Seminararbeit:

Nonparametric Estimation for Incomplete Obeservations - Kaplan and Meier, 1956

Autor: Maja Krajewska

Matrikelnummer: 10581592

Betreuer: Dr. Wiencierz

30. Mai 2014

Die Aufgabe der Teilnehmer des Bachelorseminars „Breakthroughs in Statistical Methodology“ war es, jeweils einen bedeutenden statistischen Artikel zu bearbeiten und vorzustellen. Ich habe mich für „Nonparametric Estimation from Incomplete Observations”von Kaplan und Meier entschieden, eine der meistziterten statistischen Arbeiten .
In diesem Artikel führen Kaplan und Meier den Product Limit Schätzer ein, der die Survival Funktion von Überlebensdaten schätzt. Der große Vorteil dieser Methode besteht darin, dass ebenfalls zensierte Daten in der Auswertung mit einbezogen werden können. Der Schätzer macht es möglich die Eintrittswahrscheinlichkeit von Ereignissen wie zum Beispiel der Erholung eines Patienten nach einer bestimmen Behandlung zu berechnen, auch wenn nicht alle Versuchspersonen bis zum festgelegten Endereignis an der Studie teilnehmen. In dieser Arbeit werde ich die Methode von Kaplan und Meier erklären, anhand von Beispielen zeigen wie Überlebenswahrscheinlichkeiten berechnet werden können und werde die wichtigsten Eigenschaften des Schätzers herleiten.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

1.1 Problemstellung

1.1.1 Beispiel zur Veranschaulichung der Problemstellung

In Versuchen, in denen die Individuen von einem bestimmten Ereignis bis zu beispielsweise dem Tod beobachtet werden, tritt oft das Problem auf, dass einige Versuchspersonen vor Eintritt des beendenden Ereignisses aus der Studie aussteigen. Als Beispiel betrachte ich den Datensatz von John Caplehorn et al. (1994), der die Überlebenszeiten von Heroinabhängigen von dem Eintritt in eine Klinik an bis zum Tod dokumentierte. Wie erwartet traten viele Patienten aus der Klinik und somit auch aus der Studie aus, ohne dass der Zeitpunkt des Todes bekannt ist. Lange Zeit konnten nur die unzensierten Aufzeichnungen statistisch verwertet werden, wodurch zusätzlich zu den verlorenen Daten auch ein Kostenproblem auftritt. In der vorliegenden Arbeit soll der Kaplan-Meier Schätzer vorgestellt werden, der ebenfalls die Verwertung von zensierten Daten in Lebensdaueranalysen ermöglicht.

1.1.2 Survival Analyse

Ein Versuch, bei dem über einen festen Zeitraum hinweg Individuen beobachtet werden bis hin zu einem bestimmten finalen Erignis, fällt eindeutig in den Bereich der Survival Analyse. Dies ist die Lehre der Verteilungen von Lebensdauern, den Zeiten ab einem beginnenden Ereignis an (beispielsweise dem Beginn einer Behandlung) bis zu einem finalen Ereignis (beispielsweise dem Tod). Eine charakterisierende Eigenschaft der Survival Daten ist das unumgehbare Auftreten unvollständiger Beobachtungen, insbesondere in dem bestimmten Fall, dass das beendende Ereignis für einige Individuen nicht beobachtet werden kann. Stattdessen ist lediglich bekannt, dass dieses Ereignis mindestens nach einem gegebenen Zeitpunkt eintritt.
Im Rahmen des Bachelorstudiums wird nicht auf dieses spezielle Themengebiet eingegangen, somit bin ich noch nicht mit dem Kaplan-Meier Schätzer oder anderen Methoden der Survival Analyse in Berührung gekommen.

1.2 Aufbau der Arbeit

Im Rahmen dieser Arbeit werde ich zuerst den Kaplan-Meier Schätzer definieren und berechenen, was zunächst im Grundprinzip und dann ausführlich und zuletzt anhand

zweier Beispiele geschehen wird. Anschließend wird die Varianz und das Mittel hergeleitet, darauf folgend werde ich besondere Eigenschaften des Schätzer erläutern und zuletzt auf die Kovarianz und die erwarteten Lebensdauern eingehen. Abschließend werden Kritikpunkte angesprochen, aber auch Zusammenhänge mit anderen Theorien hergeleitet und ein Überblick über die Verwendungsfelder in heutiger Zeit gegeben .

2 Der Kaplan Meier Schätzer

2.1 Begriffserklärung und Notation

Um den Kaplan-Meier Schätzer definieren zu können, muss man im Vorfeld den Begriff der nichtparametrischen Schätzung klären und einige Notationen sowie Konventionen einführen.

2.1.1 Nichtparametrische Schätzung

Die meist verwendete Methode der parametrischen Schätzung für Verteilungen von Überlebenszeiten ist wahrscheinlich die Anpassung der Normalverteilung an die Beobachtungen beziehungsweise an deren Logarithmen, wo hingegen eine nichtparametrische Schätzung der funktionalen Form des Schätzers erlaubt flexibel zu sein. Es werden also zur Herleitung des Schätzers keine Annahmen über die parametrische Klasse von Wahrscheinlichkeitsverteilungen der Daten getroffen. Eine wichtige Eigenschaft der nichtparametrischen Schätzer ist hierbei, dass im Falle eine Umformung der Zeitachse von t zu $t^* = f(t)$, wobei f eine streng monoton wachsende Funktion ist, die entsprechenden geschätzten Verteilungsfunktionen in folgender Beziehung stehen: $\widehat{F}^*(f(t)) = \widehat{F}(t)$ (Kaplan and Meier, 1958).

