Bachelorarbeit

Vergleich von verschiedenen getrimmten Schätzungen bei zensierten Daten

von

Alexandra Höller

Der Fakultät Statistik
der Technischen Universität Dortmund
vorgelegt am 15. August 2015

Betreuerin: Prof. Dr. Christine Müller

Inhaltsverzeichnis

1. Einleitung

Die Maximum-Likelihood-Methode ist eines der wichtigsten Konzepte, um Schätzungen für die Parameter einer Verteilung zu erhalten. Jedoch sind Maximum-Likelihood-Schätzer oft nicht robust. Bei der Analyse von Ereigniszeiten wird der Einfluss von Ausreißern häufig durch die Zensierung der Daten begrenzt. Dies wirft die Frage auf, ob der Maximum-Likelihood-Schätzer bei rechtszensierten Daten als robust angesehen werden kann oder ob alternative Schätzer angewendet werden sollten.

Eine wichtige Verteilung für die Beschreibung von Ereigniszeiten ist die Exponentialverteilung. Sie hat den Vorteil dass die Likelihood-Funktion auch bei zensierten Daten eine einfache Form besitzt. Das ermöglicht, auch getrimmte Likelihood-Schätzungen für den Parameter der Exponentialverteilung zu betrachten. Während bei der Maximum-Likelihood-Methode die ganze Likelihood-Funktion maximiert wird, wird bei getrimmten Likelihood-Schätzungen die Likelihood-Funktion durch eine getrimmte Version ersetzt.

Das Hauptaugenmerk der vorliegenden Arbeit liegt auf dem Vergleich von verschiedenen getrimmten Schätzungen bei rechtszensierten Daten. Dabei soll der klassische Maximum-Likelihood-Schätzer mit der getrimmten Version der Likelihood-Funktion sowie zwei neuen Schätzern, die von Clarke et al. (2014) vorgestellt wurden, mittels Simulationen verglichen werden. Die vorgeschlagenen Schätzer von Clarke et al. (2014) beziehen dabei das nach oben getrimmte Mittel, das als sehr effizient gilt, in die Maximum-Likelihood-Schätzung ein.

In Kapitel 2 werden zunächst die grundlegenden Begriffe und Konzepte der Ereigniszeitanalyse, die für die vorliegende Arbeit Anwendung finden, beschrieben. Grundlegende yMethoden der Parameterschätzung werden anschließend in Kapitel 3 vorgestellt. Dabei wird insbesondere auf Verfahren eingegangen, die für die Parameterschätzung bei rechtszensierten Ereigniszeitdaten relevant sind. In Kapitel 4 werden vier Schätzer für den Parameter der Exponentialverteilung vorgestellt, die

auf der Maximierung der Likelihood-Funktion oder deren getrimmter Version basieren. Die vier Schätzer sind für die Parameterschätzung bei Ereigniszeiten, die mit fester Zensierungszeit rechtszensiert wurden, geeignet. Zwei der Schätzer sind auch bei zufällig rechtszensierten Ereigniszeitdaten anwendbar.

In Kapitel 5 werden die vorgestellten Schätzer mittels Simulationen verglichen. Dafür werden sowohl Ereigniszeiten, die mit einer festen Zensierungszeit rechtszensiert wurden sowie zufällig rechtszensierte Ereigniszeiten simuliert und die Schätzer darauf angewendet. Die Simulationsergebnisse werden vergleichend dargestellt. Abschließend werden die Schätzer auf einen Datensatz aus einer klinischen Studie angewendet. Die Arbeit schließt in Kapitel 6 mit einer Zusammenfassung.

2. Analyse von Ereigniszeiten

Dieses Kapitel dient der Einführung von Begriffen und Konzepten der Ereigniszeitanalyse, die in dieser Arbeit Anwendung finden. In Abschnitt 2.1 wird der Begriff der Ereigniszeiten definiert. Zudem wird die Zensierung als eine Besonderheit von Ereigniszeiten vorgestellt. Die Survivalfunktion und die Hazardrate sind zwei Funktionen, die der Beschreibung von Ereigniszeitverteilungen dienen; sie werden in Abschnitt 2.2 eingeführt. In Abschnitt 2.3 schließlich wird die Exponentialverteilung als Verteilungsmodell stetiger Ereigniszeiten dargestellt.

2.1. Ereigniszeiten und Zensierung

In vielen Anwendungsbereichen interessieren Zeiten bis zum Eintreten eines wohldefinierten Ereignisses. Die beobachteten Daten werden als *Ereigniszeiten* bezeichnet (vgl. Schumacher und Schulgen (2008), S. 77). In klinischen Studien können dies beispielsweise die Überlebenszeiten nach Transplantationen oder die Zeiten bis zum Nachlassen von Krankheitssymptomen sein. Weitere Beispiele für interessierende Ereigniszeiten sind die Zeitspanne, die in einem psychologischen Experiment für die Erledigung einer Aufgabe benötigt wird, die Dauer bis zum Konkurs von Kleinunternehmen oder die Ausfallzeiten von technischen Geräten.

Ereigniszeiten können oft nur unvollständig beobachtet werden. Die Ereigniszeit eines Untersuchungssobjekts wird als *zensiert* bezeichnet, wenn von dem Zielereignis nur bekannt ist, dass es in einen bestimmten Zeitbereich fällt (vgl. Klein und Moeschberger (2003), S. 63). Es gibt verschiedene Zensierungsmodelle, von denen in der vorliegenden Arbeit nur die sogenannte *Rechtszensierung* betrachtet wird. Eine Rechtszensierung liegt vor, wenn nur die Information bekannt ist, dass das Zielereignis bis zu einem gewissen Zeitpunkt noch nicht eingetreten war. Eine rechtszensierte

Ereigniszeit ist also kleiner als die tatsächliche, aber unbekannte Ereigniszeit (vgl. Collett (2003), S. 2).

Ein häufiger Grund für die Rechtszensierung von Ereigniszeiten ist, dass eine Studie endet, bevor das Zielereignis bei allen Untersuchungsobjekten eingetreten ist. In klinischen Studien kann eine rechtszensierte Ereigniszeit auch darin begründet sein, dass im Verlauf der Studie der Kontakt zum Studienteilnehmer abgebrochen ist. Dies wird als auch als *Drop-Out* oder *Loss to Follow-Up* bezeichnet (vgl. Schumacher und Schulgen (2008), S. 80).

In den Abbildungen 2.1 und 2.2 sind zwei typische Studiensituationen, die zu rechtszensierten Beobachtungen führen, dargestellt. In Abbildung 2.1 beginnt und endet die Studie zu festgelegten Zeitpunkten. Die Untersuchungsobjekte 1 und 3 sind rechtszensiert, da das Zielereignis bis zum Studienende nicht eingetreten ist, die Untersuchungsobjekte 2 und 4 sind unzensiert. Diese Art der Rechtszensierung wird im Folgenden auch als *Rechtszensierung mit fester Zensierungszeit* bezeichnet.

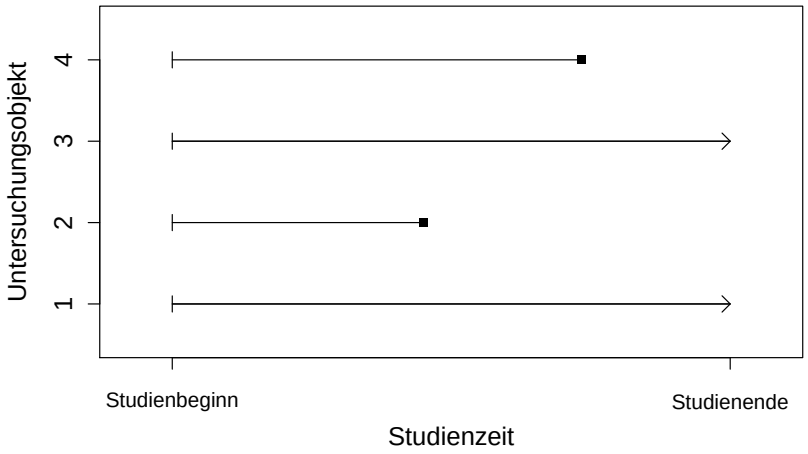

Abbildung 2.1.: Beispielhafte Darstellung der Ereigniszeiten einer Studie mit zensierten (Objekt 1 und 3) und unzensierten (Objekt 2 und 4) Beobachtungen.

In Abbildung 2.2 sind beispielhaft die Ereigniszeiten einer klinischen Studie dargestellt, die ebenfalls über einen festen Zeitraum läuft. Die Patienten können jedoch zu unterschiedlichen Zeitpunkten in die Studie eintreten oder aus der Studie ausscheiden. Jeder Patient hat also eine eigene Zensierungszeit. In dem Beispiel sind die Patienten 1, 3 und 4 zensiert, wobei die Zensierungszeiten der Patienten 1 und 4 durch das Studienende bedingt sind und bei Patient 3 beispielsweise durch einen

Drop-Out hervorgerufen wurde. Die Patienten 2 und 5 sind unzensiert. Diese Art der Rechtszensierung wird im Folgenden als *Zufällige Rechtszensierung* bezeichnet.

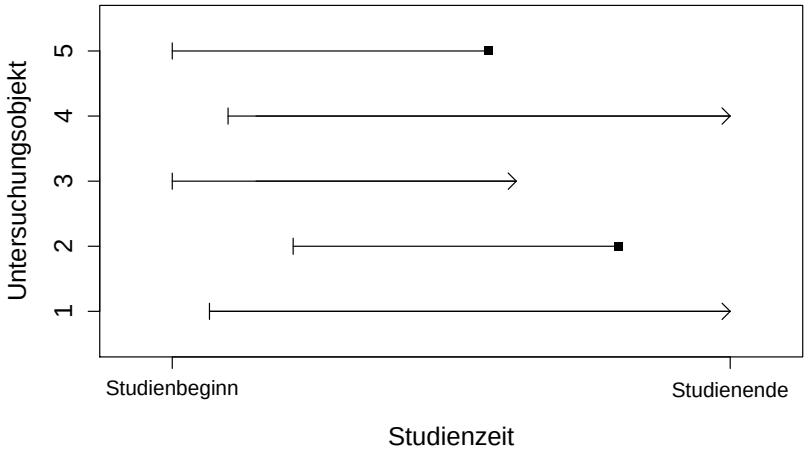

Abbildung 2.2.: Beispielhafte Darstellung der Ereigniszeiten einer klinischen Studie mit zensierten (Patienten 1, 3 und 4) und unzensierten (Patienten 2 und 5) Beobachtungen.

Bei rechtszensierten Ereigniszeiten werden das Minimun einer Ereigniszeit und einer Zensierungszeit beobachtet. Liegt eine Rechtszensierung mit fester Zensierungszeit vor, so sind die beobachteten Ereigniszeiten aller zensierten Untersuchungsobjekte gleich einer festen Zensierungskonstante. Bei zufälliger Zensierung sind die beobachteten Ereigniszeiten der zensierten Untersuchungsobjekte individuell. Eine wichtige Annahme bei der Analyse der Ereigniszeiten ist die Unabhängigkeit des Zensierungsmechanismus von den Ereigniszeiten (vgl. Schumacher und Schulgen (2008), S. 80).

2.2. Beschreibung von Ereigniszeitverteilungen

Sei T die Zeit bis zum Eintreten eines definierten Ereignisses. T ist dann eine nichtnegative, stetige Zufallsvariable. Die zugehörige Dichtefunktion sei mit $f(t)$, $t \geq 0$, bezeichnet. Die Verteilungsfunktion von T ist gegeben durch

$$F(t) = P(T < t) = \int_0^t f(u)\,\mathrm{d}u,$$

und beschreibt die Wahrscheinlichkeit, dass die Ereigniszeit einen Wert kleiner t annimmt.

In der Analyse von Ereigniszeiten sind zwei Funktionen von besonderem Interesse: die *Survivalfunktion* und die *Hazardrate*. Die Survivalfunktion ist definiert als die Wahrscheinlichkeit, dass das Ereignis bis zu einem Zeitpunkt t noch nicht eingetreten ist:

$$S(t) = P(T \geq t) = 1 - F(t). \tag{2.1}$$

$S(t)$ ist eine streng monoton fallende Funktion mit $S(0) = 1$ und $\lim_{t\to\infty} S(t) = 0$.

Die Hazardrate beschreibt das infinitesimale Risiko, dass das Ereignis im nächsten Moment eintritt, wenn es bis zum Zeitpunkt t noch nicht eingetreten war. Sie ist definiert durch

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

$h(t)$ ist nicht normiert und es gilt $0 \leq h(t) < \infty$.

Zwischen Hazardrate und Survivalfunktion besteht der folgende Zusammenhang:

$$\begin{aligned}
h(t) &= \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\
&= \lim_{\Delta t\to 0} \frac{P(\{t \leq T < t + \Delta t\} \cap \{T \geq t\})}{\Delta t\ P(T \geq t)} \\
&= \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{P(T \geq t)} \\
&= \lim_{\Delta t\to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} \\
&= \frac{f(t)}{S(t)}
\end{aligned}$$

und mit

$$f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}$$

ergibt sich

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\ log\ S(t). \tag{2.2}$$

Survivalfunktion und Hazardrate werden aus den beobachteten Ereigniszeiten geschätzt. Neben nichtparametrischen Methoden für die Schätzung der beiden Funktionen gibt es parametrische Methoden, die auf der Annahme einer bestimmten Verteilung der Ereigniszeiten beruhen.

(Vergleiche dazu Collett (2003), S. 11 f.)

2.3. Exponentialverteilung zur Modellierung von stetigen Ereigniszeiten

Ein wichtiges Verteilungsmodell für stetige Ereigniszeiten ist die Exponentialverteilung mit Parameter λ. Die Dichtefunktion und die Verteilungsfunktion der Exponentialverteilung sind gegeben durch

$$f_\lambda(t) = \frac{1}{\lambda} \mathrm{e}^{-t/\lambda}$$

$$F_\lambda(t) = 1 - \mathrm{e}^{-t/\lambda},$$

für $t > 0$. Der Parameter λ ist eine positive Konstante und wird aus den beobachteten Daten geschätzt.

Aus (2.1) und (2.2) ergeben sich für die Survivalfunktion und die Harazrdrate der Exponentialverteilung

$$S_\lambda(t) = 1 - (1 - \mathrm{e}^{-t/\lambda}) = \mathrm{e}^{-t/\lambda}$$

$$h_\lambda(t) = \frac{\frac{1}{\lambda}\mathrm{e}^{-t/\lambda}}{\mathrm{e}^{-t/\lambda}} = \frac{1}{\lambda}$$

Die Hazardrate ist konstant über die Zeit. Das Risiko, dass das Zielereignis im nächsten Moment eintritt, ist also zu jedem Zeitpunkt $t > 0$ gleich, unabhängig davon, wieviel Zeit bereits verstrichen ist. Dies motiviert auch den Begriff der gedächtnislosen Verteilung.

In Abbildung 2.3 sind die Dichtefunktion, die Survivalfunktion und die Hazardrate der Exponentialverteilung für vier verschiedene Werte des Paramters λ dargestellt.

Für den Erwartungswert einer exponentialverteilten Ereigniszeit T gilt $\mathbb{E}(T) = \lambda$. Je kleiner die zu erwartende Ereigniszeit ist, desto größer ist also das durch die Hazardrate ausgedrückte konstante Risiko, dass das Zielereignis im nächsten Moment

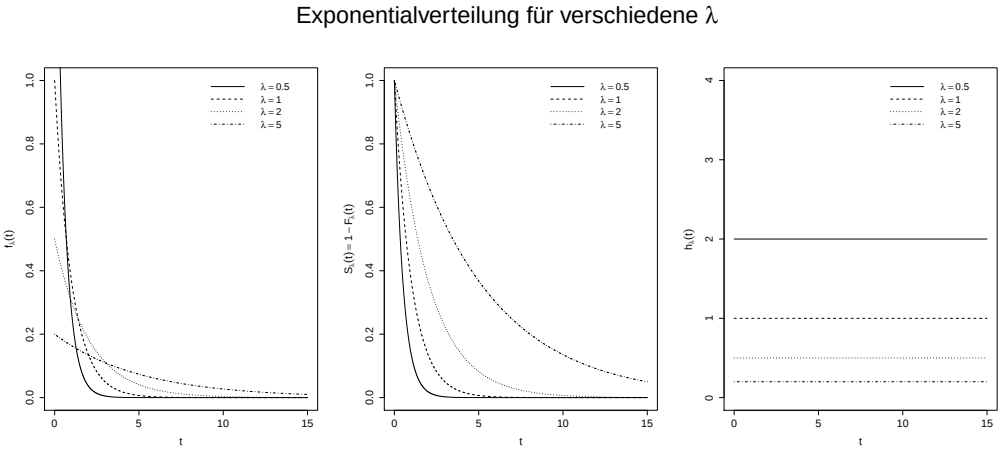

Abbildung 2.3.: Dichtefunktion, Survivalfunktion und Hazardrate der Exponentialverteilung mit den Parametern 0.5, 1, 2 und 5.

eintritt. Die Varianz von T ist gegeben durch $Var(T) = \lambda^2$, was im Mittel eine größere Variation der Ereigniszeiten bedeutet, je größer die zu erwartende Ereigniszeit ist.

(Vergleiche dazu Collett (2003), S. 152 f.)

3. Parameterschätzung

In diesem Kapitel werden grundlegende Konzepte für die Parameterschätzung bei rechtszensierten Ereigniszeitdaten vorgestellt. In Abschnitt 3.1 wird zunächst die Maximum-Likelihood-Methode beschrieben, die nicht nur bei zensierten Daten als klassischer Ansatz für die Schätzung des Parameters einer Verteilung gilt. Im Rahmen der Maximum-Likelihood-Methode wird die sogenannte Likelihood-Funktion maximiert. Die Konstruktion der Likelihood-Funktion bei zensierten Daten weist jedoch die Besonderheit auf, dass bedingte Verteilungen zu berücksichtigen sind. Darauf wird in Abschnitt 3.2 eingegangen. Eine Alternative zur Maximierung der Likelihood-Funktion stellt die Maximierung einer getrimmten Version der Likelihood-Funktion dar. Das Prinzip der Getrimmte-Likelihood-Schätzung wird in Abschnitt 3.3 dargestellt. Schließlich werden in Abschnitt 3.4 der mittlere quadratische Fehler und Effizienz als Vergleichskriterien für Parameterschätzer eingeführt.

3.1. Maximum-Likelihood-Methode

Die Maximum-Likelihood-Methode ist ein Verfahren zur Konstruktion von Schätzern für den oder die Parameter einer Verteilung. Die Idee ist es herauszufinden, welcher Parameterwert bzw. welche Parameterwerte unter den realisierten Daten am plausibelsten erscheinen.

