

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, im Februar 2015

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

Betreuer:

Prof. Dr. Manfred Riedel
HD Dr. Walter Warmuth

Kurzzusammenfassung

PLATZHALTER

Danksagung

PLATZHALTER

PLATZHALTER2

Inhaltsverzeichnis

1. Einführung in das Thema	5
1.1. Mitnahme von Alterungsrückstellung in der PKV - Auswirkungen auf das Versichertenkollektiv	5
1.2. Kohorten Modell	7
1.2.1. Lebensabschnitt: „0-jährige“	9
1.2.2. Lebensabschnitt: „Prämortalitätsphase“	9
1.2.3. Lebensabschnitt: „Phase regen Lebens“	9
1.2.4. Ereignis: Isolierte Ereignisse	9
1.2.5. Ereignis: Singuläre Ereignisketten	9
2. Theoretische Grundlagen für die Modellierung	11
2.1. Verwendete Verteilungen	11
2.2. Poisson Prozess	17
2.3. Markov-Kette	26
3. Modellierung Isolierte Ereignisse	33
3.1. Grundlegende Betrachtungen	33
3.2. Modell - Poisson-Prozess	34
3.3. Alternativer Modellansatz - Markov-Kette	37
3.4. Anwendung und Test des Modells	39
3.4.1. Datengrundlage und Aufbereitung	40
3.4.2. Identifizierung von Unfällen	42
3.4.3. Anwendung und Test der Modelle	44
A. Anhang	46
A.1. Zitierte Erkenntnisse	47
A.2. Abkürzungsverzeichnis und Zeichenerklärungen	48
A.3. Codierung nach § 301 Abs. 3 SGB V	49
A.4. Verwendete Abfragen	51
A.4.1. SQL-Skripte	51

1. Einführung in das Thema

1.1. 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). Im Jahr 2011 sind zusammengekommen Kosten in Höhe von 294 Mrd. Euro¹ entstanden. Aufgrund der Krankenversicherungspflicht müssen alle Personen, mit Wohnsitz in Deutschland, in einem der beiden Systeme versichert sein. Für die GKV besteht deshalb eine Aufnahmespflicht und aus diesem Grunde ist der Großteil der Bevölkerung (69,86 Mio.²) 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 Gesundheitsfonds 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.

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 bei Versicherungseintritt für jeden Versicherten eine individuelle Prämie berechnet, welche die Kosten des Versicherten bis zu seinem Tod abdecken sollen³. Diese wird auf Basis des Erwartungswerts der künftigen Leistungsausgaben (unternehmensinterne oder bundesweite Statistiken) und der erwarteten

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

ten Sterblichkeit bestimmt. Das erwarteten Kündigungsverhalten und der Rechnungszins haben außerdem Einfluss auf die Höhe der Prämie. Versicherte mit dem gleichen Tarif⁴ und aus derselben Altersgruppe bilden dabei ein Kollektiv, wobei individuelle Gesundheitsmerkmale durch Risikoaufschläge berücksichtigt werden.

Die Prämie ist so kalkuliert, dass sie, trotz der im Alter erwartungsgemäß steigenden Leistungskosten, über die gesamte Versicherungsdauer konstant bleibt. Deshalb liegen in den ersten Jahren die Prämien über den erwarteten Leistungen. Aus dieser Differenz wird eine Rücklage (die Alterungsrückstellung) gebildet, die dazu verwendet wird, die im Alter steigenden Kosten zu decken. 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 Tarif getroffen werden und damit eine Erhöhung nur dann zulässig ist, wenn die tatsächlichen Anstiege die prognostizierten Werte noch übertreffen.

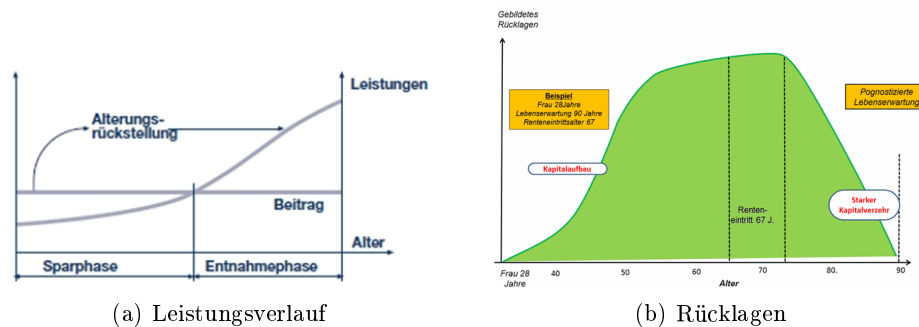


Abbildung 1.1.: Grafik Alterungsrückstellung

Seit dem 1. Januar 2009 sind die privaten Krankenversicherer 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 er in der Regel einen Teil seiner Alterungsrückstellung.

Ein Wechsel hat nicht nur Auswirkungen auf ein einzelnes Individuum, sondern betrifft

⁴Der 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.

auch das Kollektiv von Versicherten, welches der Einzelne verlässt. In dem Zeitraum vom Tarifaabschluss 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“⁵ war, entsteht dem Kollektiv damit ein Schaden, da die Überschüsse aus seiner Prämie nicht mehr dafür verwendet werden können, die Kosten morbidere 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

Die Gesundheitsforen Leipzig haben ein Modell zur Bestimmung der Höhe der noch zu erwartenden Leistungskosten eines Versicherten, basierend auf seinen Morbiditätsinformationen, entwickelt. Damit ist es möglich, für einen Versicherten zum Zeitpunkt des Kassenwechsels seine restliche Risikolast⁶ 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“⁷.

Das Ziel der Modellierung war es, möglichst wenige und möglichst gleichartige „Bausteine“ zu finden, aus denen sich die vielen Krankenversicherungsbiographien⁸ 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:

⁵Das heißt der Versicherte beansprucht weniger Leistungen als der Durchschnitt.

⁶Die 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.

⁷Quelle: Verweis auf Literaturverzeichnis.

⁸Gesundheitskosten in Zusammenhang mit den Leistungszeitpunkten im Verlauf des Lebens werden als Krankenversicherungsbiographie bezeichnet.

