universität
wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

Exponentielle Familien, Suffizienz

und das Theorem von Darmois-Koopman-Pitman

verfasst von / submitted by
Harald Führer

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 190 406 884
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Lehramtsstudium UF Mathematik und UF Informatik und Informatikmanagement
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Ing. Dr. Otmar Scherzer

Danksagung

Dank gilt meinen Betreuer Univ.-Prof. Dipl.-Ing. Dr. Otmar Scherzer für die Idee zu dieser Arbeit und seine fachliche Unterstützung. Auch möchte ich mich für seine Ausdauer und Geduld bei der Betreuung dieser Arbeit bedanken.

für Romana

Zusammenfassung

Im Mittelpunkt dieser Arbeit steht die Analyse des klassischen Satzes von (Fisher-)Darmois-Koopman-Pitman. Drei Arbeiten von G. Darmois, B. O. Koopman und E. J. G. Pitman aus den Jahren 1935/36 haben diesen Satz unabhängig voneinander, aufbauend auf Arbeiten R. A. Fishers, publiziert.

Die vorliegende Arbeit analysiert diese drei klassischen Arbeiten. Dabei wird deren Inhalt dargestellt und mit der statistischen Theorie aus heutiger Sicht verglichen. Zudem erfolgt ein Abriss der Theorie aus heutiger Sicht. Neben grundlegenden Begriffen der Maß- und Wahrscheinlichkeitstheorie werden statistische Grundlagen präsentiert. Exponentielle Verteilungsfamilien sowie suffiziente Statistiken werden ausführlich dargestellt. Des Weiteren erfolgt ein Abriss der Punktschätzung hinsichtlich der Begriffe Erwartungstreue, Fisher-Information und Maximum-Likelihood. Den Hauptteil der Arbeit bildet die Analyse der Arbeiten von G. Darmois, B. O. Koopman und E. J. G. Pitman, sowie der Vorarbeit von R. A. Fisher.

Abstract

This diploma thesis deals with the classical theorem of (Fisher-)Darmois-Koopman-Pitman. Three papers by G. Darmois, B. O. Koopman and E. J. G. Pitman from 1935/36 have published this theorem independently of one another, building on the works of R. A. Fisher.

The present work analyzes these three classical works. Their content is presented and compared with contempoary statistical theory. An overview of the theory takes place from today's perspective. In addition to basic concepts of measurement and probability theory, statistical foundations are presented. Exponential families as well as sufficient statistics are presented in detail. In addition, a brief presentation of point estimation is given in terms of expectation, Fisher information, and maximum likelihood. The main part of the diploma thesis is the analysis of the work of G. Darmois, B. O. Koopman and E. J. G. Pitman, as well as the preliminary work of R. A. Fisher.

Inhaltsverzeichnis

0 Einleitung

Im Zentrum dieser Arbeit steht der klassische Satz von (Fisher-)Darmois-Koopman-Pitman. Dabei handelt es sich um einen Satz aus der mathematischen Statistik. Dieser beschreibt ein unmittelbares Zusammenspiel zwischen suffizienten Statistiken und exponentiellen Verteilungsfamilien.

Der Satz Satz von (Fisher-)Darmois-Koopman-Pitman sagt aus, dass für ein statistisches Modell mit einer (glatten nirgends verschwindenden) Wahrscheinlichkeitsdichte genau dann eine endlich-dimensionale suffiziente Statistik, die nicht von der Stichprobengröße abhängt, existiert, wenn die Wahrscheinlichkeitsdichte einer exponentiellen Familie entstammt.

Häufig wird in der statistischen Literatur auf die ersten Arbeiten, die diesen Satz beschreiben, verwiesen, aber der konkrete Inhalt dieser Arbeiten wird nicht genannt. Drei Arbeiten, die die Aussage des Satzes unabhängig voneinander zum ersten Mal explizit publizierten, stehen dabei im Mittelpunkt. Es sind dies Arbeiten von G. Darmois, B. O. Koopman und E. J. G. Pitman aus den Jahren 1935/36. Alle drei Arbeiten bauen auf der Arbeit R. A. Fishers auf. Das Verdienst dieser drei Arbeiten ist, die Aussage des (später dann (Fisher-)Darmois-Koopman-Pitman Theorem genannten) Satzes, die in einer Arbeit R. A. Fisher von 1934 schon implizit vorhanden war, explizit zu präsentieren und auch zu beweisen.

Die Motivation zu dieser Arbeit war, diese drei klassischen Arbeiten zu analysieren. Dabei wird deren Inhalt dargestellt und mit der statistischen Theorie aus heutiger Sicht verglichen. Zudem erfolgt ein Abriss der Theorie aus heutiger Sicht. Im Laufe dieser Arbeit entwickelte sich auch eine umfangreiche Literatursammlung. Moderne Recherche-Instrumente, wie der digitale Zugriff auf viele klassische Publikationen über die Recherche-Plattform der Universität Wien, haben dabei die Arbeit sehr erleichtert.

Der Aufbau dieser Arbeit gestaltet sich daher entsprechend. Ziel und Hintergrund des Aufbaus ist es, Material und Werkzeuge zum Verständnis des Satzes von (Fisher-)Darmois-Koopman-Pitman zu präsentieren.
Das erste Kapitel enthält grundlegende Begriffe der Maß- und Wahrscheinlichkeitstheorie.
Im zweiten Kapitel werden statistische Grundlagen präsentiert. Gefolgt von der Definition exponentieller Verteilungsfamilien, sowie von einer detailierten Darstellung suffizienter Statistiken.

Im dritten Kapitel erfolgt ein Abriss der Punktschätzung hinsichtlich der Begriffe Erwartungstreue, Fisher-Information und Maximum-Likelihood.
Das vierte Kapitel widmet sich dann dem Hauptthema der Arbeit, nämlich den Arbeiten von G. Darmois, B. O. Koopman und E. J. G. Pitman, sowie der Vorarbeit durch R. A. Fisher.
Und das letzte Kapitel liefert noch einen kurzen Abriss über Literatur, die den Satz von (Fisher-)Darmois-Koopman-Pitman rezipiert, sowie aktuelle Anwendungen exponentieller Familien, wo der Satz von (Fisher-)Darmois-Koopman-Pitman noch gewissen Einfluss hat.

Erwähnen möchte ich im Vorfeld, dass viele Bereiche der mathematischen Statistik keine Erwähnung in dieser Arbeit finden. So wird z.B. weder auf statistische Testtheorie, Intervallschätzung oder Bayes'sche Statistik eingegangen. In all diesen klassischen Bereichen spielen suffiziente Statistiken und exponentielle Verteilungsfamilien eine große Rolle. Auch auf das asymptotische Verhalten von Schätzern wird nicht eingegangen. Denn das Ziel der Arbeit ist die Aufarbeitung der drei klassischen Arbeiten von G. Darmois, B. O. Koopman und E. J. G. Pitman bzw. eine Darstellung des Satzes von (Fisher-)Darmois-Koopman-Pitman. Eine vollständige Darstellung aller Anwendungen suffizienter Statistiken und exponentieller Verteilungsfamilien hätte ganz einfach den Umfang der Arbeit gesprengt. Eine gute elementare Darstellung der klassischen mathematischen Statistik aus dem Blickwinkel exponentieller Familien bietet das Buch [Oli14] von David J. Olive.

1 Grundlagen

An dieser Stelle sollen Begriffe, die für das Verständnis der weiteren Kapitel nötig bzw. hilfreich sind, in übersichtlicher Form präsentiert werden. Die Notation der maß- bzw. wahrscheinlichkeitstheoretischen Grundlagen folgt im Wesentlichen dem Standardwerk von Lehmann [LR05]. Deutsche Begriffe und Bezeichnungen orientieren sich an [Sch15] und an [Fel14], wo ebenfalls die Notation von [LR05] verwendet wird.

1.1 Maßtheoretische Grundlagen

Definition 1.1 (σ-Algebra)**.** *Eine σ-Algebra $\mathcal{A}$ ist eine Familie von Teilmengen einer nichtleeren Menge Ω, d.h. $\mathcal{A}$ ist eine Teilmenge der Potenzmenge $\mathcal{A} \subset \mathscr{P}(\Omega)$, die folgende drei Bedingungen erfüllt:*

(1) $\Omega \in \mathcal{A}$

(2) $A \in \mathcal{A} \quad \Rightarrow \quad A^c := \Omega \setminus A \in \mathcal{A}$

(3) $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A} \quad \Rightarrow \quad \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$

Eine Menge $A \in \mathcal{A}$ heißt messbar. (siehe [Sch15] S.4)

Beispiel 1.1 (Borel σ-Algebra)**.** *Die von den offenen Mengen $\mathcal{O} \subseteq \mathbb{R}^k$ erzeugte σ-Algebra $\sigma(\mathcal{O})$ heißt Borel(sche) σ-Algebra $\mathscr{B}(\mathbb{R}^k)$. Eine Menge $B \in \mathscr{B}(\mathbb{R}^k)$ heißt Borelmenge bzw. Borel-messbar. (siehe [Sch15, S.6])*

Bemerkung: Die obige Definition einer Borel-σ-Algebra in Beispiel 1.1 lässt sich in analoger Weise auf allgemeine topologische Räume $(\Omega, \mathcal{O})$ übertragen. (siehe [Sch15, S.6])

Bemerkung: Die Borel σ-Algebra $\mathscr{B}(\mathbb{R}^k)$ wird auch von anderen Mengensystemen wie den abgeschlossenen oder den kompakten Mengen auf $\mathbb{R}^k$ erzeugt. (siehe [Kle08, S.10])

Definition 1.2 (Maß)**.** *Eine Abbildung μ auf einer σ-Algebra $\mathcal{A}$ über der Menge Ω heißt Maß, wenn folgende drei Bedingungen erfüllt sind:*

1) $\forall A \in \mathcal{A}$ *ist* $0 \leq \mu(A) \leq \infty$, *d.h.* $\mu : \mathcal{A} \to [0, \infty]$

2) $\mu(\emptyset) = 0$

3) $\forall A_i, A_j \in \mathcal{A}$ *mit* $i \neq j$ *und* $A_i \cap A_j = \emptyset$ *gilt*

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

Ein Maß heißt endlich, wenn $\forall A \in \Omega$*:* $\mu(A) < \infty$.
Ein Maß heißt σ*-endlich, wenn es eine Folge* $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ *gibt, sodass* $A_n \uparrow \Omega$ *und* $\mu(A_n) < \infty$. *(siehe [Sch15, S.9])*

Bemerkung: Alle in weiterer Folge verwendeten Maße sollen σ-endlich sein.

Definition 1.3 (Produktmaß)**.** *Seien* μ *und* ν σ*-endliche Maße, dann gibt es genau ein* σ*-endliches Maß* $\mu \otimes \nu$*:* $\mathcal{A} \otimes \mathcal{B} \to [0, \infty]$ *mit*

$$\mu \otimes \nu(A \times B) = \mu(A)\nu(B) \quad (A \in \mathcal{A}, B \in \mathcal{B}), \tag{1.1}$$

das sog. Produktmaß, auf der Produkt σ*-Algebra* $\mathcal{A} \otimes \mathcal{B} := \{A \times B : A \in \mathcal{A}, B \in \mathcal{B}\}$. *(vgl. [Els11, S.167])*

Die zwei folgenden Beispiele für Maße nehmen in der Wahrscheinlichkeitstheorie einen besonderen Platz ein. Diese Maße sind Grundlage stetiger bzw. diskreter Verteilungen.

Beispiel 1.2 (Lebesgue-(Borel-)Maß)**.** *Die Mengenfunktion* λ^k *auf* $(\mathbb{R}^k, \mathscr{B}(\mathbb{R}^k))$*, die jedem halboffenen k-dimensionalem Hyper-Rechteck* $\bigotimes_{i=1}^{k}[a_i, b_i)$ *mit* $a_i \leq b_i$ *den Wert*

$$\lambda^k(\bigotimes_{i=1}^{k}[a_i, b_i)) = \prod_{i=1}^{k}(b_i - a_i) \tag{1.2}$$

zuordnet, heißt k-dimensionales Lebesgue-(Borel-)Maß. (vgl. [Sch15, S.12])

Bemerkung: Das Lebesgue-(Borel-)Maß ordnet Intervallen, Rechtecken und Quadern ihre Länge, Flächen und Volumen zu. Der Raum $(\mathbb{R}^k, \mathscr{B}(\mathbb{R}^k))$ wird dann auch euklidisch genannt.

Bemerkung: Die Borel-σ-Algebra ist die kleinste σ-Algebra, die alle (Hyper-) Rechtecke enthält. Etwas verkürzt ausgedrückt, ergibt eine Vervollständigung der Borel-σ-Algebra um alle Teilmengen von Nullmengen, die sog. Lebesgue-σ-Algebra $\mathscr{L}^k$ (siehe [Els11, S.63ff]). Das dann wie oben definierte vollständige Maß auf der

Lebesgue-σ-Algebra wird Lebesgue-Maß genannt. Allerdings wird oft nicht streng zwischen den Bezeichnungen Lebesgue-(Borel-)Maß bzw. Lebesgue-Maß unterschieden. (vgl. [LR05, S.29])

Beispiel 1.3 (Zählmaß)**.** *Wenn Ω eine abzählbare Menge ist, dann wird durch die σ-Algebra $\mathcal{A} = \mathscr{P}(\Omega)$ (Potenzmenge von Ω) und $\mu(A) = \#A$ (= Anzahl der Elemente von A) $\forall A \subseteq A$ ein Maß, das sog. Zählmaß, definiert. (siehe [Kee10, S.2f])*

Definition 1.4 (Messbarkeit)**.**

(i) Ein Messraum ist ein Paar $(\Omega, \mathcal{A})$ bestehend aus einer Menge $\Omega \neq \emptyset$ und einer σ - Algebra $\mathcal{A} \subset \mathscr{P}(\Omega)$.

(ii) Ein Messraum $(\Omega, \mathcal{A})$ heißt diskret, falls Ω höchstens abzählbar ist und $\mathcal{A} = \mathscr{P}(\Omega)$.

(iii) Ein Maßraum ist ein Tripel $(\Omega, \mathcal{A}, \mu)$, wobei $(\Omega, \mathcal{A})$ ein Messraum und μ ein Maß auf $\mathcal{A}$ ist.

(iv) Falls $\mu(\mathcal{A}) = 1$ heißt $(\Omega, \mathcal{A}, \mu)$ Wahrscheinlichkeitsraum mit Wahrscheinlichkeitsmaß μ. Im weiteren Verlauf werden Wahrscheinlichkeitsmaße auch als Wahrscheinlichsverteilungen bezeichnet und mit P bezeichnet. Die Mengen $A \in \mathcal{A}$ werden dann Ereignisse genannt. Und $P(A)$ bezeichnet die Wahrscheinlichkeit, dass A eintritt.

(v) Seien $(\Omega, \mathcal{A})$ und $(\mathcal{X}, \mathcal{B})$ Messräume. Eine Funktion $f : \Omega \to \mathcal{X}$ heißt messbar bzw. $\mathcal{A}$-$\mathcal{B}$ messbar, falls $\forall B \in \mathcal{B} : \ f^{-1}(B) \in \Omega$.

(siehe [Kle08, S.17] und [Kus14, S.46])

1.2 Grundlagen zur Integration

Hier werden einige wichtige Begriffe der Integrationstheorie, die für den weiteren Aufbau wichtig sind, exemplarisch dargestellt.

Definition 1.5 (Lebesgue-Integral)**.** *Sei f eine reellwertige messbare Funktion auf $(\mathcal{X}, \mathcal{A})$, und sei μ ein Maß auf $(\mathcal{X}, \mathcal{A})$.*
Falls f einfach ist, d.h. f hat nur eine endliche Menge an Funktionswerten und f nehme auf den messbaren Mengen $A_1, \ldots, A_m$ aus $\mathcal{A}$ die jeweils unterschiedlichen Werte $a_1, \ldots, a_m$ an, dann gilt für das Integral bezüglich μ

$$\int f d\mu = \sum_i a_i \mu(A_i), \tag{1.3}$$

mit $\mu(A_i) < \infty$, falls $a_i \neq 0$ für $i = 1, \ldots, m$.
Sei f eine nichtnegative messbare Funktion, dann existiert eine Folge f_n von nichtnegativen monoton steigenden einfachen Funktionen, die gegen f konvergiert. Das

Integral von f kann dann wie folgt definiert werden:

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu$$

und mit folgender Zerlegung in positiv- und negativ-Teil

$$\int f^+ d\mu + \int f^- d\mu = \int |f| d\mu$$

nennt man f Lebesgue-integrierbar wenn $\int |f| d\mu < \infty$. (vgl. [LR05, S.31])

Definition 1.6 (Nullmengen)**.** *Sei μ ein Maß auf $(\mathcal{X}, \mathcal{A})$. Eine Menge N heißt Nullmenge bezüglich μ, wenn $\mu(N) = 0$.*
Eine Aussage gilt fast überall bezüglich μ (f.ü. [μ]), falls sie $\forall x \in (\mathcal{X} - N)$ gilt, wobei $x \in (A - B) := (x \in A) \wedge (x \notin B)$. (vgl. [Kee10, S.6])

Satz 1.1 (Eigenschaften von (fast überall [μ]))**.**

- *Wenn $f = 0$ (f.ü. [μ]) $\Rightarrow \int f d\mu = 0$*
- *Wenn $f \geq 0$ und $\int f d\mu = 0 \Rightarrow f = 0$ (f.ü. [μ])*
- *Wenn $f = g$ (f.ü. [μ]) $\Rightarrow \int f d\mu = \int g d\mu$, falls die Integrale existieren*
- *Wenn $f > g$ (f.ü. [μ]) $\Rightarrow \int f d\mu > \int g d\mu$, falls $\mu \neq 0$*

(vgl. [Kee10, S.6])

Der folgende Satz ist für Produktdichten von unabhängigen Zufallsvariablen (vgl. Abschnitt 1.3.2) von großer Bedeutung. Durch diesen Satz lassen sich Integrale von Funktionen, die auf einem Produktmaß definiert sind (vgl. Definition 1.3), resp. mehrdimensionale Integrale auf eindimensionale Integrale zurückführen.

Satz 1.2 (Satz von Fubini)**.** *Seien $(\mathcal{X}, \mathcal{A}, \mu)$ und $(\mathcal{Y}, \mathcal{B}, \nu)$ zwei Maßräume und f sei eine nicht negative $\mathcal{A} \times \mathcal{B}$-messbare Funktion auf $\mathcal{X} \times \mathcal{Y}$. Dann gilt*

$$\begin{aligned}\int_{\mathcal{X}} \left[\int_{\mathcal{Y}} f(x,y) d\nu(y) \right] d\mu(x) &= \int_{\mathcal{Y}} \left[\int_{\mathcal{X}} f(x,y) d\mu(x) \right] d\nu(y) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} f d(\mu \times \nu)\end{aligned}$$

(vgl. [LC98, S.13])

Beweis. siehe [Els11, S.175ff] □

Kompakte Zusammenstellungen der Integrationstheorie zum Aufbau der mathematischen Statistik finden sich in etwa in [LC98]. Eine ausführliche Darstellung findet sich in [Kus14].

1.3 Wahrscheinlichkeitstheoretische Grundlagen

1.3.1 Zufallsvariablen und Verteilungen

Definition 1.7 (Zufallsvariable)**.** *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $(\mathcal{X}, \mathcal{B})$ ein Messraum. Eine messbare Funktion $X : \Omega \to \mathcal{X}$ heißt stochastische Größe bzw. Zufallsvariable mit Werten in $(\mathcal{X}, \mathcal{B})$. Ist $(\mathcal{X}, \mathcal{B}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, dann bezeichnet man X als (1-dimensionale) reelle Zufallsvariable. (vgl. [Kle08, S.43])*

Definition 1.8 (Verteilungsfunktion)**.** *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ ein Messraum und X eine Zufallsvariable $X : \Omega \to \mathbb{R}$. Dann wird das Wahrscheinlichkeitsmaß P_X für Borel-Mengen B*

$$P_X(B) := P(X^{-1}(B)) = P(\{\omega \in \Omega : X(\omega) \in B\}) = P(X \in B) \tag{1.4}$$

Verteilung von X genannt. Falls X die Verteilung Q hat, d.h. $P_X = Q$, schreibt man $X \sim Q$. Im Fall einer konkret vorliegenden Verteilung ergibt sich die übliche Schreibweise

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X \leq x\}) = P_X((-\infty, x]) \tag{1.5}$$

Für die Verteilungsfunktion F_X wird auch die Abkürzung cdf („cumulative distribution function“) verwendet. (vgl. [Kee10, S.6])

Für Verteilungsfamilien ist der folgende Begriff des dominierenden Maßes von zentraler Bedeutung. Dazu [Rü14, S.57]: *„Dominierte Verteilungsklassen lassen sich durch die im Vergleich zu Wahrscheinlichkeitsmaßen viel einfacheren Wahrscheinlichkeitsdichten beschreiben. Dieses führt auch in der Statistik zu starken Vereinfachungen und erweiterten Möglichkeiten zur Konstruktion von statistischen Verfahren (. . .).“*

Definition 1.9 (Dominierendes Maß)**.** *Seien P und μ Maße auf einem Messraum $(\mathcal{X}, \mathcal{A})$. Dann nennt man P absolut stetig bezüglich μ, geschrieben $P \ll \mu$, oder μ dominiert P falls*

$$\forall (A \in \mathcal{A}) : \ \mu(A) = 0 \Rightarrow P(A) = 0$$

(vgl. [Kee10, S.7])

Satz 1.3 (Radon-Nikodym)**.** *Wenn ein σ-endliches Maß P absolut stetig bezüglich eines σ-endliches Maßes μ ist, also $P \ll \mu$, dann existiert eine nichtnegative messbare Funktion f, sodass*

$$P(A) = \int_A f d\mu.$$

Diese Funktion f heißt Radon–Nikodym-Ableitung von P bezüglich μ oder die Dichte von P bezüglich μ

$$f = \frac{dP}{d\mu} \tag{1.6}$$

(vgl. [Kee10, S.7])

Bemerkung: Dichten sind nicht eindeutig, aber f.ü. [μ] eindeutig, d.h. auf Nullmengen bezüglich μ müssen Dichten nicht übereinstimmen. Aber falls X Verteilung P_X hat, also $X \sim P_X$, und $P_X \ll \mu$ gilt, d.h. P_X ist absolut stetig bezüglich μ mit Dichte $p = dP_X/d\mu$, wird p als *die* Dichte von X bezüglich μ bezeichnet. (vgl. [Kee10, S.7])

Beispiel 1.4 (Absolut stetige Zufallsvariable)**.** *Falls X eine reelle Zufallsvariable die Dichte p bezüglich des Lebesgue-Maßes μ hat, dann nennt man X bzw. die Verteilung von X absolut stetig mit Dichte p. Es wird dann p als Wahrscheinlichkeitsdichtefunktion („probability density function", pdf) bezeichnet. Aus dem Satz von Radon–Nikodym folgt*

$$F_X(x) = P(X \leq x) = P_X((-\infty, x]) = \int_{-\infty}^{x} p(u)d\mu(u) \tag{1.7}$$

Es kann daher die Dichte einer pdf durch Ableitung der Verteilungsfuntion bestimmt werden, d.h. $p(x) = F'_X(x)$, falls die Ableitung existiert. (vgl. [Kee10, S.7])

Beispiel 1.5 (Diskrete Zufallsvariable)**.** *Sei $\mathcal{X}_0$ eine abzählbare Teilmenge von $\mathbb{R}$. Dann kann durch*

$$\nu(B) = \#(\mathcal{X}_0 \cap B)$$

das Zählmaß auf Borel-Mengen $B \in \mathscr{B}(\mathbb{R})$ durch

$$\int f d\nu = \sum_{x \in \mathcal{X}_0} f(x)$$

festgelegt werden. Sei X eine Zufallsvariable und es gelte

$$P(X \in \mathcal{X}_0) = P_X(\mathcal{X}_0) = 1$$

dann ist X eine diskrete Zufallsvariable.
Sei N eine Nullmenge auf ν, d.h. $\nu(N) = 0$. Nach Definition von ν ist $\#(N \cap \mathcal{X}_0) = 0$, daher ist $N \cap \mathcal{X}_0 = 0$ und somit $N \subset \mathcal{X}_0^c (= \mathbb{R} - \mathcal{X}_0)$. Dann ist $P_X(N) = P(X \in N) \leq P(X \in \mathcal{X}_0^c) = 1 - P(X \in \mathcal{X}_0) = 0$. Daher ist N eine Nullmenge auf P_X und daraus

folgt, dass $P_X \ll \nu$. Die Dichte p von P_X bezüglich ν erfüllt

$$P(X \in A) = P_X(A) = \int_A p d\nu = \sum_{x \in \mathcal{X}_0} p(x) \mathbb{I}_A(x)$$

Daher ergibt sich für eine einelementige Menge $A = \{y\}$ mit $y \in \mathcal{X}_0$, dass $(X \in A) \Leftrightarrow (X = y)$ und

$$P(X = y) = \sum_{x \in \mathcal{X}_0} p(x) \mathbb{I}_{\{y\}}(x) = p(x)$$

Die Dichte p wird dann (diskrete) Wahrscheinlichkeitsfunktion (bzw. „probability mass function", pmf) genannt. (vgl. [Kee10, S.7f])

Bemerkung: Da $\mathcal{X}_0$ eine Nullmenge ist, könnte der Wert der Dichte $p(y)$ für $y \notin \mathcal{X}_0$ beliebig angenommen werden. Üblicherweise wird $p(y) = 0$ gesetzt, wenn $y \notin \mathcal{X}_0$ ist, und dann lässt sich $p(y) = P(Y = y)\ \forall y$ schreiben. (vgl. [Kee10, S.8])

Definition 1.10 (Verteilungsfamilien)**.** *Eine Verteilungsfamilie bzw. Verteilungsklasse ist eine Zusammenfassung einzelner Verteilungen, die auf einem gemeinsamen Messraum $(\mathcal{X}, \mathcal{A})$ definiert sind, zu einer Menge $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$*
Wenn alle Verteilungen in $\mathcal{P}$ absolut stetig bezüglich eines gemeinsamen Maßes μ sind, wird $\mathcal{P}$ als von μ dominiert bezeichnet. Die zwei wichtigsten Fälle sind:

(i) Diskrete Verteilungen: Hier besteht der Messraum $(\mathcal{X}, \mathcal{A})$ aus einer abzählbaren Menge $\mathcal{X}$ und deren Potenzmenge $\mathcal{A} = \mathscr{P}(\mathcal{X})$. Eine Familie $\mathcal{P}$ von diskreten Verteilungen wird vom Zählmaß dominiert.

(ii) Absolut stetige Verteilungen: In diesem Fall legen eine (Borel-)messbare Menge $\mathcal{X} \subset \mathbb{R}^k$ und die Borel-σ-Algebra $\mathscr{B}(\mathbb{R}^k)$ den Messraum $(\mathcal{X}, \mathcal{A})$ fest. Eine Familie $\mathcal{P}$ von absolut stetigen Verteilungen wird somit vom Lebesgue-Maß dominiert.

(vgl. [LC98, S.14])

Definition 1.11 (Träger)**.** *Der Träger bzw. support einer Funktion $f : A \to \mathbb{R}^k$ wird durch die Nicht-Nullstellenmenge festgelegt:*

$$Tr(f) := \{x \in A : f(x) \neq 0\} \quad A \subset \mathbb{R}^k.$$

Im Fall einer stetigen Zufallsvariable X mit Dichte p_X ist $Tr(p_X) := \{x \in \mathbb{R} : p_X(x) > 0\}$ und im Fall einer diskreten Zufallsvariable $Tr(P) := \{x \in \mathbb{R} : P(X = x) > 0\}$ ([Oli14, S.11 und S.30])

Bemerkung: Im Allgemeinen wird der Träger durch die abgeschlossene Hülle der Nicht-Nullstellenmenge festgelegt: $Tr(f) := \overline{\{x \in A : f(x) \neq 0\}}$. Für eine Zufallsvariable X kann man allgemein festlegen , dass der Träger (im Sinne der Definition 1.11) die kleinste (abgeschlossene) Menge $Tr \subset \mathbb{R}^k$ ist, für die $P(X \in Tr) = 1$. (vgl. [Bil95, S.23])

Bemerkung: Im Fall einer Dichte p, die von den Daten abhängt, wird $Tr(p)$ Stichprobenraum genannt. ([Oli14, S.11 und S.30])

1.3.2 Erwartungswert und Varianz

Definition 1.12 (Erwartungswert)**.** *Sei X eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, dann ist der Erwartungswert von X definiert durch*

$$\mathbb{E}X = \int_\Omega X dP,$$

falls das Integral existiert, d.h. falls $\int_\Omega |X| dP < \infty$.
Bei Vorliegen einer konkreten Verteilung, also wenn $X \sim P_X$ gilt, ist

$$\mathbb{E}X = \int_\Omega x dP_X(x)$$

bzw. für $Y = f(X)$ ist

$$\mathbb{E}Y = \mathbb{E}f(X) = \int_\Omega f dP_X.$$

Und im Fall, dass P_X eine Dichte bezüglich μ besitzt, ergibt sich

$$\int_\Omega f dP_X = \int_\Omega f p d\mu$$

(vgl. [Kee10, S.8f])

Beispiel 1.6 (Erwartungswert absolut stetiger Zufallsvariablen)**.** *Im Fall einer absolut stetigen Zufallsvariable mit Dichte p, ist der Erwartungswert*

$$\mathbb{E}X = \int_\Omega x p(x) d\mu(x)$$

(vgl. [Kee10, S.9])

Beispiel 1.7 (Exponentialverteilung)**.** *Eine Zufallsgröße ist exponentialverteilt mit dem Parameter λ, d.h. $X \sim Exp(\lambda)$, wenn sie die Dichte*

$$p(x) = \lambda e^{-\lambda x} \quad x > 0 \quad \lambda > 0$$

besitzt.

