Statistische Inferenz

Rainer Schlittgen

August 2010

Kennworte:

Mathematische Statistik, Parameterschätzung, Punktschätzung, Konfidenzintervall, optimales Konfidenzintervall, statistischer Test, multiple Tests, optimaler Test, statistische Information

Vorwort

Vorwort zur 2. Auflage

Für diese Auflage wurde der Text insgesamt durchgesehen, bekannte Fehler wurden korrigiert. Das Layout wurde erneuert. Zudem wurden Aufgaben, zum größten Teil mit Lösungen, aufgenommen. Für die intensive Fehlersuche danke ich Frau Loll ganz herzlich

Rainer Schlittgen

Vorwort zur 1. Auflage

Das Buch gibt eine an den Anwendungen orientierte Darstellung der statistischen Methoden auf mittlerem Niveau. „Mittleres Niveau“ meint dabei, dass das mathematische Rüstzeug, das zum Beispiel Wirtschaftswissenschaftler und Ingenieure in ihrem Grundstudium erwerben, für ein erfolgreiches Durcharbeiten ausreichen sollte.

Der Text ist im Rahmen von Vorlesungen entstanden, die ich unter dem Titel „Statistik nach der Grundausbildung“ bzw. „Statistik für Fortgeschrittene“ für Wirtschaftswissenschaftler und Wirtschaftsmathematiker an verschiedenen Universitäten immer wieder gehalten habe. In diesen Kursen konnte ich davon ausgehen, dass die Hörer schon über einige Grundkenntnisse verfügten. Um die Eigenständigkeit des Buches zu gewährleisten, wurden auch die Grundlagen mit aufgenommen. Die hier gewählte Anordnung des Stoffes ist dementsprechend weitgehend die übliche. Der wahrscheinlichkeitstheoretische Vorspann ist jedoch etwas knapper gehalten als sonst. (Da vergleichbare Texte im deutschen Sprachraum äußerst rar sind, ist der Bezug hier die englischsprachige Literatur.) Er reicht jedoch mit Sicherheit aus, um das Studium des Hauptteils ohne Rückgriff auf andere Bücher zu ermöglichen. Insbesondere sollte der Text auch als anwendungsorientierte Einführung in die Statistik für Mathematik- und Statistikstudenten geeignet sein.

In meiner „Einführung in die Statistik“ (EinfStat), auf die an mehreren Stellen hingewiesen wird, habe ich einen datenanalytischen Zugang zu den klassischen Methoden dargestellt. Dort werden die wichtigsten Verfahren mit gehöriger Motivation präsentiert. Hier sind die Ausführungen nun auf Konstruktionsprinzipien und Eigenschaften von Methoden ausgerichtet. Die einzelnen, konkreten Verfahren werden dabei eher als Beispiele von übergreifenden methodischen Ansätzen betrachtet. Die Anwendungsorientierung des Buches resultiert aus der Berücksichtigung der entsprechenden methodischen Fragestellungen und Zugänge. Dabei habe ich etliche neuere, sonst kaum in einem solchen Text behandelte methodische Ansätze wie Robustheit von Schätzfunktionen, die Bootstrap-Methode und multiple Tests aufgenommen; diese sind sowohl im Rahmen der theoretischen Statistik wie auch für die praktische Anwendung von großer Bedeutung.

Sigma-Algebren und Messbarkeit werden zwar angesprochen, da ich meine, dass sie für das Verständnis von Verteilungen grundlegend sind. Jedoch werden maßtheoretische Konzepte nicht weitergehend verwendet. Ich habe Beweise aufgenommen, soweit es das angestrebte Niveau zuließ. Denn Beweise sind hilfreich für ein vertieftes Verständnis. An verschiedenen Stellen habe ich Beweisskizzen gegeben, um die Grundzüge der Beweise wenigstens durchschimmern zu lassen.

Anstatt die verschiedenen Ansätze für statistische Rückschlüsse – Likelihood, Bayes- und Entscheidungstheorie – alle anzureißen, habe ich mich entschlossen, lieber den Likelihood-Ansatz tiefer zu verfolgen und in weitergehenden Einzelheiten darzustellen. Der Bayes-Ansatz ist mir wegen der benötigten, in der Praxis aber kaum (nie?) vorhandenen Vorinformation in Gestalt einer a priori Dichte immer suspekt geblieben. Bzgl. der Entscheidungstheorie ist es doch recht still geworden, das Interesse scheint sich weitgehend gelegt zu haben. Lineare Modelle werden hier nicht behandelt. Dazu gibt es ausgezeichnete Spezialliteratur. Auch ist diesen i. d. R. eine eigenständige Vorlesung gewidmet.

Diese Seiten hätten wohl kaum den Weg zwischen zwei Buchdeckel geschafft, wenn ich nicht das Glück gehabt hätte, in Thomas Noack einen Studenten zu finden, der mit unvergleichlichem Engagement die in einer sehr einfachen Textverarbeitung erstellte, vorläufige Version in LaTeX gebracht und die mehrmaligen als fundamental zu bezeichnenden Änderungen in motivierender Weise umgesetzt hätte. Ihm sage ich vor allem Dank.

Rainer Schlittgen

Inhaltsverzeichnis

1 Zufallsexperimente und -variablen

In der Statistik geht es darum, von Daten, den Ergebnissen wiederholter Beobachtungen eines Sachverhaltes, auf die zugehörige Grundgesamtheit oder den diese Daten generierenden Mechanismus zurückzuschließen. Die zugrundeliegenden Beobachtungsvorgänge werden im Folgenden als Experimente bezeichnet. Verständlicherweise wird der beabsichtigte induktive Schluss von den Beobachtungen auf den zugrundeliegenden 'Datengenerierenden Mechanismus' nicht ohne einige Voraussetzungen gelingen. Mit der formalen Beschreibung von geeigneten Experimenten und den damit zusammenhängenden Begriffsbildungen befassen wir uns in diesem Kapitel.

1.1 Wahrscheinlichkeitsverteilungen

1.1.1 Stichprobenraum und Wahrscheinlichkeit

Unsere eingangs getroffene Vereinbarung, Beobachtungsvorgänge als Experimente zu bezeichnen, ist sehr großzügig. Danach sind naturwissenschaftliche Experimente genauso enthalten wie einfache Alltagsbeobachtungen. Um die Bezeichnung 'Experiment' zu rechtfertigen, setzen wir bei solchen Wiederholungen jeweils ein exakt vorhersagbares Ergebnis voraus, so ist der induktive Schluss leicht: Wir haben es dann mit einer eindeutigen kausalen Beziehung zu tun. Unser Interesse gilt aber den Experimenten, bei denen das Ergebnis gerade nicht exakt vorhersagbar ist.

Definition 1.1.1 *Zufallsexperiment*

Ein Experiment heißt *Zufallsexperiment*, sofern es folgende Forderungen erfüllt:

- Es ist unter gleichen Bedingungen wiederholt durchführbar.
- Bei der Durchführung ist nicht exakt vorhersagbar, zu welchem Ergebnis der Beobachtungsprozess führt. Dies ist jedoch eindeutig.
- Es ist vorab angebbar, welche Ergebnisse überhaupt möglich sind.

Die Menge der möglichen Ergebnisse ω wird als *Stichprobenraum* Ω bezeichnet.

Bei einem Zufallsexperiment kommt der die Ergebnisse hervorbringende Mechanismus in der Form zum Tragen, dass geeignete Gesetzmäßigkeiten sich in der Masse der Beobachtungen niederschlägt. Für die Statistik sind dementsprechend die Ergebnisse wiederholter Durchführungen von Zufallsexperimenten relevant.

Beispiel 1.1.2 *Glücksspiele*

Die eingängigsten Beispiele für Zufallsexperimente sind Glücksspiele wie das Roulettespiel, Würfeln, Münzwürfe, das Ziehen einer Karte aus einem Stapel gut gemischter Spielkarten. Hierbei handelt es sich um Vorgänge, die in unserem Kulturkreis wohlbekannt sind und die sich wegen ihrer einfachen Struktur besonders zur Illustration von neuen Konzepten eignen.

Beim Würfeln lässt sich der Stichprobenraum mit den Zahlen Eins bis Sechs identifizieren: $\Omega = \{1,2,3,4,5,6\}$.

Beim Werfen einer deutschen 2 Euro-Münze kann die Zahlseite oder die Seite mit dem Adler nach oben zu liegen kommen. Wir können also setzen: $\Omega = \{Z, A\}$.

Ein drittes Zufallsexperiment bestehe wieder im Werfen einer deutschen 2 Euro-Münze, und zwar soll sie sooft geworfen werden, bis zum ersten Mal die Zahl-Seite nach oben zu liegen kommt. Die möglichen Ergebnisse sind dann

$$\omega_1 = Z,\ \omega_2 = AZ,\ \omega_3 = AAZ,\ \omega_4 = AAAZ,\ \omega_5 = AAAAZ, \ldots$$

und Ω ist die unendliche, aber abzählbare Menge $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$.

Beispiel 1.1.3 *Auslosung*

Ein Zufallsexperiment besteht in der Auswahl einer Person aus einer vorgegebenen, genau spezifizierten Personengruppe mittels Losverfahren. Jede Person stellt hier ein mögliches Ergebnis ω der Zufallsauswahl dar. Ω ist die endliche Gesamtheit aller Personen der betrachteten Gruppe.

Beispiel 1.1.4 *Blick auf eine Uhr*

Der Blick auf eine Uhr mit sich stetig bewegendem Sekundenzeiger kann als Zufallsexperiment angesehen werden. Das Ergebnis ist die Stellung des Sekundenzeigers. Diese kann in Bogenmaß angegeben werden. Dann besteht Ω aus allen Zahlen im Intervall $[0, 2\pi)$.

Es ist zwar bei einem Zufallsexperiment nicht möglich, exakt vorherzusagen, welches Ergebnis bei einer Durchführung beobachtet werden wird. Jedoch sind Chancen dafür von Interesse und auch oft ermittelbar, dass ein Ergebnis aus einer Teilmenge von Ω beobachtet werden wird. Solche Chancen sind unter zwei Gesichtspunkten von Interesse. Einmal ist die individuelle Chance wie die Gewinnchance bei einem einfachen Glückspiel; als zweites steht die Chance für die durchschnittliche Häufigkeit des Beobachtens einer Zusammenfassung von Ergebnissen, wenn das Zufallsexperiment sehr oft durchgeführt wird. Diese Betrachtungsweise herrscht etwa bei Versicherungen vor, wo weniger das einzelne Schicksal zählt als die Gesetzmäßigkeit, die sich bei einer großen Zahl von Versicherten ergibt.
Die Betrachtung von Häufigkeiten, mit denen zu einer Teilmenge gehörende Ergebnisse bei wiederholter Durchführung eines Zufallsexperimentes beobachtet werden, führt dazu, dass man auch Häufigkeiten anderer Teilmengen bestimmen will.

Definition 1.1.5 *absolute und relative Häufigkeit*

Sei Ω der zu einem Zufallsexperiment gehörige Stichprobenraum. Sei $A \subset \Omega$. Die *absolute Häufigkeit*, mit der A bei n Wiederholungen des Zufallsexperimentes eingetreten ist, ist die Anzahl der Wiederholungen, bei denen ein zu A gehörendes Ergebnis beobachtet wurde. Die *relative Häufigkeit* von A ist $h(A) = n(A)/n$.

Mit der relativen Häufigkeit von A lässt sich auch die des Komplementes A^c angeben:

$$h(A^c) = 1 - h(A).$$

Bei den Häufigkeiten zweier Teilmengen A und B gilt offensichtlich:

$$h(A \cup B) = h(A) + h(B) - h(A \cap B).$$

Weiter gilt:

$$h(\Omega) = 1.$$

Bedenken wir nun die oben angesprochene Verbindung von relativen Häufigkeiten und Chancen, so sollten für die Chancen verschiedener Teilmengen geeignete Berechnungen möglich sein. Da nicht für alle Teilmengen die Chancen von Interesse oder gar definierbar sind, werden die relevanten als *Ereignisse* besonders ausgezeichnet. Zudem wird verlangt, dass aus Ereignissen geeignet abgeleitete Teilmengen wieder Ereignisse sind.

Definition 1.1.6 *σ-Algebra von Ereignissen*

Ein System $\mathfrak{A}$ von Teilmengen des Stichprobenraumes Ω heißt *σ-Algebra von Ereignissen*, oder kurz *Ereignisalgebra*, falls gilt:

$$\Omega \in \mathfrak{A} \tag{1.1a}$$

$$A \in \mathfrak{A} \quad \Rightarrow \quad A^c \in \mathfrak{A} \tag{1.1b}$$

$$A_n \in \mathfrak{A} \quad \text{für } n = 1, 2, \ldots \quad \Rightarrow \quad \bigcup_{n=1}^{\infty} A_n \in \mathfrak{A}. \tag{1.1c}$$

Wir sagen, dass das Ereignis A bei einer Durchführung des Zufallsexperimentes eintritt, wenn ein Ereignis beobachtet wird, das zu A gehört.

Ω wird auch als *sicheres Ereignis* bezeichnet, da stets ein $\omega \in \Omega$ beobachtet wird, m. a. W. Ω stets eintritt. Das Ereignis $\Omega^c = \emptyset$ tritt nie ein: Jede Durchführung des Zufallsexperimentes führt zu einem Ergebnis. $\emptyset$ heißt daher auch das *unmögliche Ereignis*.
Aus den drei grundlegenden Eigenschaften einer σ-Algebra lassen sich weitere ableiten. Bevor wir dies tun, seien zur Erinnerung die folgenden Regeln für Mengenoperationen angegeben:

$$A \setminus B = A \cap B^c,$$

$$\bigcup A_n^c = \left(\bigcap A_n\right)^c,$$

$$\bigcap A_n^c = \left(\bigcup A_n\right)^c,$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

Satz 1.1.7 *Eigenschaften von σ-Algebren*

Ist $\mathfrak{A}$ eine σ-Algebra, so gilt:

$$\begin{array}{llll} (1)\ \emptyset \in \mathfrak{A} & & (2)\ A_1, A_2 \in \mathfrak{A} \Rightarrow & A_1 \cup A_2 \in \mathfrak{A} \\ (3)\ A_n \in \mathfrak{A} \text{ für } n = 1,2,\ldots \Rightarrow \bigcap_{n=1}^{\infty} A_n \in \mathfrak{A} & & (4)\ A_1, A_2 \in \mathfrak{A} \Rightarrow & A_1 \cap A_2 \in \mathfrak{A} \end{array} \tag{1.2}$$

Beweis:

(1) Wegen $\Omega^c = \emptyset$ gilt (1) aufgrund von (1.1b).

(2) Wir setzen $A_n = \emptyset$ für $n \geq 3$. Dann ist $A_n^c \in \mathfrak{A}$ für $n = 1,2,\ldots$ und es gilt

$$\bigcup_{n=1}^{\infty} A_n = A_1 \cap A_2 \in \mathfrak{A}.$$

(3) Mit $A_n \in \mathfrak{A}$ für $n = 1,2,\ldots$ ist wegen (1.1b) auch $A_n^c \in \mathfrak{A}$ für $n = 1,2,\ldots$
Wegen (1.1c) gilt somit

$$\bigcup_{n=1}^{\infty} A_n^c = \left(\bigcap_{n=1}^{\infty} A_n \right)^c \in \mathfrak{A}$$

und schließlich auch

$$\bigcap_{n=1}^{\infty} A_n \in \mathfrak{A}.$$

(4) Wir setzen $A_n = \Omega$ für $n \geq 3$. Dann ist $A_n \in \mathfrak{A}$ für $n = 1,2,\ldots$ und aufgrund von (1.1c) auch

$$A_1 \cap A_2 = \bigcap_{n=1}^{\infty} A_n \in \mathfrak{A}.$$

Beispiel 1.1.8 *σ-Algebren*

Sei $\Omega = \{1,2,3\}$ und $A = \{1,2\}$. Dann sind $\mathfrak{A}_1 = \{\emptyset, \{1,2\}, \{3\}, \Omega\}$ und
$\mathfrak{A}_2 = \mathfrak{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \Omega\}$ σ-Algebren über Ω.

Beispiel 1.1.9 *Weg zur Arbeit*

Sei Ω die Gesamtheit der Arbeitnehmer eines Betriebes. Alle Arbeitnehmer werden durch den von ihnen zurückzulegenden Weg zur Arbeit einer der folgenden Teilmengen von Ω zugeordnet:

$A_1 \widehat{=}$ Die Entfernung beträgt höchstens 1 km.
$A_2 \widehat{=}$ Die Entfernung beträgt mehr als 1 und höchstens 10 km.
$A_3 \widehat{=}$ Die Entfernung beträgt mehr als 10 und höchstens 25 km.
$A_4 \widehat{=}$ Die Entfernung beträgt mehr als 25 und höchstens 50 km.
$A_5 \widehat{=}$ Die Entfernung beträgt mehr als 50 km.

Die kleinste σ-Algebra, die diese Teilmengen enthält, ist die von $A_1,\ldots,A_5$ *erzeugte Ereignisalgebra*, die dann mit $A_1,\ldots,A_5$ gerade alle möglichen Vereinigungen sowie die leere Menge $\emptyset$ umfasst. Es gilt nämlich $\Omega = A_1 \cup A_2 \cup \ldots \cup A_5$ und für A_1^c:

$$A_1^c = A_2 \cup A_3 \cup A_4 \cup A_5\,.$$

Analog sieht man, dass mit den Vereinigungen auch die Komplemente für die anderen Ereignisse dazugehören.
Die Durchschnitte sind entweder leer ($=\emptyset$) oder führen ebenfalls auf eine Vereinigung einiger dieser fünf Teilmengen.

Beispiel 1.1.10 *Warten auf die S-Bahn*

Beim Warten auf die nächste S-Bahn ist die Wartezeit t (in Minuten) das mögliche Ergebnis eines entsprechenden ‚Experimentes'. Sind Verspätungen ausgeschlossen und verkehren die Züge in 20-Minuten-Abständen, so ist $\Omega = [0,20]$.Von Interesse sind hier Ereignisse der Form $[0,t]$. Das Eintreten des Ereignisses $[0,t]$ beinhaltet, dass man höchstens t Minuten auf die S-Bahn gewartet hat.
Für eine Ereignis-Algebra, die diese speziellen Ereignisse umfasst, gilt für geeignete t, t_i:

- Mit $[0,t]$ ist auch $(t,20]$ ein Ereignis.
- $(t_1,t_2]$ ist ein Ereignis: $(t_1,t_2] = [0,t_2] \cap (t_1,20]$.
- $\{t\}$ ist ein Ereignis: $\{t\} = \bigcap_{n=m}^{\infty}(t-1/n,t]$.
  (m ist so zu wählen, dass $(t-1/m,t] \subset (0,20]$ gilt. $\{0\}$ ist wegen $\{0\} = (0,t]^c$ ein Ereignis.)
- Alle Vereinigungen $\bigcup_{s=1}^{n}(t_{2s},t_{2s+1}]$ sind Ereignisse und entsprechend die Komplemente.

Wie die Beispiele 1.1.8 und 1.1.9 deutlich machen, braucht eine Ereignisalgebra nicht stets aus allen Teilmengen des Stichprobenraumes zu bestehen. Das Beispiel 1.1.10 wirft die Frage auf, ob mit den speziellen Intervallen nicht schließlich alle Teilmengen von $\Omega = [0,20]$ als Ereignisse zugelassen werden müssen. Dieses ist nicht der Fall. Es gibt vielmehr eine kleinere Ereignis-Algebra, die alle Intervalle enthält. Wie man zeigen kann, gibt es stets eine kleinste σ-Algebra, die eine gegebene Menge von Ereignissen umfasst. Diesen Punkt verdeutlicht der folgende Satz.

Satz 1.1.11 *Durchschnitt von σ-Algebren*

Seien $\mathfrak{A}_1,\mathfrak{A}_2$ σ-Algebren über dem Stichprobenraum Ω. Dann ist auch $\mathfrak{A} = \mathfrak{A}_1 \cap \mathfrak{A}_2$ eine σ-Algebra über Ω.

Beweis: Zur Übung dem Leser empfohlen.

Die Aussage des Satzes lässt sich auf beliebige Systeme von σ-Algebren erweitern. Da nun die Potenzmenge von Ω eine σ-Algebra ist, die jedes gegebene System von Teilmengen enthält, ist der Durchschnitt über alle σ-Algebren, die diese gegebenen Teilmengen enthalten, wohldefiniert und gibt die gesuchte kleinste σ-Algebra.

Allgemein sind für Zufallsexperimente mit Stichprobenraum $\Omega \subset \mathbb{R}$ oder $\Omega \subset \mathbb{R}^k$ Ereignisse der Form k-dimensionaler Intervalle

$$(a,b] \quad \text{bzw.} \quad (a_1,b_1] \times (a_2,b_2] \times \ldots \times (a_k,b_k]$$

von besonderem Interesse:

Definition 1.1.12 *σ-Algebra der Borelschen Mengen*

Die kleinste σ-Algebra von Teilmengen von $\mathbb{R}^k$, die alle k-dimensionalen Intervalle enthält, heißt *σ-Algebra der Borelschen Mengen* von $\mathbb{R}^k$. Sie wird mit $\mathfrak{B}^k$ bezeichnet.

Falls Ω nur aus endlich vielen oder höchstens abzählbar vielen möglichen Ergebnissen besteht, wird als Ereignisalgebra fast immer die Potenzmenge $\mathfrak{P}(\Omega)$ genommen. Es werden also alle Teilmengen als Ereignisse zugelassen. In diesen Fällen sind nämlich die einzelnen Ergebnisse ω_i von Interesse. Mit den Ereignissen $\{\omega_i\}$ sind dann nach der Definition der Ereignisalgebra alle Teilmengen als Ereignisse zugelassen.

Wir kommen nun zu der zentralen Definition.

Definition 1.1.13 *Axiomensystem von Kolmogorov*

Eine auf einer σ-Algebra von Ereignissen $\mathfrak{A}$ über Ω definierte Mengenfunktion $\mathrm{P}: \mathfrak{A} \to \mathbb{R}$ heißt eine *Wahrscheinlichkeitsverteilung*, wenn sie die *Kolmogorovschen Axiome* erfüllt:

$$\mathrm{P}(A) \geq 0 \tag{1.3a}$$

$$\mathrm{P}(\Omega) = 1 \tag{1.3b}$$

$$\mathrm{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathrm{P}(A_n), \text{ wenn } A_n \in \mathfrak{A} \text{ für } n = 1,2,\ldots \text{ mit } A_n \cap A_m = \emptyset \text{ für } n \neq m. \tag{1.3c}$$

$(\Omega, \mathfrak{A}, \mathrm{P})$ heißt dann ein *Wahrscheinlichkeitsraum*.

Die Eigenschaften einer Wahrscheinlichkeitsverteilung sind gerade die wesentlichen Eigenschaften relativer Häufigkeiten. Es war der geniale Ansatz von Kolmogorov, nicht von irgendwelchen Grenzwerten relativer Häufigkeiten auszugehen, sondern ihre strukturellen Eigenschaften als Basis für die axiomatische Formalisierung der Wahrscheinlichkeiten zu nehmen. Für die Anwendung bleibt dann die Aufgabe, den formalen Rahmen jeweils soweit zu füllen, dass die konkrete Wahrscheinlichkeitsverteilung tatsächlich die Chancen für das Eintreffen der interessierenden Ereignisse ergibt.

Nach dem ‚Prinzip der großen Zahlen' kann man davon ausgehen, dass bei einem Zufallsexperiment die relative Häufigkeit in etwa mit seiner (objektiv existierenden) Wahrscheinlichkeit, die die Chance seines Eintreffens angibt, übereinstimmt. Dieses Prinzip wird später in seine mathematische Form gebracht.

Der folgende Satz gibt einfache Folgerungen aus den Kolmogorovschen Axiomen für eine Wahrscheinlichkeitsverteilung an.

Satz 1.1.14 *Eigenschaften von Wahrscheinlichkeitsverteilungen*

Sei $(\Omega, \mathfrak{A}, \mathrm{P})$ ein Wahrscheinlichkeitsraum. Dann gilt:

(1) $\mathrm{P}(\emptyset) = 0$

(2) Sind $A_1, \ldots, A_n$ paarweise disjunkte Ereignisse aus $\mathfrak{A}$, d. h. $A_i \cap A_j = \emptyset$ für $i \neq j$, so gilt

$$\mathrm{P}(A_1 \cup \ldots \cup A_n) = \mathrm{P}(A_1) + \ldots + \mathrm{P}(A_n).$$

(3) $A \in \mathfrak{A} \quad \Rightarrow \quad \mathrm{P}(A) = 1 - \mathrm{P}(A)$

(4) $A, B \in \mathfrak{A}$ mit $A \subset B \quad \Rightarrow \quad \mathrm{P}(A) \leq \mathrm{P}(B)$

(5) $A \in \mathfrak{A} \quad \Rightarrow \quad \mathrm{P}(A) \leq 1$

(6) *(Formel von Sylvester):*
Für jede endliche Folge $A_1, \ldots, A_n$ von Ereignissen aus $\mathfrak{A}$ gilt:

$$\mathrm{P}(A_1 \cup \ldots \cup A_n) = \sum_{i=1}^{n} \mathrm{P}(A_i) - \sum_{1 \leq i < j \leq n}^{n} \mathrm{P}(A_i \cap A_j) + \ldots - \ldots + (-1)^{n+1} \mathrm{P}(A_1 \cap \ldots \cap A_n). \quad (1.4)$$

(7) *(Bonferroni-Ungleichung):*
Sind $A_1, \ldots, A_n$ Ereignisse aus $\mathfrak{A}$, so gilt:

$$\mathrm{P}(A_1 \cap \ldots \cap A_n) \geq 1 - [\mathrm{P}(A_1^c) + \ldots + \mathrm{P}(A_n^c)], \quad (1.5)$$

oder äquivalent dazu:

$$\mathrm{P}(A_1 \cup \ldots \cup A_n) \leq \mathrm{P}(A_1) + \ldots + \mathrm{P}(A_n).$$

(8) Sei $A_1, \ldots, A_n, \ldots$ eine Folge von Ereignissen aus $\mathfrak{A}$, so gilt:

$$\mathrm{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathrm{P}(A_n).$$

Beweis:

(1) Es gilt $\Omega = \Omega \cup \emptyset \cup \emptyset \cup \ldots$ Wegen $\Omega \cap \emptyset = \emptyset$ gilt mit Axiom 3:
$\mathrm{P}(\Omega) = \mathrm{P}(\Omega) + \mathrm{P}(\emptyset) + \mathrm{P}(\emptyset) + \ldots$, also $1 = 1 + \mathrm{P}(\emptyset) + \mathrm{P}(\emptyset) + \ldots$
Hieraus folgt $\mathrm{P}(\emptyset) = 0$.

(2) Sei $\mathfrak{A}_i = \emptyset$ für $i \geq n+1$.
Dann gilt wegen $A_i \cap \emptyset = \emptyset$ für $i \leq n$ und aufgrund von Axiom 3 mit (1):

$$\begin{aligned} \mathrm{P}(A_1 \cup \ldots \cup A_n) &= \mathrm{P}(A_1 \cup \ldots \cup A_n \cup \emptyset \cup \emptyset \cup \ldots) = \mathrm{P}(A_1) + \ldots + \mathrm{P}(A_n) + \mathrm{P}(\emptyset) + \mathrm{P}(\emptyset) + \ldots \\ &= \mathrm{P}(A_1) + \ldots + \mathrm{P}(A_n) \end{aligned}$$

(3) Es gilt $A \cup A^c = \Omega$ mit $A \cap A^c = \emptyset$.
Aufgrund von (2) gilt also $P(A) + P(A^c) = 1$ und somit $P(A^c) = 1 - P(A)$.

(4) Aus $A \subset B$ folgt $A \cap B = A$. Weiterhin gilt $B = (A \cap B) \cup (A^c \cap B)$ mit $A \cap B \cap A^c \cap B = \emptyset$.
Somit gilt $P(B) = P(A) + P(A^c \cap B)$. Wegen $P(A^c \cap B) \geq 0$ folgt $P(A) \leq P(B)$.

(5) Wegen $A \subset \Omega$ folgt mit $P(\Omega) = 1$ aus (4) dann $P(A) \leq 1$.

(6) Wir zeigen dies durch vollständige Induktion. Sei $n = 2$. Dann gilt $A_1 \cup A_2 = A_1 \cup (A_1^c \cap A_2)$. Somit gilt dann wegen $A_1^c \cap (A_1^c \cap A_2) = \emptyset$:

$$P(A_1 \cup A_2) = P(A_1) + P(A_1^c \cap A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

Die Behauptung gelte nun für beliebiges, festes n:

$$P(A_1 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(A_i \cap A_j) + \ldots - \ldots + (-1)^{n+1} \cdot P(A_1 \cap \ldots \cap A_n).$$

Zu zeigen ist dann, dass die Formel auch für $n+1$ richtig ist. Es gilt

$$\begin{aligned} P(A_1 \cup \ldots \cup A_{n+1}) &= P(A_1 \cup \ldots \cup A_n) + P(A_{n+1}) - P((A_1 \cup \ldots \cup A_n) \cap A_{n+1}) \\ &= \sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(A_i \cap A_j) + \ldots - \ldots + (-1)^{n+1} \cdot P(A_1 \cap \ldots \cap A_n) \\ &\quad + P(A_{n+1}) - P((A_1 \cap A_{n+1}) \cup \ldots \cup (A_n \cap A_{n+1})). \end{aligned}$$

Nun lässt sich die Induktionsvoraussetzung auf den letzten Term der rechten Seite anwenden:

$$\begin{aligned} P((A_1 \cap A_{n+1}) \cup \ldots \cup (A_n \cap A_{n+1})) &= \sum_{i=1}^{n} P(A_i \cap A_{n+1}) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(A_i \cap A_j \cap A_{n+1}) \\ &\quad + \ldots - \ldots + (-1)^{n+1} \cdot P(A_1 \cap \ldots \cap A_n \cap A_{n+1}). \end{aligned}$$

Damit erhalten wir:

$$\begin{aligned} P(A_1 \cup \ldots \cup A_{n+1}) &= \sum_{i=1}^{n+1} P(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(A_i \cap A_j) + \ldots - \ldots + (-1)^{n+1} \cdot P(A_1 \cap \ldots \cap A_n) \\ &\quad - \sum_{i=1}^{n} P(A_i \cap A_{n+1}) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(A_i \cap A_j \cap A_{n+1}) + \ldots - \ldots \\ &\quad + (-1)^{n+1} \cdot P(A_1 \cap \ldots \cap A_n \cap A_{n+1}) \\ &= \sum_{i=1}^{n+1} P(A_i) - \sum_{i=1}^{n} \sum_{j=i+1}^{n+1} P(A_i \cap A_j) + \ldots - \ldots + (-1)^{n+2} \cdot P(A_1 \cap \ldots \cap A_{n+1}). \end{aligned}$$

(7) Die zweite Form der Bonferroni-Ungleichung erhalten wir, indem in (8) $A_{n+1} = A_{n+2} = \ldots = \emptyset$ gesetzt werden. Die erste Form ergibt sich daraus durch den Übergang zu den Komplementen.

(8) Seien $B_1 = A_1,\ B_2 = A_2 \cap A_1^c,\ B_3 = A_3 \cap A_1^c \cap A_2^c,\ \ldots$
Dann gilt $B_i \subseteq A_i$ für $i = 1,2,3,\ldots$ und wegen (4) $\mathrm{P}(B_i) \leq \mathrm{P}(A_i)$ für $i = 1,2,3,\ldots$ Somit ist $\sum_{n=1}^{\infty} \mathrm{P}(B_n) \leq \sum_{n=1}^{\infty} \mathrm{P}(A_n)$.
Weiterhin gilt $B_i \cap B_j = \emptyset$ für $i \neq j$. Also folgt

$$\mathrm{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathrm{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathrm{P}(B_n) \leq \sum_{n=1}^{\infty} \mathrm{P}(A_n).$$

Auf die in den beiden folgenden Beispielen angegebenen Wahrscheinlichkeitsverteilungen werden wir im Folgenden immer wieder zurückkommen.

Beispiel 1.1.15 *minimalistische Wahrscheinlichkeitsverteilung*

Sei Ω ein Stichprobenraum und A ein uns interessierendes Ereignis. Die von A erzeugte σ-Algebra ist $\mathfrak{A} = \{\emptyset, \Omega, A, A^c\}$. Durch Festlegung von $\mathrm{P}(A) = p$ mit $0 \leq p \leq 1$ erhalten wir eine Wahrscheinlichkeitsverteilung auf $\{\Omega, \mathfrak{A}\}$.

Beispiel 1.1.16 *Wahrscheinlichkeitsverteilung über einer Potenzmenge*

Sei Ω eine abzählbare Ergebnismenge, d. h. $\Omega = \{\omega_1, \omega_2, \ldots\}$ und $\mathfrak{A} = \mathfrak{P}(\Omega)$. Durch Festlegung von $\mathrm{P}(\{\omega_i\}) = p_i$ mit $\sum_{i=1}^{\infty} p_i = 1$ erhalten wir eine Wahrscheinlichkeitsverteilung auf $\{\Omega, \mathfrak{P}(\Omega)\}$ mit $\mathrm{P}(A) = \sum_{\omega_i \in A} \mathrm{P}(\{\omega_i\})$ für jedes $A \in \mathfrak{P}(\Omega)$.

Satz 1.1.17 ***σ*-*Stetigkeit von Wahrscheinlichkeitsverteilungen***

$(\Omega, \mathfrak{A}, \mathrm{P})$ sei ein Wahrscheinlichkeitsraum, $A_1, \ldots, A_n, \ldots$ sei eine Folge von Ereignissen aus $\mathfrak{A}$.

Mit $A_1 \subset A_2 \subset \ldots \subset A_n \subset \ldots$ gilt für $A = \bigcup_{n=1}^{\infty} A_n$:

$$\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(A). \tag{1.6}$$

Mit $A_1 \supset A_2 \supset \ldots \supset A_n \supset \ldots$ gilt für $A = \bigcap_{n=1}^{\infty} A_n$:

$$\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(A). \tag{1.7}$$

Beweis: Für die erste Beziehung setzen wir $B_1 = A_1,\ B_2 = A_2 \backslash A_1, \ldots, B_n = A_n \backslash A_{n-1}, \ldots$ Dann gilt $B_i \cap B_j = \emptyset$ für $i \neq j$ und

$$\bigcup_{i=1}^{n} B_i = A_1 \cup (A_2 \backslash A_1) \cup \ldots \cup (A_n \backslash A_{n-1}) = A_n, \qquad \bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_n = A.$$

Damit folgt:

$$\mathrm{P}(A)=\mathrm{P}\left(\bigcup_{i=1}^{\infty}B_i\right)=\sum_{i=1}^{\infty}\mathrm{P}(B_i)=\lim_{n\to\infty}\sum_{i=1}^{n}\mathrm{P}(B_i)=\lim_{n\to\infty}\mathrm{P}\left(\bigcup_{i=1}^{n}B_i\right)=\lim_{n\to\infty}\mathrm{P}(A_n).$$

Die zweite Aussage erhalten wir durch den Übergang zu den Komplementen und der Anwendung der ersten.

Beispiel 1.1.18 *Regenmenge*

Sei $\Omega=[0,\infty]$ und $\mathfrak{B}_\Omega$ die kleinste σ-Algebra, die alle Intervalle $(a,b]$, $0\le a<b$ enthält.
$\omega\in[0,\infty]$ stehe für die erwartete Regenmenge in einer gegebenen Zeiteinheit.
f sei eine integrierbare Funktion mit

(i) $f(x)\ge 0$ für alle $x\ge 0$

(ii) $\int_0^\infty f(x)\,\mathrm{d}x=1-p<1.$

Jedem Intervall $(a,b]\subset\Omega$ wird die Wahrscheinlichkeit $\mathrm{P}((a,b])=\int_a^b f(x)\,\mathrm{d}x$ zugeordnet und den Intervallen $[0,c]$ die Wahrscheinlichkeit $\mathrm{P}([0,c])=p+\int_0^c f(x)\,\mathrm{d}x$.

Die Ereignisse $[a,b]$ bedeuten hier, dass die Regenmenge zwischen a und b liegt. Die Wahrscheinlichkeit für das Ereignis $\{0\}$, dass kein Regen fällt, ist positiv, wie man mit dem Satz 1.1.17 sieht:

$$\mathrm{P}(\{0\})=\mathrm{P}\left(\bigcap_{n=1}^{\infty}\left[0,\frac{1}{n}\right]\right)=\lim_{n\to\infty}\mathrm{P}\left(\left[0,\frac{1}{n}\right]\right)=\lim_{n\to\infty}\left(p+\int_0^{1/n}f(x)\,\mathrm{d}x\right)=p.$$

Die Wahrscheinlichkeit für derartige Ereignisse ist sonst Null:

$$\mathrm{P}(\{\omega\})=\mathrm{P}\left(\bigcap_{n=1}^{\infty}\left(\omega-\frac{1}{n},\omega\right]\right)=\lim_{n\to\infty}\mathrm{P}\left(\left(\omega-\frac{1}{n},\omega\right]\right)=\lim_{n\to\infty}\left(\int_{\omega-1/n}^{\omega}f(x)\,\mathrm{d}x\right)=0.$$

Über die Festlegung der Wahrscheinlichkeiten für die einzelnen Intervalle können die Wahrscheinlichkeiten für alle $A\in\mathfrak{A}$ eindeutig bestimmt werden. Würde man eine größere σ-Algebra von Ereignissen über Ω zulassen (etwa $\mathfrak{P}(\Omega)$), so wäre die eindeutige Bestimmung nicht mehr gewährleistet.

1.1.2 Gleichmöglichkeitsmodelle und Kombinatorik

Ein wichtiger Spezialfall der im Beispiel 1.1.16 betrachteten Situation liegt vor, wenn Ω endlich ist und alle Elementarereignisse $\{\omega_i\}$, $(i=1,\dots,N)$ gleichwahrscheinlich sind, d. h.

$P(\{\omega_i\}) = 1/N$. Man erhält in diesem Fall für jedes $A \in \mathfrak{P}(\Omega)$, wenn $|A|$ die Anzahl der Ergebnisse in A ist:

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\}) = \frac{|A|}{N}. \tag{1.8}$$

Man spricht dann von einem *Gleichmöglichkeitsmodell* oder einer *Laplaceschen Wahrscheinlichkeitsverteilung* über $(\Omega, \mathfrak{P}(\Omega))$. Bei Gleichmöglichkeitsmodellen besteht die oftmals gar nicht so leichte Aufgabe darin, die Anzahl der Ergebnisse des interessierenden Ereignisses und die Anzahl aller zu Ω gehörenden Ergebnisse zu bestimmen. Dazu übersetzt man das Gleichmöglichkeitsmodell gerne in ein sogenanntes *Urnenmodell*. Dabei stellt man sich seit Huygens, der im 17ten Jahrhundert das erste Lehrbuch zur Wahrscheinlichkeitsrechnung geschrieben hat, eine mit gleichartigen Kugeln gefüllte Urne vor. Die Kugeln mögen durchnummeriert sein und ggfls. (eine) weitere Markierung(en) tragen. Aus der Urne werden dann Kugeln gezogen und die Ergebnisse werden festgehalten. Dies kann jeweils auf unterschiedliche Weise geschehen.
Folgende Situationen sind von Bedeutung:

- Ziehen von n Kugeln mit bzw. ohne Zurücklegen in die Urne.
- Berücksichtigen der Reihenfolge, in der die Kugeln gezogen wurden, bzw. Vernachlässigen der Reihenfolge.

Insgesamt ergeben sich also vier Kombinationen. Bevor wir auf die entsprechenden Anzahlen von möglichen Ergebnissen eingehen, seien die relevanten Zählkoeffizienten angegeben.

Definition 1.1.19 *Fakultät und Binomialkoeffizient*

Sei $n \in \mathbb{N}$. Dann ist

$$n! = 1 \cdot \ldots \cdot n,$$

sprich n *Fakultät*. Es wird $0! = 1$ vereinbart. Der *Binomialkoeffizient* n über k, $n \geq k$, ist

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

Satz 1.1.20 *Eigenschaften des Binomialkoeffizienten*

Für Binomialkoeffizienten gilt

(i) $\binom{n}{k} = \binom{n}{n-k}$

(ii) $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$

(iii) $\sum_{k=0}^{n} \binom{n}{k} = 2^n.$

Beweis: (iii) folgt mit der binomischen Formel $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$, wenn $a = b = 1$ gesetzt werden.
(i) und (ii) ergeben sich durch Hinschreiben und Zusammenfassen. (Formel (ii) bildet die Grundlage des Pascal'schen Dreiecks.)

Satz 1.1.21 *Stirling'sche Formel*

Es gilt

$$n! \sim \sqrt{2\pi} \cdot n^{n+1/2} \cdot \exp\left[-n + \frac{1}{12n}\right], \tag{1.9}$$

wobei das Zeichen $\sim$ ausdrücken soll, dass der Quotient beider Seiten für $n \to \infty$ gegen 1 geht.

Beweis: Siehe Feller (1968).

Satz 1.1.22 *Anzahl von Ergebnissen bei Urnenexperimenten*

Sei eine Urne mit N gleichartigen, durchnummerierten Kugeln gegeben, formal $\Omega = \{1, \ldots, N\}$. Die Anzahl der unterschiedlichen Ergebnisse beim Ziehen von n Kugeln sind dann

- N^n beim Ziehen mit Zurücklegen, d. h. von $\Omega_1 = \{(\omega_1, \ldots, \omega_n) | \, \omega_i \in \Omega\}$;
- $N \cdot (N-1) \cdot \ldots \cdot (N-n+1)$ beim Ziehen ohne Zurücklegen, d. h. von $\Omega_2 = \{(\omega_1, \ldots, \omega_n) | \, \omega_i \in \Omega, \, \omega_i \neq \omega_j \text{ für } i \neq j\}$.

Beweis: Wir betrachten zunächst das Ziehen mit Zurücklegen.
Für die 1. Kugel gibt es N Möglichkeiten. Zu jeder dieser N Möglichkeiten gibt es bei der zweiten Kugel wiederum N Möglichkeiten. Bei zwei Kugeln gibt es also $N + \ldots + N = N \cdot N = N^2$ Möglichkeiten.
Zu jeder dieser N^2 Möglichkeiten gibt es bei der dritten Kugel abermals N Möglichkeiten. Bei drei Kugeln gibt es also $N^2 + N^2 + \ldots + N^2 = N \cdot N^2 = N^3$ Möglichkeiten, u. s. w.
Allgemein sind es wie behauptet N^n mögliche Ergebnisse.
Beim Ziehen ohne Zurücklegen gibt es für die erste Kugel wie in der eben betrachteten Situation N Möglichkeiten. Für die zweite bleiben noch $N-1$ übrig. Bei zwei Kugeln gibt es daher zu jeder ersten möglichen Kugel $N-1$ für die zweite. Das sind $(N-1)+\ldots+(N-1) = N \cdot (N-1)$ Möglichkeiten.
Bei n ausgewählten Kugeln sind das schließlich $N \cdot (N-1) \cdot \ldots \cdot (N-n+1)$ mögliche Ziehungen.

Die Ergebnisse wiederholter Ziehungen haben wir als Tupel $(\omega_1, \ldots, \omega_n)$ angegeben. Bei dieser Darstellung ist implizit vereinbart, dass die Reihenfolge wesentlich ist. Beispielsweise unterscheidet sich das Tupel $(1,2,3)$ von $(3,2,1)$ und $(1,3,2)$. Man bezeichnet daher auch die beiden im Satz angesprochenen Ziehungsarten als *Ziehungen mit Berücksichtigung der Anordnung*.

Beispiel 1.1.23 *Informationen in Computern*

In Computern werden Informationen durch endliche Folgen, die aus Nullen oder Einsen bestehen, dargestellt. Die Anzahl der unterschiedlichen Zeichen, wenn einem 8 Stellen zur Verfügung stehen, entspricht der Anzahl der Ziehungen mit Zurücklegen von 8 Kugeln aus einer Urne, die zwei Kugeln enthält. Es sind also $N = 2,\ n = 8$. Demnach können mit einer Folge, die aus 8 Nullen oder Einsen besteht, $2^8 = 256$ mögliche Zeichen dargestellt werden.

Beispiel 1.1.24 *Personen und Stühle*

Vier Stühle stehen nebeneinander. Auf wie viele Arten können sich vier Personen auf die vier Stühle setzen, wenn auf jedem der Stühle genau eine Person sitzen soll? Hier handelt es sich um das Ziehen ohne Zurücklegen. Es ist $N = n = 4$. Daher gilt: $4 \cdot \ldots \cdot (4-4+1) = 4! = 24$. Vier Personen können sich also auf 24 Arten auf vier Stühle setzen, wobei auf jedem Stuhl genau eine Person sitzt.

Im letzten Beispiel wird deutlich, dass wir beim Ziehen ohne Zurücklegen im Fall $n = N$ gerade die Anzahl $N!$ aller Anordnungen oder *Permutationen* von N unterscheidbaren Elementen erhalten. Häufig spielt die Reihenfolge, in der die einzelnen Kugeln gezogen werden, keine Rolle. Dann interessiert man sich für die Anzahl der unterschiedlichen Ergebnisse, wobei alle als identisch angesehen werden, die sich nur durch die Anordnung unterscheiden.

Satz 1.1.25 *Anzahl von Teilmengen von gegebenem Umfang*

Sei eine Urne mit N gleichartigen, durchnummerierten Kugeln gegeben, formal $\Omega = \{1, \ldots, N\}$. Es gibt $\binom{N}{n}$ Teilmengen vom Umfang n, d. h. mögliche Ergebnisse beim Ziehen ohne Zurücklegen von n Kugeln, wenn die Reihenfolge nicht berücksichtigt wird.

Beweis: Beim Ziehen ohne Zurücklegen erhielten wir als Anzahl der möglichen Ergebnisse $N \cdot (N-1) \cdot \ldots \cdot (N-n+1)$, wenn die Anordnung von Bedeutung ist. Wie wir im Beispiel eben gesehen haben, gibt es $n!$ unterschiedliche Ergebnisse, die sich nur durch die Anordnung unterscheiden. Es sind also für die Anzahl der n-Teilmengen einfach die Gesamtzahl der Tupel dadurch zu dividieren:

$$\frac{N \cdot (N-1) \cdot \ldots \cdot (N-n+1)}{n!} = \frac{N!}{n!(N-n)!} = \binom{N}{n}.$$

Beispiel 1.1.26 *Turing-Test*

Um das Jahr 2000, so hatte der geniale Mathematiker Turing prophezeit, würden Computer in der Lage sein, Menschen in einem als 'Turing-Test' bezeichneten Fragespiel hinters Licht zu führen. Turing hatte vorgeschlagen, Computer hinter einen Vorhang zu stellen; ein Mitspieler, der nicht wissen dürfe, ob sich ein Mensch oder eine Maschine hinter dem Vorhang verberge, könne dann auf dem ‚Tippwege' Fragen an den Unsichtbaren richten. Wenn der Fragesteller nicht zu entscheiden vermöge, ob hinter dem Vorhang ein Mensch oder eine Maschine verborgen sei, so komme dem Computer das Prädikat

einer 'denkenden Maschine' zu. In einer '100-Meter-Version des Turingschen Marathonlaufes' (Spiegel 47/1991, S.332) wurden zwei Menschen und sechs Computerprogramme hinter Vorhängen versteckt.

Acht Bürger durften Fragen aus eng begrenzten Wissensgebieten an die Verborgenen richten, sich aber nicht untereinander verständigen. Ihnen war verraten worden, dass an mindestens zwei Plätzen Menschen säßen. Wenn ein Bürger davon ausging, dass genau zwei Menschen und sechs Maschinen dabei waren, wie groß ist die Chance dass er durch einfaches Raten die richtige Zuordnung trifft?

Es sind alle möglichen Anordnungen zu bestimmen. Dieses Problem ist gleichwertig damit, dass aus einer Urne mit acht Kugeln zwei ohne Zurücklegen gezogen werden. Die Reihenfolge spielt keine Rolle. Daher sind es $\binom{8}{2} = 49$ mögliche Anordnungen. Die gesuchte Wahrscheinlichkeit ist also $1/49$.

1.1.3 Bedingte Wahrscheinlichkeit und Unabhängigkeit

Das Konzept der bedingten Wahrscheinlichkeit hat seinen Ursprung in einer einfachen experimentiellen Situation. Wenn bei einem Experiment bekannt ist, dass das Ereignis gewisse Randbedingungen erfüllt, welche Wahrscheinlichkeitsverteilung gibt dann die Chancen für die entsprechenden Ereignisse an? Zur Lösung werden zunächst die Häufigkeiten zweier Ereignisse A und B betrachtet. Es wird die Teilfolge ausgewählt, bei der jeweils B beobachtet wurde. Das sei die Randbedingung. Nun richtet sich die Frage auf die relative Häufigkeit, mit der A in dieser Teilfolge auftrat. Formal ergibt sie sich zu $n(A \cap B)/n(B)$.

Wegen $n(A \cap B)/n(B) = h(A \cap B)/h(B)$ wird folgende Definition nahegelegt:

Definition 1.1.27 *bedingte Wahrscheinlichkeit*

$(\Omega, \mathfrak{A}, \mathrm{P})$ sei ein Wahrscheinlichkeitsraum, $B \in \mathfrak{A}$ mit $\mathrm{P}(B) > 0$. Dann heißt

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \tag{1.10}$$

die *bedingte Wahrscheinlichkeit* von A gegeben B.

Wie man leicht sieht, ist $\mathrm{P}(\cdot|B)$ eine Wahrscheinlichkeitsverteilung über $(B, \mathfrak{A}_B)$, falls $\mathrm{P}(B) > 0$. $\mathfrak{A}_B$ ist dabei die auf B eingeschränkte σ-Algebra:

$$\mathfrak{A}_B = \{A \cap B | A \in \mathfrak{A}\}.$$

Beispiel 1.1.28 *TV-Show*

In einer TV-Show gibt es drei Vorhänge, einen Show-Master und einen Spieler. Hinter einem der Vorhänge ist ein wertvoller Preis verborgen, hinter den anderen beiden befindet sich nichts. Nur der Show-Master weiß, hinter welchem Vorhang sich der Preis befindet. Der Spieler gewinnt den Preis, wenn er auf den richtigen Vorhang weist. Zu Beginn wählt der Spieler einen der Vorhänge. Dann öffnet der Show-Master einen der beiden übrigen,

hinter dem sich der Preis nicht befindet und erlaubt dem Spieler neu zu wählen. Sollte der Spieler wechseln und auf den dritten Vorhang weisen? Diese Fragestellung führte seinerzeit zu einer breiten öffentlichen Debatte. Hintergrund der Debatte bildete das Problem der Unterscheidung zwischen bedingter und unbedingter Wahrscheinlichkeit.

Offensichtlich ist, dass die bedingte Wahrscheinlichkeit für jeden der beiden Vorhänge 0.5 beträgt, sobald der Show-Master erst einmal einen (ohne den Preis dahinter) geöffnet hat. Von diesem Punkt des Geschehens aus betrachet bringt das Wechseln keine Verbesserung der Gewinnchacen. Das Wechseln ist jedoch von Vorteil, wenn das Spiel unbedingt vor Beginn betrachtet wird. Dann gibt es die in der Tabelle 1.1 dargestellten Möglichkeiten. Man sieht, dass die Strategie „Wechseln“ in 6 von 9 Fällen zum Gewinn führt und der Spieler seine Ausgangs-Gewinn-Chance von 1/3 verdoppelt.

Tabelle 1.1: *Möglichkeiten bei einer Quiz-Show*

Preis ist hinter Vorhang Nr.	Spieler wählt Vorhang Nr.	Show-Master öffnet Vorhang Nr.	Spieler wechselt	Gewinn
1	1	2 oder 3	ja / nein	- / +
1	2	3	ja / nein	+ / -
1	3	2	ja / nein	+ / -
2	1	3	ja / nein	+ / -
2	2	1 oder 3	ja / nein	- / +
2	3	1	ja / nein	+ / -
3	1	2	ja / nein	+ / -
3	2	1	ja / nein	+ / -
3	3	1 oder 2	ja / nein	- / +

Die sich unmittelbar aus der Definition ergebende Beziehung

$$\mathrm{P}(A) = \mathrm{P}(A|B)\mathrm{P}(B)$$

wird bisweilen als *Multiplikationssatz* bezeichnet. Er lässt sich leicht auf $n > 2$ Ereignisse übertragen.

Satz 1.1.29 *Multiplikationssatz*

Seien $(\Omega, \mathfrak{A}, \mathrm{P})$ ein Wahrscheinlichkeitsraum und $A_1, A_2, \ldots, A_n \in \mathfrak{A}$ Ereignisse mit $\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) > 0$. Dann gilt

$$\begin{aligned} &\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_n) \\ &= \mathrm{P}(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \cdot \mathrm{P}(A_{n-1}|A_1 \cap A_2 \cap \ldots \cap A_{n-2}) \cdot \ldots \cdot \mathrm{P}(A_2|A_1) \cdot \mathrm{P}(A_1). \end{aligned} \tag{1.11}$$

Beweis: Aus $\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) > 0$ folgt mit $A_1 \cap A_2 \cap \ldots \cap A_{n-1} \subset A_1 \cap A_2 \cap \ldots \cap A_i$, für $1 \leq i \leq n-2$:

$$\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_i) > 0.$$

Also gilt

$$\begin{aligned}
&\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_n) \\
&\quad = \frac{\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_n)}{\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1})} \cdot \frac{\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1})}{\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-2})} \cdot \ldots \cdot \frac{\mathrm{P}(A_1 \cap A_2)}{\mathrm{P}(A_1)} \cdot \mathrm{P}(A_1) \\
&\quad = \mathrm{P}(A_n | A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \cdot \mathrm{P}(A_{n-1} | A_1 \cap A_2 \cap \ldots \cap A_{n-2}) \cdot \ldots \cdot \mathrm{P}(A_2 | A_1) \cdot \mathrm{P}(A_1).
\end{aligned}$$

Satz 1.1.30 *Formel von Bayes*

Sei $(\Omega, \mathfrak{A}, \mathrm{P})$ ein Wahrscheinlichkeitsraum. $A_1, \ldots, A_n \in \mathfrak{A}$ mögen eine Zerlegung von Ω bilden, d. h.

$$A_i \cap A_j = \emptyset \quad \text{für } i \neq j \quad \text{und} \quad A_1 \cup \ldots \cup A_n = \Omega.$$

Zudem sei $\mathrm{P}(A_i) > 0$ für $i = 1, \ldots, n$.

Dann gilt für $B \in \mathfrak{A}$ der *Satz der totalen Wahrscheinlichkeit*:

$$\mathrm{P}(B) = \sum_{j=1}^{n} \mathrm{P}(B|A_j)\mathrm{P}(A_j) \tag{1.12}$$

und, falls $\mathrm{P}(B) > 0$, die *Formel von Bayes*:

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i)\mathrm{P}(A_i)}{\sum_{j=1}^{n} \mathrm{P}(B|A_j)\mathrm{P}(A_j)}. \tag{1.13}$$

Beweis: Es ist nach Definition $\mathrm{P}(A_i|B) = \mathrm{P}(A_i \cap B)/\mathrm{P}(B)$. Weil $A_1, \ldots, A_n$ eine Zerlegung von Ω bilden, gilt weiter:

$$\mathrm{P}(B) = \mathrm{P}(B \cap A_1) + \ldots + \mathrm{P}(B \cap A_n) = \mathrm{P}(B|A_1)\mathrm{P}(A_1) + \ldots + \mathrm{P}(B|A_n)\mathrm{P}(A_n).$$

Damit folgt:

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B \cap A_i)}{\mathrm{P}(B \cap A_1) + \ldots + \mathrm{P}(B \cap A_n)} = \frac{\mathrm{P}(B|A_i)\mathrm{P}(A_i)}{\mathrm{P}(B|A_1)\mathrm{P}(A_1) + \ldots + \mathrm{P}(B|A_n)\mathrm{P}(A_n)}.$$

Die heute als Formel von Bayes bezeichnete Beziehung stammt in dieser Form nicht von Thomas Bayes (1702-1761). Man erhält sie aber über eine geeignete Verallgemeinerung seiner Ideen.

Beispiel 1.1.31 *codierte Nachrichten*

Wenn codierte Nachrichten gesendet werden, gibt es bisweilen Übertragungsfehler. Beim Morsen kann aus einem gesendeten Punkt ein Strich bzw. aus einem Strich ein Punkt beim Empfänger werden. Die Morsezeichen kommen in etwa im Verhältnis 3:4 (Punkte zu Strichen) vor, d. h. wir können ausgehen von

$$\mathrm{P}(\text{Punkt gesendet}) = \frac{3}{7}, \quad \mathrm{P}(\text{Strich gesendet}) = \frac{4}{7}.$$

Treten Vertauschungen aufgrund von Störungen jeweils mit der Wahrscheinlichkeit 1/8 auf, so erhalten wir mit der Bayesschen Formel:

$$\begin{aligned}&\mathrm{P}(\text{Punkt gesendet}|\text{Punkt empfangen})\\ &\qquad = \mathrm{P}(\text{Punkt empfangen}|\text{Punkt gesendet}) \cdot \frac{\mathrm{P}(\text{Punkt gesendet})}{\mathrm{P}(\text{Punkt empfangen})}.\end{aligned}$$

Nun wissen wir, dass $\mathrm{P}(\text{Punkt empfangen}|\text{Punkt gesendet}) = 7/8$. Weiter ist

$$\begin{aligned}&\mathrm{P}(\text{Punkt empfangen})\\ &\quad = \mathrm{P}(\text{Punkt empfangen} \cap \text{Punkt gesendet}) + \mathrm{P}(\text{Punkt empfangen} \cap \text{Strich gesendet})\\ &\quad = \mathrm{P}(\text{Punkt empfangen}|\text{Punkt gesendet})\mathrm{P}(\text{Punkt gesendet})\\ &\qquad + \mathrm{P}(\text{Punkt empfangen} \mid \text{Strich gesendet})\mathrm{P}(\text{Strich gesendet})\\ &\quad = \frac{7}{8}\cdot\frac{3}{7} + \frac{1}{8}\cdot\frac{4}{7} = \frac{25}{56}.\end{aligned}$$

Zusammen gibt dies:

$$\mathrm{P}(\text{Punkt gesendet} \mid \text{Punkt empfangen}) = \frac{7}{8}\cdot\frac{3/7}{25/56} = \frac{21}{25}.$$

Aus dem Wissen, dass das Ereignis B eintritt, bzw. eingetroffen ist, ergibt sich offensichtlich keine Konsequenz für die Wahrscheinlichkeit von A, wenn gilt:

$$\mathrm{P}(A|B) = \mathrm{P}(A),$$

oder, äquivalent:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B).$$

In diesem Fall werden A und B stochastisch unabhängig genannt. Die letzte Gleichung wird dabei als Definitionsgleichung genommen, weil sie ohne die Bedingung $\mathrm{P}(B) > 0$ auskommt und weil sie sich geeignet verallgemeinern lässt.

Definition 1.1.32 *stochastische Unabhängigkeit*

$(\Omega, \mathfrak{A}, \mathrm{P})$ sei ein Wahrscheinlichkeitsraum. $A_1, \ldots, A_n \in \mathfrak{A}$ heißen *stochastisch unabhängig*, wenn für alle Teilfolgen $A_{i_1}, \ldots, A_{i_k}$ von $A_1, \ldots, A_n$ gilt:

$$\mathrm{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathrm{P}(A_{i_1}) \cdot \mathrm{P}(A_{i_2}) \cdot \ldots \cdot \mathrm{P}(A_{i_k}). \tag{1.14}$$

Falls (nur) $\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i) \cdot \mathrm{P}(A_j)$ für $1 \leq i, j \leq n$ gilt, heißen die Ereignisse *paarweise unabhängig*.

Sind $A_1, \ldots, A_n$ stochastisch unabhängig, so sind sie auch paarweise unabhängig. Die Umkehrung gilt nicht.

Beispiel 1.1.33 *zweimaliges Würfeln*

Das Zufallsexperiment bestehe im zweimaligen Würfeln. Bei jeder Durchführung werden die Augenzahlen der Würfe notiert. Dann ist $\Omega = \{(i,j)|\, i,j = 1,\ldots,6\}$, $\mathfrak{A}$ ist die Potenzmenge von Ω. Eine geeignete Wahrscheinlichkeitsverteilung ist wieder die Gleichverteilung:

$$\mathrm{P}(A) = \sum_{\omega\in A} \frac{1}{36} = \frac{|A|}{|\Omega|}.$$

Dabei sind $|\Omega|$, $|A|$ die Anzahl der Elemente von Ω bzw. A.
Konkret werden drei Ereignisse betrachtet:

$$A_1 = \{(1,j)|\, j = 1,\ldots,6\}, \quad A_2 = \{(i,1)|\, i = 1,\ldots,6\}, \quad A_3 = \{(i,i)|\, i = 1,\ldots,6\}.$$

Für diese Ereignisse gilt: $\mathrm{P}(A_i) = 6/36 = 1/6$. Weiter ist für $i \neq j$:

$$\mathrm{P}(A_i \cap A_j) = \mathrm{P}(\{(1,1)\}) = \frac{1}{36} = \mathrm{P}(A_i)\cdot\mathrm{P}(A_j).$$

A_1, A_2, A_3 sind damit paarweise unabhängig. Wegen

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(\{(1,1)\}) = \frac{1}{36} \neq \mathrm{P}(A_1)\cdot\mathrm{P}(A_2)\cdot\mathrm{P}(A_3)$$

sind sie aber nicht stochastisch unabhängig.

Die stochastische Unabhängigkeit vereinfacht die Berechnung von Wahrscheinlichkeiten der Form $\mathrm{P}(A_1 \cap \ldots \cap A_n)$ stark bzw. ermöglicht sie erst. Dementsprechend spielt in der Statistik die Voraussetzung der Unabhängigkeit eine zentrale Rolle.

1.2 Zufallsvariablen

1.2.1 Univariate Zufallsvariablen

In vielen Fällen interessieren bei einem Zufallsexperiment $(\Omega, \mathfrak{A}, \mathrm{P})$ nicht die Ergebnisse ω selbst, sondern numerische Werte, die den Ergebnissen zugeordnet sind und darüber festgelegte Ereignisse. Grundlegend sind dabei Ereignisse, die sich in der Form ‚der Wert liegt in einem vorgegebenen Intervall' ausdrücken lassen. Da die kleinste, alle Intervalle umfassende σ-Algebra die Borelsche σ-Algebra $\mathfrak{B}$ ist, erscheint die folgende Definition plausibel.

Definition 1.2.1 *eindimensionale Zufallsvariable*

$(\Omega, \mathfrak{A}, \mathrm{P})$ sei ein Wahrscheinlichkeitsraum. Eine Abbildung $X : \Omega \to \mathbb{R}$ heißt (eindimensionale) *Zufallsvariable* falls für alle $B \in \mathfrak{B}$ gilt:

$$\{X \in B\} := X^{-1}(B) = \{\omega|\, \omega \in \Omega, X(\omega) \in B\} \in \mathfrak{A}. \tag{1.15}$$

Wir setzen

$$\{a < X \leq b\} := \{X \in (a,b]\}, \quad \{X = a\} := \{X \in \{a\}\}, \quad \text{etc.}$$

und sagen

für $\{X \in B\}$, dass X einen Wert aus B annimmt,
für $\{X = a\}$, dass X den Wert a annimmt etc.

Der folgende Satz zeigt, dass wir uns im Wesentlichen auf die Ereignisse der Form $\{X \leq a\}$ beschränken können, wenn es darum geht, zu überprüfen, ob eine Abbildung X eine Zufallsvariable ist. Wir werden aber diesen Bereich nicht weit ausdehnen. Die in unserem Rahmen zu betrachtenden Abbildungen sind zumeist gutartig, d. h. erfüllen die geforderte Eigenschaft. dass wir die formale Definition überhaupt angeben, rührt daher, dass wir die grundlegende Struktur für ein Verständnis des Folgenden für wichtig erachten.

Satz 1.2.2 *Charakterisierung einer Zufallsvariablen*

$(\Omega, \mathfrak{A}, \mathrm{P})$ sei ein Wahrscheinlichkeitsraum. Eine Abbildung $X : \Omega \to \mathbb{R}$ ist genau dann eine Zufallsvariable, wenn $\{X \leq a\} \in \mathfrak{A}$ für alle $a \in \mathbb{R}$.

Beispiel 1.2.3 *zwei Würfel*

Es wird mit einem roten und einem weißen Würfel gewürfelt. Sei

$$X = \text{Summe der Augenzahlen der beiden Würfel.}$$

Dann ist $\Omega = \{(i, j) | \, i, j \in \mathbb{N}, 1 \leq i, j \leq 6\}$ und die Ereignisalgebra über Ω ist $\mathfrak{P}(\Omega)$. Damit ist X eine Zufallsvariable. Denn stets ist $X^{-1}(B) \in \mathfrak{P}(\Omega)$.

Beispiel 1.2.4 *Schießen auf eine Zielscheibe*

Als Zufallsexperiment wird das Schießen auf eine kreisrunde Zielscheibe betrachtet. Zur Vereinfachung wird davon ausgegangen, dass die Zielscheibe auch getroffen wird. Dann seien die Punkte auf der Scheibe mit den Realisationsmöglichkeiten des Experimentes identifiziert. Die Punkte wiederum seien durch ihre Koordinaten bzgl. eines Koordinatensystems gegeben, das seinen Ursprung im Mittelpunkt der Scheibe hat. Der Stichprobenraum ist also, wenn der Radius der Zielscheibe gleich t ist:

$$\Omega = \{\omega | \, \omega = (\omega_1, \omega_2), \; \omega_1^2 + \omega_2^2 \leq t^2\}.$$

Als Ereignisalgebra wird naheliegender Weise $\mathfrak{B}_\Omega^2$ gewählt. Es interessiert nun die Abbildung $X =$'Entfernung des Treffers vom Mittelpunkt der Scheibe',

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

Damit X eine Zufallsvariable ist, muss nach dem Satz gelten:

$$X^{-1}((l, u]) = \{\omega | \, \omega \in \Omega, \; l < X(\omega) \leq u\} \in \mathfrak{B}_\Omega^2.$$

Für $l < 0 < u < t$ ist $X^{-1}((l, u])$ eine Kreisscheibe. Daher ist zu zeigen, dass mit den zweidimensionalen Intervallen $[a_1, b_1] \times [a_2, b_2]$ auch Kreisscheiben zu $\mathfrak{B}^2$ gehören. Dazu

wird die rechte Hälfte einer Kreisscheibe mit dem Radius u, $Kr = \{\omega | \, \omega = (\omega_1, \omega_2),\ \omega_1^2 + \omega_2^2 \leq u^2,\ \omega_1 > 0\}$, sukzessive mittels Rechtecken der Form

$$A_{i,n} = \left[u \cdot \frac{i}{n}, u \cdot \frac{i}{n} \right] \times \left[-u \cdot \sqrt{1 - \left(\frac{i}{n}\right)^2}, u \cdot \sqrt{1 - \left(\frac{i}{n}\right)^2} \right]$$

umschrieben:

$$B_n = \bigcup_{i=0}^{n-1} A_{i,n}.$$

Es gilt offensichtlich $B_1 \supset B_2 \supset B_4 \supset \ldots \supset \ldots \supset B_{2n} \supset \ldots$, so dass

$$\lim_{n \to \infty} B_n = \bigcap_{n=1}^{\infty} B_n = Kr.$$

Entsprechend erhält man die linke Hälfte.

Vermittelt durch die Zufallsvariable X wird allen Ereignissen aus $\mathfrak{B}$ eine Wahrscheinlichkeit $\mathrm{P}_X(B)$ zugeordnet:

$$\mathrm{P}_X(B) := \mathrm{P}(X \in B), \quad B \in \mathfrak{B}.$$

Durch X wird also ein Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathfrak{B})$ induziert. Das Diagramm 1.1 veranschaulicht dies.

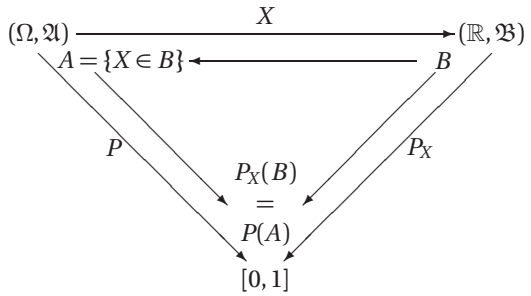


Abb. 1.1: *Zur induzierten Wahrscheinlichkeitsverteilung*

Definition 1.2.5 *Wahrscheinlichkeitsverteilung einer eindimensionalen Zufallsvariablen*

$(\Omega, \mathfrak{A}, \mathrm{P})$ sei ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Die von X auf $(\mathbb{R}, \mathfrak{B})$ induzierte Wahrscheinlichkeitsverteilung P_X,

$$\mathrm{P}_X(B) := \mathrm{P}(X \in B) \quad \text{für } B \in \mathfrak{B}, \tag{1.16}$$

heißt die *(Wahrscheinlichkeits-) Verteilung* von X.

Beispiel 1.2.6 *Weg zur Arbeit - Fortsetzung*

Im Eingangsbeispiel 1.1.9 wird durch die Entfernung der Person ω zu ihrer Arbeitsstätte eine Zufallsvariable definiert:

$$X:\Omega\to\mathbb{R} \quad \text{mit} \quad X(\omega)=i, \quad \text{falls } \omega\in A_i, \qquad i=1,\ldots,5.$$

Die durch X auf $(\mathbb{R},\mathfrak{B})$ definierte Verteilung ist dann offensichtlich schon festgelegt durch

$$\mathrm{P}_X(\{i\})=\mathrm{P}(X=i) \qquad i=1,\ldots,5.$$

Für $B\in\mathfrak{B}$ mit $\{1,2,3,4,5\}\cap B=\emptyset$ gilt speziell

$$\mathrm{P}_X(B)=0.$$

Beispiel 1.2.7 *Zerfall einer radioaktiven Substanz*

Bei dem Zerfallsprozess einer radioaktiven Substanz ist eine interessierende Zufallsvariable X die Wartezeit zwischen dem Zerfall zweier Atome. Hier gilt

$$\mathrm{P}_X([0,t])=1-e^{-\vartheta t}.$$

Dabei ist ϑ eine von der Art der Substanz abhängige Konstante. Wegen $\mathrm{P}_X([0,\infty))=1$ ist $\mathrm{P}_X((-\infty,0))=0$, so dass für $t>0$:

$$\mathrm{P}(X\le t)=\mathrm{P}_X([0,t])=1-e^{-\vartheta t}.$$

Wahrscheinlichkeiten für allgemeinere Ereignisse der Form $\{X\in B\}$ können selbstverständlich ebenfalls angegeben werden. Darauf wird später noch eingegangen.

Im letzten Beispiel wurde die Verteilung einer Zufallsvariablen angegeben, ohne den ursprünglichen Wahrscheinlichkeitsraum $(\Omega,\mathfrak{A},\mathrm{P})$ zu benennen. In der Statistik interessiert i. d. R. nicht so sehr der Wahrscheinlichkeitsraum, auf dem die Zufallsvariable definiert ist, als vielmehr ihre Verteilung. Durch diese erfolgt bereits die Bewertung der Ereignisse $B\in\mathfrak{B}^k$ entsprechend der Chance ihres Eintretens.

Man geht deshalb so vor, dass man für eine Zufallsvariable X eine Wahrscheinlichkeitsverteilung P_X über dem Stichprobenraum $(\mathbb{R},\mathfrak{B})$ vorgibt, ohne den zugrundeliegenden Wahrscheinlichkeitsraum näher anzugeben. Im Folgenden wird häufig in dieser Weise verfahren und von dem speziellen Wahrscheinlichkeitsraum $(\mathbb{R},\mathfrak{B},\mathrm{P}_X)$ oder $(\mathfrak{X},\mathfrak{D},\mathrm{P}_X)$ – mit $\mathfrak{X}\subset\mathbb{R}$ geeignet definiert und $\mathfrak{D}=\mathfrak{B}^k_{\mathfrak{X}}$ – ausgegangen. Dann brauchen wir nur noch in Ausnahmefällen die Wahrscheinlichkeitsverteilung von X mit P_X zu bezeichnen. I. d. R. reicht es, einfach das Symbol P zu verwenden, ohne dass dies zu Missverständnissen führt.

Da X nach Satz 1.2.2 eine Zufallsvariable ist, wenn $\{X\le a\}\in\mathfrak{A}$ für alle $a\in\mathbb{R}$, betrachtet man in der Statistik meist nur dadurch festgelegte Funktionen.

Definition 1.2.8 *eindimensionale Verteilungsfunktion*

X sei eine Zufallsvariable mit der Verteilung P. Dann heißt

$$F(x):=\mathrm{P}((-\infty,x])=\mathrm{P}(X\le x) \tag{1.17}$$

die *Verteilungsfunktion* von X.

Beispiel 1.2.9 *Glücksspiele - Fortsetzung von Seite 2*

Beim einmaligen Werfen einer Münze erhalten wir mit der Vereinbarung $P(\{K\}) = p$, $(0 < p < 1)$, eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathfrak{A})$. Es gilt $P(\{Z\}) = 1 - p$, $P(\emptyset) = 0$ und $P(\Omega) = 1$.

Sei X definiert durch $X(K) = 1$, $X(Z) = 0$. Dann ist X eine Zufallsvariable. Wegen

$$\{\omega | X(\omega) \leq x\} = \begin{cases} \emptyset & \text{für } x < 0 \\ \{Z\} & \text{für } 0 \leq x < 1 \\ \Omega & \text{für } x \geq 1 \end{cases}$$

hat X die Verteilungsfunktion

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{für } x < 0 \\ 0.5 & \text{für } 0 \leq x < 1 \\ 1 & \text{für } x \geq 1\,. \end{cases}$$

Satz 1.2.10 *Eigenschaften von eindimensionalen Verteilungsfunktionen*

$X : (\Omega, \mathfrak{A}, P) \to (\mathbb{R}, \mathfrak{B})$ sei eine Zufallsvariable mit der Verteilungsfunktion $F(x)$. Dann gilt:

(1) $0 \leq F(x) \leq 1$ für alle $x \in \mathbb{R}$,

(2) $F(x) \leq F(x')$ für alle $x < x'$,

(3) $\lim_{n\to\infty} F(x) = 1$, $\lim_{n\to-\infty} F(x) = 0$,

(4) $F(x)$ ist rechtsseitig stetig in x. Das bedeutet: Ist x fest und ist x'_n eine Folge mit $\lim_{n\to\infty} x'_n = x$, $x'_n \geq x$, so gilt:

$\lim_{n\to\infty} F(x'_n) = F(x)$

Beweis: Wir beweisen nur die erste Aussage von (3).

Sei dazu $x_1, x_2, \ldots$ eine Folge reeller Zahlen mit $x_n < x_{n+1}$ und $\lim_{n\to\infty} x_n = \infty$. Um den Satz 1.1.17 anwenden zu können, wird $A_n := (-\infty, x_n]$ gesetzt. Dann gilt:

$$A_1 \subset A_2 \subset \ldots \quad \text{und} \quad \bigcup_{n=1}^{\infty} A_n = \mathbb{R}.$$

Mit dem Satz 1.1.17 folgt:

$$\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} P_X((-\infty, x_n]) = P_X\left(\bigcup_{i=1}^{n} (-\infty, x_n]\right) = P_X(\mathbb{R}) = 1.$$

Die Aussage 2 des Satzes können wir sogar präzisieren zu

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1).$$

Damit lassen sich alle Wahrscheinlichkeiten für Intervalle durch die Verteilungsfunktion angeben. Wahrscheinlichkeiten für Ereignisse A, die auf kompliziertere Weise aus Intervallen resultieren, wollen wir im folgenden durch

$$\mathrm{P}(X \in A) = \int_A \mathrm{d}F(x) \tag{1.18}$$

angeben. Für Intervalle $A = (a, b]$ schreiben wir auch

$$\int_A \mathrm{d}F(x) = \int_a^b \mathrm{d}F(x) = F(b) - F(a).$$

Insbesondere gilt für die Vereinigung von disjunkten Intervallen $(a_1, b_1], \ldots, (a_k, b_k]$, $A = \bigcup_{i=1}^k (a_i, b_i]$:

$$\int_A d\,F(x) = \sum_{i=1}^k \int_{a_i}^{b_i} \mathrm{d}F(x) = \sum_{i=1}^k (F(b_i) - F(a_i)).$$

Formal ist das die Schreibweise des *Riemann-Stieltjes-Integrals*. Auf das Riemann-Stieltjes-Integral gehen wir im Anhang A ausführlicher ein.

Die Definition der Verteilungsfunktion einer Zufallsvariablen zieht einige wichtige Konzepte nach sich. Zunächst wird eine Einteilung der Zufallsvariablen vorgenommen.

Definition 1.2.11 *diskrete und stetige eindimensionale Zufallsvariablen*

Eine Zufallsvariable X heißt *diskret*, falls es eine höchstens abzählbare Teilmenge $W \subset \mathbb{R}$ und eine auf $\mathbb{R}$ definierte Funktion $f(x)$ gibt mit

$$\int_A \mathrm{d}F(x) = \sum_{x \in A \cap W} f(x). \tag{1.19}$$

Die (in der Regel nur für $x \in W$ betrachtete) Funktion $f : \mathbb{R} \to [0, 1]$,

$$f(x) = \mathrm{P}(X = x)$$

heißt *(Zähl-) Dichtefunktion* oder kurz *(Zähl-) Dichte*.

X heißt *stetig*, falls es eine als *(stetige) Dichtefunktion* oder kurz als *Dichte* bezeichnete Funktion $f : \mathbb{R} \to \mathbb{R}_+$ gibt mit

$$\int_A d\,F(x) = \int_A f(x)\,\mathrm{d}x. \tag{1.20}$$

X ist also diskret, wenn es eine höchstens abzählbare Teilmenge $W \subset \mathbb{R}$ gibt mit $F(x) = \sum_{x' \le x} \mathrm{P}(X = x')$. X ist stetig, wenn es eine Funktion $f : \mathbb{R} \to \mathbb{R}$ gibt mit

$$F(x) = \int_{-\infty}^x f(u)\,\mathrm{d}u.$$

Die Bezeichnungen 'Zähldichte' und 'stetige Dichte' sind beide etwas irreführend. Bei einer Zähldichte wird nicht nur ausgezählt, wieviele Realisationsmöglichkeiten zu der interessierenden Menge gehören. Die einzelnen Realisationsmöglichkeiten können ja durchaus verschiedene Werte zugeordnet bekommen. Auch die stetige Dichte braucht nicht überall stetig zu sein. Die Bezeichnung rührt vielmehr von der in der Maßtheorie betrachteten 'absoluten Stetigkeit' her. Allerdings ist bei einer stetigen Zufallsvariablen die Verteilungsfunktion stetig. dass wir diese Bezeichnungen dennoch verwenden, ist darin begründet, dass wir im Folgenden dann einfach von Dichten sprechen können und nicht die sonst notwendige und ermüdende Unterscheidung in Wahrscheinlichkeits- und Dichtefunktion durchhalten müssen. Auch Fallunterscheidungen bei der Formulierung etlicher Aussagen lassen sich mit dem Riemann-Stieltjes-Integral vermeiden.

Beispiel 1.2.12 *zwei Würfel - Fortsetzung von Seite 19*

Die Zähldichte von X, der Summe der Augenzahlen der beiden Würfel, geben wir in Form einer Tabelle an.

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)=\mathrm{P}(X=x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Beispiel 1.2.13 *Simpson-Verteilung*

Die *Simpson-Verteilung* ist durch eine dreieckförmige Dichtefunktion chrakterisiert:

$$f(x)=\begin{cases}\frac{1}{b}\left(1-\frac{1}{b}|x|\right) & \text{für } -b<x<b\\ 0 & \text{sonst}\end{cases}.$$

Man macht sich leicht klar, dass durch $f(x)$ tatsächlich eine Verteilung festgelegt wird.

1.2.2 Multivariate Zufallsvariablen

Wir wollen nun die Definitionen und Ergebnisse des vorigen Abschnittes auf den Fall verallgemeinern, dass in einem Experiment gleichzeitig mehrere Zufallsvariablen beobachtet werden. Zunächst erscheint die folgende Definition eine plausible Erweiterung von Definition 1.2.1 zu sein.

Definition 1.2.14 *k-dimensionale Zufallsvariable*

$(\Omega,\mathfrak{A},\mathrm{P})$ sei ein Wahrscheinlichkeitsraum. Eine Abbildung $\boldsymbol{X}=(X_1,\ldots,X_k):\Omega\to\mathbb{R}^k$ heißt *k-dimensionale Zufallsvariable* (oder für $k>1$ auch *Zufallsvektor*), falls für alle $B\in\mathfrak{B}^k$ gilt:

$$\{\boldsymbol{X}\in B\}:=\boldsymbol{X}^{-1}(B)=\{\omega|\,\omega\in\Omega,\,\boldsymbol{X}(\omega)\in B\}\in\mathfrak{A}. \tag{1.21}$$

Wir sagen entsprechend zum eindimensionalen Fall für $\{\boldsymbol{X}\in B\}$, dass $\boldsymbol{X}$ einen Wert aus B annimmt und für $\{\boldsymbol{X}=\boldsymbol{a}\}:=\{\boldsymbol{X}\in\{\boldsymbol{a}\}\}$, dass $\boldsymbol{X}$ den Wert $\boldsymbol{a}$ annimmt etc.

Satz 1.2.15 *Charakterisierung einer k-dimensionalen Zufallsvariablen*

$(\Omega, \mathfrak{A}, \mathrm{P})$ sei ein Wahrscheinlichkeitsraum. Eine Abbildung $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ ist genau dann eine Zufallsvariable, wenn $\{\boldsymbol{X} \leq \boldsymbol{a}\} \in \mathfrak{A}$ für alle $\boldsymbol{a} \in \mathbb{R}^k$.

Satz 1.2.16 *k-dimensionale Zufallsvariable aus eindimensionalen Zufallsvariablen*

$X_1, \ldots, X_k$ seien eindimensionale Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, \mathrm{P})$. Dann ist

$$\boldsymbol{X} = (X_1, \ldots, X_k) \colon \Omega \to \mathbb{R}^k, \quad \boldsymbol{X}(\omega) := (X_1(\omega), \ldots, X_k(\omega)),$$

eine k-dimensionale Zufallsvariable.

Beispiel 1.2.17 *zwei Würfel - Fortsetzung von Seite 19*

Beim Würfeln mit einem roten und einem weißen Würfel seien

$X =$ Summe der Augenzahlen der beiden Würfel
$Y =$ Differenz der Augenzahlen des roten minus und des weißen Würfels.

Es ist $\Omega = \{(i, j) | \, i, j \in \mathbb{N},\ 1 \leq i, j \leq 6\}$ und $\mathfrak{P}(\Omega)$ ist die Ereignisalgebra über Ω. Damit ist (X, Y) eine Zufallsvariable. Denn stets ist $(X, Y)^{-1}(B) \in \mathfrak{P}(\Omega)$.

So wie im eindimensionalen Fall ist auf dem Hintergrund von Satz 1.2.15 die Verteilung $\mathrm{P}_{\boldsymbol{X}}$ schon festgelegt, wenn die Wahrscheinlichkeiten

$$\mathrm{P}_{\boldsymbol{X}}((-\infty, a_1] \times \ldots \times (-\infty, a_k]) = \mathrm{P}(X_1 \leq a_1, \ldots, X_k \leq a_k) = \mathrm{P}(\boldsymbol{X} \leq \boldsymbol{a})$$

für alle $\boldsymbol{a} = (a_1, \ldots, a_k)$ bekannt sind. Daher betrachtet man auch hier meist nur entsprechend festgelegte Funktionen.

Definition 1.2.18 *Verteilungsfunktion einer k-dimensionalen Zufallsvariablen*

$\boldsymbol{X} = (X_1, \ldots, X_k)$ sei eine k-dimensionale Zufallsvariable mit der Verteilung P. Dann heißt

$$F(\boldsymbol{x}) = F(x_1, \ldots, x_k) := \mathrm{P}(\boldsymbol{X} \leq \boldsymbol{x}) \tag{1.22}$$

die *Verteilungsfunktion* von $\boldsymbol{X} = (X_1, \ldots, X_k)$.

Satz 1.2.19 *Eigenschaften von k-dimensionalen Verteilungsfunktionen*

$\boldsymbol{X} = (X_1, \ldots, X_k) : (\Omega, \mathfrak{A}, \mathrm{P}) \to (\mathbb{R}^k, \mathfrak{B}^k)$ sei eine Zufallsvariable mit der Verteilungsfunktion $F(\boldsymbol{x})$. Dann gilt:

(1) $0 \le F(\boldsymbol{x}) \le 1$ für alle $\boldsymbol{x} \in \mathbb{R}^k$.

(2) Im Fall $k = 2$ gilt für alle $x_1 < x_1'$, $x_2 < x_2'$:

$$F(x_1', x_2') + F(x_1, x_2) - F(x_1, x_2') - F(x_2', x_1) \ge 0;$$

(3) $F(\boldsymbol{x}) \le F(\boldsymbol{x}')$ falls $x_i \le x_i'$ $(i = 1, \ldots, k)$;

(4) $\lim_{x_1 \to \infty, \ldots, x_k \to \infty} F(x_1, \ldots, x_k) = F(\infty, \ldots, \infty) = 1$;

(5) $\lim_{x_i \to -\infty} F(x_1, \ldots, x_k) = F(x_1, \ldots, -\infty, \ldots, x_k) = 0$, $(i = 1, \ldots, k)$;

(6) $F(\boldsymbol{x})$ ist rechtsseitig stetig in jeder Komponente i.

Beweis: Wir beweisen nur (2). Zur Vereinfachung der Schreibweise setzen wir $(X_1, X_2) = (X, Y)$. Die folgenden Rechnungen werden durch die Skizze 1.2 verdeutlicht:

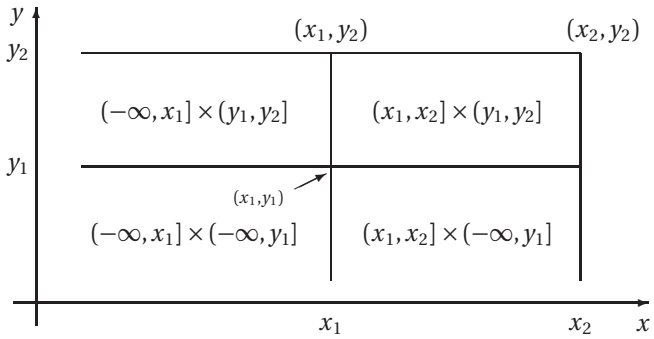


Abb. 1.2: *Zur bivariaten Verteilungsfunktion*

$$\begin{aligned}
&F(x_2, y_2) = \mathrm{P}(X \le x_2, Y \le y_2) \\
&= \mathrm{P}(X \le x_1, Y \le y_1) + \mathrm{P}(x_1 < X \le x_2, Y \le y_1) + \mathrm{P}(X \le x_1, y_1 < Y \le y_2) \\
&\quad + \mathrm{P}(x_1 < X \le x_2, y_1 < Y \le y_2) \\
&= \mathrm{P}(X \le x_1, Y \le y_1) + \mathrm{P}(X \le x_2, Y \le y_1) - \mathrm{P}(X \le x_1, Y \le y_1) + \mathrm{P}(X \le x_1, Y \le y_2) \\
&\quad - \mathrm{P}(X \le x_1, Y \le y_1) + \mathrm{P}(x_1 < X \le x_2, y_1 < Y \le y_2) \\
&= \mathrm{P}(X \le x_1, Y \le y_2) + \mathrm{P}(X \le x_2, Y \le y_1) - \mathrm{P}(X \le x_1, Y \le y_1) + \mathrm{P}(x_1 < X \le x_2, y_1 < Y \le y_2) \\
&= F(x_1, y_2) + F(x_2, y_1) - F(x_1, y_1) + \mathrm{P}(x_1 < X \le x_2, y_1 < Y \le y_2).
\end{aligned}$$

Mit $\mathrm{P}(x_1 < X \le x_2, y_1 < Y \le y_2) \ge 0$ folgt die Behauptung.

Beispiel 1.2.20

Sei

$$F(x, y) = \begin{cases} 0 & \text{für } x + y < 1 \\ 1 & \text{für } x + y \ge 1 \end{cases}.$$

$F(x,y)$ ist keine Verteilungsfunktion, denn es gilt

$$F(1,1)+F(0,0)-F(1,0)-F(0,1)=1+0-1-1=-1\,.$$

Beispiel 1.2.21 *zweidimensionale logistische Verteilung*

Einige der Eigenschaften sollen für die zweidimensionale *logistische Verteilung* verifiziert werden. Die Verteilungsfunktion ist

$$F(x,y)=\frac{1}{1+e^{-x}+e^{-y}+e^{(x+y)}}.$$

(1) Wegen $e^z>0$ für alle $z\in\mathbb{R}$ gilt: $1+e^{-x}+e^{-y}+e^{-(x+y)}>1$. Daraus folgt

$$1>\frac{1}{1+e^{-x}+e^{-y}+e^{(x+y)}}>0\,.$$

(2) Seien $x_1\le x_2,\ y_1\le y_2$. Dann ist $x_1+y_1\le x_2+y_2$ und es gilt: $e^{-x_1}\ge e^{-x_2},\ e^{-y_1}\ge e^{-y_2},\ e^{-x_1}\ge e^{-x_2}$. Daraus folgt

$$\frac{1}{1+e^{-x_1}+e^{-y_1}+e^{-x_1-y_1}}\le\frac{1}{1+e^{-x_2}+e^{-y_2}+e^{-x_2-y_2}},$$

d. h. $F(x_1,y_1)\le F(x_2,y_2)$.

(3) Es ist $\lim_{z\to\infty}e^{-z}=0$. Daher gilt

$$\lim_{x,y\to\infty}F(x,y)=\lim_{x,y\to\infty}\frac{1}{1+e^{-x}+e^{-y}+e^{-(x+y)}}=\frac{1}{1+0+0+0}=1.$$

Die Schreibweise des *Riemann-Stieltjes-Integrals* lässt sich auch bei mehrdimensionalen Zufallsvariablen anwenden. Dann ist für $A\in\mathfrak{B}^k$

$$\int_A \mathrm{d}F(\boldsymbol{x})=\mathrm{P}(\boldsymbol{X}\in A). \tag{1.23}$$

Die Einteilung der Zufallsvariablen in diskrete und stetige überträgt sich vom univariaten Fall.

Definition 1.2.22 *diskrete und stetige k-dimensionale Dichtefunktionen*

Eine Zufallsvariable $\boldsymbol{X}=(X_1,\ldots,X_k)$ heißt *diskret*, falls es eine höchstens abzählbare Teilmenge $W\subset\mathbb{R}^k$ und eine auf $\mathbb{R}^k$ definierte *(Zähl-) Dichtefunktion* oder kurz *(Zähl-) Dichte* $f(\boldsymbol{x})$ gibt mit

$$\int_A \mathrm{d}F(\boldsymbol{x})=\sum_{\boldsymbol{x}\in A\cap W} f(\boldsymbol{x}). \tag{1.24}$$

$\boldsymbol{X}$ heißt *stetig*, falls es eine *Dichtefunktion* $f:\mathbb{R}^k\to\mathbb{R}$ gibt mit

$$\int_A \mathrm{d}F(\boldsymbol{x}) = \int_A \cdots \int f(\boldsymbol{x})\,d\boldsymbol{x} = \int_A \cdots \int f(x_1,\ldots,x_k)\,\mathrm{d}x_1\ldots\mathrm{d}x_k\,. \tag{1.25}$$

Beispiel 1.2.23 *zwei Würfel - Fortsetzung von Seite 25*

Wir führen das Beispiel 1.2.17 fort. Die Zähldichte von (X,Y), der Summe und der Differenz der Augenzahlen der beiden Würfel geben wir in der Tabelle 1.2 an. Die nicht besetzten Felder haben den Wert Null.

$X\backslash Y$	−5	−4	−3	−2	−1	0	1	2	3	4	5
2						$\frac{1}{36}$					
3					$\frac{1}{36}$		$\frac{1}{36}$				
4				$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$			
5			$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		
6		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$	
7	$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$
8		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$	
9			$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		
10				$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$			
11					$\frac{1}{36}$		$\frac{1}{36}$				
12						$\frac{1}{36}$					

Tabelle 1.2: *Bivariate Zähldichte*

Beispiel 1.2.24 *Trinomialverteilung*

Die *Trinomialverteilung* ist durch die Zähldichte

$$f(x,y) = \begin{cases} \dfrac{n!}{x!y!(n-x-y)!}\cdot p_1^x\cdot p_2^y\cdot(1-p_1+p_2)^{n-x-y} & \\ \qquad\qquad\qquad\qquad \text{für } x,y=0,1,\ldots,n,\ x+y\le n & \cdot \\ 0 \qquad\qquad \text{sonst} & \end{cases}$$

definiert. Sie hängt von den Parametern p_1,p_2 mit $p_1,p_2\ge 0,\ p_1+p_2\le 1$ und n ab.

Beispiel 1.2.25 *Gleichverteilung mit dreieckförmigen Träger*

Wir betrachten folgende zweidimensionale stetige Zufallsvariable mit der Dichtefunktion

$$f(x,y) = \begin{cases} 2 & \text{für } 0\le x\le y<1 \\ 0 & \text{sonst} \end{cases}.$$

Die Verteilungsfunktion lautet dann für $0\le x\le y\le 1$:

$$F(x,y)=\int_0^x\int_u^y 2\,\mathrm{d}v\,\mathrm{d}u=\int_0^x(2\cdot y-2\cdot u)\,\mathrm{d}u=2\cdot x\cdot y-x^2,$$

für $0\le x\le 1,\ 1<y<\infty$:

$$F(x,y)=\int_0^x\int_u^1 2\,\mathrm{d}v\,\mathrm{d}u=\int_0^x(2-2\cdot u)\,\mathrm{d}u=2\cdot x-x^2$$

und für $0\le y\le 1,\ y\le x$:

$$F(x,y)=\int_0^y\int_0^v 2\,\mathrm{d}u\,\mathrm{d}v=\int_0^y 2v\,\mathrm{d}v=y^2.$$

Also gilt

$$F(x,y)=\begin{cases}2\cdot x\cdot y-x^2 & \text{für } 0\le x\le y\le 1\\ 2\cdot x-x^2 & \text{für } 0\le x\le 1,\ 1<y<\infty\\ y^2 & \text{für } 0\le y\le 1,\ y\le x\\ 1 & \text{für } x\ge 1,\ y\ge 1\\ 0 & \text{sonst}\end{cases}.$$

Beispiel 1.2.26 *bivariate Normalverteilung*

Die bivariate *Normalverteilung* hat folgende, von den fünf Parametern $\mu_1,\mu_2\in\mathbb{R},\ \sigma_1^2,\sigma_2^2>0,\ -1<\rho<1$ abhängige Dichte:

$$f(x,y)=$$

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\cdot\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2}-2\rho\frac{x-\mu_X}{\sigma_X}\cdot\frac{y-\mu_Y}{\sigma_Y}+\frac{(y-\mu_y)^2}{\sigma_Y^2}\right)\right].$$

Die Konturlinien, d. h. die Kurven konstanter Dichte, sind hier Ellipsen.

1.2.3 Randverteilungen

Oft interessiert man sich für die Verteilung einzelner Komponenten von (X,Y). Wie die Verteilung etwa von X aus der Verteilung von (X,Y) gewonnen werden kann, wird im Folgenden ausgeführt. Ausgehend davon, dass (X,Y) eine zweidimensionale Zufallsvariable ist,

$$(X,Y):\ \Omega\to\mathbb{R}^2,$$

und dass offensichtlich gilt

$$\{X\le x\}=\{\omega|\ \omega\in\Omega,X(\omega)\le x\}=\{\omega|\ \omega\in\Omega,X(\omega)\le x,Y(\omega)<\infty\},$$

erhalten wir:

$$\mathrm{P}(X \leq x) = \mathrm{P}(X \leq x, Y < \infty).$$

Die Verteilung von X ist also gegeben durch

$$\mathrm{P}_X((-\infty, x]) = \mathrm{P}_{X,Y}((-\infty, x] \times (-\infty, \infty)) = \mathrm{P}(X \leq x, Y < \infty).$$

Für die Verteilungsfunktion von X folgt somit:

$$F(x) = F_X(x) = F_{X,Y}(x, \infty).$$

Entsprechend lässt sich die Verteilung von Y aus der Verteilung von (X, Y) gewinnen. Da im Weiteren hauptsächlich die Verteilungsfunktionen bzw. Wahrscheinlichkeitsfunktionen und Dichten interessieren, gelangt man zu

Definition 1.2.27 *Randverteilung*

$\boldsymbol{X} = (X_1, \ldots, X_k)$ sei eine k-dimensionale Zufallsvariable mit der Verteilungsfunktion $\mathrm{P}_{\boldsymbol{X}}$. Die *Randverteilung* von $\boldsymbol{X}_1 = (X_1, \ldots, X_i)$ ist dann durch $F_{\boldsymbol{X}_1}$ gegeben,

$$F_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = F_{\boldsymbol{X}}(x_1, \ldots, x_i, \infty, \ldots, \infty). \tag{1.26}$$

In der Definition geschieht die Beschränkung auf die ersten i Komponenten nur aus schreibtechnischen Gründen.

Beispiel 1.2.28 *zweidimensionale logistische Verteilung - Fortsetzung von Seite 27*

Die zweidimensionale logistische Verteilung hat die Verteilungsfunktion

$$F(x, y) = \frac{1}{1 + e^{-x} + e^{-y} + e^{-(x+y)}}.$$

Hat (X, Y) diese gemeinsame Verteilung, $F_{X,Y}(x, y) = F(x, y)$, so sind die Randverteilungen von X und Y:

$$F_X(x) = F_{X,Y}(x, \infty) = \frac{1}{1 + e^{-x} + e^{-\infty} + e^{-(x+\infty)}} = \frac{1}{1 + e^{-x}},$$

$$F_Y(y) = F_{X,Y}(\infty, y) = \frac{1}{1 + e^{-\infty} + e^{-y} + e^{-(\infty+y)}} = \frac{1}{1 + e^{-y}}.$$

X und Y sind jeweils univariat logistisch verteilt.

Sofern die Verteilung von (X, Y) durch eine Dichte gegeben ist, lassen sich auch die Dichten der Randverteilungen daraus gewinnen. Betrachten wir den zweidimensionalen Fall. Für diskrete Zufallsvariablen gilt für die erste Komponente:

$$f(x) = \mathrm{P}(X = x, Y < \infty) = \sum_y \mathrm{P}(X = x, Y = y).$$

Für stetige Zufallsvariablen erhalten wir aus $F_X(x) = F(x, \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, v)\,\mathrm{d}u\,\mathrm{d}v$ sofort:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}y.$$

Beispiel 1.2.29 *zwei Würfel - Fortsetzung von Seite 28*

Für die beiden in Beispiel 1.2.17 definierten Zufallsvariablen der Augenzahlensumme und der Augenzahlendifferenz erhalten wir mit dem Beispiel 1.2.23 durch Bildung der Zeilen- und der Spaltensummen die Randverteilungen von X und Y, siehe Tabelle 1.3.

$X\backslash Y$	-5	-4	-3	-2	-1	0	1	2	3	4	5	$f_X(x)$
2						$\frac{1}{36}$						$\frac{1}{36}$
3					$\frac{1}{36}$		$\frac{1}{36}$					$\frac{2}{36}$
4				$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$				$\frac{3}{36}$
5			$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$			$\frac{4}{36}$
6		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{5}{36}$
7	$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$	$\frac{6}{36}$
8		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{5}{36}$
9			$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$			$\frac{4}{36}$
10				$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$				$\frac{3}{36}$
11					$\frac{1}{36}$		$\frac{1}{36}$					$\frac{2}{36}$
12						$\frac{1}{36}$						$\frac{1}{36}$
$f_Y(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	

Tabelle 1.3: *Bivariate Zähldichte mit Randverteilungen*

Beispiel 1.2.30 *Trinomialverteilung*

(X,Y) sei trinomial-verteilt mit den Parametern $p_1,\ p_2$ und n. Es soll die Randverteilung von Y bestimmt werden:

$$f(y)=\mathrm{P}(Y=y)=\sum_{x=0}^{n-y}\mathrm{P}(X=x,Y=y)=\sum_{x=0}^{n-y}\frac{n!}{x!y!(n-x-y)!}\,p_1^x p_2^y(1-p_1-p_2)^{n-x-y}$$

$$=\frac{n!}{y!(n-y)!}(1-p_2)^{n-y}p_2^y\cdot\sum_{x=0}^{n-y}\frac{(n-y)!}{x!(n-x-y)!}\left(\frac{p_1}{1-p_2}\right)^x\left(1-\frac{p_1}{1-p_2}\right)^{n-x-y}$$

$$=\binom{n}{y}p_2^y(1-p_2)^{n-y}.$$

Y hat eine Binomialverteilung mit den Parametern p_2 und n.

Beispiel 1.2.31 *Gleichverteilung mit dreieckförmigem Träger - Fortsetzung von Seite 28*

Die gemeinsame Dichtefunktion von X und Y ist

$$f(x,y)=\begin{cases}2 & \text{für } 0\le x\le y<1\\ 0 & \text{sonst}\end{cases}.$$

Damit folgt für $0 \le x \le 1$:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y = \int_{x}^{1} 2\,\mathrm{d}y = 2-2x\,.$$

Analog erhalten wir für $0 \le y \le 1$:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y = \int_{0}^{y} 2\,\mathrm{d}y = 2y\,.$$

Beispiel 1.2.32 *bivariate Normalverteilung - Fortsetzung von Seite 29*

(X,Y) sei bivariat normalverteilt. Um die Randverteilung von X zu erhalten, wird die Dichte, vgl. Beispiel 1.2.26, geeignet umgeformt:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\,\frac{x-\mu_X}{\sigma_X}\cdot\frac{y-\mu_Y}{\sigma_Y} + \frac{(y-\mu_y)^2}{\sigma_Y^2}\right)\right]$$

Zunächst setzen wir zur Vereinfachung $u := (x-\mu_X)/\sigma_X$ und $v := (x-\mu_Y)/\sigma_Y$. Dann gilt

$$\begin{aligned}
f(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho\,u\,v + v^2)\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_X} \cdot \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}(u^2 - (\rho\,u)^2 + (\rho\,u)^2 - 2\rho\,u\,v + v^2)\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_X} \cdot \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2\cdot(1-\rho^2)}((1-\rho^2)u^2 + (v-\rho\,u)^2\right] \\
&= \frac{1}{\sqrt{2\pi}\cdot\sigma_X}\exp(-\frac{u^2}{2}) \cdot \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{(v-\rho\cdot u)^2}{2(1-\rho^2)}\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left(-\frac{1}{2}\cdot\frac{(x-\mu_X)^2}{\sigma_X^2}\right) \cdot \frac{1}{\sqrt{2\pi}\cdot\sigma_Y\sqrt{1-\rho^2}} \\
&\qquad \cdot \exp\left[-\frac{1}{2\sigma_Y(1-\rho^2)}\cdot(y-\mu_Y - \frac{\sigma_Y}{\sigma_X}(x-\mu_X))^2\right]\,.
\end{aligned}$$

Es gilt also

$$f(x,y) = g(x)\cdot h(x,y)$$

mit

$$g(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma_X}\exp\left(-\frac{1}{2}\cdot\frac{(x-\mu_X)^2}{\sigma_X^2}\right)$$

und

$$h(x,y) = \frac{1}{\sqrt{2\pi}\cdot\sigma_Y\cdot\sqrt{1-\rho^2}}\cdot\exp\left[-\frac{1}{2\sigma_Y\cdot(1-\rho^2)}\cdot(y-\mu_Y-\frac{\sigma_Y}{\sigma_X}\cdot(x-\mu_X))^2\right].$$

Die Funktion $g(x)$ ist die Dichtefunktion einer univariaten Normalverteilung mit den Parametern μ_X und σ_X^2, während $h(x,y)$ für festes x die Dichtefunktion einer univariaten Normalverteilung mit den Parametern $\mu_Y - \sigma_Y(x-\mu_X)/\sigma_X$ und $\sigma_Y^2(1-\rho^2)$ ist.

Für festes x ist das Integral über $h(x,y)$ mit y als Integrationsvariablen folglich gleich 1. Also ist die Randdichte von X die Dichtefunktion einer Normalverteilung mit den Parametern μ_X und σ_X^2. Eine Interpretation von $h(x,y)$ folgt im nächsten Abschnitt.

1.2.4 Bedingte Verteilungen

Ein wichtiges Konzept der Wahrscheinlichkeitsrechnung, das nun auf Verteilungen übertragen werden soll, ist die bedingte Wahrscheinlichkeit. Dies führt hier zunächst auf bedingte Verteilungsfunktionen.

Definition 1.2.33 *bedingte Verteilungsfunktion bei gegebenem Ereignis*

$\boldsymbol{X} = (X_1,\ldots,X_k)$ sei eine k-dimensionale Zufallsvariable mit der Verteilung P. Für $A \in \mathfrak{B}^k$ mit $\mathrm{P}(\boldsymbol{X} \in A) > 0$ heißt

$$F(\boldsymbol{x}|A) = \mathrm{P}(\boldsymbol{X} \le \boldsymbol{x}|\boldsymbol{X} \in A)$$

die *bedingte Verteilungsfunktion* von $\boldsymbol{X}$, gegeben, dass $\boldsymbol{X}$ einen Wert aus A annimmt.

Beispiel 1.2.34 *gestutzte Poisson-Verteilung*

X sei eine *Poisson-verteilte* Zufallsvariable,

$$\mathrm{P}(X=x) = \begin{cases} e^{-\lambda}\cdot\dfrac{\lambda^x}{x!} & \text{für } x = 0,1,2,3,\ldots \\ 0 & \text{sonst}. \end{cases}$$

$\lambda > 0$ ist dabei ein fester Parameter.

In verschiedenen Anwendungen, etwa bei der Modellierung der Verteilung der Anzahl der Passagiere einer Fähre, die nur auf Anforderung fährt, ist die gestutzte Verteilung von Bedeutung:

$$\mathrm{P}(X=x|X>0) = \frac{e^{-\lambda}}{1-e^{-\lambda}}\cdot\frac{\lambda^x}{x!} \qquad x = 1,2,\ldots$$

Von besonderem Interesse sind bedingte Verteilungen mehrdimensionaler Zufallsvariablen $\boldsymbol{X} = (X_1,\ldots,X_k)$, wenn A die folgende Gestalt hat:

$$A = (-\infty,\infty)\times(-\infty,\infty)\times\ldots\times(-\infty,\infty)\times\{x_{i+1}\}\times\ldots\times\{x_k\}.$$

Diese Bedingung besagt, dass die ersten i Komponenten in Abhängigkeit davon, dass die letzten $k-i$ jeweils einen festen Wert annehmen, betrachtet werden. Natürlich ist die Beschränkung auf die ersten bzw. letzten Komponenten willkürlich und geschieht hier nur zur Vereinfachung der Schreibweise. Aus dem gleichen Grund konzentrieren wir uns nun auf zweidimensionale Zufallsvariablen.

Sei zunächst (X,Y) diskret und sei $A=(-\infty,\infty)\times\{y\}$ mit $\mathrm{P}(Y=y)>0$. Dann ist

$$F(x,y|A)=\mathrm{P}(X\le x, Y\le y\,|X<\infty, Y=y)=\frac{\mathrm{P}(X\le x, Y=y)}{\mathrm{P}(Y=y)}.$$

Dafür schreiben wir:

$$F_{X|Y}(x|y):=F(x,y\,|A)=\frac{\mathrm{P}(X\le x, Y=y)}{\mathrm{P}(Y=y)}.$$

$F_{X|Y}(x|y)$ oder einfach $F(x|y)$ ist eine Verteilungsfunktion, die Verteilungsfunktion von X bei gegebenem $\{Y=y\}$.

Im Fall $\mathrm{P}(Y=y)=0$ kann $F_{X|Y}(x|y)$ nicht wie oben definiert werden. Die Definition kann aber über einen Grenzwert erfolgen:

$$F(x|y)=\lim_{h\to 0}\frac{\mathrm{P}(X\le x, y\le Y\le y+h)}{\mathrm{P}(y\le Y\le y+h)}.$$

Die bedingte Zähldichte hat naheliegender Weise die Gestalt

$$f(x|y)=\mathrm{P}(X=x|Y=y)=\frac{\mathrm{P}(X\le x, Y\le y)}{\mathrm{P}(Y\le y)}=\frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Die Gestalt der bedingten stetigen Dichtefunktion erhalten wir aus der Definition von $F_{X|Y}(x|y)$ folgendermaßen:

$$\begin{aligned}\frac{\partial}{\partial x}F_{X|Y}(x|y)&=\lim_{dx\to 0}\frac{1}{dx}[F_{X|Y}(x+dx|y)-F_{X|Y}(x|y)]\\&=\lim_{dx,dy\to 0}\left\{\frac{[F_{X,Y}(x+dx,y+dy)-F_{X,Y}(x+dx,y)]/(dx\,dy)}{[f_Y(y+dy)-F_Y(y)]/dy}\right.\\&\qquad\left.-\frac{[F_{X,Y}(x,y+dy)-F_{X,Y}(x,y)]/(dx\,dy)}{[f_Y(y+dy)-F_Y(y)]/dy}\right\}\\&=\frac{f_{X,Y}(x,y)}{f_Y(y)}.\end{aligned}$$

Definition 1.2.35 *bedingte Verteilungsfunktion mehrdimensionaler Zufallsvariablen*

$\boldsymbol{X}=(X_1,\ldots,X_k)$ sei eine k-dimensionale Zufallsvariable mit der Verteilungsfunktion $F_{\boldsymbol{X}}(\boldsymbol{x})$. Die *bedingte Verteilungsfunktion* von $\boldsymbol{X}_1=(X_1,\ldots,X_i)$ bei gegebenen $\{\boldsymbol{X}_2=\boldsymbol{x}_2\}$, $\boldsymbol{X}_2=(X_{i+1},\ldots,X_k)$, $\boldsymbol{x}_2=(x_{i+1},\ldots,x_k)$ ist durch den Grenzwert

$$F(\boldsymbol{x}_1|\boldsymbol{x}_2)=F_{\boldsymbol{X}_1}(\boldsymbol{x}_1|\boldsymbol{x}_2)=\lim_{\boldsymbol{h}\to\boldsymbol{0}}\frac{\mathrm{P}(\boldsymbol{X}_1\le\boldsymbol{x}_1,\boldsymbol{x}_2\le\boldsymbol{X}_2\le\boldsymbol{x}_2+\boldsymbol{h})}{\mathrm{P}(\boldsymbol{x}_2\le\boldsymbol{X}_2\le\boldsymbol{x}_2+\boldsymbol{h})}\tag{1.27}$$

definiert, wenn dieser existiert. Dabei ist $\boldsymbol{h} = (h, \ldots, h)$ ein $(k-i)$-dimensionaler Vektor. Die zu einer bedingten Verteilungsfunktion gehörende Dichte

$$f(\boldsymbol{x}_1|\boldsymbol{x}_2) = f_{\boldsymbol{X}_1}(\boldsymbol{x}_1|\boldsymbol{x}_2) = \begin{cases} \dfrac{f_{\boldsymbol{X}}(\boldsymbol{x}_1,\boldsymbol{x}_2)}{f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)} & \text{für } f_{\boldsymbol{X}_2}(\boldsymbol{x}_2) > 0 \\ 0 & \text{sonst.} \end{cases} \tag{1.28}$$

wird als *bedingte Dichte(funktion)* von $\boldsymbol{X}_1$, gegeben $\{\boldsymbol{X}_2 = \boldsymbol{x}_2\}$, bezeichnet.

Beispiel 1.2.36 *Trinomialverteilung - Fortsetzung von Seite 31*

Für trinomialverteilte Zufallsvariablen (X, Y) erhalten wir:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{\dfrac{n!}{x!y!(n-x-y)!}\, p_1^x p_2^y (1-p_1-p_2)^{n-x-y}}{\dfrac{n!}{y!(n-y)!}\, p_2^y (1-p_2)^{n-y}}$$

$$= \frac{(n-y)!}{x!(n-x-y)!} \left(\frac{p_1}{1-p_2}\right)^x \left(1 - \frac{p_1}{1-p_2}\right)^{n-x-y}.$$

Die bedingte Verteilung von X, gegeben $\{Y = y\}$, ist also eine Binomialverteilung mit den Parametern $p_1/(1-p_2)$ und $(n-y)$.

Beispiel 1.2.37 *Kinder in einer Familie*

Die Anzahl X der Kinder pro Familie sei Poisson-verteilt:

$$f(x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda} \qquad x = 0, 1, 2, \ldots .$$

Die Anzahl Y der Mädchen in einer Familie mit x Kindern lässt sich als binomialverteilte Zufallsvariable auffassen. Genauer ist dies eine bedingte Verteilung:

$$f_{Y|X}(y|x) = \mathrm{P}(Y = y|X = x) = \binom{x}{y} p^y (1-p)^{x-y}.$$

Die Randverteilung von Y erhalten wir, indem wir zuerst die gemeinsame Verteilung bestimmen:

$$\begin{aligned} f_{X,Y}(x,y) &= f_{Y|X}(y|x) f_X(x) = \mathrm{P}(Y = y|X = x) \cdot \mathrm{P}(X = x) \\ &= \frac{\lambda^x}{x!} \cdot e^{-\lambda} \cdot \binom{x}{y} \cdot p^y (1-p)^{x-y} \qquad \text{für } 0 \le y \le x. \end{aligned}$$

Dann ist

$$f_Y(y) = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \cdot e^{-\lambda} \cdot \binom{x}{y} p^y (1-p)^{x-y}$$

$$= \sum_{x-y=0}^{\infty} e^{-\lambda} \cdot \frac{(\lambda(1-p))^{x-y}}{(x-y)!} \cdot \frac{\lambda^y}{y!} p^y = e^{-\lambda p} \cdot \frac{(\lambda p)^y}{y!} \sum_{z=0}^{\infty} e^{-\lambda(1-p)} \cdot \frac{(\lambda(1-p))^z}{z!}$$

$$= e^{-\lambda p} \cdot \frac{(\lambda p)^y}{y!}.$$

Y ist also wieder Poisson-verteilt mit dem Parameter λp. Dieser gibt die erwartete Anzahl der Mädchen an: Von der mittleren Zahl λ der Kinder pro Familie ist das der p-te Anteil.

Beispiel 1.2.38 *Gleichverteilung mit dreieckförmigen Träger - Fortsetzung von Seite 31*

Aus der gemeinsamen Dichtefunktion und den beiden Randdichten erhalten wir

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{1-x} & \text{für } x \le y \le 1 \\ 0 & \text{sonst} \end{cases}$$

und

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{y} & \text{für } 0 \le x \le y \\ 0 & \text{sonst} \end{cases}.$$

Beispiel 1.2.39 *bivariate Normalverteilung - Fortsetzung von Seite 32*

Für die bivariate Normalverteilung erhalten wir mit den Ergebnissen des Beispiels 1.2.32:

$$f(x,y) = g(x) \cdot h(x,y)$$

mit

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu_X)^2}{\sigma_X^2}\right)$$

und

$$h(x,y) = \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2\sigma_Y(1-\rho^2)} \cdot \left(y - \mu_Y - \frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right)^2\right].$$

Dabei ist $g(x)$ die Randdichte von X. Folglich ist $f_{Y|X}(y|x) = h(x,y)$. Das ist die oben angekündigte Interpretation dieser Funktion.
Also ist Y für gegebenes $X = x$ univariat normalverteilt mit den Parametern $\mu_Y - \sigma_Y(x-\mu_X)/\sigma_X$ und $\sigma_Y^2(1-\rho^2)$.

Wie das Konzept der bedingten Wahrscheinlichkeit lässt sich auch das der Unabhängigkeit auf Zufallsvariablen übertragen. Von der Unabhängigkeit zweier Zufallsvariablen (X, Y) werden wir naheliegender Weise verlangen: Ist A ein Ereignis, das durch eine der beiden Zufallsvariablen bestimmt ist, und ist B durch die andere Zufallsvariable festgelegt, so sind A und B unabhängig.

Speziell sollen die Ereignisse $A=(X\leq x)$ und $B=(Y\leq y)$ unabhängig sein. Mit Verteilungsfunktionen geschrieben heißt das:

$$F_{X,Y}(x,y)=F_X(x)\cdot F_Y(y).$$

Ist diese Gleichung für alle $(x,y)\in\mathbb{R}^2$ erfüllt, so ist es auch die zuvor betrachtete umfassendere Forderung.

Definition 1.2.40 *stochastische Unabhängigkeit von Zufallsvariablen*

Die Zufallsvariablen $X_1,\ldots,X_k$ mit der gemeinsamen Verteilungsfunktion $F(x_1,\ldots,x_k)$ heißen *stochastisch unabhängig*, wenn für alle $(x_1,\ldots,x_k)\in\mathbb{R}^k$ gilt:

$$F(x_1,\ldots,x_k)=F_1(x_1)\cdot\ldots\cdot F_k(x_k). \tag{1.29}$$

Dabei sind die F_i die Randverteilungsfunktionen der $X_i\quad(i=1,\ldots,k)$.

Satz 1.2.41 *Charakterisierung der Unabhängigkeit von Zufallsvariablen*

Die Zufallsvariablen $X_1,\ldots,X_k$ mit der gemeinsamen Dichtefunktion $f(x_1,\ldots,x_k)$ sind genau dann stochastisch unabhängig, wenn für alle $(x_1,\ldots,x_k)\in\mathbb{R}^k$ gilt:

$$f(x_1,\ldots,x_k)=f_1(x_1)\cdot\ldots\cdot f_k(x_k). \tag{1.30}$$

Beweis: Im Fall $k=2$ folgt für stetige Zufallsvariablen aus $f(x_1,x_2)=f_1(x_1)\cdot f_2(x_2)$ sofort $F(x_1,x_2)=F_1(x_1)\cdot F_2(x_2)$. Wegen

$$f(x,y)=\frac{\partial^2}{\partial x\partial y}F(x,y)=\frac{\partial^2}{\partial x\partial y}F_1(x)\cdot F_2(y)=f_1(x)\cdot f_2(y)$$

gilt auch die Umkehrung.
Für diskrete Zufallsvariablen ist der Nachweis elementar.

Beispiel 1.2.42 *zweidimensionale logistische Verteilung - Fortsetzung von Seite 30*

Die Verteilungsfunktion zweier Zufallsvariablen X, Y sei:

$$F(x,y)=\frac{1}{1+e^{-x}+e^{-y}+e^{-(x+y)}}.$$

Die Randverteilungen von X und Y sind gegeben durch:

$$F_X(x)=\frac{1}{1+e^{-x}},\quad F_Y(y)=\frac{1}{1+e^{-y}}.$$

Damit gilt:

$$F_{X,Y}(x,y)=F_X(x)\cdot F_Y(y).$$

X und Y sind also stochastisch unabhängig.

Beispiel 1.2.43 *bivariate Normalverteilung - Fortsetzung von Seite 36*

Für die bivariate Normalverteilung gilt, wie man sich anhand der Gestalt der Dichte sofort klar macht:

$$f(x,y) = f_X(x) \cdot f_Y(y) \quad \Leftrightarrow \quad \rho = 0.$$

Der Parameter ρ ‚steuert' also die Abhängigkeit bei der bivariaten Normalverteilung.

1.3 Transformationen von Zufallsvariablen

In vielen statistischen Überlegungen spielen Transformationen von Zufallsvariablen eine Rolle. So geht man häufig von den beobachteten zu den logarithmierten Werten über. Bei Zufallsvariablen werden durch solche Abbildungen wieder Zufallsvariablen definiert:

$$Z = g(X) \qquad \text{oder allgemeiner} \quad Z = g(X_1, \ldots, X_k).$$

Damit Z tatsächlich eine Zufallsvariable ist, muss die Funktion g geeignete Voraussetzungen erfüllen, z.B. stetig sein. In unserem Rahmen bedeutet es aber keine Einschränkung, einfach zu unterstellen, dass Z eine wohldefinierte Zufallsvariable ist. Von Interesse ist dann die Frage, wie die Verteilung der transformierten Variablen Z aus der ursprünglichen Verteilung bestimmt werden kann. Für diskrete Variablen ist das oft einfach durch Betrachtung der Wahrscheinlichkeitsfunktionen möglich.

Beispiel 1.3.1 *Trinomialverteilung - Fortsetzung von Seite 35*

(X,Y) sei eine zweidimensionale diskrete Zufallsvariable mit der Zähldichte $f(x,y)$. Gesucht werde die Verteilung von $Z = X + Y$. Es ist

$$\mathrm{P}(Z = z) = \mathrm{P}(X + Y = z) = \sum_{\substack{x,y \\ x+y=z}} f(x,y) = \sum_{y} f(z-y,y) = \sum_{x} f(x,z-x).$$

Dabei werden die Summen über alle erlaubten Realisationsmöglichkeiten gebildet. Sind (X,Y) speziell trinomialverteilt, so erhalten wir für $Z = X + Y$:

$$\begin{aligned}
f_Z(z) &= \mathrm{P}(Z = z) = \sum_{y=0}^{z} \mathrm{P}(X = z-y, Y = y) \\
&= \sum_{y=0}^{z} \frac{n!}{(z-y)!y!(n-z)!} p_1^{z-y} p_2^{y} (1-p_1-p_2)^{n-z} \\
&= \frac{n!}{z!(n-z)!} (p_1+p_2)^z (1-p_1-p_2)^{n-z} \cdot \sum_{y=0}^{z} \frac{z!}{y!(z-y)!} \left(\frac{p_2}{p_1+p_2}\right)^y \left(1 - \frac{p_1}{p_1+p_2}\right)^{z-y} \\
&= \binom{n}{z} (p_1+p_2)^z (1-p_1-p_2)^{n-z}.
\end{aligned}$$

Damit ist Z binomialverteilt mit dem Parameter $(p_1 + p_2)$.

Für stetige Zufallsvariablen wird zuerst eine einfache, aber für die Anwendung sehr wichtige Transformation betrachtet.

Satz 1.3.2 *Wahrscheinlichkeitsintegraltransformation*

Sei X eine Zufallsvariable mit einer stetigen Verteilungsfunktion $F(x)$, die für $0 < F(x) < 1$ streng monoton wachsend ist. Dann ist die Zufallsvariable $U = F(X)$ gleichverteilt über dem Intervall $[0,1] : U \sim \mathscr{R}(0,1)$.

Umgekehrt hat $X = F^{-1}(U)$ mit $F^{-1}(u) = \inf\{x | F(x) \geq u\}$ die Verteilungsfunktion $F(x)$, falls U über dem Intervall $(0,1)$ gleichverteilt ist.

Beweis: Wegen $0 \leq F(x) \leq 1$ gilt für $u < 0$: $\mathrm{P}(U \leq u) = 0$ und für $u \geq 1$: $\mathrm{P}(U \leq u) = 1$. Sei nun $0 \leq u < 1$. Da F stetig ist und $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$, gibt es zu u ein $x_0 \in \mathbb{R}$ mit

$$F(x_0) = u \quad \text{und} \quad F(x) > u \quad \text{für } x > x_0 .$$

Daher sind die Ereignisse $\{U \leq u\}$ und $\{X \leq x_0\}$ äquivalent. Somit haben wir:

$$\mathrm{P}(U \leq u) = \mathrm{P}(X) \leq x_0) = F(x_0)) = u .$$

Damit ist der erste Teil gezeigt. Der zweite folgt analog.

Satz 1.3.2 zeigt, wie wir eine (Pseudo-)Zufallszahl x aus einer Verteilung mit Verteilungsfunktion $F_X(x)$ erzeugen können, wenn uns ein (Pseudo-)Zufallszahlengenerator für gleichverteilte Zufallszahlen zur Verfügung steht:

1. Erzeuge eine $(0,1)$ gleichverteilte Zufallszahl u.
2. Bilde $x = F_X^{-1}(u)$.

Voraussetzung für dieses Vorgehen ist jedoch, dass $F_X^{-1}(x)$ in geschlossener Form vorliegt.

Beispiel 1.3.3 *logistische Verteilung*

Die Verteilungsfunktion der logistischen Verteilung lautet

$$F(x) = \frac{1}{1 + e^{-x}} .$$

Die Inverse der Verteilungsfunktion ergibt sich einfach:

$$u = \frac{1}{1 + e^{-x}} \quad \Rightarrow \quad x = \ln\left(\frac{u}{1-u}\right) = F^{-1}(u) .$$

Also erhalten wir eine Zufallszahl x aus einer logistischen Verteilung, indem wir eine gleichverteilte Zufallszahl u erzeugen und $x = \ln(u) - \ln(1-u)$ bilden.

Um die Struktur offenzulegen, wie die Dichte einer transformierten Zufallsvariablen mit der der Ausgangsvariablen zusammenhängt, betrachten wir das folgende Beispiel.

Beispiel 1.3.4 *logistischen Verteilung - Fortsetzung von Seite 39*

Wir setzen im Beispiel 1.3.3 $g(u)=\ln(u)-\ln(1-u)$. Dann wurde oben die gleichverteilte Zufallsvariable U mittels $g(u)$ transformiert. Das ergibt:

$$F_X(x)=\mathrm{P}(X\le x)=\mathrm{P}\left(\ln\left(\frac{U}{1-U}\right)\le x\right)=\mathrm{P}\left(U\le\frac{1}{1+e^{-x}}\right)=F_U(g^{-1}(x)).$$

Die Dichte von X erhalten wir durch Differenzieren:

$$f_X(x)=\frac{\partial}{\partial x}F_X(x)=\frac{\partial}{\partial x}F_U(g^{-1}(x))=f_U(g^{-1}(x))\cdot\frac{\partial}{\partial x}g^{-1}(x)=1\cdot\frac{e^{-x}}{(1+e^{-x})^2}.$$

Satz 1.3.5 *Transformationssatz für eindimensionale Variablen*

X sei eine Zufallsvariable mit der Verteilungsfunktion $F_X(x)$ und der Dichte $f(x)$. $g:\mathbb{R}\to\mathbb{R}$ sei eine streng monotone und differenzierbare Abbildung. Dann gilt für die Dichte der Zufallsvariablen $Y=g(X)$:

$$f_Y(y)=\begin{cases} f_X(g^{-1}(y))\cdot\left|\dfrac{\partial}{\partial y}g^{-1}(y)\right| & \text{für alle } y \text{ mit} f_X(g^{-1}(y))>0 \\ 0 & \text{sonst} \end{cases}. \tag{1.31}$$

Beweis: Der Beweis orientiert sich an den Ausführungen des letzten Beispiels. Sei g monoton wachsend. Dann gilt:

$$F_Y(y)=\mathrm{P}(Y\le y)=\mathrm{P}(g(X)\le y)=\mathrm{P}(X\le g^{-1}(y))=F_X(g^{-1}(y)).$$

Also gilt für alle y mit $f_X(g^{-1}(y))>0$:

$$f_Y(y)=\frac{\partial}{\partial y}F_Y(y)=f_X(g^{-1}(y))\cdot\frac{\partial}{\partial y}g^{-1}(y).$$

und ansonsten ist die Dichtefunktion 0. Ist g monoton fallend, so ist:

$$F_Y(y)=\mathrm{P}(Y\le y)=\mathrm{P}(g(X)\le y)=\mathrm{P}(X\ge g^{-1}(y))=1-\mathrm{P}(X\ge g^{-1}(y))=1-F_X(g^{-1}(y)).$$

Folglich gilt für alle y mit $f_X(g^{-1}(y))>0$:

$$f_Y(y)=\frac{\partial}{\partial y}F_Y(y)=-f_X(g^{-1}(y))\cdot\frac{\partial}{\partial y}g^{-1}(y).$$

Sonst ist die Dichtefunktion 0. Zusammen ergibt das die Behauptung.

Beispiel 1.3.6 *logarithmische Normalverteilung*

Es sei $X\sim\mathcal{N}(\mu,\sigma^2)$. Gesucht ist die Dichtefunktion von $Y=e^X$. Mit $x=g^{-1}(y)=\ln(y)$ und $\partial g^{-1}(y)/\partial y=1/y$ erhalten wir:

$$f_Y(y)=\begin{cases} \dfrac{1}{\sqrt{2\pi}\cdot\sigma\cdot y}\cdot\exp\left[-\dfrac{(\ln(y)-\mu)^2}{2\cdot\sigma^2}\right] & \text{für } y>0 \\ 0 & \text{sonst} \end{cases}. \tag{1.32}$$

Y wird als *logarithmisch normalverteilt* oder kurz als *lognormalverteilt* bezeichnet.

Beispiel 1.3.7 *Weibull-Verteilung*

Sei X exponentialverteilt mit Parameter α. Gesucht ist die Verteilung von $Y = X^{1/\beta}$.
Es gilt $x = g^{-1}(y) = y^{\beta}$ und $\partial\, g^{-1}(y)/\partial y = \beta \cdot y^{\beta-1}$ und damit:

$$f_X(g^{-1}(y)) \cdot \frac{\partial}{\partial y} g^{-1}(y) = \alpha \cdot \beta \cdot y^{\beta-1} \cdot e^{-\alpha \cdot y^{\beta}}.$$

Also ist

$$f_Y(y) = \begin{cases} \alpha \cdot \beta \cdot y^{\beta-1} \cdot e^{-\alpha \cdot y^{\beta}} & \text{für } y \geq 0 \\ 0 & \text{für } y < 0 \end{cases}. \tag{1.33}$$

Man spricht in diesem Fall von der *Weibull-Verteilung*. Diese spielt als Modell einer Lebensdauerverteilung im Rahmen der Zuverlässigkeitstheorie eine zentrale Rolle.

Beispiel 1.3.8 *Chiquadratverteilung*

Sei X standardnormalverteilt. Gesucht ist die Verteilung von $Y = X^2$. Satz 1.3.5 ist nicht anwendbar, da die Transformation nicht monoton ist. Wir können aber die Intervalle, auf denen die Transformation jeweils monoton ist, getrennt betrachten und von den bedingten Dichten ausgehen. Es gilt

$$f_{X|X\geq 0}(x|X \geq 0) = \begin{cases} \frac{2}{\sqrt{2 \cdot \pi}} \cdot e^{-0.5 \cdot x^2} & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases},$$

$$f_{X|X<0}(x|X < 0) = \begin{cases} \frac{2}{\sqrt{2 \cdot \pi}} \cdot e^{-0.5 \cdot x^2} & \text{für } x < 0 \\ 0 & \text{für } x \geq 0 \end{cases}.$$

Das ergibt, wenn $g^{-1}(y) = \sqrt{y}$ bzw. $= -\sqrt{y}$ gesetzt wird:

$$\begin{aligned} f_Y(y) &= f_{Y|X\leq 0}(y|X \leq 0) \cdot \mathrm{P}(X \leq 0) + f_{Y|X>0}(y|X > 0) \cdot \mathrm{P}(X > 0) \\ &= 2\frac{2}{\sqrt{2 \cdot \pi}\sqrt{y}} \cdot e^{-0.5 \cdot y} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{\sqrt{2 \cdot \pi}\sqrt{y}} \cdot e^{-0.5 \cdot y}. \end{aligned} \tag{1.34}$$

Man nennt die zugehörige Verteilung eine *Chiquadratverteilung* mit einem Freiheitsgrad.

Im mehrdimensionalen Fall ist der Zusammenhang ähnlich wie im Satz 1.3.5 für den eindimensionalen formuliert. Jedoch sind dann die partiellen Ableitungen zu betrachten.

Satz 1.3.9 *Transformationssatz für mehrdimensionale Variablen*

Sei $\boldsymbol{X} = (X_1, \ldots, X_k)$ eine stetige Zufallsvariable mit der Dichte $f_{\boldsymbol{X}}(\boldsymbol{x})$. Seien

$$y_1 = g_1(\boldsymbol{x}), \ldots, y_k = g_k(\boldsymbol{x})$$

Abbildungen, für die die inversen Transformationen

$$x_1 = h_1(\boldsymbol{y}), \ldots, x_k = h_k(\boldsymbol{y}),$$

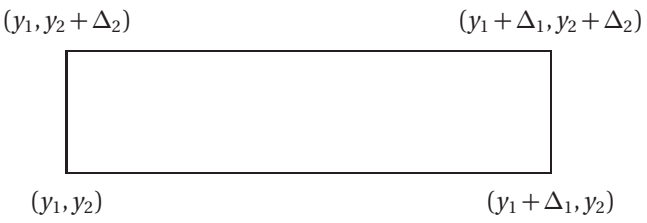
$\boldsymbol{y} = (y_1, \ldots, y_k)$, existieren. Die Abbildungen g_i, h_i, $(i = 1, \ldots, k)$, seien stetig und die partiellen Ableitungen $\partial h_i / \partial y_j$ $(i, j = 1, \ldots, k)$ mögen existieren und stetig sein.
Sofern die Jacobi-Determinante der inversen Transformation

$$\boldsymbol{J} = \det \begin{pmatrix} \dfrac{\partial h_1(\boldsymbol{y})}{\partial y_1} & \cdots & \dfrac{\partial h_1(\boldsymbol{y})}{\partial y_k} \\ \vdots & & \vdots \\ \dfrac{\partial h_k(\boldsymbol{y})}{\partial y_1} & \cdots & \dfrac{\partial h_k(\boldsymbol{y})}{\partial y_k} \end{pmatrix}$$

von Null verschieden ist, gilt für die Zufallsvariable $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(h_1(\boldsymbol{y}), \ldots, h_k(\boldsymbol{y})) \cdot |\boldsymbol{J}|. \tag{1.35}$$

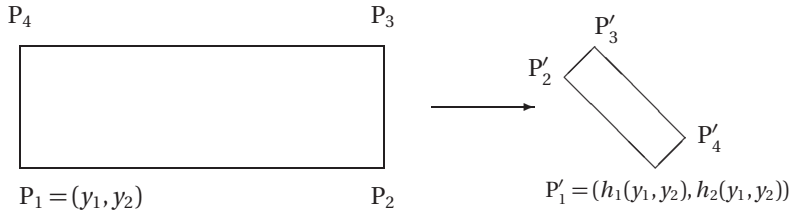
Beweisillustration: Wir betrachten hier nur den zweidimensionalen Fall. Das Auftauchen der Jacobi-Determinante lässt sich folgendermaßen veranschaulichen:
Die Wahrscheinlichkeit, dass (Y_1, Y_2) Werte aus dem Rechteck



annimmt, beträgt für kleine Δ_1, Δ_2 ungefähr

$$f_{Y_1, Y_2}(y_1, y_2) \cdot \Delta_1 \cdot \Delta_2.$$

Durch die inverse Transformation wird das Rechteck (näherungsweise) in ein Parallelogramm überführt:



Dessen Flächeninhalt hängt nun von der Kantenlänge und der Transformation $(h_1(y_1, y_2), h_2(y_1, y_2))$ ab und beträgt

$$\left| \det \begin{pmatrix} \frac{\partial h_1(y_1,y_2)}{\partial y_1} & \frac{\partial h_1(y_1,y_2)}{\partial y_2} \\ \frac{\partial h_2(y_1,y_2)}{\partial y_1} & \frac{\partial h_2(y_1,y_2)}{\partial y_2} \end{pmatrix} \right| \cdot \Delta_1 \cdot \Delta_2 = |\boldsymbol{J}| \cdot \Delta_1 \cdot \Delta_2.$$

Somit gilt, da sich bei den eineindeutigen Transformationen die Wahrscheinlichkeiten nicht ändern:

$$f_{Y_1,Y_2}(y_1,y_2) \cdot \Delta_1 \cdot \Delta_2 = f_{X_1,X_2}(h_1(y_1,y_2),\ h_2(y_1,y_2)) \cdot |\boldsymbol{J}| \cdot \Delta_1 \cdot \Delta_2.$$

Kürzen von $\Delta_1 \cdot \Delta_2$ gibt die behauptete Relation für die Dichten.

Beispiel 1.3.10 *Summe zweier exponentialverteilter Zufallsvariablen*

X, Y seien zwei unabhängige, identisch exponentialverteilte Zufallsvariablen, d. h. (X, Y) möge folgende Dichte haben :

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & \text{für } x,y > 0 \\ 0 & \text{sonst.} \end{cases}$$

Um die Verteilung der Summe von X und Y zu erhalten, betrachten wir die Transformation

$$u = x + y = g_1(x,y)$$
$$v = x = g_2(x,y).$$

Die inverse Transformation ist

$$x = u - v = h_1(u,v)$$
$$y = v = h_2(u,v).$$

Damit gilt:

$$\boldsymbol{J} = \det \begin{bmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v)}{\partial v} \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{bmatrix} = \det \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = 1$$

Es folgt :

$$f_{U,V}(u,v) = f_{X,Y}(u-v,v) = \begin{cases} \lambda^2 e^{-\lambda((u-v)+v)} & \text{für } u > v > 0 \\ \lambda^2 e^{-\lambda u} & \text{für } u > v > 0. \end{cases}$$

Für andere Werte (u,v) ist die Dichte von (U,V) null.

Daraus lässt sich nun die Verteilung der Summe $X+Y$ als Randverteilung von U bestimmen:

$$f_{X+Y}(u) = f_U(u) = \int_{-\infty}^{+\infty} f_{U,V}(u,v)\,\mathrm{d}v = \int_0^u \lambda^2 e^{-\lambda u}\,\mathrm{d}v = \lambda^2 \cdot u \cdot e^{-\lambda u}.$$

Dies ist die Dichte einer *Erlang-Verteilung*.

Das im Beispiel erhaltene Zwischenresultat

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_{X,Y}(u-v,v)\,\mathrm{d}v \tag{1.36}$$

ist allgemein gültig. Das Integral wird als *Faltungsintegral* bezeichnet.

Das Beispiel zeigt, wie wir mit Hilfe einer mehrdimensionalen Transformation einfach die Verteilung einer eindimensionalen Transformation bestimmen können. Wir setzen die interessierende Transformation gleich der ersten Komponente und wählen für die zweite (und weitere) eine Projektion oder eine andere geeignete Transformation. Die Randverteilung der ersten Komponente gibt dann die gewünschte Verteilung.

Beispiel 1.3.11 *Quotient zweier standardnormalverteilter Zufallsvariablen*

Seien X und Y unabhängige, identisch standardnormalverteilte Zufallsvariablen. Gesucht ist die Verteilung von $U = g_1(X,Y) = X/Y$. Wir setzen $V = g_2(X,Y) = Y$. Die inverse Transformation ist

$$\begin{aligned} x &= h_1(u,v) = u \cdot v \\ y &= h_2(u,v) = v\,. \end{aligned}$$

Damit gilt:

$$|\boldsymbol{J}| = \left| \det \begin{bmatrix} \dfrac{\partial h_1(u,v)}{\partial u} & \dfrac{\partial h_1(u,v)}{\partial v} \\ \dfrac{\partial h_2(u,v)}{\partial u} & \dfrac{\partial h_2(u,v)}{\partial v} \end{bmatrix} \right| = \left| \det \begin{bmatrix} v & u \\ 0 & 1 \end{bmatrix} \right| = |v|.$$

Somit haben wir:

$$f_{U,V}(u,v) = |\boldsymbol{J}| \cdot f_{X,Y}(u \cdot v, v) = \frac{1}{2 \cdot \pi} \cdot |v| \cdot \exp(-0.5 \cdot u^2 \cdot v^2 - 0.5 \cdot v^2).$$

Für die Randdichte von U erhalten wir

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} \frac{1}{2 \cdot \pi} \cdot |v| \cdot \exp(-0.5 \cdot u^2 \cdot v^2 - 0.5 \cdot v^2)\,\mathrm{d}v \\ &= \frac{1}{\pi} \int_{0}^{\infty} v \cdot \exp(-0.5 \cdot (u^2+1) \cdot v^2)\,\mathrm{d}v = \frac{1}{\pi} \cdot \left[-\frac{1}{1+u^2} \cdot \exp(-0.5 \cdot (u^2+1) \cdot v^2) \right]_0^{\infty} \\ &= \frac{1}{\pi \cdot (1+u^2)} \end{aligned}$$

Dies ist die Dichtefunktion einer *Cauchy-Verteilung*.

2 Momente von Verteilungen

2.1 Erwartungswerte

2.1.1 Grundlegende Definitionen und Eigenschaften

Im ersten Lehrbuch der Wahrscheinlichkeitsrechnung, in Christian Huygens ‚De ratiociniis in ludo aleae' aus dem Jahre 1657, wurde der Erwartungswert von Zufallsvariablen als grundlegender Begriff noch vor der Wahrscheinlichkeit selbst eingeführt. Da das Thema seines Buches das Würfelspiel war, ist Erwartungswert wörtlich zu nehmen: Eine zentrale Frage war die nach dem Betrag oder Wert, den ein Spieler mit einer bestimmten Spielstrategie als Gewinn erwarten durfte. Heutzutage wird der Erwartungswert üblicherweise als Lageparameter der Verteilung einer Zufallsvariablen eingeführt. Er ist das theoretische Analogon zum arithmetischen Mittel. Ist bei einem Zufallsexperiment der Stichprobenraum Ω endlich, $|\Omega| = N$, und P die Laplacesche Wahrscheinlichkeitsverteilung, $P(\{\omega\}) = 1/N$, so hat für jede Zufallsvariable X auf $(\Omega, \mathscr{A}, P)$ das Populationsmittel $\sum_{\omega\in\Omega} X(\omega)/N = \mu$ die zusätzliche Interpretation eines Lageparameters der Verteilung P. Das führt dazu, dass Erwartungswerte auch als Populationsmittel bezeichnet werden.

Definition 2.1.1 *Erwartungswert*

X sei eine eindimensionale Zufallsvariable mit der Verteilungsfunktion $F(x)$. Dann heißt

$$E(X) = \int_{-\infty}^{\infty} x \, dF(x) \tag{2.1}$$

der *Erwartungswert* von X oder der Verteilung von X. Vorausgesetzt wird dabei, dass das Riemann-Stieltjes-Integral absolut konvergiert:

$$\int_{-\infty}^{\infty} |x| \, dF(x) < \infty.$$

Andernfalls wird gesagt, dass der Erwartungswert nicht existiert.

Für den Erwartungswert $E(X)$ einer Zufallsvariablen X wird, wie schon erwähnt, häufig das Symbol μ verwendet: $\mu = \mu_X = E(X)$. Für die konkreten Berechnungen verwenden wir die folgenden Beziehungen: $E(X) = \sum_x x f(x)$ für diskrete und $E(X) = \int_{-\infty}^{\infty} x f(x) dx$ für stetige Verteilungen.

Beispiel 2.1.2 *Erwartungswert der Poisson-Verteilung*

X sei Poisson-verteilt,

$$f(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!} \quad \text{für} \quad x = 0, 1, 2, \ldots.$$

Den Erwartungswert von X erhalten wir leicht unter Ausnutzen der Eigenschaft einer Zähldichte, dass die Summe über alle Realisationsmöglichkeiten eins ergibt:

$$\mathrm{E}(X) = \sum_{x=0}^{\infty} x \cdot e^{-\lambda} \frac{\lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} e^{-\lambda} \frac{\lambda^{x-1}}{(x-1)!} = \lambda \sum_{y=0}^{\infty} e^{-\lambda} \frac{\lambda^y}{y!} = \lambda \cdot 1 = \lambda.$$

Beispiel 2.1.3 *Erwartungswert der Pareto-Verteilung*

Die *Pareto-Verteilung* hat die von zwei Konstanten $\alpha, k > 0$ abhängige Dichte

$$f(x) = \begin{cases} \alpha \cdot k^\alpha \cdot x^{-(\alpha+1)} & \text{für } x > k \\ 0 & \text{sonst} \end{cases}.$$

Für eine Pareto-verteilte Zufallsvariable X gilt:

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_{k}^{\infty} x \cdot \alpha \cdot k^\alpha \cdot x^{-(\alpha+1)}\, \mathrm{d}x = \int_{k}^{\infty} \alpha \cdot k^\alpha \cdot x^{-\alpha}\, \mathrm{d}x \\ &= \lim_{t\to\infty} \int_{k}^{t} \alpha \cdot k^\alpha \cdot x^{-\alpha}\, \mathrm{d}x = \lim_{t\to\infty} \left[k^\alpha \cdot \frac{\alpha}{1-\alpha} t^{1-\alpha} - k^\alpha \cdot \frac{\alpha}{1-\alpha} k^{1-\alpha} \right] \\ &= \lim_{t\to\infty} k^\alpha \cdot \frac{\alpha}{1-\alpha} [t^{1-\alpha} - k^{1-\alpha}]. \end{aligned}$$

Dieser Grenzwert existiert nur für $\alpha > 1$. Dann ist $\mathrm{E}(X) = k \cdot \alpha/(1-\alpha)$.

Es soll nun der Erwartungswert einer reellwertigen Funktion $g(X)$ einer Zufallsvariablen X bestimmt werden. Im Fall einer diskreten eindimensionalen Zufallsvariablen X ist auch $Y = g(X)$ eine diskrete Zufallsvariable, so dass formal gilt:

$$\mathrm{E}(Y) = \sum_{y} y \cdot f_Y(y) = \sum_{y} y \cdot \mathrm{P}(Y = y),$$

sofern der Erwartungswert existiert. Die Summe geht dabei über alle Realisationsmöglichkeiten von Y. Wir können $\mathrm{E}(g(X)) = \mathrm{E}(Y)$ aber auch erhalten, ohne die Verteilung von Y vorher zu ermitteln:

$$\mathrm{E}(Y) = \sum_{y} y \cdot \mathrm{P}(Y = y) = \sum_{y} y \cdot \mathrm{P}(g(X) = y) = \sum_{y} \left(y \cdot \sum_{\substack{x \\ g(x)=y}} \mathrm{P}(X = x) \right)$$

$$= \sum_y \left(\sum_{\substack{x \\ g(x)=y}} g(x) \cdot \mathrm{P}(X=x) \right) = \sum_x g(x) \cdot \mathrm{P}(X=x) = \sum_x g(x) \cdot f_X(x).$$

Diese Überlegung ist offensichtlich nicht auf eindimensionale Zufallsvariablen beschränkt, sondern auch richtig für Funktionen von mehrdimensionalen diskreten Zufallsvariablen. Stetige Variablen können analog behandelt werden. Zusammen erhalten wir

Satz 2.1.4 *Erwartungswert einer Funktion*

$\boldsymbol{X} = (X_1, \ldots, X_k)$ sei eine k-dimensionale Zufallsvariable mit der Verteilungsfunktion $F(\boldsymbol{x})$. $g : \mathbb{R}^k \to \mathbb{R}$ sei eine eindimensionale Abbildung. Dann gilt:

$$\mathrm{E}(g(\boldsymbol{X})) = \int_{\mathbb{R}^k} g(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x}), \tag{2.2}$$

sofern das Integral absolut konvergiert.

Beispiel 2.1.5 *Erwartungswert einer Linearkombination zweier stetiger Zufallsvariablen*

X und Y seien zwei stetige Zufallsvariablen mit der gemeinsamen Dichte $f(x,y)$. Falls $\mathrm{E}(X)$ und $\mathrm{E}(Y)$ existieren, gilt für $Z = aX + bY$:

$$\begin{aligned}
\mathrm{E}(aX+bY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax+by) f(x,y)\,\mathrm{d}x\mathrm{d}y \\
&= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f(x,y)\,\mathrm{d}x\mathrm{d}y + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \cdot f(x,y)\,\mathrm{d}x\mathrm{d}y \\
&= a\int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\mathrm{d}y + b\int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\mathrm{d}y \\
&= a\int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x + b\int_{-\infty}^{\infty} y \cdot f_Y(y)\,\mathrm{d}y = a \cdot \mathrm{E}(X) + b \cdot \mathrm{E}(Y).
\end{aligned}$$

Das Beispiel führt für einen Sonderfall aus, was allgemein gilt.

Satz 2.1.6 *Erwartungswert einer Linearkombination von Zufallsvariablen*

Der *Erwartungswert einer Linearkombination von Zufallsvariablen* ist gleich der Linearkombination der Erwartungswerte:

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i X_i \right) = \sum_{i=1}^{k} a_i \mathrm{E}(X_i). \tag{2.3}$$

Beweis: Analog zum Fall zweier Zufallsvariablen.