- Lebensabschnitt: „0-jährige“
- 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. Kommt es zu Überlagerungen durch einen frühen Tod (jünger als 6 Jahre), 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 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 Kohorten⁹ 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:

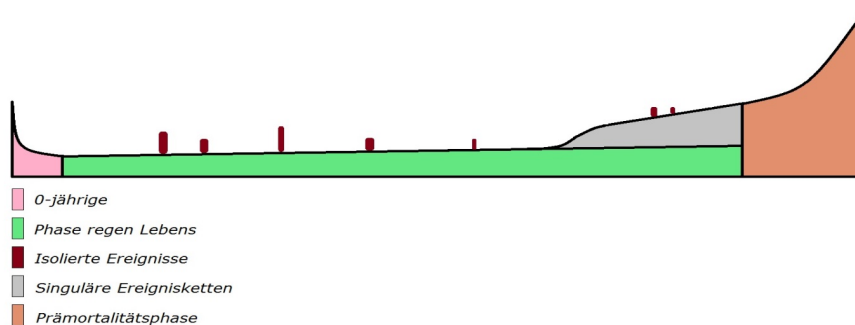


Abbildung 1.2.: Grafik Kohorten-Modell

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

⁹„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.

1.2.1. 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 restliche Lebenserwartung, zunächst stark und später abgeschwächt. Eine umfangreiche Beschreibung dieser Modellierung findet sich in folgendem Artikel: ...

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. Dieser Abschnitt ist auch das zentrale Thema des eingangs erwähnten Artikels.

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 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.

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 direkten Kosten für die Krankenversicherung verbunden sind (z. B. pflegebedürftig, da dann die Pflegeversicherung greift).

Das Ziel dieser Arbeit soll es sein, eine mögliche Ausgestaltung für den Teil der isolierten Ereignisse und der singulären Ereignisketten zu modellieren.

2. Theoretische Grundlagen für die Modellierung

In diesem Kapitel sollen die in der Arbeit verwendeten mathematischen Hilfsmittel vorgestellt werden 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 $X : \Omega \rightarrow \mathbb{R}$, die Abbildung $F_X : \mathbb{R} \rightarrow [0, 1]$ 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

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

die Schwanzfunktion von X .

2.1. Verwendete Verteilungen

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

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

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

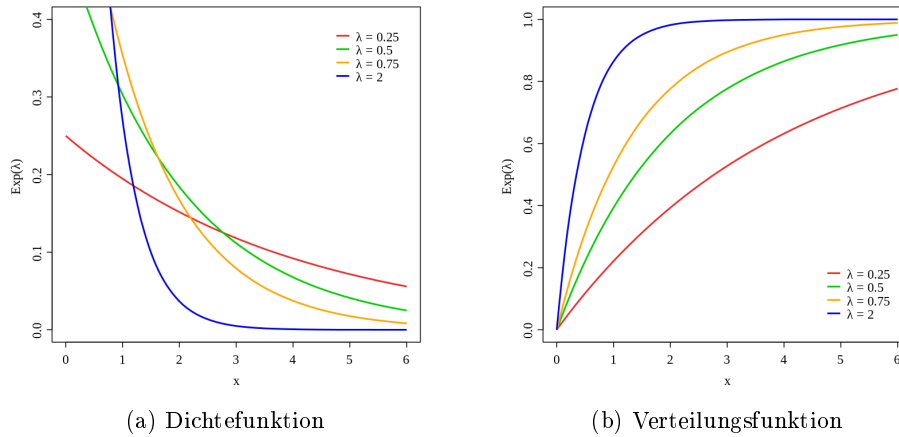


Abbildung 2.1.: Dichte- und Verteilungsfunktion der Exponentialverteilung

Folgerung 2.2. 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:

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

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

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

Die Exponentialverteilung hat die besondere Eigenschaft der Gedächtnislosigkeit und es lässt sich sogar zeigen, dass sie die einzige absolut stetige Verteilung¹ mit dieser Eigenschaft ist:

Lemma 2.3. Sei X eine absolut stetige positive Zufallsvariable, dann gilt $X \sim \exp(\lambda)$ genau dann wenn für alle $x, t > 0$ gilt, dass

$$\mathbb{P}(X > x + t \mid X > t) = \mathbb{P}(X > x) \quad (2.1)$$

Beweis: \Rightarrow Sei $X \sim \exp(\lambda)$ dann gilt:

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

\Leftarrow Sei umgekehrt X eine absolut stetige Zufallsvariable, die die Gleichung 2.1 erfüllt. Wir definieren $g(x) := \mathbb{P}(X > x)$. Für $x, y > 0$ gilt:

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

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

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

¹Im diskreten Fall ist dies die geometrische Verteilung.

und somit insbesondere auch $g(\frac{1}{n}) = (g(1))^{\frac{1}{n}}$. Da X nur positive Werte annimmt, 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

$$g(\frac{p}{q}) = g(\frac{1}{q})^p = g(1)^{\frac{p}{q}}$$

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

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

□

Definition 2.4. Eine diskrete Zufallsvariable $X : \Omega \rightarrow \mathbb{N}$ heißt **poissonverteilt zum Parameter** $\lambda \in \mathbb{R}_{>0}$ (kurz $X \sim \text{Poi}(\lambda)$), wenn gilt:

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

Folgerung 2.5. Sei $X \sim \text{Poi}(\lambda)$ dann gilt:

(i) Die Verteilungsfunktion der Poisson-Verteilung ist:

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

(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:

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

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

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

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

Das heißt die Summe 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 2.6. Sei $X \sim B_{n,p}(k)$ eine binomialverteilte Zufallsgröße. Wenn für $n \rightarrow \infty$ und $p \rightarrow 0$ gilt, dass der Erwartungswert np gegen eine von n unabhängige Konstante λ konvergiert, dann konvergiert die Verteilungsfunktion von X gegen die Verteilungsfunktion einer, zum Parameter λ , poissonverteilten Zufallsgröße.

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

$$\begin{aligned}
\lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} \mathbb{P}(X = k) &= \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} \underbrace{\left(\frac{n(n-1)(n-2)\dots(n-k+1)}{n^k}\right)}_{\rightarrow 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\rightarrow e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\rightarrow 1} \\
&= \frac{\lambda^k e^{-\lambda}}{k!}
\end{aligned}$$

□

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 2.7. Eine Zufallsvariable $X : \Omega \rightarrow \mathbb{R}$ 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 $\Gamma : (0, \infty) \rightarrow (0, \infty)$ die Gamma-Funktion ist:

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

mit $r > 0$.

Folgerung 2.8. Die Exponentialverteilung ist ein Spezialfall der Gamma-Verteilung: Sei $X \sim \gamma(\alpha, r)$ dann gilt für $r = 1$ gilt $X \sim \exp(\alpha)$. Da für $x > 0$ gilt:

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

und das ist die Dichte einer zum Parameter α exponentialverteilten Zufallsgröße.

