜR DIE VARIANZ

$$\sigma^2_{\mathrm{est}} = \frac{1}{L-1} \sum_{i=1}^{L} (x_i - \overline{x})^2 \tag{21.4}$$

ein. Der modifizierte Vorfaktor $1/(L-1)$ ist wie folgt zu verstehen. Die Größen $\Delta_i = x_i - \overline{x}$ sind nicht mehr unabhängig voneinander, da $\sum_i \Delta_i = 0$. Es gibt nur noch $L-1$ unabhängige Größen ξ_i, die aus den x_i durch Linearkombination gebildet werden können. Die Zahl der FREIHEITSGRADE wurde um eins verringert. Die ξ_i

sind dann i.u.v. nach einer Wahrscheinlichkeitsdichte $\tilde{\rho}(x)$ mit Mittelwert Null und Varianz $\mathrm{var}(x)$. D.h.

$$\sum_{i=1}^{L} \Delta_i^2 = \sum_{i=1}^{L-1} \xi_i^2 \qquad .$$

Diese Summe ergibt im Mittel den Wert $(L-1)\ \mathrm{var}(x)$.

21.1.3 Verteilung des Stichproben-Medians

Wir wollen untersuchen, wie sich der Median einer Stichprobe vom Umfang L verhält. Es sollen alle reellen kontinuierlichen Zufalls-Variablen der Stichprobe i.u.v. sein. Die zugrunde liegende Verteilung sei $\rho(x)$.

Def. 21.2 (Median) *Den Median definiert man sowohl für Wahrscheinlichkeitsverteilungen als auch für Stichproben.*

MEDIAN (ZENTRALWERT) EINER WAHRSCHEINLICHKEITSVERTEILUNG

a) Der Median $\hat{x}$ einer KONTINUIERLICHEN ZUFALLS-VARIABLEN *x mit Wahrscheinlichkeitsdichte $\rho(x)$ und Verteilungsfunktion $F(x)$ ist definiert über*

$$\begin{aligned} \frac{1}{2} &= \int_{-\infty}^{\hat{x}} \rho(t)\ dt = F(\hat{x}) \\ \hat{x} &= F^{-1}(\tfrac{1}{2}) \qquad . \end{aligned} \tag{21.5}$$

b) Der Median $\hat{x}$ einer DISKRETEN ZUFALLS-VARIABLEN *x, die mit den Wahrscheinlichkeiten P_i die Werte x_i annimmt, ist definiert über*

$$\hat{x} = \begin{cases} x_\nu & \text{wenn } \sum_{i=1}^{\nu-1} P_i < \frac{1}{2} < \sum_{i=1}^{\nu} P_i \\ \frac{x_\nu + x_{\nu+1}}{2} & \text{wenn } \sum_{i=1}^{\nu} P_i = \frac{1}{2} \end{cases} \tag{21.6}$$

Die Definition im diskreten Fall ist kompatibel zur Definition des Medians einer Stichprobe

MEDIAN (ZENTRALWERT) EINER STICHPROBE

Zunächst werden die Elemente $\{s_1, s_2, \ldots, s_L\}$ *der Stichprobe nach ansteigendem Wert* $\{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_L\}$ *sortiert. Dann ist der Median*

$$\breve{x} = \begin{cases} \tilde{s}_{n+1} & \text{wenn } L = 2n+1 \\ \\ \dfrac{\tilde{s}_n + \tilde{s}_{n+1}}{2} & \text{wenn } L = 2n \end{cases} \quad . \tag{21.7}$$

Zurück zum Fall diskreter Zufalls-Variablen mit den Wahrscheinlichkeiten P_i. Bei einer Stichprobe vom Umfang L wird es im Mittel $L\ P_i$ mal den Wert x_i in der Stichprobe geben. Wir sortieren die Elemente der Stichprobe der Größe nach $x_j \leq x_{j+1}$. Wenn es kein ν gibt, so dass $\sum_{i=1}^{\nu}\ P_i = 1/2$, dann wird die Mitte der Stichprobe in eine Sequenz von gleichen Elementen x_ν fallen.

Index	1	2	3	...	$\frac{L}{2}-2$	$\frac{L}{2}-1$	$\frac{L}{2}$	$\frac{L}{2}+1$	$\frac{L}{2}+2$	...	L
Stichprobe	x_1	x_1	x_1	...	x_ν	x_ν	x_ν	x_ν	x_ν	...	x_M

Wenn die Wahrscheinlichkeiten jedoch gerade so sind, dass es ein ν mit $\sum_{i=1}^{\nu}\ P_i = 1/2$ gibt, dann wird auch für $L \to \infty$ der Median bei ungeradzahligem L von Stichprobe zu Stichprobe zwischen x_ν und $x_{\nu+1}$ streuen und im Mittel $(x_\nu + x_{\nu+1})/2$ liefern. Wir interessieren uns für die Wahrscheinlichkeit $p(\breve{x}|L,\mathcal{B})\ d\breve{x}$, dass der Median der Stichprobe vom Umfang L Werte aus dem infinitesimalen Intervall um $\breve{x}$ annimmt. Der Median einer Stichprobe ist derjenige Wert der Zufalls-Variablen, bei dem die Verteilungsfunktion den Wert $F(\breve{x}) = 1/2$ annimmt. Das heißt, genau in der Hälfte der Fälle werden Werte kleiner bzw. größer als $\breve{x}$ sein. Die Stichprobe habe einen ungeradzahligen Umfang $L = 2n+1$. Die Stichprobe hat genau dann den Median $\breve{x}$, wenn ein Element der Stichprobe den Wert $\breve{x}$ hat, und wenn die verbleibenden $2n$ Elemente der Stichprobe zur Hälfte größere und zur Hälfte kleinere Werte als $\breve{x}$ besitzt. Die Wahrscheinlichkeit hierfür ist also proportional zu

$$P(\text{n Werte kleiner als } \breve{x}|2n+1,\mathcal{B})\cdot$$
$$\cdot\, P(\text{n Werte größer } \breve{x}|2n+1,\mathcal{B})\ p(\breve{x}|2n+1,\mathcal{B}) \quad .$$

Es handelt sich hier um nichts anderes als die Ordnungs-Statistik (Gl. (9.4)). Die Wahrscheinlichkeit, dass eine Zufallszahl einen Wert kleiner als $\breve{x}$ annimmt, ist durch

die Verteilungsfunktion $F(\breve{x})$ gegeben. Da die Element der Stichprobe i.u.v. sind, ist die gesuchte Wahrscheinlichkeit

$$p(\breve{x}|2n,\mathcal{B}) = \frac{1}{Z}\; F(\breve{x})^n\; (1-F(\breve{x}))^n\; \rho(\breve{x}) \qquad .$$

Die Normierungskonstante folgt aus

$$Z = \int\; F(\breve{x})^n\; (1-F(\breve{x}))^n\; \rho(\breve{x})\; d\breve{x} \qquad .$$

Wir führen eine Variablen-Substitution durch

$$\begin{aligned} q &:= F(\breve{x}) \\ dq &= \frac{dF(\breve{x})}{d\breve{x}} d\breve{x} = \rho(\breve{x}) d\breve{x} \\ \breve{x} &= F^{-1}(q) \qquad . \end{aligned}$$

Demnach ist die Normierung

$$Z = \int_0^1\; q^n\; (1-q)^n\; dq = B(n+1,n+1) = \frac{n!\; n!}{(2n+1)!} \qquad .$$

Diese Normierung hätten wir natürlich auch von Gl. (9.4) der Ordnungs-Statistik ablesen können. Wir gehen davon aus, dass die Dichte $\rho(x)$ keine Nullstellen hat, sonst ist F^{-1} nicht eindeutig.
Wir können nun den Erwartungswert des Stichproben-Medians ermitteln. Dazu benötigen wir

$$\begin{aligned} \left\langle \breve{x} \right\rangle &= \frac{1}{Z}\; \int\; \breve{x}\; F(\breve{x})^n\; (1-F(\breve{x}))^n\; \rho(\breve{x})\; d\breve{x} \\ &= \int_0^1\; F^{-1}(q)\; \underbrace{\frac{q^n\; (1-q)^n}{Z}}_{p_\beta(q|n+1,n+1)}\; dq \\ &= \int_0^1\; F^{-1}(q)\; p_\beta(q|n+1,n+1)\; dq := \left\langle F^{-1}(q) \right\rangle \end{aligned}$$

Die Dichte der β-Verteilung $p_\beta(q|n+1,n+1)$ hat den Mittelwert $q = 1/2$ und die Varianz $\sigma_\beta^2 = \frac{1}{4(2n+3)} = \frac{1}{4(L+2)}$. Nur wenn $F^{-1}(q)$ eine anti-symmetrische Funktion um $q = \frac{1}{2}$ ist, liefert der Median der Stichprobe im Mittel den Median[1] der zugrunde liegenden Verteilung

$$\left\langle \breve{x} \right\rangle = F^{-1}(1/2) = \hat{x} \qquad .$$

[1] Die Definition des Medians ist $F(\hat{x}) = 1/2$, bzw $\hat{x} = F^{-1}(1/2)$.

Ansonsten entwickeln wir $F^{-1}(q)$ um den Mittelwert $q = 1/2$ der β-Verteilung

$$\begin{aligned}\left\langle \breve{x} \right\rangle = \left\langle F^{-1}(q) \right\rangle &= \left\langle F^{-1}(1/2) \right\rangle + \left.\frac{dF^{-1}(q)}{dq}\right|_{q=1/2} \underbrace{\left\langle (q-1/2) \right\rangle}_{0} \\ &+ \frac{1}{2}\left.\frac{d^2F^{-1}(q)}{dq^2}\right|_{q=1/2} \underbrace{\left\langle (q-1/2)^2 \right\rangle}_{\sigma_\beta^2} + O(\sigma_\beta^4) \\ &= \left\langle F^{-1}(q) \right\rangle = F^{-1}(1/2) + \frac{1}{2}\left.\frac{d^2F^{-1}(q)}{dq^2}\right|_{q=1/2} \sigma_\beta^2 + O(\sigma_\beta^4) \qquad .\end{aligned}$$

Das heißt, im Mittel liefert der Median der Stichprobe den Median der zugrunde liegenden Wahrscheinlichkeitsdichte $\rho(x)$ nur bis auf eine systematische Abweichung der Ordnung $O(\sigma^2)$

$$\left\langle \breve{x} \right\rangle = \hat{x} + \frac{1}{2}\left.\frac{d^2F^{-1}(q)}{dq^2}\right|_{q=1/2} \sigma_\beta^2 + O(\sigma_\beta^4) \qquad .$$

21.1.4 Varianz des Stichproben-Medians

Als nächstes berechnen wir die Varianz von $p(\breve{x}|2n+1, \mathcal{B})$

$$\begin{aligned}\mathrm{var}(\breve{x}) &= \left\langle \left(F^{-1}(q) - \left\langle \breve{x} \right\rangle \right)^2 \right\rangle \\ &= \left\langle \left(\hat{x} + \left.\frac{dF^{-1}(q)}{dq}\right|_{q=1/2} (q-1/2) + O(\sigma_\beta^2) - \hat{x} - O(\sigma_\beta^2) \right)^2 \right\rangle \\ &= \left(\left.\frac{dF^{-1}(q)}{dq}\right|_{q=1/2} \right)^2 \underbrace{\left\langle (q-1/2)^2 \right\rangle}_{\sigma_\beta^2} + O(\sigma_\beta^4) \\ &= \left(\left.\frac{dF^{-1}(q)}{dq}\right|_{q=1/2} \right)^2 \sigma_\beta^2 \; + \; O(\sigma_\beta^4)\end{aligned}$$

Wir benötigen noch die erste und zweite Ableitung von $F^{-1}(q)$ nach q

$$\begin{aligned}\frac{d(F(F^{-1}(q)))}{dq} &= \frac{dq}{dq} = 1 \\ &= \frac{dF(x)}{dx}\bigg|_{x=F^{-1}(q)} \frac{dF^{-1}(q)}{dq} \\ &= \rho(F^{-1}(q)) \ \frac{dF^{-1}(q)}{dq} \\ &\Rightarrow \end{aligned}$$

$$\frac{dF^{-1}(q)}{dq} = \frac{1}{\rho(F^{-1}(q))}$$

Die zweite Ableitung folgt hieraus

$$\begin{aligned}\frac{d^2F^{-1}(q)}{dq^2} &= \frac{d}{dq}\ \frac{1}{\rho(F^{-1}(q))} \\ &= -\frac{\rho'(F^{-1}(q))}{\rho(F^{-1}(q))^2}\ \frac{dF^{-1}(q)}{dq} \\ &= -\frac{\rho'(F^{-1}(q))}{\rho(F^{-1}(q))^3} \qquad .\end{aligned}$$

Die Ableitungen an der Stelle $q = 1/2$ sind

$$\begin{aligned}\frac{dF^{-1}(q)}{dq}\bigg|_{q=1/2} &= \frac{1}{\rho(\breve{x}^{\text{Med}})} \\ \frac{d^2F^{-1}(q)}{dq^2}\bigg|_{q=1/2} &= -\frac{\rho'(\breve{x}^{\text{Med}})}{\rho(\breve{x}^{\text{Med}})^3} \qquad .\end{aligned}$$

Damit ist das Endergebnis

MITTELWERT UND VARIANZ DES MEDIANS EINER STICHPROBE VOM UMFANG $L = 2n + 1$

$$\left\langle \breve{x} \right\rangle = \begin{cases} \hat{x} & \text{für } F^{-1}(\frac{1}{2} - \Delta q) = -F^{-1}(\frac{1}{2} + \Delta q) \\ \hat{x} - \frac{\rho'(\hat{x})}{8\,(L+2)\,\rho(\hat{x})^3} + O(L^{-2}) & \text{sonst} \end{cases} \tag{21.8a}$$

$$\mathrm{var}(\breve{x}) = \left(\frac{1}{4(L+2)\,\rho(\hat{x})^2} \right) + O(L^{-2}) \quad . \tag{21.8b}$$

Wenn die Umkehrfunktion $F^{-1}(q)$ anti-symmetrisch um $q = 1/2$ ist, heißt das auch, dass der Median gleich dem Mittelwert ist. Man erkennt an diesem Ergebnis, dass der Median robuster ist als Mittelwert, da der Standardfehler nicht von der Varianz der Verteilung $\rho(x)$ abhängt. Das ist insbesondere dann wichtig, wenn die Verteilung langreichweitige Ausläufer hat.

Beispiel: Cauchy-Verteilung

Die Dichte der Cauchy-Verteilung ist

$$\rho(x) = \frac{1}{\pi\,(1 + x^2)} \quad .$$

Mittelwert und Varianz der Cauchy-Verteilung sind $\langle x \rangle = 0$ und $\mathrm{var}(x) = \infty$. Die Cauchy-Verteilung ist symmetrisch um den Mittelwert. Deshalb stimmen Median und Mittelwert überein. Wir hatten abgeleitet, dass das arithmetische Mittel einer Stichprobe um den intrinsischen Mittelwert der zugrundeliegenden Verteilung streut und somit im Mittel den richtigen Wert liefert. Die Varianz der Cauchy-Verteilung ist jedoch unendlich und somit ist auch der Standard-Fehler unendlich.
Wir wollen nun die Stichproben-Verteilung des Stichproben-Medians untersuchen. Wir benötigen hierzu die Verteilungsfunktion

$$F_C(x) = \int_{-\infty}^{x} \rho(t)\,dt = \frac{1}{2} + \frac{\arctan(x)}{\pi} \quad .$$

Diese kann leicht invertiert werden.

$$F_C^{-1}(q) = \tan(\pi(q - \frac{1}{2})) \quad .$$

Diese Funktion ist offensichtlich anti-symmetrisch um $q = 1/2$. Der Median ist demnach

$$\hat{x} = F_C^{-1}(1/2) = 0 \qquad .$$

Es bestätigt sich, dass Mittelwert und Median übereinstimmen. Das bedeutet, dass der Mittelwert des Medians einer Stichprobe gleich dem Median der Cauchy-Verteilung ist. Der Median stellt also im Fall der Cauchy-Verteilung ebenfalls einen UNBIASED ESTIMATOR des Mittelwertes dar. Die Varianz des Stichproben-Medians ist

$$\mathrm{var}(\breve{x}) = \left(\frac{\sigma_\beta}{\rho(0)}\right)^2 + O(\sigma_\beta^4) = \frac{\pi^2}{4(L+2)} + O(L^{-2}) = \frac{\pi^2}{4L} + O(L^{-2}) \qquad .$$

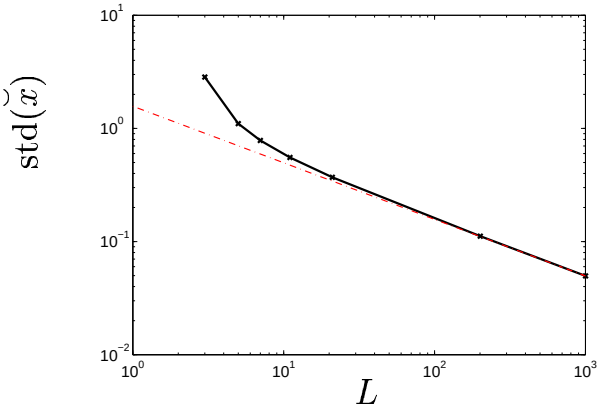

Abbildung 21.1: *Standardabweichung des Medians von Stichproben vom Umfang L als Funktion von L für die Cauchy-Verteilung. Gestrichelt eingetragen ist $\sqrt{\mathrm{var}(\breve{x})} \approx \pi/2\sqrt{L}$.*

Die Standardabweichung des Medians ist für Stichproben vom Umfang $L \geq 3$ endlich und fällt sogar mit $1/\sqrt{L}$ ab. Der Median stellt in diesem Fall einen „effizienteren" Schätzwert dar, da er einen kleinen Standardfehler besitzt.

Beispiel: Exponential-Verteilung

Für die Exponential-Verteilung liefert der Median der Stichprobe keinen UNBIASED ESTIMATOR für den intrinsischen Median der zugrunde liegenden Verteilung. Der Median der Stichprobe weicht für endliche Stichproben systematisch vom wahren Mittelwert, bzw. Median ab.
Gemäß Gl. (9.15b) ist die Verteilungsfunktion der Exponential-Verteilung

$$F_e(x|\lambda) = 1 - e^{-\lambda\, x} \qquad .$$

Die inverse Funktion ist

$$F^{-1}(q|\lambda) = -\frac{\ln(1-q)}{\lambda} \qquad .$$

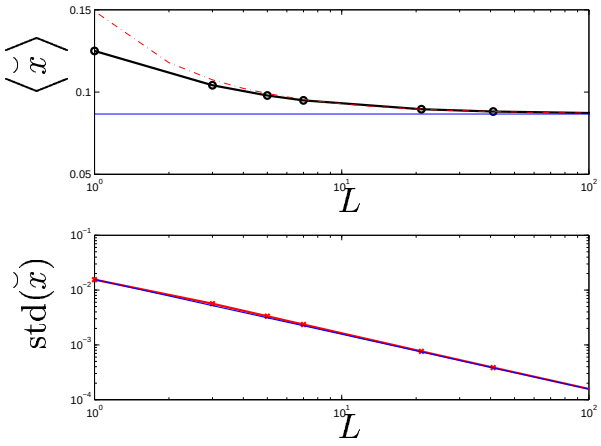

Abbildung 21.2: *Mittelwert (oberes Bild) und Standardabweichung (unteres Bild) des Medians von Stichproben vom Umfang L als Funktion von L für die Exponential-Verteilung. Der exakte Wert des Medians ist im oberen Bild zum Vergleich eingezeichnet.*

Der Median ist demnach

$$\hat{\breve{x}} = -\frac{\ln(1-1/2)}{\lambda} = \frac{\ln(2)}{\lambda} \quad .$$

Die Dichte an der Stelle des Medians ist $\rho_e(\hat{\breve{x}}) = \frac{\lambda}{2}$ und die Ableitung der Dichte an der Stelle des Medians ist $\rho_e'(\hat{\breve{x}}) = -\frac{\lambda^2}{2}$. Die Varianz des Stichproben-Medians ist nach Gl. (21.8b)

$$\mathrm{var}(\breve{x}) = \frac{1}{4L\left(\frac{\lambda}{2}\right)^2} + O(L^{-2}) = \frac{1}{L\lambda^2} + O(L^{-2})$$

Der Stichproben-Median liefert dann gemäß Gl. (21.8a) im Mittel

$$\left\langle \breve{x} \right\rangle = \hat{\breve{x}} - \frac{-\frac{\lambda^2}{2}}{8L\,(\frac{\lambda}{2})^3} + O(L^{-2}) = \frac{\ln(2)}{\lambda} + \frac{1}{2L\,\lambda} + O(L^{-2}) \quad .$$

Für Stichproben mit endlichem Umfang liegt also eine systematische Abweichung vom intrinsischen Wert des Medians vor, der allerdings wie $1/L$ verschwindet, während der Standard-Fehler nur wie $1/\sqrt{L}$ gegen Null geht.
In Abbildung 21.2 sind neben den exakten Werten zu $\lambda = 8$ auch die Näherungsformeln

$$\left\langle \breve{x} \right\rangle = \hat{\breve{x}} + \frac{1}{16\,L}$$

und

$$\mathrm{var}(\breve{x}) = \frac{1}{64\,L}$$

eingetragen.

21.2 Verteilung von Mittelwert und Varianz in normalen Stichproben

Angenommen wir haben eine Stichprobe vom Umfang N von i.u.normal-verteilte (i.u.nv.) Zufalls-Variablen x_n mit Mittelwert x_0 und Varianz σ^2. Daraus bestimmen wir den Stichproben-Mittelwert

$$\overline{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

und die unnormierte Stichproben-Varianz

$$v = \sum_{n=1}^{N} (x_n - \overline{x})^2 \qquad .$$

Wir fragen nach der Verteilung $p(\overline{x}, v|N, x_0, \sigma, \mathcal{B})$ von $\overline{x}$ und v, wobei der Bedingungskomplex $\mathcal{B}$ besagt, dass die Stichprobe aus i.u.nv. Zufalls-Variablen besteht. Die Marginalisierungsregel erlaubt es, durch Einführen der Stichprobenwerte $\underline{x} = \{x_1, \ldots, x_N\}$, die gesuchte Wahrscheinlichkeitsdichte zu berechnen

$$p(\overline{x}, v|x_0, \sigma, N, \mathcal{B}) = \int d^N x \; p(\overline{x}, v|\underline{x}, x_0, \sigma, N, \mathcal{B}) \; p(\underline{x}|x_0, \sigma, N, \mathcal{B}) \qquad . \tag{21.9}$$

Wenn die Stichprobe vorliegt, ist $\overline{x}$ und v deterministisch festgelegt

$$p(\overline{x}, v|x_1, \ldots, x_N, N, x_0, \sigma, \mathcal{B}) = \delta\left(\overline{x} - \frac{1}{N} \sum_{n=1}^{N} x_n\right) \; \delta\left(v - \sum_{n=1}^{N} (x_n - \overline{x})^2\right) . \tag{21.10}$$

Die Stichproben-Wahrscheinlichkeitsdichte ist ebenso bekannt, da die x_n i.u.nv.

$$\begin{aligned} p(\underline{x}|x_0, \sigma, N, \mathcal{B}) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{-(x_n - x_0)^2/2\sigma^2} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \; e^{-\sum_n (x_n - x_0)^2/2\sigma^2} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \; e^{-\sum_n (x_n - \overline{x} + \overline{x} - x_0)^2/2\sigma^2} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \; e^{-\sum_n (x_n - \overline{x})^2/2\sigma^2} \; e^{-N(\overline{x} - x_0)^2/2\sigma^2} \qquad . \end{aligned}$$

Zusammen mit Gl. (21.10) wird aus Gl. (21.9)

$$\begin{aligned} p(\overline{x}, v|x_0, \sigma, N, \mathcal{B}) = (2\pi\sigma^2)^{-\frac{N}{2}} \; e^{-N(\overline{x} - x_0)^2/2\sigma^2} \int d^N x \; e^{-\sum_n (x_n - \overline{x})^2/2\sigma^2} \ldots \\ \ldots \delta\left(\overline{x} - \frac{1}{N} \sum_{n=1}^{N} x_n\right) \; \delta\left(v - \sum_{n=1}^{N} (x_n - \overline{x})^2\right) \qquad . \end{aligned} \tag{21.11}$$

Wir führen nun die sogenannte Helmert-Transformation[2] durch

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \xi_i \\ \vdots \\ \xi_N \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & & & & & \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & & & & \\ \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & -\frac{3}{\sqrt{12}} & & & \\ \vdots & \vdots & & \ddots & \ddots & & \\ \frac{1}{\sqrt{i(i+1)}} & \frac{1}{\sqrt{i(i+1)}} & \frac{1}{\sqrt{i(i+1)}} & \frac{1}{\sqrt{i(i+1)}} & \frac{1}{\sqrt{i(i+1)}} & -\frac{i}{\sqrt{i(i+1)}} & \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots \\ \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_i \\ \vdots \\ x_N \end{pmatrix} \tag{21.12}$$

für die gilt, dass die Jakobi-Determinante eins ist. Aus der letzten Zeile lesen wir ab

$$\xi_N = \sqrt{N}\,\overline{x} = \sqrt{N}\,\frac{1}{N}\,\sum_{n=1}^{N} x_n \qquad .$$

Außerdem kann man zeigen, dass

$$\sum_{n=1}^{N} (x_n - \overline{x})^2 = \sum_{n=1}^{N-1} \xi_n^2 \qquad .$$

Somit haben wir

$$\begin{aligned} p(\overline{x}, v, x_0, \sigma | N, x_0, \sigma, \mathcal{B}) = (2\pi\sigma^2)^{-\frac{N}{2}}\; e^{-N\frac{(\overline{x}-x_0)^2}{2\sigma^2}} \int\; d^{N-1}\xi\; d\xi_N\; \ldots \\ \ldots\; e^{-\sum_{n=1}^{N-1}\xi_n^2/2\sigma^2} \delta\left(\overline{x} - \xi_N/\sqrt{N}\right)\; \delta\left(v - \sum_{n=1}^{N-1}\xi_n^2\right) \\ = \sqrt{N}\;(2\pi\sigma^2)^{-\frac{N}{2}}\; e^{-N\frac{(\overline{x}-x_0)^2}{2\sigma^2}}\; e^{-\frac{v}{2\sigma^2}}\; \ldots \\ \ldots \int\; d^{N-1}\xi\; \delta\left(v - \sum_{n=1}^{N-1}\xi_n^2\right) \qquad . \end{aligned}$$

Zur Berechnung des Integrals führen wir neue Variablen $\xi_n = z_n\,\sqrt{v}$ ein. Hierdurch wird aus dem Integral

$$\begin{aligned} \int\; d^{N-1}\xi\; \delta\left(v - \sum_{n=1}^{N-1}\xi_n^2\right) &= v^{\frac{N-1}{2}} \int\; d^{N-1}z\; \delta\left(v(1 - \sum_{n=1}^{N-1} z_n^2)\right) \\ &= v^{\frac{N-3}{2}} \int\; d^{N-1}z\; \delta\left(1 - \sum_{n=1}^{N-1} z_n^2\right) \\ &= a\; v^{\frac{N-3}{2}} \qquad , \end{aligned}$$

[2] *Kandall's advanced theory of statistics*, Vol. 1, Distribution Theory, §11.3.

wobei a eine Konstante ist, die wir nachträglich über die Normierung bestimmen werden. Damit lautet das Endergebnis

$$p(\overline{x}, v|N, x_0, \sigma, \mathcal{B}) = \frac{1}{Z}\, v^{\frac{N-3}{2}}\, e^{-\frac{v}{2\sigma^2}} \quad \times \quad \frac{1}{\sqrt{2\pi\sigma^2/N}} e^{-\frac{(\overline{x}-x_0)^2}{2\sigma^2/N}} \quad . \tag{21.13}$$

Wir sehen, dass die Wahrscheinlichkeitsdichte in die Wahrscheinlichkeitsdichte für den Stichproben-Mittelwert und die der Stichproben-Varianz faktorisiert.