Für nichtlineare Funktionen gilt i. d. R. $\mathrm{E}(g(X)) \neq g(\mathrm{E}(X))$. Eine allgemeine Beziehung ergibt sich aber für *konvexe Funktionen* g. Eine auf dem Intervall $[a,b]$ definierte Funktion g ist dabei konvex, wenn

$$g(\lambda x+(1-\lambda)y) \leq \lambda g(x)+(1-\lambda)g(y) \quad \text{für} \quad x,y \in [a,b],\ 0 \leq \lambda \leq 1.$$

Bei einer auf $[a,b]$ definierten konvexen Funktion g, die zudem im offenen Intervall (a,b) differenzierbar ist, verlaufen alle Tangenten unterhalb von g. Sei nämlich $x \in (a,b)$ fest und $y \in (a,b)$ beliebig. Dann gilt für $0 < \lambda < 1$: $g(x+\lambda(y-x)) \leq g(x)+\lambda(g(y)-g(x))$. Dies ist äquivalent zu

$$\frac{g(x+\lambda(y-x))-g(x)}{\lambda} \leq g(y)-g(x).$$

Für $\lambda \to 0$ geht die linke Seite gegen $g'(x)(y-x)$ und es folgt wie behauptet:

$$g(y) \geq g(x)+g'(x)(y-x). \tag{2.4}$$

Satz 2.1.7 *Jensensche Ungleichung*

Sei X eine eindimensionale Zufallsvariable mit Werten in einem offenem Intervall (a,b). g sei eine konvexe Funktion auf (a,b). Es mögen $\mathrm{E}(X)$ und $\mathrm{E}(g(X))$ existieren. Dann gilt

$$g(\mathrm{E}(X)) \leq \mathrm{E}(g(X)). \tag{2.5}$$

Beweisskizze: Wir betrachten nur die Situation einer im Punkt $\mu = \mathrm{E}(X)$ differenzierbaren Funktion $g(x)$. Für die Tangente $l(x)$ an $g(x)$ im Punkt $\mathrm{E}(X) = \mu$ gilt dann: $l(x) = g(\mu)+g'(\mu)(x-\mu)$. Wegen der Konvexität gilt die in Gleichung (2.4) angegebene Beziehung. Da Ungleichungen bei der Erwartungswertbildung erhalten bleiben, folgt:

$$\mathrm{E}(g(X)) \geq \mathrm{E}(g(\mu)+g'(\mu)(X-\mu)) = g(\mu) = g(\mathrm{E}(X)).$$

Beispiel 2.1.8 *Quadrat und Kehrwert einer Zufallsvariablen*

Sei X eine Zufallsvariable mit $\mathrm{E}(X) = \mu$ und $0 < \mathrm{Var}(X) < \infty$. Dann ist $g(x) = x^2$ konvex. Damit ist $\mathrm{E}(X)^2 < \mathrm{E}(X^2)$. Auch $g(x) = 1/x$ ist für $x > 0$ konvex. Somit gilt für positive Zufallsvariablen $\mathrm{E}(1/X) \leq 1/\mathrm{E}(X)$, sofern die Erwartungswerte existieren.

Beispiel 2.1.9 *Ungleichung zwischen drei Mittelwerten*

Die Jensensche Ungleichung kann ausgenutzt werden, um eine Ungleichung zwischen drei bekannten Mittelwerten nachzuweisen. Seien $x_1,\ldots,x_n$ positive Zahlen und

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \bar{x}_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n} \quad , \quad \bar{x}_H = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1} .$$

(arithmetisches Mittel) (geometrisches Mittel) (harmonisches Mittel)

Dann gilt: $\bar{x}_A \geq \bar{x}_G \geq \bar{x}_H$.

Um die Ungleichung von Jensen anzuwenden, betrachten wir eine Zufallsvariable X, die diese Werte x_i jeweils mit der Wahrscheinlichkeit $1/n$ annimmt. Dann ist $\mathrm{E}(X) = \bar{x}_A$. Da $g(x) = -\ln(x)$ konvex ist, erhalten wir:

$$-\ln(\bar{x}_G) = -\frac{1}{n}\sum_{i=1}^{n}\ln(x_i) = \mathrm{E}(-\ln(X)) \geq -\ln(\mathrm{E}(X)) = -\ln(\bar{x}_A),$$

woraus sich $\bar{x}_G \leq \bar{x}_A$ ergibt. Weiter ist $g(x) = 1/x$ konvex für $x > 0$. Somit gilt:

$$\frac{1}{\bar{x}_H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} = \mathrm{E}\left(\frac{1}{X}\right) \geq \frac{1}{\mathrm{E}(X)} = \frac{1}{\bar{x}_A}.$$

Das führt zu der zweiten behaupteten Beziehung.

2.1.2 Formparameter von Verteilungen

Der Erwartungswert einer Verteilung ist ein Parameter, der die Lage der Verteilung charakterisiert. Das wird besonders durch sein Verhalten bei Translationen deutlich: $\mathrm{E}(X+a) = \mathrm{E}(X)+a$. Wir betrachen nun die Erwartungswerte einiger Funktionen, die ebenfalls gewisse Eigenschaften der zugrundeliegenden Verteilungen charakterisieren. Im Folgenden wird — soweit nichts anderes gesagt — unterstellt, dass alle vorkommenden Erwartungswerte existieren.

Definition 2.1.10 *Varianz*

X sei eine Zufallsvariable mit dem Erwartungswert $\mathrm{E}(X)$. Dann heißt der Erwartungswert von $g(X) = (X - \mathrm{E}(X))^2$ die *Varianz* von X,

$$\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2) \tag{2.6}$$

oder auch die *mittlere quadratische Abweichung* von X (vom Erwartungswert).

Für die Bestimmung der Varianz ist oft die mit Satz 2.1.6 folgende Relation vorteilhaft:

$$\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2) = \mathrm{E}(X^2 - 2X \cdot \mathrm{E}(X) + \mathrm{E}(X)^2) = \mathrm{E}(X^2) - \mathrm{E}(X)^2\,. \tag{2.7}$$

Beispiel 2.1.11 *Varianz der Poisson-Verteilung*

Ist X Poisson-verteilt, so ist

$$\begin{aligned}
\mathrm{E}(X^2) &= \sum_{x=0}^{\infty} x^2 \cdot e^{-\lambda} \cdot \frac{\lambda^x}{x!} = \sum_{x=1}^{\infty} x \cdot e^{-\lambda} \cdot \frac{\lambda^x}{(x-1)!} \\
&= \sum_{x=1}^{\infty} (x-1) \cdot e^{-\lambda} \cdot \frac{\lambda^x}{(x-1)!} + \sum_{x=1}^{\infty} e^{-\lambda} \cdot \frac{\lambda^x}{(x-1)!} \\
&= \lambda^2 \sum_{y=0}^{\infty} e^{-\lambda} \cdot \frac{\lambda^y}{y!} + \lambda \\
&= \lambda^2 + \lambda.
\end{aligned}$$

Damit folgt $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \lambda$.

Beispiel 2.1.12 *Varianz der Gleichverteilung*

Für eine über dem Intervall $(0,1)$ gleichverteilte Zufallsvariable X ist $\mathrm{E}(X) = 1/2$ und

$$\mathrm{Var}(X) = \int_0^1 \left(x - \frac{1}{2}\right)^2 \mathrm{d}x = \frac{1}{3}\left(x - \frac{1}{2}\right)^3 \Bigg|_0^1 = 2 \cdot \frac{1}{3} \cdot \frac{1}{8} = \frac{1}{12}.$$

Die Varianz ist eine Maßzahl für die Ausbreitung oder die ‚Streuung' einer Verteilung. Diese Interpretation wird durch die *Ungleichung von Tschebyschev*

$$\mathrm{P}(|X - \mathrm{E}(X)| > \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2} \tag{2.8a}$$

bzw.

$$\mathrm{P}(|X - \mathrm{E}(X)| \leq \epsilon) \leq 1 - \frac{\mathrm{Var}(X)}{\epsilon^2} \tag{2.8b}$$

fundiert. Je größer die Varianz ist, desto größer muss ϵ gewählt werden, damit die obere Schranke für die Wahrscheinlichkeit gleich groß bleibt, dass X einen Wert annimmt, der sich von $\mathrm{E}(X)$ um nicht mehr als ϵ unterscheidet.

Die Tschebyschev-Ungleichung ist ein Spezialfall der *Markoffschen Ungleichung*, die für $k > 0$ besagt:

$$\mathrm{P}(|X| > \epsilon) \leq \frac{\mathrm{E}(|X|^k)}{\epsilon^k}. \tag{2.9}$$

Der Beweis folgt aus der abermals weitergehenden, im folgenden Satz formulierten Ungleichung.

Satz 2.1.13

Sei X eine eindimensionale Zufallsvariable mit $\mathrm{P}(X > 0) = 1$. $g : \mathbb{R}_+ \to \mathbb{R}_+$ sei eine monotone Funktion für die der Erwartungswert $\mathrm{E}(g(X))$ existiert. Dann gilt für jedes $\epsilon > 0$:

$$\mathrm{P}(X > \epsilon) \leq \frac{\mathrm{E}(g(X))}{g(\epsilon)}. \tag{2.10}$$

Beweis: X habe die Verteilungsfunktion $F(x)$. Dann erhalten wir die Abschätzungen

$$\mathrm{E}(g(X)) = \int_0^\infty g(x)\, d\, F(x) \geq \int_\epsilon^\infty g(x)\, \mathrm{d}F(x) \geq g(\epsilon) \int_\epsilon^\infty \mathrm{d}F(x) = g(\epsilon) \cdot \mathrm{P}(X > \epsilon).$$

Damit ergibt sich die Behauptung.

Eine Zufallsvariable X heißt *entartet*, falls $\mathrm{P}(X = c) = 1$. Die Konstante c ist dann der Erwartungswert von X und die Varianz ist offensichtlich null. Aus der Ungleichung von Tschebyschev folgt aber auch unmittelbar, dass es keine stetige Zufallsvariable X mit $\mathrm{Var}(X) = 0$ geben kann und dass jede diskrete Zufallsvariable X mit $\mathrm{Var}(X) = 0$ entartet ist.

Die Tschebyschev-Ungleichung lässt sich nur verschärfen, wenn weitere Annahmen über die Verteilung gemacht werden. Dann gibt es jedoch zahlreiche weitere Ungleichungen. Wir führen nur eine an, die später gebraucht wird.

Lemma 2.1.14 *Ungleichung von Uspensky*

Ist X binomialverteilt, $X \sim \mathscr{B}(n,p)$, so gilt:

$$\mathrm{P}\left(\left|\frac{X}{n} - p\right| \geq c\right) < 2 \cdot \exp[-nc^2/2]. \tag{2.11}$$

Beweis: Uspensky (1937).

Speziell in der Formel für die Varianz der Summe zweier Zufallsvariablen kommt das gemischte Produkt der zentrierten Variablen vor:

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{E}(X+Y-\mu_{X+Y})^2 = \mathrm{E}((X-\mu_X)+(Y-\mu_Y))^2 \\ &= \mathrm{E}(X-\mu_X)^2 + 2\cdot\mathrm{E}((X-\mu_X)(Y-\mu_Y)) + \mathrm{E}(Y-\mu_Y)^2 \\ &= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\cdot\mathrm{E}((X-\mu_X)(Y-\mu_Y)).\end{aligned}$$

Definition 2.1.15 *Kovarianz und Korrelationskoeffizient*

Für zwei Zufallsvariablen (X,Y) ist die *Kovarianz* definiert durch

$$\mathrm{Cov}(X,Y) = \mathrm{E}((X-\mu_X)\cdot(Y-\mu_Y)). \tag{2.12}$$

Der *Korrelationskoeffizient* ρ_{XY} von X und Y ist

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}. \tag{2.13}$$

Zwei Zufallsvariablen X und Y heißen *unkorreliert*, wenn $\rho_{XY} = 0$ gilt. Für $\rho_{XY} < 0$ bzw. $\rho_{XY} > 0$ heißen sie *negativ* bzw. *positiv korreliert*. Aufgrund der Cauchy-Schwarz'schen Ungleichung ist der Korrelationskoeffizient ρ_{XY} von X und Y durch eins begrenzt:

$$-1 \leq \rho_{XY} \leq 1. \tag{2.14}$$

Satz 2.1.16 *Cauchy-Schwarz'sche Ungleichung*

Für zwei Zufallsvariablen X und Y, für die die Varianzen existieren, gilt

$$\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X)\cdot\mathrm{Var}(Y). \tag{2.15}$$

Beweis: Wir betrachten die Funktion

$$h(t) = \mathrm{E}((X-\mu_X)t + (Y-\mu_Y))^2.$$

Ausmultiplizieren ergibt

$$h(t) = t^2\mathrm{E}((X-\mu_X)^2) + 2t\,\mathrm{E}((X-\mu_X)(Y-\mu_Y)) + \mathrm{E}((Y-\mu_Y)^2) = t^2\sigma_X^2 + 2t\,\mathrm{Cov}(X,Y) + \sigma_Y^2.$$

Da $h(t)$ der Erwartungswert einer positiven Zufallsvariablen ist, gilt $h(t) \geq 0$ für alle t. Dies ist dann auch für $t = -\mathrm{Cov}(X,Y)/\sigma_X^2$ richtig. Die Umformung

$$\begin{aligned} h(t) &= t^2\sigma_X^2 + 2t\,\mathrm{Cov}(X,Y) + \frac{\mathrm{Cov}(X,Y)^2}{\sigma_X^2} - \frac{\mathrm{Cov}(X,Y)^2}{\sigma_X^2} + \sigma_Y^2 \\ &= \left(t\sigma_X + \frac{\mathrm{Cov}(X,Y)}{\sigma_X}\right)^2 + \sigma_Y^2 - \frac{\mathrm{Cov}(X,Y)^2}{\sigma_X^2} \end{aligned}$$

zeigt, dass damit gelten muss:

$$\sigma_Y^2 - \frac{\mathrm{Cov}(X,Y)^2}{\sigma_X^2} \geq 0 \quad \text{bzw.} \quad \mathrm{Cov}(X,Y)^2 \leq \sigma_X^2\sigma_Y^2 .$$

Der Korrelationskoeffizient charakterisiert den linearen Zusammenhang zweier Zufallsvariablen.

Satz 2.1.17 *Korrelation und linearer Zusammenhang*

X und Y seien zwei Zufallsvariablen mit $\mathrm{Var}(X) > 0$ und $\mathrm{Var}(Y) > 0$. Es ist $\rho = \pm 1$ genau dann, wenn $\mathrm{P}(Y = aX + b) = 1$ für geeignete Konstanten a, b.

Beweis: $|\rho(X,Y)| = 1$ gilt genau dann, wenn für die im Beweis der Cauchy-Schwarzschen Ungleichung betrachtete Funktion $h(t)$ gilt:

$$h(t) = \left(t\sigma_X + \frac{\mathrm{Cov}(X,Y)}{\sigma_X}\right)^2 .$$

Dann hat $h(t)$ die (eindeutig bestimmte) Nullstelle $t_0 = -\mathrm{Cov}(X,Y)/\sigma_X^2$. Damit ist $|\rho(X,Y)| = 1$ äquivalent zu

$$\mathrm{E}((X-\mu_X)t_0 + (Y-\mu-Y))^2 = 0;$$

dies ist wiederum äquivalent mit

$$\mathrm{P}((X-\mu_X)t_0 + (Y_\mu - Y) = 0) = 1.$$

Mit geeigneter Wahl von a und b ergibt das die Behauptung.

Aus der Unabhängigkeit folgt auch die Unkorreliertheit. Die Umkehrung dieser Aussage gilt i. A. jedoch nicht. Eine wichtige Ausnahme ist die bivariate Normalverteilung. Dies werden wir später sehen.

Sind die Zufallsvariablen X und Y unkorreliert, so gilt

$$\mathrm{E}(X \cdot Y) = \mathrm{E}(X) \cdot \mathrm{E}(Y). \tag{2.16}$$

Dies ist leicht zu sehen:

$$\begin{aligned} 0 &= \mathrm{E}((X-\mathrm{E}(X))(Y-\mathrm{E}(Y))) = \mathrm{E}(XY - \mathrm{E}(Y)X - \mathrm{E}(X)Y + \mathrm{E}(X)\mathrm{E}(Y)) \\ &= \mathrm{E}(XY) - \mathrm{E}(Y)\mathrm{E}(X) - \mathrm{E}(X)\mathrm{E}(Y) + \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(XY) - \mathrm{E}(Y)\mathrm{E}(X). \end{aligned}$$

Ein weiterer Form-Aspekt einer Verteilung ist seine Symmetrieeigenschaft. Eine Verteilung ist dabei *symmetrisch* um a, falls gilt:

$$\mathrm{P}(X < a - t) = \mathrm{P}(X > a + t).$$

Für die Verteilungsfunktion gilt dann $F(a-t) = 1-F(a+t)$, so dass bei stetigen Verteilungen die Dichte an den entsprechenden Stellen gleiche Werte hat:

$$f(a - t) = f(a + t).$$

Sofern der Erwartungswert μ einer symmetrischen Verteilung existiert, ist μ zugleich der Symmetriepunkt. Dann fällt auch der Median $\tilde{\mu}$,

$$\tilde{\mu} = F^{-1}(0.5) = \inf\{x | F(x) \geq 0.5\},$$

mit dem Erwartungswert zusammen.

Die folgende Charakterisierung des Erwartungswertes erlaubt weitergehende Aussagen zu der relativen Lage von μ und $\tilde{\mu}$.

Lemma 2.1.18 *relative Lage von μ*

Sei $F(x)$ eine stetige Verteilungsfunktion mit zugehöriger Dichte $f(x)$. $\mu = \mathrm{E}(X)$ möge existieren. Dann gilt für $m \in \mathbb{R}$:

$$m - \mu = \int_{-\infty}^{m} F(x) dx - \int_{m}^{\infty} (1 - F(x)) \mathrm{d}x. \tag{2.17}$$

Beweis: Da μ endlich ist, gilt $\lim_{x\to-\infty} xF(x) = 0$ und $\lim_{x\to\infty} x(1 - F(x)) = 0$. Partielle Integration führt dann von

$$m - \mu = \int_{-\infty}^{m} (m - x) f(x) dx + \int_{m}^{\infty} (m - x) f(x) \mathrm{d}x$$

unmittelbar zur Behauptung.

Die Beziehung (2.17) lässt sich weiter umformen zu

$$m - \mu = \int_{0}^{\infty} \{F(m - x) + F(m + x) - 1\} \mathrm{d}x.$$

Daraus ist sofort zu sehen, dass aus

$$F(\tilde{\mu} - x) \geq 1 - F(\tilde{\mu} + x) \tag{2.18}$$

folgt:

$$\mu \leq m.$$

(2.18) bedeutet in Worten, dass bei gleicher Entfernung vom Median die rechte Flanke niemals mehr Wahrscheinlichkeitsmasse hat als die linke. Die Verteilug ist rechtsschief. Zur weiteren Diskussion zu diesem Problemkreis sei auf van Zwet (1979) verwiesen.

Um nun die *Asymmetrie* einer Verteilung, den Grad der Abweichung von der Symmetrie, zu beschreiben, verwendet man den Erwartungswert $\mathrm{E}(X-\mu)^3$. Diese Wahl erscheint plausibel. Ist die Verteilung z. B. rechtsschief, 'geht die Dichte also auf der rechten Seite langsamer gegen Null als auf der linken', so wird $\mathrm{E}(X-\mu)^3$ einen positiven Wert annehmen, für linksschiefe Verteilungen einen negativen. Bei symmetrischen Verteilungen ist $\mathrm{E}(X-\mu)^3$ gleich Null. Die endgültige Maßzahl, die *Schiefe*, berücksichtigt noch die Streuung und ist definiert durch

$$\beta_1 = \frac{\mathrm{E}(X-\mu)^3}{\left(\mathrm{E}(X-\mu)^2\right)^{3/2}}. \tag{2.19}$$

Nützlich für die Bestimmung von $\mathrm{E}(X-\mu)^3$ ist häufig die Beziehung

$$\mathrm{E}(X-\mu)^3 = \mathrm{E}(X^3) - 3\mathrm{E}(X)\mathrm{E}(X^2) + 2\mathrm{E}(X)^3.$$

Um die *Wölbung* oder Steilheit einer unimodalen Dichte zu beschreiben, verwendet man

$$\beta_2 = \frac{\mathrm{E}(X-\mu)^4}{\left(\mathrm{E}(X-\mu)^2\right)^2}. \tag{2.20}$$

β_2 wird i. d. R. nur für symmetrische Verteilungen betrachtet.

Auch für $\mathrm{E}(X-\mu)^4$ gibt es eine analoge Berechnungsformel:

$$\mathrm{E}(X-\mu)^4 = \mathrm{E}(X^4) - 4\mathrm{E}(X^3)\mathrm{E}(X) + 6\mathrm{E}(X^2)\mathrm{E}(X)^2 - 3\mathrm{E}(X)^4.$$

Für die Normalverteilung ergibt sich mit der Symmetrie und den im Beispiel 2.2.8 angegebenen Formeln:

$$\beta_1 = 0, \quad \beta_2 = 3.$$

Hat eine Verteilung (ungefähr) diese Maßzahlen, so wird häufig unterstellt, dass sie in der Nähe der Normalverteilung liegt, d. h. eine Dichte besitzt, die in etwa mit der der Normalverteilung übereinstimmt. Dies braucht aber nicht zu gelten, vgl. Johnson u.a. (1980).

Verteilungen mit einem Parameter $\beta_2 > 3$ heißen dabei *leptokurtisch*, solche mit $\beta_2 < 3$ *platykurtisch*.

Beispiel 2.1.19 *Momente der Gleichverteilung und der Exponentialverteilung*

Wir betrachten die Gleichverteilung und die Exponentialverteilung und bestimmen ihre Maßzahlen für Schiefe und Wölbung.

Für die Gleichverteilung über dem Intervall $[0,1]$ gilt nach den obigen Beispielen: $\mathrm{E}(X) = 1/2$, $\mathrm{Var}(X) = 1/12$. Weiter erhalten wir

$$\mathrm{E}\left(X-\frac{1}{2}\right)^3 = \int_0^1 \left(x-\frac{1}{2}\right)^3 \mathrm{d}x = 0,$$

$$\mathrm{E}\left(X-\frac{1}{2}\right)^4 = \int_0^1 \left(x-\frac{1}{2}\right)^4 \mathrm{d}x = \frac{1}{80}.$$

Somit sind $\beta_1 = 0$ und $\beta_2 = 1.8$.

Für die Exponentialfunktion lässt sich $\mathrm{E}(X^k)$ mittels partieller Integration berechnen:

$$\mathrm{E}(X^k) = \int_0^\infty x^k \lambda e^{-\lambda x} \mathrm{d}x = \frac{k!}{\lambda^k}.$$

Das ergibt mit den angegebenen Berechnungformeln: $\beta_1 = 2$. Die Exponentialverteilung ist also rechtsschief.

2.1.3 Näherungsweise Bestimmung von Erwartungswerten

Insbesondere bei stetigen Zufallsvariablen ist die explizite Bestimmung von Erwartungswerten von transformierten Zufallsvariablen oft nicht möglich. Dann behilft man sich mit Näherungen, die auf Taylor-Approximationen beruhen.
Sei $g(x)$ eine in einem Intervall um x_0 differenzierbare Funktion mit der Ableitung $g'(x)$. Im Punkt x_0 lautet die Gleichung der Tangente $l(x)$ an $g(x)$:

$$l(x) = g(x_0) + g'(x_0)(x - x_0).$$

Um x_0 lässt sich $g(x)$ durch $l(x)$ linear approximieren, vgl. die Abbildung 2.1:

$$g(x) \approx g(x_0) + g'(x_0)(x - x_0).$$

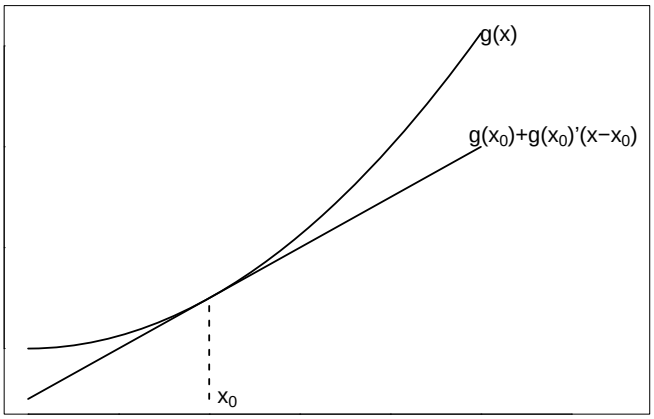


Abb. 2.1: *Zur linearen Approximation*

Eine bessere Approximation von $g(x)$ in der Nähe von x_0 erhält man aber durch eine quadratische Funktion. Der Ansatz

$$g(x) \approx a_0 + a_1(x - x_0) + a_2(x - x_0)^2$$

liefert für die Koeffizienten a_0, a_1, a_2, indem die Ableitungen der beiden Seiten an der Stelle x_0 als gleich angesetzt werden:

$$g(x_0) = a_0,$$
$$g'(x_0) = a_1,$$
$$g''(x_0) = 2a_2.$$

Somit lautet das approximierende Polynom zweiten Grades:

$$g(x_0) + g'(x_0)(x - x_0) + (x - x_0)^2.$$

Ist $g(x)$ nun in einem Intervall um x_0 $k+1$-mal differenzierbar, so kann ein entsprechendes Polynom vom Grade k zur Approximation verwendet werden. Natürlich wird es in der Regel die Funktion auch bei größerem k noch nicht genau darstellen. Vielmehr wird ein gewisser Rest bei der Darstellung bleiben. Nach Lagrange gilt mit $0 < \delta < 1$:

$$g(x) = \sum_{u=0}^{k} \frac{g^{(u)}(x_0)}{u!} \cdot (x - x_0)^u + \frac{g^{(k+1)}(x_0 + \delta \cdot (x - x_0))}{(k+1)!} \cdot (x - x_0)^{k+1}.$$

Satz 2.1.20 *Taylorreihenentwicklung*

Sofern die Funktion $g(x)$ in der Umgebung von x_0 beliebig oft differenzierbar ist, und dort für alle x und alle $k = 1, 2, 3, \ldots$ gilt: $|g^{(k)}(x)| < M$ mit einer geeigneten Konstanten M, so erhält man schließlich die *Taylorreihenentwicklung* von $g(x)$ um den Punkt x_0:

$$g(x) = \sum_{k=0}^{\infty} \frac{g^{(k)}(x_0)}{k!} \cdot (x - x_0)^k.$$

Die *Taylor-Approximation* nutzt in der Regel nur die Entwicklung bis zur ersten oder zweiten Potenz aus.

Im Fall der approximativen Bestimmung von Erwartungswerten erhält man die gewünschte Beziehung, indem die Funktion $g(x)$ den Erwartungswert μ der Zufallsvariablen X entwickelt und in der resultierenden Entwicklung anschließend zum Erwartungswert übergegangen wird:

$$\mathrm{E}(g(X)) \approx g(\mu) + g'(\mu) \cdot \mathrm{E}(X - \mu) + \frac{1}{2} \cdot g''(\mu) \cdot \mathrm{E}(X - \mu)^2 = g(\mu) + \frac{1}{2} \cdot g''(\mu) \cdot \mathrm{E}(X - \mu)^2.$$

Diese Vorgehensweise ergibt oft recht gute Resultate, jedoch ist die Rechtfertigung zu überprüfen. Insbesondere die Beschränkung auf die ersten beiden Ableitungen kann sehr unzureichend sein, wenn die Verteilung nicht eng um den Erwartungswert konzentriert ist.

Beispiel 2.1.21 *Erwartungswert einer transformierten gleichverteilten Zufallsvariablen*

Sei die Transformation der Kehrwert, $g(x) = \frac{1}{x}$. Dann ist $g'(x) = -\frac{1}{x^2}$ und $g''(x) = \frac{2}{x^3}$. Für eine über dem Intervall $[1, 1 + \vartheta]$ gleichverteilte Zufallsvariable X mit der Dichte $f(x) = 1/\vartheta$, $1 \le x \le 1 + \vartheta$, gilt

$$\mathrm{E}(X)=\mu=1+\frac{\vartheta}{2} \quad \text{und} \quad \mathrm{E}(X-\mu)^2=\frac{\vartheta^2}{12}.$$

Damit folgt die Approximation:

$$\begin{aligned}\mathrm{E}\left(\frac{1}{X}\right)=\mathrm{E}(g(X))&\approx\frac{1}{\mu}+\frac{1}{2}\cdot\frac{2}{\mu^3}\cdot\frac{\vartheta^2}{12}=\frac{1}{1+\vartheta/2}+\frac{1}{(1+\vartheta/2)^3}\cdot\frac{\vartheta^2}{12}\\&\approx\left\{\sum_{u=0}^{\infty}\left(\frac{-\vartheta}{2}\right)^u\right\}+\left\{\sum_{u=0}^{\infty}\left(\frac{-\vartheta}{2}\right)^u\right\}^3\cdot\frac{\vartheta^2}{12}\approx 1-\frac{1}{2}\vartheta+\frac{1}{3}\vartheta^2-\frac{5}{24}\vartheta^3+\frac{7}{48}\vartheta^4.\end{aligned}$$

Direktes Ausrechnen ergibt zum Vergleich:

$$\begin{aligned}\mathrm{E}\left(\frac{1}{X}\right)&=\int_1^{1+\vartheta}\frac{1}{x}\cdot\frac{1}{\vartheta}\,\mathrm{d}x=\frac{1}{\vartheta}[\ln(1+\vartheta)-\ln(1)]=\frac{1}{\vartheta}\ln(1+\vartheta)=\frac{1}{\vartheta}\cdot\left\{\sum_{u=1}^{\infty}\left(-\frac{1}{u}\right)^{u+1}\vartheta^u\right\}\\&\approx 1-\frac{1}{2}\vartheta+\frac{1}{3}\vartheta^2-\frac{1}{4}\vartheta^3+\frac{1}{5}\vartheta^4.\end{aligned}$$

Entsprechend dem Ansatz stimmt die Approximation bis zur zweiten Potenz mit dieser Reihenentwicklung überein. Offensichtlich ist die Approximation umso besser, je kleiner ϑ ist.

2.2 Momenterzeugende Funktion

Allgemein werden Erwartungswerte der Form $\mathrm{E}(X^r)$ als *Momente* und die Erwartungswerte $\mathrm{E}((X-\mu)^r)$ als *zentrale Momente* bezeichnet. Als Bezeichnungen sind weiter üblich:

$$\begin{aligned}\mu_r&=\mathrm{E}(X^r)\\\mu_r'&=\mathrm{E}((X-\mu)^r).\end{aligned}\tag{2.21}$$

Natürlich gilt $\mu_1=\mu$. Wenn das k-te Moment einer Verteilung existiert, d. h. dass $\mathrm{E}(X^k)$ wohldefiniert ist, so existieren auch die r-ten Momente für $r=1,\ldots,k-1$.

Die Bedeutung der Momente liegt darin, dass in verschiedenen, für die Praxis relevanten Fällen die Momente die gesamte Verteilung festlegen. Das wird dann benutzt, um z. B. durch Gleichsetzung der ersten (zentralen) Momente eine Verteilung auszuwählen. Darauf wird später noch weiter eingegangen.

Ein wichtiges Hilfsmittel zur Bestimmung von Momenten ist die momenterzeugende Funktion.

Definition 2.2.1 *momenterzeugende Funktion*

Sei X eine eindimensionale Zufallsvariable. Sofern es ein $h>0$ gibt, so dass für alle $t\in(-h;h)$ gilt $\mathrm{E}(e^{tX})<\infty$, heißt

$$M(t)=M_X(t)=\mathrm{E}(e^{tX})\tag{2.22}$$

die *momenterzeugende Funktion* von X.

Beispiel 2.2.2 *Binomialverteilung*

X sei binomialverteilt mit den Parametern p und n, d.h.

$$f(x)=\binom{n}{x}p^x(1-p)^{n-x} \quad \text{für} \quad x=0,1,\ldots,n.$$

Dann gilt

$$\begin{aligned}M(t)=\mathrm{E}(e^{tX})^{tx}&=\sum_{x=0}^{n}e^{tx}\binom{n}{x}\cdot p^x(1-p)^{n-x}=\sum_{x=0}^{n}\binom{n}{x}(e^t p)^x(1-p)^{n-x}\\&=(e^t p+1-p)^n.\end{aligned}$$

Hier existiert $M(t)$ für alle $t\in\mathbb{R}$.

Beispiel 2.2.3 *Poisson-Verteilung*

Für eine Poisson-verteilte Zufallsvariable X gilt:

$$M(t)=\mathrm{E}(e^{tX})=\sum_{x=0}^{\infty}e^{tx}e^{-\lambda}\frac{\lambda^x}{x!}=e^{\lambda e^t}e^{-\lambda}\sum_{x=0}^{\infty}e^{-\lambda e^t}\cdot\frac{(\lambda e^t)^x}{x!}=e^{\lambda e^t-\lambda}.$$

Beispiel 2.2.4 *Normalverteilung*

X sei normalverteilt mit den Parametern μ und σ^2. Dann erhalten wir durch geeignete Umformungen:

$$\begin{aligned}M(t)&=\int_{-\infty}^{\infty}e^{tx}\sqrt{\frac{1}{2\pi\sigma^2}}\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\mathrm{d}x\\&=\int_{-\infty}^{\infty}\exp\left[\mu t+\frac{\sigma^2t^2}{2}\right]\sqrt{\frac{1}{2\pi\sigma^2}}\cdot\exp\left[-\frac{1}{2\sigma^2}\cdot(x-(\mu+\sigma^2t))^2\right]\mathrm{d}x\\&=\exp\left[\mu t+\frac{\sigma^2t^2}{2}\right].\end{aligned}$$

Ist X eine Zufallsvariable mit der momenterzeugenden Funktion $M_X(t)$ und $Y=g(X)$ eine Transformation, so dass $\mathrm{E}(e^{g(X)t})$ für alle t aus einem Intervall um Null existiert, so ist die momenterzeugende Funktion durch diesen Erwartungswert gegeben: $M_Y(t)=\mathrm{E}(e^{g(X)t})$.

Beispiel 2.2.5 *momenterzeugende Funktion des Quadrates einer standardnormalverteilten Zufallsvariablen*

Sei X normalverteilt, $X\sim\mathcal{N}(\mu,\sigma^2)$. Dann gilt für $Y=(X-\mu)^2/\sigma^2$:

$$M_Y(t) = \int_{-\infty}^{\infty} \exp\left[\left(\frac{x-\mu}{\sigma}\right)^2 \cdot t\right] \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \mathrm{d}x$$

$$= \int_{-\infty}^{\infty} \frac{\sqrt{1-2t}}{\sqrt{2\pi}\sigma\sqrt{1-2t}} \exp\left[\frac{(x-\mu)^2}{2\sigma^2}(1-2t)\right] \mathrm{d}x = \frac{1}{\sqrt{1-2t}}.$$

Die zugehörige Verteilung ist eine χ^2-Verteilung mit einem Freiheitsgrad.

Satz 2.2.6 *Reihendarstellung von momenterzeugenden Funktionen*

Sei X eine Zufallsvariable mit momenterzeugender Funktion $M(t)$. Dann gibt es ein $h > 0$, so dass für alle $t \in (-h;h)$ gilt:

$$M(t) = \sum_{r=0}^{\infty} \frac{\mathrm{E}(X^r)}{r!} \cdot t^r.$$

Beweis: Zuerst folgen aus der Reihendarstellung der Exponentialfunktion

$$\exp(x) = e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$$

unmittelbar die Ungleichungen für $x \in \mathbb{R}$ und $r \in \mathbb{N}$, falls $h > 0$:

$$|x|^r < \frac{r! \cdot \exp(h|x|)}{h^r} < \frac{r!}{h^r} \cdot (e^{hx} + e^{-hx}).$$

Nun sei h so gewählt, dass $\mathrm{E}(e^{tX}) < \infty$ für $t \in [-2h;2h]$. Die Polynomapproximation mit dem Lagrange'schen Restglied ergibt für ein festes t aus dem angegebenen Intervall:

$$e^{tx} = \sum_{r=0}^{k} \frac{t^r x^r}{r!} + \frac{t^{k+1} e^{\delta(x,k)} x^{k+1}}{(k+1)!},$$

wobei $\delta(x,k)$ zwischen 0 und x liegt. Weiter gilt wegen obiger Ungleichung:

$$\left|\frac{t^{k+1} e^{\delta(x,k)} x^{k+1}}{(k+1)!}\right| < \frac{|t|^{k+1} \exp(h|x|) \exp(h|x|)}{h^{k+1}} < \left(\frac{|t|}{h}\right)^{k+1} \cdot (e^{2hx} + e^{-2hx}).$$

Damit folgt leicht:

$$\left|\sum_{r=0}^{k} \frac{\mathrm{E}(X^r)}{r!} t^r - M(t)\right| \leq \mathrm{E}\left|\sum_{r=0}^{k} \frac{t^r X^r}{r!} - e^{tX}\right| < \left(\frac{|t|}{h}\right)^{k+1} \cdot [M(2h) + M(-2h)].$$

Mit $k \to \infty$ geht der rechte Term gegen null. Also folgt

$$M(t) = \lim_{k\to\infty} \sum_{r=0}^{k} \frac{\mathrm{E}(X^r)}{r!} \cdot t^r = \sum_{r=0}^{\infty} \frac{\mathrm{E}(X^r)}{r!} \cdot t^r.$$

Dies gilt für alle $t \in (-h;h)$. Über die momenterzeugende Funktion erhalten wir die Momente durch sukzessives Differenzieren und Bestimmen der Werte von $M^{(k)}(t)$ an der Stelle $t = 0$.

Beispiel 2.2.7 *erstes und zweites Moment*

Die Gewinnung der Momente über die Ableitung von $M(t)$ soll für $r=1$ und $r=2$ direkt verifiziert werden. Für $r=1$ gilt, wenn Differentation und Integration vertauscht werden dürfen:

$$M^{(1)}(0) = \frac{\partial}{\partial t}\int_{-\infty}^{\infty} e^{tx}\,\mathrm{d}F(x)\bigg|_{t=0} = \int_{-\infty}^{\infty}\frac{\partial}{\partial t}e^{tx}\,\mathrm{d}F(x)\bigg|_{t=0} = \int_{-\infty}^{\infty} x e^{tx}\,\mathrm{d}F(x)\bigg|_{t=0}$$
$$= \int_{-\infty}^{\infty} x e^{0\cdot x}\,\mathrm{d}F(x) \qquad = \int_{-\infty}^{\infty} x\,\mathrm{d}F(x) = \mathrm{E}(X).$$

Für $r=2$ folgt:

$$M^{(2)}(0) = \frac{\partial^2}{\partial t^2}\int_{-\infty}^{\infty} e^{tx}\,\mathrm{d}F(x)\bigg|_{t=0} = \int_{-\infty}^{\infty}\frac{\partial^2}{\partial t^2}e^{tx}\,\mathrm{d}F(x)\bigg|_{t=0} = \int_{-\infty}^{\infty} x^2 e^{tx}\,\mathrm{d}F(x)\bigg|_{t=0}$$
$$= \int_{-\infty}^{\infty} x^2\,\mathrm{d}F(x) = \mathrm{E}(X^2).$$

Beispiel 2.2.8 *Normalverteilung*

Für eine normalverteilte Zufallsvariable X mit den Parametern $\mu=0$ und $\sigma^2>0$ gilt:

$$M(t) = e^{\sigma^2 t^2/2} = \sum_{r=0}^{\infty}\frac{(\sigma^2 t^2/2)^r}{r!} = \sum_{r=0}^{\infty}\frac{(\sigma^2/2)^r}{r!}(2r)!\frac{t^{2r}}{(2r)!}.$$

Somit ist im Fall $\mu=0$:

$$\mathrm{E}(X^k) = \begin{cases} 0 & \text{falls k ungerade} \\ \left(\dfrac{\sigma^2}{2}\right)^{k/2}\cdot\dfrac{k!}{(k/2)!} & \text{falls k gerade}. \end{cases}$$

Satz 2.2.9 *Gleichheit momenterzeugender Funktion*

Seien X und Y zwei Zufallsvariablen mit momenterzeugenden Funktionen $M_X(t)$, $M_Y(t)$. Gilt

$$M_X(t) = M_Y(t) \qquad \text{für alle } t \text{ aus einem Intervall um Null},$$

so sind die Verteilungsfunktionen von X und Y gleich: $F_X(t) = F_Y(t)$.

Für $Y = aX+b$ gilt:

$$M_Y(t) = e^{bt}\cdot M_X(at).$$

Beweis: Der Beweis der ersten Aussage übersteigt unseren Rahmen; der der zweiten Aussage erfolgt einfach durch Hinschreiben und Umformen der jeweiligen Erwartungswerte.

Der Satz bildet einerseits die Grundlage für viele theoretische Verteilungsaussagen; andererseits wir auch in der angewandten Statistik häufig daraus die Rechtfertigung gezogen, sich auf die ersten Momente zurückzuziehen, wenn es um Verteilungseigenschaften geht. Dabei ist die erste, fundamentale Aussage des Satzes unter praktischen Gesichtspunkten eher als dubios anzusehen, wie das folgende Beispiel zeigt.

Beispiel 2.2.10 *modulierte Standardnormalverteilung*

Sei $\phi(x)$ die Dichte der Standardnormalverteilung, $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$, und sei $f(x) = \phi(x) \cdot \left(1 + \frac{1}{2}\sin(2\pi x)\right)$.

$f(x)$ ist also eine modulierte Version die Dichte der Standardnormalverteilung. Es lässt sich zeigen, dass $f(x)$ tatsächlich eine Dichte ist, $f(x) \geq 0$ und $\int_{infty}^{\infty} f(x)\mathrm{d}x = 1$. Die beiden Dichten sind in der Abbildung 2.2 gegenübergestellt.

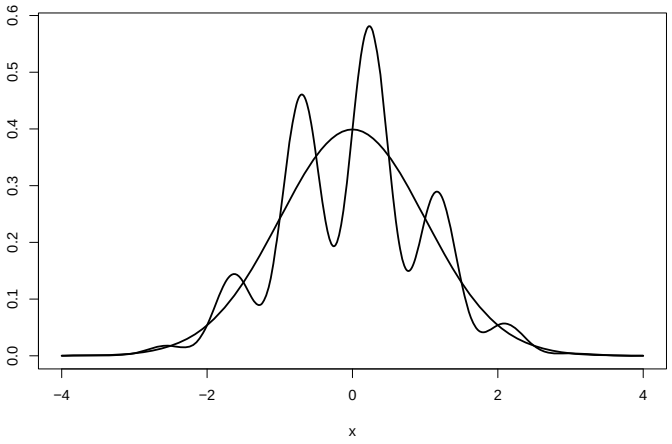


Abb. 2.2: *$\phi(x)$ und $f(x)$*

Für die momenterzeugenden Funktionen $M_1(t)$ und $M_2(t)$ gilt, vgl. McCullach (1994):

$$\ln(M_1(t)) = \frac{1}{2}t^2,$$
$$\ln(M_2(t)) = \frac{1}{2}t^2 + \ln\left(1 + \frac{1}{2}e^{-2\pi^2}\sin(2\pi t)\right).$$

Die maximale Differenz $|\ln(M_1(t)) - \ln(M_2(t))|$ beträgt $1.34 \cdot 10^{-9}$. Diese ist unter praktischen Gesichtspunkten vernachlässigbar.

Im folgenden Lemma stellen wir einige einfache Eigenschaften der momenterzeugenden Funktion zusammen.

Lemma 2.2.11 *Eigenschaften momenterzeugender Funktionen*

X sei eine Zufallsvariable mit der momenterzeugenden Funktion $M(t)$, die für $t \in (-T, T)$, $T > 0$, definiert ist. Dann hat $M(t)$ folgende Eigenschaften:

(i) $M(t)$ ist strikt positiv, stetig und stetig differenzierbar für $|t| < T$.

(ii) Weiter gilt $M(0) = 1$ und $M'(0) = \mathrm{E}(X)$.

(iii) Falls $\mathrm{E}(X) < 0$, dann gibt es ein $T_1 > 0$, $T_1 < T$, mit $M(t) < 1$ für $0 < t < T_1$.

(iv) Für $0 < t < T$ gilt die Ungleichung $\mathrm{P}(X > 0) \leq M(t)$.

Beweis: (i) ergibt sich über die Definition, (ii) wiederholt nur Bekanntes und (iii) folgt mit den ersten beiden Teilen. Wir zeigen nun (iv).
X habe die Verteilungsfunktion $F(x)$. Stets gilt $e^{tx} \geq 0$. Für $x \geq 0$ und $0 < t < T$ ist sogar $e^{tx} \geq 1$. Damit erhalten wir

$$M(t) = \int_{-\infty}^{0} e^{tx}\, dF(x) + \int_{0}^{\infty} e^{tx}\, \mathrm{d}F(x) \geq \int_{0}^{\infty} e^{tx}\, \mathrm{d}F(x) \geq \int_{0}^{\infty} \mathrm{d}F(x) = \mathrm{P}(X > 0).$$

Die Definition der momenterzeugenden Funktion lässt sich leicht auf den mehrdimensionalen Fall ausdehnen.

Definition 2.2.12 *gemeinsame momenterzeugende Funktion*

Sei $\boldsymbol{X} = (X_1, \ldots, X_k)$ eine k-dimensionale Zufallsvariable. Gibt es Intervalle $(-h_i, h_i)$, $h_i > 0$ $(i = 1, \ldots, k)$, so dass für alle $t_i \in (-h_i, h_i)$ der Erwartungswert $\mathrm{E}(e^{t_1 X_1 + \ldots + t_k X_k})$ existiert, so heißt

$$M(\boldsymbol{t}) = \mathrm{E}(e^{t_1 X_1 + \ldots + t_k X_k}) \tag{2.23}$$

die *gemeinsame momenterzeugende Funktion* von $\boldsymbol{X}$.

Die folgenden Sätze werden für zweidimensionale Zufallsvariablen formuliert. Sie gelten in Verallgemeinerung entsprechend für den k-dimensionalen Fall.

Satz 2.2.13 *momenterzeugende Funktion und Unabhängigkeit*

Seien X und Y Zufallsvariablen mit gemeinsamer momenterzeugender Funktion $M_{XY}(t_1, t_2)$. Dann gilt:

$$M_X(t_1) = M_{XY}(t_1, 0),$$
$$M_Y(t_2) = M_{XY}(0, t_2).$$

X und Y sind genau dann unabhängig, wenn

$$M_{XY}(t_1, t_2) = M_X(t_1) M_Y(t_2).$$

Satz 2.2.14 *momenterzeugende Funktion einer Summe*

Seien X und Y unabhängige Zufallsvariablen mit momenterzeugenden Funktionen $M_X(t)$ und $M_Y(t)$. Dann gilt für $Z = X + Y$:

$$M_Z(t) = M_X(t) \cdot M_Y(t).$$

Beweis: Der Beweis erfolgt wieder durch Hinschreiben und Umformen der jeweiligen Erwartungswerte. Dabei wird noch ausgenutzt, dass bei Unabhängigkeit der Erwartungswert eines Produktes gleich dem Produkt der Erwartungswerte ist.

Bei manchen Verteilungen ist die Summe zweier unabhängiger Zufallsvariablen, die jeweils eine Verteilung dieses Typs haben, wieder eine Verteilung des gleichen Typs. Dann sprechen wir von der *Reproduktivitätseigenschaft* dieser Verteilung.

Beispiel 2.2.15 *binomialverteilte Zufallsvariablen*

Sind X und Y unabhängige binomialverteilte Zufallsvariablen mit den Parametern n und p bzw. m und p, so ist $X + Y$ binomialverteilt mit den Parametern $n + m$ und p. Wie wir im Beispiel 2.2.2 gesehen haben, gilt

$$M_X(t) = (e^t p + 1 - p)^n, \quad M_Y(t) = (e^t p + 1 - p)^m .$$

Folglich ist

$$M_{X+Y}(t) = (e^t p + 1 - p)^{n+m} ,$$

was die Verteilungsaussage für $X + Y$ beweist.

Auch unabhängige Poisson-verteilte Zufallsvariablen sind wieder Poisson-verteilt. Mit Beispiel 2.2.3 gilt:

$$M_X(t) = e^{\lambda e^t - \lambda}, \quad M_Y(t) = e^{\mu e^t - \mu} \quad \Rightarrow \quad M_{X+Y}(t) = e^{(\lambda+\mu)e^t - (\lambda+\mu)}.$$

Beispiel 2.2.16 *normalverteilte Zufallsvariablen*

X und Y seien unabhängige normalverteilte Zufallsvariablen, $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ und $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Dann ist $Z = X + Y$ wieder normalverteilt. Nach Beispiel 2.2.4 gilt:

$$M_X(t) = \exp\left[\mu_X \cdot t + \frac{\sigma_X^2 t^2}{2}\right], \quad M_Y(t) = \exp\left[\mu_Y \cdot t + \frac{\sigma_Y^2 t^2}{2}\right]$$

$$\Rightarrow M_{X+Y}(t) = \exp\left[(\mu_X + \mu_Y) \cdot t + \frac{(\sigma_X^2 + \sigma_Y^2) t^2}{2}\right].$$

Beispiel 2.2.17 *χ^2-Verteilung*

Die allgemeine χ^2-Verteilung mit k Freiheitsgraden erhalten wir als Summe von k unabhängigen Zufallsvariablen, die alle χ^2-verteilt sind mit einem Freiheitsgrad. Die zugehörige momenterzeugende Funktion ist offenbar, vgl Beispiel 2.2.5:

$$M(t) = \frac{1}{(1-2t)^{k/2}}\,.$$

2.3 Bedingte Erwartungswerte

Die Berücksichtigung zusätzlicher Information bei Zufallsexperimenten kann durch die Betrachtung von ‚bedingten Ereignissen' bzw. bedingten Wahrscheinlichkeiten erfolgen. Dies führt u. a. zu den bedingten Verteilungen. Hier werden nun die Erwartungswerte solcher bedingter Verteilungen betrachtet. Sie werden in aufsteigendem Schwierigkeitsgrad eingeführt. Die einzelnen Stufen, die zu ihrer Definition führen, sind dabei von eigenständiger Bedeutung und werden an entsprechender Stelle wieder aufgegriffen.

Definition 2.3.1 *bedingter Erwartungswert bei gegebenem Ereignis*

X sei eine diskrete Zufallsvariable mit der Dichte $f(x)$. A sei ein Ereignis mit $\mathrm{P}(A) > 0$. Der *bedingte Erwartungswert* von X bei gegebenem A ist definiert durch

$$\mathrm{E}(X|A) = \sum_x x \cdot \mathrm{P}(X = x|A). \tag{2.24}$$

Dabei wird vorausgesetzt, dass $\mathrm{E}(X|A)$ existiert.

Satz 2.3.2 *Erwartungswert und Zerlegung des Stichprobenraumes*

Sei $\Omega_1, \ldots, \Omega_G$ eine Zerlegung von Ω, d. h. $\Omega_g \cap \Omega_h = \emptyset$ für $g \neq h$ und $\Omega_1 \cup \ldots \cup \Omega_G = \Omega$. Dann gilt für eine Zufallsvariable mit der Verteilungsfunktion $F(x)$:

$$\mathrm{E}(X) = \sum_{g=1}^{G} \mathrm{E}(X|\Omega_g) \cdot \mathrm{P}(\Omega_g).$$

Beweis:

$$\int_\Omega x\,\mathrm{d}F(x) = \sum_{g=1}^{G} \int_{\Omega_g} x\,\mathrm{d}F(x) = \sum_{g=1}^{G} \int_{\Omega_g} x\,\mathrm{d}F(x|\Omega_g) \cdot \mathrm{P}(\Omega_g) = \sum_{g=1}^{G} \mathrm{E}(X|\Omega_g) \cdot \mathrm{P}(\Omega_g)\,.$$

Man denke sich nun das Vorgehen zweistufig, so dass auf der ersten Stufe ein Ereignis Ω_g aus einer Zerlegung von Ω eintritt und auf der zweiten der bedingte Erwartungswert $\mathrm{E}(X|\Omega_g)$ beobachtet wird. Dann ist einsichtig, dass das Zufallsexperiment einen von insgesamt G möglichen Werten ergeben wird. Mit anderen Worten wird dadurch eine neue Zufallsvariable definiert.

Von besonderer Bedeutung ist nun der Fall, dass die Zerlegung durch eine Zufallsvariable Y induziert wird, die Ω_g also folgende Form haben:

$$\Omega_g = Y^{-1}(y) = \{\omega | \, \omega \in \Omega, \; Y(\omega) = y\}.$$

Der resultierende bedingte Erwartungswert ist der Erwartungswert der bedingten Verteilung von X, gegeben dass Y den Wert y annimmt. Der bedingte Erwartungswert $\mathrm{E}(X|Y=y)$ ist erst dann ermittelbar, wenn $\{Y=y\}$ beobachtet wurde. Vorher hängt er offensichtlich von der Zufallsvariablen Y ab. Somit ist $\mathrm{E}(X|Y)$ selbst eine Zufallsvariable, genauer eine Funktion der Zufallsvariablen Y.

Definition 2.3.3 *bedingter Erwartungswert und bedingte Erwartung*

(X,Y) sei eine zweidimensionale Zufallsvariable mit der gemeinsamen Verteilungsfunktion $F(x,y)$. Der *bedingte Erwartungswert* von X bei gegebenem $\{Y=y\}$ ist definiert durch

$$\mathrm{E}(X|Y=y) = \int_{-\infty}^{\infty} x \, \mathrm{d}F_{X|Y}(x|y). \tag{2.25}$$

Dabei wird vorausgesetzt, dass $\mathrm{E}(X|Y=y)$ existiert.

Die *bedingte Erwartung* von X gegeben Y, $\mathrm{E}(X|Y)$, ist die Zufallsvariable $g(Y)$, die festgelegt ist durch $g(y) = \mathrm{E}(X|Y=y)$.

Für diskrete und stetige Zufallsvariablen lässt sich der bedingte Erwartungswert auch durch

$$\sum_x x \cdot \frac{\mathrm{P}(X=x, Y=y)}{\mathrm{P}(Y=y)} \qquad \text{bzw. durch} \qquad \int_{-\infty}^{\infty} x \cdot \frac{f(x,y)}{f_Y(y)} \, \mathrm{d}x$$

angeben.

Beispiel 2.3.4 *zwei exponentialverteilte Zufallsvariablen - Fortsetzung von Seite 43*

X und Y seien zwei unabhängige exponentialverteilte Zufallsvariablen mit der gleichen Dichte

$$f(t) = \lambda e^{-\lambda t} \quad \text{für} \quad t > 0\,.$$

Es soll der bedingte Erwartungswert $\mathrm{E}(X+Y|Y=y)$ bestimmt werden. Nach Beispiel 1.3.10 gilt

$$f_{X+Y,Y}(u,y) = \lambda^2 e^{-\lambda u} \quad \text{für} \quad u > y > 0.$$

Damit folgt:

$$f_{X+Y|Y}(u|y) = \begin{cases} \dfrac{\lambda^2 e^{-\lambda u}}{\lambda e^{-\lambda y}} & \text{für } u > y > 0 \\ 0 & \text{sonst} \end{cases} = \begin{cases} \lambda e^{-\lambda(u-y)} & \text{für } u > y > 0 \\ 0 & \text{sonst} \end{cases}$$

und weiter:

$$\mathrm{E}(X+Y|Y=y)=\int_y^\infty u\cdot f_{X+Y|Y}(u|y)\,\mathrm{d}u=\int_y^\infty u\cdot\lambda\cdot e^{-\lambda(u-y)}\mathrm{d}u=y+\frac{1}{\lambda}.$$

Die bedingte Erwartung von $X+Y$ bei gegebenem Y ist dann

$$\mathrm{E}(X+Y|Y)=Y+\frac{1}{\lambda}.$$

Das ist die um den Erwartungswert von X verschobene Zufallsvariable Y selbst.

Es werden ohne Beweis die folgenden Eigenschaften des bedingten Erwartungswertes angegeben.

Satz 2.3.5 *Eigenschaften des bedingten Erwartungswertes*

X, Y und Z seien Zufallsvariablen, so dass die folgenden Erwartungswerte existieren.

(i) Für feste Zahlen a und b gilt:

$$\mathrm{E}(aX+bY|Z=z)=a\mathrm{E}(X|Z=z)+b\mathrm{E}(Y|Z=z).$$

(ii) Sind X und Y unabhängig, so ist

$$\mathrm{E}(X|Y=y)=\mathrm{E}(X).$$

(iii) Für jede reellwertige Funktion $g(y)$ gilt

$$\mathrm{E}(X\cdot g(Y)|Y=y)=g(y)\cdot\mathrm{E}(X|Y=y).$$

Die Eigenschaft (iii) ist folgendermaßen zu verstehen: Ist (X,Y) eine zweidimensionale Zufallsvariable, so auch $(X\cdot g(Y),Y)$. Die Formel stellt den Zusammenhang mit dem bedingten Erwartungswert von X bei gegebenem $\{Y=y\}$ her. Auch wenn sie sehr plausibel erscheint, ist die Formel nicht trivial. Für eine diskrete Zufallsvariable (X,Y) ist z. B. die gemeinsame Verteilung von $(X\cdot g(Y),Y)$ zu bestimmen:

$$\mathrm{P}(X\cdot g(Y)=u,Y=y)=\sum_{\substack{x\\ x\cdot g(y)=u}}\mathrm{P}(X=x,Y=y).$$

Erst damit kann der bedingte Erwartungswert der Zufallsvariablen $X\cdot g(Y)$ bei gegebenem $\{Y=y\}$ bestimmt werden.

Die im Satz angegebenen Eigenschaften übertragen sich natürlich auf die bedingte Erwartung:

$$\mathrm{E}(aX+bY|Z)=a\mathrm{E}(X|Z)+b\mathrm{E}(Y|Z)$$

$$\mathrm{E}(X\cdot g(Y)|Y)=g(Y)\cdot\mathrm{E}(X|Y).$$

Wesentlich bei der bedingten Erwartung $\mathrm{E}(X|Y)$ ist die Tatsache, dass nur die Verteilung, nicht die Beobachtung von X, eine Rolle bei der Bestimmung von $\mathrm{E}(X|Y)$ spielt. Das hat einige Konsequenzen für die Verteilung der bedingten Erwartung.

Beispiel 2.3.6 *eine zweidimensionale diskrete Zufallsvariable*

Gegeben sei die zweidimensionale diskrete Zufallsvariable (X,Y) auf Ω, deren Realisationsmöglichkeiten den Werten der Tupel aus Ω entsprechen:

$$\Omega = \left\{ \begin{array}{l} (3,-2),(4,-2),(5,-2),(0,-1),(1,-1),(2,-1), \\ (-1,0),\ (0,0),\ (1,0),\ (0,1),\ (1,1),\ (2,1) \end{array} \right\}$$

Die Zufallsvariable (X,Y) habe über Ω eine Gleichverteilung, $\mathrm{P}(X=x,Y=y)=1/12$ für $(x,y)\in\Omega$.
Um $\mathrm{E}(X|Y)$ zu bestimmen, muss für jedes y der bedingte Erwartungswert $\mathrm{E}(X|Y=y)$ bestimmt werden. Wegen der Symmetrie gilt:

$$\mathrm{E}(X|Y=-2)=4, \quad \mathrm{E}(X|Y=-1)=1,$$
$$\mathrm{E}(X|Y=0)=0, \quad \mathrm{E}(X|Y=1)=1;$$

zusammmengefasst ist

$$\mathrm{E}(X|Y)=Y^2.$$

Da $\mathrm{E}(X|Y)$ eine Zufallsvariable ist, kann man sich für ihren Erwartungswert interessieren. Weil (X,Y) und damit auch $Y^2=\mathrm{E}(X|Y)$ diskret ist, folgt :

$$\mathrm{E}(Y^2)=\mathrm{E}(\mathrm{E}(X|Y))=0\frac{1}{4}+1\frac{2}{4}+4\frac{1}{4}=1.5.$$

Für den Erwartungswert von X erhält man direkt:

$$\mathrm{E}(X)=-1\frac{1}{12}+0\frac{3}{12}+1\frac{3}{12}+2\frac{2}{12}+3\frac{1}{12}+4\frac{1}{12}+5\frac{1}{12}=1.5.$$

Die Erwartungswerte von X und von der bedingten Erwartung $\mathrm{E}(X|Y)$ stimmen also überein.

Das im Beispiel erhaltene Resultat gilt allgemein.

Satz 2.3.7 *Erwartungswert der bedingten Erwartung*

Für die Zufallsvariablen X und Y gilt, wenn $\mathrm{E}(X)$ existiert:

$$\mathrm{E}(\mathrm{E}(X|Y))=\mathrm{E}(X).$$

Beweis: Es wird der diskrete Fall betrachtet. Für stetige Zufallsvariablen geht der Beweis analog.

Sei $g(Y)=\mathrm{E}(X|Y)$. Dann gilt

$$\begin{aligned}\mathrm{E}(g(Y)) &= \sum_y g(y)\cdot\mathrm{P}(Y=y) = \sum_{\substack{y \\ \mathrm{P}(Y=y)>0}} g(y)\cdot\mathrm{P}(Y=y) \\ &= \sum_{\substack{y \\ \mathrm{P}(Y=y)>0}} \left[\sum_x x\cdot\mathrm{P}(X=x|Y=y)\right]\cdot\mathrm{P}(Y=y) = \sum_{(x,y)} x\cdot\mathrm{P}(X=x,Y=y) \\ &= \mathrm{E}(X).\end{aligned}$$

Die letzte Gleichung gilt dabei wegen Satz 2.1.4 mit $g(X,Y)=X$. Damit ist $\mathrm{E}(g(Y))=\mathrm{E}(\mathrm{E}(X|Y))=\mathrm{E}(X)$.

Die Varianz einer bedingten Verteilung ist im diskreten Fall offensichtlich

$$\mathrm{Var}(X|Y=y) = \mathrm{E}(X^2|Y=y) - \mathrm{E}(X|Y=y)^2.$$

Das führt zu der folgenden Definition.

Definition 2.3.8 *bedingte Varianz*

Die *bedingte Varianz von X gegeben Y* ist die Zufallsvariable, die definiert ist durch

$$\mathrm{Var}(X|Y) = \mathrm{E}(X^2|Y) - \mathrm{E}(X|Y)^2. \tag{2.26}$$

Satz 2.3.9 *Varianz und bedingte Varianz*

Es gilt

$$\mathrm{Var}(X) = \mathrm{E}(\mathrm{Var}(X|Y)) + \mathrm{Var}(\mathrm{E}(X|Y)). \tag{2.27}$$

Beweis:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \mathrm{E}(\mathrm{E}(X^2|Y)) - \mathrm{E}(\mathrm{E}(X|Y))^2 \\ &= \mathrm{E}(\mathrm{E}(X^2|Y)) - \mathrm{E}(\mathrm{E}(X|Y))^2 - \mathrm{E}(\mathrm{E}(X|Y)^2) + \mathrm{E}(\mathrm{E}(X|Y)^2) \\ &= \mathrm{E}(\mathrm{E}(X^2|Y)) - \mathrm{E}(\mathrm{E}(X|Y)^2) + \mathrm{Var}(\mathrm{E}(X|Y)) \\ &= \mathrm{E}(\mathrm{Var}(X|Y)) + \mathrm{Var}(\mathrm{E}(X|Y)).\end{aligned}$$

Im Zusammenhang mit den Stichproben aus endlichen Grundgesamtheiten wird bisweilen statt der durch Y erzeugten Zerlegung des Stichprobenraumes Ω eine nicht durch eine Zufallsvariable gegebene Zerlegung $\Omega_1, \ldots, \Omega_G$ benötigt. Das lässt sich auf die betrachtete Situation zurückführen, indem eine Zufallsvariable so definiert wird, dass sie gerade auf den Teilmengen konstant ist. Wir erhalten damit:

$$\mathrm{Var}(X|\Omega_1, \ldots, \Omega_G) = \mathrm{E}(X^2|\Omega_1, \ldots, \Omega_G) - \mathrm{E}(X|\Omega_1, \ldots, \Omega_G)^2$$

und

$$\mathrm{Var}(X) = \mathrm{E}(\mathrm{Var}(X|\Omega_1, \ldots, \Omega_G)) + \mathrm{Var}(\mathrm{E}(X|\Omega_1, \ldots, \Omega_G)).$$

2.4 Aufgaben

Aufgabe 1

Bestimmen Sie die momenterzeugenden Funktionen $M(t)$ der folgenden Verteilungen. Geben Sie jeweils den Bereich an, für den $M(t)$ definiert ist und bestimmen Sie die ersten vier Momente.

(i) Exponentialverteilung: $f(x) = \lambda \exp[-\lambda x]$ für $x > 0$
(ii) Gleichverteilung: $f(x) = 1/\theta$ für $0 < x < 1$
(iii) geometrische Verteilung: $f(x) = p(1-p)^x$ für $x = 0, 1, 2, 3, \ldots$
(iv) Simpson-Verteilung: $f(x) = 1 - |x|$ für $-1 < x < 1$.

Aufgabe 2

i) Die Wahrscheinlichkeitsfunktion der Zufallsvariablen X sei gegeben durch:

$$P(X=-1)=\frac{a^2}{2},\quad P(X=0)=1-a^2,\quad P(X=1)=\frac{a^2}{2}.$$

Skizzieren Sie die Wahrscheinlichkeiten $P(|X-\mu|\geq k\mu)$ als Funktion von $k\geq 0$ für $a=\frac{1}{4},\frac{1}{2},\frac{3}{4}$ und in das gleiche Diagramm die Höchstwahrscheinlichkeiten nach der Tschebyscheff–Ungleichung.
Was lässt sich daraus schließen?

ii) X sei eine diskrete Zufallsvariable. $\bar{X}$ sei aus einer Stichprobe vom Umfang n ermittelt. Zeigen Sie:
Bei der Tschebyschev–Ungleichung $P(|\bar{X}-\mu|\geq K\sigma)\leq\frac{1}{nK^2}$ gilt, falls $0<\sigma<\infty$:

a) Für $n=1$ kann das Gleichheitszeichen für $K\leq 1$ angenommen werden.

b) Für $n=2$ kann das Gleichheitszeichen nur für $K=1$ angenommen werden.

c) Für $n>2$ kann das Gleichheitszeichen für kein $K>0$ angenommen werden.

Hinweis: Überlegen Sie sich die Abschätzung beim Beweis der Tschebyschev–Ungleichung.

Aufgabe 3

X sei Pareto-verteilt, $f(x)=\alpha k^\alpha x^{-(\alpha+1)}$ für $x>k$. Bestimmen Sie soweit wie möglich die Momente von X.

Aufgabe 4

Sei X eine Zufallsvariable mit momenterzeugender Funktion $M(t)$, und $E(X)=\mu$, $Var(X)=\sigma^2$. Sei $C(t)=\ln M(t)$. Zeigen Sie, dass $C'(0)=\mu$ und $C''(0)=\sigma^2$.

Aufgabe 5

Seien X und Y unabhängige Poisson-verteilte Zufallsvariablen mit Parametern λ und ν. Zeigen Sie, dass $X+Y$ wieder Poisson-verteilt ist mit dem Parameter $\lambda+\nu$.

Aufgabe 6

(X,Y) sei eine bivariate Zufallsvariable mit der Dichte

$$f_\vartheta(x,y)=y^2e^{-y(x+\vartheta)}\quad\text{für}\quad x,y\geq 0.$$

Dabei ist $\vartheta>0$. Bestimmen Sie die Regressionsfunktion $E(Y|X=x)$ und die bedingten Varianzen $Var(Y|X=x)$.

Aufgabe 7

Bestimmen Sie die Dichte einer $\chi^2(1)$-Verteilung aus der Beziehung $f(x)=F'(x)$ unter Verwendung der Tatsache, dass das Quadrat einer $\mathcal{N}(0,1)$-verteilten Zufallsvariablen $\chi^2(1)$-verteilt ist.

3 Statistische Modelle

3.1 Verteilungsfamilien

3.1.1 Grundlagen

In der Wahrscheinlichkeitsrechnung geht man in der Regel von einer nicht näher spezifizierten Wahrscheinlichkeitsverteilung aus und untersucht Konsequenzen struktureller Eigenschaften der Verteilung. In der Statistik dagegen besteht eine wesentliche Fragestellung darin, dass nach Festlegung einer relevanten Zufallsvariablen X Aussagen über deren Verteilung gemacht werden sollen, die sich auf eine Anzahl von Beobachtungen dieser Zufallsvariablen stützen. Dabei wird in zahlreichen Fällen davon ausgegangen, dass die Verteilung von X bis auf einige Konstanten oder Parameter bekannt ist, mit anderen Worten, dass die Verteilung von X zu einer Familie von Verteilungen gehört.

Definition 3.1.1 *Verteilungsfamilie*

Eine *Verteilungsfamilie* ist eine indizierte Menge von Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}^k, \mathfrak{B}^k)$:

$$\mathscr{P} = \{\mathrm{P}_\vartheta | \vartheta \in \Theta\}, \text{ mit } \mathrm{P}_\vartheta \neq \mathrm{P}_{\vartheta'} \text{ für } \vartheta \neq \vartheta'.$$

$\mathscr{P}$ kann auch durch die zugehörigen k-dimensionalen Verteilungsfunktionen charakterisiert werden:

$$\mathscr{F} = \{F(x;\vartheta) | \vartheta \in \Theta\}.$$

Die Menge Θ heißt *Parameterraum*. Wir setzen $|\Theta| > 1$ voraus.

Wir sprechen von *parametrischen Modellen*, wenn Θ eine Teilmenge des $\mathbb{R}^p$ ist. Dann ist der Parameter ein p-dimensionaler Vektor. Sind die Elemente von Θ nicht durch endlich viele Zahlenangaben zu charakterisieren, so sprechen wir von einer *nichtparametrischen Verteilungsfamilie*.

Zu Verteilungsfamilien gelangt man auf verschiedene Weisen. Einmal gibt es die Möglichkeit, aus plausiblen Annahmen über den zugrundeliegenden Zufallsmechanismus die Gestalt oder den Typ der Verteilung abzuleiten. Für diskrete Verteilungen ist das z. B. beim Urnenmodell möglich. Dann können durch Grenzübergänge auch stetige Verteilungen gewonnen werden. Bei stetigen Verteilungen ist die Ableitung aus plausiblen Annahmen jedoch seltener erfolgreich.

Andere Verteilungsfamilien haben sich im Rahmen der Beschreibung von empirischen Verteilungen als geeignet erwiesen. Hier sind vor allem solche Verteilungen interessant, die geeignet flexibel sind. Damit ist gemeint, dass sie in einer Weise von Parametern abhängen, die es erlaubt, sie in einer Vielzahl von praktischen Situationen zu verwenden. Eine solche Art von Verteilungsfamilien bilden die Lage-Skalen-Familien. Sie sind zugleich sehr allgemein.

Definition 3.1.2 *Lage-Skalen-Familie*

Eine *Lage-Skalen-Familie* ist eine Familie von eindimensionalen Verteilungen, die durch Verschiebung und Standardisierung ineinander überführt werden können:

$$\mathscr{F} = \left\{ F(x;\vartheta,\tau) \,\middle|\, F\left(\frac{x-\vartheta}{\tau};\vartheta,\tau\right) = F_0(x),\ (\vartheta,\tau)\in\mathbb{R}\times\mathbb{R}^+ \right\}. \tag{3.1}$$

Analog liegt eine *Lage-Familie* bzw. eine *Skalen-Familie* vor, wenn

$$\mathscr{F} = \{F(x;\vartheta) | F(x-\vartheta;\vartheta) = F_0(x), \vartheta\in\mathbb{R}\} \quad \text{bzw.} \quad \mathscr{F} = \{F(x;\tau) | F(x/\tau) = F_0(x), \tau\in\mathbb{R}_+\}.$$

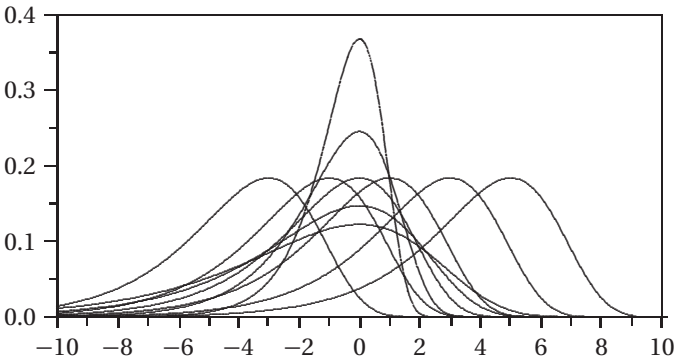


Abb. 3.1: *Dichten einer Lage-Skalen-Familie*

Der Name dieser parametrischen Verteilungsfamilien wird durch die Bedeutung der Parameter ϑ und τ erklärt: X besitze die Verteilungsfunktion $F(x)$ mit $F((x-\vartheta)/\tau) = F_0(x)$. Der Erwartungswert und die Varianz der Zufallsvariablen Z mit der Verteilungsfunktion F_0 mögen existieren. Dann gilt:

$$\mathrm{E}(X) = \vartheta + \mathrm{E}(Z), \qquad \mathrm{Var}(X) = \tau^2 \cdot \mathrm{Var}(Z).$$

ϑ charakterisiert also die Lage und τ die Streuung der Verteilung.

Gehört die Verteilung der Zufallsvariablen X zu einer Lage-Skalen-Familie, so hängt die Verteilung von $(X-\vartheta)/\tau$ nicht von den Parametern ϑ und τ ab. Für die Quantile von Lage-Skalen-Familien haben wir damit:

$$F(q_\alpha;0,1) = \alpha \quad \Leftrightarrow \quad F(\vartheta + q_\alpha\tau;\vartheta,\tau) = \alpha.$$

Beispiel 3.1.3 *Normalverteilungen als Lage-Skalen-Familie*

Die univariaten Normalverteilungen bilden eine Lage-Skalen-Familie. Für $X \sim \mathscr{N}(\mu,\sigma^2)$ gilt:

$$F_X(x) = \mathrm{P}(X\le x) = \mathrm{P}\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

wobei $\Phi(z)$ die Verteilungsfunktion der Standardnormalverteilung ist. Ist $\varphi(z)$ die zu $\Phi(z)$ gehörige Dichte, so gilt weiter

$$f_X(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right).$$

Einen weiteren Anwendungsbereich, der zu eigenen Verteilungen führt, bilden die statistischen Schätz- und Testverfahren. Hier erweisen sich vor allem abgeleitete Verteilungen als relevant. Wir kommen im Folgenden und in den weiteren Kapiteln ausführlicher darauf zurück.

Beispiel 3.1.4 *Familie der symmetrischen Verteilungen*

In der nichtparametrischen Statistik, wo man sich nicht auf spezielle Verteilungsfamilien festlegen möchte, werden u. a. symmetrische Verteilungen betrachtet. Die Familie der um μ symmetrischen Verteilungen ist

$$\mathscr{F}_s = \left\{ F \,\middle|\, F \text{ ist eine Verteilungsfunktion mit } F(\mu - x) = 1 - F(\mu + x) \right\}. \tag{3.2}$$

Zu dieser Familie gehören z. B. die Normalverteilungen mit Erwartungswert μ.

Bei festem μ und beliebigem F gibt es keinen endlich-dimensionalen Parameter, der diese Verteilungsfamilie charakterisiert.

Schließlich wurden Verteilungen auch unter dem Gesichtspunkt der Simulation kreiert.

Beispiel 3.1.5 *Ramberg-Schmeiser-Tukey-Verteilungsfamilie*

Unter Rückgriff auf einen Vorschlag von Tukey et al. (1947) propagierten Ramberg & Schmeiser (1972) eine Klasse von symmetrischen Verteilungen, die *RST-Verteilungsfamilie*. Sie sind über die Inverse der Verteilungsfunktionen definiert:

$$F^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_3}}{\lambda_2} \qquad 0 \le u \le 1. \tag{3.3}$$

Die Parameter λ_2 und λ_3 müssen dabei das gleiche Vorzeichen haben und ungleich Null sein. λ_1 ist ein Lageparameter, $\mathrm{E}(X) = \lambda_1$ für eine Zufallsvariable X mit einer $\mathscr{RST}(\lambda_1, \lambda_2, \lambda_3)$-Verteilung. λ_2 und λ_3 bestimmen die Varianz und λ_3 allein charakterisiert die Wölbung.

Über die geeignete Wahl der Parameter lassen sich sehr unterschiedliche Verteilungen recht gut approximieren. Für die Approximation der t_n- bzw. die Standardnormalverteilung sind folgende Parameterwerte geeignet ($\lambda_1 = 0$):

	t_5	t_{10}	t_{25}	t_{50}	$\mathscr{N}(0,1)$
λ_2	−0.248	0.023	0.134	0.167	0.198
λ_3	−0.136	0.015	0.089	0.112	0.135

Ramberg & Schmeiser (1974) verallgemeinerten diese Verteilungsfamilie noch, um auch nicht-symmetrische Verteilungen approximieren zu können. Da der Zugang der gleiche ist, lassen sich in jedem Fall mit der Wahrscheinlichkeitsintegraltransformation auf einfache Weise Zufallszahlen erzeugen.

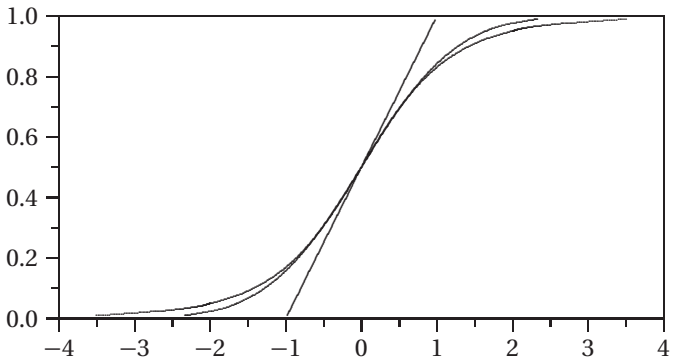


Abb. 3.2: *Einige Verteilungsfunktionen der $\mathscr{RST}$-Familie*

3.1.2 Einige univariate Verteilungen

Die hier betrachteten diskreten Verteilungen lassen sich über einen von zwei Zufallsmechanismen gewinnen. Dies sind einmal das Urnenmodell und dann der sogenannte Bernoulli-Prozess. Das Urnenmodell wurde im Abschnitt 1.1.2 besprochen.

Als *Bernoulli-Prozess* bezeichnet man die wiederholte Durchführung eines Zufallsexperimentes, wobei

- jeweils nur das Eintreten eines Ereignisses A interessiert;
- die Wahrscheinlichkeit für das Eintreten von A sich nicht ändert;
- die einzelnen Durchführungen voneinander unabhängig sind.

Diskrete Gleichverteilung

Ziehen wir zufällig eine Kugel von N gleichartigen, durchnummerierten aus einer Urne, so hat die Zufallsvariable X 'Kugelnummer' die Zähldichte

$$f(x;N)=\frac{1}{N} \qquad x=1,\ldots,N. \tag{3.4}$$

X heißt dann diskret gleich- oder rechteckverteilt, i.Z. $X\sim\mathscr{DR}(N)$.
Mit den Summenformeln

$$\sum_{i=1}^{n} i=\frac{n(n+1)}{2}, \qquad \sum_{i=1}^{n} i^2=\frac{n(n+1)(2n+1)}{6}, \qquad \sum_{i=1}^{n} i^3=\left(\frac{n(n+1)}{2}\right)^2,$$
$$\sum_{i=1}^{n} i^4=\frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

erhalten wir leicht:

$$\mathrm{E}(X)=\frac{n+1}{2},$$
$$\mathrm{Var}(X)=\frac{n(n+1)}{12}.$$

Hypergeometrische Verteilung

Gibt es in der Urne eine feste Anzahl N von gleichartigen Kugeln, von denen M auf eine bestimmte Weise markiert sind, so ergibt sich für die Anzahl X der markierten Kugeln unter n ohne Zurücklegen zufällig gezogenen eine hypergeometrische Verteilung. (Siehe EinfStat.) Die Zähldichte ist

$$f(x;N,M,n)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad x=\max\{0,n+M-N\},\ldots,\min\{n,M\}. \tag{3.5}$$

Wir schreiben $X\sim\mathscr{H}(N,M,n)$, um zu kennzeichnen, dass X hypergeometrisch mit den entsprechenden Parametern verteilt ist. Es gilt

$$\begin{aligned}\mathrm{E}(X)&=n\cdot\frac{M}{N},\\ \mathrm{Var}(X)&=n\cdot\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}.\end{aligned}$$

Binomialverteilung

Wird das eben beschriebene Ziehungsverfahren mit Zurücklegen durchgeführt, so stellt es einen Bernoulli-Prozess dar. Die Anzahl X der 'Erfolge' bei einem Bernoulli-Prozess mit n Durchführungen hat die Zähldichte der Binomialverteilung

$$f(x;n,p)=\binom{n}{x}p^x(1-p)^{n-x} \qquad x=0,1,2,\ldots,n;\quad 0<p<1. \tag{3.6}$$

(Siehe EinfStat.) Dafür schreiben wir $X\sim\mathscr{B}(n,p)$. Die $\mathscr{B}(1,p)$-Verteilung heißt auch *Bernoulli-Verteilung*.

Die Binomialverteilung hat die momenterzeugende Funktion

$$M(t)=\left(1-p+pe^t\right)^n.$$

Daraus erhalten wir:

$$\begin{aligned}\mathrm{E}(X)&=np\\ \mathrm{Var}(X)&=np(1-p).\end{aligned}$$

Weiter folgt aus den Eigenschaften der momenterzeugenden Funktion, dass die Summe unabhängiger mit dem gleichen Parameter p binomialverteilter Zufallsvariablen wieder binomialverteilt ist.

Poisson-Verteilung

Betrachten wir bei der Binomialverteilung den Grenzübergang $n\to\infty$, $p_n\cdot n=\lambda$, so gelangen wir zur Poisson-Verteilung $\mathscr{PV}(\lambda)$ mit der Zähldichte

$$f(x;\lambda)=e^{-\lambda}\cdot\frac{\lambda^x}{x!} \qquad x=0,1,2,\ldots;\quad \lambda>0. \tag{3.7}$$

Siehe dazu EinfStat. Die momenterzeugende Funktion ist nach Beispiel 2.2.3:

$$M(t) = e^{\lambda e^t - \lambda}.$$

Weiter gilt

$$\mathrm{E}(X) = \lambda,$$

$$\mathrm{Var}(X) = \lambda.$$

Der Grenzübergang, der von der Binomial- zur Poisson-Verteilung führt, legt nahe, dass letztere ein gutes Modell ist, wenn ein Ereignis nur eine geringe Wahrscheinlichkeit hat, aber viele Durchführungen des entsprechenden Zufallsexperimentes betrachtet werden. Es gibt noch eine andere Herleitung der Poisson-Verteilung, die deutlich macht, warum oft die Anzahl der Vorkommnisse pro Zeitintervall der Länge t durch eine Poisson-Verteilung mit dem Parameter $\lambda \cdot t$ modeliert wird. Diese Herleitung etabliert die Poisson-Verteilung als Basis des einfachsten Modells in der Bedienungstheorie. Dort interessiert man sich für die Verteilung der Anforderungen pro Zeitintervall. Beispiele dafür sind die Anzahl von Kunden vor einem Abfertigungsschalter und die Anzahl der ankommenden Telefongespräche in einer Telefonzentrale.

Satz 3.1.6 *Herleitung der Poisson-Verteilung*

Für jedes $t \geq 0$ sei N_t eine Zufallsvariable mit Werten in $\mathbb{N}_0$ mit den folgenden Eigenschaften:

(i) $N_0 = 0$.

(ii) $s < t \quad \Rightarrow \quad N_s$ und $N_t - N_s$ sind unabhängig.

(iii) N_s und $N_{t+s} - N_t$ sind identisch verteilt.

(iv) $\lim\limits_{t \to 0} \dfrac{\mathrm{P}(N_t = 1)}{t} = \lambda.$

(v) $\lim\limits_{t \to 0} \dfrac{\mathrm{P}(N_t > 1)}{t} = 0.$

Dann gilt für jedes $n \in \mathbb{N}_0$:

$$\mathrm{P}(N_t = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

Mit anderen Worten hat N_t eine Poisson-Verteilung mit dem Parameter $\lambda \cdot t$.

Beweis: Siehe Feller (1968).

Interpretieren wir die Zufallsvariable N_t des Satzes als Anzahl der ankommenden Forderungen, so lassen sich die Eigenschaften folgendermaßen deuten:
(i) Zunächst wird ohne jede Ankunft gestartet. (ii) Ankünfte in disjunkten Zeitintervallen sind unabhängig. (iii) Die Anzahl der Ankünfte hängt nur von der Länge des Zeitintervalls ab. (iv) Für kleine Intervalle ist die Wahrscheinlichkeit für eine Ankunft proportional zu ihrer Länge. (v) Es können nicht mehrere Ankünfte gleichzeitig vorkommen.

Geometrische Verteilung

Betrachten wir bei einem Bernoulli-Prozess die Anzahl X der Versuche, die vor dem erstmaligen Eintreten des interessierenden Ereignisses durchgeführt werden, so hat X eine geometrische Verteilung. Sie ist ein Spezialfall der unten angegebenen negativen Binomialverteilung. Daher schreiben wir $X \sim \mathcal{NB}(1,p)$.

$$f(x;p) = p(1-p)^x \qquad x = 0,1,2,3,\ldots; \quad 0 < p < 1. \tag{3.8}$$

Es gilt:

$$F(x;p) = 1-(1-p)^{x+1} \quad \text{für} \quad x = 0,1,2,3,\ldots,$$

$$M(t) = \frac{p}{1-(1-p)e^t},$$

$$\mathrm{E}(X) = \frac{1-p}{p},$$

$$\mathrm{Var}(X) = \frac{1-p}{p^2}.$$

Die geometrische Verteilung ist eine diskrete Wartezeitverteilung. Sie hat kein Gedächtnis, d.h.:

$$\mathrm{P}(X = x+y \,|\, X \geq y) = \frac{p(1-p)^{x+y}}{(1-p)^y} = \mathrm{P}(X = x).$$

Wenn also ein Ereignis eingetreten ist, ist die Verteilung der Anzahl der Versuche bis zum nächsten Erfolg bei einer neuen Versuchsserie unverändert.

Beispiel 3.1.7 *Zombies*

Die (diskrete) Wartezeit, bis der nächste Kunde eine Bank betritt, wurde durch folgendes Modell simuliert. Aus naheliegenden Gründen griff der Autor dabei auf Zombies zurück.

Genau im zeitlichen Abstand von einer Sekunde wird jeweils ein Zombie losgeschickt. Vor dem Eingang zur Bank befindet sich eine Mauer von einer Längeneinheit. In dieser ist eine Lücke der Breite p, $(0 < p < 1)$. Die Zombies treffen nun mit gleicher Wahrscheinlichkeit auf jeden Mauerabschnitt. Wenn ein Zombie auf die Lücke trifft, gelangt er durch das Hindernis und in die Bank. Es ist offensichtlich, dass die Zahl der Sekunden zwischen dem Eintreffen zweier Zombies in der Bank einer geometrischen Verteilung folgt.

Negative Binomialverteilung

Wartet man bei einem Bernoulli-Prozess bis zum Eintreten von k Ereignissen, so hat die Anzahl X der Versuche vor dem entscheidenden Erfolg die Zähldichte

$$f(x) = \binom{k-1+x}{x} p^k (1-p)^x \qquad x = 0,1,2,\ldots \quad 0 < p < 1,\ k > 0. \tag{3.9}$$

X heißt dann negativ binomialverteilt, i.Z. $X \sim \mathcal{NB}(k,p)$.

Erwartungswert und Varianz sind

$$\mathrm{E}(X) = \frac{k(1-p)}{p},$$
$$\mathrm{Var}(X) = \frac{k(1-p)}{p^2}.$$

Stetige Gleichverteilung

Die Gleich- oder Rechteckverteilung $\mathscr{R}(a,b)$ über dem Intervall (a,b) hat die Dichte

$$f(x;a,b) = \frac{1}{b-a} \qquad a < x < b. \tag{3.10}$$

Ihre Bedeutung beruht wesentlich auf der Wahrscheinlichkeitsintegraltransformation, dem Satz 1.3.2. Hier gilt:

$$\mathrm{E}(X) = \frac{a+b}{2},$$
$$\mathrm{Var}(X) = \frac{(b-a)^2}{12},$$
$$M(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}.$$

Normalverteilung

Die Poisson-Verteilung ergibt sich aus der Binomialverteilung bei einem speziellen Grenzübergang, bei dem sich p mit n ändert. Wir können auch p erst einmal festhalten und nur n gegen unendlich streben lassen. Das ist natürlicher, weil es die Frage der Berechnung von Binomialwahrscheinlichkeiten bei großen n berührt. Bereits 1718 wurde das entsprechende Resultat erstmalig veröffentlicht. Obwohl wir es als Spezialfall des zentralen Grenzwertes im Kapitel 5 erhalten, wollen wir die Approximation hier zumindest zitieren.

Satz 3.1.8 *Grenzwertsatz von DeMoivre-Laplace*

Sei $0 < p < 1$ fest und sei

$$x_{nk} = \frac{k - np}{\sqrt{np(1-p)}}, \qquad 0 \le k \le n.$$

Ist dann $A > 0$ eine beliebige Konstante, so gilt für alle k mit $|x_{nk}| < A$:

$$\binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{\sqrt{2\pi np(1-p)}} \cdot \exp\left[-\frac{x_{nk}^2}{2}\right].$$

Beweisskizze: Für die zugelassenen $k = np + \sqrt{np(1-p)} \cdot x_{nk}$ gilt

$$k \sim np, \qquad n - k \sim n(1-p).$$

Damit kann die Stirlingsche Formel auf den Binomialausdruck angewendet werden:

$$
\begin{aligned}
&\binom{n}{k} p^k (1-p)^{n-k} \\
&\quad \approx \frac{\sqrt{2\pi n}\cdot n^n e^n}{\sqrt{2\pi k}\, k^k e^k \sqrt{2\pi(n-k)}(n-k)^{n-k} e^{n-k}} p^k (1-p)^{n-k} \\
&\quad = \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} \\
&\quad \approx \frac{1}{\sqrt{2\pi np(1-p)}} \cdot \left(1 - \frac{\sqrt{np(1-p)}\cdot x_{nk}}{k}\right)^k \left(1 + \frac{\sqrt{np(1-p)}\cdot x_{nk}}{n-k}\right)^{n-k}.
\end{aligned}
$$

Logarithmieren des zweiten und dritten Faktors, Entwickeln in eine Taylorreihe und Berücksichtigen nur der relevanten Glieder führt dann zu der behaupteten Relation. Dies wird bei Feller (1968) und bei Chung (1974) ausgeführt.

Der Satz sagt nichts anderes, als dass sich die Binomialverteilung für große n durch die Normalverteilung approximieren lässt.

Die Normalverteilung hat die Dichte

$$
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}\right] \qquad -\infty < x < \infty. \tag{3.11}
$$

Wie erwähnt bilden die $\mathcal{N}(\mu,\sigma^2)$-Verteilungen eine Lage-Skalen-Familie. Daher reicht es, die nicht in geschlossener Form angebbare Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung $\mathcal{N}(0,1)$ zu tabellieren.

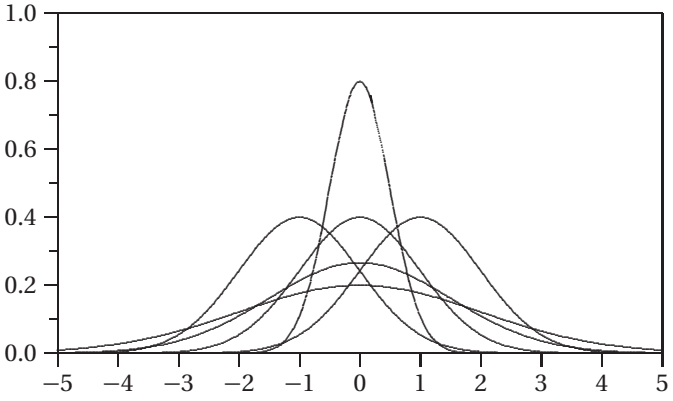


Abb. 3.3: *Normalverteilungsdichten*

Für eine $\mathcal{N}(\mu,\sigma^2)$-verteilte Zufallsvariable gilt:

$$
\begin{aligned}
\mathrm{E}(X) &= \mu, \\
\mathrm{Var}(X) &= \sigma^2.
\end{aligned}
$$

Die momenterzeugende Funktion ist:

$$M(t) = \exp\left[\mu t - \frac{\sigma^2 t^2}{2}\right]. \tag{3.12}$$

Alle ungeraden Momente um μ sind gleich Null:

$$\mathrm{E}((X-\mu)^{2k+1}) = 0 \qquad k = 0, 1, \ldots$$

Für die geraden Momente gilt:

$$\mathrm{E}((X-\mu)^{2k}) = \frac{(2\sigma^2)^k}{\sqrt{\pi}}\Gamma\left(\frac{2k+1}{2}\right).$$

Es ist $\beta_1 = 0$, die Normalverteilung ist symmetrisch. Zudem folgt speziell $\beta_2 = 3$.

Beispiel 3.1.9 *Körpergröße von Wehrpflichtigen*

Die Normalverteilung wird häufig zur Beschreibung von Beobachtungen verwendet, bei denen natürliche Größen gemessen werden. Ein Beispiel dafür bilden etwa die Körpergrößen von 220653 im Jahr 1990 erstuntersuchten Wehrpflichtigen des Geburtsjahrganges 1971 in der Bundesrepublik (IWB, 1994), siehe die Abbildung 3.4.

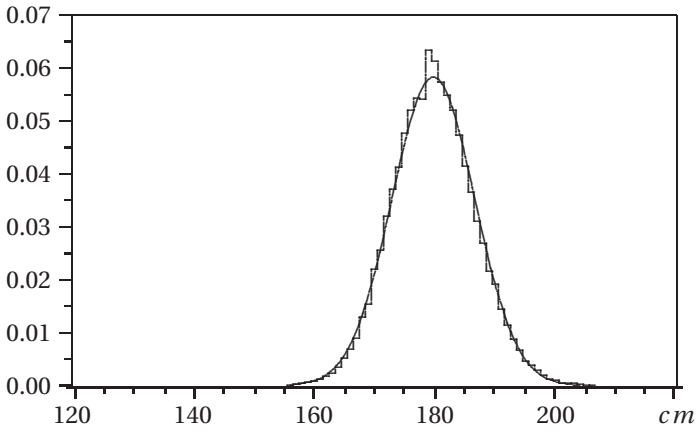


Abb. 3.4: *Körpergrößen von Wehrpflichtigen*

Eine einfache Weise, normalverteilte Zufallszahlen zu erzeugen, ist die *Box-Muller-Methode*. Sie basiert auf der Entdeckung, dass zwei unabhängige rechteckverteilte Zufallsvariablen in zwei unabhängige standardnormalverteilte Zufallsvariablen transformiert werden können.

Lemma 3.1.10 *Box-Muller-Methode*

U_1 und U_2 seien unabhängig $\mathscr{R}(0,1)$-verteilt. Dann sind

$$X_1 = \sqrt{-2\ln(U_1)}\cdot\cos(2\pi U_2) \quad \text{und} \quad X_2 = \sqrt{-2\ln(U_1)}\cdot\sin(2\pi U_2)$$

unabhängige $\mathcal{N}(0,1)$-verteilte Zufallsvariablen.

Beweis: Es ist leicht zu sehen, dass $V=\sqrt{-2\ln(U_1)}$ die Dichte

$$f_V(v)=v\cdot\exp\left[-\frac{v^2}{2}\right]\qquad v>0$$

hat. Wir betrachten daher die Transformation $g(V,U_2)=(X_1,X_2)$. Die inverse Transformation $h(X_1,X_2)=g^{-1}(X_1,X_2)$ ist gegeben durch:

$$V=\sqrt{X_1^2+X_2^2}$$
$$U_2=\frac{1}{2\pi}\tan^{-1}\left(\frac{X_1}{X_2}\right).$$

Die Jacobi-Determinante erhalten wir mit $\partial\tan^{-1}(x)/\partial x=1/(1+x^2)$ zu:

$$\boldsymbol{J}=\left|\det\begin{pmatrix}\dfrac{x_1}{\sqrt{x_1^2+x_2^2}} & \dfrac{x_2}{\sqrt{x_1^2+x_2^2}}\\ \dfrac{1}{2\pi}\cdot\dfrac{1/x_2}{1+x_1^2/x_2^2} & \dfrac{1}{2\pi}\cdot\dfrac{-x_1/x_2^2}{1+x_1^2/x_2^2}\end{pmatrix}\right|=\frac{1}{2\pi\sqrt{x_1^2+x_2^2}}.$$

Nach Satz 1.3.9 ist also

$$f_{X_1}(x_1,x_2)=f_V\left(\sqrt{x_1^2+x_2^2}\right)\cdot 1\cdot\frac{1}{2\pi\sqrt{x_1^2+x_2^2}}=\frac{1}{2\pi}\exp\left[-\frac{x_1^2+x_2^2}{2}\right]$$
$$=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{x_1^2}{2}\right]\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{x_2^2}{2}\right].$$

Lognormalverteilung

Ist X eine Zufallsvariable, so dass $Y=\ln(X)$ normalverteilt ist, so nennt man X logarithmisch normalverteilt, $X\sim\mathcal{LN}(\mu_L,\sigma_L^2)$. Die Dichte ist

$$f(x;\mu_L,\sigma_L^2)=\frac{1}{\sqrt{2\pi}\sigma_L x}\cdot\exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\mu_L}{\sigma_L}\right)^2\right]\quad\text{für}\quad x>0. \tag{3.13}$$

Die Parameter μ_L und σ_L^2 sind hier nicht gleich dem Erwartungswert und der Varianz. Über die Beziehung zur Normalverteilung, $X=\ln(Y)$, erhalten wir aber sofort die Momente

$$\mathrm{E}(X^r)=\mathrm{E}(\exp(r\cdot\ln X))=\mathrm{E}(\exp(r\cdot Y))=\exp\left[r\mu_L+\frac{r^2\sigma_L^2}{2}\right].$$

Damit folgt:

$$\mathrm{E}(X)=\exp\left[\mu_L+\frac{\sigma_L^2}{2}\right],$$
$$\mathrm{Var}(X)=\exp[2\mu_L+\sigma_L^2]\left(\exp[\sigma_L^2]-1\right),$$
$$\beta_1=\sqrt{e^{\sigma_L^2}-1}\cdot\left(e^{\sigma_L^2}+2\right).$$

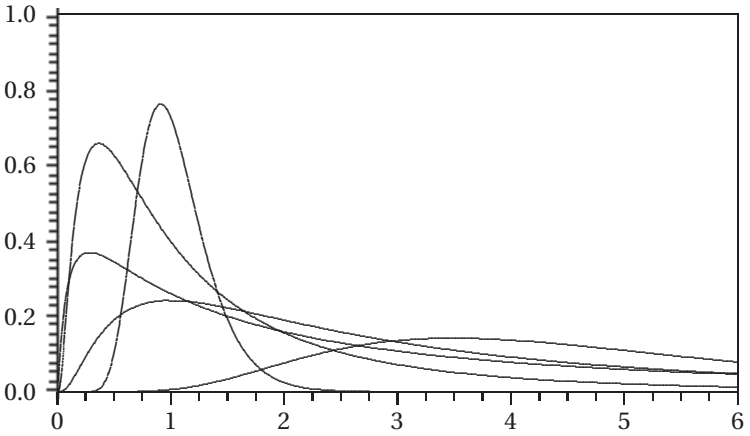


Abb. 3.5: *Dichten der Lognormalverteilung*

Exponentialverteilung

Verkürzen wir in dem Zombie-Beispiel 3.1.7 zur geometrischen Verteilung die zeitlichen Abstände und machen proportional dazu die Mauerlücke schmäler, so gelangen wir zur Exponentialverteilung $\mathcal{EX}(\lambda)$ mit der Dichte

$$f(x;\lambda) = \lambda e^{-\lambda x} \quad \text{für} \quad x > 0,\ \lambda > 0. \tag{3.14}$$

Dieser Übergang (vgl. EinfStat) macht einsichtig, dass die Exponentialverteilung eine große Rolle als Wartezeitverteilung spielt. Wir haben hier

$$\begin{aligned} \mathrm{E}(X) &= \frac{1}{\lambda}, \\ \mathrm{Var}(X) &= \frac{1}{\lambda^2}, \\ M(t) &= \frac{\lambda}{\lambda - t} \quad \text{für} \quad t < \lambda. \end{aligned}$$

Die Familie $\mathcal{EX}(\lambda)$ der Exponentialverteilungen bildet eine Skalen-Familie. Mit der Standardexponentialverteilung $\mathcal{EX}(1)$ mit der Verteilungsfunktion

$$F_0(x) = F(x;1) = \begin{cases} 0 & \text{für } x < 0 \\ 1 - e^{-x} & \text{für } x \geq 0 \end{cases}$$

gilt für die Exponentialverteilung mit dem Parameter λ:

$$F(x;\lambda) = F_0(\lambda \cdot x).$$

Folglich ist hier $\tau = 1/\lambda$ der Skalenparameter.

Durch Einführung eines Schwellenparameters ϑ gelangen wir zu der Lage-Skalen-Familie der verschobenen Exponentialverteilungen:

$$F(x;\vartheta,\lambda)=\begin{cases} 0 & \text{für } x<\vartheta \\ 1-e^{-\lambda(x-\vartheta)} & \text{für } x\geq\vartheta \end{cases}.$$

Zur Charakterisierung von Lebensdauerverteilungen wird weniger die Verteilungsfunktion $F(x)$ als die *Survivorfunktion* $S(t)=1-F(x)$ verwendet. Daneben spielt die *Hazardrate* die herausragende Rolle. Sie gibt an, wie sich die Wahrscheinlichkeit ändert, dass die Lebensdauer in einem kleinen Zeitintervall um x endet, unter der Voraussetzung, dass sie bis zum Zeitpunkt x andauert:

$$h(x)=\frac{f(x)}{S(x)}=\frac{f(x)}{1-F(x)}. \tag{3.15}$$

Für die Exponentialverteilung sehen wir leicht, dass die Hazardrate konstant ist:

$$h(x)=\lambda.$$

Gammaverteilung

Die Gammaverteilung $\mathcal{G}(\beta,\lambda)$ erhalten wir aus der negativen Binomialverteilung auf die gleiche Weise wie die Exponentialverteilung aus der geometrischen. Auch sie wird gerne zur Modellierung von Wartezeiten verwendet. Ihre Dichte hat die Gestalt

$$f(x;\beta,\lambda)=\frac{\lambda^\beta}{\Gamma(\beta)}\cdot x^{\beta-1}\cdot e^{-\lambda x} \quad \text{für} \quad x>0,\ \lambda,\beta>0. \tag{3.16}$$

Dabei ist die *Gammafunktion* $\Gamma(u)$ für $u>0$ definiert durch

$$\Gamma(u)=\int_0^\infty t^{u-1}e^{-t}\,\mathrm{d}t\,. \tag{3.17}$$

Es gilt $\Gamma(1)=1$, $\Gamma(u+1)=u\cdot\Gamma(u)$ für $u>0$ und $\Gamma(u+1)=u!$ für $u\in\mathbb{N}_0$.

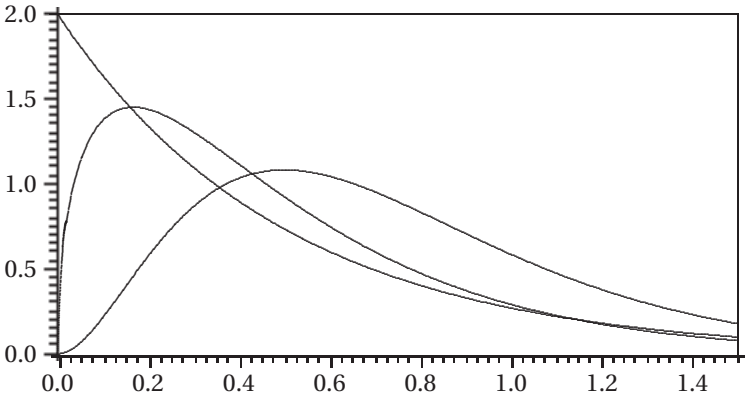


Abb. 3.6: *Einige Dichten der Gamma-Verteilung*

Die Exponentialverteilungen bilden eine Teilfamilie der Gammaverteilungen. Es ist lediglich $\beta = 1$ zu setzen.

Die momenterzeugende Funktion der Standard-Gammaverteilung mit $\lambda = 1$ lautet:

$$M(t) = \frac{1}{(1-t)^{\beta}}.$$

Damit gilt für die $\mathcal{G}(\beta, 1)$-Verteilung:

$$\mathrm{E}(X) = \mathrm{Var}(X) = \beta.$$

Zudem sind

$$\beta_1 = 2\beta^{-\frac{1}{2}},$$
$$\beta_2 = 3 + 6\beta^{-1}.$$

Weibull-Verteilung

Bei Lebensdaueranalysen erweisen sich empirische Hazardraten oft als nicht konstant. Über eine einfache Potenztransformation erhalten wir aber aus der Exponentialverteilung die Familie der Weibull-Verteilungen $\mathcal{W}(\lambda, \beta)$, die sowohl fallende als auch wachsende Verläufe der Hazardrate erfassen kann. Ihre Dichte ist

$$f(x; \lambda, \beta) = \lambda \cdot \beta \cdot x^{\beta-1} \cdot \exp\left[-\lambda x^{\beta}\right] \quad \text{für} \quad \text{für} \quad x > 0,\ \lambda > 0,\ \beta > 0. \tag{3.18}$$

Bei Lebensdaueranalysen wird häufig die sogenannte *Survivorfunktion* $S(t) = 1 - F(t)$ anstelle der Verteilungsfunktion betrachtet:

$$S(x) = 1 - F(x) = \exp\left[-\lambda x^{\beta}\right] \quad \text{für} \quad x > 0.$$

Dazu gehört die Hazardfunktion

$$h(x) = \lambda \cdot \beta \cdot x^{\beta-1}.$$

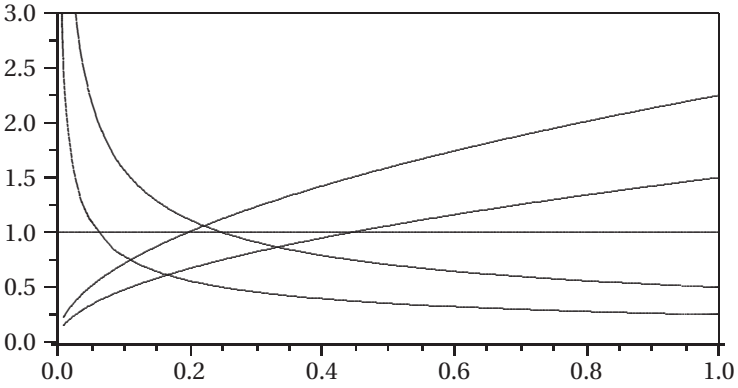


Abb. 3.7: *Einige Hazardfunktionen der Weibull-Verteilung*

$h(x)$ ist monoton wachsend für $\beta > 1$, konstant für $\beta = 1$ und monoton fallend für $\beta < 1$.

Die logarithmische Transformation überführt die Weibull-Verteilung in eine *Gumbel-* oder *Extremwertverteilung*. Die Dichte ist, wie leicht mit dem Transformationssatz 1.3.5 für Dichten nachvollzogen werden kann:

$$f(t;\vartheta,\tau) = \frac{1}{\tau}\exp\left[\frac{x-\vartheta}{\tau} - \exp\left(\frac{x-\vartheta}{\tau}\right)\right]. \tag{3.19}$$

Dabei wurde eine geeignete Reparametrisierung vorgenommen, um den Charakter als Lage-Skalen-Familie deutlicher hervorzuheben. Die Gestalt der Dichten ist aus Abb. 3.1 zu ersehen.

Logistische Verteilung

Die logistische Verteilung liefert ein Beispiel für den Ansatz, aus einfachen Annahmen zu einer Verteilungsfamilie zu gelangen.

Beispiel 3.1.11 *Ausbreitung von Gerüchten*

Wir betrachten die Ausbreitung eines Gerüchtes in einer abgeschlossenen Population. Mit $F(x)$ sei der Anteil derjenigen bezeichnet, die zum Zeitpunkt x schon informiert sind. Die Weitergabe gehe über persönliche Kontakte vonstatten, wobei sich Informierte und Uninformierte zufällig treffen. Der Zuwachs des Anteils von Informierten sei proportional zu dem Treffen von Informierten und Uninformierten, formal:

$$\frac{dF(x)}{dx} = c \cdot F(x)(1-F(x)).$$

Am Beginn ist der Anstieg relativ schwach, da $F(x)$ klein ist. Dann nimmt der Anteil der Informierten immer rascher zu, bis er wieder abklingt, weil nicht mehr so viele Personen uninformiert sind.

Die Lösung dieser Differentialgleichung ist die Verteilungsfunktion der *logistischen Verteilung*:

$$F(x;a,b) = \frac{1}{1+e^{-b(x-a)}}. \tag{3.20}$$

Bei den gemachten Voraussetzen sind die möglichen Verteilungen auf solche vom logistischen Typ beschränkt. Aussagen sind aber noch über den Parameter $\vartheta = (a,b) \in \mathbb{R}\times\mathbb{R}^+ = \Theta$ zu machen. Erst durch ihn wird die endgültige Verteilung festgelegt.

Die logistischen Verteilungen lassen sich jeweils ineinander überführen: Die Verteilungsfunktion von $Y = bX - a$ ist

$$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(bX - a \le y) = \mathrm{P}\left(X \le \frac{y+a}{b}\right) = F\left(\frac{y+a}{b};a,b\right) = \frac{1}{1+e^{-y}}.$$

Das ist die standardisierte logistische Verteilung mit den Parametern $(0,1)$.

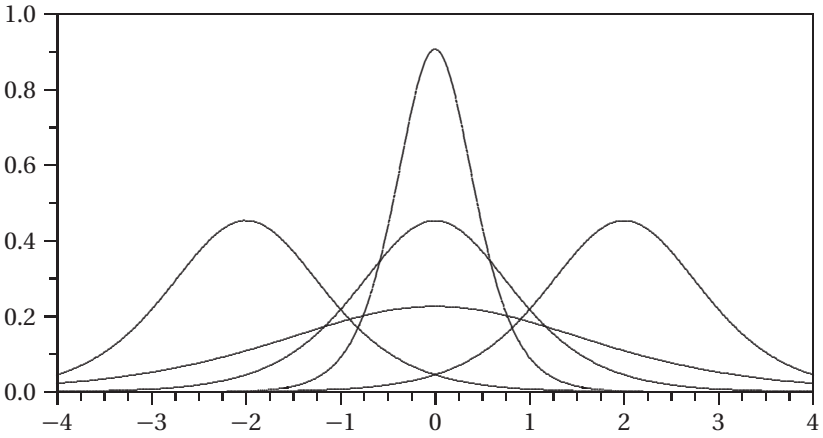


Abb. 3.8: *Einige Dichten der logistischen Verteilung*

Laplace-Verteilung

Die $\mathscr{L}(\mu,\lambda)$-Verteilungen sind gegeben durch ihre Dichten

$$f(x)=\frac{\lambda}{2}\exp[-\lambda|x-\mu|] \qquad -\infty<x<\infty. \tag{3.21}$$

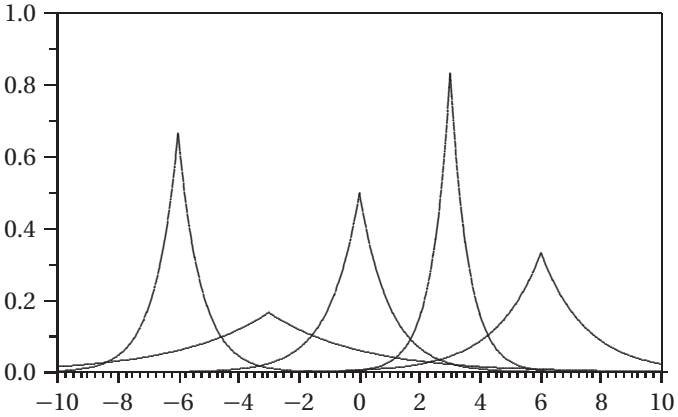


Abb. 3.9: *Einige Dichten von Laplace-Verteilungen*

Die Laplace-Verteilungen bilden eine Lage-Skalen-Familie mit

$$\begin{aligned}
\mathrm{E}(X)&=\mu,\\
\mathrm{Var}(X)&=\frac{2}{\lambda^2},\\
M(t)&=\frac{e^{\mu t}}{1-(t/\lambda)^2}.
\end{aligned}$$

Die Laplace-Verteilungen mit $\mu = 0$ ergeben sich als Differenz zweier unabhängiger exponentialverteilter Zufallsvariablen.

Beispiel 3.1.12 *Wasserstandsdifferenzen*

Bain & Engelhardt (1973) schlugen eine Laplace-Verteilung als Modell für die Differenzen der (allerdings nicht unabhängigen) Wasserstände an zwei Orten am Fox-River in den USA vor. Die Gegenüberstellung von empirischer und theoretischer Verteilungsfunktion zeigt, dass die Laplace-Verteilung ein vernünftiges Modell zu sein scheint.

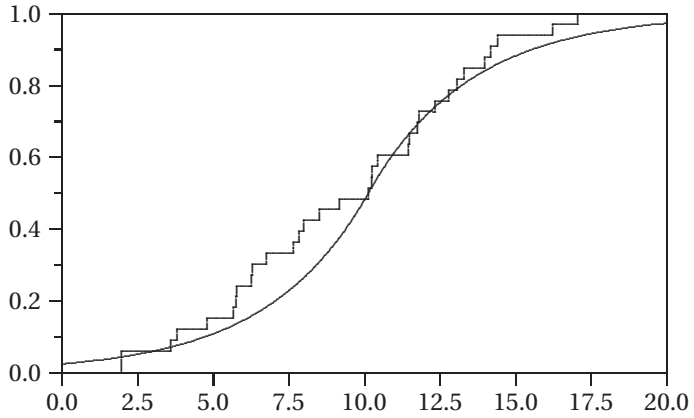


Abb. 3.10: *Gegenüberstellung von empirischer und theoretischer Verteilungsfunktion*

Cauchy-Verteilung

Die $\mathscr{C}(\mu,\lambda)$-Verteilung hat die Dichte

$$f(x;\mu,\lambda) = \frac{1}{\lambda\pi}\frac{1}{1+((x-\mu)/\lambda)^2} \qquad -\infty < x < \infty. \tag{3.22}$$

Diese Dichte hat eine ähnliche Gestalt wie die Normalverteilung, vgl Abb. 3.11. Dennoch gibt es einen fundamentalen Unterschied. Die Cauchy-Verteilung besitzt keine Momente. Genauer divergiert das Integral $\int_{-\infty}^{\infty} |x| f(x;\mu,\lambda)\,\mathrm{d}x = \infty$.

Wegen dieser Nichtexistenz des Erwartungswertes dient sie oft als Beispiel für kuriose Erscheinungen und als Modell für eine Verteilung mit starken Flanken, die sehr viele extreme Beobachtungen hervorbringt. Die Genese als Verteilung des Quotienten zweier unabhängiger, standardnormalverteilter Zufallsvariablen zeigt aber, dass sie auch als Modell für gewisse Phänomene, speziell solcher, wo Quotienten betrachtet werden, geeignet sein kann.

Betaverteilung

Die $\mathscr{BE}(p,q)$-Verteilung hat die über $(0,1)$ definierte Dichte

$$f(x;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1} \qquad 0 < x < 1,\ p > 0,\ q > 0. \tag{3.23}$$

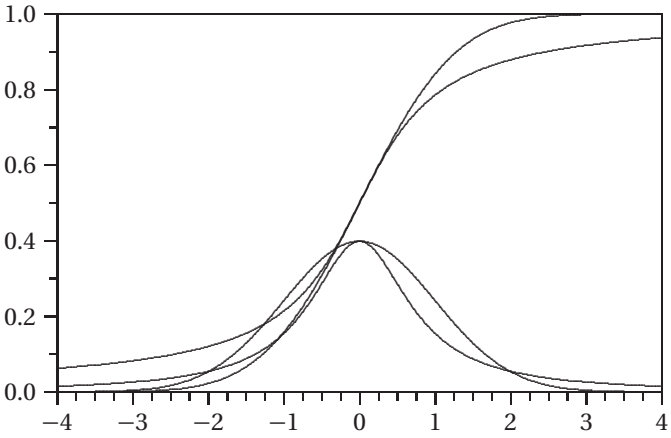


Abb. 3.11: *Dichte und VF von* $\mathcal{N}(0,1)$ *und* $\mathcal{C}(0,\sqrt{2/\pi})$

Ihr Erwartungswert und ihre Varianz sind:

$$\mathrm{E}(X) = \frac{p}{p+q},$$
$$\mathrm{Var}(X) = \frac{pq}{(p+q)^2(1+p+q)}.$$

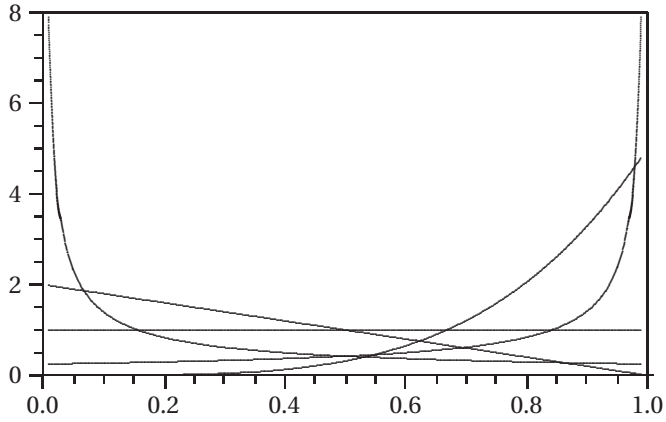


Abb. 3.12: *Einige Dichten von Betaverteilungen*

Chiquadrat-, t- und F-Verteilung

Der Spezialfall der $\mathcal{G}(1/2, n/2)$-Verteilung hat für viele Methoden der statistischen Inferenz eine große Bedeutung und wird als *Chiquadrat-Verteilung* mit n Freiheitsgraden, kurz als χ^2_n-Verteilung, bezeichnet. Ihre Dichte ist

$$f(x;n) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n-2)/2} \cdot e^{-x/2} \quad \text{für} \quad x > 0. \tag{3.24}$$

Es gilt:

$$\begin{aligned} &\mathrm{E}(X) = n, \\ &\mathrm{Var}(X) = 2n, \\ &M(t) = \frac{1}{(1-2t)^{n/2}} \quad \text{für} \quad t < 1/2. \end{aligned}$$

Aus der momenterzeugenden Funktion ist unmittelbar zu ersehen, dass die Summe unabhängiger chiquadrat-verteilter Zufallsvariablen wieder chiquadrat-verteilt ist.

Die $\mathscr{T}_n$-*Verteilung* erhalten wir als Verteilung des Quotienten zweier unabhängiger Zufallsvariablen, wobei der Zähler $\mathscr{N}(0,1)$- und der Nenner Chiquadrat-verteilt ist. Sie spielt eine große Rolle bei den statistischen Methoden, weniger als Modell zur Beschreibung von Datensätzen oder zur Modellierung von Zufallsvorgängen. Ihre Dichte ist

$$f(x;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}. \tag{3.25}$$

Wie die $\mathscr{T}_n$-Verteilung wird die $\mathscr{F}_n(n,m)$-*Verteilung* mit der Dichte

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{n}{m}\right)^{n/2} x^{n/2-1} \left(1 + \frac{n}{m}x\right)^{-(n+m)/2} \qquad x > 0 \tag{3.26}$$

im Rahmen von Methoden der statistischen Inferenz bedeutsam. Sie ist die Verteilung des mit einem Faktor versehenen Quotienten zweier unabhängiger χ^2-verteilter Zufallsvariablen.

3.1.3 Multivariate Verteilungen

Multivariate Verteilungen erhalten wir einmal durch geeignete Verallgemeinerungen von univariaten Verteilungen. Zudem können auch univariate Verteilungen so kombiniert werden, dass die Multivariaten diese als Randverteilungen haben und zusätzlich eine geeignete Abhängigkeitsstruktur aufweisen. Ein dritter Ansatz besteht in der Umsetzung der Forderung, dass sie spezielle strukturelle Eigenschaften haben sollen. Dabei werden diese Eigenschaften allerdings i.d.R. durch bereits bekannte multivariate Verteilungen nahegelegt.

Beispiel 3.1.13 *Multinomialverteilung*

Die Binomialverteilung resultiert aus der Modellvorstellung, dass ein Versuch n mal durchgeführt wird und jeweils das Eintreten eines Ereignisses A, das die Wahrscheinlichkeit p besitzt, registriert wird. Die Anzahl X der Versuchsdurchführungen, bei denen A eintritt, ist dann $\mathscr{B}(n,p)$-verteilt. Zerlegen wir nun den Stichprobenraum in drei paarweise disjunkte Ereignisse A_1, A_2 und A_3 mit den Wahrscheinlichkeiten p_1, p_2 und $1-p_1-p_2$, so hat die bivariate Zufallsvariable (X_1, X_2), die die Anzahlen der Versuche angibt, bei denen das Ereignis A_1 bzw. A_2 eintritt, eine Trinomialverteilung. Über die Verallgemeinerung auf $k+1$ Ereignisse erhalten wir die *Multinomialverteilung*.

$\boldsymbol{X} = (X_1, \ldots, X_k)$ heißt multinomialverteilt, i.Z. $\boldsymbol{X} \sim \mathscr{M}(n, \boldsymbol{p})$, $\boldsymbol{p} = (p_1, \ldots, p_k)$, falls

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}) = \frac{n!}{x_1! \cdot \ldots \cdot x_k! \cdot (n - x_1 - \ldots - x_k)!} \cdot p_1^{x_1} \cdot \ldots \cdot p_k^{x_k} (1 - p_1 - \ldots - p_k)^{n - x_1 - \ldots - x_k}, \quad (3.27)$$

wobei $x_i \geq 0$ und $\sum_{i=1}^k x_i \leq n$. Der Parameter $\boldsymbol{p}$ liegt in $\Theta = \left\{ \boldsymbol{p} \,|\, \boldsymbol{p} \in [0,1]^k, \sum_{i=1}^k p_i \leq 1 \right\}$.

Für die Komponenten X_i einer $\mathscr{M}(n, \boldsymbol{p})$-verteilten Zufallsvariablen $\boldsymbol{X}$ gilt:

$$X_i \sim \mathscr{B}(n, p_i),$$

so dass speziell $\mathrm{E}(X_i) = n p_i$ und $\mathrm{Var}(X_i) = n p_i (1 - p_i)$. Weiter ist für zwei Komponenten X_i und X_j mit $i \neq j$:

$$\mathrm{Cov}(X_i, X_j) = -n \cdot p_i p_j.$$

Die bedingte Verteilung von $\boldsymbol{X}_1 = (X_1, \ldots, X_r)$, gegeben dass $\boldsymbol{X}_2 = (X_{r+1}, \ldots, X_k)$ den Wert $(x_{r+1}, \ldots, x_k)$ annimmt, ist wieder eine Multinomialverteilung, und zwar eine mit den Parametern $n^* = n - (x_{r+1} + \cdots + x_k)$ und $p_i^* = p_i / (1 - p_{r+1} - \cdots - p_k) \quad (i = 1, \ldots, r)$.

Beispiel 3.1.14 *Lebenszeiten zweier abhängiger Komponenten*

Die Exponentialverteilung spielt eine zentrale Rolle im Bereich der statistischen Zuverlässigkeitstheorie. Bei der Betrachtung der Lebenszeiten zweier abhängiger Komponenten gelangt man aufgrund folgender Modellvorstellungen zu einer bivariaten Exponentialverteilung.

Es wird angenommen, dass drei unabhängige Quellen von Schocks vorhanden sind. Die erste Quelle zerstört die Komponente 1, die zweite Komponente 2 und die dritte beide Komponenten. Die Schocks der drei Komponenten kommen zu den Zeiten T_1, T_2 bzw. T_3 vor, wobei die Verteilungen der T_i jeweils Exponentialverteilungen sind:

$$\mathrm{P}(T_i > t) = e^{-\lambda_i t} \qquad t > 0.$$

Für die Verteilung der Lebensdauern X_1 und X_2 der beiden Komponenten gilt dann:

$$X_1 = \min(T_1, T_3), \qquad X_2 = \min(T_2, T_3).$$

Somit gilt für die gemeinsame Verteilung von X_1 und X_2:

$$\mathrm{P}(X_1 > x_1, X_2 > x_2) = \exp[-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_3 \max\{x_1, x_2\}] \qquad x_1, x_2 > 0.$$

Mit Verteilungsfunktionen formuliert lautet das:

$$F_{X_1, X_2}(x_1, x_2) = 1 - \exp[-\lambda_1 x_1] - \exp[-\lambda_2 x_2] + \exp[-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_3 \max\{x_1, x_2\}], \quad x_1, x_2 > 0.$$

In anderen Fällen ist es schwieriger, univariate Verteilungen unter dem Aspekt einer geeigneten Form von Abhängigkeit zu bivariaten zusammenzusetzen. Hier sei daher nur der Vorschlag von Farlie und Morgenstern erwähnt.

Definition 3.1.15 *Farlie-Morgenstern-Familie*

Eine *Farlie-Morgenstern-Familie* ergibt sich aus zwei univariaten Verteilungsfunktionen $F(x)$ und $G(y)$ durch Setzen von

$$H(x,y) = F(x) \cdot G(y) \cdot \{1 + \alpha \cdot [1 - F(x)] \cdot [1 - G(y)]\}. \tag{3.28}$$

Für den Abhängigkeitsparameter gilt dabei $|\alpha| < 1$.

Es ist offensichtlich, dass bei $\alpha = 0$ die Randverteilungen unabhängig sind. Die Randverteilungen stimmen auch stets mit den zur Konstruktion verwendeten überein:

$$H(x,\infty) = F(x), \qquad H(\infty,y) = G(y).$$

Beispiel 3.1.16 *eine spezielle Farlie-Morgenstern-Familie*

Seien die Randverteilungen Exponentialverteilungen mit den Parametern λ und ϑ. Dann gilt für die zugehörige Farlie-Morgenstern-Familie:

$$H(x,y) = (1 - e^{-\lambda x})(1 - e^{-\vartheta y})(1 + \alpha \cdot e^{-\lambda x} \cdot e^{-\vartheta y}) = (1 - e^{-\lambda x})(1 - e^{-\vartheta y})(1 + \alpha e^{-(\lambda x + \vartheta y)}).$$

Dabei muss natürlich $x, y > 0$ sein. Andernfalls ist $H(x,y) = 0$.

3.1.4 Die multivariate Normalverteilung

Im Beispiel 1.2.26 wurde bereits die Dichte der bivariaten Normalverteilung angegeben. Sie hat die Form

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2\cdot(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\cdot\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right]. \tag{3.29}$$

Die Dichte ist durch die 5 Parameter $\mu_1, \mu_2, \sigma_1, \sigma_2$ und ρ festgelegt. Dabei steuert der Parameter ρ die Form des Zusammenhanges. Dieser ist ja bei der bivariaten Normalverteilung extrem durchsichtig: Abhängigkeit bedeutet stets lineare Abhängigkeit.

Die Konturlinien, d.h. die Kurven konstanter Dichte, sind hier Ellipsen. Dies war eine der wichtigen Eigenschaften, die Galton bei seinen Untersuchungen über den Zusammenhang der Größe von Vätern und Söhnen entdeckte.

Aus seinen aufgezeichneten Beobachtungen erkannte er:

1. Die Randverteilungen sind Normalverteilungen.
2. Die bedingten Erwartungswerte $\mathrm{E}(Y|X=x)$ sind lineare Funktionen von x.
3. Die bedingten Varianzen $\mathrm{Var}(Y|X=x)$ sind konstant in x.
4. Die Menge der Punkte auf der Ebene, für die jeweils $f_{X,Y}(x,y) = c$ gilt, sind Ellipsen.

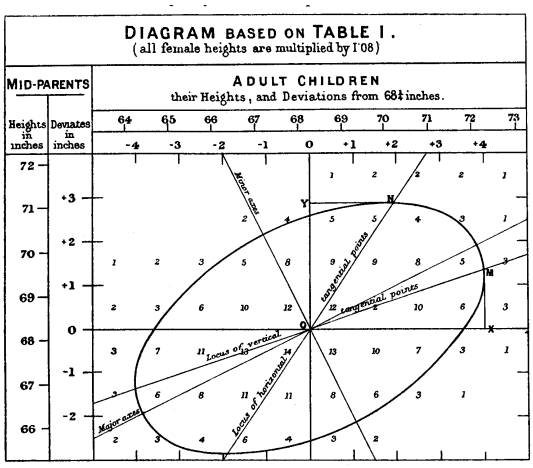


Abb. 3.13: *Galtons Konturlinien*

Insgesamt hat Galton also aus seinen Beobachtungen die zweidimensionale Normalverteilung als geeignetes Modell abgeleitet.

Um die allgemeine multivariate Normalverteilung anzugeben, ist der Übergang zur matriziellen Schreibweise nötig. Für den Exponenten der bivariaten Dichte können wir zunächst schreiben:

$$
\begin{aligned}
&-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\,\frac{x-\mu_X}{\sigma_X}\cdot\frac{y-\mu_Y}{\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\\
&\qquad = -\frac{1}{2(1-\rho^2)}(x-\mu_X, y-\mu_Y)\begin{pmatrix} \frac{1}{\sigma_X^2} & -\frac{\rho}{\sigma_X\sigma_Y} \\ -\frac{\rho}{\sigma_X\sigma_Y} & \frac{1}{\sigma_Y^2}\end{pmatrix}\begin{pmatrix} x-\mu_X \\ y-\mu_Y\end{pmatrix}\\
&\qquad = -\frac{1}{2}(x-\mu_X, y-\mu_Y)\begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}^{-1}\begin{pmatrix} x-\mu_X \\ y-\mu_Y\end{pmatrix}\\
&\qquad = -\frac{1}{2}(x-\mu_X, y-\mu_Y)\begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2\end{pmatrix}^{-1}\begin{pmatrix} x-\mu_X \\ y-\mu_Y\end{pmatrix},
\end{aligned}
$$

wobei σ_{XY} die Kovarianz von X und Y bezeichnet. Zudem gilt

$$\sigma_X^2\sigma_Y^2(1-\rho^2) = \det\begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}.$$

Schreiben wir also für die Kovarianzmatrix kurz Σ, so lässt sich die bivariate Normalverteilung

lung kompakt angeben durch

$$f_{XY}(x,y) = \frac{1}{(2\pi)^{1/2}(\det(\Sigma))^{1/2}} \cdot \exp\left[-\frac{1}{2}(x-\mu_X, y-\mu_Y)\Sigma^{-1}\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}\right].$$

Definition 3.1.17 *multivariate Normalverteilung*

Die k-dimensionale Zufallsvariable $\boldsymbol{X} = (X_1, \ldots, X_k)$ mit dem Erwartungswertvektor $\boldsymbol{\mu}$ und der nichtsingulären Kovarianzmatrix Σ heißt *multivariat normalverteilt*, wenn die gemeinsame Dichte gegeben ist durch

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}(\det(\Sigma))^{1/2}} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})'\right]. \tag{3.30}$$

Einige wichtige Eigenschaften der multivariaten Normalverteilung sind:

1. Linearkombinationen multivariat normalverteilter Zufallsvariablen sind wieder (multivariat) normalverteilt.
2. Jede Lineartransformation $\boldsymbol{Y} = \mathbf{A}\boldsymbol{X} + \boldsymbol{b}$ mit $\mathbf{A} \neq \mathbf{0}$ ergibt wieder eine Normalverteilung.
3. Alle aus Komponenten gemeinsam multivariat normalverteilter Zufallsvariablen gebildeten Untervektoren sind wieder multivariat normalverteilt.
4. Komponenten mit verschwindender Korrelation sind stochastisch unabhängig.
5. Die bedingten Verteilungen von Komponenten sind Normalverteilungen.

3.1.5 Exponentialfamilien

Viele der standardmäßig eingesetzten Verteilungsmodelle weisen eine einheitliche Struktur auf. Damit ist es möglich, bei der Analyse statistischer Verfahren für die Parameter solcher Verteilungen eine Vielzahl von Einzelmodellen mit einem einheitlichen Zugang zu behandeln. Dieser Zugang hat sich in der modernen Statistik als sehr bedeutsam erwiesen. Dementsprechend wird er an zahlreichen Stellen wieder aufgegriffen.

Beispiel 3.1.18 *verschiedene Exponentialfamilien*

Die einheitliche Struktur der Dichtefunktionen ist bisweilen erst auf den zweiten Blick ersichtlich. Verschiedene Beispiele machen dies deutlich:

Exponentialverteilung:

$$f(x) = \lambda e^{-\lambda x} = \lambda \cdot \exp[-\lambda x] = \exp[-\lambda x + \ln(\lambda)], \qquad x > 0;$$

Normalverteilung:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \exp\left[-\frac{1}{2}\cdot\frac{1}{\sigma^2}\cdot x^2 + \frac{\mu}{\sigma^2}\cdot x - \frac{1}{2}\left(\frac{\mu}{\sigma}\right)^2 - \ln\left(\sqrt{2\pi}\sigma\right)\right];$$

Gammaverteilung:

$$f(x)=\frac{\beta^{\lambda}}{\Gamma(\lambda)}\cdot x^{\lambda-1}\cdot e^{-\beta x}=\exp[-\beta\cdot x+(\lambda-1)\cdot\ln(x)+\lambda\cdot\ln(\beta)-\ln(\Gamma(\lambda)]\,,\qquad x>0;$$

Poisson-Verteilung:

$$f(x)=e^{-\lambda}\cdot\frac{\lambda^{x}}{x!}=\exp[x\cdot\ln(\lambda)-\lambda-\ln(x!)]\,,\qquad x=0,1,2,\ldots;$$

Binomialverteilung:

$$f(x)=\binom{n}{x}p^{x}(1-p)^{n-x}=\exp\left[\ln\left(\frac{p}{1-p}\right)\cdot x+n\cdot\ln(1-p)+\ln\left(\binom{n}{x}\right)\right],$$
$$x=0,1,\ldots,n.$$

Die im Beispiel angegebenen Dichtefunktionen können in ähnlicher Weise mit Hilfe der Exponentialfunktion angegeben werden. Wir betrachten bei der Verallgemeinerung zuerst den Fall, dass die Verteilungen nur von einem Parameter abhängen. Gegebenenfalls ist der weitere Parameter vorerst als fest zu betrachten.

Definition 3.1.19 *eindimensionale, einparametrige Exponentialfamilie*

Eine Familie von Verteilungen $\mathscr{F}$ sei durch ihre Dichten gegeben, $\mathscr{F}=\{f(x;\vartheta)|\vartheta\in\Theta\}$. $\mathscr{F}$ heißt eine *eindimensionale, einparametrige Exponentialfamilie,* falls $f(x;\vartheta)$ in folgender Form dargestellt werden kann:

$$f(x;\vartheta)=\exp[c(\vartheta)\cdot T(x)+d(\vartheta)+S(x)]\cdot I_A(x).$$

Dabei sind $c(\vartheta)$ und $d(\vartheta)$ reellwertige Funktionen auf Θ, $T(x)$ und $S(x)$ Abbildungen von $\mathbb{R}$ in $\mathbb{R}$ und schließlich ist $I_A(x)$ die Indikatorvariable der festen Teilmenge $A\subset\mathbb{R}$: $I_A(x)=1$ für $x\in A$ und $I_A(x)=0$ sonst.

Beispiel 3.1.20 *verschiedene Exponentialfamilien - Fortsetzung von Seite 93*

Wir identifizieren die in der Definition angegebenen Funktionen bei den im letzten Beispiel angegebenen Verteilungen:

Exponentialverteilung:

$$\vartheta=\lambda,\quad c(\lambda)=-\lambda,\quad T(x)=x,\quad d(\lambda)=\ln(\lambda),\quad S(x)=0,\quad A=(0,\infty).$$

Normalverteilung mit festem σ^2:

$$\vartheta=\mu,\quad c(\mu)=\frac{\mu}{\sigma^2},\quad T(x)=x,\quad d(\mu)=-\frac{\mu^2}{2\sigma^2},$$
$$S(x)=-\left(\frac{x^2}{2\sigma^2}+\ln\left(\sqrt{2\pi}\sigma\right)\right),\quad A=\mathbb{R}.$$

Gammaverteilung mit festem β:

$$\vartheta = \lambda, \quad c(\lambda) = \lambda - 1, \quad T(x) = \ln(x), \quad d(\lambda) = \lambda \cdot \ln(\beta) - \ln(\Gamma(\lambda),$$
$$S(x) = -\beta \cdot x, \quad A = \mathbb{R}_+ .$$

Poisson-Verteilung:

$$\vartheta = \lambda, \quad c(\lambda) = \ln(\lambda), \quad T(x) = x, \quad d(\lambda) = -\lambda, \quad S(x) = -\ln(x!), \quad A = \mathbb{N}.$$

Binomialverteilung:

$$\vartheta = p, \quad c(p) = \ln\left(\frac{p}{1-p}\right), \quad T(x) = x, \quad d(p) = n \cdot \ln(1-p),$$
$$S(x) = \ln\left(\binom{n}{x}\right), \quad A = \{0, 1, \ldots, n\}.$$

Anmerkung 3.1.21

In vielen Fällen ist es günstig, nicht von dem ursprünglichen Parameter ϑ der einparametrigen Exponentialfamilie $f(x;\vartheta) = \exp[c(\vartheta)T(x) + d(\vartheta) + S(x)] \times I_A(x)$ auszugehen, sondern $\eta := c(\vartheta)$ als neuen, *natürlichen Parameter* zu betrachten. Dann hat die Exponentialfamilie die *natürliche Form*

$$f(x;\eta) = \exp[\eta \cdot T(x) + d_0(\eta) + S(x)] \cdot I_A(x).$$

$c(\vartheta)$ muss eine streng monotone Funktion von ϑ sein, da sonst die eindeutige Korrespondenz der Verteilungen und der Parameter nicht gewährleistet ist. Daher sind die beiden Parametrisierungen gleichwertig.

Gilt $T(x) = x$, so sagt man, die Exponentialfamilie liege in *kanonischer Form* vor. A, der Träger der Verteilung, ist eine parameterunabhängige Menge. Daher bilden Verteilungen wie die Gleichverteilung, $f(x) = 1/\vartheta$ für $0 < x < \vartheta$ und $f(x) = 0$ sonst, oder die verschobene Exponentialverteilung, $f(x) = \lambda \exp[-\lambda(x-a)]$ für $x > a$ und $f(x) = 0$ sonst, keine Exponentialfamilien.

Beispiel 3.1.22 *verschiedene Exponentialfamilien - Fortsetzung von Seite 94*

Für die Darstellung in der natürlichen Form müssen bei den in den Beispielen 3.1.18 und 3.1.20 betrachteten Exponentialfamilien die folgenden Reparametrisierungen vorgenommen werden:

Exponentialverteilung: $\eta = -\lambda, \quad d_0(\eta) = \ln(-\eta)$;

Normalverteilung mit festem σ^2: $\eta = \frac{\mu}{\sigma^2}, \quad d_0(\eta) = -\sigma^2 \frac{\eta^2}{2}$;

Gammaverteilung mit festem β: $\eta = \lambda - 1, \quad d_0(\eta) = (\eta + 1) \cdot \ln(\beta) - \ln(\Gamma(\eta + 1)$;

Poisson-Verteilung: $\eta = \ln(\lambda), \quad d_0(\eta) = -e^{\eta}$;

Binomialverteilung: $\eta = \ln\left(\frac{p}{1-p}\right), \quad d_0(\eta) = n \cdot \ln\left(\frac{e^{\eta}}{1+e^{\eta}}\right)$.

Für die Binomialverteilung spielt der natürliche Parameter $\eta = \ln(p/(1-p))$ als logarithmierte *Odds-Ratio* eine große Rolle in der Epidemiologie.

Der Parameter ϑ einer Exponentialfamilie ist in natürlicher Weise mit der transformierten Zufallsvariablen $T(X)$ verbunden. Diese ist in verschiedener Weise ausgezeichnet. Hier soll zunächst gezeigt werden, dass i.d.R. die Verteilung von $T(X)$ wieder zu einer Exponentialfamilie gehört.

Satz 3.1.23 *Verteilungseigenschaft der Statistik $T(X)$*

Sei $\mathcal{F} = \{f(x;\vartheta)|\vartheta \in \Theta\}$ eine Exponentialfamilie mit den üblichen Funktionen $c(\vartheta)$, $d(\vartheta)$, $T(x)$, $S(x)$ und dem Träger A. Die Verteilung von $T(X)$ sei vom gleichen Typ wie die von X, also gleichermaßen stetig bzw. diskret. $g(t;\vartheta)$ sei die Dichtefunktion von $T(X)$. Dann bildet $\{g(t;\vartheta)|\vartheta \in \Theta\}$ eine einparametrige Exponentialfamilie:

$$g(t;\vartheta) = \exp[c(\vartheta) \cdot t + d(\vartheta) + S^*(t)] \cdot I_{A^*}(t).$$

Dabei sind $S^*(t)$ und $A^*(t)$ geeignet definiert.

Beweis: Der Beweis wird für den diskreten Fall geführt. Hier gilt nach Definition:

$$g(t;\vartheta) = \mathrm{P}_\vartheta(T(X) = t) = \sum_{\substack{x \\ T(x)=t}} \exp[c(\vartheta) \cdot T(x) + d(\vartheta) + S(x)] \cdot I_A(x)$$

$$= \exp[c(\vartheta) \cdot t + d(\vartheta)] \cdot \left\{ \sum_{\substack{x \in A \\ T(x)=t}} \exp[S(x)] \right\} \cdot I_{A^*}(t),$$

wobei $A^* = \{t|T(x) = t,\ x \in A\}$.

Wird nun $S^*(t) = \ln\left\{\sum_{\substack{x \in A \\ T(x)=t}} \exp[S(x)]\right\}$ für $t \in A^*$ gesetzt, so ist die gewünschte Darstellung der Dichte von $T(X)$ erreicht.

Beispiel 3.1.24 *Pareto-Verteilung*

X sei *Pareto-verteilt*, $f(x;\alpha) = \alpha \cdot x^{-(\alpha+1)}$ $(x > 1)$. Der Parameter α wird größer als 0 vorausgesetzt. Damit gehört die Verteilung von X zu einer Exponentialfamilie:

$$f(x;\alpha) = \exp[-\alpha \cdot \ln(x) - \ln(x) + \ln(\alpha)] \cdot I_{[1,\infty)}(x).$$

Die transformierte Zufallsvariable $T(X) = \ln(X)$ hat nach dem Transformationssatz 1.3.5 die Dichte

$$\begin{aligned} g(t;\alpha) &= f(T^{-1}(t);\alpha) \cdot \left| \frac{\partial}{\partial t} T^{-1}(t) \right| \\ &= \exp[-\alpha \cdot t - t + \ln(\alpha)] \cdot I_{[0,\infty)}(t) \cdot \exp[t] = \exp[-\alpha \cdot t + \ln(\alpha)] \cdot I_{[0,\infty)}(t) \\ &= \begin{cases} \alpha \cdot e^{-\alpha t} & \text{für } t > 0 \\ 0 & \text{für } t \leq 0 \end{cases}. \end{aligned}$$

Die logarithmische Transformation führt bei der Pareto-Verteilung also zur Exponentialverteilung.

Die momenterzeugende Funktion und damit speziell Erwartungswert und Varianz einer Zufallsvariablen X, deren Verteilung zu einer in kanonischer Form vorliegenden Exponentialfamilie gehört, hängen in einfacher Weise vom natürlichen Parameter η und von $d_0(\eta)$ ab. Wir formulieren die Aussage allgemeiner als Aussage über die Verteilung von $T(X)$. Die wesentliche Voraussetzung ist dabei, dass der Parameterraum genügend reichhaltig ist, genauer, dass der Bereich H, in dem der natürliche Parameter η variieren kann, ein Intervall oder ganz $\mathbb{R}$ ist.

Satz 3.1.25 *momenterzeugende Funktion bei Exponentialfamilien*

Hat X die Dichte $f(x;\eta) = \exp[\eta \cdot T(x) + d_0(\eta) + S(x)] \cdot I_A(x)$ und ist η ein innerer Punkt von H, so existiert die momenterzeugende Funktion von $T(X)$ und ist für s aus einem geeigneten Intervall um Null gegeben durch:

$$M_{T(X)}(s) = \exp[d_0(\eta) - d_0(s+\eta)]. \tag{3.31}$$

Weiter gilt:

$$\mathrm{E}(T(X)) = -d_0'(\eta), \quad \mathrm{Var}(T(X)) = -d_0''(\eta). \tag{3.32}$$

Beweis: Wir schreiben den Beweis für den stetigen Fall auf. Es gilt:

$$\begin{aligned} M_{T(X)}(s) = \mathrm{E}\left[e^{s \cdot T(X)}\right] &= \int_A \exp[s \cdot T(x)] \cdot \exp[\eta \cdot T(x) + d_0(\eta) + S(x)] \, \mathrm{d}x \\ &= \int_A \exp[(s+\eta) \cdot T(x) + d_0(\eta) + S(x)] \, \mathrm{d}x \\ &= \exp[d_0(\eta) - d_0(s+\eta)] \cdot \int_A \exp[(s+\eta) \cdot T(x) + d_0(s+\eta) + S(x)] \, \mathrm{d}x \\ &= \exp[d_0(\eta) - d_0(s+\eta)]. \end{aligned}$$

Die letzte Gleichung ergibt sich daraus, dass das Integral über eine Dichte Eins ist.

Die ersten beiden Ableitungen von $M_{T(X)}(s)$ ergeben die benötigten Momente:

$$\begin{aligned}
\mathrm{E}(T(X)) &= \frac{\partial}{\partial s}\exp[d_0(\eta)-d_0(s+\eta)]\Big|_{s=0} = \exp[d_0(\eta)-d_0(s+\eta)]\cdot(-d_0'(s+\eta))\big|_{s=0} \\
&= -d_0'(\eta); \\
\mathrm{E}(T(X)^2) & \\
&= \frac{\partial^2}{\partial s^2}\exp[d_0(\eta)-d_0(s+\eta)]\Big|_{s=0} = \frac{\partial}{\partial s}\exp[d_0(\eta)-d_0(s+\eta)]\cdot(-d_0'(s+\eta))\Big|_{s=0} \\
&= \{\exp[d_0(\eta)-d_0(s+\eta)]\cdot(-d_0'(s+\eta))^2 + \exp[d_0(\eta)-d_0(s+\eta)]\cdot(-d_0''(s+\eta))\}\big|_{s=0} \\
&= d_0'(\eta)^2 - d_0''(s+\eta).
\end{aligned}$$

So erhalten wir auch die angegebene Form der Varianz.

Beispiel 3.1.26 *Pareto-Verteilung - Fortsetzung von Seite 96*

X sei eine Pareto-verteilte Zufallsvariable. Für den Parameter gilt: $\eta = -\alpha$. Die transformierte Zufallsvariable $T(X) = \ln(X)$ hat nach dem Satz daher die momenterzeugende Funktion

$$M_{T(X)}(s) = \exp[d_0(\eta)-d_0(s+\eta)] = \exp[\ln(-\eta)-\ln(s+\eta)] = \frac{\alpha}{s-\alpha}.$$

Das ergibt speziell $\mathrm{E}(T(X)) = 1/\alpha$ und $\mathrm{Var}(T(X)) = 1/\alpha^2$.

Zu beachten ist hier insbesondere, dass nichts über die Momente von X ausgesagt wird. Die Pareto-Verteilung hat ja keine momenterzeugende Funktion!

Sind $X_1,\ldots,X_n$ unabhängige identisch verteilte Zufallsvariablen, deren Verteilung zu einer Exponentialfamilie gehört, so gehört auch die gemeinsame Verteilung von $\boldsymbol{X} = (X_1,\ldots,X_n)$ zu einer Exponentialfamilie. Dann ist:

$$f(x_1,\ldots,x_n;\vartheta) = \exp\left[c(\vartheta)\sum_{i=1}^{n}T(x_i) + n\cdot d(\vartheta) + \sum_{i=1}^{n}S(x_i)\right]\cdot I_{A\times\ldots\times A}(x_1,\ldots,x_n).$$

Eine Verallgemeinerung der einparametrigen Exponentialfamilie stellt die mehrparametrige dar.

Definition 3.1.27 *k-parametrige Exponentialfamilie*

Eine Familie von Verteilungen heißt eine ***k*-parametrige Exponentialfamilie**, wenn es reellwertige Funktionen $c_1,\ldots,c_k$ auf Θ gibt, so dass die Dichte $f(x;\vartheta)$ die Gestalt

$$f(x;\vartheta) = \exp[c_i(\vartheta)\cdot T_i(x) + d(\vartheta) + S(x)]\cdot I_A(x) \tag{3.33}$$

besitzt. $T_1(x),\ldots,T_k(x)$ sind Abbildungen von $\mathbb{R}$ in $\mathbb{R}$ und $d(\vartheta)$, $S(x)$ sowie $I_A(x)$ sind wie in Definition 3.1.19 festgelegt.