2.1.2 Notation und eingeführte Konventionen

Notation

Für die folgenden Arbeitsabschnitte werden folgende Funktionen definiert:

$P(t) = Pr(T > t)$
$\widehat{P}(t)$ = Kaplan-Meier Schätzer (PL) von $P(t)$
$n(t)$ = Anzahl der Individuen, die den Zeitpunkt t überleben
$N(t)$ = erwartete Anzahl der Individuen, die den Zeitpunkt t überleben $\mathbb{E}(n(t))$
$N^0(t)$ = Anzahl der Individuen, die die Observationsgrenze $\mathcal{L}$ besitzen, so dass $\mathcal{L} \leq t$

Konventionen

Kaplan und Meier führten zwei Konventionen ein, die die Grundlage für den Kaplan-Meier Schätzer bilden:
Tode, die zum Zeitpunkt t aufgezeichnet wurden, werden behandelt als ob sie kurz vor

dem Zeitpunkt t eingetreten wären. Verluste, die zum Zeitpunkt t aufgezeichnet wurden, werden hingegen behandelt als ob sie kurz nach dem Zeitpunk t stattgefunden hätten. Diese Konventionen haben mehrere Vorteile, denn somit sind *P(t)* und $\widehat{P}$*(t)* rechts-stetig, N^0*(t)* links-stetig und *n(t)* weder rechts- noch links-stetig. Weitere Vorteile sind dass einerseits die Tode behandelt werden als ob sie zum Zeitpunkt t stattgefunden haben obwohl sie in Wahrheit kurze Zeit früher stattfanden. Andererseits gilt *P(0)* = 1 genau dann, wenn kein Individuum zum Zeitpunkt des beginnenden Ereignisses stirbt.

2.2 Definition und Berechnung

2.2.1 Grundprinzip

Zur Berechnung des Kaplan-Meier Schätzers lässt sich folgendes Grundprinzip aufstellen:

1. Die Zeitachse wird in entsprechend gewählte Intervalle (0,u_1), (u_1,u_2), ... , (u_{j-1},u_j) unterteilt.

2. Für jeden Intervall (u_{j-1}, u_j) wird $p_j = P_j$ / P_{j-1} geschätzt. Dies ist die Proportion der lebenden Individuen kurz nach u_{j-1} , die den Zeitpunkt u_j überlebt haben.

3. Falls t der Trennungspunkt zweier Intervalle ist, wird die Proportion P(t) der Anzahl der Individuen, die den Zeitpunkt t überleben, von dem Produkt der geschätzten p_j für alle Intervalle vor t geschätzt.

2.2.2 Ausführliche Berechnung

Man erhält den Schätzer indem man die Intervalle in Schritt (1.) so wählt, dass die Schätzung in Schritt (2.) ein einfaches Binomial wird, ohne jegliche Annahmen über die Verteilung des Schätzers zu treffen. In jedem dieser Intervalle müssen die Tode und die Verluste auf bekannte Weise getrennt werden. Zur Vereinfachung kann anfangs angenommen werden, dass kein Intervall gleichzeitig sowohl einen Tod als auch einen Verlust beeinhaltet. Man bezeichnet nun die Anzahl der Individuen kurz nach u_{j-1} als n_j und die Anzahl der Tode, die im Intervall (u_{j-1},u_j) stattfanden, als δ_j . Somit ergibt sich als Formel für den Schätzer der Überlebenswahrscheinlichkeit in Zeitpunkt *t*:

$$\widehat{p_j} = (n_j - \delta_j)/n_j = n'_j/n_j \tag{2.1}$$

Wobei n'_j die Anzahl der Beobachtungen kurz nach den δ_j Toden ist. Falls der Intervall nur Verluste beinhaltet (aber mindestens ein Individuum den Intervall überlebt) ist der Schätzer $\widehat{p_j} = 1$.
Der Kaplan-Meier Schätzer für die Überlebenswahrscheinlichkeit bis zu dem Zeitpunkt *t*, *t* eingeschlossen, ergibt sich nun aus dem Produkt der Überlebenswahrscheinlichkeiten der Zeitpunkte t_0 bis t_j:

$$\widehat{p}(t) = \prod_{j=1}^{n} (n'_j / n_j) \tag{2.2}$$

wobei $n'_j = n_j - \delta_j$. Falls der letzte Zeitpunkt t^* einem Verlust entspricht, sollte die Formel (2.2) nicht mit einem $t > t^*$ benutzt werden. In diesem Fall wird angenommen, dass $\widehat{P}(t)$ zwischen 0 und $\widehat{P}(t^*)$ liegt, jedoch nicht näher definierbar ist. Falls der Eintritt neuer Individuen in die Stichprobe nach dem Beginn ihrer Lebenszeiten erlaubt werden soll, so behandelt man diese als negative Verluste.

Es wird angenommen, dass nichts über Individuen bekannt ist, die vor dem beginnenden Zeitpunkt gestorben sind (bzw. das finale Ereignis erreicht haben). Somit ist die Beobachtung rechts-zensiert, aber links-beschränkt.
Die Formel (2.2) wurde so gewählt, dass die kleinstmögliche Anzahl an elementaren Faktoren und die größtmögliche Anzahl an Gruppierungen der Beobachtungen benutzt werden. Es steht, frei beliebig viele Intervalle zu bestimmen, bis hin zu einem Intervall pro jeden einzelnen Tod. Nun ordnet man die N Zeitpunkte t_i, in denen ein Tod oder ein Verlust liegt, der Größe nach ansteigend an und kennzeichnet sie durch
$t'_1 \leq t'_2 \leq ... \leq t'_N$. Damit ergibt sich für den Schätzer:

$$\widehat{P}(t) = \prod_{r} [(N - r)/(N - r + 1)] \tag{2.3}$$

Wobei r Zahlen durchläuft, für die gilt $t'_r \leq t$ und t'_r der Zeitpunkt eines Todes ist.
Falls keine Verluste vorliegen, die Daten also nicht zensiert sind, reduziert sich der Schätzer auf $n(t)/N$, den gewöhnlichen Schätzer der Binomialverteilung.
Diese Formel zeigt, dass $\widehat{P}(t)$ eine Treppenfunktion ist, die sich nur in den Zeitpunkten ändert, in denen mindestens ein Tod eintritt. Für diese Zeitpunkte ist die Funktion unstetig.

