Universität Leipzig
Fakultät für Mathematik und Informatik
Mathematisches Institut

Diplomarbeit zum Thema:

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

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

Leipzig, im Februar 2014

Betreuer: HD Dr. Walter Warmuth
Betreuender Hochschullehrer: Prof. Dr. Manfred Riedel

Kurzzusammenfassung

PLATZHALTER

Danksagung

PLATZHALTER

PLATZHALTER2

Inhaltsverzeichnis

1. Einführung in das Thema

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

Im deutschen Gesundheitssystem entstehen jährlich Kosten in Höhe von 294 Mrd. Euro[1]. Um diese Summe abzudecken, existieren in Deutschland zwei Systeme: Die gesetzliche Krankenversicherung (GKV) und die private Krankenversicherung (PKV). Jeder Versicherte fällt aufgrund der Krankenversicherungspflicht in eines der beiden Systeme. In der GKV erfolgt die Finanzierung seit 2009 aus dem Gesundheitsfond. Jeder gesetzlich Versicherte zahlt monatlich einen Teil seines Arbeitseinkommens (z.B. Gehalt bzw. Rente) in diesen Fonds ein. Andere Einkommensquellen, wie zum Beispiel Mieteinnahmen, werden dabei nicht berücksichtigt. Dieses Geld wird anschließend an die verschiedenen 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 Beitrag jährlich, auf Basis des prognostizierten Behandlungsbedarfs, neu berechnet. Der Großteil der Bevölkerung (69,86 Mio.[2]) ist gesetzlich krankenversichert.

In Deutschland waren im Jahr 2011 fast neun Millionen Menschen (siehe Tabelle 1) privat krankenversichert. Im Gegensatz zur GKV handelt es beim PKV-System um eine Individualversicherung. Beim Versicherungseintritt wird für jeden Versicherten eine individuelle Prämie berechnet, welche die Kosten des Versicherten bis zu seinem Tod abdecken sollen. Diese wird auf Basis des Erwartungswerts der künftigen Leistungsausgaben (unternehmensinterne oder bundesweite Statistiken), der erwarteten Sterblichkeit, des erwarteten Kündigungsverhaltens und des Rechnungszinses bestimmt. Versicherte mit dem gleichen Tarif[3] und aus derselben Altersgruppe bilden dabei ein Kollektiv, wobei individuelle Ge-

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 1.1.: Endgültige Werte für das Geschäftsjahr 2011, Stand: November 2012

[1]siehe: Statistisches Bundesamt Zahlen in GKV für 2011 (www.destatis.de).
[2]siehe: http://bundesgesundheitsministerium.de Zahlen aus der KM1 für November 2012.
[3]Der Eintrittszeitpunkt ist dabei entscheidend. Aufgrund sich ändernder Rahmenbedingungen wer-

sundheitsmerkmale durch Risikoaufschläge berücksichtigt werden.

Die Prämie ist so kalkuliert, dass in den ersten Jahren voraussichtlich weniger Leistungen in Anspruch genommen werden als an Prämien bezahlt wird. Diese Ansparphase hält an bis die Leistungen die Prämieneinnahmen übersteigen. Dann wird das so gesparte Geld (die Alterungsrückstellung) dazu verwendet, die steigenden Kosten zu decken. Dieser Zeitraum wird als 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 Kostensteigerungen im Gesundheitswesen, oder ein Ansteigen des durchschnittlichen Lebensalters. Dabei ist zu berücksichtigen, dass gewisse Annahmen über die Entwicklung solcher Faktoren bereits bei der Kalkulation eines Tarif getroffen werden und damit eine Erhöhung nur dann zulässig ist, wenn die tatsächlichen Anstiege die prognostizierten Werte noch übertreffen.

