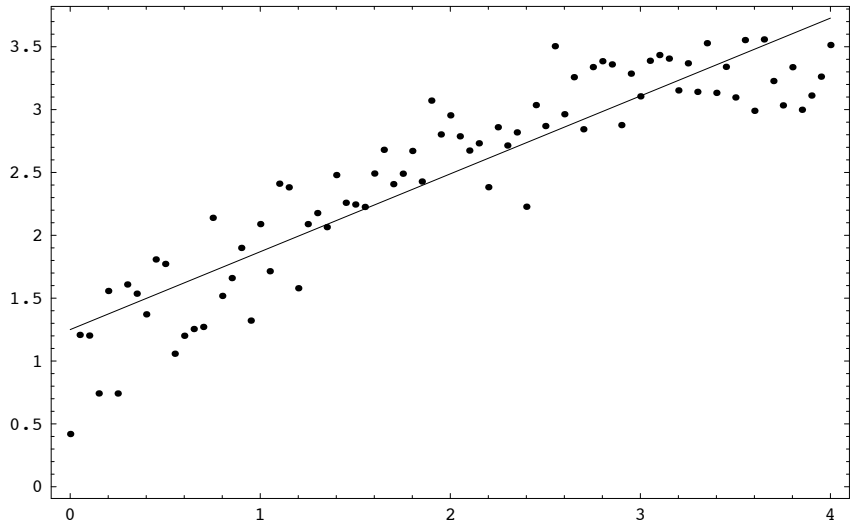
Gerold Alsmeyer

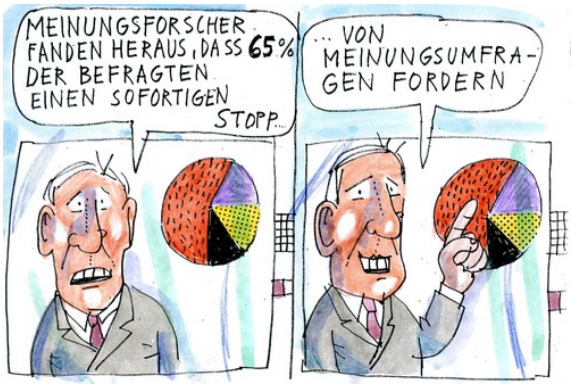
Einführung in die Mathematische Statistik

4. Dezember 2013



Skripten zur Mathematischen Statistik
Nr. 36 (4. Auflage, 2014)

Kapitel 1
Grundbegriffe der statistischen Entscheidungstheorie



Cartoon von Jan Tomaschoff

1.1 Begriffliche Abgrenzungen und ein Beispiel

Die mathematische Untersuchung zufallsabhängiger Phänomene einschließlich der Entwicklung des hierfür benötigten Rüstzeugs bildet den Inhalt der W-Theorie. Am Anfang steht immer die Vorgabe eines stochastischen Modells, das im Fall eines gegebenen realen Phänomens natürlich eine adäquaten Beschreibung desselben liefern sollte. Dabei verstehen wir unter "adäquat" einen hinreichenden Grad von Plausibilität der gemachten Annahmen. Erst dann tritt die Mathematik als exakte Wissenschaftauf den Plan, und zwar in Gestalt einer Analyse des Modells, zwecks Herleitung seiner wesentlichen Eigenschaften. Wohl jedem ist klar, dass eine korrekte Modellierung realer Phänomene aufgrund ihrer Komplexität nur in Ausnahmefällen möglich ist. Ein Modell liefert so gut wie immer nur eine Approximation der Realität, wobei seine Güte davon abhängt, inwieweit es die jeweils wesentlichen Einflussparameter berücksichtigt. Während der Wahrscheinlichkeitstheoretiker (Probabilist) nach Auswahl eines Modells nur noch reiner Mathematiker ist und das tut, was er immer tut, nämlich mathematische Sätze unter nunmehr festgelegten Voraussetzungen beweist, hat der Statistiker einen den Probabilisten oft nur wenig interessierenden, aber nicht minder gewichtigen Aspekt im Auge, nämlich den der Modellvalidierung. Gemeint ist das Bestreben, über die Analyse eines Modells in Abhängigkeit gegebener Modellparameter hinaus auch eine Antwort auf die quantitative Fragestellung zu geben, welche *Werte* diese Parameter in der realen Situation tatsächlich annehmen. Dabei fischt er in der Regel im Trüben und muss sich deshalb

darauf konzentrieren, auf der Basis der ihm zur Verfügung stehenden Information, gemeint sind Beobachtungswerte (Daten), eine möglichst gute Schätzung abzugeben. Für den Statistiker tritt also ein entscheidungstheoretischer Aspekt in den Vordergrund, den es wiederum zunächst zu mathematisieren gilt, da ja keineswegs klar ist, was unter "möglichst gut" überhaupt zu verstehen ist. Wer also geglaubt hat, in dieser Vorlesung von der Mathematik Abstand nehmen zu müssen oder zu können, der sei beruhigt oder enttäuscht: Statistik, in der hier vorgestellten Ausrichtung auch *"Mathematische Statistik"* genannt, bildet eine mathematische Teildisziplin mit einem lediglich anderen Schwerpunkt, wobei zum Verständnis die W-Theorie eine unabdingbare Voraussetzung darstellt.

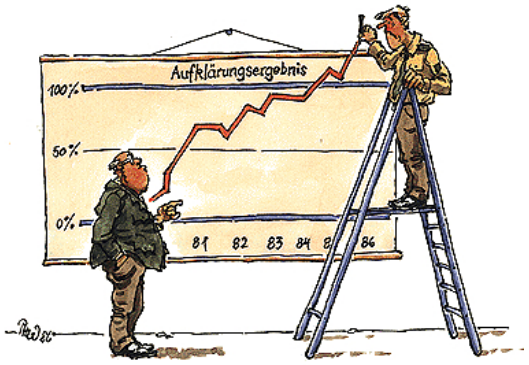


Abb. 1.1 Deskriptive Statistik: Ein Cartoon von der Webseite der hessischen Polizei

Die Bedeutung der Statistik weit über die Mathematik hinaus als eine Disziplin mit starkem Anwendungsbezug wird schon dadurch deutlich, dass der Begriff heutzutage zum Grundwortschatz der meisten Menschen unserer modernen Industriegesellschaft gehört und uns in vielfachen Bedeutungen und Zusammensetzungen begegnet, z.B. in *"statistisches Jahrbuch"*, *"Außenhandelsstatistik"*, *"statistische Auswertung von Versuchen"* oder in Formulierungen wie *"eine statistische Erhebung hat gezeigt, dass ..."* und *"statistisch ist bewiesen, dass ...* Würde man jedoch verschiedene Personen nach einer möglichst präzisen Definition fragen, so entstünde vermutlich ein sehr diffuses Bild, wenn es sich nicht gerade um lauter Experten handelte. Wir wollen daher mit einigen Klarstellungen und begrifflichen Abgrenzungen beginnen.

Die ersten beiden zuvor erwähnten Begriffe "statistisches Jahrbuch" und "Außenhandelsstatistik" betreffen die *Erhebung und Zusammenstellung gewisser Daten*. Die Aufbereitung von Daten für einen schnellen und leichten Zugang zu den darin enthaltenen Informationen fällt in das Gebiet der *beschreibenden (deskriptiven) Statistik* und ist nicht Gegenstand dieser Vorlesung. Gleiches gilt für die *explorative Datenanalyse*, die sich der Untersuchung von Daten widmet, über deren Zusammenhang nur geringes Wissen vorliegt. Im Zeitalter leistungsstarker Computer hat

sie sich zu einem wertvollen Instrument entwickelt, dessen Techniken insbesondere verwendet werden im sogenannten *Data-Mining*, wörtlich etwa *"Daten-Bergbau"* oder *"Daten-Ausbeutung"*, das die Erkennung von Mustern in großen Datenmengen mittels statistischer Techniken zum Ziel hat. Dagegen betreffen die oben genannte "statistische Auswertung von Versuchen" sowie die Formulierungen "eine statistische Erhebung hat gezeigt, dass ..." und "statistisch ist bewiesen, dass ..." schon eher den Gegenstand der von uns zu behandelnden, *schließenden (inferentiellen) Statistik*, in der es um die mathematische Analyse von Daten im Hinblick auf Entscheidungsprobleme geht, wie wir zuvor bereits angedeutet haben.

Ein einfaches Beispiel, das uns noch an verschiedenen Stellen wiederbegegnen wird, soll die Problematik näher veranschaulichen:

Beispiel 1.1. Ein Unternehmen hat die Umstellung eines Produktionsverfahrens vorgenommen und möchte diese bewerten. Vor der Umstellung hielten 12 % der erzeugten Produkte einer anschließenden Qualitätskontrolle nicht stand und bildeten somit Ausschuss. Zur Überprüfung, ob der Ausschussanteil, also der Anteil mangelhafter Produkte, nach der Umstellung gesunken ist, werden 100 Produkte des modifizierten Verfahrens derselben Qualitätsprüfung unterzogen. Wir nehmen einmal an, dass sich dabei 90 als "gut" und 10 als "mangelshaft" ergeben. Genauer besteht hier das Datenmaterial natürlich aus 100 Einstufungen "gut" oder "'mangelhaft", später meistens mit "0" oder "1" codiert, die wir aber aus Platzgründen nicht einzeln auflisten.

Kann die Firma aufgrund dieser Beobachtungswerte schließen, dass die Umstellung für sie einen positiven Effekt hat, oder sollte sie lieber zum alten Verfahren zurückkehren?

Der für die Statistik charakteristische Aspekt dieser Fragestellung besteht darin, dass das Eintreten von "gut"- bzw. "mangelhaft"-Ergebnissen einer einzelnen Überprüfung nicht nur von der Qualität des Verfahrens sondern noch von einer Vielzahl anderer, von uns nicht überschaubarer Einflüsse abhängt, die wir in ihrer Gesamtheit als zufällig ansehen können. Auf eine derartige Problematik treffen wir beispielsweise bei der Entscheidung über die Einführung neuer Medikamente oder neuer Düngemittel, bei dem Einsatz neuer Maschinen, bei der Verwendung eines neuen Materials, das gewissen Anforderungen genügen soll, etc.

Die Aufgabe der Statistik besteht darin, mathematische Modelle und Methoden zu ihrer Analyse zu entwickeln, um aus einer durch Erhebung oder Messung gewonnenen Stichprobe Entscheidungen abzuleiten, die die wahre, aber unbekannte Verteilung der beobachteten zufälligen Größe betreffen.

Zu diesem Zweck bedarf es zunächst einer Reihe notwendiger Formalisierungen und Begriffsbildungen, die wir deshalb als nächstes vornehmen wollen.

1.2 Statistische Experimente und Entscheidungsfunktionen

Grundlage der mathematischen Behandlung einer statistischen Problemstellung bildet das *statistische Experiment*, welches das den Beobachtungen zugrundegelegte Modell beschreibt.

Definition 1.2. Ein *statistisches Experiment* ist ein Tripel

$$\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta}),$$

bestehend aus

(1) einer nichtleeren Menge $\mathfrak{X}$, genannt *Stichprobenraum*, die die Menge der möglichen Beobachtungsergebnise bildet,
(2) einer σ-Algebra $\mathscr{A}$ über $\mathfrak{X}$, die die beobachtbaren Ereignisse enthält,
(3) einer Familie $(W_\theta)_{\theta\in\Theta}$ von W-Maßen auf $(\mathfrak{X}, \mathscr{A})$, die mit den Elementen des *Parameterraums* Θ parametrisiert ist. Für $\theta \neq \theta'$ gilt i.A. $W_\theta \neq W_{\theta'}$.

Mit jedem statistischen Experiment kann die Beobachtung einer Zufallsvariablen

$$X : (\Omega, \mathfrak{A}) \to (\mathfrak{X}, \mathscr{A})$$

und eine Familie $(\mathbb{P}_\theta)_{\theta\in\Theta}$ von W-Maßen auf $(\Omega, \mathfrak{A})$ verbunden werden, so dass

$$\mathbb{P}_\theta^X = W_\theta$$

für alle $\theta \in \Theta$. Wir sprechen dann von einem statistischen Experiment mit *Verteilungsannahme* $(W_\theta)_{\theta\in\Theta}$ bei Beobachtung von X. Die explizite Gestalt von $(\Omega, \mathfrak{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ spielt keine Rolle. Wird ein Wert $x \in \mathfrak{X}$ beobachtet, so interpretieren wir diesen als den von X angenommenen Wert, auch dessen *Realisierung* genannt. Eine messbare Funktion T auf $(\mathfrak{X}, \mathscr{A})$ heißt *Statistik*.

Erläuterungen anhand von Beispiel 1.1. Für die Auswertung der Qualitätskontrolle der getesteten 100 Produkte sind zwei Modellbildungen denkbar:

Modellbildung (A): Für $i = 1, ..., 100$ setzen wir

$$X_i := \begin{cases} 1, & \text{falls } i\text{-tes Produkt mangelhaft,} \\ 0, & \text{falls } i\text{-tes Produkt gut,} \end{cases}$$

so dass $X = (X_1, ..., X_{100})$ die Gesamtheit der Kontrollergebnisse (Versuchsresultate) beschreibt. Nehmen wir an, dass die X_i stochastisch unabhängig und identisch $Bern(\theta)$-verteilt sind mit unbekanntem Parameter $\theta \in [0,1]$, d.h. $\mathbb{P}_\theta^{X_i} = Bern(\theta)$, so genügt X einer $\bigotimes_{i=1}^{100} Bern(\theta)$-Verteilung unter $\mathbb{P}_\theta$, d.h.

$$W_\theta = \mathbb{P}_\theta^X = \bigotimes_{i=1}^{100} Bern(\theta).$$

Wir erhalten damit als statistisches Experiment

$$\mathscr{E}_1 = \left(\{0,1\}^{100}, \mathfrak{P}(\{0,1\}^{100}), \left(\bigotimes_{i=1}^{100} Bern(\theta) \right)_{\theta \in [0,1]} \right).$$

Modellbildung (B): Registrieren wir nur die Anzahl der mangelhaften Produkte, beobachten wir also $X = \sum_{i=1}^{100} X_i$, so genügt X unter den Modellannahmen von (A) einer $Bin(100,\theta)$-Verteilung unter $\mathbb{P}_\theta$, d.h.

$$W_\theta = \mathbb{P}_\theta^X = Bin(100,\theta).$$

Als statistisches Experiment ergibt sich in diesem Fall

$$\mathscr{E}_2 = (\{0,1,...,100\}, \mathfrak{P}(\{0,1,...,100\}), (Bin(100,\theta))_{\theta \in [0,1]}).$$

Unter Zugrundelegung eines der beiden Modelle soll nun eine "vernünftige" Antwort auf die Frage "Ist das neue Produktionsverfahren dem alten vorzuziehen?" gefunden werden. Zieht man als Bewertungskriterium nur den Parameter θ heran, so lässt sich die Frage präzisieren zu:

$$\text{“}\theta < 0.12\text{”} \qquad \text{oder} \qquad \text{“}\theta \geq 0.12\text{”?}$$

Die Entscheidung für eine der beiden Alternativen soll aufgrund der Realisierung x von X geschehen. Es ist also eine Vorschrift anzugeben, die jeder möglichen Beobachtung $x \in \mathfrak{X}$ eine dann zu treffende Entscheidung zuordnet. Dies wird wiederum mathematisch formalisiert:

Definition 1.3. Ist D eine nichtleere Menge, deren Elemente die möglichen Entscheidungen bilden, und $\mathfrak{D}$ eine σ-Algebra über D, so heißt jede messbare Abbildung $\delta : (\mathfrak{X}, \mathscr{A}) \to (D, \mathfrak{D})$ *Entscheidungsfunktion*. $(D, \mathfrak{D})$ wird *Entscheidungsraum* genannt.

Erläuterungen anhand von Beispiel 1.1. In Beispiel 1.1 sind die beiden Entscheidungen

$$\begin{array}{lcll} d_1 & \widehat{=} & \text{“}\theta < 0.12\text{”} & \widehat{=} \quad \text{neuer Prozess ist besser} \\ d_2 & \widehat{=} & \text{“}\theta \geq 0.12\text{”} & \widehat{=} \quad \text{neuer Prozess ist nicht besser} \end{array}$$

zu betrachten, also

$$(D, \mathfrak{D}) = (\{d_1, d_2\}, \mathfrak{P}(\{d_1, d_2\})).$$

Mögliche Entscheidungsfunktionen bei Modellbildung (A) sind etwa

$$\delta_1 : \{0,1\}^{100} \to D, \quad \delta_1(x) := \begin{cases} d_1, & \text{falls } \sum_{i=1}^{100} x_i \leq 8, \\ d_2, & \text{falls } \sum_{i=1}^{100} x_i > 8, \end{cases}$$

$$\delta_2 : \{0,1\}^{100} \to D, \quad \delta_2(x) := \begin{cases} d_1, & \text{falls } x_i \neq 1 \text{ für } i = 1,...,50, \\ d_2, & \text{sonst}, \end{cases}$$

wobei $x = (x_1,...,x_{100})$, und bei Modellbildung (B)

$$\widehat{\delta}_1 : \{0,1,...,100\} \to D, \quad \widehat{\delta}_1(x) := \begin{cases} d_1, & \text{falls } x \leq 8, \\ d_2, & \text{falls } x > 8, \end{cases}$$

$$\widehat{\delta}_2 : \{0,1,...,100\} \to D, \quad \widehat{\delta}_2(x) := \begin{cases} d_1, & \text{falls } x \text{ gerade}, \\ d_2, & \text{falls } x \text{ ungerade}. \end{cases}$$

Dabei ist anzumerken, dass δ_1 und $\widehat{\delta}_1$ die gleiche Entscheidung liefern und dass δ_2 und $\widehat{\delta}_2$ wenig sinnvoll sind.

1.3 Risikofunktionen und Optimalitätskriterien

Nachdem der Begriff der Entscheidungsfunktion mathematisch formalisiert worden ist, soll nun ein Maßstab gefunden werden, der eine Bewertung dieser Funktionen erlaubt. So soll z.B. die Aussage "die Entscheidungsfunktion δ_1 ist besser als die Entscheidungsfunktion δ_2" einen mathematischen Inhalt erhalten.

Wir legen für das Folgende ein statistisches Experiment $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta \in \Theta})$ und einen Entscheidungsraum $(D, \mathfrak{D})$ als gegeben zugrunde.

Definition 1.4. Eine *Verlustfunktion* ist eine Abbildung $L : \Theta \times D \to [0,\infty]$, so dass $L(\theta,\cdot) : d \mapsto L(\theta,d)$ für alle $\theta \in \Theta$ messbar ist.

$L(\theta,d)$ gibt den Verlust an, den man beim Treffen der Entscheidung d und beim Vorliegen der Verteilung W_θ erleidet. Im Allgemeinen nimmt man an, dass $L(\theta,d) = 0$, falls d bei Vorliegen von W_θ eine "richtige" Entscheidung bildet. Die Messbarkeit von $d \mapsto L(\theta,d)$ ist eine technische Voraussetzung, um die im Folgenden auftretenden Integrale bilden zu können.

Erläuterungen anhand von Beispiel 1.1. Falls $\theta < 0.12$, so ist d_1 die "richtige" Entscheidung, andernfalls d_2. Eine Verlustfunktion lässt sich z.B. angeben, indem man zwei Konstanten $L_1, L_2 > 0$ wählt und

$$L(\theta,d_1) := \begin{cases} 0, & \text{falls } \theta < 0.12 \\ L_1, & \text{falls } \theta \geq 0.12 \end{cases} \quad \text{und} \quad L(\theta,d_2) := \begin{cases} L_2, & \text{falls } \theta < 0,12 \\ 0, & \text{falls } \theta \geq 0.12 \end{cases}$$

definiert. Will man auch die Abweichung des wahren Parameters vom kritischen Wert 0.12 im Fall einer "falschen" Entscheidung berücksichtigen, so ist beispielsweise

$$L(\theta,d_1) := \begin{cases} 0, & \text{falls } \theta < 0.12, \\ L_1|\theta - 0.12|, & \text{falls } \theta \geq 0.12 \end{cases}$$

und

$$L(\theta,d_2) := \begin{cases} L_2|\theta - 0.12|, & \text{falls } \theta < 0.12, \\ 0, & \text{falls } \theta \geq 0.12 \end{cases}$$

möglich.

Definition 1.5. Ein *statistisches Modell* ist ein Tripel

$$\mathscr{S} = (\mathscr{E},(D,\mathfrak{D}),L),$$

bestehend aus einem statistischen Experiment $\mathscr{E}$, einem Entscheidungsraum $(D,\mathfrak{D})$ und einer Verlustfunktion L.

Um Entscheidungsfunktionen vergleichen zu können, wird als nächstes das *Risiko* einer solchen Funktion eingeführt. Den Erwartungswert bezüglich $\mathbb{P}_\theta$ bezeichnen wir im Folgenden im $\mathbb{E}_\theta$.

Definition 1.6. Sei $\mathscr{S} = (\mathscr{E},(D,\mathfrak{D}),L)$ ein statistisches Modell und

$$\mathscr{F} := \{\delta : (\mathfrak{X},\mathscr{A}) \to (D,\mathfrak{D})\}$$

die Menge der Entscheidungsfunktionen. Dann heißt $R : \Theta \times \mathscr{F} \to [0,\infty]$, definiert durch

$$R(\theta,\delta) := \int_{\mathfrak{X}} L(\theta,\delta(x))W_\theta(dx) = \mathbb{E}_\theta L(\theta,\delta(X)), \tag{1.1}$$

Risikofunktion. Für festes $\delta \in \mathscr{F}$ nennt man $R(\cdot,\delta)$ *Risikofunktion zu* δ.

Die Messbarkeit des in (1.1) auftretenden Integranden ist nach Annahme gesichert.

Der Vergleich von Entscheidungsfunktionen basiert nun auf dem *Prinzip der Entscheidungstheorie* und verwendet als Maßstab für deren Güte ihre Risikofunktionen. In der Praxis wird man natürlich auch andere Kriterien als die Risikofunktion berücksichtigen müssen, wie z.B. die einfache Berechenbarkeit eines Verfahrens

oder dessen Verhalten, wenn die wahre Verteilung nicht zu der im Modell angenommenen Verteilungsklasse $(W_\theta)_{\theta\in\Theta}$ gehört (Robustheit). Dennoch ist es von großem Wert, Lösungen in der hier angenommenen, idealisierten Situation zu bestimmen, auch wenn sie in der Praxis noch abgeändert werden sollten, denn die Güte eines Verfahrens wird sicher an dem gemessen, was im Idealfall optimal ist.

Wir kommen nun zur Formalisierung des Gütebegriffs von Entscheidungsfunktionen.

Definition 1.7. Sei $\mathscr{S} = (\mathscr{E}, (D, \mathfrak{D}), L)$ ein statistisches Modell und $\mathscr{F}$ die Menge aller zugehörigen Entscheidungsfunktionen. Auf $\mathscr{F}$ wird folgende Ordnungsstruktur eingeführt:

$$\delta_1 \preceq \delta_2 \quad \Leftrightarrow \quad R(\theta, \delta_1) \leq R(\theta, \delta_2) \text{ für alle } \theta \in \Theta,$$

gelesen "δ_1 *ist mindestens so gut wie (nicht riskanter als)* δ_2";

$$d_1 \prec \delta_2 \quad \Leftrightarrow \quad \delta_1 \preceq \delta_2 \text{ und } R(\theta, \delta_1) < R(\theta, \delta_2) \text{ für mindestens ein } \theta \in \Theta,$$

gelesen "δ_1 *ist besser (weniger riskant) als* δ_2".

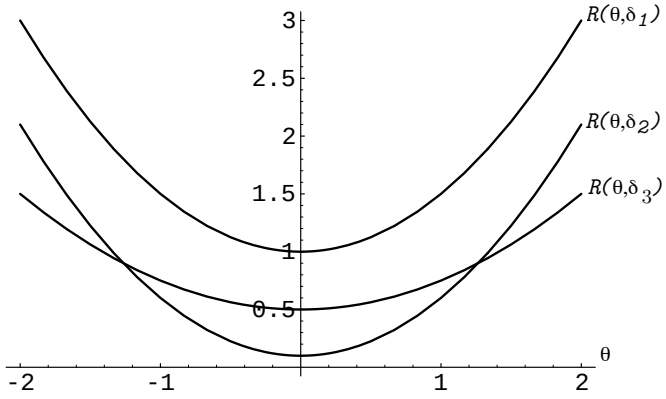


Abb. 1.2 Vergleich des Riskos von drei Entscheidungsfunktionen

Bezüglich dieser Ordnungsstruktur werden zwei Entscheidungsfunktionen jedoch i.A. *nicht vergleichbar* sein, wie Abb. 1.2 verdeutlicht. Dort sind δ_2 und δ_3 bzgl. "$\preceq$" nicht miteinander vergleichbar, während $\delta_2 \prec \delta_1$ und $\delta_3 \prec \delta_1$ gilt.

Definition 1.8. Sei $\mathscr{K} \subset \mathscr{F}$. Eine Funktion δ_0 heißt *gleichmäßig beste Entscheidungsfunktion in* $\mathscr{K}$, falls

$$\delta_0 \in \mathscr{K} \quad \text{und} \quad \delta_0 \preceq \delta \text{ für alle } \delta \in \mathscr{K}$$

gilt. Im Fall $\mathscr{K} = \mathscr{F}$ heißt δ_0 einfach *gleichmäßig beste Entscheidungsfunktion*.

In nichttrivialen Situationen existiert i.A. keine gleichmäßig beste Entscheidungsfunktion, wie das anschließende Beispiel verdeutlicht:

Beispiel 1.9. Um die Wirkung eines Düngemittels auf das Wachstum einer bestimmten Pflanzensorte zu untersuchen, wobei die erzielte Ertragsänderung pro Anbauflächeneinheit als Kriterium dient, werden die Erträge von n Paaren von Flächeneinheiten gemessen. Jedes Paar besteht aus jeweils einer Einheit mit und einer Einheit ohne Düngemittel bei sonst gleichen Anbaubedingungen. Die durch das Düngemittel resultierende Ertragsänderung soll unter Benutzung dieses Versuchsplans *geschätzt* werden. Als statistisches Modell könnte in dieser Situation das folgende dienen:

Sei $X_{1,j}$ der Ertrag der Einheit mit Düngemittel im j-ten Paar und $X_{2,j}$ der Ertrag der Einheit ohne Düngemittel im j-ten Jahr. Sei außerdem $X_j = X_{1,j} - X_{2,j}$ die Ertragsdifferenz. Der Vektor $X = (X_1, ..., X_n)$ dieser Differenzen ist eine Zufallsvariable mit Werten in $\mathfrak{X} = \mathbb{R}^n$, so dass wir als σ-Algebra $\mathscr{B}(\mathbb{R}^n)$ wählen. Unter der Annahme, dass die Ertragsdifferenzen für jedes Paar $Normal(\mu, \sigma^2)$-verteilt und voneinander unabhängig sind – eine in der Praxis sehr oft gemachte Annahme – folgt

$$(W_\theta)_{\theta \in \Theta} \;=\; \left(\bigotimes_{i=1}^{n} Normal(\mu, \sigma^2) \right)_{(\mu,\sigma^2) \in \mathbb{R} \times (0,\infty)},$$

wobei $\theta = (\mu, \sigma^2)$ und $\Theta = \mathbb{R} \times (0, \infty)$. Wählen wir den Mittelwert als Maß für die Ertragsänderung, so ergibt sich als Entscheidungsraum die Menge aller möglichen Mittelwerte, also $D = \mathbb{R}$, versehen mit der Borelschen σ-Algebra $\mathscr{B}(\mathbb{R})$. Eine gebräuchliche Verlustfunktion ist in diesem Fall die *quadratische Verlustfunktion*

$$L((\mu, \sigma^2), d) \;:=\; (d - \mu)^2.$$

Wir zeigen im Folgenden, dass in dem hiermit festgelegten statistischen Modell keine gleichmäßig beste Entscheidungsfunktion existiert. Für jedes $\mu \in \mathbb{R}$ setzen wir $\delta_\mu \equiv \mu$. Dann folgt offensichtlich

$$R((\mu, \sigma^2), \delta_\mu) \;=\; \int_{\mathbb{R}^n} L((\mu, \sigma^2), \delta_\mu(x))\, W_{(\mu,\sigma^2)}(dx) \;=\; 0$$

für alle $\sigma^2 > 0$. Für eine gleichmäßig beste Entscheidungsfunktion δ^* müsste also

$$R((\mu, \sigma^2), \delta^*) \;\leq\; R((\mu, \sigma^2), \delta_\mu) \;=\; 0$$

für alle $(\mu,\sigma^2)\in\Theta$ gelten, d.h.

$$\begin{aligned}0 &= R((\mu,\sigma^2),\delta^*)\\ &= \int_{\mathbb{R}^n}(\delta^*(x)-\mu)^2\,W_{(\mu,\sigma^2)}(dx)\\ &= \int_{\mathbb{R}^n}(\delta^*(x)-\mu)^2\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\right)\,\lambda^n(dx).\end{aligned}$$

Dies impliziert weiter, dass der Integrand λ^n-f.ü. verschwindet und somit

$$\delta^*(x)=\mu \quad \lambda^n\text{-f.ü.}$$

für jedes $(\mu,\sigma^2)\in\Theta$ gilt, was offensichtlich unmöglich ist.

Eine unter allen Entscheidungsfunktionen gleichmäßig beste finden zu wollen, macht also keinen Sinn. Andererseits gibt es in der vorliegenden Situation eine sehr naheliegende Entscheidungsfunktion: Das starke Gesetz der großen Zahlen besagt, dass das arithmetische Mittel $n^{-1}(X_1+\ldots+X_n)$ von unabhängigen und identisch verteilten Zufallsgrößen für $n\to\infty$ mit Wahrscheinlichkeit 1 gegen den Mittelwert $\mathbb{E}X_1$ konvergiert. In unserem Fall ist μ der unbekannte Mittelwert, und es liegt deshalb nahe, als Entscheidungsfunktion das sogenannte *Stichprobenmittel*

$$\overline{\delta}(x) \;=\; \bar{x} \;:=\; \frac{x_1+\ldots+x_n}{n}$$

zu wählen. Dann gilt

$$\begin{aligned}\int_{\mathbb{R}^n}\overline{\delta}(x)\,W_{(\mu,\sigma^2)}(dx) &= \int_{\mathbb{R}^n}\overline{\delta}(x)\,\mathbb{P}^X_{(\mu,\sigma^2)}(dx)\\ &= \mathbb{E}_{(\mu,\sigma^2)}\overline{\delta}(X) \;=\; \frac{1}{n}\sum_{i=1}^n\mathbb{E}_{(\mu,\sigma^2)}X_i \;=\; \mu.\end{aligned}$$

Im Mittel liefert $\overline{\delta}$ also stets das "richtige" Ergebnis; man bezeichnet eine solche Entscheidungsfunktion als *erwartungstreu* [☞ Definition 1.13]. Als dessen Risiko berechnen wir unter Benutzung der stochastischen Unabhängigkeit der X_i, $1\le i\le n$:

$$\begin{aligned}R((\mu,\sigma^2),\overline{\delta}) &= \int_{\mathbb{R}^n}(\overline{\delta}(x)-\mu)^2\,W_{(\mu,\sigma^2)}(dx) \;=\; \mathbb{E}_{(\mu,\sigma^2)}(\overline{\delta}(X)-\mu)^2\\ &= \mathbb{V}\mathrm{ar}_{(\mu,\sigma^2)}\overline{\delta}(X) \;=\; \mathbb{V}\mathrm{ar}_{(\mu,\sigma^2)}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)\\ &= \frac{1}{n^2}\sum_{i=1}^n\mathbb{V}\mathrm{ar}X_i \;=\; \frac{\sigma^2}{n}.\end{aligned}$$

Je mehr Beobachtungen man zur Verfügung hat, desto geringer wird demnach bei "vernünftigen" Entscheidungsfunktionen das Risiko.

Wir halten fest, dass das Finden einer gleichmäßig besten Entscheidungsfunktion i.A. nicht möglich ist und daher zunächst eine Klasse "vernünftiger" Entscheidungsfunktionen festgelegt werden muss, in Beispiel 1.9 etwa die erwartungstreuen Entscheidungsfunktionen. Eine andere Möglichkeit besteht darin, von dem vektoriellen Gütemaß $R(\cdot,\delta)$ zu einem reellwertigen Gütemaß $f(R(\cdot,\delta))$ überzugehen, so dass nur noch eine reellwertige Größe zu minimieren ist.

Definition 1.10. Ein $\delta_0 \in \mathscr{F}$ heißt *Minimaxverfahren*, falls gilt:

$$\sup_{\theta\in\Theta} R(\theta,\delta_0) \;\le\; \sup_{\theta\in\theta} R(\theta,\delta) \quad \text{für alle } \delta \in \mathscr{F}. \tag{1.2}$$

Bei Gebrauch eines Minimaxverfahrens will man sich also gegen den ungünstigsten Fall absichern. Die gewählte Bezeichnung erklärt sich, wenn man (1.2) umschreibt zu

$$\max_{\theta\in\Theta} R(\theta,\delta_0) \;=\; \min_{\delta\in\mathscr{F}} \max_{\theta\in\Theta} R(\theta,\delta),$$

natürlich vorausgesetzt, dass die auftretenden Suprema tatsächlich Maxima sind.

Definition 1.11. Auf Θ seien eine σ-Algebra und ein W-Maß ξ gegeben. ξ wird als *Vorbewertung* oder auch *a priori Verteilung* bezeichnet. Dann heißt $\delta_0 \in \mathscr{F}$ ein *Bayes-Verfahren zu* ξ, falls

$$\int_\Theta R(\theta,\delta_0)\,\xi(d\theta) \;\le\; \int_\Theta R(\theta,\delta)\,\xi(d\theta) \quad \text{für alle } \delta \in \mathscr{F}.$$

(Da Risikofunktionen stets nichtnegativ sind, existieren die auftretenden Integrale, können aber den Wert ∞ haben.)

Beim Bayes-Ansatz wird unterstellt, dass der Statistiker gewisse Vorkenntnisse über das Auftreten des Parameters θ besitzt, die sich in einer W-Verteilung ξ auf der Menge Θ aller möglichen Parameter niederschlagen. Natürlich muss man neben der Wahl einer geeigneten σ-Algebra auf Θ sicherstellen, dass die integrierten Risikofunktionen messbar sind. Die Benutzung von Bayes-Verfahren ist unter Statistikern nicht unumstritten und hat zu einer Unterscheidung von *Bayesianern* und *Frequentisten* geführt, die sich früher teilweise heftig befehdeten.

Eine minimale Anforderung an eine Entscheidungsfunktion besteht darin, dass es keine bessere im Sinne von Definition 1.7 gibt.

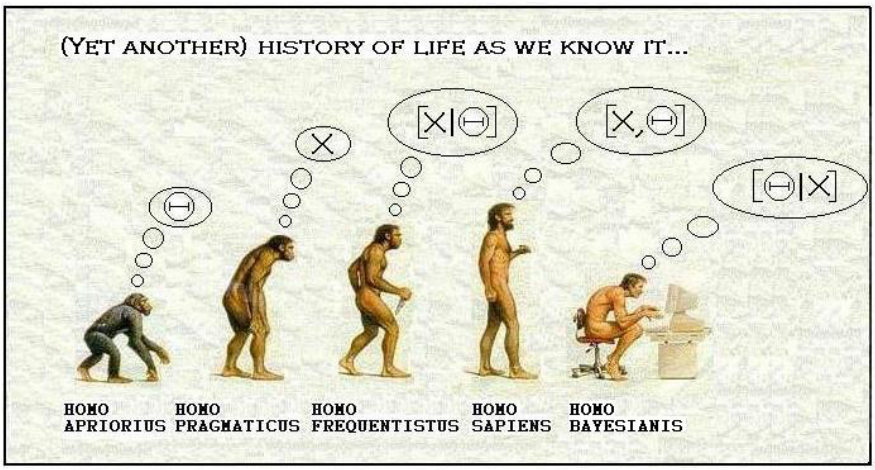


Abb. 1.3 Bayesian Learning: Eine etwas eigenwillige Sicht der Entstehungsgeschichte des Menschen. Quelle: http://stats.stackexchange.com

Definition 1.12. Eine Entscheidungsfunktion $\delta_0 \in \mathscr{F}$ heißt *zulässig*, falls kein $\delta \in \mathscr{F}$ mit $\delta \prec \delta_0$ existiert.

Obgleich die Forderung der Zulässigkeit vernünftig erscheint, kann ihr Nachweis für eine gegebene Entscheidungsfunktion ein schwieriges mathematisches Problem darstellen. Darüber hinaus kann es passieren, dass eine in einer vorgegebenen sinnvollen Teilklasse beste Entscheidungsfunktion nicht zulässig ist [☞ Satz 2.61 für eine derartige Situation].

1.4 Typen von statistischen Modellen

Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ hiernach immer ein gegebenes statistisches Experiment.

1.4.1 Schätzprobleme

Das statistische Modell der *Punktschätzungen* ist eine Abstraktion der Situation des Beispiels 1.9. Sei $\gamma : \Theta \to \mathbb{R}^m$ eine *Parameterfunktion*, die nach Beobachtung von $x \in \mathfrak{X}$ durch ein $d \in \mathbb{R}^m$ geschätzt werden soll. Es ist also $D \subset \mathbb{R}^m$ und $\mathfrak{D} = \mathscr{B}(\mathbb{R}^m)_D$ die zugehörige Borelsche (Spur-)σ-Algebra. Man benutzt in diesem Fall häufig eine Verlustfunktion der Gestalt

$$L(\theta,d) \;:=\; (\gamma(\theta)-d)^\top \Gamma(\gamma(\theta)-d),$$

in der Γ eine positiv definite Matrix bildet. Speziell für $m=1$ und $\Gamma=1$ heißt

$$L(\theta,d) \;=\; (\gamma(\theta)-d)^2$$

quadratische oder auch *Gauß-Markovsche Verlustfunktion*.

Entscheidungsfunktionen werden in der vorliegenden Situation als *Schätzer* oder *Schätzfunktionen* (für $\gamma(\theta)$) bezeichnet. Beispiel 1.9 hatte gezeigt, dass es i.A. keinen Sinn macht, einen optimalen unter allen Schätzern zu suchen, da ein solcher nicht existiert. Stattdessen schränkt man die Suche auf eine Teilklasse "vernünftiger", etwa aller erwartungstreuen Schätzer ein.

Definition 1.13. Ein Schätzer $g:\mathfrak{X}\to\mathbb{R}^m$ für $\gamma(\theta)$ heißt *erwartungstreu*, falls

$$\mathbb{E}_\theta g(X) \;=\; \int g\, dW_\theta \;=\; \gamma(\theta)$$

für alle $\theta\in\Theta$.

Für einen erwartungstreuen Schätzer ergibt sich im Fall $m=1$ bei quadratischer Verlustfunktion

$$R(\theta,g) \;=\; \int (g(x)-\gamma(\theta))^2\, W_\theta(dx) \;=\; \int (g(X)-\mathbb{E}_\theta g(X))^2\, d\mathbb{P}_\theta \;=\; \mathbb{V}\mathrm{ar}_\theta g(X).$$

In diesem Fall hat also das Risiko die anschauliche Interpretation der mittleren quadratischen Abweichung vom wahren Wert der Parameterfunktion.

1.4.2 Bereichsschätzungen

Nehmen wir an, dass in Beispiel 1.9 aufgrund der Beobachtung x die Benutzung des Schätzers $\overline{\delta}(x)=\overline{x}$ den Wert 112.5 liefert. Eine derartige Aussage wird uns häufig nicht ausreichen, denn wir wünschen uns eine Vorstellung über die Präzision bzw. den möglichen Fehler in der Schätzung. Dies führt dann zu Angaben der Form $\overline{x}\pm y$ als Schätzwert, im Beispiel etwa eine Ertragsänderung von 112.5 ± 5.3. Wie ist nun eine solche Aussage zu interpretieren? Tatsächlich hat man keine Garantie, dass der zu schätzende Wert zwischen den Grenzen $112.5-5.3$ und $112.5+5.3$ liegt, und muss daher eine wahrscheinlichkeitstheoretische Interpretation zugrundelegen. Grenzen der vorliegenden Art werden so gewählt, dass mit hoher vorgegebener Wahrscheinlichkeit, z.B. 95 %, der unbekannte Parameter tatsächlich zwischen diesen Grenzen, d.h. in der durch sie beschriebenen Menge liegt. Eine Formalisierung dieser Überlegungen führt zum Begriff des *Konfidenzbereichs*.

Gegeben sei wieder eine Parameterfunktion $\gamma : \Theta \to \mathbb{R}^m$, deren wahrer Wert $\gamma(\theta)$ mit gewisser Präzision zu schätzen ist. Als Entscheidungen wählt man nun Teilmengen von $\mathbb{R}^m$, wobei die getroffene Wahl $C \subset \mathbb{R}^m$ gerade die Entscheidung bedeutet, dass $\gamma(\theta)$ in C liegt.

Definition 1.14. Ein *Konfidenzbereich* für $\gamma(\theta)$, auch *Bereichsschätzfunktion* genannt, ist eine Abbildung $C : \mathfrak{X} \to \mathfrak{P}(\mathbb{R}^m)$ mit der Messbarkeitseigenschaft

$$\{x \in \mathfrak{X} : \gamma(\theta) \in C(x)\} \in \mathscr{A}$$

für alle $\theta \in \Theta$. Die Menge aller Konfidenzbereiche für $\gamma(\theta)$ wird mit $\mathfrak{C}$ bezeichnet. Für $\alpha \in [0,1]$ heißt ferner C ein *Konfidenzbereich zum Niveau* $1-\alpha$ oder auch $(1-\alpha)100\%$-*Konfidenzbereich*, falls

$$W_\theta(\{x \in \mathfrak{X} : \gamma(\theta) \in C(x)\}) \;\geq\; 1-\alpha$$

für alle $\theta \in \Theta$. Die Menge aller Konfidenzbereiche zum Niveau $1-\alpha$ bezeichnen wir mit $\mathfrak{C}_{1-\alpha}$.

Die Bezeichnung $(1-\alpha)\%$ anstelle von $\alpha\%$ hat ihren Grund in der statistischen Tradition, das Symbol α für *kleine* Irrtumswahrscheinlichkeiten zu verwenden und demgemäß durch $1-\alpha$ eine Wahrscheinlichkeit nahe bei 1 für eine korrekte Entscheidung zu suggerieren, ohne α tatsächlich zu spezifizieren. $(1-\alpha)100\%$-Konfidenzbereiche sollen also mit *hoher* Wahrscheinlichkeit $1-\alpha$ gewährleisten, dass der wahre Parameter $\gamma(\theta)$ in der gewählten Entscheidung liegt. Natürlich ist dies immer der Fall, sogar mit $1-\alpha = 1$, wenn man $C(x) \equiv \mathbb{R}^m$ wählt; andererseits erscheint diese Bereichsschätzfunktion wenig sinnvoll. Um wiederum zu einem Optimierungsproblem zu gelangen, wählt man zu jedem $\theta \in \Theta$ eine Teilmenge $F(\theta)$ von "falschen" oder "besonders ungünstigen" Parameterwerten, z B. $F(\theta) = \{\theta' \in \Theta : \theta' \neq \theta\}$. Gesucht wird dann eine Bereichsschätzfunktion, die bei wahrem θ die für θ "falschen" Parameter mit möglichst geringer Wahrscheinlichkeit enthält.

Definition 1.15. C_0 heißt *gleichmäßig bester Konfidenzbereich in* $\mathfrak{C}^* \subset \mathfrak{C}$ *bzgl.* F, falls $C_0 \in \mathfrak{C}^*$ und

$$W_\theta(\{x \in \mathfrak{X} : \gamma(\theta') \in C_0(x)\}) \;=\; \min_{C \in \mathfrak{C}^*} W_\theta(\{x \in \mathfrak{X} : \gamma(\theta') \in C(x)\})$$

für alle $\theta \in \Theta$ und $\theta' \in F(\theta)$ gilt. Im Fall $\mathfrak{C}^* = \mathfrak{C}_{1-\alpha}$ nennt man C_0 *gleichmäßig besten Konfidenzbereich zum Niveau* $1-\alpha$ oder auch *gleichmäßig besten* $(1-\alpha)100\%$-*Konfidenzbereich*.

1.4.3 Testprobleme

Das allgemeine Testproblem bildet eine Abstraktion der in Beispiel 1.1 vorliegenden Situation: Gegeben seien disjunkte $H, K \subset \Theta$, genannt *Hypothese* bzw. *Alternative*[1], mit $H + K = \Theta$. In unserem Beispiel bestand der Entscheidungsraum D aus zwei Elementen d_H und d_K. Es ist nun nützlich, eine Vergrößerung von D vorzunehmen, die auch die Alltagssituation der Unentschlossenheit modelliert. In einer solchen Situation werfe man eine möglicherweise unfaire Münze und entscheide sich, je nach Resultat des Münzwurfs, für d_H oder für d_K. Dies nennt man *Randomisieren mit Wahrscheinlichkeit* γ, wobei $\gamma \in (0,1)$ die Wahrscheinlichkeit für d_K angibt.

Um das beschriebene Vorgehen zu berücksichtigen, erweitern wir den Entscheidungsraum zu $D = [0,1]$, wobei

$$\begin{aligned} 0 &\mathrel{\widehat{=}} d_H \mathrel{\widehat{=}} \text{Annahme der Hypothese,} \\ 1 &\mathrel{\widehat{=}} d_K \mathrel{\widehat{=}} \text{Annahme der Alternative} \mathrel{\widehat{=}} \text{Verwerfen der Hypothese,} \\ (0,1) \ni \gamma &\mathrel{\widehat{=}} \text{Randomisieren mit Wahrscheinlichkeit } \gamma. \end{aligned}$$

Als σ-Algebra über D wählt man natürlich die Borelsche σ-Algebra $\mathscr{B}([0,1])$.

Definition 1.16. Jede Entscheidungsfunktion $\varphi : \mathfrak{X} \to [0,1]$ in einem Testproblem heißt *Test* oder *Testfunktion*. Nimmt ein Test nur die Werte 0 oder 1 an, nennt man ihn *nichtrandomisiert*, andernfalls *randomisiert*.

Als Verlustfunktion benutzt man die *Neyman-Pearsonsche Verlustfunktion*

$$L(\theta,\gamma) := \begin{cases} \gamma, & \text{falls } \theta \in H, \\ 1-\gamma, & \text{falls } \theta \in K \end{cases}$$

für alle $\gamma \in [0,1]$, speziell

$$L(\theta,0) = \begin{cases} 0, & \text{falls } \theta \in H, \\ 1, & \text{falls } \theta \in K \end{cases} \quad \text{und} \quad L(\theta,1) = \begin{cases} 1, & \text{falls } \theta \in H, \\ 0, & \text{falls } \theta \in K. \end{cases}$$

Ein Test φ hat damit die Risikofunktion

$$R(\theta,\varphi) = \begin{cases} \int \varphi \, dW_\theta = \mathbb{E}_\theta \varphi(X), & \text{falls } \theta \in H, \\ \int (1-\varphi) \, dW_\theta = 1 - \mathbb{E}_\theta \varphi(X), & \text{falls } \theta \in K. \end{cases}$$

Die Funktion $\beta_\varphi : \theta \mapsto \mathbb{E}_\theta \varphi(X)$ nennt man die *Gütefunktion von* φ, manchmal auch *Operationscharakteristik* oder abgekürzt *OC-Funktion*.

[1] Ebenfalls gebräuchliche Bezeichnungen für H und K in der Statistikliteratur sind *Nullhypothese* bzw. *Alternativhypothese*.

Hat man eine Entscheidung getroffen, kann diese entweder richtig sein oder man begeht einen der folgenden Fehler:

$$\begin{aligned} \text{Irrtümliches Verwerfen der Hypothese} &= \textit{Fehler 1. Art.} \\ \text{Irrtümliche Annahme der Hypothese} &= \textit{Fehler 2. Art.} \end{aligned}$$

Die Konsequenzen dieser beiden Fehler sind häufig sehr unterschiedlich. Untersucht man z.B. einen Patienten auf das Vorliegen einer bestimmten gefährlichen Krankheit, so führt die Fehlentscheidung für das Vorliegen der Krankheit, und damit die Durchführung einer Behandlung, häufig nur zu einer gewissen zeitlichen und/oder finanziellen Belastung des Patienten, während eine Nichtdiagnostizierung der tatsächlich vorliegenden Krankheit weitaus ernstere Folgen haben kann.

Man kann mittels eines Testverfahrens i.A. nicht erreichen, dass die Wahrscheinlichkeiten für einen Fehler 1. Art und einen Fehler 2. Art simultan klein werden. Es ist daher üblich, sich eine Schranke für den Fehler 1. Art vorzugeben, und Testfunktionen zu suchen, die den Fehler 2. Art unter dieser Nebenbedingung minimieren. Dies bedeutet für den Entscheidungsträger,

> als Alternative stets diejenige Entscheidung zu wählen, die bei irrtümlicher Annahme als riskanter eingeschätzt wird.

In der obigen Situation würde man demnach die Alternative als “der Patient ist gesund” wählen.

Für $\theta \in H$ gibt $R(\theta,\varphi) = \mathbb{E}_\theta \varphi(X)$ gerade die Wahrscheinlichkeit für den Fehler 1. Art bei Benutzung von φ an, während die Wahrscheinlichkeit für den Fehler 2. Art durch $R(\theta,\varphi) = 1 - \mathbb{E}_\theta \varphi(X)$ für $\theta \in K$ gegeben ist.

Definition 1.17. Sei $\alpha \in [0,1]$ ein vorgegebenes *Irrtums-* oder *Signifikanzniveau*. Dann heißt φ_0 ein *Test zum Niveau* α, falls

$$\mathbb{E}_\theta \varphi_0(X) \;\leq\; \alpha \quad \text{für alle } \theta \in H.$$

Die Menge all dieser Tests wird mit Φ_α bezeichnet. Minimiert φ_0 außerdem den Fehler 2. Art unter allen Tests in Φ_α, d.h.

$$\mathbb{E}_\theta \varphi_0(X) \;=\; \max_{\varphi \in \Phi_\alpha} \mathbb{E}_\theta \varphi(X) \quad \text{für alle } \theta \in K,$$

so heißt φ_0 *gleichmäßig bester Test zum Niveau* α.

Erläuterungen anhand von Beispiel 1.1. In der hier vorliegenden Situation hat die Entscheidung, die Produktion auf das neue Verfahren umzustellen, obwohl das alte besser war, i.A. unangenehmere Konsequenzen als die Entscheidung für die Beibehaltung des alten Verfahrens, obwohl eine Umstellung besser gewesen wäre. Man

wird deshalb die erstere als Alternative wählen und das Testproblem wie folgt festlegen:

$$\Theta = [0,1], \quad H = [0.12,1] \quad \text{und} \quad K = [0,0.12).$$

Bei Zugrundelegung von Modellbildung (B) entspricht die dort angegebene Entscheidungsfunktion $\widehat{\delta}_1$ (☞ im Anschluss an Definition 1.3) dem nichtrandomisierten Test

$$\varphi_1(x) \;=\; \begin{cases} 1, & \text{falls } x \le 8, \\ 0, & \text{falls } x \ge 9. \end{cases}$$

Ebenso möglich ist beispielsweise ein randomisierter Test der Form

$$\varphi_2(x) \;=\; \begin{cases} 1, & \text{falls } x \le 7, \\ \frac{1}{2}, & \text{falls } x = 8, \\ 0, & \text{falls } x \ge 9. \end{cases}$$

Als Gütefunktion von φ_1 erhalten wir offensichtlich

$$\mathbb{E}_\theta \varphi_1(X) \;=\; \mathbb{P}_\theta(X \le 8) \;=\; \sum_{j=0}^{8} \binom{100}{j} \theta^j (1-\theta)^{100-j},$$

und als Gütefunktion von φ_2 entsprechend

$$\begin{aligned} \mathbb{E}_\theta \varphi_2(X) \;&=\; \mathbb{P}_\theta(X \le 7) + \frac{1}{2}\,\mathbb{P}_\theta(X = 8) \\ &=\; \sum_{j=0}^{7} \binom{100}{j} \theta^j (1-\theta)^{100-j} \;+\; \frac{1}{2}\binom{100}{8} \theta^8 (1-\theta)^{92}. \end{aligned}$$

1.5 Dominierte Experimente und Exponentialfamilien

Wir haben im vorherigen Abschnitt eine Gliederung der statistischen Theorie hinsichtlich der Art der statistischen Entscheidungsprobleme vorgenommen. Eine weitere Unterscheidung lässt sich nach der Art der betrachteten Familie $(W_\theta)_{\theta\in\Theta}$ von W-Maßen geben. Dazu definieren wir zunächst

Definition 1.18. Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ ein statistisches Experiment. Die Familie $(W_\theta)_{\theta\in\Theta}$ wird ebenso wie das Experiment $\mathscr{E}$ als *dominiert* bezeichnet, wenn ein σ-endliches Maß ν auf $(\mathfrak{X}, \mathscr{A})$ existiert, so dass $W_\theta \ll \nu$ für alle $\theta \in \Theta$ gilt.

Nach dem Satz von Radon-Nikodym existieren in diesem Fall die ν-Dichten

$$f_\theta := \frac{dW_\theta}{d\nu}, \quad \left(\text{d.h. } W_\theta(A) = \int_A f_\theta \, d\nu \text{ für alle } A \in \mathscr{A}\right)$$

für alle $\theta \in \Theta$.

Die Untersuchung von statistischen Experimenten mit Parameterraum

$$\Theta \subset \mathbb{R}^d \quad \text{für ein } d \geq 1$$

wird als *parametrische Statistik* bezeichnet und zuerst Gegenstand dieses Textes sein. Bildet Θ dagegen eine Teilmenge eines unendlichdimensionalen Raums, so handelt es sich um ein Problem der *nichtparametrischen Statistik*, in der eine typische Familie von W-Maßen z.B. durch $(W_{(Q_1,Q_2)})_{(Q_1,Q_2)\in\Theta}$ mit $W_{(Q_1,Q_2)} = Q_1^n \otimes Q_2^n$ und

$$\Theta = \{(Q_1,Q_2) : Q_i \text{ stetiges W-Maß auf } (\mathbb{R},\mathscr{B}(\mathbb{R}))\}$$

gegeben ist.

Beispiel 1.19. Die Familie $(Normal(\mu,\sigma^2))_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)}$ der Normalverteilungen ist bekanntlich durch $\nu = \lambda$, das Lebesgue-Maß auf $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ dominiert. Dasselbe gilt für die Familie $(Exp(\theta))_{\theta\in(0,\infty)}$ der Exponentialverteilungen und viele andere bekannte Verteilungsklassen.

Beispiel 1.20. Im Fall eines *abzählbaren* Stichprobenraums $\mathfrak{X}$ mit $\mathscr{A} = \mathfrak{P}(\mathfrak{X})$ ist jede auf diesem definierte Verteilungsklasse $(W_\theta)_{\theta\in\Theta}$ dominiert, und zwar durch das Zählmaß auf $\mathfrak{X}$. Als konkretes Beispiel wählen wir $\mathfrak{X} = \{0,...,n\}$ und die Familie $(Bin(n,\theta))_{\theta\in(0,1)}$ der Binomialverteilungen. Die Zähldichte von $W_\theta = Bin(n,\theta)$ hat die Gestalt

$$f_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad \text{für jedes } x \in \mathfrak{X}.$$

Die für die Anwendung besonders bedeutsamen dominierten Verteilungsklassen der parametrischen Statistik werden durch folgende Definition gegeben.

Definition 1.21. Sei $\mathscr{E} = (\mathfrak{X},\mathscr{A},(W_\theta)_{\theta\in\Theta})$ ein dominiertes statistisches Experiment mit dominierendem Maß ν. Man nennt $(W_\theta)_{\theta\in\Theta}$ eine *Exponentialfamilie*, falls die ν-Dichten $f_\theta = \frac{dW_\theta}{d\nu}$ folgende Gestalt besitzen: Es existieren in $k \in \mathbb{N}$ und Abbildungen $C, Q_1, ..., Q_k : \Theta \to \mathbb{R}$ und messbare Abbildungen $h, T_1, ..., T_k : \mathfrak{X} \to \mathbb{R}$, so dass für jedes $\theta \in \Theta$

$$f_\theta \;=\; C(\theta)\exp\left(\sum_{i=1}^k Q_i(\theta)T_i\right)h \quad \nu\text{-f.ü.} \tag{1.3}$$

Setzen wir $\nu^* = h\nu$, so folgt unter Benutzung von Korollar 13.5 in [2]

$$\frac{dW_\theta}{d\nu^*} \;=\; C(\theta)\exp\left(\sum_{i=1}^k Q_i(\theta)T_i\right) > 0 \quad \nu^*\text{-f.ü.} \tag{1.4}$$

und daraus weiter [☞ Satz 13.11 in [2]]:

Lemma 1.22. *In einer Exponentialfamiie $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ sind alle Elemente paarweise äquivalent, und zwar gilt für $N \in \mathscr{A}$*

$$W_\theta(N) = 0 \ \textit{für alle/ein } \theta \in \Theta \quad \Leftrightarrow \quad \nu^*(N) = 0.$$

Eine Aussage ist also genau dann ν^*-f.ü. gültig, wenn sie W_θ-f.s. für jedes $\theta \in \Theta$ gilt, und für letzteres schreiben wir im Folgenden auch kurz $\mathscr{W}$-f.s. Mit (1.4) sieht man, dass die in der Definition der f_θ auftretende Funktion h stets dem dominierenden Maß ν "zugeschlagen" werden kann und sich deshalb die grundlegenden statistischen Eigenschaften von $\mathscr{W}$ allein aus der Gestalt der Exponentialterme $\exp\left(\sum_{i=1}^k Q_i(\theta)T_i\right)$ ergeben. Für jedes $\theta \in \Theta$ gibt $C(\theta)$ den Normierungsfaktor an und erfüllt die Beziehung

$$C(\theta) = \left(\int \exp\left(\sum_{i=1}^k Q_i(\theta)T_i(x)\right) h(x)\nu(dx)\right)^{-1}. \tag{1.5}$$

Man nennt $\mathscr{W}$ daher auch *k-parametrige Exponentialfamilie in* $(Q_1,...,Q_k)$ *und* $(T_1,...,T_k)$. Wie man sich leicht mittels Übergang zu einem anderen dominierenden Maß überlegt, sind weder k noch die Vektoren $(Q_1,...,Q_k)$ und $(T_1,...,T_k)$ eindeutig bestimmt. Wir gehen darauf mangels Bedeutung für unsere Zwecke jedoch nicht weiter ein und verweisen auf WITTING [21, Korollar 1.154 auf S. 146f.].