Lemma 2.9. 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:

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

Der zweite Teil der Aussage folgt direkt aus dieser Beziehung. \square

2.2. Poisson Prozess

Ein weiteres Hilfsmittel, dass zur Modellierung verwendet wird, sind Stochastische Prozesse. Diese eignen sich sehr gut dazu geordnete zufällige Vorgänge zu beschreiben.

Definition 2.10. 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 $X_t : \Omega \rightarrow Z$. Für die Indexmenge T gilt in der Regel $T \in \{\mathbb{R}_{\geq 0}, \mathbb{N}_0\}$. Das heißt X ist eine Abbildung

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

sodass $X_t : \omega \mapsto X_t(\omega)$ für alle $t \in T$ eine messbare Abbildung ist. Z heißt dann die **Zustandsmenge** und (Z, \mathcal{Z}) der **Zustandsraum**.

- (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 Z **wertdiskret** oder auch **Punktprozess**.
- (ii) Ein stochastische Prozess heißt **stationär**, wenn für alle $s > 0$, sowie $t_1, t_2, \dots, t_k \in T$ und $x_1, x_2, \dots, x_k \in Z$ gilt:

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

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

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

Definition 2.11. Sei $T_1, T_2, \dots : \omega \rightarrow [0, \infty)$ eine Folge von unabhängig und identisch verteilten Zufallsvariablen und $S_n := T_1 + \dots + T_n$ für alle $n \in \mathbb{N}$, dann ist $N := \{N_t, t \geq 0\}$ mit

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

ein stochastischer Prozess und wird als **Zählprozess** bezeichnet.

Prozesse dieser Art werden z.B. in der Zuverlässigkeitstheorie eingesetzt um die Ausfälle einer Komponente in einem bestimmten Zeitraum zu zählen. Deshalb werden die T_n häufig als **Zwischenankunftszeiten** bezeichnet und die S_n als n -te **Sprungzeit**.

Definition 2.12. 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 2.13. Seien X_1, X_2, \dots, X_n , mit $n \in \mathbb{N}$ Zufallsvariablen, dann bezeichnen wir die geordneten Variablen $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ als die **Ordnungsstatistik** der Variablen $\{X_i : 1 \leq i \leq n\}$.

Satz 2.14. Seien X_1, X_2, \dots, X_n , mit $n \in \mathbb{N}$, unabhängige und identisch verteilte Zufallsvariablen mit Dichte f und sei $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ die entsprechende Ordnungsstatistik. Dann gilt für die gemeinsame Dichte:

$$f_{X_{1:n}, X_{2:n}, \dots, X_{n:n}}(t_1, t_2, \dots, t_n) = \begin{cases} n! f(t_1) f(t_2) \dots f(t_n) & \text{falls } t_1 \leq t_2 \leq \dots \leq t_n \\ 0 & \text{sonst} \end{cases} \quad (2.2)$$

Beweis: Da aufgrund der Definition der Ordnungsstatistik die Werte aufsteigend geordnet sind, ist die Dichte gleich 0, wenn die Bedingung $t_1 \leq t_2 \leq \dots \leq t_n$ nicht erfüllt ist. Sei nun diese Bedingung erfüllt. Dann existieren genau $n!$ Möglichkeiten die Zufallsvariablen X_1, X_2, \dots, 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

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

da die Zufallsvariablen X_1, X_2, \dots, X_n unabhängig voneinander sind. Nach Multiplikation mit $n!$ erhält man 2.2. \square

Theorem 2.15. *Die folgenden Aussagen sind äquivalent²:*

- (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 $n = 1, 2, \dots$ der Zufallsvektor (S_1, S_2, \dots, 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 $n = 1, 2, \dots$ der Zufallsvektor (S_1, S_2, \dots, 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 Zuwächse und ist stationär und es gilt für $h \rightarrow 0$:

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

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

Beweis: Da der Beweis sehr umfangreich ist, werden an dieser Stelle nur die wesentliche Schritte vorgestellt.

- (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 2.9 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:

²Quelle siehe: <http://www.mathematik.uni-ulm.de/stochastik/lehre/ss05/wt/skript/node15.html>

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

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 $f_{S_1, \dots, S_{n+1}}(t_1, \dots, t_{n+1})$ von S_1, \dots, S_{n+1} . Für beliebige $t_0 = 0 \leq t_1 \leq \dots \leq t_n \leq t_{n+1}$ gilt aufgrund des Transformationssatzes (vgl. A.1)

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

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 $0 \leq t_1 \leq \dots \leq t_n \leq t$ für die gemeinsame bedingte Dichte

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

und $f_{S_1, \dots, S_n}(t_1, \dots, t_n | N_t = n) = 0$ sonst. Nach 2.2 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 $\mathbb{E}N_1 = \lambda$. Seien $x_1, \dots, x_n \in \mathbb{N}$ und $t_0 = 0 \leq t_1 \leq \dots \leq t_n$, dann gilt für $x = x_1 + \dots + x_n$

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

und damit hat der Zählprozess N_t unabhängige Zuwächse.

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

Aus (iii) folgt, dass der Zufallsvektor (S_1, \dots, 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 $x_1, \dots, x_n \in \mathbb{N}$, $t_0 = 0 \leq t_1 \leq \dots \leq t_n$ und $h > 0$

$$\mathbb{P}\left(\bigcap_{k=1}^n \{N_{t_k+h} - N_{t_{k-1}+h} = x_k\} \mid N_{t_n} + h = m\right) = \mathbb{P}\left(\bigcap_{k=1}^n \{N_{t_k} - N_{t_{k-1}} = x_k\} \mid N_{t_n} + h = m\right)$$

Aufgrund der Formel der totalen Wahrscheinlichkeit folgt damit, dass N_t stationär ist. Die Gleichverteilungseigenschaft aus (iii) liefert außerdem für $0 < h < 1$

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

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 $k = 1, 2, \dots$ 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 Summe und Grenzwert, ergibt sich

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

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

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

was äquivalent zur zweiten Bedingung in (iv) ist.