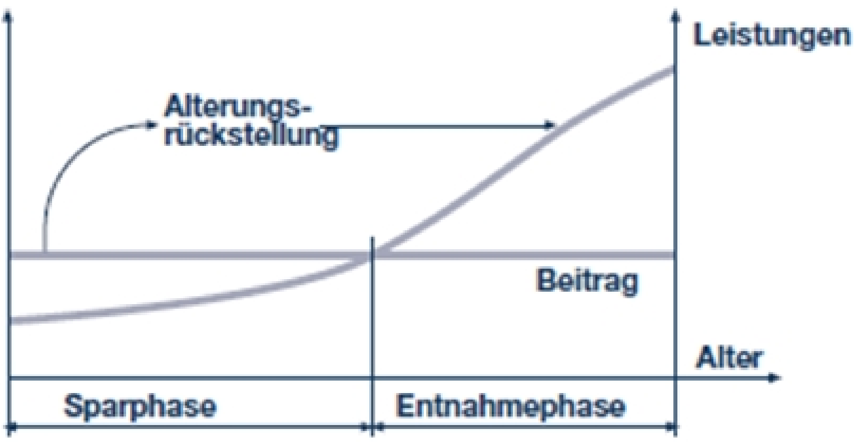


Abbildung 1.1.: Grafik Alterungsrückstellung

Seit dem 1. Januar 2009 sind die privaten Krankenversicherer dazu verpflichtet, im Fall eines Krankenkassenwechsels, die angesparte Alterungsrückstellung des Versicherten 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 er 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

den die einzelnen Tarife regelmäßig neu berechnet und dadurch entstehen verschiedene Tarifgenerationen. In einem Kollektiv werden nur Versicherte aus einer Tarifgeneration zusammengefasst.

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 im Gegenteil überhaupt keine nennenswerten Leistungen verursachen. Das Kollektiv, welches die Versicherten einer Tarifgeneration bilden, ist normalerweise dafür gedacht, 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.

Die Höhe des Schadens ist dabei schwer zu ermitteln. Man benötigt ein Modell, das auf Basis des aktuellen Gesundheitszustands eines Versicherten eine Prognose erstellt, wie viele Leistungen er bis zu seinem Tod noch verursachen wird. Ein solches Modell wird im nächsten Kapitel vorgestellt.

1.2. Kohorten Modell

Bei den Gesundheitsforen Leipzig entsteht zurzeit ein Modell zur Bestimmung der Höhe der noch zu erwartenden Leistungskosten eines Versicherten, basierend auf seinen Morbiditätsinformationen. Mit diesem Modell wäre es möglich, für einen Versicherten zum Zeitpunkt des Wechsels seine restliche Risikolast[5] zu berechnen und damit zu entscheiden, wie dieser das Kollektiv beeinflusst. Die nachfolgende Beschreibung des Modells stützt sich dabei in erster Linie auf den in der „Zeitschrift für Versicherungswesen“ (Heft 23/2011 und 24/2011) erschienenen Artikel „Auf Leben und Tod- Spezifische Implikationen eines vermeintlich längeren Lebens für die Versicherungswirtschaft“[6].

Das Ziel der Modellierung war es, möglichst wenige und möglichst gleichartige "Bausteineßu finden, aus denen sich die vielen Krankenversicherungsbiographien[7] 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“

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

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

[6] Quelle: Verweis auf Literaturverzeichnis.

[7] Gesundheitskosten in Zusammenhang mit den Leistungszeitpunkten im Verlauf des Lebens werden als Krankenversicherungsbiographie bezeichnet.

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

Die Abgrenzung der Lebensabschnitte erfolgt dabei vom Rand der Krankenversicherungsbiographie her (Geburtsdatum, Sterbedatum) jeweils Tag genau. Die Längen dieser Abschnitte sind möglichst kurz gewählt und in ganzen Jahren angegeben. Die „Phase des regen Lebens“ ist dabei in der Regel die längste Phase und wird von zwei Arten von Ereignissen überlagert:

- Ereignis: Isolierte Ereignisse
- Ereignis: Singuläre Ereignisketten

Die restliche Risikolast ergibt sich aus einer Mischung von Kohorten[8] der Restbiographien. Eine Restbiographie bezeichnet dabei den Anteil einer Versichertenbiographie, der noch nicht eingetreten ist, das heißt, 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:

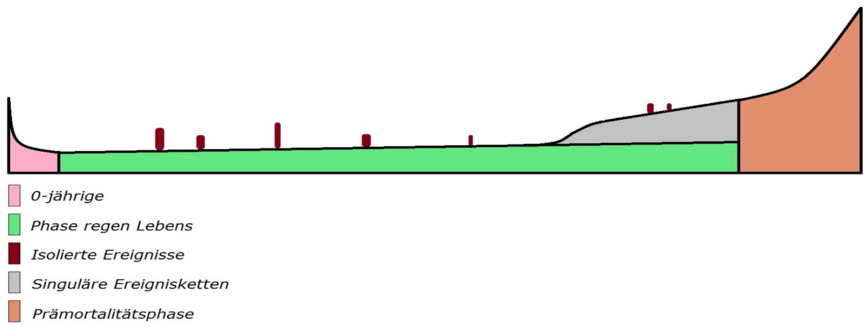


Abbildung 1.2.: Grafik Kohorten-Modell

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

1.2.1. 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 Entfernung vom Tod, zunächst stark und später abgeschwächt.