Sei X eine Zufallsvariable mit Dichtefunktion $f_\theta(x)$, wobei θ den Parameter der Verteilung von X bezeichne. $X_1, \ldots, X_n$ seien unabhängige Ausprägungen von X mit identischer Dichtefunktion $f_\theta(x_i)$, $i = 1, \ldots, n$. Die gemeinsame Dichtefunktion von $X_1, \ldots, X_n$ ist dann gegeben durch

$$f_\theta(x_1, \ldots, x_n) = f_\theta(x_1) \cdot \ldots \cdot f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Für einen festen Parameter θ wird die gemeinsame Dichtefunktion $f_\theta(x_1, \ldots, x_n)$ als eine Funktion der Daten $x_1, \ldots, x_n$ aufgefasst, die als zufällige Realisationen von $X_1, \ldots, X_n$ angesehen werden. Werden die Rollen des Parameters θ und der Daten $x_1, \ldots, x_n$ vertauscht, so erhält man die *Likelihood-Funktion*

$$L_{x_1,\ldots,x_n}(\theta) = f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} L_{x_i}(\theta).$$

Die Likelihood-Funktion ist eine Funktion des Parameters θ für feste Realisationen $x_1, \ldots, x_n$. Sie gibt für jeden Parameterwert θ an, wie plausibel das Zustandekommen der beobachteten Daten $x_1, \ldots, x_n$ ist.

Die Maximum-Likelihood-Methode zur Konstruktion eines Schätzers für den Parameter θ beruht auf der Maximierung der Likelihood-Funktion. Zu den Realisationen $x_1, \ldots, x_n$ wird somit derjenige Parameterwert $\hat{\theta}$ gewählt, für den es am plausibelsten erscheint, dass gerade die realisierten Daten $x_1, \ldots, x_n$ auftreten. Der Wert $\hat{\theta}_{ML}$, für den die Likelihood-Funktion maximal ist, d.h.

$$L_{x_1,\ldots,x_n}(\hat{\theta}_{ML}) = \max_{\theta} \; L_{x_1,\ldots,x_n}(\theta),$$

wird *Maximum-Likelihood-Schätzung* für θ genannt.

Das Maximum der Likelihood-Funktion wird durch Ableiten und Nullsetzen der Ableitung bestimmt. Ob es sich bei der Lösung tatsächlich um ein Maximum und kein Minimum handelt, lässt sich mit Hife der zweiten Ableitung überprüfen. Ein Maximum liegt vor, wenn die zweite Ableitung der Likelihood-Funktion kleiner als Null ist.

Wegen der Produkte in $L_{x_1,\ldots,x_n}(\theta)$ ist es oftmals einfacher, anstelle der Likelihood-Funktion die so genannte *Log-Likelihood-Funktion*

$$l_{x_1,\ldots,x_n}(\theta) = \ln L_{x_1,\ldots,x_n}(\theta) = \ln \prod_{i=1}^{n} f_\theta(x_i) = \sum_{i=1}^{n} \ln f_\theta(x_i) = \sum_{i=1}^{n} l_{x_i}(\theta).$$

zu maximieren. Aufgrund der strengen Monotonie des Logarithmus haben $L_{x_1,\ldots,x_n}(\theta)$ und $l_{x_1,\ldots,x_n}(\theta)$ bei festem $x_1, \ldots, x_n$ die gleichen Maximalstellen. (Vergleiche dazu Genschel und Becker (2005), S. 115 ff.)

3.2. Likelihood-Funktion bei rechtszensierten Daten

Rechtszensierte Ereigniszeitdaten sind eine Mischung aus Ereigniszeiten und Zensierungszeiten. Beim Aufstellen der Likelihood-Funktion ist folglich eine bedingte Verteilung heranzuziehen, wobei zwischen zensierten und unzensierten Beobachtungen zu unterscheiden ist.

Sei $X_1, \ldots, X_n$ eine Zufallsstichprobe von rechtszensierten Ereigniszeiten. Die Ereigniszeit des i-ten Untersuchungsobjekts, $i = 1, \ldots, n$, sei mit T_i und die Zensierungszeit, die als zufällig angenommen wird, mit C_i bezeichnet. Beobachtet werden das Minimum X_i der Ereigniszeit T_i und der Zensierungszeit C_i sowie ein Ereignisindikator Δ_i, der genau dann den Wert 1 annimmt, wenn das Ereignis eingetreten ist, und 0, wenn die Ereigniszeit zensiert wurde. Für jedes Untersuchungsobjekt werde die Realisierung (x_i, δ_i) erfasst.

Die Dichte- und die Survivalfunktion der Ereigniszeiten seien mit $f_\theta(x_i)$ bzw. $S_\theta(x_i)$ und die der Zensierungszeiten mit $f_C(x_i)$ bzw. $S_C(x_i)$ bezeichnet. Die Ereigniszeiten und die Zensierungszeiten seien stochastisch unabhängig und die Zensierung sei nicht-informativ. Die Verteilung der Zensierungszeiten enthalte also keine Informationen über die Verteilung der Ereigniszeiten.

Der Beitrag einer unzensierten Realisierung $(x_i, \delta_i = 1)$ zur Likelihood-Funktion ergibt sich zu

$$
\begin{aligned}
L_{x_i,\, \delta_i=1}(\theta) &= \lim_{\Delta x_i \to 0} \frac{P(x_i \leq X_i < x_i + \Delta x_i,\, \delta_i = 1)}{\Delta x_i} \\
&= \lim_{\Delta x_i \to 0} \frac{P(x_i \leq T_i < x_i + \Delta x_i,\, C_i > x_i)}{\Delta x_i} \\
&= \lim_{\Delta x_i \to 0} \frac{P(x_i \leq T_i < x_i + \Delta x_i)}{\Delta x_i}\, P(C_i > x_i) \\
&= f_\theta(x_i)\, S_C(x_i)\,.
\end{aligned}
$$

Entsprechend ergibt sich der Beitrag einer zensierten Beobachtung $(x_i, \delta_i = 0)$ zur Likelihood-Funktion zu

$$
\begin{aligned}
L_{x_i,\,\delta_i=0}(\theta) &= \lim_{\Delta x_i \to 0} \frac{P(x_i \le X_i < x_i + \Delta x_i,\, \delta_i = 0)}{\Delta x_i} \\
&= \lim_{\Delta x_i \to 0} \frac{P(x_i \le C_i < x_i + \Delta x_i,\, T_i > x_i)}{\Delta x_i} \\
&= \lim_{\Delta x_i \to 0} \frac{P(x_i \le C_i < x_i + \Delta x_i)}{\Delta x_i}\, P(T_i > x_i) \\
&= f_C(x_i)\, S_\theta(x_i)\,.
\end{aligned}
$$

Insgesamt ist die Likelihood-Funktion rechtszensierter Ereigniszeitdaten also gegeben durch

$$
\begin{aligned}
L_{x_1,\ldots,x_n}(\theta) &= \prod_{i=1}^{n} \big(f_\theta(x_i)\; S_C(x_i)\big)^{\delta_i}\; \big(f_C(x_i)\; S_\theta(x_i)\big)^{1-\delta_i} \\
&= \prod_{i=1}^{n} f_\theta(x_i)^{\delta_i}\, S_\theta(x_i)^{1-\delta_i} \prod_{i=1}^{n} S_C(x_i)^{\delta_i}\, f_C(x_i)^{1-\delta_i}\,. \qquad (3.1)
\end{aligned}
$$

Aufgrund der Annahme nicht-informativer Zensierung ist es ausreichend, im Rahmen des Maximum-Likelihood-Ansatzes den ersten Term in (3.1) zu maximieren. Der zweite Term hängt nicht von dem Parameter der Ereigniszeitverteilung ab und kann daher als Konstante angesehen werden. Der relevante Teil der Likelihood-Funktion ist also gegeben durch

$$
L_{x_1,\ldots,x_n}(\theta) = \prod_{i=1}^{n} f_\theta(x_i)^{\delta_i}\, S_\theta(x_i)^{1-\delta_i}
$$

und die zugehörige Log-Likelihood-Funktion durch

$$
\begin{aligned}
l_{x_1,\ldots,x_n}(\theta) &= \ln\Big(\prod_{i=1}^{n} f_\theta(x_i)^{\delta_i}\, S_\theta(x_i)^{1-\delta_i}\Big) \\
&= \sum_{i=1}^{n} \big(\delta_i\, \ln f_\theta(x_i) + (1-\delta_i)\, \ln S_\theta(x_i)\big)\,.
\end{aligned}
$$

Äquivalent sind die Likelihood-Funktion und die Log-Likelihood-Funktion auch darstellbar durch

$$L_{x_1,\ldots,x_n}(\theta) = \prod_{i=1}^{n} f_\theta(x_i)^{\delta_i}\, S_\theta(x_i)^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \Big(\frac{f_\theta(x_i)}{S_\theta(x_i)}\Big)^{\delta_i} S_\theta(x_i)$$

$$= \prod_{i=1}^{n} h_\theta(x_i)^{\delta_i}\, S_\theta(x_i) \tag{3.2}$$

bzw.

$$l_{x_1,\ldots,x_n}(\theta) = \ln\Big(\prod_{i=1}^{n} h_\theta(x_i)^{\delta_i}\, S_\theta(x_i)\Big)$$

$$= \sum_{i=1}^{n} \big(\delta_i\, \ln h_\theta(x_i) + \ln S_\theta(x_i)\big)\,. \tag{3.3}$$

Werden nicht zufällig, sondern mit einer festen Zensierungszeit rechtszensierte Ereigniszeiten betrachtet, so ergibt sich die gleiche Likelihood-Funktion.

(Vergleiche dazu Collett (2003), S. 357 f.)

3.3. Getrimmte-Likelihood-Schätzung

Sei X eine Zufallsvariable mit Dichtefunktion $f_\theta(x)$ und Log-Likelihood-Funktion $l_x(\theta) = \ln f_\theta(x)$. Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängige Ausprägungen von X, deren Realisierungen mit $x_1, \ldots, x_n$ bezeichnet seien. Die Log-Likelihood-Funktion von $X_1, \ldots, X_n$ ist gegeben durch

$$l_{x_1,\ldots,x_n}(\theta) = \sum_{i=1}^{n} \ln f_\theta(x_i) = \sum_{i=1}^{n} l_{x_i}(\theta),$$

wobei $l_{x_i}(\theta)$ für den Beitrag der i-ten Beobachtung zur Log-Likelihood-Funktion steht.

Der *Getrimmte-Likelihood-Schätzer* maximiert eine getrimmte Version der Likelihood-Funktion, bei der die am wenigsten wahrscheinlichen Beobachtungen, also die Beobachtungen mit den kleinsten Beiträgen zur Likelihood-Funktion, getrimmt wurden. Aufgrund der strengen Monotonie des Logarithmus sind dies gerade diejenigen Beobachtungen mit den kleinsten Beiträgen $l_{x_i}(\theta)$ zur Log-Likelihood-Funktion.

Der Getrimmte-Likelihood-Schätzer ist gegeben durch

$$\hat{\theta}_{TL} = \arg\max_{\theta} \sum_{i=n-r+1}^{n} l_{(i)}(\theta).$$

Dabei bezeichnen $l_{(1)}(\theta) \leq \cdots \leq l_{(n)}(\theta)$ die für ein gegebenes θ geordneten Beiträge der Beobachtungen zur Log-Likelihood-Funktion und $n - r$ Beobachtungen wurden getrimmt.

(Vergleiche dazu Müller und Neykov (2003).)

3.4. Mittlerer quadratischer Fehler und Effizienz

Als Vergleichskriterium für Punktschätzer kann der *mittlere quadratische Fehler*, kurz MSE für *mean squared error*, herangezogen werden. Er ist definiert als der erwartete quadrierte Abstand des Punktschätzers $\hat{\theta}$ vom zu schätzenden Parameter θ:

$$MSE(\hat{\theta}) = \mathbb{E}\big((\hat{\theta} - \theta)^2\big).$$

Der mittlere quadratische Fehler kann zerlegt werden in die Summe aus der Varianz des Schätzers und dem quadrierten Bias,

$$MSE = \text{Varianz} + \text{Bias}^2,$$

wobei der *Bias*, auch *Verzerrung* genannt, die mittlere Abweichung eines Punktschätzers vom wahren Parameter beschreibt:

$$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

Die Zerlegung des mittleren quadratischen Fehlers erfolgt zu

$$\begin{aligned}
MSE(\hat{\theta}) &= \mathbb{E}\big([\hat{\theta} - \theta]^2\big) \\
&= \mathbb{E}\big([\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta]^2\big) \\
&= \mathbb{E}\big([\hat{\theta} - \mathbb{E}(\hat{\theta})]^2\big) + 2\,\mathbb{E}\big([\hat{\theta} - \mathbb{E}(\hat{\theta})][\mathbb{E}(\hat{\theta}) - \theta]\big) + \mathbb{E}\big([\mathbb{E}(\hat{\theta}) - \theta]^2\big) \\
&= \mathbb{E}\big([\hat{\theta} - \mathbb{E}(\hat{\theta})]^2\big) + \big[\mathbb{E}(\hat{\theta}) - \theta\big]^2 \\
&= Var(\hat{\theta}) + Bias(\hat{\theta})^2.
\end{aligned}$$

Schätzer mit einem kleinen mittleren quadratischen Fehler sind vorzuziehen, was die Forderung nach einer kleinen Verzerrung mit der nach einer kleinen Varianz des Schätzers kombiniert. (Vergleiche dazu Genschel und Becker (2005), S. 66 und 71 f. und Fahrmeir et al. (2006), S. 370.)

Ein Schätzer $\hat{\theta}_1$ für einen Parameter θ heißt *MSE-effizienter* als ein Schätzer $\hat{\theta}_2$, falls für alle θ gilt

$$MSE(\hat{\theta}_1) \leq MSE(\hat{\theta}_2)\,.$$

Werden nur unverzerrte Schätzer, also Schätzer mit einem Bias von Null betrachtet, so reduziert sich der Vergleich der mittleren quadratischen Fehler auf den Vergleich ihrer Varianzen. Von zwei unverzerrten Schätzern $\hat{\theta}_1$ und $\hat{\theta}_2$ heißt $\hat{\theta}_1$ *effizienter* als $\hat{\theta}_2$, wenn für alle θ gilt

$$Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)\,.$$

Im Mittel liefert $\hat{\theta}_1$ dann genauere Schätzwerte als $\hat{\theta}_2$, da die Schätzwerte von $\hat{\theta}_1$ weniger um den wahren Parameterwert θ streuen.

(Vergleiche dazu Genschel und Becker (2005), S. 76 f.)

4. Schätzer für den Parameter der Exponentialverteilung

Das Hauptaugenmerk der vorliegenden Arbeit ist der Vergleich verschiedener getrimmter Schätzungen für den Parameter der Exponentialverteilung bei rechtszensierten Ereigniszeitdaten. In diesem Kapitel werden die ausgewählten Schätzverfahren vorgestellt.

Ein klassischer Ansatz für die Schätzung des Parameters einer Verteilung ist die Anwendung der Maximum-Likelihood-Methode. Der Maximum-Likelihood-Schätzer für das zugrundeliegende Modell wird in Abschnitt 4.1 hergeleitet. Während bei der Maximum-Likelihood-Methode die komplette Likelihood-Funktion maximiert wird, wird bei dem Getrimmte-Likelihood-Prinzip nur eine getrimmte Version der Likelihood-Funktion in die Maximierung einbezogen. Die Anwendung der Getrimmte-Likelihood-Schätzmethode auf exponentialverteilte Ereigniszeitdaten wird in Abschnitt 4.2 dargelegt. Für exponentialverteilte Ereigniszeiten, die mit einer festen Zensierungszeit rechtszensiert wurden, schlagen Clarke et al. (vgl. Clarke et al. (2014)) zwei Modifikationen des Maximum-Likelihood-Schätzers vor, die effizienter als dieser sein sollen. Diese sogenannten Pseudo-Maximum-Likelihood-Schätzer werden in Abschnitt 4.3 vorgestellt. Zuvor wird die Notation eingeführt, die in diesem Kapitel Verwendung findet.

Sei T die Zeit bis zum Eintreten eines definierten Ereignisses. T sei exponentialverteilt mit Dichtefunktion $f_\lambda(t) = {}^1\!/\!_\lambda \cdot \mathrm{e}^{-t/\lambda}$ für $t > 0$ und $\lambda > 0$. Survivalfunktion und Hazardrate von T sind dann gegeben durch

$$S_\lambda(t) = \mathrm{e}^{-t/\lambda}$$

$$h_\lambda(t) = \frac{1}{\lambda}$$

$T_1, \ldots, T_n$ sei eine Zufallstichprobe aus dieser Ereigniszeitverteilung, deren Realisierungen mit $t_1, \ldots, t_n$ bezeichnet werden. Die Ereigniszeiten seien rechtszensiert mit einer festen Zensierungszeit C. Beobachtet werde dann das Minimum X der Ereigniszeit T und der Zensierungszeit C

$$X_i = \min(T_i, C) \quad \text{für } i = 1, \ldots, n$$

sowie ein Ereignisindikator Δ, der den Wert 1 annimmt, wenn X eine Ereigniszeit ist und 0, wenn die Ereigniszeit zensiert wurde:

$$\Delta_i = \begin{cases} 1 & \text{wenn } T_i \leq C \text{ (Ereigniszeit unzensiert)} \\ 0 & \text{wenn } T_i > C \text{ (Ereigniszeit zensiert)} \end{cases}$$

$$= \mathbb{1}(T_i \leq C)$$

für $i = 1, \ldots, n$, wobei $\mathbb{1}$ die Indikatorfunktion bezeichne. Für jede Realisierung werde dann das Paar (x_i, δ_i) erfasst.

Wird anstelle einer Rechtszensierung mit fester Zensierungszeit eine zufällige Rechtszensierung der Ereigniszeiten angenommen, so seien die Zufallsstichprobe der Zensierungszeiten mit $C_1, \ldots, C_n$ und deren Realisierungen mit $c_1, \ldots, c_n$ bezeichnet. Die Zensierungszeiten seien unabhängig und identisch verteilt und stochastisch unabhängig von den Ereigniszeiten $T_1, \ldots, T_n$. Beobachtet werde dann

$$X_i = \min(T_i, C_i)$$

und

$$\Delta_i = \begin{cases} 1 & \text{wenn } T_i \leq C_i \text{ (Ereigniszeit unzensiert)} \\ 0 & \text{wenn } T_i > C_i \text{ (Ereigniszeit zensiert)} \end{cases}$$

$$= \mathbb{1}(T_i \leq C_i)$$

mit den Realisierungen (x_i, δ_i) für $i = 1, \ldots, n$.