- (iv) \Rightarrow (v):

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

$$p_0(t+h) = \mathbb{P}(N_t = 0, N_{t+h} - N_t = 0) \quad (2.3)$$

$$= \mathbb{P}(N_t = 0) \mathbb{P}(N_{t+h} - N_t = 0) \quad (2.4)$$

$$= \mathbb{P}(N_t = 0) \mathbb{P}(N_h = 0) \quad (2.5)$$

$$= p_0(t)(1 - \lambda h + o(h)) \quad (2.6)$$

und für $t \geq h > 0$

$$p_0(t) = p_0(t-h)(1 - \lambda h + o(h)) \quad (2.7)$$

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 2.3 und 2.7, dass für beliebige $h \geq -t$ gilt

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

Das zeigt das $p_0(t)$ differenzierbar ist und es ergibt sich für $t > 0$ folgende Differentialgleichung

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

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 $b_0 = 0 \leq a_1 < b_1 \leq \dots \leq a_n < b_n$, dann gilt

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

Nach (v) ist der stochastische Prozess stationär und die $N_t \sim Poi(\lambda t)$ 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

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

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

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

Dies bedeutet, dass die Zufallsvariablen $S_1, S_2 - S_1, \dots, S_n - S_{n-1}$ 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 2.16. *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 vorangegangenen Theorem haben wir gezeigt, dass für einen Poisson-Prozess $\{N_t, t \geq 0\}$ gilt $N_t \sim \text{Poi}(\lambda t)$. Da $\{N_t^1, t \geq 0\}$ und $\{N_t^2, t \geq 0\}$ unabhängig sind, gilt für die Summe $N_t^1 + N_t^2 \sim \text{Poi}((\lambda_1 + \lambda_2)t)$. 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 sich bei einem Poisson-Prozess die Zeit bis zum nächsten Ereignis, zu einem beliebigen Zeitpunkt, betrachtet.

Definition 2.17. *Sei $\{N_t, t \geq 0\}$ ein Zählprozess mit Zwischenankunftszeiten T_1, T_2, \dots und Sprungzeiten $\{S_n, n \in \mathbb{N}\}$. Dann bezeichnet W_t 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 2.18. 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:

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

D.h. 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. \square

Nachdem wir nun ausführlich den einfachen Poisson-Prozess betrachtet haben, wollen wir nun einen stochastischen Prozess einführen, der häufig in der Risikotheorie für die Modellierung von Schadenshöhen verwendet wird.

Definition 2.19. Sei $\{N_t, t \geq 0\}$ ein homogener Poisson-Prozess mit Intensität λ und seien X_1, X_2, \dots 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 2.20. 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

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

□

2.3. 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“³.

Definition 2.21. Ein Stochastischer Prozess $\{X_t, t > 0\}$ aus I -wertigen Zufallsvariablen, besitzt die **Markoveigenschaft**, wenn für $t_0, t_1, \dots, t_n, t_{n+1} \in T$ und alle $i_0, i_1, \dots, i_{n+1} \in I$ gilt:

$$\mathbb{P}(X_{t_{n+1}} = i_{n+1} \mid X_{t_n} = i_n, X_{t_{n-1}} = i_{n-1}, \dots, X_{t_0} = i_0) = \mathbb{P}(X_{t_{n+1}} = i_{n+1} \mid X_{t_n} = i_n) \quad (2.8)$$

Ein diskreter stochastischer Prozess der 2.8 erfüllt heißt **Markov-Kette**. Die **Startverteilung** der Markov-Kette ist definiert durch $v(i) := \mathbb{P}(X_0 = i)$ und die Wahrscheinlichkeiten $\mathbb{P}(X_{t_{n+1}} = i_{n+1} \mid X_{t_n} = i_n) =: p_{i_n, i_{n+1}}$ 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. D.h. die Kette wird durch die Übergangswahrscheinlichkeiten charakterisiert. Deshalb hat die Übergangsmatrix auch eine besondere Struktur. Im folgenden betrachten wir lediglich den stetigen Fall, d.h die Zustandsmenge I ist eine nichtleere, endliche oder höchstens abzählbar unendliche Menge und die Indexmenge T ist eine Teilmenge von \mathbb{N} .

Definition 2.22. Eine Matrix $P = (p_{i,j})$ heißt **stochastisch**, falls für alle $i, j \in I$ (Indexmenge) gilt $p_{i,j} \in [0, 1]$ und $\sum_{j \in I} p_{i,j} = 1$.

³Aktuelle Version unter <http://www.wias-berlin.de/people/koenig/www/Skripte.html> (Stand: 16.01.2015)

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 wird.

Lemma 2.23. *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 $\{X_t, n \in \mathbb{N}\}$ genau dann eine Markov-Kette mit Übergangsmatrix P und Startverteilung v , wenn für alle $n \in \mathbb{N}$ und alle $i_0, i_1, \dots, i_n \in I$ gilt*

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = v(i_0)p_{i_0, i_1}p_{i_1, i_2} \dots p_{i_{n-1}, i_n} \quad (2.9)$$

Beweis: Der Beweis dass die Gleichung 2.9 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:

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

Die andere Richtung folgt aus der Definition der bedingten Wahrscheinlichkeit:

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

□

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 2.24. Sei $\{X_t, t \in \mathbb{N}\}$ 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 $p_{i,j}^{(n)}$ wird als **n -stufigen Übergangswahrscheinlichkeiten** bezeichnet. Das heißt die Wahrscheinlichkeit dafür, dass 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 diese Aussage beweisen können benötigen wir noch einen kleinen Satz zu den Übergangswahrscheinlichkeiten:

Satz 2.25. (Chapman-Kolmogorov-Gleichung) $\{X_t, t \in \mathbb{N}\}$ eine Markov-Kette dann gilt für alle $m, n \in \mathbb{N}$ und alle $i, j \in I$

$$p_{i,j}^{(m+n)} = \sum_{k \in I} p_{i,k}^{(m)} * p_{k,j}^{(n)}. \quad (2.10)$$

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 2.24

Sei $P^{(n)} := \{p_{i,j}^{(n)}\}_{i,j \in I}$ die Matrix der n -stufigen Übergangswahrscheinlichkeiten. Dann gilt aufgrund von 2.10

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

Damit folgt dann

$$P^{(n)} = P^{(1+n-1)} = P * P^{(n-1)} = P^2 * P^{(n-2)} = \dots = P^n.$$

□

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

Definition 2.26. Im folgenden sei immer eine Markov-Kette $\{X_t, t \geq 0\}$ mit I -wertigen Zufallsgrößen und einer Übergangsmatrix P gegeben. Außerdem seien $i, j \in I$ beliebige Zustand 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 $T_{i,j} := \min\{n \in \mathbb{N} : X_n = j \mid X_0 = i\}$ die Wartezeit bis die Markovkette vom Zustand i aus, das erste Mal den Zustand j erreicht. Ein Zustand i heißt **rekurrent** falls $\mathbb{P}(T_i < \infty) = 1$, ansonsten heißt er **transient**. D.h. ein rekurrenter Zustand wird also mit Sicherheit in endlicher Zeit erneut erreicht.
- (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 $j \in I$. Analog 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:

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

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

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

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.