Kriterien der Formelwahl

(a) Falls die Anzahl der Tode relativ niedrig ist und die Tode in Reihenfolge der Zeitpunkte arrangiert werden können ohne gruppiert werden zu müssen, wird der PL Schätzer über Formel (2.2) ermittelt.

(b) Falls (a) zu zeitaufwendig ist aber die Anzahl der Tode ebenfalls relativ niedrig ist und die Tode in Reihenfolge arrangiert werden können, aber gruppiert sind, wird der PL Schätzer über Formel (2.3) ermittelt.

(c) Falls weder (a) noch (b) kompakt genug sind, wird folgende Formel verwendet:

$$\widehat{P} = \frac{n - \lambda/2 - \delta}{n - \lambda/2}$$

2.2.3 Beispiel

Zur Veranschaulichung habe ich ein Beispiel nach dem Diagramm Nummer 6-12 aus dem Buch "Epidemology"von Gordis (2000) gewählt: Es wird eine Studie an sechs Patienten durchgeführt. Es lässt sich Folgendes beaobachten:

Tode in den Monaten 4, 10, 14, 24
Verluste zwischen den Monaten 4 - 10, 14 - 24

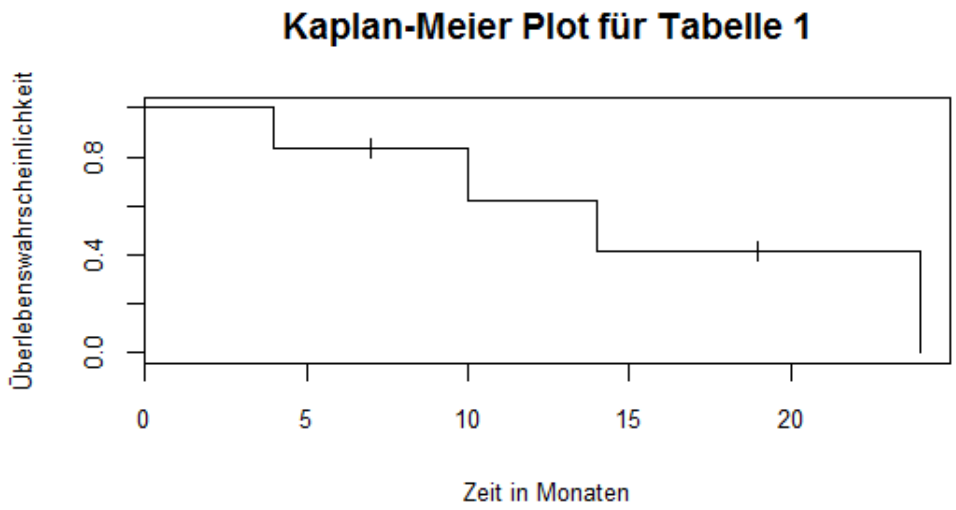
Die Berechnung der $\widehat{P}(t)$ Funktion geschieht mit Hilfe von Formel (2.2) folgendermaßen:

Tabelle 2.1: Beispiel für Formel (2.2)

u_j	n_j	n'_j	δ_j	λ_j	$\widehat{P}(u_j)$
4	6	5	1	1	5/6
10	4	3	1	0	5/8
14	3	2	1	1	5/12
24	1	0	1	0	0

Wobei λ_j die Anzahl der Verluste im j-ten Intervall darstellt. Die Kaplan-Meier Kurve lässt sich grafisch auf folgende Weise darstellen:

Abbildung 2.1: Kaplan Meier-Kurve zu Tabelle 2.1



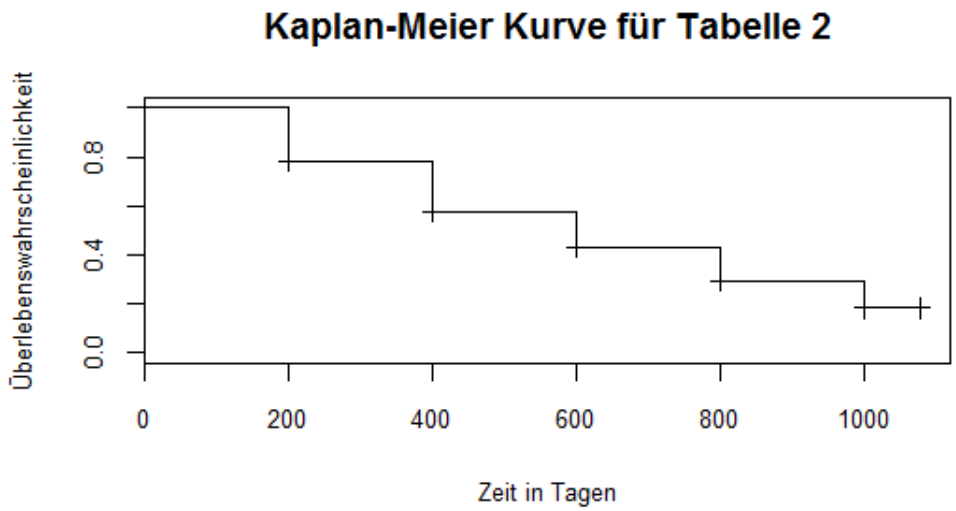
Wobei die Verluste λ_j durch senkrechte Striche auf der Funktion gekennzeichnet sind. Als zweites Beispiel wird die Stichprobe von 238 Heroinabhängigen in zwei Kliniken herangezogen. Hierbei wird aufgrund der zahlreichen Daten Formel (2.3) benutzt, die die Gruppierung von Daten erlaubt:

Tabelle 2.2: Beispiel für Formel (2.3)

u$-j$,u$_j$	δ_j	λ_j	n$_j$	n$'_j$	$\widehat{p}_j$	$\widehat{P}$(u$_j$)
0 - 200	52	18	238	181	181/238	0.76
200 - 400	44	13	168	124	124/168	0.58
400 - 600	28	26	111	83	83/111	0.43
600 - 800	18	18	57	39	39/57	0.30
800 - 1000	8	10	21	13	13/21	0.18
1000 - 1076	0	3	3	3	1	0.18

Grafisch dargestellt sieht die Funktion des Schätzers folgendermaßen aus:

Abbildung 2.2: Kaplan-Meier Kurve für Tabelle 2.2



Da aber in diesem Fall Daten aus zwei unterschiedlichen Kliniken vorliegen, liegt es nahe für jede Klinik eine eigene Überlebensfunktion zu berechenen und diese anschließend zu vergleichen. Die Werte ergeben sich demzufolge für diejenigen 163 Patienten, die in Klinik 1 behandelt wurden, als:

Tabelle 2.3: Tabelle 2.2, für Klinik = 1

u$-j$,u$_j$	δ_j	λ_j	n$_j$	n$'_j$	$\widehat{p}_j$	$\widehat{P}$(u$_j$)
0 - 200	40	13	163	123	123/163	0.76
200 - 400	34	8	110	76	76/110	0.52
400 - 600	25	13	68	43	43/68	0.33
600 - 800	16	4	30	14	14/30	0.15
800 - 1000	7	3	10	3	3/10	0.05
1000 - 1076	0	0	0	0	0	0

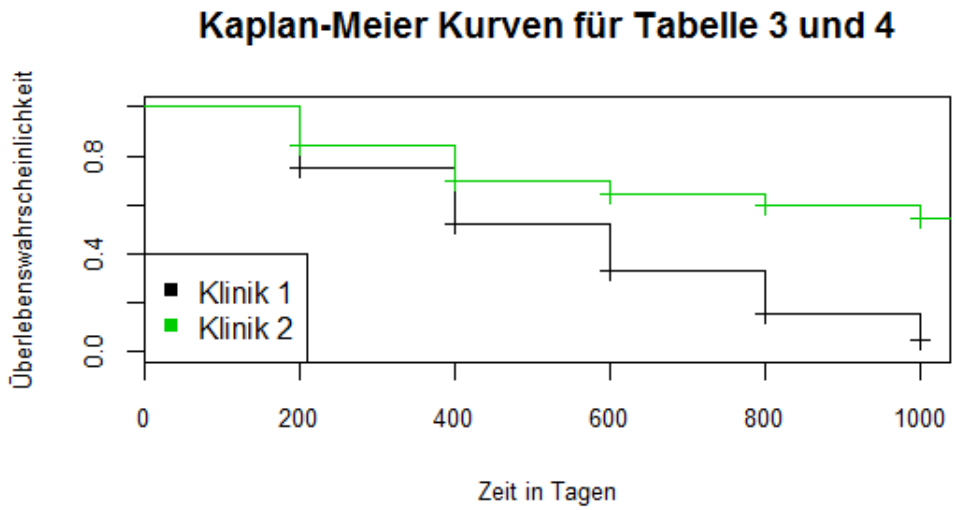
Für die 75 Patienten, die in Klinik 2 behandelt wurden, ergeben sich folgende Werte:

Tabelle 2.4: Tabelle 2.2, für Klinik = 2

u−j,u$_j$	δ_j	λ_j	n$_j$	n$'_j$	$\widehat{p}_j$	$\widehat{P}$(u$_j$)
0 - 200	12	5	75	63	63/75	0.84
200 - 400	10	5	58	48	48/58	0.70
400 - 600	3	13	43	40	40/43	0.65
600 - 800	2	14	27	25	25/27	0.60
800 - 1000	1	7	11	10	10/11	0.54
1000 - 1076	0	3	3	3	1	0.54

Wenn man beide Kurven in einer Grafik darstellt, sind nun die Unterschiede zwischen den Überlebenswahrscheinlichkeiten der beiden Kliniken leicht erkennbar:

Abbildung 2.3: Kaplan-Meier Kurven für Tabellen 2.3 und 2.4



2.3 Varianz und Mittel des Kaplan-Meier Schätzers

2.3.1 Berechnung

Als wichtige Eigenschaft von $\widehat{P}(t)$ gilt, dass der Schätzer sowohl konsistent als auch unverzerrt ist (siehe Abschnitt 2.4) und dass man einen asymptotischen Ausdruck für seine Varianz erhalten kann. Im gleichen Sinne wie der Schätzer an sich, ist der Varianzschätzer der Stichprobe unabhängig von den Grenzen der Beobachtungen, bei denen kein tatsächlicher Verlust eingetroffen ist. Die Varianz der gesamten Population hängt

jedoch von allen Grenzen der Beobachtung ab, die während des Samplings als fest angenommen werden.
In Abschnitt 2.4.2 wird gezeigt, dass die Varianz von *P(t)* approximiert werden kann durch

$$Var[\widehat{P}(t)] \doteq P^2(t) \sum_{1}^{k} (q_j / N_j p_j) \tag{2.4}$$

Wobei q_j = 1 - p_j. Wenn man in diese Formel die Schätzer der Stichprobengrößen einsetzt, erhält man

$$\widehat{Var}[\widehat{P}(t)] \doteq \widehat{P}^2(t) \sum_{1}^{k} [\delta_j / n_j (n_j - \delta_j)] = \widehat{P}^2(t) \sum_{1}^{k} \left(\frac{1}{n'_j} - \frac{1}{n_j} \right) \tag{2.5}$$

