

Skript

Mathematische Statistik

Max v. Renesse
Aufgezeichnet von Tobias Weihrauch

Sommersemester 2012
Universität Leipzig

Inhaltsverzeichnis

1	Einführung	3
1.1	Statistik als Teil der Stochastik	3
1.2	Standardfragen der Statistik in Beispielen	4
2	Parameterschätzung	4
2.1	Punktschätzer	4
2.1.1	Beispiele von (parametrischen) statistischen Modellen	7
2.1.2	Konsistenz von Schätzern	9
2.1.3	Maximum-Likelihood-Schätzer	10
2.1.4	Mittlerer quadratischer Fehler	13
2.1.5	Fisher-Information und Cramer-Rao Ungleichung	16
2.1.6	Suffizienz und Satz von Rao-Blackwell	21
2.1.7	Vollständigkeit und der Satz von Lehmann-Scheffé	25
2.1.8	Der optimale Varianzschätzer im n -fachen Gaußmodell bei unbekanntem Erwartungswert.	27
2.1.9	Bayes'sche Schätzer	31
2.2	Bereichsschätzer (Konfidenzmengen)	34
2.2.1	Konfidenzbereiche im Binomialmodell	35
2.2.2	Pivotstatistiken	41
3	Verteilungen rund um die Normalverteilung	42
4	Testen	49
4.1	Einführung in die Testproblematik	49
4.2	Gleichmäßig beste Tests	50
4.3	Das Neymann-Pearson Lemma	52
4.4	Likelihood-Quotienten-Tests	54
5	Nichtparametrische Modelle	58
5.1	Der Satz von Glivenko-Cantelli	59
5.2	Eine Quantitative Version von Glivenko-Cantelli	61
5.2.1	Konzentrationsungleichungen	61
5.2.2	Beweis von Satz 5.4	64
5.3	Der Kolmogorov-Smirnov-Anpassungstest	66
5.4	χ^2 -Anpassungstest	69
6	Lineare Modelle	73
6.1	Der Satz von Gauß-Markov	75
6.2	Konfidenzbereiche und Tests in linearen Gauß-Modellen	77
6.3	Anwendung: Einweg-Klassifizierung und ANOVA-Methode	80

Vorwort

Dies ist eine redigierte Mitschrift der Vorlesung 'Mathematische Statistik', die ich in den Sommersemestern 2011 und 2012 im Bachelor-Studiengang Mathematik an der TU Berlin bzw. im Diplomstudiengang Mathematik an der Universität Leipzig gehalten habe. Entsprechend folgt die Gliederung den für diesen Kurs typischen Modulvorgaben an deutschen Universitäten.

Bei der Vorbereitung habe ich besonders vom zweiten Teil des wunderbaren Lehrbuchs 'Stochastik' des Münchner Kollegen H.-O. Georgii profitiert, das beim de Gruyter-Verlag in dritter Auflage vorliegt. Daher stellt die Vorlesung im wesentlichen eine gewichtete Auswahl der im Georgii'schen Buch behandelten Themen dar. Als Exkurs gegenüber dieser Hauptquelle ist ein längerer Abschnitt über eine quantitative Version des Satzes von Glivenko-Cantelli eingefügt, der einen Einblick in fortgeschrittenere Fragen der nicht-parametrischen Statistik mit dem Thema Konzentrationsungleichungen verbindet¹.

Herzlichster Dank gilt Herrn Tobias Weihrauch aus Leipzig, der die Vorlesung aufmerksam verfolgt und dabei in Echtzeit den Latex-Grundstock für dieses Skript gelegt hat.

M. v. Renesse

¹Als Vorlage diente hier eine Vorlesung über maschinelles Lernen von Peter Bartlett (Berkeley)

1 Einführung

1.1 Statistik als Teil der Stochastik

Die Stochastik (Griechisch 'Kunst des Ratens') gliedert sich in zwei Teile: Wahrscheinlichkeitstheorie und Statistik. In der Wahrscheinlichkeitstheorie konstruiert und analysiert man abstrakte mathematische Modelle des Zufalls. In der Statistik hingegen wendet man diese Modelle auf konkrete Sachverhalte an.

Dabei besteht der erste und wichtigste Schritt der Statistik darin, zu den beobachteten Realisierungen eines konkreten Zufallsmechanismus aus der Menge von theoretischen Modellen ein möglichst passgenaues zu finden. Im zweiten Schritt kann das ausgewählte Modell dann für Prognosen und weitere Analysen genutzt werden.

Umgekehrt sind theoretischen Fragen im Bereich Modellbildung häufig von empirischen Beobachtungen und Experimenten motiviert. Eine strikte Trennung zwischen der Wahrscheinlichkeitstheorie und der Statistik ist daher unmöglich. Als Begriffe stehen sie jedoch für die induktive bzw. empirische Seite der Lehre vom Zufall.

Beispiel. Wir wollen das anhand des Zufallsmechanismus "Münzwurf" illustrieren.

Am Anfang steht die Beobachtung, dass der Ausgang eines Münzwurfes zufällig ist in dem Sinne, dass es keine logische Regel zur genauen Vorhersage des Einzelexperiments zu geben scheint.

- 1) Der erste Schritt zum Verstehen dieses Phänomens besteht in der Entwicklung eines geeigneten mathematischen Rahmens, in welchem der Vorgang "Zufälliger Münzwurf" beschrieben werden soll. Das führt dann z.B. auf das einfache Bernoulli-Modell eines 0-1-Experimentes $\Omega = \{1, 0\}$, $P(1) = p$, $P(0) = 1 - p$.
- 2) Damit ist die Klasse der Modelle zur Beschreibung des Münzwurfes festgelegt, es fehlt die Festlegung des Parameters p . Hierzu wird die Münze z.B. dreimal geworfen und die Folge der Ausgänge 1, 0, 1 beobachtet. Hieraus leiten wir die Annahme ab dass $p \approx \frac{2}{3}$.

In diesem Beispiel wäre der 1. Analyseschritt (Modellentwicklung) also der induktiven Seite und der 2. Analyseschritt (Modellauswahl: $p = \frac{2}{3}$) der empirischen Seite zuzuordnen. Letztere steht im Mittelpunkt dieser Vorlesung.

Literatur zur Vorlesung:

1. Hans-Otto Georgii, "Stochastik", de Gruyter (2. Auflage)
2. Vorlesungsskript "Mathematische Statistik" von Prof. Matthias Löwe, Univ. Münster
3. Vorlesungsskript "Statistik" von Prof. Volker Schmidt, Univ. Ulm
4. Achim Klenke "Wahrscheinlichkeitstheorie", Springer, 2. Auflage (als Hintergrundreferenz)

1.2 Standardfragen der Statistik in Beispielen

Beispiele.

1. *Schätzen.* Ein Zufallsgenerator produziert Zufallszahlen gleichverteilt aus $\{1, 2, \dots, M\}$. M ist dabei unbekannt. Der Apparat wird nun 10 mal betätigt mit dem Resultat

3, 32, 98, 9, 29, 4, 21, 67, 6, 44

Gesucht ist nun ein Schätzwert für den Parameter M (Punktschätzen) bzw. ein Intervall $I = [\alpha, \beta]$, welches mit sehr großer Wahrscheinlichkeit den Parameter M enthält (Bereichsschätzen).

2. *Testen.* Sind die in 1. protokollierten Beobachtungen verträglich mit der Annahme, dass der Apparat jede der Zahlen $\{1, \dots, M\}$ gleich wahrscheinlich generiert?
3. *Entscheiden.* 70 Patienten mit einer bestimmten Krankheit werden mit zwei verschiedenen Medikamenten behandelt, mit dem folgenden Resultat

Verlauf \ Medikament	A	B
schwer	20	18
leicht	22	10

Hat die Wahl des Medikamentes einen signifikanten Einfluss auf den Krankheitsverlauf?

4. *Regredieren.* Die Patientendatei einer Krankenkasse enthält 120 Patientendaten als Paare der Form (Kundenalter, Kosten p.A). Welcher quantitative Zusammenhang besteht zwischen Alter & Kosten?

2 Parameterschätzung

2.1 Punktschätzer

Beispiel (Qualitätskontrolle bei der Glühbirnenfabrikation). Die Lebensdauer einer einzelnen Glühbirne $X_i : \Omega \rightarrow \mathbb{R}_{\geq 0}$ ist zufällig. Die Annahme der Gedächtnislosigkeit² erzwingt die Modellierung der Lebensdauer X_i als exponentiell verteilte Zufallsvariable $P(X_i \geq t) = \exp^{-\vartheta t}$, wobei $\vartheta \geq 0$ ein Parameter ist, der (z.B. vom Produzenten) festgelegt werden kann. Als Empfänger einer Lieferung von Glühbirnen dieser Fabrikation wollen wir nun ϑ experimentell bestimmen.

- Wir lassen nun 100 Glühbirnen dieser Sorte abbrennen und protokollieren die Brenndauern

$$\Rightarrow (X_1, \dots, X_{100}) = (2h, 3d\ 7h, \dots, 0.3h)$$

²Gedächtnislosigkeit heißt hier z.B. $P(X > t + s | X > t) \stackrel{!}{=} P(X > s)$.

- (X_1, \dots, X_{100}) stellt das Protokoll einer 100-fachen Ziehung von exponentiell verteilten Zufallsvariablen dar. Das *starke Gesetz der großen Zahlen* besagt

$$\frac{1}{n} \cdot \sum_{i=1}^n X_i \xrightarrow{n \rightarrow \infty} \mathbb{E}(X) \text{ fast sicher,}$$

wobei die (X_i) unabhängig identisch verteilt sind. Für den Parameter ϑ finden wir somit asymptotisch fast sicher

$$\frac{1}{100} \sum_{i=1}^{100} X_i \approx \mathbb{E}_{\vartheta}(X)$$

mit dem Erwartungswert $\mathbb{E}_{\vartheta}(X)$ einer exponentiell verteilten Zufallsvariablen mit Parameter ϑ , d.h.

$$\mathbb{E}_{\vartheta}(X) = \int_0^{\infty} x \cdot \vartheta \cdot e^{-\vartheta \cdot x} dx = [-x \cdot e^{-\vartheta \cdot x}]_0^{\infty} + \int_0^{\infty} e^{-\vartheta \cdot x} dx = \frac{1}{\vartheta}.$$

- Somit ist eine erste (nicht ganz naive) Antwort, dass

$$\hat{\vartheta} = \frac{100}{\sum_{i=1}^{100} X_i}.$$

Die Vorschrift $\hat{\vartheta} = \hat{\vartheta}(X_1, \dots, X_{100}) \rightarrow \mathbb{R}$ ist eine Abbildung von dem Raum der Beobachtungen in den Raum der Parameter. Man spricht von einem *Schätzer* für ϑ (siehe unten). Da die konkrete Realisierung (X_1, \dots, X_{100}) der 100 Stichproben vom Zufall abhängt, ist auch $\hat{\vartheta} = \hat{\vartheta}(X)$ eine Zufallsgröße. Das Gesetz der großen Zahlen nährt jedoch die Hoffnung, dass $\hat{\vartheta}(X)$ für große n mit hoher Wahrscheinlichkeit nahe beim fixen aber unbekanntem Parameter ϑ liegt.

Das hier gegebene Beispiel wird im Rahmen der Statistik wie folgt systematisiert.

Definition 2.1 (Statistisches Modell).

- Ein *statistisches Modell* ist ein Tripel $M = (\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_{\vartheta}; \vartheta \in \Theta))$ bestehend aus
 - einer Menge \mathfrak{X} als *Stichprobenraum*
 - einer Sigma-Algebra \mathfrak{F} auf \mathfrak{X} als *Algebra der Beobachtungen* und
 - einer Familie von Wahrscheinlichkeitsmaßen $(\mathbb{P}_{\vartheta}; \vartheta \in \Theta)$ auf $(\mathfrak{X}, \mathfrak{F})$.
- Falls $M = (E, \mathfrak{E}, (\mathbb{Q}_{\vartheta}; \vartheta \in \Theta))$ ein statistisches Modell ist, heißt das Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_{\vartheta}; \vartheta \in \Theta))$
 - $\mathfrak{X} = E^n$
 - $\mathfrak{F} = \mathfrak{E}^{\otimes}$ (n -fache Produkt- σ -Algebra)

– $\mathbb{P}_\vartheta = \otimes_{k=1}^n \mathbb{Q}_\vartheta$ (Produkt-Wahrscheinlichkeitsmaß)

das zugehörige n -fache Produktmodell (Notation $M^{\otimes n}$).

- Falls (Σ, \mathfrak{S}) ein weiterer messbarer Raum ist, so heißt eine messbare Abbildung

$$S : (\mathfrak{X}, \mathfrak{F}) \rightarrow (\Sigma, \mathfrak{S})$$

eine *Statistik*.

- Eine Abbildung

$$\tau : \Theta \rightarrow \Sigma$$

in eine Menge Σ heißt *Kenngroße*. Eine Statistik $T : \mathfrak{X} \rightarrow \Sigma$ heißt ein *Schätzer* für τ . (Häufige Notation $T = \hat{\tau}$)

- Falls Σ ein Vektorraum ist, heißt ein Schätzer $T : \mathfrak{X} \rightarrow \Sigma$ für die Kenngroße $\tau : \Theta \rightarrow \Sigma$ *erwartungstreu* für τ , falls

$$\mathbb{E}_\vartheta(T) := \int_{\mathfrak{X}} T(x) \mathbb{P}_\vartheta(dx) = \tau(\vartheta) \quad \forall \vartheta \in \Theta.$$

Vereinbarungen zur Notation

Wir werden wir ferner von den folgenden Schreibweisen Gebrauch machen.

- x bezeichnet ein konkretes Element $x \in \mathfrak{X}$.
- X steht für die identische Zufallsvariable auf \mathfrak{X} , d.h. $X : \mathfrak{X} \rightarrow \mathfrak{X}$, $X(x) = x$.
- $\mathbb{1}_{X=x}$ ist die charakteristische Funktion der Menge $\{X = x\}$.
- Für $T : \mathfrak{X} \rightarrow \mathbb{R}$ benutzen wir die folgenden äquivalenten Schreibweisen für den Erwartungswert

$$\mathbb{E}_\vartheta(T) = \mathbb{E}_\vartheta(T(X)) = \int_{\mathfrak{X}} T(x) \mathbb{P}_\vartheta(dx).$$

- Für eine Zufallsvariable Y mit Werten in \mathfrak{X} und einem Wahrscheinlichkeitsmaß ν auf \mathfrak{X} schreiben wir $Y \simeq \nu$, falls ν die Verteilung von Y ist. Analog schreiben wir $Y \simeq Z$, falls die Zufallsvariablen Y und Z dieselbe Verteilung haben.

Entsprechend schreiben sich etwa bedingte Wahrscheinlichkeiten wie folgt

$$\mathbb{P}_\vartheta(X = x \mid S(X) = s) = \frac{\mathbb{P}_\vartheta(\{X = x\} \cap \{S(X) = s\})}{\mathbb{P}_\vartheta(S(X) = s)}.$$

2.1.1 Beispiele von (parametrischen) statistischen Modellen

Falls die Parametermenge Θ eine Teilmenge des \mathbb{R}^d ist, spricht man von einem *parametrischen Modell*. Wir wollen nun einige Exemplare vorstellen.

Beispiel (Münzwurf). Wie gesehen, wird der einmalige Wurf einer Münze mit den Seiten "1" und "0" durch das statistische Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta)) = (\{0, 1\}, \mathcal{P}(\{0, 1\}), (\mathbb{P}_\vartheta)_{\vartheta \in [0,1]})$ beschrieben, wobei $\mathcal{P}(\{0, 1\})$ die Potenzmenge von $\{0, 1\}$ und \mathbb{P}_ϑ das Bernoulli-Wahrscheinlichkeitsmaß auf $\{0, 1\}$ mit $\mathbb{P}_\vartheta = B_\vartheta$, d.h. $B_\vartheta(1) = \vartheta$ und $B_\vartheta(0) = 1 - \vartheta$ bezeichnen. Dieses Modell wird auch *Bernoulli-Modell* genannt.

Beispiel (Verlauf des n -fachen Münzwurfs). Die Münze aus dem vorigen Beispiel wird nun n mal geworfen und wir halten die Folge der Ausgänge X_1, \dots, X_n mit $X_i \in \{0, 1\}$ fest. Das zugehörige statistische Modell wäre in diesem Fall $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta)) = (\{0, 1\}^n, \mathcal{P}(\{0, 1\}^{\otimes n}), (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta \in [0,1]})$, also einfach das n -fache Produkt des obigen Bernoulli-Modells. Entsprechend nenne wir es das *n -fache Bernoulli-Produktmodell*.

Beispiel (Anzahl der Erfolge beim n -fachen Münzwurf). Wenn beim n -fachen Wurf der Münze mit unbekanntem Erfolgsparameter ϑ lediglich die Gesamtanzahl der "1"-en als Beobachtung festhalten, können wir mit dem statistischen Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ arbeiten, wobei

$$\begin{aligned}\mathfrak{X} &= \{0, \dots, n\}, \\ \mathfrak{F} &= \mathcal{P}(\mathfrak{X})\end{aligned}$$

und $(\mathbb{P}_\vartheta)_{\vartheta \in [0,1]}$ als Familie der Binomialverteilungen (bei festgehaltenem n für die Anzahl der Versuche), d.h.

$$\mathbb{P}_\vartheta(k) = \binom{n}{k} \vartheta^k (1 - \vartheta)^{n-k}.$$

Im folgenden wollen wir dieses Modell *Binomialmodell* nennen.

Wie in jedem statistischen Modell ist auch hier die identische Kenngröße

$$\tau : \Theta = [0, 1] \rightarrow \mathbb{R}, \quad \tau(\vartheta) = \vartheta,$$

d.h. der Parameter ϑ selbst, von besonderem Interesse. Im Binomialmodell ist die Statistik

$$T : \mathfrak{X} \rightarrow \mathbb{R}, \quad T(x) = \frac{x}{n}$$

ein natürlicher Schätzer für $\tau(\vartheta) = \vartheta$. T ist erwartungstreu, denn

$$\mathbb{E}_\vartheta(T) = \sum_{k=0}^n T(k) \cdot P_\vartheta(k) = \frac{1}{n} \left(\sum_{k=0}^n k \cdot \binom{n}{k} \cdot \vartheta^k \cdot (1 - \vartheta)^{n-k} \right) = \frac{1}{n} (n \cdot \vartheta) = \vartheta = \tau(\vartheta).$$

Zuletzt bemerken wir, dass das Binomialmodell (für n -Versuche) aus dem n -fachen Bernoulli-Modell hervorgeht durch Anwendung der Statistik

$$S : \{0, 1\}^n \mapsto \{0, 1, \dots, n\}, \quad S(x_1, \dots, x_n) := \sum_{i=1}^n x_i.$$

In diesem Sinne ist das Binomialmodell das *Bildmodell* des n -fachen Bernoulli-Modells unter der Statistik S .

Beispiel (Glühbirnen, Forts.). Hier können wir mit dem 100-fachen Produkt $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta)) = M^{\otimes n}$ des statistischen Modells $M = (\mathbb{R}_{\geq 0}, \mathcal{B}(\mathbb{R}_{\geq 0}), (\frac{1}{\vartheta}e^{-\vartheta x}dx)_{\vartheta \geq 0})$ für die Familie der Exponentialverteilungen arbeiten, d.h.

- $\mathfrak{X} = (\mathbb{R}_{\geq 0})^{100}$, $\mathfrak{F} = \mathcal{B}((\mathbb{R}_{\geq 0})^{100})$ ist die Borel'sche Sigma-Algebra auf $\mathbb{R}_{\geq 0}^{100}$ und \mathbb{P}_ϑ ist die 100-fache Produktverteilung der Exponential- ϑ -verteilung, d.h.

$$P_\vartheta(X_1 > t_1, \dots, X_{100} > t_{100}) = e^{-\vartheta \cdot t_1} \cdot \dots \cdot e^{-\vartheta \cdot t_{100}}.$$

- Wir interessieren uns z.B. für die Kenngröße

$$\tau : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}$$

$$\tau(\vartheta) = \frac{1}{\vartheta},$$

- für welche die Statistik

$$S : \mathfrak{X} \rightarrow \Sigma = \mathbb{R}_{\geq 0}$$

$$S(t_1, \dots, t_{100}) = \frac{1}{100} \cdot \sum_{i=1}^{100} t_i$$

wie bereits gesehen ein erwartungstreuer Schätzer ist.

Beispiel (Apfelsinenlieferung). Jemand schenkt uns N Apfelsinen, ϑ davon sind faul. Wir ziehen zufällig $n \leq N$ Apfelsinen, k sei die Anzahl der gezogenen faulen.

- $\mathfrak{X} = \{0, \dots, n\}$, $\mathfrak{F} = \mathcal{P}(\mathfrak{X})$, $(\mathbb{P}_\vartheta; \vartheta \in \{0, \dots, N\})$ mit

$$\mathbb{P}_\vartheta(k) = \begin{cases} \frac{\binom{\vartheta}{k} \cdot \binom{N-\vartheta}{n-k}}{\binom{N}{n}} & \text{falls } k \leq \vartheta \\ 0 & \text{sonst.} \end{cases}$$

- Die Statistik $T : \mathfrak{X} \rightarrow \mathbb{R}$, $T(x) = \frac{x}{n}$ ist ein erwartungstreuer Schätzer für ϑ (Übung).

Beispiel (Game-Show). In einer TV-Sendung mit zwei Kandidaten liefert ein Apparat zehn Zufallszahlen x_1, \dots, x_{10} aus einem Intervall $[0, L]$. Der Parameter L ist nur dem Moderator bekannt. Die Kandidaten sollen nun L möglichst gut raten.

Das statistische Modell für die einfache Ziehung einer auf $[0, L]$ uniform verteilten Zufallszahl ist

$$\mathfrak{X} = \mathbb{R}_{\geq 0}, \mathfrak{F} = \mathfrak{B}(\mathbb{R}_{\geq 0}), \mathbb{P}_L = U_{[0, L]}, L \geq 0,$$

wobei $U_{[0, L]}$ die Gleichverteilung auf $[0, L]$ bezeichnet

$$U_{[0, L]}(X < t) = \min\left(\frac{t}{L}, 1\right).$$

Somit lautet das Modell für die 10-fache Wiederholung

$$(\mathbb{R}_{\geq 0}^{10}, \mathcal{B}(\mathbb{R}_{\geq 0}^{10}), (U_{[0, L]}^{\otimes n})_{L \in [0, +\infty)}).$$

Gefragt ist nach der Kenngröße

$$\tau : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}, \quad L \mapsto L.$$

Die beiden Kandidaten haben nun verschiedene Ratestrategien/Schätzer. Kandidat A verwendet den Schätzer

$$\widehat{L}(\vec{x}) = \frac{2}{n} \cdot \sum_{i=1}^{10} x_i,$$

und Kandidat B benutzt

$$\widehat{M}(\vec{x}) = \max(x_1, \dots, x_{10}).$$

Die Frage, welcher von beiden die bessere Strategie hat, beantworten wir in Abschnitt 2.1.4.

2.1.2 Konsistenz von Schätzern

Die Erwartungstreue von Schätzern ist eine wünschenswerte Eigenschaft insofern als der vorgeschlagene Schätzwert $\widehat{\tau}$ im Mittel der gesuchten Kenngröße entspricht. Das Gesetz der großen Zahlen sagt dann, dass der Mittelwert der Schätzwerte $\frac{1}{k} \sum_{i=1}^k \widehat{\tau}(X_i)$ von k unabhängigen Wiederholungen desselben Schätzvorganges mit $k \rightarrow \infty$ fast sicher gegen den gesuchten Wert $\tau(\vartheta)$ konvergiert. Wenn der einzelne Schätzvorgang dabei jedoch sehr aufwendig ist, kann es aber ineffizient sein, davon viele unabhängig voneinander durchzuführen und allein den Mittelwert der Einzelschätzungen zu bilden.

Beispiel (Game-Show, Forts.). Der Schätzer \widehat{L} des Kandidaten A ist der Mittelwert der Einzelschätzer $\widehat{L}_i = 2X_i$ und ist offensichtlich erwartungstreu. $\widehat{M} = \widehat{M}^n$ könnte auch rekursiv auf der Folge der Modelle definiert werden als

$$\widehat{M}^1 = \widehat{L}^1 \quad \text{und} \quad \widehat{M}^{n+1} = \max(\widehat{M}^n, X_n).$$

Die Struktur von \widehat{M}^n ist also komplizierter als die von \widehat{L} als Mittelwert von unabhängigen Einzelschätzern. In Abschnitt 2.1.4 werden wir sehen, dass \widehat{M} in der Tat besser abschneidet als \widehat{L} . Zugleich ist \widehat{M} nicht erwartungstreu. Hierzu berechnen wir zunächst seine Verteilung gemäß

$$\mathbb{P}_L^n(\widehat{M} < t) = \begin{cases} \prod_{i=1}^n \mathbb{P}_L(X_i \leq t) = \left(\frac{1}{L}\right)^n t^n & \text{falls } t \leq L \\ 1 & \text{sonst.} \end{cases}$$

Entsprechend finden wir für den Erwartungswert (mit $G(t) := \mathbb{P}_L^n(\widehat{M} \geq t)$)

$$\mathbb{E}_L(\widehat{M}) = \int_0^\infty G(x) dx = \int_0^L \left(1 - \left(\frac{t}{L}\right)^n\right) dt = \frac{n}{n+1} L \neq L.$$

Die Menge $\widehat{M} = \widehat{M}^n$, $n \in \mathbb{N}$ aufgefasst als Folge von Schätzern, konsistent für die Kenngröße L in dem Sinne, dass die Wahrscheinlichkeit für einen Schätzfehler mit $n \rightarrow \infty$ asymptotisch verschwindet.

Definition 2.2. Sei $M^n := (\mathfrak{X}^n, \mathfrak{F}^n, (\mathbb{P}_\vartheta^n; \vartheta \in \Theta))$ eine Folge von statistischen Modellen und $\tau : \Theta \rightarrow \mathbb{R}$ eine Kenngröße, sowie $T^n : \mathfrak{X}^n \rightarrow \mathbb{R}$. Dann heißt die Folge (T^n) eine *konsistente Schätzfolge* für τ , falls

$$\forall \varepsilon > 0 \forall \vartheta \in \Theta : \mathbb{P}_\vartheta^n(|T^n(X) - \tau(\vartheta)| > \varepsilon) \rightarrow 0$$

Beispiel (Forts.). Die Folge der Schätzer $\widehat{M}^n := \max(X_1, \dots, X_n)$ definiert auf der Folge der Modelle $M_n = (R_{\geq 0}^n, \mathfrak{B}(R_{\geq 0}^n), (\mathbb{P}_L^n)_{L \geq 0})$ ist konsistent, denn für $\varepsilon > 0$ finden wir

$$\mathbb{P}_L^n(|\widehat{M}^n - L| > \varepsilon) = \mathbb{P}_L^n(\widehat{M}^n < L - \varepsilon) = \left(\frac{L - \varepsilon}{L}\right)^n \xrightarrow{n \rightarrow \infty} 0.$$

Das folgende Lemma gibt ein einfaches Kriterium für die Konsistenz einer Folge von Schätzern.

Satz 2.3. Falls in der Situation aus Definition 2.2 für alle $\vartheta \in \Theta$

- 1) $\mathbb{E}_{\mathbb{P}_\vartheta^n}(T^n(X)) \xrightarrow{n \rightarrow \infty} \tau(\vartheta)$
- 2) $\sup_{n \geq 0} \mathbb{V}_{\mathbb{P}_\vartheta^n}(T^n(X)) \rightarrow 0$,

so ist $(T^n)_n$ eine konsistente Folge von Schätzern.

Beweis. Setze $\tau^n(\vartheta) := \mathbb{E}_{\mathbb{P}_\vartheta^n}(T^n(X))$. Zu $\varepsilon > 0$ gilt

$$\begin{aligned} \mathbb{P}_\vartheta^n(|T^n(X) - \tau(\vartheta)| > \varepsilon) &\leq \mathbb{P}_\vartheta^n\left(|T^n(X) - \tau^n(\vartheta)| > \frac{\varepsilon}{2}\right) + \mathbb{P}_\vartheta^n\left(|\tau^n(\vartheta) - \tau(\vartheta)| > \frac{\varepsilon}{2}\right) \\ &= \mathbb{P}_\vartheta^n\left(|T^n(X) - \tau^n(\vartheta)| > \frac{\varepsilon}{2}\right) + \mathbb{1}_{|\tau^n(\vartheta) - \tau(\vartheta)| > \frac{\varepsilon}{2}} \\ &\leq \underbrace{\frac{1}{(\varepsilon/2)^2} \mathbb{V}_{\mathbb{P}_\vartheta^n}(T^n(X))}_{\rightarrow 0, \text{ nach 2)}} + \underbrace{\mathbb{1}_{|\tau^n(\vartheta) - \tau(\vartheta)| > \frac{\varepsilon}{2}}}_{\rightarrow 0, \text{ nach 1)}}. \quad \square \end{aligned}$$

2.1.3 Maximum-Likelihood-Schätzer

Definition 2.4 (Standardmodell). Das Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ heißt ein *Standardmodell*, falls ein *dominierendes Maß* μ_0 auf $(\mathfrak{X}, \mathfrak{F})$ existiert, so dass $\mathbb{P}_\vartheta(dx) \ll \mu_0(dx)$ für alle $\vartheta \in \Theta$. Die zugehörigen Radon-Nikodym-Dichten $\varrho(\vartheta, \cdot) = \frac{d\mathbb{P}_\vartheta}{d\mu_0} : \mathfrak{X} \rightarrow \mathbb{R}$, d.h. mit

$$\mathbb{P}_\vartheta(dx) = \varrho(\vartheta, x) \mu_0(dx)$$

heißen *Likelihood-Funktionen*.

Bemerkung. $\varrho(\vartheta, x)$ ist also die Wahrscheinlichkeit(-sdichte relativ μ_0), die Stichprobe x unter dem Wahrscheinlichkeitsmaß \mathbb{P}_ϑ zu beobachten.

Bemerkung (Invarianz gegenüber der Wahl eines dominierenden Maßes). Jedes andere zu μ_0 äquivalente Maß $\tilde{\mu}$ (d.h. mit $\tilde{\mu} \ll \mu_0$ und $\mu_0 \ll \tilde{\mu}$) ist ebenfalls ein dominierendes Maß, und mit der relativen Dichte $h = d\mu_0/d\tilde{\mu}$ gilt dann offensichtlich, dass $\mathbb{P}_\vartheta(dx) = \tilde{\rho}(\vartheta, x)\tilde{\mu}(dx)$ mit $\tilde{\rho}(\vartheta, x) = \varrho(\vartheta, x) \cdot h(x)$. – Alle in dieser Vorlesung genannten Resultate für Standardmodelle sind unabhängig von der genauen Wahl eines dominierenden Maßes μ_0 .

Beispiele.

- Game-Show:

$$\mathbb{P}_\vartheta(dx) = \underbrace{\frac{1}{\vartheta} \cdot \mathbb{1}_{[0,\vartheta]}(x)}_{=\varrho(\vartheta,x)} \cdot \lambda(dx)$$

- Binomialmodell:

$$\mathbb{P}_\vartheta(n, \vartheta, k) = \binom{n}{k} \cdot \vartheta^k (1 - \vartheta)^{n-k} = \varrho(\vartheta, k),$$

d.h. $\mathbb{P}_\vartheta(dx) = \varrho(\vartheta, x) \cdot \mu_0(dx)$, wobei $\mu_0(dx)$ das Zählmaß auf \mathbb{N}_0 ist.

Für die Konstruktion eines Schätzers von ϑ selbst (d.h. $\tau(\vartheta) = \vartheta$) in einem regulären Modell besteht ein intuitiver Ansatz darin, bei gegebener Stichprobe x ein $\hat{\vartheta} = \hat{\vartheta}(x)$ als Schätzwert vorzuschlagen, für welches die Wahrscheinlichkeit(sdichte) $\varrho(\vartheta, x)$ für die Beobachtung x unter allen $\vartheta \in \Theta$ maximiert wird.

Definition 2.5. Es sei $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta))$ ein Standardmodell mit Likelihood-Funktion $\varrho : \Theta \times \mathfrak{X} \rightarrow \mathbb{R}$. Ein Schätzer $T : \mathfrak{X} \rightarrow (\Theta, \mathfrak{G})$ für den Parameter ϑ heißt *Maximum-Likelihood-Schätzer*, falls

$$\varrho(T(x), x) = \sup_{\vartheta \in \Theta} \varrho(\vartheta, x) \quad \forall x \in \mathfrak{X}.$$

Beispiel (Binomialmodell). Hier ist die Likelihood-Funktion

$$\varrho(\vartheta, x) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x}$$

$$T(x) = \frac{x}{n}.$$

Aufgrund der Monotonie des natürlichen Logarithmus \log können wir genauso nach Maximalstellen der *log-likelihood Funktion* $\log \circ \varrho$ suchen, die sich im aktuellen Fall schreibt als

$$\log(\varrho(\vartheta, x)) = \log \binom{n}{x} + x \log \vartheta + (n - x) \log(1 - \vartheta).$$

Ableiten nach ϑ ergibt

$$\begin{aligned} \frac{\partial}{\partial \vartheta} (\log \cdot \varrho) &= \frac{x}{\vartheta} + \frac{n-x}{1-\vartheta} \stackrel{!}{=} 0 \\ \Rightarrow \hat{\vartheta} &= \frac{x}{n} = T(x). \end{aligned}$$

Somit ist der zuvor bereits gefundene Schätzer T ein Maximum-Likelihood-Schätzer.