Beispiel 3.1.28 *Normalverteilung als zweiparametrige Exponentialfamilie*

Es wird wieder die Normalverteilung betrachtet, jetzt aber mit $\boldsymbol{\vartheta} = (\mu, \sigma^2)$. $f(x;\mu,\sigma^2)$ bildet nun eine 2-parametrige Exponentialfamilie. Es gilt:

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ &= \exp\left[-\frac{1}{2}\cdot\frac{1}{\sigma^2}\cdot x^2 + \frac{\mu}{\sigma^2}\cdot x - \frac{1}{2}\left(\frac{\mu}{\sigma}\right)^2 - \ln\left(\sqrt{2\pi}\sigma\right)\right]. \end{aligned}$$

Folglich sind:

$$c_1(\boldsymbol{\vartheta}) = \mu/\sigma^2, \quad T_1(x) = x, \quad c_2(\boldsymbol{\vartheta}) = -1/2\sigma^2, \quad T_2(x) = x^2,$$
$$d(\boldsymbol{\vartheta}) = -\mu^2/2\sigma^2, \quad S(x) \equiv 0, \quad A = \mathbb{R}.$$

Satz 3.1.29 *Verteilungseigenschaft von* $\boldsymbol{T} = (T_1, \ldots, T_k)$

Die Verteilung von $\boldsymbol{X}$ gehöre zu einer mehrparametrigen Exponentialfamilie,

$$f(\boldsymbol{x};\vartheta,\zeta_1,\ldots,\zeta_k) = \exp\left[\vartheta U(\boldsymbol{x}) + \sum_{i=1}^{k}\zeta_i T_i(\boldsymbol{x}) + d(\vartheta,\zeta_1,\ldots,\zeta_k) + S(\boldsymbol{x})\right]\cdot I_A(\boldsymbol{x}).$$

Dann bilden die Verteilungen von $\boldsymbol{T} = (T_1, \ldots, T_k)$ eine Exponentialfamilie der Form

$$f(\boldsymbol{t};\vartheta,\zeta_1,\ldots,\zeta_k) = \exp\left[\sum_{i=1}^{k}\zeta_i t_i + d(\vartheta,\zeta_1,\ldots,\zeta_k) + S(\boldsymbol{t})\right]\cdot I_B(\boldsymbol{t}).$$

Die bedingten Verteilungen von U bei gegebenem $\boldsymbol{T} = \boldsymbol{t}$ bilden wiederum eine Exponentialfamilie:

$$f_{U|\boldsymbol{T}}(u|\boldsymbol{t};\vartheta) = \exp[\vartheta u + d_{\boldsymbol{t}}(\vartheta) + S_{\boldsymbol{t}}(u)]\cdot I_{C(\boldsymbol{t})}(u).$$

Diese ist insbesondere von den Parametern $\zeta_1, \ldots, \zeta_k$ unabhängig.

Beweis: Lehmann (1986, S.56).

Beispiel 3.1.30 *Poisson-verteilte Zufallsvariablen*

X und Y seien unabhängige Poisson-verteilte Zufallsvariablen mit den Parametern λ und μ. Dann ist ihre gemeinsame Verteilung gegeben durch die Dichte

$$\begin{aligned} f(x,y;\lambda,\mu) &= e^{-\lambda}\frac{\lambda^x}{x!}e^{-\mu}\frac{\mu^y}{y!} \qquad x = 0,1,\ldots;\ y = 0,1,\ldots \\ &= \exp[-\lambda + x\cdot\ln(\lambda) - \mu + y\cdot\ln(\mu) - \ln(x!) - \ln(y!)]\cdot I_A(x,y) \\ &= \exp\left[\ln\left(\frac{\lambda}{\mu}\right)\cdot x + \ln(\mu)\cdot(x+y) + d(\lambda,\mu) + S(x,y)\right]\cdot I_A(x,y) \\ &= \exp[\vartheta x - \xi(x+y) + d(\vartheta,\xi) + S(x,y)]\cdot I_A(x,y); \end{aligned}$$

dabei sind $\vartheta = \ln(\lambda/\mu)$, $\xi = \ln(\mu)$ und $A = \mathbb{N}_0^2$.

Wir haben somit eine Darstellung der gemeinsamen Dichte von X und Y als zweiparametrige Exponentialfamilie mit den Statistiken $U = X$ und $T = X + Y$. T hat wieder eine Poisson-Verteilung:

$$f(t) = \exp[-(\lambda+\mu) + t \cdot \ln(\lambda+\mu) - \ln(t!)] \cdot I_{\mathbb{N}}(t).$$

Für die bedingte Verteilung von X bei gegebenem $T = t$ erhalten wir:

$$f(x|t;\lambda,\mu) = \frac{e^{-\lambda}\frac{\lambda^x}{x!}e^{-\mu}\frac{\mu^{t-x}}{(t-x)!}}{e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^t}{t!}} = \binom{t}{x}\left(\frac{\lambda}{\lambda+\mu}\right)^x\left(1-\frac{\lambda}{\lambda+\mu}\right)^{t-x} \quad x = 0,1,\ldots,t.$$

Dies ist eine Binomialverteilung, die nur noch von dem Parameter ϑ abhängt:

$$\frac{\lambda}{\lambda+\mu} = \frac{1}{1+\exp[-\ln(\lambda/\mu)]}.$$

Dass Binomialverteilungen speziell eine Exponentialfamilie bilden, ist aus Beispiel 3.1.18 bekannt.

So wichtig einerseits Exponentialfamilien sind, so ist die Voraussetzung einer Exponentialfamilie andererseits sehr restriktiv. Das wird u. a. daran deutlich, dass die Normalverteilungen mit festem σ die einzige einparametrige Exponentialfamilie bilden, die gleichzeitig eine Lage-Familie ist. Ähnlich stellen die Gammaverteilungen mit festem Gestaltparameter die einzige Skalen-Familie unter den einparametrigen Exponentialfamilien dar. Vgl. dazu Bühler & Sehr (1987).

Satz 3.1.31 *hinreichende Bedingung für die Normalverteilungsfamilie*

Seien die Dichten der Familie $\mathscr{F} = \{f(x;\vartheta)|\vartheta \in \Theta\}$ mit $\Theta = \mathbb{R}$ gegeben durch

$$f(x;\vartheta) = \exp[c(\vartheta)\cdot x + d(\vartheta) + S(x)] \cdot I_{\mathbb{R}}(x).$$

$c(\vartheta)$ und $d(\vartheta)$ seien differenzierbar. $\mathscr{F}$ möge eine Lage-Familie bilden, d.h.

$$f(x;\vartheta) = f(x-\vartheta;0) =: f(x-\vartheta).$$

Dann ist $\mathscr{F}$ die Normalverteilungsfamilie mit festem σ^2.

Beweis: Mit $\bar{c}(\vartheta) = c(\vartheta) - c(0)$ und $\bar{d}(\vartheta) = d(\vartheta) - d(0)$ erhalten wir:

$$f(x-\vartheta) = f(x;\vartheta) = \exp[c(\vartheta)\cdot x + d(\vartheta) + S(x)]$$

$$= \exp[\bar{c}(\vartheta)\cdot x + \bar{d}(\vartheta)] \cdot \exp[c(0)\cdot x + d(0) + S(x)] = \exp[\bar{c}(\vartheta)\cdot x + \bar{d}(\vartheta)] \cdot f(x).$$

Ableiten nach ϑ und Betrachten der Ableitung bei $\vartheta = 0$ ergibt wegen der Definition von $\bar{c}(\vartheta)$ und $\bar{d}(\vartheta)$:

$$-f'(x) = [\bar{c}'(0)\cdot x + \bar{d}'(0)] \cdot f(x).$$

Da nach Definition $f(x) > 0$ für alle x, kann dies umgeschrieben werden zu

$$\frac{f'(x)}{f(x)} = -Ax + B.$$

Integrieren auf beiden Seiten ergibt

$$\ln(f(x)) = -A \cdot \frac{x^2}{2} + B \cdot x + C \quad \text{bzw.} \quad f(x) = \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \cdot \text{const}.$$

Dabei sind $\mu = B/A$, $\sigma^2 = 1/|A|$, und $const$ ist so zu bestimmen, dass $f(x)$ eine Dichte darstellt. Damit ist aber $\mathscr{F}$ die behauptete Normalverteilungsfamilie.

3.2 Strukturierte Modelle

Mit einer Familie von Wahrscheinlichkeitsverteilungen wird das allgemeine Verteilungsgesetz einer Zufallsvariablen modelliert. Die Hervorhebung spezieller Aspekte der Verteilungen führt zu strukturierten Modellen. Insbesondere lassen sich bei diesen zwei Aspekte verbinden: der Aspekt der Regularität und der der Irregularität oder zufälligen Schwankung.

3.2.1 Signal-plus-Rauschen-Modelle

Beispiel 3.2.1 *Darstellung einer normalverteilten Zufallsvariablen*

Die Zufallsvariable Y stelle eine Messung dar. Für Y wird dann gern eine Normalverteilung unterstellt: $Y \sim \mathscr{N}(\mu, \sigma^2)$. Y kann auch in der Form

$$Y = \mu + U \qquad \text{mit } U \sim \mathscr{N}(0, \sigma^2)$$

ausgedrückt werden. Diese Darstellung hat den Vorzug, das Wesentliche zu betonen: μ stellt den "wahren Wert" dar; die Messungen ergeben i.d.R. nicht diesen wahren Wert, sondern streuen um ihn. Das wird durch die Störung U ausgedrückt, für die $\mathrm{E}(U) = 0$ gilt.

Das in dem Beispiel formulierte Messmodell ist eine einfache Version der sogenannten *Signal-plus-Rauschen-Modelle*. In lockerer Schreibweise gilt für das SPR-Modell:

Variable = systematische Komponente + Zufallskomponente.

Beispiel 3.2.2 *Registrierfehler von Stromzählern*

Eine Anwendung des SPR-Modells liefert einen Untersuchung von Registrierfehlern von Stromzählern, wie sie für jede Wohnung üblich sind. Bei 144 Zählern, die den Stromverbrauch unter 'Standardbedingungen' festhielten, wurden die festgestellten Verbräuche (in kWh) mit den Ergebnissen von Zählern verglichen, die in Reihe geschaltet waren, aber geschützt gegen Umwelteinflüsse wie Temperatur, Feuchtigkeit etc. Nach einem Jahr ergab sich die in Abbildung 3.14 dargestellte Verteilung der Abweichungen (in %).

Der Modellansatz lautet hier:

Messfehler = systematische Komponente + Zufallskomponente.

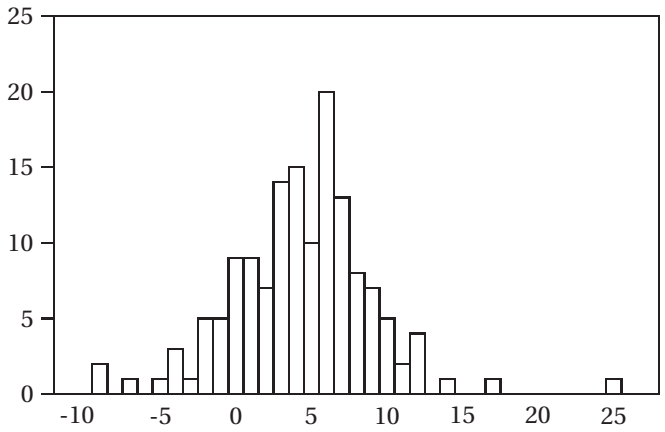


Abb. 3.14: *Messfehler von Stromzählern*

Bei vielen Anwendungen ist die systematische Komponente als abhängig von anderen Variablen X anzusehen. Dies führt zu der Formulierung der SPR-Modelle mittels bedingter Erwartungswerte:

$$\mathrm{E}(Y|X=x)=h(x). \tag{3.34}$$

In diesem Ansatz muss X keine Zufallsvariable sein, sondern kann auch eine allgemeine, systematisch veränderbare Variable darstellen.

Die Zufallskomponente wird dabei meist als unabhängig vom speziellen Wert der beeinflussenden Variablen X unterstellt. Dann wird das Modell gern in der nicht ganz präzisen Form

$$Y=h(X)+U$$

geschrieben. Zu unterschiedlichen Spezialfällen gelangt man nun, indem für die systematische Komponente $h(X)$ geeignete Festlegungen getroffen werden.

3.2.2 Einfaktorielle Varianzanalyse

Die Zufallsvariable X bezeichne die k unterschiedlichen Versuchsbedingungen eines Laborexperimentes. Dann kann X nur k Werte, die mit $1,\ldots,k$ bezeichnet werden, annehmen. Das Modell erhält die Gestalt

$$\mathrm{E}(Y|X=i)=\mu_i \qquad i=1,\ldots,k$$

oder, in offensichtlicher Modifikation der bisherigen Schreibweise:

$$Y_i=\mu_i+U\,.$$

Die üblichen Fragestellungen dieses Ansatzes zielen auf die Gleichheit der Erwartungswerte μ_i. Entweder ist die postulierte Gleichheit zu testen, oder es sind Kontraste $\mu_i-\mu_j$ zu schätzen.
Oft wird der Ansatz modifiziert und unterstellt, dass es einen mittleren Einfluß μ gibt, so dass

die verschiedenen Faktorstufen i die Abweichungen zum globalen Mittel bewirken. Dann wird geschrieben:

$$Y_i = \mu + \alpha_i + U \qquad i = 1, \ldots, k\,. \tag{3.35}$$

Diese Parametrisierung ist offensichtlich nicht eindeutig. Wir können mit einem Satz von Parametern $\mu, \alpha_1, \ldots, \alpha_k$ sofort beliebig viele weitere Parametersätze angeben:

$$\mu + c, \alpha_1 - c, \ldots, \alpha_k - c, \qquad c \in \mathbb{R}.$$

Zur Lösung dieses Problems wird die zusätzliche Reparametrisierungbedingung $\sum_{i=1}^{k} \alpha_i = 0$ eingeführt.

3.2.3 Zweifaktorielle Varianzanalyse

Bei der einfaktoriellen Varianzanalyse wird der Einfluss einer qualitativen Variablen X auf eine quantitative Zufallsvariable Y modelliert. Eine Erweiterung des Modells auf zwei beeinflussende Faktoren X_1 und X_2 mit den Faktorstufen $1, \ldots, k$ und $1, \ldots, m$ liefert

$$Y_{ij} = \mu_{ij} + U \qquad i = 1, \ldots, k,\ j = 1, \ldots, m.$$

Wie bei der einfaktoriellen Varianzanalyse betrachtet man häufig das reparametrisierte Modell

$$Y_{ij} = \mu + \alpha_i + \beta_j + U \qquad i = 1, \ldots, k,\ j = 1, \ldots, m. \tag{3.36}$$

Die Reparametrisierungsbedingungen lauten nun $\sum_{i=1}^{k} \alpha_i = 0$ und $\sum_{i=1}^{k} \beta_i = 0$.

Das Modell erfasst nur die einfache additive Überlagerung der beiden Einflussfaktoren. Können diese sich wechselseitig beeinflussen, so ist es sinnvoll das Modell zu erweitern:

$$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + U \qquad i = 1, \ldots, k,\ j = 1, \ldots, m. \tag{3.37}$$

Die Reparametrisierungsbedingungen sind nun um $\sum_{i=1}^{k} \sum_{j=1}^{m} (\alpha\beta)_{ij} = 0$ zu erweitern.

3.2.4 Lineare Regressionsmodelle

Bei stetigen Variablen X ist die Unterstellung einer linearen Abhängigkeit des bedingten Erwartungswertes $\mathrm{E}(Y|X = x)$ der wohl am meisten angewendete Ansatz:

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x.$$

Die Bedeutung dieses Ansatzes resultiert aus der Möglichkeit, formale Abhängigkeiten zumindest approximativ zu erfassen, theoretisch postulierte Zusammenhänge zu überprüfen und ggfs. Prognosen zu erstellen. Dazu sind die Koeffizienten β_0 und β_1 zu ermitteln oder zu prüfen, ob β_1 gleich Null ist. Nur im Fall $\beta_1 \neq 0$ ändert sich der Erwartungswert von Y mit X, hat der Regressor X also einen Einfluss auf den Regressanden Y.

Ein weiterer Vorteil des Modells besteht darin, dass es leicht auf die Situation mehrerer unabhängiger Variablen erweiterbar ist. Es bekommt dann die Gestalt

$$\mathrm{E}(Y|X_1 = x_1, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{3.38}$$

Geht man von n vorgegebenen Werte-Folgen der p unabhängigen Variablen aus, so erhält man n Zufallsvariablen $Y_1,\ldots,Y_n$, die sich — laut Modell — nur im Erwartungswert unterscheiden. Daher gilt nun:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + U_i. \tag{3.39}$$

Die Indizierung der Störvariablen U_i ist nötig, um das Modell korrekt in den Variablen statt in den bedingten Erwartungswerten ausdrücken zu können. Dass es sich um Replikate derselben Variablen handelt, wird durch die zusätzliche Forderung identischer Verteilungen der U_i einbezogen. Weiter werden die Störungen meist als unkorreliert und normalverteilt angesetzt:

$$U_i \sim \mathcal{N}(0;\sigma^2), \qquad \mathrm{Cov}(U_i,U_j)=0 \qquad i\neq j;\ \ i,j=1,\ldots,n.$$

3.2.5 Lineares Regressionsmodell mit stochastischen Regressoren

Wird im letzten Beispiel von Zufallsvariablen $X_1,\ldots,X_p$ mit nicht vorab festgelegten Werten ausgegangen, so beschreibt

$$\mathrm{E}(Y|X_1,\ldots,X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

ein lineares Regressionsmodell mit stochastischen Regressoren. Es stellt sich dann die Frage, bei welchen gemeinsamen Verteilungen von Y, $X_1,\ldots,X_p$ der lineare Ansatz gerechtfertigt ist. Bei gemeinsamer Normalverteilung ist der bedingte Erwartungswert von Y eine lineare Funktion der X_i. Für den Spezialfall $p=1$ ist dies in Kapitel 2 ausgeführt worden. Anhand eines konkreten Beispiels wird hier verdeutlicht, dass diese Eigenschaft keineswegs allgemein gilt.

Beispiel 3.2.3 *Regressionsfunktion einer Farlie-Morgenstern-Familie*

Sei der Einfachheit halber $p=1$ und $X=X_1$. Die gemeinsame Verteilung von (X,Y) gehöre zu der Farlie-Morgenstern-Familie

$$H(x,y) = F(x)\cdot G(y)\cdot\{1+\alpha\cdot[1-F(x)]\cdot[1-G(y)]\}$$

mit

$$F(x) = 1-e^{-\lambda x}, \qquad G(y) = 1-e^{-\vartheta y}.$$

Für die gemeinsame Dichte gilt:

$$h(x,y) = f(x)\cdot g(y)\cdot\{1+\alpha\cdot[2F(x)-1]\cdot[2G(y)-1]\}.$$

Damit ist die bedingte Dichte $h(y|x)$ unmittelbar ersichtlich, und wir erhalten:

$$\mathrm{E}(Y|X=x) = \int_0^\infty y\cdot h(y|x)\,\mathrm{d}y = \frac{1}{\vartheta}\cdot\left[1+\frac{\alpha}{2}-\alpha e^{-\lambda x}\right].$$

Somit ist

$$\mathrm{E}(Y|X) = \frac{1}{\vartheta} \cdot \left[1 + \frac{\alpha}{2} - \alpha e^{-\lambda X}\right],$$

mithin alles andere als linear.

Die Bestimmung der besten linearen Approximation von Y an X führt übrigens auf

$$\hat{Y} = \frac{1}{\vartheta} \cdot \left[1 - \frac{\alpha}{2} + \alpha \cdot \lambda \cdot X\right].$$

Beispiel 3.2.4 *Hubbles Gesetz*

Die Vorstellung vom Urknall und dem sich seit dem ersten Zeitpunkt aller Zeiten ausdehnenden und auseinanderfliegenden Universum basiert ursprünglich auf der Beobachtung der Rotverschiebung entfernter Galaxien und deren Interpretation durch Hubble im Jahre 1929. Die Teile des Universums, das sind im Wesentlichen die Galaxien, fliegen auseinander wie die Splitter nach einer großen Explosion, und zwar die weitesten am schnellsten. Daraus schließt man, dass die Urexplosion, d.h. der Urknall, in einem Zentrum stattgefunden hat und dass diejenigen Objekte, die den stärksten Impuls mitbekommen haben, eben am weitesten weggeflogen sind. Das wird zusammengefasst in der einfachen Hubble-Gleichung mit der Hubble-Konstanten H:

Fluchtgeschwindigkeit $= H \times$ Entfernung.

Das ist graphisch in der Abbildung 3.15 wiedergegeben (nach Larsen & Marx (1986)). Die Abbildung zeigt auch, dass das Gesetz nur bis auf stochastische Störungen gilt. Hier ist insgesamt ein Regressionsmodell mit einem stochastischen Regressor adäquat.

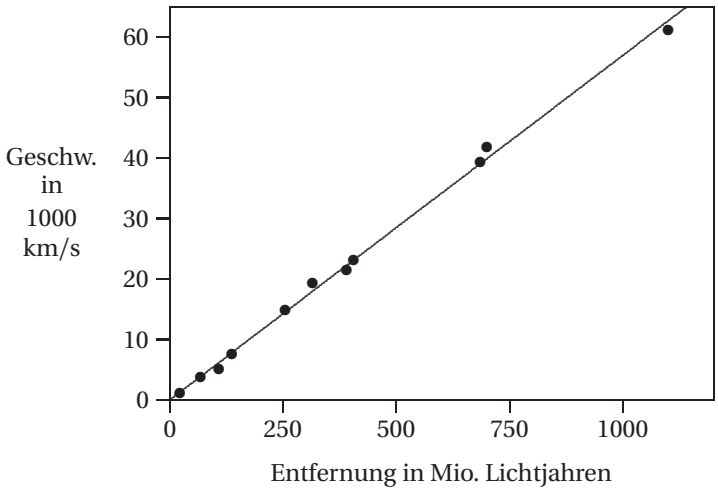


Abb. 3.15: *Hubbles Gesetz*

3.2.6 Nichtlineare Regressionsmodelle

Häufig sind nichtlineare Formen der Abhängigkeit der beobachteten Zufallsvariablen Y von der unabhängigen Variablen X sinnvoll. Das gilt etwa, wenn bei Wachstumsvorgängen von vornherein eine obere Schranke für das Wachstum gegeben ist. Die spezielle funktionale Form der nichtlinearen Abhängigkeit ist aus dem jeweiligen Kontext abzuleiten oder aufgrund der Beobachtungen aus dem "Vorrat der Modelle" auszuwählen. Für die breitere Anwendbarkeit können die Funktionen in nichtlinearer Weise von Parametern abhängen. Ziel ist es dann, diese unbekannten Parameter zu ermitteln.

Beispiel 3.2.5 *Wachstum mit Sättigungsverhalten*

Bei biologischen Wachstumsvorgängen mit Sättigungsverhalten wird vielfach eine logistische Funktion für $h(x)$ gewählt:

$$h(x;\alpha,\beta,\gamma,\delta) = \frac{\gamma}{1+\beta\cdot\exp[-\alpha(x-\delta)]}.$$

Im Fall $\alpha > 0$ strebt $h(x;\alpha,\beta,\gamma,\delta)$ mit wachsendem x gegen γ. Das Modell lautet damit:

$$Y = \frac{\gamma}{1+\beta\cdot\exp[-\alpha(x-\delta)]} + U.$$

Beispiel 3.2.6 *Cobb-Douglas-Konsumfunktion*

Eine Familie von Modellen in der Ökonomie basiert auf den Cobb-Douglas-Konsumfunktionen

$$Y = \vartheta_0 X_1^{\vartheta_1} X_2^{\vartheta_2} \cdot \ldots \cdot X_k^{\vartheta_k}.$$

Für die Berücksichtigung des Störterms wurden verschiedene Vorschläge gemacht. Das additive Modell ist

$$Y = \vartheta_0 X_1^{\vartheta_1} X_2^{\vartheta_2} \cdot \ldots \cdot X_k^{\vartheta_k} + U,$$

während das multiplikative lautet:

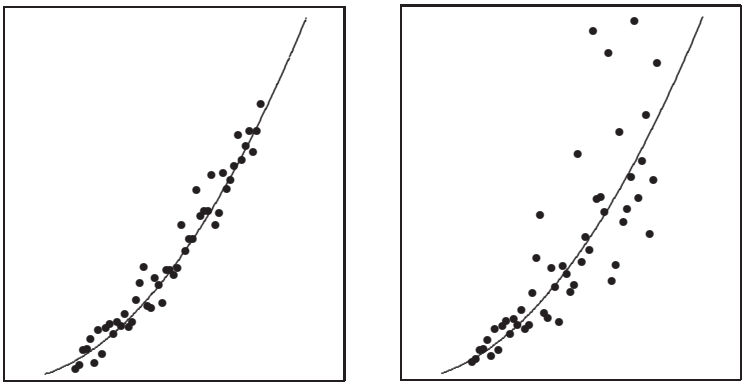
$$Y = \vartheta_0 X_1^{\vartheta_1} X_2^{\vartheta_2} \cdot \ldots \cdot X_k^{\vartheta_k} \cdot e^U.$$

In beiden Fällen wird $U \sim \mathcal{N}(0,\sigma^2)$ unterstellt.

Das multiplikative Modell lässt sich durch logarithmische Transformation in ein in den Parametern lineares Modell verwandeln:

$$\ln(Y) = \eta_0 + \eta_1 \ln(X_1) + \ldots + \eta_k \ln(X_k) + U.$$

Das vereinfacht die Anwendung des Modells erheblich. Das entscheidende Kriterium für die Modellwahl besteht jedoch in der Form, wie die Beobachtungen von der Regressionsfunktion abweichen, vgl. die Abbildung 3.16.



a) additiver Fehler b) multiplikativer Fehler

Abb. 3.16: *Nichtlineare Regressionsfunktionen*

3.2.7 Poisson-Regression

Bei den bisher betrachteten strukturierten Modellen ist die abhängige Zufallsvariable jeweils stetig. Oft soll aber die Abhängigkeit einer diskreten Zufallsvariablen von einer oder mehreren erklärenden Variablen modelliert werden. Beispiele hierfür sind die Anzahlen der Kunden in einer Warteschlange vor einer Kasse in Abhängigkeit von der Tageszeit oder die Anzahl der Krankheitsfälle in einer Population in Abhängigkeit vom Alter.

Ein leicht zu handhabendes Verteilungsmodell für Zählvariablen ist die Poisson-Verteilung. Soll, wie bei der linearen Regression, der bedingte Erwartungswert der Poisson-verteilten Zufallsvariablen Y modelliert werden, so ist der naheliegende Ansatz:

$$\mathrm{E}(Y|X_1 = x_1, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Hier sind wir aber mit dem Problem konfrontiert, dass auf der rechten Seite negative Werte vorkommen können, was bei einer Poisson-Verteilung unmöglich und allgemein bei Zählvariablen nicht sinnvoll ist. Daher wird der Ansatz modifiziert zu

$$\mathrm{E}(Y|X_1 = x_1, \ldots, X_p = x_p) = \exp[\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p]. \tag{3.40}$$

Man spricht in diesem Fall von *Poisson-Regression.*

Beispiel 3.2.7 *Verluste im Kühlwassersystem*

Die Häufigkeit Y von Verlusten im Kühlwassersystem in Kernkraftwerksanlagen ist sicherlich von der Betriebsdauer abhängig. Da es sich zum Glück um ein nicht so häufig vorkommendes Ereignis handelt, ist der Ansatz (3.40) hier adäquat.

Der Ansatz führt dazu, dass die Beobachtungen y_i als Realisationen einer Poisson-verteilten Zufallsvariablen Y angesehen werden mit

$$\mathrm{E}(Y|J = j) = \lambda e^{\beta j}.$$

Santner & Duffy (1989) geben die folgenden einschlägigen Daten an:

Anlage Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Anzahl Vorfälle	4	40	0	10	14	31	2	4	13	4	27	14	10	7	4
Betriebszeit (J)	15	12	8	8	6	5	5	4	4	3	4	4	4	2	3
Anlage Nr.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Anzahl Vorfälle	3	11	1	0	3	5	6	35	12	1	10	5	16	14	58
Betriebszeit (J)	3	2	2	2	1	1	1	5	3	1	3	2	4	3	11

3.2.8 Logistische Regression

Ein einfaches Beispiel einer Zählvariablen stellt eine binäre oder Bernoulli-Variable dar, welche nur die Werte 0 und 1 annehmen kann. Es gibt eine Vielzahl von Beispielen, bei denen die abhängige Variable dieses Typs ist. So wird ein Gut bei einem bestimmten Preis entweder gekauft oder nicht gekauft. Hier ist von Interesse, wie sich der Anteil der Käufer mit dem Preis ändert: Verschiedene Preise werden unterschiedliche Anteile von potentiellen Käufern zum entscheidenden Schritt verleiten. Eine "klassische" Anwendung von binären Variablen stellen Dosis-Wirkungsbeziehungen dar, bei der die Sterbewahrscheinlichkeit in Abhängigkeit von einer Medikamentendosis modelliert wird.

Ist Y eine binäre Zufallsvariable mit $P(Y=0)=1-p$, $P(Y=1)=p$, so gilt $E(Y)=p$. Naheliegend ist der Ansatz

$$E(Y|X_1=x_1,\ldots,X_p=x_p)=\beta_0+\beta_1x_1+\ldots+\beta_px_p.$$

Da der Wertebereich von $p=E(Y|X_1=x_1,\ldots,X_p=x_p)$ aber das Intervall $(0,1)$ ist, ist dann nicht gesichert, dass diese Beziehung für alle Werte von $X_1,\ldots,X_p$ erfüllt ist. Daher wird die Abhängigkeit nicht direkt von der Linearkombination unterstellt, sondern vermittelt über eine nichtlineare Funktion. Am naheliegendsten sind spezielle Verteilungsfunktionen:

Die Verwendung der Normalverteilungsfunktion ergibt die sogenannten *Probit-Modelle*:

$$E(Y|X_1=x_1,\ldots,X_p=x_p)=\Phi(\beta_0+\beta_1x_1+\ldots+\beta_px_p).$$

Die Verteilungsfunktion der logistischen Verteilung führt auf die *Logit-Modelle*:

$$E(Y|X_1=x_1,\ldots,X_p=x_p)=\frac{1}{1+\exp(\beta_0+\beta_1x_1+\ldots+\beta_px_p)}.$$

Beispiel 3.2.8 *Herz-Kreislauf-Erkrankungen*

Bei Herz-Kreislauf-Erkrankungen spielt das Alter eine relevante Rolle. Bezeichnet die binäre Variable Y das Vorliegen einer solchen Erkrankung und X das Alter, so erhält man bei 100 Personen die graphische Darstellung für die Y-Werte in Abhängigkeit von X-Werten, siehe Abbildung 3.17 nach Hosmer & Lemeshow (1989, S.4).

Die Bildung von Altersgruppen und die Bestimmung des Anteils der Erkrankten pro Gruppe (Kreise in Abbildung 3.17) macht deutlich, dass der Anteil tatsächlich entsprechend der Form einer logistischen Funktion zunimmt.

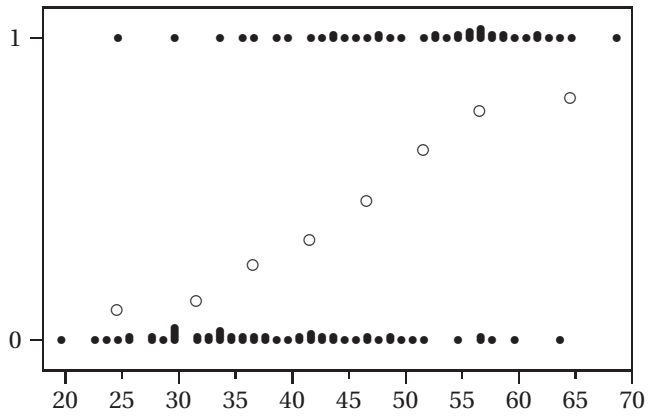


Abb. 3.17: *Herz-Kreislauf-Erkrankungen*

3.2.9 Log-Lineare Modelle für Kontingenztafeln

Es werden zwei nominal skalierte Zufallsvariablen X und Y betrachtet. Die I bzw. J Realisationsmöglichkeiten werden mit den Zahlen $i = 1,\ldots,I$ und $j = 1,\ldots,J$ identifiziert. Die gemeinsame Verteilung bildet dann die Grundlage für eine zweidimensionale Kontingenztafel:

$X\backslash Y$	1	2	$\cdots$	J
1	p_{11}	p_{12}	$\cdots$	p_{1J}
2	p_{21}	p_{22}	$\cdots$	p_{2J}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
I	p_{I1}	p_{I2}	$\cdots$	p_{IJ}

Da die Zufallsvariablen X und Y nominal skaliert sind, gibt es keine eindeutig zu präferierende Anordnung der Zeilen und Spalten. Das adäquate Verteilungsmodell für die gemeinsame Verteilung ist eine geeignete Multinomialverteilung. Ein einfaches, diese globale Erfassung einschränkendes Modell wäre das der Unabhängigkeit:

$$p_{ij} = \mathrm{P}(X=i, Y=j) = \mathrm{P}(X=i)\mathrm{P}(Y=j) = p_{i\cdot}p_{\cdot j} \quad i=1,\ldots,I,\ j=1,\ldots,J\,.$$

Darüber hinausgehend ist wie bei der zweifachen Varianzanalyse ggfs. die Erfassung von Überlagerungen der Effekte von Wertekombinationen von Interesse. Dazu geht man zu Logarithmen der erwarteten Zellhäufigkeiten $\mu_{ij} = n \cdot p_{ij}$ über und setzt genau wie bei der zweifachen Varianzanalyse das *log-lineare Modell* an:

$$\ln(\mu_{ij}) = \upsilon + \upsilon_{1(i)} + \upsilon_{2(j)} + \upsilon_{12(ij)} \qquad i=1,\ldots,I,\ j=1,\ldots,J.$$

Zunächst sehen wir, dass dies eine reine Reparametrisierung darstellt. Es wird keine zusätzliche Forderung an die Multinomialwahrscheinlichkeiten gestellt. Allerdings lassen sich über die Spezifizierungen der Effekte einsichtige Strukturen modellieren. So gilt $\ln(\mu_{ij}) = \upsilon + \upsilon_{1(i)} + \upsilon_{2(j)}$ genau dann, wenn $p_{ij} = p_{i\cdot}p_{\cdot j}$, $(i=1,\ldots,I,\ j=1,\ldots,J)$.

Bei höherdimensionalen Tafeln führt dieser Zugang zu einer sinnvollen Modellklasse, die sich in einheitlicher Weise analysieren lässt.

Beispiel 3.2.9 *Arbeitsplatzzufriedenheit*

Eine Stichprobe von 715 Arbeitern ergab die folgende drei-Wege-Tafel, vgl. Andersen (1989, S.4):

Qualität des Managements	Arbeitsplatzzufriedenheit des Vorgesetzten	eigene Arbeitsplatzzufriedenheit gering	hoch
schlecht	gering	103	87
	hoch	32	42
gut	gering	59	109
	hoch	78	205

Für eine Analyse des Zusammenhanges von eigener Arbeitsplatzzufriedenheit mit der des Vorgesetzten und der Qualität des Managements ist ein geeignetes Modell nötig. Sicherlich wäre die Annahme der Unabhängigkeit der drei binären Variablen wenig plausibel.

3.2.10 Generalisierte lineare Modelle

Bis auf die nichtlineare Regression haben alle betrachteten Modelle folgendes gemeinsam:

1. Die Beobachtungen Y_i $(i = 1,\ldots,n)$ haben Verteilungen, die alle zur gleichen Exponentialfamile gehören. Diese liegt in kanonischer Form vor:
$$f(y;\eta_i) = \exp[\eta_i \cdot y + d(\eta_i) + S(y)] \times I_A(y).$$
2. Die systematische Komponente wird gebildet durch eine Linearkombination $\beta_0+\beta_1 x_{1i} + \ldots + \beta_p x_{pi}$ der erklärenden Variablen $X_1,\ldots,X_p$.
3. Die Verbindung von zufälliger und systematischer Komponente, indem die Erwartungswerte μ_i der Y_i als Funktion des linearen Prädiktors angesetzt wird. Üblich ist jedoch die Angabe der Beziehung in der Form $g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$. Man bezeichnet g auch als Linkfunktion.

Modelle, die diese drei Bedingungen erfüllen, heißen *verallgemeinerte lineare Modelle* (generalized linear models).

Zwischen dem natürlichen Parameter und dem Erwartungswert sowie der Varianz σ^2 bestehen nach Satz 3.1.25 die Beziehungen

$$\mu_i = d'(\eta_i), \qquad \sigma^2(\mu_i) = d''(\eta_i).$$

Von besonderer Bedeutung sind Modelle mit natürlicher Linkfunktion, d.h. dass die Linearkombination gleich dem natürlichen Parameter ist:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}.$$

Bei der linearen Regression ist die zugrunde liegende Verteilung die Normalverteilung. Der natürliche Parameter ist gleich dem Erwartungswert und die Linkfunktion ist die Identität.

$$\mu_i = \eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}.$$

Die Binomialverteilung hat als natürlichen Parameter $\eta = \ln(p/(1-p))$. Der Ansatz der logistischen Regression,

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi},$$

entspricht gerade der natürlichen Linkfunktion. Bei Probit-Modellen liegt dagegen keine natürliche Linkfunktion vor.

Der natürlichen Linkfunktion entspricht auch die Wahl $\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})$ für den Parameter λ_i bei der Poisson-Regression. Denn der natürliche Parameter der Poisson-Verteilung ist gerade $\eta_i = \ln(\lambda_i)$, so dass

$$\begin{aligned} \eta_i = \ln(\lambda_i) &= \ln(\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})) \\ &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}. \end{aligned}$$

3.3 Aufgaben

Aufgabe 1

Welche der folgenden Verteilungen bilden eine Exponentialfamilie?

(i) Gleichverteilung: $f(x) = 1/\theta$ für $0 < x < 1$
(ii) geometrische Verteilung: $f(x) = p(1-p)^x$ für $x = 0, 1, 2, 3, \ldots$
(iii) Simpson-Verteilung: $f(x) = 1 - |x|$ für $-1 < x < 1$.
(iv) Pareto-Verteilung: $f(x) = \alpha k^\alpha x^{-(\alpha+1)}$ für $x > k$.

4 Stichproben und Statistiken

4.1 Stichproben aus endlichen Grundgesamtheiten

Stichproben-Ziehen ist von der Vorstellung her mit der Auswahl von Objekten aus einer Grundgesamtheit oder Population verbunden. Als Population gilt hier sowohl eine spezifizierte Bevölkerungsgruppe wie auch eine genau abgegrenzte Menge von Objekten, etwa die an einem Tag in einem Werk hergestellten Radioapparate eines speziellen Typs. Die Population wird mit dem Stichprobenraum Ω gleichgesetzt, die einzelnen Elemente werden wie sonst mit ω bezeichnet. Die Vorschrift, wie ein einzelnes Element aus Ω auszuwählen ist, legt eine Wahrscheinlichkeitsverteilung P auf einer geeigneten σ-Algebra $\mathfrak{A}$ von Ereignissen über Ω fest. $(\Omega, \mathfrak{A}, \mathrm{P})$ ist der Wahrscheinlichkeitsraum.

Da hier von einer endlichen Population Ω, $|\Omega| = N$, ausgegangen wird, ist es sinnvoll, für $\mathfrak{A}$ die Potenzmenge von Ω zu wählen, $\mathfrak{A} = \mathfrak{P}(\Omega)$. Weiter wird nur die einfache Zufallsauswahl betrachtet. Die genaue Fassung dieses Begriffs erfolgt gleich nach der Einführung einiger notwendiger Vereinbarungen. Bei Stichproben vom Umfang $n = 1$ führt das jedenfalls zu der *Laplace'schen Wahrscheinlichkeitsverteilung* P, die definiert ist durch

$$\mathrm{P}(A) = \frac{|A|}{N}.$$

Jede auf Ω definierte Zufallsvariable $X : \Omega \to \mathbb{R}$ ist wegen der Endlichkeit von Ω diskret. Mit der Laplace'schen Wahrscheinlichkeitsverteilung P gilt dann:

$$\mathrm{E}(X) = \sum_{i=1}^{I} x_i \mathrm{P}(X = x_i) = \sum_{i=1}^{I} x_i \frac{N(X = x_i)}{N} = \sum_{i=1}^{N} X(\omega_i) \cdot \frac{1}{N} = \mu. \tag{4.1}$$

Der Erwartungswert von X ist also gleich dem arithmetischen Mittel über alle Einzelwerte, oder, anders gesagt, gleich dem Populationsmittel.

Für die Varianz erhalten wir entsprechend:

$$\mathrm{Var}(X) = \sum_{i=1}^{I} (x_i - \mu)^2 \cdot \frac{N(X = x_i)}{N} = \sum_{i=1}^{N} (X(\omega_i) - \mu)^2 \cdot \frac{1}{N} = \sigma^2. \tag{4.2}$$

Auch die Varianz von X ist mit der Populationsvarianz identisch.

Nun soll eine Stichprobe vom Umfang n, d.h. n Elemente aus Ω zufällig gezogen werden. Der dem Gesamtexperiment zugrunde liegende Stichprobenraum ist dann die Menge $\Omega^{(n)}$ der n-Tupel $\omega^{(n)} = (\omega_1, \ldots, \omega_n)$, wenn das zuerst gezogene Element mit ω_1, das zweite mit ω_2 und das letzte schließlich mit ω_n bezeichnet wird.

Es werden als Grundlage zwei Ziehungsarten betrachtet: Das Ziehen mit und das ohne Zurücklegen. Wie wir aus Kapitel 1 wissen, sind dann beim Ziehen mit Zurücklegen N^n und beim Ziehen ohne Zurücklegen $N(N-1)\cdot\ldots\cdot(N-n+1)$ unterschiedliche Stichproben möglich. Stichproben mit gleichem ω an unterschiedlichen Stellen können beim Ziehen ohne Zurücklegen nicht vorkommen. Das wird erfasst, indem solchen Stichproben die Wahrscheinlichkeit null zugeordnet wird.

Definition 4.1.1 *einfache Zufallsauswahl*

Sei Ω eine endliche Menge mit N Elementen und sei $\Omega^{(n)} = \{(\omega_1,\ldots,\omega_n)|\,\omega_i \in \Omega, i = 1,\ldots,n\}$. Die Wahrscheinlichkeitsverteilung P über $(\Omega^{(n)}, \mathfrak{P}(\Omega^{(n)}))$ charakterisiert eine einfache *Zufallsauswahl mit Zurücklegen*, wenn jedem Element aus $\Omega^{(n)}$ die Wahrscheinlichkeit N^{-n} zugeordnet ist:

$$\mathrm{P}(\{(\omega_1,\ldots,\omega_n)\}) = \frac{1}{N^n} \qquad \text{für alle } (\omega_1,\ldots,\omega_n) \in \Omega^{(n)}. \tag{4.3}$$

P charakterisiert eine *einfache Zufallsauswahl ohne Zurücklegen*, wenn für alle Elemente $(\omega_1,\ldots,\omega_n) \in \Omega^{(n)}$ gilt:

$$\mathrm{P}(\{(\omega_1,\ldots,\omega_n)\}) = \begin{cases} \dfrac{1}{N(N-1)\cdots(N-n+1)} & \text{falls } \omega_i \neq \omega_j \text{für } i \neq j \\ 0 & \text{sonst}. \end{cases} \tag{4.4}$$

Im Weiteren werden beide Fälle unter dem Begriff *einfache Zufallsauswahl* subsumiert.

Das Ziehen mit Zurücklegen wird offensichtlich durch die Produktverteilung erfasst. Damit sind die einzelnen Stichprobenzüge unabhängig. Beim Ziehen ohne Zurücklegen sind sie dagegen abhängig. Wir werden sehen, dass das einen etwas aufwendigeren Formalismus mit sich bringt.

Satz 4.1.2 *Wahrscheinlichkeiten spezieller Ereignisse*

Bei der einfachen Zufallsauswahl gilt für alle Ereignisse A_ω der Form $\{\omega_1\}\times\Omega^{(n-1)}$, $\Omega^{(n-1)}\times\{\omega_n\}$, $\Omega^{(i-1)}\times\{\omega_i\}\times\Omega^{(n-i)}$, $\omega\in\Omega$, $1<i<n$:

$$\mathrm{P}(A_\omega) = \frac{1}{N}.$$

Beweis: Beim Ziehen mit Zurücklegen ist offensichtlich

$$\mathrm{P}\left(\Omega^{(i-1)}\times\{\omega_i\}\times\Omega^{(n-i)}\right) = \frac{N^{n-1}}{N^n} = \frac{1}{N^n}.$$

Bei der einfachen Zufallsauswahl ohne Zurücklegen folgt, wenn wir berücksichtigen, dass bei dem Ereignis $\Omega^{(i-1)}\times\{\omega\}\times\Omega^{(n-i)}$ noch $n-1$ Elemente aus $N-1$ 'freien' zu ziehen sind:

$$\mathrm{P}\left(\Omega^{(i-1)}\times\{\omega_i\}\times\Omega^{(n-i)}\right) = \frac{(N-1)!/(N-1-(n-1))!}{N!/(N-n)!} = \frac{1}{N}.$$

Um den Unterschied zwischen dem Ziehen mit und ohne Zurücklegen zu quantifizieren, betrachten wir die maximale Differenz von Wahrscheinlichkeiten für die möglichen Ereignisse unter den beiden Ziehungsarten der einfachen Zufallsauswahl.

Satz 4.1.3 *Differenz der Wahrscheinlichkeiten bei der Zufallsauswahl mit bzw. ohne Zurücklegen*

Sei $\Omega^{(n)}$ wie in Definition 4.1.1 gegeben und seien P und Q die mit der einfachen Zufallsauswahl mit bzw. ohne Zurücklegen assoziierten Wahrscheinlichkeitsverteilungen. Dann gilt

$$\|\mathrm{P}-\mathrm{Q}\| = \sup_{A\subseteq\Omega^{(n)}} |\mathrm{P}(A)-\mathrm{Q}(A)| = 1 - \frac{N(N-1)\cdot\ldots\cdot(N-n+1)}{N^n}$$

bzw.

$$1-\exp\left[-\frac{n(n-1)}{2N}\right] < \|\mathrm{P}-\mathrm{Q}\| < \frac{n(n-1)}{2N}.$$

Beweis: Sei $\Delta\subseteq\Omega^{(n)}$ die Teilmenge, die aus den n-Tupeln besteht, welche keine gleichen Elemente an verschiedenen Stellen enthalten. P ist konstant gleich $1/N^n$ für alle einelementigen Teilmengen, und Q ist konstant gleich $1/(N(N-1)\cdot\ldots\cdot(N-n+1))$ für einelementige Teilmengen aus Δ und gleich null für die anderen. Damit folgt:

$$\|\mathrm{P}-\mathrm{Q}\| = \|\mathrm{P}(\Delta)-\mathrm{Q}(\Delta)\| = \mathrm{Q}(\Delta)-\mathrm{P}(\Delta) = 1-\mathrm{P}(\Delta) = 1-\frac{N(N-1)\cdot\ldots\cdot(N-n+1)}{N^n}.$$

Aus den bekannten Ungleichungen

$$1-\sum_{i=1}^{n-1} x_i < \prod_{i=1}^{n-1}(1-x_i) < \exp\left[-\sum_{i=1}^{n-1} x_i\right]$$

für reelle Zahlen $x_i\in(0,1)$, $(i=1,\ldots,n)$ ergibt sich, dass folgende Schranken für

$$\frac{N(N-1)\cdot\ldots\cdot(N-n+1)}{N^n} = -\prod_{i=1}^{n-1}\left(1-\frac{i}{N}\right)$$

gelten:

$$\text{Nach unten:}\qquad 1-\sum_{i=1}^{n-1}\frac{i}{N} = 1-\frac{n(n-1)}{2N};$$

$$\text{nach oben:}\qquad \exp\left[-\sum_{i=1}^{n-1}\frac{i}{N}\right] = \exp\left[-\frac{n(n-1)}{2N}\right].$$

Das ergibt die Schranken für $1-N(N-1)\cdot\ldots\cdot(N-n+1)/N^n$.

Als eine numerische Illustration des Satzes sei $N = 100\,000\,000$. Dann gilt für $n = 1\,000$: $0.00498 < \|\mathrm{P}-\mathrm{Q}\| < 0.005$; und für $n = 5\,000$ gilt: $0.117 < \|\mathrm{P}-\mathrm{Q}\| < 0.125$. Bei größeren Stichprobenumfängen kann es also bzgl. des betrachteten Abstandes einen größeren Unterschied geben. Wahrscheinlichkeiten für speziellere Ereignisse brauchen natürlich nicht so stark zu differieren.

Um die wiederholte Beobachtung einer Zufallsvariablen formal zu erfassen, sei X eine auf $(\Omega,\mathfrak{P}(\Omega))$ definierte Zufallsvariable, wobei wir uns zunächst auf den *eindimensionalen Fall*

beschränken. x_1 bezeichne den Wert von X beim ersten Stichprobenelement, also den Wert $X(\omega_1)$. x_2 sei der Wert von X beim zweiten Stichprobenelement, $x_2 = X(\omega_2)$, usw. bis x_n schließlich den Wert von X bei ω_n, dem Ergebnis der n'ten Stichprobenziehung, $x_n = X(\omega_n)$, beschreibt. Die Beobachtungen $x_1, \ldots, x_n$ sollen nun als Realisationen von Zufallsvariablen aufgefasst werden. Das führt zu folgender

Definition 4.1.4 *Stichprobenvariable*

Sei X eine Zufallsvariable auf $(\Omega, \mathfrak{P}(\Omega))$. Die *Stichprobenvariablen* $X_1, \ldots, X_n$ sind die auf $(\Omega^{(n)}, \mathfrak{P}(\Omega^{(n)}))$ folgendermaßen definierten Zufallsvariablen: Für $\omega^{(n)} = (\omega_1, \omega_2, \ldots, \omega_n) \in \Omega^{(n)}$ sind

$$X_i(\omega^{(n)}) = X_i(\omega_1, \omega_2, \ldots, \omega_n) := X(\omega_i) \qquad i = 1, \ldots, n\,. \tag{4.5}$$

Es sollen jetzt die Verteilungen der Stichprobenvariablen bei der einfachen Zufallsauswahl bestimmt werden. Dazu werden als Hilfsgrößen Indikatorvariablen auf $\Omega^{(n)}$ eingeführt. Zu ihrer Definition werden die Elemente von Ω mit $\omega^1, \ldots, \omega^j, \ldots, \omega^N$ bezeichnet. Dann sei für $\omega^{(n)} = (\omega_1, \ldots, \omega_i, \ldots, \omega_n)$:

$$Z_{ij}(\omega^{(n)}) = \begin{cases} 1 & \text{falls } \omega_i = \omega^j \\ 0 & \text{sonst} \end{cases}.$$

Lemma 4.1.5 *Verteilung der Z_{ij}*

Bei der einfachen Zufallsauswahl sind die Z_{ij} jeweils Bernoulli-verteilt mit $p = 1/N$. Zudem gilt für die Kovarianz von Z_{ij} und Z_{uv} :

$$\mathrm{Cov}(Z_{ij}, Z_{uv}) = \begin{cases} (N-1)/N^2 & \text{falls } i = u,\ j = v \\ -1/N^2 & \text{falls } i = u,\ j \neq v \\ -f/N^2 & \text{falls } i \neq u,\ j = v \\ f/N^2(N-1) & \text{falls } i \neq u,\ j \neq v\,, \end{cases}$$

wobei

$$f = \begin{cases} 1 & \text{im Fall des Ziehens ohne Zurücklegen} \\ 0 & \text{im Fall des Ziehens mit Zurücklegen} \end{cases}.$$

Beweis: Sei das Element ω^j an der i'ten Stelle der Stichprobe fixiert. Dann gilt nach Satz 4.1.2:

$$\mathrm{P}(Z_{ij} = 1) = \frac{1}{N}.$$

Da die Z_{ij} $0-1-$Variablen sind, gilt

$$\begin{aligned} \mathrm{Cov}(Z_{ij}, Z_{uv}) &= \mathrm{P}(Z_{ij} Z_{uv} = 1) - \mathrm{P}(Z_{ij} = 1) \cdot \mathrm{P}(Z_{uv} = 1) \\ &= \mathrm{P}(Z_{ij} = 1, Z_{uv} = 1) - \mathrm{P}(Z_{ij} = 1) \cdot \mathrm{P}(Z_{uv} = 1). \end{aligned}$$

Damit ergeben sich die behaupteten Kovarianzen mit ähnlichen kombinatorischen Überlegungen wie sie der Herleitung der Randverteilungen der Z_{ij} zugrunde liegen.

Die Stichprobenvariablen X_i können mit Hilfe der Indikatorvariablen Z_{ij} ausgedrückt werden:

$$X_i = \sum_{j=1}^{N} X(\omega^j) \cdot Z_{ij}. \tag{4.6}$$

Diese Darstellung erlaubt die einfache Bestimmung der Randverteilung von X_i.

Satz 4.1.6 *Verteilung der Stichprobenvariablen*

Die Stichprobenvariablen X_i sind bei der einfachen Zufallsauswahl identisch verteilt wie X.

Beweis: Seien $x_1, \ldots, x_I$ die Realisationsmöglichkeiten der zugrunde liegenden Zufallsvariablen X. Sei die Realisationsmöglichkeit X festgehalten. Mit $X^{-1}(x) = \{\omega^{j_1}, \ldots, \omega^{j_L}\}$ ist

$$\mathrm{P}(X = x) = \mathrm{P}(\{X^{-1}(x)\}) = \mathrm{P}(\{\omega^{j_1}, \ldots, \omega^{j_L}\}) = \frac{L}{N}.$$

Wegen der Disjunktheit der Ereignisse $\{Z_{ij_k} = 1\},\ k = 1, \ldots, L$, gilt auch:

$$\mathrm{P}(X_i = x) = \mathrm{P}(\{Z_{ij_1} = 1\} \cup \ldots \cup \{Z_{ij_L} = 1\}) = \mathrm{P}(Z_{ij_1} = 1) + \ldots + \mathrm{P}(Z_{ij_L} = 1) = \frac{L}{N}.$$

Aus dem Satz ergibt sich insbesondere, dass die Stichprobenvariablen auch den gleichen Erwartungswert und die gleiche Varianz wie die Zufallsvariable X haben.

Ziel bei der Stichprobenziehung ist i.d.R. die Ermittlung eines Näherungs- oder Schätzwertes für das Populationsmittel aus der Stichprobe. Das ist ein Spezialfall der in Kapitel 6 zu behandelnden Fragestellung der *Schätzung eines Parameters*. Die Realisationen der Stichprobenvariablen $X_1, \ldots, X_n$ werden dazu verdichtet, d.h. es werden zu $x_1, \ldots, x_n$ gehörige Maßzahlen, speziell das arithmetische Mittel und die empirische Varianz, bestimmt. Dem liegt die Bildung von Stichprobenfunktionen oder Statistiken zugrunde.

Definition 4.1.7 *Stichprobenfunktion*

Sei T eine auf $\mathbb{R}^n$ definierte Abbildung und $X_1, \ldots, X_n$ auf $(\Omega^{(n)}, \mathfrak{P}(\Omega^{(n)}))$ definierte Stichprobenvariablen. Die durch

$$\begin{aligned} T(X_1, \ldots, X_n) : \Omega^{(n)} &\to \mathbb{R}^k \\ \omega^{(n)} &\to T((X_1, \ldots, X_n)(\omega^{(n)})) = T(X_1(\omega^{(n)}), \ldots, X_n(\omega^{(n)})) \end{aligned} \tag{4.7}$$

definierte Zufallsvariable $T(X_1, \ldots, X_n)$ heißt eine k-dimensionale *Stichprobenfunktion* oder *Statistik*.

Beispiel 4.1.8 *Stichprobenfunktionen arithmetisches Mittel und empirische Varianz*

Zwei spezielle Stichprobenfunktionen sind etwa das arithmetische Mittel und die empirische Varianz:

$$T_1(X_1,\ldots,X_n)=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}$$

und

$$T_2(X_1,\ldots,X_n)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2=S^2.$$

im Folgenden soll gezeigt werden, dass die Verwendung des arithmetischen Mittels $\bar{X}$ Schätzwerte für μ liefert, die um den wahren Parameter konzentriert sind, d.h. dass $\mathrm{E}(\bar{X})=\mu$ gilt. Zudem werden die Varianzen von $\bar{X}$ bestimmt.

Satz 4.1.9 *$\bar{X}$ bei der einfachen Zufallsauswahl*

Für die Stichprobenfunktion $\bar{X}$ gilt bei der einfachen Zufallsauswahl:

$$\mathrm{E}(\bar{X})=\mu;\qquad \mathrm{Var}(\bar{X})=\frac{\sigma^2}{n}\cdot\left[1-f\cdot\frac{n-1}{N-1}\right];$$

wobei wieder

$$f=\begin{cases}1 & \text{im Fall des Ziehens ohne Zurücklegen}\\ 0 & \text{im Fall des Ziehens mit Zurücklegen}\end{cases}.$$

Beweis: In beiden Fällen gilt für den Erwartungswert

$$\mathrm{E}(\bar{X})=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$$

und für die Varianz

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\left[\sum_{i=1}^{n}\mathrm{Var}(X_i)+\sum_{i\neq j}\mathrm{Cov}(X_i,X_j)\right]\\ &= \frac{1}{n}\sigma^2+\frac{1}{n^2}\cdot\sum_{i\neq j}\mathrm{Cov}(X_i,X_j).\end{aligned}$$

Ziehen mit und ohne Zurücklegen sind nun getrennt zu betrachten. Mit der Darstellung der Stichprobenvariablen mittels der Z_{ij} erhalten wir für das Ziehen mit Zurücklegen unmittelbar:

$$\mathrm{Cov}(X_i,X_j)=0\qquad\text{falls } i\neq j.$$

Damit ist

$$\mathrm{Var}(\bar{X})=\frac{1}{n}\sigma^2.$$

Beim Ziehen ohne Zurücklegen gilt für $i \neq j$:

$$\begin{aligned}
\mathrm{Cov}(X_i, X_j) &= \mathrm{Cov}(X(\omega^u)Z_{iu}, X(\omega^v)Z_{jv}) = \sum_{u=1}^{N}\sum_{v=1}^{N} X(\omega^u)\cdot X(\omega^v)\cdot \mathrm{Cov}(Z_{iu}, Z_{jv}) \\
&= \sum_{\substack{u,v=1\\ u=v}} X(\omega^u)\cdot X(\omega^v)\cdot\frac{-1}{N^2} + \sum_{\substack{u,v=1\\ u\neq v}} X(\omega^u)\cdot X(\omega^v)\cdot\frac{1}{N^2(N-1)} \\
&= \frac{1}{N-1}\sum_{u,v=1} X(\omega^u)\cdot X(\omega^v)\cdot\frac{1}{N^2} - \left(\frac{1}{N-1}+1\right)\sum_{u=1}^{N} X(\omega^u)^2\cdot\frac{1}{N^2} \\
&= -\frac{1}{N-1}\left[\sum_{u=1}^{N} X(\omega^u)^2\cdot\frac{1}{N} - \sum_{u=1}^{N}\sum_{v=1}^{N} X(\omega^u)\cdot X(\omega^v)\cdot\frac{1}{N^2}\right] = -\frac{1}{N-1}\sigma^2.
\end{aligned}$$

Das ergibt:

$$\mathrm{Var}(\bar{X}) = \frac{1}{n}\sigma^2 - \frac{n(n-1)}{N-1}\sigma^2 = \frac{\sigma^2}{n}\left[1-\frac{n-1}{N-1}\right].$$

Die Realisationen von $\bar{X}$ sind bei der einfachen Zufallsauswahl also um das Populationsmittel μ zentriert. Die Varianz gibt dann an, wie eng sie um μ konzentriert sind. Offenbar ist das für die Praxis relevantere Ziehen ohne Zurücklegen dem Vorgehen mit Zurücklegen vorzuziehen. Von Bedeutung ist das Modell des Ziehens mit Zurücklegen vor allem wegen seiner leichteren Handhabbarkeit. Zudem lässt sich das Ziehen ohne Zurücklegen bei genügend großem N durch das mit Zurücklegen annähern.

Für die Anwendung ist eine Abschätzung der Standardabweichung $\sqrt{\mathrm{Var}(\bar{X})}$ der Schätzung $\bar{X}$ von großer Bedeutung. Mit

$$\begin{aligned}
\mathrm{E}\left(\sum(X_i-\bar{X})^2\right) &= \mathrm{E}\left(\sum((X_i-\mu)-(\bar{X}-\mu))^2\right) \\
&= \sum \mathrm{E}(X_i-\mu)^2 - 2\sum \mathrm{E}((X_i-\mu)(\bar{X}-\mu)) + n\cdot\mathrm{E}(\bar{X}-\mu)^2 \\
&= n\cdot\sigma^2 - 2\sum \mathrm{E}((X_i-\mu)(\bar{X}-\mu)) + n\cdot\frac{\sigma^2}{n}\left[1-\frac{n-1}{N-1}\right]
\end{aligned}$$

und

$$\mathrm{E}((X_i-\mu)(\bar{X}-\mu)) = \frac{1}{n}\sum_j (X_i-\mu)(X_j-\mu) = \frac{\sigma^2}{n} - \frac{f}{n}\cdot\frac{n-1}{N-1}\cdot\frac{\sigma^2}{n}$$

folgt:

$$\begin{aligned}
\mathrm{E}\left(\sum(X_i-\bar{X})^2\right) &= n\cdot\sigma^2 - 2n\frac{\sigma^2}{n} - \frac{f}{n}\cdot\frac{n-1}{N-1}\cdot\frac{\sigma^2}{n} + n\cdot\frac{\sigma^2}{n}\left[1-\frac{n-1}{N-1}\right] \\
&= (n-1)\left(1-\frac{f}{N-1}\right)\sigma^2.
\end{aligned}$$

Da i.d.R. $f/(N-1)$ klein ist, sind die Werte von $S^2/n = \sum(X_i-\bar{X})^2/n(n-1)$ bei $\sigma^2/n = \mathrm{Var}(\bar{X})$ zentriert und können als Basis für eine Schätzung der Standardabweichung von $\bar{X}$ verwendet werden.

Beispiel 4.1.10 *Zufriedenheit mit einer Klinik*

Für eine Befragungsaktion bzgl. der (Un-)Zufriedenheit mit einer Klinik sollen aus den 611 Patienten, die in einem Halbjahr auf einer Fachabteilung der Klinik gelegen haben, 50 zufällig ausgewählt werden. Von Interesse ist dabei auch die Liegezeit der Patienten, da diese voraussichtlich einen Einfluss auf die Zufriedenheit hat. Aus den Krankenakten weiß man bzgl. der Liegedauer X (in Tagen) der 611 Patienten: $\mu_X = 17.282, \sigma_X = 17.037$.

Für $\bar{X}$ gilt dann:

$$\mathrm{Var}(\bar{X}) = \frac{17.037^2}{50}\left(1 - \frac{49}{610}\right) = 5.339.$$

Beim Ziehen mit Zurücklegen würden wir für die Varianz von $\bar{X}$ den Wert 5.805 erhalten.

4.2 Die mathematische Stichprobe

Bei sehr großen Grundgesamtheiten ist die Abhängigkeit der Stichprobenvariablen beim Ziehen mit Zurücklegen vernachlässigbar und die Verhältnisse sind mit denen beim Ziehen ohne Zurücklegen vergleichbar. Da offensichtlich die Behandlung von Stichproben sehr viel einfacher ist, wenn das Modell 'Mit Zurücklegen' betrachtet wird, bildet es den Ausgangspunkt für die Verallgemeinerung von Stichproben. Darauf basiert auch die Formalisierung der ‚wiederholten unabhängigen Beobachtung einer Zufallsvariablen'.

Ausgangspunkt bildet ein Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, \mathrm{P})$. Die n-malige Durchführung des Zufallsexperimentes führt auf den Stichprobenraum

$$\Omega^{(n)} = \{\omega | \, \omega = (\omega_1, \ldots, \omega_n),\ \omega_i \in \Omega,\ i = 1, \ldots, n\}.$$

Über $\Omega^{(n)}$ muss eine σ-Algebra $\mathfrak{A}^{(n)}$ definiert sein. Diese sollte aus Ereignissen bestehen, welche in geeigneter Weise aus denen des einfachen Zufallsexperimentes zusammengesetzt sind. Wir erhalten $\mathfrak{A}^{(n)}$ folgendermaßen: Für jeweils n Ereignisse $A_1 \in \mathfrak{A}, \ldots, A_n \in \mathfrak{A}$ sei

$$A_1 \times A_2 \times \ldots \times A_n = \{\omega^{(n)} | \, \omega^{(n)} \in \Omega^{(n)},\ \omega_1 \in A_1,\ \omega_2 \in A_2, \ldots, \omega_n \in A_n\}.$$

Dieses ist ein zusammengesetztes Ereignis: A_1 tritt bei der ersten Durchführung des Experimentes ein, A_2 bei der zweiten, bis schließlich A_n bei der n'ten Durchführung eintritt. Ereignisse dieser Form werden als *Rechteck-Ereignisse* bezeichnet.

Die benötigte σ-Algebra $\mathfrak{A}^{(n)}$ ist nun die kleinste σ-Algebra, die alle Rechteck-Ereignisse umfasst. Diese Konstruktion entspricht der der σ-Algebra der Borelschen Mengen, welche als kleinste σ-Algebra, die alle Intervalle umfasst, definiert ist.

Schließlich ist eine Wahrscheinlichkeitsverteilung auf $\mathfrak{A}^{(n)}$ zu definieren. Die Verteilung wird mit $\mathrm{P}^{(n)}$ bezeichnet. Für jedes Rechteck-Ereignis $A_1 \times A_2 \times \ldots \times A_n \in \mathfrak{A}^{(n)}$ wird

$$\mathrm{P}^{(n)}(A_1 \times A_2 \times \ldots \times A_n) = \prod_{i=1}^{n} \mathrm{P}(A_i) \tag{4.8}$$

gesetzt. Der Grund für diese Definition liegt auf der Hand. Das Rechteck-Ereignis beschreibt das Eintreten verschiedener einfacher Ereignisse in den einzelnen Durchführungen des Zufallsexperimentes. Diese sollten unabhängig sein, was durch diese Definition gerade erreicht

wird. Formal lässt sich das so sehen: Bzgl. $\Omega^{(n)}$ ist das Ereignis ‚A_i tritt an i'ter Stelle ein' zu beschreiben durch

$$A_{i,i}^{(n)} = \Omega \times \ldots \times \Omega \times A_i \times \Omega \times \ldots \times \Omega.$$

Somit gilt offensichtlich

$$A_1 \times A_2 \times \ldots \times A_n = \bigcap_{i=1}^{n} A_{i,i}^{(n)},$$

und es ist in Übereinstimmung mit der Definition der Unabhängigkeit von Ereignissen:

$$\mathrm{P}^{(n)}\left(\bigcap_{i=1}^{n} A_{i,i}^{(n)}\right) = \mathrm{P}^{(n)}(A_1 \times A_2 \times \ldots \times A_n) = \prod_{i=1}^{n} \mathrm{P}(A_i) = \prod_{i=1}^{n} \mathrm{P}^{(n)}(A_{i,i}^{(n)}).$$

Damit ist $\mathrm{P}^{(n)}$ für Rechteck-Ereignisse definiert. Es ist nun möglich, die Definition geeignet und eindeutig auf die σ-Algebra $\mathfrak{A}^{(n)}$ fortzusetzten. Speziell für die Eindeutigkeit ist notwendig, dass $\mathfrak{A}^{(n)}$ die kleinste σ-Algebra ist, die alle Rechteck-Ereignisse enthält. Die so auf $\mathfrak{A}^{(n)}$ definierte Wahrscheinlichkeitsverteilung wird wieder mit $\mathrm{P}^{(n)}$ bezeichnet.

Definition 4.2.1 *Produktraum und Produktverteilung*

Gegeben sei ein Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, \mathrm{P})$. Dann heißt $(\Omega^{(n)}, \mathfrak{A}^{(n)}, \mathrm{P}^{(n)})$ der *Produktraum* und speziell $\mathrm{P}^{(n)}$ die *Produktverteilung* von P.

Beispiel 4.2.2 *Laplacesche Wahrscheinlichkeitsverteilung*

Die einfache Zufallsauswahl mit Zurücklegen ist gerade durch die Produktverteilung auf $(\Omega^{(n)}, \mathfrak{P}(\Omega^{(n)}))$ charakterisiert, wobei von der Laplaceschen Wahrscheinlichkeitsverteilung über $(\Omega, \mathfrak{P}(\Omega))$ ausgegangen wird.

Stichprobenvariablen $X_1, \ldots, X_n$ sind auf $(\Omega^{(n)}, \mathfrak{A}^{(n)})$ nun ebenso definiert wie im Fall der einfachen Zufallsauswahl. Ist X eine Zufallsvariable auf $(\Omega, \mathfrak{A}, \mathrm{P})$, so sind die Zufallsvariablen $X_1, \ldots, X_n$ auf $(\Omega^{(n)}, \mathfrak{A}^{(n)})$ definiert durch: Für $\omega^{(n)} = (\omega_1, \omega_2, \ldots, \omega_n) \in \Omega^{(n)}$ sind

$$X_i(\omega^{(n)}) = X_i(\omega_1, \omega_2, \ldots, \omega_n) := X(\omega_i) \qquad i = 1, \ldots, n. \tag{4.9}$$

Satz 4.2.3 *Verteilung der Stichprobenvariablen*

Sei X eine Zufallsvariable auf $(\Omega, \mathfrak{A}, \mathrm{P})$. Auf $(\Omega^{(n)}, \mathfrak{A}^{(n)})$ sei die Produktverteilung $\mathrm{P}^{(n)}$ gegeben. Dann sind die Stichprobenvariablen $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit der gleichen Verteilungsfunktion wie X.

Beweis: Es gilt

$$\{X_i \le x_i\} = \underbrace{\Omega \times \ldots \times \Omega}_{i-1} \times \{X \le x_i\} \times \underbrace{\Omega \times \ldots \times \Omega}_{n-i}.$$

Damit sind die X_i offensichtlich Zufallsvariablen. Aufgrund der Definition von $\mathrm{P}^{(n)}$ gilt zudem:

$$\mathrm{P}^{(n)}(X_i \le x_i) = \mathrm{P}(X \le x_i).$$

Damit sind die Stichprobenvariablen identisch verteilt.

Die Unabhängigkeit folgt ebenso leicht:

$$\mathrm{P}^{(n)}(\{X_i \le x_i\}) = \mathrm{P}^{(n)}(\{X_1 \le x_1\} \cap \ldots \cap \{X_n \le x_n\}) = \mathrm{P}^{(n)}(\{X \le x_1\} \times \ldots \times \{X \le x_n\})$$
$$= \mathrm{P}(\{X \le x_1\}) \cdot \ldots \cdot \mathrm{P}(\{X \le x_n\}) = \prod_{i=1}^{n} \mathrm{P}^{(n)}(\{X_i \le x_i\}).$$

Es ist schließlich möglich, in diesem Rahmen auch unendliche Stichproben und Folgen von Stichprobenvariablen zu betrachten. Dies führt zu einem Stichprobenraum $\Omega^{(\infty)}$, der Menge aller abzählbaren Folgen

$$\omega^{(\infty)} = (\omega_1, \omega_2, \ldots, \omega_n, \ldots)$$

von Elementen aus Ω. Die σ-Algebra $\mathfrak{A}^{(\infty)}$ ist die kleinste σ-Algebra, welche alle Teilmengen der Form

$$A_1 \times A_2 \times \ldots \times A_n \times \Omega \times \Omega \times \ldots$$

enthält, wobei $A_1 \in \mathfrak{A}, \ldots, A_n \in \mathfrak{A}$ gilt und n eine beliebige ganze Zahl ist. Für solche Rechteck-Ereignisse ist die Wahrscheinlichkeitsverteilung $\mathrm{P}^{(\infty)}$ über $\mathfrak{A}^{(\infty)}$ definiert durch

$$\mathrm{P}^{(\infty)}(A_1 \times A_2 \times \ldots \times A_n \times \Omega \times \Omega \times \ldots) = \prod_{i=1}^{n} \mathrm{P}(A_i).$$

Unendliche Folgen von Stichprobenvariablen, die unabhängige Beobachtungen der Zufallsvariablen X beschreiben, lassen sich entsprechend einführen.
Für jedes $\omega^{(\infty)} = (\omega_1, \omega_2, \ldots \omega_n, \ldots)$ aus $\Omega^{(\infty)}$ wird

$$X_n(\omega^{(\infty)}) := X(\omega_n) \qquad n = 1, 2, 3, \ldots$$

gesetzt. Wie oben erhalten wir, dass die X_n identisch verteilt und unabhängig sind.

Die Stichprobenvariablen stellen das Wesentliche dar, was uns die Formalisierung der ‚wiederholten, unabhängigen Beobachtung einer Zufallsvariablen' erbracht hat. Wir heben dies durch die folgende Definition hervor. Wie meist auch im Weiteren machen wir den Stichprobenraum, auf dem die X_i definiert sind, nicht mehr extra als Produktraum kenntlich, sondern begnügen uns mit der knapperen Schreibweise $(\Omega, \mathfrak{A}, \mathrm{P})$.

Definition 4.2.4 *mathematische Stichprobe*

Eine *mathematische Stichprobe* ist eine Folge $X_1, \ldots, X_n$ von identisch verteilten unabhängigen Zufallsvariablen über einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, \mathrm{P})$. $X_1, \ldots, X_n$ werden auch als Stichprobenvariablen bezeichnet.

Als Sprechweisen vereinbaren wir, dass $X_1, \ldots, X_n$ oder $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe aus einer Grundgesamtheit mit der Verteilung $F(x)$, oder kurz aus der Verteilung $F(x)$, genannt wird, wenn die X_i unabhängig und identisch verteilt mit der Verteilungsfunktion $F(x)$ sind. Das Gleiche drücken wir auch dadurch aus, dass wir $X_1, \ldots, X_n$ als Stichprobenvariablen mit einer Verteilungsfunktion $F(x)$ bezeichnen.

Wie bei der einfachen Zufallsauswahl ist die Betrachtung von Stichprobenfunktionen oder Statistiken für die weiteren Betrachtungen zentral. Zwei bereits bekannte Stichprobenfunktionen sind das arithmetische Mittel und die empirische Varianz. Eine weitere Klasse von Stichprobenfunktionen bilden die *Ordnungs-* oder *geordneten Statistiken* $X_{i:n}$ $(i = 1,\ldots,n)$ wobei $x_{i:n}$ den i'ten Wert in der geordneten Stichprobe $x_{1:n} \leq \ldots \leq x_{n:n}$ bezeichnet.

Von entscheidender Bedeutung für die Anwendung vieler statistischer Verfahren ist die Kenntnis der Verteilung von speziellen Stichprobenfunktionen oder Statistiken. Dabei sind einmal die Verteilungen bei festem n von Interesse. Diese hängen von der jeweiligen Verteilung der Stichprobenvariablen ab. Somit sind hier nur Spezialresultate zu erzielen. Zum anderen sind asymptotische Verteilungen von Bedeutung. Für $n \to \infty$ zeigt sich nämlich, dass die Verteilung der Stichprobenvariablen u.U. nur eine geringe Rolle für das asymptotische Verteilungsgesetz spielt. Davon handelt das nächste Kapitel. Wir betrachten hier noch einige spezielle Stichprobenfunktionen.

Definition 4.2.5 *k-dimensionale Stichprobenfunktion und Stichprobenverteilung*

Sei $T(\boldsymbol{X}) = T(X_1,\ldots,X_n)$ eine k-dimensionale *Stichprobenfunktion* oder *Statistik*, d.h. T ist eine auf $\mathbb{R}^n$ definierte Abbildung, $X_1,\ldots,X_n$ sind auf $(\Omega,\mathfrak{A},\mathrm{P})$ definierte k-dimensionale Stichprobenvariablen und $T(\boldsymbol{X}) : \Omega \to \mathbb{R}^k$ ist definiert durch

$$\omega \to T(\boldsymbol{X})(\omega) = T(X_1(\omega),\ldots,X_n(\omega)). \tag{4.10}$$

Die Verteilung der Stichprobenfunktion $T(\boldsymbol{X})$ wird auch ihre *Stichprobenverteilung* genannt.

Wir bestimmen als erstes die Verteilung einer einzelnen Ordnungsstatistik $X_{k:n}$. Wegen der Bedeutung dieser Stichprobenverteilung machen wir uns zunächst heuristisch klar, welche Gestalt sie hat. Die zugrunde liegende Zufallsvariable X habe dazu die Verteilungsfunktion $F(x)$ und die Dichte $f(x)$, sei also insbesondere stetig. Das Eintreten des Ereignisses $\{x < X_{k:n} \leq x + dx\}$ ist dann für kleines dx gleichwertig damit, dass

$k-1$ Stichprobenvariablen einen Wert $\leq x$ annehmen,

1 Stichprobenvariable einen Wert im Intervall $(x; x+dx]$ annimmt, und

$n-k$ Stichprobenvariablen Werte $> x+dx$ annehmen.

Es gibt $n!/((k-1)!(n-k)!)$ Möglichkeiten, die n Beobachtungen auf die drei Intervalle zu verteilen. Die Wahrscheinlichkeit für eine dieser Möglichkeiten ist approximativ:

$$F(x)^{k-1} \cdot (f(x) \cdot dx) \cdot (1 - F(x+dx))^{n-k}.$$

Damit ergibt sich für die Dichte von $X_{k:n}$:

$$\begin{aligned} f_{X_{k:n}}(x) &= \lim_{dx \to 0} \frac{P(x < X_{k:n} \leq x + dx)}{dx} \\ &= \lim_{dx \to 0} \frac{n!}{(k-1)!1!(n-k)!} \cdot \frac{F(x)^{k-1} \cdot (f(x) \cdot dx) \cdot (1 - F(x+dx))^{n-k}}{dx} \\ &= \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} \cdot f(x) \cdot (1 - F(x))^{n-k}. \end{aligned}$$

Satz 4.2.6 *Verteilung einer geordneten Statistik*

$X_1,\ldots,X_n$ seien Stichprobenvariablen mit einer stetigen Verteilungsfunktion $F(x)$ und Dichte $f(x)$. Dann gilt für die k'te geordnete Statistik $X_{k:n}$:

$$f_{X_{k:n}}(x)=\frac{n!}{(k-1)!(n-k)!}\cdot f(x)\cdot F(x)^{k-1}\cdot[1-F(x)]^{n-k}. \tag{4.11}$$

Beweis: Der Beweis wird durch Induktion geführt, und zwar beginnend bei $k=n$. Für $X_{n:n}$ gilt

$$P(X_{n:n}\le x)=P(X_1\le x,\ldots,X_n\le x)=F(x)^n,$$

und somit

$$f_{X_{n:n}}(x)=n\cdot f(x)F(x)^{n-1}.$$

Die Behauptung gelte nun für ein beliebiges, festes $k<n$. Dann ist

$$\begin{aligned}P(X_{k-1:n}\le x) &= P(X_{k:n}\le x)+P(X_{k-1:n}\le x<X_{k:n})\\ &= F_{X_{k:n}}(x)+\binom{n}{k-1}F(x)^{k-1}(1-F(x))^{n-k+1},\end{aligned}$$

da $\{X_{k-1:n}\le x<X_{k:n}\}$ gerade das Ereignis ist, dass genau $k-1$ der Stichprobenvariablen kleiner oder gleich x sind und die anderen $n-(k-1)$ größer. Es folgt

$$\begin{aligned}f_{X_{k-1:n}}(x)&=\frac{\partial}{\partial x}F_{X_{k-1:n}}(x)\\ &=f_{X_{k:n}}(x)+\binom{n}{k-1}\cdot\big\{(k-1)f(x)F(x)^{k-2}(1-F(x))^{n-k}\\ &\qquad -(n+1-k)f(x)F(x)^{k-1}(1-F(x))^{n-k}\big\}\\ &=\frac{n!}{(k-2)!(n-(k-1))!}F(x)^{k-2}[1-F(x)]^{n-(k-1)}f(x).\end{aligned}$$

Beispiel 4.2.7 *geordnete Statistik bei einer Gleichverteilung*

Seien $X_1,\ldots,X_n$ Stichprobenvariablen mit einer Gleichverteilung über dem Intervall $(0,1)$. Dann ist

$$f_{X_{k:n}}(x)=\frac{n!}{(k-1)!(n-k)!}x^{k-1}[1-x]^{n-k}. \tag{4.12}$$

Das ist eine Betaverteilung mit den Parametern k und $n-k+1$. Speziell hat $X_{k:n}$ den Erwartungswert $k/(n+1)$ und die Varianz $k(n-k+1)/(n+1)^2(n+2)$.

Am weitestgehenden sind die Stichprobenverteilungen von Statistiken bei normalverteilten Stichprobenvariablen bekannt. Hier sollen einige Aussagen aus diesem Bereich angegeben werden.

Satz 4.2.8 *$\bar{X}$ und S^2 bei Normalverteilung*

$X_1,\ldots,X_n$ seien Stichprobenvariablen mit $X_i \sim \mathcal{N}(\mu;\sigma^2)$. Dann gilt:

i) Das arithmetische Mittel $\bar{X}$ hat eine $\mathcal{N}(\mu;\sigma^2/n)$ Verteilung.

ii) Die empirische Varianz S^2 hat nach Multiplikation mit n/σ^2 eine χ^2-Verteilung mit $n-1$ Freiheitsgraden, in Zeichen:

$$\frac{n\cdot S^2}{\sigma^2} \sim \chi^2(n-1).$$

iii) Das arithmetische Mittel $\bar{X}$ und die empirische Varianz S^2 sind unabhängig.

Beweis: Der Punkt i) ist der Vollständigkeit halber mit aufgeführt. Er gilt nach den Resultaten des Abschnitts 2.2. Der Beweis von ii) und iii) wird durch Induktion geführt.

Zuerst wird der Fall $n=2$ betrachtet. Hier gilt $\bar{X}=\bar{X}_2=(X_1+X_2)/2$ und, wie man leicht nachrechnet: $S^2=(X_1-X_2)^2/2$.

Da jede Linearkombination zweier gemeinsam normalverteilter Zufallsvariablen wieder normalverteilt ist, gilt ii). Denn $(X_1-X_2)/\sqrt{2}$ ist $\mathcal{N}(0;1)$ verteilt und das Quadrat besitzt nach Definition eine $\chi^2(1)$-Verteilung.
Punkt iii) ergibt sich aus der Unkorreliertheit von (X_1+X_2) und (X_1-X_2), die wegen der Normalverteilung auch die Unabhängigkeit von (X_1+X_2) und (X_1-X_2) nach sich zieht:

$$\mathrm{Cov}((X_1+X_2),(X_1-X_2))=0.$$

Die Unabhängigkeit überträgt sich auf die einzeln transformierten Variablen $\bar{X}$ und S^2.

Die Behauptungen mögen nun für ein festes, beliebiges n gelten. Es ist zu zeigen, dass sie dann auch für eine Stichprobe vom Umfang $n+1$ erfüllt sind. Etwas Rechnerei ergibt:

$$\bar{X}_{n+1}=\frac{1}{n+1}(n\cdot\bar{X}_n+X_{n+1}),$$
$$n\cdot S^2_{n+1}=(n-1)S^2_n+\frac{n}{n+1}(X_{n+1}-\bar{X}_n)^2.$$

Wegen der Reproduktivitätseigenschaft der χ^2-Verteilung folgt ii), wenn wir zeigen, dass die beiden Summanden unabhängige χ^2-verteilte Zufallsvariablen sind und der zweite Summand χ^2-verteilt ist mit einem Freiheitsgrad.

Die Unabhängigkeit von S^2_n und $(X_{n+1}-\bar{X}_n)^2$ erhalten wir so: S^2_n und X_{n+1} sind unabhängig, da in S^2_n nur $X_1,\ldots,X_n$ eingehen und die Stichprobenvariablen unabhängig sind. Zudem sind S^2_n und $\bar{X}_n$ nach Induktionsvoraussetzung unabhängig. Damit sind es auch S^2_n und $(X_{n+1}-\bar{X}_n)^2$.

Weiter besitzt $X_{n+1}-\bar{X}_n$ als Linearkombintion zweier normalverteilter Zufallsvariablen selbst eine $\mathcal{N}(0;(n+1)\sigma^2/n)$-Verteilung. Erwartungswert und Varianz ergeben sich dabei auf einfache Weise. $n(X_{n+1}-\bar{X}_n)^2/(n+1)\sigma^2$ ist also verteilt wie das Quadrat einer $\mathcal{N}(0;1)$ verteilten Zufallsvariablen, d.h. $\chi^2(1)$-verteilt.

Um iii) zu zeigen, gehen wir von den Zerlegungen $\bar{X}_{n+1} = (n \cdot \bar{X}_n + X_{n+1})/(n+1)$ und $n \cdot S^2_{n+1} = (n-1)S^2_n + n(X_{n+1} - \bar{X}_n)^2/(n+1)$ der beiden Stichprobenfunktionen aus und betrachten die entsprechenden Teile. Erstens ist festzuhalten, dass nach Induktionsvoraussetzung S^2_n und $\bar{X}_n$ unabhängig sind. Die Zerlegungen zeigen, dass dies auch für S^2_n und $\bar{X}_{n+1}$ gilt.

Zudem sind $(n \cdot \bar{X}_n + X_{n+1})$ und $(X_{n+1} - \bar{X}_n)$ gemeinsam normalverteilte Zufallsvariablen. Ihre Unabhängigkeit ergibt sich somit aus der Unkorreliertheit:

$$\mathrm{Cov}\left((n \cdot \bar{X}_n + X_{n+1}), (X_{n+1} - \bar{X}_n)\right) = \sigma^2 - n \cdot \frac{\sigma^2}{n} = 0.$$

Daher folgt die Unabhängigkeit von S^2_{n+1} und $\bar{X}_{n+1}$, da sie aus unabhängigen Teilen zusammengesetzt sind.

Satz 4.2.9 *Verteilung der Stichprobenfunktion T*

Unter den gleichen Voraussetzungen wie im Satz 4.2.8 hat die Stichprobenfunktion $T = (\bar{X} - \mu)/(S/\sqrt{n-1})$ eine (Student) t-Verteilung mit $n-1$ Freiheitsgraden.

Beweis: Aufgrund des vorhergehenden Satzes reicht es wegen

$$\frac{\bar{X} - \mu}{S/\sqrt{n-1}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n-1})}{S/\sigma}$$

zu zeigen, dass für zwei unabhängige Zufallsvariablen Z und Y, mit $Z \sim \mathcal{N}(0;1)$ und $Y \sim \chi^2(r)$ der Quotient $Z/\sqrt{Y/r}$ die Dichte

$$f(t) = \frac{\Gamma((r+1)/2)}{\sqrt{r\pi} \cdot \Gamma(r/2)} \cdot \left(1 + \frac{t^2}{r}\right)^{-(r+1)/2} \qquad t \in \mathbb{R}$$

besitzt. Dies erhalten wir auf standardmäßige Weise. Die gemeinsame Dichte von Z und Y ist

$$f(z,y) = f_Z(z) \cdot f_Y(y)$$

mit

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right),\ z \in \mathbb{R} \quad \text{und} \quad f_Y(y) = \begin{cases} \dfrac{1}{\Gamma(r/2)2^{r/2}} \cdot y^{(r/2)-1} e^{-y/2} & y > 0 \\ 0 & y \le 0 \end{cases}.$$

Die Transformationen

$$t = \frac{z}{\sqrt{y/r}} \quad \text{und} \quad u = y$$

haben die Umkehrfunktionen

$$z = \frac{t \cdot \sqrt{u}}{\sqrt{r}} \quad \text{und} \quad y = u.$$

Das ergibt die Jacobi-Determinante für die Transformation

$$\det(J) = \det \begin{pmatrix} \frac{\sqrt{u}}{\sqrt{r}} & \frac{t}{2\sqrt{u}\sqrt{r}} \\ 0 & 1 \end{pmatrix} = \frac{\sqrt{u}}{\sqrt{r}}.$$

Mit dem Transformationssatz 1.3.9 folgt

$$\begin{aligned} f_{T,U}(t,u) &= \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2u/2r} \cdot \frac{1}{\Gamma(r/2)2^{r/2}} \cdot u^{(r/2)-1} \cdot e^{-u/2} \cdot \frac{\sqrt{u}}{\sqrt{r}} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{r}\Gamma(r/2)2^{r/2}} \cdot u^{(r-1)/2} \cdot \exp\left[-\frac{u}{2}\left(1+\frac{t^2}{r}\right)\right]. \end{aligned}$$

Die Randverteilung von T erhalten wir nun durch Herausintegrieren von u:

$$f_T(t) = \int_0^\infty \frac{1}{\sqrt{2\pi}\sqrt{r}\Gamma(r/2)2^{r/2}} \cdot u^{(r-1)/2} \cdot \exp\left[\frac{u}{2}\left(1+\frac{t^2}{r}\right)\right] \mathrm{d}u.$$

Die Variablensubstitution $w = u(1+t^2/r)/2$ ergibt $\partial w/\partial u = (1+t^2/r)/2$ und weiter:

$$\begin{aligned} f_T(t) &= \int_0^\infty \frac{1}{\sqrt{2\pi}\sqrt{r}\Gamma(r/2)2^{r/2}} \cdot \left(\frac{2w}{1+(t^2/r)}\right)^{(r-1)/2} \cdot e^{-w} \frac{2}{1+(t^2/r)} \,\mathrm{d}w \\ &= \frac{1}{\sqrt{\pi}\sqrt{r}\Gamma(r/2)} \cdot \frac{1}{(1+(t^2/r))^{(r-1)/2}} \cdot \int_0^\infty w^{(r-1)/2} e^{-w} \,\mathrm{d}w \\ &= \frac{\Gamma((r+1)/2)}{\sqrt{\pi}\sqrt{r}\Gamma(r/2)} \cdot \frac{1}{(1+(t^2/r))^{(r-1)/2}} \qquad t \in \mathbb{R}. \end{aligned}$$

4.3 Der Informationsgehalt von Stichproben

4.3.1 Likelihood und Fisher-Information

Die für die induktive Statistik zentrale Verbindung zwischen einem Verteilungsmodell und einer Stichprobe ist durch die Dichte

$$f(x_1,\ldots,x_n;\vartheta) = f(x_1;\vartheta)\cdot\ldots\cdot f(x_n;\vartheta)$$

gegeben. Im diskreten Fall ist dies der Wert der Wahrscheinlichkeit dafür, dass die Stichprobe $\boldsymbol{x} = (x_1,\ldots,x_n)$ beobachtet werden wird. Im stetigen ist für kleines ϵ durch $\prod f(x_i;\vartheta)\epsilon^n$ näherungsweise die Wahrscheinlichkeit gegeben, dass eine Stichprobe beobachtet wird, die von $\boldsymbol{x}$ in den einzelnen Komponenten nur wenig abweicht. Da der Faktor ϵ^n fest ist, kann er bei der weiteren Betrachtung vernachlässigt werden. Jedoch macht die Wahrscheinlichkeitsinterpretation nach der Beobachtung keinen Sinn mehr. Vielmehr soll ja von den konkreten Werten x_i auf den unbekannten Parameter ϑ zurückgeschlossen werden. Daher führen wir den Begriff der Mutmaßlichkeit oder Likelihood ein.

Definition 4.3.1 *Likelihoodfunktion und Loglikelihoodfunktion*

Sei $\mathscr{F} = \{f(\boldsymbol{x};\vartheta)|\vartheta \in \Theta\}$ eine Verteilungsfamilie. Dann heißt die Funktion

$$\begin{aligned} L &: \Theta \to \mathbb{R} \\ \vartheta &\to L(\vartheta) = L(\vartheta;\boldsymbol{x}) = f(\boldsymbol{x};\vartheta) \end{aligned} \tag{4.13}$$

die *Likelihoodfunktion* von ϑ auf der Basis von $\boldsymbol{x} = (x_1,\ldots,x_n)$. Die *Loglikelihoodfunktion* von ϑ ist $\mathscr{L}(\vartheta) := \ln L(\vartheta)$.

Sei $L(\vartheta;\boldsymbol{x})$ die Likelihoodfunktion eines stetigen Parameters auf der Basis von $\boldsymbol{x}$, wobei $\boldsymbol{x}$ eine typische Beobachtung oder Stichprobe ist, also auch die Gestalt $\boldsymbol{x} = (x_1,\ldots,x_n)$ haben kann. Die Diskriminierung zwischen zwei Parameterwerten $\vartheta, \vartheta' \in \Theta = (a;b)$ aufgrund der Stichprobe ist umso leichter, je stärker sich die Werte der Likelihood-Funktion unterscheiden. Dabei wird bevorzugt von dem relativen Unterschied ausgegangen. Dies ist gleichwertig mit der Betrachtung der Loglikelihoodfunktion. Hat die Ableitung dieser Funktion an einer Stelle ϑ einen absolut gesehen großen Wert, so unterscheiden sich $\mathscr{L}(\vartheta;\boldsymbol{x})$ und $\mathscr{L}(\vartheta';\boldsymbol{x})$ stark, auch wenn ϑ' nahe bei ϑ liegt. Als Ausgangspunkt können wir also das Quadrat der Ableitung von $\mathscr{L}(\vartheta;\boldsymbol{x})$ nehmen:

$$\left(\frac{\partial \mathscr{L}(\vartheta;\boldsymbol{x})}{\partial \vartheta}\right)^2 = \left(\frac{\partial \ln L(\vartheta;\boldsymbol{x})}{\partial \vartheta}\right)^2 = \left(\frac{\partial \ln f(\boldsymbol{x};\vartheta)}{\partial \vartheta}\right)^2.$$

Ein globales, von der einzelnen Stichprobe unabhängiges Maß ist der Erwartungswert dieses Quadrates. Wir erhalten ihn, indem $\boldsymbol{x}$ durch die Zufallsvariable $\boldsymbol{X}$ ersetzt und dann von der resultierenden Stichprobenfunktion der Erwartungswert gebildet wird.

Definition 4.3.2 *Fisher-Information*

Sei $\boldsymbol{X} = (X_1,\ldots,X_k)$ eine k-dimensionale Zufallsvariable mit der Dichte $f(\boldsymbol{x};\vartheta)$. ϑ sei ein stetiger Parameter, $\vartheta \in \Theta = (a,b)$, und $f(\boldsymbol{x};\vartheta)$ sei stetig nach ϑ differenzierbar. Die *Fisher-Information* oder Fishers Maß der erwarteten Information ist, sofern der Erwartungswert existiert:

$$I(\vartheta) = \mathrm{E}\left(\left(\frac{\partial \ln(f(\boldsymbol{X};\vartheta))}{\partial \vartheta}\right)^2\right). \tag{4.14}$$

Beispiel 4.3.3 *Fisher-Information bei Poisson-Verteilung*

Sei X Poisson-verteilt, $X \sim \mathscr{P}(\lambda)$. Aus $f(x;\lambda) = e^{-\lambda}\cdot\lambda^x/x!$ erhalten wir $\ln f(x;\lambda) = -\lambda + x\cdot\ln\lambda - \ln(x!)$. Damit folgt:

$$\frac{\partial \ln f(x;\vartheta))}{\partial \lambda} = -1 + x\cdot\frac{1}{\lambda}.$$

Also ergibt sich

$$I(\lambda) = \mathrm{E}\left(\left(-1+\frac{x}{\lambda}\right)^2\right) = \frac{1}{\lambda^2}\cdot\mathrm{E}((X-\lambda)^2) = \frac{1}{\lambda}.$$

Beispiel 4.3.4 *Fisher-Information bei Normalverteilung mit bekanntem σ^2*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine Stichprobe aus einer $\mathcal{N}(\mu, \sigma^2)$ Verteilung mit bekanntem σ^2. Dann ist:

$$\begin{aligned} f(\boldsymbol{x};\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right], \\ \ln f(\boldsymbol{x};\mu) &= -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2, \\ \frac{\partial \ln f(\boldsymbol{x};\mu)}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu). \end{aligned}$$

Mit der Unkorreliertheit der Stichprobenvariablen folgt:

$$I_n(\mu) = \mathrm{E}\left(\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)\right)^2\right) = \frac{1}{\sigma^4}\sum_{i=1}^{n}\mathrm{E}((X_i-\mu)^2) = \frac{n}{\sigma^2}.$$

Das letzte Beispiel legt nahe, dass die Fisher-Information additiv ist, dass bei Stichproben jede Beobachtung den gleichen Teil zur gesamten Information beiträgt. Zur formalen Fassung dieses Sachverhaltes benötigen wir als Voraussetzung die Vertauschbarkeit von Integration bzw. Summenbildung und Differentation nach dem Parameter. Da die Voraussetzung noch öfter gemacht werden muss, formulieren wir sie in folgender Definition.

Definition 4.3.5 *Regularität einer Verteilungsfamilie*

Eine Verteilungsfamilie $\mathcal{F} = \{f(\boldsymbol{x};\vartheta) | \vartheta \in \Theta\}$ heißt *regulär*, wenn für die Dichtefunktionen folgende Bedingungen erfüllt sind:

(I) Der Träger von $\mathcal{F}$, d.h. die Menge $A = \{\boldsymbol{x} | f(\boldsymbol{x};\vartheta) > 0\}$, ist unabhängig von $\vartheta \in \Theta$.

(II) Im stetigen Fall gilt für jede Funktion $g(\boldsymbol{x})$ mit $\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}|g(\boldsymbol{x})|f(\boldsymbol{x};\vartheta)\,\mathrm{d}\boldsymbol{x} < \infty$:

$$\frac{\partial}{\partial\vartheta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(\boldsymbol{x})f(\boldsymbol{x};\vartheta)\,\mathrm{d}\boldsymbol{x} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(\boldsymbol{x})\left(\frac{\partial}{\partial\vartheta}f(\boldsymbol{x};\vartheta)\right)\,\mathrm{d}\boldsymbol{x}$$

und im diskreten die entsprechende Beziehung für die gleichbedeutende Summe.

Diese Regularitätsbedingungen können aus zwei Gründen verletzt sein. Zunächst kann der Träger der Verteilung vom Parameter abhängen. Dann ist es möglich, dass die Dichtefunktion bzw. ihre Ableitung nicht schnell genug gegen null geht, so dass die zugrundeliegenden Grenzwerte nicht existieren. Diese zweite Bedingung ist für die praktisch relevanten Verteilungen selten ein Problem. Insbesondere sind beide für Exponentialverteilungen erfüllt.

Satz 4.3.6 *Regularität von einparametrigen Exponentialfamilien*

$\mathscr{F} = \{f(\boldsymbol{x};\vartheta)|\vartheta \in \Theta\}$ sei eine einparametrige Exponentialfamilie. Dann ist $\mathscr{F}$ regulär.

Beweis: Lehmann (1986, S.59).

Der folgende Satz führt zu einer zusätzlichen Interpretation der Fisher-Information. Damit ist $I(\vartheta)$ die Varianz der Stichprobenfunktion $\partial \ln(f(x;\vartheta))/\partial\vartheta$, oder anders gesagt, der Steigungen von $\ln(f(X;\vartheta))$. Ein großer Wert von $I(\vartheta)$ bedeutet dann stark unterschiedliche Werte dieser Steigungen und somit größere Änderungen der Loglikelihoodfunktion auch bei nahe beieinander liegenden Werten des Parameters.

Mit dieser Interpretation erhalten wir leicht, dass die Fisher-Information additiv ist, dass bei einer Stichprobe jede Stichprobenvariable den gleichen Beitrag zur Information liefert.

Satz 4.3.7 *Fisher-Information bei einer regulären Verteilungsfamilie*

X sei eine Zufallsvariable mit einer Verteilung aus einer regulären Familie $\mathscr{F}$. Dann gilt

$$\mathrm{E}\left(\frac{\partial \ln(f(X;\vartheta))}{\partial\vartheta}\right) = 0. \tag{4.15}$$

Beweis: Wir betrachten den stetigen Fall.

$$\begin{aligned}\mathrm{E}\left(\frac{\partial \ln(f(X;\vartheta))}{\partial\vartheta}\right) &= \int_{-\infty}^{\infty} \frac{\partial \ln(f(x;\vartheta))}{\partial\vartheta} f(x;\vartheta)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{\partial f(x;\vartheta)/\partial\vartheta}{f(x;\vartheta)} f(x;\vartheta)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{\partial f(x;\vartheta)}{\partial\vartheta}\,\mathrm{d}x = \frac{\partial}{\partial\vartheta}\int_{-\infty}^{\infty} f(x;\vartheta)\,\mathrm{d}x = \frac{\partial}{\partial\vartheta}1 = 0.\end{aligned}$$

Satz 4.3.8 *Additivität der Fisher-Information*

Sei $X_1,\ldots,X_n$ eine Stichprobe aus einer Verteilung mit der Dichte $f(x;\vartheta)$, die zu einer regulären Familie $\mathscr{F}$ gehöre. Dann gilt, wenn $I_n(\vartheta)$ die Fisher-Information von ϑ auf Basis von n Stichprobenvariablen bezeichnet:

$$I_n(\vartheta) = n \cdot I_1(\vartheta). \tag{4.16}$$

Beweis: Mit der Unabhängigkeit der Stichprobenfunktionen $\partial \ln(f(X_i;\vartheta))/\partial\vartheta$ erhalten wir:

$$\begin{aligned}I_n(\vartheta) &= \mathrm{E}\left(\frac{\partial \ln f(\boldsymbol{X};\vartheta)}{\partial\vartheta}\right)^2 = \mathrm{E}\left(\sum_{i=1}^{n}\frac{\partial \ln(f(X_i;\vartheta))}{\partial\vartheta}\right)^2 \\ &= \mathrm{Var}\left(\sum_{i=1}^{n}\frac{\partial \ln(f(X_i;\vartheta))}{\partial\vartheta}\right) = \sum_{i=1}^{n}\mathrm{Var}\left(\frac{\partial \ln(f(X_i;\vartheta))}{\partial\vartheta}\right) = n \cdot I_1(\vartheta).\end{aligned}$$

Satz 4.3.9 *Eine Identität für die Fisher-Information*

Gegeben sei eine reguläre Familie $\mathscr{F}$ von Verteilungen, bei denen zusätzlich die Beziehung

$$\frac{\partial^2}{\partial\vartheta^2}\int\limits_{-\infty}^{\infty} f(x;\vartheta)\,dx = \int\limits_{-\infty}^{\infty}\frac{\partial^2}{\partial\vartheta^2} f(x;\vartheta)\,\mathrm{d}x\,,$$

bzw. die entsprechende für die Summen im diskreten Fall, erfüllt ist. Dann gilt folgende Idendität:

$$I(\vartheta) = \mathrm{E}\left(\left(\frac{\partial \ln f(X;\vartheta)}{\partial\vartheta}\right)^2\right) = -\mathrm{E}\left(\frac{\partial^2 \ln f(X;\vartheta)}{\partial\vartheta^2}\right). \tag{4.17}$$

Beweis: Für eine stetige Zufallsvariable X mit der Dichte $f(x;\vartheta)$ ist

$$\begin{aligned}\frac{\partial^2 \ln f(x;\vartheta)}{\partial\vartheta^2} &= \frac{\partial}{\partial\vartheta}\left(\frac{\partial}{\partial\vartheta}\ln f(x;\vartheta)\right) = \frac{\partial}{\partial\vartheta}\left(\frac{\partial f(x;\vartheta)/\partial\vartheta}{f(x;\vartheta)}\right)\\ &= \frac{(\partial^2 f(x;\vartheta)/\partial\vartheta^2)\cdot f(x;\vartheta) - \left(\partial f(x;\vartheta)/\partial\vartheta\right)^2}{f(x;\vartheta)^2}\\ &= \frac{\partial^2 f(x;\vartheta)/\partial\vartheta^2}{f(x;\vartheta)} - \left(\frac{\partial}{\partial\vartheta}\ln f(x;\vartheta)\right)^2.\end{aligned}$$

Es bleibt zu zeigen, dass der Erwartungswert des ersten Summanden im letzten Ausdruck verschwindet:

$$\begin{aligned}\mathrm{E}\left(\frac{\partial^2 f(x;\vartheta)/\partial\vartheta^2}{f(x;\vartheta)}\right) &= \int\limits_{-\infty}^{\infty}\frac{\partial^2 f(x;\vartheta)/\partial\vartheta^2}{f(x;\vartheta)}\cdot f(x;\vartheta)\,\mathrm{d}x = \int\limits_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\vartheta^2} f(x;\vartheta)\right)\mathrm{d}x\\ &= \frac{\partial^2}{\partial\vartheta^2}\int\limits_{-\infty}^{\infty} f(x;\vartheta)\,\mathrm{d}x = \frac{\partial^2}{\partial\vartheta^2}1 = 0.\end{aligned} \tag{4.18}$$

Der Satz zeigt, dass die Fisher-Information auch als die erwartete Änderung der Loglikelihoodfunktion definiert werden kann, wenn diese über die zweite Ableitung beschrieben wird. Dazu sind allerdings weitergehende Voraussetzungen an die Verteilungen nötig.

Es ist eine Erweiterung der hier angeführten Überlegungen auf den Fall mehrdimensionaler Parameter möglich. Für den Parametervektor $(\vartheta_1,\ldots,\vartheta_p)$ erhalten wir die $p\times p$ - *Informationsmatrix*

$$\boldsymbol{I}(\vartheta_1,\ldots,\vartheta_p) = \left(\mathrm{E}\left(\frac{\partial \ln f(x;\vartheta_1,\ldots,\vartheta_p)}{\partial\vartheta_i}\cdot\frac{\partial \ln f(x;\vartheta_1,\ldots,\vartheta_p)}{\partial\vartheta_j}\right)\right).$$

Für diese gilt unter geeigneten Regularitätsbedingungen:

$$\boldsymbol{I}(\vartheta_1,\ldots,\vartheta_p) = -\left(\mathrm{E}\left(\frac{\partial^2 \ln f(x;\vartheta_1,\ldots,\vartheta_p)}{\partial\vartheta_i\partial\vartheta_j}\right)\right).$$

Beispiel 4.3.10 *Fisher-Information bei Normalverteilung*

Sei X normalverteilt, $X \sim \mathcal{N}(\mu, \sigma^2)$. Dann ist

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

Das ergibt:

$$\begin{aligned}\frac{\partial \ln f(x;\mu,\sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2}(x-\mu),\\ \frac{\partial \ln f(x;\mu,\sigma^2)}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2.\end{aligned}$$

Die Elemente der Informationsmatrix sind daher:

$$\begin{aligned}\mathrm{E}\left(\frac{1}{\sigma^2}(X-\mu)\right)^2 &= \frac{1}{\sigma^2},\\ \mathrm{E}\left(-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X-\mu)^2\right)^2 &= \frac{1}{4\sigma^4} - 2\frac{\sigma^2}{2\sigma^2 2\sigma^4} + \frac{1}{4\sigma^8}\mathrm{E}(X-\mu)^4\\ &= -\frac{1}{4\sigma^4} + \frac{1}{4\sigma^8}\cdot\left(\frac{\sigma^2}{2}\right)^2 \cdot \frac{4!}{2!} = \frac{1}{2\sigma^4},\end{aligned}$$

$$\mathrm{E}\left(\left[\frac{1}{\sigma^2}(X-\mu)\right]\left[-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X-\mu)^2\right]\right) = 0.$$

Insgesamt ist

$$\boldsymbol{I}(\mu,\sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4)\end{pmatrix}.$$

4.3.2 Suffizienz

Die Ausgangsfrage bei der Suffizienz ist die nach der Erhaltung der in einer Stichprobe verfügbaren Information über einen unbekannten Parameter, wenn die Stichproben zusammengefasst werden.
Ist $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe aus der Verteilung $f(x;\vartheta)$, $\vartheta \in \Theta$, so enthält eine Statistik sicher *keine* Information über ϑ, wenn ihre Verteilung nicht von ϑ abhängt. Umgekehrt enthält die Statistik die gesamte Information, wenn die bedingte Verteilung der ursprünglichen Zufallsvariablen bei gegebenem Wert der Statistik nicht mehr von ϑ abhängt.

Beispiel 4.3.11 *Summe von geometrisch verteilten Zufallsvariablen*

X, Y seien unabhängig identisch geometrisch verteilt. Die gemeinsame Dichtefunktion ist dann:

$$f(x,y;p) = \mathrm{P}_p(X=x, Y=y) = p(1-p)^x p(1-p)^y \quad \text{für} \quad x,y = 0,1,2,3,\ldots$$

Es soll gezeigt werden, dass $X+Y$ in dem angesprochenen Sinn alle Information über $p \in (0,1)$ enthält.
Zum Nachweis wird die Verteilung von $T = X+Y$ benötigt. Wie man leicht sieht gilt:

$$\mathrm{P}_p(T=t\,|\,p) = (t+1)p^2(1-p)^t \quad \text{für} \quad t=0,1,2,3,\ldots$$

Damit erhalten wir:

$$\mathrm{P}(X=x, Y=y\,|\,T=t) = \begin{cases} \dfrac{\mathrm{P}_p(X=x, Y=y)}{\mathrm{P}_p(T=t)} & \text{für} \quad x+y=t \\ 0 & \text{für} \quad x+y\neq t \end{cases}$$

$$= \begin{cases} \dfrac{p(1-p)^x p(1-p)^y}{(t+1)p^2(1-p)^t} = \dfrac{1}{t+1} & \text{für} \quad x+y=t \\ 0 & x+y\neq t\,. \end{cases}$$

Folglich hängt $f_{X,Y|X+Y}(x,y\,|\,t)$ nicht mehr von p ab.

Das Ergebnis des Beispiels ist plausibel: Da p die Eintrittswahrscheinlichkeit des einen interessierenden Ereignisses ist, gibt die Gesamtzahl der Fehlversuche vor dem jeweiligen Eintreten des Ereignisses die relevante Information über p.

Auch im stetigen Fall stellt $f(x;\vartheta)$ die Verbindung zwischen der Stichprobe und der Statistik her. Das weist darauf hin, dass die skizzierte Idee in diesem Kontext ebenfalls sinnvoll ist.

Definition 4.3.12 *Suffizienz*

$\boldsymbol{X} = (X_1,\ldots,X_n)$ sei eine Zufallsstichprobe aus einer Verteilung mit der Dichte $f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)$, $\vartheta \in \Theta$. Dann heißt eine (eventuell mehrdimensionale) Statistik $T(\boldsymbol{X})$ *suffizient* für ϑ, wenn für alle $\vartheta \in \Theta$ und alle möglichen Stichproben $\boldsymbol{x}$ die bedingte Dichtefunktion von $\boldsymbol{X}$ gegeben $\{T(\boldsymbol{X}) = t\}$ nicht mehr von ϑ abhängt, d.h. für alle $\vartheta \in \Theta$ gilt:

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}{f_T(t;\vartheta)} = f(\boldsymbol{x}|\,t), \qquad T(\boldsymbol{x}) = t\,. \tag{4.19}$$

Eine einfache Möglichkeit, die Suffizienz einer Statistik nachzuweisen, liefert der folgende Satz.

Satz 4.3.13 *Faktorisierungssatz von Neyman*

$X_1,\ldots,X_n$ sei eine Zufallsstichprobe aus einer Verteilung mit der Dichtefunktion $f(x;\vartheta)$, $\vartheta \in \Theta$. Eine Statistik $T(\boldsymbol{X})$ ist genau dann suffizient für ϑ, wenn sich $f(\boldsymbol{x};\vartheta)$ als Produkt

$$f(\boldsymbol{x};\vartheta) = h(\boldsymbol{x}) \cdot g(T(\boldsymbol{x});\vartheta) \tag{4.20}$$

darstellen lässt, wobei h nicht von ϑ abhängt und $g(\cdot;\vartheta)$ nur über T von $\boldsymbol{X}$.

Beweis: Der Beweis wird für den diskreten Fall formuliert.

Sei $T(\boldsymbol{X})$ zunächst suffizient für ϑ. Dann ist die bedingte Wahrscheinlichkeitsfunktion $f_{\boldsymbol{X}|T}(\boldsymbol{x}|\,t)$ unabhängig von ϑ und wir können schreiben:

$$f_{\boldsymbol{X}|T}(\boldsymbol{x}|\,t) = h(\boldsymbol{x}).$$

Weiter ist nach Definition der bedingten Dichtefunktion

$$h(\boldsymbol{x}) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}{f_T(t;\vartheta)} \qquad \text{für} \quad T(\boldsymbol{x}) = t\,.$$

Der Nenner hängt von $\boldsymbol{x}$ nur über T ab. Multiplizieren mit $f_T(t|\,\vartheta)$ ergibt folglich die eine Richtung der Behauptung.

Um die andere Richtung zu zeigen, wird die Gültigkeit der Zerlegung unterstellt:

$$f(\boldsymbol{x};\vartheta) = h(\boldsymbol{x}) \cdot g(T(\boldsymbol{x});\vartheta)\,.$$

Für die bedingte Dichtefunktion folgt dann, sofern $T(\boldsymbol{x}) = t$:

$$f_{\boldsymbol{X}|T}(\boldsymbol{x}|t) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_T(t;\vartheta)} = \frac{h(\boldsymbol{x}) \cdot g(T(\boldsymbol{x});\vartheta)}{\mathrm{P}(T(\boldsymbol{X}) = t;\vartheta)} \qquad (T(\boldsymbol{x}) = t).$$

Für den Nenner gilt:

$$\mathrm{P}_\vartheta(T(\boldsymbol{X}) = t) = \sum_{T(\boldsymbol{x})=t} f_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{T(\boldsymbol{x})=t} h(\boldsymbol{x}) \cdot g(T(\boldsymbol{x});\vartheta) = \left(\sum_{T(\boldsymbol{x})=t} h(\boldsymbol{x}) \right) g(t;\vartheta).$$

Einsetzen der Zerlegung zeigt, dass sich $g(t;\vartheta)$ wegkürzt, so dass ϑ auf der rechten Seite nicht mehr vorkommt. Somit ist diese Richtung gezeigt.

Korollar 4.3.14 *Notwendige und hinreichende Bedingung für die Suffizienz*

Unter den Voraussetzungen des Satzes 4.3.13 ist die Statistik T genau dann suffizient für ϑ, wenn sich die Dichte von $\boldsymbol{X}$ in der Form

$$f(\boldsymbol{x};\vartheta) = h_{\boldsymbol{X}|T}(\boldsymbol{x}|t) \cdot f_T(t;\vartheta), \qquad T(\boldsymbol{x}) = t$$

darstellen lässt. Dabei ist der erste Faktor auf der rechten Seite die bedingte Dichte von $\boldsymbol{X}$ bei gegebenem Wert von $T = t$, und der zweite Faktor ist die Dichte von T.

Beweis: Analog zum vorstehenden Satz.

Beispiel 4.3.15 *Normalverteilung mit zweidimensionalem Parameter*

Seien $X_1, \ldots, X_n$ unabhängig $\mathcal{N}(\mu, \sigma^2)$-verteilt. $\boldsymbol{\vartheta} = (\mu, \sigma^2)$ sei der unbekannte Parameter. Dann gilt:

$$\begin{aligned} f(\boldsymbol{x};\mu,\sigma^2) &= \left(\frac{1}{2\pi \cdot \sigma^2} \right)^{n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] \\ &= \left(\frac{1}{2\pi \cdot \sigma^2} \right)^{n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{n}{2\sigma^2} \sum_{i=1}^{n} (\bar{x} - \mu)^2 \right]. \end{aligned}$$

Somit ist $(\bar{X}, \sum_{i=1}^{n} (X_i - \bar{X})^2)$ suffizient für $\boldsymbol{\vartheta}$.

Beispiel 4.3.16 *Suffizienz bei einer mehrparametrigen Exponentialfamilie*

Die $\mathcal{N}(\mu,\sigma^2)$-Verteilung im letzten Beispiel ist ein Spezialfall einer mehrparametrigen Exponentialfamilie mit der allgemeinen Dichte

$$f(\boldsymbol{x};\boldsymbol{\vartheta}) = \exp\left[\sum_{i=1}^{k} c_i(\boldsymbol{\vartheta})T_i(\boldsymbol{x}) + d(\boldsymbol{\vartheta}) + S(\boldsymbol{x})\right] \cdot 1_A(\boldsymbol{x}).$$

Nach dem Faktorisierungssatz ist $(T_1(\boldsymbol{X}),\ldots,T_k(\boldsymbol{X}))$ offensichtlich suffizient für $\boldsymbol{\vartheta}$.

Satz 4.3.17

$T(\boldsymbol{X})$ sei suffizient für den Parameter ϑ, von dem die Dichtefunktion von $\boldsymbol{X}$ abhängt. Dann ist der Informationsgehalt von $T(\boldsymbol{X})$ genau so groß wie der von $\boldsymbol{X}$:

$$I(\vartheta) = \mathrm{E}\left(\left(\frac{\partial \ln f(x_1,\ldots,x_n;\vartheta)}{\partial\vartheta}\right)^2\right) = \mathrm{E}\left(\left(\frac{\partial \ln f_T(t;\vartheta)}{\partial\vartheta}\right)^2\right).$$

Beweis: Wir führen den Beweis für den diskreten Fall. Es ist

$$f_T(t;\vartheta) = \sum_{T(\boldsymbol{x})=t} f(\boldsymbol{x};\vartheta) = \sum_{T(\boldsymbol{x})=t} h(\boldsymbol{x})\cdot g(t;\vartheta).$$

Damit folgt:

$$\begin{aligned}\frac{\partial \ln f_T(t;\vartheta)}{\partial\vartheta} &= \frac{\partial}{\partial\vartheta}\ln\left[\sum_{T(\boldsymbol{x})=t} f(\boldsymbol{x};\vartheta)\right] = \frac{\partial}{\partial\vartheta}\left\{\ln(g(t;\vartheta)) + \ln\left[\sum_{T(\boldsymbol{x})=t} h(\boldsymbol{x})\right]\right\}\\ &= \frac{\partial}{\partial\vartheta}\ln(g(t;\vartheta)).\end{aligned}$$

Offensichtlich gilt auch

$$\frac{\partial \ln f(\boldsymbol{x};\vartheta)}{\partial\vartheta} = \frac{\partial}{\partial\vartheta}\ln(g(t;\vartheta)),$$

so dass die Behauptung unmittelbar folgt.

Auch bei der Bildung des bedingten Erwartungswertes bzgl. einer suffizienten Statistik geht die Abhängigkeit vom Parameter verloren.

Satz 4.3.18 *bedingte Erwartung bei gegebener suffizienter Statistik*

$\boldsymbol{X}$ sei eine n-dimensionale Zufallsvariable mit der gemeinsamen Dichte $f(\boldsymbol{x};\vartheta)$, $\vartheta\in\Theta$. $T(\boldsymbol{X})$ sei eine für ϑ suffiziente Statistik und $S(\boldsymbol{X})$ eine eindimensionale Statistik. Dann gilt für alle $\vartheta\in\Theta$:

$$\mathrm{E}_\vartheta(S(\boldsymbol{X})|T(\boldsymbol{X})) = \mathrm{E}(S(\boldsymbol{X}_n)|T(\boldsymbol{X})), \tag{4.21}$$

d.h. die bedingte Erwartung von $S(\boldsymbol{X})$ gegeben $T(\boldsymbol{X})$ ist unabhängig von ϑ.

Beweis: Der Beweis wird für den stetigen Fall formuliert.

Wegen der Suffizienz von T ist die bedingte Dichte $f(\boldsymbol{x}|t;\vartheta)$ unabhängig von ϑ, also gleich $f(\boldsymbol{x}|t)$. Es folgt

$$\mathrm{E}_\vartheta(S(\boldsymbol{X})|T(\boldsymbol{X})=t)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}S(\boldsymbol{x})\cdot f(\boldsymbol{x}|t)\,\mathrm{d}x_1\cdots\mathrm{d}x_n.$$

Da der Integrand unabhängig von ϑ ist, gilt auch

$$\mathrm{E}_\vartheta(S(\boldsymbol{X})|T(\boldsymbol{X})=t)=\mathrm{E}(S(\boldsymbol{X})|T(\boldsymbol{X})=t).$$

Daraus folgt schließlich die Unabhängigkeit der bedingten Erwartung vom Parameter ϑ.

Beispiel 4.3.19 *zwei exponentialverteilte Zufallsvariablen - Fortsetzung von Seite 65*

Seien X, Y zwei unabhängige Zufallsvariablen mit der gleichen Exponentialverteilung. Es soll $\mathrm{E}(Y|X+Y)$ bestimmt werden. Nach Beispiel 1.3.10 gilt:

$$f_{X+Y}(u;\lambda)=\lambda^2 u\cdot e^{-\lambda u},\qquad u>0.$$

Mit Beispiel 2.3.4 ist

$$f_{X+Y,Y}(u,y;\lambda)=\lambda^2 e^{-\lambda u},\qquad u>y>0.$$

Damit folgt:

$$f_{Y|X+Y}(y|u;\lambda)=\frac{\lambda^2 e^{-\lambda u}}{\lambda^2 u\cdot e^{-\lambda u}}=\frac{1}{u},\qquad u>y>0,$$

$$\mathrm{E}_\lambda(Y|X+Y=u)=\int_0^u y\cdot f_{Y|X+Y}(y|u;\lambda)\,\mathrm{d}y=\int_0^u y\cdot\frac{1}{u}\,\mathrm{d}y=\frac{1}{u}\cdot\frac{1}{2}u^2=\frac{1}{2}u.$$

Somit ist der bedingte Erwartungswert $\mathrm{E}_\lambda(Y|X+Y=u)$ unabhängig von λ. Folglich ist auch die bedingte Erwartung $\mathrm{E}_\lambda(Y|X+Y)=u/2$ unabhängig von λ:

$$\mathrm{E}_\lambda(Y|X+Y)=\mathrm{E}(Y|X+Y)\qquad\text{für alle }\lambda>0.$$

Das ist natürlich nicht der Nachweis der Suffizienz von $X+Y$, aber eine Konsequenz daraus.

Wir betrachten nun die Möglichkeiten, suffiziente Statistiken weiter zu verdichten. Dazu starten wir mit einem Beispiel.

Beispiel 4.3.20 *Suffizienz der Ordnungsstatistik*

$X_1,\ldots,X_n$ seien unabhängige, identisch verteilte Zufallsvariablen mit der Dichte $f(x)$. Werden die zugelassenen Verteilungen nicht weiter eingeschränkt, so kann f selbst als

Parameter genommen werden. Der Parameterraum ist dann $\Theta = \{f(x) | f(x) \text{ ist eine Dichte}\}$. Die Ordnungsstatistik $\boldsymbol{X}_{:n}$ ist suffizient für $f \in \Theta$:

$$f(\boldsymbol{x}) = f(x_1) \cdot \ldots \cdot f(x_n) = f(x_{1:n}) \cdot \ldots \cdot f(x_{n:n}) = f(x_{1:n}, \ldots, x_{n:n}).$$

Beschränken wir uns nachträglich auf eine bestimmte Familie $\{P_\vartheta | \vartheta \in \Theta'\}$ von Verteilungen mit dem charakterisierenden Parameter ϑ — etwa Normalverteilungen mit $\vartheta = (\mu, \sigma^2)$, oder Exponentialverteilungen — so ist $X_{:n}$ auch suffizient für $\vartheta \in \Theta'$, da ja die Faktorisierung weiter gültig bleibt.

Das letzte Beispiel zeigt auch, dass eine Statistik suffizient für einen Parameter sein kann, die bei der Reduktion quasi ‚auf der halben Strecke' stehen bleibt. Es ist etwa $\bar{X} = \sum X_{i:n}$, so dass die bei Normalverteilung mit bekanntem σ^2 für μ suffiziente Statistik $\bar{X}$ als Funktion von $X_{1:n}, \ldots, X_{n:n}$ aufgefasst werden kann. Allgemein gilt:

Satz 4.3.21 *Suffizienz der Verfeinerungen suffizienter Statistiken*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine n-dimensionale Zufallsvariable mit der gemeinsamen Dichte $f(\boldsymbol{x}; \vartheta)$, $\vartheta \in \Theta$. Ist die Statistik $T(\boldsymbol{X})$ suffizient für ϑ, so ist es auch jede Statistik $S(\boldsymbol{X})$, für die

$$T(\boldsymbol{X}) = R(S(\boldsymbol{X}))$$

gilt, wobei R eine geeignete Funktion vom Bildbereich von S in den von T ist.

Beweis: Die Behauptung folgt sofort aus

$$f(\boldsymbol{x}; \vartheta) = h(\boldsymbol{x}) \cdot g(T(\boldsymbol{x}); \vartheta) = h(\boldsymbol{x}) \cdot g(R(S(\boldsymbol{x})); \vartheta) = h(\boldsymbol{x}) \cdot \tilde{g}(S(\boldsymbol{x}); \vartheta)$$

mit $\tilde{g}(S(\boldsymbol{x}); \vartheta) = g(R(S(\boldsymbol{x})); \vartheta)$.

Das Beispiel 4.3.19 und der Satz 4.3.21 werfen unmittelbar die Frage nach suffizienten Statistiken auf, die nicht weiter 'verdichtet' werden können, ohne dass die Eigenschaft der Suffizienz verloren geht.

Definition 4.3.22 *Minimalsuffizienz*

Die Zufallsvariable X habe eine Verteilung, die zu der Verteilungsfamilie $\mathcal{F} = \{f(x; \vartheta) | \vartheta \in \Theta\}$ gehöre. $\boldsymbol{X}$ sei eine Zufallsstichprobe aus der Verteilung von X. Eine Statistik $T(\boldsymbol{X})$ heißt *minimalsuffizient* für $\vartheta \in \Theta$, falls gilt:

(i) $T(\boldsymbol{X})$ ist suffizient für $\vartheta \in \Theta$;

(ii) $T(\boldsymbol{X})$ ist eine Funktion jeder anderen suffizienten Statistik.

Die folgende Aussage ist hilfreich, um den Punkt (ii) der Definition zu verdeutlichen.

Lemma 4.3.23 *Charakterisierung der Funktionseigenschaft*

Seien $U(\boldsymbol{t})$ und $V(\boldsymbol{t})$ zwei auf derselben Menge $D \subseteq \mathbb{R}^k$ definierte Funktionen. $U(\boldsymbol{t})$ ist genau dann eine Funktion von $V(\boldsymbol{t})$, wenn für alle $\boldsymbol{t}, \boldsymbol{t}^* \in D$ gilt:

$$V(\boldsymbol{t}) = V(\boldsymbol{t}^*) \quad \Rightarrow \quad U(\boldsymbol{t}) = U(\boldsymbol{t}^*).$$

Der nun einzuführende Begriff der Likelihood-Äquivalenz ist zentral für die Entscheidung, ob eine suffiziente Statistik auch minimalsuffizient ist. Er geht davon aus, dass der Informationsgehalt verschiedener Stichproben als gleich anzusehen ist, wenn die Likelihoodfunktion für diese Stichproben im Wesentlichen konstant ist. Dabei wird zur Vereinfachung unterstellt, dass der Stichprobenraum nur Punkte enthält, für die jeweils mindestens eine Dichtefunktion von Null verschieden ist; er erfüllt also

$$\{\boldsymbol{x} | f(\boldsymbol{x};\vartheta) > 0 \quad \text{für ein } \vartheta \in \Theta\}.$$

Definition 4.3.24 *Likelihoodäquivalenz*

Zwei Punkte $\boldsymbol{y}$, $\boldsymbol{z}$ heißen *likelihoodäquivalent* für $\mathscr{F} = \{f(\boldsymbol{x};\vartheta)|\vartheta \in \Theta\}$, wenn

$$\{\vartheta | f(\boldsymbol{y};\vartheta) > 0\} = \{\vartheta | f(\boldsymbol{z};\vartheta) > 0\}$$

und wenn für alle $\vartheta \in \Theta$ gilt:

$$f(\boldsymbol{z};\vartheta) = c(\boldsymbol{z},\boldsymbol{y}) \cdot f(\boldsymbol{y};\vartheta).$$

Beispiel 4.3.25 *Stichprobe aus einer Rechteckverteilung*

Seien $X_1,\ldots,X_n$ rechteckverteilt mit dem Parameter $\vartheta > 0$, $X_i \sim \mathscr{R}(0,\vartheta)$. Hier gilt:

$$f(\boldsymbol{x};\vartheta) = f(\boldsymbol{y};\vartheta) = \begin{cases} 1/\vartheta^n & \text{falls } x_i < \vartheta,\ y_j < \vartheta \ \ (i,j = 1,\ldots,n) \\ 0 & \text{falls } x_i > \vartheta \text{ und } y_j > \vartheta \\ & \text{für je ein } i,j \in \{1,\ldots,n\} \end{cases}$$

$$f(\boldsymbol{x};\vartheta) \neq f(\boldsymbol{y};\vartheta) \qquad \text{sonst.}$$

Damit ist

$$\{\vartheta | f(\boldsymbol{x};\vartheta) > 0\} = \{\vartheta | \max\{x_1,\ldots,x_n\} < \vartheta\}.$$

Weiter sind genau für die Stichproben mit $\max\{\boldsymbol{x}\} = \max\{\boldsymbol{y}\}$ die zulässigen Dichtefunktionen identisch. Somit sind alle Stichproben likelihoodäquivalent, bei denen der maximale Wert der Stichprobe übereinstimmt.

Beispiel 4.3.26 *Stichprobe aus einer Bernoulli-Verteilung*

$X_1,\ldots,X_n$ seien $\mathscr{B}(1,p)$-verteilt. Dann ist die Likelihoodfunktion

$$L(p;\boldsymbol{x}) = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}.$$

Somit sind alle Stichproben likelihoodäquivalent, welche die gleiche Anzahl von Einsen haben.

Satz 4.3.27 *Suffizienz und Likelihoodäquivalenz*

$\mathscr{F} = \{f(\boldsymbol{x};\vartheta)|\vartheta \in \Theta\}$ sei eine Familie von Verteilungen, $\boldsymbol{X}$ die zugrundeliegende Zufallsvariable, und $T(\boldsymbol{X})$ sei eine suffiziente Statistik. Je zwei Punkte, für die $T(\boldsymbol{X})$ den gleichen Wert annimmt, sind likelihoodäquivalent:

$$T(\boldsymbol{x}) = T(\boldsymbol{y}) \quad \Rightarrow \quad \boldsymbol{x} \text{ und } \boldsymbol{y} \text{ sind likelihoodäquivalent.}$$

$T(\boldsymbol{X})$ ist minimalsuffizient, wenn umgekehrt auch gilt:

$$\boldsymbol{x} \text{ und } \boldsymbol{y} \text{ sind likelihoodäquivalent} \quad \Rightarrow \quad T(\boldsymbol{x}) = T(\boldsymbol{y}).$$

Beweis: Sei $T(\boldsymbol{X})$ suffizient für ϑ. Dann gilt mit dem Neymanschen Faktorisierungssatz:

$$f(\boldsymbol{x};\vartheta) = h(\boldsymbol{x}) \cdot g_\vartheta(T(\boldsymbol{x})).$$

Ist also $T(\boldsymbol{x}) = T(\boldsymbol{y})$ und ist $h(\boldsymbol{y}) \neq 0$, so folgt:

$$f(\boldsymbol{x};\vartheta) = h(\boldsymbol{x}) \cdot g(T(\boldsymbol{x});\vartheta) = \frac{h(\boldsymbol{x})}{h(\boldsymbol{y})} \cdot h(\boldsymbol{y}) \cdot g(T(\boldsymbol{y});\vartheta) = c((\boldsymbol{x}),(\boldsymbol{y})) \cdot f(\boldsymbol{y};\vartheta).$$

Wäre $h(\boldsymbol{x}) = 0$, ergäbe sich mit der Faktorisierung

$$f(\boldsymbol{x};\vartheta) = h(\boldsymbol{x}) \cdot g(T(\boldsymbol{x});\vartheta) \equiv 0 \qquad \text{für } \vartheta \in \Theta.$$

Das wurde aber eingangs ausgeschlossen. $T(\boldsymbol{X})$ möge nun genau für die likelihoodäquivalenten Stichproben konstant sein. Nach Lemma 4.3.22 ist $T(\boldsymbol{X})$ dann eine Funktion aller suffizienten Statistiken, mithin also minimalsuffizient.

Der Satz legt einen Algorithmus nahe, wie eine minimalsuffiziente Statistik bestimmt werden kann. Er besteht aus zwei Schritten:

Schritt 1: Für jede Stichprobe $\boldsymbol{x}$ ist die Menge

$$\Theta_{\boldsymbol{x}} = \{\vartheta | \vartheta \in \Theta,\ f(\boldsymbol{x};\vartheta) > 0\}$$

in Abhängigkeit von $\boldsymbol{x}$ zu bestimmen.

Schritt 2: Bestimme für $\boldsymbol{x}$ diejenigen Stichproben $\boldsymbol{y}$, für die gilt:

$$\Theta_{\boldsymbol{y}} = \Theta_{\boldsymbol{x}},$$
$$\frac{f(\boldsymbol{y};\vartheta)}{f(\boldsymbol{x};\vartheta)} = \text{ const.} \qquad \text{auf } \Theta_{\boldsymbol{x}}.$$

Die minimalsuffiziente Statistik $T(\boldsymbol{X})$ ergibt sich daraus, indem $T(\boldsymbol{x})$ so definiert wird, dass $T(\boldsymbol{x})$ auf genau den likelihoodäquivalenten Stichproben, die ja im Schritt 2 identifiziert werden, als jeweils konstant gesetzt wird.

Beispiel 4.3.28 *Stichprobe aus einer Bernoulli-Verteilung - Fortsetzung von Seite 138*

Offensichtlich sind zwei Stichproben $\boldsymbol{x}$, $\boldsymbol{y}$ aus einer Bernoulli-Verteilung genau dann likelihoodäquivalent, wenn für alle p, $0 < p < 1$ gilt:

$$f_p(\boldsymbol{x}) = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i} = p^{\Sigma y_i}(1-p)^{n-\Sigma y_i} = f_p(\boldsymbol{y}).$$

Dies ist aber nur im Fall $\sum x_i = \sum y_i$ erfüllt. Somit ist $T(\boldsymbol{X}) = \sum X_i$ minimalsuffizient.

Beispiel 4.3.29 *Stichprobe aus einer verschobenen Exponentialverteilung*

Sei $\boldsymbol{X}$ eine Zufallsstichprobe aus einer verschobenen Exponentialverteilung,

$$f(\boldsymbol{x};\alpha)=\begin{cases}\exp\left[-\left(\sum x_i - n\alpha\right)\right], & x_1,\ldots,x_n>\alpha\\ 0 & \text{sonst}\end{cases}.$$

Hier gilt offensichtlich

$$f(\boldsymbol{x};\alpha)>0 \qquad \text{falls } x_{(1)}>\alpha,$$

und für eine zweite Stichprobe $(\boldsymbol{y})$ mit $y_{(1)}>\alpha$:

$$\frac{f(\boldsymbol{x};\alpha)}{f(\boldsymbol{y};\alpha)}=\frac{\exp\left[-\left(\sum x_i - n\alpha\right)\right]}{\exp\left[-\left(\sum y_i - n\alpha\right)\right]}=\exp\left[-\left(\sum x_i-\sum y_i\right)\right] \qquad \text{falls } x_{(1)}=y_{(1)}.$$

Folglich ist $T(\boldsymbol{X})=\min\{X_1,\ldots,X_n\}$ minimalsuffizient für α.

Eine Entscheidungshilfe für die Minimalsuffizienz einer Statistik bietet die Vollständigkeit. Die Vollständigkeit ist eine Eigenschaft der Verteilungsfamilie. Ist der Parameterraum Θ klein, so gibt es viele von null verschiedene Funktionen mit $\mathrm{E}_\vartheta(g(X))=0$. Je größer der Parameterraum aber wird, umso kleiner wird diese Menge der Funktionen, bis nur noch eine übrig bleibt, die dies immer schafft. Das ist dann die konstante Funktion. Für die Suffizienz heißt das, dass es keine echte Funktion der suffizienten Statistik gibt, die noch alle Information enthält.

Definition 4.3.30 *Vollständigkeit*

Die Verteilungsfamilie $\mathscr{F}=\{f_X(x;\vartheta)|\vartheta\in\Theta\}$ der Zufallsvariablen X heißt *vollständig*, wenn für jede Funktion $h(s)$ mit $\mathrm{E}_\vartheta(h(X))=0$ für alle $\vartheta\in\Theta$ folgt:

$$\mathrm{P}_\vartheta(h(X)=0)=1 \qquad \text{für alle } \vartheta\in\Theta.$$

Wir sagen auch, dass die Zufallsvariable X vollständig ist.

Beispiel 4.3.31 *diskrete Gleichverteilung*

Die Zufallsvariable X sei auf $\mathbb{N}$ definiert und besitze eine diskrete Gleichverteilung mit der Zähldichte

$$f_T(t;N)=\begin{cases}\dfrac{1}{N} & \text{für } t=1,2,\ldots,N\\ 0 & \text{sonst}\end{cases}.$$

Der Parameterraum ist hier $\mathbb{N}$.
Für den Nachweis der Vollständigkeit von X muss für jedes $N\in\mathbb{N}$ gezeigt werden:

$$\mathrm{E}_N(h(X))=0 \quad\Rightarrow\quad \mathrm{P}_N(h(X)=0)=1.$$

Wir zeigen dies mit vollständiger Induktion.

Im Fall $N=1$ folgt für h aus $\mathrm{E}(h(X))=1\cdot h(1)+0\cdot\sum_{x=2}^{\infty}h(x)=0$ sofort $h(1)=0$.

Die Behauptung gelte nun für ein festes, beliebiges N.

Sei $h(x)$ gegeben mit $\mathrm{E}_{N+1}(h(X))=0$, d.h.

$$\begin{aligned}\mathrm{E}_{N+1}(h(X)) &= h(1)\frac{1}{N+1}+h(2)\frac{1}{N+1}+\ldots+h(N)\frac{1}{N+1}+h(N+1)\frac{1}{N+1}+0\cdot\sum_{x=N+2}^{\infty}h(x)\\ &=0.\end{aligned}$$

Aufgrund der Induktionsvoraussetzung gilt

$$h(1)\frac{1}{N}+h(2)\frac{1}{N}+\ldots+h(N)\frac{1}{N}=0 \quad\Rightarrow\quad h(1)=\ldots=h(N)=0.$$

Damit erhalten wir auch

$$h(1)\frac{1}{N+1}+h(2)\frac{1}{N+1}+\ldots+h(N)\frac{1}{N+1}=0,$$

und somit

$$h(N+1)\frac{1}{N+1}=0 \quad \text{bzw.} \quad h(N+1)=0.$$

Also folgt aus $\mathrm{E}_{N+1}(h(X))=0$, dass $h(1)=h(2)=\ldots=h(N)=h(N+1)=0$ und weiter $\mathrm{P}(h(X)=0)=1$ gilt.

Entfernt man auch nur ein Element des Parameterraumes, so ist die Verteilungsfamilie nicht mehr vollständig. Das ist anhand des folgenden Falles zu sehen.

Der Parameterraum sei $N=\{2,3,\ldots\}=\mathbb{N}\backslash\{1\}$. Die Zähldichte von X ist weiterhin

$$f(x;N)=\begin{cases}\frac{1}{N} & \text{für } x=1,2,\ldots,N\\ 0 & \text{sonst}\end{cases}.$$

Sei

$$h(x)=\begin{cases}0.5 & \text{für } x=1\\ -0.5 & \text{für } x=2\\ 0 & \text{sonst}\end{cases}.$$

Für $N=2$ gilt:

$$\mathrm{E}_2(h(X))=0.5\cdot 0.5+(-0.5)\cdot 0.5=0.$$

Offensichtlich gilt auch für $N>2$: $\mathrm{E}_N(h(X))=0$. Wegen $\mathrm{P}(h(X)=0)\neq 1$ ist X nicht vollständig.

Man kann zeigen, dass *Exponentialfamilien* vollständig sind. Wie in Satz 3.1.23 gezeigt wurde, gehört $T(X)$ zu einer Exponentialfamilie. Wenn die einzelnen X_i unabhängig sind und dieselbe Exponentialfamilie besitzen, gehört auch $S=\sum_{i=1}^{n}X_i$ zu einer Exponentialfamilie. Somit ist S bei der Exponentialfamilie eine vollständige Statistik. Wie wir gesehen haben, ist S auch suffizient.

Eine vollständige suffiziente Statistik ist auch minimalsuffizient.

Lemma 4.3.32 *Hinreichende Bedingung für die Minimalsuffizienz*

Wenn eine vollständige und suffiziente Statistik T für den Parameter ϑ existiert, ist sie minimalsuffizient.

Beweis: Wir führen den Beweis für den diskreten Fall. Sei T minimalsuffizient und sei S eine andere suffiziente Statistik, die nicht minimalsuffizient ist. Wir zeigen, dass S dann nicht vollständig sein kann.

Da T eine Funktion von S ist, gibt es zwei Mengen A und B, so dass

$$\begin{aligned} &\mathrm{P}_\vartheta(A)\neq 0,\ \mathrm{P}_\vartheta(B)\neq 0 && \forall\,\vartheta\in\Theta\\ &S(a)=s_1\neq s_2=S(b) && \forall\,a\in A,\,b\in B\\ &T(a)=T(b)=t && \forall\,a\in A,\,b\in B. \end{aligned}$$

Wir betrachten nun die Funktion

$$g(s)=\begin{cases} 1 & s=s_1\\ \dfrac{-\mathrm{P}(S=s_1|T=t)}{\mathrm{P}(S=s_2|T=t)} & s=s_2\\ 0 & \text{sonst}\end{cases}.$$

Die Funktion $g(s)$ ist nicht identisch gleich null und unabhängig von ϑ, da T suffizient ist. Andererseits gilt für alle $\vartheta\in\Theta$:

$$\begin{aligned}\mathrm{E}_\vartheta[g(S)] &= \mathrm{P}_\vartheta(S=s_1)+\frac{-\mathrm{P}(S=s_1|T=t)}{\mathrm{P}(S=s_2|T=t)}\mathrm{P}_\vartheta(S=s_2)\\ &= \mathrm{P}_\vartheta(S=s_1)-\frac{\mathrm{P}(S=s_1,T=t)}{\mathrm{P}(S=s_2,T=t)}\mathrm{P}_\vartheta(S=s_2)=0.\end{aligned}$$

Beispiel 4.3.33 *minimalsuffiziente Statistik bei Gleichverteilung*

$\boldsymbol{X}$ sei eine Stichprobe aus einer $\mathscr{R}(\vartheta-1/2,\vartheta+1/2)$-Verteilung. Dann ist $(X_{1:n},X_{n:n})$ minimalsuffizient für ϑ. Da

$$\mathrm{E}_\vartheta(X_{1:n})=\vartheta-\frac{n-1}{2(n+1)},\qquad \mathrm{E}_\vartheta(X_{n:n})=\vartheta+\frac{n-1}{2(n+1)}$$

ist

$$g(s,t)=t-s-\frac{n-1}{n+1}$$

eine nicht überall verschwindende Funktion mit $\mathrm{E}_\vartheta(g(X_{1:n},X_{n:n}))=0$ für alle ϑ. $(X_{1:n},X_{n:n})$ ist nicht vollständig und damit existiert überhaupt keine vollständige suffiziente Statistik.

4.4 Aufgaben

Aufgabe 1

Bei Umfragen ist es oft schwierig, richtige Antworten auf sensible Fragen zu erhalten. – Man denke etwa an eine Frage der Art: ‚Haben Sie jemals Rauschgift genommen?‘ – Warner hat 1965 daher die Methode der randomisierten Antwort eingeführt, um solche Situationen zu bewältigen. Der Frager weiß bei dieser Methode nicht, auf welche der beiden einfach alternativen Fragen der Befragte antwortet. Das wird erreicht, indem der Befragte (ohne dass der Frager es sehen kann) zufällig eine Kugel aus einer Urne zieht, welche Kugeln mit zwei unterschiedlichen Farben enthält. Die eine Farbe repräsentiert ‚Ich habe die Eigenschaft A‘, die andere repräsentiert ‚Ich habe die Eigenschaft A nicht‘. Auf dieses Statement antwortet der Befragte dann mit ja bzw. nein.

Sei

- R der Anteil der Befragten, die mit ‚ja‘ antworten,
- p die Chance, dass eine Kugel mit der Farbe, welche ‚Ich habe die Eigenschaft A‘ repräsentiert, gezogen wird,
- q der Anteil der Personen mit der Eigenschaft A in der Grundgesamtheit,
- r die Wahrscheinlichkeit, dass der Befragte mit ‚ja‘ antwortet.

Zeigen Sie:

i) $r = (2p-1)q + (1-p)$
Hinweis: P(‚$j\,a$‘)|Frage1)P(Frage2) + P(‚$j\,a$‘|Frage2)P(Frage2).

ii) $\mathrm{E}(R) = r$ und $\mathrm{Var}(R) = r(n-1)/n$ (unter Vernachlässigung des Faktors für das Ziehen ohne Zurücklegen).

Aufgabe 2

Bestimmen Sie suffiziente Statistiken für Stichproben vom Umfang n, wenn folgende Verteilungen zugrunde liegen:

$$\begin{aligned}
\text{Pareto mit bekanntem } k\text{:}\quad & f(x) = \alpha k^{\alpha} x^{-(\alpha+1)} \quad \text{für} \quad x > k \\
\text{Laplace:}\quad & f(x,\lambda) = 0.5\,\lambda \exp[-\lambda|x|] \\
\text{Gamma:}\quad & f(x;\lambda,k) = \lambda^k x^{k-1} \exp[-\lambda x]/\Gamma(k) \quad \text{für} \quad x > 0 \\
\text{verschobene Exponentialverteilung:}\quad & f(x,\theta,\lambda) = \lambda \exp[-\lambda(x-\theta)] \quad \text{für} \quad x > \theta \\
\text{zentrierte Gleichverteilung:}\quad & f(x) = 1/2\theta \quad \text{für} \quad -\theta < x < \theta\,.
\end{aligned}$$

Aufgabe 3

$X_1, \ldots, X_n$ seien unabhängige identisch verteilte Zufallsvariablen mit einer stetigen Verteilung. Bestimmen Sie die Verteilung von $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Zeigen Sie mittels Induktion, dass die Dichte der k'ten geordneten Statistik $X_{(k)}$ gegeben ist durch

$$g_k(x) = \frac{n!}{(k-1)!(n-k)!} f(x)[F(x)]^{k-1}[1-F(x)]^{n-k}.$$

Dabei ist f die Dichte der X_i und F die zugehörige Verteilungsfunktion.

Aufgabe 4

Eine Münze wird solange geworfen, bis zum ersten Mal ‚Zahl' erscheint. Ist dies beim $k$ten Wurf der Fall, dann wird der Betrag 2^{k-1} ausgezahlt. Das Spiel endet nach höchstens n Würfen. Ist n-mal ‚Kopf' eingetreten, wird nichts ausbezahlt.

Übersetzen Sie dieses Spiel in ein Stichprobenverfahren; gehen Sie dazu von der Grundgesamtheit $\{\omega_1, \ldots, \omega_n\}$ aus, wobei $\omega_i = 0$ falls Kopf und $\omega_i = 1$ falls Zahl kommt.

Geben Sie die Verteilung der Zufallsvariablen $X =$'Gewinn' an. Wie groß ist der erwartete Gewinn? Wie groß ist die Varianz von X?

5 Grenzwertsätze und asymptotische Verteilungen

5.1 Formen der Konvergenz

In vielen Anwendungen der Statistik geht es um das Herstellen bzw. Ausnützen eines Zusammenhanges zwischen empirischen Häufigkeits- und theoretischen Wahrscheinlichkeitsverteilungen. Bei endlichen Stichprobenumfängen erschwert die natürliche Zufallsschwankung das Erkennen des Zusammenhanges. Bei wachsendem Stichprobenumfang verbessert sich die Situation in vielen Fällen. Es gibt zur Präzisierung dieser Aussage verschiedene Konzepte, die hier zunächst vorgestellt werden sollen.