4.1. Maximum-Likelihood-Schätzer

Ein klassischer Ansatz, ein parametrisches Modell an beobachtete Ereigniszeitdaten anzupassen, besteht in der Anwendung der Maximum-Likelihood-Methode, die

in Kapitel 3.1 vorgestellt wurde. Bei rechtszensierten Ereigniszeitdaten, die einer Exponentialverteilung folgen, ist die Likelihood-Funktion gegeben durch

$$\begin{aligned} L_{x_1,\ldots,x_n}(\lambda) &= \prod_{i=1}^{n} h_\lambda(x_i)^{\delta_i}\, S_\lambda(x_i) \\ &= \prod_{i=1}^{n} \left(\frac{1}{\lambda}\right)^{\delta_i} \mathrm{e}^{-x_i/\lambda} \end{aligned}$$

und die Log-Likelihood-Funktion durch

$$l_{x_1,\ldots,x_n}(\lambda) = \ln\, L_{x_1,\ldots,x_n}(\lambda) = -\sum_{i=1}^{n} \delta_i \ln(\lambda) - \frac{1}{\lambda}\sum_{i=1}^{n} x_i. \tag{4.1}$$

Die erste Ableitung der Log-Likelihood-Funktion nach λ ergibt sich zu

$$\frac{\mathrm{d}\, l_{x_1,\ldots,x_n}(\lambda)}{\mathrm{d}\lambda} = -\frac{1}{\lambda}\sum_{i=1}^{n} \delta_i + \frac{1}{\lambda^2}\sum_{i=1}^{n} x_i.$$

Nullsetzen und Auflösen nach λ führt zum Maximum-Likelihood-Schätzer

$$\hat{\lambda}_{mle} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} \delta_i}\,. \tag{4.2}$$

Die zweite Ableitung der Log-Likelihood-Funktion nach λ ist gegeben durch

$$\frac{\mathrm{d}^2\, l_{x_1,\ldots,x_n}(\lambda)}{\mathrm{d}\lambda^2} = \frac{1}{\lambda^2}\sum_{i=1}^{n} \delta_i - \frac{2}{\lambda^3}\sum_{i=1}^{n} x_i.$$

Einsetzen von $\hat{\lambda}_{mle}$ in die zweite Ableitung führt zu

$$\begin{aligned} \frac{1}{\hat{\lambda}_{mle}^2}\sum_{i=1}^{n} \delta_i - \frac{2}{\hat{\lambda}_{mle}^3}\sum_{i=1}^{n} x_i &= \frac{\left(\sum_{i=1}^{n} \delta_i\right)^3}{\left(\sum_{i=1}^{n} x_i\right)^2} - \frac{2\left(\sum_{i=1}^{n} \delta_i\right)^3}{\left(\sum_{i=1}^{n} x_i\right)^2} \\ &= -\frac{\sum_{i=1}^{n} \delta_i}{\left(\sum_{i=1}^{n} x_i\right)^2} < 0\,. \end{aligned}$$

Bei der in (4.2) ausgewiesenen Lösung handelt es sich also tatsächlich um ein Maximum der Likelihood-Funktion.

Der Maximum-Likelihood-Schätzer $\hat{\lambda}_{mle}$ ist hier gleich der Summe der beobachteten Ereigniszeiten dividiert durch die Anzahl der beobachteten Ereignisse. Er hat somit den Charakter einer mittleren Ereigniszeit zensierter Ereigniszeitdaten.

Die in Kapitel 3.2 hergeleitete Likelihood-Funktion bei rechtszensierten Daten gilt gleichermaßen für Studien mit fester Zensierungsschranke und Studien mit zufälliger Zensierung. Der Maximum-Likelihood-Schätzer ist somit für beide Zensierungsmodelle gleich.

(Vergleiche dazu Collett (2003) S. 160 f.)

4.2. Getrimmte-Likelihood-Schätzer

Der Getrimmte-Likelihood-Schätzer, der in Kapitel 3.3 beschrieben wurde, ist bei rechtszensierten Ereigniszeitdaten, die einer Exponentialverteilung folgen, gegeben durch

$$\hat{\lambda}_{tle} = \arg\max_{\lambda} \sum_{i=n-r+1}^{n} l_{(i)}(\lambda). \tag{4.3}$$

Dabei bezeichnen $l_{(1)}(\lambda) \leq \ldots \leq l_{(n)}(\lambda)$ die für ein gegebenes λ geordneten Beiträge der Beobachtungen zur Log-Likelihood-Funktion, während sich der Beitrag einer einzelnen Beobachtung aus (4.1) zu

$$l_{x_i}(\lambda) = -\delta_i \ln(\lambda) - \frac{1}{\lambda} x_i$$

ergibt und die $n - r$ Beobachtungen mit den kleinsten Beiträgen getrimmt wurden.

Für $r = n$ entspricht der Getrimmte-Likelihood-Schätzer dem Maximum-Likelihood-Schätzer aus Kapitel 4.1.

Für $r < n$ kann die Getrimmte-Likelihood-Schätzung berechnet werden, indem alle r-elementigen Teilmengen $M = \{i_1, \ldots, i_r\}$ von $\{1, \ldots, n\}$ und damit alle Teildatensätze $(x, \delta)(M) := ((x_{i_1}, \delta_{i_1}), \ldots, (x_{i_r}, \delta_{i_r}))^T$ von $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ gebildet und die zugehörigen Log-Likelihood-Funktionen

$$l_{(x,\delta)(M)}(\lambda) = -\sum_{j=1}^{r} \delta_{i_j} \ln(\lambda) - \frac{1}{\lambda} \sum_{j=1}^{r} x_{i_j}$$

maximiert werden. Die Getrimmte-Likelihood-Schätzung $\hat{\lambda}_{tle}$ ist dann die Maximum-Likelihood-Schätzung desjenigen Teildatensatzes, für den $l_{(x,\delta)(M)}$ maximal ist.

Das Betrachten aller r-elementigen Teilmengen von $\{1, \ldots, n\}$ ist sehr aufwendig. Es lässt sich jedoch zeigen, dass viele Teildatensätze von der Berechnung ausgenommen werden können, da sie nicht zu einem maximalen Wert der getrimmten Log-Likelihood-Funktion gemäß (4.3) führen werden.

Seien $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ Realisierungen von rechtszensierten Ereigniszeitdaten mit fester Zensierungszeit.

Behauptung: Seien M_1 und M_2 zwei r-elementige Teilmengen von $\{1, \ldots, n\}$. Die Anzahl der zensierten Beobachtungen sei kleiner als r und die Teildatensätze

$$(x, \delta)(M_1) = ((x^1_{i_1}, \delta^1_{i_1}), \ldots, (x^1_{i_r}, \delta^1_{i_r}))^T \text{ und}$$

$$(x, \delta)(M_2) = ((x^2_{i_1}, \delta^2_{i_1}), \ldots, (x^2_{i_r}, \delta^2_{i_r}))^T$$

enthalten jeweils die gleiche Anzahl an unzensierten sowie alle zensierten Beobachtungen. Für die geordneten Ereigniszeitbeobachtungen gelte $x^1_{(i_j)} \leq x^2_{(i_j)}$ für alle $j = 1, \ldots, r$ und $x^1_{(i_j)} < x^2_{(i_j)}$ für mindestens ein j. Für die getrimmte Log-Likelihood-Funktion ausgewertet an der Stelle der jeweiligen Maximum-Likelihood-Schätzung gilt dann

$$l_{(x,\delta)(M_1)}(\hat{\lambda}^1_{mle}) \; > \; l_{(x,\delta)(M_2)}(\hat{\lambda}^2_{mle})\,.$$

Beweis: Für die Maximum-Likelihood-Schätzungen $\hat{\lambda}^1_{mle}$ und $\hat{\lambda}^2_{mle}$ der Teildatensätze $(x, \delta)(M_1)$ bzw. $(x, \delta)(M_2)$ gilt

$$\hat{\lambda}^1_{mle} = \frac{\sum_{i=1}^n x_i - \sum_{i_j \notin M_1} t_{i_j}}{\sum_{i=1}^n \delta_i - (n-r)} \; < \; \frac{\sum_{i=1}^n x_i - \sum_{i_j \notin M_2} t_{i_j}}{\sum_{i=1}^n \delta_i - (n-r)} = \hat{\lambda}^2_{mle}\,,$$

da bei der Maximierung der getrimmten Log-Likelihood-Funktionen jeweils $n - r$ unzensierte Beboachtungen nicht berücksichtigt werden und die Summe der nicht berücksichtigten Ereigniszeitdaten für Teildatensatz $(x, \delta)(M_1)$ größer ist als für Teildatensatz $(x, \delta)(M_2)$. Daher gilt:

$$
\begin{aligned}
& l_{(x,\delta)(M_1)}(\hat{\lambda}^1_{mle}) \; - \; l_{(x,\delta)(M_2)}(\hat{\lambda}^2_{mle}) \\
&= -\ln(\hat{\lambda}^1_{mle})\Big(\sum_{i=1}^{n}\delta_i - (n-r)\Big) - \frac{1}{\hat{\lambda}^1_{mle}}\Big(\sum_{i=1}^{n} x_i - \sum_{i_j \notin M_1} t_{i_j}\Big) \\
&\quad - \Big(- \ln(\hat{\lambda}^2_{mle})\Big(\sum_{i=1}^{n}\delta_i - (n-r)\Big) - \frac{1}{\hat{\lambda}^2_{mle}}\Big(\sum_{i=1}^{n} x_i - \sum_{i_j \notin M_2} t_{i_j}\Big)\Big) \\
&= \big(\ln(\hat{\lambda}^2_{mle}) - \ln(\hat{\lambda}^1_{mle})\big)\Big(\sum_{i=1}^{n}\delta_i - (n-r)\Big) \\
&\quad + \Big(\sum_{i=1}^{n}\delta_i - (n-r)\Big) - \Big(\sum_{i=1}^{n}\delta_i - (n-r)\Big) \\
&= \underbrace{\ln\Big(\underbrace{\frac{\hat{\lambda}^2_{mle}}{\hat{\lambda}^1_{mle}}}_{>1}\Big)}_{>0}\;\underbrace{\Big(\sum_{i=1}^{n}\delta_i - (n-r)\Big)}_{>0} \; > \; 0.
\end{aligned}
$$

Daraus folgt die Behauptung. □

Sind $n-r$ Beobachtungen zu trimmen und es liegen mehr als $n-r$ unzensierte Realisierungen vor, so genügt es also, nur solche r-elementigen Teildatensätze für die Bestimmung des Getrimmte-Likelihood-Schätzers zu betrachten, die entweder nur zensierte oder bis zu $n-r$ der größten unzensierten Realisierungen nicht enthalten. Desweiteren haben alle zensierten Realisierungen den gleichen Wert. r-elementige Teilmengen von $\{1,\ldots,n\}$, die eine oder mehrere zensierte Realisierungen nicht enthalten, können daher zu identischen Teildatensätzen führen, was die Menge der zu berücksichtigenden Teildatensätze weiter reduziert.

Seien nun $(x_1,\delta_1),\ldots,(x_n,\delta_n)$ Realisierungen von zufällig rechtszensierten Ereigniszeitdaten. Auch hier sind nur Teildatensätze zu betrachten, die bis zu $n-r$ der größten unzensierten Realisierungen nicht enthalten. Der Beweis kann analog zum Fall rechtszensierter Daten mit fester Zensierungszeit erbracht werden. Zusätzlich lässt sich zeigen, dass bei Vorliegen von mehr als $n-r$ zensierten Realisierungen nur solche Teildatensätze zur Getrimmte-Likelihood-Schätzung führen können, die bis zu $n-r$ der größten zensierten Realisierungen nicht beinhalten.

Behauptung: Seien M_1 und M_2 zwei r-elementige Teilmengen von $\{1,\dots,n\}$. Die Anzahl der unzensierten Beboachtungen sei kleiner als r und die Teildatensätze

$$(x,\delta)(M_1) = ((x^1_{i_1},\delta^1_{i_1}),\dots,(x^1_{i_r},\delta^1_{i_r}))^T \text{ und}$$
$$(x,\delta)(M_2) = ((x^2_{i_1},\delta^2_{i_1}),\dots,(x^2_{i_r},\delta^2_{i_r}))^T$$

enthalten jeweils die gleiche Anzahl an zensierten sowie alle unzensierten Beobachtungen. Für die geordneten Ereigniszeitbeobachtungen gelte $x^1_{(i_j)} \le x^2_{(i_j)}$ für alle $j = 1,\dots,r$ und $x^1_{(i_j)} < x^2_{(i_j)}$ für mindestens ein j. Für die getrimmte Log-Likelihood-Funktion ausgewertet an der Stelle der jeweiligen Maximum-Likelihood-Schätzung gilt dann

$$l_{(x,\delta)(M_1)}(\hat\lambda^1_{mle}) \;>\; l_{(x,\delta)(M_2)}(\hat\lambda^2_{mle})\,.$$

Beweis: Für die Maximum-Likelihood-Schätzungen $\hat\lambda^1_{mle}$ und $\hat\lambda^2_{mle}$ der Teildatensätze $(x,\delta)(M_1)$ bzw. $(x,\delta)(M_2)$ gilt

$$\hat\lambda^1_{mle} = \frac{\sum_{i=1}^n x_i - \sum_{i_j\notin M_1} c_{i_j}}{\sum_{i=1}^n \delta_i} \;<\; \frac{\sum_{i=1}^n x_i - \sum_{i_j\notin M_2} c_{i_j}}{\sum_{i=1}^n \delta_i} = \hat\lambda^2_{mle}\,,$$

da bei der Maximierung der getrimmten Log-Likelihood-Funktionen die Summe der nicht berücksichtigten Ereigniszeitdaten für Teildatensatz $(x,\delta)(M_1)$ größer ist als für Teildatensatz $(x,\delta)(M_2)$. Es folgt:

$$\begin{aligned}
&l_{(x,\delta)(M_1)}(\hat\lambda^1_{mle}) \;-\; l_{(x,\delta)(M_2)}(\hat\lambda^2_{mle})\\
&= -\ln(\hat\lambda^1_{mle})\sum_{i=1}^n \delta_i - \frac{1}{\hat\lambda^1_{mle}}\Big(\sum_{i=1}^n x_i - \sum_{i_j\notin M_1} c_{i_j}\Big)\\
&\quad - \Big(-\ln(\hat\lambda^2_{mle})\sum_{i=1}^n \delta_i - \frac{1}{\hat\lambda^2_{mle}}\Big(\sum_{i=1}^n x_i - \sum_{i_j\notin M_2} c_{i_j}\Big)\Big)\\
&= \big(\ln(\hat\lambda^2_{mle}) - \ln(\hat\lambda^1_{mle})\big)\sum_{i=1}^n \delta_i + \sum_{i=1}^n \delta_i - \sum_{i=1}^n \delta_i\\
&= \underbrace{\ln\Big(\underbrace{\frac{\hat\lambda^2_{mle}}{\hat\lambda^1_{mle}}}_{>1}\Big)}_{>0}\;\underbrace{\sum_{i=1}^n \delta_i}_{>0} \;>\; 0.
\end{aligned}$$

Daraus folgt die Behauptung. □

Insgesamt lässt sich also der Aufwand zur Berechnung der Getrimmte-Likelihood-Schätzung sowohl bei rechtszensierten Ereigniszeitdaten mit fester Zensierungszeit als auch bei zufällig rechtszensierten Ereigniszeitdaten deutlich verringern, indem nur ausgewählte Teildatensätze betrachtet werden. Sollen $n-r$ Beobachtungen getrimmt werden, so sind nur solche Teildatensätze Kandidaten für den Teildatensatz, der zur Getrimmte-Likelihood-Schätzung gemäß (4.3) führt, die eine Kombination der $n-r$ größten unzensierten und der $n-r$ größten zensierten Realisierungen nicht enthalten.

Anstatt aller $\binom{n}{n-r}$ Teildatensätze $(x,\delta)(M) := ((x_{i_1},\delta_{i_1}),\ldots,(x_{i_r},\delta_{i_r}))^T$ der Realisierungen $(x_1,\delta_1),\ldots,(x_n,\delta_n)$ sind somit je nach vorliegender Datenkonstellation maximal $\binom{2(n-r)}{n-r}$ Teildatensätze zu bilden und die entsprechenden Log-Likelihood-Funktionen an der Stelle ihrer Maximum-Likelihood-Schätzungen auszuwerten.

4.3. Zwei Pseudo-Maximum-Likelihood-Schätzer

Bei rechtszensierten Ereigniszeiten ist das nach oben getrimmte Mittel ein gebräuchlicher Schätzer für den Parameter der Exponentialverteilung. Für die geordneten Ereigniszeiten $T_{(1)} \leq \ldots \leq T_{(n)}$ ist es gegeben durch

$$\overline{T}_{n,\beta} = \frac{1}{r}\sum_{i=1}^{r} T_{(i)}, \tag{4.4}$$

mit $r = n - \lfloor n\beta \rfloor$ die Anzahl der verbleibenden Beobachtungen, nachdem $\lfloor n\beta \rfloor$ Beobachtungen getrimmt wurden (vgl. Staudte und Sheather (1990), S. 28). Das nach oben getrimmte Mittel wird im Folgenden als *β-getrimmtes Mittel* bezeichnet.

Im Gegensatz zum Maximum-Likelihood-Schätzer, der in Kapitel 4.1 vorgestellt wurde, gilt das β-getrimmte Mittel in dem zugrundeliegenden Modell rechtszensierter Ereigniszeiten als robuster und effizienter Schätzer (vgl. Clarke et al. (2014)). Es kann jedoch nur dann berechnet werden, wenn mehr Beobachtungen zu trimmen sind als Beobachtungen zensiert wurden.

Clarke et al. stellen zwei Hybrid-Schätzer für den Parameter der Exponentialverteilung bei mit fester Zensierungszeit rechtszensierten Ereigniszeiten vor, die sich von dem Maximum-Likelihood-Schätzer und dem β-getrimmten Mittel ableiten (vgl.

Clarke et al. (2014)). Dazu werden beide Schätzer zunächst mittels statistischer Funktionale dargestellt.

Ein statistisches Funktional ist eine Abbildung, deren Definitionsbereich eine Funktionenmenge ist. Sei g eine reellwertige Funktion. Das Funktional $T[F] = \int g(t)\mathrm{d}F(t)$ ist dann definiert durch

$$T[F] := \sum_{i=1}^{\infty} g(t_i)p_i$$

mit p_i der Wahrscheinlichkeit an der Stelle t_i, wenn F eine diskrete Verteilung mit diskreter Dichte p ist, und durch

$$T[F] := \int g(t)f(t)\mathrm{d}t\,,$$

wenn F eine stetige Verteilung mit Dichte f ist (vgl. Staudte und Sheather (1990), S. 12 f.).