Dann ergibt sich für den Erwartungswert mit Hilfe partieller Integration

$$\begin{aligned}\mathbb{E}X &= \int_0^\infty x\lambda e^{-\lambda x}dx \\ &= \left(-xe^{-\lambda x}\right)_0^\infty + \int_0^\infty e^{-\lambda x}dx \\ &= \left(-xe^{-\lambda x} - \frac{1}{\lambda}e^{-\lambda x}\right)_0^\infty \\ &= \lim_{t\to\infty}\left(-te^{-\lambda t} - \frac{1}{\lambda}e^{-\lambda t}\right) - (0 - \frac{1}{\lambda}) \\ &= \frac{1}{\lambda}\end{aligned}$$

Beispiel 1.8 (Erwartungswert diskreter Zufallsvariablen)**.** *Sei X eine diskreten Zufallsvariable, d.h.* $P(X \in \mathcal{X}_0) = 1$ *für eine abzählbare Menge* $\mathcal{X}_0$*,* μ *das Zählmaß auf* $\mathcal{X}_0$ *und p die Wahrscheinlichkeitsfunktion* $p(x) = P(X = x)$*, dann gilt*

$$\mathbb{E}X = \int x dP_X(x) = \int xp(x)d\mu(x) = \sum_{x\in\mathcal{X}_0} xp(x)$$

(vgl. [Kee10, S.9])

Satz 1.4 (Rechenregeln für Erwartungswerte)**.** *Es seien* X, Y *Zufallsvariablen auf* $(\Omega, \mathcal{A}, P)$*,* $a, b \in \mathbb{R}$ *und die Existenz der Erwartungswerte immer vorausgesetzt, d.h.* $\mathbb{E}X < \infty$ *und* $\mathbb{E}Y < \infty$*. Dann gilt:*

(i) Linearität $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$

(ii) Monotonie $X \leq Y$ $(f.ü.[P]) \Rightarrow \mathbb{E}X \leq \mathbb{E}Y$ *mit* $\mathbb{E}X = \mathbb{E}Y$ *falls* $X = Y$ $(f.ü.[P])$

(iii) Für $X \geq 0$ $(f.ü.\ [P])$ *gilt:* $\mathbb{E}X = 0 \Leftrightarrow X = 0$ $(f.ü.[P])$

(vgl. [Kle08, S.104])

Beweis. Die Rechenregeln folgen direkt aus den Rechenregeln für Integrale. □

Definition 1.13 (Varianz, Kovarianz, Korrelation)**.** *Sei X eine Zufallsvariable mit* $\mathbb{E}X < \infty$*, dann ist die Varianz von X*

$$var(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

Falls X absolut stetig mit Dichte p ist (vgl. 1.6), kann die Varianz folgendermaßen beschrieben werden

$$var(X) = \int_\Omega (x - \mathbb{E}X)^2 p(x)d\mu(x),$$

und im Fall einer diskreten Zufallsvariable mit der Wahrscheinlichkeitsfunktion p (vgl. 1.8)

$$var(X) = \sum_{x \in \mathcal{X}_0} (x - \mathbb{E}X)^2 p(x).$$

Zur Berechnung der Varianz wird häufig die Umformung

$$var(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X^2 - 2X \cdot \mathbb{E}X + (\mathbb{E}X)^2) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

verwendet (sog. „Verschiebungssatz“). Und für die Kovarianz ergibt sich

$$cov(X,Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY - X \cdot \mathbb{E}Y - \mathbb{E}X \cdot Y + (\mathbb{E}X)(\mathbb{E}Y)] = \mathbb{E}XY - \mathbb{E}X \cdot \mathbb{E}Y$$

Die Korrelation als Maß für den linearen Zusammenhang zwischen X und Y, wird definiert durch

$$cor(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

und nimmt Werte in $[-1,1]$ an. Absolute Werte nahe dem Wert 1 bedeuten einen starken Zusammenhang, Werte nahe 0 bedeuten wenig und der Wert null bedeutet keinen (linearen) Zusammenhang. (vgl. [Kee10, S.10])

Definition 1.14 (Unabhängigkeit)**.** *Seien $(\mathcal{X}, \mathcal{A}, P_X)$ und $(\mathcal{Y}, \mathcal{B}, P_Y)$ Wahrscheinlichkeitsräume. Zwei Zufallsvariable X und Y werden (stochastisch) unabhängig genannt, wenn*

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

für alle Borel-Mengen A,B gilt. (vgl. [Kee10, S.13])

Die Verteilung von $Z \in A \times B$ ist in diesem Fall $P_Z(A \times B) = P_X(A) \times P_Y(B)$, d.h. die Verteilung von Z ist das Produktmaß $P_Z = P_X \times P_Y$. Weiters ergibt sich bei Vorhandensein von Dichten von X und Y die Dichte von Z mit Hilfe des Satzes von Fubini als das Produkt der Dichten: $p_Z(x,y) := p_X(x)p_Y(y)$

Bemerkung: Dies lässt sich leicht auf ein Produkt endlich vieler unabhängiger Zufallsvariablen erweitern: $X_1, \ldots, X_n$ sind unabhängig, wenn

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n)$$

für beliebige Borel-mengen $B_1, \cdots, B_n$. Die gemeinsame Verteilung ergibt sich aus dem eindeutigen Produktmaß

$$P(B_1 \times \cdots \times B_n) = P_{X_1}(B_1) \times \cdots \times P_{X_n}(B_n).$$

Und falls die X_i mit $i = 1, \ldots, n$ Dichten bezüglich μ_i besitzen, ist die gemeinsame

Dichte

$$p(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n) \quad \text{bezüglich } \mu = \mu_1 \times \cdots \times \mu_n$$

Falls $X_1, \ldots, X_n$ unabhängig und identisch verteilt sind, also $X_i \sim Q$, dann bezeichnet man $X_1, \ldots, X_n$ als iid (*independent and identically distributed*) (siehe Definition 2.3 im nächsten Kapitel) (vgl. [Kee10, S.14])

Bemerkung: Unkorreliertheit, d.h. $cov(X, Y) = 0$, bedeutet nicht stochastische Unabhängigkeit.

Beispiel 1.9. *Sei X stetig gleichverteilt auf* $[-1, 1]$ *und* $Y = X^2$ *zwei Zufallsvariablen, so ergibt* $cov(X, Y) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2) = 0$*, aber X und Y sind nicht unabhängig. Umgekehrt folgt aber aus der stochastischen Unabhängigkeit, dass* $cov(X, Y) = 0$*, also die Unkorreliertheit. (vgl. [Kle08, S.104])*

1.3.3 Bedingte Verteilungen und bedingte Erwartung

Die Formel der elementaren bedingten Wahrscheinlichkeit zweier Ereignisse lautet

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0.$$

Sie gibt die Wahrscheinlichkeit des Eintretens des Ereignisses A an, unter der Bedingung, dass B zuvor eingetreten ist, oder kurz: die Wahrscheinlichkeit für das Ereignis A unter Voraussetzung des Ereignisses B (vgl. [Bil95, S.427]). Diese Formel lässt sich unmittelbar auf die Definition einer bedingten diskreten Verteilung übertragen:

Definition 1.15 (Diskrete bedingte Verteilung)**.** *Seien X und Y diskrete Zufallsvariablen und sei* $p_X(x) = P(X = x)$ *die Wahrscheinlichkeitsfunktion von X. Die (diskrete) Wertemenge von X sei* $\mathcal{X}_0 = \{x : p_X(x) > 0\}$*. Dann definiert man die bedingte Verteilung von Y gegeben* $X = x$ *durch*

$$Q_x(B) = P(Y \in B | X = x) = \frac{P(Y \in B, X = x)}{P(X = x)} \tag{1.8}$$

für alle Borel-Mengen B und $x \in \mathcal{X}_0$*.* Q_x *legt* $\forall x \in \mathcal{X}_0$ *ein Wahrscheinlichkeitsmaß fest. (vgl. [Kee10, S.15])*

Für stetige Verteilungen soll hier nur der Fall von reellen Zufallsvariablen betrachtet werden. Die allgemeine Definition einer bedingten Verteilung in maßtheoretischer Darstellung findet sich in [Bil95, S.430].

Sei X eine Zufallsvariable mit Werten in $(\mathbb{R}^m, \mathscr{B}(\mathbb{R}^m))$ und Y eine Zufallsvariable mit Werten in $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ und sei $Z = (X, Y)$ eine Zufallsvariable mit Werten in $(\mathbb{R}^{m+n}, \mathscr{B}(\mathbb{R}^{m+n}))$. Die Verteilung P_Z von Z habe die Dichte p_Z bezüglich des Produktmaßes $\mu \times \nu$, wobei μ und ν σ-endliche Maße auf $\mathbb{R}^m$ bzw. $\mathbb{R}^n$ sind. Die Dichte

P_Z heißt dann gemeinsame Dichte von X und Y:

$$P(Z \in B) = \int \int \mathbb{I}_B(x,y) p_Z(x,y) d\mu(x) d\nu(y). \tag{1.9}$$

Um $P(X \in A)$ zu berechnen, wobei $(X \in A) \Leftrightarrow (Z \in A \times \mathbb{R}^n)$, erlaubt es der Satz von Fubini die Integrationsreihenfolge zu vertauschen. Daher ist mit $\mathbb{I}_{A\times\mathbb{R}^n}(x,y) = \mathbb{I}_A(x)$

$$\begin{aligned} P(X \in A) = P(Z \in A \times \mathbb{R}^n) &= \int \int \mathbb{I}_A(x) p_Z(x,y) d\nu(y) d\mu(x) \\ &= \int_A \left\{ \int p_Z(x,y) d\nu(y) \right\} d\mu(x) \end{aligned}$$

Daraus lässt sich die *Randdichte* von X bezüglich μ ablesen:

$$p_X(x) = \int p_Z(x,y) d\nu(y). \tag{1.10}$$

Analog ist die Randdichte von Y bezüglich ν

$$p_Y(y) = \int p_Z(x,y) d\mu(x). \tag{1.11}$$

(vgl. [Kee10, S.101])

Definition 1.16 (Stetige bedingte Verteilung)**.** *Die Funktion Q ist eine bedingte Verteilung von Y gegeben X, d.h.*

$$(Y|X = x) \sim Q_x,$$

wenn

i) $Q_x(\cdot)$ *ein Wahrscheinlichkeitsmaß* $\forall x$ *ist,*

ii) $Q_x(B)$ *eine messbare Funktion für alle Borel-Mengen B ist, und*

iii) für alle Borel-Mengen A und B gilt

$$P(X \in A, Y \in B) = \int_A Q_x(B) dP_X(x).$$

(vgl. [Kee10, S.102f])

Satz 1.5. *Seien X und Y wie oben definiert,* p_X *die Randdichte von X und* $Tr(p_X) = \{x : p_X(x) > 0\}$ *der Träger der Randdichte. Dann ist für* $x \in Tr(p_X)$

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \tag{1.12}$$

die Dichte der bedingten Verteilung von Y gegeben X. (vgl. [Kee10, S.103])

Beweis. (siehe [Kee10, S.103f]) □

Für den weiteren Aufbau reicht folgende Definition des Erwartungswertes:

Definition 1.17 (Bedingter Erwartungswert bei Vorliegen von Dichten)**.** *Im Fall obiger bedingter Verteilungen lässt sich die bedingte Erwartung in der Form*

$$\mathbb{E}[Y|X=x] = \int y \cdot p_{Y|X}(y|x) d\nu(y) \tag{1.13}$$

angeben.

Bemerkung: Im Allgemeinen wird der bedingte Erwartungswert über Unter-σ-Algebren definiert:

Definition 1.18 (Bedingter Erwartungswert allgemein)**.** *Sei X eine messbare Zufallsvariable auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und $\mathcal{F}$ eine Unter-σ-Algebra von $\mathcal{A}$, d.h. $\mathcal{F} \subset \mathcal{A}$ und $\mathcal{F}$ ist selbst wieder eine σ-Algebra. Dann ist die Zufallsvariable $\mathbb{E}[X|\mathcal{F}]$, d.h. der Erwartungswert der Zufallsvariablen X bezüglich $\mathcal{F}$ durch folgende zwei Bedingungen festgelegt:*

(i) $\mathbb{E}[X|\mathcal{F}]$ *ist $\mathcal{F}$-messbar*

(ii) $\mathbb{E}[X|\mathcal{F}]$ *erfüllt die Funktionalgleichung*

$$\int_F \mathbb{E}[X|\mathcal{F}] dP = \int_F X dP \quad \forall F \subset \mathcal{F}$$

(siehe [Bil95, S.445] und [Kus14, S.228f])

Diese allgemeine Form wird aber im weiteren Verlauf nicht benötigt.

Bemerkung: Die Bedingung (ii) aus Definition 1.18 könnte auch in der Form $\int_F (X - \mathbb{E}[X|\mathcal{F}]) dP = 0 \quad \forall F \subset \mathcal{F}$ geschrieben werden. Dies erlaubt folgende Interpretation: Falls ein Beobachter, dem (alle Information aus) $\mathcal{F}$ vollständig bekannt ist, die Möglichkeit erhält auf das Eintreten des Ereignis F zu wetten, wäre diese Wette bei einem Einsatz von $\mathbb{E}[X|\mathcal{F}]$ mit der Auszahlung eines Betrages der Höhe X fair, d.h. der Erwartungswert für den Gewinn/Verlust ist gleich 0. (siehe [Bil95, S.445])

Satz 1.6 (Rechenregeln für bedingte Erwartungswerte)**.** *Es seien X, Y, Z Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, $a, b \in \mathbb{R}$ und die Existenz der Erwartungswerte sei immer vorausgesetzt. Dann gilt:*

(i) Linearität: $\mathbb{E}(aZ + bY|X) = a\mathbb{E}(Z|X) + b\mathbb{E}(Y|X)$

(ii) Monotonie: $X \leq Y \ (f.\ddot{u}.[P]) \Rightarrow \mathbb{E}(Z|X) \leq \mathbb{E}(Y|X)$ *mit* $\mathbb{E}(Z|X) = \mathbb{E}(Y|X)$ *falls* $Z = Y \ (f.\ddot{u}.[P])$

(iii) Unabhängigkeit: Sind X und Y unabhängig, dann gilt: $\mathbb{E}(Y|X) = \mathbb{E}Y$

(iv) Turmeigenschaft: $\mathbb{E}[\mathbb{E}(X|\mathcal{F})|\mathcal{G}] = \mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{F}] = \mathbb{E}(X|\mathcal{F})$ *für* σ*-Algebren* $\mathcal{F}, \mathcal{G}$ *mit* $\mathcal{F} \subset \mathcal{G} \subset \mathcal{A}$

(vgl. [Kle08, S.176])

Beweis. (siehe [Kle08, S.176f]) □

Eine gute zusammenfassende Beschreibung zur Bedeutung des bedingten Erwartungswertes findet sich in [Rü14, S.391]:

> *Bedingte Erwartungswerte bilden die Grundlage für den Begriff der Suffizienz. Sie finden sich auch in der Konstruktion von verbesserten Schätzverfahren (Satz von Rao-Blackwell) und allgemeiner Entscheidungsverfahren. In der Testtheorie sind sie die Grundlage für die Methode der bedingten Tests.*

2 Exponentielle Familien und Suffizienz

2.1 Statistische Grundlagen

Ein typisches statistisches Problem besteht in der Betrachtung einer Zufallsvariable $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ als Ergebnis eines Experiments und dessen Ergebnissen (*Realisierungen, Daten*). Die Festlegung des Zustandsraums $\mathcal{X}$ erfolgt durch die (zu erwartende) Art der Realisierungen. Das eigentliche Problem besteht aber darin, die Verteilung der Zufallsvariable X zu bestimmen. Dazu wird vorab die Familie $\mathcal{P}$ der in Frage kommenden Verteilungen für die Größe X festgelegt, das sogenannte *statistische Modell*. (vgl. [Fel14] S.11)

Die statistischen Modelle lassen sich in zwei Gruppen einteilen. In sogenannte *parametrische* und *nicht-parametrische* Modelle.

Bei einem parametrischen Modell hat man Kenntnis über die Art der Verteilung. Im Fall von stetigen Verteilungen ist die Art der Dichtefunktion $f(x;\theta)$ bzw. $f_\theta(x)$ oder im diskreten Fall ist die Art der Wahrscheinlichkeitsfunktion $p(x;\theta)$ bzw. $p_\theta(x)$ bekannt. Dabei werden die Verteilungen $P_\theta \in \mathcal{P}$ durch einen Parameter $\theta \in \mathbb{R}^k$ bestimmt. Der Parameterraum Θ ist in diesem Fall endlichdimensional. Aber der Parameter θ ist unbekannt und daher ist dessen Bestimmung ein mögliches Ziel der statistischen Untersuchung.

Bei einem nicht-parametrischen Modell ist der Raum der möglichen Maße nicht endlichdimensional. Sei zum Beispiel $\mathcal{P}$ die Menge aller stetigen Verteilungen, so wäre dieser Raum nicht endlichdimensional (vgl. [Fel14] S.12). Eine Bestimmung von Informationen über die Verteilung erfolgt in diesem Fall über die *empirische Verteilungsfunktion*. Dieser Ansatz wird in dieser Arbeit aber nicht weiter verfolgt, da sich diese Arbeit vor allem auf spezielle parametrisches Modelle, die exponentiellen Familien, beschränkt.

Definition 2.1 (Statistisches Experiment)**.** *Eine Menge von Wahrscheinlichkeitsräumen* $\{(\mathcal{X}, \mathcal{B}, P), P \in \mathcal{P}\}$ *heißt statistisches Experiment. (vgl. [Fel14] S.11)*

Definition 2.2 (Statistik)**.** *Eine Statistik ist eine meßbare Abbildung* $T : (\mathcal{X}, \mathcal{B}) \to (\mathbb{R}^k, \mathscr{B}(\mathbb{R}^k))$ *mit der Borel-*σ*-Algebra* $\mathscr{B}(\mathbb{R}^k)$*. (vgl. [Fel14, S.11])*

Bemerkung: Eine Statistik ist i.A. eine beliebige messbare Funktion $T : (\mathcal{X}, \mathcal{B}) \to$

$(\mathcal{T}, \mathcal{C})$. Aber in weiterer Folge soll, wenn T Werte in einem euklidischen Raum annimmt, immer $\mathcal{C} = \mathscr{B}(\mathbb{R}^k)$ (Borel-σ-Algebra) gelten.

Definition 2.3 (Zufallsstichprobe)**.** *Seien die Zufallsvariablen $X_1, X_2, \ldots, X_n$ unabhängig und identisch verteilt (independent and identically distributed, daher kurz iid), dann bilden sie zusammen eine Zufallsstichprobe der Größe n der gemeinsamen Verteilung.*
Die Realisierungen einer Zufallsstichprobe $(X_1, X_2, \ldots, X_n)$ werden als $(x_1, x_2, \ldots, x_n)$ geschrieben. (vgl. [HMC12] S.204)

Bemerkung: Sei $(X_1, X_2, \ldots, X_n)$ eine Zufallsstichprobe der Zufallsvariable X. Dann ist eine Statistik eine meßbare Funktion $T(X_1, X_2, \ldots, X_n)$. (vgl. [HMC12] S.204)

Beispiel 2.1 (Bernoulli-Experiment)**.** *Die Zufallsstichprobe $(X_1, X_2, \ldots, X_n)$ mit Daten $(x_1, x_2, \ldots, x_n)$ entstamme einem Bernoulli-Experiment. Dies bedeutet der Zustandsraum $\mathcal{X} = \{0, 1\}^n$ besteht aus allen $0/1$-Folgen der Länge n. Die Familie $\mathcal{P}$ entspricht dem Intervall $[0, 1]$. Das Modell ist dann eine Familie von Produktmaßen P der Form $P = \bigotimes_{i=1}^n P_i$ mit identischen Verteilungen P_i. Die P_i sind gegeben durch*

$$P_i(X = 1) = p \qquad \text{mit} \quad 0 \leq p \leq 1.$$

Und hier ist der Parameterraum $\Theta = [0, 1]$

Das Stichprobenmittel resp. die relative Häufigkeit $T(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$ wäre hier eine naheliegende Statistik. (vgl. [Fel14] S.12)

Beispiel 2.2 (Beispiele für Statistiken)**.** *Sei $X = (X_1, X_2, \ldots, X_n)$ eine Zufallsstichprobe mit Daten $(x_1, x_2, \ldots, x_n)$.*

- *Stichprobenmittel $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$*
- *Stichprobenvarianz $s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ bzw. die (bias-)korrigierte Stichprobenvarianz $s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$*
- *Rangstatistik $T(x_1, x_2, \ldots, x_n) = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ mit $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)})$*
- *Ordnungsstatistik $T(x_1, x_2, \ldots, x_n) = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ mit $x_{(1)} < x_{(2)} < \cdots < x_{(n)})$ für den Fall, dass $X_1, X_2, \ldots, X_n$ unabhängig stetig verteilt sind, d.h. es gilt fast sicher $x_i \neq x_j$ für $i \neq j$, d.h. es gibt keine Bindungen. (vgl. [LR05] S.37f)*

2.2 Exponentielle Familien

Zu den exponentiellen Familien gehören viele der bekannten Verteilungen aus der Wahrscheinlichkeitstheorie und Statistik, wie z.B. Normalverteilung, Binomialverteilung, Poisson-Verteilung, Beta-Verteilung und viele andere. Bekannte Beispiele

für Verteilungen, die keine exponentielle Familien bilden, sind etwa die stetige Gleichverteilung oder die Cauchy-Verteilung.

Verteilungen, die einer exponentiellen Familie angehören, weisen viele interessante und nützliche Eigenschaften auf: Für sie, d.h. für deren Parameter, existieren suffiziente Statistiken, sie erfüllen viele sogenannte „Regularitätsbedingungen“, die sie für mathematische „Werkzeuge“ etwa aus der Analysis oder Wahrscheinlichkeitstheorie zugänglich machen, für sie sind statistische Testverfahren in gewisser Weise „optimal“ einsetzbar und viele weitere. Kurz: Exponentielle (Verteilungs-)Familien dienen in vielen Bereichen der Wahrscheinlichkeitstheorie und Statistik als Beispiele, wenn es darum geht, Eigenschaften für grundlegende Verfahren und Prinzipien zu demonstrieren.

Diese vielen „angenehmen“ Eigenschaften dürfen aber in der Anwendung (also im vorliegenden Fall in der Statistik) nicht dazu führen, darauf zu verzichten, genau zu prüfen, ob exponentielle Familien in dem jeweiligen Fall zur Modellbildung verwendet werden können.

Definition 2.4 (Exponentielle Familie)**.** *Eine Familie von Wahrscheinlichkeitsmaßen $\{P_\theta : \theta \in \Theta\}$ auf dem Messraum $(\mathcal{X}, \mathcal{B})$ mit $\theta = (\theta_1, \ldots, \theta_k)$, wobei jedes Maß von einem σ-endlichen Maß μ dominiert wird, heißt exponentielle Familie, falls ihre Radon-Nikodym-Ableitungen $p_\theta(x) = (dP_\theta/d\mu)(x)$ folgende Gestalt haben:*

$$p_\theta(x) = C(\theta) \cdot exp\left[\sum_{j=1}^{k} Q_j(\theta) T_j(x)\right] h(x). \tag{2.1}$$

Dabei sind $C(\theta) \geq 0$, $h(x) \geq 0$, $x \in \mathbb{R}^n$, $Q_j(\theta)$ und $T_j(x)$ reellwertige Funktionen und $k \in \mathbb{N}$.

Im Fall $k = 1$ spricht man von einer ein-parametrigen Exponentialfamilie, im Fall von $k > 1$ von einer k-parametrigen Exponentialfamilie. (vgl. [LR05, S.46] und [Mar15, S.30])

Durch die Definition einer exponentiellen Familie in der allgemeinen Form von 2.4 sind sowohl stetige als auch diskrete exponentielle Familien durch diese Form festgelegt. Dies entspricht der heutigen Darstellungsform. Historisch wurde die Theorie zu Exponentialfamilien zuerst für stetige Verteilungen unter sehr strikten Voraussetzungen entwickelt (siehe [Koo36]).

Definition 2.5 (Natürliche Parametrisierung einer exponentiellen Familie)**.** *Falls in obiger Definition 2.4 $Q_j(\theta) = \eta_j$ mit $\eta = (\eta_1, \ldots, \eta_k) \in \Omega$ erfüllt ist, nennt man die exponentielle Familie natürlich parametrisiert. Dann lässt sich Gleichung (2.1) in der*

Form

$$p_\eta(x) = \tilde{C}(\eta) \cdot exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] h(x) \tag{2.2}$$

mit dem (natürlichen) Parameterraum $\Omega \subset \mathbb{R}^k$, schreiben. Falls nötig, kann die Funktion $h(x)$ auch zu μ verschoben werden, d.h. $d\tilde{\mu}(x) = h(x)d\mu(x)$. (vgl. [LR05, S.48]

Definition 2.6 (Reguläre exponentielle Familie)**.** *Gegeben seien drei Bedingungen:*

(R1) $\Omega = \{\eta : \int h(x) \cdot exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] d\mu(x) < \infty\}$ *wobei* $\eta = (\eta_1, \ldots, \eta_k)$

(R2) Die $T_j(x)$ sind linear unabhängig, d.h. für $\sum_{j=1}^{k} a_j T_j(x)$ mit $a_1, \ldots, a_k \in \mathbb{R}$ gilt: $\left(\sum_{j=1}^{k} a_j T_j(x) = 0 \ \forall x\right) \Rightarrow \ (a_1 = \cdots = a_k = 0)$

(R3) Ω ist eine (k-dimensionale) nichtleere offene Menge

Falls für eine natürlich parametrisierte exponentielle Familie aus Definition 2.5 die Bedingungen (R1), (R2), (R3) erfüllt sind, nennt man diese eine reguläre exponentielle Familie, kurz REF. Falls nur (R1), (R2) erfüllt sind, hat sie vollen Rang („full (rank) exponential family“) (vgl. [Oli14, S.83f]). Der Wert k wird auch als Ordnung der exponentiellen Familie bezeichnet (vgl. [BN06a]).

Bemerkung: Manche Autoren machen obige Unterscheidung nicht, und bezeichnen eine exponentielle Familie, die (R1)-(R3) erfüllt, als *„full rank“* exponential family (vgl. [LC98, S.24])

Eine wichtige Eigenschaft zeigt folgender Satz:

Satz 2.1. *Der natürliche Parameterraum Ω ist konvex.*

Beweis. Es gelte (R1) und sei $d\tilde{\mu}(x) = h(x)d\mu(x)$ und seien $\eta, \eta' \in \Omega$ mit $\eta = (\eta_1, \ldots, \eta_k)$ bzw. $\eta' = (\eta'_1, \ldots, \eta'_k)$, dann folgt aus der Hölder'schen Ungleichung

$$\int \exp(\sum_j [\alpha\eta_j + (1-\alpha)\eta'_j]T_j(x))d\tilde{\mu}(x)$$
$$\leq \left(\int \exp(\sum_j \eta_j T_j(x))d\tilde{\mu}(x)\right)^{\alpha} \left(\int \exp(\sum_j \eta'_j T_j(x))d\tilde{\mu}(x)\right)^{1-\alpha} < \infty$$

$\forall\alpha \in (0,1) \ \Rightarrow \ (\alpha\eta + (1-\alpha)\eta') \in \Omega \quad \forall\alpha \in (0,1)$ (vgl. [LR05, S.48]) □

Beispiel 2.3 (1-dimensionale Normalverteilung mit bekannter Varianz)**.** *Der Er-*

wartungswert ist a, wobei $x, a \in \mathbb{R}$ und $\sigma \in \mathbb{R}$ const.

$$\begin{aligned} p_a(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) = \\ &= \exp\left(x\frac{a}{\sigma^2} - \frac{1}{2\sigma^2}a^2\right) \exp\left(-\frac{x^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} = \\ &= \exp\left(x\eta - \frac{\sigma^2}{2}\eta^2\right) \underbrace{\exp\left(-\frac{x^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}}}_{h(x)} \end{aligned}$$

$$\eta = \frac{a}{\sigma^2} \ldots \textit{natürlicher Parameter}$$

(vgl. [HMC12, S.398])

Beispiel 2.4 (1-dimensionale Poisson-Verteilung)**.** *Der Erwartungswert ist λ und $x \in \mathbb{N}$*

$$\begin{aligned} p_\lambda(x) &= \frac{\lambda^x \exp(-\lambda)}{x!} = \\ &= \exp[x \log \lambda - \lambda] \frac{1}{x!} = \\ &= \exp[x\eta - \exp(\eta)] \underbrace{\frac{1}{x!}}_{h(x)} \end{aligned}$$

$$\eta = \log \lambda \ldots \textit{natürlicher Parameter}$$

(vgl. [HMC12, S.399])

Satz 2.2 (Eigenschaften von exponentiellen Familien)**.**

a) Sei $\phi(x)$ eine Funktion auf dem Stichprobenraum $(\mathcal{X}, \mathcal{B})$ für die das Integral

$$Z(\eta) := \int \phi(x) \exp\left[\sum_{j=1}^{k} \zeta_j T_j(x)\right] d\mu(x)$$

als Funktion der komplexen Variable $\zeta_j = \alpha_j + i\beta_j$ mit $j = (1, \ldots, k)$ existiert und endlich ist für alle $(\alpha_1, \ldots, \alpha_k) \in \Omega$, dem natürlichen Parameterraum, und $(\beta_1, \ldots, \beta_k) \in \mathbb{R}$. Dann gilt:

(i) $Z(\eta)$ ist eine analytische Funktion

(ii) Ableitungen beliebiger Ordnung von $Z(\eta)$ nach den ζ_j können innerhalb des Integrals durchgeführt werden.