[8] „Kohorten sind Jahrgänge oder Gruppen von Jahrgängen, die der Abgrenzung von Bevölkerungsgruppen dienen. Sie sind durch ein zeitlich gemeinsames, längerfristig prägendes Startereignis definiert.“ (Quelle: Wikipedia) In diesem Fall ist das Kohorten Merkmal eine gleiche restliche Lebenserwartung.

1.2.2. 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.

1.2.3. 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.

1.2.4. Ereignis: Isolierte Ereignisse

Seltene Ereignisse, welche die Gesundheit eines Versicherten beeinträchtigen, sich nicht ankündigen und bei denen der Eintrittszeitpunkt der eigentliche Auslöser von 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.

1.2.5. Ereignis: Singuläre Ereignisketten

Eine nachhaltige, in der Regel bis zum Tod des Versicherten andauernde, beträchtliche Verschlechterung der Gesundheit wird nachfolgend als singuläre Ereigniskette bezeichnet. Es werden auch singuläre Ereignisketten berücksichtigt, die mit keinen Kosten für die Krankenversicherung verbunden sind (z. B. pflegebedürftig, da dann die Pflegeversicherung greift).

Das Ziel dieser Arbeit soll es sein, ein passendes Modell für die isolierten Ereignisse und die singulären Ereignisketten zu erstellen.

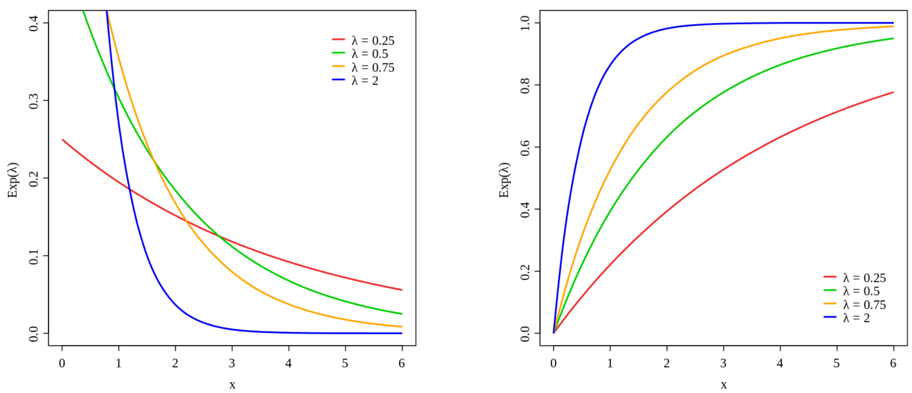
2. Fachliche Einführung

In diesem Kapitel sollen die in der Arbeit verwendeten mathematischen Hilfsmittel vorgestellt werden und die wichtigsten Eigenschaften bewiesen werden. Dabei wird vorausgesetzt das der Leser bereit Grundlegende Kenntnisse im Bereich der Wahrscheinlichkeitstheorie besitzt und Begriffe wie Wahrscheinlichkeitsraum, Zufallsvariable, Dichte- und Verteilungsfunktion bekannt sind. Im Folgenden sei stets der Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, \mathbb{P})$ gegeben.

Zunächst werden wir die Vereitlungen vorstellen, die in dieser Arbeit verwendet werden, und die wichtigsten Eigenschaften ableiten.

Definition 2.1. *Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ heißt* ***exponentialverteilt zum Parameter*** *λ (kurz: $\sim exp(\lambda)$) wenn sie die folgende Dichtefunktion besitzt:*

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



(a) Dichtefunktion

(b) Verteilungsfunktion

Abbildung 2.1.: Dichte- und Verteilungsfunktion der Exponentialverteilung

Die Exponentialverteilung hat folgende Eigenschaften:

(i) Für die Verteilungsfunktion der Exponentialverteilung gilt:

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

(ii) Der Erwartungswert ist gegeben durch:

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

(iii) Die Varianz ist gegeben durch:

$$Var(X) = \int_\infty^0 \left(x - \frac{1}{\lambda} \right)^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2}$$