21.2.1 Stichproben-Mittelwert

Der hintere Faktor von Gl. (21.13) ist die marginale Wahrscheinlichkeitsdichte für den Stichproben-Mittelwert.

WAHRSCHEINLICHKEITSDICHTE DES STICHPROBEN-MITTELWERTS NORMALVERTEILTER ZUVALLSVARIABLEN

GEGEBEN:
Stichprobe: $\{x_1, \ldots, x_N\}$ aus $\mathcal{N}(x|x_0, \sigma)$

BEKANNT:
N, x_0, σ

ZUFALLS-VARIABLE:
$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$\begin{aligned} p(\overline{x}|N, x_0, \sigma, \mathcal{B}) &= \frac{1}{\sqrt{2\pi\sigma^2/N}} e^{-\frac{(\overline{x}-x_0)^2}{2\sigma^2/N}} \\ &= \mathcal{N}(\overline{x}|x_0, \sigma/\sqrt{N}) \quad . \end{aligned} \tag{21.14}$$

Mittelwert $\langle\overline{x}\rangle = x_0$ und Varianz $\mathrm{var}(\overline{x}) = \frac{\mathrm{var}(x)}{N}$ haben wir bereits kennengelernt.

21.2.2 Stichproben-Varianz, χ^2-Statistik

Die Wahrscheinlichkeitsdichte für die Stichproben-Varianz ist mit der richtigen Normierung

WAHRSCHEINLICHKEITSDICHTE DER STICHPROBEN-VARIANZ NORMALVERTEILTER ZUVALLSVARIABLEN

GEGEBEN:
Stichprobe: $\{x_1, \ldots, x_N\}$ aus $\mathcal{N}(x|x_0, \sigma)$

BEKANNT:
N, x_0, σ

ZUFALLS-VARIABLE:
$v = \sum_{n=1}^{N} (x_n - \overline{x})^2$

$$\begin{aligned} p(v|N, x_0, \sigma, \mathcal{B}) &= \frac{(2\sigma^2)^{-\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})} \, v^{(\frac{N-1}{2}-1)} \, e^{-v/2\sigma^2} \\ &= p_\Gamma\left(v|\alpha = \frac{N-1}{2}, \beta = \frac{1}{2\sigma^2}\right) \quad . \end{aligned} \tag{21.15}$$

Im Zusammenhang mit der χ^2-Verteilung ist es üblich, auf normierte Zufalls-Variablen

$$\tilde{x}_n = \frac{x_n}{\sigma}$$

überzugehen. Die Stichproben-Varianz v geht dabei über in

$$z = \frac{v}{\sigma^2} \quad .$$

Die Stichproben-Verteilung der Varianz kann dann als χ^2-Verteilung Gl. (9.13a) mit $N-1$ Freiheitsgraden formuliert werden

$$p(z|N, x_0, \sigma, \mathcal{B}) = p_{\chi^2}(z|N-1) \quad . \tag{21.16}$$

Wir sind von N Freiheitsgraden x_n ausgegangen. Da in die Definition von v der Stichproben-Mittelwert eingeht, ging in Gl. (21.11) bei der Integration über den Stichproben-Mittelwert (ξ_N) ein Freigheitsgrad verloren. Das passiert nicht, wenn die x_n i.u.nv. Zufalls-Variablen mit bekanntem Mittelwert, z.B. $x_0 = 0$, sind.
Da die meisten Verteilungen, die wir bisher untersucht haben, im Grenzfall großer Stichproben gegen eine Normal-Verteilung konvergieren, bietet die χ^2-Verteilung eine weitverbreitete Möglichkeit, Signifikanz-Tests durchzuführen. Voraussetzung ist, dass wir unabhängige Zufalls-Variablen x_n vorliegen haben, die einer Normal-Verteilung $\mathcal{N}(x_n|x_{0,n}, \sigma_n)$ genügen. Mittelwert und Varianzen können hierbei von n abhängen. Angenommen, die zu untersuchende Hypothese impliziert die Werte für

$x_{0,n}$ und σ_n, dann können wir hieraus i.u.v. Zufalls-Variable $\tilde{x}_n$ konstruieren

$$\tilde{x}_n = \frac{x_n - x_{0,n}}{\sigma_n} \quad ,$$

die der normierten Normal-Verteilung $\mathcal{N}(\tilde{x}_n|0,1)$ gehorchen. Somit ist die Summe

$$\chi^2 = \sum_{n=1}^{N} \frac{(x_n - x_{0,n})^2}{\sigma_n^2} \tag{21.17}$$

χ^2-verteilt. Man nennt diesen Ausdruck auch χ^2-Statistik. Eine Statistik ist allgemein ein Funktional, das von den Stichproben-Elementen und weiteren durch Hypothesen festgelegten Parametern abhängt. Es ist wichtig zu unterscheiden, dass wir i.d.R sehr große Stichproben benötigen, damit die x_n normal-verteilt sind, N in Gl. (21.17) kann und wird in der Regel jedoch klein sein.

WAHRSCHEINLICHKEITSDICHTE DER χ^2-STATISTIK

GEGEBEN:

Stichprobe: $\{x_1, \ldots, x_N\}$ mit x_i aus $\mathcal{N}(x_i|x_0^i, \sigma_i)$

BEKANNT:

N, $x_0^1, \ldots, x_0^N$, $\sigma_1, \ldots, \sigma_N$

ZUFALLS-VARIABLE:

$\chi^2 = \sum_{n=1}^{N} \frac{(x_n - x_0^n)^2}{\sigma_n^2}$

$$\begin{aligned} p(\chi^2|N, x_{1,0}, \ldots, x_{N,0}, \sigma_1, \ldots, \sigma_N, \mathcal{B}) &= \frac{(2)^{-\frac{N}{2}}}{\Gamma(\frac{N}{2})} (\chi^2)^{(\frac{N}{2}-1)} e^{-\chi^2/2} \\ &= p_{\chi^2}(\chi^2|N) \quad . \end{aligned} \tag{21.18}$$

21.2.3 Beispiel für χ^2-Test

Ein Würfel werde $L \gg 1$ mal geworfen, dabei erscheine die Augenzahl i m_i mal. Wenn die Ausgangs-Hypothese besagt, dass der Würfel symmetrisch ist, dann sollte jede Augenzahl gleich-wahrscheinlich sein. Wie wir wissen, geht die Multinomial-Verteilung für $L \gg 1$ in eine Normal-Verteilung über, und die Voraussetzungen für die χ^2-Verteilung sind somit erfüllt. Wir werden dieses Beispiel später weiter verfolgen.

21.3 z-Statistik

Gegeben seien zwei Stichproben vom Umfang N_α ($\alpha = 1, 2$) sowie deren Varianzen σ_α. Als Statistik betrachten wir

$$\begin{aligned} z &= \frac{\overline{x}_2 - \overline{x}_1}{\sigma} \\ \sigma^2 &= \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} \qquad . \end{aligned} \tag{21.19}$$

Wir verwenden die Abkürzungen $\underline{N} = \{N_1, N_2\}$ und $\underline{\sigma} = \{\sigma_1, \sigma_2\}$. Die Wahrscheinlichkeitsdichte von z erhalten wir aus der Marginalisierungsregel

$$\begin{aligned} p(z|\underline{N}, \underline{\sigma}, \mathcal{B}) &= \int d\overline{x}_1 \, d\overline{x}_2 \, p(z|\overline{x}_1, \overline{x}_2, \underline{N}, \underline{\sigma}, \mathcal{B}) \, p(\overline{x}_1, \overline{x}_2|\underline{N}, \underline{\sigma}, \mathcal{B}) \\ &= \int d\overline{x}_1 \, d\overline{x}_2 \, \delta(z - \frac{\overline{x}_2 - \overline{x}_1}{\sigma}) \, p(\overline{x}_1, \overline{x}_2|\underline{N}, \underline{\sigma}, \mathcal{B}) \\ &= \sigma \int d\overline{x}_1 \, d\overline{x}_2 \, \delta(\overline{x}_2 - (\overline{x}_1 + z\sigma)) \, p(\overline{x}_1, \overline{x}_2|\underline{N}, \underline{\sigma}, \mathcal{B}) \\ &= \sigma \int d\overline{x}_1 \, p(\overline{x}_1, \overline{x}_2 = \overline{x}_1 + z\sigma|\underline{N}, \underline{\sigma}, \mathcal{B}) \\ &= \sigma \int dx_0 \int d\overline{x}_1 \, p(\overline{x}_1, \overline{x}_2 = \overline{x}_1 + z\sigma|\underline{N}, \underline{\sigma}, x_0, \mathcal{B}) \cdot \\ &\qquad \cdot p(x_0|\underline{N}, \underline{\sigma}, \mathcal{B}) \qquad . \end{aligned}$$

Im letzten Schritt haben wir noch über die Marginalisierungsregel den intrinsischen Mittelwert x_0 der Verteilung eingeführt. Da die Elemente der beiden Stichproben unabhängig voneinander sind, faktorisiert die Wahrscheinlichkeitsdichte der Mittelwer-

te Gl. (21.14)

$$
\begin{aligned}
p(z|\underline{N},\underline{\sigma},\mathcal{B}) &= \sigma \int dx_0\; p(x_0|\underline{N},\underline{\sigma},\mathcal{B}) \;\ldots \\
&\quad \ldots \int d\overline{x}_1\; p(\overline{x}_1|N_1,\sigma_1,x_0,\mathcal{B})\; p(\overline{x}_2=\overline{x}_1+z\sigma|N_2,\sigma_2,x_0,\mathcal{B}) \\
&= \sigma\; (2\pi)^{-1}\; \sqrt{\tfrac{N_1\,N_2}{\sigma_1^2\,\sigma_2^2}} \int dx_0\; p(x_0|\underline{N},\underline{\sigma},\mathcal{B}) \;\ldots \\
&\quad \ldots \int d\overline{x}_1\; e^{-\frac{1}{2}\left(\frac{(\overline{x}_1-x_0)^2}{\sigma_1^2/N_1}+\frac{(\overline{x}_1+z\sigma-x_0)^2}{\sigma_2^2/N_2}\right)} \\
&= \sigma\; (2\pi)^{-1}\; \sqrt{\tfrac{N_1\,N_2}{\sigma_1^2\,\sigma_2^2}} \underbrace{\int dx_0\; p(x_0|\underline{N},\underline{\sigma},\mathcal{B})}_{=1} \;\ldots \\
&\quad \ldots \sqrt{2\pi}\; e^{-\frac{z^2}{2}}\; \frac{1}{\sqrt{\frac{N_1}{\sigma_1^2}+\frac{N_2}{\sigma_2^2}}} \\
&= \sigma\; \frac{1}{\sqrt{2\pi}}\; e^{-\frac{z^2}{2}}\; \underbrace{\frac{1}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}}}_{\sigma} \qquad .
\end{aligned}
$$

WAHRSCHEINLICHKEITSDICHTE DER z-STATISTIK

GEGEBEN:

Stichprobe 1: $\{x_1^1,\ldots,x_{N_1}^1\}$ aus $\mathcal{N}(x|x_0,\sigma_1)$

Stichprobe 2: $\{x_1^2,\ldots,x_{N_2}^2\}$ aus $\mathcal{N}(x|x_0,\sigma_2)$

BEKANNT:

N_1, N_2, σ_1, σ_2

ZUFALLS-VARIABLE:

$z = \frac{\overline{x}_2-\overline{x}_1}{\sigma}$

mit $\sigma^2 = \frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}$

$$
\begin{aligned}
p_z(z) &= \frac{1}{\sqrt{2\pi}}\; e^{-z^2/2} \\
&= \mathcal{N}(z|0,1)
\end{aligned}
\tag{21.20}
$$

Zur Berechnung von z benötigen wir N_α und σ_α. Ansonsten gehen diese Größen nicht in die Wahrscheinlichkeitsdichte ein. Die z-Statistik ist so konstruiert, dass keine Information über den intrinsischen Mittelwert x_0 benötigt wird.

Die Größe

$$\sigma_{\overline{x}_\alpha} = \sqrt{\frac{\sigma_\alpha^2}{N_\alpha}} \qquad ,$$

die in die Berechnung der z-Statistik eingeht, stellt den Standard-Fehler des Stichproben-Mittelwertes $\overline{x}_\alpha$ dar. $\overline{x}_1 - \overline{x}_2$ ist die Differenz zweier normal-verteilter Größen mit den individuellen Standard-Fehlern $\sigma_{\overline{x}_\alpha}$. Die Fehler addieren sich zu

$$\sigma_{\overline{x}}^2 = \sigma_{\overline{x}_1}^2 + \sigma_{\overline{x}_2}^2 \qquad .$$

Die z-Statistik ist allgemein nichts anderes als

z-STATISTIK

$$z = \frac{\text{Stichproben-Mittelwert} - \text{wahrer Mittelwert}}{\text{Standard-Fehler}} = \frac{\overline{x} - x_0}{\text{SF}} \qquad . \tag{21.21}$$

Im obigen Fall ist der Stichproben Mittelwert $\overline{x}$, die Differenz zweier Stichproben-Mittelwerte $\overline{x} = \overline{x}_1 - \overline{x}_2$. Der wahre Mittelwert dieser Differenz ist Null und der Standardfehler ist durch $\sigma_{\overline{x}}$ gegeben. Man könnte auch den Spezialfall betrachten, dass nur eine Stichprobe vorliegt von i.u.v. Zufallszahlen der Verteilung $\mathcal{N}(x|x_0, \sigma)$. In diesem Fall ist $\overline{x}$ der Mittelwert der Stichprobe und $\sigma_{\overline{x}}^2 = \sigma^2/N$. Man kann diesen Fall auch aus den beiden Stichproben erhalten, wenn wir den Fehler der zweiten Stichprobe auf Null setzen

$$\sigma_2 = 0 \qquad .$$

Da beide Verteilungen denselben Mittelwert haben sollen, ist deren Differenz wieder Null $\overline{x} = 0$. Das ergibt

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sigma_1^2/N_1} \qquad .$$

Wir können nun $\overline{x}_2$ mit dem wahren Mittelwert $\overline{x}$ des Problems gleichsetzen.

21.4 Student-t Statistik

Wir wollen nun die Verteilung der sogenannten Student-t Statistik

$$t = \frac{(\overline{x} - x_0)}{\tilde{\text{SF}}} = \frac{\sqrt{N(N-1)}(\overline{x} - x_0)}{\sqrt{v}} \tag{21.22}$$

untersuchen, wobei wie zuvor $\overline{x}$ der Stichproben-Mittelwert und v die unnormierte Stichproben-Varianz darstellt. Außerdem sollen die Elemente der Stichprobe i.u.nv.

sein. Im Unterschied zu vorher soll nun der wahre Wert der Varianz nicht bekannt sein. Das ist der Grund, warum durch den Schätzwert der Standard-Abweichung ($\tilde{\text{SF}}$) geteilt wird. Wir bestimmen die gesuchte Wahrscheinlichkeitsdichte $p(t|N,x_0,\mathcal{B})$ wie zuvor über die Marginalisierungsregel. Allerdings benötigen wir nicht die individuellen Werte der Stichprobe, sondern lediglich $\overline{x}$ und v

$$\begin{aligned}
p(t|N,x_0,\mathcal{B}) &= \int d\sigma \int d\overline{x}\; dv\; p(t|\overline{x},v,\sigma,N,x_0,\mathcal{B})\; p(\overline{x},v,\sigma|N,x_0,\mathcal{B}) \\
&= \int d\sigma \int d\overline{x}\; dv\; \delta\left(t - \frac{\sqrt{N(N-1)}(\overline{x}-x_0)}{\sqrt{v}}\right)\ldots \\
&\quad \ldots p(\overline{x},v|\sigma,N,x_0,\mathcal{B})\; p(\sigma|N,x_0,\mathcal{B}) \\
&= \int d\sigma\; p(\sigma|N,x_0,\mathcal{B}) \int dv\;\; p(v|\sigma,N,x_0,\mathcal{B})\;\ldots \\
&\quad \ldots \int d\overline{x}\; \delta\left(\overline{x} - x_0 - t\frac{\sqrt{v}}{\sqrt{N(N-1)}}\right) \frac{\sqrt{v}}{\sqrt{N(N-1)}}\; p(\overline{x}|\sigma,N,x_0,\mathcal{B}) \\
&= \frac{1}{\sqrt{N(N-1)}} \int d\sigma\; p(\sigma|N,x_0,\mathcal{B}) \int dv\; \sqrt{v}\; p(v|\sigma,N,x_0,\mathcal{B})\;\ldots \\
&\quad \ldots\; p(\overline{x} = x_0 + t\frac{\sqrt{v}}{\sqrt{N(N-1)}}|\sigma,N,x_0,\mathcal{B}) \qquad .
\end{aligned}$$

Wir setzen Gl. (21.14) und Gl. (21.15) ein und erhalten

$$\begin{aligned}
p(t|N,x_0,\mathcal{B}) &= \frac{1}{\sqrt{N(N-1)}} \int d\sigma\; p(\sigma|N,x_0,\mathcal{B}) \int dv\; \sqrt{v}\; p(v|\sigma,N,x_0,\mathcal{B})\;\ldots \\
&\quad \ldots\; \frac{\sqrt{N}}{\sqrt{2\pi\sigma^2}}\; e^{-\frac{t^2 v}{2\frac{\sigma^2}{N}N(N-1)}} \\
&= \frac{1}{\sqrt{2\pi(N-1)}} \int d\sigma\; \sigma^{-1}\; p(\sigma|N,x_0,\mathcal{B})\ldots \\
&\quad \ldots \int dv\; \sqrt{v}\; \frac{(2\sigma^2)^{-\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})}\; v^{\frac{N-3}{2}}\; e^{-v\frac{1}{2\sigma^2}}\; e^{-v\frac{t^2}{2\sigma^2(N-1)}} \\
&= \frac{2^{-\frac{N-1}{2}}}{\sqrt{2\pi(N-1)}}\; \frac{1}{\Gamma(\frac{N-1}{2})} \int d\sigma\; \sigma^{-N}\; p(\sigma|N,x_0,\mathcal{B})\ldots \\
&\quad \ldots \int \frac{dv}{v}\; v^{\frac{N}{2}}\; e^{-v\left(\frac{1+t^2/(N-1)}{2\sigma^2}\right)} \qquad .
\end{aligned}$$

Im letzten Integral substituieren wir

$$z = v\left(\frac{1+t^2/(N-1)}{2\sigma^2}\right)$$

und erhalten für das Integral

$$\begin{aligned}
\int \frac{dv}{v}\; v^{\frac{N}{2}}\; e^{-v\left(\frac{1+t^2/(N-1)}{2\sigma^2}\right)} &= \left(\frac{1+t^2/(N-1)}{2\sigma^2}\right)^{-\frac{N}{2}} \int \frac{dz}{z}\; z^{\frac{N}{2}}\; e^{-z} \\
&= 2^{\frac{N}{2}}\; \sigma^N\; \left(1 + \frac{t^2}{N-1}\right)^{-\frac{N}{2}} \Gamma(\frac{N}{2}) \qquad .
\end{aligned}$$

Damit haben wir schließlich

$$p(t|N,x_0,\mathcal{B}) = \frac{1}{\sqrt{\pi(N-1)}}\,\frac{\Gamma(\frac{N}{2})}{\Gamma(\frac{N-1}{2})}\,\underbrace{\int d\sigma\; p(\sigma|N,x_0,\mathcal{B})}_{=1}\,\left(1+\frac{t^2}{N-1}\right)^{-\frac{N}{2}}$$
$$= \frac{1}{\sqrt{\pi(N-1)}}\,\frac{\Gamma(\frac{N}{2})}{\Gamma(\frac{N-1}{2})}\,\left(1+\frac{t^2}{N-1}\right)^{-\frac{N}{2}}$$

Das Endergebnis ist die Student-t Verteilung aus Gl. (9.26a).

WAHRSCHEINLICHKEITSDICHTE DER STUDENT-t-STATISTIK

GEGEBEN:
Stichprobe: $\{x_1,\ldots,x_N\}$ aus $\mathcal{N}(x|x_0,\sigma)$

BEKANNT:
N, x_0

ZUFALLS-VARIABLE:
$t = \frac{\sqrt{N(N-1)}(\overline{x}-x_0)}{\sqrt{v}}$
mit $v = \sum_{i=1}^{N}(x_i-\overline{x})^2$

$$p(t|N,x_0,\mathcal{B}) = \frac{1}{\sqrt{\pi(N-1)}}\,\frac{\Gamma(\frac{N}{2})}{\Gamma(\frac{N-1}{2})}\,\left(1+\frac{t^2}{N-1}\right)^{-\frac{N}{2}} \qquad (21.23)$$
$$= p_t(t|\nu = N-1)$$

Das bemerkenswerte an dem Ergebnis ist, dass keine Kenntnis über σ nötig war. Wir haben die Varianz zwar eingeführt, konnten sie am Ende aber wieder ausintegrieren, ohne $p(\sigma|N,x_0,\mathcal{B})$ angeben zu müssen. Darüber hinaus hängt die Student-t-Verteilung nicht vom wahren Mittelwert x_0 ab. Wie wir später sehen werden, verwendet man die Student-t-Statistik bei Signifikanz-Tests um zu überprüfen, ob x_0 der wahre Mittelwert einer Stichprobe ist oder ob die Mittelwerte zweier Stichproben übereinstimmen, wenn man die zugehörige Varianz nicht kennt.

21.5 Snedecors F-Statistik

Der t-Test wird u.a. verwendet, um auf der Basis der Stichproben-Mittelwerte zu entscheiden, ob die wahren Mittelwerte zweier Stichproben gleich sind. Damit kann man

überlegen, ob zwei Datensätze denselben physikalischen Ursprung haben. Die Stichproben mögen unterschiedliche Umfänge N_1 und N_2 besitzen. Alternativ kann man auch die Stichproben-Varianzen heranziehen. Die F-Statistik ist definiert als das Verhältnis zweier Stichproben-Varianzen v_1 und v_2

$$f = \frac{\tilde{\sigma_1}^2}{\tilde{\sigma_2}^2} = \frac{1}{r}\,\frac{v_1}{v_2} \tag{21.24}$$

$$r = \frac{N_1 - 1}{N_2 - 1} \quad . \tag{21.25}$$

Diese Statistik ist auch wieder so konstruiert, dass der wahre Wert der Varianz nicht benötigt wird und die Verteilung der F-Statistik ohne Kenntnis dieses Wertes angegeben werden kann. Das ermöglicht es, diese Größe für Signifikanz-Tests heranzuziehen. Außerdem wird nicht angenommen, dass die wahren Mittelwerte in beiden Fällen übereinstimmen. Die gesuchte Wahrscheinlichkeitsdichte folgt wieder aus der Marginalisierungsregel

$$\begin{aligned}
p(f|N_1,N_2,\mathcal{B}) &= \int d\sigma \int dv_1\, dv_2\; p(f|v_1,v_2,N_1,N_2,\mathcal{B})\; p(v_1,v_2,\sigma|N_1,N_2,\mathcal{B}) \\
&= \int d\sigma\; p(\sigma|N_1,N_2,\mathcal{B}) \int dv_1\, dv_2\; \delta(f - \tfrac{1}{r}\frac{v_1}{v_2})\cdot \\
&\qquad\qquad \cdot\, p(v_1,v_2|\sigma,N_1,N_2,\mathcal{B}) \\
&= \int d\sigma\; p(\sigma|N_1,N_2,\mathcal{B}) \int dv_2\; r v_2 \cdot \\
&\qquad\qquad \cdot\, p(v_1 = f r v_2, v_2|\sigma,N_1,N_2,\mathcal{B}) \qquad .
\end{aligned}$$

Nun sind die Stichproben-Varianzen unabhängig voneinander, da alle Zufalls-Variablen, die in die Summen eingehen, unabhängig sind

$$\begin{aligned}
p(f|N_1,N_2,\mathcal{B}) &= \int dv_2\; r v_2\; p(v_1 = f r v_2|N_1,\mathcal{B})\; p(v_2|N_2,\mathcal{B}) \\
&= C_1\, C_2 \int d\sigma\; p(\sigma|N_1,N_2,\mathcal{B})\; \sigma^{-(N_1+N_2-2)} \;\ldots \\
&\ldots \int dv_2\; r v_2\; (f r v_2)^{\frac{N_1-3}{2}}\; e^{-\frac{f r v_2}{2\sigma^2}}\; v_2^{\frac{N_2-3}{2}}\; e^{-\frac{v_2}{2\sigma^2}}
\end{aligned}$$

$$\text{mit} \qquad C_\alpha = \frac{2^{-\frac{N_\alpha-1}{2}}}{\Gamma(\frac{N_\alpha-1}{2})}$$

$$\begin{aligned}
p(f|N_1,N_2,\mathcal{B}) &= r\, C_1\, C_2 \int d\sigma\; p(\sigma|N_1,N_2,\mathcal{B})\; \sigma^{-(N_1+N_2-2)} \;\ldots \\
&\ldots\; (fr)^{\frac{N_1-3}{2}} \int \frac{dv_2}{v_2}\; (v_2)^{\frac{N_1+N_2-2}{2\sigma^2}}\; e^{-v_2\,\frac{1+fr}{2}} \qquad .
\end{aligned}$$

Wir substituieren $\frac{v_2}{\sigma^2} \rightarrow v$ und erhalten

$$p(f|N_1,N_2,\mathcal{B}) = r\ C_1\ C_2 \underbrace{\int d\sigma\ p(\sigma|N_1,N_2,\mathcal{B})}_{=1} (rf)^{\frac{N_1-3}{2}} \times \int_0^\infty \frac{dv}{v}\ v^{\frac{N_1+N_2-2}{2}}\ e^{-v\ \frac{1+fr}{2}} \quad .$$

Das verbleibende Integral kann wie im Fall der t-Verteilung leicht berechnet werden und liefert

$$\begin{aligned} p(f|N_1,N_2,\mathcal{B}) &= r\ C_1\ C_2\ (rf)^{\frac{N_1-3}{2}}\ 2^{\frac{N_1+N_2-2}{2}}\ (1+fr)^{-\frac{N_1+N_2-2}{2}}\ \Gamma(\tfrac{N_1+N_2-2}{2}) \\ &= \frac{\Gamma(\frac{N_1+N_2-2}{2})}{\Gamma(\frac{N_1-1}{2})\Gamma(\frac{N_2-1}{2})}\ r\ (rf)^{\frac{N_1-3}{2}}\ (1+fr)^{-\frac{N_1+N_2-2}{2}} . \end{aligned} \tag{21.26}$$

WAHRSCHEINLICHKEITSDICHTE DER F-STATISTIK

GEGEBEN:

Stichprobe 1: $\{x_1^1,\ldots,x_{N_1}^1\}$ aus $\mathcal{N}(x|x_0^1,\sigma)$

Stichprobe 2: $\{x_1^2,\ldots,x_{N_2}^2\}$ aus $\mathcal{N}(x|x_0^2,\sigma)$

BEKANNT:

N_1, N_2

ZUFALLS-VARIABLE:

$f = \frac{1}{r}\frac{v_1}{v_2}$
mit $v_\alpha = \sum_{i=1}^{N_\alpha}(x_i^\alpha - \overline{x}_\alpha)^2$
und $r = \frac{N_1-1}{N_2-1}$

$$p(f|N_1,N_2,\mathcal{B}) = \frac{\Gamma(\frac{N_1+N_2-2}{2})}{\Gamma(\frac{N_1-1}{2})\Gamma(\frac{N_2-1}{2})}\ r\ (rf)^{\frac{N_1-3}{2}}\ (1+fr)^{-\frac{N_1+N_2-2}{2}} \tag{21.27}$$

Qualitativ sehen die Kurven der $F-Verteilung$ wie die der χ^2-Verteilung aus. Der Mittelwert der Verteilung kann leicht durch direkte Integration berechnet werden und liefert

$$\langle f\rangle = \frac{N_2-1}{N_2-3} \quad .$$

Interessanterweise hängt das Ergebnis nicht von N_1 ab. Der Mittelwert ist immer größer Eins und nähert sich für $N_2 \gg 1$ dem Wert eins. Das ist allerdings nicht überraschend, denn die Stichproben-Varianzen sind χ^2-verteilt, wobei der Mittelwert mit dem wahren Wert übereinstimmt (unbiased). Der Mittelwert des Zählers liefert also immer den wahren Wert der Varianz. Nur die nichtlineare Transformation $1/v_2$ führt einen Bias ein, da

$$\left\langle \left(\frac{1}{v_2} \right) \right\rangle \neq \frac{1}{\langle v_2 \rangle} \qquad ,$$

außer für $N_2 \to \infty$. Die Verteilungsfunktion der F-Verteilung führt auf hypergeometrische Funktionen, die wir hier nicht weiter diskutieren wollen.