Definition 5.1.1 *Konvergenzbegriffe*

Eine Folge $X_1, X_2, \ldots, X_n, \ldots$ von Zufallsvariablen heißt konvergent gegen die Zufallsvariable X

in Wahrscheinlichkeit oder *stochastisch,* in Zeichen $X_n \xrightarrow{\mathrm{P}} X$, falls

$$\lim_{n\to\infty} \mathrm{P}(|X_n - X| > \epsilon) = 0 \quad \text{für alle } \epsilon > 0; \tag{5.1}$$

mit Wahrscheinlichkeit eins oder *fast sicher,* in Zeichen $X_n \xrightarrow{f.s.} X$, falls

$$\mathrm{P}\left(\lim_{n\to\infty} |X_n - X| = 0\right) = 1; \tag{5.2}$$

im r'ten Mittel, in Zeichen $X_n \xrightarrow{r} X$, falls

$$\lim_{n\to\infty} \mathrm{E}[|X_n - X|^r] = 0; \tag{5.3}$$

in Verteilung oder *schwach,* in Zeichen $X_n \xrightarrow{V} X$, falls für die zugehörigen Verteilungsfunktionen $F_n(x)$ bzw. $F(x)$ gilt:

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \text{für alle } x. \tag{5.4}$$

Um die Begriffe zu illustrieren, denken wir uns die Zufallsvariable X vorübergehend als entartet, d. h. als Konstante c. Die Konvergenz in Wahrscheinlichkeit sagt, dass mit wachsendem Stichprobenumfang immer seltener große Abweichungen der Realisationen der X_n von

c zu erwarten sind. Dazu sagt man häufig auch, die Folge der Beobachtungen stabilisiere sich bei der Konstanten c. Die fast sichere Konvergenz beinhaltet dann sogar, dass alle Folgen bis auf eine 'praktisch irrelevante Menge' tatsächlich gegen c konvergieren.

Die schwache Konvergenz erlaubt schließlich, Verteilungsaussagen über die Zufallsvariablen X_n mit Hilfe der Verteilung von X zu machen. Bei genügend großem n ist der Fehler für Anwendungszwecke vernachlässigbar.

Lemma 5.1.2 *Schwaches Gesetz der großen Zahlen*

X_n, $n = 1,2,\ldots$ seien unabhängige identisch verteilte Zufallsvariablen mit $\mathrm{E}(X) = \mu$ und $\mathrm{Var}(X) = \sigma^2$. Dann konvergiert $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ stochastisch gegen μ.

Beweis: Mit der Tschebyschev-Ungleichung erhalten wir für festes $\epsilon > 0$:

$$\mathrm{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \text{für } n \to \infty.$$

Wird das Lemma 5.1.2 auf Zufallsvariablen X_n mit $\mathrm{P}(X_n = 1) = p = 1 - \mathrm{P}(X_n = 0)$ angewandt, so erhält man das *Theorem von Bernoulli.*

Satz 5.1.3 *Beziehungen zwischen den Formen von Konvergenz*

Bzgl. der in der Definition 5.1.1 formulierten Konvergenzbegriffe gilt:

$$(i) \quad X_n \xrightarrow{f.s.} X \Rightarrow X_n \xrightarrow{\mathrm{P}} X$$

$$(ii) \quad X_n \xrightarrow{\mathrm{P}} X \Rightarrow X_n \xrightarrow{V} X$$

$$(iii) \quad X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{\mathrm{P}} X$$

$$(iv) \quad X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{s} X \quad \text{sofern } s < r.$$

Die im Satz nicht aufgeführten Folgebeziehungen gelten nicht allgemein. Für weitere Beziehungen und Beispiele sei auf Rohatgi (1975) verwiesen.

Das in Lemma 5.1.2 präsentierte schwache Gesetz der großen Zahlen lässt sich zu dem starken Gesetz der großen Zahlen verschärfen, bei dem an die Stelle der stochastischen Konvergenz die fast sichere Konvergenz tritt. Zum Nachweis brauchen wir das Lemma von Borel und Cantelli, für das wiederum die folgende einfache Aussage aus der Theorie der unendlichen Reihen wichtig ist.

Hilfssatz 5.1.4

Sei (a_n) eine Folge von reellen Zahlen mit $0 < a_n < 1$ für alle $n \geq 1$. Dann folgt aus $\sum_{n\geq 1} a_n = +\infty$:

$$\prod_{k\geq 1}(1 - a_k) = 0.$$

Beweis: Seien

$$K_n = \prod_{k=1}^{n}(1-a_k) \qquad \text{und} \qquad L_n = \sum_{k=1}^{n} a_n \qquad n \geq 1.$$

Da die a_n zwischen null und eins liegen, erhalten wir:

$$\ln K_n = \sum_{k=1}^{n} \ln(1-a_k) = -\sum_{k=1}^{n}\left(a_k + \frac{a_k^2}{2} + \frac{a_k^3}{3} + \ldots\right) \leq -\sum_{k=1}^{n} a_k = -L_n.$$

Für $n \to \infty$ folgt nun $-L_n = \ln K_n \to -\infty$. Das heißt aber $K_n \to 0$.

Lemma 5.1.5 *Lemma von Borel-Cantelli*

Sei (A_n) eine Folge von Ereignissen mit den jeweiligen Wahrscheinlichkeiten $\mathrm{P}(A_n)$. A sei das Ereignis, dass unendlich viele der A_n eintreten, $A = \cap_{n\geq 1} \cup_{m\geq n} A_m$. Dann gilt

$$\sum_{n=1}^{\infty} \mathrm{P}(A_n) < \infty \quad \Rightarrow \quad \mathrm{P}(A) = 0.$$

Sind die Ereignisse zudem unabhängig, so gilt auch

$$\sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty \quad \Rightarrow \quad \mathrm{P}(A) = 1.$$

Beweis: Sei $B_n = \cup_{m\geq n} A_m$, $n \geq 1$. Dann ist

$$A = \bigcap_{n\geq 1} \bigcup_{m\geq n} A_m = \bigcap_{n\geq 1} B_n \subseteq B_n \qquad n \geq 1.$$

Daraus folgt:

$$\mathrm{P}(A) = \mathrm{P}\left(\bigcap_{n\geq 1} B_n\right) \leq \mathrm{P}(B_N) = \mathrm{P}\left(\bigcup_{m\geq N} A_m\right) \leq \sum_{m=N}^{\infty} \mathrm{P}(A_m) \qquad N \geq 1.$$

Aus der Konvergenz der Reihe $\sum \mathrm{P}(A_m)$ folgt, dass der rechts stehende Term für $N \to \infty$ gegen null geht, so dass der erste Teil gezeigt ist. Die in der letzten Zeile verwendete Boolesche Ungleichung $\mathrm{P}(\cup A_m) \leq \sum \mathrm{P}(A_m)$ gilt für jede Folge von Ereignissen, also auch wenn diese abhängig sind.

Sind die A_n unabhängig, so sind es auch die komplementären Ereignisse A_n^c und wir erhalten:

$$1 - \mathrm{P}(A) = \mathrm{P}(A^c) = \mathrm{P}\left(\bigcup_{n\geq 1} \bigcap_{m\geq n} A_m^c\right) \leq \sum_{n\geq 1} \mathrm{P}\left(\bigcap_{m\geq n} A_m^c\right)$$

$$= \sum_{n\geq 1} \prod_{m\geq n} \mathrm{P}(A_m^c) = \sum_{n\geq 1} \prod_{m\geq n} [1 - \mathrm{P}(A_m)].$$

Nach dem Hilfssatz ist der letzte Term gleich null, wenn $\sum \mathrm{P}(A_n) = \infty$. Damit ist auch der zweite Teil gezeigt.

Satz 5.1.6 *Starkes Gesetz der großen Zahlen*

$X_1, X_2, \ldots, X_n, \ldots$ sei eine Folge von unabhängigen Zufallsvariablen mit $\mathrm{E}(X_i) = \mu$, $\mathrm{E}(X_i - \mu)^2 = \sigma^2$ und $\mathrm{E}(X_i - \mu)^4 = \mu_4 < \infty$. Dann gilt für $\bar{X}_n$ das *starke Gesetz der großen Zahlen*, d.h.

$$\bar{X}_n \overset{f.s.}{\to} \mu.$$

Beweis: Wir setzen $Y_i = X_i - \mu$ und $A_n = \{|\bar{Y}_n| > \epsilon\}$ für festes, beliebiges $\epsilon > 0$. Mit der Markoffschen Ungleichung erhalten wir:

$$\mathrm{P}(A_n) = \mathrm{P}(|\bar{Y}_n| > \epsilon) \leq \frac{1}{\epsilon^4}\mathrm{E}(\bar{Y}_n^4) = \frac{1}{\epsilon^4} \cdot \frac{1}{n^4} \sum_{i,j,k,l} \mathrm{E}(Y_i Y_j Y_k Y_l).$$

Wegen der Unabhängigkeit und wegen $\mathrm{E}(Y_i) = 0$ sind nur die Summanden von Null verschieden, bei denen alle Indizes gleich sind oder bei denen zwei Paare gleicher Indizes vorkommen. Das sind n bzw. $3n(n-1)$ Summanden. Das ergibt:

$$\mathrm{E}(\bar{Y}_n^4) = \frac{1}{n^4}(n\mu_4 + 3n(n-1)\sigma^2) = \frac{\mu_4}{n^3} + \frac{3\sigma^4}{n^2} - \frac{3\sigma^4}{n^3}.$$

Da $\sum n^{-\tau} < \infty$ für $\tau > 1$, erhalten wir mit dem Lemma von Borel-Cantelli:

$$\mathrm{P}\left(\bigcup_{n\geq 1} \bigcap_{m \geq n} A_m\right) = 0.$$

Das war aber zu zeigen. Denn das Ereignis $A = \cap_{n\geq 1} \cup_{m \geq n} A_m$ bedeutet gerade, dass eine Realisation der Folge $|\bar{X} - \mu|$ divergiert: Für die Divergenz muss es ein $\epsilon > 0$ geben, so dass zu jedem n mindestens ein $k > n$ existiert mit $|\bar{X}_k - \mu| > \epsilon$. Da die Wahrscheinlichkeit für die Divergenz null ist, haben wir mit anderen Worten

$$\mathrm{P}\left(\lim_{n\to\infty} |\bar{X}_n - \mu| = 0\right) = 1.$$

Die Aussage des Satzes ist speziell für Anteile gültig, da für dichotome Variablen $X_i \sim \mathscr{B}(1, p)$ gilt: $\bar{X} = \sum_{i=1}^n X_i/n$. Wir haben damit die *starke Fassung des Theorems von Bernoulli*. Auch auf die empirische Verteilungsfunktion ist das Ergebnis anwendbar, so dass bei einer Folge von Stichprobenvariablen $X_1, X_2, \ldots, X_n, \ldots$ mit der theoretischen Verteilungsfunktion $F(x)$ und der empirischen $\hat{F}_n(x) = \sum_{i=1}^n I_{[X_i \leq x]}/n$ gilt:

$$\mathrm{P}\left(\lim_{n\to\infty} |\hat{F}_n(x) - F(x)| = 0\right) = 1.$$

Tatsächlich lässt sich aber noch viel mehr zeigen, und zwar dass $\hat{F}_n(x)$ gleichmäßig gegen $F(x)$ konvergiert. Wegen seiner Wichtigkeit für die Statistik wird dieses Resultat auch als *Hauptsatz der Statistik* bezeichnet.

Satz 5.1.7 *Satz von Glivenko-Cantelli*

Sei $X_1, X_2, \ldots$ eine Folge von unabhängigen identisch verteilten Zufallsvariablen mit der Verteilungsfunktion $F(x)$. Sei $\hat{F}_n(x)$ die empirische Verteilungsfunktion. Dann gilt:

$$\mathrm{P}\left(\lim_{n\to\infty} \sup_{-\infty<x<+\infty} |\hat{F}_n(x) - F(x)| = 0\right) = 1.$$

Beweis: Wir zeigen, dass

$$\sum_{k\geq n} \mathrm{P}\left(\left\{\sup_{-\infty<x<\infty} |\hat{F}_k(x) - F(x)| > \frac{1}{m}\right\}\right)$$

für festes $m \geq 1$ endlich ist. Dann folgt mit dem Lemma von Borel-Cantelli

$$\mathrm{P}\left(\bigcap_{n\geq 1}\bigcup_{k\geq n}\left\{\sup_{-\infty<x<\infty} |\hat{F}_k(x) - F(x)| > \frac{1}{m}\right\}\right) = 0.$$

Das ist, wie schon bei dem starken Gesetz der großen Zahlen, gerade die zu zeigende Beziehung.

Für die einzelnen Summanden betrachten wir die Zerlegung der reellen Achse mittels der $1/n'$-Quantile $x_{k:n'}$, $n' = \sqrt{n}$, von F:

$$F(x_{k:n'}) = \frac{k}{n'}, \quad k = 1, \ldots, n' \quad \text{und} \quad x_{0:n} = -\infty,\ x_{n'+1:n'} = \infty.$$

Dann gilt für x mit $x_{k-1:n'} < x < x_{k:n'}$, wenn $\hat{F}_n(x_{k:n'}) - F(x) \leq 0$:

$$0 \geq \hat{F}_n(x) - F(x) \geq \hat{F}_n(x_{k-1:n'}) - F(x_{k:n'}) = \hat{F}_n(x_{k-1:n'}) - F(x_{k-1:n'}) - \frac{1}{n'},$$

und wenn $\hat{F}_n(x) - F(x) > 0$:

$$0 < \hat{F}_n(x) - F(x) \leq \hat{F}_n(x_{k:n}) - F(x_{k-1:n}) = \hat{F}_n(x_{k:n}) - F(x_{k:n}) + \frac{1}{n'}.$$

Zusammen folgt:

$$|\hat{F}_n(x) - F(x)| \leq \frac{1}{n'} + \max\{|\hat{F}_n(x_{k:n-1}) - F(x_{k-1:n})|, |\hat{F}_n(x_{k:n}) - F(x_{k:n})|\}.$$

Dies ergibt:

$$\mathrm{P}\left(\sup_{-\infty<x<+\infty} |\hat{F}_n(x) - F(x)| \geq \frac{1}{m}\right) \leq \mathrm{P}\left(\max_k |\hat{F}_n(x_{k:n}) - F(x_{k:n})| \geq \frac{1}{m} - \frac{1}{n'}\right)$$

$$= \mathrm{P}\left(\bigcup_{k=1}^{n'}\left\{|\hat{F}_n(x_{k:n}) - F(x_{k:n})| \geq \frac{1}{m} - \frac{1}{n'}\right\}\right) \leq \sum_{k=1}^{n'} \mathrm{P}\left(\left\{|\hat{F}_n(x_{k:n}) - F(x_{k:n})| \geq \frac{1}{m} - \frac{1}{n'}\right\}\right)$$

$$\leq \sqrt{n} \cdot 2 \cdot \exp\left[-\frac{n}{2}\left(\frac{1}{m} - \frac{1}{\sqrt{n}}\right)^2\right].$$

Bei der letzten Abschätzung haben wir die Abschätzung von Uspensky verwendet, siehe Lemma 2.1.14.

Nun ist also

$$\sum_{k\geq n} \mathrm{P}\left(\left\{\sup_{-\infty<x<+\infty} |\hat{F}_k(x)-F(x)| > \frac{1}{m}\right\}\right) \leq \sum_{k\geq n} \sqrt{k}\cdot 2\cdot \exp\left[-\frac{k}{2}\left(\frac{1}{m}-\frac{1}{\sqrt{k}}\right)^2\right].$$

Da $\sum_{k=0}^{\infty} e^{-\alpha k} = \frac{1}{1-e^{\alpha}}$ und folglich

$$\sum_{k=0}^{\infty} k\, e^{-\alpha k} = -\frac{d}{d\alpha}\frac{1}{1-e^{\alpha}} = \frac{e^{\alpha}}{(1-e^{\alpha})^2}$$

endlich ist, gilt die Behauptung.

Von herausragender Bedeutung für die Statistik ist vielfach die Kenntnis der Verteilungen von Summen von Zufallsvariablen. Hier spielt die Normalverteilung eine zentrale Rolle. Denn bei großen Stichprobenumfängen lassen sich die Verteilungen von Summen vielfach durch eine Normalverteilung approximieren. Eine Beweismethode basiert wesentlich auf einer Eigenschaft der momenterzeugenden Funktionen.

Satz 5.1.8 *schwache Konvergenz als Folge der Konvergenz momenterzeugender Funktionen*

Sei $X_1, X_2, \ldots, X_n, \ldots$ eine Folge von Zufallsvariablen mit momenterzeugenden Funktionen $M_n(t)$, die alle für $t \in (-h, h)$ existieren mögen. Gibt es dann eine Zufallsvariable X mit der momenterzeugenden Funktion $M(t)$, die für alle $t \in (-h', h')$ existiert, wobei $h' \leq h$, und gilt $M_n(t) \to M(t)$ für $n \to \infty$, so gilt auch

$$X_n \xrightarrow{V} X.$$

Beispiel 5.1.9 *asymptotische Poisson-Verteilung von binomialverteilten Zufallsvariablen*

Sei $X_n \sim \mathcal{B}(n, p_n)$ wobei $np_n = \lambda$ für alle $n \geq 1$. Dann gilt

$$M_n(t) = (1-p_n+p_n e^t)^n = \left(1-\frac{\lambda}{n}+\frac{\lambda}{n}e^t\right)^n = \left(1+\frac{\lambda(e^t-1)}{n}\right)^n \to \exp[\lambda(e^t-1)].$$

Das ist gerade die momenterzeugende Funktion einer Poisson-verteilten Zufallsvariablen. Somit sind die X_n asymptotisch Poisson-verteilt.

Nun zu der bereits angesprochenen Normalverteilung.

Definition 5.1.10 *zentraler Grenzwertsatz*

Sei $X_1, \ldots, X_n, \ldots$ eine Folge von unabhängigen Zufallsvariablen mit Erwartungswerten μ_n und Varianzen σ_n^2, $0 < \sigma_n^2 < \infty$. Für die Folge gilt *der zentrale Grenzwertsatz*, wenn die Verteilungen der Folge der standardisierten Summen

$$S_n := \frac{\sum_{i=1}^{n}(X_i-\mu_i)}{\sigma_1^2+\ldots+\sigma_n^2} \qquad n = 1, 2, \ldots$$

gegen die Standard-Normalverteilung konvergiert:

$$\lim_{n\to\infty} \mathrm{P}(S_n \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-x^2}{2}\right] \mathrm{d}x. \tag{5.5}$$

Wir zeigen, dass der zentrale Grenzwertsatz gilt, wenn die zugrunde liegende Folge aus identisch verteilten Zufallsvariablen besteht, welche zudem eine momenterzeugende Funktion besitzen.

Satz 5.1.11 *Gültigkeit des zentralen Grenzwertsatzes bei iid Zufallsvariablen*

Sei $X_1, X_2, \ldots, X_n, \ldots$ eine Folge von identisch verteilten unabhängigen Zufallsvariablen mit einer momenterzeugender Funktion, die für alle $t \in (-h, h)$ existiere. Dann gilt für diese Folge der zentrale Grenzwertsatz.

Beweis: Sei $\mathrm{E}(X_i) = \mu$ und $\mathrm{Var}(X_i) = \sigma^2$. $(X_i - \mu)/\sigma$ habe die momenterzeugende Funktion $M(t)$. Dann hat

$$S_n := \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^{n} (X_i - \mu)$$

die momenterzeugende Funktion $M(t/\sqrt{n})^n$.

Wir haben nun zu zeigen, dass

$$\lim_{n\to\infty} M\left(\frac{t}{\sqrt{n}}\right)^n = \exp\left[\frac{t^2}{2}\right].$$

Aufgrund der Standardisierung der X_i gilt für $M(t)$: $M(0) = 1$, $M'(0) = \mathrm{E}((X_i - \mu)/\sigma) = 0$, $M''(0) = \mathrm{E}((X_i - \mu)/\sigma)^2 = 1$. Die Taylorentwicklung von $M(t)$ um 0 bis zum zweiten Glied gibt:

$$M(t) = 1 + M'(0) \cdot t + \frac{1}{2} M''(r) \cdot t^2 = 1 + \frac{1}{2} t^2 M''(r)$$

für einen Zwischenwert r, $|r| < |t|$. Damit und mit der Stetigkeit der e-Funktion folgt:

$$\begin{aligned}
\lim_{n\to\infty} M\left(\frac{t}{\sqrt{n}}\right)^n &= \lim_{n\to\infty} \left(1 + \frac{t^2}{2n} M''(s)\right)^n, \qquad |s| < \frac{|t|}{\sqrt{n}} \\
&= \exp\left[\lim_{n\to\infty} n \cdot \ln\left(1 + \frac{t^2}{2n} M''(s)\right)\right] \\
&= \exp\left[\lim_{n\to\infty} \frac{t^2}{2} M''(s) \cdot \frac{\ln(1 + t^2 \cdot M''(s)/2n) - \ln(1)}{t^2 \cdot M''(s)/2n}\right].
\end{aligned}$$

Da mit der Existenz der momenterzeugenden Funktion auch alle Ableitungen existieren, ist insbesondere $M''(t)$ stetig, so dass $\lim_{t\to 0} M''(t) = M''(0) = 1$. Wegen $|s| < |t|/\sqrt{n}$ gilt $s \to 0$ für $n \to \infty$. Das ergibt

$$\lim_{n\to\infty} M''(s) = M''(0) = 1.$$

Weiter gilt damit für $h = t^2 \cdot M''(s)/2n$:

$$\lim_{n\to\infty} h = 0$$

Zusammen folgt, da die Ableitung des Logarithmus bei eins gleich eins ist:

$$\lim_{n\to\infty} M\left(\frac{t}{\sqrt{n}}\right)^n = \exp\left[\frac{t^2}{2} \cdot \lim_{n\to\infty} M''(s) \cdot \lim_{h\to\infty} \frac{\ln(1+h)-\ln(1)}{h}\right] = \exp\left[\frac{t^2}{2}\right].$$

Beispiel 5.1.12 *Grenzwertsatz von De Moivre und Laplace*

Bei nicht zu kleiner Wahrscheinlichkeit p kann die Binomialverteilung für große n durch die Normalverteilung approximiert werden. Dies ist so zu sehen: Sind die X_i unabhängig $B(1,p)$-verteilt, so gilt:

$$S_n = (X_1 + \ldots + X_n) \text{ ist } \mathscr{B}(n,p)\text{-verteilt}$$

und es ist

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\right) = \Phi(x). \tag{5.6}$$

Sofern also n genügend groß ist, kann die Differenz

$$\left|\mathrm{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\right) - \Phi(x)\right|$$

vernachlässigt werden.

Der Grenzwertsatz von De Moivre und Laplace gibt also die Möglichkeit, die Binomialverteilung durch die Normalverteilung zu approximieren. Er kann etwas verallgemeinert werden. Mit dieser Verallgemeinerung kann dann die asymptotische Verteilung eines Stichprobenquantils recht einfach hergeleitet werden.

Lemma 5.1.13 *Verallgemeinerung des Grenzwertsatzes von De Moivre und Laplace*

Sei $\vartheta(n)$ eine Folge mit $\epsilon < \vartheta(n) < 1-\epsilon$, wobei $\epsilon > 0$. Weiter sei $k(n)$ eine Folge, $0 < k(n) < n$, so dass

$$\lim_{n\to\infty} \frac{\sqrt{n}(k(n)/n - \vartheta(n))}{\sqrt{\vartheta(n)(1-\vartheta(n))}} = -c, \qquad |c| < \infty.$$

Dann gilt

$$\lim_{n\to\infty} \sum_{j=k(n)}^{n} \binom{n}{k} \vartheta(n)^j (1-\vartheta(n))^{n-j} = 1 - \Phi(-c) = \Phi(c). \tag{5.7}$$

Beweisskizze: Die Aussage des Grenzwertsatzes von DeMoivre-Laplace bzgl. der binomialverteilten Zufallsvariablen S_n,

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\right) = \Phi(x),$$

lässt sich auch in die folgende Form bringen:

$$\lim_{n\to\infty} \sum_{0\le j\le k(n)} \binom{n}{j} p^j (1-p)^{n-j} = \Phi(c) = 1-\Phi(-c),$$

wobei

$$k(n) = n\left(\frac{c}{\sqrt{n}} - \sqrt{p(1-p)} + p\right) \quad \text{bzw.} \quad \frac{\sqrt{n}(k(n)/n-p)}{\sqrt{p(1-p)}} = c.$$

Lassen wir c nun von n abhängen, so dass $\lim_{n\to\infty} c_n = c$, so ist einsichtig, dass die Grenzwertaussage richtig bleibt. Das gilt auch, wenn dazu p in der angegebenen Weise durch eine Folge $p(n)$ ersetzt wird.

Satz 5.1.14 *asymptotische Verteilung eines Stichprobenquantils*

X sei eine Zufallsvariable mit der Dichte $f(x)$ und der Verteilungsfunktion $F(x)$. Das Verteilungsquantil q_α sei eindeutig definiert durch $F(q_\alpha) = \alpha$, $0 < \alpha < 1$. Weiter sei $X_1, X_2, \ldots, X_n, \ldots$ eine Zufallsstichprobe und $X_{k:n}$, $k = k(n) = [n\alpha] + 1$, die k'te geordnete Statistik aus den jeweils ersten n Beobachtungen. Dann gilt:

$$\mathrm{P}\left((X_{k:n} - q_\alpha)\cdot f(q_\alpha)\cdot\sqrt{\frac{n}{\alpha(1-\alpha)}} < z\right) = \Phi(z), \tag{5.8}$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist.

Beweis: Mit

$$X_{k:n} \le x \quad \Leftrightarrow \quad \sum_{j=1}^{n} I_{[X_j\le x]} \ge k(n).$$

haben wir die Darstellung

$$\mathrm{P}(X_{k:n} \le x) = \sum_{j=k(n)}^{n} \binom{n}{j} F(x)^j (1-F(x))^{n-j}.$$

Wir betrachten nun die Verteilung von $X_{k:n}$ in einer Umgebung von q_α. Genauer lassen wir x^* mit n gegen q_α wandern. Dazu setzen wir $x^* = q_\alpha + x/\sqrt{n}$. Zunächst gilt:

$$\mathrm{P}(I_{[X_j\le q_\alpha + x/\sqrt{n}]} = 1) = F\left(q_\alpha + \frac{x}{\sqrt{n}}\right) = \alpha + \frac{x}{\sqrt{n}} f(q_\alpha) + r_n = \vartheta(n).$$

Dabei wurde $F(x)$ in eine Taylorreihe um q_α entwickelt. Für das Restglied r_n gilt, dass $r_n\sqrt{n} \to 0$ für $n\to\infty$. Somit ist

$$\mathrm{P}\left(\sqrt{n}(X_{k:n} - q_\alpha) \le x\right) = \sum_{j=k(n)}^{n} \binom{n}{j} \vartheta(n)^j (1-\vartheta(n))^{n-j}.$$

Um das Lemma anzuwenden, haben wir $\vartheta(n)$ und $\sqrt{n}[k(n)/n - \vartheta(n)]/\sqrt{\vartheta(n)(1-\vartheta(n))}$ zu untersuchen. Es gilt $\vartheta(n) \to \alpha$ für $n\to\infty$. Daher ist für den Fall $0 < \alpha < 1$ die Bedingung an $\vartheta(n)$ erfüllt. Für den Nenner gilt eben damit: $\sqrt{\vartheta(n)(1-\vartheta(n))} \to \sqrt{\alpha(1-\alpha)}$.

Für den Zähler erhalten wir aufgrund der Definition von $k(n)$:

$$\sqrt{n}\left(\frac{k(n)}{n}-\vartheta(n)\right)=\frac{[n\alpha]+1-n\alpha}{\sqrt{n}}-xf(q_\alpha)-r_n\sqrt{n}.$$

Der Ausdruck geht offensichtlich gegen $-xf(q_\alpha)$ für $n\to\infty$. Insgesamt folgt

$$\frac{\sqrt{n}(k(n)/n-\vartheta(n))}{\sqrt{\vartheta(n)(1-\vartheta(n))}}\to\frac{-xf(q_\alpha)}{\sqrt{\alpha(1-\alpha)}}=-c.$$

Das impliziert

$$\mathrm{P}(\sqrt{n}(X_{k:n}-q_\alpha)\le x)\to\Phi(c)$$

bzw.

$$\mathrm{P}\left(\frac{\sqrt{n}(X_{k:n}-q_\alpha)}{\sqrt{\alpha(1-\alpha)/f(q_\alpha)}}\le c\right)\to\Phi(c).$$

5.2 Die Delta-Methode

Die δ-Methode ist eine allgemeine Technik zur Bestimmung von asymptotischen Verteilungen und damit auch von asymptotischen Erwartungswerten, Varianzen und Kovarianzen. Insbesondere ist sie geeignet, um die Asymptotik nichtlinearer Funktionen von Folgen von Statistiken, die asymptotisch normalverteilt sind, zu behandeln. Diese nichtlinearen Funktionen haben die gleiche asymptotische Verteilung wie ihre mittels Taylor-Entwicklung linearisierte Versionen.

Für die folgenden Aussagen benötigen wir noch die folgende Definition.

Definition 5.2.1 *Landausche Symbole*

Für zwei Funktionen $g(x)$ und $f(x)$ schreiben wir

$$g(x)=o(f(x)) \tag{5.9}$$

(lies: $g(x)$ ist klein oh von $f(x)$), wenn gilt:

$$\lim_{x\to 0}\frac{g(x)}{f(x)}=0.$$

Wir schreiben

$$g(x)=O(f(x)) \tag{5.10}$$

(lies: $g(x)$ ist groß oh von $f(x)$), wenn es eine Konstante k gibt, so dass gilt:

$$\lim_{x\to 0}\left|\frac{g(x)}{f(x)}\right|\le k.$$

Satz 5.2.2 *univariate Delta-Methode*

$X_1, X_2, \ldots$ sei eine Folge von Zufallsvariablen, so dass $\sqrt{n}(X_n - \mu)$ asymptotisch $\mathcal{N}(0, \sigma^2)$-verteilt ist. $g(x)$ sei eine Funktion mit nichtverschwindender Ableitung in μ, $g'(\mu) \neq 0$. Dann ist $\sqrt{n}(g(X_n) - g(\mu))$ asymptotisch normalverteilt:

$$\sqrt{n}(g(X_n) - g(\mu)) \dot{\sim} \mathcal{N}(0, (g'(\mu))^2 \sigma^2). \tag{5.11}$$

Beweis: Einen exakten Beweis findet der Leser z. B. bei Serfling (1980). Wir beschränken uns hier auf eine Beweisskizze, die die wichtigsten Punkte illustriert.

Nach dem Mittelwertsatz der Differentialrechnung hat $g(x)$ in einer δ-Umgebung von μ die Taylor-Entwicklung

$$g(x) = g(\mu) + (x - \mu) g'(\mu) + r(x),$$

wobei für das Restglied gilt:

$$r(x) = o(|x - \mu|).$$

Wegen $\sqrt{n}(X_n - \mu) \dot{\sim} \mathcal{N}(0, \sigma^2)$ nehmen die X_n asymptotisch nur Werte aus dieser δ-Umgebung an:

$$\mathrm{P}(|X_n - \mu| < \delta) \to 1.$$

Bei Vernachlässigung des Ereignisses $|X_n - \mu| \geq \delta$, dessen Wahrscheinlichkeit ja gegen n ull geht, gilt somit für genügend großes n

$$\sqrt{n} \frac{g(X_n) - g(\mu)}{g'(\mu)} = \sqrt{n}(X_n - \mu) + \sqrt{n} \frac{r(X_n)}{g'(\mu)}. \tag{5.12}$$

Der Term $\sqrt{n} \cdot r(X_n)/g'(\mu)$ ist asymptotisch vernachlässigbar. Aus $r(x) = o(|x - \mu|)$ folgt nämlich $|r(x)| < \epsilon |x - \mu|$ für alle x aus einer von ϵ abhängigen δ-Umgebung von μ. Somit ist für genügend großes n:

$$\mathrm{P}(|\sqrt{n} \cdot r(X_n)| \leq \epsilon) \geq \mathrm{P}(|\sqrt{n}(X_n - \mu)| \leq 1) \to 1.$$

D. h. $\sqrt{n}\mathrm{Var}(X_n)$ konvergiert in Verteilung gegen den Wert null. Damit hat die linke Seite von Gleichung 5.12 dieselbe asymptotische Verteilung wie $\sqrt{n}(X_n - \mu)$:

$$\sqrt{n}(g(X_n) - g(\mu)) \dot{\sim} \mathcal{N}(0, (g'(\mu))^2 \sigma^2).$$

Beispiel 5.2.3 *heuristische Anwendung der Delta-Methode*

Oft wird dieser Satz in recht heuristischer Weise angewendet, um Varianzen von nichtlinearen Funktionen approximativ zu berechnen.

Es sei etwa $X \sim \mathcal{N}(\mu, \sigma^2)$ mit $\mu \neq 0$, und es soll approximativ Erwartungswert und Varianz von $Y = X^2$ angegeben werden, d. h. von $Y = g(X)$ mit $g(x) = x^2$. Wir entwickeln dazu g in einer Taylor-Reihe um den Erwartungswert μ von X. Dies gibt, wenn nach dem linearen Glied abgebrochen wird:

$$g(x) \doteq g(\mu) + g'(\mu)(x - \mu) = \mu^2 + 2\mu(x - \mu),$$

d. h. eine lineare Funktion von x. Näherungsweise kann daher $Y = X^2$ durch die lineare Funktion $\mu^2 + 2\mu(X-\mu)$ ersetzt werden. Erwartungswert und Varianz dieser linearen Funktion sind:

$$\mathrm{E}(\mu^2 + 2\mu(X-\mu)) = \mu^2 = g(\mu),$$
$$\mathrm{Var}(\mu^2 + 2\mu(X-\mu)) = 4\mu^2\mathrm{Var}(X-\mu) = 4\sigma^2\mu^2 = (g'(\mu))^2\sigma^2.$$

Diese Ergebnisse sind dann Approximationen für $\mathrm{E}(Y)$ und $\mathrm{Var}(Y)$. Streng genommen ist das jedoch nur unter den Bedingungen von Satz 5.2.2 korrekt.

Natürlich kann die Taylor-Reihe auch weiter fortgesetzt werden, wodurch man u. U. bessere Approximation erhält. Entwickeln wir z. B. bis zum zweiten Glied,

$$g(x) \doteq g(\mu) + g'(\mu)(x-\mu) + \frac{1}{2}g''(\mu)(x-\mu)^2,$$

so haben wir approximativ:

$$\mathrm{E}(g(X)) \doteq g(\mu) + \frac{1}{2}g''(\mu)\sigma^2,$$
$$\begin{aligned}\mathrm{Var}(g(X)) &\doteq g'(\mu)^2\mathrm{Var}(X-\mu) + \frac{1}{4}g''(\mu)^2\mathrm{Var}((X-\mu)^2) + g'(\mu)g''(\mu)\mathrm{Cov}(X-\mu,(X-\mu)^2)\\ &\doteq g'(\mu)^2\sigma^2 + \frac{1}{4}g''(\mu)^2\mathrm{E}((X-\mu)^4) - \frac{1}{4}g''(\mu)^2\sigma^2 + g'(\mu)g''(\mu)\mathrm{E}((X-\mu)^3).\end{aligned}$$

Häufig wird diese Näherung nur für den Erwartungswert benutzt.

Beispiel 5.2.4 *asymptotische Verteilung der empirischen Varianz bei $\mathscr{B}(1,\mu)$-Verteilung*

$X_1, X_2, \ldots, X_n, \ldots$, sei eine Folge von unabhängigen $\mathscr{B}(1,\mu)$-verteilten Zufallsvariablen, d. h. $\mathrm{P}(X_i = 0) = 1-\mu$, $\mathrm{P}(X_i = 1) = \mu$. Der Anteil $\bar{X}_n = \sum_{i=1}^n X_i/n$ der Einsen (Erfolge) unter den ersten n Beobachtungen ist nach dem zentralen Grenzwertsatz asymptotisch normalverteilt

$$\sqrt{n}(\bar{X}_n - \mu) \dot{\sim} \mathscr{N}(0, \mu(1-\mu)).$$

Wir wollen die asymptotische Verteilung der empirischen Varianz

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

bestimmen. Hier ist $g(x) = x(1-x)$ und $g'(x) = 1-2x$, d. h. $g'(\mu) = 1-2\mu$.

Nach Satz 5.2.2 gilt:

$$\sqrt{n}(Y - \mu(1-\mu)) \dot{\sim} \mathscr{N}(0, \mu(1-\mu)(1-2\mu)^2).$$

Beispiel 5.2.5 *Varianzstabilisierende Transformationen*

Eine wichtige Anwendung sind sogenannte varianzstabilisierende Transformationen: Bei vielen Zufallsvariablen hängt die Varianz σ^2 von X funktional vom Erwartungswert μ ab. Man sucht dann eine Transformation $Y = g(x)$, so dass die approximative Varianz $g'(\mu)^2\sigma^2$ von Y konstant wird. Die wichtigsten Beispiele sind:

(1) Ist X *Poisson-verteilt* mit $\mathrm{E}(X) = \mu$, so gilt auch $\mathrm{Var}(X) = \mu$. Dann hat $Y = \sqrt{X}$ die approximative Varianz $\mu(2\sqrt{\mu})^{-2} = 1/4$.

(2) Für *binomialverteilte* X_n aus Beispiel 5.2.4 gilt $\mathrm{E}(\bar{X}_n) = \mu$, $\mathrm{Var}(\bar{X}_n) = \mu(1-\mu)/n$. Hier hat $Y_n = \arcsin(X_n)$ die asymptotische Varianz $n/4$.

(3) Ist X *exponentialverteilt* mit $\mathrm{E}(X) = \mu$, so ist $\mathrm{Var}(X) = \mu^2$. Die varianzstabilisierende Transformation ist $Y = \ln(X)$ mit $\mathrm{Var}(Y) \doteq (1/\mu)^2\mu^2 = 1$.

(4) Generell erhalten wir für Zufallsvariablen X mit $\mathrm{Var}(X) = \mu^c$ die *Box-Cox-Transformation* $Y = (X^\lambda - 1)/\lambda$ bzw. äquivalent dazu $Y = X^\lambda$ als varianzstabilisierende Transformation. Dabei ist λ aus dem Ansatz

$$(\mu^{\lambda-1})^2\mu^c \stackrel{!}{=} \text{const.}$$

zu ermitteln, d. h. $\lambda = 1 - c/2$. Dies enthält (1) und (3) als Spezialfälle, wenn für $\lambda = 0$ die Transformation $y = \ln(x)$ als Grenzwert einbezogen wird.

(5) Der *empirische Korrelationskoeffizient* $\hat{\rho}$ aus n unabhängigen Beobachtungspaaren einer bivariaten Normalverteilung hat die asymptotische Verteilung:

$$\hat{\rho} \sim \left[\rho, \frac{(1-\rho^2)^2}{n}\right].$$

Die varianzstabilisierende Transformation ist

$$\tilde{\rho} = \operatorname{arctanh}\hat{\rho} = \frac{1}{2}\ln\frac{1+\hat{\rho}}{1-\hat{\rho}}$$

mit $\mathrm{Var}(\tilde{\rho}) \doteq 1/n$. Erfahrungsgemäß wird durch diese Transformation auch die Annäherung an die Normalverteilung wesentlich beschleunigt.

Der folgende Satz verallgemeinert das Resultat von Satz 5.2.2 auf den mehrdimensionalen Fall.

Satz 5.2.6 *multivariate Delta-Methode*

$\boldsymbol{X}_n = (X_{n1}, \ldots, X_{nk})'$, $n = 1, 2, 3, \ldots$ sei eine Folge von k-dimensionalen Zufallsvektoren, die asymptotisch multivariat normalverteilt ist:

$$\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{\mu}) \dot{\sim} \mathcal{N}(\mathbf{0}, \Sigma).$$

$g(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x}))$ sei eine differenzierbare Abbildung des $\mathbb{R}^k$ in den $\mathbb{R}^m$, deren partielle Ableitungen im Punkt $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ nicht sämtlich verschwinden. Dann ist

$g(\boldsymbol{X}_n)$ asymptotisch normalverteilt,

$$\sqrt{n}(g(\boldsymbol{X}_n) - g(\boldsymbol{\mu})) \dot{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}\Sigma\boldsymbol{D}^T),$$

wobei $\boldsymbol{D}$ die Matrix der partiellen Ableitungen an der Stelle $\boldsymbol{x} = \boldsymbol{\mu}$ ist:

$$\boldsymbol{D} = \left[\left. \frac{\partial g_i}{\partial x_j} \right|_{\boldsymbol{x}=\boldsymbol{\mu}} \right].$$

Beweis: Siehe Serfling (1980, S.122).

Beispiel 5.2.7 *Fehlerfortpflanzungsgesetz*

Die heuristische Anwendung des Satzes besteht wieder in der Taylor-Entwicklung bis zu den Gliedern ersten Grades und der Bestimmung des Erwartungswervektors und der Kovarianzmatrix der linearen Approximationen. Sei z. B. $Y = X_1 X_2$ mit $\mathrm{E}(X_i) = \mu_i$, $\mathrm{Var}(X_i) = \sigma_i^2$, $\mathrm{Cov}(X_1, X_2) = \sigma_{12}$.

Dann ist wegen $\partial x_1 x_2 / \partial x_1 = x_2$:

$$Y \doteq \mu_1\mu_2 + \mu_2(X_1 - \mu_1) + \mu_1(X_2 - \mu_2)$$

und

$$\mathrm{E}(Y) \doteq \mu_1\mu_2$$

$$\mathrm{Var}(Y) \doteq \mu_2^2\sigma_1^2 + \mu_1^2\sigma_2^2 + 2\mu_1\mu_2\sigma_{12}.$$

Für den speziellen Fall $\sigma_{12} = 0$ kann das Resultat in der übersichtlichen Form

$$\frac{\mathrm{Var}(Y)}{\mathrm{E}(Y)^2} = \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} \tag{5.13}$$

festgehalten werden. Dies ist Basis des *Fehlerfortpflanzungsgesetzes*: Bei der Bildung von (unabhängigen) Summen addieren sich die *absoluten Fehler* ($\hat{=}\,\sigma_i^2$), bei der Multiplikation die *relativen Fehler* ($\hat{=}\,\sigma_i^2/\mu_i^2$).

Beispiel 5.2.8 *gemeinsame Verteilung zweier Quotienten*

Wir bestimmen die gemeinsame Verteilung zweier Quotienten. Seien Y_{0n}, Y_{1n}, Y_{2n} ($n = 0, 1, 2, \ldots$) gemeinsam asymptotisch normalverteilt mit

$$\mu_i := \mathrm{E}(Y_{in}) = a_i + \frac{b_i}{n} + O(n^{-2}) \qquad a_n \neq 0$$

und

$$\sigma_{ij} := \mathrm{Cov}(Y_{in}, Y_{jn}) = \frac{c_{ij}}{n} + O(n^{-2}).$$

Dann sind Y_{1n}/Y_{0n}, Y_{2n}/Y_{0n} ebenfalls gemeinsam asymptotisch normalverteilt mit

$$\mathrm{E}\left[\frac{Y_{1n}}{Y_{0n}}\right] = \frac{a_1}{a_0} + \frac{1}{n}\left\{\frac{b_1 a_0 - a_1 b_0}{a_0^2} - \frac{c_{01}}{a_0^2} + \frac{a_1 c_{00}}{a_0^3}\right\} + O(n^{-2})$$

$$\mathrm{Var}\left[\frac{Y_{1n}}{Y_{0n}}\right] = \frac{1}{n}\left\{\frac{c_{11}}{a_0^2} + \frac{c_{00}a_1^2}{a_0^4} - \frac{2c_{01}a_1}{a_0^3}\right\} + O(n^{-2})$$

$$\mathrm{Cov}\left[\frac{Y_{1n}}{Y_{0n}}, \frac{Y_{2n}}{Y_{0n}}\right] = \frac{1}{n}\left\{\frac{c_{12}}{a_0} + \frac{c_{00}a_1a_2}{a_0^4} - \frac{c_{01}a_2 + c_{02}a_1}{a_0^3}\right\} + O(n^{-2}).$$

Zur Berechnung des Erwartungswertes wurde die Funktion y_1/y_0 in ein Taylorpolynom um $y_1 = \mu$, $y_0 = \mu_0$ entwickelt, das nach den Termen zweiten Grades abgebrochen wurde. Der Erwartungswert von Y_{1n}/Y_{0n} und der Erwartungswert des Taylorpolynoms unterscheiden sich dann nur noch um $O(\mathrm{E}|Y_{in} - \mu_i|^3)$-Terme. Da für asymptotisch normalverteilte Zufallsvariablen $\sqrt{N}(Y_{in} - \mu_i)$ gilt $\mathrm{E}(|\sqrt{n}(Y_{in} - \mu_i)|^3) = O(N^{-1/2})$, ist $\mathrm{E}(|Y_{in} - \mu_i|^3) = O(n^{-2})$. Die Kovarianzen erhält man durch Abbrechen der Taylorpolynome nach den Gliedern ersten Grades und Benutzung von $\mathrm{E}(|\sqrt{n}(Y_{in} - \mu_i)|^3) = O(1)$.

Die entstehenden Ausdrücke wurden dabei mit Hilfe der Entwicklung

$$\frac{a + b/n + c/n^2}{d + e/n + f/n^2} = \frac{a}{d} + \frac{bd - ae}{d^2}\frac{1}{n} + O(n^{-2})$$

weiter vereinfacht.

5.3 Aufgaben

Aufgabe 1

(i) Gegeben sei die Folge von Zufallsvariablen $X_n : (0,1] \to \mathscr{R}$

$$X_n(\omega) = \begin{cases} 2^k & \text{für } \dfrac{n}{2^k} < \omega \leq \dfrac{n+1}{2} - 1 \\ 0 & \text{sonst} \end{cases}$$

$$P(X_n = 2^k) = 2^{-k}, \quad P(X_n = 0) = 1 - 2^{-k},$$

wobei k eindeutig bestimmt ist durch $n = 2^k + m$ mit $0 \leq m < 2^k$.
Zeigen Sie, dass diese Folge zwar in Wahrscheinlichkeit, aber nicht mit Wahrscheinlichkeit eins konvergiert.

(ii) Die Folge von unabhängigen Zufallsvariablen X_n sei definiert durch:

$$P(X_n = 0) = 1 - 1/n, \quad P(X_n = 1) = 1/n.$$

Zeigen Sie, dass die Folge zwar im quadratischen Mittel, aber nicht mit Wahrscheinlichkeit eins konvergiert.

6 Punktschätzung

Ausgangsüberlegungen

In diesem Kapitel wird das Problem betrachtet, den unbekannten Wert des Parameters ϑ zu schätzen, oder allgemeiner den Wert $g(\vartheta)$ einer Funktion von ϑ. Den Ausgangspunkt dafür bildet, dass eine Vorstellung über die Verteilung einer Zufallsvariablen X in Form der Zugehörigkeit zu einer Verteilungsfamilie $\mathscr{F} = \{F_\vartheta : \vartheta \in \Theta\}$ vorliegt. Die Basis für Aussagen über den unbekannten Parameter ϑ bildet dann eine unabhängige Stichprobe vom Umfang n. Mit der Verteilung von X wird dann auch die gemeinsame Verteilung der X_i durch den Parameter ϑ festgelegt, so dass die Aussage über ϑ tatsächlich auf Basis einer Beobachtung von $X_1, \ldots, X_n$ gemacht werden kann.

Hier wird weitgehend nur der Fall eindimensionaler Funktionen $g(\vartheta)$ berücksichtigt. Neben dem punktuellen Eingehen auf das Schätzen mehrdimensionaler Parameter gehen wir im Abschnitt 6.1.7 jedoch weit darüber hinaus und betrachten die nichtparametrische Schätzung von stetigen Dichtefunktionen.

Definition 6.1 *Schätzfunktion*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine Stichprobe aus einer Verteilung, die zu der Verteilungsfamilie $\mathscr{F} = \{F_\vartheta : \vartheta \in \Theta\}$ gehöre. Eine Statistik $T(\boldsymbol{X})$ heißt *Schätzfunktion* oder kurz *Schätzer* für $g(\vartheta)$, wenn T den Stichprobenraum in den Wertebereich von g abbildet, und wenn der Wert von $T(\boldsymbol{X})$ als Stichprobennäherung für $g(\vartheta)$ genommen werden soll.

Beispiel 6.1 *Schätzfunktion für den Parameter der Poisson-Verteilung*

Sei $\mathscr{F}$ die Familie der Poisson-Verteilungen, die durch den Parameter λ charakterisiert wird. Dann ist $g_1(\lambda) = \lambda$ von primärem Interesse. Eine naheliegende Schätzfunktion für λ ist wegen $E(X) = \lambda$ das arithmetische Mittel. Aber auch die Standardabweichung $g_2(\lambda) = \sqrt{\lambda}$ ist in vielen Fällen von Bedeutung. Dafür bieten sich zwei Schätzer an. Einmal die empirische Standardabweichung und dann, da $\sigma = \sqrt{\lambda}$, auch $\sqrt{\bar{X}}$.

6.1 Schätzmethoden

6.1.1 Substitutionsprinzipien

Viele theoretische Parameter besitzen empirische Gegenstücke, so etwa die Wahrscheinlichkeit eines Ereignisses die relative Häufigkeit, der Erwartungswert das arithmetische Mittel.

Um zu Schätzfunktionen zu gelangen, liegt es dann nahe, die theoretischen Größen durch die entsprechenden empirischen zu ersetzen. Weiter gibt es auch Fälle, in denen eine solche Ersetzung zumindest einen Zugang zu einem theoretischen Parameter liefert.

Im Fall $X \sim \mathcal{B}(n,p)$ ist der Stichprobenanteil $T(X) = X/n$ die naheliegende Schätzung für p. Nahegelegt wird $T(X)$ dabei durch das schwache Gesetz der großen Zahlen, das bei großem n garantiert, dass $T(X)$ mit großer Wahrscheinlichkeit einen Wert in der Nähe vom wahren Parameterwert p annimmt. Sollen allgemeiner bei einer Multinomialverteilung $\mathcal{M}(n,p_1,\ldots,p_k)$ nicht die Wahrscheinlichkeiten $p_1,\ldots,p_k$ selbst geschätzt werden, sondern der Wert einer stetigen Funktion $q(p_1,\ldots,p_k)$, so ist die Schätzfunktion

$$T(X_1,\ldots,X_n) = q\left(\frac{N_1}{n},\ldots,\frac{N_k}{n}\right)$$

plausibel. Dabei ist N_i/n der mit der Wahrscheinlichkeit p_i korrespondierende Stichprobenanteil.

Hängen umgekehrt die Wahrscheinlichkeiten $p_1,\ldots,p_k$ stetig von einem (mehrdimensionalen) Parameter ϑ ab, so kann ϑ oder eine Funktion von ϑ über die Substitution der Häufigkeiten für die Wahrscheinlichkeiten geschätzt werden, wenn die Inverse existiert, ϑ also selbst als stetige Funktion der $p_1,\ldots,p_k$ dargestellt werden kann.

Definition 6.1.1 *Prinzip der Häufigkeitssubstitution*

Seien die Wahrscheinlichkeiten $p_1,\ldots,p_k$ stetige Funktionen des Parameters ϑ und sei $g(\vartheta)$ eine stetige Funktion von ϑ, deren Wert geschätzt werden soll. Es gelte

$$g(\vartheta) = h(p_1(\vartheta),\ldots,p_k(\vartheta)).$$

Dann ist der Häufigkeitssubstitutionsschätzer gegeben durch

$$\widehat{g(\vartheta)} = T(X_1,\ldots,X_n) = h\left(\frac{N_1}{n},\ldots,\frac{N_k}{n}\right). \tag{6.1}$$

Beispiel 6.1.2 *Geschlecht von Kindern*

Das Geschlecht von Kindern innerhalb einer Familie kann kaum als unabhängig angesehen werden. Ein realistisches Modell für die Verteilung des Geschlechts in n Familien mit zwei Kindern geht aus von der multinomialverteilten 2-dimensionalen Zufallsvariablen

X_1 = Anzahl der Familien mit zwei Jungen
X_2 = Anzahl der Familien mit einem Mädchen und einem Jungen.

Dann gilt:

$$\mathrm{P}((X_1,X_2)=(x_1,x_2)) = \frac{n!}{x_1!x_2!(n-x_1-x_2)!}\cdot p_1^{x_1}p_2^{x_2}(1-p_1-p_2)^{n-x_1-x_2}.$$

Die Wahrscheinlichkeiten p_1,p_2 hängen nun von zwei Parametern ab:

$$p_1 = \vartheta_1 \cdot \vartheta_2, \qquad p_2 = 2(1-\vartheta_1)\vartheta_2\,.$$

Einfache Umformungen ergeben:

$$\vartheta_1 = q_1(p_1,p_2) = \frac{p_1}{p_1 + p_2/2}, \qquad \vartheta_2 = q_2(p_1,p_2) = \frac{p_1}{\vartheta_1} = p_1 + \frac{p_2}{2}.$$

Mithin sind ϑ_1 und ϑ_2 stetige Funktionen von (p_1, p_2).

ϑ_2 kann als die Wahrscheinlichkeit für eine Jungengeburt angesehen werden. ϑ_1 erfasst den ‚Häufungseffekt': Im Fall $\vartheta_1 = \vartheta_2$ sind die Geburten innerhalb der Familien unabhängig, für $\vartheta_1 > \vartheta_2$ gibt es überproportional viele gleichgeschlechtliche Kinder.
Für eine bestimmte Bevölkerungsgruppe erhielt man bei $n = 1275$ Familien für die jeweils beiden ältesten Kinder:

$$x_1 = 350, \qquad x_2 = 587.$$

Dies ergibt

$$\hat{\vartheta}_1 = \frac{350/1275}{(350 + 0.5 \cdot 587)/1275} = 0.544 \quad \text{und} \quad \hat{\vartheta}_2 = \frac{350 + 0.5 \cdot 587}{1275} = 0.505.$$

Dass $\hat{\vartheta}_1 > \hat{\vartheta}_2$ gilt, zeigt den positiven Häufungseffekt.

Die Gewinnung von Parameterschätzern mittels Ersetzung der theoretischen durch die empirischen Momente rangiert unter dem Namen der Momentenmethode.

Definition 6.1.3 *Momentenmethode*

Die zu schätzende Funktion $g(\boldsymbol{\vartheta})$ des Parameters $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_p)$ sei eine stetige Funktion der ersten k Momente $\mu_j = \mathrm{E}_{\boldsymbol{\vartheta}}(X^j)\,(j = 1, \ldots, k)$:

$$g(\boldsymbol{\vartheta}) = h(\mu_1(\vartheta), \ldots, \mu_k(\vartheta)).$$

Mit $\hat{\mu}_j = \sum_{j=1}^{n} X_i^j / n$ ist der *Momentenschätzer* für $g(\boldsymbol{\vartheta})$ definiert durch

$$\widehat{g(\boldsymbol{\vartheta})} = T(X_1, \ldots, X_n) = h(\hat{\mu}_1, \ldots, \hat{\mu}_k). \tag{6.2}$$

Es werden also die theoretischen Momente durch die empirischen ersetzt.

Beispiel 6.1.4 *Momentenschätzer bei Negativer Binomialverteilung*

Die Zufallsvariable X sei negativ binomialverteilt, $X \sim \mathcal{NB}(k, p)$; ihre Zähldichte ist entsprechend:

$$f(x; p, k) = \binom{x+k-1}{x} p^k (1-p)^x \qquad x = 0, 1, 2, \ldots$$

Dann gilt:

$$\mathrm{E}(X) = \frac{k \cdot (1-p)}{p}, \qquad \mathrm{E}(X^2) = \frac{k \cdot (1-p)(1+k+kp)}{p^2}\,.$$

Für p und k folgt:

$$p = \frac{\mathrm{E}(X)}{\mathrm{E}(X^2) - \mathrm{E}(X)^2}, \qquad k = \frac{\mathrm{E}(X)^2}{\mathrm{E}(X^2) - \mathrm{E}(X)^2 - \mathrm{E}(X)}.$$

Die Momentenschätzer für p und k lauten also, wenn wir zur Abkürzung $\overline{x^2}$ für $\sum x_i^2/n$ schreiben:

$$\hat{p} = \frac{\bar{x}}{\overline{x^2} - \bar{x}^2}, \qquad \hat{k} = \frac{\bar{x}^2}{\overline{x^2} - \bar{x}^2 - \bar{x}}.$$

Für Substitutions-Schätzer spricht, dass sie i.d.R. einfach zu berechnende Schätzer ergeben, die auch oft intuitiv naheliegen. Leider können sie auch nicht-plausible Schätzungen liefern. Schon das zeigt, dass unter Umständen andere Schätzmethoden sinnvoller sind.

Beispiel 6.1.5 *Momentenschätzer bei Rechteckveteilung*

Es sei eine konkrete Stichprobe aus einer $\mathscr{R}(0,\vartheta)$-Verteilung gegeben:

$$x_1 = 0.1, \qquad x_2 = 0.8, \qquad x_3 = 2.1\,.$$

Der Momentenschätzer für ϑ ist dann wegen $\mathrm{E}(X) = \vartheta/2$:

$$\hat{\vartheta} = 2 \cdot \bar{x} = 2.0\,.$$

Diese Schätzung ist offensichtlich indiskutabel, da bei $\vartheta = 2$ der Wert x_3 nicht möglich wäre.

6.1.2 Die Methode der Kleinsten Quadrate

Die Methode der Kleinsten Quadrate ist neben der im folgenden Abschnitt zu besprechenden Maximum-Likelihood-Methode die wohl verbreitste Schätzmethode. Sie wird in der Regressionsanalyse und davon abgeleiteten Modellen eingesetzt.

Im Signal-plus-Rauschen-Modell,

$$Y = \mu + U,$$

mit $\mathrm{E}(U) = 0$ ist ein plausibler Ansatz, den Parameter μ aus einer Stichprobe $y_1, \ldots, y_n$ so zu bestimmen, dass die Residuen

$$\hat{u}_i = y_i - \hat{\mu}$$

möglichst eng um Null konzentriert sind. Das führt auf das Zielkriterium

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \stackrel{!}{=} \min.$$

Der resultierende Schätzer heißt *Kleinste-Quadrate-Schätzer* für μ.

Im allgemeineren Modell

$$Y_i = g_i(\vartheta_1, \ldots, \vartheta_p) + U_i, \qquad i = 1, \ldots, n,$$

bei dem die Funktionen g_i bekannt sind, kann zur Bestimmung der unbekannten Parameter $\vartheta_1,\ldots,\vartheta_p$ der gleiche Ansatz herangezogen werden. Damit er aber sinnvoll bleibt, sind für die Störungen U_i geeignete Annahmen zu machen:

$$\begin{aligned} &\mathrm{E}(U_i)=0 && 1\le i\le n\\ &\mathrm{Var}(U_i)=\sigma^2 && 1\le i\le n\\ &\mathrm{Cov}(U_i,U_j)=0 && 1\le i<j\le n\,. \end{aligned}$$

Sind die Funktionen g_i differenzierbar und ist der Parameterraum Ω ein offenes p-dimensionales Intervall, so muss die Lösung von

$$\sum_{i=1}^{n}(y_i-g_i(\vartheta_1,\ldots,\vartheta_p))^2 \stackrel{!}{=} \min \tag{6.3}$$

auch die folgenden *Normalgleichungen* erfüllen:

$$\frac{\partial}{\partial\vartheta_j}\sum_{i=1}^{n}(y_i-g_i(\vartheta_1,\ldots,\vartheta_p))^2=0 \qquad j=1,\ldots,p\,,$$

bzw.

$$\sum_{i=1}^{n}[y_i-g_i(\vartheta_1,\ldots,\vartheta_p)]\cdot\frac{\partial}{\partial\vartheta_j}g_i(\vartheta_1,\ldots,\vartheta_p)=0 \qquad j=1,\ldots,p\,. \tag{6.4}$$

Ausgangspunkt für diese Regressionsmodelle bilden die linearen Regressionsmodelle, die auch von der strukturellen Seite her von besonderem Interesse sind.

Beispiel 6.1.6 *lineares Regressionsmodell*

In dem linearen Regressionsmodell

$$Y_i=\vartheta_0+\vartheta_1x_i+\ldots+\vartheta_px_{pi}+U_i, \qquad i=1,\ldots,n,$$

mit $\mathrm{E}(U_i)=0$, $\mathrm{Var}(U_i)=\sigma^2$ für $1\le i\le n$ und $\mathrm{Cov}(U_i,U_j)=0$ für $i\ne j$ führt das Zielkriterium

$$\sum_{i=1}^{n}(y_i-\vartheta_0-\vartheta_1x_i-\ldots-\vartheta_px_{pi})^2 \stackrel{!}{=} \min$$

auf die Normalgleichungen

$$\begin{aligned} &-2\sum_{i=1}^{n}(y_i-\hat\vartheta_0-\hat\vartheta_1x_i-\ldots-\hat\vartheta_px_{pi})=0\\ &-2\sum_{i=1}^{n}(y_i-\hat\vartheta_0-\hat\vartheta_1x_i-\ldots-\hat\vartheta_px_{pi})x_{ji}=0 \qquad j=1,\ldots,p\,. \end{aligned}$$

Diese lassen sich kompakt angeben, wenn wir die Matrizenschreibweise verwenden:

$$\begin{pmatrix} 1 & 1 & \ldots & 1\\ x_{11} & x_{12} & \ldots & x_{1n}\\ \vdots & \vdots & & \vdots\\ x_{p1} & x_{p2} & \ldots & x_{pn}\end{pmatrix}\left[\begin{pmatrix} y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}-\begin{pmatrix} 1 & x_{11} & \ldots & x_{p1}\\ 1 & x_{12} & \ldots & x_{p2}\\ \vdots & \vdots & & \vdots\\ 1 & x_{1n} & \ldots & x_{pn}\end{pmatrix}\right]\begin{pmatrix}\hat\vartheta_0\\ \hat\vartheta_1\\ \vdots\\ \hat\vartheta_0\end{pmatrix}=\begin{pmatrix} 0\\ 0\\ \vdots\\ 0\end{pmatrix}.$$

Mit

$$\mathbf{X} = \begin{pmatrix} 1\, x_{11} \ldots x_{p1} \\ 1\, x_{12} \ldots x_{p2} \\ \vdots\ \vdots \qquad \vdots \\ 1\, x_{1n} \ldots x_{pn} \end{pmatrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$$

wird das lineare Modell zu

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\vartheta} + \boldsymbol{U}\,. \tag{6.5}$$

Die Normalgleichungen lauten mit $\hat{\boldsymbol{\vartheta}} = (\hat{\vartheta}_0, \hat{\vartheta}_1, \ldots, \hat{\vartheta}_p)^T$ und $\mathbf{0} = (0, 0, \ldots, 0)^T$:

$$\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X})\hat{\boldsymbol{\vartheta}} = \mathbf{0}$$

oder

$$\mathbf{X}^T\boldsymbol{X}\hat{\boldsymbol{\vartheta}} = \mathbf{X}^T\boldsymbol{Y}\,. \tag{6.6}$$

Sofern die Inverse von $\mathbf{X}^T\boldsymbol{X}$ existiert, erhalten wir für $\hat{\boldsymbol{\vartheta}}$:

$$\hat{\boldsymbol{\vartheta}} = (\mathbf{X}^T\boldsymbol{X})^{-1}\mathbf{X}^T\boldsymbol{Y}\,. \tag{6.7}$$

Im Fall $p = 1$ konkretisiert sich die vorletzte Gleichung zu

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \hat{\vartheta}_0 \\ \hat{\vartheta}_1 \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}.$$

Daraus erhalten wir die expliziten Lösungen

$$\hat{\vartheta}_0 = \bar{Y} - \hat{\vartheta}_1\bar{x}, \qquad \hat{\vartheta}_1 = \frac{\sum x_i Y_i/n - \bar{x}\cdot\bar{Y}}{\sum x_i^2/n - \bar{x}^2}.$$

Beispiel 6.1.7 *Bremsweg von Kraftfahrzeugen*

Es ist wohlbekannt, dass eine Strecke, die ein Kraftfahrzeug zum Anhalten braucht, von seiner Geschwindigkeit (X) abhängt. Diese Abhängigkeit ist dergestalt, dass der Bremsweg (Y) mit dem Quadrat der Geschwindigkeit wächst. Es wird daher erst nach einer Transformation von X ein lineares Regressionsmodell unterstellt:

$$Y_i = \vartheta_0 + \vartheta_1 x_i^2 + U_i \qquad i = 1, \ldots, 18.$$

Eine Serie von 18 Messungen ergab die folgenden Werte (aus: Mosteller, Fienberg & Rourke (1983)):

Geschw. (mph)	Geschw.^{2}	Bremsweg (Fuß)	Geschw. (mph)	Geschw.^{2}	Bremsweg (Fuß)
20	400	12.3	40	1600	68.9
20	400	16.8	50	2500	121.0
20	400	14.5	50	2500	130.0
30	900	27.0	50	2500	123.2
30	900	34.8	50	2500	116.6
30	900	31.4	60	3600	179.2
30	900	38.4	60	3600	171.6
40	1600	70.6	60	3600	151.3
40	1600	75.5	60	3600	166.4

Die Kleinste-Quadrate-Methode liefert die Schätzwerte $\hat{\vartheta}_0 = -7.241$ und $\hat{\vartheta}_1 = 0.0494$.

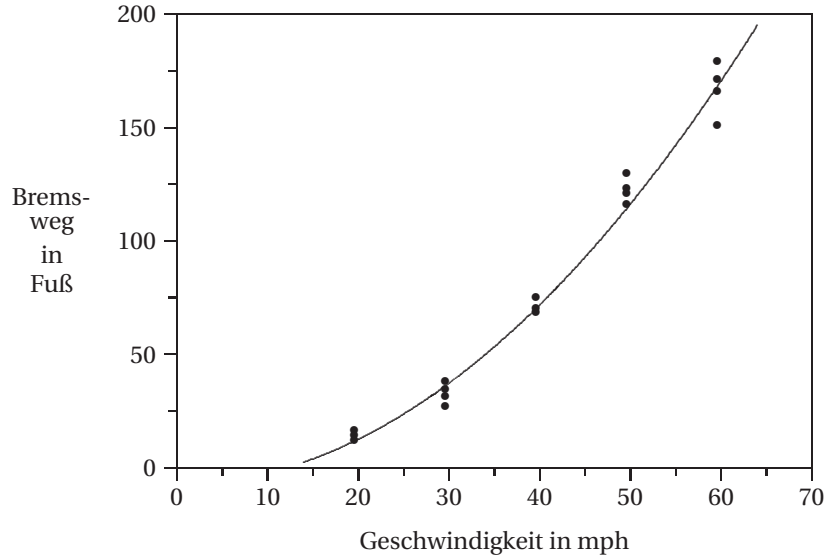


Abb. 6.1: *Geschwindigkeit und Bremsweg*

In vielen Anwendungen ist es unrealistisch, von gleichen Varianzen der Störungen U_i im Regressionsmodell auszugehen. Wenn die Abhängigkeit der Varianzen von den Trägerpunkten x_i sich durch einen multiplikativen Faktor erfassen lässt, $\mathrm{Var}(U_i) = w_i\sigma^2$, spricht man von einem *heteroskedastischen Modell* (im Gegensatz zu einem *homoskedastischen*, bei dem die w_i alle gleich sind). Dann führt die Methode der Kleinsten Quadrate nicht mehr zu optimalen Schätzungen, vgl. Satz 6.2.23. Wenn aber die w_i bekannte Gewichte sind, kann das Modell transformiert werden, so dass wieder optimale Schätzungen erhältlich sind. Für das einfache lineare Regressionsmodell

$$Y_i = \vartheta_0 + \vartheta_1 x_i + U_i, \qquad i = 1, \ldots, n,$$

führt das zu

$$Z_i = \frac{Y_i}{\sqrt{w_i}} = \frac{\vartheta_0 + \vartheta_1 x_i}{\sqrt{w_i}} + \frac{U_i}{\sqrt{w_i}} \qquad i = 1, \ldots, n.$$

Dann ist $\mathrm{Var}(U_i/\sqrt{w_i}) = w_i\sigma^2/(\sqrt{w_i})^2 = \sigma^2$ und die Z_i erfüllen die Voraussetzungen an das Regressionsmodell. Die Methode der Kleinsten Quadrate führt dann auf das Zielkriterium

$$\sum_{i=1}^{n} \frac{1}{w_i}(y_i - \vartheta_0 - \vartheta_1 x_i)^2 \stackrel{!}{=} \min. \tag{6.8}$$

Aus offensichtlichen Gründen spricht man dann von einer *gewichteten Kleinste-Quadrate-Schätzung*. Mit einer Verallgemeinerung der GKQ-Methode lässt sich auch die Situation korrelierter Störungen meistern.

Beispiel 6.1.8 *Durchmesser von Erbsen*

Viele Ideen der Regressionsanalyse erschienen zuerst in den Arbeiten von Sir Francis Galton. Im Rahmen der Diskussion der genetischen Vererbungslehre betrachtete er auch die folgenden Daten über die Größe von Erbsen. Die Erbsen wurden nach ihren Durchmessern in verschiedenen Gruppen eingeteilt. Daraus wurden neue Pflanzen gezogen, und die Durchmesser der Erbsen der nachfolgenden Generation wurden ermittelt. Nur die mittleren Durchmesser und Standardabweichungen wurden mitgeteilt (nach Weisberg (1985)). Die Angaben sind in 1/100 inch.

Durchmesser der Erbsen der Elterngeneration	Mittlerer Durchmesser der Erbsen der Tochtergeneration	Standard-abweichung
21	17.16	1.988
20	17.07	1.938
19	16.37	1.896
18	16.40	2.037
17	16.13	1.654
16	16.17	1.594
15	15.98	1.763

Da zu jedem x-Punkt unterschiedliche y-Punkte gehören, für die auch die Streuungen mitgeteilt sind, ist hier der Ansatz

$$\bar{Y}_i = \vartheta_0 + \vartheta_1 x_i + U_i\,, \qquad \mathrm{Var}(U_i) = w_i^2\sigma^2 \qquad i = 1,\ldots,7$$

mit $w_i = s_i^2$ sinnvoll, da die empirischen Varianzen sinnvolle Schätzungen der theoretischen sind.

6.1.3 Maximum-Likelihood-Methode

Im Kapitel 4 wurde dargelegt, dass die Likelihoodfunktion das Verbindungsglied zwischen einer realisierten Stichprobe $x_1,\ldots,x_n$ und dem zugrunde liegenden Verteilungsmodell darstellt. Das Likelihood-Prinzip sagt, dass alle Information über den unbekannten Parameter ϑ in der Likelihoodfunktion enthalten ist.

Der *Maximum-Likelihood-Ansatz* fordert, als Schätzung für ϑ den ‚plausibelsten' Parameterwert zu wählen; das ist derjenige Wert, bei dem die Likelihoodfunktion ihr Maximum annimmt.

Definition 6.1.9 *Maximum-Likelihood-Schätzer*

Sei $X_1,\ldots,X_n$ eine Zufallsstichprobe aus $f(x;\vartheta)$, $\vartheta\in\Theta$, und sei $L(\vartheta)$ die zugehörige Likelihoodfunktion. Jeder Parameterwert $\hat{\vartheta}$, der die Likelihoodfunktion maximiert,

$$L(\hat{\vartheta}) \geq L(\vartheta) \quad \text{für alle } \vartheta\in\Theta,$$

heißt *Maximum-Likelihood-Schätzwert* für ϑ.

Die *Maximum-Likelihood-Schätzfunktion* $\hat{\vartheta}(X_1,\ldots,X_n)$ ist die Stichprobenfunktion, die jeder Stichprobe den Maximum-Likelihood-Schätzwert zuordnet.

Den *ML-Schätzer für eine Funktion $g(\vartheta)$ des Parameters ϑ* erhalten wir als entsprechende Funktion $g(\hat{\vartheta})$ des ML-Schätzers $\hat{\vartheta}$ von ϑ.

Beispiel 6.1.10 *ML-Schätzer bei einer Gleichverteilung*

Sei $X_1,\ldots,X_n$ eine Zufallsstichprobe aus einer Gleichverteilung,

$$f(x;\vartheta)=\begin{cases}\dfrac{1}{\vartheta} & 0<x<\vartheta \\ 0 & \text{sonst}\end{cases} \qquad \vartheta>0.$$

Für eine Zufallsstichprobe $x_1,\ldots,x_n$ hat die Likelihoodfunktion die folgende Gestalt:

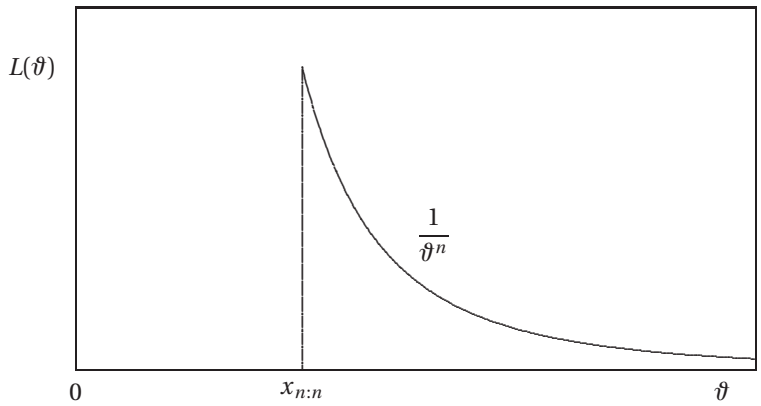


Abb. 6.2: *Likelihoodfunktion der Gleichverteilung*

Daraus ist elementar zu ersehen, dass der ML-Schätzer hier lautet:

$$\hat{\vartheta}(X_1,\ldots,X_n)=X_{n:n}=\max_{1\le i\le n}\{X_i\}\,.$$

Um den ML-Schätzwert numerisch zu bestimmen, wird i.d.R. zur *Loglikelihoodfunktion* übergegangen:

$$\mathscr{L}(\vartheta,x_1,\ldots,x_n)=\ln L(\vartheta,x_1,\ldots,x_n)=\ln(f(x_1;\vartheta)\cdot\ldots\cdot f(x_n;\vartheta))=\sum_{i=1}^{n}\ln f(x_i;\vartheta).$$

$\mathscr{L}$ nimmt sein Maximum an der gleichen Stelle an wie L. Diese Stelle wird durch Nullsetzen der ersten (partiellen) Ableitung(en) von L nach ϑ gefunden. Der ML-Schätzer $\hat{\vartheta}$ ist dann Lösung der *Likelihoodgleichung*

$$\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\ln f(x_i;\vartheta)=\sum_{i=1}^{n}\frac{\partial f(x_i;\vartheta)/\partial\vartheta}{f(x_i;\vartheta)}=0\,. \tag{6.9}$$

Die Likelihoodgleichung bildet die Basis für die Bestimmung der ML-Schätzer, auch wenn das Verschwinden der ersten Ableitungen nur eine notwendige Bedingung für Extrema von differenzierbaren Funktionen ist. Hinreichend dafür, dass an der gefundenen Stelle tatsächlich ein Maximum vorliegt, ist bekanntlich, dass die zweite Ableitung dort kleiner als null ist. Das Erfülltsein dieser hinreichenden Bedingung werden wir im Folgenden i.d.R. nicht weiter nachvollziehen.

Ist der Parameter $\boldsymbol{\vartheta}$ mehrdimensional, so bleibt die Definition der ML-Schätzfunktion natürlich die gleiche. Zur Bestimmung von $(\hat{\vartheta}_1,\ldots,\hat{\vartheta}_p)$ sind dann die partiellen Ableitungen der Loglikelihoodfunktion zu berechnen und das aus dem Nullsetzen resultierende System der *Likelihoodgleichungen* zu lösen:

$$\sum_{i=1}^{n} \frac{\partial f(x_i;\vartheta_1,\ldots,\vartheta_p)/\partial\vartheta_j}{f(x_i;\vartheta_1,\ldots,\vartheta_p)} = 0 \qquad j = 1,\ldots,p\,. \tag{6.10}$$

Beispiel 6.1.11 *ML-Schätzer bei einer Normalverteilung*

Seien $X_1,\ldots,X_n$ unabhängige $\mathcal{N}(\mu,\sigma^2)$-verteilte Zufallsvariablen. Die Loglikelihoodfunktion ist

$$\mathcal{L}(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Die Likelihoodgleichungen sind:

$$\begin{aligned} &\frac{1}{\hat{\sigma}^2}\textstyle\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0,\\ &-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\textstyle\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = 0. \end{aligned}$$

Die ML-Schätzer sind also in diesem Fall $\hat{\mu} = \bar{X}$ und $\hat{\sigma}^2 = \sum(X_i-\bar{X})^2/n$.

Beispiel 6.1.12 *Geschlecht von Kindern - Fortsetzung von Seite 162*

Die gemeinsame Verteilung von (X_1,X_2) mit $X_1 =$ Anzahl der Familien mit zwei Jungen und $X_2 =$ Anzahl der Familien mit einem Mädchen und einem Jungen ist:

$$\mathrm{P}((X_1,X_2)=(x_1,x_2)) = \frac{n!}{x_1!x_2!(n-x_1-x_2)!}\cdot p_1^{x_1}p_2^{x_2}(1-p_1-p_2)^{n-x_1-x_2}$$

wobei $p_1 = \vartheta_1\cdot\vartheta_2,\ p_2 = 2(1-\vartheta_1)\vartheta_2$. Die Loglikelihoodfunktion lautet:

$$\begin{aligned} &\mathcal{L}(\vartheta_1,\vartheta_2,x_1,x_2)\\ &\quad= \ln\left(\frac{n!}{x_1!x_2!(n-x_1-x_2)!}\right) + x_1\cdot\ln(p_1) + x_2\cdot\ln(p_2) + (n-x_1-x_2)\cdot\ln(1-p_1-p_2)\\ &\quad= \ln\left(\frac{n!}{x_1!x_2!(n-x_1-x_2)!}\right) + x_1(\ln\vartheta_1+\ln\vartheta_2)\\ &\qquad + x_2(\ln(1-\vartheta_1)+\ln\vartheta_2) + (n-x_1-x_2)\ln(1+\vartheta_1\vartheta_2-2\vartheta_2). \end{aligned}$$

Somit sind die partiellen Ableitungen:

$$\frac{\partial}{\partial\vartheta_1}\mathscr{L}(\vartheta_1,\vartheta_2,x_1,x_2)=\frac{x_1}{\vartheta_1}-\frac{x_2}{1-\vartheta_1}+\frac{n-x_1-x_2}{1+\vartheta_1\vartheta_2-2\vartheta_2}\cdot\vartheta_2$$
$$\frac{\partial}{\partial\vartheta_2}\mathscr{L}(\vartheta_1,\vartheta_2,x_1,x_2)=\frac{x_1}{\vartheta_2}+\frac{x_2}{\vartheta_2}+\frac{n-x_1-x_2}{1+\vartheta_1\vartheta_2-2\vartheta_2}\cdot(\vartheta_1-2).$$

Nullsetzen ergibt die Likelihoodgleichungen:

$$\frac{x_1}{\hat\vartheta_1}-\frac{x_2}{1-\hat\vartheta_1}+\frac{n-x_1-x_2}{1+\hat\vartheta_1\hat\vartheta_2-2\hat\vartheta_2}\cdot\hat\vartheta_2=0$$
$$\frac{x_1}{\hat\vartheta_2}+\frac{x_2}{\hat\vartheta_2}+\frac{n-x_1-x_2}{1+\hat\vartheta_1\hat\vartheta_2-2\hat\vartheta_2}\cdot(\hat\vartheta_1-2)=0$$

mit den Lösungen:

$$\hat\vartheta_1=\frac{2x_1}{2x_1+x_2},\qquad \hat\vartheta_2=\frac{2x_1+x_2}{2n}.$$

Im letzten Beispiel wurde nicht verifiziert, dass die Likelihoodfunktion bei der gefundenen Lösung der Normalgleichungen tatsächlich ein Maximum annimmt. Dies ist aber im Prinzip möglich. Im folgenden Beispiel wird ein einfaches Maximierungsargument verwendet, um für einen Spezialfall das Vorliegen eines Maximums zu begründen. Der dort anzusprechende *empirische Median* $\tilde x$ ist definiert durch

$$\tilde x=\begin{cases} x_{(n+1)/2:n} & \text{falls } n \text{ ungerade}\\ \dfrac{x_{n/2:n}+x_{1+n/2:n}}{2} & \text{falls } n \text{ gerade}\end{cases};\qquad (6.11)$$

dabei ist $x_{i:n}$ der it'-kleinste Wert von $x_1,\ldots,x_n$.

Beispiel 6.1.13 *ML-Schätzer bei einer Laplace-Verteilung*

Sei $X_1,\ldots,X_n$ eine Zufallsstichprobe aus

$$f(x;\vartheta)=\frac{1}{2}\exp[-|x-\vartheta|],\qquad x\in\mathbb{R},\ \vartheta\in\Theta=\mathbb{R}.$$

Gesucht ist der ML-Schätzer für ϑ. Die Loglikelihoodfunktion lautet, wenn $x_1,\ldots,x_n$ eine feste Stichprobe ist:

$$\mathscr{L}(\vartheta)=-n\cdot\ln(2)-\sum_{i=1}^n|x_i-\vartheta|=-n\cdot\ln(2)-\sum_{i=1}^n\left((x_i-\vartheta)^2\right)^{1/2}.$$

Die Funktion $\mathscr{L}(\vartheta)$ ist überall stetig und bis auf die Stellen $\vartheta=x_1,\ldots,x_n$ differenzierbar. Sofern die Ableitung existiert, gilt:

$$\frac{\partial}{\partial\vartheta}\mathscr{L}(\vartheta)=\mathscr{L}'(\vartheta)=\sum_{i=1}^n\left((x_i-\vartheta)^2\right)^{-1/2}(x_i-\vartheta)=\sum_{i=1}^n\frac{x_i-\vartheta}{|x_i-\vartheta|}.$$

Sei nun $x_{1:n},\ldots,x_{n:n}$ die geordnete Statistik. Für $\vartheta < x_{1:n}$ gilt $\mathscr{L}'(\vartheta) = n$, da $(x_i - \vartheta)/|x_i - \vartheta| = 1$ für alle i. Wenn $x_{1:n} < \vartheta < x_{2:n}$, so ist $\mathscr{L}'(\vartheta) = (n-1)\cdot(+1) + 1\cdot(-1) = n-2$. Bei $x_{2:n} < \vartheta < x_{3:n}$ wird $\mathscr{L}'(\vartheta) = n-4$ usw. $\mathscr{L}'(\vartheta)$ ist also positiv für $i < n/2$ und für $i > (n+1)/2$ negativ.

Da der Graph von $\mathscr{L}(\vartheta)$ ein stetiger Polygonzug ist, folgt, dass $\mathscr{L}(\vartheta)$ für ungerades n auf dem Intervall $(-\infty; x_{(n+1)/2:n}]$ streng monoton wachsend ist und dann monoton fallend. Für gerades n liegt der höchste Punkt auf dem horizontalen Segment $\{(\vartheta, \mathscr{L}(\vartheta)) | x_{n/2:n} < \vartheta < x_{1+n/2:n}\}$. Somit ist der Stichprobenmedian der ML-Schätzer für ϑ.

Satz 6.1.14 *Eindeutigkeit von ML-Schätzern*

Sei T eine suffiziente Statistik für den Parameter ϑ der Verteilungsfamilie $\{f(x_1,\ldots,x_n;\vartheta) | \vartheta \in \Theta\}$. Wenn ein ML-Schätzer $\hat{\vartheta}$ für ϑ existiert und eindeutig ist, so ist er eine Funktion von T.

Beweis: Da T suffizient ist, gilt nach dem Faktorisierungssatz:

$$f(x_1,\ldots,x_n;\vartheta) = h(x_1,\ldots,x_n)\cdot g(T(x_1,\ldots,x_n);\vartheta).$$

Die Maximierung der Likelihoodfunktion bzgl. ϑ ist daher äquivalent zur Maximierung von $g(T(x_1,\ldots,x_n);\vartheta)$. Diese Funktion hängt aber nur über T von $x_1,\ldots,x_n$ ab.

Die Bedeutung der Eindeutigkeit des ML-Schätzers in der vorigen Aussage wird durch das folgende Beispiel deutlich.

Beispiel 6.1.15 *Nichteindeutigeit von ML-Schätzern*

Die Wahrscheinlichkeitsfunktion der Zufallsvariablen X sei

$$f(x;\vartheta) = \begin{cases} 0.25 & x = 1,2 \\ 0.25 + 0.25\cdot\vartheta & x = 3 \\ 0.25 - 0.25\cdot\vartheta & x = 4 \end{cases},$$

wobei $\vartheta \in \Theta = [0;1]$. Bei $n = 1$ sind alle Stichprobenfunktionen

$$\hat{\vartheta}(x) = \begin{cases} a & x = 1 \\ b & x = 2 \\ 1 & x = 3 \\ 0 & x = 4 \end{cases}$$

mit $0 \le a, b \le 1$ ML-Schätzer für ϑ. Für $a \neq b$ ist $\hat{\vartheta}(x)$ keine Funktion der suffizienten Statistik

$$T(x) = \begin{cases} 2 & x = 1,2 \\ 1 & x = 3 \\ 0 & x = 4 \end{cases}.$$

6.1.4 Numerische Bestimmung von ML-Schätzern

Anhand eines Beispiels soll zuerst verdeutlicht werden, dass der ML-Schätzer in vielen praktisch relevanten Anwendungen nicht explizit angegeben werden kann. Dann hilft nur noch die Bestimmung mittels einer geeigneten numerischen Routine.

Beispiel 6.1.16 *gestutzte Poisson-Verteilung*

X habe eine gestutzte Poisson-Verteilung:

$$f(x;\lambda)=\frac{e^{-\lambda}}{1-e^{-\lambda}}\cdot\frac{\lambda^x}{x!}\qquad x=1,2,3,\ldots$$

Der Parameter λ soll mit der Maximum-Likelihood-Methode aus einer Stichprobe vom Umfang n geschätzt werden.

Die Likelihoodfunktion lautet:

$$L(\lambda)=\prod_{i=1}^{n}f(x_i;\lambda)=\prod_{i=1}^{n}\frac{e^{-\lambda}}{1-e^{-\lambda}}\cdot\frac{\lambda^{x_i}}{x_i!}.$$

Das führt auf die Loglikelihoodfunktion:

$$\mathscr{L}(\lambda)=-n\lambda-n\cdot\ln\left(1-e^{-\lambda}\right)+\left(\sum_{i=1}^{n}x_i\right)\cdot\ln(\lambda)-\sum_{i=1}^{n}\ln(x_i!).$$

Nullsetzen der ersten Ableitung liefert:

$$-n-n\cdot\frac{\exp(-\hat{\lambda})}{1-\exp(-\hat{\lambda})}+\sum_{i=1}^{n}x_i\cdot\frac{1}{\lambda}=0\quad\Rightarrow\quad\frac{\hat{\lambda}}{1-\exp(-\hat{\lambda})}=\bar{x}.$$

Daraus ist $\hat{\lambda}$ nicht explizit bestimmbar.

Numerische Lösungen liefert etwa das Newton-Verfahren, das iterativ vorgeht, um die Nullstelle einer Funktion $h(x)$ zu berechnen. Das Newton-Verfahren beruht auf einer Taylor-Approximation, bei der nur die erste Ableitung berücksichtigt wird.

Ist x_0 eine Nullstelle von $h(x)$ und x ein in der Nähe von x_0 liegender Wert, so gilt approximativ:

$$0=h(x_0)\cong h(x)+h'(x)\cdot(x_0-x).$$

Umstellen ergibt

$$x_0=x-\frac{h(x)}{h'(x)}.$$

Wird nun mit einem Näherungswert $x^{(0)}$ gestartet, so ergibt sich mittels dieser Beziehung eine neue Näherung $x^{(1)}$ die (hoffentlich!) dichter bei der Nullstelle liegt. $x^{(1)}$ wird dann erneut verwendet, um eine abermalige Verbesserung zu erhalten:

$$x^{(k+1)}=x^{(k)}-\frac{h(x^{(k)})}{h'(x^{(k)})}\qquad k=0,1,2,\ldots \tag{6.12}$$

Das Verfahren stoppt, wenn die Abweichung von $h(x^{(k)})$ von Null bzw. die Änderung nur noch unterhalb einer vorgegebenen Toleranz liegt.

Beispiel 6.1.17 *gestutzte Poisson-Verteilung - Fortsetzung von Seite 173*

Wir setzen das letzte Beispiel fort. Es ist die Nullstelle der Funktion $h(\lambda)=\bar{x}\cdot(1-\exp[-\hat{\lambda}])-\hat{\lambda}$ zu berechnen.

Ausgehend von einem Startwert $\hat{\lambda}^{(0)}$ erhalten wir sukzessiv bessere Näherungen entsprechend der obigen Iterationformel. Wird hier $\hat{\lambda}^{(0)}=\bar{x}$ gewählt, so gilt für $\hat{\lambda}^{(1)}$:

$$\hat{\lambda}^{(1)}=\bar{x}+\frac{\exp(-\bar{x})}{\bar{x}\cdot(\exp(-\bar{x})-1))}.$$

Wir betrachten nun die numerische Bestimmung der Parameter in verallgemeinerten linearen Modellen mit natürlicher Linkfunktion, vgl. Abschnitt 3.2.10.

Sei $Y_1,\ldots,Y_n$ eine Zufallsstichprobe aus einer Verteilung mit der Dichte

$$f(y;\boldsymbol{z}_i,\boldsymbol{\beta})=\exp[\eta_i y+d(\eta_i)+S(y)]\times I_A(y),$$

wobei $\eta_i=\beta_0+\beta_1 z_{1i}+\ldots+\beta_p z_{pi}=\boldsymbol{z}_i\boldsymbol{\beta}^T$ mit bekannten Vektoren $\boldsymbol{z}_i=(z_{i1},\ldots,z_{ip})$ $(i=1,\ldots,n)$ von unabhängigen nichtstochastischen Variablen sind und $\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_p)$ ein Vektor unbekannter Parameter ist.

Die Loglikelihoodfunktion lautet

$$\begin{aligned}\mathscr{L}(\beta_0,\beta_1,\ldots,\beta_p) &= \sum_{i=1}^n \eta_i y_i+\sum_{i=1}^n d(\eta_i)+\sum_{i=1}^n S(y_i)\\ &= \sum_{i=1}^n y_i\boldsymbol{z}_i\boldsymbol{\beta}^T+\sum_{i=1}^n d(\boldsymbol{z}_i\boldsymbol{\beta}^T)+\sum_{i=1}^n S(y_i).\end{aligned}$$

Die partiellen Ableitungen sind, wenn wir $z_{0i}=0$ für $i=1,\ldots,n$ setzen:

$$\frac{\partial\mathscr{L}}{\partial\beta_j}=\sum_{i=1}^n y_i z_{ij}+\sum_{i=1}^n d'(\boldsymbol{z}_i\boldsymbol{\beta}^T)z_{ij}\qquad j=0,1,\ldots,p.$$

Als Likelihoodgleichungen erhalten wir

$$\sum_{i=1}^n y_i z_{ij}+\sum_{i=1}^n d'(\boldsymbol{z}_i\boldsymbol{\beta}^T)z_{ij}=0\qquad j=0,1,\ldots,p.$$

Wegen $\mathrm{E}(Y_i)=-d'(\boldsymbol{z}_i\boldsymbol{\beta}^T)$, vgl. Abschnitt 3.1.5, können wir das unter Verwendung der üblichen Symbolik $\mathrm{E}(X_i)=\mu_i$ auch in der Form

$$\sum_{i=1}^n z_{ij}(y_i-\mu_i)=0\qquad j=0,1,\ldots,p \tag{6.13}$$

angeben.

In der Regel hängt μ_i nichtlinear von den interessierenden Parametern β_j ab. (Die große Ausnahme bildet das lineare Regressionsmodell mit normalverteilten Störungen.) Um das Gleichungssystem zu lösen, geht man wieder iterativ vor:

1. Zuerst werden Startwerte $\beta_j^{(0)}$ für die β_j bestimmt.

2. Dann ersetzt man $\mu_i = -d'(\boldsymbol{z}_i\boldsymbol{\beta}^{(0)T})$ durch eine Taylor-Approximation erster Ordnung um $\beta^{(0)}$.

3. Die Lösung des linearen Gleichungssystems (6.13) ergibt eine Korrektur, sagen wir $d\boldsymbol{\beta}$, der letzten Schätzung. Damit wird die Verbesserung $\boldsymbol{\beta}^{(k+1)} := \boldsymbol{\beta}^{(k)} + d\boldsymbol{\beta}$ bestimmt.

4. Anschließend wird $k := k+1$ gesetzt und nach 2. zurückgegangen, sofern ein geeignetes Stoppkriterium noch nicht erfüllt ist.

Die resultierende Schätzung ist nicht notwendig eine Lösung des Ausgangsproblems. Auch kann es zu Problemen bei der numerischen Bestimmung kommen, sofern die Startwerte nicht hinreichend gut waren.

Beispiel 6.1.18 *Verluste im Kühlwassersystem - Fortsetzung von Seite 107*

Sei $Y_1,\ldots,Y_n$ eine Zufallsstichprobe aus einer Poisson-Verteilung, wobei die Erwartungswerte μ_i der Y_i gemäß

$$\mu_i = \exp[\beta_0 + \beta_1 z_i]$$

von einer unabhängigen deterministischen Variablen abhängen. Dann sind die Likelihoodgleichungen gegeben durch:

$$\begin{aligned}\frac{\partial}{\partial\beta_0}\mathcal{L}(\beta_0,\beta_1) &= \sum_{i=1}^{n}(y_i - \exp[\beta_0+\beta_1 z_i]) = 0,\\ \frac{\partial}{\partial\beta_1}\mathcal{L}(\beta_0,\beta_1) &= \sum_{i=1}^{n} z_i(y_i - \exp[\beta_0+\beta_1 z_i]) = 0.\end{aligned}$$

Um dieses Gleichungssystem zu lösen, approximieren wir $\exp[\beta_0+\beta_1 z_i]$ linear:

$$\begin{aligned}\exp[\beta_0+\beta_1 z_i] \cong \exp\left[\beta_0^{(0)}+\beta_1^{(0)} z_i\right] &+ \exp\left[\beta_0^{(0)}+\beta_1^{(0)} z_i\right]\left(\beta_0-\beta_0^{(0)}\right)\\ &+ z_i\exp\left[\beta_0^{(0)}+\beta_1^{(0)} z_i\right]\left(\beta_1-\beta_1^{(0)}\right).\end{aligned}$$

Diese Approximation wird oben eingesetzt, die Lösung des resultierenden linearen Gleichungssystems ergibt die Korrekturen $d\beta_i$.

Startwerte erhalten wir etwa mit der KQ-Methode über den Ansatz

$$\ln(y_i) \cong \beta_0 + \beta_1 z_i + U_i \qquad i = 1,\ldots,n.$$

6.1.5 M-Schätzer

In vielen Anwendungen wird unter Hinweis auf den zentralen Grenzwertsatz für die betrachtete Zufallsvariable eine Normalverteilung unterstellt. Bei Datensätzen zeigen diagnostische

Grafiken, wie etwa QQ-Diagramme[1], in der Tat im mittleren Bereich oft eine recht gute Übereinstimmung mit der Normalverteilung. An den Rändern besteht dagegen häufig eine größere Diskrepanz. Besonders deutlich wird sie, wenn sich ein größerer Anteil extremerer Werte in der Stichprobe befindet, als bei Gültigkeit der Normalverteilung zu erwarten wäre.

Bei diesen Gegebenheiten ist eine ML-Schätzung auf Basis der Normalverteilungsannahme nicht mehr adäquat. Besser geeignet sind dann Schätzer, die einerseits analog zu den ML-Schätzern konstruiert sind, andererseits aber dem skizzierten Erscheinungsbild Rechnung tragen. Um diese *Schätzer vom Maximum-Likelihood-Typ*, kurz *M-Schätzer*, zu motivieren, betrachten wir zwei Beispiele.

Beispiel 6.1.19 *ML-Schätzer für den Lageparameter der Huber-Familie*

Die Verteilung der Zufallsvariablen X gehöre zu einer Lage-Familie $\mathscr{F} = \{f(x)|f(x-\vartheta) = f_0(x),\ \vartheta \in \mathbb{R}\}$, wobei die Dichte $f_0(x)$ gegeben ist durch

$$f_0(x) = \begin{cases} \dfrac{1-\epsilon}{\sqrt{2\pi}} \cdot e^{-x^2/2} & |x| \le k \\ \dfrac{1-\epsilon}{\sqrt{2\pi}} \cdot e^{k^2/2} \cdot e^{-k|x|} & |x| > k\,. \end{cases} \tag{6.14}$$

Diese Verteilungen wurden von Huber vorgeschlagen; wir bezeichnen sie als *Huber-Familie*. Die Verteilung ist im zentralen Bereich eine Normalverteilung und an den Rändern entspricht sie der Laplace-Verteilung. Die Parameter ϵ und k sind verknüpft gemäß:

$$\frac{\epsilon}{1-\epsilon} = \frac{2\phi(k)}{k} - 2\Phi(-k).$$

Der ML-Schätzer für ϑ ergibt sich wie üblich als Lösung von

$$\sum_{i=1}^{n} \ln f(x_i - \vartheta) \stackrel{!}{=} \max$$

oder äquivalent als Lösung von

$$-\sum_{i=1}^{n} \ln f(x_i - \vartheta) \stackrel{!}{=} \min.$$

Diese Umformung wird nahegelegt durch die Interpretation von $-\ln f(x-\vartheta)$ als Abstand. Bis auf eine Konstante gilt nämlich:

$$-\ln f(x-\vartheta) = \begin{cases} (x-\vartheta)^2 & |x-\vartheta| \le k \\ k|x-\vartheta| & |x-\vartheta| > k \end{cases}.$$

Das Minimierungsproblem führt somit auf die Lösung der Gleichung

$$\sum_{i=1}^{n} \frac{\partial f(x_i-\vartheta)/\partial\vartheta}{f(x_i-\vartheta)} = 0.$$

[1] Bei QQ-Diagrammen werden die geordneten Werte $x_{i:n}$ in Abhängigkeit von den theoretischen $F^{-1}((i-0.5)/n)$ dargestellt. Vgl. EinfStat.

Die Summanden $\psi(x_i-\vartheta)=(\partial f(x_i-\vartheta)/\partial\vartheta)/f(x_i-\vartheta)$ der letzten Gleichung zeigen, mit welchem Gewicht die einzelnen Beobachtungen bei der Bestimmung von $\hat{\vartheta}$ eingehen. Es ist

$$\psi(x_i-\vartheta)=\begin{cases} x_i-\vartheta & |x_i-\vartheta|\le k \\ k\cdot\operatorname{sign}(x_i-\vartheta) & |x_i-\vartheta|>k \end{cases}. \tag{6.15}$$

Damit reflektiert die ML-Schätzung genau die Konstruktion der Dichte. Extreme Werte werden wie bei der Laplace-Verteilung nur entsprechend ihrer rechts-links-Lage vom Zentrum berücksichtigt. Werte im mittleren Teil des Datensatzes kommen wie bei der Normalverteilung mit dem vollen Gewicht zur Geltung.

Beispiel 6.1.20 *ML-Schätzer für den Lageparameter der Cauchy-Verteilung*

X habe eine Cauchy-Verteilung mit dem Skalenparameter $\tau=1$,

$$f(x;\vartheta)=\frac{1}{\pi}\cdot\frac{1}{1+(x-\vartheta)^2}.$$

Der ML-Ansatz führt hier zu

$$\sum_{i=1}^{n}\frac{\partial f(x_i-\vartheta)/\partial\vartheta}{f(x_i-\vartheta)}=\sum_{i=1}^{n}\psi(x_i-\vartheta)=0,$$

wobei

$$\psi(x-\vartheta)=\frac{2(x-\vartheta)}{1+(x-\vartheta)^2}.$$

Die Cauchy-Verteilung produziert mehr sehr extreme Werte als die Normalverteilung. Das führt dazu, dass sehr weit vom Zentrum weg liegende Beobachtungen ein mit wachsender Entfernung abnehmendes Gewicht bzgl. der Bestimmung des Schätzwertes von ϑ erhalten. Im Zentrum jedoch ist $\psi(u)$ fast linear, also fast gleichwertig mit der Normalverteilung.

Die beiden Beispiele geben Anlass, eine Klasse von Lage-Schätzern dadurch zu definieren, dass ein geeigneter Abstand $\rho(x-\vartheta)$ vorgegeben wird und die Aufgabe $\sum\rho(x_i-\vartheta)\stackrel{!}{=}\min$ durch Lösung der entsprechenden Gleichung $\sum_{i=1}^{n}\psi(x_i-\vartheta)=0$, wobei $\psi(u)=\rho'(u)$ gilt, erledigt wird. Einen Schritt weiter gehend kann man statt der ρ-Funktion auch gleich die ψ-Funktion vorgeben und die Lageparameter durch Lösung dieser Gleichung ermitteln.

Definition 6.1.21 *M-Schätzer*

Sei $X_1,\ldots,X_n$ eine Zufallsstichprobe aus einer Verteilung mit dem Lageparameter ϑ. Ein *M-Schätzer* für ϑ ist dann jede Lösung von

$$\sum_{i=1}^{n}\psi(x_i-\vartheta)=0.$$

Dabei ist die ψ-Funktion geeignet gewählt.

M-Schätzer enthalten ML-Schätzer als Spezialfälle. Sie unterscheiden sich darin, dass die ψ-Funktion nicht mehr durch das Modell festgelegt wird, sondern aufgrund gewisser Eigenschaften, die die resultierenden Schätzer haben sollen. Der Hintergrund für diese Betrachtungsweise ist die Einschätzung, dass das zugrunde liegende Modell i.d.R. nicht bekannt und die Unterstellung einer speziellen Verteilung wie der Normalverteilung nicht adäquat ist.

Die Einteilung der Schätzer erfolgt in der Regel über ihre ψ-Funktionen, die auch für die Bestimmung der Schätzwerte eine zentrale Bedeutung haben. Bei Hubers ψ-Funktion werden 'zu große' Abstände etwas heruntergewichtet; bei der aus der Cauchy-Verteilung abgeleiteten geschieht das noch stärker. Letztere ist ein Beispiel für eine sogenannte 'redescending' ψ-Funktion, bei denen $\psi(u) \to 0$ für $|u| \to \infty$. Diese wurden von verschiedenen Autoren vorgeschlagen, um sich stärker gegen extreme Werte zu schützen. Dazu gehört z.B. auch Tukeys Biweight

$$\psi(u) = \begin{cases} u\left(6^2 - u^2\right)^2 & |u| \le 6 \\ 0 & |u| > 6 \end{cases},$$

sowie die ψ-Funktion von Hampel:

$$\psi(u) = \begin{cases} u & |\mathrm{u}| < 1.5 \\ 1.5 \cdot \mathrm{sign}(u) & 1.5 \le |\mathrm{u}| < 3.6 \\ 0.341 \cdot (8 - |u|) \cdot \mathrm{sign}(u) & 3.6 \le |\mathrm{u}| < 8.0 \\ 0 & 8.0 \le |\mathrm{u}| \end{cases}.$$

Die ψ-Funktionen hängen in den meisten Fällen von geeigneten Tuning-Konstanten ab. Diese kommen ins Spiel, weil bei den ϱ-Funktionen, die ja als Abstände interpretierbar sind, eine Skalierung nicht automatisch passiert. M-Schätzer sind in der Regel nicht Lage- und Skalen-äquivariant, d.h. für sie gilt i.d.R. nicht $T(a + bx_1, \ldots, a + bx_n) = a + bT(x_1, \ldots, x_n)$. Das Abstandsquadrat und der Betrag bilden eine Ausnahme.

Daher ist die Berücksichtigung der Streuung in den meisten Fällen wesentlich. Einmal kann das geschehen, indem die Daten zunächst standardisiert werden. Dazu wird i.d.R. der *MAD*, der Median der absoluten Abstände vom Median verwendet:

$$\tilde{\tau} = 1.483 \cdot \mathrm{med}\left\{|x_j - \mathrm{med}\{x_i : 1 \le i \le n\}| : 1 \le j \le n\right\}. \tag{6.16}$$

Der Faktor soll diesen Schätzer im Fall der Normalverteilung zu einem geeigneten Schätzer für die Standardabweichung σ machen.

Zum anderen gibt es auch M-Schätzer für Skalen-Parameter. Für die Motivation der Definition betrachten wir die Likelihoodgleichungen einer Lage-Skalen-Familie $\mathscr{F} = \{f((x - \vartheta)/\tau)/\tau \,|\, \vartheta \in \mathbb{R}, \tau > 0\}$. Mit $\psi(u) = f'(u)/f(u)$ können diese in der Form

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \vartheta}{\hat{\tau}}\right) = 0,$$

$$\sum_{i=1}^{n} \left[\frac{x_i - \vartheta}{\hat{\tau}} \cdot \psi\left(\frac{x_i - \vartheta}{\hat{\tau}}\right) - 1\right] = 0$$

geschrieben werden. Für die simultane M-Schätzung des Lage- und des Skalenparameters ist also ein Gleichungssystem der Form

$$\sum_{i=1}^{n} \psi\left(\frac{x_i-\vartheta}{c\hat{\tau}}\right)=0 \quad \text{und} \quad \sum_{i=1}^{n} \chi\left(\frac{x_i-\vartheta}{c\hat{\tau}}\right)=0$$

zu lösen. Die Tuningkonstante c wird dabei i.d.R. so gewählt, dass unter der Normalverteilung der Schätzer $\hat{\tau}$ günstig für die Schätzung der Standardabweichung ist. Die Funktion $\chi(u)$ ist gerade, $\chi(-u)=\chi(u)$. Oft wird $\chi(u)=\psi(u)^2-1$ empfohlen.

Simulationsstudien haben allerdings gezeigt, dass die Vorabschätzung mittels MAD oft nicht schlechter oder sogar besser ist als die simultane M-Schätzung von Lage- und Skalenparameter.

In noch größerem Ausmaß als bei den ML-Schätzern müssen iterative Verfahren zur numerischen Bestimmung eingesetzt werden. Für die numerische Bestimmung der M-Schätzer der Lage können iterativ neu gewichtete Mittelwerte herangezogen werden.

Sei dazu $w(u)$ definiert durch $w(u) := u \cdot \psi(u)$. Setzen wir dies in die Bestimmungsgleichung für den Lageparameter ein, so erhalten wir:

$$\sum_{i=1}^{n} \frac{x_i-\vartheta}{c\hat{\tau}} \cdot w\left(\frac{x_i-\vartheta}{c\hat{\tau}}\right)=0.$$

Umsortieren ergibt:

$$\hat{\vartheta}=\frac{\sum_{i=1}^{n} x_i\, w((x_i-\vartheta)/c\hat{\tau})}{\sum_{i=1}^{n} w((x_i-\vartheta)/c\hat{\tau})}. \tag{6.17}$$

Haben wir einen Startschätzer für $\hat{\vartheta}$, etwa den Median der Stichprobe, so können wir mit diesem die Residuen $u_i = x_i - \hat{\vartheta}$ bestimmen, diese auf der rechten Seite von (6.17) einsetzen und erhalten so eine neue Schätzung für ϑ. Das Verfahren wird iteriert, bis sich keine wesentlichen Änderungen mehr ergeben.

Die so bestimmten M-Schätzer werden bisweilen auch als *W-Schätzer* bezeichnet.

6.1.6 L-Schätzer

Wie die M-Schätzer ist die Klasse der L-Schätzer im Rahmen der Betrachtung von Methoden, die unempfindlich gegen einzelne extreme Beobachtungen sind, ins Blickfeld des Interesses gerückt. Einen ihrer Ausgangspunkte haben sie u.a. in der von vielen Anwendern praktizierten Vorgehensweise, zu extreme Werte aus dem Datensatz zu eliminieren. Das führt sofort zu dem *getrimmten Mittel*, bei dem jeweils $\alpha \cdot 100\%$ der Beobachtungen an den Rändern weggelassen werden. Wir definieren L-Schätzer zunächst allgemein.

Definition 6.1.22 *L-Schätzer*

Ein *L-Schätzer* für den Parameter ϑ ist eine Statistik L, die sich als Linearkombination der geordneten Statistiken $X_{r:n}$ darstellen lässt,

$$L = L(X_1,\ldots,X_n) = \sum_{i=1}^{n} c_{n,i} X_{i:n}, \tag{6.18}$$

und deren Wert als Schätzwert für ϑ verwendet werden soll.

Beispiel 6.1.23

(i) Das α-getrimmte Mittel ist für $0 < \alpha \leq 0.5$ gegeben durch

$$T = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{i:n}.$$

(ii) Das α-winsorisierte Mittel mit $0 < \alpha < 0.5$ lautet:

$$\bar{X}_{W,\alpha} = \frac{1}{n} \left([n\alpha] X_{[n\alpha]:n} + \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{i:n} + [n\alpha] X_{n-[n\alpha]:n} \right).$$

An den Rändern werden also jeweils etwa $(1-\alpha)100\%$ der Beobachtungen durch den zugehörigen innersten Wert ersetzt.

(iii) Der Quartilsabstand ist – bis auf einen konstanten Faktor – ein L-Schätzer für den Skalenparameter:

$$T = X_{[3n/4]:n} - X_{[n/4]:n}.$$

Eine die meisten verwendeten Schätzer umfassende Teilklasse der L-Schätzer ist gegeben durch

$$L = \frac{1}{n} \sum_{i=1}^{n} J\left(\frac{i}{n+1}\right) X_{i:n} + \sum_{j=1}^{m} a_j X_{[np_j]:n}. \tag{6.19}$$

Hier repräsentiert $J(u)$, $0 \leq u \leq 1$, eine Gewichte generierende Funktion. Weiter wird angenommen, dass $0 < p_1 < p_2 < \ldots < p_m < 1$ und dass $a_1, \ldots, a_m$ von Null verschiedene Konstanten sind.

Die Ausgangsform der L-Schätzer erhalten wir also, indem die $c_{n,i}$ gleich den $J(i/(n+1))/n$ gesetzt werden plus einer additiven Konstanten a_j, falls $i = [np_j]$ für ein geeignetes $j \in \{1, 2, \ldots, m\}$. Somit sind die Schätzer der Teilklasse charakterisiert als Summe zweier spezieller Formen von L-Schätzern; $J(u)$ ist typischerweise eine recht glatte Funktion, während der andere Teil die gewichtete Summe einer festen Anzahl von Quantilen ist. Oft interessiert nur einer der beiden Teile, so dass dann $J(u) \equiv 0$ bzw. $a_j \equiv 0$ gesetzt wird.

Beispiel 6.1.24 *Einige L-Schätzer*

(i) Das α-getrimmte Mittel T ist asymptotisch von der angegebenen Form mit $m = 0$ und

$$J(u) = \begin{cases} 1/(1-2\alpha) & \text{für } \alpha < u < 1-\alpha \\ 0 & \text{sonst} \end{cases}.$$

(ii) Das α-winsorisierte Mittel ist asymptotisch äquivalent zu einem Schätzer der angegebenen Form mit $m = 2$, $p_1 = \alpha$, $p_2 = 1-\alpha$ und $a_1 = a_2 = \alpha$ und

$$J(u)=\begin{cases}1 & \text{für } \alpha<u<1-\alpha\\ 0 & \text{sonst}\end{cases},$$

(iii) Der Quartilsabstand ist im Wesentlichen von der angegegbenen Gestalt. Dazu sind $J(u)=0$, $m=2$, $p_1=0.25$, $p_2=0.75$, $a_1=-1$ und $a_2=1$ zu setzen.

6.1.7 Dichteschätzung

Um sich einen Eindruck über die Gestalt der zugrunde liegenden Verteilung zu verschaffen, wird man in aller Regel erst einmal ein Histogramm zeichnen. Dies ist die verbreitetste nichtparametrische Dichteschätzung. Dabei geht man von einer vorgegebenen Klasseneinteilung $x_0^*<x_1^*<\ldots<x_m^*$ aus. Dann nimmt es für eine Stichprobe $x_1,\ldots,x_n$ die folgende Form an:

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n} I_{(x_{k-1}^*,x_k^*]}(x_i)\frac{1}{x_k^*-x_{k-1}^*} \quad \text{für } x\in(x_{k-1}^*,x_k^*] \quad k=1,\ldots,m.$$

Dabei ist $I_{(x_{k-1}^*,x_k^*]}(x)$ die Indikatorfunktion für die i'te Klasse, $I_A(u)=1$ für $u\in A$ und $=0$ sonst.

Das Histogramm ist unstetig und hängt sehr stark von der Wahl der Klassengrenzen ab. Als naheliegende Verbesserung bietet sich daher an, die Klassen über den Bereich der Beobachtungen 'gleiten' zu lassen. Das führt bei einer festen Klassenbreite h zu

$$\hat{f}(x)=\frac{1}{h\cdot n}\sum_{i=1}^{n} I_{(x-h/2,x+h/2]}(x_i)=\frac{1}{h\cdot n}\sum_{i=1}^{n} I_{(-1/2,1/2]}\left(\frac{x-x_i}{h}\right).$$

$\hat{f}(x)$ hat eine enge Verwandtschaft mit einer Approximation der Ableitung der empirischen Verteilungsfunktion, $\hat{f}(x)=(\hat{F}(x-h/2)-\hat{F}(x+h/2))/h$. Das macht erklärlich, dass das Resultat dieser Schätzung i.d.R. sehr unruhig ist. Eine Verbesserung erhalten wir, indem nicht alle für die Schätzung von f an der Stelle x einbezogenen Beobachtungen das gleiche, sondern ein mit der Entfernung von x abnehmendes Gewicht bekommen. Ist $K(u)$ eine Funktion mit den Eigenschaften einer Dichtefunktion, so bekommt dann die Dichteschätzung die Gestalt

$$\hat{f}(x)=\frac{1}{h\cdot n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right). \tag{6.20}$$

$K(u)$ wird als Kern bezeichnet. Der Kern charakterisiert die *Kerndichteschätzung* $\hat{f}(x)$. Diese Schätzer zeichnen sich dadurch aus, dass sich Stetigkeits- und Differenzierbarkeitseigenschaften der Kerne auf die Schätzer übertragen. Die Wahl der *Bandbreite* h ist wie beim (gleitenden) Histogramm der kritische Punkt.

Beispiel 6.1.25 *Illustration zur Kerndichteschätzung*

Für eine univariate Stichprobe vom Umfang $N=10$ ist in der folgenden Abbildung die Kerndichteschätzung unter Verwendung des Kernes der Normalverteilungsdichte mit $h=1$ dargestellt. Zudem zeigt die Abbildung die Beiträge der einzelnen Kerne, die bei den Datenpunkten zentriert sind.

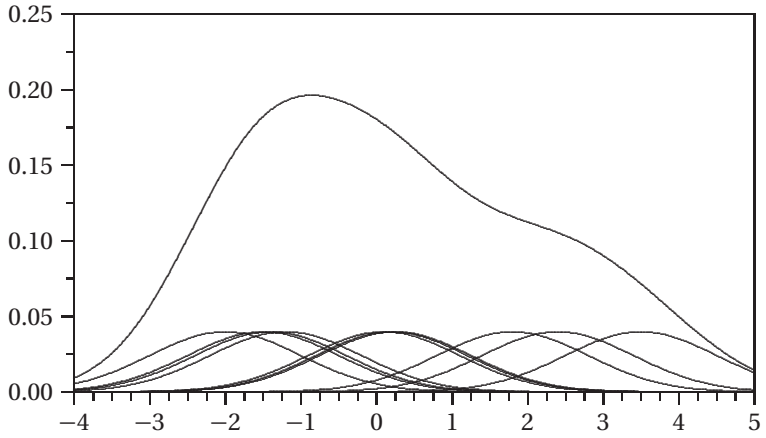


Abb. 6.3: *Kerndichteschätzung*

Die Verallgemeinerung der Kerndichteschätzung auf den Fall multivariater Daten ist leicht. Dazu nimmt man eine auf $\mathbb{R}^p$ definierte Kernfunktion und setzt in Analogie zu (6.20) für $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ bei gegebener Stichprobe $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$:

$$\hat{f}(\boldsymbol{x}) = \frac{1}{h^p n} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right). \tag{6.21}$$

Verbreitet ist die Verwendung von *Produktkernen*, bei denen sich die p-dimensionale Funktion $K_h(\boldsymbol{u})$ als Produkt von p univariaten Kernfunktionen ergibt:

$$\hat{f}(\boldsymbol{x}) = \hat{f}(x_1, \ldots, x_p) = \frac{1}{h_1 \cdots h_p n} \sum_{i=1}^{n} \left[\prod_{j=1}^{p} K_j\left(\frac{x_i - x_{ij}}{h_j}\right) \right]. \tag{6.22}$$

Eine umfassende Darstellung der multivariaten Dichteschätzung gibt Scott (1992).

6.2 Eigenschaften von Schätzfunktionen

Wie wir gesehen haben, können verschiedene Methoden zur Gewinnung von Schätzern zu unterschiedlichen Schätzfunktionen führen. Wir werden nun solche Schätzfunktionen vorziehen, von denen wir erwarten können, dass sie den unbekannten Wert $g(\vartheta)$ ‚relativ gut' treffen. Diese Vorstellung lässt sich auf verschiedene Weise präzisieren.

6.2.1 Konsistenz

Eine erste Forderung an gute Schätzer wäre, dass die Verteilung der Schätzfunktionen relativ eng um den Parameterwert konzentriert ist, etwa:

$$P_\vartheta(|T(\boldsymbol{X}) - g(\vartheta)| \leq \varepsilon) \approx 1$$

für nicht zu kleine $\varepsilon > 0$.

Diese Forderung lässt sich für feste Stichprobenumfänge i.d.R. nicht erfüllen. Mit wachsendem Stichprobenumfang gelingt es aber bei vernünftigen Schätzfunktionen immer besser. Wenn wir die Diskussion in Kapitel 5 betrachten, führt dieser Ansatz auf die Konvergenz in Wahrscheinlichkeit. Die fast sichere Konvergenz ist ein stärkeres Konzept, das die stochastische Konvergenz nach sich zieht. Damit haben wir entsprechend einen zweiten Ansatz für die Fassung dieser Eigenschaft von Schätzfunktionen.

Eine weitere Form als die oben angegebene, die Konzentration einer Verteilung um den Parameterwert zu messen, ist der *mittlere quadratische Fehler*, die erwartete quadratische Abweichung des Schätzers vom Parameterwert:

$$\mathrm{MQF}(T,\vartheta) = \mathrm{E}_\vartheta\left[(T-\vartheta)^2\right]. \tag{6.23}$$

Der MQF ist das verbreitetste Maß zur Erfassung der Güte von Schätzern. Er gibt offensichtlich auch bei festen Stichprobenumfängen eine sinnvolle Charakterisierung der Qualität eines Schätzers. Bevor aber darauf weiter eingegangen wird, sollen die angesprochenen Konzepte für große Stichproben betrachtet werden.

Definition 6.2.1 *Konsistenz*

$X_1, X_2, \ldots, X_n, \ldots$ seien Stichprobenvariablen mit der Verteilungsfunktion $F_X(x,\vartheta)$, wobei $\vartheta \in \Theta$ ein unbekannter Parameter ist. Eine Folge $T_1(X_1)$, $T_2(X_1,X_2), \ldots, T_n(X_1,\ldots,X_n), \ldots$, kurz $T_n(\boldsymbol{X})$, von Schätzfunktionen für $g(\vartheta)$ heißt

- *schwach konsistent* für $g(\vartheta)$ (relativ zu $\{F(x;\vartheta)|\vartheta \in \Theta\}$), wenn für jedes zu $F(x;\vartheta)$ gehörige $P_\vartheta, \vartheta \in \Theta$, gilt

$$\lim_{n\to\infty} P_\vartheta(|T_n(\boldsymbol{X}) - g(\vartheta)| \le \varepsilon) = 1 \qquad \text{für alle } \varepsilon > 0; \tag{6.24}$$

- *konsistent im (quadratischen) Mittel* oder einfach *konsistent* für $g(\vartheta)$ (relativ zu $\{F(x;\vartheta)|\vartheta \in \Theta\}$), wenn für jedes $F(x;\vartheta), \vartheta \in \Theta$, der mittlere quadratische Fehler gegen null geht:

$$\lim_{n\to\infty} \mathrm{E}_\vartheta\left((T_n(\boldsymbol{X}) - g(\vartheta))^2\right) = 0; \tag{6.25}$$

- *stark konsistent* für $g(\vartheta)$ (relativ zu $\{F(x;\vartheta)|\vartheta \in \Theta\}$), wenn für jedes zu $F(x,\vartheta)$ gehörige $P_\vartheta, \vartheta \in \Theta$, gilt

$$P_\vartheta\left(\lim_{n\to\infty} |T_n(\boldsymbol{X}) - g(\vartheta)| = 0\right) = 1. \tag{6.26}$$

Beispiel 6.2.2 *Konsistenz des arithmetischen Mittels und der empirischen Varianz*

Die Existenz einer endlichen Varianz einer Zufallsvariable X, $\mathrm{Var}(X) = \sigma^2$, sichert wegen $\mathrm{Var}(\bar{X}) = \sigma^2/n$ die Konsistenz des arithmetischen Mittels einer Zufallsstichprobe. Hat X ein endliches viertes Moment, so ist nach dem starken Gesetz der großen Zahlen, Satz 5.1.6, $\bar{X}_n = \sum_{i=1}^n X_i/n$ stark konsistent für $\mu = \mathrm{E}(X)$:

$$\bar{X}_n \overset{f.s.}{\to} \mu.$$

Setzen wir zudem $\mathrm{E}((X-\mu)^8) < \infty$ voraus, so erhalten wir auch die starke Konsistenz der empirischen Varianz $S_n^2 = \sum (X_i - \bar{X})^2/n$, $S_n^2 \overset{f.s.}{\to} \sigma^2$, mit der gleichen Argumentation wie beim Beweis von Satz 5.1.6.

Satz 6.2.3 *Funktion einer konsistenten Schätzfunktion*

Sei T_n eine schwach konsistente Folge von Schätzfunktionen für den Parameter $\vartheta \in \Theta$. Ist dann g eine auf Θ definierte stetige Funktion, so ist $g(T_n)$ konsistent für $g(\vartheta)$.

Beweis: Die Stetigkeit von g bei ϑ bedeutet:

$$\forall \varepsilon > 0 : \exists \delta_\varepsilon > 0 : \forall \vartheta' : |\vartheta - \vartheta'| < \delta_\varepsilon \quad \Rightarrow \quad |g(\vartheta) - g(\vartheta')| < \varepsilon.$$

Daher gilt für beliebiges $\varepsilon > 0$:

$$P(|T_n - \vartheta| < \delta_\varepsilon) \le P(|g(T_n) - g(\vartheta)| < \varepsilon).$$

Da wegen der schwachen Konsistenz von T_n die linke Seite mit $n \to \infty$ gegen eins geht, folgt die Behauptung.

Beispiel 6.2.4 *konsistente Schätzung des Parameters einer Gleichverteilung*

X sei eine über dem Intervall $(0, \vartheta)$ gleichverteilte Zufallsvariable. Der Endpunkt ϑ sei nicht bekannt und Gegenstand des Schätzproblems. $\{F_\vartheta | \vartheta > 0\}$ sei die Menge der möglichen Verteilungen von X,

$$F(x;\vartheta) = \begin{cases} 0 & \text{für } x \le 0, \\ \dfrac{1}{\vartheta} x & \text{für } x \in (0,\vartheta), \\ 1 & \text{für } x \ge \vartheta. \end{cases}$$

Es soll gezeigt werden, dass $T_n = \max\{X_1, \ldots, X_n\}$ eine konsistente Folge von Schätzern für ϑ ist. Dabei ist $X_1, \ldots, X_n$ eine Stichprobe aus $F(x;\vartheta)$.

Wie man leicht sieht, gilt:

$$P_\vartheta(T_n \le t) = \left(\frac{1}{\vartheta} \cdot t\right)^n \text{ für } 0 < t < \vartheta \quad \text{und} \quad f_{T_n}(t;\vartheta) = \begin{cases} \dfrac{n}{\vartheta^n} \cdot t^{n-1} \text{ für } 0 < t < \vartheta \\ 0 \qquad \text{sonst}. \end{cases}$$

Damit folgt:

$$\mathrm{E}_\vartheta (T_n - \vartheta)^2 = \int_0^\vartheta (t-\vartheta)^2 \cdot \frac{n}{\vartheta^n} \cdot t^{n-1} \mathrm{d}t = \frac{n}{\vartheta^n} \int_0^\vartheta \left(t^{n+1} - 2\vartheta t^n + \vartheta^2 t^{n-1}\right) \mathrm{d}t$$

$$= \frac{n}{\vartheta^n} \left[\frac{1}{n+2}\vartheta^{n+2} - \frac{2\vartheta}{n+1}\vartheta^{n+1} + \frac{\vartheta^2}{n}\vartheta^n\right] = \frac{2\vartheta^2}{(n+1)(n+2)}.$$

Der mittlere quadratische Fehler $\mathrm{E}_\vartheta\left[(T_n - \vartheta)^2\right]$ geht also für $n \to \infty$ gegen null.

Beispiel 6.2.5 *Konsistenz eines Stichprobenquantils*

Das α-Stichprobenquantil $X_{k:n}$, $k = k(n) = [\alpha n] + 1$, ist ein Schätzer für das theoretische Quantil q_α einer stetigen Verteilung mit der Dichte $f(x)$. Es ist unter den Voraussetzungen von Satz 5.1.14 konsistent im quadratischen Mittel. Denn dann gilt:

$$\frac{\sqrt{n}(X_{k:n} - q\alpha) \cdot f(q\alpha)}{\sqrt{\alpha(1-\alpha)}} \quad \xrightarrow{V} \quad \mathcal{N}(0,1).$$

Also haben wir

$$\lim_{n\to\infty} \mathrm{E}(\sqrt{n}(X_{k:n} - q\alpha))^2 = \frac{\alpha(1-\alpha)}{f(q\alpha)^2} \quad \text{oder} \quad \lim_{n\to\infty} \mathrm{E}((X_{k:n} - q_\alpha)^2) = 0.$$

Die Eigenschaft der Konsistenz wird häufig als Minimalanforderung an Schätzer genannt. Sie wird i.d.R. auch von den Schätzern erfüllt, die mit den angesprochenen Schätzprinzipien gewonnen werden. So sind Momentenschätzer meist konsistent. Für M-Schätzer, und damit auch für ML-Schätzer, gilt der folgende Satz.

Satz 6.2.6 *Konsistenz von M-Schätzern*

Sei $X_1, \ldots, X_n$ eine Zufallsstichprobe aus F und $\psi(x - \vartheta)$ eine schwach monoton fallende Funktion von ϑ. Sei ϑ_0 eine eindeutige Nullstelle von

$$H(\vartheta) = \int \psi(x - \vartheta) f(x)\,\mathrm{d}x = 0.$$

Weiter gelte:

i) In einer Umgebung von ϑ_0 sei $H'(\vartheta) \neq 0$.

ii) $\int \psi^2(x - \vartheta) f(x)\,\mathrm{d}x < \infty$ und stetig in einer Umgebung von ϑ_0.

Ist dann $\hat{\vartheta}_n$ eine Lösung der Schätzgleichung

$$H_n(\vartheta) = \sum_{i=1}^{n} \psi(X_i - \vartheta) = 0,$$

so folgt:

a) $\hat{\vartheta}_n$ ist konsistent für ϑ_0.

b) $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \xrightarrow{V} \mathcal{N}(0, \sigma^2(\vartheta_0)/H'(\vartheta_0))$, wobei $\sigma^2(\vartheta_0) = \mathrm{Var}(\psi(X_1 - \vartheta_0))$.

Beweis: Siehe Sen & Singer (1993, S. 224).

6.2.2 Erwartungstreue

Die üblichen Maßzahlen für Lage und Ausbreitung bzw. Konzentration von Verteilungen sind der Erwartungswert und die Varianz. Ihre Kombination ergibt gerade den mittleren quadratischen Fehler einer Schätzfunktion $T = T(\boldsymbol{X})$:

$$\mathrm{MQF}(T,\vartheta) = \mathrm{E}_\vartheta\left((T-\vartheta)^2\right) = (\mathrm{E}_\vartheta(T)-\vartheta)^2 + \mathrm{Var}_\vartheta(T).$$

Der erste Term auf der rechten Seite misst den Abstand des Zentrums der Verteilung von T zum Parameterwert ϑ. Wenn sich bei beschränktem Stichprobenumfang schon eine beliebig starke Konzentration der Verteilung des Schätzers um den zu schätzenden Parameter nicht erreichen lässt, werden wir zumindest verlangen, dass sie um den wahren Parameterwert konzentriert ist. D.h., wir fordern, dass diese Differenz, die als systematischer Fehler interpretiert wird, verschwindet.

Definition 6.2.7 *Erwartungstreue*

$\boldsymbol{X} = (X_1,\ldots,X_n)$ sei eine Zufallsstichprobe zu der Dichtefunktion $f(x;\vartheta), \vartheta \in \Theta$. Eine Schätzfunktion $T(\boldsymbol{X})$ von $g(\vartheta)$, $g(\vartheta) \in \mathbb{R}$, heißt *erwartungstreu* oder *unverfälscht*, wenn gilt

$$\mathrm{E}_\vartheta(T(\boldsymbol{X})) = g(\vartheta) \qquad \text{für alle } \vartheta \in \Theta. \tag{6.27}$$

Allgemein heißt die Differenz

$$b(T, g(\vartheta)) = \mathrm{E}_\vartheta(T(\boldsymbol{X})) - g(\vartheta) \tag{6.28}$$

der *Bias* von T. Im Fall $b(T, g(\vartheta)) \neq 0$ heißt T verfälscht.

Beispiel 6.2.8 *Schätzung des Parameters der Poisson-Verteilung*

X sei Poisson-verteilt mit dem Parameter λ:

$$P_\lambda(X = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!} \qquad x = 0,1,2,3,\ldots$$

Der Erwartungswert von X ist $\mathrm{E}_\lambda(X) = \lambda$. Daher liegt es nahe, als Schätzfunktion für λ das arithmetische Mittel zu nehmen. Ist $X_1,\ldots,X_n$ eine Stichprobe aus der Verteilung von X, so hat $\bar{X} = \sum_{i=1}^n X_i/n$ wieder den Erwartungswert λ:

$$\mathrm{E}_\lambda(\bar{X}) = \lambda.$$

Beispiel 6.2.9 *Schätzung des Parameters σ^2 der Normalverteilung*

$\boldsymbol{X} = (X_1,\ldots,X_n)$ sei eine Stichprobe aus einer $\mathcal{N}(\mu,\sigma^2)$-Verteilung. Der Parameter ist $\vartheta = (\mu,\sigma^2)$. Es interessiere eine Schätzung von $g(\vartheta) = \sigma^2$. Betrachtet werden die beiden Schätzfunktionen

$$T_1(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

und

$$T_2(X_1,\ldots,X_n)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Mit

$$\begin{aligned}
\mathrm{E}(X_i-\bar{X})^2 &= \mathrm{Var}(X_i-\bar{X})=\frac{1}{n^2}\mathrm{Var}(n\cdot X_i-X_1-\ldots-X_n)\\
&=\frac{1}{n^2}\mathrm{Var}((n-1)\cdot X_i-X_1-\ldots-X_{i-1}-X_{i+1}-\ldots-X_n)\\
&=\frac{1}{n^2}[\mathrm{Var}(X_1)+\ldots+\mathrm{Var}(X_{i-1})+\mathrm{Var}(X_{i+1})+\ldots+\mathrm{Var}(X_n)+(n-1)^2\mathrm{Var}(X_i)]\\
&=\frac{1}{n^2}[(n-1)\cdot\sigma^2+(n-1)^2\cdot\sigma^2]=\frac{1}{n^2}(n-1)\cdot n\cdot\sigma^2\\
&=\frac{n-1}{n}\sigma^2
\end{aligned}$$

erhalten wir:

$$\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=\sum_{i=1}^{n}\mathrm{E}(X_i-\bar{X})^2=(n-1)\sigma^2.$$

Also ist $T_1(X_1,\ldots,X_n)$ eine verfälschte und $T_2(X_1,\ldots,X_n)$ eine unverfälschte Schätzfunktion.

Die Eigenschaft der Erwartungstreue darf nicht überbewertet werden. Es gibt Situationen, wo keine (sinnvolle) erwartungstreue Schätzfunktion existiert.

Beispiel 6.2.10 *erwartungstreue Schätzung bei der geometrischen Verteilung*

X sei geometrisch verteilt, $P(X=x)=p(1-p)^x$ $(x=0,1,\ldots)$. Ist $T(X)$ erwartungstreu für p, so folgt:

$$p=\mathrm{E}_p(T(X))=\sum_{x=0}^{\infty}T(x)p(1-p)^x.$$

Dies ergibt für alle p, $0<p<1$:

$$\sum_{x=0}^{\infty}T(x)p(1-p)^x=p.$$

Die einzige Funktion $T(x)$, die diese Beziehung erfüllt, ist als Schätzfunktion offensichtlich nicht akzeptabel:

$$T(x)=\begin{cases}1 & x=0\\ 0 & x>0\end{cases}.$$

6.2.3 Effizienz

Wie wir im vorigen Abschnitt gesehen haben, stellt der Bias die eine Komponente des mittleren quadratischen Fehlers dar. Schön wäre es natürlich, unverzerrte Schätzfunktionen mit möglichst kleiner Varianz zu haben, die insgesamt den mittleren quadratischen Fehler minimieren. Dass dies nicht möglich ist, zeigen schon die konstanten Statistiken, die an einer Stelle den MQF null hätten, aber sicher nicht allgemein sinnvoll sind. Auch bei anderen Situationen gibt es verzerrte Schätzer, die einen kleineren MQF haben als naheliegende, unverzerrte Schätzer.

Beispiel 6.2.11 *verschiedene Schätzer für σ^2 bei Normalverteilung*

Seien $X_1,\ldots,X_n$ eine Stichprobe aus einer $\mathcal{N}(\mu,\sigma^2)$-Verteilung, wobei es um die Schätzung von σ^2 geht. Man kann zeigen, dass

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

einen größeren MQF hat als $(n-1)S^2/n$. (Der von $(n-1)S^2/(n+1)$ ist sogar noch kleiner!)

Wir beschränken uns nun auf die Betrachtung von erwartungstreuen Schätzfunktion. Für diese ist der MQF durch die Varianz gegeben. Bei zwei (oder mehr) erwartungstreuen Schätzfunktionen für einen Parameter wählen wir also die Schätzfunktion mit der kleineren Varianz aus.

Definition 6.2.12 *Effizienz und relative Effizienz*

Eine erwartungstreue Schätzfunktion T_1 heißt *effizient*, wenn für jede andere erwartungstreue Schätzfunktion T gilt:

$$\forall\vartheta\in\Theta: \quad \mathrm{Var}_\vartheta(T_1)\le\mathrm{Var}_\vartheta(T).$$

Die *relative Effizienz* eines erwartungstreuen Schätzers T_1 bzgl. eines zweiten T_2 ist der vom Parameter abhängige Quotient der Varianzen

$$\mathrm{Var}_\vartheta(T_2)/\mathrm{Var}_\vartheta(T_1)\ .$$

Eine effiziente Schätzfunktion hat also minimale Varianz unter allen erwartungstreuen Schätzern. In der englisch sprachigen Literatur nennt man einen effizienten Schätzer daher *UMVUE*, uniformly minimum variance unbiased estimator.

Beispiel 6.2.13 *erwartungstreue Schätzfunktionen bei der Gleichverteilung*

X sei gleichverteilt über dem Intervall $(0;\vartheta)$, $\vartheta>0$ sei unbekannt. $X_1,\ldots,X_n$ sei eine Zufallsstichprobe für X. Es sollen zwei erwartungstreue Schätzfunktionen für ϑ bestimmt werden.

Zunächst gilt

$$\mathrm{E}_\vartheta(X) = \int_0^\vartheta x \cdot \frac{1}{\vartheta}\,\mathrm{d}x = \frac{1}{\vartheta}\cdot\frac{1}{2}\vartheta^2 = \frac{\vartheta}{2}.$$

Damit ist $T_1 = 2\cdot\bar{X}$ eine erwartungstreue Schätzfunktion für ϑ.

Für die Varianz von T_1 erhalten wir:

$$\mathrm{Var}_\vartheta(T_1) = \mathrm{Var}_\vartheta(2\bar{X}) = \frac{4}{n}\mathrm{Var}_\vartheta(X) = \frac{4}{n}\cdot\frac{\vartheta^2}{12} = \frac{\vartheta^2}{3n}.$$

Die zweite Schätzfunktion für ϑ basiert auf dem Maximum der Stichprobe. Wegen

$$\mathrm{E}_\vartheta(\max(X_1,\ldots,X_n)) = \frac{n}{n+1}\cdot\vartheta$$

ist $T_2(X_1,\ldots,X_n) = (n+1)\cdot\max\{X_1,\ldots,X_n\}/n$ erwartungstreu.

Weiter gilt:

$$\mathrm{Var}_\vartheta(T_2) = \frac{(n+1)^2}{n^2}\mathrm{Var}_\vartheta(\max\{X_1,\ldots,X_n\}) = \frac{(n+1)^2}{n^2}\cdot\frac{n}{(n+1)^2(n+2)}\vartheta^2 = \frac{\vartheta^2}{n(n+2)}.$$

Ein Vergleich der Varianzen von T_1 und T_2 zeigt, dass für $n > 1$ die Verteilung von T_1 weniger eng um ϑ konzentriert ist als die von T_2:

$$\mathrm{Var}_\vartheta(T_1) = \frac{\vartheta^2}{3n} > \frac{\vartheta^2}{n(n+2)} = \mathrm{Var}_\vartheta(T_2).$$

Im Anschluss an die Definition stellt sich natürlicherweise die Frage nach der Existenz einer effizienten Schätzfunktion bzw. nach der Möglichkeit, die Effizienz nachzuweisen. Diese Frage wird im Wesentlichen durch den folgenden Satz beantwortet.

Satz 6.2.14 *Informationsungleichung oder Ungleichung von Rao-Cramér*

Die Zufallsvariable X habe die Dichte $f(x;\vartheta)$, $\vartheta\in\Theta$, die zu einer regulären Familie gehöre, für die weiter gilt: $0 < I(\vartheta) < \infty$. Sei $\boldsymbol{X} = (X_1,\ldots,X_n)$ eine Stichprobe aus der Verteilung von X und $T(\boldsymbol{X})$ eine erwartungstreue Schätzfunktion für $\psi(\vartheta)$, $\mathrm{E}_\vartheta(T(\boldsymbol{X})) = \psi(\vartheta)$. Es sei $\mathrm{Var}_\vartheta(T(\boldsymbol{X})) < \infty$. Dann gilt

$$\mathrm{Var}_\vartheta(T(\boldsymbol{X})) \geq \frac{[\psi'(\vartheta)]^2}{I_n(\vartheta)}. \tag{6.29}$$

Beweis: Mit Satz 6.2.6 gilt:

$$\begin{aligned}\mathrm{E}_\vartheta\left[\frac{\partial}{\partial\vartheta}\ln(f(\boldsymbol{X};\vartheta))\right] &= \mathrm{E}_\vartheta\left[\frac{\partial}{\partial\vartheta}\sum_{i=1}^n\ln(f(X_i;\vartheta))\right] = \sum_{i=1}^n\mathrm{E}_\vartheta\left[\frac{\partial}{\partial\vartheta}\ln(f(X_i;\vartheta))\right]\\ &= n\cdot\mathrm{E}_\vartheta\left[\frac{\partial}{\partial\vartheta}\ln(f(X;\vartheta))\right] = 0\,.\end{aligned} \tag{6.30}$$

Somit folgt einmal

$$I_n(\vartheta) = \mathrm{E}_\vartheta\left[\left(\frac{\partial}{\partial\vartheta}\ln f(\boldsymbol{X};\vartheta)\right)^2\right] = \mathrm{Var}_\vartheta\left[\frac{\partial}{\partial\vartheta}\ln f(\boldsymbol{X};\vartheta)\right],$$

und weiter

$$\mathrm{Cov}_\vartheta\left[T(\boldsymbol{X}), \frac{\partial}{\partial\vartheta}\ln f(\boldsymbol{X};\vartheta)\right] = \mathrm{E}_\vartheta\left[T(\boldsymbol{X})\cdot\frac{\partial}{\partial\vartheta}\ln f(\boldsymbol{X};\vartheta)\right].$$

Mit der Cauchy-Schwarz'schen Ungleichung gilt daher:

$$\left(\mathrm{E}_\vartheta\left[T(\boldsymbol{X})\cdot\frac{\partial}{\partial\vartheta}\ln f(\boldsymbol{X};\vartheta)\right]\right)^2 \le \mathrm{Var}_\vartheta(T(\boldsymbol{X}))\cdot I_n(\vartheta). \tag{6.31}$$

Wenn noch gezeigt werden kann, dass der Erwartungswert auf der linken Seite gleich $\psi'(\vartheta)$ ist, folgt die Behauptung durch Einsetzen in diese Ungleichung.

Nun ist

$$\mathrm{E}_\vartheta(T(\boldsymbol{X})) = \psi(\vartheta) = \int_{\mathbb{R}^n} T(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x};\vartheta).$$

Damit erhalten wir durch Differentiation im stetigen Fall:

$$\begin{aligned}\psi'(\vartheta) &= \frac{\partial}{\partial\vartheta}\int_{\mathbb{R}^n} T(\boldsymbol{x})f(\boldsymbol{x};\vartheta)\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^n} T(\boldsymbol{x})\left(\frac{\partial}{\partial\vartheta}f(\boldsymbol{x};\vartheta)\right)\mathrm{d}\boldsymbol{x}\\ &= \int_{\mathbb{R}^n} T(\boldsymbol{x})\left(\frac{\partial f(\boldsymbol{x};\vartheta)/\partial\vartheta}{f(\boldsymbol{x};\vartheta)}\right)f(\boldsymbol{x};\vartheta)\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^n} T(\boldsymbol{x})\left(\frac{\partial}{\partial\vartheta}\ln(f(\boldsymbol{x};\vartheta))\right)f(\boldsymbol{x};\vartheta)\,\mathrm{d}\boldsymbol{x}\\ &= \mathrm{E}_\vartheta\left[T(\boldsymbol{X})\cdot\frac{\partial}{\partial\vartheta}\ln f(\boldsymbol{X};\vartheta)\right].\end{aligned}$$

Für Statistiken $T(\boldsymbol{X})$, deren Erwartungswert gerade gleich ϑ ist, nimmt die Ungleichung eine vereinfachte Gestalt an:

$$\mathrm{Var}_\vartheta(T(\boldsymbol{X})) \ge \frac{1}{I_n(\vartheta)}. \tag{6.32}$$

Beispiel 6.2.15 *Schätzung des Parameters p bei der Binomialverteilung*

Sei X binomialverteilt, $X \sim \mathcal{B}(n,p)$. Ein Schätzer für p ist $T = X/n$. Die Varianz von T soll mit der unteren Schranke von Cramér-Rao verglichen werden. Zunächst ist

$$\mathrm{Var}(T) = \frac{1}{n^2}\mathrm{Var}(X) = \frac{p(1-p)}{n}.$$

Weiterhin gilt

$$\ln f(x;p) = \ln\left[\binom{n}{x}\cdot p^x(1-p)^{n-x}\right] = \ln\binom{n}{x} + x\cdot\ln(p) + (n-x)\cdot\ln(1-p)$$

und

$$\frac{\partial}{\partial p}\ln f(x;p) = 0 + x\cdot\frac{1}{p} + (n-x)\frac{1}{1-p}(-1) = \frac{x}{p} - \frac{n-x}{1-p}.$$

Damit erhalten wir:

$$I(p) = \mathrm{E}\left(\left(\frac{\partial}{\partial p}\ln f(X;p)\right)^2\right) = \mathrm{E}\left(\frac{X}{p} - \frac{n-X}{1-p}\right)^2 = \mathrm{E}\left(\frac{X-np}{p(1-p)}\right)^2 = \frac{n}{p(1-p)}.$$

Also ist $\mathrm{Var}(T) = I(p)^{-1}$, und T ist eine effiziente Schätzfunktion.

Beispiel 6.2.16 *effiziente Schätzung von* $1/\lambda$ *bei der Exponentialverteilung*

$X_1,\ldots,X_n$ sei eine Zufallsstichprobe aus einer Exponentialverteilung mit dem Parameter λ,

$$f(x;\lambda) = \lambda e^{-\lambda x} \qquad \text{für } x > 0.$$

Dann ist

$$I(\lambda) = \mathrm{E}\left[\left(\frac{\partial}{\partial\lambda}\ln f(x;\lambda)\right)^2\right] = \mathrm{E}\left[\left(\frac{1}{\lambda} - X\right)^2\right] = \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

$\bar{X}$ ist eine erwartungstreue Schätzfunktion für $g(\lambda) = 1/\lambda$. Damit erhalten wir:

$$\mathrm{Var}(\bar{X}) \geq \frac{[g'(\lambda)]^2}{n\cdot I(\lambda)} = \frac{(-1/\lambda^2)^2}{n/\lambda^2} = \frac{1}{n\lambda^2}.$$

Also ist $\bar{X}$ effizient.

Der folgende Satz zeigt die Bedeutung der Suffizienz für die Suche nach erwartungstreuen Schätzfunktionen mit minimaler Varianz auf. Genauer können wir uns bei der Suche nach solchen Schätzfunktionen auf Funktionen suffizienter Statistiken beschränken.

Satz 6.2.17 *Rao-Blackwell*

$\boldsymbol{X} = (X_1,\ldots,X_n)$ sei eine n-dimensionale Zufallsvariable mit der Dichtefunktion $f(\boldsymbol{x};\vartheta)$, $\vartheta \in \Theta$. $g(\vartheta)$ sei eine reellwertige Funktion von ϑ und $T(\boldsymbol{X})$ sei eine erwartungstreue Schätzfunktion für $g(\vartheta)$. Ist dann $S(\boldsymbol{X})$ eine suffiziente Statistik für ϑ, so gilt:

(i) $T_1(\boldsymbol{X}) = \mathrm{E}_\vartheta(T(\boldsymbol{X})|S(\boldsymbol{X}))$ ist eine erwartungstreue Schätzfunktion für $g(\vartheta)$.

(ii) $\mathrm{Var}_\vartheta(T(\boldsymbol{X})) \geq \mathrm{Var}_\vartheta(T_1(\boldsymbol{X}))$ für alle $\vartheta \in \Theta$, d.h. die Varianz von $T_1(\boldsymbol{X})$ ist gleichmäßig nicht größer als die von $T(\boldsymbol{X})$.

Beweis: Für Punkt (i) ist zunächst zu zeigen, dass $T_1(\boldsymbol{X})$ nur eine Funktion von $S(\boldsymbol{X})$ ist und nicht von ϑ abhängt. Dies gilt aber offensichtlich wegen der Suffizienz von S.

Die Erwartungstreue von $T_1(\boldsymbol{X})$ erhalten wir mit Satz 2.3.7:

$$\mathrm{E}_\vartheta(T_1(\boldsymbol{X})) = \mathrm{E}_\vartheta(\mathrm{E}_\vartheta(T(\boldsymbol{X})|S(\boldsymbol{X}))) = \mathrm{E}_\vartheta(T(\boldsymbol{X})) = g(\vartheta).$$

Nun zu Punkt (ii). Allgemein gilt

$$h_1(X) \geq h_2(X) \geq 0 \quad \Rightarrow \quad \mathrm{E}(h_1(X)) \geq \mathrm{E}(h_2(X)),$$

und wegen $\mathrm{Var}(h(X)) = \mathrm{E}(h(X)^2) - \mathrm{E}(h(X))^2 \geq 0$:

$$\mathrm{E}(h(X)^2) \geq \mathrm{E}(h(X))^2.$$

Diese Ungleichungen gelten auch für bedingte Erwartungswerte und folglich auch für bedingte Erwartungen. Anwendung der zweiten auf $h(\boldsymbol{X}) = T(\boldsymbol{X}) - g(\vartheta)$ ergibt:

$$\begin{aligned}\mathrm{E}_\vartheta\left[(T(\boldsymbol{X}) - g(\vartheta))^2 | S(\boldsymbol{X})\right] &\geq (\mathrm{E}_\vartheta[T(\boldsymbol{X}) - g(\vartheta)|S(\boldsymbol{X})])^2 \\ = (\mathrm{E}_\vartheta[T(\boldsymbol{X})|S(\boldsymbol{X})] - g(\vartheta))^2 &= (T_1(\boldsymbol{X}) - g(\vartheta))^2\end{aligned}$$

Wird die erste der beiden obigen Ungleichungen hierauf angewendet, so erhalten wir:

$$\begin{aligned}\mathrm{Var}_\vartheta(T(\boldsymbol{X})) = \mathrm{E}_\vartheta\left[(T(\boldsymbol{X}) - g(\vartheta))^2\right] &= \mathrm{E}_\vartheta\left[\mathrm{E}_\vartheta\left[(T(\boldsymbol{X}) - g(\vartheta))^2 | S(\boldsymbol{X})\right]\right] \\ \geq \mathrm{E}_\vartheta\left[(T_1(\boldsymbol{X}) - g(\vartheta))^2\right] &= \mathrm{Var}_\vartheta(T_1(\boldsymbol{X})).\end{aligned}$$

Der Satz legt folgende Vorgehensweise bei der Suche nach einer ‚guten' erwartungstreuen Schätzfunktion nahe: Gehe aus von einer erwartungstreuen Schätzfunktion und wähle eine suffiziente Statistik. Ist die erwartungstreue Schätzfunktion keine Funktion der suffizienten Statistik, so bestimme die bedingte Erwartung der erwartungstreuen Statistik, gegeben die suffiziente Statistik. Die so gewonnene Statistik ist ebenfalls erwartungstreu und hat eine kleinere Varianz.