Beispiel 1.23. Die Familie $(Normal(\mu,\sigma^2))_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)}$ ist eine 2-parametrige Exponentialfamilie, denn für die zugehörigen λ-Dichten gilt

$$\begin{aligned} f_{(\mu,\sigma^2)}(x) &= \frac{1}{(2\pi)^{1/2}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \underbrace{\frac{1}{(2\pi)^{1/2}\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)}_{=C(\mu,\sigma^2)}\exp\left(-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x\right), \end{aligned}$$

d.h. (1.4) mit $k=2$, $\nu^* = \lambda$,

$$Q_1(\mu,\sigma^2) = -\frac{1}{2\sigma^2}, \quad Q_2(\mu,\sigma^2) = \frac{\mu}{\sigma^2}, \quad T_1(x) = x^2 \quad \text{und} \quad T_2(x) = x.$$

Beispiel 1.24. Die Familie $(Bin(n,\theta))_{\theta\in(0,1)}$ ist für festes $n \in \mathbb{N}$ eine einparametrige Exponentialfamilie, denn für die zugehörigen Zähldichten auf $\{0,...,n\}$ gilt

$$f_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \underbrace{(1-\theta)^n}_{=C(\theta)}\exp\left(\log\left(\frac{\theta}{1-\theta}\right)x\right)\binom{n}{x},$$

d.h. (1.3) mit $k=1$,

$$Q_1(\theta) \;=\; \log\left(\frac{\theta}{1-\theta}\right), \quad T_1(x) \;=\; x \quad \text{und} \quad h(x) \;=\; \binom{n}{x}.$$

Eine einfache Anwendung des Transformationssatzes liefert

Satz 1.25. *Ist $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ eine k-parametrige Exponentialfamilie in $Q = (Q_1,...,Q_k)$ und $T = (T_1,...,T_k)$, so ist $\mathscr{W}^T := (W_\theta^T)_{\theta\in\Theta}$ eine k-parametrige Exponentialfamilie in Q und der Identität auf $\mathbb{R}^k$.*

Beweis. Es genügt der Hinweis, dass offensichtlich für alle $\theta \in \Theta$

$$\frac{dW_\theta^T}{d\nu^{*T}}(t) \;=\; C(\theta)\exp\left(\sum_{i=1}^k Q_i(\theta)t_i\right) \quad \mathscr{W}^T\text{-f.s.}, \tag{1.6}$$

gilt, wobei $t = (t_1,...,t_k)$. □

Eine weitere nützliche Eigenschaft von Exponentialfamilien bildet ihre Konsistenz bei der Bildung von Produktmaßen. Wir sagen, dass eine Zufallsvariable eine Eigenschaft unter $(\mathbb{P}_\theta)_{\theta\in\Theta}$ besitzt, wenn sie diese Eigenschaft unter *jedem* $\mathbb{P}_\theta$, $\theta \in \Theta$, besitzt.

Satz 1.26. *Es seien $X_1,...,X_n$ unter $(\mathbb{P}_\theta)_{\theta\in\Theta}$ unabhängige Zufallsvariablen und $X = (X_1,...,X_n)$.*

(a) Bildet $(\mathbb{P}_\theta^{X_i})_{\theta\in\Theta}$ für jedes $i = 1,...,n$ eine Exponentialfamilie, so gilt dies auch für die Familie $(\mathbb{P}_\theta^X)_{\theta\in\Theta}$.

(b) Sind $X_1,...,X_n$ ferner identisch verteilt und bildet $(\mathbb{P}_\theta^{X_1})_{\theta\in\Theta}$ eine k-parametrige Exponentialfamilie in $Q = (Q_1,...,Q_k)$ und $T = (T_1,...,T_k)$, so ist auch $(\mathbb{P}_\theta^X)_{\theta\in\Theta}$ eine solche, und zwar in Q und

$$T_\Sigma(x) \;:=\; \sum_{i=1}^n T(x_i) \;=\; \left(\sum_{i=1}^n T_1(x_1),...,\sum_{i=1}^n T_k(x_i)\right), \quad x = (x_1,...,x_n).$$

Beweis. Da (a) sehr leicht nachzurechnen ist, betrachten wir nur Teil (b) und notieren, dass

$$\begin{aligned}\frac{d\mathbb{P}_\theta^X}{d\nu^{*n}}(x) \;=\; \prod_{j=1}^n \frac{d\mathbb{P}_\theta^{X_j}}{d\nu^*}(x_j) \;&=\; C(\theta)^n \prod_{j=1}^n \exp\left(\sum_{i=1}^k Q_i(\theta)T_i(x_j)\right)\\ &=\; C(\theta)^n \exp\left(\sum_{i=1}^k \left(Q_i(\theta)\sum_{j=1}^n T_i(x_j)\right)\right),\end{aligned}$$

was offensichtlich die Behauptung beweist. □

Als nächstes wollen wir auf den Übergang zu einer anderen, sehr naheliegenden Parametrisierung einer Exponentialfamilie $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ in $Q = (Q_1,...,Q_k)$ und $T = (T_1,...,T_k)$ eingehen. Setzen wir

$$\zeta \;=\; (\zeta_1,...,\zeta_k) \;:=\; (Q_1(\theta),...,Q_k(\theta))$$

und beachten in (1.5), dass $C(\theta)$ von θ nur über $Q(\theta)$ abhängt, wir also o.B.d.A. $C(Q(\theta)) = C(\zeta)$ schreiben können, so folgt nach Umparametrisierung $W_\theta \rightsquigarrow W_\zeta$

$$\frac{dW_\zeta}{d\nu^*} \;=\; C(\zeta)\exp\left(\sum_{j=1}^k \zeta_j T_j\right) \quad \nu^*\text{-f.ü.}$$

Wir nennen $\zeta = \zeta(\theta)$ *natürlichen Parameter* und sagen in diesem Fall, dass $\mathscr{W} = (W_\zeta)_{\zeta\in Q(\Theta)}$ in *natürlicher Parametrisierung* gegeben ist. Letztere erweist sich bei der Untersuchung analytischer Eigenschaften von $\mathscr{W}$ als besonders geeignet [☞ Satz 1.27]. Im Allgemeinen bildet $Q(\Theta)$ nur eine Teilmenge von

$$\mathfrak{Z} \;:=\; \left\{\xi = (\xi_1,...,\xi_k)\in\mathbb{R}^k : 0 < \int \exp\left(\sum_{j=1}^k \xi_j T_j\right) d\nu^* < \infty\right\}, \tag{1.7}$$

die als *natürlicher Parameterraum von* $\mathscr{W}$ bezeichnet wird. $\mathfrak{Z}$ besteht aus genau denjenigen Parametern ξ, für die $\exp\left(\sum_{j=1}^k \xi_j T_j\right)$ nach Normierung durch

$$C(\xi) = \left(\int \exp\left(\sum_{j=1}^k \xi_j T_j\right) d\nu^*\right)^{-1}$$

eine ν^*-Dichte bildet, d.h. ein W-Maß W_ξ mit dieser Dichte auf $(\mathfrak{X},\mathscr{A})$ definiert.

Die wichtigsten analytischen Eigenschaften von Exponentialfamilien in natürlicher Parametrisierung fasst der nachfolgende Satz zusammen, wobei $i = \sqrt{-1}$ die imaginäre Einheit bezeichnet.

Satz 1.27. *Sei $\mathscr{W} = (W_\zeta)_{\zeta\in\mathfrak{Z}}$ eine k-parametrige Exponentialfamilie in ζ und T mit natürlichem Parameterraum $\mathfrak{Z}$. Dann gilt:*

(a) *$\mathfrak{Z}$ ist konvex.*

(b) *Besitzt $\mathfrak{Z}$ innere Punkte, d.h. $\overset{\circ}{\mathfrak{Z}} \neq \emptyset$, so sei*

$$\mathfrak{R} \;:=\; \left\{\xi = \zeta + iy : \zeta\in\overset{\circ}{\mathfrak{Z}},\ y\in\mathbb{R}^k\right\} \;\subset\; \mathbb{C}^k.$$

Dann wird für jede bzgl. aller W_ζ, $\zeta\in\overset{\circ}{\mathfrak{Z}}$, integrierbare Funktion $\varphi : \mathfrak{X}\to\mathbb{R}$ durch $F : \mathfrak{R}\to\mathbb{C}$,

$$F(\xi) = \int \varphi(x) \exp\left(\sum_{j=1}^{k} \xi_j T_j(x) \right) \nu^*(dx)$$

eine Funktion definiert, die in jedem ξ_j bei festgehaltenen $\xi_1, ..., \xi_{j-1}, \xi_{j+1}, ..., \xi_k$ holomorph ist.

(c) *Es gilt für alle* $(\alpha_1, ..., \alpha_k) \in \mathbb{N}_0^k$

$$\begin{aligned} &\frac{\partial^{\alpha_1}}{\partial \xi_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_k}}{\partial \xi_k^{\alpha_k}} F(\xi) \\ &\quad = \int \varphi(x) T_1(x)^{\alpha_1} \cdot \ldots \cdot T_k(x)^{\alpha_k} \exp\left(\sum_{j=1}^{k} \xi_j T_j(x) \right) \nu^*(dx). \end{aligned} \tag{1.8}$$

Beweis. Wir verweisen auf das Buch von WITTING [21, Satz 1.161, Hilfssatz 1.162 und Korollar 1.163 auf S. 150ff] sowie auch Satz 40.4 in [2]. □

Im Hinblick auf die Frage, wann der natürliche Parameterraum innere Punkte besitzt, geben wir zunächst die folgende

Definition 1.28. Eine k-parametrige Exponentialfamilie $\mathscr{W} = (W_\theta)_{\theta \in \Theta}$ in $(Q_1, ..., Q_k)$ und $(T_1, ..., T_k)$ besitzt *vollen Rang*, wenn $1, Q_1, ..., Q_k$ linear unabhängig auf Θ und $1, T_1, ..., T_k$ linear unabhängig auf N^c für jede $\mathscr{W}$-Nullmenge $N \in \mathscr{A}$ sind, wobei letzteres bedeutet:

$$c_0 + \sum_{j=1}^{k} c_j T_j \;=\; 0 \quad \mathscr{W}\text{-f.s.} \quad \Rightarrow \quad c_0 = ... = c_k \;=\; 0.$$

Für k-parametrige Exponentialfamilien vollen Rangs in natürlicher Parametrisierung gilt schließlich:

Satz 1.29. *Sei $\mathscr{W} = (W_\zeta)_{\zeta \in \mathfrak{Z}}$ eine k-parametrige Exponentialfamilie vollen Rangs in ζ und T mit natürlichem Parameterraum $\mathfrak{Z}$. Dann gilt:*

(a) *$\mathfrak{Z}$ besitzt ein nichtleeres Inneres.*

(b) *Die Funktion*

$$\zeta \;\mapsto\; -\log C(\zeta) \;=\; \log\left(\int \exp\left(\sum_{j=1}^{k} \zeta_j T_j \right) d\nu^* \right)$$

ist strikt konvex.

Beweis. Wir verweisen erneut auf WITTING [21, Satz 1.161 auf S. 150]. □

1.6 Und zum Ende noch ein paar Formalitäten

Dieser Abschnitt, der beim ersten Lesen getrost übersprungen werden kann, dient der Klarstellung einiger Formalitäten, auf die der Leser im Bedarfsfall zurückgreifen kann, wenn er zu einem späteren Zeitpunkt im bisweilen schwer passierbaren Dickicht notwendiger Bezeichungen und Definitionen den Überblick zu verlieren droht oder schon verloren hat.

Das folgende Schema beschreibt die typische Grundsituation der in den nachfolgenden Kapiteln behandelten statistischen Modelle.

$$(\Omega,\mathfrak{A},(\mathbb{P}_\theta)_{\theta\in\Theta}) \xrightarrow{X} (\mathfrak{X},\mathscr{A},\underbrace{(\mathbb{P}_\theta^X)_{\theta\in\Theta}}_{=(W_\theta)_{\theta\in\Theta}}) \xrightarrow{T} (\mathfrak{X}',\mathscr{A}',\underbrace{(\mathbb{P}_\theta^{T(X)})_{\theta\in\Theta}}_{=(W_\theta^T)_{\theta\in\Theta}})$$

Die Beobachtungsvariable X ist auf einem nicht näher spezifizierten Raum $(\Omega,\mathfrak{A})$ mit einer Familie $(\mathbb{P}_\theta)_{\theta\in\Theta}$ von W-Maßen definiert und induziert das gegebene statistische Experiment $\mathscr{E}=(\mathfrak{X},\mathscr{A},(W_\theta)_{\theta\in\Theta})$ in der Mitte des Schemas. Eine Statistik T auf dem Stichprobenraum $(\mathfrak{X},\mathscr{A})$ mit Werten in einem messbaren Raum $(\mathfrak{X}',\mathscr{A}')$ dient häufig der *Datenreduktion* (etwa der Übergang von $x=(x_1,...,x_n)\in\mathbb{R}^n$ zu $T(x)=x_1+...+x_n\in\mathbb{R}$) und führt zu dem *reduzierten Experiment* [☞ Definition 2.18] $\mathscr{E}^T=(\mathfrak{X}',\mathscr{A}',(W_\theta^T)_{\theta\in\Theta})$.

Bei drei Familien von W-Maßen auf i.A. verschiedenen messbaren Räumen stellt sich die Frage nach einer sinnvollen Notation für die entsprechenden Erwartungswerte (Integrale). Mit $\mathbb{E}_\theta$ bezeichnen wir immer den Erwartungswert bezüglich $\mathbb{P}_\theta$, also $\mathbb{E}_\theta(\bullet)=\int\bullet\,d\mathbb{P}_\theta$, was in Analogie zur W-Theorie geschieht, wo $\mathbb{E}(\bullet)=\int\bullet\,d\mathbb{P}$. Entsprechend benutzen wir $\mathbb{V}\mathrm{ar}_\theta$ für die Varianz von ZG auf $(\Omega,\mathfrak{A})$ unter $\mathbb{P}_\theta$. Für das Integral einer messbaren numerischen Funktion $g:(\mathfrak{X},\mathscr{A})\to(\overline{\mathbb{R}},\mathscr{B}(\overline{\mathbb{R}}))$ bezüglich W_θ führen wir kein eigenes Erwartungswertsymbol ein, sondern schreiben einfach $W_\theta(g)$, was insbesondere $W_\theta(A)=W_\theta(\mathbf{1}_A)$ impliziert. Eine entsprechende Notation gilt für bedingte Erwartungswerte unter $\mathbb{P}_\theta$ und W_θ, also $\mathbb{E}_\theta(\bullet|\bullet)$ bzw. $W_\theta(\bullet|\bullet)$. Keine eigene Kurzschreibweise benutzen wir dagegen für Integrale bezüglich W_θ^T. Schließlich sei noch auf die folgenden Zusammenhänge zwischen

den verschiedenen Integralen hingewiesen:

$$W_\theta(g) \;=\; \mathbb{E}_\theta g(X), \tag{1.9}$$

$$\int h \, dW_\theta^T \;=\; W_\theta(h(T)) \;=\; \mathbb{E}_\theta h(T(X)) \tag{1.10}$$

$$W_\theta(g|S=\cdot)\circ S(X) \;=\; W_\theta(g|S)\circ X \;=\; \mathbb{E}_\theta(g(X)|S(X)), \tag{1.11}$$

wobei die letzte Zeile $\mathbb{P}_\theta$-f.s. gilt und äquivalent ist zu der W_θ^S-f.s. gültigen Identität

$$W_\theta(g|S=\cdot) \;=\; \mathbb{E}_\theta(g(X)|S(X)=\cdot) \tag{1.12}$$

Die auftretenden Funktionen und Zufallsvariablen sind natürlich auf jeweils geeigneten messbaren Räumen definiert. (1.9) und (1.10) folgen direkt mittels des Transformationssatzes, während wir für (1.11) noch notieren, dass

$$\begin{aligned}\int_{\{S(X)\in B\}} g(X)\, d\mathbb{P}_\theta \;&=\; \int_{\{S\in B\}} g \, dW_\theta \;=\; \int_{\{S\in B\}} W_\theta(g|S)\, dW_\theta \\ &=\; \int_B W_\theta(g|S=s)\, W_\theta^S(ds) \;=\; \int_B W_\theta(g|S=s)\, \mathbb{P}_\theta^{S(X)}(ds) \\ &=\; \int_{\{S(X)\in B\}} W_\theta(g|S=\cdot)\circ S(X)\, d\mathbb{P}_\theta\end{aligned}$$

für alle messbaren Teilmengen des Bildraums von S gilt.

Statistikerwitze zum Ersten

Statistiker zu sein heißt niemals sagen zu müssen, man sei sich sicher.

Zwei Statistiker befinden sich auf dem Weg von LA nach New York. Nach etwa einer Stunde Flugzeit meldet sich der Pilot mit der Nachricht, dass gerade ein Triebwerk ausgefallen sei, aber kein Grund zur Sorge bestehe, da sie noch drei intakte Triebwerke hätten. Die Flugzeit verlängere sich allerdings von ursprünglich 5 auf 7 Stunden.

Nur wenig später teilt er mit, dass ein weiteres Triebwerk ausgefallen sei, fügt aber beruhigend hinzu, dass ja noch zwei funktionieren würden. Lediglich die Flugzeit betrage nun 10 Stunden, um in New York anzukommen.

Wiederum eine Weile später meldet sich der Pilot erneut über Sprechfunk, um den Ausfall des dritten Triebwerk mitzuteilen. Grund zur Panik bestehe aber nicht, denn die Maschine könne problemlos mit einem Triebwerk fliegen. Die Flugzeit betrage nun aber leider 18 Stunden bis New York.

An diesem Punkt dreht sich der eine Statistiker zum anderen und seufzt “Meine Güte, ich hoffe, dass nicht auch noch das letzte Triebwerk ausfällt, sonst müssen wir nämlich für immer hier oben bleiben!”

Statistik spielt in der Genetik eine wichtige Rolle. Beispielsweise kann man mit Statistik beweisen, dass die Anzahl der Nachkommen eine Erbeigenschaft ist. Wenn deine Eltern kinderlos geblieben sind, ist die Wahrscheinlichkeit sehr groß, dass du auch keine bekommst.

Kapitel 2
Parametrische Schätztheorie und Suffizienz

2.1 Drei Methoden zum Finden von Schätzern

Im Folgenden seien ein statistisches Experiment $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ mit $\Theta \subset \mathbb{R}^d$ und eine zu schätzende Parameterfunktion $\gamma : \Theta \to \mathbb{R}^m$ zugrundegelegt. Zum Einstieg stellen wir uns die Frage, ob es einfache und/oder schnelle Verfahren gibt, zu einem "vernünftigen" Schätzer für $\gamma(\theta)$ zu gelangen, auch wenn dieser nicht unbedingt gewisse Optimalitätskriterien erfüllt. In manchen Fällen ist dies sehr einfach, wie das folgende Standardbeispiel zeigt: Sei $\gamma(\theta) = \mu(\theta)$ der unbekannte Mitelwert einer Verteilung $\widehat{W}_\theta$, der auf der Basis von n stochastisch unabhängigen, identisch $\widehat{W}_\theta$-verteilten Beobachtungen $X_1, ..., X_n$ geschätzt werden soll. Es gilt dann $W_\theta = \mathbb{P}_\theta^{(X_1,...,X_n)} = \widehat{W}_\theta^n$ für alle $\theta \in \Theta$. Als guter und zudem intuitiv naheliegender Schätzer für $\mu(\theta)$ erweist sich das *Stichprobenmittel*

$$g(x_1, ..., x_n) \;:=\; \overline{x}_n \;=\; \frac{x_1 + ... + x_n}{n},$$

denn er ist erwartungstreu, also

$$\mathbb{E}_\theta\, g(X_1, ..., X_n) \;=\; \mathbb{E}_\theta \overline{X}_n \;=\; \mu(\theta)$$

für alle $\theta \in \Theta$, und besitzt nach dem starken Gesetz der großen Zahlen außerdem die asymptotische Eigenschaft

$$\lim_{n\to\infty} \overline{X}_n \;=\; \mu(\theta) \quad \mathbb{P}_\theta\text{-f.s.}$$

für alle $\theta \in \Theta$. Bei wachsender Zahl von Beobachtungen strebt der Schätzer also stets – genauer mit Wahrscheinlichkeit 1 – gegen den wahren Parameter. Dies nennt man *starke Konsistenz*.

2.1.1 Die Momentenmethode

Statistik mag langweilig sein, aber es hat seine Momente.

Verweilen wir bei der zuvor betrachteten speziellen Situation und nehmen zusätzlich an, dass $E_\Theta|X_1|^d < \infty$ für alle $\theta \in \Theta \subset \mathbb{R}^d$ gilt. Dann bilden offensichtlich

$$\begin{aligned} \bar{x}_n \;=\; \bar{x}_n^{(1)} &:= \frac{x_1 + \ldots + x_n}{n}, \\ \bar{x}_n^{(2)} &:= \frac{x_1^2 + \ldots + x_n^2}{n}, \\ &\vdots \\ \bar{x}_n^{(d)} &:= \frac{x_1^d + \ldots + x_n^d}{n} \end{aligned}$$

erwartungstreue und stark konsistente Schätzer für

$$\mu_1(\theta) \;=\; \mathbb{E}_\theta X_1, \quad \mu_2(\theta) \;=\; \mathbb{E}_\theta X_1^2, \quad \ldots \quad \mu_d(\theta) = \mathbb{E}_\theta X_1^d.$$

Die *Momentenmethode* beinhaltet nun folgendes Vorgehen:
Stelle für den d-dimensionalen Parameter $\theta = (\theta_1, \ldots, \theta_d)$ die Gleichungen

$$\begin{aligned} [1] \qquad \bar{x}_n^{(1)} &= \mu_1(\theta_1, \ldots, \theta_d), \\ [2] \qquad \bar{x}_n^{(2)} &= \mu_2(\theta_1, \ldots, \theta_d), \\ &\vdots \\ [d] \qquad \bar{x}_n^{(d)} &= \mu_d(\theta_1, \ldots, \theta_d) \end{aligned} \tag{2.1}$$

auf und versuche, diese für $\theta_1, \ldots, \theta_d$ zu lösen.

Definition 2.1. Gegeben seien Realisierungen $x_1, \ldots, x_n$ von n unabhängigen, identisch verteilten ZG $X_1, \ldots, X_n$ zum Schätzen von $\theta = (\theta_1, \ldots, \theta_d)$, $d \geq 1$. Es gelte $\mathbb{E}_\theta|X_1|^d < \infty$. Besitzt das Gleichungssystem (2.1) eine Lösung $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$, so heißt diese ein *Momentenmethode-Schätzer (MMS) für* θ und jede Komponente $\widehat{\theta}_j = \widehat{\theta}_j(x_1, \ldots, x_n)$ ein *MMS für* θ_j, $j = 1, \ldots, d$, bei gegebenem $x = (x_1, \ldots, x_n)$.

Der Leser beachte bei dieser Definition, dass ein MMS als Funktion von $x = (x_1, \ldots, x_n)$ nicht notwendig für alle $x \in \mathfrak{X}$ zu existieren braucht. Aus entscheidungstheoretischer Sicht liefert die Momentenmethode also möglicherweise einen nur partiell, d.h. auf einer echten Teilmenge $\mathfrak{X}'$ des Stichprobenraums definierten Schätzer. In diesem Fall ist eine Berechnung seiner Risikofunktion und damit der Vergleich mit anderen Schätzern offenbar unmöglich, es sei denn, $W_\theta((\mathfrak{X}')^c) = 0$

für alle $\theta \in \Theta$ oder man vervollständigt die Definition des Schätzers in irgendeiner vernünftigen Weise auf $(\mathfrak{X}')^c$. Andererseits wird diese mathematisch unbefriedigende Feststellung aus praktischer Sicht hinnehmbar, wenn man bedenkt, dass die Momentenmethode dem stets mit einer fest gegebenen Beobachtung (Datensatz) konfrontierten Anwender lediglich ein mögliches Verfahren liefert, zu einer Schätzung des unbekannten Parameters zu gelangen. Führt sie zu keinem Ergebnis, muss er sich eben nach einer Alternative umschauen.

Zur Veranschaulichtung der Methode betrachten wir eine Reihe von Beispielen.

Beispiel 2.2. (Normalverteilungen) Nehmen wir an, dass $X_1,...,X_n$ stochastisch unabhängig und jeweils *Normal*(μ,σ^2)-verteilt sind mit unbekanntem Parameter $\theta = (\mu,\sigma^2) \in \mathbb{R}\times(0,\infty)$. Dann gilt $\mu_1(\theta) = \mathbb{E}_\theta X_1 = \mu$ und $\mu_2(\theta) = \mathbb{E}_\theta X_1^2 + \mathbb{V}\mathrm{ar}_\theta(X_1) = \mu^2+\sigma^2$. Folglich ist das Gleichungssystem

$$\begin{aligned} \bar{x}_n &= \mu, \\ \bar{x}_n^{(2)} &= \mu^2+\sigma^2, \end{aligned}$$

in μ und σ^2 zu lösen. Wir erhalten

$$\widehat{\mu} = \widehat{\mu}(x_1,...,x_n) = \bar{x}_n \quad \text{und}$$
$$\widehat{\sigma}^2 = \widehat{\sigma}^2(x_1,...,x_n) = \bar{x}_n^{(2)} - \bar{x}_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x}_n)^2.$$

In diesem einfachen Beispiel liefert die Momentenmethode Schätzer für die unbekannten Parameter, die mit unserer Intuition übereinstimmen und zudem auf dem ganzen Stichprobenraum $\mathfrak{X}$ definiert sind. Von größerem Nutzen sind jedoch Anwendungen auf Schätzprobleme, in denen keine offensichtlichen Schätzer für die unbekannten Parameter existieren.

Beispiel 2.3. (Gammaverteilungen) Seien $X_1,...,X_n$ stochastisch unabhängig und jeweils $\Gamma(\alpha,\beta)$-verteilt mit unbekanntem Parameter $\theta = (\alpha,\beta) \in (0,\infty)^2$. Die Lebesgue-Dichte von X_1 unter $\mathbb{P}_{(\alpha,\beta)}$ lautet also

$$f_{(\alpha,\beta)}(x) = \Gamma(\alpha)^{-1}\beta^\alpha x^{\alpha-1}e^{-\beta x}\mathbf{1}_{(0,\infty)}(x).$$

Ferner gilt

$$\mathbb{E}_{(\alpha,\beta)}X_1 = \frac{\alpha}{\beta} \quad \text{und} \quad \mathbb{V}\mathrm{ar}_{(\alpha,\beta)}X_1 = \frac{\alpha}{\beta^2},$$

also $\mathbb{E}_{(\alpha,\beta)}X_1^2 = \frac{\alpha(1+\alpha)}{\beta^2}$. Demnach ist hier das Gleichungssystem

$$\begin{aligned} \bar{x}_n &= \frac{\alpha}{\beta}, \\ \bar{x}_n^{(2)} &= \frac{\alpha(1+\alpha)}{\beta^2} \end{aligned}$$

in α und β zu lösen. Wir setzen $\mathsf{v}_n^2 = \overline{x}_n^{(2)} - \overline{x}_n^2 = n^{-1}\sum_{j=1}^n (x_j - \overline{x}_n)^2$. Wie im vorherigen Beispiel gesehen, ist v_n^2 der MMS für die Varianz der X_j. Das Gleichungssystem (2.1) kann zu

$$\begin{aligned} \overline{x}_n &= \frac{\alpha}{\beta}, \\ \mathsf{v}_n^2 &= \frac{\alpha}{\beta^2} \end{aligned}$$

umgeformt werden, und man erhält daraus leicht als MMS für α und β

$$\widehat{\alpha} = \left(\frac{\overline{x}_n}{\mathsf{v}_n}\right)^2 \quad \text{und} \quad \widehat{\beta} = \frac{\overline{x}_n}{\mathsf{v}_n^2}.$$

Das zuvor gewählte Vorgehen, Gleichung [2] mit $\overline{x}_n^{2)}$ durch eine mit $\mathsf{v}_n^2 = \overline{x}_n^{(2)} - \overline{x}_n^2 = n^{-1}\sum_{j=1}^n (x_j - \overline{x}_n)^2$ zu ersetzen, ist natürlich immer möglich, indem man in (2.1) die quadrierte Gleichung [1] von Gleichung [2] abzieht. Anstelle von [2] tritt dort folglich der MMS für die Varianz

$$[2'] \qquad \mathsf{v}_n^2 = \mu_2(\theta_1, ..., \theta_d) - \mu_1(\theta_1, ..., \theta_d)^2.$$

Beispiel 2.4. (Binomial-Verteilungen) Als letztes Beispiel betrachten wir die Situation n unabhängiger $Bin(m,p)$-verteilter Beobachtungen $X_1, ..., X_n$ mit unbekannten Parametern $m \in \mathbb{N}$ und $p \in (0,1)$, d.h. $\theta = (m,p)$.

Da bekanntlich $\mathbb{E}_\theta X_1 = mp$ und $\mathbb{V}\mathrm{ar}_\theta X_1 = mp(1-p)$, erhalten wir das Gleichungssystem

$$\begin{aligned} \overline{x}_n &= mp, \\ \mathsf{v}_n^2 &= mp(1-p) \end{aligned}$$

in m und p und daraus nach ein wenig Algebra die MMS

$$\widehat{m} = \frac{\overline{x}_n^2}{\overline{x}_n - \mathsf{v}_n^2} \quad \text{und} \quad \widehat{p} = \frac{\overline{x}_n}{\widehat{m}} = 1 - \frac{\mathsf{v}_n^2}{\overline{x}_n}$$

für m bzw. p. Der aufmerksame Leser wird feststellen, dass $\widehat{m}, \widehat{p}$ auch negative Werte annehmen und dass $\widehat{m}$ i.A. nicht ganzzahlig ist. Dieses Problem, dass der Wertebereich eines Schätzers nicht mit dem des zu schätzenden Parameters übereinstimmt, tritt häufig auf und bildet neben der Möglichkeit eines nur partiell definierten Schätzers ein weiteres unerfreuliches Phänomen, das bei Verwendung der Momentenmethode auftreten kann. Der Fairness halber muss allerdings festgehalten werden, dass sich hier negative Werte für $\widehat{m}$ und $\widehat{p}$ nur dann ergeben, wenn die Stichprobenvarianz v_n^2 den Wert des Stichprobenmittels $\overline{v}_n$ übersteigt, was eine hohe Schwankungsbreite der Daten (relativ zum Mittelwert) bedeutet und damit generell eine hohe Unzuverlässigkeit jedweder Schätzung nach sich zieht. Die Mo-

mentenmethode hat uns aber immerhin zwei Kandidaten von Schätzern für m und p geliefert, was im Fall von m ein keineswegs einfaches Problem darstellt.

2.1.2 Die Maximum-Likelihood-Methode

Die Maximum-Likelihood-Methode ist eine der populärsten Techniken zum Finden von Schätzfunktionen, wenngleich auch sie von den zuvor angesprochenen unerfreulichen Phänomenen begleitet wird. Zu ihrer Beschreibung benötigen wir zunächst folgende

Definition 2.5. Sei $\mathscr{E}$ ein durch ν dominiertes Experiment mit Dichten $f_\theta = \frac{dW_\theta}{d\nu}$. Gegeben den Beobachtungswert $X = x \in \mathfrak{X}$, wird $\mathbf{L}(\cdot|x) : \Theta \to [0,\infty)$, definiert durch

$$\mathbf{L}(\theta|x) := f_\theta(x),$$

als *Likelihood-Funktion von X (gegeben x)* bezeichnet. Entsprechend heißt $\ell(\theta|x) := \log \mathbf{L}(\theta|x)$ *Log-Likelihood-Funktion von X (gegeben x).*

Im Fall einer diskreten Verteilungsfamilie $(W_\theta)_{\theta\in\Theta}$ mit dominierendem Zählmaß gibt $\mathbf{L}(\theta|x)$ also nichts anderes als die Wahrscheinlichkeit des Beobachtungswertes x unter W_θ an, und das Maximum-Likelihood-Prinzip bedeutet, diejenige Verteilung $W_{\widehat{\theta}}$ zu finden, unter der diese Wahrscheinlichkeit maximiert wird. Man kann die Interpretation auf den allgemeinen Fall übertragen, wenn man den Wert $f_\theta(x)$ bis auf einen von θ unabhängigen Faktor als Wahrscheinlichkeit einer infinitesimalen Umgebung von x auffasst, was zumindest im Fall $\nu = \lambda^n$ und stetiger f_θ gerechtfertigt ist, weil in diesem Fall

$$\mathbb{P}_\theta(X \in \mathbb{B}_\varepsilon(x)) \;\approx\; f_\theta(x)\,\lambda^n(\mathbb{B}_\varepsilon(x))$$

für kleine $\varepsilon > 0$ gilt, wobei $\mathbb{B}_\varepsilon(x)$ die n-dimensionale ε-Kugel um x bezeichnet. Wir definieren nun weiter

Definition 2.6. Existiert zu gegebenem Beobachtungswert $x \in \mathfrak{X}$ ein Parameterwert $\widehat{\theta}(x) \in \Theta$, in dem $\mathbf{L}(\cdot|x)$ sein Maximum annimmt, so heißt $\widehat{\theta}(x)$ *Maximum-Likelihood-Schätzer (MLS) für θ auf der Basis von x*. Existiert $\widehat{\theta}(x)$ für alle oder zumindest für $(W_\theta)_{\theta\in\Theta}$-fast alle $x \in \mathfrak{X}$, so heißt die Funktion $\widehat{\theta}$ einfach *Maximum-Likelihood-Schätzer für θ*.

Wir unterstellen in dieser letzten Definition implizit, dass $\widehat{\theta}$ als Abbildung von $\mathfrak{X}$ nach $\Theta \subset \mathbb{R}^d$ messbar ist, sofern über Θ die Borelsche Spur-σ-Algebra $\mathscr{B}(\mathbb{R}^d)_\Theta$ zugrundegelegt wird.

Es spielt selbstverständlich für das Ergebnis keine Rolle, ob wir die Likelihood-Funktion $\mathbf{L}(\theta|x)$ oder die Log-Likelihood-Funktion $\ell(\theta|x)$ maximieren. Mathematisch ist häufig die Funktion $\ell(\theta|x)$ leichter zu analysieren, und zwar vor allem dann, wenn X aus n unabhängigen und identisch verteilten Komponenten $X_1,...,X_n$ besteht, wobei $(\mathbb{P}_\theta^{X_1})_{\theta\in\Theta} \ll \rho$ für ein σ-endliches Maß ρ gilt. In diesem Fall ergibt sich nämlich mit $g_\theta = d\mathbb{P}_\theta^{X_1}/d\rho$ und $f_\theta = d\mathbb{P}_\theta^{X}/d\rho^n$

$$\mathbf{L}(\theta|x) = \prod_{j=1}^{n} g_\theta(x_j), \quad x = (x_1,...,x_n),$$

und folglich

$$\ell(\theta|x) = \sum_{j=1}^{n} \log g_\theta(x_j).$$

Wird nun das Maximum mittels Differentiation nach θ bestimmt, erweist sich dies bei einer Summe meistens leichter als bei einem Produkt.

Betrachten wir als nächstes wieder eine Reihe von Beispielen.

Beispiel 2.7. (Normalverteilung) Gegeben sei die Situation von Beispiel 2.2 mit unabhängigen, $Normal(\mu,\sigma^2)$-verteilten $X_1,...,X_n$. Nehmen wir zunächst an, dass σ^2 *bekannt* ist, so dass nur μ als unbekannter Parameter auftritt. Als Log-Likelihood-Funktion erhalten wir dann

$$\ell(\mu|x) \;=\; -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2, \tag{2.2}$$

deren Ableitung durch

$$\frac{d}{d\mu}\ell(\mu|x) \;=\; \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j-\mu)$$

gegeben ist. Setzen wir diese gleich 0, so ergibt sich $\widehat{\mu}(x) = \bar{x}_n = (x_1+...+x_n)/n$ als (eindeutige) Lösung der *Log-Likelihood-Gleichung* $\ell(\mu|x) = 0$ und damit als MLS für μ, denn $\frac{d^2}{d\mu^2}\ell(\bar{x}_n|x) = -n/\sigma^2 < 0$.

Vertauschen wir die Rollen, indem wir μ *als bekannt und* σ^2 *als unbekannt* annehmen, so ergibt sich natürlich dieselbe Log-Likelihood-Funktion wie in (2.2), nur dieses Mal mit $\sigma^2 > 0$ als Argument. In diesem Fall schreiben wir also $\ell(\sigma|x)$. Die Log-Likelihood-Gleichung nimmt dann die Form

$$\frac{d}{d\sigma^2}\ell(\sigma|x) \;=\; -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{n}(x_j-\mu)^2 \;=\; 0$$

an und liefert den MLS $\widehat{\sigma}^2(x) = n^{-1}\sum_{j=1}^{n}(x_j-\mu)^2$, wie man leicht nachrechnet.

Kommen wir schießlich zu dem Regelfall, dass *sowohl* μ *als auch* σ^2 *unbekannte* Parameter bilden. Die Log-Likelihood-Funktion in (2.2), nun $\ell(\mu,\sigma^2|x)$,

ist also in (μ,σ^2) zu maximieren. Dazu betrachten wir die partiellen Ableitungen $\frac{\partial}{\partial\mu}\ell(\mu,\sigma^2|x)$ und $\frac{\partial}{\partial\sigma^2}\ell(\mu,\sigma^2|x)$, genauer die Log-Likelihood-Gleichungen

$$\begin{aligned}\frac{\partial}{\partial\mu}\ell(\mu,\sigma^2|x) &= \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j-\mu) = 0,\\ \frac{\partial}{\partial\sigma^2}\ell(\mu,\sigma^2|x) &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{j=1}^{n}(x_j-\mu)^2 = 0,\end{aligned}$$

und erhalten wie zuvor $\widehat{\mu}(x)=\bar{x}_n$ als MLS für μ und $\widehat{\sigma}^2(x)=\mathsf{v}_n^2$ als MLS für σ^2, wobei v_n^2 wie in 2.3 definiert ist. Auf den Nachweis, dass die Jacobi-Matrix der zweiten partiellen Ableitungen von $\ell(\mu,\sigma|x)$ an der Stelle $(\bar{x}_n,\mathsf{v}_n^2)$ negativ definit ist, haben wir verzichtet. In der vorliegenden Situation stimmen also MLS und MMS überein.

Beispiel 2.8. (Poisson-Verteilungen) Im Fall n stochastisch unabhängiger, jeweils *Poisson*(θ)-verteilter Beobachtungen mit unbekanntem Parameter $\theta\in(0,\infty)$ gilt für $x=(x_1,\ldots,x_n)\in\mathbb{N}_0^n$

$$\mathbf{L}(\theta|x) = e^{-n\theta}\frac{\theta^{n\bar{x}_n}}{x_1!\cdot\ldots\cdot x_n!}, \quad \text{d.h.} \quad \ell(\theta|x) = -n\theta+n\bar{x}_n\log\theta-\sum_{j=1}^{n}\log x_j!.$$

Differenziert man wieder $\ell(\theta,x)$ nach θ, so ergibt sich leicht $\widehat{\theta}(x)=\bar{x}_n$ als MLS für θ. Dieser stimmt erneut mit dem MMS überein. Der Leser beachte, dass $\widehat{\theta}(x)$ im Fall $x=(0,\ldots,0)$ nur dann zu einem zulässigen Schätzwert führt, wenn wir den Parameterraum Θ um den entarteten Parameter $\theta=0$ erweitern, wobei *Poisson*$(0):=\delta_0$.

Beispiel 2.9. (Binomial-Verteilungen) In der Situation von Beispiel 2.4 sei $m\in\mathbb{N}$ bekannt und $p\in(0,1)$ unbekannt. Dann gilt für $x=(x_1,\ldots,x_n)\in\{0,\ldots,m\}^n$

$$\mathbf{L}(p|x) = \prod_{j=1}^{n}\left(\binom{m}{x_j}p^{x_j}(1-p)^{m-x_j}\right) = \left(\prod_{j=1}^{n}\binom{m}{x_j}\right)(1-p)^{mn}\left(\frac{p}{1-p}\right)^{n\bar{x}_n}$$

und damit

$$\ell(p|x) = \left(\sum_{j=1}^{n}\log\binom{m}{x_j}\right)+mn\log(1-p)+n\bar{x}_n\log p-n\bar{x}_n\log(1-p).$$

Das gewohnte Vorgehen liefert $\widehat{p}(x)=\bar{x}_n/m$ als MLS für p. Dabei sind $\widehat{p}(0,\ldots,0)=0$ und $\widehat{p}(m,\ldots,m)=1$ nur dann zulässige Schätzwerte, wenn wir den Parameterraum um $p=0$ und $p=1$ erweitern, wobei $Bin(m,0):=\delta_0$ und $Bin(m,1):=\delta_m$.

Nehmen wir als nächstes an, dass p bekannt und m unbekannt ist, so erhalten wir dieselbe Likelihood-Funktion wie vorher, jedoch mit m anstelle von p als Argument, wobei $\binom{m}{x}:=0$ für $x\notin\{0,\ldots,m\}$. In diesem Fall liegt ein diskretes Maximierungsproblem vor, da $m\in\mathbb{N}$, und wir notieren als erstes, dass $\mathbf{L}(m|x)=0$ für $m<\max_{1\le j\le n}x_j$. Für $m\ge\max_{1\le j\le n}x_j$ gilt offensichtlich

$$\frac{\mathbf{L}(m-1|x)}{\mathbf{L}(m|x)} \;=\; \frac{\prod_{j=1}^{n}(m-x_j)}{(m(1-p))^n} \;=\; (1-p)^{-n}\prod_{j=1}^{n}\left(1-\frac{x_j}{-m}\right).$$

Da diese Funktion streng monoton wachsend in m ist, lautet die Bedingung an den MLS $\widehat{m}(x)$

$$\frac{\mathbf{L}(\widehat{m}-1|x)}{\mathbf{L}(\widehat{m}|x)} \;<\; 1 \quad \text{und} \quad \frac{\mathbf{L}(\widehat{m}|x)}{\mathbf{L}(\widehat{m}+1|x)} \;\geq\; 1,$$

was äquivalent ist zu

$$\prod_{j=1}^{n}\left(1-\frac{x_j}{\widehat{m}}\right) \;<\; (1-p)^n \quad \text{und} \quad \prod_{j=1}^{n}\left(1-\frac{x_j}{\widehat{m}+1}\right) \;\geq\; (1-p)^n.$$

Man kann dieses Maximierungsproblem lösen, indem man beispielsweise numerisch die eindeutige Nullstelle $\widehat{z}$ des Polynoms

$$f_{x,p}(z) \;=\; \prod_{j=1}^{n}(1-x_j z)-(1-p)^n$$

in $[0,1/\max_{1\leq j\leq n}x_j]$ bestimmt. Der MLS $\widehat{m}(x)$ ist dann die größte ganze Zahl $\leq 1/\widehat{z}$.

Kommen wir schließlich zu dem Fall, dass sowohl m als auch p unbekannte Parameter bilden. Differentiation der Log-Likelihood-Funktion nach p ergibt analog zum eingangs betrachteten Fall (m bekannt), dass $\ell(p,m|x)$ für jedes feste $m \geq \max_{1\leq j\leq n}x_j$ an der Stelle $p_m(x)=\overline{x}_n/m$ ihr Maximum annimmt, was insbesondere die Beziehung $\widehat{p}(x)=\overline{x}_n/\widehat{m}$ liefert. Es bleibt also lediglich $\widehat{m}$ zu bestimmen, wozu wir gleich voraussetzen, dass nicht alle x_j denselben Wert haben, da sich andernfalls offenbar $\widehat{p}(x)=1$ und $\widehat{m}(x)=x_1$ als MLS ergeben. Statt jedoch den Quotienten wie zuvor zu betrachten, ist es hier günstiger, den Parameter m als kontinuierlich anzusehen, die partielle Ableitung $\frac{\partial}{\partial m}\ell(p,m|x)$ zu bestimmen und schließlich eine Nullstelle von $\frac{\partial}{\partial m}\ell(p_m,m|x)$ etwa mittels des Newton-Verfahrens zu berechnen. Natürlich bleibt dann noch zu zeigen, dass diese tatsächlich einen MLS für m bildet. Wir wollen darauf nicht weiter eingehen, erwähnen allerdings noch eine Arbeit aus dem Jahr 1981 von OLKIN, PETKAU & ZIDEK [12], die gezeigt haben, dass der MLS $\widehat{m}$ sehr instabil sein kann. Zur Illustration betrachten sie zwei Datensätze von fünf Beobachtungen einer $Bin(m,p)$-Verteilung, deren Parameter m und p beide unbekannt sind. Der erste Datensatz lautet (16, 18, 22, 25, 27), der zweite (16, 18, 22, 25, 28) Der Unterschied zwischen beiden besteht als lediglich darin, dass beim zweiten als letzte Beobachtung 27 durch 28 ersetzt wurde. Für den ersten Datensatz ergibt sich als MLS $\widehat{m}=99$, für den zweiten dagegen $\widehat{m}=190$. Der Grund für diesen gewaltigen Unterschied liegt darin, dass die Likelihood-Funktion in einem sehr großen, den MLS enthaltenden Intervall nahezu konstant ist, was zur Folge hat, dass schon eine geringfügige Änderung ihres Verlaufs durch Veränderung der Daten die Maximalstelle erheblich verschiebt.

Eine nützliche Eigenschaft von MLS ist ihre Invarianz unter Parametertransformationen. Zur Präzision dieser Feststellung sei $\gamma : \Theta \to \Theta'$ eine Parameterfunktion. Zum Schätzen von $\gamma(\theta)$ (bei Beobachtung von $X = x$) mit Hilfe der Maximum-Likelihood-Methode definieren wir zunächst die von γ *induzierte Likelihood-Funktion* $\mathbf{L}_\gamma(\cdot|x) : \gamma(\Theta) \to [0,\infty)$ durch

$$\mathbf{L}_\gamma(\eta|x) \;=\; \sup_{\theta\in\Theta:\gamma(\theta)=\eta} \mathbf{L}(\theta|x).$$

Ein Wert $\widehat{\eta}(x) \in \gamma(\Theta)$ heißt MLS für $\eta = \gamma(\theta)$ zur Beobachtung $x \in \mathfrak{X}$, falls dieser $\mathbf{L}_\gamma(\cdot|x)$ maximiert. Die naheliegende Frage, ob $\gamma(\widehat{\theta}(x))$ einen MLS für η gegeben x bildet, beantwortet

Satz 2.10. *Seien $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ ein statistisches Experiment, $\gamma : \Theta \to \Theta'$ eine Parameterfunktion und $\widehat{\theta}(x)$ ein MLS für θ zu gegebenem $x \in \mathfrak{X}$. Dann ist $\widehat{\eta}(x) = \gamma(\widehat{\theta}(x))$ ein MLS für $\eta = \gamma(\theta)$ zu x.*

Beweis. Die Behauptung ergibt sich leicht aus der Gleichungskette

$$\begin{aligned}\mathbf{L}_\gamma(\gamma(\widehat{\theta}(x))|x) \;&=\; \max_{\theta\in\Theta:\gamma(\theta)=\gamma(\widehat{\theta}(x))} \mathbf{L}(\theta|x) \;=\; \mathbf{L}(\widehat{\theta}(x)|x) \;=\; \max_{\theta\in\Theta} \mathbf{L}(\theta|x)\\ &=\; \max_{\eta\in\gamma(\Theta)} \max_{\theta:\gamma(\theta)=\eta} \mathbf{L}(\theta|x) \;=\; \max_{\eta\in\gamma(\Theta)} \mathbf{L}_\gamma(\eta|x). \qquad \square\end{aligned}$$

Dieser Satz liefert insbesondere die formale Bestätigung dafür, dass es in Beispiel 2.7 keine Rolle spielt, ob wir den MLS für die Varianz σ^2 durch Maximierung der Likelihood-Funktion in σ^2 bestimmen, wie dort geschehen, oder durch Maximierung in σ und anschließender Quadrierung des Resultats.

2.1.3 Die Bayes-Methode

Beispiel 1.9 hatte gezeigt, dass es in nichttrivialen Situationen aussichtslos ist, einen unter allen Schätzern gleichmäßig besten finden zu wollen, da dieser für jeden Parameter das Risiko 0 besitzen müsste. Der Bayessche Ansatz bietet einen Ausweg aus diesem Dilemma, indem er die Risikofunktion $R(\theta,\delta)$ nicht für jedes θ zu minimieren versucht, sondern lediglich deren Mittelung bezüglich einer *a priori Verteilung* ξ, d.h.

$$B(\xi,\delta) \;:=\; \int_\Theta R(\theta,\delta)\; \xi(d\theta).$$

Ein Schätzer δ mit minimalem Bayes-Risiko $B(\xi,\delta) = B(\xi)$ heißt *Bayes-Schätzer (BS)* bzgl. ξ [☞ auch Definition 1.11].

Das Problem der Berechnung von Bayes-Schätzern taucht in verschiedenen Kontexten auf. Hier geht es uns allerdings nur darum, eine weitere Möglichkeit des

Findens interessanter Schätzer vorzustellen; für eine ausführlichere Diskussion verweisen wir auf die Monographie von LEHMANN [11, S. 336ff]. Die generelle Schwierigkeit besteht offenkundig darin, die a priori Verteilung ξ auszuwählen. Eine Möglichkeit bildet die Auswahl eines plausiblen Elements aus einer mathematisch günstigen und zugleich hinreichend flexiblen parametrischen Verteilungsfamilie. Bevor wir dies anhand von zwei Beispielen demonstrieren, wollen wir zunächst das zugrundegelegte Modell in geeigneter Weise festlegen und das allgemeine Vorgehen zur Berechnung von Bayes-Schätzern beschreiben:

Die Zuordnung $\theta \mapsto W_\theta(A)$ sei messbar für jedes $A \in \mathscr{A}$, eine technische Voraussetzung, die uns keine Kopfschmerzen zu machen braucht, weil sie von den allermeisten gebräuchlichen Verteilungsfamilien erfüllt wird. Wir können dann annehmen, dass der unbekannte Parameter die Realisierung einer weiteren Zufallsvariablen Λ bildet, die zusammen mit der Variablen X auf einem W-Raum $(\Omega, \mathfrak{A}, \mathbb{P})$ definiert ist, wobei

$$\mathbb{P}^\Lambda \;=\; \xi \quad \text{und} \quad \mathbb{P}^{X|\Lambda=\theta} \;=\; W_\theta$$

für alle $\theta \in \Theta$. Man wähle etwa $(\Omega, \mathfrak{A}) = (\Theta \times \mathfrak{X}, \mathscr{B}(\Theta) \otimes \mathscr{A})$, Λ, X als Projektionen auf die erste bzw. zweite Komponente und definiere $\mathbb{P}(A \times B) := \int_A W_\theta(B)\, \xi(d\theta)$. Die unbedingte Verteilung $\mathbb{P}^X$ von X lautet demnach

$$W_\xi(\cdot) \;:=\; \int_\Theta W_\theta(\cdot)\; \xi(d\theta) \;=\; \mathbb{E} W_\Lambda(\cdot).$$

Für das Bayes-Risiko eines beliebigen Schätzers δ unter ξ können wir nun auch schreiben:

$$\begin{aligned} B(\xi, \delta) \;&=\; \int_\Theta R(\theta, \delta)\; \xi(d\theta) \\ &=\; \int_\Theta \int_{\mathfrak{X}} L(\theta, \delta(x))\; W_\theta(dx)\; \xi(d\theta) \\ &=\; \int_\Theta \int_{\mathfrak{X}} L(\theta, \delta(x)) \mathbb{P}^{X|\Lambda=\theta}(dx)\; \mathbb{P}^\Lambda(d\theta) \qquad (2.3) \\ &=\; \int_{\mathfrak{X}} \int_\Theta L(\theta, \delta(x))\; \mathbb{P}^{\Lambda|X=x}(d\theta)\; \mathbb{P}^X(dx) \\ &=\; \int_{\mathfrak{X}} \mathbb{E}(L(\Lambda, \delta(x)) | X = x)\; \mathbb{P}^X(dx). \end{aligned}$$

Die Berechnung eines Bayes-Schätzers ist im Prinzip sehr einfach. Betrachten wir zunächst die Situation *vor* Erhalt einer Realisierung der Beobachtungsvariablen X. Gegeben eine zu schätzende Parameterfunktion $\gamma : \Theta \to \mathbb{R}^m$ und eine hierfür gewählte Verlustfunkion L, ist dann jedes $\widehat{d} \in \mathbb{R}^m$, das

$$d \;\mapsto\; \mathbb{E} L(\Lambda, d) \;=\; \int_\Theta L(\theta, d)\; \xi(d\theta)$$

minimiert, ein BS für $\gamma(\theta)$. *Nach* Erhalt eines Datums x wird die a priori Verteilung ξ durch die sogenannte *a posteriori Verteilung* von Λ, d.h. durch $\mathbb{P}^{\lambda|X=x}$ ersetzt,

und einen BS bildet jedes $\widehat{\delta}(x) \in \mathbb{R}^m$, welches das *a posteriori Risiko*

$$d \mapsto \mathbb{E}(L(\Lambda,d)|X=x) = \int_\Theta L(\theta,d)\, \mathbb{P}^{\Lambda|X=x}(d\theta) \tag{2.4}$$

minimiert [☞ (2.3)]. Mathematisch präzise lautet das Ergebnis:

Satz 2.11. *Neben den zuvor gemachten Voraussetzungen gelte:*

(1) Es gibt einen Schätzer δ_0 mit endlichem Bayes-Risiko.

(2) Es gibt einen Schätzer $\widehat{\delta}$, so dass $\widehat{\delta}(x)$ für W_ξ-fast alle x das a posteriori Risiko (2.4) *minimiert.*

Dann ist $\widehat{\delta}$ ein BS für $\gamma(\theta)$ (vgzl. ξ).

Beweis. Ist δ irgendein Schätzer mit endlichem Bayes-Risiko, so hat dieser auch endliches a posteriori Risiko $\mathbb{E}(L(\Lambda,\delta(x))|X=x)$ für W_ξ-fast alle x, und es folgt vermöge (2)

$$\infty > \mathbb{E}(L(\Lambda,\delta(x))|X=x) \geq \mathbb{E}(L(\Lambda,\widehat{\delta}(x))|X=x) \quad W_\xi\text{-f.s.}$$

Die Behauptung ergibt sich nun durch Integration bzgl. $\mathbb{P}^X$ auf beiden Seiten [☞ (2.3)]. □

Voraussetzung (1) sichert lediglich, dass jeder BS eine endliche Risikofunktion besitzt. Im speziellen Fall der quadratischen Verlustfunktion $L(\theta,d) = (\gamma(\theta)-d)^2$ ergibt sich:

Korollar 2.12. *Gegeben die Voraussetzung von Satz 2.11, bildet*

$$\widehat{\delta}(x) = \mathbb{E}(\gamma(\Lambda)|X=x)$$

einen BS für $\gamma(\theta)$ unter quadratischer Verlustfunktion.

Beweis. Nach Satz 2.11 ergibt sich $\widehat{\delta}$ durch minimierung von

$$d \mapsto \mathbb{E}((\gamma(\Lambda)-d)^2|X=x)$$

und damit die Behauptung als definierende Eigenschaft des bedingten Erwartungswerts [☞ [2, Satz 51.1]]. □

Eine in vielen Situationen interessante Frage lautet, wann ein BS eindeutig bestimmt ist. Diesbetreffend notieren wir:

Korollar 2.13. *Gegeben seien die Voraussetzungen von Satz 2.11 und eine in d strikt konvexe Verlustfunktion $L(\theta,d)$ $\left[z.B.\ L(\theta,d)=(\gamma(\theta)-d)^2\right]$. Dann ist der BS $\widehat{\delta}$ $\mathscr{W}$-f.s eindeutig bestimmt, falls $\mathscr{W}=(W_\theta)_{\theta\in\Theta}\ll W_\xi$.*

Beweis. Ist $L(\theta,\cdot)$ strikt konvex, gilt offenbar dasselbe für $\mathbb{E}(L(\Lambda,\cdot)|X=x)$ auf $I_x:=\{d:\mathbb{E}(L(\Lambda,d)|X=x)<\infty\}$. $\widehat{\delta}(x)$ bildet deshalb die eindeutige Minimalstelle dieser Funktion, sofern $I_x\neq\emptyset$. Nach den Voraussetzungen in Satz 2.11 (unter Hinweis auf $\mathbb{E}L(\Lambda,\widehat{\delta}(X))<\infty$) ist letzteres aber W_ξ-f.s. und somit $\mathscr{W}$-f.s. der Fall. □

In allen drei weiter unten behandelten Beispielen sind die Voraussetzungen des Korollars erfüllt und der jeweilige BS deshalb eindeutig. Für die notwendige Berechnung der a posteriori Verteilung $\mathbb{P}^{\Lambda|X=x}$ erweist sich dort, wie in vielen anderen Beispielen, der folgende Satz als nützlich, der im Grunde schon aus der W-Theorie bekannt ist:

Satz 2.14. *Sei μ ein σ-endliches Maß und ξ eine a priori Verteilung mit μ-Dichte g. Sei ferner $(W_\theta)_{\theta\in\Theta}$ eine Verteilungsfamilie auf $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ und ebenfalls dominiert durch ein σ-endliches Maß ν. Schließlich sei $f_\theta:=dW_\theta/d\nu$ und $f_\xi(x):=\int_\Theta f_\theta(x)\xi(d\theta)$. Dann bildet*

$$f^{\Lambda|X=x}(\theta)\ :=\ \frac{g(\theta)f_\theta(x)}{f_\xi(x)}\mathbf{1}_{\{f_\xi>0\}}+g(\theta)\mathbf{1}_{\{f_\xi=0\}} \tag{2.5}$$

eine μ-Dichte der a posteriori Verteilung $\mathbb{P}^{\Lambda|X=x}$ für W_ξ-fast alle x.