(b) Gegeben sei die Dichte einer exponentiellen Familie mit natürlicher Parametrisierung aus Definition 2.5

$$p_\eta(x) = \tilde{C}(\eta) \cdot exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] h(x)$$

mit Parametervektor $\eta = (\eta_1, \ldots, \eta_k) \in \Omega$.

(i) Die momenterzeugende Funktion von T ist

$$\Psi(t) = \mathbb{E}(\exp(t^\top T)) = \frac{\tilde{C}(\eta_1, \ldots, \eta_k)}{\tilde{C}(\eta_1 + t_1, \ldots, \eta_k + t_k)}$$

mit $t = (t_1, \ldots, t_k) \in \mathbb{R}^k$ und dem Standardskalarprodukt $x^\top x$ (vgl. [Fel14, S.24])

(ii) Der Erwartungswert von T ist gegeben durch

$$\mathbb{E}_\eta T_j = -\frac{\partial \log \tilde{C(\eta)}}{\partial \eta_j} \tag{2.3}$$

und die Kovarianz durch

$$cov(T_i, T_j) = -\frac{\partial^2 \log \tilde{C(\eta)}}{\partial \eta_i \partial \eta_j}$$

wobei sich für $i = j$ die Varianz ergibt $var(T_i) = cov(T_i, T_i)$.

(c) Gegeben sei eine exponentielle Familie mit natürlicher Parametrisierung aus Definition 2.5 in der Form

$$dP_\eta(x) = \tilde{C}(\eta) \cdot exp\left[\sum_{j=1}^{r} \eta_j T_j(x) + \sum_{i=r+1}^{k} \eta_i T_i(x)\right] d\tilde{\mu}(x)$$

mit Parametervektor $\eta = (\eta_1, \ldots, \eta_k) \in \Omega$, $\tilde{\mu}$ ein σ-endliches Maß in $\mathbb{R}^n$, $T_1, \ldots, T_k$ Funktionen von $\mathbb{R}^n \to \mathbb{R}$ und $1 \leq r < k$. Dann existieren Maße λ'_r in $\mathbb{R}^r$ und λ''_r in $\mathbb{R}^{k-r}$, sodass

(i) die Verteilung eines Teilvektors $T' = (T_1, \ldots, T_r)$ von $T = (T_1, \ldots, T_k)$ der Form

$$dP_\eta(x) = \tilde{C}(\eta) \cdot exp\left[\sum_{j=1}^{r} \eta_j T_j(x)\right] d\lambda_{r'}(x)$$

wieder eine exponentielle Familie ist, und

(ii) die bedingte Verteilung $T'' = (T_{r+1}, \ldots, T_k)$ bei gegebenen $T_1 = t_1, \ldots, T_r =$

t_r *mit der Form*

$$dP_{\eta''}(x) = \tilde{C}(\eta'') \cdot exp\left[\sum_{i=r+1}^{k} \eta_i T_i(x)\right] d\lambda_{r''}(x)$$

mit $\eta'' = (\eta_{r+1}, \ldots, \eta_k)$ *ebenfalls eine exponentielle Familie ist.*

Beweis. (a) siehe [LR05, S.49]
(b)(i) siehe [Oli14, S.91f]
(b)(ii) Sei μ ein (σ-endliches Maß. Für die Dichte $p_\eta(x)$ gilt

$$1 = \int \tilde{C}(\eta) \cdot exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] h(x) d\mu(x) \tag{2.4}$$

$$\tilde{C}(\eta)^{-1} = \int exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] h(x) d\mu(x) \tag{2.5}$$

Die Ableitung beider Seiten nach η_i und nach Vertauschung von Integration und Ableitung auf der rechten Seite (nach Punkt (a) dieses Satzes) ergibt:

$$\begin{aligned}
\frac{\partial}{\partial \eta_j} \tilde{C}(\eta)^{-1} &= \int \frac{\partial}{\partial \eta_j} exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] h(x) d\mu(x) \\
-\tilde{C}(\eta)^{-2} \frac{\partial \tilde{C}(\eta)}{\partial \eta_j} \tilde{C}(\eta)^{-1} &= \int T_j(x) exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] h(x) d\mu(x) \\
-\tilde{C}(\eta)^{-2} \frac{\partial \tilde{C}(\eta)}{\partial \eta_j} &= \int T_j(x) exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] h(x) d\mu(x) \quad | \cdot \tilde{C}(\eta) \\
\underbrace{-\tilde{C}(\eta)^{-1} \frac{\partial \tilde{C}(\eta)}{\partial \eta_j}}_{\frac{\partial \log \tilde{C}(\eta)}{\partial \eta_j}} &= \int T_j(x) \underbrace{\tilde{C}(\eta) exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right]}_{p_\eta(x)} h(x)\, d\mu(x)
\end{aligned}$$

Und mit der Definition des Erwartungswertes (siehe Definition 1.12) auf der rechten Seite erhält man:

$$-\frac{\partial \log \tilde{C}(\eta)}{\partial \eta_j} = \mathbb{E}_\eta T_j \tag{2.6}$$

Zweimaliges partielles Ableiten, also $\frac{\partial^2}{\partial \eta_i \partial \eta_j}$ von (2.5), bzw. Ableiten von (2.6) nach η_i, führt auf die Kovarianz (siehe [Oli14, S.90ff]).
(c) siehe [LR05, S.48]

□

Bemerkung: Punkt (a) aus obigem Satz bedeutet insbesondere, dass die Dichte f aus Definition 2.5 reell stetig differenzierbar bezüglich der natürlichen Parameter η_j ist.

2.3 Suffizienz

„Die Suffizienz beschreibt die Möglichkeit einer Datenreduktion ohne Informationsverlust für Entscheidungsprobleme. Datenreduktion lässt sich durch eine Statistik $T(x)$ beschreiben. Anstelle des Beobachtungsvektors $x \in \mathcal{X}$ wird nur die reduzierte Größe $T(x)$ zur Konstruktion von Entscheidungsverfahren verwendet.“ ([Rü14, S.82])

Suffiziente Statistiken enthalten dabei immer die „wesentliche“ Information aus den Daten für die statistische Entscheidung.

Suffizienz ist eine grundlegende Eigenschaft, die in bei vielen anderen statistischen Eigenschaften in notwendiger oder hinreichender Form auftritt.

Definition 2.7 (Suffizienz)**.** *Eine Statistik T ist suffizient für eine Familie $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ von Verteilungen auf dem Stichprobenraum $(\mathcal{X}, \mathcal{A})$, wenn $\forall\ A \in \mathcal{A}$ eine von θ unabhängige Version von $P_\theta(A|T = t) = P(A|T = t)$ existiert. (vgl. [LR05, S.44])*

Bemerkung: Im Allgemeinen sind bedingte Verteilungen nicht eindeutig, sondern nur f.ü. [P]. Eine spezielle bedingte Verteilung wird dann Version genannt (vgl. [Bil95, S.430]).

Bemerkung: Sind $X_1, \ldots, X_n$ iid (*independent identically distributed*) mit $p_\theta(x)$ und dem unbekannten Parameter θ, wobei $p_\theta(x)$ eine Wahrscheinlichkeitsdichtefunktion (*pdf*) (siehe Definition 1.4) oder Wahrscheinlichkeitsfunktion (*pmf*) (siehe Definition 1.5) sein kann. Dann ist eine Statistik $T = T(X_1, \ldots, T_n)$ suffizient für θ bzw. $p_\theta(x)$, wenn die bedingte gemeinsame Dichte der $X_1, \ldots, X_n$, gegeben $T = t$ unabhängig von θ ist.
Suffizienz bedeutet, dass es bei Kenntnis des Werts der Statistik T möglich ist, eine Menge $X_1', \ldots, X_n'$ zu konstruieren mit der gleichen Verteilung wie $X_1, \ldots, X_n$ (vgl. [LR05, S.19]).

Beispiel 2.5 (Bernoulli-Experiment)**.** *Seien $X_1, \ldots, X_n$ iid Bernoulli-verteilt mit Parameter θ und $T(x) = \sum_{i=1}^n x_i$. Dann ist*

$$P_\theta(X = x|T(x) = t) = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}$$

Und $\binom{n}{t}^{-1}$ ist unabhängig von θ, daher ist T suffizient. [Mar15, S.33]

Satz 2.3 (Neyman-Fisher-Faktorisieruns-Theorem)**.** *Die Familie von Wahrscheinlichkeitsmaßen $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Omega\}$ auf $(\mathcal{X}, \mathcal{A})$ sei durch das σ-endliche Maß μ mit Dichten $p_\theta = d\mathbf{P}_\theta/d\mu$ dominiert. Dann ist die Statistik $T(X)$ mit Bildmenge $(\mathcal{T}, \mathcal{B})$ genau dann suffizient, wenn eine $\mathcal{B}$-meßbare Funktionen $g_\theta > 0$ und eine $\mathcal{A}$-messbare Funktion $h > 0$ existieren, sodass*

$$p_\theta(x) = g\left(\theta, T(x)\right) h(x) \tag{2.7}$$

vgl. [Kee10, S.45] und [Fel14, S.26]

Beweis. Der allgemeine Beweis ist technisch und findet sich beispielsweise in [LR05, S.45f]. Für (euklidische) Räume, auf denen die Dichte der bedingten Verteilung existiert (siehe [LR05, S. 44 Theorem 2.6.1]), lässt sich 2.3 unmittelbar nachrechnen. Da die Statistik T eine Funktion von X ist, ist die gemeinsame Dichte von (X, T) gegeben durch $g_\theta(T(x))h(x)$ und die Randdichte von T ist $g_\theta(t)$. Die bedingte Dichte ist der Quotient dieser beiden Dichten und die von θ abhängigen Teile können gekürzt werden. Daher ist die bedingte Verteilung bei gegebenem T unabhängig von θ, wenn T suffizient ist. (siehe [Mar15, S.34] bzw. [Fel14, S.26]) □

Daraus folgt unmittelbar:

Korollar 2.1. *Unter den Voraussetzungen von Satz 2.3 ist eine Statistik U genau dann suffizient, wenn für festes θ und θ' der Quotient*

$$\frac{p_\theta(x)}{p_{\theta'}(x)}$$

nur eine Funktion von $U(x)$ ist. (vgl.[LC98, S.37])

Beispiel 2.6 (Stetige Gleichverteilung)**.** *Seien $X_1, \ldots, X_n$ iid stetig-gleichverteilt mit Parameter θ und Dichte*

$$p_\theta(x) = \begin{cases} 1, & \text{wenn } x \in (\theta, \theta + 1) \\ 0, & \text{sonst} \end{cases}$$

d.h. $p_\theta(x) = \mathbb{I}_{(\theta,\theta+1)}$ ($\mathbb{I}$ ... Indikatorfunktion). Die gemeinsame Dichte ist dann

$$p_\theta(x) = \prod_{i=1}^{n} \mathbb{I}_{(\theta,\theta+1)}(x_i),$$

und diese ergibt 1 genau dann, wenn $\max_i x_i < \theta + 1$ und $\min_i x_i > \theta$. Daraus folgt, dass

$$p_\theta(x) = \mathbb{I}_{(\theta,\infty)} \min_i(x_i) \cdot \mathbb{I}_{(-\infty,\theta+1,)} \max_i(x_i).$$

Und daraus folgt nach dem Faktorisierungstheorem (Satz 2.3), dass $T(X) = (\min_i x_i, \max_i x_i)$ suffizient ist, da $p_\theta(x)$ nur durch $T(X)$ von θ abhängt. (vgl. [Kee10, S.46])

Beispiel 2.7 (Exponentialfamilie)**.** *Insbesondere ist $T = (T_1, \dots, T_k)$ suffizient für $\eta = (\eta_1, \dots, \eta_k)$ aus Definition 2.5 für alle natürlich parametrisierten Exponentialfamilien, da*

$$\begin{aligned} p_\theta(x) &= \tilde{C}(\eta) \cdot exp\left[\sum_{j=1}^{k} \eta_j T_j(x)\right] h(x) \\ &= exp\left[\sum_{j=1}^{k} \eta_j T_j(x) + \log \tilde{C}(\eta)\right] h(x). \end{aligned}$$

Suffiziente Statistiken sind nicht eindeutig:

Korollar 2.2. *Falls T eine suffiziente Statistik für eine Familie von Verteilungen $\mathcal{P}$ ist und $T = H(U)$ gilt, dann ist auch U suffizient. (vgl. [LC98, S.37])*

Beweis. Folgt unmittelbar aus dem Faktorisierungskriterium:

$$p_\theta(x) = \frac{d\mathbf{P}_\theta}{d\mu} = g_\theta\left[(H(U(x))\right] h(x) = g_\theta \circ H(U(x))h(x)$$

□

Es ist daher wünschenswert eine suffiziente Statistik kleinstmöglicher Dimension zu finden. D.h. diese bringt in gewissem Sinn die größtmögliche Reduktion der Daten:

Definition 2.8 (Minimalsuffizienz)**.** *Eine Statistik T ist minimalsuffizient, falls T suffizient ist und für alle suffizienten Statistiken U eine Funktion H existiert, sodass $T = H(U)$ (fast überall auf $\mathcal{P}$). (vgl. [Kee10, S.46])*

Minimal suffiziente Statistiken existieren auf jeden Fall für $X \in \mathbb{R}^n$ und bei gleichzeitigem Vorliegen von Dichten (vgl. [LC98, S.37]).

Satz 2.4. *Gegeben sei eine endliche Familie $\mathcal{P}$ von $k+1$ Verteilungen auf demselben Träger mit Dichten $p_i,\ i = 1, \dots, k$. Dann ist*

$$T(X) = \left(\frac{p_1(X)}{p_0(X)}, \frac{p_2(X)}{p_0(X)}, \dots, \frac{p_k(X)}{p_0(X)}\right)$$

minimal suffizient. (siehe [LC98, S.37])

Beweis. Sei U eine beliebige suffiziente Statistik, dann ist $\frac{p_i(X)}{p_0(X)} = H_i(U)$ eine Funktion von U nach Korollar 2.1. Daher gilt:

$$T(X) = (H_1(U), \dots, H_k(U))$$

und T ist minimal suffizient. (vgl. [Fel14, S.29]) □

Dieser Satz lässt sich unmittelbar auf abzählbare Familien erweitern:

Lemma 2.1. *Sei $\mathcal{P}$ eine (abzählbare) Familie von Verteilungen auf demselben Träger und $\mathcal{P}_0 \subset \mathcal{P}$. Dann gilt: Wenn T minmal suffizient für $\mathcal{P}_0$ und suffizient für $\mathcal{P}$ ist, dann ist T minimal suffizient für $\mathcal{P}$. (vgl. [LC98, S.38])*

Beweis. Wenn eine Statistik U suffizient für $\mathcal{P}$ ist, ist sie auch suffizient für $\mathcal{P}_0$, und da T eine Funktion von U ist, folgt die Behauptung. (vgl. [Fel14, S.29]) □

Der Satz 2.4 angewandt auf exponentielle Familien ergibt:

Satz 2.5. *Sei X verteilt mit Dichte einer regulären exponentiellen Familie (siehe Definition 2.6). Die Statistiken $T_j,\ j = 1, \dots, k$ sind dann minimal suffizient für den natürlichen Parametervektor $\eta = (\eta_1, \dots, \eta_k)$. (vgl. [LC98, S.39])*

Beweis. Aus Satz 2.3 folgt, dass die T_j suffizient sind. Sei $\mathcal{P}_0$ eine Teilfamilie mit $k+1$ Verteilungen mit $\eta^{(j)} = (\eta_1^{(j)}, \dots, \eta_k^{(j)})$. Nach Satz 2.4 sind die Dichtequotienten minimal suffizient, und äquivalent zu

$$\sum_i (\eta_i^{(1)} - \eta_i^{(0)}) T_i(X), \dots, \sum_i (\eta_i^{(k)} - \eta_i^{(0)}) T_i(X). \tag{2.8}$$

Der natürliche Parameterraum ist nach Satz 2.1 eine konvexe Teilmenge des $\mathbb{R}^k$, und daher kann die Matrix

$$(\eta^{(1)} - \eta^{(0)}; \dots; \eta^{(k)} - \eta^{(0)})$$

als regulär angenommen werden. Da die $T_j,\ j = 1, \dots, k$ linear unabhängig sind, ist (2.8) äquivalent zu $T = (T_1(X), \dots, T_k(X))$. Dieser Vektor ist für $\mathcal{P}_0$ minimal suffizient und und wegen Lemma (2.1) auch minimal suffizient für $\mathcal{P}$. (vgl. [LC98, S.39] und [Fel14, S.29f]) □

Beispiel 2.8 (Normalverteilung)**.** *Seien $X_1, \dots, X_n$ iid mit $N(\mu, \sigma^2)$, dann ergibt deren gemeinsame Dichte*

$$\begin{aligned} p_\theta(x_1, \dots, x_n) &= \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/(2\sigma^2)} \right] \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left[\frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 - n\left(\frac{\mu^2}{2\sigma^2} + \log\sigma \right) \right] \end{aligned}$$

mit $\theta = (\mu, \sigma^2)$. Das ist eine zwei-parametrige exponentielle Familie mit natürlicher Parametrisierung, wobei $T_1(x) = \sum_{i=1}^n x_i,\ T_2(x) = \sum_{i=1}^n x_i^2,\ \eta_1(\theta) = \mu/\sigma^2,\ \eta_2(\theta) =$

$-1/(2\sigma^2)$, $\log C(\theta) = -n[\mu^2/(2\sigma^2) + \log\sigma]$ *und* $h(x) = 1/(2\pi)^{n/2}$.
D.h. $T = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)$ *ist minimalsuffizient für* $\eta = (\mu/\sigma^2, -1/(2\sigma^2))$

Bemerkung: Seien $X_1, \ldots, X_n$ iid, jeweils mit Dichte einer regulären exponentiellen Familie (REF) (siehe Definition 2.6. Die gemeinsame Dichte der $X_i s$ bildet dann wieder eine REF mit $T = (T_1^*, \ldots, T_s^*)$ wobei die $T_i^* = \sum_{j=1}^n T_i(X_j)$ sind. D.h. in einer Stichprobe aus einer REF können die Daten zu einer s-dimensionalen Statistik reduziert werden, unabhängig von der Stichprobengröße. Dabei bezeichnet man T als stetige s-dimensionale suffiziente Statistik über dem Euklidischen-Stichprobenraum $\mathcal{X}$, wenn unter den Voraussetzungen des Faktorisierungssatzes 2.3 $T(X) = (T_1(x), \ldots, T_s(x))$ stetig ist und Satz 2.3 $\forall x \in \mathcal{X}$ (nicht f.ü.!) gilt (vgl. [LC98, S.40]). Die Umkehrung ergibt den klassischen Satz von (Fisher-)Darmois-Koopman-Pitman:

Satz 2.6 (Darmois-Koopman-Pitman-Theorem)**.** *Seien $X_1, \ldots, X_n$ reellwertig unanhängig identisch verteilt (iid) mit einer Dichte $f_\theta(x_i)$, die vom Lebesgue-Maß μ dominiert wird, stetig in x_i ist und deren Träger für alle θ ein Interval I ist. Angenommen die gemeinsame Verteilung der $X = (X_1, \ldots, X_n)$*

$$p_\theta(x) = f_\theta(x_1) \cdots f_\theta(x_n) \tag{2.9}$$

besitzt eine stetige k-dimensionale suffiziente Statistik, dann gilt:

(i) Wenn $k = 1$, dann gibt es Funktionen Q_j, C und h, die die Gleichung (2.1) erfüllen, also eine exponentielle Familie bilden.

(ii) Für $k > 1$, und wenn die Dichten $f_\theta(x_i)$ stetige partielle Ableitungen nach x_i besitzen, dann gibt es Funktionen Q_j, C und h, die die Gleichung (2.1) erfüllen mit $s \leq k$.

(vgl. [LC98, S.40] und [BNP68, S.198])

Beweis. siehe [BNP68] □

Bemerkung: In [And70] wird ein Beweis für den diskreten Fall mit $k = 1$ gegeben.

Bemerkung: Dar Satz von Darmois-Koopman-Pitman sagt also, dass unter den „glatten“ absolut stetigen Verteilungsfamilien die exponentiellen Familien, die einzigen sind, die eine Dimensionsreduktion durch Suffizienz ermöglichen. Allerdings muss der Träger der Verteilung unabhängig vom Parameter θ sein.

2.4 Nebensächliche Statistiken und Vollständigkeit

Definition 2.9 (Nebensächliche Statistik)**.** *Eine Statistik $V(X)$ heißt nebensächliche („ancilliary") Statistik , wenn ihre Verteilung nicht vom Parameter θ abhängt. (vgl. [LC98, S.41])*

Bemerkung: Eine nebensächliche Statistik an sich enthält keinerlei Information über θ, aber minimal suffiziente Statistiken können durchaus nebensächliche Information enthalten.(vgl. [LC98, S.41])

Definition 2.10 (Vollständige Statistik)**.** *Eine Statistik T heißt vollständig für eine Familie $\mathcal{P} = \{P_\theta,\ \theta \in \Omega\}$, wenn*

$$\mathbb{E}_\theta h(T) = 0 \quad \forall \theta \in \Omega \quad \Rightarrow h(t) = 0 \quad (f.\ \ddot{u}.\ \mathcal{P})$$

(vgl. [LC98, S.42])

Satz 2.7 (Satz von Basu)**.** *Sei T eine vollständige suffiziente Statistik für $\mathcal{P} = \{P_\theta,\ \theta \in \Omega\}$, dann ist jede nebensächliche Statistik V unabhängig von T. (vgl. [LC98, S.42])*

Beweis. siehe [Kee10, S.50] □

Satz 2.8. *Wenn X nach einer regulären exponentiellen Familie (siehe Definition 2.6) verteilt ist, dann ist $T = (T_1(X), \ldots, T_k(X))$ vollständig.*

Beweis. siehe [LR05, S.117] oder [Bro86, S.43] □

Beispiel 2.9 (Normalverteilung mit bekannter Standardabweichung)**.** *Seien die $X_1, \ldots, X_n$ iid mit $N(\mu, \sigma^2)$ und sei $\mathcal{P}_\sigma = \{N(\mu, \sigma^2)^n : \mu \in \mathbb{R}\}$, d.h. $\mathcal{P}_\sigma$ ist die Familie der Normalverteilungen mit bekannter Varianz. Mit $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ kann die gemeinsame Dichte (Produktdichte) in der Form*

$$\begin{aligned}
p_\mu(x_1, \ldots, x_n) &= \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\sum_{i=1}^n \frac{x_i^2 - 2x_i\mu + \mu^2}{2\sigma^2}\right] \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[\frac{-\sum_{i=1}^n x_i^2 + 2\mu\sum_{i=1}^n x_i - n\mu^2}{2\sigma^2}\right] \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[\frac{n\mu}{\sigma^2}\bar{x} - \frac{n\mu^2}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right]
\end{aligned}$$

geschrieben werden. Dann ist $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *nach Satz 2.8 eine vollständige suffiziente Statistik für* $\mathcal{P}_\sigma$*. Die (korrigierte) Stichprobenvarianz*

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

ist nebensächlich für $\mathcal{P}_\sigma$*, da nach Definition 2.9 ihre Verteilung nicht von* μ *abhängt. Um dies zu zeigen, wählt man* $Y_i = X_i - \mu,\ i = 1, \dots, n$ *und erhält die Verteilung der* Y_i

$$\begin{aligned} p_\mu(Y_i \leq y) = P_\mu(X_i \leq y + \mu) &= \int_{-\infty}^{y+\mu} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \frac{dx}{\sqrt{2\pi\sigma^2}} \\ &= \int_{-\infty}^{y} \exp\left[-\frac{1}{2\sigma^2}u^2\right] \frac{du}{\sqrt{2\pi\sigma^2}}. \end{aligned}$$

Dabei ist der Integrand die Dichte für $N(0, \sigma^2)$*,* $Y_i \sim N(0, \sigma^2)$*. Somit sind die* $Y_1, \dots, Y_n$ *iid mit* $N(0, \sigma^2)$*. Weiters ist* $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{X} - \mu,\ X_i - \bar{X} = Y_i - \bar{Y},\ i = 1, \dots, n$ *und*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Da die gemeinsame Dichte der Y_i *von* σ*, aber nicht von* μ *abhängt, ist* S^2 *nebensächlich für* $\mathcal{P}_\sigma$*. Daher sind nach dem Satz von Basu* $\bar{X}$ *und* S^2 *unabhängig. (vgl. [Kee10, S.50f])*

3 Punktschätzung und Optimalität

„Inferential statistics can be viewed as the science or art of learning about an unknown parameter θ from data X“ [Kee10, S.39]. Der Satz charakterisiert weite Bereiche der Inferenzstatistik bzw. schließenden Statistik, wie Punktschätzung, Intervallschätzung, Hypothesentestung, d.h. also Daten sollen Informationen über ein zugrunde liegendes Modell liefern.

Dieser Abschnitt soll dazu dienen, die Begriffe Maximum-Likelihood-Schätzer, Fisher-Information und deren Bedeutung für exponentielle Familien im Hinblick auf die klassischen Beweise des Satzes von Darmois-Koopman-Pitman in Kapitel 4 zu skizzieren.

Es werden hier nur Schätzverfahren aus dem Blickwinkel von exponentiellen Familien präsentiert. Eine umfassende breit angelegte Einführung in die Schätztheorie ist nicht das Ziel. Dazu sei auf [LC98] verwiesen.

3.1 Punktschätzung

Zu Beginn sollen grundlegende Voraussetzungen besprochen werden.

Eine statistische Entscheidung zur Bestimmung eines Parameters heißt *statistische Schätzung*. Weiters benötigt man eine Entscheidungsfunktion $d(.)$, d.h. eine Statistik vom Stichprobenraum in den Parameterraum. Und zur Beurteilung der Güte der Schätzung wird eine sog. Verlustfunktion (Abstandsmaß, Metrik, ...) herangezogen. (vgl. [Fel14, S.51])

Der Parameter θ einer Familie von Verteilungen $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ und die Daten X sind also durch das X zugrunde liegende Modell resp. Verteilung P_θ verbunden. Ziel ist es also eine Statistik $\delta(X)$ als „gute“ Schätzung für $g(\theta)$ zu finden. $\delta(X)$ wird dann *Schätzer für* $g(\theta)$ genannt, wobei auch $g(\theta) = \theta$ zulässig ist. Zur Beurteilung der Güte der Schätzung benötigt man eine Verlustfunktion $L(\theta, d)$. Diese soll den Verlust einer Schätzung von $g(\theta)$ durch d beschreiben. Dabei soll $L(\theta, d) \geq 0$ für alle θ und d sein. Und für eine exakte Schätzung muss $L(\theta, g(\theta)) = 0$ gelten. (vgl. [Kee10, S.39ff])

Da X eine Zufallsgröße ist, ist auch $L(\theta, \delta(X))$ eine Zufallsgröße, und man definiert den durchschnittlichen Verlust, d.h. den Erwartungswert der Verlustfunktion als Risiko der Schätzung:

Definition 3.1 (Risiko der Schätzung)**.**

$$R(\theta, \delta) := \mathbb{E}_\theta L(\theta, \delta(X))$$

(vgl. [Kee10, S.41])

Definition 3.2 (Erwartungstreuer Schätzer)**.** *Ein Schätzer $\delta(X)$ heißt erwartungstreu oder unverzerrt für $g(\theta)$, wenn der Erwartungswert des Schätzers $\delta(X)$ der zu schätzenden Funktion $g(\theta)$ entspricht, d.h. wenn $\forall\theta \in \Omega$*

$$\mathbb{E}_\theta \delta(X) = g(\theta)$$

erfüllt ist. Falls ein erwartungstreuer Schätzer existiert, so nennt man $g(\theta)$ erwartungstreu schätzbar. (vgl. [Kee10, S.61])

„Die Frage nach der erwartungstreuen Schätzbarkeit von Funktionen g ist i.A. schwer zu beantworten. Für manche natürlich erscheinenden Parameterfunktionen existieren keine, wenige oder auch nur skurrile erwartungstreue Schätzer.“ [Rü14, S.125]

Beispiel 3.1 (Stichprobenmittel)**.** *Seien $X_1, \ldots, X_n$ iid mit $\mathbb{E}[X_i] = \mu,\ \forall i = 1, \ldots, n$, dann ist das Stichprobenmittel $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ erwartungstreu:*

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^n \mu = \frac{1}{n}\cdot n\mu = \mu$$

Beispiel 3.2 (Poissonverteilung)**.** *Sei $\theta \in \Theta = \mathbb{R}^+$ und $P_\theta = \mathcal{P}(\theta)$ die Poisson-Verteilung mit Parameter θ,*

$$P_\theta(\{k\}) = \begin{cases} e^{-\theta}\frac{\theta^k}{k!}, & k \in \mathbb{N}_0 \\ 0, & sonst \end{cases}$$

$\mathcal{P}$ beschreibt das Auftreten von seltenen Ereignissen, z.B. die Anzahl der Besucher einer entlegenen Berghütte pro Woche. Angenommen $g(\theta) = e^{-3\theta}$ sei die Wahrscheinlichkeit, dass in drei voneinander unabhängigen Wochen kein Besucher erscheint und sei $\delta(k)$ mit $k \in \mathbb{N}_0$ ein Schätzer für $g(\theta)$. Falls $\delta(k)$ ein erwartungstreuer Schätzer

für $g(\theta)$ ist, gilt:

$$e^{-3\theta} = \sum_{k=0}^{\infty} d(k) e^{-\theta} \frac{\theta^k}{k!}, \ \theta \in \mathbb{R}^+$$
$$e^{-2\theta} = \sum_{k=0}^{\infty} d(k) \frac{\theta^k}{k!}$$

Mit $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ auf der linken Seite ergibt sich:

$$\sum_{k=0}^{\infty} (-2)^k \frac{\theta^k}{k!} = \sum_{k=0}^{\infty} d(k) \frac{\theta^k}{k!}$$

Und nach Koeffizientenvergleich erhält man:

$$d(k) = (-2)^k, \ k \in \mathbb{N}_0$$

D.h., nur ein offensichtlich unsinniger Schätzer ist erwartungstreu. (vgl. [Rü14, S.126])

Das Beispiel 3.2 zeigt, dass das Prinzip der Erwartungstreue ein sehr beschränktes Kriterium zur Beurteilung von Schätzern ist. Daher besteht eine Möglichkeit zur Verbesserung, erwartungstreue Schätzer auf „optimale“ erwartungstreue Schätzer einzuschränken. Zur Beurteilung der Optimalität wird das Risiko unter einer Verlustfunktion herangezogen.