Die Gleichgewichtsverteilung kann nicht immer explizit angegeben werden, aber in einigen Sonderfällen ist dies möglich.

Satz 2.27. *Sei $\{X_t, t \geq 0\}$ 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, dass die Gleichgewichtsverteilung v eindeutig bestimmt ist und gilt $v(i) = \frac{1}{\mu_i}$.

Beweis:

- (iii) \Rightarrow (ii): Diese Richtung ist trivial.
- (ii) \Rightarrow (i): Wir definieren ein Maß, das für eine, 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:

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

Durch Anwendung des Satzes der monotonen Konvergenz⁴ und ausnutzen der

⁴siehe Anhang

Markov-Eigenschaft, können wir zeigen, dass dies ist ein invariantes Maß⁵:

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

Nun zeigen wir, dass γ auch zu einer Verteilung normiert werden kann und damit dann eine Gleichgewichtsverteilung ist.

$$\begin{aligned}
\sum_{j \in I} \gamma_k(j) &= \sum_{j \in I} \mathbb{E} \left(\sum_{n=1}^{T_{k,k}} \mathbb{1}_{\{X_n=j\}} \right) \\
&= \mathbb{E} \left(\sum_{n=1}^{T_{k,k}} \underbrace{\sum_{j \in I} \mathbb{1}_{\{X_n=j\}}}_1 \right) \\
&= \mathbb{E}(T_{k,k}) = \mu_k < \infty
\end{aligned}$$

- (i) \Rightarrow (iii): Sei v eine Gleichgewichtsverteilung und sei $k \in I$, dann ist $\gamma = \frac{v}{v(k)}$ ein invariantes Maß mit $\gamma(k) = 1$ und es gilt $\gamma(k) = \gamma^6$. Damit folgt

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

⁵d.h. $\gamma P = \gamma$

⁶Vergleiche Script König Satz 9.5.3

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 2.28. *Sei $\{X_t, t \geq 0\}$ eine irreduzible Markov-Kette mit aperiodischen und positiv rekurrenten Zuständen und Gleichgewichtsverteilung π . Dann gilt für alle $i, j \in I$*

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

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

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

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. 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 Krankenkassen Abrechnungsdaten werden dann die Modellparameter für einen konkreten Anwendungsfall geschätzt und die Modellgüte bewertet.

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 **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. 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 keine besondere Betrachtung mehr benötigen, 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 anhängen wird, ob ein Zusammenhang besteht und wie stark dieser Zusammenhang 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 sagen, dass die zu modellierende Unfallwahrscheinlichkeit unabhängig von bereits eingetretenen Unfällen ist, aber je nach Unfalltyp, vom Alter

und Geschlecht des Versicherten abhängen und einer Saisonalität unterliegen kann. Für die Modellierung werden wir die Saisonalität allerdings nicht besonders 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:

3.2. Modell - Poisson-Prozess

Die erwartenden Kosten für Unfälle in der Phase regen Lebens, setzt 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 nehmen wir an, dass aufgrund des Charakters der Unfälle als seltene Ereignisse, die Genesungsdauer vernachlässigt werden kann. Das heißt die begrenzte Leistungsfolge, die bei jedem 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 $t_0, t_1, t_2, t_3, \dots$, 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 kumulierten Zwischenankunftszeiten werden mit $S_n = \sum_{k=0}^n T_k$ bezeichnet.
- (v) Die Anzahl der Unfälle wird bezeichnet durch

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

Man sieht leicht, dass $N_t, t \geq 0$ der gesuchte stochastische Prozess ist und gemäß Definition 2.11 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 $\{T_k\}_{k \geq 0}$, dass die Zeit bis zum nächsten Unfall unabhängig davon ist, wie viel Zeit s bereits ohne 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 2.3, $T_k \sim \exp(\lambda)$ für alle $k \geq 0$. Damit ist, innerhalb eines Kalenderjahres, $N_t, t \geq 0$, laut Definition 2.12, ein homogener Poisson-Prozess und $N_t \sim \text{Poi}(\lambda)$. Da ein Poisson-Prozess, gemäß Theorem 2.15 unabhängige Zuwächse besitzt und stationär ist, passt dieser auch zu den Eingangs aufgestellten Modellannahmen. An dieser Stelle sollte auch erwähnt werden, dass aufgrund von Lemma 2.18 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 stationär ist, 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 $\lambda_{\text{Altersgruppe, Geschlecht}}$ geschätzt werden. Dazu betrachten wir die relative Häufigkeit eines Unfalls für die jeweilige Kombination aus Alter und Geschlecht und setzen

$$\lambda_{\text{Altersgruppe, Geschlecht}} := \frac{\#\text{Unfall}_{\text{Altersgruppe, Geschlecht}}}{\#\text{Versicherte}_{\text{Altersgruppe, Geschlecht}}}. \quad (3.1)$$

Dabei ist $\#\text{Versicherte}_{\text{Altersgruppe, Geschlecht}}$ 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 $\#\text{Unfall}_{\text{Altersgruppe, Geschlecht}}$ die Anzahl der Unfälle aller 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 diese sich hinsichtlich ihrer Kosten sehr stark unterscheiden können. Das Lemma 2.16 erlaubt es uns, den allgemeinen Unfall Prozess als eine Überlagerung von Teilprozesse darzustellen. Seien U_1, U_2, \dots, U_n mit $n \in \mathbb{N}$ die verschiedenen relevanten Unfalltypen und $\lambda_{U_1}, \lambda_{U_2}, \dots, \lambda_{U_n}$ die zugehörigen Unfallwahrscheinlichkeiten. Dann gilt für die daraus

resultierenden Prozesse

$$\begin{aligned} N_t^{\lambda_{U_i}} &\sim \text{poi}(\lambda_{U_i}) \\ N_t^\lambda &:= \sum_{i=1}^n N_t^{\lambda_{U_i}} \\ N_t^\lambda &\sim \text{poi}\left(\sum_{i=1}^n \lambda_{U_i}\right) \end{aligned}$$

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, \dots , 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), 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 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-Verteilung¹ 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 Veränderungen, durch Preisveränderungen oder die Einführung alternativer Behandlungsmethoden, unterliegen. Es ist kann versucht werden, dies durch jährliche Änderungsraten auszugleichen, trotzdem hat man in der Regel nicht genug Information 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

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

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

Für einen Versicherten ist das Risiko also gleich dem Barwerten der vorhergesagten Kosten, in der Phase regen Lebens.

