Universität Leipzig
Fakultät für Mathematik und Informatik
Mathematisches Institut

Diplomarbeit zum Thema:

Modellierung eines Risikoäquivalentes für isolierte Ereignisse und singuläre Ereignisketten in Krankenversicherungsbiographien - mit Anwendung

Leipzig, den 16.2.2015

vorgelegt von:
Tobias, Riedel
geb. am: 25. 10. 1984
Studiengang Mathematik

Betreuer:

Prof. Dr. Manfred Riedel
HD Dr. Walter Warmuth

Kurzzusammenfassung

In der Welt der Krankenversicherung gibt es immer wieder die Frage, wie sich die zukünftigen Leistungskosten eines Versicherten abschätzen lassen. Bei den Gesundheitsforen Leipzig ist zu diesem Zweck ein mathematisches Modell entstanden, dass diese Fragestellung beantworten soll: Das Kohorten-Modell. Bei diesem wird das Leben eines Versicherten in drei Phasen unterteilt, welche zusätzlich von zwei Ereignisarten überlagert werden. Die verschieden Lebensphasen wurden unter anderen in den Artikeln „Leben und Tod- Spezifische Implikationen eines vermeintlich längeren Lebens für die Versicherungswirtschaft“ ([8] und [8]) und „Demographischer Wandel in Deutschland- Legenden und Mythen“ ([10]) ausführlich diskutiert. Diese Diplomarbeit widmet sich deshalb der Beschreibung und Modellierung der überlagernden Ereignisarten: Den isolierten Ereignissen und den singulären Ereignisketten. Es wird vorgeführt, welche mathematischen Methoden verwendet werden können, um geeignete Modelle zu konstruieren. Außerdem werden die Besonderheiten im deutschen Gesundheitswesen erläutert, die dabei beachtet werden müssen. Die vorgestellten Modelle werden dann praktisch und theoretisch implementiert und bewertet. Als Resultat kann ein Risikoäquivalent für dieses Teile des Kohorten-Modells angegeben werden.

Inhaltsverzeichnis

1. Einleitung

Die Gesundheitsforen Leipzig haben ein Modell entwickelt, welches, basierend auf den Morbitätsinformationen von Versicherten, die Höhe ihrer zu erwartenden Leistungskosten, bis zum Lebensende, für die Krankenversicherung bestimmt. Dieses Modell bildet den thematischen Rahmen dieser Diplomarbeit. Das Leben eines Krankenversicherten wird in diesem Ansatz in drei Phasen unterteilt, welche von den titelgebenden isolierten Ereignissen und singulären Ereignisketten überlagert werden. Das Ziel dieser Arbeit ist es ein Risikoäquivalent für diese beiden Modellbestandteile zu bestimmen.

Im zweiten Kapitel wird die Motivation für die Erstellung des Kohorten-Modells erläutert und es wird in diesem Zusammenhang einen kurzer Einblick in das deutsche Krankenversicherungssystem gegeben. Anschließend werden die einzelnen Modellbestandteile vorgestellt und gegeneinander abgegrenzt. Diese Zusammenfassung basiert auf den bisherigen Veröffentlichungen zum Kohorten-Modell: „Leben und Tod- Spezifische Implikationen eines vermeintlich längeren Lebens für die Versicherungswirtschaft“ ([8] und [8]) und „Demographischer Wandel in Deutschland- Legenden und Mythen“ ([10]).

Das dritte Kapitel stellt die benötigten, mathematischen Grundlagen für die Modellierung vor. Zu Beginn werden die in dieser Arbeit verwendeten Standardverteilungen eingeführt und die jeweiligen Eigenschaften gezeigt. Anschließend werden die stochastischen Prozesse, die in den späteren Modellen verwendet werden, betrachtet. Dies sind insbesondere der Poisson-Prozess und die Markov-Kette.

Im vierten und fünften Kapitel werden die isolierten Ereignisse und die singulären Ereignisketten behandelt. Die vorgestellten mathematischen Methoden werden verwendet, um, unter Berücksichtigung der Rahmenbedingungen durch das Kohorten-Modell und den Gegebenheiten im deutschen Gesundheitswesen, jeweils einen Modellansatz zu konstruieren, der eine Prognose der erwarteten Leistungen ermöglicht. Im Anschluss wird darauf eingegangen, wie isolierte Ereignisse und singuläre Ereignisketten in den Abrechnungsdaten der Krankenkassen erkannt werden können und gegenüber den restlichen Leistungen abgrenzt werden sollten. Der Modellansatz für die isolierten Ereignisse wird schließlich am Beispiel von Gehirnerschütterungen umgesetzt und die prognostizierten Kosten werden analysiert. Für die singulären Ereignisketten wird lediglich eine systematische Beschreibung des Vorgehens bei der Umsetzung angegeben, da, aufgrund der Anforderungen an die Datengrundlage, eine konkrete Umsetzung nicht möglich war.

In beiden Kapiteln wird versucht, durch Beispiele aus der Praxis, die verschiedenen Eigenschaften von isolierten Ereignissen und singulären Ereignisketten zu veranschaulichen.

Diese Beispiele basieren auf den Erfahrungen aus verschiedenen Projekten bei den Gesundheitsforen Leipzig sowie aus Gesprächen mit den Mitarbeitern.

Zum Abschluss der Arbeit erfolgt eine Zusammenfassung der gewonnenen Erkenntnisse sowie eine Einordnung der Resultate im Kohorten-Modell.

2. Einführung in das Thema

2.1. Mitnahme von Alterungsrückstellung in der PKV - Auswirkungen auf das Versichertenkollektiv

Im deutschen Gesundheitssystem gibt es zwei Systeme: Die Gesetzliche Krankenversicherung (GKV) und die Private Krankenversicherung (PKV). Aufgrund der Krankenversicherungspflicht müssen alle Personen, die einen Wohnsitz in Deutschland haben, in einem der beiden Systeme versichert sein. Für die GKV gilt eine Aufnahmepflicht und aus diesem Grund ist der Großteil der Bevölkerung[1] in der GKV versichert. Personen, die bestimmte Aufnahmekriterien erfüllen, können auch in die PKV wechseln. Seit 2009 zahlt jeder gesetzlich Versicherte monatlich einen Teil seines Einkommens (z.B. Gehalt bzw. Rente) in den Gesundheitsfond ein. Andere Einkommensquellen, wie zum Beispiel Mieteinnahmen, werden dabei nicht berücksichtigt. Dieses Geld wird anschließend an die verschiedenen gesetzlichen Krankenkassen verteilt. Bei der Verteilung werden, neben dem Alter und dem Geschlecht der Versicherten, auch ausgewählte, besonders kostenintensive Krankheiten berücksichtigt. Bei diesem System wird der Beitragsatz jährlich, auf Basis des prognostizierten Behandlungsbedarfs, neu berechnet.

Versicherte mit Krankenvollversicherung	8,98 Mio.
Zusatzversicherungen	22,50 Mio.
Beitragseinnahmen	34,67 Mrd. Euro
Ausgezahlte Versicherungsleistungen	22,77 Mrd. Euro
Alterungsrückstellungen Bestand	169,43 Mrd. Euro

Tabelle 2.1.: Endgültige Werte für das Geschäftsjahr 2011, Stand: November 2012, vergleiche [11]

In Deutschland waren im Jahr 2011 fast neun Millionen Menschen (siehe Tabelle 1.1.) privat krankenversichert. Im Gegensatz zur GKV handelt es beim PKV-System um eine Individualversicherung. Im Allgemeinen wird beim Versicherungseintritt für jeden Versicherten eine individuelle Prämie berechnet, welche die Kosten des Versicherten bis zu seinem Tod abdecken soll[2]. Diese wird auf Basis des Erwartungswerts der künftigen Leistungsausgaben (unternehmensinterne oder bundesweite Statistiken) und der erwarteten

[1] 69,86 Mio. (vergleiche: [11] - Zahlen aus der KM1 für November 2012)

[2] Gilt nur für Tarife nach Art der Lebensversicherung. Kollektivtarife z.B. für Priester werden anderes kalkuliert.

Sterblichkeit bestimmt. Das erwartete Kündigungsverhalten und der Rechnungszins haben außerdem Einfluss auf die Höhe der Prämie. Individuelle Gesundheitsmerkmale werden durch Risikoaufschläge berücksichtigt. Versicherte mit dem gleichen Tarif[3] und die derselben Altersgruppe angehören, bilden dabei ein Kollektiv.

Die Prämie ist so kalkuliert, dass sie, trotz der im Alter erwartungsgemäß steigenden Leistungskosten, über die gesamte Versicherungsdauer hinweg konstant bleibt. Deshalb liegen die Prämien in den ersten Jahren über den zu erwartenden Leistungen. Aus dieser Differenz wird eine Rücklage (die Alterungsrückstellung) gebildet, welche die steigenden Kosten im Alter abdecken soll. Diese beiden Prozesse werden als Anspar- und Entnahmephase bezeichnet. Damit ein weitestgehend konstanter Beitrag sichergestellt werden kann, muss die Versicherungsprämie in einer Weise kalkuliert sein, die zum einen die dauernde Erfüllbarkeit der vom Versicherer versprochenen Leistungen sicherstellt, und zum anderen Prämiensteigerungen nur aus solchen Gründen zulässt, die vom Versicherer nicht zu beeinflussen sind. Dazu zählen beispielsweise nicht vorhersehbare Kostensteigerungen im Gesundheitswesen. Dabei ist zu berücksichtigen, dass gewisse Annahmen über die Entwicklung solcher Faktoren bereits bei der Kalkulation eines Tarifs getroffen werden und damit eine Erhöhung nur dann zulässig ist, wenn die tatsächlichen Anstiege die prognostizierten Werte noch übertreffen.

Seit dem 1. Januar 2009 sind die privaten Krankenversicherer gesetzlich dazu verpflichtet, im Fall des Krankenkassenwechsels eines Versicherten innerhalb der PKV, ihm die angesparte Alterungsrückstellung im Umfang des Basistarifs mitzugeben. Für den Versicherten hat das den Vorteil, dass er die Krankenkasse wechseln kann, ohne eine komplett neue Rückstellung aufbauen zu müssen. Da die mitgegebene Rückstellung allerdings nur den Basistarif umfasst, verliert der Kassenwechsler in der Regel einen Teil seiner Alterungsrückstellung.

Ein Wechsel hat nicht nur Auswirkungen auf ein einzelnes Individuum, sondern betrifft auch das Kollektiv von Versicherten, welches der Einzelne verlässt. In dem Zeitraum vom Tarifabschluss bis zum Wechsel können sich die Risikomerkmale des Versicherten verändern. Er kann z.B. eine chronische Krankheit entwickeln, nach einem Unfall zum Pflegefall werden, oder überhaupt keine nennenswerten Leistungen verursachen. Das Kollektiv hat normalerweise die Funktion, diese Schwankungen auszugleichen. Dies erfolgt dadurch, indem die Gesünderen mit ihren Prämien die Kosten der Morbideren abdecken. Falls der Versicherte zum Zeitpunkt des Wechsels „besonders gesund“[4] war, entsteht dem Kollektiv damit ein Schaden, da die Überschüsse aus seiner Prämie nicht mehr dafür verwendet werden können, die Kosten morbiderer Versicherter abzudecken.

[3] Der Eintrittszeitpunkt ist dabei entscheidend, da die einzelnen Tarife regelmäßig neu berechnet werden und dadurch verschiedene Tarifgenerationen entstehen. In einem Kollektiv werden nur Versicherte aus einer Tarifgeneration zusammengefasst.

[4] Das heißt der Versicherte beansprucht weniger Leistungen als der Durchschnitt.

Die Höhe des Schadens ist dabei schwer zu ermitteln. Es wird ein Modell benötigt, das auf Basis des aktuellen Gesundheitszustands eines Versicherten eine Prognose erstellt, wie viele Leistungen er bis zu seinem Tod noch verursachen wird. Dies war eine Motivation für die Entwicklung des Kohorten-Modells.

2.2. Kohorten Modell

Die Gesundheitsforen Leipzig haben ein Modell entwickelt, welches auf Basis der Morbiditätsinformationen eines Versicherten die Höhe seiner zu erwartenden Leistungskosten bestimmt. Damit ist es möglich für einen Versicherten zum Zeitpunkt des Kassenwechsels seine restliche Risikolast[5] zu berechnen und damit zu ermitteln, welche Auswirkungen der Wechsel auf das Kollektiv hat. Die nachfolgende Beschreibung des Modells stützt sich dabei auf den in der „Zeitschrift für Versicherungswesen“ erschienenen Artikel „Auf Leben und Tod - Spezifische Implikationen eines vermeintlich längeren Lebens für die Versicherungswirtschaft“(vergleiche [8] und [9]).

Das Ziel der Modellierung war es möglichst wenige und möglichst gleichartige „Bausteine“ zu finden, aus denen sich die verschiedenen Krankenversicherungsbiographien[6] zusammensetzen. Aus Millionen von Leistungsfällen, die über viele Jahre beobachtet wurden, konnten Milliarden von „individuellen“ Leistungspfaden analysiert werden. Dafür wurden Daten verwendet, die überwiegend aus dem Umfeld gesetzlich Versicherter stammen. Diese wurden durch Schätzungen, und vielfach durch einen bilanziellen Abgleich mit öffentlichen Gesamtdaten, vervollständigt.

Dabei ließ sich eine gewisse Homogenität, jeweils innerhalb von drei Lebensphasen, erkennen:

- Lebensabschnitt: „0-jährige“
- Lebensabschnitt: „Phase regen Lebens“
- Lebensabschnitt: „Prämortalitätsphase“

Die Abgrenzung der Lebensabschnitte erfolgt dabei vom Rand der Krankenversicherungsbiographie her (Geburtsdatum, Sterbedatum) jeweils auf den Tag genau. Die Längen dieser Abschnitte sind möglichst kurz gewählt und in ganzen Jahren angegeben. Kommt es zu Überlagerungen durch einen frühen Tod, so wird zuerst versucht den Abschnitt der „0-jährige“ vollständig abzubilden. Die verbleibenden Lebensjahre fallen anschließend in die „Prämortalitätsphase“. Bei dem Tod eines Neugeborenen, gibt es demzufolge nur einen

[5] Die restliche Risikolast ist die Summe aller Krankheitskosten, die eine Person im Laufe ihres restlichen Lebens verursacht. Eine zeitunabhängige Vergleichbarkeit wird durch den Übergang zu Barwerten der zukünftigen (zufälligen) Leistungen erreicht.

[6] Gesundheitskosten in Zusammenhang mit den Leistungszeitpunkten im Verlauf des Lebens eines Versicherten, werden als Krankenversicherungsbiographie bezeichnet.

Lebensabschnitt der betrachtet werden kann. Die „Phase des regen Lebens“ ist aber in der Regel die längste Phase und wird zusätzlich von zwei Arten von Ereignissen überlagert:

- Ereignis: Isolierte Ereignisse
- Ereignis: Singuläre Ereignisketten

Die restliche Risikolast ergibt sich aus einer Mischung von Kohorten[7] der Restbiographien. Eine Restbiographie bezeichnet dabei den Anteil einer Versichertenbiographie, der noch nicht eingetreten ist, das heißt, der noch in der Zukunft liegt. Jede Restbiographie setzt sich aus entsprechenden Anteilen aus der Phase der 0-jährigen, der Phase regen Lebens, aus isolierten Ereignissen, aus singulären Ereignisketten und der Prämortalitätsphase zusammen.

Die folgende Grafik veranschaulicht die Zusammensetzung einer Krankenversicherungsbiographie aus den drei Lebensabschnitten mit den überlagernden Ereignissen:

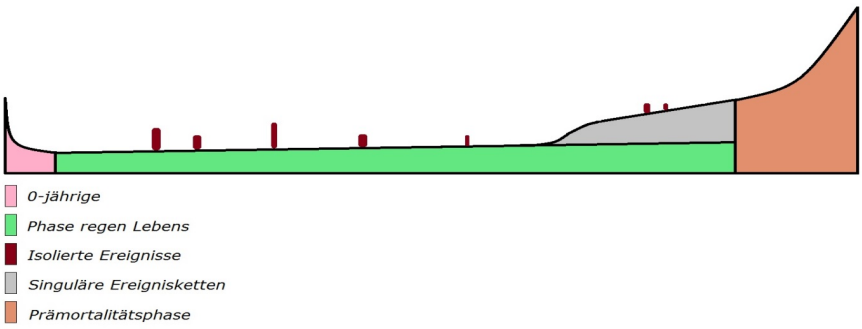


Abbildung 2.1.: Schema des Kostenverlaufs im Kohorten-Modell

Im Folgenden sollen die einzelnen Lebensabschnitte und Ergebnisse noch einmal im Detail beleuchtet und deren Besonderheiten dargestellt werden.

2.2.1. Lebensabschnitt: „0-jährige“

Dieser Lebensabschnitt beginnt mit dem „Tag der Geburt“ und dauert maximal ein Jahr. Für den Lebensabschnitt der 0-jährigen gilt: Mit jedem Tag des Überlebens steigt die restliche Lebenserwartung zunächst stark und später abgeschwächt. Dieser Teil der Modellierung ist Gegenstand des Artikels *„Demographischer Wandel“ in Deutschland - Mythen und Legenden* ([10]).

[7] Das Kohorten Merkmal ist eine gleiche restliche Lebenserwartung.

2.2.2. Lebensabschnitt: „Prämortalitätsphase“

Dieser Lebensabschnitt endet mit dem „Tag des Todes“. Der Abschnitt wird alters- und geschlechtsunabhängig auf maximal fünf Jahre virtuell „rückwärts“ ausgedehnt. Zu Beginn dieser Zeitspanne kann dadurch sehr gut an die Kosten der Vorphase angeknüpft werden. Die eigentlichen Kostenentwicklungen „kurz vor dem Tod“ sind über alle Altersbereiche in diesem Abschnitt abgebildet. Dieser Abschnitt ist auch das zentrale Thema des eingangs erwähnten Artikels.

2.2.3. Lebensabschnitt: „Phase regen Lebens“

Der Zeitraum vom ersten Tag des 2. Lebensjahres bis zum letzten Tag vor der Prämortalitätsphase entspricht dem dritten Lebenszeitabschnitt. Zur Kennzeichnung findet das Wort „rege“ Verwendung. Reges Leben findet in der Regel über eine Länge von vielen Jahren statt. In diesem Lebensabschnitt finden sich die Bereiche einer eher unauffälligen, „gleichbleibenden“ Kostenstruktur. Allerdings sind zwei Ereignisarten mit einem spezifischen Überlagerungscharakter aus dieser Zeitspanne abzugrenzen.

2.2.4. Ereignis: Isolierte Ereignisse

Seltene Ereignisse, welche die Gesundheit eines Versicherten beeinträchtigen, sich nicht ankündigen und bei denen der Eintrittszeitpunkt der eigentliche Auslöser von zeitlich begrenzten Leistungsabfolgen sind, werden nachfolgend als isolierte Ereignisse bezeichnet. Solche Ereignisse (z. B. Unfälle) treten selten, und in der Regel unabhängig voneinander, auf. Isolierte Ereignisse können auch in der Phase der 0-jährigen und in der Prämortalitätsphase auftreten. Diese werden für die Modellierung der Alters- und Geschlechtsabhängigkeit isolierter Ereignisse zwar hinzugezogen, aber als jeweilige Phasen-Leistungen in der Krankenversicherungsbiographie kumuliert. Auf diese Weise „verschwinden“ keine Leistungen und das Phänomen der isolierten Ereignisse kann innerhalb der Phase „regen Lebens“ separiert beschrieben werden.

2.2.5. Ereignis: Singuläre Ereignisketten

Singuläre Ereignisketten bezeichnen eine nachhaltige und beträchtliche Verschlechterung der Gesundheit des Versicherten, die in der Regel bis zu seinem Tod andauert. Es werden auch singuläre Ereignisketten berücksichtigt, die mit keinen direkten Kosten für die Krankenversicherung verbunden sind. Das sind zum Beispiel Erkrankungen bei denen der Patient pflegebedürftig wird. Diese Kosten werden dann von der Pflegeversicherung abgedeckt.

Das Ziel dieser Arbeit ist es, ein Modell zur Beschreibung der isolierten Ereignisse und der singulären Ereignisketten zu erstellen.

3. Theoretische Grundlagen für die Modellierung

In diesem Kapitel sollen die in der Arbeit verwendeten mathematischen Hilfsmittel vorgestellt und die wichtigsten Eigenschaften bewiesen werden. Dabei werden Grundlegende Kenntnisse im Bereich der Wahrscheinlichkeitstheorie vorausgesetzt und dass Begriffe wie Wahrscheinlichkeitsraum, Zufallsvariable, Dichte- und Verteilungsfunktion bekannt sind. Im Folgenden sei stets der Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, \mathbb{P})$ gegeben. Wir bezeichnen für eine Zufallsvariable $X : \Omega \to \mathbb{R}$, die Abbildung $F_X : \mathbb{R} \to [0, 1]$, welche definiert wird durch

$$F_X(x) := \mathbb{P}(X \leq x),$$

als **Verteilungsfunktion** von X. Wir schreiben $X \sim F_X$, wenn F_X die Verteilungsfunktion von X ist. Des Weiteren ist

$$\overline{F}_X(x) := 1 - F_X(x) = \mathbb{P}(X > x)$$

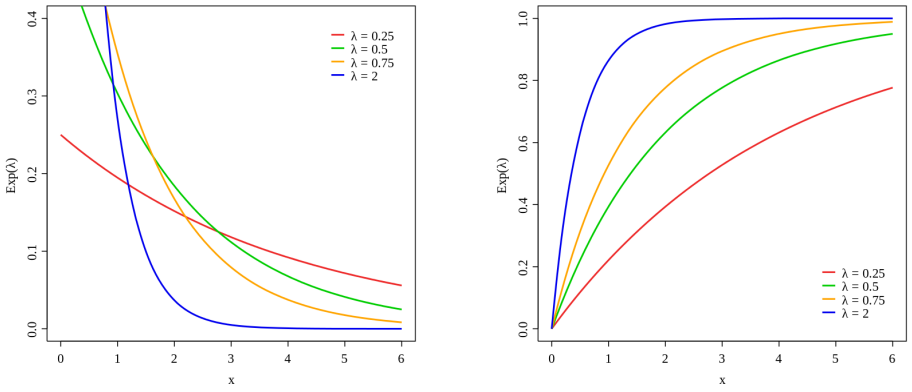
die Schwanzfunktion von X.

3.1. Verwendete Verteilungen

Zunächst werden wir die Verteilungen vorstellen, die in dieser Arbeit verwendet werden, und deren wichtigste Eigenschaften vorstellen.

Definition 3.1. *Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ heißt **exponentialverteilt zum Parameter** λ (kurz: $X \sim exp(\lambda)$), wenn sie die folgende Dichtefunktion besitzt:*

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases}$$



Dichtefunktion Verteilungsfunktion

Abbildung 3.1.: Dichte- und Verteilungsfunktion der Exponentialverteilung

Folgerung 3.2. *Sei $X \sim exp(\lambda)$ dann gilt:*

(i) Die Verteilungsfunktion von X ist:

$$F_X(x) = \int_{-\infty}^{x} f_\lambda(t)dt = \begin{cases} 1 - e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases}$$

(ii) Der Erwartungswert ist:

$$\mathbb{E}(X) = \int_0^\infty \lambda x e^{-\lambda x} dx = \left[-\frac{e^{-\lambda x}(\lambda x + 1)}{\lambda} \right]_0^\infty = \frac{1}{\lambda}$$

(iii) Die Varianz ist:

$$\begin{aligned} Var(X) &= \int_0^\infty \left(x - \frac{1}{\lambda} \right)^2 \lambda e^{-\lambda x} dx \\ &= \int_0^\infty \left(x^2 - 2x\frac{1}{\lambda} + \frac{1}{\lambda^2} \right) \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^\infty x^2 e^{-\lambda x} dx - 2 \int_0^\infty x e^{-\lambda x} dx + \frac{1}{\lambda} \int_0^\infty e^{-\lambda x} dx \\ &= \lambda \int_0^\infty x^2 e^{-\lambda x} dx - \frac{2}{\lambda^2} + \frac{1}{\lambda^2} \\ &= \lambda \Big(\underbrace{\left[-\frac{1}{\lambda} x^2 e^{-\lambda x} \right]_0^\infty}_{0} + \frac{2}{\lambda} \int_0^\infty x e^{-\lambda x} dx \Big) - \frac{1}{\lambda^2} \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\ &= \frac{1}{\lambda^2} \end{aligned}$$

(iv) Die Exponentialverteilung ist gedächtnislos (auch Nichtalterungseigenschaft genannt), d.h. es gilt für alle $x, t > 0$:

$$\begin{aligned}\mathbb{P}(X > x + t \mid X > t) &= \frac{\mathbb{P}(X > x + t, X > t)}{\mathbb{P}(X > t)} \\ &= \frac{\mathbb{P}(X > x + t)}{\mathbb{P}(X > t)} \\ &= \frac{e^{-\lambda(x+t)}}{e^{-\lambda t}} \\ &= e^{-\lambda x} = \mathbb{P}(X > x)\end{aligned}$$

Die Exponentialverteilung hat die besondere Eigenschaft der Gedächtnislosigkeit und es lässt sich sogar zeigen, dass sie die einzige, absolut stetige Verteilung[1] mit dieser Eigenschaft ist.

