Institut für Statistik

Ludwig-Maximilians-Universität
München

Bachelorarbeit

Evidential frameworks zur statistischen Inferenz mit Anwendung auf die Analyse von Einflüssen auf Sectio-Raten

Autor: *Jui Andreas Tang*

BetreuerIn: *Almond Stöcker & Prof. Dr. Sonja Greven*

12. April 2018

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit eigenständig und ohne fremde Hilfe angefertigt habe. Textpassagen, die wörtlich oder dem Sinn nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, den 12. April 2018

Jui Andreas Tang

Inhaltsverzeichnis

1 Einleitung

Aufgrund der rasanten Entwicklung in der Technik gewinnen Informationen und Daten in der heutigen Zeit immer mehr an Bedeutung. Gleichzeitig steigen die bestehenden Datenmengen rapide an, da anhaltend neue Daten generiert und gesammelt werden. Um diese Fülle an Informationen adäquat zu handhaben, wird dabei die Statistik herangezogen.
Im Zuge dessen ist es nicht überraschend, dass die Statistik, insbesondere die evidenzbasierte Statitik, während der letzten Jahrhundertwende immer mehr an Bedeutung gewann. Das Heranziehen und Verfeinern von Elementen aus verschiedenen statistischen Inferenzkonzepten ist eine statistisch moderne Darstellung, mit der wir problemlos unterschiedlichste Modellanalysen durchführen können, wie unter anderem die Unsicherheit eines Modells, der Vergleich von verschiedenen Modellen, die Schätzung von Parametern und deren Unsicherheiten. Deshalb und wegen vieler anderer Gründe können wir behaupten, dass die evidenzbasierte Statistik momentan eine essentielle Rolle für die Wissenschaft im 21. Jahrhundert einnimmt.(Taper und Ponciano; 2016, Abstract)
Im wissenschaftlichem Umfeld werden statistische Methoden für eine sinnvolle Interpretation der Daten verwendet. Die Statistik bietet hierbei Möglichkeiten an, um auf effiziente Weise objektive Alternativen neben der eigenen Beurteilung zu finden, damit es möglich ist, die Evidenz aus Untersuchungs- und Beobachtungsstudien angemessen zu deuten.(Royall; 1997, Preface)
Eine dieser Wissenschaften, in den die evidenzbasierte Statistik eine zentrale Rolle spielt, ist die Medizin. Im Rahmen von Untersuchungen zur Wirksamkeit von medizinischen Methoden oder Medikamenten werden Ergebnisse meist anhand der Evidenz erschlossen. Mit Hilfe dieser evidenzbasierten Möglichkeiten werden subjektive Intuitionen und unsystematische klinische Studien vermieden, damit rationale Entscheidungsfindungen getroffen werden können.(Müllner; 2005, Kapitel 1), (Guyatt et al.; 1992, Abstract)
Trotz der Bedeutung der Evidenz weisen die gängigen statistischen Methoden, die für den Zweck der Evidenzbestimmung genutzt werden wie unter anderem der Hypothesentest, kein konkret definiertes Evidenzkonzept auf. Ferner liefern sie dadurch keine Antwort

auf die grundlegenden Fragen, wann es richtig wäre zu sagen, dass die gegebenen Daten eine Evidenz zugunsten einer Hypothese gegenüber einer anderen aufweist oder ob wir ein objektives Maß haben, um die Stärke dieser Evidenz auszudrücken.(Royall; 2000, Kapitel 1)

Einen möglichen Ansatz, um dieses Problem der Ungenauigkeit zu bewältigen, bietet das *evidential framework* von Jeffrey D. Blume an. Dieses allgemeine Framework soll uns den Vergleich und die Evaluation statistischer Paradigmen ermöglichen, die augenscheinlich die Stärke der statistischen Evidenz in den Daten messen. Dabei soll die Evidenz nicht mehr nur auf einen Wert beschränkt werden, sondern wird in drei essentiellen Größen aufgeteilt. Im Folgenden wird dargelegt werden, wie jede einzelne Größe für das Verständnis und die Bewertung der statistischen Evidenz relevant ist. Außerdem wird sich zeigen, dass das Fehlen eines wohldefinierten Frameworks zu verschiedenen Kontroversen führen kann. Das *evidential framework* wird auf einen medizinischen Datensatz angewendet, um den Einfluss von verschiedenen Faktoren auf die Sectiorate herauszufinden.(Blume; 2011, Kapitel 1)

2 Statistische Inferenzkonzepte

Ein wichtiger Nutzen, den wir aus statistischen Inferenzkonzepten ziehen können, ist die Konstruktivität dieser Konzepte, die uns eine universelle Anwendbarkeit ermöglicht. Neben der bekannten klassischen bzw. frequentistischen Inferenz gibt es zum einen die Likelihood-Inferenz und zum anderen die Bayes-Inferenz. Der Kern des baysianischen Konzepts ist es, die Likelihood-Funktion mit Vorwissen zu verbinden, um daraus neue Erkenntnisse zu ziehen.(Held; 2008, Kapitel 1.1)
Es gibt in breiterem Sinne drei Problembereiche, für die wir statistische Inferenzkonzepte benötigen. Der Erste ist das *Schätzproblem*[1]. Unter der Bedingung, dass eine bestimmte Modellannahme existiert und die Daten gegeben sind, wollen wir versuchen, Aussagen über den unbekannten Modellparametern zu treffen, d.h. wir möchten hierbei diese Parameter schätzen. Der zweite Bereich ist das *Modellwahlproblem*[1], bei dem wir das Modell aus verschiedenen Modellen herausfinden wollen, welches die gegebenen Daten am besten beschreibt. Der letzte Problembereich ist das *Prognoseproblem*[1]. Hier interessieren wir uns dafür, die Erkenntnisse aus den vorliegenden Daten zu nutzen, um zukünftige Beobachtungen sinnvoll zu prognostizieren.(Held; 2008, Kapitel 1.1)
Im Laufe dieses Kapitels werden sowohl die Likelihood-Inferenz als auch die Bayes-Inferenz genauer betrachtet, da diese neben der frequentistischen Inferenz die Grundlage für das *evidential framework* im nächstem Kapitel bilden.

2.1 Likelihood-Inferenz

Eine der bekanntesten Methoden zur statistischen Inferenz ist die von Sir Ronald A. Fisher eingeführte Likelihood-Inferenz. Dabei kann der englische Begriff *Likelihood* am ehesten mit „Plausabilität“ übersetzt werden. Die Basis dieses Inferenzkonzepts bildet die Likelihood-Funktion.(Held; 2008, Kapitel 2)

[1]Ein Begriff, der vom Autor Held eingeführt wurde.

2.1.1 Likelihood-Funktion

Wir nehmen an, dass $X = x$ die beobachtete Realisation einer Zufallsvariable X mit dazugehöriger Dichtefunktion $f(x|\theta)$ ist. Die Funktion $f(x|\theta)$ beschreibt die Verteilung der Zufallsvariable X für einen festen Parameter θ. Das Ziel besteht darin, Aussagen über den unbekannten Parameter θ aus dem Parameterraum Θ zu folgern, wobei die Funktion $f(x|\theta)$ bekannt ist. Die Likelihood-Funktion mit festem x

$$L(\theta) = f(x|\theta), \qquad \theta \in \Theta$$

bildet hierbei die Hauptgröße.(Held; 2008, Kapitel 2.1), (Held und Bové; 2014, Kapitel 2.1)

2.1.2 Law of Likelihood

Die statistische Analyse hat in der Wissenschaft die wichtige Aufgabe, die Evidenz aus den beobachteten Daten zu deuten. Obwohl es dafür momentan durchaus gängige Methoden gibt, wie beispielsweise den Hypothesentest oder die Interpretation der Konfidenzintervalle, beinhaltet die Theorie dieser Methoden kein konkretes Evidenzkonzept und kann somit nicht die zentrale Frage beantworten, wann die gegebenen Daten eine evidente Unterstützung einer statistischen Hypothese gegenüber einer anderen repräsentiert. Diese Unzulänglichkeit führt zu einer Kontroverse über die ordnungsgemäße Anwendung und Interpretation der p-Werte. Doch die *Law of Likelihood* füllt die Lücke im fehlendem Evidenzkonzept.(Royall; 2000, Kapitel 1)
Die Antwort auf die grundlegende Frage, wie statistische Daten als Evidenz zu interpretieren sind, liefert uns die Definition der *Law of Likelihood*:
Wenn eine Hypothese H_1 impliziert, dass eine Zufallsvariable X den Wert x mit der dazugehörigen Wahrscheinlichkeit $f_1(x)$ annimmt, während eine andere Hypothese H_2 impliziert, dass die Wahrscheinlichkeit $f_2(x)$ ist, dann ist die Beobachtung $X = x$ eine Evidenz für die Hypothese H_1 gegenüber H_2, falls $f_1(x) > f_2(x)$ gilt. Dabei ist die Likelihood Ratio $\frac{f_1(x)}{f_2(x)}$ ein Maß für die Stärke dieser Evidenz.(Royall; 2000, Kapitel 1.1)
Daraus ergibt sich, dass die *Law of Likelihood* eine Leitlinie zur Interpretation der statistischen Daten als Evidenz ist. Sobald durch dieses Axiom eine Hypothese über einer Anderen steht, bedeutet es, dass die bevorzugte Hypothese eine genauere Prädiktion liefert, d.h. sie erzielt eine größere Wahrscheinlichkeit zu dem beobachteten Punkt x. Dadurch wird nicht die Frage beantwortet, ob die Evidenz für oder gegen eine einzelne Hypothese ist, sondern es gibt uns eine Interpretationsweisung, wie wir die Evidenz

für eine Hypothese gegenüber einer anderen deuten sollen. Wir können festhalten, dass die *Law of Likelihood* eine objektive Evaluation der Daten als Evidenz unabhängig von Vorwissen ermöglicht.(Royall; 2000, Kapitel 1.1)

2.1.3 Likelihood Principle

Seien ein Wahrscheinlichkeitsmodell für eine Zufallsvariable X, dessen Verteilungsfamilie durch den Parameter θ indiziert ist, und eine Beobachtung, die eine Likelihood-Funktion $L(\theta)$ erzeugt, gegeben. Mittels der *Law of Likelihood* erhält diese Funktion ihre Bedeutung, d.h. für zwei Parameterwerte θ_1 und θ_2 misst deren Likelihood Ratio $\frac{L(\theta_1)}{L(\theta_2)}$ die Evidenzstärke $X = x$ zugunsten von θ_1 gegenüber θ_2.(Royall; 2000, Kapitel 1.2)
Die Definition der *Likelihood Principle* lautet wie folgt:
Nachdem $X = x$ bereits beobachtet wurde, sind alle Informationen über θ in der Likelihood-Funktion für θ enthalten. Sofern zwei Likelihood-Funktion für θ proportional zueinander sind, enthalten sie zudem die selbe Information über θ.(Berger et al.; 1988)
Nehmen wir nun an, dass wir zwei Datenszenarios mit erhobenen Daten zu statistischer Evidenz haben. Diese erzeugen eine äquivalente statistsche Evidenz, was bedeutet, dass alle Likelihood Ratios gleich sein müssen. Im Umkehrschluss bedeutet es aber auch, dass beide Likelihood-Funktion aus den Datenszenarios identisch sind. Daraus können wir den Schluss ziehen, dass für jedes Paar von θ_1 und θ_2 die Stärke der Evidenz zugunsten von θ_1 gegenüber θ_2 in beiden Szenarien gleich ist. Mit Hilfe der Definition der *Likelihood Principle* können wir schlussfolgern:
Zwei Fälle von statistischer Evidenz sind äquivalent und haben die selbe evidente Bedeutung, wenn und nur wenn sie die selbe Likelihood-Funktion erzeugen.
Somit dient die Likelihood-Funktion als mathematische Verkörperung der statistischen Evidenz an sich und die Likelihood Ratio misst die dazugehörige Evidenzstärke.(Royall; 2000, Kapitel 1.2), (Birnbaum; 1962, Kapitel 5)