Sei $F_n(t)$ die empirische Verteilungsfunktion der Ereigniszeitdaten:

$$F_n(t) = \frac{\#T_i \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(T_i \leq t)\,.$$

Die Anzahl der unzensierten Beobachtungen n_{uc} kann somit geschrieben werden als

$$n_{uc} = \#\{T_i \leq C\} = nF_n(C)\,.$$

Der in Kapitel 4.1 hergeleitete Maximum-Likelihood-Schätzer $\hat{\lambda}_{mle}$ lässt sich dann überführen in

$$\begin{aligned}
\hat{\lambda}_{mle} &= \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} \delta_i} \\
&= \frac{\sum_{i=1}^{n} T_i\,\mathbb{1}(T_i \leq C) + (n - n_{uc})C}{n_{uc}} \\
&= \frac{n\left(\frac{1}{n}\sum_{i=1}^{n} T_i\,\mathbb{1}(T_i \leq C)\right) + (n - n_{uc})C}{n_{uc}} \\
&= \frac{n\int_0^C x\mathrm{d}F_n(x) + (n - nF_n(C))C}{nF_n(C)}\,.
\end{aligned} \tag{4.5}$$

Das β-getrimmte Mittel kann ebenfalls in funktionaler Form geschrieben werden. Ist F eine stetige Verteilungsfunktion mit dem $(1-\beta)$-Quantil $t_{1-\beta} = F^{-1}(1-\beta)$, so ist das β-getrimmte Mittel allgemein darstellbar durch

$$\begin{aligned} T_\beta[F] &= \mathbb{E}(T < t \,|\, T < t_{1-\beta}) \\ &= \frac{1}{1-\beta} \int_0^{t_{1-\beta}} t \mathrm{d}F(t) \end{aligned}$$

(vgl. Staudte und Sheather (1990), S. 53).

Das Funktional T_β angewendet auf die empirische Verteilungsfunktion F_n führt dann zum Schätzer für das β-getrimmte Mittel

$$T_\beta[F_n] = \overline{T}_{n,\beta} = \frac{1}{r} \sum_{i=1}^{r} T_{(i)}$$

mit $\lfloor n\beta \rfloor$ dem Anteil der getrimmten und $r = n - \lfloor n\beta \rfloor$ den verbleibenden Beobachtungen.

Da T hier exponentialverteilt ist mit Verteilungsfunktion $F_\lambda(t) = 1 - \mathrm{e}^{-t/\lambda}$, ist das $(1-\beta)$-Quantil $t_{1-\beta}$, das die Gleichung $F_\lambda(t_{1-\beta}) = 1 - \beta$ erfüllt, gegeben durch $t_{1-\beta} = -\lambda \ln(\beta)$ und das β-getrimmte Mittel ergibt sich zu

$$T_\beta[F_\lambda] = \frac{1}{1-\beta} \int_0^{-\lambda \ln(\beta)} t \frac{1}{\lambda} \mathrm{e}^{-t/\lambda} \mathrm{d}t \,.$$

Partielle Integration führt zu

$$\begin{aligned} T_\beta[F_\lambda] &= \frac{1}{1-\beta} \Big(\Big[-t\mathrm{e}^{-t/\lambda} \Big]_0^{-\lambda \ln(\beta)} - \int_0^{-\lambda \ln(\beta)} -\mathrm{e}^{-t/\lambda} \mathrm{d}t \Big) \\ &= \frac{1}{1-\beta} \Big(\lambda \ln(\beta)\beta - \Big[\lambda \mathrm{e}^{-t/\lambda} \Big]_0^{-\lambda \ln(\beta)} \Big) \\ &= \frac{1}{1-\beta} \big(\lambda \ln(\beta)\beta - (\lambda\beta - \lambda) \big) \\ &= \frac{1}{1-\beta} \big(1 - \beta + \beta \ln(\beta) \big) \lambda \,. \end{aligned}$$

Damit ist $T_\beta[F_\lambda]$ aber nicht Fisher-konsistent für λ, denn $T[F_\lambda] = \lambda$ ist nicht für alle λ erfüllt (vgl. Staudte und Sheather (1990), S. 50). Ein konsistenter Schätzer für den Parameter λ der Exponentialverteilung ist folglich gegeben durch

$$\hat{\lambda}_\beta = \frac{1-\beta}{1-\beta+\beta\ln(\beta)}\,\overline{T}_{n,\beta}\,. \tag{4.6}$$

Ersetzen von F_n durch $F_{\hat{\lambda}_\beta}$ in (4.5) führt schließlich zum Pseudo-Maximum-Likelihood-Schätzer

$$\hat{\lambda}_{pmle} = \frac{n\int_0^C x\mathrm{d}F_{\hat{\lambda}_\beta}(x) + (n - nF_{\hat{\lambda}_\beta}(C))C}{nF_{\hat{\lambda}_\beta}(C)}\,. \tag{4.7}$$

$\hat{\lambda}_\beta$ kann jedoch nur berechnet werden, wenn $F_\lambda^{-1}(1-\beta) \leq C$ ist. Clarke et al. schlagen daher vor, für $F_\lambda^{-1}(1-\beta) > C$ den Maximum-Likelihood-Schätzer $\hat{\lambda}_{mle}$ zu verwenden. Ein neuer Schätzer für den Parameter der Exponentialverteilung bei rechtszensierten Daten mit fester Zensierungszeit könnte also die folgende Form haben:

$$\hat{\lambda}_{pmle1} = \begin{cases} \hat{\lambda}_{pmle}, & \text{wenn } 1-\beta \leq F_n(C) \\ \hat{\lambda}_{mle}, & \text{wenn } 1-\beta > F_n(C) \end{cases} \tag{4.8}$$

Berechnen des Integrals in (4.7) mittels partieller Integration führt zu

$$\begin{aligned} \int_0^C x\mathrm{d}F_{\hat{\lambda}_\beta}(x) &= \int_0^C x\,\frac{1}{\hat{\lambda}_\beta}\mathrm{e}^{-x/\hat{\lambda}_\beta}\mathrm{d}x \\ &\quad \Big[x(1-\mathrm{e}^{-x/\hat{\lambda}_\beta})\Big]_0^C - \int_0^C 1-\mathrm{e}^{-x/\hat{\lambda}_\beta}\mathrm{d}x \\ &\quad = C - C\mathrm{e}^{-C/\hat{\lambda}_\beta} - \Big[x+\hat{\lambda}_\beta\mathrm{e}^{-x/\hat{\lambda}_\beta}\Big]_0^C \\ &\quad = C - C\mathrm{e}^{-C/\hat{\lambda}_\beta} - (C+\hat{\lambda}_\beta\mathrm{e}^{-C/\hat{\lambda}_\beta} - \hat{\lambda}_\beta) \\ &\quad = \hat{\lambda}_\beta - (C+\hat{\lambda}_\beta)\mathrm{e}^{-C/\hat{\lambda}_\beta} \end{aligned} \tag{4.9}$$

und Einsetzen in (4.7) ergibt

$$\hat{\lambda}_{pmle} = \frac{n\,(\hat{\lambda}_\beta - (C+\hat{\lambda}_\beta)\mathrm{e}^{-C/\hat{\lambda}_\beta}) + (n - nF_{\hat{\lambda}_\beta}(C))C}{nF_{\hat{\lambda}_\beta}(C)}\,.$$

Ersetzen von $F_{\hat{\lambda}_\beta}(C) = 1 - \mathrm{e}^{-C/\hat{\lambda}_\beta}$ führt schließlich zu

$$\begin{aligned}
\hat{\lambda}_{pmle} &= \frac{n\,\hat{\lambda}_\beta - n(C + \hat{\lambda}_\beta)\mathrm{e}^{-C/\hat{\lambda}_\beta} + \big(n - n(1 - \mathrm{e}^{-C/\hat{\lambda}_\beta})\big)C}{n(1 - \mathrm{e}^{-C/\hat{\lambda}_\beta})} \\
&= \frac{n\,\hat{\lambda}_\beta - nC\mathrm{e}^{-C/\hat{\lambda}_\beta} - n\hat{\lambda}_\beta\mathrm{e}^{-C/\hat{\lambda}_\beta} + nC\mathrm{e}^{-C/\hat{\lambda}_\beta}}{n(1 - \mathrm{e}^{-C/\hat{\lambda}_\beta})} \\
&= \frac{n\,\hat{\lambda}_\beta(1 - \mathrm{e}^{-C/\hat{\lambda}_\beta})}{n(1 - \mathrm{e}^{-C/\hat{\lambda}_\beta})} \\
&= \hat{\lambda}_\beta\,,
\end{aligned}$$

was genau dem konsistenten Schätzer $\hat{\lambda}_\beta$ in (4.6) enspricht.

Der erste Vorschlag für einen neuen Schätzer ist also definiert durch

$$\hat{\lambda}_{pmle1} = \begin{cases} \hat{\lambda}_\beta, & \text{wenn } 1 - \beta \leq F_n(C) \\ \hat{\lambda}_{mle}, & \text{wenn } 1 - \beta > F_n(C) \end{cases} .$$

Ein zweiter Vorschlag für die Herleitung eines neuen Schätzers ergibt sich bei erneuter Betrachtung des Integrals in (4.5)

$$\int_0^C t\mathrm{d}F_n(t) = \int_0^{F_n^{-1}(1-\beta)} t\mathrm{d}F_n(t) + \int_{F_n^{-1}(1-\beta)}^C t\mathrm{d}F_n(t)\,,$$

das wegen $(1-\beta)\mathrm{T}_\beta[F_n] = \int_0^{F_n^{-1}(1-\beta)} t\mathrm{d}F_n(t)$ zu

$$\int_0^C t\mathrm{d}F_n(t) = (1-\beta)\mathrm{T}_\beta[F_n] + \int_{F_n^{-1}(1-\beta)}^C t\mathrm{d}F_n(t) \tag{4.10}$$

führt. Das Integral auf der rechten Seite von (4.10) könnte bedingt durch große Beobachtungen instabil sein, daher wird F_n durch $F_{\hat{\lambda}_\beta}$ ersetzt. Dadurch und mittels partieller Integration ergibt sich

$$\begin{aligned}
\int_{F_n^{-1}(1-\beta)}^{C} t\mathrm{d}F_n(t) &\approx \int_{F_n^{-1}(1-\beta)}^{C} t\mathrm{d}F_{\hat{\lambda}_\beta}(t) \\
&= \int_{F_n^{-1}(1-\beta)}^{C} t\frac{1}{\hat{\lambda}_\beta}\mathrm{e}^{-t/\hat{\lambda}_\beta}\mathrm{d}t \\
&= \Big[-t\mathrm{e}^{-t/\hat{\lambda}_\beta}\Big]_{F_n^{-1}(1-\beta)}^{C} - \Big[\hat{\lambda}_\beta \mathrm{e}^{-t/\hat{\lambda}_\beta}\Big]_{F_n^{-1}(1-\beta)}^{C} \\
&= -C\mathrm{e}^{-C/\hat{\lambda}_\beta} + F_n^{-1}(1-\beta)\mathrm{e}^{-F_n^{-1}(1-\beta)/\hat{\lambda}_\beta} \\
&\quad - \hat{\lambda}_\beta \mathrm{e}^{-C/\hat{\lambda}_\beta} + \hat{\lambda}_\beta \mathrm{e}^{-F_n^{-1}(1-\beta)/\hat{\lambda}_\beta} \\
&= -(C+\hat{\lambda}_\beta)\mathrm{e}^{-C/\hat{\lambda}_\beta} + \big(F_n^{-1}(1-\beta) + \hat{\lambda}_\beta\big)\mathrm{e}^{-F_n^{-1}(1-\beta)/\hat{\lambda}_\beta}
\end{aligned}$$

Das Integral in (4.7) führt somit zu

$$\begin{aligned}
\int_0^C t\mathrm{d}F_n(t) &\approx (1-\beta)\mathrm{T}_\beta[F_n] - (C+\hat{\lambda}_\beta)\mathrm{e}^{-C/\hat{\lambda}_\beta} \\
&\quad + \big(F_n^{-1}(1-\beta) + \hat{\lambda}_\beta\big)\mathrm{e}^{-F_n^{-1}(1-\beta)/\hat{\lambda}_\beta} \\
&\equiv \mathrm{Correction}(F_n, \beta, C)\,.
\end{aligned}$$

Hieraus ergibt sich der Pseudo-korrigierte-Maximum-Likelihood-Schätzer

$$\hat{\lambda}_{pcmle} = \frac{n\,\mathrm{Correction}(F_n, \beta, C) + \big(n - nF_n(C)\big)C}{nF_n(C)}$$

und der zweite Vorschlag für einen neuen Schätzer ist definiert durch

$$\hat{\lambda}_{pmle2} = \begin{cases} \hat{\lambda}_{pcmle}, & \text{wenn } 1-\beta \leq F_n(C) \\ \hat{\lambda}_{mle}, & \text{wenn } 1-\beta > F_n(C) \end{cases}$$

(Vergleiche dazu Clarke et al. (2014).)

Die beiden Schätzer $\hat{\lambda}_{p1}$ und $\hat{\lambda}_{p2}$ für den Parameter der Exponentialverteilung wurden von Clarke et al. explizit für den Fall rechtszensierter Ereigniszeiten mit fester

Zensierungszeit hergeleitet. Eine Anwendung der Schätzer bei zufällig rechtszensierten Ereigniszeiten ist somit nicht sinnvoll. Die Annahme zufällig rechtszensierter Ereigniszeiten würde vielmehr eine Herleitung entsprechender Schätzer über bivariate Funktionale $T[F_\lambda, F_C]$ erforderlich machen, wobei F_λ die Verteilung der Ereigniszeiten und F_C die Verteilung der Zensierungszeiten ist.

5. Vergleich der Schätzer

Das Verhalten der in Kapitel 4 vorstellten Schätzer für den Parameter der Exponentialverteilung bei rechtszensierten Daten soll nun in einem kontaminierten Modell untersucht werden. Von besonderem Interesse ist dabei das Abschneiden der getrimmten Schätzungen im Vergleich zum klassischen Maximum-Likelihood-Ansatz.

In Abschnitt 5.1 werden für verschiedene Stichprobengrößen rechtszensierte Ereigniszeitdaten mit fester Zensierungszeit simuliert und alle vier Schätzer darauf angewendet. Die Zensierungszeiten werden dabei variiert und die Simulationsergebnisse in Abhängigkeit von den Zensierungszeiten ausgewiesen.

Analog dazu werden in Abschnitt 5.2, ebenfalls für verschiedene Stichprobenumfänge, zufällig rechtszensierte Ereigniszeitdaten simuliert. Dabei werden unterschiedliche Zensierungsanteile vorgegeben. Für die simulierten Daten werden dann die Maximum-Likelihood-Schätzung sowie die Getrimmte-Likelihood-Schätzung berechnet und in Abhängigkeit von den Zensierungsanteilen dargestellt.

Abschließend werden in Abschnitt 5.3 alle Schätzverfahren auf einen Datensatz aus einer klinischen Studie angewendet.

Alle Simulationen und Berechnungen wurden mit der Open Source Software R (vgl. R Core Team (2014)) in der Version 3.1.2 durchgeführt. Auch die graphischen Darstellungen der Simulationsergebnisse sowie alle Abbildungen in der gesamten Arbeit wurden mir R erstellt. Der R-Code sowie die Simulationsergebisse können dem beiliegenden Datenträger entnommen werden.

5.1. Simulation bei rechtszensierten Daten mit fester Zensierungszeit

Bei einer Rechtszensierung mit fester Zensierungszeit hängt der Anteil der zensierten Beobachtungen von der zugrundeliegenden Verteilung der Ereigniszeiten sowie von der gewählten Zensierungszeit ab. In der Simulation wird für die Verteilung der Ereigniszeiten ein kontaminiertes Modell mit Kontaminierung am oberen Rand der Verteilung angenommen. Sowohl die interessierenden Ereigniszeiten als auch die Ausreißer werden dabei als exponentialverteilt erachtet. Die simulierten Zufallszahlen stammen also aus einer Verteilung der Form

$$F(t) = (1 - \epsilon)F_{\lambda_1}(t) + \epsilon F_{\lambda_2}(t), \tag{5.1}$$

wobei ϵ gleich dem Anteil der Kontaminierung ist. F_{λ_1} bezeichnet die Verteilung der interessierenden Ereigniszeiten mit dem Erwartungswert λ_1 und F_{λ_2} die Exponentialverteilung der Ausreißer mit dem Erwartungswert λ_2 und es gilt $\lambda_1 < \lambda_2$. Neben den Parametern für die kontaminierte Verteilung ist noch ein Trimmungsanteil für die Berechnung der getrimmten Schätzwerte festzulegen.

In Anlehnung an Clarke et al. (2014), die in ihrem Arbeitspapier Simulationen zum Vergleich der von ihnen vorgeschlagenen Pseudo-Maximum-Likelihood-Schätzer mit dem Maximum-Likelihood-Schätzer vorstellen, werden die Parameter der interessierenden Ereigniszeitverteilung mit $\lambda_1 = 1$ und der Ereigniszeitverteilung der Ausreißer mit $\lambda_2 = 5$ bei einem Kontaminierungsanteil von $\epsilon = 0.05$ gewählt, sowie der Anteil der zu trimmenden Beobachtungen auf $\beta = 0.1$ festgesetzt (vgl. Clarke et al. (2014)).

Um das Verhalten der Schätzer für unterschiedlich große Stichproben vergleichen zu können, wurden alle gewählten Parameterkombinationen für einen kleinen Stichprobenumfang von 20 Untersuchungsobjekten, für einen mittleren Stichprobenumfang von 100 Untersuchungsobjekten und für einen großen Stichprobenumfang von 500 Untersuchungsobjekten simuliert. Für die betrachteten Szenarien wurden jeweils $m = 10.000$ Durchläufe generiert.

Als Vergleichskriterium für die Schätzer wird der mittlere quadratische Fehler (MSE) herangezogen, der neben der Varianz auch die Verzerrung der Schätzer einbezieht. Der MSE wird dabei durch $1/m \sum_{i=1}^{m} (\hat{\lambda}_i - \lambda)^2$ geschätzt, wobei $\hat{\lambda}_i$ den Schätzwert des i-ten Durchlaufs und λ den wahren Parameter im nicht kontaminierten Modell

bezeichnet. λ entspricht hier also dem Parameter λ_1 im kontaminierten Modell. Die Verzerrung wird mit $\bar{\hat{\lambda}} - \lambda$ geschätzt und die Varianz durch die empirische Stichprobenvarianz der Schätzwerte.

Die Ergebnisse der Simulationen sind für den Stichprobenumfang $n = 20$ in Abbildung 5.1, für den Stichprobenumfang $n = 100$ in Abbildung 5.2 und für den Stichprobenumfang $n = 500$ in Abbildung 5.3 ersichtlich. Für jedes Szenario besteht die Ergebnisdarstellung aus vier Grafiken. Diese umfassen die mittleren Schätzwerte der vier Schätzer, die geschätzten mittleren quadratischen Fehler sowie deren Komponenten, also die geschätzten Varianzen und die quadrierten Verzerrungen. Alle Schätzwerte werden in Abhängigkeit von der Zensierungszeit dargestellt, die einen Bereich zwischen dem Erwartungswert im nicht-kontaminierten Modell und sehr großen Werten abdeckt.