Lemma 3.3. *Sei X eine positive Zufallsvariable mit absolut stetiger Verteilungsfunktion, dann gilt $X \sim exp(\lambda)$ genau dann, wenn für alle $x, t > 0$ gilt, dass*

$$\mathbb{P}(X > x + t \mid X > t) = \mathbb{P}(X > x) \tag{3.1}$$

Beweis: $\Rightarrow$ Wir haben bereit gezeigt, dass wenn $X \sim exp(\lambda)$, dann gilt 3.1.
$\Leftarrow$ Sei umgekehrt X eine positive Zufallsvariable mit absolut stetiger Verteilungsfunktion, die die Gleichung 3.1 erfüllt. Wir definieren $g(x) := \overline{F}_X(x) = \mathbb{P}(X > x)$. Da g stetig ist, gilt für $x, y > 0$ gilt:

$$\begin{aligned}g(x + y) &= \mathbb{P}(X > x + y) \\ &= \mathbb{P}(X > x + y \mid X > y)\mathbb{P}(X > y) \\ &= \mathbb{P}(X > x)\mathbb{P}(X > y) = g(x)g(y)\end{aligned}$$

Durch n-fache Anwendung folgt für alle $n \in \mathbb{N}$:

$$g(1) = g(\underbrace{\frac{1}{n} + ... + \frac{1}{n}}_{n-mal}) = \left(g(\frac{1}{n})\right)^n$$

und somit insbesondere auch $g(\frac{1}{n}) = (g(1))^{\frac{1}{n}}$. Da $X > 0$ fast sicher, existiert ein $n \in \mathbb{N}$ mit $g(1/n) > 0$. Außerdem existiert wegen $0 < g(1) \leq 1$, ein $\lambda \geq 0$ mit $g(1) = e^{-\lambda}$. Für beliebige $p, q \in \mathbb{N}$ gilt

$$g(\frac{p}{q}) = g(\frac{1}{q})^p = g(1)^{\frac{p}{q}} = e^{-\lambda * \frac{p}{q}}$$

[1] Im diskreten Fall ist dies die geometrische Verteilung.

und somit $g(r) = e^{-\lambda r}$ für alle $r \in \mathbb{Q}^+$. Aufgrund der Stetigkeit folgt daraus für alle $x \in \mathbb{R}^+$

$$g(x) = e^{-\lambda x}.$$

□

Als nächstes wollen wir eine Wahrscheinlichkeitsverteilung vorstellen, die für unabhängige, zufällige Ereignisse mit konstanter Eintrittrate modelliert, wie viel davon in ein festes Zeitintervall fallen. Ein Beispiel dafür wäre die Anzahl der Patienten die innerhalb einer Stunde in eine Praxis kommen, wenn die Ankunftszeiten zufällig und unabhängig voneinander und die Patienten mit einer konstanten Rate die Praxis aufsuchen.

Definition 3.4. *Eine diskrete Zufallsvariable $X : \Omega \to \mathbb{N}$ heißt **poissonverteilt zum Parameter** $\lambda \in \mathbb{R}_{>0}$ (kurz $X \sim Poi(\lambda)$), wenn gilt:*

$$Poi_\lambda(k) := \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, 2, ...$$

Folgerung 3.5. *Sei $X \sim Poi(\lambda)$ dann gilt:*

(i) Die Verteilungsfunktion der Poisson-Verteilung ist:

$$F_\lambda(n) = \sum_{k=0}^{n} Poi_\lambda(k) = e^{-\lambda} \sum_{k=0}^{n} \frac{\lambda^k}{k!}$$

(ii) Der Erwartungswert ist:

$$\begin{aligned} \mathbb{E}(X) &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = 0 + \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned}$$

Der Parameter λ der Poisson-Verteilung kann also als die erwartete Ereignishäufigkeit pro Zeiteinheit interpretiert werden.

(iii) Die Varianz ist:

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} \frac{((k-1)+1)\lambda^k}{(k-1)!} \\
&= e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} + e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\
&= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda e^{-\lambda} \sum_{1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda^2 + \lambda
\end{aligned}$$

$$Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

(iv) Seien X_1 und X_2 unabhängige poissonverteilte Zufallsvariablen mit $X_1 \sim Poi(\lambda_1)$ und $X_2 \sim Poi(\lambda_2)$, dann gilt für $X := X_1 + X_2$:

$$\begin{aligned}
\mathbb{P}(X = x) &= \sum_{k=0}^{x} \mathbb{P}(X_1 = k)\mathbb{P}(X_2 = x - k) \\
&= e^{-\lambda_1} e^{-\lambda_2} \sum_{k=0}^{x} \frac{\lambda_1^k}{k!} \frac{\lambda_2^{x-k}}{(x-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{x!} \sum_{k=0}^{x} \frac{x!}{k!(x-k)!} \lambda_1^k \lambda_2^{x-k} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^x}{x!} \\
\Rightarrow X &\sim Poi(\lambda_1 + \lambda_2)
\end{aligned}$$

Das heißt, die Summe von poissonverteilten Zufallsvariablen ist wieder poissonverteilt.

Wir wollen uns an dieser Stelle kurz überlegen, weshalb gerade die Poisson-Verteilung für unser Beispiel mit den Patienten in der Praxis angewendet werden kann. Dazu unterteilen wir den betrachteten Zeitraum t in n gleiche Teilstücke der Länge $\frac{t}{n}$. Dabei ist n so klein gewählt, dass in jedem dieser Teilintervalle maximal ein Patient in die Praxis kommt. Die Wahrscheinlichkeit, dass in einem dieser Intervalle ein Patient eintrifft, bezeichnen wir mit p_n. Wie bereits beschrieben nehmen wir an, dass die Ankunftszeiten unabhängig voneinander sind. Das heißt, wir haben ein Bernoulli-Experiment der Länge

n mit Erfolgswahrscheinlichkeit p_n. Wählen wir nun p_n so, dass für große n die erwartete Anzahl von Patienten konstant ist, dann ist die Anzahl der zufälligen Ankunftszeitpunkte poissonverteilt. Das folgende Lemma fasst diese Überlegung zusammen.

Lemma 3.6. *Sei $\{p_n\}_{n\in\mathbb{N}} \in (0,1)$ eine Folge mit der Eigenschaft $\lim_{n\to\infty} np_n = \lambda$ für eine beliebige Konstante $\lambda \in (0,\infty)$, dann gilt für alle $k \in \mathbb{N}_0$*

$$\lim_{n\to\infty} Bi_{n,p_n}(k) = Poi_\lambda(k).$$

Dabei bezeichnet Bi_{n,p_n} die Verteilungsfunktion der Binomialverteilung mit Parametern n und p_n.

Beweis: Wir zeigen, dass der Grenzwert $n \to \infty$ der Verteilungsfunktion $Bi_{n,p_n}(k)$ an der Stelle k gleich dem Wert der Poissonverteilung an der Stelle k ist.

$$\begin{aligned}
\lim_{n\to\infty} Bi_{n,p_n}(k) &= \lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \underbrace{\left(\frac{n(n-1)(n-2)...(n-k+1)}{n^k}\right)}_{\to 1} \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1} \\
&= \frac{\lambda^k e^{-\lambda}}{k!}
\end{aligned}$$

□

An dieser Stelle wollen wir noch eine weitere Verteilung einführen, die in einer interessanten Beziehung sowohl zur Exponentialverteilung, als auch zur Poisson-Verteilung steht.

Definition 3.7. *Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ heißt **gammaverteilt zu den Parametern α und r** (kurz: $X \sim \gamma(\alpha, r)$), wenn sie die folgende Dichtefunktion besitzt:*

$$\gamma_{\alpha,r}(x) = \begin{cases} \frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x} & \text{für } x > 0 \\ 0 & \text{für } x \leq 0 \end{cases}$$

Wobei $\Gamma : (0,\infty) \to (0,\infty)$ die Gamma-Funktion ist:

$$\Gamma(r) = \int_0^\infty y^{r-1} e^{-y} dy$$

mit $r > 0$.

Es lässt sich leicht zeigen, dass die Exponentialverteilung ein Spezialfall der Gamma-Verteilung ist.

Folgerung 3.8. *Im Fall $r = 1$, gilt für alle $x > 0$*

$$\gamma_{\alpha,1}(x) = \frac{\alpha^1}{\Gamma(1)} x^0 e^{-\alpha x} = \frac{\alpha}{\int_0^\infty e^{-y} dy} e^{-\alpha x} = \alpha e^{-\alpha x}.$$

Das ist die Dichte Exponentialverteilung.

Lemma 3.9. *Die Summe zweier unabhängiger gammaverteilter Zufallsgrößen X_1, X_2 mit Parametern α und r_1 bzw. α und r_2 ist gammaverteilt mit Parametern α und $r_1 + r_2$. Insbesondere ist für $k \in \mathbb{N}$ die Gamma-Verteilung mit Parametern α und k identisch zur Verteilung der Summe von k unabhängigen, zum Parameter α exponentialverteilten Zufallsgrößen.*

Beweis: Es gilt

$$\begin{aligned}
\mathbb{P}(X_1 + X_2 \leq x) &= \int_0^x \gamma_{\alpha,r_1}(t)\gamma_{\alpha,r_2}(x-t)dt \\
&= \frac{\alpha^{r_1}}{\Gamma(r_1)} \frac{\alpha^{r_2}}{\Gamma(r_2)} e^{-\alpha x} \int_0^x t^{r_1-1}(s-t)^{r_2-1} dt \\
&= \gamma_{\alpha,r_1+r_2}(t) \frac{\Gamma(r_1+r_2)}{\Gamma(r_1)\Gamma(r_2)} \int_0^x t^{r_1-1}(s-t)^{r_2-1} dt \\
&= \gamma_{\alpha,r_1+r_2}(t) \frac{\Gamma(r_1+r_2)}{\Gamma(r_1)\Gamma(r_2)} \underbrace{\int_0^1 u^{r_1-1}(1-u)^{r_2-1} du}_{Eulersche Beta-Integral} \\
&= \gamma_{\alpha,r_1+r_2}(t) \frac{\Gamma(r_1+r_2)}{\Gamma(r_1)\Gamma(r_2)} \frac{\Gamma(r_1)\Gamma(r_2)}{\Gamma(r_1+r_2)} \\
&= \gamma_{\alpha,r_1+r_2}(t)
\end{aligned}$$

mit $u = \frac{s}{t}$. Der zweite Teil der Aussage folgt direkt aus dieser Beziehung. □

3.2. Poisson Prozess

Ein weiteres Hilfsmittel, dass zur Modellierung verwendet wird, sind Stochastische Prozesse. Diese eignen sich sehr gut dazu, geordnete zufällige Vorgänge zu beschreiben. Allgemein lässt sich ein stochastischer Prozess wie folgt definieren.

Definition 3.10. *Sei $(Z, \mathcal{Z})$ ein mit einer σ-Algebra versehener Raum, dann ist ein* ***Stochastischer Prozess*** *eine Familie von Zufallsvariablen $\{X_t, t \in T\}$ mit $X_t : \Omega \to Z$ und einer beliebigen nichtleeren Indexmenge T. Das heißt X ist eine Abbildung*

$$X : \Omega \times T \to Z,\ (\omega, t) \mapsto X_t(\omega),$$

so dass $X_t : \omega \mapsto X_t(\omega)$ für alle $t \in T$ eine messbare Abbildung ist. Z heißt dann die ***Zustandsmenge*** *und $(Z, \mathcal{Z})$ der* ***Zustandsraum****. Ein stochastische Prozess heißt* ***zeitdiskret****, wenn T abzählbar ist, z.B. $T = \mathbb{N}_0$. Ansonsten heißt er* ***zeitstetig****. Weiterhin heißt ein Prozess mit diskreten Zustandsraum Z* ***wertdisktret*** *oder auch* ***Punktprozess****.*

Für unsere Zwecke reicht es aus als Zustandsraum die reellen Zahlen mit der borelschen σ-Algebra zu betrachten. Da wir eine zeitliche Entwicklung untersuchen wollen reicht es außerdem, wenn T eine Teilmenge der reellen Zahlen ist und als Zeit betrachtet wird. Im Folgenden sind die $\{X_t, t \geq 0\}$ deshalb reellwertige Zufallsvariablen. Mit dieser Einschränkung werden wir nun einige Eigenschaften vorstellen, die ein stochastischer Prozess haben kann.

(i) Ein stochastische Prozess heißt **stationär**, wenn für alle $n \geq 0, s \geq 0$, sowie alle $0 \leq t_1 \leq t_2 \leq ... \leq t_k$ und $x_1, x_2, ..., x_k \in \mathbb{R}$ gilt:

$$\mathbb{P}(X_{t_1+s} \leq x_1, X_{t_2+s} \leq x_2, ..., X_{t_k+s} \leq x_k) = \mathbb{P}(X_{t_1} \leq x_1, X_{t_2} \leq x_2, ..., X_{t_k} \leq x_k)$$

Das bedeutet, das zufällige Verhalten des Prozesses hängt nicht vom Zeitpunkt der Beobachtung ab.

(ii) Analog besitzt ein stochastischer Prozess **stationäre Zuwächse**, wenn für alle $n \geq 0, s \geq 0$ und für alle $0 \leq t_1 \leq t_2 \leq ... \leq t_k \in T$ die Verteilung des Zufallsvektors $(X_{t_1+s} - X_{t_0+s}, X_{t_2+s} - X_{t_1+s}, ..., X_{t_n+s} - X_{t_{n-1}+s})$ nicht von s abhängt.

(iii) Ein stochastischer Prozess besitzt **Unabhängige Zuwächse**, wenn die Zufallsvariablen $X_{t_0}, X_{t_1} - X_{t_0}, ..., , X_{t_n} - X_{t_{n-1}}$ für alle n=1,2,... und $0 \leq t_0 < t_1 < ... < t_n$ unabhängig sind.

Stochastische Prozesse werden häufig zur Modellierung der Eintrittszeitpunkte von zufälligen Ereignissen verwendet. Mit Blick auf die isolierten Ereignisse interessiert uns allerdings eher, wie viele Ereignisse in einem bestimmten Zeitraum eintreten werden. Dies führt zu der folgenden Definition.

Definition 3.11. *Sei $T_1, T_2, ... : \Omega \to \mathbb{R}, \omega \to [0, \infty)$ eine Folge von unabhängig und identisch verteilten Zufallsvariablen und $S_n := T_1 + ... + T_n$ für alle $n \in \mathbb{N}$, dann ist $N := \{N_t, t \geq 0\}$ mit*

$$N_t = \sum_{k=1}^{\infty} \mathbb{1}(S_k \leq t)$$

ein stochastischer Prozess und wird als Zählprozess bezeichnet.

Betrachten wir wieder unser Beispiel mit den Patienten die in eine Praxis kommen, dann können wir die T_n als Wartezeit auf den nächsten Patienten auffassen, nachdem gerade einer angekommen ist. Die S_n hingegen sind die konkreten Ankunftszeitpunkte. Deshalb ist es naheliegend die T_n als **Zwischenankunftszeiten** zu bezeichnen und die S_n als n-te **Sprungzeit**. Prozesse dieser Art werden z.B. auch in der Zuverlässigkeitstheorie eingesetzt, um die Ausfälle einer Komponente in einem bestimmten Zeitraum zu zählen. Der Zusammenhang zur vorher eingeführten Poisson-Verteilung wird im Folgenden deutlich.

Definition 3.12. *Ein Zählprozess $\{N_t, t \geq 0\}$ mit exponentialverteilten Zwischenankunftszeiten $T_n \sim exp(\lambda)$ heißt* ***homogener Poisson-Prozess mit der Intensität*** *λ.*

In dem nachfolgendem Theorem werden die wichtigsten Eigenschaften und äquivalenten Definitionen des Poisson-Prozesses deutlich. Vorher benötigen wir jedoch folgende Definition:

Definition 3.13. *Seien $X_1, X_2, ..., X_n$, mit $n \in \mathbb{N}$ Zufallsvariablen, dann bezeichnen wir die geordneten Variablen $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ als die* ***Ordnungsstatistik*** *der Variablen $\{X_i : 1 \leq i \leq n\}$.*

Satz 3.14. *Seien $X_1, X_2, ..., X_n$, mit $n \in \mathbb{N}$, unabhängige und identisch verteilte Zufallsvariablen mit Dichte f und sei $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ die entsprechende Ordnungsstatistik. Dann gilt für die gemeinsame Dichte:*

$$f_{X_{1:n}, X_{2:n}, ..., X_{n:n}}(t_1, t_2 ..., t_n) = \begin{cases} n! f(t_1) f(t_2) ... f(t_n) & \text{falls } t_1 \leq t_2 \leq ... \leq t_n \\ 0 & \text{sonst} \end{cases} \tag{3.2}$$

Beweis: Da aufgrund der Definition der Ordnungsstatistik die Werte aufsteigend geordnet sind, ist die Dichte gleich 0, wenn die Bedingung $t_1 \leq t_2 \leq ... \leq t_n$ nicht erfüllt ist. Sei nun diese Bedingung erfüllt. Dann existieren genau $n!$ Möglichkeiten die Zufallsvariablen $X_1, X_2, ..., X_n$ anzuordnen. Zum Beispiel gilt für $n = 2$, dass $\{X_{1:2} = t_1, X_{2:2} = t_2\}$ genau dann eintritt wenn entweder $\{X_1 = t_1, X_2 = t_2\}$, oder $\{X_2 = t_1, X_1 = t_2\}$ eintritt. Diese Möglichkeiten unterscheiden sich nur durch Permutation und besitzen somit die gleiche Dichte. Deshalb reicht es aus nur eine Möglichkeit zu betrachten und das Ergebnis

anschließend mit der Anzahl der möglichen Permutationen $n!$ zu multiplizieren. Es gilt für die einfachste Möglichkeit

$$f_{X_1,X_2,...,X_n}(t_1,t_2...,t_n) = f(t_1)f(t_2)...f(t_n)$$

da die Zufallsvariablen $X_1, X_2, ..., X_n$ unabhängig voneinander sind. Durch Multiplikation mit $n!$, erhält man 3.2. □

Theorem 3.15. *Die folgenden Aussagen sind äquivalent ([15]):*

(i) $\{N_t, t \geq 0\}$ *ist ein Poisson-Prozess mit der Intensität* λ

(ii) Die Zufallsvariablen N_t *sind poissonverteilt zum Parameter* λt *für alle* $t \geq 0$.

Unter der Bedingung $\{N_t = n\}$, *hat für beliebige* $n = 1, 2, ...$ *der Zufallsvektor* $(S_1, S_2, ..., S_n)$, *die gleiche Verteilung wie die Ordnungsstatistik von n unabhängigen, in* $[0, t]$ *gleichverteilten Zufallsvariablen.*

(iii) Der stochastische Prozess $\{N_t, t \geq 0\}$ *hat unabhängige Zuwächse und es gilt* $\mathbb{E}(N_1) = \lambda$.

Unter der Bedingung $\{N_t = n\}$, *hat für beliebige* $n = 1, 2, ...$ *der Zufallsvektor* $(S_1, S_2, ..., S_n)$, *die gleiche Verteilung wie die Ordnungsstatistik von n unabhängigen, in* $[0, t]$ *gleichverteilten Zufallsvariablen.*

(iv) Der stochastische Prozess $\{N_t, t \geq 0\}$ *hat unabhängige und stationäre Zuwächse und es gilt für* $h \to 0$*:*

$$\begin{aligned} \mathbb{P}(N_h = 0) &= 1 - \lambda h + o(h), \text{ und} \\ \mathbb{P}(N_h = 1) &= \lambda h + o(h) \end{aligned}$$

(v) Der stochastische Prozess $\{N_t, t \geq 0\}$ *hat unabhängige und stationäre Zuwächse. Außerdem gilt für jedes* $t \geq 0$ *das* $N_t \sim Poi(\lambda t)$.

Beweis:

- $(i) \Rightarrow (ii)$: Seien $T_{ii} \in \mathbb{N}$ die Zwischenankunftszeiten des Poisson-Prozesses $\{N_t, t \geq 0\}$, dann folgt, dass $S_n = \sum_{i=1}^{n} T_i$ eine Summe von n unabhängigen und zum Parameter λ exponentialverteilten Zufallsvariablen ist. Nach 3.9 gilt also $S_n \sim \gamma_{\lambda,n}$. Hieraus folgt $\mathbb{P}(N_t = 0) = \mathbb{P}(S_1 > t) = e^{-\lambda t}$ und damit gilt:

$$\begin{aligned}
\mathbb{P}(N_t = n) &= \mathbb{P}(N_t \geq n) - \mathbb{P}(N_t \geq n+1) \\
&= \mathbb{P}(S_n \leq t) - \mathbb{P}(S_{n+1} \leq t) \\
&= \int_0^t \frac{\lambda^n v^{n-1}}{(n-1)!} e^{-\lambda v} \mathrm{d}v - \int_0^t \frac{\lambda^{n+1} v^n}{n!} e^{-\lambda v} \mathrm{d}v \\
&= \int_0^t \frac{\mathrm{d}}{\mathrm{d}v} \left(\frac{(\lambda v)^n}{n!} e^{-\lambda v} \right) \mathrm{d}v \\
&= \frac{(\lambda t)^n}{n!} e^{-\lambda t}
\end{aligned}$$

Dies gilt für jedes $n \geq 1$, und damit folgt, dass $N_t \sim Poi(\lambda t)$. Dies ist der erste Teil von (ii) und für den zweiten Teil betrachten wir die gemeinsame Dichte $f_{S_1,...,S_{n+1}}(t_1, ..., t_{n+1})$ von $S_1, ..., S_{n+1}$. Für beliebige $t_0 = 0 \leq t_1 \leq ... \leq t_n \leq t_{n+1}$ gilt aufgrund des Transformationssatzen (vgl. A.1)

$$\begin{aligned}
f_{S_1,...,S_{n+1}}(t_1, ..., t_{n+1}) &= f_{T_1,T_2,...,T_{n+1}}(t_1, t_2 - t_1, ..., t_{n+1} - t_n) * |det(DA)| \\
&= \prod_{k=1}^{n+1} \lambda e^{-\lambda(t_k - t_{k-1})} = \lambda^{n+1} e^{-\lambda t_{n+1}}
\end{aligned}$$

und 0 sonst. Dabei ist A die Transformation $t_i \mapsto t_i - t_{i-1}$ und damit gilt $|det(DA)| = 1$. Somit gilt unter der Bedingung $N_t = n$ und $0 \leq t_1 \leq ... \leq t_n \leq t$ für die gemeinsame bedingte Dichte

$$\begin{aligned}
f_{S_1,...,S_n}(t_1, ..., t_n \mid N_t = n) &= f_{S_1,...,S_n}(t_1, ..., t_n \mid S_1 \leq t, ..., S_n \leq t, S_{n+1} > t) \\
&= \frac{\int_t^\infty \lambda^{n+1} e^{-\lambda x_{n+1}} dx_{n+1}}{\int_0^t \int_{x_1}^t ... \int_{x_{n-1}}^t \int_t^\infty \lambda^{n+1} e^{-\lambda x_{n+1}} dx_{n+1} ... dx_1} \\
&= \frac{n!}{t^n}
\end{aligned}$$

und $f_{S_1,...,S_n}(t_1, ..., t_n | N_t = n) = 0$ sonst. Nach 3.2 ist dies die Dichte der Ordnungsstatistik von n unabhängigen in[0,t] gleichverteilten Zufallsvariablen und damit der zweite Teil dieses Beweisschritts.

- $(ii) \Rightarrow (iii)$:

  Aufgrund von (ii) gilt $N_t \sim Poi(\lambda t)$ und damit ist $\mathbb{E}N_1 = \lambda$. Nun zeigen wir, dass der Prozess unabhängige Zuwächse hat. Dazu seien $x_1, ..., x_n \in \mathbb{N}$ und $t_0 = 0 \leq t_1 \leq ... \leq t_n$, weiterhin sei $x = x_1 + ... + x_n$. Es gilt für die Zuwächse

$$\begin{aligned}
\mathbb{P}(N_{t_1} - N_{t_0} = x_1) &= \mathbb{P}\big(\sum_{k=1}^{\infty} \mathbb{1}_{\{S_k \leq t_1\}} - \mathbb{1}_{\{S_k \leq t_0\}} = x_1\big) \\
&= \mathbb{P}\big(\sum_{k=1}^{\infty} \mathbb{1}_{\{S_k \in (t_0, t_1]\}} = x_1\big) \\
&= \mathbb{P}(S_1 \in (t_0, t_1], ..., S_{x_1} \in (t_0, t_1])
\end{aligned}$$

  Damit können, unter Ausnutzung der Nebenbedingung, schreiben

$$\begin{aligned}
& \mathbb{P}(N_{t_1} - N_{t_0} = x_1, ..., N_{t_n} - N_{t_{n-1}} = x_n \mid N_{t_n} = x) \\
= \; & \mathbb{P}(S_1 \in (t_0, t_1], ...S_{x_1} \in (t_0, t_1], ..., S_{x-x_n+1} \in (t_{n-1}, t_n], ..., S_x \in (t_{n-1}, t_n] \mid N_{t_n} = x) \\
= \; & \int_{(t_0,t_1]^{x_1} \times ... \times (t_{n-1},t_n]^{x_n}} f_{S_1,...,S_x}(y_1, ..., y_x \mid N_{t_n} = x) dy_x ... dy_1 \\
= \; & \frac{x!}{t_n^x} \int_{(t_0,t_1]^{x_1}} \mathbb{1}_{\{y_1 \leq y_2\}} ... \mathbb{1}_{\{y_{x_1-1} \leq y_{x_1}\}} dy_{x_1} ... dy_1 * ... \\
& * \int_{(t_{n-1},t_n]^{x_n}} \mathbb{1}_{\{y_{x-x_n+1} \leq y_{x-x_n+2}\}} ... \mathbb{1}_{\{y_{x-1} \leq y_x\}} dy_x ... dy_{x-x_n+1} \\
= \; & \frac{x!}{t_n^x} \int_{t_0}^{t_1} \int_{y_1}^{t_1} ... \int_{y_{x_1-1}}^{t_1} dy_{x_1} ... dy_1 * ... * \int_{t_{n-1}}^{t_n} \int_{y_{x-x_n+1}}^{t_n} ... \int_{y_{x-1}}^{t_n} dy_x ... dy_{x-x_n+1}
\end{aligned}$$

  Nach Nebenrechnung A.3 können wir die Integrale direkt angeben und damit ist

$$\begin{aligned}
& \mathbb{P}(N_{t_1} - N_{t_0} = x_1, ..., N_{t_n} - N_{t_{n-1}} = x_n \mid N_{t_n} = x) \\
&= \frac{x!}{t_n^x} \prod_{k=1}^{n} \frac{(t_k - t_{k-1})^{x_k}}{x_k!} \\
&= \frac{x!}{x_1! x_2! ... x_n!} \prod_{k=1}^{n} \big(\frac{t_k - t_{k-1}}{t_n}\big)^{x_k}.
\end{aligned} \tag{3.3}$$

  Diese Beziehung werden wir noch an einigen Stellen im weiteren Beweis benötigen.

Damit gilt

$$\begin{aligned}
\mathbb{P}(\bigcap_{k=1}^{n}\{N_{t_k}-N_{t_{k-1}}=x_k\}) &= \mathbb{P}(N_{t_1}-N_{t_0}=x_1,...,N_{t_n}-N_{t_{n-1}}=x_n)\\
&= \mathbb{P}(N_{t_1}-N_{t_0}=x_1,...,N_{t_n}-N_{t_{n-1}}=x_n \mid N_{t_n}=x)\mathbb{P}(N_{t_n}=x)\\
&= \frac{(\lambda t_n)^x}{x!}e^{-\lambda t_n}\frac{x!}{x_1!...x_n!}\prod_{k=1}^{n}\left(\frac{t_k-t_{k-1}}{t_n}\right)^{x_k}\\
&= \prod_{k=1}^{n}\frac{(\lambda(t_k-t_{k-1}))^{x_k}}{x_k!}e^{-\lambda(t_k-t_{k-1})}
\end{aligned}$$

und damit hat der Zählprozess N_t unabhängige Zuwächse.