2.2 Bayes-Inferenz

Im Gegensatz zur klassischen Statistik hat die bayesianische Variante den Vorteil, dass sie auf intuitiver Weise anschaulicher und einfacher zu begründen ist. Probleme im Bereich der Hypothesenprüfung oder der Bereichsschätzung, die in der frequentistischen Statistik nicht zu lösen sind, können mit der Bayes-Statistik entwirrt werden. Der Grund dafür ist, dass das bayesianische Konzept auf dem Bayes-Theorem basiert, wodurch wir den

unbekannten Parametern Wahrscheinlichkeitsverteilungen zuordnen können. Alle Fragestellungen der Parameterschätzung, der Hypothesenprüfung und der Bereichsschätzung werden auf Basis des Bayes-Theorems durchgeführt.(Koch; 2000, Kapitel 1)
In der klassischen Statistik betrachten wir die Daten X als zufällig. Wir interessieren uns hier für die frequentistischen Eigenschaften von daraus abgeleiteten Statistiken, insbesondere von möglichen Punkt- und Intervallschätzern. Der aus dem Wahrscheinlichkeitsmodell gegebene Parameter θ ist in der klassischen Inferenz unbekannt, aber fest, weshalb er hier keine Zufallsvariable bildet. Im Gegensatz dazu ist der Parameter θ im bayesianischem Pendant eine Zufallsvariable mit einer Priori-Verteilung $f(\theta)$. Nach der Untersuchung der Daten $X = x$ ist vor allem die Betrachtung der Posteriori-Verteilung $f(\theta|x)$ von großer Wichtigkeit. Beide Verteilungen nehmen in der Bayes-Inferenz einen sehr hohen Stellenwert ein, weshalb wir im Laufe dieses Kapitels näher auf die Priori- und die Posteriori-Verteilung eingehen werden.(Held; 2008, Kapitel 5)

2.2.1 Bayes-Theorem

Der Satz von Bayes bildet das zentrale Element in der Bayes-Statistik. Nehmen wir an, dass Ω eine Grundmenge mit disjunkten Zerlegungen $A_1, A_2, ..., A_k \subset \Omega$ ist und für die Wahrscheinlichkeiten $P(A_i) > 0$ und $P(B|A_i) > 0$ gilt, wobei $i = 1, ..., k$ und $B \subset \Omega$ gilt. Nun interessiert uns die Wahrscheinlichkeit von A_i unter der Bedingung einer bereits bekannten Information für Ereignis B. Unter den gegebenen Annahmen kann die bedingte Wahrscheinlichkeit als

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)}, \qquad i = 1, ..., k \tag{2.1}$$

dargestellt werden. Ferner gilt mit dem Satz von der totalen Wahrscheinlichkeit für

$$P(B) = \sum_{i=1}^{k} P(B|A_i) \cdot P(A_i),$$

wodurch wir den Satz von Bayes auch in der Form

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)}, \qquad i = 1, ..., k$$

für ein bestimmtes Ereignis darstellen können. Hierbei ist die Priori-Wahrscheinlichkeit $P(A_i)$ und die Posteriori-Wahrscheinlichkeit ist dargestellt als $P(A_i|B)$. Das heißt, dass

$P(A_i)$ die Wahrscheinlichkeit für das Aufkommen des Ereignisses A_i angibt, wobei wir hier keine Informationen über das Ereignis B besitzen. Sobald B eintritt, nehmen wir diese Erkenntnis, die die Priori-Wahrscheinlichkeit verändert, als „Hintergrundwissen“ auf und erhalten dadurch die Posteriori-Wahrscheinlichkeit $P(A_i|B)$.(Fahrmeir, Künstler, Pigeot und Tutz; 2007, Kapitel 4.6 & 4.7), (Koch; 2000, Kapitel 2.1.8)

2.2.2 Posteriori-Verteilung

Die Posteriori-Verteilung umfasst die gesamte Dateninformation über den unbekannten Parameter θ nach der Beobachtung der Daten $X = x$, weshalb sie die wichtigste Größe in der Bayes-Statistik bildet.(Held und Bové; 2014, Kapitel 6.2) Die Definition der Posteriori-Verteilung lautet wie folgt:
Gegeben sei die Beobachtung x einer Zufallsvariable bzw. eines Zufallsvektors X mit der Dichtefunktion $f(x|\theta)$. Nachdem die Festlegung einer Priori-Verteilung mit der Dichtefunktion $f(\theta)$ erfolgt ist, ergibt sich aus dem Satz von Bayes (siehe Kapitel 2.2.1) bei einem stetigem Parameterraum Θ die Dichtefunktion

$$f(\theta|x) = \frac{f(x|\theta) \cdot f(\theta)}{\int\limits_{\Theta} f(x|\theta) \cdot f(\theta) d\theta}$$

der Posteriori-Verteilung, wobei $f(x|\theta)$ die Likelihood-Funktion $L(\theta)$ darstellt. Nach der Integration des Nenners

$$\int f(x|\theta) \cdot f(\theta) d\theta = \int f(x|\theta) d\theta = f(x)$$

können wir feststellen, dass der Nenner nicht von θ abhängt. Daraus ergibt sich die Erkenntnis, dass die Posteriori-Verteilung mit der Proportionalitätskonstante $\frac{1}{f(x)}$ zum Produkt von Priori-Verteilung und Likelihood proportional ist, also

$$\text{Posteriori-Verteilung} \propto \text{Priori-Verteilung} \cdot \text{Likelihood}.$$

Eine weitere Notationsform ist:

$$f(\theta|x) \propto f(\theta) \cdot f(x|\theta)$$

wobei auch hier die Proportionalitätskonstante $\frac{1}{f(x)}$ ist und die Eigenschaft der Dichte $\int f(\theta|x) d\theta = 1$ gegeben sein muss.(Held; 2008, Definition 5.1)

2.2.3 Priori-Verteilung

Die Bayes-Statistik ermöglicht es uns mittels einer festgelegten Priori-Verteilung Aussagen über die Wahrscheinlichkeit unbekannter Parameter zu treffen. Obwohl die Festlegung der Priori-Verteilung im Allgemeinen eher subjektiv ist, gibt es einige Methoden, um diesen Grad an Subjektivität zu verringern. Der Grund dafür ist, dass die Priori-Verteilung definitionsgemäß vor der Beobachtung bereits festgelegt sein muss. Würden wir diese Festlegung erst nach der Beobachtung der zu analysierenden Daten durchführen, wäre dadurch die logische Kohärenz verletzt. Die Frage ist jedoch, wie wir diese Verteilung für differenzierte Anwendungen mit einer bestimmten Likelihood-Funktion festlegen können. Für diese Verteilungswahl existieren mehrere Möglichkeiten. (Held; 2008, Kapitel 5.2), (Gelman und Hennig; 2017, Kapitel 5.3)

Konjugierte Priori-Verteilung

Wir versuchen die Priori-Verteilung so zu wählen, dass die resultierende Posteriori-Verteilung einer gängigen Verteilungsfamilie folgt. Im Idealfall nimmt die Posteriori-Verteilung die selbe Verteilungsklasse an wie die Priori-Verteilung. (Held; 2008, Kapitel 5.2.1) Man spricht dann von einer konjugierten Priori-Verteilung mit folgender Definition:
Wir nehmen an, dass $L(\theta) = f(x|\theta)$ eine Likelihood-Funktion basierend auf der Beobachtung $X = x$ ist. Eine Klasse $\mathcal{G}$ von Verteilungen heißt *konjugiert bezüglich* $L(\theta)$, wenn für alle x die Posterior-Verteilung $f(\theta|x)$ in $\mathcal{G}$ ist, wann immer die Prior-Verteilung $f(\theta)$ ebenfalls in $\mathcal{G}$ enthalten ist.(Held; 2008, Definition 5.5), (Held und Bové; 2014, Definition 6.5)
Daraus ergibt sich, dass die Menge $\mathcal{G} = \{\text{alle Verteilungen}\}$ immer zu einer beliebigen Likelihood-Funktion $L(\theta)$ konjugiert. Da dieses Vorgehen in der Praxis wenig sinnvoll ist, werden in der Regel kleinere Mengen $\mathcal{G}$ gesucht.(Held; 2008, Kapitel 5.2.1)
Um dieses Vorgehen genauer darzustellen, illustrieren wir es anhand einer Binomialverteilung. Sei $X|\pi \sim \mathcal{B}(n, \pi)$. Wenn wir nun die Beta-Verteilung für π als Priori-Verteilung annehmen, ist $\pi \sim Be(\alpha, \beta)$ konjugiert bezüglich der Likelihood-Funktion $L(\pi)$, da die Posteriori-Verteilung ebenfalls wieder Beta-verteilt, $\pi|x \sim Be(\alpha+x, \beta+n-x)$, ist. Denn für *a priori* $\pi \sim Be(\alpha, \beta)$ und $\alpha, \beta > 0$ ist die Likelihood-Funktion

$$L(\pi) = f(x|\pi) = \begin{pmatrix} n \\ x \end{pmatrix} \pi^x (1-\pi)^{n-x}, \qquad x = 0, 1, ..., n$$

und für die Priori-Verteilung gilt

$$f(\pi) = \frac{1}{B(\alpha, \beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}, \qquad 0 < \pi < 1.$$

Für die Dichtefunktion der Posteriori-Verteilung ergibt sich somit (vgl. Kapitel 2.2.2)

$$\begin{aligned} f(\pi|x) &\propto L(\pi) \cdot f(\pi) \\ &\propto \pi^x(1-\pi)^{n-x} \cdot \pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &= \pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}, \end{aligned}$$

d.h. die Posteriori-Verteilung folgt der Beta-Verteilung, jedoch mit anderen Parametern: $\pi|x \sim Be(\alpha + x, \beta + n - x)$.(Held; 2008, Beispiel 5.1 & 5.3)

Uneigentliche und Nichtinformative Priori-Verteilungen

Wollen wir hingegen den Einfluss durch die Wahl der Priori-Verteilung minimieren, müssen wir eine Verteilung mit sehr großer Varianz wählen. Im Extremfall kann dies dazu führen, dass die gewählte Priori-Verteilung nicht mehr integrierbar ist. In diesem Fall spricht man von einer uneigentlichen Priori-Verteilung. Doch solange die dazugehörige Posteriori-Verteilung integrierbar bleibt, ist die Verwendung von uneigentlichen Priori-Verteilungen durchaus möglich.(Held; 2008, Kapitel 5.2.2) Eine uneigentliche Priori-Verteilung mit der Dichtefunktion $f(\theta) \geq 0$ ist gegeben, wenn

$$\int_\Theta f(\theta)d\theta = \infty \quad \text{oder} \quad \sum_{\theta \in \Theta} f(\theta) = \infty$$

für einen entsprechend stetigen oder diskreten Parameter θ gilt.(Held und Bové; 2014, Definition 6.6)
Mit der Absicht so wenig Hintergrundwissen wie möglich in die Wahl der Priori-Verteilung $f(\theta)$ einfließen zu lassen, gibt es den naiven Ansatz, für den Parameter θ eine stetige Gleichverteilung anzunehmen. Dies ist die sogenannte nichtinformative Priori-Verteilung. Es besteht jedoch die Möglichkeit, dass im Anschluss die zugehörige Dichtefunktion der Priori-Verteilung $f(\theta)$ nicht mehr integrierbar ist.(Held; 2008, Kapitel 5.2.3)