Beispiel (n -faches Gauß-Produktmodell). Im statistischen Modell $(\mathbb{R}^n, \mathfrak{B}^n, (\nu_{m,v}^{\otimes n}; m \in \mathbb{R}, v \geq 0))$ mit

$$\nu_{m,v}(dx) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}} dx$$

$$\Theta = \mathbb{R} \times \mathbb{R}_{\geq 0}$$

der n -fachen Wiederholung eines Gauß'schen Experiments mit unbekanntem Parametern m und v für Erwartungswert und Varianz ist die likelihood-Funktion

$$\varrho((m, v), (x_1, \dots, x_n)) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x_i - m)^2}{2v}}.$$

Also ergibt sich die log-likelihood-Funktion

$$-\sum_{i=1}^n \frac{1}{2} \log(2\pi v) - \sum_{i=1}^n \frac{(x_i - m)^2}{2v} =: \eta(m, v).$$

Der ML-Schätzer für $\vartheta = (m, v)$ entspricht der Maximalstelle (\hat{m}, \hat{v}) der obigen log-likelihood-Funktion bei festem $\vec{x} = (x_1, \dots, x_n)$. Hierzu suchen wir die Nullstellen (\hat{m}, \hat{v}) von $\nabla \eta = \left(\frac{\partial \eta}{\partial m}, \frac{\partial \eta}{\partial v}\right)$ und erhalten

$$\hat{m} = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \hat{m}(x) = \bar{x}, \quad \hat{v} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \hat{v}(x).$$

Man überzeugt sich leicht, dass dies in der Tat eine Maximalstelle ist, d.h. (\hat{m}, \hat{v}) ist ein ML-Schätzer für den unbekanntem Parameter (m, v) .

Satz 2.6. *Im n -fachen Gauß-Produktmodell*

$$(\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n), (\nu_{m,v}^{\otimes n}; m \in \mathbb{R}, v \geq 0))$$

ist

$$\hat{m}(x) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

ein ML-Schätzer für m und

$$\hat{v}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

ein ML-Schätzer für v . Der Schätzer \hat{m} ist erwartungstreu, der Schätzer \hat{v} nicht, denn

$$\mathbb{E}_{(m,v)}(\hat{v}) = \frac{n-1}{n} \cdot v \neq v.$$

Beweis. Es wurde bereits gezeigt, dass $(\hat{m}, \hat{v}) =: \hat{\vartheta}$ ein ML-Schätzer für $\vartheta = (m, v)$ ist. Bleibt die Erwartungstreue zu untersuchen. Für (m, v) beliebig ist

$$\mathbb{E}_{(m,v)}(\hat{m}) = \mathbb{E}_{(m,v)}\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{(m,v)}(x_i) = \frac{1}{n} \sum_{i=1}^n m = m.$$

Ebenso rechnen wir den Erwartungswert von \hat{v}

$$\mathbb{E}_{(m,v)}(\hat{v}) = \mathbb{E}_{(m,v)}\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{(m,v)}((x_i - \bar{x})^2).$$

Mit

$$\mathbb{E}_{(m,v)} \left(\underbrace{(x_i - m)}_{=: z_i} - \underbrace{(\bar{x} - m)}_{=: \bar{z}} \right)^2 = \mathbb{E}_{(m,v)} ((z_i)^2 - 2 \cdot z_i \cdot \bar{z} + (\bar{z})^2)$$

gilt

$$\begin{aligned} \mathbb{E}_{(m,v)} ((z_i)^2) &= \mathbb{E}_{(m,v)} ((x_i - m)^2) = v, \\ \mathbb{E}_{(m,v)} (z_i \bar{z}) &= \frac{1}{n} \mathbb{E}_{(m,v)} \left(\sum_{j=1}^n z_i z_j \right) = \frac{1}{n} \sum_{j=1}^n \underbrace{\mathbb{E}_{(m,v)} (z_i z_j)}_{=0 \text{ für } i \neq j} = \frac{v}{n}, \end{aligned}$$

und

$$\mathbb{E}_{(m,v)} ((\bar{z})^2) = \frac{1}{n^2} \mathbb{E}_{(m,v)} \left(\sum_{i,j} z_j z_j \right) = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_{(m,v)} ((z_i)^2) = \frac{1}{n^2} \cdot n \cdot v = \frac{v}{n}.$$

Damit folgt

$$\begin{aligned} \mathbb{E}_{(m,v)} ((x_i - \bar{x})^2) &= v - \frac{2v}{n} + \frac{v}{n} = \left(1 - \frac{1}{n}\right) v, \\ \Rightarrow \mathbb{E}_{(m,v)} (\hat{v}) &= \frac{1}{n} \cdot n \cdot \left(1 - \frac{1}{n}\right) v = \frac{n-1}{n} \cdot v. \end{aligned} \quad \square$$

Bemerkung. Durch Reskalieren

$$\tilde{v}(x) := \frac{n}{n-1} \hat{v}(x)$$

erhalten wir einen neuen, erwartungstreuen Schätzer \tilde{v} für v .

2.1.4 Mittlerer quadratischer Fehler

Zum Vergleich von Schätzern führt man Genauigkeits- bzw. Fehlermaße ein.

Definition 2.7. Seien $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein statistisches Modell, $\tau : \Theta \rightarrow \mathbb{R}$ eine Kenngröße und $T : \mathfrak{X} \rightarrow \mathbb{R}$ ein Schätzer für τ .

1. Dann heißt

$$\mathbb{B}_\vartheta(T) := |\mathbb{E}_\vartheta(T) - \tau(\vartheta)|$$

systematischer Fehler des Schätzers T (engl. *Bias*).

2. Die Größe

$$\mathbb{V}_\vartheta(T) := \mathbb{E}_\vartheta((T - \mathbb{E}_\vartheta(T))^2)$$

ist die *Varianz* (bzw. *Streuung*) von T .

3. und

$$\mathbb{F}_\vartheta(T) := \mathbb{E}_\vartheta((T - \tau(\vartheta))^2)$$

heißt (*mittlerer quadratischer*) *Fehler* von T .

Bemerkung.

1. Gelegentlich werden auch andere Fehlermaße studiert, etwa von der Form

$$\mathbb{F}_\vartheta^\Psi(T) = \mathbb{E}_\vartheta[\Psi(\tau(\vartheta) - T(X))]$$

mit einer gewissen 'Fehlerfunktion' $\Psi : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$.

2. Der mittlere quadratische Fehler spielt (ähnlich wie die verwandten Fehlermaße $\mathbb{F}_\vartheta^\Psi(T)$ mit $\Psi(t) = t^p$) eine wichtige Rolle bei der Anwendung der Tschebyschev-Ungleichung in der Form

$$\mathbb{P}_\vartheta(|T(X) - \tau(\vartheta)| \geq \epsilon) \leq \frac{1}{\epsilon^2} \mathbb{F}_\vartheta(T).$$

D.h. je kleiner $\mathbb{F}_\vartheta(T)$ umso höher ist die Wahrscheinlichkeit, dass der Schätzwert $T(x)$ nahe bei $\tau(\vartheta)$ liegt.

3. Offensichtlich gilt

$$T \text{ erwartungstreu} \Leftrightarrow \mathbb{B}_\vartheta(T) = 0 \quad \forall \vartheta \in \Theta.$$

Allgemein finden wir die Zerlegung

$$\mathbb{F}_\vartheta(T) = (\mathbb{B}_\vartheta(T))^2 + \mathbb{V}_\vartheta(T)$$

Demnach entspricht der Fehler von T der Summe aus dem quadrierten systematischen Fehler und der Streuung.

Beispiel. Im Binomialmodell

$$(\{0, \dots, n\}, \mathcal{P}(\{0, \dots, n\}), (B(n, \vartheta); \vartheta \in [0, 1]))$$

ist die Statistik $T(x) := \frac{x}{n}$ ein erwartungstreuer (ML-)Schätzer für $\tau(\vartheta) = \vartheta$, denn

$$\mathbb{E}_\vartheta(T) = \vartheta \quad \forall \vartheta \in [0, 1].$$

Für den quadratischen Fehler erhalten wir somit

$$\begin{aligned} \mathbb{F}_\vartheta(T) &= \underbrace{(\mathbb{B}_\vartheta(T))^2}_{=0} + \mathbb{V}_\vartheta(T) = \text{Var}_\vartheta\left(\frac{x}{n}\right) = \frac{1}{n^2} \text{Var}_\vartheta(x) \\ &= \frac{1}{n^2} \cdot n \cdot \vartheta \cdot (1 - \vartheta) = \frac{\vartheta \cdot (1 - \vartheta)}{n} \end{aligned} ,$$

folglich

$$\Rightarrow \mathbb{F}_\vartheta \leq \frac{1}{4n} \quad \forall \vartheta \in [0, 1].$$

Wir wollen nun als alternativen Schätzer betrachten

$$S(x) = \frac{x+1}{n+2},$$

dessen systematischer Fehler

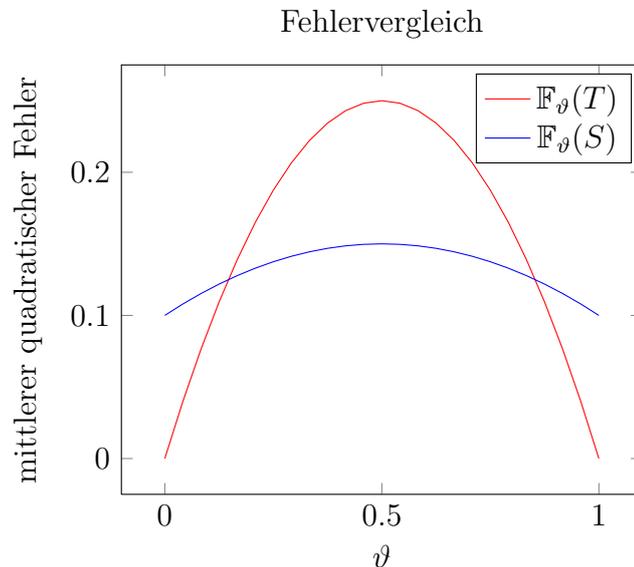
$$\mathbb{B}_\vartheta(S) = \mathbb{E}_\vartheta \left(\frac{x+1}{n+2} \right) - \vartheta = \frac{\mathbb{E}_\vartheta(x) + 1}{n+2} - \vartheta = \frac{n \cdot \vartheta + 1}{n+2} - \vartheta = \frac{1-2\vartheta}{n+2}$$

beträgt (insbesondere ist S also nicht erwartungstreu). Seine Varianz lässt sich ebenfalls leicht ausrechnen

$$\mathbb{V}_\vartheta(S) = \text{Var}_\vartheta \left(\frac{x+1}{n+2} \right) = \frac{1}{(n+2)^2} \cdot n \cdot \vartheta \cdot (1-\vartheta)$$

Für den mittleren quadratischen Fehler erhalten wir also

$$\mathbb{F}_\vartheta(S) = (\mathbb{B}_\vartheta(S))^2 + \mathbb{V}_\vartheta = \left(\frac{1-2\vartheta}{n+2} \right)^2 + \frac{n\vartheta(1-\vartheta)}{(n+2)^2}.$$



Im Vergleich ergibt sich, dass für den Parameterbereich $\left\{ \left| \vartheta - \frac{1}{2} \right|^2 \leq \left(1 + \frac{1}{n}\right) \vartheta(1+\vartheta) \right\}$ im Sinne des mittleren quadratischen Fehlers S ein besserer Schätzer ist als T . Grund hierfür ist die große Streuung des Schätzers T , die ihn ungeachtet seiner Erwartungstreue ungenau macht.

Beispiel (Game-Show, Forts.). Wir wollen zeigen, dass der Schätzer $\widehat{M} = \max(X_1, \dots, X_n)$ für den unbekannt Parameter L besser ist als $\widehat{L} = \frac{2}{n} \sum_{i=1}^n X_i$. Der Schätzer \widehat{L} ist erwartungstreu. Also stimmen Varianz und mittlerer quadratischer Fehler überein. Die Varianz einer auf $[0, L]$ gleichverteilten Zufallsgröße ist

$$\mathbb{V}_L(X_1) = \mathbb{E}_L X_1^2 - (\mathbb{E}(X_1))^2 = \frac{1}{L} \int_0^L x^2 dx - \left(\frac{L}{2}\right)^2 = \frac{L^2}{12}.$$

Somit ist

$$\mathbb{F}_L(\widehat{L}) = \frac{4}{n} \frac{L^2}{12} = \frac{L^2}{3n}.$$

Den Erwartungswert von \widehat{M} hatten wir bereits bestimmt als $\mathbb{E}_L(\widehat{M}) = \frac{n}{n+1}L$, somit ist

$$(\mathbb{B}_L(\widehat{M}))^2 = \frac{L^2}{(n+1)^2}.$$

Zur Berechnung der Varianz benutzen wir die Formel

$$\mathbb{E}(X^p) = p \int_0^\infty t^{p-1} \mathbb{P}(X \geq t) dt$$

für die Berechnung des p -ten Moments einer nichtnegativen Zufallsvariable. Wir hatten bereits gesehen, dass $G(t) := \mathbb{P}_L^n(\widehat{M} \geq t) = 1 - (\frac{t}{L})^n$, folglich

$$\mathbb{E}_L(\widehat{M}^2) = 2 \int_0^\infty t \cdot G(t) dt = 2 \int_0^L t(1 - (\frac{t}{L})^n) dt = \frac{n}{n+2}L^2.$$

Für den Fehler von \widehat{M} erhalten wir damit

$$\begin{aligned} \mathbb{F}_L(\widehat{M}) &= \mathbb{V}_L(\widehat{M}) + [\mathbb{B}_L(\widehat{M})]^2 = \mathbb{E}_L(\widehat{M}^2) - [\mathbb{E}_L(\widehat{M})]^2 + [\mathbb{B}_L(\widehat{M})]^2 \\ &= L^2 \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2} + \frac{1}{(n+1)^2} \right) = \frac{2}{(n+2)(n+1)}L^2. \end{aligned}$$

Der mittlere quadratische Fehler von \widehat{M} fällt im Gegensatz zu \widehat{L} sogar quadratisch in n .

2.1.5 Fisher-Information und Cramer-Rao Ungleichung

Definition 2.8. Seien $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein statistisches Modell und T ein erwartungstreuer Schätzer für eine Kenngröße $\tau : \Theta \rightarrow \mathbb{R}$. Dann heißt T *varianzminimierend* (bzw. *gleichmäßig bester Schätzer* (oder UMV für 'uniformly minimizing variance'), wenn für alle erwartungstreuen Schätzer $S : \mathfrak{X} \rightarrow \mathbb{R}$ gilt:

$$\forall \vartheta \in \Theta : \mathbb{V}_\vartheta(T) \leq \mathbb{V}_\vartheta(S).$$

Unter den erwartungstreuen Schätzern für eine Kenngröße sind also die mit minimaler Varianz vorzuziehen. Die nun folgende Cramer-Rao-Schranke gibt eine allgemeine (aber in vielen Fällen noch immer zu optimistische) Abschätzung, wie klein die Varianz bestenfalls sein kann.

Definition 2.9. Sei $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein Standardmodell. Dann heißt $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ *regulär*, falls folgende Bedingungen gelten:

- i) $\Theta \subseteq \mathbb{R}$ ist ein offenes Intervall.
- ii) $\varrho(\vartheta, \cdot) : \mathfrak{X} \rightarrow \mathbb{R}$ ist strikt positiv auf $\text{spt}(\mu_0)$ für alle $\vartheta \in \Theta$.
- iii) Falls $S : \mathfrak{X} \mapsto \mathbb{R}$ mit $\mathbb{E}_{\vartheta_0}(S^2(X)) < \infty$ für ein $\vartheta_0 \in \Theta$, so gilt

$$\frac{d}{d\vartheta} \Big|_{\vartheta=\vartheta_0} \int_{\mathfrak{X}} S(x) \varrho(\vartheta, x) \mu_0(dx) = \int_{\mathfrak{X}} S(x) \frac{\partial}{\partial \vartheta} \Big|_{\vartheta=\vartheta_0} \varrho(\vartheta, x) \mu_0(dx).$$

Satz 2.10 (Cramer-Rao Informationsungleichung). *In einem regulären statistischen Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ sei $\tau : \Theta \rightarrow \mathbb{R}$ eine C^1 -Kenngröße und $T : \mathfrak{X} \rightarrow \mathbb{R}$ ein erwartungstreuer Schätzer für τ , dann gilt*

$$\mathbb{V}_\vartheta(T) \geq \frac{|\tau'(\vartheta)|^2}{I(\vartheta)} \quad \forall \vartheta \in \Theta,$$

wobei

$$I(\vartheta) = \int_{\mathfrak{X}} \left(\frac{\partial}{\partial \vartheta} \log \varrho(\vartheta, x) \right)^2 \varrho(\vartheta, x) \mu_0(dx).$$

Beweis. O.B.d.A. sei $\mathbb{V}_\vartheta(T) < \infty$, denn anderenfalls ist nichts zu beweisen. Wir definieren die Statistik $\mathfrak{X} \ni x \mapsto u_\vartheta(x) := \frac{\partial}{\partial \vartheta} \log(\varrho(\vartheta, x))$, dann liefert die Anwendung der Regularitätseigenschaft iii) auf die konstante Statistik $\mathfrak{X} \ni x \mapsto S(x) = 1$, dass

$$\begin{aligned} \mathbb{E}_\vartheta(u_\vartheta) &= \int_{\mathfrak{X}} \frac{\partial}{\partial \vartheta} \log(\varrho(\vartheta, x)) \cdot \varrho(\vartheta, x) \mu_0(dx) \\ &= \int_{\mathfrak{X}} \frac{\frac{\partial}{\partial \vartheta} \varrho(\vartheta, x)}{\varrho(\vartheta, x)} \varrho(\vartheta, x) \mu_0(dx) = \frac{\partial}{\partial \vartheta} \underbrace{\int_{\mathfrak{X}} \varrho(\vartheta, x) \mu_0(dx)}_{=1 \quad \forall \vartheta \in \Theta} = 0 \end{aligned}$$

Somit erhalten wir

$$\begin{aligned} \text{Cov}_\vartheta(T, u_\vartheta) &= \mathbb{E}_\vartheta(T \cdot u_\vartheta) = \int_{\mathfrak{X}} T(x) \cdot \frac{\partial}{\partial \vartheta} \varrho(\vartheta, x) \mu_0(dx) \\ &= \frac{\partial}{\partial \vartheta} \int_{\mathfrak{X}} T(x) \cdot \varrho(\vartheta, x) \mu_0(dx) = \frac{\partial}{\partial \vartheta} \mathbb{E}_\vartheta(T) = \frac{\partial}{\partial \vartheta} \tau(\vartheta) = \tau'(\vartheta) \end{aligned}$$

wobei wir beim drittletzten Schritt wieder die Regularitätseigenschaft iii) mit $S = T$ ausgenutzt haben. Andererseits gilt mit der Cauchy-Schwarz-Ungleichung, dass

$$\text{Cov}(T, u_\vartheta) \leq \sqrt{\mathbb{V}_\vartheta(T)} \cdot \sqrt{\mathbb{V}_\vartheta(u_\vartheta)} = \sqrt{\mathbb{V}_\vartheta(T)} \cdot \sqrt{I(\vartheta)}.$$

Die Behauptung ergibt sich nun durch Quadrieren und Umstellen nach $\mathbb{V}_\vartheta(T)$. \square

Bemerkung. Die Funktion

$$\begin{aligned} u &: \Theta \times \mathfrak{X} \rightarrow \mathbb{R} \\ u(\vartheta, x) &= \frac{\partial}{\partial \vartheta} \log(\varrho(\vartheta, x)) = \frac{\frac{\partial}{\partial \vartheta} \varrho(\vartheta, x)}{\varrho(\vartheta, x)} \end{aligned}$$

wird auch 'Score-Funktion' genannt, die Funktion $\Theta \ni \vartheta \rightarrow I(\vartheta) \geq 0$ heißt 'Fischer-Information'. $I(\cdot)$ hängt nur vom statistischen Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ab (und nicht von τ oder T).

Definition 2.11. Sei $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein reguläres statistisches Modell, $\tau : \Theta \rightarrow \mathbb{R}$ eine Kenngröße und $T : \mathfrak{X} \rightarrow \mathbb{R}$ ein erwartungstreuer Schätzer für τ . Dann heißt T *Cramer-Rao optimal*, falls

$$\mathbb{V}_\vartheta(T) = \frac{|\tau'(\vartheta)|^2}{I(\vartheta)} \quad \forall \vartheta \in \Theta$$

Bemerkung. Ein Cramer-Rao optimaler Schätzer ist somit insbesondere UMV.

Lemma 2.12 (Rechenregel).

$$I(\vartheta) = \mathbb{E}_{\vartheta} \left(-\frac{\partial^2}{\partial \vartheta^2} \log(\varrho(\vartheta, \cdot)) \right).$$

Beweis. Übungsaufgabe. □

Lemma 2.13 (Additivität der Fischer-Information). *Es sei $M = (\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_{\vartheta}; \vartheta \in \Theta))$ ein reguläres statistisches Modell mit Fischer-Information $I_M : \Theta \rightarrow \mathbb{R}_{\geq 0}$. Dann gilt für das Produktmodell $M^{\otimes n}$:*

$$I_{M^{\otimes n}}(\vartheta) = n \cdot I_M(\vartheta).$$

Beweis. Übungsaufgabe. □

Beispiel. Im n -fachen Gauß-Produktmodell

$$M = (\mathbb{R}^n, \mathfrak{B}^n, (\nu_{0,v}^{\otimes n}; v \in (0, +\infty)))$$

mit unbekanntem Parameter v ist

$$T := \sqrt{\frac{\pi}{2}} \cdot \frac{1}{n} \cdot \sum_{i=1}^n |x_i|$$

ein erwartungstreuer Schätzer für die Kenngröße $\tau(v) = \sqrt{v} = \sigma$ (also die Standardabweichung). In der Tat gilt (Übung)

$$\mathbb{E}_{(0,v)}(T) = \sqrt{v} = \tau(v) \quad \forall v > 0$$

und ferner

$$\mathbb{V}_{(0,v)}(T) = \left(\frac{\pi}{2} - 1 \right) \frac{v}{n}.$$

Wir bestimmen nun die Cramer-Rao Schranke für τ . Im einfachen Gauß-Modell ist

$$\varrho(v, x) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v}},$$

$$\log(\varrho(v, x)) = -\frac{1}{2} \log(2\pi v) - \frac{x^2}{2v},$$

mit den Ableitungen

$$\frac{\partial}{\partial v} \log(\varrho(v, x)) = -\frac{1}{2v} + \frac{x^2}{2v^2},$$

$$\frac{\partial^2}{\partial v^2} \log(\varrho(v, x)) = \frac{1}{2v} - \frac{x^2}{v^3}.$$

Daher

$$\mathbb{E}_v \left(-\frac{\partial^2}{\partial v^2} \log(\varrho(v, x)) \right) = \frac{1}{2v^2},$$

woraus wir $I(\vartheta) = \frac{n}{2v^2}$ für das n -fache Produktmodell erhalten. Für den Zähler der Cramer-Rao Schranke finden wir

$$\tau'(v) = \frac{1}{2\sqrt{v}},$$

woraus wir für einen beliebigen erwartungstreuen Schätzer \widehat{T} von \sqrt{v} die Schranke erhalten

$$\mathbb{V}_v(\widehat{T}) \geq \frac{\frac{1}{4v}}{\frac{n}{2v^2}} = \frac{1}{2} \frac{v}{n}.$$

Für den aktuellen Schätzer T finden wir

$$\mathbb{V}_v(T) = \left(\frac{\pi}{2} - 1\right) \frac{v}{n} > \frac{1}{2} \frac{v}{n},$$

d.h. T ist nicht Cramer-Rao optimal.

Satz 2.14 (Cramer-Rao, scharfe Version). *Sei $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein reguläres statistisches Modell mit $\tau : \Theta \rightarrow \mathbb{R}$ und $T : \mathfrak{X} \rightarrow \mathbb{R}$ ein erwartungstreuer Schätzer für τ . Dann ist die Eigenschaft*

$$\mathbb{V}_\vartheta(T) = \frac{|\tau'(\vartheta)|^2}{I(\vartheta)} \quad \forall \vartheta \in \Theta$$

äquivalent dazu, dass sich die Likelihood-Funktion des Modells in der Form

$$\varrho(\vartheta, x) = e^{a(\vartheta) \cdot T(x) - b(\vartheta)} \cdot h(x) \quad (*)$$

darstellen lässt mit gewissen Funktionen $a, b : \Theta \rightarrow \mathbb{R}$ und messbarem $h : \mathfrak{X} \rightarrow \mathbb{R}$.

Beweis. Im Beweis der Cramer-Rao Schranke tauchte nur bei der Verwendung von Cauchy-Schwarz ein Ungleichheitszeichen auf. Es gilt also genau dann Gleichheit, wenn in der Cauchy-Schwarz-Ungleichung Gleichheit gilt. Das ist aber gleichbedeutend mit $(T(x) - \tau(\vartheta)) = c(\vartheta) \cdot u_\vartheta(\vartheta, x)$ für \mathbb{P}_ϑ -fast alle $x \in \mathfrak{X}$ und eine gewisse Konstante $c = c(\vartheta)$, d.h.

$$\begin{aligned} T(x) \frac{1}{c(\vartheta)} - \frac{\tau(\vartheta)}{c(\vartheta)} &= \frac{\partial}{\partial \vartheta} \log(\varrho(\vartheta, x)), \\ \Rightarrow T(x) \underbrace{\int_{\vartheta_0}^{\vartheta} \frac{1}{c(s)} ds}_{=: a(\vartheta)} - \underbrace{\int_{\vartheta_0}^{\vartheta} \frac{\tau(s)}{c(s)} ds}_{=: b(\vartheta)} &= \log(\varrho(\vartheta, x)) - \log(\varrho(\vartheta_0, x)), \\ \Rightarrow e^{T(x)a(\vartheta) - b(\vartheta)} &= \underbrace{\varrho(\vartheta, x) / \varrho(\vartheta_0, x)}_{=: h(x)}. \quad \square \end{aligned}$$

Bemerkung. Wollen wir etwa davon ausgehen, dass $h \equiv 1$, wählen wir als dominierendes Maß $\tilde{\mu}_0(dx) = h(x)\mu_0(dx)$ und schreiben

$$\mathbb{P}_\vartheta(dx) = \varrho(\vartheta, x) \cdot \mu_0(dx) = \underbrace{\varrho(\vartheta, x) / h(x)}_{=: \tilde{\varrho}(\vartheta, x)} \cdot \underbrace{h(x) \cdot \mu_0(dx)}_{=: \tilde{\mu}_0(dx)}$$

Die Fisher-Information bleibt von diesem Wechsel der Darstellung des Modells unberührt. (Übung).

Definition 2.15 (Exponentielle Familie). Ein reguläres statistisches Modell heißt eine *exponentielle Familie*, wenn die Likelihood-Funktion von der Gestalt (*) ist, bzw. äquivalent hierzu nach Wahl eines geeigneten dominierenden Maßes $\tilde{\mu}_0$ gilt

$$\mathbb{P}_\vartheta(dx) = \frac{1}{z(\vartheta)} e^{a(\vartheta) \cdot T(x)} \tilde{\mu}_0(dx),$$

mit gewissen Funktionen $a, z : \Theta \mapsto \mathbb{R}$ und einer Statistik $T : \mathfrak{X} \rightarrow \mathbb{R}$.

Bemerkung. In der obigen Darstellung von \mathbb{P}_ϑ ist $z(\vartheta)$ als *Normierungskonstante* durch $a(\vartheta), T(\cdot), \tilde{\mu}_0$ und die Bedingung $\int_{\mathfrak{X}} \mathbb{P}_\vartheta(dx) = 1$ eindeutig bestimmt, d.h.

$$z(\vartheta) = \int_{\mathfrak{X}} e^{a(\vartheta) \cdot T(x)} \tilde{\mu}_0(dx).$$

Korollar 2.16. In einem exponentiellen Modell mit

$$\varrho(\vartheta, x) = e^{a(\vartheta)T(x) - b(\vartheta)} \cdot h(x)$$

ist $T : \mathfrak{X} \rightarrow \mathbb{R}$ ein (erwartungstreuer) Cramer-Rao optimaler Schätzer für die Kenngröße

$$\tau(\vartheta) = \frac{b'(\vartheta)}{a'(\vartheta)},$$

und in diesem Fall gilt:

$$\mathbb{V}_\vartheta(T) = \frac{\tau'(\vartheta)}{a'(\vartheta)}$$

da $I(\vartheta) = a'(\vartheta) \cdot \tau'(\vartheta)$.

Beweis. Übung. □

Beispiele exponentieller Familien.

1. Normalverteilung mit bekanntem Erwartungswert m :

$$\varrho(v, x) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}}$$

2. Normalverteilung mit bekannter Varianz v :

$$\varrho(m, x) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}} = \underbrace{\frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v}}}_{=: h(x)} \cdot \underbrace{e^{-\frac{m^2}{2v}}}_{=: \frac{1}{z(m)}} \cdot \underbrace{e^{\frac{xm}{v}}}_{=: e^{a(m) \cdot T(x)}}$$

3. Poisson-Verteilung mit dem Zählmaß μ_0 auf \mathbb{N}_0

$$\varrho(\vartheta, x) = e^{-\vartheta} \frac{\vartheta^x}{x!} = e^{(\log \vartheta)x - \vartheta} \cdot \frac{1}{x!}$$

4. Binomialverteilung mit n fest und Parameter $\vartheta \in [0, 1]$, $\mu_0(dx)$ Zählmaß auf $\{0, \dots, n\}$.

$$\varrho(\vartheta, x) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x} = \binom{n}{x} e^{x(\log \vartheta - \log(1-\vartheta))} (1 - \vartheta)^n$$

Bemerkung. Alternativ ist eine exponentielle Familie dadurch gekennzeichnet, dass sich nach geeigneter Wahl von μ_0 so dass $h(\cdot) \equiv 1$ die log-likelihood Funktionen darstellen lassen als eine durch $\vartheta \in \Theta$ parametrisierte Familie von affinen Transformationen einer gewissen Statistik $T : \mathfrak{X} \rightarrow \mathbb{R}$. Im Fall des einfachen Gaußmodells mit unbekanntem Parameter m und v erhalten wir $\mu_0(dx) = dx$ und

$$\log \rho((m, v), x) = \log\left(\frac{1}{\sqrt{2\pi v}}\right) + \frac{(x - m)^2}{2v}$$

Somit ist das einfache Gaußmodell mit unbekanntem Parameter $\vartheta = (m, v)$ keine exponentielle Familie, da es eine Darstellung der Form

$$\frac{1}{2v}(x - m)^2 = \alpha(\vartheta)T(x) + \beta(\vartheta) \quad \forall x, m \in \mathbb{R}, v \geq 0$$

gewissen Funktionen α, β und T nicht geben kann (Übung).

Zusammenfassung

- Unter allen erwartungstreuen Schätzern sind diejenigen mit minimaler Varianz vorzuziehen (bzw. andernfalls die mit minimalem quadratischen Fehler).
- Die Cramer-Rao Schranke ist eine untere Schranke für den minimal möglichen mittleren quadratischen Fehler eines erwartungstreuen Schätzers.
- Ein statistisches Modell ist genau dann eine exponentielle Familie, wenn in der Cramer-Rao Abschätzung für alle $\vartheta \in \Theta$ Gleichheit gilt.
- Im nicht-exponentiellen Fall ist die Cramer-Rao Schranke i.A. zu optimistisch. Dennoch kann man nach varianzoptimalen erwartungstreuen Schätzern fragen. Eine Antwort hierauf gibt der Satz von Rao-Blackwell, den wir als nächstes besprechen werden.

2.1.6 Suffizienz und Satz von Rao-Blackwell

Definition 2.17. Sei $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein statistisches Modell und $S : (\mathfrak{X}, \mathfrak{F}) \rightarrow (\Sigma, \mathfrak{G})$ eine Statistik. Dann heißt S *suffizient*, falls für alle messbaren, beschränkten Funktionen

$$f : \mathfrak{X} \rightarrow \mathbb{R}$$

die bedingte Erwartung

$$\mathbb{E}_\vartheta(f(\cdot) | \sigma(S))$$

nicht mehr von $\vartheta \in \Theta$ abhängt. Dabei bezeichnet $\sigma(S)$ die von S erzeugte σ -Algebra $\sigma(\{S^{-1}(G) \mid G \in \mathfrak{G}\})$.

Bemerkung (Diskreter Spezialfall). Im Falle, dass S eine diskrete Abbildung (also mit diskretem Wertebereich) ist, ist die obige Bedingung äquivalent dazu, dass für alle beschränkt messbaren f und $s \in \Sigma$

$$\mathbb{E}_\vartheta(f(X)|S(X) = s)$$

nicht von ϑ abhängt.