- $(iii) \Rightarrow (iv)$:

  Aus (iii) folgt, dass der Zufallsvektor $(S_1,...,S_m)$, unter der Bedingung $N(t_n+h)=m$, die gleiche Verteilung hat, wie die Ordnungstatistik vom m unabhängigen in $[0,t_n+h]$ gleichverteilten Zufallsvariablen hat. Deshalb gilt für beliebige $x_1,...,x_n \in \mathbb{N}$ mit $x_1+...+x_n \leq m$, $t_0=0\leq t_1\leq ...\leq t_n$, $h>0$ und $k=m-(x_1+...+x_n)$, unter Verwendung von 3.3

$$\begin{aligned}
&\mathbb{P}(\bigcap_{k=1}^{n}\{N_{t_k+h}-N_{t_{k-1}+h}=x_k\} \mid N_{t_n+h}=m)\\
=\ &\mathbb{P}(\bigcap_{k=1}^{n}\{N_{t_k+h}-N_{t_{k-1}+h}=x_k\}\cap\{N_n-N_0=k\}\\
=\ &\frac{m!}{x_1!x_2!...x_n!}\prod_{k=1}^{n}\left(\frac{t_k+h-t_{k-1}-h}{t_n+h}\right)^{x_k}\frac{1}{k!}\left(\frac{n-0}{t_n+h}\right)^{k}\\
=\ &\frac{m!}{x_1!x_2!...x_n!}\prod_{k=1}^{n}\left(\frac{t_k-t_{k-1}}{t_n+h}\right)^{x_k}\frac{1}{k!}\left(\frac{T_n+h-t_n}{t_n+h}\right)^{k}\\
=\ &\mathbb{P}(\bigcap_{k=1}^{n}\{N_{t_k}-N_{t_{k-1}}=x_k\}\cap\{N_{t_n+h}-N_{t_n}=k\}\\
=\ &\mathbb{P}(\bigcap_{k=1}^{n}\{N_{t_k}-N_{t_{k-1}}=x_k\} \mid N_{t_n+h}=m)
\end{aligned}$$

Aufgrund der Formel der totalen Wahrscheinlichkeit folgt damit, dass N_t stationäre Zuwächse besitzt. Die Gleichverteilungseigenschaft aus (iii) liefert außerdem für $0<h<1$, unter Verwendung von 3.3

$$\begin{aligned}\mathbb{P}(N_h = 0) &= \sum_{k=0}^{\infty} \mathbb{P}(N_h = 0, N_1 - N_h = k) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(N_1 = k)\mathbb{P}(N_h = 0, N_1 - N_h = k \mid N_1 = k) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(N_1 = k)(1-h)^k\end{aligned}$$

und somit gilt

$$\begin{aligned}\frac{1}{h}(1 - \mathbb{P}(N_h = 0)) &= \frac{1}{h}\left(1 - \sum_{k=0}^{\infty} \mathbb{P}(N_1 = k)(1-h)^k\right) \\ &= \sum_{k=1}^{\infty} \mathbb{P}(N_1 = k)\frac{1-(1-h)^k}{h}\end{aligned}$$

Da $(1-h)^k \geq 1 - kh$ für beliebige $0 < h < 1$ und $k = 1, 2, ...$ gilt, folgt, dass die Funktionen $g_h(k) = \frac{1-(1-h^k)}{h}$ die gemeinsame Schranke $g(k) = k$ besitzen. Diese Schranke ist integrierbar, da gilt

$$\sum_{k=1}^{\infty} k\mathbb{P}(N_1 = k) = \mathbb{E}(N_1) = \lambda < \infty.$$

Durch die Vertauschung von Summation und Grenzwert ergibt sich

$$\lim_{h\to 0} \frac{1}{h}\mathbb{P}(N_h > 0) = \lambda$$

und damit der erste Grenzwert von (iv). Analog dazu gilt

$$\lim_{h\to 0} \frac{1}{h}\mathbb{P}(N_h = 1) = \lim_{h\to 0} \sum_{k=1}^{\infty} \mathbb{P}(N_1 = k)k(1-h)^{k-1} = \lambda$$

was äquivalent zur zweiten Bedingung in (iv) ist.

- $(iv) \Rightarrow (v)$:

  Sei $p_n(t) := \mathbb{P}(N_t = n)$ mit $n \in \mathbb{N}$ und $t \geq 0$, dann gilt für $h > 0$

$$\begin{aligned} p_0(t+h) &= \mathbb{P}(N_t = 0, N_{t+h} - N_t = 0) \\ &= \mathbb{P}(N_t = 0)\mathbb{P}(N_{t+h} - N_t = 0) \\ &= \mathbb{P}(N_t = 0)\mathbb{P}(N_h = 0) \\ &= p_0(t)(1 - \lambda h + o(h)) \end{aligned} \tag{3.4}$$

  und für $t \geq h > 0$

$$p_0(t) = p_0(t-h)(1 - \lambda h + o(h)) \tag{3.5}$$

  Damit ist $p_0(t)$ stetig in $(0, \infty)$ und rechtsstetig im Punkt $t = 0$. Da $p_0(t-h) = p_0(t) + o(1)$ folgt aus 3.4 und 3.5, dass für beliebige $h \geq -t$ gilt

$$\frac{p_0(t+h) - p_0(t)}{h} = -\lambda p_0(t) + o(1)$$

  Das zeigt, dass $p_0(t)$ differenzierbar ist und es ergibt sich für t>0 folgende Differenzialgleichung

$$p_0'(t) = -\lambda p_0(t).$$

  Durch die Randbedingung $p_0(0) = \mathbb{P}(N_0 = 0) = 1$ ist die eindeutig bestimmte Lösung

$$p_0(t) = e^{-\lambda t}, t \geq 0.$$

  Für beliebige $n \in \mathbb{N}$ gilt

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, t \geq 0.$$

  Dies lässt sich analog zum vorherigen Fall zeigen und durch vollständige Induktion nach n folgt (v).

- $(v) \Rightarrow (i)$:
  Sei $b_0 = 0 \leq a_1 < b_1 \leq ... \leq a_n < b_n$, dann gilt

$$\mathbb{P}\left(\bigcap_{k=1}^{n}\{a_k < S_k \leq b_k\}\right) =$$
$$\mathbb{P}\left(\bigcap_{k=1}^{n-1}\{N_{a_k} - N_{b_{k-1}} = 0, N_{b_k} - N_{a_k} = 1\} \cap \{N_{a_n} - N_{b_{n-1}} = 0, N_{b_n} - N_{a_n} \geq 1\}\right).$$

  Nach (v) besitzt der stochastische Prozess $\{N_t, t \geq 0\}$ stationär Zuwächse und $N_t \sim Poi(\lambda t)$ deshalb gilt

$$\mathbb{P}(N_{a_k} - N_{b_{k-1}} = 0) = \mathbb{P}(N_{a_k - b_{k-1}} = 0) = e^{-\lambda(a_k - b_{k-1})}$$

  und

$$\mathbb{P}(N_{b_k} - N_{a_k} = 1) = \mathbb{P}(N_{b_k - a_k} = 1) = \lambda(b_k - a_k)e^{-\lambda(b_k - a_k)}.$$

  Da die Intervalle $\{(b_{k-1}, a_k)\}_{1 \leq k \leq n}$ und $\{(a_k, b_k)\}_{1 \leq k \leq n}$ disjunkt sind, und der Prozess unabhängige Zuwächse hat, gilt

$$\begin{aligned}
& \mathbb{P}\left(\bigcap_{k=1}^{n}\{a_k < S_k \leq b_k\}\right) \\
= \; & e^{-\lambda(a_n - b_{n-1})}(1 - e^{-\lambda(b_n - a_n)}) \prod_{k=1}^{n-1} e^{-\lambda(a_k - b_{k-1})}\lambda(b_k - a_k)e^{-\lambda(b_k - a_k)} \\
= \; & (e^{-\lambda a_n} - e^{-\lambda b_n})\lambda^{n-1} \prod_{k=1}^{n-1}(b_k - a_k) \\
= \; & \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} \lambda^n e^{-\lambda y_n} dy_n ... dy_1 \\
= \; & \int_{a_1}^{b_1} \int_{a_2 - x_1}^{b_2 - x_1} ... \int_{a_n - x_1 - ... - x_{n-1}}^{b_n - x_1 - ... - x_{n-1}} \lambda^n e^{-\lambda(x_1 + ... + x_n)} dx_n ... dx_1.
\end{aligned}$$

  Dabei wurde im letzten Schritt wieder der Transformationssatz angewendet. Die gemeinsame Dichte von $S_1, S_2 - S_1, ..., S_n - S_{n-1}$ ist somit gegeben durch

$$f_{T_1,..T_n}(x_1, ..., x_n) = f_{S_1, S_2 - S_1, ..., S_n - S_{n-1}}(x_1, ..., x_n) = \lambda^n e^{-\lambda(x_1 + ... + x_n)}.$$

Dies bedeutet, dass die Zufallsvariablen $T_1, T_2 ..., T_n$ unabhängig und exponentialverteilt zum Parameter λ sind, d.h., $\{N_t\}$ ist ein Poisson-Prozess mit der Intensität λ.

□

Die folgende, sehr nützliche Eigenschaft, ist bereits von der Poisson-Verteilung bekannt:

Lemma 3.16. *Die Überlagerung von zwei unabhängigen Poisson-Prozessen* $\{N_t^1, t \geq 0\}$ *und* $\{N_t^2, t \geq 0\}$ *mit der Intensität* λ_1 *bzw.* λ_2 *ist wieder ein Poisson-Prozess mit Intensität* $\lambda = \lambda_1 + \lambda_2$.

Beweis: Im vorangegangen Theorem hab wir gezeigt, dass für einen Poisson-Prozess $\{N_t, t \geq 0\}$ gilt $N_t \sim Poi(\lambda t)$. Da $\{N_t^1, t \geq 0\}$ und $\{N_t^2, t \geq 0\}$ unabhängig sind gilt für die Summe $N_t^1 + N_t^2 \sim Poi((\lambda_1 + \lambda_2)t)$. Damit ist die Überlagerung der beiden Poisson-Prozesse wieder ein Poisson-Prozess mit Intensität $\lambda = \lambda_1 + \lambda_2$. □

Ein interessantes Phänomen ist zu beobachten, wenn man bei einem Poisson-Prozess, zu einem beliebigen Zeitpunkt, die Wartezeit bis zum nächsten Ereignis betrachtet.

Definition 3.17. *Sei* $\{N_t, t \geq 0\}$ *ein Zählprozess mit Zwischenankunftszeiten* $T_1, T_2, ...$ *und Sprungzeiten* $\{S_n, n \in \mathbb{N}\}$. *Weiterhin bezeichne* W_t *die Wartezeiten bis zum nächsten Sprung und* V_t *die Zeit seit dem letzten Sprung. Des Weiteren bezeichne* $T_{N_{t+1}}$ *die* t *enthaltende Zwischenankunftszeit und Analog sind* S_{N_t} *bzw.* $S_{N_{t+1}}$ *die letzte Sprungzeit vor* t *bzw. die nächste nach* t. *Dann gilt:*

$$\begin{aligned} V_t + W_t &= T_{N_{t+1}} = S_{N_{t+1}} - S_{N_t} \\ V_t &= t - S_{N_t} \\ W_t &= S_{N_{t+1}} - t \end{aligned}$$

Lemma 3.18. *Sei* $\{N_t, t \geq 0\}$ *homogener Poisson-Prozess mit Intensität* λ. *Dann ist* $W_t \sim exp(\lambda)$ *für alle* $t > 0$.

Beweis: Da die Zwischenankunftszeiten eines Poisson-Prozesses exponentialverteilt sind, folgt der Beweis direkt aus der Gedächtnislosigkeit der Exponentialverteilung:

$$\begin{aligned} \mathbb{P}(W_t \leq s) &= \mathbb{P}(T_{N_{t+1}} \leq V_t + s) \\ &= \mathbb{P}(T_{N_{t+1}} \leq V_t + s \mid T_{N_{t+1}} > V_t) \\ &= \mathbb{P}(T_{N_{t+1}} \leq s) = 1 - e^{-\lambda s} \end{aligned}$$

Das heißt die Wartezeit zum Zeitpunkt t ist genauso verteilt wie die Zwischenankunftszeiten und damit unabhängig davon, wie viel Zeit bereits seit dem letzten Ereignis vergangen ist. □

Für unser Patienten-Beispiel bedeutet das, dass egal zu welchem Zeitpunkt wir die Beobachtung beginnen, die Wartezeit auf den nächsten Patienten hat immer die gleiche Verteilung. Nachdem wir ausführlich den Poisson-Prozess betrachtet haben, wollen wir nun einen stochastischen Prozess einführen, der häufig in der Risikotheorie für die Modellierung von Schadenshöhen verwendet wird.

Definition 3.19. *Sei $\{N_t, t \geq 0\}$ ein homogener Poisson-Prozess mit Intensität λ und seien $X_1, X_2, ...$ unabhängige, nichtnegative und identisch verteilte Zufallsvariablen. Dann heißt der Prozess*

$$Y_t = \sum_{i=1}^{N_t} X_i$$

zusammengesetzter Poisson-Prozess.

Der Erwartungswert eines zusammengesetzten Poisson-Prozesses lässt sich einfach bestimmen:

Lemma 3.20. *Sei $Y_t = \sum_{i=1}^{N_t} X_i$ ein zusammengesetzter Poisson-Prozess und $\mathbb{E}(X_i) = \mu$, dann gilt für alle $t \geq 0$*

$$\mathbb{E}(Y_t) = \mu\lambda t$$

Beweis: Da die X_i unabhängig sind, gilt nach der Satz der totalen Erwartung

$$\begin{aligned}
\mathbb{E}(Y_t) &= \mathbb{E}\left(\sum_{i=1}^{N_t} X_i\right) \\
&= \sum_{n=1}^{\infty} \mathbb{E}\left(\sum_{i=1}^{N_t} X_i \mid N_t = n\right) \mathbb{P}(N_t = n) \\
&= \sum_{n=1}^{\infty} \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \mathbb{P}(N_t = n) \\
&= \sum_{n=1}^{\infty} \mu n * \mathbb{P}(N_t = n) \\
&= \mu\mathbb{E}(N_t) = \mu\lambda t
\end{aligned}$$

□

3.3. Markov-Kette

Als nächstes werden wir eine weitere wichtige Klasse an stochastischen Prozessen vorstellen. Dieser Abschnitt basiert zum Großteil auf dem Skript von Professor Wolfgang König, „Wahrscheinlichkeitstheorie I und II“ ([7]). Für den gesamten Abschnitt gilt, dass I eine nichtleere, endliche oder höchstens abzählbar unendliche Menge ist.

Definition 3.21. *Ein Stochastischer Prozess $\{X_n, n \in \mathbb{N}\}$ aus I-wertigen Zufallsvariablen besitzt die **Markoveigenschaft**, wenn für alle $n \in \mathbb{N}$ und alle $i_0, i_1, ..., i_{n+1} \in I$ gilt:*

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) \quad (3.6)$$

*Ein diskreter stochastischer Prozess, der (3.6) erfüllt, heißt **Markov-Kette**. Die **Startverteilung** der Markov-Kette ist definiert durch $v(i) := \mathbb{P}(X_0 = i)$ und die Wahrscheinlichkeiten $\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) =: p_{i_n, i_{n+1}}$ werden als **Übergangswahrscheinlichkeiten** bezeichnet. Die Matrix $P = (p_{i,j})_{i,j \in I}$, die sich aus den Übergangswahrscheinlichkeiten ergibt, heißt **Übergangsmatrix**.*

Der nächste Zustand einer Markov-Kette hängt also immer nur von dem aktuellen Zustand ab. Das heißt, die Kette wird durch die Übergangswahrscheinlichkeiten charakterisiert. Deshalb hat die Übergangsmatrix auch eine besondere Struktur.

Definition 3.22. *Eine Matrix $P = (p_{i,j})$ heißt **stochastisch**, falls für alle $i, j \in I$ (Indexmenge) gilt $p_{i,j} \in [0, 1]$ und $\sum_{j \in I} p_{i,j} = 1$.*

Die Übergangsmatrix ist also eine stochastische Matrix, welche für jeden Zustand eine Zeile besitzt, in der die möglichen Übergänge des entsprechenden Zustands in andere Zustände und die dazugehörigen Wahrscheinlichkeiten angegeben werden.

Lemma 3.23. *Sei $\{X_n, n \in \mathbb{N}\}$ eine Folge von I-wertigen Zufallsgrößen, v eine Verteilung auf I und P eine stochastische Matrix, dann ist $\{X_n, n \in \mathbb{N}\}$ genau dann eine Markov-Kette mit Übergangsmatrix P und Startverteilung v, wenn für alle $n \in \mathbb{N}$ und alle $i_0, i_1, ..., i_n \in I$ gilt*

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, ..., X_n = i_n) = v(i_0) p_{i_0, i_1} p_{i_1, i_2} ... p_{i_{n-1}, i_n} \quad (3.7)$$

Beweis: Der Beweis, dass die Gleichung (3.7) für eine Markov-Kette gilt, erfolgt leicht mithilfe von vollständiger Induktion nach n zusammen mit der Definition der Übergangswahrscheinlichkeiten:

Induktionsanfang:

$$\begin{aligned} \mathbb{P}(X_0 = i_0) &= v(i_0) \\ \mathbb{P}(X_0 = i_0, X_1 = i_1) &= v(i_0) * \mathbb{P}(X_1 = i_1 \mid X_0 = i_0) = v(i_0) * p_{i_0, i_1} \end{aligned}$$

Induktionsschritt:

$$\begin{aligned}
& \mathbb{P}(X_0 = i_0, X_1 = i_1, ..., X_{n+1} = i_{n+1}) \\
= & \frac{\mathbb{P}(X_0 = i_0, X_1 = i_1, ..., X_{n+1} = i_{n+1})}{\mathbb{P}(X_0 = i_0, X_1 = i_1, ..., X_n = i_n)} \mathbb{P}(X_0 = i_0, X_1 = i_1, ..., X_n = i_n) \\
= & \mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, ..., X_n = i_n) * v(i_0) p_{i_0,i_1} p_{i_1,i_2} ... p_{i_{n-1},i_n} \\
= & \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) * v(i_0) p_{i_0,i_1} p_{i_1,i_2} ... p_{i_{n-1},i_n} \\
= & v(i_0) p_{i_0,i_1} p_{i_1,i_2} ... p_{i_{n-1},i_n} * p_{i_n,i_{n+1}}
\end{aligned}$$

Die andere Richtung folgt aus der Definition der bedingten Wahrscheinlichkeit:

$$\begin{aligned}
\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, ..., X_n = i_n) &= \frac{\mathbb{P}(X_0 = i_0, ..., X_n = i_n, X_{n+1} = i_{n+1})}{\mathbb{P}(X_0 = i_0, ..., X_n = i_n)} \\
&= \frac{v(i_0) p_{i_0,i_1} p_{i_1,i_2} ... p_{i_{n-1},i_n} p_{i_n,i_{n+1}}}{v(i_0) p_{i_0,i_1} p_{i_1,i_2} ... p_{i_{n-1},i_n}} \\
&= p_{i_n,i_{n+1}} = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n)
\end{aligned}$$

□

Als Nächstes wollen wir mithilfe der Übergangsmatrix die Wahrscheinlichkeit dafür bestimmen, dass sich der Prozess nach n Schritten in einem bestimmten Zustand $j \in I$ befindet.

Lemma 3.24. *Sei $\{X_n, n \in \mathbb{N}\}$ eine Markov-Kette im Zustand i mit Übergangsmatrix P. Dann gilt für alle $n \in \mathbb{N}$ und alle $i, j \in I$*

$$p_{i,j}^{(n)} := \mathbb{P}(X_n = j, X_0 = i) = (P^n)_{i,j}$$

und die $p_{i,j}^{(n)}$ werden als ***n-stufigen Übergangswahrscheinlichkeiten*** *bezeichnet. Das heißt die Wahrscheinlichkeit dafür, dass sich die Markov Kette in n Schritten vom Zustand i in den Zustand j bewegt, entspricht der n-ten Potenz der Übergangsmatrix an der Stelle (i, j).*

Bevor wir dieses Lemma beweisen können, benötigen wir noch folgenden Satz zu den Übergangswahrscheinlichkeiten:

Satz 3.25. ***(Chapman-Kolmogorov-Gleichung)*** *$\{X_n, n \in \mathbb{N}\}$ eine Markov-Kette dann gilt für alle $m, n \in \mathbb{N}$ und alle $i, j \in I$*

$$p_{i,j}^{(m+n)} = \sum_{k \in I} p_{i,k}^{(m)} * p_{k,j}^{(n)}. \tag{3.8}$$

Beweis:

$$\begin{aligned}
p_{i,j}^{(m+n)} &= \mathbb{P}(X_{n+m} = j, X_0 = i) \\
&= \sum_{k\in I} \mathbb{P}(X_{m+n} = j, X_m = k \mid X_0 = i) \\
&= \sum_{k\in I} \frac{\mathbb{P}(X_{m+n} = j, X_m = k, X_0 = i)}{\mathbb{P}(X_0 = i)} \\
&= \sum_{k\in I} \frac{\mathbb{P}(X_{m+n} = j, X_m = k, X_0 = i)}{\mathbb{P}(X_m = k, X_0 = i)} \frac{\mathbb{P}(X_m = k, X_0 = i)}{\mathbb{P}(X_0 = i)} \\
&= \sum_{k\in I} \mathbb{P}(X_{m+n} = j \mid X_m = k, X_0 = i)\mathbb{P}(X_m = k \mid X_0 = i) \\
&= \sum_{k\in I} \mathbb{P}(X_{m+n} = j \mid X_m = k)\mathbb{P}(X_m = k \mid X_0 = i) \\
&= \sum_{k\in I} \mathbb{P}(X_n = j \mid X_0 = k)\mathbb{P}(X_m = k \mid X_0 = i) \\
&= \sum_{k\in I} p_{i,k}^{(m)} * p_{k,j}^{(n)}
\end{aligned}$$

□

Mit Hilfe der Chapman-Kolmogorov-Gleichung können wir beweisen, dass die **n-stufigen Übergangswahrscheinlichkeiten** der n-ten Potenz der Übergangsmatrix an der Stelle (i, j) entsprechen.

Beweis: Lemma 3.24
Sei $P^{(n)} := \{p_{i,j}^{(n)}\}_{i,j\in I}$ die Matrix der n-stufigen Übergangswahrscheinlichkeiten, wobei $P^{(0)} := \mathbb{I}$ die Einheitsmatrix ist. Dann gilt aufgrund von (3.8)

$$P^{(n+m)} = P^{(n)} * P^{(m)}.$$

Für n>0 folgt dann

$$P^{(n)} = P^{(1+n-1)} = P * P^{(n-1)} = ... = P^n.$$

□

Nun wollen wir die verschiedenen Eigenschaften vorstellen, die eine Markov-Kette haben kann, wenn die Übergangsmatrix eine besondere Struktur hat.

Definition 3.26. *Im folgenden sei immer eine Markov-Kette $\{X_n, n \in \mathbb{N}\}$ mit I-wertigen Zufallsgrößen und einer Übergangsmatrix P gegeben. Außerdem seien $i, j \in I$ beliebige Zustände des Zustandsraums.*

(i) Eine Markov-Kette heißt irreduzibel oder ergodisch, wenn für alle $i, j \in I$ ein $n \in \mathbb{N}$ existiert, so dass

$$\mathbb{P}(X_n = j \mid X_0 = i) = p_{i,j}^{(n)} > 0.$$

Das heißt jeder Zustand der Kette kann jeden anderen Zustand mit positiver Wahrscheinlichkeit erreichen.

*(ii) Sei $T_{i,j} := min\{n \in \mathbb{N} : X_n = i \mid X_0 = i\}$ die Wartezeit bis die Markov-Kette vom Zustand i aus, das erste Mal den Zustand j erreicht. Ein Zustand i heißt **rekurrent** falls $\mathbb{P}(T_{i,i} < \infty) = 1$, ansonsten heißt er **transient**. D.h. ein rekurrenter Zustand wird also fast sicher in endlicher Zeit erneut erreicht. Eine Markov-Kette heißt rekurrent wenn alle ihre Zustände rekurrent sind.*

*(iii) Sei $\mu_i := \mathbb{E}(T_{i,i}) = \sum_{n \in \mathbb{N}} n\mathbb{P}(T_{i,i} = n)$ die erwartete Rückkehrzeit zu einem Zustand i bei Start in i. Ein Zustand i heißt **positiv rekurrent**, falls $\mu_i < \infty$ und **nullrekurrent**, wenn i rekurrent ist, aber nicht positiv rekurrent.*

(iv) Ein Zustand i heißt absorbierend, wenn $\mathbb{P}(X_{n+m} = j \mid X_n = i) = 0$ für alle $m \in \mathbb{N}$ und alle $j \in I$. Anlog heißt eine Menge $A \subset I$ absorbierend, wenn $\mathbb{P}(X_{n+m} \notin A \mid X_n \in A) = 0$. Das heißt eine Markov-Kette, die einen absorbierenden Zustand bzw. eine absorbierende Teilmenge von I erreicht, kann diese nicht mehr verlassen.

(v) Die Periode eines Zustands i ist definiert als:

$$d_i = ggT\{n \geq 1 \mid p_{i,i}^{(n)} > 0\}.$$

*Ein Zustand heißt **aperiodisch**, wenn $d_i = 1$ und **periodisch** sonst.*

*(vi) Die Startverteilung v einer Markov-Kette heißt **stationär** oder **Gleichgewichtsverteilung**, wenn für alle $n \in \mathbb{N}$ und alle i gilt*

$$\mathbb{P}(X_n = i) = v(i). \tag{3.9}$$

*Das heißt die Wahrscheinlichkeit hängt zu jedem Zeitpunkt nur von der Startverteilung ab. Anders ausgedrückt gilt $vP = v$, d.h. v ist ein Eigenvektor der Übergangsmatrix zum Eigenwert 1. Ist v ein beliebiges Maß mit der Eigenschaft (3.9), dann bezeichnen wir es als **invariantes Maß**.*

Die Gleichgewichtsverteilung kann nicht immer explizit angegeben werden, aber in einigen Sonderfällen ist dies möglich. Doch vorher benötigen wir noch folgenden Hilfsatz.