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.

3.3. Alternativer Modellansatz - Markov-Kette

Ein alternativer Modellansatz für die isolierten Ereignisse, kann mit Hilfe von Markov-Ketten konstruiert werden. Im Gegensatz zum Poisson-Prozess 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 3.1 beschriebenen Grundannahmen. Wie bereits in Abschnitt 2.3 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 bieten sich dabei ein Tag an, da Informationen zum Unfallzeitpunkt und zur Genesungsdauer in der Regel nur Tag genau vorliegen. Die folgende Grafik veranschaulicht die möglichen Übergänge.

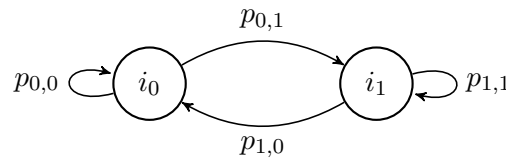


Abbildung 3.1.: Einfaches Markov-Modell

Die Übergangswahrscheinlichkeiten sind dabei wie folgt definiert. 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 man wieder die mittlere Genesungsdauer t_{Genesung} als Grundlage nehmen und damit definieren, dass und $p_{1,1} := 1/t_{\text{Genesung}}$ und $p_{1,0} = 1 - p_{1,1}$. In diesem Fall ist die Verweildauer im Zustand i_1 im stetigen Fall geometrisch verteilt und exponentialverteilt im kontinuierlichen Fall.

An dieser Stelle wird die Schwäche dieses einfachen Ansatzes deutlich. In einer Markov-Kette sind die Verweilzeit immer geometrisch- bzw. 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 sauber abbilden zu können, ist es notwendig neue Zustände einzuführen. Deshalb wird für jeden Genesungstag ein eigener Zustand angelegt, der modellieren, ob die Behandlung abgeschlossen ist oder einen weiteren Tag dauert. Die daraus resultierenden Übergänge werden in der folgenden Grafik veranschaulicht.

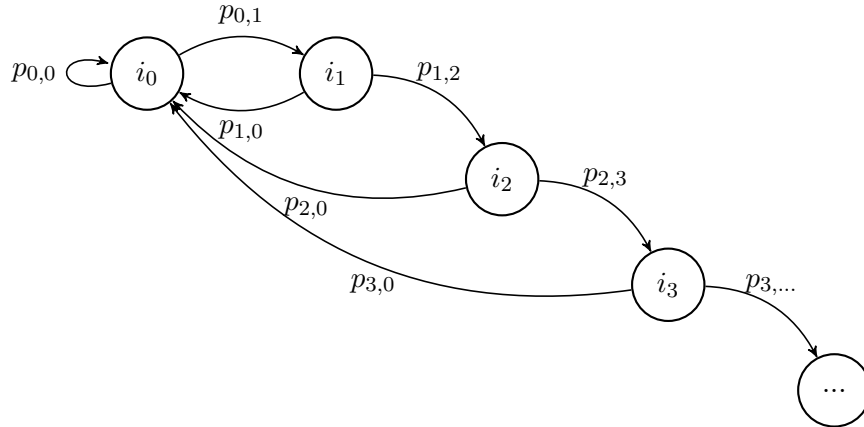


Abbildung 3.2.: Erweitertes Markov-Modell

Dieses Konzept ließe sich sogar noch erweitern, um verschiedene Genesungspfade zu modellieren. Zum Beispiel eine notwendige Reha des Patienten oder eine ambulante Betreuung. Es muss allerdings beachtet werden, dass je feiner das Modell aufgegliedert wird, es immer schwieriger wird für jeden Pfad eine ausreichende Menge an Daten zu finden um die Parameter zu schätzen. Wenn die Datenbasis zu klein wird, ist die Gefahr groß durch Extremfälle das Modell zu verzerren.

Bisher haben wir nur einen Unfalltyp betrachtet. Weitere Unfalltypen können als neue Zustände, i_{U_n} für $n \in \mathbb{N}$, in das bestehende 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 möglich 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 modelliert 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 eine neuer Unfall passieren kann. Um dies zu vermeiden, müssten für jeden Tag der Genesung Übergangswahrscheinlichkeiten zu den verschiedenen Unfalltypen modelliert werden. Dadurch würde sich die Komplexität des Modells und der Modellierungsaufwand sehr stark erhöhen. Allerdings sind Unfälle, wie Eingangs beschrieben, seltene Ereignisse und damit kann der Fall, dass zwei Unfälle direkt hintereinander eintreten, vernachlässigt werden, da der daraus resultierende Fehler entsprechend klein ist.

Die nächste notwendige Erweiterung dieses Modellansatzes ist der Wechsel von konstanten Übergangswahrscheinlichkeiten zu Alters- und Geschlechtsabhängigen. Wie bereits beim Poisson-Modell ist die praktikabelste Lösung die Wahrscheinlichkeiten als Treppenfunktion auf Basis der relativen Häufigkeit zu modellieren. Dabei ist es möglich nur die Unfallwahrscheinlichkeiten $\{p_{0,U_n}\}_{n \in \mathbb{N}}$ als Funktion von Alter und Geschlecht darzustellen. Dies hängt natürlich immer vom jeweiligen Unfalltyp ab.

Die Kosten die für die Behandlung eines Unfalls entstehen, treten in diesem Modellansatz immer dann auf, wenn ein Patient aus dem Krankenhaus entlassen wird. Für die Modellierung der Kostenfunktion gelten die selben 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 benötigt hat. Dadurch lässt sich die Streuung der einzelnen Kostenverteilungen reduzieren.

Dieser Modellansatz erzeugt sehr komplexe Modelle, die sich am Besten mit einem Simulationsansatz auswerten lassen. Dafür eignet sich das Monte-Carlo-Verfahren sehr gut, da auf diese Weise empirische Kostenverteilungen erzeugt werden können, die sich dann einfach weiterverarbeiten lassen. Außerdem sollte erwähnt werden, dass unabhängig von der Startverteilung die Markov-Kette nach ausreichend vielen Schritten immer wieder in den Ausgangszustand i_0 zurückkehrt. Es gibt keine absorbierenden Zustände, da Tod und Pflegefälle ausgeschlossen wurden. Außerdem ist jeder erreichbare Zustand rekurrent und kann somit immer, in endlich Zeit, erneut erreicht werden.