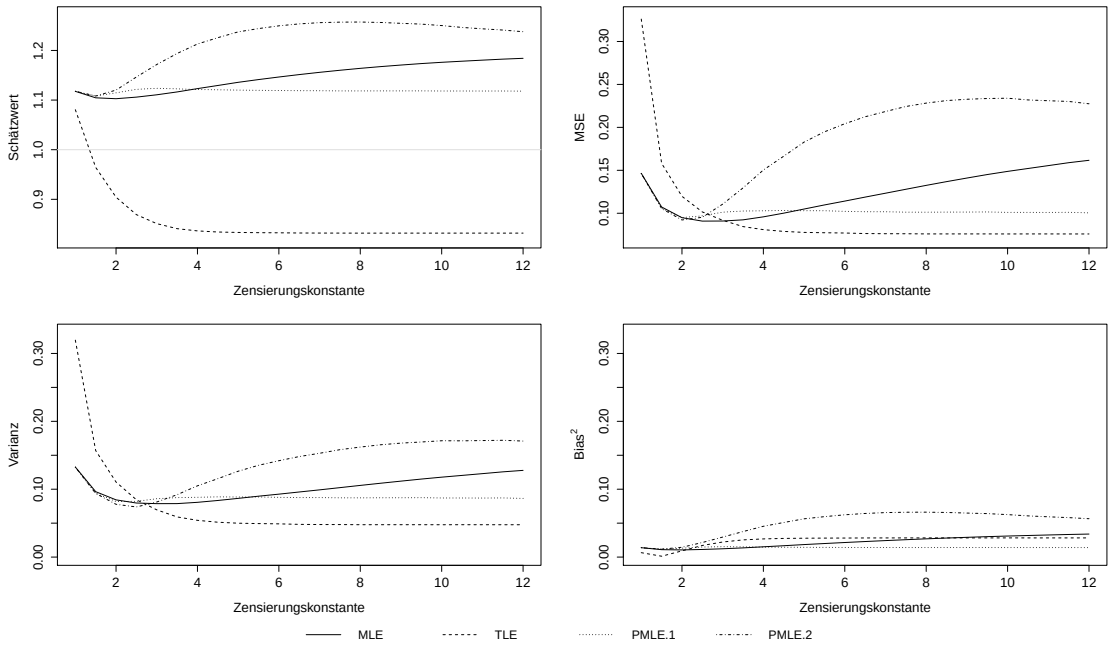

Abbildung 5.1.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_5(t)$ für einen Stichprobenumfang von $n = 20$ und einen Trimming-Anteil von $\beta = 0.1$.

Der zu schätzende Parameter wird bei allen Stichprobenumfängen überschätzt. Eine Ausnahme stellt der Getrimmte-Likelihood-Schätzer dar, der den gesuchten Parameter bei einem kleinen Stichprobenumfang unterschätzt.

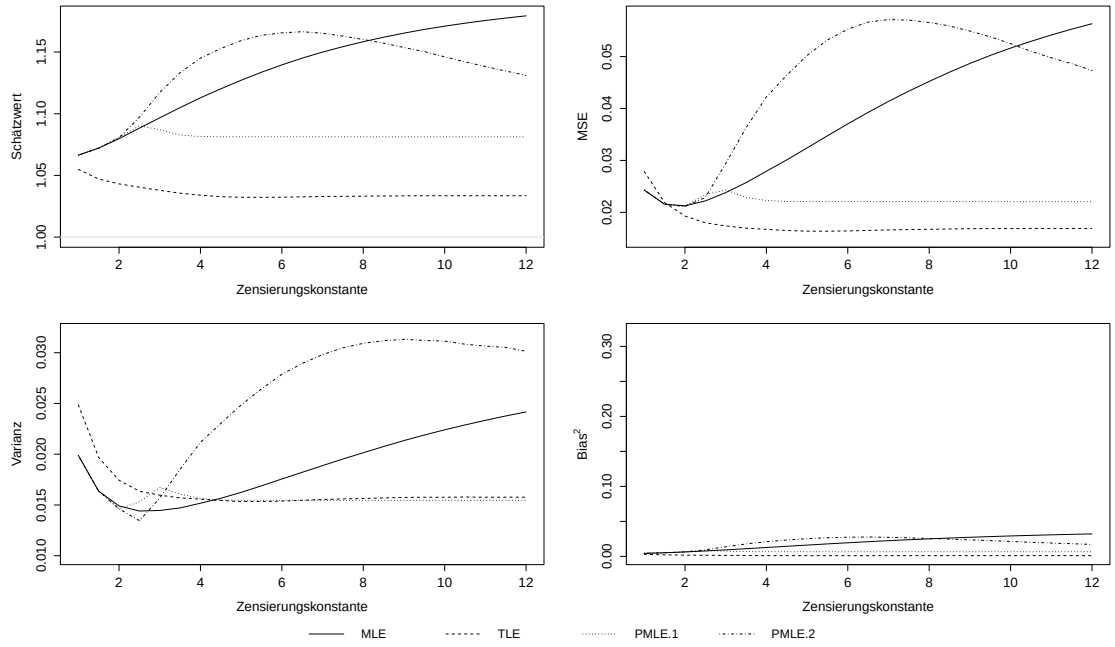

Abbildung 5.2.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_5(t)$ für einen Stichprobenumfang von $n = 100$ und einen Trimming-Anteil von $\beta = 0.1$.

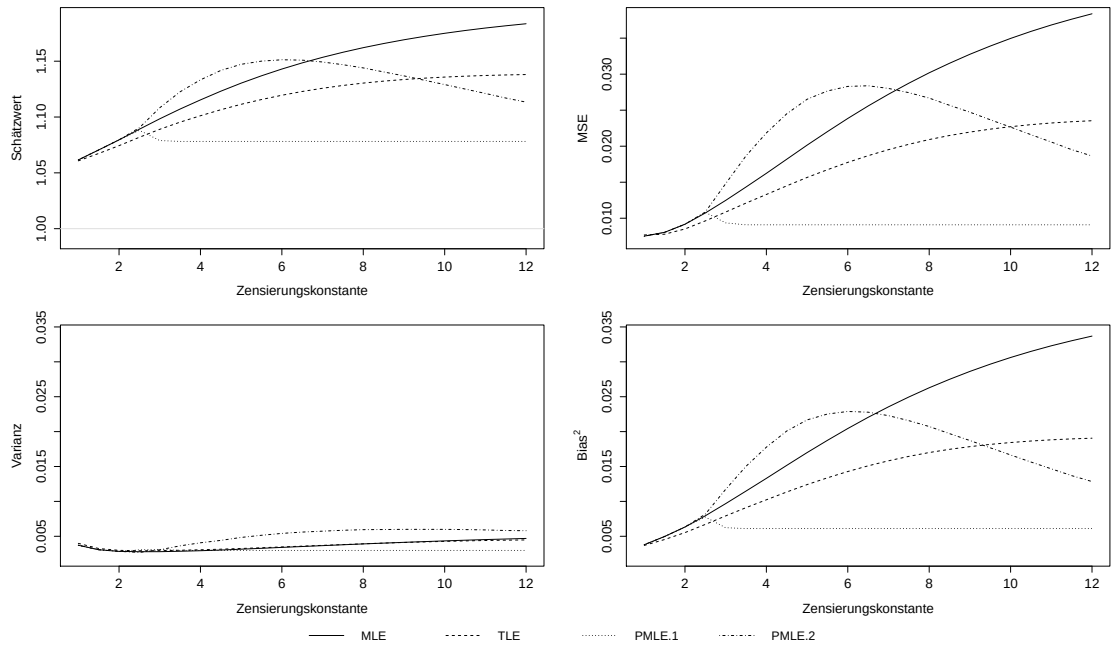

Abbildung 5.3.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_5(t)$ für einen Stichprobenumfang von $n = 500$ und einen Trimming-Anteil von $\beta = 0.1$.

Der Maximum-Likelihood-Schätzer ist insbesondere für sehr kleine Zensierungszeiten nicht weniger effizient als die getrimmten Schätzer. Er wird jedoch um so unzuverlässiger, je größer die Zensierungszeit und je größer der Stichprobenumfang sind.

Der Getrimmte-Likelihood-Schätzer ist in der hier gewählten Parameterkombination insbesondere bei kleinen und mittleren Stichprobengrößen für die meisten Zensierungszeiten der Schätzer mit dem kleinsten MSE. Nur für sehr kleine Zensierungszeiten und damit große Zensierungsanteile schneidet er schlechter als alle anderen Schätzer ab.

Auch der Pseudo-Maximum-Likelihood-Schätzer aus Vorschlag 1 liefert für die gewählte Parameterkombination sehr gute Ergebnisse. Bei großem Stichprobenumfang löst er den Getrimmte-Likelihood-Schätzer als MSE-effizientester der hier betrachteten Schätzer ab.

Der Pseudo-Maximum-Likelihood-Schätzer aus Vorschlag 2 schneidet bei kleinen Stichproben schlechter ab als der Maximum-Likelihood-Schätzer. Erst bei sehr großen Zensierungskonstanten und großen Stichproben verbessert er sich.

Auffällig ist zudem, dass in dem hier betrachteten Szenario bei einem mittleren Stichprobenumfang die Schätzungen für den mittleren quadratischen Fehler vorrangig durch die Varianzen getragen werden, während bei kleinen und großen Stichprobenumfängen die quadrierten Verzerrungen dominieren.

Im Anhang sind die Simulationsergebnisse für weitere Parameterkombinationen zu finden. In einem Szenario wurde der Anteil der Kontaminierung erhöht, in einem weiteren der Trimming-Anteil verringert. Zudem wurden die Parameter λ_1 und λ_2 der kontaminierten Verteilung der Ereigniszeiten verändert. Hervorzuheben ist, dass für größere Parameter der kontaminierten Verteilung der Pseudo-Maximum-Likelihood-Schätzer aus Vorschlag 1 nun auch bei mittlerem Stichprobenumfang einen kleineren MSE hat als der Getrimmte-Likelihood-Schätzer. Nur bei kleinem Stichprobenumfang bleibt der Getrimmte-Likelihood-Schätzer für die meisten Zensierungszeiten der effizienteste der hier betrachteten Schätzer.

5.2. Simulation bei zufällig rechtszensierten Daten

Die Simulation bei zufällig rechtszensierten Ereigniszeitdaten umfasst einen Vergleich des Maximum-Likelihood-Schätzers mit dem Getrimmte-Likelihood-Schätzer. Dabei werden wie im vorherigen Abschnitt Zufallszahlen aus einer kontaminierten Verteilung der Form (5.1) gezogen. Diese werden jedoch nicht mit einer festen Zensierungszeit, sondern zufällig zensiert. Für die Zensierungszeiten wird eine Exponentialverteilung angenommen, da somit für jedes Untersuchungsobjekt ein konstantes Risiko besteht, zensiert zu werden.

Die Wahrscheinlichkeit, dass eine zufällig ausgewählte Beobachtung i $(i = 1, \dots n)$ vor Eintritt des Zielereignisses zensiert wird, ist gegeben durch

$$\begin{aligned} p &= P(C_i < T_i) \\ &= \int_0^\infty f_\lambda(t_i)\, P(C_i < T_i) \mathrm{d}t_i \\ &= \int_0^\infty f_\lambda(t_i)\, F_{\lambda_C}(t_i) \mathrm{d}t_i \\ &= \int_0^\infty f_\lambda(t_i)\, (1 - \mathrm{e}^{-t_i/\lambda_C}) \mathrm{d}t_i \,. \end{aligned} \tag{5.2}$$

Dabei bezeichnen T_i die Ereigniszeit von Untersuchungsobjekt i mit Dichte $f_\lambda(t_i)$ und C_i dessen Zensierungszeit. Die Zensierungszeit ist exponentialverteilt mit Verteilungsfunktion $F_{\lambda_C}(c_i)$ und λ_C steht für deren Parameter. Der Parameter der Exponentialverteilung der Zensierungszeiten kann für einen angestrebten Zensierungsanteil p durch numerisches Lösen der Gleichung in (5.2) ermittelt werden.

Für die Simulation in diesem Abschnitt werden die gleichen Parameterkombinationen wie in Abschnitt 5.1 gewählt. Dies sind $\lambda_1 = 1$ für den Parameter der interessierenden Ereigniszeitverteilung und $\lambda_2 = 5$ für den Parameter der Ereigniszeitverteilung der Ausreißer bei einem Kontaminierungsanteil von $\epsilon = 0.05$ sowie ein Anteil der zu trimmenden Beobachtungen von $\beta = 0.1$. Auch die zuvor betrachteten Stichprobenumfänge werden beibehalten. Die Darstellung der Simulationsergebnisse erfolgt nun aber in Abhängigkeit von dem Zensierungsanteil, der einen Bereich von 5 % bis 75 % abdeckt.

Die Ergebnisse der Simulationen sind für den Stichprobenumfang $n = 20$ in Abbildung 5.4, für den Stichprobenumfang $n = 100$ in Abbildung 5.5 und für den Stichprobenumfang $n = 500$ in Abbildung 5.6 dargestellt.

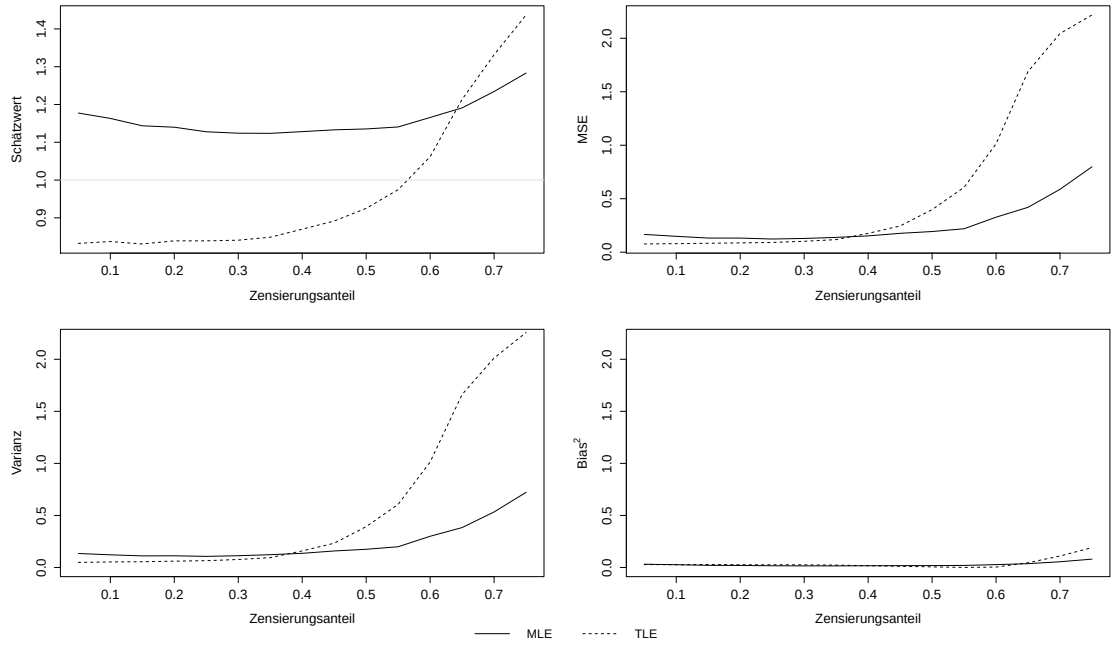

Abbildung 5.4.: Simulationsergebnisse bei zufällig zensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_5(t)$ für einen Stichprobenumfang von $n = 20$ und einen Trimming-Anteil von $\beta = 0.1$.

Bei kleinem Stichprobenumfang und für die gewählte Parameterkombination haben die Getrimmte-Likelihood-Schätzung und die Maximum-Likelihood-Schätzung bis zu einem Zensierungsanteil von etwa 0.4 einen ähnlichen geschätzten MSE, wobei die Getrimmte-Likelihood-Schätzung den gesuchten Parameter unterschätzt und die Maximum-Likelihood-Schätzung diesen überschätzt. Ab einem Zensierungsanteil von 0.4 verschlechtert sich jedoch der geschätzte mittlere quadratische Fehler der Getrimmte-Likelihood-Schätzung und nimmt zunehmend schlechtere Werte als für die ungetrimmte Schätzung an.

Bei mittlerem und großem Stichprobenumfang ist der getrimmte Schätzwert dem ungetrimmten deutlich überlegen. In Analogie zu der Situation bei kleinem Stichprobenumfang fällt auch bei einem mittleren Stichprobenumfang auf, dass sich der geschätzte mittlere quadratische Fehler des Getrimmte-Likelihood-Schätzwerts ab einem mittleren Zensierungsanteil verschlechtert. Bei großem Stichprobenumfang ist dies nicht zu beobachten.

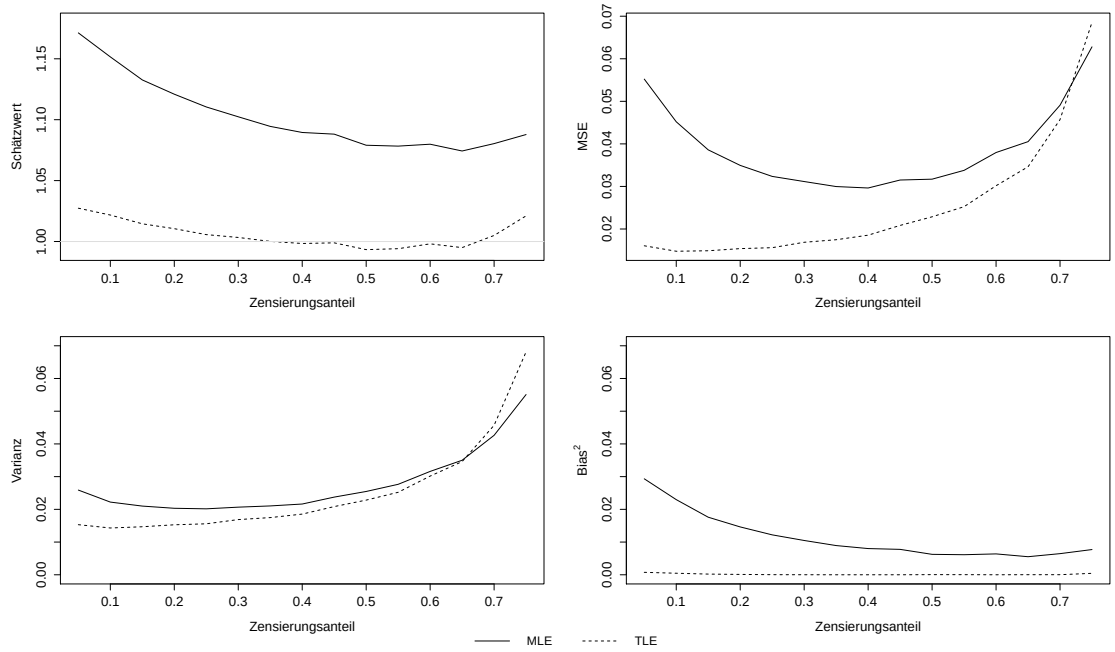

Abbildung 5.5.: Simulationsergebnisse bei zufällig zensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_5(t)$ für einen Stichprobenumfang von $n = 100$ und einen Trimming-Anteil von $\beta = 0.1$.