Beispiel. Im n -fachen Bernoulli-Modell $(\{0, 1\}^n, \mathcal{P}(\{0, 1\}^{\otimes n}), B_\vartheta^{\otimes n})$ mit der Bernoulli-Verteilung B_ϑ auf $\{0, 1\}$ zum Erfolgsparameter $\vartheta \in [0, 1]$ ist die Statistik

$$S(x_1, \dots, x_n) = \sum_{i=1}^n x_i$$

suffizient. Zum Nachweis hiervon reicht es, die Bedingung aus 2.1.6 für f von der Form $f(x) = \mathbb{1}_B(x)$ zu überprüfen. Die Bedingung bedeutet dann

$$\mathbb{P}_\vartheta(X \in B|S(X) = s) \text{ hängt nicht von } \vartheta \text{ ab.}$$

Sei also o.B.d.A. $B = \{b = (b_1, \dots, b_n)\} = \{(0, 0, 1, \dots, 1, 0, \dots, 0, 1, \dots)\}$ eine einelementige Menge, dann

$$\begin{aligned} \mathbb{P}_\vartheta(X \in B|S(X) = m) &= \frac{\mathbb{P}_\vartheta(X \in B, S(X) = m)}{\mathbb{P}_\vartheta(S(X) = m)} \\ &= \begin{cases} 0 & , \text{ falls } S(b) \neq m \\ \frac{\mathbb{P}_\vartheta(X \in B)}{\mathbb{P}_\vartheta(S(X) = m)} & , \text{ sonst} \end{cases} \\ &= \begin{cases} 0 & , \text{ falls } S(b) \neq m \\ \frac{\vartheta^m \cdot (1-\vartheta)^{n-m}}{\binom{n}{m} \cdot \vartheta^m \cdot (1-\vartheta)^{n-m}} & , \text{ sonst} \end{cases} \\ &= \begin{cases} 0 & , \text{ falls } S(b) \neq m \\ \frac{1}{\binom{n}{m}} & , \text{ sonst} \end{cases} \end{aligned}$$

Diese Zahl hängt nun in der Tat nicht mehr von ϑ ab, somit ist S eine suffiziente Statistik.

Bemerkung (Anschauliche Bedeutung einer suffizienten Statistik). Durch Bedingen/Festlegen auf das Ergebnis der Hilfsbeobachtung $S = S(X)$ verbleibt zwar noch ein gewisser 'Restzufall' in der Beobachtung X , allerdings hängt dessen Verteilung nicht mehr von ϑ ab. Die Bezeichnung *suffizient* wird durch das folgende Resultat gerechtfertigt, das wir hier nur zur Illustration anfügen.

Satz 2.18. *In einem parametrischen Modell mit suffizienter Statistik S ist das Maß \mathbb{P}_ϑ durch seine Bildverteilung $\mathbb{P}_\vartheta \circ S^{-1}$ unter S festgelegt.*

Beweis. Die Verteilung \mathbb{P}_ϑ ist durch die Integrale $\mathbb{E}_\vartheta(f(X))$ von nichtnegativen messbaren Funktionen $f : \mathfrak{X} \rightarrow \mathbb{R}$ festgelegt. Die bedingte Erwartung $\mathbb{E}_\vartheta(f(X)|\sigma(S))$ ist eine $\sigma(S)$ -messbare Zufallsvariable, und daher darstellbar in der Form

$$\mathbb{E}_\vartheta(f(X)|\sigma(S)) = \tilde{f}_\vartheta(S)$$

mit einer gewissen messbaren Funktion $\tilde{f}_\vartheta : \Sigma \rightarrow \mathbb{R}$. $\mathbb{E}_\vartheta(f(X))$. Aufgrund der Suffizienz von S hängt die Funktion $\tilde{f}_\vartheta = \tilde{f}$ in der Tat nicht vom Parameter ϑ ab. D.h. für jede messbare Funktion $f : \mathfrak{X} \rightarrow \mathbb{R}$ existiert eine messbare Funktion $\tilde{f} : \Sigma \rightarrow \mathbb{R}$, so dass

$$\mathbb{E}_\vartheta(f(X)|\sigma(S)) = \tilde{f}(S) \quad \mathbb{P}_\vartheta - \text{fast sicher.}$$

Mit $\nu_\vartheta := \mathbb{P}_\vartheta \circ S^{-1}$ gilt folglich

$$\mathbb{E}_\vartheta(f(X)) = \mathbb{E}_\vartheta[\mathbb{E}_\vartheta(f(X)|\sigma(S))] = \mathbb{E}_\vartheta(\tilde{f}(S)) = \int_\Sigma \tilde{f}(s)\nu_\vartheta(ds). \quad \square$$

Satz 2.19 (Rao-Blackwell). *Seien $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein statistisches Modell, $\tau : \Theta \rightarrow \mathbb{R}^d$ eine Kenngröße, $S : \mathfrak{X} \rightarrow \Sigma$ eine suffiziente Statistik und $T : \mathfrak{X} \rightarrow \mathbb{R}^d$ ein erwartungstreuer Schätzer für τ . Weiter sei*

$$T' : \mathfrak{X} \rightarrow \mathbb{R}^d, \quad T' = \mathbb{E}(T|\sigma(S)),$$

dann ist T' ebenfalls ein erwartungstreuer Schätzer für τ mit

$$\mathbb{V}_\vartheta(T') \leq \mathbb{V}_\vartheta(T).$$

Dabei gilt genau dann Gleichheit, wenn $T = T'$ \mathbb{P}_ϑ -fast überall gilt.

Bemerkung. Im Spezialfall dass X diskret ist, kann man T' auch beschreiben durch

$$T'(x) = \mathbb{E}_\vartheta(T(X)|S(X) = S(x)).$$

Beweis von 2.19. 1) Aufgrund der Suffizienz von T hängt die Konstruktion von T' nicht von ϑ ab. T' ist somit insbesondere wohldefiniert.

2) T' ist erwartungstreu, denn die Projektivität der bedingten Erwartung liefert

$$\mathbb{E}_\vartheta(T') = \mathbb{E}_\vartheta(\mathbb{E}_\vartheta(T|\sigma(S))) = \mathbb{E}_\vartheta(T) = \tau(\vartheta).$$

3) Mit der Jensen'schen Ungleichung für bedingte Erwartungen in Schritt (*) gilt schließlich

$$\begin{aligned} \mathbb{V}_\vartheta(T') &= \mathbb{E}_\vartheta([T' - \tau(\vartheta)]^2) \\ &= \mathbb{E}_\vartheta([\mathbb{E}_\vartheta(T|\sigma(S)) - \tau(\vartheta)]^2) \\ &= \mathbb{E}_\vartheta([\mathbb{E}_\vartheta(T - \tau(\vartheta)|\sigma(S))]^2) \\ &\stackrel{(*)}{\leq} \mathbb{E}_\vartheta(\mathbb{E}_\vartheta([T - \tau(\vartheta)]^2|\sigma(S))) \\ &= \mathbb{E}_\vartheta([T - \tau(\vartheta)]^2) = \mathbb{V}_\vartheta(T). \end{aligned} \quad \square$$

Bemerkung. Der Satz von Rao-Blackwell liefert ein konstruktives Verfahren zur Verbesserung von erwartungstreuen Schätzern durch Integration entlang der Hilfsbeobachtung S .

Beispiel. Im n -fachen Bernoulli-Modell $(\{0, 1\}^{\otimes n}, \mathcal{P}(\{0, 1\}^{\otimes n}), B_\vartheta^{\otimes n})$ ist

$$S(x_1, \dots, x_n) = \sum_{i=1}^n x_i$$

eine suffiziente Statistik. Weiter sei $\tau(\vartheta) := \vartheta$ und

$$T(x_1, \dots, x_n) := x_1,$$

dann ist T erwartungstreu, denn

$$\mathbb{E}_\vartheta(T) = \mathbb{E}_\vartheta(x_1) = \vartheta.$$

Wir konstruieren einen neuen Schätzer T' aus T vermöge S nach Rao-Blackwell wie folgt.

$$\begin{aligned} T'(x_1, \dots, x_n) &= \mathbb{E}_\vartheta(T | S = \underbrace{S(x_1, \dots, x_n)}_{=:s}) \\ &= \frac{\mathbb{E}_\vartheta(T \cdot \mathbf{1}_{S=s})}{\mathbb{P}_\vartheta(S=s)} \\ &= \frac{\mathbb{P}_\vartheta\left(x_1 = 1, \sum_{i=1}^n x_i = s\right)}{\mathbb{P}_\vartheta(S=s)} \\ &= \frac{\vartheta \cdot \binom{n-1}{s-1} \vartheta^{s-1} \cdot (1-\vartheta)^{n-s}}{\mathbb{P}_\vartheta(S=s)} \\ &= \frac{\binom{n-1}{s-1} \vartheta^s (1-\vartheta)^{n-s}}{\binom{n}{s} \vartheta^s (1-\vartheta)^{n-s}} = \frac{s}{n} \end{aligned}$$

Also haben wir den neuen Schätzer

$$T'(x_1, \dots, x_n) = \frac{\sum_{i=1}^n x_i}{n}.$$

Da

$$\mathbb{V}_\vartheta(T) = \vartheta(1-\vartheta)$$

$$\mathbb{V}_\vartheta(T') = \frac{\vartheta(1-\vartheta)}{n}$$

Ist der Schätzer T' im Sinne des quadratischen Fehlers echt besser als T .

Das folgende praktische Lemma liefert eine Charakterisierung all der statischen Modelle, die eine suffiziente Statistik aufweisen.

Satz 2.20 (Neyman-Fischer Faktorisierungslemma). *Sei $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein reguläres statistisches Modell. Es existiert genau dann eine suffiziente Statistik $S : \mathfrak{X} \rightarrow \Sigma$, wenn*

$$\exists h : \Theta \times \Sigma \rightarrow \mathbb{R} : \varrho(\vartheta, x) = h(\vartheta, S(x)) \cdot k(x) \quad \mu_0\text{-fast überall}$$

und eine für eine gewisse messbare Funktion $k : \mathfrak{X} \rightarrow \mathbb{R}$.

Beweis. "⇒": (Für den Fall, dass \mathfrak{X} diskret ist.) O.B.d.A. ist $\mu_0(dx)$ das Zählmaß.

$$\begin{aligned} \mathbb{P}_\vartheta(X = x | S(X) = s) &= \mathbb{1}_{S(x)=s} \cdot \frac{\mathbb{P}_\vartheta(X = x)}{\mathbb{P}_\vartheta(S(X) = s)} \\ &= \mathbb{1}_{S(x)=s} \cdot \frac{\overbrace{\varrho(\vartheta, x) \mu_0(x)}^{=1}}{\sum_{y \in \mathfrak{X}, S(y)=s} \varrho(\vartheta, y) \cdot \underbrace{\mu_0(y)}_{=1}} \\ &= \mathbb{1}_{S(x)=s} \cdot \frac{\varrho(\vartheta, x)}{\sum_{y \in S^{-1}(s)} \varrho(\vartheta, y)} \stackrel{\text{Suffizienz}}{=} f(x, s) \\ \Rightarrow \mathbb{1}_{S(x)=s} \cdot \varrho(\vartheta, x) &= f(x, s) \cdot \underbrace{\sum_{y \in S^{-1}(s)} \varrho(\vartheta, y)}_{h(\vartheta, s)} \end{aligned}$$

Wahl von $s = S(x)$ führt somit zu

$$\Rightarrow \varrho(x, \vartheta) = f(x, S(x)) \cdot h(\vartheta, S(x)) = k(x) \cdot h(\vartheta, S(x))$$

"⇐": Übung. □

2.1.7 Vollständigkeit und der Satz von Lehmann-Scheffé

Es stellt sich nun die Frage, ob man das Rao-Blackwell Verfahren unbegrenzt iterieren oder ob es irgendwann abbricht? Eine Antwort gibt der Satz von Lehmann-Scheffé. Dafür brauchen wir allerdings noch einen weiteren Begriff.

Definition 2.21. Sei $M = (\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein statistisches Modell. $S : \mathfrak{X} \rightarrow (\Sigma, \mathfrak{G})$ heißt *vollständig*, falls für alle messbaren $h : \Sigma \rightarrow \mathbb{R}$ mit $\mathbb{E}_\vartheta(h^2(X)) < \infty \forall \vartheta \in \Theta$ die Implikation gilt

$$\left(\mathbb{E}_\vartheta(h(S)) = 0 \forall \vartheta \in \Theta \right) \Rightarrow \left(h(S) = 0 \text{ } \mathbb{P}_\vartheta\text{-fast überall } \forall \vartheta \in \Theta \right).$$

Beispiel. Sei M wieder das n -fache Bernoulli-Modell. Die Statistik

$$S(x_1, \dots, x_n) := \sum_{i=1}^n x_i$$

ist vollständig, denn sei $h : \mathbb{Z} \rightarrow \mathbb{R}$ beschränkt, dann gilt

$$\begin{aligned} \mathbb{E}_\vartheta(h(S)) &= \sum_{m=0}^n \binom{n}{m} \vartheta^m (1-\vartheta)^{n-m} h(m) \\ &= (1-\vartheta)^n \cdot \sum_{m=0}^n \binom{n}{m} \left(\frac{\vartheta}{1-\vartheta} \right)^m \cdot h(m) \end{aligned}$$

Somit folgt aus

$$\sum_{m=0}^n \binom{n}{m} \left(\frac{\vartheta}{1-\vartheta} \right)^m h(m) = 0 \quad \forall \vartheta \in]0, 1[$$

durch Variablentransformation $y = \frac{\vartheta}{1-\vartheta}$ dass

$$\sum_{m=0}^n \binom{n}{m} h(m) y^m = 0 \quad \forall y \in \mathbb{R}_{>0}$$

das Nullpolynom in der Variablen y ist, d.h. $h(m) = 0 \quad \forall m$, also $h \equiv 0$.

Satz 2.22 (Lehmann-Scheffé). *Falls $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein statistisches Modell, $\tau : \Theta \rightarrow \mathbb{R}$ eine Kenngröße, $T : \mathfrak{X} \rightarrow \mathbb{R}$ ein erwartungstreuer Schätzer mit endlichem zweiten Moment bzgl. aller $\mathbb{P}_\vartheta, \vartheta \in \Theta$ und $S : \mathfrak{X} \rightarrow (\Sigma, \mathfrak{G})$ eine suffiziente und vollständige Statistik ist, so ist*

$$T' = \mathbb{E}_\vartheta(T \mid \sigma(S))$$

der eindeutige varianzminimierende erwartungstreue Schätzer.

Bemerkung. Es könnte jedoch einen im Sinne des mittleren quadratischen Fehlers besseren nicht-erwartungstreuen Schätzer geben.

Beweis von 2.22. Nach Rao-Blackwell ist T' erwartungstreu und

$$\mathbb{V}_\vartheta(T') \leq \mathbb{V}_\vartheta(T).$$

Sei H ein weiterer erwartungstreuer Schätzer mit $H' = \mathbb{E}(H|S)$. Sowohl T' als auch H' sind dann $\sigma(S)$ -messbare Zufallsvariablen. Somit existieren messbare Funktion $t, h : \Sigma \rightarrow \mathbb{R}$ so, dass

$$H'(x) = h(S(x))$$

$$T'(x) = t(S(x))$$

und

$$0 = \mathbb{E}_\vartheta(H' - T') = \mathbb{E}_\vartheta(h(S(X)) - t(S(X))) = \mathbb{E}_\vartheta((h - t)(S(X))) \quad \forall \vartheta \in \Theta.$$

Aus der Vollständigkeit von S folgt nun

$$(h - t)(S) = 0 \quad \mathbb{P}_\vartheta\text{-fast überall} \quad \forall \vartheta \in \Theta,$$

d.h. $H'(x) = h(S(x)) = t(S(x)) = T'(x)$ \mathbb{P}_ϑ -fast überall. □

Die einmalige Ausführung der Rao-Blackwell-Konstruktion entlang einer suffizienten und vollständigen Statistik S führt somit für verschiedene erwartungstreue Schätzer derselben Kenngröße τ stets auf den gleichen (optimalen) erwartungstreuen Schätzer. Insbesondere können wir noch das folgende Ergebnis festhalten.

Korollar 2.23. *Ein erwartungstreuer Schätzer der Form $T(x) = t(S(x))$ mit einer vollständigen und suffizienten Statistik S ist UMV.*

2.1.8 Der optimale Varianzschätzer im n -fachen Gaußmodell bei unbekanntem Erwartungswert.

Im folgenden geben wir ein Beispiel für einen UMV-Schätzer, der nicht Cramer-Rao optimal ist. Wir arbeiten dazu im n -fachen Gauß-Produktmodell $\mathbb{P}_\vartheta = (\nu_{m,v}^{\otimes n}; m \in \mathbb{R}; v \geq 0)$ auf $(\mathfrak{X}, \mathfrak{F}) = (\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n))$ mit unbekanntem Parameter m und v . Gesucht sei ein Schätzer für den Parameter v . Wie am Ende von Abschnitt 2.1.5 gesehen, ist die scharfe Version des Satzes von Cramer-Rao hier nicht anwendbar. Dennoch gilt das folgende Resultat.

Satz 2.24. *Im n -fachen Gauß-Produktmodell $\mathbb{P}_\vartheta = (\nu_{m,v}^{\otimes n}; m \in \mathbb{R}; v \geq 0)$ auf $(\mathfrak{X}, \mathfrak{F}) = (\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n))$ mit unbekanntem Parameter m und v ist*

$$d_2(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

der gleichmäßig beste Schätzer für v .

Zum Beweis benötigen wir einige Vorbereitungen.

Definition 2.25 (Mehrdimensionale exponentielle Familien). Ein statistisches Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ mit $\Theta \subseteq \mathbb{R}^d$ heißt *mehrdimensionale exponentielle Familie*, falls die Likelihood-Funktionen darstellbar sind in der Form

$$\varrho(\vartheta, x) = \frac{1}{z(\vartheta)} \cdot e^{A(\vartheta) \cdot T(x)}$$

mit $z(\vartheta) \in \mathbb{R}_{>0}$, $A(\vartheta) \in \mathbb{R}^d$ und $T : \mathfrak{X} \rightarrow \mathbb{R}^d$ messbar.

Bemerkung. Nach Umparametrisierung des Modells gemäß sämtlichen auftretenden Vektoren in $\mathfrak{A} := \{A(\vartheta) \mid \vartheta \in \Theta\}$, d.h.

$$\varrho(\vartheta, x) = \tilde{\varrho}(A, x)$$

mit

$$\tilde{\varrho}(A, x) = \frac{1}{\tilde{z}(A)} \cdot e^{A \cdot T(x)}$$

ist dann die Menge der auftretenden Wahrscheinlichkeitsverteilungen auch beschrieben als \mathfrak{X}

$$\{\varrho(\vartheta, x) \mid \vartheta \in \Theta\} = \{\tilde{\varrho}(A, x) \mid A \in \mathfrak{A}\}.$$

Diese Darstellung wird auch *natürliche Parametrisierung* genannt. Von ihr wollen wir im Folgenden ausgehen.

Satz 2.26. *Im mehrdimensionalen Fall einer exponentiellen Familie*

$$\mathbb{P}_\vartheta(dx) = \frac{1}{z(\vartheta)} e^{\vartheta \cdot T(x)} \mu_0(dx)$$

mit $\vartheta \in \Theta \subseteq \mathbb{R}^d$ ist die Statistik

$$T : \mathfrak{X} \rightarrow \mathbb{R}^d$$

suffizient und vollständig, sofern Θ mindestens einen inneren Punkt hat.

Beweis. Die Suffizienz von T folgt aus (einer mehrdimensionalen Fassung von) dem Neyman-Fisher Faktorialisierungslemma (Satz 2.20). Zur Vollständigkeit sei $Q = \{\eta \in \mathbb{R}^d \mid \|\eta - \vartheta_0\|_\infty < \varepsilon\} \subseteq \Theta$ für ein $\vartheta_0 \in \Theta$. Sei also $f : \mathbb{R}^d \rightarrow \mathbb{R}$ beschränkt messbar mit

$$\mathbb{E}_\vartheta[f(T(X))] = 0 \quad \forall \vartheta \in \Theta.$$

Sei $f(x) = f_+(x) - f_-(x)$ die Zerlegung von f in Positiv- und Negativteil. Folglich ist

$$\mathbb{E}_\vartheta(f_+(T(X))) = \int_{\mathfrak{X}} \frac{1}{z(\vartheta)} e^{\vartheta \cdot T(x)} f_+(T(x)) \mu_0(dx).$$

Sei $\nu_0(dt) = (T)^{-1} \mu_0(dx)$ das Bildmaß auf \mathbb{R}^d von μ_0 unter der Abbildung T . Dann gilt

$$\mathbb{E}_\vartheta(f_+(T(X))) = \int_{\mathbb{R}^d} \frac{1}{z(\vartheta)} e^{\vartheta \cdot t} f_+(t) \nu_0(dt)$$

und

$$\mathbb{E}_\vartheta(f_-(T(X))) = \int_{\mathbb{R}^d} \frac{1}{z(\vartheta)} e^{\vartheta \cdot t} f_-(t) \nu_0(dt).$$

Insbesondere gilt für alle $\vartheta \in Q$:

$$\int_{\mathbb{R}^d} e^{\vartheta \cdot t} f_-(t) \nu_0(dt) = \int_{\mathbb{R}^d} e^{\vartheta \cdot t} f_+(t) \nu_0(dt).$$

Nach der Eindeutigkeit der Laplace-Transformation von Maßen gilt damit

$$f_+(t) \nu_0(dt) = f_-(t) \nu_0(dt)$$

im Sinne von Maßen auf \mathbb{R}^d . Nach Definition von $\nu_0 = (T^{-1}) \mu_0$ ist dies äquivalent zu

$$f_+(T(x)) \mu_0(dx) = f_-(T(x)) \mu_0(dx)$$

im Sinne von Maßen auf \mathfrak{X} also $(T(x)) = f_+(T(x)) - f_-(T(x)) = 0$ μ_0 -fast überall. \square

Korollar 2.27. Jeder Schätzer der Form $\Phi(T(x))$ für

$$\tau(\vartheta) = \mathbb{E}_\vartheta(\Phi(T(X)))$$

ist ein varianzminimierender erwartungstreuer Schätzer für τ .

Beweis von Satz 2.24. Es handelt sich hierbei um eine mehrdimensionale exponentielle Familie mit

$$T = T(x) = \left(\underbrace{\sum_{i=1}^n (x_i)^2}_{=: T_1}, \underbrace{\sum_{i=1}^n x_i}_{=: T_2} \right) \in \mathbb{R}^2$$

und wir hatten bereits gesehen, dass

$$d_2(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n (x_i)^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right)$$

ein erwartungstreuer Schätzer für v ist. Zudem ist T eine suffiziente und vollständige Statistik für (\mathbb{P}_ϑ) . Da wir d_2 darstellen können als

$$d_2(x) = \frac{1}{n-1} \left(T_1(x) - \frac{1}{n} (T_2(x))^2 \right) = \tilde{d}_2(T(x))$$

folgt aus Korollar 2.27, dass ist $d_2(x)$ ein UMV-Schätzer für v ist. \square

Vergleich mit der Cramer-Rao-Schranke

Wir wollen jetzt noch die Varianz des Schätzers d_2 mit der Cramer-Rao-Schanke für mehrdimensionale parametrische Modelle vergleichen. Diese lautet wie folgt.

Satz 2.28 (Cramer-Rao in \mathbb{R}^d). *Seien $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein reguläres³ statistisches Modell mit $\Theta \subseteq \mathbb{R}^d$ offen, $\tau : \Theta \rightarrow \mathbb{R}$ und $T : \mathfrak{X} \rightarrow \mathbb{R}$ ein erwartungstreuer Schätzer für τ . Dann gilt*

$$\mathbb{V}_\vartheta(T) \geq \langle \nabla \tau, I^{-1} \nabla \tau \rangle.$$

Dabei ist $I(\vartheta)$ die Informationsmatrix

$$(I(\vartheta)_{ij}) = \left(\mathbb{E}_\vartheta \left(\frac{\partial}{\partial \vartheta_i} \log(\varrho(\vartheta, X)) \cdot \frac{\partial}{\partial \vartheta_j} \log(\varrho(\vartheta, X)) \right) \right).$$

Beweis. Zunächst gilt wie zuvor in 1d hier für die partiellen Ableitungen

$$\begin{aligned} \frac{\partial}{\partial \vartheta_j} \tau(\vartheta) &= \frac{\partial}{\partial \vartheta_j} \mathbb{E}_\vartheta(T) \\ &= \int_{\mathfrak{X}} T(x) \cdot \frac{\partial}{\partial \vartheta_j} \varrho(\vartheta, x) \mu_0(dx) \\ &= \int_{\mathfrak{X}} T(x) \frac{\partial}{\partial \vartheta_j} \log(\varrho(\vartheta, x)) \cdot \varrho(\vartheta, X) dx \\ &= \mathbb{E}_\vartheta \left(T(X) \cdot \frac{\partial}{\partial \vartheta_j} \log(\varrho(\vartheta, X)) \right) \end{aligned}$$

Ferner gilt

$$\mathbb{E}_\vartheta \left(\frac{\partial}{\partial \vartheta_j} \log(\varrho(\vartheta, X)) \right) = 0$$

³Die Regularitätsbedingung 3) aus dem eindimensionalen Fall $\Theta \subset \mathbb{R}$ für $\frac{d}{d\vartheta}$ muss hier hier durch die entsprechende Regularitätsbedingung für sämtliche partielle Ableitungen $\frac{\partial}{\partial \vartheta_i}, i = 1, \dots, d$, ersetzt werden.

und

$$\begin{aligned}
\sqrt{\text{Var}_\vartheta(T)} &= \sqrt{\mathbb{V}_\vartheta(T)} = \sup_{\mathbb{V}_\vartheta(Z)=1} \text{Cov}_\vartheta(T, Z) \\
&= \sup_{\substack{\mathbb{V}_\vartheta(Z)=1 \\ \mathbb{E}_\vartheta(Z)=0}} \mathbb{E}_\vartheta(T \cdot Z) \\
&= \sup_{\mathbb{E}_\vartheta(Z)=0} \frac{\mathbb{E}_\vartheta(T \cdot Z)}{\sqrt{\mathbb{V}_\vartheta(Z)}} \\
&\geq \sup_{\eta \in \mathbb{R}^d} \frac{\mathbb{E}_\vartheta\left(T \cdot \overbrace{\sum \eta_k \frac{\partial}{\partial \vartheta_k} \log(\varrho(\vartheta, X))}^{R(\eta)}\right)}{\sqrt{\mathbb{V}_\vartheta(R(\eta))}} \\
&= \frac{\langle \nabla \tau, \eta \rangle}{\sqrt{\langle \eta, I \eta \rangle}}.
\end{aligned}$$

Somit

$$\begin{aligned}
\sqrt{\mathbb{V}_\vartheta(T)} &\geq \sup_{\eta \in \mathbb{R}^d} \frac{\langle \nabla \tau, \eta \rangle}{\sqrt{\langle \eta, I \eta \rangle}} = \sup_{\langle \eta, I \eta \rangle \leq 1} \langle \nabla \tau, \eta \rangle = \sup_{\langle \eta, I \eta \rangle \leq 1} \langle I^{-1} \nabla \tau, I \eta \rangle \\
&= \sqrt{\langle I^{-1} \nabla \tau, I I^{-1} \nabla \tau \rangle} = \sqrt{\langle \nabla \tau, I^{-1} \nabla \tau \rangle} \quad \square
\end{aligned}$$

Beispiel. Für die Anwendung von Satz 2.28 im n -fachen n -fachen Gauß-Produktmodell $\mathbb{P}_\vartheta = (\nu_{m,v}^{\otimes n}; m \in \mathbb{R}; v \geq 0)$ auf $(\mathfrak{X}, \mathfrak{F}) = (\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n))$ mit unbekanntem Parametern m und v für die Kenngröße

$$\tau(m, v) = v \quad \Rightarrow \quad \nabla \tau = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

berechnen wir

$$\frac{\partial \log \varrho}{\partial v} = -\frac{n}{2v} + \sum_{i=1}^n \frac{(x_i - m)^2}{2v^2} = \sum_{i=1}^n \frac{(x_i - m)^2 - v}{2v^2} = u_v$$

und

$$\frac{\partial \log \varrho}{\partial m} = \sum_{i=1}^n \frac{x_i - m}{v} = u_m.$$

Folglich

$$\begin{aligned}
I_{22}(\vartheta) &= \mathbb{E}_\vartheta(u_v(X)^2) = \frac{n}{2v^2} \\
I_{11}(\vartheta) &= \mathbb{E}_\vartheta(u_m(X)^2) = \frac{n}{v} \\
I_{12}(\vartheta) &= I_{21}(\vartheta) = \mathbb{E}_\vartheta(u_m(X) \cdot u_v(X)) = 0.
\end{aligned}$$

Also

$$I(\vartheta) = \begin{pmatrix} \frac{n}{v} & 0 \\ 0 & \frac{n}{2v^2} \end{pmatrix}$$

und damit

$$I(\vartheta)^{-1} = \begin{pmatrix} \frac{v}{n} & 0 \\ 0 & \frac{2v^2}{n} \end{pmatrix},$$

so dass

$$\mathbb{V}(T) \geq \langle \nabla \tau, I^{-1} \nabla \tau \rangle = I_{22}^{-1} = \frac{2v^2}{n}.$$

Ferner rechnet man direkt nach (Übung), dass

$$\mathbb{V}_{(m,v)}(d_2) = \frac{2v^2}{n-1} > \frac{2v^2}{n}.$$

Da d_2 UMV-optimal ist, gibt es also keinen Cramer-Rao optimalen Schätzer für v im n -fachen Gaußmodell, sofern m ebenfalls unbekannt ist.

2.1.9 Bayes'sche Schätzer

Der sogenannte Bayes'sche Ansatz liefert ein weiteres Konstruktionsverfahren für Schätzer. Allerdings hängt diese Methode von einer weiteren Größe ab, die in einem statistischen Modell zunächst nicht enthalten ist, nämlich einer zuvor gewählten (*a-priori*-) Wahrscheinlichkeitsverteilung ν auf der Parametermenge Θ .

Beispiel (Münzsack). Wir finden großen Sack mit Münzen unterschiedlicher Art, d.h. insbesondere mit unterschiedlichen Erfolgsparametern p . Wir nehmen an, dass jeder Erfolgsparameter $p \in [0, 1]$ gleich häufig vorkommt. Wir ziehen nun eine Münze und werfen sie n mal.

- Die Wahrscheinlichkeit, mit der gezogenen Münze k mal Erfolg zu haben, beträgt (mit $X \hat{=}$ Anzahl der Erfolge)

$$\mathbb{P}(X = k) := \int_{[0,1]} B(n, \vartheta)[k] \nu(d\vartheta),$$

mit $B(n, \vartheta)[k] = \binom{n}{k} \vartheta^k (1 - \vartheta)^{n-k}$ (Binomialverteilung zum Erfolgsparameter $\vartheta \in [0, 1]$) und $\nu(d\vartheta) = d\theta$ (Gleichverteilung auf $[0, 1]$).

- Angenommen, wir hatten beim n -maligen Werfen mit der gezogenen Münze k mal Erfolg, so können wir uns nun ('a-posteriori') fragen, welche Wahrscheinlichkeitsverteilung sich hieraus für den unbekanntem Erfolgsparameter p der gezogenen Münze ergibt. Dies wäre zum Beispiel bedeutsam, wenn wir ein neues Spiel mit dieser gezogenen Münze spielen wollten. Hierzu stellen wir fest, dass

$$\begin{aligned} \mathbb{P}(\vartheta \in d\vartheta | X = k) &= \frac{\mathbb{P}(X = k, \vartheta \in d\vartheta)}{\mathbb{P}(X = k)} = \frac{\mathbb{P}(X = k | \vartheta \in d\vartheta) \cdot \mathbb{P}(\vartheta \in d\vartheta)}{\mathbb{P}(X = k)} \\ &= \frac{B(n, \vartheta)[k] \nu(d\vartheta)}{\int_{[0,1]} B(n, \vartheta)[k] \nu(d\vartheta)} =: \nu_{X=k}(d\vartheta). \end{aligned}$$

Wenn wir also wie zuvor von der Gleichverteilung $\nu(d\vartheta) = d\vartheta$ auf $[0, 1]$ für die im Münzsack enthaltenen Münzen bzw. Erfolgsparameter ausgehen, ergibt sich

$$\nu_{X=k}(d\vartheta) = \frac{\binom{n}{k} \vartheta^k (1 - \vartheta)^{n-k} d\vartheta}{\int_{[0,1]} \binom{n}{k} \vartheta^k (1 - \vartheta)^{n-k} d\vartheta} = \frac{1}{B(k, n-k)} \cdot \vartheta^k \cdot (1 - \vartheta)^{n-k} d\vartheta.$$

Dies ist nicht mehr die Gleichverteilung sondern die sogenannte Beta-Verteilung⁴ auf $[0, 1]$ mit den Parametern $(k + 1, n - k + 1)$.

Die in diesem Beispiel auftretenden Verteilungen ν und ν_x heißen *a-priori* bzw. *a-posteriori*-Verteilungen für den unbekanntem Parameter $\vartheta \in \Theta$. Allgemein verwenden wir die folgende Sprechweise.

Definition 2.29. Sei $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein Standardmodell mit Likelihood-Funktion $\varrho(\vartheta, x)$. Dann heißt ein Wahrscheinlichkeitsmaß ν auf (Θ, τ) *a-priori-Verteilung* für den Parameter $\vartheta \in \Theta$. Zu $x \in \mathfrak{X}$ heißt

$$\nu_x(d\vartheta) = \frac{\varrho(\vartheta, x)}{\int_{\Theta} \varrho(\vartheta, \tilde{\vartheta}) \nu(d\tilde{\vartheta})} \nu(d\vartheta)$$

die *a-posteriori-Verteilung* für den Parameter $\vartheta \in \Theta$ gegeben die Beobachtung x .