Bei quadratischer Verlustfunktion wird das Risiko

$$R(\theta, \delta) = \mathbb{E}_\theta(\delta(X) - g(\theta))^2 = MSE(\delta) \tag{3.1}$$

mittlerer quadratischer Fehler *„Mean Squared Error“* (MSE) genannt (vgl. [Fel14, S.58]). Falls δ erwartungstreu ist, gilt

$$R(\theta, \delta) = \mathbb{E}_\theta(\delta(X) - g(\theta))^2 = var_\theta(\delta(X)).$$

D.h. in diesem Fall ist die Varianz zu minimieren (vgl. [Kee10, S.62]). Und dies führt zu folgender Definition:

Definition 3.3 (UMVU-Schätzer)**.** *Ein erwartungstreuer Schätzer δ heißt Schätzer mit überall minimaler Varianz bei Unverzerrtheit („uniformly minimum variance unbiased“, UMVU), wenn*

$$var_\theta(\delta) \leq var_\theta(\delta^*) \quad \forall \theta \in \Omega$$

für alle erwarungstreuen Schätzer δ^* gilt. (vgl. [Kee10, S.62])

Definition 3.4 (LMVU-Schätzer)**.** *Ein erwartungstreuer Schätzer δ heißt Schätzer mit lokaler minimaler Varianz bei Unverzerrtheit („locally minimum variance unbiased", LMVU) an der Stelle $\theta_0 = \theta$, wenn*

$$var_{\theta_0}(\delta) \leq var_{\theta_0}(\delta^*)$$

für alle erwarungstreuen Schätzer δ^* gilt. (vgl. [LC98, S.85])

Im Allgemeinen muss kein UMVU-Schätzer existieren. Aber unter den Voraussetzungen, dass $g(\theta)$ erwartungstreu schätzbar und $T(X)$ eine vollständige suffiziente Statistik ist, gibt es einen eindeutigen (UMVU)-Schätzer basierend auf T. (siehe Satz 3.4 unten)

Bemerkung: Die Risikofunktion einer Schätzung bleibt nur für Schätzungen mit endlichem zweiten Moment endlich (vgl. [Fel14, S.60]). Im Weiteren wird endliche Varianz für Schätzer verlangt.

Bemerkung: Eine Verallgemeinerung von quadratischen zu konvexen Verlustfunktionen, d.h. zu Verlustfunktionen $L(\theta, d)$, die konvex in d sind, ist für weite Teile der Theorie ohne größeren Aufwand möglich. (vgl. [LC98, S.7])

Definition 3.5 (Konvexität)**.** *Eine reellwertige Funktion f auf einer konvexen Menge $\mathcal{C} \in \mathbb{R}^k$ ist konvex, wenn für alle $x, y \in \mathcal{C}$ mit $x \neq y$ und für alle $\gamma \in (0, 1)$*

$$f[\gamma x - (1-\gamma)y] \leq \gamma f(x) - (1-\gamma)f(y) \tag{3.2}$$

gilt. Falls in (3.2) echt kleiner gilt, dann wird f streng konvex genannt. (siehe [Kee10, S.51])

Satz 3.1. *Sei f eine konvexe Funktion auf einem offenen Intervall $\mathcal{C}$ und $t \in \mathcal{C}$ beliebig, dann existiert eine Konstante $c = c_t$ sodass*

$$f(t) + c(x - t) \leq f(x), \quad \forall x \in \mathcal{C}. \tag{3.3}$$

Falls f streng konvex ist, gilt $f(t) + c(x - t) < f(x), \quad \forall x \in \mathcal{C}, x \neq t$. (siehe [Kee10, S.51])

Bemerkung: Geometrisch definiert $g(x) := f(t) + c(x - t)$ eine Gerade unterhalb der Funktion f im Punkt $(t, f(t))$, d.h. $g(x) \leq f(x), \forall x \in \mathcal{C}$.

Die Definition einer konvexen Funktion f bedeutet, dass jeder Funktionswert zwischen zwei Punkten von f nicht größer als der gewichte Durchschnitt der beiden Funktionswerte der zwei Punkte sein kann. Erweitert man dies auf einen gewichteten Durchschnitt einer unendlichen Menge von Punkten, so erhält man die sog. Jensen'sche Ungleichung (in maßtheoretischer Version). (vgl. [LC98, S.46])

Satz 3.2 (Jensen'sche Ungleichung)**.** *Für eine konvexe Funktion auf einem offenen Intervall C und einer Zufallsvariablen X mit $P(X \in \mathcal{C}) = 1$ und $\mathbb{E}(X) < \infty$ gilt*

$$f[\mathbb{E}(X)] \leq \mathbb{E}[f(X)] \tag{3.4}$$

Falls f streng konvex ist, gilt die Ungleichung sogar im strengen Sinn, außer X ist (fast sicher) konstant. (siehe [Kee10, S.52])

Beweis. Nach Satz 3.1 mit $T = \mathbb{E}(X)$ gilt

$$f(\mathbb{E}(X)) + c(x - \mathbb{E}(X)) \leq f(x), \quad \forall x \in \mathcal{C},$$

und daher

$$f(\mathbb{E}(X)) + c(X - \mathbb{E}(X)) \leq f(X), \quad \text{f.ü. [P]}.$$

Und der Erwartungswert beider Seiten ergibt den ersten Teil des Satzes. Wenn f streng konvex ist, dann ist die vorige Ungleichung streng im Fall $X \neq \mathbb{E}(X)$ (f.ü. [P]), da für $f \geq 0$ und $\int f d\mu = 0$ $f = 0$ (f.ü. $[\mu]$) folgt (siehe [Kee10, S.6 und S.52]). □

Bermerkung: Die Jensen'sche Ungleichung gilt auch in höheren Dimensionen mit X als Zufallsvektor (vgl. [Kee10, S.53]).

Bei Vorliegen einer suffizienten Statistik T und einer konvexen Verlustfunktion sagt folgender Satz, dass es für einen Schätzer, der nicht von T abhängt, einen „besseren" Schätzer gibt, der nur von T abhängt. (vgl. [LC98, S.47]).

Satz 3.3 (Rao-Blackwell)**.** *Sei X eine Zufallsvariable mit Verteilung $P_0 \in \mathcal{P}$ aus einer Familie von Verteilungen $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, und sei T eine suffiziente Statistik für θ. Weiters sei δ ein Schätzer für $g(\theta)$ unter der in d konvexen Verlustfunktion $L(\theta, d)$ mit endlicher Erwartung und endlichem Risiko,*

$$R(\theta, \delta) = \mathbb{E}L[\theta, \delta(X)] < \infty,$$

und wenn

$$\eta(T) := \mathbb{E}[\delta(X)|T],$$

dann erfüllt das Risiko vom Schätzer $\eta(T)$

$$R(\theta, \eta) \leq R(\theta, \delta).$$

Wenn L streng konvex ist, dann gilt obige Ungleichung ebenso streng, außer $\delta(X) = \eta(T)$ mit Wahrscheinlichkeit 1 (f.ü. $[P_\theta]$). (vgl. [LC98, S.47]).

Beweis. Die Anwendung der Jensen'schen Ungleichung auf die bedingten Erwartungswerte von $\delta(X)$ bei gegebenen T bzw. $L(\theta, \delta(X)|T)$ ergibt

$$L(\theta, \eta(T)) \leq \mathbb{E}[L(\theta, \delta(X))|T]$$

Die Bildung der Erwartungswerte auf beiden Seiten ergibt

$$R(\theta, \eta) \leq R(\theta, \delta).$$

(vgl. [LC98, S.48]) □

Bemerkung: Die Suffizienz von T stellt in obigem Beweis nur sicher, dass $\eta(T)$ nicht von θ abhängt (vgl. [LC98, S.48]). Weiters ist aus dem Satz von Rao-Blackwell ersichtlich, dass es sich bei einer einzigen erwartungstreuen Schätzung für $g(x)$, die nur von der suffizienten Statistik T abhängt, um die beste handeln muss. Dies erhält man, wenn von T zusätzlich Vollständigkeit verlangt wird (vgl. [LC98, S.87] und (vgl. [Fel14, S.60])):

Satz 3.4 (Lehmann-Scheffé)**.** *Sei X eine Zufallsvariable mit einer Verteilung aus einer Familie von Verteilungen $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, sei weiters T eine suffiziente und vollständige Statistik und es existiere eine erwartungstreue Schätzung δ für $g(\theta)$. Dann gilt:*

a) Es existiert ein erwartungstreuer Schätzer für $g(\theta)$ mit gleichmäßigem minimalem Risiko bei einer in d konvexen Verlustfunktion $L(\theta, d)$, nämlich:

$$\eta(T) = \mathbb{E}[\delta(X)|T]$$

Im Speziellen (bei quadratischer Verlustfunktion) ist dies der UMVU-Schätzer.

b) η aus a) ist der einzige erwartungstreue Schätzer, der nur von T abhängt. Er ist der einzige erwartungstreue Schätzer mit minimalem Risiko, wenn das Risiko von δ endlich ist und wenn die Verlustfunktion L streng konvex ist.

(vgl. [LC98, S.88]

Beweis. Nach einer Eigenschaft des bedingten Erwartungswerts (vgl. [Kus14, S.229 Satz 14.8]) gilt $\mathbb{E}[\eta(T)] = \mathbb{E}[\mathbb{E}[\delta(X)|T]] = \mathbb{E}[\delta(X)]$. Da δ nach Voraussetzung erwartungstreu ist, ist daher auch η erwartungstreu.

Angenommen es gibt einen weiteren erwartungstreuen Schätzer $\eta'(T)$, der nur von T abhängt, dann gilt

$$\mathbb{E}_\theta(\eta(T) - \eta'(T)) = 0 \quad \forall \theta \in \Omega.$$

Da T nach Voraussetzung vollständig ist, gilt $\eta(T) = \eta'(T)$ fast sicher. Daher ist η (f.s.) der einzige erwartungstreue Schätzer, der nur von T abhängt, und jeder andere kann durch Bildung der bedingten Erwartung verbessert werden.

Der letzte Punkt folgt direkt aus Satz 3.3, da bei streng konvexer Verlustfunktion jeder von η verschiedene Schätzer ein höheres Risiko hätte. (vgl. [Fel14, S.61]) □

Korollar 3.1. *Wenn $\mathcal{P}$ eine reguläre exponentielle Familie ist, dann gilt Satz 3.4 für $T = (T_1, \dots, T_k)$ und der UMVU-Schätzer ist $\eta(T) = \mathbb{E}[\delta(X)|T]$.*

Beweis. folgt unmittelbar aus Satz 2.8. (vgl. [Fel14, S.61]) □

Bemerkung: Der Satz 3.4 und das daraus folgende Korollar stellen für viele Klassen von Verteilungsfamilien beste erwartungstreue Schätzer bereit. Obwohl die Verlustfunktion eine beliebige konvexe Funktionen sein kann, spricht man trotzdem meist von einen UMVU-Schätzer. In dem Fall, wo ein erwartungstreuer Schätzer δ für $g(\theta)$ vorliegt, der nur von einer vollständigen suffizienten Statistik abhängt, ist dies nach obigem Satz der UMVU-Schätzer. (vgl. [LC98, S.88])

Beispiel 3.3 (Normalverteilung)**.** *Seien $X_1, \dots, X_n$ iid bezüglich $N(\xi, \sigma^2)$ und es gilt die Varianz zu schätzen. Der standardmäßige erwartungstreue Schätzer ist dann*

$$\delta = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}) \quad \textit{mit} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Da dieser eine Funktion der vollständig suffizienten Statistik

$$T = (\sum_i X_i, \sum_i (X_i - \bar{X})^2)$$

ist, folgt nach Satz 3.4, dass δ der UMVU-Schätzer ist. (vgl. [LC98, S.88])

Neben dem direkten Vorliegen des UMVU-Schätzers, wie im obigen Beispiel, gibt es zwei Strategien, um einen UMVU-Schätzer zu bestimmen (vgl. [LC98, S.88f]):

(Methode 1) *[Auflösen nach δ]* Ist eine vollständige suffiziente Statistik T gegeben, dann ist der UMVU-Schätzer jeder erwartungstreu schätzbaren Funktion $g(\theta)$ durch Definition 3.2

$$\mathbb{E}_\theta \delta(X) = g(\theta) \quad \forall \theta \in \Omega \tag{3.5}$$

bestimmt.

Beispiel 3.4 (UMVU-Schätzer mit Binomialverteilung)**.** *Sei T binomialverteilt mit der pmf*

$$p_{n,\theta}(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

wobei $n, x \in \mathbb{N},\ 0 \leq x \leq n,\ 0 < \theta < 1$. *Weiters sei* $g(\theta) = \theta(1-\theta)$. *Dann wird (3.5) zu*

$$\sum_{t=0}^{n} \binom{n}{t} \delta(t) \theta^t (1-\theta)^{n-t} = \theta(1-\theta), \quad 0 < \theta < 1 \tag{3.6}$$

Mit $r = \frac{\theta}{1-\theta}$ *ergibt sich* $\theta = \frac{r}{1+r}$ *und* $(1-\theta) = \frac{1}{1+r}$ *und somit lässt sich (3.6) schreiben als*

$$\sum_{t=0}^{n} \binom{n}{t} \delta(t) \frac{r^t}{(1+r)^n} = \frac{r}{(1+r)^2} \tag{3.7}$$

$$\sum_{t=0}^{n} \binom{n}{t} \delta(t) r^t = r(1+r)^{n-2} \tag{3.8}$$

Mit dem binomischen Lehrsatz

$$(r+1)^{n-2} = \sum_{k=0}^{n-2} \binom{n-2}{k} r^k = \sum_{t=1}^{n-1} \binom{n-2}{t-1} r^{t-1}$$

eingesetzt in (3.8) erhält man

$$\sum_{t=0}^{n} \binom{n}{t} \delta(t) r^t = \sum_{t=1}^{n-1} \binom{n-2}{t-1} r^t, \quad 0 < r < \infty.$$

Und Koeffizientenvergleich ergibt

$$\delta(t) = \frac{t(n-t)}{n(n-1)}, \quad 0 < t < n$$

als UMVU-Schätzer. (vgl. [LC98, S.88f])

(Methode 2) *[Bedingen]* Wenn eine beliebige erwartungstreue Schätzung $\delta(X)$ vorhanden ist, folgt aus dem Satz 3.4, dass der UMVU-Schätzer $\delta'(X)$ durch die bedingte Erwartung von $\delta(X)$ unter der Bedingung T berechnet werden kann, d.h. $\delta'(X) = E[\delta(X)|T]$. Eine günstige Wahl des erwartungstreuen Schätzers $\delta(X)$ kann dabei die Berechnung erleichtern.

Beispiel 3.5 (UMVU-Schätzer mit Gleichverteilung)**.** *Seien* $X_1, \ldots, X_n$ *mit Gleichverteilung* $U(0, \theta)$ *iid und sei* $g(\theta) = \theta/2$. *Das Maximum der* X_i *ist eine vollständige suffiziente Statistik, also* $T = X_{(n)}$.
Der Erwartungswert für X_i *ist:*

$$\mathbb{E}(X_i) = \int_0^\theta x \frac{1}{\theta} dx = \frac{1}{\theta} \frac{x^2}{2} \Big|_0^\theta = \theta/2$$

Da eine suffiziente und vollständige Statistik vorliegt, sowie eine erwartungstreue Schätzung für $g(\theta) = \theta/2$ *existiert, ist nach Satz 3.4 der UMVU-Schätzer von* $g(\theta) = \theta/2$ *durch* $\mathbb{E}[X_1|X_{(n)} = t]$ *zu finden. Wenn also* $X_{(n)} = t$ *ist, dann nimmt* X_i *mit Wahrscheinlichkeit* $1/n$ *den Wert* t *an und* X_i *ist gleichverteilt auf* $(0, t)$ *mit Wahrscheinlichkeit* $(n-1)/n$*. Daraus ergibt sich*

$$\mathbb{E}[X_i|t] = \frac{1}{n} \cdot t + \frac{n-1}{n} \cdot \frac{t}{2} = \frac{n+1}{n} \cdot \frac{t}{2}. \tag{3.9}$$

Der UMVU-Schätzer für $g(\theta) = \theta/2$ *ist daher* $[(n+1)/n] \cdot T/2$*, bzw.* $[(n+1)/n] \cdot T$ *für* $g(\theta) = \theta$*. (vgl. [LC98, S.89])*

3.2 Fisher Information

Die Größe bzw. das Ausmaß des Risikos einer Schätzung ist ein weiteres wichtiges Beurteilungskriterium einer Parameterschätzung. Das Risiko (Definition 3.1) unter einer quadratischem Verlustfunktion ist die Varianz $V_L(\theta)$ für alle θ. Allerdings lässt sich die Varianz $V_L(\theta)$ oftmals nur schwer direkt bestimmen. Insbesondere dann, wenn keine „optimalen“ Schätzer, wie z.B. ein UMVU-Schätzer, existieren.

Existiert z.B. kein UMVU-Schätzer (*„uniformly minimum variance unbiased“* siehe Definition 3.3), aber ein LMVU-Schätzer (*„locally minimum variance unbiased“* siehe Definition 3.4) an der Stelle θ_0, so wäre dessen Varianz $V_L(\theta_0)$ die kleinste Varianz, die ein erwartungstreuer Schätzer an der Stelle θ_0 erreichen kann. Somit wäre das eine mögliche untere Schranke für eine erwartungstreue Schätzung.

Es werden also untere Schranken gesucht, die von gewissen Klassen von Schätzern nicht unterschritten werden. Die Güte der Schätzung kann durch Erreichen der Schranke bzw. durch den Abstand zwischen Schätzer und Schranke beurteilt werden. (vgl. [LC98, S.113] und [Fel14, S.64f]). In weiterer Folge wird die Informationsungleichung (Cramér-Rao-Ungleichung) hergeleitet und deren Untergrenze bildet die sog. Cramér-Rao-Schranke.

Definition 3.6 (Regularitätsbedingungen)**.** *Es sei* θ *ein unbekannter Parameter einer Familie von Verteilungen* $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ *und* $p_\theta(x)$ *die zugehörige Dichte (bezüglich eines* σ*-endlichen Maßes) von* X*.*

(RB1) Ω *ist ein offenes Intervall.*

(RB2) Die Verteilungen P_θ *haben einen gemeinsamen Träger unabhängig von* θ*, d.h. oBdA ist die Menge* $A = \{x : p_\theta(x) > 0\}$ *unabhängig von* θ*.*

(RB3) $\forall x \in A$ *und* $\forall \theta \in \Omega$ *existiert die Ableitung*

$$p'_\theta(x) = \frac{\partial}{\partial \theta} p_\theta(x) < \infty$$

und ist stetig.

(RB4) Die Reihenfolge von Ableitung und Integration nach θ *darf getauscht werden:*

$$\frac{\partial}{\partial \theta} \int d(x) p_\theta(x) dx = \int d(x) \frac{\partial}{\partial \theta} p_\theta(x) dx$$

wobei $d(.)$ *eine integrierbare Funktion ist.*

(RB5) Für die stochastische Größe Ψ *(Score-Funktion)*

$$\Psi(X, \theta) := \frac{\partial}{\partial \theta} \log p_\theta(X) \tag{3.10}$$

gilt, dass $\forall \theta$ *die Funktion* Ψ *ein endliches zweites Moment*

$$\frac{\partial^2}{\partial \theta^2} \log p_\theta(x)$$

besitzt, stetig ist und

$$\frac{\partial^2}{\partial \theta^2} \int d(x) \log p_\theta(x) dx = \int d(x) \frac{\partial^2}{\partial \theta^2} \log p_\theta(x) dx$$

für alle integrierbaren Funktionen $d(.)$ *erfüllt.*

(vgl. [LC98, S.115f] und [Rü14, S.156f])

Bemerkung: Eine Familie von Verteilungen $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ wird von manchen Autoren auch regulär genannt, falls Bedingungen (RB1)-(RB5) erfüllt sind ([Rü14, S.157]). In weiterer Folge wird hier darauf verzichtet, um nicht in Konflikt mit regulären exponentiellen Familien (REF) (vgl. Definition 2.6) zu gelangen.

Bemerkung: Eine reguläre exponentielle Familie nach Definition 2.6 erfüllt die Bedingungen (RB1)-(RB5) (siehe [Fel14, S.]).

Bemerkung: *„Die Score-Funktion* $\Psi(X, \theta) = \frac{\partial}{\partial \theta} \log p_\theta(X)$ *beschreibt die Änderung der Dichte in Abhängigkeit von der Änderung des Parameters. Ist diese lokale Änderung groß, dann erhält man in Konsequenz aus einer Beobachtung viel „Information" über den Parameter. Dieses motiviert den Begriff der Fisher-Information."* [Rü14, S.157]

Definition 3.7 (Fisher-Information)**.** *Der Erwartungswert von* Ψ^2

$$I(\theta) := \mathbb{E}_\theta \Psi^2(X,\theta) = \mathbb{E}_\theta \left[\frac{\partial \log p_\theta(x)}{\partial \theta}\right]^2 = \\ = \int \left(\frac{\partial \log p_\theta(x)}{\partial \theta}\right)^2 p_\theta(x) d\mu(x) = \int \left(\frac{\partial p_\theta(x)}{\partial \theta}\right)^2 \frac{1}{p_\theta(x)} d\mu(x)$$

wird als Fisher-Information des Parameters θ *von* P_θ *bezeichnet. (vgl. [Rü14, S.157] und [Fel14, S.66])*

Bemerkung: $I(\theta)$ hängt von der jeweiligen Parametrisierung ab (vgl. [LC98, S.115]). Sei $\theta = h(\xi)$, wobei h differenzierbar vorausgesetzt wird, dann ergibt sich für $\tilde{I}(\xi)$ mit Hilfe von $h'(\xi) = \partial\theta/\partial\xi$

$$\tilde{I}(\xi) = \tilde{\mathbb{E}}_\xi \left[\frac{\partial \log p_{h(\xi)}(x)}{\partial \xi}\right]^2 = \tilde{\mathbb{E}}_\xi \left[\frac{\partial \log p_{h(\xi)}(x)}{\partial \xi}\right]^2 = \left[h'(\xi)\right]^2 \mathbb{E}_\theta \left[\frac{\partial \log p_\theta(x)}{\partial \theta}\right]^2 = \\ = \left[h'(\xi)\right]^2 I(\theta) \tag{3.11}$$

(vgl. [Kee10, S.74])

Lemma 3.1. *Falls* $\mathcal{P}$ *die Bedingungen (RB1)-(RB5) erfüllt, dann gilt für die Score-Funktion* $\Psi(X,\theta)$ *(siehe Definition 3.6 Zeile (3.10)) bzw. für die Fisher-Information* $I(\theta)$ *(siehe Definition 3.7) :*

$$\mathbb{E}_\theta \Psi(X,\theta) = 0 \tag{3.12}$$

und

$$I(\theta) = var_\theta \Psi(X,\theta) \tag{3.13}$$

sowie

$$I(\theta) = -\mathbb{E}_\theta \frac{\partial^2}{\partial \theta^2} \log p_\theta(X) \tag{3.14}$$

(vgl. [Rü14, S.157] und [Fel14, S.66])

Beweis. Da nach (RB4)

$$\int \frac{\partial p_\theta(x)}{\partial \theta} d\mu(x) = \frac{\partial}{\partial \theta} \int p_\theta(x) d\mu(x) = \frac{\partial}{\partial \theta} 1 = 0$$

gilt, folgt

$$\mathbb{E}_\theta \Psi(X,\theta) = \int \frac{\partial p_\theta(x)}{\partial \theta} \frac{1}{p_\theta(x)} p_\theta(x) d\mu(x) = 0$$

Daraus folgt sofort, dass $I(\theta) = var_\theta \Psi(X,\theta)$, da $V(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
Die Ableitung der Score-Funktion ergibt

$$\frac{\partial^2}{\partial \theta^2} \log p_\theta(x) = \frac{1}{p_\theta(x)} \left(\frac{\partial^2}{\partial \theta^2} p_\theta(x)\right) - \frac{1}{p_\theta(x)^2} \left(\frac{\partial p_\theta(x)}{\partial \theta}\right)^2.$$

Nach Integration resp. durch Bildung der Erwartungswerte auf beiden Seiten ist

$$\int \frac{\partial^2}{\partial\theta^2} \log p_\theta(x) p_\theta(x) d\mu(x) = \int \left(\frac{\partial^2}{\partial\theta^2} p_\theta(x) \right) d\mu(x) - \int \frac{1}{p_\theta(x)} \left(\frac{\partial p_\theta(x)}{\partial\theta} \right)^2 d\mu(x).$$

Und da wegen der Vertauschbarkeit von zweiter Ableitung und Integration (RB5)

$$\int \left(\frac{\partial^2}{\partial\theta^2} p_\theta(x) \right) d\mu(x) = \frac{\partial^2}{\partial\theta^2} \left(\int p_\theta(x) d\mu(x) \right) = \frac{\partial^2}{\partial\theta^2} (1) = 0$$

ist, folgt die Behauptung. (vgl. [Rü14, S.157] und [Fel14, S.66]) □

Beispiel 3.6 (Exponentialfamilie). *Sei $\mathcal{P} = \{P_\eta : \eta \in \Omega\}$ eine einparametrige reguläre Exponentialfamilie (siehe Definition 2.5 und 2.6) mit $\tilde{C}(\eta) = \exp[-A(\eta)]$, d.h.*

$$p_\eta(x) = exp\left[\eta T(x) - A(\eta)\right] h(x) \tag{3.15}$$

dann ist

$$\frac{\partial}{\partial\eta} \ln p_\eta(x) = T(x) - \frac{\partial}{\partial\eta} A(\eta) = T(x) - A'(\eta)$$
$$\frac{\partial^2}{\partial\eta^2} \ln p_\eta(x) = -\frac{\partial^2}{\partial\eta^2} A(\eta) = -A''(\eta)$$

Daraus folgt nach Lemma 3.1, dass

$$I(\eta) = \mathbb{E}[T(x) - A'(\eta)]^2 \underset{(3.13)}{=} var_\eta(T(x) - A'(\eta)) \underset{(3.12)}{=} var_\eta(T(x)) \underset{(3.14)}{=} A''(\eta) \tag{3.16}$$

Bei sog. Mittelwertsparametrisierung, d.h. $\tau(\eta) := A'(\eta) = \mathbb{E}\eta T$ (vgl. (2.3)), ergibt sich durch (3.11) und (3.16), dass

$$A''(\eta) = \left[A''(\eta)\right]^2 I(\tau(\eta))$$

ist. Mit

$$var_\eta(T) = A''(\eta)$$

aus (3.16) folgt daher

$$I(\tau(\eta)) = \frac{1}{var_\eta(T)} \tag{3.17}$$

(vgl. [Kee10, S.74])

Satz 3.5. *Seien X und Y unabhängig verteilt mit Dichten p_θ und q_θ, bezüglich σ-additiver Maße μ und ν, die die Regularitätsbedingungen (RB1)-(RB3) erfüllen. Weiters bezeichne $I_X(\theta)$, $I_Y(\theta)$ und $I_{(X,Y)}(\theta)$ die Fisher-Information von X, Y und (X,Y), dann gilt:*

$$I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta).$$

(vgl. [LC98, S.119])

Beweis. Die Fisher-Information der zusammengesetzten Größe (X, Y) ist

$$I_{X,Y}(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log p_\theta(X, Y) \right)^2$$

Aufgrund der Unabhängigkeit gilt

$$p_\theta(X, Y) = p_\theta(X) q_\theta(Y) \quad \text{bzw.} \quad \log p_\theta(X, Y) = \log p_\theta(X) + \log q_\theta(Y)$$

und somit

$$\begin{aligned} I_{X,Y}(\theta) &= \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log p_\theta(X) + \frac{\partial}{\partial \theta} \log q_\theta(Y) \right)^2 \\ &= I_X(\theta) + 2\mathbb{E}_\theta \Psi(X, \theta) \mathbb{E}_\theta \Psi(Y, \theta) + I_Y(\theta) \end{aligned}$$

Und nach Lemma 3.12 ist $\mathbb{E}_\theta \Psi(X, \theta) = 0$, daher folgt die Behauptung. (vgl. [Kee10, S.75]) □

Korollar 3.2. *Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt, ihre Dichten $p_\theta(X_i)$ erfüllen (RB1)-(RB5) und jedes X_i hat die Fisher-Information $I(\theta)$, dann ist die Fisher-Information von $X = (X_1, \ldots, X_n)$ gleich $nI(\theta)$. (vgl. [LC98, S.119])*

Satz 3.6 (Informationsungleichung)**.** *Seien $X_1, \ldots, X_n$ unabhängig identisch verteilt mit Dichten $p_\theta(x)$ aus einer Familie $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ mit dominierendem Maß μ, die die Regularitätsbedingungen (RB1)-(RB5) erfüllen und mit $I(\theta) > 0$. Sei δ eine Statistik mit $\mathbb{E}_\theta(\delta^2) < \infty$ für die $\frac{\partial}{\partial \theta} \mathbb{E}_\theta(\delta)$ existiert und Ableitung und Integral folgendermaßen vertauscht werden können:*

$$\frac{\partial}{\partial \theta} \mathbb{E}_\theta(\delta) = \int \frac{\partial}{\partial \theta} \delta \; p_\theta \; d\mu.$$

Dann gilt:

$$var_\theta \delta \geq \frac{\left[\frac{\partial}{\partial \theta} \mathbb{E}_\theta(\delta) \right]^2}{I(\theta)} \tag{3.18}$$

$I(\theta)$ . . . Fisher-Information (siehe Definition 3.7)
(vgl. [LC98, S.120])

Beweis. Die Kovarianz von $\Psi(X, \delta)$ und $\delta(X)$ ist

$$cov_\theta(\Psi, \delta) = \mathbb{E}_\theta[\Psi\delta] - \mathbb{E}_\theta \Psi \mathbb{E}_\theta \delta = \mathbb{E}_\theta[\Psi\delta],$$

da nach Lemma 3.12 $\mathbb{E}_\theta \Psi = 0$.

$$\begin{aligned}\mathbb{E}_\theta[\Psi\delta] &= \int \delta(x) \left(\frac{\partial}{\partial\theta} \log p_\theta(x) \right) p_\theta(x) d\mu(x) \\ &= \int \delta(x) \left(\frac{\partial}{\partial\theta} p_\theta(x) \right) d\mu(x) \\ &= \frac{\partial}{\partial\theta} \int \delta(x) p_\theta(x) d\mu(x) \\ &= \frac{\partial}{\partial\theta} \mathbb{E}_\theta \delta\end{aligned}$$

Die Anwendung der Schwarz'schen Ungleichung

$$cov_\theta(\Psi, \delta)^2 = [\frac{\partial}{\partial\theta}\mathbb{E}_\theta\delta]^2 \leq var_\theta\Psi var_\theta\delta = I(\theta) var_\theta\delta \tag{3.19}$$

ergibt dann die Behauptung, da nach Lemma 3.12 $I(\theta) = var_\theta\Psi$. (vgl. [Fel14, S.68]) □

Bemerkung: Die Informationsungleichung bzw. Cramér-Rao-Ungleichung liefert eine untere Schranke für die Varianz einer Schätzfunktion δ.