21.6 Fehler-Fortpflanzung

Wir betrachten eine lineare Funktion

$$g = a + \sum_n b_n \, x_n \tag{21.28}$$

von Zufalls-Variablen x_n die normal-verteilt sind, aber mit unterschiedlichen Mittelwerten x_n^0 und Varianzen σ_n^2. Diese Situation liegt häufig in der Praxis vor, wenn Größen aus unterschiedlichen Experimenten kombiniert werden. Den Mittelwert von g erhalten wir unmittelbar

$$\langle g \rangle = a + \sum_n b_n \, \langle x_n \rangle = a + \sum_n b_n \, x_n^0 \qquad .$$

Hier interessiert uns insbesondere der „Fehler" des Mittelwertes. Hierzu benötigen wir die Varianz

$$\begin{aligned} \mathrm{var}(g) = \left\langle (g - \langle g \rangle)^2 \right\rangle &= \left\langle \left(a + \sum_n b_n \, x_n - a - \sum_n b_n \, \langle x_n \rangle \right)^2 \right\rangle \\ &= \left\langle \left(\sum_n b_n \, (x_n - x_n^0) \right)^2 \right\rangle \\ &= \sum_{n,m} b_n \, b_m \, \langle \Delta_n \, \Delta_m \rangle \end{aligned}$$

mit $\qquad \Delta_n := (x_n - x_n^0) \qquad .$

Wenn die Zufalls-Variablen unkorreliert sind gilt

$$\langle \Delta_n \, \Delta_m \rangle = \delta_{n,m} \, \sigma_n^2 \qquad .$$

Somit gilt

FEHLER-FORTPFLANZUNG VON UNKORRELIERTEN FEHLERN

$$\sigma^2 = \mathrm{var}(g) = \sum_n b_n^2 \, \sigma_n^2 \qquad . \tag{21.29}$$

Wenn die Fehler korreliert sind und die Kovarianz-Matrix

$$C_{nm} = \langle \Delta_n \, \Delta_m \rangle$$

bekannt ist, lautet der Fehler

FEHLER-FORTPFLANZUNG VON KORRELIERTEN FEHLERN

$$\sigma^2 = \mathrm{var}(g) = b^T C b \qquad , \tag{21.30}$$

wobei b der Vektor der Koeffizienten b_n ist.

Wenn die Kovarianz nicht bekannt ist, kann man den Fehler über die Schwarzsche Ungleichung abschätzen.

SCHWARZSCHE UNGLEICHUNG DER KOVARIANZ-MATRIX

$$|\langle \Delta_n \, \Delta_m \rangle| \leq \sigma_n \, \sigma_m \qquad . \tag{21.31}$$

Zum Beweis gehen wir von der Zufalls-Variablen

$$x = \Delta_n - \frac{C_{nm}}{\sigma_m^2} \Delta_m$$

aus. Es gilt

$$\begin{aligned}
0 \leq \langle x^2 \rangle &= \left\langle \left(\Delta_n - \frac{C_{nm}}{\sigma_m^2} \Delta_m \right)^2 \right\rangle \\
&= \langle \Delta_n^2 \rangle - 2 \frac{C_{nm}}{\sigma_m^2} \langle \Delta_n \, \Delta_m \rangle + \frac{C_{nm}^2}{(\sigma_m^2)^2} \langle \Delta_m^2 \rangle \\
&= \sigma_n^2 - 2 \frac{C_{nm}^2}{\sigma_m^2} + \frac{C_{nm}^2}{\sigma_m^2} \\
&= \sigma_n^2 - \frac{C_{nm}^2}{\sigma_m^2} \\
C_{nm}^2 &\leq \sigma_n^2 \sigma_m^2 \qquad \text{q.e.d.}
\end{aligned}$$

Damit können wir den Fehler σ abschätzen

$$\sigma^2 = \sum_{nm} b_n \, b_m \, C_{nm} \leq \sum_{nm} |b_n| \; |b_m| \; |C_{nm}|$$
$$\leq \sum_{nm} |b_n| \; |b_m| \; \sigma_n \, \sigma_m \leq \left(\sum_{n} |b_n| \, \sigma_n \right)^2 \qquad .$$

FEHLER-ABSCHÄTZUNG BEI KORRELIERTEN FEHLERN

$$\sigma \leq \sum_{n} |b_n| \; \sigma_n \tag{21.32}$$

Bislang hatten wir nur lineare Funktionen der Zufalls-Variablen x_n betrachtet. Unter der Annahme, dass die Fehler klein sind, kann man eine beliebige Funktion $f(x)$ der Zufallsvariablen $x = \{x_1, x_2 \ldots, x_N\}$, deren erste Ableitung existiert, um den Mittelwert $x^0 = \{x_1^0, x_2^0 \ldots, x_N^0\}$ entwickeln

$$f(x) \simeq f(x^0) + \sum_{n} \left. \frac{\partial f(x)}{\partial x_n} \right|_{x=x^0} \Delta_n \qquad .$$

Damit liegt wieder ein lineares Modell wie in Gl. (21.28) mit den Koeffizienten

$$a = f(x^0) \; , \qquad b_n = \left. \frac{\partial f(x)}{\partial x_n} \right|_{x=x^0}$$

vor. Dann lässt sich mit Gl. (21.29), Gl. (21.30), oder Gl. (21.32) der Fehler $\sqrt{\mathrm{var}(f)}$ berechnen.

Kapitel 22

Orthodoxe Hypothesen Tests

Die Idee hinter den Signifikanz-Tests hatten wir bereits im Abschnitt 5.2.2 andiskutiert und im Kapitel über Stichproben-Verteilungen (21) die Wahrscheinlichkeitsdichten der für uns wichtigsten Statistiken besprochen.
Die Grundidee ist sehr einfach. Bevor wir sie allgemein formulieren, wollen wir sie an einem Beispiel illustrieren.

22.1 Einführung am Beispiel des z-Tests

Wir behaupten, dass eine Steuergröße S keinen Einfluss auf eine Messgröße D habe. Diese Behauptung nennt man NULL-HYPOTHESE. Um die Hypothese zu testen, führen wir eine Serie von Experimenten zum Wert der Steuergröße $S = S_1$ durch und ermitteln hierzu den Mittelwert $\overline{d}_1$ der Messwerte D. Anschließend wiederholen wir das Experiment zu $S = S_2$ und erhalten $\overline{d}_2$. Wir bilden hiervon die Differenz $\overline{d}_2 - \overline{d}_1$. Wenn die Hypothese stimmt, sollte die Differenz, abgesehen vom statistischen Fehler, null ergeben. Je größer die Differenz ist, desto unwahrscheinlicher ist es, dass die Hypothese stimmt. Nun hängt die mit dem statistischen Fehler kompatible Differenz vom Standardfehler ab. Deshalb ist es sinnvoll, folgende Statistik zu verwenden

$$z = \frac{\overline{d}_2 - \overline{d}_1}{\mathrm{SF}} \quad .$$

Der z-Test setzt voraus, dass der Standardfehler bekannt ist. Der gemessene Wert (Realisierung) sei z_0. Da nach dem zentralen Grenzwertsatz der Mittelwert von i.u.v. Zufalls-Variablen normal-verteilt ist, ist die Zufalls-Variable z normal-verteilt mit Mittelwert Null und Varianz Eins.
Man kann im Rahmen der orthodoxen Statistik nicht die Wahrscheinlichkeit angeben, dass eine Hypothese zutrifft. Man könnte nach der Wahrscheinlichkeit fragen, dass genau der beobachtete Wert z_0 auftritt, wenn die Null-Hypothese richtig ist. Bei Problemen mit kontinuierlichen Freiheitsgraden ist diese Wahrscheinlichkeit jedoch null. Aber auch bei diskreten Problemen macht die Wahrscheinlichkeit für z_0 bei Tests wenig Sinn.

Nehmen wir z.B. das Bernoulli-Problem. Eine Münze werde N-mal geworfen. Dabei erscheint n-mal „Kopf". Die Wahrscheinlichkeit hierfür ist unter der Null-Hypothese, dass die Münze symmetrisch ist,

$$P(n|N,H) = \binom{N}{n} 2^{-N} \quad .$$

Wir hatten im Zusammenhang mit dem „Gesetz der großen Zahlen" (Gl. (4.15)) gefunden, dass selbst für $n = N/2$, dem Wert der am verträglichsten mit der Null-Hypothese ist, die Wahrscheinlichkeit lediglich

$$P(N/2|N,H) = \frac{1}{\sqrt{\pi N/2}} \quad \underset{N\to\infty}{\longrightarrow} 0$$

beträgt und somit für große Stichproben immer kleiner wird. Um überhaupt Aussagen machen zu können, müssen wir über ein geeignet gewähltes Intervall I integrieren/summieren. Wir wählen das Intervall z.B. so

$$I = (-z^*, z^*) \quad ,$$

dass die Wahrscheinlichkeit

$$P(z_0 \in I) = 0.99$$

ist. Wenn die Null-Hypothese richtig ist, heißt das, dass in 99% der Fälle der experimentelle Wert z_0 in dieses Intervall fällt. Das bedeutet aber umgekehrt nicht, dass die Null-Hypothese die Wahrscheinlichkeit 0.99 hat, denn es könnte noch beliebig viele andere Hypothesen geben, die genauso gut oder noch besser mit den Daten verträglich sind. Hypothesen können auf diesem Weg nicht verifiziert werden. Wir können sie aber falsifizieren.
Wenn wir weiterhin von der Korrektheit der Null-Hypothese ausgehen, ist die Wahrscheinlichkeit, dass ein z_0 nicht in das angegebene Intervall fällt, lediglich 0.01. In diesem Fall können wir, unabhängig von alternativen Hypothesen, feststellen, dass die Null-Hypothese nicht mit den Daten verträglich ist, und wir können mit großer Zuversicht die Null-Hypothese verwerfen.

22.1.1 Indirekter Schluss

Bei den statistischen Signifikanz-Tests zieht man deshalb einen indirekten Schluss. Man geht davon aus, dass die Hypothese korrekt ist. Dann gibt man ein sogenanntes SIGNIFIKANZ-NIVEAU (Irrtumswahrscheinlichkeit) α vor, z.B. $\alpha = 0.01, 0.05, 0.1$ und legt damit den ZURÜCKWEISUNGSBEREICH (KRITISCHEN BEREICH) I_r^α fest, für den gilt

$$P(z \in I_r^\alpha | H) = \alpha.$$

Im vorliegenden Beispiel wählen wir z.B. $I_r^\alpha = (z_\alpha, \infty)$. Wenn der Messwert z_0 in den kritischen Bereich fällt, wird die Hypothese verworfen. Die Situation ist in Abb. 22.1

skizziert. Die kritische Region entspricht dem rechten schraffierten Bereich, dessen Fläche gerade α beträgt.
Für unser Beispiel bedeutet das, dass wir die Null-Hypothese, *die Steuergröße S habe keinen Einfluss auf die Messgröße* bei einem Signifikanz-Niveau von beispielsweise $\alpha = 0.01$ immer dann verwerfen werden, wenn $z_0 > 2.33$ ist.

TEST OB ZWEI STICHPROBEN DENSELBEN MITTELWERT HABEN BEI BEKANNTER VARIANZ: z-TEST

Die Größe

$$z = \frac{(\overline{d^{(2)}} - \overline{d^{(1)}})}{\mathrm{SF}} \tag{22.1}$$

genügt der z-Verteilung

$$p_z(z|\nu) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} = \mathcal{N}(z, 0, 1) \tag{22.2}$$

22.1.2 Fehler erster und zweiter Art

Die Wahrscheinlichkeit, dass bei Zutreffen der Null-Hyppothese Messwerte in die kritische Region fallen, ist per definitionem gleich dem Signifikanz-Niveau. Damit ist α gleichbedeutend mit der (Irrtums-)Wahrscheinlichkeit, die Hypothese zu verwerfen, obwohl sie richtig ist. Das hatten wir als FEHLER ERSTER ART bezeichnet. Man kann diesen Fehler natürlich beliebig klein machen, indem man $\alpha \to 0$ wählt. Dann wächst aber der FEHLER ZWEITER ART, dass die Hypothese akzeptiert wird, obwohl sie falsch ist. Diese Wahrscheinlichkeit lässt sich nun nicht im Rahmen der Hypothese H bestimmen. Der Fehler zweiter Art besagt, eine Alternative H_1 ist richtig, die wir nun irrtümlich verwerfen. Die zugehörige Wahrscheinlichkeit ist

$$P(z_0 \notin I_r^\alpha | H_1) = 1 - P(z_0 \in I_r^\alpha | H_1) \quad .$$

Die kritische Region ist durch das Signifikanz-Niveau nicht eindeutig festgelegt. Es besagt nur, dass im Zurückweisungsbereich die Wahrscheinlichkeitsmasse α liegt. Neben den eben behandelten EINSEITIGEN TESTS sind auch ZWEISEITIGE TESTS weit verbreitet. Im zweiseitigen Test schaut man sich die Wahrscheinlichkeit für Abweichungen in beide Richtungen vom hypothetischen Wert (in unserem Beispiel null) an. Die Situation ist in Abb. 22.1 dargestellt. Der kritische Bereich ist im zweiseitigen Test.

$$I_r^\alpha = (-\infty, z_{\alpha/2}^u] \cup [z_{\alpha/2}^o, \infty) \quad ,$$

wobei die Cutoffs $z_{\alpha/2}^{u/o}$ so bestimmt sind, dass die Fläche in beiden schraffierten Bereichen jeweils $\alpha/2$ beträgt.

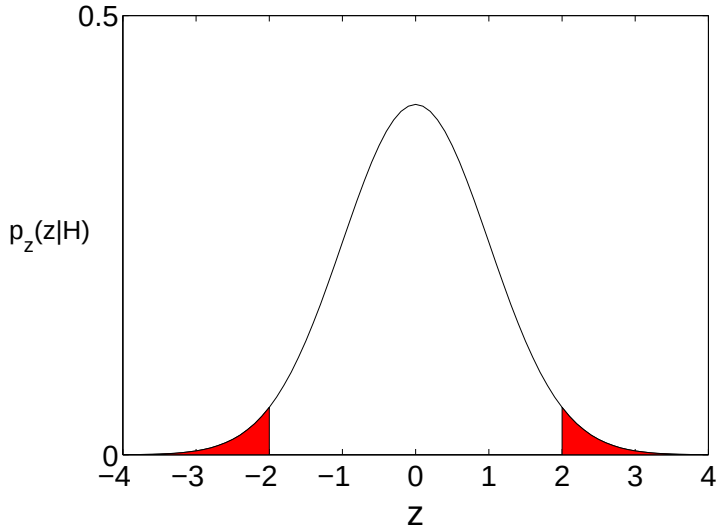

Abbildung 22.1: *Wahrscheinlichkeitsdichte des z-Tests mit den Bereichen, die einer Irrtumswahrscheinlichkeit von* 5% *entsprechen.*

Im Falle des z-Tests, der eine symmetrische Wahrscheinlichkeitsdichte $p_z(z)$ besitzt, gilt $z_\alpha^u = -z_\alpha^o = -z_\alpha$. Bei einem Signifikanz-Niveau von $\alpha = 0.05$ ist bei einem einseitigen/zweiseitigen (es/zs) Test

$$\begin{aligned} z_{5\%}^{\text{es}} &= 1.65 \\ z_{5\%}^{\text{zs}} &= 1.96 \end{aligned} \quad .$$

Entsprechend gilt für das Signifikanz-Niveau $\alpha = 0.01$

$$\begin{aligned} z_{1\%}^{\text{es}} &= 2.33 \\ z_{1\%}^{\text{zs}} &= 2.58 \end{aligned} \quad .$$

Wir wollen die kritische Region anhand des Beispiels der z-Statistik noch etwas genauer untersuchen. Die Null-Hypothese besagt, dass die Zufalls-Variable z um den Mittelwert $\mu_0 = 0$ normal-verteilt ist, da die Steuergröße keinen Einfluss auf die Messgröße haben soll. Wir geben ein Signifikanz-Niveau $\alpha = 0.1$ vor. Der gemessene z-Wert sei z_0. Den Bereich, komplementär zu I_r, bezeichnen wir mit $I_{\overline{r}}$. Falls $z_0 \in I_{\overline{r}}$ kann die Hypothese nicht verworfen werden. Wir untersuchen vier unterschiedliche Zurückweisungsbereiche (ZB), die in der folgenden Tabelle angegeben und in Ab-

bildung 22.2 schraffiert dargestellt sind.

ZB	I_r	$I_{\bar{r}}$
1	$[z_\alpha, \infty)$	$(-\infty, z_\alpha)$
2	$(-\infty, -z_\alpha]$	$(-z_\alpha, \infty,)$
3	$(-\infty, -z_{\frac{\alpha}{2}}] \cup [z_{\frac{\alpha}{2}}, \infty)$	$(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}})$
4	$(-z^*_\alpha, z^*_\alpha)$	$(-\infty, -z^*_\alpha] \cup [z^*_\alpha, \infty)$

Der Cutoff z_β ist über die Gleichung

$$\int_{z_\beta}^{\infty} p_z(z)\, dz = \beta$$

definiert. Wir benötigen dieses Integral für die Werte $\beta = \alpha$ und $\beta = \alpha/2$. Der Cutoff z^*_α folgt aus

$$\int_{-z^*_\alpha}^{z^*_\alpha} p_z(z)\, dz = \alpha \qquad .$$

Die Wahrscheinlichkeit, einen Fehler erster Art zu machen, ist in allen Fällen α. In den Rechnungen zur Abbildung 22.2 wurde $\alpha = 0.1$ verwendet. Die Fälle unterscheiden sich jedoch in der Wahrscheinlichkeit für einen Fehler zweiter Art $P(F_2|\sigma = 1, \mu, \mu_0 = 0, ZB, \mathcal{B})$. Diese Wahrscheinlichkeit ist

$$P(F_2|\sigma = 1, \mu, \mu_0 = 0, ZB, \mathcal{B}) =$$
$$\int dz_0\; P(F_2|z_0, \sigma = 1, \mu, \mu_0 = 0, ZB, \mathcal{B})\; \underbrace{p(z_0|\sigma = 1, \mu, \mu_0 = 0, ZB, \mathcal{B})}_{= p(z_0|\sigma=1,\mu,\mathcal{B})}.$$

Der erste Faktor hängt vom Signifikanz-Bereich ab. Man macht einen Fehler zweiter Art, wenn man die Null-Hypothese akzeptiert, obwohl sie falsch ist. Bis auf den Fall $\mu = 0$ ist die Null-Hypothese in unserem Beispiel immer falsch. Das heißt, immer dann, wenn wir die Hypothese trotzdem akzeptieren, begehen wir einen Fehler zweiter Art. Wir akzeptieren die Null-Hypothese, wenn

$$P(F_2|z_0, \sigma = 1, \mu, \mu_0 = 0, ZB, \mathcal{B}) = \theta(z_0 \in I^{ZB}_{\bar{r}}) \quad ZB = 1, 2, 3, 4 \qquad .$$

Damit haben wir also

$$P(F_2|\sigma = 1, \mu, \mu_0 = 0, ZB, \mathcal{B}) = \frac{1}{\sqrt{2\pi}} \int_{I^{ZB}_{\bar{r}}} dz_0\; e^{-\frac{(z_0-\mu)^2}{2}}$$
$$= P(z \in I^{ZB}_{\bar{r}}|\sigma = 1, \mu, \mu_0 = 0, ZB, \mathcal{B}) \qquad .$$

In Abbildung 22.2 ist die Wahrscheinlichkeit für einen Fehler zweiter Art als Funktion der wahren Differenz μ aufgetragen. Man nennt einen Test unverzerrt, wenn

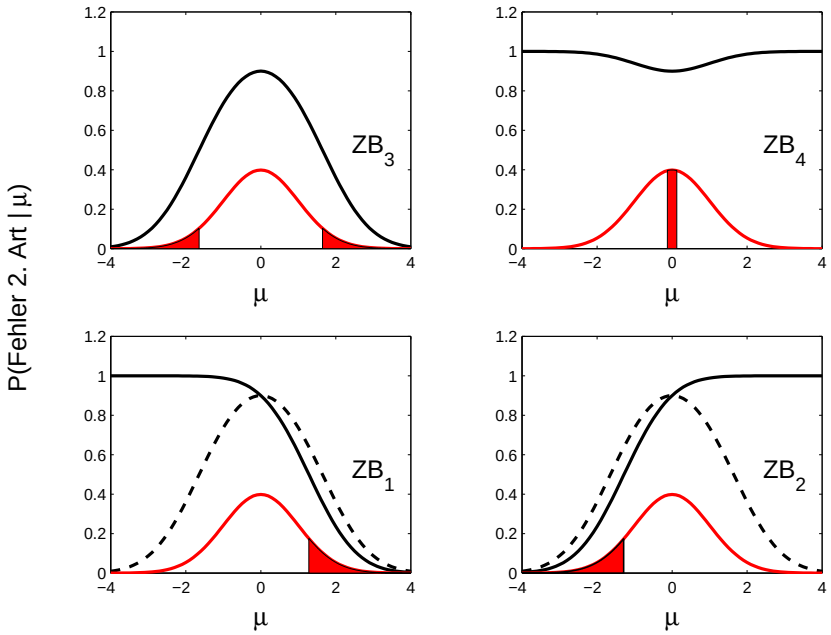

Abbildung 22.2: *Vergleich der vier im Text beschriebenen kritischen Regionen. Im Bild oben links befindet sich der übliche zweiseitige Test. Im unteren Teil sind die beiden einseitigen Tests dargestellt und mit dem Ergebnis des zweiseitigen Tests (gestrichelte Kurve) verglichen. Im Bild rechts oben ist das Ergebnis zu ZB_4 dargestellt.*

$$P(z_0 \in I_r|\mu) \geq \alpha = P(z_0 \in I_r|\mu = \mu_0) \qquad .$$

Diese Definition ist sinnvoll, weil die Wahrscheinlichkeit, die Null-Hypothese zu verwerfen dann am kleinsten ist, wenn die Null-Hypothese wahr ist.
In Abbildung 22.2 sind die Ergebnisse zu den vier kritischen Regionen dargestellt. Der zweiseitige Test ($ZB = 3$) verhält sich qualitativ, wie man es erwartet. Die Wahrscheinlichkeit, einen Fehler zweiter Art zu machen sinkt monoton mit zunehmender Abweichung des wahren Wertes μ vom hypothetischen Wert $\mu_0 = 0$. Je größer diese Diskrepanz, desto einfacher ist es zu erkennen, dass die Null-Hypothese falsch ist. Dementsprechend wird sie in zunehmendem Maße verworfen. Für $\mu \to 0$ hingegen werden immer häufiger Stichproben vorkommen, die mit dem hypothetischen Wert verträglich sind. Die Null-Hypothese wird also akzeptiert, obwohl sie falsch ist. Einen Sonderfall stellt $\mu = 0$ dar, für den die Null-Hypothese korrekt ist. Entsprechend ist die Wahrscheinlichkeit für $\mu = 0$

$$P(z \in I_{\bar{r}}|\sigma, \mu_0 = \mu, ZB, \mathcal{B}) = 1 - P(z \in I_r|\sigma, \mu_0 = \mu, ZB, \mathcal{B}) = 1 - \alpha \qquad .$$

Dieser Wert wird in der Abbildung angenommen. In diesem Ausnahmefall liegt kein Fehler zweiter Art sondern eine korrekte Entscheidung vor.