Im Beispiel des Münzsackes war also $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ das Binomialmodell für n Versuche und unbekanntem Erfolgsparameter p , sowie $\vartheta \in [0, 1]$, $\nu(d\vartheta) \hat{=}$ das uniforme Maß auf $[0, 1]$. (Letzteres entsprang unserer Annahme, dass im Münzsack jeder Erfolgsparameter gleich wahrscheinlich auftritt. Wir hätten also auch ein anderes a-priori Maß wählen können.) Mit dieser Wahl vom a-priori Maß ν erhalten wir die durch $k \in \mathfrak{X}$ parametrisierte Familie von a-posteriori Verteilungen auf $[0, 1]$

$$\nu_k(d\vartheta) = \frac{\vartheta^k (1 - \vartheta)^{n-k}}{\int_0^1 \tilde{\vartheta}^k (1 - \tilde{\vartheta})^{n-k} d\tilde{\vartheta}} d\vartheta$$

Bemerkung (Mathematische Bedeutung des a-posteriori-Maßes). Durch zufällige Wahl aus der Menge $(\mathbb{P}_\vartheta)_\vartheta$ gemäß der Verteilung ν ergibt sich ein Wahrscheinlichkeitsmaß

$$\mathbb{P}(dx d\vartheta) = \mathbb{P}_\vartheta(dx) \nu(d\vartheta)$$

für das Auftreten von (x, ϑ) -Paaren. Mit anderen Worten, für $f = f(\vartheta, x)$

$$\mathbb{E}_{\mathbb{P}}(f(X, \vartheta)) = \int_{\Theta} \int_{\mathfrak{X}} f(x, \vartheta) \mathbb{P}_\vartheta(dx) \nu(d\vartheta)$$

Die Verteilungen des statischen Modells

$$\mathbb{P}_\vartheta(dx) = \mathbb{P}(dx|\vartheta)$$

sind somit die bedingte Verteilung von X unter $\mathbb{P}(\vartheta, dx)$ gegeben ϑ . Die a-posteriori Verteilungen ergeben sich jetzt aus $\mathbb{P}(d\vartheta, dx)$ durch Vertauschung der Bedingung, d.h.

$$\nu_x(d\vartheta) = \mathbb{P}(d\vartheta|x),$$

⁴Die Familie der Beta-Verteilungen

$$\frac{1}{B(p, q)} x^{p-1} (1-x)^{q-1} = f(x), \quad p, q > 0$$

mit $B(p, q) = \int_0^1 x^{p-1} (1-x)^{q-1} dx$ sind Wahrscheinlichkeitsdichten auf $[0, 1]$ (siehe Kapitel 3).

bzw. äquivalent hierzu durch Vertauschung der Integrationsreihenfolge, d.h. für $f = f(\vartheta, x)$

$$\mathbb{E}_{\mathbb{P}}(f(X, \vartheta)) = \int_{\mathfrak{X}} \int_{\Theta} f(x, \vartheta) \nu_x(d\vartheta) \mathbb{P}_{\nu}(dx).$$

Die Randverteilung $\mathbb{P}_{\nu}(dx) := \int_{\Theta} \mathbb{P}_{\vartheta}(dx) \nu(d\vartheta)$ für die Zufallsvariable X unter \mathbb{P} wird auch als *Mischung* der Familie (\mathbb{P}_{ϑ}) mit der Verteilung ν bezeichnet.

Definition 2.30. In der Situation wie in Definition 2.29 sei ferner $\tau : \Theta \rightarrow \mathbb{R}$ eine Kenngröße und sei $T : \mathfrak{X} \rightarrow \mathbb{R}$ ein Schätzer. Weiter sei

$$\mathbb{E}_{\nu}((\tau(\vartheta))^2) < \infty.$$

Dann heißt der Schätzer T ein *Bayes-Schätzer* zur Kenngröße τ , falls

$$\mathbb{F}_{\nu}(T) = \mathbb{E}_{\nu}(\mathbb{F}_{\vartheta}(T)) = \int_{\Theta} \mathbb{F}_{\vartheta}(T) \nu(d\vartheta) = \int_{\Theta} \int_{\mathfrak{X}} |\tau(\vartheta) - T(x)|^2 \mathbb{P}_{\vartheta}(dx) \nu(d\vartheta)$$

minimal ist unter allen Schätzern $T : \mathfrak{X} \rightarrow \mathbb{R}$.

Satz 2.31. *Unter den Voraussetzungen von Definition 2.30 ist*

$$\widehat{T}(x) := \mathbb{E}_{\nu_x}(\tau(\vartheta)) = \int_{\Theta} \tau(\vartheta) \nu_x(d\vartheta)$$

ist der (bis auf Nullmengen) eindeutig bestimmte Bayes-Schätzer zur Kenngröße τ .

Beweis. Sei T ein Schätzer. Dann wird der Ausdruck

$$\mathbb{F}_{\nu}(T) = \int_{\mathfrak{X}} \underbrace{\int_{\Theta} |\tau(\vartheta) - T(x)|^2 \nu_x(d\vartheta)}_{:= H_x(T(x))} \mathbb{P}_{\nu}(dx)$$

mit

$$H_x(T) = \int_{\Theta} |\tau(\vartheta) - T|^2 \nu_x(d\vartheta)$$

minimal, wenn für jedes $x \in \mathfrak{X}$ die Funktion $T \rightarrow H_x(T)$ für bzgl. der Variablen T minimiert wird. Nun wissen wir aus der Wahrscheinlichkeitstheorie dass allgemein ein Ausdruck der Form $\mathbb{E}(|X - c|^2)$ für $c = \mathbb{E}(X)$ minimal wird, d.h. der Ausdruck $H_x(T)$ ist minimal für $T = T(x) := \int_{\Theta} \tau(\vartheta) \nu_x(d\vartheta) = \mathbb{E}_{\nu_x}(\tau)$. \square

Der Bayes-Schätzer $T(x)$ zur Kenngröße τ bei gegebener Realisierung $x \in \mathfrak{X}$ ist also einfach der Erwartungswert von τ unter dem a-posteriori Maß $\nu_x(d\vartheta)$ gegeben x .

Beispiel (Münzsack, Forts.). Die a-priori-Verteilung auf $[0, 1]$ war

$$\nu = U([0, 1]).$$

Hieraus ergeben sich die a-posteriori-Verteilungen auf $[0, 1]$ gegeben $x \in \{0, \dots, n\}$

$$\nu_k(d\vartheta) = \frac{\vartheta^k(1-\vartheta)^{n-k}}{\underbrace{\int_0^1 \tilde{\vartheta}^k(1-\tilde{\vartheta})^{n-k} d\tilde{\vartheta}}_{:=B(k+1, n-k+1)}} d\vartheta.$$

Sei nun $\tau(\vartheta) = \vartheta$. Dann ist der Bayes-Schätzer für ϑ

$$\begin{aligned} T(x) &= \mathbb{E}_{\nu_x}(\tau(\vartheta)) = \frac{1}{B(x+1, n-x+1)} \int_0^1 \tau(\vartheta) \vartheta^x (1-\vartheta)^{n-x} d\vartheta \\ &= \frac{1}{B(x+1, x-k+1)} \int_0^1 \vartheta^{x+1} (1-\vartheta)^{n-x} d\vartheta = \frac{B(x+2, x-k+1)}{B(x+1, x-k+1)} = \frac{x+1}{n+1}, \end{aligned}$$

wobei man im letzten Schritt ausnutzt, dass $B(p, q)$ ein Quotient von Gamma-Funktionen ist (siehe Kapitel 3).

Im Münzsack-Beispiel schätzt man also mit der Gleichverteilung auf $[0, 1]$ als a-priori Verteilung nach der Bayes-Methode den Erfolgsparameter der gezogenen Münze auf $(x+1)/(n+1)$, sofern man beim n -maligen Ausprobieren x Erfolge beobachtet hat. Mit $T(x) = \frac{x+1}{n+1} =: T^n(x)$ ist die Folge der Schätzer T^n zwar nicht erwartungstreu, aber nach Satz 2.3 konsistent, denn

$$\mathbb{E}(T^n) = \frac{n\vartheta + 1}{n+1} \rightarrow \vartheta$$

und

$$\mathbb{V}(T^n) = \frac{n \cdot \vartheta \cdot (1-\vartheta)}{(n+1)^2} \rightarrow 0.$$

2.2 Bereichsschätzer (Konfidenzmengen)

In vielen Fällen wird ein Schätzer T für eine Kenngröße τ diesen fast sicher nicht treffen, selbst wenn T erwartungstreu ist.

Beispiel. Im einfachen Gauß-Modell $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), (\nu_{m,1}; m \in \mathbb{R}))$ mit unbekanntem Erwartungswert $m \in \mathbb{R}$ ist $T(x) := x$ ein erwartungstreuer Schätzer für die Kenngröße $\tau(m) = m$ aber $T(X) \neq m$ fast sicher, da X eine stetige Zufallsvariable ist und somit einen fest vorgegebenen Punkt fast sicher nie trifft, d.h.

$$\mathbb{P}_m(X = c) = 0 \quad \forall c \in \mathbb{R}$$

Als Lösung für dieses Problem führt man mengenwertige Schätzer ein, sogenannte Konfidenzbereiche.

Definition 2.32. Sei $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein statistisches Modell und $\tau : \Theta \rightarrow \Sigma$ eine Kenngröße und $\alpha \in (0, 1)$. Dann heißt eine Abbildung

$$C : \mathfrak{X} \rightarrow \mathcal{P}(\Sigma)$$

Konfidenzbereich für τ zum Niveau α , falls

$$\forall \vartheta \in \Theta : \mathbb{P}_\vartheta(C(X) \ni \tau(\vartheta)) \geq 1 - \alpha$$

Vereinbarungen.

1. Solang nicht anders definiert, beschränken wir uns in diesem Abschnitt stets auf den Fall $\tau(\vartheta) = \vartheta$.
2. Für eine Menge $C \subset \mathfrak{X} \times \Theta$ führen wir die folgenden Schreibweisen ein:

$$C := \{(x, \vartheta) \in \mathfrak{X} \times \Theta \mid \vartheta \in C(x)\}$$

und

$$C_\vartheta := \{x \in \mathfrak{X} \mid (x, \vartheta) \in C\}$$

$$C_x = C(x) = \{\vartheta \in \Theta \mid (x, \vartheta) \in C\}$$

3. Der Vollständigkeit halber müsste die Definition 2.32 noch um die Messbarkeitsvoraussetzung

$$C_\vartheta \in \mathfrak{F} \quad \forall \vartheta \in \Theta$$

erweitert werden, aber wir werden diesen technischen Punkt im folgenden ignorieren.

Bemerkung.

1. In Worten lautet die Bedingung an einen Konfidenzbereich wie folgt. *Die Wahrscheinlichkeit, dass $C(X)$ die gesuchte Kenngröße $\tau(\vartheta)$ enthält, beträgt mindestens $1 - \alpha$.*
2. Die Notation

$$\mathbb{P}_\vartheta(\tau(\vartheta) \in C(X)) \geq 1 - \alpha$$

ist formell äquivalent, könnte aber leicht zu logischen Fehlern führen.

3. Trivialer Weise ist die konstante Abbildung $\mathfrak{X} \ni x \rightarrow C(x) = \Theta$ ein Konfidenzbereich für jedes Niveau $\alpha > 0$, aber leider vollkommen wertlos, weil aus der Realisierung x keine Information über die Lage des Parameters ϑ gewonnen wird. Man ist also an möglichst kleinen Konfidenzmengen interessiert. Falls etwa $C(x) \subset \tilde{C}(x)$ für alle $x \in \mathfrak{X}$ für zwei Konfidenzbereiche C und \tilde{C} zum Niveau α , so wäre C vorzuziehen.

2.2.1 Konfidenzbereiche im Binomialmodell

Wir diskutieren drei Ansätze zur Bestimmung von Konfidenzbereichen im Binomialmodell $(\{0, \dots, n\}, \mathcal{P}(\mathfrak{X}), (\mathbb{P}_\vartheta; \vartheta \in [0, 1]))$ mit $\mathbb{P}_\vartheta = B_{n,\vartheta}$.

1. Methode. (Vermöge Tschebyschev-Abschätzung)

Wir benutzen den Schätzer $T(x) = \frac{x}{n}$ und wählen den Ansatz $C(x) = (\frac{x}{n} - \varepsilon, \frac{x}{n} + \varepsilon)$. Hierbei ist $\varepsilon > 0$ so zu wählen dass $B_{n,\vartheta}(|\frac{X}{n} - \vartheta| \geq \varepsilon) \leq \alpha$. Nun gilt mit der Tschebyschev-Ungleichung

$$B_{n,\vartheta} \left(\left| \frac{X}{n} - \vartheta \right| \geq \varepsilon \right) \leq \frac{1}{\varepsilon^2} \text{Var}_\vartheta \left(\frac{X}{n} \right) = \frac{\vartheta(1-\vartheta)}{n^2 \varepsilon^2} \leq \frac{1}{n4\varepsilon^2} \stackrel{!}{\leq} \alpha,$$

wobei wir $\vartheta(1 - \vartheta) \leq \frac{1}{4}$ benutzt haben, und die letzte Ungleichung zutrifft, falls

$$\varepsilon \geq \frac{1}{2\sqrt{n\alpha}}.$$

Wir halten das Ergebnis in einem Satz fest.

Satz 2.33. *Im Binomialmodell $(\{0, \dots, n\}, \mathcal{P}(\mathfrak{X}), (\mathbb{P}_\vartheta; \vartheta \in [0, 1]))$ definiert die Vorschrift*

$$\{0, \dots, n\} \ni x \rightarrow C(x) := \left(\frac{x}{n} - \varepsilon, \frac{x}{n} + \varepsilon\right) \subset \mathbb{R}$$

mit $\varepsilon = 1/2\sqrt{n\alpha}$ einen Konfidenzbereich zum Niveau $\alpha > 0$ für den Parameter ϑ .

Falls wir also eine unbekannte Münze $n = 1000$ mal werfen, so überdeckt das zufällige Intervall

$$C(x) = (x/1000 - 0.1, x/1000 + 0.1)$$

einer Wahrscheinlichkeit von mehr als $0.9975 = 1 - \alpha$ (für $\alpha = 2.5\%$) den unbekanntem Erfolgsparameter ϑ .

2. Methode. (Clopper-Pearson oder Quantil-Methode)

Hier konstruieren wir die Menge $C \subset \mathfrak{X} \times \Theta$ in ϑ -Schnitten wie folgt.

1. Bestimme zu $\vartheta \in \Theta$ ein C_ϑ , sodass

$$\mathbb{P}_\vartheta(X \in C_\vartheta) \geq 1 - \alpha$$

2. Setze $C := \bigcup_{\vartheta \in \Theta} C_\vartheta \times \{\vartheta\}$

3. Definiere $C(x) := \{\vartheta \in \Theta \mid (x, \vartheta) \in C\}$

Bevor wir diesen Ansatz weiterverfolgen, erinnern wir noch an die nützliche Quantil-Sprechweise für reelle Verteilungen.

Definition 2.34. Q sei ein Wahrscheinlichkeitsmaß auf \mathbb{R} , $\alpha \in (0, 1)$. Dann heißt $q \in \mathbb{R}$ α -Quantil zur Verteilung Q , falls

$$Q((-\infty, q]) \geq \alpha \text{ und } Q([q, \infty)) \geq 1 - \alpha.$$

Bemerkung. \circ $\frac{1}{2}$ -Quantile heißen *Mediane*.

- \circ Die $\frac{1}{4}$ -Quantile und $\frac{3}{4}$ -Quantile heißen auch 1. bzw. 3. *Quartil*.
- \circ Ein $(1 - \alpha)$ -Quantil heißt α -*Fraktile*.

In diesem Sinne handelt es sich bei der Clopper-Pearson-Methode um einen *Quantilansatz*, d.h.

$C_\vartheta = \{x_-(\vartheta), \dots, x_+(\vartheta)\}$ mit

$$x_-(\vartheta) = \max \left\{ x \in \mathfrak{X} \mid B_{n,\vartheta}(\{0, \dots, x-1\}) \leq \frac{\alpha}{2} \right\}$$

$$x_+(\vartheta) = \min \left\{ x \in \mathfrak{X} \mid B_{n,\vartheta}(\{x+1, \dots, n\}) \leq \frac{\alpha}{2} \right\}$$

$x_-(\vartheta)$ ist also das größte $\frac{\alpha}{2}$ -Quantil der $B_{n,\vartheta}$ -Verteilung und $x_+(\vartheta)$ ist das kleinste $\frac{\alpha}{2}$ -Fraktile der $B_{n,\vartheta}$ -Verteilung. Gemäß unserem Ansatz definieren wir nun

$$C := \bigcup_{\vartheta \in (0,1)} C_\vartheta \times \{\vartheta\},$$

so dass

$$C(x) = \{\vartheta \in \Theta \mid (x, \vartheta) \in C\} = \{\vartheta \in \Theta \mid x_-(\vartheta) \leq x \leq x_+(\vartheta)\},$$

Um die Intervallgrenzen von $C(x)$ zu erhalten, müssen wir also die Funktionen $\vartheta \mapsto x_-(\vartheta)$ und $\vartheta \mapsto x_+(\vartheta)$ invertieren. Das folgende Lemma gibt eine (mehr oder weniger) explizite Darstellung der Inversen dieser beiden Funktionen in Form von Quantilen von geeigneten Beta-Verteilungen an.

Lemma 2.35.

1. Für $x \in \{1, \dots, n\}$ ist die Funktion

$$\vartheta \mapsto B_{n,\vartheta}(\{x, \dots, n\})$$

stetig und strikt wachsend auf $[0, 1]$. Ferner gilt

$$B_{n,\vartheta}(\{x, \dots, n\}) = \beta_{x,n-x+1}([0, \vartheta])$$

2. Es gilt

$$x \leq x_+(\vartheta) \Leftrightarrow \beta_{x,n-x+1}([0, \vartheta]) > \frac{\alpha}{2}$$

bzw.

$$x \geq x_-(\vartheta) \Leftrightarrow \beta_{x,n-x+1}([0, \vartheta]) > \frac{\alpha}{2}.$$

3. Die Bedingung

$$x_-(\vartheta) \leq x \leq x_+(\vartheta)$$

ist äquivalent zu

$$p_-(x) \leq \vartheta \leq p_+(x)$$

wobei $p_-(x)$ das $\frac{\alpha}{2}$ -Quantil der $\beta_{x,n-x+1}$ -Verteilung und $p_+(x)$ das $\frac{\alpha}{2}$ -Fraktile der $\beta_{x+1,n-x}$ -Verteilung bezeichnen.

Der Beweis ist elementar, aber durchaus aufwendig, so dass wir ihn hier nicht führen wollen (die Details findet man im Buch von Georgii auf S. 230.) sondern halten als Konsequenz das folgende Ergebnis fest.

Satz 2.36. Im Binomialmodell $(\{0, \dots, n\}, \mathcal{P}(\{0, \dots, n\}), (B_{n,\vartheta}; \vartheta \in [0, 1]))$ seien zu $\alpha \in (0, 1)$ die Funktionen

$$\circ p_-(x) := \frac{\alpha}{2}\text{-Quantil der } \beta_{x,n-x+1}\text{-Verteilung}$$

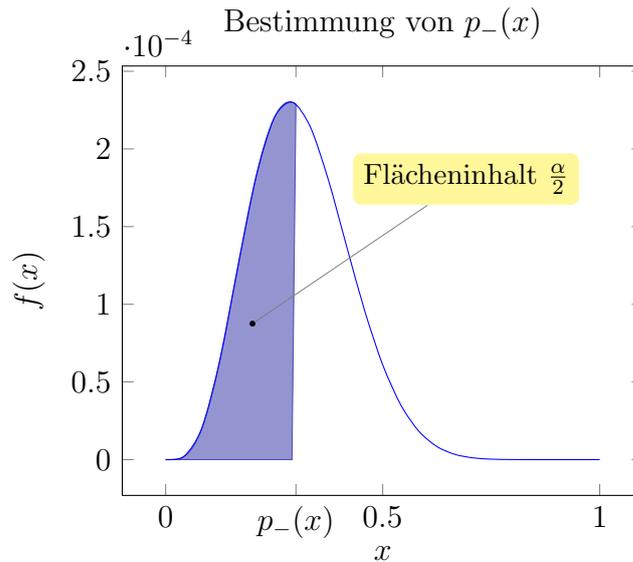
◦ $p_+(x) := \frac{\alpha}{2}$ -Fraktile der $\beta_{x+1, n-x}$ -Verteilung

definiert. Dann ist die Vorschrift

$$x \mapsto C(x) := [p_-(x), p_+(x)] \subset [0, 1]$$

ein Konfidenzbereich für $\tau(\vartheta) = \vartheta$ zum Niveau α .

Bemerkung. Die Zahlwerte von $p_-(x)$ bzw. $p_+(x)$ ermittelt man z.B. durch numerische Integration.



3. Methode. (Approximation durch Normalverteilung/Zentraler Grenzwertsatz)

Wir wählen wieder den Ansatz $C(x) = (\bar{x} - \varepsilon, \bar{x} + \varepsilon)$ mit $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, sodass

$$\mathbb{P}_\vartheta(|\bar{x} - \vartheta| < \varepsilon) > 1 - \alpha \quad \text{für } n \text{ hinreichend groß.}$$

Der Zusatz hinter der Ungleichung ist wichtig, denn wir wollen nun Gebrauch vom Zentralen Grenzwertsatz machen. Dieser ist eine asymptotische Aussage, so dass die Ungleichung in der Tat nur für hinreichend große n gilt.

Zunächst erinnern wir an dieses fundamentale Resultat der Wahrscheinlichkeitstheorie.

Satz 2.37 (ZGS). *Es sei (X_i) eine Folge von unabhängigen Realisierungen eines reellwertigen Zufallsexperimentes $X_i \simeq X_0 \forall i$ mit $\mathbb{E}(X_0^2) < \infty$, so gilt*

$$\frac{\sum_i (X_i - \mathbb{E}(X_0))}{\sqrt{n \cdot \text{Var}(X_0)}} \xrightarrow{\text{in Verteilung}} \nu_{0,1} \quad \text{falls } n \rightarrow \infty.$$

Für die Anwendung des ZGS in unserem Fall sei

$$\xi_n := \frac{\sum_{i=1}^n (x_i - \vartheta)}{\sqrt{n \cdot \vartheta \cdot (1 - \vartheta)}}$$

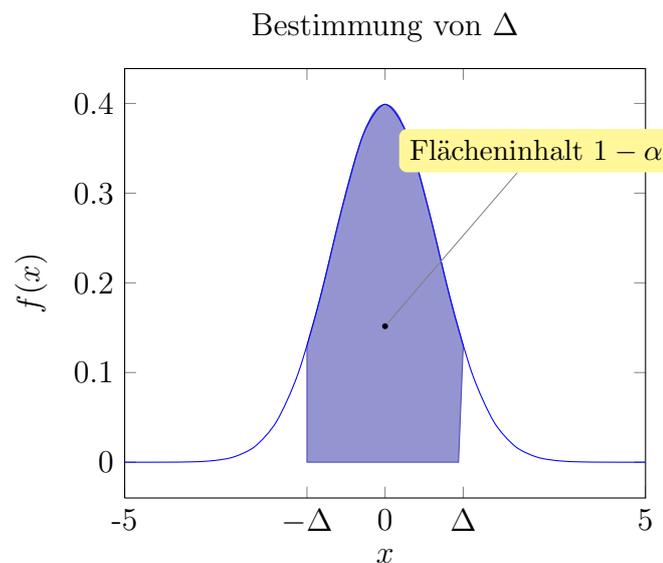
die normierte Summenvariable, so finden wir, dass für $\kappa > 0$ und $n \rightarrow \infty$

$$\lim_n \mathbb{P}_\vartheta (|\xi_n| < \kappa) = \mathbb{P}(|N| < \kappa)$$

wobei N eine standardnormalverteilte Zufallsvariable bezeichnet. Folglich gilt für hinreichend große n

$$\mathbb{P}_\vartheta (|\xi_n| < \Delta) > 1 - \alpha,$$

falls $\Delta > 0$ so gewählt ist, dass $\Phi(\Delta) - \Phi(-\Delta) > 1 - \alpha$ für die Verteilungsfunktion $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-s^2/2} ds$ der Normalverteilung.



Aufgrund der Symmetrie der Normalverteilung ist dies äquivalent zu der Bedingung

$$2(1 - \Phi(\Delta)) < \alpha.$$

Falls $\Delta > \Phi^{-1}(1 - \frac{\alpha}{2})$, also größer als das $\alpha/2$ -Fraktile der Standardnormalverteilung gewählt wurde, tritt somit für hinreichend große n das Ereignis $\{|\xi_n| < \Delta\}$ mit Wahrscheinlichkeit nicht größer als α auf. Nach Definition von ξ_n

$$|\xi_n| < \Delta \Leftrightarrow |\bar{x} - \vartheta| < \Delta \sqrt{\frac{\vartheta(1-\vartheta)}{n}}.$$

Da $\vartheta(1-\vartheta) \leq \frac{1}{4}$ folgt hieraus auch, dass dann $|\bar{x} - \vartheta| < \frac{\Delta}{2\sqrt{n}}$. Somit finden wir abschließend, dass unter \mathbb{P}_ϑ mit $\vartheta \in]0, 1[$ das Ereignis

$$\{|\bar{x} - \vartheta| < \frac{\Delta}{2\sqrt{n}}\} \text{ mit Wahrscheinlichkeit mindestens } 1 - \alpha$$

eintritt, sofern $\Delta > \Phi(1 - \frac{\alpha}{2})$ und hinreichend groß, d.h. falls $n > N = N(\alpha, \vartheta)$.

Für die Zusammenfassung dieser Überlegungen führen wir nun noch die folgende Sprechweise ein.

Definition 2.38. Für eine Folge von statistischen Modellen $(\mathfrak{X}_n, \mathfrak{F}_n, (\mathbb{P}_\vartheta^n)_{\vartheta \in \Theta})$ mit gemeinsamer Parametermenge Θ heißt eine Folge von Abbildungen $C_n : \mathfrak{X}_n \rightarrow \mathfrak{P}(\Theta)$ ein *asymptotischer Konfidenzbereich für ϑ zum Niveau $\alpha \in [0, 1]$* , falls für alle $\vartheta \in \Theta$

$$\liminf_{n \rightarrow \infty} \mathbb{P}_\vartheta^n(C_n \ni \vartheta) \geq 1 - \alpha.$$

Mit dieser Sprechweise liest sich das Ergebnis in dieses Abschnitts wie folgt.

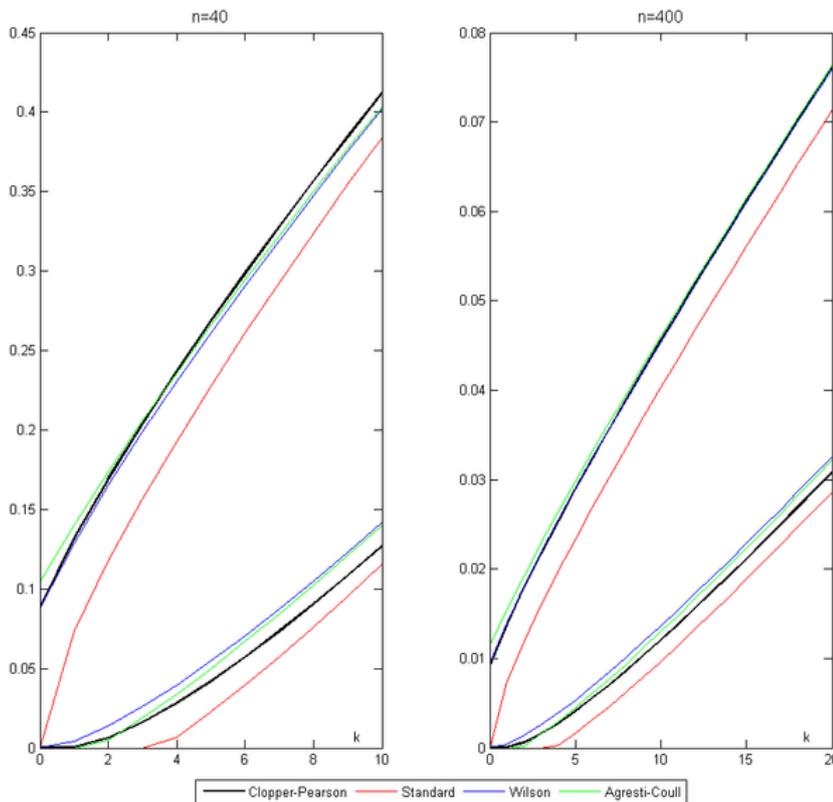
Satz 2.39. In der Folge der Binomialmodelle $(\{0, \dots, n\}, \mathcal{P}(\{0, \dots, n\}), (B_{n,\vartheta}; \vartheta \in [0, 1]))$ definiert die Vorschrift

$$x \rightarrow C_n(x) = (\bar{x} - \varepsilon_n, \bar{x} + \varepsilon_n) \subset \mathbb{R}$$

einen asymptotischen Konfidenzbereich zum Niveau $\alpha \in [0, 1]$ für den Parameter ϑ , sofern

$$\varepsilon_n > \frac{\Phi^{-1}(1 - \frac{\alpha}{2})}{2\sqrt{n}}.$$

Es gibt mehr als diese drei Methoden zur Konstruktion von Konfidenzbereichen im Binomialmodell. Zur Illustration zeigen wir noch einen gemeinsamen Plot der Randkurven $\vartheta_-(\cdot)$ und $\vartheta_+(\cdot)$ von $x \rightarrow C(x) = [\vartheta_-(x), \vartheta_+(x)]$ für die verschiedenen Konfidenzbereiche C im Falle von $n = 40$ und $n = 100$ (Quelle: Wikipedia). (Die Randkurven zur 'Tschebyschev-Methode' wären Geraden und sind nicht eingezeichnet.)



2.2.2 Pivotstatistiken

Abschließend wollen wir noch eine allgemeine Methode zur Konstruktion von Konfidenzbereichen darstellen, der wir später im Zusammenhang mit Tests wiederbegegnen werden.

Definition 2.40. Sei $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein statistisches Modell und $\pi : \Theta \times \mathfrak{X} \rightarrow \Sigma$ eine Abbildung. Dann heißt π eine *Pivot-Statistik* oder auch nur *Pivot*, falls für $\vartheta \in \Theta$ die Verteilung von

$$\pi(\vartheta, \cdot) : \mathfrak{X} \rightarrow \Sigma$$

unter \mathbb{P}_ϑ nicht von ϑ abhängt, d.h. es existiert ein Wahrscheinlichkeitsmaß η auf Σ , s.d.

$$\mathbb{P}_\vartheta(\pi(\vartheta, X) \in \mathfrak{G}) = \eta(\mathfrak{G}) \quad \forall \vartheta \in \Theta.$$

Beispiel. Im statischen Modell einer auf einem reellen Einheitsintervall unbekannter Lage gleichverteilten Zufallsvariable $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta)) = (\mathbb{R}, \mathfrak{B}, (dx \upharpoonright_{[\vartheta - \frac{1}{2}, \vartheta + \frac{1}{2}]}; \vartheta \in \mathbb{R}))$ ist $\pi(\vartheta, x) := (x - \vartheta)$ ein Pivor, denn für eine beliebige messbare Menge $J \subset \mathbb{R}$ gilt

$$\begin{aligned} \mathbb{P}_\vartheta(\pi(\vartheta, X) \in J) &= \mathbb{P}_\vartheta((X - \vartheta) \in J) \\ &= U_{[-\frac{1}{2}, \frac{1}{2}]((X + \vartheta) - \vartheta \in J) \\ &= U_{[-\frac{1}{2}, \frac{1}{2}]}(X \in J) \\ &= U_{[-\frac{1}{2}, \frac{1}{2}]}(J). \end{aligned}$$

Beispiel. Im einfachen Gauß-Modell mit unbekanntem Erwartungswert $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta)) = (\mathbb{R}, \mathfrak{B}, (\nu_{m,1}; m \in \mathbb{R}))$ ist $\pi(m, x) := x - m$ unter \mathbb{P}_m stets $\nu_{0,1}$ verteilt. D.h. die Verteilung von $\pi(X, m)$ hängt nicht mehr von m ab. Somit ist π Pivot.

Satz 2.41. Sei $\pi = \pi(\vartheta, x)$ ein Pivot für das statistische Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ mit Pivot-Verteilung η (gemäß Definition 2.40). Zu $\alpha \in (0, 1)$ sei $\mathfrak{G} \subseteq \Sigma$ eine messbare Menge mit

$$\eta(\mathfrak{G}) \geq 1 - \alpha$$

Dann ist die Menge

$$C := \{(\vartheta, x) \in \Theta \times \mathfrak{X} \mid \pi(\vartheta, x) \in \mathfrak{G}\}$$

ein Konfidenzbereich für ϑ zum Niveau α .

Beweis. Sei $x \mapsto C(x) := \{\vartheta \in \Theta \mid \pi(\vartheta, x) \in \mathfrak{G}\}$. Dann gilt:

$$\begin{aligned} \mathbb{P}_\vartheta(C(X) \ni \vartheta) &= \mathbb{P}_\vartheta(\{x \in \mathfrak{X} \mid \pi(\vartheta, x) \in \mathfrak{G}\}) \\ &= \mathbb{P}_\vartheta(\pi^{-1}(\vartheta, \cdot)(\mathfrak{G})) \\ &= \eta(\mathfrak{G}) \\ &\geq 1 - \alpha. \end{aligned} \quad \square$$

Beispiel (Gauß'sches Modell mit bekannter Varianz). $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta)) = (\mathbb{R}, \mathfrak{B}, (\nu_{m,1}; m \in \mathbb{R}))$, $\pi(m, x) = x - m$ ist Pivot. Sei $\alpha \in (0, 1)$. Dann gilt z.B.:

$$\nu_{0,1} \left[-\Phi^{-1} \left(1 - \frac{\alpha}{2} \right); \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right] \geq 1 - \alpha,$$

wobei $\Phi(t) = 1/\sqrt{2\pi} \int_0^t e^{-u^2/2} du$ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. D.h. mit (also dem $\alpha/2$ -Fraktile $\Delta := \Phi^{-1}(1 - \frac{\alpha}{2})$ von $\nu_{0,1}$ gilt

$$\mathfrak{S} = [-\Delta, +\Delta], \quad \eta(\mathfrak{S}) = \nu_{0,1}(\mathfrak{S}) \geq 1 - \alpha.$$

Entsprechend definiert $C(x) = \{m \mid x - m \in \mathfrak{S}\} = [x - \Delta, x + \Delta]$ einen Konfidenzbereich für den unbekannt Parameter m .