(iv) Die Exponentialverteilung ist gedächtnislos (auch Nichtalterungseigenschaft genannt), d.h.:

$$\begin{aligned} F(x|t) = \mathbb{P}(x+t \leq X, t < X) &= \frac{\mathbb{P}(t < X < x+t)}{\mathbb{P}(t < X)} \\ &= \frac{\mathbb{P}(X \leq x+t) - \mathbb{P}(t \leq X)}{1 - \mathbb{P}(t \leq X)} \\ &= \frac{1 - e^{-\lambda(x+t)} - (1 - e^{-\lambda t})}{1 - (1 - e^{-\lambda t})} \\ &= \frac{-e^{-\lambda x - \lambda t} + e^{-\lambda t}}{e^{-\lambda t}} \\ &= \frac{e^{-\lambda t}(1 - e^{-\lambda x})}{e^{-\lambda t}} \\ &= 1 - e^{-\lambda x} = \mathbb{P}(x \leq X) = F(x) \end{aligned}$$

Lemma 2.2. *Wenn die Überlebenswahrscheinlichkeit ($\overline{F}(x) := 1 - F(x)$) einer nicht ausgearteten, nicht negativen Zufallsvariable X der folgenden Beziehung genügt:*

$$\mathbb{P}(x+t \leq X) = \mathbb{P}(x \leq X)\mathbb{P}(t \leq X)$$

dann gilt $X \sim exp(\lambda)$ für ein $\lambda > 0$.

Beweis: Mit $f(x) = \mathbb{P}(x < X)$ ergibt sich folgende Gleichung

$$f(x+t) = f(x)f(t), \text{mit} f(0) = \mathbb{P}(0 \leq X) = 1$$

die sich wie folgt umformen lässt:

$$ln(f(x+t)) = ln(f(x)) = ln(f(t))$$

Unter der Voraussetzung, das die Dichte $p_X(t) = -f(t+0), t \geq 0$ existiert, ergibt sich durch Differentiation nach t:

$$\frac{f'(x+t)}{f(x+t)} = \frac{f'(t)}{f(t)}$$

Für $t = 0$ folgt damit:

$$\frac{f^{'}(x)}{f(x)} = f^{'}(t)$$

Da $f(0) = 1$ der maximale Wert der Funktion f(x) ist, gilt $0 \geq f^{'}(0) := -\lambda$ und damit erhalten wir die bekannte Differentialgleichung:

$$\begin{aligned}
&f^{'}(x) = -\lambda f(x) \\
\Rightarrow\ &f(x) = e^{-\lambda x} \\
\Rightarrow &F(x) = 1 - f(x) = 1 - e^{-\lambda x} \\
\Rightarrow \quad &X \sim exp(\lambda)
\end{aligned}$$

□

Definition 2.3. *Eine diskrete Zufallsvariable $X : \Omega \rightarrow \mathbb{N}$ heißt **Poisson-verteilt zum Parameter** $\lambda \in \mathbb{R}_{>0}$ (kurz $X \sim Poi(\lambda)$), wenn gilt:*

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

Die Poisson-Verteilung hat folgende Eigenschaften:

(i) Für die Verteilungsfunktion der Poisson-Verteilung gilt:

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

(ii) Der Erwartungswert ist gegeben durch:

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

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

(iii) Die Varianz ist gegeben durch:

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

(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 Summer von poissonverteilten Zufallsvariablen ist wieder poissonverteilt.

Die Poisson-Verteilung hat außerdem eine besondere Bedeutung, da sie unter den richtigen Voraussetzungen die Grenzverteilung der Binomialverteilung ist. Dieser Zusammenhang wird in folgendem Lemma verdeutlicht:

Lemma 2.4. *Die Poisson-Verteilung ist die Grenzverteilung der Binomial-Verteilung ($B_{n,p}(x)$ mit einer geringen Erfolgswahrscheinlichkeit $p \to 0$ und dafür einer hohen Anzahl an Versuchen $n \to \infty$, wenn gilt das $\lambda = np$ weder 0 noch ∞ ist.*

Beweis: Wir zeigen, dass der Grenzwert $n \to \infty$ einer binomialverteilten Zufallsvariable an der Stelle x, gegen den Wert einer poissonverteilten Zufallsvariablen an der Stelle x geht.

$$
\begin{aligned}
\lim_{x \to \infty} (P)(X = x) &= \lim_{x \to \infty} \frac{n!}{x!(n-x)!} \frac{\lambda^x}{n} \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{x \to \infty} \frac{\lambda^x}{x!} \frac{n(n-1)(n-2)...(n-x+1)}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!} * 1 * e^{-\lambda} * 1 \\
&= \frac{\lambda^x e^{-\lambda}}{x!}
\end{aligned}
$$

□

Definition 2.5. *Ein **Stochastischer Prozess** X ist eine Familie von Zufallsvariablen $\{X_t\}$ mit $X_t : \Omega \to \mathfrak{A}, t \in T$. Für die Indexmenge T gilt in der Regel $T \in \{\mathbb{R}_{\geq 0}, \mathbb{N}_0\}$. Das heißt X ist eine Abbildung*

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

sodass $X_t : \omega \mapsto X_t(\omega)$ für alle $t \in T$ eine messbare Abbildung ist.