Der einseitige Test mit Cutoff rechts ($ZB = 1$) ist sogar noch diskriminierender, wenn die wahre Differenz μ größer Null ist. Allerdings ist der Test für $\mu < 0$ VERZERRT. Die Irrtumswahrscheinlichkeit zweiter Art geht sogar sehr schnell gegen eins. Genau umgekehrt verhält sich der andere einseitige Test ($ZB = 2$). Der Test mit der kritischen Region ($ZB = 4$) im Maximum der Wahrscheinlichkeitsdichte ist, wie nicht anders zu erwarten, völlig unbrauchbar. Er ist überall verzerrt und liefert gerade dann große Fehler zweiter Art, wenn große Diskrepanzen zwischen dem wahren Wert und dem hypothetischen vorliegen.

Wenn $z_0 \in I_r^\alpha$, so sagt man, dass die Null-Hypothese verworfen werden muss, da die Daten unter dieser zu unwahrscheinlich sind. Das Umgekehrte gilt nicht. Wenn die Hypothese nicht verworfen werden kann, heißt das noch nicht, dass sie richtig ist. Denn es kann sehr wohl andere Hypothesen geben, die die Daten genauso gut oder besser erklären können. Mit den statistischen Tests kann man nur eine Theorie falsifizieren. Denn wenn sie die Daten nicht erklärt, tut sie das unabhängig von den anderen Hypothesen.

22.1.3 Allgemeines Prinzip der statistischen Tests

Wir haben am Beispiel des z-Tests das Prinzip des Signifikanz-Tests eingeführt und wollen nun die allgemeine Vorgehensweise zusammenfassen. Man unterscheidet zwischen EINFACHEN und ZUSAMMENGESETZTEN Tests. Bei einfachen Tests legt die Null-Hypothese alle Parameter fest. Bei zusammengesetzten Tests hingegen, gibt es noch unbestimmte Parameter. Z.B. könnte die Null-Hypothese besagen: *Die Spannung ist proportional zur Stromstärke*. Das heißt, es soll gelten

$$U \propto I \qquad .$$

Die Proportionalitätskonstante (hier der Widerstand) ist ein freier Parameter. Freie Parameter werden über eine Maximum-Likelihood-Schätzung festgelegt. Die Zahl der Freiheitsgrade reduziert sich hierbei um die Zahl der freien Parameter, da diese aus der Stichprobe ermittelt werden.

GENERELLE VORGEHENSWEISE BEI SIGNIFIKANZ TESTS

1. Formulierung der Null-Hypothese.
2. Wahl der Test-Statistik (x).
3. Bestimmung der freien Parameter aus dem ML-Prinzip.
4. Festlegen des Signifikanz-Niveaus (z.B. $\alpha = 0.01$).
5. Ermittlung des kritischen Bereiches I_r^α.
6. Auswertung der Stichprobe $\Rightarrow x_0$.
7. Feststellen, ob Daten signifikant $x_0 \in I_r^\alpha$
   und ob die Null-Hypothese verworfen werden muss.

Wenn ein Test dazu führt, dass die Null-Hypothese verworfen werden muss, dann ist das losgelöst von anderen Alternativen. Die Hypothese erklärt einfach die Daten nicht. Es wird dann sicher eine andere Hypothese geben, die mit den Daten konform ist.
Anders ist das, wenn die Null-Hypothese aufgrund der Daten nicht verworfen werden kann. Selbst wenn die Daten noch so gut passen, heißt das nicht, dass die Hypothese richtig sein muss.
Hypothesen können in diesem Sinne nur falsifiziert und niemals verifiziert werden.
Statt den kritischen Bereich vorab zu bestimmen, kann man den sogenannten P-Wert berechnen. Das ist bei einseitigen Tests die Wahrscheinlichkeit

$$P = P(x \geq x_0) \qquad ,$$

dass x-Werte so groß wie oder größer als der beobachtete Wert vorkommen, wenn die Null-Hypothese korrekt ist. Ist der P-Wert kleiner als α, so wird die Hypothese verworfen, da sie im Rückweisungsbereich (kritische Region) liegt.
Man sagt dann, x_0 weicht signifikant vom hypothetischen Wert ab. Wenn die Daten SIGNIFIKANT sind, versteht man darunter, dass die beobachteten Abweichungen vom hypothetischen Wert so groß sind, dass man sie nicht mit statistischen Fluktuationen erklären kann. Das Signifikanz-Niveau gibt den Schwellwert an, ab dem Diskrepanzen vom hypothetischen Wert als signifikant und nicht als zufällig betrachtet werden.

22.2 χ^2-Test

Man geht beim χ^2-Tests davon aus ($\mathcal{B}$), dass die Stichprobe $\{x_1, \ldots, x_N\}$ aus N Zufalls-Variablen x_i besteht, die alle unabhängig voneinander normal-verteilt sind.

Mittelwert μ_i und Varianz σ_i^2 der x_i sind bekannt, können aber alle unterschiedlich sein. Die χ^2-Statistik x lautet dann

$$x = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \quad .$$

Die Statistik x besitzt die Wahrscheinlichkeitsdichte der χ^2-Verteilung

$$p_{\chi^2}(x|N) = \frac{2^{-\frac{N}{2}}}{\Gamma(\frac{N}{2})}\, x^{\frac{N}{2}-1}\, e^{-\frac{x}{2}} \quad ,$$

mit N Freiheitsgraden. Der Erwartungswert $\langle x \rangle$ der χ^2-Statistik ist N. Das bedeutet, dass im Mittel jeder Datenpunkt eine Standard-Abweichung vom wahren Wert entfernt sein wird. Beim einseitigen Test ist die kritische Region

$$I_r^\alpha = [x_\alpha, \infty)$$

mit $$\alpha \stackrel{!}{=} P(x \geq x_\alpha|N) = 1 - F_{\chi^2}(x_\alpha|N)$$

bzw. $$x_\alpha = F_{\chi^2}^{-1}(1 - \alpha|N) \quad .$$

N	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
10	15.987	18.307	23.209
20	28.412	31.410	37.566
30	40.256	43.773	50.892
40	51.805	55.758	63.691
50	63.167	67.505	76.154
60	74.397	79.082	88.379
70	85.527	90.531	100.425
80	96.578	101.879	112.329
90	107.565	113.145	124.116
100	118.498	124.342	135.807

Tabelle 22.1: *Grenzwerte x_α in Abhängigkeit von α und N.*

χ^2-TEST

Die Größe

$$x = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{22.3}$$

genügt der χ^2-Verteilung

$$p_{\chi^2}(x|N) = \frac{2^{-\frac{N}{2}}}{\Gamma(\frac{N}{2})}\, x^{\frac{N}{2}-1}\, e^{-\frac{x}{2}} \tag{22.4}$$

mit N Freiheitsgraden.

22.2.1 Vergleich mit theoretischen Modellen

Wir betrachten ein Situation, die wir in der Physik häufig antreffen. Zu gewissen Steuergrößen $s_i \in \mathbb{R}$ $(i = 1, \ldots, N)$ (z.B. angelegte Spannung) werden Experimente durchgeführt, die Werte d_i für eine Messgröße (z.B. Strom) liefern. Es existiere ein theoretisches Modell $y(s|a)$, das für die Steuergröße s_i den Messwert $y(s_i|a)$ vorhersagt. Das Modell hängt von Parametern $a \in \mathbb{R}^{N_p}$ ab, die bekannt sein sollen. Wenn das Modell stimmt, das ist die Null-Hypothese, dann sollten die experimentellen Werte nur aufgrund von Messfehlern von den theoretischen Werten abweichen

$$d_i = y(s_i|a) + \eta_i \quad .$$

Die Abweichungen η_i seien normalverteilt mit bekannter Varianz σ_i^2, die von der Steuergröße s_i, oder dem exakten Wert $y(s_i|a)$ abhängen kann. Die χ^2-Statistik ist dann

$$x_0 = \sum_{i=1}^{N} \frac{(d_i - y(s_i|a))^2}{\sigma_i^2} \quad .$$

Diese Größe ist χ^2-verteilt zu N Freiheitsgraden.

Reduzierte Zahl von Freiheitsgraden

Wir betrachten dieselbe Situation wie zuvor, nur dass nun die Modell-Parameter nicht oder nur teilweise bekannt sein sollen. Die Zahl der unbekannten Parameter sei r. Man bestimmt diese Parameter aus dem Maximum-Likelihood-Prinzip (siehe Abschnitt 20.2)

$$a^{\mathrm{ML}} = a^{\mathrm{ML}}(d) \quad .$$

Der Schätzwert hängt nur von der Stichprobe und bekannten Größen ab. Der Wert der χ^2-Statistik

$$x_0 = \sum_{i=1}^{N} \frac{(d_i - y(s_i|a^{\mathrm{ML}}(d))^2}{\sigma_i^2}$$

ist somit tatsächlich eine Statistik. Man kann auch zeigen, dass sie χ^2-verteilt ist. Allerdings hat sich die Zahl der Freiheitsgrade verringert, da aufgrund der Abhängigkeit der ML-Parameter von der Stichprobe nicht mehr alle Summanden voneinander unabhängig sind. Die Zahl der Freiheitsgrade ist nun $N-r$. Die Hypothese wird wieder verworfen, wenn

$$1 - F_{\chi^2}(x_0|N-r) < \alpha$$

bzw.

$$F_{\chi^2}(x_0|N-r) > 1-\alpha \qquad .$$

Natürlich ist ein Wert $x = 0$ ebenfalls extrem unwahrscheinlich. Man wird deshalb den zweiseitigen Test verwenden, der Abweichungen vom Mittelwert $\langle x \rangle$ in beide Richtungen untersucht.

22.2.2 Test von Verteilungen

Häufig stellt sich die Frage, ob gemessen Daten einer bestimmten Verteilung genügen. Z.B. ob beobachtete radioaktive Zerfälle poisson-verteilt sind.
Die zu testende Verteilung kann diskret oder kontinuierlich sein. In letzterem Fall wird man das Problem diskretisieren. Das soll an einem eindimensionalen Problem erläutert werden. Die Zufalls-Variable x nehme alle Werte aus dem Intervall $x \in [0, \infty)$ an. Die zugehörige Wahrscheinlichkeitsdichte sei $\rho(x)$. Wir diskretisieren die x-Achse

$$0 = x_0 < x_1 < x_2 < \ldots < \infty \qquad .$$

Die Unterteilung wird häufig äquidistant sein, muss es aber nicht. Z.B. kann es bei einer Exponential-Verteilung sinnvoller sein, die Intervall-Breite exponentiell anwachsen zu lassen.
Wir definieren zu diesen Intervallen die Wahrscheinlichkeiten

$$p_i = P(x \in [x_i, x_{i+1})|\mathcal{B}) = \int_{x_i}^{x_{i+1}} \rho(x)\, dx, \qquad i = 0, 1, \ldots \qquad .$$

Nach der Diskretisierung sind kontinuierliche Probleme wie diskrete zu behandeln. Wenn eine Stichprobe vom Umfang N vorliegt, erwartet man, im Intervall I_i im Mittel

$$\tilde{n}_i = N p_i$$

Ereignisse (Teilchen) anzutreffen. Der Standard-Fehler dieses Mittelwertes ist

$$\mathrm{SF} = \sqrt{N p_i (1-p_i)} \qquad .$$

Somit ist die χ^2-Statistik

$$x_0 = \sum_{i=1}^{N} \left(\frac{n_i - \tilde{n}_i}{\sqrt{\tilde{n}_i \, (1 - p_i)}} \right)^2 = \sum_{i=1}^{N} \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i \, (1 - p_i)} \qquad . \tag{22.5}$$

Nun sind die Beiträge zur Summe nicht alle unabhängig, da $\sum_i \tilde{n}_i = N$. Man verwendet deshalb

$$x_0 = \sum_{i=1}^{N} \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i} \qquad . \tag{22.6}$$

Diese Zufalls-Variable ist wieder χ^2-verteilt mit $N-1$ Freiheitsgraden, da einer wegen der Nebenbedingung verlorengegangen ist.

Beispiel: Münze

Das einfachste Beispiel für den χ^2-Test ist das Münz-Experiment. Eine Münze wird N-mal geworfen. Die Null-Hypothese lautet: die Münze ist symmetrisch, das heißt $P_1 = P_2 = 1/2$. In der Stichprobe komme n_1-mal Kopf und n_2-mal Zahl vor. Die theoretischen Werte sind $n_1^* = n_2^* = N/2$. Der Wert der χ^2-Statistik ist

$$x_0 = \frac{(n_1 - N/2)^2}{N/2} + \frac{(n_2 - N/2)^2}{N/2} \qquad .$$

Nun ist aber $n_1 + n_2 = N$, bzw. $n_2 = N - n_1$. Das heißt,

$$x_0 = \frac{(n_1 - N/2)^2}{N/2} + \frac{(N - n_1 - N/2)^2}{N/2} = \frac{2(n_1 - N/2)^2}{N/2} = \frac{(2n_1 - N)^2}{N} \qquad .$$

Die Zahl der Freiheitsgrade ist, da kein freier Parameter vorkommt,

$$\nu = 2 - 1 = 1 \qquad .$$

Die zugehörigen Grenzwerte sind

$$\begin{aligned} x_{5\%} &= 3.84 \\ x_{1\%} &= 6.63 \qquad . \end{aligned}$$

Bei $N = 100$ heißt das, dass bei 5% Signifikanz-Schwelle akzeptable Werte für n_1 folgende Bedingung erfüllen müssen

$$\begin{aligned} \tfrac{1}{2} \left(N - \sqrt{3.84 \, N} \right) \leq n_1 \leq \tfrac{1}{2} \left(N + \sqrt{3.84 \, N} \right) \\ 41 \leq n_1 \leq 59 \qquad . \end{aligned}$$

Bei Werten außerhalb dieses Intervalls sind die Daten mit einem Niveau von $\alpha = 5\%$ signifikant. Aus Symmetriegründen gilt dasselbe für n_2.

Einschub: Sufficient Statistics

Wir haben gerade festgestellt, dass es bei binären Problemen ausreicht, eine der beiden Zufalls-Variablen (z.B. n_1) in der Statistik zu berücksichtigen. Das gilt generell, da die zweite Zufalls-Variable über $n_2 = N - n_1$[1] festgelegt ist. Es stellt sich die Frage, ob wir anstelle der χ^2-Statistik eine andere hätten verwenden können. Wir gehen von der allgemeinsten Form einer alternativen Statistik

$$y = f(n_1, N, P_1) \tag{22.7}$$

aus, die von der Zufalls-Variablen n_1, dem Stichprobenumfang und der a-priori-Wahrscheinlichkeit P_1 abhängt. Eine Möglichkeit ist die χ^2-Statistik mit

$$\begin{aligned} f &= \frac{(n_1 - NP_1)^2}{NP_1} + \frac{(n_2 - NP_2)^2}{NP_2} \\ &= \frac{(n_1 - NP_1)^2}{NP_1} + \frac{(N - n_1 - N(1 - P_1))^2}{NP_2} \\ &= \frac{(n_1 - NP_1)^2}{NP_1} + \frac{(n_1 - NP_1)^2}{NP_2} \\ &= \frac{(n_1 - NP_1)^2 (P_1 + P_2)}{NP_1P_2} \\ f &= \frac{(n_1 - NP_1)^2}{NP_1P_2} \quad . \end{aligned} \tag{22.8}$$

Wir führen für die y-Statistik eine für alle f gleiche Signifikanz-Schwelle $P(y \geq y_s) = P$ ein. Nun lösen wir Gl. (22.7) nach n_1 auf. Für die folgenden Überlegungen gehen wir davon aus, dass diese Lösung eindeutig ist. Über Gl. (22.7) korrespondiert dann zu jedem n_1 eindeutig ein Wert y. Es soll insbesondere n_s zu y_s korrespondieren

$$y_s = f(n_s, N, P_1) \quad . \tag{22.9}$$

Dann gilt offenbar

$$P(n_1) = P(y) \quad .$$

Daraus folgt unmittelbar durch Summieren

$$P(y \geq y_s) = P(n_1 \geq n_s) = \sum_{n_1 = n_s}^{\infty} P(n_1) \quad .$$

Die rechte Seite ist natürlich unabhängig von der Wahl der Statistik f und somit muss auch die linke Seite unabhängig sein. Das heißt, die Irrtumswahrscheinlichkeit $P(y \geq y_s)$ ist unabhängig von f wenn nur y_s über Gl. (22.9) ermittelt wurde und die Statistik f eine eineindeutig Funktion von n_1 ist. Das heißt, der einzige Test, der im binären Problem wirklich zählt, ist die Wahrscheinlichkeit

$$P(n_1 \geq n_s) \quad .$$

[1] N ist keine Zufalls-Variable!

Wie passt die χ^2-Statistik ins Bild? Mit f aus Gl. (22.8) erhalten wir bei der Umkehrung

$$n_1 = NP_1 \pm \sqrt{NP_1P_2\, y} \qquad .$$

Diese Lösung verletzt die Annahme, dass die Umkehrung eindeutig ist. Der Bereiche $y \geq y_s$ korrespondiert, wegen des quadratischen Zusammenhangs

$$y = \frac{(n_1 - NP_1)^2}{NP_1P_2} \qquad ,$$

zu den beiden Bereichen

$$n_1 \leq \max(0, NP_1 - \sqrt{NP_1P_2\, y_s})$$
$$\text{und}$$
$$n_1 \geq NP_1 + \sqrt{NP_1P_2\, y_s} \qquad .$$

Wir definieren $n_s = NP_1 + \sqrt{NP_1P_2\, y_s}$. Somit gilt

$$P(\chi^2 \geq \chi^2_s) = P(n_1 \geq n_s) + P(n_1 \leq \max(0,\ 2NP_1 - n_s)) \qquad .$$

Der χ^2-Test ist in diesem Fall äquivalent zu dem beidseitigen Test, dass der beobachtete Wert für n_1 unwahrscheinlich weit vom Mittelwert NP_1 entfernt ist. Hierbei werden die Abweichungen in beide Richtungen berücksichtigt.
Oder besser umgekehrt. Ausgehend vom sinnvollerweise beidseitigen Test in n_1, erhält man den einseitigen χ^2-Test. Das ist aber ein Spezialfall des binären Problems.
Wir haben gesehen, dass für binäre Probleme alle Tests äquivalent sind. Man nennt die zugehörige Statistik SUFFICIENT STATISTICS.

Beispiel: Mustererkennung

Wir betrachten ein weiteres Anwendungsgebiet der χ^2-Statistik. Gegeben sei ein Bild aus L Pixel. Es soll entschieden werden, ob sich auf dem Bild ein Objekt befindet (z.B: Tumor, extra-terrestrische Objekte (schwache Sterne)) oder ob nur ein verrauschter Untergrund vorliegt. Das Bild liege in Form von Photonen-Zählraten

$$n_i\,(i = 1, \ldots, L)$$

der einzelnen Pixel vor. Die Null-Hypothese lautet, *es liegt kein Objekt vor*[2]. Dann ist

$$P_i = \frac{1}{L} \qquad ,$$

[2]Eine aussagekräftigere Hypothese könnte auch sein, *Es liegt ein ganz bestimmtes, bekanntes Objekt vor*

und man würde nach dieser Hypothese im Mittel $\frac{N}{L}$ Photonen pro Pixel erwarten, wobei N die Gesamtzahl der Photonen im Bild ist. Der Wert der χ^2-Statistik ist somit

$$x_0 = \sum_{l=1}^{L} \frac{(n_l - \frac{N}{L})^2}{\frac{N}{L}} \quad .$$

Das soll an einem Beispiel demonstriert werden. Wir erzeugen hierzu künstliche Da-

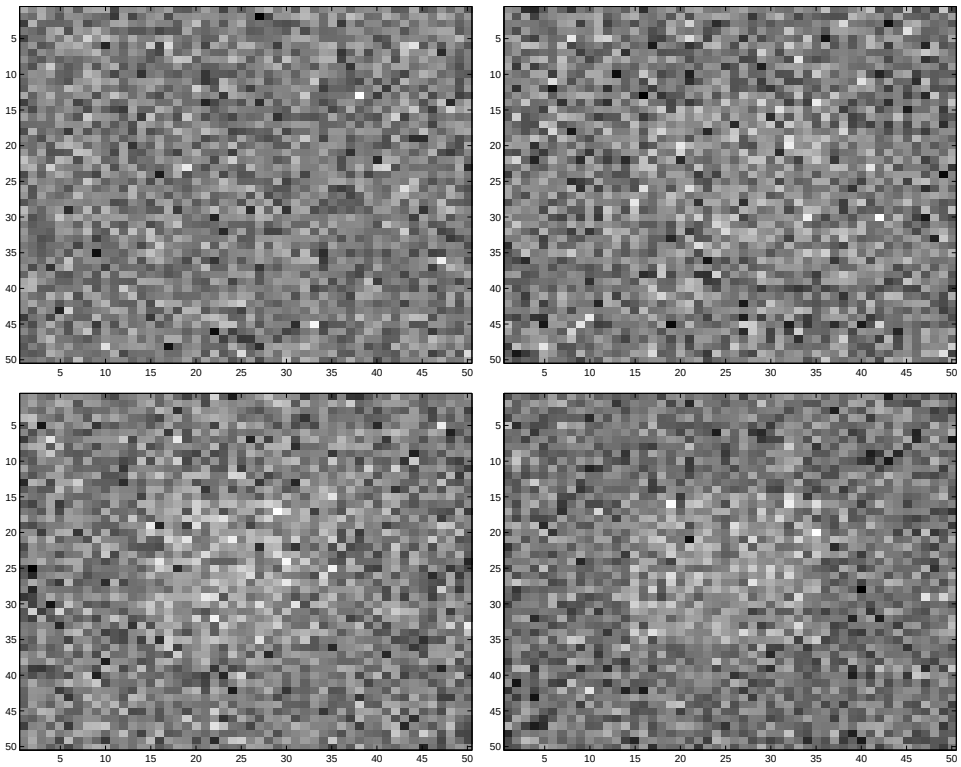

Abbildung 22.3: *Muster-Erkennung über χ^2-Statistik. Auf einem Untergrund von $N = 100$ wurde ein Objekt mit zusätzlich ΔN angebracht und poisson-verrauscht. Die Werte ΔN sind von links nach rechts und oben nach unten 0, 4, 8, 10.*

ten (Mock-Data). Auf einem quadratischen Pixel-Gitter der Größe 50×50 setzen wir in die Mitte ein quadratisches Objekt der Größe 20×20. Die mittlere Photonen-Zahl des Untergrundes beträgt 100, und darauf sitzt ein Objekt mit zusätzlich ΔN Zählern pro Pixel. Da es sich sich um ein Zähl-Experiment handelt, ist der Messwert um die angegebenen Werte poisson-verteilt. Die Abbildung 22.3 zeigt einige repräsentative Bilder für $\Delta N = 0, 4, 8, 10$. Insgesamt wurden Simulationen für die Werte $\Delta N = 0, 1, \ldots, 10$ durchgeführt und dazu aus der χ^2-Statistik der P-Wert ermittelt. Zu jedem Wert ΔN wurde die Simulation 100-mal wiederholt und Mittelwert und Standard-Fehler von P ermittelt. Die Ergebnisse sind in Abb. 22.4 dargestellt. Man erkennt, dass der P-Wert bei ungefähr $\Delta N = 8$ in den Prozentbereich eintaucht. Wenn man die Bilder in Abbildung 22.3 vergleicht, stellt man interessanterweise fest, dass man auch mit bloßem

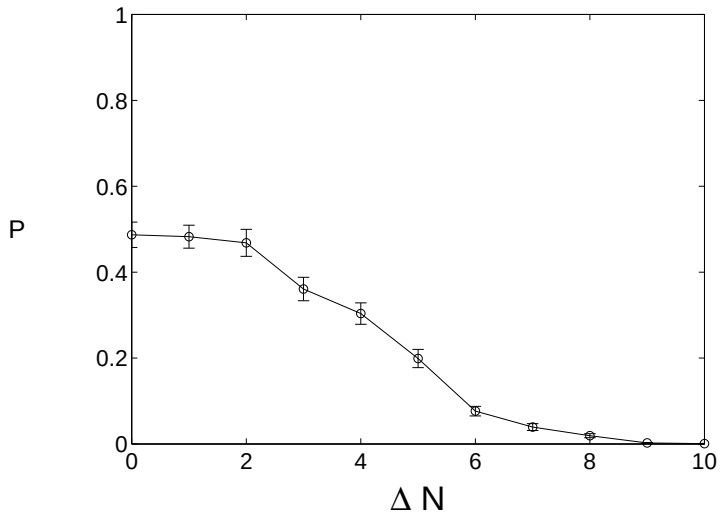

Abbildung 22.4: *P-Wert als Funktion von ΔN. Der Untergrund beträgt $N = 100$. Es wurde über 100 Konfigurationen gemittelt.*

Auge (und der Bildverarbeitungssoftware des menschlichen Gehirns) für $\Delta N = 8$ erstmals erkennt, dass ein Objekt vorhanden ist.