2.3 Kurze Konklusion beider Inferenzkonzepte

Zusammenfassend können wir sehen, dass beide Inferenzkonzepte zwar nicht auf die selbe Art und Weise agieren, aber in manchen Punkten besteht doch ein Zusammenhang zwischen beiden Konzepten. Während das zentrale Element der Likelihood-Inferenz sich in der *Law of Likelihood* widerspiegelt, ist es aus bayesianischer Sicht das Bayes-Theorem. Wie wir aber in Kapitel 2.2.2 gesehen haben, spielt die Likelihood-Funktion in der Bayes-Statistik ebenfalls eine Rolle.
Auch bei der Informationsgewinnung aus Daten unterscheidet sich die Vorgehensweise in beiden Inferenzkonzepten. Aus bayesianischer Sicht beinhaltet die Posteriori-Verteilung die gesamte Dateninformation über den Parameter θ, der aus dem Wahrscheinlichkeitsmodell hervorgeht, und dient somit in der Bayes-Statistik als Basis für die Dateninterpretation. Dagegen ist in der Likelihood-Inferenz das zentrale Element der Dateninterpretation die Likelihood-Funktion. Sie bildet die Leitlinie zur Interpretation der statistischen Daten als Evidenz. Außerdem gilt nach der Definition der *Likelihood Principle*, dass alle Informationen über den Parameter θ in der Likelihood-Funktion für θ enthalten sind, sofern die Daten bereits bekannt sind. Zudem besagt die *Likelihood Principle*, dass zwei Likelihood-Funktion von θ die gleiche Information über θ enthalten, wenn sie proportional zueinander sind. Das bedeutet hier aber auch, dass der Effekt auf jede Priori-Verteilung für θ in beiden Fällen derselbe ist. (Royall; 2000, Kapitel 1.2).
Letzten Endes können wir also sagen, dass die *Law of Likelihood* mit dem Likelihood Ratio ein Maß für die Evidenzstärke zwischen zwei Hypothesen bietet. Demgegenüber gibt es noch die *Likelihood Principle*, die die Konditionen, unter denen zwei Experimente die selbe äquivalente statistische Evidenz erzielt, festlegt. Diese Bedingung ist erfüllt, wenn und nur wenn beide Experimente die selbe Likelihood-Funktion erzeugen (vgl. Kapitel 2.1.3).(Blume; 2011, Kapitel 2)

3 Evidential Frameworks

Da die Statistik in der Wissenschaft eine wichtige Rolle für die korrekte Interpretation der Daten im Sinne einer wissenschaftlichen Evidenz spielt, gibt es ein breites Spektrum an statistischer Literatur, die sich mit diesem Thema beschäftigt. Aufgrund der Komplexität dieses Themengebietes ist eine Vielfalt von Sichtweisen und Meinungen diesbezüglich entstanden. Bedingt durch diese Diversität fehlt uns ein allgemein anerkanntes Framework zur Charakterisierung und Bewertung von Paradigmen, welche vorgeben, statistische Evidenz zu messen, d.h. uns fehlt, wie in Kapitel 2.1.2 beschrieben, ein konkretes Evidenzkonzept.(Blume; 2011, Kapitel 1)
Ein mögliches allgemeines Framework bietet uns der Ansatz von Blume. Er ermöglicht uns den Vergleich und die Beurteilung von statistischen Paradigmen, die behaupten, die Stärke der Evidenz in den Daten zu messen. Die Schlüsselkomponenten für dieses Framework setzen sich aus drei Größen zusammen, die aus den drei bekannten Inferenzkonzepten der Statistik, der frequentistischen, der bayesianischen und der Likelihood-Inferenz, hervorgehen. Das Ziel besteht darin, eine kritischere Beurteilung der statistischen Evidenz zu ermöglichen.(Blume; 2011, Kapitel 1)
Das Fehlen eines wohldefinierten Frameworks kann zu verschiedenen Kontroversen führen, wie die ordnungsgemäße Anwendung und Interpretation der p-Werte (vgl. Kapitel 2.1.2). Auch die Bayes-Inferenz ist nicht frei von Ungewissheit. Hier soll sowohl die Posteriori-Verteilung als auch der Bayes-Faktor ein Maß für die Evidenzstärke in den Daten darstellen. Daran lässt sich erkennen, dass auch hier ein klares Evidenzkonzept fehlt und es stellt sich uns die Frage, welches Maß die Evidenzstärke in den Daten besser repräsentiert. Im Laufe des Kapitels wird sich zeigen, dass die drei Schlüsselkomponenten einen wichtigen Beitrag zum Verständnis des Schemas zur Messung der statistischen Evidenz leisten.(Blume; 2011, Kapitel 1 & 1.3)

3.1 Drei Größen zur Evidenz

Die drei essentiellen Größen zur Bewertung und Interpretation der Evidenz in den Daten (im weiteren Verlauf EQ^2 genannt) setzen sich zusammen aus:

- **EQ1:** das Maß für die Stärke der Evidenz
- **EQ2:** die Wahrscheinlichkeit, dass ein bestimmtes Studiendesign eine irreführende Evidenz hervorbringt
- **EQ3:** die Wahrscheinlichkeit, dass die beobachtete Evidenz selbst irreführend ist.

EQ2 informiert uns über den Sammelprozess der Daten, während EQ1 und EQ3 uns Informationen über die statistische Evaluation der Daten als wissenschaftliche Evidenz geben. Alle drei Größen sind daher in der Wissenschaft und in der Statistik unentbehrlich. Die zeitliche Reihenfolge, in der die EQs während einer wissenschaftlichen Forschung bestimmt werden, ist: EQ2, EQ1 und EQ3, wobei EQ2 vor dem Sammelprozess der Daten bereits bestimmt wird.(Blume; 2011, Kapitel 1.1)
Jede einzelne Evidenzgröße beinhaltet die Antwort auf eine kritische Frage. EQ1 beantwortet die Frage, wie stark die Evidenz für oder gegen eine Hypothese in den Daten ist. EQ2 liefert uns Ergebnisse zur Wahrscheinlichkeit, mit der eine Studie Daten hervorbringen wird, die irreführend sind. EQ3 wiederum zeigt die Wahrscheinlichkeit an, dass die bereit beobachteten Daten irreführend sind. Daraus können wir schließen, dass EQ1 und EQ3 von den beobachteten Daten abhängen und sich auf diese beziehen. EQ2 hängt hingegen vom gewähltem Studiendesign ab und liefert keine Informationen zur Dateninterpretation, weil EQ2 bereits vor dem Sammelprozess der Daten bestimmt wird.(Blume; 2011, Kapitel 1.1)
Jedes einzelne Evidenzmaß bietet einzigartige Informationen bezüglich der Interpretation (EQ1), des Sammelprozesses (EQ2) und der Zuverlässigkeit (EQ3) an. Ein wohldefiniertes *evidential framework* ist erst dann gegeben, wenn alle drei EQs eindeutig definiert sind und klar voneinander unterschieden werden.(Blume; 2011, Kapitel 1.1)
In Kapitel 3.2 wird anhand einer Analogie der Unterschied zwischen den Wahrscheinlichkeiten EQ2 und EQ3 klarer beschrieben.

[2]Abkürzung aus dem englischem *evidential quantity*

3.2 Analogie zum Unterschied zwischen EQ2 und EQ3

Sowohl EQ2 als auch EQ3 stellen jeweils eine Wahrscheinlichkeit dar. Das kann zu einer Verwechslung der beiden Größen führen. Doch die genaue Unterscheidung der beiden EQs ist erforderlich, da jede für sich eine ausschlaggebende Information über die statistische Evidenz enthält. Wir wissen aus Kapitel 3.1, dass EQ2 bereits vor dem Sammelprozess bestimmt wird. EQ2 charakterisiert also die Wahrscheinlichkeit, dass das gewählte Studiendesign ein irreführendes Ergebnis erzielen wird. Doch sobald ein Datensatz zusammengetragen wurde, verliert EQ2 ihre Bedeutung, denn die beobachteten Daten sind entweder irreführend bezüglich der Evidenz oder nicht. Unser Hauptinteresse gilt nun der Wahrscheinlichkeit, dass die gerade gesammelten Daten irreführend sind. Genau diese Wahrscheinlichkeit wird im *evidential framework* als EQ3 bezeichnet. Daraus geht hervor, dass EQ3 die Wahrscheinlichkeit, mit der ein bereits beobachtetes Ergebnis irreführend ist, beschreibt.(Blume; 2011, Kapitel 1.2)
Mithilfe eines simplen Beispiels soll der genaue Unterschied beider Evidenzgrößen illustriert werden. Max und Moritz nehmen an einer Lotterie teil. Die Regeln lauten wie folgt: Es gibt 59 weiße und 39 rote Kugeln, wobei die Weißen von 1 bis 59 und die Roten von 1 bis 39 durchnummeriert sind. Um den Hauptgewinn zu erhalten, müssen fünf richtige weiße Kugeln und eine richtige rote Kugel ohne Zurücklegen gezogen werden, wobei die Reihenfolge bei den weißen Kugeln irrelevant ist. Daraus ergibt sich eine Gewinnchance von 1 zu 195,249,054.(Blume; 2011, Kapitel 1.2)
Max kauft sich ein Lotterielos und Moritz zehn. Jedoch ist ein Los aus diesen zehn identisch mit dem von Max, d.h. falls Max gewinnen sollte, gewinnt Moritz ebenfalls. Da Moritz aber mit seinen neun anderen Losen ebenso gewinnen könnte, hat er eine zehnmal so hohe Wahrscheinlichkeit auf den Hauptgewinn. Aufgrund der unterschiedlichen Spielstrategien, haben beide unterschiedliche Gewinnwahrscheinlichkeiten und genau diese Wahrscheinlichkeit auf den Hauptgewinn entspricht EQ2, d.h. die Wahrscheinlichkeit mit den gewählten Lotterielosen zu gewinnen. Dabei stellt das Gewinnen der Lotterie die irreführende Evidenz dar und die Lotterielose das gewählte Studiendesign.(Blume; 2011, Kapitel 1.2)
Am Tag nach der Auslosung schauen beide in die Zeitung, um die Gewinnnummern zu erfahren. Unglücklicherweise ist die Nummer der roten Kugel verschmiert, sodass diese nicht mehr zu erkennen ist. Doch die verbliebenen Nummern der weißen Kugeln stimmen alle mit Max' Lotterielos überein. Weil Moritz ein Los besitzt, dass identisch ist mit dem vom Max, haben beide zum jetzigen Zeitpunkt dieselbe Gewinnwahrscheinlichkeit

von 2.5%. Diese neue Wahrscheinlichkeit auf den Hauptgewinn entspricht EQ3, welche die Wahrscheinlichkeit zu Gewinnen darstellt, nachdem alle Kugeln bis auf die Rote übereinstimmen. Die Tatsache, dass Moritz zuvor zehn Lotterielose gekauft und somit eine zehnmal so hohe Gewinnwahrscheinlichkeit gegenüber Max hatte, ist zum jetzigen Zeitpunkt komplett irrelevant geworden. Das zeigt, wie bereits erwähnt, den Verlust der Relevanz von EQ2 nach der Datenerhebung.(Blume; 2011, Kapitel 1.2)

3.3 Ansatz auf Basis der Likelihood-Inferenz

Der Ansatz dieses Frameworks basiert auf der *Law of Likelihood*, d.h. die Daten unterstützen eher die Hypothese, die die beobachteten Ereignisse besser vorhersagt, und die Likelihood Ratio misst dabei den Grad, in dem eine Hypothese besser unterstützt wird als die andere (vgl. Kapitel 2.1.2). Dabei ist die Likelihood Ratio niemals negativ.(Blume; 2011, Kapitel 2)

3.3.1 Maß zur Evidenzstärke: Likelihood Ratio

Aus Kapitel 3.1 wissen wir, dass EQ1 ein Maß für die Evidenzstärke in unserem Framework ist und die *Law of Likelihood* bietet uns mit dem Likelihood Ratio eine konkrete Größe dazu an (vgl. Kapitel 2.1.2). Nun fehlt noch die Erkenntnis, wie wir einen konkreten Likelihood Ratio-Wert interpretieren können.
Wir nehmen an, dass die Beobachtungen $X_1, ..., X_n$ unabhängig und entsprechend einer Dichte $f(X_i|\theta)$ identisch verteilt sind. Zudem gibt es zum einen die Nullhypothese $H_0 : \theta = \theta_0$ und zum anderen die Alternativhypothese $H_1 : \theta = \theta_1$. Daraus ergibt sich, dass die Likelihood Ratio $LR = \frac{L_n(\theta_1)}{L_n(\theta_0)}$ die Evidenzstärke für H_1 gegenüber H_0 misst. Für eine beobachtete Likelihood Ratio wird zwischen drei Bereichen unterschieden:

- $LR \in \left[0, \frac{1}{k}\right]$: weist Evidenz für H_0 über H_1 auf
- $LR \in \left(\frac{1}{k}, k\right)$: schwache Evidenz für beide Hypothesen
- $LR \in [k, \infty)$: weist Evidenz für H_1 über H_0 auf

Konventionell wird $k = 8$ oder 32 gesetzt. Eine $LR = 8$ in den Beobachtungen deutet auf eine „ziemlich starke“ Evidenz hin und bei $LR = 32$ sprechen wir im Allgemeinen von einer „starken“ Evidenz. Je nach dem, ob wir eine „gemäßigte“ oder eine „harte“ Grenze setzen wollen, wir das entsprechende k gewählt.(Blume; 2011, Kapitel 2.3), (Royall; 2000, Kapitel 1.3)

3.3.2 Illustration der drei Evidenzgrößen

Mithilfe eines Beispiels angelehnt an dem bekannten Diagnosetestbeispiel von Royall (vgl. Royall; 1997, Kapitel 1.2) werden im Folgenden die drei Evidenzgrößen beschrieben. Nehmen wir an, die Krankheit Diabetes mellitus einer werdenden Mutter sei im Krankenhaus ein potentieller Faktor für die Schnittentbindung. Um zu erkennen, ob eine zukünftige Mutter Diabetikerin ist, werden Blutuntersuchungen durchgeführt. Das Bluttestergebnis dient in diesem Beispiel als Evidenz für eine Diabetes mellitus. Die

		Bluttestergebnis (B)	
		Positiv (+)	Negativ (-)
Diabetes mellitus (D)	Ja (+)	0.94	0.06
	Nein (-)	0.02	0.98

Tabelle 3.1: Wahrscheinlichkeiten der Blutuntersuchung bezüglich einer Diabetes mellitus Erkrankung

Wahrscheinlichkeiten in unserem Krankenhausbeispiel mit fiktiven Zahlen sind in der Tabelle 3.1 aufgeführt. Daraus erkennen wir, dass in unserem Beispiel die Sensitivität[3] $0.94 = P(B+|D+)$ und die Spezifität[4] $0.98 = P(B-|D-)$ ist. Legen wir nun die Hypothesen fest. H_+ bedeutet, dass die Mutter Diabetikerin ist, und H_- bedeutet wiederum, dass dies nicht der Fall ist. Bei einem positiven Bluttestergebnisses haben wir eine Likelihood Ratio von

$$LR = \frac{P(B+|D+)}{P(B+|D-)} = \frac{0.94}{0.02} = 47$$

und bei einem Negativen beträgt die Likelihood Ratio

$$LR = \frac{P(B-|D-)}{P(B-|D+)} = \frac{0.98}{0.06} = 16.3.$$

Mit Hilfe der Bereichsübergänge aus Kapitel 3.3.1 können die Likelihood Ratios nun genauer interpretieren werden. Für $k = 8$ stellen wir bei einem positiven Untersuchungsergebnis fest, dass hier mit $LR = 47$ eine starke Evidenz für H_+ gegenüber H_- vorliegt. Im Falle eines negativem Ergebnisses wird mit $LR = 16.3$ von einer starken Evidenz für H_- gegenüber H_+ gesprochen. Daraus erschließt sich, dass die Likelihood Ratio den

[3]Richtig-Positiv-Rate: Fähigkeit zu erkennen, ob Mütter mit einem positivem Blutuntersuchungsergebnis tatsächlich Diabetes mellitus haben

[4]Richtig-Negativ-Rate: Fähigkeit zu erkennen, ob Mütter mit einem negativem Blutuntersuchungsergebnis tatsächlich kein Diabetes mellitus haben

Grad misst, an dem die Daten eine Hypothese über eine andere unterstützt und somit entspricht die Likelihood Ratio dem Maß EQ1.(Blume; 2011, Kapitel 2.1)
Gleichwohl besteht die Möglichkeit, dass dieser Bluttest eine irreführende Evidenz generiert. Ein positives Untersuchungsergebnis wird korrekterweise als Evidenz für H_+ gegenüber H_- interpretiert, doch in 2% der Fälle tritt ein positives Testergebnis auch auf, obwohl die getestete Mutter keine Diabetikerin ist. Sofern dieses Szenario auftritt, hat die Blutuntersuchung eine irreführende Evidenz erzeugt. Auch ein negatives Ergebnis führt in unserem Beispiel in 6% der Fälle zu einer fehlgeleiteten Evidenz. Diese beiden Wahrscheinlichkeitswerte entsprechen der zweiten Evidenzgröße (EQ2). Sie sind analog zu den Fehlerraten im Hypothesentest und bilden wichtige Kennwerte für die Qualität der Blutuntersuchung und für den Sammelprozess der Daten. Ein guter Bluttest zeichnet sich durch die Maximierung von Sensitivität und Spezifität aus, was hier gleichbedeutend ist mit der Minimierung von EQ2. Durch das Verringern von EQ2 wird das Potential, ein irreführendes Bluttestergebnis zu beobachten, minimiert.(Blume; 2011, Kapitel 2.1)
Obwohl wir durch die Likelihood Ratio (EQ1) wissen, wie stark die Evidenz in den Daten ist, können wir dennoch keine sichere Aussage treffen, ob ein beobachtetes Testergebnis irreführend ist oder nicht. Allerdings besteht die Möglichkeit herauszufinden, ob ein beobachtetes Testergebnis dazu neigt, in die Irre zu führen. Voraussetzung dafür ist die Bereitschaft, eine bestimmte Annahme über die Priori Wahrscheinlichkeit der Hypothesen zu treffen. Zusammenfassend können wir für unser Beispiel sagen, dass ein positives Ergebnis der Blutuntersuchung irreführend ist, wenn und nur wenn die getestete Mutter nicht unter der Krankheit Diabetes mellitus leidet. Dabei ist $P(D-|B+)$ die Wahrscheinlichkeit dafür, dass die werdende Mutter keine Diabetikerin ist. Aus Kapitel 2.2.2 wissen wir, dass die Wahrscheinlichkeit $P(D-|B+)$ allgemein als Posteriori-Wahrscheinlichkeit bekannt ist.(Blume; 2011, Kapitel 2.1)
Damit die Berechnung dieser Posteriori-Wahrscheinlichkeit mit Hilfe des Bayes-Theorems (s. Kapitel 2.2.1) möglich ist, müssen zuvor die Priori-Wahrscheinlichkeiten festgelegt werden. Sei in unserem Beispiel $\pi_+ = P(H_+)$ (die Wahrscheinlichkeit, mit der die Patientin unter Diabetes mellitus leidet) und $\pi_- = P(H_-)$ (die Wahrscheinlichkeit, dass die Patientin nicht unter Diabetes mellitus leidet) unsere Priori-Wahrscheinlichkeiten sind. Wir wissen jedoch aus Kapitel 2.2.3, dass es verschiedene Wege gibt, um die Priori-Wahrscheinlichkeit zu bestimmen. Jedoch müssen wir in unserem Krankenhausbeispiel nicht die komplexen Methoden für die Priori-Wahrscheinlichkeiten nutzen, weil es sich hier um eine binäre Aussage handelt. Ferner ist unser Beispiel angelehnt an dem Diagnosetestbeispiel von Royall (vgl. Royall; 1997, Kapitel 1.2), weshalb wir auch in unse-

rem Szenario von einem Spezialfall sprechen können, da es hier bereits eine allgemeine Übereinstimmung bezüglich der Priori-Wahrscheinlichkeit gibt. Sofern es angemessen ist anzunehmen, dass die getestete Mutter zufällig aus einer Population gezogen wurde, bildet die Krankheitsprävalenz bzw. in unserem Beispiel die Diabetes mellitus Prävalenz die Priori-Wahrscheinlichkeit.(Blume; 2011, Kapitel 2.1)
Sei unsere Prävalenz für Diabetes mellitus $\pi_+ = 0.015$. Der komplementäre Wert dazu ist demnach $\pi_- = 1 - \pi_+ = 0.985$. Daraus ergibt sich die Wahrscheinlichkeit, mit der die werdende Mutter trotz positivem Bluttestergebnis kein Diabetes mellitus hat, durch

$$P(D-|B+) \overset{(2.1)}{=} \frac{P(B+|D-) \cdot P(D-)}{P(B+)} = ... = \left(1 + LR \cdot \frac{\pi_+}{\pi_-}\right)^{-1} \overset{\text{LR}=47}{=} 0.583$$

und analog dazu die Wahrscheinlichkeit, dass eine werdende Mutter trotz negativem Bluttestergebnis Diabetes mellitus hat, durch

$$P(D+|B-) \overset{(2.1)}{=} \frac{P(B-|D+) \cdot P(D+)}{P(B-)} = ... = \left(1 + LR \cdot \frac{\pi_-}{\pi_+}\right)^{-1} \overset{\text{LR}=16.3}{=} 0.0009.$$

Diese beiden Posteriori-Wahrscheinlichkeiten entsprechen der dritten Evidenzgröße (EQ3). Anhand der EQ3 Werte können wir sehen, dass ein positives Untersuchungsergebnis nicht so sicher ist wie ein Negatives. Tatsächlich führt in unserer Beispielpopulation ein beobachtetes positives Bluttestergebnis in mehr als der Hälfte der Fälle in die Irre. Es ist hier aber nicht falsch, das positive Testergebnis als Evidenz für die Präsens von Diabetes mellitus zu interpretieren. Es bedeutet lediglich, dass unsere Evidenzstärke in den Daten nicht stark genug ist, um unser Vorwissen über die Präsenz von Diabetes mellitus, also die Priori Wahrscheinlichkeiten, aufzuwiegen. $P(D-|B+) = 0.583$ bedeutet keineswegs, dass ein positives Bluttestergebnis eine Abwesenheit der Diabetes mellitus beweist. Interessant ist die Wahrscheinlichkeit für eine Diabetes mellitus Erkrankung vor der Blutuntersuchung, die bei $1.5\% = \pi_+$ liegt. Sobald aber ein positives Untersuchungsergebnis erzielt wurde, steigt diese Wahrscheinlichkeit auf $P(D+|B+) = 1 - P(D-|B+) = 41.7\%$. Der Grund für diese extreme Steigerung liegt in der großen Likelihood Ratio ($LR = 47$).(Blume; 2011, Kapitel 2.1)
Zusammenfassend können wir für EQ3 sagen, die Evidenzgröße hängt vom Kontext ab, da sie auf der Priori-Wahrscheinlichkeit aufbaut. Zudem kann auch eine starke Evidenz fehlgeleitet sein, aber in der Regel ist eine größere Likelihood Ratio ein eher sicheres Anzeichen dafür, dass eine bereits beobachtete Evidenz weniger wahrscheinlich in die Irre führt. Je stärker also die Evidenz ist, desto unwahrscheinlicher ist das Potential einer

Irreführung. Deswegen haben EQ1 und EQ3 eine inverse Beziehung zueinander.(Blume; 2011, Kapitel 2.1)

3.4 Problematik beim Fehlen eines wohldefinierten Frameworks

Wie bereits in Kapitel 3 erwähnt, kann das Fehlen eines wohldefinierten *evidential frameworks* zu verschiedenen Auseinandersetzungen kommen, wenn es um die Interpretation der Evidenz geht. Im Folgenden werden bestimmte statistische Situationen beschrieben, in denen diese Problematiken bezüglich der Evidenz auftreten.