Beweis. Da X Werte in $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ annimmt und $\mathscr{B}(\mathbb{R}^n)$ abzählbar erzeugt wird, können wir nach Satz 53.13 in [2] die f_θ so wählen, dass die Abbildung $(\theta,x)\mapsto f_\theta(x)$ messbar ist. Damit bildet aber $g(\theta)f_\theta(x)$ eine $\mu\otimes\nu$-Dichte von (Λ,X) und f_ξ eine ν-dichte von X, wie man sofort nachprüft. Es folgt die Behauptung gemäß Satz 53.11 in [2]. □

Beispiel 2.15. (Bernoulli-Verteilungen) Die Beobachtungsvariable X bestehe aus n unabhängigen identisch $Bern(\theta)$-verteilten Komponenten $X_1,...,X_n$ mit unbekanntem Parameter $\theta\in\Theta:=(0,1)$ Eine mathematisch günstige und zugleich flexible Familie von a priori Verteilungen für θ bilden die *Betaverteilungen* $\beta(a,b)$, $(a,b)\in(0,\infty)^2$, mit λ-Dichten

$$g_{a,b}(x)\ :=\ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}\mathbf{1}_{(0,1)}(x).$$

Abb. 2.1 illustriert dies anhand sechs ausgewählter Parameterkombinationen. Beachte, dass $\beta(1,1)=Unif(0,1)$ gilt. Wir notieren außerdem, dass

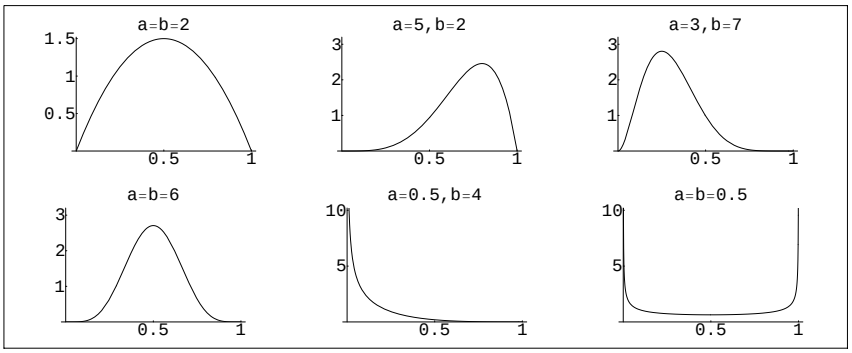


Abb. 2.1 Dichten der $\beta(a,b)$-Verteilung

$$\mathbb{E}\beta(a,b) = \frac{a}{a+b} \qquad \text{und} \qquad \mathbb{V}\mathrm{ar}\,\beta(a,b) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{2.6}$$

Um den BS $\widehat{\theta}$ für θ bei quadratischer Verlustfunktion zu bestimmen, müssen wir zunächst die a posteriori Verteilung $\mathbb{P}^{\Lambda|X=x}$ berechnen, wobei $x = (x_1,...,x_n) \in \{0,1\}^n$ und $\Lambda \stackrel{d}{=} \beta(a,b)$ für irgendeine feste Parameterkombination $(a,b) \in (0,\infty)^2$. Wir bedienen uns des zuvor gezeigten Satzes 2.14 mit $\mu = \lambda$ und $\nu =$ Zählmaß auf $\{0,1\}^n$. Es gilt

$$f_\theta(x) \;=\; \mathbb{P}_\theta(X_1 = x_1,...,X_n = x_n) = \theta^s(1-\theta)^{n-s}$$

für alle $\theta \in (0,1)$ und $x = (x_1,...,x_n) \in \{0,1\}^n$, wobei $s := \sum_{j=1}^n x_j$. Hieraus folgt offenbar $f_\xi > 0$ auf $\{0,1\}^n$, und wir erhalten vermöge (2.5) als bedingte λ-Dichte

$$f^{\Lambda|X=x}(\theta) \;=\; C(a,b,x)\,\theta^{a+s-1}(1-\theta)^{b+n-s-1}\mathbf{1}_{(0,1)}(\theta),$$

also wieder eine Betaverteilung, nämlich $\beta(a+s,b+n-s)$, als a posteriori Verteilung. Für die auftretende Konstante ergibt sich damit natürlich ohne weitere Rechnung

$$C(a,b,x) \;=\; \frac{\Gamma(a+b+n)}{\Gamma(a+s)\Gamma(b+n-s)}$$

aus der Normierung. Gemäß Korollar 2.12 und mit (2.6) folgt als BS für $\gamma(\theta) = \theta$ bei quadratischer Verlustfunktion

$$\widehat{\theta}(x) \;=\; \mathbb{E}(\Lambda|X=x) \;=\; \mathbb{E}\beta(a+s,b+n-s) \;=\; \frac{a+s}{a+b+n},$$

den wir wegen $\overline{x} = s/n$ auch in der Form

$$\widehat{\theta}(x) \;=\; \left(\frac{a+b}{a+b+n}\right)\frac{a}{a+b} + \left(\frac{n}{a+b+n}\right)\overline{x} \tag{2.7}$$

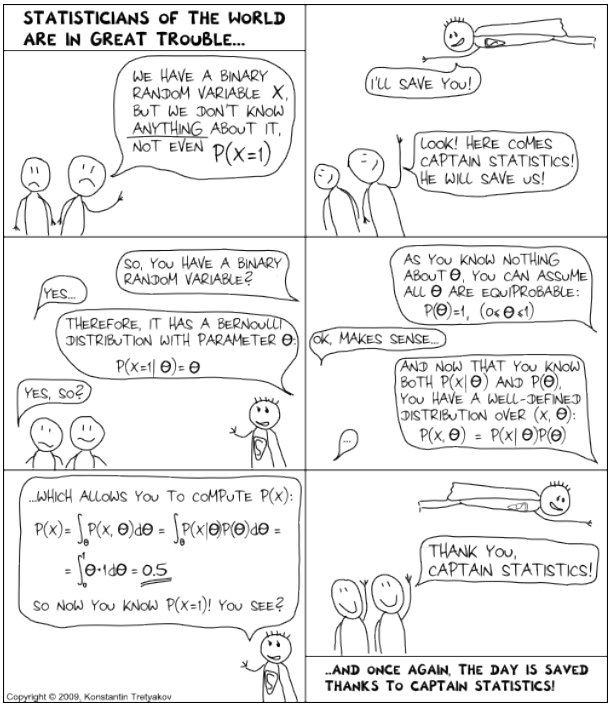


Abb. 2.2 Quelle: http://stats.stackexchange.com

schreiben können. Der BS ergibt sich demnach als gewichtetes Mittel von $\frac{a}{a+b}$, dem a priori Schätzer für θ vor Erhalt der Daten, und dem kanonischen Schätzer $\bar{x}$ für den Mittelwert θ ohne Benutzung einer a priori Verteilung. Für große n ist hierbei offenkundig der "a priori Anteil" klein, was auch der Intuition entspricht, denn mit wachsender Zahl von Beobachtungen wird man den Daten mehr und mehr Gewicht geben und das a priori Mittel $\frac{a}{a+b}$ gleichsam korrigieren. Setzen wir $a = b = 0$ in (2.7), ist $\bar{x}$ sogar die exakte Bayes-Lösung des Problems, die zugehörige Vorbewertung mit λ-Dichte $\theta^{-1}(1-\theta)^{-1}\mathbf{1}_{(0,1)}(\theta)$ jedoch kein W-Maß mehr. Man spricht dann von einer *verallgemeinerten Vorbewertung* (engl. *improper prior*). Abschließend sei noch erwähnt, dass sich derselbe BS im Fall einer $Bin(n,\theta)$-verteilten Beobachtung ergibt, d.h., wenn $X = \sum_{i=1}^{n} X_i$.

Beispiel 2.16. (Normalverteilungen, unbekannter Mittelwert) Gegeben n unabhängige, jeweils $Normal(\mu,\sigma^2)$-verteilte Zuallsgrößen, sei $\sigma^2 > 0$ *bekannt* und der Mittelwert $\mu \in \mathbb{R}$ der zu schätzende Parameter. Als günstige Familie von a priori Verteilungen erweisen sich hier Normalverteilungen. Falls $\Lambda \stackrel{d}{=} Normal(\nu,\tau^2)$, erhalten wir unter erneuter Benutzung von Satz 2.14 nach einiger Rechnung

$$\mathbb{P}^{\Lambda|X=x} \;=\; N\left(\frac{n\bar{x}/\sigma^2+\nu/\tau^2}{n/\sigma^2+1/\tau^2}, \frac{1}{n/\sigma^2+1/\tau^2}\right).$$

Die Details überlassen wir dem Leser zur Übung. Bei quadratischer Verlustfunktion ergibt sich als BS für μ

$$\widehat{\mu}(x) \;=\; \left(\frac{1/\tau^2}{n/\sigma^2+1/\tau^2}\right)\nu \;+\; \left(\frac{n/\sigma^2}{n/\sigma^2+1/\tau^2}\right)\bar{x},$$

d.h. analog zum vorherigen Beispiel ein gewichtetes Mittel des kanonischen Schätzers $\bar{x}$ und dem a priori-Schätzer ν.

Beispiel 2.17. (Normalverteilungen, unbekannte Varianz) Sei nun μ bekannt, o.E. $\mu = 0$ (andernfalls gehe zum Beobachtungsvektor $X - \mu = (X_1 - \mu, ..., X_n - \mu)$ über), und $\sigma^2 > 0$ der unbekannte Parameter. Auch in dieser Situation lässt sich eine günstige Klasse von a priori Verteilungen angeben. Nehmen wir an, dass

$$\Lambda' \;:=\; \frac{1}{2\Lambda} \;\stackrel{d}{=}\; \Gamma(\alpha,\beta)$$

für ein $(\alpha,\beta) \in (1,\infty)\times(0,\infty)$. Mit anderen Worten: Wir betrachten zunächst eine wohlbekannte Vorbewertung für den Parameter $1/2\sigma^2$. Dann gilt mit der Bezeichnung $g_{\alpha,\beta}$ für die Dichte der $\Gamma(\alpha,\beta)$-Verteilung und der Beziehung $\frac{1}{2y}g_{\alpha,\beta}(y) = \frac{\beta}{2(\alpha-1)}g_{\alpha-1,\beta}(y)$

$$\mathbb{E}(\Lambda) \;=\; \mathbb{E}\left(\frac{1}{2\Lambda'}\right) \;=\; \frac{\beta}{2(\alpha-1)}\int_{-\infty}^{\infty} g_{\alpha-1,\beta}(y)\,dy \;=\; \frac{\beta}{2(\alpha-1)}.$$

Als a posteriori Verteilung für Λ' ergibt sich unter Beachtung von $\mu = 0$

$$\mathbb{P}^{\Lambda'|X=x} \;=\; \Gamma(\alpha+n/2, \beta+ns^2), \quad s^2 := \frac{1}{n}\sum_{j=1}^{n} x_j^2. \tag{2.8}$$

Bei quadratischer Verlustfunktion ist deshalb gemäß Korollar 2.12 und (2.8)

$$\begin{aligned}\widehat{\sigma}^2(x) \;&=\; \mathbb{E}\left(\frac{1}{2\Lambda'}\middle| X = x\right) \;=\; \frac{\beta+s^2}{2\alpha+n-2}\\ &=\; \left(\frac{2\alpha-2}{2\alpha+n-2}\right)\frac{\beta}{2\alpha-2} \;+\; \left(\frac{n}{2\alpha+n-2}\right)s^2\end{aligned}$$

der BS für σ^2 und einmal mehr eine gewichtetes Mittel des a priori Schätzers $\frac{\beta}{2(\alpha-1)}$ und des kanonischen Schätzers s^2 [gemäß 2.7 der MLS für σ^2 im Fall $\mu = 0$].

In jedem der drei diskutierten Beispiele hatten wir, zur gegebenen Verteilungsklasse $\mathscr{W}$ für den Beobachtungsvektor X, die Klasse der a priori Verteilungen $\mathscr{V}$ so gewählt, dass die a posteriori Verteilungen wieder zur selben Klasse $\mathscr{V}$ gehören. Man nennt $\mathscr{V}$ in diesem Fall die zu $\mathscr{W}$ *konjugierte Klasse* von a priori Verteilungen, deren mathematische Definition jedoch noch einiger Sorgfalt bedürfte. Wir gehen darauf nicht weiter ein.

2.2 Datenreduktion und Suffizienz

Betrachten wir wieder ein statisisches Experiment $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ und nehmen zur Motivation des Folgenden an, dass der Stichprobenraum $\mathfrak{X}$ von hoher Dimension ist. In diesem Fall besteht zweifellos ein Interesse darin, irrelevante Informationen, die in den Daten vorhanden sind, auszusondern (Einsparung von Speicherplatz, bessere Berechenbarkeit). Man denke etwa an die Situation in Beispiel 1.1, in der die Stichprobe aus einem Vektor $x = (x_1, ..., x_{100}) \in \{0,1\}^{100}$ mit 100 Komponenten besteht und sich die Frage stellt, ob nicht die Erhebung von $s_{100} = x_1 + ... + x_{100}$ zu ebenso guten statistischen Entscheidungen führt. Was unter einer derartigen statistisch gleichwertigen *Datenreduktion* zu verstehen ist, soll in diesem Abschnitt formalisiert und damit präzisiert werden. Wir erinnern daran, dass jede messbare Abbildung T auf $(\mathfrak{X}, \mathscr{A})$ *Statistik* genannt wird.

Definition 2.18. Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ ein statistisches Experiment und $T : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}', \mathscr{A}')$ eine Statistik. Als *durch T reduziertes Experiment* bezeichnet man das Tripel

$$\mathscr{E}^T \;=\; (\mathfrak{X}', \mathscr{A}', (W_\theta^T)_{\theta\in\Theta}).$$

Der Definition liegt folgende Interpretation zugrunde: Benutzt der Statistiker in der Untersuchung eines Experiments eine Statistik T, so zieht er zur Entscheidungsfindung nicht den beobachteten Wert x heran, der eventuell eine Vielzahl irrelevanter Informationen enthält, sondern den "reduzierten Wert" $T(x) = t$, im zuvor genannten Beispiel 1.1 etwa $T(x_1, ..., x_{100}) = \sum_{j=1}^{100} x_j$. Natürlich sollte eine Datenreduktion so durchgeführt werden, *dass keine wesentlichen Informationen verloren gehen*. Um zu einer mathematischen Präzisierung dieser Forderung in unserem Modell zu gelangen, stellen wir zunächst folgende Überlegung an: Sei eine Familie $(\mathbb{P}_\theta)_{\theta\in\Theta}$ von W-Maßen auf einem messbaren Raum $(\Omega, \mathfrak{A})$ gegeben. Um Aussagen über den unbekannten Parameter zu gewinnen, entwerfen zwei Statistiker jeweils Experimente und beobachten Realisierungen der Zufallsvariablen X bzw. X' aus *demselben* Stichprobenraum $(\mathfrak{X}, \mathscr{A})$. Sie gelangen somit zu statistischen Experimenten

$$(\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta}) \quad \text{mit} \quad W_\theta \;=\; \mathbb{P}_\theta^X$$

bzw.

$$(\mathfrak{X}, \mathscr{A}, (W'_\theta)_{\theta\in\Theta}) \quad \text{mit} \quad W'_\theta \;=\; \mathbb{P}_\theta^{X'}.$$

Vom Standpunkt der statistischen Entscheidungstheorie sind beide Vorgehensweisen äquivalent, wenn $W_\theta = W'_\theta$ für alle $\theta \in \Theta$ gilt, weil in diesem Fall die resultierenden statistischen Experimente übereinstimmen. Betrachten wir vor diesem Hintergrund das Problem der Datenreduktion: Die Beobachtung des ersten Statistikers werde wieder repräsentiert durch die Zufallsvariable X, d.h. $W_\theta = \mathbb{P}_\theta^X$. Typischerweise nimmt X Werte im $\mathbb{R}^n$ an, $n \geq 2$, so dass W_θ ein W-Maß auf $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ bildet. Eine Datenreduktion hat häufig das Ziel, eine niedrigere Dimensionalität zu

erreichen, indem eine Statistik $T : \mathbb{R}^n \to \mathbb{R}^k$, $k < n$, benutzt wird. Die Beobachtung in dem durch T reduzierten Experiment wird dann durch $T \circ X$ repräsentiert und hat die Verteilung $\mathbb{P}_\theta^{T\circ X} = W_\theta^T$. Letztere bildet ein W-Maß auf $(\mathbb{R}^k, \mathscr{B}(\mathbb{R}^k))$, und es gilt $W_\theta^T \neq W_\theta$.

Wir werden nun eine Datenreduktion als *suffizient* bezeichnen, falls der zweite Statistiker unabhängig vom Ausgangsexperiment, unter bloßer Benutzung der reduzierten Daten einen Zufallsmechanismus derart konstruieren kann (z.B. mit einem Zufallszahlengenerator, beschrieben durch die Zufallsvariable X'), dass $\mathbb{P}_\theta^{X'} = W_\theta$ für alle $\theta \in \Theta$ gilt. Dabei ist zu beachten, dass dieser Mechanismus *nicht mehr vom Parameter* θ, der ja unbekannt ist, abhängen darf. Dies führt zu folgender

Definition 2.19. Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ ein statistisches Experiment. Eine Statistik $T : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}', \mathscr{A}')$ heißt *suffizient (für $\mathscr{E}$ oder $(W_\theta)_{\theta\in\Theta}$)*, falls für jedes $A \in \mathscr{A}$ eine messbare Abbildung $W(A|T = \cdot) : \mathfrak{X}' \to [0,1]$ existiert, so dass $W(A|T = \cdot)$ für jedes $\theta \in \Theta$ eine Version der faktorisierten bedingten Wahrscheinlichkeit $W_\theta(A|T = \cdot)$ bildet, d.h., wenn

$$W_\theta(A \cap T^{-1}(A')) \;=\; \int_{A'} W(A|T = t)\, W_\theta^T(dt)$$

für alle $A \in \mathscr{A}$, $A' \in \mathscr{A}'$ und $\theta \in \Theta$ gilt.

Die Suffizienz einer Statistik T bedeutet also, dass die bedingte Wahrscheinlichkeit von $X \in A$ gegeben $T(X) = t$ unter jedem $\mathbb{P}_\theta$ dieselbe ist, d.h. von θ nicht mehr abhängt.

Anmerkung 2.20. Äquivalent zu der Existenz einer von θ unabhängigen Version der *faktorisierten* bedingten Wahrscheinlichkeit $W(A|T = \cdot)$ ist natürlich die Existenz einer von θ unabhängigen Version des nicht faktorisierten Pendants $W(A|T) = W(A|\sigma(T))$ für jedes $A \in \mathscr{A}$. Demzufolge kann man auch die Suffizienz für σ-Algebren definieren:

Eine Unter-σ-Algebra $\mathscr{F}$ von $\mathscr{A}$ heißt *suffizient*, falls es für jedes $A \in \mathscr{A}$ eine von θ unabhängige Version $W(A|\mathscr{F})$ der bedingten Wahrscheinlichkeit von A gegeben $\mathscr{F}$ gibt.

Anmerkung 2.21. Im Fall eines polnischen Raums $\mathfrak{X}$, insbesondere also im Fall $\mathfrak{X} = \mathbb{R}^n$, kann $W(\cdot|T = \cdot)$ als stochastischer Kern gewählt werden. Der Beweis verläuft ähnlich dem von Satz 53.4 in [2], wobei wir auf Witting [18], Satz 3.15 auf S. 341, verweisen. Aus

$$W_\theta(A \cap T^{-1}(A')) \;=\; \mathbb{P}_\theta(X \in A, T(X) \in A')$$

für alle $A \in \mathscr{A}$ und $A' \in \mathscr{A}'$ folgt dann die Beziehung

$$W(\cdot|T=\cdot) \;=\; \mathbb{P}_\theta^{X|T(X)=\cdot} \quad W_\theta^T\text{-f.s.}$$

für alle $\theta \in \Theta$. Suffizienz der Statistik T kann also in diesem Fall auch dadurch ausgedrückt werden, dass eine von θ unabhängige Version der bedingten Verteilung von X gegeben $T(X)$ existiert.

Anmerkung 2.22. Kehren wir noch einmal zu den Überlegungen vor Definition 2.19 zurück. Der zweite Statistiker, dem nur der Wert t der Statistik T als Beobachtung gegeben ist, kann nun – zumindest theoretisch – folgende Vorgehensweise wählen: Er führt ein weiteres Zufallsexperiment auf $(\mathfrak{X}, \mathscr{A})$ mit zugrundeliegender bedingter Verteilung $W(\cdot|T=t)$ durch. Dessen Ausgang wird durch eine Zufallsvariable $X' : (\Omega, \mathfrak{A}) \to (\mathfrak{X}, \mathscr{A})$ repräsentiert, wobei

$$\mathbb{P}_\theta^{X'|T(X)=t} \;=\; W(\cdot|T=t)$$

für alle $\theta \in \Theta$ gelte. Dann sind aber $(\mathfrak{X}, \mathscr{A}, (\mathbb{P}_\theta^X)_{\theta\in\Theta})$ und $(\mathfrak{X}, \mathscr{A}, (\mathbb{P}_\theta^{X'})_{\theta\in\Theta})$ in der Tat äquivalente Experimente, denn

$$\begin{aligned}\mathbb{P}_\theta(X' \in A) \;&=\; \int_{\mathfrak{X}'} \mathbb{P}_\theta^{X'|T(X)=t}(A)\, \mathbb{P}_\theta^{T(X)} \\ &=\; \int_{\mathfrak{X}'} W(A|T=t)\, W_\theta^T(dt) \;=\; W_\theta(A) \;=\; \mathbb{P}_\theta(X \in A)\end{aligned}$$

für alle $\theta \in \Theta$ und $A \in \mathscr{A}$.

Anmerkung 2.23. Im Folgenden benutzen wir $W_\theta(\cdot|T=\cdot)$ sowohl für die bedingte Verteilung als auch für den zugehörigen bedingten Erwartungswert. Gegeben eine unter W_θ quasi-integrierbare Funktion $g : \mathfrak{X} \to \overline{\mathbb{R}}$, gelte also

$$W_\theta(g|T=t) := \int g(x)\, W_\theta(dx|T=t).$$

Man erhält in der üblichen Weise unter Benutzung des Funktions-Erweiterungsarguments: Ist $g : \mathfrak{X} \to \overline{\mathbb{R}}$ eine unter jedem W_θ quasi-integrierbare Abbildung, d.h. $\int g\, dW_\theta = \mathbb{E}_\theta g(X)$ existiert für jedes $\theta \in \Theta$, so gibt es eine messbare Abbildung $W(g|T=\cdot) : \mathfrak{X}' \to \overline{\mathbb{R}}$, die für jedes θ eine Version des faktorisierten bedingten Erwartungswertes $W_\theta(g|T=\cdot) = \mathbb{E}_\theta(g(X)|T(X)=\cdot)$ bildet. Kann man $W(\cdot|T=\cdot)$ als bedingte Verteilung gemäß Anmerkung 2.21 wählen, so folgt nach [2, Satz 53.6]

$$W(t|T=t) = \int_{\mathfrak{X}} g(x) W(dx|T=t) \quad W_\theta^T\text{-f.s.}$$

für alle $\theta \in \Theta$.

Beispiel 2.24. Betrachten wir nochmals die in Beispiel 1.1 beschriebene Situation. In Abschnitt 1.2 hatten wir zwei verschiedene Modellbildungen vorgestellt. Bei der

ersten bestand die Beobachtung aus einem Zufallsvektor $X = (X_1, ..., X_n)$, $n = 100$, mit Komponenten

$$X_i := \begin{cases} 1, & \text{falls } i\text{-tes Produkt mangelhaft,} \\ 0, & \text{falls } i\text{-tes Produkt gut,} \end{cases}$$

die unter jedem $\mathbb{P}_\theta$ stochastisch unabhängig und jeweils $Bern(\theta)$-verteilt sind, d.h. $W_\theta = \mathbb{P}_\theta^X = \bigotimes_{i=1}^n Bern(\theta) = Bern(\theta)^n$. W_θ hat also die Zähldichte

$$f_\theta(x) \;=\; W_\theta(\{x\}) = \theta^{s_n}(1-\theta)^{n-s_n}$$

für alle $x = (x_1, ..., x_n) \in \{0,1\}^n$, wobei $s_n = \sum_{j=1}^n x_j$. Das zugrundeliegende statistische Experiment hat die Form

$$\mathscr{E}_1 \;=\; \big(\{0,1\}^n, \mathfrak{P}(\{0,1\}^n),\, (Bern(\theta)^n)_{\theta\in[0,1]}\big).$$

Als zweite Modellbildung hatten wir das statistische Experiment gewählt, welches lediglich die Summe $T(X) = \sum_{i=1}^n X_i$ beobachtet, d.h.

$$\mathscr{E}_2 \;=\; \big(\{0,1,...,n\}, \mathfrak{P}(\{0,1,...,n\}),\, (Bin(n,\theta))_{\theta\in[0,1]}\big)$$

unter Beachtung von $T(X) \stackrel{d}{=} Bin(n,\theta)$ unter $\mathbb{P}_\theta$. In diesem Fall liegt offensichtlich eine Datenreduktion vor, $\mathscr{E}_2$ ist nämlich das durch T reduzierte Experiment $\mathscr{E}_1^T$, und es stellt sich die Frage nach deren Suffizienz, genauer nach der Suffizienz von $T : \{0,1\}^n \to \{0,...,n\}$, $x = (x_1, ..., x_n) \mapsto s_n$ für $\mathscr{E}_1$. Dazu berechnen wir $W_\theta(A|T=t)$ für beliebiges $A \subset \{0,1\}^n$, wobei wir uns o.B.d.A. auf $A = \{x\}$ beschränken. Es gilt

$$\begin{aligned} W_\theta(\{x\})|T=t) \;&=\; \frac{W_\theta(\{x\} \cap \{x' \in \{0,1\}^n : T(x') = t\})}{W_\theta(x' \in \{0,1\}^n : T(x' = t)} \\ &= \begin{cases} \dfrac{W_\theta(\{x\})}{W_\theta^T(\{t\})}, & \text{falls } s_n = t, \\ 0, & \text{falls } s_n \neq t \end{cases} \\ &= \begin{cases} \dfrac{\theta^{s_n}(1-\theta)^{n-s_n}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}, & \text{falls } s_n = t, \\ 0, & \text{falls } s_n \neq t \end{cases} \\ &= \begin{cases} \dfrac{1}{\binom{n}{t}}, & \text{falls } s_n = t, \\ 0, & \text{falls } s_n \neq t \end{cases} \end{aligned}$$

Wir erhalten somit tatsächlich die Suffizienz von T mit

$$W(A|T=t) = \frac{1}{\binom{n}{t}} |A \cap \{x : T(x) = t\}|.$$

2.3 Das Neyman-Kriterium für Suffizienz

Um die Suffizienz einer Statistik nachzuweisen, sind bei Benutzung der Definition bedingte Wahrscheinlichkeiten zu berechnen. Da dies mit viel Mühsal verbunden sein kann, stellt sich die Frage nach einem handlichen Kriterium für Suffizienz. Ein solches bildet das *Faktorisierungs-Kriterium* von Neyman, kurz auch *Neyman-* oder *Neyman-Fisher-Kriterium* genannt, das im Folgenden hergeleitet wird. Als Hilfsresultat benötigen wir

Lemma 2.25. *Sei $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ eine dominierte Familie von W-Maßen auf dem messbaren Raum $(\mathfrak{X},\mathscr{A})$. Dann gilt:*

(a) Es gibt eine abzählbare Teilfamilie $\mathscr{W}' = (W_\theta)_{\theta\in\Theta'}$ von $\mathscr{W}$, die zu $\mathscr{W}$ äquivalent ist, d.h. $W_\theta(N) = 0$ für alle $\theta\in\Theta$ genau dann, wenn $W_\theta(N) = 0$ für alle $\theta\in\Theta'$.

(b) Es gibt ein zu $\mathscr{W}$ äquivalentes W-Maß ν, zum Beispiel $\nu = \sum_{j\geq 1} 2^{-j} W_{\theta_j}$ mit $\{\theta_1,\theta_2,\ldots\} = \Theta'$ gemäß (a).

Beweis. Zu zeigen ist nur Teil (a). Es bezeichne μ das o.B.d.A. als endlich vorausgesetzte dominierende Maß von $\mathscr{W}$, f_θ die μ-Dichte von W_θ und $K_\theta := \{f_\theta > 0\}$. Setze

$$\mathscr{F} := \left\{ \bigcup_{\theta\in I} K_\theta : \Theta \supset I \text{ abzählbar} \right\}$$

und $\rho := \sup\{\mu(C) : C\in\mathscr{F}\}$. Da $\mathscr{F}$ unter der Bildung abzählbarer Vereinigungen abgeschlossen ist, gibt es eine abzählbare Teilmenge Θ' von Θ, so dass $\mu(D) = \rho$ für $D := \bigcup_{\theta\in\Theta'} K_\theta \in \mathscr{F}$. Wir zeigen die Äquivalenz von $\mathscr{W}' := (W_\theta)_{\theta\in\Theta'}$ und $\mathscr{W}$. Dazu notieren wir als erstes, dass $\mu(K_\theta \cap D^c) = 0$ für alle $\theta\in\Theta$. Andernfalls gäbe es nämlich ein θ_0 mit $\mu(K_{\theta_0}\cap D^c) > 0$, was für $D' = D + K_{\theta_0}\cap D^c \in \mathscr{F}$ dann $\mu(D') > \mu(D) = \rho$ lieferte, also einen Widerspruch zur Maximalität von ρ. Nach dieser Feststellung folgt weiter

$$W_\theta(A) \;=\; W_\theta(A\cap K_\theta^c) + W_\theta(A\cap K_\theta\cap D^c) + W_\theta(A\cap K_\theta\cap D) \;=\; W_\theta(A\cap K_\theta)\cap D)$$

für alle $\theta\in\Theta$ und $A\in\mathscr{A}$. Um nun $\mathscr{W}\ll\mathscr{W}'$ nachzuweisen (die umgekehrte Dominiertheit ist trivial), sei $N\in\mathscr{A}$ eine $\mathscr{W}'$-Nullmenge, d.h. $W_\theta(N) = 0$ für alle $\theta\in\Theta'$. Es folgt $\mu(N\cap K_\theta) = 0$ für alle $\theta\in\Theta'$ vermöge

$$0 \;=\; W_\theta(N\cap K_\theta) \;=\; \int_{N\cap\{f_\theta>0\}} f_\theta \, d\mu$$

[☞ Satz 10.2 in [2]] und dann weiter

$$\mu(N\cap K_\theta\cap D) \;=\; \mu\left(N\cap K_\theta\cap \bigcup_{\vartheta\in\Theta'} K_\vartheta\right) \;\leq\; \sum_{\vartheta\in\Theta'} \mu(N\cap K_\theta\cap K_\vartheta) \;=\; 0$$

und wegen $\mathscr{W} \ll \mu$ demnach $W_\theta(N \cap K_\theta \cap D) = W_\theta(N) = 0$ für alle $\theta \in \Theta$. □

Mit Hilfe dieses Lemmas beweisen wir nun zuerst eine Vorstufe des angekündigten Faktorisierungs-Kriteriums.

Satz 2.26. *Sei $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ eine dominierte Familie von W-Maßen auf $(\mathfrak{X}, \mathscr{A})$ und ν ein zu $\mathscr{W}$ äquivalentes W-Maß der Form $\nu = \sum_{j\geq 1} 2^{-j} W_{\theta_j}$. Genau dann ist eine Statistik $T : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}', \mathscr{A}')$ suffizient (für $\mathscr{W}$), wenn für jedes $\theta \in \Theta$ eine $\mathscr{A}'$-messbare Funktion g_θ existiert, so dass $g_\theta \circ T$ eine ν-Dichte von W_θ bildet.*

Beweis. "⇒" Sei T als suffizient vorausgesetzt und $W(A|T) = W(A|T = \cdot) \circ T$ für $A \in \mathscr{A}$. Aufgrund der speziellen Gestalt von ν sieht man sofort ein, dass $W(A|T)$ auch eine Version von $\nu(A|T) = \sum_{j\geq 1} 2^{-j} W_{\theta_j}(A|T)$ bildet, also $\nu(g|T)\nu$-f.s. für jede nicht-negative Funktion g gilt. Weiter folgt aus der Äquvalenz von $(W_\theta)_{\theta\in\Theta}$ und ν die Äquivalenz ihrer Einschränkungen auf $\sigma(T)$, bezeichnet mit $(W_{\theta|\sigma(T)})_{\theta\in\Theta}$ bzw. $\nu_{|\sigma(T)}$. Nach dem Satz von Radon-Nikodym [13.9 in [2]] zusammen mit dem Faktorisierungslemma [52.1 in [2]] existiert deshalb eine $\sigma(T)$-messbare $\nu_{|\sigma(T)}$-Dichte $f_\theta = g_\theta \circ T$ von $W_{\theta|\sigma(T)}$. Wir folgern aus der Suffizienz von T

$$\begin{aligned}
W_\theta(A) &= \int_{\mathfrak{X}} W(A|T)\, dW_\theta \\
&= \int_{\mathfrak{X}} W(A|T)\, dW_{\theta|\sigma(T)} \\
&= \int_{\mathfrak{X}} W(A|T)(g_\theta \circ T)\, d\nu_{|\sigma(T)} \\
&= \int_{\mathfrak{X}} \nu(\mathbf{1}_A \cdot (g_\theta \circ T)|T)\, d\nu \\
&= \int_{\mathfrak{X}} \mathbf{1}_A \cdot (g_\theta \circ T)\, d\nu \\
&= \int_A g_\theta \circ T\, d\nu
\end{aligned}$$

für alle $A \in \mathscr{A}$, wobei die besagte ν-f.s. gültige Beziehung $\nu(g|T) = W(g|T)$ für die vierte Zeile benutzt wurde. Damit ist $g_\theta \circ T$ aber sogar eine Dichte von W_θ bzgl. ν und die Behauptung bewiesen.

"⇐" Existiere für jedes θ eine $\mathscr{A}'$-messbare Funktion g_θ mit $g_\theta \circ T = dW_\theta/d\nu$, und sei $h_A(t) = \nu(A|T = t)$ für $A \in \mathscr{A}$. Dann folgt

$$\begin{aligned}
\int_{A'} h_A(t) W_\theta^T(dt) &= \int_{\{T\in A'\}} (h_A \circ T)(g_\theta \circ T)\, d\nu \\
&= \int_{\{T\in A'\}} \nu(\mathbf{1}_A|T)(g_\theta \circ T)\, d\nu \\
&= \int_{\{T\in A'\}} \nu(\mathbf{1}_A \cdot (g_\theta \circ T)|T)\, d\nu
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\{T\in A'\}} \mathbf{1}_A \cdot (g_\theta \circ T)\, d\nu \\
&= \int_{\{T\in A'\}} \mathbf{1}_A \, dW_\theta \\
&= W_\theta(A \cap \{T \in A'\})
\end{aligned}
$$

für alle $\theta \in \Theta$, $A \in \mathscr{A}$ und $A' \in \mathscr{A}'$. $h_A(t)$ bildet somit gemäss Satz 46.2 in [2] eine von θ unabhängige Version der faktorisierten bedingten Wahrscheinlichkeit $W_\theta(A|T=t)$, d.h., T ist suffizient. □

Wir sind nun in der Lage, das angekündigte Neyman-Kriterium zu beweisen.

Satz 2.27. (Neyman-Kriterium für Suffizienz) *Sei $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ eine dominierte Familie von W-Maßen auf $(\mathfrak{X},\mathscr{A})$ mit dominierendem Maß μ. Genau dann ist eine Statistik $T : (\mathfrak{X},\mathscr{A}) \to (\mathfrak{X}',\mathscr{A}')$ suffizient (für $\mathscr{W}$), wenn für jedes $\theta \in \Theta$ eine $\mathscr{A}'$-messbare Funktion g_θ und eine $\mathscr{A}$-messbare Funktion h existieren, so dass durch $(g_\theta \circ T)\cdot h$ eine μ-Dichte von W_θ gegeben ist.*

Beweis. Sei ν das nach Lemma 2.25 existierende zu $\mathscr{W}$ äquivalente W-Maß der Form $\nu = \sum_{j\geq 1} 2^{-j} W_{\theta_j}$ für geeignete $\{\theta_1,\theta_2,\ldots\} \subset \Theta$.

"⇒" Nach dem vorherigen Satz gibt es dann $\mathscr{A}'$-messbare Funktionen g_θ, so dass $g_\theta \circ T = dW_\theta/d\nu$. Aufgrund der Äquivalenz von $\mathscr{W}$ und ν folgt natürlich auch $\nu \ll \mu$ und somit die Existenz einer μ-Dichte h von ν. Insgesamt erhalten wir unter Verwendung von Korollar 13.5 in [2]

$$
\frac{dW_\theta}{d\mu} \;=\; \frac{dW_\theta}{d\nu}\cdot\frac{d\nu}{d\mu} \;=\; (g_\theta \circ T)\cdot h.
$$

"⇐" Wir können gleich $g_\theta,\ h \geq 0$ annehmen, da andernfalls der Übergang zu $|g_\theta|$, $|h|$ das Gewünschte liefert. Sei $\mu' = h\mu$, woraus wegen $W_\theta = ((g_\theta \circ T)\cdot h)\mu$ sofort $dW_\theta/d\mu' = g_\theta \circ T$ folgt. Wir erhalten

$$
\frac{d\nu}{d\mu'} \;=\; \sum_{j\geq 1} 2^{-j}\,\frac{dW_{\theta_j}}{d\mu'} \;=\; \sum_{j\geq 1} 2^{-j} g_{\theta_j} \circ T.
$$

Nun ergibt sich aber unter nochmaliger Benutzung von Korollar 13.5 in [2]

$$
\frac{dW_\theta}{d\nu} \;=\; \frac{dW_\theta}{d\mu'}\Big/\frac{d\nu}{d\mu'} \;=\; \frac{g_\theta \circ T}{\sum_{j\geq 1} 2^{-j} g_{\theta_j} \circ T} \;=\; \left(\frac{g_\theta}{\sum_{j\geq 1} 2^{-j} g_{\theta_j}}\right) \circ T
$$

für alle $\theta \in \Theta$ und daraus die Suffizienz von T vermöge Satz 2.26. □

Die anschließenden Beispiele sollen demonstrieren, wie einfach es mit Hilfe des Neyman-Kriteriums oft ist, suffiziente Statistiken für ein statistisches Experiment zu finden.

Beispiel 2.28. Kehren wir noch einmal zu der in Beispiel 1.1 beschriebenen Situation zurück und betrachten das Experiment $(\{0,1\}^n, \mathfrak{P}(\{0,1\}^n), (Bern(\theta)^n)_{\theta\in[0,1]})$. Die Zähldichte von W_θ auf $\{0,1\}^n$ lautet

$$f_\theta(x_1,...,x_n) \;=\; \theta^{s_n}(1-\theta)^{n-s_n}, \quad s_n \;=\; x_1+...+x_n,$$

woraus jetzt mit dem Neyman-Kriterium die Suffizienz der Statistik $T(x_1,...,x_n) = s_n$ folgt (setze $g_\theta(t) = \theta^t(1-\theta)^{n-t}$ und $h = 1$). In Beispiel 2.24 hatten wir dies durch direktes Nachrechnen gezeigt.

Beispiel 2.29. (a) Sind $X_1,...,X_n$ stochastisch unabhängige, jeweils $Normal(\mu,\sigma^2)$-verteilte ZG mit unbekanntem Parameter $\theta = (\mu,\sigma^2) \in \mathbb{R}\times(0,\infty)$, so hat $X = (X_1,...,X_n)$ unter $\mathbb{P}_\theta$ die $\lambda\!\!\!\lambda^n$-Dichte

$$\begin{aligned} f_\theta(x) \;&=\; \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right) \\ &=\; \frac{1}{(\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma^2}\right) \\ &=\; g_\theta\circ(T_1(x),T_2(x))\cdot\mathbf{1}, \end{aligned}$$

wobei

$$T(x) \;:=\; (T_1(x),T_2(x)) \;=\; \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)$$

und

$$g_\theta(t_1,t_2) \;:=\; \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}t_2 + \frac{\mu}{\sigma^2}t_1 - \frac{n\mu^2}{2\sigma^2}\right).$$

Die Statistik T ist somit suffizient für $(Normal(\mu,\sigma^2)^n)_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)}$.

(b) Nehmen wir an, dass ledigich $\theta = \sigma^2$ unbekannt ist, μ dagegen eine bekannte Konstante, so bleibt zwar die zuvor definierte Statistik T suffizient, doch dasselbe gilt auch für $U(x) = \sum_{i=1}^n (x_i-\mu)^2$, da $f_{\sigma^2}(x) = \widehat{g}_{\sigma^2}\circ U(x)$ mit

$$\widehat{g}_{\sigma^2}(u) \;:=\; \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{u}{2\sigma^2}\right).$$

(c) Ist schließlich μ unbekannt und σ^2 eine gegebene Konstante, so erhält man als suffiziente Statistik neben T auch T_1, da nun $f_\mu(x) = (\overline{g}_\mu\circ T_1(x))\cdot h(x)$ mit

$$\overline{g}_\mu(t) \;:=\; \exp\left(\frac{\mu t}{2\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right) \quad\text{und}\quad h(x) := \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{T_2(x)}{2\sigma^2}\right).$$

Beispiel 2.30. Sind $X_1,...,X_n$ stochastisch unabhängige $Unif(a,b)$-verteilte ZG mit unbekanntem Parameter $\theta = (a,b)$, $-\infty < a < b < \infty$, so hat $X = (X_1,...,X_n)$ die $\lambda\!\!\!\lambda^2$-Dichte

$$\begin{aligned} f_\theta(x) &= \frac{1}{(b-a)^n}\prod_{i=1}^n \mathbf{1}_{(a,b)}(x_i) \\ &= \frac{1}{(b-a)^n}\mathbf{1}_{(a,\infty)}\left(\min_{1\le i\le n} x_i\right)\mathbf{1}_{(-\infty,b)}\left(\max_{1\le i\le n} x_i\right) \\ &= \frac{1}{(b-a)^n}\mathbf{1}_{(a,\infty)\times(-\infty,b)}\left(\min_{1\le i\le n} x_i, \max_{1\le i\le n} x_i\right). \end{aligned}$$

Nach dem Neyman-Kriterium bildet somit $T(x) := (\min_{1\le i\le n} x_i,\ \max_{1\le i\le n} x_i)$ eine suffiziente Statistik für $(Unif(a,b)^n)_{-\infty<a<b<\infty}$.

Ebenso wie die vorherigen Beispiele ergibt sich das folgende Korollar als direkte Konsequenz des Neyman-Kriteriums.

Korollar 2.31. *Sei $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ eine k-parametrige Exponentialfamilie mit Dichten*

$$\frac{dW_\theta}{d\nu} = C(\theta)\exp\left(\sum_{i=1}^k Q_i(\theta)T_i\right)h \quad \nu\text{-f.ü.} \tag{2.9}$$

bezüglich eines dominierenden Maßes ν. Dann ist $(T_1,...,T_k)$ suffizient für $\mathscr{W}$.

Gegeben unabhängige, identisch verteilte $X_1,...,X_n$ mit Werten in einem messbaren Raum $(\mathfrak{X},\mathscr{A})$ und Verteilungen $\mathbb{P}_\theta^{X_1} = W_\theta$ für alle $\theta\in\Theta$, die eine k-parametrige Exponentialfamilie wie in 2.31 bilden, impliziert das Korollar insbesondere die Suffizienz der Statistik

$$T_\Sigma(x) = \left(\sum_{i=1}^n T_1(x_i),...,\sum_{i=1}^n T_k(x_i)\right), \quad x = (x_1,...,x_n),$$

für $\mathscr{W}^n = (W_\theta^n)_{\theta\in\Theta}$ bzw. das zugehörige *Produkt-Experiment* $\mathscr{E}^n = (\mathfrak{X}^n,\mathscr{A}^n,\mathscr{W}^n)$, wie sich sofort unter Hinweis auf Satz 1.26(b) ergibt. Die Dimension dieser Statistik hängt offenkundig nicht von der Anzahl der Beobachtungen ab.

Das nächste Korollar zeigt, dass es zu einem statistischen Experiment i.A. "viele" suffiziente Statistiken gibt.

Korollar 2.32. *Gegeben eine suffiziente Statistik $T:\mathfrak{X}\to\mathfrak{X}'$ für das dominierte statistische Experiment $\mathscr{E} = (\mathfrak{X},\mathscr{A},(W_\theta)_{\theta\in\Theta})$, ist jede weitere Statistik $S:\mathfrak{X}\to\mathfrak{X}''$ mit $T = k\circ S$ für eine messbare Abbildung $k:\mathfrak{X}''\to\mathfrak{X}'$ ebenfalls suffizient.*

Beweis. Bezeichnet μ ein dominierendes Maß für $(W_\theta)_{\theta\in\Theta}$ und f_θ die μ-Dichte von W_θ, so existieren gemäß dem Neyman-Kriterium Funktionen g_θ und h, so dass

$$f_\theta = (g_\theta\circ T)\cdot h = (g_\theta\circ k\circ S)\cdot h = (\widetilde{g}_\theta\circ S)\cdot h$$

für alle $\theta \in \Theta$ gilt, wobei $\widetilde{g}_\theta = g_\theta \circ k$. Damit folgt aber auch die Suffizienz von S nach demselben Kriterium. □

Beispiel 2.33. Wir hatten in 2.29 gesehen, dass $T(x) = (\sum_{j=1}^n x_j, \sum_{j=1}^n x_j^2)$ eine suffiziente Statistik für $(Normal(\mu, \sigma^2)^n)_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)}$ definiert. Mit Hilfe von Korollar 2.32 ergibt sich nun sofort, dass dies auch für

$$S(x) \;=\; (S_1(x), S_2(x)) \;:=\; \left(\overline{x}_n, \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2\right),$$

bestehend aus Stichprobenmittel und Stichprobenvarianz, der Fall ist, denn $T(x) = (nS_1(x), nS_2(x) + nS_1(x)^2)$.

2.4 Minimalsuffizienz

In diesem Abschnitt wollen wir kurz auf die Frage eingehen, in welchem Umfang eine Datenreduktion in einem gegebenen statistischen Modell durchgeführt werden kann, also auf die Frage nach einer *maximalen Reduktion*. Im Hinblick auf suffiziente Statistiken wollen wir demnach wissen, ob es eine solche gibt, die unter allen am "gröbsten" oder "einfachsten" ist. Dies führt zum Begriff der *Minimalsuffizienz*.

Definition 2.34. Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, \mathscr{W})$ ein statistisches Experiment. Eine suffiziente Statistik $T^* : \mathfrak{X} \to \mathfrak{X}'$ heißt *minimalsuffizient (für $\mathscr{E}$ oder $\mathscr{W}$)*, wenn sie über jeder weiteren suffizienten Statistik T faktorisiert, d.h., wenn

$$T^* \;=\; h \circ T \quad \mathscr{W}\text{-f.s.}$$

für eine geeignete messbare Funktion h gilt.

Minimalsuffiziente Statistiken existieren unter sehr schwachen Voraussetzungen [☞ z.B. BAHADUR [3]], insbesondere wenn das Experiment $\mathscr{E}$ dominiert und der Stichprobenraum euklidisch ist und versehen mit der Borelschen σ-Algebra, d.h., wenn $(\mathfrak{X}, \mathscr{A}) = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ für ein $n \geq 1$. Es gibt aber auch Ausnahmen, wie die Arbeiten von LANDERS & ROGGE [10] und PITCHER [16] zeigen.

Bei der Frage nach der Konstruktion von minimalsuffizienten Statistiken wollen wir uns im Folgenden auf den Fall beschränken, dass die zugrundeliegende Verteilungsfamilie $\mathscr{W}$ aus äquivalenten Maßen besteht. Diese Voraussetzung gilt insbesondere für Exponentialfamilien.

Satz 2.35. *Sei $\mathscr{W} = (W_j)_{0\le j\le n}$ eine endliche Familie äquivalenter Verteilungen auf einem messbaren Raum $(\mathfrak{X}, \mathscr{A})$ mit Dichten $f_0, ..., f_n$ bzgl. eines dominierenden Maßes μ. Dann definiert*

$$T \;=\; \left(\frac{f_1}{f_0}, ..., \frac{f_n}{f_0}\right)$$

eine minimalsuffiziente Statistik für $\mathscr{W}$.

Beweis. Da alle W_j äquivalent sind, stimmen die Mengen $\{f_j > 0\}$ μ-f.ü. überein, und T ist ferner wohldefiniert, falls wir $\frac{0}{0} := 0$ vereinbaren. Nun gilt aber für jedes $j \in \{0, ..., n\}$ auch

$$\frac{dW_j}{dW_0} \;=\; \frac{dW_j}{d\mu} \Big/ \frac{dW_0}{d\mu} \;=\; \frac{f_j}{f_0} \;=\; p_j \circ T \quad W_0\text{-f.s.},$$

wobei $p_j(x_1, ..., x_n) = x_j$ die j-te Projektion bezeichnet. Die Suffizienz folgt deshalb gemäß dem Neyman-Kriterium 2.27 mit W_0 als dominierendem Maß. Nach diesem existieren aber außerdem für jede weitere suffiziente Statistik S Funktionen $h, g_0, ..., g_n$, so dass $f_j = (g_j \circ S) \cdot h$, also $f_j/f_0 = (g_j/g_0) \circ S$ μ-f.ü. Dies impliziert schließlich

$$T \;=\; \left(\frac{g_1}{g_0}, ..., \frac{g_n}{g_0}\right) \circ S \quad W_0\text{-f.s.}$$

und somit die Minimalsuffizienz von T. □

Für beliebige Familien äquivalenter Verteilungen kann man die Minimalsuffizienz einer Statistik nun oft durch Kombination des vorherigen Satzes mit dem Folgenden Lemma zeigen.

Lemma 2.36. *Sei $\mathscr{W}$ eine Familie äquivalenter Verteilungen und $\mathscr{W}_0$ eine endliche Teilfamilie. Dann ist jede Statistik T, die minimalsuffizient für $\mathscr{W}_0$ und suffizient für $\mathscr{W}$ ist, bereits minimalsuffizient für $\mathscr{W}$.*

Beweis. Sei T eine Statistik wie im Lemma gefordert und S eine weitere für $\mathscr{W}$ suffiziente Statistik. Dann ist S natürlich auch suffizient für $\mathscr{W}_0$, und wegen der Minimalsuffizienz von T für $\mathscr{W}_0$ existiert eine messbare Funktion h, derart, dass $T = h \circ S$ $\mathscr{W}_0$-f.s. und somit $\mathscr{W}$-f.s., denn $\mathscr{W}$ und $\mathscr{W}_0$ sind nach Voraussetzung äquivalent. □

Für Exponentialfamilien können wir nun zeigen:

Satz 2.37. *Sei $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ eine k-parametrige Exponentialfamilie von Verteilungen auf $(\mathfrak{X},\mathscr{A})$ mit ν-Dichten gemäß (2.9). Dann ist $T = (T_1,...,T_k)$ minimalsuffizient für $\mathscr{W}$, wenn $\mathscr{Q}(\Theta) := \{(Q_1(\theta),...,Q_k(\theta)) : \theta \in \Theta\} \subset \mathbb{R}^k$ innere Punkte besitzt.*

Beachte, dass die Bedingung an $\mathscr{Q}(\Theta)$ vermöge Satz 1.29 insbesondere gilt, wenn $\mathscr{W}$ vollen Rang besitzt [☞ Definition 1.28] und $\mathscr{Q}(\Theta)$ dem natürlichen Parameterraum $\mathfrak{Z}$ von $\mathscr{W}$, definiert in (1.7), entspricht.

Beweis. Die Suffizienz von T für $\mathscr{W}$ gilt gemäß Korollar 2.31. Sei $\mathscr{W}_0 = (T_\theta)_{0\le j\le k}$ eine Teilfamilie von $\mathscr{W}$ mit $k+1$ Elementen. Nach Satz 2.35 und einer einfachen bijektiven, von $\theta_0,...,\theta_k$ abhängigen Transformation folgt, dass

$$\widehat{T}(x) \;=\; \left(\sum_{j=1}^{k}(Q_j(\theta_1)-Q_j(\theta_0))T_j(x),...,\sum_{j=1}^{k}(Q_j(\theta_k)-Q_j(\theta_0))T_j(x)\right)$$

minimalsuffizient für $\mathscr{W}_0$ ist, und damit wegen

$$\widehat{T} \;=\; \Delta Q\cdot T \;=\; \begin{pmatrix} Q_1(\theta_1)-Q_1(\theta_0) & \dots & Q_k(\theta_1)-Q_k(\theta_0)\\ \vdots & \ddots & \vdots\\ Q_1(\theta_k)-Q_1(\theta_0) & \dots & Q_k(\theta_k)-Q_k(\theta_0)\end{pmatrix}\cdot\begin{pmatrix}T_1\\ \vdots\\ T_k\end{pmatrix}$$

auch die Minimalsuffizienz von T für $\mathscr{W}_0$, sofern ΔQ eine reguäre Matrix bildet. Dies kann aber durch geeignete Wahl von $\theta_0,...,\theta_k$ immer erreicht werden, weil $\mathscr{Q}(\Theta)$ nach Voraussetzung innere Punkte besitzt. Die Minimalsuffizienz von T für $\mathscr{W}$ ergibt sich nun aus Lemma 2.36. □

Jedes der drei folgenden Beispiele bildet eine Exponentialfamilie, für die die Bedingung an $\mathscr{Q}(\Theta)$ trivialerweise erfüllt ist.

Beispiel 2.38 (vgl. Beispiel 2.28). $\mathscr{W} = (Bin(m,p)^n)_{p\in(0,1)}, m \ge 1$ fest: Dann sind $T(x) = s_n = x_1 + ... + x_n$ und $\widehat{T}(x) = \bar{x}_n$ minimalsuffiziente Statistiken.

Beispiel 2.39. $\mathscr{W} = (Poisson(\theta)^n)_{\theta>0}$: Dann sind wiederum $T(x) = s_n$ und $\widehat{T}(x) = \bar{x}_n$ minimalsuffiziente Statistiken.

Beispiel 2.40 (vgl. Beispiele 2.29 und 2.33). $\mathscr{W} = (Normal(\mu,\sigma^2)^n)_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)}$: Dann sind $T(x) = (s_n, \sum_{j=1}^n x_j^2)$ sowie $\widehat{T}(x) = (\bar{x}_n, \frac{1}{n}\sum_{j=1}^n (x_j-\bar{x}_n)^2)$ minimalsuffiziente Statistiken.

2.5 Vollständigkeit, Verteilungsfreiheit und der Satz von Basu

Im weiteren Verlauf unserer Betrachtungen werden wir neben der Suffizienz einen weiteren Begriff benötigen, und zwar den der *Vollständigkeit* einer Statistik. Wir

beginnen mit einer kurzen Motivation einschließlich der Definition von *Verteilungsfreiheit*, die wir dem ausgezeichneten Werk von LEHMANN [11, S. 45f] entnommen haben. Diese ist umso bemerkenswerter, weil in den meisten Lehrbüchern die Definition der Vollständigkeit eher auf den Leser "niederkommt" nach dem Motto: Hier ist die Bedingung, die wir jetzt benötigen, aber fragen Sie mich nicht, warum!

Erinnern wir uns: Die wesentliche Konsequenz bei Vorliegen einer suffizienten Statistik T besteht darin, dass die zugrunde liegende unbekannte Verteilung W_θ faktoriell in einen von θ unabhängigen Anteil $W(\cdot|T=\cdot)$ und einen von θ abhängigen Anteil W_θ^T zerlegt werden kann, ausgedrückt durch die Beziehung

$$W_\theta(dx) \;=\; W(dx|T=t)\, W_\theta^T(dt).$$

Der von θ unabhängige Anteil $W(\cdot|T=t)$ enthält also lediglich zusätzliche Informationen über $\mathbb{W}_\theta$, die den Parameter nicht mehr betreffen, und erlaubt daher den Übergang ohne Informationsverlust von dem statistischen Experiment $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ zu dem durch T reduzierten Experiment $\mathscr{E}^T$, das in 2.18 definiert wurde. Die interessante Frage, in welchem Umfang eine Datenreduktion in einem gegebenen Experiment $\mathscr{E}$ möglich ist, scheint somit davon abzuhängen, wieviel zusätzliche, den Parameter nicht betreffende Information in den Verteilungen W_θ enthalten ist. Dies führt zu folgender

Definition 2.41. Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ ein statistisches Experiment und wie üblich $W_\theta = \mathbb{P}_\theta^X$. Eine Statistik $T : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}', \mathscr{A}')$ heißt

(a) *verteilungsfrei* (engl. *ancillary*) *(für $\mathscr{E}$ oder $(W_\theta)_{\theta\in\Theta}$)*, falls $W_\theta^T = \mathbb{P}_\theta^{T(X)}$ unabhängig von θ ist.

(b) *verteilungsfrei 1. Ordnung (für $\mathscr{E}$ oder $(W_\theta)_{\theta\in\Theta}$)*, falls $\mathbb{E}_\theta T(X)$ nicht von θ abhängt.

Verteilungsfreie Statistiken bilden also den Gegenpol zu suffizienten Statistiken, indem sie überhaupt keine Auskunft über den unbekannten Parameter liefern. Andererseits können selbst minimalsuffiziente Statistiken immer noch sehr viel verteilungsfreies Material enthalten, denn es gibt Situationen, in denen zu einer minimalsuffizienten Statistik T die Transformation $f(T)$ mit einer geeigneten nichtkonstanten Abbildung f verteilungsfrei ist. Dennoch erscheint der Umkehrschluss plausibel: Eine Datenreduktion mittels einer suffizienten Statistik T kann nicht weiter verbessert werden, wenn es *keine* nichtkonstante Funktion f gibt derart, dass $f(T)$ verteilungsfrei ist. Unter leichter Verschärfung der Voraussetzung, indem wir "verteilungsfrei" durch "verteilungsfrei 1. Ordnung" ersetzen, werden wir diese Aussage tatsächlich in Satz 2.43 bestätigen. Die verschärfte Voraussetzung lässt sich offenbar auch folgendermaßen formulieren:

$$\mathbb{E}_\theta f(T(X)) = c \;\text{ f.a. } \theta\in\Theta \quad\Rightarrow\quad f = c \;\; \mathbb{P}_\theta^{T(X)}\text{-f.s. f.a. } \theta\in\Theta \tag{2.10}$$

für beliebige $c \in \mathbb{R}$. Subtrahiert man die Konstante c, so erweist sich (2.10) als äquivalent zu

$$\mathbb{E}_\theta f(T(X)) = 0 \ \text{ f.a } \theta \in \Theta \quad \Rightarrow \quad f = 0 \ \mathbb{P}_\theta^{T(X)}\text{-f.s. f.a. } \theta \in \Theta \tag{2.11}$$

und rechtfertigt die anschließende Definition.