22.2.3 Test von Verteilungen mit unbekannten Parametern

Wir wollen nun untersuchen, ob Datenpunkte einer bestimmten Verteilung genügen. Die Parameter, von denen die Verteilung abhängt, seien nicht bekannt. Der Einfachheit halber betrachten wir nur diskrete (diskretisierte) Ereignisse, mit den Wahrscheinlichkeiten

$$p_l(a)\,, \qquad l = 1, 2, \ldots, L \qquad ,$$

wobei $a \in \mathbb{R}^{N_p}$ unbekannte Parameter darstellen.
Es liege eine Stichprobe

$$\{n_1, \ldots, n_L\}$$

vom Umfang L vor.
Unter der Annahme, dass die Hypothese korrekt ist, lautet die Likelihood-Funktion

$$P(\underline{n}|a, L, \mathcal{B}) = \frac{N}{n_1! \cdots n_L!} \prod_{l=1}^{L} p_l^{n_l}(a) \qquad .$$

Die unbekannten Modell-Parameter werden zunächst aus dem Maximum-Likelihood-Prinzip bestimmt. Da der Logarithmus monoton ist, kann man auch das

Maximum der Log-Likelihood-Funktion (vgl. Abschnitt 20.2) ermitteln

$$\begin{aligned} 0 &= \frac{\partial}{\partial a_i} \sum_{l=1}^{L} n_l \ \ln(p_l(a)) \\ &= \sum_{l=1}^{L} n_l \ \frac{\partial}{\partial a_i} \ \ln(p_l(a)) \\ &\Rightarrow a^{\mathrm{ML}} \qquad . \end{aligned} \tag{22.10}$$

Die ML-Parameter verwenden wir, um damit die theoretischen Werte

$$n_l^* = N p_l(a^{\mathrm{ML}})$$

zu ermitteln und die χ^2-Statistik zu berechnen

$$\chi_0^2 = \sum_{l=1}^{L} \frac{(n_l - n_l^*)^2}{n_l^*} \qquad . \tag{22.11}$$

Die Zahl der Freiheitsgrade ist nun nur noch

$$\nu = L - 1 - N_p \qquad ,$$

da N_p Freiheitsgrade für die Bestimmung der Parameter aufgebraucht worden sind. Ein weiterer Freiheitsgrad geht auf das Konto der Summenregel $\sum_i n_i = \sum_i n_i^*$.

Beispiel: Poisson-Verteilung

Die Daten der nachstehenden Tabelle sollen dahingehend untersucht werden, ob sie einer Poisson-Verteilung entstammen. Die Daten sind in Abbildung 22.5 abgebildet und mit dem Maximum-Likelihood-Fit an eine Poisson-Verteilung (Gl. (12.3)) verglichen.

l	0	1	2	3	4	5	6	7
n_l	153	278	288	184	103	33	25	4

Wir benötigen zunächst die ML-Lösung für den Parameter λ. Gemäß Gl. (22.10) ist die Bestimmungsgleichungen

$$\begin{aligned} 0 &= \sum_{l=0}^{L} n_l \ \frac{\partial}{\partial \lambda} \ \ln\left(e^{-\lambda} \ \frac{\lambda^l}{l!}\right) \\ &= \sum_{l=0}^{L} n_l \ \left(-1 \ + \frac{l}{\lambda}\right) = -N \left(1 - \frac{\langle l \rangle}{\lambda}\right) \qquad . \end{aligned}$$

Der ML-Wert für den Parameter λ der Poisson-Verteilung ist also gleich dem Stichproben-Mittelwert $\langle l \rangle = 2.023$. Der Umfang der Stichprobe ist 8. Da noch ein

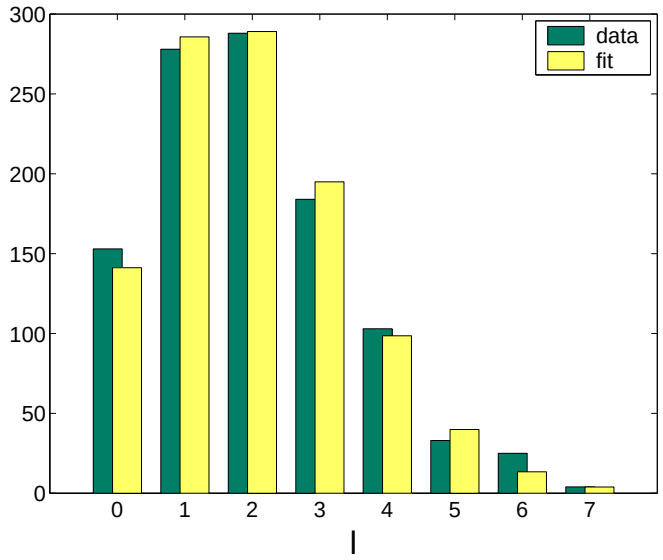

Abbildung 22.5: *Daten eines Zählexperimentes und Maximum-Likelihood-Fit einer Poisson-Verteilung an die Daten.*

Parameter aus den Daten bestimmt wurde, ist die Zahl der Freiheitsgrade $\nu = 6$. Die Analyse der Daten liefert

$$\begin{aligned} x_0 &= 13.104870 \\ P(x > x_0|H, \mathcal{B}) &= 0.041 \end{aligned} \quad .$$

Mit einer Signifikanz-Schwelle von 5% würden wir die Null-Hypothese verwerfen und die Daten als nicht poisson-verteilt abtun. Die Null-Hypothese ist jedoch noch mit einem Signifikanz-Niveau von 1% verträglich.

22.2.4 Kontingenz-Tabellen

Ein weiteres Anwendungsgebiet des χ^2-Tests sind die sogenannten KONTINGENZ-TABELLEN. Dabei geht es um folgende Fragestellung. Es werden N Versuche durchgeführt, deren Werte durch die Zufalls-Variablen x und y gekennzeichnet sind. Beide Variablen seien diskret und können die Werte $x_1, x_2, \ldots, x_l$ bzw. $y_1, y_2, \ldots, y_k$ annehmen. Die Anzahl der Versuche, die das Werte-Paar (x_i, y_j) liefert, sei n_{ij}. Man ordnet diese Zahlen nun in der sogenannten KONTINGENZ-TABELLE, wie sie in Tabelle 22.2 illustriert ist, an.

	y_1	y_1	$\dots$	y_k	Summe
x_1	n_{11}	n_{12}	$\dots$	n_{1k}	$n_{1,\cdot}$
x_2	n_{21}	n_{22}	$\dots$	n_{2k}	$n_{2,\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_l	n_{l1}	n_{l2}	$\dots$	n_{lk}	$n_{l,\cdot}$
Summe	$n_{\cdot,1}$	$n_{\cdot,2}$	$\dots$	$n_{\cdot,k}$	N

Tabelle 22.2: *Kontingenz Tabelle.*

Es wurden folgende Abkürzungen verwendet

$$
\begin{aligned}
n_{i,\cdot} &= \sum_j n_{ij} \\
n_{\cdot,j} &= \sum_i n_{ij} \\
N &= \sum_{ij} n_{ij} \qquad .
\end{aligned}
$$

Wir bezeichnen die Wahrscheinlichkeit für das Auftreten des Wertes x_i mit p_i und des Wertes y_j mit q_j. Die Null-Hypothese soll bedeuten, dass die Wahrscheinlichkeiten $P(x_i, y_j)$ unkorreliert sind, das heißt

$$
P(x_i, y_j) = p_i \; q_j \qquad .
$$

Wenn die Hypothese korrekt ist, sollte im Mittel der Eintrag in der Kontingenz-Tabelle an der Stelle ij den Wert Np_iq_j haben. Nun kennen wir die Werte q_i und p_j nicht. Sie müssen aus der Likelihood-Funktion bestimmt werden. Die Likelihood ist im vorliegenden Fall

$$
p(\underline{n}|\underline{p}, \underline{q}, N) \propto \prod_{ij} (p_i q_j)^{n_{ij}} \qquad .
$$

Die ML-Lösung erhalten wir aus der Ableitung der Log-Likelihood mit den Nebenbedingungen

$$
\begin{aligned}
\sum_i p_i &= 1 \\
\sum_j q_j &= 1 \qquad .
\end{aligned}
$$

Die Nebenbedingungen binden wir über Lagrange-Parameter an

$$\begin{aligned} 0 &= \frac{\partial}{\partial p_\nu} \left(\sum_{ij} n_{ij} \left(\ln(p_i) + \ln(q_j) \right) - \lambda_p \sum_i p_i - \lambda_q \sum_j q_j \right) \\ &= \sum_j \frac{n_{\nu j}}{p_\nu} - \lambda_p \\ &\Rightarrow \\ p_i^{\mathrm{ML}} &= \frac{1}{N} n_{i,\cdot} \qquad . \end{aligned}$$

Im letzten Schritt wurde die Normierung auf Eins berücksichtigt. Analog erhalten wir

$$q_j^{\mathrm{ML}} = \frac{1}{N} n_{\cdot,j} \qquad .$$

Die χ^2-Statistik ist dann

$$x = \sum_{ij} \frac{(n_{ij} - N p_i^{\mathrm{ML}} q_j^{\mathrm{ML}})^2}{N(p_i^{\mathrm{ML}} q_j^{\mathrm{ML}})} \qquad .$$

Die Zahl der Freiheitsgrade beträgt, da neben der Summenregel $(k-1) + (l-1)$ Parameter berechnet wurden,

$$\nu = l * k - 1 - k - l + 2 = (l-1)(k-1) \qquad .$$

Alles andere ist wie in den zuvor besprochenen Tests.

22.2.5 Vierfelder-Test

Besonders weit verbreitet ist der Vierfelder-Test, bei dem $k = l = 2$ ist. In diesem Fall ist die Zahl der Freiheitsgrade $\nu = 1$ und die χ^2-Statistik vereinfacht sich zu

$$x_0 = \frac{(n_{11} n_{22} - n_{12} n_{21})^2}{N^3 p_1^{\mathrm{ML}} (1 - p_1^{\mathrm{ML}}) q_1^{\mathrm{ML}} (1 - q_1^{\mathrm{ML}})} \qquad .$$

Man erkennt, dass $\chi^2 = 0$ ist, wenn die Determinante der Kontingenz-Matrix verschwindet. Das ist genau dann der Fall, wenn die Matrix-Elemente die Produktform $n_{ij} = N p_i q_j$ haben. Generell bedeutet ein großer χ^2-Wert, dass die Daten korreliert sind.

Beispiel: Medizinischer Test

Der Vierfelder-Test wird häufig in der Medizin eingesetzt, um zu entscheiden, ob eine von zwei Behandlungsmethoden erfolgreicher ist. x_1 steht für die Behandlungsmethode A und x_2 für die Behandlungsmethode B. y_1 gibt an, dass die Behandlung

	erfolglos	erfolgreich	Summe
Behandlung A	500	1000	1500
Behandlung B	1060	1940	3000
Summe	1560	2940	4500

	erfolglos	erfolgreich
Behandlung A	.33	.67
Behandlung B	.35	.65

Tabelle 22.3: *Kontingenz-Tabelle (oben) und relative Erfolgs/Misserfolgs-Häufigkeiten (unten).*

erfolgreich war und y_2, dass sie nichts gebracht hat. Es soll untersucht werden, ob eine der Behandlungen erfolgreicher ist. Tabelle 22.3 enthält Werte einer solchen Untersuchung. Wir erhalten hieraus $x_0 = 1.77$. Das entspricht $P(x \geq x_0|H) = 0.18$. Das bedeutet mit den üblichen Signifikanz-Niveaus von 1% oder 5%, dass diese oder größer Diskrepanzen nicht so unwahrscheinlich sind, als dass man die Hypothese verwerfen kann. Das bedeutet, dass wir nicht sagen können, dass eine Methoden besser ist als die andere.
Aus Tabelle 22.3 enthält auch die relativen Anteile von Erfolg/Misserfolg der beiden Methoden. Hieraus geht hervor, dass beide Methoden erfolgreich sind, und A nur marginal besser ist als B.

22.2.6 Simpsons Paradoxon

Es soll nun das obige Experiment genauer analysiert werden. Die Untersuchungen wurden an zwei unterschiedlichen Patientengruppen P_1 und P_2 durchgeführt, z.B. Männer/Frauen oder Kinder/Erwachsene. Wir betrachten nun die Kontingenz-Tabellen der beiden Gruppen getrennt.

Wir erhalten für die Patientengruppe P_1

	erfolglos	erfolgreich	Summe
Behandlung A	100	400	500
Behandlung B	600	1400	2000
Summe	700	1800	2500

	erfolglos	erfolgreich
Behandlung A	.2	.8
Behandlung B	.3	.7

Tabelle 22.4: *Kontingenz-Tabelle für die Patientengruppe P_1 (oben) und relative Erfolgs-/Misserfolgsanteile der Behandlungsmethoden.*

Hieraus ergibt sich $x_0 = 19.84$. Das entspricht $P(x \geq x_0|H) = 0.00000841$. Das heißt, im Unterschied zu den gepoolten Daten, muss die Null-Hypothese nun verworfen werden. Es gibt offensichtlich Korrelationen, und die relativen Erfolgshäufigkeiten zeigen ganz klar, dass die Behandlungsmethode A besser ist.
Die Kontingenz-Tabelle für die Patientengruppe P_2 ergibt

	erfolglos	erfolgreich	Summe
Behandlung A	400	600	1000
Behandlung B	460	540	1000
Summe	860	1140	2000

	erfolglos	erfolgreich
Behandlung A	.4	.6
Behandlung B	.46	.54

Tabelle 22.5: *Kontingenz-Tabelle für die Patientengruppe P_2 (oben) und relative Erfolgs-/Misserfolgsanteile der Behandlungsmethoden.*

Hierbei erhalten wir $x_0 = 7.34$. Das entspricht $P(x \geq x_0|H) = 0.007$. Auch hier ist die Wahrscheinlichkeit, dass solche oder größere Diskrepanzen unter der Null-Hypothese zufällig auftreten, nur 0.007 und somit kleiner als die gängigen Schwellwerte von 5% oder 1%. Die Null-Hypothese, dass keine Korrelation zwischen den Behandlungstypen und dem Ergebnis besteht, wird somit verworfen. Auch für diese Patientengruppe zeigen die relativen Erfolgshäufigkeiten, dass die Behandlungsmethode A besser ist als B.
Das Ergebnis erscheint paradox. Für beide Patientengruppen ist die Methode A besser als die Methode B und dennoch trifft das nicht mehr zu, wenn die Daten gemeinsam analysiert werden. Der Grund wird augenfällig, wenn man die Tabellen 22.4, 22.5 und 22.3 vergleicht. In der nachstehenden Tabelle ist das Verhältnis von Erfolg zu Misserfolg nach Patientengruppe und Behandlungsmethode aufgeschlüsselt. Man sieht, dass beide Behandlungsmethoden bei der Patientengruppe P_1 erfolgrei-

	P_1	P_2
Behandlung A	4.0	1.5
Behandlung B	2.3	1.2

Tabelle 22.6: *Tabelle des relativen Erfolgs.*

cher sind als bei der Gruppe P_2. Insbesondere ist die schlechtere Behandlungsmethode B an P_1 erfolgreicher als die bessere an P_2. Im gepoolten Datensatz kommen verhältnismäßig viele Daten von B an P_2 vor.

22.3 t-Test

Die Grundannahme des t-Tests ist, dass eine Stichprobe $\{x_1, \ldots, x_N\}$ i.u.nv. Zufalls-Variablen gemessen wurde. Die Varianz σ^2 der Verteilung ist nicht bekannt. Es soll getestet werden (Null-Hypothese), ob ein bestimmter Wert ξ der wahre Mittelwert der Verteilung ist. Die t-Statistik lautet

$$t_0 = \frac{\overline{x} - \xi}{\tilde{SF}}$$

$$\tilde{\mathrm{SF}} = \left(\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N(N-1)} \right)^{1/2} , \qquad \overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad .$$

Die Zahl der Freiheitsgrade ist $\nu = N - 1$. Man wird in diesem Fall i.d.R. einen zweiseitigen Test durchführen. Wir geben die Irrtumswahrscheinlichkeit α vor. Die zugehörige kritische Region $|t| \geq t_\alpha$ erhalten wir aus

$$P(|t| \geq t_\alpha|\nu) = 2\, P(t \geq t_\alpha|\nu) = 2 - 2 \underbrace{P(t \leq t_\alpha|\nu)}_{F_t(t_\alpha|\nu)} \stackrel{!}{=} \alpha$$

$$t_\alpha = F_t^{-1}(1 - \tfrac{\alpha}{2}|\nu) \quad .$$

Der Unterschied zum z-Test besteht darin, dass man die Varianz nicht kennt. Deshalb sind größere Abweichungen vom hypothetischen Wert nötig, um eine Hypothese verwerfen zu können. Beim z-Test hatten wir für den zweiseitigen Test gefunden, dass die Null-Hypothese zu verwerfen ist, wenn bei einem Signifikanz-Niveau von $\alpha = 0.01$ $z^{\mathrm{zs}}_{1\%} = 2.58$ ist. Bei demselben Signifikanz-Niveau ist der Grenzwert der t-Statistik $t^{\mathrm{zs}}_{1\%} = 3.25$, wenn die Stichprobe den Umfang 10 hat. In Abbildung 22.6 ist der Grenzwert $t^{\mathrm{zs}}_{1\%}$ zum zweiseitigen t-Test als Funktion des Stichprobenumfangs aufgetragen. Die kleinste zulässige Stichprobe hat den Umfang 2.

22.3.1 Vergleich von Mittelwerten

Man kann den t-Test auch heranziehen, um zu testen, ob die wahren Mittelwerte zweier Stichproben gleich sind. Voraussetzungen sind

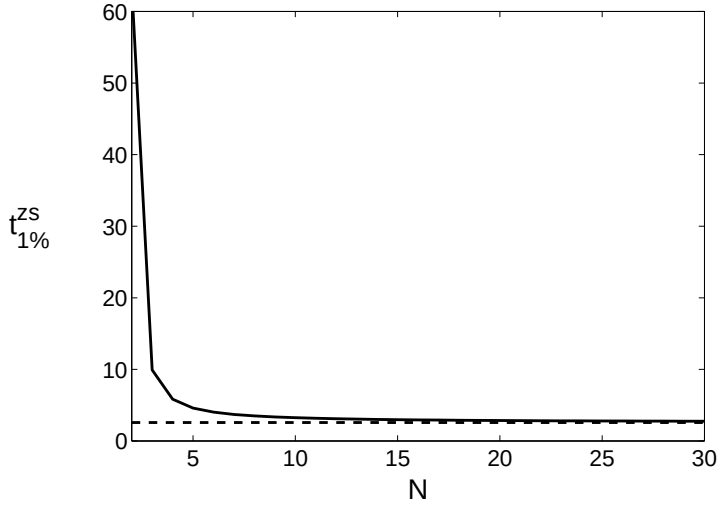

Abbildung 22.6: *Grenzwert $t^{zs}_{1\%}$ zum zweiseitigen t-Test als Funktion des Stichprobenumfangs.*

- Alle Zufalls-Variablen sind unabhängig voneinander.
- Die Zufalls-Variable innerhalb der beiden Stichproben sind jeweils i.u.nv.
- Die Varianz ist in beiden Fällen gleich.
- Null-Hypothese: Die Mittelwerte, die den beiden Stichproben zugrundeliegen, sind gleich.

Die Stichproben $\{d^\beta_1, \ldots, d^\beta_{L_\beta}\}, \quad (\beta = 1, 2)$ haben den Umfang L_1 und L_2. Die Gesamtzahl der Stichproben-Elemente ist

$$L = L_1 + L_2 \qquad .$$

Mittelwerte und Varianzen der beiden Stichproben sind

$$\overline{d^{(\beta)}} = \frac{1}{L_\beta} \sum_{i=1}^{L_\beta} d^\beta_i$$

$$\overline{(\Delta d^{(\beta)})^2} = \frac{1}{L_\beta} \sum_{i=1}^{L_\beta} (d^\beta_i - \overline{d^\beta})^2 \qquad .$$

Die Null-Hypothese besagt, dass die Differenz der Stichproben Mittelwerte

$$\Delta = \overline{d^{(2)}} - \overline{d^{(1)}}$$

um Null normal-verteilt ist. Der Schätzwert für die Varianz der Differenz Δ ist

$$\tilde{\sigma}^2 = \frac{L}{L^{(1)} L^{(2)}} \left(L^{(1)} \, \overline{(\Delta d^{(1)})^2} + L^{(2)} \, \overline{(\Delta d^{(2)})^2} \right) \qquad . \tag{22.12}$$

Hierbei wurden wieder die Varianzen der beiden Stichproben quadratische addiert. Es ist sehr sinnvoll, dass die Schätzwerte für die Varianzen aus den Stichproben mit ihren Größen gewichtet eingehen, da auf diese Weise ihre Zuverlässigkeit korrekt berücksichtigt werden. Die t-Statistik ist dann

$$t = \frac{\Delta}{\tilde{\mathrm{SF}}} = \sqrt{L-2}\ \frac{(\overline{d^{(2)}} - \overline{d^{(1)}})}{\tilde{\sigma}} \quad ,$$

wobei die Zahl der Freiheitsgrade $\nu = L - 2$ beträgt, da jede Stichprobe wegen der Bestimmung des jeweiligen Mittelwerts einen Freiheitsgrad verliert.

TEST OB ZWEI STICHPROBEN DENSELBEN MITTELWERT HABEN BEI UNBEKANNTER VARIANZ: t-TEST

Die Größe

$$t = \sqrt{L-2}\ \frac{(\overline{d^{(2)}} - \overline{d^{(1)}})}{\tilde{\sigma}}$$

mit

$$\tilde{\sigma} = \frac{L}{L^{(1)}L^{(2)}}\ \left(L^{(1)}\ \overline{(\Delta d^{(1)})^2} + L^{(2)}\ \overline{(\Delta d^{(2)})^2}\right) \tag{22.13}$$

genügt der Student-t Verteilung

$$p_t(t|\nu) = \frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}\ \frac{1}{\sqrt{\pi\nu}}\ \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{22.14}$$

mit $\nu = L - 2$ Freiheitsgraden.

Beispiel zum Vergleich von Mittelwerten

Die Tabelle 22.7 enthält zwei Stichproben der Lebensdauer in willkürlichen Einheiten von zwei radioaktive Elementen X und Y. Die Null-Hypothese besagt, dass es sich bei beiden Präparaten um dasselbe Element handelt. Aus den Werten in der Tabelle erhält man eine Standardfehler für die Differenz der Mittelwerte von

$$\tilde{\mathrm{SF}} = \frac{\tilde{\sigma}}{\sqrt{L-2}} = 1.95 \quad .$$

Damit ist der Wert der t-Statistik

$$t_0 = \frac{\overline{d^X} - \overline{d^Y}}{\tilde{\mathrm{SF}}} = 0.62 \quad .$$

i	d_i^X	d_i^Y
1	11	18
2	14	11
3	8	9
4	12	7
5	9	14
6		10
7		15
$\overline{d^{(\beta)}}$	10.8	12.0
$\overline{(\Delta d^{(\beta)})^2}$	4.56	12.57

Tabelle 22.7: *Messungen der Lebensdauer (in willkürlichen Einheiten) zweier radioaktiver Elemente X und Y.*

Die Zahl der Freiheitsgrade beträgt $\nu = 10$. Bei einem zweiseitigen Test zum Signifikanz-Niveau von 1% beginnt die kritische Region bei

$$t_{1\%} = 3.2 \qquad .$$

Damit kann die Null-Hypothese nicht verworfen werden. Wir berechnen noch den P-Wert

$$P = P(|t| \geq t_0|\nu) = 2F_t(-|t_0|\,|\nu) = 2 * 0.276 = 0.552 \qquad ,$$

der besagt, dass Diskrepanzen in den Stichprobenmittelwerten, die mindestens so groß sind wie die beobachtete, aufgrund von zufälligen Fluktuationen mit einer Wahrscheinlichkeit 0.55 auftreten können und die Null-Hypothese von daher nicht verworfen werden kann.

22.4 F-Test

Unter der F-Statistik versteht man das Verhältnis der Schätzwerte zweier Stichproben-Varianzen.

$$f_0 = \frac{\tilde{\sigma_1}^2}{\tilde{\sigma_2}^2} \tag{22.15}$$

Voraussetzung ist, dass die Elemente der Stichproben i.u.nv. sind, und wir nehmen an, dass beide Stichproben dieselbe intrinsische Varianz σ^2 besitzen. Wir hatten bereits in 21.5 gezeigt, dass die $F-$Statistik einer Wahrscheinlichkeitsverteilung genügt, die unabhängig von σ^2 ist und nur von den Umfängen der beiden Stichproben abhängt. Weiters hatten wir gesehen, dass die Wahrscheinlichkeitsdichte der F-Statistik nicht von den Mittelwerten der beiden Stichproben abhängen. Der Umfang der Stichproben sei N_β. Die Zahl der Freiheitsgrade in den Stichproben ist dann $\nu_\beta = N_\beta - 1$.

Die Null-Hypothese ist, dass beiden Stichproben (ungeachtet der Mittelwerte) dieselbe intrinsische Varianz haben und deshalb auf denselben physikalischen Ursprung zurückgehen.

Das wesentliche am F-Test ist, dass weder zur Berechnung der Statistik noch für die Wahrscheinlichkeitsdichte der intrinsische Wert der Varianz σ^2 benötigt wird. Deshalb kann der F-Test in vielen Fällen angewandt werden.

Wenn die Hypothese richtig ist, sollte der berechnete Wert (f_0) der Statistik nahe bei eins liegen. Wenn er wesentlich davon abweicht, heißt das, dass die Hypothese wahrscheinlich nicht korrekt ist. Wie im χ^2-Test verwendet man als Signifikanz-Niveau 5% oder 1% und verwirft die Hypothese, wenn die Wahrscheinlichkeit $P(f \geq f_0)$, dass ein Wert f größer-gleich f_0 vorkommt, wenn die Hypothese korrekt ist, größer ist als das Signifikanz-Niveau. Bei einem Signifikanz-Niveau α verwerfen wir die Hypothese, wenn

$$P(f \geq f_0) = \int_{f_0}^{\infty} df \; p_F(f|N_1, N_2, \mathcal{B}) \leq \alpha \qquad .$$

Die Hypothese ist auch zu verwerfen, wenn f_0 deutlich kleiner als eins ist. In diesem Fall muss gelten

$$P(f \leq f_0) = \int_{0}^{f_0} df \; p_F(f|N_1, N_2, \mathcal{B}) \leq \alpha \qquad .$$

Da aber im Fall $f_0 < 1$ nur die Indizes der Stichproben vertauscht werden müssen, um zu $f_0 > 1$ zu gelangen, wählt man die Indizes i.d.R. so, dass $f_0 > 1$.

F-TEST
Die Größe $$f = \frac{\tilde{\sigma_1}^2}{\tilde{\sigma_2}^2} \tag{22.16}$$ *genügt der F-Statistik mit der Wahrscheinlichkeitsdichte* $$p_F(f|\nu_1, \nu_2) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \; r \; (rf)^{\frac{\nu_1-2}{2}} \; (1+fr)^{-\frac{\nu_1+\nu_2}{2}} \tag{22.17}$$ *mit* $\quad r = \frac{\nu_1}{\nu_2} \qquad .$ *mit* $\nu_\beta = N_\beta - 1$.