3.4.1 Problematik im frequentistischem Inferenzkonzept

Zu den wichtigsten Werkzeugen der klassischen Inferenz gehören zum einen der Hypothesentest und zum anderen der Signifikanztest. Während der statistischen Untersuchung eines Experimentes werden typischerweise beide Testmethoden durchgeführt, wobei der Hypothesentest das Studiendesign festlegt und im Signifikanztest die Analyse stattfindet. Da aber diese Kombination von Testmethoden nicht wohldurchdacht ist, entsteht Verwirrung bei der sogenannten *tail area* Wahrscheinlichkeit[5]. Während im Hypothesentest die *tail area* Wahrscheinlichkeit die zweite Evidenzgröße (EQ2) als Fehler erster Art verkörpert, soll sie im Signifikanztest jedoch die Stärke der Evidenz (EQ1) messen. Ferner existiert im Signifikanztest kein EQ2 und im Hypothesentest lässt sich keine Größe finden, die die Evidenzstärke (EQ1) misst. Dies kann fälschlicherweise zu der Annahme führen, es wäre vernünftig, beide Evidenzgrößen zusammenzufassen. Doch in der Wissenschaft ist es immer ratsam, alle drei Evidenzgrößen auszumachen und voneinander abzugrenzen, denn jede einzelne Größe birgt eine Information, die für den wissenschaftlichen Prozess unerlässlich ist.(Blume; 2011, Kapitel 2.2)
Um diese Problematik genauer zu illustrieren, überlegen wir uns einen passenden Hypothesen- und Signifikanztest basierend auf dem Beispiel aus Kapitel 3.3.2. Für den Hypothesentest werden zunächst die Hypothesen formuliert. Die Nullhypothese H_0 ist hier, dass die Patientin nicht unter Diabetes mellitus leidet, und die Alternativhypothese H_1 steht für eine Diabetes mellitus Erkrankung der Patientin. Sofern ein positives Blutuntersuchungsergebnis beobachtet wird, kann die Nullhypothese abgelehnt werden und

[5]Die *tail area* Wahrscheinlichkeit ist ein wahrscheinlichkeitstheoretischer Begriff, der die Kernberechnung beim p-Wert und beim Fehler erster Art repräsentiert

analog dazu die Nullhypothese nicht abgelehnt werden, wenn ein negatives Untersuchungsergebnis auftritt. Dabei soll der Fehler erster Art bei 2% und der Fehler zweiter Art bei 6% liegen. Im Allgemeinen geht man hier von einem guten Test aus, da der Fehler erster Art kleiner ist als die konventionelle Grenze von 5%. Problematisch wird es erst, wenn wir versuchen, die resultierenden Testergebnisse als statistische Evidenz zu interpretieren. Falls wir die Nullhypothese nicht ablehnen können, impliziert dies keineswegs eine Evidenz für die Nullhypothese. Ferner kann ein negatives Ergebnis, wie beispielsweise das Scheitern der Ablehnung einer Nullhypothese, niemals als Evidenz für das Fehlen der Diabetes mellitus Erkrankung interpretiert werden. Das Fehlen einer Evidenz bedeutet nämlich nicht, dass es eine Evidenz für das Fehlen ist. Sollte es nicht möglich sein, die Nullhypothese abzulehnen, wird das Testergebnis stattdessen als statistisch ergebnislos interpretiert. Außerdem können wir unter diesen Umständen nirgendwo die Stärke der Evidenz wiedergeben, d.h. EQ1 existiert hier nicht. Die einzige Erkenntnis, die wir hieraus ziehen können, ist die Entscheidung, ob wir anhand der Fehlerrate aus unserer Entscheidungsregel die Nullhypothese ablehnen können oder nicht.(Blume; 2011, Kapitel 2.2)
Diese Information allein ist jedoch aus wissenschaftlichem Standpunkt heraus unzureichend, besonders wenn wir eine konkrete Evidenzstärke zu der Hypothese wollen, die uns interessiert. Um dies zu ermöglichen, wird am Ende einer Studie ein Signifikanztest durchgeführt. Dieser Test beinhaltet die Berechnung der p-Werte, die wir als Maß für die Evidenzstärke gegen die Nullhypothese verwenden. In unserem Beispiel liegt der p-Wert bei 2%, was hinsichtlich des konventionellen Maßstabs von 5% als eine starke Evidenz gegen die Nullhypothese betrachtet wird, da unser p-Wert kleiner ist als 5%. Auch hier sehen wir, dass es nicht möglich ist, eine Evidenz zugunsten der Nullhypothese zu bekommen. Große p-Werte können nämlich nicht als Evidenz für die Nullhypothese interpretiert werden, sondern sie deuten darauf hin, dass ein Ergebnis nicht beweiskräftig ist. Daraus ergibt sich die Erkenntnis, dass im Signifikanztest zwar mit dem p-Wert ein EQ1 vorhanden ist, aber es können weder EQ2 noch EQ3 bestimmt werden.(Blume; 2011, Kapitel 2.2)

3.4.2 Problematik im bayesianischem Inferenzkonzept

Im bayesianischem Ansatz liegt der Mittelpunkt bei der Evidenz vor allem auf der Posteriori-Wahrscheinlichkeit. Nehmen wir wieder das Beispiel aus Kapitel 3.3.2, dann liegt die Posteriori Wahrscheinlichkeit für Diabetes mellitus nach einem positiven Bluttestergebnis bei $P(D+|B+) = 0.417$. Sofern diese Wahrscheinlichkeit als Maß für die

Evidenzstärke (EQ1) dienen soll, bleibt unklar, wie dieses positive Bluttestergebnis zu interpretieren ist. Denn hier liegt die Situation vor, dass es nach der Beobachtung eines positiven Bluttestergebnisses wahrscheinlicher ist, kein Diabetes mellitus zu haben, da $P(D+|B+) = 41.7\% < 50\%$.(Blume; 2011, Kapitel 2.2)
Sollte also ein positives Blutuntersuchungsergebnis als Evidenz für das Fehlen der Diabetes mellitus Erkrankung betrachtet werden? Falls ja, kann die Blutuntersuchung niemals eine statistische Evidenz für die Präsenz einer Diabetes mellitus erzeugen, da die Posteriori-Wahrscheinlichkeit für eine Diabtis mellitus Erkrankrung nach einem negativem Bluttestergebnis mit $P(D+|B-) = 0.0009$ sehr klein ist. Falls nicht, basierend auf welchem Maßstab und Kontext sollen wir dann unsere Posteriori Wahrscheinlichkeit interpretieren? Es ist also nötig, EQ1 genau zu definieren. Außerdem bleibt im bayesianischem Konzept unklar, ob es sinnvoll ist, EQ2 in Abhängigkeit von der Priori Wahrscheinlichkeit zu definieren. Also können wir auch hier sagen, dass das Fehlen eines *evidential frameworks* zu keiner klaren Aussage bezüglich der Evidenz führt.(Blume; 2011, Kapitel 2.2)

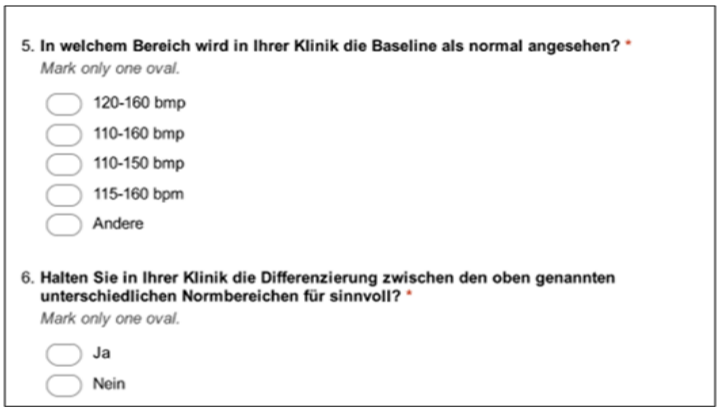
4 Analyse von Einflussfaktoren auf die Sectiorate anhand eines evidential frameworks

Nachdem die Grundidee eines *evidential frameworks* in der Theorie beschrieben wurde, wenden wir dieses Konzept nun auf einen Datensatz aus dem medizinischen Bereich an. Dieser Datensatz wurde von Herrn Dr. Martin Daumer vom „Sylvia Lawry Centre for Multiple Sclerosis Research e.V.“ bereitgestellt. Bevor wir zu der Anwendung kommen, werden zunächst der Datensatz und das verwendete Regressionsmodell genauer betrachtet.

4.1 Datensatz

Um herauszufinden, inwiefern verschiedene Faktoren die Sectiorate in den deutschen Krankenhäusern beeinflussen, wurde mittels *Crowdsourcing* über das Internet eine offene Umfrage durchgeführt, die mehrere kategoriale Fragen über die gegebenen Bedingungen und Richtlinien im gynäkologischen Bereich der jeweiligen Kliniken beinhalten.
Der Grund für das Interesse an diesem Thema ist die rasant gestiegene Rate an Kaiserschnitt Operationen in Deutschland, die 2010 bei 31.9% lag und somit innerhalb eines Jahrzehnts um zehn Prozentpunkte zugenommen hat. Damit unnötige Kaiserschnitt Operationen zukünftig vermieden werden, besteht die Nachfrage, herauszufinden, welche Faktoren den Anstieg der Sectiorate beeinflussen. Obwohl die Risiken einer Sectio für Mutter und Kind deutlich gesunken sind, wäre ein Anstieg an medizinisch unnötigen Kaiserschnitt Operationen dennoch nicht erstrebenswert. Problematisch ist hierbei nicht nur das allgemeine Risiko eines operativen Eingriffs, es existieren auch Hinweise auf mögliche langfristige gesundheitliche Folgen für das Kind (z.B. höheres Risiko auf Diabetes Typ 1, Asthma und Übergewicht) oder mögliche psychische Folgen für Mutter und Kind.(Kolip et al.; 2012, Kapitel 1 & 2.1)

In der Abbildung 4.1 sehen wir einen Auszug des Fragebogens aus der offenen Umfra-


5. **In welchem Bereich wird in Ihrer Klinik die Baseline als normal angesehen?** *
Mark only one oval.
- 120-160 bmp
- 110-160 bmp
- 110-150 bmp
- 115-160 bpm
- Andere

6. **Halten Sie in Ihrer Klinik die Differenzierung zwischen den oben genannten unterschiedlichen Normbereichen für sinnvoll?** *
Mark only one oval.
- Ja
- Nein

Abbildung 4.1: Auszug aus dem Fragebogen

ge, der an die „Deutsche Gesellschaft für Gynäkologie und Geburtshilfe e.V.“ gerichtet wurde, wobei Krankenhäuser die Zielgruppe sind. Mithilfe dieser Umfrage wurde ein Datensatz mit einem Stichprobenumfang von 97 generiert.
Die Variablen, auf die wir uns im Laufe dieses Kapitels bei der Anwendung des *evidential frameworks* beziehen werden, sind in Tabelle 4.1 aufgelistet. Da wir die verschiedenen

Variable	**Erklärung**
Sectio	Primäre und sekundäre Sectiorate in der Klinik *7 Kategorien:* <20%, 21-25%, 26-30%, 31-35%, 36-40%, 41-45%, >45%
Entbindung_Jahr	Anzahl der Entbindungen in der Klinik im Jahr 2014 *5 Kategorien:* <500 Geburten/Jahr, 500-1000 Geburten/Jahr, 1001-1500 Geburten/Jahr, 1501-2000 Geburten/Jahr, > 2001 Geburten/Jahr
Software	Hersteller der Zentralüberwachungssoftware *4 Kategorien:* Andere, Nexus, Philips, Trium
Baseline	Klinik-intern festgelegter Normalbereich für die Baseline *5 Kategorien:* 110-150 bmp, 110-160 bmp, 115-160 bmp, 120-160 bmp, Andere
DiffBase	Differenzierung von Normbereichen für die Baseline sinnvoll? (aus Sicht der Klinik) *binär:* Ja, Nein

Tabelle 4.1: Variablenliste

Einflüsse auf die Sectiorate[6][7] herausfinden möchten, ist unsere Zielvariable „Sectio“. In der Abbildung 4.2 sehen wir, wie sich die 97 Krankenhäuser aus der Umfrage auf die Kategorien der Zielvariable verteilen.