Definition 2.42. Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, \mathscr{W})$ ein statistisches Experiment, wobei $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ und $W_\theta = \mathbb{P}_\theta^X$. Eine Statistik $T : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}', \mathscr{A}')$ heißt *vollständig (für $\mathscr{E}$ oder $\mathscr{W}$)*, falls sie der Bedingung (2.11) genügt.

Hier nun das angekündigte Ergebnis, das im Wesentlichen auf BAHADUR [3] zurückgeht.

Satz 2.43. (von Bahadur) *Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ ein statistisches Experiment. Dann folgt aus der Suffizienz und Vollständigkeit einer Statistik $T : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}', \mathscr{A}')$ stets deren Minimalsuffizienz, sofern $(\mathfrak{X}', \mathscr{A}')$ ein Borel-Raum ist.*

Beweis. Da $(\mathfrak{X}', \mathscr{A}')$ ein Borel-Raum, also das bimessbare Bild einer Borelschen Teilmenge von $[0,1]$ ist, können wir T o.B.d.A. als reellwertig annehmen[1]. Sei $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ und $S : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}'', \mathscr{A}'')$ eine beliebige suffiziente Statistik. Wir müssen $T = f \circ S$ $\mathscr{W}$-f.s. für eine geeignete messbare Funktion f zeigen, was vermöge (einer leichten Verallgemeinerung) des Faktorisierungslemmas 52.1 in [2] damit gleichbedeutend ist, dass für alle $A \in \mathscr{A}'$ ein geeignetes $B \in \mathscr{A}''$ existiert mit $\mathbf{1}_A(T) = \mathbf{1}_B(S)$ $\mathscr{W}$-f.s.[2] Dies wiederum gilt, wenn $W(T \in A|S) - \mathbf{1}_A(T) = 0$ $\mathscr{W}$-f.s. für alle $A \in \mathscr{A}'$. Bedingen dieser Differenz unter T und Integration bzgl. W_θ liefert

$$\int_{\mathfrak{X}} (W(W(T \in A|S)|T) - \mathbf{1}_A(T))\, dW_\theta \ = \ W_\theta(T \in A) - W_\theta(T \in A) \ = \ 0$$

für alle $\theta \in \Theta$ und daher $W(W(T \in A|S)|T) = \mathbf{1}_A(T)$ $\mathscr{W}$-f.s. wegen der Vollständigkeit von T. Damit erhalten wir aber weiter

$$\begin{aligned} 0 \ &\leq \ W((W(T \in A|S) - \mathbf{1}_A(T))^2|T) \\ &= \ W(W(T \in A|S)^2) - 2\,\mathbf{1}_A(T)W(W(T \in A|S)|T) + \mathbf{1}_A(T)^2 \end{aligned}$$

[1] Ist nämlich $\phi : (\mathfrak{X}', \mathscr{A}') \to (C, \mathscr{B}(C))$ eine bimessbare Bijektion für ein $C \in \mathscr{B}([0,1])$, so gilt offenkundig $T^{-1}(\mathscr{A}') = (\phi \circ T)^{-1}(\mathscr{B}(C))$ und folglich, dass mit T auch die *reellwertige* Abbildung $\phi \circ T$ vollständig und suffizient ist.

[2] Dann folgt nämlich für geeignete primitive Funktionen g_n, f_n, dass

$$T \ = \ \lim_{n\to\infty} g_n \circ T \ = \ \lim_{n\to\infty} f_n \circ S \quad \mathscr{W}\text{-f.s.}$$

und daraus die $\mathscr{W}^S$-f.s. Konvergenz der f_n gegen ein f, was $T = f \circ S$ $\mathscr{W}$-f.s. impliziert.

$$\begin{aligned} &= W(W(T \in A|S)^2|T) - \mathbf{1}_A(T)^2 \\ &\leq W(W(T \in A|S)|T) - \mathbf{1}_A(T) \;=\; 0 \quad \mathscr{W}\text{-f.s.}, \end{aligned}$$

was offenkundig $W(T \in A|S) = \mathbf{1}_A(T)$ $\mathscr{W}$-f.s. beweist. □

Wir betonen nochmals, dass eine minimalsuffiziente Statistik i.A. nicht vollständig sein muss. Die Umkehrung von Satz 2.43 ist folglich falsch. Richtig ist dagegen, dass entweder keine oder alle minimalsuffizienten Statistiken für ein Experiment vollständig sind. Dies ergibt sich aus der Tatsache, dass jede minimalsuffiziente Statistik über jeder anderen solchen faktorisiert sowie dem anschließenden einfachen

Lemma 2.44. *Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ ein statistisches Experiment und T eine vollständige Statistik für $\mathscr{E}$. Dann ist auch $f \circ T$ für jedes messbare f vollständig für $\mathscr{E}$.*

Beweis. Sei $S = f \circ T$. Aus

$$\mathbb{E}_\theta g(S(X)) \;=\; \mathbb{E}_\theta (g \circ f)(T(X)) \;=\; 0$$

für alle $\theta \in \Theta$ folgt wegen der Vollständigkeit von T

$$W_\theta^{f\circ T}(\{g = 0\}) \;=\; W_\theta^T(\{g \circ f = 0\}) \;=\; 1,$$

d.h. $g = 0$ W_θ^S-f.s. für alle $\theta \in \Theta$. Dies zeigt die Vollständigkeit von S. □

Unsere eingangs angestellten Überlegungen hatten zum Begriff der Vollständigkeit als Kriterium für das Fehlen weiterer verteilungsfreier Information geführt. Dies rechtfertigt die Vermutung, dass eine vollständige suffiziente Statistik von jeder verteilungsfreien Statistik unabhängig ist und wird im folgenden Satz von BASU [4] bestätigt.

Satz 2.45. (von Basu) *Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ ein statistisches Experiment und T eine vollständige und suffiziente Statistik für $\mathscr{E}$. Dann ist jede verteilungsfreie Statistik unabhängig von T.*

Beweis. Sei $S : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}', \mathscr{A}')$ eine verteilungsfreie Statistik und Q die Verteilung von S unter jedem W_θ. Für $A' \in \mathscr{A}'$ definieren wir $f_{A'}(t) := W(S \in A'|T = t)$. Dann gilt

$$\begin{aligned} \mathbb{E}_\theta(f_{A'}(T(X)) - Q(A')) \;&=\; \int (W(S \in A'|T) - Q(A'))\, dW_\theta \\ &=\; \int (\mathbf{1}_{\{S\in A'\}} - Q(A'))\, dW_\theta = 0 \end{aligned}$$

für alle $\theta \in \Theta$ und $A' \in \mathscr{A}'$. Da T vollständig ist, folgt $f_{A'} = W(S \in A'|T = \cdot) = Q(A')\ W_\theta^T$-f.s. für jedes $\theta \in \Theta$ und somit die Behauptung. □

Bevor wir eine interessante Anwendung des Satzes von Basu geben, notieren wir das folgende Resultat für Exponentialfamilien.

Satz 2.46. *$\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ sei eine k-parametrige Exponentialfamiie von Verteilungen auf $(\mathfrak{X}, \mathscr{A})$ mit ν-Dichten gemäß (2.9). Besitzt $\mathscr{Q}(\Theta) := \{(Q_1(\theta), ..., Q_k(\theta)) : \theta \in \Theta\} \subset \mathbb{R}^k$ innere Punkte, so ist $T = (T_1, ..., T_k)$ vollständig für $\mathscr{W}$.*

Beweis. Wir verweisen auf WITTING [21, Satz 3.39 auf S. 356f]. □

Es sei an dieser Stelle nochmals erwähnt [vgl. Satz 2.37], dass im obigen Satz innere Punkte auf jeden Fall existieren, wenn $\mathscr{W}$ vollen Rang besitzt und $\mathscr{Q}(\Theta)$ dem natürlichen Parameterraum $\mathfrak{Z}$ von $\mathscr{W}$ entspricht.

Beispiel 2.47. (Unabhängigkeit von Stichprobenmittel und Stichprobenvarianz bei Normalverteilungen) Gegeben sei ein Zufallsvektor $X = (X_1, ..., X_n)$, bestehend aus stochastisch unabhängigen, unter $\mathbb{P}_{(\mu,\sigma^2)}$ jeweils $Normal(\mu, \sigma^2)$-verteilten ZG mit unbekanntem Parameter $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, folglich

$$W_{(\mu,\sigma^2)} \;=\; \mathbb{P}^X_{(\mu,\sigma^2)} \;=\; Normal(\mu, \sigma^2)^n.$$

Wir hatten in 2.40 festgehalten, dass

$$\widehat{T}(x) \;=\; \left(\bar{x}_n, \frac{1}{n}\sum_{j=1}^n (x_j - \bar{x}_n)^2\right) \;=:\; \big(S(x), V^2(x)\big)$$

für die Verteilungsfamilie $(Normal(\mu, \sigma^2)^n)_{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)}$ minimlasuffizient ist, und werden im Anschluss zeigen, dass das Stichprobenmittel $S(x)$ und die Stichprobenvarianz $V^2(x)$ außerdem unter jedem $W_{(\mu,\sigma^2)}$ stochastisch unabhängig sind. Dazu halten wir ein beliebiges $\sigma^2 > 0$ fest und erinnern daran, dass S in diesem Fall suffizient ist für die reduzierte Verteilungsfamilie $(Normal(\mu, \sigma^2)^n)_{\mu\in\mathbb{R}}$ (☞ 2.29(c)). Eine Anwendung von Satz 2.46 liefert ferner die Vollständigkeit von S. Es reicht deshalb unter Hinweis auf den Satz von Basu zu zeigen, dass V^2 für die reduzierte Familie verteilungsfrei ist, wozu wir uns folgendes überlegen: Sei

$$Y_j \;=\; Y_j(\mu) \;:=\; X_j - \mu$$

für $j = 1, ..., n$. Dann sind $Y_1, ..., Y_n$ unter $\mathbb{P}_{(\mu,\sigma^2)}$ stochastisch unabhängig und jeweils $Normal(0, \sigma^2)$-verteilt[3]. Darüber hinaus liefert eine einfache Rechnung

[3] Funktionen $g(\theta, X)$ der Beobachtungsvariablen X und des unbekannten Parameters θ, deren Verteilung für jedes θ dieselbe ist, werden in der englischsprachigen Literatur *pivotal element* oder einfach *pivot* genannt. Eine verteilungsfreie Statistik ist natürlich immer auch ein pivot, aber die Umkehrung gilt nicht.

$$V^2(X) \;=\; \frac{1}{n}\sum_{j=1}^{n}(X_j-\mu+\mu-\overline{X}_n)^2 \;=\; V^2(Y),$$

wobei Y natürlich den Vektor der Y_j bezeichnet. Mit $\mathbb{P}^Y_{(\mu,\sigma^2)}$ ist aber auch

$$\mathbb{P}^{V^2(Y)}_{(\mu,\sigma^2)} \;=\; \mathbb{P}^{V^2(X)}_{(\mu,\sigma^2)} \;=\; W^{V^2}_{(\mu,\sigma^2)}$$

unabhängig vom Parameter μ (bei festem σ^2) und somit $V^2(X)$ in der Tat verteilungsfrei für $(\mathit{Normal}(\mu,\sigma^2)^n)_{\mu\in\mathbb{R}}$.

Dass das Stichprobenmittel $S(x)=\bar{x}_n$ auch für die Verteilungsfamilien

$$(\mathit{Bin}(m,p)^n)_{p\in(0,1)}, \quad (\mathit{Poisson}(\theta)^n)_{\theta>0} \quad \text{und} \quad (\mathit{NBin}(m,p)^n)_{p\in(0,1)}$$

jeweils eine vollständige Statistik bildet, kann der Leser unter Benutzung von Satz 2.46 leicht selbst nachweisen.

2.6 Gleichmäßig beste erwartungstreue Schätzer

Im Folgenden sei ein statistisches Modell

$$\mathscr{S} \;=\; ((\mathfrak{X},\mathscr{A},(W_\theta)_{\theta\in\Theta}),(D,\mathfrak{D}),L)$$

zum Schätzen einer *reellen* Parameterfunktion $\gamma:\Theta\to\mathbb{R}$ gegeben, wobei $(D,\mathfrak{D})=(\mathbb{R},\mathscr{B}(\mathbb{R}))$ und $L(\theta,d)=(d-\gamma(\theta))^2$ [☞ auch Unterabschnitt 1.4.1]. Die Risikofunktion eines Schätzers $g:(\mathfrak{X},\mathscr{A})\to(\mathbb{R},\mathscr{B}(\mathbb{R}))$ hat somit die Form

$$R(\theta,g) \;=\; \int (g(x)-\gamma(\theta))^2 W_\theta(dx) \;=\; \mathbb{E}_\theta(g(X)-\gamma(\theta))^2.$$

Wie anhand von Beispiel 1.9 gesehen, macht die Suche nach einem gleichmäßig besten Schätzer in nichttrivialen Situationen keinen Sinn, da ein solcher nicht existiert. Man muss deshalb zunächst eine Einschränkung der zugelassenen Schätzfunktionen auf eine vernünftige Teilklasse vornehmen. Eine derartige Einschränkung bildet die Forderung der *Erwartungstreue* an eine Schätzfunktion g:

$$\int g(x)W_\theta(dx) \;=\; \mathbb{E}_\theta g(X) \;=\; \gamma(\theta)$$

für alle $\theta\in\Theta$ [☞ Definition 1.13]. Für das Risiko von g folgt dann

$$R(\theta,g) \;=\; \mathbb{E}_\theta(g(X)-\mathbb{E}_\theta g(X)) \;=\; \mathbb{V}\mathrm{ar}_\theta g(X)$$

für alle $\theta \in \Theta$. Im Anschluss wollen wir uns mit dem Problem beschäftigen, ob und wie eine optimale, gemäß der vorherigen Zeile also *varianzminimierende* Schätzfunktion unter allen erwartungstreuen gefunden werden kann. Wir definieren zuvor:

Definition 2.48. Es bezeichne $\mathscr{U}_\gamma$ die Menge aller erwartungstreuen (engl. *unbiased*) Schätzer für $\gamma(\theta)$. Ein Element $g^* \in \mathscr{U}_\gamma$ heißt *gleichmäßig bester erwartungstreuer Schätzer (GBES) für* $\gamma(\theta)$, wenn für alle $\theta \in \Theta$ und $g \in \mathscr{U}$ gilt:

$$R(\theta, g^*) \;=\; \mathbb{V}\mathrm{ar}_\theta\, g^*(X) \;\leq\; \mathbb{V}\mathrm{ar}_\theta\, g(X) \;=\; R(\theta, g).$$

Wir stellen im Anschluss die kanonischen erwartungstreuen Schätzer für den Mittelwert und die Varianz bei unabhängigen, identisch verteilten Beobachtungen vor. Wir werden später zeigen, dass dieser unter gewissen Voraussetzungen auch GBES bilden.

Beispiel 2.49 (Stichprobenmittel). Beobachtet werde der Zufallsvektor $(X_1,...,X_n)$ mit stochastisch unabhängigen, identisch verteilten ZG $X_1,...,X_n$ unter jedem $\mathbb{P}_\theta$. Sei $\widehat{W}_\theta = \mathbb{P}_\theta^{X_1}$, d.h. $W_\theta = \widehat{W}_\theta^n$. Sofern $\mathbb{E}_\theta|X_1| = \int |x|\,\widehat{W}_\theta(dx) < \infty$ für alle $\theta \in \Theta$, definiert, wie bereits am Anfang von Abschnitt 2.1 festgestellt, das Stichprobenmittel $g_1(x) = \bar{x}_n$ einen erwartungstreuen Schätzer für $\gamma_1(\theta) = \int x\,\widehat{W}_\theta(dx)$ mit (möglicherweise unendlichem) Risiko

$$R(\theta, \bar{x}_n) \;=\; \mathbb{V}\mathrm{ar}_\theta \overline{X}_n \;=\; \frac{1}{n}\,\mathbb{V}\mathrm{ar}_\theta X_1.$$

Beispiel 2.50 (Stichprobenvarianz). Im Fall $\mathbb{E}_\theta X_1^2 = \int x^2 \widehat{W}_\theta(dx) < \infty$ für alle $\theta \in \Theta$ erwarten sicherlich viele, dass die Stichprobenvarianz $\sum_{j=1}^n (x_j - \bar{x}_n)^2/n$ erwartungstreu für die Varianz $\gamma_2(\theta) = \mathbb{V}\mathrm{ar}_\theta X_1$ ist. Diese Vermutung bedarf jedoch einer kleinen Korrektur, die darauf der beruht, dass in diesem Schätzer der Mittelwert $\gamma_1(\theta)$ implizit durch $\bar{x}_n$ mit geschätzt wird. Tatsächlich gilt

$$\begin{aligned}
\mathbb{E}_\theta \left(\sum_{j=1}^n (X_j - \overline{X}_n)^2 \right) \;&=\; \mathbb{E}_\theta \left(\sum_{j=1}^n X_j^2 - 2\overline{X}_n \sum_{j=1}^n X_j + n\overline{X}_n^2 \right) \\
&=\; \mathbb{E}_\theta \left(\sum_{j=1}^n X_j^2 - n\overline{X}_n^2 \right) \\
&=\; n\,\mathbb{E}_\theta X_1^2 - n\,\mathbb{V}\mathrm{ar}_\theta \overline{X}_n - n(\mathbb{E}_\theta \overline{X}_n)^2 \\
&=\; n\,\gamma_2(\theta) + n\gamma_1(\theta)^2 - \gamma_2(\theta) - n\,\gamma_1(\theta)^2 \\
&=\; (n-1)\gamma_2(\theta),
\end{aligned}$$

und somit definiert nicht $\sum_{j=1}^n (x_j - \bar{x}_n)^2/n$, sondern

$$g_2(x) \;=\; (n-1)^{-1}\sum_{j=1}^{n}(x_j-\bar{x}_n)^2$$

für $n \geq 2$ einen erwartungstreuen Schätzer für $\gamma_2(\theta)$. Ist der Mittelwert $\gamma_1(\theta) \equiv \mu$ konstant und bekannt, so definiert auch $g_3(x) = n^{-1}\sum_{j=1}^{n}(x_j-\mu)^2$ einen erwartungstreuen Schätzer für $\gamma_2(\theta)$, wie man sofort nachrechnet. Da in der Literatur die Bezeichnung "Stichprobenvarianz" sowohl für $n^{-1}\sum_{j=1}^{n}(x_j-\bar{x}_n)^2$ als auch für $(n-1)^{-1}\sum_{j=1}^{n}(x_j-\bar{x}_n)^2$ verwendet wird, sollte man sich beim Lesen von Statistik-Texten stets der jeweils gewählten Definition vergewissern.

Um als nächstes der Frage nachzugehen, wie zu einem erwartungstreuen Schätzer g für $\gamma(\theta)$ ein besserer konstruiert werden kann, nehmen wir an, dass im betrachteten Schätzmodell $\mathscr{S}$ eine suffiziente Statistik $T : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}', \mathscr{A}')$ gegeben ist. Gemäß Bemerkung 2.23 können wir dann die Funktion $g^* : \mathfrak{X}' \to \mathbb{R}$ durch

$$g^*(t) \;=\; W(g|T=t) \;=\; \mathbb{E}(g(X)|T(X)=t)$$

definieren und erhalten mit $g^* \circ T : \mathfrak{X} \to \mathbb{R}$ i.A. eine neue Schätzfunktion für $\gamma(\theta)$, die tatsächlich eine Verbessserung bildet, wie der nachfolgende auf C.R. RAO [17] und D. BLACKWELL [6] zurückgehende Satz zeigt.

Satz 2.51. (von Rao-Blackwell) *Für den zuvor definierten Schätzer $g^* \circ T$ für $\gamma(\theta)$ gilt:*

(a) $g^* \circ T \in \mathscr{U}_\gamma$, *d.h. $g^* \circ T$ ist ebenfalls erwartungstreu.*

(b) $\mathbb{V}\mathrm{ar}_\theta g^* \circ T(X) \leq Var_\theta g(X)$ *für alle $\theta \in \Theta$.*

(c) *Für alle $\theta \in \Theta$ mit $Var_\theta g(X) < \infty$ gilt:*

$$\mathbb{V}\mathrm{ar}_\theta g^* \circ T(X) = Var_\theta g(X) \quad \Leftrightarrow \quad g^* \circ T = g \;\; W_\theta\text{-f.s.}$$

Beweis. Aussage (a) ist trivial, und dasselbe gilt für (b), falls $\mathbb{V}\mathrm{ar}_\theta g(X) = \infty$. Sei also ein $\theta \in \Theta$ mit $\mathbb{V}\mathrm{ar}_\theta g(X) < \infty$ gegeben. Dann erhalten wir

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_\theta g(X) \;&=\; \mathbb{E}_\theta(g(X)-\gamma(\theta))^2 \\
&=\; \mathbb{E}_\theta\{(g(X)-g^*\circ T(X))+(g^*\circ T(X)-\gamma(\theta))\}^2 \\
&=\; \mathbb{E}_\theta(g(X)-g^*\circ T(X))^2 + \mathbb{V}\mathrm{ar}_\theta g^*\circ T(X) \qquad (2.12)\\
&\quad\; + 2\,\mathbb{E}_\theta\{(g(X)-g^*\circ T(X))(g^*\circ T(X)-\gamma(\theta))\} \\
&\geq\; \mathbb{V}\mathrm{ar}_\theta g^*\circ T(X) + 2\mathbb{E}_\theta\{(g(X)-g^*\circ T(X))(g^*\circ T(X)-\gamma(\theta))\}.
\end{aligned}$$

Die Behauptung ergibt sich nun, weil der zweite Term in der letzten Zeile verschwindet. Es gilt nämlich

$$\mathbb{E}_\theta\{(g(X)-g^*\circ T(X))(g^*\circ T(X)-\gamma(\theta))\}$$

$$= \mathbb{E}_\theta\{\mathbb{E}(g(X) - g^* \circ T(X))|T(X))(g^* \circ T(X) - \gamma(\theta))\}$$
$$= \mathbb{E}_\theta\{(g^* \circ T(X) - g^* \circ T(X))(g^* \circ T(X) - \gamma(\theta))\} = 0$$

unter Ausnutzung der Definition von g^* beim vorletzten Gleichheitszeichen.

Zum Nachweis von (c) genügt der Hinweis, dass aus (2.12) weiter

$$\begin{aligned} \mathbb{V}\mathrm{ar}_\theta g^* \circ T(X) = \mathbb{V}\mathrm{ar}_\theta g(X) \quad &\Leftrightarrow \quad \mathbb{E}_\theta\{(g(X) - g^* \circ T(X))\}^2 = 0 \\ &\Leftrightarrow \quad g - g^* \circ T = 0 \ \ W_\theta\text{-f.s.} \end{aligned}$$

folgt, sofern $\mathbb{V}\mathrm{ar}_\theta g(X) < \infty$ vorausgesetzt wird. □

Satz 2.51 zeigt also, dass zu gegebenem erwartungstreuen Schätzer g bei Vorliegen einer suffizienten Statistik T durch Bedingen stets ein mindestens ebenso guter erwartungstreuer Schätzer $g^* \circ T$ konstruiert werden kann. Aus Teil (c) ergibt sich außerdem, dass $g^* \circ T$ sogar eine echte Verbesserung darstellt, falls für wenigstens ein θ die Varianz von g unter W_θ endlich ist und g nicht bereits selbst schon W_θ-f.s. eine Funktion von T bildet. Bevor wir die Nützlichkeit dieses Satzes anhand von Beispielen demonstrieren, wollen wir ein weiteres Resultat beweisen, das sogar die Optimalität eines so erhaltenen Schätzers $g^* \circ T$ zeigt, sofern T auch noch vollständig ist. Es geht zurück auf LEHMANN & SCHEFFÉ [12] und ist deshalb nach ihnen benannt.

Satz 2.52. (von Lehmann-Scheffé) *In der Situation des Satzes von Rao-Blackwell sei T außerdem vollständig. Dann gilt für jeden Schätzer $g \in \mathscr{U}_\gamma$, dass $g^* \circ T$ einen GBES bildet, wobei $g^* = W(g|T = \cdot)$.*

Beweis. Sei $g \in \mathscr{U}_\gamma$ und $g^* = W(g|T = \cdot)$. Wir müssen zeigen, dass für alle $\theta \in \Theta$ und $h \in \mathscr{U}_\gamma$

$$\mathbb{V}\mathrm{ar}_\theta g^* \circ T(X) \leq \mathbb{V}\mathrm{ar}_\theta h(X)$$

gilt. Für beliebiges $h \in \mathscr{U}_\gamma$ setzen wir $h^* = W(h|T = \cdot)$. Dann folgt aus Satz 2.51, dass auch $h^* \circ T \in \mathscr{U}_\gamma$ sowie

$$\mathbb{V}\mathrm{ar}_\theta h^* \circ T(X) \leq \mathbb{V}\mathrm{ar}_\theta h(X)$$

für alle $\theta \in \Theta$. Zudem liefert die Erwartungstreue für alle $\theta \in \Theta$

$$\int_{\mathfrak{X}'} h^*(t)\, W_\theta^T(dt) = \mathbb{E}_\theta h^* \circ T(X) = \gamma(\theta) = \mathbb{E}_\theta g^* \circ T(X) = \int_{\mathfrak{X}'} g^*(t)\, W_\theta^T(dt),$$

also $\int_{\mathfrak{X}'}(h^*(t) - g^*(t))W_\theta^T(dt) = 0$, so dass $h^* = g^*(W_\theta^T)_{\theta\in\Theta}$-f.s. aufgrund der Vollständigkeit von T folgt. Die Schätzer $h^* \circ T$ und $g^* \circ T$ stimmen somit $(W_\theta)_{\theta\in\Theta}$-f.s. überein, und bezüglich des Risikos erhalten wir für alle $\theta \in \Theta$

$$\mathbb{V}\mathrm{ar}_\theta g^* \circ T(X) = \mathbb{V}\mathrm{ar}_\theta h^* \circ T(X) \leq \mathbb{V}\mathrm{ar}_\theta h(X),$$

was die gewünschte Optimalität beweist, da h beliebig gewählt war. □

Mittels Satz 2.51 erhält man auch leicht die folgende Eindeutigkeitsaussage:

Korollar 2.53. *Existiert in einem Schätzmodell mit vollständiger und suffizienter Statistik $T : (\mathfrak{X}, \mathscr{A}) \to (\mathfrak{X}', \mathscr{A}')$ ein GBES g für $\gamma(\theta)$, so ist dieser bereits W_θ-f.s. eindeutig bestimmt für alle $\theta \in \Theta$ mit $\mathbb{V}\mathrm{ar}_\theta g(X) < \infty$.*

Beweis. Gegeben zwei GBES g, h für $\gamma(\theta)$ mit folglich identischen Varianzen unter jedem W_θ, liefert zunächst Satz 2.51 $g = g^* \circ T$ und $h = h^* \circ T$ W_θ-f.s. für alle $\theta \in \Theta := \{\theta \in \Theta : \mathbb{V}\mathrm{ar}_\theta g(X) < \infty\}$. Wie im vorherigen Beweis gezeigt, impliziert die Erwartungstreue beider Schätzer in Verbindung mit der Vollständigkeit von T außerdem $g^* = h^*(W_\theta^T)_{\theta\in\Theta}$-f.s. und somit $g = h$ $(W_\theta)_{\theta\in\Theta}$-f.s. □

Mit Hilfe des Satzes von Lehmann und Scheffé und dessen Korollar zur Eindeutigkeit kann man nun in einer großen Zahl von Anwendungen den GBES auf eine der beiden folgenden Weisen bestimmen:

(1) Durch Angabe eines beliebigen erwartungstreuen Schätzers, der über einer suffizienten und vollständigen Statistik faktorisiert.
(2) Durch Bedingen eines beliebigen erwartungstreuen Schätzers unter einer vollständigen und suffizienten Statistik.

Dabei kann die zweite Methode allerdings mit einigem Rechenaufwand verbunden sein. In den ersten drei der anschließenden Beispielen benutzen wir Methode 1, während das vierte eine Anwendung von Methode 2. bildet.

Im Folgenden sei $X = (X_1, ..., X_n)$ stets ein Zufallsvektor mit unter jedem $\mathbb{P}_\theta$ unabhängigen und identisch verteilten Komponenten. Ferner seien wie schon früher $\widehat{W}_\theta = \mathbb{P}_\theta^{X_1}$, d.h. $W_\theta = \widehat{W}_\theta^n$, und $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$, $x = (x_1, ..., x_n)$ und $s_n = x_1 + ... + x_n$.

Beispiel 2.54. Bildet $(\widehat{W}_\theta)_{\theta\in\Theta}$ eine 1-parametrige Exponentialfamilie in $Q(\theta)$ und $\widehat{T}(x) = x$ bzgl. ν, d.h.

$$\frac{d\widehat{W}_\theta}{d\nu}(x) \;=\; C(\theta)e^{Q(\theta)x} \quad \nu\text{-f.ü.},$$

so ist $\mathscr{W}$ gemäß Satz 1.26(b) eine 1-parametrige Exponentialfamilie in $Q(\theta)$ und der suffizienten Statistik $T(x) = s_n$ [☞ Korollar 2.31]. Setzen wir voraus, dass die Menge $\{Q(\theta) : \theta \in \Theta\}$ innere Punkte hat und somit T auch vollständig ist [☞ Satz 2.46], so folgt nun sofort aus Satz 2.52, dass das Stichprobenmittel $g(x) = \overline{x}_n$ der GBES für $\gamma(\theta) = \mathbb{E}_\theta X_1$ ist. Dieses allgemeine Resultat können wir beispielsweise auf die Spezialfälle $(Bin(m,p)^n)_{p\in(0,1)}$, $(Poisson(\theta)^n)_{\theta>0}$, $(NBin(m,p)^n)_{p\in(0,1)}$, $(Exp(\theta)^n)_{\theta>0}$ oder auch $(Normal(\mu,\sigma^2)^n)_{\mu\in\mathbb{R}}$ mit festem $\sigma^2 > 0$ anwenden.

Beispiel 2.55 (Normalverteilungen). Falls $\widehat{W}_\theta = \mathit{Normal}(\mu,\sigma^2)$ für $\theta = (\mu,\sigma^2) \in \mathbb{R}\times(0,\infty)$, so bildet $\mathscr{W}$ offenkundig eine 2-parametrige Exponentialfamilie, die den Voraussetzungen in Satz 2.52 genügt, und deshalb $T(x) = (\bar{x}_n, n^{-1}\sum_{j=1}^n (x_j-\bar{x}_n)^2)$ eine vollständige und suffiziente Statistik für $\mathscr{W}$ [☞ Beispiel 2.40]. In Beispiel 2.49 wurde gezeigt, dass

$$g_1(x) \;=\; \bar{x}_n \quad \text{und} \quad g_2(x) \;=\; \frac{1}{n-1}\sum_{j=1}^n (x_j-\bar{x}_n)^2$$

erwartungstreue Schätzer für $\mathbb{E}_\theta X_1 = \mu$ bzw. $\mathbb{V}\mathrm{ar}_\theta X_1 = \sigma^2$ sind. Da diese offenkundig von x nur über $T(x)$ abhängen, liefert erneut Satz 2.52, dass g_1 und g_2 bereits die GBES für μ bzw. σ^2 bilden.

Beispiel 2.56 (Gleichverteilungen). Betrachten wir als nächstes den Fall, dass $\widehat{W}_\theta = \mathit{Unif}(0,\theta)$ für $\theta > 0$. Dieses Beispiel unterscheidet sich von den vorherigen darin, dass $\mathscr{W} = (\mathit{Unif}((0,\infty)^n))_{\theta>0}$ *keine* Exponentialfamilie bildet, wie sich sofort aus der Nicht-Äquivalenz der W_θ ergibt. Wegen

$$f_\theta(x) \;:=\; \frac{dW_\theta}{d\lambda^n}(x) \;=\; \frac{1}{\theta^n}\prod_{j=1}^n \mathbf{1}_{(0,\theta)}(x_j) \;=\; \frac{1}{\theta^n}\mathbf{1}_{(0,\theta)}(x_{(n)}) \quad \lambda^n\text{-f.ü.}$$

folgern wir aus dem Neyman-Kriterium die Suffizienz von $T(x) = x_{(n)} := \max_{j\le n} x_j$. Als nächstes zeigen wir, dass T sogar vollständig ist für $\mathscr{W}$. Offensichtlich gilt

$$W_\theta(T\le t) \;=\; \mathbb{P}_\theta(X_{(n)}\le t) \;=\; \mathbb{P}_\theta(X_1\le t,\ldots,X_n\le t) \;=\; \mathbb{P}_\theta(X_1\le t)^n \;=\; \frac{t^n}{\theta^n}$$

für alle $\theta > 0$ und $t\in(0,\theta)$, so dass

$$\frac{dW_\theta^T}{d\lambda}(t) \;=\; \frac{nt^{n-1}}{\theta^n}\mathbf{1}_{(0,\theta)}(t) \quad \lambda\text{-f.ü.} \tag{2.13}$$

Sei nun f eine messbare numerische Funktion mit $\mathbb{E}_\theta f\circ T(X) = 0$ für alle $\theta > 0$, was nach dem soeben Gezeigten äquivalent ist zu

$$\int_{(0,\theta)} f^+(t)t^{n-1}\,\lambda(dt) \;=\; \int_{(0,\theta)} f^-(t)t^{n-1}\,\lambda(dt)$$

für alle $\theta > 0$. Dann gilt aber auch

$$\begin{aligned} \nu^+((a,b]) \;&:=\; \int_{(a,b]} f^+(t)t^{n-1}\,\lambda(dt) \\ &=\; \int_{(a,b]} f^-(t)t^{n-1}\,\lambda(dt) \;=:\; \nu^-((a,b]) \end{aligned} \tag{2.14}$$

für alle $0\le a<b<\infty$, wobei ν^-,ν^+ die Borel-Maße auf $(0,\infty)$ mit λ-Dichten $f^-(t)t^{n-1}$ bzw. $f^+(t)t^{n-1}$ bezeichnen. Da die halboffenen Intervalle einen $\cap$-stabilen Erzeuger von $\mathscr{B}((0,\infty))$ bilden, impliziert (2.14) die Gleichheit von ν^- und

ν^+ sowie folglich auch die λ-f.ü. gültige Gleichheit ihrer Dichten, was schließlich $f^+ = f^-$ λ-f.ü. und damit die Vollständigkeit von T beweist. Unter Benutzung von (2.13) erhalten wir leicht

$$\mathbb{E}_\theta T(X) \;=\; \int_0^\theta \frac{nt^n}{\theta^n}\, dt \;=\; \frac{n\theta}{n+1}$$

und somit die Erwartungstreue des Schätzers $g(x) = (1+\frac{1}{n})x_{(n)}$ für $\gamma(\theta) = \theta$. Satz 2.52 liefert einmal mehr, dass g bereits der GBES für θ ist.

Wer Lust hat, kann zur Übung folgendes rechenaufwendigere Vorgehen wählen: Wegen $\mathbb{E}_\theta X_1 = \theta/2$ definiert $\widehat{g}(x) = 2x_1$ einen ersten erwartungstreuen Schätzer für θ. Mit Methode (2) ergibt sich dann wiederum der obige Schätzer g als GBES nach Berechnung von $\widehat{g}^*(t) = W(\widehat{g}|T=t)$ und Setzen von $g(x) = \widehat{g}^* \circ T(x)$.

Beispiel 2.57 (Poisson-Verteilungen). Als letztes Beispiel wollen wir mittels Methode (2) für die Familie $\widehat{\mathscr{W}} = (\mathit{Poisson}(\theta))_{\theta>0}$ den GBES für

$$\gamma_k(\theta) \;=\; \widehat{W}_\theta(\{k\}) \;=\; e^{-\theta}\,\frac{\theta^k}{k!}, \quad k \in \mathbb{N}_0,$$

berechnen. Wie man direkt sieht, bildet $\mathscr{W} = (\mathit{Poisson}(\theta)^n)_{\theta>0}$ eine 1-parametrige Exponentialfamilie vollen Rangs in θ und der Statistik $T(x) := s_n$, die daher suffizient und vollständig für das betrachtete Experiment ist. Ein offensichtlicher erwartungstreuer Schätzer für $\gamma_k(\theta)$ lautet $g(x) = \mathbf{1}_{\{k\}}(x_1)$, denn

$$\mathbb{E}_\theta g(X) \;=\; \mathbb{P}_\theta(X_1 = k) \;=\; e^{-\theta}\,\frac{\theta^k}{k!} \;=\; \gamma_k(\theta)$$

für alle $\theta > 0$. Im trivialen Fall $n = 1$ ist g als Funktion von $T(x) = T(x_1) = x_1$ bereits der GBES. Wir notieren im Hinblick auf das allgemeine Ergebnis weiter unten, dass g auch in der Form $g(x) = \mathit{Bin}(x_1, \frac{1}{n})(\{k\})$ geschrieben werden kann, sofern wir die übliche Vereinbarung $\mathit{Bin}(m,0) := \delta_0$ und $\mathit{Bin}(m,1) := \delta_m$ für alle $m \in \mathbb{N}_0$ treffen. Sei hiernach $n \geq 2$ vorausgesetzt. Wir müssen dann zur Bestimmung des GBES

$$g^*(t) \;=\; W(g(x)|T=t) \;=\; \mathbb{P}(X_1 = k|S_n = t)$$

für beliebige $t \in \mathbb{N}_0$ berechnen. Unter Benutzung von $W_\theta^T = \mathbb{P}_\theta^{S_n} = \mathit{Poisson}(n\theta)$ für alle $\theta > 0$ und $n \geq 1$ erhalten wir

$$\begin{aligned}
\mathbb{P}(X_1 = k|S_n = t) \;&=\; \frac{\mathbb{P}_1(X_1 = k, S_n = t)}{\mathbb{P}_1(S_n = t)} \\
&=\; \frac{\mathbb{P}_1(X_1 = k, X_2 + \ldots + X_n = t-k)}{\mathbb{P}_1(S_n = t)} \\
&=\; \frac{\mathbb{P}_1(X_1 = k)\mathbb{P}_1(S_{n-1} = t-k)}{\mathbb{P}_1(S_n = t)}
\end{aligned}$$

$$= \frac{\left(e^{-1}\frac{1}{k!}\right)\left(e^{-(n-1)}\frac{(n-1)^{t-k}}{(t-k)!}\right)}{e^{-n}\frac{n^t}{t!}}\,\mathbf{1}_{\mathbb{N}_0}(t-k)$$

$$= \begin{cases} 0, & \text{falls } t<k, \\ \binom{t}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{t-k}, & \text{falls } t\geq k \end{cases}$$

$$= Bin\left(t,\frac{1}{n}\right)(\{k\}).$$

Der GBES für $\gamma_k(\theta)=e^{-\theta}\theta^k/k!$ lautet demnach für alle $k\in\mathbb{N}_0$ und $n\geq 1$

$$g^*\circ T(x) \;=\; Bin\left(T(x),\frac{1}{n}\right)(\{k\}) \;=\; Bin\left(n\bar{x}_n,\frac{1}{n}\right)(\{k\}),$$

d.h., die Poisson-Wahrscheinlichkeiten $\widehat{W}_\theta(\{k\})$ werden erwartungstreu am besten mittels Binomial-Wahrscheinlichkeiten der angegebenen Form geschätzt. Da vermöge des Poissonschen Grenzwertsatzes 29.4 in [2] (beachte

$$\lim_{n\to\infty} n\overline{X}_n\cdot\frac{1}{n} \;=\; \lim_{n\to\infty}\overline{X}_n \;=\; \theta \quad \mathbb{P}_\theta\text{-f.s.}$$

für alle $\theta>0$ nach dem starken Gesetz der großen Zahlen)

$$\lim_{n\to\infty} Bin\left(n\overline{X}_n,\frac{1}{n}\right)(\{k\}) \;=\; Poisson(\theta)(\{k\}) \quad \mathbb{P}_\theta\text{-f.s.}$$

für alle $\theta>0$ und $k\in\mathbb{N}_0$ gilt, folgt die starke Konsistenz (☞ zu Beginn von Abschnitt 2.1) des erhaltenen Schätzers $g^*\circ T(x)$ bei gegen ∞ strebendem Stichprobenumfang n.

Vergleichen wir unser Ergebnis zum Abschluss mit dem im Fall einer Maximum-Likelihood-Schätzung. In 2.8 hatten wir gesehen, dass $\bar{x}_n$ den MLS für θ bildet. Gemäß Satz 2.10 ist deshalb

$$\gamma_k(\bar{x}_n) \;=\; e^{-\bar{x}_n}\frac{\bar{x}_n^k}{k!} \;=\; Poisson(\bar{x}_n)(\{k\})$$

der MLS für $\gamma_k(\theta)$ für alle $k\in\mathbb{N}_0$. Mittels der Maximum-Likelihood-Methode werden also die unbekannten Poisson-Wahrscheinlichkeiten durch solche vom selben Typ geschätzt, wobei man den unbekannten Parameter θ durch den MLS $\bar{x}_n$ ersetzt ($Poisson(0):=\delta_0$). Auch hier ergibt sich leicht die starke Konsistenz, weil die Funktionen $\gamma_k(\theta)$ stetig in θ sind.

Anmerkung 2.58. Die in 2.57 gewählte Vorgehensweise kann man wie folgt allgemein bescheiben: Gegeben eine Verteilungsfamilie $\mathscr{W}=(W_\theta)_{\theta\in\Theta}$ mit einem *eindimensionalen* Parameter θ, finde zunächst eine vollständige und suffiziente Statistik

T (sofern möglich) und berechne dann

$$f(\theta) \;:=\; \mathbb{E}_\theta T(X).$$

Erweist sich $f : \Theta \to \mathbb{R}$ als bijektive (= nicht konstante) *lineare* Funktion in θ, so definiert offensichtlich $g(x) = f^{-1} \circ T(x)$ den GBES für θ.

Sofern $|\Theta| \geq 2$, $\mathscr{W}$ dominiert ist und $W_\theta \neq W_{\theta'}$, falls $\theta \neq \theta'$, sichert übrigens die Suffizienz und Vollständigkeit von T die Bijektivität von f, denn: Gemäß Lemma 2.25 existiert ein zu $\mathscr{W}$ äquivalentes W-Maß ν. Nehmen wir an, $f \equiv c$ für ein $c \in \mathbb{R}$. Dann folgt $T(X) = c$ $\mathscr{W}$-f.s. und somit ν-f.s. aus der Vollständigkeit von T. Wegen der Suffizienz von T existieren nach dem Neyman-Kriterium ferner Funktionen g_θ und h, so dass

$$\frac{dW_\theta}{d\nu} \;=\; g_\theta \circ T \cdot h \;=\; g_\theta(c) \cdot h \quad \nu\text{-f.ü.}$$

für alle $\theta \in \Theta$, was offenbar $g_\theta(c) = (h(x)\nu(dx))^{-1}$ und damit $W_\theta = W_{\theta'}$ für alle $\theta, \theta' \in \Theta$ impliziert.

In allen vorherigen Betrachtungen hatten wir entweder unterstellt oder sofort eingesehen, dass zu einer gegebenen Parameterfunktion γ überhaupt ein erwartungstreuer Schätzer existiert, also $\mathscr{U}_\gamma \neq \emptyset$ gilt. Wir nennen γ dann *schätzbar*. Das anschließende Beispiel soll zeigen, dass dies keineswegs immer der Fall ist.

Beispiel 2.59 (Hypergeometrische Verteilungen und inverses Bernoulli-Sampling). Ein Biologen-Team will die Anzahl θ einer bestimmten Fischart in einem See schätzen und fängt zu diesem Zweck r Fische dieser Art, markiert sie und setzt sie wieder im See aus. Nach einiger Zeit fängt das Team erneut s Fische derselben Art und registriert die Anzahl X der markierten Exemplare, die sich darunter befinden.

Dieses Experiment entspricht unter gewissen idealisierenden Voraussetzungen offenkundig dem Ziehen ohne Zurücklegen von s Kugeln aus einer Urne mit einer unbekannten Gesamtzahl θ von Kugeln, von denen genau r eine Markierung tragen und die anderen $\theta - r$ nicht. Wir wählen daher als statistisches Experiment

$$\widetilde{\mathscr{E}} \;=\; \big(\{0,...,s\}, \mathfrak{P}(\{0,...,s\}), (HGeom(\theta,r,s))_{\theta \geq r \vee s}\big),$$

wobei $HGeom(\theta,r,s)$ die hypergeometrische Verteilung mit Parametern θ (= Anzahl der Kugeln in der Urne), r (= Anzahl der markierten Kugeln) und s (= Anzahl der gezogenen Kugeln) bezeichnet. Es gilt somit

$$\begin{aligned}\mathbb{P}_\theta(X = n) \;&=\; HGeom(\theta,r,s)(\{n\}) \\ &=\; \begin{cases} \dfrac{\binom{r}{n}\binom{\theta-r}{s-n}}{\binom{\theta}{s}}, & \text{falls } (r+s-\theta \vee 0) \leq n \leq (r \wedge s), \\ 0, & \text{sonst.} \end{cases}\end{aligned}$$

Will man nun θ schätzen, so kann dies auf keinen Fall erwartungstreu geschehen, denn jeder Schätzer $g : \{0,...,s\} \to \mathbb{R}$ ist notwendigerweise beschränkt, wohingegen der Parameterraum Θ unbeschränkt ist. Aus demselben Grund gibt es auch für keine andere unbeschränkte Parameterfunktion γ einen erwartungstreuen Schätzer.

Die interessante Frage lautet daher: Kann das Biologen-Team dennoch zu einer erwartungstreuen Schätzung von θ gelangen, indem es sein Vorgehen bei der Erhebung der zweiten Stichprobe modifiziert? Um eine Antwort zu erhalten, sucht das Team einen Statistiker auf, der ihnen zu folgendem sequentiellen Vorgehen rät, genannt *inverses Bernoulli-Sampling*: Nach Vorgabe einer Zahl $m \geq 1$ beginnt das Team erneut, Fische der betrachteten Art zu fangen, und zwar einen nach dem anderen. Jeder Fang wird registriert als "1" oder "0", je nachdem, ob er eine Markierung aufweist oder nicht, und dann in den See zurückgeworfen. Das Team hört auf, sobald der m-te markierte Fisch ins Netz geht. Es bezeichne N die Zufallsgröße, die die Anzahl notwendiger Fänge angibt, und X_n für $1 \leq n \leq N$ die Bernoulli-Variable, die das n-te Fangergebnis beschreibt. Folglich ist $X = (N, X_1, ..., X_N)$ der zugrundeliegende Beobachtungsvektor. Unter der idealisierenden Annahme, dass für jeden gefangenen Fisch die Wahrscheinlichkeit $p(\theta) = r/\theta$ beträgt, eine Markierung aufzuweisen, gelangen wir zu dem statistischen Experiment $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ mit

$$\mathfrak{X} = \bigcup_{n\geq m} \{n\} \times \{0,1\}^{n-1} \times \{1\}, \quad \mathscr{A} = \mathfrak{P}(\mathfrak{X}), \quad \Theta = \{r, r+1, ...\}$$

und

$$\begin{aligned} &W_\theta(\{(n, x_1, ..., x_{n-1}, 1)\}) \\ &\quad = \mathbb{P}_\theta(N = n, X_1 = x_1, ..., X_{n-1} = x_{n-1}, X_n = 1) \\ &\quad = \begin{cases} \mathbb{P}_\theta(X_1 = x_1, ..., X_{n-1} = x_{n-1}, X_n = 1), & \text{falls } \sum_{j=1}^{n-1} x_j = m-1, \\ 0, & \text{sonst} \end{cases} \\ &\quad = \begin{cases} \left(\frac{r}{\theta}\right)^m \left(1 - \frac{r}{\theta}\right)^{n-m}, & \text{falls } \sum_{j=1}^{n-1} x_j = m-1, \\ 0, & \text{sonst.} \end{cases} \end{aligned}$$

$N - m$, die Anzahl nicht markierter Fänge bis zum Fang des m-ten markierten Fisches, besitzt unter $\mathbb{P}_\theta$ eine $NBin(m, r/\theta)$-Verteilung, wie man leicht nachrechnet, d.h.

$$\mathbb{P}_\theta(N - m = n) \;=\; \binom{m+n-1}{m-1} \left(\frac{r}{\theta}\right)^m \left(1 - \frac{r}{\theta}\right)^n$$

für $k \in \mathbb{N}_0$. Mit Hilfe des Neyman-Kriteriums sieht man, dass $T(n, x_1, ..., x_n) := n - m$ suffizient für $\mathscr{E}$ ist. Weiter folgt

$$\mathbb{E}_\theta T(X) \;=\; \mathbb{E}_\theta(N - m) \;=\; \frac{m(1 - r/\theta)}{r/\theta} \;=\; \frac{m(\theta - r)}{r},$$

und daraus die Erwartungstreue des Schätzers

$$g(n,x_1,...,x_n) \;=\; \frac{rn}{m}$$

für θ [☞ Anm. 2.58]. Um diesen auch als GBES zu identifizieren, benötigt man noch die Vollständigkeit von T. Hierbei tritt allerdings ein Problem auf, das uns zwingt, unser Ziel ein wenig bescheidener zu stecken. Betrachten wir eine beliebige messbare Funktion $f : \mathbb{N}_0 \to \mathbb{R}$ derart, dass

$$\mathbb{E}_\theta f \circ T(X) \;=\; \left(\frac{r}{\theta}\right)^m \sum_{n\geq 0} f(n) \binom{n+m-1}{m-1} \left(1-\frac{r}{\theta}\right)^n \;=\; 0$$

für alle $\theta \in \Theta$. Dann wäre es naheliegend, $f(n)\binom{n+m-1}{m-1} = 0$ und somit $f(n) = 0$ für alle $n \geq 0$ unter Benutzung des Identitätssatzes für Potenzreihen zu schließen. Unglücklicherweise hat die Menge $\{1-\frac{r}{\theta} : \theta \in \Theta\}$ lediglich den Häufungspunkt 1, der i.A. nicht im Inneren des Konvergenzkreises liegt. Um dennoch zu einer Optimalitätseigenschaft von g zu gelangen, betrachten wir die erweiterte Verteilungsfamilie $\mathscr{W}' = (W_\theta)_{\theta\in\Theta'}$, wobei $\Theta' := [r,\infty)$. In diesem Fall folgt $\{1-\frac{r}{\theta} : \theta \in \Theta'\} = [0,1)$ und deshalb in der Tat die Vollständigkeit von T für $\mathscr{W}'$ nach dem Identitätssatz. Ferner bleibt T natürlich suffizient. Nun bildet $g(n,x_1,...,x_n) = rn/m$ aber auch im zugehörigen erweiterten Experiment $\mathscr{E}' = (\mathfrak{X},\mathscr{A},\mathscr{W}')$ einen erwartungstreuen Schätzer für θ, wie man sofort einsieht, also nach dem eben Gezeigten sogar den gleichmäßig besten. Wir können aber nicht ausschließen, und hierin besteht die Einschränkung, dass es einen besseren erwartungstreuen Schätzer für das ursprüngliche Modell mit kleinerem Parameterraum gibt, der nicht mehr erwartungstreu im erweiterten Experiment $\mathscr{E}'$ ist.

Nachdem wir eingesehen haben, dass die erwartungstreuen Schätzer eine vernünftige Klasse bilden, innerhalb derer die Berechnung eines gleichmäßig besten Schätzers bei Vorliegen einer suffizienten und vollständigen Statistik eine lösbare Optimierungsaufgabe darstellt, wollen wir den Abschnitt mit einer Enttäuschung beenden: Wir zeigen, gleichsam als Mahnung und zur Schärfung des kritischen Urteilsvermögens, *dass ein GBES nicht zulässig* [☞ Definition 1.12] *zu sein braucht*. Dabei verblüfft besonders, dass wir hierfür keineswegs ein pathologisches Beispiel bemühen müssen, sondern bereits bei der wohlbekannten Stichprobenvarianz bei normalverteilgen Beobachtungen fündig werden. Es folgt zunächst ein Lemma, mit dessen Hilfe sich das Risiko der Stichprobenvarianz bei quadratischer Verlustfunktion berechnen lässt.

Lemma 2.60. *Gegeben unabhängige, identisch verteilte $X_1,...,X_n$ mit Varianz σ^2, endlichem vierten zentralen Moment $\mu_4 = \mathbb{E}(X_1 - \mathbb{E}X_1)^4$ und Stichprobenmittel $\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$, gilt*

$$\mathbb{E}\left(\sum_{k=1}^{n}(X_k-\overline{X}_n)^2\right)^2 \;=\; \frac{(n-1)^2}{n}\,\mu_4 \;+\; \frac{(n-1)(n(n-2)+3)}{n}\,\sigma^4. \qquad (2.15)$$

Beweis. Wir dürfen o.B.d.A. $\mathbb{E}X_1 = 0$ voraussetzen. Eine einfache Rechnung liefert

$$\begin{aligned}\mathbb{E}\left(\sum_{k=1}^{n}(X_k-\overline{X}_n)^2\right)^2 \;&=\; \mathbb{E}\left(\sum_{k=1}^{n}X_k^2-2\sum_{k=1}^{n}X_k\overline{X}_n+n\overline{X}_n^2\right)^2 \\ &=:\; I_1-I_2+I_3,\end{aligned} \qquad (2.16)$$

wobei

$$I_1 := \mathbb{E}\left(\sum_{k=1}^{n}X_k^2\right)^2,\; I_2 := \frac{2}{n}\,\mathbb{E}\left(\sum_{k=1}^{n}X_k^2\right)\left(\sum_{k=1}^{n}X_k\right)^2 \text{ und } I_3 := \frac{1}{n^2}\,\mathbb{E}\left(\sum_{k=1}^{n}X_k\right)^4.$$

Für die drei Ausdrücke I_1, I_2 und I_3 erhalten wir unter Benutzung von $\mathbb{E}X_1 = 0$, der Unabhängigkeit der X_k sowie $\sum_{1\le i<j\le n}1 = \sum_{i=1}^{n}(i-1) = \frac{n(n-1)}{2}$:

$$I_1 \;=\; \mathbb{E}\left(\sum_{k=1}^{n}X_k^4\right) \;+\; 2\,\mathbb{E}\left(\sum_{1\le i<j\le n}X_i^2X_j^2\right) \;=\; n\mu_4+n(n-1)\sigma^4,$$

$$I_2 \;=\; \frac{2}{n}\,\mathbb{E}\left(\sum_{k=1}^{n}X_k^2\right)^2 \;+\; \frac{4}{n}\,\mathbb{E}\left(\sum_{k=1}^{n}X_k^2\left(\sum_{1\le i<j\le n}X_iX_j\right)\right) \;=\; \frac{2}{n}I_1+\frac{4}{n}0 \;=\; \frac{2}{n}I_1$$

und

$$I_3 \;=\; \frac{1}{n^2}\,\mathbb{E}\left(\sum_{k=1}^{n}X_k^4\right) \;+\; \frac{1}{n^2}\binom{4}{2}\,\mathbb{E}\left(\sum_{1\le i<j\le n}X_i^2X_j^2\right) \;=\; \frac{1}{n}\mu_4 \;+\; \frac{3(n-1)}{n}\sigma^4,$$

woraus insgesamt durch Einsetzen in (2.16) die Behauptung folgt. □

Sei nun $X = (X_1,...,X_n)$, $n \ge 2$, ein Beobachtungsvektor mit unter $\mathbb{P}_\theta$ unabhängigen, jeweils *Normal*(μ,σ^2)-verteilten Komponenten bei unbekanntem Parameter $\theta = (\mu,\sigma^2) \in \mathbb{R}\times(0,\infty)$. Wir setzen

$$\widehat{\sigma}^2_{n,c}(x) := \frac{1}{c}\sum_{k=1}^{n}(x_k-\overline{x}_n)^2$$

für $c > 0$ und erinnern daran, dass $\widehat{\sigma}^2_{n,n-1}(x)$ den GBES und $\widehat{\sigma}^2_{n,n}(x)$ den MLS für die Varianz σ^2 bildet [☞ 2.56 bzw. 2.7].

Satz 2.61. *Unter den soeben getroffenen Voraussetzungen und bei quadratischer Verlustfunktion $L(\theta,d) = (d-\sigma^2)^2$ gilt für das Risiko $R_n(c) := R(\theta,\widehat{\sigma}^2_{n,c})$ des Schätzers $\widehat{\sigma}^2_{n,c}$:*

$$R_n(c) \;=\; \sigma^4\left((n-1)(n+1)\left(\frac{1}{c}-\frac{1}{n+1}\right)^2+\frac{2}{n+1}\right). \tag{2.17}$$

Es ist (für jedes θ) minimal für $c = n+1$, wobei insbesondere

$$\left(1+\frac{2}{n-1}\right)\frac{2\sigma^4}{n+1} \;=\; R_n(n-1) \;>\; R_n(n) \;>\; R_n(n+1) \;=\; \frac{2\sigma^4}{n+1}. \tag{2.18}$$

Weder der GBES $\widehat{\sigma}^2_{n,n-1}$ noch der MLS $\widehat{\sigma}^2_{n,n}$ sind somit unter der obigen Verlustfunktion zulässig.