Satz 3.27. *Sei $\{X_n, n \in \mathbb{N}\}$ eine irreduzibel und rekurrente Markov-Kette, und sei $k \in I$. Weiterhin sei γ_k ein Maß, das für die, in einem beliebigen Zustand $k \in I$ gestartete Markov-Kette, die erwartete Anzahl an Besuchen in einem Zustand $i \in I$ bis zur ersten Rückkehr der Kette nach k zählt. Also ist*

$$\gamma_k(i) := \mathbb{E}\left(\sum_{n=1}^{T_{k,k}} \mathbb{1}_{\{X_n=i\}}\right).$$

Dann gilt:

(i) γ_k *ist ein invariantes Maß*

(ii) Für alle $i \in I$ *gilt* $0 < \gamma_k(i) < \infty$

(iii) γ_k *ist das einzige invariante Maß mit Wert* 1 *in* k.

Beweis:

- (i): Durch Anwendung des Satzes der monotonen Konvergenz[2] und Ausnutzung der Markov-Eigenschaft, können wir zeigen, dass γ_k ein invariantes Maß ist. Dazu sei $i \in I$ ein beliebiger Zustant, dann gilt

$$\begin{aligned}
\gamma_k(i) &= \mathbb{E}\left(\sum_{n=1}^{T_{k,k}} \mathbb{1}_{\{X_n=i, n\leq T_{k,k}\}}\right) \\
&= \sum_{n\in\mathbb{N}} \mathbb{P}(X_n = i, n \leq T_{k,k}) \\
&= \sum_{n\in\mathbb{N}}\sum_{j\in I} \mathbb{P}(X_n = i, X_{n-1} = j, n \leq T_{k,k}) \\
&= \sum_{n\in\mathbb{N}}\sum_{j\in I} \mathbb{P}(X_{n-1} = j, n \leq T_{k,k})\mathbb{P}(X_n = i | X_{n-1} = j) \\
&= \sum_{n\in\mathbb{N}}\sum_{j\in I} \mathbb{P}(X_{n-1} = j, n-1 \leq T_{k,k}-1)p_{j,i} \\
&= \sum_{j\in I} p_{j,i} \sum_{n\in\mathbb{N}_0} \mathbb{P}(X_n = j, n \leq T_{k,k}-1) \\
&= \sum_{j\in I} p_{j,i}\mathbb{E}\left(\sum_{n=0}^{T_{k,k}-1} \mathbb{1}_{\{X_n=j\}}\right) \\
&= \sum_{j\in I} p_{j,i}\mathbb{E}\left(\sum_{n=1}^{T_{k,k}} \mathbb{1}_{\{X_n=j\}}\right) \\
&= \sum_{j\in I} \gamma_k(j)p_{j,i}.
\end{aligned}$$

[2] siehe Anhang A.2

- (ii): Wir haben gezeigt, dass γ_k ein invariantes Maß ist. Also folgt insbesondere für alle $n \in \mathbb{N}$ und $j \in I$

$$1 = \gamma_k(k) \geq \gamma_k(j) p_{j,k}^{(n)}.$$

Aufgrund der Irreduzibilität existiert für jedes j ein n mit $p_{j,k}^{(n)} > 0$ und deshalb ist $\gamma_k(j) < \infty$ für jedes j. Außerdem ist

$$\gamma_k(j) \geq \gamma_k(k) p_{k,j}^{(n)} = p_{k,j}^{(n)} > 0$$

für ein geeignetes n.

- (iii): Sei λ ein invariantes Maß mit $\lambda(k) = 1$. Aufgrund der Invarianz gilt für jedes $j \in I$:

$$\begin{aligned}
\lambda(j) &= \sum_{i \in I \setminus \{k\}} \lambda(i) p_{i,j} + p_{k,j} \\
&= \sum_{i \in I \setminus \{k\}} \Big(\sum_{i_1 \in I \setminus \{k\}} \lambda(i_1) p_{i_1,i} + p_{k,i} \Big) p_{i,j} + p_{k,j} \\
&= \sum_{i,i_1 \in I \setminus \{k\}} \lambda(i_1) p_{i_1,i} + \sum_{i \in I \setminus \{k\}} p_{k,i} p_{i,j} + p_{k,j} \\
&= \sum_{i,i_1 \in I \setminus \{k\}} \lambda(i_1) p_{i_1,i} + \mathbb{P}(T_{k,k} \geq 2, X_2 = j) + \mathbb{P}(T_{k,k} \geq 1, X_1 = j) \\
&= \dots \\
&= \sum_{i,i_1,\dots,i_n \in I \setminus \{k\}} \lambda(i_n) \Big(\prod_{r=1}^{n} p_{i_r,i_{r-1}} \Big) p_{i,j} + \sum_{r=1}^{n+1} \mathbb{P}(T_{k,k} \geq r, X_r = j) \\
&\geq \sum_{r=1}^{n+1} \mathbb{P}(T_{k,k} \geq r, X_r = j) \\
&= \mathbb{E}\Big(\sum_{r=1}^{min\{T_{k,k}, n+1\}} \mathbb{1}_{\{X_r = j\}} \Big) \\
&\xrightarrow{n \to \infty} \gamma_k(j)
\end{aligned}$$

Das heißt $\lambda(j) \geq \gamma_k(j)$ für alle $j \in I$. Deshalb ist $\lambda - \gamma_k$ auch ein invariantes Maß mit einer Nullstelle in k. Da die Kette aber irreduzibel ist, muss $\lambda - \gamma_k = 0$ sein und dies beweist die Eindeutigkeit von γ_k.

□

Satz 3.28. *Sei $\{X_n, n \in \mathbb{N}\}$ eine irreduzible Markov-Kette mit Übergangsmatrix P, dann sind folgende Aussagen äquivalent:*

(i) Es existiert eine Gleichgewichtsverteilung.

(ii) Es existiert ein positiv rekurrenter Zustand $i \in I$.

(iii) Alle Zustände in I sind positiv rekurrent.

Außerdem gilt, falls eine dieser Bedingungen erfüllt ist, dass die Gleichgewichtsverteilung π eindeutig bestimmt ist und es ist $\pi(i) = \frac{1}{\mu_i}$.

Beweis:

- $(iii) \Rightarrow (ii)$: Diese Richtung ist trivial.
- $(ii) \Rightarrow (i)$: Wir zeigen, dass das im letzten Satz definierte invariante Maß γ auch zu einer Verteilung normiert werden kann und damit dann eine Gleichgewichtsverteilung ist. Dazu sei im folgenden $k \in I$ der positiv rekurrente Zustand, damit gilt

$$\begin{aligned}\sum_{j\in I}\gamma_k(j) &= \sum_{j\in I}\mathbb{E}\Big(\sum_{n=1}^{T_{k,k}}\mathbb{1}_{\{X_n=j\}}\Big)\\ &= \mathbb{E}\Big(\sum_{n=1}^{T_{k,k}}\underbrace{\sum_{j\in I}\mathbb{1}_{\{X_n=j\}}}_{1}\Big)\\ &= \mathbb{E}(T_{k,k}) = \mu_k < \infty.\end{aligned}$$

- $(i) \Rightarrow (iii)$: Sei π eine Gleichgewichtsverteilung und sei $k \in I$, dann ist $\gamma = \frac{\pi}{\pi(k)}$ ein invariantes Maß mit $\gamma(k) = 1$. Wir haben im vorherigen Satz gezeigt, dass es in diesem Fall nur ein invariantes Maß mit Wert 1 in k gibt, das heißt $\gamma = \gamma_k$. Damit folgt

$$\mu_k = \sum_{j\in I}\gamma_k(j) = \sum_{j\in I}\gamma(j) = \frac{1}{\pi(k)}\sum_{j\in I}\pi(j) = \frac{1}{\pi(k)} < \infty$$

  Da dies für alle k gilt, sind alle Zustände positiv rekurrent. Der letzte Teil des Satzes folgt direkt aus dem diesem Beweisschritt.

□

Leider ist es gerade bei komplexeren Markov-Ketten nicht immer möglich, die stationäre Verteilung analytisch zu bestimmen. Unter bestimmten Voraussetzungen kann sie aber zumindest angenähert werden.

Satz 3.29. *Sei $\{X_n, n \in \mathbb{N}\}$ eine irreduzible Markov-Kette mit aperiodischen und positiv rekurrenten Zuständen und Gleichgewichtsverteilung π. Dann gilt für alle $i, j \in I$*

$$\lim_{n\to\infty} p_{i,j}^{(n)} = \pi(j).$$

Das heißt durch wiederholtes Potenzieren der Matrix der Übergangswahrscheinlichkeiten, konvergiert jede Zeile gegen die Gleichgewichtsverteilung.

Auf den Beweis dieser Aussage wird an dieser Stelle verzichtet. Eine vollständige Ausführung des Satzes findet sich in [5, Kapitel 4 -Theorem 4.1.4].

4. Modellierung Isolierte Ereignisse

In diesem Kapitel soll ein Risikomodell für die isolierten Ereignisse in der Phase „regen Lebens“ (im Folgenden alternativ auch als Unfälle bezeichnet) erstellt werden. Die Definition der Isolierten Ereignisse wird deshalb genauer analysiert, um die Anforderungen an das spätere Risikomodell abzuleiten. Es werden zwei Ansätze vorgestellt, die diese Anforderungen erfüllen. Auf Basis von Krankenkassenabrechnungsdaten werden dann die Modellparameter für einen konkreten Anwendungsfall geschätzt und die Modellgüte bewertet.

4.1. Grundlegende Betrachtungen

Im Kohorten-Modell werden die Unfälle als **seltene Ereignisse**, welche die Gesundheit eines Versicherten beeinträchtigen, **sich nicht ankündigen** und bei denen der Eintrittszeitpunkt der eigentliche **Auslöser von begrenzten Leistungsabfolgen** ist, beschrieben. Außerdem sollen sie selten, und **unabhängig voneinander**, auftreten. Das heißt die Eintrittswahrscheinlichkeit für einen Unfall hängt nicht davon ab, ob der Versicherte bereits einen Unfall hatte. Es gibt also insbesondere **keinen Lerneffekt** und **keine Folgeschäden**, die im Modell berücksichtigt werden müssen. Außerdem ist die durch den Unfall hervorgerufene Leistungsfolge begrenzt. Das bedeutet, dass **keine Chronizität** durch einen Unfall entstehen kann. Dadurch sind die isolierten Ereignisse klar zu den singulären Ereignisketten abgegrenzt, die im nächsten Kapitel behandelt werden.

Für das Kohorten-Modell müssen die isolierten Ereignisse in der Phase „regen Lebens“ beschrieben werden. Deshalb werden für die Modellierung zwar alle Unfälle hinzugezogen, aber die Kosten von Unfällen aus den anderen beiden Phasen werden als jeweilige Phasen-Leistungen in der Krankenversicherungsbiographie kumuliert. Das heißt insbesondere, dass Unfälle mit Todesfolge nicht betrachtet werden müssen, da sie bereits in der dritten Phase des Modells erfasst wurden.

Die Erfahrung zeigt, dass die Unfallwahrscheinlichkeit sowohl vom Alter, als auch vom Geschlecht des Versicherten abhängen kann. Dabei ist zu beachten, dass es vom Unfalltyp abhängt, ob ein Zusammenhang besteht und wie stark dieser ist. Des Weiteren können bestimmte Unfälle auch von der Saison abhängen und so zum Beispiel häufiger im Winter auftreten als im Sommer.

Zusammenfassend lässt sich feststellen, dass die zu modellierende Unfallwahrscheinlichkeit unabhängig von bereits eingetretenen Unfällen ist. Jedoch kann sie, je nach Unfalltyp, vom Alter und Geschlecht des Versicherten abhängen und einer Saisonalität unterliegen. Für die Modellierung wird die Saisonalität allerdings nicht weiter betrachten, da aufgrund des in der Regel sehr langen Beobachtungszeitraums, in der Phase „regen Lebens“, diese Effekte nur einen sehr geringen Einfluss auf das Modell haben werden. Auf Basis dieser Erkenntnisse werden nun zwei Modelle vorgestellt:

4.2. Modell - Poisson-Prozess

Die erwartenden Kosten für Unfälle in der Phase regen Lebens, setzen sich aus zwei Teilen zusammen. Zum einen müssen die Kosten für einen Unfall modelliert werden und zum anderen benötigen wir einen Stochastischen Prozess, der die Häufigkeit für einen Unfall in dem beobachteten Zeitraum zählt. Wir werden zuerst den Stochastischen Prozess betrachten. Eine Modellannahme ist, dass die Wahrscheinlichkeit für einen Unfall unabhängig davon ist, wann der vorherige Unfall eingetreten ist. Das heißt der Stochastische Prozess hat unabhängige Zuwächse. Außerdem wird angenommen, dass aufgrund des Charakters der Unfälle als seltene Ereignisse, die Genesungsdauer vernachlässigt werden kann. Das heißt die begrenzte Leistungsfolge, die bei jeden Unfall ausgelöst wird, wird auf den Zeitpunkt des Unfalls konzentriert und die Zeit zwischen zwei Unfällen, hängt nur von der Unfallwahrscheinlichkeit ab. Bevor wir den Prozess näher beschreiben, führen wir folgende Bezeichnungen ein:

(i) Die Unfallwahrscheinlichkeit wird mit λ bezeichnet.

(ii) Die Unfallzeitpunkte seien gegeben durch $t_1, t_2, t_3, ...$ und weiterhin sei $t_0 = 0$ der Beginn der Beobachtung.

(iii) Die Zeit zwischen den einzelnen Unfällen sei definiert als $\{T_k\}_{k\geq 1}$ mit $T_k := t_k - t_{k-1}$.

(iv) Die kumulierten Zwischenankunftszeiten werden mit $S_n = \sum_{k=1}^{n} T_k$ bezeichnet.

(v) Die Anzahl der Unfälle wird bezeichnet durch

$$N_t := max\{k : t_k \leq t\} = \sum_{k=1}^{\infty} \mathbb{1}(S_k \leq t).$$

Es ist leicht zu erkennen, dass $\{N_t, t \geq 0\}$ der gesuchte stochastische Prozess ist und gemäß Definition 3.11 ein Zählprozess ist. Laut Voraussetzung hängt die Unfallwahrscheinlichkeit nur vom Alter und Geschlecht der Versicherten ab. Ohne Einschränkung der Allgemeinheit können wir das Alter eines Versicherten als die Differenz aus dem aktuellen Jahr und seinem Geburtsjahr definieren. Das heißt alle Versicherten werden immer zum 1.1. ein Jahr älter. Da aufgrund der Modellannahmen keine weiteren Einflussfaktoren berücksichtigt werden müssen, ist somit die Unfallwahrscheinlichkeit eines Versicherten

innerhalb eines Kalenderjahres konstant und die Unfallzeitpunkte sind gleichmäßig verteilt. Dann gilt für die Zwischenankunftszeiten $\{T_k\}_{k\geq 1}$, dass die Zeit bis zum nächsten Unfall unabhängig davon ist, wie viel Zeit $s \geq 0$ bereits ohne einen Unfall verstrichen ist

$$(T_k > x + s | T_k > s) = (T_k > x).$$

Das heißt die Verteilung der Zwischenankunftszeiten sind gedächtnislos und somit gilt, gemäß Lemma 3.3, $T_k \sim exp(\lambda)$ für alle $k \geq 1$. Damit ist, innerhalb eines Kalenderjahres, $\{N_t, t \geq 0\}$, laut Definition 3.12, ein homogener Poisson-Prozess und $N_t \sim Poi(\lambda t)$. Da ein Poisson-Prozess, gemäß Theorem 3.15 unabhängige und stationäre Zuwächse besitzt, passt dieser auch zu den eingangs aufgestellten Modellannahmen. An dieser Stelle sollte auch erwähnt werden, dass aufgrund von Lemma 3.18 der Zeitpunkt, zu dem die Beobachtung beginnt, keinen Einfluss auf den Zeitpunkt des nächsten beobachteten Unfalls hat. Deshalb ist auch die Zerlegung des Beobachtungszeitraumes in Jahresscheiben gerechtfertigt und die Summe der Unfälle in den einzelnen Jahresscheiben ergibt dann die Anzahl an Unfälle im Beobachtungszeitraum.

Als Nächstes müssen wir uns überlegen, wie wir die Wahrscheinlichkeit für einen Unfall schätzen können. Da der Poisson-Prozess stationäre Zuwächse hat, können wir die Parameter auf Basis von historischen Informationen schätzen. Aufgrund der erwarteten Abhängigkeit von Alter und Geschlecht, muss für jede Kombination aus Altersgruppe und Geschlecht ein separates $\lambda_{Altersgruppe,Geschlecht}$ geschätzt werden. Dazu betrachten wir die relative Häufigkeit eines Unfalls für die jeweilige Kombination aus Alter und Geschlecht und setzen

$$\bar{\lambda}_{Altersgruppe,Geschlecht} := \frac{\#\{\text{Unfälle}\}_{Altersgruppe,Geschlecht}}{\#\{\text{Versicherte}\}_{Altersgruppe,Geschlecht}}. \tag{4.1}$$

Dabei ist $\#\{\text{Versicherte}\}_{Altersgruppe,Geschlecht}$ die Anzahl aller Versicherten, die in die jeweilige Alters- und Geschlechtsgruppe fallen und in dem betrachteten Zeitraum bei der Kasse versichert waren. Des Weiteren ist $\#\{\text{Unfälle}\}_{Altersgruppe,Geschlecht}$ die Anzahl der Unfälle, im betrachteten Zeitraum, von Personen, die in die jeweilige Alters- und Geschlechtsgruppe fallen. Es werden dabei alle Fälle betrachtet, die in dem jeweiligen Kalenderjahr begonnen haben, da wir Unfälle auf den Eintrittszeitpunkt konzentrieren. Die Einteilung der Altersgruppen ist dabei flexibel.

Bisher haben wir nur die allgemeine Wahrscheinlichkeit für einen Unfall untersucht. Wie bereits am Anfang erwähnt ist es sinnvoll verschiedene Unfalltypen zu betrachten, da sich diese hinsichtlich ihrer Kosten sehr stark unterscheiden können. Das Lemma 3.16 erlaubt es uns, den allgemeinen Unfall Prozess als eine Überlagerung von Teilprozessen darzustellen. Seien $U_1, U_2, ..., U_n$ mit $n \in \mathbb{N}$ die verschiedenen, relevanten Unfalltypen und $\lambda_{U_1}, \lambda_{U_2}, ..., \lambda_{U_n}$ die zugehörigen Unfallwahrscheinlichkeiten, dann gilt für die daraus

resultierenden Prozesse

$$\begin{aligned} N_t^{\lambda_{U_i}} &\sim Poi(\lambda_{U_i} t) \\ N_t^{\lambda} &:= \sum_{i=1}^{n} N_t^{\lambda_{U_i}} \\ N_t^{\lambda} &\sim Poi(\sum_{i=0}^{n} \lambda_{U_i} t) \end{aligned}$$

Das heißt der Prozess lässt sich bei Bedarf beliebig in Unterprozesse zerlegen, die jeweils eine Teilmenge der Unfälle beschreiben. Die Voraussetzung dafür ist, dass die $N_{\lambda_{U_i}}$ unabhängig voneinander sind. Deshalb müssen die einzelnen Unfalltypen so abgrenzt werden, dass es nicht zu Überschneidungen kommt.

Damit ist der erste Teil der Modellierung abgeschlossen und wir können nun die Kosten betrachten. Für jeden Unfalltyp sind diese aufgrund der Modellannahmen unabhängige und gleich verteilte Zufallsgrößen $X_1, X_2, ...$, wobei X_i die Kosten für den i-ten Unfall angibt. Bei Leistungskosten in der Krankenversicherung ist zu beobachten, dass es viele Fälle mit niedrigen Kosten und einige mit sehr hohen Kosten (Hochkostenfällen) gibt, aber nur einen sehr geringen Teil mit durchschnittlichen Kosten. Aus diesem Grund ist es wichtig eine Verteilungsfunktion für die $\{X_i\}_{i>0}$ zu wählen, die zum einen die Standardfälle abdeckt und zum anderen noch genug Masse in den hohen Kostenbereichen hat, um die Hochkostenfälle nicht zu vernachlässigen. Verteilungsfunktionen mit dieser Eigenschaft werden auch als Heavy-Tail-Verteilung[1] bezeichnet. Alternativ könnte als Näherung für die Kostenverteilung auch die empirische Verteilung auf Basis der historischen Informationen verwendet werden. Dabei werden die Kosten für alle Leistungen, die aus dem Unfall resultieren, zusammenaddiert, auch wenn die Leistungszeitpunkte weit auseinander liegen oder sogar in verschiedene Jahre fallen.

Bei der Modellierung der Leistungskosten ist außerdem zu beachten, dass diese einer kontinuierlichen Veränderung, durch Preisveränderungen oder die Einführung alternativer Behandlungsmethoden, unterliegen. Es ist kann versucht werden, dies durch jährliche Änderungsraten auszugleichen, trotzdem sind in der Regel nicht genug Information vorhanden, um alle Einflüsse auf die Kostenentwicklung zu berücksichtigen. Das optimale Vorgehen bei der Modellierung der Kostenverteilung hängt letztendlich vom Unfalltyp ab.

Kombinieren wir jetzt diese beiden Modellbestandteile, dann erhalten wir einen zusammengesetzten Poisson-Prozess, der die zukünftigen Kosten aufgrund von Unfällen beschreibt. Dieser ist für jeden Unfalltyp U definiert durch

$$Y_t^U = \sum_{i=1}^{N_t^U} X_i^U.$$

[1] z.B. Weibull-Verteilung mit Formparameter<1 oder Log-Normalverteilung

Das Risikoäquivalent, für die singulären Ereignisse eines Versicherten zum Zeitpunkt der Berechnung, entspricht somit der Summe der Barwerte[2], der für die Phase regen Lebens vorhergesagten Unfallkosten.

Im weiteren Verlauf dieses Kapitels werden wir diesen Modellansatz beispielhaft für einen Unfalltyp implementieren und auswerten. Allerdings wird vorher noch ein alternativer Modellansatz vorgestellt, mit dem Unfälle beschrieben werden können.

4.3. Alternativer Modellansatz - Markov-Kette

Ein alternativer Modellansatz für die isolierten Ereignisse kann mit Hilfe von Markov-Ketten konstruiert werden. Im Gegensatz zum Poisson-Ansatz können wir mit einer Markov-Kette die Genesungszeiten abbilden. Dadurch kann ausgeschlossen werden, dass ein Patient, der im Krankenhaus behandelt wird, in dieser Zeit einen weiteren Unfall erleidet. Ansonsten gelten weiterhin die unter 4.1 beschriebenen Grundannahmen. Wie bereits in Abschnitt 3.3 gezeigt wurde, ist eine Markov-Kette durch ihre Übergangsmatrix eindeutig definiert. Deshalb werden wir beschreiben, wie die Zustände und Übergangswahrscheinlichkeiten modelliert werden können.

Der Zustandsraum besteht im einfachsten Fall aus zwei Zuständen: $i_0 :=$ Patient ist gesund und $i_1 :=$ Patient erholt sich von einem Unfall. Als Zeitraum für einen Prozessschritt bietet sich dabei ein Tag an, da Informationen zum Unfallzeitpunkt und zur Genesungsdauer in der Regel nur auf den Tag genau vorliegen. Die folgende Grafik veranschaulicht die möglichen Übergänge.

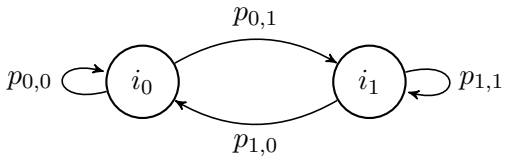


Abbildung 4.1.: Einfaches Markov-Modell

Die Übergangswahrscheinlichkeiten können dabei wie folgt definiert werden: Die Eintrittswahrscheinlichkeit für einen Unfall $p_{0,1}$ ist die, auf den Tag heruntergerechnete, relative Unfallhäufigkeit, die auf Basis der empirischen Daten ermittelt werden kann. Damit gilt für die Wahrscheinlichkeit gesund zu bleiben $p_{0,0} = 1 - p_{0,1}$. Die Genesungswahrscheinlichkeit hängt von der Verweildauer, also der Genesungszeit, ab. Intuitiv könnte wieder die mittlere Genesungsdauer $t_{Genesung}$ als Grundlage genommen und damit definieren werden, dass und $p_{1,1} := 1/t_{Genesung}$ und $p_{1,0} = 1 - p_{1,1}$. Doch dann ist die Verweildauer im Zustand i_1 geometrisch verteilt.

[2]Der Barwert ist der Wert, den eine zukünftige Zahlung in der Gegenwart besitzt.

An dieser Stelle wird die Schwäche dieses einfachen Ansatzes deutlich. In einer Markov-Kette sind die Verweilzeiten immer geometrisch- bzw. im stetigen Fall exponentialverteilt. Aber gerade nach einem schweren Unfall kann es passieren, dass der Patient eine gewisse Zeit im Krankenhaus bleiben muss, bevor die Möglichkeit einer Entlassung besteht. Um auch diese Fälle korrekt abbilden zu können, ist es notwendig neue Zustände einzuführen. Deshalb wird für jeden Genesungstag ein eigener Zustand angelegt, der modelliert, ob die Behandlung abgeschlossen ist, oder einen weiteren Tag andauert. Die daraus resultierenden Übergänge werden in der folgenden Grafik veranschaulicht.

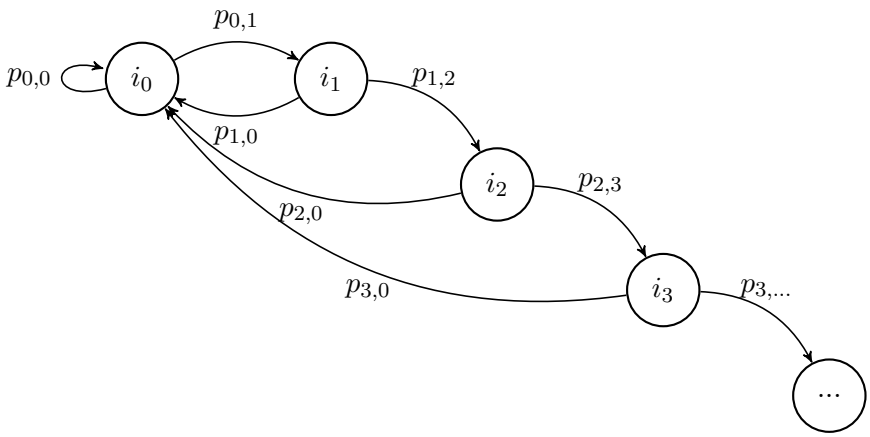


Abbildung 4.2.: Erweitertes Markov-Modell

Dieses Konzept ließe sich sogar noch erweitern, um verschiedene Genesungspfade zu modellieren. Zum Beispiel eine notwendige Reha des Patienten oder eine ambulante Betreuung. Allerdings eignet sich dieses Vorgehen besser dafür, einzelne Krankheiten oder Unfalltypen gezielt zu untersuchen, als für eine gesamte Betrachtung der Unfallkosten.