(i) Ein stochastische Prozess heißt **zeitdiskret** wenn T abzählbar ist, z.B. $T = \mathbb{N}_0$. Ansonsten heißt er **zeitstetig**. Analog heißt ein Prozess mit diskreten Zustandsraum $\mathfrak{A}$ **wertdisktret** oder auch **Punktprozess**.

(ii) Ein stochastische Prozess heißt **stationär** wenn für alle $s > 0$, sowie $t_1, t_2, ..., t_k \in T$ und $x_1, x_2, ..., x_k \in \mathfrak{A}$ gilt:

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

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

(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.

(iv)

Definition 2.6. *Sei $T_1, T_2, ... : \omega \to [0, \infty)$ eine Folge von unabhängig und identisch verteilten Zufallsvariablen, dann ist $N := \{N_t, t \geq 0\}$ mit*

$$N_t = \sum_{k=1}^{\infty} \mathbb{I}(S_k \leq t) \text{ und } S_n = T_1 + ... + T_n$$

*ein stochastischer Prozess und wird als **Zähl- bzw. Erneuerungsprozess** bezeichnet.*

Prozesse dieser Art werden z.B. in der Systemtheorie eingesetzt um die Ausfälle einer Komponente in einem bestimmten Zeitraum zu zählen. Deshalb werden die T_n häufig als **Zwischenankunftszeiten** bezeichnet.

Definition 2.7. *Ein Zählprozess $\{N_t, t \geq 0\}$ mit exponentialverteilten Zwischenankunftszeiten $T_n \sim exp(\lambda)$ heißt **homogener Poisson-Prozess mit der Intensität** λ.*

In folgendem Theorem werden die wichtigsten Eigenschaften des Poisson deutlich:

Theorem 2.8. *Die folgenden Aussagen sind äquivalent*[1]*:*

(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 $\{N_t\}$ 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 $\{N_t\}$ hat unabhängige Zuwächse und ist stationär und es gilt für $h \downarrow 0$:

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

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

Beweis:

- $(i) \Rightarrow (ii)$: Aus (i) folgt, dass $S_n = \sum_{i=1}^n T_i$ eine Summe von n unabhängigen und zum Parameter λ exponentialverteilten Zufallsvariablen ist, d.h., $S_n \sim Erl(n, \lambda)$, wobei $Erl(n, \lambda)$ die Erlang-Verteilung mit den Parametern n und λ bezeichnet. Hieraus folgt $\mathbb{P}(N_t = 0) = \mathbb{P}(S_1 > t) = e^{-\lambda t}$ und damit folgt:

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

[1] siehe: `http://www.mathematik.uni-ulm.de/stochastik/lehre/ss05/wt/skript/node15.html`

Dies gilt für jedes $n \geq 1$, und damit folgt, dass $N_t \sim Poi(\lambda t)$, womit der erste.

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

$\square$

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

Lemma 2.9. *Die Überlagerung von zwei unabhängigen Poisson-Prozessen* $\{N_t^1, t \geq 0\}$ *und* $\{N_t^1, t \geq 0\}$ *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 $\{N_t^1, t \geq 0\}$ 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 $\lambda = \lambda_1 + \lambda_2$. $\square$

Als nächstes werden wir eine weitere wichtige Klasse an stochastischen Prozessen vorstellen, doch vorher müssen wir noch den Begriff der stochastischen Matrix einführen:

Definition 2.10. *Eine Matrix* $P = (p_{i,j})$ *heißt* ***stochastisch****, falls für alle* $i, j \in I$ *(Indexmenge) gilt* $p_{i,j} \in [0,1]$ *und* $\sum_{j \in I} p_{i,j} = 1$.