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 und 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 Variablen die geschätzt werden müssen. Dadurch ist eine effiziente Auswertung nur mit Hilfe eines Simulationsansatzes möglich. Im weiteren Verlauf dieses Kapitels werden wir uns deshalb auf den Poisson-Ansatz konzentrieren.

3.4. 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. Aus diesen werden dann im letzten Teil die Parameter des Modells geschätzt und die Modellgüte bewertet. In diesem Abschnitt werden nur stationäre Fälle betrachtet, also Fälle die zu einer Aufnahme im Krankenhaus geführt haben.

3.4.1. Datengrundlage und Aufbereitung

Die Datengrundlage für die Anwendung des Modells besteht aus Versicherten- und Krankenhausabrechnungsdaten, die über einen Zeitraum von 4 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 abgerechnete DRG's und die gestellten Diagnosen (ICD's). Deshalb wird, bevor die Datengrundlage näher betrachtet, im Folgenden ein kurzer Einblick in das deutsche Abrechnungssystem im Krankenhaus gegeben werden, um diese Abkürzungen zu erklären.

Seit 2003 wird in Deutschland die Vergütung im Krankenhausbereich über ein Fallpauschalensystem realisiert. Dieses basiert auf den sogenannten Diagnosis Related Groups (DRG) oder auf deutsch diagnosebezogene Fallgruppen. 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 vorgenommenen Prozeduren. Komplikationen oder erschwerende Begleiterkrankungen (Komorbiditäten) werden über verschiedene Schweregrade berücksichtigt. Für jede DRG gibt es eine Fallpauschale. Abhängig von der Verweildauer können zu dieser Pauschale noch Zu- bzw. Abschläge hinzukommen. Der endgültige Betrag ergibt dann die Kosten für einen Behandlungsfall. Das System wird vom „Institut für das Entgeltsystem im Krankenhaus“² (kurz InEK) gepflegt und weiterentwickelt.

Die während eines Krankenhausaufenthaltes 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 ICD³). Der zugehörige Katalog wird vom „deutschen Institut für Medizinische Dokumentation und Information“⁴, kurz DIMDI, jährlich aktualisiert und angepasst. Er ist hierarchisch strukturiert und enthält 22 Krankheitskapitel, die sich in Gruppen, Katego-

²www.g-drg.de

³Aufgrund des englischen Namens: „International Statistical Classification of Diseases and Related Health Problems“

⁴www.dimdi.de

rien 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 Operationen- und Prozedurenschlüssel (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 Krankenhaufalles gestellten Diagnosen erfasst werden. Das heißt es können auch Diagnosen vorkommen, die mit der DRG nicht in Zusammenhang stehen⁵. Die Datenstruktur der aufbereiteten Daten wird in folgendem Diagramm veranschaulicht:

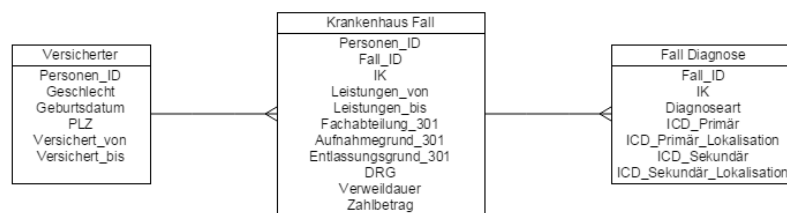


Abbildung 3.3.: ER-Diagramm Datengrundlage

- (i) **Versicherter:** In der ersten Tabelle sind Information 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. Leider besitzt jeder Versicherte mehrere Einträge in dieser Tabelle, da für jede Änderungsmeldung eine neue Zeile angelegt wurde.
- (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 Einsatz⁶. Für jeden Behandlungsfall gibt es Daten zum Zeitpunkt und Grund der Aufnahme und Analog für die Entlassung. Außerdem wird die abgerechnete DRG und die damit verbundenen Kosten angegeben.

⁵ z.B. ein Herzinfarkt während einer Operation

⁶Das ist notwendig da jedes Krankenhaus die Fall_ID selbst vergibt und es so vorkommt das zwei Krankenhäuser die selbe 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 beschriebenen 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. Das ist z.B. notwendig um zu spezifizieren, ob der linke oder der rechte Arm gebrochen ist. In einigen wenigen Fällen gibt es auch noch eine Sekundäre ICD.

3.4.2. 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 beschriebenen Methodik erfasst werden können, trotzdem bilanziell im Kohorten Modell erfasst sind. Diese werden dann der Phase regen Lebens zugeordnet und somit gehen keine Kosten verloren.

Das wichtigste Erkennungsmerkmal der Unfälle ist, dass sie sich nicht ankündigen. Deshalb ist der Aufnahmegrund im Krankenhaus ein guter Indikator dafür, ob der Patient mit Überweisung oder Termin aufgenommen wurde oder nicht. In den Daten ist der Aufnahmegrund entsprechend den gesetzlichen Vorgaben codiert⁷. Die für uns relevanten Information sind in der 3. und 4. Stelle codiert. Die folgenden 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 ⁸)	0,001%
	<i>Keine Angabe (Feld nicht gefüllt)</i>	4,598%

Tabelle 3.1.: 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) behandelt oder aufgrund eines Gewaltverbrechens(04). Für uns interessant sind also die Notfälle und die Fälle die tatsächlich als Unfall aufgenommen werden (02 und 03). Letztere machen leider nur einen sehr kleinen

⁷ siehe Anhang A.3

Teil aus, was auf die Natur der Daten zurückzuführen ist. Diese Schlüssel werden nur vergeben, wenn es 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 wurde⁹. 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 3.2.: 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 Patient stirbt oder muss in ein Hospiz verlegt werden. Diese Fälle werden in der Prä mortalitätsphase des Kohorten-Modells erfasst und sind somit für die Singulären Ereignisse, zumindest kostentechnisch, nicht relevant. Allerdings kann die Tatsache, dass ein entsprechender Unfall passiert ist, bei der Modellierung der alters und geschlechtsabhängigen Unfallwahrscheinlichkeiten verwendet werden. 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 einem Kassenwechsel des Patient 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.