Bisher haben wir nur einen Unfalltyp betrachtet. Weitere Unfalltypen können als neue Zustände, i_{U_n} für $n \in \mathbb{N}$, in das bestehendes Modell integriert werden. Dabei ist wieder zu beachten, dass die Unfälle so abgegrenzt sind, dass es zu keinen Überschneidungen kommt, da sonst die Übergangsmatrix keine stochastische Matrix mehr ist. Für jeden neu hinzugefügten Unfalltyp verringert sich die Wahrscheinlichkeit gesund zu bleiben $p_{0,0}$ genau um die jeweilige Eintrittswahrscheinlichkeit p_{0,U_n}. Die restlichen Übergangswahrscheinlichkeiten bleiben unverändert. Bei der Menge an möglichen Unfällen, und unter Berücksichtigung der benötigten Zustände um die Genesungszeiten zu modellieren, führt dies zu einem sehr großen Zustandsraum. In der Übergangsmatrix sind allerdings die meisten Einträge 0 und müssen somit nicht betrachtet werden.

Der Vorteil dieser Vorgehensweise ist, dass nachvollziehbar dargestellt wird, was bei einem Unfall passiert. Jedes Mal, wenn die Markov-Kette den Zustand i_0 verlässt, ist ein Unfall passiert und es wird ein kleiner Teilprozess gestartet, der die Genesung simuliert.

Dieser Teilprozess ist abgeschlossen, wenn die Markov-Kette wieder in den Zustand i_0 zurückkehrt. Der Nachteil jedoch ist, dass bevor ein neuer Unfall eintreten kann, die Markov-Kette in den Zustand i_0 zurückkehren muss. Dies führt dazu, dass nach jedem Unfall ein gesunder Tag eintreten muss, bevor ein neuer Unfall passieren kann. Um dies zu vermeiden, müssten für jeden Tag der Genesung Übergangswahrscheinlichkeiten zu den verschiedenen Unfalltypen modelliert werden. Allerdings sind Unfälle, wie eingangs beschrieben, seltene Ereignisse und damit ist der resultierende Fehler, ohne Anpassung, relativ gering.

Die nächste notwendige Erweiterung dieses Modellansatzes ist der Wechsel von konstanten Übergangswahrscheinlichkeiten zu Alters- und Geschlechtsabhängigen. Wie bereits beim Poisson-Modell können die Wahrscheinlichkeiten auf Basis der relativen Häufigkeit modelliert werden. Dabei ist es ausreichend, nur die Unfallwahrscheinlichkeiten $\{p_{0,U_n}\}_{n\in\mathbb{N}}$ als Funktion von Alter und Geschlecht darzustellen und die Genesungszeiten konstant zu lassen.

Die Kosten, die für die Behandlung eines Unfalls entstehen, treten in diesem Modellansatz immer dann auf, wenn die Behandlung eines Patienten abgeschlossen ist. Für die Modellierung der Kostenfunktion gelten dieselben Überlegungen, wie im Poisson-Modell. Allerdings ermöglicht es der Markov-Ansatz, unterschiedliche Kostenverteilungen für einen Unfalltyp zu verwenden und zwar abhängig davon, wie lange die Genesung angedauert hat. Dadurch lässt sich die Streuung der einzelnen Kostenverteilungen reduzieren.

Dieser Modellansatz erzeugt sehr komplexe Modelle, insbesondere durch die alters- und geschlechtsabhängigen Unfallwahrscheinlichkeiten. Es gibt immer mindestens einen positiv rekurrent Zustand. Dies ist i_0 und die erwartete Rückkehrzeit entspricht der mittleren Genesungszeit aller möglichen Unfälle und ist damit endlich ist. Es gibt keine absorbierenden Zustände, da Tod und Pflegefälle ausgeschlossen wurden. Trotzdem ist nicht sichergestellt, dass die Markov-Kette immer irreduzibel ist, da es möglich ist, dass bestimmte Unfälle nur Frauen bzw. Männer betreffen oder in bestimmten Altersbereichen nicht eintreten können. Ein möglicher Ansatz zur Auswertung ist eine Monte-Carlo Simulation. Dies ist ein Verfahren, dass durch die wiederholte Durchführung eines Zufallsexperiments Aussagen über das Verhalten des Systems ableitet. Dieses Vorgehen basiert auf dem Gesetz der großen Zahlen. Das heißt die relativen Häufigkeiten nähern sich im Grenzwert den Wahrscheinlichkeiten der zugrunde liegenden Verteilung an. Alternativ könnten auch hinreichend kleine Zeitintervalle betrachtet werden, in denen die Übergangswahrscheinlichkeit konstant ist. Die erreichbaren Zustände und die entsprechenden Übergangswahrscheinlichkeiten bilden dann eine irreduzible Markov-Kette, für die nach Satz 3.28 eine Gleichgewichtsverteilung existiert, die ausgewertet werden kann.

Zusammenfassend lässt sich sagen, dass auch mit Hilfe von Markov-Ketten das Phänomen der isolierten Ereignisse beschrieben werden kann. Diese Herangehensweise erlaubt dabei eine detaillierte Modellierung des Genesungsprozesses. Sie ist dabei hinsichtlich der modellierten Unfalltypen ähnlich leicht skalierbar, wie der Poisson-Ansatz. Der Nachteil ist

die Komplexität der Modelle und die große Anzahl an Parametern, die geschätzt werden müssen. Dadurch ist eine effiziente Auswertung in der Regel nur mit Hilfe eines Simulationsansatzes möglich. Außerdem kann es, je nach Datengrundlage, schwierig werden alle benötigten Werte stabil zu schätzen. Der wichtigste Kritikpunkt ist aber, dass dieser Ansatz durch die Genesungszeiten nur schwer mit der Phase regen Lebens kombiniert werden kann, die im Kohorten-Modell von den Unfällen überlagert werden soll. Im weiteren Verlauf dieses Kapitels werden wir uns deshalb auf den Poisson-Ansatz konzentrieren.

4.4. Anwendung und Test des Modells

Im folgenden Abschnitt wird zunächst die Datengrundlage vorgestellt, auf der das Modell später getestet werden soll. Anschließend wird die Vorgehensweise zur Identifizierung der Unfälle beschrieben. Im letzten Teil werden dann die Parameter des Modells geschätzt und die Modellgüte bewertet. Unfälle können entweder im ambulanten Bereich abgerechnet werden, oder im stationären. In beiden Fällen ist es notwendig, das zugrunde liegende Abrechnungssystem nachzuvollziehen und dann alle relevanten Daten zu einem Fall zusammenzufassen. Für diese Auswertung betrachten wir ausschließlich die stationären Fälle, also solche, die zu einer Aufnahme im Krankenhaus geführt haben.

4.4.1. Erläuterung der Datengrundlage

Die Datengrundlage für die Anwendung des Modells besteht aus Versicherten- und Krankenhausabrechnungsdaten, die über einen Zeitraum von drei Jahren erfasst wurden. Für die Aufbereitung der Daten wurden Duplikate entfernt, die Datenformate vereinheitlicht und fehlerhafte bzw. unvollständige Datensätze herausgefiltert. Die wichtigsten Informationen in den Krankenhausdaten sind die abgerechneten Krankenhausbehandlungen und die gestellten Diagnosen. Deshalb werden wir, bevor wir die Datengrundlage näher betrachten, im Folgenden einen kurzen Einblick in das deutsche Abrechnungssystem im Krankenhausbereich geben.

Seit 2003 wird in Deutschland die Vergütung im Krankenhausbereich über ein Fallpauschalensystem realisiert. Dieses basiert auf den sogenannten Diagnosis Related Groups (kurz: DRG)[3]. Behandlungsfälle, die medizinisch und hinsichtlich des Ressourcenverbrauchs ähnlich sind, sind dort zu Fallgruppen zusammengefasst. Die Zuordnung erfolgt auf Basis der Patienten- und Falldaten, die während eines Krankenhausaufenthaltes gesammelt werden. Dazu gehören, neben dem Alter und Geschlecht des Patienten, auch die gestellten Diagnosen und die vorgenommen Prozeduren. Komplikationen, oder erschwerende Begleiterkrankungen (Komorbiditäten), werden über verschiedene Schweregrade berücksichtigt. Auf diese Weise wird jeder DRG ein relatives Kostengewicht zugeordnet. Außerdem gibt es für jede DRG abhängig von der mittleren Verweildauer, eine untere

[3] auf Deutsch: diagnosebezogene Fallgruppen

und eine obere Grenzverweildauer. Liegt die tatsächliche Verweildauer eines Falles außerhalb dieses Bereichs, gibt es Zu- bzw. Abschläge auf das Kostengewicht.

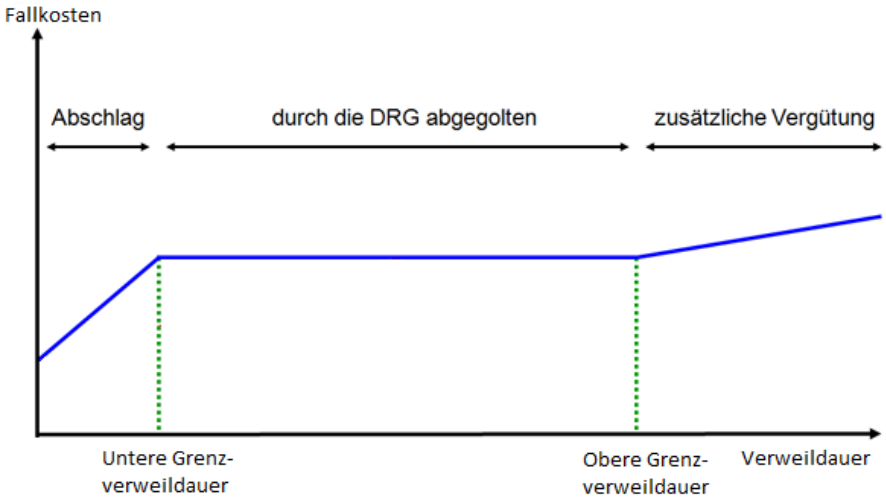


Abbildung 4.3.: DRG-Fallkostenentwicklung (eigene Darstellung)

Am Ende wird das resultierende Kostengewicht mit dem Basisfallwert multipliziert, um die Kosten des Krankenhausfalles zu bestimmen. Der Basisfallwert wird dabei jährlich für jedes Bundesland neu berechnet. Aktuell liegt dieser Wert zwischen 3.190,81 Euro und 3.311,98 Euro[4]. Das System wird vom „Institut für das Entgeltsystem im Krankenhaus“([13]), kurz InEK, gepflegt und weiterentwickelt.

Die während eines Krankenhausaufenthalts gestellten Diagnosen, und die durchgeführten Prozeduren, werden nach den medizinischen Klassifikationen ICD-10-GM und OPS kodiert. Das ICD-10-GM ist eine für Deutschland angepasste Version der „internationalen statistischen Klassifikation der Krankheiten und verwandter Gesundheitsprobleme“(kurz: ICD[5]). Der dazugehörige Katalog wird vom „deutschen Institut für Medizinische Dokumentation und Information“([14]),kurz DIMDI, jährlich aktualisiert und angepasst. Er ist hierarchisch strukturiert und enthält 22 Krankheitskapitel, die sich in Gruppen, Kategorien und Subkategorien aufsplitten. Das DIMDI schätzt die Anzahl der Schlüsselnummern auf ca. 13.400. Die jährlichen Anpassungen stellen auch eine besondere Herausforderung bei der Arbeit mit ICD-Schlüsseln dar. Es kann vorkommen, dass in zwei unterschiedlichen Jahren einem ICD-Schlüssel unterschiedliche Diagnosen zugeordnet werden. Der

[4]Quelle: `http://www.gkv-spitzenverband.de/krankenversicherung/krankenhaeuser/budgetverhandlungen/bundesbasisfallwert/bundesbasisfallwert.jsp` (Stand 8.2.2015)

[5]Aufgrund des englischen Namens: „International Statistical Classification of Diseases and Related Health Problems“

Operationen- und Prozedurenschlüssel (kurz: OPS) wird ebenfalls vom DIMDI gepflegt und ist die amtliche Klassifikation zum Verschlüsseln von Operationen, Prozeduren und allgemein medizinischen Maßnahmen im stationären Bereich und beim ambulanten Operieren.

Sowohl die DRG- als auch ICD-Informationen sind, in den Grunddaten, für jeden einzelnen Krankenhausfall aufgeschlüsselt. Es ist zu beachten, dass alle, während eines Krankenhausfalles gestellten, Diagnosen erfasst werden. Das heißt es können auch Diagnosen vorkommen, die mit der DRG nicht in Zusammenhang stehen[6]. Die Datenstruktur der aufbereiteten Daten wird in folgendem Diagramm veranschaulicht:

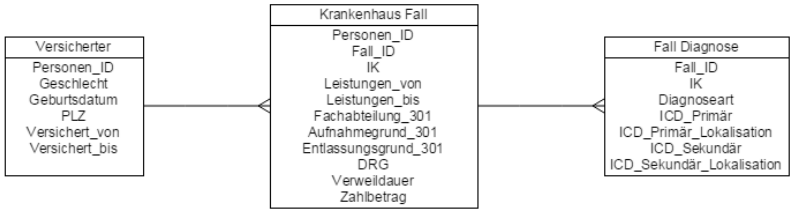


Abbildung 4.4.: ER-Diagramm Datengrundlage

(i) **Versicherter:** In der ersten Tabelle sind Informationen zu den Versicherten hinterlegt. Neben Geschlecht und Geburtsdatum gibt es auch Information über den Wohnort (Postleitzahl) und den Zeitraum, in dem er versichert war. Letzteres ist wichtig, da nur während dieses Zeitraums Leistungsdaten über den entsprechenden Versicherten vorliegen. Als Primärschlüssel wird in dieser Tabelle die Personen_ID verwendet, wodurch dann auch eine Zuordnung zu den Krankenhaus-Falldaten möglich ist.

(ii) **Krankenhaus_Fall:** Ein Krankenhausfall umfasst alle Maßnahmen, die von der Einweisung bis zu Entlassung eines Versicherten fällig werden. In dieser Tabelle kommt ein zusammengesetzter Primärschlüssel aus Fall_ID und Institutionskennzeichen des Krankenhauses (IK) zum Einsatz [7]. Für jeden Behandlungsfall gibt es Daten zum Zeitpunkt und zum Grund der Aufnahme beziehungsweise für die Entlassung. Außerdem werden die abgerechnete DRG und die damit verbundenen Kosten angegeben.

[6] z.B. ein Herzinfarkt während einer Operation

[7] Das ist notwendig da jedes Krankenhaus die Fall_ID selbst vergibt und es so vorkommt das zwei Krankenhäuser dieselbe ID vergeben.

(iii) **Fall_Diagnose:** Zu jedem Fall sind außerdem die dazugehörigen Diagnosen hinterlegt. Diese können wieder über die Kombination aus Institutionskennzeichen und Fall_ID zugeordnet werden. Die Diagnosen werden nach dem oben beschrieben ICD-Katalog codiert und in Haupt- und Nebendiagnosen unterteilt, durch das Feld Diagnoseart. Jeder Datensatz enthält eine primäre ICD zu der, falls notwendig, auch die Lokalisation angegeben wird. Dies dient der Spezifizierung, zum Beispiel, ob der linke oder der rechte Arm gebrochen ist. In einigen wenigen Fällen gibt es auch noch eine sekundäre ICD.

4.4.2. Identifizierung von Unfällen

Eine Herausforderung dieser Arbeit war es aus den sehr umfangreichen Abrechnungsdaten die Fälle zu identifizieren, die als Unfälle im Sinne des Kohorten-Modells interpretiert werden können. In diesem Zusammenhang sollte nochmal erwähnt werden, dass alle Fälle, die nicht über die in diesem Abschnitt beschrieben Methodik erfasst werden können, trotzdem bilanziell im Kohorten Modell abgebildet sind. Diese werden dann der Phase regen Lebens zugeordnet, wodurch keine Kosten verloren gehen.

Das wichtigste Erkennungsmerkmal von Unfällen ist, dass sie sich nicht ankündigen. Deshalb ist der Aufnahmegrund im Krankenhaus ein guter Indikator dafür, ob der Patient per Überweisung oder mit Termin aufgenommen wurde. In den Daten ist der Aufnahmegrund entsprechend den gesetzlichen Vorgaben kodiert[8]. Die für uns relevanten Informationen sind in der 3. und 4. Stelle kodiert. Die folgende Tabelle liefert eine Übersicht der möglichen Ausprägungen und der relativen Häufigkeiten in den Daten:

Code	Aufnahmegrund	Anteil
01	Normalfall	56,704%
07	Notfall	38,566%
03	Verkehrsunfall / Sportunfall / Sonstiger Unfall	0,105%
02	Arbeitsunfall / Wegeunfall / Berufskrankheit	0,021%
06	Kriegsbeschädigten-Leiden / BVG-Leiden	0,003%
04	Hinweis auf Einwirkung von äußerer Gewalt	0,002%
05	frei (früher Hinweis auf Selbstmord / Selbstbeschädigung)	0,001%
	Keine Angabe (Feld nicht gefüllt)	4,598%

Tabelle 4.1.: Aufnahmegrund und relative Häufigkeiten

Mehr als die Hälfte der Krankenhausfälle werden als Normalfall klassifiziert und kommen somit über eine Überweisung oder mit Termin ins Krankenhaus. Ein sehr kleiner Teil wird aufgrund einer Kriegsbeschädigung(06) oder aufgrund eines Gewaltverbrechens(04) behandelt. Für uns sind die Notfälle und Fälle interessant, die tatsächlich als Unfall aufgenommen wurden (02 und 03). Letztere machen nur einen sehr kleinen Teil aus, was auf

[8] siehe Anhang A.3

die Natur der Daten zurückzuführen ist. Diese Schlüssel werden nur vergeben, wenn diese Angabe versicherungsrelevant ist. Damit konnten wir den Anteil der möglich Unfälle bereits auf ca. 40% der Krankenhausfälle einschränken.

Den nächsten Anhaltspunkt liefert der Entlassungs- bzw. Verlegungsgrund. Allerdings gibt es 25 verschiedene Ausprägungen und zusätzlich die Information, ob der Versicherte arbeitsfähig entlassen wurde[9]. Deshalb erläutern wir lediglich die interessanten Fälle, welche in der nachfolgenden Tabelle zusammengefasst sind. Die Anteile beziehen sich dabei bereits nur auf Notfälle und Unfälle:

Entlassungsgrund	Anteil
Behandlung beendet	83,210%
Tod oder Entlassung in ein Hospiz	5,796%
Entlassung in eine Pflegeeinrichtung	3,199%
Entlassung in eine REHA-Einrichtung	2,219%
Sonstige	5,575%

Tabelle 4.2.: Entlassungsgrund und relative Häufigkeiten für Not- und Unfälle

Für den Großteil der Not- und Unfälle endet mit dem Krankenhausaufenthalt auch die Behandlung. Ein Teil der Patienten stirbt oder muss in ein Hospiz verlegt werden. Diese Fälle werden in der Prämortalitätsphase des Kohorten-Modells erfasst und sind somit für die singulären Ereignisse nicht relevant. Fälle mit anschließender Pflege sind für die singulären Ereignisketten von Relevanz. Diese werden im nächsten Kapitel behandelt. Unter „Sonstige“ sind insbesondere die nicht trivialen Fälle zusammengefasst. Das bedeutet in erster Linie eine Verlegung oder eine Entlassungen aus abrechnungstechnischen Gründen. In diesen Fällen muss der weitere Verlauf aufwändig rekonstruiert werden, was nicht immer gelingt. Insbesondere bei einen Kassenwechsel des Patienten stehen die Information i.d.R. nicht zur Verfügung. Hier muss dann abhängig vom jeweiligen Unfalltyp entschieden werden, ob sich der Aufwand lohnt. Gelingt die Rekonstruktion der Behandlungspfade, müssten die Kosten aller relevanten Anschlussbehandlung zu den Unfallkosten hinzugezählt werden. Gelingt dies nicht, dürfen die Fälle trotzdem nicht vernachlässigt werden. Da allerdings keine Aussage über die Kosten getroffen werden kann, findet der Fall nur bei der Modellierung der Unfallwahrscheinlichkeit Beachtung, nicht aber bei der Modellierung der Kostenverteilung.

Nachdem die möglichen Unfälle, aufgrund des Aufnahme und Entlassungsgrundes, eingegrenzt wurden, sind in den Daten nur noch die DRG und ICD Informationen vorhanden. Zu jedem Fall gibt es eine DRG und mehrere Diagnosen, wobei eine davon als Hauptdiagnose ausgezeichnet wird. Das INeK definiert die Hauptdiagnose als „Die Diagnose, die nach Analyse als diejenige festgestellt wurde, die hauptsächlich für die Veranlassung

[9] siehe Anhang A.3

des stationären Krankenhausaufenthaltes des Patienten verantwortlich ist.“[10]. Deshalb kann die Hauptdiagnose dazu verwendet werden, um verschiedene Unfälle voneinander abzugrenzen. Es kommt aber vor, dass während der Behandlung andere Erkrankung entdeckt werden und dadurch die Hauptdiagnose am Ende für die DRG, und damit für die Kosten, kaum eine Rolle spielt. An dieser Stelle sollte erwähnt werden, dass ein Krankenhaus ein Unternehmen ist, welches darauf ausgerichtet ist Gewinn zu erwirtschaften. Deshalb besteht die Gefahr, dass Informationen verändert werden, um höhere Fallpauschalen abzurechnen oder um Patienten aufnehmen zu können, die eigentlich an andere Einrichtungen verwiesen werden müssten.

Aufgrund der Komplexität der möglichen Krankenhausfälle ist es schwierig eine allgemeine Vorgehensweise zu beschreiben, um die Unfälle exakt abzugrenzen. Die vergebenen Diagnosen liefern einen guten Anhaltspunkt dafür, was mit dem Patienten passiert und warum er im Krankenhaus ist. Insbesondere sind dabei die Kapitel S00-T98: „Verletzungen, Vergiftungen und bestimmte andere Folgen äußerer Ursachen“ hervorzuheben, da diese Diagnosen gut zu unserer Definition von Unfällen passen. Eine Abgrenzung nur auf Basis der Diagnosen ist allerdings schwierig, da viele Fälle mehrere Diagnosen und wir sicherstellen müssen, dass es bei der Definition der Unfalltypen nicht zu Überschneidungen kommt. Außerdem hängen die gestellten Diagnosen nicht immer mit der gewählten Behandlung zusammen und somit sind auch keine Rückschlüsse auf die Kosten möglich. Die DRG drückt hingegen nur aus, welche Behandlung vom Krankenhaus abgerechnet wurde. Ein DRG deckt allerdings einen großen Bereich an Fällen ab und dabei auch Fälle, die wir nicht als Unfälle definieren würden.

Zur weiteren Abgrenzung von Unfällen ist also eine Kombination aus DRG und Diagnose nötig. Diese lassen sich teilweise nur mit medizinischen Hintergrundwissen identifizieren.

Für die Auswahl des Testfalls im nächsten Abschnitt wurden die häufigsten Hauptdiagnosen betrachtet. Das sind Hypertonie, Herzinsuffizienz, Ohnmacht oder eine Gehirnerschütterung. Hypertonie oder eine Herzinsuffizienz sind im Sinne des Kohorten-Modells keine Unfälle, da es häufig bereits im Vorfeld Anzeichen für eine solche Erkrankung gibt. Eine Ohnmacht kann dagegen sehr vielfältige Ursachen haben, wovon einige mit unserer Unfalldefinition konform sind und andere nicht. Die Gehirnerschütterung ist dahingegen eine nachvollziehbare Unfalldiagnose und wurde deshalb für unser Anwendungsbeispiel ausgewählt.

4.4.3. Anwendung und Test des Poisson-Modells

In diesem Abschnitt werden wir, am Beispiel vom Unfall Gehirnerschütterung, den vorgestellten Poisson-Ansatz implementieren und bewerten. Dazu ist es notwendig eine saubere Datenbasis auszuwählen, auf der die notwendigen Parameter geschätzt werden können. Außerdem werden wird eine Testmenge benötigt, die später zur Bewertung des fertigen

[10] vergleiche [13] - Kodierrichtlinien Seite 4 `http://www.g-drg.de/cms/content/view/full/5064`

Modells verwendet werden kann. Es existieren Daten von drei Jahren (2009, 2010 und 2011). Deshalb werden wir das erste Jahr (2009) für die Schätzung der Modellparameter verwenden und die anderen beiden Jahre für den Test. Jedem Jahr ordnen wir alle Unfälle zu, die in dem jeweiligen Jahr begonnen haben. Alle Kosten eine Falles werden auf den Tag der Krankenhauseinweisung konzentriert, auch wenn sie erst in den nachfolgenden Jahren abgerechnet wurden.

Im vorangegangenen Abschnitt haben wir bereits erläutert, wie die Abrechnungsdaten nach Aufnahme und Entlassungsgrund gefiltert werden können, um die relevanten Fälle zu identifizieren. Dieselben Filter wurden auch verwendet, um die Grundtabelle mit allen potentiellen Unfällen zu erstellen[11]. Dazu wurde der MS SQL Server als Datenbank verwendet und die benötigten Abfragen per SQL formuliert. Anschließend wurden alle Fälle betrachtet, die als Hauptdiagnose die Diagnose S06.0 "Gehirnerschütterung" hatten, um zu untersuchen, welche DRGs in diesen Fällen abgerechnet wurden. Die häufigste DRG (Anteil $97,5\%$) ist die B80Z: „Andere Kopfverletzung", also die Behandlung wegen einer Kopfverletzung. Es gab auch Fälle die mit der Diagnose Gehirnerschütterung eingeliefert wurden und bei denen dann ein Herzschrittmacher eingesetzt wurde, was zu einer ganz anderen DRG geführt hat. Daran zeigt sich, weshalb die Kombination aus Diagnose und DRG wichtig für die Abgrenzung ist. Für unseren Test beschränken wir uns deshalb auf alle Fälle, die wegen einer Gehirnerschütterung aufgenommen wurden und auch wegen einer Kopfverletzung behandelt wurden.