Falls wir weder m noch v kennen, wird die Konstruktion eines Konfidenzbereichs z.B. für v erheblich komplizierter. Hierbei tritt die sogenannte Student'sche t -Verteilung auf, die wir im nächsten Kapitel zusammen mit anderen Verwandten der Normalverteilung systematisch kennenlernen werden.

Satz 2.42 (Gauß'sches Modell mit unbekannte (m, v)). *Sei M das statistische Modell des n -fachen Gaußversuches mit unbekannter Varianz v und unbekanntem Erwartungswert m . Seien ferner $\alpha \in (0, 1)$ und $t_{\alpha/2}$ das $\frac{\alpha}{2}$ -Fraktile der t_{n-1} -Verteilung (Student-Verteilung mit $n - 1$ Freiheitsgraden). Dann ist*

$$C(x) := \left(\bar{x} - t_{\alpha/2} \sqrt{\frac{v^*}{n}}, \bar{x} + t_{\alpha/2} \sqrt{\frac{v^*}{n}} \right)$$

mit $v^* := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ ein Konfidenzbereich für den Parameter m zum Niveau α .

Beweis. Im nächsten Abschnitt der Vorlesung werden wir sehen, dass die Größe $\pi(\vartheta, x) = T_m(x) := \frac{\bar{x} - m}{\sqrt{v^*}} \sqrt{n}$ unter $\mathbb{P}_{(m,v)}$ verteilt ist nach der t_{n-1} -Verteilung (siehe Satz 3.9). Insbesondere ist T_m also ein Pivot, und wir erhalten aufgrund der Symmetrie der Student-Verteilung

$$\mathbb{P}_{\vartheta}(T_m \in (-t_{\alpha/2}, t_{\alpha/2})) \geq 1 - \alpha$$

und

$$T_m(x) \in (-t_{\alpha/2}, t_{\alpha/2}) \Leftrightarrow m \in \underbrace{\left(\bar{x} - t_{\alpha/2} \sqrt{\frac{v^*}{n}}, \bar{x} + t_{\alpha/2} \sqrt{\frac{v^*}{n}} \right)}_{:=C(x)}. \quad \square$$

3 Verteilungen rund um die Normalverteilung

Die fundamentale Bedeutung der ein- oder mehrdimensionalen Normalverteilungen ergibt sich aus dem zentralen Grenzwertsatz. Entsprechend spielen Normalverteilungen auch in der Statistik eine herausragende Rolle. Bei der Analyse von Gauß'schen Modellen treten dann weitere Verteilungen natürlich auf, die wir in diesem Abschnitt als Vorbereitung systematisch studieren.

Vereinbarung zur Schreibweise. Verteilungen auf \mathbb{R} (bzw. auf Teilmenge von \mathbb{R}), die absolut stetig gegenüber den Lebesgue-Maß sind, werden im Folgenden geschrieben in der Form

$$\nu(dx) = \frac{1}{Z} g(x) dx$$

mit $Z = \int_{\mathbb{R}} g(x) dx$, d.h. Z ist nur eine Normierungskonstante. Der explizite Zahlwert von Z hängt im Einzelnen von g ab. Diese Abhängigkeit wird in der Notation fortan nicht berücksichtigt.

Satz 3.1. Falls $X \simeq \nu_{0,1}$, so gilt

$$X^2 \simeq \Gamma_{1/2,1/2}$$

Dabei ist $\Gamma_{(b,p)}(dx) = \frac{1}{Z} \cdot \mathbf{1}_{\mathbb{R}_{\geq 0}}(x) e^{-bx} x^{p-1} dx$ die sogenannte Gamma-Verteilung zu den Parametern $p, q > 0$.

Beweis. Sei $X \simeq \nu_{0,1}$. Dann gilt für eine beschränkt messbare Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$:

$$\mathbb{E}(f(X^2)) = \frac{1}{Z} \int_0^\infty e^{-\frac{x^2}{2}} f(x^2) dx \stackrel{y=x^2}{=} \frac{1}{Z} \int_0^\infty e^{-\frac{y}{2}} f(y) \frac{1}{2\sqrt{y}} dy = \mathbb{E}(f(Y))$$

sofern Y eine reelle Zufallsvariable mit $Y \simeq \Gamma_{1/2,1/2}$ ist. □

Bemerkung. Die Normierungskonstante der Γ -Verteilung zu Parametern b, p ist

$$Z(b, p) = \int_{\mathbb{R}_{\geq 0}} e^{-bx} x^{p-1} dx = \frac{\Gamma(p)}{b^p}.$$

(Die Funktion $\Gamma(p) = \int_0^\infty e^{-t} t^{p-1}$ für $p > 0$ ist die Gamma-Funktion. Es gilt $\Gamma(p+1) = (p+1)\Gamma(p)$.)

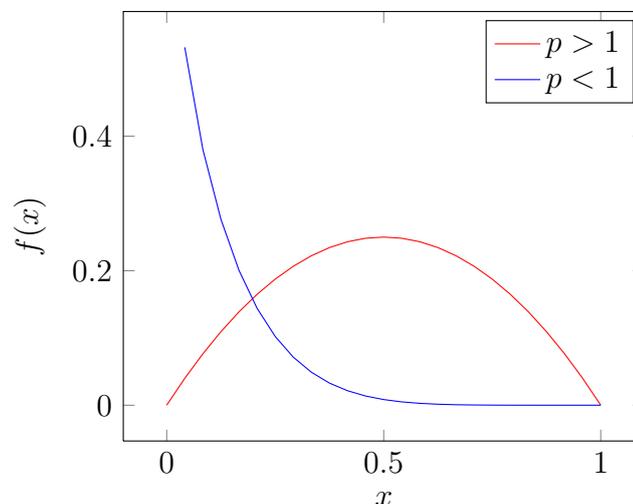
Satz 3.2. Seien $X \simeq \Gamma_{\alpha,r}$ und $Y \simeq \Gamma_{\alpha,s}$ unabhängig. Dann sind die Zufallsgrößen $X+Y$ und $\frac{X}{X+Y}$ stochastisch unabhängig. $\Gamma_{\alpha,r+s}$ bzw. $\beta(r, s)$ verteilt.

Bemerkung. Die Familie der Beta-Verteilungen

$$\beta(p, q)(dx) = \frac{1}{B(p, q)} x^{p-1} (1-x)^{q-1} dx|_{[0,1]}$$

für Parameter $p, q > 0$ und $B(p, q) = \int_0^1 x^{p-1} (1-x)^{q-1} dx$, ist eine Familie von Wahrscheinlichkeitsmaßen auf $[0, 1]$. Je nachdem ob p bzw. q größer oder kleiner als 1 ist, hat die Dichte eine Null- oder eine integrierbare Polstelle bei $x = 0$ bzw. $x = 1$.

Beta-Verteilungen



Beweis von Satz 3.2. Sei $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ beschränkt messbar. Dann gilt:

$$\begin{aligned}
\mathbb{E} \left(f \left(X + Y, \frac{X}{X + Y} \right) \right) &= \frac{1}{Z} \int_{\mathbb{R}_{\geq 0}} \int_{\mathbb{R}_{\geq 0}} f \left(x + y, \frac{x}{x + y} \right) e^{-\alpha x} x^{r-1} \cdot e^{-\alpha y} y^{s-1} dx dy \\
&= \frac{1}{Z} \int_{\mathbb{R}_{\geq 0}} \int_{\mathbb{R}_{\geq 0}} f \left(\underbrace{x + y}_{=:u}, \underbrace{\frac{x}{x + y}}_{=:v} \right) e^{-\alpha(x+y)} \left(\frac{x}{x + y} \right)^{r-1} \left(1 - \frac{x}{x + y} \right)^{s-1} (x + y)^{r+s-2} dx dy \\
&= \frac{1}{Z} \int_{\mathbb{R}_{\geq 0}} \int_{\mathbb{R}_{\geq 0}} f(u, v) e^{-\alpha u} v^{r-1} (1 - v)^{s-1} u^{r+s-2} \cdot \left| \frac{d(x, y)}{d(u, v)} \right| dudv \\
&= \frac{1}{Z} \int_{\mathbb{R}_{\geq 0}} \int_{\mathbb{R}_{\geq 0}} f(u, v) e^{-\alpha u} u^{r+s-1} v^{r-1} (1 - v)^{s-1} dudv = \mathbb{E}(f(U, V)),
\end{aligned}$$

falls U und V zwei stochastisch unabhängige nach $\Gamma_{\alpha, r+s}$ bzw. $\beta_{r,s}$ verteilte Zufallsvariablen sind. Dass U und V hierbei unabhängig zu wählen sind, folgt aus der Struktur der Integrationsdichte $e^{-\alpha u} u^{r+s-1} v^{r-1} (1 - v)^{s-1}$ als Produkt zweier Funktionen in u und v . \square

Bemerkung. Wählen wir $f(x, y) = \varphi(y)$, so gilt

$$\begin{aligned}
\mathbb{E}(\varphi(V)) &= \frac{\alpha^r \alpha^s}{\Gamma(r)\Gamma(s)} \int_{\mathbb{R}_{\geq 0}} e^{-\alpha u} u^{r+s-1} du \int_{[0,1]} \varphi(v) v^{r-1} (1 - v)^{s-1} dv \\
&= \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} \cdot \frac{\Gamma(r + s)}{\alpha^{r+s}} \int_{[0,1]} \varphi(v) v^{r-1} (1 - v)^{s-1} dv
\end{aligned}$$

Als Normierungskonstante der Beta-Verteilung finden wir somit in der Tat

$$B(r, s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r + s)}.$$

Korollar 3.3. Sind X_1, \dots, X_N i.i.d. ('identically independent distributed' bzw. 'identisch unabhängig verteilt') gemäß $\nu_{0,1}$, so gilt

$$\sum_{i=1}^n (X_i)^2 \simeq \Gamma_{1/2, n/2}$$

Beweis. Zunächst liefert Satz 3.2, dass

$$(X_1)^2 \simeq \Gamma_{1/2, 1/2}, \quad (X_2)^2 \simeq \Gamma_{1/2, 1/2}$$

Wenn nun $U \simeq \Gamma_{1/2, n/2}$ und $X_{n+1} \simeq \Gamma_{1/2, 1/2}$, so schließen wir induktiv mit Satz 3.2

$$\sum_{i=1}^{n+1} (X_i)^2 = U + (X_{n+1})^2 \simeq \Gamma_{1/2, n/2+1/2} = \Gamma_{1/2, (n+1)/2}. \quad \square$$

Definition 3.4. $\Gamma_{1/2, n/2}$ heißt χ^2 -Verteilung mit n Freiheitsgraden (kurz χ_n^2 -Verteilung).

Satz 3.5. Seien $X_1, \dots, X_m, Y_1, \dots, Y_n$ i.i.d. $\nu_{0,1}$ -verteilt, dann ist die Zufallsgröße

$$F_{m,n} := \frac{\frac{1}{m} \sum_{i=1}^m (X_i)^2}{\frac{1}{n} \sum_{i=1}^n (Y_i)^2}$$

auf $\mathbb{R}_{\geq 0}$ verteilt gemäß $f_{m,n}(x)dx$, wobei

$$f_{m,n}(x) := \frac{1}{Z} \frac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{m+n}{2}}} \cdot \mathbf{1}_{\{x \geq 0\}}$$

und

$$Z = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)}.$$

Definition 3.6. Die Verteilung zur Dichte $f_{m,n}$ heißt *Fischer-Verteilung* mit (m, n) Freiheitsgraden.

Beweis von Satz 3.5. Die Zufallsgrößen $\tilde{X} := \sum_{i=1}^m X_i^2$ und $\tilde{Y} := \sum_{i=1}^n Y_i^2$ sind unabhängig mit

$$\tilde{X} \simeq \Gamma_{1/2, m/2},$$

$$\tilde{Y} \simeq \Gamma_{1/2, n/2}.$$

Folglich

$$\tilde{Z} := \frac{\tilde{X}}{\tilde{X} + \tilde{Y}} \simeq \beta_{m/2, n/2}.$$

Da

$$F_{m,n} = \frac{n\tilde{X}}{m\tilde{Y}} = \frac{n}{m} \frac{\tilde{Z}}{1-\tilde{Z}}$$

können wir schreiben

$$\begin{aligned} \mathbb{E}(\varphi(F_{m,n})) &= \mathbb{E}\left(\varphi\left(\frac{n}{m} \cdot \frac{\tilde{Z}}{1-\tilde{Z}}\right)\right) \\ &= \frac{1}{Z} \int_0^1 \varphi\left(\underbrace{\frac{n}{m} \cdot \frac{u}{1-u}}_{=:w}\right) u^{\frac{m}{2}-1} (1-u)^{\frac{n}{2}-1} du \\ &= \frac{1}{Z} \int_{\mathbb{R}_{\geq 0}} \varphi(w) u(w)^{\frac{m}{2}-1} (1-u(w))^{\frac{n}{2}-1} \frac{du}{dw} dw, \end{aligned}$$

mit $u = u(w) = \frac{w}{\frac{n}{m} + w}$ und $\frac{du}{dw} = \frac{n \cdot m}{(n+mw)^2}$, was eingesetzt die Behauptung ergibt. \square

Satz 3.7. Seien X, Y_1, \dots, Y_n i.i.d. $\nu_{0,1}$ -verteilt. Dann ist

$$T := \frac{X}{\sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i)^2}}$$

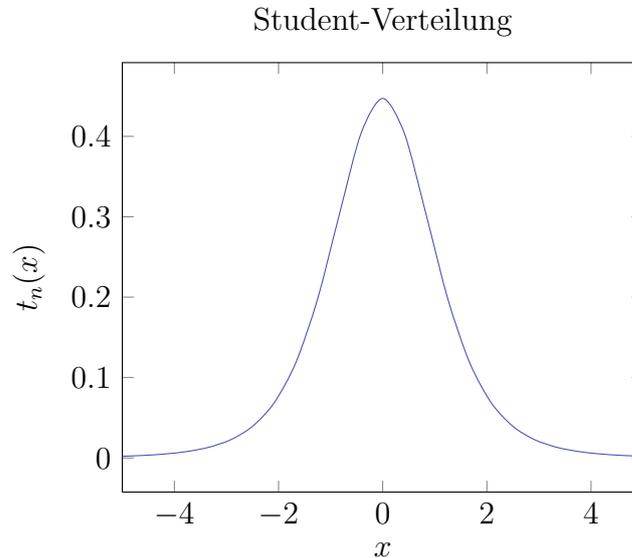
auf \mathbb{R} verteilt mit Dichte

$$t_n(x) = \frac{1}{Z} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

und

$$Z = B(1/2, n/2)\sqrt{n}.$$

Definition 3.8. Die Verteilung $t_n(x)dx$ zur Dichte $t_n(x)$ auf \mathbb{R} heißt *Student-Verteilung* mit n -Freiheitsgraden.



Beweis zu Satz 3.7. Wir stellen zunächst fest, dass T auf \mathbb{R} symmetrisch verteilt ist. Weiter gilt

$$T^2 = \frac{X^2}{\frac{1}{n} \sum_{i=1}^n (Y_i)^2} \simeq F_{1,n}.$$

Damit folgt

$$\begin{aligned} \mathbb{P}(|T| \leq \alpha) &= \mathbb{P}(T^2 \leq \alpha^2) = \mathbb{P}(F_{1,n} \leq \alpha^2) \\ &\Rightarrow \int_{-\alpha}^{\alpha} t_n(x) dx = \int_0^{\alpha^2} f_{1,n}(x) dx. \end{aligned}$$

Da t_n eine gerade Funktion ist, folgt hieraus durch Differentiation nach dem Parameter α , dass

$$t_n(\alpha) = \frac{1}{2} 2\alpha f_{1,n}(\alpha^2) = \alpha f_{1,n}(\alpha^2). \quad \square$$

Satz 3.9 (Student, 1908). Im n -fachen Gauß-Produktmodell $\nu_{m,v}^{\otimes n}$ gilt

1. $\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i$ und $V^* := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ sind stochastisch unabhängig,
2. $\bar{X} \simeq \nu_{m, \frac{v}{n}}$ und $\frac{n-1}{v} V^* \simeq \chi_{n-1}^2$,
3. $T_m = \frac{\sqrt{n}(\bar{X}-m)}{\sqrt{V^*}}$ ist t_{n-1} -verteilt.

Für den Beweis erinnern wir zunächst an einige Eigenschaften von mehrdimensionalen Gaußverteilungen.

Definition 3.10 (Multivariate Normalverteilung). Eine \mathbb{R}^d -wertige Zufallsgröße X heißt Normal-(bzw. Gauß-) verteilt, falls jede Komponente $X_i, i = 1, \dots, d$, von X normalverteilt ist.

Satz 3.11 (Charakterisierung multivariater Normalverteilungen). 1) $X \in \mathbb{R}^d$ ist genau dann normalverteilt, falls gilt

$$\mathbb{P}(X \in dx) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det C}} e^{-\frac{1}{2}\langle (x-m), C^{-1}(x-m) \rangle} dx_1 \dots dx_d$$

mit $C \in \mathbb{R}_{symm, \geq 0}^{d \times d}$ (Kovarianzmatrix) und $m \in \mathbb{R}^d$ (Erwartungswert).

2) Wenn X normalverteilt ist, dann ist die Verteilung eindeutig bestimmt durch

$$\mathbb{E}(X) = m \text{ und } \text{Kov}(X_i, X_j) = \mathbb{E}((x_i - m_i)(x_j - m_j)) = C_{ij} \quad \forall i, j \in \{1, \dots, d\}.$$

Beweis. Siehe Standardliteratur, z.B. Bauer 'Wahrscheinlichkeitstheorie'. \square

Notation. Wir schreiben $\nu_{m,C}$ für eine multivariate Gaußverteilung mit Erwartungswert $m \in \mathbb{R}^d$ und Kovarianzmatrix $C \in \mathbb{R}_{symm, \geq 0}^{d \times d}$.

Korollar 3.12. Falls $X \simeq \nu_{m,C}$ und $A \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^d) = \mathbb{R}^{d \times d}$, gilt

$$Y := AX \simeq \nu_{Am, ACA^T}$$

Beweis. Folgt direkt aus Satz 3.11 (Übung). \square

Beweis von Satz 3.9. $\mathbb{P}_\vartheta = \nu_{\vec{m}, v}^{\otimes n}$ ist als n-faches Produkt von Gauß-Maßen ein multivariates Gauß-Maß mit $\vec{m} = (m, \dots, m)$ und Kovarianz-Matrix $C = \text{diag}(v, \dots, v)$, denn die Komponenten $(X_i)_{i=1, \dots, n}$ sind unabhängig standardnormalverteilt. Sei $B = \{b^1, \dots, b^n\}$ eine Orthonormalbasis von \mathbb{R}^n mit $b_1 = \frac{1}{\sqrt{n}}(1, \dots, 1)$ und sei $T := B^T$ die Matrix mit den Zeilenvektoren aus B . T ist die Matrix für den Basiswechsel im \mathbb{R}^n von der euklidischen Standardbasis zu B .

1. Sei nun $Y := TX$. Da B orthogonal ist, gilt

$$Y \simeq \nu_{T\vec{m}, \text{diag}(v, \dots, v)}$$

Insbesondere sind die Komponenten (Y_1, \dots, Y_n) wieder unabhängig. Zudem

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i = \frac{1}{\sqrt{n}} \langle b^1, X \rangle = \frac{1}{\sqrt{n}} Y_1.$$

Ferner

$$(n-1)V^* = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n \cdot \bar{X}^2 = |X|^2 - Y_1^2 = |Y|^2 - (Y_1)^2 = \sum_{i=2}^n (Y_i)^2.$$

Somit folgt die Unabhängigkeit von \bar{X} und V^* aus der von Y_1, \dots, Y_n .

2. Summen von normalverteilten Zufallsvariablen sind wieder normalverteilt. Somit folgt aus $\mathbb{E}(\bar{X}) = m$, $\text{Var}(\bar{X}) = \frac{1}{n^2} \cdot n \cdot v = \frac{v}{n}$, dass $\bar{X} \simeq \nu_{m, \frac{v}{n}}$.

Zur Bestimmung der Verteilung von V^* können wir wobei o.B.d.A. davon ausgehen, dass $m = 0$, denn bei Ersetzung von X_i durch $\tilde{X}_i = X_i - m$ ändert V^* sich nicht.

Hiermit finden wir

$$\frac{n-1}{v} \cdot V^* = \frac{1}{v} \sum_{i=2}^n (Y_i)^2 = \sum_{i=2}^n \underbrace{(Y_i/\sqrt{v})^2}_{\simeq \nu_{0,1}}$$

Damit folgt

$$\frac{n-1}{v} V^* \simeq \chi_{n-1}^2$$

3. Es gilt

$$T_m = \frac{\sqrt{n}(\bar{X} - m)}{\sqrt{V^*}} = \frac{\sqrt{\frac{n}{v}}(\bar{X} - m)}{\sqrt{\frac{1}{(n-1)v} \sum_{i=2}^n (Y_i)^2}} \simeq \frac{\nu_{0,1}^{(1)}}{\sqrt{\frac{1}{n-1} \sum_{i=2}^n (\nu_{0,1}^{(i)})^2}}$$

mit $\{\nu^{(k)} \mid i = 1, \dots, n\}$ unabhängig $\nu_{0,1}$ -verteilt. Damit ist $T_m \simeq t_{n-1}$. \square

Mit diesen Aussagen folgt nun die Behauptung im Beweis von Satz 2.42.

Satz 2.42. *Im statistischen Modell des n -fachen Gauß-Experiments mit bekannter Varianz v und unbekanntem Erwartungswert m ist für $\alpha \in (0, 1)$ ein Konfidenzbereich für den Parameter m gegeben durch*

$$C(x) = \left(\bar{x} - \eta_{\alpha/2} \sqrt{\frac{v}{n}}; \bar{x} + \eta_{\alpha/2} \sqrt{\frac{v}{n}} \right)$$

mit dem $\frac{\alpha}{2}$ -Fraktile $\eta_{\alpha/2}$ der $\nu_{0,1}$ -Verteilung.

Bemerkung. Man beachte die Unterschiede zwischen den Aussagen von Satz 2.42 und Satz 2.42. Obwohl sich die Formeln für die Konfidenzbereiche auf den ersten Blick gleichen, werden bei unbekannter Varianz v als Ersatz v^* und das $\frac{\alpha}{2}$ -Fraktile der Student-Verteilung benutzt, wohingegen bei bekannter Varianz v und das $\frac{\alpha}{2}$ -Fraktile der Standardnormalverteilung benutzt wird.

Korollar 3.13 (aus Satz 3.9). *Im statistischen Modell des n -fachen Gauß-Experiments mit unbekannter Varianz v und unbekanntem Erwartungswert m ist für $\alpha \in (0, 1)$ und $\alpha_1, \alpha_2 \in (0, 1)$ mit $\alpha_1 + \alpha_2 = \alpha$ ein Konfidenzbereich für den Parameter v gegeben durch*

$$C(x) = \left(\frac{n-1}{t_{\alpha_2}} v^*; \frac{n-1}{s_{\alpha_1}} v^* \right)$$

mit dem α_1 -Quantil s_{α_1} von χ_{n-1}^2 und dem α_2 -Fraktile t_{α_2} von χ_{n-1}^2 .

Beweis. $\pi_{\vartheta}(X) = \frac{n-1}{v}V^*$ ist χ_{n-1}^2 -verteilt, also insbesondere ein Pivot. Damit gilt

$$\mathbb{P}_{\vartheta}(\pi_{\vartheta}(X) \in (s_{\alpha_1}, t_{\alpha_2})) = \chi_{n-1}^2((s_{\alpha_1}, t_{\alpha_2})) \geq 1 - \alpha,$$

gemäß Wahl von s_{α_1} und t_{α_2} . Ferner gilt

$$\pi_{\vartheta}(X) = \frac{n-1}{v}V^* \in (s_{\alpha_1}, t_{\alpha_2}) \Leftrightarrow v \in \left(\frac{n-1}{t_{\alpha_2}}V^*; \frac{n-1}{s_{\alpha_1}}V^* \right). \quad \square$$

Bemerkung. Die Größe $\tilde{\pi}_{\vartheta}(X) = \frac{\sqrt{n}(\bar{X}-m)}{\sqrt{v}} \simeq \nu_{0,1}$ ist ebenfalls ein Pivot (bei unbekanntem v und m).

$$\Rightarrow \mathbb{P}_{\vartheta}(\underbrace{\tilde{\pi}_{\vartheta}(X) \in (-\eta_{\alpha/2}; \eta_{\alpha/2})}_{=:(*)}) \geq 1 - \alpha$$

Dieses Pivot ist aber weniger für die Konstruktion von Konfidenzbereichen geeignet, da sich die hiermit gebildete Bedingung (*) nicht gut nach m bzw. v auflösen lässt.

4 Testen

4.1 Einführung in die Testproblematik

Beispiel (Erfolgparameter einer Münze). Jemand bietet uns ein Wettspiel auf einen Münzwurf an. Für einen Wetteinsatz von einem Euro erhalten wir bei erfolgreichem Ausgang des Münzwurfs drei Euro zurück. Da der Erfolgparameter $p \in [0, 1]$ der Münze uns jedoch unbekannt ist, dürfen wir die Münze zuvor einige Male ausprobieren.

Nach welchem Verfahren sollen wir entscheiden, ob wir die Wette annehmen (d.h. das Spiel spielen)? Offensichtlich wären wir bereit, das Wettspiel zu spielen, falls $p \geq \frac{1}{3}$. Um letzteres zu überprüfen, haben wir nun im wesentlichen lediglich die Möglichkeit, die Münze n mal zu werfen (mit n möglichst groß) und in Abhängigkeit von der Anzahl der beobachteten Erfolge die Vermutung $p < \frac{1}{3}$ oder die Vermutung $p \geq \frac{1}{3}$ aufzustellen und entsprechend das Spiel abzulehnen bzw. anzunehmen.

Eine *Formalisierung* dieses Vorgehens könnte dann wie folgt aussehen:

- Im statistischen Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_{\vartheta}; \vartheta \in \Theta)) = (\{0, \dots, n\}, (\{0, \dots, n\}), (B(n, p))_{p \in [0, 1]})$ (Binomialmodell)
- formulieren wir als *Nullhypothese* die Aussage H_0 : “ $p \in \Theta_0 := [0, \frac{1}{3}]$ ”
- und als *Alternativhypothese* die Aussage H_1 : “ $p \in \Theta_1 := [\frac{1}{3}, 1]$ ”.
- Wir verwenden eine *Teststatistik* $t : \mathfrak{X} \rightarrow [0, 1]$, $t(x) = x/n = \bar{x}$
- für den *Test* $\varphi : \mathfrak{X} \rightarrow [0, 1]$, $\varphi(x) = \mathbb{1}_{t(x) > c}$ mit einem geeigneten Schwellwert c .

Die Aussagen H_0 und H_1 sind einander widersprechende Behauptungen über die Lage des unbekanntem Parameters. Nur eine von beiden kann also richtig sein. Der Test wertet eine Statistik t auf der beobachteten Stichprobe aus und liefert nach Vergleich mit einem

geeignet gewähltem Schwellenwertes c den Indikator 0 oder 1 zurück, je nachdem ob H_0 oder H_1 unterstützt werden.

Zur Vereinfachung der Sprechweise werden wir fortan zwischen Θ_i als Teilmengen von Θ und den entsprechenden Aussage “ $\vartheta \in \Theta_i$ ” nicht unterscheiden. Man lässt im Allgemeinen nun auch noch sogenannte *randomisierte* Tests zu, deren Rückgabewert im gesamten Intervall $\varphi \in [0, 1]$ liegen kann; bei gegebener Realisierung von $\varphi(X)$ entscheidet man sich schließlich noch einmal stochastisch unabhängig mit Wahrscheinlichkeit $\varphi(X)$ für Θ_1 und mit Wahrscheinlichkeit $1 - \varphi(X)$ für Θ_0 .

Definition 4.1 (Parametrischer Test). Seien $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein statistisches Modell und zwei disjunkte Teilmengen $\Theta_0, \Theta_1 \subseteq \Theta$ gegeben. Dann heißt eine Statistik $\varphi : \mathfrak{X} \rightarrow [0, 1]$ *Test* von Θ_0 gegen Θ_1 . Falls φ hierbei Werte im Innern $(0, 1)$ des Einheitsintervalls annimmt, so heißt φ auch ein *randomisierter Test*.

4.2 Gleichmäßig beste Tests

Das Testergebnis hängt gemäß $\varphi(X)$ von der beobachteten Stichprobe X ab und ist somit zufällig. Hierdurch ergeben sich zwei Möglichkeiten, aus der gegebenen Realisierung X eine falsche Schlussfolgerung zu ziehen:

- *Die Nullhypothese wird verworfen, obwohl sie zutrifft.*
Im Münzbeispiel würden wir das Spiel also spielen, obwohl die Münze keinen hinreichend großen Erfolgsparameter hat. (Dies könnte z.B. der Fall sein, wenn wir beim n -fachen Ausprobieren der Münze uncharakteristisch viele Erfolge beobachtet haben.)
- *Die Nullhypothese wird beibehalten, obwohl sie nicht zutrifft.*
Im Münzbeispiel würden wir das Spiel ablehnen, obwohl die Münze einen hinreichend großen Erfolgsparameter hat. (Dies könnte z.B. der Fall sein, wenn wir beim n -fachen Ausprobieren der Münze uncharakteristisch viele Misserfolge beobachtet haben.)

Man könnte nun den ersten Fehler radikal minimieren, indem man als $\varphi(x) \equiv 0$ wählt (0-Test). Dieser würde jedoch niemals die Alternativhypothese Θ_1 vorschlagen, also insbesondere auch dann nicht, wenn Θ_1 zutrifft. Damit ist der 0-Test gänzlich nutzlos. Stattdessen könnte man auch den zweiten Fehler radikal minimieren, indem man als $\varphi(x) \equiv 1$ wählt (1-Test). Dieser würde also stets die Alternativhypothese vorschlagen, also insbesondere auch dann, wenn diese gar nicht zutrifft. Damit ist auch der 1-Test gänzlich nutzlos.

Die beiden Fehler sind also komplementär zueinander in dem Sinne, dass im Allgemeinen die Minimierung der Wahrscheinlichkeit des einen zu einem Ansteigen für das Auftreten des anderen bewirkt. Aus diesem Grund führt man schließlich eine Rangordnung der Fehlerarten ein. Als der gravierendere (‘peinliche’) von beiden Fehlern wird dabei das irrtümliche Verwerfen der Nullhypothese festgelegt. Ein irrtümliches Festhalten an der Nullhypothese (Nichterkennen der Alternativhypothese) wird dagegen als weniger gravierend eingestuft.

Definition 4.2. Eine Entscheidung für Θ_1 , obwohl Θ_0 richtig ist, nennt man *Fehler erster Art*. Eine Entscheidung für Θ_0 , obwohl Θ_1 richtig ist, nennt man *Fehler zweiter Art*.

Zur besseren Unterscheidung werden wir gelegentlich vom Fehler erster Art als 'peinlichen' und dem Fehler zweiter Art als 'verzeihlichen' Fehler sprechen.

Im vorausgegangenen Beispiel des Spiels mit der unbekanntem Münze hatten wir als Nullhypothese " $p < \frac{1}{3}$ " und entsprechend " $p \geq \frac{1}{3}$ " als Alternativhypothese gewählt. Damit ist der peinliche Fehler festgelegt darin, das Spiel zu spielen, obwohl die Münze nicht gut ist. Die Vermeidung dieses Fehlers soll also einen höheren Stellenwert haben als des Fehlers zweiter Art, nämlich das Spiel abzulehnen, obwohl die Münze gut ist. Hätten wir die Rollen von Null- und Alternativhypothese vertauscht, bestünde entsprechend der peinliche Fehler darin, das Spiel abzulehnen, obwohl die Münze gut ist.

Die Zuordnung der beiden disjunkten Teilbereiche von Θ als als Null- bzw. Alternativhypothese kann also beliebig vorgenommen werden und dient allein der Festlegung, welcher der beiden (zunächst symmetrischen) Testfehler als peinlich bzw. als verzeihlich angesehen wird.

Definition 4.3. Ein Test φ heißt *zulässig zum Irrtumsniveau* $\alpha \in [0, 1]$, falls

$$\sup_{\vartheta \in \Theta_0} \mathbb{E}_{\vartheta}(\varphi) \leq \alpha.$$

Ein zulässiger Test zum Niveau α ist mithin dadurch charakterisiert, dass einen peinlichen Fehler mit maximaler Wahrscheinlichkeit α produziert. Unter allen zu einem gegebenen Niveau α zulässigen Tests interessieren wir uns nun für diejenigen, die mit minimaler Wahrscheinlichkeit verzeihliche Fehler produzieren, d.h. welche auf der 'schlechten' Parametermenge Θ_1 besonders häufig einen Wechsel zur (dann korrekten) Alternativhypothese Θ_1 vorschlagen.

Definition 4.4. φ ist zum Niveau α ein *gleichmäßig bester Test*, falls

$$\forall \vartheta \in \Theta_1 : \mathbb{E}_{\vartheta}(\varphi) \geq \mathbb{E}_{\vartheta}(\psi)$$

für alle zulässigen Test $\psi : \mathfrak{X} \rightarrow [0, 1]$ von Θ_0 gegen Θ_1 zum Niveau α .