Beispiel

Wir betrachten zwei Bildausschnitte, die jeweils 4 Pixel enthalten, eines Bildes, das im wesentlichen aus Untergrund besteht. Im ersten Ausschnitt finden wir die Werte $\{2.0, 3.0, 1.0, 2.0\}$ und im zweiten $\{2.0, 2.5, 1.5, 2.0\}$. Die Null-Hypothese sei:
in beiden Fällen sehen wir nur Untergrund,
das heißt, die Varianzen sollen gleich sein. Die Stichproben-Mittelwerte sind in beiden Fällen gleich 2.0. Die Schätzwerte der Varianzen liefern

$$\tilde{\sigma}_1^2 = \frac{2}{3}$$

und

$$\tilde{\sigma}_2^2 = 0.25\ \frac{2}{3} \qquad .$$

Damit ist der Wert der F-Statistik

$$f_0 = 4.0 \qquad .$$

Wir vereinbaren, dass wir die Stichproben immer so nummerieren, dass der Wert der F-Statistik größer eins ist. Bei einem Signifikanz-Niveau α liegt die Grenze zur kritischen Region bei

$$\int_{f_\alpha}^{\infty} p_F(f|\nu_1, \nu_2)\ df = 1 - F_F(f_\alpha|\nu_1, \nu_2) \stackrel{!}{=} \alpha \qquad .$$

Daraus folgt

$$f_\alpha = F_F^{-1}(1 - \alpha|\nu_1, \nu_2) \qquad .$$

Wenn wir speziell $\alpha = 0.1$ wählen ist in unserem Beispiel

$$f_{10\%} = F_F^{-1}(0.95|3, 3) = 5.3908 \qquad .$$

Das heißt, selbst bei einem Signifikanz-Niveau von 10% (in dem wir erlauben, dass die Null-Hypothese in 10% der Fälle irrtümlich verworfen wird) kann die Null-Hypothese nicht verworfen werden. Der P-Wert ist

$$P = 1 - F_F(f_0|3, 3) = 0.14 \qquad ,$$

was zeigt, dass im Bereich $f \geq f_0$ 14% der Wahrscheinlichkeitsmasse steckt.

22.5 Kritik an der Test-Logik

Grundlage der Signifikanz-Tests ist die Likelihood-Funktion

$$P = P(s \geq s_0|H_0) \qquad ,$$

das heißt, die Wahrscheinlichkeit, dass der Werte der untersuchten Statistik größergleich dem beobachteten Wert sind. Zum einen ist die Wahl der kritischen Bereiche abhängig von den Alternativen. Diese werden aber in den typischen Anwendungen nicht berücksichtigt. In den meisten Fällen wird man einseitige oder zweiseitige Test durchführen. Über die Fehler zweiter Art weiß man i.d.R. wenig. Wir hatten bereits festgestellt, dass ein großer P-Wert bedeutet, dass die Null-Hypothese nicht verworfen werden kann. Das heißt aber umgekehrt nicht, dass die Null-Hypothese stimmt. Für den inversen Schluss benötigen wir ALLE Alternativ-Hypothesen. Extremes Beispiel, H_0 besagt Herr X wäscht den Wagen. Die Beobachtung z_0 besagt, dass die Straße nass ist. Damit ist

$$P(z_0 \in I_{\overline{r}}|H_0) = 1 \qquad .$$

Das heißt aber noch lange nicht, dass Herr X den Wagen wäscht. Genausogut kann es regnen.
Im eben angeführten Beispiel kann man leicht falsifizierende Schlüsse ziehen. Wenn die Straße nämlich nicht nass ist, wissen wir, dass die Hypothese H_0 falsch ist. Diese eindeutige Konsequenz liegt jedoch daran, dass $P(\overline{z_0} \in I|H_0) = 0$. Wenn H_0 keine deterministischen Schlüsse zulässt, ist auch dieser Umkehrschluss schwierig. Wenn z.B. $P(\overline{z_0} \in I|H_0) = 0.01$, dann bedeutet das für den Fall, dass die Straße nicht nass ist, dass nur in 1% der Fälle beim Autowaschen die Straße nicht nass wird. Wir können also ziemlich sicher sein, dass Herr/Frau X den Wagen nicht waschen. Voraussetzung ist, dass wir außer der Feuchtigkeit auf der Straße nichts wissen. Es könnte aber sein, dass zusätzliches Vorwissen existiert, das besagt, dass andere Alternativen auszuschließen sind.
In naturwissenschaftlichen Problemen sind die vorliegenden Daten sicherlich nicht die einzige Information, die für das betrachtete Problem bekannt ist. Es wird in der Regel schon sehr viel Vorwissen in Form von früheren experimentellen Daten, allgemeinem Wissen oder theoretischen Fakten geben, wie zum Beispiel Positivitätsforderungen, Summenregeln, Grenzwert-Verhalten.
EIN WISSENSCHAFTLER STEHT NIEMALS VOR DER SITUATION, DASS NUR DIE GEGENWÄRTIGEN DATEN ZÄHLEN.
Insofern benötigen wir unbedingt eine konsistente Theorie, die es erlaubt, alles bekannte Wissen einfließen zu lassen. Es wurde von Kritikern bemängelt, dass das bedeutet, dass die Wahrscheinlichkeit subjektiv sei. Wir haben mathematisch streng bewiesen, dass die Wahrscheinlichkeitstheorie die einzig konsistente Theorie zur Behandlung von Teilwahrheiten ist. Die Theorie ist in sich konsistent und deterministisch. Das einzige, was als subjektiv bezeichnet werden kann, ist das Vorwissen, dass in die Prior-Wahrscheinlichkeiten einfließt. DIESES MASS DER SUBJEKTIVITÄT IST ABER DAS FUNDAMENT ALLER WISSENSCHAFTEN. DAS IST DIE EXPERTISE, DIE IN DER JEWEILIGEN DISZIPLIN VORLIEGT. NATÜRLICH HÄNGT ES VOM ENTWICKLUNGSGRAD DES JEWEILIGEN FORSCHUNGSFELDES AB.
Die Durchführung bzw. Generierung von experimentellen Daten oder die Ermittlung von Stichproben – und somit die orthodoxen Statistiken – ist jedoch im gleichen Maße subjektiv motiviert vom Vorwissen und präjudiziert teilweise das Ergebnis. I.d.R.

wird man, geleitet von theoretischen Modellen, Daten, die dem Vorwissen entsprechen, schnell akzeptieren und solche, die absolut nicht ins Bild passen, wiederholt messen. All das, was bei der Daten- und Stichproben-Erzeugung in großem Maße unbewusst abläuft, wird mit der Wahrscheinlichkeitstheorie in einen mathematisch strengen und konsistenten Rahmen gebracht.
Wenn z.B. die Auslenkung eines harmonischen Oszillators als Funktion der Zeit zu analysieren ist, wird ein Physiker niemals auf die Idee kommen, für $x(t)$ ein Potenzgesetz der Form

$$x_P(t) = a + b\,t + c\,t^2$$

anzusetzen. Er wird immer die Form einer gedämpften Schwingung verwenden

$$x_O(t) = A\cos(\omega t + \varphi)\; e^{-\lambda t} \qquad .$$

Auch wenn ein Signifikanz-Test liefern sollte, dass x_O zu verwerfen und x_P zu favorisieren ist, ist unser Vorwissen doch so dominant, die Aussage der Daten nicht zu ernst zu nehmen. Insbesondere wissen wir, dass wir dem Modell $x_P(t)$ definitiv nicht für große Zeiten trauen werden.
Das heißt, da es in den Tests immer eine Restunsicherheit gibt (Fehler erster, bzw. zweiter Art), dass auch das Falsifizieren einer Hypothese nicht zwingend ist. Im Rahmen der Wahrscheinlichkeitstheorie verschwindet die Asymmetrie zwischen Falsifizieren und Verifizieren. Beide Fälle werden gleichwertig behandelt, indem man die Wahrscheinlichkeit $P(H|D, I, \mathcal{B})$ für die Hypothese bestimmt, im Lichte der neuen Daten und des Vorwissens des Fachgebietes und des Bedingungskomplexes.

Kapitel 23

Wahrscheinlichkeitstheoretische Hypothesen Tests

Der Hypothesen-Test im Rahmen der Bayessche Wahrscheinlichkeitstheorie ist denkbar einfach. Wir wollen zwei Hypothesen H_1 und H_2[1] im Lichte von neuen experimentellen Daten D vergleichen. Dazu benötigen wir die Wahrscheinlichkeiten

$$P(H_\alpha|D,\mathcal{B})$$

für die beiden Hypothesen H_α, gegeben die Daten D und weiteres Vorwissen, das im Bedingungskomplex $\mathcal{B}$ zusammengefasst ist. In der Rechnung werden wir nach Bedarf auf die in $\mathcal{B}$ enthaltene Information zurückgreifen und auch die Daten D genauer spezifizieren. Das Bayessche Theorem liefert

$$P(H_\alpha|D,\mathcal{B}) = \frac{P(D|H_\alpha,\mathcal{B})\; P(H_\alpha|\mathcal{B})}{P(D|\mathcal{B})} \quad .$$

Es ist sinnvoll, das sogenannte Odds-Ratio zu berechnen

ODDS-RATIO

$$o = \frac{P(H_1|D,\mathcal{B})}{P(H_2|D,\mathcal{B})} = \underbrace{\frac{P(D|H_1,\mathcal{B})}{P(D|H_2,\mathcal{B})}}_{o_{\mathrm{BF}}} \underbrace{\frac{P(H_1|\mathcal{B})}{P(H_2|\mathcal{B})}}_{o_P} \quad . \tag{23.1}$$

Man nennt den ersten Term den BAYES-FAKTOR und den zweiten PRIOR-ODDS. Der Vorteil der Odds-Ratios ist, dass er Normierungsfaktor $P(D|\mathcal{B})$, die sogenannte DATEN-EVIDENZ, herausfällt. Die Daten-Evidenz ist hier irrelevant. Prinzipiell gibt

[1] Ein Spezialfall ist natürlich $H_2 = \overline{H_1}$.

sie an, wie wahrscheinlich die gemessenen Daten für eine bestimmte Problemstellung ($\mathcal{B}$) sind, ungeachtet der speziellen Modell-Parameter. Z.B. könnte $\mathcal{B}$ besagen, dass wir optische Experimente im sichtbaren Bereich machen und Wellenlängenmessungen durchführen. Dann wird die Datenevidenz für Längen im Bereich um 555 nm maximal sein und nach 380nm bzw 780nm hin auf Null abfallen.
Aus dem Odds-Ratio erhalten wir die Wahrscheinlichkeiten über die Formel

$$P(H_1|D,\mathcal{B}) = \frac{o}{1+o} \qquad . \tag{23.2}$$

Diese Form gilt, wenn es nur zwei Alternativhypothesen gibt. Die Verallgemeinerung auf mehrere Hypothesen werden wir später besprechen.
Die Formel Gl. (23.1) quantifiziert, wie sich unser Wissensstand durch die neuen experimentellen Daten, erweitert hat. Der Bayes-Faktor stellt den Erkenntnisgewinn aufgrund des neuen Daten dar. Der Odds-Ratio gibt an, was vor dem Experiment bereits bekannt war. In der Regel werden Experimente von Fach-Experten durchgeführt und nicht von Laien. Das heißt, vor dem Experiment liegt bereits Wissen vor, z.B. aus früheren Experimenten, das im Prior-Odds zu berücksichtigen ist. Dieser Faktor ist zwar nicht immer leicht zu erhalten, es ist aber unseriös, wenn man auf einem Gebiet wissenschaftlich tätig ist und sich nicht die Mühe macht, sich in das bereits bestehende Fachwissen einzuarbeiten. Natürlich kann es auch Probleme geben, bei denen noch kein Vorwissen vorhanden ist. Dann muss man $o_P = 1$ wählen, da jede andere Wahl bedeutet, dass man doch etwas weiß. Wenn das nicht der Fall ist, ist der Index an den Hypothesen völlig willkürlich, und aus Symmetriegründen ist $o_P = 1$ zu wählen. Der Prior-Odds gibt den Wissensstand vor dem Experiment an und nicht welche Hypothese tatsächlich richtig ist oder besser zutrifft.
Wir wollen z.B. überprüfen, ob jemand telepathische Fähigkeiten besitzt. Hierzu werde eine symmetrische Münze N-mal verdeckt geworfen und man überprüft die Übereinstimmungen. Von den N Würfen sollen n Vorhersagen zutreffen. Die Hypothese ist

$$H : \textit{Die Person hat telepatische Fähigkeiten.}$$

Wenn die komplementäre Hypothese $\overline{H}$ zutrifft, sind die Übereinstimmungen zufällig und sollten die Wahrscheinlichkeit $q = 1/2$ haben. Bei Vorliegen telepathischer Fähigkeiten ist die Wahrscheinlichkeit für Übereinstimmungen $q > 1/2$. Die Marginalisierungsregel liefert

$$P(n|N,H,\mathcal{B}) = \int_0^1 P(n|q,N,H,\mathcal{B})\; p(q|N,H,\mathcal{B})\; dq \qquad .$$

Nun besagt die Hypothese H, dass die Person telepathisch veranlagt ist ohne genauer zu spezifizieren, wie stark diese Fähigkeit ausgeprägt ist. Demnach gilt

$$p(q|N,H,\mathcal{B}) = 2\;\theta(\tfrac{1}{2} < q \leq 1) \qquad .$$

Der Likelihood-Term $P(n|q,N,H,\mathcal{B}) = P_B(n|q,N)$ ist nichts anderes als die Bernoulli-Verteilung

$$P(n|q,N,H,\mathcal{B}) = \binom{N}{n}\, q^n\, (1-q)^{N-n} \qquad .$$

Somit liefert das Ergebnis

$$\begin{aligned} P(n|N,H,\mathcal{B}) &= 2\binom{N}{n} \int_{1/2}^{1} q^n\,(1-q)^{N-n}\, dq \qquad \overset{p=1-q}{\Longrightarrow} \\ &= 2\binom{N}{n} \int_{0}^{1/2} p^{N-n}\,(1-p)^n\, dp \\ &= 2\binom{N}{n}\, B(\tfrac{1}{2}; N-n+1, n+1) \qquad , \end{aligned}$$

wobei der letzte Faktor die unvollständige Beta-Funktion Gl. (9.9b) ist. Im komplementären Fall $\overline{H}$ ist die Berechnung einfacher

$$\begin{aligned} P(n|N,\overline{H},\mathcal{B}) &= P(n|q=1/2,N,\mathcal{B}) \\ &= \binom{N}{n}\, 2^{-N} \qquad . \end{aligned}$$

Damit lautet der Bayes-Faktor

$$o_{\mathrm{BF}} = 2^{N+1}\, B(\tfrac{1}{2}; N-n+1, n+1) \qquad .$$

Nehmen wir an, es sei $N = 10$ und $n = 7$, dann erhalten wir $o_{\mathrm{BF}} = 1.38$ und wenn wir den Prior-Odds-Faktor gleich Eins wählen folgt daraus

$$P(n|N,H,\mathcal{B}) = \frac{o_{\mathrm{BF}}}{1+o_{\mathrm{BF}}} = 0.58 \qquad .$$

Diese Wahrscheinlichkeit ist nicht so groß, dass wir von den telepathischen Fähigkeiten der Versuchsperson überzeugt sind, $n = 7$ bei 10 Versuchen kann auch eine statistische Fluktuation sein. Wir machen deshalb ein umfangreicheres Experiment $N = 100$ und wir finden $n = 77$. Die daraus resultierenden Werte sind $o_{\mathrm{BF}} = 1\,009\,520$ bzw. $P = 0.9999990$. Hier können wir aufgrund der Daten mit an Sicherheit grenzender Wahrscheinlichkeit, davon ausgehen, dass die Person telepathisch veranlagt ist. Trotzdem wird kaum jemand von dem Ergebnis überzeugt sein und fortan an telepathische Fähigkeiten glauben, da wir bereits in so vielen Situationen erlebt haben, dass diese Übereinstimmungen nicht auf telepathischen Fähigkeiten sondern auf irgendeiner Art von Manipulation beruhen. Und wir werden von daher einen extrem kleinen Wert für den Prior-Odds-Faktor ansetzen. Wie können wir ihn abschätzen. Wenn auf der Erde jemand nachweislich und reproduzierbare telepathische Fähigkeiten besäße, wäre das mittlerweile bekannt. Es gibt auf der Erde $\approx 10^{10}$ Menschen.

Bei welcher Zahl ist es denkbar, dass ihre telepathischen Fähigkeiten unbeobachtet geblieben sind. Vielleicht $O(10^2)$. Demnach würden wir für den Prior-Odds-Faktor grob den Wert

$$o_P = \frac{10^{-8}}{1 - 10^{-8}} \approx 10^{-8}$$

ansetzen. Damit ist im Falle von $N = 100$ und $n = 77$, $o = 0.001$ und $p = 0.001$. Das macht mehr Sinn. Wir sehen aber auch, wie sich aufgrund der Daten die verschwindend kleinen Prior-Wahrscheinlichkeit von 10^{-8} deutlich vergrößert hat. Das ist gerade der LERNEFFEKT, den die Wahrscheinlichkeitstheorie quantifiziert.
Natürlich sollte man den Prior-Odds-Faktor nicht Null oder Unendlich wählen, ansonsten sind Experimente überflüssig.

23.1 Stimmt der vorgegebene Mittelwert?

Wir beginnen mit dem einfachsten Beispiel. Wir gehen von I.U.NV. ZUFALLSZAHLEN MIT BEKANNTER VARIANZ σ^2 aus. Das ist Teil des Bedingungskomplexes ($\mathcal{B}$). Außerdem enthält er die Information, um welches physikalische Problem es sich handelt. Wenn wir Längen untersuchen, ist es sicherlich ein Unterschied, ob wir extraterrestrische Objekte oder Atome untersuchen. Diese Information legt ein Intervall $I = [m_1, m_2]$ fest, innerhalb dessen sich der noch unbekannte Mittelwert bewegen wird.
Es liege eine Stichprobe $\underline{x} = \{x_1, \ldots, x_N\}$ vom Umfang N vor. Die zu untersuchende Hypothese lautet

$$H : \textit{Der wahre Mittelwert ist } m.$$

Das Komplement besagt, dass der wahre Mittelwert eben nicht m ist. Gemäß des Bedingungskomplexes kann er dann irgendeinen Wert $m \in I$ annehmen

$$P(m|\overline{H}, \mathcal{B}) = \frac{1}{V_m}\, \theta(m_1 \leq m \leq m_2) \qquad .$$

Hierbei ist $V_m = m_2 - m_1$. Genauso gut könnten wir natürlich auch andere Hypothesen aufstellen und evaluieren. Wir benötigen die marginale Likelihood-Funktion

$$\begin{aligned} P(\underline{x}|H, \mathcal{B}) &= P(\underline{x}|m, \sigma, \mathcal{B}) = (2\pi\sigma^2)^{-N/2}\; e^{-\frac{1}{2\sigma^2}\;\sum_{i=1}^{N}\,(x_i - m)^2} \\ &= (2\pi\sigma^2)^{-N/2}\; e^{-\frac{N}{2\sigma^2}\;\mathrm{var}(x)}\; e^{-\frac{N}{2\sigma^2}\;(m-\overline{x})^2} \\ &= C\; e^{-\frac{N}{2\sigma^2}\;(m-\overline{x})^2} \qquad . \end{aligned} \tag{23.3}$$

Die marginale Likelihood von $\overline{H}$ erhalten wir aus der Marginalisierungsregel

$$\begin{aligned} P(\underline{x}|\overline{H}, \mathcal{B}) &= \int\; P(\underline{x}|m, \sigma, \mathcal{B})\; P(m|\overline{H}, \mathcal{B})\; dm \\ &= \frac{C}{V_m} \int_{m_1}^{m_2} e^{-\frac{N}{2\sigma^2}\;(m-\overline{x})^2}\; dm \qquad . \end{aligned}$$

Der Einfachheit halber nehmen wir an, dass der Stichproben-Mittelwert mehrere Standard-Fehler

$$\mathrm{SF} = \frac{\sigma}{\sqrt{N}}$$

von den Grenzwerten m_α entfernt ist. Das wird bei einem gut konzipierten Experiment der Fall sein, da das Experiment ansonsten keine wesentlich über unser Vorwissen hinausgehende Information liefert. Dann gilt

$$\begin{aligned} P(\underline{x}|\overline{H}, \mathcal{B}) &= \frac{C}{V_m} \int_{-\infty}^{\infty} e^{-\frac{N}{2\sigma^2}\,(m-\overline{x})^2}\, dm \\ &= \frac{C\ \mathrm{SF}\ \sqrt{2\pi}}{V_m} \end{aligned}$$

und der Bayes-Faktor ergibt

$$o_{\mathrm{BF}} = \frac{V_m}{\mathrm{SF}}\ \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}z^2} \qquad . \tag{23.4}$$

Das Endergebnis ist also

BAYESSCHER Z-TEST

$$\begin{aligned} o_{\mathrm{BF}} &= \frac{V_m}{\mathrm{SF}}\ p_Z(z) \\ z &:= \frac{m - \overline{x}}{\mathrm{SF}} \qquad . \end{aligned} \tag{23.5}$$

Das Ergebnis ist aus mehreren Gründen äußerst befriedigend

- Es geht nur die Wahrscheinlichkeitsdichte für den Wert der Statistik ein, der zur Stichprobe gehört.
- Es geht das Vorwissen über den Prior-Odds ein.
- Es muss kein willkürlich gewähltes Signifikanz-Niveau vorgegeben werden. Diese Funktion wird gewissermaßen von V_m, dem Prior-Volumen übernommen. Das ist sehr befriedigend, denn es fließt nun fachspezifisches Wissen ein.

Interessanterweise ist der Unterschied zwischen dem P-Wert (Wert der Verteilungsfunktion) und dem Wert der Wahrscheinlichkeitsdichte der z-Statistik am Messwert z_0 gar nicht so groß. Wir betrachte hierzu das Verhältnis

$$q = \frac{p_z(z_0)}{\int_{z_0}^{\infty}\ p_z(z)\ dz} \qquad .$$

Dieses Verhältnis bewegt sich im relevanten Bereich, das heißt wenn die P-Werte zwischen 0.5 und 0.0001 liegen, zwischen 1 und 4. Der wesentliche Unterschied zwischen Wahrscheinlichkeitstheorie und orthodoxer Statistik liegt hier vielmehr im Ockham-Faktor.

23.1.1 Der Ockham-Faktor

Der Faktor

$$\frac{V_m}{\mathrm{SF}}$$

wird OCKHAM-FAKTOR genannt. Er ist das Verhältnis der Wahrscheinlichkeitsmasse der von der Likelihood erlaubten Werte von m zu der des Priors. Dieser Faktor „bestraft" komplexe Modelle. Im vorliegenden Beispiel enthält das erste Modell keine freien Parameter und ist somit weniger komplex als das zweite Modell, bei dem der Mittelwert m alle Werte aus I annehmen kann. Solange die Daten kompatibel mit dem ersten Modell sind, wird dieses favorisiert. Erst wenn das erste Modell eine zu großen Diskrepanz zu den Daten aufweist, gewinnt das zweite, komplexe Modell. Die Regeln der Wahrscheinlichkeitstheorie quantifizieren eindeutig die Idee von Ockham's Razor.
Wir wollen den Ockham-Faktor an einem einfachen Urnen-Beispiel diskutieren. Wir betrachten zwei Arten von Urnen (zwei Modelle). Jede Urne enthält Kugeln, auf denen ganze Zahlen aufgedruckt sind. Die erste Urne (einfaches Modell (I)) ist wenig flexibel und enthält nur Kugeln mit den Zahlen 1 und 2, die in gleicher Häufigkeit vorkommen. Die zweite Urne beschreibt ein wesentlich komplexeres Modell (II) und enthält mit gleicher Häufigkeit die Zahlen $1, 2, \ldots 10$.
Die a-priori Wahrscheinlichkeit der beiden Modelle sei gleich. Das bedeutet, dass in der Natur beide Urnen-Arten gleich häufig vorkommen. Nehmen wir an, basierend auf den beiden Modellen kommen in der Natur (Population) 20000 Ereignisse vor, $N = 10000$ von jedem Modell. Das Modell I liefert im Mittel 5000 mal eine eins und 5000 mal eine zwei. Das zweite Modell liefert die natürlichen Zahlen $1 - 10$ im Mittel je 1000 mal. Die eins kommt in der Grundgesamtheit also insgesamt 6000 mal vor, davon stammen 5000 vom ersten Modell. Das heißt, wenn wir eine eins messen, ist die Wahrscheinlichkeit $5000/6000 = 0.83$, dass die Ursache für die eins das Modell I ist (bzw., dass diese eins aus der Urne I stammt). Obgleich beide Modelle die eins erklären können, wird das Modell II „bestraft", da es eine größere Vielfalt von natürlichen Zahlen erzeugt (Ockham's Razor). Das gleiche gilt für die Zahl zwei. Beobachten wir hingegen eine drei, so muss es sich um Modell II handeln. Hier kommt dann ins Spiel, dass dieses Modell aufgrund seiner Komplexität auch diese Zahl erklären kann, die im anderen Modell nicht vorkommt. Dieser Fall zeigt extrem die Wirkung des Daten-Konstraints (Likelihood-Anteils).
Zusammenfassend gilt, wenn die beobachtete Zahl in beiden Urnen vorkommt, ist die Wahrscheinlichkeit größer, dass sie aus der Urne mit der geringeren Vielfalt stammt, da mehr Kugeln mit der beobachteten Zahl von ihr im Umlauf sind als von der anderen.

Auf die Datenanalyse übertragen, bedeutet das: wenn die Daten kompatibel mit beiden Modellen sind, ist es wahrscheinlicher, dass das weniger flexible Modell das zutreffende ist, da es die beobachteten Daten häufiger in der Natur realisiert als das flexiblere Modell, das die von ihm erzeugten Elemente der Population auf eine große Vielfalt verteilt.

Wir wollen die Idee des Ockham-Faktors noch an einem anderen Problem diskutieren. Wir betrachten wieder zwei Modelle. Das einfachere Modell besagt, alle Massen-Punkte bewegen sich auf einer Geraden. Das komplexere Modell sagt aus, dass die Massen-Punkte sich in einer Ebene bewegen. Beobachtet werden Orte von Teilchen, die im Rahmen der Messgenauigkeit auf einer Geraden liegen. Natürlich kommt diese Konfiguration auch in der Ebene vor nur ist sie dort extrem unwahrscheinlich.

Der Ockham-Faktor tritt hier und in der z-Statistik in einfacher Gestalt auf. In anspruchsvolleren Datenanalyse-Problemen kann er wesentlich komplexer in Erscheinung treten.

Ockham's Razor liegt den meisten Theorie-Bildungen zugrunde, wenn auch größten Teils unbewusst und aus Mangel an der Kenntnis komplexerer mathematischer Theorien.