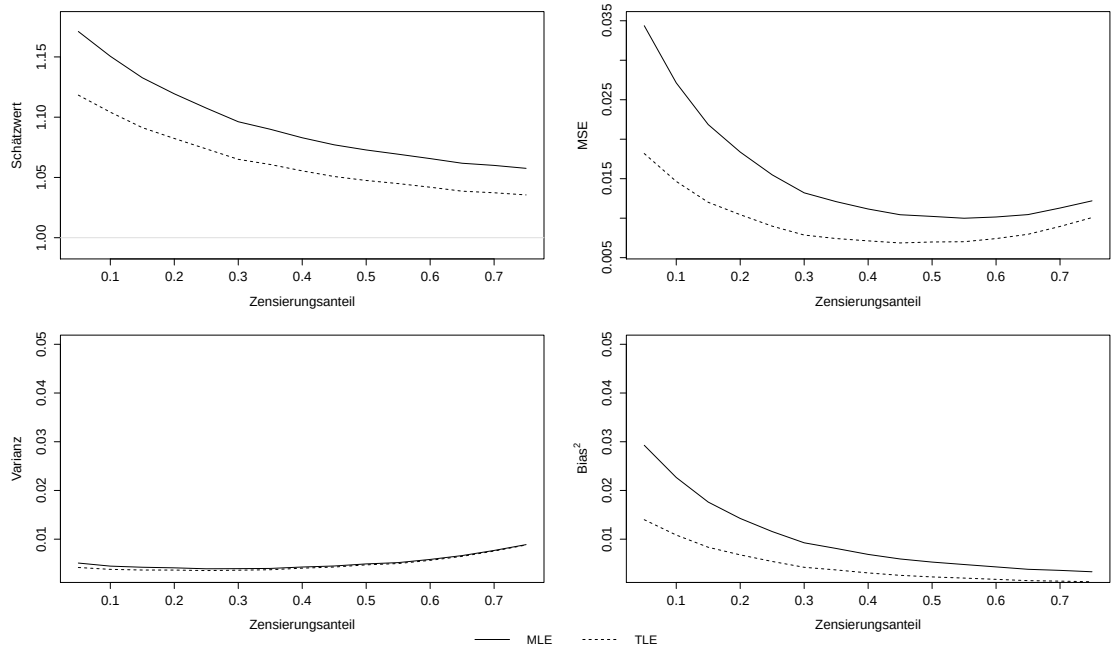

Abbildung 5.6.: Simulationsergebnisse bei zufällig zensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_5(t)$ für einen Stichprobenumfang von $n = 500$ und einen Trimming-Anteil von $\beta = 0.1$.

Für die Simulation bei zufällig rechtszensierten Ereigniszeiten finden sich im Anhang ebenfalls Ergebnisse für weitere Parameterkombinationen. Den Ergebnissen ist gemein, dass bei mittleren und großen Stichprobenumfängen der Getrimmte-Likelihood-Schätzer stets deutlich besser abschneidet als der Maximum-Likelihood-Schätzer. Nur bei kleinen Stichproben verschlechtert sich die Effizienz des Getrimmte-Likelihood-Schätzers ab einem Zensierungsanteil von etwa 0.4.

5.3. Anwendung der Schätzer auf einen Datensatz

Im Rahmen einer klinischen Studie des Royal Free Hospital in London wurden in den 1960er und 1970er Jahren Überlebenszeiten von 44 Patienten mit chronischer Hepatitis B erhoben. Die Patienten waren gleichmäßig auf zwei Gruppen aufgeteilt, von denen eine mit dem Medikament Prednisolone behandelt wurde. Die zweite Gruppe erfuhr keine Behandlung und diente als Kontrollgruppe. Die Überlebenszeiten in Monaten sind in Tabelle 5.1 aufgeführt (vgl. Hand et al. (1994), S. 343).

Um einen ersten Überblick über die Ereigniszeiten der Patienten zu erhalten und die Eignung der Daten für die in Kapitel 4 vorgestellten Schätzverfahren beurteilen zu können, wurden die Survivalfunktionen und die Hazardraten mittels nicht-parametrischer Verfahren geschätzt. Die Schätzung der Survivalfunktionen erfolgte nach Kaplan-Meier, die der Hazardraten nach einer an den Kaplan-Meier-Schätzer angelehnten Schätzmethode. Zu den nicht-parametrischen Schätzverfahren sei zum Beispiel auf Collett (2003), S. 19 ff. und S. 30 ff. verwiesen.

Die geschätzten Survivalfunktionen sind in Abbildung 5.7 ersichtlich, wobei die senkrechten Striche die zensierten Beobachtungen andeuten. Die geschätzten Hazardraten sind in Abbildung 5.8 dargestellt. Es ist nicht anzunehmen, dass alle Patienten zur gleichen Zeit in die Studie eingetreten sind; dadurch bedingt sind auch die Zensierungszeiten individuell. Da in der Kontrollgruppe jedoch nur die größten Beobachtungen zensiert wurden, soll für die Kontrollgruppe hier das Modell der Rechtszensierung mit fester Zensierungszeit unterstellt werden. Dies erfordert nur eine leichte Modifikation der Daten. Für die Behandlungsgruppe ist das Modell der zufälligen Rechtszensierung zutreffend.

Die geschätzte Hazardrate der Kontrollgruppe ist nahezu konstant. Die Annahme exponentialverteilter Ereigniszeiten ist also gerechtfertigt. In der Behandlungsgrup-

Tabelle 5.1.: Überlebenszeiten (in Monaten) von Patienten einer klinischen Studie zur Behandlung von chronischer Hepatitis B (* = zensierte Überlebenszeit).

Kontrollgruppe	Behandlungsgruppe
2	2
3	6
4	12
7	54
10	56*
22	68
28	89
29	96
32	96
37	125*
40	128*
41	131*
54	140*
61	141*
63	143
71	145*
127*	146
140*	148*
146*	162*
158*	168
167*	173*
182*	181*

pe ist die geschätzte Hazardrate während der ersten 125 Monate ebenfalls konstant. Dann steigt sie sprunghaft an, um weiterhin beinahe konstant zu verlaufen. Möglicherweise war die Patientengruppe nicht homogen, was einer kontaminierten Verteilung entspricht, wie sie in den Simulationen der vorherigen Abschnitte erzeugt wurde. Insgesamt kann daher auch für die Behandlungsgruppe eine Exponentialverteilung der Ereigniszeiten unterstellt werden.

Um die Schätzer, die für das Modell der Rechtszensierung mit fester Zensierungszeit hergeleitet wurden, nun auf die Daten der Kontrollgruppe anwenden zu können, wird

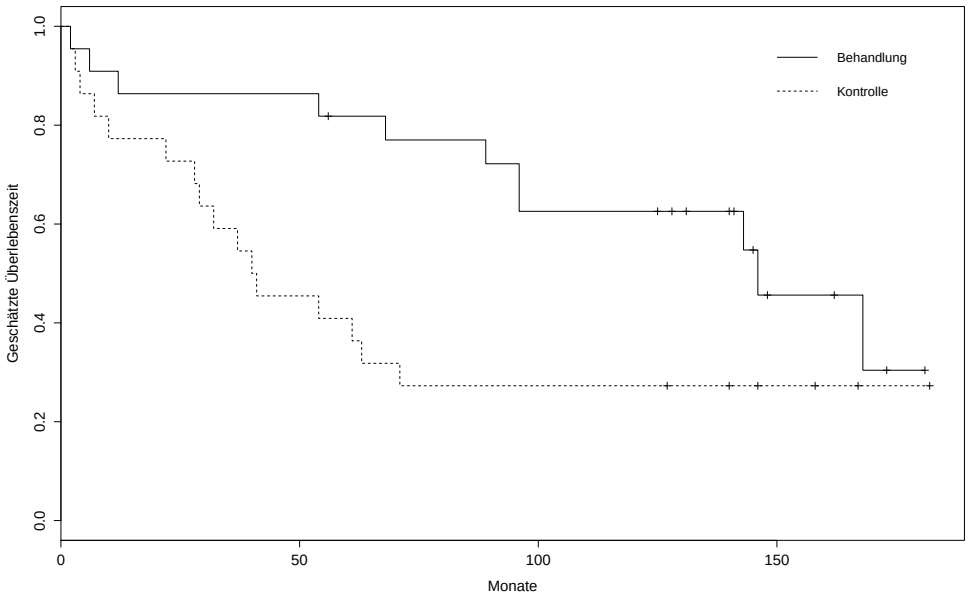

Abbildung 5.7.: Geschätzte Survivalfunktionen einer klinischen Studie für Patientengruppen mit chronischer Hepatitis B.

eine künstliche Zensierungszeit von 126 Monaten eingeführt, da die kleinste zensierte Beobachtung 127 Monate beträgt. Die Realisierungen der 6 zensierten Beobachtungen reduzieren sich dadurch auf 126. Für den so veränderten Datensatz ergibt sich für den Parameter der Exponentialverteilung eine Maximum-Likelihood-Schätzung von 78.75 Monaten. Das Trimmen von 10 % der Daten führt zu einer Getrimmte-Likelihood-Schätzung von 80.43 Monaten, wobei zur Berechnung des Schätzwertes die beiden größten unzensierten Beobachtungen getrimmt werden. Der Zensierungsanteil in der Kontrollgruppe beträgt etwa 27 %. Damit sind mehr Beobachtungen zensiert als zu trimmen sind. Die Pseudo-Maximum-Likelihood-Schätzungen sind also identisch mit dem Maximum-Likelihood-Schätzwert.

Alle vier Schätzverfahren liefern also einen ähnlichen Schätzwert für den Parameter der Exponentialverteilung, was gemäß der Simulationsergebnisse für eine so kleine Stichprobe und einen Zensierungsanteil von etwa 27 % zu erwarten war. Der Erwartungswert der Überlebenszeit in der Kontrollgruppe (nach Modifikation) beträgt demnach etwa 80 Monate und die Hazardrate liegt bei etwa 0.013.

Der Maximum-Likelihood-Schätzer und der Getrimmte-Likelihood-Schätzer werden nun auf die Daten der Behandlungsgruppe angewendet. Bei einem Zensierungsanteil von genau 50 % ergibt die Maximum-Likelihood-Schätzung für den Parameter

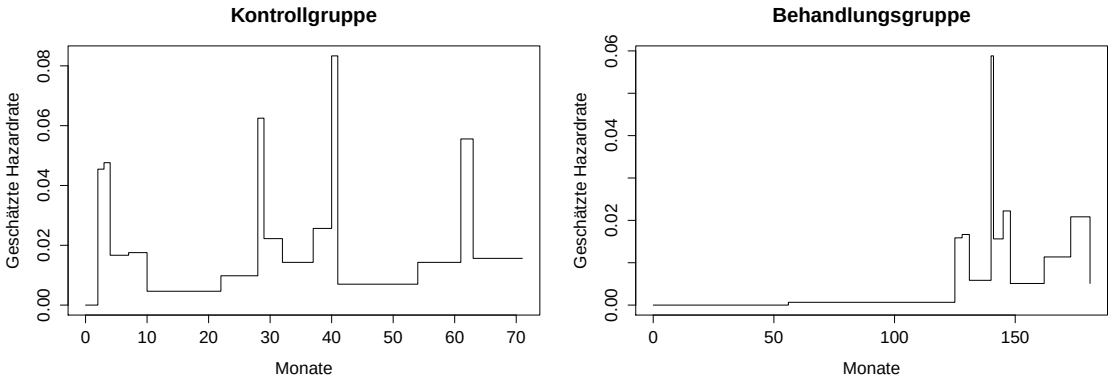

Abbildung 5.8.: Geschätzte Hazardraten einer klinischen Studie für Patientengruppen mit chronischer Hepatitis B.

der Exponentialverteilung 219.09 Monate. Das Trimmen von 10 % der Beobachtungen führt zu einer Getrimmte-Likelihood-Schätzung von 232.89 Monaten; auch hierbei wurden die beiden größten unzensierten Beobachtungen getrimmt. Für das angenommene Modell entspricht dies einer Hazardrate von etwa 0.0046 bzw. 0.0043.

6. Zusammenfassung

In der vorliegenden Arbeit wurden verschiedene auf der Likelihood-Funktion basierende getrimmte Schätzungen für den Parameter der Exponentialverteilung bei rechtszensierten Daten mit dem klassichen Maximum-Likelihood-Schätzer verglichen. Der Vergleich erfolgte mittels Simulationen, wobei die Ereigniszeitdaten aus einer kontaminierten Verteilung gezogen wurden. Die Simulationsergebnisse wurden anhand des mittleren Schätzwertes für den gesuchten Parameter sowie anhand des geschätzten mittleren quadratischen Fehlers und dessen Komponenten, empirischer Varianz und geschätzter Verzerrung, verglichen. Desweiteren wurden die Schätzer auf einen Datensatz aus einer klinischen Studie angewendet.

Bei rechtszensierten Daten, die mit einer festen Zensierungszeit zensiert waren, konnten alle in Kapitel 4 vorgestellten Schätzer berechnet werden: der klassische Maximum-Likelihood-Schätzer, der Getrimmte Likelihood-Schätzer sowie zwei von Clarke et al. (2014) vorgestellte Hybrid-Schätzer, in deren Herleitung auch das nach oben getrimmte Mittel einbezogen wurde. Es zeigte sich, dass keiner der Schätzer bei allen gewählten Parameterkombinationen und Stichprobenumfängen als der Schätzer mit dem kleinsten MSE aller betrachteten Schätzer ausgemacht werden konnte. Gleichwohl lieferten der Getrimmte-Likelihood-Schätzer und der Pseudo-Maximum-Likelihood-Schätzer aus Vorschlag 1 in den Simulationen insgesamt im Mittel eher kleine Schätzungen für den mittleren quadratischen Fehler. Der Maximum-Likelhood-Schätzer war den anderen Schätzern insbesondere bei sehr kleinen Zensierungszeiten nicht unterlegen, während der Getrimmte-Likelihood-Schätzer bei kleinen Zensierungszeiten und damit hohen Zensierungsanteilen den Schätzwert im Mittel eher unzuverlässig schätzt. Nicht überzeugen konnte hingegen der Pseudo-Likelihood-Schätzer aus Vorschlag 2, der häufig größerer geschätzte mittlere quadratische Fehler aufwies als der Maximum-Likelihood-Schätzer.

Mittels einer Simulation von zufällig rechtszensierten Daten wurden sodann der Maximum-Likelihood-Schätzer und der Getrimmte-Likelihood-Schätzer miteinander

vergleichen, wobei die Simulationsergebnisse in Abhängigkeit des Zensierungsanteils dargestellt wurden. Es zeigte sich, dass die Getrimmte-Likelihood-Schätzung in nahezu allen gewählten Parameterkombinationen dem Maximum-Likelihood-Schätzer überlegen war. Waren jedoch viele Daten zensiert, so ließ insbesondere bei kleinen Stichprobenumfängen die Überlegenheit des Getrimmte-Likelihood-Schätzers nach und der geschätzte mittlere quadratische Fehler wurde größer. Als kritische Grenze für den Zensierungsanteil konnte etwa 0.4 beobachtet werden.

In dieser Arbeit wurden nur Schätzer für exponentialverteilte Ereigniszeiten betrachtet. In vielen Studiensituationen, z.B. bei der Untersuchung der Lebensdauern von technischen Produkten, kann es sinnvoll sein, eine Exponentialverteilung für Ereigniszeiten anzunehmen. Für die Analyse menschlicher Ereigniszeiten, beispielsweise in klinischen Studien, scheint dies jedoch oft nicht gerechtfertigt. Daher wäre es erstrebenswert zu untersuchen, ob die hier betrachteten getrimmten Schätzer auch auf andere Ereigniszeitverteilungen übertragbar sind.

Die in dieser Arbeit gewählten Parameterkombinationen waren an die Arbeit von Clarke et al. (2014) angelehnt, in der insbesondere das Verhalten der Schätzer bei sehr großen Zensierungszeiten untersucht wurde. Große Zensierungszeiten gehen jedoch mit kleinen Zensierungsanteilen einher. Eine differenziertere Betrachtung auch von größeren Zensierungsanteilen wäre wünschenswert. Wie der Datensatz aus einer klinischen Studie, der für die Anwendung der Schätzer ausgewählt wurde, zeigt, sind sehr kleine Zensierungsanteile in Studien zur Analyse von Ereigniszeiten eher selten.

A. Anhang

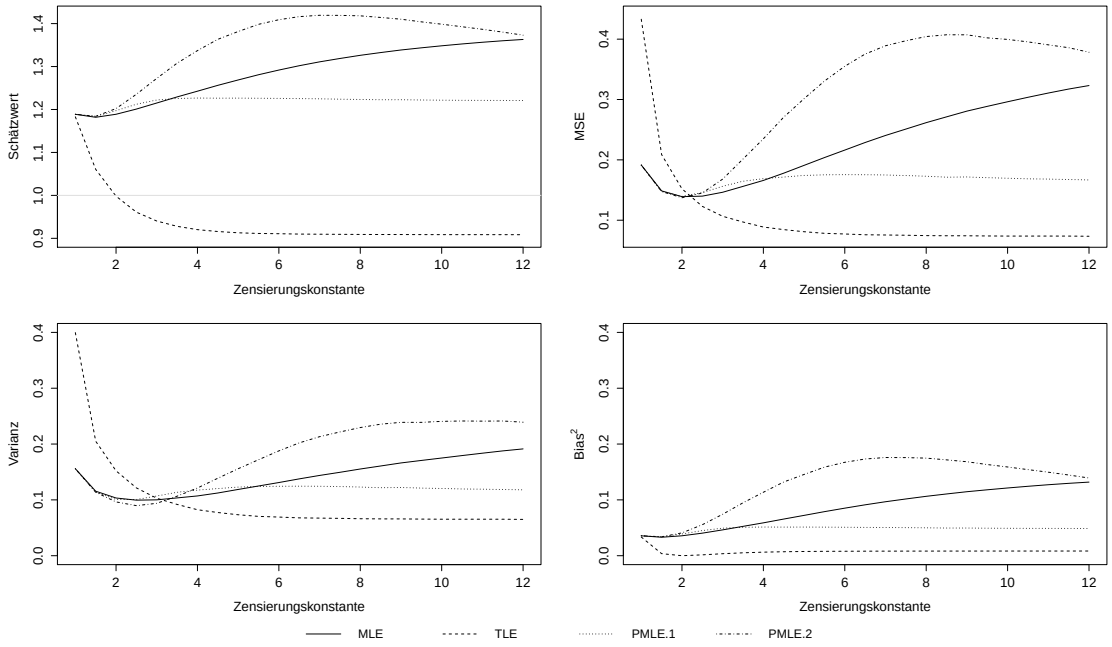

Abbildung A.1.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.9F_1(t) + 0.1F_5(t)$ für einen Stichprobenumfang von $n = 20$ und einen Trimming-Anteil von 0.1.