Korollar 3.3. *Seien $X_1, \ldots, X_n$ unabhängig identisch verteilt mit Dichten $p_\theta(x)$ aus einer Familie $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, die die Regularitätsbedingungen (RB1)-(RB5) erfüllen und mit Fisher-Information $0 < I(\theta) < \infty$. Sei δ eine erwartungstreue Schätzung für θ, dann gilt*

$$var_\theta\delta \geq \frac{1}{nI(\theta)} \tag{3.20}$$

(vgl. [HMC12, S.331] und [LC98, S.440])

Definition 3.8 (effizienter Schätzer)**.** *Eine erwartungstreue Schätzfunktion T heißt (CR-)effizient, wenn T die Cramér-Rao-Schranke erreicht, d.h. in (3.20) gilt Gleichheit. Das Verhältnis von der CR-Schranke zur tatsächlichen Varianz eines erwartungstreuen Schätzers wird Effizienz genannt. (vgl. [HMC12, S.332])*

Bemerkung: Selbst UMVU-Schätzer erreichen nicht zwingend die CR-Schranke. Der Beweis der Informationsungleichung lässt die Frage nach der Gestalt von effizienten Schätzern offen. Das Gleichheitszeichen in der Schwarz'schen Ungleichung gilt nur im Fall, wenn $\delta(X)$ und $\Psi(X, \theta)$ linear abhängig sind, d.h. wenn es Funktionen $A(\theta)$ und $B(\theta)$ gibt, sodass

$$\delta(X) = A(\theta)\psi(X, \theta) + B(\theta). \tag{3.21}$$

Diese Form, die unmittelbar zu exponentiellen Familien führt, findet bereits Erwähnung in [Fis34, S.293], allerdings in sehr informeller Weise. (vgl. [LC98,

S.121], [Rü14, S.160] und [Fel14, S.70])

Satz 3.7. *Wenn X der Verteilung einer 1-dimensionalen exponentiellen Familie mit Dichte*

$$p_\theta(x) = C(\theta) \cdot \exp\left[g(\theta)T(x)\right] h(x) \tag{3.22}$$

und Mittelwertparametrisierung

$$\mathbb{E}T = \theta$$

folgt, dann gilt

$$I(\theta) = \frac{1}{var(T)} \tag{3.23}$$

Beweis. Aus

$$\Psi(X,\theta) = T(X)g'(\theta) + \frac{C'(\theta)}{C(\theta)}$$

ergibt sich mit

$$A(\theta) := \frac{1}{g'(\theta)}$$

und

$$B(\theta) := -\frac{C'(\theta)}{C(\theta)g'(\theta)}$$

die Form (3.21). Die Korrelation zwischen Ψ und T ist 1 und in (3.19) gilt Gleichheit. Die Gleichung (3.22) erfüllt die (RB1)-(RB5) aus 3.6. (vgl. [Fel14, S.70]) □

Beispiel 3.7 (Poissonverteilung)**.** *Sei X Poisson-verteilt mit Parameter λ (siehe Beispiel 3.2). Der Erwartungswert der Poissonverteilung ist $\mathbb{E}(X) = \lambda$. Daher ist nach Satz 3.7 die Fisher-Information $I[\lambda] = 1/\lambda$. Für die (monoton steigende) Funktion $\tau(\lambda) = \log\theta$ ist $I[\log\lambda] = \lambda$. Das bedeutet die Information von X über λ ist indirekt proportional zu der über $\log\lambda$. Im Gegensatz zu λ kann also $\log\lambda$ auch für große Werte genau geschätzt werden. Allerdings kehrt sich dieser Effekt bei Werten nahe null um. Interessanterweise gibt es eine Funktion deren Informationsgehalt sozusagen zwischen dem von λ und $\log\lambda$ liegt. Für $\tau(\theta) = \sqrt{\lambda}$ ist der Informationsgehalt unabhängig von λ. (vgl. [LC98, S.118])*

Bemerkung: Eine Erweiterung der Theorie auf mehrere Parameter lässt sich mit der sog. Fisher-Informationsmatrix bzw. der mehrparametrigen Informationsungleichung durchführen. Eine kompakte Darstellung dazu findet sich beispielsweise in [LC98, Kapitel 2.6 S.124ff].

3.3 Maximum Likelihood

Die Maximum-Likelihood-Methode ist eine weitere Möglichkeit einen unbekannten Parameter einer Verteilung zu schätzen. Dabei werden die Daten einer Zufallsstichprobe in die gemeinsame Dichte- bzw. Wahrscheinlichkeitsfunktion eingesetzt, und dadurch erhält man eine Funktion abhängig von Parameter, die Likelihood-Funktion. Der Wert, der diese Funktion maximiert resp. der Wert mit der größten Likelihood, ist dann der Maximum-Likelihood-Schätzer. Es sollte dabei immer von der größten Likelihood bzw. der größten Plausibilität gesprochen werden, und nicht von größter Wahrscheinlichkeit. Der Begriff der Likelihood und das dahinter stehende Konzept war lange Zeit Gegenstand ausführlicher Diskussionen. Siehe dazu [Ald97] und [Pfa17]. Dem Kapitel vorangestellt sei noch ein Zitat von R. A. Fisher, der die Maximum-Likelihood-Methode maßgeblich entwickelt hat.

„What has now appeared is that the mathematical concept of probability is, in most cases, inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term "Likelihood" to designate this quantity.“ [Fis50b, S.10]

Definition 3.9 (Likelihood-Funktion)**.** *Seien $X_1, \dots, X_n$ unabhängig identisch verteilte Zufallsvariablen mit Dichte $p_\theta(x)$ aus einer Familie $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ mit dominierendem Maß μ. Der Parameter θ sei unbekannt. Dann ist die Likelihood-Funktion als Funktion von θ gegeben durch*

$$L(\theta) = \prod_{i=1}^{n} p(\theta, x_i), \quad \theta \in \Omega \tag{3.24}$$

für $x = (x_1, \dots, x_n)$. Die Funktion

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log p(\theta, x_i), \quad \theta \in \Omega \tag{3.25}$$

wird als Log-Likelihood-Funktion bezeichnet. (vgl. [HMC12, S.321])

Definition 3.10 (Maximum-Likelihood-Schätzer)**.**

a) Eine messbare Abbildung $\hat{\theta}(X) : (\mathcal{X}, \mathcal{B}) \to (\Omega, \mathcal{A})$ heißt Maximum-Likelihood-

Schätzer (MLS) bzw. maximum likelihood estimator (MLE), wenn

$$L(\hat{\theta}(X)) = \sup_{\theta \in \Omega} L(\theta) \tag{3.26}$$

$$\Leftrightarrow \quad \log L(\hat{\theta}(X)) = \sup_{\theta \in \Omega} \log L(\theta) \tag{3.27}$$

b) Sei $\Omega \subset \mathbb{R}^k$ offen und $L(\theta)$ partiell differenzierbar, dann erfüllt ein MLS die Likelihood-Gleichungen

$$\frac{\partial}{\partial \theta_j} \log L(\hat{\theta}_j(X)) = 0 \quad j = 1, \ldots, s \tag{3.28}$$

(vgl. [Rü14, S.162])

Beispiel 3.8 (Normalverteilung). *Sei $\Omega = \mathbb{R} \times \mathbb{R}^+$, $\theta = (\mu, \sigma)$ und $P_\theta = \bigotimes_{i=1}^n N(\mu, \sigma^2)$, dann ist für $x \in \mathcal{X} = \mathbb{R}^n$*

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

bzw. für die Log-Likelihood-Funktion

$$\ln L(\theta) = -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

Die partiellen Ableitungen

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)$$

$$\frac{\partial}{\partial (\sigma^2)} \ln L(\mu, \sigma^2) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2$$

führen auf die Likelihood-Gleichungen

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$-\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0.$$

Die Lösungen ergeben

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\hat{\mu})=0$$
$$\sum_{i=1}^{n}x_i-n\hat{\mu}=0$$
$$\hat{\mu}=\bar{x}_n=\frac{1}{n}\sum_{i=1}^{n}x_i$$

und

$$-n+\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2=0 \quad (\sigma^2>0)$$
$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Und da

$$\frac{\partial^2}{\partial\mu^2}\ln L(\mu,\sigma^2)=-\frac{n}{\sigma^2}<0$$
$$\frac{\partial^2}{\partial(\sigma^2)^2}\ln L(\mu,\sigma^2)=\frac{n}{2}\frac{1}{\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\bar{x})^2=-\frac{n}{2}\frac{1}{\hat{\sigma}^4}<0$$

ist $\hat{\theta}=(\hat{\mu},\hat{\sigma}^2)$ der MLS. $\hat{\theta}$ ist suffizient, aber nicht erwartungstreu.
(vgl. [Rü14, S.163])
Anm.: Eine erwartungstreue Schätzung der Varianz wäre die (korrigierte) Stichprobenvarianz $s^2_{n-1}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

Bemerkung: MLS müssen i.A. nicht existieren. Auch eine explizite Berechnung ist nicht immer möglich. Allerdings können dann (Existenz vorausgesetzt) numerische Lösungsverfahren eingesetzt werden.

Der MLS muss nicht immer suffizient sein. Im Fall, dass der MLS suffizient ist, gibt es keine einfacheren suffizienten Statistiken:

Satz 3.8. *Wenn der MLS $\hat{\theta}(X)$ eine suffiziente Statistik hat und eindeutig ist, dann ist er auch minimal suffizient. (vgl. [Fel14, S.58])*

Beweis. Sei T eine beliebige suffiziente Statistik für θ, dann ist die Likelihood-Funktion $L_x(\theta)$ faktorisierbar

$$L_x(\theta)=g[T(x),\theta]h(x).$$

$\hat{\theta}$ maximiert somit $g[T(x), \theta]$ und hängt nur über $T(x)$ von den Daten x ab. Der MLS $\hat{\theta}$ ist eine Funktion von T und nach Voraussetzung eindeutig und daher minimal suffizient. (vgl. [Fel14, S.58]) □

Bemerkung: Wie im Beispiel 3.8 zu sehen ist, muss der ML-Schätzer nicht erwartungstreu sein. Ebensowenig muss der ML-Schätzer eindeutig sein:

Beispiel 3.9 (Cauchy-Verteilung)**.** *Sei X_1, X_2 eine unabhängige Zufallsstichprobe aus einer Cauchy-Verteilung mit Dichte*

$$p(x, \theta) = \frac{1}{\pi(1 + (x - \theta)^2)}$$

wobei $-\infty < x < \infty$ und $-\infty < \theta < \infty$.
Die Likelihood-Funktion bzw. die Log-Likelihood ist dann

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{2} \frac{1}{\pi(1 + (x_i - \theta)^2)} \\ l(\theta) &= \ln L(\theta) = \ln \prod_{i=1}^{2} \frac{1}{\pi(1 + (x_i - \theta)^2)} \\ l(\theta) &= -2 \ln \pi - \sum_{i=1}^{2} \ln(1 + (x_i - \theta)^2) \\ \frac{\partial l(\theta)}{\partial \theta} &= -\sum_{i=1}^{2} \frac{2(x_i - \theta)}{(1 + (x_i - \theta)^2)}. \end{aligned}$$

Und die Likelihood-Gleichung $\frac{\partial l(\theta)}{\partial \theta} = 0$ ergibt oBdA mit $X = (-x, x)$

$$\begin{aligned} 0 &= -\sum_{i=1}^{2} \frac{2(x_i - \theta)}{(1 + (x_i - \theta)^2)} \\ 0 &= \frac{2(x + \theta)}{(1 + (x + \theta)^2)} - \frac{2(x - \theta)}{(1 + (x - \theta)^2)} \\ &\vdots \\ 0 &= 2\theta(\theta^2 - (x^2 - 1)) \end{aligned}$$

Das bedeutet, dass für $|x| \leq 1$ der ML-Schätzer $\hat{\theta} = 0$ ist. Und für $|x| > 1$ ist der ML-Schätzer $\hat{\theta} = \pm\sqrt{x^2 - 1}$ ($\hat{\theta} = 0$ ergibt in diesem Fall ein lokales Minimum, da $\frac{\partial^2 l(\theta)}{\partial \theta^2}(0) > 0$) und somit nicht eindeutig.

4 Satz von Darmois-Koopman-Pitman

4.1 Einleitung

In diesem Abschnitt werden die Inhalte jener drei Publikationen von Darmois [Dar35], Koopman [Koo36] und Pitman [Pit36] analysiert, die in fast allen Standardwerken über exponentielle Familien und suffiziente Statistiken genannt werden. Auch heutige Fachartikel, in denen exponentielle Familien und Suffizienz eine wichtige Rolle spielen, enthalten noch Verweise auf diese drei Artikel, um auf die Ursprünge der Begriffe zu verweisen. So werden die Entdecker ihrer Verdienste gewürdigt.

Auffallend dabei ist, dass zwar diese drei Arbeiten immer wieder zitiert werden, aber nur bis etwa Ende der 1970er-Jahre wird auch auf Inhalte der drei Artikel eingegangen. So fasst zum Beispiel der Artikel [BM63] die Ergebnisse von Darmois, Koopman und Pitman hervorragend zusammen, und das Buch von [Huz76] geht an mehreren Stellen (vgl. [Huz76] S.16, S.98 und S.200) auf die Arbeiten von Koopman [Koo36] und Pitman [Pit36] ein.

Der Grund liegt darin, dass etwa 1970 die Theorie zu Suffizienz und exponentiellen Familien weitestgehend ausgearbeitet war (vgl. [Pfa17, S.1 und S.107]). So konnten viele Autoren die Inhalte in ihrer Gesamtheit präsentieren. Der Aufbau und Struktur früherer Arbeiten hatte sich dadurch überholt. Stellvertretend für viele andere umfassende Darstellungen seien hier nur die Arbeiten von Barndorff-Nielsen [BN78], Brown ([Bro86] oder Lehmann ([LR05] und [LC98]) genannt. (Die späteren Jahreszahlen ergeben sich dadurch, dass die Autoren die Ergebnisse ihrer wissenschaftlichen Publikationen erst einige Jahre nach Erscheinen in Lehrbuchform publizierten bzw. entstanden durch neuere Auflagen.)

Hier werden also die drei Publikationen von Darmois [Dar35], Koopman [Koo36] und Pitman [Pit36] aus heutigem Blickwinkel mit Verweis auf die heutige mathematische Theorie untersucht. Doch zuvor wird auch noch der Beitrag von R. A. Fisher dargestellt. Denn auf ihn gehen nicht nur Begriffe wie Suffizienz, Information, Likelihood und viele andere zurück, sondern er hatte auch schon tiefe Einblicke über suffiziente Statistiken, Maximum-Likelihodd-Schätzer und dem Informationsbegriff. In Fischers Arbeiten finden sich Herleitungen, die das Zusammenwirken zwischen suffizienten Statistiken und speziellen Formen von Verteilungen geben (vgl. [Fis34]). Fisher hatte aber deren konkrete Form (nämlich die Gestalt

der exponentiellen Familien) nicht explizit angegeben.

Bemerkung: Die Notation der historischen Artikel und die Form der Herleitungen wird weitestgehend beibehalten. Es werden nur Hinweise, Erklärungen oder Verweise zu der heutigen Darstellungsform gegeben.

4.2 Ronald Aylmer Fisher

R. A. Fisher war einer der wichtigsten Wegbereiter der modernen Statistik. Viele Begriffe und Methoden gehen maßgeblich auf ihn zurück. Fisher hat Begriffe wie Konsistenz, Effizienz, Suffizienz, Validität und vor allem Likelihood geprägt (vgl. [Han15, S.2]). Viele Gebiete der heutigen Wissenschaft und Forschung, in denen quantitative Methoden Anwendung finden, wie Biologie, Medizin, Psychologie, Sozialwissenschaften etc., sind eng mit dem Namen Fisher verbunden. Dies zeigt sich einerseits dadurch, dass viele Methoden mit seinem Namen bezeichnet sind, wie (exakter) Fisher-Test, F(isher)-Verteilung, usw., und anderseits sind viele Konzepte und Methoden, wie Varianzanalyse, Hypothesenprüfung, Diskriminazanalyse uvm., von ihm angeregt oder entwickelt worden, sodass sein Name immer bei diesen Methoden genannt wird. Weiters haben zu seiner Bekanntheit seine populären Bücher zu quantitativen Forschungsmethoden, wie [Fis50b, Fis56], beigetragen. Das Buch „*Statistical Methods for Research Workers*“ [Fis50b] erschien von 1925 bis 1970 in 14 Auflagen! Und nach wie vor haben diese Werke Gültigkeit. Alle Bücher (die Ausnahme ist einzig [Fis50a], aber dies ist ein Sammelband seiner statistischen Artikel) und viele von Fishers Arbeiten waren für die Anwender quantitativer Methoden geschrieben, also Biologen, Mediziner, Psychologen, Sozialwissenschaftler usw., aber nicht für Mathematiker. Dazu ist anzumerken, dass Fisher selbst viele Beiträge auf dem Gebiet der Genetik, Eugenik und Evolutionsbiologie leistete.

Diese Ausrichtung haftet aber auch seinen rein statistischen Arbeiten an. Keine dieser Arbeiten ist eine „typisch“ mathematische Arbeit, mit exakten Definitionen, Voraussetzungen, Beweisen. Das Verständnis seiner Arbeiten wird auch dadurch erschwert, dass Fisher Bezeichnungen und Konzepte in seinen Arbeiten verwendet, die er in früheren Arbeiten eingeführt hat, und dann in späteren Arbeiten wieder verwendet, ohne sie ein weiteres Mal zu erläutern oder auf die früheren Arbeiten zu referenzieren. Und doch verwendet Fisher mit erstaunlicher Sicherheit und Klarheit, die von ihm eingeführten Begriffe. D.h. er leitet tiefliegende Zusammenhänge oft anhand von Spezialfällen her, und bewegt sich ohne formale Strenge hin zu allgemeingültigen Schlüssen. Diesen Mangel an Stringenz und Strenge füllen dann nachfolgende Autoren. Und genau dies geschieht im Fall des Satzes von Darmois-Koopman-Pitman. Hier hat Fisher die Herleitung (in seiner intuiti-

ven Darstellungsweise) angegeben (siehe [Fis34, S.293]), ohne jedoch den letzten entscheidenden Schritt zu vollziehen und die besondere Form der exponentiellen Familie hervorzuheben.

Im nächsten Abschnitt werden Fishers Vorarbeiten zum Satz von Darmois-Koopman-Pitman beschrieben.

4.2.1 Beiträge Fishers zu D-K-P

Im Wesentlichen leisten drei Arbeiten R. A. Fishers ([Fis22, Fis25, Fis34]) einen Beitrag zum Satz von Darmois-Koopman-Pitman. Auf diese drei Arbeiten verweisen auch spätere Autoren immer im Zusammenhang mit dem Satz von Darmois-Koopman-Pitman.

1. Die Arbeit *„On the Mathematical Foundations of Theoretical Statistics“* [Fis22] ist ein Meilenstein in der Geschichte der Statistik. Auf 60 Seiten legt hier Fisher einen Grundstein der modernen Statistik. Eingereicht wurde diese Arbeit am 25.6.1921 bei der Royal Society und veröffentlicht wurde sie am 19.4.1922 (vgl. [Han15, S.1]). Dadurch erklärt sich, dass Fisher bei anderer Gelegenheit von seiner Arbeit von 1921 spricht, wohingegen die Arbeit in den Literaturverzeichnissen immer unter 1922 geführt wird. In weiterer Folge werden nur Aspekte der Arbeit Fishers besprochen, die die Entwicklung des Satzes von Darmois-Koopman-Pitman beeinflusst haben.

   Die Arbeit beginnt mit einer Liste von 15 Begriffsbildungen. Darunter so grundlegende Begriffe wie Konsistenz, Effizienz, Schätzung, Fisher-Information (*„intrinsic accuracy“*), Likelihood, Suffizienz und Validität. Nicht alle Begriffe wurden hier erstmalig präsentiert, aber alle diese Begriffe erfahren hier eine (wenn auch nur informelle) Definition und zeigen Fishers tiefes Verständnis dieser Begriffsbildungen. Herausgegriffen und im Original zitiert, soll hier nur der Begriff der Suffizienz werden (siehe [Fis22, S.310]). Auch sind die folgenden Originalzitate gute Beispiele für die Ausdrucksweise Fishers.

   Suffizienz: *„A statistic satisfies the criterium of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated.“* (siehe [Fis22, S.310])

   Suffizienzkriterium: *„That the statistik chosen should summarise the whole information of the relevant information supplied by the sample.“* und weiter *„In mathematical language we may interpret this statement by saying that if θ be the parameter to be estimated, θ_1 a statistik which contains the whole of the information as to the value of θ, which the sample supplies, and θ_2 any other statistic, then the surface of the distribution of pairs of*

> *values θ_1 and θ_2, for a given value of θ, is such that for a given value of θ_1, the distribution of θ_2, does not involve θ. In other words, when θ_1 is known, knowledge of the value of θ_2 throws no further light upon the value of θ*". (siehe [Fis22, S.316])

Dies ist ein typisches Beispiel für eine Definition bei Fisher. Sein Stil bemüht sich um Klarheit, daher die ausführliche Beschreibung, aber nicht um mathematischen Formalismus. Der Eindruck, er möchte sich nicht mit formalen Details aufhalten oder bremsen lassen, ist auch nicht ganz von der Hand zu weisen. Allerdings ist zu bedenken, dass der geeignete Begriffsapparat aus der Maßtheorie für die Statistik und die Wahrscheinlichkeitsrechnung, wie heutzutage üblich, 1922 noch nicht zur Verfügung gestanden wäre.

Im Abschnitt „*7.Satisfaction of the Criterion of Sufficiency*" analysiert Fisher das Suffizienzkriterium. Fisher postuliert, dass das Suffizienzkriterium für einen Maximum-Likelihood-Schätzer erfüllt ist. (Dies gilt aber nur, falls der ML-Schätzer eindeutig ist (vgl. [Moo71, S.51] bzw. [Gor16, S.420]). Ausgehend von einem Maximum-Likelihood-Schätzer $\hat{\theta}$ und einer weiteren nicht näher spezifizierten Schätzung θ_1 für einen unbekannten Parameter θ definiert er für eine Stichprobe die „*frequency*"-Funktion f

$$f(\theta, \hat{\theta}, \theta_1) d\hat{\theta} d\theta_1$$

mit $\hat{\theta}$ und θ_1 in den Bereichen $d\hat{\theta}$ bzw. $d\theta_1$. Der Bereich $d\hat{\theta}d\theta_1$ liegt im „isostatischten" Bereich $\hat{\theta}$ (*...isostatistical regions provide identical sets of statistics* [Fis22, 330]). Daher gilt

$$\frac{\partial}{\partial \theta} \log f(\theta, \hat{\theta}, \theta_1) = 0 \tag{4.1}$$

unabhängig von θ_1 für $\theta = \hat{\theta}$. Das ist der Fall, wenn

$$f(\theta, \hat{\theta}, \theta_1) = \phi(\theta, \hat{\theta}) \cdot \phi'(\hat{\theta}, \theta') \tag{4.2}$$

und weiters

$$\frac{\partial}{\partial \theta} \log f = \frac{\partial}{\partial \theta} \log \phi$$

gilt. So vereinfacht sich (4.1) zu

$$\frac{\partial}{\partial \theta} \log \phi(\theta, \hat{\theta}) = 0$$

unabhängig von θ_1. Fisher merkt an, dass die Faktorisierung (4.2) eine mathematische Definition der Suffizienz ist. Dies entspricht dem Faktorisierungskriterium. (vgl. [LR05, S.19])

Allerdings zeigt Fisher nicht, dass der ML-Schätzer suffizient ist, sondern die Umkehrung, dass (4.1) aus (4.2) folgt (siehe [Fis22, S.330f] und vgl. [Gor16, S.390]). Auch geht Fisher implizit davon aus, dass der Parameter θ die gleiche Dimension hat, wie eine suffiziente Statistik dieses Parameters. Dies muss aber nicht der Fall sein. Es kann auch eine Menge $T_1, \ldots, T_m$ von (linear unabhängigen) Statistiken suffizient für $\theta_1, \ldots, \theta_q$ Parameter ($q \leq m$) sein (vgl. [Pit36, S.577] und [Ols40]).

Weitere Abschnitte dieser bemerkenswerten Arbeit, wie zum Beispiel Punkt 6 in [Fis22, S.324], wo zum ersten Mal die „Maximum-Likelihood-Methode“ vorgestellt wird), finden hier keine Betrachtung, da sie einerseits nur in spezieller Form anhand von Beispielen vorgestellt werden, und anderseits für die Entwicklung des Darmois-Koopman-Pitman-Theorems nur untergeordnete Bedeutung haben.

2. In *„Theory of Statistical Estimation“* [Fis25] unternimmt Fisher Klarstellungen und Verfeinerungen zu seiner Arbeit [Fis22]. Aber es werden auch neue Konzepte erstmals eingeführt wie nebensächliche (anziliäre) Statistiken und Verlustfunktionen für die Information. So erwähnt er explizit Statistiken von nicht normal-verteilten Daten, die aber ebenso suffizient sind „*. . . which contain in themselves the whole of the relevant information available in the data*“ ([Fis25, S.712]).

   In dieser Arbeit gibt Fisher eine Definition einer suffizienten Statistik ohne Bezug zu ML-Schätzern, aber analog zu Gleichung (4.2):

   Suffizienz: *„In general, if θ is any parameter, T_1 a statistic sufficient in estimating that parameter, and T_2 any other statistic, the sampling distribution of simultaneous values of T_1 and T_2 must be such that for any given value of T_1, the distribution of T_2 does not involve θ.“* (siehe [Fis25, S.713])

   Und dazu folgt das Faktorisierungskriterium im Anschluss.