Wobei $\mathcal{L}_1 < \mathcal{L}_2 < ... < \mathcal{L}_{k-1}$ die verschiedenen Grenzen der Beobachtungen sind, die kleiner sind als der Zeitpunkt $t = \mathcal{L}_k$, in dem P(t) geschätzt wird.
Diese Formel (2.5) bleibt ebenfalls bei einer Reduzierung der Parameter, wie in Formel (2.2), oder einer Erweiterung auf ein Intervall pro Tod, wie in Formel (2.3), gültig. Da Formel (2.5) nach Einsetzten von Formel (2.2) als

$$\widehat{Var}[\widehat{P}(t)] \doteq \widehat{P}^2(t) * \widehat{p}(t)$$

darstellbar ist, ergibt sich für den Schätzer aus Formel (2.3)

$$\widehat{Var}[\widehat{P}(t)] \doteq \widehat{P}^2(t) * \widehat{P}(t)$$

und somit lässt sich die Varianz von $\widehat{P}$*(t)* ebenfalls darstellen als:

$$\widehat{Var}[\widehat{P}(t)] \doteq \widehat{P}^2(t) \sum_{r} [(N - r)(N - r + 1)]^{+1} \tag{2.6}$$

wobei r aus den positiven ganzen Zahlen, für die $t'_r \leq t$, wobei t'_r den Zeitpunkt eines Todes kennzeichnet.
Mit der berechneten Varianz lassen sich nun auch punktweise (1-α)-Konfidenzintervalle für $\widehat{P}$*(t)* berechnen (Hedderich and Sachs, 2006) :

$$\mathrm{KI}_{\widehat{P}(t)} = [\widehat{P}(t) \pm z_{1-\alpha/2} \sqrt{Var(\widehat{P}(t))}] \tag{2.7}$$

2.3.2 Beispiel

Für das Beispiel der 6 Patienten aus Tabelle 1 ergab sich für $\widehat{P}(10) = 5/8 = 0.625$. Mit Formel (2.5) ergibt sich der Schätzer der Varianz von $P(10)$ als:

$$\widehat{Var}[\widehat{P}(10)] = (0.625)^2\left(\left(\frac{1}{5} - \frac{1}{6})\right) + \left(\frac{1}{3} - \frac{1}{4}\right)\right) = 0.046$$

Der 95%-Konfidenzintervall lässt sich berechnen als:

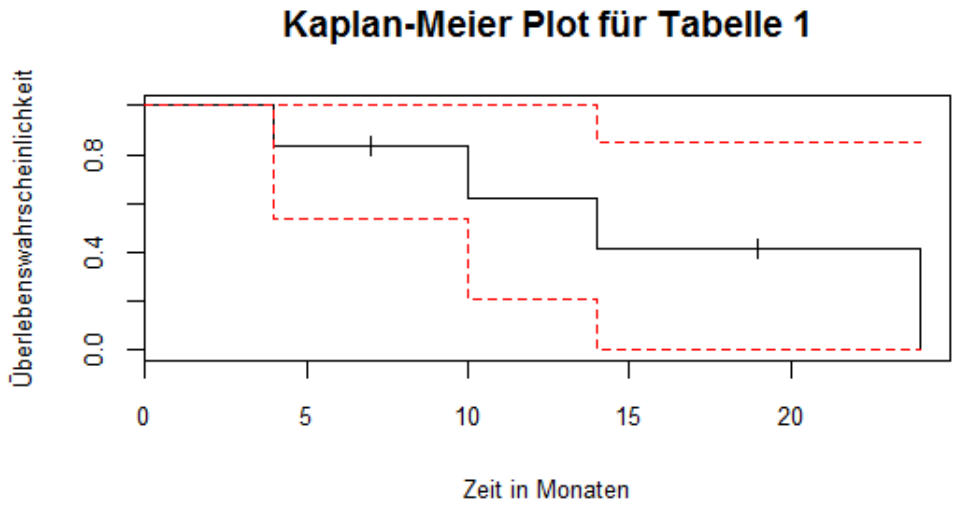
$$KI_{\widehat{P}(10)} = [0.625 \pm 1.96\sqrt{0.046}] = [0.205; 1.045]$$

Jedoch ist hierbei zu beachten, dass der Konfidenzintervall für $\widehat{P}(t)$ wiederrum im Intervall [0;1] liegen muss. Somit ergibt sich:

$$KI_{\widehat{P}(10)} = [0.205; 1.000]$$

Berechnet man diesen für alle Zeitpunkte t so lassen sich die 95%-Konfidenzintervallsgrenzen grafisch (in diesem Fall durch rote Linien) wie folgt darstellen:

Abbildung 2.4: Kaplan-Meier Kurve für Tabelle 2.1 mit markiertem 95%-Konfidenzintervall



2.4 Eigenschaften des Kaplan-Meier Schätzers

2.4.1 Unverzerrtheit

Mit den Formeln (2.1) und (2.2) lässt sich zeigen, dass man den Kaplan-Meier Schätzer auch darstellen kann als:

$$\widehat{P}(t) = \prod_{j=1}^{k} \widehat{p}_j \tag{2.8}$$

wobei $\widehat{p}_j = ((n_j \text{ - } \delta_j) \ / \ n_j)$.
Nun sei $\mathbb{E}_j$ der feste bedingte Erwartungswert für $n_1, \ \ldots \ , n_j$. Dann gilt $\mathbb{E}_j(\widehat{p}_j) = p_j$, unter der Bedingung dass $n_j > 0$.
Falls $n_j > 0$ vernachlassbar, ergibt sich für den Erwartungswert des Schätzers:

$$\begin{aligned} \mathbb{E}\widehat{P}(t) &= \mathbb{E}[\widehat{p}_j ... \widehat{p}_{k-1} \mathbb{E}_k \widehat{p}_k] = \mathbb{E}[\widehat{p}_1 ... \widehat{p}_{k-1} p_k] \\ &= p_k \mathbb{E}[\widehat{p}_1 ... \widehat{p}_{k-2} \mathbb{E}_{k-1} \widehat{p}_{k-1}] \\ &= ... = p_k p_{k-1} ... p_1 = P(t) \end{aligned} \tag{2.9}$$

Somit gilt, dass $\widehat{P}(t)$ ein unverzerrter Schätzer ist.