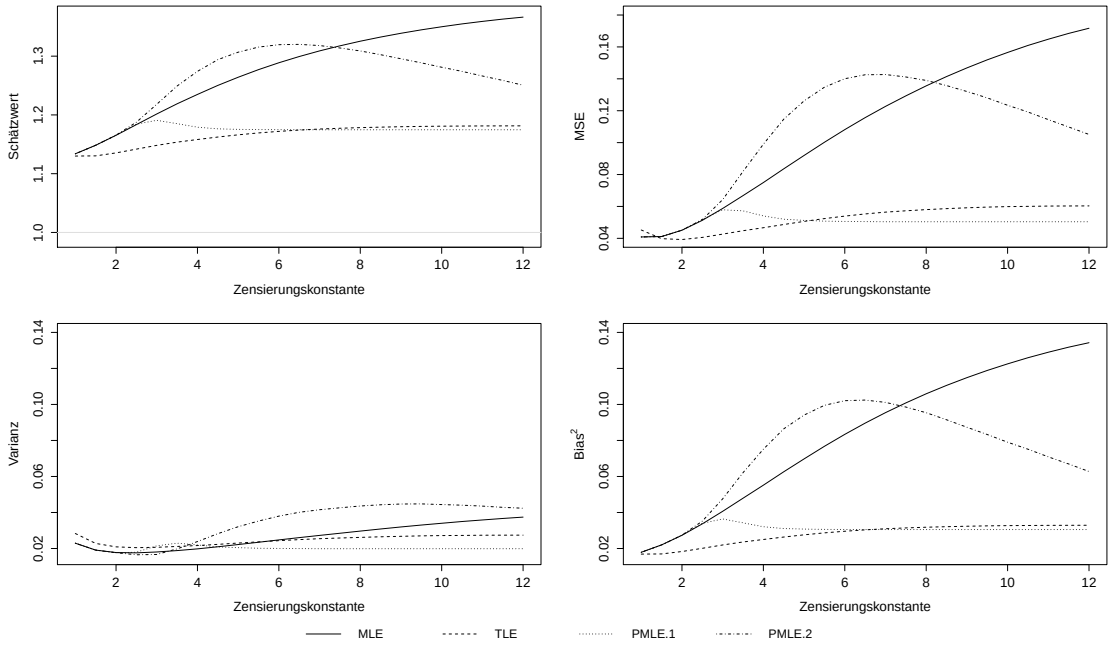

Abbildung A.2.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.9F_1(t) + 0.1F_5(t)$ für einen Stichprobenumfang von $n = 100$ und einen Trimming-Anteil von 0.1.

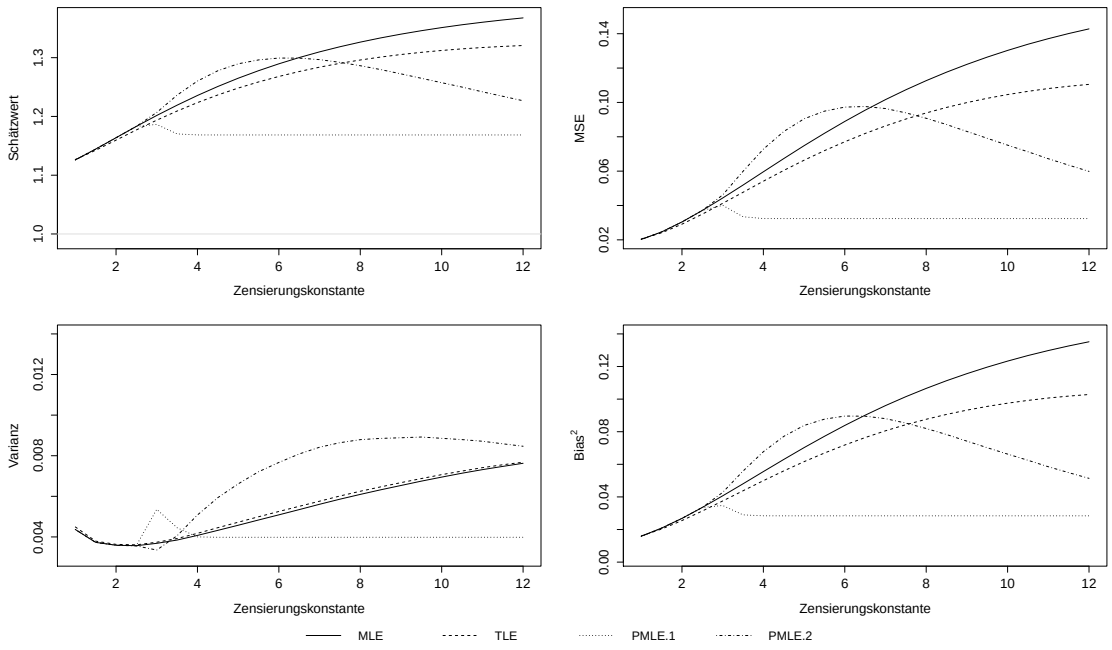

Abbildung A.3.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.9F_1(t) + 0.1F_5(t)$ für einen Stichprobenumfang von $n = 500$ und einen Trimming-Anteil von $\beta = 0.1$.

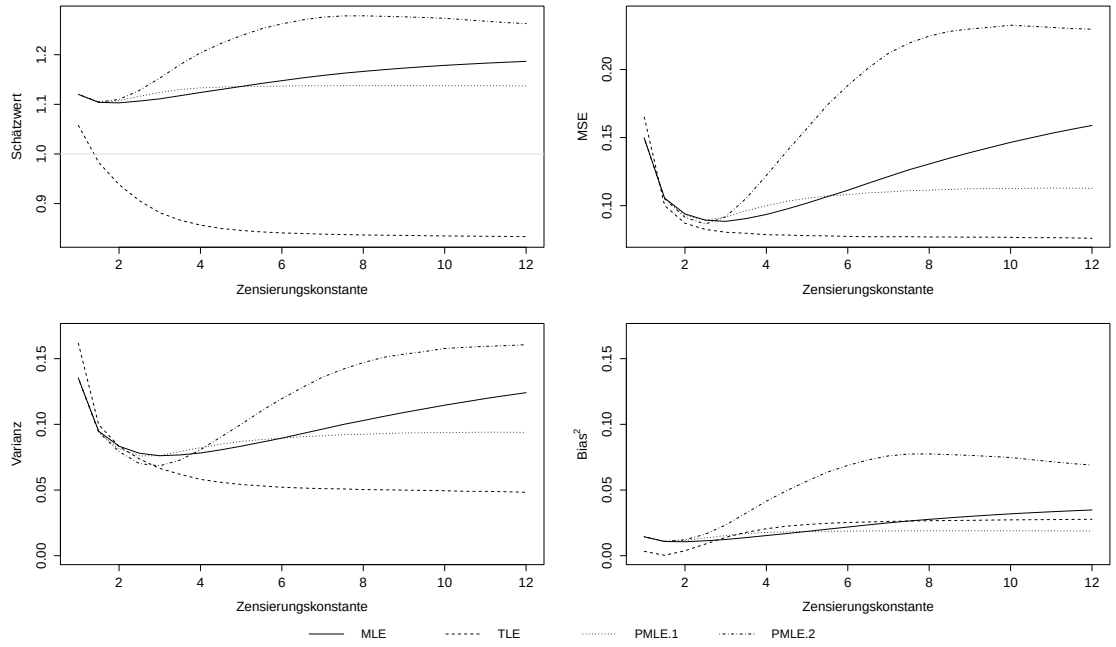

Abbildung A.4.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t)+0.05F_5(t)$ für einen Stichprobenumfang von $n = 20$ und einen Trimming-Anteil von $\beta = 0.05$.

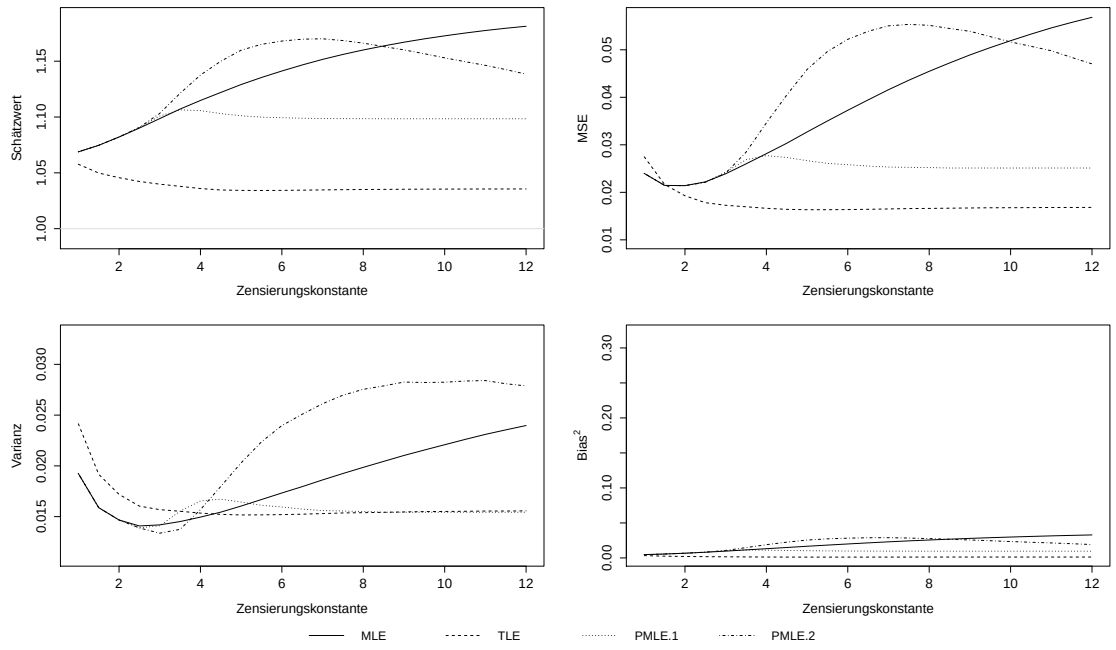

Abbildung A.5.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t)+0.05F_5(t)$ für einen Stichprobenumfang von $n = 100$ und einen Trimming-Anteil von 0.05.

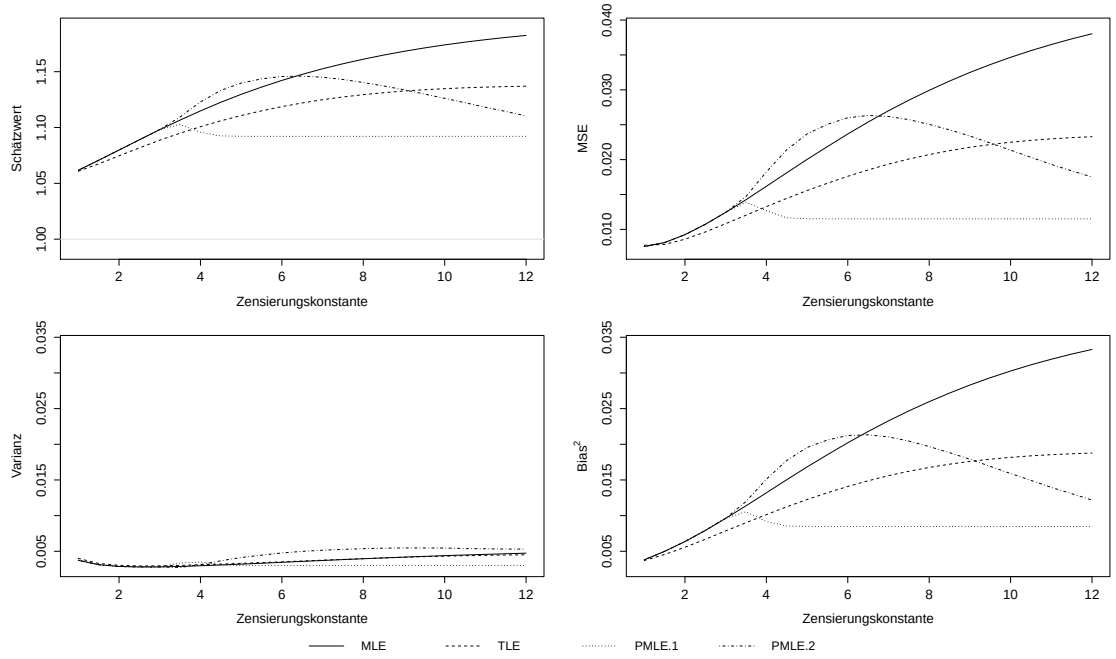

Abbildung A.6.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t)+0.05F_5(t)$ für einen Stichprobenumfang von $n = 500$ und einen Trimming-Anteil von 0.05.

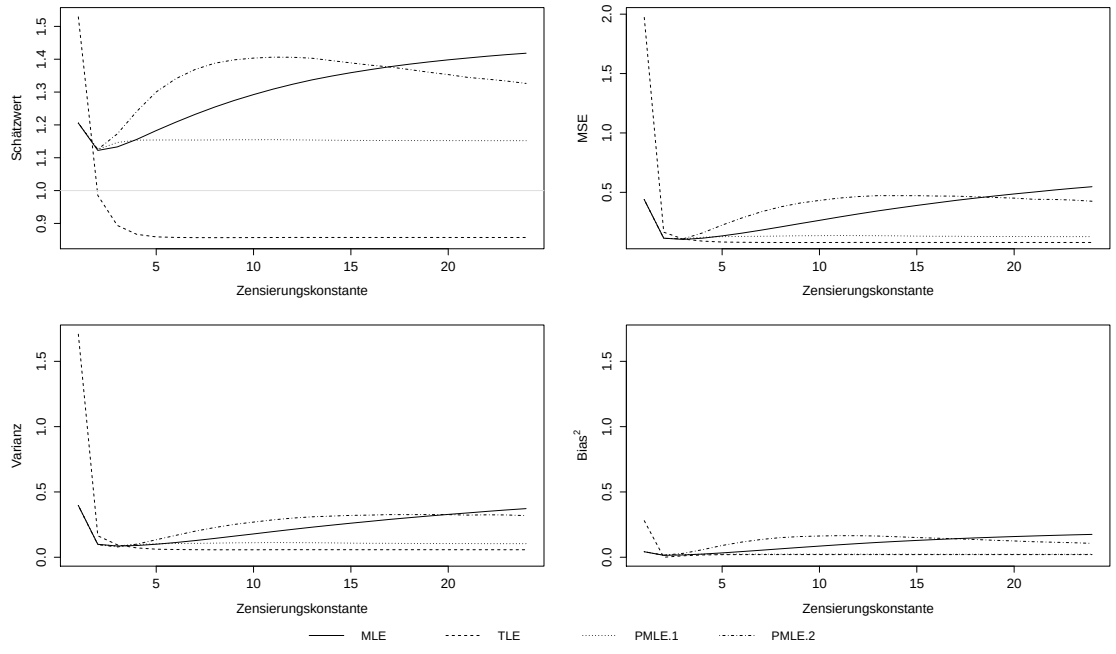

Abbildung A.7.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_{10}(t)$ für einen Stichprobenumfang von $n = 20$ und einen Trimming-Anteil von 0.1.

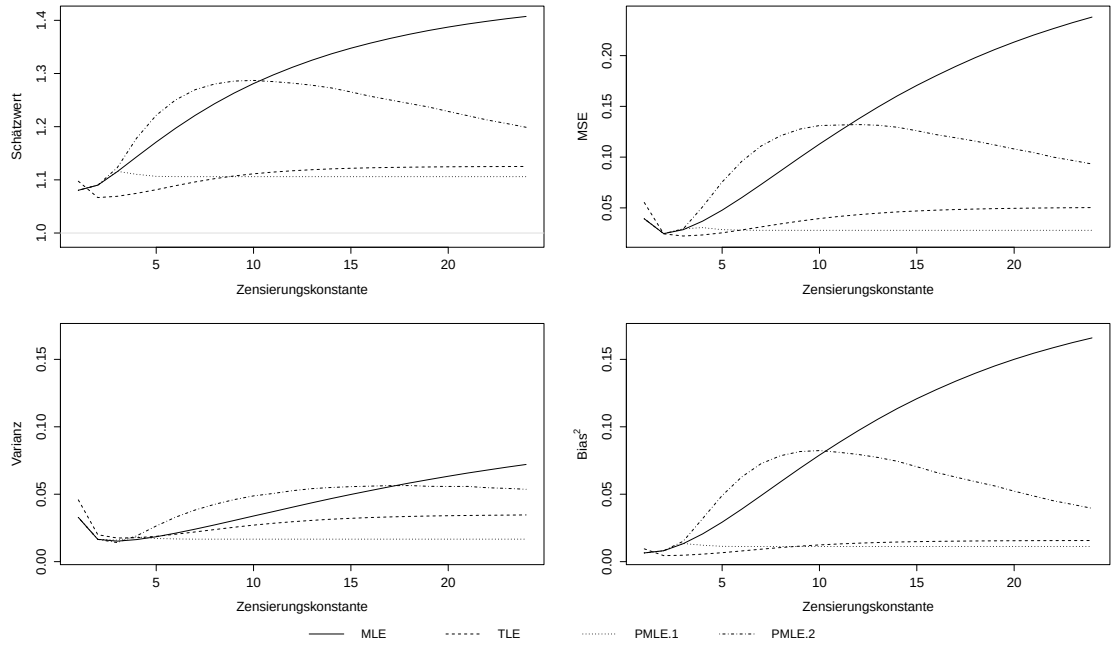

Abbildung A.8.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_{10}(t)$ für einen Stichprobenumfang von $n = 100$ und einen Trimming-Anteil von 0.1.