Beweis. Unter Benutzung der Transformationseigenschaften der Normalverteilung folgt sofort

$$R_n(c) \;=\; R((\mu,\sigma^2),\widehat{\sigma}^2_{n,c}) \;=\; R((0,\sigma^2),\widehat{\sigma}^2_{n,c}).$$

Ferner gilt unter Benutzung von $\mathbb{E}_{(\mu,\sigma^2)}\widehat{\sigma}^2_{n,c}(X) = (n-1)\sigma^2/c$ [☞ **??**]

$$\begin{aligned} R((0,\sigma^2),\widehat{\sigma}^2_{n,c}) \;&=\; \mathbb{E}_{(0,\sigma^2)}(\widehat{\sigma}^2_{n,c}(X)-\sigma^2)^2 \\ &=\; \mathbb{E}_{(0,\sigma^2)}\widehat{\sigma}^4_{n,c}(X) - 2(n-1)c^{-1}\sigma^4+\sigma^4 \end{aligned}$$

für jedes $c > 0$. Wenden wir schließlich auf

$$\mathbb{E}_{(0,\sigma^2)}\widehat{\sigma}^4_{n,c}(X) \;=\; \frac{1}{c^2}\,\mathbb{E}_{(0,\sigma^2)}\left(\sum_{k=1}^n (X_k-\overline{X}_n)^2\right)^2$$

das zuvor bewiesene Lemma an und beachten, dass $\mathbb{E}_{(0,\sigma^2)}X_1^4 = 3\sigma^4$, so ergibt sich nach einiger Rechnung (2.17) und daraus weiter (2.18). Wir verzichten auf weitere Details. □

Anmerkung 2.62. Betrachtet man in der Situation von Lemma 2.60 den für σ^2 erwartungstreuen [☞ **??**] Schätzer $\widehat{\sigma}^2_{n,n-1}(x) = \frac{1}{n-1}\sum_{k=1}^n (x_k-\overline{x}_n)^2$, sofern $n \geq 2$, so ergibt sich für diesen bei quadratischem Verlust das Risiko

$$\mathbb{V}\mathrm{ar}\left(\frac{1}{n-1}\sum_{k=1}^n (X_k-\overline{X}_n)^2\right) \;=\; \frac{\mu_4}{n}-\frac{(n-2)\sigma^4}{n(n-1)},$$

wie der Leser mittels (2.15) leicht nachprüft.

2.7 Die Informations-Ungleichung

Im Folgenden werden wir zunächst den Begriff der *Fisher-Information* einführen und dann mit dessen Hilfe eine untere Schranke für die Varianz (für jedes θ) erwartungstreuer Schätzer herleiten. Auf der Suche nach einem GBES bietet dieses Resultat in einigen Situationen eine Alternative zu dem im vorherigen Abschnitt vorgestellten Verfahren. Sofern wir nämlich für einen erwartungstreuen Schätzer nachweisen können, dass seine Varianz die untere Schranke (für jedes θ) annimmt, ist dieser natürlich ein GBES. Andererseits werden wir zeigen, dass diese Situationen im wesentlichen jene sind, in denen die zugrundeliegende Verteilungsfamilie eine einparametrige Exponentialfamilie bildet [☞ Satz 2.75]. Unsere Darstellung orientiert sich weitestgehend an den Darstellungen in [5, Abschnitt 4.3] und [10, Abschnit 2.6].

Gegeben sei im Folgenden ein dominiertes statistisches Experiment $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ mit dominierendem Maß ν, dessen Verteilungsfamilie $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ folgende Zusatzvoraussetzungen erfüllt:

(A1) Der Parameterraum Θ ist eine offene Teilmenge von $\mathbb{R}$.

(A2) Es existieren Versionen $f(\theta,\cdot) = dW_\theta/d\nu,\ \theta\in\Theta$, so dass

$$\mathfrak{X}^* \ := \ \{x\in\mathfrak{X} : f_\theta(x) > 0\}$$

nicht von θ abhängt, was insbesondere die paarweise Äquivalenz der W_θ impliziert.

(A3) Für alle $\theta\in\Theta$ und $x\in\mathfrak{X}^*$ existiert $\frac{\partial}{\partial\theta}f(\theta,x)$ und ist stetig in θ.

(A4) Für jede Statistik $S : (\mathfrak{X},\mathscr{A}) \to (\mathbb{R},\mathscr{B}(\mathbb{R}))$ mit $\mathbb{E}_\theta|S(X)| < \infty$ für alle $\theta\in\Theta$ gilt

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta S(X) \ = \ \frac{\partial}{\partial\theta}\int_{\mathfrak{X}^*} S(x)f(\theta,x)\ \nu(dx) \ = \ \int_{\mathfrak{X}^*} S(x)\frac{\partial}{\partial\theta}f(\theta,x)\ \nu(dx),$$

sofern das letzte Integral existiert und endlich ist.

Zur Abkürzung nennen wir $\mathscr{W}$ ebenso wie $\mathscr{E}$ unter diesen Voraussetzungen *regulär*. Wir weisen darauf hin, dass aus der Messbarkeit von $f(\theta,x)$ und $\frac{\partial}{\partial\theta}f(\theta,x)$ in x bei festem θ und der Stetigkeit in θ bei festem $x\in\mathfrak{X}^*$ (Bedingung (A3)) ferner folgt:

(A5) $f(\theta,x)$ und $\frac{\partial}{\partial\theta}f(\theta,x)$ sind als Funktionen von $(\Theta\times\mathfrak{X},\mathscr{B}(\Theta)\otimes\mathscr{A})$ nach $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ produkt-messbar.

Diese wohlbekannte Tatsache wird an einigen Stellen ohne weitere Erwähnung benutzt. Für einen Beweis mag der Leser z.B. [13, Kapitel IV, Theorem T47] konsultieren. Da Bedingung (A4) für praktische Zwecke offenkundig wertlos ist, geben wir zunächst folgendes

Lemma 2.63. *Gegeben sei ein statistisches Experiment $\mathscr{E}$, das die Bedingungen (A1–3) erfüllt. Für jede Statistik $S : (\mathfrak{X}, \mathscr{A}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ mit $\mathbb{E}_\theta |S(X)| < \infty$ für alle $\theta \in \Theta$ seien außerdem die Integrale*

$$\int_{\mathfrak{X}^*} S(x) \frac{\partial}{\partial \theta} f(\theta, x)\ \nu(dx) \quad \textit{und} \quad \int_{\mathfrak{X}^*} \left| S(x) \frac{\partial}{\partial \theta} f(\theta, x) \right|\ \nu(dx)$$

stetig in θ. Dann gilt Bedingung (A4).

Beweis. Da Θ offen ist, genügt es, die Behauptung auf jedem Intervall (θ_0, θ_1) mit $I = [\theta_0, \theta_1] \subset \Theta$ zu zeigen. Nach Voraussetzung ist $\vartheta \mapsto \int_{\mathfrak{X}^*} |S(x) \frac{\partial}{\partial \theta} f(\vartheta, x)|\, \nu(dx)$ auf I stetig und somit beschränkt. Der Satz von Fubini sichert daher

$$\begin{aligned} \int_{\theta_0}^{\theta} \int_{\mathfrak{X}^*} S(x) \frac{\partial}{\partial \vartheta} f(\vartheta, x)\ \nu(dx)\, d\vartheta &= \int_{\mathfrak{X}^*} S(x) \int_{\theta_0}^{\theta} \frac{\partial}{\partial \vartheta} f(\vartheta, x) d\vartheta\ \nu(dx) \\ &= \int_{\mathfrak{X}^*} S(x)(f(\theta, x) - f(\theta_0, x))\ \nu(dx) \\ &= \mathbb{E}_\theta S(X) - \mathbb{E}_{\theta_0} S(X). \end{aligned}$$

Differenziert man nun den ersten und den letzten Ausdruck nach θ, so ergibt sich offenkundig die in (A4) geforderte Identität. □

Als einfache Konsequenz erhalten wir:

Satz 2.64. *Eine 1-parametrige Exponentialfamilie $\mathscr{W} = (W_\theta)_{\theta \in \Theta}$ mit offenem Parameterraum Θ, dominierendem Maß ν und ν-Dichten der Form*

$$f(\theta, x) = C(\theta) \exp(Q(\theta) T(x)) h(x)$$

ist regulär, falls $Q(\Theta)$ offen und $Q(\theta)$ auf ganz Θ stetig differenzierbar ist.

Wie der Leser sofort einsieht, liefert der Satz insbesondere die Regularität, wenn $\mathscr{W}$ in natürlicher Parametrisierung gegeben ist und der natürliche Parameterraum $\mathfrak{Z}$ ein nichtleeres Inneres besitzt, welches Θ enthält.

Beweis. Wir müssen offenkundig nur Bedingung (A4) nachprüfen. Sei S eine beliebige, o.B.d.A. nichtnegative Statistik mit $\mathbb{E}_\theta S(X) < \infty$ für alle $\theta \in \Theta$. Wir definieren die W-Maße $\widetilde{W}_\zeta$ durch

$$\widetilde{W}_\zeta(B) := \frac{1}{F(\zeta)} \int_B S(x) e^{\zeta T(x)} h(x)\ \nu(dx), \quad B \in \mathscr{A},$$

für $\zeta \in \widetilde{\mathfrak{Z}} := \{\xi \in \mathbb{R} : \int S(x) e^{\xi T(x)} h(x) \nu(dx) < \infty\}$, wobei

$$F(\zeta) \;:=\; \int S(x)e^{\xi T(x)}h(x)\;\nu(dx).$$

$(\widetilde{W}_\zeta)_{\zeta\in\widetilde{\Im}}$ bildet daher ebenfalls eine 1-parametrige Exponentialfamilie mit natürlichem Parameterraum $\widetilde{\Im}$ und $Q(\Theta)$, weil nach Voraussetzung offen, eine Teilmenge des Inneren von $\widetilde{\Im}$. Mittels Satz 1.27 erhalten wir, dass sowohl $F(\zeta)$ als auch $G(\zeta) := \int e^{\zeta T(x)}h(x)\,\nu(dx) > 0$ auf $Q(\Theta)$ unendlich oft differenzierbar ist. Beachte, dass $C(\theta) = G\circ Q(\theta)^{-1}$. Es folgt die Behauptung wegen

$$\begin{aligned}\int S(x)\frac{\partial}{\partial\theta}f(\theta,x)\nu(dx) \;&=\; \int \frac{\partial}{\partial\theta}\left(\frac{S(x)\exp(Q(\theta)T(x))h(x)}{G\circ Q(\theta)}\right)\;\nu(dx)\\ &=\; \frac{\partial}{\partial\theta}\left(\frac{F\circ Q(\theta)}{G\circ Q(\theta)}\right)\end{aligned}$$

und der stetigen Differenzierbarkeit von $Q(\theta)$. □

Der Leser kann leicht nachprüfen, dass jede der folgenden 1-parametrigen Exponentialfamilien die Bedingungen in Satz 2.64 erfüllt und somit regulär ist ($n \geq 1$ beliebig):

$(Bin(m,\theta)^n)_{\theta\in(0,1)}$ und $(NBin(m,\theta)^n)_{\theta\in(0,1)}$ für festes $m\in\mathbb{N}$,
$(Poisson(\theta)^n)_{\theta>0}$,
$(Normal(\mu,\sigma^2)^n)_{\mu\in\mathbb{R}}$ für festes $\sigma^2>0$,
$(Normal(\mu,\sigma^2)^n)_{\sigma^2>0}$ für festes $\mu\in\mathbb{R}$,
$(\Gamma(a,b)^n)_{a>0}$ für festes $b>0$ und $(\Gamma(a,b)^n)_{b>0}$ für festes $a>0$,
$(\beta(a,b)^n)_{a>0}$ für festes $b>0$ und $(\beta(a,b)^n)_{b>0}$ für festes $a>0$.

Definition 2.65. Sei $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ eine Verteilungsfamilie, die den Voraussetzungen (A1–3) genügt, wobei wie üblich $W_\theta = \mathbb{P}_\theta^X$ für eine Zufallsvariable X unterstellt wird. Dann heißt

$$I(\theta) \;:=\; \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\theta,x)\right)^2 \;=\; \int_{\mathfrak{X}^*}\left(\frac{\frac{\partial}{\partial\theta}f(\theta,x)}{f(\theta,x)}\right)^2 f(\theta,x)\;\nu(dx) \qquad (2.19)$$

die *Fisher-Information von $\mathscr{W}$ in θ*. Es gilt $0\leq I(\theta)\leq\infty$.

Der Quotient $\frac{\partial}{\partial\theta}f(\theta,x)/f(\theta,x)$ gibt offenbar die relative Rate an, mit der sich die Dichte $f(\theta,x)$ im Punkt x ändert. Aus diesem Grund erscheint folgende Heuristik erlaubt: Je größer der Wert von $I(\cdot)$ an einer Stelle θ_0 ist, desto leichter lässt sich der Parameter θ_0 von Nachbarwerten θ unterscheiden und desto genauer ist eine Schätzung von θ, wenn θ_0 den wahren Parameter bezeichnet.

Da die W_θ in 2.65 paarweise äquivalent sind, können wir als dominierendes Maß auch W_{θ_0} für irgendein $\theta_0\in\Theta$ wählen, wobei unter Hinweis auf Korollar 13.5 in [2]

$$g(\theta,x) \;:=\; \frac{f(\theta,x)}{f(\theta_0,x)}\,\mathbf{1}_{\mathfrak{X}^*}(x)$$

eine W_{θ_0}-Dichte von W_θ für jedes θ bildet. Wegen $\frac{\partial}{\partial\theta}g(\theta,x) = f(\theta_0,x)^{-1}\frac{\partial}{\partial\theta}f(\theta,x)$ für alle $(\theta,x)\in\Theta\times\mathfrak{X}^*$ erhalten wir

$$\begin{aligned}\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log g(\theta,X)\right)^2 \;&=\; \int_{\mathfrak{X}^*}\left(\frac{\frac{\partial}{\partial\theta}g(\theta,x)}{g(\theta,x)}\right)^2 g(\theta,x)\,W_{\theta_0}(dx)\\ &=\; \int_{\mathfrak{X}^*}\left(\frac{\frac{\partial}{\partial\theta}f(\theta,x)}{f(\theta,x)}\right)^2 f(\theta,x)\,\nu(dx)\\ &=\; \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\theta,x)\right)^2\end{aligned}$$

und folglich die Unabhängigkeit der Fisher-Information vom speziell gewählten dominierenden Maß und der damit resultierenden Dichten. Dagegen ist es wichtig festzustellen, dass die Fisher-Information sehr wohl von der jeweils gewählten Parametrisierung abhängt: Gilt $\theta = h(\xi)$ für eine streng monotone, differenzierbare Funktion $h:\Xi\to\Theta$, so ergibt sich unter der neuen Parametrisierung für die Fisher-Information $I^*(\xi)$ von $\mathscr{W}$ in ξ

$$I^*(\xi) \;=\; \mathbb{E}_{h(\xi)}\left(\frac{\partial}{\partial\xi}\log f(h(\xi),X)\right)^2 \;=\; I(h(\xi))h'(\xi)^2 \tag{2.20}$$

für alle $\xi\in\Xi$. Sofern mit verschiedenen Parametrisierungen einer Verteilungsfamilie gearbeitet wird, ist die Bezeichnung $I(\theta)$ demnach inadäquat. Für viele Zwecke reicht sie allerdings aus, insbesondere erweist sich die untere Schranke der angestrebten Informationsungleichung als von der gewählten Parametrisierung unabhängig [☞ im Anschluss an Satz 2.72].

Zwei alternative Darstellungen der Fisher-Information gibt das nächste

Lemma 2.66. *Gegeben eine reguläre Verteilungsfamilie $(W_\theta)_{\theta\in\Theta}$, gilt für alle θ*

$$\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\theta,X)\right) \;=\; 0 \tag{2.21}$$

und

$$I(\theta) \;=\; \mathbb{V}\mathrm{ar}_\theta\left(\frac{\partial}{\partial\theta}\log f(\theta,X)\right). \tag{2.22}$$

Falls außerdem $\frac{\partial^2}{\partial\theta^2}f(\theta,x)$ für alle $(\theta,x)\in\Theta\times\mathfrak{X}^*$ existiert und die Beziehung

$$\int\frac{\partial^2}{\partial\theta^2}f(\theta,x)\nu(dx) \;=\; \frac{\partial^2}{\partial\theta^2}\int f(\theta,x)\nu(dx)\quad(=0)$$

erfüllt ist, folgt weiter

$$I(\theta) \;=\; -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(\theta,X)\right) \tag{2.23}$$

für alle $\theta\in\Theta$.

Beweis. Unter Benutzung der Voraussetzung (A4) ergibt sich bei Differentiation der Beziehung $\int f(\theta,x)\nu(dx)=1$ nach θ

$$\begin{aligned}
0 \;&=\; \int_{\mathfrak{X}^*}\frac{\partial}{\partial\theta}f(\theta,x)\;\nu(dx)\\
&=\; \int_{\mathfrak{X}^*}\frac{\frac{\partial}{\partial\theta}f(\theta,x)}{f(\theta,x)}f(\theta,x)\;\nu(dx\\
&=\; \int_{\mathfrak{X}^*}\left(\frac{\partial}{\partial\theta}\log f(\theta,x)\right)f(\theta,x)\;\nu(dx)\\
&=\; \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\theta,X)\right),
\end{aligned}$$

d.h. (2.21). Behauptung (2.22) ist damit offensichtlich. Zum Nachweis von (2.23) notieren wir

$$\frac{\partial^2}{\partial\theta^2}\log f(\theta,x) \;=\; \frac{\frac{\partial^2}{\partial\theta^2}f(\theta,x)}{f(\theta,x)}-\left(\frac{\frac{\partial}{\partial\theta}f(\theta,x)}{f(\theta,x)}\right)^2,$$

woraus das Resultat per Integration nach $W_\theta(dx)=f(\theta,x)\,\nu(dx)$ unter Beachtung der Zusatzvoraussetzung folgt. □

Beispiel 2.67. Sei $m\geq 1$ fest gewählt. Falls $W_\theta = Bin(m,\theta)$ für $\theta\in(0,1)$ mit Zähldichte

$$f(\theta,x) \;=\; \binom{m}{x}\theta^x(1-\theta)^{m-x} \;=\; \binom{m}{x}\exp\left(\log\left(\frac{\theta}{1-\theta}\right)x+m\log(1-\theta)\right)$$

für $x\in\mathfrak{X}=\mathfrak{X}^*=\{0,\ldots,m\}$, so gilt

$$\frac{\partial}{\partial\theta}\log f(\theta,x) \;=\; \frac{x}{\theta(1-\theta)}-\frac{m}{1-\theta}$$

und folglich

$$I(\theta) \;=\; \mathbb{V}\mathrm{ar}_\theta\left(\frac{X}{\theta(1-\theta)}+\frac{m}{1-\theta}\right) \;=\; \frac{\mathbb{V}\mathrm{ar}_\theta X}{\theta^2(1-\theta)^2} \;=\; \frac{m}{\theta(1-\theta)}$$

für alle $\theta\in(0,1)$.

Beispiel 2.68. Falls $W_\theta = \mathit{Poisson}(\theta)$ für $\theta > 0$ mit Zähldichte

$$f(\theta,x) \;=\; e^{-\theta}\frac{\theta^x}{x!} \;=\; \frac{1}{x!}\exp(x\log\theta-\theta)$$

für $x \in \mathfrak{X} = \mathfrak{X}^* = \mathbb{N}_0$, so gilt

$$\frac{\partial}{\partial\theta}\log f(\theta,x) \;=\; \frac{x}{\theta}-1$$

und folglich

$$I(\theta) \;=\; \frac{\mathbb{V}\mathrm{ar}_\theta X}{\theta^2} \;=\; \frac{1}{\theta}$$

für alle $\theta > 0$.

Beispiel 2.69. Sei $\sigma^2 > 0$ fest gewählt. Falls $W_\mu = \mathit{Normal}(\mu,\sigma^2)$ für $\mu \in \mathbb{R}$ mit $\mathfrak{B}$-Dichte

$$f(\mu,x) \;=\; \frac{1}{(2\pi)^{1/2}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \;=\; \frac{1}{(2\pi)^{1/2}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}+\frac{\mu x}{\sigma^2}-\frac{\mu^2}{2\sigma^2}\right)$$

für $x \in \mathfrak{X} = \mathfrak{X}^* = \mathbb{R}$, so gilt

$$\frac{\partial}{\partial\mu}\log f(\mu,x) \;=\; \frac{x-\mu}{\sigma^2}$$

und folglich

$$I(\mu) \;=\; \frac{\mathbb{V}\mathrm{ar}_\mu X}{\sigma^4} \;=\; \frac{1}{\sigma^2}$$

für alle $\mu \in \mathbb{R}$.

Bevor wir uns unserem Hauptanliegen, einer unteren Schranke für die Varianz erwartungstreuer Schätzer, zuwenden, geben wir noch eine wichtige Eigenschaft der Fisher-Information, nämlich ihre *Additivität unter der Bildung von Produktverteilungen.*

Satz 2.70. *Gegeben zwei reguläre Verteilungsfamilien $\mathscr{W}_j = (W_{j,\theta})_{\theta\in\Theta},\ j = 1,2$, sei $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ die zugehörige Produktfamilie, d.h. $W_\theta = W_{1,\theta}\otimes W_{2,\theta}$ für $\theta \in \Theta$. Dann ist auch $\mathscr{W}$ regulär, und es gilt für alle $\theta \in \Theta$*

$$I(\theta) \;=\; I_1(\theta)+I_2(\theta) \tag{2.24}$$

für die Fisher-Information $I_1(\cdot), I_2(\cdot), I(\cdot)$ von $\mathscr{W}_1, \mathscr{W}_2$ bzw. $\mathscr{W}$.

Beweis. Dass $\mathscr{W}$ mit $\mathscr{W}_1$ und $\mathscr{W}_2$ ebenfalls die Bedingungen (A1–4) erfüllt, sieht man sofort ein. Kommen wir deshalb gleich zum Nachweis von (2.24). Sei $X =$

(Y,Z) mit stochastisch unabhängigen Zufallsvariablen Y und Z, wobei $\mathbb{P}_\theta^Y = W_{1,\theta}$ und $\mathbb{P}_\theta^Z = W_{2,\theta}$. Für geeignete σ-endliche Maße ν_1, ν_2 bezeichne ferner $f_1(\theta,\cdot)$ die ν_1-Dichte von $W_{1,\theta}$ und $f_2(\theta,\cdot)$ die ν_2-Dichte von $W_{2,\theta}$, also $f(\theta,\cdot) = f_1 \otimes f_2(\theta,\cdot)$ die $\nu_1 \otimes \nu_2$-Dichte von W_θ. Wir erinnern daran, dass die Fisher-Information nicht vom speziell gewählten dominierenden Maß abhängt. Es folgt nun unter Hinweis auf (2.22)

$$\begin{aligned} I(\theta) &= \mathbb{V}\mathrm{ar}_\theta\left(\frac{\partial}{\partial\theta}\log f(\theta,X)\right) \\ &= \mathbb{V}\mathrm{ar}_\theta\left(\frac{\partial}{\partial\theta}\log(f_1(\theta,Y)f_2(\theta,Z))\right) \\ &= \mathbb{V}\mathrm{ar}_\theta\left(\frac{\partial}{\partial\theta}\log f_1(\theta,Y)+\frac{\partial}{\partial\theta}\log f_2(\theta,Z)\right) \\ &= \mathbb{V}\mathrm{ar}_\theta\left(\frac{\partial}{\partial\theta}\log f_1(\theta,Y)\right) + \mathbb{V}\mathrm{ar}_\theta\left(\frac{\partial}{\partial\theta}\log f_2(\theta,Z)\right) \\ &= I_1(\theta) + I_2(\theta), \end{aligned}$$

wobei für die vorletzte Zeile die Unabhängigkeit von Y und Z benutzt wurde. □

Für die Standardsituation, in der $X = (X_1,\ldots,X_n)$ einen Zufallsvektor mit unabhängigen, identisch verteilten Komponenten bildet, liefert Satz 2.70 zusammen mit einer einfachen Induktion:

Korollar 2.71. *Sei $n \geq 1$ und $\mathscr{V} = (V_\theta)_{\theta\in\Theta}$ eine Verteilungsfamilie, welche die Bedingungen (A1–3) erfüllt. Dann genügt auch die zugehörige Familie $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ der n-fachen Produktverteilungen (d.h. $W_\theta = V_\theta^n$) diesen Bedingungen, und es gilt für alle $\theta \in \Theta$*

$$I(\theta) = nJ(\theta),$$

wobei $I(\cdot)$ und $J(\cdot)$ die Fisher-Information von $\mathscr{W}$ bzw. $\mathscr{V}$ bezeichnen.

Wir kommen nun zu der angekündigten Ungleichung für die Varianz erwartungstreuer Schätzer, die in der Literatur unter dem Namen *Informations-Ungleichung* oder auch *Cramér-Rao-Ungleichung* bekannt ist:

Satz 2.72. (Informations-Ungleichung) *Sei $\mathscr{E} = (\mathfrak{X},\mathscr{A},(W_\theta)_{\theta\in\theta})$ ein reguläres statistisches Experiment, $T : (\mathfrak{X},\mathscr{A}) \to (\mathbb{R},\mathscr{B}(\mathbb{R}))$ eine Statistik mit $\mathbb{V}\mathrm{ar}_\theta T(X) < \infty$ für alle $\theta \in \Theta$ und $\psi(\theta) := \mathbb{E}_\theta T(X)$. Ist die Fisher-Information $I(\theta)$ für alle $\theta \in \Theta$ positiv und endlich, so ist ψ differenzierbar und*

$$\mathbb{V}\mathrm{ar}_\theta T(X) \geq \frac{\psi'(\theta)^2}{I(\theta)} \tag{2.25}$$

für alle $\theta \in \Theta$.

Wie bereits erwähnt, ist die Schranke in (2.25) invariant unter Parametertransformationen der Form $\theta = h(\xi)$ mit streng monotoner, differenzierbarer Funktion $h : \Xi \to \Theta$. Da zum einen $\psi^*(\xi) := \mathbb{E}_{h(\xi)} T(X)$ die Ableitung $\psi^{*\prime}(\xi) = \psi'(h(\xi))h'(\xi)$ besitzt und zum anderen $I^*(\xi) = I(h(\xi))h'(\xi)^2$ für die Fisher-Information $I^*(\xi)$ von $\mathscr{W}$ in ξ [☞ (2.20)] gilt, folgt nämlich in (2.25)

$$\frac{\psi'(\theta)^2}{I(\theta)} = \frac{\psi'(h(\xi))^2}{I(h(\xi))} = \frac{\psi^{*\prime}(\xi)^2}{I^*(\xi)}.$$

Beweis. In den üblichen Bezeichnungen folgt unter Benutzung von Bedingung 4. die Differenzierbarkeit von $\psi(\theta)$ sowie

$$\begin{aligned}
\psi'(\theta) &= \int_{\mathcal{X}^*} T(x) \frac{\partial}{\partial \theta} f(\theta, x) \nu(dx) \\
&= \int_{\mathcal{X}^*} T(x) \left(\frac{\frac{\partial}{\partial \theta} f(\theta, x)}{f(\theta, x)} \right) f(\theta, x) \nu(dx) \\
&= \mathbb{E}_\theta \left(T(X) \frac{\partial}{\partial \theta} \log f(\theta, X) \right)
\end{aligned}$$

für alle $\theta \in \Theta$. Da ferner $\mathbb{E}_\theta \frac{\partial}{\partial \theta} \log f(\theta, x) = 0$ gemäß (2.21) in Lemma 2.66, erhalten wir weiter

$$\psi'(\theta) = \mathrm{Cov}_\theta \left(T(X), \frac{\partial}{\partial \theta} \log f(\theta, X) \right)$$

und daraus vermöge der Cauchy-Schwarz-Ungleichung

$$|\psi'(\theta)| \leq \left(\mathbb{V}\mathrm{ar}_\theta T(X) \mathbb{V}\mathrm{ar}_\theta \left(\frac{\partial}{\partial \theta} \log f(\theta, X) \right) \right)^{1/2}$$

für alle $\theta \in \Theta$. Eine einfache Umformung dieser Ungleichung liefert schließlich (2.25), wenn man noch $I(\theta) = \mathbb{V}\mathrm{ar}_\theta (\frac{\partial}{\partial \theta} \log f(\theta, X)$ gemäß (2.21) in Lemma 2.66 beachtet. □

Die untere Schranke der Informations-Ungleichung hängt von der Statistik T noch durch $\psi'(\theta)$ ab. Betrachtet man speziell erwartungstreue Schätzer für $\psi(\theta) = \theta$, gelangt man zu einer universellen Schranke:

Korollar 2.73. *Gegeben die Situation von Satz 2.72, sei T ein erwartungstreuer Schätzer für θ. Dann gilt*

$$\mathbb{V}\mathrm{ar}_\theta T(X) \geq \frac{1}{I(\theta)}$$

für alle $\theta \in \Theta$, wobei $1/I(\theta)$ wird als ***untere Cramér-Rao-Schranke*** *(in θ) bezeichnet wird.*

Den wichtigen Spezialfall, in dem die Beobachtung X aus n unabhängigen, identisch verteilten Zufallsvariablen $X_1, ..., X_n$ besteht, behandelt

Korollar 2.74. *In der Situation von Satz 2.72 sei außerdem $W_\theta = V_\theta^n$ für ein $n \geq 1$ und alle $\theta \in \Theta$. Die Familie $\mathscr{V} = (V_\theta)_{\theta\in\Theta}$ sei regulär mit zugehöriger Fisher-Information $J(\cdot)$. Dann gilt*

$$\mathbb{V}\mathrm{ar}_\theta T(X) \;\geq\; \frac{\psi'(\theta)^2}{nJ(\theta)}$$

für alle $\theta \in \Theta$.

Beweis. Es genügt der Hinweis auf Korollar 2.71, nach dem $I(\theta) = nJ(\theta)$ gilt. □

In jedem der Spezialfälle

$$\begin{aligned}\mathscr{W} &= (Bin(m,\theta)^n)_{\theta\in(0,1)},\ m \geq 1 \text{ fest},\\ \mathscr{W} &= (Poisson(\theta)^n)_{\theta>0},\\ \mathscr{W} &= (Normal(\theta,\sigma^2)^n)_{\theta\in\mathbb{R}},\ \sigma^2 > 0 \text{ fest}\end{aligned}$$

mit Fisher-Informationen $I(\theta) = \frac{mn}{\theta(1-\theta)}, = \frac{n}{\theta}$ bzw. $= \frac{n}{\sigma^2}$ (☞ 2.67, 2.68 und 2.69) bildet $T(x) = \bar{x}_n$ einen erwartungstreuen Schätzer für $\psi(\theta) = \mathbb{E}_\theta X_1$, d.h. $\psi(\theta) = m\theta$ im ersten Fall und $\psi(\theta) = \theta$ in den anderen beiden Fällen. Wie man sofort nachprüft, nimmt dieser Schätzer jeweils für jedes θ die Schranke der Informationsungleichung (2.25) an, d.h.

$$\mathbb{V}\mathrm{ar}_\theta \overline{X}_n \;=\; \frac{\psi'(\theta)^2}{I(\theta)} < \infty$$

für alle $\theta \in \Theta$, und bildet folglich einen GBES. Dies hatten wir auf alternative Weise bereits in 2.54 eingesehen, dort im allgemeinen Kontext 1-parametriger Exponentialfamilien in T. Dass auch hier die Tatsache, dass jede der obigen Familien eine 1-parametrige Exponentialfamilie bildet, kein Zufall ist, belegt unser letztes Resultat dieses Abschnitts:

Satz 2.75. *Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, \mathscr{W})$, $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$, ein statistisches Experiment mit offenem Parameterraum Θ.*

(a) Bildet $\mathscr{W}$ eine 1-parametrige Exponentialfamilie in $Q(\theta)$ und $T(x)$, die den Bedingungen in Satz 2.64 genügt ($Q(\Theta)$ offen und Q auf ganz Θ stetig differenzierbar), so nimmt $T(x)$ in jedem θ die Schranke der Informationsunleichung an und ist folglich ein GBES für $\psi(\theta) = \mathbb{E}_\theta T(X)$.

(b) Ist umgekehrt $\mathscr{E}$ regulär und existiert eine Statistik $T : (\mathfrak{X}, \mathscr{A}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, welche für jedes θ endliche, positive Varianz besitzt und die Schranke der Informationsungleichung annimmt, also

$$0 \;<\; \mathbb{V}\mathrm{ar}_\theta T(X) \;=\; \frac{\psi'(\theta)^2}{I(\theta)} \;<\; \infty, \quad \psi(\theta) \;:=\; \mathbb{E}_\theta T(X),$$

für alle $\theta \in \Theta$, so existiert eine auf ganz Θ stetig differenzierbare Funktion $Q : \Theta \to \mathbb{R}$, so dass $\mathscr{W}$ eine 1-parametrige Exponentialfamiie in Q und T bildet.

Dieses Resultat wurde übrigens erst 1973 von WIJSMAN [18] bewiesen. Etwas später zeigte JOSHI [9], dass bei Abschwächung der Regularitätsannahmen (A1–4) die Schranke der Informationsungleichung auch dann angenommen werden kann, wenn die zugrundeliegende Verteilungsfamilie keine 1-parametrige Exponentialfamilie bildet.

Zum Beweis des Satzes benötigen wir die folgende, leicht zu beweisende Ergänzung zur Cauchy-Schwarz-Ungleichung:

Für zwei quadratisch integrierbare ZG X, Y auf einem W-Raum $(\Omega, \mathfrak{A}, \mathbb{P})$ gilt genau dann

$$|\mathrm{Cov}\,(X,Y)| = (\mathbb{V}\mathrm{ar} X)^{1/2} (\mathbb{V}\mathrm{ar} Y)^{1/2},$$

wenn $Y = aX + b$ $\mathbb{P}$-f.s. für geeignete $a, b \in \mathbb{R}$.

Beweis. Wie der Beweis der Informationsgleichung gezeigt hat, gilt dort genau dann die Gleichheit in einem $\theta \in \Theta$, wenn

$$\left| \mathrm{Cov}_\theta \left(T(X), \frac{\partial}{\partial \theta} \log f(\theta, X) \right) \right| \;=\; (\mathbb{V}\mathrm{ar}_\theta T(X))^{1/2} \left(\mathbb{V}\mathrm{ar}_\theta \left(\frac{\partial}{\partial \theta} \log f(\theta, x) \right) \right)^{1/2},$$

also nach der obigen Ergänzung zur Cauchy-Schwarz-Ungleichung, wenn

$$\frac{\partial}{\partial \theta} \log f(\theta, x) \;=\; a(\theta) T(x) + b(\theta) \quad W_\theta\text{-f.s.} \tag{2.26}$$

für geeignete $a(\theta), b(\theta) \in \mathbb{R}$ gilt.

(a) Wir notieren als erstes, dass $\mathbb{V}\mathrm{ar}_\theta T(X) < \infty$ für alle $\theta \in \Theta$ durch Satz 1.27 gesichert ist [☞ speziell (1.8) mit $\varphi \equiv 1$]. Aus $f(\theta, x) = C(\theta) \exp(Q(\theta) T(x)) h(x)$ für alle $\theta \in \Theta$ und $x \in \mathfrak{X}$, wobei o.B.d.A. $h \equiv 1$ gewählt werden kann, folgt

$$\log f(\theta,x) \;=\; Q(\theta)T(x)+\log C(\theta)$$

und dann (2.26) mit $a(\theta)=Q'(\theta)$ und $b(\theta)=\frac{C'(\theta)}{C(\theta)}$. Die Differenzierbarkeit von $Q(\theta)$ gilt hierbei nach Voraussetzung und impliziert mit Satz 1.27 die von $C(\theta)=(\int e^{Q(\theta)T(x)}\,\nu(dx))^{-1}$.

(b) Kommen wir nun zu dem interessanteren, aber auch schwierigeren Umkehrschluss: Der Ausgangspunkt ist diesmal die Gültigkeit von (2.26) für alle $\theta\in\Theta$, wobei das Problem darin besteht, dass die W_θ-Nullmenge

$$\left\{x:\frac{\partial}{\partial\theta}\log f(\theta,x)\neq a(\theta)T(x)+b(\theta)\right\}$$

i.A. von θ abhängt. Wir werden deshalb als nächstes zeigen, dass

$$\mathfrak{X}^{**} \;:=\; \left\{x\in\mathfrak{X}^*:\frac{\partial}{\partial\theta}\log f(\theta,x)=a(\theta)T(x)+b(\theta)\text{ für alle }\theta\in\Theta\right\}$$

Wahrscheinlichkeit 1 unter jedem W_θ besitzt und dass $a(\theta),b(\theta)$ stetige Funktionen sind. Es folgt dann die Behauptung, weil offenbar für beliebiges $\theta_0\in\Theta$

$$f(\theta,x) \;=\; \exp\left(\left(\int_{\theta_0}^{\theta}a(\vartheta)d\vartheta\right)T(x)+\int_{\theta_0}^{\theta}b(\vartheta)d\vartheta\right)f(\theta_0,x)$$

für alle $\theta\in\Theta$ und $x\in\mathfrak{X}^{**}$ gilt und $Q(\theta):=\int_{\theta_0}^{\theta}a(\vartheta)d\vartheta$ stetig differenzierbar ist.

Wegen $\mathbb{V}\mathrm{ar}_\theta T(X)>0$ für alle $\theta\in\Theta$ existieren $x,y\in\mathfrak{X}^*$ mit $T(x)\neq T(y)$. $a(\theta)$ und $b(\theta)$ ergeben sich dann als Lösungen des linearen Gleichungssystems

$$\begin{aligned}\frac{\partial}{\partial\theta}\log f(\theta,x) \;&=\; a(\theta)T(x)+b(\theta),\\ \frac{\partial}{\partial\theta}\log f(\theta,y) \;&=\; a(\theta)T(y)+b(\theta).\end{aligned}$$

Sie sind stetig, weil $\frac{\partial}{\partial\theta}\log f(\theta,x)$ für jedes $x\in\mathfrak{X}^*$ stetig in θ ist (Voraussetzung (A2)). Wegen der paarweisen Äquivalenz der W_θ liefert (2.26)

$$W_\theta\left(\left\{x\in\mathfrak{X}^*:\frac{\partial}{\partial\theta}\log f(\vartheta,x)=a(\vartheta)T(x)+b(\vartheta)\right\}\right) \;=\; 1$$

für alle $\theta,\vartheta\in\Theta$. Sei nun Θ^* eine beliebige abzählbare, dichte Teilmenge von Θ. Dann folgt zum einen

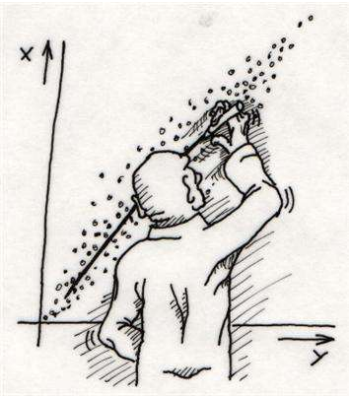
$$W_\theta\left(\left\{x\in\mathfrak{X}^*:\frac{\partial}{\partial\theta}\log f(\vartheta,x)=a(\vartheta)T(x)+b(\vartheta)\text{ für alle }\vartheta\in\Theta^*\right\}\right) \;=\; 1$$

für alle $\theta\in\Theta$ und zum anderen aus der Stetigkeit der Funktionen $a(\theta)$, $b(\theta)$ und $\frac{\partial}{\partial\theta}\log f(\theta,x)$ für $x\in\mathfrak{X}^*$ offensichtlich

$$\mathfrak{X}^{**} = \left\{ x \in \mathfrak{X}^* : \frac{\partial}{\partial \vartheta} \log f(\vartheta, x) = a(\vartheta) T(x) + b(\vartheta) \text{ für alle } \vartheta \in \Theta^* \right\},$$

was insbesondere die Messbarkeit von $\mathfrak{X}^{**}$ beweist. □

2.8 Lineare Modelle und die Methode der kleinsten Quadrate



Cartoon von G. Meixner ©1998
Quelle: VIAS Science Cartoons

In diesem Abschnitt werden wir uns mit einer Klasse von statistischen Modellen beschäftigen, die aufgrund ihrer mannigfaltigen Anwendungen von großer praktischer Relevanz sind und deren Untersuchung Inhalt der sogenannten *Regressionsanalyse* bildet. Die dort zum Tragen kommende *Methode der kleinsten Quadrate* geht zurück auf CARL FRIEDRICH GAUSS [7], der sie für Schätzungen bei astronomischen Messungen benutzte.

2.8.1 Homoskedastische Modelle

Wir beginnen mit der Beschreibung des einfachsten Modelltyps, dem sogenannten *linearen Regressionsmodell*. Dabei geht es um die Bestimmung einer unbekannten linearen Beziehung zwischen einer interessierenden Größe x und einer Kontrollvariablen k, die jedoch aufgrund zufälliger Schwankungen, verursacht durch Messfehler oder weitere nicht kontrollierbare Einflussgrößen, nur "verrauscht" feststellbar ist und den Einsatz statistischer Methoden notwendig macht. Als Beispiel denke man an die Untersuchung der Aufnahmemenge x einer Düngechemikalie in Nutzpflanzen in Abhängigkeit von der eingesetzten Menge k des Düngemittels. Letztere ist hier ein *quantativer Faktor*, der – mit gewissen Einschränkungen – vom Versuchsleiter frei variiert werden kann. Durch Wahl verschiedener sinnvoller Werte von k (Düngemittelmengen) für gewisse Pflanzen und Messung der jeweiligen Aufnahmemenge x soll die lineare Beziehung zwischen x und k, die sich durch

$$x = \theta_1 + \theta_2 k + \text{"zufällige Schwankung"}$$

ausdrücken lässt, empirisch geschätzt werden, und zwar durch die unbekannten Parameter θ_1 und θ_2. Der Versuchsleiter führt dazu n Experimente unter Variation der Düngemittelmenge k durch und gelangt so zur Beobachung von n Zufallsgrößen

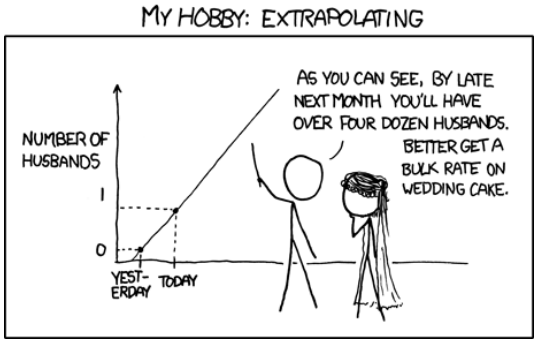


Abb. 2.3 Quelle: XKCD (A Webcomic of Romance, Sarcasm, Math, and Language).

$$\begin{aligned} X_1 &= \theta_1 + \theta_2 k_1 + \varepsilon_1, \\ &\vdots \\ X_n &= \theta_1 + \theta_2 k_n + \varepsilon_n, \end{aligned} \tag{2.27}$$

wobei wir annehmen, dass $\varepsilon_1, ..., \varepsilon_n$ unabhängig sind mit $\mathbb{E}\varepsilon_j = 0$ und $\mathbb{V}\mathrm{ar}\,\varepsilon_j = \sigma^2 \in (0,\infty)$ für $j = 1, ..., n$. Es bezeichne Q_j die Verteilung von ε_j und Q_j^a die Verteilung von $a + \varepsilon_j$, d.h. $Q_j^a(B) = Q_j(B-a)$ für alle $B \in \mathscr{B}(\mathbb{R})$. Die Voraussetzung gleicher Varianzen für die sogenannten *Fehler* $\varepsilon_1, ..., \varepsilon_n$ ist wesentlich für die im Anschluss entwickelte Theorie; man nennt ein Modell in diesem Fall *homoskedastisch*[4] und sonst *heteroskedastisch* [☞ auch Definition 2.77].

Das zugrundeliegende statistische Experiment können wir nun wie folgt beschreiben: Als unbekannten Parameter betrachten wir

$$\vartheta := (\theta_1, \theta_2, Q_1, ..., Q_2),$$

wobei $\theta_1, \theta_2 \in \mathbb{R}$ und $Q_1, ..., Q_n$ W-Maße auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ bilden mit jeweiligem Erwartungswert 0 und Varianz σ^2. Es gilt also $\Theta = \widetilde{\Theta} \times \mathscr{Q}$ mit $\widetilde{\Theta} \subset \mathbb{R}^2$ und

$$\begin{aligned} \mathscr{Q} := \big\{ &(Q_1, ..., Q_n) : Q_j \text{ W-Maß auf } (\mathbb{R}, \mathscr{B}(\mathbb{R})) \text{ mit } \textstyle\int x\, Q_j(x) = 0 \\ &\text{f.a. } 1 \le j \le n \text{ und } 0 < \textstyle\int x^2\, Q_1(dx) = ... = \int x^2\, Q_n(dx) < \infty \big\}. \end{aligned} \tag{2.28}$$

Als Stichprobenraum liegt hier natürlich $(\mathfrak{X}, \mathscr{A}) = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ vor, und die unbekannten Verteilungen W_ϑ haben die Form

$$W_\vartheta = \mathbb{P}_\vartheta^X = Q_1^{\theta_1 + \theta_2 k_1} \otimes ... \otimes Q_n^{\theta_1 + \theta_2 k_n}.$$

Niemand gibt jedoch lineare Modelle in dieser Form an, sondern man beschränkt sich auf Angaben wie in (2.27). Zu schätzen sind, wie schon gesagt, die reellen Parameterfunktionen $\gamma_1(\vartheta) = \theta_1$ und $\gamma_2(\vartheta) = \theta_2$. Man beachte jedoch, dass hier

[4] Abgeleitet vom griechischen Substantiv $\sigma\kappa\varepsilon\delta\alpha\sigma\iota\varsigma$, das "Zerstreuung" bedeutet.

wegen der zusätzlich auftretenden nichtparametrischen Komponente $(Q_1,...,Q_n)$ in ϑ kein parametrisches Modell im bisherigen Sinne vorliegt, sondern ein sogenanntes *semiparametrisches Modell*.

Die *Methode der kleinsten Quadrate* [5] besagt nun, dass wir in Abhängigkeit von dem beobachteten $x = (x_1,...,x_n)$ die Schätzer $\widehat{\theta}_1(x)$ und $\widehat{\theta}_2(x)$ für θ_1 und θ_2 so wählen sollen, dass gilt

$$\sum_{j=1}^{n}(x_j - \widehat{\theta}_1(x) - \widehat{\theta}_2(x)k_j)^2 \;=\; \min_{(\theta_1,\theta_2)\in\widetilde{\Theta}} \sum_{j=1}^{n}(x_j - \theta_1 - \theta_2 k_j)^2.$$

Differenzieren wir die rechtsstehende, zu minimierende Funktion partiell nach θ_1 sowie θ_2 und setzen die Ableitungen gleich 0, so ergeben sich die *Normalengleichungen*

$$\frac{\partial}{\partial\theta_i}\sum_{j=1}^{n}(x_j - \theta_1 - \theta_2 k_j)^2 \;=\; 0 \quad (i=1,2),$$

also

$$\sum_{j=1}^{n}(x_j - \theta_1 - \theta_2 k_j) \;=\; 0 \quad \text{und} \quad \sum_{j=1}^{n} k_j(x_j - \theta_1 - \theta_2 k_j) \;=\; 0. \tag{2.29}$$

Sofern nicht alle k_j identisch sind und die folgenden Lösungen in $\widetilde{\Theta}$ liegen, erhalten wir

$$\widehat{\theta}_1(x) \;=\; \bar{x}_n - \widehat{\theta}_2(x)\bar{k}_n \quad \text{und} \quad \widehat{\theta}_2(x) \;=\; \frac{\sum_{j=1}^{n}(k_j-\bar{k}_n)(x_j-\bar{x}_n)}{\sum_{j=1}^{n}(k_j-\bar{k}_n)^2}, \tag{2.30}$$

wobei natürlich wie bisher $\bar{x}_n = n^{-1}\sum_{j=1}^{n} x_j$ und $\bar{k}_n = n^{-1}\sum_{j=1}^{n} k_j$ gilt.

Beispiel 2.76. Neun Bodenstücke (Parzellen) wurden mit verschiedenen Mengen k_j eines Phosphordüngers behandelt und anschließend die Aufnahmemengen x_j von Phosphor in Kornpflanzen gemessen, die nach 38 Tagen auf den jeweiligen Bodenstücken gewachsen waren. Die Ergebnisse zeigt die umseitige Tabelle, gefolgt von einer Graphik, die sowohl die Messwerte (k_j,x_j) als auch die geschätzte Regressionsgerade enthält.

---

[5] Wie schon zu Beginn dieses Abschnitts erwähnt, geht die Methode auf CARL FRIEDRICH GAUSS zurück, dem es mit ihrer Hilfe gelang, die *elliptische* Bahn des Zwergplaneten Ceres aus den Beobachtungen des ital. Astronomen GUISEPPE PIAZZI sehr genau zu berechnen. Piazzi hatte den Planeten am Neujahrstag 1801 entdeckt und daraufhin 40 Tage lang beobachtet, bevor er hinter der Sonne verschwand. Sein Ansehen litt anschließend deutlich, weil seine Beobachtungen nicht zu einer unter Experten erwarteten kreisförmigen Bahn passen wollten. Erst als FRANZ XAVER VON ZACH und HEINRICH WILHELM OLBERS im Dezember 1801 den Kleinplaneten genau an dem von Gauß vorhergesagten Ort wiederfanden, war Piazzas Ruf wiederhergestellt. Quelle: [19]

k_j	1	4	5	9	11	13	23	23	28
x_j	64	71	54	81	76	93	77	95	109

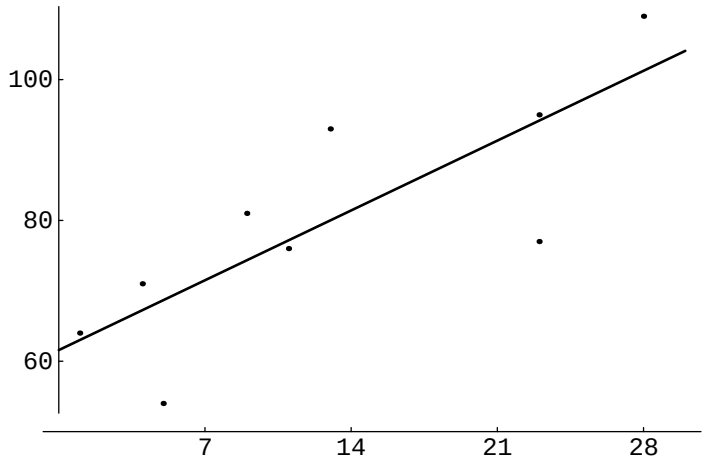


Abb. 2.4 Die gemäß obiger Tabelle gemessenen Aufnahmenmengen x_j an Phosphor in Kornpflanzen als Funktion der eingesetzten Menge k_j mit resultierender Regressionsgeraden.

Wir kommen nun zur allgemeinen Definition des linearen statistischen Modells sowie des Kleinste-Quadrate-Schätzers, wobei wir anstelle einer Darstellung wie in (2.27) besser auf Matrizen und Vektoren zurückgreifen. Ein Vektor $x \in \mathbb{R}^n$ wird im Folgenden stets als *Spalten*vektor interpretiert und dessen durch Transposition $^\top$ resultierenden Zeilenvektor mit $x^\top$ bezeichnet.

Definition 2.77. Ein *lineares (statistisches) Modell* liegt vor, wenn der zugehörige Beobachtungsvektor $X = (X_1, ..., X_n)^\top$ mit Werten im $\mathbb{R}^n$ die *Regressionsgleichung*

$$X = A\theta + \varepsilon, \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{2.31}$$

erfüllt mit unbekanntem θ aus einem linearen Unterraum $\widetilde{\Theta}$ des $\mathbb{R}^p$, bekannter $n \times p$-Matrix $A = (a_{ij})$, die *Design-Matrix* genannt wird, sowie unabhängigen ZG $\varepsilon_1, ..., \varepsilon_n$, genannt *Fehler* oder *Residuen*, deren Verteilungen $Q_1, ..., Q_n$ jeweils mit Erwartungswert 0 und Varianzen $\sigma_1^2, ..., \sigma_n^2$, ebenfalls unbekannt sind. Das Modell heißt *homoskedastisch*, falls $\sigma_1^2 = ... = \sigma_n^2$, und sonst *heteroskedastisch*.

Eine Funktion $\widehat{\theta} = (\widehat{\theta}_1, ..., \widehat{\theta}_p)^\top : \mathbb{R}^n \to \mathbb{R}^p$ heißt dann *Kleinste-Quadrate-Schätzer (KQS) für* θ (engl. *Least Square Estimator (LSE)*), wenn

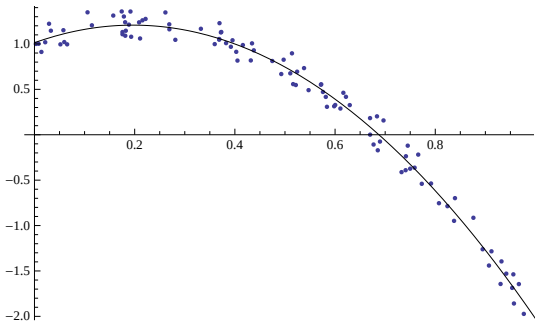


Abb. 2.5 Beispiel einer quadratischen Regression

$$(x-A\widehat{\theta}(x))^\top(x-A\widehat{\theta}(x)) \;=\; \min_{\theta\in\widetilde{\Theta}}(x-A\theta)^\top(x-A\theta)$$

für alle $x\in\mathbb{R}^n$ gilt.

Bis auf weiteres unterstellen wir ein homoskedastisches Modell, ohne dies immer wieder explizit zu erwähnen. Die damit identische Fehlervarianz bezeichnen wir mit σ^2.

Das zuvor betrachtete *lineare Regressionsmodell* bildet offensichtlich ein spezielles lineares Modell mit $p=2$ und

$$A \;=\; \begin{pmatrix} 1 & k_1 \\ \vdots & \vdots \\ 1 & k_n \end{pmatrix}.$$

Dasselbe gilt auch für dessen Verallgemeinerung, dem *polynomialen Regressionsmodell*, in dem die Daten x_j durch ein Polynom $\sum_{i=0}^{r}\theta_{i+1}k_j^i$ vom Grad r approximiert werden, wobei in diesem Fall offensichtlich $X=A\theta+\varepsilon$ gilt mit

$$A \;=\; \begin{pmatrix} 1 & k_1 & k_1^2 & \dots & k_1^r \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & k_n & k_n^2 & \dots & k_n^r \end{pmatrix}.$$

Abb. 2.5 zeigt ein Beispiel einer *quadratischen Regression* $(r=2)$.

Wie bereits für das lineare Regressionsmodell eingeführt, setzen wir im Folgenden $\vartheta=(\theta^\top,Q_1,\dots,Q_n)$, wobei $\Theta=\widetilde{\Theta}\times\mathcal{Q}$ mit $\mathcal{Q}$ gemäß (2.28). Ferner bezeichne $\mathfrak{L}_A$ das Bild von $\widetilde{\Theta}$ unter A, d.h.

$$\mathfrak{L}_A \;=\; \{A\theta : \theta\in\widetilde{\Theta}\}.$$

Da $\widetilde{\Theta}$ als Unterraum des $\mathbb{R}^p$ vorausgesetzt wird, bildet $\mathfrak{L}_A$ einen Unterraum des $\mathbb{R}^n$, und $\widehat{\theta}(x)$ definiert genau dann einen KQS für θ, falls es die Beziehung

$$\|x - A\widehat{\theta}(x)\| \;=\; \min_{y\in\mathfrak{L}} \|x-y\|, \quad \|x\| \;:=\; \left(\sum_{j=1}^{n} x_j^2\right)^{1/2},$$

erfüllt, d.h., falls $A\widehat{\theta}(x)$ der *Projektion von x auf* $\mathfrak{L}_A$, definiert durch $\mathbf{P}_{\mathfrak{L}_A}(x)$, entspricht. Bekanntlich gilt

$$A\widehat{\theta}(x) \;=\; \mathbf{P}_{\mathfrak{L}_A}(x) \quad\Leftrightarrow\quad x - A\widehat{\theta}(x) \;\perp\; \mathfrak{L}_A. \tag{2.32}$$

Nehmen wir an, dass $\widetilde{\Theta} = \mathbb{R}^p$ gilt, also $\mathfrak{L}_A = \text{Bild}(A)$, und bezeichnet e_j den j-ten Einheitsvektor des $\mathbb{R}^p$, so folgt aus (2.32) weiter

$$\begin{aligned} A\widehat{\theta}(x) \;=\; \mathbf{P}_{\mathfrak{L}_A}(x) \quad &\Leftrightarrow \quad (Ae_j)^\top (x - A\widehat{\theta}(x)) \;=\; 0 \text{ für } j = 1,\ldots,p \\ &\Leftrightarrow \quad e_j^\top A^\top (x - A\widehat{\theta}(x)) \;=\; 0 \text{ für } j = 1,\ldots,p \\ &\Leftrightarrow \quad A^\top x - A^\top A\widehat{\theta}(x) \;=\; 0 \\ &\Leftrightarrow \quad A^\top A\widehat{\theta}(x) \;=\; A^\top x. \end{aligned}$$

Die letzte Gleichung heißt *Normalengleichung* und entspricht im einfchen Regressionsmodell den gleichgenannten Gleichungen (2.29). Wir hatten für die Herleitung vorausgesetzt, dass $\widetilde{\Theta} = \mathbb{R}^p$ gilt, was wir im Folgenden stets tun wollen.

Definition 2.78. Ein lineares Modell der Form $X = A\theta + \varepsilon$ besitzt *vollen Rang*, falls der Rang von A gleich p ist und somit $\dim(\mathfrak{L}_A) = p$ gilt.

In einem linearen Modell vollen Rangs sind also die Spalten von A linear unabhängig, was insbesondere $p \leq n$ (A injektiv) impliziert. Der KQS kann dann sehr einfach angegeben werden.

Satz 2.79. *In einem linearen Modell $X = A\theta + \varepsilon$ vollgen Rangs gilt: Der KQS $\widehat{\theta} : \mathbb{R}^n \to \mathbb{R}^p$ für θ wird durch*

$$\widehat{\theta}(x) \;=\; (A^\top A)^{-1} A^\top x$$

gegeben.