Hängt die erwartungstreue Statistik schon von der suffizienten Statistik ab, so ist auf diesem Wege keine Verbesserung möglich.

Beispiel 6.2.18 *Unfälle pro Jahr*

Ein vernünftiges Modell für die Zahl der Unfälle pro Jahr in einem bestimmten Betrieb ist die Poisson-Verteilung. Beschreiben $X_1, \ldots, X_n$ die Zahl der Unfälle in n aufeinanderfolgenden Jahren, so soll der Einfachheit halber angenommen werden, dass die X_i unabhängig sind und sich der Parameter nicht ändert. Dann gilt

$$f_\lambda(\boldsymbol{x}) = e^{-\lambda} \cdot \frac{\lambda^{x_1}}{x_1!} \cdot \ldots \cdot e^{-\lambda} \cdot \frac{\lambda^{x_n}}{x_n!}.$$

Gesucht sei nun eine gute Schätzfunktion für $e^{-\lambda}$, für die Wahrscheinlichkeit also, dass in einem gegebenen Jahr kein Unfall passiert.

Nach obigem Satz wird nur eine unverfälschte Schätzfunktion für $e^{-\lambda} = g(\lambda)$ benötigt. Dann erhalten wir eine bessere, wenn wir die bedingte Erwartung bzgl. der suffizienten Statistik $S(\boldsymbol{X}) = \sum X_i$ bilden. Wir wählen die Schätzfunktion

$$T(\boldsymbol{X}) = \begin{cases} 1 & X_1 = 0 \\ 0 & \text{sonst}. \end{cases}$$

$T(\boldsymbol{X})$ ist also erwartungstreu für $g(\lambda)$:

$$\mathrm{E}_\lambda(T(\boldsymbol{X})) = 0 \cdot P_\lambda(T(\boldsymbol{X}) = 0) + 1 \cdot P_\lambda(T(\boldsymbol{X}) = 1) = 1 \cdot P_\lambda(X_1 = 0) = e^{-\lambda}.$$

Die Verbesserung $T_1(S(\boldsymbol{X}))$ von T erhalten wir über den bedingten Erwartungswert:

$$\begin{aligned}
\mathrm{E}_\lambda(T(\boldsymbol{X})|S(\boldsymbol{X}) = s) &= \mathrm{E}_\lambda(T(\boldsymbol{X})|X_1 + \ldots + X_n = s) \\
&= 0 \cdot P_\lambda(T(\boldsymbol{X}) = 0|X_1 + \ldots + X_n = s) + 1 \cdot P_\lambda(T(\boldsymbol{X}) = 1|X_1 + \ldots + X_n = s) \\
&= \frac{P_\lambda(X_1 = 0, X_2 + \ldots + X_n = s)}{P_\lambda(X_1 + \ldots + X_n = s)} = \frac{e^{-\lambda} \cdot e^{-(n-1)\lambda}((n-1)\lambda)^s/s!}{e^{-n\cdot\lambda}(n\cdot\lambda)^s/s!} \\
&= \left(\frac{n-1}{n}\right)^s = \left(1 - \frac{1}{n}\right)^s.
\end{aligned}$$

Die neue, erwartungstreue Schätzfunktion für $g(\lambda) = e^{-\lambda}$ ist also

$$T_1(\boldsymbol{X}) = \left(1 - \frac{1}{n}\right)^{X_1 + \ldots + X_n}.$$

Der im Beispiel gefundene Schätzer ist effizient. Um dies zeigen zu können, müssen wir die Vollständigkeit der Statistik zeigen. Eine Statistik S ist ja genau dann vollständig, wenn die einzige von ihr abhängige erwartungstreue Schätzfunktion der 0 die Funktion ist, die mit Wahrscheinlichkeit 1 identisch gleich 0 ist. Diese Eigenschaft lässt sich für den Nachweis der Effizienz ausnutzen. Der folgende Satz zeigt, dass eine für $g(\vartheta)$ erwartungstreue Funktion einer vollständigen und suffizienten Statistik effizient für $g(\vartheta)$ ist.

Satz 6.2.19 *Lehmann-Scheffé*

Seien $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit Dichte $f_X(x;\vartheta)$ und $S(\boldsymbol{X})$ eine vollständige und für $g(\vartheta)$ suffiziente Statistik.
Ist $T = t(S(\boldsymbol{X}))$ erwartungstreu für $g(\vartheta)$, dann ist T effizient für $g(\vartheta)$.

Beweis: Wir zeigen zunächst, dass es nur eine erwartungstreue Schätzfunktion für $g(\vartheta)$ gibt, die von $S(\boldsymbol{X})$ abhängt.

Sei $\tilde{T} = \tilde{t}(S(\boldsymbol{X}))$ eine erwartungstreue Schätzfunktion für $g(\vartheta)$, dann folgt mit der Erwartungstreue von T:

$$\mathrm{E}_\vartheta(\tilde{T}) - \mathrm{E}_\vartheta(T) = 0.$$

Für $T^* = t^*(S(\boldsymbol{X})) = \tilde{t}(S(\boldsymbol{X})) - t(S(\boldsymbol{X}))$ gilt also

$$\mathrm{E}_\vartheta(T^*) = \mathrm{E}_\vartheta(t^*(S(\boldsymbol{X}))) = 0.$$

Da die Statistik $S(\boldsymbol{X})$ vollständig ist, folgt $P_\vartheta(t^*(S(\boldsymbol{X})) = 0) = 1$, d.h.,

$$P_\vartheta(\tilde{t}(S(\boldsymbol{X})) = t(S(\boldsymbol{X}))) = 1.$$

Also gibt es nur eine erwartungstreue Schätzfunktion für $g(\vartheta)$, die von $S(\boldsymbol{X})$ abhängt.

Aus dem Satz von Rao-Blackwell folgt weiter, dass es keine erwartungstreue Schätzfunktion von $g(\vartheta)$ gibt, die nicht von S abhängt und eine kleinere Varianz als T hat. Wäre nämlich $U(\boldsymbol{X})$ eine erwartungstreue Schätzfunktion von $g(\vartheta)$, die nicht von $S(\boldsymbol{X})$ abhängt, so wäre $\mathrm{E}(U(\boldsymbol{X})|S(\boldsymbol{X}))$ eine erwartungstreue Schätzfunktion von $g(\vartheta)$ mit kleinerer Varianz als $U(\boldsymbol{X})$.

Insgesamt ist also T effizient.

Aufgrund dieses Satzes ist die in Beispiel 6.2.18 angegebene Schätzfunktion für $e^{-\lambda}$ bei Poisson-Verteilung ein effizienter Schätzer. Die Poisson-Verteilungen bilden eine Exponentialfamilie und S ist die zugehörige vollständige suffiziente Statistik.

Beispiel 6.2.20 *effiziente Schätzung von λ bei der Exponentialverteilung*

Seien $X_1,\ldots,X_n$ unabhängige identisch mit Parameter λ exponentialverteilte Zufallsvariablen. Dann ist $S=\sum_{i=1}^n X_i$ gammaverteilt mit den Parametern n und λ. Es gilt also:

$$f_S(s;\lambda)=\begin{cases}\dfrac{1}{\Gamma(n)}\cdot\lambda^n\cdot s^{n-1}\cdot e^{-\lambda\cdot s} & \text{für } s>0\\ 0 & \text{sonst}\end{cases}.$$

Nun ist $T=(n-1)/S$ eine erwartungstreue Schätzfunktion für λ:

$$\begin{aligned}\mathrm{E}(T)=\mathrm{E}\left(\frac{n-1}{S}\right)&=(n-1)\cdot\int_0^\infty\frac{1}{s}\cdot\frac{1}{\Gamma(n)}\cdot\lambda^n\cdot s^{n-1}\cdot e^{-\lambda\cdot s}\mathrm{d}s\\ &=\frac{n-1}{\Gamma(n)}\cdot\int_0^\infty\lambda^n\cdot s^{n-2}\cdot e^{-\lambda\cdot s}\mathrm{d}s=\frac{n-1}{\Gamma(n)}\cdot\lambda\cdot\int_0^\infty\lambda^{n-1}\cdot s^{n-2}\cdot e^{-\lambda\cdot s}\mathrm{d}s\\ &=\frac{n-1}{\Gamma(n)}\cdot\lambda\cdot\Gamma(n-1)=\lambda,\end{aligned}$$

da $\Gamma(n)=(n-1)\cdot\Gamma(n-1)$.

Weil T erwartungstreu für λ ist und eine Funktion der vollständigen und für λ suffizienten Statistik $S=\sum_{i=1}^n X_i$, ist T effizient für λ.

Zur Bestimmung der Varianz von T berechnen wir $\mathrm{E}(T^2)$:

$$\begin{aligned}\mathrm{E}(T^2)=\mathrm{E}\left(\left(\frac{n-1}{S}\right)^2\right)&=(n-1)^2\cdot\int_0^\infty\frac{1}{s^2}\cdot\frac{1}{\Gamma(n)}\cdot\lambda^n\cdot s^{n-1}\cdot e^{-\lambda\cdot s}\mathrm{d}s\\ &=(n-1)^2\cdot\frac{1}{\Gamma(n)}\cdot\int_0^\infty\lambda^n\cdot s^{n-3}\cdot e^{-\lambda\cdot s}\mathrm{d}s=(n-1)^2\cdot\frac{\lambda^2}{\Gamma(n)}\cdot\int_0^\infty\lambda^{n-2}\cdot s^{n-3}\cdot e^{-\lambda\cdot s}\mathrm{d}s\\ &=(n-1)^2\cdot\frac{\lambda^2}{\Gamma(n)}\cdot\Gamma(n-2)=(n-1)^2\cdot\frac{\lambda^2}{(n-1)\cdot(n-2)}=\frac{n-1}{n-2}\cdot\lambda^2.\end{aligned}$$

Also gilt

$$\mathrm{Var}(T) = \frac{n-1}{n-2} \cdot \lambda^2 - \lambda^2 = \frac{\lambda^2}{n-2}.$$

Nun gilt bei Exponentialverteilung:

$$I(\lambda) = \lambda^{-2}.$$

Damit ist die untere Schranke für die Varianz einer erwartungstreuen Schätzfunktion von $g(\lambda) = \lambda$ gleich λ^2/n. Somit wird die untere Schranke in unserem Beispiel nicht angenommen.

Um den Satz von Lehmann-Scheffé anzuwenden, muss man eine vollständige suffiziente Statistik für den Parameter ϑ finden. D.h. zunächst einmal, dass spezielle Statistiken auf diese Eigenschaften hin untersucht werden müssen. Hilfe bietet das Lemma 4.3.32. Es schränkt die Kandidaten, die betrachtet werden müssen, auf eine minimalsuffiziente Statistik ein, da eine vollständige, suffiziente Statistik minimalsuffizient ist. Folglich kann die Suche nach einer vollständigen und suffizienten Statistik in zwei Schritten erfolgen: (1) Bestimmung einer minimalsuffizienten Statistik. (2) Überprüfung der Vollständigkeit der Verteilungsfamilie dieser Statistik.

6.2.4 Eigenschaften spezieller Klassen von Schätzern

Maximum-Likelihood-Schätzer sind nicht allgemein unverfälscht. Jedoch kann oft durch eine leichte Modifikation die Unverfälschtheit erreicht werden. Weiter stimmt in einer 'regulären' Situation, in denen ein unverfälschter Schätzer existiert, dessen Varianz zudem die Rao-Cramér-Schranke erreicht, der ML-Schätzer mit diesem überein.

Satz 6.2.21 *Charakterisierung eines effizienten Schätzers*

Ist $T(\boldsymbol{X})$ ein effizienter Schätzer für $g(\vartheta)$, so ist T Lösung der Likelihoodgleichung

$$\frac{\partial}{\partial \vartheta} \ln f(\boldsymbol{x}; \vartheta) = 0.$$

Beweis: Ist $T(\boldsymbol{X})$ effizient, so ist $T(\boldsymbol{X})$ unverfälscht und es gilt nach dem Beweis von Satz 6.2.14, Formel (6.31), für $Z = \partial \ln f(\boldsymbol{X}; \vartheta)/\partial \vartheta$:

$$\mathrm{Cov}(T(\boldsymbol{X}), Z) = 1.$$

Mit $\mathrm{E}(T(\boldsymbol{X})) = \vartheta$ und $\mathrm{E}(Z) = 0$, siehe 6.2.9, folgt

$$\mathrm{P}(Z = T(\boldsymbol{X} - \vartheta)) = 1.$$

Also ist $\partial \ln f(\boldsymbol{x}; \hat{\vartheta})/\partial \vartheta = 0$ (mit Wahrscheinlichkeit eins), genau dann, wenn $T(\boldsymbol{X}) = \hat{\vartheta}$.

Es ist nicht nur so, dass sich effiziente Schätzer im Wesentlichen als ML-Schätzer ergeben. Auch umgekehrt lässt sich zeigen, dass ML-Schätzer zumindest asymptotisch effizient sind. Dafür muss allerdings die zugrunde liegende Dichte einige Regularitätsbedingungen erfüllen, die pathologische Fälle ausschließen und u.a. die Taylor-Approximationen ermöglichen. Von den verschiedenen Bedingungsgruppen greifen wir nur eine heraus und skizzieren den Beweis der Aussage auch nur. Den genauen Beweis des folgenden Satzes findet man z.B. bei Giri (1993, S. 290 ff.).

Satz 6.2.22 *Asymptotik der Lösung der Likelihoodgleichung*

$X_1,\ldots,X_n$ sei eine Stichprobe aus einer Verteilung mit der Dichte $f(x;\vartheta)$, $\vartheta\in\Theta$. Θ sei ein offenes Intervall in $\mathbb{R}$. Die Dichte möge mindestens dreimal nach ϑ differenzierbar sein und mit integrierbaren Ableitungen. Weiter seien

$$\mathrm{E}\left(\left|\frac{\partial^3}{\partial\vartheta^3}f(X;\vartheta)\right|\right)<M \quad\text{und}\quad 0<\mathrm{E}\left(\left(\frac{\partial}{\partial\vartheta}\ln f(X;\vartheta)\right)^2\right)<\infty.$$

Dann hat die Likelihoodgleichung

$$\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\ln f(X_i;\vartheta)=0$$

eine Lösung $\hat\vartheta_n=\hat\vartheta_n(X_1,\ldots,X_n)$, die für $n\to\infty$ in Wahrscheinlichkeit gegen den wahren Parameterwert ϑ_0 konvergiert. Zudem ist $\sqrt{n}(\hat\vartheta_n-\vartheta_0)$ asymptotisch normalverteilt mit Erwartungswert 0 und Varianz $1/I(\vartheta_0)$, wobei

$$I(\vartheta_0)=\mathrm{E}\left(\left(\frac{\partial}{\partial\vartheta}\ln f(X;\vartheta)\right)^2\right).$$

Beweisskizze: Nach dem schwachen Gesetz der großen Zahlen gilt

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\ln f(X_i;\vartheta)\quad\xrightarrow{\mathrm{P}}\quad \mathrm{E}_{\vartheta_0}\left(\frac{\partial}{\partial\vartheta}\ln f(X;\vartheta_0)\right)=0.$$

Daher gilt für beliebige $\varepsilon,\eta>0$ und genügend große n:

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\ln f(X_i;\vartheta_0)\right|<\varepsilon\right)>1-\eta.$$

Wir setzen nun die Monotonie von $\partial\ln f(X;\vartheta)/\partial\vartheta$ (bzgl. ϑ!) in einem Intervall um ϑ_0 voraus und wählen $\vartheta_1<\vartheta_0<\vartheta_2$ so, dass

$$\left.\frac{\partial}{\partial\vartheta}\ln f(x;\vartheta)\right|_{\vartheta=\vartheta_1}<\left.\frac{\partial}{\partial\vartheta}\ln f(x;\vartheta)\right|_{\vartheta=\vartheta_0}<\left.\frac{\partial}{\partial\vartheta}\ln f(x;\vartheta)\right|_{\vartheta=\vartheta_2}.$$

Die entsprechenden Relationen gelten für die Erwartungswerte bzgl. ϑ_0:

$$\mathrm{E}_{\vartheta_0}\left(\frac{\partial}{\partial\vartheta}\ln f(X;\vartheta_1)\right)<0<\mathrm{E}_{\vartheta_0}\left(\frac{\partial}{\partial\vartheta}\ln f(X;\vartheta_2)\right).$$

Sei δ so, dass der linke Erwartungswert $<-\delta$ und der rechte $>\delta$ ist. Dann können wir n so groß wählen, dass

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\ln f(X_i;\vartheta_1)<-\delta,\ \delta<\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\ln f(X_i;\vartheta_2)\right)>1-\eta.$$

Diese Wahrscheinlichkeit geht wegen der beliebigen Wahl von η gegen 1. Daraus folgt mit der Stetigkeit von $\partial \ln f(x;\vartheta)/\partial\vartheta$, dass es eine Lösung $\hat{\vartheta}_n$ in der Nähe von ϑ_0 gibt. Mit wachsendem n dürfen zudem ϑ_1 und ϑ_2 immer dichter bei ϑ_0 gewählt werden. Das zeigt, dass es einen ML-Schätzer gibt mit

$$\hat{\vartheta}_n \xrightarrow{\text{P}} \vartheta_0 .$$

Für die Verteilungsaussage betrachten wir die Taylor-Reihenentwicklung

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\vartheta} \ln f(X_i;\hat{\vartheta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\vartheta} \ln f(X_i;\vartheta_0) + (\hat{\vartheta}-\vartheta_0)\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\vartheta^2} \ln f(X_i;\vartheta_0) + \frac{1}{2}(\hat{\vartheta}-\vartheta_0)^2 \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3}{\partial\vartheta^3} \ln f(X_i;\vartheta_0).$$

Umformen ergibt

$$\sqrt{n}(\hat{\vartheta}_n-\vartheta_0) = \frac{-\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\vartheta} \ln f(X_i;\vartheta_0)\right)}{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\vartheta^2} \ln f(X_i;\vartheta_0) + \frac{1}{2}(\hat{\vartheta}_n-\vartheta_0)^2 \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3}{\partial\vartheta^3} \ln f(X_i;\vartheta_0)} .$$

Der Zähler des rechten Ausdruckes ist nach dem zentralen Grenzwertsatz asymptotisch normalverteilt mit Erwartungswert 0 und Varianz $I(\vartheta_0)$, vgl. (4.14) und (4.15).

Der erste Term des Nenners geht nach dem schwachen Gesetz der großen Zahlen gegen $I(\vartheta_0)$. Da $\text{E}(|\partial^3 \ln f(X_i;\vartheta)/\partial\vartheta^3|) < K$, folgt mit der Konsistenz von $\hat{\vartheta}_n$, dass der zweite Term des Nenners gegen null geht.

Insgesamt hat $\sqrt{n}(\hat{\vartheta}_n-\vartheta_0)$ also asymptotisch eine Normalverteilung mit Erwartungswert 0 und Varianz $I(\vartheta_0)/I(\vartheta_0)^2 = I(\vartheta_0)^{-1}$.

Die *Kleinste-Quadrate-Schätzer* weisen unter allen linearen unverfälschten Schätzfunktionen minimale Varianz auf. Diese Optimalitätseigenschaft wird als BLUE (best linear unbiased estimator) bezeichnet.

Satz 6.2.23 *Gauß-Markoff-Theorem*

Gegeben sei das lineare Regressionsmodell

$$Y_i = \vartheta_0 + \vartheta_1 x_{1i} + \ldots + \vartheta_p x_{pi} + U_i \qquad i = 1,\ldots,n,$$

bzw. in Matrix-Schreibweise

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\vartheta} + \boldsymbol{U} \qquad \text{mit} \quad \text{E}(U_i) = 0; \quad \text{Cov}(U_i,U_j) = \begin{cases} \sigma^2 & i=j \\ 0 & i\neq j \end{cases} .$$

Die Matrix $\mathbf{X}^T\mathbf{X}$ habe vollen Rang. $\hat{\boldsymbol{\vartheta}}$ sei der Kleinste-Quadrate-Schätzer für $\boldsymbol{\vartheta}$. Dann hat für jedes $\boldsymbol{l}$ die Schätzfunktion $\boldsymbol{l}^T\hat{\boldsymbol{\vartheta}} = l_0\hat{\vartheta}_0 + \ldots + l_p\hat{\vartheta}_p$ die kleinste Varianz unter allen erwartungstreuen Schätzern für $\boldsymbol{l}^T\boldsymbol{\vartheta}$.

Beweis: Wir betrachten nur das einfache lineare Regressionsmodell ($p = 1$). Die folgende Beweisführung ist etwas umständlich. Eleganter und allgemeiner geht es, wenn man zur Matrizenformulierung des linearen Regressionsmodells übergeht. Dies übersteigt aber unseren gewählten Rahmen.

Nach Beispiel 6.1.6 gilt für den KQ-Schätzer $(\hat{\vartheta}_0, \hat{\vartheta}_1)^T$:

$$\hat{\vartheta}_0 = \bar{Y} - \hat{\vartheta}_1 \bar{x} \qquad \hat{\vartheta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n x_i Y_i - \bar{x}\bar{Y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}.$$

Dies lässt sich leicht umschreiben zu

$$\hat{\vartheta}_0 = \sum_{i=1}^n \left(\frac{1}{n} - \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} \bar{x} \right) Y_i, \qquad \hat{\vartheta}_1 = \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} Y_i.$$

Damit sind die KQ-Schätzer linear in den Y_i.

Etwas Rechnerei ergibt zudem $\mathrm{E}(\hat{\vartheta}_i) = \vartheta_i$, $i = 1,2$; somit sind die KQ-Schätzer auch erwartungstreu. Diese Eigenschaft zieht sich selbstverständlich durch, so dass $\boldsymbol{l}^T \hat{\boldsymbol{\vartheta}} = l_0 \hat{\vartheta}_0 + l_1 \hat{\vartheta}_1$ erwartungstreu für $\boldsymbol{l}^T \boldsymbol{\vartheta}$ ist.

Sei nun $T = \sum_{i=1}^n t_i Y_i$ ein beliebiger erwartungstreuer Schätzer für $\boldsymbol{l}^T \boldsymbol{\vartheta}$. Wir zeigen, dass dessen Varianz minimal wird für

$$t_i = l_0 \left(\frac{1}{n} - \frac{x_i \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) + l_1 \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad i = 1, \ldots, n.$$

Das sind gerade die Gewichte des KQ-Schätzers.

Da T erwartungstreu ist für $\boldsymbol{l}^T \boldsymbol{\vartheta}$, gilt

$$l_0 \vartheta_0 + l_1 \vartheta_1 = \mathrm{E}(T) = \mathrm{E}\left(\sum_{i=1}^n t_i Y_i \right) = \sum_{i=1}^n t_i \mathrm{E}(Y_i) = \left(\sum_{i=1}^n t_i \right) \vartheta_0 + \left(\sum_{i=1}^n t_i x_i \right) \vartheta_1$$

für alle ϑ_0, ϑ_1. Damit muss gelten:

$$\sum_{i=1}^n t_i = l_0, \qquad \sum_{i=1}^n t_i x_i = l_1. \tag{6.33}$$

Nun haben wir die Varianz von T, $\mathrm{Var}(T) = \sum_{i=1}^n t_i^2 \sigma^2$, unter den Nebenbedingungen (6.33) zu minimieren. Dafür verwenden wir den Lagrange-Ansatz. Wir setzen

$$Q(t_1, \ldots, t_n, \lambda_0, \lambda_1) = \sum_{i=1}^n t_i^2 \sigma^2 + \lambda_0 \left(\sum_{i=1}^n t_i - l_0 \right) + \lambda_1 \left(\sum_{i=1}^n t_i x_i - l_1 \right).$$

Die partiellen Ableitungen sind

$$\frac{\partial}{\partial t_i}Q = 2\sigma^2 t_i + \lambda_0 + \lambda_1 x_i \qquad i = 1,\ldots,n$$
$$\frac{\partial}{\partial \lambda_0}Q = \sum_{i=1}^{n} t_i - l_0$$
$$\frac{\partial}{\partial \lambda_1}Q = \sum_{i=1}^{n} t_i x_i - l_1.$$

Nullsetzen dieser Ableitungen ergibt ein notwendiges System von Gleichungen. Summation der ersten Gleichung ergibt dann

$$2\sigma^2 \bar{t} + \lambda_0 + \lambda_1 \bar{x} = 0. \tag{6.34}$$

Dann multiplizieren wir die Gleichungen $i = 1,\ldots,n$ jeweils mit x_i und addieren anschließend wieder. Dies ergibt

$$2\sigma^2 \overline{tx} + \lambda_0 \bar{x} + \lambda_1 \overline{x^2} = 0, \tag{6.35}$$

wobei $\overline{tx} = \sum_{i=1}^{n} t_i x_i/n$, $\overline{x^2} = \sum_{i=1}^{n} x_i^2/n$. Aus den beiden Gleichungen erhalten wir

$$\frac{l_0}{n} = \bar{t} \qquad \text{und} \qquad \frac{l_1}{n} = \overline{tx}.$$

Einsetzen in (6.34) und (6.35) führt auf

$$2\sigma^2 \frac{l_0}{n} + \lambda_0 + \lambda_1 \bar{x} = 0 \quad \text{und} \quad 2\sigma^2 \frac{l_1}{n} + \lambda_0 \bar{x} + \lambda_1 \overline{x^2} = 0.$$

Auflösen nach λ_0 und λ_1 bringt

$$\lambda_0 = -\frac{2\sigma^2}{n}\left(l_0 - \frac{l_1 - l_0\bar{x}}{\overline{x^2} - \bar{x}^2}\bar{x}\right) \quad \text{und} \quad \lambda_1 = \frac{2\sigma^2}{n}\frac{l_0\bar{x} - l - 1}{\overline{x^2} - \bar{x}^2}.$$

Dies wird wieder oben eingesetzt:

$$2\sigma^2 t_i - \frac{2\sigma^2}{n}\left(l_0 - \frac{l_1 - l_0\bar{x}}{\overline{x^2} - \bar{x}^2}x_i\right) = 0.$$

Wir erhalten

$$t_i = l_0\left(\frac{1}{n} - \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x}\right) + l_1\frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Also ist T der Schätzer $\boldsymbol{l}^T\hat{\boldsymbol{\theta}}$.

Beispiel 6.2.24 *Regression durch den Nullpunkt*

Wir wollen anhand einer einfachen Situation die Optimalität des KQ-Schätzers verdeutlichen. Dazu sei das folgende Modell gegeben:

$$Y_i = \vartheta x_i + U_i \qquad i = 1, \ldots, n,$$

für das die üblichen Bedingungen $\mathrm{E}(U_i) = 0$, $\mathrm{Var}(U_i) = \sigma^2$, $\mathrm{Cov}(U_i, U_j) = 0$, $i \neq j$ gelten. Der KQ-Schätzer für ϑ ist dann

$$\hat{\vartheta} = \sum_{i=1}^{n} \frac{x_i}{\sum_{j=1}^{n} x_j^2} Y_i.$$

Für ihn gilt:

$$\mathrm{E}(\hat{\vartheta}) = \mathrm{E}\left(\sum_{i=1}^{n} \frac{x_i}{\sum_{j=1}^{n} x_j^2} Y_i \right) = \sum_{i=1}^{n} \frac{x_i}{\sum_{j=1}^{n} x_j^2} \vartheta x_i = \vartheta,$$

$$\mathrm{Var}(\hat{\vartheta}) = \mathrm{Var}\left(\sum_{i=1}^{n} \frac{x_i}{\sum_{j=1}^{n} x_j^2} Y_i \right) = \sum_{i=1}^{n} \frac{x_i^2}{\left(\sum_{j=1}^{n} x_j^2 \right)^2} \mathrm{Var}(Y_i) = \frac{\sigma^2}{\sum_{j=1}^{n} x_j^2}.$$

Einen anderen linearen Schätzer für ϑ erhalten wir, wenn wir nicht von vornherein berücksichtigen, dass der Achsenabschnitt Null ist. Dann ist

$$\tilde{\vartheta} = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} Y_i.$$

Es ist

$$\mathrm{E}(\tilde{\vartheta}) = \mathrm{E}\left(\sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} Y_i \right) = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \vartheta x_i = \vartheta.$$

$$\mathrm{Var}(\tilde{\vartheta}) = \mathrm{Var}\left(\sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} Y_i \right) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\left(\sum_{j=1}^{n} (x_j - \bar{x})^2 \right)^2} \sigma^2 = \frac{\sigma^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}.$$

Wegen $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (x_i - 0)^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$ ist die Varianz von $\hat{\vartheta}$ kleiner (oder gleich) der Varianz von $\tilde{\vartheta}$.

Das Modell ohne Achsenabschnitt wird i.d.R. für positive x_i verwendet. Dann kann der Varianzunterschied beträchtlich sein.

Wir betrachten nun die *L*- und *M-Schätzer* für Lageparameter einer Lage-Skalen-Familie $\mathscr{F} = \{F | F(x) = F_0((x - \vartheta)/\tau),\ -\infty < \vartheta < \infty,\ \tau > 0\}$. Wie bei der Einführung dieser Schätzer erwähnt wurde, sind sie vor allem unter dem Gesichtspunkt von Interesse, dass das zugehörige Modell nicht (genau) bekannt ist, bzw. dass man sich gegen einzelne extreme Werte schützen möchte. Darauf gehen wir im nächsten Abschnitt noch genauer ein.

Bei kleinen Stichproben können bei bekannter Verteilung F_0 mit dem Gauß-Markoff-Theorem unter Verwendung der Erwartungswerte der geordneten Statistiken und ihrer Varianzen und Kovarianzen optimale Gewichte bestimmt werden. Bei großen Stichprobenumfängen ist es sinnvoll, die Gewichte über eine auf $(0;1)$ definierte Gewichtsfunktion $J(u)$ auszudrücken:

$$c_{n,i} = \frac{1}{n} \cdot J\left(\frac{i}{n+1}\right).$$

Sie sind dann gegeben durch

$$J(u) \propto g(F^{-1}(u)),$$

wobei

$$g(x) = -\frac{\partial}{\partial x}\left(\frac{f'(x)}{f(x)}\right) = -\frac{\partial^2}{\partial x^2}\ln(f(x)).$$

Man verlangt nur Proportionalität von $J(u)$ und $g(F^{-1}(u))$, um den Erwartungswert mit dem zu schätzenden Parameter in Übereinstimmung bringen zu können.

Beispiel 6.2.25 *Gewichte für L-Schätzer bei verschiedenen Verteilungen*

Ist die zugrunde liegende Verteilungsfamilie die Normalverteilung mit festem σ^2, so ist

$$g(x) = -\frac{\partial^2}{\partial x^2}\ln(f(x)) = \frac{1}{\sigma^2}.$$

D.h. die Gewichte sind konstant; wegen der Normierung gilt, $c_{n,i} = 1/n$, bzw. $J(u) \equiv 1$. Der optimale Schätzer ist das arithmetische Mittel.

Bei der Laplace-Verteilung erhalten wir

$$\frac{1}{n} \cdot J(u) = \begin{cases} 1 & u = 0.5 \\ 0 & \text{sonst.} \end{cases}$$

Der resultierende Schätzer ist der Median.

Bei der logistischen Verteilung gilt

$$J(u) = u(1-u).$$

Die Gewichte sind symmetrisch um das mittlere Quantil und nehmen zum Rand hin ab.

Unter verschiedenen Restriktionen an die zugrunde liegende Verteilung und die Funktion $J(u)$ sind L-Schätzer asymptotisch normalverteilt. Siehe dazu Serfling (1980). Dort finden sich auch Aussagen zur asymptotischen Normalverteiltheit von M-Schätzern. Siehe aber auch den Satz 6.2.6.

Zwischen M- und L-Schätzern gibt es eine enge Verwandschaft, die in Hall/Joiner (1983) aufgezeigt wurde. Um diese zu erkennen, gehen wir aus von einer stetigen Verteilung mit einer um ϑ symmetrischen Dichte $f(x)$. Der asymptotisch effiziente ML-Schätzer ist dann der optimale M-Schätzer; er hat dann die ψ-Funktion $\psi(x) = -f'(x)/f(x)$ und ist Lösung von

$$\sum \psi(X_{i:n} - \vartheta) = \sum \frac{f'(X_{i:n} - \vartheta)}{f(X_{i:n} - \vartheta)} = 0. \tag{6.36}$$

Nun nehmen wir eine spezielle Abweichung $(X_{i:n}-\vartheta)$ und approximieren f'/f linear bei $F^{-1}(i/(n+1))$. Das ergibt

$$\frac{f'(X_{i:n}-\vartheta)}{f(X_{i:n}-\vartheta)} \cong \frac{f'(F^{-1}(i/(n+1)))}{f(F^{-1}(i/(n+1)))} + \left\{(X_{i:n}-\vartheta) - F^{-1}\left(\frac{i}{n+1}\right)\right\} \cdot \frac{\partial}{\partial x}\left(\frac{f'(x)}{f(x)}\right)\Big|_{x=F^{-1}(i/(n+1))}.$$

Setzen wir nun

$$J(u) = \frac{\partial}{\partial x}\left(\frac{f'(x)}{f(x)}\right)\Big|_{x=F^{-1}(u)},$$

und summieren wir über i, so erhalten wir mit den Eigenschaften von f gerade den optimalen L-Schätzer:

$$\hat{\vartheta} = \frac{\frac{1}{n}\sum_{i=1}^{n} J\left(\frac{i}{n+1}\right) X_{i:n}}{\sum_{i=1}^{n} J\left(\frac{i}{n+1}\right)}. \tag{6.37}$$

Die beiden Schätzer sind also ganz ähnlich definiert. Der M-Schätzer durch

$$-\sum \frac{f'(X_{i:n}-\vartheta)}{f(X_{i:n}-\vartheta)} = 0$$

und der L-Schätzer analog:

$$-\sum \left[\text{lin.Approx von } \frac{f'}{f}\right](X_{i:n}-\vartheta) = 0\,.$$

Wie die ψ-Funktion beim M-Schätzer wird die Gewichtsfunktion J eines L-Schätzers i.d.R. nicht aufgrund einer konkreten Modellvorstellung sondern eher unter dem Gesichtspunkt der Eigenschaften des resultierenden Schätzers gewählt.

Beispiel 6.2.26 *Hubers M-Schätzer und das winsorisierte Mittel*

Der zu Hubers ψ-Funktion passende L-Schätzer ist das α-winsorisierte Mittel mit $0 < \alpha < 1/2$,

$$\bar{X}_{W,\alpha} = \frac{1}{n}\left\{[n\alpha]X_{[n\alpha]:n} + \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{i:n} + [n\alpha]X_{n-[n\alpha]:n}\right\}. \tag{6.38}$$

An den Rändern werden also jeweils etwa $(1-\alpha)100\%$ der Beobachtungen durch den zugehörigen innersten Wert ersetzt.

Zur Beurteilung von *nichtparametrischen Dichteschätzern* $\hat{f}$ in einem Punkt x bietet sich der mittlere quadratische Fehler an:

$$\mathrm{MSE}_x(\hat{f}) = \mathrm{E}([\hat{f}(x)-f(x)]^2) = \int_{-\infty}^{+\infty} (\hat{f}(x)-f(x))^2 f(x)\,\mathrm{d}x = [\mathrm{E}(\hat{f}(x))-f(x)]^2 + \mathrm{Var}(\hat{f}(x)). \tag{6.39}$$

Das verbreitetste globale Gütemaß ist *der mittlere integrierte quadratische Fehler*, kurz *MISE*:

$$\mathrm{MISE}(\hat{f}) = \mathrm{E}\left[\int_{-\infty}^{+\infty} (\hat{f}(x)-f(x))^2\,\mathrm{d}x\right].$$

Dies kann umgeformt werden zu

$$\begin{aligned}\mathrm{MISE}(\hat{f}) &= \int_{-\infty}^{+\infty} \mathrm{E}[(\hat{f}(x)-f(x))^2]\,\mathrm{d}x = \int_{-\infty}^{+\infty} \mathrm{MSE}_x(\hat{f})\,\mathrm{d}x \\ &= \int_{-\infty}^{+\infty} (\mathrm{E}(\hat{f}(x))-f(x))^2\mathrm{d}x + \int_{-\infty}^{+\infty} \mathrm{Var}(\hat{f}(x))\,\mathrm{d}x.\end{aligned} \tag{6.40}$$

Der MISE ist also die Summe des integrierten quadrierten Bias und der integrierten Varianz.

Für Kern-Dichteschätzer $\hat{f}(x) = \sum_{i=1}^{n} K((x-X_i)/h)/(h\cdot n)$ erhalten wir mit der Unabhängigkeit der Stichprobenvariablen:

$$\mathrm{E}(\hat{f}(x)) = \frac{1}{h\cdot n}\sum_{i=1}^{n} \mathrm{E}\left[K\left(\frac{x-X_i}{h}\right)\right] = \int_{-\infty}^{+\infty} \frac{1}{h}K\left(\frac{x-z}{h}\right) f(z)\,\mathrm{d}z$$

und

$$n\cdot\mathrm{Var}(\hat{f}(x)) = \int_{-\infty}^{+\infty} \frac{1}{h^2}K\left(\frac{x-z}{h}\right)^2 f(z)\,\mathrm{d}z - \left(\int_{-\infty}^{+\infty} \frac{1}{h}K\left(\frac{x-z}{h}\right) f(z)\,\mathrm{d}z\right)^2.$$

Im Prinzip können wir diese Ausdrücke in (6.40) einsetzen und damit explizite Formeln für den MISE erhalten. Jedoch ergeben sich außer in wenigen Spezialfällen sehr komplizierte und wenig aufschlussreiche Formeln.

Die hier angegebenen Resultate zeigen jedoch, dass Kerndichteschätzungen geglättete Versionen der zugrunde liegenden Dichtefunktionen schätzen. Um wenigstens asymptotisch die Dichte unverzerrt zu schätzen, muss mit $n \to \infty$ die Bandbreite h gegen Null gehen. – Andererseits erhöht natürlich eine kleine Bandbreite die Varianz und damit den MISE der Kerndichteschätzung.

Optimalität von Kerndichteschätzern bedeutet Minimierung des MISE. Allerdings hängt der MISE von der Bandbreite, dem Kern und der zugrunde liegenden Dichte selbst ab. Unter gewissen Annahmen ist jedoch der sogenannte *Epanechnikoff-Kern* optimal:

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right) & -\sqrt{5} \le t \le \sqrt{5} \\ 0 & \text{sonst.} \end{cases} \tag{6.41}$$

Zu weiteren Diskussionen hierzu sei auf Scott (1990) und auf Silverman (1986) verwiesen.

6.2.5 Robustheit

Die bisher vorgestellten Gütekriterien gehen von einer durch einen Parameter charakterisierten Verteilungsfamilie aus. Solche Verteilungsannahmen sind aber nach neuerdings allgemein akzeptierter Erkenntnis vielfach dubios. Das wirft aber sofort die Frage auf, wie sich Schätzer verhalten, wenn das unterstellte Modell nicht exakt gilt, sondern nur ‚in etwa'. Die näherungsweise Gültigkeit eines Modells wird durch das folgende Beispiel verdeutlicht.

Beispiel 6.2.27 *kontaminierte Normalverteilung*

Empirische Datensätze, die zum größten Teil aus der Beobachtung einer Normalverteilung resultieren, zu einem kleinen Teil aus 'verschmutzten' Daten bestehen, können mittels einer *kontaminierten Normalverteilung* modelliert werden. Deren Dichte lautet

$$f(x)=(1-\varepsilon)\cdot\frac{1}{\sigma_1}\cdot\varphi\left(\frac{x-\mu_1}{\sigma_1}\right)+\varepsilon\cdot\frac{1}{\sigma_2}\cdot g\left(\frac{x-\mu_2}{\sigma_2}\right). \tag{6.42}$$

ε ist der 'Verschmutzungsgrad'. Man geht davon aus, dass nicht selten $\varepsilon = 0.1$ gilt, also 10% der Daten verfälscht sind. Die verfälschten Werte stammen aus einer Verteilung mit der Dichte $g(x)$, die i.d.R. als symmetrisch um null angenommen wird. Um Ausreißer, d.h. sehr extreme Werte zu modellieren, unterstellt man gern für $g(x)$ ebenfalls eine Normalverteilung, allerdings eine mit wesentlich größerer Varianz.

Schätzer reagieren auf Ausreißer, einige stärker als andere. Um diese qualitative Einschätzung zu quantifizieren, bedarf es der Erfassung (mindestens) zweier Aspekte, die hier von der Daten-Seite aus angegangen werden. Dann spricht man auch von der Resistenz von Schätzern und reserviert den Begriff Robustheit für die Unempfindlichkeit bei Modellabweichungen. Die beiden Ansätze liefern weitgehend gleiche Einschätzungen. Daher ist es gerechtfertigt, im Folgenden die beiden Begriffe weitgehend synonym zu verwenden.

Bzgl. der angesprochenen Aspekte ist einmal die Reaktion der Schätzer auf die Größe von Ausreißern zu untersuchen. Bei resistenten Schätzern sollten wenige, beliebig weit von den restlichen Werten weg liegende Ausreißer nur einen beschränkten Einfluss haben. Die Beschreibung dieser Eigenschaft geschieht mit Hilfe der Einflussfunktion. Der zweite Aspekt betrifft die Anzahl der Ausreißer, die ein Schätzer verkraften kann, bevor er beliebig unsinnige Schätzwerte liefert. Er wird durch den Bruchpunkt erfasst.

Ein weiterer wesentlicher Gesichtspunkt bei der Betrachtung und dem Einsatz resistenter Schätzer ist ihre Effizienz. Ein 'beliebig' resistenter Schätzer ist bei weitem nicht so günstig wie einer, der zwar auch resistent ist, aber zudem noch in der Ausreißer-freien Situation einen im Verhältnis kleinen MQF aufweist. Das naheliegendste Beispiel dafür ist eine konstante Schätzfunktion. Diese ist sehr resistent - noch so viele extreme Werte können sie nicht beeinflussen; allerdings ist sie i.d.R. auch sehr ineffizient.

Zu den einsichtigsten Konzepten der robusten Statistik gehört der von Hampel (1971) eingeführte Bruchpunkt eines Schätzers. Grob gesprochen ist er der kleinste Anteil an 'verfälschten' Beobachtungen, der bewirken kann, dass der Schätzwert beliebig fern von dem zu schätzenden Parameterwert zu liegen kommen kann. Seine Bedeutung ist vor allem im Bereich endlicher Stichproben zu sehen. Für diese Situation geben Donoho & Huber (1987) eine neue Definition. Ihr Konzept ist sehr datenanalytisch ausgerichtet und verzichtet auf modellmäßige Aspekte. Die obige rohe Formulierung des Bruchpunktes macht schon deutlich, dass dieses Konzept nur für nicht beschränkte Schätzer geeignet ist. Die Unbeschränktheit erlaubt eine eindeutige Fassung der angesprochenen 'beliebigen Ferne'. Bei beschränkten Schätzfunktionen gibt es diesbezüglich Schwierigkeiten. So ist etwa die Korrelation ρ_{XY} zweier Zufallsvariablen X und Y durch ± 1 beschränkt. Es sind aber nicht nur diese Werte von Bedeutung, sondern auch der Wert null: Schätzwerte um null deuten auf das Fehlen linearer Abhängigkeiten hin. Wechselt die Korrelation gar ihr Vorzeichen, so wird auf eine falsche Abhängigkeitsrichtung geschlossen.

Ausgangspunkt für die Definition des Bruchpunktes im Fall unbeschränkter Parameter ist die

Definition 6.2.28 *ε-Ersetzung*

Sei $\boldsymbol{x} = (x_1, \ldots, x_n)$ eine gegebene Folge von Beobachtungen. Eine ε-Ersetzung von $\boldsymbol{x}$ ist eine Beobachtungsfolge $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)$, die sich an höchstens $\varepsilon \cdot n$ Stellen von $\boldsymbol{x}$ unterscheidet. Ihr Abstand ist definiert durch

$$d_n(\boldsymbol{x}, \boldsymbol{x}^*) := \frac{1}{n} \cdot \# \left\{ i | x_i \neq x_i^* \quad i = 1, \ldots, n \right\} \leq \varepsilon.$$

Die ε-Ersetzung ist ein einfach zu handhabendes Verschmutzungsmodell, das weitgehend der Kontamination von Verteilungen entspricht.

Definition 6.2.29 *Bruchpunkt*

Sei T_n eine nicht-beschränkte Schätzfunktion. Der *maximale Bias* von T_n bei $\boldsymbol{x} = (x_1, \ldots, x_n)$, der durch eine ε-Ersetzung bewirkt werden kann, ist

$$b(\varepsilon, \boldsymbol{x}, T_n) = \sup\{||T_n(x) - T_n(x^*)|| \, | x^* \in \mathbb{R}^n, \; d_n(\boldsymbol{x}, \boldsymbol{x}^*) \leq \varepsilon\}.$$

Der *Bruchpunkt* von T_n bei $\boldsymbol{x}$ ist definiert als

$$\varepsilon^*(\boldsymbol{x}, T_n) = \inf\{\varepsilon | b(\varepsilon, \boldsymbol{x}, T_n) = \infty\}. \tag{6.43}$$

Wie oben angedeutet, ist der Bruchpunkt eines Schätzers grob gesprochen der kleinste Anteil verfälschter Werte, der zum ‚Zusammenbrechen' der Schätzung führen kann. Der höchste, sinnvoll erreichbare Bruchpunkt ist der Wert 1/2. Denn bei der Hälfte oder sogar mehr verfälschter Werte können die richtigen nicht mehr eindeutig identifiziert werden.

Beispiel 6.2.30 *Bruchpunkt des arithmetischen Mittels*

Der Bruchpunkt des arithmetischen Mittels beträgt $\varepsilon^* = 1/n$, d. h. ein Ausreißer unter n Werten kann die Schätzung vollständig zerstören. Um das einzusehen, sei

$$x_i^* = \begin{cases} x_i & i = 2, \ldots, n \\ x_i + c & i = 1 \end{cases}.$$

Damit gilt für $T_n(\boldsymbol{x}) = \bar{x}$:

$$T_n(\boldsymbol{x}) - T_n(\boldsymbol{x}^*) = \bar{x} - \bar{x^*} = -\frac{c}{n}.$$

Bei festem n kann die Differenz über die Wahl von c beliebig groß gemacht werden, so dass

$$b\left(\frac{1}{n}, \boldsymbol{x}, T_n\right) = \infty$$

und folglich $\varepsilon^* = 1/n$.

Beispiel 6.2.31 *Bruchpunkt des Median*

Der Median ist ein Schätzer für das Niveau eines Datensatzes, der einen Bruchpunkt von 1/2 besitzt.

Nach der Definition des Median, siehe (6.11), können nämlich auf einer Seite $(n-1)/2$ bei ungeradem n, bzw. $(n-2)/2$ bei geradem n Werte verfälscht sein, ohne dass sich der Median ändert. Die Ablenkung eines weiteren Wertes reicht aber aus, um einen beliebig großen Bias zu erzeugen. Somit bricht der Median als Schätzer zusammen, wenn der Anteil der verfälschten Werte

$$\frac{1}{2}+\frac{1}{2n} \quad \text{bei ungeradem } n \quad \text{und} \quad \frac{1}{2} \quad \text{bei geradem } n$$

beträgt. Der Bruchpunkt ist, kurz gesprochen, gleich 1/2.

Einen zweiten wichtigen Aspekt der Robustheit stellt die (mögliche) Reaktion eines Schätzers auf einen geringfügigen Verschmutzungsgrad einer Stichprobe dar. Zu ihrer Beschreibung führte Hampel (1974) die Influenzfunktion ein. Einen adäquaten Eindruck von der Influenzfunktion erhält man in vielen Fällen durch die für endliche Stichproben definierte *Sensitivitätskurve* von Tukey:

$$\mathrm{SC}(x;T_n;x_1,\ldots,x_n)=n\cdot[T_n(x_1,\ldots,x_{n-1},x)-T_n(x_1,\ldots,x_n)].$$

Die Sensitivitätskurve zeigt, wie sich ein Schätzwert bei Verändern einer einzelnen Beobachtung in einer konkreten Stichprobe mit dem Grad der Änderung selbst verändert. Um uns von konkreten Stichproben zu lösen, wird der Erwartungswert dieser Differenzen betrachtet:

$$\mathrm{E}_\vartheta(n\cdot[T_n(X_1,\ldots,X_{n-1},x)-T_n(X_1,\ldots,X_n)]).$$

Die Einflussfunktion IF erhält man dann als Grenzwert. Damit dies sinnvoll ist, wird vorausgesetzt, dass T konsistent für ϑ ist. Zudem kann für $\mathrm{E}_\vartheta[T_n(X_1,\ldots,X_{n-1},x)]$ mit festgehaltenem x auch $\mathrm{E}_\vartheta[T_n(X_1,\ldots,X_n)|X_n=x]$ geschrieben werden.

Definition 6.2.32 *Einflussfunktion*

X sei eine Zufallsvariable mit einer vom Parameter ϑ abhängigen Verteilung, $\boldsymbol{X}=(X_1,\ldots,X_n)$ sei eine Stichprobe aus der Verteilung von X und $T(\boldsymbol{X})$ ein konsistenter Schätzer für ϑ. Die *Einflussfunktion* von $T(\boldsymbol{X})$ bei ϑ ist der Grenzwert

$$\mathrm{IF}(x;T;\vartheta)=\lim_{n\to\infty}(n\cdot[\mathrm{E}_\vartheta(T_n(\boldsymbol{X})|X_n=x)-\mathrm{E}_\vartheta(T_n(\boldsymbol{X}))]), \tag{6.44}$$

sofern dieser existiert.

Dies ist nicht die standardmäßige, auf Hampel zurückgehende Definition. Eingeführt wurde die Einflussfunktion für Schätzer, die sich in der Form $T(\hat{F}_n)$ schreiben lassen. Ein Beispiel bildet das arithmetische Mittel von transformierten Stichprobenvariablen:

$$T(\hat{F}_n)=\int g(x)\,\mathrm{d}\hat{F}_n(x)=\frac{1}{n}\sum_i g(X_i). \tag{6.45}$$

Diese Schätzer seien konsistent für den entsprechenden Verteilungsparameter $T(F)$. Die Funktionalschreibweise erlaubt die Betrachtung des Parameters bei einer verschmutzten Verteilung gemäß $T((1-\epsilon)F+\epsilon H_x)$; hier ist $H_x(x')=1$ für $x'\geq x$ und $=0$ sonst. Der Einfluss einer ‚infinitesimalen' Verschmutzung wird dann gemessen durch

$$\lim_{\epsilon\to\infty}\frac{T((1-\epsilon)F+\epsilon H_x)-T(F)}{\epsilon}.$$

Die Verwandtschaft mit unserer Definition wird deutlich, wenn wir für Schätzer der Form $T(\hat{F}_n)$ mit festgehaltenem Wert $X_n=x$ $T((1-1/n)\hat{F}_{n-1}+H_x/n)$ schreiben. Dann ist

$$\mathrm{SC}(x;T_n;X_1,\ldots,X_n)=\frac{T\left(\left(1-\frac{1}{n}\right)\hat{F}_{n-1}+\frac{1}{n}H_x\right)-T(\hat{F}_n)}{1/n}.$$

Für große n ist dies mit der Konsistenz der Zähler approximativ gleich

$$T\left(\left(1-\frac{1}{n}\right)F+\frac{1}{n}H_x\right)-T(F).$$

Setzen wir noch $\epsilon=1/n$, so sind wir für $n\to\infty$ bei der klassischen Definition.

Einflussfunktionen lassen sich in folgender Weise linear kombinieren: Wenn T_n und S_n zwei konsistente Folgen von Schätzern für zwei Parameter ϑ_1 und ϑ_2 sind, so gilt:

$$\mathrm{IF}(x,a+bT+cS,a+b\vartheta_1+c\vartheta_2)=b\cdot\mathrm{IF}(x,T,\vartheta_1)+c\cdot\mathrm{IF}(x,S,\vartheta_2).$$

Beispiel 6.2.33 *Einflussfunktion des arithmetischen Mittels*

Für das arithmetische Mittel gilt:

$$n\cdot[T_n(X_1,\ldots,X_{n-1},x)-T_n(X_1,\ldots,X_n)]=n\cdot\left[\frac{1}{n}\left(\sum_{i=1}^{n-1}X_i+x\right)-\frac{1}{n}\sum_{i=1}^{n}X_i\right]=x-X_n.$$

Da x fest ist, gilt, sofern μ der Erwartungswert der zugrunde liegenden Verteilung ist:

$$\mathrm{IF}(x;\bar{X};\mu)=x-\mu.$$

Der Einfluss eines einzelnen fehlerhaften Wertes auf die Schätzung von μ mittels $\bar{X}$ ist linear. Je weiter ein Ausreißer von dem Datensatz weg lieg, desto schlimmer. Er kann die Schätzung vollständig verfälschen, wie die Unbeschränktheit von $\mathrm{IF}(x;\bar{X};\mu)$ zeigt. $\bar{X}$ ist nicht robust.

Wir betrachten nun die Einflussfunktion von M-Schätzern.

Satz 6.2.34 *Einflussfunktion von M-Schätzern*

$X_1,\ldots,X_n$ sei eine Stichprobe aus einer von dem Parameter ϑ abhängigen Verteilung, $X_i\sim F(x;\vartheta)$, $\vartheta\in\Theta$. ψ sei eine Psi-Funktion mit beschränkter zweiter Ableitung, $|\partial^2\psi(x,\vartheta)/\partial\vartheta^2|<K$, für die weiter gilt: $\mathrm{E}_\vartheta(\psi(X,\vartheta))=0$.

Der M-Schätzer $\hat{\vartheta}_n$, d.h. die Lösung von $\sum_{i=1}^{n} \psi(X_i, \vartheta) = 0$, sei konsistent.

Weiter sei auch für jeweils festes x die Lösung $\tilde{\vartheta}_n$ von $\sum_{i=1}^{n-1} \psi(X_i, \vartheta) + \psi(x, \vartheta) = 0$ konsistent. Dann ist die Einflussfunktion von $\hat{\vartheta}_n$ gegeben durch

$$\mathrm{IF}(x, \hat{\vartheta}_n, \vartheta) = -\frac{\psi(x, \vartheta)}{\mathrm{E}(\partial \psi(X, \vartheta)/\partial \vartheta)}. \tag{6.46}$$

Beweis: Sei ϑ_0 der wahre Parameterwert und sei $\hat{\vartheta}_n$ Lösung von

$$\sum_{i=1}^{n} \psi(X_i, \vartheta) = 0.$$

Der Mittelwertsatz der Differentialrechnung liefert mit $0 \le c_i \le 1$:

$$\psi(X_i, \hat{\vartheta}_n) = \psi(X_i, \vartheta_0) + \left(\frac{\partial}{\partial \vartheta} \psi(X_i, \vartheta_0 + c_i(\hat{\vartheta}_n - \vartheta_0)) \right) (\hat{\vartheta}_n - \vartheta_0).$$

Damit folgt:

$$0 = \sum_{i=1}^{n} \psi(X_i, \hat{\vartheta}_n) = \sum_{i=1}^{n} \psi(X_i, \vartheta_0) + \sum_{i=1}^{n} \left(\frac{\partial}{\partial \vartheta} \psi(X_i, \vartheta_0 + c_i(\hat{\vartheta}_n - \vartheta_0)) \right) (\hat{\vartheta}_n - \vartheta_0),$$

und weiter:

$$\hat{\vartheta}_n - \vartheta_0 = -\frac{\displaystyle\sum_{i=1}^{n} \psi(X_i, \vartheta_0)}{\displaystyle\sum_{i=1}^{n} \frac{\partial}{\partial \vartheta} \psi(X_i, \vartheta_0 + c_i(\hat{\vartheta}_n - \vartheta_0))}.$$

Analog erhalten wir für festgehaltenes $X_n = x$, wenn $\tilde{\vartheta}_n$ Lösung von $\sum_{i=1}^{n-1} \psi(X_i, \vartheta) + \psi(x, \vartheta) = 0$ ist:

$$\tilde{\vartheta}_n - \vartheta_0 = -\frac{\displaystyle\sum_{i=1}^{n-1} \psi(X_i, \vartheta_0) + \psi(x, \vartheta_0)}{\displaystyle\sum_{i=1}^{n-1} \frac{\partial}{\partial \vartheta} \psi(X_i, \vartheta_0 + d_i(\tilde{\vartheta}_n - \vartheta_0)) + \frac{\partial}{\partial \vartheta} \psi(x, \vartheta_0 + d(\tilde{\vartheta}_n - \vartheta_0))}.$$

Auch hier ist $0 \le d,\ d_i \le 1$. Folglich gilt:

$$n(\tilde{\vartheta}_n - \hat{\vartheta}) = -\frac{\psi(x, \vartheta_0)}{\frac{1}{n} \sum_{i=1}^{n-1} \frac{\partial}{\partial \vartheta} \psi(X_i, \vartheta_0 + d_i(\tilde{\vartheta}_n - \vartheta_0)) + \frac{1}{n} \cdot \frac{\partial}{\partial \vartheta} \psi(x, \vartheta_0 + d(\tilde{\vartheta}_n - \vartheta_0))}$$

$$- \frac{\sum_{i=1}^{n-1} \psi(X_i, \vartheta)}{\frac{1}{n}\sum_{i=1}^{n-1} \frac{\partial}{\partial \vartheta}\psi(X_i, \vartheta_0 + c_i(\hat{\vartheta}_n - \vartheta_0)) + \frac{1}{n} \cdot \frac{\partial}{\partial \vartheta}\psi(x, \vartheta_0 + d(\tilde{\vartheta}_n - \vartheta_0))}$$

$$+ \frac{\sum_{i=1}^{n} \psi(X_i, \vartheta)}{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \vartheta}\psi(X_i, \vartheta_0 + c_i(\hat{\vartheta}_n - \vartheta_0))}.$$

Aus der Beschränktheit der zweiten Ableitungen folgt:

$$\left| \frac{\partial}{\partial \vartheta}\psi(X_i, \vartheta_0 + c_i(\hat{\vartheta}_n - \vartheta_0)) - \frac{\partial}{\partial \vartheta}\psi(X_i, \vartheta_0) \right| \leq K|\vartheta_0 - (\vartheta_0 + c_i(\hat{\vartheta}_n - \vartheta_0))| = K c_i |\hat{\vartheta}_n - \vartheta_0|.$$

Da

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \vartheta}\psi(X_i, \vartheta_0) \quad \xrightarrow{P} \quad \mathrm{E}\left(\frac{\partial}{\partial \vartheta}\psi(X, \vartheta_0)\right),$$

erhalten wir mit der Konsistenz von $\hat{\vartheta}_n$ auch:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \vartheta}\psi(X_i, \vartheta_0 + c_i(\hat{\vartheta}_n - \vartheta_0)) \quad \xrightarrow{P} \quad \mathrm{E}\left(\frac{\partial}{\partial \vartheta}\psi(X, \vartheta_0)\right).$$

Das gleiche Verhalten gilt für die Terme mit $\tilde{\vartheta}_n$. Da zudem

$$\frac{1}{n} \cdot \frac{\partial}{\partial \vartheta}\psi(x, \vartheta_0 + d(\tilde{\vartheta}_n - \vartheta_0)) \quad \xrightarrow{P} \quad 0$$

und $\mathrm{E}\psi(X, \vartheta_0) = 0$, ist schließlich

$$\lim_{n\to\infty} n \cdot \mathrm{E}(\tilde{\vartheta}_n - \hat{\vartheta}_n) = -\frac{\psi(x, \vartheta)}{\mathrm{E}(\partial\psi(X, \vartheta)/\partial\vartheta)}.$$

Die im Satz geforderte Eigenschaft $\mathrm{E}_\vartheta(\psi(X, \vartheta)) = 0$ geht konform mit der üblichen Bestimmung von ML-Schätzern aus den Normalgleichungen:

$$\mathrm{E}\left(\frac{\partial}{\partial \vartheta} \ln f(X_1, \ldots, X_n; \vartheta)\right) = 0.$$

Auch die Konsistenz von $\hat{\vartheta}_n$ und $\tilde{\vartheta}_n$ ist in vielen Anwendungen gegeben, vgl. den letzten Satz. Dieser zeigt auch den Zusammenhang:

$$\mathrm{Var}\left(\sqrt{n}(\hat{\vartheta} - \vartheta)\right) \doteq \frac{\mathrm{E}(\psi(X_i, \vartheta)^2)}{\mathrm{E}(\partial\psi(X_i, \vartheta)/\partial\vartheta)^2} = \mathrm{E}(\mathrm{IF}(X, \hat{\vartheta}, \vartheta)^2). \tag{6.47}$$

Die Einflussfunktion von L-Schätzern leiten wir in mehreren Schritten ab. Zuerst betrachten wir die Schätzung eines einzelnen Quantils.

Beispiel 6.2.35 *L-Schätzer eines Quantils*

Wir betrachten die Schätzung eines Verteilungsquantils $q_\alpha = F^{-1}(\alpha)$ einer stetigen Verteilung mittels der geordneten Statistik $X_{n,[\alpha n]+1}$. Stichprobenquantile sind bekanntlich konsistent für die theoretischen Quantile, sofern diese eindeutig sind. Das soll im Weiteren unterstellt sein. Für die Einflussfunktion benötigen wir den bedingten Erwartungswert $\mathrm{E}(X_{r:n}|X_n = z) = \int x \cdot f_{X_{r:n}|X_n}(x|z)\,\mathrm{d}x$. Die bedingte Dichte erhalten wir leicht mit der gleichen heuristischen Methode, die wir vor dem Satz 4.2.6 erwähnt haben:

$$f_{X_{r:n}|X_n}(x|z) = \begin{cases} \dfrac{(n-1)!}{(r-1)!(n-1-r)!} f(x)F(x)^{r-1}(1-F(x))^{n-1-r} & \text{für } x<z,\ r<n \\ [1-F(x)]^{n-1} & \text{für } x=z,\ r=1 \\ \dfrac{(n-1)!}{(r-1)!(n-r)!} F(x)^{r-1}(1-F(x)]^{n-r} & \text{für } x=z,\ r>1 \\ \dfrac{(n-1)!}{(r-2)!(n-r)!} f(x)F(x)^{r-2}(1-F(x))^{n-r} & \text{für } x>z,\ r>1 \\ 0 & \text{sonst.} \end{cases}$$

Ein Vergleich mit der unbedingten Dichte der r-ten geordneten Statistik liefert

$$f_{X_{r:n}|X_n}(x|z) = \begin{cases} f_{X_{r:n-1}}(x) & x<z,\ r<n \\ f_{X_{r-1:n-1}}(x) & x>z,\ r>1 \end{cases}.$$

Wir betrachten nun zunächst die Situation gleichverteilter Stichprobenvariablen $U_1,\ldots,U_n$, $U_r \sim \mathscr{R}(0;1)$. Sei $1<r<n$. Dann erhalten wir, wenn wir Gebrauch machen von der bekannten Beziehung der Beta- und der Binomialverteilung

$$\int_z^1 \frac{(a+b-1)!}{(a-1)!(b-1)!} u^{a-1}(1-u)^{b-1}\,\mathrm{d}u = \sum_{i=0}^{a-1} \binom{a+b-1}{i} z^i(1-z)^{a+b-1-i}:$$

$$\begin{aligned} \mathrm{E}(U_{r:n}|U_n = z) &= \int_0^1 u \cdot f_{U_{r:n}|U_n}(u|z)\,\mathrm{d}u \\ &= \int_0^z u \cdot f_{U_{r:n}|U_n}(u|z)\,\mathrm{d}u + \int_z^1 u \cdot f_{U_{r:n}|U_n}(u|z)\,\mathrm{d}u + z\binom{n-1}{r-1} z^{r-1}(1-z)^{n-r} \end{aligned}$$

$$
\begin{aligned}
&= \frac{r}{n} - \frac{r}{n} \int_z^1 \frac{n!}{r!(n-1-r)!} u^r (1-u)^{n-1-r} \, \mathrm{d}u \\
&\quad + \frac{r-1}{n} \int_z^1 \frac{n!}{(r-1)!(n-r)!} u^{r-1} (1-u)^{n-r} \, \mathrm{d}u + \binom{n-1}{r-1} z^r (1-z)^{n-r} \\
&= \frac{r}{n} - \frac{1}{n} \sum_{i=0}^{r-1} \binom{n}{i} z^i (1-z)^{n-i}.
\end{aligned}
$$

Mit $\mathrm{E}(U_{r:n}) = r/(n+1)$ ist also:

$$
n[\mathrm{E}(U_{r:n} | U_n = z) - \mathrm{E}(U_{r:n})] = \frac{1}{n+1} r - \sum_{i=0}^{r-1} \binom{n}{i} z^i (1-z)^{n-i}.
$$

Für $r = [\alpha n] + 1$, d.h. für die Schätzung des α-Quantils, ergibt sich

$$
\sum_{i=0}^{r-1} \binom{n}{i} z^i (1-z)^{n-i} \quad \longrightarrow \quad \Phi\left(\sqrt{n} \frac{[n\alpha]/n - z}{\sqrt{z(1-z)}} \right).
$$

Für $z = \alpha$ geht das Argument von Φ, $\sqrt{n}([n\alpha]/n - z)/\sqrt{z(1-z)}$, gegen Null, für $z > \alpha$ ($z < \alpha$) gegen $-\infty$ ($+\infty$). Dies ergibt schließlich

$$
\lim_{n\to\infty} n[\mathrm{E}(U_{r:n} | U_n = z) - \mathrm{E}(U_{r:n})] = \begin{cases} \alpha - 1 & \text{für } z < \alpha \\ \alpha - 1/2 & \text{für } z = \alpha \\ \alpha & \text{für } z > \alpha \end{cases}. \tag{6.48}
$$

Sei nun X eine stetige Zufallsvariable mit der streng monotonen Verteilungsfunktion F und der stetigen Dichte f. Die geordnete Statistik $X_{r:n}$ ist dann darstellbar als Transformierte einer geordneten Statistik einer Stichprobe aus einer $\mathscr{R}(0,1)$-Verteilung:

$$
X_{r:n} = F^{-1}(U_{r:n}).
$$

Eine Taylorentwicklung um $r/(n+1)$ liefert für $1 < r < n$, wenn $G(u) = F^{-1}(u)$ gesetzt wird:

$$
\begin{aligned}
X_{r:n} \;=\; & G\left(\frac{r}{n+1} \right) + G^{(1)}\left(\frac{r}{n+1} \right) \cdot \left(U_{r:n} - \frac{r}{n+1} \right) \\
& + \frac{1}{2} G^{(2)}\left(\frac{r}{n+1} + c\left(U_{r:n} - \frac{r}{n+1} \right) \right) \cdot \left(U_{r:n} - \frac{r}{n+1} \right)^2
\end{aligned}
$$

mit $0 < c < 1$. Damit folgt:

$$
\begin{aligned}
&n[\mathrm{E}(X_{r:n}|X_n = x) - \mathrm{E}(X_{r:n})] \\
&\quad = G^{(1)}\left(\frac{r}{n+1}\right) \cdot n \cdot [\mathrm{E}(U_{r:n}|U_n = F(x)) - \mathrm{E}(U_{r:n})] \\
&\qquad + \frac{n}{2}\left\{ \mathrm{E}\left[G^{(2)}\left(\frac{r}{n+1} + c\left(U_{r:n} - \frac{r}{n+1}\right)\right) \cdot \left(U_{r:n} - \frac{r}{n+1}\right)^2 \middle| U_n = F(x)\right] \right. \\
&\qquad \left. - \mathrm{E}\left[G^{(2)}\left(\frac{r}{n+1} + c\left(U_{r:n} - \frac{r}{n+1}\right)\right) \cdot \left(U_{r:n} - \frac{r}{n+1}\right)^2 \right] \right\}.
\end{aligned}
$$

Wegen $G^{(1)}(r/(n+1)) = 1/f(F^{-1}(r/(n+1))) \to 1/f(q_\alpha)$ folgt mit den Eigenschaften der Einflussfunktion

$$
\mathrm{IF}(x, \{X_{([\alpha n]+1):n}\}, q_\alpha) = \begin{cases} \dfrac{\alpha - 1}{f(q_\alpha)} & \text{für } z < q_\alpha \\ \dfrac{\alpha - 1/2}{f(q_\alpha)} & \text{für } z = q_\alpha \\ \dfrac{\alpha}{f(q_\alpha)} & \text{für } z > q_\alpha \end{cases},
$$

sofern der zweite Term gegen null geht. Dies folgt mit einigen umständlichen Umformungen und der Ungleichung, vgl. Reiss (1989, S. 89):

$$
\mathrm{E}\left[\left|U_{r:n} - \frac{r}{n+1}\right|^j\right] \le \frac{2 \cdot j! \cdot 5^j \left[\dfrac{r}{n+1}\left(1 - \dfrac{r}{n+1}\right)\right]^{j/2}}{n^{j/2}}
$$

Analog zu den M-Schätzern erhalten wir:

$$
\begin{aligned}
\mathrm{E}[\mathrm{IF}(X, \{X_{([\alpha n]+1):n}\}, q_\alpha)^2] &= \left(\frac{\alpha - 1}{f(q_\alpha)}\right)^2 P(X < q_\alpha) + \left(\frac{\alpha - 0.5}{f(q_\alpha)}\right)^2 P(X = q_\alpha) + \left(\frac{\alpha}{f(q_\alpha)}\right)^2 P(X > q_\alpha) \\
&= \frac{\alpha(1-\alpha)}{f(q_\alpha)^2} = \lim_{n\to\infty} n \cdot \mathrm{Var}[X_{([\alpha n]+1):n}].
\end{aligned}
$$

Für einen L-Schätzer $L = \sum_{j=1}^m a_j X_{[np_j]}$ für den Parameter $\vartheta = \sum_{j=1}^m a_j F^{-1}(p_j)$ folgt mit der Additivität der Einflussfunktionen aus dem letzten Beispiel:

$$
\mathrm{IF}(x, L, \vartheta) = \sum_{j=1}^m a_j \cdot \mathrm{IF}(x, X_{[np_j]:n}, F^{-1}(p_j)). \tag{6.49}
$$

Beispiel 6.2.36 *Einflussfunktion des Quartilsabstandes*

Für den Quartilsabstand $T = X_{[n/4]:n} - X_{[3n/4]:n}$ erhalten wir mit den Einflussfunktionen für die Stichprobenquantile, wenn die zugrunde liegende Verteilung symmetrisch ist, oder zumindest für die theoretischen Quartile $q_{1/4}$ und $q_{3/4}$ die Gleichheit $f(q_{1/4}) = f(q_{3/4})$ gilt:

$$\mathrm{IF}(x, T, F^{-1}(3/4) - q_{1/4}) = \begin{cases} \dfrac{1}{2f(q_{1/4})} & \text{für } x < q_{1/4} \\ 0 & \text{für } x = q_{1/4} \\ -\dfrac{1}{2f(q_{1/4})} & \text{für } q_{1/4} < x < q_{3/4} \\ 0 & \text{für } x = q_{3/4} \\ \dfrac{1}{2f(q_{1/4})} & \text{für } x > q_{3/4} \end{cases}$$

Für L-Schätzer der Gestalt $L = \sum_{r=1}^{n} J(r/n+1)X_{r:n}/n$ unterstellen wir die Integrierbarkeit von $J(u)$. Wir gehen wieder zuerst von gleichverteilten Stichprobenvariablen aus:

$$\begin{aligned} & n[\mathrm{E}(L_n|U_n = z) - \mathrm{E}(L_n)] \\ &= \frac{1}{n}\sum_{r=2}^{n-1} J\left(\frac{r}{n+1}\right) \cdot \left(\frac{1}{n+1}r - \sum_{i=0}^{r-1}\binom{n}{i} z^i (1-z)^{n-i}\right) \\ &\quad + \frac{1}{n} J\left(\frac{1}{n+1}\right)\left(\frac{1}{n+1} + (n-1)(1-z)^n\right) + \frac{1}{n} J\left(\frac{n}{n+1}\right)\left(z^n - \frac{1}{n+1}\right) \\ &= \sum_{r=2}^{n-1} J\left(\frac{r}{n+1}\right) \cdot \frac{r}{n+1} \cdot \frac{1}{n} - \frac{1}{n}\sum_{r=2}^{n-1}\left(J\left(\frac{r}{n+1}\right)\sum_{i=0}^{r-1}\binom{n}{i} z^i (1-z)^{n-i}\right) + o(1). \end{aligned}$$

Die erste Summe ergibt für $n \to \infty$ das Riemann-Integral $\int_0^1 J(u)u\,\mathrm{d}u$. Für die zweite Summe gilt wegen

$$\sum_{i=0}^{r-1}\binom{n}{i} z^i (1-z)^{n-i} \quad \to \quad \begin{cases} 0 & \text{für } \dfrac{r}{n} < z \\ 1 & \text{für } \dfrac{r}{n} \geq z \end{cases} :$$

$$\frac{1}{n}\sum_{r=2}^{n-1}\left(J\left(\frac{r}{n+1}\right)\sum_{i=0}^{r-1}\binom{n}{i} z^i (1-z)^{n-i}\right) \quad \to \quad \int_z^1 J(u)\,\mathrm{d}u.$$

Also ist hier

$$\mathrm{IF}(z, L, \vartheta) = \int_0^1 J(u)u\,\mathrm{d}u - \int_z^1 J(u)\,\mathrm{d}u. \tag{6.50}$$

Für stetige Zufallsvariablen X mit der streng monotonen Verteilungsfunktion F und der stetigen Dichte f erhalten wir mittels der Transformation $X_{r:n} = F^{-1}(U_{r:n})$ wieder, wenn p_r so gewählt ist, dass $r/(n+1) = p_r + o(n^{-1})$:

$$\begin{aligned} X_{r:n} &= F^{-1}(p_r) + [F^{-1}(p_r + c_r(U_{r:n} - p_r))]' \cdot (U_{r:n} - p_r) \\ &= F^{-1}(p_r) + \frac{1}{f(F^{-1}(p_r + c_r(U_{r:n} - p_r)))} \cdot (U_{r:n} - p_r) \end{aligned}$$

mit $0 < c_r < 1$.

Mit der Stetigkeit von f und der Konsistenz von $U_{r:n}$ für p_r ergibt sich im Grenzwert:

$$\begin{aligned} \mathrm{IF}(x, L, \vartheta) &= \int_0^1 J(u) \frac{u}{f(F^{-1}(u))} \, d\,u - \int_{F(x)}^1 J(u) \frac{1}{f(F^{-1}(u))} \, \mathrm{d}u \\ &= \int_{-\infty}^{\infty} J(F(t)) \cdot F(t) \, \mathrm{d}t - \int_x^{\infty} J(F(t)) \, \mathrm{d}t. \end{aligned}$$

Für die M-Schätzer in einer Lage-Familie und für die Stichprobenquantile hatten wir folgenden Zusammenhang zwischen der Einflussfunktion und der Varianz des Schätzers:

$$\mathrm{E}[\mathrm{IF}(X, \hat{\vartheta}_n, \vartheta)^2] = \lim_{n \to \infty} n \cdot \mathrm{Var}(\hat{\vartheta}_n). \tag{6.51}$$

Dieser Zusammenhang gilt unter geeigneten Regularitätsvoraussetzungen für verschiedene Klassen von Schätzern.

6.3 Jackknife und Bootstrap

Das Jackknife und das Bootstrap sind zwei Verfahren der wiederholten Stichprobenziehung. Darunter verstehen wir das erneute Ziehen von Stichproben aus der vorliegenden Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)$.

Die Idee, auf (Teile der) vorhandenen Beobachtungen erneut zuzugreifen und darüber weitere Größen neben den interessierenden Parametern zu schätzen, geht auf Quenouille (1956) zurück. Er hat das folgende, später als *Jackknife* (Taschenmesser-) bezeichnete Verfahren zur Reduzierung des Bias einer Schätzfunktion entwickelt.

Als Hintergrund betrachten wir die univariate δ-Methode, vgl. Satz 5.2.2. Ist der Schätzer $\hat{\vartheta}_n$ für ϑ asymptotisch normalverteilt, $\sqrt{n}(\hat{\vartheta}_n - \vartheta) \sim \mathcal{N}(0, \sigma^2)$, so gilt dies auch für $g(\hat{\vartheta}_n)$ bzgl. $g(\vartheta)$ für geeignete Funktionen $g(t)$. Zum Beweis wird auf die Taylor-Entwicklung bis zum ersten Glied zurückgegriffen. Das ist aber nur möglich, weil für solche Schätzer die Glieder höherer Ordnung in der Taylorreihen-Entwicklung schneller als $1/n$ gegen null gehen. Die Entwicklung kann oft in der Form

$$\mathrm{E}(g(\hat{\vartheta}_n)) = g(\vartheta) + \frac{a_1(\vartheta)}{n} + \frac{a_2(\vartheta)}{n^2} + \frac{a_3(\vartheta)}{n^3} + \ldots$$

geschrieben werden, wobei die $a_i(\vartheta)$ nur von der zugrunde liegenden Verteilung abhängen. Gilt nun diese Entwicklung für endliches n, so erhalten wir für $n-1$ entsprechend

$$\mathrm{E}(g(\hat{\vartheta}_{n-1})) = g(\vartheta) + \frac{a_1(\vartheta)}{n-1} + \frac{a_2(\vartheta)}{(n-1)^2} + \frac{a_3(\vartheta)}{(n-1)^3} + \ldots$$

Zusammenfassen ergibt:

$$\begin{aligned} n \cdot \mathrm{E}(g(\hat{\vartheta}_{n-1})) - (n-1) \cdot \mathrm{E}(g(\hat{\vartheta}_{n-1})) &= g(\vartheta) + \frac{a_2(\vartheta)}{n} - \frac{a_2(\vartheta)}{(n-1)} + \frac{a_3(\vartheta)}{n^2} - \frac{a_3(\vartheta)}{(n-1)^2} + \ldots \\ &= g(\vartheta) + \frac{a_2(\vartheta)}{n(n-1)} + a_3(\vartheta)\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right) + \ldots \end{aligned}$$

Hier ist der $a_1(\vartheta)$ enthaltene Summand herausgefallen, so dass die Abweichung von $g(\vartheta)$ nur noch von der Ordnung $O(n^{-2})$ ist statt von der Ordnung $O(n^{-1})$.

Die Umsetzung dieser Betrachtung führt nun zu folgender Definition.

Definition 6.3.1 *Jackknife-Schätzer*

Sei $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ eine Stichprobe aus einer Verteilung mit der Verteilungsfunktion $F(x;\vartheta)$. Sei $\hat{g}(\vartheta) = g(X_1, \ldots, X_n)$ ein Schätzer für $g(\vartheta)$. Ein *Jackknife-Schätzer* für $g(\vartheta)$ auf der Basis von $\hat{g}(\vartheta)$ ist definiert als

$$J(\hat{g}(\vartheta)) = n \cdot \hat{g}(\vartheta) - (n-1) \cdot \bar{g}_{(\cdot)}(\vartheta), \tag{6.52}$$

wobei

$$\bar{g}_{(\cdot)}(\vartheta) = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_{(i)}(\vartheta)$$

mit $\hat{g}_{(i)}(\vartheta) = g(\boldsymbol{X}_{(i)}) = g(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$; d.h. $\hat{g}_{(i)}(\vartheta)$ ist die Schätzung von $g(\vartheta)$, die auf den $n-1$ Beobachtungen nach Ausschluss der i'ten basiert.

Wie zuvor angemerkt, reduziert das Jackknife Verfahren den Bias, bringt ihn aber i. Allg. nicht zum Verschwinden. Daher sind die beiden folgenden Beispiele eher untypisch. Der hauptsächliche Einsatzbereich liegt in komplexen Situationen, die analytisch nicht zu durchdringen sind, wie z.B. dem Quotientenschätzer bei Stichproben aus endlichen Grundgesamtheiten. Zudem setzt das Verfahren voraus, dass der Koeffizient $a_2(\vartheta)$ in der unterstellten Taylor-Reihen-Entwicklung von der gleichen Größenordnung ist wie $a_1(\vartheta)$. Sonst ist es möglich, dass die Bemühung ins Leere läuft.

Beispiel 6.3.2 *auf $\bar{X}$ basierende Jackknife-Schätzung*

Für das arithmetische Mittel gilt:

$$\begin{aligned} n \cdot \bar{X} - (n-1)\frac{1}{n}\sum_{i=1}^{n} \bar{X}_{(i)} &= n \cdot \bar{X} - (n-1)\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j\neq i} X_j\right) \\ &= n \cdot \bar{X} - \frac{1}{n}\sum_{i=1}^{n}(n-1)X_i = \bar{X}. \end{aligned}$$

Der Jackknife-Schätzer auf der Basis von $\bar{X}$ ist $\bar{X}$ selbst. Das ist insofern kein Wunder, als $\bar{X}$ ja ein unverfälschter Schätzer für μ ist.

Beispiel 6.3.3 *auf S^2 basierende Jackknife-Schätzung*

Sei S^2 die empirische Varianz, $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$. Bekanntlich ist $\mathrm{E}(S^2) = (n-1)\sigma^2/n$, wenn σ^2 die Varianz der zugrunde liegenden Verteilung ist. Somit ist der Bias von S^2 gleich $-\sigma^2/n$. Für den auf S^2 basierenden Jackknife-Schätzer erhalten wir:

$$S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n X_i X_j.$$

Weiter ist:

$$\begin{aligned}
S_{(i)}^2 &= \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n (X_j - \bar{X}_i)^2 = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n X_j^2 - \left(\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n X_j\right)^2 \\
&= \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n X_j^2 - \left(\frac{1}{n-1}\right)^2 \sum_{\substack{j=1\\ j\neq i}}^n\sum_{\substack{k=1\\ k\neq i}}^n X_j X_k, \\
\frac{1}{n}\sum_{i=1}^n S_{(i)}^2 &= \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n X_j^2\right) - \frac{1}{n}\sum_{i=1}^n \left(\left(\frac{1}{n-1}\right)^2 \sum_{\substack{j=1\\ j\neq i}}^n\sum_{\substack{k=1\\ k\neq i}}^n X_j X_k\right) \\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{n-1}{n(n-1)^2}\sum_{i=1}^n X_i^2 - \frac{n-2}{n(n-1)^2}\sum_{i=1}^n\sum_{\substack{j=1\\ j\neq i}}^n X_i X_j \\
&= \frac{n-2}{(n-1)^2}\sum_{i=1}^n X_i^2 - \frac{n-2}{n(n-1)^2}\sum_{i=1}^n\sum_{j=1}^n X_i X_j.
\end{aligned}$$

Zusammenfassen ergibt:

$$\begin{aligned}
J(S^2) &= \left(\sum_{i=1}^n X_i^2 - \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n X_i X_j\right) - \left(\frac{n-2}{(n-1)}\sum_{i=1}^n X_i^2 - \frac{n-2}{n(n-1)}\sum_{i=1}^n\sum_{j=1}^n X_i X_j\right) \\
&= \frac{1}{n-1}\sum_{i=1}^n X_i^2 - \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j=1}^n X_i X_j = \frac{n}{n-1}S^2.
\end{aligned}$$

Auch hier ist das Ergebnis ein unverfälschter Schätzer. Da der Bias von S^2 von der Ordnung $O(n^{-1})$ ist, aber keine Glieder höherer Ordnung vorkommen, wird er vollständig eliminiert.
Wollen wir aber die Standardabweichung σ schätzen und nehmen dazu $\sqrt{\hat{\sigma}^2}$, dann ist das Jackknife tatsächlich von Bedeutung, da der Schätzer $\sqrt{\hat{\sigma}^2}$ verfälscht ist und der Bias nicht so elementar angebbar.

Das *Bootstrap-Verfahren* von Efron hat als Hintergrund, dass die Simulation von den verschiedensten Verteilungen heutzutage sehr einfach ist und damit computermäßig Probleme bearbeitet werden können, die für eine analytische Behandlung zu komplex sind. Ist es z.B. nicht möglich, den Standardfehler eines Schätzers analytisch anzugeben, so lässt sich einfach künstlich über die Erzeugung vieler Stichproben gleichen Umfanges eine empirische Verteilung des Schätzers gewinnen. Diese kann als Näherung der theoretischen angesehen werden. Somit gibt die empirische Standardabweichung der Bootstrap-Verteilung eine Schätzung der theoretischen.

Ist die zugrunde liegende Verteilung bis auf einen gegebenenfalls mehrdimensionalen Parameter bekannt, so können unter Verwendung des geschätzten Parameters die Bootstrap-Stichproben aus dieser Verteilung simuliert werden. Dies ist das *parametrische Bootstrap*. Ist die zugrunde liegende Verteilung nicht bekannt, so verwendet *das nichtparametrische Bootstrap* die empirische Verteilungsfunktion, um die Stichproben zu erzeugen. Die empirische Verteilungsfunktion $\hat{F}_n(x)$ wird ja oft als nichtparametrischer Maximum-Likelihood-Schätzer der theoretischen Verteilungsfunktion bezeichnet, da sie den Ausdruck

$$L(F) = \prod_{i=1}^{n} (F(x_i) - F(x_i - 0)), \tag{6.53}$$

mit $F(x-0) = \lim_{t \uparrow x} F(t)$, über alle Verteilungsfunktionen F maximiert. Zudem ist die empirische Verteilungsfunktion ein konsistenter Schätzer für die theoretische, so dass sich $\hat{F}_n(x)$ von $F(x)$ für große n nur wenig unterscheiden sollte.

Definition 6.3.4 *Bootstrap-Algorithmus*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine Stichprobe aus einer Verteilung mit der Verteilungsfunktion $F(x;\vartheta)$. $\hat{\vartheta}(\boldsymbol{X})$ sei ein Schätzer für ϑ. Der Bootstrap-Algorithmus zur Schätzung von Verteilungsparametern der Schätzfunktion $\hat{\vartheta}(\boldsymbol{X})$ geht aus von einer beobachteten Stichprobe $\boldsymbol{x}$ und ist dabei durch folgende Schritte festgelegt:

1) Bestimme die empirische Verteilungsfunktion $\hat{F}_n(x)$ zu der Stichprobe $\boldsymbol{x}$:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(x_i \leq x]}.$$

2) Für $b = 1, \ldots, N$: Ziehe eine Stichprobe $\boldsymbol{x}_b^* = (x_{b1}^*, \ldots, x_{bn}^*)$ vom Umfang n (mit Zurücklegen) aus der empirischen Verteilungsfunktion und bestimme $\hat{\vartheta}(\boldsymbol{x}_b^*)$.

3) Schätze den interessierenden Verteilungsparameter von $\hat{\vartheta}(\boldsymbol{X})$ mittels eines geeigneten Schätzers aus den N Werten $\hat{\vartheta}(\boldsymbol{x}_b^*)$.

Auch wenn der Wert des Bootstrap vor allem in der Anwendbarkeit auf endliche Stichproben liegt, ist die theoretische Absicherung asymptotischer Natur. Wir wollen kurz den Rahmen dafür angeben. Gegeben ist die Situation der Definition. Zudem seien $X_1^*, \ldots, X_n^*$ Stichprobenvariablen aus der empirischen Verteilungsfunktion $\hat{F}_n(x)$. $X_1^*, \ldots, X_n^*$ sind also bei gegebenem $\boldsymbol{X} = \boldsymbol{x}$ unabhängig.

Wir beobachten nun B solcher Stichproben $\boldsymbol{X}_b^*$ und die sich daraus ergebenden Stichprobenfunktionen $\hat{\vartheta}_{nb}^* = \hat{\vartheta}(\boldsymbol{X}_b^*)$ $(b = 1, \ldots, B)$. Seien nun für $t \in \mathbb{R}$

$$\hat{F}_{nB}^*(t) = \frac{1}{B} \sum_{b=1}^{B} I_{(\sqrt{n}(\hat{\vartheta}_{nb}^* - \hat{\vartheta}_n) \le t]} \quad \text{und} \quad F_{\hat{\vartheta}_n}(t) = \mathrm{P}(\sqrt{n}(\hat{\vartheta}_n - \vartheta) \le t)$$

sowie

$$v_{nB}^2 = \frac{1}{B} \sum_{b=1}^{B} n(\hat{\vartheta}_{nb}^* - \hat{\vartheta}_n)^2.$$

Dann gilt unter geeigneten Regularitätsvoraussetzungen für große B (und n):

$$\sup_{t \in \mathbb{R}} |\hat{F}_{nB}^*(t) - F_{\hat{\vartheta}_n}(t)| \quad \xrightarrow{P} \quad 0, \tag{6.54}$$

und für den Bootstrap-Varianz-Schätzer

$$|v_{nB}^2 - \mathrm{E}(n(\hat{\vartheta}_n - \vartheta)^2)| \quad \xrightarrow{P} \quad 0. \tag{6.55}$$

Es ist ziemlich leicht zu akzeptieren, dass durch eine sehr große Wahl der Anzahl B der Bootstrapstichproben die Bootstrap-Varianz die bedingte Varianz $\mathrm{E}((\hat{\vartheta}_{n1}^* - \hat{\vartheta}_n)^2 | X_1 = x_1, \ldots, X_n = x_n)$ mit $\hat{\vartheta}_{n1}^* = \hat{\vartheta}(X_1^*, \ldots, X_n^*)$ hinreichend genau approximiert wird. Dazu ist nur die Konsistenz noch einmal genau zu überdenken. Dass dann weiterhin auch (6.55) gilt, stellt schon höhere Anforderungen an die Vorstellungskraft. Deshalb sei angemerkt, dass für die bedingte Erwartung der Varianz einer Bootstrap-Stichprobe gilt:

$$\mathrm{E}\left(\frac{1}{n} \sum_{j=1}^{n} (X_j^* - \bar{X}_n)^2 \,\middle|\, X_1, \ldots, X_n \right) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = S_n^2 \quad \xrightarrow{f.s} \quad \sigma^2.$$

Hier wurde das arithmetische Mittel $\bar{X}_n = \sum_{i=1}^n X_i / n$ der ursprünglichen Stichprobe, das dem Erwartungswert der X_j^* entpricht, als Lageparameter genommen. Dies ist generell bei der Bestimmung von Standardfehlern von Schätzern mit dem Bootstrap anzuraten. Sonst wird dieser i.d.R. unterschätzt. Dies ist derselbe Effekt, den wir von der empirischen Varianz kennen. Die beobachteten Werte sind enger um das arithmetische Mittel als um den theoretischen Erwartungswert konzentriert.

Bickel & Freedman (1981) zeigen für verschiedene Schätzer, dass die Bootstrap-Versionen asymptotisch normalverteilt sind. Wir betrachten nur das arithmetische Mittel in einer einfach zu behandelnden Situation.

Satz 6.3.5 *asymptotische Veteilung des Bootstrap-Mittels*

Sei X eine Zufallsvariable mit Erwartungswert μ, Varianz σ^2 und momenterzeugender Funktion $M(t)$. $X_1, X_2, \ldots, X_n, \ldots$ sei eine Stichprobe aus der Verteilung von X. Mit den oben eingeführten Bezeichnungen $\bar{X}_n$, S_n^2 und mit $\bar{X}_n^* = \sum_{j=1}^n X_j^* / n$ gilt:

$$\lim_{n\to\infty} \mathrm{P}\left(\sqrt{n}\frac{\bar{X}_n^* - \bar{X}_n}{\sigma} \le z \,\middle|\, X_1, X_2, \ldots, X_n \right) = \Phi(z).$$

Beweis: Wir wählen h so, dass $M(t)$ für $|t| < 8h$ existiert. Für die bedingte momenterzeugende Funktion von $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ gilt:

$$\begin{aligned}
\mathrm{E}(\exp[t\sqrt{n}(\bar{X}_n^* - \bar{X}_n)] | X_1, X_2, \ldots, X_n) &= \mathrm{E}\left(\exp\left[t \sum_{j=1}^{n} \left(\frac{X_j^*}{\sqrt{n}} - \frac{\bar{X}_n}{\sqrt{n}} \right) \right] \,\middle|\, X_1, X_2, \ldots, X_n \right) \\
&= \left(\frac{1}{n} \sum_{j=1}^{n} \exp\left[t \cdot \frac{X_j - \bar{X}_n}{\sqrt{n}} \right] \right)^n = \left(1 + \frac{1}{2} \cdot \frac{1}{n} \sum_{j=1}^{n} \frac{t^2 (X_j - \bar{X}_n)^2}{n} + R_n \right)^n \qquad (6.56) \\
&= \left(1 + \frac{t^2 S^2 + n R_n}{2n} \right)^n .
\end{aligned}$$

Nach der Abschätzung im Beweis von Satz 2.2.6 haben wir für R_n:

$$\begin{aligned}
|R_n| &< \left(\frac{|t|}{h} \right)^3 \frac{1}{n^{3/2}} \cdot \frac{1}{n} \sum_{j=1}^{n} \left(\exp[2h(X_i - \bar{X}_n)] + \exp[-2h(X_i - \bar{X}_n)] \right) \\
&= \left(\frac{|t|}{h} \right)^3 \frac{1}{n^{3/2}} \left(e^{2h\bar{X}_n} \frac{1}{n} \sum_{j=1}^{n} \exp[2hX_i] + e^{-2h\bar{X}_n} \frac{1}{n} \sum_{j=1}^{n} \exp[-2hX_i] \right) .
\end{aligned} \qquad (6.57)$$

Nun gilt $e^{2h\bar{X}_n} \overset{f.s.}{\to} e^{2h\mu}$, $e^{-2h\bar{X}_n} \overset{f.s.}{\to} e^{-2h\mu}$. Wir haben also noch zu zeigen, dass

$$\frac{1}{n} \sum_{j=1}^{n} \exp[\pm 2hX_i] \quad \overset{f.s.}{\longrightarrow} \quad M(\pm 2h).$$

Dann ist nämlich der Ausdruck in der Klammer von (6.57) beschränkt, so dass die rechte Seite von (6.57) auch nach Multiplikation mit n gegen null geht. Somit geht auch $n \cdot R_n$ fast sicher gegen null, und die rechte Seite von (6.56) konvergiert gegen $\exp[1 + \sigma^2 t^2/2]$, woraus sich die Behauptung dann unmittelbar ergibt.

Seien $Y_j = \exp[2hX_j] - M(2h)$ und $A_n = \{|\bar{Y}_n| > \epsilon\}$. Dann gilt aufgrund der Markoffschen Ungleichung

$$\mathrm{P}(A_n) \le \frac{1}{\epsilon^4} \mathrm{E}(\bar{Y}_n^4),$$

und wie im Beweis des starken Gesetzes der großen Zahlen folgt:

$$\begin{aligned}
\mathrm{P}(A_n) &\le \frac{1}{\epsilon^4} \cdot \frac{1}{n^4} \left\{ n \cdot \mathrm{E}\left[(\exp[2hX_1] - M(2h))^4 \right] + 3n(n-1) \left(\mathrm{E}[\exp[2hX_1] - M(2h)]^2 \right)^2 \right\} \\
&= \frac{1}{\epsilon^4} \cdot \frac{1}{n^3} \left\{ M(8h) - 4MM(6h)M(2h) + 6M(4h)M(2h)^2 - 3M(2h)^4 \right. \\
&\qquad \left. + 3(n-1)(M(4h) - M(2h))^2 \right\} .
\end{aligned}$$

Also gilt

$$\mathrm{P}\left(\bigcup_{n\geq 1}\bigcap_{m\geq n} A_m\right) = 0.$$

Damit ist gezeigt, dass

$$\frac{1}{n}\sum_{j=1}^{n}\exp[2hX_i] \quad \xrightarrow{f.s.} \quad M(2h).$$

Analog folgt dies für $-h$. Zusammen ergibt das die Behauptung.

6.4 Auswahl von Schätzern für die Anwendung

Die Eigenschaften von Schätzern legen die Anwendung von ML- bzw. von robusten Schätzern nahe. Robuste Schätzer sind vorzuziehen, wenn es um Formparameter von Verteilungen wie Lage und Streuung geht. Denn oft ist die zugrunde liegende Verteilung gar nicht genau spezifiziert, so dass ML-Schätzer von vornherein dubios sind.

Ein anderer Punkt ist, dass die Regularitätsbedingungen, die zum Nachweis der guten Eigenschaften der ML-Schätzer benötigt werden, nicht vorzuliegen brauchen. Wir geben hier nur ein Beispiel. Weitere finden sich bei Le Cam (1990).

Beispiel 6.4.1 *Varianzschätzung bei Beobachtungspaaren mit unabhängigen Komponenten*

$(X_i, Y_i)\,(i = 1,\ldots,n)$ seien unabhängige Zufallsvektoren, wobei X_i und Y_i ebenfalls unabhängig und $\mathcal{N}(\mu_i,\sigma^2)$ -verteilt seien. Wir wollen σ^2 schätzen. Ein naheliegender Ansatz ist, $Z_i = X_i - Y_i$ zu bilden, um die störenden Parameter μ_i zu eliminieren. Es gilt offensichtlich $Z_i \sim \mathcal{N}(0, 2\sigma^2)$. Daher ist

$$S^2 = \frac{1}{2n}\sum_{i=1}^{n} Z_i^2$$

ein naheliegender Schätzer.
Der ML-Ansatz

$$-n(\ln(2\pi)+\ln(\sigma^2)) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_i)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu_i)^2 \stackrel{!}{=} \max$$

führt dagegen zu

$$\hat{\sigma}^2 = \frac{1}{2}S^2.$$

Das ist offensichtlich vollständig unplausibel. Der Grund ist darin zu sehen, dass hier zu viele Parameter geschätzt werden müssen.

Die verschiedenen Beispiele führen Le Cam zu den folgenden Prinzipien, deren Beherzigung für die Anwendung wärmstens empfohlen wird.

Prinzip 1: Vertraue keinem Schätzprinzip.

Prinzip 2: Mach Dir klar, was Du schätzen möchtest.

Prinzip 3: Versuche festzulegen, welchen Grad an Präzision Du brauchst (oder erreichen kannst) und was Du mit der Schätzung vorhast.

Prinzip 4: Vor dem Einsatz eines Schätzers überprüfe den Grund für seine Auswahl und die Anwendbarkeit auf die vorliegenden Daten.

Prinzip 5: Wenn bis jetzt alles in Ordnung ist, versuche einen kruden aber zuverlässigen Schätzer, um zu bestimmen, in welcher Region der Parameter liegt.

Prinzip 6: Wenn Prinzip 4 erfüllt ist, verbessere Deine Schätzung durch Ausnutzen Deiner theoretischen Voraussetzungen.

Prinzip 7: Vertraue keiner Schätzung, die ein völlig anderes Ergebnis ergibt, wenn nur eine Beobachtung aus dem Datensatz herausgenommen wird.

Prinzip 8: Wenn ein asymptotisches Argument herangezogen wird, vergiss nicht, die Anzahl der Beobachtungen gegen unendlich gehen zu lassen.

Prinzip 9: J.Bertrand formulierte es so: 'Gib mir vier Parameter und ich beschreibe einen Elefanten; mit fünfen schwenkt er den Rüssel'.

Das Prinzip 4 bringt die robusten Schätzer ins Spiel. Le Cam hat von diesem Prinzip aber schon 1946 gehört. Angewandte Statistiker waren also schon immer vorsichtige Menschen.

6.5 Aufgaben

Aufgabe 1

Ein Buch von 800 Seiten soll auf die Verteilung der Druckfehler hin untersucht werden. Dazu sollen 10 Seiten durchgegangen werden.

i) Welche Seiten wählt man aus, wenn man aus einer Zufallszahlentabelle die folgende Sequenz auswählt:
03 47 43 73 86 36 96 47 36 61 46 98 63 71 62 16 80 45 ?

ii) Schätzen Sie die durchschnittliche Fehlerzahl pro Seite und geben Sie eine Schätzung des Standardfehlers dieser Schätzung an.
(Der Standardfehler eines Schätzers ist seine Standardabweichung.)

iii) Geben Sie Schätzungen der Gesamtfehlerzahl und des zugehörigen Standardfehlers an.

iv) Schätzen Sie den Anteil der Seiten mit Fehlern in dem Buch und geben Sie ein Konfidenzintervall zum Niveau 0.75 für diesen Anteil an.

Aufgabe 2

$X_1,\ldots,X_n$ sei eine Stichprobe aus einer Verteilung mit $E(X)=\mu$ und $Var(X)<\infty$. Bestimmen Sie die BLU-Schätzfunktion für μ, d.h. die Koeffizienten $a_1,\ldots,a_n$, so dass $\hat{\mu}=a_1X_1+\ldots+a_nX_n$ erwartungstreu ist und die kleinste Varianz unter allen derartigen Schätzern hat.

Aufgabe 3

Gegeben sei eine Grundgesamtheit $\{\omega_1,\ldots,\omega_N\}$ vom Umfang $N=40$ mit den folgenden Werten $X(\omega_j)$ der Zufallsvariablen X:

0 1 1 2 5 4 7 7 8 6 6 8 9 10 13 12 15 16 16 17 18 19 20 20 24 23 25 28 29 27 26
30 31 31 33 32 35 37 38 38.

Aus dieser Grundgesamtheit wird eine systematische Stichprobe vom Umfang $n=4$ und der Schrittweite $k=10$ gezogen. Bei dieser Ziehungsart wird von einer seriellen Anordnung der Elemente ausgegangen. Es wird eine zufällige Startzahl r $(1\le r\le k)$ bestimmt und dann wird mit r auch $k+r, 2k+r, 3k+r,\ldots,(n-1)k+r$ gezogen.

Berechnen Sie Erwartungswert und Varianz von $\bar{X}$ bei dem durch diese Ziehungsart auf $\Omega^{(n)}$ induzierten Wahrscheinlichkeitsmaß.

Aufgabe 4

In einer Population von N Elementen ist der Wert x_1 der Variablen X für das Element ω_1 bekannt. Aus den übrigen $N-1$ Einheiten wird eine einfache Zufallsauswahl mit Zurücklegen vom Umfang n gezogen und hieraus der Mittelwert $\bar{x}^*$ bestimmt.

i) Bestimmen Sie den Erwartungswert und die Varianz des Schätzers

$$\tilde{\mu}=\frac{1}{N}[x_1+(N-1)\bar{X}^*].$$

ii) Zeigen Sie, dass dieser Schätzer von μ besser ist als der Schätzer $\bar{X}_n$, der auf der Basis einer einfachen Zufallsauswahl mit Zurücklegen aus allen N Einheiten ermittelt wird.

Aufgabe 5

X sei exponentialverteilt mit dem Parameter $\lambda>0$, $f(x)=\lambda\,e^{-\lambda x}$, $x>0$. Zeigen Sie, dass $\hat{\lambda}=1/\bar{X}$ schwach konsistent für λ ist.

Aufgabe 6

X sei stetig verteilt mit der Dichte $f(x)$. Dann hat der Stichprobenmedian $\tilde{X}$ für große n approximativ eine Normalverteilung mit Erwartungswert $\tilde{\mu}$ $(P(X<\tilde{\mu})=0.5$ und Varianz $(n\,4f(\tilde{\mu})^2)^{-1}$. Bestimmen Sie die relative Effizienz von $\tilde{X}$ bzgl. $\bar{X}$ für die Fälle:

(i) $f(x) = 0.5e^{-|x-\mu|}$, d. h. X ist Laplace-verteilt.

(ii) X ist normalverteilt mit festem σ^2.

Aufgabe 7

(X, Y) sei eine bivariate Zufallsvariable mit der Dichte $f_\vartheta(x,y) = y^2 e^{-y(x+\vartheta)}, x, y \geq 0$.

Welches wäre ein adäquater Ansatz für eine Schätzung von ϑ nach der Methode der Kleinsten Quadrate?

Aufgabe 8

Eine (vereinfachte Stichprobe) aus den Versicherungsschäden eines bestimmten Schadentypes ergab folgende Beobachtungen:

x_i	750	1500	2500	3500	4500	5500
n_i	8	5	4	1	1	1

.

Es wird unterstellt, dass die Stichprobe aus einer Pareto-Verteilung stammt:

$$f_\vartheta(x) = \vartheta\, 600^\vartheta\, x^{-\vartheta+1} \text{für } x > 600, \quad f_\vartheta(x) = 0 \text{ sonst.}$$

Dabei ist $\vartheta > 0$.

(i) Zeichnen Sie die Loglikelihoodfunktion für den Bereich $0 < \vartheta < 4$.

(ii) Bestimmen Sie den ML-Schätzwert.

Aufgabe 9

Es wird ein Gen mit zwei Allelen betrachtet. Drei Genotypen $i = 1,2,3$ sind möglich, die entsprechend dem sogenannten Hardy-Weinberg-Gesetz vorkommen:

$$p_1 = \vartheta^2, \quad p_2 = 2\vartheta(1-\vartheta), \quad p_3 = (1-\vartheta)^2.$$

Dabei ist $0 < \vartheta < 1$. Bei einer Stichprobe vom Umfang n werden N_i Personen des Genotyps i beobachtet, $i = 1,2,3$.
Bestimmen Sie den ML-Schätzer für ϑ!
Ist er konsistent? erwartungstreu? effizient?
(Hinweis: (N_1, N_2) ist trinomialverteilt.)

Aufgabe 10

$x_1 = 0.5$, $x_2 = -2.0$ sei eine Stichprobe aus einer Cauchy-Verteilung mit der Dichte

$$f(x;\mu) = \pi^{-1}[1 + (x-\mu)]^{-1}.$$

Zeichnen Sie die Likelihoodfunktion im Intervall $[-2.5; 1.0]$.

Welche Forderungen ergeben sich daraus für die ML-Schätzung?

Aufgabe 11

Die Zufallsvariable X habe die vom Parameter ϑ, $1 < \vartheta < 2.5$, abhängige Wahrscheinlichkeitsfunktion

$$f_\vartheta(x) = \frac{1}{3}(5-2\vartheta)^{(3-x)(2-x)/2}\,(\vartheta-1)^{(x-1)(x-4)/2} \quad \text{für } x = 1,2,3$$

und $f_\vartheta(x) = 0$ sonst.

Bestimmen Sie den ML-Schätzer für ϑ! Ist er konsistent? erwartungstreu? effizient?

Aufgabe 12

Bei Umfragen ist es oft schwierig, richtige Antworten auf sensible Fragen zu erhalten. – Man denke etwa an eine Frage der Art: ‚Haben Sie jemals Rauschgift genommen?' – Warner hat 1965 daher die Methode der randomisierten Antwort eingeführt, um solche Situationen zu bewältigen. Der Frager weiß bei dieser Methode nicht, auf welche der beiden einfach alternativen Fragen der Befragte antwortet. Das wird erreicht, indem der Befragte (ohne dass der Frager es sehen kann) zufällig eine Kugel aus einer Urne zieht, welche Kugeln mit zwei unterschiedlichen Farben enthält. Die eine Farbe repräsentiert ‚Ich habe die Eigenschaft A', die andere repräsentiert ‚Ich habe die Eigenschaft A nicht'. Auf dieses Statement antwortet der Befragte dann mit ja bzw. nein.
Sei

- R der Anteil der Befragten, die mit ‚ja' antworten,
- p die Chance, dass eine Kugel mit der Farbe, welche ‚Ich habe die Eigenschaft A' repräsentiert, gezogen wird,
- q der Anteil der Personen mit der Eigenschaft A in der Grundgesamtheit,
- r die Wahrscheinlichkeit, dass der Befragte mit ‚ja' antwortet.

Es gilt: $r = (2p-1)q + (1-p)$.
Hinweis: P(‚ja')|Frage1)P(Frage2) + P(‚ja'|Frage2)P(Frage2) sowie $\mathrm{E}(R) = r$ $\mathrm{Var}(R) = r(n-1)/n$ (unter Vernachlässigung des Faktors für das Ziehen ohne Zurücklegen).

Geben Sie eine Schätzfunktion Q für q an; weisen sie die Unverfälschtheit von Q nach.

Aufgabe 13

507 Personen wurden bzgl. ihrer Einstellung zum Cholesterin befragt. (Die Woche vom 19.8.93)

	gefährlich	nicht gefährtlich
Frauen	165	85
Männer	149	108

Der Anteil der Männer in der Grundgesamtheit ist dabei 48%.

Schätzen Sie den Anteil aller Personen, die das Cholesterin für gefährlich halten. Welche Verbesserung ergibt sich schätzungsweise, wenn das Ergebnis als Resultat einer vorher geplanten Schichtung angesehen wird? (Wegen der Größe der Grundgesamtheit darf von einer Zufallsauswahl mit Zurücklegen ausgegangen werden.)

Aufgabe 14

Eine Erhebung des StaLa bei 10 000 landwirtschaftlichen Betrieben eines Bundeslandes hat im Frühjahr ergeben, dass insgesamt 80 000 ha mit Weizen bestellt waren. Um sich unmittelbar nach der Ernte einen Überblick über die Erntemenge zu verschaffen, wählt das Amt 200 Betriebe zufällig aus und erfragt die geerntete Weizenmenge X sowie die Weizenanbaufläche Y. Es ergibt sich (X in t, Y in ha):

$$\sum x_i = 20\,000 \quad \sum x_i^2 = 4\,960\,000 \quad \sum y_i = 2\,000$$

$$\sum y_i^2 = 50\,000 \quad \sum x_i y_i = 490\,000.$$

Berechnen Sie zwei Schätzungen für die Ernte; vergleichen Sie soweit wie möglich.

Aufgabe 15

$X_1, \ldots, X_n$ sei eine Stichprobe zu X, wobei $E(X) = \mu$ und $Var(X) < \infty$. Bestimmen Sie die BLU-Schätzfunktion für μ, d.h. die Koeffizienten $a_1, \ldots, a_n$, so dass $\hat{\mu} = a_1 X_1 + \ldots + a_n X_n$ erwartungstreu ist und die kleinste Varianz unter allen derartigen Schätzern hat.

Aufgabe 16

Ein verliebter Student muss oft zu lange auf den Anruf seiner Freundin warten. In den Wartezeiten hat er sich naturgemäß seine Zeit damit vertrieben, eine Wahrscheinlichkeitsmodell für die Wartezeiten aufzustellen. Er kam zu folgender Wahrscheinlichkeitsfunktion für die Wartezeit X (in Tagen):

x	0	1	2	3	4
$f(x;\vartheta)$	0.1ϑ	0.2ϑ	0.3ϑ	0.4ϑ	$1-\vartheta$

Helfen Sie ihm nun, für den unbekannten Parameter ϑ, $0 < \vartheta < 1$, eine geeignete Schätzfunktion zu finden!

(i) Bestimmen Sie a und b so, dass $\hat{\vartheta}_1 = a + b\bar{X}$ erwartungstreu ist. Ist $\hat{\vartheta}$ dann konsistent?

(ii) Untersuchen Sie die Schätzfunktionen $\hat{\vartheta}_2 = 10\,h(X=0)$, $\hat{\vartheta}_3 = 2.5\,h(X=3)$ und $\hat{\vartheta}_4 = 1 - h(X=4)$ auf Erwartungstreue und Konsistenz.

(iii) Bestimmen Sie die relativen Effizienzen der Schätzfunktionen. Ist eine von ihnen effizient?
(Hinweis: Satz von Rao-Cramér)

Aufgabe 17

X sei exponentialverteilt mit dem Parameter $\lambda > 0$, $f(x) = \lambda e^{-\lambda x}$, $x > 0$. Zeigen Sie, dass $\hat{\lambda} = 1/\bar{X}$ schwach konsistent für λ ist.

Aufgabe 18

Zeigen Sie, dass bei der Gleichverteilung $\mathscr{R}(0,\vartheta)$ der Schätzer $X_{(n)} = \max\{X_1,\ldots,X_n\}$ stark konsistent für ϑ ist.

Aufgabe 19

X_1, X_2 seien geometrisch verteilt, $f(x) = p(1-p)^x$, $x = 0,1,2,\ldots$ Bestimmen Sie mit Hilfe des Satzes von Rao-Blackwell eine erwartungstreue Schätzfunktion für p, die besser ist als $T(X_1,X_2) = 1$ falls $X_1 = 0$ und $T(X_1,X_2) = 0$ falls $X_1 > 0$.

Können Sie etwas über die Effizienz Ihrer Schätzfunktion aussagen?

Aufgabe 20

$X_1, X_2 \ldots$ sei eine Stichprobe aus einer $\mathscr{R}(0,\vartheta)$-Verteilung. Zeigen Sie, dass die auf $X_{(n)}$ basierende erwartungstreue Schätzfunktion UMVUE ist.

Aufgabe 21

Bestimmen Sie die Einflussfunktion der empirischen Varianz $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Aufgabe 22

Eine Zufallsvariable X heißt logarithmisch normalverteilt mit den Parametern μ_L und σ_L^2, wenn $Y = \ln(X)$ normalverteilt ist, also $Y \sim N(\mu_L, \sigma_L^2)$.
(Die logarithmische Normalverteilung dient u.a. als Modell für Einkommensverteilungen, für die Verteilung der Größen von Steinen, die in einem Steinbruch abgebaut werden und für die Verteilungen der Biegefestigkeiten von Fasern.)
Bestimmen Sie die ML-Schätzfunktionen für μ_L und σ_L^2.

7 Konfidenzschätzung

7.1 Grundlagen

Die Punktschätzung hat das Problem, dass bei einer stetigen Verteilung die Zufallsvariable jeden Wert x mit Wahrscheinlichkeit null annimmt: $P(X = x) = 0$ für alle $x \in \mathbb{R}$. Daher ist nur sicher, dass der Punktschätzer den wahren Parameterwert *nicht* trifft. Es ist in Anwendungen folglich nicht ausreichend, wenn das Ergebnis einer Punktschätzung mitgeteilt wird. I.d.R. wird zusätzlich eine Angabe über die Präzision oder den möglichen Fehler der Schätzung gewünscht. Oft findet man in der Praxis den Schätzer zusammen mit einer Schätzung $\hat{\sigma}(\hat{\vartheta})$ seines Standardfehlers angegeben. Wir können uns $\hat{\vartheta} \pm \hat{\sigma}(\hat{\vartheta})$ als eine Art Grenzen für ϑ vorstellen. Jedoch bereitet die Interpretation dieser Grenzen Schwierigkeiten. Wir haben keine Gewähr, dass $\hat{\vartheta} \pm \hat{\sigma}(\hat{\vartheta})$ den wahren Parameterwert tatsächlich enthält. Und Wahrscheinlichkeiten sind bei konkreten Werten wie z.B. 10 ± 2 nicht mehr angebracht. Auch auf der Ebene der Schätzfunktionen ist i.d.R. die Wahrscheinlichkeit kleiner als eins, dass ϑ von solchen Grenzen eingeschlossen wird:

$$P(\hat{\vartheta} - \hat{\sigma}(\hat{\vartheta}) < \vartheta < \hat{\vartheta} + \hat{\sigma}(\hat{\vartheta})) < 1.$$

Beispiel 7.1.1 *Grenzen für μ bei Normalverteilung*

Sei $X_1, \ldots, X_n$ eine Zufallsstichprobe aus einer $\mathcal{N}(\mu, \sigma^2)$-Verteilung. σ^2 sei bekannt. Dann gilt für den Schätzer $\bar{X}$ von μ:

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}}\right) = 0.6827.$$

Auch wenn wir die Grenzen weiter ziehen, etwa ein Mehrfaches der Standardabweichung als Grenzen nehmen, bleibt die Wahrscheinlichkeit kleiner als eins, dass μ in den Grenzen $\bar{X} \pm k\sigma/\sqrt{n}$ liegen wird. Es bleibt daher als Ziel, nach der Ermittlung der Grenzen wenigstens mit einem einigermaßen sicheren Gefühl behaupten zu können, der unbekannte wahre Parameterwert liege dazwischen. Daher werden wir die Grenzen so bestimmen, dass sie den unbekannten Wert mit einer geeignet großen Wahrscheinlichkeit einschließen. Dazu müssen wir natürlich die Diskussion auf der Ebene von Zufallsvariablen oder konkreter auf der von Stichprobenfunktionen führen.

Definition 7.1.2 *Konfidenzintervall und Konfidenzschranke*

Gegeben seien die Stichprobe $\boldsymbol{X} = (X_1, \ldots, X_n)$ aus einer Verteilung mit der Dichtefunktion $f(x; \vartheta)$ und zwei Statistiken $T_1(\boldsymbol{X})$ und $T_2(\boldsymbol{X})$. Für die Statistiken gelte für alle $\vartheta \in \Theta$:

(1) $P_\vartheta(T_1(\boldsymbol{X}) \le T_2(\boldsymbol{X})) = 1$,

(2) $P_\vartheta(T_1(\boldsymbol{X}) \le g(\vartheta) \le T_2(\boldsymbol{X})) \ge 1-\alpha \qquad (0<\alpha<1)$.

Dabei ist α unabhängig von ϑ und $g(\vartheta)$ ist eine Funktion des Parameters.

Dann heißt $\{g(\vartheta)|T_1(\boldsymbol{X}) \le g(\vartheta) \le T_2(\boldsymbol{X})\}$, oder kurz $[T_1(\boldsymbol{X}), T_2(\boldsymbol{X})]$, ein $(1-\alpha)$-*Konfidenzintervall* oder ein Konfidenzintervall zum Niveau $(1-\alpha)$ oder $(1-\alpha)100\%$ für $g(\vartheta)$. Das sich durch die beobachtete Stichprobe $\boldsymbol{x}$ ergebende Intervall $[T_1(\boldsymbol{x}), T_2(\boldsymbol{x})]$ nennen wir *realisiertes Konfidenzintervall*.

Im Fall $T_1(\boldsymbol{X}) = -\infty$ heißt $T_2(\boldsymbol{X})$ eine *obere Konfidenzschranke*; $T_1(\boldsymbol{X})$ heißt *untere Konfidenzschranke*, falls $T_2(\boldsymbol{X}) = \infty$.

Aus der Beziehung $P_\vartheta(T_1(\boldsymbol{X}) \le g(\vartheta) \le T_2(\boldsymbol{X})) \ge 1-\alpha$ darf nicht gefolgert werden, dass der wahre Parameterwert mit Wahrscheinlichkeit $1-\alpha$ in einem realisierten Konfidenzintervall $[T_1(\boldsymbol{x}), T_2(\boldsymbol{x})]$ liegt. Man kann für ein spezielles realisiertes Konfidenzintervall keine Aussage treffen, ob es ϑ überdeckt oder nicht. Der Grund ist, dass das Zufallsgeschehen sich nicht auf den Parameter bezieht, sondern auf das Beobachten der Stichprobenvariablen. (Der Parameter ist zwar unbekannt, unterliegt aber in der von uns behandelten Theorie keiner Wahrscheinlichkeitsverteilung. Denkbar ist der Fall, dass auf der Parametermenge Θ eine Verteilung gegeben ist. Dann sind Intervallschätzungen über den Bayes-Ansatz möglich. Dieser Themenbereich soll hier aber ausgespart bleiben.)

Die Aussagekraft einer Intervallschätzung liegt in folgender Interpretation: Wenn aus der zugrunde liegenden Verteilung wiederholt Stichproben gezogen und zu jeder realisierten Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)$ die Werte der Stichprobenfunktionen $T_1(\boldsymbol{x}), T_2(\boldsymbol{x})$ errechnet werden, und so eine Folge von konkreten Intervallen entsteht, dann enthalten diese den unbekannten, wahren Parameterwert ϑ mit einer relativen Häufigkeit, die etwa dem Konfidenzniveau entspricht. Andererseits enthalten approximativ $100 \cdot \alpha\%$ dieser Intervalle den Parameter nicht.

Das Konfidenzniveau kann als eine Art Maßzahl für den Einsatz bei einer fairen Wette interpretiert werden. Wenn bei einem Niveau von 95% gesagt wird, dass der Parameter im realisierten Intervall liegt, so ist es fair, einen Wetteinsatz von 95 Einheiten gegen 5 Einheiten zu setzen. Die Wiederholung der Wette führt dann zu einem ausgeglichenen Verhältnis von Gewinn und Verlust.

Beispiel 7.1.3 *Konfidenzintervall für μ bei Normalverteilung*

Sei $X_1, \ldots, X_n$ eine Zufallsstichprobe aus einer $\mathcal{N}(\mu, \sigma^2)$-Verteilung. σ^2 sei bekannt. Wird als Konfidenzniveau $1-\alpha = 0.95$ vorgegeben, so erhalten wir wegen

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

das Konfidenzintervall

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

Für $1-\alpha = 0.99$ erhalten wir mit $z_{1-0.005} = 2.576$ das breitere Intervall

$$\left[\bar{X} - 2.576\frac{\sigma}{\sqrt{n}}, \bar{X} + 2.576\frac{\sigma}{\sqrt{n}}\right].$$

Das Beispiel macht deutlich, dass ein Konfidenzintervall umso breiter ist, je größer das Konfidenzniveau gewählt wird. Ein hoher Grad an ‚Vertrauen' geht mit einer wenig präzisen Angabe über die Lage des unbekannten Parameters einher. In der Anwendung müssen wir also jeweils zwischen der Präzision und der Zuverlässigkeit abwägen.

Häufig sind die Angaben von Konfidenzintervallen nur über asymptotische Resultate möglich. Aufgrund des zentralen Grenzwertsatzes können wir in vielen Fällen von der approximativen Normalverteilung ausgehen. Bisweilen führen geeignete, etwa varianzstabilisierende Transformationen zu verbesserten Approximationen.

Beispiel 7.1.4 *asymptotisches Konfidenzintervall bei Poisson-Verteilung*

Für Poisson-verteilte Zufallsvariablen X gilt $\mathrm{E}(X) = \lambda$, $\mathrm{Var}(X) = \lambda$. Die varianzstabilisierende Transformation ist hier die Quadratwurzel. Für eine Stichprobe aus einer $\mathscr{P}(\lambda)$-Verteilung ist nach den Ergebnissen aus Kapitel 5.2 $\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda})$ approximativ $\mathscr{N}(0, 1/4)$ -verteilt, so dass ein approximatives $(1-\alpha)$-Konfidenzintervall für λ gegeben ist durch

$$\left[\left(\max\left\{0, \sqrt{\bar{X}} - \frac{z_{1-\alpha/2}}{2\sqrt{n}}\right\}\right)^2, \left(\sqrt{\bar{X}} + \frac{z_{1-\alpha/2}}{2\sqrt{n}}\right)^2\right].$$

7.2 Konstruktion von Konfidenzintervallen

7.2.1 Die Pivot- und die statistische Methode

Das dem Beispiel 7.1.3 zugrunde liegende Vorgehen besteht in der Umformung

$$-z \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le z \quad \Leftrightarrow \quad \bar{X} - z\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z\frac{\sigma}{\sqrt{n}}.$$

Es wird also ausgenutzt, dass die standardisierte Stichprobenfunktion $\bar{X}$, $\sqrt{n}(\bar{X} - \mu)/\sigma$, eine von μ unabhängige Verteilung besitzt. Das ist ein Beispiel der allgemein anwendbaren Pivot-Methode. (Pivot,(frz.): Angelpunkt/Drehpunkt, d.h. μ ist diejenige Größe, um die die obige Ungleichung gedreht wird.)

Definition 7.2.1 *Pivot*

Gegeben seien die Stichprobenvariablen $X_1, \ldots, X_n$ mit der Dichtefunktion $f(\cdot;\vartheta)$, $\vartheta \in \Theta$. Die Funktion $Q(\boldsymbol{X}, \vartheta)$ heißt genau dann *Pivot-Größe* oder kurz *Pivot*, wenn sie eine Verteilung besitzt, die nicht von ϑ abhängt.

Definition 7.2.2 *Pivot-Methode*

Die Pivot-Methode besteht aus folgenden Schritten:

1. Für den unbekannten Parameter ϑ wird ein Schätzer $T(\boldsymbol{X})$ gesucht.
2. Es wird eine Funktion des Schätzers und des Parameters, das sogenannte Pivot $Q(T(\boldsymbol{X}),\vartheta)$, gesucht, dessen Verteilung unabhängig von ϑ ist.
3. Es werden zwei Grenzen q_1, q_2 bestimmt, so dass

$$\mathrm{P}(q_1 \le Q(T(\boldsymbol{X}),\vartheta) \le q_2) \ge 1-\alpha. \tag{7.1}$$

   q_1, q_2 werden so gewählt, dass das vorgegebene Konfidenzniveau möglichst gut ausgeschöpft wird.
4. Die Ungleichungen $q_1 \le Q(T(\boldsymbol{X}),\vartheta) \le q_2$ werden so umgeformt, dass ϑ isoliert steht (Pivotisierung).

Dann ist

$$\{\vartheta | q_1 \le Q(T(\boldsymbol{X}),\vartheta) \le q_2\} = \{\vartheta | T_1(\boldsymbol{X}) \le \vartheta \le T_2(\boldsymbol{X})\} \tag{7.2}$$

ein $(1-\alpha)$-Konfidenzintervall für ϑ.

Naheliegender Weise sucht man vor allem solche Pivots, die von suffizienten Statistiken abhängen.

Beispiel 7.2.3 *KI für den Parameter einer Gleichverteilung*

Sei $X_{1:n},\ldots,X_{n:n}$ die geordnete Statistik zu einer Stichprobe aus einer Gleichverteilung über dem Intervall $[0,\vartheta]$. Die minimalsuffiziente Statistik ist hier $T=\max\{X_1,\ldots,X_n\}$. T hat die Dichte $f_T(t;\vartheta) = n t^{n-1}/\vartheta^n$ für $0<t<1$. Eine auf T basierende Pivot-Größe ist daher $Q=T/\vartheta$. Q ist $\mathcal{R}(0,1)$-verteilt. Die Grenzen des Konfidenzintervalls sind dann aus folgender Gleichung zu bestimmen:

$$\mathrm{P}(a \le Q \le b) = \mathrm{P}_\vartheta(a\vartheta \le T \le b\vartheta) = \mathrm{P}\left(\frac{T}{b} \le \vartheta \le \frac{T}{a}\right) = 1-\alpha.$$

Wegen $\mathrm{P}(T \le \vartheta) = 1$ liegt es nahe, für b den Wert 1 zu wählen. Aufgrund von

$$\mathrm{P}(a \le Q \le 1) = 1 - \left(\frac{a\vartheta}{\vartheta}\right)^n = 1-a^n = 1-\alpha$$

ist dann das Konfidenzintervall gleich $[T, T\alpha^{-1/n}]$.

Beispiel 7.2.4 *KI für σ^2 bei Normalverteilung*

$X_1,\ldots,X_n$ sei eine Stichprobe aus einer $\mathcal{N}(\mu,\sigma^2)$ -Verteilung. Wir bestimmen ein $(1-\alpha)$-Konfidenzintervall für σ^2 mit der Pivot-Methode. Dazu gehen wir aus von der Schätzfunktion $S^2 = \sum(X_i-\bar{X})^2/n$, für die ja gilt:

$$\frac{n}{\sigma^2}S^2 \sim \chi^2_{n-1}.$$

Damit ist nS^2/σ^2 ein Pivot. Für geeignete Quantile a und b ist

$$\mathrm{P}\left(a \le \frac{n}{\sigma^2}S^2 \le b\right) = 1-\alpha.$$

Elementare Umformungen ergeben:

$$1-\alpha = \mathrm{P}\left(a \le \frac{n}{\sigma^2}S^2 \le b\right) = \mathrm{P}\left(\frac{n}{b}S^2 \le \sigma^2 \le \frac{n}{a}S^2\right).$$

Damit ist $[nS^2/b, nS^2/a]$ ein $(1-\alpha)$-Konfidenzintervall. Üblicherweise werden a und b als $\alpha/2$- und $(1-\alpha/2)$ -Quantile der χ^2_{n-1}-Verteilung gewählt.

Es können an jedem der Schritte 1. bis 4. Probleme mit der Pivot-Methode auftauchen. Das geringste Problem ist dabei, einen (erwartungstreuen) Schätzer zu finden. Dann daraus aber ein Pivot zu entwickeln, ist schon weit schwieriger, in einigen Fällen sogar unmöglich. Allgemein gilt aber: Wenn ϑ ein Lageparameter ist, so ist $X-\vartheta$ unabhängig von ϑ verteilt und somit sind $X_i-\vartheta$ Pivots. Für Skalenparameter ϑ erhält man über X/ϑ immer ein Pivot.

Hat man eine Funktion gefunden, deren Verteilung von ϑ unabhängig ist, so kann es noch vorkommen, dass man den Term zur Intervallbestimmung nicht nach ϑ auflösen kann, dass also das eigentliche Pivotisieren nicht zum Ziel führt. Hier müssen dann numerische Verfahren herangezogen werden, um eine angenäherte Lösung zu finden.

Schließlich sprechen wir von einem *approximativen Pivot*, wenn die Funktion eine Limesverteilung besitzt, die für $n \to \infty$ gegen eine bestimmte, von ϑ unabhängige Verteilung strebt. In diesem Fall kann das approximative Pivot für genügend große n wie ein eigentliches Pivot behandelt werden.

Beispiel 7.2.5 *KI für den Parameter einer Exponentialverteilung*

Für eine Stichprobe aus einer Exponentialverteilung gilt approximativ für die suffiziente Statistik $\bar{X}$:

$$\sqrt{n}\frac{\bar{X}-1/\lambda}{1/\lambda} \quad \dot{\sim} \quad \mathcal{N}(0,1).$$

Diese Funktion ist also ein approximatives Pivot. Wir erhalten:

$$-z \le \frac{\bar{X}-1/\lambda}{1/\lambda\sqrt{n}} \le z \quad \Leftrightarrow \quad \frac{1}{\bar{X}}\left(1-\frac{z}{\sqrt{n}}\right) \le \lambda \le \frac{1}{\bar{X}}\left(1+\frac{z}{\sqrt{n}}\right).$$

Damit ist mit der Wahl $z = z_{1-\alpha/2}$ durch die rechte Seite ein $(1-\alpha)$-Konfidenzintervall für den Parameter λ der Exponentialverteilung gegeben. Konkret erhalten wir für $1-\alpha = 0.95,\ n = 50,\ \bar{X} = 3$ das realisierte Konfidenzintervall $[0.24; 0.426]$.

Wie gesagt, gibt es nicht für jede Situation ein Pivot, mit Hilfe dessen man ein Konfidenzintervall für den interessierenden Parameter ϑ aufstellen kann. Lässt sich keine Funktion $Q(\boldsymbol{X}, \vartheta)$ finden, deren Verteilungsfunktion unabhängig von ϑ ist, können wir doch unter gewissen Bedingungen Schätzer für die Konstruktion von Konfidenzintervallen verwenden.

Sei $T(\boldsymbol{X})$ eine Schätzfunktion für ϑ. Dann hängen in der Gleichung $\mathrm{P}(a \le T(\boldsymbol{X}) \le b) = 1-\alpha$ dir Grenzen a und b vom Parameter ab. Wir können also schreiben:

$$\mathrm{P}_\vartheta(a(\vartheta) \le T(\boldsymbol{X}) \le b(\vartheta)) = 1-\alpha.$$

Kennen wir nun die Funktionen $a(\vartheta)$ und $b(\vartheta)$ explizit, und sind sie monoton (wachsend) und invertierbar, so können wir die beiden Ungleichungen

$$\mathrm{P}(a(\vartheta) \leq T(\boldsymbol{X})) = \alpha/2 \quad \text{und} \quad \mathrm{P}(T(\boldsymbol{X}) \leq b(\vartheta)) = \alpha/2$$

umformen zu

$$\mathrm{P}(\vartheta \leq a^{-1}(T(\boldsymbol{X}))) = \alpha/2 \quad \text{und} \quad \mathrm{P}(b^{-1}(T(\boldsymbol{X})) \leq \vartheta) = \alpha/2.$$

Damit ist es also unter geeigneten Bedingungen möglich, die Form $T_1(\boldsymbol{X}) < \vartheta < T_2(\boldsymbol{X})$ zu erreichen, wobei T_1 und T_2 unabhängig von ϑ sind und $[T_1(\boldsymbol{X}), T_2(\boldsymbol{X})]$ somit ein Konfidenzintervall ist.

Satz 7.2.6 *statistische Methode zur Konstruktion von Konfidenzintervallen*

Gegeben sei eine Stichprobe $\boldsymbol{X} = (X_1, \ldots, X_n)$ aus einer von dem Parameter ϑ, $\vartheta \in \Theta$ abhängigen Verteilung. Sei $T = T(\boldsymbol{X})$ eine Schätzfunktion für ϑ mit der Verteilungsfunktion $F_T(t;\vartheta)$. Seien α, α_1 und α_2 mit $0 < \alpha, \alpha_1, \alpha_2 < 1$ und $\alpha_1 + \alpha_2 = \alpha$ gegeben. $F_T(t;\vartheta)$ sei

(i) für jedes feste t eine monoton fallende Funktion von ϑ oder

(ii) für jedes feste t eine monoton wachsende Funktion von ϑ.

Es seien zwei Funktionen $h_1(t)$ und $h_2(t)$ folgendermaßen definierbar:

$$\int_t^{\infty} \mathrm{d}F_T(s; h_1(t)) = \mathrm{P}_{h_1(t)}(T > t) = \alpha_1, \tag{7.3a}$$

$$\int_{-\infty}^{t} \mathrm{d}F_T(s; h_2(t)) = \mathrm{P}_{h_2(t)}(T < t) = \alpha_2. \tag{7.3b}$$

Dann ist im Fall (i) $[h_1(T), h_2(T)]$ ein Konfidenzintervall zum Niveau $1-\alpha$ für ϑ und im Fall (ii) gilt dies für $[h_2(T), h_1(T)]$.

Beweis: Wir beweisen nur die Aussage (i) für stetige Verteilungen der Statistik T. Da $F_T(t;\vartheta)$ monoton fallend ist für jedes feste t und $1-\alpha_1 > \alpha_2$, gilt $h_1(t) < h_2(t)$ und die Werte $h_1(t)$ und $h_2(t)$ sind eindeutig. Weiter folgt aus der Monotonie:

$$\begin{aligned} F_T(t;\vartheta) < 1-\alpha_1 \quad &\Leftrightarrow \quad \vartheta > h_1(t), \\ F_T(t;\vartheta) > \alpha_2 \quad &\Leftrightarrow \quad \vartheta < h_2(t). \end{aligned}$$

Unter Ausnutzung der Integraltransformation erhalten wir

$$\mathrm{P}_\vartheta(h_1(T) \leq \vartheta \leq h_2(T)) = \mathrm{P}_\vartheta(\alpha_2 \leq F_T(T;\vartheta) \leq 1-\alpha_1) = 1-\alpha_1-\alpha_2 = 1-\alpha.$$

Die Aussage (ii) folgt für stetige Verteilungen analog. Für diskrete Verteilungen ist der Integraltransformationssatz nicht anwendbar. Hier muss mit den Wahrscheinlichkeiten direkt argumentiert werden.

Ist die Statistik T diskret und besteht auch der Parameterraum nur aus abzählbar vielen Punkten, so brauchen die im Satz angegebenen Funktionen nicht zu existieren. Dann sind die ‚=' durch ‚≤' zu ersetzen und es ist zu verlangen, dass für jeden Wert der Statistik jeweils mindestens ein Parameterwert die angegebenen Bedingungen erfüllt. Sonst braucht das erhaltene Konfidenzintervall nicht wohldefiniert zu sein. Siehe dazu Peskun (1990).

Beispiel 7.2.7 *Konfidenzintervall für den Parameter der geometrischen Verteilung*

Sei X geometrisch verteilt, $f(x;p) = p(1-p)^x$ $(x = 0,1,\ldots)$. Wir wollen mit der statistischen Methode ein Konfidenzintervall auf der Basis einer einzelnen Beobachtung bestimmen.

Die ML-Schätzfunktion für p ist

$$T(X) = \frac{1}{X+1}.$$

Wegen $\mathrm{P}\left(\frac{1}{X+1} \le t\right) = \mathrm{P}\left(X \ge \frac{1}{t} - 1\right) = (1-p)^{1/t}$ für $\frac{1}{t} = 1,2,3,\ldots$ ist für jedes feste t die Verteilungsfunktion von T monoton fallend in p. Damit liegt die Situation (i) des Satzes vor.

Wir bestimmen $h_1(t)$ gemäß (7.3a). Der Ansatz

$$\frac{\alpha}{2} = \mathrm{P}_{h_1(t)}(T > t) = \mathrm{P}_{h_1(t)}\left(X+1 < \frac{1}{t}\right) = 1-(1-h_1(t))^{\frac{1}{t}-1}$$

ergibt

$$h_1(t) = 1 - \left(1 - \frac{\alpha}{2}\right)^{t/(1-t)}.$$

Dies ist nur für $t = 1/2, 1/3, 1/4, \ldots$ sinnvoll. Den Fall $t = 0$ stellen wir zurück. Analog erhalten wir $h_2(t)$ gemäß (7.3b) aus

$$\frac{\alpha}{2} = \mathrm{P}_{h_2(t)}(T < t) = \mathrm{P}_{h_2(t)}\left(X+1 > \frac{1}{t}\right) = 1-(1-h_2(t))^{1/t}$$

zu

$$h_2(t) = 1 - \left(\frac{\alpha}{2}\right)^t.$$

Das $(1-\alpha)$-Konfidenzintervall für p mit $\alpha_1 = \alpha_2 = \alpha/2$ lautet folglich für $x = 1,2,3,\ldots$

$$\left[1 - \left(1 - \frac{\alpha}{2}\right)^{1/X}; 1 - \left(\frac{\alpha}{2}\right)^{1/(X+1)}\right].$$

Für $X = 0$ bzw. $T = 1$ geben die beiden Funktionen offensichtlich keine vernünftigen Grenzen. Daher setzen wir $h_1(1) = 0.95$ und $h_2(1) = 1$. Dies ist wegen

$$\mathrm{P}_{0.95}(X = 0) = \mathrm{P}_{0.95}(T = 1) = p \ge 0.95$$

für $h_1(1) \le p \le h_2(1)$ sinnvoll.

Beispiel 7.2.8 *Konfidenzintervall für den Parameter p der Binomialverteilung*

Sei $X \sim \mathcal{B}(n,p)$ mit festem n. Hier ist $F_X(x;p)$ monoton fallend in p für jedes feste x. Das erhalten wir durch die Betrachtung der ersten Ableitung von $F_X(x;p)$. Sei dazu x fest mit $0 < x < n$. (Für $x = n$ ist nichts zu zeigen und für $x = 0$ ist das offensichtlich.) Dann gilt:

$$\frac{\partial}{\partial p} F_X(x;p) = \frac{\partial}{\partial p} \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i} = \frac{1}{(1-p)p} \sum_{i=0}^{x} (i-np) \binom{n}{i} p^i (1-p)^{n-i}.$$

Die rechte Seite ist kleiner oder gleich Null:

Für $x < np$ gilt $\partial F_X(x;p)/\partial p < 0$, da wegen $(i-np) < 0$ für $i \le x$ alle Summanden kleiner als null sind. Für $x \ge np$ gilt

$$\begin{aligned} &\frac{1}{(1-p)p} \sum_{i=0}^{x} (i-np) \binom{n}{i} p^i (1-p)^{n-i} \\ &\qquad \le \frac{1}{(1-p)p} \sum_{i=0}^{n} (i-np) \binom{n}{i} p^i (1-p)^{n-i} = \frac{np-np}{(1-p)p} = 0. \end{aligned}$$

Die statistische Methode zur Bestimmung eines Konfidenzintervalles für p führt nun dazu, dass zu dem beobachteten x die beiden Gleichungen

$$\sum_{i=x}^{n} \binom{n}{i} p^i (1-p)^{n-i} = \alpha_1 \quad \text{und} \quad \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i} = \alpha_2 \tag{7.4}$$

gelöst werden müssen. Daraus erhalten wir die Unter- und die Obergrenze des Konfidenzintervalls, p_u und p_o. Das Vorgehen ist in der Abbildung 7.1 illustriert.

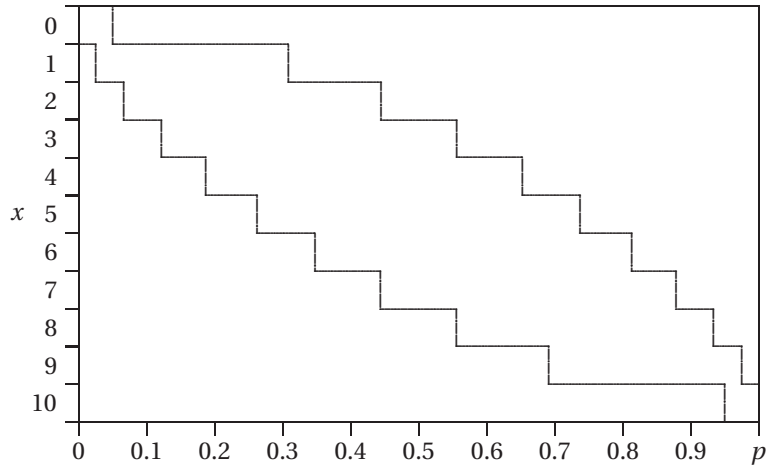


Abb. 7.1: *Clopper-Pearson-Konfidenzintervalle für p* $(n = 10,\ 1-\alpha = 0.95)$

Bei der Bestimmung von p_u und p_o aus (7.4) können bekannte Beziehungen zur $\mathcal{F}$-Verteilung ausgenutzt werden. Werden wie üblich $\alpha_1 = \alpha_2 = \alpha/2$ gewählt, so führt dies

für $0 < x \leq n$ zu

$$T_1(x) = \frac{x\mathscr{F}_{2x,2(n-x+1);\alpha/2}}{(n-x+1)+x\mathscr{F}_{2x,2(n-x+1);\alpha/2}} = \frac{x}{x+(n-x+1)F_{2(n-x+1),2x;\alpha/2}} \tag{7.5a}$$

und für $0 \leq x < n$ zu

$$T_2(x) = \frac{(x+1)\mathscr{F}_{2(x+1),2(n-x);\alpha/2}}{(n-x)+(x+1)\mathscr{F}_{2(x+1),2(n-x);\alpha/2}} = \frac{x+1}{(x+1)+(n-x)\mathscr{F}_{2(n-x),2(x+1);\alpha/2}}. \tag{7.5b}$$

Die resultierenden Grenzen $T_1(x)$ und $T_2(x)$ werden nach *Clopper und Pearson* benannt und sind in verschiedenen Tafelwerken angegeben. Das tatsächliche Konfidenzniveau kann hier sehr viel größer sein als $1-\alpha$, nämlich bis zu $1-\alpha/2$, siehe Angus & Schafer (1984).

7.2.2 Konfidenzintervalle auf der Basis der Likelihoodfunktion

Die große Attraktivität der Maximum-Likelihood-Schätzung legt nahe, bei der Konstruktion von Konfidenzintervallen von der Likelihoodfunktion auszugehen und mittels der Likelihoodfunktion zu bestimmen, welche Parameterwerte in geeigneter Nähe des ML-Schätzers $\hat{\vartheta}$ liegen. Wie wir in Kapitel 4 gesehen haben, basiert der adäquate Vergleich auf dem Verhältnis der Likelihoodfunktion für verschiedene Werte, hier für $\hat{\vartheta}$ und einem Wert ϑ. Wir betrachten also den Likelihoodquotienten

$$\ell(\vartheta;\boldsymbol{X}) = \frac{f(X_1;\vartheta)\cdot\ldots\cdot f(X_n;\vartheta)}{f(X_1;\hat{\vartheta})\cdot\ldots\cdot f(X_n;\hat{\vartheta})}.$$

Auf der Basis der asymptotischen Verteilung von $\ell(\vartheta)$ kann nun eine Klasse von Konfidenzintervallen gewonnen werden. Diese Methode wird selten verwendet, da sie die asymptotische Normalverteilung des ML-Schätzers voraussetzt. Dann kann aber im Fall explizit angebbarer Schätzer auch einfach ein Konfidenzintervall über die approximative Pivotisierung gewonnen werden. Bei nicht explizit angebbaren ML-Schätzern ist diese Methode wohl aufgrund der numerischen Aspekte unberücksichtigt geblieben. Das ist heute allerdings kein relevanter Gesichtspunkt mehr. Wir geben die Methode hier auch deshalb an, weil eine nichtparametrische Weiterentwicklung dieses Ansatzes eine große Attraktivität besitzt.

Satz 7.2.9 *asymptotische Verteilung des logarithmierten Likelihoodquotienten*

Die Komponenten von $\boldsymbol{X}_n = (X_1,\ldots,X_n)$ mögen eine Folge von unabhängigen identisch verteilten Zufallsvariablen aus einer Verteilung mit der Dichte $f(x;\vartheta)$, $\vartheta \in \Theta$ bilden. $f(xt;\vartheta)$ sei genügend regulär, um speziell die asymptotische Normalverteiltheit des ML-Schätzers $\hat{\vartheta}$ zu sichern. ϑ_0 sei der wahre Parameterwert. Dann gilt für den Likelihoodquotienten

$$\ell(\vartheta_0;\boldsymbol{X}_n) = \frac{f(X_1;\vartheta_0)\cdot\ldots\cdot f(X_n;\vartheta_0)}{f(X_1;\hat{\vartheta})\cdot\ldots\cdot f(X_n;\hat{\vartheta})}: \quad -2\ln(\ell(\vartheta_0;\boldsymbol{X}_n)) \dot{\sim} \chi^2_1.$$

$-2\ln(\ell(\vartheta_0;\boldsymbol{X}_n))$ ist also approximativ chiquadratverteilt mit einem Freiheitsgrad.

Beweisskizze: Wir unterdrücken im Folgenden weitgehend den Index n bei $\boldsymbol{X}$. Sei $\mathscr{L}(\vartheta;\boldsymbol{X})$ die Loglikelihoodfunktion,

$$\mathscr{L}(\vartheta;\boldsymbol{X}) = \ln(f(X_1;\vartheta)\cdot\ldots\cdot f(X_n;\vartheta)) = \sum_{i=1}^{n} \ln f(X_i;\vartheta).$$

Ist ϑ_0 der wahre Parameterwert, so konvergiert $\mathscr{L}(\vartheta_0;\boldsymbol{X})/n$ aufgrund des schwachen Gesetzes der großen Zahlen in Wahrscheinlichkeit gegen $\mathrm{E}(\ln f(X;\vartheta_0))$. Das Gleiche gilt mit Satz 6.2.3 für $\mathscr{L}(\hat{\vartheta};\boldsymbol{X})/n$, wenn der ML-Schätzer $\hat{\vartheta}$ konsistent ist.

Wir können dann bei genügend großem n den Ausdruck $\sum_{i=1}^{n} \ln f(X_i;\vartheta)/n$ bei $\hat{\vartheta}$ in einer Taylorreihe entwickeln und erhalten

$$\frac{1}{n}\sum_{i=1}^{n} \ln f(X_i;\vartheta_0) = \frac{1}{n}\sum_{i=1}^{n} \ln f(X_i;\hat{\vartheta}) + \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\vartheta} \ln f(X_i;\vartheta^*)\right)(\hat{\vartheta}-\vartheta_0) + \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\vartheta^2} \ln f(X_i;\vartheta^*)\right)(\hat{\vartheta}-\vartheta_0)^2.$$

Dabei ist ϑ^* ein Wert zwischen $\hat{\vartheta}$ und ϑ_0.

Der zweite Term auf der rechten Seite verschwindet, da $\hat{\vartheta}$ konsistent ist und auch der andere Faktor gegen den Erwartungswert der Summanden, also gegen null geht. Nun konvergiert $\left(\sum_{i=1}^{n} \partial^2 \ln f(X_i;\vartheta^*)/\partial\vartheta^2\right)/n$ in Wahrscheinlichkeit gegen $\mathrm{E}(\partial^2 \ln f(X;\vartheta_0)/\partial\vartheta^2) = -I(\vartheta_0)$. Damit erhalten wir:

$$-2\sum_{i=1}^{n} \ln\frac{f(X_i;\vartheta_0)}{f(X_i;\hat{\vartheta})} = -2\left[\sum_{i=1}^{n} \ln f(X_i;\vartheta_0) - \sum_{i=1}^{n} \ln f(X_i;\hat{\vartheta})\right] \cong I(\vartheta_0)(\sqrt{n}(\hat{\vartheta}-\vartheta_0))^2.$$

$\sqrt{n}(\hat{\vartheta}-\vartheta_0)/\sqrt{I(\vartheta_0)^{-1}}$ ist asymptotisch $\mathscr{N}(0,1)$ -verteilt, so dass der Ausdruck auf der rechten Seite χ^2_1-verteilt ist. Links steht aber gerade $-2\ln(\ell(\vartheta_0;\boldsymbol{X}))$.