Es zeigt sich, dass die Kategorien „41-45%“ und „>45%“ mit je zwei Krankenhäusern

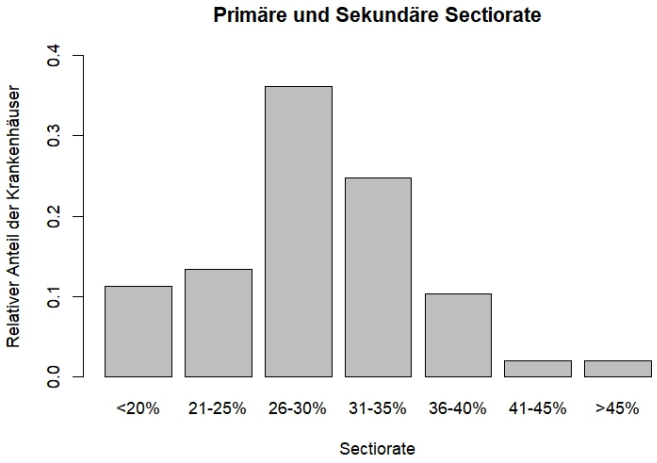


Abbildung 4.2: Relative Häufigkeiten der Zielvariable „Sectio“

vergleichsweise unterrepräsentiert sind, weshalb hier beide Kategorien mit der Kategorie „36-40%“ zusammengefasst werden. Daraus ergibt sich die neue Kategorie „>35%“. Durch diese Verknüpfung in eine gemeinsame Kategorie verhindern wir mögliche Komplikationen bei der späteren Modellierung, da unterrepräsentierte Kategorien wahrscheinlicher einen nicht repräsentativen Effekt erzielen können.

In Abbildung 4.3 sehen wir einen deskriptiven Überblick unserer Kovariablen. Um auch bei den Kovariablen das Problem mit unterrepräsentierten Kategorien zu vermeiden, wurden bei der Kovariable „Baseline“ die Kategorien „115-160 bmp“ und „120-160 bmp“ in die Kategorie „Andere“ aufgenommen. Somit ergibt sich eine neue Konstellation der Kategorien einiger Variablen, die wir in Tabelle 4.2 sehen können.

Bevor diese Zusammenfassung der Kategorien durchgeführt wird, stellt sich die Frage, ob gerade ein Krankenhaus, das eine Differenzierung von Baseline-Normbereichen für redundant hält, bei der Kovariable „Baseline“ in die Kategorie „Andere“ fällt. Aus dem Mosaikplot aus Abbildung 4.4 können wir jedoch erkennen, dass Krankenhäuser, die

[6] Primäre Sectio: Kaiserschnitt Operation wird vor dem Einsetzen der Wehen und bei intakter Fruchtblase durchgeführt.(Kolip et al.; 2012, Kapitel 3.2)

[7] Sekundäre Sectio: Kaiserschnitt Operation wird aufgrund einer Notfallsituation oder wegen der mütterlichen oder kindlichen Indikation (z.B. Geburtsstillstand) durchgeführt.(Kolip et al.; 2012, Kapitel 3.2)

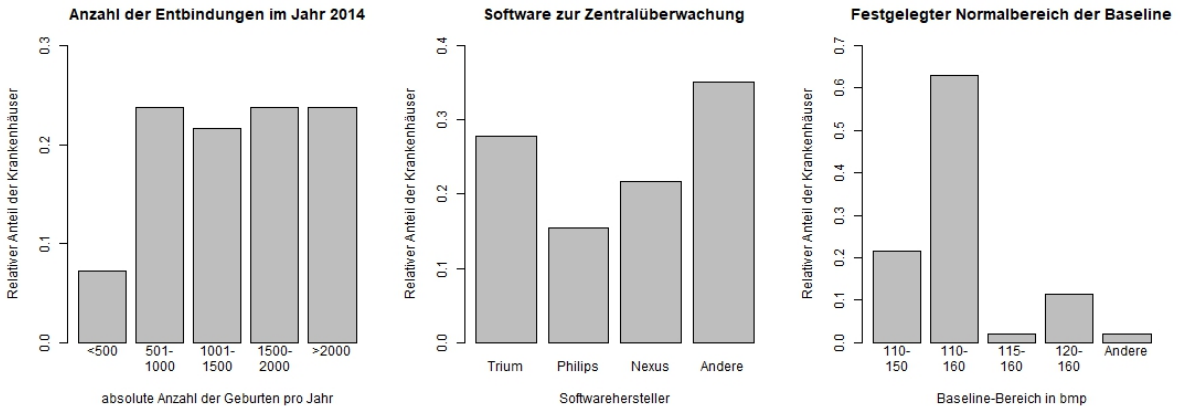


Abbildung 4.3: Relative Häufigkeiten der Kovariablen
links: „Entbindung_Jahr“, mitte: „Software“, rechts: „Baseline“

eine Differenzierung als nicht sinnvoll erachten, meistens einen geregelten Normbereich für die Baseline haben. So sind etwa 39.4% der Kliniken, die einen Normbereich für die Baseline von 110-160bmp festgelegt haben, diejenigen, die eine Differenzierung für redundant halten. Von allen befragten Kliniken, die einen reglementierten Normbereich von 120-160bmp haben, sind es sogar 72.7%, die eine Differenzierung des Normbereichs als nicht sinnvoll erachten.

4.2 Proportional Odds Model

Das Modell der proportionalen kumulativen Chancen oder auch *proportional odds model* (s. Fahrmeir, Kneib und Lang (2007), Kapitel 5.3 & Harrell (2015), Kapitel 13.3) ist das ordinal logistische Modell, welches am häufigsten angewendet wird. Die Anwendung dieses Modells ist dann geeignet, wenn die abhängige Variable ordinalskaliert ist, d.h. wenn sie Werte in geordneten Kategorien annimmt. Wir nehmen an, dass in diesem Modell hinter den beobachteten Kategorien eine latente unbeobachtbare Variable U steht, die zur Kovariablenvektor x_i durch

$$U_i = -x_i'\beta + \varepsilon_i, \qquad i = 1, ..., n$$

bestimmt ist, wobei β einen Parametervektor, n die Anzahl der Beobachtungen und ε_i eine Störvariable mit Verteilungsfunktion F darstellt. Der Zusammenhang zwischen der latenten Variable U und der Beobachtung Y sei bestimmt durch das Schwellenwertkon-

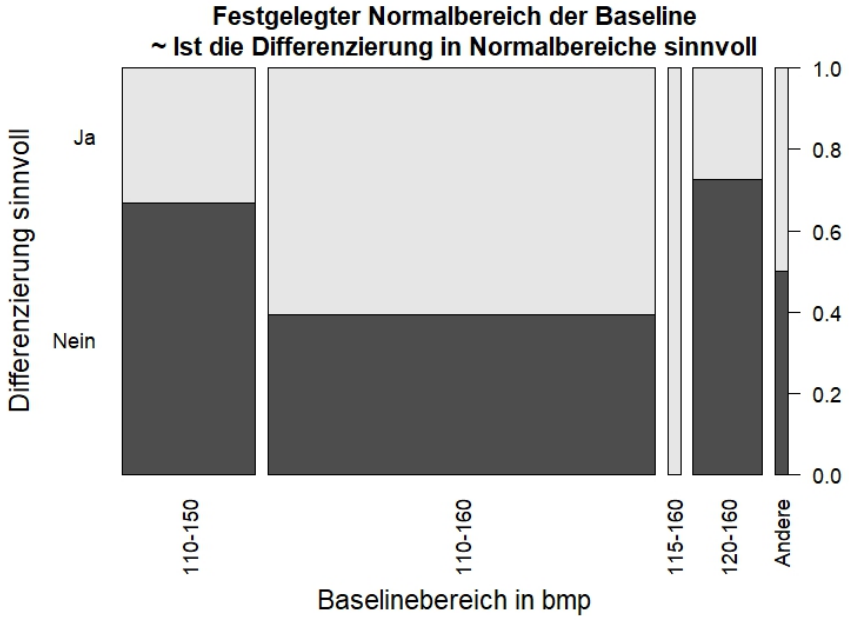


Abbildung 4.4: Zusammenhang zwischen den Kovariablen „Baseline“ und „DiffBase“

zept

$$Y_i = r \Longleftrightarrow \alpha_{r-1} < U_i \leq \alpha_r, \qquad r = 1, ..., q,$$

wobei $-\infty = \alpha_0 < \alpha_1 < ... < \alpha_c = \infty$ die Schwellwerte sind, die auf dem latenten Kontinuum liegen.(Fahrmeir, Kneib und Lang; 2007, Kapitel 5.3)
Daraus ergibt sich das kumulative Modell mit der Verteilungsfunktion F durch

$$P(Y_i \leq r|x_i) = P(U_i \leq \alpha_r) = F(\alpha_r + x_i'\beta), \qquad r = 1, ..., c.$$

Wir sehen, dass wir nun ein Regressionsmodell mit den Regressoren x_i, den Parametern $\alpha_1, ..., \alpha_q$ und β haben, wobei die latente Variable im Modell nicht mehr enthalten ist. Daraus ergeben sich auch die Wahrscheinlichkeiten

$$\begin{aligned} P(Y_i = 1|x_i) &= P(Y_i \leq 1|x_i) = F(\alpha_1 + x_i'\beta), \\ P(Y_i = r|x_i) &= P(Y_i \leq r|x_i) - P(Y_i \leq r-1|x_i) \\ &= F(\alpha_r + x_i'\beta) - F(\alpha_{r-1} + x_i'\beta), \qquad r = 2, ..., q. \end{aligned}$$

Um das *proportional odds model* bzw. das kumulative Logit-Modell zu erhalten, nehmen wir für F die logistische Verteilungsfunktion an, wodurch wir

$$P(Y_i \leq r|x_i) = \frac{exp(\alpha_r + x_i'\beta)}{1 + exp(\alpha_r + x_i'\beta)}$$

oder äquivalent dazu

$$ln\left(\frac{P(Y_i \leq r|x_i)}{P(Y_i > r|x_i)}\right) = logit(P(Y_i \leq r|x_i)) = \alpha_r + x_i'\beta$$

erhalten. Der Name des kumulativen Logit-Modells leitet sich von der Eigenschaft der proportional über alle Kategorien hinweg bestehenden kumulierten Chancen im Modell ab. Dies zeigt sich beim Verhältnis der kumulativen Chancen bezüglich zweier Subpopulationen x_i und $\tilde{x}_i$:

$$\frac{P(Y_i \leq r|x_i)/P(Y_i > r|x_i)}{P(Y_i \leq r|\tilde{x}_i)/P(Y_i > r|\tilde{x}_i)} = \frac{exp(\alpha_r + x_i'\beta)}{exp(\alpha_r + \tilde{x}_i'\beta)} = exp\left((x - \tilde{x})'\beta\right)$$

Es zeigt sich nämlich, dass dieses Verhältnis nicht mehr von der Kategorie r abhängig ist.(Fahrmeir, Kneib und Lang; 2007, Kapitel 5.3)

Variable	**Erklärung**
Sectio	Primäre und sekundäre Sectiorate in der Klinik *5 Kategorien:* <20%, 21-25%, 26-30%, 31-35%, >35%
Entbindung_Jahr	Anzahl der Entbindungen in der Klinik im Jahr 2014 *5 Kategorien:* <500 Geburten/Jahr, 500-1000 Geburten/Jahr, 1001-1500 Geburten/Jahr, 1501-2000 Geburten/Jahr, > 2001 Geburten/Jahr
Software	Hersteller der Zentralüberwachungssoftware *4 Kategorien:* Andere, Nexus, Philips, Trium
Baseline	Klinik-intern festgelegter Normalbereich für die Baseline *3 Kategorien:* 110-150 bmp, 110-160 bmp, Andere
DiffBase	Differenzierung von Normbereichen für die Baseline sinnvoll? (aus Sicht der Klinik) *binär:* Ja, Nein