Bemerkung. Zu einem Test φ heißt die Funktion $G_{\varphi} : \Theta \rightarrow [0, 1]$, $G_{\varphi}(\vartheta) = \mathbb{E}_{\vartheta}(\varphi)$ *Gütefunktion*. Ein Test φ ist somit dann "gut", falls $G \upharpoonright_{\Theta_0}$ "klein" und $G \upharpoonright_{\Theta_1}$ "groß" ist.

Beispiel. Betrachte das n -fache Gauß-Produktmodell mit bekannter Varianz v und unbekanntem Erwartungswert m , $\Theta = \mathbb{R}$. Seien nun $\Theta_0 = \{m_0\}$ und $\Theta_1 = \mathbb{R} \setminus \{m_0\}$.

$$T(x) := \sqrt{n} \frac{\bar{x} - m_0}{\sqrt{v}}$$

ist unter $\mathbb{P}_{m_0} \nu_{0,1}$ -verteilt. Sei η das $\frac{\alpha}{2}$ -Fraktile von $\nu_{0,1}$ und

$$\varphi(x) := \mathbb{1}_{\{T(x) \notin (-\eta, \eta)\}}$$

Dann ist φ ein Test von Θ_0 gegen Θ_1 zum Niveau α , denn

$$\sup_{\vartheta \in \Theta_0} \mathbb{E}_{\vartheta}(\varphi) = \mathbb{E}_{m_0}(\varphi) = \mathbb{P}_{m_0}(\varphi = 1) = \mathbb{P}_{m_0}(T(x) \notin (-\eta, \eta)) = \alpha \leq \alpha$$

Die Gütefunktion dieses Tests berechnet sich dann wie folgt

$$\begin{aligned}
G(\vartheta) &= G(m) = \mathbb{E}_m(\varphi) = \mathbb{P}_m(|T| \geq \eta) = \mathbb{P}_m\left(\left|\sqrt{n}\frac{\bar{x} - m}{\sqrt{v}} - \sqrt{n}\frac{m_0 - m}{\sqrt{v}}\right| \geq \eta\right) \\
&= 1 - \mathbb{P}_m\left(\left|\sqrt{n}\frac{\bar{x} - m}{\sqrt{v}} - \sqrt{n}\frac{m_0 - m}{\sqrt{v}}\right| < \eta\right) \\
&= 1 - \left[\Phi\left(\sqrt{n}\frac{m - m_0}{\sqrt{v}} + \eta\right) - \Phi\left(\sqrt{n}\frac{m - m_0}{\sqrt{v}} - \eta\right)\right] \\
&= \Phi\left(\sqrt{n}\frac{m - m_0}{\sqrt{v}} - \eta\right) + \Phi\left(-\sqrt{n}\frac{m - m_0}{\sqrt{v}} - \eta\right) \\
&= \Phi\left(\sqrt{n}\frac{m - m_0}{\sqrt{v}} - \eta\right) + \Phi\left(\sqrt{n}\frac{m_0 - m}{\sqrt{v}} - \eta\right)
\end{aligned}$$

wobei Φ wieder die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Aus $\Phi(t) \rightarrow 1$ für $t \rightarrow \infty$ und $\Phi(t) \rightarrow 0$ für $t \rightarrow -\infty$ folgert man, dass in der Tat $G(m) \rightarrow 1$, falls $|m - m_0| \rightarrow \infty$.

4.3 Das Neymann-Pearson Lemma

Das folgende Resultat liefert (gemeinsam mit der anschließenden Ergänzung) eine Charakterisierung aller gleichmäßig besten Tests in dem Spezialfall, dass die Null- und Alternativhypothesen zu einelementigen Parametermengen (d.h. $\#\Theta_0 = \#\Theta_1 = 1$) formuliert sind.

Satz 4.5 (Neyman-Pearson). *Sei $M = (\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_0, \mathbb{P}_1))$ ein reguläres statistisches Modell mit dominantem Maß μ_0 , d.h.*

$$\mathbb{P}_0(dx) = \varrho(0, x)\mu_0(dx)$$

$$\mathbb{P}_1(dx) = \varrho(1, x)\mu_0(dx)$$

Sei $R(x) := \frac{\varrho(1, x)}{\varrho(0, x)}$, dann ist für $c \geq 0$

$$\varphi(x) := \mathbb{1}_{\{R(x) \geq c\}}$$

ein gleichmäßig bester Test von $\Theta_0 = \{0\}$ gegen $\Theta_1 = \{1\}$ zum Niveau $\alpha = \mathbb{E}_0(\varphi)$.

Bemerkung. Der Ausdruck $R(x) = \frac{\varrho(1, x)}{\varrho(0, x)}$ heißt *likelihood-Quotient*. Entsprechend nennt man einen Test der Form $\varphi(x) = \mathbb{1}_{\{R(x) \geq c\}}$ mit $R = \frac{\varrho_1}{\varrho_0}$ einen *likelihood-Quotienten-Test*.

Beweis von Satz 4.5. Nach Definition von α ist φ ein Test zum Niveau α . Sei also $\psi : \mathfrak{X} \rightarrow [0, 1]$ ein weiterer Test von Θ_0 gegen Θ_1 zum Niveau α . Für $x \in \mathfrak{X}$ beliebig machen wir eine Fallunterscheidung.

1. Fall: $\varrho(1, x) - c \cdot \varrho(0, x) \geq 0$
 $\Rightarrow \varphi(x) = 1 \Rightarrow \varphi(x) - \psi(x) \geq 0$
2. Fall: $\varrho(1, x) - c \cdot \varrho(0, x) < 0$
 $\Rightarrow \varphi(x) = 0 \Rightarrow \varphi(x) - \psi(x) \leq 0$

Insgesamt gilt also

$$f(x) := [\varphi(x) - \psi(x)] \cdot [\varrho(1, x) - c \cdot \varrho(0, x)] \geq 0 \quad \forall x \in \mathfrak{X}.$$

Integration bzgl. μ_0 ergibt

$$0 \leq \int_{\mathfrak{X}} f \, d\mu_0 = \mathbb{E}_1(\varphi(X)) - \mathbb{E}_1(\psi(X)) - c \cdot [\mathbb{E}_0(\varphi(X)) - \mathbb{E}_0(\psi(X))].$$

Also gilt

$$\mathbb{E}_1(\varphi(X)) - \mathbb{E}_1(\psi(X)) \geq \underbrace{c \mathbb{E}_0(\varphi(X))}_{=\alpha} - \underbrace{c \mathbb{E}_0(\psi(X))}_{\leq \alpha} \geq 0,$$

so dass schließlich $\mathbb{E}_1(\varphi(X)) \geq \mathbb{E}_1(\psi(X))$ wie behauptet. \square

Satz 4.6 (Struktur von gleichm. besten Tests). 1. Für jeden anderen gleichmäßig besten Test mit $\mathbb{E}_0(\psi) = \alpha$ gilt

$$\varphi(x) = \psi(x) \quad \text{für } \mu_0\text{-fast alle } x \in \{R \neq c\}.$$

2. Zu $\alpha \in (0, 1)$ existieren $c \geq 0, \gamma \in [0, 1]$ so, dass

$$\varphi(x) = \mathbf{1}_{\{R(x) > c\}} + \gamma \mathbf{1}_{\{R(x) = c\}}$$

ein gleichmäßig besten Test zum Niveau α definiert. Dabei sind

$$c := \inf \{t \in \mathbb{R} \mid \nu((t, \infty)) < \alpha\},$$

$$\gamma := \begin{cases} 0 & \text{falls } \nu(\{c\}) = 0 \\ \frac{\alpha - \nu((c, \infty))}{\nu(\{c\})} & \text{sonst} \end{cases}$$

mit dem Bildmaß $\nu = \mathbb{P}_0 \circ R^{-1}$ von \mathbb{P}_0 unter $R : \mathfrak{X} \mapsto \mathbb{R}$.

Beweis. 1) Falls $\mathbb{E}_0(\psi) = \alpha$ und ψ ebenfalls ein bester Test zum Niveau α ist, so gilt offensichtlich $\mathbb{E}_1(\psi) = \mathbb{E}_1(\varphi)$. Somit finden wir für das μ_0 -Integral der nichtnegativen Funktion $f(x) = [\varphi(x) - \psi(x)] \cdot [\varrho(1, x) - c \cdot \varrho(0, x)]$, dass $\int_{\mathfrak{X}} f(x) \mu_0(dx) = 0$, also $f(x) = 0$ für μ_0 -fast alle $x \in \mathfrak{X}$. Nach Definition von f heißt das

$$\varphi(x)(\varrho(1, x) - c\varrho_0(x)) = \psi(x)(\varrho(1, x) - c\varrho_0(x)) \quad \mu_0\text{-fast sicher,}$$

also auch $\varphi(x) = \psi(x)$ für μ_0 -fast alle $x \in \mathfrak{X} \setminus \{R = c\}$.

2) Sei $\nu = \mathbb{P}_0 \circ R^{-1}$ das Bildmaß von \mathbb{P}_0 unter der Abbildung \mathbb{R} . Sei $c \in \mathbb{R}$ das minimale α -Quantil von ν , d.h.

$$\mathbb{P}_0(R \geq c) \geq \alpha \quad \text{und} \quad \mathbb{P}_0(R > c) < \alpha.$$

Setze

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

wobei

$$\gamma = \begin{cases} 0 & \text{falls } \mathbb{P}_0(R = c) = 0 \\ \frac{\alpha - \mathbb{P}_0(R > c)}{\mathbb{P}_0(R = c)} & \text{falls } \mathbb{P}_0(R = c) > 0. \end{cases}$$

Dann ist φ ein zulässiger Test zum Niveau α , denn

$$\mathbb{E}_0(\varphi) = \gamma \mathbb{P}_0(R = c) + \mathbb{P}_0(R > c) = \alpha.$$

Sei ψ ein weiterer zulässiger Test von \mathbb{P}_0 vs. \mathbb{P}_1 zum Niveau α , dann gilt (s.o.) dass $f(x) = [\varphi(x) - \psi(x)] \cdot [\varrho(1, x) - c \cdot \varrho(0, x)] \geq 0$, woraus wie gesehen folgt, dass $\mathbb{E}_1(\varphi) - \mathbb{E}_1(\psi) \geq c(\mathbb{E}_0(\varphi) - \mathbb{E}_0(\psi)) \geq 0$, d.h. $\mathbb{E}_1(\varphi) \geq \mathbb{E}_1(\psi)$, d.h. φ ist ein gleichm. bester Test zum Niveau α . \square

4.4 Likelihood-Quotienten-Tests

Eine erste Verallgemeinerung des Neyman-Pearson Lemmas auf kompliziertere als einelementige Parametermengen ist im Fall von monotonen statistischen Modellen möglich.

Definition 4.7. Ein reguläres statistisches Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ mit $\Theta \subseteq \mathbb{R}$ heißt *monoton wachsend* bzgl. der Statistik $T : \mathfrak{X} \rightarrow \mathbb{R}$, falls für alle $\vartheta_1, \vartheta_2 \in \Theta$ mit $\vartheta_1 \geq \vartheta_2$ eine streng monoton wachsende Funktion

$$f_{\vartheta_1, \vartheta_2} : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$$

existiert mit

$$\frac{\varrho(\vartheta_1, x)}{\varrho(\vartheta_2, x)} = f_{\vartheta_1, \vartheta_2}(T(x)) \mu_0\text{-fast überall.}$$

Beispiel. Im Falle einer exponentiellen Familie $\varrho(\vartheta, x) = \frac{1}{z(\vartheta)} e^{\lambda(\vartheta) \cdot t(x)}$ finden wir

$$\frac{\varrho(\vartheta_1, x)}{\varrho(\vartheta_2, x)} = \frac{z(\vartheta_2)}{z(\vartheta_1)} \cdot e^{(\lambda(\vartheta_1) - \lambda(\vartheta_2)) \cdot t(x)}.$$

Die exponentielle Familie ist monoton wachsend bzgl. t , falls λ streng monoton fällt. Für λ in t streng monoton wachsend, ist die exponentielle Familie wachsend bzgl. $-t$.

Satz 4.8. Sei $\Theta \subseteq \mathbb{R}$ und $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein bzgl. $T : \mathfrak{X} \rightarrow \mathbb{R}$ monoton wachsendes statistisches Modell. Sei

$$\Theta_0 := \{\vartheta \in \Theta \mid \vartheta \leq \vartheta_0\}$$

und

$$\Theta_1 := \{\vartheta \in \Theta \mid \vartheta \geq \vartheta_1\}$$

mit $\vartheta_0 < \vartheta_1$. Für $\vartheta_2 > \vartheta_0$, $\vartheta_2 \in \Theta$ sei

$$R(x) = \frac{\varrho(\vartheta_2, x)}{\varrho(\vartheta_0, x)}$$

und

$$\varphi(x) = \mathbf{1}_{\{R \geq c\}}$$

Dann ist φ ein gleichmäßig bester Test von Θ_0 gegen Θ_1 zum Niveau

$$\alpha := \mathbb{E}_{\vartheta_0}(\varphi).$$

Bemerkung. Man beachte, dass die Konstruktion von φ nicht von ϑ_1 abhängt.

Beweis von 4.8.

$$1. R(x) \geq c \Leftrightarrow T(x) \geq f_{\vartheta_2, \vartheta_0}^{-1}(c) =: k$$

$$2. \text{ Sei } \tilde{\vartheta} > \vartheta_0$$

$$T(x) \geq k \Leftrightarrow f_{\tilde{\vartheta}, \vartheta_0}(T(x)) \geq f_{\tilde{\vartheta}, \vartheta_0}(k) = \tilde{c}$$

$$T(x) \geq k \Leftrightarrow \frac{\varrho(\tilde{\vartheta}, x)}{\varrho(\vartheta_0, x)} \geq \tilde{c}$$

$$\varphi = \mathbb{1}_{\{T(x) \geq k\}} = \mathbb{1}_{\left\{\frac{\varrho(\tilde{\vartheta}, x)}{\varrho(\vartheta_0, x)} \geq \tilde{c}\right\}}$$

$\Rightarrow \varphi$ ist ein Neyman-Pearson Test für P_{ϑ_0} gegen $P_{\tilde{\vartheta}}$ zum Niveau $\mathbb{E}_{\vartheta_0}(\varphi) = \alpha$.

$\Rightarrow \varphi$ ist simultan ein Neyman-Pearson-optimaler Test für alle Testprobleme

$$P_{\vartheta_0} \text{ vs. } P_{\tilde{\vartheta}} \quad \forall \tilde{\vartheta} > \vartheta_0$$

$\Rightarrow \varphi$ ist ein gleichmäßig bester Test für $\{P_{\vartheta_0}\}$ gegen $\{P_{\vartheta} \mid \vartheta > \vartheta_1\}$.

3. Behauptung: $\vartheta \mapsto \mathbb{E}_{\vartheta}(\varphi)$ ist nicht-fallend.

Seien $\vartheta < \vartheta'$, $\vartheta, \vartheta' \in \Theta$.

$$T(x) \geq k \Leftrightarrow \frac{\varrho(\vartheta', x)}{\varrho(\vartheta, x)} \geq c'$$

Ferner gilt

$$\begin{aligned} & (\varrho(\vartheta', x) - c' \varrho(\vartheta, x)) \cdot (\varphi(x) - b) \geq 0 \quad \forall b \in [0, 1] \\ \Leftrightarrow & \varrho(\vartheta', x) - c' \varrho(\vartheta, x) \geq (\varrho(\vartheta', x) - c' \varrho(\vartheta, x)) b \\ \Rightarrow & \mathbb{E}_{\vartheta'}(\varphi) - c' \mathbb{E}_{\vartheta}(\varphi) \geq (1 - c') b \end{aligned}$$

Wähle nun $b = \mathbb{E}_{\vartheta}(\varphi)$

$$\Rightarrow \mathbb{E}_{\vartheta'}(\varphi) \geq \mathbb{E}_{\vartheta}(\varphi)$$

4. Aus 2. und 3. folgt dass φ ist ein gleichmäßig bester Test von $\Theta_0 = \{P_{\vartheta} \mid \vartheta \leq \vartheta_0\}$ gegen $\Theta_1 = \{P_{\vartheta} \mid \vartheta \geq \vartheta_1\}$ zum Niveau $\alpha = \mathbb{E}_{\vartheta_0}(\varphi)$, denn

$$(i) \sup_{\vartheta \in \Theta_0} \mathbb{E}_{\vartheta}(\varphi) \stackrel{3.}{=} \mathbb{E}_{\vartheta_0}(\varphi) = \alpha$$

(ii) Falls ψ zulässig, d.h.

$$\sup_{\vartheta \in \Theta_0} \mathbb{E}_{\vartheta}(\psi) \leq \alpha \Rightarrow \mathbb{E}_{\vartheta_0}(\psi) \leq \alpha$$

Insbesondere ist ψ zulässig für $\{P_{\vartheta_0}\}$ gegen $\{P_{\vartheta} \mid \vartheta \geq \vartheta_1\}$. Hierfür ist φ jedoch optimal (2.)

$$\Rightarrow \mathbb{E}_{\vartheta}(\varphi) \geq \mathbb{E}_{\vartheta}(\psi) \quad \forall \vartheta \geq \vartheta_1$$

Somit ist φ optimal. □

Die vorausgehenden Resultate motivieren nun den folgenden generellen Ansatz zur Konstruktion von Tests, von denen man dann im Einzelfall die Optimalität prüfen muss.

Definition 4.9 (Verallgemeinerte likelihood-ratio-Methode). Sei $M = (\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ ein reguläres statistisches Modell, $\Theta_0 \subseteq \Theta$, $\Theta_1 \subseteq \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$. Dann heißt $R : \mathfrak{X} \rightarrow \mathbb{R}_{\geq 0} \cup \{\infty\}$

$$R(x) = \frac{\sup_{\vartheta \in \Theta_1} \varrho(\vartheta, x)}{\sup_{\vartheta \in \Theta_0} \varrho(\vartheta, x)}$$

verallgemeinerter likelihood-Quotient. Ein Test der Form

$$\varphi(x) = \mathbb{1}_{\{R(x) \geq c\}}$$

heißt verallgemeinerter likelihood-Quotienten-Test von Θ_0 gegen Θ_1 .

Beispiel. Als Beispiel für dieses Schema betrachten wir das n -fache Gauß'sche Produktmodell M mit unbekanntem (m, v) und

$$\Theta_0 = \{v \leq v_0\},$$

$$\Theta_1 = \{v > v_0\}.$$

Der verallgemeinerte likelihood-Quotient ist hier

$$R(x) = \frac{\sup_{\substack{m \in \mathbb{R} \\ v > v_0}} \varrho((m, v), x)}{\sup_{\substack{m \in \mathbb{R} \\ v \leq v_0}} \varrho((m, v), x)}.$$

Wegen

$$\sup_{m \in \mathbb{R}} e^{-\frac{\sum (x_i - m)^2}{2v}} = e^{-\frac{nV(x)}{2v}}$$

mit $V(x) = \frac{1}{n} \sum_i (x_i - \bar{x})^2$, also bei $m = \bar{x}$, ist

$$R(x) = \frac{\sup_{v > v_0} v^{-n/2} e^{-\frac{nV(x)}{2v}}}{\sup_{v \leq v_0} v^{-n/2} e^{-\frac{nV(x)}{2v}}} = \begin{cases} e^{\frac{n}{2} \left[\frac{V(x)}{v_0} - \ln\left(\frac{V(x)}{v_0}\right) - 1 \right]} & \text{falls } V(x) > v_0 \\ e^{-\frac{n}{2} \left[\frac{V(x)}{v_0} - \ln\left(\frac{V(x)}{v_0}\right) - 1 \right]} & \text{falls } V(x) \leq v_0. \end{cases}$$

$R(x)$ ist somit eine strikt monoton wachsende Funktion von $V(x) = \frac{1}{n} \sum_i (x_i - \bar{x})^2$.

Der verallgemeinerte likelihood-Quotienten-Ansatz für das Testproblem $\{v \leq v_0\}$ gegen $\{v > v_0\}$ liefert also im n -fachen Gauß-Modell einen Test der Form

$$\varphi(x) = \mathbb{1}_{\{V^*(x) \geq \hat{c}\}}$$

mit $V^* = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$. Der Beweis des folgenden Satzes veranschaulicht, dass der Nachweis der Optimalität mitunter sehr kompliziert sein kann.

Satz 4.10. *Im n -fachen Gaußmodell mit (m, v) unbekannt ist*

$$\varphi(x) = \mathbb{1}_{\{\sum (x_i - \bar{x})^2 \geq v_0 \cdot \eta\}}$$

ist ein gleichmäßig bester Test für $\{v \leq v_0\}$ gegen $\{v > v_0\}$ zum Niveau α , wobei η das α -Fraktile der χ_{n-1}^2 -Verteilung sei.

Beweis. φ ist zulässig, denn falls $\vartheta = (m, v) \in \Theta_0$, gilt

$$\begin{aligned}\mathbb{E}_{\vartheta}(\varphi) &= \mathbb{P}_{(m,v)} \left(\sum_{i=1}^n (X_i - \bar{X})^2 \geq v_0 \cdot \eta \right) \\ &= \mathbb{P}_{(m,v)} \left(\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{v} \geq \frac{v_0}{v} \eta \right) \\ &= \chi_{n-1}^2 \left(\left[\frac{v_0}{v} \eta; \infty \right) \right) \\ &\leq \chi_{n-1}^2 ([\eta, +\infty)) \\ &= \alpha\end{aligned}$$

da $\frac{v_0}{v} \geq 1$.

φ ist optimal. Denn sei $\vartheta_1 = (m_1, v_1) \in \Theta_1$. Für $v \leq v_1$ sei

$$\bar{P}_v(dx) = \int_{\mathbb{R}} w_v(dm) P_{(m,v)}(dx)$$

mit

$$w_v = \begin{cases} \nu_{m_1, \frac{v_1-v}{n}} & \text{falls } v < v_1 \\ \delta_{m_1} & \text{falls } v = v_1. \end{cases}$$

(Mischung der Familie $(P_{(m,v)})$ bezüglich dem Parameter $m \in \mathbb{R}$ mit dem Maß $w_v(dm)$)

Wir haben dadurch ein neues statistisches Modell

$$M' = (\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n), (\bar{P}_v; v \leq v_1))$$

Ferner gilt (Übung)

$$\bar{P}_v(dx) = \begin{cases} \frac{1}{z(v)} e^{-\frac{n-1}{v} V^*(x) - \frac{(m_1 - \bar{x})^2}{2v_1/n}} & , \text{ falls } v < v_1 \\ \nu_{m_1, v_1}^{\otimes n} & , \text{ falls } v = v_1 \end{cases}$$

M' ist also eine exponentielle Familie zur Statistik $t(x) = V^*(x)$ mit wachsender Koeffizientenfunktion

$$\lambda(v) = -\frac{n-1}{2v}.$$

Somit befinden wir uns in der Neyman-Pearson-Situation mit monotonem likelihood-Verhältnis, d.h. es existiert ein eindeutiger optimaler Test in $(\bar{P}_v, v \leq v_1)$ für $\{\bar{P}_v, v \leq v_0\}$ vs. $\{\bar{P}_{v_1}\}$, nämlich

$$\psi(x) = \mathbf{1}_{\{V^*(x) \geq c\}}.$$

Sei α das Niveau, dann soll gelten

$$\alpha = \bar{P}_{v_0}(\psi(X) = 1) = \bar{P}_{v_0}(V^*(X) \geq c) = P_{v_0, m}(V^* \geq c)$$

falls $c = \frac{v_0 \cdot \eta}{n-1}$.

Folglich ist $\psi(x) = \varphi(x)$ der optimale Test von $\{\bar{P}_v; v \leq v_0\}$ gegen $\{\bar{P}_{v_1}\}$ zum Niveau α

im modifizierten statistischen Modell M' . Sei nun $\zeta : \mathfrak{X} \rightarrow [0, 1]$ ein zulässiger Test von $\{P_{(m,v)}; v \leq v_0\}$ gegen $\{P_{(m,v)}; v > v_0\}$ zum Niveau α . Wir interpretieren ζ als Test von $\{\bar{P}_v; v \leq v_0\}$ gegen $\{\bar{P}_{v_1}\}$. Dann ist ζ als solcher zulässig (da $\mathbb{E}_{\bar{P}_v}(\zeta) \leq \alpha$). Wegen der Optimalität von φ gilt ferner

$$\mathbb{E}_{P_{(m_1, v_1)}}(\varphi) = \mathbb{E}_{\bar{P}_{v_1}}(\varphi) \geq \mathbb{E}_{\bar{P}_{v_1}}(\zeta) = \mathbb{E}_{P_{(m_1, v_1)}}(\zeta).$$

Da $v_1 > v_0$ beliebig gewählt wurde gilt also auch

$$\mathbb{E}_{P_{\vartheta_1}}(\varphi) \geq \mathbb{E}_{P_{\vartheta_1}}(\zeta) \quad \forall \vartheta_1 \in \Theta_1.$$

Also ist φ auch optimal für das ursprüngliche Testproblem. \square

Zusammenfassung

1. Es wird eine Priorisierung von peinlichen versus verzeihlichen Fehlern eines Tests vorgenommen, die letztlich durch die Zuordnung von Null- und Alternativhypothese zweier disjunkter Teilbereiche von Θ festgelegt wird.
2. Das Niveau eines Tests ist die maximale Wahrscheinlichkeit für einen peinlichen Fehler.
3. Das Niveau sagt nichts über die Alarmgenauigkeit (im Sinne der Wahrscheinlichkeit, die Alternativvermutung korrekt anzuzeigen) des Tests aus.
4. Die Nachweis der Optimalität (im Sinne von Alarmgenauigkeit) ist im Allgemeinen sehr schwierig (\rightsquigarrow Neyman-Pearson)!

5 Nichtparametrische Modelle

Definition 5.1. Ein statistisches Modell $(\mathfrak{X}, \mathfrak{F}, (\mathbb{P}_\vartheta; \vartheta \in \Theta))$ heißt *parametrisch*, falls $\Theta \subseteq \mathbb{R}^d$, andernfalls heißt es *nichtparametrisch*, d.h. also insbesondere falls $\Theta \subseteq F$ lediglich in einer abstrakten oder unendlich-dimensionalen Menge F enthalten ist.

Beispiel. Seien X_1, \dots, X_n unabhängige Realisierungen bzw. Stichproben für eine Wahrscheinlichkeitsverteilung ϱ auf einem messbaren Raum $(\mathfrak{E}, \mathcal{E})$. Falls keine weiteren Strukturannahmen an die Verteilung von X_i gemacht werden, wäre das zugehörige statistische Modell das n -fache Produktmodell $M^{\otimes n}$ von

$$M = (\mathfrak{E}, \mathcal{E}, \{\rho \mid \rho \text{ ist ein W-Maß auf } (\mathfrak{E}, \mathcal{E})\}).$$

Im Spezialfall eines diskreten Raumes $\mathfrak{E} \subset \mathbb{N}$ ist die Menge $\{\rho \mid \rho \text{ ist ein W-Maß auf } (\mathfrak{E}, \mathcal{E})\}$ gleich dem unendlich-dimensionalen Simplex aller Folgen $\varrho : \mathfrak{E} \rightarrow [0, 1]$ mit $\sum_{i \in \mathfrak{E}} \varrho(i) = 1$.

Um nun von den Realisierungen X_1, \dots, X_n auf ϱ zu schließen, benutzt man z.B. das *starke Gesetz der Großen Zahlen*. Sei hierzu $f : E \rightarrow \mathbb{R}$ eine beschränkte messbare

Testfunktion. Dann liefert die Anwendung des Gesetzes der großen Zahlen auf die Folge der unabhängigen Zufallsvariablen $Z_i := f(X_i)$, dass

$$\frac{1}{n} \sum_{i=1}^n Z_i \xrightarrow{n \rightarrow \infty} \mathbb{E}(Z) \text{ fast sicher}$$

mit $\mathbb{E}(Z) = \mathbb{E}(Z_1) = \dots = \sum_{i \in E} f(i) \varrho(i)$. Wählt man also z.B.

$$f(i) = \begin{cases} 1, & \text{falls } i = i_0 \\ 0, & \text{falls } i \neq i_0 \end{cases}$$

für ein beliebiges aber festes $i_0 \in E$, so erhält man

$$\frac{1}{n} \sum_{i=1}^n Z_i \rightarrow \mathbb{E}(Z) = \varrho(i_0) \text{ fast sicher.}$$

Durch die Auswertung unendlich vieler unabhängiger Stichproben erhält man die vollständige Information über die zugrundeliegende Verteilung ρ .

5.1 Der Satz von Glivenko-Cantelli

Im kontinuierlichen Fall $\mathfrak{E} \subset \mathbb{R}$ kann man Analog zum diskreten Fall vorgehen durch Wahl

$$f : \mathfrak{E} \rightarrow \{0, 1\}, f(s) = \mathbb{1}_{]-\infty, t]}(s),$$

zu beliebig aber fest gewählten $t \in \mathbb{R}$. Die Anwendung des starken Gesetzes der Großen Zahlen auf die i.i.d. Folge $Z_i := F(X_i)$, $i \in \mathbb{N}$, liefert in diesem Fall sofort die folgende gute Nachricht.

Satz 5.2. *Es sei X_i eine i.i.d. Folge von reellen Zufallsvariablen mit Verteilung $X_1 \simeq \varrho$ und $t \in \mathbb{R}$, dann gilt für $n \rightarrow \infty$*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{]-\infty, t]}(X_i) \longrightarrow \rho(]-\infty, t]) \text{ fast sicher.}$$

Anders ausgedrückt besagt dieses Resultat, dass die Folge der (zufälligen) *empirischen Verteilungsfunktionen*

$$F_n : \mathbb{R} \rightarrow \mathbb{R}, \quad F_n(t) := \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{(-\infty, t]}(X_k)$$

fast sicher punktweise gegen die zugrundeliegende Verteilungsfunktion $F(t) := \rho(]-\infty, t])$ konvergiert. In der Tat liegt sogar gleichmäßige Konvergenz vor.

Satz 5.3 (Glivenko-Cantelli). *Es seien X_1, X_2, \dots unabhängig identisch verteilte Zufallsvariablen auf \mathbb{R} mit der Verteilungsfunktion*

$$F(x) = \mathbb{P}(X_1 \leq x)$$

Sei

$$F_n(t) := \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{(-\infty, t]}(X_k),$$

dann gilt

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \rightarrow 0 \quad \text{fast sicher.}$$

Bemerkung. Im Sinne unserer bisherigen Systematik von statistischen Modellen wäre für festes $n \in \mathbb{N}$

$$M = (\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n), (\mathbb{P}_F^{\otimes n})_{F \in \mathfrak{F}})$$

mit $\mathbb{P}_F(X \leq x) = F(x)$. \mathfrak{F} ist dabei die Menge aller Verteilungsfunktionen auf \mathbb{R} . Da die fast sichere Konvergenz die Konvergenz in Wahrscheinlichkeit impliziert, können wir aus dem Satz von Glivenko-Cantelli folgern, dass durch $(F_n)_n$ eine konsistente Folge von Schätzern für F definiert ist, sofern Schätzfehler in der Supremumsnorm auf reellen Funktionen gemessen werden.

Beweis von Satz 5.3. Zu $t \in \mathbb{R}$ sei $Y_i := \mathbf{1}_{(-\infty, t)}(X_i)$ und $Z_i := \mathbf{1}_{(-\infty, t]}(X_i)$, dann gilt nach dem starken Gesetz der großen Zahlen, dass für $n \rightarrow \infty$

$$F_n(t-) := \frac{1}{n} \sum_{i=1}^n Y_i \longrightarrow F(t-)$$

und

$$F_n(t) := \frac{1}{n} \sum_{i=1}^n Z_i \longrightarrow F(t)$$

fast sicher. Mit der Konvention, dass $F(-\infty) := 0$ und $F(\infty) := 1$ und $N \in \mathbb{N}$ fest wählen wir die Punkte

$$x_j := \inf\{x \in [-\infty, \infty] \mid F(x) \geq \frac{j}{N}\}, \quad j = 0, \dots, N$$

und

$$R_n := \max_{j=1, \dots, N-1} (|F_n(x_j) - F(x_j)| + |F_n(x_{j-}) - F(x_{j-})|).$$

Dann gilt $R_n \rightarrow 0$ fast sicher, und für $x \in (x_{j-1}, x_j)$

$$F_n(x) \leq F_n(x_{j-}) \leq F(x_{j-}) + R_n \leq F(x) + R_n + \frac{1}{N}.$$

Analog gilt

$$F_n(x) \geq F_n(x_{j-1}) \geq F(x_{j-1}) - R_n \geq F(x) - R_n - \frac{1}{N}.$$

Zusammen ergibt sich, dass für $x \in \mathbb{R}$ beliebig

$$\limsup_n |F_n(x) - F(x)| \leq \limsup_n R_n + \frac{1}{N} = \frac{1}{N} \quad \text{fast sicher.}$$

Mit $N \rightarrow \infty$ erhalten wir die Behauptung. □

5.2 Eine Quantitative Version von Glivenko-Cantelli

Das vorausgehende Resultat liefert keine quantitativen Informationen über den Approximationsfehler $D_n := \|F_n - F\|_\infty$. Der folgende Satz sagt, dass diese Approximation (im Sinne der stochastischen Konvergenz) exponentiell in der Anzahl der Versuche n fällt, und zwar unabhängig von der zugrundeliegenden Verteilungsfunktion F .