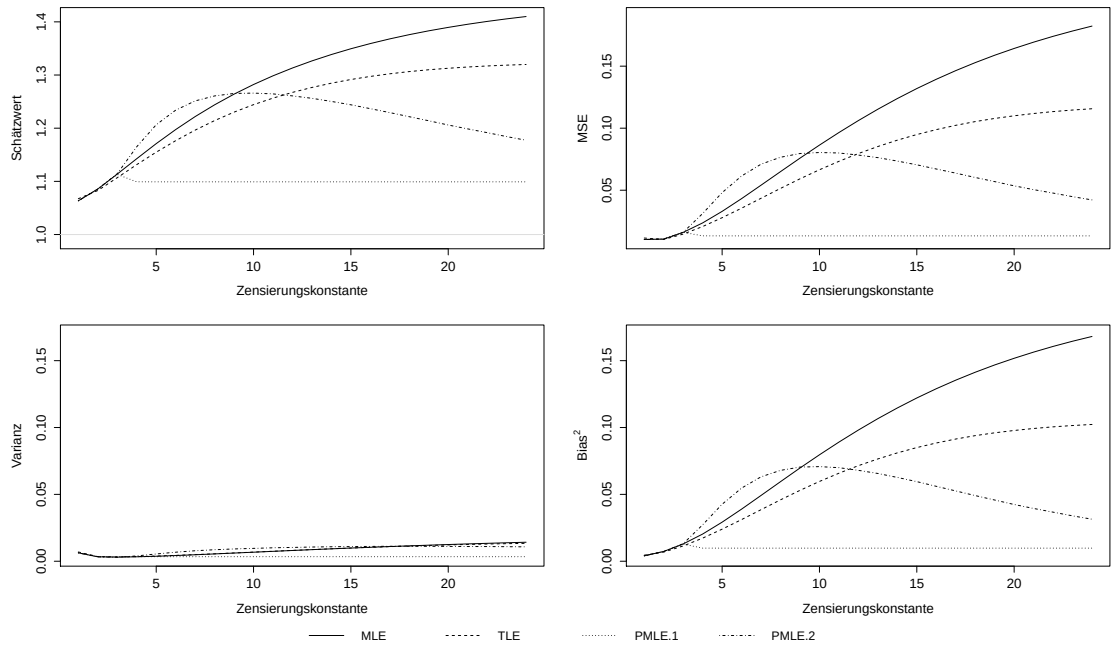

Abbildung A.9.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_{10}(t)$ für einen Stichprobenumfang von $n = 500$ und einen Trimming-Anteil von 0.1.

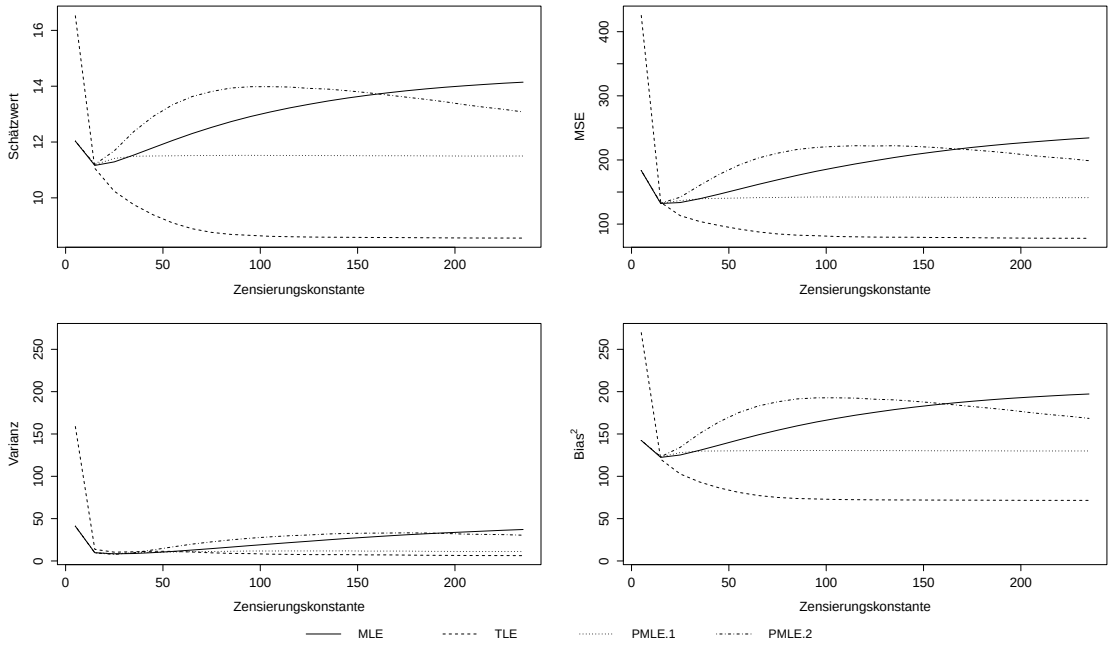

Abbildung A.10.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_{10}(t) + 0.05F_{100}(t)$ für einen Stichprobenumfang von $n = 20$ und einen Trimming-Anteil von 0.1.

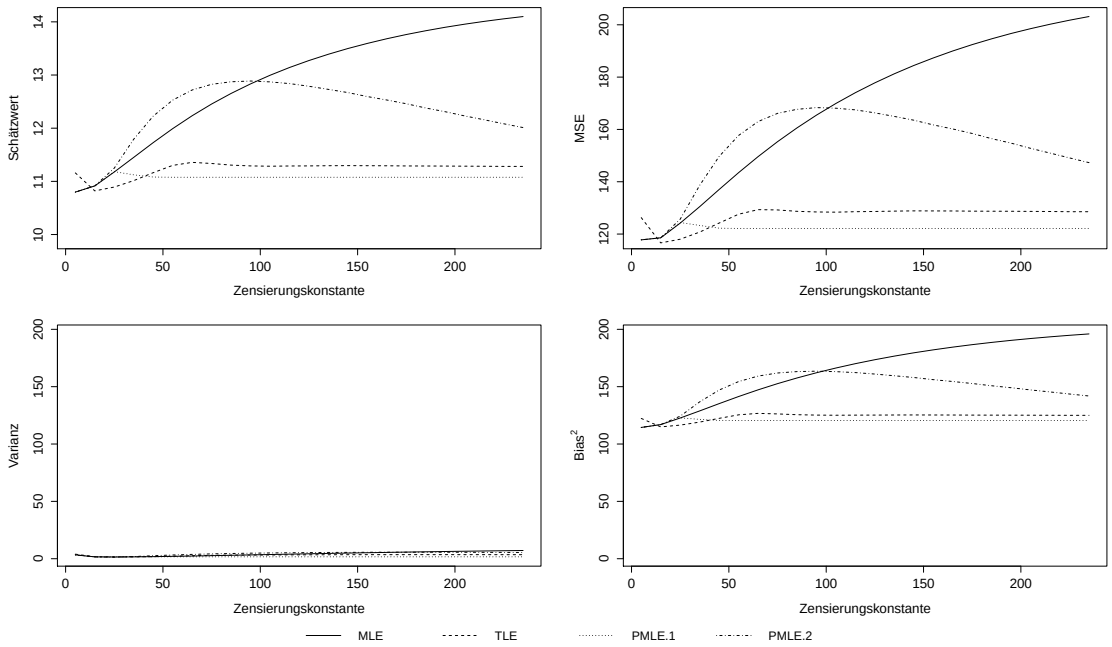

Abbildung A.11.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_{10}(t) + 0.05F_{100}(t)$ für einen Stichprobenumfang von $n = 100$ und einen Trimming-Anteil von 0.1.

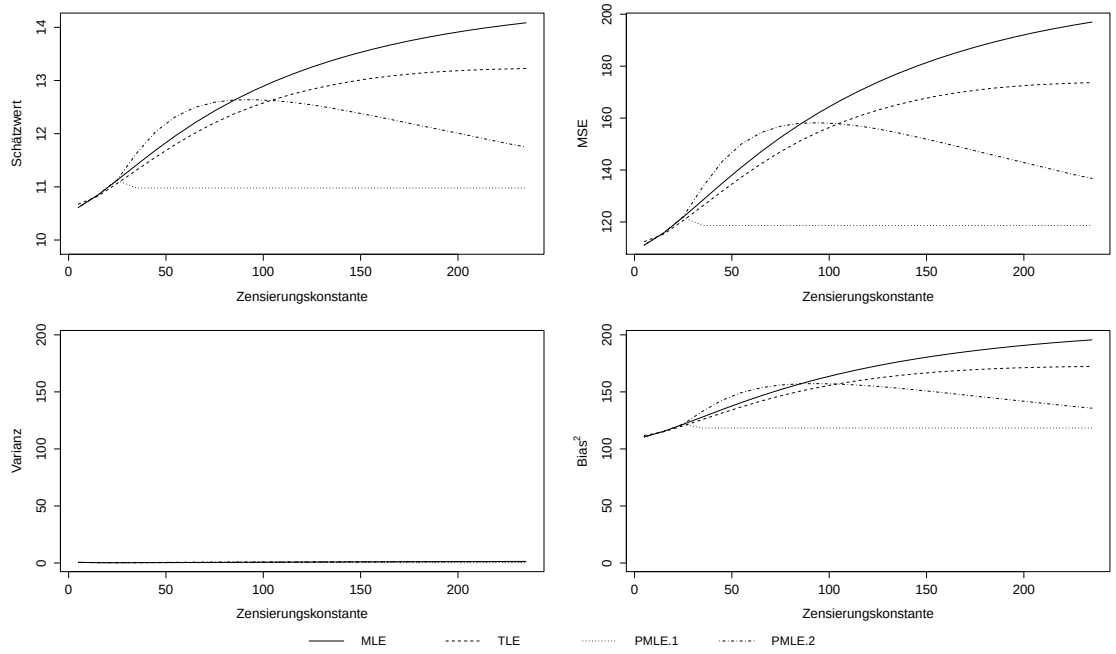

Abbildung A.12.: Simulationsergebnisse bei rechtszensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_{10}(t) + 0.05F_{100}(t)$ für einen Stichprobenumfang von $n = 500$ und einen Trimming-Anteil von 0.1.

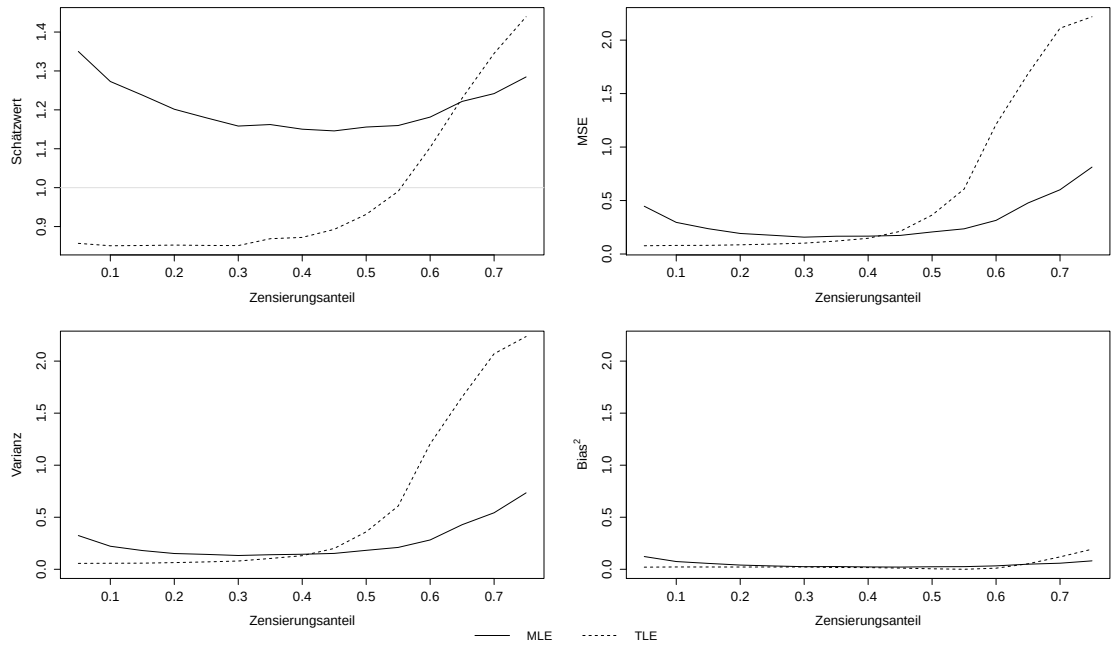

Abbildung A.13.: Simulationsergebnisse bei zufällig zensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_{10}(t)$ für einen Stichprobenumfang von $n = 20$ und einen Trimming-Anteil von 0.1.

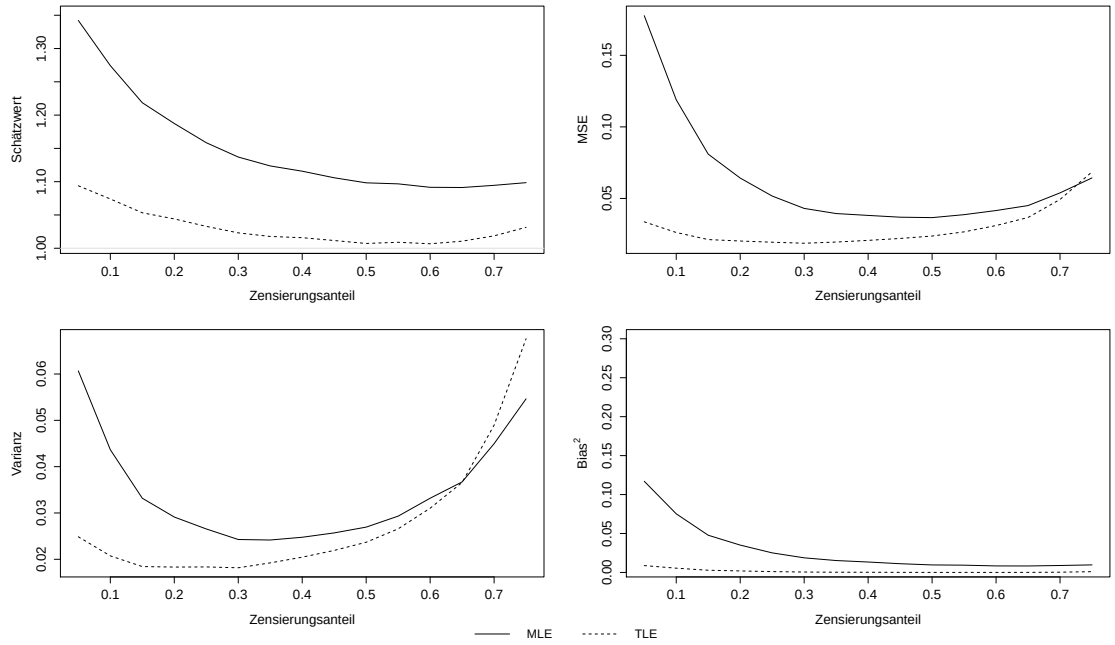

Abbildung A.14.: Simulationsergebnisse bei zufällig zensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_{10}(t)$ für einen Stichprobenumfang von $n = 100$ und einen Trimming-Anteil von 0.1.

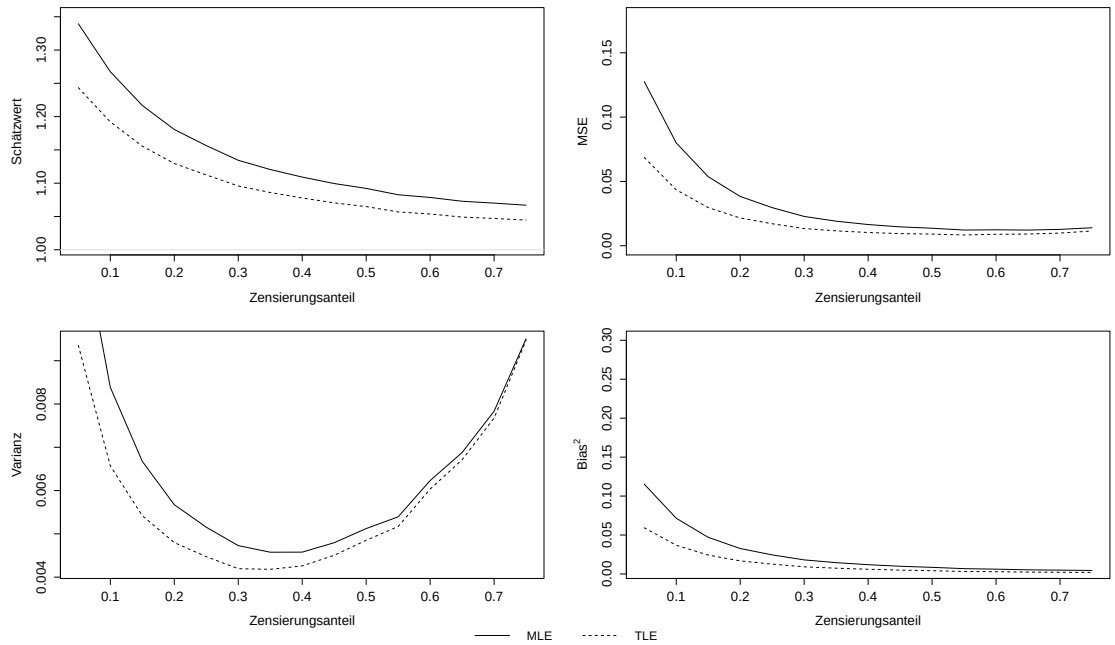

Abbildung A.15.: Simulationsergebnisse bei zufällig zensierten Ereigniszeitdaten aus der kontaminierten Verteilung $F(t) = 0.95F_1(t) + 0.05F_{10}(t)$ für einen Stichprobenumfang von $n = 500$ und einen Trimming-Anteil von 0.1.

Literaturverzeichnis

Clarke, B. R., Müller, C. H., Keppler, J. und Wamahiu, K. (2014). „Investigation of the performance of the trimmed likelihood of life time distributions with censoring“. Working Paper.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. 2. Auflage. Chapman & Hall/CRC, Boca Raton.

Fahrmeir, L., Künstler, R., Pigeot, I. und Tutz, G. (2006). *Statistik: Der Weg zur Datenanalyse*. 6. Auflage. Springer, Berlin.

Genschel, U. und Becker, C. (2005). *Schließende Statistik: Grundlegende Methoden*. Springer, Berlin.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. und Ostrowski, E. (1994). *A Handbook of Small Data Sets*. Chapman & Hall, London.

Klein, J. P. und Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. 2. Auflage. Springer, New York.

Müller, C. H. und Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference* 116.2, S. 503–519.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL: `http://www.R-project.org/`.

Schumacher, M. und Schulgen, G. (2008). *Methodik klinischer Studien*. 3. Auflage. Springer, Berlin.

Staudte, R. G. und Sheather, S. J. (1990). *Robust Estimation and Testing*. John Wiley & Sons, Inc., New York.

Eidesstattliche Versicherung

Höller, Alexandra	[redacted]
Name, Vorname	Matr.-Nr.

Ich versichere hiermit an Eides statt, dass ich die vorliegende Bachelorarbeit/~~Masterarbeit~~* mit dem Titel

Vergleich von verschiedenen getrimmten Schätzungen bei zensierten Daten

selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Dortmund, den	
Ort, Datum	Unterschrift

*Nichtzutreffendes bitte streichen

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -)

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird gfls. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin“) zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Dortmund, den	
Ort, Datum	Unterschrift