Definition 2.11. *Ein Stochastischer Prozess* $\{X_t, t \geq 0\}$ *besitzt die* ***Markoveigenschaft****, wenn für alle* $i_0, i_1, ..., i_{n+1} \in \mathfrak{A}$ *gilt:*

$$\mathbb{P}(X_{t+1} = i_{t+1} | X_t = i_t, X_{t-1} = i_{t-1}, ..., X_0 = i_0) = \mathbb{P}(X_{t+1} = i_{t+1} | X_t = i_t)$$

Ein solcher Prozess heißt ***Markovscher Prozess*** *oder* ***Markov-Kette****. Die* ***Startverteilung*** *der Markov-Kette ist definiert durch* $v(i) = \mathbb{P}(X_0 = i)$ *und die Wahrscheinlichkeiten* $\mathbb{P}(X_{t+1} = i_{t+1} | X_t = i_t) =: p_{i_t, i_{t+1}}$ *werden als* ***Übergangswahrscheinlichkeiten*** *bezeichnet. Die Matrix* $P = (p_{i,j})_{i,j \in \mathfrak{A}}$ *die sich aus den Übergangswahrscheinlichkeiten ergibt, ist eine stochastische Matrix und heißt* ***Übergangsmatrix****.*

Der nächste Zustand einer Markov-Kette hängt also immer nur von dem aktuellen Zustand ab. D.h. die Kette wird durch die Übergangswahrscheinlichkeiten charakterisiert.
TODO: Eigenschaften der Markov-Kette charakterisieren Pfade durch Potenzen der Übergangsmatrix Irreduzibilität Rekurrenz und Transienz Absorbierende Zustände Stationäre Verteilung Ergodizität

3. 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. Anschließend soll dieses Modell auf einen konkreten Datenbestand angewendet und überprüft werden. Dazu werden zwei Ansätze vorgestellt um ein passendes Risikomodell zu erstellen. Einleitend soll die Definition der Isolierten Ereignisse genauer analysiert werden, um die Anforderungen an das spätere Risikomodell abzuleiten:

3.1. 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 in der Regel **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 was bedeutet, dass **keine Chronizität** durch einen Unfall entstehen kann.[1]

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 keine besondere Betrachtung mehr benötigen, da sie bereits in der dritten Phase des Modells erfasst wurden.

Die Erfahrung zeigt, dass die Unfallwahrscheinlichkeit in der Regel sowohl vom Alter- als auch vom Geschlecht des Versicherten abhängt. Einen guten Beleg dafür liefert die Statistik über die „Unfallentwicklung auf deutschen Straßen 2012“ vom Statistischen Bundesamt[2]. Daraus geht hervor das besonders junge Erwachsene im Alter zwischen 18 und 24 Jahren in schwere Verkehrsunfälle mit Personenschäden verwickelt sind. Hinzukommt das in 86% aller Motorradunfälle in diesem Altersbereich der Verunglückte männlich war.

[1] Dieser Fall wird durch die singulären Ereignisketten im nächsten Kapitel abgedeckt.

[2] Quelle: Statistisches Bundesamt unter `https://www.destatis.de/DE/Publikationen/Thematisch/TransportVerkehr/Verkehrsunfaelle/PK_Unfallentwicklung.html`

Neben Alter und Geschlecht wird auch die Saison in bestimmten Fällen Einfluss auf die Unfallwahrscheinlichkeit haben. Dies spiegelt sich auch bei den Verkehrsunfällen wieder, denn im Zeitraum von Mai bis September gab es die meisten Unfälle mit Personenschäden, da es aufgrund der höheren Geschwindigkeiten bei schönen Wetter zu schwereren Verkehrsunfällen kam. Dabei ist zu beachten, dass die Einflüsse von Alter, Geschlecht und Saison auf die Unfallwahrscheinlichkeit auch vom Typ des Unfalls abhängen.

Zusammenfassend lässt sich sagen, dass die Unfallwahrscheinlichkeit unabhängig von bereits eingetretenen Unfällen ist, aber je nach Unfalltyp, vom Alter und Geschlecht des Versicherten abhängen und einer Saisonalität unterliegen kann. Darauf aufbauend werden nun zwei Modelle vorgestellt:

3.2. Modell - Poisson-Prozess

Für den ersten Modellansatz nehmen wir an, dass aufgrund des Charakters der Unfälle als seltene Ereignisse, die Genesungsdauer vernachlässigt werden kann. Das heißt die aus dem Unfall resultierenden Folgeleistung werden auf den Zeitpunkt des Unfalls konzentriert und es gibt keinen Zeitraum nach einem Unfall, in dem die Wahrscheinlichkeit für einen Weiteren angepasst wird. Laut Voraussetzung hängt die Unfallwahrscheinlichkeit vom Alter und Geschlecht der Versicherten, sowie von der Saison ab. Doch zunächst werden wir ein Modell entwickeln für den Fall das die Unfallwahrscheinlichkeit konstant ist. *Diese ist intuitiv definiert als die Anzahl der Unfälle pro Zeiteinheit.* Anschließend kann dieses Modell erweitert werden um die verschieden Abhängigkeiten zu berücksichtigen.