2.4.2 Greenwood Formel

Unter der Annahme, dass die konditionellen Varianzen von n_j fest sind in N_j, wird $\widehat{p}_j$ als unabhängig behandelt und es ergibt sich:

$$\begin{aligned} \mathbb{E}[\widehat{P}^2(t)] &= \prod_{j=1}^{k} \left(p_j^2 + \frac{p_j q_j}{N_j} \right) = p_1^2 ... p_k^2 \prod_{j=k}^{k} \left(1 + \frac{q_j}{N_j p_j} \right) \\ &= p^2(t) \prod_{j=1}^{k} \left(1 + \frac{q_j}{N_j p_j} \right) \end{aligned} \tag{2.10}$$

Falls die Reihenfolge ${N_j}^{-2}$ vernachlassbar ist, so lässt sich die Varianz des Schätzers approximieren durch:

$$Var[\widehat{P}(t)] \doteq P^2(t) \sum_{j=1}^{k} \frac{q_j}{N_j p_j} \tag{2.11}$$

Diese Formel wird als „Greenwood Formel“ bezeichnet (Greenwood, 1926).

2.4.3 Sonstige Eigenschaften

Zusätzlich wurden folgende Eigenschaften des Kaplan-Meier Schätzers nach der Veröffentlichung des Artikels bewiesen:

- $\widehat{P}$(t) ist der generalisierte Maximum-Likelihood Schätzer von *F(t)* (Johansen, 1978)
- $\widehat{P}$(t) ist schwach konvergent. (Efron, 1967)
- $\widehat{P}$(t) ist stark konsistent (Peterson, 1977), beziehungsweise stark uniform konsistent (Foeldes and Rejtoe, 1981).

2.5 Kovarianz und mittlere Lebensdauer

2.5.1 Mittlere Lebensdauer

Der Kaplan-Meier Schätzer $\widehat{\mu}$ der mittleren Lebensdauer μ ist definiert als das Mittel des Schätzers der Verteilung. Es ist bekannt, dass das Mittel einer nicht-negativen unabhängigen Variable der Fläche unter der entsprechenden Überlebensfunktion entspricht:

$$\widehat{\mu} = \int_0^\infty \widehat{P}(t)d(t) \tag{2.12}$$

Selbstverständlich gilt, dass falls $\widehat{P}$(t) nicht an allen Stellen bestimmt ist, $\widehat{\mu}$ undefiniert ist.
Falls man auf eine gruppierte Tabelle beschränkt ist (wie zum Beispiel Tabelle 2) so ist es nötig, aktuarial-typische Annahmen zu treffen (zum Beispiel die Trapezregel) um $\widehat{\mu}$ und dessen Varianz schätzen zu können.

Beispiel

Für das Beispiel aus Tabelle 1 ergibt sich:

$$\begin{aligned}\widehat{\mu} = 1 * 4 + (5/6) * (4-1) + (5/8) * (10-4) + (5/12) * (14-10) \\ = 4 + 2.5 + 3.75 + 1.667 = 11.917\end{aligned}$$

Für das Beispiel aus Tabelle 2 ergibt sich mit Anwendung der Trapezregel:

$$\begin{aligned}\widehat{\mu} = \frac{1}{2}(1 + 0.76) * (200 - 0) + (0.76 + 0.58) * (400 - 200) \\ +... + (0.18 + 0.18) * (1079 - 1000) \\ = 545.68\end{aligned}$$

2.5.2 Kovarianz

Mit der Methode von Irwin (1949) lässt sich die Varianz von $\widehat{\mu}$ darstellen als:

$$\begin{aligned}\mathbb{E}(\mu^2) &= \int_0^\infty \int_0^\infty \widehat{P}(u)\widehat{P}(v)dvdu \\ &= 2\int_0^\infty \int_u^\infty \mathbb{E}[\widehat{P}(u)\widehat{P}(v)]dvdu\end{aligned} \tag{2.13}$$

Seien nun $u < v$ und $\mathcal{L}_h$ und $\mathcal{L}_k$ bezeichnen entsprechend u und v. Somit ergibt sich:

$$\begin{aligned}\mathbb{E}[\widehat{P}(u)\widehat{P}(v)] &= \mathbb{E}[\widehat{P}^2(u)\mathbb{E}_{h+1}(\widehat{p}_{h+1}\widehat{p}_{h+2}...\widehat{p}_k)] \doteq p_{h+1}p_{h+2}...p_k\mathbb{E}[\widehat{P}^2(u)] \\ &= [P(v)/P(u)]\mathbb{E}[\widehat{P}^2(u)] \doteq P(u)P(v)[1+U(u)]\end{aligned} \tag{2.14}$$

Wobei *U(u)* $\doteq \sum$ *[(N-r)(N-r+1)]*$^{-1}$, also die Summe der Tode in all den Zeitpunkten, die u nicht überschreiten und somit ein Schätzer der Stichprobengröße ist. Für die Varianz von $\widehat{\mu}$ ergibt sich demnach approximativ:

$$\begin{aligned}Var(\widehat{\mu}) \doteq 2\int_0^\infty \int_u^\infty P(u)P(v)U(u)du &= 2\int_0^\infty A(u)P(u)U(u)du \\ &= \int_0^\infty A^2(u)dU(u),\end{aligned} \tag{2.15}$$

$$\text{Mit } A(u) = \int_u^\infty P(v)dv$$

2.6 Kritik und Zusammenhänge mit späteren Methoden

Viele spätere Theorien und Methoden greifen die Idee des Kaplan-Meier Schätzers auf und versuchen unter anderem diesen zu verbessern, da es seit der Publizierung des Artikels durchaus auch Kritik gab.