   Faktorisierungskriterium: *„This will evidently be the case, if $f(\theta, T_1, T_2)dT_1dT_2$ be the probability that T_1 and T_2 should fall in the ranges dT_1, dT_2, and if*

$$f(\theta, T_1, T_2) = \phi(\theta, T_1)\phi'(T_1, T_2).$$

   If this condition is fulfilled for all possible statistics T_2, then will T_1 be a sufficient statistic.“. (siehe [Fis25, S.713])

   Und Fisher gibt noch eine informelle Beschreibung *„When a sufficient statistic exists it is equivalent, for all subsequent purposes of estimation, to the original data from which it was derived.“*. (siehe [Fis25, S.713f])

3. *„Two New Properties of Mathematical Likelihood“* [Fis34] enthält zwei Weiterentwicklungen der Arbeiten [Fis22, Fis25]. Zum einen zeigt Fisher, dass Verteilungen von einer bestimmen Gestalt sein müssen, um suffiziente Statistiken (unter gewissen Voraussetzungen) aufzuweisen und zum anderen beschreibt er den Gebrauch von nebensächlichen (anziliären) Statistiken zur Verbesserung der Schätzung von Parametern. Im Weiteren wird nur auf den ersten Punkt eingegangen.

   Im zweiten Abschnitt („2. The Distribution of Sufficient Statistics“) von [Fis34] gibt Fisher eine Definition der Suffizienz und des Faktorisierungskriteriums wie in [Fis25, S.713]. Danach beschreibt er sehr klar sein Vorhaben:

   > *„The condition that $\frac{\partial L}{\partial \theta}$ should be constant over the same sets of samples for all values of θ, which has been shown to establish the existence of a sufficient estimate of θ, thus requires that the likelihood is a function of θ, which, apart from a factor dependent on the sample, is of the same form for all samples yielding the same estimate T. The sufficiency of sufficient statistics may thus be traced to the fact that in such cases the value of T itself alone determines the form of the likelihood as a function of θ. If a conventional value such as unity is given to the maximum likelihood for any sample, the likelihood is thus expressible as a function of θ and T only, if T is the sufficient estimate. We shall use this property in obtaining a general form for the sampling distribution of sufficient statistics“* [Fis34, S.289]

   Und die Herleitung erfolgt bei Fisher folgendermaßen: Sei die Lösung einer Maximum-Likelihood-Gleichung

   $$\phi(T) = A$$

   mit einer symmetrischen Funktion A, die nur von den Daten, aber nicht vom Parameter abhängt. Der Ausdruck für $\frac{\partial}{\partial \theta} \log L$ kann dann als

   $$C\{A\psi'(\theta) - \phi(\theta) \cdot \psi'(\theta)\}$$

   geschrieben werden. Dabei ist C entweder konstant oder eine Funktion von A und ψ eine differenzierbare Funktion. Daher ist $\log L$ von der Form

   $$CA\psi(\theta) - C \int \phi(\theta) d\psi(\theta) + B$$

   mit einer Funktion B, die nur von den Daten abhängt und mit vom Parameter unabhängigen Dichten. Dass die (nur von den Daten abhängige und symmetrische) Funktion C gleich n ist, ergibt sich daraus, dass $\log L$ die Summe von

Ausdrücken der einzelnen Beobachtungen ist. Daher setzt Fisher

$$CA = S(X), \quad B = S(X_1)$$

wobei X und X_1 nur von den Daten abhängen. Und so kann die Likelihood-Funktion L in der Form

$$e^{-n\int \phi(\theta)d\psi(\theta)} e^{\psi(\theta)\cdot S(X)} e^{S(X_1)} \tag{4.3}$$

geschrieben werden (siehe [Fis34, S.293f] und vgl. [Gor16, S.594]).

Das bedeutet aber, dass die Dichte jeder Beobachtung in der Form

$$c(\theta)e^{q(\theta)t(x)}h(x)$$

geschrieben werden kann, wobei $c(\theta)$ und $q(\theta)$ nur vom Parameter abhängen, $t(x)$ und $h(x)$ nur von den Daten abhängen und der Träger der Dichten unabhängig von θ ist. Und das ist die Gestalt einer ein-parametrigen Exponentialfamilie. Jedoch hat Fisher dieser Form nicht weiter Bedeutung beigemessen, sondern hat weitere Überlegungen mit der charakteristischen Funktion angestellt.

Am Ende des Kapitels verweist Fisher noch auf den Zusammenhang mit dem Neyman-Pearson-Lemma bei UMP-Tests („uniformly most powerfull tests“). Für ein-parametrige Exponentialfamilien existieren (unverfälschte) UMP-Tests. Insbesondere durch die Existenz suffizienter Statistiken und Likelihood-Funktionen, die dem Faktorisierungskriterium genügen (vgl. [Oli14, S.185ff] und [CB01, S.388ff]).

4.3 Georges Darmois

Der Artikel von Darmois [Dar35] erschien 1935 in den *Comptes Rendus de l'Académie des sciences (kurz C. R. Acad. Sci.)* der Académie des sciences (französische Akademie der Wissenschaften). Dies ist eine wissenschaftliche Zeitschriftenreihe, die Tagungsberichte der Académie des sciences herausgibt. Diese Reihe hat eine lange Tradition. Sie wird in dieser Form seit 1835 herausgegeben. Ein Vorläufer dieser Schriftenreihe existierte schon ab 1666.

Dadurch erschließt sich auch der Stil der Publikation: Es ist ein Bericht von nicht ganz zwei Seiten eines Forschungsergebnisses, welches in einer Sitzung am 8.April 1935 vorgestellt wurde, enthält aber keinerlei Beweise oder Herleitungen. In der Publikationsliste von Darmois (siehe [Dug61]) lässt sich auch keine Publikation finden, die die Herleitung des Resultats aus [Dar35] näher beschreibt.

Darmois nimmt zwar im 1936 erschienen Buch (siehe [Dar36] S.24-S.26) wieder Bezug auf seine Resultate aus [Dar35], aber auch hier bleibt die Herleitung im Dunklen.

Aber es ist die erste Veröffentlichung, die den Zusammenhang zwischen suffizienten Statistiken und exponentiellen Familien darstellt und erlangt dadurch ihre Bedeutung.

4.3.1 Analyse Artikel Darmois (1935)

Der Titel der Arbeit *„Sur les lois de probabilité à estimation exhaustive“* bedeutet etwa *„Über die Wahrscheinlichkeitsgesetze zu einer erschöpfenden Schätzung“* und beschreibt den Inhalt sehr treffend.

Darmois beschreibt am Anfang seiner Arbeit [Dar35] den Urspungs des Begriffs der *sufficient estimation* von Fisher aus dem Jahr 1922 [Fis22], und erwähnt auch noch die Fortführung in [Fis34].

Zunächst hält er fest, dass die (Punkt-)schätzung eines unbekannten Parameters α einer bekannten Verteilung, durch n unabhängige Beobachtungen erfolgen kann. Dann wirft er die Frage auf, ob die alleinige Kenntnis der n unabhängigen Beobachtungen alle Information in Bezug auf den unbekannten Parameter ausschöpft?

Darmois beantwortet diese Frage sogleich, dass bei Vorliegen einer Zufallsstichprobe $X_1, X_2, \ldots, X_n$ und einer daraus gebildeten „erschöpfenden“(suffizienten) Statistik, die zugrunde liegende Verteilung eine bestimmte „Struktur“ haben muss. Des Weiteren hält er fest, dass die Struktur der Verteilungen notwendig und hinreichend für das Vorliegen einer „erschöpfenden“ Schätzung des unbekannten Parameters α ist. D.h. die „erschöpfende“ Schätzung liefert das gleiche Maß an Information, wie der Parameter selbst.

Bemerkung: Darmois verwendet nicht den Begriff einer Statistik, sondern stellt die Funktion der n Zufallsvariablen als Punkt im euklidischen Raum dar, bzw. verwendet Darmois Hyperebenen zur Beschreibung.

Hernach postuliert er die Form, die die Wahrscheinlichkeitsdichte $f(x\alpha)$ in diesem Fall haben muss :

$$f = e^{u(x)v(\alpha)+r(x)+s(\alpha)}$$

wobei $s(\alpha)$ von den Funktionen u,v und r abhängt. Dies präzisiert er erst in [Dar36], dass dies im Sinne der Normierung $\int f(x\alpha)dx = 1$ gemeint ist. Weiters ist die Gleichung der Hyperebene $u(x_1) + u(x_2) + \cdots + u(x_n) = const.$

Unmittelbar daran anschließend gibt Darmois die Form für k Variablen und h Parameter an:

$$\log f = U_1(x, y, z, \ldots)A_1(\alpha, \beta, \ldots)+\cdots+U_h(x, y, z, \ldots)A_h(\alpha, \beta, \ldots)+V(x, y, z, \ldots)+B(\alpha, \beta, \ldots)$$

wobei in diesem Fall $B(\alpha, \beta, \dots)$ von den Funktionen U_i, A_j und V abhängt.

Den Schluss des Berichts bildet ein Hinweis, dass sowohl stetige als auch diskrete Verteilungen durch die Definition beschrieben werden und ein Hinweis auf eine mehrdimensionale Gauss-Verteilung als Beispiel.

Diese Arbeit legt zwar neue Erkenntnisse aufbauend auf die vagen Formulierungen Fishers über den Zusammenhang zwischen exponentiellen Familien und suffizienten Statistiken dar, enthält aber keinerlei Beiweise oder irgendwelche Herleitungen. Auch die verwendeten Größen werden nicht näher erklärt. Darmois macht keinerlei Angaben über die Definitionsbereiche der vorkommenden Variablen noch gibt er einschränkende Bedingungen für die verwendeten Funktionen an.

4.4 Bernard Osgood Koopman

Unabhängig von der Publikation von Darmois [Dar35] hat Koopman seine Arbeit *„On Distributions Admitting a Sufficient Statistic“* [Koo36] am 16. März 1935 bei der *American Mathematical Society* eingereicht. Präsentiert wurde die Arbeit am 20. April 1935, und dann 1936 in den *Transactions of the American Mathematical Society* veröffentlicht. Es ist davon auszugehen, dass Koopman keine Kenntnis von der Arbeit Darmois hatte. Zumindest findet sich kein Hinweis darauf in seiner Arbeit. Koopman referenziert nur auf Arbeiten Fishers [Fis22, Fis25]. Seine Arbeit ist mathematisch sehr stringent und auch in ihrer Darstellung sehr analytisch und klar. Sie wird in dieser Hinsicht auch von anderen Autoren hervorgehoben (vgl. [Huz76, S.100f, S.108f], [BM63, S.218]).

4.4.1 Analyse Artikel Koopman (1936)

Dieser Abschnitt ist eine Analyse der Arbeit von [Koo36] hinsichtlich der Voraussetzungen, Definitionen und Beweise und folgt [Koo36] durch alle Abschnitte. Alle Formulierungen stammen, wenn nicht anders erwähnt aus [Koo36].

Koopman beginnt seine Arbeit [Koo36] mit einer detailierten Darstellung der verwendeten mathematischen Größen deren Bezeichnungen und Voraussetzungen. Am Beginn der Ausführungen steht die Erklärung einer Dichtefunktion $f(\theta_1, \dots, \theta_\nu, x)$, wobei $x \in \mathbb{R}^n$ und die Parameter $(\theta_1, \dots, \theta_\nu)$ zwar unbekannt sind, aber einem vorgegebenen Bereich $\Omega \subseteq \mathbb{R}^\nu$ entstammen .

Die Frage, die er dann stellt ist: Wie können n unabhängige Beobachtungen $(x_1, \dots, x_n)$ der stochastischen Größe X zur Festlegung der Parameter $(\theta_1, \dots, \theta_\nu)$ verwendet werden?

Bezugnehmend auf die Arbeit von Fisher [Fis22] stellt Koopman die Antwort auf diese Frage wie folgt dar: Es müssen ν Funktionen anhängig von n Argumenten $\phi_j(x_1, \dots, x_n)(j = 1, \dots, \nu)$ festgelegt werden, sodass die Funktionswerte als

Schätzwerte für $(\theta_1, \ldots, \theta_\nu)$ herangezogen werden können. Mit Verweis auf Fisher nennt Koopman die ϕ_j *Schätzer* für θ_j bzw. *Statistiken*.

Im nächsten Schritt stellt Koopman die Frage, ob diese eine Punktschätzung $(\phi_1(x_1, \ldots, x_n), \ldots, \phi_n(x_1, \ldots, x_n)) \in \Omega$ die gesamte „relevante Information" aus der Zufallsstichprobe $(x_1, \ldots, x_n)$ über die Lage der Parameter $(\theta_1, \ldots, \theta_\nu) \in \Omega$ enthält? Oder geht „relevante Information"(Koopman) verloren? Dabei wird immer $\nu < n$ vorausgesetzt.

Dann verweist Koopman darauf, dass Fisher im Fall, dass keine „relevante Information" verlorengeht, die Menge $\phi_1(x_1, \ldots, x_n), \ldots, \phi_n(x_1, \ldots, x_n)$ als Menge suffizienter Statisiken bezeichnet. Und der Begriff „relevante Information" ist auch im Sinne Fishers als bestimmtes Integral zu verstehen (vgl. Definition 3.7).

An dieser Stelle erfolgt bei Koopman die mathematische Definition einer suffizienten Statistik in Sinne Fishers, wie Koopman es nennt. Diese Definition soll die intuitive Formulierung des Begriffs von Fisher präzisieren.

Definition 4.1 (Suffiziente Statistiken nach Koopman)**.** *Für eine Verteilung $f(\theta_1, \ldots, \theta_\nu, x)$ ist das System von Statistiken $\phi_j(x_1, \ldots, x_n)(j = 1, \ldots, \nu)$ suffizient, wenn aus dem Gleichungssystem*

$$\phi_j(x_1, \ldots, x_n) = \phi_j(x'_1, \ldots, x'_n) \quad (j = 1, \ldots, \nu) \tag{4.4}$$

für $(\theta_1, \ldots, \theta_\nu) \in \Omega, (\theta'_1, \ldots, \theta'_\nu) \in \Omega$ folgt:

$$\frac{\prod_{i=1}^n f(\theta_1, \ldots, \theta_\nu, x_i)}{\prod_{i=1}^n f(\theta_1, \ldots, \theta_\nu, x'_i)} = \frac{\prod_{i=1}^n f(\theta'_1, \ldots, \theta'_\nu, x_i)}{\prod_{i=1}^n f(\theta'_1, \ldots, \theta'_\nu, x'_i)} \tag{4.5}$$

Diese Definition ist verglichen mit der heutzutage üblichen Definition von Suffizienz (vgl. 2.7) enger gefasst, da sie eine Aussage über Likelihood-Quotienten macht. Koopman gibt auch zusätzlich eine Begründung seiner Definition. Sehr ähnliche aber präziser formulierte Argumente zu den Eigenschaften von Likelihood-Quotienten finden sich beispielsweise in [LR05, S. 59f]. Dadurch enthält die Definition in gewisser Weise auch schon die Aussage des Faktorisierungs-Theorems (siehe Satz 2.3).

Auch wird angenommen, dass es verschiedene (stetige) Versionen der Dichte gibt (siehe Bemerkung zu Satz 1.3 und [Kee10, S.7]).

Bemerkung: Koopman verwendet im unten folgenden Satz 4.1 die Notation $(R - T)$. Dabei bedeutet R immer $\mathbb{R}^n$, d.h. im Weiteren sei $R := \mathbb{R}^n$ um bei der Analyse von Satz 4.1 bei der Notation Koopmans zu bleiben. Zur Menge T macht Koopman keinerlei Angabe bzw. es finden sich keine direkten Hinweise auf die Gestalt der Menge T. Die Notation bedeutet, dass $(R - T)$ eine (nichtleere) offene

Teilmenge des $\mathbb{R}^n$ ist. Die Einschränkung auf offene Teilmengen des $\mathbb{R}^n$ wird für die Definitionsmenge der in Satz 4.1 verwendeten (Dichte)-Funktion f benötigt.

Mit dieser Definition einer suffizienten Statistik formuliert Koopman sein erstes Theorem:

Satz 4.1. *Sei $f(\theta_1, \ldots, \theta_\nu, x)$ analytisch und sei $f \neq 0$ auf $\Omega \times (R - T) \subset \Omega \times R$ (vgl. obige Bemerkung). Weiters sei $\phi_j(x_1, \ldots, x_n) = \phi_j(x'_1, \ldots, x'_n)$ $(j = 1, \ldots, \nu)$ stetig auf $\mathbb{R}^n$ sowie $n > \nu$.*

Dann ist eine notwendige Bedingung, dass $(\phi_1, \ldots, \phi_\nu)$ ein System suffizienter Statistiken für die obige Verteilung ist, und dass für alle Punkte $(a_1, \ldots, a_\nu, b) \in \Omega \times (R - T)$ eine Umgebung $N_{ab} = \omega_{ab} \times r_{ab} \subset \Omega \times (R - T)$ existiert, wobei

$$\begin{aligned} \omega_{ab} &:| \theta_j - a_j |< h_j \qquad (j = 1, \ldots, \nu), \\ r_{ab} &:| x - b |< h, \end{aligned}$$

sodass

$$f(\theta_1, \ldots, \theta_\nu, x) = \exp\left[\sum_{k=1}^{\mu} \Theta_k X_k + \Theta + X\right]. \tag{4.6}$$

Die dabei auftretenden Funktionen Θ_k, Θ sind reellwertige analytische Funktionen der jeweiligen Parameter $(\theta_1, \ldots, \theta_\nu)$ in ω_{ab}, die X_k, X sind reellwertige analytische Funktionen von x in r_{ab} und $\mu \leq \nu$ ($\mu = 0$ bedeutet, dass Θ_k und X_k fehlen).

Darüberhinaus gilt für den kleinsten Wert von μ, für den die Gleichung (4.6) noch zutrifft:

$$\sum_{i=1}^{n} X_k(x_i) = V_k[\phi_1(x_1, \ldots, x_n), \ldots, \phi_\nu(x_1, \ldots, x_n)] \quad (k = 1, \ldots, \mu)$$

Wobei diese minimale Darstellung immer erreicht werden kann.

Beweis. Im Beweis beschränkt sich Koopman auf $\nu = 2$, also auf zwei Parameter mit der Begründung, dass dies „sufficiently illustrative“ sei [Koo36, 402]. OBdA wählt er $n = 3$ unter Einhaltung der Voraussetzung $n > \nu$, und für $n > 3$ setzt er $\phi_j(x_1, x_2, x_3) = \phi_j(x_1, x_2, x_3, b, \ldots, b)$ bzw. $b = x_4 = \cdots = x_n = x'_4 = \cdots = x'_n$.

Da N_{ab} eine Teilmenge von $\Omega \times (R - T)$ ist und $f(\theta_1, \theta_2, x)$ eine reelle einwertige analytische Funktion $\neq 0$ auf N_{ab} ist, ist auch die Funktion

$$\log \frac{f(\theta_1, \theta_2, x)}{f(\theta'_1, \theta'_2, x)}$$

eine reelle einwertige analytische Funktion für alle $(\theta_1, \theta_2, \theta'_1, \theta'_2, x)$ in einer Umge-

bung von $\omega_{ab}^2 \times r_{ab}$ von (a_1, a_2, a_1, a_2, b).

Schrittweises Einsetzen von $(\theta_1', \theta_2') = (a_{j1}, a_{j2})$ $(j = 1, 2, 3)$ in (4.5) ergibt folgendes Gleichungssystem:

$$\sum_{i=1}^{n} \log \frac{f(\theta_1, \theta_2, x_i)}{f(a_{j1}, a_{j2}, x_i)} = \sum_{i=1}^{n} \log \frac{f(\theta_1, \theta_2, x_i')}{f(a_{j1}, a_{j2}, x_i')} \quad (j = 1, 2, 3) \tag{4.7}$$

mit darin vorkommenden reellen einwertigen analytischen Funktion
für alle Punkte $(\theta_1, \theta_2, x_1, x_2, x_3)$ und $(\theta_1, \theta_2, x_1', x_2', x_3')$ aus $\omega_{ab} \times r_{ab}^3$. Da nach Voraussetzung ϕ_1, ϕ_2 suffizient sind, ergibt sich (4.7) als Konsequenz von (4.4):

$$\phi_1(x_1, x_2, x_3) = \phi_1(x_1', x_2', x_3')$$
$$\phi_2(x_1, x_2, x_3) = \phi_2(x_1', x_2', x_3')$$

An dieser Stelle hält Koopman fest, dass die Determinante der Jacobi-Matrix der linken Seite von (4.7) abgeleitet nach x_1, x_2, x_3 nicht vollen Rang besitzt ("...identical vanishing of the jacobian..."; vgl. [Koo36, S.403]). Er begründet dies indirekt unter Zuhilfenahme der Stetigkeitseigenschaften der suffizienten Statistiken aus seiner Definition 4.1.

Dann folgt eine Darstellung der (führenden) Hauptminoren der Jacobi-Matrix der linken Seite von (4.7), um den Rang ρ der Jacobi-Matrix zu untersuchen. Der Rang einer Matrix ist dann die größte Zahl ρ, für die eine von Null verschiedene ρ-reihige Unterdeterminante existiert.

$$\det \mathbf{M}_3 = |M_{ij}|_{(i,j=1,2,3)} = \left| \frac{\partial}{\partial x_i} \log \frac{f(\theta_1, \theta_2, x_i)}{f(a_{j1}, a_{j2}, x_i)} \right|_{(i,j=1,2,3)}$$
$$\det \mathbf{M}_2 = |M_{ij}|_{(i,j=1,2)}$$
$$\det \mathbf{M}_1 = |M_{11}|$$

Der Rang ρ kann noch die Werte $0, 1, 2$ annehmen, der Fall $\rho = 3$ wurde bereits zuvor bei der Überlegung ausgeschlossen, dass die Jacobi-Matrix nicht vollen Rang besitzt.

Fall 1: $\rho = 0$. Die diesen Fall definierende Gleichung

$$\det \mathbf{M}_1 = \frac{\partial}{\partial x_i} \log \frac{f(\theta_1, \theta_2, x_1)}{f(a_{11}, a_{12}, x_1)} = 0 \tag{4.8}$$

kann nach x_1 über dem Intervall $[b; x] \subseteq r_{ab}$ integriert werden:

$$\log \frac{f(\theta_1, \theta_2, x)}{f(a_{11}, a_{12}, x)} = \log \frac{f(\theta_1, \theta_2, b)}{f(a_{11}, a_{12}, b)} \tag{4.9}$$

Und nach $f(\theta_1, \theta_2, x)$ aufgelöst ergibt sich

$$f(\theta_1, \theta_2, x) = \exp\left[\log \frac{f(\theta_1, \theta_2, b)}{f(a_{11}, a_{12}, b)} + f(a_{11}, a_{12}, x)\right]$$

und mit $\mu = 0$ sowie

$$\Theta(\theta_1, \theta_2) = \log \frac{f(\theta_1, \theta_2, b)}{f(a_{11}, a_{12}, b)})$$
$$X(x) = f(a_{11}, a_{12}, x)$$

entspricht dies der Darstellung (4.6) mit allen geforderten analytischen Eigenschaften.

Fall 2: $\rho = 1$. In diesem Fall ist der Rang der Jacobi-Matrix $\rho = 1$ und es muss $\det \mathbf{M}_2 = 0$ und $\det \mathbf{M}_1 \neq 0$ gelten.

$$\det \mathbf{M}_2 = \begin{vmatrix} \frac{\partial}{\partial x_1} \log \frac{f(\theta_1, \theta_2, x_1)}{f(a_{11}, a_{12}, x_1)} & \frac{\partial}{\partial x_1} \log \frac{f(\theta_1, \theta_2, x_1)}{f(a_{21}, a_{22}, x_1)} \\ \frac{\partial}{\partial x_2} \log \frac{f(\theta_1, \theta_2, x_2)}{f(a_{21}, a_{12}, x_2)} & \frac{\partial}{\partial x_2} \log \frac{f(\theta_1, \theta_2, x_2)}{f(a_{21}, a_{12}, x_2)} \end{vmatrix} = 0$$

Subtraktion der 1.Spalte von der 2.Spalte und Herausziehen von $\frac{\partial}{\partial x_1}$ führt auf:

$$\det \mathbf{M}_2 = \frac{\partial}{\partial x_1} \begin{vmatrix} \log \frac{f(\theta_1, \theta_2, x_1)}{f(a_{11}, a_{12}, x_1)} & \log \frac{f(\theta_1, \theta_2, x_1)}{f(a_{21}, a_{22}, x_1)} - \log \frac{f(\theta_1, \theta_2, x_1)}{f(a_{11}, a_{12}, x_1)} \\ \frac{\partial}{\partial x_2} \log \frac{f(\theta_1, \theta_2, x_2)}{f(a_{21}, a_{12}, x_2)} & \frac{\partial}{\partial x_2} \log \frac{f(\theta_1, \theta_2, x_2)}{f(a_{21}, a_{12}, x_2)} - \frac{\partial}{\partial x_2} \log \frac{f(\theta_1, \theta_2, x_2)}{f(a_{21}, a_{12}, x_2)} \end{vmatrix} = 0$$

Diese Gleichung kann wieder nach x_1 über dem Intervall $[b; x] \subseteq r_{ab}$ (hier steht bei Koopman [Koo36, S.404] fälschlicherweise $\cdots \subseteq \omega_{ab}$) integriert werden, und auf diese Form gebracht werden:

$$\begin{vmatrix} \log \frac{f(\theta_1, \theta_2, x)}{f(a_{11}, a_{12}, x)} & \log \frac{f(a_{11}, a_{12}, x)}{f(a_{21}, a_{22}, x)} \\ \frac{\partial}{\partial x_2} \log \frac{f(\theta_1, \theta_2, x_2)}{f(a_{11}, a_{12}, x_2)} & \frac{\partial}{\partial x_2} \log \frac{f(a_{11}, a_{12}, x_2)}{f(a_{21}, a_{22}, x_2)} \end{vmatrix} = \begin{vmatrix} \log \frac{f(\theta_1, \theta_2, b)}{f(a_{11}, a_{12}, b)} & \log \frac{f(a_{11}, a_{12}, b)}{f(a_{21}, a_{22}, b)} \\ \frac{\partial}{\partial x_2} \log \frac{f(\theta_1, \theta_2, x_2)}{f(a_{11}, a_{12}, x_2)} & \frac{\partial}{\partial x_2} \log \frac{f(a_{11}, a_{12}, x_2)}{f(a_{21}, a_{22}, x_2)} \end{vmatrix} \quad (4.10)$$

Für eine weitere übersichtlichere Darstellung werden folgende (bei Koopman nicht verwendete Abkürzungen) eingeführt:

$$
\begin{aligned}
B &:= \log \frac{f(a_{11}, a_{12}, x)}{f(a_{21}, a_{22}, x)} \\
C &:= \frac{\partial}{\partial x_2} \log \frac{f(\theta_1, \theta_2, x_2)}{f(a_{11}, a_{12}, x_2)} \\
D &:= \frac{\partial}{\partial x_2} \log \frac{f(a_{11}, a_{12}, x_2)}{f(a_{21}, a_{22}, x_2)} \\
G &:= \begin{vmatrix} \log \frac{f(\theta_1, \theta_2, b)}{f(a_{11}, a_{12}, b)} & \log \frac{f(a_{11}, a_{12}, b)}{f(a_{21}, a_{22}, b)} \\ \frac{\partial}{\partial x_2} \log \frac{f(\theta_1, \theta_2, x_2)}{f(a_{11}, a_{12}, x_2)} & \frac{\partial}{\partial x_2} \log \frac{f(a_{11}, a_{12}, x_2)}{f(a_{21}, a_{22}, x_2)} \end{vmatrix}
\end{aligned}
$$

Dadurch lässt sich die Gleichung (4.10) schematisch wie folgt darstellen:

$$
\begin{vmatrix} \log \frac{f(\theta_1, \theta_2, x)}{f(a_{11}, a_{12}, x)} & B \\ C & D \end{vmatrix} = G \tag{4.11}
$$

Ziel ist es wiederum, die Gleichung (4.10) nach $f(\theta_1, \theta_2, x)$ aufzulösen. Es ergibt sich (in vereinfachter Darstellung):

$$
f(\theta_1, \theta_2, x) = \exp\left[\frac{G}{D} + \frac{B \cdot C}{D} + \log f(a_{11}, a_{12}, x)\right] \tag{4.12}
$$

Dies entspricht wieder der in (4.6) geforderten Form (in verkürzter Darstellung), wobei in diesem Fall für $\mu = 1$

$$
\begin{aligned}
\Theta(\theta_1, \theta_2) &= \frac{G}{D} \\
X(x) &= \log f(a_{11}, a_{12}, x) \\
\Theta_1(\theta_1, \theta_2) &= \frac{C}{D} \\
X_1(x) &= B
\end{aligned}
$$

gesetzt werden kann. Voraussetzung dafür ist, dass

$$
D = \frac{\partial}{\partial x_2} \log \frac{f(a_{11}, a_{12}, x_2)}{f(a_{21}, a_{22}, x_2)} \neq 0
$$

gilt, wie sich bei der Umformung nach $f(\theta_1, \theta_2, x)$ gezeigt hat, da sonst im Nenner 0 auftreten würde. Die notwendige Bedingung $D \neq 0$ bedeutet, dass es einen Wert $x_2 \in r_{ab}$ gibt, der dies leistet. Dies lässt sich indirekt zeigen:

Angenommen es gelte

$$D = \frac{\partial}{\partial x_2} \log \frac{f(a_{11}, a_{12}, x_2)}{f(a_{21}, a_{22}, x_2)} = 0$$

Diese Gleichung entspricht Gleichung (4.9) vom Fall $\rho = 0$, wenn man (4.9) mit -1 multipliziert und für (θ_1, θ_2) die Werte (a_{21}, a_{22}) einsetzt.