Zunächst werden wir grundlegende Bezeichnungen einführen:

(i) Die Unfallwahrscheinlichkeit wird mit λ bezeichnet.

(ii) Die Unfallzeitpunkte seien gegeben durch $t_0, t_1, t_2, t_3, ...$, wobei $t_0 = 0$ den Beginn der Beobachtung kennzeichnet.

(iii) Die Zeit zwischen den einzelnen Unfällen sei definiert als $\{T_k\}_{k\geq 0}$ mit $T_k := t_k - t_{k-1}$ und $T_0 := 0$.

(iv) Die Anzahl der Unfälle wird bezeichnet durch $N_t := max\{k : t_k \leq t\}$

Sei nun die Wahrscheinlichkeit für einen Unfall λ konstant. Dann gilt für die Wartezeiten $\{T_k\}_{k\geq 0}$, dass die Zeit bis zum nächsten Unfall unabhängig davon ist, wie viel Zeit bereits ohne Unfall verstrichen ist. Sei $F(x)$ die gemeinsame Verteilung der Wartezeiten, dann gilt also:

$$\begin{aligned} F(x|t) &= F(x) \\ => F(x+t) &= F(x)F(t) \end{aligned}$$

Damit folgt gemäß Lemma 2.2 das $F(x) \sim Exp(\lambda)$. Damit ist $N_\lambda = \{N_t\}_{t>0}$ gemäß Definition 2.7 ein Poisson-Prozess mit Parameter λ. Wie bereits am Anfang erwähnt, ist die Unfallwahrscheinlichkeit von verschieden Faktoren abhängen und dabei in erster Linie natürlich vom Unfalltyp. Deshalb zerlegen wir den Prozess in mehrere Teilprozesse:

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 $N_\lambda = \sum_{i=1}^{n} N_{\lambda_{U_i}}$. Das heißt der Prozess lässt sich beliebig in Unterprozesse zerlegen die jeweils einen Typ bzw. auch eine Familie von Typen beschreiben. Dadurch lässt sich das Modell beliebig verfeinern, falls dies in einem Bereich notwendig ist.

Da Unfälle sowohl vom Alter und Geschlecht, sowie von der aktuellen Jahreszeit abhängen können, reicht es nicht aus einen Homogenen Poisson-Prozess zu verwenden. Deshalb ist es notwendig das unfallspezifische λ als Funktion dieser drei Einflüsse zu berücksichtigen. Die Stärke der jeweiligen Einflussfaktoren hängt natürlich in erster Linie vom Unfalltyp ab. Als praktikable Lösung zur Bestimmung der einzelnen Funktionen bietet es sich deshalb an, λ als Treppenfunktion zu modellieren, bei der die einzelnen Stufen von der empirischen Verteilung der Unfallwahrscheinlichkeit abhängen. Das detaillierte Vorgehen bei der Bestimmung von λ wird im weiteren Verlauf dieses Kapitels anhand eines Beispiels erläutert.

Ein analoges Vorgehen (empirische Verteilung) würde sich auch bei der Modellierung der Kostenverteilung für die Unfälle anbieten, insbesondere wenn diese nur schlecht durch eine Verteilungsfunktion beschrieben werden können. Schwierigkeiten können entstehen, da bei Leistungskosten in der Krankenversicherung häufig zu beobachten ist, dass es viele Fälle mit niedrigen und durchschnittlichen Kosten gibt aber auch eine nicht zu vernachlässigende Anzahl an Hochkostenfällen (Verweis), die um ein vielfaches teurer sind. Aus diesem Grund ist es wichtig eine Verteilungsfunktion zu wählen, die zu 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[3] bezeichnet. Das optimale Vorgehen bei der Modellierung der Kostenverteilung hängt letztendlich vom Unfalltyp und den zur Verfügung stehenden Daten ab.
- Zusammenfassung

3.3. Modell - Markov-Kette

- Ansatz der den Prozess eines Unfalls mit den verschiedenen Phasen der Genesung gut abbildet - Grundannahme: kein Lerneffekt und Gleichverteilte Eintritts- bzw. Übergangswahrscheinlichkeiten - Ergodisch (nur bei Unfällen) => Startverteilung spielt auf längere Sicht keine Rolle - Übergangswahrscheinlichkeiten werden aus den Daten geschätzt (Mittelwert) - Zur Bestimmung der zu Erwartenden Anzahlen evtl. Simulation?