2.6.1 Kritik

Ein wichtiger Kritikpunkt ist, dass der Schätzer erst ab einem Stichprobenumfang von mindestens 15 Individuen sinnvoll ist. Andernfalls kann das Ergebnis stark verzerrt werden, da es bei einer kleineren Stichprobe schwieriger wird eventuelle Abhängigkeiten zu

erkennen.
Auch wenn die Stichprobe größtenteils aus zensierten Daten besteht oder falls in dem vorhergehenden Intervall eine hohe Anzahl an Toden eintrat und somit die Zahl der beobachtbaren Individuen sehr klein ist, kann der Schätzer verzerrte Ergebnisse liefern. In diesem Fall neigt $\widehat{P}(t)$ dazu, die tatsächliche Varianz zu unterschätzen. Dies ist vor allem in den späteren Intervallen, in denen für die meisten Individuen bereits entweder ein Tod oder ein Verlust eingetreten ist, der Fall. Somit treffen Schätzer für spätere Intervalle generell weniger verlässliche Prognosen als die für frühere.
Außerdem ist der Kaplan-Meier Schätzer nicht verwendbar, falls beidseitig zensierte Daten vorliegen.
Ein weiterer großer Schwachpunkt der Methode von Kaplan und Meier besteht darin, dass es meistens nicht aussagekräftig genug ist, nur die Unterschiede zweier Überlebensfunktionen zu betrachten, da man im Rahmen einer Studie in den meisten Fällen ebenfalls an den Auslösern einer höheren beziehungsweise einer niedrigeren Überlebenswahrscheinlichkeit interessiert ist. Der Kaplan-Meier Schätzer ist als univariate Methode jedoch nicht im Stande zusätzliche Faktoren zu berechnen und zu bestimmen. Auch Breslow hat in dem Vorwort zu dem Artikel von Kaplan und Meier das Problem angemerkt, dass die Kaplan-Meier Kurve im Gegensatz zur Form der unmittelbaren Hazard-Funktion wenig informativ ist. (Breslow, 1992)

2.6.2 Zusammenhänge mit späteren Theorien und Methoden

Cox-Regression

3 Schluss

3.1 Zusammenfassung

Der Kaplan-Meier Schätzer ist vor allem bei ausreichend großem Stichprobenumfang ein effektiver unverzerrter und konsistenter Schätzer, der bei der Schätzung von Lebensdauern auch zensierte Daten in die Berechnung mit einbezieht. Das Mittel kann wie gewohnt als Fläche unter der entsprechenden Kurve des Schätzers berechnet werden und die Varianz lässt sich über die Greenwood Formel approximieren. Der Schätzer hat die Form einer Treppenfunktion und eignet sich zum Vergleich der Schätzungen mehrer Überlebensfunktionen.
Falls keine zensierten Daten vorliegen, reduziert sich der Schätzer auf die empirische Überlebensfunktion.

3.2 Verwendung in heutiger Zeit

Der Kaplan-Meier Schätzer ist bis heute vor allem in der Medizin und der Pharamazie eine der meistverwendeten Methoden zur Modellierung von Überlebenskurven. Kaum eine medizinische Studie verwendet heutzutage keine Kaplan-Meier Kurven zur Interpretation und Veranschaulichung der Ergebnisse. Aber auch in anderen Gebieten, unter anderem der Wirtschaft (zum Beispiel zur Messung des Zeitraums, in dem ein Arbeitsloser eine Neueinstellung findet) oder auch dem Ingenieurwesen (beispielsweise zur Messung der Zeit, die bis zum Versagen eines Bauteiles vergeht) kann der Kaplan-Meier Schätzer verwendet werden. Wenn es jedoch darum geht, die Überlebenswahrscheinlichkeiten mehrerer Gruppen zu vergleichen, geht das mithilfe dieses Schätzers nur mit reinem Auge. Um die Ergebnisse genau zu interpretieren ist die Durchführung eines log-rank Test (Mantel, 1966) nötig.

R-Code zur Erstellung der Grafiken

Grafik 1

```
zeit <- c(4,7,10,14,19,24)
status <-c(1,0,1,1,0,1)
pat.surv <- Surv(zeit, status)
fit.pat<- survfit(pat.surv ~ 1, conf.type="none")
plot(fit.pat, main="Kaplan-Meier Plot für Tabelle 1",
xlab="Zeit in Monaten", ylab="Überlebenswahrscheinlichkeit",
cex.lab=0.8,cex.axis=0.8)
```

Grafik 2

```
 install.packages("survival")
library(survival)
addicts <- read.table("C:/Users/Maja/Downloads/addicts.txt")
attach(addicts)


int1 <- addicts[(addicts$Time<201),]
int2 <- addicts[(addicts$Time>200)&(addicts$Time<401),]
int3 <- addicts[(addicts$Time>400)&(addicts$Time<601),]
int4 <- addicts[(addicts$Time>600)&(addicts$Time<801),]
int5 <- addicts[(addicts$Time>800)&(addicts$Time<1001),]
int6 <- addicts[(addicts$Time>1000),]
int1o <- int1[order(int1$Time),]
int2o <- int2[order(int2$Time),]
int3o <- int3[order(int3$Time),]
int4o <- int4[order(int4$Time),]
int5o <- int5[order(int5$Time),]
int6o <- int6[order(int6$Time),]


length(int1o$Status[int1o$Status == 1])
length(int1o$Status[int1o$Status == 0])
length(int2o$Status[int2o$Status == 1])
length(int2o$Status[int2o$Status == 0])
length(int3o$Status[int3o$Status == 1])
length(int3o$Status[int3o$Status == 0])
length(int4o$Status[int4o$Status == 1])
length(int4o$Status[int4o$Status == 0])
length(int5o$Status[int5o$Status == 1])
length(int5o$Status[int5o$Status == 0])
length(int6o$Status[int6o$Status == 1])
```

```
length(int6o$Status[int6o$Status == 0])


zeit2 <- c(rep(200,70), rep(400,57), rep(600,54), rep(800,36),
rep(1000,18),rep(1078,3))
status2 <- c(rep(c(1,0),18), rep(1,34), rep(1, 44), rep(0,13), rep(1,28),
rep(0,26), rep(1,18), rep(0,18), rep(1,8), rep(0,13))

surv2 <- Surv(zeit2, status2)
fit.h <- survfit(surv2 ~, conf.type="none")
plot(fit2, main="Kaplan-Meier Kurve für Tabelle 2 mit Gruppierungen",
ylab="Überlebenswahrscheinlichkeit", xlab="Zeit in Tagen",
cex.axis=0.8, cex.lab=0.8)
```