Beweis. Wie soeben gesehen, erfüllt $\widehat{\theta}$ die Normalengleichung

$$A^\top A\widehat{\theta}(x) \;=\; A^\top x$$

für alle $\theta \in \mathbb{R}^n$. Da A den Rang p besitzt, gilt dasselbe für die $p \times p$-Matrix $A^\top A$[6], d.h. $A^\top$ ist als $p \times p$-Matrix invertierbar. Die Behauptung ergibt sich folglich durch Auflösen der Normalengleichung nach $\widehat{\theta}(x)$. □

Wenden wir uns als nächstes dem Problem zu, in einem linearen Modell reelle Parameterfunktion4n $\gamma(\vartheta)$ erwartungstreu zu schätzen, wobei die Schätzfunktionen außerdem linear seien. Dazu:

Definition 2.80. In einem linearen Schätzmodell für die reelle Parameterfunktion $\gamma(\vartheta)$ mit Stichprobenraum $\mathfrak{X} = \mathbb{R}^n$ wird jeder Schätzer der Form

$$g_b : \mathbb{R}^n \to \mathbb{R}, \quad g_b(x) \;=\; b^\top x \quad \text{für ein festes } b \in \mathbb{R}^n$$

als *linearer Schätzer* bezeichnet. Außerdem heißt g^* *gleichmäßig bester linearer erwartungstreuer Schätzer (GBLES)* (engl. *BLUE* für *"best linear unbiased estimator"*), falls g^* erwartungstreu und linear ist sowie varianzminimierend unter allen weiteren erwartungstreuen linearen Schätzern, d.h.

$$\mathbb{V}\mathrm{ar}_\vartheta\, g^*(X) \;=\; \min_{b \in \mathbb{R}^n : \mathbb{E}_\vartheta g_b(X) = \gamma(\vartheta)} \mathbb{V}\mathrm{ar}_\vartheta\, g_b(X)$$

für alle $\vartheta \in \Theta$.

Der nächste Satz zeigt, dass sich für *lineare Parameterfunktionen*

$$\gamma(\vartheta) \;=\; \beta^\top \theta, \quad \beta \in \mathbb{R}^p,$$

der GBLES sofort mit Hilfe der KQS $\widehat{\theta}$ für θ berechnen lässt.

Satz 2.81. (Satz von Gauß-Markov) *In einem linearen Modell $X = A\theta + \varepsilon$ vollen Rangs sei $\widehat{\theta}(x) = (A^\top A)^{-1} A^\top x$ der KQS für θ. Dann ist, zu gegebenem $\beta \in \mathbb{R}^p$, $\beta^\top \widehat{\theta}$ der GBLES für $\gamma(\vartheta) = \beta^\top \theta$ mit Risiko*

$$R(\vartheta, \beta^\top \widehat{\theta}) \;=\; \mathbb{V}\mathrm{ar}_\vartheta\, \beta^\to \widehat{\theta}(X) \;=\; \sigma^2 \beta^\top (A^\top A)^{-1} \beta \tag{2.33}$$

für alle $\vartheta \in \Theta$.

Wir notieren zunächst, dass für einen Zufallsvektor $X = (X_1, ..., X_n)^\top$ mit Werten im $\mathbb{R}^n$

$$\mathbb{E}X \;:=\; \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{pmatrix} \quad \text{und} \quad \mathrm{Cov}(X) \;:=\; (\mathrm{Cov}(X_i, X_j))_{1 \le i,j \le n}$$

[6] denn: $a^\top A x = 0 \;\Rightarrow\; 0 = x^\top A^\top A x = (Ax)^\top Ax = \|Ax\|^2 \;\Rightarrow\; Ax = 0 \;\Rightarrow\; x = 0$, da A injektiv ist.

dessen Erwartungswertvektor und Kovarianzmatrix bezeichnen. Für jede $m \times n$-Matrix B gilt dann [☞ (30.15) in [2]]

$$\mathbb{E}(BX) \;=\; B\,\mathbb{E}X \quad \text{und} \quad \mathrm{Cov}(BX) \;=\; B\,\mathrm{Cov}(X)B^\top.$$

Beweis. Unter Beachtung der vorstehenden Regeln sowie von $\mathbb{E}_\vartheta X = \mathbb{E}_\vartheta(A\theta+\varepsilon) = A\theta$ erhalten wir nun für alle $\vartheta \in \Theta$

$$\begin{aligned}\mathbb{E}_\vartheta\,\widehat{\theta}(X) \;&=\; \mathbb{E}_\vartheta(A^\top A)^{-1}A^\top X \;=\; (A^\top A)^{-1}A^\top \mathbb{E}X \\ &=\; (A^\top A)^{-1}A^\top(A\theta) \;=\; (A^\top A)^{-1}(A^\top A)\theta \;=\; \theta\end{aligned}$$

und damit für jedes $\beta \in \mathbb{R}^p$

$$\mathbb{E}_\vartheta\beta^\top\widehat{\theta}(X) \;=\; \beta^\top\mathbb{E}_\vartheta\widehat{\theta}(X) \;=\; \beta^\top\theta,$$

d.h. die Erwartungstreue von $\beta^\top\widehat{\theta}(x)$.

Sei nun g ein weiterer linearer erwartungstreuer Schätzer für $\beta^\top\theta$, also $g(x) = b^\top x$ für ein $b \in \mathbb{R}^n$ und $\mathbb{E}_\vartheta b^\top X = \beta^\top\theta$ für alle $\vartheta \in \Theta$. Aus

$$\beta^\top\theta \;=\; \mathbb{E}_\vartheta b^\top X \;=\; b^\top\mathbb{E}_\vartheta X \;=\; b^\top A\theta$$

für alle $\vartheta \in \Theta$ folgt weiter aufgrund des vollen Rangs $b^\top A = \beta^\top$. Mit Hilfe der Zerlegung

$$b^\top X \;=\; (b^\top - \beta^\top(A^\top A)^{-1}A^\top)X + \beta^\top(A^\top A)^{-1}A^\top X$$

erhalten wir als nächstes

$$\begin{aligned}\mathbb{V}\mathrm{ar}_\vartheta\, b^\top X \;=\; &\underbrace{\mathbb{V}\mathrm{ar}_\vartheta(b^\top - \beta^\top(A^\top A)^{-1}A^\top)X}_{\geq 0} + \underbrace{\mathbb{V}\mathrm{ar}_\vartheta\,\beta^\top(A^\top A)^{-1}A^\top X}_{=\mathbb{V}\mathrm{ar}_\vartheta\,\beta^\top\widehat{\theta}(X)} \\ &+\; 2\,\mathrm{Cov}_\vartheta((b^\top - \beta^\top(A^\top A)^{-1}A^\top)X, \beta^\top(A^\top A)^{-1}A^\top X),\end{aligned}$$

so dass es für den Nachweis der Optimalität von $\beta^\top\widehat{\theta}$ offenkundig reicht, dass die auftretende Kovarianz verschwindet. Zu diesem Zweck bemerken wir als erstes, dass aufgrund der Unabhängigkeit der $\varepsilon_1, \ldots, \varepsilon_n$

$$\begin{aligned}\mathrm{Cov}_\vartheta(c^\top X, d^\top X) \;&=\; \mathrm{Cov}_\vartheta\left(\sum_{j=1}^n c_j X_j, \sum_{j=1}^n d_j X_j\right) \\ &=\; \mathbb{E}_\vartheta\left(\left(\sum_{j=1}^n c_j(X_j - \mathbb{E}_\vartheta X_j)\right)\left(\sum_{j=1}^n d_j(X_j - \mathbb{E}_\vartheta X_j)\right)\right) \\ &=\; \mathbb{E}_\vartheta\left(\sum_{j=1}^n c_j d_j (X_j - \mathbb{E}_\vartheta X_j)^2\right)\end{aligned}$$

$$
\begin{aligned}
&= \sigma^2 \sum_{j=1}^{n} c_j d_j \\
&= \sigma^2 c^\top d
\end{aligned}
$$

für $c, d \in \mathbb{R}^n$. Damit ergibt sich nun $(b^\top A = \beta^\top)$

$$
\begin{aligned}
&\mathrm{Cov}_\vartheta((b^\top - \beta^\top (A^\top A)^{-1} A^\top)X, \beta^\top (A^\top A)^{-1} A^\top X) \\
&\quad = \sigma^2 (b^\top - \beta^\top (A^\top A)^{-1} A^\top)(\beta^\top (A^\top A)^{-1} A^\top)^\top \\
&\quad = \sigma^2 (b^\top - \beta^\top (A^\top A)^{-1} A^\top) A((A^\top A)^{-1})^\top \beta \\
&\quad = \sigma^2 (b^\top A((A^\top A)^{-1})^\top \beta - \beta^\top (A^\top A)^{-1}(A^\top A)((A^\top A)^{-1})^\top \beta) \\
&\quad = \sigma^2 (\beta^\top ((A^\top A)^{-1})^\top \beta - \beta^\top ((A^\top A)^{-1})^\top \beta) \;=\; 0
\end{aligned}
$$

und weiter für das Risiko von $\beta^\top \widehat{\theta}$ unter Benutzung von $\mathrm{Cov}_\vartheta(X) = \sigma^2 I_n$, I_n die n-dimensionale Einheitsmatrix,

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}_\vartheta \beta^\top \widehat{\theta}(X) &= \mathrm{Cov}_\vartheta(\beta^\top (A^\top A)^{-1} A^\top X, \beta^\top (A^\top A)^{-1} A^\top X) \\
&= (\beta^\top (A^\top A)^{-1} A^\top) \mathrm{Cov}_\vartheta(X) (\beta^\top (A^\top A)^{-1} A^\top)^\top \\
&= \sigma^2 \beta^\top (A^\top A)^{-1} A((A^\top A)^{-1})^\top \beta = \sigma^2 \beta^\top ((A^\top A)^{-1})^\top \beta \\
&= \sigma^2 (\beta^\top (A^\top A)^{-1} \beta)^\top = \sigma^2 \beta^\top (A^\top A)^{-1} \beta.
\end{aligned}
$$

Damit ist der Satz vollständig bewiesen. □

Wenden wir uns als nächstes einem wichtigen Spezialfall zu und betrachten ein *lineares Modell mit normalverteilten Fehlern* (und vollem Rang). Wir nehmen also an, dass $\varepsilon_1, ..., \varepsilon_n$ jeweils $Normal(0, \sigma^2)$-verteilt sind, also $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^\top$ eine (n-dimensionale) $Normal_n(0, \sigma^2 I_n)$-Verteilung besitzt. Damit folgt

$$W_\vartheta = \mathbb{P}_\vartheta^X = \mathbf{N}_n(A\theta, \sigma^2 I_n),$$

wobei wir jetzt als Parameter

$$\vartheta = (\theta^\top, \sigma^2) \in \Theta = \mathbb{R}^p \times (0, \infty)$$

betrachten und folglich wieder ein parametrisches Modell im üblichen Sinne erhalten. Für die λ^n-Dichten von W_ϑ gilt

$$
\begin{aligned}
\frac{dW_\vartheta}{d\lambda^n}(x) &= \frac{d\,Normal_n(A\theta, \sigma^2 I_n)}{d\lambda^n}(x) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(x - A\theta)^\top (x - A\theta)\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(x_i - \sum_{j=1}^{p} a_{ij}\theta_j\right)^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} \left(x_i^2 - 2x_i \sum_{j=1}^{n} a_{ij}\theta_j + \left(\sum_{j=1}^{p} a_{ij}\theta_j \right)^2 \right) \right) \right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} a_{ij}\theta_j \right)^2 \right) \\
&\qquad\qquad \times \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \sum_{j=1}^{p} \frac{\theta_j}{\sigma^2} \sum_{i=1}^{n} a_{ij} x_i \right) \\
&= C(\theta) \exp\left(\sum_{j=1}^{p+1} Q_j(\vartheta) T_j(x) \right)
\end{aligned}
$$

$$
\text{mit} \quad \big(Q_1(\vartheta), \dots, Q_p(\vartheta), Q_{p+1}(\vartheta)\big) := \left(\frac{\theta_1}{\sigma^2}, \dots, \frac{\theta_p}{\sigma^2}, -\frac{1}{2\sigma^2} \right),
$$

$$
\big(T_1(x), \dots, T_p(x), T_{p+1}(x)\big) := \left(\sum_{i=1}^{n} a_{i1} x_i, \dots, \sum_{i=1}^{n} a_{ip} x_i, \sum_{i=1}^{n} x_i^2 \right).
$$

Die Statistik $T = (T_1, \dots, T_{p+1})^\top$ ist also nach dem Neyman-Kriterium [bzw. Korollar 2.31] suffizient und gemäß Satz 2.46 auch vollständig, denn die Menge

$$
\{(Q_1(\vartheta), \dots, Q_{p+1}(\vartheta)) : \vartheta \in \Theta\} = \mathbb{R}^p \times (-\infty, 0)
$$

hat innere Punkte. In vektorieller Schreibweise gilt

$$
(T_1(x), \dots, T_p(x))^\top = A^\top x, \quad \text{also} \quad T(x) = (A^\top x, x^\top x)^\top.
$$

Nach diesen Vorüberlegungen können wir nun leicht die folgende Verschärfung des Satzes von Gauß-Markov zeigen:

Satz 2.82. *In einem linearen Modell $X = A\theta + \varepsilon$ vollen Rangs, normalverteilten Fehlern $\varepsilon_1, \dots, \varepsilon_n$ und KQS $\widehat{\theta}(x) = (A^\top A)^{-1} A^\top x$ für θ gilt: Für jedes $\beta \in \mathbb{R}^p$ ist $\beta^\top \widehat{\theta}$ der GBES für $\gamma(\vartheta) = \beta^\top \vartheta$ (statt nur GBLES).*

Beweis. Für den KQS $\widehat{\theta}(x) = (A^\top A)^{-1} A^\top x$ für θ gilt offensichtlich $\widehat{\theta} = g \circ T^\top$, wobei T wie oben und $g : \mathbb{R}^{p+1} \to \mathbb{R}^p$ durch

$$
g(y_1, \dots, y_{p+1}) = (A^\top A)^{-1} \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}
$$

definiert ist. Damit folgt $\beta^\top \widehat{\theta} = \beta^\top (g \circ T^\top)$ für alle $\beta \in \mathbb{R}^p$ und daraus weiter die Behauptung des Satzes per Kombination des Satzes von Gauß-Markov mit dem von Lemann-Scheffé, denn T ist vollständig und suffizient. □

Um eine Vorstellung von der *Güte einer Regressionsschätzung* zu bekommen, ist es notwendig, auch die unbekannte Varianz der Fehler zu schätzen, d.h. σ^2. Dies zeigt sich insbesondere daran, dass das Risiko eines GBLES gemäß (2.33) linear von σ^2 abhängt. Kehren wir noch einmal zu Beispiel 2.76 zurück, so erkennen wir bereits an der dortigen Abb. 2.4, dass hierfür die *Summe der Fehlerquadrate (SFQ)*

$$\mathrm{SFQ} : \mathbb{R}^n \to [0,\infty), \quad x \mapsto \sum_{j=1}^{n} (x_j - \widehat{\theta}_1(x) - \widehat{\theta}_2(x) k_j)^2 \tag{2.34}$$

herangezogen werden sollte. Dasselbe gilt für jedes allgemeine lineare Modell $X = A\theta + \varepsilon$, wobei dann analog zu (2.34)

$$\mathrm{SFQ}(x) \;=\; \sum_{j=1}^{n} (x_j - (A\widehat{\theta}(x))_j)^2 \;=\; (x - A\widehat{\theta}(x))^\top (x - A\widehat{\theta}(x))$$

gesetzt wird. Wir zeigen nun:

Satz 2.83. *In einem linearen Modell $X = A\theta + \varepsilon$ vollen Rangs mit $n > p$ und KQS $\widehat{\theta}(x) = (A^\top A)^{-1} A^\top x$ für θ definiert $(n-p)^{-1}$ SFQ(x) einen erwartungstreuen Schätzer für σ^2, der im Fall normalverteilter Fehler sogar der GBES ist.*

Beweis. Wir beginnen mit einer weiteren Umformung von SFQ(x):

$$\begin{aligned}
\mathrm{SFQ}(x) \;&=\; (x - A(A^\top A)^{-1}A^\top x)^\top (x - A(A^\top A)^{-1}A^\top x) \\
&=\; x^\top x - x^\top (A(A^\top A)^{-1}A^\top x \;-\; (A(A^\top A)^{-1}A^\top x)^\top x \\
&+\; (A(A^\top A)^{-1}A^\top x)^\top A(A^\top A)^{-1}A^\top x \\
&=\; x^\top x - x^\top A(A^\top A)^{-1}A^\top x \;-\; \underbrace{x^\top A(A^\top A)^{-1}A^\top x}_{\text{denn: } ((A^\top A)^{-1})^\top = ((A^\top A)^\top)^{-1} = (A^\top A)^{-1}} \\
&+\; x^\top A(A^\top A)^{-1}A^\top A(A^\top A)^{-1}A^\top x \\
&=\; x^\top x - x^\top A(A^\top A)^{-1} x \\
&=\; x^\top (I_n - A(A^\top A)^{-1}A^\top) x.
\end{aligned}$$

Wir haben damit

$$\mathrm{SFQ}(X) = X^\top (I_n - A(A^\top A)^{-1}A^\top) X$$

erhalten, und wie man leicht nachrechnet, definiert $C := A(A^\top A)^{-1}A^\top$ eine symmetrische und idempotente ($C^2 = C$) $n \times n$-Matrix vom Rang p[7]. Als nächstes beachte, dass für die Matrix $B = I_n - C$ gilt

[7] *Beweis:* Da A injektiv und $A^\top A$ bijektiv ist, folgt $A^\top x = 0$ direkt aus $Cx = 0$ für jedes $x \in \mathbb{R}^n$ und damit $\mathrm{Kern}(C) = \mathrm{Kern}(A^\top)$. Dies impliziert aber $\mathrm{Rang}(C) = n - \dim \mathrm{Kern}(C) = n - \dim \mathrm{Kern}(A^\top) = \mathrm{Rang}(A^\top) = \mathrm{Rang}(A) = p$.

$$\begin{aligned}\mathbb{E}_\vartheta X^\top BX &= \mathbb{E}_\vartheta \mathrm{Spur}(BXX^\top) = \mathrm{Spur}(\mathbb{E}_\vartheta BXX^\top) = \mathrm{Spur}(B\mathbb{E}_\vartheta XX^\top)\\ &= \mathrm{Spur}\big(B(\mathrm{Cov}_\vartheta(X)+(\mathbb{E}_\vartheta X)(\mathbb{E}_\vartheta X)^\top)\big)\\ &= \mathrm{Spur}(B\mathrm{Cov}_\vartheta(X))+\mathrm{Spur}(B(\mathbb{E}_\vartheta X)(\mathbb{E}_\vartheta X)^\top)\\ &= \mathrm{Spur}(\mathrm{Cov}_\vartheta(X)+(\mathbb{E}_\vartheta X)^\top B(\mathbb{E}_\vartheta X).\end{aligned}$$

Damit folgt nun unter Beachtung von $\mathbb{E}_\vartheta X = A\theta$ und $\mathrm{Cov}_\vartheta(X) = \sigma^2 I_n$

$$\begin{aligned}\mathbb{E}_\vartheta \mathrm{SFQ}(X) &= \mathbb{E}_\vartheta X^\top(I_n-C)X = \mathrm{Spur}((I_n-C)\sigma^2 I_n)+(A\theta)^\top(I_n-C)A\theta\\ &= \sigma^2(\mathrm{Spur}(I_n)-\mathrm{Spur}(C))+\theta^\top A^\top \underbrace{(A-A(A^\top A)^{-1}A^\top A)}_{=0}\theta\\ &= \sigma^2(n-\mathrm{Rang}(C)) = \sigma^2(n-p),\end{aligned}$$

wobei für die vorletzte Gleichheit benutzt wurde, dass Spur und Rang einer symmetrischen idempotenten Matrix übereinstimmen[8]. Unter der Voraussetzung $p<n$ bildet also $(n-p)^{-1}$ $\mathrm{SFQ}(x)$ einen erwartungstreuen Schätzer für σ^2, der im Fall normalverteilter Fehler sogar GBES ist, weil $\mathrm{SFQ}(x) = x^\top x-(A^\top x)^\top(A^\top A)^{-1}A^\top x$ von x dann nur über die vollständige und suffiziente Statistik $T(x) = (A^\top x, x^\top x)$ abhängt. [☞ Überlegungen vor Satz 2.82]. □

Zum Abschluss betrachten wir ein weiteres Beispiel der Regressionsnalayse.

Beispiel 2.84. (Lineares Modell mit zwei qualitativen Faktoren) Gegeben sei eine Situation, in der ein zufallsabhängiges Ergebnis von zwei qualitativen Faktoren F_1 und F_2 abhängt. Denken wir etwa an die Auswirkung verschiedener Lernmethoden (Faktor 1) in verschiedenen Klassenstufen (Faktor 2) auf die Testergebnisse von Schülern oder auch die Auswirkung bestimmter Behandlungsmethoden (Faktor 1) auf die Heilungsergebnisse von Patienten bei unterschiedlicher Schwere der Erkrankung (Faktor 2). Die Beobachtung besteht hier aus Zufallsgrößen

$$X_{ijk} = \theta_{ij}+\varepsilon_{ijk},\quad i=1,...,I,\ j=1,...,J,\ k=1,...,n_{ij},$$

wobei

I,J die Anzahl der Stufen zu Faktor 1 bzw. Faktor 2,
n_{ij} die Anzahl der Beobachtungen zu Faktorkombination (i,j)

[8] *Beweis:* Für eine symmetrische Matrix C existiert nach der Hauptachsentransformation eine orthogonale Matrix U derart, dass $U^\top CU = \mathrm{diag}(\alpha_1,...,\alpha_n)$, wobei die α_j die Eigenwerte von C bezeichnen. Ist C außerdem idempotent, so liefert $\alpha x = Cx = C^2x = \alpha Cx = \alpha^2 x$ für jeden Eigenwert α mit Eigenvektor $x\neq 0$, dass $\alpha\in\{0,1\}$. Nun folgt

$$\begin{aligned}\mathrm{Rang}(C) &= \mathrm{Rang}(U^\top CU) = \mathrm{Rang}(\mathrm{diag}(\alpha_1,...,\alpha_n)) = |\{j:\alpha_j=1\}|\\ &= \mathrm{Spur}(\mathrm{diag}(\alpha_1,...,\alpha_n)) = \mathrm{Spur}(U^\top CU) = \mathrm{Spur}(CUU^\top) = \mathrm{Spur}(C).\end{aligned}$$

bezeichnen. Der Parameter $\theta = (\theta_{11}, \theta_{12}, ..., \theta_{IJ})^\top \in \widetilde{\Theta} = \mathbb{R}^{I \cdot J}$ ist unbekannt, ebenso wie die Verteilungen $Q_{ijk} \in \mathscr{Q}$, $\mathscr{Q}$ analog zu (2.28), der stochastisch unabhängigen Fehler ε_{ijk}. Es liegt damit ein lineares Modell der Form $X = A\theta + \varepsilon$ vor mit

$$\begin{aligned} X &= (X_{111}, X_{112}, ..., X_{11n_{11}}, X_{121}, ..., X_{12n_{12}}, ..., X_{IJ1}, ..., X_{1Jn_{IJ}})^\top, \\ \varepsilon &= (\varepsilon_{111}, \varepsilon_{112}, ..., \varepsilon_{11n_{11}}, \varepsilon_{121}, ..., \varepsilon_{12n_{12}}, ..., \varepsilon_{IJ1}, ..., \varepsilon_{IJn_{IJ}})^\top, \end{aligned}$$

$$A = \begin{pmatrix} \left.\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}\right\} n_{11}\text{-mal} & & \mathbf{0} \\ & \left.\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}\right\} n_{12}\text{-mal} & \\ & \ddots & \\ \mathbf{0} & & \left.\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}\right\} n_{IJ}\text{-mal} \end{pmatrix} \in \mathbb{R}^{\sum_{i,j} n_{ij} \times IJ}$$

Wie man sofort erkennt, gilt $\mathrm{Rang}(A) = IJ$, d.h. das Modell besitzt vollen Rang. Ferner folgt

$$A^\top A = \mathrm{diag}(n_{11}, n_{12}, ..., n_{IJ} \quad \text{und} \quad (A^\top A)^{-1} = \mathrm{diag}(n_{11}^{-1}, n_{12}^{-1}, ..., n_{IJ}^{-1}).$$

Im Folgenden greifen wir auf die nachstehende, in der Statistik oft benutzte Schreibweise zurück:

$$x_{ij\bullet} := \sum_{k=1}^{n_{ij}} x_{ijk} \quad \text{und} \quad \overline{x}_{ij\bullet} := \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} x_{ijk} = \frac{x_{ij\bullet}}{n_{ij}}.$$

Für den KQS $\widehat{\theta}(x)$ von θ ergibt sich dann

$$\widehat{\theta}(x) = (A^\top A)^{-1} A^\top x = (A^\top A)^{-1} \begin{pmatrix} x_{11\bullet} \\ x_{12\bullet} \\ \vdots \\ x_{IJ\bullet} \end{pmatrix} = \begin{pmatrix} \overline{x}_{11\bullet} \\ \overline{x}_{12\bullet} \\ \vdots \\ \overline{x}_{IJ\bullet} \end{pmatrix}$$

und somit, wie zu erwarten war, $\overline{x}_{ij\bullet}$ als GBLES für θ_{ij} nach dem Satz von Gauß-Markov. Für die unbekannte Fehlervarianz σ^2 lautet hier der in Satz 2.83 angegebene erwartungstreue Schätzer

$$\frac{1}{n - IJ} \sum_{i,j,k} (x_{ijk} - \overline{x}_{ij\bullet})^2, \quad n := \sum_{i,j} n_{ij},$$

vorausgesetzt, dass mindestens ein n_{ij} grösser als 1 ist. In vielen Fällen findet man das betrachtete Modell jedoch in einer anderen Parametrisierung vor, die auf einer Zerlegung des Mittleren Effekts θ_{ij} bei Faktorkombination (i,j) beruht. Es gilt nämlich

$$\theta_{ij} \;=\; \overline{\theta}_{\bullet\bullet} + (\overline{\theta}_{i\bullet} - \overline{\theta}_{\bullet\bullet}) + (\overline{\theta}_{\bullet j} - \overline{\theta}_{\bullet\bullet}) + ((\theta_{ij} - \overline{\theta}_{\bullet\bullet}) - (\overline{\theta}_{i\bullet} - \overline{\theta}_{\bullet\bullet}) - (\overline{\theta}_{\bullet j} - \overline{\theta}_{\bullet\bullet})),$$

wobei nach derselben Konvention wie oben

$$\overline{\theta}_{i\bullet} \;:=\; \frac{1}{J}\sum_{j=1}^{J}\theta_{ij}, \quad \overline{\theta}_{\bullet j} \;:=\; \frac{1}{J}\sum_{i=1}^{I}\theta_{ij} \quad \text{und} \quad \overline{\theta}_{\bullet\bullet} \;:=\; \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\theta_{ij}.$$

Definieren wir nun

$$\begin{aligned} \mu \;&:=\; \overline{\theta}_{\bullet\bullet}, \quad \alpha_i := \overline{\theta}_{i\bullet} - \overline{\theta}_{\bullet\bullet}, \quad \beta_j \;:=\; \overline{\theta}_{\bullet j} - \overline{\theta}_{\bullet\bullet} \\ \text{und} \quad \gamma_{ij} \;&:=\; (\theta_{ij} - \overline{\theta}_{\bullet\bullet}) - (\overline{\theta}_{i\bullet} - \overline{\theta}_{\bullet\bullet}) - (\overline{\theta}_{\bullet j} - \overline{\theta}_{\bullet\bullet}), \end{aligned}$$

so gilt offensichtlich

$$\theta_{ij} \;=\; \mu + \alpha_i + \beta_j + \gamma_{ij}$$

mit den Nebenbedingungen

$$\sum_{i=1}^{I}\alpha_i \;=\; \sum_{j=1}^{J}\beta_j \;=\; \sum_{i=1}^{I}\gamma_{ij} \;=\; \sum_{j=1}^{J}\gamma_{ij} \;=\; 0$$

und der Interpretation

$$\begin{aligned} \mu \;&=\; \text{mittlerer Gesamteffekt,} \\ \alpha_i \;&=\; \text{mittlerer Zusatzeffekt von Faktor } F_1 \text{ in Stufe } i, \\ \beta_j \;&=\; \text{mittlerer Zusatzeffekt von Faktor } F_2 \text{ in Stufe } j, \\ \gamma_{ij} \;&=\; \text{mittlere Wechselwirkung von Faktor } F_1 \text{ in Stufe } i \text{ mit Faktor } F_2 \text{ in Stufe } j. \end{aligned}$$

Nach den obigen Ausführungen und linearen Darstellungen der Parameter μ, α_i, β_j und γ_{ij} durch die θ_{ij} ergeben sich nun sofort als deren GBLES

$$\begin{aligned} \widehat{\mu}(x) \;&=\; \frac{1}{IJ}\sum_{i,j}\overline{x}_{ij\bullet}, \\ \widehat{\alpha}_i(x) \;&=\; \frac{1}{J}\sum_{j}\overline{x}_{ij\bullet} - \widehat{\mu}(x), \\ \widehat{\beta}_j(x) \;&=\; \frac{1}{I}\sum_{i}\overline{x}_{ij\bullet} - \widehat{\mu}(x), \\ \widehat{\gamma}_{ij}(x) \;&=\; \overline{x}_{ij\bullet} - \widehat{\alpha}_i(x) - \widehat{\beta}_j(x) + \widehat{\mu}(x). \end{aligned}$$

Diese Formeln vereinfachen sich noch im Fall $n_{ij} = k$ für alle i, j, und zwar zu

$$\widehat{\mu}(x) \;=\; \bar{x}_{\bullet\bullet\bullet} \;:=\; \frac{1}{IJK}\sum_{i,j,k} x_{ijk}, \quad \widehat{\alpha}_i(x) \;=\; \bar{x}_{i\bullet\bullet} - \bar{x}_{\bullet\bullet\bullet},$$

$$\widehat{\beta}_j(x) \;=\; \bar{x}_{\bullet j\bullet} - \bar{x}_{\bullet\bullet\bullet} \quad \text{und} \quad \widehat{\gamma}_{ij}(x) \;=\; \bar{x}_{ij\bullet} - \bar{x}_{i\bullet\bullet} - \bar{x}_{\bullet j\bullet} + \bar{x}_{\bullet\bullet\bullet},$$

wobei die nicht definierten Bezeichnungen $\bar{x}_{i\bullet\bullet}$ und $\bar{x}_{\bullet j\bullet}$ die kanonische Bedeutung besitzen.

2.8.2 Heteroskedastische Modelle

Es leuchtet ein, dass die bisher gemachte Voraussetzung gleicher Fehlervarianzen sowohl in Beispiel 2.76 als auch in vielen anderen Situationen unangebracht sein kann. So erscheint in 2.76 die Möglichkeit vorstellbar, dass bei einer großen eingesetzten Düngemittelmenge k_j die Schwankung des aufgenommenen Phosphors durch die Kornpflanzen entsprechend größeren Schwankungen unterliegt als bei kleinen Mengen k_j. Andererseits kennt der Versuchsleiter eventuell die Abhängigkeit der Varianz des Fehlers ε_j von dem Wert k_j zumindest bis auf eine multiplikative Konstante. Gegeben ein allgemeines lineares Modell $X = A\theta + \varepsilon$, lautet die dementsprechende Annahme in präziser Form

$$\mathbb{V}\mathrm{ar}_\vartheta \varepsilon_j \;=\; w_j\sigma^2, \quad j = 1,...,n, \tag{2.35}$$

für unbekanntes $\sigma^2 > 0$, aber bekannten Gewichten $w_1,...,w_n \in (0,\infty)$. Obwohl der KQS für θ auch in dieser Situation wie zuvor berechnet werden kann, verliert er i.A. seine in den Sätzen 2.81 und 2.82 nachgewiesenen Optimalitätseigenschaften. In der Tat ist die Methode der kleinsten Quadrate nicht mehr direkt anwendbar, sondern erst nach einer Transformation der Beobachtungsvariablen $X = (X_1,...,X_n)^\top$. Wir setzen

$$Y_j \;:=\; w_j^{-1/2} X_j$$

für $j = 1,...,n$, gehen also über zu dem Beobachtungsvektor $Y = W^{-1/2}X$ mit der Matrix

$$W \;:=\; \begin{pmatrix} w_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_n \end{pmatrix},$$

die wir als *Gewichts-Matrix* bezeichnen. Wir gelangen so zu dem neuen linearen Modell

$$Y \;=\; B\theta + \eta \tag{2.36}$$

mit Design-Matrix $B := W^{-1/2}A$ und Fehlervektor $\eta = W^{-1/2}\varepsilon$, d.h.

$$\eta \;=\; (\eta_1,...,\eta_n)^\top \quad \text{wobei} \quad \eta_j \;=\; w_j^{-1/2}\varepsilon_j \;\text{ für }\; j = 1,...,n.$$

Unter Hinweis auf (2.35) folgt sofort

$$\mathbb{E}_\vartheta \eta_j \;=\; 0 \quad \text{und} \quad \operatorname{Var}_\vartheta \eta_j \;=\; \sigma^2$$

für $j = 1,\ldots,n$. Ferner haben A und B denselben Rang, weil W invertierbar ist. Wir können demnach festhalten:

Lemma 2.85. *Gegeben ein heterosketastisches Modell $X = A\theta + \varepsilon$, dessen Fehler der Bedingung* (2.35) *für positive Gewichte $w_1,\ldots,w_n$ genügen, ist das zugehörige transformierte lineare Modell $Y = B\theta + \eta$ in* (2.36) *homoskedastisch. $X = A\theta + \varepsilon$ hat außerdem genau dann vollen Rang, wenn dies für $Y = B\theta + \varepsilon$ gilt.*

Nach dieser Feststellung ist das weitere Vorgehen kanonisch und kann in Kürze abgehandelt werden: Der KQS in dem transformierterten Modell lautet

$$\phi(y) \;:=\; (B^\top B)^{-1} B^\top y.$$

Er ist auf die transformierte Beobachtung $y = W^{-1/2}x$ anzuwenden, was unter Beachtung von $B = W^{-1/2}A$ und $(W^{-1/2})^\top = W^{-1/2}$ zu dem Schätzer

$$\widehat{\theta}(x) \;:=\; \phi(y) = \phi(W^{-1/2}x) \;=\; (A^\top W^{-1} A)^{-1} A^\top W^{-1} x$$

für das Ausgangsmodell $X = A\theta + \varepsilon$ führt. Wegen

$$\begin{aligned}(y - B\phi(y))^\top (y - B\phi(y)) \;&=\; \min_{\theta\in\widetilde{\Theta}} (y - B\theta)^\top (y - B\theta) \\ &=\; \min_{\theta\in\widetilde{\Theta}} (W^{1/2}y - A\theta)^\top W^{-1} (W^{1/2}y - A\theta)\end{aligned}$$

gemäß (2.35) folgt

$$\begin{aligned}(x - A\widehat{\theta}(x))^\top W^{-1} (x - A\widehat{\theta}(x)) \;&=\; (W^{1/2}y - W^{1/2}B\phi(y)^\top W^{-1} (W^{1/2}y - W^{1/2}B\phi(y)) \\ &=\; \min_{\theta\in\widetilde{\Theta}} (W^{1/2}y - A\theta)^\top W^{-1} (W^{1/2}y - A\theta) \\ &=\; \min_{\theta\in\widetilde{\Theta}} (x - A\theta)^\top W^{-1} (x - A\theta).\end{aligned}$$

Der Schätzer $\widehat{\theta}(x)$ löst somit anstelle der ursprünglichen Minimierungsaufgabe

$$(x - A\theta)^\top (x - A\theta) \;\to\; \min$$

die modifizierte Minimierungsaufgabe

$$(x - A\theta)^\top W^{-1} (x - A\theta) \;\to\; \min,$$

in der das gewöhnliche euklidische Skalarprodukt $u^\top v$ durch die quadratische Form $u^\top W^{-1} v$ ersetzt wird. Wegen der auftretenden Gewichte $w_1,\ldots,w_n$ bezeichnet man $\widehat{\theta}(x)$ als *gewichteten Kleinste-Quadrate-Schätzer (GKQS)*.

Eine einfache Überlegung liefert weiter, dass statt der Summe der Fehlerquadrate hier die *gewichtete Summe der Fehlerquadrate (GSFQ)*

$$\mathrm{GSFQ}(x) \;:=\; (x-A\widehat{\theta}(x))^\top W^{-1}(x-A\widehat{\theta}(x)) \;=\; (y-B\phi(y))^\top (y-B\phi(y))$$

zur Schätzung des unbekannten Parameters σ^2 heranzuziehen ist. Unter Verwendung der Sätze 2.81 (von Gauß-Markov), 2.82 und 2.83 für das transformierte Modell kann man nun leicht deren Verallgemeinerung auf den heteroskedatischen Fall der beschriebenen Form zeigen. Wir beschränken uns daher auf die Zusammenfassung der Ergebnisse in einem einzigen Satz und überlassen den Beweis dem geneigten Leser.

Satz 2.86. *In einem heteroskedastischen linearen Modell $X = A\theta + \varepsilon$ vollen Rangs mit Gewichtsmatrix $W = diag(w_1,...,w_n)$, $w_1,...,w_n > 0$, sei*

$$\widehat{\theta}(x) \;=\; (A^\top W^{-1}A)^{-1}A^\top W^{-1}x$$

der GKQS für θ und

$$GSFQ(x) = (x-A\widehat{\theta}(x))^\top W^{-1}(x-A\widehat{\theta}(x))$$

die gewichtete Summe der Fehlerquadrate.

(a) Zu gegebenem $\beta \in \mathbb{R}^p$ ist dann $\beta^\top\widehat{\theta}$ der GBLES für $\gamma(\vartheta) = \beta^\top\theta$ mit Risiko

$$R(\vartheta,\beta^\top\widehat{\theta}) \;=\; \mathbb{V}\mathrm{ar}_\vartheta\,\beta^\top\widehat{\theta}(X) = \sigma^2\beta^\top(A^\top W^{-1}A)^{-1}\beta$$

für alle $\vartheta \in \Theta$. Er ist sogar der GBES im Fall normalverteilter Fehler, d.h. $Q_j = Normal(0, w_j\sigma^2)$ für $j = 1,...,n$.

(b) Sofern $n > p$, bildet $(n-p)^{-1}$ $GSFQ(x)$ einen erwartungstreuen Schätzer für σ^2, der im Fall normalverteilter Fehler sogar der GBES ist.

Statistikerwitze zum Zweiten

Ein Mann in einem Heißluftballon hat sich verirrt. Über einem Feld lässt er den Ballon deshalb absinken und ruft einem Mann am Boden zu:
“Können sie mir vielleicht sagen, wo ich mich gerade befinde und in welche Richtung ich fliege?”
“Selbstverständlich!” entgegnet der Gefragte. “Sie befinden sich bei 43 Grad, 12 Minuten, 21,2 Sekunden Nord und 123 Grad, 8 Minuten, 12,8 Sekunden West. Ihre Höhe beträgt 212 Meter über dem Meeresspiegel. Momentan schweben sie in der Luft, aber auf dem Weg hierhin betrug ihre Geschwindigkeit 1,83 Meter pro Sekunde bei einer Winkelfrequenz von 1,929.”
“Besten Dank! Ach sagen sie, sind sie ein Statistiker?”
“Das bin ich! Aber wie haben sie das bemerkt?”
“Nun ja, alles, was sie mir mitgeteilt haben, ist so vollkommen präzise. Zudem haben sie mir viel mehr Details gegeben als ich benötige, and dies in einer Weise, die mir nicht im geringsten weiterhilft!”
“Autsch! Aber dann sagen sie mir doch, sind sie etwa ein Teilprojektleiter in einer Forschungseinrichtung?”
“Wow! Wie haben sie das denn erkannt?”
“Nun, sie wissen nicht, wo sie sich gerade befinden und wohin sie fliegen. Sie sind bis hierhin gelangt, indem sie heiße Luft verbreitet haben, sie stellen lauter Fragen, nachdem sie in Schwierigkeiten geraten sind, und sie sind noch immer an derselben Stelle wie vor einigen Minuten, aber jetzt ist es natürlich mein Fehler!”

Ein Statistiker versuchte vertrauensvoll einen Fluss zu durchqueren, dessen durchschnittliche Tiefe einen Meter betrug. Er ertrank!

Ein Mathematiker, ein Physiker und ein Statistiker begeben sich auf die Hirschjagd. Als schließlich ein Bock in ihr Visier gerät, schießt der Mathematiker als erstes und verfehlt die Nase des Hirsches nur um Zentimeter. Danach feuert der Physiker eine Kugel ab, die aber hauchdünn am Schwanz des Bocks vorbeisaust. Daraufhin springt der Statistiker freudig von einem Bein aufs andere und jubelt “Wir haben ihn erwischt, wir haben ihn erwischt!”

Kapitel 3
Parametrische Testtheorie

3.1 Beste Tests zum Niveau α bei einfachen Hypothesen: Das Neyman-Pearson-Lemma

Die grundlegende Beschreibung von Testproblemen einschließlich der wichtigsten Begriffsbildungen hatten wir in Unterabschnitt 1.4.3 vorgenommen. Zur Erinnerung fassen wir noch einmal das Wichtigste zusammen: Auf der Basis eines statistischen Experiments $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ mit Parameterraum Θ soll entschieden werden, ob der unbekannte Parameter θ in einer vorgegebenen Teilmenge H von Θ liegt, genannt *Hypothese* oder auch *Null-Hypothese*, oder in deren Komplement $K = \Theta\backslash H$, genannt *Alternative*. Dabei soll die Wahrscheinlichkeit, sich irrtümlich für die Alternative zu entscheiden *(Fehler 1. Art)*, ein vorgegebenes *Signifikanzniveau* nicht überschreiten. Als Entscheidungsfunktionen betrachtet man sämgliche messbaren Abbildungen $\varphi : \mathfrak{X} \to [0,1]$, bezeichnet als *Tests* oder *Testfunktionen*, wobei $\varphi(x)$ die Wahrscheinlichkeit angibt, sich bei Beobachtung von x für K zu entscheiden. Im Fall $\varphi(x) \in (0,1)$ wird ein weiteres Bernoulli-Experiment durchgeführt, das mit Wahrscheinlichkeit $\varphi(x)$ eine "1" und damit eine Entscheidung für K liefert. Dieser Vorgang heißt *"Randomisieren"* und jeder Test φ, der nicht nur die Werte 0 und 1 annimmt, *randomisierter Test*.

Als Verlustfunktion wird die *Neyman-Pearsonsche Verlustfunktion* benutzt:

$$L(\theta,\gamma) \;=\; \begin{cases} \gamma, & \text{falls } \theta \in H, \\ 1-\gamma, & \text{falls } \theta \in K \end{cases}$$

für alle $\gamma \in [0,1]$. Ein Test φ hat dann die Risikofunktion

$$R(\theta,\varphi) \;=\; \begin{cases} \int \varphi \, dW_\theta \;=\; \mathbb{E}_\theta \varphi(X), & \text{falls } \theta \in H, \\ \int (1-\varphi) \, dW_\theta \;=\; 1 - \mathbb{E}_\theta \varphi(X), & \text{falls } \theta \in K. \end{cases}$$

Die Funktion $\beta_\varphi = \theta \mapsto \mathbb{E}_\theta \varphi(X)$ nennt man die *Gütefunktion, OC-Funktion* oder auch *Operationscharakteristik* von φ. Für $\theta \in H$ gibt sie offensichtlich die Irrtums-

wahrscheinlichkeit unter W_θ an (= Fehler 1. Art), während diese im Fall $\theta \in K$ gerade durch $1-\beta_\varphi(\theta)$ gegeben ist (= Fehler 2. Art). Erfüllt die Gütefunktion bei gegebenem $\alpha \in [0,1]$

$$\beta_\varphi(\theta) \;=\; \mathbb{E}_\theta \varphi(X) \;\leq\; \alpha$$

für alle $\theta \in H$, heißt φ *Test zum Niveau* α (für H gegen K), deren Gesamtheit wir mit Φ_α bezeichnen. Weiter heißt φ *gleichmäßig bester Test zum Niveau* α (für H gegen K), wenn er unter allen Tests zum Niveau α den Fehler 2. Art minimiert, d.h., wenn

$$\mathbb{E}_\theta \varphi(X) \;=\; \max_{\psi \in \Phi_\alpha} \mathbb{E}_\theta \psi(X)$$

für alle $\theta \in K$ gilt.

Die Philosophie hinter der Betrachtung von Tests zum Niveau α besteht darin, dass in Ermangelung eines global besten Tests, also eines solchen, der sowohl den Fehler 1. als auch 2. Art unter allen Tests minimiert, die Entscheidung für die – zumeist riskantere – Alternative Priorität genießt und nur mit einer Wahrscheinlichkeit von höchstens α irrtümlich getroffen werden soll. Unter allen Tests, die diese Voraussetzung erfüllen, wird dann derjenige gewählt, der außerdem die Alternative mit minimaler Wahrscheinlichkeit irrtümlich verwirft.

Quelle: XKCD (A Webcomic of Romance, Sarcasm, Math, and Language).

Wie aber lassen sich solche Tests finden? Dazu betrachten wir zunächst den elementaren Fall *einfacher Hypothesen* $H = \{\theta_0\}$ und $K = \{\theta_1\}$, der zwar aus praktischer Sicht nur wenig relevant ist, der aber grundlegende Ekenntnisse liefert, die später zur Behandlung interessanterer Probleme benutzt werden können. Fundamental für alles Weitere ist das folgende, von J. NEYMAN & E. PEARSON [14] stammende und nach ihnen benannte Ergebnis:

Lemma 3.1. (von Neyman-Pearson) *Seien W_0 und W_1 W-Maße auf $(\mathfrak{X}, \mathscr{A})$ mit Dichten f_0 bzw. f_1 bzgl. eines dominierenden Maßes μ (etwa $\mu = W_0 + W_1$). Ferner seien $\alpha \in (0,1)$ und $\Phi_\alpha = \{\varphi : \varphi$ Test mit $\int \varphi dW_0 \leq \alpha\}$. Dann gilt:*

(a) ***(Hinreichende Bedingung)*** *Sei ψ ein Test mit:*

(1) $\int \psi \, dW_0 = \alpha$ $(\Rightarrow \psi \in \Phi_\alpha)$.

(2) Es existiert ein $k \in [0,\infty)$ derart, dass

$$\psi(x) \;=\; \begin{cases} 1, & \text{falls } f_1 \gtrless k f_0(x) \quad \mu\text{-f.ü.} \\ 0, & \end{cases} \tag{3.1}$$

Dann folgt

$$\int \psi \, dW_1 \;=\; \max_{\varphi \in \Phi_\alpha} \int \varphi \, dW_1. \tag{3.2}$$

(b) ***(Existenz)*** *Es existiert ein Test ψ, der (a.1) und (a.2) erfüllt.*

(c) ***(Notwendige Bedingung)*** *Für jeden Test $\psi \in \Phi_\alpha$, der* (3.2) *erfüllt, folgt: ψ ist von der Form in* (3.1)*. Falls außerdem $\int \psi \, dW_1 < 1$ gilt, so erfüllt ψ auch (a.1)*

Beweis. (a) Sei $\varphi \in \Phi_\alpha$ ein beliebiger Test zum Niveau α. Es gilt dann

$$(\psi(x) - \varphi(x))(f_1(x) - k f_0(x)) \;\geq\; 0 \quad \mu\text{-f.ü.}, \tag{3.3}$$

da für μ-fast alle $x \in \mathfrak{X}$

$$\begin{aligned} f_1(x) - k f_0(x) > 0 \quad &\Rightarrow \quad \psi(x) = 1 \quad \Rightarrow \quad \psi(x) - \varphi(x) \geq 0, \\ f_1(x) - k f_0(x) < 0 \quad &\Rightarrow \quad \psi(x) = 0 \quad \Rightarrow \quad \psi(x) - \varphi(x) \leq 0. \end{aligned}$$

Integriert man Ungleichung (3.3), so folgt weiter

$$\int \psi f_1 \, d\mu \;-\; \int \varphi f_1 \, d\mu \;-\; k \left(\int \psi f_0 \, d\mu \;-\; \int \varphi f_0 \, d\mu \right) \;\geq\; 0,$$

also

$$\int \psi \, dW_1 \;-\; \int \varphi \, dW_1 \;\geq\; k \left(\underbrace{\int \psi \, dW_0}_{=\alpha} \;-\; \underbrace{\int \varphi \, dW_0}_{\leq \alpha} \right) \;\geq\; 0.$$

Die gewünschte Gleichheit resultiert nun aus $\psi \in \Phi_\alpha$.

(b) Betrachte im folgenden die Tests

$$\varphi_{k,\gamma}(x) \;=\; \begin{cases} 1, & \\ \gamma, & \text{falls } f_1 \;\gtreqless\; k f_0(x) \\ 0, & \end{cases}$$

für $k \in [0, \infty)$ und $\gamma \in [0,1]$, also $\varphi_{k,\gamma} = \mathbf{1}_{\{f_1 > k f_0\}} + \gamma \mathbf{1}_{\{f_1 = k f_0\}}$. Diese Tests erfüllen offenkundig (a.2). Wir zeigen nun, dass es ein Paar (k, γ) mit $\int \varphi_{k,\gamma} dW_0 = \alpha$ gibt. Dazu setzen wir

$$T(x) = \begin{cases} \dfrac{f_1(x)}{f_0(x)}, & \text{falls } f_0(x) > 0 \\ \infty, & \text{falls } f_0(x) = 0 \end{cases}$$

und erhalten für alle x mit $f_0(x) > 0$

$$f_1(x) \;\gtreqless\; k f_0(x) \quad \Leftrightarrow \quad T(x) \;\gtreqless\; k,$$

also

$$\begin{aligned}\int \varphi_{k,\gamma}\, dW_0 &= \int_{\{f_0>0\}} \varphi_{k,\gamma} f_0\, d\mu \\ &= \int_{\{f_0>0\}} \big(\mathbf{1}_{\{T>k\}} + \gamma \mathbf{1}_{\{T=k\}}\big) f_0\, d\mu \\ &= \int \big(\mathbf{1}_{\{T>k\}} + \gamma \mathbf{1}_{\{T=k\}}\big)\, dW_0 \\ &= W_0(T>k) + \gamma W_0(T=k).\end{aligned}$$

Es sind daher $k \in [0,\infty)$ und $\gamma \in [0,1]$ so zu bestimmen, dass

$$W_0(T>k) + \gamma W_0(T=k) \;=\; \alpha. \tag{3.4}$$

Sei

$$k \;:=\; \inf\{y \geq 0 : W_0(T>y) \leq \alpha\} \;=\; \inf\{y \geq 0 : W_0(T \leq k) \geq 1-\alpha\}.$$

Wegen $\alpha \in (0,1)$ gilt $k < \infty$, und aus der rechtsseitigen Stetigkeit von $y \mapsto W_0(T>y)$ folgt $W_0(T>k) \leq \alpha$ sowie im Fall $W_0(T>k) < \alpha$ außerdem

$$\begin{aligned}W_0(T=k) &= W_0(T \geq k) - W_0(T>k) \\ &> W_0(T \geq k) - \alpha \\ &= \lim_{y \uparrow k} (W_0(T>y) - \alpha) \;\geq\; 0.\end{aligned}$$

Wir setzen deshalb weiter

$$\gamma \;:=\; \begin{cases} 0, & \text{falls } W_0(T>k) = \alpha, \\ \dfrac{\alpha - W_0(T>k)}{W_0(T=k)}, & \text{falls } W_0(T>k) < \alpha. \end{cases}$$

und erhalten dann in der Tat (3.4) und damit (b), denn

$$0 \;\leq\; \gamma \;\leq\; \frac{W_0(T \geq k) - W_0(T<k)}{W_0(T=k)} \;=\; 1.$$

(c) Sei $\psi \in \Phi_\alpha$ ein Test, der (3.2) genügt, und φ ein weiterer, gemäß Teil (b) existierender Test, der (a.1) und (a.2) erfüllt. Betrachte die Menge

$$A := \; \{x \in \mathfrak{X} : \psi(x) = \varphi(x) \text{ oder } f_1(x) = k f_0(x)\}.$$

Um die Gestalt (3.1) für ψ nachzuweisen, genügt es offensichtlich, $\mu(A^c) = 0$ zu zeigen. Nehmen wir dazu das Gegenteil an, also $\mu(A^c) > 0$. Dann folgt

$$\int (\varphi - \psi)(f_1 - k f_0)\, d\mu \;=\; \int_A (\varphi - \psi)(f_1 - k f_0)\, d\mu \;>\; 0,$$

denn

$$(\varphi-\psi)(f_1-kf_0) \;=\; (1-\psi)(f_1-kf_0) \;>\; 0$$

auf $A^c\cap\{f_1>kf_0\}$ und

$$(\varphi-\psi)(f_1-kf_0) \;=\; -\psi(f_1-kf_0) \;>\; 0$$

auf $A^c\cap\{f_1<kf_0\}$. Damit erhalten wir aber den Widerspruch

$$\begin{aligned}\int\varphi\,dW_1 \;-\; \int\psi\,dW_1 \;&>\; \left(\int\varphi\,dW_0 \;-\; \int\psi\,dW_0\right)\\ &=\; k\left(\alpha \;-\; \int\psi\,dW_0\right) \;\geq\; \alpha.\end{aligned}$$

Sei schließlich $\int\psi\,dW_1<1$ und $\int\psi\,dW_0<\alpha$ angenommen. Dann folgt $W_1(B)>0$ für $B=\{\psi<1\}$, und wir können $\varepsilon>0$ mit $\varepsilon W_0(B)\leq\alpha-\int\psi\,dW_0$ wählen. Wir definieren nun einen besseren Test $\widehat{\psi}\in\Phi_\alpha$ als ψ durch

$$\widehat{\psi}(x) \;=\; \psi(x)\mathbf{1}_{B^c}(x) \;+\; \min\{\psi(x)+\varepsilon,1\}\,\mathbf{1}_B(x).$$

In der Tat gilt dann $\int\widehat{\psi}\,dW_1>\int\psi\,dW_1$ sowie

$$\int\widehat{\psi}\,dW_0 \;\leq\; \int\psi\,dW_0 \;+\; \varepsilon W_0(B) \;\leq\; \alpha,$$

also $\widehat{\psi}\in\Phi_\alpha$, womit ebenfalls ein Widerspruch erzeugt ist. □

Anmerkung 3.2. Ist $\alpha\in(0,1)$ und Q ein W-Maß auf $(\mathbb{R},\mathscr{B}(\mathbb{R}))$, so wird die Konstante

$$c(Q,\alpha) \;:=\; \inf\{y\in\mathbb{R}:Q((-\infty,y])\geq 1-\alpha\} \;=\; \inf\{y\in\mathbb{R}:Q((y,\infty))\leq\alpha\}$$

als α*-Fraktil von* Q bezeichnet. Für das im Beweis von 3.1(b) definierte k gilt dann

$$k \;=\; \inf\{y:W_0(T>y)\leq\alpha\} \;=\; \inf\{y:W_0^T((y,\infty))\leq\alpha\} \;=\; c(W_0^T,\alpha),$$

und man nennt k in diesem Fall auch das α*-Fraktil von* T *bzgl.* W_0.

3.2 Einseitige Testprobleme bei monotonen Dichtequotienten

Zur Motivation der nachfolgenden Theorie kehren wir einmal mehr zurück zu unserem Einführungsbeispiel 1.1.

Beispiel 3.3. Betrachten wir wieder die dortige Situation (Untersuchung eines neuen Produktionsprozesses), jedoch in allgemeiner Form. Dazu nehmen wir an, dass unsere Beobachtung aus n unabhängigen, jeweils $Bern(\theta)$-verteilten ZG $X_1,\ldots,X_n$

(Ergebnisse einer Qualitätskontrolle von n produzierten Stücken) besteht. Der Vektor $X = (X_1, ..., X_n)$ hat dann unter $\mathbb{P}_\theta$, $\theta \in \Theta = (0,1)$, die Zähldichte

$$f_\theta(x) \;=\; \theta^{s_n}(1-\theta)^{n-s_n}, \quad x \;=\; (x_1,...,x_n) \;\in\; \mathfrak{X} \;=\; \{0,1\}^n,\; s_n \;=\; \sum_{j=1}^{n} x_j.$$

Wir wollen das Testproblem $H = [\theta_0, 1)$ gegen $K = (0, \theta_0)$ betrachten, wobei $\theta_0 \in (0,1)$ ein kritischer Parameterwert sei; in der Situation von 1.1 ist θ_0 der Ausschussanteil bei Verwendung des alten Produktionsverfahrens. Die Menge K stellt demnach gerade die Menge der Parameterwerte dar, bei der eine Umstellung auf das neue Verfahren gerechtfertigt ist.