[3] z.B. Weibull-Verteilung mit Formparameter<1

Monte-Carlo-Methode - Verweildauer und Kosten evtl. aus DRG Katalog?

3.4. Anwendung und Test eines 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. Aus diesen werden dann die Parameter des Modells geschätzt und die Modellgüte bewertet.

3.4.1. Datengrundlage + Aufbereitung

Als Datengrundlage stehen Versicherten- und Krankenhausabrechnungsdaten zur Verfügung, die über einen Zeitraum von 4 Jahren erfasst wurden. Die Grunddaten wurden aufbereitet und auf drei Tabellen verteilt. Während der Aufbereitung wurden Duplikate entfernt, die Datenformate vereinheitlicht und fehlerhafte bzw. unvollständige Datensätze herausgefiltert. Das Resultat wird in folgendem Diagramm veranschaulicht:

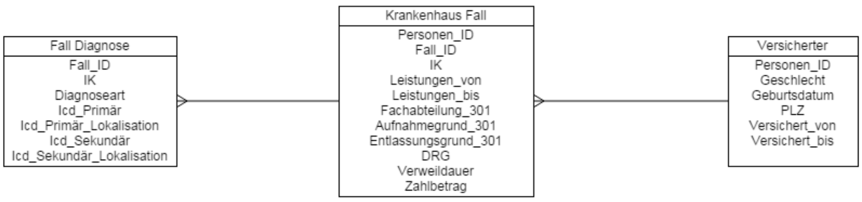


Abbildung 3.1.: ER-Diagramm Datengrundlage

Es ist zu sehen, das für jeden Versicherten das Geschlecht und Geburtsdatum hinterlegt sind. Außerdem gibt es Information über den Wohnort(Postleitzahl) und der Zeitraum in dem er versichert war. Letzteres ist wichtig, da logischerweise nur während dieses Zeitraums Daten über den entsprechenden Versicherten vorliegen. Jeder Versicherte kann über die Personen$_{I}DmehrerenKrankenhausfllenzugewiesensein.EinFallistdurchdieFall_{I}Dunddurchda$
- Anzahl Datensätze und Fälle - Tabellen Vorstellen o Beziehung (ER-Diagramm) o Schlüssel o Inhalt + Bedeutung der wichtigsten Spalten - ICD Codierung erläutern o Schwierigkeit mit Jahreswechsel und Änderungen in der Codierung - Datenaufbereitung o Zusammenführen der Tabellen o Datenfehler entfernen o Duplikate entfernen o Beschreibungstexte anspielen

3.4.2. Identifizierung von Unfällen

- Unfälle kündigen sich nicht an - Nur versicherte werden betrachtet die im Gesamtenzeitraum da waren - Einweisungsgrund + Entlassungsgrund - Aufenthaltsdauer - Nicht relevante Fälle herausfiltern - Liste der häufigsten Unfälle - Hochkostenfälle - Spezialfälle: Unfall mit Todesfolge o Wird vernachlässigt Versicherte mit diesen schweren Unfällen fallen in die letzte Phase (Walter fragen: Fälle gehören zwar zum Phänomen Unfall aber die entsprechenden Versicherten sind nicht die ganze Zeit da) - Spezialfälle: Unfall mit Folgeschäden o Pflegekosten => Singuläre Ereigniskette o Keine Folgekosten = Vernachlässigen

3.4.3. Anwendung und Test der Modelle

- Modell ableiten am Beispiel von gehirnerschüterrung - Saissonalität und Geschlechts-/Altersabhängigkeit testen und Parameter schätzen - Ergodizität voraussetzen - Eingrenzung der Fälle - Analyse der Daten o Häufigkeitsverteilung o Kostenverteilung o Zeitpunktverteilung o Altersverteilung o Hochkostenfälle - Anpassungstest an Poissonverteilung - Verknüpfung mit Kosten - Test des Modells

A. Anhang

A.1. 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, $\mathbb{P}(X = k)$
$\mathbb{I}$	Identische Abbildung

Abbildungsverzeichnis

Tabellenverzeichnis