Neben den Fallinformationen benötigen wir außerdem die Personeninformationen. In der entsprechenden Grundtabelle sind für jeden Versicherten mehrere Einträge vorhanden. Jeder dieser Einträge hat dabei dieselben Grundinformationen, wie Alter und Geschlecht, aber unterschiedliche Angaben zu dem Versicherungszeitraum. Um herauszufinden, ob ein Versicherter in einem bestimmten Zeitraum versichert war, müssen die einzelnen Zeiträume zusammengeführt werden. Da dies mit SQL sehr aufwendig ist, wurde ein JAVA Programm geschrieben, das diese Aufgabe erfüllt. Als Eingabe dienen die Personeninformationen und das Jahr, welches ausgewertet werden soll. Anschließend prüft das Programm für jeden Versicherten, welche Zeitintervalle in das beobachtete Jahr fallen und gibt die Anzahl der Tage aus, die der Versicherte in diesem Jahr bei der Kasse war. Zusätzlich wird noch das Alter[12] des Versicherten im jeweiligen Jahr bestimmt. Diese Informationen werden dann im einem letzten Schritt an die Unfalldaten angespielt und ausgegeben. Die so erzeugte Grundtabelle bietet die Basis für das weitere Vorgehen.

Im ersten Schritt betrachten wir die Unfallwahrscheinlichkeit λ für eine Gehirnerschütterung. Da ein Poisson-Prozess stationäre Zuwächse besitzt, können wir diese auf Basis der relativen Häufigkeit in den historischen Daten schätzen.

$$\bar{\lambda} = \frac{\#\text{Unfälle im Beobachtungszeitraum}}{\#\text{Versicherte im Beobachtungszeitraum}}.$$

[11] siehe Anhang A.4

[12] Differenz aus betrachtetem Jahr und Geburtsjahr

Dabei ist #Unfälle im Beobachtungszeitraum in unserem Beispiel die Anzahl alle Gehirnerschütterung, die in 2009 begonnen haben. Dieser Wert wird durch die Anzahl aller Personen geteilt, die in 2009 versichert waren. Das führt zu der Frage wie Versicherte behandelt werden, die nicht über den gesamten betrachteten Zeitraum hinweg versichert waren. Diese dürfen nicht ignoriert werden, da sonst evtl. relevante Informationen verloren gehen und das Modell verzerrt wird. Deshalb werden wir eine Gewichtung auf Basis der versicherten Tage vornehmen. Dementsprechend wird jemand, der nur 6 Monate versichert war, nur mit einem Gewicht von 0,5 in der Zählung berücksichtigt. Die Anzahl der Unfälle wird dabei nicht gewichtet. Für unsere Testdaten ergibt sich in 2009 eine geschätzte Unfallwahrscheinlichkeit von $\bar{\lambda} = 0,189\%$ für das seltene Ereignis „Gehirnerschütterung".

Als Nächstes untersuchen wir, ob dieser Unfalltyp vom Alter und Geschlecht der Versicherten abhängt. Dazu unterteilen wir den Versichertenbestand in verschiedene Altersgruppen, die jeweils 5 Jahre umfassen. Die Neugeborenen sind dabei in einer extra Gruppe zusammengefasst, da diese in der ersten Phase des Kohorten-Modells betrachtet werden. Außerdem bilden Versicherte, die älter sind als 90, eine Gruppe, da es in den hohen Altersbereichen nicht genug Versicherte gibt um allgemeine Aussagen ableiten zu können. Anschließend werden diese Gruppe noch nach Geschlecht getrennt. Wir bezeichnen diese Gruppen im weiteren kurz als AGG für Alters- und Geschlechtsgruppe. Die folgende Grafik zeigt die Aufteilung der Grundgesamtheit in diese Gruppen.

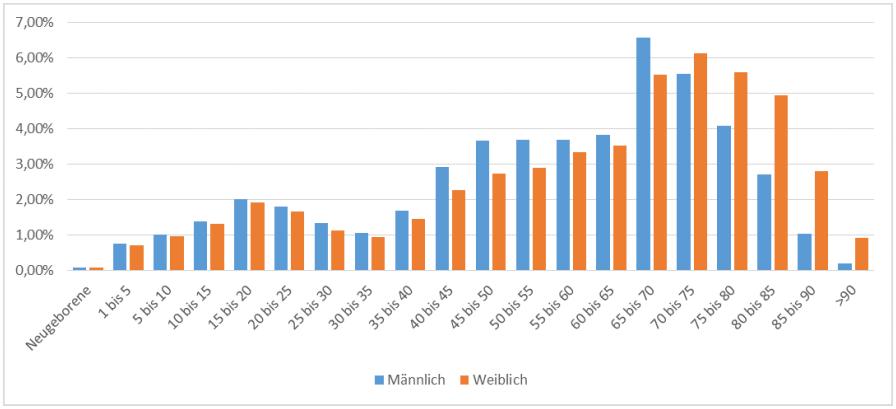


Abbildung 4.5.: Anteile in den Alters- und Geschlechtsgruppen

Nun können wir für jede AGG die jeweiligen $\bar{\lambda}_{AGG}$ schätzen. Die folgende Grafik zeigt die relativen Häufigkeiten für eine Gehirnerschütterung in 2009, auf Basis der vorliegenden Daten.

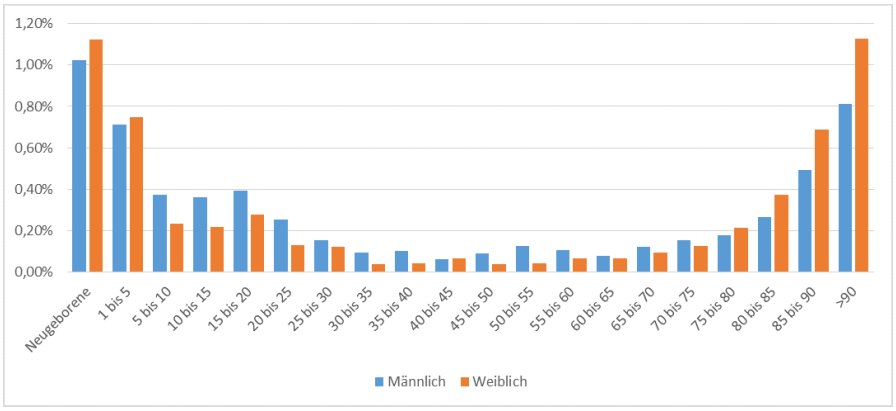


Abbildung 4.6.: Geschätzte Unfallwahrscheinlichkeiten

Diese typische Badewannenkurve ist häufig im Zusammenhang mit Leistungen im Gesundheitswesen zu beobachten. Es ist klar zu sehen, dass im Vergleich zur durchschnittlichen Unfallwahrscheinlichkeit $\bar{\lambda}$, junge und alte Menschen zwei bis viermal so häufig wegen einer Gehirnerschütterung behandelt wurden. Dahingegen liegt die Häufigkeit im Altersbereich $30 - 65$, deutlich unter dem Durchschnitt.

Durch die sehr geringe Fallzahl bei Unfällen entsteht beim Schätzen der Unfallwahrscheinlichkeiten ein hoher Standardfehler. Je kleiner die Altersgruppen werden, umso größer wird dieser. Da wir annehmen, dass die Anzahl der Unfälle für einen Versicherten Poisson-Verteilt ist, ergibt sich der Standardfehler für eine AGG durch

$$\sigma_{\bar{\lambda}_{AGG}} = \sqrt{\frac{\bar{\lambda}_{AGG}}{\#\text{Versicherte}_{AGG}}}.$$

Insbesondere in den schwach besetzten Gruppen ist deshalb die Genauigkeit der Schätzfunktion limitiert, da der relative Schätzfehler groß ist. Aus diesem Grund werden wir im weiteren nur noch drei Altersgruppen betrachten. Entsprechend den vorangegangen Beobachtungen unterteilen wir die Versicherten in junge Leute, Erwachsene und Rentner. Diese Gruppen bilden die Grundlage für die Bewertung. Damit ergeben sich folgende Werte für die geschätzten Unfallwahrscheinlichkeiten und die entsprechenden Standardfehler.

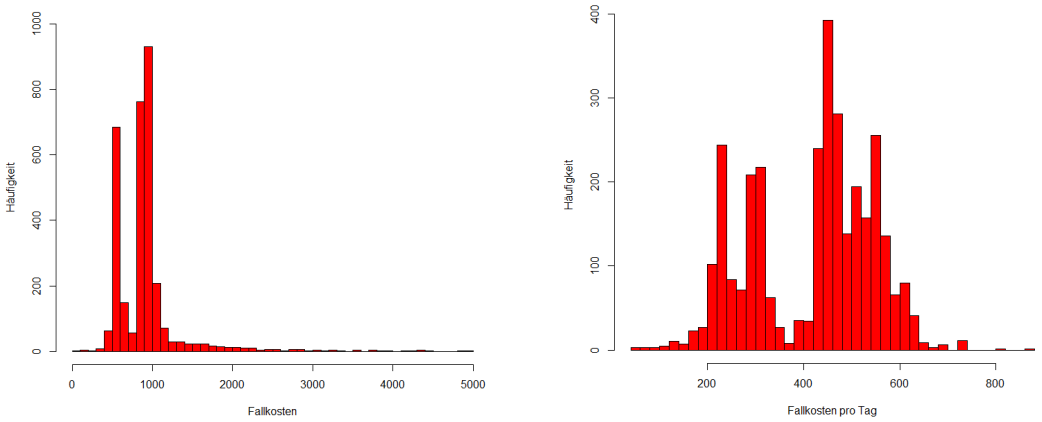
Gruppe	$\bar{\lambda}$ Männer	$\bar{\lambda}$ Frauen	σ Männer	σ Frauen
Junge Menschen(1-25 Jahre)	0,3829%	0,2720%	0,0180%	0,0156%
Erwachsene(26-65 Jahre)	0,0973%	0,0585%	0,0051%	0,0043%
Rentner(>65 Jahre)	0,1867%	0,2820%	0,0074%	0,0080%

Tabelle 4.3.: Übersicht Unfallwahrscheinlichkeiten Gehirnerschütterung

Diese Unfallwahrscheinlichkeiten werden später dazu verwendet, um die erwarteten Kosten eines Versicherten zu schätzen. Die Unterschiede zwischen den einzelnen Gruppen

sind dabei deutlich zu erkennen.

Als Nächstes wollen wir die Kosten in den historischen Daten untersuchen, um für das Modell eine passende Kostenverteilung zu ermitteln. Zur Auswertung wurde in diesem Fall R[13] gewählt. Die folgende Grafiken zeigen zum einen die Verteilung der Gesamtkosten und zum anderen die Kosten pro Tag in 2009.



Gesamtkosten Kosten pro Tag

Abbildung 4.7.: Übersicht Leistungskosten Gehirnerschütterung

In beiden Grafik ist deutlich zu erkennen, dass die Kostenverteilung zwei Häufungspunkte besitzt. Das deutet auf eine Überlagerung von mindestens zwei Verteilungen hin. Die Ursache dafür ist das Abrechnungssystem im Krankenhaus. Wie bereits beschrieben, entsprechen die Kosten für einen Fall dem Produkt aus Kostengewicht und Basisfallwert. Der bundeslandspezifische Basisfallwert bleibt innerhalb eines Jahres konstant. Das Kostengewicht hängt jedoch in erster Linie von der Verweildauer des Patienten im Krankenhaus ab. Der daraus resultierende Zusammenhang zwischen Kosten und Verweildauer lässt sich für die DRG B80Z[14] wie folgt darstellen.

[13] `http://www.r-project.org/`

[14] Quelle: `http://drg.uni-muenster.de/index.php?option=com_drgsystematik&view=DRGSystematik&Itemid=49&mdc=01` (Stand: 8.2.2015)

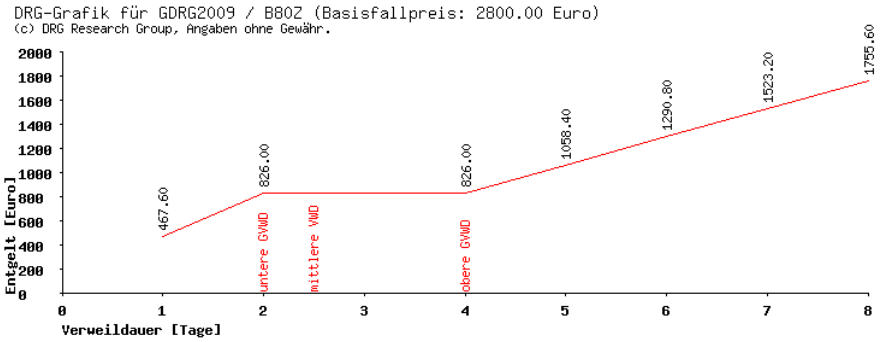


Abbildung 4.8.: DRG B80Z Kostenentwicklung

Das heißt die untere Grenzverweildauer beträgt zwei Tage und die obere Grenzverweildauer vier Tage. In diesem Zeitraum ist das Kostengewicht konstant und außerhalb davon gibt es Zu- und Abschläge. Die Verteilung der Verweildauer ist dabei immer diskret, da das Abrechnungssystem nur ganze Tage berücksichtigt. Betrachten wir die Verweildauer für unsere Fälle wird deutlich, dass die Mehrheit nur ein bis drei Tage im Krankenhaus aufgenommen wurde. Die mittlere Verweildauer beträgt 1,8 Tage.

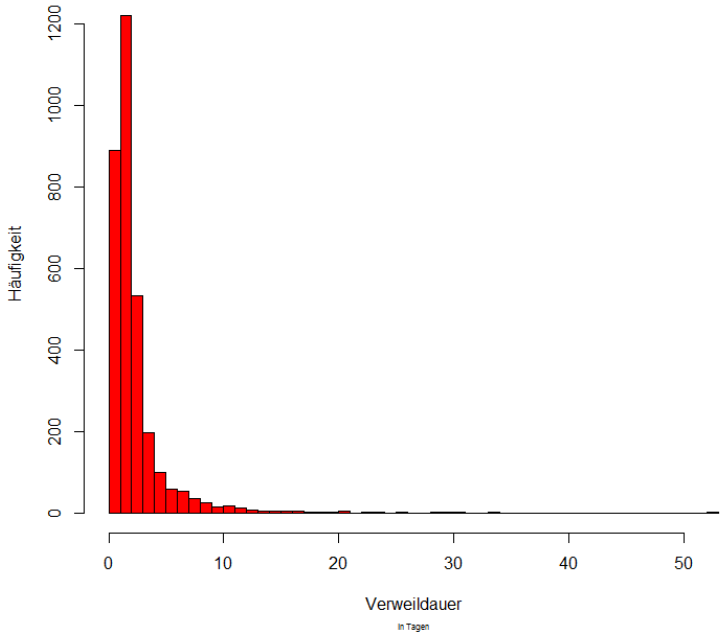


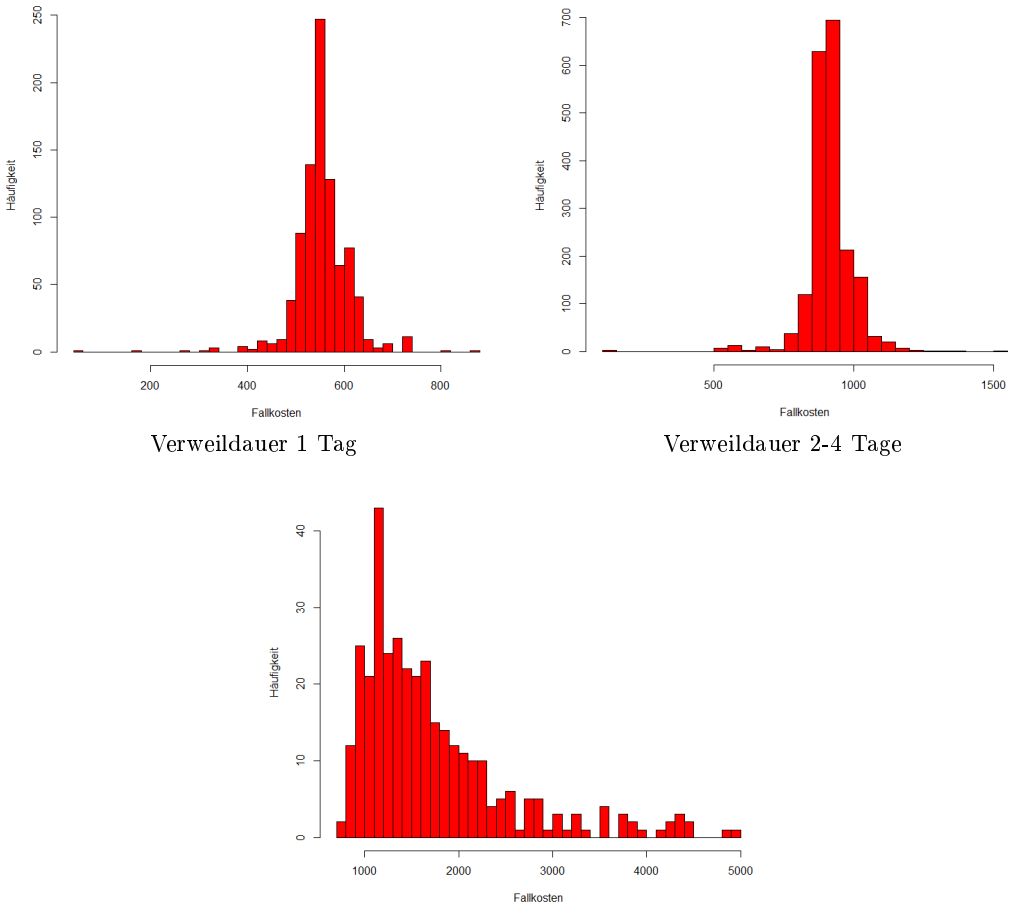
Abbildung 4.9.: Übersicht Verweildauer Gehirnerschütterung

Durch dieses Zusammenhang zwischen der diskret verteilten Verweildauer und den Kosten für einen Fall, ist es sinnvoll die Kostenverteilung in drei Teile zu zerlegen. Jeweils eine Verteilung für die Zeit mit Abschlägen auf das Kostengewicht (Verweildauer ein Tag), mit konstantem Kostengewicht (Verweildauer zwei bis vier Tage)und mit Zuschlä-

gen. Bezeichne F_A, F_K und F_Z jeweils die entsprechende Verteilungsfunktion und sei V die Verweildauer, dann lässt sich die Kostenverteilung wie folgt darstellen.

$$F_{Gesamt}(x) = \mathbb{P}(V \leq 1)F_A(x) + \mathbb{P}(1 < V \leq 4)F_K(x) + \mathbb{P}(V > 4)F_Z(x) \tag{4.2}$$

Betrachten wir nun erneut die Verteilung der Kosten in den Daten, allerdings unterteilt nach Verweildauer, dann ergibt sich folgendes Bild.



Verweildauer 1 Tag

Verweildauer 2-4 Tage

Verweildauer >4 Tage

Abbildung 4.10.: Übersicht Leistungskosten unterteilt nach Verweildauer

Eine Approximation durch stetige Verteilung ist für keine der drei Verteilungen gelungen. Zum Testen wurde der Kolmogorow-Smirnow-Test verwendet und als mögliche Verteilungen wurden Normalverteilung, Weibullverteilung, und im letzten Fall die Exponentialverteilung bzw. Gammaverteilung, untersucht. In jedem Fall konnte die Nullhypothese bei einem Signifikanzniveau von 1% verworfen werden. Aus diesem Grund verwenden wir

im weiteren die empirische Verteilung.

Der Erwartungswert der empirischen Kostenkostenverteilung lässt sich mit Hilfe des Mittelwert-Schätzers bestimmen. Damit ergeben sich durchschnittliche Kosten in Höhe von 922,85 Euro. Zusammen mit den geschätzten Unfallwahrscheinlichkeiten können wir somit das Risikoäquivalent für einen Versicherten bestimmen. Dieses ergibt sich als Summe der erwarteten Kosten für alle Folgejahre, bis zum Eintritt in die Prämortalitätsphase. Die erwarteten Kosten in einem Jahr entsprechen dem Erwartungswert des zusammengesetzten Poisson-Prozesses für dieses Jahr. Bezeichne Y_t^{AGG} den zusammengesetzten Poisson-Prozess und N_t^{AGG} den Poisson-Prozess für die Anzahl der Gehirnerschütterungen, jeweils für einen Versicherten in der entsprechenden AGG und seien weiterhin X_i die Kosten der i-ten Gehirnerschütterung, dann gilt

$$\mathbb{E}(Y_t^{AGG}) = \mathbb{E}(\sum_{i=1}^{N_t^{AGG}} X_i) = \lambda_{AGG} * t\mathbb{E}(X_i) = \lambda_{AGG} * t * 922,85 \text{ Euro} \tag{4.3}$$

Beispielsweise betragen die zu erwarteten Kosten für einen 30-jährigen Versicherten für das nächste Jahr rund 0,90 Euro.

Für die Bewertung der Modellgüte werden wir untersuchen, wie sich die realen Kosten im Vergleich zu diesen Erwartungswerten entwickeln. Dazu betrachten wir die anderen beiden Datenjahre und beginnen mit dem Jahr 2010. Im ersten Schritt testen wir, ob sich die relevanten Modellparameter signifikant verändert haben. Dazu führen wir einen Einstichproben-t-Test für die mittleren Fallkosten und einen gewichteten Einstichproben-t-Test für die alters- und geschlechtsabhängigen Unfallwahrscheinlichkeiten durch[15]. Die Nullhypothese ist dementsprechend, dass der Mittelwert der jeweiligen Daten aus 2010 gleich dem im Vorjahr geschätzten Modellparameter ist. Als Signifikanzniveau wählen wir 5%. Die folgende Tabelle zeigt eine Übersicht der Entwicklung der einzelnen Parameter und das Ergebnis des Tests.

Parameter	Wert 2009	Wert 2010	p-Wert des t-Tests
Mittlere Fallkosten	922,85 Euro	922,07 Euro	0,9254
λ Gesamt	0,1890%	0,1891%	0,9838
λ Frauen 1-25 Jahre	0,2720%	0,2897%	0,2971
λ Frauen 26-65 Jahre	0,0585%	0,0631%	0,3356
λ Frauen >65 Jahre	0,2820%	0,2762%	0,4831
λ Männer 1-25 Jahre	0,3829%	0,3574%	0,1751
λ Männer 26-65 Jahre	0,0973%	0,0930%	0,4090
λ Männer >65 Jahre	0,1867%	0,2016%	0,0675

Tabelle 4.4.: Übersicht Testergebnisse 2010

[15] Die entsprechenden Abfragen für R befinden sich im Anhang A.4.

Da keiner der resultierenden p-Werte $< 0{,}05$ ist, konnte die Nullhypothese für keinen der getesteten Parameter verworfen werden. Dementsprechend scheinen die geschätzten Parameter auch für 2010 korrekt zu sein.

Um eine detailliertere Auswertung zu ermöglichen, werden wir zusätzlich einen Simulationsansatz verfolgen. Anstatt einmal den gesamten Bestand auszuwerten, werden zufällig gezogene Stichproben untersucht. Da eine Stichprobe allein nur eine geringe Aussagekraft hat werden wir entsprechend, eines Monte-Carlo-Ansatzes, 10 000 zufällige Stichproben analysieren. Die geschätzten und realen Kosten werden dann, für alle Versicherten in der Stichprobe, aufsummiert und verglichen. So erhalten wir ein Maß für die Vorhersagegenauigkeit.

Zur Durchführung der Stichprobenziehung, und zur Berechnung der erwarteten Unfallkosten, wurde wieder ein Java Programm erstellt. Für jeden Versicherten werden dabei die erwarteten Kosten, entsprechend der Formel 4.3, bestimmt. Da wir λ auf Basis der jährlichen Unfallrate geschätzt haben, entspricht der Zeitraum der Schätzung t, dem Anteil des Jahres, den der Versicherte bei der Kasse war. In jedem Durchlauf werden Stichproben mit einen Umfang von 100 000 Versicherten gezogen. Anschließend werden die realen und die geschätzten Kosten aufsummiert und in eine separate Datei herausgeschrieben. Zur Auswertung der Ergebnisse kam wieder R zum Einsatz.

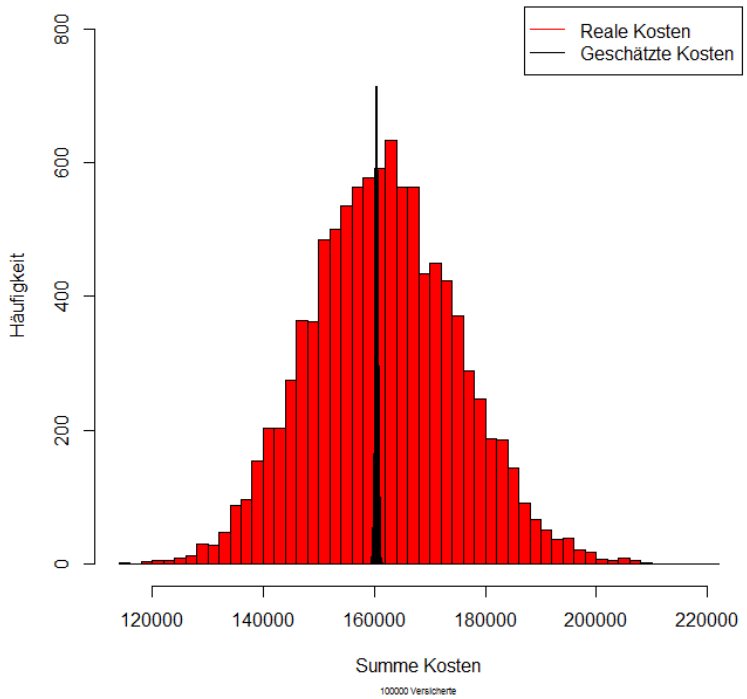


Abbildung 4.11.: Ergebnis Simulation 2010

Aufgrund der geringen Varianz der Erwartungswertschätzung, sind die geschätzten Kosten auf ein kleines Intervall konzentriert. Trotzdem ist gut zu sehen, dass die Mittelwerte der beiden Verteilungen sehr dicht beieinander liegen. Für die realen Kosten ist dieser Wert 161 695,80 Euro und für die geschätzten Kosten 160 408,70 Euro. Das bestätigt die Ergebnisse des t-Test für die Modellparameter.

Der Vorteil des Simulationsansatzes ist, dass wir die Differenz zwischen realen und geschätzten Kosten untersuchen können, um ein Verteilung der Abweichung von den Erwarteten Kosten zu erhalten. Das folgende Histogramm zeigt die Verteilung der Kostendifferenzen in den 10 000 Stichproben für 2010.