Tabelle 4.2: Variablenliste mit den angepassten Kategorien

4.3 Anwendung des evidential frameworks

Bevor wir die einzelnen Evidenzgrößen (EQs) bestimmen können, stellen wir unser unrestringiertes *proportional odds model* (vgl. Kapitel 4.2) auf. Wenn wir die Variablen aus der Tabelle 4.2 verwenden und „Sectio“ als Zielvariable festlegen, erhalten wir folgendes Modell:

$$logit(P(Y \leq r|x_i)) = \alpha_r + \beta_{Entbindung\_Jahr} + \beta_{Software} + \beta_{Baseline} + \beta_{DiffBase} \quad (4.1)$$

mit $r = 1, ..., 4$. Die Indizes von β sind so zu verstehen, dass sie jeweils für eine Kategorie aus der jeweiligen Kovariable stehen (s. Tabelle 4.2), wobei die Referenzkategorien ausgenommen wurden. Die Referenzkategorien für die einzelnen Kovariablen lauten wie folgt: für Kovariable „Entbindung_Jahr“ ist es die Kategorie „<500 Geburten/Jahr“, für Kovariable „Software“ ist es die Kategorie „Andere“, für Kovariable „Baseline“ ist es die Kategorie „110-150 bmp“ und für Kovariable „DiffBase“ ist es die Kategorie „Ja“.
Bei der Wahl der Hypothesen gibt es typischerweise verschiedene Untersuchungsvarianten, die von den Interessen eines Anwenders abhängen. In unserem Fall möchten wir überprüfen, ob die Festlegung eines Normalbereichs für die Baseline einen Einfluss auf die Sectiorate hat, da wir nicht erkennen können, auf welcher Basis diese Normbereiche entstehen. Unsere Vermutung ist, dass diese Normalbereiche der Baseline je nach Klinik eher willkürlich festgelegt werden. Daraus entstehen unsere folgenden Hypothesen:

- Nullhypothese H_0: Die Kovariable „Baseline“ hat keinen Einfluss auf die Zielvariable (restringiertes Modell ohne die Kovariable „Baseline“)
- Alternativhypothese H_1: Die Kovariable „Baseline“ hat einen Einfluss auf die Zielvariable (unrestringiertes Modell)

oder äquivalent dazu H_0: $\beta_{Baseline} = 0$ und H_1: $\beta_{Baseline} \neq 0$.
Nachdem das Modell und die Hypothesen festgelegt sind, können wir unser *evidential framework* definieren. Dabei ist EQ1 die Likelihood Ratio zwischen unserem restringierten Modell ohne der Kovariable „Baseline“ und dem unrestringierten Modell. EQ2 funktioniert analog zum sogenannten Likelihood Ratio Test (s. Huelsenbeck und Crandall (1997), Kapitel *Likelihood Ratio Tests In Phylogenetics* & Harrell (2015), Kapitel 9.2.1) und liefert uns mit dem Fehler erster Art die Wahrscheinlichkeit für eine irreführende Evidenz im gewählten Studiendesign.
Anders als im Krankenhausbeispiel aus Kapitel 3.3.2 unterscheidet sich in unserem Fall

die Berechnung der Likelihood Ratio. Bedingt durch die verschiedene Parametervektoren βs unseres *proportional odds models* und unter Berücksichtigung unserer Hypothesen H_0: $\beta_{Baseline} = 0$ und H_1: $\beta_{Baseline} \neq 0$ ergibt sich für unsere Likelihood Ratio (Banerjee und Wellner; 2001, Kapitel 2.2)

$$\begin{aligned} LR &= \frac{sup_{H_1} f(y_i|\beta_{Entbindung\_Jahr}, \beta_{Software}, \beta_{Baseline}, \beta_{DiffBase})}{sup_{H_0} f(y_i|\beta_{Entbindung\_Jahr}, \beta_{Software}, \beta_{Baseline}, \beta_{DiffBase})} \\ &= \frac{sup_{H_1} f(y_i|\beta_{gesamt})}{sup_{H_0} f(y_i|\beta_{gesamt})} \end{aligned} \tag{4.2}$$

mit $i = 1, ..., n$, wobei n die Anzahl der befragten Kliniken ist. Die Funktion $f(\cdot)$ ist eine diskrete Dichtefunktion bzw. unsere Likelihood-Funktion und liefert die Wahrscheinlichkeit für die Beobachtung y_i unter den gegebenen Parametervektoren (β_{gesamt}).
Für EQ3 müssen wir unser bisheriges Modell in ein bayesianisches Modell umformen, womit wir die Posteriori-Wahrscheinlichkeit (EQ3) bestimmen können. Wie bereits in Kapitel 3.1 erwähnt, lautet die Reihenfolge für die Berechnung der einzelnen Evidenzfolgen EQ2, EQ1 und EQ3.

4.3.1 Berechnung von EQ1 und EQ2

Wie wir nun wissen, funktioniert unsere EQ2 analog zum Likelihood Ratio Test (LR-Test). Die Likelihood Ratio Statistik folgt bei großem Stichprobenumfang approximativ der χ^2-Verteilung und die Differenz der Parameter in beiden zu vergleichenden Modellen entspricht der Anzahl an Freiheitsgraden. Dabei werden unsere *proportional odds models* in `R` mithilfe der Funktion `polr()` aus dem `R`-Paket `MASS` (s. Venables und Ripley (2002)) dargestellt, wobei hier beachten werden müssen, dass die Funktion `polr()` für die *odds* bzw. für die Chancen statt $\theta_r + x_i'\beta$ die Form $\theta_r - x_i'\beta$ verwendet. Die Teststatistik zum LR-Test ist (Harrell; 2015, Kapitel 9.2.1)

$$T = -2(ln(sup_{H_0} f(y_i|\beta_{gesamt})) - ln(sup_{H_1} f(y_i|\beta_{gesamt}))) = -2(l_0 - l_1).$$

Damit wir den Fehler erster Art (EQ2) bestimmen können, nutzen wir die Eigenschaft der approximativen Verteilungsannahme der Teststatistik T. Dabei muss folgende Transformation durchgeführt werden, damit T approximativ der χ^2-Verteilung mit $T \overset{a}{\sim} \chi^2_2$

folgt:

$$\begin{aligned} exp(T) &= exp(-2(l_0 - l_1)) \\ &= exp(l_0 - l_1)^{-2} \\ &= \left(\frac{exp(l_0)}{exp(l_1)}\right)^{-2} \\ &= \left(\frac{sup_{H_0} f(y_i|\beta_{gesamt})}{sup_{H_1} f(y_i|\beta_{gesamt})}\right)^{-2} \\ &= \left(\frac{sup_{H_1} f(y_i|\beta_{gesamt})}{sup_{H_0} f(y_i|\beta_{gesamt})}\right)^{2} \\ &\overset{4.2}{=} LR^2. \end{aligned}$$

Wie aus Kapitel 3.3.1 bekannt ist, muss $LR \geq k$ sein, damit eine Evidenz zugunsten von H_1 über H_0 vorliegt. In unserem Fall wählen wir die konventionelle Grenze $k = 8$. Bezogen auf die Teststatistik T und der Transformation ergibt sich ein Grenzwert von $ln(LR^2) = ln(8^2)$. Wenn wir das auf unsere gewählten Hypothesen anwenden, erhalten wir für $H_0 : T \leq ln(64)$ und für $H_1 :> ln(64)$. Da LR stetig ist, und somit die Wahrscheinlichkeit für einen konkreten Punkt gleich 0 ist, können wir bei der Nullhypothese statt $T < ln(64)$ auch unsere jetzige Definition wählen, damit die Korrektheit der Hypothesenbildung berücksichtigt wird. Unser EQ2 entspricht nun dem Fehler erster Art, d.h. die Wahrscheinlichkeit, dass unser gewähltes Studiendesign eine Evidenz für H_1 aufweist, obwohl H_0 wahr ist, $P(T > ln(64)|H_0$ wahr). Daraus ergibt sich für EQ2 (Berechnung mit `R`, s. elektronischer Anhang)

$$P(T > ln(64)|H_0 \text{ wahr}) = 1 - P(T \leq ln(64)|H_0 \text{ wahr}) \approx 0.125,$$

d.h. unsere Wahrscheinlichkeit, dass wir mit unserem Studiendesign eine irreführende Evidenz erhalten (EQ2), beträgt etwa 12.5%.
Im nächsten Schritt bestimmen wir die Stärke der Evidenz (EQ1) anhand der Likelihood Ratio (vgl. Kapitel 3.3.1). Daraus ergibt sich (Berechnung mit `R`, s. elektronischer Anhang)

$$LR \overset{4.2}{=} \frac{sup_{H_1} f(y_i|\beta_{gesamt})}{sup_{H_0} f(y_i|\beta_{gesamt})} \approx 36.04,$$

d.h. wir haben eine starke Evidenz zugunsten des unrestringierten Modells (H_1) gegenüber dem restringierten Modell ohne die Kovariable „Baseline“ (H_0).

4.3.2 Berechnung von EQ3 mit spike-and-slab Prioris

Um schlussendlich die Posteriori-Wahrscheinlichkeit (EQ3) bestimmen zu können, müssen wir unser frequentistisches Modell (4.1) in ein bayesianisches konvertieren. Das ermöglicht uns, die notwendige Priori-Wahrscheinlichkeit zu bestimmen, die für die Berechnung der Posteriori-Wahrscheinlichkeit benötigt werden. In Kapitel 2.2.3 haben wir gesehen, dass es verschiedene Methoden der Priori-Bestimmung gibt. Aufgrund der Komplexität unseres *proportional odds models* können wir hier keine konjugierten Priori-Verteilungen benutzen. Doch in unserem Fall bietet sich die sogenannte „*spike-and-slab*“ Priori-Verteilung an.
Im Folgenden werden die Verteilungsannahmen der verschiedenen Parameter und Variablen aus unserem bayesianischen Modell beschrieben. Sei die Priori-Verteilung der Koeffizienten $\beta_{Entbindung\_Jahr}, \beta_{Software}$ und $\beta_{DiffBase}$ eine Normalverteilung mit $\beta_{gesamt} \sim \mathcal{N}(0,1)$. Da wir durch unsere Hypothesen H_0: $\beta_{Baseline} = 0$ und H_1: $\beta_{Baseline} \neq 0$ den Einfluss der Kovariable „Baseline“ überprüfen wollen, hat $\beta_{Baseline}$ eine andere Priori-Verteilung. Sie folgt der „*spike-and-slab*“ Priori-Verteilung.
Sei $\beta_{Baseline}$ definiert als $\beta_{Baseline} = \vartheta \cdot \beta^{\star}_{Baseline}$. Dabei gibt der Parameter ϑ an, ob die Kovariable „Baseline“ einen Einfluss auf die Zielvariable hat, d.h. $\vartheta = 1$, wenn $\beta_{Baseline} \neq 0$ und $\vartheta = 0$, wenn $\beta_{Baseline} = 0$. Also ist der Parameter $\vartheta \in \{0,1\}$ und folgt einer Bernoulli-Verteilung mit $\vartheta \sim \mathcal{B}(1, 0.5)$. Die Verteilung von ϑ bildet dabei den „*spike*“ Teil und repräsentiert somit die Wahrscheinlichkeit, dass ein Koeffizient aus dem Modell gleich 0 ist. Der „*slab*“ Part bildet hingegen die Verteilung für die Werte, deren Koeffizienten nicht 0 sind, bedingt auf das Wissen, welche Koeffizienten ungleich 0 sind. In unserem Fall ist es die Verteilung von $\beta^{\star}_{Baseline}$, die analog zu den anderen Koeffizienten normalverteilt ist mit $\beta^{\star}_{Baseline} \sim \mathcal{N}(0,1)$.(Scott und Varian; 2015, Kapitel 4.2.2)
Nun fehlen noch die Verteilungsannahmen der *Intercepts* und der Zielvariable. Die *Intercepts* α_r mit $r = 1, ..., 4$ folgen der Normalverteilung: $\alpha_r \sim N(0, 1000)$. Die Zielvariable Y_i mit $i = 1, ..., n$ ist multinomialverteilt mit $Y_i \sim \mathcal{M}(n, p_z)$. Dabei steht n für die Anzahl der befragten Krankenhäuser aus dem Datensatz, z stellt eine Kategorie aus der Zielvariable „Sectio“ (s. Tabelle 4.2) dar und p_z ist die Wahrscheinlichkeit, mit der eine Beobachtung in die Kategorie z fällt.
Die Posteriori-Verteilung unseres bayesianischen Modells kann nun mithilfe eines *Markov Chain Monte Carlo* (MCMC) Algorithmus simuliert werden. Dieser Algorithmus wird sehr oft wiederholt, sodass sich eine Kette von Ziehungen ergibt, aus der wir die Verteilung der Posteriori-Wahrscheinlichkeit von ϑ empirisch schätzen können. Da ϑ ein