Satz 5.4. *Für die Folge der empirischen Verteilungsfunktion F_n einer i.i.d. Folge von reellen Zufallsvariablen (X_i) mit Verteilung F gilt für $\epsilon > 0$*

$$\mathbb{P} \left(\|F_n - F\|_\infty > \epsilon + 2 \cdot \sqrt{\frac{2 \ln(n+1)}{n}} \right) \leq e^{-2\epsilon^2 \cdot n}$$

Bemerkung. Insbesondere folgt $\mathbb{P}(\|F_n - F\|_\infty > 2\epsilon) \leq e^{-2\epsilon^2 \cdot n}$, falls $\sqrt{\frac{2 \ln(n+1)}{n}} < \epsilon$.

5.2.1 Konzentrationsungleichungen

Als Vorbereitung für den Beweis von Satz 5.4 besprechen wir jetzt drei Lemmata, die in die Familie von sogenannten Konzentrationsungleichungen gehören. Dabei geht es letztlich um möglichst genaue Abschätzungen für die Wahrscheinlichkeit, dass eine gegebene Zufallsvariable von ihrem Erwartungswert abweicht. Das simple aber fundamentale Resultat für diesen Abschnitt ist das folgende.

Lemma 5.5 (Hoeffding). *Falls X eine reelle Zufallsvariable mit $\mathbb{E}(X) = 0$ und $X \in [a, b]$ fast sicher ist, gilt*

$$\mathbb{E}(e^{sX}) \leq e^{\frac{s^2 \cdot (b-a)^2}{8}} \quad \forall s > 0.$$

Beweis. $x \mapsto e^{s \cdot x}$ ist konvex, also

$$e^{sX} \leq \underbrace{\frac{X-a}{b-a}}_{=: \alpha} e^{s \cdot b} + \underbrace{\frac{b-X}{b-a}}_{=: 1-\alpha} e^{s \cdot a}$$

mit $X = \alpha \cdot b + (1 - \alpha) \cdot a$. Durch Übergang zum Erwartungswert erhalten wir

$$\mathbb{E}(e^{sX}) \leq \frac{\mathbb{E} - a}{b-a} e^{sb} + \frac{b - \mathbb{E}(X)}{b-a} e^{sa} = \frac{-a}{b-a} \cdot e^{sb} + \frac{b}{b-a} e^{sa} \stackrel{!}{\leq} e^{\frac{s^2(b-a)^2}{8}},$$

wobei man für den letzten Schritt benutzt, dass

$$\frac{-a}{b-a} \cdot e^{sb} + \frac{b}{b-a} e^{sa} = e^{L(h)}$$

mit $h = s \cdot (b - a)$, $\eta = \frac{-a}{b-a}$ und $L(h) = -h\eta + \ln(1 - \eta + \eta e^h)$. Durch Nachrechnen verifiziert man $L(0) = 0$, $L'(0) = 0$, $L''(h) \leq \frac{1}{4}$. Durch Entwicklung von $L(\cdot)$ um den Punkt $0 \in \mathbb{R}$ sieht man hieraus

$$L(h) \leq \frac{1}{8} h^2. \quad \square$$

Bemerkung. Die Abschätzung des Hoeffding-Lemmas ist wertvoll für s nahe bei 0.

Lemma 5.6 (McDiarmids Lemma). *Es seien $X_1, \dots, X_n \in X$ unabhängige Zufallsvariablen, $f : X^n \rightarrow \mathbb{R}$ eine Abbildung, sodass*

$$|\phi(X_1, \dots, X_i, \dots, X_n) - \phi(X_1, \dots, X'_i, \dots, X_n)| \leq c_i$$

Dann gilt

$$\mathbb{P}(\{\phi(X_1, \dots, X_n) - \mathbb{E}(\phi(X_1, \dots, X_n)) \geq t\}) \leq \exp\left(-\frac{2t^2}{\sum c_i^2}\right).$$

Beweis. Sei $V_i := \mathbb{E}(\phi(X_1, \dots, X_n) | X_1, \dots, X_i) - \mathbb{E}(\phi(X_1, \dots, X_n) | X_1, \dots, X_{i-1})$. Dann gelten

$$\sup_{X_1, \dots, X_i} V_i(X_1, \dots, X_i) - \inf_{X_1, \dots, X_i} V_i(X_1, \dots, X_i) \leq c_i \quad . \quad (1)$$

$$\mathbb{E}(V_i | X_1, \dots, X_{i-1}) = 0 \quad (2)$$

und

$$\sum_{i=1}^n V_i = \phi(X_1, \dots, X_n) - \mathbb{E}(\phi(X_1, \dots, X_n)) \quad (3)$$

Damit gilt

$$\begin{aligned} \mathbb{P}(\phi(X_1, \dots, X_n) - \mathbb{E}(\phi) \geq t) &\stackrel{s>0 \text{ bel.}}{\leq} \mathbb{P}(s \cdot (\phi(X_1, \dots, X_n) - \mathbb{E}(\phi)) \geq s \cdot t) \\ &\leq \frac{1}{e^{st}} \mathbb{E}(e^{s \cdot [\phi(X_1, \dots, X_n) - \mathbb{E}(\phi)]}) \\ &\stackrel{(3)}{=} \frac{1}{e^{st}} \mathbb{E}\left(\prod_{i=1}^n e^{s \cdot V_i}\right) \\ &= \frac{1}{e^{st}} \mathbb{E}\left(\prod_{i=1}^{n-1} e^{s \cdot V_i} \cdot e^{s \cdot V_n}\right) \\ &= \frac{1}{e^{st}} \mathbb{E}\left(\prod_{i=1}^{n-1} e^{s \cdot V_i} \cdot \mathbb{E}(e^{s \cdot V_n} | X_1, \dots, X_{n-1})\right) \quad (*) \end{aligned}$$

Wegen (1) und (2) können wir das Hoeffding-Lemma anwenden und erhalten

$$\mathbb{E}(e^{s \cdot V_n} | X_1, \dots, X_{n-1}) \leq e^{\frac{s^2 \cdot c_n^2}{8}}.$$

Somit gilt

$$\begin{aligned} \mathbb{P}(\phi(X_1, \dots, X_n) - \mathbb{E}(\phi) \geq t) &\stackrel{(*)}{\leq} \frac{1}{e^{st}} \mathbb{E}\left(\prod_{i=1}^{n-1} e^{s \cdot V_i} \cdot \mathbb{E}(e^{s \cdot V_n} | X_1, \dots, X_{n-1})\right) \\ &\leq \frac{1}{e^{st}} \mathbb{E}\left(\prod_{i=1}^n e^{s \cdot V_i}\right) \cdot e^{\frac{s^2 \cdot c_n^2}{8}} \\ &\stackrel{\text{Induktion}}{\leq} \frac{1}{e^{st}} \cdot \exp\left(s^2 \frac{\sum_{i=1}^n c_i^2}{8}\right) \quad \forall s > 0. \end{aligned}$$

Also

$$\mathbb{P}(\phi(X_1, \dots, X_n) - \mathbb{E}(\phi) \geq t) \leq \inf_{s>0} \frac{1}{e^{st}} \cdot \exp\left(\frac{\sum_{i=1}^n c_i^2}{8s^2}\right) = \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \quad \square$$

Lemma 5.7 (Endliche Klassen-Lemma von Massart). *Sei $A \subseteq \mathbb{R}^n$ endlich, $A \subseteq B_R(0)$. Dann gilt*

$$\mathbb{E}\left(\max_{a \in A} \frac{1}{n} \sum_{i=1}^n \sigma_i a_i\right) \leq R \cdot \frac{\sqrt{2 \ln(\#A)}}{n}$$

wenn $(\sigma_i)_{i=1, \dots, n}$ unabhängig symmetrisch Bernoulli- $\{-1, +1\}$ -verteilt sind.

Beweis. Sei $A \subseteq \mathbb{R}^n$ endlich, $A \subseteq B_R(0)$, $\vec{\sigma} = (\sigma_1, \dots, \sigma_n)$ mit $\sigma_i = \pm 1$ mit Wahrscheinlichkeit $\frac{1}{2}$. Weiter sei

$$Z_{\vec{a}} = \sum_{i=1}^n \sigma_i a_i = \langle \vec{\sigma}, \vec{a} \rangle, \quad \vec{a} \in A$$

Für $s > 0$ haben wir dann

$$\begin{aligned} \mathbb{E}_{\vec{\sigma}}(\exp(s \cdot \max_{a \in A} Z_{\vec{a}})) &= \mathbb{E}_{\vec{\sigma}}(\max_{a \in A} \exp(s \cdot Z_{\vec{a}})) \\ &\leq \mathbb{E}_{\vec{\sigma}}\left(\sum_{a \in A} \exp(s \cdot Z_{\vec{a}})\right) \\ &= \sum_{a \in A} \mathbb{E}_{\vec{\sigma}}(\exp(s \cdot Z_{\vec{a}})) \\ &\stackrel{(*)}{\leq} \sum_{a \in A} \exp\left(\frac{s^2 R^2}{2}\right) \\ &= (\#A) \cdot \exp\left(\frac{s^2 R^2}{2}\right). \end{aligned}$$

In (*) haben wir das Hoeffding-Lemma angewandt, da

$$\mathbb{E}_{\vec{\sigma}}(Z_{\vec{a}}) = 0$$

und

$$|Z_{\vec{a}}(\sigma_1, \dots, \sigma_i, \dots, \sigma_n) - Z_{\vec{a}}(\sigma_1, \dots, \hat{\sigma}_i, \dots, \sigma_n)| \leq 2R$$

gilt. Die Jensen'sche Ungleichung liefert nun (aufgrund der Konvexität von $x \mapsto e^x$)

$$\exp\left(\mathbb{E}_{\vec{\sigma}}(s \cdot \max_{a \in A} Z_{\vec{a}})\right) \leq \mathbb{E}_{\vec{\sigma}}(\exp(s \cdot \max_{a \in A} Z_{\vec{a}})).$$

Daraus folgt

$$\mathbb{E}_{\vec{\sigma}}\left(\max_{a \in A} Z_{\vec{a}}\right) \leq \frac{\ln(\#A)}{s} + \frac{s \cdot R^2}{2} \quad \forall s > 0,$$

und daher

$$\mathbb{E}_{\vec{\sigma}}\left(\max_{a \in A} Z_{\vec{a}}\right) \leq \inf_{s>0} \left(\frac{\ln(\#A)}{s} + \frac{s R^2}{2}\right) = R \cdot \sqrt{2 \ln(\#A)}. \quad \square$$

5.2.2 Beweis von Satz 5.4

Sei $\mathfrak{G} := \{g_t : \mathbb{R} \rightarrow \mathbb{R} \mid g_t(x) = \mathbb{1}_{(-\infty, t]}(x), t \in \mathbb{R}\}$

$$\Rightarrow F(t) = \mathbb{E}(g_t(X_1)) = \mathbb{P}(X_1 \leq t)$$

sowie analog

$$F_n(t) = \frac{1}{n} \sum_{k=1}^n g_t(X_k) = E_n(g_t)$$

mit dem Erwartungswert E_n bezüglich dem empirischen Maß auf \mathbb{R}

$$\mu_n = \frac{1}{n} \sum_{k=1}^n \delta_{X_k},$$

d.h. für $f : \mathbb{R} \rightarrow \mathbb{R}$ gilt $E_n(f) = \frac{1}{n} \sum_{k=1}^n f(X_k) = \int_{\mathbb{R}} f(t) \mu_n(dt)$. Daher können wir schreiben

$$\sup |F_n(t) - F(t)| = \sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)|.$$

$E_n(g)$ hängt dabei von den Beobachtungen X_1, \dots, X_n ab, $\mathbb{E}(g)$ nicht. Setze nun

$$f(X_1, \dots, X_n) := \sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)|.$$

Dann gilt

$$|f(X_1, \dots, X_i, \dots, X_n) - f(X_1, \dots, X'_i, \dots, X_n)| \leq \frac{1}{n}$$

Nach dem McDiarmid-Lemma 5.6 ergibt sich mit Wahrscheinlichkeit größer als $1 - e^{-2\varepsilon^2 n}$

$$\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| \leq \mathbb{E} \left(\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| \right) + \varepsilon$$

Wir zeigen nun, dass der rechte Erwartungswert gegen Null konvergiert. Dazu seien X'_1, \dots, X'_n unabhängige Kopien von X_1, \dots, X_n . Weiter sei $\vec{\sigma} = (\sigma_1, \dots, \sigma_n)$, sodass die Zufallsvariablen $\{\sigma_i, X_i, X'_i, i = 1, \dots, n\}$ unabhängig sind und $\sigma_i = \pm 1$ jeweils mit Wahrscheinlichkeit $\frac{1}{2}$ sei. Dann gilt

$$\begin{aligned} \mathbb{E} \left(\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| \right) &= \mathbb{E}_{X_1, \dots, X_n} \left(\sup_{g \in \mathfrak{G}} \left| E_n(g) - \mathbb{E}_{X'_1, \dots, X'_n} \left(\frac{1}{n} \sum_{k=1}^n g(X'_k) \right) \right| \right) \\ &\leq \mathbb{E}_{X_1, \dots, X_n} \left(\sup_{X'_1, \dots, X'_n} \left| \frac{1}{n} \sum_{k=1}^n g(X_k) - g(X'_k) \right| \right) \\ &\stackrel{(*)}{=} \mathbb{E}_{\vec{X}, \vec{X}', \vec{\sigma}} \left(\sup_{g \in \mathfrak{G}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i (g(X_i) - g(X'_i)) \right| \right) \\ &\leq 2 \cdot \mathbb{E}_{\vec{X}, \vec{\sigma}} \left(\sup_{g \in \mathfrak{G}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right| \right) \rightarrow 0. \end{aligned}$$

Zu (*): Die Multiplikation der Differenzen $g(X_i) - g(X'_i)$ mit σ_i ist gleichwertig zu einer zufälligen Vertauschung von X_i und X'_i , was die Verteilung der Größe innerhalb des Erwartungswertes nicht ändert. Ferner können wir beim letzten Ausdruck dieser Ungleichungskette ohne Verlust zur sogenannten Ordnungsstatistik übergehen, d.h.

$$\mathbb{E}_{\vec{X}, \vec{\sigma}} \left(\sup_{g \in \mathfrak{G}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right| \right) = \mathbb{E}_{\vec{X}, \vec{\sigma}} \left(\sup_{g \in \mathfrak{G}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_{(i)}) \right| \right),$$

wobei $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ die Umordnung nach Größe der Realisierungen X_1, \dots, X_n bezeichnet. Auf den letzten Ausdruck können wir nun Lemma 5.7 anwenden, da

$$(g(X_{(i)}))_{i=1}^n \in \left\{ \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\} =: A.$$

Es gilt dann $\#A = n + 1$, $A \subseteq B_{\sqrt{n}}(0)$. Damit gilt

$$\begin{aligned} 2 \cdot \mathbb{E}_{\vec{X}, \vec{\sigma}} \left(\sup_{g \in \mathfrak{G}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right| \right) &= 2 \cdot \mathbb{E}_{\vec{X}} \left[\mathbb{E}_{\vec{\sigma}} \left(\sup_{g \in \mathfrak{G}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right| \right) \right] \\ &\stackrel{5.7}{\leq} 2 \cdot \mathbb{E}_{\vec{X}} \left(\sqrt{\frac{2 \ln(n+1)}{n}} \right) \rightarrow 0. \end{aligned}$$

Insgesamt erhalten wir

$$\begin{aligned} &\mathbb{P} \left(\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| > \varepsilon + 2 \cdot \sqrt{\frac{2 \ln(n+1)}{n}} \right) \\ &\leq \mathbb{P} \left(\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| > \varepsilon + \mathbb{E} \left[\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| \right] \right) \\ &= \mathbb{P} \left(\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| - \mathbb{E} \left[\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| \right] > \varepsilon \right) \leq e^{-2\varepsilon^2 \cdot n}. \quad \square \end{aligned}$$

Bemerkung. Aus Satz 5.4 lässt sich die Aussage von Glivenko-Cantelli ebenfalls wie folgt ableiten. Sei $\varepsilon > 0$, dann gilt

$$\begin{aligned} \mathbb{P} \left(\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| \geq 2\varepsilon \right) &\leq \mathbb{P} \left(\sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| > \varepsilon + 2\sqrt{\frac{2 \ln(n+1)}{n}} \right) \\ &\leq e^{-2\varepsilon^2 n}, \end{aligned}$$

falls n so groß ist, dass

$$\sqrt{\frac{2 \ln(n+1)}{n}} < \varepsilon.$$

Sei $A_n := \left\{ \sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| > 2\varepsilon \right\}$. Dann gilt nach dem obigen, dass

$$\mathbb{P}(A_n) \leq e^{-2\varepsilon^2 n} \quad \forall n \geq N_0,$$

$$\Rightarrow \sum_{n=1}^{\infty} \mathbb{P}(A_n) \leq \sum_{n=1}^{\infty} (e^{-2\varepsilon^2})^n < \infty.$$

Anwendung des 1. Lemmas von Borel-Cantelli (s. Wahrscheinlichkeitstheorie 1) liefert

$$\mathbb{P}(\limsup_{n \rightarrow \infty} A_n) = 0,$$

wobei

$$\limsup_{n \rightarrow \infty} A_n = \bigcap_{n \geq 0} \bigcup_{m \geq n} A_m$$

D.h. für beliebiges $\epsilon > 0$ gilt

$$\limsup_n \sup_{g \in \mathfrak{G}} |E_n(g) - \mathbb{E}(g)| \leq 2\epsilon \quad \text{fast sicher,}$$

also die gewünschte Aussage, dass $\|F_n - F\|_{\infty} \rightarrow 0$ fast sicher. \square

Bemerkung. Zu Illustration erinnern wir noch einmal Unterschied zwischen Konvergenz in Wahrscheinlichkeit und fast sicherer Konvergenz mit dem folgenden klassischen Beispiel.

Sei $\Omega = [0, 1]$, $\mathbb{P} = dx$, (a_n) eine Abzählung der rationalen Zahlen in $[0, 1]$. Weiter sei

$$X_n := \mathbb{1}_{[a_n - \frac{1}{n}, a_n + \frac{1}{n}]}$$

Dann konvergiert die Folge (X_n) in Wahrscheinlichkeit gegen Null, aber nicht dx -fast sicher, denn

$$\mathbb{P}(X_n > \varepsilon) \leq \frac{2}{n} \rightarrow 0,$$

aber für jedes $x \in [0, 1]$ ist

$$\limsup_{n \rightarrow \infty} X_n(x) = 1.$$

5.3 Der Kolmogorov-Smirnov-Anpassungstest

Beispiel. Beobachtet werden X_1, \dots, X_n Realisierungen desselben Zufallsexperimentes. Wir vermuten eine bestimmte Verteilungsfunktion $F: \mathbb{R} \rightarrow [0, 1]$ als Zufallsgesetz hinter den Ziehungen, d.h. den Zusammenhang $F(t) \stackrel{?}{=} \mathbb{P}(X_1 \leq t)$. Wie kann man eine solche Vermutung als Nullhypothese im Sinne eines statistischen Tests überprüfen?

Definition 5.8. Sei X eine reelle Zufallsvariable mit Verteilungsfunktion $F(t) = \mathbb{P}(X \leq t)$. Dann heißt die Funktion

$$F^{-1}: [0, 1] \rightarrow \mathbb{R}$$

$$F^{-1}(s) = \inf \{t \in \mathbb{R} \mid F(t) > s\}$$

verallgemeinerte (rechtsstetige) Inverse von F .

Das folgende Lemma ist eine beliebte Übungsaufgabe aus der Wahrscheinlichkeitstheorie. (Eigenschaft 2 wird genutzt, um mit dem Computer Zufallsvariablen zu vorgegebener Verteilungsfunktion zu generieren.)

Lemma 5.9.

1. F^{-1} ist die (klassische) Inverse von F , falls F strikt monoton ist.
2. Falls F stetig und $Z \simeq U([0, 1])$, so ist

$$X := F^{-1}(Z)$$

auf \mathbb{R} gemäß F verteilt.

3. F^{-1} ist rechtsstetig.

Beweis. Übung. □

Die folgende Beobachtung erklärt, warum die Abschätzung bei der quantitativen Variante vom Glivenko-Cantelli nicht von der unterliegenden Verteilungsfunktion F abhängt, jedenfalls sofern F stetig ist. Zur Erleichterung der Schreibweise bezeichnen wir diese Größe im folgenden mit $D_n := \|F_n - F\|_\infty$.

Lemma 5.10 (Pivot-Eigenschaft). *Seien X_1, \dots, X_n i.i.d. reellwertige Zufallsvariablen mit $X_i \simeq F$, F stetig, dann hängt die Verteilung von*

$$D_n(\vec{X}) := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

nicht mehr von F ab.

Beweis. O.B.d.A. gelte $X_i = F^{-1}(U_i)$ mit $U_i \simeq U([0, 1])$. Dann gilt:

$$\begin{aligned} F_n(x) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{F^{-1}(U_i) \leq x\}} \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i \leq F(X_i)\}} \\ &= \widehat{F}_n(F(x)) \end{aligned}$$

mit der empirischen Verteilungsfunktion $\widehat{F}_n(t)$ von (U_1, \dots, U_n) auf $[0, 1]$. Weiter gilt

$$\begin{aligned} D_n &= \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \\ &= \sup_{x \in \mathbb{R}} |\widehat{F}_n(F(x)) - F(x)| \\ &= \sup_{t \in [0, 1]} |\widehat{F}_n(t) - t| \end{aligned}$$

Da in die Konstruktion von \widehat{F}_n als empirische Verteilungsfunktion von $U([0, 1])$ -verteilten Zufallsvariablen die Funktion F nicht eingeht, folgt die Behauptung. □

Lemma 5.11. *Unter den Voraussetzungen wie in Lemma 5.10 sei $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ die realisierten Werte der Stichprobe (X_1, \dots, X_n) nach Umordnung gemäß Lage in \mathbb{R} . Dann gilt*

$$D_n = \max \left\{ \max \left(F(X_{(i)}) - \frac{i-1}{n}; \frac{i}{n} - F(X_{(i)}) \right) \mid 1 \leq i \leq n \right\}.$$

Beweis. Da F monoton und F_n stückweise konstant ist, können maximale Abweichungen von $F(x)$ und $F_n(x)$ nur an den Sprungstellen $x = X_i$ von F_n auftreten. Die rechts- und linksseitigen Limiten von F_n sind hier i/n oder $(i-1)/n$. \square

Korollar 5.12. *Wegen $F(X_i) \simeq U_i$ uniform auf $[0, 1]$ verteilt sind, gilt in Verteilung*

$$D_n = \max \left\{ \max \left(U_{(i)} - \frac{i-1}{n}; \frac{i}{n} - U_{(i)} \right) \mid 1 \leq i \leq n \right\}$$

wobei $U_{(1)} \leq \dots \leq U_{(n)}$ die geordneten Realisierungen von i.i.d. uniform auf $[0, 1]$ verteilten Zufallsvariablen U_1, \dots, U_n sind.

Definition 5.13. Die Verteilung der Größe D_n (als universelle Verteilung) wollen wir *diskrete Kolmogorov-Verteilung* mit n Freiheitsgraden nennen.

Als Korollar zu Pivot-Eigenschaft der Fehlerstatistik D_n erhalten wir sofort einen zulässigen Test für Hypothesen der Form

$$H_0 \hat{=} \text{“Die } (X_i)_i \text{ sind i.i.d. mit Verteilungsfunktion } F \text{ verteilt.”}$$

Satz 5.14 (Kolmogorov-Smirnov-Anpassungstest, exakte Version). *Im nichtparametrischen Modell $M = (\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n), (\nu_F^{\otimes n}))$, wobei F die Menge der stetigen Verteilungsfunktion auf \mathbb{R} durchläuft, sei $H_0 := \{F = F_0\}$ und $H_1 := \{F \neq F_0\}$ mit einer gewissen stetigen Verteilungsfunktion F_0 . Zu $\alpha \in (0, 1)$ sei $K_{n,1-\alpha}$ das α -Fraktile der diskreten Kolmogorov-Verteilung mit n Freiheitsgraden. Dann definiert*

$$\varphi(\vec{X}) := \mathbb{1}_{D_n(\vec{X}) \geq K_{n,1-\alpha}}$$

einen zulässigen Test von H_0 gegen H_1 zum Niveau α .

Beweis. Nach Wahl von $K_{n,1-\alpha}$ gilt

$$H_0 = \{F_0\} \Rightarrow \sup_{F \in H_0} \mathbb{E}_F(\varphi) = \mathbb{E}_{F_0}(\varphi) = \mathbb{E}(D_n \geq K_{n,1-\alpha}) = \mathbb{P}(D_n \geq K_{n,1-\alpha}) \leq \alpha. \quad \square$$

Beispiel. Behauptet wird, dass die Schuhgrößen im Hörsaal normalverteilt sind mit Erwartungswert $m = 50$ und Varianz $v = 200$. Wir wollen diese Hypothese zum Niveau $\alpha = 5\% = 0.05$ testen, wobei

$$F_0(t) := F_{m,v}(t) := \int_{-\infty}^t \nu_{m,v}(dt).$$

Die 22 Hörer der Vorlesung haben die Schuhgrößen $X_1 = 43, X_2 = 39, \dots, X_{22} = 40$. Nun bilden wir die Testgröße

$$D_n(\vec{X}) = \max \left\{ F_{m,v}(X_{(i)}) - \frac{i-1}{22}; \frac{i}{22} - F_{m,v}(X_{(i)}) \mid 1 \leq i \leq 22 \right\} \stackrel{?}{\geq} K_{22,0.95}$$

Gilt diese Ungleichung, so entscheiden wir uns für die Alternativhypothese.

Bemerkung. Satz 5.14 trifft keine Aussage über die Güte des Tests. Er besagt nur dass der D_n -Test nicht besonders viele *peinliche* Fehler produziert.

Satz 5.15 (Kolmogorov-Smirnov). *Sei (X_i) eine Folge von i.i.d. reellen Zufallsgrößen verteilt gemäß der stetigen Verteilungsfunktion $F : \mathbb{R} \rightarrow [0, 1]$. Dann gilt:*

$$\lim_{n \rightarrow \infty} \mathbb{P}(\sqrt{n} \cdot D_n \leq x) = k(x)$$

mit der Kolmogorov-Verteilungsfunktion $k : \mathbb{R} \rightarrow [0, 1]$:

$$k(x) = \begin{cases} 0 & \text{falls } x \leq 0 \\ 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 x^2}, & \text{falls } x > 0. \end{cases}$$

Beweis. Siehe z.B Matthias Löwe, Skript, Univ. Münster. □

Definition 5.16 (Kolmogorov-Smirnov-Test, Asymptotische Variante). Sei $\alpha \in (0, 1)$, n 'hinreichend groß', $H_0 := \{F = F_0\}$, $H_1 := \{F \neq F_0\}$. Dann wird H_0 abgelehnt, falls

$$D_n(\vec{X}_n) \geq \frac{\xi_{1-\alpha}}{\sqrt{n}}$$

wobei $\xi_{1-\alpha}$ das α -Fraktil der Kolmogorov-Verteilung ist.

Bemerkung. Da der Test allein die Stetigkeit der vermuteten Verteilungsfunktion voraussetzt, kann er breit eingesetzt werden. Unbeachtet bleibt dabei meistens die Frage, wie stark sich dieser Test vom 0-Test unterscheidet.

5.4 χ^2 -Anpassungstest

Der χ^2 -Anpassungstest ist ein zweiter sehr häufig in der Praxis genutzter nichtparametrischer Test für $X \simeq \varrho$ (Nullhypothese H_0) vs. $X \simeq \eta \neq \varrho$ (Alternativhypothese H_1) bei einer gegebene Menge von Stichproben X_1, \dots, X_n auf einem diskreten Zustandsraum $E = \{1, \dots, s\}$.

Beispiel (Schuhgrößen, Forts.). $E = \{1, \dots, 100\}$. X_1, \dots, X_{123} Stichproben. Vermutung (H_0): 'Jede Schuhgröße ist gleich wahrscheinlich', d.h. $\varrho(i) = \frac{1}{100} \forall i = 1, \dots, 100$.

Satz 5.17. *Zu X_1, \dots, X_n i.i.d. Zufallsvariablen mit Verteilung $X_1 \simeq \varrho$ auf E sei*

$$\vec{h}_n := (h_n(1), \dots, h_n(s))$$

der Vektor der absoluten Häufigkeiten, d.h.

$$h_n(i) = \# \{X_j = i\}.$$

Dann ist h_n multinomial verteilt mit Parametern n und ϱ , d.h.

$$\mathbb{P}(\vec{h}_n = (k_1, \dots, k_s)) = \frac{n!}{(k_1)!(k_2)! \cdot \dots \cdot (k_s)!} (\varrho_1)^{k_1} (\varrho_2)^{k_2} \cdot \dots \cdot (\varrho_s)^{k_s} =: \mu_{n,\varrho}(\vec{k}).$$

Beweis. Folgt aus elementarer Kombinatorik, analog zur Binomialverteilung. \square

Korollar 5.18 (Multinomial-Anpassungstest). *Sei $M = (E^n, \mathcal{P}(E^n), (\varrho^{\otimes n}))$ wobei ϱ alle Wahrscheinlichkeitsmaße auf E durchlaufe. Weiter sei $H_0 := \{\varrho = \varrho_0\}$ und $H_1 := \{\varrho \neq \varrho_0\}$ für ein gewisses ϱ_0 . Zu $\alpha \in (0, 1)$ sei $A \subseteq \mathbb{R}^s$ mit*

$$\mu_{n, \varrho_0}(A) \geq 1 - \alpha$$

Dann ist

$$\varphi(\vec{X}) := \mathbb{1}_{\{\vec{h}_n(\vec{X}) \notin A\}}$$

ein Test von H_0 gegen H_1 zum Niveau α .

Beweis. Folgt unmittelbar aus dem vorausgehenden Korollar. \square

Satz 5.19 (Pearson). *In der Situation wie oben und falls $(X_i)_{i \in \mathbb{N}}$ i.i.d. mit $X_1 \simeq \varrho$ (Verteilung auf E), so gilt für alle $\alpha \in (0, 1)$*

$$\lim_{n \rightarrow \infty} \mathbb{P}(T_n(X_1, \dots, X_n) > \chi_{s-1, 1-\alpha}^2) = \alpha,$$

wobei

$$T_n(X_1, \dots, X_n) = \sum_{i=1}^s \frac{1}{n \cdot \varrho_i} (h_n(i) - n \cdot \varrho_i)^2$$

und $\chi_{s-1, 1-\alpha}^2$ das α -Fraktile der χ_{s-1}^2 -Verteilung bezeichnet.

Bemerkung. Der Satz sagt aus, dass die Statistiken (T_n) für $n \rightarrow \infty$ in Verteilung gegen eine χ_{s-1}^2 -verteilte Zufallsgröße konvergieren.

Beweis von Satz 5.19. Aus den Eigenschaften der Multinomialverteilung folgt

$$\mathbb{E}(h_n(i)) = n \cdot \varrho_i,$$

$$\text{Kov}(h_n(i), h_n(j)) = \begin{cases} -n \cdot \varrho_i \cdot \varrho_j & \text{falls } i \neq j \\ n \cdot \varrho_i \cdot (1 - \varrho_i) & \text{falls } i = j. \end{cases}$$

Weiter gilt

$$h_n(i) = \sum_{k=1}^n \mathbb{1}_{\{X_k=i\}}.$$

Mit dem zentralen Grenzwertsatz in der vektorwertigen Version folgt dann, dass

$$\widetilde{h}_n := (\widetilde{h}_n(i))_{i=1, \dots, s-1}$$

mit

$$\widetilde{h}_n(i) = \frac{h_n(i)}{\sqrt{n}} - \sqrt{n} \cdot \varrho_i$$

für $n \rightarrow \infty$ in Verteilung gegen die Normalverteilung auf \mathbb{R}^{s-1}

$$\widetilde{h} \simeq \nu_{0, K}$$

konvergiert mit Kovarianzmatrix

$$K = (K_{ij}) \in \mathbb{R}^{s-1 \times s-1},$$

$$K_{ij} = \text{Cov}(\tilde{h}(i), \tilde{h}(j)) = \begin{cases} -n \cdot \varrho_i \cdot \varrho_j & \text{falls } i \neq j \\ n \cdot \varrho_i \cdot (1 - \varrho_i) & \text{falls } i = j. \end{cases}$$

Die Inverse $A := K^{-1}$ ist gegeben durch

$$A_{ij} = \begin{cases} \frac{1}{\varrho_s} & i = j \\ \frac{1}{\varrho_i} + \frac{1}{\varrho_j} & i \neq j. \end{cases}$$

Sei $A^{1/2} = \sqrt{A}$, so folgt mit der Charakterisierung der multivariaten Normalverteilungen Satz 3.11, dass

$$\Rightarrow A^{1/2} \cdot \tilde{h}_n \xrightarrow{\text{in Verteilung}} \nu_{0, Id}.$$

Damit gilt auch

$$\langle A^{1/2} \tilde{h}_n, A^{1/2} \tilde{h}_n \rangle_{\mathbb{R}^{s-1}} \xrightarrow{\text{in Verteilung}} \chi_{s-1}^2$$

gemäß Definition der χ_{s-1}^2 -Verteilung. Andererseits ist

$$\begin{aligned} \langle A^{1/2} \tilde{h}_n, A^{1/2} \tilde{h}_n \rangle &= \langle A \tilde{h}_n, \tilde{h}_n \rangle \\ &= n \cdot \sum_{k=1}^{s-1} \frac{1}{\varrho_k} \left(\frac{h_n(k)}{n} - \varrho_k \right)^2 + \frac{n}{\varrho_s} \left(\frac{h_n(s)}{n} - \varrho_s \right)^2 \\ &= T_n(X_1, \dots, X_n). \end{aligned} \quad \square$$

Definition 5.20 (χ^2 -Anpassungstest). Zu $\alpha \in (0, 1)$ und n hinreichend groß wird $H_0 \hat{=} \{\varrho\}$ verworfen, falls

$$T_n(X_1, \dots, X_n) > \chi_{s-1, 1-\alpha}^2,$$

wobei $E = \{1, \dots, s\}$ und $\chi_{s-1, 1-\alpha}^2$ das α -Fraktile der χ_{s-1}^2 -Verteilung sei.