1. Schritt: Zur Anpassung der Testgestalt in (3.1) an die jetzige Situation wählen wir zunächst zwei beliebige $\eta_0, \eta_1 \in (0,1)$ mit $\eta_0 > \eta_1$. Es gilt dann für $k \in (0,\infty)$

$$\begin{aligned}
& f_{\eta_1}(x) \;\gtreqless\; k f_{\eta_0}(x) \\
\Leftrightarrow \quad & \left(\frac{\eta_1}{\eta_0}\right)^{s_n} \left(\frac{1-\eta_1}{1-\eta_0}\right)^{n-s_n} \;\gtreqless\; k \\
\Leftrightarrow \quad & \left(\frac{\eta_1(1-\eta_0)}{\eta_0(1-\eta_1)}\right)^{s_n} \;\gtreqless\; \left(\frac{1-\eta_0}{1-\eta_1}\right)^{n} \\
\Leftrightarrow \quad & s_n \log\Bigg(\underbrace{\frac{\eta_1(1-\eta_0)}{\eta_0(1-\eta_1)}}_{<1}\Bigg) \;\gtreqless\; \log(k) + n\log\left(\frac{1-\eta_0}{1-\eta_1}\right) \\
\Leftrightarrow \quad & s_n \;\lesseqgtr\; \frac{\log(k) + n\log(\frac{1-\eta_0}{1-\eta_1})}{\log(\frac{\eta_1(1-\eta_0)}{\eta_0(1-\eta_1)})}.
\end{aligned}$$

2. Schritt: Es wird nun ein Test φ^* für H gegen K konstruiert, der sich später als gleichmäßig bester zum Niveau α herausstellt. Wir definieren

$$c \;:=\; \max\{t \in \{0,...,n\} : \mathbb{P}_{\theta_0}(S_n < t) \le \alpha\} \quad \text{und} \quad \gamma \;:=\; \frac{\alpha - \mathbb{P}_{\theta_0}(S_n < c)}{\mathbb{P}_{\theta_0}(S_n = c)},$$

wobei $S_n = \sum_{j=1}^n X_j$. Dann gilt $\mathbb{P}_{\theta_0}(S_n \le c) > \alpha$ und somit

$$0 \;\le\; \gamma \;=\; \frac{\alpha - \mathbb{P}_{\theta_0}(S_n < c)}{\mathbb{P}_{\theta_0}(X_n = c)} \;<\; \frac{\mathbb{P}_{\theta_0}(S_n \le c) - \mathbb{P}_{\theta_0}(S_n < c)}{\mathbb{P}_{\theta_0}(S_n = c)} \;=\; 1.$$

Sei

$$\varphi^*(x) \;:=\; \begin{cases} 1, & \\ \gamma, & \text{falls } s_n \;\lesseqgtr\; c \\ 0, & \end{cases}$$

Für diesen Test folgt

$$\begin{aligned}\mathbb{E}_{\theta_0}\varphi^*(X) &= \mathbb{P}_{\theta_0}(\varphi^*(X)=1)+\gamma\mathbb{P}_{\theta_0}(\varphi^*(X)=\gamma)\\ &= \mathbb{P}_{\theta_0}(S_n<c)+\gamma\mathbb{P}_{\theta_0}(S_n=c)\\ &= \mathbb{P}_{\theta_0}(S_n<c)+\frac{\alpha-\mathbb{P}_{\theta_0}(S_n<c)}{\mathbb{P}_{\theta_0}(S_n=c)}\mathbb{P}_{\theta_0}(S_n=c) \;=\; \alpha,\end{aligned}$$

d.h. φ^* erfüllt Bedingung (a.1) aus dem Neyman-Pearson-Lemma mit $W_0 = W_{\theta_0}$.

3. Schritt: Aus nächstes zeigen wir, dass φ^* Bedingung (a.2) des Neyman-Pearson-Lemmas mit $W_0 = W_{\eta_0}$ und $W_1 = W_{\eta_1}$ für alle $\eta_0, \eta_1 \in (0,1)$ mit $\eta_0 > \eta_1$ erfüllt. Definiere dazu $k = k(\eta_0, \eta_1)$ für beliebige solche η_0, η_1 durch

$$c \;=\; \frac{\log(k)+n\log(\frac{1-\eta_0}{1-\eta_1})}{\log(\frac{\eta_1(1-\eta_0)}{\eta_0(1-\eta_1)})},$$

also

$$k \;:=\; \exp\left\{c\log\left(\frac{\eta_1(1-\eta_0)}{\eta_0(1-\eta_1)}\right)-n\log\left(\frac{1-\eta_0}{1-\eta_1}\right)\right\} \;>\; 0.$$

Gemäß Schritt 1 gilt dann

$$\varphi^*(x) \;=\; \left\{\begin{array}{l}1,\\ \gamma, \quad \text{falls } s_n \lessgtr c\\ 0,\end{array}\right\} \;=\; \left\{\begin{array}{l}1,\\ \gamma, \quad \text{falls } f_{\eta_1}(x) \gtrless k f_{\eta_0}(x)\\ 0,\end{array}\right.$$

und damit insbesondere

$$\varphi^*(x) \;=\; \begin{cases}1,\\ 0,\end{cases} \quad \text{falls } f_{\eta_1}(x) \gtrless k f_{\eta_0}(x).$$

4. Schritt: Sei nun $\theta \in K$ beliebig. Gemäß Schritt 2 und 3 erfüllt φ^* die Bedingungen in Lemma 3.1(a) mit $W_0 = W_{\theta_0}$ und $W_1 = W_\theta$. Daher folgt

$$\mathbb{E}_{\theta\varphi^*}(X) \;=\; \max_{\varphi:\mathbb{E}_{\theta_0}\varphi(X)\le\alpha} \mathbb{E}_{\theta\varphi}(X)$$

für alle $\theta \in K$.

5. Schritt: Um abschließend nachzuweisen, dass φ^* einen gleichmäßig besten Test zum Niveau α für $H = [\theta_0, 1)$ gegen $K = (0, \theta_0)$ definiert, bleibt also nur noch $\varphi^* \in \Phi_\alpha$, d.h.

$$\mathbb{E}_\theta\varphi^*(X) \;\le\; \alpha$$

für alle $\theta \in H$ zu zeigen. Wähle dazu ein beliebiges $\vartheta > \theta_0$. Gemäß Schritt 2 hat φ^* die Gestalt (3.1) aus Lemma 3.1 mit $W_0 = W_\vartheta$ und $W_0 = W_{\theta_0}$. Außerdem erfüllt φ^* trivialerweise auch (a.1) mit α ersetzt durch $\beta := \mathbb{E}_\vartheta\varphi^*(X)$. Aus diesem Grund folgt wiederum gemäß Lemma 3.1

$$\mathbb{E}_{\theta_0}\varphi^*(X) \;=\; \max_{\varphi:\mathbb{E}_\vartheta\varphi(X)\le\beta} \mathbb{E}_{\theta_0}\varphi(X) \;\ge\; \beta$$

(wähle $\varphi \equiv \beta$) und damit

$$\alpha \;=\; \mathbb{E}_{\theta_0}\varphi^*(X) \ge \beta \;=\; \mathbb{E}_\vartheta\varphi^*(X).$$

6. Schritt: Analog können wir das Testproblem $H = (0,\theta_0]$ gegen $K = (\theta_0,1)$ behandeln und erhalten für dieses einen gleichmäßig besten Test φ^* zum Niveau α der Form

$$\varphi^*(x) \;:=\; \begin{cases} 1, & \\ \gamma, & \text{falls } s_n \;\gtreqless\; c \\ 0, & \end{cases}$$

mit $c := \min\{t \in \{0,...,n\} : \mathbb{P}_{\theta_0}(S_n > t) \le \alpha\} = c(\mathbb{P}_{\theta_0}^{S_n}, \alpha) = c(Bin(n,\theta_0),\alpha)$ und

$$\gamma \;:=\; \frac{\alpha - \mathbb{P}_{\theta_0}(S_n > c)}{\mathbb{P}_{\theta_0}(S_n = c)}.$$

7. Schritt: Der Wert c, durch den der Test φ^* festgelegt wird, heißt *kritischer Wert*. Sofern der Stichprobenumfang n nicht zu groß ist, kann seine Bestimmung in der hier vorliegenden Situation unter Heranziehung einer Statistik-Software am PC oder mittels eines *Tafelwerks* für die Binomial-Verteilung erfolgen. Für große Werte n (Fausregel: $n\theta_0(1-\theta_0) \ge 9$) benutzt man dagegen eine Approximation durch die Normalverteilung *(Normalapproximation)*. Nach dem zentralen Grenzwertsatz gilt dann nämlich

$$\mathbb{P}_{\theta_0}\left(\frac{S_n - n\theta_0}{(n\theta_0(1-\theta_0))^{1/2}} \le z\right) \;\approx\; \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{z} e^{-y^2/2}dy \tag{3.5}$$

für alle $z \in \mathbb{R}$. Sei $u_\alpha := c(Normal(0,1),\alpha)$, d.h.

$$\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{u_\alpha} e^{-y^2/2}dy \;=\; 1-\alpha.$$

Dann liefert (3.5)

$$\mathbb{P}_{\theta_0}\left(\frac{S_n - n\theta_0}{(n\theta_0(1-\theta_0))^{1/2}} \le u_\alpha\right) \;\approx\; 1-\alpha$$

und folglich

$$\mathbb{P}_{\theta_0}(S_n > n\theta_0 + u_\alpha(n\theta_0(1-\theta_0))^{1/2}) \;\approx\; \alpha.$$

Für das Testproblem in Schritt 6 würde man also $c = n\theta_0 + u_\alpha(n\theta_0(1-\theta_0))^{1/2}$ und $\gamma = 0$ wählen, während man für das eingangs betrachtete Testproblem analog mit $u_{1-\alpha}$ statt u_α vorgeht. Dazu später mehr.

Wir wollen nun die angestellten Überlegungen des Beispiels auf allgemeinere Situationen übertragen, wozu wir mit einer Definition beginnen.

Definition 3.4. Sei $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$ ein dominiertes statistisches Experiment mit $\Theta \subset \mathbb{R}$. $(W_\theta)_{\theta\in\Theta}$ wird als *Familie mit monotonem Dichtequotienten* bezeichnet, wenn für sie ein dominierendes Maß μ existiert, so dass die μ-Dichten $f_\theta = dW_\theta/d\mu$ folgende Eigenschaft besitzen: Es existieren eine Statistik $T : \mathfrak{X} \to \mathbb{R}$ sowie für alle $(\theta_0, \theta_1) \in \Theta^2$ mit $\theta_0 < \theta_1$ eine monoton wachsende Abbildung $g_{\theta_0,\theta_1} : \mathbb{R} \to [0,\infty]$, so dass auf $\{x : (f_{\theta_0}(x), f_{\theta_1}(x) \neq (0,0)\}$

$$\frac{f_{\theta_1}}{f_{\theta_0}} = g_{\theta_0,\theta_1} \circ T \quad \mu\text{-f.ü.} \tag{3.6}$$

gilt. Wir sprechen dann von einer *Familie mit monotonem Dichtequotienten in T*.

Beispiel 3.5. Sei $(W_\theta)_{\theta\in\Theta}$ eine Exponentialfamilie mit $\Theta \subset \mathbb{R}$ und ν-Dichten

$$f_\theta(x) \;=\; C(\theta)e^{Q(\theta)T(x)}h(x),$$

wobei $C, Q : \Theta \to \mathbb{R}$ und $T, h : \mathfrak{X} \to \mathbb{R}$. Dann gilt $\{f_\theta = 0\} = \{h = 0\}$ für alle $\theta \in \Theta$ und ferner

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \;=\; \frac{C(\theta_1)}{C(\theta_0)}e^{(W(\theta_1)-Q(\theta_0))T(x)} \;=\; g_{\theta_0,\theta_1} \circ T(x)$$

für alle $x \in \{h \neq 0\}$ und $(\theta_0, \theta_1) \in \Theta^2$ mit $\theta_0 < \theta_1$, wobei

$$g_{\theta_0,\theta_1}(t) \;=\; \frac{C(\theta_1)}{C(\theta_0)}e^{(Q(\theta_1)-Q(\theta_0))t}.$$

Sofern $Q : \Theta \to \mathbb{R}$ monoton wachsend oder fallend ist, liegt also eine Familie mit monotonem Dichtequotienten in T bzw. $-T$ vor.

Sind speziell $X_1, ..., X_n$ stochastisch unabhängig und jeweils $Normal(\mu, \sigma^2)$-verteilt mit unbekanntem Mittelwert $\mu \in \mathbb{R}$ und bekannter Varianz $\sigma^2 > 0$, so hat $X = (X_1, ..., X_n)$ die λ^n-Dichte

$$f_{\mu,\sigma^2}(x) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \exp\left(\frac{\mu}{\sigma^2}\sum_{j=1}^n s_j\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^n x_j^2\right)$$

für $x = (x_1, ..., x_n) \in \mathbb{R}^n$ und somit $(Normal(\mu, \sigma^2)^n)_{\mu\in\mathbb{R}}$ einen monotonen Dichtequotienten $T(x) = \sum_{j=1}^n x_j$.

Ist dagegen der Mittelwert μ bekannt, jedoch die Varianz $\sigma^2 \in (0,\infty)$ unbekannt, so erhält man wegen

$$f_{\mu,\sigma^2}(x) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^n (s_j - \mu)^2\right),$$

dass $(Normal(\mu,\sigma^2)^n)_{\sigma^2>0}$ einen monotonen Dichtequotienten in der Statistik $T(x) = \sum_{j=1}^n (x_j-\mu)^2$ besitzt.

Beispiel 3.6. Sind $X_1,...,X_n$ stochastisch unabhängig und jeweils $Unif(0,\theta)$-verteilt mit unbekanntem Parameter $\theta > 0$, so hat $X = (X_1,...,X_n)$ die λ^n-Dichte

$$f_\theta(x) \;=\; \theta^{-n}\prod_{j=1}^n \mathbf{1}_{(0,\theta)}(x_j) \;=\; \theta^{-n}\mathbf{1}_{(0,\infty)}\left(\min_{1\le j\le n} x_j\right)\mathbf{1}_{(0,\theta)}\left(\max_{1\le j\le n} x_j\right).$$

Für $0<\theta_0<\theta_1$ und $x\in\{y: f_{\theta_0}(y)\vee f_{\theta_1}(y))>0\} = \{y: 0<\min_j y_j<\theta_1\}$ gilt dann

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \;=\; \left\{\begin{matrix}\infty, & \text{falls } \max_j x_j \ge \theta_0\\ (\theta_0/\theta_1)^n, & \text{falls } \max_j x_j < \theta_0\end{matrix}\right\} \;=\; g_{\theta_0,\theta_1}\left(\max_j x_j\right),$$

wobei $g_{\theta_0,\theta_1}:\mathbb{R}\to[0,\infty]$ durch

$$g_{\theta_0,\theta_1}(t) \;:=\; \begin{cases}\infty, & \text{falls } t\ge\theta_0,\\ (\theta_0/\theta_1)^n, & \text{falls } 0<t<\theta_0,\\ 0, & \text{falls } t\le 0\end{cases}$$

definiert ist. $(Unif(0,\theta)^n)_{\theta>0}$ besitzt also einen monotonen Dichtequotienten in der Statistik $T(x)=\max_j x_j$.

In Verallgemeinerung der Ergebnisse aus Beispiel 3.3 zeigen wir nun

Satz 3.7. *Sei $\mathcal{E} = (\mathfrak{X},\mathcal{A},(W_\theta)_{\theta\in\Theta})$ ein dominiertes statistisches Experiment mit $\Theta\subset\mathbb{R}$. Die Familie $(W_\theta)_{\theta\in\Theta}$ besitze einen monotonen Dichtequotienten in der Statistik $T:\mathfrak{X}\to\mathbb{R}$. Seien ferner $H=\{\theta\in\Theta:\theta\le\theta_0\}$ und $K=\{\theta\in\Theta:\theta>\theta_0\}$ für ein $\theta_0\in\Theta$ und $\alpha\in(0,1)$. Dann gelten die folgenden Aussagen:*

(a) Es existieren $c\in\mathbb{R}$ und $\gamma\in[0,1]$ mit

$$W_{\theta_0}(T>c) \;+\; \gamma W_{\theta_0}(T=c) \;=\; \alpha. \tag{3.7}$$

(b) Bildet (c^*,γ^*) eine Lösung von (3.7) und definiert man

$$\varphi^*(x) \;:=\; \begin{cases}1,\\ \gamma^*, & \text{falls } T(x) \gtreqless c^*\\ 0,\end{cases} \tag{3.8}$$

so folgt:

(1) $R(\theta,\varphi^*) = \min\limits_{\varphi:\mathbb{E}_{\theta_0}\varphi(X)=\alpha} R(\theta,\varphi)$ für alle $\theta\ne\theta_0$.

(2) φ^* ist gleichmäßig bester Test zum Niveau α für H gegen K.

Beweis. (a) Man wähle $c := c(W_{\theta_0}^T, \alpha)$ und dann

$$\gamma := \begin{cases} 0, & \text{falls } W_{\theta_0}(T > c) = \alpha, \\ \dfrac{\alpha - W_{\theta_0}(T > c)}{W_{\theta_0}(T = c)}, & \text{falls } W_{\theta_0}(T > c) < \alpha. \end{cases}$$

(b) Sei nun (c^*, γ^*) eine Lösung von (3.7) und φ^* gemäß (3.8) definiert. Dann folgt

$$\mathbb{E}_{\theta_0}\varphi^*(X) \;=\; W_{\theta_0}(T > c^*) + \gamma^* W_{\theta_0}(T = c^*) \;=\; \alpha.$$

Wähle zu $(W_\theta)_{\theta\in\Theta}$ die Funktionen $g_{\theta,\vartheta}$, $(\theta,\vartheta) \in \Theta^2$, gemäß Definition 3.4. Wir zeigen als nächstes, dass für alle $\theta > \theta_0$

$$\varphi^*(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } f_\theta(x) \;\gtrless\; k f_{\theta_0}(x) \quad \mu\text{-f.ü.}$$

mit $k := g_{\theta_0,\theta}(c^*) \in [0,\infty)$ gilt. Wäre $k = \infty$, folgte

$$\begin{aligned} \alpha \;&\le\; W_{\theta_0}(T \ge c^*) \\ &\le\; W_{\theta_0}(g_{\theta_0,\theta} \circ T \ge g_{\theta_0,\theta}(c^*)) \\ &=\; W_{\theta_0}(g_{\theta_0,\theta} \circ T = \infty) \\ &=\; W_{\theta_0}\left(\frac{f_\theta}{f_{\theta_0}} = \infty\right) \\ &\le\; W_{\theta_0}(f_{\theta_0} = 0) \\ &=\; \int_{\{f_{\theta_0}=0\}} f_{\theta_0}\, d\mu \;=\; 0, \end{aligned}$$

was offensichtlich einen Widerspruch bildet.

Wir erhalten ferner für μ-fast alle $x \in \mathfrak{X}$ nach Definition von φ^*:

$$\begin{aligned} f_\theta(x) \;\gtrless\; k f_{\theta_0}(x) \quad &\Rightarrow \quad f_\theta(x) \vee f_{\theta_0}(x)) > 0 \text{ und } \frac{f_\theta(x)}{f_{\theta_0}(x)} \;\gtrless\; k \\ &\Rightarrow \quad g_{\theta_0,\theta}(T(x)) \;\gtrless\; g_{\theta_0,\theta}(c^*) \\ &\Rightarrow \quad T(x) \;\gtrless\; c^* \\ &\Rightarrow \quad \varphi^*(x) \;=\; \begin{cases} 1 & \\ 0 & \end{cases}. \end{aligned}$$

Für $\theta \in K$ gilt $\theta > \theta_0$, und nach den obigen Überlegungen genügt φ^* den Bedingungen (a.1) und (a.2) des Neyman-Pearson-Lemmas 3.1 mit $W_0 = W_{\theta_0}$ und $W_1 = W_\theta$. Daher erhalten wir

$$1 - R(\theta, \varphi^*) \;=\; \mathbb{E}_\theta \varphi^*(X) \;=\; \max_{\varphi:\mathbb{E}_{\theta_0}\varphi(X)\le\alpha} \mathbb{E}_\theta \varphi(X) \;=\; \max_{\varphi:\mathbb{E}_{\theta_0}\varphi(X)\le\alpha} (1 - R(\theta,\varphi))$$

und somit Behauptung (1) für alle $\theta > \theta_0$, denn

$$R(\theta,\varphi^*) = \min_{\varphi:\mathbb{E}_{\theta_0}\varphi(X)\le\alpha} R(\theta,\varphi) = \min_{\varphi:\mathbb{E}_{\theta_0}\varphi(X)=\alpha} R(\theta,\varphi),$$

wobei die letzte Gleichheit wegen $\mathbb{E}_{\theta_0}\varphi^*(X) = \alpha$ gilt.

Für Behauptung (2) bleibt jetzt nur noch nachzuweisen, dass φ^* einen Test zum Niveau α bildet, d.h., dass $\mathbb{E}_\theta\varphi^*(X) \le \alpha$ für alle $\theta \in H$ gilt. Für solche θ, also $\theta < \theta_0$, ergibt sich analog zu vorher

$$\varphi^*(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } f_\theta(x) \;\lesseqgtr\; k' f_{\theta_0}(x) \quad \mu\text{-f.ü.}$$

mit $k' := g_{\theta,\theta_0}(c^*)^{-1} \in [0,\infty)$. Aus $k' = \infty$ folgt nämlich der Widerspruch

$$1 \;>\; \alpha \;=\; \mathbb{E}_{\theta_0}\varphi^*(X) \ge W_{\theta_0}(f_{\theta_0} > 0) \;=\; \int_{\{f_{\theta_0}>0\}} f_{\theta_0}\, d\mu \;=\; 1.$$

Außerdem erhalten wir ähnlich wie oben für μ-fast alle $x \in \mathfrak{X}$:

$$\begin{aligned} f_\theta(x) \;\lesseqgtr\; k' f_{\theta_0}(x) \quad &\Rightarrow \quad f_\theta(x), \vee f_{\theta_0}(x)) > 0 \text{ und } \frac{f_{\theta_0}(x)}{f_\theta(x)} \;\gtreqless\; \frac{1}{k'} \\ &\Rightarrow \quad g_{\theta,\theta_0}(T(x)) \;\gtreqless\; c^* \\ &\Rightarrow \quad \varphi^*(x) \;=\; \begin{cases} 1 & \\ 0 & . \end{cases} \end{aligned}$$

Für $\psi^* = 1 - \varphi^*$ liefert dies

$$\psi^*(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } f_\theta(x) \;\gtreqless\; k' f_{\theta_0}(x) \quad \mu\text{-f.ü.}$$

sowie $\mathbb{E}_{\theta_0}\psi^*(X) = 1-\alpha$. Nach Lemma 3.1(a) folgt deshalb

$$\mathbb{E}_\theta\psi^*(X) \;=\; \max_{\psi:\mathbb{E}_{\theta_0}\psi(X)\le 1-\alpha} \mathbb{E}_\theta\psi(X) \;=\; \max_{\psi:\mathbb{E}_{\theta_0}\psi(X)=1-\alpha} \mathbb{E}_\theta\psi(X)$$

und daraus weiter sowohl (1) für $\theta < \theta_0$ als auch (2), denn

$$\begin{aligned} R(\theta,\varphi^*) \;&=\; \mathbb{E}_\theta\varphi^*(X) \;=\; 1 - \mathbb{E}_\theta\psi^*(X) \\ &=\; 1 - \max_{\psi:\mathbb{E}_{\theta_0}\psi(X)=1-\alpha} \mathbb{E}_\theta\psi(X) \\ &=\; \min_{\psi:\mathbb{E}_{\theta_0}\psi(X)=1-\alpha} \mathbb{E}_\theta(1-\psi(X)) \;=\; \min_{\varphi:\mathbb{E}_{\theta_0}\varphi(X)=\alpha} \mathbb{E}_\theta\varphi(X) \;\le\; \alpha \end{aligned}$$

(wähle $\varphi \equiv \alpha$ für die letzte Ungleichung). □

Anmerkung 3.8. Der Satz hat gezeigt, dass für das einseitige Testproblem

$$H = \{\theta \in \Theta : \theta \leq \theta_0\} \quad \text{gegen} \quad K = \{\theta \in \Theta : \theta > \theta_0\}$$

bei Vorliegen eines monotonen Dichtequotienten in T durch φ^* gemäß (3.8) ein gleichmäßig bester Test zum Niveau α definiert wird, sofern man c^* als das α-Fraktil von T unter W_{θ_0} wählt, d.h.

$$c^* = c(W_{\theta_0}^T, \alpha) = \inf\{t \in \mathbb{R} : T > t) \leq \alpha\}.$$

Falls $W_{\theta_0}(T > c^*) = \alpha$, so erkennen wir aus dem Beweis von (a), dass ferner $\gamma^* = 0$ gesetzt werden kann, was insbesondere der Fall ist, wenn $W_{\theta_0}^T$ eine stetige Verteilung bildet, d.h. $W_{\theta_0}^T(\{t\}) = 0$ für alle $t \in \mathbb{R}$ gilt.

Anmerkung 3.9. Sofern die Verteilung $W_{\theta_0}^T$ zu den "populären" der Statistik zählt, lassen sich zur expliziten Bestimmung von c^*, wie bereits am Ende von Beispiel 3.3 (7. Schritt) bemerkt, Statistik-Software oder statistische Tafelwerke wie z.B. [15] oder [1] heranziehen. Andererseits erweist sich in manchen Fällen folgende Überlegung als nützlich: Ist $h : \mathbb{R} \to \mathbb{R}$ eine streng monoton wachsende stetige Funktion, so gilt offensichtlich

$$\varphi^*(x) \;=\; \begin{cases} 1, & \\ \gamma^*, & \text{falls } h(T(x)) \;\gtreqless\; h(c^*), \\ 0, & \end{cases}$$

wobei

$$\begin{aligned} h(c^*) &= h(\inf\{t \in \mathbb{R} : W_{\theta_0}(T > t) \leq \alpha\}) \\ &= \inf\{h(t) : t \in \mathbb{R}, W_{\theta_0}(h \circ T > h(t)) \leq \alpha\} \\ &= \inf\{y \in \mathbb{R} : W_{\theta_0}(h \circ T > y) \leq \alpha\} \\ &= c(W_{\theta_0}^{h \circ T}, \alpha). \end{aligned}$$

Sind die α-Fraktile von $W_{\theta_0}^T$ nicht vertafelt, so lässt sich macnhmal ein h finden, so dass dies für die α-Fraktile von $W_{\theta_0}^{h \circ T}$ der Fall ist.

Anmerkung 3.10. (Normalapproximation) Für große Stichprobenumfänge n liegen i.A. zur Bestimmung der α-Fraktile von $W_{\theta_0}^T$ keine Tafelwerke mehr vor. Es tritt dann aber oft die Situation auf, dass $W_{\theta_0}^{h \circ T}$ für eine geeignete streng monoton wachsende Funktion $h = h_{n,\theta_0}$ mit ausreichender Genauigkeit durch die Standard-Normalverteilung approximiert wird und somit ein Rückgriff auf deren α-Fraktile u_α erlaubt ist. Diese wiederum sind gut vertafelt, und als näherungsweise gleichmäßig bester Test φ^* ergibt sich dann

$$\varphi^*(x) \;=\; \begin{cases} 1, & \text{falls } h(T(x)) \;\lessgtr\; u_\alpha. \\ 0, & \end{cases}$$

[vgl. Beispiel 3.3, Schritt 7, dort: $h(t) = (n\theta_0(1-\theta_0))^{-1/2}(t - n\theta_0)$ und $T(x) = s_n$.]

Anmerkung 3.11. Betrachten wir das Testproblem

$$H = \{\theta \in \Theta : \theta \geq \theta_0\} \quad \text{gegen} \quad K = \{\theta \in \Theta : \theta < \theta_0\},$$

so ergibt sich analog zu Satz 3.7 als gleichmäßig bester Test zum Niveau α

$$\varphi^*(x) := \begin{cases} 1, & \\ \gamma^*, & \text{falls } T(x) \lessgtr c^* \\ 0, & \end{cases}$$

wobei nun c^*, γ^* aus

$$W_{\theta_0}(T < c^*) + \gamma^* W_{\theta_0}(T = c^*) = \alpha$$

zu bestimmen sind. Es gilt somit

$$\begin{aligned} c^* &= \sup\{t \in \mathbb{R} : W_{\theta_0}(T < t) \leq \alpha\} \\ &= \sup\{t \in \mathbb{R} : W_{\theta_0}(-T > -t) \leq \alpha\} \\ &= -\inf\{-t \in \mathbb{R} : W_{\theta_0}(-T > -t) \leq \alpha\} \\ &= -\inf\{s \in \mathbb{R} : W_{\theta_0}(-T > s) \leq \alpha\} \\ &= -c(W_{\theta_0}^{-T}, \alpha). \end{aligned}$$

Besitzt T unter W_{θ_0} eine symmetrische Verteilung, d.h. $W_{\theta_0}^T = W_{\theta_0}^{-T}$, so liefert dies weiter

$$c^* = -c(W_{\theta_0}^T, \alpha).$$

Da die Standard-Normalverteilung symmetrisch ist, folgt insbesondere bei einer Normalapproximation gemäß 3.10, d.h., falls $W_{\theta_0}^{h\circ T} \approx \mathit{Normal}(0,1)$ für ein $h = h_{n,\theta_0}$,

$$\varphi^*(x) = \begin{cases} 1, & \\ 0, & \text{falls } h(T(x)) \gtrless -u_\alpha. \end{cases}$$

Anmerkung 3.12. Sofern θ_0 ein (topologischer) Randpunkt von H ist, kann dieser in der Situation von 3.11 und Satz 3.7 auch jeweils K zugeordnet werden, ohne dass sich der gleichmäßig beste Test zum Niveau α ändert.

Beispiel 3.13 (Einseitiger Gauß-Test). Seien $X_1, \ldots, X_n$ stochastisch unabhängig und jeweils $\mathit{Normal}(\mu, \sigma^2)$-verteilt mit unbekanntem Mittelwert μ und bekannter Varianz $\sigma^2 > 0$. Für das Testproblem $H = (-\infty, \mu_0]$ gegen $K = (\mu_0, \infty)$ ergibt sich dann als gleichmäßig bester Test zum Niveau α

$$\varphi^*(x) = \begin{cases} 1, & \\ 0, & \text{falls } s_n = \sum_{k=1}^n x_k \gtrless c^* \end{cases}$$

mit c^*, so dass $\mathbb{E}_{\mu_0}\varphi^*(X) = \alpha$. Gemäß Beispiel 3.5 liegt nämlich ein monotoner Dichtequotient in $T(x) = s_n$ vor. Beachtet man nun, dass $(\sigma^2 n)^{-1/2}\sum_{j=1}^n (X_j - \mu_0)$ unter $\mathbb{P}_{\mu_0}$ eine $Normal(0,1)$-Verteilung besitzt [☞ Bemerkung 30.3(c) in [2]], so folgt weiter

$$\begin{aligned}\varphi^*(x) &= \begin{cases}1, & \text{falls } \frac{1}{n^{1/2}}\sum_{k=1}^n \frac{x_k-\mu_0}{\sigma} \overset{>}{\leqslant} u_\alpha\\ 0, & \end{cases}\\ &= \begin{cases}1, & \text{falls } s_n \overset{>}{\leqslant} n\mu_0 + \sigma u_\alpha n^{1/2}\\ 0, & \end{cases}\\ &= \begin{cases}1, & \text{falls } \bar{x}_n \overset{>}{\leqslant} \mu_0 + \sigma u_\alpha n^{-1/2}.\\ 0, & \end{cases}\end{aligned}$$

φ^* wird als *einseitiger Gauß-Test* bezeichnet.

Beispiel 3.14. Betrachten wir nun die Situation, dass μ bekannt und σ^2 unbekannt ist. Dann liegt gemäß Beispiel 3.5 ein monotoner Dichtequotient in $T(x) = \sum_{j=1}^n (x_j - \mu)^2$ vor. Für das Testproblem $H = (0, \sigma_0^2]$ gegen $K = (\sigma_0^2, \infty)$ erhalten wir deshalb als gleichmäßig besten Test zum Niveau α

$$\varphi^*(x) = \begin{cases}1, & \text{falls } \sum_{k=1}^n (x_k-\mu)^2 \overset{>}{\leqslant} c^*\\ 0, & \end{cases}$$

mit c^*, so dass $\mathbb{E}_{\sigma_0^2}\varphi^*(X) = \alpha$. Da die $\sigma_0^{-1}(X_j - \mu)$ unter $\mathbb{P}_{\sigma_0}$ standard-normalverteilt sind, besitzt die ZG

$$\sigma_0^{-2}T(X) = \sum_{j=1}^n \sigma_0^{-2}(X_j-\mu)^2$$

unter $\mathbb{P}_{\sigma_0}$ als Summe von n uanbhängigen quadrierten $Normal(0,1)$-verteilten ZG eine *Chi-Quadrat-Verteilung mit n Freiheitsgraden*, kurz χ_n^2-Verteilung. Diese entspricht der $\Gamma(n/2, 1/2)$-Verteilung und hat die λ-Dichte

$$\frac{d\chi_n^2}{d\lambda}(t) = \frac{1}{2^{n/2}\Gamma(n/2)} t^{(n/2)-1} e^{-t/2} \mathbf{1}_{(0,\infty)}(t),$$

[☞ Satz 31.7 in [2]]. Bezeichnen wir ihr α-Fraktil mit $\chi_{n,\alpha}^2$, so ergibt sich schließlich

$$\begin{aligned}\varphi^*(x) &= \begin{cases}1, & \text{falls } \sum_{k=1}^n \left(\frac{x_k-\mu}{\sigma_0}\right)^2 \overset{>}{\leqslant} \chi_{n,\alpha}^2\\ 0, & \end{cases}\\ &= \begin{cases}1, & \text{falls } \frac{1}{n}\sum_{k=1}^n (x_k-\mu)^2 \overset{>}{\leqslant} \frac{\sigma_0^2\chi_{n,\alpha}^2}{n}.\\ 0, & \end{cases}\end{aligned}$$

3.3 Das verallgemeinerte Neyman-Pearson-Lemma

In diesem Abschnitt zeigen wir eine Verallgemeinerung des Neyman-Pearson-Lemmas, die für die Behandlung zweiseitiger Testprobleme benötigt wird. Dabei geht es um folgende Problemstellung: Seien μ ein σ-endliches Maß auf dem Stichprobenraum $(\mathfrak{X}, \mathscr{A})$, $g_1, ..., g_{m+1}$ μ-integrierbare Funktionen sowie $\widetilde{\alpha} = (\alpha_1, ..., \alpha_m) \in \mathbb{R}^m$. Gesucht ist eine Testfunktion φ^* mit

$$\int \varphi^* g_j \, d\mu \leq \alpha_j \quad \text{für} \quad j = 1, ..., m,$$

$$\int \varphi^* g_{m+1} \, d\mu \;=\; \sup\left\{ \int \varphi g_{m+1} \, d\mu : \varphi \text{ Test mit } \int \varphi g_j \, d\mu \leq \alpha_j \text{ für } j = 1, ..., m \right\}.$$

Dabei ist beabsichtigt, dass die $g_1, ..., g_{m+1}$ nicht notwendig W-Dichten bilden.

Lemma 3.15. (Verallgemeinertes Neyman-Pearson-Lemma) *In der zuvor beschriebenen Situation sei Φ die Menge aller Tests auf $(\mathfrak{X}, \mathscr{A})$ und*

$$\Phi(\widetilde{\alpha}) \;:=\; \left\{ \varphi \in \Phi : \int \varphi g_j \, d\mu \leq \alpha_j \text{ für } j = 1, ..., m \right\},$$

$$\Phi[\widetilde{\alpha}] \;:=\; \left\{ \varphi \in \Phi : \int \varphi g_j \, d\mu = \alpha_j \text{ für } j = 1, ..., m \right\}.$$

Dann gilt

(a) ***(Hinreichende Bedingung)*** *Sei ψ ein Test mit:*
(1) $\psi \in \Phi[\widetilde{\alpha}]$.
(2) Es existieren $k_1, ..., k_m \in \mathbb{R}$ derart, dass

$$\psi(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } g_{m+1} \gtrless \sum_{j=1}^{m} k_j g_j(x) \quad \mu\text{-f.ü.} \tag{3.9}$$

Für einen solchen Test folgt

$$\int \psi g_{m+1} \, d\mu \;=\; \max\left\{ \int \varphi g_{m+1} \, d\mu : \varphi \in \Phi[\widetilde{\alpha}] \right\} \tag{3.10}$$

und im Fall $k_1, ..., k_m \geq 0$ sogar

$$\int \psi g_{m+1} \, d\mu \;=\; \max\left\{ \int \varphi g_{m+1} \, d\mu : \varphi \in \Phi(\widetilde{\alpha}) \right\}. \tag{3.11}$$

(b) ***(Existenz)*** *Liegt $\widetilde{\alpha}$ im Inneren von $Q_m := \{(\int \varphi g_1 d\mu, ..., \int \varphi g_m d\mu) : \varphi \in \Phi\}$, so existiert ein Test ψ, der die Voraussetzungen (a.1) und (a.2) erfüllt.*

(c) ***(Notwendige Bedingung)*** *Liegt $\widetilde{\alpha}$ im Inneren von Q_m, so ist jeder Test $\psi \in \Phi[\widetilde{\alpha}]$, der* (3.10) *erfüllt, von der Gestalt (3.9).*

Beweis. Wir zeigen nur Teil (a) und verweisen für (b) und (c) auf WITTING [21, Satz 2.67 auf S. 255f].

Erfülle ψ also (a.1) und (a.2). Dann folgt für jeden weiteren Test $\varphi \in \Phi[\widetilde{\alpha}]$ mit demselben Argument wie im Beweis von Lemma 3.1

$$\int (\psi - \varphi) \left(g_{m+1} - \sum_{j=1}^{m} k_j g_j \right) d\mu \;\geq\; 0$$

und daraus weiter

$$\int \psi g_{m+1} d\mu - \int \varphi g_{m+1} d\mu \;\geq\; \sum_{j=1}^{m} k_j \left(\int \psi g_j d\mu - \int \varphi g_j d\mu \right) \;=\; 0,$$

was (3.10) beweist. Sind sämtliche k_j nichtnegativ, so ist die obige Summe von Integraldifferenzen offenkundig sogar für jedes $\varphi \in \Phi(\widetilde{\alpha})$ nichtnegativ. Dies impliziert (3.11). □

Korollar 3.16. *Für $m \geq 1$ seien $W_1, \ldots, W_{m+1}$ W-Maße auf $(\mathfrak{X}, \mathscr{A})$ und W_{m+1} keine Linearkombination von $W_1, \ldots, W_m$. Dann existiert für jedes $\alpha \in (0,1)$ ein Test ψ mit $\int \psi \, dW_j = \alpha$ für $j = 1, \ldots, m$ und $\int \psi \, dW_{m+1} > \alpha$.*

Beweis. Sei μ ein σ-endliches dominierendes Maß für $W_1, \ldots, W_{m+1}$, etwa $\mu = \sum_{j=1}^{m+1} W_j$, und $f_j = dW_j/d\mu$ für $j = 1, \ldots, m+1$. Wir führen den Beweis per Induktion über m:

Induktionsanfang: Für $m = 1$ existiert nach dem Neyman-Pearson-Lemma 3.1 ein Test ψ der Form (3.1) mit $\int \psi \, dW_1 = \alpha$ und $\int \psi \, dW_2 = \sup\{\int \varphi dW_2 : \varphi \text{ Test mit } \int \varphi \, dW_1 \leq \alpha\}$. Wäre dieses Supremum gleich α, so müsste der Test $\varphi_\alpha \equiv \alpha$ gemäß 3.1(c) ebenfalls von der Form (3.1) für eine Konstante k sein, was weiter implizierte

$$\mu(f_2 > k f_1) + \mu(f_2 < k f_1) \;=\; 0, \quad \text{d.h.} \quad \mu(f_2 \neq k f_1) = 0.$$

Nun sind f_1, f_2 aber W-Dichten, so dass $\int k f_1 d\mu = k = \int f_2 d\mu = 1$, also $f_1 = f_2$ μ-f.ü. und damit $W_1 = W_2$ folgte, was nach Voraussetzung ausgeschlossen ist.

Induktionsschritt $m-1 \to m\,(\geq 2)$:

1. Fall: $W_1, \ldots, W_m$ sind linear abhängig, o.B.d.A. $W_m = \sum_{j=1}^{m-1} r_j W_j$ für geeignete $r_j \in \mathbb{R}$. Da W_m ein W-Maß bildet, folgt $\sum_{j=1}^{m-1} r_j = 1$. Nach Induktionsvoraussetzung existiert ein Test ψ mit $\int \psi dW_j = \alpha$ für $j = 1, \ldots, m-1$ und $\int \psi \, dW_{m+1} > \alpha$. Es folgt das Gewünschte, da $\int \psi \, dW_m = \sum_{j=1}^{m-1} r_j \int \psi \, dW_j = \alpha \sum_{j=1}^{m-1} r_j = \alpha$.

2. Fall: $W_1, \ldots, W_m$ sind linear unabhängig. Wir zeigen zuerst, dass $(\alpha, \ldots, \alpha)$ einen inneren Punkt von $Q_m = \{(\int \varphi \, dW_1, \ldots, \int \varphi \, dW_m) : \varphi \in \Phi\}$ bildet. Nach Indukti-

onsvoraussetzung existieren für jedes $i \in \{1,...,m\}$ Tests φ_i und ϕ_i derart, dass $\int \varphi_i\, dW_j = \alpha,\ \int \phi_i\, dW_j = 1-\alpha$ für alle $1 \le j \le m,\ j \ne i$, und $\int \varphi_i\, dW_i > \alpha$, $\int \phi_i\, dW_i > 1-\alpha$. Setzen wir also $\psi_i = 1-\phi_i$, so folgt offensichtlich $\int \psi_i\, dW_j = \alpha$ für alle $1 \le j \le m,\ j \ne i$, und $\int \psi_i\, dW_i < \alpha$. Damit gilt $\int \psi_i\, dW_i < \alpha < \int \varphi_i\, dW_i$ für jedes $i = 1,...,m$, und da Q_m konvex ist, muss $(\alpha,...,\alpha)$ in der Tat ein innerer Punkt sein.

Wäre nun $\int \varphi\ dW_{m+1} \le \alpha$ für jeden Test φ mit $\int \varphi\ dW_j = \alpha$ für $j = 1,...,m$, so folgte für $\varphi_\alpha \equiv \alpha$

$$\int \varphi_\alpha\ dW_{m+1} \ = \ \sup\left\{ \int \varphi\ dW_{m+1} : \varphi \in \Phi \text{ mit } \int \varphi\ dW_j = \alpha \text{ für } j = 1,...,m \right\}$$

und daraus weiter nach Lemma 3.15(c)

$$\varphi_\alpha(x) \ = \ \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } f_{m+1} \ \gtrless \ \sum_{j=1}^{m} k_j f_j(x) \quad \mu\text{-f.ü.}$$

für geeignete $k_1,...,k_m \in \mathbb{R}$. Aus dieser Tatsache ergäbe sich aber mit demselben Argument wie im Induktionsanfang

$$\mu\left(f_{m+1} \ne \sum_{j=1}^{m} k_j f_j \right) \ = \ 0, \quad \text{d.h.} \quad W_{m+1} \ = \ \sum_{j=1}^{m} k_j W_j,$$

was im Widerspruch zur Wahl von $W_1,...,W_{m+1}$ steht. □

Als einfache Folgerung für die im vorherigen Abschnitt behandelten einseitigen Testprobleme halten wir noch fest:

Korollar 3.17. *In der Situation von Satz 3.7 gilt für den dortigen optimalen Test φ^*: Die Gütefunktion $\mathbb{E}_\theta \varphi^*(X)$ ist streng monoton wachsend auf der Menge $\{\theta \in \Theta : 0 < \mathbb{E}_\theta \varphi^*(X) < 1\}$.*

Beweis. Seien $\theta_1, \theta_2 \in \Theta$ mit $\theta_1 < \theta_2$ und $0 < \beta := \mathbb{E}_{\theta_1} \varphi^*(X) < 1$. Eine analoge Argumentation wie zu Beginn von Satz 3.7 (ersetze dort θ_0 durch θ_1 und bestimme c^* mit β anstelle von α) zeigt, dass φ^* die Form

$$\varphi^*(x) \ = \ \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } f_{\theta_2}(x) \ \gtrless \ k f_{\theta_1}(x)$$

hat mit $k := g_{\theta_1,\theta_2}(c^*) \in [0,\infty)$. Nach dem Neyman-Pearson-Lemma 3.1 ist φ^* deshalb ein bester Test zum Niveau β für $H = \{\theta_1\}$ gegen $K = \{\theta_2\}$, und es folgt aus Korollar 3.16 (β innerer Punkt von $Q_1 = [0,1]$)

$$\mathbb{E}_{\theta_2} \varphi^*(X) \ > \ \beta \ = \ \mathbb{E}_{\theta_1} \varphi^*(X),$$

also die Behauptung. □

Ist $\mathscr{W} = (W_\theta)_{\theta\in\Theta}$ in der Situation von Satz 3.7 eine Exponentialfamilie, so gilt $0 < \mathbb{E}_\theta\varphi^*(X) < 1$ *für alle* $\theta \in \Theta$ und somit die strenge Monotonie der Gütefunktion von φ^* auf ganz Θ. Aus $\mathbb{E}_\theta\varphi^*(X) = 0$ oder $= 1$, also $\varphi^* = 0$ bzw. $= 1$ W_θ-f.s. für ein θ folgte nämlich dasselbe $\mathscr{W}$-f.s. wegen der paarweisen Äquivalenz aller W_θ. Dies wiederum implizierte aber $\mathbb{E}_\theta\varphi^*(X) = 0$ oder $= 1$ für alle $\theta \in \Theta$, was wegen $\mathbb{E}_{\theta_0}\varphi^*(X) = \alpha \in (0,1)$ unmöglich ist.

3.4 Gleichmäßig beste Tests für zweiseitige Hypothesen

Inhalt dieses Abschnitts bilden Testprobleme mit zweiseitigen Hypothesen der Form

$$H = \{\theta \in \Theta : \theta \le \theta_1\} \quad \text{gegen} \quad K = \{\theta \in \Theta : \theta_1 < \theta < \theta_2\} \tag{3.12}$$

für beliebige $\theta_1 < \theta_2$, wobei $\Theta \subset \mathbb{R}$. Zwar sind diese von geringerer Bedeutung als Testprobleme mit zweiseitigen Alternativen, auf die wir im Anschluss eingehen werden, bedürfen aber noch nicht der Einführung *unverfälschter Tests* und werden deshalb vorgezogen. Darüber hinaus vermitteln sie bereits eine Vorstellung, von welcher Form ein optimaler Test bei zweiseitigen Alternativen zu sein hat.

Im Folgenden liege ein statistisches Experiment $\mathscr{E} = (\mathfrak{X}, cA, (W_\theta)_{\theta\in\Theta})$ mit einparametriger Exponentialfamilie $(W_\theta)_{\theta\in\Theta}$ in *natürlicher Parametrisierung* vor, d.h. Θ sei der natürliche Parameterraum (Satz Satz 1.27 ein Intervall) und

$$\frac{dW_\theta}{d\nu}(x) \;=\; C(\theta)e^{\theta T(x)}h(x)$$

mit einem geeigneten dominierenden Maß ν für $(W_\theta)_{\theta\in\Theta}$. Obgleich eine solche Parametrisierung immer vorgenommen werden kann [☞ Abschnitt 1.5], bildet diese Voraussetzung in Bezug auf die zu betrachtenden Testprobleme eine Einschränkung, weil bei einer anderen Parametrisierung Hypothese und Alternative nicht mehr notwendig von der geforderten Form sein müssen. Andererseits entfällt dieser Einwand im Fall von ν-Dichten der Form $C(\theta)e^{Q(\theta)T(x)}h(x)$ mit streng monotoner Parameterfunktion $Q(\theta)$, eine Bedingung, die in allen typischen Anwendungsbeispielen erfüllt ist.

Satz 3.18. *Neben den zuvor gemachten Annahmen sei* $\alpha \in (0,1)$. *Dann gilt für das Testproblem (3.12):*

(a) *Es existieren* $c_1, c_2 \in \mathbb{R}$ *und* $\gamma_1, \gamma_2 \in [0,1]$, *so dass für* $j \in \{1,2\}$

$$W_{\theta_j}(c_1 < T < c_2) + \gamma_1(W_{\theta_j}(T = c_1) + \gamma_2 W_{\theta_j}(T = c_2) = \alpha. \tag{3.13}$$

(b) *Sei* $(c_1^*, c_2^*, \gamma_1^*, \gamma_2^*)$ *eine Lösung von* (3.13). *Definiert man*

$$\varphi^*(x) := \begin{cases} 1, & \text{falls} \quad \in (c_1^*, c_2^*) \\ \gamma_i^*, & \text{falls } T(x) = c_i^* \ (i=1,2), \\ 0, & \text{falls} \quad \notin [c_1^*, c_2^*] \end{cases} \tag{3.14}$$

so folgt:

(1) Für alle $\theta \notin \{\theta_1, \theta_2\}$ gilt

$$R(\theta, \varphi^*) = \min_{\varphi: \mathbb{E}_{\theta_1}\varphi(X) = \mathbb{E}_{\theta_2}\varphi(X) = \alpha} R(\theta, \varphi).$$

(2) φ^* ist gleichmäßig bester Test zum Niveau α für H gegen K.

Zum Beweis des Satzes bedarf es des folgenden einfachen Lemmas:

Lemma 3.19. *Es seien $b_1, b_2 \in \mathbb{R}$ mit $b_1 < 0 < b_2$. Dann gilt:*

(a) Für $a_1, a_2 \in (0, \infty)$ bildet die Menge $\{y \in \mathbb{R} : a_1 e^{b_1 y} + a_2 e^{b_2 y} < 1\}$ ein beschränktes offenes Intervall.

(b) Zu $c_1, c_2 \in \mathbb{R}$ mit $c_1 < c_2$ existieren stets $a_1, a_2 \in (0, \infty)$, so dass

$$\{y \in \mathbb{R} : a_1 e^{b_1 y} + a_2 e^{b_2 <} < 1\} = (c_1, c_2).$$

Beweis. Die Behauptungen ergeben sich sofort aus der Tatsache, dass die Funktion $f(y) := a_1 a e^{b_1 y} + a_2 e^{b_2 y}$ strikt konvex ist ($f''(y) = a_1 b_1^2 e^{b_1 y} + a_2 b_2^2 e^{b_2 y} > 0$) mit $\lim_{|y| \to \infty} f(y) = \infty$ und $f(0) = a_1 + a_2$ [☞ Abb. 3.1]. □

Beweis (von Satz 3.18). (b) Sei φ^* ein Test der Form (3.14). Für beliebiges $\theta \in K$ gilt $\theta_1 - \theta < 0 < \theta_2 - \theta$, so dass nach Lemma 3.19(b) positive a_1, a_2 existieren derart, dass $\{y \in \mathbb{R} : a_1 e^{(\theta_1 - \theta) y} + a_2 e^{(\theta_2 - \theta) y} < 1\} = (c_1^*, c_2^*)$ und somit

$$\begin{aligned} \varphi^*(x) &= \begin{cases} 1, & \text{falls } a_1 e^{(\theta_1 - \theta) T(x)} + a_2 e^{(\theta_2 - \theta) T(x)} \leqslant 1 \\ 0, & \end{cases} \\ &= \begin{cases} 1, & \text{falls } C(\theta) e^{\theta T(x)} \geqslant \sum_{j=1}^{2} k_j C(\theta_j) e^{\theta_j T(x)}, \\ 0, & \end{cases} \end{aligned}$$

wobei $k_j := a_j C(\theta) / C(\theta_j) > 0$ für $j = 1, 2$. Unter Anwendung von Lemma 3.15(a) folgt daher

$$\mathbb{E}_\theta \varphi^*(X) = \max_{\varphi: \mathbb{E}_{\theta_1}\varphi(X) = \mathbb{E}_{\theta_2}\varphi(X) \leq \alpha} \mathbb{E}_\theta \varphi(X) = \max_{\varphi: \mathbb{E}_{\theta_1}\varphi(X) = \mathbb{E}_{\theta_2}\varphi(X) = \alpha} \mathbb{E}_\theta \varphi(X),$$

d.h. die Behauptung (b.1) für $\theta \in (\theta_1, \theta_2)$.

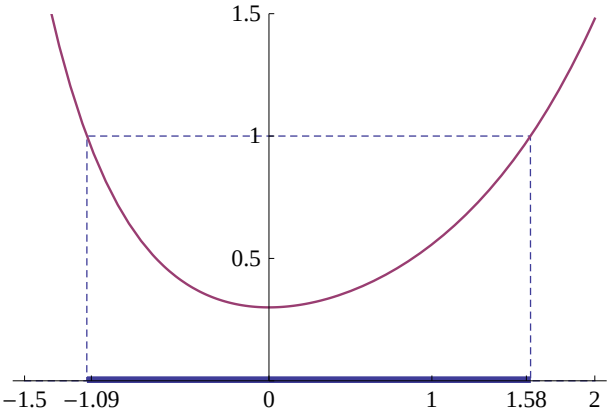


Abb. 3.1 Die Menge $\{y \in \mathbb{R} : \frac{1}{10}e^{-2y} + \frac{1}{5}e^{y} < 1\} = (-1.09, 1.58)$.

Zu zeigen bleibt (b.1) für $\theta \notin [\theta_1, \theta_2]$, und für (b.2) genügt der Nachweis, dass φ^* einen Test zum Niveau α bildet, d.h. $\varphi^* \in \Phi_\alpha$. Sei dazu $\theta < \theta_1$ beliebig. In diesem Fall liefert Lemma 3.19(b) wegen $\theta - \theta_1 < 0 < \theta_2 - \theta_1$ die Existenz von positiven a_1, a_2, so dass

$$\begin{aligned} \varphi^*(x) &= \begin{cases} 1, \\ 0, \end{cases} \quad \text{falls } a_1 e^{(\theta-\theta_1)T(x)} + a_2 e^{(\theta_2-\theta_1)T(x)} \lessgtr 1 \\ &= \begin{cases} 1, \\ 0, \end{cases} \quad \text{falls } C(\theta)e^{\theta T(x)} \lessgtr \sum_{j=1}^{2} k_j C(\theta_j) e^{\theta_j T(x)}, \end{aligned}$$

wobei hier $k_1 := C(\theta)/a_1 C(\theta_1)$ und $k_2 := -a_2 C(\theta)/a_1 C(\theta_2)$ gesetzt wird. Damit hat φ^* die Gestalt (3.9) in Lemma 3.15, nur mit vertauschten Ungleichheitszeichen. Der Übergang zu $\psi^* = 1 - \varphi^*$ liefert deshalb unter Beachtung von $\mathbb{E}_{\theta_1}\psi^*(X) = \mathbb{E}_{\theta_2}\psi^*(X) = 1 - \alpha$

$$\mathbb{E}_\theta \psi^*(X) = \max_{\psi : \mathbb{E}_{\theta_1}\psi(X) = \mathbb{E}_{\theta_2}\psi(X) = 1-\alpha} \mathbb{E}_\theta \psi(X) \geq 1 - \alpha$$

(wähle $\psi \equiv 1 - \alpha$), was offensichtlich sowohl

$$R(\theta, \varphi^*) = \min_{\varphi : \mathbb{E}_{\theta_1}\varphi(X) = \mathbb{E}_{\theta_2}\varphi(X) = \alpha} R(\theta, \varphi)$$

als auch $\mathbb{E}_\theta \varphi^*(X) \leq \alpha$ zeigt. Den Fall "$\theta > \theta_2$" behandelt man analog.

(a) Wie man mit Korollar 3.16 [☞ dessen Beweis, 2. Fall] leicht zeigen kann, bildet (α, α) einen inneren Punkt von $Q_2 = \{(\mathbb{E}_{\theta_1}\varphi(X), \mathbb{E}_{\theta_2}\varphi(X)) : \varphi \in \Phi\}$. Für ein beliebig gewähltes $\theta \in K$ existiert daher nach Lemma 3.15(b) ein Test ψ^* mit $\mathbb{E}_{\theta_1}\psi^*(X) = \mathbb{E}_{\theta_2}\psi^*(X) = \alpha$ der Form

$$\psi^*(x) \;=\; \begin{cases} 1, \\ 0, \end{cases} \quad \text{falls } C(\theta)e^{\theta T(x)} \;\gtrless\; \sum_{j=1}^{2} k_j C(\theta_j) e^{\theta_j T(x)} \quad (k_1, k_2 \in \mathbb{R})$$

$$= \begin{cases} 1, \\ 0, \end{cases} \quad \text{falls } a_1 e^{b_1 T(x)} + a_2 e^{b_2 T(x)} \;\lessgtr\; 1,$$

wobei $a_j := k_j C(\theta_j)/C(\theta)$ und $b_2 := \theta_1 - \theta < 0 < b_2 := \theta_2 - \theta$.

Es gilt $a_1, a_2 > 0$, denn: Aus $a_1, a_2 \leq 0$ folgte $\psi^* \equiv 1$ und damit $\mathbb{E}_{\theta_1} \psi^*(X) = \mathbb{E}_{\theta_2} \psi^*(X) = 1 \neq \alpha$. Wäre dagegen $a_1 > 0$ und $a_2 \leq 0$, so wäre $f(y) = a_1 e^{b_1 y} + a_2 e^{b_2 y}$ monoton fallend und somit ψ^* von der Form

$$\psi^*(x) \;=\; \begin{cases} 1, \\ 0, \end{cases} \quad \text{falls } T(x) \;\gtrless\; c$$

für eine geeignete Konstante c. Nach 3.5 besitzt $(W_\theta)_{\theta \in \Theta}$ einen monotonen Dichtequotienten in T, so dass ψ^* dann gemäß Satz 3.7 einen gleichmäßig besten Test zum Niveau α für $\{\theta \in \Theta : \theta \leq \theta_1\}$ gegen $\{\theta \in \Theta : \theta > \theta_1\}$ definierte und gemäß Korollar 3.17 eine streng monotone Gütefunktion auf $\{\theta : 0 < \mathbb{E}_\theta \psi^*(X) < 1\}$ besäße, was wegen $0 < \mathbb{E}_{\theta_1} \psi^*(X) = \mathbb{E}_{\theta_2} \psi^*(X) = \alpha < 1$ nicht sein kann. Auf dieselbe Weise führt man die Annahme $a_1 \leq 0$ und $a_2 > 0$ zum Widerspruch.

Aus $a_1, a_2 > 0$ folgt nun nach Lemma 3.19(a) die Existenz reeller c_1, c_2 mit $c_1 < c_2$, so dass

$$\psi^*(x) \;=\; \begin{cases} 1, \\ 0, \end{cases} \quad \text{falls } T(x) \;\begin{matrix} \in \\ \notin \end{matrix}\; \begin{matrix} (c_1, c_2) \\ [c_1, c_2] \end{matrix}.$$

Sei $\nu^* = h\nu$ das zu $(W_\theta)_{\theta \in \Theta}$ äquivalente Maß . Definieren wir dann für $j = 1, 2$

$$\gamma_j \;:=\; \begin{cases} \frac{1}{\nu^*(T = c_j)} \int_{\{T = c_j\}} \psi^*(x) \nu^*(dx), & \text{falls } \nu^*(T = c_j) > 0, \\ 0, & \text{falls } \nu^*(T = c_j) = 0, \end{cases}$$

so rechnet man leicht nach, dass $(c_1, c_2, \gamma_1, \gamma_2)$ die Gleichungen (3.13) erfüllt. □

Anmerkung 3.20. Lässt man die Alternative in K in (3.12) auf einen Punkt $\theta_0 \in \Theta$ zusammenschrumpfen, so ergibt sich das Testproblem $H = \{\theta \in \Theta : \theta \neq \theta_0\}$ gegen $K = \{\theta_0\}$. Hierfür ist es jedoch nicht möglich, einen sinnvollen Test zum Niveau α anzugeben. Nach Satz 1.27 hat nämlich jeder Test φ eine stetige Gütefunktion, so dass aus $\mathbb{E}_\theta \varphi(X) \leq \alpha$ für alle $\theta \neq \theta_0$ bereits $\mathbb{E}_{\theta_0} \varphi(X) \leq \alpha$ folgt, da θ_0 kein isolierter Punkt des Intervalls Θ ist. Letzteres impliziert aber, dass der unsinnige Test $\varphi_\alpha \equiv \alpha$, der unabhängig von der Beobachtung immer mit Wahrscheinlichkeit α zugunsten der Alternative randomisiert, einen gleichmäßig besten Test zum Niveau α für H gegen K definiert.