23.2 Stimmt der angegebene Mittelwerte bei unbekannter Varianz

Wir betrachten nun den Fall i.u.nv. Zufallsvariablen unbekannter Varianz. Als weiteres Vorwissen soll bekannt sein, dass der Mittelwerte m im Intervall $I = (m_1, m_2)$ liegt. Ebenso seien Obergrenze σ_o und Untergrenze σ_u der Varianz bekannt. Gegeben sei wieder eine Stichprobe $\underline{x} = \{x_1, \ldots, x_N\}$ vom Umfang N.

Die Hypothese H besagt, dass die Stichprobe einer Normal-Verteilung mit vorgegebenem Mittelwert m entstammt. Das heißt, die marginale Likelihood ist

$$p(\underline{x}|H, N, \mathcal{B}) = p(\underline{x}|N, m, \mathcal{B}) = \int p(\underline{x}|m, \sigma, N, \mathcal{B})\; p(\sigma|m, N, \mathcal{B})\; d\sigma \qquad .$$

Die Likelihood $p(\underline{x}|m, \sigma, N, \mathcal{B})$ ist nun vollständig spezifiziert. Wir benötigen noch den Prior für die Varianz. Es handelt sich um eine Skalen-Variable, für die Jeffreys-Prior zutrifft. Allerdings hatten wir aus physikalischen Gründen Grenzwerte $\sigma_{o/u}$ festgelegt. Somit lautet der normiert Jeffreys-Prior

$$\begin{aligned} p_J(\sigma) &= \frac{1}{V_\sigma}\; \theta(\sigma_u \leq \sigma \leq \sigma_o)\; \frac{1}{\sigma} \\ V_\sigma &= \ln\left(\sigma_o/\sigma_u\right) \qquad . \end{aligned} \tag{23.6}$$

Die Marginalisierung über die Varianz kann nun unter Ausnutzung von Gl. (23.3)

durchgeführt werden

$$p(\underline{x}|N,H,\mathcal{B}) = \frac{(2\pi)^{-\frac{N}{2}}}{V_\sigma} \int\limits_{\sigma_u}^{\sigma_o} \sigma^{-N}\, e^{-\frac{N\left(\mathrm{var}(x)+(m-\overline{x})^2\right)}{2\sigma^2}}\, \frac{d\sigma}{\sigma} \qquad .$$

Wir gehen der Einfachheit halber wieder davon aus[2], dass die Daten die Varianz wesentlich stärker einschränken, als die Integrationsgrenzen, so dass wir sie gegen 0 bzw. ∞ schieben können

$$\begin{aligned} p(\underline{x}|N,H,\mathcal{B}) &= \frac{(2\pi)^{-\frac{N}{2}}}{V_\sigma} \int\limits_{0}^{\infty} \sigma^{-N}\, e^{-\frac{N\left(\mathrm{var}(x)+(m-\overline{x})^2\right)}{2\sigma^2}}\, \frac{d\sigma}{\sigma} \\ &= \frac{(2\pi)^{-\frac{N}{2}}}{2\,V_\sigma} \left(\frac{N\ (\mathrm{var}(x)+(m-\overline{x})^2)}{2}\right)^{-\frac{N}{2}} \Gamma(\tfrac{N}{2}) \\ &= \frac{\pi^{-\frac{N}{2}}}{2\,V_\sigma}\ (N\ \mathrm{var}(x))^{-\frac{N}{2}} \left(1+\frac{(m-\overline{x})^2}{\mathrm{var}(x)}\right)^{-\frac{N}{2}} \Gamma(\tfrac{N}{2}) \qquad . \end{aligned}$$

In Anlehnung an den t-Test führen wir die Zufalls-Variable

$$t := \frac{|m-\overline{x}|}{\tilde{\mathrm{SF}}} = \frac{|m-\overline{x}|}{\sqrt{\mathrm{var}(x)/(N-1)}} \tag{23.7}$$

ein und erhalten

$$p(\underline{x}|N,H,\mathcal{B}) = C\,\left(1+\frac{t^2}{N-1}\right)^{-\frac{N}{2}} \qquad . \tag{23.8}$$

Die Konstante C wird später herausfallen. Wir wenden uns nun der alternativen Hypothese $\overline{H}$ zu. In diesem Fall ist der Mittelwert eben nicht bekannt und kann irgendeinen Wert $m \in I$ annehmen. Das heißt, die Prior-Verteilung ist

$$p(m|N,\overline{H},\mathcal{B}) = \frac{1}{V_m}\,\theta(m_1 \le m \le m_1) \qquad .$$

Wir erinnern uns, dass $p(\underline{x}|N,H,\mathcal{B}) = p(\underline{x}|N,m,\mathcal{B})$ ist. Die marginale Likelihood er-

[2] Diese Annahme ist im Rahmen der Wahrscheinlichkeitstheorie nicht notwendig. Sie macht jedoch den Vergleich mit der orthodoxen Statistik transparenter. Wenn man auf diese Näherung verzichtet, erhält man im Endergebnis unvollständige Γ-Funktionen.

halten wir aus der Marginalisierungsregel. Im Fall $\overline{H}$ heißt das

$$\begin{aligned} p(\underline{x}|N,\overline{H},\mathcal{B}) &= \int p(\underline{x}|m,\overline{H},N,\mathcal{B})\; p(m|N,\overline{H},\mathcal{B})\; dm \\ &= \frac{1}{V_m} \int\limits_{m_1}^{m_2} p(\underline{x}|m,\overline{H},N,\mathcal{B})\; dm \end{aligned}$$

In $p(\underline{x}|m,\overline{H},N,\mathcal{B})$ ist die Proposition $\overline{H}$ überflüssig, da der Wert für m bereits explizit angegeben ist. Diese Größe hatten wir in Gl. (23.8) bereits ermittelt. Damit gilt

$$p(\underline{x}|N,\overline{H},\mathcal{B}) = C\; \frac{1}{V_m} \int\limits_{m_1}^{m_2} \left(1 + \frac{(m-\overline{x})^2}{\mathrm{var}(x)}\right)^{-\frac{N}{2}}\; dm$$

Wir schieben auch in diesem Fall die Integrationsgrenzen auf $\mp\infty$. Das ist nicht nötig, macht aber die folgende Analyse transparenter. Im anderen Fall, müssen die Integrale numerisch gelöst werden. Das Integral kann nun analytisch berechnet werden

$$\begin{aligned} p(\underline{x}|N,\overline{H},\mathcal{B}) &= C\; \frac{1}{V_m} \int\limits_{-\infty}^{\infty} \left(1 + \frac{y^2}{\mathrm{var}(x)}\right)^{-\frac{N}{2}}\; dy \\ &= C\; \frac{\sqrt{\mathrm{var}(x)}}{V_m}\; \sqrt{\pi}\; \frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N}{2})} \qquad . \end{aligned}$$

Daraus folgt der Bayes-Faktor

$$o_{\mathrm{BF}} = \frac{V_m}{\sqrt{\frac{\mathrm{var}(x)}{N-1}}}\; \frac{\Gamma(\frac{N}{2})}{\Gamma(\frac{N-1}{2})}\; \frac{1}{\sqrt{\pi\;(N-1)}} \left(1 + \frac{t^2}{N-1}\right)^{-\frac{N}{2}} \qquad .$$

BAYESSCHER t-TEST

$$\begin{aligned} o_{\mathrm{BF}} &= \frac{V_m}{\tilde{\mathrm{SF}}}\; p_t(t|N-1) \\ t &= \frac{m-\overline{x}}{\tilde{\mathrm{SF}}} \\ \tilde{\mathrm{SF}} &= \sqrt{\frac{\mathrm{var}(x)}{N-1}} \end{aligned} \tag{23.9}$$

Das entspricht einer Student-t-Verteilung mit $\nu = N - 1$ Freiheitsgraden. Ein Freiheitsgrad ist mit der Integration über m verlorengegangen.
Es kommt gar nicht auf das Prior-Volumen V_σ an. In diesem Fall können wir also im Limes zum uneigentlichen Prior übergehen und erhalten dennoch ein wohldefiniertes Ergebnis.
Wir haben wieder eine ähnliche Struktur: die Wahrscheinlichkeitsdichte am Wert der zugehörigen Statistik (hier t) multipliziert mit dem Verhältnis der Volumina von Prior und Likelihood.
Auch in diesem Fall liefert die Wahrscheinlichkeitstheorie mit wenig Aufwand und durch dieselbe Vorgehensweise wie in den anderen Fällen die relevante Statistik und die zugehörige Wahrscheinlichkeitsdichte.
Der Unterschied zum Ergebnis der orthodoxen Statistik ist zum einen der Ockham-Faktor. Zum anderen geht hier die Wahrscheinlichkeitsdichte und nicht die Verteilungsfunktion ein. Auch in diesem Fall liegt das Verhältnis Wahrscheinlichkeitsdichte/P-Wert im Bereich $0.8 - 4$ für Stichproben mit Umfang größer 10.

23.3 Sind die Mittelwerte gleich?

Es werde z.B. in einem Massenspektrometer die Masse von zwei unbekannten Spezies unmittelbar hintereinander gemessen. Man weiß, dass die Apparatur häufig nachjustiert werden muss, so dass der Wert der Massen in beiden Fällen einen systematischen Fehler aufweisen kann. Auch verstellt sich die Messgenauigkeit im Laufe der Zeit, so dass auch die Varianz der Fehler-Statistik unbekannt ist. Jedoch sollte die Varianz in beiden Messungen dieselbe sein, da die Messung unmittelbar nacheinander durchgeführt worden sind. Die Fehler seien Gauß-verteilt.
$\mathcal{B}$: *Daten der Stichprobe i.u.nv. mit unbekannter Varianz.*
H : *Die zugrundeliegenden Mittelwerte sind in beiden Fällen gleich.*
Das heißt im angeführten Beispiel, dass die beiden Spezies dieselbe Masse haben.
Die beiden Datensätze seien $\underline{d}^{(\alpha)} = \{d_1^{(\alpha)}, d_2^{(\alpha)}, \ldots, d_{L^\alpha}^{(\alpha)}\}$ mit jeweils L^α Elementen. Wir benötigen die marginale Likelihood

$$p(\underline{d}^{(1)}, \underline{d}^{(2)}|A, \mathcal{B}) \qquad ,$$

für $A \in \{H, \overline{H}\}$.
Wir beginnen mit $A = H$.

23.3.1 Berechnung der marginalen Likelihood zu H

Wie zuvor, verwenden wir die Marginalisierungsregel

$$p(\underline{d}^{(1)}, \underline{d}^{(2)}|H, \mathcal{B}) = \int p(\underline{d}^{(1)}, \underline{d}^{(2)}|m, \sigma, H, \mathcal{B})\; p(m, \sigma|H, \mathcal{B})\; dm\; d\sigma \qquad .$$

Da die Daten gemäß $\mathcal{B}$ unkorreliert sind, wird daraus

$$\begin{aligned}
p(\underline{d}^{(1)}, \underline{d}^{(2)}|H, \mathcal{B}) &= \int p(\underline{d}^{(1)}|m, \sigma, H, \mathcal{B})\; p(\underline{d}^{(2)}|m, \sigma, H, \mathcal{B})\; p(m, \sigma|H, \mathcal{B})\; dm\; d\sigma \\
&= \int \left(2\pi\sigma^2\right)^{-\frac{L^{(1)}+L^{(2)}}{2}}\; e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{L^{(1)}}\left(d_i^{(1)}-m\right)^2+\sum_{i=1}^{L^{(2)}}\left(d_i^{(2)}-m\right)^2\right)} \\
&\qquad \times\; p(m, \sigma|H, \mathcal{B})\; dm\; d\sigma \\
&= (2\pi)^{-\frac{L}{2}} \int d\sigma\; \sigma^{-L}\; p(\sigma|H, \mathcal{B}) \\
&\qquad \times \left(\int dm\; p(m|H, \mathcal{B})\; e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{L^{(1)}}\left(d_i^{(1)}-m\right)^2\right.} \right. \\
&\qquad\qquad \left. \left. +\sum_{i=1}^{L^{(2)}}\left(d_i^{(2)}-m\right)^2\right) \right) \qquad , \qquad (23.10)
\end{aligned}$$

wobei $L = L^{(1)} + L^{(2)}$. Wir haben ausgenutzt, dass der Mittelwert und die Varianz logisch unabhängig sind. Wir formen nun das Argument der Exponential-Funktion um.

$$\begin{aligned}
&\sum_{i=1}^{L^{(1)}}\left(d_i^{(1)}-m\right)^2+\sum_{i=1}^{L^{(2)}}\left(d_i^{(2)}-m\right)^2 = \\
&\quad = \sum_{\alpha=1}^{2}\sum_{i=1}^{L^{(\alpha)}} {d_i^{(\alpha)}}^2 - 2\; m\; \sum_{\alpha=1}^{2}\sum_{i=1}^{L^{(\alpha)}} d_i^{(\alpha)} + L\; m^2 \\
&\quad = L\left(\overline{d^2} - 2\; m\; \overline{d} + m^2\right) \\
&\quad = L\left(\overline{(\Delta d)^2} + (m-\overline{d})^2\right) \qquad .
\end{aligned}$$

Hierbei ist $\overline{d^\nu}$ der Stichproben-Erwartungswert ungeachtet der Experiment-Zuordnung. Wir setzen dieses Ergebnis in das Integral über m in Gl. (23.10) ein. Als Prior verwenden wir mit derselben Begründung wie zuvor den flachen Prior

$$p(m|\mathcal{B}) = \frac{1}{V_m}\; \theta(m_u \le m \le m_o) \qquad .$$

Ebenso schieben wir die Grenzen des Integrals nach $\mp\infty$, da die Beiträge der hinzugefügten Integrationsbereiche vernachlässigbar sind, wenn die Daten aussagekräftig

sind. Damit erhalten wir

$$\begin{aligned}
&\int p(m|H,\mathcal{B})\; e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{L^{(1)}}\left(d_i^{(1)}-m\right)^2+\sum_{i=1}^{L^{(2)}}\left(d_i^{(2)}-m\right)^2\right)}\; dm \\
&\qquad = \frac{1}{V_m} e^{-\frac{L}{2\sigma^2}\,\overline{(\Delta d)^2}} \int_{-\infty}^{\infty} dm\; e^{-\frac{L}{2\sigma^2}\,(m-\overline{d})^2} \\
&\qquad = \frac{1}{V_m}\; e^{-\frac{L}{2\sigma^2}\,\overline{(\Delta d)^2}}\; \sqrt{\frac{2\pi\sigma^2}{L}} \qquad . \qquad (23.11)
\end{aligned}$$

Dieses Ergebnis setzen wir in Gl. (23.10) ein und erhalten mit dem normierten Jeffreys-Prior für σ, nachdem wir wie vorhin die Integrationsgrenzen von σ nach 0 bzw. ∞ geschoben haben,

$$\begin{aligned}
p(\underline{d}^{(1)},\underline{d}^{(2)}|H,\mathcal{B}) &= \frac{(2\pi)^{-\frac{L-1}{2}}}{V_\sigma\; V_m\; \sqrt{L}} \int_0^{\infty} \frac{d\sigma}{\sigma}\; e^{-\frac{L}{2\sigma^2}\,\overline{(\Delta d)^2}}\; \sigma^{-(L-1)} \\
&= \frac{(2\pi)^{-\frac{L-1}{2}}}{2\; V_\sigma\; V_m\; \sqrt{L}} \left(\frac{L}{2}\overline{(\Delta d)^2}\right)^{-\frac{L-1}{2}} \Gamma\left(\tfrac{L-1}{2}\right) \\
&= \frac{\pi^{-\frac{L-1}{2}}\; L^{-\frac{L}{2}}}{2\; V_\sigma\; V_m}\; \Gamma\left(\tfrac{L-1}{2}\right)\; \left(\overline{(\Delta d)^2}\right)^{-\frac{L-1}{2}} \qquad (23.12)
\end{aligned}$$

23.3.2 Berechnung der marginalen Likelihood zu $\overline{H}$

In diesem Fall sind beide Mittelwerte m_α unabhängig voneinander. In Analogie zu Gl. (23.10) erhalten wir

$$\begin{aligned}
p(\underline{d}^{(1)},\underline{d}^{(2)}|\overline{H},\mathcal{B}) = (2\pi)^{-\frac{L}{2}} &\int d\sigma\; p(\sigma|\overline{H},\mathcal{B})\; \sigma^{-L} \\
&\times \left(\int e^{-\frac{1}{2\sigma^2}\,\sum_{i=1}^{L^{(1)}}\left(d_i^{(1)}-m_1\right)^2}\; p(m_1|\overline{H},\mathcal{B})\; dm_1\right. \\
&\times \left.\int e^{-\frac{1}{2\sigma^2}\,\sum_{i=1}^{L^{(2)}}\left(d_i^{(2)}-m_2\right)^2}\; p(m_2|\overline{H},\mathcal{B})\; dm_2\right) \qquad . \qquad (23.13)
\end{aligned}$$

Die beiden Integrale über m_α sind beide vom selben Typ wie Gl. (23.11). Wir müssen lediglich wahlweise $L^{(1)}$ oder $L^{(2)}$ auf Null setzen und erhalten

$$\int e^{-\frac{1}{2\sigma^2}\,\sum_{i=1}^{L^{(\alpha)}}\left(d_i^{(\alpha)}-m_\alpha\right)^2}\; p(m_\alpha,\sigma|\overline{H},\mathcal{B})\; dm_\alpha = \frac{1}{V_m}\sqrt{\frac{2\pi}{L^{(\alpha)}}}\; e^{-\frac{L^{(\alpha)}}{2\sigma^2}\,\overline{(\Delta d^{(\alpha)})^2}}\; \sigma$$

Damit wird Gl. (23.13) zu

$$
\begin{aligned}
&p(\underline{d}^{(1)}, \underline{d}^{(2)} | \overline{H}, \mathcal{B}) \\
&= \frac{(2\pi)^{-\frac{L-2}{2}}}{V_\sigma \, V_m^2 \, \sqrt{L^{(1)} L^{(2)}}} \int\limits_0^\infty \frac{d\sigma}{\sigma} \; \sigma^{-(L-2)} \; e^{-\frac{1}{2\sigma^2} \left(L^{(1)} \, \overline{(\Delta d^{(1)})^2} + L^{(2)} \, \overline{(\Delta d^{(2)})^2} \right)} \\
&= \frac{\pi^{-\frac{L-2}{2}}}{2 \, V_\sigma \, V_m^2 \, \sqrt{L^{(1)} L^{(2)}}} \left(L^{(1)} \, \overline{(\Delta d^{(1)})^2} + L^{(2)} \, \overline{(\Delta d^{(2)})^2} \right)^{-\frac{L-2}{2}} \Gamma(\tfrac{L-2}{2}) \qquad . \qquad (23.14)
\end{aligned}
$$

Das Verhältnis von Gl. (23.12) zu Gl. (23.14) liefert den Bayes-Faktor

$$
\begin{aligned}
o_{BF} &= \frac{\frac{\pi^{-\frac{L-1}{2}} \, L^{-\frac{L}{2}}}{2 \, V_\sigma \, V_m} \, \Gamma(\tfrac{L-1}{2}) \, \left(\overline{(\Delta d)^2} \right)^{-\frac{L-1}{2}}}{\frac{\pi^{-\frac{L-2}{2}}}{2 \, V_\sigma \, V_m^2 \, \sqrt{L^{(1)} L^{(2)}}} \left(L^{(1)} \, \overline{(\Delta d^{(1)})^2} + L^{(2)} \, \overline{(\Delta d^{(2)})^2} \right)^{-\frac{L-2}{2}} \Gamma(\tfrac{L-2}{2})} \\
&= \frac{V_m \, \sqrt{L^{(1)} L^{(2)}} \, L^{-\frac{L}{2}} \, \Gamma(\tfrac{L-1}{2})}{\sqrt{\pi} \, \Gamma(\tfrac{L-2}{2})} \, \frac{\left(\overline{(\Delta d)^2} \right)^{-\frac{L-1}{2}}}{\left(L^{(1)} \, \overline{(\Delta d^{(1)})^2} + L^{(2)} \, \overline{(\Delta d^{(2)})^2} \right)^{-\frac{L-2}{2}}} \\
&= V_m \, \frac{\Gamma(\frac{L-1}{2})}{\Gamma(\frac{L-2}{2})} \, \frac{1}{\sqrt{\pi (L-2)}} \, \frac{\sqrt{L-2} \, \sqrt{L^{(1)} L^{(2)}}}{\sqrt{L}} \, \times \\
&\quad \frac{\left(L \, \overline{(\Delta d)^2} \right)^{-\frac{L-1}{2}}}{\left(L^{(1)} \, \overline{(\Delta d^{(1)})^2} + L^{(2)} \, \overline{(\Delta d^{(2)})^2} \right)^{-\frac{L-2}{2}}} \qquad .
\end{aligned}
$$

In weiser Voraussicht führen wir die Abkürzung

$$
\tilde{\sigma}^2 = \frac{L}{L^{(1)} L^{(2)}} \left(L^{(1)} \, \overline{(\Delta d^{(1)})^2} + L^{(2)} \, \overline{(\Delta d^{(2)})^2} \right) \qquad (23.15)
$$

ein. Damit vereinfacht sich der obige Ausdruck zu

$$
o_{BF} = \frac{V_m}{\frac{\tilde{\sigma}}{\sqrt{L-2}}} \, \frac{\Gamma(\frac{L-1}{2})}{\Gamma(\frac{L-2}{2})} \, \frac{1}{\sqrt{\pi (L-2)}} \left(\frac{L^2 \, \overline{(\Delta d)^2}}{L^{(1)} \, L^{(2)} \, \tilde{\sigma}^2} \right)^{-\frac{L-1}{2}} \qquad . \qquad (23.16)
$$

Wegen $\overline{\Delta d^2} = \overline{d^2} - \overline{d}^2$ gilt

$$
\begin{aligned}
L^2\ \overline{(\Delta d)^2} &= L^2\ \overline{d^2} - L^2\ \overline{d}^2 \\
&= L\ \left(L^{(1)}\overline{d^{(1)}{}^2} + L^{(2)}\overline{d^{(2)}{}^2}\right) - \left(L^{(1)}\overline{d^{(1)}} + L^{(2)}\overline{d^{(2)}}\right)^2 \\
&= L\ \left(L^{(1)}\overline{\Delta d^{(1)}{}^2} + L^{(2)}\overline{\Delta d^{(2)}{}^2}\right) + L\ \left(L^{(1)}\overline{d^{(1)}}^2 + L^{(2)}\overline{d^{(2)}}^2\right) \\
&\quad - \left(L^{(1)}\overline{d^{(1)}} + L^{(2)}\overline{d^{(2)}}\right)^2 \\
&= L^{(1)}\ L^{(2)}\ \tilde{\sigma}^2 + \left(L^{(1)} + L^{(2)}\right)\ \left(L^{(1)}\overline{d^{(1)}}^2 + L^{(2)}\overline{d^{(2)}}^2\right) \\
&\quad - \left(L^{(1)}\overline{d^{(1)}} + L^{(2)}\overline{d^{(2)}}\right)^2 \\
&= L^{(1)}\ L^{(2)}\ \tilde{\sigma}^2 + L^{(1)}\ L^{(2)}\ \left(\overline{d^{(2)}} - \overline{d^{(1)}}\right)^2 \qquad .
\end{aligned}
$$

Damit haben wir schließlich

$$
\frac{L^2\ \overline{(\Delta d)^2}}{L^{(1)}\ L^{(2)}\ \tilde{\sigma}^2} = 1 + \frac{\left(\overline{d^{(2)}} - \overline{d^{(1)}}\right)^2}{\tilde{\sigma}^2} \qquad .
$$

Das setzen wir noch in Gl. (23.16) ein und erhalten

$$
o_{BF} = \frac{V_m}{\frac{\tilde{\sigma}}{\sqrt{L-2}}}\ \frac{\Gamma(\frac{L-1}{2})}{\Gamma(\frac{L-2}{2})}\ \frac{1}{\sqrt{\pi(L-2)}}\ \left(1 + \frac{\left(\overline{d^{(2)}} - \overline{d^{(1)}}\right)^2}{\tilde{\sigma}^2}\right)^{-\frac{L-1}{2}} \qquad .
$$

Damit ist auch hier der Kontakt zur orthodoxen Statistik hergestellt. Es liegen $\nu = L-2$ Freiheitsgrade vor. Die t-Statistik für die Differenz zweier Stichproben-Mittelwert war

$$
t^2 = \frac{\left(\overline{d^{(2)}} - \overline{d^{(1)}}\right)^2}{\tilde{\sigma}^2/(L-2)} \qquad .
$$

Wir identifizieren

$$
\tilde{\mathrm{SF}} = \frac{\tilde{\sigma}}{\sqrt{L-2}}
$$

mit dem modifizierten Standard-Fehler der Differenz der Mittelwerte und erhalten schließlich

$$
o_{BF} = \frac{V_m}{\tilde{\mathrm{SF}}}\ p_T(t|\nu = L-2)
$$

formal dasselbe Ergebnis wie im vorherigen Abschnitt nur mit den entsprechend angemaßten Ausdrücken für t und SF und einem Freiheitsgrad weniger, da nun zwei Integrale über m_α durchgeführt worden sind.
Es sei noch erwähnt, dass $\tilde{\sigma}/\sqrt{L-2}$ den unverzerrten Schätzwert für den Standard-Fehler der Differenz der Stichproben-Mittelwerte darstellt.