Bemerkung. Der χ^2 -Anpassungstest ist ein asymptotischer Test, basierend auf

$$T_n \xrightarrow{n \rightarrow \infty} \chi_{s-1}^2 \text{ in Verteilung,}$$

sofern $X_i \simeq \varrho$. Wenn nun eine beobachtete Realisierung von $T_n(X_1, \dots, X_n)$ einen Wert ergibt, der gemäß der χ_{s-1}^2 -Verteilung extrem unwahrscheinlich ist, so wird die Annahme $X_i \simeq \varrho$ verworfen.

Beispiel (Pannini-Sammelalbum). Wir sammeln Fußball-Sammelbilder der deutschen Nationalmannschaft.

$$E = \{1, \dots, 25\} \hat{=} \{\text{Neuer}, \dots, \text{Boateng}\}$$

der Fußball-Nationalmannschaft. Gekauft wurden $n = 2000$ verschlossene Tüten mit jeweils einem enthaltenen Sammelbild. Um zu testen, ob alle Bilder der Spieler in den Tüten gleich wahrscheinlich vorkommen (Herstellergarantie) bilden wir den Vektor $\overrightarrow{h_{2000}}$ der absoluten Häufigkeiten ($h_{2000}(1) =$ Anzahl der gekauften Tüten mit Neuer, $h_{2000}(2) =$ Anzahl der gekauften Tüten mit Lahm, \dots , $h_{2000}(25) =$ Anzahl der gekauften Tüten mit

Boateng). Unter der Annahme/Nullhypothese dass die Verteilung der Spielerkarten ρ in den die Gleichverteilung auf E ist, d.h. $\rho(k) = \frac{1}{25} \forall k \in E$, erhalten wir die Teststatistik

$$T_{2000} = \sum_{i=1}^{25} \frac{(h_{2000}(k) - \frac{2000}{25})^2}{\frac{2000}{25}}$$

Falls nun

$$T_{2000} > \chi_{24; 99\%}^2$$

so wird die Annahme/Nullhypothese verworfen.

Bemerkung. Es gibt diverse Varianten des χ^2 -Tests.

Beispiel (χ^2 -Test auf Unabhängigkeit). Nikotinsucht vs. Geschlecht:

	Mann	Frau	
Raucher	12	8	20
Nichtraucher	16	9	25
	28	17	45

Ein solche Kreuztabelle wird auch *Kontingenztafel* genannt.

Definition 5.21. Falls $E^1 = \{1, \dots, a\}$, $E^2 = \{1, \dots, b\}$ und $X_i = (X_i^1, X_i^2) \in E^1 \times E^2$,

$$h_n(i, j) = \sum_{k=1}^n \mathbb{1}_{\{X_k=(i,j)\}},$$

$$h_n^1(i) = \sum_{k=1}^n \mathbb{1}_{\{X_k^1=i\}},$$

$$h_n^2(j) = \sum_{k=1}^n \mathbb{1}_{\{X_k^2=j\}},$$

dann ist

$$D_n = \sum_{(i,j) \in E^1 \times E^2} \frac{\left(h_n(i, j) - \frac{h_n^1(i)h_n^2(j)}{n} \right)^2}{\frac{h_n^1(i)h_n^2(j)}{n}}.$$

Bemerkung. Angenommen h_n^1 und h_n^2 sind stochastisch unabhängig. Dann gilt

$$\mathbb{P}(X = (i, j)) = \rho^1(i)\rho^2(j)$$

und es gilt weiter

$$h_n(i, j) \cong n \cdot \mathbb{P}(X = (i, j)) = n \cdot \rho^1(i)\rho^2(j) = \frac{1}{n} h_n^1(i)h_n^2(j).$$

Die Größe D_n kann daher als ein Maß interpretiert werden, inwieweit die gemeinsamen Verteilung von (X^1, X^2) eine Produktstruktur hat.

Satz 5.22. Falls $X_i = (X_i^1, X_i^2) \in E^1 \times E^2$, (X_i) eine Folge von i.i.d. Ziehungen einer auf $E^1 \times E^2$ verteilten Zufallsgröße mit

$$\mathbb{P}(X = (i, j)) = \varrho^1(i)\varrho^2(j)$$

ist, so gilt

$$D_n \xrightarrow{\text{in Verteilung}} \chi_{(a-1) \cdot (b-1)}^2$$

wobei $a = \#E^1$ und $b = \#E^2$ die Kardinalitäten der Zustandsräume E^1 und E^2 bezeichnet.

Beweis. Siehe z.B. Georgii □

Definition 5.23 (χ^2 -Test auf Unabhängigkeit). In der Situation von Satz 5.22 für n hinreichend groß und $\alpha \in (0, 1)$ wird die Annahme "i $\in E^1$ und $j \in E^2$ sind unabhängig" verworfen, falls

$$D_n > \chi_{(a-1) \cdot (b-1), 1-\alpha}^2,$$

wobei $\chi_{(a-1) \cdot (b-1), 1-\alpha}^2$ das α -Fraktile der $\chi_{(a-1) \cdot (b-1)}^2$ -Verteilung auf $\mathbb{R}_{\geq 0}$ bezeichnet.

6 Lineare Modelle

Als letzten Themenbereich diskutieren wir nun noch die sogenannten lineare Modelle, die eine Verallgemeinerung der aus der Schule oder Physikpraktikum vielleicht bekannten linearen Regression darstellen. Bei letzterer hat man durch Messung n -Datenpunkte vom Typ $(Y_i, X_i) \in \mathbb{R}^2$ gewonnen und fragt nun nach einem systematischen quantitativen Zusammenhang zwischen den Beobachtungsgröße Y und X der Form

$$X_i = \alpha + \beta \cdot Y_i + \sqrt{v}\epsilon_i$$

mit gewissen Konstanten $\alpha, \beta \in \mathbb{R}$, $v > 0$ und einem zufälligen 'Messfehler' ϵ_i für die jeweilige Messung. Die Parameter α und β werden dann nach der Methode der kleinsten Fehlerquadrate bestimmt, d.h.

$$(\alpha, \beta) := \operatorname{argmin}_{\alpha, \beta} \left[\sum_{i=1}^n (\alpha + \beta \cdot Y_i - X_i)^2 \right].$$

Dieses Vorgehen soll nun etwas verallgemeinert und in die Sprache der Statistik überführt werden.

Definition 6.1. Seien $s, n \in \mathbb{N}$ mit $s \leq n$ und $A \in \mathbb{R}^{n \times s}$ gegeben. Weiter sei

$$\vec{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^t \in \mathbb{R}^n$$

ein Zufallsvektor, wobei (ϵ_i) stochastisch unabhängig sei mit

$$\mathbb{E}(\vec{\epsilon}) = \begin{pmatrix} \mathbb{E}(\epsilon_1) \\ \vdots \\ \mathbb{E}(\epsilon_n) \end{pmatrix} = 0 \text{ und } \operatorname{Kov}(\vec{\epsilon}) = (\operatorname{Kov}(\epsilon_i, \epsilon_j))_{1 \leq i, j \leq n} = I_n$$

Außerdem sei $\vec{\gamma} \in \mathbb{R}^s$ und $v \geq 0$. Dann heißt der Zufallsvektor $\vec{X} \in \mathbb{R}^n$

$$\vec{X} = A\vec{\gamma} + \sqrt{v} \cdot \vec{\varepsilon}$$

ein *lineares Modell*. – Das zugehörige statistische Modell $M = (\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n), (\mathbb{P}_\vartheta; \vartheta = (\gamma, v) \in \mathbb{R}^s \times \mathbb{R}_{\geq 0}))$ mit

$$\mathbb{P}_\vartheta := \text{Verteilung von } (A\vec{\gamma} + \sqrt{v} \cdot \vec{\varepsilon})$$

heißt *lineares statistisches Modell*.

Bemerkung. Häufig werden die folgenden Bezeichnungen verwendet.

- X – 'Beobachtungsvektor'
- γ – 'Verschiebungsvektor'
- A – 'Designmatrix'
- $\vec{\varepsilon}$ – 'Fehlervektor'
- v – 'Skalenparameter'

Bemerkung. Im folgenden werden wir stets davon ausgehen, dass $\text{rang}(A) = s$. Dies hat die Invertierbarkeit der Matrix A^*A zur Folge. Insbesondere ist die Abbildung $\vartheta \mapsto \mathbb{P}_\vartheta$ injektiv (Übung).

Beispiel ((Inhomogene) Lineare Regression). Ausgehend von n beobachteten Datenpunkten $(X_i, Y_i) \in \mathbb{R}^2$ geht man aus von einem affin linearen Zusammenhang der Form

$$x_i = \gamma_0 + y_i \cdot \gamma_1 + \sqrt{v}\varepsilon_i.$$

Also ist

$$\vec{X} = \begin{pmatrix} 1 & y_1 \\ \vdots & \vdots \\ 1 & y_n \end{pmatrix} \cdot \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} + \sqrt{v} \cdot \vec{\varepsilon}.$$

$s = 2$ mit $\text{rang}(A) = s = 2 \Leftrightarrow \exists i, j, i \neq j, y_i \neq y_j$.

Beispiel (Polynomiale Regression). Ausgehend von n beobachteten Datenpunkten (X_i, Y_i) geht man aus von einem quantitativen Zusammenhang der Form

$$X_i = \gamma_0 + \gamma_1 Y_i + \frac{1}{2} \gamma_2 Y_i^2 \cdots + \frac{1}{k!} \gamma_k Y_i^k + \sqrt{v}\varepsilon_i.$$

In diesem Fall ist $s = k$, $n = n$, $A = (a_{ij}) \in \mathbb{R}^{n \times k}$ mit Zeilenvektoren

$$a_i = (1 \quad Y_i \quad \frac{1}{2} Y_i^2 \quad \cdots \quad \frac{1}{k!} Y_i^k).$$

Beispiel (Einweg-Klassifizierung). Zur Behandlung einer Krankheit stehen k verschiedene Medikamente zur Auswahl. Eine klinische Studie sammelt die Genesungsdauer $x_{ij} \in \mathbb{R}$ für die jeweiligen Patienten, gruppiert nach dem erhaltenen Medikament.

Medikament	Genesungsdauern Patienten
1	x_{11}, \dots, x_{1n_1}
2	x_{21}, \dots, x_{2n_2}
\vdots	\vdots
k	x_{k1}, \dots, x_{kn_k}

In der sog. Einweg-Klassifizierung wählt man nun den Ansatz

$$X_{il} = \gamma_i + \sqrt{v}\varepsilon_{il}, \quad l = 1, \dots, n_i, \quad i = 1, \dots, k,$$

d.h. (mit der Noation $1_{d \times 1} = (1, \dots, 1)^T \in \mathbb{R}^d$)

$$\begin{aligned} \vec{X} &= (X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2}, \dots, X_{k1}, \dots, X_{kn_k})^T \\ &= \begin{pmatrix} 1_{n_1 \times 1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1_{n_k \times 1} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_k \end{pmatrix} + \sqrt{v} \vec{\varepsilon}. \end{aligned}$$

6.1 Der Satz von Gauß-Markov

Die Methode der kleinsten Fehlerquadrate geht bereits auf Gauß zurück. Es folgt die Verallgemeinerung auf beliebige lineare Modelle. Wir starten mit einem Hilfresultat aus der linearen Algebra.

Lemma 6.2. *Es seien $\text{rang}(A) = s$, $L = \{A\vec{\gamma} \mid \vec{\gamma} \in \mathbb{R}^s\}$. Dann hat die orthogonale Projektion*

$$\pi_L : \mathbb{R}^n \rightarrow L$$

folgende Charakterisierung:

$$z = \pi_L(x) \Leftrightarrow z \in L, \quad \|x - z\| = \min_{y \in L} \|y - x\| \Leftrightarrow z = A(A^*A)^{-1}A^*x.$$

Beweis. Setze $Q := A(A^*A)^{-1}A^*$. Dann gilt $x = Qx + (I - Q)x$, $Q^2 = Q$, $Q^T = Q$ und

$$\begin{aligned} \|y - x\|^2 &= \langle y - Qx - (I - Q)x, y - Qx - (I - Q)x \rangle \\ &= \|y - Qx\|^2 + \|(I - Q)x\|^2 - 2 \underbrace{\langle y - Qx, (I - Q)x \rangle}_{:=D}, \end{aligned}$$

wobei

$$\begin{aligned} D &= (y - Qx)^T \cdot (I - Q)x = (y^T - x^T Q^T)(I - Q)x \\ &= y^T x - x^T Q^T x - y^T Qx + x^T Q^T Qx \\ &= y^T x - y^T Qx = y^T x - (Qy)^T x = 0. \end{aligned}$$

Es ist also $\|y - x\|^2$ minimal, wenn $\|y - Qx\|^2$ minimal ist, d.h. wenn $y = Qx$. \square

Satz 6.3 (Gauß-Markov). *Im linearen Modell $(\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n), (\mathbb{P}_\vartheta, \vartheta = (\gamma, v) \in \mathbb{R}^s \times \mathbb{R}_{\geq 0}))$ gilt*

1. $\hat{\gamma} := (A^*A)^{-1}A^*\vec{X}$ ist ein erwartungstreuer Schätzer für γ .
2. Unter allen unter allen linearen erwartungstreuen Schätzern für γ ist $\hat{\gamma}$ ist Varianz-optimal (engl. 'best linear unbiased estimator' (BLUE)).
3. Die Statistik

$$V^* = \frac{\|X\|^2 - \|\pi_L X\|^2}{n-s} = \frac{\|X - \pi_L X\|^2}{n-s} = \frac{\|X - A\hat{\gamma}\|^2}{n-s}$$

ist ein erwartungstreuer Schätzer für v .

Beweis. 1) Für $\vartheta = (\gamma, v)$ gilt wegen $\mathbb{E}(\vec{\epsilon}) = \vec{0}$.

$$\begin{aligned} \mathbb{E}_\vartheta(\hat{\gamma}) &= \mathbb{E}_\vartheta((A^*A)^{-1}A^*\vec{X}) \\ &= \mathbb{E}((A^*A)^{-1}A^*(A\gamma + \sqrt{v}\vec{\epsilon})) \\ &= (A^*A)^{-1}A^*A\gamma + \sqrt{v}(A^*A)^{-1}A^*\mathbb{E}(\vec{\epsilon}) = \gamma. \end{aligned}$$

2) Sei $\lambda : \mathbb{R}^n \rightarrow \mathbb{R}^s$ ein erwartungstreuer linearer Schätzer für γ , dann ist also $\lambda(X) = CX$ mit einer Matrix $C \in \mathbb{R}^{s \times n}$, so dass

$$\mathbb{E}_\vartheta(\lambda(X)) = C \cdot A \cdot \gamma = \gamma \quad \forall \gamma \in \mathbb{R}^s,$$

letzteres wegen der Erwartungstreue von λ . Also ist C linksinvers zu A . Für die Varianz ergibt sich

$$\begin{aligned} \mathbb{V}_\vartheta(\lambda) &= \mathbb{E}_\vartheta(\|CX - \gamma\|^2) \\ &= \mathbb{E}(\|CA\gamma - C\sqrt{v}\vec{\epsilon} - \gamma\|^2) \\ &= v\mathbb{E}(\|C\epsilon\|^2) \\ &= v\mathbb{E}\left(\sum_{i,j,k} C_{ij}C_{ik}\epsilon_j\epsilon_k\right) \\ &= v\sum_{i,j} C_{ij}^2 = v\|C\|_{HS}^2, \end{aligned}$$

mit der (sog. 'Hilbert-Schmidt'-) Norm $\|C\|_{HS} := \sqrt{\text{spur}(C^*C)}$. Wir behaupten nun, dass die Matrix $(A^*A)^{-1}A^*$ die eindeutig bestimmte Linksinverse zu A mit minimaler $\|\cdot\|_{HS}$ -norm ist. In der Tat, aus

$$\mathbb{R}^s \xrightarrow{A} \mathbb{R}^n \xrightarrow{C} \mathbb{R}^s, \quad C \cdot A = \text{Id}_{\mathbb{R}^s}$$

und der Minimalität von C bzgl. $\|\cdot\|_{HS}$ folgt, dass

$$C = C \cdot \pi_L,$$

denn andernfalls könnte man C durch $\tilde{C} := C \cdot \pi_L$ ersetzen, um ein Linksinverses mit geringerer HS-norm $\|\cdot\|_{HS}$ zu erhalten. Mit Lemma 6.2 ergibt sich hieraus

$$C = C \cdot \pi_L = C \cdot A(A^*A)^{-1}A^* = (A^*A)^{-1}A^*,$$

wobei wir im letzten Schritt erneut benutzt haben, dass C linksinvers zu A ist.

3) Sei nun u_1, \dots, u_n eine Orthonormalbasis mit $\text{span}\{u_1, \dots, u_s\} = L$ und O die orthogonale Matrix mit Spaltenvektoren $(u_1 \cdots u_n)$, dann gilt $\pi_L = O \cdot E_s \cdot O^*$ mit E_s die orthogonale Projektion auf die ersten s Koordinatenrichtungen in \mathbb{R}^n . Gemäß Definition von V^* erhalten wir

$$(n-s)V^* = \|X - A(A^*A)^{-1}A^*X\|^2$$

und somit laut Definition von X unter P_ϑ

$$\begin{aligned} \mathbb{E}_\vartheta((n-s)V^*) &= \mathbb{E}(\|A\gamma + \sqrt{v}\vec{\epsilon} - A(A^*A)^{-1}A^*(A\gamma + \sqrt{v}\vec{\epsilon})\|^2) \\ &= v\mathbb{E}(\|\vec{\epsilon} - A(A^*A)^{-1}A^*\vec{\epsilon}\|^2) \\ &= v\mathbb{E}(\|\vec{\epsilon} - \pi_L\epsilon\|^2) \\ &= v\mathbb{E}(\|\vec{\epsilon} - OE_sO^*\epsilon\|^2) \\ &= v\mathbb{E}(\|O^*\vec{\epsilon} - E_sO^*\epsilon\|^2) \\ &= v\mathbb{E}\left(\sum_{k=s+1}^n (\eta_k)^2\right) \quad \text{mit } \vec{\eta} = O^*\vec{\epsilon} \\ &= v(n-s), \end{aligned}$$

da

$$\mathbb{E}(\eta_k^2) = \mathbb{E}\left(\sum_{i,j} O_{ik}O_{jk}\epsilon_i\epsilon_j\right) = \sum_l O_{lk}O_{lk} = 1.$$

V^* ist somit ein erwartungstreuer Schätzer für v . □

6.2 Konfidenzbereiche und Tests in linearen Gauß-Modellen

In dem Fall, dass der Fehlervektor $\vec{\epsilon}$ Gauß'sch modelliert wird, können wir in erheblichem Maße von unseren Kenntnissen aus Abschnitt 3 profitieren, um Konfidenzbereiche und Tests für die zugehörigen linearen Modelle zu konstruieren.

Definition 6.4. Falls $\vec{X} = A\vec{\gamma} + \sqrt{v}\vec{\epsilon}$ mit $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ mit unabhängig $\nu_{0,1}$ -verteilten ϵ_i , so heißt das Modell ein *Lineares Gauß-Modell*.

Satz 6.5. Im linearen Gauß-Modell gilt unter $\mathbb{P}_\vartheta = \mathbb{P}_{(\vec{\gamma}, v)}$:

1. $\hat{\gamma}$ ist (s -dimensional) $\nu_{\gamma, v, (A^*A)^{-1}}$ -verteilt.
2. $\frac{n-s}{v} \cdot v^*$ ist χ_{n-s}^2 -verteilt.
3. $\frac{\|A\hat{\gamma} - \gamma\|^2}{v} = \frac{\|\pi_L X - \mathbb{E}_\vartheta(X)\|^2}{v}$ ist χ_s^2 -verteilt und stochastisch unabhängig von v^* .

Beweis.

1. Linearkombinationen von normalverteilten Zufallsvariablen sind normalverteilt. Also ist $\hat{\gamma} = (A^*A)^{-1}A^*X$ ein s -dimensional normalverteilter Zufallsvektor mit

$$\mathbb{E}(\hat{\gamma}) = \gamma$$

Sei nun o.B.d.A. $\mathbb{E}(\hat{\gamma}) = 0$, dann finden wir

$$\begin{aligned} \text{Cov}(\hat{\gamma})_{ij} &= \mathbb{E}(\hat{\gamma}_i \hat{\gamma}_j) = \mathbb{E} \left(\sum_k c_{ik} X_k \sum_l c_{jl} X_l \right) \\ &= \sum_{k,l} c_{ik} c_{jl} \cdot v \cdot \delta_{lk} = v \cdot \sum_k c_{ik} c_{jk} = (v \cdot C \cdot C^*)_{ij} \end{aligned}$$

mit

$$C \cdot C^* = (A^* A)^{-1} A^* A [(A^* A)^{-1}]^* = (A^* A)^{-1}.$$

2. Sei $L := \{A\gamma \mid \gamma \in \mathbb{R}^s\} \subseteq \mathbb{R}^n$ und seien u_1, \dots, u_s eine Orthonormalbasis von \mathbb{R}^n , s.d.

$$\text{span}(u_1, \dots, u_s) = L$$

Sei weiter $\eta = \sum_{i=1}^n \xi_i u_i$ mit ξ_1, \dots, ξ_n i.i.d., $\xi_i \simeq \nu_{0,1}$, dann ist $\eta \in \mathbb{R}^n$ standardnormalverteilt, d.h. $\eta \simeq \nu_{0, E_n} \simeq \varepsilon$ und somit gilt für die Größe

$$\tilde{X} := A\gamma + \sqrt{v} \cdot \eta \simeq X.$$

Entsprechend gilt für die davon abgeleitete Statistik

$$v^* = \frac{1}{n-s} \|X - A\hat{\gamma}\|^2 = \frac{1}{n-s} \|X - \pi_L(X)\|^2 \simeq \frac{1}{n-s} \|\tilde{X} - \pi_L(\tilde{X})\|^2.$$

Wegen $\eta = \sum_{i=1}^s \eta_i + \sum_{i=s+1}^n \eta_i = \pi_L(\eta) + \eta^\perp$ gilt

$$\pi_L(\tilde{X}) = A\gamma + \sqrt{v} \pi_L(\eta),$$

so dass

$$\|\tilde{X} - \pi_L(\tilde{X})\|^2 = v \cdot \|\eta^\perp\|^2 = v \cdot \sum_{i=s+1}^n \|\eta_i\|^2 = v \cdot \sum_{i=s+1}^n \xi_i^2,$$

also nach Definition der χ^2 -Verteilung

$$\frac{n-s}{v} \cdot v^* \simeq \chi_{n-s}^2.$$

3. $\|A(\hat{\gamma} - \gamma)\|^2 = \|\pi_L(X) - A\gamma\|^2 \simeq \|\pi_L(\tilde{X}) - A\gamma\|^2 = v \cdot \|\pi_L(\eta)\|^2 = v \cdot \sum_{i=1}^s \xi_i^2$
 $\Rightarrow \frac{\|A(\hat{\gamma} - \gamma)\|^2}{v} \simeq \chi_s^2. \quad \square$

Korollar 6.6. *Im linearen Gauß-Modell ist*

$$\frac{\|A(\hat{\gamma} - \gamma)\|^2}{s \cdot v^*} \simeq f_{s, n-s},$$

d.h. Fisher-verteilt mit s und $n-s$ Freiheitsgraden. \square

Satz 6.7 (Konfidenzbereiche im linearen Gauß-Modell). *Zu $\alpha \in (0, 1)$ definieren im linearen Gauß-Modell*

1. *die Menge*

$$C(X) = \{\gamma \in \mathbb{R}^s \mid |A(\gamma - \hat{\gamma})|^2 < s \cdot f_{1-\alpha} v^*\}$$

einen Konfidenzbereich für γ zum Niveau α , wobei $f_{1-\alpha}$ das α -Fraktile der Fischer-Verteilung $\mathfrak{F}_{s, n-s}$ bezeichnet,

2. *und die Menge*

$$\tilde{C}(X) := \left(\frac{(n-s)v^*}{q_+}, \frac{(n-s)v^*}{q_-} \right)$$

ein Konfidenzbereich für v zum Niveau α , wobei $q_- = \chi_{n-s, \alpha/2}^2$, $q_+ = \chi_{n-s, 1-\alpha/2}^2$ das $\alpha/2$ -Quantil bzw. $\alpha/2$ -Fraktile der χ_{n-s}^2 -Verteilung bezeichnen.

Beweis.

1. Nach Korollar 6.6 ist für $\vartheta = (\gamma, v)$ beliebig $T(X) := \frac{\|A(\hat{\gamma} - \gamma)\|^2}{sv^*}$ $\mathfrak{F}_{s, n-s}$ -verteilt. Somit ist T ein Pivot und nach Wahl von $f_{1-\alpha}$ ist

$$\mathbb{P}_\vartheta(T > f_{1-\alpha}) \leq \alpha$$

bzw.

$$\mathbb{P}_\vartheta(T < f_{1-\alpha}) \geq 1 - \alpha$$

Hieraus folgt der Schluss mit dem Pivot-Prinzip.

2. Analog. □

Satz 6.8. *Im linearen Gauß-Modell $X = A\gamma + \sqrt{v}\varepsilon$*

1) *ist für $c \in \mathbb{R}^s$ die Größe*

$$\frac{\frac{\langle c, \hat{\gamma} - \gamma \rangle}{\sqrt{v \cdot \langle c, (A^*A)^{-1}c \rangle}}}{\sqrt{\frac{v^*}{v}}} \simeq \frac{\nu_{0,1}}{\sqrt{\frac{1}{n-s} \chi_{n-s}^2}}$$

Student-verteilt gemäß t_{n-1} mit $n-s$ Freiheitsgraden und

2) *für einen Unterraum $H \subseteq L$ mit $\dim H = r < s$, $A\gamma \in H$ ist die Größe*

$$\frac{n-s}{s-r} \frac{\|\pi_L(X) - \pi_H(X)\|^2}{\|X - \pi_L(X)\|^2} \simeq \frac{\frac{1}{s-r} \sum_{i=r+1}^s \xi_i^2}{\frac{1}{n-s} \sum_{i=s+1}^n \xi_i^2}$$

Fisher-verteilt gemäß $f_{s-r, n-s}$ mit $s-r$ und $n-r$ Freiheitsgraden.

Beweis. Analog zum Beweis von Satz 6.5. □

Wir können nun auch sofort eine Reihe von Tests in linearen Gaußmodellen zu gegebenem Niveau α angeben.

Satz 6.9 (Tests im linearen Gauß-Modell). *Im linearen Gauß-Modell $X = A\gamma + \sqrt{v}\varepsilon$ gilt:*

1. (*t*-Test für $\langle c, \gamma \rangle = m$) Für $c \in \mathbb{R}^s$ ist

$$\varphi(X) := \mathbb{1}_{\{|\langle c, \gamma \rangle - m| > t_{n-s; 1-\alpha/2} \sqrt{\langle c, (A^*A)^{-1}c \rangle v^*}\}}$$

ein Test von $H_0 = \{\langle c, \gamma \rangle = m\}$ gegen $H_1 = \{\langle c, \gamma \rangle \neq m\}$ zum Niveau α , wobei $t_{n-s; 1-\alpha/2}$ das $\alpha/2$ -Fraktile der t_{n-s} -Verteilung bezeichnet.

2. (χ^2 -Test für die Varianz) Für $v_0 > 0$ ist

$$\varphi(X) := \mathbb{1}_{\{(n-s)v^* > v_0 \chi_{n-s; 1-\alpha}^2\}}$$

ein Test von $H_0 = \{v \leq v_0\}$ gegen $H_1 = \{v > v_0\}$ zum Niveau α , wobei $\chi_{n-s; 1-\alpha}^2$ das α -Fraktile von χ_{n-s}^2 bezeichnet.

3. (*F*-Test für $A\gamma \in H$) Für $H \subseteq L$ mit $\dim H =: r < s$ ist

$$\varphi(X) = \mathbb{1}_{\{F_{H;L}(X) > f_{s-r, n-s; 1-\alpha}\}}$$

ein Test von $H_0 = \{A\gamma \in H\}$ gegen $H_1 = \{A\gamma \notin H\}$ zum Niveau α , mit dem α -Fraktile $f_{s-r, n-s; 1-\alpha}$ der der Fischer-Verteilung $F_{s-r, n-s}$ und

$$F_{H,L}(X) = \frac{n-s}{s-r} \frac{\|\pi_L(X) - \pi_H(X)\|^2}{\|X - \pi_L(X)\|^2}.$$

Beweis. Folgt unmittelbar aus den Sätzen 6.5 und 6.8 für die hier verwendeten Pivot-Statistiken. \square

6.3 Anwendung: Einweg-Klassifizierung und ANOVA-Methode

Zum Abschluss eine Anwendung im Bereich der Medikamentenforschung. Die folgenden Daten von Patienten mit einer gewissen Erkrankung wurden gesammelt und nach dem verabreichten Medikament geordnet.

Medikamente	Beobachtungen	Mittelwerte
1	X_{11}, \dots, X_{1n_1}	$m_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}$
2	X_{21}, \dots, X_{2n_2}	$m_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$
\vdots	\vdots	\vdots
s	X_{s1}, \dots, X_{sn_s}	$m_s = \frac{1}{n_s} \sum_{i=1}^{n_s} X_{si}$

Wir gehen nun von der Modellannahme aus, dass die Merkmalsausprägungen je nach verabreichtem Medikament innerhalb der entsprechenden Gruppe um einen medikamentenspezifischen Mittelwert schwanken, d.h.

$$X_{ki} = \gamma_k + \sqrt{v}\varepsilon_{ki} \quad \forall k = 1, \dots, s \quad \forall i = 1, \dots, n_k,$$

bzw.

$$X = A\gamma + \sqrt{v} \cdot \varepsilon$$

mit

$$X = (X_{11}, \dots, X_{1n_1}, \dots, X_{s1}, \dots, X_{sn_s})^T, \quad \gamma = (\gamma_1, \dots, \gamma_s)^T,$$

$$\varepsilon = (\varepsilon_{11}, \dots, \varepsilon_{1n_1}, \dots, \varepsilon_{s1}, \dots, \varepsilon_{sn_s})^T$$

und

$$A = \begin{pmatrix} 1_{n_1 \times 1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1_{n_s \times 1} \end{pmatrix}.$$

Dann gilt

$$(A^*A) = \begin{pmatrix} n_1 & & 0 \\ & \ddots & \\ 0 & & n_s \end{pmatrix},$$

$$A^*X = (n_1 m_1, \dots, n_s m_s)^T,$$

$$\hat{\gamma}(X) = (A^*A)^{-1} A^*X = (m_1, \dots, m_s)^T,$$

sowie

$$v^* = \frac{1}{n-s} \|\pi_L(X) - X\|^2 = \frac{1}{n-s} \sum_{k,i} (X_{ki} - m_k)^2 =: v_{iG}^*.$$

Ferner machen wir von der elementaren Zerlegung Gebrauch

$$(n-1)v_{tot}^* = (n-s)v_{iG}^* + (s-1)v_{zG}^*,$$

wobei

$$n = n_1 + \dots + n_s$$

$$v_{tot}^* = \frac{1}{n-1} \sum_{k,i} (X_{ki} - \mu)^2, \quad \mu = \frac{1}{n} \sum_{k,i} X_{ki}$$

und

$$v_{zG}^* = \frac{1}{s-1} \sum_{k=1}^s n_k (m_k - \mu)^2.$$

v_{tot}^* ist die totale Streuung aller Messwerte, v_{iG}^* als die mittlere Streuung innerhalb der Patientengruppen und v_{zG}^* die mittlere Streuung der Gruppenmittelwerte.

Die Anwendung der Tests aus dem vorigen Abschnitt bietet nun die Möglichkeit zum Vergleich von Medikamenten wie folgt.

1. Im Fall $s = 2$ (Vergleich zweier Medikamente) wählt man als Nullhypothese

$$H_0 = \text{”Beide Medikamente sind gleich gut.”}$$

d.h.

$$\Theta_0 = \left\{ \gamma \mid \left\langle \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \gamma \right\rangle = 0 \right\}.$$

Hierauf können wir den t -Test für Lineare-Gaußmodelle anwenden, d.h. mit einem Niveau α lautet der Test hier

$$\varphi(X) = \mathbb{1}_{\left\{ |m_1 - m_2| > t_{n-2; 1-\alpha/2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) v_{iG}^*} \right\}}.$$

2. Im Fall $s \geq 3$ (Vergleich mehrerer Medikamente) wählt man die Nullhypothese

$$H_0 = \text{”Alle Medikamente sind gleich gut”}$$

in der Form des F -Tests behandeln mit

$$\Theta_0 = \left\{ \gamma \mid A\gamma \in \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mathbb{R} =: H \right\}.$$

Der F -Test lautet also für diesen Fall

$$\varphi(X) = \mathbb{1}_{\left\{ \frac{v_{zG}^*}{v_{iG}^*} > f_{s-1, n-s; 1-\alpha} \right\}}$$

In der Teststatistik tauchen lediglich die beobachteten Streugrößen v_{zG}^* und v_{iG}^* auf, was den Namen dieses Verfahrens als ANOVA-Methode (für ”Analysis of Variances”) begründet.