binärer Parameter ist, entspricht der Erwartungswert von ϑ der Wahrscheinlichkeit, für die $\vartheta = 1$ ist, d.h. die Wahrscheinlichkeit, dass die Kovariable „Baseline“ einen Einfluss auf die Zielvariable hat. Also ergibt sich unsere gesuchte Posteriori-Wahrscheinlichkeit (EQ3) aus dem Erwartungswert der gezogenen ϑs.(Scott und Varian; 2014, Kapitel 4.1) Für eine genauere Erläuterung der Theorie zu den *spike-and-slab* Prioris wird neben der verwendeten Literatur auf folgende Quellen verwiesen: Mitchell und Beauchamp (1988), George und McCulloch (1993), Ishwaran und Rao (2005). Die Berechnung der EQ3 wurde in `R` mit der Funktion `run.jags()` aus dem `R`-Paket `runjags` (s.(Denwood; 2016)) durchgeführt. Aus der Simulation des MCMC Algorithmus mit 2 Ketten, in denen jeweils 10,000 Zufallsziehungen stattfinden, erhalten wir für den Erwartungswert von ϑ $\mathbb{E}(\vartheta) \approx 0.794$ (Berechnung mit `R`, s. elektronischer Anhang). Diese Posteriori zeigt aber die Wahrscheinlichkeit für einen Einfluss der Kovariable „Baseline“ (H_1) an. Da unser EQ1 eine Evidenz zugunsten von H_1 gegenüber H_0 aufweist und wir die Wahrscheinlichkeit, mit der die beobachtete Evidenz irreführend ist, suchen, muss für EQ3 die Gegenwahrscheinlichkeit gebildet werden. Daraus ergibt sich eine Posteriori-Wahrscheinlichkeit von $1 - 0.794 = 0,206$, d.h. die Wahrscheinlichkeit, mit der unsere beobachtete Evidenz irreführend ist (EQ3), beträgt 20.6%, wobei mit beobachteter Evidenz die Likelihood Ratio (EQ1) aus Kapitel 4.3.1 gemeint ist.
Zusammenfassend ergibt sich für das untersuchte Hypothesenpaar H_0: $\beta_{Baseline} = 0$ und H_1: $\beta_{Baseline} \neq 0$ folgendes Ergebnis: Die Wahrscheinlichkeit, dass unser gewähltes Studiendesign eine irreführende Evidenz erzielt (EQ2), beträgt etwa 12.5%. Die berechnete Likelihood Ratio (EQ1) von etwa 36.04 zeigt, dass die Daten eine starke Evidenz zugunsten von H_1 gegenüber H_0 aufweist, wobei die Wahrscheinlichkeit, dass diese beobachtete Evidenz irreführend ist, 20.6% beträgt.

5 Fazit und Ausblick

Wie aufgezeigt wurde, ist ein wohldefiniertes Framework für die Evidenz durchaus von Vorteil. Es zeigt sich, dass ein alleiniger Wert, der die Evidenz in den Daten repräsentieren soll, nicht ausreicht, um vollkommene Klarheit zu erlangen. Aufgrund des mangelnden Informationsspektrums der herkömmlichen Methoden zur Festlegung der Evidenz stellt der Autor Jeffrey Blume sein *evidential framework* vor. In seinem Ansatz wird die Information über die Evidenz nicht mehr ausschließlich an einem Wert gemessen. Weil jede einzelne der drei Evidenzgrößen unerlässliche Informationen enthält, kann diesem Mangel entgegen gewirkt werden. Nur wenn alle drei Evidenzgrößen klar voneinander unterschieden werden, spricht man von einem wohldefinierten Framework. Dadurch ist es nicht nur möglich zu überprüfen, ob Evidenz in den Daten ist, sondern es gibt uns die Möglichkeit, eine Aussage darüber zu treffen, wie wahrscheinlich es ist, eine irreführende Evidenz vor der Datenerhebung zu erzielen. Zusätzlich ist eine konkrete Bezifferung der Evidenzstärke und die Wahrscheinlichkeit, dass diese beobachtete Evidenz irreführend ist, ein weiterer Vorteil des Frameworks. Dieses *evidential framework* ermöglicht es, genauere Behauptungen über die Evidenz zu treffen.
Es besteht zudem die Möglichkeit, die einzelnen Evidenzgrößen (EQs) des *evidential frameworks* neben der bisherigen Konstellation zu verändern. Dadurch ist es möglich, sich unterschiedlichen Situationen anzupassen. Im Bezug zu einer Studie der Bertelsmann Stiftung (Kolip et al.; 2012) bestünde die Möglichkeit, nicht die Likelihood Ratio (LR) sondern die Odds Ratio (OR) als EQ1 zu wählen, da sie in dieser Studie als Bezugswert für statistische Evidenz dient. Resultierend daraus müssten auch die Hypothesen für EQ2 geändert werden. Eine mögliche Hypothesenwahl wäre, dass $H_0 : OR = 1$ und $H_1 : OR \neq 1$ ist. Eine OR von 1 bedeutet nämlich, dass sich die Odds bzw. die Chancen der verglichenen Gruppen sich nicht voneinander unterscheiden.
Eine weitere Gestaltungsmöglichkeit des *evidential frameworks* wäre eine auf den AIC Wert basierende Evidenzstärke. Es bestünde die Möglichkeit, die Differenz der AIC Werte aus zwei Modellen (Δ_{AIC}), die verglichen werden sollen, als EQ1 zu wählen. Der Vorteil des AICs gegenüber dem LR-Ansatz ist der Strafterm. Er hängt von der Anzahl der

geschätzten Parameter ab und bei einer Zunahme der geschätzten Parameter wird das Modell härter bestraft.(Fahrmeir, Kneib und Lang; 2007, Kapitel 4.1.4)
Um genauere Aussagen über die tatsächliche Evidenzstärke zu treffen, müsste man hier, ähnlich wie bei der Likelihood Ratio, Grenzen festlegen, die zeigen, ab wann Δ_{AIC} eine starke Evidenz für oder gegen ein Modell aufweist (vgl. Kapitel 3.3.1). Auch bei dieser Wahl von EQ1 müssten die Hypothesen für EQ2 neu formuliert werden. Eine mögliche Gestaltung der Hypothesen könnte sein, dass $H_0 : \Delta_{AIC} = 0$ und $H_1 : \Delta_{AIC} \neq 0$, da $\Delta_{AIC} = 0$ bedeuten würde, dass es weder für das eine noch das andere Modell eine Evidenz gibt.
Abschließend muss angemerkt werden, dass der Einfachheit halber bei der Modellierung im Kapitel 4.3 die Betrachtung der Interaktionen zwischen den einzelnen Kovariablen vernachlässigt wurde, da der Fokus dieser Arbeit vor allem auf die Methodik zum *evidential framework* und deren Anwendung lag. Bei einer genaueren Analyse in der Praxis sollte dieser Schritt berücksichtigt werden.
Aufgrund der Flexibilität und des Informationsgewinns ist das *evidential framework* von Jeffrey Blume durchaus ein aussichtsreicher Ansatz, um zukünftig die Durchführung von evidenzbasierten Studien akkurater zu gestalten.

Literaturverzeichnis

Banerjee, M. und Wellner, J. A. (2001). Likelihood ratio tests for monotone functions, *Annals of Statistics* pp. 1699–1731.

Berger, J. O., Wolpert, R. L., Bayarri, M. J., DeGroot, M. H., Hill, B. M., Lane, D. A. und LeCam, L. (1988). The likelihood principle, *Lecture Notes-Monograph Series* **6**: iii–199.

Birnbaum, A. (1962). On the foundations of statistical inference, *Journal of the American Statistical Association* **57**(298): 269–306.

Blume, J. D. (2011). Likelihood and its evidential framework, *Handbook of the philosophy of science: philosophy of statistics* **7**: 493–511.

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS, *Journal of Statistical Software* **71**(9): 1–25.

Fahrmeir, L., Kneib, T. und Lang, S. (2007). *Regression: Modelle, Methoden und Anwendungen*, Statistik und ihre Anwendungen, Springer Berlin Heidelberg.

Fahrmeir, L., Künstler, R., Pigeot, I. und Tutz, G. (2007). *Statistik: Der Weg zur Datenanalyse*, 6. edn, Springer Berlin Heidelberg.

Gelman, A. und Hennig, C. (2017). Beyond subjective and objective in statistics, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **180**(4): 967–1033.

George, E. I. und McCulloch, R. E. (1993). Variable selection via gibbs sampling, *Journal of the American Statistical Association* **88**(423): 881–889.

Guyatt, G., Cairns, J., Churchill, D. und et al (1992). Evidence-based medicine: A new approach to teaching the practice of medicine, *JAMA* **268**(17): 2420–2425.

Harrell, F. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer Series in Statistics, Springer International Publishing.

Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes*, 1. edn, Spektrum Akademischer Verlag.

Held, L. und Bové, D. S. (2014). *Applied Statistical Inference: Likelihood and Bayes*, Springer Berlin Heidelberg.

Huelsenbeck, J. P. und Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood, *Annual Review of Ecology and Systematics* **28**(1): 437–466.

Ishwaran, H. und Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies, *The Annals of Statistics* **33**(2): 730–773.

Koch, K.-R. (2000). *Einführung in die Bayes-Statistik*, Springer Berlin Heidelberg.

Kolip, P., Nolting, H.-D. und Zich, K. (2012). Kaiserschnittgeburten - entwicklung und regionale verteilung, *Faktencheck Gesundheit Kaiserschnitt* .

Mitchell, T. J. und Beauchamp, J. J. (1988). Bayesian variable selection in linear regression, *Journal of the American Statistical Association* **83**(404): 1023–1032.

Müllner, M. (2005). *Erfolgreich wissenschaftlich arbeiten in der Klinik: Evidence Based Medicine*, SpringerLink: Springer e-Books, Springer Vienna.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*, Vol. 71, CRC press.

Royall, R. (2000). On the probability of observing misleading statistical evidence, *Journal of the American Statistical Association* **95**(451): 760–768.

Scott, S. L. und Varian, H. R. (2014). Predicting the present with bayesian structural time series, *International Journal of Mathematical Modelling and Numerical Optimisation* **5**(1-2): 4–23.

Scott, S. L. und Varian, H. R. (2015). Bayesian variable selection for nowcasting economic time series, *Economic analysis of the digital economy*, University of Chicago Press, pp. 119–135.

Taper, M. L. und Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science, *Population Ecology* **58**(1): 9–29.

Venables, W. N. und Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4. edn, Springer, New York.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*