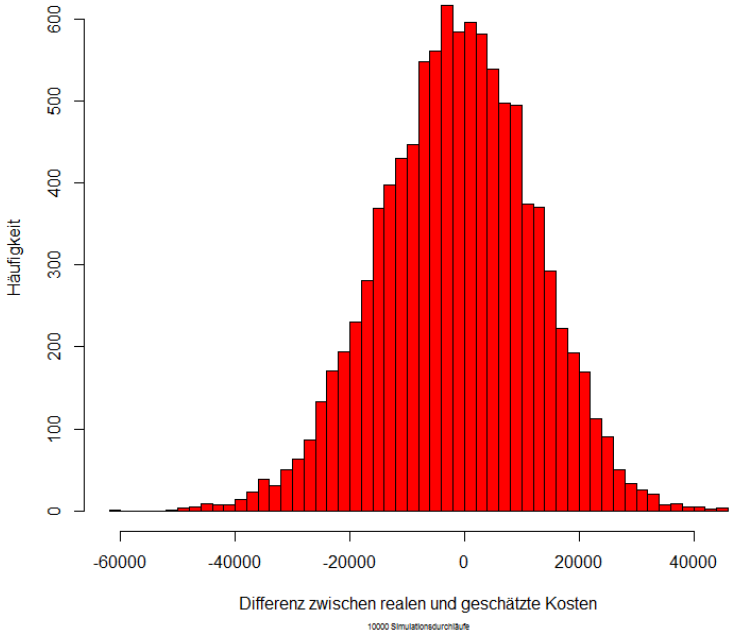


Abbildung 4.12.: Differenzen der aufsummierten Kosten in 2010

Der Mittelwert dieser Differenzen beträgt -1287,04 Euro. Das heißt wir unterschätzen im Mittel die eintretenden Kosten für 2010 geringfügig. Zurückzuführen ist dies auf eine leicht erhöhte Unfallrate in 2010. Obwohl die Differenzen normalverteilt aussehen, ließ sich dies nicht durch einen entsprechenden Test belegen. Ein angepasster Kolomogorov-Smirnow-Test (Lilliefors-Test) liefert einen p-Wert von 0,01326. Das heißt bei einem Signifikanzniveau von 5% muss die Nullhypothese, dass die Daten normalverteilt sind, verworfen werden. Trotzdem kann die empirische Verteilung dazu verwendet werden Risikozuschläge zu bestimmen, damit wir die eintretenden Kosten nicht unterschätzen.

Beispielsweise könnte der Risikozuschlag bestimmt werden, damit die Differenz aus realen Kosten und geschätzten Kosten in 95% der Fälle positiv ist. Dazu betrachten wir im

einfachsten Fall das 5% Quantil der Verteilung in Abbildung 4.12: $Q_{0,05} = -23554,70$ Euro. Da jede Stichprobe 100 000 Versicherte umfasst, würde ein einheitlicher Risikozuschlag von rund 24 Cent für jeden Versicherten ausreichen. Diese einfache Berechnung ist lediglich als Beispiel dafür gedacht, wofür die Verteilung der Abweichung verwendet werden kann.

Grundsätzlich lässt sich sagen, dass die Ergebnisse im Jahr 2010 im Mittel stabil waren. Als Nächsten werden wir das Jahr 2011 betrachten. Dazu untersuchen wir wieder zu Beginn die geschätzten Parameter mit Hilfe eines t-Test. Das Vorgehen ist dabei das Gleiche, wie für das Jahr 2010. Der Test liefert folgende Ergebnisse.

Parameter	Wert 2009	Wert 2011	p-Wert des t-Tests
Mittlere Fallkosten	922,85 Euro	933,43 Euro	0,1382
λ Gesamt	0,1890%	0,2056%	0,00002
λ Frauen 1-25 Jahre	0,2720%	0,2941%	0,1909
λ Frauen 26-65 Jahre	0,0585%	0,0611%	0,7179
λ Frauen >65 Jahre	0,2820%	0,3227%	0,00001
λ Männer 1-25 Jahre	0,3829%	0,3882%	0,7801
λ Männer 26-65 Jahre	0,0973%	0,0954%	0,7221
λ Männer >65 Jahre	0,1867%	0,2180%	0,0004

Tabelle 4.5.: Übersicht Testergebnisse 2011

Es ist gut zu sehen, dass sich die Unfallwahrscheinlichkeit, bei einem Signifikanzniveau von 5%, signifikant verändert hat. Dies betrifft in erster Linie die Versicherten, die älter als 65 sind. Das kann verschiedene Ursachen haben. Im einfachsten Fall ist es ein Phänomen, dass sich auf das Jahr beschränkt, zum Beispiel ein besonders schnee- und eisreicher Winter[16], der durch die erhöhte Glatteisgefahr zu mehr Stürzen bei älteren Menschen geführt hat. In diesem Fall spricht dagegen, dass die Unfallwahrscheinlichkeit in den anderen Gruppen weitestgehend stabil geblieben ist. Eine Möglichkeit zum Ausgleich von Effekten, die auf einen begrenzten Zeitraum beschränkt sind, wäre zum Beispiel ein Risikoaufschlag der von der Häufigkeit dieser Effekte abhängt.

Es besteht jedoch auch die Möglichkeit, dass die Änderungen permanent sind und, dass sich das beobachte System verändert hat. Zum Beispiel könnten Anpassungen im Abrechnungssystem dazu geführt haben, dass ältere Versicherte eher in DRG B80Z eingestuft wurden. Es könnte sich die Definition einer DRG geändert haben, die besonders die Altersgruppe über 65 betrifft, oder durch strengere Prüfkriterien könnten Krankenhäuser gezwungen werden die B80Z abzurechnen, anstatt einer ähnlichen möglicherweise teureren DRG. Bei einem solchen Effekt ist der Einsatz eines Korrekturfaktors das übliche Vorgehen in der PKV[17].

[16] Im Jahr 2011 waren die Wintermonate jedoch im Schnitt zu warm und eher regnerisch. (Quelle: `http://www.wetterprognose-wettervorhersage.de` Stand 11.02.15)

[17] Siehe Projektunterlagen GFL: „Hochrechnung, Risikozuschlagsverrechnung, Schwangerschaftskos-

Eine Dritte mögliche Ursache für diese Entwicklung könnte ein sogenannter Drift in den Daten sein. Das heißt durch äußere Einflussfaktoren, die bisher nicht im Modell berücksichtigt wurden, verändern sich die Modellparameter kontinuierlich. Denkbare Faktoren wären zum Beispiel, medizinischer Fortschritt, der zu einer Kostensenkung bzw. zu einer Kostensteigerung führt oder Entwicklungen innerhalb der Bevölkerung, die das Risiko für einen Unfall beeinflussen. Auch in diesem Fall kann versucht werden, diese Effekte mit Hilfe eines Korrekturfaktors auszugleichen, der jedoch jedes Jahr erneut angewendet wird.

Zusammenfassend können wir festhalten, dass der Poisson-Ansatz zur Modellierung der isolierten Ereignisse die erwartenden Unfallkosten gut vorhersagen kann und in der Regel stabile Ergebnisse produziert. Die größte Herausforderung bei der Modellierung ist die klare Abgrenzung der Unfälle, da diese sehr verschiedene Ausprägungen haben können. Äußere Effekte, die die Modellparameter beeinflussen, können durch Korrekturfaktoren oder Risikozuschläge ausgeglichen werden.

tenbewertung, Kopfschadenstatitik und Auswertung für den PKV Bundesverband“

5. Modellierung Singuläre Ereignisketten

Nachdem wir ausführlich die isolierten Ereignisse untersucht haben, soll in diesem Kapitel ein Risikomodell für die singulären Ereignisketten (im Folgenden alternativ auch als chronische Erkrankungen bezeichnet) vorgestellt werden. Zuerst werden die allgemeinen Eigenschaften der chronischen Erkrankungen analysiert und anschließend wird ein Modellansatz vorgestellt, der diese abbildet. Darauf aufbauend wird erläutert, was bei der Umsetzung dieses Ansatzes zu beachten ist. Außerdem werden die allgemeine Vorgehensweise bei der Implementierung und beim Test des Risikomodells vorgestellt.

5.1. Grundlegende Betrachtungen

Als singuläre Ereigniskette wird im Kohorten-Modell eine **beträchtliche Verschlechterung der Gesundheit** bezeichnet, **die bis zum Tod des Versicherten andauert**. Im Gegensatz zu den isolierten Ereignissen können nach dem Eintreten immer wieder neue Kosten anfallen. Da die chronische Erkrankung bis zum Tod andauert, ist eine Genesung ausgeschlossen. Typische Beispiele sind z.B. Diabetes oder verschiedene Formen von Krebs. Auch psychische Erkrankungen sind in den meisten Fällen nicht heilbar und treten bis zum Tod immer wieder in Schüben auf.

Für die Modellierung sind dementsprechend zwei Prozesse zu betrachten. Zum einen der Eintritt in die singuläre Ereigniskette und zum anderen das regelmäßige Auftreten von Kosten. Für einen Versicherten beginnt die singuläre Ereigniskette in dem Moment, wenn bei ihm eine chronischen Krankheit diagnostiziert wird. Wir werden für die Modellierung annehmen, dass sich dieses Ereignis nicht ankündigt und unabhängig von der medizinischen Vergangenheit des Versicherten ist. Die Wahrscheinlichkeit für den Eintritt hängt hauptsächlich vom Typ der Erkrankung ab. Außerdem werden auch Alter und Geschlecht des Versicherten einen Einfluss auf die Wahrscheinlichkeit haben. Eine Abhängigkeit von der Saison ist dagegen nicht zu erwarten, da Chroniker in der Regel das ganze Jahr über an der Erkrankung leiden.

Inwiefern, und in welcher Form, nach dem Eintritt in die singuläre Ereigniskette Kosten auftreten, kann sehr unterschiedlich sein. Im einfachsten Fall sind es regelmäßig verordnete Arzneimittel sowie Routineuntersuchungen. Eine andere Möglichkeit sind regelmäßige Eskalationen in Form von längeren Krankenhausaufenthalten und anschließenden REHA-Behandlungen. Des Weiteren kann sich der Zustand des Versicherten im Verlauf der chronischen Erkrankung kontinuierlich verschlechtern, was zu häufigeren und aufwendigeren Behandlungen, und damit höheren Kosten, führt. Alle diese Effekte können

sich auch überlagern. Außerdem ist es denkbar, dass weitere chronische Erkrankungen hinzukommen.

Diese Komorbiditäten sorgen dafür, dass die verschiedenen chronischen Erkrankungen in der Regel nicht unabhängig voneinander sind. Chroniker entwickeln zum Beispiel häufiger Depressionen. Bewegungsmangel und falsche Ernährung führen oftmals zu Adipositas, Diabetes und Bluthochdruck, sowie zu deren Folgeerkrankungen, wie etwa Koronaren Herzkrankheiten. Das heißt sobald in der Versichertenbiografie ein Ereignis den Eintritt in die singuläre Ereigniskette auslöst, entstehen nicht nur zusätzliche Kosten für die Behandlung der auslösenden chronischen Erkrankung, sondern nach und nach entstehen ebenso Kosten für Begleit- und Folgeerkrankungen. Aus diesem Grund ist es sinnvoll nicht den parallelen Verlauf verschiedener chronischer Erkrankungen zu betrachten, sondern die Kostenverläufe, ausgehend von der initialen chronischen Erkrankung, zu bewerten.

Dieses Vorgehen hat einen weiteren Vorteil. Im Gegensatz zu Unfällen lässt sich bei chronischen Erkrankungen im Allgemeinen nicht eindeutig identifizieren, welche Kosten der Behandlung der einzelnen Erkrankungen zuzuordnen sind. Wird etwa ein chronisch Kranker ambulant bei seinem Hausarzt behandelt, so lassen sich die dabei entstehenden Behandlungskosten, aufgrund der pauschalen Vergütung des Arztkontaktes, nicht einzelnen Krankheitsbildern zuordnen. Dies trifft ebenso im stationären Bereich zu. Auch bei den Arzneimittelkosten kann im Allgemeinen nicht exakt zugeordnet werden, welches Krankheitsbild durch das Arzneimittel behandelt werden soll. So werden etwa Statine vorwiegend bei Hyperlipidämie[1]verordnet, gleichzeitig dienen diese jedoch sowohl zur Prophylaxe der Arteriosklerose, als auch zur Behandlung bei Multipler Sklerose.

Zusammenfassend ist dies ein geeigneter Weg, unter den gegebenen Rahmenbedingungen die Kosten zu betrachten. Untersucht wird im Folgenden also die Verteilung der Leistungskosten von chronisch erkrankten Versicherten in einem Jahr.

5.2. Modellierung

Zur Abbildung der singulären Ereignisketten müssen, wie beschrieben, zwei Prozesse modelliert werden. Wir untersuchen zuerst den Eintritt in den Zustand der chronischen Erkrankung und überlegen uns anschließend, wie sich die Kosten, die in der Zeit als Chroniker entstehen, abbilden lassen. Die grundlegende Modellannahme ist, dass sich das Ereignis des Eintritts nicht ankündigt und unabhängig von eventuellen Vorerkrankungen ist. Trotz der Ähnlichkeiten zu den Charakteristiken eines Unfalls, können wir die Erkenntnisse aus dem vorangegangen Kapitel nicht anwenden, da jeder Versicherte nur einmal in die singuläre Ereigniskette eintreten kann.

[1] zu hoher Cholesterinwert

Wir können uns den Eintritt in die singuläre Ereigniskette als einfache Markov-Kette vorstellen. Diese hat zwei Zustände i_0 :=„gesund“ und i_1 :=„Chroniker“.

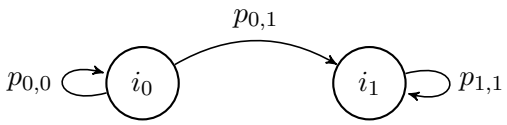


Abbildung 5.1.: Markov-Modell für isolierte Ereignisse

Da es es laut Definition der singulären Ereignisketten keine Genesungschance gibt, existiert kein Pfad von i_1 nach i_0. Das heißt i_1 ist ein absorbierender Zustand und die Markov-Kette ist nicht irreduzibel. Wir gehen im Folgenden von einem Jahr als Zeitraum für einen Prozessschritt aus. Bei Bedarf können jedoch auch kleinere Intervalle betrachtet werden. Das Verweilen im Zustand i in einer Markov-Kette mit konstanten Übergangswahrscheinlichkeiten, lässt sich als Bernoulli-Experiment mit Erfolgswahrscheinlichkeit $p_{i,i}$ interpretieren. Damit ist die Verweildauer T_i geometrisch verteilt mit Parameter $p_{i,i}$ und die Wahrscheinlichkeit dafür, dass die Markov-Kette bei Start im Zustand i_0 nach n Schritte im Zustand i_1 ist, gegeben durch

$$p_{0,1}(n) = \mathbb{P}(T_0 < n) = 1 - \mathbb{P}(T_0 >= n) = 1 - \mathbb{P}(T_0 = n) = 1 - p_{0,0}^n = 1 - p_{0,0}^{(n)}.$$

Es gilt $p_{0,0} = 1 - p_{0,1}$ und deshalb reicht es aus, wenn wir die Übergangswahrscheinlichkeit in den Zustand i_1 schätzen. Da i_1 gleichbedeutend damit ist, dass bei dem Versicherten eine chronischen Erkrankungen diagnostiziert wurde, zerfällt die Wahrscheinlichkeit $p_{0,1}$ in eine Summe aus einzelnen Eintrittswahrscheinlichkeiten. Die verschiedenen, möglichen chronischen Erkrankungen bezeichnen wir im Folgenden als $C_1, ..., C_n$ und die Ersteintrittswahrscheinlichkeit in eine dieser Erkrankungen als $p_{0,C_1}, ..., p_{0,C_n}$. Dabei ist n die Anzahl aller diagnostizierbaren chronischen Erkrankungen. Damit ist

$$p_{0,1} = \sum_{k=1}^{n} p_{0,C_k}.$$

Dadurch ergibt sich eine Markov-Kette mit n absorbierenden Zuständen. Die Abgrenzung der Ereignisse „Eintritt in Krankheit C1“, „Eintritt in Krankheit C2“ usw. wird im späteren Verlauf des Kapitels erläutert.

Die Wahrscheinlichkeit dafür, dass eine bestimmte chronische Erkrankung bei einem Versicherten zum ersten Mal diagnostiziert wird, kann wieder auf Basis der historischen Daten bestimmt werden. Analog zum Vorgehen beim Poisson-Modell für die Unfälle, betrachten wir dafür die relative Häufigkeit in vorher definierten Alters- und Geschlechtsgruppen (AGG). Da wir uns nur für Versicherte interessieren, die neu im Zustand „Chroniker“ für eine bestimmte Erkrankung sind, müssen wir zur Bestimmung der Übergangswahrscheinlichkeiten zwei Jahre betrachten. Als Grundmenge dienen alle Versicherten,

die im ersten Jahr keine chronische Erkrankung hatten. Durch Auswertung des zweiten Jahres können dann die Wahrscheinlichkeiten für alle $k \in [1, n]$ wie folgt bestimmt werden

$$\bar{p}_{0,C_k}(AGG) := \frac{\#\{\text{neu an } C_k \text{ erkrankte Chroniker}\}_{AGG}}{\#\{\text{Versicherte in Grundmenge}\}_{AGG}}. \tag{5.1}$$

Da diese Übergangswahrscheinlichkeiten nicht mehr konstant sind, ist die Verweilzeit im Zustand i_0 nicht mehr geometrisch verteilt und es müssen die n-stufigen Übergangswahrscheinlichkeiten verwendet werden. Damit ist klar beschrieben, wie der Eintrittszeitpunkt in den Zustand Chroniker im Modell bestimmt werden kann. Ab dem Jahr des Eintritts in die singuläre Ereigniskette entstehen Kosten. Der entsprechende Kostenprozess hängt in erster Linie davon ab, welche Krankheit zuerst diagnostiziert wurde. Chronische Erkrankungen können auf sehr unterschiedliche Weise und in verschiedenen Phasen verlaufen. Wir werden die möglichen grundsätzlichen Verläufe vorstellen und beschreiben, wie diese modelliert werden können. Dabei liegt der Fokus auf den resultierenden Kosten.

Der einfachste Fall ist ein stabiler Krankheitsverlauf. Durch Medikamente und regelmäßige Behandlungen werden die Symptome der Krankheit unter Kontrolle gehalten und es kommt zu keiner weiteren Verschlechterung des Gesundheitszustands. Ein gutes Beispiel dafür sind Patienten mit einer Depression. Diese erhalten Antidepressiva und sind angehalten regelmäßig einen Arzt zu konsultieren. In dieser Phase entstehen gleichmäßige jährliche Kosten, deren Verteilungsfunktion, anlog zum Vorgehen im vorangegangen Kapitel, modelliert werden kann. Das heißt solange der Versicherte in dieser Phase ist, sind die Kosten ein einfacher Stochastischer Prozess $\{X_n, n \in \mathbb{N}\}$, wobei die X_i unabhängig und identisch verteilt sind und jeweils den Kosten im i-ten Jahr entsprechen.

Es besteht die Möglichkeit, dass sich, mit fortschreitenden Krankheitsverlauf, der Gesundheitszustand des Patienten verschlechtert, so dass häufigere und intensivere Behandlungen notwendig werden. Dadurch steigen die Kosten über die Jahre an und sind dementsprechend nicht mehr gleich verteilt. Stattdessen hängen davon ab, wie lange der Versicherte bereits an der Krankheit leidet. Beispiele sind Krankheiten wie Multiple Sklerose oder Alzheimer, bei denen durch den voranschreitenden Verfall der Nerven immer weitere Therapien, oder sogar eine Pflegebetreuung notwendig werden. Da wir in diesem Modellansatz Jahresscheiben betrachten, kann das Voranschreiten der jeweiligen Krankheit als eine Aneinanderreihung von Phasen mit stabilem Kostenverlauf interpretiert werden. Je nach Krankheitsverlauf dauern diese Phasen ein bis mehrere Jahre. Die Abgrenzung der Phasenübergänge kann anhand der historischen Informationen vorgenommen werden. Um jedoch die Krankheitsverläufe abbilden zu können, ist die Analyse von Versicherungsdaten über einen längeren Zeitraum nötig.

Neben den normalen Kosten können auch Eskalationen auftreten. Diese sind zeitlich begrenzte Verschlechterungen des Gesundheitszustands, die zusätzliche Kosten verursachen. Diese treten selten auf, kündigen sich nicht an und haben keine Auswirkungen auf

den allgemeinen Krankheitsverlauf. Eine häufige Ursachen dafür ist ein Fehlverhalten des Patienten, wie zum Beispiel eine unzureichende Dosierung der Medikamente oder ein falscher Lebensstil. Insbesondere bei Psychopharmaka kann eine Vernachlässigung der Medikation zu einer akuten Krankheitsepisode führen, die im schlimmsten Fall eine stationäre Behandlung erforderlich macht. Es handelt sich also um isolierte Ereignisse, die nur auftreten, wenn der Versicherte an einer chronischen Erkrankung leidet. Dementsprechend können wir die Erkenntnisse aus dem vorangegangen Kapitel dazu verwenden, um diesen Teil der Kosten zu modellieren. Damit gilt, unter Verwendung derselben Bezeichner wie in Kapitel 4, dass die Eskalationskosten durch einen zusammengesetzten Poisson-Prozess der Form

$$Y_t^E = \sum_{i=1}^{N_t^E} X_i^E$$

abgebildet werden können. Dabei bezeichnet E den Eskalationstyp. Für eine chronische Erkrankung sind auch mehrere Eskalationstypen denkbar.

Der Kostenprozess setzt sich also aus zwei Teilen zusammen, den Grundkosten $\{X_n, n \in \mathbb{N}\}$ und einem Anteil für die Eskalationen $\{Y_t^E\}_E$.

5.3. Erläuterungen zur Anwendung und Test des Modells

Chronische Erkrankungen verlaufen über einen großen Zeitraum, in dem die Patienten ambulante, stationäre und Arzneimittel-Kosten verursachen. Aus diesem Grund ist die auf stationäre Leistungen beschränkte Datenbasis aus dem vorangegangen Kapitel nicht ausreichend, um das Phänomen der singulären Ereignisketten beispielhaft zu untersuchen. Deshalb werden wir in diesem Kapitel nur das theoretische Vorgehen bei der Umsetzung des oben beschriebenen Modellsatzes erläutern. Dazu gehen wir zu Beginn auf die Anforderungen an die Datengrundlage ein und erläutern die Vorgehensweise zur Identifikation von Chronikern. Anschließend beschreiben wir, wie die Modellparameter geschätzt und die Ergebnisse getestet werden können.

5.3.1. Identifikation von Chronischen Erkrankungen und Abgrenzung der Leistungen

Die Datengrundlage für die Anwendung des Modells sollte aus Versicherten- und Abrechnungsdaten bestehen, die über einen Zeitraum von mehreren Jahren erfasst wurden. Dabei sollten Informationen zu stationären und ambulanten Leistungen sowie Informationen zu den Arzneimittel vorliegen. Die Daten müssen bereinigt sein und über die Jahre hinweg konsistent sein. Des Weiteren wird für eine stabile Schätzung der Modellparameter eine ausreichende Anzahl an Versicherten benötigt.

Eine der großen Herausforderungen bei der Anwendung dieses Modellansatzes ist es, chronische Erkrankungen in den Daten zu identifizieren und klar gegeneinander abzugrenzen.

Eine Möglichkeit dafür bieten die Diagnosen im ICD-10-GM Katalog. Chronische Erkrankung sind dort zwar nicht speziell gekennzeichnet. Es werden jedoch von verschiedenen Verbänden Listen mit chronischen Diagnosen[2] veröffentlicht, die als Ausgangspunkt für eine Abgrenzung dienen können. Aufgrund der hierarchischen Struktur des Diagnosekatalogs sind zusammengehörige Diagnosen auch gut zu erkennen. Auf diese Weise können Diagnosegruppen identifiziert werden, welche die Zugehörigkeit zu den, im Modellansatz mit $C_1, ..., C_n$ bezeichneten, chronischen Erkrankungen definieren.

Es reicht jedoch nicht aus, Versicherte als Chroniker zu definieren, die einmal eine entsprechende Diagnose hatten. Zum einen besteht immer die Gefahr einer Fehldiagnose und zum anderen gibt es Erkrankungen die nicht zwangsläufig chronisch verlaufen müssen. Beispiele dafür wären Depressionen oder Schizophrenie, die auch als einzelne Episode auftreten können und danach nicht wieder vorkommen. Deshalb ist notwendig zu überprüfen, ob die Diagnose wiederkehrt, bevor die Zuordnung des Versicherten zu einer der Chronikergruppen erfolgen kann. Der Zeitraum, der dafür betrachtet werden muss, hängt von der Art der chronischen Erkrankung ab. In der Regel sollte es jedoch ausreichen das Jahr nach der ersten Diagnose zu überprüfen. Als Eintrittszeitpunkt in die singuläre Ereigniskette gilt dann das Jahr der ersten Diagnose.

Aufgrund des Abrechnungssystems in Deutschland kann für einzelne Leistungsdaten, nicht entschieden werden, welcher Erkrankung sie zuzuordnen sind. Deshalb werden wir alle Leistungen, die ein Versicherter nach dem Eintritt in eine singuläre Ereigniskette verursacht, der entsprechenden chronischen Krankheit zuweisen. Davon ausgenommen sind alle Leistungen, die als Unfälle im Sinne der isolierten Ereignisse definiert sind.

Dieses Vorgehen erlaubt es für jeden Versicherten, in den Daten, zu entscheiden, wer in welchen Jahren ein Chroniker war und welche Kosten dadurch entstanden sind. Diese Information bilden die Basis für die Umsetzung des obigen Modellansatzes im nächsten Abschnitt.

5.3.2. Anwendung und Test des Modells

In diesem Abschnitt wird das allgemeine Vorgehen bei der Anwendung des oben beschriebenen Modellansatzes erläutert. Außerdem wird darauf eingegangen, wie die

Zunächst werden wir beschreiben, was beim Schätzen der Eintrittswahrscheinlichkeit in den Zustand „Chroniker“ beachtet werden muss. Im Modell ist bereits beschrieben, dass dafür die Chronikerinformationen aus zwei Jahren benötigt werden. Die Entscheidung, ob ein Versicherter im zweiten Jahr ein Chroniker geworden ist, erfordert zusätzlich die Auswertung von mindestens einem weiteren Jahr. Das heißt es werden Daten aus mindestens drei Jahren benötigt. Aufgrund dieses großen Zeitraums bietet es sich an, die vorhandenen Daten in eine Lern- und eine Testmenge zu unterteilen, damit für die

[2] z.B. `http://www.kvhb.de/sites/default/files/icd_chronische_krankheiten_2013.pdf` (Stand: 13.02.15)

spätere Bewertung der Modellgüte, nicht noch weitere Datenjahre benötigt werden. Die Unterteilung sollte zufällig sein und es sollte sichergestellt werden, dass beide Mengen hinreichend viele Versicherte enthalten.