Aber (4.9) ist äquivalent zu $\det \mathbf{M}_1 = 0$. Und dies ist ein Widerspruch zur Voraussetzung des 2.Falls für $\rho = 1$, dass $\det \mathbf{M}_1 \neq 0$. Also $D = 0$ kann in diesem Fall nicht eintreten.

Fall 3: $\rho = 2$. In diesem Fall ist die Jacobi-Matrix von Rang 2 und es muss $\det \mathbf{M}_3 = 0$ und $\det \mathbf{M}_1, \det \mathbf{M}_2 \neq 0$ gelten. Die Beweisschritte sind analog zu Fall 2:

Im Ansatz $\det \mathbf{M}_3 = 0$ kann wiederum die erste Spalte von der zweiten und dritten Spalte subtrahiert werden, $\frac{\partial}{\partial x_1}$ vorgezogen und danach nach x_1 von b bis a in r_{ab} integriert werden. Die resultierende Gleichung kann wiederum nach $f(\theta_1, \theta_2, x)$ aufgelöst werden. Damit die Gleichung nicht lösbar wäre, müssten 2-reihige Unterdeterminanten den Wert 0 annehmen. Aber diese Fälle können auf $\det \mathbf{M}_2 = 0$ zurückgeführt werden. Dies ist aber für den Fall 3 ausgeschlossen.

Bemerkung: Für diesen Fall gibt aber Koopman keine explizite Darstellung an, sondern nur obige Erklärung.

Der letzte Teil des Beweises zeigt noch, dass für minimales μ in (4.6)

$$\sum_{i=1}^{n} X_k(x_i) = V_k[\phi_1(x_1, \ldots, x_n), \ldots, \phi_\nu(x_1, \ldots, x_n)] \quad (k = 1, \ldots, \mu)$$

gilt.

Dazu wird die Darstellung (4.6) mit der Annahme, dass μ minimal ist, in (4.5) eingesetzt und nach Logarithmieren zur Folgerung von (4.4):

$$\begin{aligned} \sum_{k=1}^{\mu} \left\{ [\Theta_k(\theta_1, \ldots, \theta_\nu) - [\Theta_k(\theta'_1, \ldots, \theta'_\nu)] \cdot \sum_{i=1}^{n} X_k(x_i) \right\} = \\ = \sum_{k=1}^{\mu} \left\{ [\Theta_k(\theta_1, \ldots, \theta_\nu) - [\Theta_k(\theta'_1, \ldots, \theta'_\nu)] \cdot \sum_{i=1}^{n} X_k(x'_i) \right\} \end{aligned} \tag{4.13}$$

Durch schrittweises Einsetzen von $(\theta'_1 \ldots, \theta'_\nu) = (a_{j1}, \ldots, a_{j\nu})(j = 1, \ldots, \mu)$, jeweils auf ω^r_{ab}, erhält man μ Gleichungen, die auf

$$\sum_{i=1}^{n} X_k(x_i) = \sum_{i=1}^{n} X_k(x'_i) \quad (k = 1, \ldots, \mu) \tag{4.14}$$

als Folgerung von (4.4) führen. Dies entspricht aber der letzten Behauptung von Satz 4.1, vorbehaltlich, dass die Determinante

$$\Delta = |\Delta_{kj}| = |\Theta_k(\theta_1, \ldots, \theta_\nu) - \Theta_k(a_{j1}, \ldots, a_{j\nu})| \quad (k, j = 1, \ldots, \mu) \tag{4.15}$$

verschieden von null ist.

Dies lässt sich indirekt zeigen:

Angenommen $\Delta = 0$, und sei Δ' jene Determinante, die man erhält, wenn man die erste Spalte von den anderen Spalten von Δ subtrahiert. Dann gilt ebenfalls $\Delta' = 0$. Dann betrachtet Koopman die Determinante

$$|\Theta_k(a_{11}, \ldots, a_{1\nu}) - \Theta_k(a_{l1}, \ldots, a_{l\nu})| \quad (k = 1, \ldots, \mu - 1; l = 2, \ldots, \mu) \tag{4.16}$$

welche durch streichen der ersten Spalte und der letzten Zeile von Δ' entsteht, oder anders ausgedrückt den Minor $M_{\mu,1}$ von Δ'. Falls es eine Belegung $(a_{l1}, \ldots, a_{l\nu})$ gibt, sodass die obige Determinante gleich null ist, also $M_{\mu,1} = 0$, dann würde daraus

$$|\Delta_{kj}|_{k,j=1,\ldots,\mu-1} = 0 \tag{4.17}$$

folgen.

Aber zuerst zum Fall, dass $M_{\mu,1} \neq 0$: D.h. es gibt eine Belegung für $(a_{l1}, \ldots, a_{l\nu})$, sodass $M_{\mu,1} \neq 0$. Da nach Voraussetzung $\Delta' = 0$ gilt, muss daher die erste Spalte von Δ' linear abhängig von den restlichen Spalten von Δ' sein. Da diese restlichen Spalten unabhängig von den $\{\theta_1, \ldots, \theta_\nu\}$ sind, erhält man $\mu - 1$ reelle einwertige und auf ω_{ab}^{ν} analytische Funktionen $\mathcal{T}_s$ und $\mu(\mu - 1)$ reelle Konstanten C_{ks}, sodass

$$\Theta_k(\theta_1, \ldots, \theta_\nu) = \sum_{s=1}^{\mu-1} C_{ks} \mathcal{T}_s(\theta_1, \ldots, \theta_\nu) \Theta_k(a_{11}, \ldots, a_{1\nu}) \quad (k = 1, \ldots, \mu). \tag{4.18}$$

Durch Einsetzen dieses Ausdrucks in (4.6) zeigt sich unmittelbar, dass (4.6) mit weniger als μ nicht identisch verschwindenden Funktionen $\Theta_k(\theta_1, \ldots, \theta_\nu) X_k(x)$ geschrieben werden könnte. Dies widerspricht aber der Annahme, dass μ bereits minimal ist.

Im Fall, dass (4.17) gilt, also dass $|\Delta_{kj}|_{k,j=1,\ldots,\mu-1} = 0$, setzt man $|\Delta| := |\Delta_{kj}|$ und argumentiert analog wie oben, dass μ nicht reduziert werden kann.

□

Daraus ergibt sich unmittelbar folgendes Korollar:

Korollar 4.1. *Falls von den Parametern $\theta_1, \ldots, \theta_\nu$ aus Satz 4.1 $\theta_1, \ldots, \theta_\lambda \quad (1 < \lambda < \nu)$ unbekannt und $\theta_{\lambda+1}, \ldots, \theta_\nu$ hingegen bekannt sind, dann kann die Voraussetzung, dass $\phi_1, \ldots, \phi_\nu$ suffiziente Statistiken sind, durch die Voraussetzung, dass $\phi_1, \ldots, \phi_\lambda$*

suffiziente Statistiken von der Form $\phi_j(\theta_{\lambda+1},\dots,\theta_\nu;x_1,\dots,x_n) \quad (j=1,\dots,\lambda)$ *sind, ersetzt werden. Es bleiben für* $\mu>\lambda$ *die Aussagen von Satz 4.1 gültig, wobei* X_k, X, V_k *zusätzlich zu* $x_1,\dots,x_n$ *auch von* $\theta_{\lambda+1},\dots,\theta_\nu$ *abhängen.*

Bemerkung: Die bekannten Parameter $\theta_{\lambda+1},\dots,\theta_\nu$ werden also den Daten $x_1,\dots,x_n$ hinzugefügt. Der Beweis erfolgt analog zum Beweis von Satz 4.1 mit λ anstelle von ν und den Daten $(\theta_{\lambda+1},\dots,\theta_\nu;x_1,\dots,x_n)$.

Als nächstes gibt Koopman folgenden Satz für eine disjunkte Zerlegung von $R = R^* \cup R^{**}$ mit $\emptyset = R^* \cap R^{**}$, dabei ist wie in Satz 4.1 $R := \mathbb{R}^n$. (Die Schreibweise bei Koopman ist $R = R^* + R^{**}$, $R^*R^{**} = 0$.)

Satz 4.2. *Eine hinreichende Bedingung für die Suffizienz des Systems von Statistiken* $\phi_j(x_1,\dots,x_n) \quad (j=1,\dots,\nu)$ *der Verteilungen* $f(\theta_1,\dots,\theta_\nu,x)$*, bezüglich einer Zerlegung* $R = R^* \cup R^{**}$ *mit* $\emptyset = R^* \cap R^{**}$*, ist, wenn für*

$$f(\theta_1,\dots,\theta_\nu,x) = \begin{cases} 0 & \forall\ (\theta_1,\dots,\theta_\nu,x)\ \textit{auf}\ \Omega\times R^{**} \\ \exp\left[\sum_{k=1}^{\mu}\Theta_k X_k + \Theta + X\right] & \forall\ (\theta_1,\dots,\theta_\nu,x)\ \textit{auf}\ \Omega\times R^{*} \end{cases} \tag{4.19}$$

$$\sum_{i=1}^{n} X_k(x_i) = V_k[\phi_1(x_1,\dots,x_n),\dots,\phi_\nu(x_1,\dots,x_n)] \quad (k=1,\dots,\mu)$$
$$\forall\ (\theta_1,\dots,\theta_\nu,x)\ \textit{auf}\ \Omega\times R^* \tag{4.20}$$

gilt. Dabei sind Θ_k, Θ *reelle einwertige analytische Funktionen der jeweiligen Parameter* $(\theta_1,\dots,\theta_\nu)$ *für alle Punkte in* Ω *und* X_k, X *sind reelle einwertige analytische Funktionen von x (fast überall auf* R^**), sodass für jedes* $(\theta_1,\dots,\theta_\nu)$ *aus* Ω

$$\int_R f(\theta_1,\dots,\theta_\nu,x)dx = 1$$

(bezüglich des Lebesgue-Maßes) erfüllt ist. Die Funktionen V_k *sind reell und einwertig für alle Werte ihrer* ν *Argumente auf dem Definitionsbereich von* $(\phi_1),\dots,\phi_\nu)$ *und* $(x_1,\dots,x_n)\in(R^*)^n$*.*

Bemerkung: Als Beweis des obigen Satzes verweist Koopman nur auf direktes Einsetzen und Umformen um aus (4.4) (4.5) zu folgern. Weiters merkt hier Koopman an, dass obige Sätze auch leicht für multivariate Verteilungen hergeleitet werden können, d.h. für $x\in R^n$

Satz 4.3. *Als Voraussetzung gelte (4.19) aus Satz 4.2 unter gleichen Bedingungen. Der Ausdruck* $\prod_{i=1}^{n} f(\theta_1,\dots,\theta_\nu,x_i)$ *besitze als Funktion von* $(\theta_1,\dots,\theta_\nu)$ *für jedes* $(x_1,\dots,x_n)\in(R^*)^n$ *ein eindeutiges Maximum* $(\hat\theta_1,\dots,\hat\theta_\nu)$ *im Inneren von* Ω*, wobei obige Funktion differnzierbar nach* $(\theta_1,\dots,\theta_\nu)$ *bei* $(\hat\theta_1,\dots,\hat\theta_\nu)$ *ist und einen positiven*

Wert annimmt. Die Determinante $|\partial\Theta_k/\partial\theta_j|$ *sei* $\neq 0$ *am Punkt* $(\hat{\theta}_1, \ldots, \hat{\theta}_\nu)$ *und die Funktionen* $\hat{\theta}_j = \hat{\theta}_j(x_1, \ldots, x_n)$ $(j = 1, \ldots, \nu)$ *sollen ein System von Statistiken von* θ_j $(j = 1, \ldots, \nu)$ *bilden.*

Dann ist dieses System von Statistiken ein System von suffizienten Statistiken.

Bemerkung: Die Aussage bedeutet verkürzt ausgedrückt: Ein eindeutiger Maximum-Likelihood-Schätzer kann als Funktion suffizienter Statistiken geschrieben werden (vgl. [Moo71]).

Beweis. Mit Berücksichtigung von Satz 4.2 genügt es zu zeigen, dass unter den Voraussetzungen des obigen Satzes die Gleichung (4.20) mit $\phi_j(x_1, \ldots, x_n) := \hat{\theta}_j(x_1, \ldots, x_n)$ erfüllt ist. Unter den gegebenen Voraussetzungen der stetigen Differenzierbarkeit muss für das Maximum bei $(\theta_1, \ldots, \theta_\nu) = (\hat{\theta}_1, \ldots, \hat{\theta}_\nu)$

$$\frac{\partial}{\partial\theta_j} \log \prod_{i=1}^{n} f((\theta_1, \ldots, \theta_\nu), x) = 0 \quad (j = 1, \ldots, \nu) \tag{4.21}$$

gelten. Dann erhält man durch Einsetzen von (4.19) aus Satz 4.2:

$$\sum_{k=1}^{\nu} \left\{ \frac{\partial\Theta_k}{\partial\theta_j} \left[\sum_{i=1}^{n} X_k(x_i) \right] \right\} = -n \frac{\partial\Theta}{\partial\theta_j} \quad (j = i, \ldots, \nu). \tag{4.22}$$

Nun bleibt noch zu zeigen, dass Werte $(\hat{\theta}_1, \ldots, \hat{\theta}_\nu)$ existieren, die diese ν Ausdrücke $[\sum_{i=1}^{n} X_k(x_i)]$ eindeutig festlegen. Dass zumindest eine Belegung für $\hat{\theta}_1, \ldots, \hat{\theta}_\nu)$ existiert, folgt direkt aus der Definition der $\hat{\theta}_j$. Und die Eindeutigkeit folgt aus der Annahme

$$\left| \frac{\partial\Theta_k}{\partial\theta_j} \right|_{\hat{\theta}} \neq 0.$$

□

An dieser Stelle referenziert Koopman wieder auf R. A. Fisher: Die Statistik $(\hat{\theta}_1, \ldots, \hat{\theta}_\nu)$ wird nach Fisher „maximum likelihood"-Statistik bzw. „ML-Schätzer" genannt, und Satz 4.3 präzisiert Fishers Behauptung, dass wenn suffiziente Statistiken existieren, die „maximum likelihood" Statistik eine solche liefert.

Bemerkung: Die strenge Definition der Suffizienz bei Koopman erlaubt Stetigkeitsargumente. Bei vielen Beweisen wird Regularität (stetige Differenzierbarkeit, Existenz innerer Punkte) vorausgesetzt bzw. implizit angenommen.

4.5 Edwin James George Pitman

Die Arbeit Pitmans *„Suffient Statistics and Intrinsic Accuracy"* [Pit36] wurde im Juni 1936 bei den *„Mathematical Proceedings of the Cambridge Philosophical Society"*

eingereicht und im Dezember des selben Jahres veröffentlicht. Diese Arbeit referenziert auf die drei Artikel R. A. Fishers [Fis22, Fis25, Fis34] und ist unabhängig von den Artikeln von Darmois [Dar35] und Koopman [Koo36] erschienen.

Der Titel der Arbeit nimmt also direkt Bezug auf [Fis34], da die Begriffe Suffizienz und Fisher-Information („intrinsic accuracy“) den Kern von [Fis34] bilden. Der Ansatz der Herleitung von Pitman unterscheidet sich gänzlich von dem Koopmans in [Koo36]. Pitman gibt zwar mathematische Herleitungen in seiner Arbeit, allerdings in einer sehr informellen Art. Dieser Umstand findet auch in späteren Arbeiten (vgl. z.B. [Huz76, S. 101] oder [BM63]), die sich auf diesen Artikel beziehen, kritische Erwähnung.

Pitman untersucht den Zusammenhang zwischen Suffizienz und „intrinsic accuracy“, d.h. der Fisher-Information und nimmt die Fisher-Information als Ausgangspunkt seiner Herleitungen. Pitman stellt die Frage nach der Gestalt von Verteilungen, die suffiziente Statistiken zur Parameterschätzung liefern. Weiters untersucht er Verteilungen deren Definitionsbereiche von den Parametern der Verteilungen abhängen. Die Unabhängigkeit der Parameter einer Verteilung vom Definitionsbereich ist eine der Regularitätsbedingungen (vgl. Definition 3.6), die im Vorfeld (z.B. bei der Bestimmung von Maximum-Likelihood-Schätzern) an die Verteilungen gestellt werden. Solche Fragestellungen, wo gewisse Regularitätsbedingungen nicht erfüllt sind, werden daher *nicht-reguläre* Probleme genannt. Dies führt allerdings bei der Lösung im Allgemeinen nicht auf exponentielle Verteilungsfamilien. Eine genaue und übersichtliche Darstellung dieser sogenannten *nicht-regulären* Probleme findet sich in [Huz76]. Der letzte Punkt, den Pitman behandelt, ist der Fall, dass für einen Parameter einer Verteilung nicht eine einzelne suffiziente Statistik existiert, es aber eine Menge von Statistiken gibt, die den Parameter suffizient schätzen.

4.5.1 Analyse Artikel Pitman (1936)

Der Artikel gliedert sich übersichtlich in sieben Abschnitte:

1. Einleitung: Zu Beginn verweist Pitman auf die Arbeiten Fishers [Fis22, Fis25, Fis34] und im Speziellen auf dessen Beitrag zu den Definitionen einer suffizienten Statistik und der (Fisher-)Information. Pitman weist darauf hin, dass Fisher in [Fis34, S. 294] exemplarisch die Gestalt von Verteilungen mit suffizienten Statistiken angedeutet, aber nicht explizit angegeben hat. Pitman weist auch darauf hin, dass Fisher immer davon ausgeht, dass der Wertebereich (genauer der Träger der Dichtefunktion) nicht vom zu schätzenden Parameter abhängt.

2. Dieser Abschitt behandelt die Fisher-Information im Fall, dass der Wertebereich einer Verteilung unabhängig vom Parameter θ der Verteilung ist.

   Pitman definiert für eine Dichtefunktion $f(x,\theta)$ (im Original als *frequency function* bezeichnet) auf einem Bereich (a,b)

$$J = \mathbb{E}\left\{\frac{\partial}{\partial\theta}\log f\right\} = \mathbb{E}\left\{\frac{f'}{f}\right\} = \int_a^b f'dx \tag{4.23}$$

$$I = \mathbb{E}\left\{\left(\frac{\partial}{\partial\theta}\log f\right)^2\right\} = \mathbb{E}\left\{\left(\frac{f'}{f}\right)^2\right\} = \int_a^b \frac{(f')^2}{f}dx \tag{4.24}$$

$$\tag{4.25}$$

   f' soll die Ableitung von f nach θ bedeuten und im Fall diskreter Verteilungen (bei Pitman *„discontinuous distributions“*) sei das Integral durch Summation zu ersetzen. Präzisere Definitionen und weitere Voraussetzungen gibt Pitman nicht an. Er knüpft direkt an den Artikel von R. A. Fisher an. Allerdings findet sich auch dort keine genaue Darstellung. Beide lassen offen, welche Bedingungen für obige Definition bzw. für die weiteren Berechnungen nötig sind. Implizit setzt er allerdings die Regularitätsbedingungen (siehe Definition 3.6) voraus.

   Da für f

$$\int_a^b f dx = 1$$

   gilt, und nach Voraussetzung a und b von θ unabhängig sind, erhält man durch differenzieren nach θ, dass $J = 0$ ist und somit ist I die Varianz von $(\frac{\partial}{\partial\theta}\log f)$. Und I kann in Form

$$I = -\mathbb{E}\left\{\frac{\partial^2}{\partial\theta^2}\log f\right\}$$

   geschrieben werden. (Pitman schreibt zur Herleitung nur *„it is easy to show“*. Die Herleitung findet sich in [Kee10, S.72f] bzw. (3.12).)
   An dieser Stelle verweist Pitman darauf, dass I Fishers *„intrinsic accuracy“* ist, und I im Fall einer Normalverteilung mit Mittelwert θ der Kehrwert der Varianz ist (vgl. [Kee10, S.74], wo gezeigt wird, dass dieses Resultat sogar auf einparametrige exponentielle Familien erweiterbar ist).

   Dann wird für eine Zufallsstichprobe $(x_1,\ldots,x_n)$) und deren gemeinsamer Dichte $\phi = \prod f(x_r)$ die Fischer-Information angegeben:

$$\mathbb{E}\left\{\left(\frac{\partial}{\partial\theta}\log\phi\right)^2\right\} = \mathbb{E}\left\{\left(\frac{\phi'}{\phi}\right)^2\right\} = nI$$

da $\left(\frac{\phi'}{\phi}\right)$ die Summe von n unabhängigen Variablen mit gleicher Varianz ist (siehe [HMC12, S. 330] und [Fel14, S. 67]).

Im letzten Teil dieses Abschnitts betrachtet Pitman noch eine (erwartungstreue) Schätzung H für einen unbekannten Parameter θ. Für diese Schätzung H gibt Pitman noch eine Herleitung, dass deren Fischer-Information durch nI beschränkt ist. Dabei verweist Pitman auf [Fis25], wo Fischer bereits diesen Umstand exemplarisch zeigt. Dies entspricht einem Korollar der Informationsungleichung (siehe [HMC12, S.332]). Es folgt die Herleitung für eine Schätzung H mit Wahrscheinlichkeitsfunktion $h(H, \theta)$ und Fischer-Information

$$\mathbb{E}\left\{\left(\frac{\partial}{\partial\theta}\log h\right)^2\right\}$$

bezüglich einer diskreten Grundgesamtheit (in [Pit36] als *„discontinuous population“* bezeichnet). Daher ergibt sich die Wahrscheinlichkeit, dass H einen beliebigen Wert annimmt, aus der Summe der Wahrscheinlichkeiten von Stichproben, die diesen Wert ergeben:

$$h = \sum \phi$$

(wobei die Summation über die Gruppe von Stichproben zu verstehen ist, die diesen jeweiligen Wert von H ergeben). Daher erhält man mit $h' = \sum \phi'$

$$\frac{h'}{h} = \frac{\sum \phi'}{\sum \phi} = \frac{\sum(\phi'/\phi)\phi}{\sum \phi} = \mathbb{E}_H(\phi'/\phi)$$

den Erwartungswert innerhalb der Gruppe von ϕ'/ϕ. Für die Varianz V_H innerhalb der Gruppe ergibt sich mit Hilfe des Verschiebungssatzes (vgl. [Kee10, S.10])

$$\left(\frac{h'}{h}\right)^2 = \{\mathbb{E}_H(\phi'/\phi)\}^2 = \mathbb{E}_H\{(\phi'/\phi)^2\} - V_H(\phi'/\phi).$$

Und nach Multiplikation mit h und Summation über alle Werte von H erhält man die Fischer-Information von H

$$I' = \mathbb{E}\{(\phi'/\phi)^2\} - V^*(\phi'/\phi)$$

wobei V^* der Mittelwert aller Gruppenvarianzen ist. Daher gilt

$$I' \leq \mathbb{E}\{(\phi'/\phi)^2\} = nI \tag{4.26}$$

und Gleichheit gilt nur im Fall, dass alle Gruppenvarianzen den Wert null

haben. D.h. ϕ'/ϕ ist konstant, wenn H konstant ist. Und daraus folgert Pitman, dass ϕ'/ϕ auch von H abhängt und nicht ausschließlich eine Funktion von θ sein kann.

Im Fall, dass für eine Schätzfunktion eines Parameters resp. Statistik θ in (4.26) Gleichheit gilt, wird dieser Schätzer auch effizient genannt (vgl [HMC12, S.332]).

Bemerkung: Die Notation Pitmans, wie z.B. die Ableitung der Score-Funktion $\left(\frac{\partial}{\partial\theta}\log f\right)$ als (f'/f) zu schreiben, ist der aus [Fis25] ähnlich. Allerdings ist Pitman in seinen mathematischen Ausführungen stringenter als R. A. Fisher in [Fis25], obwohl beide kaum Voraussetzungen oder exakte mathematische Definitionen angeben.

3. Suffiziente Statistiken im Fall, dass der Wertebereich einer Verteilung unabhängig vom Parameter θ ist.

   Dieser Abschnitt enthält Pitmans Beitrag zum Satz von Darmois-Koopman-Pitman. Nämlich Pitmans Herleitung der speziellen Gestalt einer Verteilungsfamilie bei Vorliegen einer Statistik mit maximaler Fisher-Information. Pitman verweist hier zu Beginn wieder auf R. A. Fisher, wonach in [Fis25, S.713] gezeigt wurde, dass eine Statistik mit maximaler Fisher-Information auch suffizient ist. Ist also T_1 eine vorliegende Statistik mit maximaler Fisher-Informatin nI bzgl. des Parameters θ, dann ist jede andere Statistik T_2 unabhängig von θ. D.h. T_2 enthält keine weitere Information über θ.

   Bemerkung: Eine Darstellung zu Statistiken, die die Informationsungleichung exakt erfüllen, findet sich in [LC98, S.121].

   Ohne detailierte Beschreibung setzt Pitman ausgehend von einer Zufallsstichprobe $(x_1,\ldots,x_n)$ und deren Produktdichte $\phi=\prod_{r=1}^{n} f(x_r)$

$$L=\log\phi=\sum\log f(x_r,\theta) \tag{4.27}$$

   und erhält durch die Ableitung des Logarithmus

$$\frac{\phi'}{\phi}=\frac{\partial L}{\partial\theta}=\sum\frac{f'(x_r,\theta)}{f(x_r,\theta)}=\Phi(T_1,\theta). \tag{4.28}$$

   T_1 ist unabhängig von θ daher wählt Pitman für θ einen zulässigen festen Wert, z.B. oBdA $\theta=0$ und dann lässt sich aus (4.28) T_1 explizit darstellen:

$$T_1=\psi\left(\sum g(x_r)\right).$$

Weiters erhält man durch Reparametrisierung

$$T = \sum g(x_r)$$

(das entspricht der natürlichen Parametrisierung bei exponentiellen Familien; vgl. [Rü14, S.68]). Dann hängt $\partial L/\partial\theta$ nur von T und θ ab:

$$\frac{\partial L}{\partial \theta} = H(T, \theta)$$

und daher ist nach der Kettenregel (mit Pitmans Schreibweise für $H'_T := \frac{\partial H}{\partial T}$)

$$\frac{\partial^2 L}{\partial\theta\partial x_r} = H'_T(T,\theta)\frac{\partial T}{\partial x_r}. \tag{4.29}$$

Die linke Seite von (4.29) hängt von x_r und θ ab, während $\frac{\partial T}{\partial x_r}$ nur von x_r abhängt, d.h.

$$H'_T(T,\theta)$$

hängt von x_r und θ ab. Da T eine symmetrische Funktion der Stichprobe $(x_1,\ldots,x_n)$ ist (siehe [Fis34, S.294]), und H'_T nur von T und θ abhängt, hängt H'_T nur von θ ab. So ergibt Integration nach T mit beliebigen Funktionen $p(\theta)$ und $p(\theta)$

$$H(T,\theta) = p(\theta)T + q(\theta) \tag{4.30}$$

$$\sum \frac{f'(x_r,\theta)}{f(x_r,\theta)} = p(\theta)\sum g(x_r) + q(\theta). \tag{4.31}$$

Und daraus folgt mit $p_1(\theta) := \int p(\theta)d\theta$, $q_1(\theta) := \int q(\theta)d\theta$ und einer Funktion $k(x_r)$

$$\frac{f'(x_r,\theta)}{f(x_r,\theta)} = p(\theta)g(x_r) + q(\theta)/n \tag{4.32}$$

$$\log f(x_r,\theta) = p_1(\theta)g(x_r) + q_1(\theta)/n + k(x_r), \tag{4.33}$$

sodass $f(x_r,\theta)$ in der Form

$$f(x_r,\theta) = u(\theta)h(x_r)e^{p_1(\theta)g(x_r)} \tag{4.34}$$

geschrieben werden kann. Und für die Werte $(x_1,\ldots,x_n)$ ergibt sich

$$\phi = \prod f(x_r,\theta) = u(\theta)^n e^{p_1(\theta)T}\prod h(x_r). \tag{4.35}$$

Das bedeutet, dass (unter den gegebenen Voraussetzungen) für eine feste Sta-

tistik T und wenn der Wertebereich nicht von θ abhängt, jede andere Statistik nicht von θ abhängt, d.h. T ist eine suffiziente Statistik. Eine Funktion, die nur von T abhängt, wie beispielsweise T_1 von oben, ist wieder eine suffiziente Statistik.

Im Anschluss beschreibt Pitman, dass bekannte Verteilungen wie etwa Poisson- oder Normalverteilung von der soeben Hergeleiteten „Form“ (Pitman verwendet noch keine spezielle Bezeichnung für exponentielle Familien) sind. Er verwendet auch (nach heutiger Bezeichnung) die kanonischer Form

$$f(x,\theta) = \frac{h(x)e^{\theta g(x)}}{K(\theta)}$$

um im Fall, dass der Wertebereich der Verteilung das Intervall (a,b) ist, $K(\theta)$ wie folgt darzustellen

$$K(\theta) = \int_a^b h(x)e^{\theta g(x)}dx \quad \text{bzw.} \quad K(\theta) = \sum_a^b h(x)e^{\theta g(x)}.$$

Und da in diesem Fall nur $K(\theta)$ von den Grenzen a und b abhängt, besitzen auch gestutzte (*truncated*) Verteilungen suffiziente Statistiken.

Den Abschluss bildet noch die Herleitung der charakteristischen Funktion für exponentielle Familien. Diese folgt aber im Wesentlichen [Fis34, S.294]. Eine moderne Darstellung für die momenterzeugende Funktion von exponentiellen Familien findet sich in [LC98, S.28f].