Die Ausformulierung dieser Beweisskizze findet sich bei Wilks (1962).

Dieser Satz liefert die *Likelihood-Konfidenzintervalle*

$$R_{1-\alpha} = \left\{\vartheta \mid -2\ln(\ell(\vartheta;\boldsymbol{X})) \le \chi^2_{1;1-\alpha}\right\}. \tag{7.6}$$

Für den wahren Parameterwert gilt ja asymptotisch $\mathrm{P}_\vartheta(-2\ln(\ell(\vartheta;\boldsymbol{X})) \le \chi^2_{1;1-\alpha}) = 1-\alpha$, so dass er mit der entsprechenden Wahrscheinlichkeit zu dem Intervall $R_{1-\alpha}$ gehört.

Beispiel 7.2.10 *Likelihood-Konfidenzintervall bei Exponentialverteilung*

Seien X_i $(i = 1,2,\ldots,n,\ldots)$ exponentialverteilt mit dem Parameter λ. Dann ist $\hat{\lambda} = \bar{X}^{-1}$, und es gilt:

$$-2\left[\sum_{i=1}^{n} \ln f(X_i;\lambda) - \sum_{i=1}^{n} \ln f(X_i;\hat{\lambda})\right] = -2[n\cdot\ln(\lambda) - \lambda n\bar{X} + n\cdot\ln(\bar{X}) + n].$$

Das Konfidenzintervall besteht also aus allen Parameterwerten λ, die die Ungleichung

$$-2[n\cdot\ln(\lambda)-\lambda n\bar{X}+n\cdot\ln(\bar{X})+n]\leq\chi^2_{1;1-\alpha}$$

erfüllen.
Für den konkreten Fall $1-\alpha=0.95,\ n=50,\ \bar{x}=3$ erhalten wir die in der Abbildung 7.2 dargestellte Funktion $-2\ln(\ell(\lambda))$ mit der zugehörigen Grenze $\chi^2_{1;0.95}=3.841$.

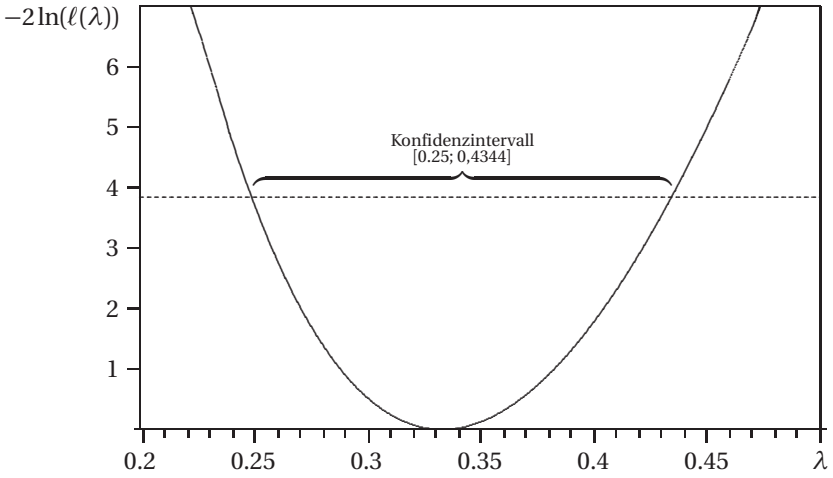


Abb. 7.2: *Likelihood-Konfidenzintervall bei der Exponentialverteilung*

Das ergibt das realisierte Konfidenzintervall $[0.25;0.4344]$. Das ist mit dem Konfidenzintervall $[0.24;0.426]$ aus Beispiel 7.2.5 zu vergleichen.

Für beliebige Verteilungen wird die empirische Verteilungsfunktion $\hat{F}_n(x)$ als nichtparametrischer ML-Schätzer für die zugrunde liegende Verteilungsfunktion angesehen, da $\hat{F}_n(x)$ die Funktion

$$L(F,\boldsymbol{x})=\prod_{i=1}^{n}\{F(X_i)-F(X_i-)\}$$

maximiert. $F(x-)$ bezeichnet dabei den linksseitigen Grenzwert von F an der Stelle x, für theoretische Verteilungen also $P(X<x)$. Der analog zum parametrischen Fall gebildete Likelihoodquotient

$$\ell(F,\boldsymbol{X})=\frac{L(F,\boldsymbol{X})}{L(\hat{F}_n)}$$

kann dann genauso wie das parametrische Gegenstück zu Konfidenzintervallen für den Formparameter $\vartheta(F_0)$ einer Verteilung F_0 führen, wenn die asymptotische Verteilung von $\ell(F_0,\boldsymbol{X})$ bzw. von $-2\ln(\ell(F_0,\boldsymbol{X}))$ bekannt ist. Die Konfidenzintervalle sind wieder von der Gestalt

$$\left\{\vartheta(F)|\,F\text{ ist eine Verteilungsfunktion mit }-2\ln\left(\frac{L(F,\boldsymbol{X})}{L(\hat{F}_n)}\right)\leq c\right\}. \tag{7.7}$$

Natürlich kann dieser Ansatz nicht für alle Verteilungen und alle Formparameter funktionieren. So ist bei $\vartheta(F) = \mathrm{E}(X)$ das durch (7.7) gegebene Intervall die ganze Zahlengerade, sofern nicht einschränkende Bedingungen an die Verteilung gestellt werden.

Eine offensichtliche Einschränkung bei der Betrachtung von $L(F,\boldsymbol{x})/L(\hat{F}_n)$ beruht darauf, dass dieser Quotient null ist, falls einer der Faktoren $\{F(x_i) - F(x_i-)\}$ verschwindet. Dann ist $-2\ln(L(F,\boldsymbol{x})/L(\hat{F}_n))$ unendlich groß, und die Verteilung kann zu der Menge (7.7) nichts mehr beitragen. Alle in Betracht kommenden Verteilungen sind also in Abhängigkeit von den Stichprobenvariablen $\boldsymbol{X} = (X_1,\ldots,X_n)$ gegeben durch

$$F(x) = \sum_{i=1}^{n} w_i I_{\{X_i \le x\}}. \tag{7.8}$$

Dabei gilt $w_i \ge 0$, $\sum w_i = 1$.
Sofern die Beobachtungen $x_1,\ldots,x_n$ alle verschieden sind, wird der Likelihoodquotient $L(F,\boldsymbol{x})/L(\hat{F}_n)$ bei einer gemäß (7.8) gegebenen Verteilung einfach zu

$$\frac{L(F,\boldsymbol{x})}{L(\hat{F}_n)} = \prod_{i=1}^{n} \frac{w_i}{1/n} = \prod_{i=1}^{n} n w_i.$$

Für eine mittels (7.8) gegebene Stichprobenfunktion $\vartheta(F,\boldsymbol{X})$, die das Stichprobengegenstück eines Formparameters $\vartheta(F)$ der Verteilung F darstellt, schreiben wir auch $\vartheta(\boldsymbol{w})$, wobei $\boldsymbol{w} = (w_1,\ldots,w_n)$.

Satz 7.2.11 *Asymptotik des empirischen Likelihoodquotienten*

Sei $X_1,\ldots,X_n,\ldots$ eine Folge von unabhängigen identisch verteilten Zufallsvariablen aus einer Verteilung mit der Verteilungsfunktion $F_0(x)$, für die die vierten Momente existieren. Sei $\vartheta_0 = \vartheta(F_0)$ der Erwartungswert, die Varianz oder eine stetige Funktion einer dieser beiden Formparameter. Dann gilt für den *empirischen Likelihoodquotienten* $\ell(\vartheta(F,\boldsymbol{X})) = \dfrac{L(\vartheta(F,\boldsymbol{X}),\boldsymbol{X})}{L(\vartheta(\hat{F}_n))} = \max\limits_{\boldsymbol{w}:\vartheta(\boldsymbol{w})=\vartheta(F),\, w_i\ge 0,\sum w_i=1} \prod\limits_{i=1}^{n} n w_i$:

$$-2\ln(\ell(\vartheta(F_0))) \quad \dot{\sim} \quad \chi^2_1. \tag{7.9}$$

Damit ist $R_{1-\alpha} = \{\vartheta(F,\boldsymbol{X}) | -2\ln(\ell(\vartheta(F))) \le \chi^2_{1;1-\alpha}\}$ ein asymptotisches $(1-\alpha)$-Konfidenzintervall für ϑ_0:

$$\mathrm{P}(\vartheta_0 \in R_c) \doteq 1-\alpha.$$

Beweis: Siehe Owen (1988), Hall & LaScala (1990).

Für die praktische Bestimmung der Konfidenzintervalle sind diejenigen Werte ϑ_U und ϑ_O zu bestimmen, für die gilt:

$$\vartheta_U = \inf\left\{\vartheta(\boldsymbol{w}) | \boldsymbol{w} = (w_1,\ldots,w_n),\ w_i \ge 0,\ \sum w_i = 1, -2\sum \ln(n w_i) \le \chi^2_{1;1-\alpha}\right\},$$
$$\vartheta_O = \sup\left\{\vartheta(\boldsymbol{w}) | \boldsymbol{w} = (w_1,\ldots,w_n),\ w_i \ge 0,\ \sum w_i = 1, -2\sum \ln(n w_i) \le \chi^2_{1;1-\alpha}\right\}.$$

Sei nun $\vartheta(F)$ der Erwartungswert und somit $\vartheta(\boldsymbol{w}) = \sum w_i X_i$. Dann erhalten wir Ausdrücke für die Gewichte w_i mit dem Lagrange-Multiplikatoren-Ansatz für die Extremwertbestimmung unter Nebenbedingungen. Da beide Extrema ϑ_U und ϑ_O für Gewichtsvektoren mit $-2\sum \ln(n w_i) = \chi^2_{1;1-\alpha} = c$ angenommen werden, betrachten wir die Funktion

$$G(\boldsymbol{w},\gamma,\delta) = \sum_{i=1}^{n} w_i X_i + \beta\left(\frac{1}{2}c + \sum_{i=1}^{n}\ln(n w_i)\right) + \gamma\left(\sum_{i=1}^{n} w_i - 1\right).$$

Nullsetzen der partiellen Ableitungen $\partial G(\boldsymbol{w},\gamma,\delta)/\partial w_i$ liefert

$$w_i = \frac{\beta}{X_i - \gamma} \qquad i = 1,\ldots,n.$$

Die Reparametrisierung

$$\delta = \frac{1}{n\beta}, \quad \mu = n\beta\left(1 + \frac{\gamma}{n\beta}\right)$$

liefert:

$$w_i = \frac{1}{n(1+\delta(X_i-\mu))} \qquad i = 1,\ldots,n.$$

Durch Einsetzen sehen wir, dass nun die Nebenbedingung $\sum w_i - 1 = 0$ äquivalent ist zu $\sum w_i(X_i - \mu) = 0$. Weiter ist $-2\sum \ln(n w_i) = c$ gleichwertig mit $2\sum \ln(1+\delta(X_i-\mu)) = c$.

Es sind also die beiden Lösungen (δ_1,μ_1) und (δ_2,μ_2) der Gleichungen

$$\sum_{i=1}^{n} \frac{1}{1+\delta(X_i-\mu)}(X_i - \mu) = 0$$
$$2\sum_{i=1}^{n} \ln(1+\delta(X_i-\mu)) - c = 0$$

zu bestimmen und für μ_U den kleineren und für μ_O den größeren zu nehmen.

Für die Bestimmung eines Konfidenzintervalles für die Varianz $\mathrm{Var}(X)$ der zugrunde liegenden Verteilung beachten wir die Darstellung $\delta(F) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$, die zu $\vartheta(\boldsymbol{w}) = \sum w_i X_i^2 - (\sum w_i X_i)^2$ führt. Der gleiche Ansatz wie beim Erwartungswert führt unter Setzung von $\tau = \sum w_i(X_i-\mu)^2$ zu den drei Gleichungen

$$\sum_{i=1}^{n} \frac{1}{1+\delta[(X_i-\mu)^2-\tau]}(X_i - \mu) = 0$$
$$\sum_{i=1}^{n} \frac{1}{1+\delta[(X_i-\mu)^2-\tau]}[(X_i-\mu)^2 - \tau] = 0$$
$$2\sum_{i=1}^{n} \ln(1+\delta[(X_i-\mu)^2-\tau]) - c = 0\,.$$

Das Konfidenzintervall für die Varianz wird aus dem kleinsten und dem größten Wert von τ der Lösungen gebildet.

Die auf der empirischen Likelihood basierenden Konfidenzintervalle haben einen engen Bezug zu den *Bayes'schen Bootstrap-Konfidenzintervallen*, siehe Rubin (1981). Zunächst hat Efron (1979) *Bootstrap Perzentil-Konfidenzintervalle* in der Weise eingeführt, dass die Bootstrapverteilung eines Parameters auf der Basis einer Stichprobe ermittelt wird und dann die $\alpha/2$- und $1-\alpha/2$-Quantile dieser Bootstrapverteilung als Konfidenzgrenzen genommen werden. Beim Bayesschen Bootstrap gelangen die beobachteten Werte nicht mit der Wahrscheinlichkeit $1/n$ wieder in die Bootstrapstichprobe, sondern mit einer beliebigen, sich nach jedem Bootstrapzug ändernden Wahrscheinlichkeit; dabei ist natürlich die Randbedingung $\sum p_i = 1$ zu beachten. Damit kann das Bayessche Bootstrap als eine Variante der empirischen Likelihood-Konfidenzintervalle angesehen werden. Letztere basieren ja auf einer asymptotischen Verteilungtheorie.

7.3 Eigenschaften von Konfidenzintervallen

Da i.d.R. mehrere Konfidenzintervalle für einen Parameter möglich sind, benötigen wir Kriterien, nach denen eines davon ausgewählt werden kann. Da die Wahrscheinlichkeit, mit der der wahre Parameterwert von den beiden Zufallsvariablen eingeschlossen wird, vorgegeben ist, liegt es nahe, zur Beurteilung eines Konfidenzintervalles die Wahrscheinlichkeit heranzuziehen, mit der vom wahren Wert verschiedene Parameterwerte überdeckt werden.

Definition 7.3.1 *gleichmäßig bestes Konfidenzintervall*

Ein Konfidenzintervall $[T_1, T_2]$ für einen Parameter $g(\vartheta) \in \Theta$ ist *optimal* oder *gleichmäßig bestes Konfidenzintervall* zum vorgegebenem Niveau $1-\alpha$, wenn für alle anderen Konfidenzintervalle $[T_1', T_2']$ zu diesem Niveau $1-\alpha$ gilt:

$$\mathrm{P}_\vartheta(T_1 \le g(\vartheta') \le T_2) \le \mathrm{P}_\vartheta(T_1' \le g(\vartheta') \le T_2') \qquad \text{für alle } \vartheta, \vartheta' \in \Theta,\ \vartheta' \neq \vartheta\,.$$

In der Regel gibt es kein optimales Konfidenzintervall zu einem vorgegebenem Niveau. Daher liegt es nahe, die zur Konkurrenz zugelassenen einzuschränken und in der eingeschränkten Klasse nach einem besten zu suchen. In lockerer Analogie zur Punktschätzung führen wir das Konzept der Unverfälschtheit ein.

Definition 7.3.2 *unverfälschtes Konfidenzintervall*

Ein $(1-\alpha)$-Konfidenzintervall $[T_1, T_2]$ für den Parameter $g(\vartheta)$ heißt *unverfälscht*, wenn für alle ϑ, ϑ' gilt:

$$\mathrm{P}_\vartheta(T_1 \le g(\vartheta) \le T_2) \ge \mathrm{P}_\vartheta(T_1 \le g(\vartheta') \le T_2).$$

Die Forderung der Unverfälschtheit besagt im stetigen Fall, dass die falschen Parameterwerte höchstens mit der Wahrscheinlichkeit $1-\alpha$ überdeckt werden, während ja der wahre Parameterwert mindestens mit der Wahrscheinlichkeit $1-\alpha$ von den Intervallgrenzen eingeschlossen wird.

Beispiel 7.3.3 *KI für den Parameter einer Gleichverteilung - Fortsetzung von Seite 230*

Für den Parameter ϑ einer rechteckverteilten Zufallsvariablen erhielten wir das Konfidenzintervall $\left[T, T\alpha^{-1/n}\right]$. Wir zeigen, dass dieses unverfälscht ist. Sei dazu $\vartheta' > 0$ ein anderer Parameterwert, $\vartheta' = f \cdot \vartheta$. Dann gilt:

$$\mathrm{P}_\vartheta(T \le \vartheta' \le T\alpha^{-1/n}) = \mathrm{P}_\vartheta(T \le f\cdot\vartheta \le T\alpha^{-1/n}) = \mathrm{P}_\vartheta\left(f\alpha^{1/n} \le \frac{T}{\vartheta} \le f\right).$$

Für $f > 1$ folgt

$$\mathrm{P}_\vartheta\left(f\alpha^{1/n} \le \frac{T}{\vartheta} \le f\right) = \mathrm{P}_\vartheta\left(f\alpha^{1/n} \le \frac{T}{\vartheta} \le 1\right) = 1 - (f\alpha^{1/n})^n = 1 - f^n\alpha < 1-\alpha.$$

Für $f < 1$ erhalten wir:

$$\mathrm{P}_\vartheta\left(f\alpha^{1/n} \le \frac{T}{\vartheta} \le f\right) = \mathrm{P}_\vartheta\left(f\alpha^{1/n} \le \frac{T}{\vartheta} \le 1\right) = f^n - (f\alpha^{1/n})^n = f^n(1-\alpha) < 1-\alpha.$$

Das Konzept der gleichmäßig besten unverfälschten Konfidenzintervalle hängt eng mit dem optimaler Tests zusammen. Daher verfolgen wir diese Optimalitätseigenschaft nun nicht weiter, sondern greifen sie später noch einmal auf.

Weil es nicht gar soviele Situationen mit gleichmäßig besten unverfälschten Konfidenzintervallen gibt, werden auch andere Gütekriterien herangezogen. Die Forderung, dass ein Konfidenzintervall möglichst kurz sein sollte, entspricht der Intuition. Die Kürze eines Konfidenzintervalles ist aber im Wesentlichen als Substitut für die eigentliche Optimalität zu sehen. Denn es ist einsichtig, dass ein kürzeres Intervall auch andere Parameterwerte mit nur kleinerer Wahrscheinlichkeit überdeckt. Mit den Intervallgrenzen ist meist auch die Länge $T_2(\boldsymbol{X}) - T_1(\boldsymbol{X})$ selbst eine Zufallsvariable. Daher wird an ihrer Stelle die erwartete Länge $\mathrm{E}_\vartheta(T_2 - T_1)$ betrachtet. Leider gibt es auch hier i.d.R. kein Konfidenzintervall zu einem vorgegebenem Niveau mit minimaler erwarteter Länge für alle ϑ. Man gibt sich daher häufig mit weniger zufrieden. Hat man mit der Pivot- oder mit der statistischen Methode ein Konfidenzintervall bestimmt, so sucht man nur das kürzeste, das sich mit diesem Ansatz ergibt.

Im Beispiel 7.2.4 hängt die Länge des Konfidenzintervalles nur von der Differenz der 'einfachen Funktion' $1/x$ ab, die an zwei Stellen a und b zu bestimmen ist. Diese Differenz kann als Integral ausgedrückt werden:

$$\frac{1}{a} - \frac{1}{b} = \int_a^b \frac{1}{t^2}\,\mathrm{d}t.$$

Das ist typisch für viele derartige Anwendungen. Somit gelangen wir mit diesem Prinzip zu folgender Formulierung des Problems der Bestimmung eines kürzesten Konfidenzintervalles:

Bestimme die Werte a und b, die

$$\int_a^b g(t)\,\mathrm{d}t$$

unter Berücksichtigung der Nebenbedingung an die zugrunde liegende Dichte:

$$\int_a^b f(t)\,\mathrm{d}t = 1-\alpha$$

minimieren. In einer geeigneten Verallgemeinerung erhalten wir die Lösung dieser Aufgabe mit folgendem Satz.

Satz 7.3.4

Sei $f(t)$ eine stetige Dichte und $g(t)$ eine stetige positive Funktion. Um

$$\int_C g(t)\,\mathrm{d}t \quad \stackrel{!}{=} \quad \min_C \tag{7.10}$$

unter der Nebenbedingung

$$\int_C f(t)\,\mathrm{d}t = 1-\alpha \tag{7.11}$$

zu lösen, ist $C = \{t | f(t)/g(t) > k\}$ zu wählen, wobei k durch die Nebenbedingung festgelegt ist.

Beweis: Für jede Menge D, die (7.11) erfüllt, erhalten wir:

$$\begin{aligned}
0 &= \int_D f(t)\,\mathrm{d}t - \int_C f(t)\,\mathrm{d}t = \int_{C^c\cap D} f(t)\,\mathrm{d}t - \int_{D^c\cap C} f(t)\,\mathrm{d}t \\
&\le \int_{C^c\cap D} k\cdot g(t)\,dt - \int_{D^c\cap C} k\cdot g(t)\,\mathrm{d}t \\
&= k\cdot\left\{\int_{C^c\cap D} g(t)\,\mathrm{d}t - \int_{D^c\cap C} g(t)\,\mathrm{d}t\right\} = k\cdot\left\{\int_D g(t)\,\mathrm{d}t - \int_C g(t)\,\mathrm{d}t\right\}.
\end{aligned}$$

Da k positiv ist, gilt $\int_D g(t)\,\mathrm{d}t - \int_C g(t)\,\mathrm{d}t \ge 0$. Somit ist C die gesuchte Menge, die (7.10) minimiert.

Beispiel 7.3.5 *KI für σ^2 bei Normalverteilung - Fortsetzung von Seite 230*

Wir betrachten wieder die Bestimmung eines $(1-\alpha)$-Konfidenzintervalles für den Parameter σ^2 einer $\mathcal{N}(\mu,\sigma^2)$-Verteilung. Wie wir gesehen haben gilt für geeignete Quantile a und b:

$$\mathrm{P}_{\sigma^2}\left(a \le \frac{n}{\sigma^2}S^2 \le b\right) = 1-\alpha.$$

Damit ergibt die Pivot-Methode das $(1-\alpha)$-Konfidenzintervall $[nS^2/b, nS^2/a]$. Im Beispiel 7.2.4 haben wir darauf hingewiesen, dass es üblich ist, a und b als $\alpha/2$- und $(1-\alpha/2)$-Quantile der χ^2_{n-1}-Verteilung zu wählen.

Konkret erhalten wir dann für $n=21,\ 1-\alpha=0.95$ das Konfidenzintervall:

$$\chi^2_{20;0.025} = 9.59, \quad \chi^2_{20;0.975} = 34.17 \quad \Rightarrow \quad [nS^2/34.17, nS^2/9.59].$$

Um mit dem Satz das kürzeste auf der Pivot-Größe nS^2/σ^2 basierende Konfidenzintervall zu gewinnen, haben wir mit $g(t) = 1/t^2$ und der Dichte $f(t; n-1)$ der χ^2_{n-1}-Verteilung den Bereich

$$C = \left\{ t \,\middle|\, \frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)} t^{(n-3)/2} e^{-t/2} > \frac{k}{t^2} \right\}.$$

Offenbar ist C ein Intervall, $C = [a, b]$. Die Grenzen a und b sind so zu bestimmen, dass

$$f(a; n-1) = k/a^2 \quad \text{und} \quad f(b; n-1) = k/b^2,$$

siehe Abbildung 7.3, bzw.

$$a^2 \cdot f(a; n-1) = b^2 \cdot f(b; n-1) \quad \Leftrightarrow \quad a^{n/2-1} e^{-a/2} = b^{n/2-1} e^{-b/2}. \tag{7.12}$$

Dabei muss natürlich gelten:

$$F(b; n-1) - F(a; n-1) = \int_a^b f(t; n-1)\,\mathrm{d}t = 1 - \alpha. \tag{7.13}$$

Dabei ist $F(t; n-1)$ die Verteilungsfunktion der χ^2-Verteilung mit $n-1$ Freiheitsgraden.

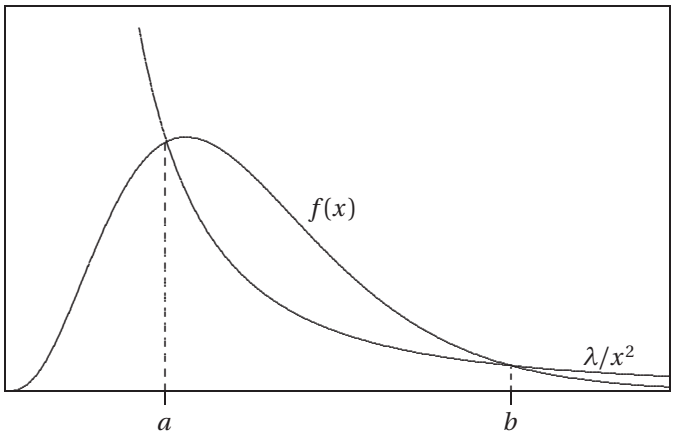


Abb. 7.3: *Zur Bestimmung eines Konfidenzintervalls*

Für die angegebene konkrete Situation erhalten wir: $a = 10.611874$, $b = 39.561042$. Die Länge des entsprechenden Konfidenzintervalles ist dann $nS^2 \cdot 0.06896$.

Eine Tabelle mit numerischen Lösungen für dieses Problem geben Tate & Klett (1969).

Schließlich fragen wir noch nach einem unverfälschten Konfidenzintervall auf der Basis dieser Statistik. Hier muss gelten:

$$\mathrm{P}_{\tilde{\sigma}^2}\left(a \le \frac{n}{\sigma^2} S^2 \le b \right) \le 1 - \alpha \quad \text{für } \tilde{\sigma}^2 \ne \sigma^2.$$

Wegen

$$\mathrm{P}_{\tilde{\sigma}^2}\left(a \le \frac{n}{\sigma^2}S^2 \le b\right) = \mathrm{P}_{\tilde{\sigma}^2}\left(\frac{\sigma^2}{\tilde{\sigma}^2}a \le \frac{n}{\tilde{\sigma}^2}S^2 \le \frac{\sigma^2}{\tilde{\sigma}^2}b\right)$$

sind a und b also so zu bestimmen, dass mit $\lambda = \sigma^2/\tilde{\sigma}^2$ gilt:

$$F(\lambda b) - F(\lambda a) \begin{cases} \le 1-\alpha & \text{für } \lambda \neq 1 \\ = 1-\alpha & \text{für } \lambda = 1. \end{cases}$$

Das erreichen wir über Nullsetzen der Ableitung $d(F(\lambda b) - F(\lambda a))/d\lambda$ an der Stelle $\lambda = 1$ unter Beachtung der Nebenbedingung $F(b) - F(a) = 1-\alpha$.

Für unser Zahlenbeispiel ergibt sich: $a = 9.958,\ b = 35.227$.

Beispiel 7.3.6 *KI für den Parameter einer Gleichverteilung - Fortsetzung von Seite 241*

Für den Parameter ϑ einer rechteckverteilten Zufallsvariablen erhielten wir das Konfidenzintervall

$$[T, T\alpha^{-1/n}].$$

Wir zeigen, dass das aus heuristischen Überlegungen heraus gewählte Intervall das kürzeste auf der Pivot-Größe $Q = T/\vartheta$ basierende ist.

Die auf dieser Pivot-Größe aufbauenden Konfidenzintervalle sind gegeben durch:

$$\mathrm{P}_{\vartheta}\left(\frac{T}{b} \le \vartheta \le \frac{T}{a}\right) = 1-\alpha.$$

Die Länge ist also $T\cdot(1/a - 1/b)$, so dass die Funktion g aus dem Satz gleich $1/t^2$ ist. Das kürzeste Konfidenzintervall ist folglich von der Form

$$\left\{ t \,\middle|\, \frac{n\,t^{n-1}}{\vartheta^n} > \frac{k}{t^2},\ 0 < t < \vartheta \right\} = \left\{ t \,|\, n\,t^{n+1} > k',\ 0 < t < \vartheta \right\},$$

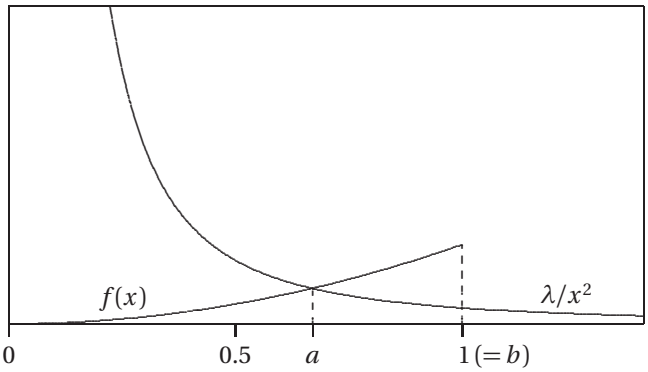


Abb. 7.4: *Zur Bestimmung des kürzesten Konfidenzintervalls bei $\mathcal{R}(0,\vartheta)$-Verteilung, vgl. Text*

vgl. Abbildung 7.4. Da t^{n+1} stärker als linear monoton wachsend ist für $0 < t < \vartheta$, erhalten wir das kürzeste Intervall, wenn a möglichst groß wird. Das führt zu

$$P_\vartheta(a \le Q \le 1) = 1 - \left(\frac{a\vartheta}{\vartheta}\right)^n = 1 - a^n = 1 - \alpha,$$

also zu $a = \alpha^{1/n}$ und $b = 1$.

Wir zeigten in Beispiel 7.2.3, dass das Konfidenzintervall $[T, T\alpha^{-1/n}]$ auch unverfälscht ist. Somit führen die verschiedenen Konzepte hier zum gleichen Ergebnis.

Bisweilen gibt es mehrere Pivot-Größen für ein Problem. Es scheint, dass dann diejenige Pivot-Größe die günstigste ist, die nur noch von einer suffizienten Statistik abhängt. Formal ist das aber noch nicht nachgewiesen.

Bei der statistischen Methode erhalten wir das kürzeste Konfidenzintervall auf der Basis der Statistik T durch die Minimierung von $h_2^{-1}(T) - h_1^{-1}(T)$ unter Berücksichtigung der Nebenbedingung $P_\vartheta(h_1^{-1}(T) \le \vartheta \le h_2^{-1}(T)) = 1 - (\alpha_1 + \alpha_2)$.

Beispiel 7.3.7 *statistische Methode für eine spezielle Verteilung*

Gegeben sei die Verteilung

$$F(x;\vartheta) = \begin{cases} 0 & \text{für } x \le 0 \\ x^\vartheta & \text{für } 0 < x \le 1 \\ 1 & \text{für } x > 1 \end{cases}.$$

Wir betrachten der Einfachheit halber eine Stichprobe vom Umfang $n = 1$. Dann führt die statistische Methode mit $T = X_1$ auf die beiden Beziehungen

$$F(h_1(\vartheta);\vartheta) = \alpha_1 \quad \Leftrightarrow \quad h_1(\vartheta) = \alpha_1^{1/\vartheta},$$

$$F(h_2(\vartheta);\vartheta) = 1 - \alpha_2 \quad \Leftrightarrow \quad h_2(\vartheta) = (1-\alpha_2)^{1/\vartheta}.$$

Die inversen Funktionen sind

$$h_1^{-1}(t) = \frac{\ln(\alpha_1)}{\ln(t)}, \quad h_2^{-1}(t) = \frac{\ln(1-\alpha_2)}{\ln(t)}.$$

Das Konfidenzintervall hat somit die Gestalt

$$\left[\frac{\ln(1-\alpha_2)}{\ln(X)}; \frac{\ln(\alpha_1)}{\ln(X)}\right].$$

Wir können hier noch α_1 und α_2 mit $\alpha_1 + \alpha_2 = \alpha$ geeignet festlegen, etwa so, dass die Intervalllänge minimiert wird. Das geschieht durch die Bestimmung des Minimums von

$$g(\alpha_2) = \frac{\ln(\alpha - \alpha_2)}{\ln(X)} - \frac{\ln(1-\alpha_2)}{\ln(X)} \qquad 0 \le \alpha_2 \le \alpha.$$

Das wird bei $\alpha_2 = 0$ angenommen. Das resultierende Konfidenzintervall ist also einseitig.

7.4 Konfidenzbereiche für mehrdimensionale Parameter

Wir können das Konzept der Konfidenzschätzung auf mehrdimensionale Parameter ausweiten. Sei dazu $g(\vartheta) = (g_1(\vartheta), \ldots, g_p(\vartheta))$ eine p-dimensionale Transformation des Parametervektors und seien $V_j(\boldsymbol{X}), U_j(\boldsymbol{X})$ $(j = 1, \ldots, p)$ eindimensionale Statistiken. Dann heißt der p-dimensionale Würfel

$$W(\boldsymbol{X}) = \left\{ g(\vartheta) | V_j(\boldsymbol{X}) \leq g_j(\vartheta) \leq U_j(\boldsymbol{X}),\ j = 1, \ldots, p \right\} \tag{7.14}$$

ein *Konfidenzbereich zum Niveau* $1-\alpha$, wenn der unbekannte, aber feste Parametervektor $(g_1(\vartheta), \ldots, g_p(\vartheta))$ mindestens mit der Wahrscheinlichkeit $1-\alpha$ überdeckt wird. Dafür schreiben wir

$$\mathrm{P}(W(\boldsymbol{X}) \ni g(\vartheta)) \geq 1 - \alpha. \tag{7.15}$$

Sind $[V_j(\boldsymbol{X}), U_j(\boldsymbol{X})]$ univariate $(1-\alpha_j)$-Konfidenzintervalle für $g_j(\vartheta)$, so gilt bei Unabhängigkeit der Paare von Statistiken $(V_1(\boldsymbol{X}), U_1(\boldsymbol{X})), \ldots, (V_p(\boldsymbol{X}),\ U_p(\boldsymbol{X}))$ für den zugehörigen p-dimensionalen Bereich $W(\boldsymbol{X})$:

$$\mathrm{P}(W(\boldsymbol{X}) \ni g(\vartheta)) = \mathrm{P}(V_1(\boldsymbol{X}) \leq g_1(\vartheta) \leq U_1(\boldsymbol{X})) \cdot \ldots \cdot \mathrm{P}(V_p(\boldsymbol{X}) \leq g_p(\vartheta) \leq U_p(\boldsymbol{X})) = \prod_{j=1}^{p} (1 - \alpha_j).$$

Damit lassen sich aus univariaten Konfidenzintervallen mehrdimensionale gewinnen.

In der Regel ist die Forderung der Unabhängigkeit der univariaten Konfidenzintervalle nicht erfüllt. Dann hilft die *Bonferroni-Ungleichung*, aus univariaten Konfidenzintervallen mehrdimensionale zu gewinnen. Die Ungleichung führt hier zu der Aussage:

$$\mathrm{P}(W(\boldsymbol{X}) \ni g(\vartheta)) \geq 1 - \sum_{j=1}^{p} \mathrm{P}([V_j(\boldsymbol{X}), U_j(\boldsymbol{X})] \not\ni g_j(\vartheta)) = 1 - \sum_{j=1}^{p} \alpha_j.$$

Wählen wir also $\alpha_j = \alpha/p$, so hält der p-dimensionale Bereich $W(\boldsymbol{X})$ das Niveau $1-\alpha$ ein.

Die Bonferroni-Intervalle sind i. Allg. etwas länger als Intervalle bei Unabhängigkeit. Dies ist offensichtlich wegen

$$1 - \frac{\alpha}{p} > (1-\alpha)^{1/p}.$$

Beispiel 7.4.1 *simultane Konfidenzaussage bei Normalverteilung*

Sei $X_1, \ldots, X_n$ eine Stichprobe aus einer $\mathcal{N}(\mu, \sigma^2)$- Verteilung. Wollen wir eine simultane Konfidenzaussage für μ und σ^2 machen, so können wir auf die univariaten Konfidenzintervalle für die beiden Parameter zurückgreifen. Das $(1-\alpha)$-Konfidenzintervall für μ ist

$$I_1(\boldsymbol{X}) = \left[\bar{X} - \frac{t_{n-1;1-\alpha/2} \cdot \hat{\sigma}}{\sqrt{n}},\ \bar{X} + \frac{t_{n-1;1-\alpha/2} \cdot \hat{\sigma}}{\sqrt{n}} \right],$$

und das auf den gleichen Randwahrscheinlichkeiten basierende für σ^2 lautet:

$$I_2(\boldsymbol{X}) = \left[\frac{n-1}{\chi^2_{n-1;1-\alpha/2}} \hat{\sigma}^2,\ \frac{n-1}{\chi^2_{n-1;\alpha/2}} \hat{\sigma}^2 \right].$$

Ersetzen wir in den beiden Intervallen α durch $\alpha/2$, so erhalten wir mit dem Bonferroni-Ansatz das Konfidenzrechteck

$$\left[\bar{X} - \frac{t_{n-1;1-\alpha/4} \cdot \hat{\sigma}}{\sqrt{n}},\ \bar{X} + \frac{t_{n-1;1-\alpha/4} \cdot \hat{\sigma}}{\sqrt{n}} \right] \times \left[\frac{n-1}{\chi^2_{n-1;1-\alpha/4}} \hat{\sigma}^2,\ \frac{n-1}{\chi^2_{n-1;\alpha/4}} \hat{\sigma}^2 \right],$$

das den Parametervektor (μ, σ^2) mit der (Mindest-)Wahrscheinlichkeit $1-\alpha$ überdeckt.

Unter Ausnutzen der asymptotischen multivariaten Normalverteilung verschiedener Schätzer lassen sich auch andere Konfidenzbereiche als die Bonferroni-Intervalle für die entsprechenden mehrdimensionalen Parameter angeben. Wir führen dies für den zweidimensionalen Fall aus (nach Douglas (1993)).

$\hat{\vartheta}_1$ und $\hat{\vartheta}_2$ seien Schätzer für ϑ_1 und ϑ_2 auf der Basis einer Stichprobe vom Umfang n. Diese Abhängigkeit machen wir nicht weiter kenntlich. Es gelte

$$\sqrt{n}(\hat{\vartheta}_1 - \vartheta_1,\ \hat{\vartheta}_2 - \vartheta_2) \quad \dot{\sim} \quad \mathcal{N}((0,0), V(\hat{\vartheta}_1, \hat{\vartheta}_2)).$$

Dann ist in vielen Fällen

$$\frac{1}{2}(\hat{\vartheta}_1 - \vartheta_1,\ \hat{\vartheta}_2 - \vartheta_2)\, \hat{V}(\hat{\vartheta}_1,\ \hat{\vartheta}_2)^{-1} \begin{pmatrix} \hat{\vartheta}_1 - \vartheta_1 \\ \hat{\vartheta}_2 - \vartheta_2 \end{pmatrix}$$

asymptotisch $\mathcal{F}_{2,n-2}$-verteilt. Wir haben hier eine mehrdimensionale Pivot-Größe vorzuliegen. Die Pivotisierung besteht einfach im Vertauschen von $\hat{\vartheta}_i$ und ϑ_i. Folglich ist dann

$$\left\{ (\vartheta_1, \vartheta_2) \,\middle|\, (\vartheta_1 - \hat{\vartheta}_1,\ \vartheta_2 - \hat{\vartheta}_2)\, \hat{V}(\hat{\vartheta}_1, \hat{\vartheta}_2)^{-1} \begin{pmatrix} \hat{\vartheta}_1 - \vartheta_1 \\ \hat{\vartheta}_2 - \vartheta_2 \end{pmatrix} \leq 2\mathcal{F}_{2,n-2;1-\alpha} \right\}$$

ein $(1-\alpha)$-Konfidenzbereich für $(\vartheta_1, \vartheta_2)$. Er hat die Form einer Ellipse mit dem Zentrum $(\hat{\vartheta}_1, \hat{\vartheta}_2)$.

Bei Erfüllen von standardmäßigen Regularitätsbedingungen ist dieser Konfidenzbereich bei Maximum-Likelihood-Schätzern angemessen. Hier ist

$$V(\hat{\vartheta}_1, \hat{\vartheta}_2)^{-1} = -\mathrm{E}\left(\frac{\partial^2 L(\boldsymbol{X}, \vartheta)}{\partial \vartheta_i\, \partial \vartheta_j} \right).$$

Ebenfalls gilt dies für Momentenschätzer $(\hat{\vartheta}_1, \hat{\vartheta}_2)$, die sich als Lösungen von

$$\bar{X} \stackrel{!}{=} \mathrm{E}_{\hat{\vartheta}_1, \hat{\vartheta}_2}(X) \quad \text{und} \quad \overline{X^2} \stackrel{!}{=} \mathrm{E}_{\hat{\vartheta}_1, \hat{\vartheta}_2}(X^2)$$

ergeben. Die Kovarianzmatrix ist approximativ

$$V(\hat{\vartheta}_1,\hat{\vartheta}_2)=\begin{pmatrix} \frac{\partial \mathrm{E}_{\vartheta_1,\vartheta_2}(X)}{\partial \vartheta_1} & \frac{\partial \mathrm{E}_{\vartheta_1,\vartheta_2}(X)}{\partial \vartheta_2} \\ \frac{\partial \mathrm{E}_{\vartheta_1,\vartheta_2}(X^2)}{\partial \vartheta_1} & \frac{\partial \mathrm{E}_{\vartheta_1,\vartheta_2}(X^2)}{\partial \vartheta_2} \end{pmatrix}^{-1} \cdot V_M(\bar{X},\overline{X^2})^{-1} \cdot \begin{pmatrix} \frac{\partial \mathrm{E}_{\vartheta_1,\vartheta_2}(X)}{\partial \vartheta_1} & \frac{\partial \mathrm{E}_{\vartheta_1,\vartheta_2}(X^2)}{\partial \vartheta_1} \\ \frac{\partial \mathrm{E}_{\vartheta_1,\vartheta_2}(X)}{\partial \vartheta_2} & \frac{\partial \mathrm{E}_{\vartheta_1,\vartheta_2}(X^2)}{\partial \vartheta_2} \end{pmatrix}^{-1}$$

mit

$$V_M(\bar{X},\overline{X^2})=\frac{1}{n}\begin{pmatrix} \mathrm{E}(X^2)-\mathrm{E}(X)^2 & \mathrm{E}(X^3)-\mathrm{E}(X)\mathrm{E}(X^2) \\ \mathrm{E}(X^3)-\mathrm{E}(X)\mathrm{E}(X^2) & \mathrm{E}(X^4)-\mathrm{E}(X^2)^2 \end{pmatrix}.$$

Beispiel 7.4.2 *Konfidenzellipse für zwei Erwartungswerte*

Auf mehreren Überfahrten wurde jeweils das (durchschnittliche) Stampfen X einer Fähre bestimmt und der Anteil Y der Fahrgäste, denen übel wurde, ermittelt. Die Daten (x_i,y_i) (X in $1/rads^2$, Y in %) sind in Abbildung 7.5 zusammen mit einer 95%-Konfidenzellipse für (μ_X,μ_Y) dargestellt.

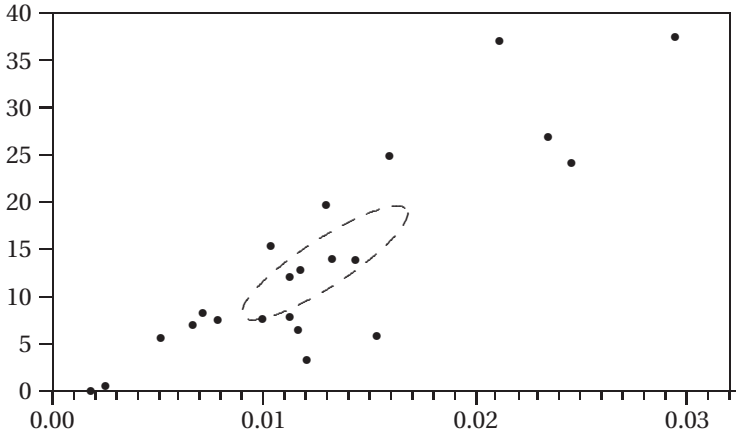


Abb. 7.5: *Schiffsbewegung und Reisekrankheit*

Das in 7.2 behandelte Konzept der empirischen Likelihood-Konfidenzintervalle lässt sich leicht auf die Situation mehrdimensionaler Formparameter erweitern. Die Basis dafür bildet eine Verallgemeinerung des Satzes 7.2.6.

Satz 7.4.3 *asymptotische Verteilung des Loglikelihoodquotienten*

Sei $\boldsymbol{X}=(X_1,\ldots,X_n)$ eine Zufallsstichprobe aus einer Dichte $f(x;\boldsymbol{\vartheta})$. $\boldsymbol{\vartheta}\in\Theta$ sei ein r-dimensionaler Parameter und Θ sei eine offene Teilmenge von $\mathbb{R}^r$. $f(x;\boldsymbol{\vartheta})$ sei hinreichend regulär, so dass insbesondere der Schätzer $\hat{\boldsymbol{\vartheta}}$ asymptotisch multivariat normalverteilt ist. Ist $\boldsymbol{\vartheta}_0$ der wahre Parametervektor, so ist

$$-2\ln(\ell(\boldsymbol{\vartheta}_0;\boldsymbol{X}))=-2\ln\left(\frac{f(X_1;\boldsymbol{\vartheta}_0)\cdot\ldots\cdot f(X_n;\boldsymbol{\vartheta}_0)}{f(X_1;\hat{\boldsymbol{\vartheta}}_0)\cdot\ldots\cdot f(X_n;\hat{\boldsymbol{\vartheta}}_0)}\right)$$

approximativ χ^2_r verteilt.

Beweis: Siehe Rao (1973, S. 350f.)

Wie im eindimensionalen Fall erhalten wir *empirische Likelihood-Konfidenzregionen* der Form

$$R_{1-\alpha} = \{\vartheta | -2\ln(\ell(\vartheta; \boldsymbol{X})) \leq \chi^2_{r;1-\alpha}\}. \tag{7.16}$$

Der große Vorteil dieser Konfidenzintervalle gegenüber denen, die auf der asymptotischen Normalverteilung beruhen, ist in der Tatsache zu sehen, dass die Gestalt der Konfidenzbereiche nur von den Daten selbst abhängt und nicht von vornherein die Form eines Ellipsoids aufweist.

Beispiel 7.4.4 *Überschreitungsgeschwindigkeit und Alter*

Im Rahmen einer umfangreichen Erhebung wurden verschiedene Charakteristika von Verkehrssündern festgestellt. Bei 40 zufällig ausgwählten Fahrern ergab sich der in Abbildung 7.6 dargestellte Zusammenhang von Alter X (in Jahren) und Überschreitungsgeschwindigkeit Y (in km/h). Das Alter ist durch die Volljährigkeitsgrenze nach unten begrenzt. Die Überschreitungsgeschwindigkeit betrug mindestens 20[km/h], da bei geringerer Überschreitung keine Anzeige erstattet wurde.

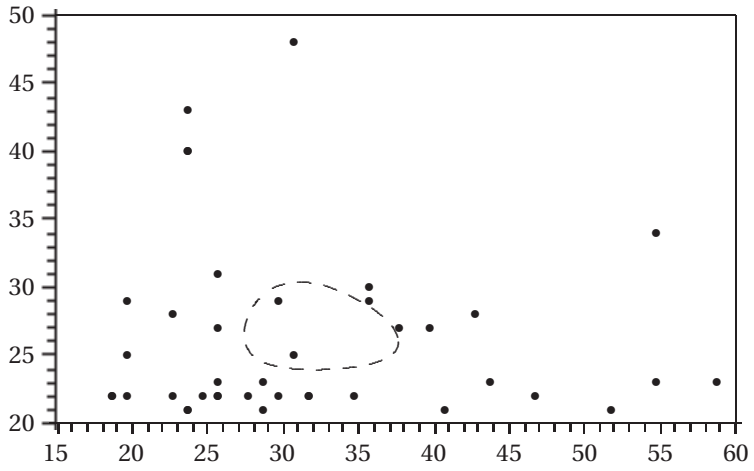


Abb. 7.6: *Alter und Geschwindigkeitsüberschreitung*

Der 95%-Konfidenzbereich auf der Basis der empirischen Likelihoodfunktion hat dementsprechend eine von der Ellipsenform abweichende Gestalt.

Neben Konfidenzschätzungen für ein- oder mehrdimensionale Parameter gibt es auch die Möglichkeit, einen Konfidenzbereich für eine ganze Verteilungsfunktion anzugeben. Grundlage hierfür ist der folgende Satz.

Satz 7.4.5 *Satz von Kolmogorov*

Ist die Verteilungsfunktion F der Zufallsvariablen X stetig, so gilt

$$\lim_{n\to\infty} \mathrm{P}(\sqrt{n} \sup_{-\infty<x<+\infty} |\hat{F}_n(x) - F(x)| \leq z) = H(z) \tag{7.17}$$

mit

$$H(z) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2z^2} & \text{für } z > 0 \\ 0 & \text{für } \leq 0. \end{cases}$$

Beweis: Siehe Fisz (1978, S. 462f.) für eine Beweisskizze und Sen & Singer (1993, Kapitel 8) für eine detailliertere Darstellung.

Einige ausgewählte Quantile der Grenzverteilung sind wie folgt:

z	1.00	1.07	1.22	1.36	1.63
$H(z)$	0.73	0.80	0.90	0.95	0.99

Die Statistik $D_n = \sqrt{n} \sup_{-\infty<x<+\infty} |\hat{F}_n(x) - F(x)|$ kann als Pivot angesehen werden, da $D_n \leq z$ gleichwertig damit ist, dass für alle $x \in \mathbb{R}$ simultan gilt:

$$\max\left\{0, \hat{F}_n(x) - \frac{z}{\sqrt{n}}\right\} \leq F(x) \leq \min\left\{1, \hat{F}_n(x) + \frac{z}{\sqrt{n}}\right\}.$$

Folglich schließen $\hat{F}_n(x) - z_{1-\alpha}/\sqrt{n}$ und $\hat{F}_n(x) + z_{1-\alpha}/\sqrt{n}$ mit Wahrscheinlichkeit $1-\alpha$ die gesamte Verteilungsfunktion ein.

Beispiel 7.4.6 *Wasserstandsdifferenzen - Fortsetzung von Seite 87*

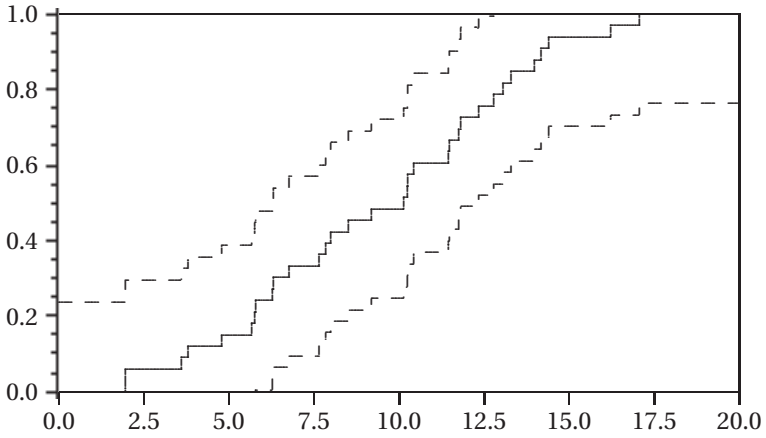


Abb. 7.7: *Konfidenzband der Verteilungsfunktion*

Für die Wasserstandsdifferenzen zu Abbildung 3.10 erhalten wir das in der Abbildung 7.7 illustrierte 95%-Konfidenzband

$$\max\left\{0, \hat{F}_n(x) - \frac{1.36}{\sqrt{33}}\right\} \leq F(x) \leq \min\left\{1, \hat{F}_n(x) + \frac{1.36}{\sqrt{33}}\right\}.$$

7.5 Aufgaben

Aufgabe 1

Die Versicherungsschäden eines bestimmten Schadentypes folgen der von dem Parameter ϑ abhängigen Pareto-Verteilung:

$$f_\vartheta(x) = \vartheta k^\vartheta x^{-(\vartheta+1)} \quad \text{für } x > k, \quad f_\vartheta(x) = 0 \text{ sonst.}$$

Dabei ist $\vartheta > 0$.

Überprüfen Sie mittels eines 0.95-Konfidenzbandes, ob die folgenden, durch Hurrikane in den USA verursachten Schäden (in Mio \$), die 6 Mio \$ übersteigen, als Zufallsstichprobe aus einer Pareto-Verteilung mit $k = 6$, $\vartheta = 0.4$ angesehen werden können.

6.7	7.1	10.6	14.5	15.4	17.0	18.4	19.0	25.3	29.1	30.1
33.7	40.6	41.4	47.9	49.4	52.6	59.9	63.1	77.8	102.9	103.2
123.7	140.1	192.0	198.4	227.3	329.5	428.7	513.6	545.8	750.4	863.9
1638.0										

8 Grundzüge der Testtheorie

8.1 Grundlegende Definitionen

8.1.1 Das Testproblem

Beim Schätzen von Parametern wird davon ausgegangen, dass die Verteilung der Zufallsvariablen X einem bestimmten Verteilungstyp angehört. Dann wird die tatsächliche Verteilung schließlich durch einen Parameter spezifiziert. Somit ist $\mathscr{P} = \{P_\vartheta | \vartheta \in \Theta\}$ die Menge der möglichen Verteilungen für X und ϑ der interessierende Parameter, der die richtige Verteilung festlegt.
Beim Testen gehen wir ebenfalls von dieser Situation aus. Wir haben hier aber noch weitere Vermutungen oder Angaben über den Parameter ϑ. Diese lassen sich dadurch beschreiben, dass ϑ in einer von zwei Teilmengen von Θ liegt, $\vartheta \in \Theta_0$ oder $\vartheta \in \Theta_1$, $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 \subseteq \Theta$. Ziel ist es dann zu entscheiden, welche der beiden Möglichkeiten zutrifft. Üblicherweise werden diese Möglichkeiten als *Hypothesen* bezeichnet, und zwar als *Nullhypothese* H_0 und als *Alternativ-* oder *Gegenhypothese* H_1:

$$H_0 : \vartheta \in \Theta_0, \qquad H_1 : \vartheta \in \Theta_1.$$

Das Testproblem lässt sich nun so formulieren, dass zwischen H_0 und H_1 entschieden werden soll. Es wird im Folgenden einfach durch das Hypothesenpaar angeben. Bei der Formulierung der Hypothesen wurde zugelassen, dass $\Theta_0 \cup \Theta_1$ nicht den ganzen Parameterraum aufspannt. Diese etwas größere Allgemeinheit erlaubt eine vernünftige Behandlung von praktisch relevanten Situationen, wo es Parameterbereiche geben kann, bei denen eine Indifferenz der Entscheidungen akzeptabel ist.

Für die Entscheidung zwischen den beiden Hypothesen ist eine Entscheidungsregel nötig. Eine solche Regel wird als *Test* bezeichnet. Ein Test gibt an, bei welchen Stichproben wir uns für die eine und bei welchen wir uns für die andere Hypothese entscheiden sollen. Die Menge der möglichen Stichproben ist also aufzuteilen in zwei disjunkte Teilmengen H und K. Erhalten wir eine Stichprobe aus dem *Annahmebereich* H, so wird H_0 akzeptiert. Bei einer Stichprobe aus K, dem *Ablehn-* oder *kritischen Bereich,* wird H_0 abgelehnt; die Entscheidung fällt für H_1. Anstelle der beiden Bereiche wird aus technischen Gründen einfach die Indikatorfunktion des kritischen Bereiches für die formale Definition eines Tests herangezogen.

Definition 8.1.1 *Testfunktion*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine Stichprobe aus der Verteilung P_ϑ, $\vartheta \in \Theta$. Gegeben sei das Testproblem

$$H_0 : \vartheta \in \Theta_0, \qquad H_1 : \vartheta \in \Theta_1.$$

Ein *Test* für H_0 gegen H_1 ist eine Partition des Stichprobenraumes, die durch eine *Testfunktion* $\varphi : \mathbb{R}^n \to \{0,1\}$ gegeben ist. φ wird auch selbst als Test bezeichnet. Entsprechend allgemeiner Vereinbarung wird $\varphi(\boldsymbol{x}) = 1$ mit der Entscheidung für H_1 gleichgesetzt.

In den meisten Anwendungen werden Annahme- und Ablehnbereich über eine *Teststatistik*, d.h. eine geeignete Stichprobenfunktion, festgelegt. Dann entsprechen oft Werte der Teststatistik T in einer Richtung (z.B. kleine t) Stichproben im Annahmebereich und die in der anderen (z.B. große t) solchen im Ablehnbereich.

Beispiel 8.1.2 *Test von $H_0 : \mu = \mu_0$ gegen $H_1 : \mu = \mu_1$ bei Normalverteilung*

Auf der Basis der Beobachtung einer Stichprobe $\boldsymbol{X} = (X_1, \ldots, X_n)$ aus einer $\mathcal{N}(\mu, 1)$-Verteilung soll zwischen folgenden Hypothesen entschieden werden:

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu = \mu_1 \qquad (\mu_0 < \mu_1).$$

Dazu sind der Annahmebereich H und der kritische Bereich K festzulegen. Da das arithmetische Mittel eine gute Schätzfunktion für μ ist, liegt es nahe, H_0 abzulehnen, wenn $\bar{X}$ zu groß ist bzw. beizubehalten, wenn $\bar{X}$ noch als verträglich mit H_0 angesehen wird. Entsprechend ist eine Konstante c zu wählen, so dass

$$H = \left\{ \boldsymbol{x} \,\middle|\, \frac{1}{n}\sum_{i=1}^{n} x_i \le c \right\} \quad \text{und} \quad K = \left\{ \boldsymbol{x} \,\middle|\, \frac{1}{n}\sum_{i=1}^{n} x_i > c \right\}.$$

Mit der Testfunktion formuliert lautet der Test:

$$\varphi(\boldsymbol{x}) = \begin{cases} 0 & \text{für } \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i \le c \\[2ex] 1 & \text{für } \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i > c \end{cases}.$$

Die im Beispiel auftauchende Konstante c wird als *kritischer Wert* bezeichnet. Der Test wird erst durch die Wahl des Schwellenwertes c für die Teststatistik endgültig festgelegt. Er trennt die Stichproben des Annahmebereiches von denen des Ablehnbereiches. Die Festlegung von c und generell des kritischen Bereiches geschieht unter dem Gesichtspunkt der Häufigkeit, mit der dann richtige Entscheidungen gefällt werden. Da Zufallsstichproben als Basis der Entscheidungen dienen, können wegen der Zufallsschwankungen ja auch Fehlentscheidungen vorkommen.

		Realität	
		μ_0	μ_1
Entscheidung	μ_0	richtige Entscheidung	Fehler 2. Art
	μ_1	Fehler 1. Art	richtige Entscheidung

Die Fehlentscheidung, H_0 abzulehnen, obwohl H_0 richtig ist, wird als *Fehler 1. Art* bezeichnet. Der *Fehler 2. Art* besteht darin, H_1 abzulehnen, obwohl H_1 richtig ist. Die Möglichkeiten richtiger und falscher Entscheidungen lassen sich entsprechend obigem Schema darstellen. Die Wahrscheinlichkeit, sich richtig zu entscheiden bzw. einen der beiden Fehler zu begehen, hängt nun von der Wahl des kritischen Wertes c ab. Der Zusammenhang der Fehlerwahrscheinlichkeiten wird im folgenden Beispiel deutlich gemacht.

Beispiel 8.1.3 *Test von $H_0: \mu = \mu_0$ gegen $H_1: \mu = \mu_1$ bei Normalverteilung - Fortsetzung von Seite 254*

Bei dem oben angegebenen Test für $H_0: \mu = \mu_0$ gegen $H_1: \mu = \mu_1$ auf der Basis einer Stichprobe vom Umfang n ist die Wahrscheinlichkeit für einen Fehler 1. Art bei gegebenem c:

$$P_{\mu_0}(\varphi(\boldsymbol{X}) = 1) = P_{\mu_0}(\bar{X} > c) = P_{\mu_0}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right).$$

Ein Fehler 2. Art wird mit der folgenden Wahrscheinlichkeit begangen:

$$P_{\mu_1}(\varphi(\boldsymbol{X}) = 0) = P_{\mu_1}(\bar{X} \leq c) = P_{\mu_1}\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} \leq \frac{c - \mu_1}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{c - \mu_1}{\sigma/\sqrt{n}}\right).$$

Seien konkret $n = 9$, $\sigma = 1$, $\mu_0 = 0$ und $\mu_1 = 1$. Dann ergibt sich die folgende Tabelle über den Zusammenhang von c und den Wahrscheinlichkeiten für die beiden Fehlerarten.

c	$P_{\mu_0}(\bar{X} > c)$	$P_{\mu_1}(\bar{X} \leq c)$
0.00	0.5	0.00
0.25	0.2265	0.0122
0.50	0.0665	0.0665
0.75	0.0122	0.2265
1.00	0.00	0.50

Die Tabelle lehrt, dass β umso größer ist, je kleiner α gewählt wird. Der Zusammenhang wird durch den kritischen Wert c hergestellt. Da c über die standardisierte Teststatistik $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ bestimmt wird, können wir die Testfunktion auch in folgender Form angeben:

$$\varphi(\boldsymbol{x}) = \begin{cases} 0 & \text{für } \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq c' \\ 1 & \text{für } \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > c' \end{cases}.$$

Das hat den Vorteil, dass c' direkt aus der Standardnormalverteilung bestimmt werden kann.

Wie das Beispiel verdeutlicht, können die Wahrscheinlichkeiten für die beiden Fehlerarten nicht gleichzeitig klein gemacht werden. Dies führt dazu, dass für den Fehler 1. Art eine obere Schranke angegeben wird. Diese wird als Irrtumswahrscheinlichkeit α bezeichnet. Übliche Werte für α sind etwa 0.05, 0.01, 0.001. Die Ablehnung von H_0 ist folglich eine relativ sichere Sache. Entweder ist es eine richtige Entscheidung, oder sie geschieht fälschlicherweise

nur mit einer kleinen Wahrscheinlichkeit. Wir werden sehen, wie es mit der Entscheidung für H_0 aussieht.

Häufig ist analog zum Vorgehen im letzten Beispiel die Standardisierung der Teststatistik notwendig, um den kritischen Bereich bestimmen zu können. Das Zusammenspiel von kritischem Wert und Wahrscheinlichkeit des Fehlers 1. Art führt ja dazu, dass die Verteilung der Teststatistik unter der Nullhypothese bekannt sein muss. Daher wird oft gleich die standardisierte Version einer Stichprobenfunktion als Teststatistik genommen.

Im Beispiel wurden durch H_0 und H_1 nur jeweils ein Parameter und damit eine Verteilung festgelegt. Beide Hypothesen waren also *einfach*, d.h. einelementig. Ist nach dem allgemeinen Sprachgebrauch H_0 *zusammengesetzt*, besteht Θ_0 also aus mehr als einem Parameter, so ist natürlich zu fordern, dass für kein $\vartheta \in \Theta_0$ die Irrtumswahrscheinlichkeit α überschritten wird.

Definition 8.1.4 *Test zum Niveau α, Umfang eines Tests*

$X_1, \ldots, X_n$ seien unabhängige identisch verteilte Zufallsvariable mit der Verteilung P_ϑ, $\vartheta \in \Theta$. Gegeben sei das Testproblem

$$H_0 : \vartheta \in \Theta_0, \qquad H_1 : \vartheta \in \Theta_1.$$

Ein Test $\varphi : \mathbb{R}^n \to \{0,1\}$ für H_0 gegen H_1 ist ein *Test zum Niveau α*, falls sein *Umfang* $\sup_{\vartheta \in \Theta_0} E_\vartheta \varphi(\boldsymbol{X})$ höchstens α beträgt:

$$\sup_{\vartheta \in \Theta_0} E_\vartheta \varphi(\boldsymbol{X}) \leq \alpha.$$

$K = \{\boldsymbol{x} | \varphi(\boldsymbol{x}) = 1\}$ heißt auch Ablehnbereich zum Niveau α für H_0 gegen H_1.

Der Fehler 1. Art wird durch die vorgegebene Irrtumswahrscheinlichkeit α unter Kontrolle gehalten. Es bleibt also, die Wahrscheinlichkeit für den Fehler 2. Art ins Visier zu nehmen, um Tests weiter zu untersuchen. Dabei wird es das Ziel sein, Tests zu finden, bei denen die oft mit β bezeichnete Wahrscheinlichkeit für einen solchen Fehler möglichst klein ist. Andersherum formuliert präferieren wir Tests mit einer großen Chance $1 - \beta$ für die richtige Entscheidung für H_1.

Die Wahrscheinlichkeit für einen Fehler 2. Art bei dem Testproblem

$$H_0 : \vartheta \in \Theta_0, \qquad H_1 : \vartheta \in \Theta_1$$

ist bei dem Test $\varphi : \mathbb{R}^n \to \{0,1\}$ gegeben durch

$$P_\vartheta(\varphi(\boldsymbol{X}) = 0) \qquad \vartheta \in \Theta_1.$$

Dafür können wir schreiben:

$$P_\vartheta(\varphi(\boldsymbol{X}) = 0) = 1 - P_\vartheta(\varphi(\boldsymbol{X}) = 1) = 1 - E_\vartheta \varphi(\boldsymbol{X}).$$

Ein Test sollte also den letzten Ausdruck für $\vartheta \in \Theta_1$ möglichst klein bzw. den Erwartungswert $E_\vartheta \varphi(\boldsymbol{X})$ möglichst groß machen. Dies ist die Wahrscheinlichkeit, sich für H_1 zu entscheiden, wenn $\vartheta \in \Theta$ der wahre Parameterwert ist. Für $\vartheta \in \Theta \backslash \Theta_0$ ist es die Wahrscheinlichkeit, Abweichungen von der Hypothese aufzudecken. Im Folgenden wird vor allem dieser Ausdruck betrachtet.

Definition 8.1.5 *Gütefunktion*

Der als Funktion von $\vartheta \in \Theta$ aufgefasste Erwartungswert der Testfunktion $\varphi(\boldsymbol{X})$,

$$\mathrm{E}_\vartheta \varphi(\boldsymbol{X}),$$

heißt *Gütefunktion* des Tests φ.

Die Gütefunktion ist das adäquate Instrument, um die Eigenschaften von statistischen Tests zu beschreiben und zu analysieren.

Beispiel 8.1.6 *einseitiges Testproblem für μ*

Wir modifizieren unser Ausgangsbeispiel und betrachten das einseitige Testproblem

$$H_0 : \mu \le \mu_0, \qquad H_1 : \mu > \mu_0.$$

Wir wollen sehen, inwieweit der oben angegebene Test sinnvoll bleibt. Wir betrachten also die Testfunktion

$$\varphi(\boldsymbol{x}) = \begin{cases} 0 & \text{für } \dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \le c \\ 1 & \text{für } \dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > c \end{cases}.$$

Für die Gütefunktion des Tests erhalten wir:

$$\begin{aligned} \mathrm{E}_\mu \varphi(\boldsymbol{X}) &= \mathrm{P}_\mu(\varphi(\boldsymbol{X}) = 1) = \mathrm{P}_\mu\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > c\right) = \mathrm{P}_\mu\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > c + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(c + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right). \end{aligned}$$

Wird c so bestimmt, dass $1-\Phi(c) = \alpha$, so folgt aus der Monotonie von $\Phi(\cdot)$, dass $\mathrm{E}_\mu \varphi(\boldsymbol{X}) \le$

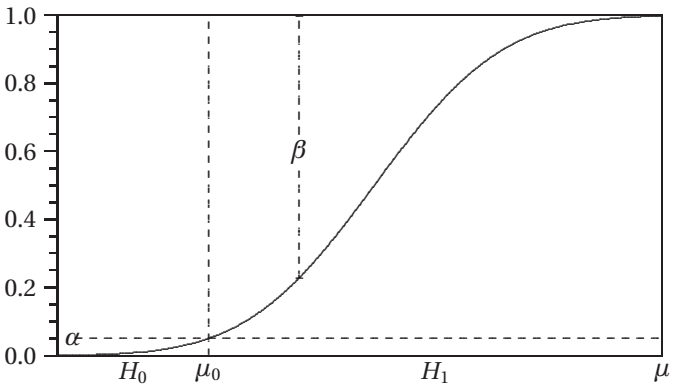


Abb. 8.1: *Gütefunktion*

α für alle $\mu \le \mu_0$. Der Test ist also ein Test zum Niveau α. In der Abbildung 8.1 ist die Gütefunktion für die konkreten Werte aus dem Beispiel 8.3.6 und $\alpha = 0.05$ dargestellt.

Die Darstellung der Gütefunktion in dem Beispiel macht deutlich, dass der Fehler zweiter Art groß sein kann, wenn der Parameter dicht bei den durch H_0 spezifizierten Werten liegt. Dementsprechend ist eine Entscheidung für H_0 stets wenig befriedigend. Es kann ja H_1 richtig sein, jedoch der Parameter sich nicht stark von μ_0 unterscheiden. Aus der Gütefunktion ist aber zu erkennen, dass der im Beispiel betrachtete Test *konsistent* ist, d.h. dass mit $n \to \infty$ für jeden durch H_1 festgelegten Parameterwert gilt: $\mathrm{E}_\mu \varphi(\boldsymbol{X}) \to 1$. Daher wird oft im Fall einer Entscheidung für H_0 gesagt, dass die Daten nicht ausgereicht hätten, um die Nullhypothese abzulehnen. Zu beachten ist dabei, dass zwischen einer *substanzwissenschaftlichen Relevanz* und einer *statistischen Signifikanz* unterschieden werden muss. Mit einem genügend großen Stichprobenumfang lässt sich bei Verwendung eines konsistenten Tests jede Abweichung von der Nullhypothese aufdecken. Das ist aber unter Anwendungsgesichtspunkten nicht unbedingt sinnvoll. Wie eingangs erwähnt, erlaubt man oft eine gewisse Indifferenz für Parameterwerte, die nahe an der Nullhypothese liegen. Dann wird beispielsweise bei einem Test der Hypothesen $H_0 : \vartheta \leq \vartheta_0, \quad H_1 : \vartheta > \vartheta_0$ ein Parameterwert $\vartheta_1 > \vartheta_0$ sowie eine Fehlerwahrscheinlichkeit β vorgegeben und der Stichprobenumfang n so bestimmt, dass $\mathrm{E}_{\vartheta_1} \varphi(\boldsymbol{X}) \geq 1 - \beta$.

8.1.2 Randomisierte Tests

Die Betrachtung der Gütefunktion im Beispiel lehrt weiter, dass die vorgegebene Irrtumswahrscheinlichkeit α ausgeschöpft wird; es ist speziell $\mathrm{E}_{\mu_0} \varphi(\boldsymbol{X}) = \alpha$. Bei diskreten Verteilungen ist es aber möglich, dass keine Aufteilung des Stichprobenraumes gefunden werden kann, bei der die Gütefunktion des zugehörigen Tests zumindest an einer Stelle die vorgegebene Grenze α erreicht. Solche Tests bezeichnen wir als *konservativ*; sie unterbieten das vorgegebene Niveau, führen folglich seltener zur fälschlichen Ablehnung der Nullhypothese als es aufgrund der Vorgaben erlaubt wäre. Bei einer stetig vom Parameter abhängigen Gütefunktion, ist dann aber auch die Wahrscheinlichkeit für den Fehler 2. Art größer als nötig. Wenn wir in solchen Fällen die Gütefunktion geeignet ‚anheben' könnten, würden wir einen besseren Test zum Niveau α erhalten. Dieses Anheben der Gütefunktion ist im Fall diskreter Verteilungen möglich, indem von der diskreten Teststatistik zu einer stetigen übergegangen wird. Dies soll im Rahmen eines Beispiels ausgeführt werden.

Beispiel 8.1.7 *Testproblem bzgl. des Parameters p bei Binomialverteilung*

Die Zufallsvariable X sei binomialverteilt mit $n = 10$. Für p liege das Testproblem $H_0 : p \leq p_0 = 0.1$, $H_1 : p > p_0 = 0.1$ vor, das zum Niveau $\alpha = 0.05$ getestet werden soll.

Die Entscheidung soll auf Basis einer Beobachtung von X getroffen werden. Da unter H_1 größere Werte als unter H_0 zu erwarten sind, wird H_0 abgelehnt, wenn ein zu großer Wert von X beobachtet wird. Genauer erhalten wir wegen

$$\sup_{p \leq p_0} \mathrm{P}_p(X > 3) = \mathrm{P}_{p_0}(X > 3) = 0.0128 \quad \text{und} \quad \sup_{p \leq p_0} \mathrm{P}_p(X > 2) = \mathrm{P}_{p_0}(X > 2) = 0.0702$$

den Ablehnbereich $\{4, 5, \ldots, 10\}$ mit der zugehörigen Testfunktion

$$\varphi_1(x) = \begin{cases} 1 & x > 3 \\ 0 & x \leq 3 \end{cases}.$$

Der Test hält zwar das Niveau α ein, schöpft es aber nicht aus. Um für das Testproblem einen Test zu erhalten, der das Niveau ausschöpft, gehen wir zu einer stetigen Zufallsvariablen Y über.

Sei R eine über dem Intervall $[0;1)$ gleichverteilte, von X unabhängige Zufallsvariable. (Realisationen von R erhalten wir z.B., indem wir zufällig eine Zahl aus einer Zufallszahlentafel auswählen. Diese Vorgehensweise sichert zugleich die Unabhängigkeit von X und R.)

Dann sei $Y := X + R$. Y ist eine stetige Zufallsvariable, deren Verteilung ebenso wie die von X vom Parameter p abhängt. Die Wahrscheinlichkeiten $\mathrm{P}_p(X = x)$ werden bei Y über die entsprechenden Intervalle $[x; x+1)$ verschmiert. Wegen $\mathrm{P}_p(X < 0) = \mathrm{P}_p(X > n) = 0$ gilt mit $\mathrm{P}(0 \le R < 1) = 1$ nämlich:

(i) $\mathrm{P}_p(Y < 0) = 0$,

(ii) $\mathrm{P}_p(Y \le n+1) = 1$.

Für $y \in [0, n+1)$ gilt mit $y = x + r$, $0 \le r < 1$:

$$\begin{aligned}\text{(iii)}\quad \mathrm{P}_p(Y \le y) &= \mathrm{P}_p(X + R \le x + r) = \mathrm{P}_p(X \le x-1, R \le 1) + \mathrm{P}_p(X = x, R \le r)\\ &= \mathrm{P}_p(X \le x-1) \cdot \mathrm{P}(R \le 1) + \mathrm{P}_p(X = x) \cdot \mathrm{P}(R \le r)\\ &= \mathrm{P}_p(X \le x-1) + r \cdot \mathrm{P}_p(X = x).\end{aligned}$$

Das obige Testproblem kann also auch bei Zugrundelegen von Y betrachtet werden. H_0 wird selbstverständlich wieder abgelehnt, wenn Y zu große Werte annimmt. Es ist also ein Wert c' zu bestimmen mit

$$\mathrm{P}_{p_0}(Y > c') = 0.05.$$

Wegen $\mathrm{P}_{p_0}(Y > 3) = \mathrm{P}_{p_0}(X > 2) = 0.0702$, $\mathrm{P}_{p_0}(Y > 4) = \mathrm{P}_{p_0}(X > 3) = 0.0128$ liegt c' zwischen 3 und 4. c' lässt sich bestimmen aus:

$$\begin{aligned}0.05 &= \mathrm{P}_{p_0}(Y > c') = 1 - \mathrm{P}_{p_0}(Y \le c') = 1 - \mathrm{P}_{p_0}(Y \le 3 + r')\\ &= 1 - [\mathrm{P}_{p_0}(X \le 2) + r' \cdot \mathrm{P}_{p_0}(X = 3)] = 1 - [0.93 + r' \cdot 0.0574]\\ &= 0.07 - r' \cdot 0.0574\,.\end{aligned}$$

Dies ergibt $r' = 0.35$. Folglich ist $c' = 3.35$ und der somit bestimmte Test lautet:

$$\varphi_2(y) = \begin{cases} 1 & y > 3.35 \\ 0 & y \le 3.35 \end{cases}.$$

Aufgrund der Beziehung $Y = X + R$ können die beiden Tests höchstens im Fall $x = 3$ zu unterschiedlichen Entscheidungen führen. Dies geschieht, wenn R einen Wert ≥ 0.35 annimmt. Werden X und R nicht gleichzeitig, sondern nacheinander beobachtet, so braucht das Randomisierungsexperiment 'Beobachten von R' nur bei Eintreten des Ereignisses $\{X = 3\}$ durchgeführt zu werden. H_0 wird dann noch bei Eintreten von $\{R > 0.35\}$ abgelehnt. Dies geschieht mit der Wahrscheinlichkeit $\mathrm{P}(R > 0.35) = 0.65 = \gamma$. Bei Eintreten von $\{X = 3\}$ ist zunächst also nur die Wahrscheinlichkeit bekannt, mit der H_0 abgelehnt wird. Wir erhalten wieder eine einheitliche Interpretation, wenn wir allgemein

$\varphi(x)$ als Wahrscheinlichkeit auffassen, mit der H_0 bei Beobachten des Wertes x abzulehnen ist. $\varphi(x)=1$ bedeutet dann, dass H_0 mit Wahrscheinlichkeit 1 abgelehnt wird. Dies ist gleichwertig mit der bisherigen Interpretation, bei der $\varphi(x)=1$ bedeutete: H_0 wird abgelehnt, da x im Ablehnbereich liegt. Entsprechend wird H_0 eben nicht abgelehnt, wenn die Wahrscheinlichkeit für das Ablehnen gleich null ist.

Mit dieser Interpretation kann der Test $\varphi_2(y)$ auch als Funktion von x angegeben werden:

$$\varphi_2(x)=\begin{cases} 1 & x>3 \\ \gamma=0.65 & x=3 \\ 0 & x<3 \end{cases}.$$

Gehen wir von der eben dargestellten Interpretation aus, so sprechen wir von randomisierten Tests.

Die Wahrscheinlichkeit, mit der H_0 bei Vorliegen von $p\in[0,1]$ abgelehnt wird, setzt sich bei einem randomisierten Test zusammen aus der Wahrscheinlichkeit $\mathrm{P}_p(X=x)$, die Beobachtung x zu erhalten, und der bedingten Wahrscheinlichkeit $\varphi(x)$, mit der H_0 beim Vorliegen der Beobachtung X abgelehnt wird. Dass beides eintritt, hat die Wahrscheinlichkeit $\varphi(x)\cdot\mathrm{P}_p(X=x)$.

Für die Wahrscheinlichkeit, irgendein x zu beobachten und dann jeweils H_0 abzulehnen, erhalten wir

$$\sum_{x=0}^{n}\varphi(x)\cdot\mathrm{P}_p(X=x)=\mathrm{E}_p\varphi(x).$$

Damit lässt sich die Gütefunktion des randomisierten Tests wieder als Erwartungswert in Abhängigkeit vom Parameterwert darstellen. Die Abbildung 8.2 zeigt, dass das Randomisieren einen beträchtlichen Gewinn bzgl. der Güte bringen kann.

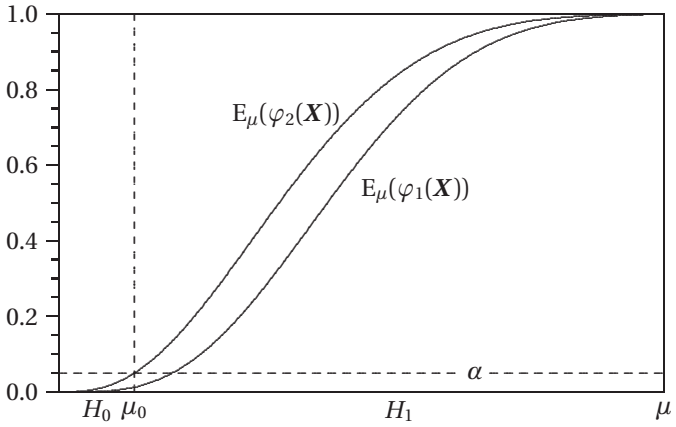


Abb. 8.2: *Gütevergleich zweier Tests*

In Verallgemeinerung der Diskussion des Beispiels können wir jede Funktion, die nur Werte zwischen Null und Eins annehmen kann, als Test ansehen. Wenn nur die Werte Null und Eins möglich sind, liegt ein nichtrandomisierter Test vor.

Definition 8.1.8 *randomisierter Test*

$X_1,\ldots,X_n$ seien unabhängige identisch verteilte Zufallsvariablen mit der Verteilung P_ϑ, $\vartheta\in\Theta$. Ein *(randomisierter) Test zum Niveau* α für das Testproblem

$$H_0:\ \vartheta\in\Theta_0,\qquad H_1:\ \vartheta\in\Theta\backslash\Theta_0$$

ist eine Funktion $\varphi:\mathbb{R}^n\to[0,1]$, für die gilt:

$$E_\vartheta\varphi(\boldsymbol{X})\le\alpha\qquad\text{für alle }\vartheta\in\Theta_0.$$

$\varphi(\boldsymbol{x})$ ist dabei die Wahrscheinlichkeit, mit der H_0 bei Beobachten der Stichprobe $\boldsymbol{x}$ abgelehnt wird. Die Gütefunktion eines randomisierten Tests φ ist der als Funktion von $\vartheta\in\Theta$ aufgefasste Erwartungswert $E_\vartheta\varphi(\boldsymbol{X})$.

Beispiel 8.1.9 *ein spezieller randomisierter Test*

Gegeben sei eine Stichprobe $\boldsymbol{X}=(X_1,\ldots,X_n)$ aus der Verteilung P_ϑ, $\vartheta\in\Theta$. Ein randomisierter Test für das Testproblem $H_0:\ \vartheta\in\Theta_0$; $H_1:\ \vartheta\in\Theta_1$ ist dann gegeben durch:

$$\varphi(\boldsymbol{x})\equiv\alpha.$$

Egal, welche Beobachtung gemacht wird, man entscheidet sich aufgrund eines Randomisierungsexperimentes zwischen H_0 und H_1. Dieses führt dann mit Wahrscheinlichkeit α zur Ablehnung von H_0 und mit Wahrscheinlichkeit $1-\alpha$ zur Entscheidung, H_0 beizubehalten. Wegen

$$E_\vartheta\varphi(\boldsymbol{X})=\alpha\qquad\text{für alle }\vartheta\in\Theta$$

ist auch die Gütefunktion konstant gleich α.

Im Folgenden werden randomisierte Tests nur noch ausnahmsweise betrachtet. Der Grund dafür ist, dass das Vorgehen der Randomisierung von den Anwendern weitestgehend abgelehnt wird. Das ist ja auch verständlich. Die gleichen Daten können einmal zur Ablehnung und zum anderen Mal zur Beibehaltung der Nullhypothese führen! Diese Ablehnung der randomisierten Tests hat nichts damit zu tun, dass ihre Konstruktion dem 'theoretischen Überbau' nicht entspräche. Das Problem liegt vielmehr in der von den Anwendern geforderten eindeutigen Reproduzierbarkeit der Ergebnisse.

8.1.3 Einige gängige Hypothesen

Liegt eine *parametrische Verteilungsfamilie* vor und ist der interessierende Parameter eindimensional, $\Theta\subseteq\mathbb{R}$, so sprechen wir von *einseitigen Testproblemen*, wenn Θ_0 und Θ_1 Intervalle sind, $\Theta_0=(a,\vartheta_0],\Theta_1=(\vartheta_0,b)$ bzw. $\Theta_0=[\vartheta_0,b),\Theta_1=(a,\vartheta_0)$. Gegebenenfalls werden die Hypothesen in der Form

$$H_0:\ \vartheta\le\vartheta_0,\ H_1:\ \vartheta>\vartheta_0 \tag{8.1}$$

bzw.

$$H_0:\ \vartheta\ge\vartheta_0,\ H_1:\ \vartheta<\vartheta_0 \tag{8.2}$$

angegeben. *Zweiseitige Testprobleme* entsprechen einer Partition von Θ gemäß

$$H_0:\ \vartheta=\vartheta_0,\ H_1:\ \vartheta\ne\vartheta_0. \tag{8.3}$$

Bei mehrdimensionalen Parametern wird oft nur eine eindimensionale Funktion betrachtet. Auch dann sprechen wir ggf. von einseitigen bzw. zweiseitigen Testproblemen. Für die Testkonstruktion ist natürlich der gesamte Parameterraum zu berücksichtigen.

Beispiel 8.1.10 *drei Testpobleme für μ*

Sei $\boldsymbol{X} = (X_1, \ldots, X_m)$ eine Stichprobe aus einer $\mathcal{N}(\mu, \sigma^2)$-Verteilung. Ist σ^2 unbekannt, so sind die Hypothesen der drei Testpobleme für μ zusammengesetzt:

$$\begin{aligned} &H_0 : \mu \le \mu_0, \quad \sigma^2 > 0 \qquad H_1 : \mu > \mu_0, \quad \sigma^2 > 0; \\ &H_0 : \mu \ge \mu_0, \quad \sigma^2 > 0 \qquad H_1 : \mu < \mu_0, \quad \sigma^2 > 0; \\ &H_0 : \mu = \mu_0, \quad \sigma^2 > 0 \qquad H_1 : \mu \ne \mu_0, \quad \sigma^2 > 0. \end{aligned}$$

Da nur der Lageparameter μ interessiert, sind die ersten beiden Testprobleme einseitig und das dritte ist zweiseitig.

Es ist vom Ansatz statistischer Tests her im Prinzip unmöglich, die Gleichheit eines Parameters mit einem vorgegebenem Wert oder die Gleichheit zweier Parameter statistisch 'nachzuweisen'. Dazu wären die Hypothesen im Einstichprobenfall von der Form

$$H_0 : \vartheta \ne \vartheta_0, \qquad H_1 : \vartheta = \vartheta_0$$

zu wählen, da die Ablehnung von H_0 die als zuverlässig angesehene Entscheidung darstellt. Dass es für dieses Testproblem i.d.R. keinen sinnvollen Test gibt, illustriert das folgende Beispiel.

Beispiel 8.1.11 *zweiseitiger Test für μ*

Sei $\boldsymbol{X} = (X_1, \ldots, X_m)$ eine Stichprobe aus einer $\mathcal{N}(\mu, \sigma^2)$-Verteilung. σ^2 sei bekannt. Das Testproblem sei

$$H_0 : \mu \ne \mu_0, \qquad H_1 : \mu = \mu_0.$$

Ein naheliegender Test basiert auf der Teststatistik $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ und hat die Gestalt

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{falls } -c < \sqrt{n}(\bar{X} - \mu_0)/\sigma < c \\ 0 & \text{sonst} \end{cases}.$$

Damit φ ein Test zum Niveau α wird, muss gelten:

$$\mathrm{E}_\mu \varphi(\boldsymbol{X}) = \mathrm{P}_\mu(-c < \sqrt{n}(\bar{X} - \mu_0)/\sigma < c) \le \alpha \qquad (\mu \ne \mu_0).$$

Es ist leicht zu sehen, dass dies nur erreicht wird, wenn c so gewählt wird, dass

$$\mathrm{P}_{\mu_0}(-c < \sqrt{n}(\bar{X} - \mu_0)/\sigma < c) = \alpha,$$

wenn also $c = z_{0.5+\alpha/2}$. Dann ist φ zwar ein Test zum Niveau α, jedoch ist α auch die Chance, sich für H_1 zu entscheiden, wenn H_1 richtig ist. Wir könnten uns gleich auf der Basis eines Würfelexperimentes entscheiden!

In manchen Anwendungen möchte man trotz der eben dargestellten Problematik das Vorliegen eines Parameterwertes oder die Gleichheit der Parameter der Verteilungen zweier unabhängiger Stichproben nachweisen. Das gilt etwa für Bioäquivalenzstudien, in denen man sich für die Wirkung eines Präparates bzw. den biologischen Abbau im Blut interessiert. Das Problem löst man dann so, dass unter H_1 nicht nur ein Wert sondern ein geeignetes Intervall spezifiziert wird. Die Gleichheit wird also in eine ungefähre Übereinstimmung aufgelöst. Zum Beispiel tritt an die Stelle von $H_0 : \mu \neq \mu_0$, $H_1 : \mu = \mu_0$ das Testproblem

$$H_0 : \mu \notin [\mu_0 - \Delta_1, \mu_0 + \Delta_2], \qquad H_1 : \mu \in [\mu_0 - \Delta_1, \mu_0 + \Delta_2]. \tag{8.4}$$

Für nichtparametrische Verteilungsfamilien sind im Wesentlichen Hypothesen bzgl. der Lage von Interesse. Die Lage betreffende Hypothesen beziehen sich im Rahmen der bisher betrachten einfachen Stichproben auf den Median. Bei solch allgemeinen Verteilungsklassen ist die Existenz des Erwartungswertes ja nicht ohne Weiteres gegeben.

Wir bezeichnen mit $\mathscr{F}$ die Klasse aller stetigen eindimensionalen Verteilungsfunktionen F und mit $\tilde{\mu}_F$ den zu F gehörigen Median. Dann lautet das zweiseitige Testproblem

$$H_0 : \tilde{\mu}_F = \tilde{\mu}_0, \quad F \in \mathscr{F}; \qquad H_1 : \tilde{\mu}_F \neq \tilde{\mu}_0, \quad F \in \mathscr{F}. \tag{8.5}$$

Daneben spielt für einfache Stichproben die Hypothese der Symmetrie um Null noch eine große Rolle. Diese resultiert aus der folgenden experimentellen Situation zum Vergleich zweier Behandlungsmethoden, Verfahren oder Ähnlichem. Bei gleichartigen Paaren von Versuchseinheiten werden jeweils Beobachtungen X_i bzw. Y_i $(i = 1, \ldots, n)$ gemacht. Die paarweisen Differenzen $X_i - Y_i$ werden verwendet, um Aussagen über die Unterschiede der Behandlungen zu machen.

Beispiel 8.1.12 *Steigerung der Lernfähigkeit*

Zwei Behandlungsmethoden A und B zur Steigerung der Lernfähigkeit sollen miteinander verglichen werden. Dazu werden 10 Paare von Schülern zufällig ausgewählt, die sich jeweils bzgl. der sozialen Stellung, der schulischen Leistung etc. möglichst stark gleichen. Jeder Paarling wird einer der beiden Behandlungen unterzogen. Am Ende wird der Lernerfolg mittels eines geeigneten Vorgehens festgestellt.

Paarnr.	1	2	3	4	5	6	7	8	9	10
Behandlung A (X)	72.6	70.1	66.0	95.0	57.2	60.8	83.3	67.3	68.8	70.0
Behandlung B (Y)	76.0	73.1	70.5	94.0	60.1	58.5	81.1	72.1	73.0	72.0
Differenz $A-B$ $(X-Y)$	-3.4	-3.0	-4.5	+1.0	-2.9	+2.3	+2.2	-4.8	-4.	-2.0

Speziell interessiert, ob die zugrundeliegenden Verteilungen von X und Y die gleiche Lage haben. Sofern die Verteilungen identisch sind, ist die Differenz von $X - Y$ symmetrisch um Null verteilt. Mit der Unabhängigkeit folgt nämlich sofort:

$$\mathrm{P}(X - Y \leq a) = \mathrm{P}(Y - X \leq a) = \mathrm{P}(X - Y \geq -a).$$

Für die Verteilungsfunktion von $Z = X - Y$ bedeutet das:

$$F_Z(-z) = 1 - F_Z(z) \qquad z \in \mathbb{R}.$$

Ist $\mathscr{F}_s = \{F | F \in \mathscr{F},\ F(\tilde{\mu} - x) + F(\tilde{\mu} + x) = 1, x \in \mathbb{R}, \tilde{\mu} \in \mathbb{R}\}$, so lautet das interessierende Testproblem:

$$H_0 : \tilde{\mu} = 0 \text{ und } F \in \mathscr{F}_s; \qquad H_1 : \tilde{\mu} \neq 0 \text{ oder } F \in \mathscr{F} \setminus \mathscr{F}_s. \tag{8.6}$$

Für das nichtparametrische Streuungsproblem gibt es nur den Ansatz, Differenzen von Quantilen, speziell den Quartilsabstand $q_{0.75} - q_{0.25}$, zu betrachten.

Eine weitere Fragestellung betrifft das Vorliegen einer speziellen Verteilung oder Verteilungsfamilie:

$$H_0 : F = F_0, \qquad H_1 : F \neq F_0,$$

bzw.

$$H_0 : F \in \mathscr{F}_0, \qquad H_1 : F \notin \mathscr{F}_0.$$

Häufig ist dies damit verbunden, dass überprüft werden soll, ob auf einen Datensatz weitere statistische Auswertungsverfahren, die gewisse Verteilungsvoraussetzungen haben, angewendet werden dürfen. Hier soll also eine Zufallsstichprobe für mehrere Tests oder Konfidenzintervalle herhalten. *Dies ist zunächst einmal nicht erlaubt.* Ein Eckpfeiler der Theorie der Konfidenzschätzung und der Tests ist es, dass jede Stichprobe nur für einen induktiven Schluss verwendet werden darf. Modifikationen, die dann doch mehrere Schlüsse erlauben, sprechen wir in Abschnitt 8.5 an.

Beim Vergleich von zwei oder mehr Stichproben handelt es sich im Rahmen parametrischer Verteilungsfamilien darum, Vorstellungen oder Angaben bzgl. der Relation eines Parameters in den beiden zugrundeliegenden Verteilungen zu überprüfen. Auch wenn wir im Folgenden nur ein Beispiel für den Lagevergleich angeben, sind in parametrischen Verteilungsfamilien die Vergleiche bzgl. aller Parameter möglich. Ihre Sinnhaftigkeit hängt vom Kontext ab.

Beispiel 8.1.13 *Vergleich von Erwartungswerten*

Sei $\boldsymbol{X} = (X_1, \ldots, X_m)$ eine Stichprobe aus einer $\mathscr{N}(\mu_1, \sigma_1^2)$-Verteilung und $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ eine aus einer $\mathscr{N}(\mu_2, \sigma_2^2)$-Verteilung. Die Testprobleme

$$\begin{aligned} H_0 &: \mu_1 \leq \mu_2, \qquad H_1 : \mu_1 > \mu_2 \\ H_0 &: \mu_1 \geq \mu_2, \qquad H_1 : \mu_1 < \mu_2 \\ H_0 &: \mu_1 = \mu_2, \qquad H_1 : \mu_1 \neq \mu_2 \end{aligned}$$

können bei bekannten sowie – in den meisten Fällen realistischer – unbekannten σ_1^2 und σ_2^2 betrachtet werden. Bei unbekannten Varianzen ist zu unterscheiden, ob beide als gleich angesehen werden können oder ob Ungleichheit zuzulassen ist. Einen 'guten' Test für den letzten Fall zu finden, war lange Zeit Gegenstand intensiver Forschung. Es handelt sich hierbei um das sogenannte *Behrens-Fisher-Problem*.

Für den nichtparametrischen Lagevergleich gehen wir von dem Vergleich der Erwartungswerte zweier normalverteilter Zufallsvariablen aus. Wenn die Varianzen als gleich angesehen werden, ist die Hypothese $H_0 : \mu_1 \leq \mu_2$ gleichwertig mit $F_1(x) > F_2(x)\,(x \in \mathbb{R})$, wenn F_1 und F_2 die zugehörigen Verteilungsfunktionen sind.

$\mu_1 < \mu_2$ bedeutet also, dass die aus F_1 stammenden Werte nicht nur im Mittel, sondern in der generellen Tendenz kleiner sind als die aus F_2. Das führt zu der folgenden Definition.

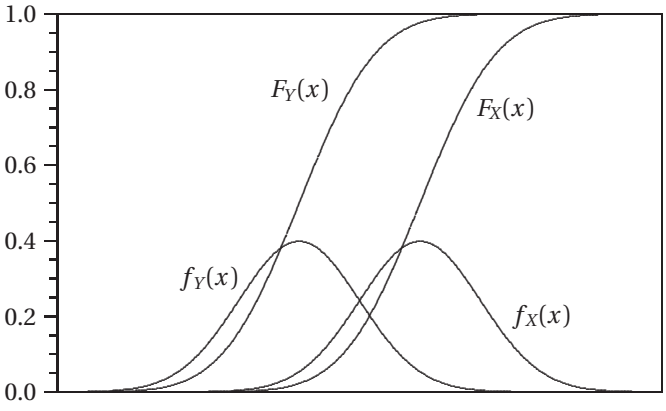


Abb. 8.3: *X stochastisch größer als Y*

Definition 8.1.14 *stochastische Größer-Relation*

Seien X und Y zwei Zufallsvariablen mit den Verteilungsfunktionen F_X und F_Y. X heißt *stochastisch größer* als Y, falls

$$F_X(x) \leq F_Y(x) \qquad x \in \mathbb{R}.$$

Dafür schreiben wir auch $F_X \leq F_Y$. $F_X < F_Y$ bedeutet $F_X(x) \leq F_Y(x)$ $(x \in \mathbb{R})$ und $F_X(x) < F_Y(x)$ für mindestens ein x.

Der Lagevergleich resultiert also in der Formulierung der Hypothesen

$$H_0 : F_X \leq F_Y, \quad (F_X, F_Y) \in \mathscr{G}, \qquad H_1 : F_X > F_Y, \quad (F_X, F_Y) \in \mathscr{G}; \tag{8.7}$$

dabei ist

$$\mathscr{G} = \{(F,G) | F,G \in \mathscr{F},\ F \leq G \text{ oder } G \leq F\}.$$

Den Vergleich zweier Mediane erhalten wir aus der Einschränkung der zugelassenen Verteilungen:

$$\begin{aligned} H_0 &: F_X(x) = F_Y(x+\Delta), \qquad \Delta \geq 0,\ F_Y \in \mathscr{F}; \\ H_1 &: F_X(x) = F_Y(x-\Delta), \qquad \Delta > 0,\ F_Y \in \mathscr{F}. \end{aligned} \tag{8.8}$$

8.1.4 Überschreitungswahrscheinlichkeiten

Der Festlegung des Niveaus, zu dem ein Test durchgeführt werden soll, haftet etwas Willkürliches an. Es ist zwar Konvention, für α die Werte 0.05, 0.01, 0.001 etc. zu wählen, aber letztlich liegt die Wahl in der Hand und im Belieben des Anwenders. Für die gleiche Situation können verschiedene Personen den gleichen Test durchaus zu unterschiedlichen Niveaus durchführen. Daher ist es inzwischen verbreitet, bei einer Anwendung nicht nur die Ablehnung der Nullhypothese zum vorgegebenen Niveau, sondern auch die *Überschreitungswahrscheinlichkeit*, auch ***P**-Wert* oder *empirisches Signifikanzniveau* genannt, mit zu berichten. (Veröffentlichungen ohne signifikantes Testergebnis sind sehr selten.) Die Überschreitungswahrscheinlichkeit ist definiert als das kleinste Signifikanzniveau, zu dem der

verwendete Test bei der vorliegenden Beobachtung $\boldsymbol{x}$ noch zur Ablehnung der Nullhypothese führt. In der Regel beruhen die praktisch verwendeten Tests auf Teststatistiken $T(\boldsymbol{X})$. Führen große Werte der Teststatistik zur Ablehnung von $H_0 : \vartheta \in \Theta_0$, so ist es bei bei festem Beobachtungsvektor $\boldsymbol{x}$:

$$p(\boldsymbol{x}) = \sup_{\vartheta \in \Theta_0} \mathrm{P}_\vartheta(T(\boldsymbol{X}) \geq t(\boldsymbol{x})). \tag{8.9}$$

Natürlich können wir die Überschreitungswahrscheinlichkeiten auch formal als Varianten zu den bisher betrachteten Tests betrachten. Dann sind sie jeweils mit den vorgegebenen Niveaus zu vergleichen. Ist eine Überschreitungswahrscheinlichkeit kleiner oder gleich dem vorgegebenem α, so führt der zugehörige Test zur Ablehnung der Nullhypothese.

Beispiel 8.1.15 *Test auf den Parameter einer Poisson-Verteilung*

Sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe aus einer Poisson-Verteilung. Für den Parameter λ der $\mathscr{P}(\lambda)$-Verteilung liege das Testproblem $H_0 : \lambda \leq \lambda_0$, $H_1 : \lambda > \lambda_0$ vor. Die Teststatistik $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ ist wieder Poisson-verteilt, und zwar bei Gültigkeit von λ_0 mit dem Parameter $n \cdot \lambda_0$. Mit dem Test

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & T(\boldsymbol{x}) > c \\ 0 & T(\boldsymbol{x}) \leq c \end{cases},$$

bei dem c festgelegt ist als kleinster Wert mit

$$\sum_{j=c+1}^{\infty} e^{-n\lambda_0} \frac{(n\lambda_0)^j}{j!} \leq \alpha,$$

korrespondiert der auf der Überschreitungswahrscheinlichkeit basierende Test:

$$\tilde{\varphi}(\boldsymbol{x}) = \begin{cases} 1 & \displaystyle\sum_{j=T(\boldsymbol{x})}^{\infty} e^{-n\lambda_0} \frac{(n\lambda_0)^j}{j!} \leq \alpha \\ 0 & \text{sonst} \end{cases}.$$

Die Verwendung von Überschreitungswahrscheinlichkeiten bringt in der Praxis verschiedene Vorteile mit sich:

- Viele Computerprogramme berechnen die p-Werte $p(\boldsymbol{x})$ und ermöglichen damit den direkten Vergleich mit dem selbstgewählten Signifikanzniveau α. Das Nachschlagen nach den kritischen Werten für die Prüfgrößen entfällt.
- Teststatistiken sind oft nicht direkt miteinander vergleichbar, weil sie verschiedene Verteilungen unter der Nullhypothese haben. Sind die zugrundeliegenden Verteilungen stetig, so sind die p-Werte zumindest für die ausgezeichneten Verteilungen, die zur Bestimmung der kritischen Werte herangezogen werden, rechteckverteilt über dem Intervall $[0,1]$. Sie sind also quasi standardisierte Versionen, die sich gut vergleichen lassen.

- Es ist aus der vorangegangenen Diskussion klar, dass sich die p-Werte gut zur Datenexploration eignen. An die Stelle der Signifikanz tritt dann die ‚Merkwürdigkeit' in dem Sinne, dass ein ‚signifikantes Ergebnis' die Aufmerksamkeit auf sich ziehen sollte, um über Hintergründe weiter zu spekulieren. (Aber nicht, um etwas als ‚statistisch bewiesen' zu berichten!)

8.2 Zur Konstruktion von Tests

Es sollen hier einige Vorgehensweisen für die Bestimmung von Ablehn- bzw. von Annahmebereichen aufgrund heuristischer Argumente besprochen werden. Diese Konstruktionen sind mehr oder weniger systematisch. Oft ist man bei konkreten Fragestellungen schon zufrieden, wenn man überhaupt für ein Testproblem eine entsprechende Entscheidungsregel findet.

8.2.1 Tests, die von Schätzfunktionen ausgehen

Eine naheliegende Möglichkeit, Tests zu konstruieren, beruht auf folgender Überlegung: Legt die Nullhypothese H_0 einen einzelnen Wert fest, d.h. $\vartheta_0 \in \Theta_0$ und ist $\hat{\vartheta}(\boldsymbol{X})$ eine geeignete Schätzfunktion für ϑ, so deuten große Unterschiede zwischen Schätzwerten $\hat{\vartheta}(\boldsymbol{x})$ und dem hypothetischen Wert ϑ_0 darauf hin, dass der wahre Parameterwert verschieden von ϑ_0 ist. Diese Überlegung ist im Einzelfall jeweils zu konkretisieren und entsprechend zu modifizieren, wenn Θ_0 aus mehr als einem Parameterwert besteht.

Beispiel 8.2.1 *Test von $H_0: \mu = \mu_0$ gegen $H_1: \mu = \mu_1$ bei Normalverteilung - Fortsetzung von Seite 254*

Im Beispiel 8.1.2 wurde der Ablehnbereich des Tests über die Schätzfunktion $\bar{X}$ festgelegt. Bei der Bestimmung von c wurde die Beziehung

$$\mathrm{P}_{\mu_0}(\bar{X} > c) = \mathrm{P}_{\mu_0}\left(\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} > c'\right) = \alpha$$

ausgenutzt. Die Zufallsvariable $\sqrt{n}(\bar{X}-\mu_0)/\sigma$ ist unter H_0 standardnormalverteilt, so dass $c' = \sqrt{n}(\bar{X}-\mu_0)/\sigma$ aus der entsprechenden Tabelle bestimmt werden konnte.
Bei der Stichprobenfunktion $\sqrt{n}(\bar{X}-\mu_0)/\sigma$ geht der Vergleich zwischen Schätzwert und hypothetischem Wert direkt ein. Da außerdem ihre Verteilung unter H_0 bekannt und tabelliert ist, wird sie als Teststatistik genommen.

Beispiel 8.2.2 *Test für σ^2 bei Normalverteilung*

Neben der Frage, ob bei der Produktion ein Normwert im Mittel eingehalten wird, interessiert man sich auch dafür, ob die Abweichung von diesem Wert nicht zu groß wird. Dies führt auf Tests für Skalenparameter.

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine Zufallsstichprobe aus einer Normalverteilung mit unbekanntem Erwartungswert μ. Es liege als Testproblem vor

$$H_0: \sigma^2 \le \sigma_0^2, \qquad H_1: \sigma^2 > \sigma_0^2.$$

Zur Bestimmung eines geeigneten Ablehnbereiches zum Niveau α gehen wir von der Schätzfunktion $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ aus. Wegen $\mathrm{E}_{\sigma^2}(\hat{\sigma}^2) = \sigma^2$ wird H_0 abgelehnt, wenn $\hat{\sigma}^2$ zu große Werte annimmt. Um den zu der vorgegebenen Irrtumswahrscheinlichkeit α gehörigen kritischen Wert zu bestimmen, wird die Verteilung von $\hat{\sigma}^2$ bzw. die Verteilung einer aus $\hat{\sigma}^2$ gebildeten Prüfgröße unter H_0 benötigt.

$(n-1)\cdot\hat{\sigma}^2/\sigma_0^2$ ist χ^2-verteilt mit $n-1$ Freiheitsgraden. Zudem nimmt diese Statistik genau dann große Werte an, wenn $\hat{\sigma}^2$ dies tut. Ist die Varianz der X_i gleich dem hypothetischen Wert σ_0^2, so ist die Verteilung von $(n-1)\cdot\hat{\sigma}^2/\sigma_0^2$ bekannt und tabelliert. Somit kann ein kritischer Wert c mit

$$\mathrm{P}_{\sigma_0^2}\left(\frac{n-1}{\sigma_0^2}\hat{\sigma}^2 > c\right) = \alpha$$

aus der $\chi^2(n-1)$-Tabelle bestimmt werden.
Die Nullhypothese legt hier aber nicht nur einen Wert fest, sondern mehrere. Bevor $(n-1)\hat{\sigma}^2/\sigma_0^2$ als Prüfgröße etabliert werden kann, ist also noch nachzuprüfen, ob das vorgegebene Niveau eingehalten wird, d.h. ob für $\sigma^2 \le \sigma_0^2$ ebenfalls gilt:

$$\mathrm{P}_{\sigma^2}\left(\frac{n-1}{\sigma_0^2}\hat{\sigma}^2 > c\right) \le \alpha.$$

Wegen $\sigma_0^2/\sigma^2 \ge 1$ folgt für die Parameter aus dem unter H_0 festgelegten Bereich:

$$\mathrm{P}_{\sigma^2}\left(\frac{n-1}{\sigma_0^2}\hat{\sigma}^2 > c\right) = \mathrm{P}_{\sigma^2}\left(\frac{n-1}{\sigma^2}\hat{\sigma}^2 > \frac{\sigma_0^2}{\sigma^2}c\right) \le \mathrm{P}_{\sigma^2}\left(\frac{n-1}{\sigma^2}\hat{\sigma}^2 > c\right) = \alpha.$$

Das Niveau wird also eingehalten. Der kritische Bereich des Tests ist:

$$K = \left\{ \boldsymbol{x} \,\middle|\, \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_0^2} \ge c \right\}.$$

Beispiel 8.2.3 *Marktstellung*

Eine Firma untersucht, ob sich eine Marktstellung geändert hat. Bei der letzten Untersuchung wurde festgestellt, dass von dem interessierenden Produkt 25 % der verkauften Produkte von der Firma stammten.

Dieses Problem lässt sich formal folgendermaßen formulieren. Es wird eine Zufallsstichprobe $X_1, \ldots, X_n$ gezogen mit

$$X_i = \begin{cases} 1 & \text{Das verkaufte Produkt stammt von der Firma.} \\ 0 & \text{Das verkaufte Produkt stammt nicht von der Firma.} \end{cases}$$

Wenn $p = \mathrm{P}(X_i = 1)$ gesetzt wird, ist dann zu testen:

$$H_0 : p = 1/4, \qquad H_1 : p \neq 1/4.$$

$\sum_{i=1}^n X_i/n$ ist eine erwartungstreue Schätzfunktion für p. Da $\sum_{i=1}^n X_i$ binomialverteilt ist mit den Parametern n und p, verwenden wir diese Stichprobenfunktion als Teststatistik. Offensichtlich deuten zu kleine und zu große Werte von $\sum X_i$ darauf hin, dass H_0 falsch ist. Dementsprechend sind hier zwei kritische Werte c_1 und c_2 zu bestimmen, so dass

$$\mathrm{P}_{p_0}\left(\sum_{i=1}^n X_i < c_1\right) + \mathrm{P}_{p_0}\left(\sum_{i=1}^n X_i > c_2\right) = \alpha.$$

Nun gibt es zwei Probleme.
Zunächst hat $\sum X_i$ eine diskrete Verteilung, so dass es unter Umständen nicht möglich ist, c_1 und c_2 so zu wählen, dass oben die Wahrscheinlichkeiten zusammen genau den Wert α ergeben. Dieses Problem lässt sich lösen, indem c_1 und c_2 so gewählt werden, dass das Niveau eingehalten wird:

$$\mathrm{P}_{p_0}\left(\sum_{i=1}^n X_i < c_1\right) + \mathrm{P}_{p_0}\left(\sum_{i=1}^n X_i > c_2\right) \leq \alpha.$$

Unter Umständen wird es also nicht ausgeschöpft. Das zweite Problem besteht in der konkreten Wahl von c_1 und c_2. Da Abweichungen nach unten genauso schwer wiegen wie Abweichungen nach oben, wählt man schließlich c_1 und c_2 so, dass

$$\mathrm{P}_{p_0}\left(\sum_{i=1}^n X_i < c_1\right) \leq \frac{\alpha}{2}, \qquad \mathrm{P}_{p_0}\left(\sum_{i=1}^n X_i > c_2\right) \leq \frac{\alpha}{2}.$$

Es ist klar, dass die kritischen Werte so vom jeweiligen Rand so weit weg wie möglich liegen, ohne diese Bedingungen zu verletzen. Bei einem Stichprobenumfang von $n = 16$ erhalten wir, wenn $\alpha = 0.05$ bzw. $\alpha/2 = 0.025$ vorgegeben ist:

$$\left.\begin{aligned} \mathrm{P}_{1/4}\left(\sum_{i=1}^{16} X_i < 1\right) &= 0.0100 \\ \mathrm{P}_{1/4}\left(\sum_{i=1}^{16} X_i < 2\right) &= 0.0635 \end{aligned}\right\} \quad \Rightarrow \quad c_1 = 1 \quad \text{(maximales } c_1\text{)}$$

$$\left.\begin{aligned} \mathrm{P}_{1/4}\left(\sum_{i=1}^{16} X_i > 7\right) &= 0.0271 \\ \mathrm{P}_{1/4}\left(\sum_{i=1}^{16} X_i > 8\right) &= 0.0074 \end{aligned}\right\} \quad \Rightarrow \quad c_2 = 8 \quad \text{(minimales } c_2\text{)}$$

8.2.2 Tests und Konfidenzintervalle

Punktschätzer bilden einen naheliegenden Ausgangspunkt, um Tests zu bestimmen. Konfidenzintervalle können sogar direkt zum Testen von Hypothesen eingesetzt werden: Sei

$\mathscr{F} = \{F(\cdot;\vartheta)|\vartheta \in \Theta\}$ eine vom Parameter ϑ abhängige Verteilungsfamilie. $[T_1(\boldsymbol{X}), T_2(\boldsymbol{X})]$ sei ein $(1-\alpha)$-Konfidenzintervall für ϑ. Da das Konfidenzintervall den wahren Parameterwert mit einer Wahrscheinlichkeit $\geq 1-\alpha$ überdeckt, ist durch

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & T_1(\boldsymbol{x}) > \vartheta_0 \text{ oder } T_2(\boldsymbol{x}) < \vartheta_0 \\ 0 & T_1(\boldsymbol{x}) \leq \vartheta_0 \leq T_2(\boldsymbol{x}) \end{cases} \tag{8.10}$$

ein Test zum Niveau α für das Testproblem

$$H_0 : \vartheta = \vartheta_0, \qquad H_1 : \vartheta \neq \vartheta_0$$

gegeben.

Für einseitige Testprobleme können die untere bzw. die obere Konfidenzschranke herangezogen werden. Bei $H_0 : \vartheta \leq \vartheta_0, \quad H_1 : \vartheta > \vartheta_0$ ist die untere Konfidenzschranke $T_1(\boldsymbol{X})$ adäquat. Mit

$$\mathrm{P}_\vartheta(T_1(\boldsymbol{X}) \leq \vartheta) \geq 1-\alpha \qquad \vartheta \in \Theta$$

ist

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & T_1(\boldsymbol{x}) > \vartheta_0 \\ 0 & T_1(\boldsymbol{x}) \leq \vartheta_0 \end{cases}$$

ein Test zum Niveau α. Für die andersherum formulierten Hypothesen sind die oberen Konfidenzschranken zu nehmen.

Haben wir umgekehrt eine Schar von Tests $\varphi_{\vartheta_0}(\boldsymbol{X})$ zum Niveau α für die Testprobleme $H_0 : \vartheta = \vartheta_0$, $H_1 : \vartheta \neq \vartheta_0$, so ergibt sich ein Konfidenzbereich für ϑ aus

$$C(\boldsymbol{X}) = \{\vartheta' | \varphi_{\vartheta'}(\boldsymbol{X}) = 0\}.$$

Ist nämlich ϑ_0 der wahre Parameterwert, so gilt:

$$\mathrm{P}_{\vartheta_0}(C(\boldsymbol{X}) \ni \vartheta_0) = \mathrm{P}_{\vartheta_0}(\varphi_{\vartheta_0}(\boldsymbol{X}) = 0) \geq 1-\alpha.$$

Der realisierte Konfidenzbereich besteht also aus allen Parameterwerten ϑ', bei denen der zugehörige Test $\varphi_{\vartheta'}$ aufgrund der Stichprobe $\boldsymbol{x}$ nicht zur Ablehnung führt.

Konfidenzbereiche haben gegenüber Teststatistiken den Vorteil, dass man bei Ablehnung der Nullhypothese gleich über die Lage des unbekannten Parameterwertes informiert ist. Die Pivot-Methode legt eine zentrale Verbindung zwischen Konfidenzintervallen und standardisierten Teststatistiken offen. Letztere enthalten ja in der Regel einen unter H_0 festgelegten Parameter, sind spezielle Pivots. Natürlich gilt der angegebene Zusammenhang zwischen Tests und Konfidenzintervallen allgemein. Er führt nicht selten dazu, dass Konfidenzintervalle in den Lehrbüchern sehr am Rande oder sogar nur im Zusammenhang mit Tests behandelt werden.

Beispiel 8.2.4 *Marktstellung - Fortsetzung von Seite 268*

Betrachten wir das Beispiel 8.2.3 im Zusammenhang mit dem Beispiel 7.2.8, so sehen wir, dass hier die Äquivalenz des Tests auf eine Wahrscheinlichkeit mit dem über die statistische Methode gewonnenen Konfidenzintervall vorliegt.

Auch die in Abschnitt 7.2.2 betrachteten Konfidenzintervalle auf der Basis der Likelihoodfunktion führen sofort zu statistischen Tests. Da das damit verbundene Testprinzip aber von großer und weitreichender Bedeutung ist, soll es in einem eigenen Abschnitt behandelt werden.

8.2.3 Likelihood-Quotienten-Tests

Wir haben bereits in Abschnitt 7.2.2 ausgenutzt, dass der Likelihoodquotient

$$\ell(\vartheta;\boldsymbol{X}) = \frac{f(X_1;\vartheta)\cdot\ldots\cdot f(X_n;\vartheta)}{f(X_1;\hat{\vartheta})\cdot\ldots\cdot f(X_n;\hat{\vartheta})} \tag{8.11}$$

eine sinnvolle Basis für die Einschätzung der Nähe von bzw. des Unterschiedes zwischen dem Parameterwert ϑ und dem ML-Schätzer $\hat{\vartheta}$ darstellt. Daher liegt es nahe, darauf einen Test aufzubauen. Diese Methode, Tests zu konstruieren, wird zunächst im Rahmen eines Beispiels betrachtet.

Beispiel 8.2.5 *Marktstellung - Fortsetzung von Seite 270*

Wir gehen von der Situation des Beispiels 8.2.3 aus. Für das Testproblem $H_0 : p = p_0$, $H_1 : p \neq p_0$ bestimmen wir einen Ablehnbereich unter Verwendung des Likelihoodquotienten. Natürlich ist für den Parameter ϑ in (8.11) jetzt der unter H_0 spezifizierte Wert zu nehmen.

Bei einer Stichprobe $\boldsymbol{x} = (x_1,\ldots,x_n)$ ist der Likelihoodquotient, da $\sum x_i/n$ der Maximum-Likelihood-Schätzwert für p ist:

$$\ell(p_0;\boldsymbol{x}) = \frac{p_0^{\sum x_i}(1-p_0)^{n-\sum x_i}}{\left(\frac{1}{n}\sum x_i\right)^{\sum x_i}\left(1-\frac{1}{n}\sum x_i\right)^{n-\sum x_i}} = \frac{p_0^{\sum x_i}(1-p_0)^{n-\sum x_i}}{\max\limits_{0<p<1} p^{\sum x_i}(1-p)^{n-\sum x_i}}.$$

Über die allgemeinen Eigenschaften, die $\ell(p_0;\boldsymbol{x})$ als Teststatistik vernünftig erscheinen lassen, erhellt die folgende konkrete Interpretation diese Wahl.

Ist $H_0 : p = p_0$ richtig, so hat eine Stichprobe $\boldsymbol{x}$ mit $x = \sum x_i$ die Wahrscheinlichkeit $p_0^x(1-p_0)^{n-x}$. Diese wird mit der Wahrscheinlichkeit verglichen, die die Stichprobe bei anderen Parameterwerten p hat. Ist sie bei p_0 im Verhältnis zu den anderen Wahrscheinlichkeiten klein, so werden wir diese Stichprobe in der Regel entsprechend seltener unter H_0 als unter H_1 beobachten. Nun unterstellen wir wie schon bei der Maximum-Likelihood-Schätzung auch im Einzelfall, dass wir beobachten, was in der Regel häufiger eintritt. Konsequenter Weise akzeptieren wir einen Parameterwert als bessere Erklärung der Realität, wenn bei ihm das beobachtete Ergebnis häufiger eintritt als bei einem anderen. Folglich wird der Ablehnbereich aus den möglichen Stichproben gebildet, die im Verhältnis unter H_0 die geringste Chance gegenüber H_1 haben.

Für den Vergleich mit $p_0^x(1-p_0)^{n-x}$ kann das Maximum von $p^x(1-p)^{n-x}$ über alle $p \in (0,1)$ genommen werden. Dies hat den Vorteil, dass das Maximum dann unter Ausnutzung des ML-Schätzers bestimmt werden kann. Die Argumentation bleibt dabei immer noch gültig. Für $x = 0,1,\ldots,n$ wird also das Verhältnis $\ell(p_0;x)$ gebildet. Dann sind die

Werte von $\ell(p_0;x)$ der Größe nach zu ordnen und die Stichproben mit den kleinsten $\ell(p_0;x)$ zum Ablehnbereich zusammenzufassen.

Für die im Beispiel 8.2.3 angegebene Konstellation $n=16,\ p_0=1/4$ erhält man die in der Tabelle 8.1 angegebene Verteilung.

x	$\ell(p_0;x)$	$P_{p_0}(X=x)$	$P_{p_0}(\ell(p_0;x)\leq\ell(p_0;x))$
16	$2.3\cdot10^{-10}$	0.0000	0.0000
15	$3.0\cdot10^{-8}$	0.0000	0.0000
14	$8.7\cdot10^{-7}$	0.0000	0.0000
13	$1.4\cdot10^{-5}$	0.0000	0.0000
12	0.00015	0.0000	0.0000
11	0.00117	0.0002	0.0002
10	0.00877	0.0014	0.0016
0	0.01	0.0100	0.0116
9	0.029	0.0058	0.0174
8	0.10	0.0197	0.0371
1	0.14	0.0535	0.0906
7	0.25	0.0524	0.1430
2	0.5	0.1336	0.2766
6	0.5	0.1101	0.3867
3	1	0.2079	0.5946
5	1	0.1802	0.7748
4	1	0.2252	1.0000

Tabelle 8.1: *Verteilung des Likelihoodquotienten*

Da das vorgegebene Niveau $\alpha=0.05$ nicht überschritten werden darf, ergibt sich als Ablehnbereich

$$K'=\left\{(x_1,\ldots,x_{16})\,\middle|\,x_i\in\{0,1\},\ \sum x_i<1\quad\text{oder}\quad\sum x_i>7\right\}.$$

K' unterscheidet sich von dem im Beispiel 8.2.3 erhaltenen Ablehnbereich K. Es gilt: $K\subset K',\ K\neq K'$.

Im obigen Beispiel wurde ein Ablehnbereich mit Hilfe von $\ell(p_0;x)$ konstruiert. Es ist einsichtig, dass dieses Prinzip für alle Verteilungen mit einer Dichte anwendbar ist. Im stetigen Fall ist die Interpretation wie bei der Maximum-Likelihood-Schätzung natürlich nur noch näherungsweise gültig.
Für den Fall, dass die Nullhypothese aus mehr als einem Parameter besteht, können wir im Zähler das Maximum über alle $\vartheta\in\Theta$ bilden. Die Interpretation aus dem Beispiel bleibt dann im Wesentlichen erhalten. Anstelle des Maximums nehmen wir aus formalen Gründen jedoch das Supremum. (Das Maximum braucht nicht angenommen zu werden.) Als Teststatistik ergibt sich damit:

$$\ell(\boldsymbol{x})=\frac{\sup\limits_{\vartheta\in\Theta_0} f(\boldsymbol{x};\vartheta)}{\sup\limits_{\vartheta\in\Theta} f(\boldsymbol{x};\vartheta)}.$$

Definition 8.2.6 *Likelihood-Quotienten-Test*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine n-dimensionale Zufallsvariable mit der gemeinsamen Dichte $f(\boldsymbol{x}; \vartheta)$, $\vartheta \in \Theta$. Ein Test (zum Niveau α) mit einem Ablehnbereich K für das Testproblem

$$H_0 : \vartheta \in \Theta_0, \qquad H_1 : \vartheta \in \Theta \backslash \Theta_0,$$

der gegeben ist durch

$$K = \left\{ \boldsymbol{x} \;\middle|\; \ell(\boldsymbol{x}) = \frac{\sup\limits_{\vartheta \in \Theta_0} f(\boldsymbol{x}; \vartheta)}{\sup\limits_{\vartheta \in \Theta} f(\boldsymbol{x}; \vartheta)} < c \right\}, \tag{8.12}$$

heißt *Likelihood-Quotienten-Test* (zum Niveau α).

Auch wenn das Prinzip der Likelihood-Quotienten-Tests einleuchtend ist, kann im Einzelfall die Bestimmung des Ablehnbereichs problematisch sein, da es oft nicht leicht ist, die Verteilung von $\ell(\boldsymbol{X})$ zu bestimmen. Bisweilen lässt sich die Verteilung einer geeigneten Transformation von ℓ ermitteln. Dies wird im folgenden Beispiel vorgeführt.

Beispiel 8.2.7 *Likelihood-Quotienten-Test für μ*

Sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Zufallsstichprobe aus einer Normalverteilung mit den unbekannten Parametern μ und σ^2. Wir betrachten das Testproblem

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0.$$

Dafür soll ein Likelihood-Quotienten-Test zum Niveau α hergeleitet werden.

Es ist

$$\ell(\boldsymbol{x}) = \frac{\sup\limits_{\sigma^2>0} (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^n \frac{(x_i-\mu_0)^2}{\sigma^2}\right]}{\sup\limits_{\sigma^2>0,\mu} (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^n \frac{(x_i-\mu)^2}{\sigma^2}\right]} = \frac{(2\pi S^{*2})^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^n \frac{(x_i-\mu_0)^2}{S^{*2}}\right]}{(2\pi S^2)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^n \frac{(x_i-\bar{x})^2}{S^2}\right]}.$$

Dabei sind $S^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$ und $S^{*2} = \sum_{i=1}^n (x_i - \mu_0)^2/n$. Vereinfachen ergibt:

$$\ell(\boldsymbol{x}) = \left(\frac{S^2}{S^{*2}}\right)^{n/2},$$

bzw.

$$\begin{aligned}\ell(\boldsymbol{x})^{2/n} = \frac{S^2}{S^{*2}} &= \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{\sum_{i=1}^n (x_i-\mu_0)^2} = \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{\sum_{i=1}^n x_i^2 - 2\mu_0 \sum_{i=1}^n x_i + n\mu_0^2} \\ &= \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2 + n\bar{x}^2 - 2\mu_0 n\bar{x} + n\mu_0^2} = \frac{1}{1 + \dfrac{(\bar{x}-\mu_0)^2}{S^2}}.\end{aligned}$$

Daher ist die Relation $\ell(\boldsymbol{x}) > c$ gleichwertig mit $\sqrt{n-1}(\bar{x}-\mu_0)^2/S^2 > c'$ oder mit $\sqrt{c'} < \sqrt{n-1}|\bar{x}-\mu_0|/\sqrt{S^2}$. $\sqrt{n-1}(\bar{X}-\mu_0)/\sqrt{S^2}$ ist bekanntlich t-verteilt, falls μ_0 der wahre Parameterwert ist. Dementsprechend heißt der auf dieser Teststatistik basierende Test auch t-Test. Er ist hier also äquivalent zum Likelihood-Quotienten-Test.

Wir geben noch einen Likelihood-Quotienten-Test für ein speziell strukturiertes Modell, das einer einfachen Varianzanalyse mit der gleichen Anzahl von Beobachtungen auf jeder Stufe des Einflussfaktors entspricht, an. Dieses Beispiel ist exemplarisch für den Bereich der linearen Modelle. In gewisser Weise verallgemeinert das folgende Beispiel das letzte.

Beispiel 8.2.8 *Einfaktorielle ANOVA*

In einem Versuch werden k verschiedene Bedingungen vorgegeben. Es interessiert, ob die k Bedingungen alle die gleiche Auswirkung auf die Zielvariable haben. Zur Konkretion kann man sich verschiedene Formen der Werbung in unterschiedlichen Gebieten vorstellen, die sich möglicherweise auf den Umsatz eines speziellen Produktes auswirken.

Als Modellannahmen unterstellen wir unabhängige normalverteilte Zufallsvariablen X_{11}, $\ldots, X_{1m}, X_{21}, \ldots, X_{2m}, \ldots, X_{k1}, \ldots, X_{km}$, für die gelte:

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \qquad i = 1, \ldots, k,\ j = 1, \ldots, n.$$

Innerhalb einer Stufe sind also die Erwartungswerte gleich. Die unbekannten Varianzen werden generell als gleich vorausgesetzt. Als Testproblem ergibt sich:

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \ldots = \mu_k, \quad \sigma^2 > 0, \\ &H_1 : \text{Die Erwartungswerte sind nicht alle gleich,} \quad \sigma^2 > 0. \end{aligned}$$

Mit $\Theta = \{(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2) | -\infty < \mu_i < \infty,\ i = 1, \ldots, k,\ \sigma^2 > 0\}$ und
$\Theta_0 = \{(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2) | (\mu_1, \mu_2, \ldots, \mu_k, \sigma^2) \in \Theta \wedge \mu_1 = \mu_2 = \ldots = \mu_k\}$ lautet das Testproblem in der allgemeinen Schreibweise:

$$H_0 : \vartheta \in \Theta_0, \qquad H_1 : \vartheta \in \Theta \backslash \Theta_0.$$

Für die Bestimmung des Likelihood-Quotienten-Tests werden die Suprema in Zähler und Nenner von $\ell(\boldsymbol{x})$, $\boldsymbol{x} = (x_{11}, \ldots, x_{km})$ getrennt ermittelt. Zunächst wird der Zähler betrachtet:

$$\sup_{\vartheta \in \Theta_0} \left(\frac{1}{2\pi\sigma^2} \right)^{k \cdot m/2} \exp \left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{k} \sum_{j=1}^{m} (x_{ij} - \mu)^2 \right) \right].$$

Das Supremum wird über die Nullsetzen der partiellen Ableitungen von $\ln f(\boldsymbol{x}; \vartheta)$ bestimmt. Wir erhalten

$$\begin{aligned} \frac{\partial \ln f(\boldsymbol{x}; \vartheta)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{m} (x_{ij} - \mu) \\ \frac{\partial \ln f(\boldsymbol{x}; \vartheta)}{\partial \sigma^2} &= \frac{k \cdot m}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{k} \sum_{j=1}^{m} (x_{ij} - \mu)^2 . \end{aligned}$$

Nullsetzen dieser partiellen Ableitungen ergibt die Werte $\tilde{\mu}$ und $\tilde{\sigma}^2$, bei denen das Supremum unter H_0 angenommen wird. Diese sind:

$$\tilde{\mu} = \frac{1}{k \cdot m} \sum_{i=1}^{k} \sum_{j=1}^{m} x_{ij} = \bar{x}, \qquad \tilde{\sigma}^2 = \frac{1}{k \cdot m} \sum_{i=1}^{k} \sum_{j=1}^{m} (x_{ij} - \bar{x})^2 = s^2.$$

Einsetzen ergibt:

$$\sup_{\vartheta\in\Theta_0} f(\boldsymbol{x};\vartheta) = \left(\frac{k\cdot m}{2\pi\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij}-\bar{x})^2}\right)^{k\cdot m/2} \exp\left[-\frac{k\cdot m\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij}-\bar{x})^2}{2\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij}-\bar{x})^2}\right]$$
$$= \left(\frac{k\cdot m}{2\pi\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij}-\bar{x})^2}\right)^{k\cdot m/2} \exp\left[-\frac{k\cdot m}{2}\right].$$

Analog erhalten wir bei der Bestimmung des Nenners die Werte $\tilde{\mu}_1,\tilde{\mu}_2,\ldots,\tilde{\mu}_k$ und σ^2, bei denen der Zähler den maximalen Wert annimmt:

$$\frac{\partial \ln f(\boldsymbol{x};\vartheta)}{\partial \mu_i} = \frac{1}{\sigma^2}\sum_{j=1}^{m}(x_{ij}-\mu_i) \qquad i=1,\ldots,k$$
$$\frac{\partial \ln f(\boldsymbol{x};\vartheta)}{\partial \sigma^2} = \frac{k\cdot m}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij}-\mu_i)^2.$$

Nullsetzen ergibt die Lösungen:

$$\hat{\mu}_i = \frac{1}{m}\sum_{j=1}^{m}x_{ij}=\bar{x}_i, \qquad i=1,\ldots,k,$$
$$\hat{\sigma}^2 = \frac{1}{k\cdot m}\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij}-\bar{x}_i)^2.$$

Einsetzen ergibt:

$$\sup_{\vartheta\in\Theta} f(\boldsymbol{x};\vartheta) = \left(\frac{k\cdot m}{2\pi\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij}-\bar{x}_i)^2}\right)^{k\cdot m/2} \exp\left[-\frac{k\cdot m}{2}\right].$$

Damit ist schließlich der Likelihood-Quotient:

$$\ell(\boldsymbol{x}) = \left(\frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij}-\bar{x}_i)^2}{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij}-\bar{x})^2}\right)^{k\cdot m/2}.$$

Für den entsprechenden Ablehnbereich ist ein c zu finden mit $\mathrm{P}_\vartheta(\ell(\boldsymbol{X})<c)\le\alpha$ für alle $\vartheta\in\Theta_0$.
Die Verteilung einer mit $\ell(\boldsymbol{X})$ zusammenhängenden Stichprobenfunktion ist bekannt falls H_0 gilt; dies ist

$$\mathscr{L}(\boldsymbol{X}) = \frac{k(m-1)}{k-1}\left(\ell(\boldsymbol{X})^{-1/(km)}-1\right) = \frac{\frac{1}{k-1}}{\frac{1}{k(m-1)}}\cdot\frac{\sum_{i=1}^{k}m(\bar{X}_i-\bar{X})^2}{\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{ij}-\bar{X})^2}.$$

Es ist eine F-Verteilung mit $(k-1, k\cdot(m-1))$ Freiheitsgraden. Speziell ist die Verteilung von $\mathscr{L}(\boldsymbol{X})$ unabhängig von $\vartheta \in \Theta_0$. Aus der Tabelle der F-Verteilung kann also ein c' bestimmt werden mit $P_\vartheta(\mathscr{L}(\boldsymbol{X}) > c') = \alpha$ für alle $\vartheta \in \Theta_0$.
Da die Transformation $\ell \to \mathscr{L}$ monoton ist, ist der resultierende Ablehnbereich $K = \{\boldsymbol{x} | \mathscr{L}(\boldsymbol{x}) \geq c'\}$ äquivalent zu dem direkt auf ℓ basierenden.

Die Fragestellung dieses Beispiels wird als einfaktorielle *Varianzanalyse*, kurz als *ANOVA* bezeichnet.

In Fällen, wo man nicht direkt oder mittels geeigneter Transformation die Verteilung der Teststatistik unter H_0 bestimmen kann, muss man auf asymptotische Resultate zurückgreifen und sich mit einen näherungsweisen Ablehnbereich begnügen. Die Grundlage dafür bildet der folgende Satz, der den Satz 7.4.3 auf die schon im Beispiel betrachtete Situation zusammengesetzter Nullhypothesen erweitert.

Satz 8.2.9 *approximative Verteilung der Loglikelihoodfunktion*

Sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Zufallsstichprobe aus einer Dichte $f(x;\vartheta)$. $\boldsymbol{\vartheta} \in \Theta$ sei ein r-dimensionaler Parameter und Θ sei eine offene Teilmenge von $\mathbb{R}^r$. $f(x;\vartheta)$ sei hinreichend regulär, so dass insbesondere der Schätzer $\hat{\vartheta}$ asymptotisch (multivariat) normalverteilt ist.

Die Hypothese Θ_0 habe die Form $\Theta_0 = \{\vartheta | \vartheta \in \Theta,\ h_j(\vartheta) = 0,\ j = 1, \ldots, q\}$, wobei die Restriktionen nicht redundant seien. Sei dann $\ell(\boldsymbol{X})$ die Likelihood-Quotienten-Statistik zum Testen von

$$H_0 : \vartheta \in \Theta_0, \qquad H_1 : \vartheta \in \Theta \backslash \Theta_0.$$

Ist dann $H_0 : \vartheta \in \Theta_0$ wahr, so ist $-2\ln(\ell(\boldsymbol{X}))$ approximativ χ^2_q verteilt.

Beweis: Siehe Rao (1973, S. 350 ff.)

Wird auf das asymptotische Resultat, also auf die Teststatistik $-2\ln(\ell(\boldsymbol{X}))$, zurückgegriffen, so ist zu beachten, dass bei dieser transformierten Teststatistik *große Werte* zur Ablehnung der Nullhypothese führen.

Beispiel 8.2.10 *Likelihood-Quotienten-Test bei Exponentialverteilung*

Für den Parameter λ einer exponentialverteilten Zufallsvariablen X liege das Testproblem $H_0 : \lambda = \lambda_0,\ H_1 : \lambda \neq \lambda_0$ vor. Auf der Basis einer Stichprobe vom Umfang n soll ein Likelihood-Quotienten-Test zum Niveau α bestimmt werden.

Mit der Maximum-Likelihood-Schätzfunktion $\hat{\vartheta} = 1/\bar{x}$ erhalten wir:

$$\ell(\boldsymbol{x}) = \frac{f(\boldsymbol{x};\lambda_0)}{\sup\limits_{\lambda>0} f(\boldsymbol{x};\lambda)} = \frac{\lambda_0^n \exp\left[-\lambda_0 \sum_{i=1}^n x_i\right]}{\sup\limits_{\lambda>0} \lambda^n \exp\left[-\lambda \sum_{i=1}^n x_i\right]} = \frac{\lambda_0^n e^{-\lambda_0 n\bar{x}}}{\bar{x}^{-n} e^{-n}}.$$

Logarithmieren ergibt

$$-2\cdot\ln\ell(\boldsymbol{x}) = 2n\cdot(\lambda_0\bar{x} - 1) - 2n\cdot\ln(\lambda_0\bar{x}).$$

Da $\bar{x}$ konsistent für $1/\lambda$ ist, weicht $\lambda_0\bar{x}$ für große n nur mit einer kleinen Wahrscheinlichkeit von 1 ab. Somit gilt approximativ:

$$\begin{aligned}-2\cdot\ln\ell(\boldsymbol{x}) &= 2n\cdot(\lambda_0\bar{x}-1)-2n\cdot\ln(1+(\lambda_0\bar{x}-1))\\ &= 2n(\lambda_0\bar{x}-1)-2n\sum_{\nu=1}^{\infty}(-1)^{\nu-1}\frac{(\lambda_0\bar{x}-1)^\nu}{\nu} = 2n\sum_{\nu=2}^{\infty}(-1)^{\nu}\frac{(\lambda_0\bar{x}-1)^\nu}{\nu}\\ &= n(\lambda_0\bar{x}-1)^2+2n\sum_{\nu=3}^{\infty}(-1)^{\nu}\frac{(\lambda_0\bar{x}-1)^\nu}{\nu}.\end{aligned}$$

Aufgrund des zentralen Grenzwertsatzes ist $\sqrt{n}(\lambda_0\bar{X}-1)$ asymptotisch standardnormalverteilt, so dass $n(\lambda_0\bar{X}-1)^2$ asymptotisch χ^2-verteilt ist mit einem Freiheitsgrad. Der vorstehende Satz sagt gerade, dass der zweite Term des letzten Ausdruckes asymptotisch vernachlässigbar ist.

Anhand dieses Beispiels lässt sich auch die Bedeutung der Bedingung an die Nullhypothese ersehen. Für das Testproblem $H_0: \lambda\le\lambda_0$, $H_1: \lambda>\lambda_0$ erhalten wir als Teststatistik des Likelihood-Quotienten-Tests:

$$\ell(\boldsymbol{x}) = \frac{\sup\limits_{0<\lambda\le\lambda_0} f(\boldsymbol{x};\lambda)}{\sup\limits_{\lambda>0} f(\boldsymbol{x};\lambda)} = \begin{cases} 1 & \text{falls } \dfrac{1}{\bar{x}}\le\lambda_0 \\ \dfrac{\lambda_0^n e^{-\lambda_0 n\bar{x}}}{\bar{x}^{-n}e^{-n}} & \text{falls } \dfrac{1}{\bar{x}}>\lambda_0 \end{cases}.$$

Es ist offensichtlich, dass $-2\ln\ell(\boldsymbol{X})$ hier asymptotisch keine Standardverteilung haben kann.

8.2.4 Der Wald- und der Score-Test

Um diese beiden Tests, die in jüngerer Zeit eine größere Bedeutung erlangt haben, zu motivieren, wird von einer anderen Interpretation der asymptotischen Variante des Likelihood-Quotienten-Tests ausgegangen.[1] Seine Prüfgröße lässt sich auch in der Form

$$\mathrm{LQ} = 2\cdot[\ln \mathrm{L}(\hat{\vartheta}) - \ln \mathrm{L}(\tilde{\vartheta})] \tag{8.13}$$

schreiben, wobei $\mathrm{L}(\vartheta) = f(x;\vartheta)$ gilt, $\hat{\vartheta}$ der uneingeschränkte Maximum-Likelihood-Schätzer für ϑ ist und $\tilde{\vartheta}$ der bedingte ML-Schätzer, der die Nullhypothese spezifizierenden Bedingungen erfüllt.

Zur besseren Veranschaulichung wird unterstellt, dass ϑ ein eindimensionaler Parameter ist und die Hypothese H_0 einfach, $H_0: \vartheta=\vartheta_0$. Die Alternative sei $H_1: \vartheta\ne\vartheta_0$. Wird dann die Loglikelihoodfunktion gezeichnet, so kann der Wert von

$$\frac{1}{2}\mathrm{LQ} = \ln(\mathrm{L}(\hat{\vartheta})) - \ln(\mathrm{L}(\vartheta_0))$$

direkt aus der Grafik abgelesen werden, siehe Abbildung 8.4.

[1] Die Dartstellung in diesem Abschnitt basiert auf Buse (1982)

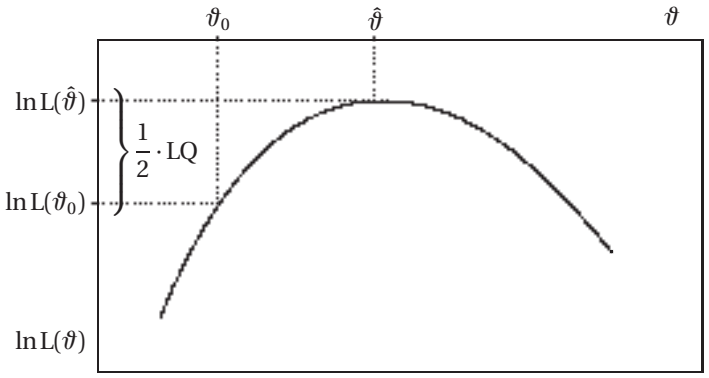


Abb. 8.4: *Der Likelihood-Quotienten-Test*

Aus der Abbildung ist ersichtlich, dass der Abstand LQ/2 von der Differenz $\hat{\vartheta} - \vartheta_0$ und von der Krümmung der Log-Likelihood-Funktion abhängt. Die Krümmung wird durch $K(\hat{\vartheta})$, dem absoluten Wert von

$$\left. \frac{\partial^2 \ln L(\vartheta)}{\partial \vartheta^2} \right|_{\vartheta=\hat{\vartheta}}$$

gemessen. Bei gegebener Krümmung $K(\hat{\vartheta})$ ist $\ln L(\vartheta_0)$ umso weiter vom Maximum $\ln L(\hat{\vartheta})$ entfernt, mithin LQ/2 umso größer, je größer $\hat{\vartheta} - \vartheta_0$ ist. Umgekehrt gilt, dass bei gegebener Differenz $\hat{\vartheta} - \vartheta_0$ die Entfernung $\ln L(\hat{\vartheta}) - \ln L(\vartheta_0)$ mit wachsender Krümmung größer wird.

Der *Wald-Test* geht nun nicht von dem Abstand $\ln L(\hat{\vartheta}) - \ln L(\vartheta_0)$ aus. Stattdessen basiert er auf der intuitiv einleuchtenden Differenz $\hat{\vartheta} - \vartheta_0$. Große Unterschiede zwischen $\hat{\vartheta}$ und ϑ_0 werden als Evidenz gegen die Nullhypothese gewertet. Entsprechend der oben geführten Diskussion ist es von Vorteil, die quadrierte Differenz $(\hat{\vartheta} - \vartheta_0)^2$ noch mit der Krümmung $K(\hat{\vartheta})$ zu gewichten. Ergeben sich bei zwei Stichproben A und B gleiche Werte der quadrierten Differenzen, so ist diejenige Schätzung $\hat{\vartheta}$ als unverträglicher mit H_0 anzusehen, bei der die Krümmung $K(\hat{\vartheta})$ größer ist. In der Abbildung 8.5 ist dies verdeutlicht. Die Krümmung bei

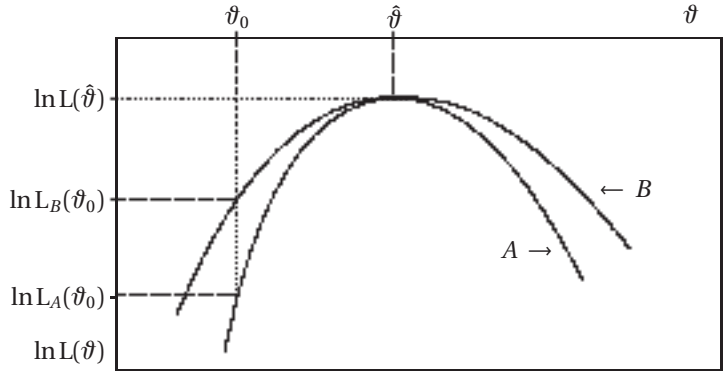


Abb. 8.5: *Der Wald-Test*

der Stichprobe A ist offenbar größer als bei der Stichprobe B, die Hypothese $H_0 : \vartheta = \vartheta_0$ ist

bei A eher als unplausibel anzusehen.

Allerdings ist es üblich, bei der Bildung der Prüfgröße die Krümmung durch die erwartete Krümmung zu ersetzen. Dies ergibt

$$W = (\hat{\vartheta} - \vartheta_0)^2 I(\hat{\vartheta}), \tag{8.14}$$

wobei $I(\vartheta)$ die Fisher-Information ist, $I(\vartheta) = \mathrm{E}((\partial^2 \ln \mathrm{L}(\vartheta))/\partial \vartheta^2)$.

Der Wald-Test verwendet offensichtlich nur die ML-Schätzung aus dem vollen Parameterraum. Er hat daher unter dem Aspekt der Berechnung einen gewissen Vorteil gegenüber dem Likelihood-Quotienten-Test.

Die Idee des *Score-Tests* (auch *Lagrange-Multiplikator-Test* genannt) besteht darin, die Steigung der Log-Likelihood-Funktion, $S(\vartheta) = \partial \ln \mathrm{L}(\vartheta)/\partial \vartheta$, an der Stelle ϑ_0 als Indikator für den Abstand zu nehmen. Bei $\vartheta = \hat{\vartheta}$ gilt $S(\hat{\vartheta}) = 0$. Je mehr $S(\vartheta_0)$ von Null abweicht, desto mehr unterscheiden sich Schätzung und hypothetischer Wert. Wie die Abbildung 8.6 deutlich macht, ist aber auch hier die Einbeziehung der Krümmung der Log-Likelihood angebracht. Jedoch wird nun die Krümmung an der Stelle ϑ_0 betrachtet. Das ergibt ein qualitativ anderes Bild als beim Wald-Test: Beim Datensatz A, wo die Likelihoodfunktion die größere Krümmung in ϑ_0 aufweist, ist der Abstand zwischen ϑ_0 und $\hat{\vartheta}$ geringer als bei B. Da hier das Vorzeichen der Steigung irrelevant ist, stellt $S(\vartheta_0)^2$ die geeignete Basisgröße für den Test dar. Nach den

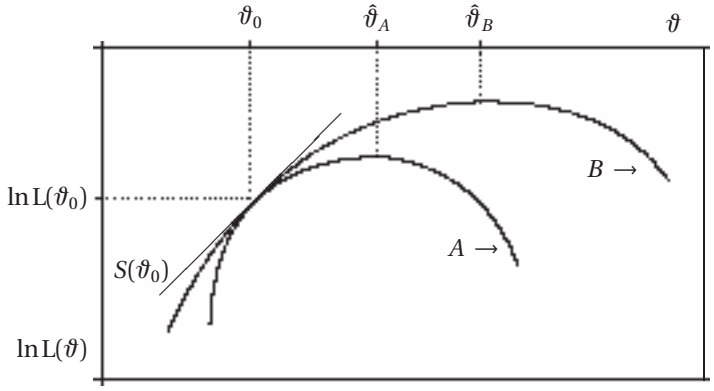


Abb. 8.6: *Der Score-Test*

Überlegungen zur Krümmung der Log-Likelihood ist $S(\vartheta_0)^2$ mit der Inversen der Krümmung zu gewichten. Als Teststatistik erhalten wir also

$$\mathrm{LM} = S(\vartheta_0)^2 I(\vartheta_0)^{-1}. \tag{8.15}$$

Beide Teststatistiken, W und LM, sind unter H_0 asymptotisch verteilt wie die Teststatistik des Likelihood-Quotienten-Tests, nämlich χ^2_1. Bei mehrdimensionalen Parametern ändern sich die Freiheitsgrade der asymptotischen Verteilung entsprechend, vgl. Satz 8.2.9.