Auf der Lernmenge können nun die Versicherten untersuchen werden, die im ersten Jahr keiner chronischen Erkrankung zugewiesen wurden. Diese bilden die Grundmenge für die Schätzung der Eintrittswahrscheinlichkeiten. Es können nur Versicherte betrachtet werden, für die über den gesamten Zeitraum vollständige Daten vorliegen. Ansonsten bestünde die Gefahr die Eintrittswahrscheinlichkeit zu unterschätzen. Für jede chronische Erkrankung müssen anschließen die in der Formel 5.1 definierten alters- und geschlechtsabhängigen, relativen Häufigkeiten ausgewertet werden. Als erste Einteilung für die AGGs haben sich Intervalle von fünf Jahren bewährt. Das ermöglicht einen guten Überblick darüber, ob und in welcher Form die Wahrscheinlichkeiten von Alter und Geschlecht abhängen. Dies bildet dann die Grundlage zur Bildung hinreichend großer Gruppen, um den relativen Fehler bei der Schätzung zu minimieren.

Die Stabilität der Schätzung der Eintrittswahrscheinlichkeiten kann anschließend auf der Testmenge evaluiert werden. Dies kann zum Beispiel, analog zum Vorgehen in Abschnitt 4.4.3, durch einen einseitigen t-Test überprüft. Eine signifikante Abweichung der Testdaten, zu einem Signifikanzniveau von 5% von den geschätzten Eintrittswahrscheinlichkeiten, deutet darauf hin, dass die Datenmenge nicht für eine stabile Schätzung ausreicht. Mögliche Ursachen wären dann eine zu feine Einteilung in Alters- und Geschlechtsgruppen, oder generell zu wenige Datensätze.

Nachdem die Eintrittswahrscheinlichkeiten bestimmt wurden, können die Leistungskosten analysiert werden. Dazu werden für jeden Krankheitstyp die Kosten alle Chroniker betrachtet, die dem jeweiligen Typ zugeordnet wurden. Die Leistungskosten werden im Modell als Jahreskosten betrachtet, die durch Eskalationen in Gestalt von isolierten Ereignissen ergänzt werden. Die Identifikation dieser Eskalationskosten ist dabei, wie schon bei den Unfällen, eine Herausforderung. Gelingt es einzelne Abrechnungspositionen zu erkennen, die für einen Eskalationsprozess in Frage kommen, dann kann der Poisson-Ansatz aus dem vorangegangen Kapitel angewendet werden, um diese Kosten abzubilden. Alle verbleibenden Leistungskosten der betrachteten Chroniker werden zusammengefasst und für die jeweiligen Jahre aggregiert. Dabei sind die Kosten von Versicherten, die nicht über den gesamten Beobachtungszeitraum versichert waren, entsprechend hochzurechnen.

Die empirische Verteilung dieser Jahreskosten kann verwendet werden um, zusammen mit den Eskalationskosten, die jährlichen Kosten der betrachteten Chroniker zu beschreiben. Es ist zu erwarten, dass diese Verteilung eine hohe Varianz und verschiedene Häufungspunkte hat. Das ist darauf zurückzuführen, dass wir fast alle Behandlungskosten, nach Eintritt in den singuläre Ereigniskette für die Modellierung verwenden müssen, da eine Abgrenzung nicht möglich ist. Außerdem können, mit lediglich zwei Datenjahren, keine Krankheitsverläufe abgebildet werden. Letzteres ließe sich durch die Betrachtung von längeren Zeitreihen ausgleichen. Dann könnten die im Modell beschriebenen Krankheits-

phasen identifiziert und die entsprechenden Übergange abgebildet werden.

Zur Bestimmung der erwartenden Kosten, im nächsten Jahr, für einen Chroniker in der Krankheit C_k, mit $k \in [1, n]$, müssen also die erwarteten Grundkosten und die erwarten Eskalationskosten betrachtet werden. Sei $\{X_i(C_k), i \in \mathbb{N}\}$ der Kostenprozess der Grundkosten, dann sind die Kosten im nächsten Jahr gegeben durch $X_1(C_k)$. Diese sind entsprechend der empirischen Verteilung der Jahreskosten verteilt. Den Erwartungswert der $\{X_i\}$ bezeichnen wir mit μ_{C_k}. Dieser kann mit Hilfe des Mittelwert-Schätzers auf den historischen Daten bestimmt werden. Seien weiterhin die Eskalationskosten, für jeden Eskalationstyp E_{C_k}, gegeben durch den zusammengesetzten Poisson-Prozess $Y_t^{E_{C_k}}$, mit erwarteten Kosten für eine Eskalation $\bar{E}_{C_k}$ und erwarteter Eskalationshäufigkeit $\bar{E}_{C_k}$. Dann können die erwarteten Kosten für einen Versicherten mit der chronischen Erkrankung C_k für das nächste Jahr bestimmt werden durch

$$\mathbb{E}(X_1 + \sum_{E_{C_k}} Y_1^{E_{C_k}}) = \mathbb{E}(X_1) + \mathbb{E}(\sum_{E_{C_k}} Y_1^{E_{C_k}}) = \mu_{C_k} + \sum_{E_{C_k}} \lambda_{E_{C_k}} * \bar{E}_{C_k} \tag{5.2}$$

Die erwarteten Chroniker-Kosten, aufgrund der Krankheit C_k, für einen gesunden Versicherten, für das nächste Jahr ergeben sich dann, aus dem Produkt dieses Erwartungswerts mit der geschätzten Übergangswahrscheinlichkeit $\bar{p}_{0,C_k}^{AGG}$. Auf diese Weise können für die Testmenge die erwarten Kosten durch Chroniker bestimmt werden. Für die Bewertung der Genauigkeit dieser Schätzung bietet sich der gleiche Simulationsansatz an, den wir bei den isolierten Ereignissen angewendet haben.

Zusammenfassend konnte also ein Risikomodell für die singulären Ereignisketten angegeben werden. Aufgrund der Vorgehensweise bei der Bestimmung der Modellparameter, ist zu erwarten, dass die Vorhersage erwartungstreu ist, wenn sich die Rahmenbedingen nicht ändern. Die Qualität der Vorhersage, hängt dabei stark von der, für die Modellierung, verwendeten Datenbasis ab. Je mehr Information zur Verfügung stehen, umso präziser lassen diese sich abbilden. Dabei sind insbesondere längere Zeitreihen interessant, um den Krankheitsverlauf besser untersuchen zu können. Im Fall sich ändernder Rahmenbedingung können die im vorangegangenen Kapitel vorgestellten Verfahren verwendet werden, um diese Änderung auszugleichen.

6. Zusammenfassung und Ausblick

Das Kohorten-Modell bietet eine Möglichkeit die restliche Risikolast durch einen Versicherten, für eine Krankenkasse abzuschätzen. Zwei wichtige Bestandteile dieses Modells, die isolierte Ereignisse und die singuläre Ereignisketten, wurden im Rahmen dieser Arbeit untersucht. Die Zielstellung war es, ein Risikoäquivalent zu modellieren, um die Kosten aufgrund von Unfällen und chronischen Erkrankungen abschätzen zu können.

Für die isolierten Ereignisse wurde ein zusammengesetzter Poisson-Prozess konstruiert, der die Unfallkosten eines Versicherten abbildet. In diesem Modell werden Unfälle als punktuelle und unabhängige Ereignisse, ohne bleibende Auswirkung auf den Versicherten, interpretiert. Dieser Modellansatz konnte erfolgreich, auf der Basis von stationären Abrechnungsdaten, für das Beispiel „Gehirnerschütterung“ umgesetzt werden. Im ersten Testjahr lieferte diese Umsetzung stabile Ergebnisse. Im zweiten Testjahr gab es signifikante Abweichungen von den vorhergesagten Kosten. Deshalb wurden mögliche Ursachen dafür diskutiert, und Verfahren vorgestellt um die Abweichung auszugleichen. Eine große Herausforderung bei der Umsetzung des Modellansatzes war es, die Unfälle voneinander sowie von den restlichen Leistungen abzugrenzen. Die Ursache dafür ist das Fallpauschalen basierende Abrechnungssystem im Krankenhausbereich. Dadurch ist schwierig die einzelnen Fälle den verschiedenen Unfalltypen zuzuordnen. Da jedoch im Kohorten-Modell alle Leistungskosten einem der fünf Bausteine zugeordnet werden gehen keine Kosten verloren.

Neben dem Poisson-Ansatz wurde zusätzlich die Möglichkeit untersucht, die isolierten Ereignisse über eine Markov-Kette abzubilden. Dieser Ansatz ermöglichte zwar die detaillierte Modellierung des Genesungsprozesses nach einem Unfall. Er führte aber andererseits zu sehr komplexen Modellen. Der wichtigste Kritikpunkt war jedoch, dass die isolierten Ereignisse als Überlagerung der Phase „regen Lebens“ definiert sind. Deshalb lässt sich der Markov-Ansatz in dem explizit Genesungszeiten berücksichtigt werden, nur schwer in das Kohorten-Modell integrieren.

Die singulären Ereignisketten wurden in zwei Prozesse aufgeteilt: Zum einen wurde das Ereignis des Eintritts in die Kette untersucht und zum anderen in welcher Form nach diesem Eintritt Kosten entstehen. Für diese Betrachtung wurden Jahresscheiben gewählt und dadurch konnte der Eintrittsprozess als Markov-Kette mit absorbierenden Zuständen dargestellt werden. Eine größere Herausforderung war die Modellierung der Kosten. Die pauschale Vergütung, im ambulanten und stationären Bereich, verhinderte eine klare Zuordnung von einzelnen Behandlungen zu bestimmten Krankheitsbildern. Dadurch war es nicht möglich die Kosten, die ein Chroniker aufgrund seiner Krankheit verursacht,

von den restlichen Leistungskosten abzugrenzen. Deshalb wurde folgender Weg gewählt: Alle, ab dem Eintritt in die singuläre Ereigniskette, anfallenden Kosten wurden der initialen chronischen Erkrankung zugeordnet. Davon ausgenommen waren die Kosten für isolierte Ereignisse. Untersucht wurde anschließend die Verteilung der Leistungskosten von chronisch erkrankten Versicherten in einem Jahr. Diese konnten im einfachsten Fall als Stochastischer Prozess mit unabhängigen und identisch verteilten jährlichen Leistungskosten dargestellt werden. Für den Fall das die Möglichkeit besteht, das Kostenverhalten von Chronikern, über einen längere Zeitraum, zu untersuchen, wurden zusätzlich verschiedene Krankheitsphasen modelliert. Neben den Grundkosten können außerdem Eskalationen auftreten, diese wurden, analog zu den isolierten Ereignissen, als zusammengesetzter Poisson-Prozess abgebildet.

Aufgrund der umfangreichen Anforderungen an die Datenbasis war es im Rahmen der Arbeit nicht möglich, den vorgestellten Modellansatz für die singulären Ereignisketten umzusetzen. Deshalb wurde theoretisch erläutert, was bei der Umsetzung beachtet werden muss und wie die Modellparameter geschätzt und getestet werden können.

Zusammenfassend konnte gezeigt werden, dass es möglich war für beide Phänomene ein Risikoäquivalent zu modellieren, welches unter den gegeben Voraussetzungen erwartungstreu ist. Für die Einbettung der Ergebnisse in das Kohorten-Modell, sollte an dieser Stelle noch einmal der Umgang mit der Prämortalitätsphase erläutert werden. Die Phase der Neugeboren ist klar abgegrenzt, so dass es zu keiner Überschneidung kommen kann. Die Kosten in der Phase „regen Lebens“ werden so aufgeteilt, dass jede Leistung, die während dieser Phase, auftritt entweder den isolierten Ereignissen, den singulären Ereignisketten oder der Phase selbst zugeordnet wird. Die Abgrenzung der Prämortalitätsphase ist dagegen nicht so einfach zu erkennen. Bei der Betrachtung eines Ausschnitts aus einer Krankenversicherungsbiographie ist, mit Ausnahme für Neugeborene, nicht entscheidbar, ob der Versicherte bereits in der Prämortalitätsphase ist oder sich noch in Phase „regen Lebens“ befindet. Somit ist nicht klar welchem Teil des Modells seine Kosten zugeordnet werden müssen. Für diese Abgrenzung ist es notwendig längere Zeitreihen zu analysieren, um auf dieser Basis zu entscheiden, wie die Kosten zuzuordnen sind.

Die in dieser Arbeit vorgestellten Ansätze könnten durch die Betrachtung zusätzlicher Daten und längerer Zeiträume noch verfeinert werden. Insbesondere durch aussagekräftigere Abrechnungsdaten könnte die Zuordnung, der Kosten zu den einzelnen Modellbestandteilen, verbessert werden. Eine klarere Abgrenzung würde die Vorhersagegenauigkeit verbessern und eventuell zusätzliche Verteilungsaussagen ermöglichen. Auf dieser Basis könnten weiterführende risikotheoretische Erkenntnisse gewonnen werden.

A. Anhang

A.1. Zitierte Erkenntnisse und Nebenrechnungen

Satz A.1. *(Transformationssatz)*
Seien U und V offene Teilmengen des $\mathbb{R}^n$ und sei weiterhin $T : U \to V$ ein Diffeomorphismus. Das ist die Funktion f auf V genau dann über V integrierbar, wenn die Funktion $(f \circ T)|detDT|$ über U integrierbar ist. Es gilt dann:

$$\int_U f(T(x))|det(D * T(x))|dx = \int_V f(y)dy.$$

Dabei ist D die Jacobi-Matrix und $detDT(x)$ die Funktionaldeterminante von T.
Vergleiche [3, Kapitel 9 - Abschnitt 1].

Satz A.2. *(Satz von der monotonen Konvergenz)*
Seien $(\Omega, \mathfrak{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\{X_n\}_{n\in\mathbb{N}}$ eine nicht negative, fast sicher monoton wachsende Folge von Zufallsvariablen, dann gilt für ihre Erwartungswerte

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(\lim_{n\to\infty} X_n).$$

Das heißt Integration und Grenzwertbildung können vertauscht werden.
Vergleiche [4, Seite 109].

Nebenrechnung A.3. *Wir zeigen mit Hilfe vollständiger Induktion, dass für alle $y_1 ... y_n \in \mathbb{R}$ und $t_0, t_1 \in \mathbb{R}$ gilt*

$$\int_{t_0}^{t_1} \int_{y_1}^{t_1} ... \int_{y_{n-1}}^{t_1} dy_n ... dy_1 = \frac{(t_1 - t_0)^n}{n!}.$$

Induktionsanfang:

$$\int_{y_{n-1}}^{t_1} dy_n = (t_1 - y_{n-1})$$

Induktionsschritt: Sei

$$\int_{y_1}^{t_1} ... \int_{y_{n-1}}^{t_1} dy_n ... dy_2 = \frac{(t_1 - y_1)^{n-1}}{(n-1)!}$$

Dann ist

$$\begin{aligned}\int_{y_1}^{t_1} \dots \int_{y_n}^{t_1} dy_{n+1} \dots dy_2 &= \int_{y_1}^{t_1} \frac{(t_1 - y_2)^{n-1}}{(n-1)!} dy_2 \\ &= \big[-\frac{1}{n!}(t - y_2)^n \big]_{y_1}^{t} \\ &= \frac{1}{n!}(t - y_1)^n\end{aligned}$$

□

A.2. Abkürzungsverzeichnis und Zeichenerklärungen

$\mathbb{N}$	Natürliche Zahlen
$\mathbb{N}_0$	Natürliche Zahlen *inklusive* 0
$\mathbb{R}$	Reelle Zahlen
$\mathbb{R}^+$	die Menge reeller, nicht negativer Zahlen
F_X	Verteilungsfunktion der indizierten Zufallsgröße X
$\mathbb{E}(X)$	Erwartungswert der Zufallsgröße X
$Var(X)$	Varianz der Zufallsgröße X
$(\Omega, \mathfrak{A}, \mathbb{P})$	Wahrscheinlichkeitsraum
$X \sim F$	die Zufallsgröße X habe die Verteilung F
p_k	Wahrscheinlichkeitsfunktion diskreter Zufallsgrößen, $\mathbb{P}(X = k)$
$\mathbb{1}$	Identische Abbildung
# Menge	Anzahl der Elemente in der Menge

A.3. Codierung nach § 301 Abs. 3 SGB V

Im Folgenden wird ein Auszug aus der Anlage 2 zur § 301-Vereinbarung wiedergegeben[1]

Schlüssel 1: Aufnahmegrund

1. u. 2. Stelle	01	Krankenhausbehandlung, vollstationär
	02	Krankenhausbehandlung vollstationär mit vorausgegangener vorstationärer Behandlung
	03	Krankenhausbehandlung, teilstationär
	04	Vorstationäre Behandlung ohne anschließende vollstationäre Behandlung
	05	Stationäre Entbindung
	06	Geburt
	07	Wiederaufnahme wegen Komplikationen (Fallpauschale) nach KFPV 2003
	08	Stationäre Aufnahme zur Organentnahme
	09	- frei -
3. u. 4. Stelle	01	Normalfall
	02	Arbeitsunfall / Berufskrankheit (§ 11 Abs. 5 SGB V)
	03	Verkehrsunfall / Sportunfall / Sonstiger Unfall (z. B. § 116 SGB X)
	04	Hinweis auf Einwirkung von äußerer Gewalt
	05	- frei -
	06	Kriegsbeschädigten-Leiden / BVG-Leiden
	07	Notfall

[1] Quelle: GKV - Spitzenverband `http://www.gkv-datenaustausch.de/leistungserbringer/krankenhaeuser/krankenhaeuser.jsp`

Schlüssel 5: Entlassungs-/Verlegungsgrund

1.u. 2. Stelle	01	Behandlung regulär beendet
	02	Behandlung regulär beendet, nachstationäre Behandlung vorgesehen
	03	Behandlung aus sonstigen Gründen beendet
	04	Behandlung gegen ärztlichen Rat beendet
	05	Zuständigkeitswechsel des Kostenträgers
	06	Verlegung in ein anderes Krankenhaus
	07	Tod
	08	Verlegung in ein anderes Krankenhaus im Rahmen einer Zusammenarbeit (§ 14 Abs. 5 Satz 2 BPflV in der am 31.12.2003 geltenden Fassung)
	09	Entlassung in eine Rehabilitationseinrichtung
	10	Entlassung in eine Pflegeeinrichtung
	11	Entlassung in ein Hospiz
	12	interne Verlegung
	13	externe Verlegung zur psychiatrischen Behandlung
	14	Behandlung aus sonstigen Gründen beendet, nachstationäre Behandlung vorgesehen
	15	Behandlung gegen ärztlichen Rat beendet, nachstationäre Behandlung vorgesehen
	16	externe Verlegung mit Rückverlegung oder Wechsel zwischen den Entgeltbereichen der DRG-Fallpauschalen, nach der BPflV oder für besondere Einrichtungen nach § 17b Abs.1 Satz 15 KHG mit Rückverlegung
	17	interne Verlegung mit Wechsel zwischen den Entgeltbereichen der DRG-Fallpauschalen, nach der BPflV oder für besondere Einrichtungen nach § 17b Abs.1 Satz 15 KHG
	18	Rückverlegung
	19	Entlassung vor Wiederaufnahme mit Neueinstufung
	20	Entlassung vor Wiederaufnahme mit Neueinstufung wegen Komplikation
	21	Entlassung oder Verlegung mit nachfolgender Wiederaufnahme
	22	Fallabschluss (interne Verlegung) bei Wechsel zwischen voll- und teilstationärer Behandlung
	23	Beginn eines externen Aufenthalts mit Abwesenheit über Mitternacht (BPflV-Bereich – für verlegende Fachabteilung)
	24	Beendigung eines externen Aufenthalts mit Abwesenheit über Mitternacht (BPflV-Bereich – für Pseudofachabteilung 0003)
	25	Entlassung zum Jahresende bei Aufnahme im Vorjahr (für Zwecke der Abrechnung – PEPP*)
3. Stelle	1	arbeitsfähig entlassen
	2	arbeitsunfähig entlassen
	9	keine Angabe

A.4. Verwendete Abfragen

A.4.1. SQL-Scripte

- Abfrage zur Eingrenzung der Unfälle nach Aufnahme- und Entlassunggrund

```
SELECT KHD.[PERSONEN_ID]
                                    ,KHD.[FALL_ID]
                                    ,KHD.[IK]
                                    ,KHD.[DRG]
                                    ,KHD.[LEISTUNG_VON]
                                    ,KHD.[LEISTUNG_BIS]
                                    ,KHD.[VERWEILDAUER]
                                    ,KHD.[AUFNAHMEGRUND_301]
                                    ,KHD.[ENTLASSUNGSGRUND_301]
                                    ,KHD.[FACHABTEILUNG_301]
                                    ,KHD.[ZAHLBETRAG]
                                    ,KH_ICD.[DIAGNOSEART]
                                    ,KH_ICD.[ICD_PRIMAER]
INTO [falldaten_unfaelle]
FROM
          (SELECT * FROM [Krankenhaus_Fall]
          WERE (AUFNAHMEGRUND_301 LIKE '__03%' OR
                    AUFNAHMEGRUND_301 LIKE '__02%' OR
                    AUFNAHMEGRUND_301 LIKE '__07%')
          AND (ENTLASSUNGSGRUND_301 NOT LIKE '07%' AND
                    ENTLASSUNGSGRUND_301 NOT LIKE '11%'))
          AS KHD
INNER JOIN [Fall_Diagnose] AS KH_ICD
ON KHD.FALL_ID =KH_ICD.FALL_ID AND KHD.IK = KH_ICD.IK;
```

A.4.2. R-Scripte

- Abfrage zur Eingrenzung der Unfälle nach Aufnahme- und Entlassunggrund

```
## Test der Parameter 2010
data <- read.csv(file="UnfallData2010_final.csv", sep=';', dec=',')
t.test(data$ZAHLBETRAG, mu=922.851)

data <- read.csv(file="vers2010Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "W" & data$Alter>0 & data$Alter<26,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.002720257,data$Gewicht)

data <- read.csv(file="vers2010Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "W" & data$Alter>25 & data$Alter<66,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.000584902,data$Gewicht)

data <- read.csv(file="vers2010Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "W" & data$Alter>65,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.002820436,data$Gewicht)

data <- read.csv(file="vers2010Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "M" & data$Alter>0 & data$Alter<26,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.003829016,data$Gewicht)

data <- read.csv(file="vers2010Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "M" & data$Alter>25 & data$Alter<66,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.000972688,data$Gewicht)

data <- read.csv(file="vers2010Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "M" & data$Alter>65,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.001866967,data$Gewicht)

## Test der Parameter 2011
data <- read.csv(file="UnfallData2011_final.csv", sep=';', dec=',')
t.test(data$ZAHLBETRAG, mu=922.851)

data <- read.csv(file="vers2011Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "W" & data$Alter>0 & data$Alter<26,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.002720257,data$Gewicht)

data <- read.csv(file="vers2011Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "W" & data$Alter>25 & data$Alter<66,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.000584902,data$Gewicht)

data <- read.csv(file="vers2011Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "W" & data$Alter>65,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.002820436,data$Gewicht)

data <- read.csv(file="vers2011Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "M" & data$Alter>0 & data$Alter<26,]
```

```
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.003829016,data$Gewicht)

data <- read.csv(file="vers2011Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "M" & data$Alter>25 & data$Alter<66,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.000972688,data$Gewicht)

data <- read.csv(file="vers2011Detail.csv", sep=';', dec='.')
data <- data[data$Geschlecht == "M" & data$Alter>65,]
wtd.t.test(data$AnzahlUnfaelle/data$Gewicht,0.001866967,data$Gewicht)
```

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Elke Warmuth. *Diskrete Simulation - Mathematische Modelle diskreter stochastischer Systeme*, ZFH, 1997.

[2] J.A. Rosanow. *Stochastische Prozesse*, Akademie-Verlag, 1975.

[3] Konrad Königsberger. *Analysis 2*, Springer-Verlag, 2004.

[4] Stefan Tappe. *Einführung in die Wahrscheinlichkeitstheorie*, Springer-Verlag, 2013.

[5] John G. Kemeny, J. Laurie Snell. *Finite Markov Chains*, Springer-Verlag, 1976.

[6] R.R. Barlow, F. Proschan. *Statistische Theorie der Zuverlässigkeit*, Akademie-Verlag, 1981.

[7] Wolfgang König, `http://www.wias-berlin.de/people/koenig/www/Skripte.html`, 16.01.2015.

[8] Björn Degenkolbe, Elena Gette, Thomas Höpfner, Walter Warmuth. *Auf Leben und Tod- Spezifische Implikationen eines vermeintlich längeren Lebens für die Versicherungswirtschaft(1)* in: *Zeitschrift für Versicherungswesen*, 23/2011, S. 819-822.

[9] Björn Degenkolbe, Elena Gette, Thomas Höpfner, Walter Warmuth. *Auf Leben und Tod- Spezifische Implikationen eines vermeintlich längeren Lebens für die Versicherungswirtschaft(2)* in: *Zeitschrift für Versicherungswesen*, 24/2011, S. 853-856.

[10] Björn Degenkolbe, Elena Gette, Thomas Höpfner, Walter Warmuth. *Demographischer Wandel in Deutschland- Legenden und Mythen* in: B. Mühlbauer, D. Matusiewicz, F.Kellerhoff *Zukunftsperspektiven der Gesundheitswirtschaft*; Schriftenreihe *Gesundheitsökonomie: Management und Politik*, LIT Verlag, 2011, S.382 ff..

[11] Bundesministerium für Gesundheit, `http://bundesgesundheitsministerium.de`, (13.2.2015).

[12] Statistisches Bundesamt, `http://www.destatis.de`, (13.2.2015).

[13] Institut für das Entgeltsystem im Krankenhaus, `www.g-drg.de`, (13.2.2015).

[14] Deutsches Institut für Medizinische Dokumentation und Information, `www.dimdi.de`, (13.2.2015).

[15] Universität Ulm, `http://www.mathematik.uni-ulm.de/stochastik/lehre/ss05/wt/skript/node15.html`, (13.2.2015).

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Leipzig, den 16.02.2015

_____

Tobias Riedel