Grafik 3

```
 int1o <- int1[order(int1$Time),]
int2o <- int2[order(int2$Time),]
int3o <- int3[order(int3$Time),]
int4o <- int4[order(int4$Time),]
int5o <- int5[order(int5$Time),]
int6o <- int6[order(int6$Time),]


int11 <- int1o[int1o$Clinic==1,]
int12 <- int1o[int1o$Clinic==2,]
int21 <- int2o[int2o$Clinic==1,]
int22 <- int2o[int2o$Clinic==2,]
int31 <- int3o[int3o$Clinic==1,]
int32 <- int3o[int3o$Clinic==2,]
int41 <- int4o[int4o$Clinic==1,]
int42 <- int4o[int4o$Clinic==2,]
int51 <- int5o[int5o$Clinic==1,]
int52 <- int5o[int5o$Clinic==2,]
int61 <- int6o[int6o$Clinic==1,]
int62 <- int6o[int6o$Clinic==2,]


Clinic 1
length(int11$Status[int11$Status == 1])
length(int11$Status[int11$Status == 0])
length(int21$Status[int21$Status == 1])
length(int21$Status[int21$Status == 0])
```

```
length(int31$Status[int31$Status == 1])
length(int31$Status[int31$Status == 0])
length(int41$Status[int41$Status == 1])
length(int41$Status[int41$Status == 0])
length(int51$Status[int51$Status == 1])
length(int51$Status[int51$Status == 0])
length(int61$Status[int61$Status == 1])
length(int61$Status[int61$Status == 0])


Clinic 2
length(int12$Status[int12$Status == 1])
length(int12$Status[int12$Status == 0])
length(int22$Status[int22$Status == 1])
length(int22$Status[int22$Status == 0])
length(int32$Status[int32$Status == 1])
length(int32$Status[int32$Status == 0])
length(int42$Status[int42$Status == 1])
length(int42$Status[int42$Status == 0])
length(int52$Status[int52$Status == 1])
length(int52$Status[int52$Status == 0])
length(int62$Status[int62$Status == 1])
length(int62$Status[int62$Status == 0])


zeit3 <- c(rep(200,53),rep(400,42), rep(600,38),
rep(800,20), rep(1000,10))
status3 <- c(rep(1,40), rep(0,13), rep(1,34), rep(0,8), rep(1,25),
rep(0,13), rep(1,16), rep(0,4), rep(1,7), rep(0,3))
zeit4 <- c(rep(200,17), rep(400,15), rep(600,16), rep(800,16),
rep(1000,8), rep(1078,3))
status4 <- c(rep(1,12), rep(0,5), rep(1,10), rep(0,5), rep(1,3),
rep(0,13), rep(1, 2), rep(0,14), rep(1,1), rep(0,10))


surv3 <- Surv(zeit3, status3)
fit3 <- survfit(surv3 ~ 1, conf.type="none")
surv4 <- Surv(zeit4, status4)
fit4 <- survfit(surv4 ~ 1, conf.type="none")
plot(fit3, main="Kaplan-Meier Kurven für Tabelle 3 und 4",
ylab="Überlebenswahrscheinlichkeit", xlab="Zeit in Tagen",
cex.lab=0.8, cex.axis=0.8)
lines(fit4,col=3)
legend("bottomleft", c("Klinik 1", "Klinik 2"),col=c(1,3), pch=15)
```

Grafik 4

```
 fit.pat2<- survfit(pat.surv t∼1, conf.int=.95, conf.type="plain")
plot(fit.pat2, col=c(1,2,2), main="Kaplan-Meier Plot für Tabelle 1",
xlab="Zeit in Monaten", ylab="Überlebenswahrscheinlichkeit",
cex.lab=0.8,cex.axis=0.8)
```

Literaturverzeichnis

Andersen, P. and N. Keiding (Eds.) (2006). *Survival and Event History Analysis*. John Willey and Sons Ltd.

Breslow, N. (1992). Introduction to kaplan and meier (1958) nonparametric estimation from incomplete observations. In *Breakthroughs in Statistics*, Springer Series in Statistics, pp. 311–318. Springer New York.

Caplehorn, J., S. Dalton, C. Cluff, and A. Petrenas (1994). Retention in methadone maintenance and heroin addicts risk of death. *Addiction 89*, 203–207.

Efron, B. (1967). The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 4*, 831–853.

Foeldes, A. and L. Rejtoe (1981). Strong uniform consistency for nonparametric survival curve estimators from randomly censored data. *Annals of Statistics 9*, 122–129.

Gordis, L. (2000). *Epidemiology*. Saunders.

Greenwood, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects 33*.

Hedderich, J. and N. Sachs (2006). *Angewandte Statistik*. Springer.

Irwin, J. (1949). The standard error of an estimate of expectational life. *Journal of Hygiene 47*, 188–9.

Johansen, S. (1978). The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics 5*, 195–199.

Kaplan, E. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*, 457–481.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports 50*, 163–170.

Peterson, A. (1977). Expressing the kaplan-meier estimator as a function of empirical suvsurvival functions. *Journal of the American Statistical Association 72*, 854–858.

Turnbull, B. (1974). Nonparamtric estimation of a survivorship with doubly censored data. *Journal of the American Statistical Association 69*, 169–173.

Erklärung zur Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Seminar-Arbeit selbststa”ndig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 30. Mai 2014

(Maja Krajewska)