Beispiel 8.2.11 *Score-Test bei einfacher linearer Regression*

Gegeben sei das Regressionsmodell

$$Y_i = \beta_0 + \beta_1 x_i + U_i, \quad i = 1, \ldots, n,$$

wobei die U_i unabhängig seien mit $U_i \sim \mathcal{N}(0,\sigma^2)$.

Die Loglikelihoodfunktion lautet hier:

$$\ln L(\sigma^2,\beta_0,\beta_1) = -\frac{n}{2}\ln(\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\beta_0-\beta_1 x_i)^2.$$

Das ergibt mit $\boldsymbol{\vartheta} = (\sigma^2,\beta_0,\beta_1)^T$ den Score-Vektor

$$S(\boldsymbol{\vartheta}) = \begin{pmatrix} \frac{\partial \ln L(\boldsymbol{\vartheta})}{\partial \vartheta_1} \\ \frac{\partial \ln L(\boldsymbol{\vartheta})}{\partial \vartheta_2} \\ \frac{\partial \ln L(\boldsymbol{\vartheta})}{\partial \vartheta_3} \end{pmatrix} = \begin{pmatrix} -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(y_i-\beta_0-\beta_1 x_i)^2 \\ \frac{1}{\sigma^2}\sum(y_i-\beta_0-\beta_1 x_i) \\ \frac{1}{\sigma^2}\sum(y_i-\beta_0-\beta_1 x_i)x_i \end{pmatrix}$$

und die Fisher-Informationsmatrix

$$I(\boldsymbol{\vartheta}) = \left(-\frac{\partial^2 \ln L(\boldsymbol{\vartheta})}{\partial\vartheta_i \partial\vartheta_j}\right) = \frac{1}{\sigma^2}\begin{pmatrix} -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}\sum z_i^2 & \frac{1}{\sigma^2}\sum z_i & \frac{1}{\sigma^2}\sum z_i x_i \\ \frac{1}{\sigma^2}\sum z_i & n & \sum x_i \\ \frac{1}{\sigma^2}\sum z_i x_i & \sum x_i & \sum x_i^2 \end{pmatrix}$$

mit $z_i = (y_i - \beta_0 - \beta_1 x_i)$.
Es interessiere nun die Nullhypothese $H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0 = (1,0,0)^T$. Dann ist

$$I(\boldsymbol{\vartheta}_0) = \begin{pmatrix} -\frac{n}{2} + \sum x_i^2 & \sum y_i & \sum y_i x_i \\ \sum y_i & n & \sum x_i \\ \sum y_i x_i & \sum x_i & \sum x_i \end{pmatrix},$$

und die approximativ χ^2_3-verteilte Teststatistik des Score-Tests lautet

$$\begin{aligned} &S(\boldsymbol{\vartheta}_0)^T I(\boldsymbol{\vartheta}_0)^{-1} S(\boldsymbol{\vartheta}_0) \\ &= \left(-\tfrac{n}{2} + \tfrac{1}{2}\sum Y_i^2, \sum Y_i, \sum Y_i x_i\right) \cdot \begin{pmatrix} -\frac{n}{2} + \sum Y_i^2 & \sum Y_i & \sum Y_i x_i \\ \sum Y_i & n & \sum x_i \\ \sum Y_i x_i & \sum x_i & \sum x_i \end{pmatrix}^{-1} \begin{pmatrix} -\frac{n}{2} + \frac{1}{2}\sum Y_i^2 \\ \sum Y_i \\ \sum Y_i x_i \end{pmatrix}. \end{aligned}$$

Nun betrachten wir die Nullhypothese $H_0 : \beta_1 = \beta_1^*$. Für den Score-Test ist der Teilvektor (σ^2,β_0) mit der restringierten ML-Methode zu schätzen. Dazu wird in $\boldsymbol{\vartheta}$ der unter H_0 spezifizierte Wert eingesetzt und dann werden die restlichen Normalgleichungen gelöst. Bezeichnen wir die restringierte ML-Schätzung von ϑ_1 unter H_0 mit $\tilde{\vartheta}_1$, und setzen wir $\tilde{\boldsymbol{\vartheta}} = (\tilde{\vartheta}_1', \vartheta_2^*)^T$, so hat die Score-Funktion $S(\tilde{\boldsymbol{\vartheta}})$ i.d.R. an den Stellen Nullen, an denen

die Koeffizienten geschätzt wurden. Für die unbekannten Parameter werden in der Informationsmatrix naheliegender Weise die restringierten Schätzungen eingesetzt. Man kann zeigen, dass dies unter Regularitätsbedingungen erlaubt ist. Die Teststatistik des Score-Tests ist dann $S(\tilde{\boldsymbol{\vartheta}})^T I(\tilde{\boldsymbol{\vartheta}})^{-1} S(\tilde{\boldsymbol{\vartheta}})$.

Sei konkret $H_0 : \beta_1 = 0$. Dann erhalten wir die restringierten ML-Schätzwerte für σ^2 und β_0 als Lösungen von

$$\begin{aligned} -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\beta_0)^2 = 0, \\ \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0) = 0 \end{aligned} \quad .$$

Das ergibt $\tilde{\beta}_0 = \bar{y}$ und $\tilde{\sigma}^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2/n$. Die Teststatistik des Score-Tests ist damit

$$\begin{aligned} &S(\tilde{\boldsymbol{\vartheta}})^T I(\tilde{\boldsymbol{\vartheta}})^{-1} S(\tilde{\boldsymbol{\vartheta}}) \\ &= \tilde{\sigma}^2 \cdot \left(0, 0, \tfrac{1}{\tilde{\sigma}^2}\sum(Y_i-\bar{Y})x_i\right) \begin{pmatrix} n/2\tilde{\sigma}^2 & 0 & \sum(Y_i-\bar{Y})x_i \\ 0 & n & \sum x_i \\ \sum(Y_i-\bar{Y})x_i & \sum x_i & \sum x_i \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ \frac{1}{\tilde{\sigma}^2}\sum(Y_i-\bar{Y})x_i \end{pmatrix} \\ &= \frac{n\cdot(\overline{Y\cdot x}-\bar{Y}\cdot\bar{X})^2}{\tilde{\sigma}^2(\overline{x^2}-\bar{x}^2)-2\tilde{\sigma}^4(\overline{Y\cdot x}-\bar{Y}\cdot\bar{x})^2}. \end{aligned}$$

Diese ist äquivalent zu der bekannten, auf der Steigung basierenden F-Teststatistik. $S(\tilde{\boldsymbol{\vartheta}})^T I(\tilde{\boldsymbol{\vartheta}})^{-1} S(\tilde{\boldsymbol{\vartheta}})$ ist asymptotisch χ^2_1-verteilt.

8.2.5 Bedingte Tests

In mehrparametrischen Verteilungen liegt häufig der Fall vor, dass das Testproblem nur eine Komponente des Parameters betrifft. Als Beispiel sei nur der Fall der Normalverteilung genannt, bei dem es um das Testen einer Hypothese über den Erwartungswert geht, und wo die Varianz unbekannt und beliebig ist. Bisweilen ist es möglich, von der Verteilung der Stichprobenvariablen $\boldsymbol{X}$ zu einer bedingten Verteilung überzugehen, wobei der Wert einer geeigneten Statistik $T(\boldsymbol{X})$ gegeben ist. Bei diesem Übergang verschwindet dann die Abhängigkeit von dem *nuisance Parameter*, d. h. von den nicht relevanten Komponenten des Parameters. Das formale Problem, wann die einzelnen bedingten Tests zu einem Gesamttest zuammengesetzt werden dürfen, soll hier nicht diskutiert werden. Der Leser sei dazu auf Nölle (1967) verwiesen. Wenn das allerdings möglich ist, so ist der ‚zusammengesetzte' Test ein Test zum Niveau α:

$$\mathrm{E}_\vartheta\,\varphi(\boldsymbol{X}) = \mathrm{E}_\vartheta(\mathrm{E}_\vartheta(\varphi(\boldsymbol{X})|T(\boldsymbol{X}))) \le \mathrm{E}_\vartheta(\alpha) = \alpha. \tag{8.16}$$

Der resultierende Test wird als *bedingter Test* bezeichnet.

Beispiel 8.2.12 *verschobene Exponentialverteilung*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine Stichprobe aus einer verschobenen Exponentialverteilung, habe also die Dichte

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

Für den Parameter λ sei das Testproblem

$$H_0 : \lambda \leq \lambda_0, \qquad H_1 : \lambda > \lambda_0$$

gegeben. Ausführlich müssten wir dafür schreiben:

$$H_0 : (\lambda, \vartheta) \in \Lambda_0 \times \mathbb{R}, \qquad H_1 : (\lambda, \vartheta) \in \Lambda_1 \times \mathbb{R}$$

mit $\Lambda_0 = (\lambda_0, \infty)$, $\Lambda_1 = (-\infty, \lambda_0]$.

$(U, T) = (X_{1:n}, \bar{X})$ ist eine suffiziente Statistik für diese Verteilungsfamilie. Da U zudem der ML-Schätzer für ϑ ist, betrachten wir die bedingte Verteilung von $\boldsymbol{X}$ bei gegebenem $U = u$. Mit der Randverteilung von U,

$$f_U(u; \lambda, \vartheta) = n(1 - F_X(u; \lambda, \vartheta))^{n-1} \lambda e^{-\lambda(u-\vartheta)} = n\lambda e^{-n\lambda(u-\vartheta)} \qquad (u > \vartheta),$$

erhalten wir:

$$f_{\boldsymbol{X}|U}(\boldsymbol{x}|u) = \frac{\lambda^n e^{-n\lambda(\bar{x}-\vartheta)}}{n\lambda e^{-n\lambda(u-\vartheta)}} = \frac{1}{n} \lambda^{n-1} e^{-n\lambda(\bar{x}-u)}.$$

Die bedingte Verteilung von $\boldsymbol{X}$ bei gegebenem Wert von $U = u$ hängt nicht mehr von ϑ ab. Zudem ist $T = \bar{X}$ für den Parameter λ der bedingte Verteilung suffizient. Somit können wir nach Satz 8.1.10 für jedes feste u einen nur von T abhängigen Test für unser Testproblem angeben:

$$\varphi(t, u) = \begin{cases} 1 & t - u > c(u) \\ 0 & t - u \leq c(u) \end{cases}.$$

Bestimmen wir $c(u)$ jeweils so, dass der bedingte Test ein Test zum Niveau α ist, $\mathrm{E}_\lambda(\varphi(T, u)) \leq \alpha$, so gilt auch:

$$\mathrm{E}_{\lambda,\vartheta}(\varphi(T, U)) = \mathrm{E}_{\lambda,\vartheta}(\mathrm{E}_{\lambda,\cdot}(\varphi(T, U)|U)) \leq \mathrm{E}_{\lambda,\vartheta}(\alpha) = \alpha.$$

8.2.6 Permutationstests

Oft ist es unmöglich, eine parametrische Familie von Verteilungen anzugeben, die das interessierende Phänomen adäquat erfasst. Speziell die Annahme einer Normalverteilung ist in vielen Fällen dubios. Dann mag es sinnvoll sein, bei einer stetigen Zufallsvariablen alle stetigen Verteilungen als mögliche Modelle zuzulassen. Bei der Betrachtung von Differenzen lässt sich diese Klasse auf die symmetrischen stetigen Verteilungen eingrenzen. Beide Verteilungsklassen sind nichtparametrisch, vgl. Abschnitt 8.1.3. Wir wollen in diesem und dem nächsten Abschnitt Prinzipien für die Konstruktion von Tests für solche nichtparametrischen Verteilungsfamilien angeben. Wir unterstellen, dass die zugrundeliegenden Verteilungen alle stetig sind. Dies ist in diesem Abschnitt nicht unbedingt erforderlich, erleichtert aber die Diskussion erheblich. Am Ende sollte der geneigte Leser imstande sein, die notwendigen Änderungen für die Abschwächung dieser Voraussetzung selbst vorzunehmen.

Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen, die unter der Nullhypothese H_0 identisch verteilt sind und die stetige Verteilungsfunktion $F \in \mathscr{F}_0 \subset \mathscr{F}$ mit der Dichte f haben. Die

Alternativhypothese lege eine geeignete Abweichung von der unter H_0 spezifizierten Verteilungsform bzw. -aussage fest.

$S(\boldsymbol{X})$ sei suffizient für $F \in \mathscr{F}_0$. $\mathscr{F}_0$ sei nun so umfangreich, dass die bedingte Verteilung jeder Statistik $T(\boldsymbol{X})$ gegeben den Wert $S(\boldsymbol{X}) = s$ lediglich über Enumeration aller dann noch möglichen Realisationen bestimmbar ist. Der Ablehnbereich eines bedingten Tests $\varphi(\boldsymbol{X}|S = s)$, der auf einer Teststatistik $T(\boldsymbol{X})$ basiert, kann dann ebenfalls über ein Abzählen der möglichen Werte von $T(\boldsymbol{X})$ bei jeweils gegebenem $S(\boldsymbol{X}) = s$ ermittelt werden. Oft basieren diese Abzählungen auf Permutationen. Daher sprechen wir von *Permutationstests*. Da die bedingte Verteilung von $T(\boldsymbol{X})$ unabhängig von dem Parameter F ist, sind Permutationstests *verteilungsfrei*.

Oft nimmt man als Teststatistik eine, die man aus der parametrischen Theorie 'gewohnt' ist. Da zudem die Bestimmung aller Permutationen sehr aufwendig ist oder sogar praktisch unmöglich sein kann, behilft man sich bei nicht so kleinen Stichproben mit einer Monte-Carlo-Variante. Dazu werden ausreichend viele Permutationen (z.B. 10 000) zufällig erzeugt und der Wert der Teststatistik dafür berechnet. Sei α^* der Anteil der so erzeugten Werte von T, die extremer sind als der aus der aktuellen Stichprobe erhaltene. Falls α^* kleiner ist als das vorgegebene Niveau, so wird H_0 abgelehnt.

Natürlich ist zu beachten, dass nicht alle ‚gewohnten' Teststatistiken hier sinnvoll sind und nicht alle Testprobleme mittels dieses Ansatzes behandelbar. Bei dem univariaten Lageproblem z.B. können wir nicht von dem arithmetischen Mittel ausgehen. Denn dieses ergibt ja unter allen Permutationen denselben Wert.

Beispiel 8.2.13 *Zweistichproben-Lageproblem*

Beim Zweistichproben-Lageproblem liegen Stichprobenvariablen $X_1, \ldots, X_m$, $X_{m+1}, \ldots, X_n$ vor. Die Verteilung der ersten m Zufallsvariablen sei $F(x)$, die der letzten $n-m$ sei $F(x-\Delta)$, wobei F wieder eine stetige Verteilungsfunktion ist, $F \in \mathscr{F}$.

Wir betrachten die Hypothesen

$$H_0 : \Delta = 0; \qquad H_1 : \Delta \neq 0.$$

Unter H_0 ist dann $\boldsymbol{X}_{:n} = (X_{1:n}, \ldots, X_{n:n})$ eine minimalsuffiziente Statistik. Damit gilt für jede Statistik $T(\boldsymbol{X})$, wenn $(i_1, \ldots, i_n)$ eine beliebige Permutation von $(1, \ldots, n)$ bezeichnet:

$$\mathrm{P}(T(\boldsymbol{X}) = t | \boldsymbol{X}_{:n} = \boldsymbol{x}_{:n}) = \frac{|\{(i_1, \ldots, i_n) | T(x_{i_1}, \ldots, x_{i_n}) = t\}|}{n!}. \tag{8.17}$$

Es liegt nahe, für den Betrag der Differenz $\bar{X}_1 - \bar{X}_2$ zu wählen, wobei $\bar{X}_1 = \sum_{i=1}^{m} X_i/m$ und $\bar{X}_2 = \sum_{i=m+1}^{n} X_i/(n-m)$.

Wir haben also alle Permutationen der beobachteten Werte zu bestimmen und jeweils aus den ersten m Komponenten $\bar{x}_1$ und aus den letzten $(n-m)$ Komponenten $\bar{x}_2$ zu berechnen. H_0 wird bei einer vorgegebenen Irrtumswahrscheinlichkeit α abgelehnt, wenn

$$\alpha^* = \mathrm{P}(|\bar{X}_1 - \bar{X}_2| \geq t_{\text{empirisch}} | X_{1:n} = x_{1:n}, \ldots, X_{n:n} = x_{n:n}) = \frac{k(|\bar{x}_1 - \bar{x}_2|)}{n!} \leq \alpha, \tag{8.18}$$

wobei $k(|\bar{x}_1 - \bar{x}_2|)$ die Anzahl der Permutationen ist, bei denen die Teststatistik einen größeren Wert als den aus der vorliegenden Stichprobe $x_1, \ldots, x_n$ berechneten $t_{\text{empirisch}}$ hat.

Beispiel 8.2.14 *Symmetrieproblem*

Wir betrachten das Symmetrieproblem aus dem Abschnitt 8.1.3. Sei $\mathscr{F}_s = \{F | F \in \mathscr{F},\ F(\tilde{\mu} - x) + F(\tilde{\mu} + x) = 1,\ x \in \mathbb{R},\ \tilde{\mu} \in \mathbb{R}\}$. Das interessierende Testproblem laute:

$$H_0 : \tilde{\mu} = 0 \text{ und } F \in \mathscr{F}_s; \qquad H_1 : \tilde{\mu} \neq 0 \text{ oder } F \in \mathscr{F} \setminus \mathscr{F}_s.$$

Ein Test basiert auf Beobachtungen $x_1, \ldots, x_n$ einer geeigneten Zufallsvariablen. $X_{1:n}, \ldots, X_{n:n}$ ist suffizient für die unter H_0 festgelegte Verteilungsfamilie $\mathscr{F}_s^0 = \{F | F \in \mathscr{F},\ F(-x) + F(x) = 1,\ x \in \mathbb{R}\}$. $T = \bar{X}$ ist wegen der Symmetrie eine sinnvolle Teststatistik, jedoch ist die Verwendung für feste Werte $x_{1:n}, \ldots, x_{n:n}$ sinnlos. $\bar{x}$ ist ja für alle Permutationen der Beobachtungen gleich groß. Die adäquate suffiziente Statistik ist hier wegen der Symmetrie auch eine andere. Mit der zu F gehörigen Dichte f gilt mit der Symmetrie um Null unter H_0, wenn $x_{|i|:n}$ die i'te geordnete Statistik der Beträge $|x_1|, \ldots, |x_n|$ bedeutet:

$$f(x_1) \cdot \ldots \cdot f(x_n) = f(|x_1|) \cdot \ldots \cdot f(|x_n|) = n! \cdot f(x_{|1|:n}) \cdot \ldots \cdot f(x_{|n|:n}).$$

Halten wir also $x_{|1|:n}, \ldots, x_{|n|:n}$ fest, so erhalten wir die verschiedenen Werte von $\bar{X}$ dadurch, dass diesen Beträgen zufällig positive oder negative Vorzeichen zugeordnet werden. Nach den Überlegungen im Abschnitt zur Kombinatorik gibt es 2^n mögliche Vorzeichenfolgen $--\ldots--$, $--\ldots-+$, $--\ldots+-$, ..., $++\ldots++$.

Damit ist die bedingte Verteilung von $\bar{X}$ unter H_0 gegeben durch

$$\mathrm{P}_0(\bar{X} = \bar{x} | X_{|1|:n} = x_{|1|:n}, \ldots, X_{|n|:n} = x_{|n|:n}) = \frac{|\{(i_1, \ldots, i_n) | i_j \in \{0,1\}, \sum (-1)^{i_j} x_{|i|:n} = \bar{x}\}|}{2^n}.$$

Zu große und zu kleine Werte von $\bar{X}$ führen zur Ablehnung von H_0.

Im Beispiel 8.1.12 sind bei $n = 10$ Beobachtungen $2^{10} = 1024$ Werte zu berechnen.

8.2.7 Rangtests

Eine wichtige spezielle Klasse von Permutationstests sind diejenigen, bei denen die Teststatistik auf Rängen basiert. Zu einer Stichprobe $X_1, \ldots, X_n$ gehörige *Ränge* $R_1, \ldots, R_n$ erhalten wir, indem die zu der geordneten Statistik $X_{1:n}, \ldots, X_{n:n}$ komplementäre Information betrachtet wird: $X_i = X_{R_i:n}$. Somit nimmt $R_i = R(X_i)$ den Wert 1 an, wenn x_i die kleinste Beobachtung ist, den Wert 2 bei der zweitkleinsten usw. bis zum Wert n bei der größten. Wegen der vorausgesetzten Stetigkeit der zugrundeliegenden Verteilungen sind alle Beobachtungen — theoretisch — verschieden.

Ränge benutzen nur die Information der Anordnung der beobachteten Werte. Damit sind auf Rängen basierende Tests auch für ordinal skalierte Variablen geeignet

Beispiel 8.2.15 *Rangbildung*

Eine Stichprobe vom Umfang $n = 3$ ergab folgende Werte: $x_1 = 10,\ x_2 = 8.5,\ x_3 = 15$. Dann sind $x_{1:3} = 8.5,\ x_{2:3} = 10,\ x_{3:3} = 15$ und $r_1 = r(x_1) = 2,\ r_2 = r(x_2) = 1,\ r_3 = r(x_3) = 3$.

Lemma 8.2.16 *Gleichverteilung der Ranganordnungen*

Sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe aus einer stetigen Verteilung F mit der Dichte f. Dann sind die Zufallsvektoren $\boldsymbol{X}_{:n} = (X_{1:n}, \ldots, X_{n:n})$ und $\boldsymbol{R} = (R_1, \ldots, R_n)$ stochastisch unabhängig mit

$$\mathrm{P}(\boldsymbol{R} = \boldsymbol{r}) = \frac{1}{n!} \tag{8.19}$$

für alle Permutationen $\boldsymbol{r} = (r_1, \ldots, r_n)$ von $(1, \ldots, n)$.

Beweis: Wir zeigen, dass die Wahrscheinlichkeit für eine spezielle Rangfolge gleich $1/n!$ ist. O.B.d.A. wählen wir $(1, \ldots, n)$.

$$\begin{aligned}\mathrm{P}(\boldsymbol{R} = (1, \ldots, n)) &= \mathrm{P}(X_1 < \cdots < X_n) \\ &= \int\limits_{x_1 < \cdots < x_n} \cdots \int f(x_1) \cdot \ldots \cdot f(x_n)\, \mathrm{d}x_1 \cdot \ldots \cdot \mathrm{d}x_n = \int\limits_{x_1 < \cdots < x_n} \cdots \int \frac{1}{n!} f_{\boldsymbol{X}_{:n}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ &= \frac{1}{n!} \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} f_{\boldsymbol{X}_{:n}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \frac{1}{n!}.\end{aligned}$$

Die Unabhängigkeit ist leicht zu sehen: Da $\boldsymbol{R}$ und $\boldsymbol{X}_{:n}$ den Ausgangsvektor $\boldsymbol{X}$ eindeutig bestimmen, gilt

$$f(\boldsymbol{r}, \boldsymbol{x}_{:n}) = f(\boldsymbol{x}) = f(x_1) \cdot \ldots \cdot f(x_n) = \frac{1}{n!} n!\, f(x_{1:n}) \cdot \ldots \cdot f(x_{n:n}) = f(\boldsymbol{r}) \cdot f(\boldsymbol{x}_{:n}).$$

Beispiel 8.2.17 *ein Rangtest für das Zweistichproben-Lageproblem*

Wir wollen einen plausiblen Test für das Zweistichproben-Lageproblem angeben. $X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m}$ seien dazu die Stichprobenvariablen, deren Verteilungen zu einer Lage-Familie $\mathscr{F} = \{F(\cdot; \Delta) | F(x; \Delta) = F_0(x - \Delta)\}$ gehören. Die Frage ist, ob die beiden Stichproben, bestehend aus den ersten n bzw. den letzten m Stichprobenvariablen, aus gegeneinander verschobenen Verteilungen stammen. Als einseitiges Testproblem können wir also etwa formulieren, wobei die Zuordnungen der Δ offensichtlich sind:

$$H_0 : \Delta_1 \leq \Delta_2, \quad F \in \mathscr{F}; \qquad H_1 : \Delta_1 > \Delta_2, \quad F \in \mathscr{F}.$$

Unter H_1 werden wir eher größere Werte der ersten Stichprobe erwarten. Folglich führen große Werte der Summe der Ränge der ersten Stichprobe

$$W = \sum_{i=1}^{n} R_i \tag{8.20}$$

zur Ablehnung von H_0. Dies ist die Teststatistik des *Wilcoxon-Rangsummentests*. Es lässt sich leicht nachrechnen, dass für $\Delta = 0$ gilt:

$$\mathrm{E}(W) = \frac{n \cdot m(n+m)}{2}, \qquad \mathrm{Var}(W) = \frac{n \cdot m \cdot (n+m+1)}{12}.$$

Zudem ist W asymptotisch normalverteilt, so dass die kritischen Werte approximativ bestimmt werden können. Für kleine Stichprobenumfänge geben wir weiter unten ein Verfahren zur Berechnung der exakten Verteilung an.

Beispiel 8.2.18 *Rangtests für das Zweistichproben-Skalenproblem*

Einen plausiblen Test für das Zweistichproben-Skalenproblem erhalten wir folgendermaßen. Bei der gleichen Ausgangssituation wie im letzten Beispiel sei die Verteilungsfamilie jetzt eine Skalen-Familie, $\mathscr{F} = \{F(\cdot;\tau)|F(x;\tau) = F_0(x/\tau)\}$. Die Stichprobenvariablen seien zudem positiv, $F_0(0) = 0$. Haben die Verteilungen der ersten n Variablen den Parameter τ_1 und die letzten m den Parameter τ_2, so beinhaltet die Alternativhypothese

$$H_1, : \tau_1 > \tau_2,$$

dass die Werte der ersten Stichprobe unter H_1 stärker streuen als unter $H_0 : \tau_1 \leq \tau_2$.

Ordnen wir extremen Rängen R_i kleine Werte $a(R_i)$ zu und mittleren Rängen größere, so deutet ein großer Wert von

$$T = \sum_{i=1}^{n} a(R_i)$$

auf die Gültigkeit von H_1 hin. Die Wahl

$$a(i) = \frac{n+m+1}{2} - \left| i - \frac{n+m+1}{2} \right|$$

ergibt den *Ansary-Bradley-Test.* Der Vorschlag von *Mood* lautet:

$$a(i) = \left(i - \frac{n+m+1}{2} \right)^2 .$$

Auf die asymptotische Normalverteilung und die exakte Berechnung der Verteilungen unter H_0 für kleine Stichprobenumfänge gehen wir weiter unten ein.

Die in den beiden Beispielen betrachteten Tests sind spezielle *lineare Rangtests.* Diese haben die allgemeine Form

$$L = \sum_{i=1}^{n+m} c_i a(R_i). \tag{8.21}$$

Dabei heißen die Konstanten $a(1), \ldots, a(n+m)$ die *Scores* und die $c_1, \ldots, c_{n+m}$ die *Regressionskonstanten.*

Aus der allgemeinen Form erhalten wir z.B. W durch die Wahl

$$c_i = \begin{cases} 1 & i = 1, \ldots, n \\ 0 & i = n+1, \ldots, n+m \end{cases}, \qquad a(i) = i \qquad i = 1, \ldots, n+m.$$

Unterschiedliche Scores und Regressionskonstanten ergeben Tests mit unterschiedlichen Eigenschaften. Wir wollen hier die Idee des Score-Tests zur Konstruktion von Rang-Tests für den Vergleich von zwei Verteilungen ausnutzen. Dazu gehen wir von Lage- bzw. Skalenfamilien aus und leiten die Form der Rangtests her. Konkrete Wahlen der Verteilungen ergeben Tests, die zwar alle verteilungsfrei sind, aber – wie wir später sehen werden – geeignet für unterschiedliche Situationen.

Die Verteilung sei für den Lagevergleich durch die Dichte $f(x-\Delta)$ bzw. durch $f(x)$ charakterisiert, je nachdem, ob eine Stichprobenvariable zur ersten Stichprobe gehört oder aus der zweiten stammt. Für die Zähldichte des Rangvektors $\boldsymbol{R}=(R_1,\ldots,R_{n+m})$ erhalten wir:

$$f_{\boldsymbol{R}}(\boldsymbol{r};\Delta)=\int_A\cdots\int\left[\prod_{i=1}^{n}f(x_i-\Delta)\right]\left[\prod_{i=n+1}^{n+m}f(x_i)\right]\mathrm{d}x_1\cdots\mathrm{d}x_{n+m}.$$

Dabei ist $A=\{(x_1,\ldots,x_{n+m})|r(x_1)=r_1,\ldots,r(x_{n+m})=r_{n+m}\}$. Setzen wir $y_{r_i}=x_{r_i:n+m}$, so ergibt dies weiter:

$$\begin{aligned}f_{\boldsymbol{R}}(\boldsymbol{r};\Delta)&=\int_{y_1<\cdots<y_{n+m}}\cdots\int\left[\prod_{i=1}^{n}\frac{f(y_{r_i}-\Delta)}{f(y_{r_i})}\right]\left[\prod_{i=1}^{n+m}f(y_{r_i})\right]\mathrm{d}y_1\cdots\mathrm{d}y_{n+m}\\&=\frac{1}{(n+m)!}\int_{y_1<\cdots<y_{n+m}}\cdots\int\left[(n+m)!\prod_{i=1}^{n}\frac{f(y_{r_i}-\Delta)}{f(y_{r_i})}\right]\cdot\left[\prod_{i=1}^{n+m}f(y_{r_i})\right]\mathrm{d}y_1\cdots\mathrm{d}y_{n+m}\\&=\frac{1}{(n+m)!}\mathrm{E}\left[\prod_{i=1}^{n}\frac{f(Y_{r_i}-\Delta)}{f(Y_{r_i})}\right].\end{aligned}$$

Da $(n+m)!\prod_{i=1}^{n+m}f(y_{r_i})$ die Dichte der geordneten Statistik von $n+m$ unabhängigen identisch verteilten Zufallsvariablen ist, ist der Erwartungswert E_0 entsprechend zu interpretieren.

Die Score-Teststatistik auf der Basis der Ränge erhalten wir als Ableitung an der Stelle Null der logarithmierten Zähldichte. Zunächst ist

$$\begin{aligned}U(\Delta)&=\frac{\partial\ln f_{\boldsymbol{R}}(\boldsymbol{r};\Delta)}{\partial\Delta}\\&=\frac{\frac{1}{(n+m)!}\mathrm{E}\left(\frac{\partial}{\partial\Delta}\prod_{i=1}^{n}\frac{f(Y_{r_i}-\Delta)}{f(Y_{r_i})}\right)}{\frac{1}{(n+m)!}\mathrm{E}\left(\prod_{i=1}^{n}\frac{f(Y_{r_i}-\Delta)}{f(Y_{r_i})}\right)}=\frac{\sum_{i=1}^{n}\mathrm{E}\left(\frac{-f'(Y_{r_i}-\Delta)}{f(Y_{r_i})}\prod_{j=1,j\neq i}^{n}\frac{f(Y_{r_j}-\Delta)}{f(Y_{r_j})}\right)}{\frac{1}{(n+m)!}\mathrm{E}\left(\prod_{i=1}^{n}\frac{f(Y_{r_i}-\Delta)}{f(Y_{r_i})}\right)}.\end{aligned}\tag{8.22}$$

An der Stelle $\Delta=0$ ist das gleich

$$U(0)=\sum_{i=1}^{n}\mathrm{E}\left[\frac{-f'(Y_{r_i})}{f(Y_{r_i})}\right].\tag{8.23}$$

Mit $a(r_i)=\mathrm{E}\left[-f'(Y_{r_i})/f(Y_{r_i})\right]$ bzw. $a(i)=\mathrm{E}\left[-f'(X_{i:n})/f(X_{i:n})\right]$ lässt sich die Teststatistik auch in der Form $\sum_{i=1}^{n}c_i a(R_i)$ angeben.

Wenn wir von der Likelihoodfunktion selbst ausgehen, erhalten wir übrigens für die Scores $-f'(x_i)/f(x_i)$.

Beispiel 8.2.19 *Rangscores auf Basis der logistischen Verteilung*

Wir betrachten den Fall, dass die Scores durch die logistische Verteilung mit der Dichte

$$f(x;\mu)=\frac{e^x}{(1+e^{x-\mu})^2} \qquad -\infty<x<\infty$$

erzeugt werden. Hier ist $f'(x)/f(x)=1-2e^x/(1+e^x)=1-2F(x)$. Dabei ist $F(x)$ die zu $f(x;0)$ gehörige Verteilungsfunktion. $F(X)$ ist nun gleichverteilt über $[0,1]$. Daher gilt $\mathrm{E}(F(X_{i:n+m}))=i/(n+m+1)$. Folglich erhalten wir

$$a(i)=\frac{2i}{n+m+1}-1.$$

Diese Scores sind offensichtlich gleichwertig mit denen der Wilcoxon-Teststatistik.

Die Behandlung des Skalenproblems erfolgt ganz analog zum Lageproblem. An die Stelle von $f(x-\Delta)$ tritt $f(x/\sigma)$. Das führt dann zu

$$U(1)=\sum_{i=1}^{n}\mathrm{E}\left[Y_{r_i}\frac{-f'(Y_{r_i})}{f(Y_{r_i})}\right]. \tag{8.24}$$

Mit $a(i)=\mathrm{E}\left[X_{i:n+m}\dfrac{-f'(X_{i:n+m})}{f(X_{i:n+m})}\right]$ lässt sich die Teststatistik auch in der Form $\sum\limits_{i=1}^{n+m}c_i a(R_i)$ angeben.

Beispiel 8.2.20 *Savage-Scores*

Sei X exponentialverteilt mit dem Parameter λ, $f(x;\lambda)=\lambda e^{-\lambda x}$, $x>0$. λ ist hier ein Skalenparameter. Für $\lambda=1$ erhalten wir:

$$\frac{f'(y)}{f(y)}=\frac{-e^{-x}}{e^{-x}}=-1.$$

Damit ist:

$$a(i)=\mathrm{E}(X_{i:n+m})=\sum_{j=1}^{i}\frac{1}{n+m-j+1}-1.$$

Der auf diesen *Savage-Scores* basierende Test wird als *Lograng-Test* bezeichnet.

Satz 8.2.21 *Erwartungswert und Varianz einer linearen Rangstatistik*

Seien $(c_1,\ldots,c_n)$ und $(a(1),\ldots,a(n))$ zwei gegebene Vektoren von Konstanten mit

$$\bar{a}=\frac{1}{n}\sum_{i=1}^{n}a(i),\quad \bar{c}=\frac{1}{n}\sum_{i=1}^{n}c_i,\quad \sigma_a^2=\frac{1}{n-1}\sum_{i=1}^{n}[a(i)-\bar{a}]^2.$$

Weiter seien $(r_1,\ldots,r_n)$ eine Zufallspermutation von $(1,\ldots,n)$ und

$$T=\sum_{i=1}^{n}c_i a(r_i).$$

Dann gilt

$$\mathrm{E}(T) = n \cdot \bar{a} \cdot \bar{c}, \quad \mathrm{Var}(T) = \sigma_a^2 \sum_{i=1}^{n} [c_i - \bar{c}]^2,$$

wobei Erwartungswert und Varianz bzgl. der Gleichverteilung über alle Permutationen gebildet werden.

Beweis: Siehe Hajek/Sidak (1967, S. 61).

Lineare Rangstatistiken sind unter H_0 bei nur schwachen Annahmen asymptotisch normalverteilt. Für eine exakte Formulierung der asymptotischen Aussagen sind die Regressionskonstanten sowie die Scores vom Stichprobenumfang abhängig zu machen. Wir schreiben daher für die Regressionskonstanten $c_n(i)$ statt c_i. Für sie wird nur gefordert, dass keine einzelne Konstante die anderen dominiert. Die entsprechende *Noether-Bedingung* lautet

$$\frac{\sum_{i=1}^{n} [c_n(i) - \bar{c}_n]^2}{\max\limits_{1 \le i \le n} [c_n(i) - \bar{c}_n]^2} \to \infty. \tag{8.25}$$

Für die Scores $a(i)$ unterstellen wir die folgende Form:

$$a(i) = u + v \cdot \varphi\left(\frac{i}{n+1}\right) \qquad 1 \le i \le n.$$

Dabei ist $\varphi : (0,1) \to \mathbb{R}$ eine schwach monoton wachsende Funktion oder schwach monoton fallend auf $(0, t_0)$ und schwach monoton wachsend auf $(t_0, 1)$. Weiter wird für φ die Integrationsbedingung

$$0 < \int_0^1 [\varphi(t) - \bar{\varphi}]^2 \mathrm{d}t < \infty$$

mit $\bar{\varphi} = \int_0^1 \varphi(t)\mathrm{d}t$ vorausgesetzt. Wir bezeichnen dann φ als *quadratisch integrierbare Scorefunkion.*

Beispiel 8.2.22 *zum Rangsummentest von Wilcoxon*

Für den Rangsummentest von Wilcoxon ist

$$c_n(i) = \begin{cases} 1 & i = 1, \ldots, n \\ 0 & i = n+1, \ldots, n+m \end{cases}.$$

Die Noether-Bedingung wird hier mit $\bar{c} = (n-m)/n$ zu

$$\frac{n\left(1 - \frac{n}{n+m}\right)^2 + m\left(0 - \frac{n}{n+m}\right)^2}{\max\left\{\left(1 - \frac{n}{n+m}\right)^2, \left(0 - \frac{n}{n+m}\right)^2\right\}} = \frac{n \cdot m \cdot (n+m)}{\max\{n^2, m^2\}} \to \infty.$$

Das ist erfüllt, wenn n und m gegen unendlich gehen.
Für die zu den Scores des Wilcoxon Tests gleichwertigen Scores aus dem Beispiel 8.2.15 erhalten wir über die Scorefunktion $\varphi(t)=2t-1$:

$$\bar{\varphi}=\int_0^1(2t-1)\,\mathrm{d}t=0, \qquad \int_0^1(2t-1)^2\mathrm{d}t=\frac{4}{3}.$$

Damit ist hier die Voraussetzung der quadratischen Integrierbarkeit der monoton wachsenden Scorefunktion erfüllt.

Beispiel 8.2.23 *zum Ansary-Bradley-Test*

Die Regressionskonstanten der Rangtests im Beispiel 8.2.18 für das Zweistichproben-Skalenproblem sind die gleichen wie im vorangehenden Beispiel. Die Scores des Ansary-Bradley-Tests sind gleichwertig mit den Scores

$$a(i)=\frac{1}{2}-\left|t-\frac{1}{2}\right|,$$

zu denen die Scorefunktion $\varphi(t)=|t-1/2|$ gehört. $\varphi(t)$ erfüllt die andere Monotonieforderung mit $t_0=1/2$.

Die Scores des Lograng-Tests sind jedoch nicht von der hier angegebenen Form.

Satz 8.2.24 *asymptotische Normalverteilung linearer Rangstatistiken*

Sei T_n eine lineare Rangstatistik der Form

$$T_n=\sum_{i=1}^{n}c_n(i)\left(u+v\cdot\varphi\left(\frac{r_i}{n+1}\right)\right) \tag{8.26}$$

mit einer quadratisch integrierbaren Scorefunktion, die die oben angegebene Monotonieeigenschaft besitzt. Wenn die Regressionskonstanten die Noether-Bedingung erfüllen, ist T asymptotisch normalverteilt,

$$\mathrm{P}\left(\frac{T_n-n\cdot\bar{a}_n\cdot\bar{c}_n}{\sigma_{na}}\le t\right)=\Phi(t).$$

Dabei ist

$$\sigma_{na}^2=\sum_{i=1}^{n}[c_n(i)-\bar{c}_n]^2\int_0^1[\varphi(t)-\bar{\varphi}]^2\mathrm{d}t. \tag{8.27}$$

Beweis: Hajek/Sidak (1967).

Siehe Hajek (1969, S. 84) zur asymptotischen Normalverteilung der Lograng-Teststatistik unter H_0.

Für kleinere Stichprobenumfänge lässt sich die exakte Verteilung der linearen Rangteststatistiken zum Vergleich zweier Stichproben leicht berechnen. Grundlage dafür ist die Gleichverteilung der Permutationen der Ränge unter der Nullhypothese $H_0 : \Delta = 0$ bzw. $H_0 : \sigma = 1$.

$$\mathrm{P}((R_1,\ldots,R_{n+m}) = (r_1,\ldots,r_{n+m})) = \frac{1}{(n+m)!}. \tag{8.28}$$

Daher gilt

$$\mathrm{P}(T_{n,m} = t) = \frac{n!\,m!\,h_{n+m}(t,n)}{(n+m)!}. \tag{8.29}$$

Dabei ist n die Anzahl der 1'en im Vektor $\boldsymbol{x}$ und m die der 0'en. $h_{n+m}(t,m)$ bezeichnet die Anzahl der Teilmengen $\{i_1,\ldots,i_n\}$ vom Umfang n aus $\{1,\ldots,n+m\}$ mit

$$t = a(i_1) + \ldots + a(i_n).$$

Zur rekursiven Berechnung der Verteilung von $T_{n,m}$ bei gegebenen n, m betrachten wir $n+m$ unabhängige $\mathscr{B}(1,1/2)$-verteilte Zufallsvariablen X_i. Dann ist mit $T = \sum X_i a(i)$:

$$\mathrm{P}(T_{n,m} = t) = \mathrm{P}(T = t | \Sigma X_i = n).$$

Für die gemeinsame Verteilung von T und $\sum X_i$ gilt:

$$\begin{aligned}\mathrm{P}\left(T = t, \sum X_i = n\right) &= \mathrm{P}\left(T = t \,\middle|\, \sum X_i = n\right) \mathrm{P}\left(\sum X_i = n\right) \\ &= \frac{n!\,m!\,h_{n+m}(t,n)}{(n+m)!}\binom{n+m}{n}\left(\frac{1}{2}\right)^{n+m} = h_{n+m}(t,n)\left(\frac{1}{2}\right)^{n+m}.\end{aligned}$$

Zudem ist

$$\begin{aligned}\mathrm{P}\left(\sum_{i=1}^{n+m} a(i)X_i = t, \sum_{i=1}^{n+m} X_i = n\right) =& \mathrm{P}\left(\sum_{i=1}^{n+m-1} a(i)X_i = t - a(n+m), \sum_{i=1}^{n+m-1} X_i = n-1, X_{n+m} = 1\right) \\ &+ \mathrm{P}\left(\sum_{i=1}^{n+m-1} a(i)X_i = t, \sum_{i=1}^{n+m-1} X_i = n, X_{n+m} = 0\right) \\ =& \frac{1}{2}\left\{\mathrm{P}\left(\sum_{i=1}^{n+m-1} a(i)X_i = t - a(n+m), \sum_{i=1}^{n+m-1} X_i = n-1\right)\right. \\ &\left. + \mathrm{P}\left(\sum_{i=1}^{n+m-1} a(i)X_i = t, \sum_{i=1}^{n+m-1} X_i = n\right)\right\}.\end{aligned}$$

Das bedeutet für die Häufigkeiten $h_n(u,n)$:

$$h_{n+m}(t,n) = h_{n+m-1}(t - a(n+m), n-1) + h_{n+m-1}(t,n). \tag{8.30}$$

8.2.8 Anpassungstests

Als eine der gängigen Hypothesen haben wir die Festlegung einer Verteilung angegeben, ohne dass die zugelassenen Verteilungen sich auf eine parametrische Klasse einschränken ließen. Für dieses Anpassungsproblem kann nun ein Test auf der Basis des Satzes 7.4.5, des Satzes von Kolmogorov, angegeben werden. Die verteilungsfreie Grenzverteilung lässt sich zum Testen von Hypothesen vollständig spezifizierter Verteilungen ausnutzen. Der Satz von Glivenko-Cantelli sichert dann, dass der Test konsistent ist.

Für das anwendungsrelevante Problem des Testens einer Verteilungsfamilie ist zu beachten, dass die Teststatistik nicht mehr verteilungsfrei ist. Daher gehören zu jeder Verteilungsfamilie andere Grenzverteilungen. Speziell für Normalverteilungen ohne vorgegebene Werte für μ und σ^2 und für Exponentialverteilungen ohne bekannten Parameter λ sind kritische Werte mittels Simulationen bestimmt worden. Zu anderen, auf der empirischen Verteilungsfunktion basierenden Tests sei auf Durbin (1973) verwiesen.

Ein weiterer Anpassungstest geht von einem Vergleich von erwarteten und beobachteten Häufigkeiten aus. Er ist für diskrete bzw. für klassierte stetige Verteilungen sinnvoll. Die Basis liefert

Satz 8.2.25 *Pearsons X^2-Statistik*

$(N_1,\dots,N_k)$ sei $\mathcal{M}(n,p_1,\dots,p_k)$-verteilt. Dann ist die Statistik

$$X^2=\sum_{i=1}^{k}\frac{(N_i-np_i)^2}{np_i} \tag{8.31}$$

für $n\to\infty$ asymptotisch χ^2-verteilt mit $k-1$ Freiheitsgraden.

Beweis: Wir führen den Beweis mittels Induktion nach der Anzahl der Klassen k. Für $k=2$ haben wir die Umformung

$$\begin{aligned} X^2 &= \sum_{i=1}^{2}\frac{(N_i-np_i)^2}{np_i}=\frac{(N_1-np_1)^2}{np_1}+\frac{(n-N_1-n(1-p_1))^2}{n(1-p_1)} \\ &= \frac{(N_1-np_1)^2}{np_1(1-p_1)}=\left(\frac{\frac{N_1}{n}-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}\right)^2. \end{aligned}$$

Mit dem zentralen Grenzwertsatz folgt die Behauptung in diesem Fall unmittelbar.

Die Induktionsvoraussetzung für festes $k>2$ hat als Konsequenz, dass bei $k+1$ Klassen mit festgehaltenem N_1 die Summe

$$S=\sum_{i=2}^{k+1}\frac{(N_i-(n-N_1)p_i')^2}{(n-N_1)p_i'}$$

mit $p_i' = p_i/(1-p_1)$ für große $(n-N_1)$ approximativ χ^2_{k-1}-verteilt ist. Um diese Induktionsvoraussetzung verwenden zu können, zerlegen wir X^2 in S und ein Restglied:

$$X^2 = S + \frac{(N_1 - np_1)^2}{np_1(1-p_1)} + \frac{(N_1-np_1)^2}{np_1} + \frac{p_1^{3/2}(1-p_1)^{1/2}\sqrt{n}}{n-N_1}\left(\frac{N_1-np_1}{\sqrt{np_1(1-p_1)}}\right)^3$$
$$- \frac{\sqrt{n}}{n-N_1}\sqrt{p_1(1-p_1)}\frac{N_1-np_1}{\sqrt{np_1(1-p_1)}}\sum_{i=2}^{k+1}(1-p_i)\frac{(N_i-np_i)^2}{np_i(1-p_i)}.$$

Aufgrund der Induktionsvoraussetzung ist S asymptotisch unabhängig von $(N_1-np_1)^2/(np_1(1-p_1))$. Mit der asymptotischen χ^2_1-Verteilung dieses Summanden folgt die Behauptung aus der Reproduktionseigenschaft der χ^2-Verteilung, wenn der dritte und vierte Term auf der rechten Seite für $n \to \infty$ gegen Null gehen. Dies ist aber erfüllt, weil dann $\sqrt{n}/(n-N_1)$ gegen Null geht und es für jedes $\epsilon > 0$ eine Konstante r gibt, so dass die Zufallsvariablen A und B in den beiden relevanten Termen die Beziehung $P(|A| \le r) \ge 1-\epsilon$, $P(|B| \le r) \ge 1-\epsilon$ erfüllen.

Will man mit diesem Test die Nullhypothese einer vollständig spezifizierten stetigen Verteilung überprüfen, so ist von einer geeigneten Klasseneinteilung auszugehen, um das Problem in den Rahmen einer Multinomialverteilung zu stellen. Dann gilt $p_i = F_0(x_i^*) - F_0(x_{i-1}^*)$. Für Anwendungszwecke gibt es die Faustregel, dass die Approximation hinreichend genau ist, wenn der Stichprobenumfang n so groß ist, dass für alle Wahrscheinlichkeiten gilt $np_i \ge 5$. Gegebenenfalls müssen auch bei diskreten Verteilungen Kategorien zusammengelegt werden.

Der große Vorteil des χ^2-Anpassungstests besteht darin, dass die asymptotische Null-Verteilung auch bekannt ist, wenn Parameter geschätzt werden müssen. Grob gesprochen reduziert sich dann die Anzahl der Freiheitgrade um die Anzahl der geschätzten Parameter. Siehe dazu Fisz (1978, S. 512f).

8.2.9 Testkonstruktion und Testanwendung

Wir haben in diesem Abschnitt verschiedene Konstruktionsprinzipien für statistische Tests vorgestellt. Wesentlich sind dabei natürlich auch die zuvor überblicksmäßig angegebenen Hypothesen im Zusammenhang mit den zugrundeliegenden Verteilungen. Im folgenden Abschnitt werden wir uns mit der Güte von Tests auseinandersetzen. Zuvor soll aber noch darauf hingewiesen werden, dass der Testkonstruktion und der Suche nach einem optimalen Test die Testanwendung gegenübersteht. Wie die auf Wilrich (1979) zurückgehende tabellarische Übersicht deutlich macht, sind für die Anwendung verschiedene Fragen zu klären, für die die theoretischen Ergebnisse grundlegend sind. So etwa die Frage nach dem Stichprobenumfang, die ohne Kenntnis der Gütefunktion nicht beantwortbar ist. Daneben gibt es zahlreiche weitere Probleme wie etwa die Umsetzung der Stichprobenziehung. Diese sparen wir hier aus.

Testkonstruktion	*Testanwendung*	
1. Voraussetzungen über die Grundgesamtheit	Sachproblem formulieren a. Voraussetzungen klären b. zu testende Hypothesen aufstellen	Problem-formulierung
2. Hypothesen formulieren	Zu verwendendes Testverfahren auswählen	Planung
3. Teststatistik T konstruieren	Signifikanzniveau α festlegen (Berücksichtigung der Folgen des Fehlers 1. Art)	
4. kritischen Bereich abgrenzen	Stichprobenumfang festlegen (Berücksichtigung der Folgen des Fehlers 2. Art bzw. der Kosten)	
5. Entscheidungsregel formulieren	Stichprobe ziehen	Daten-gewinnung
6. Gütefunktion bestimmen	Wert der Teststatistik errechnen	Auswertung
7. Eigenschaften (Konsistenz,...) untersuchen	Entscheidungsregel anwenden	
8. Effizienz mit anderen Tests vergleichen	Entscheidung deuten	

8.3 Gleichmäßig beste Tests

Wir haben in Abschnitt 8.1 gesehen, dass die Gütefunktion eines Tests die zentrale Größe für die Bewertung seiner Eigenschaften ist. Die dortige Diskussion weist auch darauf hin, dass jeweils der Test mit der größeren Güte vorzuziehen ist; bei seiner Verwendung besteht ja die größere Chance, tatsächliche Abweichungen von der Nullhypothese auch zu entdecken.

Ist die Alternative Θ_1 zusammengesetzt, so ist die Entscheidung zwischen zwei Tests nur dann leicht, wenn sich die Gütefunktionen in Θ_1 nicht schneiden, sondern die Güte des einen Tests gleichmäßig größer ist als die des anderen.

Beispiel 8.3.1 *einseitiges Testproblem für μ - Fortsetzung von Seite 257*

Für den Parameter μ einer $\mathcal{N}(\mu,\sigma)$-Verteilung wurde im Beispiel 8.1.6 der Gauß-Test

$$\varphi_1(\boldsymbol{x}) = \begin{cases} 0 & \text{für } \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \leq c \\ 1 & \text{für } \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > c \end{cases} \tag{8.32}$$

für das Testproblem $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$ diskutiert. Für die Gütefunktion des Tests erhielten wir:

$$\mathrm{E}_\mu \varphi_1(\boldsymbol{X}) = 1 - \Phi\left(c + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right).$$

c ist so zu bestimmen, dass $1 - \Phi(c) = \alpha$.

Bei einer Stichprobe vom Umfang n ist die Anzahl Y der Beobachtungen, die größer sind als μ_0, binomialverteilt mit den Parametern n und $p = \mathrm{P}(X > \mu_0)$. Ist der wahre Parameterwert μ größer als μ_0, so werden wegen $\mathrm{P}(\mu_0 < X < \mu) > 0$ etliche Beobachtungen zwischen μ und μ_0 liegen. Folglich werden unter H_1 mehr Beobachtungen größer als μ_0 sein als unter H_0. Damit erhalten wir einen zweiten Test für unser Testproblem:

$$\varphi_2(\boldsymbol{x}) = \begin{cases} 1 & \text{für } Y = \sum 1_{(-\infty,\mu_0]}(x_i) = y > c' \\ 0 & \text{für } Y = \sum 1_{(-\infty,\mu_0]}(x_i) = y \leq c' \end{cases}. \tag{8.33}$$

c' ist entsprechend dem vorgegebenem Niveau α und dem Stichprobenumfang n aus der $\mathscr{B}(n, 0.5)$-Verteilung zu ermitteln. Bei der Normalverteilung ist ja μ zugleich der Median, $\mathrm{P}_\mu(X > \mu) = 0.5$.

Der Test (8.33) wird als *Zeichentest* bezeichnet. Der Zeichentest ist natürlich nicht an die Voraussetzung der Normalverteilung gebunden.

Y hat unter der Alternative dann wieder eine Binomialverteilung, jedoch mit dem Parameter $p = \mathrm{P}_\mu(X > \mu_0)$. Die Gütefunktion von φ_2 ist dann

$$\mathrm{E}_\mu \varphi_2(\boldsymbol{X}) = \mathrm{P}_\mu(\varphi_2(\boldsymbol{X}) = 1) = \mathrm{P}_\mu(Y > c') = \sum_{i=c'+1}^{n} \binom{n}{i} p^i (1-p)^{n-i}.$$

Somit liegen nunmehr mit φ_1 und φ_2 zwei Tests zum Niveau $\alpha = 0.002$ für das Testproblem $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$ vor.

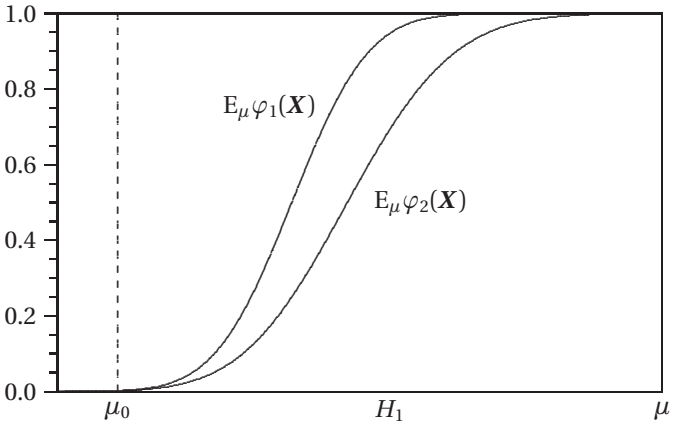


Abb. 8.7: *Gütefunktionen des Gauß- und des Zeichentests*

Die Gütefunktion der Tests ist für die Werte $\alpha = 0.002$, $n = 25$ und den resultierenden Wert $c' = 18$ in der Abbildung 8.7 dargestellt. Die Graphik zeigt, dass die Gütefunktion des Tests φ_1 im ganzen Bereich der Alternative über der des Tests φ_2 liegt. Egal welcher Wert von μ unter H_1 richtig ist, stets ist die Chance dies zu erkennen, mit φ_1 größer.

Allgemein zeichnen wir Tests, die alle Konkurrenten bzgl. der Güte dominieren, als optimal aus.

Definition 8.3.2 *gleichmäßig bester Test*

Gegeben sei eine Stichprobe $\boldsymbol{X} = (X_1, \ldots, X_n)$ aus der vom Parameter $\vartheta \in \Theta$ abhängigen Verteilung P_ϑ. Ein Test φ^* heißt *gleichmäßig bester Test zum Niveau α*, kurz *GB-Test*, für das Testproblem

$$H_0 : \vartheta \in \Theta_0, \qquad H_1 : \vartheta \in \Theta \backslash \Theta_0,$$

wenn er ein Test zum Niveau α ist, d.h. wenn er für alle $\vartheta \in \Theta_0$

$$\text{(i)} \quad \mathrm{E}_\vartheta \varphi^*(\boldsymbol{X}) \leq \alpha$$

erfüllt und wenn für alle Tests $\varphi(\boldsymbol{X})$ zum Niveau α für dieses Testproblem gilt

$$\text{(ii)} \quad \mathrm{E}_\vartheta \varphi^*(\boldsymbol{X}) \geq \mathrm{E}_\vartheta \varphi(\boldsymbol{X}) \qquad \text{für alle } \vartheta \in \Theta \backslash \Theta_0.$$

Im Fall $\Theta \backslash \Theta_0 = \{\vartheta_1\}$ sprechen wir von einem besten Test zum Niveau α.

Die Frage, für welche Situationen es GB-Tests gibt, studieren wir in den folgenden Abschnitten. Dies geschieht in der Form, dass wir in aufsteigendem Schwierigkeitsgrad Situationen angeben, für die geeignete Tests existieren. Wir werden sehen, dass wir zum Teil Einschränkungen bei den Konkurrenten vornehmen müssen, um die Existenz optimaler Tests zeigen zu können.

Auch im Bereich der Tests spielt die Suffizienz eine wichtige Rolle. Sie wird der Charakterisierung, dass eine suffiziente Statistik die gesamte Information einer Stichprobe erhält, hier dadurch gerecht, dass es zu jedem Test auf der Basis einer Stichprobe $\boldsymbol{X}$ einen Test gibt, der nur von der suffizienten Statistik abhängt und die gleiche Gütefunktion wie der auf $\boldsymbol{X}$ basierende besitzt. Bei der Konstruktion von und der Suche nach besten Test werden wir uns also auf solche beschränken können, die von einer suffizienten Statistik abhängen. Allerdings ist dies nur für randomisierte Tests korrekt. Da die Randomisierung nur bei diskreten Verteilungen tatsächlich bedeutsam ist, stellt es bei stetigen Verteilungen keine Einschränkung dar, wenn wir hier vor allem an nichtrandomisierte Tests denken.

Satz 8.3.3 *Testfunktion und Suffizienz*

Sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe aus einer Verteilung, die von einem Parameter $\vartheta \in \Theta$ abhängt. $T(\boldsymbol{X})$ sei eine für ϑ suffiziente Statistik. Ist $\varphi(\boldsymbol{x})$ ein Test für das Testproblem $H_0 : \vartheta \in \Theta_0$, $H_1 : \vartheta \in \Theta_1$, so gibt es einen Test ψ, der nur von T abhängt und der die gleiche Gütefunktion wie φ besitzt:

$$\forall \vartheta \in \Theta : \mathrm{E}_\vartheta \varphi(\boldsymbol{X}) = \mathrm{E}_\vartheta \psi(T).$$

Beweis: Wir setzen $\psi(t) = \mathrm{E}.(\varphi(\boldsymbol{X})|T(\boldsymbol{X}) = t)$. Wegen der Suffizienz von $T(\boldsymbol{X})$ ist ψ eine vom Parameter ϑ unabhängige Testfunktion, für die natürlich gilt: $0 \le \psi(t) \le 1$. Zudem gilt:

$$\forall \vartheta \in \Theta : \mathrm{E}_\vartheta \varphi(\boldsymbol{X}) = \mathrm{E}_\vartheta \mathrm{E}.(\varphi(\boldsymbol{X})|T(\boldsymbol{X})) = \mathrm{E}_\vartheta \psi(T(\boldsymbol{X})).$$

8.3.1 Einfache Hypothesen

Bei der Entwicklung der Situationen, für die optimale Tests existieren, wird zunächst das Problem des Testens einer *einfachen Hypothese* gegen eine *einfache Alternative* betrachtet. Dies ist keine so sehr praktisch relevante Problemstellung. Relevant sind aber Folgerungen aus seiner Lösung.

Das Testproblem sei also

$$H_0 : f(x) = f_0(x), \qquad H_1 : f(x) = f_1(x),$$

d.h. es soll auf Basis einer Stichprobe $\boldsymbol{X} = (X_1, \ldots, X_n)$ entschieden werden, ob die zugrundeliegende Zufallsvariable X die Dichtefunktion $f_0(x)$ oder $f_1(x)$ hat. Es ist klar, dass das Problem gleichwertig mit der gemeinsamen Dichte der Stichprobenvariablen formuliert werden kann:

$$H_0 : f(\boldsymbol{x}) = f_0(\boldsymbol{x}), \qquad H_1 : f(\boldsymbol{x}) = f_1(\boldsymbol{x}).$$

Der folgende Satz beinhaltet, dass für das hier betrachtete Testproblem stets ein bester Test existiert. Dieser ist ein Likelihood-Quotienten-Test. Weil $f_0(\boldsymbol{x})/\max\{f_0(\boldsymbol{x}), f_1(\boldsymbol{x})\} < c$ äquivalent ist mit $f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) > c'$, können wir den Ablehnbereich auch mittels des zweiten Quotienten beschreiben.

Satz 8.3.4 *Fundamentallemma von Neyman-Pearson*

Sei $\varphi^*(\boldsymbol{x})$ der Test, der im oben formulierten Testproblem H_1 ablehnt, wenn der Likelihood-Quotient den Wert k, $0 \le k \le \infty$, überschreitet:

$$\varphi^*(\boldsymbol{x}) = \begin{cases} 1 & \dfrac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} > k \\ 0 & \dfrac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} \le k \end{cases} . \tag{8.34}$$

Dann ist φ^* der beste Test zum Niveau $\alpha := \mathrm{P}_0(f_1(\boldsymbol{X})/f_0(\boldsymbol{X}) > k)$, für jeden Test ψ mit $\mathrm{E}_0\psi(\boldsymbol{X}) \le \alpha$ gilt also

$$\mathrm{E}_1\psi(\boldsymbol{X}) \le \mathrm{E}_1\varphi^*(\boldsymbol{X}).$$

Beweis: Es wird gezeigt, dass für jede Funktion $\psi(\boldsymbol{x})$ mit $0 \le \psi(\boldsymbol{x}) \le 1$ und $\mathrm{E}_0\psi(\boldsymbol{X}) \le \alpha$ die behauptete Ungleichung erfüllt ist. Damit gilt die Aussage auch für randomisierte Tests. Wir führen den stetigen Fall aus. Der diskrete geht analog.

Sei also ψ entsprechend gegeben. Der Stichprobenraum wird gemäß der unterschiedlichen Definition von φ^* in zwei Bereiche zerlegt:

$$A = \left\{\boldsymbol{x} | f_1(\boldsymbol{x}) > k \cdot f_0(\boldsymbol{x})\right\} \quad \text{und} \quad B = \left\{\boldsymbol{x} | f_1(\boldsymbol{x}) \le k \cdot f_0(\boldsymbol{x})\right\}.$$

Die Differenz der Erwartungswerte wird entsprechend aufgespalten:

$$\begin{aligned}\mathrm{E}_1[\varphi^*(\boldsymbol{X})-\psi(\boldsymbol{X})] &= \int_{\mathbb{R}^n}[\varphi^*(\boldsymbol{x})-\psi(\boldsymbol{x})]f_1(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &= \int_A[\varphi^*(\boldsymbol{x})-\psi(\boldsymbol{x})]f_1(\boldsymbol{x})\,d\boldsymbol{x}+\int_B[\varphi^*(\boldsymbol{x})-\psi(\boldsymbol{x})]f_1(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=a+b.\end{aligned}$$

Für $\boldsymbol{x}\in A$ gilt nun $\varphi^*(\boldsymbol{x})=1\geq\psi(\boldsymbol{x})$. Mit der Definition von A folgt daher:

$$a\geq k\int_A[\varphi^*(\boldsymbol{x})-\psi(\boldsymbol{x})]f_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Wegen $\varphi^*(\boldsymbol{x})=0$ für $\boldsymbol{x}\in B$ ist

$$\varphi^*(\boldsymbol{x})-\psi(\boldsymbol{x})\leq 0\quad\text{für alle }\boldsymbol{x}\in B.$$

Außerdem ist $f_1(\boldsymbol{x})\leq k\cdot f_0(\boldsymbol{x})$ für $\boldsymbol{x}\in B$ und $k>0$. Somit folgt:

$$b\geq k\int_B[\varphi^*(\boldsymbol{x})-\psi(\boldsymbol{x})]f_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Zusammen ergibt dies die Behauptung:

$$\begin{aligned}\mathrm{E}_1\varphi^*(\boldsymbol{X})-\mathrm{E}_1\psi(\boldsymbol{X}) &= \mathrm{E}_1[\varphi^*(\boldsymbol{X})-\psi(\boldsymbol{X})]=a+b\\ &\geq k\int_{\mathbb{R}^n}[\varphi^*(\boldsymbol{x})-\psi(\boldsymbol{x})]f_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=k\cdot\mathrm{E}_0[\varphi^*(\boldsymbol{X})-\psi(\boldsymbol{X})]\\ &= k\cdot\{\alpha-\mathrm{E}_0\psi(\boldsymbol{X})\}\geq k\cdot 0=0.\end{aligned}$$

Die letzte Ungleichung folgt dabei aus der Voraussetzung, dass $\psi(\boldsymbol{x})$ ein Test zum Niveau α ist, also $\mathrm{E}_0\psi(\boldsymbol{X})\leq\alpha$ gilt.

Die Formulierung des Fundamentallemmas erfolgte unter Berücksichtigung des Sachverhaltes, dass bei diskreten Zufallsvariablen der Likelihood-Quotient nur diskrete Werte annehmen kann und somit nicht zu jedem Niveau $0\leq\alpha\leq 1$ ein bester (nichtrandomisierter) Test existiert. Im stetigen Fall ist es dagegen möglich, zu vorgegebenem α die Schranke k mit $\mathrm{P}_0(f_1(\boldsymbol{X})>k\cdot f_0(\boldsymbol{X}))=\alpha$ zu bestimmen. Dann gibt es zu jedem vorgegebenem Niveau einen besten Test.

Korollar 8.3.5 *Unverfälschtheit des besten Tests für einfache Hypothesen*

Der beste Test aus Satz 8.3.4 ist *unverfälscht*, d.h. es gilt:

$$\mathrm{E}_1\varphi^*(\boldsymbol{X})\geq\mathrm{E}_0\varphi^*(\boldsymbol{X}).\tag{8.35}$$

Beweis: Wir setzen im Beweis von Satz 8.3.4 $\psi(\boldsymbol{x})\equiv\alpha$. Dann folgt:

$$\mathrm{E}_1\varphi^*(\boldsymbol{X})\geq\mathrm{E}_1\psi(\boldsymbol{X})=\mathrm{E}_1\alpha=\alpha.$$

Beispiel 8.3.6 *Einfache Hypothesen bei Normalverteilung - Fortsetzung von Seite 267*

Im Beispiel 8.1.2 wurde das Testproblem bzgl. des Parameters μ einer normalverteilten Zufallsvariablen X betrachtet:

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu = \mu_1 \qquad (\mu_0 < \mu_1).$$

Die Varianz war als bekannt vorausgesetzt, $\sigma^2 = 1$. Es wurde von einer Stichprobe vom Umfang n ausgegangen.

Das Problem kann umformuliert werden zu

$$H_0 : f(\boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu_0)^2\right)$$
$$H_1 : f(\boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu_1)^2\right).$$

Um den besten Test zum Niveau α für dieses Problem zu ermitteln, wird zunächst $k > 0$ bestimmt mit

$$\mathrm{P}_0(f_1(\boldsymbol{X}) > k \cdot f_0(\boldsymbol{X})) = \alpha.$$

Es gilt

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu_1)^2\right) > k \cdot \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu_0)^2\right)$$
$$\iff -\sum_{i=1}^n (x_i - \mu_1)^2 > \ln(k) - \sum_{i=1}^n (x_i - \mu_0)^2 \iff \frac{1}{n}\sum_{i=1}^n x_i > \frac{\ln(k) + n(\mu_1^2 - \mu_0^2)}{2n(\mu_1 - \mu_0)} = c.$$

Dann ist c aus $\mathrm{P}_0(\sum_{i=1}^n X_i > c) = \alpha$ zu bestimmen. Dies ist gerade der im Beispiel 8.1.2 heuristisch motivierte Test.

8.3.2 Einseitige Hypothesen

Das Fundamentallemma von Neyman-Pearson soll nun ausgenutzt werden, um gleichmäßig beste Tests für Testprobleme der Form

$$H_0 : \vartheta \le \vartheta_0, \qquad H_1 : \vartheta > \vartheta_0 \qquad (\text{bzw. } H_0 : \vartheta \ge \vartheta_0, \quad H_1 : \vartheta < \vartheta_0)$$

zu bestimmen. Zunächst behandeln wir diese Frage in Form eines Beispiels.

Beispiel 8.3.7 *Testproblem bei Poisson-Verteilung*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine Zufallsstichprobe aus einer Poisson-Verteilung. Als Testproblem liege vor:

$$H_0 : \lambda = \lambda_0, \qquad H_1 : \lambda > \lambda_0.$$

Mit dem Fundamentallemma erhalten wir den besten Test zum Niveau α für das reduzierte Problem

$$H_0 : \lambda = \lambda_0, \qquad H_1' : \lambda = \lambda_1,$$

wobei $\lambda_1 > \lambda_0$ gewählt wird. Dafür ist ein $k > 0$ zu bestimmen mit

$$\mathrm{P}_{\lambda_0}(f(\boldsymbol{X};\lambda_1) > k \cdot f(\boldsymbol{X};\lambda_1)) \leq \alpha.$$

Es gilt, wenn $s = \sum_{i=1}^n x_i$ gesetzt wird:

$$f(\boldsymbol{x};\lambda) = e^{-n\lambda} \frac{\lambda^s}{x_1! \cdot \ldots \cdot x_n!}.$$

Damit folgt:

$$e^{-n\lambda_1} \frac{\lambda_1^s}{x_1! \cdot \ldots \cdot x_n!} > k \cdot e^{-n\lambda_0} \frac{\lambda_0^s}{x_1! \cdot \ldots \cdot x_n!} \iff \left(\frac{\lambda_1}{\lambda_0}\right)^s > k \cdot e^{n(\lambda_1 - \lambda_0)}$$

$$\iff s \cdot \ln\left(\frac{\lambda_1}{\lambda_0}\right) > \ln(k) + n(\lambda_1 - \lambda_0) \iff s > \frac{\ln(k) + n(\lambda_1 - \lambda_0)}{\ln(\lambda_1/\lambda_0)} = k'.$$

Da $S = \sum_{i=1}^n X_i$ eine Poisson-Verteilung hat mit dem Parameter $n \cdot \lambda$, kann k' entsprechend bestimmt werden aus

$$\mathrm{P}_{n\lambda_0}(S > k') \leq \alpha.$$

Der beste Test für H_0 gegen H_1' ist dann:

$$\varphi^*(\boldsymbol{x}) = \begin{cases} 1 & s = \sum_{i=1}^n x_i > k' \\ 0 & s = \sum_{i=1}^n x_i \leq k' \end{cases}.$$

λ_1 wurde mit $\lambda_1 > \lambda_0$ bei der Bestimmung von $\varphi^*(\boldsymbol{x})$ als fest vorgegeben vorausgesetzt. Betrachten wir die Herleitung, die zu φ^* führt, und den Test selbst, so sehen wir, dass zwei Punkte wesentlich sind:

- Die Form des Ablehnbereichs ergibt sich mit $\lambda_0 < \lambda_1$ wegen der Äquivalenz $f(\boldsymbol{x};\lambda_1) \gtrless k \cdot f(\boldsymbol{x};\lambda_0) \Leftrightarrow s = \sum x_i \gtrless k'$.
- Die Wahl von k' hängt nur vom Niveau α und vom hypothetischen Wert λ_0 ab. (k hängt auch von λ_1 ab. Aber verschiedene λ_1 führen zum gleichen k'.

Daraus folgt nun, dass für jede Wahl von λ_1 mit $\lambda_1 > \lambda_0$ $\varphi^*(\boldsymbol{x})$ als bester Test resultiert. Die Gütefunktion von φ^* hat daher für alle $\lambda_1 > \lambda_0$ jeweils den größten Wert von allen Tests zum Niveau α für das jeweilige Problem $H_0 : \lambda = \lambda_0$ gegen $H_1' : \lambda = \lambda_1$. Somit ist φ^* sogar gleichmäßig bester Test zum Niveau α für das Testproblem

$$H_0 : \lambda = \lambda_0, \qquad H_1 : \lambda > \lambda_0.$$

Die nächste Frage ist, ob auch die Nullhypothese ‚aufgeblasen' werden kann zu $H_0 : \lambda \leq \lambda_0$. Dann wäre der nach dem Fundamentallemma bestimmte Test für einfache Hypothesen sogar gleichmäßig bester Test für die einseitigen Hypothesen. Dazu ist zu zeigen:

$$\mathrm{E}_{\lambda'} \varphi^*(\boldsymbol{X}) \leq \alpha \qquad \text{für alle } \lambda' < \lambda_0.$$

Dies ergibt sich aber sofort aus der Tatsache, dass $\varphi^*(\boldsymbol{x})$ aufgrund seiner Struktur nach den bisherigen Überlegungen bester Test ist für das Testproblem:

$$H_0' : \lambda = \lambda', \qquad H_1' : \lambda = \lambda_0.$$

Das Niveau ist dabei $\alpha' = \mathrm{E}_{\lambda'}\varphi^*(\boldsymbol{X})$. Da φ^* bester Test zum Niveau α' für H_0' gegen H_1' ist, folgt mit dem Korrollar 8.3.5:

$$\mathrm{E}_{\lambda'}\varphi^*(\boldsymbol{X}) \leq \mathrm{E}_{\lambda_0}\varphi^*(\boldsymbol{X}).$$

Für Poisson-Verteilungen wurde in dem Beispiel gezeigt, dass der beste Test für einfache Hypothesen sogar gleichmäßig bester Test für die einseitigen Hypothesen ist. Die Frage, die sich anschließt, ist natürlich folgende: Welche Eigenschaft der Poisson-Verteilung ist wesentlich, um die Erweiterung der Hypothesen zu ermöglichen?

Wie eine genauere Betrachtung zeigt, wurde ausschließlich verwendet, dass die besten Tests zu beliebigem Niveau α für

$$H_0 : \lambda = \lambda_0, \qquad H_1 : \lambda = \lambda_1$$

bei jeder Wahl von λ_0 und λ_1 mit $\lambda_0 < \lambda_1$ die gleiche, auf der Äquivalenz

$$\frac{f(\boldsymbol{x};\lambda_1)}{f(\boldsymbol{x};\lambda_0)} \gtreqless k \quad \Longleftrightarrow \quad \sum_{i=1}^{n} x_i \gtreqless k'$$

basierende Form haben. Die Äquivalenz beruht darauf, dass der Quotient auf der linken Seite monoton von der Prüfgröße $\sum x_i$ abhängt. Für zwei Stichproben $\boldsymbol{x}$ und $\boldsymbol{x}'$ gilt:

$$\sum_{i=1}^{n} x_i < \sum_{i=1}^{n} x_i' \quad \Longleftrightarrow \quad \frac{f(\boldsymbol{x};\lambda_1)}{f(\boldsymbol{x};\lambda_0)} < \frac{f(\boldsymbol{x}';\lambda_1)}{f(\boldsymbol{x}';\lambda_0)}.$$

Liegt diese Eigenschaft bei einer Verteilung (mit eventuell anderer Statistik $T(\boldsymbol{X})$) vor, so kann das für Poisson-Verteilungen erhaltene Resultat offenbar übertragen werden.

Definition 8.3.8 *monotoner Likelihoodquotient*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine n-dimensionale Zufallsvariable mit der gemeinsamen Dichtefunktion $f(\boldsymbol{x};\vartheta)$, $\vartheta \in \Theta \subset \mathbb{R}$. Dann haben die Verteilungen von $\boldsymbol{X}$ – oder kurz $\boldsymbol{X}$ selbst – einen *monotonen Likelihoodquotienten* in $T(\boldsymbol{X})$, wenn für jedes Paar ϑ_0, ϑ_1 mit $\vartheta_0 < \vartheta_1$ gilt:

$$\frac{f(\boldsymbol{x};\vartheta_1)}{f(\boldsymbol{x};\vartheta_0)} < \frac{f(\boldsymbol{x}';\vartheta_1)}{f(\boldsymbol{x}';\vartheta_0)} \quad \Longleftrightarrow \quad T(\boldsymbol{x}) < T(\boldsymbol{x}').$$

Beispiel 8.3.9 *monotoner Likelihoodquotient bei Exponentialfamilien*

Als eine große Klasse von Verteilungsfamilien haben einparametrige Exponentialfamilien einen monotonen Likelihoodquotienten in $\sum T(x_i)$, wenn bei der Dichtefunktion $f(x) = \exp[c(\vartheta)\cdot T(x) + d(\vartheta) + S(x)]\cdot I_A(x) \quad c(\vartheta)$ monoton wachsend in ϑ ist. Für $\vartheta_0 < \vartheta_1$ gilt dann nämlich für $(x_1,\ldots,x_n)\in A$:

$$\begin{aligned}\frac{f(\boldsymbol{x};\vartheta_1)}{f(\boldsymbol{x};\vartheta_0)} &= \frac{\exp[c(\vartheta_1)\cdot\sum T(x_i) + n\cdot d(\vartheta_1) + \sum S(x_i)]}{\exp[c(\vartheta_0)\cdot\sum T(x_i) + n\cdot d(\vartheta_0) + \sum S(x_i)]}\\ &= \exp\left[(c(\vartheta_1) - c(\vartheta_0))\cdot\sum T(x_i)\right]\cdot\exp[n\cdot(d(\vartheta_1) - d(\vartheta_0))].\end{aligned}$$

Mit der Monotonie der e-Funktion und mit $c(\vartheta_1) - c(\vartheta_0) > 0$ folgt die Behauptung unmittelbar.

Ist $c(\vartheta)$ monoton fallend, so hat die Familie einen monotonen Likelihoodquotienten in $-\sum T(x_i)$.

Ein bei Heyer (1973, S. 88) angegebener Satz zeigt, dass die Exponentialfamilien sich durch einen besonders ausgeprägten monotonen Likelihoodquotienten auszeichnen.

Nun kann die allgemeine Lösung der einseitigen Testprobleme formuliert werden.

Satz 8.3.10 *gleichmäßig bester Test bei monotonem Likelihoodquotienten*

$\boldsymbol{X} = (X_1,\ldots,X_n)$ sei eine n-dimensionale Zufallsvariable mit der Dichtefunktion $f(\boldsymbol{x};\vartheta)$, $\vartheta\in\Theta$ und $\Theta\subset\mathbb{R}$. Die Verteilung von $\boldsymbol{X}$ möge einen monotonen Likelihoodquotienten in $T(\boldsymbol{X})$ haben. Dann gibt es bei gegebenem α einen gleichmäßig besten Test $\varphi^*(\boldsymbol{x})$ zum Niveau $\alpha^*\le\alpha$ für das Testproblem

$$H_0:\ \vartheta\le\vartheta_0, \qquad H_1:\ \vartheta>\vartheta_0.$$

Er ist gegeben durch

$$\varphi^*(\boldsymbol{x}) = \begin{cases} 1 & T(\boldsymbol{x}) > k\\ 0 & T(\boldsymbol{x}) \le k\,.\end{cases}$$

Dabei wird k als größte Zahl bestimmt, für die

$$\mathrm{E}_{\vartheta_0}\varphi^*(\boldsymbol{X}) = \mathrm{P}_{\vartheta_0}(T(\boldsymbol{X}) > k) = \alpha^* \le \alpha.$$

Beweis: Es sind lediglich die Ergebnisse von Beispiel 8.3.7 allgemein zu formulieren. Seien also $\vartheta_1 > \vartheta_0$ beliebig. Dann gibt es einen besten Test für

$$H_0':\ \vartheta=\vartheta_0, \qquad H_1':\ \vartheta=\vartheta_1.$$

Wegen $\vartheta_0 < \vartheta_1$ gilt für gegebenes k:

$$\frac{f(\boldsymbol{x};\vartheta_1)}{f(\boldsymbol{x};\vartheta_0)} > k' \quad\Longleftrightarrow\quad T(\boldsymbol{x}) > k.$$

Der optimale Test hat nach Satz 8.3.4 die Form

$$\varphi^*(\boldsymbol{x}) = \begin{cases} 1 & T(\boldsymbol{x}) > k \\ 0 & T(\boldsymbol{x}) \le k \end{cases},$$

wobei k als der größte Wert bestimmt wird mit

$$\mathrm{P}_{\vartheta_0}(T(\boldsymbol{X}) > k) = \alpha^* \le \alpha.$$

Die Bestimmung von k ist unabhängig von der speziellen Wahl von ϑ_1. φ^* ist somit für alle $\vartheta_1 > \vartheta_0$ bester Test zum Niveau α^*, also GB-Test für

$$H_0' : \vartheta = \vartheta_0, \qquad H_1 : \vartheta > \vartheta_0.$$

Weiter gilt $\mathrm{E}_\vartheta \varphi^*(\boldsymbol{X}) \le \mathrm{E}_{\vartheta_0} \varphi^*(\boldsymbol{X})$ für $\vartheta \le \vartheta_0$. Dazu wird bei speziellem $\vartheta' < \vartheta_0$ das Testproblem

$$H_0'' : \vartheta = \vartheta', \qquad H_1'' : \vartheta = \vartheta_0$$

zum Niveau $\alpha' = \mathrm{E}_{\vartheta'} \varphi^*(\boldsymbol{X})$ betrachtet.

Der beste Test ist von der gleichen Struktur wie φ^*. Die Wahl von α' ergibt sogar, dass φ^* der beste Test zum Niveau α' für H_0'' gegen H_1'' ist. Nach Korollar 8.3.5 muss daher gelten:

$$\alpha' = \mathrm{E}_{\vartheta'} \varphi^*(\boldsymbol{X}) \le \mathrm{E}_{\vartheta_0} \varphi^*(\boldsymbol{X}) = \alpha^*.$$

φ^* ist also GB-Test für $H_0' : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$ und hält zusätzlich das Niveau ein für $\vartheta < \vartheta_0$. Daraus folgt die Behauptung.

Haben die Verteilungen von $\boldsymbol{X} = (X_1, \ldots, X_n)$ einen monotonen Likelihoodquotienten in $T(\boldsymbol{X})$, so auch in $\tilde{T}(\boldsymbol{X}) = h(T(\boldsymbol{X}))$, wenn $h(t)$ selbst eine streng monoton wachsende Funktion ist. Wir werden die Prüfgröße nach Möglichkeit so wählen, dass der kritische Wert k aus einer geeigneten Verteilungstabelle bestimmt werden kann. Dass der gleiche Test φ^* resultiert, ist offensichtlich.

Wir haben uns bei der Betrachtung der einseitigen Testprobleme auf solche der Art $H_0 : \vartheta \le \vartheta_0, \quad H_1 : \vartheta > \vartheta_0$ beschränkt. Es ist klar, dass unter denselben Voraussetzungen auch optimale Tests für $\tilde{H}_0 : \vartheta \ge \vartheta_0$ gegen $\tilde{H}_1 : \vartheta < \vartheta_0$ angegeben werden können, die von analoger Gestalt sind.

Beispiel 8.3.11 *Ausschussstücke in einem Fertigungslos*

In der Wirtschaft bezeichnet das Wort Los oder Fertigungslos eine Menge eines Erzeugnisses, die hintereinander ohne Unterbrechung durch andere Produkte erstellt wird. Bei einem Los vom Umfang $N = 200$ mögen M Ausschussstücke sein. Sei X die Zahl der schlechten Stücke, die in einer Stichprobe vom Umfang $n = 5$ vorgefunden werden. Dabei werde die Stichprobe ohne Zurücklegen gezogen. Dann ist X hypergeometrisch verteilt,

$$f(x; M) = P_M(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

Es soll für das Testproblem

$$H_0 : M \leq 10, \qquad H_1 : M > 10$$

ein GB-Test zum Niveau $\alpha = 0.01$ bestimmt werden.

Zunächst ist zu untersuchen, ob es eine Statistik $T(X)$ gibt, so dass die hypergeometrischen Verteilungen einen monotonen Likelihoodquotienten in $T(X)$ haben. Sei dazu $M_0 < M_1$. Dann ist

$$\begin{aligned}\frac{f(x;M_1)}{f(x;M_0)} &= \frac{\binom{M_1}{x}\binom{N-M_1}{n-x}\binom{N}{n}^{-1}}{\binom{M_0}{x}\binom{N-M_0}{n-x}\binom{N}{n}^{-1}} \\ &= \frac{M_1!(N-M_1)!(N-M_0-n+x)\cdot\ldots\cdot(N-M_1-n+x+1)}{M_0!(N-M_0)!(M_1-x)\cdot\ldots\cdot(M_0-x+1)}.\end{aligned}$$

Je größer x wird, desto größer wird $(N-M_0-n+x)\cdot\ldots\cdot(N-M_1-n+x+1)$ und desto kleiner wird $(M_1-x)\cdot\ldots\cdot(M_0-x+1)$. Daher hängt $f(x;M_1)/f(x;M_0)$ monoton von x ab. Der beste Test für das Problem hat also die Form

$$\varphi^*(x) = \begin{cases} 1 & x > k \\ 0 & x \leq k \end{cases}.$$

k ist dabei konkret für $M_0 = 10$ zu bestimmen aus

$$\mathrm{P}_{M_0}(X > k) \leq 0.01, \qquad \mathrm{P}_{M_0}(X > k-1) > 0.01.$$

Wegen

$$\mathrm{P}_{M_0}(X > 1) = 0.02084, \qquad \mathrm{P}_{M_0}(X > 2) = 0.00087$$

ist $k = 2$.

Um zu verdeutlichen, dass die Voraussetzung des monotonen Likelihoodquotienten in der Tat der Schlüssel für die Existenz gleichmäßig bester Tests für einseitige Probleme ist, wird das folgende Beispiel betrachtet. Es zeigt, dass Cauchy-Verteilungen keinen monotonen Likelihoodquotienten haben, und dass es auch zu gewissen Niveaus keine gleichmäßig besten Tests für einseitige Testprobleme gibt.

Beispiel 8.3.12 *einseitiges Testproblem bei Cauchy-Verteilung*

X habe eine Cauchy-Verteilung mit dem Parameter $\mu \in \mathbb{R}$,

$$f(x;\mu) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\mu)^2} \qquad -\infty < x < \infty.$$

Die Cauchy-Verteilungen haben keinen monotonen Likelihoodquotienten in X. Für den Quotienten $f(x;\mu_1)/f(x;\mu_0)$ gilt nämlich im Fall $\mu_0 < \mu_1$:

$$\frac{f(x;\mu_1)}{f(x;\mu_0)} = \frac{\frac{1}{\pi}\cdot\frac{1}{1+(x-\mu_1)^2}}{\frac{1}{\pi}\cdot\frac{1}{1+(x-\mu_0)^2}} = \frac{1+(x-\mu_0)^2}{1+(x-\mu_1)^2} \begin{cases} \to 1 & \text{für } x \to -\infty \\ < 1 & \text{für } x = \mu_0 \\ = 1 & \text{für } x = (\mu_0+\mu_1)/2 \\ > 1 & \text{für } x = \mu_1 \\ \to 1 & \text{für } x \to +\infty \end{cases}.$$

Also ist der Likelihoodquotient nicht monoton in x. Es gibt auch keinen gleichmäßig besten Test für das einseitige Testproblem. Sei zur Vereinfachung

$$H_0 : \mu = 0, \qquad H_1 : \mu \geq 0.$$

Wir betrachten wieder ein vereinfachtes Testproblem:

$$H_0' : \mu = 0, \qquad H_1' : \mu = 1.$$

Nach dem Fundamentallemma hat der beste Test hier die Gestalt :

$$\varphi^*(\boldsymbol{x}) = \begin{cases} 1 & \frac{f(x;\mu_1)}{f(x;\mu_0)} > k \\ 0 & \frac{f(x;\mu_1)}{f(x;\mu_0)} \leq k \end{cases}.$$

Der Likelihoodquotient ist in der folgenden Abbildung dargestellt. Schon die Grafik macht deutlich, dass der Ablehnbereich ein endliches, abgeschlossenes Intervall ist, wenn k größer als 1.5 gewählt wird.

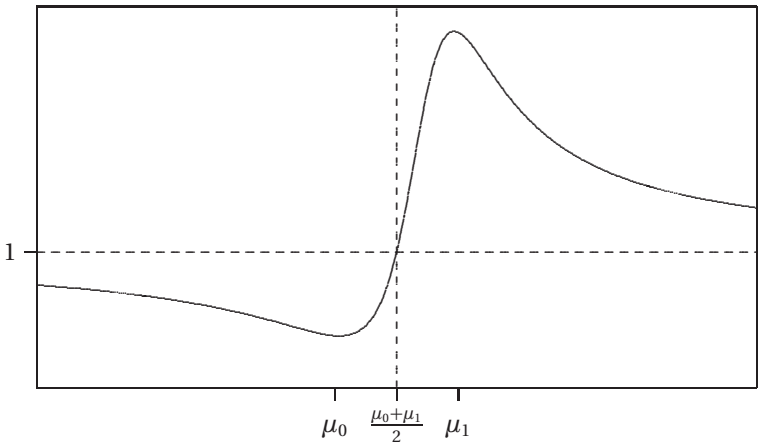


Abb. 8.8: *Likelihoodquotient bei Cauchy-Verteilung*

Speziell erhalten wir für $k=2$:

$$f(x;\mu_1) > 2\cdot f(x;\mu_0) \Longleftrightarrow 1+x^2 > 2\cdot\left(1+(x-1)^2\right) \Longleftrightarrow x^2-4x+3<0 \Longleftrightarrow (x-2)^2<1$$
$$\Longleftrightarrow 1<x<3.$$

Es folgt

$$P_{\mu_0}(1<X<3) = \frac{1}{\pi}\cdot\arctan(3) - \frac{1}{\pi}\cdot\arctan(1) = 0.14758.$$

Geben wir also dieses Niveau vor, so erhalten wir einen Test mit dem Ablehnbereich $(1;3)$. Da mit μ die Verteilung verschoben wird, ist klar, dass die Wahrscheinlichkeit für dieses Intervall gegen Null geht:

$$\lim_{\mu\to\infty} \mathrm{P}_\mu(1<X<3) = 0.$$

Der beste Test für $H_0: \mu = 0$ gegen $H_1: \mu = 1$ ist also nicht gleichmäßig bester Test für $H_0': \mu \le 0$ gegen $H_1': \mu > 0$; dies zeigt sich schon daran, dass er für das allgemeinere Problem verfälscht ist.

Der GB-Test für die einseitigen Hypothesen wurde über das Neyman-Pearson Lemma gewonnen, also aus einem Likelihood-Quotienten-Test. Den Zusammenhang zwischen allgemeinen Likelihood-Quotienten-Tests und gleichmäßig besten Tests für einseitige Fragestellungen beleuchtet der folgende Satz.

Satz 8.3.13 *Optimalität des Likelihood-Quotienten-Tests*

Sei $\boldsymbol{X} = (X_1,\ldots,X_n)$ eine Zufallsstichprobe aus einer Verteilung mit der Dichte $f(x;\vartheta)$ und sei ϑ ein reellwertiger Parameter. $T(\boldsymbol{X})$ sei eine suffiziente Statistik. Weiter gelte:

- $f(\boldsymbol{x};\vartheta) = \prod_{i=1}^n f(x_i;\vartheta)$ habe einen monotonen Likelihood-Quotienten in $T(\boldsymbol{X})$.
- $f(\boldsymbol{x};\vartheta_1)/f(\boldsymbol{x};\vartheta_0) = f(t;\vartheta_1)/f(t;\vartheta_0) =: L(t;\vartheta)$ sei eine nicht-fallende Funktion von ϑ für $\vartheta < \hat{\vartheta}$. Dabei ist $\hat{\vartheta}$ der Maximum-Likelihood-Schätzer.
- $\hat{\vartheta} = \hat{\vartheta}(t)$ ist eine nicht-fallende Funktion von t.

Dann ist der Likelihood-Quotienten-Test für $H_0: \vartheta \le \vartheta_0$ gegen $H_1: \vartheta > \vartheta_0$ der gleichmäßig beste Test zu seinem Niveau für dieses Testproblem.

Beweis: Siehe Birkes (1990)

Beispiel 8.3.14 *GB-Tests bei Exponentialfamilien*

Exponentialfamilien erfüllen die Voraussetzungen des Satzes, wenn sie die im Beispiel 8.3.9 genannten Eigenschaften aufweisen.

Beispiel 8.3.15 *GB-Tests bei Gleichverteilung*

Die Gleichverteilung über dem Intervall $[0;\vartheta]$ hat die Dichte $f(x;\vartheta) = \vartheta^{-1}\cdot I_{[0;\vartheta]}(x)$. Man überlegt leicht, dass die Voraussetzungen des Satzes erfüllt sind.

8.3.3 Einfache Hypothesen gegen zweiseitige Alternativen

Es werden nun Testprobleme für eindimensionale Parameter $\vartheta \in \Theta$ untersucht, die folgende Struktur haben:

$$H_0 : \vartheta = \vartheta_0, \qquad H_1 : \vartheta \neq \vartheta_0.$$

Zunächst wird mittels eines Beispiels gezeigt, dass i. Allg. für solche Probleme kein gleichmäßig bester Test zu einem vorgegebenen Niveau existiert.

Beispiel 8.3.16 *Normalverteilung mit bekannter Varianz*

X sei normalverteilt mit bekannter Varianz σ^2. Als Testproblem liege vor:

$$H_0 : \mu = 0, \qquad H_1 : \mu \neq 0.$$

Dieses soll bei vorgegebenem Niveau α auf Basis einer Stichprobe vom Umfang n entschieden werden. Für

$$H_0 : \mu = 0, \qquad H_1' : \mu > 0$$

gibt es einen gleichmäßig besten Test $\varphi_1^*(\boldsymbol{x})$ zum Niveau α.

Da jeder Test zum Niveau α für H_0 gegen H_1' auch ein Test zum Niveau α für H_0 gegen H_1 ist und umgekehrt, kann kein Test φ für das ursprüngliche Problem bei $\mu > 0$ eine größere Güte haben als $\varphi_1^*(\boldsymbol{x})$. Allerdings ist φ_1^* dem analog gebildeten, für das Problem $H_0 : \mu = 0$ gegen $H_1'' : \mu < 0$ optimalen Test $\varphi_2^*(\boldsymbol{x})$ für $\mu < 0$ unterlegen. Offensichtlich gibt es keinen gleichmäßig besten Test zum Niveau α.

Wie in dem Beispiel existiert für viele Testprobleme kein gleichmäßig bester Test zu einem vorgegebenem Niveau. Um dann noch einen möglichst guten auszuzeichnen, wird die Menge Φ aller zur Konkurrenz zugelassenen Tests zum Niveau α geeignet eingeschränkt. Naheliegend ist es, bei zweiseitigen Testproblemen nur unverfälschte Tests zuzulassen, da gleichmäßig beste Tests ja stets unverfälscht sind. Die auf unverfälschte Tests eingeschränkte Menge von Tests sei Φ^e. Unter den Tests aus Φ^e suchen wir den optimalen, d.h. den gleichmäßig besten unter allen unverfälschten Tests, kurz den *gleichmäßig besten unverfälschten Test* oder *GBU-Test*. Für andere Testprobleme sind andere Restriktionen zu wählen, um in den eingeschränkten Mengen von Tests sinnvoll nach einem besten suchen zu können.

Mit der Beschränkung auf unverfälschte Tests werden auch die Tests, die für die einseitigen Probleme optimal sind, ausgeschieden. Dies ist sinnvoll, da sie nur formal Tests für die zweiseitigen Probleme sind. Konstruiert wurden sie jedoch nur zur Aufdeckung von Abweichungen nach einer Richtung.

Um einen gleichmäßig besten unverfälschten Test zum Niveau α für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$ zu konstruieren, können aber die optimalen einseitigen Tests ausgenutzt werden. Die Verteilungen von $\boldsymbol{X} = (X_1, \ldots, X_n)$ mögen wieder einen monotonen Likelihoodquotienten in $T(\boldsymbol{X})$ haben. $\varphi_1(\boldsymbol{x})$, $\varphi_2(\boldsymbol{x})$ seien die gleichmäßig besten Tests zu den Niveaus α_1 bzw. α_2 für $H_0 : \vartheta = \vartheta_0$ gegen $H_1' : \vartheta < \vartheta_0$ bzw. $H_1'' : \vartheta > \vartheta_0$. Einen Test, der Abweichungen nach beiden Seiten aufdeckt, erhalten wir offensichtlich, wenn die Ablehnbereiche von φ_1 und φ_2 zu einem vereinigt werden. Der daraus resultierende Test φ ist

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & T(\boldsymbol{x}) < k_1,\ T(\boldsymbol{x}) > k_2 \\ 0 & k_1 \leq T(\boldsymbol{x}) \leq k_2 \end{cases}.$$

Die Niveaus α_1 und α_2 und entsprechend die kritischen Werte k_1 und k_2 sind dabei so festzulegen, dass für $i = 1,2$ gilt:

$$\mathrm{E}_{\vartheta_0}\varphi_i(\boldsymbol{X}) \leq \alpha_i \text{ und } \mathrm{E}_{\vartheta}\varphi_i(\boldsymbol{X}) \geq \alpha_i \text{ für alle } \vartheta \in \Theta \backslash \Theta_0.$$

Die Frage ist nun, ob es eine derartige Wahl von φ_1 und φ_2 überhaupt gibt und wenn, ob der Test φ dann ein gleichmäßig bester unverfälschter Test ist.

Das folgende Beispiel zeigt, dass die Voraussetzung des monotonen Likelihoodquotienten zu schwach ist, um die Existenz eines optimalen Tests zu sichern.

Beispiel 8.3.17 *eine Verteilung mit drei möglichen Parameterwerten*

X sei eine stetige Zufallsvariable, deren mögliche Verteilungen durch die Dichten $f(x;\vartheta)$, $\vartheta \in \{-1,0,1\}$ festgelegt sind:

$$\begin{aligned} f(x;-1) &= -2x+2 && 0<x<1, \\ f(x;0) &= 1 && 0<x<1, \\ f(x;1) &= \begin{cases} x & 0<x\leq 0.5, \\ 5x-2 & 0.5<x<1. \end{cases} \end{aligned}$$

Es ist leicht zu verifizieren, dass die Verteilungen monotonen Likelihoodquotienten in X haben.

Gegeben sei nun das zweiseitige Testproblem $H_0 : \vartheta = 0, \quad H_1 : \vartheta \neq 0$, für das ein Test zum Niveau $\alpha = 0.1$ auf der Basis einer Beobachtung bestimmt werden soll.

Die optimalen Tests für die beiden Testproblemen $H_0' : \vartheta = 0, \quad H_1' : \vartheta = -1$ und $H_0'' : \vartheta = 0, \quad H_1'' : \vartheta = 1$, seien φ_1 und φ_2. Sie sind gegeben durch

$$\varphi_1(x) = \begin{cases} 1 & \text{für } x < \alpha \\ 0 & \text{für } x \geq \alpha \end{cases} \quad \text{und} \quad \varphi_2(x) = \begin{cases} 1 & \text{für } x > 1-\alpha \\ 0 & \text{für } x \leq 1-\alpha \end{cases}.$$

Der zusammengesetzte Test für das Ausgangsproblem hat dann die Form

$$\varphi(x) = \begin{cases} 1 & \text{für } x < k_1,\ x > k_2 \\ 0 & \text{für } k_1 \leq x \leq k_2 \end{cases}.$$

k_1 und k_2 sind nun so festzulegen, dass das Niveau eingehalten wird und $\varphi(x)$ nicht nur unverfälscht ist, sondern auch möglichst große Güte in $\vartheta = 1$ bzw. $\vartheta = -1$ hat. Die Forderung des Einhaltens der vorgegebenen Irrtumswahrscheinlichkeit führt zu

$$\alpha = \mathrm{E}_0\varphi(X) = \mathrm{P}_0(X < k_1) + \mathrm{P}_0(X > k_2) = k_1 + (1-k_2)$$

bzw. $k_2 = k_1 + 1 - \alpha$.

Die Werte der Gütefunktion an den Stellen $\vartheta = \pm 1$ sind, wenn $0 < \alpha < 1/8$ vorausgesetzt wird:

$$\mathrm{E}_{-1}\varphi(X) = 2(1-\alpha)k_1 + \alpha^2,$$
$$\mathrm{E}_1\varphi(X) = -2k_1^2 + (5\alpha - 3)k_1 + 3\alpha - 2.5\alpha^2.$$

Der Test soll noch unverfälscht sein, d.h. es muss gelten: $E_{-1}\varphi(X) \geq \alpha, \quad E_1\varphi(X) \geq \alpha$.

Die erste Ungleichung ergibt sofort: $k_1 \geq \alpha/2$. $\mathrm{E}_1\varphi(X) \geq \alpha$ führt zu

$$\frac{5\alpha - 3}{4} - \frac{\sqrt{5\alpha^2 - 14\alpha + 9}}{4} \leq k_1 \leq \frac{5\alpha - 3}{4} + \frac{\sqrt{5\alpha^2 - 14\alpha + 9}}{4}.$$

Allerdings folgt aus $\alpha < 1/8$ bereits $k_1 \leq 1/2$. Damit ist $\varphi(x)$ nicht eindeutig bestimmt.

Für das Testproblem $H_0' : \vartheta = 0, \quad H_1' : \vartheta = 1$ erhalten wir, dass k_1 möglichst klein gewählt werden muss, um eine große Güte in $\vartheta = 1$ zu erreichen. Für das Testproblem $H_0' : \vartheta = 0, \quad H_1' : \vartheta = -1$ muss k_1 dagegen möglichst groß sein. Das führt bei $\alpha = 0.1$ zu folgenden konkreten Tests mit sich 'schneidenden' Gütefunktionen:

$$\varphi'(x) = \begin{cases} 1 & x < 0.05,\ x > 0.95 \\ 0 & 0.05 \leq x \leq 0.95 \end{cases}$$
$$\varphi''(x) = \begin{cases} 1 & x < 0.066,\ x > 0.934 \\ 0 & 0.066 \leq x \leq 0.934\,. \end{cases}$$

Der heuristische Ansatz führt also nicht allgemein für Verteilungen mit monotonem Likelihoodquotienten zu einem gleichmäßig besten unverfälschten Test. Im obigem Beispiel liegt das daran, dass $f(x;0)/f(x;-1)$ anders von x abhängt als $f(x;1)/f(x;0)$.

Um mit dem Ansatz zum Ziel zu kommen, muss eine stärkere Voraussetzung gemacht werden, die eine geeignete 'Gleichartigkeit' der Likelihoodquotienten sichert. Diese ist bei einparametrigen Exponentialfamilien gegeben. Darauf beschränken sich die weiteren Betrachtungen. Zudem unterstellen wir, dass die Exponentialfamilie in Abhängigkeit von ihrem natürlichen Parameter betrachtet wird, also in kanonischer Form vorliegt:

$$f(x_1,\ldots,x_n;\eta) = \exp\left[\eta \sum_{i=1}^{n} T(x_i) + d(\eta) + \sum_{i=1}^{n} S(x_i)\right] \times I_A(x_1,\ldots,x_n).$$

Bei Exponentialfamilien ist die Gütefunktion jedes Tests differenzierbar. Die Forderung der Unverfälschtheit des zweiseitigen Tests, $\mathrm{E}_\eta\varphi(\boldsymbol{X}) \geq \alpha$ für $\eta \neq \eta_0$, ergibt mit der Bedingung $\mathrm{E}_{\eta_0}\varphi(\boldsymbol{X}) = \alpha$, d.h. dass das vorgegebene Niveau ausgeschöpft wird:

$$\frac{\partial}{\partial \eta}\mathrm{E}_\eta\varphi(\boldsymbol{X})\bigg|_{\eta=\eta_0} = 0. \tag{8.36}$$

Anstelle des GBU-Tests wird der beste Test unter allen Tests zum Niveau α bestimmt, deren Ableitung im Punkt $\eta = \eta_0$ gleich Null ist. Die Menge dieser Tests umfasst nach dem eben Gesagten die unverfälschten.

Zur Vorbereitung benötigen wir noch ein technisches Lemma.

Lemma 8.3.18

Gegeben sei eine einparametrige Exponentialfamile mit der Dichte

$$f(\boldsymbol{x};\eta) = \exp[\eta U(\boldsymbol{x}) + d(\eta) + V(\boldsymbol{x})] \times I_A(x_1,\ldots,x_n).$$

Mit $U(\boldsymbol{x}) = \sum_{i=1}^n T(x_i)$ und $V(\boldsymbol{x}) = \sum_{i=1}^n S(x_i)$. Dann gilt für jeden Test $\varphi(\boldsymbol{X})$:

$$\frac{\partial}{\partial\eta}\mathrm{E}_\eta\varphi(\boldsymbol{X})\bigg|_{\eta=\eta_0} = \mathrm{E}_{\eta_0}[\varphi(\boldsymbol{X})\cdot U(\boldsymbol{X})] - \mathrm{E}_{\eta_0}[U(\boldsymbol{X})]\mathrm{E}_{\eta_0}\varphi(\boldsymbol{X}). \tag{8.37}$$

Beweis: Zur Vereinfachung der Schreibweise wird der Beweis für den stetigen Fall ausgeführt:

$$\begin{aligned}
\frac{\partial}{\partial\eta}\mathrm{E}_\eta\varphi(\boldsymbol{X}) &= \frac{\partial}{\partial\eta}\int_A \varphi(\boldsymbol{x})\cdot f(\boldsymbol{x};\eta)\,\mathrm{d}\boldsymbol{x} = \int_A \varphi(\boldsymbol{x})\cdot\frac{\partial}{\partial\eta}f(\boldsymbol{x};\eta)\,\mathrm{d}\boldsymbol{x}\\
&= \int_A \varphi(\boldsymbol{x})\cdot\frac{\partial}{\partial\eta}\exp[\eta U(\boldsymbol{x}) + d(\eta) + V(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x}\\
&= \int_A \varphi(\boldsymbol{x})\cdot(U(\boldsymbol{x}) + d'(\eta))\cdot\exp[\eta U(\boldsymbol{x}) + d(\eta) + V(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x}\\
&= \int_A \varphi(\boldsymbol{x})\cdot U(\boldsymbol{x})\cdot f(\boldsymbol{x};\eta)\,\mathrm{d}\boldsymbol{x} + \int_A \varphi(\boldsymbol{x})\cdot d'(\eta)\cdot f(\boldsymbol{x};\eta)\,\mathrm{d}\boldsymbol{x}\\
&= \mathrm{E}_\eta[\varphi(\boldsymbol{X})\cdot U(\boldsymbol{X})] - \mathrm{E}_\eta U(\boldsymbol{X})\cdot\mathrm{E}_\eta\varphi(\boldsymbol{X})
\end{aligned}$$

Dabei wurde die in Satz 3.1.19 bewiesene Relation $d'(\eta) = -\mathrm{E}_\eta T(\boldsymbol{X})$ ausgenutzt.

Nun kann das Resultat über die Existenz von gleichmäßig besten unverfälschten Tests bei zweiseitigen Testproblemen formuliert werden. Dabei berücksichtigen wir wieder, dass es bei diskreten Verteilungen zu Problemen mit dem Ausschöpfen des Niveaus kommen kann.

Satz 8.3.19 *optimale Tests bei einparametrigen Exponentialfamilien*

$\boldsymbol{X} = (X_1,\ldots,X_n)$ sei eine n-dimensionale Zufallsvariable mit einer Verteilung, die zu einer einparametrigen Exponentialfamilie gehöre. Die gemeinsame Dichte sei

$$f(\boldsymbol{x};\eta) = \exp[\eta T(\boldsymbol{x}) + d(\eta) + S(\boldsymbol{x})]\cdot I_A(\boldsymbol{x}) \qquad \eta\in\Theta.$$

Für das Testproblem

$$H_0:\ \eta = \eta_0, \qquad H_1:\ \eta\neq\eta_0$$

ist ein Test von der Form

$$\varphi^*(\boldsymbol{x}) = \begin{cases} 1 & T(\boldsymbol{x}) \le k_1,\ T(\boldsymbol{x}) > k_2\\ 0 & k_1 \le T(\boldsymbol{x}) \le k_2 \end{cases}$$

gleichmäßig bester unter allen Tests zum Niveau $\alpha = \mathrm{E}_{\eta_0}(\varphi^*(\boldsymbol{X}))$, für die gilt

$$\frac{\partial}{\partial\eta}\mathrm{E}_\eta\varphi(\boldsymbol{X})\bigg|_{\eta=\eta_0} = \mathrm{E}_{\eta_0}[\varphi(\boldsymbol{X})\cdot T(\boldsymbol{X})] - \alpha\cdot\mathrm{E}_{\eta_0}T(\boldsymbol{X}) = 0.$$

Beweisskizze: Die wesentlichen Punkte des Beweises sind, für eine eindimensionale Zufallsvariable X formuliert, folgende.

Zunächst ist nachzuweisen, dass es zwei Konstanten k_1 und k_2 gibt, so dass für den zugehörigen Test φ^* gilt:

$$\begin{aligned}&\mathrm{E}_{\eta_0}\varphi^*(X) = \mathrm{P}_{\eta_0}(T(X) < k_1) + \mathrm{P}_{\eta_0}(T(X) > k_2) = \alpha,\\ &\mathrm{E}_{\eta_0}\varphi^*(X)\cdot T(X) - \alpha\cdot \mathrm{E}_{\eta_0} T(X) = 0.\end{aligned}$$

Dann ist ihre Eindeutigkeit zu zeigen. Dieser Teil wird nicht weiter ausgeführt. Seien k_1 und k_2 als entsprechend bestimmt vorausgesetzt.

Sei ψ ein anderer Test zum Niveau α mit $\mathrm{E}_{\eta_0}\psi(X)\cdot T(X) = \alpha\cdot \mathrm{E}_{\eta_0} T(X)$. Dann ist für $\eta_1 \in \Theta$ zu zeigen:

$$\mathrm{E}_{\eta_1}\varphi^*(X) \geq \mathrm{E}_{\eta_1}\psi(X).$$

O.B.d.A. sei $\eta_1 > \eta_0$. Analog zum Beweis des Fundamentallemmas 8.3.4 wird der Stichprobenraum in die Teilmengen zerlegt:

$$A = \{x | T(x) < k_1\}, \quad B = \{x | T(x) > k_2\}, \quad C = \{x | k_1 \leq T(x) \leq k_2\}.$$

Mit der Konvexität des Likelihoodquotienten

$$\frac{f(x;\eta_1)}{f(x;\eta_0)} = \exp[(\eta_1 - \eta_0)\cdot T(x)]$$

folgt die Existenz zweier Konstanten $a > 0, b > 0$, so dass für die Gerade $a + b\,T(x)$ gilt:

$$\begin{aligned}&\frac{f(x;\eta_1)}{f(x;\eta_0)} > (a + b\cdot T(x)) \qquad \text{für alle } x \in A \cup B,\\ &\frac{f(x;\eta_1)}{f(x;\eta_0)} \leq (a + b\cdot T(x)) \qquad \text{für alle } x \in C.\end{aligned}$$

Wir erhalten weiter:

$$\begin{aligned}\mathrm{E}_{\eta_1}\varphi^*(X) - \mathrm{E}_{\eta_1}\psi(X) &= \int_{-\infty}^{\infty} (\varphi^*(x) - \psi(x)) f(x;\eta_1)\,\mathrm{d}x\\ &= \int_{A\cup B} (\varphi^*(x) - \psi(x)) f(x;\eta_1)\,\mathrm{d}x + \int_C (\varphi^*(x) - \psi(x)) f(x;\eta_1)\,\mathrm{d}x.\end{aligned}$$

Wegen $\varphi^*(x) = 1$ für $x \in A \cup B$ folgt mit $\varphi^*(x) - \psi(x) \geq 0$:

$$\int_{A\cup B} (\varphi^*(x) - \psi(x)) f(x;\eta_1)\,\mathrm{d}x \geq \int_{A\cup B} (\varphi^*(x) - \psi(x))(k_1' + k_2'\cdot T(x)) f(x;\eta_0)\,\mathrm{d}x.$$

Da $\varphi^*(x) - \psi(x) < 0$ für $x \in C$ gilt, ist auch hier

$$\int_C (\varphi^*(x) - \psi(x)) f(x;\eta_1)\,\mathrm{d}x \geq \int_C (\varphi^*(x) - \psi(x))(k_1' + k_2'\cdot T(x)) f(x;\eta_0)\,\mathrm{d}x.$$

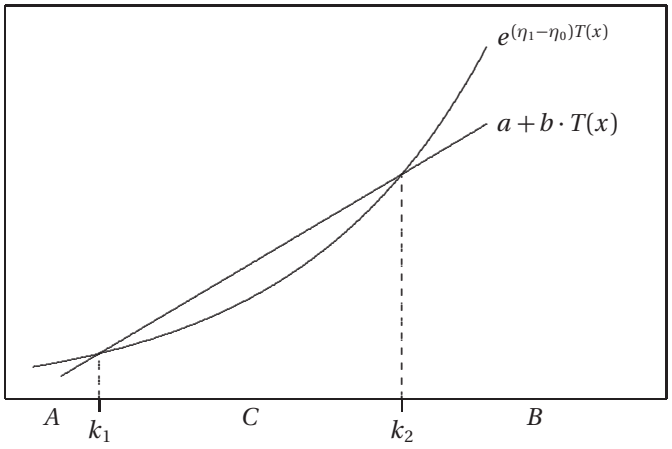


Abb. 8.9: *Illustration zur Konvexität des Likelihoodquotienten*

Zusammen ergibt das unter Verwendung des vorangegangenen Lemmas:

$$\begin{aligned}\mathrm{E}_{\eta_1}\varphi^*(X)-\mathrm{E}_{\eta_1}\psi(X) &\geq a\cdot\mathrm{E}_{\eta_0}[\varphi^*(X)-\psi(X)]+b\cdot\mathrm{E}_{\eta_0}[(\varphi^*(X)-\psi(X))\cdot T(X)]\\ &\geq a(\alpha-\alpha)+b(\alpha\cdot\mathrm{E}_{\eta_0}T(X)-\alpha\cdot\mathrm{E}_{\eta_0}T(X))=0.\end{aligned}$$

Beispiel 8.3.20 *zweiseitiger Test auf μ*

X sei eine normalverteilte Zufallsvariable mit bekannter Varianz $\sigma^2=1$. Auf Basis einer Stichprobe vom Umfang n soll bei vorgegebenem α zwischen den Hypothesen $H_0:\mu=0$ und $H_1:\mu\neq 0$ entschieden werden.

Die gemeinsame Verteilung der Stichprobenvariablen $X_1,\ldots,X_n$ ist

$$\begin{aligned}f(\boldsymbol{x};\mu) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^n(x_i-\mu)^2\right]\\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^n x_i^2+\frac{n}{2}\bar{x}^2-\frac{n}{2}(\bar{x}-\mu)^2\right].\end{aligned}$$

Sie gehört zu einer Exponentialfamilie mit dem Parameter μ und mit $T(\boldsymbol{x})=\bar{x}$.

Der gleichmäßig beste unverfälschte Test für H_0 gegen H_1 hat also die Gestalt:

$$\varphi(\boldsymbol{x})=\begin{cases}1 & \bar{x}<k_1,\ \bar{x}>k_2\\ 0 & k_1\leq\bar{x}\leq k_2\end{cases}.$$

k_1 und k_2 sind zu bestimmen aus den beiden Bedingungen

$$\mathrm{E}_{\mu_0}\varphi(\boldsymbol{X})=\mathrm{P}_{\mu_0}(\bar{X}<k_1)+\mathrm{P}_{\mu_0}(\bar{X}>k_2)=\alpha \tag{8.38a}$$

$$\mathrm{E}_{\mu_0}\varphi(\boldsymbol{X})\cdot T(\boldsymbol{X})-\alpha\cdot\mathrm{E}_{\mu_0}T(\boldsymbol{X})=0\,. \tag{8.38b}$$

Die Symmetrie der Normalverteilung legt die Wahl $k_1 = -k_2$ nahe. Da $T(\boldsymbol{X}) = \bar{X}$ wieder normalverteilt ist mit dem Erwartungswert μ und der Varianz $1/n$, ergibt dies: $k_2 = z_{1-\alpha/2}/\sqrt{n}$. Damit ist (8.38a) offensichtlich erfüllt. Für (8.38b) ergibt sich zusätzlich zu $\mathrm{E}_{\mu_0} T(\boldsymbol{X}) = 0$ mit der Symmetrie der Verteilung und des kritischen Bereiches $\mathrm{E}_{\mu_0} \varphi(\boldsymbol{X}) \cdot T(\boldsymbol{X}) = 0$.

Birkes (1990) untersuchte auch den Zusammenhang zwischen Likelihoodquotiententests und gleichmäßig besten unverfälschten Tests für das zweiseitige Problem. Für einparametrige Exponentialfamilien zeigte er, dass ein Likelihoodquotiententest genau dann gleichmäßig bester unverfälschter Test seines Umfanges ist, wenn er unverfälscht ist.

8.3.4 GBU-Tests in mehrparametrigen Exponentialfamilien

Einen wichtigen Teil der angewandten Probleme stellen die Testprobleme für einen eindimensionalen Parameter in mehrparametrigen Verteilungsfamilien dar. Die anderen Parameter sind dann *Nuisance-Parameter*, d. h. störend. Sind diese Familien nun Exponentialfamilien, so lassen sich auch für diese erweiterten Probleme GBU-Tests angeben. Wir skizzieren hier nur die wesentlichen Punkte der Theorie. Für Details sei auf Lehmann (1986) verwiesen.

$\boldsymbol{X} = (X_1, \ldots, X_n)$ sei eine Stichprobe, deren gemeinsame Verteilung eine zweiparametrige Exponentialfamilie bilde, also eine Dichte der Form

$$f(\boldsymbol{x}; \vartheta, \xi) = \exp[\vartheta U(\boldsymbol{x}) + \xi V(\boldsymbol{x}) + d(\vartheta, \xi) + S(\boldsymbol{x})] \cdot I_A(\boldsymbol{x})$$

habe. Von Interesse sei ein Testproblem für die erste Komponente des natürlichen Parameters (ϑ, ξ). Wir nehmen als Beispiel das Testproblem

$$H_0 : \vartheta \leq \vartheta_0, \qquad H_1 : \vartheta > \vartheta_0,$$

das ausführlicher geschrieben lautet, wenn Ξ der gesamte Bereich der zweiten Komponente des Parameters ist:

$$H_0 : \vartheta \leq \vartheta_0,\ \xi \in \Xi, \qquad H_1 : \vartheta > \vartheta_0,\ \xi \in \Xi.$$

Da (U, V) eine suffiziente Statistik ist, reicht es, das Testproblem bzgl. der Verteilung dieser Statistik zu betrachten. Diese gehört wieder zu einer Exponentialfamilie:

$$f(u, v; \vartheta, \xi) = \exp[\vartheta u + \xi v + d(\vartheta, \xi) + S(u, v)] \cdot I_A(u, v).$$

Die in Abschnitt 8.2.5 eingeführten bedingte Tests resultierten gerade aus solchen Situationen. Nach der dort geführten Diskussion liegt es nahe, V als bedingende Statistik zu wählen und zunächst bedingte Tests bei jeweils gegebenen $V = v$ zu betrachten.

Die bedingte Verteilung von U bei gegebenem $V = v$ bildet jeweils eine einparametrige Exponentialfamilie:

$$f(u|v; \vartheta) = \exp[\vartheta u + d_v(\vartheta) + S(u, v)] \times I_A(u, v).$$

Daher können die Ergebnisse der beiden vorangehenden Abschnitte für die bedingten Situationen ausgenutzt werden. Es existieren somit GB (randomisierte) Tests für die jeweiligen bedingten Situationen:

$$\varphi_v(u) = \begin{cases} 1 & u > c(v) \\ \gamma(v) & u = c(v) \\ 0 & u \leq c(v) \end{cases}.$$

Interpretieren wir nun die Schaar $\varphi_v(u)$ der bedingten Tests als einen Test, der von u und v abhängt, dann ist der so erhaltene Gesamttest ein GBU-Test für das Ausgangsproblem:

$$\varphi(u,v)=\begin{cases} 1 & u>c(v) \\ \gamma(v) & u=c(v) \\ 0 & u\le c(v) \end{cases}.$$

Die Randomisierung bei den bedingten Tests ist wesentlich, da bei diskreten Verteilungen die einzelnen bedingten Tests ohne Randomisierung ja alle ein unterschiedliches Niveau $\le\alpha$ haben können. Um sie zu einen GBU-Test zusammensetzen zu können, müssen sie alle den gleichen Umfang haben.

Analoges Vorgehen führt bei anderen Testprobleme für den Parameter ϑ zu entsprechenden GBU-Tests.

Satz 8.3.21 *GBU-Tests bei zweiparametrigen Exponentialfamilien*

$\boldsymbol{X}=(X_1,\ldots,X_n)$ sei eine Stichprobe aus einer Verteilung, die zu einer zweiparametrigen Exponentialfamilie gehöre, also eine Dichte der Form

$$f(\boldsymbol{x};\vartheta,\xi)=\exp[\vartheta U(\boldsymbol{x})+\xi V(\boldsymbol{x})+d(\vartheta,\xi)+S(\boldsymbol{x})]\cdot I_A(\boldsymbol{x}).$$

Für die Testprobleme

$$H_0^{(i)}:\ \vartheta\in\Theta_0^{(i)},\qquad H_1^{(i)}:\ \vartheta\notin\Theta_0^{(i)},$$

mit $\Theta_0^{(1)}=(-\infty,\vartheta_0]$, $\Theta_0^{(2)}=[\vartheta_0,\infty)$ und $\Theta_0^{(3)}=\{\vartheta_0\}$ gibt es GBU-Tests zum Niveau α. Dies ergeben sich aus den jeweiligen randomisierten GBU-Tests zum Niveau α für die bedingten Situationen, bei denen $V(\boldsymbol{x})$ festgehalten wird. Sie haben die Gestalt:

$$\varphi_1(u,v)=\begin{cases} 1 & u>k(v) \\ \gamma(v) & u=k(v) \\ 0 & u<k(v) \end{cases},$$

$$\varphi_2(u,v)=\begin{cases} 1 & u<k(v) \\ \gamma(v) & u=k(v) \\ 0 & u>k(v) \end{cases},$$

$$\varphi_3(u,v)=\begin{cases} 1 & u<k_1(v),\ u>k_2(v) \\ \gamma_1(v) & u=k_1(v) \\ \gamma_2(v) & u=k_2(v) \\ 0 & k_2(v)<u<k_2(v) \end{cases},$$

mit $u=U(\boldsymbol{x})$, $v=V(\boldsymbol{x})$.

Beweis: Siehe Lehmann (1986, S. 147).

Beispiel 8.3.22 *Vergleich zweier Wahrscheinlichkeiten*

X und Y seien unabhängige binomialverteilte Zufallsvariablen mit den Parametern n, p bzw. m, q. Als Testproblem liege vor:

$$H_0 : p \le q, \qquad H_1 : p > q.$$

Um dies in dem Rahmen dieses Abschnittes zu behandeln, identifizieren wir die gemeinsame Verteilung von X und Y als zugehörig zu einer Exponentialfamilie: Für $x = 0,1,\ldots,n,\ y = 0,1,\ldots,m$ gilt:

$$\begin{aligned} f(x,y;p,q) &= \binom{n}{x}\binom{m}{y} p^x (1-p)^{n-x} q^y (1-q)^{m-y} \\ &= \exp\left[\ln\left(\frac{p(1-q)}{(1-p)q}\right) x + \ln\left(\frac{q}{1-q}\right)(x+y) + n\cdot\ln(1-p)\right. \\ &\qquad \left. + m\cdot\ln(1-q) + \ln\binom{n}{x} + \ln\binom{m}{y}\right] \cdot I_A(x,y) \\ &= \exp\left[\ln\left(\frac{p(1-q)}{(1-p)q}\right)\cdot x - \ln\left(\frac{q}{1-q}\right)\cdot(x+y)\right. \\ &\qquad \left. + d\left(\frac{p(1-q)}{(1-p)q}, \frac{q}{1-q}\right) + S(x,y)\right] \cdot I_A(x,y) \\ &= \exp[\vartheta x - \xi(x+y) + d(\vartheta,\xi) + S(x,y)] \cdot I_A(x,y); \end{aligned}$$

dabei sind $\vartheta = \ln(p(1-q)/(1-p)q),\ \xi = \ln(q/1-q)$ und $A = \{(x,y)|x = 0,1,\ldots,n,\ y = 0,1,\ldots,m\}$.

Unser Testproblem lässt sich äquivalent mit dem Parameter ϑ formulieren:

$$H_0 : \vartheta \le 0, \qquad H_1 : \vartheta > 0.$$

Wir können somit den letzten Satz ausnutzen und den GBU-Test über die bedingten Situationen bestimmen. Mit den Statistiken $U = X$ und $T = X + Y$ haben wir die bedingte Dichte von X bei gegebenem $T = t$ zu bestimmen.

Die Verteilung von T ist, wie leicht nachzurechnen ist:

$$\begin{aligned} f(t) &= \sum_{j=0}^{t} \binom{n}{j}\binom{m}{t-j} p^j (1-p)^{n-j} q^{t-j} (1-q)^{m-(t-j)} \\ &= \left(\frac{q}{1-q}\right)^t \sum_{j=0}^{t} \binom{n}{j}\binom{m}{t-j} \left(\frac{p(1-q)}{(1-p)q}\right)^j (1-p)^n (1-q)^m. \end{aligned}$$

Für die bedingte Verteilung von X bei gegebenem $T=t$ folgt:

$$f(x|t;p,q) = \frac{\binom{n}{x}\binom{m}{t-x}\left(\frac{p(1-q)}{(1-p)q}\right)^x (1-p)^n(1-q)^m}{\sum_{j=0}^{t}\binom{n}{j}\binom{m}{t-j}\left(\frac{p(1-q)}{(1-p)q}\right)^j (1-p)^n(1-q)^m}$$
$$= \frac{\binom{n}{x}\binom{m}{t-x}\left(\frac{p(1-q)}{(1-p)q}\right)^x}{\sum_{j=0}^{t}\binom{n}{j}\binom{m}{t-j}\left(\frac{p(1-q)}{(1-p)q}\right)^j}.$$

Sie hängt also nur von $\vartheta = p(1-q)/q(1-p)$ ab. Zudem bilden diese Verteilungen bei festem t eine Exponentialfamilie:

$$f(x|t) = \exp[\vartheta \cdot x + d(\vartheta) + S(x)] \cdot I_A(x)$$

mit $d(\vartheta) = -\ln\left(\sum_{j=0}^{t}\binom{n}{j}\binom{m}{t-j}\left(\frac{p(1-q)}{(1-p)q}\right)^j\right)$, $S(x) = \ln\left(\binom{n}{x}\binom{m}{t-x}\right)$ und $A = \{0,1,\ldots,t\}$.

Der bedingte Test ist also von der Gestalt:

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

$c(t)$ und $\gamma(t)$ sind dabei so festzulegen, dass für $\vartheta = 0$ gilt:

$$P(X > c(t)|T=t) + \gamma(t) \cdot P(X = c(t)|T=t) = \alpha.$$

$\vartheta = 0$ ist ja nun äquivalent mit $p = q$. Dann ist

$$f(x|t;p,q) = \frac{\binom{n}{x}\binom{m}{t-x}}{\sum_{j=0}^{t}\binom{n}{j}\binom{m}{t-j}} = \frac{\binom{n}{x}\binom{m}{t-x}}{\binom{n+m}{t}}.$$

Die bedingte Verteilung von X ist für $\vartheta = 0$ also eine hypergeometrische Verteilung. Das macht die Bestimmung von $c(t)$ und $\gamma(t)$ einfach.

In den Anwendungen wird wie erwähnt nicht gerne auf die Randomisierung zurückgegriffen. Dann verliert der Test seine Optimalitätseigenschaft. Dennoch ist der nichtrandomisierte Test bei kleinen Stichprobenumfängen der am häufigsten durchgeführte für die vorliegende Situation. Je größer die Stichprobenumfänge sind, desto geringer ist natürlich die Rolle, die die Randomisierung spielt. Bei großen Stichproben kann er schließlich über eine entsprechende Normalapproximation als approximativer Test durchgeführt werden.

In manchen Situationen ist die direkte Angabe des GBU-Tests über die entsprechenden bedingten Tests recht kompliziert und unpraktisch. Dann ist es bisweilen möglich, den Test in einer unbedingten Form in Abhängigkeit von einer geeigneten Teststatistik anzugeben. Die Basis dafür bildet der folgende Satz.

Satz 8.3.23 *GBU-Test bei einer Exponentialfamilie*

Die Verteilung von $\boldsymbol{X}$ sei gegeben durch die Dichte $f(\boldsymbol{x};\vartheta,\xi) = \exp[\vartheta U(\boldsymbol{x}) + \xi V(\boldsymbol{x}) + d(\vartheta,\xi) + S(\boldsymbol{x})] \cdot I_A(\boldsymbol{x})$. Die Statistik $T = h(U,V)$ sei für $\vartheta = \vartheta_0$ unabhängig von V. Dann gilt:

(i) Ist h monoton wachsend in u für jedes feste v, so ist der Test

$$\varphi_1(t) = \begin{cases} 1 & \text{für } t > k_0 \\ \gamma_0 & \text{für } t = k_0 \\ 0 & \text{für } t < k_0 \end{cases},$$

für geeignet bestimmte Größen k_0 und γ_0 der GBU-Test für das Testproblem $H_0^1 : \vartheta \le \vartheta_0, \quad H_1^1 : \vartheta_0 > \vartheta_0$.

(ii) Der GBU-Test für $H_0^2 : \vartheta \ge \vartheta_0, \quad H_1^2 : \vartheta < \vartheta_0$ ergibt sich analog zu (i) unter der gleichen Voraussetzung an h.

(iii) Falls $t = h(u,v) = a(v)u + b(v)$ gilt, ist

$$\varphi_3(t) = \begin{cases} 1 & \text{für } t < k_1 \text{ oder } t > k_2 \\ \gamma_1 & \text{für } t = k_1 \\ \gamma_2 & \text{für } t = k_2 \\ 0 & \text{für } k_1 < t < k_2 \end{cases}$$

der GBU-Test für $H_0^3 : \vartheta = \vartheta_0, \; H_1^3 : \vartheta \ne \vartheta_0$. k_1, k_2, γ_1 und γ_2 sind dabei wieder geeignet zu bestimmen.

Beweis: Siehe Lehmann (1986, S. 190).

Beispiel 8.3.24 *Test auf μ bei unbekanntem σ^2*

In dem Beispiel 8.3.16 haben wir Testprobleme für den Parameter μ einer Normalverteilung bei bekanntem σ^2 betrachtet. Realistischer ist es i.d.R. von einem unbekannten σ^2 auszugehen. Dies wollen wir nun tun. Sei also X normalverteilt mit dem Parameter $\vartheta = (\mu,\sigma^2)$. Als Testproblem liege vor:

$$H_0 : \mu \le \mu_0,\ \sigma^2 > 0, \qquad H_1 : \mu > \mu_0,\ \sigma^2 > 0. \tag{8.39}$$

Der Übergang zu $\boldsymbol{X} - \mu_0 = (X_1 - \mu_0, \ldots, X_n - \mu_0)$ zeigt, dass wir o.B.d.A. $\mu_0 = 0$ annehmen können.

Um den Satz 8.3.23 anwenden zu können, schreiben wir die Dichte der Zufallsstichprobe $\boldsymbol{X} = (X_1, \ldots, X_n)$ in der Form

$$\begin{aligned} f(\boldsymbol{x};\mu,\sigma^2) &= \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right] \\ &= \exp\left[\frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - \frac{n\mu}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right]. \end{aligned}$$

Mit $\vartheta = \mu/\sigma^2$, $\xi = -1/(2\sigma^2)$, $U = \sum X_i$ und $V = \sum X_i^2$ haben wir die Ausgangssituation des Satzes vorzuliegen. Das Testproblem 8.39 lautet nun äquivalenter Weise:

$$H_0 : \vartheta \leq 0, \qquad H_1 : \vartheta > 0.$$

Die Statistik

$$T = h(U,V) = \frac{U/\sqrt{n}}{\sqrt{(V - U^2/n)/(n-1)}} = \frac{\sqrt{n}\bar{X}}{\sqrt{(\sum X_i^2 - n\bar{X}^2)/(n-1)}}$$

ist für $\mu = 0$ bzw. $\vartheta = 0$ unabhängig von $V = \sum X_i^2$ und für jedes feste v ist h monoton wachsend in u. Daher ist der GBU-Test für das einseitige Problem der klassische t-Test:

$$\varphi_1(t) = \begin{cases} 1 & \text{für } t > t_{n-1;\alpha} \\ 0 & \text{für } t \leq t_{n-1;\alpha} \end{cases}.$$

Für das zweiseitige Problem betrachten wir die Statistik $S = U/n\sqrt{V}$. Diese ist für $\mu = 0$ ebenfalls unabhängig von V. Zusätzlich ist S linear in U. Wegen der offensichtlichen Symmetrie der Verteilung von S in diesem Fall hat der Test die Form

$$\varphi_3(t) = \begin{cases} 1 & \text{für } |s| > c \\ 0 & \text{für } |s| \leq c \end{cases}.$$

Weil schließlich für

$$T = \frac{\sqrt{(n-1)n}\cdot S}{\sqrt{1 - nS^2}} = \frac{\sqrt{n}\bar{X}}{\sqrt{(\sum X_i^2 - n\bar{X}^2)/(n-1)}}$$

gilt: $|T|$ ist eine monoton wachsende Funktion von $|S|$, erhalten wir auch für das zweiseitige Testproblem den t-Test.

8.4 Weitere Eigenschaften von Tests

Die Ausführungen in Abschnitt 8.3 zeigen, dass die Existenz gleichmäßig bester Tests nur in ausgezeichneten Situationen gesichert ist. Dazu sind teilweise die zur Konkurrenz zugelassenen Tests einzuschränken. Zum anderen können wir uns auf Tests beschränken, die nur über eine suffiziente Statistik von den Beobachtungen abhängt, vgl. Satz 8.3.3. Neben dieser, alle Information in der Stichprobe erhaltenden Reduktion durch Suffizienz spielt die *Reduktion durch Invarianz* eine große Rolle bei der Suche nach besten Tests. Diese Form der Reduktion bringt tatsächlich einen Informationsverlust mit sich. Jedoch ist für manche Testprobleme ein angemessener Informationsverlust nicht problematisch.

Um anhand einer konkreten Situation zu diskutieren: Die relevante Information bei einem Lageproblem $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$ wird dann erhalten bleiben, wenn die Stichprobe — und die zugehörigen Parameter μ — linear transformiert werden. Die Transformation $Y = a + bX$ führt zu dem gleichwertigen Testproblem $H_0 : a + b\mu \leq a + b\mu$, $H_1 : a + b\mu > a + b\mu$,

und es ist offensichtlich egal, welche Form des Testproblems man betrachtet. Das bedeutet, dass man sich bei der Suche nach besten Tests auf die Klasse von Tests zurückziehen kann, die von äquivarianten Stichprobenfunktionen ausgehen, von Stichprobenfunktionen also, für die gilt $T(a+bX) = a+bT(X)$.

Genaueres zu der Reduktion durch Invarianz findet sich bei Witting (1985).

Lässt sich kein bester Test zum Niveau α für das u.U. eingeschränkte Testproblem finden, so ist es plausibel, wenigstens die maximale Fehlerwahrscheinlichkeit für den Fehler 2. Art möglichst klein zu halten, sich also gegen den schlimmsten Fall so gut es geht abzusichern. Dies ist äquivalent dazu, $\inf_{\vartheta\in\Theta_1} \mathrm{E}_\vartheta \varphi(\boldsymbol{X})$ zu maximieren. Ein Test φ^*, der das leistet, also

$$\inf_{\vartheta\in\theta_1} \mathrm{E}_\vartheta \varphi^*(\boldsymbol{X}) = \sup_{\varphi\in\Phi} \inf_{\vartheta\in\Theta_1} \mathrm{E}_\vartheta \varphi(\boldsymbol{X}) \tag{8.40}$$

erfüllt, wobei zu Φ nur Tests zum Niveau α gehören, wird als *Maximin-Test* zum Niveau α bezeichnet.

Damit dieser Ansatz sinnvoll ist, müssen die Null- und Alternativhypothese geeignet weit auseinander liegen. In unserem Standardbeispiel des Tests auf den Erwartungswert einer Normalverteilung ist die Gütefunktion der Tests ja stetig, so dass bei der üblichen Formulierung des einseitigen Testproblems $H_0 : \mu \le \mu_0,\ H_1 : \mu > \mu_0$ für Tests mit $\mathrm{E}_{\mu_0}\varphi(\boldsymbol{X}) = \alpha$ gilt:

$$\inf_{\mu>\mu_0} \mathrm{E}\varphi_\mu(\boldsymbol{X}) = \alpha.$$

Damit wären alle diese Tests Maximin-Tests. Hier ist also ein Indifferenzbereich $\mu_0 < \mu < \mu_1$ zwischen der Nullhypothese $H_0 : \mu \le \mu_0$ und der Alternativhypothese $H_1 : \mu > \mu_1$ festzulegen.

Diese Form der Optimalität und geeignete Verallgemeinerungen davon werden vor allem in der auf A. Wald zurückgehenden *Entscheidungstheorie* betrachtet. Dort erlaubt man auch die Berücksichtigung von Vorinformationen bzgl. des Parameters, die die Form einer Wahrscheinlichkeitsverteilung haben. Das führt zu Bayes'schen Verfahren; siehe dazu Wald (1950) und Berger (1980).

In Fällen, wo kein optimaler Test existiert, bleibt nur, Tests paarweise zu vergleichen und ihre relative Effizienz geeignet zu bestimmen. Das ist für endliche Stichproben oft nicht einfach, da sich die Gütefunktionen der Tests ja schneiden können und der endliche Stichprobenumfang i.d.R. eine allgemeinere Aussage erschwert. Wir betrachten daher nur die *asymptotische relative Effizienz*, kurz ARE.

Dazu seien (φ_n) und (ψ_n) zwei vom Stichprobenumfang n abhängige Folgen von Tests für das Testproblem $H_0 : \vartheta \in \Theta_0$, $H_1 : \vartheta \in \Theta_1$. Es ist nun nicht sinnvoll, die Tests für festgehaltene Parameterwerte unter H_1 direkt zu vergleichen, da ja für konsistente Tests für $\vartheta \in \Theta_1$ gilt:

$$\mathrm{E}_\vartheta \varphi(\boldsymbol{X}) \to 1, \qquad \mathrm{E}_\vartheta \psi(\boldsymbol{X}) \to 1.$$

Da es auch bei großen Stichproben am schwierigsten ist, Abweichungen von der Nullhypothese zu entdecken, die in der Nähe von H_0 liegen, betrachtet man Folgen von Parameterwerten $\vartheta_n \in \Theta_1$, die gegen die ϑ_0 konvergieren. Sind nun, wie es für zahlreiche Tests gilt, die zugrundeliegenden Teststatistiken auch unter der Alternative asymptotisch normalverteilt,

so können die Gütefunktionen durch geeignete Wahl der Stichprobenumfänge ineinander überführt werden. (Dies gilt speziell für lineare Rangtests und auch für andere, in der nichtparametrischen Statistik betrachtete Tests. Daher wird das Konzept der ARE auch vor allem dort behandelt.) Das Verhältnis der Stichprobenfunktionen ist dann die ARE. Der Wilcoxon-Vorzeichen-Rangtest hat beispielsweise im Verhältnis zum t-Test eine ARE von 0.95, wenn die zugrundeliegende Verteilung die Normalverteilung ist. Das bedeutet, dass man, vereinfacht gesagt, einen Stichprobenumfang von 100 benötigt, um mit diesem nichtparametrischen Test die gleiche Güte zu haben wie bei dem t-Test bei einem Stichprobenumfang von 95. Im Übrigen sinkt die ARE bei keiner stetigen Verteilung unter 0.864 und ist bei Verteilungen mit mehr Wahrscheinlichkeitsmasse in den Flanken wesentlich größer als eins. Analoge Aussagen gelten für den Wilcoxon-Rangsummentes im Vergleich mit dem Zweistichproben-t-Test.

Die ARE wird bei Büning & Trenkler (1994) und ausführlicher bei Randles & Wolfe (1979) behandelt.

Genauso wie bei der Paramerterschätzung stellt sich für statistische Tests die Frage nach der Auswirkung von Abweichungen von den Modellvoraussetzungen und dem Einfluss einzelner, extremer Beobachtungen. Auch hier gibt es einen quantitativ und einen qualitativ orientierten Zugang zu der *Robustheit statistischer Tests*.

Den Ausgangspunkt für die quantitative Untersuchung von Robustheitseigenschaften und den Begriff der Robustheit überhaupt bildet eine Arbeit von Box (1953) über die Varianzanalyse bei nicht-normalverteilten Störungen. Den Bedürfnissen der Anwendung entsprechend wird einmal die Auswirkung auf das Niveau von Tests untersucht. So ist der t-Test als Test auf den Erwartungswert einer normalverteilten Zufallsvariablen relativ α-robust beim zweiseitigen Testproblem. Auf einzelne, extreme Werte reagiert er aber mit einem drastischen Güteverlust. Auch sonst ist er nicht effizienz-robust, wenn die zugrundeliegende Verteilung stärkere Flanken hat als die Normalverteilung. Bei der Laplace-Verteilung sinkt seine relative Effizienz zum Zeichentest auf $1/2$ (von $\pi/2$ bei der Normalverteilung)! Da Rangtests keinerlei Schwierigkeiten mit dem Einhalten des Niveaus haben und etliche von ihnen, wie z.B. die Wilcoxon-Tests, sowohl unter der Normalverteilung als auch für Verteilungen mit viel Wahrscheinlichkeitsmasse in den Flanken eine hohe Effizienz aufweisen, bilden sie unter Robustheitsgesichtspunkten eine adäquate Alternative zu den jeweiligen besten parametrischen Tests. Als andere Möglichkeit wird die Entfernung von Ausreißern aus der Stichprobe empfohlen. Eine ausgezeichnete Behandlung der praktisch relevanten Gesichtspunkte findet sich bei Miller (1995). Büning (1991) gibt einen breiten Überblick über den quantitativen Zugang zur Robustheit von Tests und betrachtet auch robustifizierte Varianten klassischer parametrischer Verfahren.

Der qualitative Zugang geht von Umgebungen der Verteilungen aus, welche unter H_0 und H_1 festgelegt sind und untersucht, wie sich die Gütefunktionen der Tests verhalten. Qualitative Robustheit liegt bei einem statistischen Test vor, wenn bei Verteilungen aus geeignet kleinen Umgebungen nur kleine Änderungen passieren. Diesbezüglich sind Rangtests robust, vgl. Rieder (1982).

Um den qualitativen Zugang etwas zu verdeutlichen, sei ein Resultat von Huber (1965) angeführt. Er betrachtet das Testproblem einer einfachen Hypothese gegen eine einfache Alternative:

$$H_0 : f(x) = f_0(x), \qquad H_1 : f(x) = f_1(x).$$

Es sei die Monotonie des Likelihood-Quotienten $f_1(x)/f_0(x)$ vorausgesetzt. Der nach dem Neyman-Pearson Lemma optimale Test hat die Gestalt

$$\varphi(\boldsymbol{x}) = \begin{cases} 0 & \prod_{i=1}^{n} \frac{f_1(x_i)}{f_0(x_i)} \leq k \\ 1 & \prod_{i=1}^{n} \frac{f_1(x_i)}{f_0(x_i)} > k. \end{cases}$$

Dieser Test wird durch einzelne extreme Werte stark beeinflusst, da extreme Werte mit Faktoren $f_1(x_i)/f_0(x_i)$ in der Nähe von Null oder von extremer Größe leicht den Wert von $\varphi(\boldsymbol{x})$ und damit die Testentscheidung determinieren können.

Lassen wir nun auch Verteilungen zu, die jeweils 'in der Nähe' von $f_0(x)$ bzw. $f_1(x)$ liegen. Damit wird aus dem Testproblem einfacher Hypothesen eines mit zusammengesetzten. Zudem sind die in Abschnitt 8.3 unterstellten Strukturen hier nicht vorhanden, da wir z.B. als Umgebung von $f_j(x)$ mit der Verteilungsfunktion $F_j(x)$ alle Verteilungen zulassen können, die einen geeigneten Kolmogorov-Abstand $\sup_{-\infty<x<\infty} |F_j(x) - F(x)| < \delta$ haben. Jedoch hat Huber gezeigt, dass der modifizierte Test

$$\varphi(\boldsymbol{x}) = \begin{cases} 0 & \prod_{i=1}^{n} q(x_i) \leq C \\ 1 & \prod_{i=1}^{n} q(x_i) > C, \end{cases} \tag{8.41}$$

mit $q(x) = \max\{c', \min\{c'', f_1(x)/f_0(x)\}\}$, $0 < c' < c'' < \infty$, ein Maximin-Test zum Niveau α ist, wenn die Konstanten c', c'' und C geeignet gewählt werden. Die Modifikation des Likelihoodquotienten entspricht gerade der von ihm vorgeschlagenen ψ-Funktion für M-Schätzer.

Ausführlich ist dies bei Huber (1981) und Witting (1985) dargestellt. Rieder (1994) gibt eine geschlossene Abhandlung der robusten Statistik vom asymptotischen Blickwinkel her.

8.5 Multiple Tests[1]

Selten findet sich der Statistiker in der glücklichen Lage, dass ein Versuch geplant, durchgeführt und ausgewertet werden kann mit dem Ziel und dem endgültigen Ergebnis, dass nur eine Frage — eine statistische Hypothese — zu beantworden ist. Führt etwa ein Test für das Problem $H_0 : \vartheta = 0, \quad H_1 : \vartheta \neq 0$ bei einem mehrdimensionalen Parameter zur Ablehnung der Nullhypothese, so möchte der Anwender natürlich auch wissen, welche der Komponenten signifikant von Null verschieden sind. Dann stellt sich aber die Frage nach einer Irrtumswahrscheinlichkeit für die Gesamtaussage, die am Ende steht.

[1] Die Ausführungen dieses Abschnittes halten sich ganz eng an die beiden Arbeiten von Sonnemann (1981,1982).

Wir sprechen von einem *multiplen Testproblem*, wenn für eine Verteilungsfamilie $\mathscr{F} = \{F(\cdot,\vartheta)|\vartheta \in \Theta\}$ $k \geq 2$ Testprobleme gegeben sind:

$$\begin{array}{ll} H_0^1 : \vartheta \in \Theta_0^1, & H_1^1 : \vartheta \in \Theta \backslash \Theta_0^1, \\ \ldots & \\ H_0^k : \vartheta \in \Theta_0^k, & H_1^k : \vartheta \in \Theta \backslash \Theta_0^k. \end{array} \tag{8.42}$$

Mit einem multiplen Testproblem ist nun die Aufgabe verbunden, eine Entscheidungsvorschrift zu bestimmen, die für jede Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)$ zu jedem einzelnen der k Testprobleme eine Testentscheidung liefert.

Definition 8.5.1 *multipler Test*

Ein *multipler Test* oder auch eine *multiple Testprozedur* ist eine Abbildung

$$\varphi : \mathbb{R}^n \to \{0,1\}^k \tag{8.43}$$

mit den Komponenten $\varphi_i : \mathbb{R}^n \to \{0,1\}$. Bei $\varphi_i(\boldsymbol{x}) = 0$ wird die Hypothese H_0^i beibehalten, bei $\varphi_i(\boldsymbol{x}) = 1$ entscheiden wir für H_1^i.