23.4 Sind die Varianzen gleich, ungeachtet der Mittelwerte?

Gegeben seien zwei Stichproben $\underline{d}^{(\alpha)}$ vom Umfang L_α, $\alpha = 1, 2$. Die Elemente der beiden Stichproben seien i.u.nv.. Allerdings sollen die zugehörigen Mittelwert unbekannt sein. Die a-priori Wahrscheinlichkeit des Mittelwerts m sei für beide Stichproben

$$p(m) = \frac{1}{V_m}\, \theta(m_u \leq m \leq m_o) \qquad . \tag{23.17}$$

Ebenso soll der möglichen Bereich der Varianzen für beide Stichproben eingeschränkt sein

$$\sigma_u \leq \sigma \leq \sigma_o \tag{23.18}$$

Die Hypothese H, die uns hier interessiert, besagt

$$H : \textit{Die intrinsischen Varianzen beider Stichproben sind gleich.}$$

Der Wert der intrinsischen Varianz ist allerdings nicht bekannt. Als Prior muss die Jeffreys-Verteilung Gl. (23.6) gewählt werden. Die marginale Likelihood unter der Hypothese H lautet

$$p(\underline{d}^{(1)}, \underline{d}^{(2)} | L_1, L_2, H, \mathcal{B}) \qquad .$$

Um die Likelihood zu vervollständigen, müssen wir über die Marginalisierungsregel die beiden Mittelwert m_1 und m_2 und die gemeinsame Standardabweichung σ einführen

$$\begin{aligned} p(\underline{d}^{(1)}, \underline{d}^{(2)} | L_1, L_2, H, \mathcal{B}) &= \int p(\underline{d}^{(1)}, \underline{d}^{(2)} | m_1, m_2, \sigma, L_1, L_2, H, \mathcal{B}) \\ &\quad \times\; p(m_1, m_2, \sigma | \mathcal{B})\; dm_1\; dm_2\; d\sigma \\ &= \int p(\underline{d}^{(1)} | m_1, \sigma, L_1, \mathcal{B})\; p(\underline{d}^{(2)} | m_2, \sigma, L_2, \mathcal{B}) \\ &\quad \times\; p(m_1 | \mathcal{B})\; p(m_2 | \mathcal{B})\; p(\sigma | \mathcal{B})\; dm_1\; dm_2\; d\sigma \qquad . \end{aligned} \tag{23.19}$$

Wir nutzen nun aus, dass die Likelihood-Funktion gemäß Gl. (23.3) in zwei Teile faktorisiert

$$p(\underline{d}^{(\alpha)} | m_\alpha, \sigma, L_\alpha, \mathcal{B}) = (2\pi\sigma^2)^{-\frac{L_\alpha}{2}}\; e^{-\frac{v_\alpha}{2\sigma^2}}\; e^{-\frac{L_\alpha}{2\sigma^2}\,(m_\alpha - \overline{d^{(\alpha)}})^2} \qquad .$$

Wir haben der Übersicht halber die Abkürzung

$$v_\alpha = L_\alpha\, \mathrm{var}(d^{(\alpha)})$$

eingeführt. Wir können deshalb die Integrale über m_1 und m_2 ausführen. Wie zuvor erstrecken wir die Integrale wieder bis $\mp\infty$. Wir werden hier auch die Varianzen mit

einem Index versehen, da diese Unterscheidung später benötigt wird.

$$\begin{aligned}
&\int p(\underline{d}^{(\alpha)}|m_\alpha,\sigma_\alpha,L_\alpha,\mathcal{B})\; p(m_\alpha|\mathcal{B})\; dm_\alpha \\
&\quad = \frac{1}{V_m}\;(2\pi\sigma_\alpha^2)^{-\frac{L_\alpha}{2}}\; e^{-\frac{v_\alpha}{2\sigma_\alpha^2}} \int\limits_{-\infty}^{\infty} e^{-\frac{L_\alpha}{2\sigma_\alpha^2}\,(m_\alpha-\overline{d^{(\alpha)}})^2}\; dm_\alpha \\
&\quad = \frac{1}{V_m}\;(2\pi\sigma_\alpha^2)^{-\frac{L_\alpha}{2}}\; e^{-\frac{v_\alpha}{2\sigma_\alpha^2}}\; \sqrt{2\pi\frac{\sigma_\alpha^2}{L_\alpha}} \\
&\quad = \frac{(2\pi)^{-\frac{L_\alpha-1}{2}}}{V_m\,\sqrt{L_\alpha}}\;\sigma_\alpha^{-(L_\alpha-1)}\; e^{-\frac{v_\alpha}{2\sigma_\alpha^2}} \qquad .
\end{aligned}$$

Demnach haben wir

$$\int p(\underline{d}^{(\alpha)}|m_\alpha,\sigma_\alpha,L_\alpha,\mathcal{B})\; p(m_\alpha|\mathcal{B})\; dm_\alpha = C_\alpha\;\sigma_\alpha^{-(L_\alpha-1)}\; e^{-\frac{v_\alpha}{2\sigma_\alpha^2}} \qquad .$$

Die Konstanten C_α fallen im Odds-Ratio heraus. Nun können wir die Integration über σ in Gl. (23.19) ausführen

$$\begin{aligned}
&p(\underline{d}^{(1)},\underline{d}^{(2)}|L_1,L_2,H,\mathcal{B}) \\
&\quad = \frac{C_1C_2}{V_\sigma}\int\limits_{\sigma_u}^{\sigma_o}\sigma^{-(L_1+L_2-2)}\; e^{-\frac{(v_1+v_2)}{2\sigma^2}}\;\frac{d\sigma}{\sigma} \\
&\quad = \frac{C_1C_2}{2V_\sigma}\;2^{\frac{L_1+L_2-2}{2}}\;(v_1+v_2)^{-\frac{L_1+L_2-2}{2}}\;\Gamma(\tfrac{L_1+L_2-2}{2})
\end{aligned}$$

Wir haben die Integration mit derselben Begründung wie zuvor über $(0,\infty)$ erstreckt. Die Alternative zu H lautet gemäß des Bedingungskomplexes, dass die beiden Varianzen unabhängig voneinander Jeffreys-verteilt sind. Die zugehörigen Integrale faktorisieren

$$\begin{aligned}
&p(\underline{d}^{(1)},\underline{d}^{(2)}|L_1,L_2,\overline{H},\mathcal{B}) \\
&\quad = \prod_\alpha \frac{C_\alpha}{V_\sigma}\int\limits_{\sigma_u}^{\sigma_o}\sigma^{-(L_\alpha-1)}\; e^{-\frac{v_\alpha}{2\sigma^2}}\;\frac{d\sigma}{\sigma} \\
&\quad = \frac{C_1C_2}{4V_\sigma^2}\;2^{\frac{L_1+L_2-2}{2}}\prod_\alpha v_\alpha^{-\frac{L_\alpha-1}{2}}\;\Gamma(\tfrac{L_\alpha-1}{2}) \qquad .
\end{aligned}$$

Der Bayes-Faktor lautet somit

$$o_{\mathrm{BF}} = \frac{2V_\sigma}{(v_1+v_2)^{\frac{L_1+L_2-2}{2}}} \frac{v_1^{\frac{L_1-1}{2}}\, v_2^{\frac{L_2-1}{2}}}{\Gamma(\frac{L_1-1}{2})\,\Gamma(\frac{L_2-1}{2})} \frac{\Gamma(\frac{L_1+L_2-2}{2})}{}$$

$$= 2V_\sigma \frac{1}{B(\frac{L_1-1}{2}, \frac{L_2-1}{2})} \left(\frac{v_1}{v_2}\right)^{\frac{L_1-1}{2}} \left(1+\frac{v_1}{v_2}\right)^{-\frac{L_1+L_2}{2}+1} \quad .$$

Um den Kontakt zur F-Statistik noch klarer zu machen, benutzen wir die Definition

$$f = \frac{L_2-1}{L_1-1} \frac{L_1\ \mathrm{var}(x)_1}{L_2\ \mathrm{var}(x)_2} = \frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_2^2} \quad ,$$

die das Verhältnis der unverzerrten Schätzwerte der Varianzen darstellt. Mit der zusätzlichen Abkürzung

$$r = \frac{L_1-1}{L_2-1}$$

gilt $\frac{v_1}{v_2} = rf$, und somit haben wir

BAYESSCHER F-TEST

$$\begin{aligned} o_{\mathrm{BF}} &= 2V_\sigma\ f\ \frac{f^{-1}}{B(\frac{L_1-1}{2}, \frac{L_2-1}{2})}\ (rf)^{\frac{L_1-1}{2}}\ (1+rf)^{-\frac{L_1+L_2}{2}+1} \\ &= 2V_\sigma\ f\ p_F(f|\nu_1 = L_1-1, \nu_2 = L_2-1) \quad . \end{aligned}$$

Hier taucht auch die F-Statistik als relevante (sufficient) Statistik auf. Der Ockham-Faktor sieht etwas anders als in den bisher untersuchten Fällen.

23.4.1 Beispiel: Hat sich die Messapparatur verstellt?

Eine Messapparatur habe, wenn sie optimal funktioniert, eine Gaußsche Fehler-Statistik mit konstanter Varianz und Mittelwert Null. Das heißt, die Messwerte werden mit fester aber unbekannter Varianz um den Wert einer physikalischen Größe streuen. Es werden an zwei Substanzen Messungen durchgeführt (Stichproben entnommen) mit jeweils $L_\alpha = 20$ Messwerten. Man möchte neben der Analyse der physikalischen Fragestellung zudem wissen, ob die Apparatur noch optimal funktioniert, d.h. in beiden Fällen dieselbe Varianz vorliegt. Die Erfahrung im Umgang mit der Messapparatur besagt, dass sie, wenn sie außer Kontrolle gerät,

Standardabweichungen mit $\sigma_o/\sigma_u = 100$ liefern kann. D.h.

$$V_\sigma = \ln(\frac{\sigma_o}{\sigma_u}) = 4.605 \qquad .$$

Ebenso wissen wir aus langjähriger Erfahrung mit der Apparatur, dass die Ausfall-Wahrscheinlichkeit $ca.1\%$ beträgt

$$o_P = \frac{P(H|\mathcal{B})}{P(\overline{H}|\mathcal{B})} = \frac{0.99}{0.01} = 99 \qquad .$$

Das Odds-Ratio ist somit

$$o = 99 \times 2 \times 4.6 \times \ f\ p_F(f|L_1 - 1, L_2 - 1) = 911.8\ fp_F(f|L_1 - 1, L_2 - 1) \qquad .$$

Das Ergebnis der Experimente soll sein $\tilde{\sigma}_1 = 4.4$ und $\tilde{\sigma}_2 = 2.5$. Daraus ergibt sich

$$f = \left(\tfrac{4.4}{2.5}\right)^2 = 3.1 \qquad .$$

Der P-Wert ergibt sich aus

$$P = 1 - F_F(3.1|\nu_1 = 19, \nu_2 = 19) = 0.0088.$$

Das heißt, die Hypothese ist bei den üblichen Signifikanz-Niveaus (1%, 5%) zu verwerfen. Der F-Test besagt, die Abweichungen des F-Werte von eins sind signifikant, die Apparatur hat sich verstellt.
Die wahrscheinlichkeitstheoretische Analyse ergibt hingegen

$$o = 911.8 * 3.1 * p_F(3.1|\nu_1 = 19, \nu_2 = 19) = 911.8 * 3.1 * 0.015 = 43.5 \qquad .$$

Daraus resultiert

$$P(H|D, \mathcal{B}) = \frac{o}{1 + o} = 0.978 \qquad .$$

Diese Wahrscheinlichkeit ist so nahe an eins, dass man davon ausgehen kann, dass die Apparatur noch einwandfrei funktioniert.

Kapitel 24

Modell-Vergleich

Gegeben sind zwei oder mehrere Modelle M_α, die zu vergleichen sind. Der Bedingungskomplex beinhaltet jegliche Information über die Art und Definition des zu untersuchenden wissenschaftlichen Problems, das durch die Modelle beschrieben werden soll. Damit sind auch die Prior-Wahrscheinlichkeiten $P(M_\alpha|\mathcal{B})$ spezifiziert. Der Bedingungskomplex definiert Wahrscheinlichkeitsraum, Borel-Körper, Prior-Wahrscheinlichkeiten und Grundgesamtheit. Ein Polynom niedrigen Grades mag gut sein zur Beschreibung von Trajektorien aber völlig ausgeschlossen sein, wenn Zeitreihen periodischer Prozesse analysiert werden. Im Gegensatz zu herkömmlichen statistischen Methoden erlaubt es die Wahrscheinlichkeitstheorie, diese essentiellen Zusatzinformationen explizit und konsistent zu berücksichtigen.
Wie beim Hypothesen-Vergleich, die Unterschiede sind sehr gering, betrachten wir das Odds-Ratio

$$o_{21} = \frac{P(M_2|D,\mathcal{B})}{P(M_1|D,\mathcal{B})} = \frac{p(D|M_2,\mathcal{B})\ P(M_2|\mathcal{B})}{p(D|M_1,\mathcal{B})\ P(M_1|\mathcal{B})} \quad .$$

Wir haben wieder den Bayes-Faktor und den Prior-Odds-Faktor. Wenn es mehr als zwei Modelle gibt, berechnen sich die Wahrscheinlichkeiten etwas anders. Angenommen, es gäbe L alternative Modelle und wir haben alle o_{i1} ($i = 2, 3, \ldots, L$) berechnet. Aus der Normierung folgt

$$\begin{aligned} 1 &= \sum_{i=1}^{L} P(M_i|D,\mathcal{B}) \\ &= P(M_1|D,\mathcal{B}) \ + \ \sum_{i=2}^{L} o_{i1}\ P(M_1|D,\mathcal{B}) \\ &= P(M_1|D,\mathcal{B}) \ \left(1 + \ \sum_{i=2}^{L} o_{i1}\right) \quad . \end{aligned}$$

Daraus folgt

$$P(M_1|D,\mathcal{B}) = \frac{1}{1 + \ \sum_{i=2}^{L} \ o_{i1}} \quad . \tag{24.1}$$

Gemessen werden die Größen $d \in \mathbb{R}^{N_d}$. Das können mehr-dimensionale Objekte sein. Die Steuergrößen seien $s \in \mathbb{R}^{N_s}$. Zudem werden die Modelle von Parametern $x \in \mathbb{R}^{n_\alpha}$ abhängen, deren Anzahl wiederum vom jeweiligen Modell abhängen wird. Man will z.B. testen, ob die Daten durch eine Gerade oder eine Parabel zu beschreiben sind. Es besteht die Beziehung

$$d = f(s|x) + \eta \qquad . \tag{24.2}$$

Die Größe $\eta \in \mathbb{R}^{N_d}$ ist der Messfehler. Die Wahrscheinlichkeitstheorie erlaubt die Behandlung beliebiger Fehler-Verteilungen. In Anlehnung an weit verbreitete Verfahren der orthodoxen Statistik gehen wir davon aus, dass die Fehler i.u.nv. sind.
Es werde eine Stichprobe $\underline{d} = \{d_1, \ldots, d_N\}$ vom Umfang N zu den Steuergrößen $\underline{s} = \{s_1, \ldots, s_N\}$ ermittelt. Die marginale Likelihood-Funktion, die in die Bayes-Faktoren eingeht, ist somit

$$p(\underline{d}|\underline{s}, M_\alpha, \mathcal{B}) \qquad .$$

Damit die Ausdrücke übersichtlich bleiben, haben wir die Größen N, N_d, N_s, n_α mit in den Bedingungskomplex aufgenommen, da sie sich während der gesamten Rechnung ohnehin nicht ändern. Wir müssen nun unterscheiden, ob die Fehler-Varianz durch das Experiment gegeben ist, oder nicht. Wir behandeln hier nur den Fall bekannter Varianzen.

24.1 Bekannte Varianzen

Wir fangen damit an, dass σ auch bekannt ist. Die marginale Likelihood lautet dann

$$p(\underline{d}|\sigma, \underline{s}, M_\alpha, \mathcal{B}) \qquad .$$

Um die marginale Likelihood berechnen zu können, benötigen wir die noch nicht spezifizierten Modell-Parameter x, die wir über die Marginalisierungsregel einführen

$$\begin{aligned} & p(\underline{d}|\sigma, \underline{s}, M_\alpha, \mathcal{B}) \\ & \quad = \int p(\underline{d}|x, \sigma, \underline{s}, M_\alpha, \mathcal{B})\; p(x|\sigma, \underline{s}, M_\alpha, \mathcal{B})\; dx \\ & \quad = \int p(\underline{d}|x, \sigma, \underline{s}, \mathcal{B})\; p(x|M_\alpha, \mathcal{B})\; dx \qquad . \end{aligned}$$

In einigen Problemen lassen sich diese Integrale exakt analytisch lösen. Wir wollen hier alternativ eine weitverbreitete Näherungsmethode vorstellen.

24.1.1 Steepest Descent Näherung

Hierzu schreiben wir den Integranden um in

$$p(\underline{d}|x, \sigma, \underline{s}, M_\alpha, \mathcal{B})\; p(x|\sigma, \underline{s}, M_\alpha, \mathcal{B}) = e^{\Phi(x)} \qquad . \tag{24.3}$$

Dieser Ansatz ist exakt und naheliegend, da Likelihood-Funktionen häufig exponentielle Funktionen sind. Wir stellen fest, dass der Integrand

$$\begin{aligned} & p(\underline{d}|x, \sigma, \underline{s}, M_\alpha, \mathcal{B})\ p(x|\sigma, \underline{s}, M_\alpha, \mathcal{B}) \\ & \quad = p(x, \underline{d}|\sigma, \underline{s}, M_\alpha, \mathcal{B}) \\ & \quad = p(x|\underline{d}, \sigma, \underline{s}, M_\alpha, \mathcal{B})\ p(\underline{d}, |\sigma, \underline{s}, M_\alpha, \mathcal{B}) \end{aligned}$$

was die x-Abhängigkeit angeht, proportional zur A-Posteriori-Wahrscheinlichkeit

$$p(x|\underline{d}, \sigma, \underline{s}, M_\alpha, \mathcal{B})$$

ist. Wir maximieren zunächst die A-Posteriori-Wahrscheinlichkeit bzgl. x und erhalten so die MAP-Lösung x^{MAP}.
Nun entwickeln wir die Funktion $\Phi(x)$ um das Maximum bis zu quadratischen Termen

$$\begin{aligned} \Phi(x) &\simeq \Phi^{\mathrm{MAP}} + \frac{1}{2}\ \Delta x^T\ H\ \Delta x \\ \Delta x &= x - x^{\mathrm{MAP}} \\ H^\alpha_{ij} &= \left.\frac{\partial^2 \Phi(x)}{\partial x_i \partial x_j}\right|_{x=x^{\mathrm{MAP}}} \end{aligned} \quad . \tag{24.4}$$

Die Matrix H ist die Hesse-Matrix, die sowohl in der Dimension, als auch in der MAP-Lösung vom Modell α abhängt. Die Prior-Wahrscheinlichkeit beinhaltet i.d.R. noch Schranken für die erlaubten Parameter. In der Steepest-Descent-Näherung setzt man voraus, dass die A-Posteriori-Wahrscheinlichkeit an den Schranken bereits so vernachlässigbar klein ist, so dass man die verbleibenden Integrale über den gesamten $\mathbb{R}^{n_\alpha}$ erstrecken kann. Somit lautet die gesuchte Likelihood

$$\begin{aligned} & p(\underline{d}|\sigma, \underline{s}, M_\alpha, \mathcal{B}) \\ & \quad = p(\underline{d}|x_\alpha^{\mathrm{MAP}}, \sigma, \underline{s}, \mathcal{B})\ p(x_\alpha^{\mathrm{MAP}}|M_\alpha, \mathcal{B})\ \int\ e^{-\frac{1}{2}\ \Delta x^T H^\alpha \Delta x}\ dx \\ & \quad = p(\underline{d}|x_\alpha^{\mathrm{MAP}}, \sigma, \underline{s}, \mathcal{B})\ p(x_\alpha^{\mathrm{MAP}}|M_\alpha, \mathcal{B})\ (2\pi)^{\frac{n_\alpha}{2}}\ \ |H^\alpha|^{-\frac{1}{2}} \end{aligned} \quad . \tag{24.5}$$

Wir können nun das Odds-Ratio der Modelle 1,2 angeben

$$o_{21} = \frac{p(\underline{d}|x_2^{\mathrm{MAP}}, \sigma, \underline{s}, \mathcal{B})\ p(x_2^{\mathrm{MAP}}|M_2, \mathcal{B})\ (2\pi)^{\frac{n_2}{2}}\ \ |H^2|^{-\frac{1}{2}}}{p(\underline{d}|x_1^{\mathrm{MAP}}, \sigma, \underline{s}, \mathcal{B})\ p(x_1^{\mathrm{MAP}}|M_1, \mathcal{B})\ (2\pi)^{\frac{n_1}{2}}\ \ |H^1|^{-\frac{1}{2}}}\ \frac{P(M_2|\mathcal{B})}{P(M_1|\mathcal{B})} \quad .$$

Für eine qualitative Diskussion ist es sinnvoll, einen flachen Prior anzunehmen. Die MAP-Lösung geht dann über in die Maximum-Likelihood (ML) Lösung. In diesem Fall ist

$$p(x_\alpha^{\mathrm{MAP}}|M_\alpha, \mathcal{B}) = \frac{1}{V_\alpha^P}$$

gleich dem inversen Prior-Volumen. Das ist der Bereich im Parameter-Raum, den der Prior zulässt. Dieses Volumen wird bei komplexen Modellen größer sein als bei einfachen Modellen. Außerdem kann man

$$(2\pi)^{\frac{n_\alpha}{2}} \ |H^\alpha|^{-\frac{1}{2}} = V_\alpha^L$$

als effektives Volumen betrachten, das von der Likelihood erlaubt ist. Das sieht man folgendermaßen

$$(2\pi)^{\frac{n_\alpha}{2}} \ |H^\alpha|^{-\frac{1}{2}} = \int e^{-\frac{1}{2}\,\Delta x^T H^\alpha \Delta x}\, dx = \int\limits_{V_\alpha^L} dx \qquad .$$

Damit ist das Odds-Ratio

$$o_{21} = \frac{p(\underline{d}|x_2^{\mathrm{ML}}, \sigma, \underline{s}, \mathcal{B})}{p(\underline{d}|x_1^{\mathrm{ML}}, \sigma, \underline{s}, \mathcal{B})} \ \frac{V_2^L \ V_1^P}{V_2^P \ V_1^L} \ \frac{P(M_2|\mathcal{B})}{P(M_1|\mathcal{B})} \qquad . \tag{24.6}$$

Es besteht aus dem Verhältnis der Likelihood-Funktionen am ML-Wert, dem Ockham-Faktor und dem Prior-Odds. Der Ockham-Faktor gibt das Verhältnis der Volumina von Likelihood zu Prior der beiden Modelle an. Es sei L^P das effektive Intervall des Priors, so dass

$$V_\alpha^P = (L^P)^{n_\alpha} \qquad .$$

Analog führen wir für die Likelihood-Funktion eine effektive Länge L^L ein, mit der

$$V_\alpha^L = (L^L)^{n_\alpha}$$

gilt. Dann wird der Ockham-Faktor zu

$$o_{\mathrm{Ockham}} = \left(\frac{L^L}{L^P}\right)^{n_2 - n_1} \qquad .$$

Der Daten-Konstraint wird immer dafür sorgen, dass $\frac{L^L}{L^P} \ll 1$. Wir wählen das Modell zwei als das komplexere, d.h. dasjenige mit mehr Freiheitsgraden $n_2 > n_1$. Dann ist offensichtlich, dass der Ockham-Faktor Komplexität bestraft. Das gleiche ist zu beobachten, wenn das komplexere Modell zwar dieselbe Zahl an Freiheitsgraden besitzt, aber der Prior ein größeres Volumen erlaubt $V_2^P > V_1^P$.
Soweit haben wir die anfangs gemachte Annahme, dass die Daten i.u.nv. sind überhaupt nicht benutzt. Bisher war alles allgemein. Wir wollen nun ausnutzen, dass die Daten d_i i.u.nv. um den wahren Wert

$$y_i = f(s_i|x)$$

streuen. Weiterhin gehen wir von schwachen Prior-Verteilungen aus, so dass Gl. (24.6) gültig bleibt. Die Likelihood wird dann zu

$$\begin{aligned} p(\underline{d}|x_\alpha^{\mathrm{ML}}, \sigma, \underline{s}, \mathcal{B}) &= (2\pi)^{-\frac{N_d}{2}} \ \sigma^{-N_d} \ e^{-\frac{v_\alpha}{2\sigma^2}} \\ v_\alpha &= \sum_{i=1}^{N_d} \left(d_i - f(s_i|x_\alpha^{\mathrm{ML}})\right)^2 \end{aligned} \qquad . \tag{24.7}$$

Wenn wir das in das Odds-Ratio einsetzen erhalten wir

$$o_{21} = e^{-\frac{(v_2 - v_1)}{2\sigma^2/N}} \frac{V_2^L}{V_2^P} \frac{V_1^P}{V_1^L} \frac{P(M_2|\mathcal{B})}{P(M_1|\mathcal{B})} \quad .$$

Im vorliegenden Fall, wo die Likelihood eine Normal-Verteilung und der Prior vergleichsweise flach ist, ist die Steepest-Descent-Näherung keine Näherung, sondern exakt.

Teil VI

Literatur

Literatur

1. Devinder S. Sivia: DATA ANALYSIS: A BAYESIAN TUTORIAL, Oxford Science Publications (1998).

2. Dieter Wickmann: BAYES-STATISTIK,B.I. Wissenschaftsverlag, Mannheim/Wien/Zürich (1990).

3. B. Roy Frieden: PROBABILITY, STATISTICAL OPTICS, AND DATA TESTING, Springer-Verlag (1991).

4. Herbert Meschkowski: WAHRSCHEINLICHKEITSTHEORIE, B.I. Hochschultaschenbücher, Mannheim/Wien/Zürich (1968).

5. Athanasios Papoulis PROBABILITY, RANDOM VARIABLES, AND STOCHASTIC PROCESSES, McGraw-Hill (1984).

6. Siegmund Brandt: DATENANALYSE, Spektrum Akademischer Verlag, Heidelberg/Berlin (1999).

7. John P. Taylor: AN INTRODUCTION TO ERROR ANALYSIS, University Science Books, Sausolito (1997).

8. Brian Buck and Vincent A, Macaulay: MAXIMUM ENTROPY IN ACTION, Oxford Science Publications (1991).

9. J.N. Kapur and H.K. Kesavan ENTROPY OPTIMIZATION PRINCIPLES WITH APPLICATIONS, Academic Press (1992).

10. Darrell Huff HOW TO LIE WITH STATISTICS, W.W. Norton & Company, New York/ London, (1993)

11. David Freedman, Robert Pisani, Roger Purves STATISTICS, W.W. Norton & Company, New York (1998).

12. Myron Tribus RATIONAL DESCRIPTIONS, DECISIONS, AND DESIGNS, Pergamon Press, (1969).

13. Norman T.J. Bailey THE ELEMENTS OF STOCHASTIC PROCESSES, John Wiley& Sons, (1990).

14. William Feller AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS I/II, John Wiley& Sons, (1957).

15. A.O'Hagan: KANDALL'S ADVANCED THEORY OF STATISTICS 2B, 'BAYESIAN INFERENCE', Halsted Press (1994).

16. W.T. Grandy,jr NEW FOUNDATION OF STATISTICAL MECHANICS I/II, D.Reidel Publishing Company, Member of Kluwer Academic Publishers.

17. E.T. Jaynes Probability Theory: The Logic of Science, Bayes.wustl.edu, dir: pub/jaynes.book