3.5 Gleichmäßig beste unverfälschte Tests für zweiseitige Alternativen

Vertauschen wir die Rollen von H und K im vorherigen Abschnitt, wobei die Randpunkte θ_1 und θ_2 wiederum zur Hypothese H gehören, so gelangen wir zu den beiden Testproblemen

$$H = \{\theta \in \Theta : \theta_1 \leq \theta \leq \theta_2\} \quad \text{gegen} \quad K = \{\theta \in \Theta : \theta < \theta_1 \text{ oder } \theta > \theta_0\},$$
$$H = \{\theta_0\} \quad \text{gegen} \quad K = \{\theta \in \Theta : \theta \neq \theta_0\}$$

für $\theta_0, \theta_1, \theta_2 \in \Theta$ mit $\theta_1 < \theta_2$. Bevor wir jedoch für diese Probleme gleichmäßig beste Tests zum Niveau α herleiten können, bedarf es einer weiteren Einschränkung der zugelassenen Testfunktionen, wie am folgenden Beispiel erläutert wird.

Beispiel 3.21. Man nehme an, eine etablierte Theorie postuliert, dass der Wert einer bestimmten physikalischen Konstante gleich θ_0 ist. Ein Wissenschaftler glaubt jedoch gute Gründe dafür zu haben, diese Theorie in der vorliegenden Form anzuzweifeln und entwirft deshalb einen Versuchsaufbau zum Messen der physikalischen Konstante. Er führt n Messungen durch, und seine Kenntnisse über den Versuchsaufbau führen ihn zu der Annahme, dass er stochastisch unabhängige, jeweils $Normal(\theta, \sigma^2)$-verteilte ZG mit unbekanntem Mittelwert $\theta \in \mathbb{R}$ und bekannter Varianz $\sigma^2 > 0$ beobachtet. Der Physiker betrachtet nun das Testproblem

$$H = \{\theta_0\} \quad \text{gegen} \quad K = \{\theta \in \mathbb{R} : \theta \neq \theta_0\}$$

und sucht einen gleichmäßig besten Test zum Niveau α für H gegen K. Die folgenden Überlegungen zeigen, dass er vergeblich sucht, weil ein solcher Test *nicht* existiert.

Das Gegenteil annehmend, sei φ^* ein gleichmäßig bester Test zum Niveau $\alpha \in (0,1)$ für H gegen K. Dann gilt dasselbe auch für das Testproblem H gegen $K' = (\theta_0, \infty)$, denn die Klasse der Tests zum Niveau α stimmt für beide Probleme überein. Analoge Überlegungen wie in Beispiel 3.3 unter Benutzung des Neyman-Pearson-Lemmas zeigen, dass φ^* dann die Form

$$\varphi^*(x) = \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } \sum_{j=1}^{n} x_j \gtrless c \quad \lambda^n\text{-f.ü.}$$

für ein $c \in \mathbb{R}$ hat und folglich nach Satz 3.7 bereits einen gleichmäßig besten Test zum Niveau α für $H' = (-\infty, \theta_0]$ gegen K' definiert (beachte $\lambda^n(\sum_{j=1}^n x_j = c) = 0$). Nach Korollar 3.17 besitzt φ^* also eine streng monotone Gütefunktion auf $\{\theta \in \mathbb{R} : 0 < \mathbb{E}_\theta \varphi^*(X) < 1\} = \mathbb{R}$, was man im hier vorliegenden Spezialfall auch direkt nachrechnen kann. Es gilt demnach

$$\mathbb{E}_\theta \varphi^*(X) < \mathbb{E}_{\theta_0} \varphi^*(X) \leq \alpha$$

für alle $\theta < \theta_0$. Andererseits hat φ^* als gleichmäßig bester Test zum Niveau α für H gegen K für jedes $\theta \neq \theta_0$ mindestens die gleiche Güte wie $\varphi_\alpha \equiv \alpha$, was insbesondere

$$\mathbb{E}_\theta \varphi^*(X) \;\geq\; \mathbb{E}_\theta \varphi_\alpha(X) \;=\; \alpha$$

für alle $\theta < \theta_0$ bedeutet und im Widerspruch zur vorherigen Ungleichung steht.

Die beschriebene Argumentation gilt generell für einparametrige Exponentialfamilien und in diesem Fall ebenso für das Testproblem

$$H = \{\theta \in \Theta : \theta_1 \leq \theta \leq \theta_2\} \quad \text{gegen} \quad K = \{\theta \in \Theta : \theta < \theta_1 \text{ oder } \theta > \theta_2\}.$$

Ohne darauf näher einzugehen, sei aber erwähnt, dass es auch Verteilungsfamilien gibt, z.B. die Klasse $(Unif(0,\theta)^n)_{\theta>0}$, für die ein gleichmäßig bester Test zum Niveau α für "$\theta = \theta_0$" gegen "$\theta \neq \theta_0$" existiert.

Kehren wir noch einmal zurück zum *einseitigen Gauß-Test* für $H = (-\infty, \theta_0]$ gegen $k = (\theta_0, \infty)$, definiert durch

$$\widehat{\varphi}(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } s_n \begin{matrix} > \\ \leqslant \end{matrix} n\theta_0 + \sigma u_\alpha n^{1/2}$$

[☞ Beispiel 3.13]. Dieser lehnt die Hypothese ab, wenn das Stichprobenmittel $\bar{x}_n$ *"signifikant größer"* als θ_0 ist.

Im Fall $H = \{\theta_0\}$ gegen $K = \{\theta \neq \theta_0\}$ liegt es nahe, die Hypothese abzulehnen, wenn $\bar{x}_n$ *"signifikant größer"* oder *"signifikant kleiner"* als θ_0 ist, was zu dem Test

$$\varphi^*(x) \;=\; \begin{cases} 1, & \text{falls } \bar{x}_n < \theta_0 - \sigma c_1 n^{-1/2} \text{ oder } \bar{x}_n > \theta_0 + \sigma c_2 n^{-1/2}, \\ 0, & \text{falls } \theta_0 - \sigma c_1 n^{-1/2} \leq \bar{x}_n \leq \theta_0 + \sigma c_2 n^{-1/2} \end{cases}$$

führt, wobei $c_1, c_2 > 0$ so gewählt werden, dass $\mathbb{E}_{\theta_0} \varphi^*(X) = \alpha$ gilt. Durch die Symmetrie der Normalverteilung um ihren Mittelwert wird eine symmetrische Wahl der rechten und linken Grenze sinnvoll, d.h. $c_1 = c_2 = c$. Es ergibt sich dann

$$\begin{aligned} \alpha \;&=\; \mathbb{E}_{\theta_0} \varphi^*(X) \\ &=\; \mathbb{P}_{\theta_0}\left(\frac{1}{\sigma n^{1/2}} \sum_{j=1}^n (X_j - \theta_0) < -c \right) \;+\; \mathbb{P}_{\theta_0}\left(\frac{1}{\sigma n^{1/2}} \sum_{j=1}^n (X_j - \theta_0) > c \right) \\ &=\; Normal(0,1)([-c,c]^c) \;=\; 2\,Normal(0,1)((c,\infty)), \end{aligned}$$

also $Normal(0,1)((c,\infty)) = \alpha/2$. Mit anderen Worten, c ist das $\alpha/2$-Fraktil $u_{\alpha/2}$ der Standard-Normalverteilung. Als plausibles Testverfahren ergibt sich folglich

$$\varphi^*(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } |\bar{x}_n - \theta_0| \begin{matrix} > \\ \leqslant \end{matrix} \sigma u_{\alpha/2} n^{-1/2},$$

genannt *zweiseitiger Gauß-Test*. Für dessen Gütefunktion $\mathbb{E}_\theta \varphi^*(X)$ für $\theta \neq \theta_0$ notieren wir zum Abschluss

$$\begin{aligned}
&1 - \mathbb{E}_\theta \varphi^*(X) \\
&= \mathbb{P}_\theta \left(-u_{\alpha/2} \leq \frac{1}{\sigma n^{1/2}} \sum_{j=1}^n (X_j - \theta_0) \leq u_{\alpha/2} \right) \\
&= \mathbb{P}_\theta \left(-u_{\alpha/2} + \frac{n^{1/2}(\theta_0 - \theta)}{\sigma} \leq \frac{1}{\sigma n^{1/2}} \sum_{j=1}^n (X_j - \theta) \leq u_{\alpha/2} + \frac{n^{1/2}(\theta_0 - \theta)}{\sigma} \right) \\
&= \mathit{Normal}(0,1) \left(\left[-u_{\alpha/2} + \frac{n^{1/2}(\theta_0 - \theta)}{\sigma}, u_{\alpha/2} + \frac{n^{1/2}(\theta_0 - \theta)}{\sigma} \right] \right) \\
&< \mathit{Normal}(0,1)([-u_{\alpha/2}, u_{\alpha/2}]) \;=\; 1 - \alpha.
\end{aligned}$$

Für die Ungleichung in der vorletzten Zeile beachte man, dass die Funktion $x \mapsto \mathit{Normal}(0,1)([-u_{\alpha/2} + x, u_{\alpha/2} + x])$ in $x = 0$ ihr eindeutiges Maximum besitzt, wie mit Mitteln der Analysis leicht gezeigt werden kann. Es folgt also für den zweiseitigen Gauß-Test

$$\mathbb{E}_\theta \varphi^*(X) > \alpha \tag{3.15}$$

für alle $\theta \neq \theta_0$, eine Bedingung, die der einseitige Gauß-Test nicht erfüllt.

Es überrascht nun nicht mehr, dass es gerade die Klasse der Tests mit der Eigenschaft (3.15) (ersetze dort "$>$" durch "$\geq$") ist, auf die in Testproblemen mit zweiseitigen Alternativen die Suche nach einem gleichmäßig besten Test eingeschränkt wird. Wir definieren:

Definition 3.22. Gegeben sei ein Testproblem H gegen K. Ein Test φ heißt *unverfälscht zum Niveau* α, falls $\mathbb{E}_\theta \varphi(X) \leq \alpha$ für alle $\theta \in H$ (d.h. $\varphi \in \Phi_\alpha$) und $\mathbb{E}_\theta \varphi(X) \geq \alpha$ für alle $\theta \in K$ (d.h. $R(\theta, \varphi) \leq 1 - \alpha$ für alle $\theta \in K$) gilt. Es bezeichne Φ_α^u die Gesamtheit aller unverfälschter Tests zum Niveau α. Ein Test φ^* heißt *gleichmäßig bester unverfälschter Test zum Niveau* α, falls $\varphi^* \in \Phi_\alpha^u$ und

$$\mathbb{E}_\theta \varphi^*(X) \;=\; \max_{\varphi \in \Phi_\alpha^u} \mathbb{E}_\theta \varphi(X)$$

für alle $\theta \in K$.

Wir werden schon bald sehen [Beispiel 3.28], dass der zweiseitige Gauß-Test nicht nur vernünftig, sondern bereits der gleichmäßig beste unverfälschte Test zum Niveau α für das Testproblem $H = \{\theta_0\}$ gegen $K = \{\theta \neq \theta_0\}$ ist.

Wie in Abschnitt 3.4 liege im Folgenden ein Experiment $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta \in \Theta})$ mit einparametriger Exponentialfamilie $(W_\theta)_{\theta \in \Theta}$ in natürlicher Parametrisierung vor. Wir betrachten zunächst wieder das Testproblem

$$H = \{\theta_0\} \quad \text{gegen} \quad K = \{\theta \in \Theta : \theta \neq \theta_0\}, \tag{3.16}$$

wobei θ_0 ein *innerer Punkt* von Θ sei, d.h. $(\theta_0 - \varepsilon, \theta_0 + \varepsilon) \subset \Theta$ für ein $\varepsilon > 0$. Gemäß Satz 1.27 ist dann die Gütefunktion $\mathbb{E}_\theta \varphi(X)$ für jeden Test φ in $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ beliebig oft differenzierbar, was zu folgendem Lemma führt.

Lemma 3.23. *Unter den obigen Annahmen gilt*

$$\Phi_\alpha^u \subset \Psi_\alpha := \{\varphi \in \Phi : \mathbb{E}_{\theta_0} \varphi(X) = \alpha \text{ und } \mathbb{E}_{\theta_0} \varphi(X) T(X) = \alpha \mathbb{E}_{\theta_0} T(X)\}. \tag{3.17}$$

Beweis. Sei $\varphi \in \Phi_\alpha^u$ mit Gütefunktion $\beta_\varphi(\theta) = \mathbb{E}_\theta \varphi$. Dann folgt $\mathbb{E}_{\theta_0} \varphi(X) = \alpha$ sofort aus der Stetigkeit von β_φ zusammen mit der Unverfälschtheit, denn θ_0 ist ein innerer Punkt von Θ. Ferner hat diese Funktion in θ_0 ein relatives Minimum, so dass unter Benutzung von Satz 1.27(c)

$$\begin{aligned} 0 \;=\; \beta_\varphi'(\theta_0) \;&=\; \frac{d}{d\theta} C(\theta) \int \varphi(X) e^{\theta T(x)} \nu^*(dx)\big|_{\theta=\theta_0} \\ &=\; \mathbb{E}_{\theta_0} \varphi(X) T(X) \;+\; \frac{C'(\theta_0)}{C(\theta_0)} \mathbb{E}_{\theta_0} \varphi(X) \end{aligned}$$

und daraus (3.17) folgt, da $1 = C(\theta) \int e^{\theta T(x)} \nu^*(dx)$ leicht $\frac{C'(\theta_0)}{C(\theta_0)} = -\mathbb{E}_{\theta_0} T(X)$ impliziert. □

Satz 3.24. *In der zuvor beschriebenen Situation sei $\alpha \in (0,1)$. Dann gilt für das Testproblem* (3.16)*:*

(a) Es existieren $c_1, c_2 \in \mathbb{R}$ und $\gamma_1, \gamma_2 \in [0,1]$, so dass

$$\begin{aligned} &W_{\theta_0}(T \notin [c_1, c_2]) \;+\; \gamma_1 W_{\theta_0}(T = c_1) + \gamma_2 W_{\theta_0}(T = c_2) \;=\; \alpha, \\ &\int_{[c_1,c_2]^c} x W_{\theta_0}^T(dx) \;+\; \sum_{j=1}^{2} \gamma_j c_j W_{\theta_0}(T = c_j) \;=\; \alpha \mathbb{E}_{\theta_0} T(X). \end{aligned} \tag{3.18}$$

(b) Bildet $(c_1^*, c_2^*, \gamma_1^*, \gamma_2^*)$ eine Lösung von (3.18), und definiert man

$$\varphi^*(x) \;:=\; \begin{cases} 1, & \text{falls} \quad \notin [c_1^*, c_2^*] \\ \gamma_i^*, & \text{falls } T(x) = c_i^* \quad (i = 1,2), \\ 0, & \text{falls} \quad \in (c_1^*, c_2^*) \end{cases} \tag{3.19}$$

so ist φ^* ein gleichmäßig bester unverfälschter Test zum Niveau α für H gegen K.

Als Hilfsresultat für den Beweis des Satzes benötigen wir [vgl. Lemma 3.19]

Lemma 3.25. *Für jedes $b \in \mathbb{R}\setminus\{0\}$ gilt:*

(a) Für $a_1, a_2 \in \mathbb{R}$ mit $a_2 b > 0$ ist die Menge $\{y \in \mathbb{R} : a_1 + a_2 y > e^{by}\}$ ein beschränktes offenes Intervall.

(b) Zu $c_1, c_2 \in \mathbb{R}$ mit $c_1 < c_2$ existieren stets $a_1, a_2 \in \mathbb{R}$, so dass $a_2 b > 0$ und

$$\{y \in \mathbb{R} : a_1 + a_2 y > e^{by}\} \;=\; (c_1, c_2).$$

Beweis. Wegen $a_2 b > 0$ besitzen a_2 und b dasselbe Vorzeichen. Die Funktion $f(y) := e^{by} - a_1 - a_2 y$ erfüllt deshalb in jedem Fall $\lim_{|y|\to\infty} f(y) = \infty$. Da f ferner strikt konvex ist ($f''(y) = b^2 e^{by} > 0$) und $f(0) = 1 - a_1$, ergeben sich die Behauptungen wie in Lemma 3.19. □

Beweis (von Satz 3.24). Der Beweis verläuft in seinen Grundzügen analog zu dem von Satz 3.18, wobei wir wieder zuerst Teil (b) zeigen wollen.

(b) Sei also φ^* ein Test der Gestalt (3.19) und $\theta \in K$ beliebig, d.h. $\theta - \theta_0 \neq 0$. Aufgrund der Wahl von $c_1^*, c_2^*, \gamma_1^*, \gamma_2^*$ folgt $\varphi^* \in \Psi_\alpha$. Nach Lemma 3.25(b) existieren $a_1, a_2 \in \mathbb{R}$, so dass $a_2(\theta - \theta_0) > 0$ und

$$\begin{aligned}\varphi^*(x) \;&=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } a_1 + a_2 T(x) \;\lessgtr\; e^{(\theta-\theta_0)T(x)} \\ &=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } C(\theta) e^{(\theta T(x)} \;\gtrless\; k_1 C(\theta_0) e^{\theta_0 T(x)} + k_2 C(\theta_0) e^{\theta_0 T(x)} T(x),\end{aligned}$$

wobei $k_j := a_j C(\theta)/C(\theta_0)$ für $j = 1, 2$. Wir wenden nun wieder das verallgemeinerte Neyman-Pearson-Lemma 3.15 an, und zwar mit $m = 2$, $\widetilde{\alpha} = (\alpha, \alpha \mathbb{E}_{\theta_0} T(X))$ sowie

$$\begin{aligned} g_1(x) \;&:=\; C(\theta_0) e^{\theta_0 T(x)}, \quad g_2(x) \;:=\; C(\theta_0) e^{\theta_0 T(x)} T(x) \\ \text{und} \quad g_3(x) \;&:=\; C(\theta) e^{\theta T(x)}, \end{aligned} \tag{3.20}$$

was $\Phi[\widetilde{\alpha}] = \Psi_\alpha$ ergibt. Es folgt

$$\mathbb{E}_\theta \varphi^*(X) \;=\; \max_{\varphi \in \Psi_\alpha} \mathbb{E}_\theta \varphi(X)$$

und daraus insbesondere $\varphi^* \in \Phi_\alpha^u$ wegen $\varphi_\alpha \equiv \Psi_\alpha$. Unter Hinweis auf Lemma 3.23 zeigt dies weiter $\max_{\varphi\in\Psi_\alpha} \mathbb{E}_\theta \varphi(X) = \max_{\varphi\in\Phi_\alpha^u} \mathbb{E}_\theta \varphi(X)$, was den Beweis von Teil (b) abschließt.

(a) Wir zeigen zuerst, dass $(\alpha, \alpha \mathbb{E}_{\theta_0} T(X))$ einen inneren Punkt der offensichtlich konvexen Menge

$$Q_2 \;=\{(\mathbb{E}_{\theta_0}\varphi(X),\mathbb{E}_{\theta_0}\varphi(X)T(X)) : \varphi\in\Phi\}$$

bildet. Seien dazu φ_1 und φ_2 die gleichmäßig besten Tests zum Niveau α für die Testprobleme

$$H:\theta\le\theta_0 \text{ gegen } K:\theta>\theta_0 \quad \text{bzw.} \quad H:\theta\ge\theta_0 \text{ gegen } K:\theta<\theta_0$$

und $\beta_{\varphi_1},\beta_{\varphi_2}$ ihre gemäß Korollar 3.17 und dem anschließenden Hinweis auf ganz Θ streng monotonen Gütefunktionen. Dann gilt $\beta_{\varphi_1}(\theta_0)=\beta_{\varphi_2}(\theta_0)=\alpha$ und aufgrund der strengen Monotonie ferner $\beta'_{\varphi_2}(\theta_0)<0<\beta'_{\varphi_1}(\theta_0)$. Wegen

$$\beta'_{\varphi_j}(\theta_0) \;=\; \mathbb{E}_{\theta_0}\varphi_j(X)T(X)-\mathbb{E}_{\theta_0}\varphi_j(X)\,\mathbb{E}_{\theta_0}T(X) \;=\; \mathbb{E}_{\theta_0}\varphi_j(X)T(X)-\alpha\,\mathbb{E}_{\theta_0}T(X)$$

für $j=1,2$ [☞ Beweis von Lemma 3.23] liefert dies

$$\mathbb{E}_{\theta_0}\varphi_2(X)T(X) \;<\; \alpha\,\mathbb{E}_{\theta_0}T(X) \;<\; \mathbb{E}_{\theta_0}\varphi_1(X)T(X).$$

Wie man vermöge Abb. 3.2 erkennt, folgt nun $(\alpha,\alpha\mathbb{E}_{\theta_0}T(X))\in\overset{\circ}{Q}_2$ aus der Konvexität von Q_2 und

$$(0,0),\ (\alpha,\mathbb{E}_{\theta_0}\varphi_1(X)T(X)),\ (1,\mathbb{E}_{\theta_0}T(X)),\ (\alpha,\mathbb{E}_{\theta_0}\varphi_2(X)T(X)) \;\in\; Q_2. \qquad (3.21)$$

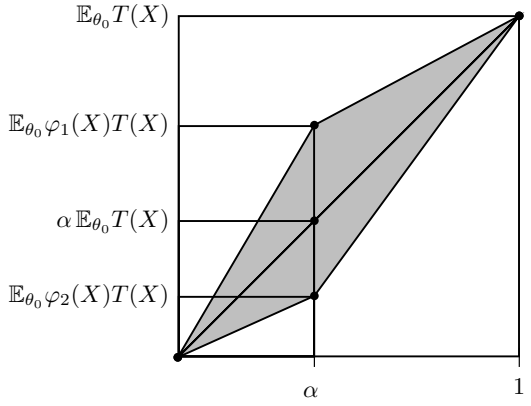


Abb. 3.2 Die von den 4 Punkten aus (3.21) aufgespannte konvexe Teilmenge von Q_2.

Seien als nächstes $\theta\in K$ mit $\theta-\theta_0>0$ und g_1,g_2,g_3 gemäß (3.20) definiert. Wie schon im Beweis von Satz 3.18(a) liefert hier Lemma 3.15(b) die Existenz eines Tests $\psi^*\in\Psi_\alpha$ der Gestalt

$$\psi^*(x) \;=\; \begin{cases}1, & \\ 0, & \end{cases} \quad \text{falls } C(\theta)e^{\theta T(x)} \;\gtrless\; k_1C(\theta_0)e^{\theta_0T(x)}+k_2C(\theta_0)e^{\theta_0T(x)}T(x)$$

$$= \begin{cases} 1, \\ 0, \end{cases} \quad \text{falls falls } a_1 + a_2 T(x) \lessgtr e^{(\theta-\theta_0)T(x)},$$

wobei $a_j := k_j C(\theta_0)/C(\theta)$ für $j = 1,2$. Ein Widerspruchsargument wie im Beweis von Satz 3.18(a) (Fall "$a_1 > 0$, $a_2 \leq 0$") liefert hier $a_2 > 0$, also $a_2(\theta - \theta_0) > 0$, und dann vermöge Lemma 3.25(a) die Existenz von c_1, c_2 mit $c_1 < c_2$, so dass

$$\psi^*(x) = \begin{cases} 1, & \text{falls } T(x) \notin [c_1, c_2] \\ 0, & \text{falls } T(x) \in (c_1, c_2) \end{cases}.$$

Setzen wir schließlich für $j = 1,2$

$$\gamma_j := \begin{cases} \frac{1}{\nu^*(T=c_j)} \int_{\{T=c_j\}} \psi^*(x) \nu^*(dx), & \text{falls } \nu^*(T=c_j) > 0, \\ 0, & \text{falls } \nu^*(T=c_j) = 0, \end{cases}$$

so rechnet man leicht nach, dass $c_1, c_2, \gamma_1, \gamma_2$ die Gleichungen in (3.18) erfüllen. □

Anmerkung 3.26. Wie der vorherige Satz aussagt, werden die kritischen Konstanten $c_1^*, c_2^*, \gamma_1^*, \gamma_2^*$ für den gleichmäßig besten unverfälschten Test φ^* in (3.19) aus den Gleichungen (3.18)) bestimmt, was offensichtlich mühsamer ist als im Fall einseitiger Testprobleme. Doch auch für diese Aufgabe stehen die schon füher genannten Tafelwerke [15] oder [1]) zur Verfügung. Besitzt $W_{\theta_0}^T$ eine stetige Verteilung, so vereinfacht sich (3.18) zu

$$\begin{aligned} W_{\theta_0}(T < c_1^*) + W_{\theta_0}(T > c_2^*) &= \alpha, \\ \int_{\{T<c_1^*\}} T \, dW_{\theta_0} + \int_{\{T>c_2^*\}} T \, dW_{\theta_0} &= \alpha \, e_{\theta_0} T(X). \end{aligned}$$

Eine weitergehende Vereinfachung ergibt sich, falls $W_{\theta_0}^{T-a}$ für ein $a \in \mathbb{R}$ symmetrisch ist, d.h., wenn $W_{\theta_0}^{T-a} = W_{\theta_0}^{-(T-a)}$. Für diesen Fall notieren wir:

Korollar 3.27. *In der Situation von Satz 3.24 sei $W_{\theta_0}^{T-a}$ symmetrisch für ein $a \in \mathbb{R}$. Für $c \in \mathbb{R}$ und $\gamma^* \in [0,1]$ mit $W_{\theta_0}(T - a > c^*) + \gamma^* W_{\theta_0}(T - a = c^*) = \alpha/2$ sei*

$$\varphi^*(x) := \begin{cases} 1, \\ \gamma^*, & \textit{falls } |T(x) - a| \gtreqless c^* \\ 0, \end{cases} \tag{3.22}$$

Dann ist φ^* gleichmäßig bester unverfälschter Test zum Niveau α für H gegen K.

Beweis. Man hat lediglich nachzurechnen, dass φ^* in (3.22) von der Gestalt (3.19) mit $c_1^* = a - c^*$, $c_2^* = a + c^*$, $\gamma_1^* = \gamma_2^* = \gamma^*$ ist und ein Element von Ψ_α bildet, also

die Gleichungen (3.18) erfüllt. Wir überlassen dies dem Leser zur Übung und geben hierzu noch den Hinweis, dass $\mathbb{E}_{\theta_0}\varphi(X)(T(X)-a)=0$ gilt. □

Beispiel 3.28 (Zweiseitiger Gauß-Test). Kehren wir noch einmal zurück zu der Situation in Beispiel 3.21 mit n unabhängigen, jeweils $Normal(\theta,\sigma^2)$-verteilten ZG $X_1,...,X_n$, auf deren Basis $H=\{\theta_0\}$ gegen $K=\{\theta\neq\theta_0\}$ getestet werden soll, wobei $\sigma^2>0$ als bekannt vorausgesetzt wird. Es liegt eine Exponentialfamilie in $T(x)=\sum_{j=1}^n x_j$ vor, und die Verteilung

$$W_{\theta_0}^{T-n\theta_0} \;=\; \mathbb{P}_{\theta_0}^{\sum_{j=1}^n (X_j-\theta_0)} \;=\; Normal(0,n\sigma^2)$$

ist symmetrisch. Wir erhalten daher als gleichmäßig besten unverfälschten Test zum Niveau α

$$\varphi^*(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls} \left|\sum_{j=1}^n (x_j-\theta_0)\right| \begin{matrix} > \\ \leqslant \end{matrix} c(Normal(0,n\sigma^2),\alpha/2)$$

Eine einfache Umformung liefert weiter

$$\begin{aligned} \varphi^*(x) \;&=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } \frac{1}{\sigma n^{1/2}} \left|\sum_{j=1}^n (x_j-\theta_0)\right| \begin{matrix} > \\ \leqslant \end{matrix} \frac{c^*}{\sigma n^{1/2}} \;=\; u_{\alpha/2} \\ &=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } |\bar{x}_n-\theta_0| \begin{matrix} > \\ \leqslant \end{matrix} \sigma u_{\alpha/2} n^{-1/2}, \end{aligned}$$

d.h. den zweiseitigen Gauß-Test in 3.21.

Kommen wir abschließend, unter Beibehaltung der zuvor gegebenen Situation, zu dem Testproblem

$$H=\{\theta:\theta_1\leq\theta\leq\theta_2\} \quad \text{gegen} \quad K=\{\theta:\theta<\theta_1 \text{ oder } \theta>\theta_2\} \tag{3.23}$$

mit $\theta_1<\theta_2$, das sich sehr einfach unter Benutzung der Ergebnisse des vorherigen Abschnitts behandelt lässt. Wir nehmen an, dass θ_1 und θ_2 innere Punkte von Θ sind. Dann folgt aus der Stetigkeit der Gütefunktion leicht für die Menge der unverfälschten Tests zum Niveau α

$$\Phi_\alpha^u \;\subset\; \{\varphi\in\Phi:\mathbb{E}_{\theta_1}\varphi(X)=\mathbb{E}_{\theta_2}\varphi(X)=\alpha\}. \tag{3.24}$$

Satz 3.29. *Neben den zuvor gemachten Annahmen sei $\alpha\in(0,1)$. Dann gilt für das Testproblem* (3.23)*:*

(a) Es existieren $c_1,c_2\in\mathbb{R}$ und $\gamma_1,\gamma_2\in[0,1]$, so dass für $j\in\{1,2\}$

$$W_{\theta_j}(T\notin[c_1,c_2])+\gamma_1 W_{\theta_j}(T=c_1)+\gamma_2 W_{\theta_j}(T=c_2)=\alpha. \tag{3.25}$$

(b) Bildet $(c_1^*, c_2^*, \gamma_1^*, \gamma_2^*)$ eine Lösung von (3.25), und definiert man

$$\varphi^*(x) := \begin{cases} 1, & \text{falls} \quad \notin [c_1^*, c_2^*] \\ \gamma_i^*, & \text{falls } T(x) = c_i^* \quad (i=1,2), \\ 0, & \text{falls} \quad \in (c_1^*, c_2^*) \end{cases} \tag{3.26}$$

(also genauso wie in (3.19)*) so folgt:*

(1) $R(\theta, \varphi^*) = \min_{\varphi, \mathbb{E}_{\theta_1}\varphi(X)=\mathbb{E}_{\theta_2}\varphi(X)=\alpha} R(\theta,\varphi)$ für alle $\theta \notin \{\theta_1, \theta_2\}$.

(2) φ^* ist ein gleichmäßig bester unverfälschter Test zum Niveau α für H gegen K.

Beweis. (a) Nach Satz 3.18(a) existieren $c_1, c_2 \in \mathbb{R}$ und $\gamma_1, \gamma_2 \in [0,1]$, so dass

$$W_{\theta_j}(c_1 < T < c_2) + (1-\gamma_1) W_{\theta_j}(T = c_1) + (1-\gamma_2) W_{\theta_j}(T = c_2) \;=\; 1-\alpha$$

für $j \in \{1,2\}$ gilt, was offensichtlich zu (3.25) äquivalent ist.

(b) Hier ergeben sich sämtliche Behauptungen unter Benutzung von Satz 3.18(b), denn $\psi^* = 1 - \varphi^*$ erfüllt die dortigen Voraussetzungen mit $1-\alpha$ anstelle von α. Für den Nachweis der Optimalität von φ^* unter den unverfälschten Tests zum Niveau α benutze man (3.24), die Unverfälschtheit von φ^* sichert (b.1) des genannten Satzes, wenn man $\varphi_\alpha \equiv \alpha \in \Phi_u^\alpha$ beachtet. □

3.6 Der t-Test für den Mittelwert bei Normalverteilungen: Eine heuristische Herleitung

Betrachten wir wieder ein statistisches Experiment, das auf n stochastisch unabhängigen und jeweils $Normal(\mu, \sigma^2)$-verteilungen Beobachtungen $X_1, ..., X_n$ basiert, wobei jedoch im Gegensatz zu den vorhergehenden Abschnitten sowohl der Mittelwert μ als auch die Varianz σ^2 unbekannt seien. Als Parameterraum liegt also

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

vor. Man will die Vermutung $\mu > \mu_0$ überprüfen und betrachtet dazu das einseitige Testproblem

$$H = \{(\mu, \sigma^2) : \mu \le \mu_0, \sigma^2 > 0\} \quad \text{gegen} \quad K = \{(\mu, \sigma^2) : \mu > \mu_0, \sigma^2 > 0\}.$$

Dieses Problem lässt sich mit den bisherigen Methoden nicht behandeln: In der Vermutung wird zwar nur etwas über den unbekannten Mittelwert μ geäußert, jedoch geht die ebenfalls unbekannte Varianz σ^2 als sogenannter *Störparameter (nuisance parameter)* in die Gestalt der Dichten ein. Hypothese und Alternative sind jetzt

zweidimensionale Mengen, und um sinnvoll etwas über die Vermutung "$\mu > \mu_0$" aussagen zu können, bedarf es auch einer Berücksichtigung der Information, die die Messwerte über σ^2 liefern.

Um zu einem sinnvollen Test für das Problem zu gelangen, wollen wir in diesem Abschnitt ein *heuristisches Vorgehen* wählen: Wäre σ^2 bekannt, würden wir den einseitigen Gauß-Test

$$\varphi(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } \frac{n^{1/2}(\bar{x}_n - \mu_0)}{\sigma} \begin{matrix} > \\ \leqslant \end{matrix} u_\alpha$$

aus Beispiel 3.13 benutzen. Da wir σ nicht kennen, erscheint es naheliegend, für σ einen geeigneten Schätzwert $\widehat{\sigma}(x)$ einzusetzen, und wir erhalten

$$\widehat{\varphi}(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } \frac{n^{1/2}(\bar{x}_n - \mu_0)}{\widehat{\sigma}(x)} \begin{matrix} > \\ \leqslant \end{matrix} \widehat{c}.$$

Dabei sind der Schätzer $\widehat{\sigma}$ und die Konstante $\widehat{c}$ möglichst so zu wählen, dass ein unverfälschter Test zum Niveau α resultiert. Aus der Schätztheorie wissen wir [☞ 2.50 und 2.55], dass

$$\widehat{\sigma}^2(x) \;=\; \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x}_n)^2$$

einen GBES für die Varianz definiert, was folgenden Test bei geeigneter Wahl von $\widehat{c}$ nahelegt:

$$\widehat{\varphi}(x) \;=\; \begin{cases} 1, & \\ 0, & \end{cases} \quad \text{falls } \frac{n^{1/2}(\bar{x}_n - \mu_0)}{\left(\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2\right)^{1/2}} \begin{matrix} > \\ \leqslant \end{matrix} \widehat{c}. \tag{3.27}$$

Um zu untersuchen, ob wir so zu einem unverfälschten Test zum Niveau α gelangen, haben wir uns mit der Verteilung der Teststatistik

$$T(x) = \frac{n^{1/2}(\bar{x}_n - \mu_0)}{(\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2)^{1/2}}$$

unter jedem W_θ zu beschäftigen, wobei wir zunächst die unter $W_{(\mu_0,\sigma^2)}$ betrachten wollen, d.h.

$$W^T_{(\mu_0,\sigma^2)} = \mathbb{P}_{(\mu_0,\sigma^2)}\left(\frac{n^{1/2}(\bar{X}_n - \mu_0)}{(\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2)^{1/2}} \in \cdot\right).$$

Die folgende Überlegung zeigt, dass diese Verteilung *nicht von μ_0 und σ^2 abhängt*. Definieren wir nämlich

$$Y_j := \frac{X_j - \mu_0}{\sigma} \quad \text{für} \quad 1 \leq j \leq n,$$

so sind $Y_1,...,Y_n$ unter $W_{(\mu_0,\sigma^2)}$ stochastisch unabhängig und jeweils $Normal(0,1)$-verteilt. Ferner gilt aber auch

$$\frac{n^{1/2}(\bar{X}_n-\mu_0)}{(\frac{1}{n-1}\sum_{j=1}^n(X_j-\bar{X}_n)^2)^{1/2}} \;=\; \frac{n^{1/2}\bar{Y}_n}{(\frac{1}{n-1}\sum_{j=1}^n(Y_j-\bar{Y}_n)^2)^{1/2}},$$

wie eine einfache Rechnung zeigt, so dass

$$\mathbb{P}_{(\mu_0,\sigma^2)}\left(\frac{n^{1/2}(\bar{X}_n-\mu_0)}{(\frac{1}{n-1}\sum_{j=1}^n(X_j-\bar{X}_n)^2)^{1/2}}\in\cdot\right) = \mathbb{P}_{(0,1)}\left(\frac{n^{1/2}\bar{X}_n}{(\frac{1}{n-1}\sum_{j=1}^n(X_j-\bar{X}_n)^2)^{1/2}}\in\cdot\right).$$

Als nächstes erinnern wir uns daran, dass in Beispiel 2.47 als Folgerung aus dem Satz von Basu gezeigt wurde, dass $\bar{X}_n$ und $\frac{1}{n-1}\sum_{j=1}^n(X_j-\bar{X}_n)^2$ unter jedem $\mathbb{P}_{(\mu,\sigma^2)}$ stochastisch unabhängig sind. Weitergehende Auskunft gibt der folgende

Satz 3.30. *Es seien $Y_0,...,Y_n$ stochastisch unabhängige und jeweils $Normal(0,1)$-verteilte ZG mit $\bar{Y}_n=\frac{1}{n}\sum_{j=1}^n Y_j$, wobei $n\geq 2$. Dann gilt:*

$$\sum_{j=1}^n(Y_j-\bar{Y}_n)^2 \;\stackrel{d}{=}\; \sum_{j=1}^{n-1}Y_j^2 \;\stackrel{d}{=}\; \chi^2_{n-1}, \tag{3.28}$$

$$\frac{n^{1/2}\bar{Y}_n}{(\frac{1}{n-1}\sum_{j=1}^n(Y_j-\bar{Y}_n)^2)^{1/2}} \;\stackrel{d}{=}\; \frac{Y_0}{(\frac{1}{n-1}\sum_{j=1}^{n-1}Y_j^2)^{1/2}} \;\stackrel{d}{=}\; t_{n-1}, \tag{3.29}$$

wobei t_n für $n\geq 1$ die $\mathfrak{R}$-Dichte

$$f_n(x) \;=\; \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})(\pi n)^{1/2}}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2} \tag{3.30}$$

*besitzt und als **(Studentsche) t-Verteilung mit n Freiheitsgraden** oder auch kurz als t_n-**Verteilung** bezeichnet wird.*

In Form und Lage sehen sich die Dichten der t-Verteilung und der Standard-Normalverteilung sehr ähnlich, wie Abb. 3.3 illustriert. Allerdings hat die t-Verteilung lediglich polynomial gegen 0 fallende Flanken (engl. *tails*), während die der Standard-Normalverteilung exponentiell gegen 0 konvergieren. Als einfache Gedächtnishilfe für die t-Verteilung mag die symbolische Schreibweise

$$t_n \;=\; \frac{Normal(0,1)}{\sqrt{\chi^2_n/n}} \tag{3.31}$$

dienen, wobei Zähler und Nenner stochastisch unabhängig sind.

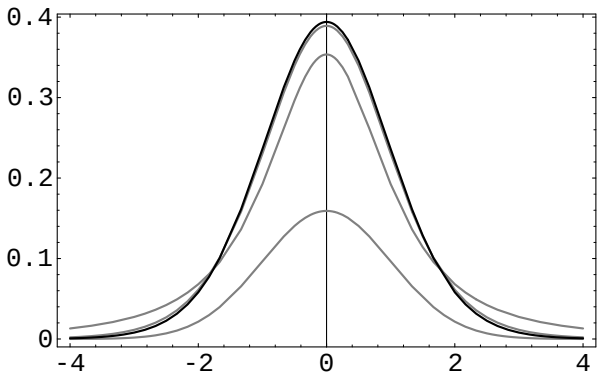


Abb. 3.3 Dichten der t_2-, t_{10}-, t_{20}- (grau, von unten nach oben) und $Normal(0,1)$-Verteilung (schwarz)

Beweis (von Satz 3.30). Sei $Z = (Z_1, ..., Z_n)^\top$ ein Zufallsvektor mit n-dimensionaler Normalverteilung mit Mittelwertvektor $\nu \in \mathbb{R}^n$ und symmetrischer, positiv definiter Kovarianzmatrix Σ, d.h. $Z \stackrel{d}{=} Normal_n(\nu, \Sigma)$. Nach Satz 30.9 in [2] gilt für jede orthogonale $n \times n$-Matrix C:

$$CZ \stackrel{d}{=} Normal_n(C\nu, C\Sigma C^\top).$$

Wenden wir dies in der vorliegenden Situation auf $Y = (Y_1, ..., Y_n)^\top$ an, so folgt aus $Y \stackrel{d}{=} Normal_n(0, I_n)$ offensichtlich

$$CY \stackrel{d}{=} Normal_n(C0, CI_nC^\top) = Normal_n(0, CC^\top) = Normal_n(0, I_n)$$

für jedes orthogonale C, d.h. mit $Y_1, ..., Y_n$ sind auch die Komponenten von CY wieder unabhängig und jeweils $Normal(0,1)$-verteilt. Wähle nun irgendeine orthogonale Matrix C mit erstem Zeilenvektor $(n^{-1/2}, ..., n^{-1/2})$. Es folgt für $X = CY$ offensichtlich $X_1 = n^{-1/2} \sum_{j=1}^n Y_j$ sowie $X_j = c^{(j)}Y$ für $2 \le j \le n$, wobei $c^{(j)}$ den j-ten Zeilenvektor von C bezeichnet. Aus der Längentreue orthogonaler Abbildungen ergibt sich aber weiter

$$\sum_{j=1}^n X_j^2 = \sum_{j=1}^n Y_j^2$$

und daraus unter Beachtung von $X_1^2 = n^{-1}(\sum_{j=1}^n Y_j)^2 = n\bar{Y}_n^2$

$$\sum_{j=2}^n X_j^2 = \sum_{j=1}^n Y_j^2 - n\bar{Y}_n^2 = \sum_{j=1}^n (Y_j - \bar{Y}_n)^2.$$

$\sum_{j=1}^n (Y_j - \bar{Y}_n)^2$ besitzt also dieselbe Verteilung wie die Summe von $n-1$ unabhängigen quadrierten $Normal(0,1)$-verteilten ZG, was (3.28) unter Hinweis auf Satz 3.19

in [2] beweist. Außerdem folgt erneut die Unabhängigkeit von Stichprobenmittel und Stichprobenvarianz [☞ 2.47], denn

$$\bar{Y}_n \;=\; n^{-1/2}X_1 \quad \text{und} \quad \frac{1}{n-1}\sum_{j=1}^{n}(Y_j-\bar{Y}_n)^2 \;=\; \frac{1}{n-1}\sum_{j=2}^{n}X_j^2.$$

Zum Beweis von (3.29) müssen wir nach den vorhergehenden Überlegungen und unter Beachtung von $n^{1/2}\bar{Y}_n \stackrel{d}{=} \mathit{Normal}(0,1)$ nur noch zeigen, dass die t_n-Verteilung tatsächlich die in (3.30) angegebene Dichte besitzt. Unter Hinweis auf (3.31) betrachten wir dazu die ZG $U/\sqrt{V/n}$, wobei U und V stochastisch unabhängig sind mit $U \stackrel{d}{=} \mathit{Normal}(0,1)$ und $V \stackrel{d}{=} \chi^2_n$. Durch Differentiation von

$$\mathbb{P}(\sqrt{V/n} \le t) = \mathbb{P}(V \le nt^2), \quad t \in (0,\infty),$$

nach t folgt leicht, dass $\sqrt{V/n}$ die $\mathfrak{R}$-Dichte

$$\begin{aligned} h(t) \;&=\; \frac{2nt}{2^{n/2}\Gamma(n/2)}(nt^2)^{(n/2)-1}e^{-nt^2/2}\mathbf{1}_{(0,\infty)}(t) \\ &=\; \frac{n^{n/2}}{2^{(n/2)-1}\Gamma(n/2)}t^{n-1}e^{-nt^2/2}\mathbf{1}_{(0,\infty)}(t) \end{aligned}$$

besitzt. Vermöge Satz 24.5 in [1] erhalten wir nun

$$\begin{aligned} f_n(x) \;&=\; \int_0^\infty \frac{t}{(2\pi)^{1/2}}e^{-(xt)^2/2}\frac{n^{n/2}}{2^{(n/2)-1}\Gamma(n/2)}t^{n-1}e^{-nt^2/2}\,dt \\ &=\; \frac{n^{(n+1)/2}}{(n\pi)^{1/2}2^{(n-1)/2}\Gamma(n/2)}\int_0^\infty t^n e^{-t^2(x^2+n)/2}\,dt \\ &=\; \frac{n^{(n+1)/2}}{(n\pi)^{1/2}2^{(n-1)/2}\Gamma(n/2)}\int_0^\infty \left(\frac{2y}{x^2+n}\right)^{n/2}\frac{e^{-y}}{x^2+n}\left(\frac{2y}{x^2+n}\right)^{-1/2}dy \\ &\left[\text{per Substitution } t \;=\; \left(\frac{2y}{x^2+n}\right)^{1/2} \text{ und } dt \;=\; \frac{1}{x^2+n}\left(\frac{2y}{x^2+n}\right)^{-1/2}dy\right] \\ &=\; \frac{1}{(n\pi)^{1/2}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}\int_0^\infty y^{(n-1)/2}e^{-y}\,dy. \end{aligned}$$

Es folgt (3.30), da das letzte Integral bekanntlich $\Gamma(\frac{n+1}{2})$ entspricht. □

Gemäß Satz 3.30 besitzt die in (3.29) gegebene Teststatistik T also unter jedem $W_{(\mu_0,\sigma^2)}$, $\sigma^2 > 0$, eine t_{n-1}-Verteilung [beachte (3.30)], deren α-Fraktil im Folgenden mit $t_{n-1,\alpha}$ bezeichnet wird. Sei nun $\widehat{\varphi}$ der in (3.27) definierte Test mit $\widehat{c} = t_{n-1,\alpha}$, d.h.

$$\widehat{\varphi}(x) \;=\; \begin{cases} 1, & \text{falls } \dfrac{n^{1/2}(\bar{x}_n-\mu_0)}{\left(\frac{1}{n-1}\sum_{j=1}^n(x_j-\bar{x}_n)^2\right)^{1/2}} \;\begin{matrix}>\\ \leqslant\end{matrix}\; t_{n-1,\alpha}. \\ 0, & \end{cases}$$

Dann heißt $\widehat{\varphi}$ *einseitiger t-Test* für H: "$\mu \leq \mu_0$" gegen K: "$\mu > \mu_0$". Ganz entsprechend kann man für das zweiseitige Testproblem H: "$\mu = \mu_0$" gegen K: "$\mu \neq \mu_0$" den *zweiseitigen t-Test*

$$\widehat{\varphi}(x) \;=\; \begin{cases} 1, & \text{falls } \dfrac{n^{1/2}|\bar{x}_n - \mu_0|}{\left(\frac{1}{n-1}\sum_{j=1}^n (x_j - \bar{x}_n)^2\right)^{1/2}} \begin{matrix} > \\ \leqslant \end{matrix} t_{n-1,\alpha/2} \\ 0, & \end{cases}$$

heuristisch aus dem zweiseitigen Gauß-Test herleiten. Es gilt dann:

Satz 3.31. *$\widehat{\varphi}$ und $\widetilde{\varphi}$ bilden unverfälschte Tests zum Niveau α für die Testprobleme $H = (-\infty, \mu_0] \times (0,\infty)$ gegen $K = (\mu_0,\infty) \times (0,\infty)$ bzw. $H = \{\mu_0\} \times (0,\infty)$ gegen $K = (\mathbb{R}\backslash\{\mu_0\}) \times (0,\infty)$.*

W.S. GOSSET (1876–1937).

Die t-Verteilung ebenso wie den t-Test verdanken wir dem engl. Statistiker WILLEM SEALY GOSSETT, der nach seinem Studium der Chemie und Mathematik in Oxford 1899 an der Dubliner Guiness-Brauerei zu arbeiten begann. Guinness war ein fortschrittlicher agro-chemischer Betrieb, und Gosset wandte sein statistisches Wissen, das er zuvor bei Studien und Versuchen im biometrischen Labor von KARL PEARSON erworben hatte, sowohl in der Brauerei als auch in der Landwirtschaft an, um die beste Gersten-Qualität für die Bierherstellung zu erzeugen. Nachdem ein anderer Wissenschaftler bei der Brauerei eine Arbeit publiziert hatte, die – zum Schaden der Brauerei – Betriebsgeheimnisse enthielt, verbot die Brauerei ihren Mitarbeitern, irgendwelche Arbeiten zu veröffentlichen. Aus diesem Grund war Gosset gezwungen, seine Ergebnisse unter einem Pseudonym zu veröffentlichen, wobei er den Namen "Student" wählte [8]. Dies erklärt auch die Bezeichnung "Studentsche t-Verteilung".

KARL PEARSON (1857–1936) hatte ein sehr gutes Verhältnis zu Gosset und half ihm bei der mathematischen Kleinarbeit in seinen Schriften. Obgleich dies auch für die o.g. Arbeit [8] von 1908 der Fall war, entging ihm dabei ihre Bedeutung, da sie sich mit kleinen Stichprobengrößen befasste, ein typisches Problem einer Brauerei, nicht dagegen eines Biometers, der üblichweise hunderte von Stichproben zu Verfügung hat.

Karl Pearson war übrigens der Vater von EGON PEARSON (1895–1980), der 1933 gemeinsam mit JERZY NEYMAN das für die Testtheorie fundamentale, nach ihnen benannte Lemma 3.1 bewies. Quelle: [20].

Beweis (von Satz 3.31). Wir zeigen zunächst die Behauptung für $\widehat{\varphi}$. Für alle $\mu \in \mathbb{R}$ aund $\sigma^2 > 0$ gilt nach Satz 3.30 und den zugehörigen Vorüberlegungen

$$\mathbb{E}_{(\mu,\sigma^2)}\widehat{\varphi}(X) \;=\; \mathbb{P}_{(\mu,\sigma^2)}\left(\frac{n^{1/2}(\bar{X}_n - \mu_0)}{(\frac{1}{n-1}\sum_{j=1}^n (X_j - \bar{X}_n)^2)^{1/2}} > t_{n-1,\alpha}\right)$$

$$\lessgtr \mathbb{P}_{(\mu,\sigma^2)} a\left(\frac{n^{1/2}(\bar{X}_n-\mu)}{(\frac{1}{n-1}\sum_{j=1}^n (X_j-\bar{X}_n)^2)^{1/2}} > t_{n-1,\alpha}\right) = \alpha,$$

falls $\mu \lessgtr \mu_0$, was offensichtlich das Gewünschte liefert.

Das Argument für $\widetilde{\varphi}$ ist etwas schwieriger und bedarf der Ausnutzung der Unabhängigkeit von Stichprobenmittel und Stichprobenvarianz. Bezeichne Q die von (μ,σ^2) unabhängige Verteilung von $(\frac{1}{n-1}\sum_{j=1}^n (\frac{X_j-\bar{X}_n}{\sigma})^2)^{1/2}$, Φ die Verteilungsfunktion der $Normal(0,1)$-Verteilung, und sei $\Delta := \mu-\mu_0$. Dann ergibt sich

$$\begin{aligned}
\mathbb{E}_{(\mu,\sigma^2)}\widetilde{\varphi}(X) &= \mathbb{P}_{(\mu,\sigma^2)}\left(\frac{n^{1/2}|\bar{X}_n-\mu_0|}{(\frac{1}{n-1}\sum_{j=1}^n (X_j-\bar{X}_n)^2)^{1/2}} > t_{n-1,\alpha/2}\right)\\
&= \int_0^\infty \mathbb{P}_{(\mu,\sigma^2)}\left(\frac{n^{1/2}|\bar{X}_n-\mu+\Delta|}{\sigma} > x t_{n-1,\alpha/2}\right) Q(dx)\\
&= \int_0^\infty \mathbb{P}_{(0,1)}\left(\left|X_1+\frac{n^{1/2}\Delta}{\sigma}\right| > x t_{n-1,\alpha/2}\right) Q(dx)\\
&= \int_0^\infty \left(1-\Phi\left(-\frac{n^{1/2}\Delta}{\sigma}+x t_{n-1,\alpha/2}\right)+\Phi\left(-\frac{n^{1/2}\Delta}{\sigma}-x t_{n-1,\alpha/2}\right)\right) Q(dx).
\end{aligned}$$

Benutzt man nun noch die (schon vor (3.15) erwähnte) Tatsache, dass die Funktion $r\mapsto 1-\Phi(r+s)+\Phi(r-s)$ für jedes $s\geq 0$ ihr Minimum im Punkt $r=0$ annimmt, so lässt sich die letzte obere Zeile weiter nach unten abschätzen durch

$$\begin{aligned}
\mathbb{E}_{(\mu,\sigma^2)}\widetilde{\varphi}(X) &\geq \int_0^\infty (1-\Phi(x t_{n-1,\alpha/2})+\Phi(-x t_{n-1,\alpha/2}))\, Q(dx)\\
&= \mathbb{P}_{(\mu,\sigma^2)}\left(\frac{n^{1/2}|\bar{X}_n-\mu|}{(\frac{1}{n-1}\sum_{j=1}^n (X_j-\bar{X}_n)^2)^{1/2}} > t_{n-1,\alpha/2}\right) = \alpha,
\end{aligned}$$

was den Beweis abschließt. □

Dass $\widehat{\varphi}$ und $\widetilde{\varphi}$ sogar gleichmäßig beste unverfälschte Tests zum Niveau α für die jeweiligen Testprobleme bilden, ist keineswegs klar und bedarf zunächst einer geeigneen Theorie für Modelle mit mehrdimensionalem Parameter. Diese wird in den nächsten drei Abschnitten bereitgestellt.

3.7 Bedingte Tests

Gegeben sei wieder ein beliebitges statistisches Experiment $\mathscr{E} = (\mathfrak{X}, \mathscr{A}, (W_\theta)_{\theta\in\Theta})$, jetzt allerdings mit *mehrdimensionalem* Parameterraum $\Theta \subset \mathbb{R}^d$, $d \geq 2$. Betrachten wir dann ein Testproblem H gegen $K = \Theta - H$, wobei H und K einen gemeinsamen topologischen Rand J besitzen, so muss die Gütefunktion jedes unverfälschten Tests

Literaturverzeichnis

1. Biometrika tables for statisticians. Vol. I. Edited by E. S. Pearson and H. O. Hartley. 3^{rd} Edition. Published for the Biometrika Trustees by the Cambridge University Press, London-New York-Ibadan (1966)
2. Alsmeyer, G.: Wahrscheinlichkeitstheorie, *Skripten zur Mathematischen Statistik, 5. Auflage*, vol. 30. Institut f. Math. Statistik, Universität Münster, Münster (2007)
3. Bahadur, R.R.: Sufficiency and statistical decision functions. Ann. Math. Statistics **25**, 423–462 (1954)
4. Basu, D.: On statistics independent of a complete sufficient statistic. Sankhyā **15**, 377–380 (1955)
5. Bickel, P.J., Doksum, K.A.: Mathematical statistics. Basic ideas and selected topics. Holden-Day Inc., San Francisco, Calif. (1977)
6. Blackwell, D.: Conditional expectation and unbiased sequential estimation. Ann. Math. Statistics **18**, 105–110 (1947)
7. Gauss, C.: Theoria combinationis observationum erroribus minimis obnoxiae (2 Teile), *Commentationes Societatis Regiae Scientiarum Gottingensis recentiores, classis mathematicae*, vol. 5/6 (1821–23). Deutsche Übersetzung: Abhandlungen zur Methode der kleinsten Quadrate. Hrsg. Anton Börsch; Paul Simon. Berlin 1887
8. Gosset, W.S.: The probable error of a mean. Biometrika **6**(1), 1–25 (1908). Originally published under the pseudonym "Student".
9. Joshi, V.M.: On the attainment of the Cramér-Rao lower bound. Ann. Statist. **4**, 998–1002 (1976)
10. Landers, D., Rogge, L.: Minimal sufficient σ-fields and minimal sufficient statistics. Two counterexamples. Ann. Math. Statist. **43**, 2045–2049 (1972)
11. Lehmann, E.L.: Theory of point estimation. John Wiley & Sons Inc., New York (1983)
12. Lehmann, E.L., Scheffé, H.: Completeness, similar regions, and unbiased estimation. I. Sankhyā **10**, 305–340 (1950)
13. Meyer, P.A.: Probability and potentials. Blaisdell Publishing Company Ginn and Co., Waltham, Mass.-Toronto, Ont.-London (1966)
14. Neyman, J., Pearson, E.: On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London, Series A **231**, 289–337 (1933)
15. Owen, D.B.: Handbook of statistical tables. Addison-Wesley Publishing Co., Inc., Reading, Mass.-London (1962)
16. Pitcher, T.S.: Sets of measures not admitting necessity and sufficient statistics or subfields. Ann. Math. Statist. **28**, 267–268 (1957)
17. Rao, C.R.: Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81–91 (1945)
18. Wijsman, R.A.: On the attainment of the Cramér-Rao lower bound. Ann. Statist. **1**, 538–542 (1973)
19. Wikipedia-Community: Methode der kleinsten Quadrate. http://de.wikipedia.org/wiki/Methode_der_kleinsten_Quadrate
20. Wikipedia-Community: William Sealy Gosset. http://de.wikipedia.org/wiki/William_Sealy_Gosset
21. Witting, H.: Mathematische Statistik I. B.G. Teubner, Stuttgart (1985)

Sachverzeichnis