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. Die Hauptdiagnose ist in der Regel der Grund für die Hospitalisierung. Es kommt vor, dass während der Behandlung andere Erkrankung entdeckt werden und dadurch die Hauptdiagnose am Ende für die DRG und damit für die Kosten,

⁹ siehe Anhang A.3

kaum eine Rolle spielt¹⁰.

Aufgrund der Komplexität der möglichen Krankenhausfälle ist es schwierig eine allgemeine Vorgehensweise zu beschreiben um weitere Unfälle einzugrenzen. Um geeignete Kandidaten für ein Unfallmodell zu finden, kann man allerdings die häufigsten oder teuersten Fälle betrachten, die nach den Filter des Aufnahme- und Entlassungsgrundes übrig bleiben. Der Vorteil des Kohorten-Modells ist es, dass unabhängig davon wie vollständig man die Unfalltypen modelliert keine Kosten verloren gehen. Außerdem lässt sich das Phänomen der isolierten Ereignisse jederzeit durch weitere Modelle genauer zu beschreiben, falls die erforderlich ist. Die restlichen Kosten werden alle in der Phase regen Lebens erfasst.

In den verwendeten Daten waren die häufigsten Unfälle Hypertonie, Herzinsuffizienz, Ohnmacht oder eine Gehirnerschütterung. Die konkreten DRGs und ICDs, die man einem Unfalltyp zuordnen möchte, lassen sich teilweise nur mit medizinischen Hintergrundwissen identifizieren. Aus diesem Grund wurde für die Erstellung der Modelle ein einfacher Fall ausgewählt: die Gehirnerschütterung.

3.4.3. Anwendung und Test der Modelle

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 wir eine Testmenge definieren, die später zur Bewertung des fertigen Modells verwendet wird.

Im vorangegangenen Abschnitt haben wir bereits erläutert, wie die Abrechnungsdaten nach Aufnahme und Entlassungsgrund gefiltert werden können, um die relevanten Fälle zu identifizieren. Die selben Filter wurden auch verwendet, um die Grundtabelle mit allen potentiellen Unfällen zu erstellen¹¹. Dazu wurde MS SQL Server als Datenbank verwendet und die benötigten Abfragen per SQL formuliert.

Datenselktion

SQL: Eingrenzung der Fälle Jahre - bevor parameter geschätzt werden können müssen unfalldaten eingegrenzt und aufbereitet werden. Für die Modellierung werden dabei nur Versicherte betrachtet, die über den gesamten Zeitraum bei der Kasse versichert waren und somit vollständige Datensätze besitzen Probleme bei der Eingrenzung

Datenzusammenführung und aggregation

Problem mit personenDaten, Java Lösung => Fertige Datenbasis

Schätzen der Modellparameter schätzen der parameter über excel allgemeine Unfallwahrscheinlichkeit Geschlechts-/Altersabhängigkeit testen und Parameter schätzen => Verteilung

Schätzden der Kostenverteilung evtl mit anpassungstest

¹⁰Z.B. wenn ein Patient wegen einer Ohnmacht eingeliefert wird, kommt es vor, dass am Ende das einsetzen eines Herzschrittmachers abgerechnet wird.

¹¹siehe Anhang A.4

Prognose auf Basis der Schätzwerte auf Basis der Sterbewahrscheinlichkeit

- anschließend müssen die für das Modell benötigten personendaten den unfällen zugewiesen werden -

- Selection der Daten.... - auswahl per sql - aufbereitung der personendaten per java -

- Hochkostenfälle ?
- Fälle von Versicherten die nicht die ganze Zeit da sind gehören zwar zum Phänomen Unfall aber die Kosten können nicht berücksichtigt werden
- Modell ableiten am Beispiel von gehirnerschütterung
- Saisonalität und Geschlechts-/Altersabhängigkeit testen und Parameter schätzen
- Ergodizität voraussetzen
- Eingrenzung der Fälle
- Analyse der Daten
- Häufigkeitsverteilung
- Kostenverteilung
- Zeitpunktverteilung
- Altersverteilung
- Hochkostenfälle
- Anpassungstest an Poissonverteilung
- Verknüpfung mit Kosten
- Test des Modells

A. Anhang

A.1. Zitierte Erkenntnisse

Satz A.1. (*Transformationssatz*)

Seien U und V offene Teilmengen des \mathbb{R}^n und sei weiterhin $T : U \rightarrow V$ ein Diffeomorphismus. Das ist die Funktion f auf V genau dann über V integrierbar, wenn die Funktion $(f \circ T)|\det DT|$ 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 $\det DT(x)$ die Funktionaldeterminante von T .
Vergleiche [Königsberger, 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

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

Das heißt Integration und Grenzwertbildung können vertauscht werden.
Vergleiche TODO:.

A.2. Abkürzungsverzeichnis und Zeichenerklärungen

\mathbb{N}	Natürliche Zahlen
\mathbb{N}_0	Natürliche Zahlen <i>inklusive</i> 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
$\text{Var}(X)$	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

Definition A.3. *Ein Verfahren, dass durch die wiederholte Durchführung eines Zufallsexperiments Aussagen über das Verhalten des Systems ableitet, wird als **Monte-Carlo-Verfahren** oder **Monte-Carlo-Simulation** bezeichnet. 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. Die Verteilung der Häufigkeiten wird dabei als **empirische Verteilung** bezeichnet.*

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

Im Folgenden wird ein Auszug aus der Anlage 2 zur § 301-Vereinbarung wiedergegeben¹

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

¹Quelle: 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-Skripte

- Abfrage zur Eingrenzung der Unfälle nach Aufnahme- und Entlassungsgrund

```
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 LIKE '01%' OR
          ENTLASSUNGSGRUND_301 LIKE '03%'))
    AS KHD
INNER JOIN [Fall_Diagnose] AS KH_ICD
ON KHD.FALL_ID =KH_ICD.FALL_ID AND KHD.IK = KH_ICD.IK ;
```

Abbildungsverzeichnis

1.1. Grafik Alterungsrückstellung	6
1.2. Grafik Kohorten-Modell	8
2.1. Dichte- und Verteilungsfunktion der Exponentialverteilung	12
3.1. Einfaches Markov-Modell	37
3.2. Erweitertes Markov-Modell	38
3.3. ER-Diagramm Datengrundlage	41

Tabellenverzeichnis

1.1. Endgültige Werte für das Geschäftsjahr 2011, Stand: November 2012 . . .	5
3.1. Aufnahmegrund und relative Häufigkeiten	42
3.2. Entlassungsgrund und relative Häufigkeiten für Not- und Unfälle	43

Literaturverzeichnis

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

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