4. Suffiziente Statistiken im Fall, dass der Wertebereich einer Verteilung abhängig vom Parameter θ ist.

   Ausgehend vom Intervall (a,b) als Wertebereich einer Verteilung $f(x,\theta)$ mit Parameter θ, zeigt Pitman, dass suffiziente Statistiken i.A. (d.h. wenn $\exists h : a \leq h < b$) Funktionen der Ordnungsstatistiken sind. Ist z.B. die rechte Intervallgrenze b fest und die linke Intervallgrenze a abhängig von θ, d.h. a ist eine Funktion $a(\theta)$, dann ist eine suffiziente Statistik eine Funktion vom Minimum $x_0 := \min(x_1,\ldots,x_n)$ der Stichprobe. Eine Übersichtliche Darstellung für suffiziente Statistiken der nicht regulären Fälle gibt [Huz76, S.126 ff]. Ebenfalls in [Huz76, S. 101f] findet sich eine Kritik zu diesem Abschnitt von Pitmans Arbeit, vor allem hinsichtlich der Sprünge in Pitmans Ausführungen.

5. Die Fisher-Information im Fall, dass der Wertebereich einer Verteilung abhängig vom Parameter θ der Verteilung ist.

Auf diesen Abschnitt wird nicht näher eingegangen, da sein Inhalt nicht Teil des Darmois-Koopman-Pitman-Theorems ist. Jedoch erscheint folgender Aspekt, den Pitman hier behandelt, erwähnenswert:

Für eine stetige Verteilung mit geeigneten Voraussetzungen (siehe Definition 3.6) und mit Dichte $f(x,\theta)$ ergibt sich aus

$$\int_a^b f(x,\theta)dx = 1$$

durch Ableiten nach θ

$$\int_a^b f'dx - a'f(a,\theta) + b'f(b,\theta) = 0$$

und somit (vgl. (4.23) für die Definition von J)

$$J = \int_a^b f'dx = a'f(a,\theta) - b'f(b,\theta).$$

Das bedeutet in weiterer Folge, dass $J \neq 0$ (außer in Fällen, wenn z.B. $f(a,\theta) = 0 = f(b,\theta)$ und $J = 0$), und dass die drei Größen

$$\mathbb{E}\left\{\left(\frac{\partial}{\partial\theta}\log f\right)^2\right\}, \quad Varianz\left(\frac{\partial}{\partial\theta}f\right), \quad -\mathbb{E}\left\{\frac{\partial^2}{\partial\theta^2}\log f\right\}$$

i.A. nicht mehr den gleichen Wert aufweisen. Weiters hält Pitman (ohne Herleitung) fest, dass

$$\mathbb{E}\left\{\left(\frac{\partial}{\partial\theta}\log f\right)^2\right\} = nI + n(n-1)J^2.$$

D.h. im Fall $J \neq 0$ wächst der Informationsgehalt einer Verteilung schneller als die Stichprobengröße.

6. Bündel suffizienter Statistiken für einen Parameter

   Hier bespricht Pitman noch den Fall, dass, wenn keine einzelne suffiziente Statistik für einen Parameter existiert, es ein (minimales) Bündel resp. Menge von r Statistiken $T_1, T_2, \ldots, T_r$ geben kann, die zusammen die volle Information eines Parameters θ ausschöpfen. Eine genaue Darstellung findet sich in [Huz76].

   So bilden beispielsweise im nicht-regulären Fall, wenn eine Verteilung als Wertebereich das Intervall (a,b) besitzt und beide Grenzen vom Parameter θ abhängen, also $a(\theta), b(\theta)$ Funktionen von θ sind, das Minimum und das

Maximum der Stichprobe $(x_1, \ldots, x_n)$ ein minimales Paar suffizienter Statistiken für den Parameter θ. In [Huz76] werden Fälle, die nur von den Ordnungsstatistiken bzw. nur vom Maximum und Minimum abhängen, als reine nicht-reguläre Fälle *„pure type non-regular cases“*) bezeichnet (siehe [Huz76, S.126]).

Im Fall, dass z.B. nur die untere Grenze a vom Parameter θ abhängt, also $a(\theta) \leq x < b$ und die Verteilung von der Form

$$f(x, \theta) = \frac{h(x)e^{p(\theta)g(x)}}{K(\theta)}$$

ist, bilden das Minimum der Stichprobe $x_0 := \min(x_1, \ldots, x_n)$ und $T = \sum g(x_r)$ ein Paar suffizienter Statistiken. Dies ist ein Beispiel für ein minimales Paar von suffizienten Statistiken eines *„mixed type non-regular case“* aus [Huz76, S.126].

In regulären Fall, wenn also beide Intervallgrenzen a und b von θ unabhängig sind, und es ein (minimales) Paar suffizienter Statistiken U und V für den Parameter θ gibt, leitet Pitman die daraus resultierende Form der Verteilung analog zum Fall einer einzelen suffizienten Statistik her: Ausgehend von der Dichte ϕ der Stichprobe $(x_1, \ldots, x_n)$ mit der Funktion $K(x_r)$ unabhängig von θ

$$\phi = \Psi(U, V, \theta)K$$

bzw.

$$L = \log \phi$$

erhält man durch ableiten nach θ

$$\frac{\partial L}{\partial \theta} = \Psi(U, V, \theta).$$

Wählt man oBdA für θ zulässige Werte, so ergeben sich folgende Gleichungen:

$$\sum g_1(x_r) = \Psi(U, V, 0)$$
$$\sum g_2(x_r) = \Psi(U, V, 1).$$

Somit lässt sich erkennen, dass U und V nur von

$$\sum g_1(x_r) \quad \text{und} \quad \sum g_2(x_r)$$

abhängen. Setzt man weiters

$$T_1 = \sum g_1(x_r) \quad \text{und} \quad T_2 = \sum g_2(x_r)$$

dann hängt $\frac{\partial L}{\partial \theta}$ nur von T_1, T_2, θ ab, d.h.

$$\frac{\partial L}{\partial \theta} = H(T_1, T_2, \theta).$$

Es erfolgen die gleichen Umformungen wie im Fall einer einzelnen Statistik und da $\frac{\partial H}{\partial T_1}$ bzw. $\frac{\partial H}{\partial T_2}$ nur von θ abhängen, kann die Dichte in der Form

$$f(x, \theta) = \frac{h(x)e^{p_1(\theta)g_1(x)+p_2(\theta)g_2(x)}}{F(\theta)}$$

geschrieben werden. In diesem Fall bilden $T_1 = \sum g_1(x_r)$ und $T_2 = \sum g_2(x_r)$ ein Paar suffizienter Statistiken.

Zum Abschluss verweist Pitman noch darauf, dass sich in analoger Weise eine Darstellung für ein Bündel von r linear unabhängiger (*„algebraically independent“*) Statistiken $T_1, T_2, \ldots, T_r$ der Stichprobe $x_1, \ldots, x_n$ mit $r < n$ für den Parameter θ in der Form

$$f(x, \theta) = \frac{h(x)e^{\sum_{i=1}^{r} p_i(\theta)g_i(x)}}{F(\theta)} \tag{4.36}$$

angeben lässt, wobei $T_1, T_2, \ldots, T_r$ eine Menge suffizienter Statistiken für θ bildet. Im diesem Fall, wenn ϕ'/ϕ nur von $T_1, \ldots, T_r, \theta$ abhängen, wird die Informationsungleichung exakt, d.h. die Fisher-Information ist in diesem Fall nI.

Bemerkung: Die Voraussetzung, dass es sich um *„algebraically independent functions“* handelt, stellt sicher, dass die Anzahl r der T_r nicht verringert werden kann.

7. Es folgt noch eine kurze Schlussbemerkung, wo Pitman noch einmal die entscheidenden Punkte seiner Arbeit zusammenfasst.

5 Erweiterungen zum Satz von Darmois-Koopman-Pitman

5.1 Rezeption der Arbeiten von Darmois, Koopman und Pitman

Hier eine Auswahl von Arbeiten, die unmittelbar an die Arbeiten von Darmois, Koopman und Pitman anschließen bzw. diese erweitern.

5.1.1 E. W. Barankin und A. P. Maitra (1963)

Ein Vergleich der Arbeiten von Darmois, Koopman und Pitman hinsichtlich der Tragweite der Resultate ist im Vorwort der Arbeit „Generalization of the Fisher-Darmois-Koopman-Pitman Theorem on Sufficient Statistics“ [BM63] von E. W. Barankin und A. P. Maitra zu finden. Als Grundlage zur Besprechung im Text wird in [BM63] grob das Theorem von Darmois-Koopmam-Pitman folgendermaßen beschrieben: Ausgehend von einer Familie von Dichten einer Verteilung $p_0(\xi, \theta)$ auf der ξ-Achse, einem ν-dimensionalem Parametervektor $\theta := (\theta_1, \ldots, \theta_\nu)$ und einer n-dimensionalen Zufallsstichprobe $x := (x_1, \ldots, x_n)$, hat die Produktdichte $\prod_{i=1}^{n} p_0(x_i, \theta)$ suffiziente Statistiken von Dimension $s < n$, genau dann wenn die logarithmierte Produktdichte von der Form

$$b_0(\theta) + \psi_0(x) + \sum_{\lambda=1}^{r} b_\lambda(\theta)\psi_\lambda(x) \tag{5.1}$$

mit $r \leq s$ und $\nu < n$ ist.

Bemerkung: Mit der Darstellung ist eine exponentielle Familie von vollem Rang gemeint, d.h. u.a. sind die $\psi_\lambda(x)$ vollständig suffizient (siehe (2.8) bzw. [LC98, S.23])

Die Zusamenfassung in [BM63] ist dann wie folgt: Darmois postuliert sein Resultat mit $r = s$, d.h. die Frage nach der Minimalsuffizienz bleibt offen, und $s = \nu$, d.h. er geht davon aus, dass jeder Parameter durch eine einzelne Statistik geschätzt werden kann. Koopman nimmt implizit auch $s = \nu$ an, allerdings setzt er $r \leq s$, d.h. er betrachtet minimal suffiziente Statistiken. Bei Pitman ist wieder $r = s$ allerdings lässt er auch $s \neq \nu$ zu, falls Bündel von Statistiken für eine erschöpfende

Schätzung nötig sind, allerdings führt das auf nicht-reguläre Probleme (d.h. der Träger einer Dichte ist parameterabhängig) und i. A. nicht zu exponentiellen Familien. (vgl. [BM63, S.217])

Die Arbeit [BM63] bietet eine hervorragende Darstellung des Satzes von Darmois-Koopman-Pitman und erweitert die Annahme von Darmois, Koopman und Pitman, dass alle Elemente der Zufallsstichprobe dieselbe zugrunde liegende Verteilung resp. dieselbe Dichte $p(x_r, \theta)$ haben, auf den Fall, dass die einzelnen Dichten $p_m(x_r, \theta)$ lediglich einer gemeinsamen Verteilungsfamilie, einer exponentiellen Familie von Verteilungen, entstammen müssen, aber nicht unbedingt identisch verteilt sein müssen. Es werden für diesen Fall notwendige und hinreichende Bedingungen angegeben.

Diese Arbeit gibt eine ausführliche Darstellung der Voraussetzungen und führt präzise Beweise. Dies geht sogar soweit, dass die Autoren, um Klarheit zu schaffen, ihre früheren Arbeiten, die wichtige Vorarbeiten für die Beweisschritte enthalten, in einem eigenen Kapitel zusammenfassen und nicht nur darauf referenzieren! Der Beweis für den Satz von Darmois-Koopman-Pitman wird zuerst lokal (*„local result“*), d.h. für einen Punkt $x_0 \in \mathbb{R}^n$ geführt, und dann auf ein n-dimensionales offenes Intervall des $\mathbb{R}^n$ erweitert (*„global result“*). Die Resultate Koopmans (*„classical Koopman results“* [BM63, S.219]), also mit identen Dichten, folgen als Korollare im Anschluss.

Im Verlauf von [BM63] wird an mehreren Stellen auf die Arbeit von Koopman verwiesen, da die Art und Strenge der Beweisführung einander ähnlich sind. Beide Arbeiten beginnen mit einer Definition für suffiziente Statistiken. Allerdings ist die Definition Koopmans viel einschränkender als jene von [BM63, S.221]. Die Definition in [BM63] entspricht dem Faktorisierungskriterium (vgl. [LC98, S.35]), wohingegen Koopmans Definition einer Formalisierung des Suffzienzprinzips entspricht. Koopmans Voraussetzung an die Statistiken ist daher nur Stetigkeit, wohingegen [BM63] stetig differenzierbare Statistiken voraussetzt. Das hat aber zur Folge, dass Koopman einen Satz mit notwendigen und einen Satz mit hinreichenden Bedingungen erhält. Die Autoren von [BM63] können aufgrund der stärkeren Voraussetzungen Koopmans Resultate zu notwendigen und hinreichenden Bedingungen zusammenfassen (vgl. [BM63, S.242ff]).

5.1.2 Weitere Arbeiten

Die Arbeiten [Bro64, BNP68, Hip74] befassen sich mit verallgemeinernden Regularitätsbedingungen für den Satz von Darmois-Koopman-Pitman. Die Arbeit [BNP68] gibt einen übersichtlichen Beweis für den Satz von Darmois-Koopman-Pitman (siehe Satz 2.6. Interessant ist die Arbeit [Hip74], in der die Voraussetzung der stetigen partiellen Differnzierbarbeit der Dichten im Stichprobenraum auf lokale Lipschitz-

Stetigkeit reduziert werden konnte.

Im Artikel [And70] wird eine Version des Satzes von Darmois-Koopman-Pitman für diskrete Dichten resp. Wahrscheilichkeitsfunktionen für die Ordnung $k = 1$ gezeigt. Die Bedingungen für diesen Fall sind, dass auf dem Stichprobenraum der suffizienten Statistik eine Totalordnung definiert werden kann, die zusätzlich eine Regularitätsbedingung erfüllt (vgl. [And70, S. 1250]).

5.1.3 Betrachtungen zum Satz von Darmois-Koopman-Pitman aus historischer Sicht

Ein herausragendes Buch über die Geschichte der Statistik aus fachlicher Sicht ist [Pfa17] von Johann Pfanzagl. Nach Themengebieten aufgebaut, gibt es fachlich fundiert eine historische Einführung in die jeweiligen Abschnitte. Dort findet sich in der Einleitung zur asymptotischen Theorie eine Abgrenzung zur der klassischen parametrischen Statistik, die auch ein Licht auf exponentielle Familien aus heutiger Betrachtung wirft: *„Nonasymptotic theory, which flourished between 1940 and 1960, ended up as a collection of results that does not amount to a consistent theory. Important parts of nonasymptotic theory are, in fact, confined to exponential families. Even in this restricted domain, the applicability depends on accidental features of the model.“* [Pfa17, S.107]

Das Buch [Gor16] von P. Gorroochurn stellt die Entwicklung der Statistik in chronologischer Reihenfolge übersichtlich und gut nachvollziehbar dar. Präsentiert wird von S.593 - S.602 ein Abriss der Arbeiten von Fisher, Darmois, Koopman und Pitman im historischen Kontext. Allerdings werden hier nur Teile ihrer Arbeiten präsentiert, ohne Verweise oder mathematische Querverbindungen zur heutigen statistischen Theorie.

5.2 Ausblick

Ohne Anspruch auf Vollständikeit seien zum Abschluss noch einige Anwendungsfelder von exponentiellen Familien erwähnt:

Große Bedeutung haben suffiziente Statistiken und exponentielle Familien bei der Untersuchung von Lageparameter- und Skalenparameterfamilien. Eine umfassende Darstellung dazu findet sich in [Huz76].

Für exponentielle Familien existieren „optimale“ d.h. unverfälscht trennscharfe Tests. Hier spielen vollständige suffiziente Statistiken eine wesentliche Rolle (vgl. [LR05, S.110ff])

Des Weiteren haben Konzepte wie Suffizienz, Faktorisierungsatz, exponentielle Familien etc. weitreichende Anwendung in der Bayes-Statistik.

Aktuelle Anwendung finden exponentielle Familien in verallgemeinerten linearen Modellen (Generalized Linear Models, kurz GLM). Dabei handelt es sich um eine Erweiterung des klassischen linearen Regressionsmodells. Beim klassischen Regressionsmodell wird angenommen, dass der Fehlerterm normalverteilt ist. In einem verallgemeinerten linearen Modell kann diese Einschränkung auf exponentielle Familien erweitert werden.

Eine weitere zeitgemäße Anwendung haben exponentielle Verteilungsfamilien im *Machine Learning*, siehe dazu [Das11].

Allerdings treten klassische parametrische Verfahren gegenüber (computergestützten) asymptotischen, robusten oder parameterfreien Verfahren immer mehr in den Hintergrund.

Literaturverzeichnis

[Ald97] ALDRICH, John: R.A. Fisher and the making of maximum likelihood 1912-1922. In: Statist. Sci. 12 (1997), 09, Nr. 3, 162–176. – DOI 10.1214/ss/1030037906

[And70] ANDERSEN, Erling B.: Sufficiency and Exponential Families for Discrete Sample Spaces. In: Journal of the American Statistical Association 65 (1970), Nr. 331, 1248–1255. http://www.jstor.org/stable/2284291. – ISSN 01621459

[Arn06] ARNOLD, Steven F.: Sufficient Statistics. In: Encyclopedia of Statistical Sciences (2006). – DOI 10.1002/0471667196.ess2638.pub2. – Stand: 22.08.2019

[BD01] BICKEL, Peter J. ; DOKSUM, Kjell A.: Mathematical Statistics: Basic Ideas and Selected Topics. 2. New Jersey : Prentice-Hall, 2001. – ISBN 0–13–850363–X

[Bil95] BILLINGSLEY, Patrick: Probability and Measure. 3. New York : Wiley & sons, 1995. – ISBN 0–471–00710–2

[BM63] BARANKIN, Edward W. ; MAITRA, Ashok P.: Generalization of the Fisher-Darmois-Koopman-Pitman Theorem on Sufficient Statistics. In: Sankhya: The Indian Journal of Statistics, Series A 25 (1963), April, Nr. 3, 217–244. – ISSN 0581572X

[BN78] BARNDORFF-NIELSEN, Ole E.: Information and Exponential Families. New York : John Wiley & Sons Ltd, 1978. – ISBN 0–471–99545–2

[BN06a] BARNDORFF-NIELSEN, O.: Exponential Families. In: Encyclopedia of Statistical Sciences (2006). – DOI 10.1002/0471667196.ess0582.pub2. – Stand: 22.08.2019

[BN06b] BARNDORFF-NIELSEN, O. E.: General Exponential Families. In: Encyclopedia of Statistical Sciences (2006). – DOI 10.1002/0471667196.ess0296.pub2. – Stand: 22.08.2019

[BNP68] BARNDORFF-NIELSEN, O. E. ; PEDERSEN, K.: Sufficient Data Reduction and Exponential Families. In: MATHEMATICA SCANDINAVICA 22 (1968), Jun., 197-202. – DOI 10.7146/math.scand.a–10883

[Bro64] BROWN, L.: Sufficient Statistics in the Case of Independent Random Variables. In: The Annals of Mathematical Statistics 35 (1964), Nr. 4, 1456–1474. http://www.jstor.org/stable/2238285. – ISSN 00034851

[Bro86] BROWN, Lawrence D.: Fundamentals of Statistical Exponential Families. Hayward, California : IMS(Institute of Mathematical Sciencies), 1986. – ISBN 0–940600–10–2

[CB01] CASELLA, George ; BERGER, Roger: Statistical Inference. 2. Duxbury : Duxbury Resource Center, 2001. – ISBN 0534243126

[Dar35] DARMOIS, M. G.: Sur les lois de probabilites a estimation exhaustive. In: C.R. Acad. Sci. Paris 200 (1935), S. 1265–1266

[Dar36] DARMOIS, M. G.: L'emploi des observations statistiques, méthodes d'estimation. Paris : Hermann, 1936

[Das11] DASGUPTA, Anirban: Probability for Statistics and Machine Learning. 1. New York : Springer, 2011. – ISBN 978–1–4419–9634–3

[Den67] DENNY, J. L.: Sufficient Conditions for a Family of Probabilities to be Exponential. In: Proceedings of the National Academy of Sciences of the USA 57 (1967), Nr. 5, S. 1184–1188. – DOI 10.1073/pnas.57.5.1184

[Dug61] DUGUE, D.: Georges Darmois, 1888-1960. In: The Annals of Mathematical Statistics 32 (1961), 357–360. – DOI 10.1214/aoms/1177705046

[DY00] DYNKIN, Evgenij B. (Hrsg.) ; YUSHKEVICH, A. A. (Hrsg.): Selected papers of E. B. Dynkin. Providence, RI : American Mathematical Soc. [u.a.], 2000 . – ISBN 0–8218–1065–0

[Dyn51] DYNKIN, Evgenij B.: Necessary and Sufficient Statistics for a Family of Probability Distributions. In: [DY00], S. 393–416

[Efr06] EFRON, Brad: Special Exponential Families. In: Encyclopedia of Statistical Sciences (2006). – DOI 10.1002/0471667196.ess1162.pub2. – Stand: 22.08.2019

[Els11] ELSTRODT, Jürgen: Maß- und Integrationstheorie. 7. Berlin Heidelberg : Springer, 2011. – ISBN 978–3–642–17904–4

[Fel14] FELSENSTEIN, Klaus: Mathematische Statistik. Vorlesungsskriptum, 2014. – TU Wien

[Fis22] FISHER, Ronald A.: On the Mathematical Foundations of Theoretical Statistics. In: Phil. Trans. R. Soc. Lond. A 222 (1922), S. 309–368. – DOI 10.1098/rsta.1922.0009

[Fis25] FISHER, Ronald A.: Theory of Statistical Estimation. In: Mathematical Proceedings of the Cambridge Philosophical Society 22 (1925), Nr. 5, S. 700–725. – DOI 10.1017/S0305004100009580

[Fis34] FISHER, Ronald A.: Two New Properties of Mathematical Likelihood. In: Phil. Trans. R. Soc. Lond. A 144 (1934), S. 285–307. – DOI 10.1098/rspa.1934.0050

[Fis50a] FISHER, R. A.: Contributions to Mathematical Statistics. New York : John Wiley & Sons, 1950

[Fis50b] FISHER, R. A.: Statistical Methods for Research Workers. 11. Edinburgh : Oliver and Boyd, 1950

[Fis56] FISHER, R. A.: Statistical Methods and Scientific Inference. 1. New York : Hafner Publishing Company, 1956

[Gho88] GHOSH, J. K. (Hrsg.): Statistical Information and Likelihood: A Collection of Critical Essays by Dr. D. Basu. NewYork : Springer, 1988 (Lecture Notes in Statistics 45). – ISBN 978–0–387–96751–6

[Gor16] GORROOCHURN, Prakash: Classic Topics on the History of Modern Mathematical Statistics: From Laplace to More Recent Times. Hoboken, New Jersey : John Wiley & Sons, 2016. – ISBN 978–1–119–12792–5

[Han15] HAND, David J.: From evidence to understanding: a commentary on Fisher (1922) 'On the mathematical foundations of theoretical statistics. In: Philosophical Transactions of the Royal Society A 373 (2015). – DOI 10.1098/rsta.2014.0252. – ISSN 1364–503X

[Hip74] HIPP, Christian: Sufficient Statistics and Exponential Families. In: The Annals of Statistics 2 (1974), Nr. 6, 1283–1292. http://www.jstor.org/stable/2958344. – ISSN 00905364

[HMC12] HOGG, Robert V. ; MCKEAN, Joseph W. ; CRAIG, Allen T.: Introduction to Mathematical Statistics. 7. International : Pearson, 2012. – ISBN 978–0–321–79543–4

[Huz76] HUZURBAZAR, Vasant S.: Sufficient Statistics. 1. New York : Marcel Dekker, 1976. – ISBN 0–8247–6296–7

[Kee10] KEENER, Robert W.: Theoretical Statistics. NewYork : Springer, 2010. – ISBN 978–0–387–93838–7

[Kle08] KLENKE, Achim: Wahrscheinlichkeitstheorie. 2. Berlin Heidelberg : Springer, 2008. – ISBN 978–3–540–76317–8

[Koo36] KOOPMAN, Bernard: On Distributions Admitting a Sufficient Statistic. In: Transactions of the American Mathematical Society 39 (1936), Nr. 3, S. 399–409. – DOI 10.2307/1989758

[Kru80] KRUSKAL, William: The Significance of Fisher: A Review of R.A. Fisher: The Life of a Scientist. In: Journal of the American Statistical Association 75 (1980), Nr. 372, 1019–1030. http://www.jstor.org/stable/2287199. – ISSN 01621459

[Kus14] KUSOLITSCH, Norbert: Maß- und Wahrscheinlichkeitstheorie. 2. Berlin Heidelberg : Springer, 2014. – ISBN 978–3–642–45386–1

[LC98] LEHMANN, E.L. ; CASELLA, George: Theory of Point Estimation. 2. New York : Springer, 1998. – ISBN 0–387–98502–6

[LR05] LEHMANN, E.L. ; ROMANO, Joseph P.: Testing Statistical Hypotheses. 3. New York : Springer, 2005. – ISBN 0–387–98864–5

[Mar15] MARTIN, Ryan: Lecture Notes on Advanced Statistical Theory. http://homepages.math.uic.edu/~rgmartin/Teaching/Stat511/Notes/511notes.pdf, 2015. – Stand: 22.08.2019

[Mar16] MARTIN, Ryan: Lecture Notes on Statistical Theory. http://homepages.math.uic.edu/~rgmartin/Teaching/Stat411/Notes/411notes.pdf, 2016. – Stand: 22.08.2019

[Moo71] MOORE, D. S.: Maximum Likelihood and Sufficient Statistics. In: The American Mathematical Monthly 78 (1971), Nr. 1, 50–52. http://www.jstor.org/stable/2317488. – ISSN 00029890, 19300972

[Mor82] MORSE, Philip M.: In Memoriam: Bernard Osgood Koopman, 1900-1981. In: Operations Research 30 (1982), Nr. 3, viii–427. http://www.jstor.org/stable/170181. – ISSN 0030364X, 15265463

[Mor06] MORRIS, Carl N.: Natural Exponential Families. In: Encyclopedia of Statistical Sciences (2006). – DOI 10.1002/0471667196.ess1759.pub2. – Stand: 22.08.2019

[Nor72] NORDEN, R. H.: A Survey of Maximum Likelihood Estimation. In: International Statistical Review / Revue Internationale de Statistique 40 (1972), Nr. 3, 329–354. http://www.jstor.org/stable/1402471. – ISSN 03067734, 17515823

[Nor73] NORDEN, R. H.: A Survey of Maximum Likelihood Estimation: Part 2. In: International Statistical Review / Revue Internationale de Statistique 41 (1973), Nr. 1, 39–58. http://www.jstor.org/stable/1402786. – ISSN 03067734, 17515823

[Oli08] OLIVE, David J.: Using Exponential Families in an Inference Course. http://lagrange.math.siu.edu/Olive/siexpf.pdf, 2008. – Stand: 30.08.2015

[Oli14] OLIVE, David J.: Statistical Theory and Inference. Heidelberg New York : Springer International Publishing, 2014. – ISBN 978–3–319–04972–4

[Ols40] OLSHEVSKY, Louis: Two Properties of Sufficient Statistics. In: The Annals of Mathematical Statistics 11 (1940), Nr. 1, 104–106. http://www.jstor.org/stable/2235975. – ISSN 00034851

[Pfa17] PFANZAGL, Johann: Mathematical Statistics. 1. Berlin Heidelberg : Springer, 2017. – ISBN 978–3–642–31083–6

[Pit36] PITMAN, E. J. G.: Suffcient Statistics and Intrinsic Accuracy. In: Mathematical Proceedings of the Cambridge Philosophical Society 32 (1936), Nr. 4, S. 567–579. – DOI 10.1017/S0305004100019307

[Pit17] PITMAN, E. J. G.: Some Basic Theory for Statistical Inference: Monographs on Applied Probability and Statistics. 1. Chapman & Hall/-CRC, 2017. – ISBN 978–1–315–89767–7

[Rü14] RÜSCHENDORF, Ludger: Mathematische Statistik. Berlin Heidelberg : Springer, 2014. – ISBN 978–3–642–41996–6

[RS86] ROMANO, Joseph P. ; SIEGEL, Andrew F.: Counterexamples in Probability and Statistics. 1. Monterey : Wadsworth & Brooks / Cole, 1986. – ISBN 0–534–05568–0

[Sch15] SCHILLING, René L.: Maß und Integral. Berlin : DeGruyter, 2015. – ISBN 978–3–11–034814–9

[Spr94] SPRENT, P.: E. J. G. Pitman, 1897-1993. In: Journal of the Royal Statistical Society. Series A (Statistics in Society) 157 (1994), Nr. 1,

153–154. http://www.jstor.org/stable/2983511. – ISSN 09641998, 1467985X

[Sti73] STIGLER, Stephen M.: Studies in the History of Probability and Statistics. XXXII: Laplace, Fisher and the Discovery of the Concept of Sufficiency. In: Biometrika 60 (1973), Nr. 3, 439–445. http://www.jstor.org/stable/2334992. – ISSN 00063444

[Zac81] ZACKS, Shelemyahu: Parametric Statistical Inference. 1. Oxford : Elsevier Ltd, Pergamon Press, 1981. – ISBN 978–0–08–026468–4