Beispiel 8.5.2 *Qualitätskontrolle von Uhrenfedern*

Bei einem Versuch der Qualitätskontrolle über die Güte von Uhrenfedern sollten gleichartige Federn von drei Firmen verglichen werden. Von jeder Firma wurden vier Federn untersucht. Das Messverfahren bestand darin, an die Uhrfedern genormte Gewichte zu hängen und die Dehnung der Feder zu messen. Es ergaben sich folgende Werte (in inches):

		Feder Nr.			
		1	2	3	4
Firma	A	0.053	0.055	0.061	0.055
	B	0.055	0.055	0.057	0.050
	C	0.043	0.052	0.061	0.050

Es soll ein Test zum Niveau α durchgeführt werden, um zu entscheiden, ob die Federn der drei Firmen im Mittel dieselbe Dehnung aufweisen ($\alpha = 0.05$).
Als Modell werden unabhängige normalverteilte Zufallsvariablen X_{ij} $i = 1,2,3$; $j = 1,2,3,4$ mit $X_{ij} \sim \mathscr{N}(\mu_i, \sigma^2)$ unterstellt. Als Testproblem ergibt sich nun, dass unter H_0 die Erwartungswerte der Firmen gleich sind und unter H_1, dass dies nicht gilt. Mit $\vartheta = (\mu_1, \mu_2, \mu_3, \sigma^2)$ und $\Theta = \mathbb{R}^3 \times \mathbb{R}_+$ lauten die Hypothesen:

$$H_0 : \vartheta \in \Theta_0 = \{(\mu_1, \mu_2, \mu_3, \sigma^2) | (\mu_1, \mu_2, \mu_3, \sigma^2) \in \Theta,\ \mu_1 = \mu_2 = \mu_3\},$$
$$H_1 : \vartheta \in \Theta \backslash \Theta_0.$$

Kurz können wir dafür schreiben:

$$H_0 : \mu_1 = \mu_2 = \mu_3,$$
$$H_1 : \mu_1 \neq \mu_2 \quad \text{oder} \quad \mu_1 \neq \mu_3 \quad \text{oder} \quad \mu_2 \neq \mu_3.$$

Dies kann auch als ein multiples Testproblem formuliert werden. $\mu_1 = \mu_2 = \mu_3$ gilt ja genau dann, wenn $\mu_1 = \mu_2$, $\mu_1 = \mu_3$ und $\mu_2 = \mu_3$. Somit lässt sich die Frage auch in die folgende Form bringen:

$$\begin{aligned} H_0^1 &: \mu_1 = \mu_2, \qquad H_1^1 : \mu_1 \neq \mu_2; \\ H_0^2 &: \mu_1 = \mu_3, \qquad H_1^2 : \mu_1 \neq \mu_3; \\ H_0^3 &: \mu_2 = \mu_3, \qquad H_1^3 : \mu_2 \neq \mu_3. \end{aligned}$$

In dieser Form ist das Testproblem für den Anwender viel interessanter. Kann doch am Ende (hoffentlich) detaillierter erkannt werden, welche der Firmen sich von welcher unterscheidet, nicht nur, dass ein globaler Unterschied besteht.

Im einzelnen Testproblem $H_0^i : \vartheta \in \Theta_0^i, \quad H_1^i : \vartheta \in \Theta_1^i$ wird ein Fehler 1. Art begangen, wenn man sich für H_0^i entscheidet, obwohl der wahre Parameter in Θ_1^i liegt. Im multiplen Testproblem ist es naheliegend, die Formulierungen im Beispiel rückwärts zu lesen und einen Fehler 1. Art dadurch zu charakterisieren, dass die globale Nullhypothese H_0, die genau dann gilt, wenn alle H_0^i gleichzeitig gelten, fälschlich abgelehnt wird. Zur Vereinfachung unterstellen wir, dass die globale Nullhypothese nicht leer ist, dass also $\cap_{i=1}^k \Theta_0^i \neq \emptyset$. Diese Vereinbarung, die auch im weiteren gelten soll, bedeutet, dass diese Nullhypothesen sich nicht gegenseitig ausschließen.

Dann heißt ein multipler Test φ für das multiple Testproblem $H_0^i : \vartheta \in \Theta_0^i, \quad H_1^i : \vartheta \in \Theta \backslash \Theta_0^i$, $(i = 1, \ldots, k)$ ein *Test zum globalen Niveau* α, wenn gilt

$$\forall\, \vartheta \in \Theta_0 = \bigcap_{i=1}^k \Theta_0^i : \quad \mathrm{P}_\vartheta \left(\bigcup_{i=1}^k \{\varphi_i(\boldsymbol{X}) = 1\} \right) \leq \alpha. \tag{8.44}$$

Die Aufteilung des Parameterraums in nur zwei disjunkte Teilmengen Θ_0 und $\Theta \backslash \Theta_0$, wie sie bei Tests zum globalen Niveau erfolgt, erscheint beim multiplen Testproblem aber wenig angebracht. Durch den Rückgriff auf ein einfaches Testproblem geht ja wieder die Möglichkeit einer Detailbetrachtung verloren. Ein multipler Test ist im Gegensatz zum einfachen Test keine Abbildung in die zweielementige Menge $\{0, 1\}$, sondern in die 2^k-elementige Menge $\{0, 1\}^k$.

Eine alternative Beschreibung der möglichen Fehler erhalten wir, wenn wir von den richtigen Entscheidungen ausgehen. Bei einem multiplen Testproblem $H_0^i : \vartheta \in \Theta_0^i, \quad H_1^i : \vartheta \in \Theta_1^i = \Theta \backslash \Theta_0^i$, $(i = 1, \ldots, k)$ liefert ein multipler Test $\varphi = (\varphi_1, \ldots, \varphi_k)$ genau dann auf allen Komponenten eine richtige Entscheidung, wenn der wahre Parameterwert ϑ in $\Theta_{i_1}^1 \cap \ldots \cap \Theta_{i_k}^k$ liegt, $(i_1, \ldots, i_k) \in \{0, 1\}^k$, und wenn gilt: $\varphi(\boldsymbol{x}) = (i_1, \ldots, i_k)$. Es ist bei einem multiplen Testproblem offensichtlich möglich, auf einer Komponente einen Fehler 1. Art zu begehen und gleichzeitig bei einer anderen einen Fehler 2. Art. Lassen wir dieses zu, so gelangen wir zu der folgenden Definition.

Definition 8.5.3 *multiple Fehler*

Gegeben sei ein multiples Testproblem $H_0^i : \vartheta \in \Theta_0^i, \quad H_1^i : \vartheta \in \Theta_1^i = \Theta \backslash \Theta_0^i$, $(i = 1, \ldots, k)$ und ein multipler Test $\varphi = (\varphi_1, \ldots, \varphi_k)$. Ein *multipler Fehler 1. Art* tritt genau dann auf,

wenn für den wahren Parameterwert ϑ gilt

$$\exists\, i \in \{1,\ldots,k\}: \quad \vartheta \in \Theta_0^i \quad \text{und} \quad \varphi_i(\boldsymbol{x}) = 1.$$

Ein *multipler Fehler 2. Art* tritt genau dann auf, wenn für den wahren Parameterwert ϑ gilt

$$\exists\, i \in \{1,\ldots,k\}: \quad \vartheta \in \Theta_1^i \quad \text{und} \quad \varphi_i(\boldsymbol{x}) = 0.$$

Da nun die Wahrscheinlichkeit für einen Fehler 1. Art höchstens α betragen soll, darf die Wahrscheinlichkeit, dass mindestens eine wahre Nullhypothese (fälschlicherweise) abgelehnt wird, α nicht überschreiten. Etwas formaler gesprochen: Für jede mögliche Parameterkonstellation, d.h. für jedes $\vartheta \in \Theta$ mit $\vartheta \in \Theta_0^i$ für mindestens ein i, muss die Ablehnung der dazugehörigen Nullhypothesen insgesamt durch α beschränkt sein. Andersherum lässt sich das auch so ausdrücken, dass die Wahrscheinlichkeit, alle wahren Nullhypothesen anzunehmen, gleichgültig welche und wie viele gelten, mindestens $1-\alpha$ betragen muss.

Definition 8.5.4 *Test zum multiplen Niveau*

Gegeben sei ein multiples Testproblem (8.42) und ein zugehöriger Test φ gemäß Definition 8.5.1. Sei $I = \{1,\ldots,k\}$ und $I(\vartheta) := \{i | i \in I, \vartheta \in \Theta_0^i\}$. Dann heißt φ ein *(multipler) Test zum multiplen Niveau* α, wenn er

$$\forall\, \vartheta \in \Theta,\ I(\vartheta) \neq \emptyset: \quad \mathrm{P}_\vartheta\left(\bigcup_{i \in I(\vartheta)} \{\varphi_i(\boldsymbol{X}) = 1\}\right) \leq \alpha \tag{8.45}$$

erfüllt.

Da ein multipler Test zum multiplen Niveau α auch dann das Niveau einhält, wenn nur eine der Nullhypothesen gilt, $\sum_{j=1} i_j = 1$ für $(i_1,\ldots,i_k) \in \{0,1\}^k$, ist jede Komponente φ_i ein Test zum Niveau α für das zugehörige Testproblem.

Es ist offensichtlich, dass die Bedingung (8.45) die Forderung, die an einen Test mit globalem Niveau gestellt wurde, umfasst. Ein simultaner Test zum multiplen Niveau α ist immer auch ein Test zum globalen Niveau α. Die Definition 8.5.4 stellt jedoch wesentlich höhere Anforderungen an einen Test.

Lemma 8.5.5 *Vereinigungs-Durchschnittsprinzip von Roy*

Gegeben sei das multiple Testproblem $H_0^i: \vartheta \in \Theta_0^i$, $H_1^i: \vartheta \in \Theta_1^i = \Theta \backslash \Theta_0^i$, $i \in I = \{1,\ldots,k\}$, und ein multipler Test $\varphi = (\varphi_1,\ldots,\varphi_k)$. φ ist genau dann ein Test zum multiplen Niveau α, wenn für jede Durchschnittshypothese $H_0^J: \vartheta \in \Theta_0^J = \cap_{j \in J} \Theta_0^j$ mit $J \neq \emptyset$ und $\Theta_0^J \neq \emptyset$ der nach dem allgemeinen *Vereinigungs-Durchschnittsprinzip von Roy* abgeleitete Test φ_J mit dem kritischen Bereich

$$\{\varphi_J = 1\} := \bigcup_{j \in J} \{\varphi_j = 1\}$$

ein Test zum Niveau α ist. Formal:

$$\begin{aligned} &\forall\, J \subseteq I,\ J \neq \emptyset: \quad \forall\, \vartheta \in \Theta_0^J := \textstyle\bigcap_{j\in J} \Theta_0^j: \\ &\mathrm{P}_\vartheta\left(\textstyle\bigcup_{j\in J}\{\varphi_j = 1\}\right) \leq \alpha. \end{aligned} \tag{8.46}$$

Weiter ist φ genau dann ein Test zum multiplen Niveau α, wenn für jede Durchschnittshypothese $H_0^J : \vartheta \in \Theta_0^J = \cap_{j\in J}\Theta_0^j$ mit $J \neq \emptyset$ und $\Theta_0^J \neq \emptyset$ der nach dem allgemeinen *Vereinigungs-Durchschnittsprinzip von Roy* abgeleitete Test φ_J mit dem kritischen Bereich

$$\{\varphi_J = 1\} := \bigcup_{j\in J}\{\varphi_j = 1\}$$

ein Test zum Niveau α ist.

Beweis: Sei φ ein Test zum multiplen Niveau α, und sei $\vartheta \in \Theta_0^J$ für eine nichtleere Teilmenge J von I. Dann gilt $J \subseteq I(\vartheta)$

$$\mathrm{P}_\vartheta\left(\bigcup_{j\in J}\{\varphi_j = 1\}\right) \leq \mathrm{P}_\vartheta\left(\bigcup_{j\in I(\vartheta)}\{\varphi_j = 1\}\right) \leq \alpha.$$

Umgekehrt folgt aus (8.46) mit der jeweiligen Wahl $J = I(\vartheta)$ sofort die Eigenschaft (8.45). Die zweite Aussage folgt offensichtlich aus der ersten.

Beispiel 8.5.6 *Bonferroni-Test*

Gegeben sei ein multiples Testproblem der Form (1). Zu jedem einzelnen Testproblem $H_0^i : \vartheta \in \Theta_0^i$, $H_1^i : \vartheta \in \Theta_1^i$ $(i = 1,\ldots,k)$ möge ein Test φ_i zum Niveau α/k existieren:

$$\forall\, i \in \{1,\ldots,k\}: \quad \forall\, \vartheta \in \Theta_0^i: \quad \mathrm{P}_\vartheta(\varphi_i(\boldsymbol{X}) = 1) \leq \frac{\alpha}{k}.$$

Wir fassen nun die einzelnen Tests φ_i zu einem Test zusammen: $\varphi = (\varphi_1,\ldots,\varphi_k)$. Damit haben wir einen multiplen Test. Dieser ist auch ein Test zum multiplen Niveau, der sogenannte Bonferroni-Test.

Der Nachweis des multiplen Niveaus ist elementar. Wir benötigen lediglich die Bonferroni-Ungleichung und die Voraussetzung, dass die einzelnen Tests das Niveau α/k einhalten. Sei $J \subset \{1,\ldots,k\}$ und $\vartheta \in \bigcap_{i\in J}\Theta_0^i$. Dann gilt:

$$P_\vartheta\left(\bigcup_{i\in J}\{\varphi_i = 1\}\right) \leq \sum_{i\in J} P_\vartheta(\{\varphi_i = 1\}) \leq |J| \cdot \frac{\alpha}{k} \leq \alpha.$$

Die verwendete Bonferroni-Abschätzung ist scharf, wenn die Ereignisse disjunkt sind, sie ist sehr grob, wenn sich beispielsweise alle Wahrscheinlichkeit im Durchschnitt der Ereignisse konzentriert. Auch bei der Abschätzung $(|J| \cdot \alpha/k) \leq \alpha$ verschenkt man umso mehr, je kleiner $|J|$ ist. Dies führt zu der verbesserten Methode, die weiter unten angegeben wird.

Beispiel 8.5.7 *Einseitige und zweiseitige Tests*

Als konkretes Beispiel für den Bonferroni-Test diene ein Zweistichprobenproblem, in dem die Erwartungswerte μ_1 und μ_2 zweier Normalverteilungen mit gleichen, aber unbekannten Varianzen auf der Basis der Beobachtung der Stichprobenvariablen $X_1,\dots,X_m$, $X_{m+1},\dots,X_n$ verglichen werden sollen. Es wird also unterstellt, dass $X_i \sim \mathcal{N}(\mu_1,\sigma^2)$ $(i = 1,\dots,m)$ und $X_i \sim \mathcal{N}(\mu_2,\sigma^2)$ $(i = m+1,\dots,n)$. Das Ziel sei nicht nur das Prüfen von $H_0 : \mu_1 = \mu_2$ gegen $H_1 : \mu_1 \neq \mu_2$, vielmehr soll eine Entscheidung getroffen werden zwischen den drei Aussagen:

$$\mu_1 < \mu_2, \quad \mu_1 = \mu_2, \quad \mu_1 > \mu_2.$$

Dann fassen wir dies als multiples Testproblem auf:

$$H_0^1 : \mu_1 \leq \mu_2, \qquad H_1^1 : \mu_1 > \mu_2;$$
$$H_0^2 : \mu_1 \geq \mu_2, \qquad H_1^2 : \mu_1 < \mu_2.$$

Als Tests verwenden wir die t-Tests, die auf der Teststatistik

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}} \cdot \sqrt{\frac{n}{m(n-m)}}$$

mit

$$\bar{X}_1 = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \bar{X}_2 = \frac{1}{m}\sum_{i=m+1}^{n} X_i, \quad \hat{\sigma} = \frac{1}{n-2}\left\{\sum_{i=1}^{m}(X_i - \bar{X}_1)^2 + \sum_{i=m+1}^{n}(X_i - \bar{X}_2)^2\right\}$$

basieren. Für $\mu_1 = \mu_2$ ist sie t-verteilt mit $n-2$ Freiheitsgraden.

Soll das multiple Niveau α eingehalten werden, so sind die beiden einzelnen Tests zum Niveau $\alpha/2$ durchzuführen:

$$\varphi_1(\boldsymbol{x}) = \begin{cases} 1 & T > t_{n-2;1-\alpha/2} \\ 0 & \text{sonst} \end{cases} \quad \text{und} \quad \varphi_2(\boldsymbol{x}) = \begin{cases} 1 & T < -t_{n-2;1-\alpha/2} \\ 0 & \text{sonst}. \end{cases}$$

Wir erhalten also die gleichen kritischen Werte wie beim zweiseitigen Test.

Von einem multiplen Test sollte man nicht nur verlangen, dass er ein Test zum multiplen Niveau ist. Er soll natürlich auch keine widersprüchlichen Aussagen liefern. Gabriel (1969) hat in seiner grundlegenden Arbeit dafür die Begriffe der Kohärenz und Konsonanz geprägt. Einmal soll mit einer Nullhypothese H_i kohärent auch jede weitere Nullhypothese abgelehnt werden, die eine Teilmenge spezifiziert. Wenn die Nullhypothesenfamilie abgeschlossen ist gegenüber der Durchschnittsbildung, so möchte man zusätzlich Konsonanz erzielen: Wenn eine Nullhypothese abgelehnt wird, zu der andere Nullhypothesen echte Obermengen spezifizieren, so soll auch mindestens eine dieser umfassenden Nullhypothesen abgelehnt werden. Da umfassendere Nullhypothesen spezieller sind, bedeutet Konsonanz, dass man einen globaleren Unterschied auf mindestens einen speziellen zurückführen kann.

Unverfälschtheit multipler Testprozeduren wird bei Gather, Pawlitschko & Pigeot (1995) diskutiert.

Definition 8.5.8 *kohärente und konsonante Tests, durchschnittsabgeschlosses Testproblem*

Gegeben sei ein multiples Testproblem der Form (8.42) und ein zugehöriger multipler Test $\varphi = (\varphi_1, \ldots, \varphi_k)$. Der Test φ heißt *kohärent*, wenn

$$\forall\, i, j \in I: \quad \Theta_0^j \subset \Theta_0^i \quad \Rightarrow \quad \{\varphi_i = 1\} \subseteq \{\varphi_j = 1\}. \tag{8.47}$$

Das Testproblem bzw. die zugehörige Familie von Nullhypothesen heißt *durchschnittsabgeschlossen*, wenn gilt:

$$\forall\, J \subset I = \{1, \ldots, k\}: \quad \exists\, i \in I: \quad \bigcap_{j \in J} \Theta_0^j = \Theta_0^i. \tag{8.48}$$

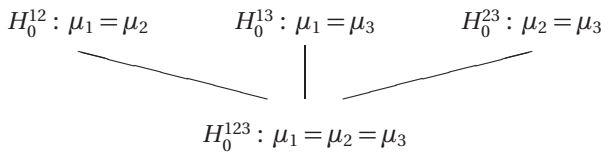
Ist das Testproblem durchschnittsabgeschlossen, so heißt φ *konsonant*, wenn

$$\forall\, i \in I: \quad \exists\, j \in I: \quad \Theta_0^i \subset \Theta_0^j \quad \Rightarrow \quad \{\varphi_i = 1\} \subseteq \bigcup_{j:\Theta_0^i \subset \Theta_0^j} \{\varphi_j = 1\}. \tag{8.49}$$

In einer Hypothesenfamilie $H_0^i: \vartheta \in \Theta_0^i$, $(i = 1, \ldots, k)$ bezeichnen wir als Elementarhypothesen diejenigen, die keine echte Obermenge haben. Genauer ist Θ_0^i eine Elementarhypothese, falls es keinen Index $j \neq i$ gibt mit $\Theta_0^i \subset \Theta_0^j$.

Beispiel 8.5.9 *Qualitätskontrolle von Uhrenfedern - Fortsetzung von Seite 322*

Gegeben seien drei elementare und eine globale Nullhypothese



Diese Familie von Nullhypothesen ist offensichtlich durchschnittsabgeschlossen. Wir bezeichnen mit φ_{ij} die zugehörigen Tests und ordnen die Entscheidungen bzw. die Realisationen der Testfunktionen entsprechend zu dem obigen Hypothesenschema an. Der Test $\varphi = (\varphi_{12}, \varphi_{13}, \varphi_{23}, \varphi_{123})$ ist kohärent, wenn – bis auf Permutationen in der oberen Zeile – nur folgende Entscheidungen möglich sind:

0 0 0	0 0 0	1 0 0	1 1 0	1 1 1
0	1	1	1	1

Die zweite Entscheidungsmöglichkeit muss ausgeschlossen sein, damit der Test auch konsonant ist. Konsonanz bedeutet nämlich, dass man einen globalen Unterschied auf mindestens einen speziellen zurückführen kann.

Wir wollen nun ein allgemeines Prinzip formulieren, wie aus einzelnen Tests zum vorgegebenem Niveau α ein kohärenter multipler Test zum multiplen Niveau α konstruiert werden kann.

Definition 8.5.10 *Abschlusstest*

Gegeben sei ein multiples Testproblem mit der durchschnitsabgeschlossenen Nullhypothesenfamilie $H_0^1 : \vartheta \in \Theta_0^1, \ldots, H_0^k : \vartheta \in \Theta_0^k$ und ein multipler Test φ mit den Komponenten φ_i. Die φ_i seien sämtlich Tests zum Niveau α für die zugehörigen Testprobleme $H_0^i : \vartheta \in \Theta_0^i, \quad H_1^i : \vartheta \in \Theta_1^i$.

Der *Abschlusstest* $\psi(\varphi) = (\psi_1, \ldots, \psi_k)$ lehnt die Hypothese $H_0^i : \vartheta \in \Theta_0^i$ genau dann ab, wenn alle Hypothesen $H_0^j : \vartheta \in \Theta_0^j$ mit $\Theta_0^j \subseteq \Theta_0^i$ durch ihre φ_j abgelehnt werden. Formal:

$$\forall\, i \in I,\ \forall\, \boldsymbol{x} \in \mathbb{R}^n : \quad \psi_i(\boldsymbol{x}) = \begin{cases} 1 & \forall\, j \in I : [\Theta_0^j \subseteq \Theta_0^i : \varphi_j(\boldsymbol{x}) = 1] \\ 0 & \text{sonst .} \end{cases} \tag{8.50}$$

Satz 8.5.11 *Kohärenz des Abschlusstests*

Es seien die Bedingungen der vorstehenden Definition gegeben. Dann gilt:

(i) Ist der Test φ kohärent, so ist φ ein Test zum multiplen Niveau α.

(ii) Der Abschlusstest $\psi(\varphi)$ ist ein kohärenter Test zum multiplen Niveau α.

Beweis: Wir bezeichnen mit I wieder die Menge $\{1, \ldots, k\}$.
zu (i):
Seien $\vartheta \in \Theta$ und $I(\vartheta) = \{i | i \in I,\ \vartheta \in \Theta_0^i\}$. Wir unterstellen, dass $I(\vartheta) \neq \emptyset$. Dann folgt aus der Durchschnittsabgeschlossenheit der Nullhypothesen:

$$\exists\, j \in I : \quad \Theta_0^j = \bigcap_{i \in I(\vartheta)} \Theta_0^i.$$

Damit gilt wegen der Kohärenz: $\forall\, i \in I(\vartheta) : \{\varphi_i = 1\} \subseteq \{\varphi_j = 1\}$.
Wegen $j \in I(\vartheta)$ folgt weiter:

$$\{\varphi_j = 1\} \subseteq \bigcup_{i \in I(\vartheta)} \{\varphi_i = 1\} \subseteq \{\varphi_j = 1\}.$$

Damit folgt die Behauptung:

$$\mathrm{P}_\vartheta \left(\bigcup_{i \in I(\vartheta)} \{\varphi_i = 1\} \right) = \mathrm{P}_\vartheta(\varphi_j = 1) \leq \alpha.$$

zu (ii):
Der Abschlusstest ist per definitionem kohärent, da aus (5) für $\Theta_0^j \subseteq \Theta_0^i$ direkt folgt:

$$\{\psi_i = 1\} = \bigcap_{k : \Theta_0^k \subseteq \Theta_0^i} \{\varphi_k = 1\} \quad \subseteq \quad \bigcap_{k : \Theta_0^k \subseteq \Theta_0^j} \{\varphi_k = 1\} = \{\psi_j = 1\}.$$

Seine Komponenten sind wegen $\{\psi_i = 1\} \subseteq \{\varphi_i = 1\}$ sämtlich Tests zum Niveau α, so dass $\psi(\varphi)$ die Voraussetzungen von Teil (i) erfüllt.

Der Abschluss-Test ist also per definitionem kohärent. Ob er auch konsonant ist, hängt von den verwendeten Niveau-α-Tests ab. Siehe dazu die Literatur.

Beispiel 8.5.12 *Zur Durchführung des Abschluss-Tests*

In den Anwendungen ist es i. Allg. nicht notwendig, die Komponenten des Abschluss-Tests explizit zu bestimmen. Bei endlichen Hypothesenfamilien führt man die gegebenen Niveau-α-Tests sequentiell durch, indem man in dem Hypothesenverband schrittweise vorgeht. Dies sei am Beispiel dreier Elementarhypothesen $H_0^i : \vartheta \in \Theta_0^i$, $H_1^i : \vartheta \in \Theta_1^i$ $(i = 1,2,3)$ verdeutlicht, wobei die Nullhypothesenfamilie auch alle Durchschnitte der Θ_0^i umfasse. Das schrittweise Vorgehen lässt sich nun folgendermaßen darstellen:

$$\left.\begin{array}{ccc} H_0^1 & H_0^2 & H_0^3 \\ \varphi_1 = 1? & \varphi_2 = 1? & \varphi_3 = 1? \end{array}\right\} \text{3.Stufe}$$

$$\left.\begin{array}{ccc} H_0^{12} & H_0^{13} & H_0^{23} \\ \varphi_{12} = 1? & \varphi_{13} = 1? & \varphi_{23} = 1? \end{array}\right\} \text{2.Stufe}$$

$$\left.\begin{array}{c} H_0^{123} \\ \varphi_{123} = 1? \end{array}\right\} \text{1.Stufe}$$

Führt auf diesem Stufenweg ein Test zur Beibehaltung der zugehörigen Nullhypothese, so werden auch alle Nullhypothesen beibehalten, die diese umfassen. Der Abschlusstest kann also folgende Entscheidungsmuster ergeben (abgesehen von Permutationen der H_0^i):

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	1 0 0	1 0 0	1 1 0	1 1 1
0 0 0	0 0 0	1 0 0	1 1 0	1 1 1	1 1 0	1 1 1	1 1 1	1 1 1
0	1	1	1	1	1	1	1	1

Daran ist die Kohärenz direkt abzulesen. Konsonanz gilt genau dann, wenn höchstens die Muster (1),(6),(8) und (9) auftreten.

Beispiel 8.5.13 *Einseitige und zweiseitige Tests - Fortsetzung von Seite 326*

Die beiden Nullhypothesen im Beispiel 8.5.7 bilden noch keine durchschnittsabgeschlossene Nullhypothesenfamilie. Daher erweitern wir sie um die globale Nullhypothese. Dann haben wir die Nullhypothesen:

$$H_0 : \mu_1 = \mu_2, \qquad H_0^1 : \mu_1 \le \mu_2, \qquad H_0^2 : \mu_1 \ge \mu_2.$$

Wir haben dann drei t-Tests zum Niveau α:

$$\varphi_1(\boldsymbol{x}) = \begin{cases} 1 & T > t_{n-2;1-\alpha} \\ 0 & \text{sonst} \end{cases},$$

$$\varphi_2(\boldsymbol{x}) = \begin{cases} 1 & T < -t_{n-2;1-\alpha} \\ 0 & \text{sonst} \end{cases},$$

$$\varphi_3(\boldsymbol{x}) = \begin{cases} 1 & T < -t_{n-2;1-\alpha/2},\ T > t_{n-2;1-\alpha/2} \\ 0 & \text{sonst} \end{cases}.$$

Der Abschlusstest beginnt dann mit der Anwendung von φ_3. Führt dieser Test zur Ablehnung von H_0, so führt auf der zweiten Stufe ebenfalls einer der beiden Tests zur Ablehnung seiner einseitigen Nullhypothese. Lehnt andererseits φ_3 die Hypothese H_0 nicht ab, so lehnt der Abschlusstest keine der drei Nullhypothesen ab.
Betrachtet man das Problem von einer anderen Seite, so hat der Abschlusstest zur Folge, dass man lediglich den zweiseitigen Test durchzuführen hat. Führt dieser zur Annahme, entscheidet man sich für H_0. Erhält man eine Ablehnung, so entscheidet man sich für H_1^1 bzw. H_1^2, je nachdem ob $T < -t_{n-2;1-\alpha/2}$ oder $T > t_{n-2;1-\alpha/2}$. Die Wahrscheinlichkeit für einen Fehler 1. Art bleibt durch α beschränkt

Beispiel 8.5.14 *Bonferroni-Holm-Test*

Wie im Beispiel 8.5.6 gezeigt wurde, kann man für jedes multiple Testproblem den Bonferroni-Test nutzen. Er ist ein Test zum multiplen Niveau α, und man benötigt als Voraussetzung nur die Kenntnis der kritischen Werte c_i für die Tests φ_i der einzelnen Nullhypothesen H_0^i oder die zugehörigen Überschreitungswahrscheinlichkeiten p_i. Über das Abschlussprinzip lässt er sich jedoch gleichmäßig zum Bonferroni-Holm-Test verbessern. Dieser sieht in seiner sequentiellen Form zunächst einmal sehr überraschend und unplausibel aus. Daher geben wir ihn erst einmal in dieser Form an und führen die formalen Betrachtungen im Anschluss durch.
Die Testprobleme seien wieder durch Paare nicht leerer und zueinander komplementärer Teilmengen des Parameterraums gegeben:

$$H_0^i : \vartheta \in \Theta_0^i, \qquad H_1^i : \vartheta \in \Theta_1^i \qquad i \in I = \{1, \ldots, k\}.$$

Für jedes einzelne Testproblem H_0^i gegen H_1^i gebe es eine reellwertige Teststatistik $T_i = T_i(\boldsymbol{X})$, die zu größeren Werten tendiere, wenn die Nullhypothese H_0^i falsch ist. Wir werden den Test aus Übersichtlichkeitsgründen mit den zugehörigen Überschreitungswahrscheinlichkeiten $p_i(\boldsymbol{x}) = \mathrm{P}(T_i(\boldsymbol{X}) > T_i(\boldsymbol{x}))$ formulieren; dabei ist $\boldsymbol{x} = (x_1, \ldots, x_n)$ die realisierte Stichprobe. Für die p_i gilt:

$$\forall\, i \in I = \{1, \ldots, k\} : \quad \forall\, m \in \{1, \ldots, k\} : \quad \forall\, \vartheta \in \Theta_0^i :$$

$$\mathrm{P}_\vartheta \left(p_i \leq \frac{\alpha}{m} \right) \leq \frac{\alpha}{m}.$$

Diese Voraussetzungen genügen, den Bonferroni-Holm-Test, auf die folgende Weise durchzuführen:
Die für die einzelnen Nullhypothesen im Experiment beobachteten Überschreitungswahrscheinlichkeiten $p_i = p_i(\boldsymbol{x})$, $i = 1, \ldots, k$, werden der Größe nach geordnet, so dass gilt

$$p_{(1)} < p_{(2)} < \ldots < p_{(k)}.$$

Die zugehörigen Testprobleme werden entsprechend angeordnet und dann mit

$$H_0^{(1)} : H_1^{(1)},\ H_0^{(2)} : H_1^{(2)}, \ldots, H_0^{(k)} : H_1^{(k)}$$

bezeichnet. Damit finden wir die Entscheidung sofort in dem in der Abbildung dargestellten Entscheidungsbaum.

Der durch die beschriebene Vorgehensweise und die Abbildung definierte Bonferroni-Holm-Test ist ein Test zum multiplen Niveau α. Er ist gleichmäßig besser als der klassische Bonferroni-Test in dem Sinne, dass alle die Nullhypothesen, die der Bonferroni-Test ablehnt, auch von ihm abgelehnt werden, aber im Allgemeinen noch einige zusätzlich. Bindungen in den $p_{(i)} = p_{(i)}(\boldsymbol{x})$ bereiten keine Probleme. Gilt etwa $p_{(i)}(\boldsymbol{x}) = p_{(i+1)}(\boldsymbol{x})$, so folgt aus $p_{(i)}(\boldsymbol{x}) > \alpha/(k-i+1)$, dass $H_0^{(i)}$ und damit auch $H_0^{(i+1)}$ angenommen wird, und aus $p_{(i)}(\boldsymbol{x}) \leq \alpha/(k-i+1)$ folgt die Ablehnung von $H_0^{(i)}$ und wegen

$$p_{(i+1)}(\boldsymbol{x}) = p_{(i)}(\boldsymbol{x}) \leq \frac{\alpha}{k-i+1} \leq \frac{\alpha}{k-i}$$

auch die Ablehnung von $H_0^{(i+1)}$. Gleiche Überschreitungswahrscheinlichkeiten führen also unabhängig von ihrer Anordnung zu ein und derselben Entscheidung. Wir wollen jetzt

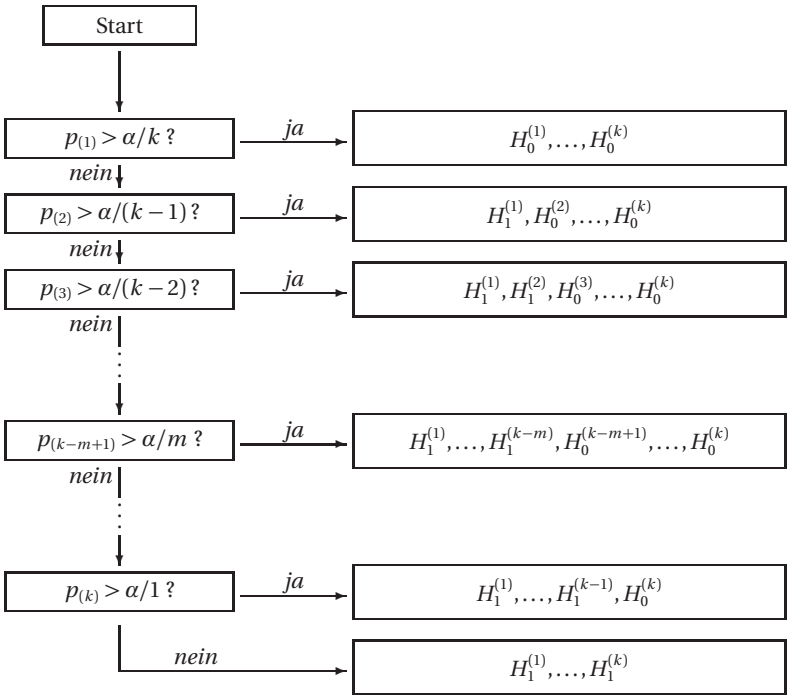


Abb. 8.10: *Entscheidungsfindung beim Bonferroni-Holm-Test*

zeigen, dass diesem Vorgehen ein Abschlusstest zugrunde liegt. Wir erweitern zunächst die k gegebenen Testprobleme zu einer durchschnittsabgeschlossenen Familie von K Testproblemen, indem wir für $J \subseteq I = \{1, \ldots, k\}$ setzen:

$$H_0^J : \vartheta \in \Theta_0^J = \bigcap_{j \in J} \Theta_0^j.$$

Gemäß dem Vereinigungs-Durchschnittsprinzip von Roy definieren wir dann die zuge-

hörigen Tests

$$\varphi_J(\boldsymbol{x}) = \begin{cases} 1 & \text{für } \min\limits_{j\in J} p_j(\boldsymbol{x}) \le \dfrac{\alpha}{|J|} \\ 0 & \text{sonst} \end{cases} .$$

Diese K Tests sind offensichtlich alle Tests zum Niveau α. Damit ist der aus den φ_J resultierende Abschlusstest ein kohärenter Test zum multiplen Niveau α. Nach dem Abschlussprinzip werden mit der ersten Hypothese $H_0^{(1)}$ auch alle Hypothesen verworfen, die aus einem Schnitt mit $\Theta_0^{(1)}$ resultieren. Insbesondere ist dies auch für die Globalhypothese $H_0 : \vartheta \in \cap_{i=1}^k \Theta_0^i$ der Fall. Im zweiten Schritt kommen nur noch solche Testprobleme vor, die aus Schnitten von höchstens $k-1$ Ausgangshypothesen resultieren. Dazu brauchen wir nur noch die Überschreitungswahrscheinlichkeiten mit $\alpha/(k-1)$ zu vergleichen. Die Fortsetzung dieser Überlegung zeigt die Behauptung.

Es ist offensichtlich, dass der Bonferroni-Holm-Test auch konsonant ist, reicht doch jede Ablehnung bis zu mindestens einer Elementarhypothese.

8.6 Aufgaben

Aufgabe 1

Der Anpassungstest von Kolmogorov ist bei diskreten Verteilungen konservativ, d.h. er darf angewendet werden, jedoch führt er unter H_0 seltener zur Ablehnung als das Niveau eigentlich zulässt.

Überprüfen Sie zum Niveau $\alpha = 0.01$ die Hypothese, dass die Verteilung der Endziffern beim Ablesen von 733 Fieberthermometern eine diskrete Gleichverteilung besitzt:

Ziffer	0	1	2	3	4	5	6	7	8	9
Häufigkeit(%)	14.5	8.2	12.5	9.9	9.9	8.0	8.5	8.8	9.5	10.2

Aufgabe 2

Ein Produkt wird auf fünf verschiedene Weisen, die mit $\vartheta = 1,2,3,4,5$ gekennzeichnet sind, hergestellt. Je nach Herstellungsweise gibt es unterschiedliche Ausschusswahrscheinlichkeiten. Genauer gilt für die Zufallsvariable

$$X = \begin{cases} 0 & \text{falls das Produkt kein Ausschuss ist} \\ 1 & \text{falls das Produkt Ausschuss ist} \end{cases} :$$

	ϑ				
x	1	2	3	4	5
0	0.9	0.7	0.5	0.3	0.1
1	0.1	0.3	0.5	0.7	0.9

In einer Lieferung werden $n = 3$ Produkte daraufhin untersucht, ob sie Ausschuss sind.

i) Bestimmen Sie den Likelihood-Quotienten-Test zum Niveau $\alpha = 0.225$ für das Testproblem
$H_0 : \vartheta \le 2$ gegen $H_1 : \vartheta > 2$.

ii) Bestimmen Sie die Gütefunktion des Tests.

Aufgabe 3

Der Kaffeegeschmack (= X) habe drei Realisationsmöglichkeiten: -1='schlecht', 0='mittel', 1='gut'.
Bei vier Kaffeesorten ($\vartheta = 1,2,3,4$) einer Firma sind die Verteilungen des Geschmacks unter den Käufern bekannt:

	ϑ			
x	1	2	3	4
-1	0.6	0.4	0.3	0.1
0	0.3	0.5	0.4	0.3
1	0.1	0.1	0.3	0.6

Bei einer Lieferung für ein Büro, das eine Sorte angefordert hatte, fehlen die Aufkleber auf den sonst bei allen Sorten gleich aussehenden Packungen. Um festzustellen, ob der gelieferte Kaffee zu einer der beiden besseren Sorten ($\vartheta > 2$) gehört, sollen zwie Büroangestellte unabhängig voneinander den Kaffee einer Packung kosten und ihre Einschätzung abgeben.

(i) Geben Sie alle Tests zum Niveau $\alpha = 0.13$ für $H_0 : \vartheta \leq 2$ gegen $H_1 : \vartheta > 2$ an.

(ii) Geben Sie die Gütefunktion der Tests an.

(iii) Welche Tests sind unverfälscht?

(iv) Gibt es einen besten unter diesen Tests?

Aufgabe 4

Kunde König ärgert sich, dass in einem Supermarkt 8 andere Kunden vor ihm an der Kasse stehen. Der Ärger rührt auch daher, dass in der Werbung versprochen wurde, dass im Mittel nur 3 andere Kunden jeweils an der Kasse stünden. Bevor sich der Kunde König beschwert, will er prüfen, ob diese lange Warteschlange eine Ausnahme ist, ein 'Ausreißer' gewesen sein kann.

Für die Überprüfung will er einen statistischen Test durchführen. Da jeder vor ihm an der Kasse stehende Kunde als Misserfolg bei dem Versuch, selbst dranzukommen, anzusehen ist, unterstellt er eine geometrische Verteilung für die Anzahl der vor ihm an der Kasse wartenden Kunden, i.Z.: $X \sim \mathcal{GEO}(p)$. Die Anzahlen $X_1, \ldots, X_n$ bei verschiedenen Einkäufen können als unabhängig angesehen werden. Die Fragestellung lautet dann formal:

$H_0 : X_i \sim \mathcal{GEO}(1/4) \quad i \in \{1, \ldots, n\}$
$H_1 : X_i \sim \mathcal{GEO}(1/4) \quad i \in \{1, \ldots, n\}/\{i_0\}, \quad X_{i_0} \sim \mathcal{GEO}(p)$ mit $p < 1/4$.

a) Geben Sie einen auf $X_{(n)}$ basierenden Test zum Niveau α für H_0 gegen H_1 an.

b) Wie lautet die Testentscheidung bei $\alpha = 0.15$ unter Verwendung der Aufzeichnungen der letzten 4 Einkäufe, die Warteschlangen von 4,2,3,8 (vor ihm wartende) Personen ergaben?

c) Herr König hat gehört, dass i.d.R. mehr Daten ein schärferes Resultat liefern. Er bezieht daher die folgenden Einkäufe bei einer erneuten Testdurchführung mit ein. Die Warteschlangenlänge von ≥ 8 hat er dabei nicht mehr beobachtet. Wie lautet nun sein Ergebnis?

d) War Kunde König gut beraten, seine Beobachtungsserie zu vergrößern, d.h. liefert der Test eher ein gerechtfertigtes signifikantes Ergebnis, wenn der Stichprobenumfang wächst? Beantworten Sie diese Frage für die Grenzsituation $n \to \infty$.

Aufgabe 5

Eine Erhebung bei 2 000 Personen ergab folgende Verteilung der Anzahl von Einheiten eines bestimmten Produktes, die in einem halben Jahr gekauft worden waren:

x_i	1	2	3	4	5	6	7	8	9	10	11	12	13
n_i	164	71	42	28	17	12	12	9	7	6	3	3	5

Zudem wurden die Werte 14,16,19,20,21,22,24,26 je einmal beobachtet.
Für die Verteilung derjenigen, die mindestens einmal diese Produkt kaufen, wird die logarithmische Verteilung als sinnvolles Modell angesehen:

$$P_\vartheta(X = x) = \frac{a\vartheta^x}{x} \quad \text{mit } x = 1,2,3,\ldots$$

Dabei ist $0 < \vartheta < 1$ und $a = -\ln(1-\vartheta)^{-1}$. Wie beurteilen Sie die Eignung?

Aufgabe 6

Das Gewicht von Kork-Reserven von 28 Bäumen (in 100steln Gramm) wurde für jede der 4 Himmelsrichtungen ermittelt. Gibt es einen signifikanten Unterschied zwischen den vier Himmelsrichtungen?

Nord	Ost	Süd	West	Nord	Ost	Süd	West	Nord	Ost	Süd	West
72	66	76	77	91	79	100	75	32	30	34	28
60	53	66	63	56	68	47	50	63	45	74	63
56	57	64	58	79	65	70	61	54	46	60	52
41	29	36	38	81	80	68	58	47	51	52	43
32	32	35	36	78	55	67	60	39	36	39	31
30	35	34	26	46	38	37	38	50	34	37	40
39	39	31	27	39	35	34	37	43	37	39	50
42	43	31	25	32	30	30	32	48	54	57	43
37	40	31	25	60	50	67	54				
33	29	27	36	35	37	48	39				

Aufgabe 7

Die Zufallsvariable X habe eine Cauchy-Verteilung,

$$f(x) = \frac{1}{\lambda\pi} \cdot \frac{1}{1+((x-\mu)/\lambda)^2}.$$

Bei bekanntem $\mu = 0$ soll der optimale Test für die Nullhypothese $H_0 : \lambda = 1$ gegen die Alternative $H_1 : \lambda = 2$ zum Niveau $\alpha = 0.1$ bestimmt werden.

Hinweis: $F_{0,\lambda}(x) = 0.5 + \frac{1}{\pi}\arctan(x/\lambda)$, $\arctan(-x) = -\arctan(x)$.

Anhang A
Das Riemann-Stieltjes-Integral

Wir wollen hier einige Grundlagen der Theorie der Riemann-Stieltjes-Integrale angeben. Es erlaubt eine einheitliche Sprech- und Schreibweise für diskrete und stetige Verteilungen. Eine ausführlichere Diskussion des Riemann-Stieltjes-Integrals findet sich bei Smirnow (1976).

Als Ausgangspunkt betrachten wir das *bestimmte Riemann-Integral*. Sei f eine auf $[a,b]$ definierte, beschränkte Funktion. Es gilt also $|f(x)| < K$ für alle $a \le x \le b$ für eine geeignete Konstante K. Dann sei eine Zerlegung $\mathscr{Z}_n$ des Intervalls $[a,b]$ in Teilintervalle $(x_{i-1}, x_i]$ $(1 \le i \le n)$ gegeben mit

$$a = x_0 < x_1 < \ldots < x_n = b.$$

Zu einer Zerlegung wird jeweils die Riemann-Summe

$$R(z_n) = \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1})$$

gebildet; ξ_i ist dabei ein Zwischenwert aus dem Intervall $[x_{i-1}, x_i]$, $x_{i-1} \le \xi_i \le x_i$. Mit $\Delta_i = x_i - x_{i-1}$ ist das (Riemann-) Integral dann definiert als der Grenzwert aller Riemann-Summen für $\Delta_i \to 0$:

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{\Delta_i \to 0} \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}).$$

Es lässt sich zeigen, dass dieser Grenzwert unter bestimmten Bedingungen unabhängig von der Wahl der Zerlegungen und der Wahl der Zwischenpunkte ξ_i ist.

Die Verallgemeinerung, die vom Riemann-Integral zum Rieman-Stieltjes-Integral führt, besteht nun darin, dass die Intervall-Längen gewichtet in die Summenbildung eingehen. Dies geschieht mittels geeigneter, auf dem gleichen Intervall definierter Funktionen.

Definition A.1 *Riemann-Stieltjes-Integral*

Gegeben seien zwei beschränkte Funktionen f und g über dem Intervall $[a,b]$. Die Zerlegungen von $[a,b]$ u. s. w. seien wie oben eingeführt. Als Riemann-Stieltjes-Summe bezeichnen wir den Ausdruck

$$\sum_{i=1}^{n} f(\xi_i)(g(x_i) - g(x_{i-1})) \qquad \text{mit} \quad x_{i-1} \le \xi_i \le x_i.$$

Sofern der Grenzwert der Riemann-Stieltjes-Summen für alle Zerlegungsfolgen mit

$$\max\{\Delta_i | 1 \leq i \leq n\} \to 0$$

und alle beliebigen Zwischenpunkte ξ_i existiert und gleich ist, heißt er dann das Riemann-Stieltjes-Integral von f bzgl. g über $[a,b]$. Wir schreiben dafür

$$\int_a^b f(x)\,\mathrm{d}g(x) = \lim_{\Delta_i \to 0} \sum_{i=1}^{n} f(\xi_i)(g(x_i) - g(x_{i-1})).$$

Man nennt g auch die integrierende Funktion oder die Belegungsfunktion. Für $g(x) = x$ liegt gerade der Fall des gewöhnlichen Riemann-Integrals vor.

Zuerst stellt sich die Frage, wann ein Riemann-Stieltjes-Integral existiert.

Satz A.2 Existenz des Riemann-Stieltjes-Integrals

Die integrierende Funktion g sei monoton und beschränkt.

(1) Ist f stetig auf $[a,b]$, dann existiert das Riemann-Stieltjes-Integral von f in Bezug auf g.

(2) Ist f stückweise stetig und fallen die Unstetigkeitspunkte von f und g nicht zusammen, so existiert das Riemann-Stieltjes-Integral von f in Bezug auf g.

Der Satz ist für uns bedeutsam, da Verteilungsfunktionen monoton und beschränkt sind.

Beim Riemann-Stieltjes-Integral ist es – anders als beim Riemann-Integral – wichtig anzugeben, ob die Intervallgrenzen mit zum Integrationsbereich gehören oder nicht. Das Integral von f bezüglich g über $[a,b]$ kann sich nämlich vom Integral von f bezüglich g über $(a,b]$ unterscheiden. Wir treffen daher die folgende Vereinbarung bzgl. der Schreibweise:

$a-0$: Die untere Intervallgrenze a gehört zum Integrationsbereich.

$a+0$: Die untere Intervallgrenze a gehört nicht zum Integrationsbereich.

Mit dieser Vereinbarung gilt

$$\int_{a-0}^{b} f(x)\,\mathrm{d}g(x) - \int_{a+0}^{b} f(x)\,\mathrm{d}g(x) = f(a)(g(a+0) - g(a-0)).$$

$g(a+0) - g(a-0)$ ist gerade der ggf. vorliegende Sprung der Funktion g im Punkt a.

Wenn nichts anderes gesagt ist, wird der rechte Eckpunkt der Intervalle zum Integrationsbereich gehörig betrachtet, der linke nicht. Es ist also, wenn $b+0$ entsprechend dahingehend interpretiert wird, dass die obere Integrationsgrenze zum Integrationsbereich gehört:

$$\int_a^b f(x)\,\mathrm{d}g(x) := \int_{a+0}^{b+0} f(x)\,\mathrm{d}g(x).$$

Für die Verbindung mit den Verteilungsfunktionen sind die beiden folgenden Aussagen essentiell:

Satz A.3

(i) Sei h beschränkt und g monoton wachsend auf $[a,b]$ und konstant auf jedem der Teilintervalle (x_{i-1},x_i) $(1\le i\le n)$ wobei $a=x_0<x_1<\ldots<x_n<b$. Dann gilt

$$\int_a^b h(x)\,\mathrm{d}g(x)=h(a)(g(a+0)-g(a)) + \sum_{i=1}^{n} h(x_i)(g(x_i+0)-g(x_i-0))+h(b)(g(b)-g(b-0)).$$

(ii) Sei h auf $[a,b]$ Riemann-Stieltjes-integrierbar bzgl. g, wobei g eine stetige Ableitung auf $[a,b]$ habe. Dann gilt:

$$\int_a^b h(x)\,\mathrm{d}g(x)=\int_a^b h(x)g'(x)\,\mathrm{d}x\,.$$

Beweis: Siehe Khuri (1993, S. 235).

Teil (ii) des Satzes bleibt richtig, wenn g monoton und g' stückweise stetig ist.

Da Verteilungsfunktionen F monoton sind, existiert nach Satz A.2 für stetige und stückweise stetige Funktionen h das Riemann-Stieltjes-Integral von h bezüglich F.

Für $h(x)=1$ gilt wegen der rechtsseitigen Stetigkeit von F

$$\int_a^b \mathrm{d}F(x)=\int_{a+0}^{b+0}\mathrm{d}F(x)=F(b)-F(a).$$

Für die Verteilungsfunktion F einer kontinuierlichen Zufallsvariablen erhalten wir mit Satz A.3, da aufgrund der Eigenschaften des Riemann-Integrals fast überall $F'(x)=f(x)$ gilt:

$$\int_a^b h(x)\,\mathrm{d}F(x)=\int_a^b h(x)f(x)\,\mathrm{d}x,$$

und speziell für $h(x)=1$:

$$\int_a^b \mathrm{d}F(x)=\int_a^b f(x)\,\mathrm{d}x.$$

Wir sind also wieder in den vertrauten Gefilden des Riemann-Integrals.

Betrachten wir nun die Verteilungsfunktion F einer diskreten Zufallsvariablen mit den Realisationsmöglichkeiten x_i, $x_i<x_{i+1}$, $(i=1,2,\ldots)$. F hat Sprünge in den Punkten x_i; zwischen

diesen Punkten ist F konstant. Nach dem Teil (i) des Satzes A.3 folgt:

$$\int_a^b h(x)\,\mathrm{d}F(x) = \sum_{a<x_i<b} h(x_i)(F(x_i+0)-F(x_i-0)).$$

Da wir die Sprünge der diskreten Verteilungsfunktion mit $\mathrm{P}(X=x_i)=f(x_i)$ bezeichnet haben, kommen wir mit der rechsseitigen Stetigkeit zu der vertrauten Formel

$$\int_a^b h(x)\,\mathrm{d}F(x) = \sum_{a<x_i<b} h(x_i)f(x_i).$$

Speziell für $h(x)=1$ finden wir

$$\int_a^b \mathrm{d}F(x) = \sum_{a<x_i\le b} f(x_i).$$

Satz A.4 Regeln für das Rechnen mit Riemann-Stieltjes-Integralen

Wir setzen voraus, dass die Integrale existieren. Dann gilt:

1. $\int_a^b c\cdot h(x)\,\mathrm{d}g(x) = c\cdot\int_a^b h(x)\,\mathrm{d}g(x)$.
2. $\int_a^b \left(\sum_i h_i(x)\right)\mathrm{d}g(x) = \sum_i \int_a^b h_i(x)\,\mathrm{d}g(x)$.
3. Ist h nicht negativ und g nicht abnehmend, so gilt: $\int_a^b h(x)\,\mathrm{d}g(x)\ge 0$.
4. $\int_a^b \sum_{i=1}^k c_i h_i(x)\,\mathrm{d}g(x) = \sum_{i=1}^k c_i \int_a^b h_i(x)\,\mathrm{d}g(x)$.
5. $\int_a^b h(x)\mathrm{d}\left(\sum_{i=1}^k c_i g_i(x)\right) = \sum_{i=1}^k c_i \int_a^b h(x)\,\mathrm{d}g_i(x)$.
6. $\int_a^b h(x)\,\mathrm{d}g(x) = \sum_{i=1}^k \int_{c_{i-1}}^{c_i} h(x)\,\mathrm{d}g(x)$ für $a=c_0<c_1<\ldots<c_n=b$.
7. Existiert $\int_a^b h(x)\,\mathrm{d}g(x)$, so existiert auch $\int_a^b g(x)\,\mathrm{d}h(x)$.
8. Es gilt folgende Verallgemeinerung der Formel der partiellen Integration:
$$\int_a^b h(x)\,\mathrm{d}g(x) + \int_a^b g(x)\,\mathrm{d}h(x) = [h(x)g(x)]_a^b.$$
Dabei folgt aus der Existenz des einen Integrals die Existenz des anderen.

9. Existiert $\int_a^b |h(x)|\,\mathrm{d}g(x)$, so existiert auch $\int_a^b h(x)\,\mathrm{d}g(x)$.

Das Riemann-Stieltjes-Integral in der Ebene und in höherdimensionalen Räumen erhalten wir ganz entsprechend zum eindimensionalen Fall, indem den üblicherweise betrachteten Flächen bzw. mehrdimensionalen Würfeln mittels einer integrierenden Funktion ein verallgemeinertes Volumen zugeordnet wird. Für den Fall einer zweidimensionalen Funktion haben wir zunächst die Riemann-Stieltjes-Summen

$$\sum_{i=1}^{n}\sum_{j=1}^{m} h(\xi_i,\eta_j)(g(x_i,y_j)-g(x_{i-1},y_j)-g(x_i,y_{j-1})+g(x_{i-1},y_{j-1}))$$
$$\text{mit}\quad x_{i-1}\le\xi_i\le x_i,\quad y_{j-1}\le\eta_j\le y_j,$$

wobei der Einfachheit halber das Integral über $A=(a_1,b_1]\times(a_2,b_2]$ bestimmt werden soll. Es ist dann der Grenzwert dieser Summen, wenn alle Differenzen x_i-x_{i-1}, y_j-y_{j-1} gegen null gehen. Wir schreiben dann für das Integral

$$\int_A h(x,y)\,\mathrm{d}g(x,y)=\int_{a_1}^{b_1}\int_{a_2}^{b_2} h(x,y)\,\mathrm{d}g(x,y).$$

Das Integral existiert wie oben unter geeigneten Stetigkeits- und Beschränktheitsbedingungen.

Die kurzen Ausführungen machen deutlich, dass die Theorie des Riemann-Stieltjes-Integrals weit ausgereift ist und für die Praxis relevante Ergebnisse bereitstellt. Da wir nur diskrete bzw. stetige Verteilungen betrachten, die integrierende Funktion als Verteilungsfunktionen stetig oder Treppenfunktionen sind, ist es in unserem Rahmen möglich, Riemann-Stieltjes-Integrale einfach als Abkürzungen zu lesen. Ist $\boldsymbol{X}=(X_1,\ldots,X_k)$ eine mehrdimensionale Zufallsvariable mit Verteilungsfunktion $F(\boldsymbol{x})$ und mit der Dichte $f(\boldsymbol{x})$, so gilt:

$$\int_A h(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x})=\begin{cases}\sum\cdots\sum_{x\in A} h(\boldsymbol{x})f(\boldsymbol{x}) & \text{falls } \boldsymbol{X} \text{ diskret}\\ \int_A\cdots\int h(\boldsymbol{x})f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} & \text{falls } \boldsymbol{X} \text{ stetig}\end{cases}.$$

Anhang B Lösungen zu den Aufgaben

Aufgaben aus Kapitel 2

Aufgabe 1

(i) Exponentialverteilung $f(x)=\lambda\exp(-\lambda x)\quad x>0$

$$M(t)=\int_0^\infty e^{xt}\lambda e^{-\lambda x}\mathrm{d}x=\int_0^\infty \lambda e^{-(\lambda-t)x}\mathrm{d}x=\frac{\lambda}{\lambda-t}\int_0^\infty(\lambda-t)e^{-(\lambda-t)x}\mathrm{d}x$$
$$=\frac{\lambda}{\lambda-t}\qquad\text{für}\quad |t|<\lambda$$

1. Moment: $\mathrm{E}(X)=M'(0)=\lambda/(\lambda-t)^2|_{t=0}=1/\lambda$
2. Moment: $\mathrm{E}(X^2)=M''(0)=2\lambda/(\lambda-t)^3|_{t=0}=2/\lambda^2$
3. Moment: $\mathrm{E}(X^3)=M'''(0)=6\lambda/(\lambda-t)^4|_{t=0}=6/\lambda^3$
4. Moment: $\mathrm{E}(X^4)=M^{(4)}(0)=24\lambda/(\lambda-t)^5|_{t=0}=24/\lambda^4$.

(ii) Gleichverteilung $f(x)=1/\vartheta\quad 0<x<\vartheta$

$$M(t)=\int_0^\vartheta e^{xt}\frac{1}{\vartheta}\,\mathrm{d}x=\frac{1}{t\vartheta}\left(e^{\vartheta t}-1\right)=\frac{1}{t\vartheta}\sum_{n=1}^\infty\frac{(\vartheta t)^n}{n!}=\sum_{n=1}^\infty\frac{(\vartheta t)^{(n-1)}}{n!}=\sum_{n=0}^\infty\frac{\vartheta^n}{n+1}\cdot\frac{t^n}{n!}$$
$$\Longrightarrow \mathrm{E}(X^r)=\frac{\vartheta^r}{r+1}.$$

(iii) geometrische Verteilung $f(x)=p(1-p)^x\quad x=0,1,2,3,\ldots$

$$M(t)=\sum_{x=0}^\infty e^{tx}p(1-p)^x=p\sum_{x=0}^\infty\left(e^t(1-p)\right)^x=\frac{p}{1-e^t(1-p)}\quad\text{für}\quad t<-\ln(1-p)$$

$$\mathrm{E}(X^r)=\sum_{x=0}^\infty x^r p(1-p)^x$$
$$\Rightarrow\quad\frac{\partial}{\partial p}\mathrm{E}(X^r)=\sum_{x=0}^\infty x^r[(1-p)^x+px(1-p)^{(x-1)}]$$
$$=\frac{1}{p}\sum_{x=0}^\infty x^r p(1-p)^x-\frac{1}{1-p}x^{r+1}p(1-p)^x=\frac{1}{p}\mathrm{E}(X^r)-\frac{1}{1-p}\mathrm{E}(X^{r+1})$$

Damit ist

$$\mathrm{E}(X^{r+1}) = \frac{1-p}{p}\mathrm{E}(X^r) - (1-p)\frac{\partial}{\partial p}\mathrm{E}(X^r)$$

und es folgt

$$\mathrm{E}(X) = \frac{1-p}{p}$$

$$\mathrm{E}(X^2) = \left(\frac{1-p}{p}\right)^2 - (1-p)\frac{-p-(1-p)}{p^2} = \frac{2-3p+p^2}{p^2}$$

$$\dots$$

(iv) Simpson-Verteilung $f(x) = 1-|x|$ für $-1 < x < 1$

$$M(t) = \int_{-1}^{1} e^{tx} f(x)\, d\,\mathrm{d}x = \int_{-1}^{0} e^{tx}(1+x)\, d\, x + \int_{0}^{1} e^{tx}(1-x)\,\mathrm{d}x = \frac{e^t + e^{-t} - 2}{t^2}$$

$$= \frac{1}{t^2}\left(\sum_{t=0}^{\infty}\frac{t^n}{n!} + \sum_{t=0}^{\infty}\frac{(-t)^n}{n^!} - 2\right) = \sum_{n=2}^{\infty}\frac{t^{n-2}+(-t)^{n-2}}{n!} = \sum_{n=0}^{\infty}\frac{1+(-1)^n}{(n+2)(n+1)}\,\frac{t^n}{n!}$$

also $\mathrm{E}(X^r) = \dfrac{1+(-1)^r}{(r+2)(r+1)}$.

Aufgabe 2

(i)

$$\mathrm{P}(|X-\mu| \geq k\sigma) = \mathrm{P}(|X| \geq ka) = \begin{cases} 1 & \text{falls } k \leq 0 \\ a^2 & \text{falls } 0 < k \leq a^{-1} \\ 0 & \text{falls } k > a^{-1} \end{cases}$$

Die Tschebyschev-Ungleichung sagt (wegen $\sigma = a$): $\mathrm{P}(|X-\mu| \geq ka) \leq 1/k^2$.

Wir setzen $k = a^{-1}$ und sehen, dass für diese Wahl die Abschätzung exakt ist. Da mit $0 < a \leq 1 \quad k = a^{-1}$ Werte aus dem ganzen Intervall $[1, \infty)$ annehmen kann, ist die Ungleichung nicht zu verschärfen.

(ii) Für $n = 1$ folgt die Behauptung schon aus Teil (i).

Für die anderen Fälle betrachten wir den Beweis der Tschebyschev-Ungleichung. Zur Vereinfachung setzen wir $(\bar{X}-\mu)^2 = Y$. Dann gilt: $Y \geq 0$.

$$\mathrm{E}(Y) = \sum y_i \mathrm{P}(Y = y_i) = \sum_{y_i < a} y_i \mathrm{P}(Y = y_i) + \sum_{y_i \geq a} y_i \mathrm{P}(Y = y_i) \geq \sum_{y_i \geq a} a\,\mathrm{P}(Y = y_i)$$

$$= a\,\mathrm{P}(Y \geq a).$$

Also kann Gleichheit nur gelten, wenn

$$(1) \quad \sum_{y_i < a} y_i \mathrm{P}(Y = y_i) = 0 \qquad \text{und} \qquad (2) \quad \mathrm{P}(Y > a) = 0.$$

Aus (1) folgt, dass nur $P(Y=0)>0$ sein darf und sonst $P(0<Y<a)=0$ gilt.

Aus (2) folgt, dass nur $P(Y=a)>0$ gelten kann.

Insgesamt darf also Y nur die Werte 0 und a annehmen, daraus ergeben sich 3 mögliche Werte für $\bar{X}_n$, nämlich μ, $\mu+a$ und $\mu-a$.

Da $\mathrm{Var}(X)=E(X-\mu)^2>0$ vorausgesetzt ist, muss auch $P(Y=a)>0$ gelten.

Wegen der positven Varianz nimmt X mindestens 2 verschiedene Werte an. Seien diese x_1 und x_2. Bei $n=3$ hat $\bar{X}$ dann schon folgende mögliche Werte: x_1, x_2, $(2x_1+x_2)/3$, $(x_1+2x_2)/3$.

Damit ist die Gleichheit im Fall $n\geq 3$ nicht mehr möglich.

Für den Fall $n=2$ gehen wir wieder von den möglichen Werten für $\bar{X}$ aus. Dies sind μ, $\mu+a$ und $\mu-a$. Damit sind für X die beiden Werte $\mu+a$ und $\mu-a$ möglich; sonst könnte $\bar{X}$ wieder mehr Werte annehmen.

Wegen der Symmetrie muss dann $P(X=\mu-a)=P(X=\mu+a)=1/2$ gelten.

Somit ist $\sigma=a$ und $P(\bar{X}=\mu-a)=P(\bar{X}=\mu+a)=1/4$, $P(\bar{X}=\mu)=1/2$.

$$P(|\bar{X}-\mu|\geq k\sigma)=P(|\bar{X}-\mu|\geq ka)=\begin{cases}1 & \text{falls } k\leq 0\\ 1/2 & \text{falls } 0<k\leq 1\\ 0 & \text{falls } k>1\end{cases}$$

$P(|\bar{X}-\mu|\geq k\sigma)=1/2k^2$ ist damit nur für $k=1$ möglich.

Aufgabe 3

1. Moment
$$E(X)=\int_k^\infty x\,\alpha k^\alpha x^{-(\alpha+1)}\mathrm{d}x=\alpha k^\alpha\int_k^\infty x^{-\alpha}\mathrm{d}x$$
$$=\alpha k^\alpha\frac{1}{(\alpha-1)k^{\alpha-1}}\int_k^\infty(\alpha-1)k^{(\alpha-1)}x^{-[(\alpha-1)+1]}\mathrm{d}x=\frac{\alpha}{\alpha-1}\,k\ \text{ für }\ \alpha>1$$

2. Moment
$$E(X^2)=\int_k^\infty x^2\,\alpha k^\alpha x^{-(\alpha+1)}\mathrm{d}x=\alpha k^\alpha\int_k^\infty x^{-\alpha+1}\mathrm{d}x$$
$$=\alpha k^\alpha\frac{1}{(\alpha-2)k^{\alpha-2}}\int_k^\infty(\alpha-2)k^{(\alpha-2)}x^{-[(\alpha-2)+1]}\mathrm{d}x=\frac{\alpha}{\alpha-2}\,k^2\ \text{ für }\ \alpha>2;$$

analog das n-te Moment:
$$E(X^n)=\int_k^\infty x^n\,\alpha k^\alpha x^{-(\alpha+1)}\mathrm{d}x=\alpha k^\alpha\int_k^\infty x^{-\alpha+n-1}\mathrm{d}x$$
$$=\alpha k^\alpha\frac{1}{(\alpha-n)k^{\alpha-n}}\int_k^\infty(\alpha-n)k^{(\alpha-n)}x^{-[(\alpha-n)+1]}\mathrm{d}x=\frac{\alpha}{\alpha-n}\,k^n\ \text{ für }\ \alpha>n\,.$$

Aufgabe 4

Sei $C(t)=\ln(M(t))$, dann folgt

$$C'(t)=\frac{1}{M(t)}M'(t) \qquad \Rightarrow \quad C'(0)=\frac{1}{M(0)}M'(0)=\mathrm{E}(X)/\sum_{k=0}^{\infty}\frac{\mathrm{E}(X^k)}{k!}0^k=\mathrm{E}(X).$$

$$\begin{aligned} C''(t)&=-\frac{M'(t)}{M(t)^2}M'(t)+\frac{1}{M(t)}M''(t)\\ \Rightarrow \quad C''(0)&=-\frac{M'(0)}{M(0)^2}M'(0)+\frac{1}{M(0)}M''(0)=-\frac{\mathrm{E}(X)}{1}\mathrm{E}(X)+M''(0)=-\mathrm{E}(X)^2+\mathrm{E}(X^2)\\ &=\mathrm{Var}(X). \end{aligned}$$

Aufgabe 5

$$\begin{aligned} X\sim\mathscr{P}(\lambda) \quad \Rightarrow \quad M^X(t)&=\sum_{x=0}^{\infty}e^{tx}e^{-\lambda}\frac{\lambda^x}{x!}=\sum_{x=0}^{\infty}e^{-\lambda}\frac{(\lambda e^t)^x}{x!}=e^{-\lambda+\lambda e^t}\sum_{x=0}^{\infty}e^{-(\lambda e^t)}\frac{(\lambda e^t)^x}{x!}\\ &=\exp(\lambda(e^t-1))\\ \Rightarrow \quad M^{X+Y}(t)&=M^X(t)\,M^Y(t)=\exp(\lambda(e^t-1)+\nu(e^t-1))=\exp((\lambda+\nu)(e^t-1)). \end{aligned}$$

Dies ist die momenterzeugende Funktion einer Poisson-Verteilung mit Parameter $\lambda+\nu$.

Aufgaben aus Kapitel 3

Aufgabe 1

Verteilungen bilden eine Exponentialfamilie, wenn die Dichte $f_\vartheta(x)$ in der Form

$$f_\vartheta(x)=\exp(c(\vartheta)T(x)+d(x)+e(\vartheta))\cdot h(x)$$

darstellbar ist.

(i) $f_\vartheta(x)=1/\vartheta=\exp(-\ln\vartheta)\cdot 1_{(0,\vartheta)}(x)$.
Da der Träger von ϑ abhängt, ist dies keine Exponentialfamilie.

(ii) $f_p(x)=p(1-p)^x=\exp(\underbrace{x\ln(1-p)}_{T(x)c(p)}+\underbrace{\ln p}_{e(p)})\cdot 1_{\mathbb{N}_0}(x)$; Exponentialfamilie mit $d(x)=0$.

(iii) $f(x)=1-|x|$. Diese Verteilung bildet überhaupt keine Familie.

(iv) $f_\alpha(x)=\alpha k^\alpha x^{-(\alpha+1)}=\exp(\underbrace{-(\alpha+1)\ln x}_{c(\alpha)T(x)}+\underbrace{\ln\alpha+\alpha\ln k}_{e(\alpha)})\cdot 1_{(k,\infty)}(x)$
bildet für festes k eine Exponentialfamilie in α mit $d(x)=0$.

Aufgaben aus Kapitel 4

Aufgabe 1

Setze

$$X = \begin{matrix} 1 & \text{falls Antwort ‚ja‘} \\ 0 & \text{falls Antwort ‚nein‘ ’} \end{matrix} \qquad \text{und} \qquad Y = \begin{matrix} 1 & \text{falls Frage nach Eigenschaft A} \\ 0 & \text{sonst.} \end{matrix}$$

Gegeben ist:

$\mathrm{P}(Y=1)=p$, $\mathrm{P}(Y=0)=1-p$, $\mathrm{P}(X=1|Y=1)=q$, $\mathrm{P}(X=1|Y=0)=1-q$, $\mathrm{P}(X=1)=r$.

(i)

$$\begin{aligned} r &= \mathrm{P}(X=1) = \mathrm{P}(X=1|Y=1)\mathrm{P}(X=1) + \mathrm{P}(X=1|Y=0)\mathrm{P}(X=1) = qp + (1-q)(1-p) \\ &= (2p-1)q + (1-p). \end{aligned}$$

(ii)

Es ist $\mathrm{P}(X=1)=r$, $\mathrm{P}(X=0)=1-r$ und $R = \frac{1}{n}\sum_{i=1}^{n} X_i$, wobei die X_i unabhängig sind (Ziehen mit Zurücklegen).

Also hat nR eine $\mathscr{B}(n,r)$–Verteilung. Damit folgt $\mathrm{E}(R)=r$ und $\mathrm{Var}(R)=r(1-r)/n$.

Weiter ist

$$\hat{Q} = \frac{1}{2p-1}R + \frac{p-1}{2p-1} \quad \text{mit} \quad \mathrm{E}(\hat{Q}) = \frac{1}{2p-1}\mathrm{E}(R) + \frac{p-1}{2p-1} = r.$$

Aufgabe 2

(i) Pareto-Verteilung mit bekanntem k

$$f_{\alpha,k}(x_1,\ldots,x_n) = \alpha^n k^{n\alpha} \left(\prod_{i=1}^{n} x_i\right)^{-\alpha+1} \qquad x_i > k \quad \Rightarrow \quad \prod_{i=1}^{n} X_i \quad \text{ist suffizient.}$$

(ii) Laplace-Verteilung: $f(x,\lambda) = 0.5\lambda \exp[-\lambda|x|]$

$$f_\lambda(x_1,\ldots,x_n) = \left(\frac{\lambda}{2}\right)^n \exp\left(-\lambda \sum_{i=1}^{n} |x_i|\right) \quad \Rightarrow \quad \sum_{i=1}^{n} |X_i| \quad \text{ist suffizient.}$$

(iii) Gamma-Verteilung: $f(x;\lambda,k) = \lambda^k x^{k-1} \exp[-\lambda x]/\Gamma(k)$ für $x>0$

$$f_{\lambda,k}(x_1,\ldots,x_n) = \left(\frac{\lambda^k}{\Gamma(k)}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{k-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

$$\Rightarrow \quad \left(\prod_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i\right) \quad \text{sind gemeinsam suffizient.}$$

(iv) verschobene Exponentialverteilung: $f(x,\vartheta,\lambda)=\lambda\exp[-\lambda(x-\vartheta)]$ für $x>\vartheta$

$$\begin{aligned} f_{\vartheta,\lambda}(x_1,\ldots,x_n) &= \lambda^n \exp\left(-\lambda\left(\sum_{i=1}^n -n\vartheta\right)\right) \quad \text{für} \quad x_i>\vartheta \\ &= \lambda^n \exp\left(-\lambda\left(\sum_{i=1}^n -n\vartheta\right)\right) 1_{[\min(x_1,\ldots,x_n)>\vartheta]} \\ \Rightarrow \quad & \left(\min(X_1,\ldots,X_n), \sum_{i=1}^n X_i\right) \quad \text{sind gemeinsam suffizient.} \end{aligned}$$

(v) zentrierte Gleichverteilung: $f(x)=1/2\vartheta$

$$\begin{aligned} f_\vartheta(x_1,\ldots,x_n) &= \left(\frac{1}{2\vartheta}\right)^n \quad \text{für} \quad -\vartheta<\min(x_1,\ldots,x_n)\le\max(x_1,\ldots,x_n)<\vartheta \\ \Rightarrow \quad & \left(X_{(1)},X_{(n)}\right) \quad \text{sind gemeinsam suffizient.} \end{aligned}$$

Aufgabe 3

Der Beweis wird durch Induktion geführt, und zwar beginnend bei $k=n$.

Für $X_{(n)}$ gilt

$$\mathrm{P}(X_{(n)}\le x)=\mathrm{P}(X_1\le x,\ldots,X_n\le x)=F(x)^n,$$

und somit

$$g_n(x)=nf(x)F(x)^{n-1}.$$

Die Behauptung gelte nun für ein beliebiges, festes $k<n$. Dann ist

$$\begin{aligned} \mathrm{P}(X_{(k-1)}\le x) &= \mathrm{P}(X_{(k)}\le x)+\mathrm{P}(X_{(k-1)}\le x<X_{(k)}) \\ &= F^{X_{(k)}}(x)+\binom{n}{k-1}F(x)^{k-1}(1-F(x))^{n-k}, \end{aligned}$$

da $\{X_{(k-1)}\le x<X_{(k)}\}$ gerade das Ereignis ist, dass genau $k-1$ der Stichprobenvariablen kleiner oder gleich x sind, die anderen $n-k+1$ größer. Es folgt

$$\begin{aligned} g_{(k-1)}(x) &= \frac{\partial}{\partial x}F^{X_{(k-1)}}(x) \\ &= g_{(k)}(x)+\binom{n}{k-1}\{(k-1)f(x)F(x)^{k-2}(1-F(x))^{n-k} \\ &\quad -(n+1-k)f(x)F(x)^{k-1}(1-F(x))^{n-k}\} \\ &= \frac{n!}{(k-2)!(n-(k-1))!}F(x)^{k-2}(1-F(x))^{n-(k-1)}. \end{aligned}$$

Aufgabe 4

Grundgesamtheit der Ereignisse: $(\omega_1,\ldots,\omega_n)\in\{0,1\}^n$
mit $\omega_i=0$ falls Kopf
und $\omega_i=1$ falls Zahl beim i-ten Wurf erscheint.

Interessierende Ereignisse: $A_i:=\{0,\ldots,0,1\}\times\{0,1\}^{n-i}$ für $i=1,\ldots,n$.

Sei X die Zufallsvariable ‚Gewinn', dann gilt

$$\mathrm{P}(X=2^{i-1})=\mathrm{P}(A_i)=\frac{2^{n-i}}{2^n}=\frac{1}{2^i}\quad\text{und}\quad\mathrm{P}(X=0)=\frac{1}{2^n}.$$

Für den Erwartungswert und die Varianz von X gelten:

$$\mathrm{E}(X)=0\cdot\frac{1}{2^n}+\sum_{i=1}^{n}2^{i-1}\,\mathrm{P}(X=2^{i-1})=\sum_{i=1}^{n}2^{i-1}\frac{1}{2^i}=\sum_{i=1}^{n}\frac{1}{2}=\frac{n}{2}$$

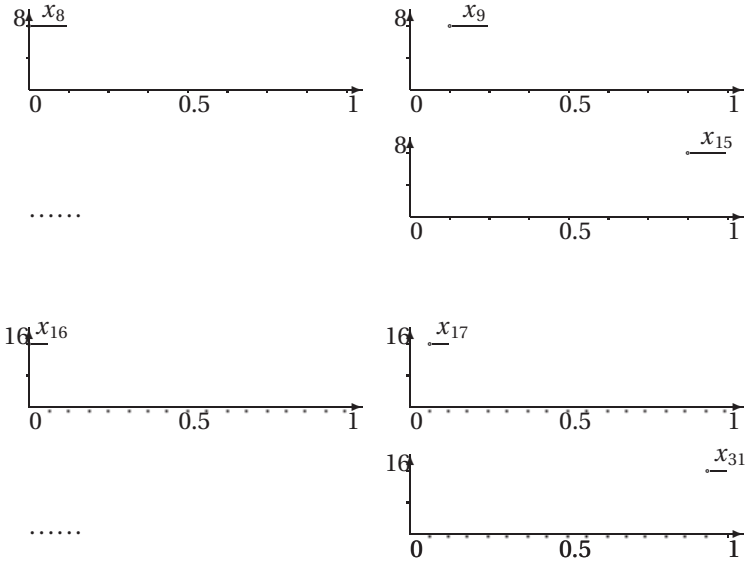
$$\mathrm{Var}(X)=\mathrm{E}(X^2)-E^2(X)=\sum_{i=1}^{n}2^{2i-2}\frac{1}{2^i}-\frac{2^n}{4}=\frac{1}{2}\sum_{i=1}^{n}2^{i-1}-\frac{2^n}{4}$$

$$=\frac{1}{2}\,\frac{1-2^n}{1-2}-\frac{2^n}{4}=2^{n-1}-\frac{1}{2}-\frac{2^n}{4}=\frac{2^{n+1}-n^2-2}{4}.$$

Aufgaben aus Kapitel 5

Aufgabe 1

i) Die Folge von Zufallsvariablen ist charakterisiert durch



Für kein $\omega \in (0,1)$ konvergiert die Folge $X_n(\omega)$: (Cauchy-Kriterium)

$$\bigwedge_{\epsilon>0} \bigvee_{n_0} \bigwedge_{n,m>n_0} |x_n - x_m| < \epsilon$$

Sei $\epsilon > 0$ und n_0 beliebig.bWeiter sei k_1 so, dass $2^{k_1} > n_0$ und m_1, m_2 so, dass $n_1 = 2^{k_1} + m_1$, $n_2 = 2^{k_1+l} + m_2$. Dann führt

$$\frac{2^{k_1}+m_1}{2^{k_1}} - 1 < \omega \leq \frac{2^{k_1}+m_1+1}{2^{k_1}} - 1 \quad \text{und} \quad \frac{2^{k_1+l}+m_2}{2^{k_1+l}} - 1 < \omega \leq \frac{2^{k_1+l}+m_2+1}{2^{k_1+l}} - 1$$

zu

$$|X_{n_1}(\omega) - X_{n_2}(\omega)| = 2^{k_1+l} - 2^{k_1} = 2^{k_1}(2^l - 1) > \epsilon$$

ii) Konvergenz im quadratischen Mittel gegen Null:

$$\mathrm{E}(X_n - 0)^2 = \mathrm{E}(X_n)^2 = 1/n^2 \qquad \text{also} \qquad \lim_{n\to\infty} \mathrm{E}(X_n - 0)^2 = \lim_{n\to\infty} 1/n^2 = 0$$

Konvergenz nicht mit Wahrscheinlichkeit eins:

Es ist zu zeigen, dass die Wahrscheinlichkeit für Divergenz größer null ist.

$$X_n \quad \text{divergiert} \iff \bigvee_m \bigwedge_n \bigvee_{k>n} X_k > 1/m$$

zu zeigen

$$\mathrm{P}(\bigcup_{n=1}^{\infty} A_m) > 0 \qquad \text{mit} \quad A_m := \bigcap_{n=1}^{\infty} \bigcup_{k\geq n} \{X_k > 1/m\}$$

Es reicht dabei zu zeigen $\mathrm{P}(A) > 0$ mit $A := \bigcap_{n=1}^{\infty} \bigcup_{k\geq n} \{X_k > 0\}$.
Es gilt $\mathrm{P}(X_k > 0) = 1/n$ und weiterhin

$$\mathrm{P}(A^C) = \mathrm{P}(\bigcap_{n=1}^{\infty} \bigcup_{k\geq n} \{X_k = 0\}) \leq \sum_{n=1}^{\infty} \mathrm{P}(\bigcup_{k\geq n} \{X_k = 0\}) = \sum_{n=1}^{\infty} \prod_{k\geq n} (1 - \frac{1}{n}) = 0\,.$$

Also ist $\mathrm{P}(A) = 1$.

Aufgaben aus Kapitel 6

Aufgabe 1

(i) 1. Vorschlag: Schichtung, d.h. 800 Seiten werden in 10 Schichten à 80 Seiten aufgeteilt. Mit Hilfe der Zufallszahlen < 80 wird dann eine Seite pro Schicht ausgewählt.
2. Vorschlag: Je drei Ziffern werden zu einer Seitenzahl zusammengefasst. Ungültige Ziffern werden dabei einfach übergangen.

(ii) Schätzung der durchschnittlichen Fehlerzahl mit Hilfe des arithmetischen Mittels:

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} \text{Fehleranzahl Seite } i \; = 1.2$$

Standardfehler dieses Schätzers:

$$s^2 = \frac{n}{n-1}(\overline{x^2} - \overline{x}^2) = \frac{10}{9}(2.8 - 1.44) = 1.511$$

Unterscheidung von Ziehen $\underline{mit}\,(f=0)$ und $\underline{ohne}\,(f=1)$ Zurücklegen:

$$\hat{s}^2_{\bar{x}} = \frac{1}{10}\,1.511\,\left(1 - f\frac{n-1}{N-1}\right) = 0.1511\,\left(1 - f\frac{9}{799}\right)$$

$$\rightarrow \hat{\sigma}_{\bar{x}} = 0.3865 \text{ ohne Zurücklegen}$$
$$\hat{\sigma}_{\bar{x}} = 0.3887 \text{ mit Zurücklegen}$$

(iii) Schätzung der Gesamtfehlerzahl mit zugehörigem Standardfehler:

$$y = 800\,\bar{x} = 960 \qquad \hat{\sigma}_y = \begin{cases} 309.2 & \text{ohne Z.} \\ 310.96 & \text{mit Z.} \end{cases}$$

(iv) Fehlerseitenanteil:

$$\hat{p} = \frac{\text{fehlerhafte Seiten}}{n} = \frac{6}{10} = 0.6$$

$$P(|\hat{p} - p| \le k) \ge 1 - \frac{p(1-p)}{n\,k^2} \ge 1 - \frac{1}{4\,n\,k^2} \qquad \text{da } \max_{0\le p\le 1} p(1-p) = 1/4$$

$$\text{Niveau} \Rightarrow \quad 0.75 = 1 - \frac{1}{4\,n\,k^2} \quad \Rightarrow n\,k^2 = 1 \quad \Rightarrow k = \sqrt{1/10} = 0.316$$

Konfidenzintervall:

$$\hat{p} - 0.316 \le p \le \hat{p} + 0.316 \quad \Rightarrow 0.284 \le p \le 0.916$$

Aufgabe 2

Aufgabe 3

Liste der möglichen Stichproben mit zugehörigen Mitteln:

r	x_1	x_2	x_3	x_4	$\bar{x}$	$\bar{x}^2$
1	0	6	18	26	12.50	156.25
2	1	8	19	30	14.50	210.25
3	1	9	20	31	15.25	232.5625
4	2	10	20	31	15.75	248.0625
5	5	13	24	33	18.75	351.5625
6	4	12	23	32	17.75	315.0625
7	7	15	25	35	20.50	420.25
8	7	16	28	37	22.00	484
9	8	16	29	38	22.75	517.5625
10	6	17	27	38	22.00	484

Daraus ergibt sich:

$$\mathrm{E}(\bar{X}) = 18.175 \qquad \mathrm{Var}(\bar{X}) = \mathrm{E}(\bar{X}^2) - E^2(\bar{X}) = 11.625625$$

Aufgabe 4

μ_{N-1} sei das Populationsmittel der Teilgesamtheit aus den verbliebenen $N-1$ Elementen und μ_N das gesamte Populationsmittel. Entsprechend seien σ^2_{N-1} und σ^2_N die Varianzen.

i)

$$\mathrm{E}(\tilde{\mu}) = \frac{1}{N}x_1 + \frac{N-1}{N}\mu_{N-1}\,, \qquad \mathrm{Var}(\tilde{\mu}) = \left(\frac{N-1}{N}\right)^2 \frac{\sigma^2_{N-1}}{n}.$$

ii)

$$\mathrm{E}(\bar{X}) = \mu_N, \qquad \mathrm{Var}(\bar{X}) = \frac{1}{n}\left(\frac{N-1}{N}\sigma^2_{N-1} + \frac{N-1}{N}(\mu_{N-1}-\mu_N)^2 + \frac{1}{N}(x_1-\mu_N)^2\right).$$

Der Vergleich der beiden Varianzen zeigt sofort, dass $\tilde{\mu}$ besser ist.

Aufgabe 5

Lemma $\bar{X}_n$ sei schwach konsistent für μ. Weiterhin sei $g(x)$ stetig. Dann gilt:

$$g(\bar{X}_n) \quad \text{ist schwach konsistent für } g(\mu)$$

Beweis

$$\text{Stetigkeit:} \quad \forall\epsilon > 0 \quad \exists\delta_\epsilon > 0 \quad \forall|x-\mu| < \delta \quad \Rightarrow \quad |g(x) - g(\mu)| < \epsilon$$

Aus

$$|\bar{X}_n - \mu| < \delta \quad \Rightarrow \quad |g(\bar{X}_n) - g(\mu)| < \epsilon$$

folgt

$$\Rightarrow \quad \mathrm{P}(|\bar{X}_n - \mu| < \delta) \;\le\; \mathrm{P}(|g(\bar{X}_n) - g(\mu)| < \epsilon) \;\le\; 1$$

$$\lim = 1 \qquad \Rightarrow \qquad \lim = 1$$

Im Fall einer Exponentialverteilung ist $\bar{X}$ schwach konsistent für $1/\lambda$, also folgt die Behauptung.

Aufgabe 6

i) $\mathrm{E}(X) = \mu, \quad \mathrm{Var}(X) = 2$

$$\frac{\mathrm{Var}(\bar{X})}{\mathrm{Var}(\tilde{X})} = \frac{2/n}{1/4n(0.5)^2} = 2 \qquad \Rightarrow \qquad \tilde{X} \quad \text{ist doppelt so effizient wie } \bar{X}$$

ii)

$$\frac{\mathrm{Var}(\bar{X})}{\mathrm{Var}(\tilde{X})} = \frac{\sigma^2/n}{1/4n(\frac{1}{\sqrt{2\pi}\sigma})^2} = 2/\pi = 0.637 \qquad \Rightarrow \qquad \bar{X} \quad \text{ist ca. 1.57mal so effizient wie } \tilde{X}$$

Aufgabe 7

Es ist $f_\vartheta(x,y) = y^2 e^{-(x+\vartheta)}$ $x, y \geq 0$.

$$\begin{aligned} f_\vartheta(x) &= \int_0^\infty f_\vartheta(x,y)\,\mathrm{d}y = \int_0^\infty y^2 e^{-(x+\vartheta)}\mathrm{d}y = \frac{1}{x+\vartheta}\int_0^\infty y^2 (x+\vartheta) e^{-(x+\vartheta)}\mathrm{d}y \\ &= \frac{1}{x+\vartheta}\mathrm{E}(X^2) \qquad \text{mit } X \sim \mathcal{EX}(x+\vartheta) \\ &= \frac{1}{x+\vartheta}\frac{2}{(x+\vartheta)^2} = \frac{2}{(x+\vartheta)^3} \end{aligned}$$

Dies führt zu:

$$f_\vartheta(y|x) = \frac{f_\vartheta(x,y)}{f_\vartheta(x)} = 0.5\, y^2 (x+\vartheta)^3 e^{-(x+\vartheta)} \qquad x, y \geq 0$$

$$\Longrightarrow \mathrm{E}(Y|X=x) = \int_0^\infty 0.5\, y^3 (x+\vartheta)^3 e^{-(x+\vartheta)}\mathrm{d}y$$

Mit der Substitution $u = y(x+\vartheta), dy = \frac{1}{x+\vartheta} du$ folgt weiter:

$$\mathrm{E}(Y|X=x) = \frac{0.5}{x+\vartheta}\int_0^\infty u^3 e^{-u}\,\mathrm{d}u = \frac{3}{x+\vartheta}$$

$$\begin{aligned} \mathrm{E}(Y^2|X=x) &= \int_0^\infty y^2 f_\vartheta(y|x)\,\mathrm{d}y = \int_0^\infty y^2 0.5 (x+\vartheta)^3 y^2 e^{-y(x+\vartheta)}\mathrm{d}y \\ &= \frac{0.5}{(x+\vartheta)^2}\int_0^\infty u^4 e^{-u}\,\mathrm{d}u = \frac{12}{(x+\vartheta)^2} \end{aligned}$$

$$\mathrm{Var}(Y|X=x) = \mathrm{E}(Y^2|X=x) - E^2(Y|X=x) = \frac{3}{(x+\vartheta)^2}.$$

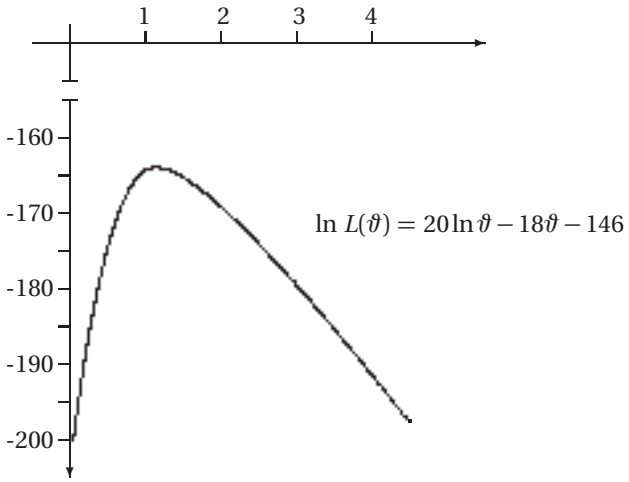
Ein adäquater Ansatz ist ein gewichteter KQ-Ansatz mit dem Zielkriterium

$$\sum_i (x_i+\vartheta)^2 (y_i - 3/(x_i+\vartheta))^2 \quad \rightarrow \quad \min!$$

Aufgabe 8

i)

$$\begin{aligned} \ln(L(\vartheta)) &= \ln\left(\prod_{i=1}^6 f_\vartheta(x_i)^{n_i}\right) = \ln(\vartheta^{20} 600^{20\vartheta} (750^8 \cdot 1500^5 \cdot 2500^4 \cdot 3500 \cdot 4500 \cdot 5500)^{-(\vartheta+1)} \\ &= 20\ln(\vartheta) + 20\vartheta(600) - (\vartheta+1)(8\ln(750) + \ldots + \ln(5500)) \\ &= 20\ln(\vartheta) - 18.069\vartheta - 146.008 \end{aligned}$$



Nullsetzen der Ableitung:

$$0 \stackrel{!}{=} \frac{\partial}{\partial \vartheta} \ln(L(\vartheta)) = \frac{20}{\vartheta} - 18.069 \Longrightarrow \quad \hat{\vartheta} = \frac{20}{18.069} = 1.107\,.$$

Wegen

$$\frac{\partial^2}{\partial \vartheta^2} \ln(L(\vartheta))\,|_{\vartheta=\hat{\vartheta}} = -\frac{n}{\vartheta^2}\,|_{\vartheta=\hat{\vartheta}} < 0$$

liegt ein Maximum vor.

Aufgabe 9

$$\begin{aligned} \mathrm{P}(\{N_1,N_2\}{=}(n_1,n_2)) &= \frac{n!}{n_1!\,n_2!(n-n_1-n_2)!}\,\vartheta^{2n_1}(2\vartheta(1-\vartheta))^{n_2}(1-\vartheta)^{2(n-n_1-n_2)} \\ &= \frac{n!\,2^{n-2}}{n_1!\,n_2!(n-n_1-n_2)!}\,\vartheta^{2n_1+n_2}(1-\vartheta)^{2n-2n_1-n_2} \end{aligned}$$

$$\ln p(n_1,n_2) = \ln\left(\frac{n!\,2^{n-2}}{n_1!\,n_2!(n-n_1-n_2)!}\right) d + (2n_1+n_2)\ln\vartheta + (2n-2n_1-n_2)\ln(1-\vartheta)$$

$$\frac{\partial}{\partial \vartheta} \ln p(n_1,n_2) = \frac{2n_1+n_2}{\vartheta} - \frac{2n-2n_1 n_2}{1-\vartheta} \stackrel{!}{=} 0$$

$$\Rightarrow \quad \frac{2n_1+n_2}{\hat{\vartheta}} = \frac{2n-2n_1-n_2}{1-\hat{\vartheta}} \Rightarrow \quad \hat{\vartheta} = \frac{2n_1+n_2}{2n} = \frac{n_1+n_2/2}{n}$$

Erwartungstreue:

$$\mathrm{E}(\hat{\vartheta}) = E\left(\frac{N_1}{n}\right) + \frac{1}{2}\mathrm{E}\left(\frac{N_2}{n}\right) = \vartheta^2 + \frac{1}{2}2\vartheta(1-\vartheta) = \vartheta$$

Konsistenz:

$$\left.\begin{array}{l} \frac{N_1}{n} \longrightarrow \vartheta^2 \\ \frac{N_2}{n} \longrightarrow 2\vartheta(1-\vartheta) \end{array}\right\} \quad \text{Gesetz der großen Zahlen !}$$

$$\begin{aligned} \mathrm{Var}(\hat{\vartheta}) &= E\left[\left(\frac{N_1}{n} + \frac{1}{2}\cdot\frac{N_2}{n} - \vartheta^2 - \vartheta(1-\vartheta)\right)^2\right] \\ &= \mathrm{Var}\left(\frac{N_1}{n}\right) + \mathrm{Var}\left(\frac{N_2}{2n}\right) + 2\,\mathrm{Cov}\left(\frac{N_1}{n}, \frac{N_2}{2n}\right) \\ &\leq \mathrm{Var}\left(\frac{N_1}{n}\right) + \mathrm{Var}\left(\frac{N_2}{2n}\right) + \sqrt{\mathrm{Var}\left(\frac{N_1}{n}\right) + \mathrm{Var}\left(\frac{N_2}{n}\right)} \quad \longrightarrow \quad 0 \end{aligned}$$

Effizienz:

$$\frac{\partial^2}{\partial\vartheta^2}\ln p(n_1, n_2) = -\frac{2n_1 + n_2}{\vartheta^2} - \frac{2n - 2n_1 - n_2}{(1-\vartheta)^2}$$

Damit folgt:

$$\begin{aligned} -\mathrm{E}\left(\frac{\partial^2}{\partial\vartheta^2}\ln p(N_1, N_2)\right) &= \mathrm{E}\left(\frac{2N_1 + N_2}{\vartheta^2} + \frac{2n - 2N_1 - N_2}{(1-\vartheta)^2}\right) \\ &= \frac{1}{\vartheta^2}(2n\vartheta^2 + 2n\vartheta(1-\vartheta)) + \frac{2n}{(1-\vartheta)^2} - \frac{2n\vartheta^2}{(1-\vartheta)^2} - \frac{2\vartheta(1-\vartheta)}{(1-\vartheta)^2} \\ &= 2n + 2n\frac{1-\vartheta}{\vartheta} + \frac{2n(1-\vartheta}{(1-\vartheta)^2} - \frac{2n\vartheta}{1-\vartheta} = 2n\left(\frac{1}{\vartheta} + \frac{1}{1-\vartheta}\right) = \frac{2n}{\vartheta(1-\vartheta)} \end{aligned}$$

Also ist

$$\frac{1}{-\mathrm{E}\left(\frac{\partial^2}{\partial\vartheta^2}\ln p(N_1, N_2)\right)} = \frac{\vartheta(1-\vartheta)}{2n}.$$

$$\begin{aligned} \mathrm{Var}(\hat{\vartheta}) &= \frac{1}{n^2}\mathrm{Var}\left(N_1 + \frac{1}{2}N_2\right) = \frac{1}{n^2}\left(\mathrm{Var}(N_1) + \frac{1}{4}\mathrm{Var}(N_2) + \frac{2}{2}\mathrm{Cov}(N_1, N_2)\right) \\ &= \frac{1}{n^2}\left(n\vartheta^2(1-\vartheta^2) + \frac{1}{4}n2\vartheta(1-\vartheta)(1-2\vartheta(1-\vartheta)) - n\vartheta^2 2\vartheta(1-\vartheta)\right) \\ &= \frac{1}{n}\vartheta\left(\vartheta - \vartheta^3 + \frac{1}{2}(1-\vartheta) - \vartheta(1-\vartheta)^2 - 2\vartheta^2(1-\vartheta)\right) \\ &= \frac{1}{n}\vartheta\left(\vartheta + \frac{1}{2} - \frac{3}{2}\vartheta\right) = \frac{1}{2n}\vartheta(1-\vartheta). \end{aligned}$$

Die Varianz von $\hat{\vartheta}$ ist gleich der Rao-Cramér-Schranke, $\hat{\vartheta}$ ist effizient.

Aufgabe 10

$$\ln L(\mu) = 2\ln\pi - \ln\left(1+\left(\frac{1}{2}-\mu\right)^2\right) - \ln(1+(2+\mu)^2)$$

$$\frac{\partial \ln L}{\partial \mu} = 2\left(\frac{\frac{1}{2}-\mu}{1+\left(\frac{1}{2}-\mu\right)^2} - \frac{2+\mu}{1+(2+\mu)^2}\right)$$

$$\begin{aligned}
0 &\stackrel{!}{=} \left(\frac{1}{2}-\mu\right) + \left(\frac{1}{2}-\mu\right)(2+\mu)^2 - (2+\mu) - (2+\mu)\left(\frac{1}{2}-\mu\right)^2 \\
&= -\frac{3}{2} - 2\mu + \frac{1}{2}(4+4\mu+\mu^2) - \mu(4+4\mu+\mu^2) - 2\left(\frac{1}{4}-\mu+\mu^2\right) - \mu\left(\frac{1}{4}-\mu+\mu^2\right) \\
&= -\frac{9}{4}\mu - \frac{9}{2}\mu^2 - 2\mu^3
\end{aligned}$$

$$\Rightarrow \quad 0 = \mu^3 + \frac{9}{4}\mu^2 + \frac{9}{4}\mu$$

$$\mu_1 = 0$$

$$\mu_{2,3} = -\frac{9}{8} \pm \sqrt{\left(\frac{9}{8}\right)^2 - \frac{9}{8}} = \frac{-9\pm 3}{8} \quad \Rightarrow \quad \mu_2 = -\frac{3}{4} \qquad \mu_3 = -\frac{3}{2}$$

Aufgabe 11

$$L(\vartheta) = \left(\frac{1}{3}\right)^n (5-2\vartheta)^{\sum(3-x_i)(2-x_i)/2} (\vartheta-1)^{\sum(x_i-1)(4-x_i)/2}$$

$$\begin{aligned}
\ln L(\vartheta) = {} & -n\ln 3 + \frac{1}{2}\sum_{i=1}^{n}(6-5x_i+x_i^2)\ln(5-2\vartheta) + \frac{1}{2}\sum_{i=1}^{n}(6-5x_i+x_i^2)\ln(5-2\vartheta) \\
& + \frac{1}{2}\sum_{i=1}^{n}(-4+5x_i-x_i^2)\ln(\vartheta-1)
\end{aligned}$$

$$\frac{\partial \ln L(\vartheta)}{\partial \vartheta} = \frac{-2}{5-2\vartheta}\,\frac{1}{2}\sum_{i=1}^{n}(6-5x_i+x_i^2) + \frac{1}{\vartheta-1}\,\frac{1}{2}\sum_{i=1}^{n}(-4+5x_i-x_i^2) \stackrel{!}{=} 0$$

Umformen führt zu:

$$\begin{aligned}
&\frac{1}{5-2\hat{\vartheta}}(6n-5n\bar{x}+n\bar{x^2}) = \frac{1}{\hat{\vartheta}-1}\,\frac{1}{2}\sum_{i=1}^{n}(-4n+5n\bar{x}-n\bar{x^2}) \\
&\Rightarrow \quad \hat{\vartheta}(6-5\bar{x}+\bar{x^2})-(6-5\bar{x}+\bar{x^2}) = -\hat{\vartheta}(-4+5\bar{x}-\bar{x^2})+\frac{5}{2}(-4+5\bar{x}-\bar{x^2}) \\
&\Rightarrow \quad 2\hat{\vartheta} = 6-5\bar{x}+\bar{x^2}-10+\frac{25}{2}\bar{x}-\frac{5}{2}\bar{x^2} = -4+\frac{15}{2}\bar{x}-\frac{3}{2}\bar{x^2} \\
&\Rightarrow \quad \hat{\vartheta} = -2+\frac{15}{4}\bar{x}-\frac{3}{4}\bar{x^2}.
\end{aligned}$$

Es ist

$$\mathrm{E}(X^k) = \frac{1^k}{3}(5-2\vartheta)+\frac{2^k}{3}(\vartheta-1)+\frac{3^k}{3}(\vartheta-1).$$

Das ergibt:

$$\mathrm{E}(X) = \vartheta, \quad \mathrm{E}(X^2) = -\frac{8}{3}+\frac{11}{3}\vartheta, \quad \mathrm{E}(X^3) = -\frac{30}{3}+\frac{33}{3}\vartheta, \quad \mathrm{E}(X^4) = -\frac{92}{3}+\frac{95}{3}\vartheta \quad \cdots$$

$$\begin{aligned}
\mathrm{E}(\hat{\vartheta}) &= \mathrm{E}\left(-2+\frac{15}{4}\bar{X}-\frac{3}{4}\bar{X^2}\right) = -2+\frac{15}{4}\mathrm{E}(X)-\frac{3}{4}\mathrm{E}(X^2) = -2+\frac{15}{4}\vartheta-\frac{3}{4}\left(-\frac{8}{3}+\frac{11}{3}\vartheta\right) \\
&= \vartheta.
\end{aligned}$$

$\hat{\vartheta}$ ist auch konsistent, da $\bar{X} \to \mathrm{E}(X)$, $\bar{X^2} \to \mathrm{E}(X^2)$.

Effizienz:

$$\begin{aligned}
\mathrm{Var}(\hat{\vartheta}) &= \mathrm{Var}\left(-2+\frac{15}{4}\bar{X}-\frac{3}{4}\bar{X^2}\right) \\
&= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{15}{4}X_i-\frac{3}{4}X_i^2\right)\right) \\
&= \frac{1}{n}\mathrm{Var}\left(\frac{15}{4}X-\frac{3}{4}X^2\right) \\
&= \frac{1}{n}E\left(\left(\frac{15}{4}X-\frac{3}{4}X^2\right)^2\right)-\frac{1}{n}\underbrace{\left(E\left(\frac{15}{4}X-\frac{3}{4}X^2\right)\right)^2}_{=(\vartheta+2)^2}
\end{aligned}$$

$$
\begin{aligned}
E\left(\left(\frac{15}{4}X-\frac{3}{4}X^2\right)^2\right) &= E\left(\frac{225}{16}X^2-2\frac{45}{16}X^3+\frac{9}{16}X^4\right)\\
&= \frac{225}{16}\left(-\frac{8}{3}+\frac{11}{3}\vartheta\right)-\frac{90}{16}\left(-\frac{30}{0}+\frac{33}{3}\vartheta\right)+\frac{9}{16}\left(-\frac{92}{3}+\frac{95}{3}\vartheta\right)\\
&= \frac{1}{48}[(-1800+2700-828)+(2475-2970+855)\vartheta]\\
&= \frac{3}{2}+\frac{15}{2}\vartheta
\end{aligned}
$$

$$
\mathrm{Var}(\hat{\vartheta})=\frac{1}{n}\left(\frac{3}{2}+\frac{15}{2}\vartheta-(\vartheta+2)^2\right)=\frac{1}{n}\left(-\frac{5}{2}+\frac{7}{2}\vartheta-\vartheta^2\right)=\frac{1}{2n}(-5+7\vartheta-2\vartheta^2);
$$

andererseits gilt

$$
\begin{aligned}
\ln f_\vartheta(x) &= -\ln 3+\frac{1}{2}(3-x)(2-x)\ln(5-2\vartheta)+\frac{1}{2}(x-1)(4-x)\ln(\vartheta-1)\\
\frac{\partial}{\partial\vartheta}\ln f_\vartheta(x) &= \frac{-2}{2(5-2\vartheta)}(6-5x+x^2)+\frac{1}{2(\vartheta-1)}(-4+5x-x^2)\\
\frac{\partial^2}{\partial\vartheta^2}\ln f_\vartheta(x) &= \frac{-4}{2(5-\vartheta)^2}(6-5x+x^2)-\frac{1}{2(\vartheta-1)^2}(-4+5x-x^2)
\end{aligned}
$$

$$
\begin{aligned}
-\mathrm{E}\left(\frac{\partial^2}{\partial\vartheta^2}\ln f_\vartheta(X)\right) &= \frac{2}{(5-2\vartheta)^2}\left(6-5\vartheta-\frac{8}{3}+\frac{11}{3}\vartheta\right)+\frac{1}{2(\vartheta-1)^2}\left(5\vartheta-4+\frac{8}{3}-\frac{11}{3}\vartheta\right)\\
&= \frac{2}{(5-2\vartheta)^2}\left(\frac{10}{3}-\frac{4}{3}\vartheta\right)+\frac{1}{2(\vartheta-1)^2}\left(\frac{4}{3}\vartheta-\frac{4}{3}\right)\\
&= \frac{4}{3(5-2\vartheta)}+\frac{2}{3(\vartheta-1)}=\frac{2}{(5-2\vartheta)(\vartheta-1)}
\end{aligned}
$$

Damit ist

$$
\frac{1}{-n\,\mathrm{E}\left(\frac{\partial^2}{\partial\vartheta^2}\ln f_\vartheta(X)\right)}=\frac{1}{2n}(5-2\vartheta)(\vartheta-1)=\frac{1}{2n}(-5+7\vartheta-2\vartheta^2)
$$

$\hat{\vartheta}$ ist also effizient.

Aufgabe 12

Setze

$$
X=\begin{matrix}1 & \text{falls Antwort ‚ja'}\\ 0 & \text{falls Antwort ‚nein' ,}\end{matrix}\qquad\text{und}\qquad Y=\begin{matrix}1 & \text{falls Frage nach Eigenschaft A}\\ 0 & \text{sonst}\end{matrix}
$$

Gegeben ist:

$\mathrm{P}(Y=1)=p$, $\mathrm{P}(Y=0)=1-p$, $\mathrm{P}(X=1|Y=1)=q$, $\mathrm{P}(X=1|Y=0)=1-q$, $\mathrm{P}(X=1)=r$.

(i)
$$\begin{aligned} r &= \mathrm{P}(X=1) = \mathrm{P}(X=1|Y=1)\mathrm{P}(X=1) + \mathrm{P}(X=1|Y=0)\mathrm{P}(X=1) \\ &= qp + (1-q)(1-p) = (2p-1)q + (1-p). \end{aligned}$$

(ii) Es ist $\mathrm{P}(X=1)=r$, $\mathrm{P}(X=0)=1-r$ und $R=\frac{1}{n}\sum_{i=1}^{n} X_i$, wobei die X_i unabhängig sind (Ziehen mit Zurücklegen).

Also hat nR eine $\mathscr{B}(n,r)$–Verteilung.Damit folgt $\mathrm{E}(R)=r$ und $\mathrm{Var}(R)=r(1-r)/n$.

Weiter ist
$$\hat{Q} = \frac{1}{2p-1}R + \frac{p-1}{2p-1} \quad \text{mit} \quad \mathrm{E}(\hat{Q}) = \frac{1}{2p-1}\mathrm{E}(R) + \frac{p-1}{2p-1} = r.$$

Aufgabe 13

Schätzung für den Anteil der Personen, die Cholesterin für gefährlich halten

...ohne Schichtung
$$\tilde{p} = \frac{1}{507}(165+149) = 0.61933$$

...mit Schichtung (Anteil der Männer 48 %)
$$\hat{p} = \frac{165}{250}0.52 + \frac{149}{257}0.48 = 0.62149$$

Vergleich der Varianzen:
$$\widehat{\mathrm{Var}}(\tilde{p}) = \frac{0.61933(1-0.61933)}{507} = 0.0004650$$
$$\widehat{\mathrm{Var}}(\hat{p}) = 0.52^2\,\frac{\frac{165}{250}(1-\frac{165}{250})}{250} = 0.48^2\,\frac{\frac{149}{257}(1-\frac{149}{257})}{257} = 0.0004611$$

Aufgabe 14

1. Schätzung: $\bar{X} = \frac{1}{200}\sum x_i = 100$
2. Schätzung: $\frac{\bar{X}}{\bar{Y}}\mu_Y = \frac{100}{10}\cdot\frac{80000}{10000} = 80$

Vergleich der Varianzen:
$$\widehat{\mathrm{Var}}(\bar{X}) = \frac{1}{200}\left(\frac{1}{200}4\,960\,000 - 100^2\right)\left(1-\frac{n}{N}\right) = \frac{1}{200}14\,800\cdot\frac{200}{10\,000} = 72.52$$

$$\begin{aligned} &\widehat{\mathrm{Var}}\left(\frac{\bar{X}}{\bar{Y}}\mu_Y\right) \\ &= \frac{1}{n}\left(1-\frac{n}{N}\right)\left(14\,800 - 2\frac{100}{10}\left(\frac{490\,000}{200} - 100\cdot 10\right)\right) + \left(\frac{100}{10}\right)^2\left(\frac{50\,000}{200} - 10^2\right) \\ &= \frac{1}{200}\left(1-\frac{200}{10\,000}\right)(14\,800 - 20\cdot 1\,450 + 100\cdot 150) = \frac{1}{200}0.98\cdot 800 = 3.92 \end{aligned}$$

Aufgabe 15

Sei $\hat{\mu} = a_1 X_1 + \cdots + a_n X_n$. Dann gilt

$$\mathrm{E}(\hat{\mu}) = \sum_{i=1}^{n} a_i \mathrm{E}(X_i) = \mu \sum_{i=1}^{n} a_i .$$

Damit folgt

$$\hat{\mu} \quad \text{ist erwartungstreu} \qquad \Longleftrightarrow \qquad \sum_{i=1}^{n} a_i = 1 .$$

Weiter ist

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) = \sigma^2 \sum_{i=1}^{n} a_i^2 .$$

Damit die Varianz minimal ist, muss also gelten $\sum_{i=1}^{n} a_i^2 \Rightarrow \min!$

Sei nun $h(a) := \sum_{i=1}^{n} a_i^2$ und $f(a) := \sum_{i=1}^{n} a_i - 1$ mit $a \in \mathbb{R}^n$, d. h. minimiere $h(a)$ unter der Nebenbedingung $f(a) = 0$!

Sei nun a^* der gesuchte Punkt, dann gilt

$$\begin{aligned} &\operatorname{grad} h(a^*) = \lambda \operatorname{grad} f(a^*) && \lambda \in \mathbb{R} \\ \rightarrow \quad & 2a^* = \lambda(1, \ldots, 1) \\ \rightarrow \quad & a_i^* = \lambda/2 && \text{für } i = 1, \ldots, n \end{aligned}$$

Aus $\sum_{i=1}^{n} a_i^* = \sum_{i=1}^{n} \lambda/2 = 1$ folgt $\lambda = 2/n$, also $a_i^* = 1/n \quad i = 1, \ldots, n$.

$\mathrm{Var}(\hat{\mu})$ ist also minimal für

$$\hat{\mu} = \frac{1}{n} X_1 + \ldots + \frac{1}{n} X_n = \bar{X}_n .$$

Aufgabe 16

i)

x	p_x	$x\,p_x$	$x^2 p_x$
0	$0.1\,\vartheta$	0	0
1	$0.2\,\vartheta$	$0.2\,\vartheta$	$0.2\,\vartheta$
2	$0.3\,\vartheta$	$0.6\,\vartheta$	$1.2\,\vartheta$
3	$0.4\,\vartheta$	$1.2\,\vartheta$	$3.6\,\vartheta$
$\sum$	ϑ	$4 - 2\,\vartheta$	$16 - 11\,\vartheta$

$$\begin{aligned} \Rightarrow \quad \mathrm{E}(X) &= 4 - 2\,\vartheta \\ \bar{X} &= 4 - 2\,\vartheta \\ \hat{\vartheta}_1 &= \frac{4 - \bar{X}}{2} = 2 - 1/2\bar{X} \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X) &= (16 - 11\,\vartheta) - (4 - 2\,\vartheta)^2 \\ &= 16 - 11\,\vartheta - 16 + 16\,\vartheta - 4\,\vartheta^2 = 5\,\vartheta - 4\,\vartheta^2 \end{aligned}$$

Konsistenz:

$$\begin{aligned}\mathrm{E}((\hat{\vartheta}_1-\vartheta)^2) &= \mathrm{Var}(\hat{\vartheta}_1) = \mathrm{Var}(2-1/2\bar{X}) = 1/4\,\mathrm{Var}(\bar{X}) \\ &= \frac{1}{4}\cdot\frac{1}{n}\mathrm{Var}(X) = \frac{1}{4}\cdot\frac{1}{n}(5\vartheta-4\vartheta^2) = \frac{1}{n}\vartheta(1.25-\vartheta) \\ \Rightarrow \qquad & \lim_{n\to\infty}\mathrm{E}((\hat{\vartheta}_1-\vartheta)^2) = 0\end{aligned}$$

ii) Sei $Y_1 := \#\{X=0\}$ mit $Y_1 \sim \mathcal{B}(n, 0.1\,\vartheta)$, dann gilt $\hat{\vartheta}_2 = 10\,Y_1/n$.

$$\Rightarrow \quad \mathrm{E}(\hat{\vartheta}_2) = 10/n\,\mathrm{E}(Y_1) = \frac{10}{n}(n\cdot 0.1\,\vartheta) = \vartheta$$

$$\mathrm{Var}(\hat{\vartheta}_2) = \frac{100}{n^2}\mathrm{Var}(Y_1) = \frac{100}{n}0.1\vartheta(1-0.1\,\vartheta) = \frac{1}{n}\vartheta(10-\vartheta).$$

Analog sind $\hat{\vartheta}_3$ und $\hat{\vartheta}_4$ erwartungstreu und

$$\mathrm{Var}(\hat{\vartheta}_3) = \frac{2.5^2}{n}0.4\,\vartheta(1-0.4\,\vartheta) = \frac{1}{n}\vartheta(2.5-\vartheta) \quad \text{sowie} \quad \mathrm{Var}(\hat{\vartheta}_4) = \frac{1}{n}\vartheta(1-\vartheta).$$

Relative Effizienzen:

$$\frac{\mathrm{Var}(\hat{\vartheta}_2)}{\mathrm{Var}(\hat{\vartheta}_4)} = \frac{10-\vartheta}{1-\vartheta} > 1 \qquad \frac{\mathrm{Var}(\hat{\vartheta}_3)}{\mathrm{Var}(\hat{\vartheta}_4)} = \frac{2.5-\vartheta}{1-\vartheta} > 1 \qquad \frac{\mathrm{Var}(\hat{\vartheta}_1)}{\mathrm{Var}(\hat{\vartheta}_4)} = \frac{1.25-\vartheta}{1-\vartheta} > 1$$

$\Rightarrow$ $\hat{\vartheta}_4$ ist der effizienteste unter den vier Schätzern.
$\hat{\vartheta}_4$ ist auch effizient:

$$\frac{\partial}{\partial\vartheta}\ln f_\vartheta(x) = \begin{cases} 1/\vartheta & x = 0,1,2,3 \\ -1/(1-\vartheta) & x = 4 \end{cases}$$

$$\mathrm{E}_\vartheta\left((\frac{\partial}{\partial\vartheta}\ln f_\vartheta(x))^2\right) = \frac{1}{\vartheta^2}\mathrm{P}(X\le 3) + \frac{1}{(1-\vartheta)^2}\mathrm{P}(X=4) = \frac{1}{\vartheta} + \frac{1}{1-\vartheta} = \frac{1}{\vartheta(1-\vartheta)}$$

Die untere Schranke nach Rao-Cramér ist:

$$\frac{1}{n\,E_\vartheta[(\frac{\partial}{\partial\vartheta}\ln f_\vartheta(x))^2]} = \frac{\vartheta(1-\vartheta)}{n} = \mathrm{Var}(\hat{\vartheta}_4)$$

Aufgabe 17

$$\mathrm{P}(X_{(n)}\le x) = \left(\frac{1}{\vartheta}x\right)^n \qquad 0\le x\le\vartheta$$

$$\lim_{n\to\infty} X_{(n)} = \vartheta \qquad \Leftrightarrow \qquad \forall m\,\exists n\,\forall k\ge n \quad \vartheta - X_{(k)} < 1/m$$

wegen Monotonie von $X_{(n)}$: $\lim_{n\to\infty} X_{(n)} = \vartheta \quad \Leftrightarrow \quad \forall m\, \exists n \quad \vartheta - X_{(n)} < 1/m$
$\Rightarrow$ Divergenz bedeutet: $\exists m\, \forall n \quad \vartheta - X_{(n)} \geq 1/m$.

$$\mathrm{P}\left(\bigcap_n \{\vartheta - X_{(n)} \geq 1/m\}\right) = \mathrm{P}\left(\bigcap_n \{X_{(n)} \leq \vartheta - 1/m\}\right) \leq \mathrm{P}(X_{(n)} \leq \vartheta - 1/m) \qquad \forall n$$

$$= \left(\frac{\vartheta - 1/m}{\vartheta}\right)^n \quad \longrightarrow \quad 0 \quad (\text{da } \vartheta - 1/m > 0)$$

$$\Rightarrow \qquad \mathrm{P}\left(\bigcup_m \bigcap_n \{\vartheta - X_{(n)} \geq 1/m\}\right) \leq \sum_{m=1}^{\infty} \mathrm{P}\left(\bigcap_n \{\vartheta - X_{(n)} \geq 1/m\}\right) = 0$$

$$\Rightarrow \qquad \mathrm{P}\left(\lim_{n\to\infty} |X_{(n)} - \vartheta| = 0\right) = 1\,.$$

Aufgabe 18

Wähle $S(X_1, X_2) = X_1 + X_2$ als suffiziente Statistik für p und definiere als neue erwartungstreue Schätzfunktion nach Rao-Blackwell gemäß $T_1 := E.(T|S)$ mit $\mathrm{Var}(T_1) \leq \mathrm{Var}(T)$

Es gilt

$$\begin{aligned}
\mathrm{E.}(T|X_1 + X_2 = x) &= 1 \cdot \mathrm{P}(T = 1|X_1 + X_2 = x) + 0 \cdot \mathrm{P}(T = 0|X_1 + X_2 = x) \\
&= \frac{\mathrm{P}(T = 1, X_1 + X_2 = x)}{\mathrm{P}(X_1 + X_2 = x)} = \frac{\mathrm{P}(X_1 = 0, X_2 = x)}{(x+1)p^2(1-p)^x} = \frac{p(1-p)^0 p(1-p)^x}{(x+1)p^2(1-p)^x} \\
&= \frac{1}{x+1} \\
\Rightarrow \qquad T_1(X_1, X_2) &= \frac{1}{X_1 + X_2 + 1}
\end{aligned}$$

Verteilung von $X_1 + X_2$ ist vollständig und bildet eine Exponentialfamilie. $\Longrightarrow T_1(X_1, X_2)$ ist nach Lehmann-Scheffé UMVUE.

Aufgabe 19

Nach Lehmann-Scheffé ist zu zeigen:

1. $\{f_\vartheta(x)|\vartheta \in \Theta\}$ ist vollständig
2. $X_{(n)}$ ist suffizient für ϑ
3. Es gibt erwartungstreues $T = t(X_{(n)})$ für ϑ

zu 1)

$X_{(n)}$ hat die Dichte $f_\vartheta(t) = \frac{n}{\vartheta^n} t^{n-1} \quad 0 < t < \vartheta$.

Sei $\mathrm{E}_\vartheta(h(X_{(n)})) = 0 \qquad \forall \vartheta \quad 0 < \vartheta < \infty$. Also:

$$\mathrm{E}_\vartheta(h(X_{(n)})) = \int_0^\vartheta h(t) \frac{n}{\vartheta^n} t^{n-1} \mathrm{d}t = 0 \Leftrightarrow \int_0^\vartheta h(t) t^{n-1} \mathrm{d}t = 0$$

$$h \text{ stetig} \Rightarrow \int_{\vartheta_1}^{\vartheta_2} h(t) t^{n-1} \mathrm{d}t = 0 \qquad \forall\, \vartheta_1, \vartheta_2 \quad \vartheta_1 < \vartheta_2$$

Es gibt kein Intervall (a,b) mit $h(t) > 0 \quad \forall\, t \in (a,b)$; sonst ergäbe sich ein Widerspruch durch

$$\int_a^b h(t) t^{n-1} \mathrm{d}t \geq \int_a^b h(t) a^{n-1} \mathrm{d}t \geq a^{n-1} \int_a^b h(t) \mathrm{d}t > 0 .$$

Analog gibt es kein Intervall (a,b) mit $h(t) < 0 \quad \forall\, t \in (a,b)$. Wegen der Existenz von $\int_a^b h(t) f(t) \mathrm{d}t$ folgt: h ist an höchstens abzählbar vielen Stellen unstetig!

$$\mathrm{P}(h(X_{(n)} \leq 0) = 1 \quad \bigwedge \quad \mathrm{P}(h(X_{(n)} \geq 0) = 1$$

zu 2) $T(X_1, \ldots, X_n) = \max(X_i)$ ist unabhängig von ϑ und somit suffizient für ϑ.

zu 3)

$$\mathrm{E}(X_{(n)}) = \int_0^\vartheta x \cdot \frac{n}{\vartheta^n} x^{n-1} \mathrm{d}x = \frac{n}{\vartheta^n} \left[\frac{1}{n+1} x^{n+1} \right]_0^\vartheta = \frac{n}{n+1} \vartheta$$

$$\Rightarrow \quad t(X_{(n)}) := \frac{n+1}{n} X_{(n)} \Rightarrow \quad \mathrm{E}(t(X_{(n)})) = \vartheta .$$

Somit ist nach Lehmann-Scheffé $T = t(X_{(n)}) = \frac{n+1}{n} X_{(n)}$ UMVUE.

Aufgabe 20

Einflussfunktion der Stichprobenvarianz

$$\begin{aligned} S^2 &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n-1} X_i^2 + \frac{1}{n} X_n^2 - \frac{1}{n^2} \left(\sum_{i=1}^{n-1} X_i \right)^2 - \frac{2}{n^2} \left(\sum_{i=1}^{n-1} X_i \right) X_n - \frac{1}{n^2} X_n^2 \end{aligned}$$

Also:

$$
\begin{aligned}
& n[\mathrm{E}(S^2|X_n=x)-\mathrm{E}(S^2)] \\
&= \sum_{i=1}^{n-1}\mathrm{E}(X_i^2) + x^2 - \frac{1}{n}\mathrm{E}\left(\left(\sum_{i=1}^{n-1}X_i\right)^2\right) - \frac{2}{n}(n-1)\mu\cdot x - \frac{1}{n}x^2 \\
&- \sum_{i=1}^{n-1}\mathrm{E}(X_i^2) - \mathrm{E}(X_n^2) + \frac{1}{n}\mathrm{E}\left(\left(\sum_{i=1}^{n-1}X_i\right)^2\right) + \frac{2}{n}(n-1)\mu^2 + \frac{1}{n}\mathrm{E}(X_n^2) \\
&= \frac{n-1}{n}x^2 - 2\frac{n-1}{n}\mu(x-\mu) - \frac{n-1}{n}\mathrm{E}(X_n^2) = \frac{n-1}{n}\left[x^2-\mathrm{E}(X^2)-2x\mu+\mu^2\right] \\
&= \frac{n-1}{n}\left[x^2-(\sigma^2+\mu^2)-2x\mu+\mu^2\right] = \frac{n-1}{n}\left[x^2-\sigma^2-2x\mu+\mu^2\right]
\end{aligned}
$$

somit

$$
IF(x,T_n,(\mu,\sigma^2)) = \lim_{n\to\infty}\frac{n-1}{n}\left[x^2-\sigma^2-2x\mu+\mu^2\right] = (x-\mu)^2-\sigma^2\,.
$$

Aufgabe 21

Maximum-Likelihood-Funktion mit Dichte von $Y=\ln X$:

$$
L(\mu_L,\sigma_L^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n\left(\frac{1}{\sigma^2}\right)^n \exp\left\{\frac{1}{-2\sigma^2}\left(\sum_{i=1}^{n}(\ln x_i-\mu_L)^2\right)\right\}
$$

$$
\ln(L(\mu_L,\sigma_L^2)) = -\frac{n}{2}(\ln 2\pi - \ln\sigma_L^2) - \frac{1}{2\sigma_L^2}\sum_{i=1}^{n}(\ln x_i-\mu_L)^2
$$

$$
0 \stackrel{!}{=} \frac{\partial}{\partial\mu_L}\ln(L) = \frac{1}{\sigma_L^2}\sum_{i=1}^{n}(\ln x_i-\mu_L)
$$

$$
\Rightarrow \quad \sum_{i=1}^{n}\ln x_i = n\hat{\mu}_L \quad \Rightarrow \quad \hat{\mu}_L = \frac{1}{n}\sum_{i=1}^{n}\ln x_i
$$

$$
0 \stackrel{!}{=} \frac{\partial}{\partial\sigma_L^2}\ln(L) = -\frac{n}{2\sigma_L^2} + \frac{1}{2\sigma_L^3}\sum_{i=1}^{n}(\ln x_i-\mu_L)^2
$$

$$
\Rightarrow \quad \frac{n\,2\hat{\sigma}_L^3}{2\hat{\sigma}_L^2} = \sum_{i=1}^{n}(\ln x_i-\mu_L)^2 \qquad \Rightarrow \qquad \hat{\sigma}_L^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln x_i-\hat{\mu}_l)^2
$$

Aufgaben aus Kapitel 7

Aufgabe 1

Paretoverteilung mit $k = 6$ und $\vartheta = 0.4$:

$$F_0(x_i) = \int_k^{\infty} \vartheta\, k^{\vartheta} x^{-(\vartheta+1)} \mathrm{d}x = \int_6^{\infty} 0.4 \cdot 6^{0.4} x^{-1.4} \mathrm{d}x = 1 - (6/x_i)^{0.4}$$

empirische Verteilung: $\hat{F}(x_i) = i/35$.
Wertetabelle für Kolmogorov-Smirnov-Anpassungstest

Nummer	x_i	$\hat{F}(x_i)$	$F_0(x_i)$	$\lvert F_0(x_i) - \hat{F}(x_i)\rvert$	$\lvert \hat{F}(x_{i-1}) - F_0(x_i)\rvert$
1	6.70	0.029	0.043	0.015	0.043
2	7.10	0.057	0.065	0.008	0.037
3	10.60	0.086	0.204	0.118	0.146
4	14.50	0.114	0.297	0.183	→ 0.212 ←
5	15.40	0.143	0.314	0.171	0.200
6	17.00	0.171	0.341	0.170	0.198
			…		
30	428.70	0.857	0.819	0.038	0.010
31	513.60	0.886	0.831	0.054	0.026
32	545.80	0.914	0.835	0.079	0.050
33	750.40	0.943	0.855	0.088	0.060
34	863.90	0.971	0.863	0.108	0.080
35	1638.00	1.000	0.894	0.106	0.077

Es gilt somit:

$$\sup |\hat{F}(x) - F_0(x)| = 0.212 \quad \overset{\text{Modif.}}{\Longrightarrow} \quad \sqrt{35} \cdot 0.212 = 1.254$$

Dieser Wert liegt damit unterhalb des kritischen Wertes (= 1.36) für $\alpha = 0.05$

$$\Rightarrow \qquad H_0 : X \sim \mathscr{P}\mathscr{A}(6, 0.4) \quad \text{wird nicht abgelehnt}$$

Aufgaben aus Kapitel 8

Aufgabe 1

Aufgabe 2

$$\lambda(x_1, x_2, x_3) = \frac{\sup_{\vartheta \in \Theta} p\vartheta(x_1, x_2, x_3)}{\sup_{\vartheta \in \Theta_0} p\vartheta(x_1, x_2, x_3)}$$

Verteilung der möglichen Stichproben mit Quotienten

			ϑ					
x_1	x_2	x_3	1	2	3	4	5	$\lambda(x_1,x_2,x_3)$
0	0	0	0.729	0.343	0.125	0.027	0.001	$0.125/0.729 = 0.171$
0	0	1	0.081	0.147	0.125	0.063	0.009	$0.125/0.147 = 1.176$
0	1	0	0.081	0.147	0.125	0.063	0.009	1.176
1	0	0	0.081	0.147	0.125	0.063	0.009	1.176
0	1	1	0.009	0.063	0.125	0.147	0.081	$0.147/0.063 = 2.33$
1	0	1	0.009	0.063	0.125	0.147	0.081	2.33
0	1	1	0.009	0.063	0.125	0.147	0.081	2.33
1	1	1	0.009	0.027	0.125	0.343	0.729	$0.729/0.027 = 27.0$

H_0 wird abgelehnt, falls $\lambda(x_1,x_2,x_3) \geq 2.33$

ii)

ϑ	$P_\vartheta(\lambda(x_1,x_2,x_3) \geq 2.33)$
1	0.028
2	0.216
3	0.500
4	0.784
5	0.972

...Test zum Niveau 0.216

Aufgabe 3

Verteilung der mögliche Stichproben

		ϑ			
x_1	x_2	1	2	3	4
-1	-1	0.36	0.16	0.09	0.01
-1	0	0.18	0.20	0.12	0.03
0	-1	0.18	0.20	0.12	0.03
-1	1	0.06	0.04	0.09	0.06
1	-1	0.06	0.04	0.09	0.06
0	0	0.09	0.25	0.16	0.09
0	1	0.03	0.05	0.12	0.18
1	0	0.03	0.05	0.12	0.18
1	1	0.01	0.01	0.09	0.36

i+ii) Tests zum Niveau $\alpha = 0.13$

		$P_\vartheta(A)$			
Nr.	Ablehnbereich A	1	2	3	4
1	$(1,1),(1,0),(0,1)$	0.07	0.11	0.33	0.72
2	$(1,1),(-1,1),(1,-1)$	0.13	0.09	0.27	0.48
3	$(1,1),(1,0),(1,-1)$	0.10	0.10	0.30	0.60
4	$(1,1),(1,0),(-1,1)$	0.10	0.10	0.30	0.60
5	$(1,1),(0,1),(1,-1)$	0.10	0.10	0.30	0.60
3	$(1,1),(1,0),(-1,1)$	0.10	0.10	0.30	0.60

iii) Alle Tests sind unverfälscht, da $P_{3,4}(A) \geq P_{1,2}(A)$ für alle Tests gilt!

iv) Test Nr.1 hat die größte Güte für $\vartheta = 3$ und $\vartheta = 4$ und ist somit bester Test!

Aufgabe 4

a) Sei $X_{(1)},\ldots,X_{(n)}$ die geordnete Stichprobe. Lehne H_0 ab, wenn $X_{(n)} \geq c_\alpha$. Unter H_0 gilt:

$$\mathrm{P}(X_{(n)} \leq c_\alpha) = \mathrm{P}(X_1 \leq c_\alpha,\ldots,X_n \leq c_\alpha) = \mathrm{P}(X_1 \leq c_\alpha)\cdot\ldots\cdot\mathrm{P}(X_n \leq c_\alpha) = \left(1-(1-p)^{c_\alpha+1}\right)^n$$

$$\begin{aligned}
\mathrm{P}(X_{(n)} \geq c_\alpha) = 1-\mathrm{P}(X_{(n)} \leq c_\alpha - 1) = 1-\left(1-(1-p)^{c_\alpha}\right)^n &\overset{!}{\leq} \alpha \\
\Leftrightarrow \quad 1-\alpha \leq \left(1-(1-p)^{c_\alpha}\right)^n \Leftrightarrow \quad (1-p)^{c_\alpha} &\leq 1-(1-\alpha)^{1/n} \\
\Leftrightarrow \quad c_\alpha \ln(1-p) \leq \ln(1-(1-\alpha)^{1/n}) \Leftrightarrow \quad c_\alpha &\geq \frac{\ln(1-(1-\alpha)^{1/n})}{\ln(1-p)} \\
\text{(da } \ln(1-p) < 0\text{)}.&
\end{aligned}$$

b) $X_{(4)} = 8 \dfrac{\ln(1-0.85^{1/4})}{\ln 0.75} = \dfrac{-3.2235}{-0.2877} = 11.205 > 8$. H_0 wird nicht abgelehnt!

c)

$$\frac{\ln(1-0.85^{1/8})}{\ln 0.75} = \frac{-3.9065}{-0.2877} = 13.579 > 8$$

H_0 wird nicht abgelehnt!

d) Für $n \to \infty$ gilt

$$c_{\alpha,n} = \frac{\ln(1-(1-\alpha)^{1/n})}{\ln(1-p)} \to \infty$$

weiter

$$\begin{aligned}
\mathrm{P}(X_{(n)} \leq c_{\alpha,n}|H_1) &= \mathrm{P}(X_1 \leq c_{\alpha,n},\ldots,X_n\, c_{\alpha,n}|H_1) \\
&= (1-(1-p_0)^{c_{\alpha,n}})^{n-1}(1-(1-p)^{c_{\alpha,n}}) \qquad \text{mit} \quad p < p_0 \\
\mathrm{P}(X_{(n)} \geq c_{\alpha,n}|H_1) &= 1-(1-(1-p_0)^{c_{\alpha,n}})^{n-1}(1-(1-p)^{c_{\alpha,n}}) \\
&= 1-(1-(1-p_0)^{c_{\alpha,n}})^n \frac{1-(1-p)^{c_{\alpha,n}}}{1-(1-p_0)^{c_{\alpha,n}}}
\end{aligned}$$

Da $c_{\alpha,n} \to \infty$, gilt für festes p :

$$\frac{1-(1-p)^{c_{\alpha,n}}}{1-(1-p_0)^{c_{\alpha,n}}} \to 1 \qquad \text{und damit} \qquad P(X_{(n)} \geq c_{\alpha,n} | H_1) \to \infty.$$

Aufgabe 5

$$E(X) = \sum_{x=1}^{\infty} x \cdot P(X=x) = \sum_{x=1}^{\infty} a\vartheta^x = a\left(\sum_{x=0}^{\infty} \vartheta^x - 1\right) = a\left(\frac{\vartheta}{1-\vartheta}\right)$$

$$\Rightarrow \quad \bar{x} = \frac{\hat{\vartheta}}{-\ln(1-\hat{\vartheta})(1-\hat{\vartheta})}$$

$$E(X^2) = \sum_{x=1}^{\infty} x^2 \cdot P(X=x) = a\sum_{x=1}^{\infty} x\vartheta^x = \frac{a}{1-\vartheta} \underbrace{\sum_{x=0}^{\infty} x(1-\vartheta)\vartheta^x}_{E(X) \text{ mit } X \sim \mathcal{GEO}(1-\vartheta)} = \frac{a}{1-\vartheta}\frac{\vartheta}{1-\vartheta} = \frac{a\vartheta}{(1-\vartheta)^2}$$

$$\Rightarrow \quad \bar{x^2} = \frac{\hat{\vartheta}}{-\ln(1-\hat{\vartheta})(1-\hat{\vartheta})^2}$$

Momentenschätzer:

$$\frac{\bar{x}}{\bar{x^2}} = \frac{\hat{\vartheta}(-\ln(1-\hat{\vartheta})(1-\hat{\vartheta})^2)}{-\ln(1-\hat{\vartheta})(1-\hat{\vartheta})(\hat{\vartheta})} = 1-\hat{\vartheta} \quad \Rightarrow \quad \hat{\vartheta} = 1 - \frac{\bar{x}}{\bar{x^2}} \qquad \text{als erste Näherung}$$

Für die empirischen Momente ergibt sich ($x \geq 1$):

$$\bar{x} = \frac{1}{387}\sum_{i=1}^{26} x_i \cdot n_i = 3.9272 \bar{x^2} = \frac{1}{387}\sum_{i=1}^{26} x_i^2 \cdot n_i = 24.5271 \Rightarrow \quad \hat{\vartheta} = 0.8656.$$

Vergleich mit beobachteten Daten:

x_i	1	2	3	4	5	6	7	8	9	10
n_i	164	71	42	28	17	12	12	9	7	6
$n \cdot P_{\hat{\vartheta}}(X=x_i)$	166.93	72.24	41.69	27.06	18.74	13.52	10.03	7.60	5.84	4.55

x_i	11	12	13	14	15	16	17	18	19	20
n_i	3	3	5	1	0	1	0	0	1	1
$n \cdot P_{\hat{\vartheta}}(X=x_i)$	3.58	2.84	2.27	1.83	1.47	1.20	0.97	0.80	0.65	0.54

x_i	21	22	23	24	25	26
n_i	1	1	0	1	0	1
$n \cdot P_{\hat{\vartheta}}(X=x_i)$	0.44	0.37	0.30	0.25	0.21	0.17

Geringe Abweichungen lassen auf eine sehr gute Eignung des logarithmischen Modells mit $\vartheta = 0.8656$ schließen!

Literatur

Andersen, E.B. (1990): *The Statistical analysis of categorical data*; Berlin: Springer

Angus, J.E. (1989): A note on the central limit theorem for the bootstrap mean; *Commun. Statist.-Theory Meth.*, 18, 1979-1982

Angus, J.E. and Schafer,R.E. (1984): Improved confidence Statements for the binimal parameter; *The American Statistician*, 38, 189-191

Bain, L.J. and Engelhardt,M. (1973): Interval estimation for the two-parameter double exponential distribution; *Technometrics*, 15, 875-885

Berger, J.O. (1980): *Statistical decision theory*; Berlin: Springer

Bickel, P.J. and Friedmann,D.A. (1981): Some asymptotic theory for the bootstrap; *Annals of Statistics*, 9, 1196-1217

Birkes, D.(1990): Generalized Likelihood Ratio Test and Uniformly Most Powerful Tests. *The American Statistician*, 44, 163-166

Box, G.E.P. (1953): Non-normality and tests on variances; *Biometrika* 40, 318-335

Buse, A. (1982): The likelihood ratio, Wald, and Lagrange multiplier tests: an expository note; *The American Statistican*, 36, 153-157

Bühler, W.S. and Sehr,J. (1987): Some remarks on exponential families; *The American Statistican*, 41, 279-280

Büning, H. (1991): *Robuste und adaptive Tests*; Berlin: de Gruyter

Büning, H. and Trenkler,G. (1994): *Nichtparametrische statistische Methoden*; 2te Auflage; Berlin: de Gruyter

Charnes, A., Frome,E.L. and Yu,P.L. (1976): The equivalence of generalized least squares and maximum likelihood estimates in the exponential family; *Journal of the American Statistical Association*, 71, 169-171

Chung, K.L. (1974): *Elementary Probability Theory with Stochastic Processes*; Berlin: Springer

Donoho, D.L. and Huber,P.J. (1983): The notion of breakdown point; in: *A Festschrift for Erich Lehmann, edited by P.Bickel, K.Doksum and J.L.Hodges Jr.*; Belmont, CA: Wadsworth

Douglas, J.B. (1993): Confidence regions for parameter pairs; *The American Statistician*, 47, 43-45

Durbin, J. (1973): *Distribution theory for tests based on the sample distribution function*; Regional Conference Series in Appl. Math. 9, SIAM, Phil. Penn.

Efron, B.(1979): Bootstrap methods: another look at the jackknife; *The Annals of Statistics*, 7, 1-26

Feller, W. (1968): *An introduction to probability theory and its applications Vol.1*; New York: Wiley

Ferentios, K.(1988): On shortest confidence intervals and their relation with uniformly minimum variance unbiased estimators; *Statistical Papers*, 29, 59-75

Ferentios, K.(1990): Shortest confidence intervals for families of distributions involving truncation parameters; *The American Statistician*, 44, 167-168

Fisz, M. (1978): *Wahrscheinlichkeitsrechnung und mathematische Statistik*; 9te Auflage; Berlin: Deutscher Verlag der Wissenschaften

Freedman, D. (1977): A remark on the difference between sampling with and without replacement; *Journal of the American Statistical Association*, 22, 681

Gabriel, K.R. (1969): Simultaneous test procedures - some theory of multiple comparisons; *Annals of Mathematical Statistics*, 40, 224-250

Gather, U., Pawlitschko,J. and Pigeot (1995): Unbiasedness of multiple tests; *Scandinavian Journal of Statistics*

Giri, N.C. (1993): *Introduction to probability and statistics*; 2nd ed.; New York: Marcel Dekker

Good, P.(1994): *Permutation Tests*; Berlin: York: Springer

Guenther, W. (1969): Shortest confidence intervals; *The American Statistician*, 23, 22-25

Guenther, W. (1971): Unbiased confidence intervals; *The American Statistician*, 25, 51-53

Hajek, J. (1969): *A Course in nonparametric statistics*; San Francisco: Hol-den Day

Hajek, J. and Sidak, Z. (1967): *Theory of rank tests*; New York: Academic Press

Hall, D.L. and Joiner, B.L. (1983): The ubiquitous role of f'/f in efficient estimation of location; *The American Statistican*, 37, 128-133

Hall, P. and Scala, B.La (1990): Methodology and algorithms of empirical likelihood; *International Statistical Review*, 58, 109-127

Hampel, F.R. (1971): A general qualitative definition of robustness; *Annals of Mathematical Statistics*, 42, 1887-1896

Hampel, F.R. (1974): The influence curve and its role in robust estimation; *Journal of the American Statistical Association*, 69, 383-393

Heyer, H. (1973): *Mathematische Theorie statistischer Experimente*; Berlin: Springer

Holm, S. (1979): *A simple sequentially rejective multiple test procedure*; Scandinavian Journal of Statistics 6, 65-70

Hosmer, D.W. and Lemeshow, S. (1989): *Applied logistic regression*; New York: Wiley

Huber, P.J. (1981): *Robust Statistics*; New York: Wiley

Johnson, M.E., Tietjen,G.L. and Beckman,R.J. (1980): A new family of probability densities with applications to Monte Carlo Studies; *Journal of the American Statistical Association* 75,276-279

Johnson, N.L., Kotz,S. (1970): *Continuous multivariate distributions*; New York: Wiley

Johnson, N.L., Kotz,S. and Balakrishnan,N. (1994): *Continuous univariate distributions Vol.1, 2nd ed.*; New York: Wiley

Johnson, N.L., Kotz,S. and Balakrishnan,N. (1995): *Continuous univariate distributions Vol.2, 2nd ed.*; New York: Wiley

Johnson, N.L., Kotz,S. and Kemp,A.W. (1992): *Univariate discrete distributions, 2nd ed.*; New York: Wiley

Juola, R.C. (1993): More on shortest confidence intervals; *The American Statistician*, 47, 117-119

Khuri, A.I. (1993): *Advanced Calculus with Applications in Statiscs*; New York: Wiley

Larsen, R.J. and Marx,M.L. (1986): *An introduction to mathematical statistics and its applications*; London: Prentice Hall

LeCam, L. (1990): Maximum Likelihood: An Introduction; *International Statistical Review*, 58, 153-171

Lehmann, E.L.(1986): *Testing Statistical Hypotheses 2nd ed.*; New York: Wiley

McCullach, P. (1994): Does the moment-generating function characterize a distribution ?; *The American Statistican*, 48, 208

Miller, R.G. jr. (1996): *Grundlagen der angewandten Statistik*; München: Oldenbourg

Morgenstern, D. (1958): Elementarer Beweis fur die asymptotische χ^2-Verteilung; *Metrika*, 1, 239-241

Mosteller, F., Fienberg, S.E. and Rourke, R.E.K. (1983): *Beginning statistics with data analysis*; Reading, MA: Addison-Wesley

Nölle, G. (1967): Zur Theorie der bedingten Tests. Unveröffentlichte Dissertation an der Universität Münster

Owen, A.B. (1988): Empirical likelihood ratio confidence intervals for a single functional; *Biometrika* 75, 237 - 249

Owen, A.B. (1990): Empirical likelihood ratio confidence regions; *The Annals of Statistics* 18,90 - 120

Pearson, K. (1920): Notes on the history of correlation; *Biometrika*, 13, 25-45

Peskun, P.H. (1990): A note on a general method for obtaining confidence intervals from samples from discrete distributions; *The American Statistician*, 44,31 – 35

QuenouiUe, M.H. (1956): Notes on bias in estimation; *Biometrika 43*, 353-360

Ramberg, J.S. and Schmeiser,B.W. (1972): An approximate method for generating symmetric random variables; *Commun. ACM*, 15, 987-990

Ramberg, J.S. and Schmeiser,B.W. (1974): An approximate method for generating asymmetric random variables; *Commun. ACM*, 17, 78-82

Randles, R.H. and Wolfe,D.A. (1979): *Introduction to the theory of nonparametric statistics*; New York: Wiley

Rao, C.R. (1973): *Lineare statistische Methoden und ihre Anwendungen*; Berlin: Akademie Verlag

Reiss, R.-D. (1989): *Approximate Distributions of Order Statistics*; Berlin: Springer

Rieder, H. (1982): Qualitative robustness of rank tests: *The Annals of Statistics*, 10, 205-211

Rieder, H. (1994): *Robust asymptotic statistics*; Berlin: Springer

Rohatgi, V.K. (1979): *An introduction to probability theory and mathematical statistics*; New York: Wiley

Romanowski, M. (1979): *Random errors in observations and the influence of modulation on their distribution*; Stuttgart: K. Wittwer

Rubin, D.B. (1981): The Bayesian bootstrap; *The Annals of Statistics*, 9,130-134

Santner, T.J. and Dufy,D.E. (1989): *The statistical analysis of discrete data*; Berlin: Springer

Schlittgen, R. (2008): *Einführung in die Statistik; 11te Auflage*; München: Oldenbourg

Scott, D.W. (1992): *Multivariate density estimation*; New York: Wiley

Seber, G.A.F. and Wild,C.J. (1989): *Nonlinear Regression*; New York: Wiley

Sen, P.K. and Singer,J.M. (1993): *Large sample methods in statistics*; London: Chapman and Hall

Serfling, R.J. (1980): *Approximation theory of mathematical statistics*; New York: Wiley

Silverman, B.W. (1986): *Density estimation for statistics and data analysis*; London: Chapman and Hall

Smirnow, W.I. (1976):*Lehrgang der höheren Mathematik, Teil V, 7te Aufig.*; Berlin: Deutscher Verlag der Wissenschaften

Sonnenmann, E. (1981): *Tests zum multiplen Niveau α*; unveroffentlichtes Manuskript; Abteilung Statistik der Universitat Dortmund

Sonnenmann, E. (1982): Allgemeine Losungen multipler Testprobleme; *EDV in Medizin und Biologie* 13, 120-128

Tate, R.F. and Klett,G.W.(1969): Optimal confidence intervals for the variance of a normal distribution; *Journal of the American Statistical Association* 54, 674 - 682

Uspensky, J.V. (1937): On Ch. Jordan's series for probability; *Annals of Mathematics*, 32, 306-312

Vaart, H.R. van der (1961): A simple derivation of the limiting distribution function of a sample quantile with increasing sample size; *Statistica Neerlandica* 15, 239-242

Wald, A. (1950): *Statistical decision functions*; New York: Wiley

Weisberg, S. (1985): *Applied linear regression*; New York: Wiley

Wilks, S.S. (1962): *Mathematical Statistics*; New York: Wiley

Wilrich, P.Th. (1979): Persönliche Mitteilung

Witting, H. (1985): *Mathematische Statistik I*; Stuttgart: Teubner

Zwet, W.R. van (1979): Mean, Median, Mode II; *Statistica Neerlandica,* 33, 1-5

Index