Maximum-Likelihood-Schätzung für zensierte und getrimmte Daten

von Janine Keppler
Studiengang Mathematik auf Diplom

Diplomarbeit WS 2012/2013
Universität Kassel
Fachbereich Mathematik

Betreuende Gutachterin:
Prof. Dr. Christine Müller

Zweitgutachterin:
Prof. Dr. Doris Fiebig

INHALTSVERZEICHNIS

EINLEITUNG

MOTIVATION

Die vorliegende Diplomarbeit wurde im Rahmen einer Forschungsgruppe der Universität Kassel erstellt, in der die Ermüdung von graduierten Material untersucht wurde. Unter Laborbedingungen wurde das für die Analyse benötigte Material durch ständige Belastung und Verbiegung erstellt, so dass eine Ermüdung auftrat. Die so entstandene Materialprobe wurde dann zu verschiedenen Zeitpunkten unter einem Mikroskop fotografiert, so dass eine Bilderserie aus Einzelbildern zu jedem Zeitpunkt für die Materialprobe entstand. Zu jedem Zeitpunkt wurden 6×9 Bilder erstellt, die ein Gesamtbild des Materials ergaben. Aufgrund von Belichtungsschatten und stellenweise mangelnden Kontrast, erfolgte eine generelle Nachbearbeitung des Bildmaterials. Die so entstandenen Bilder wurden anschließend einer Segmentierung unterzogen. Im Speziellen bedeutet dies, dass die Bildpunkte in zwei Kategorien aufgeteilt wurden, einerseits in Risspunkte und andererseits in Hintergrundpunkte. Durch diese Segmentierung konnte dann ein Riss sichtbar gemacht und seine Länge bestimmt werden. Für jedes Einzelbild wurden somit Risse erfasst und dokumentiert. Des Weiteren konnte mit den ermittelten Rissen der längste Riss bestimmt werden. Mit Hilfe dieser Daten wurden dann schlussendlich mit unterschiedlichen Maximum-Likelihood-Methoden Schätzer die Lebenszeit der Materialproben ermittelt.

AUFGABENSTELLUNG

Das Hauptaugenmerk der vorliegenden Arbeit wurde auf die Segmentierung der Risspunkte und die später benötigten Maximum-Likelihood-Methoden gelegt. Beide Punkte werden ausführlich in der Theorie vorgestellt und besprochen, um sie dann schlussendlich

in der Praxis auf die vorliegenden Daten anwenden zu können und eine gute Schätzung der Lebenszeit zu bestimmen.

VORGEHEN

In der vorliegenden Arbeit werden zuerst einige wichtige Begriffe erläutert. Unter anderem sind dies der Mittelwert, sowie seine getrimmte Variante. Dazu kommt noch die Definition einer Schätzfunktion, die Konsistenz derjenigen und die stochastische Normalität. Darüber hinaus wird erläutert, was eine Zensierung ist und ihre Varianten beleuchtet. Diese Begriffe werden benötigt um in den darauf folgenden Kapiteln die weiterführende Theorie aufzubauen, die für den Praxisteil verwendet wird. Der Maximum-Likelihood-Schätzer mit seinen Varianten stellt hierbei die Grundlage der Lebenszeitanalyse dar. Mit ihm wurden speziell für die Rissentwicklung die unterschiedlichsten Schätzer ermittelt. Des Weiteren wird die Konsistenz der Schätzer überprüft und deren Robustheit festgestellt. Außerdem wird die Segmentierung erläutert, die den ersten Schritt für die Risserkennung darstellt. Mit Hilfe der Segmentierung und dem *Dijkstra's Shortest Path Algorithm* aus [MÜ09] werden dann die längsten Risse ermittelt und daraus die Lebenszeit geschätzt.

KAPITEL 1

EINLEITENDE DEFINITIONEN

In den folgenden Kapiteln wird der *getrimmte Mittelwert* vorgestellt. Wobei unterschieden werden muss, ob er zensiert oder unzensiert vorliegt. Um die Voraussetzungen zu schaffen, wird an dieser Stelle nochmal auf den Mittelwert und seine Eigenschaften eingegangen. Der Mittelwert hat die Gestalt:

$$\overline{y} = \frac{1}{N}\sum_{n=1}^{N} y_n.$$

Es wird nun untersucht, ob der Mittelwert erwartungstreu, konsistent, asymptotisch normalverteilt und effizient ist. Der Grund zur Prüfung dieser drei Eigenschaften ist, dass dies auch für den *getrimmten Mittelwert* durchgeführt werden muss, und somit gezeigt wird, dass dieser wohldefiniert ist.
Die Stichprobe $y = (y_1^{;}..., y_n)^T$ ist eine Realisierung der Zufallsvariable $Y = (Y_1, ..., Y_N)^T$ mit stochastisch unabhängigen und identisch verteilten (i.i.d) $Y_1, ..., Y_N$.

Definition 1.0.1 *Seien $Y_1, ..., Y_k$ voneinander disjunkte Zufallsvariablen und sei $P(Y = y_i)$ mit $i = 1, 2, ...$ die Wahrscheinlichkeitsfunktion. Dann heißt*

$$F_Y(y) = F_Y(y_1, ..., y_k) := P(Y_1 \leq y_1, ..., Y_k \leq y_k)$$

die Verteilungsfunktion. Sei $f(y)$ eine stetige Wahrscheinlichkeitsfunktion, dann ist die Verteilungsfunktion gegeben durch:

$$F_Y(y) = P(Y \leq y) := \int_{-\infty}^{y} f(x)dx.$$

Definition 1.0.2 *Eine Schätzfunktion $\hat{\theta}$ für θ heißt Erwartungstreu, wenn sie*

$$E(\hat{\theta}(Y)) = \theta \qquad \forall \quad \theta \in \Theta$$

erfüllt.

Für den Fall des Mittelwerts μ folgt somit:

$$E(\overline{y}) = E\left(\frac{1}{N}\sum_{n=1}^{N} y_n\right) = \frac{1}{N}\sum_{n=1}^{N} E(y_n) = \frac{1}{N}N\mu = \mu \quad \forall N$$

Definition 1.0.3 *Eine Folge von Schätzfunktionen $(\hat{\theta}_n)_{n\in\mathbb{N}}$ heißt schwach bzw. einfach konsistent für θ, falls*

$$\lim_{n\to\infty} P(\mid \hat{\theta}_n(Y) - \theta \mid > \epsilon) = 0 \quad \forall \epsilon > 0$$

gilt.

Somit ergibt sich mit dem Gesetz der großen Zahlen die Gleichung

$$\lim_{N\to\infty} \overline{y} = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} y_n = \mu.$$

Es ist leicht zu erkennen, dass der Mittelwert konsistent ist.

Definition 1.0.4 *Als asymptotisch normal wird eine Folge von Schätzfunktionen $(\hat{\theta}_n)_{n\in\mathbb{N}}$ genannt, für die der Ausdruck $\sqrt{n}(\hat{\theta}_n - \theta)$ stochastisch gegen eine Normalverteilung $\mathcal{N}(0, Var(\theta))$ konvergiert. (Vergleiche [SH90])*
Eine alternative Bedingung für asymptotische Normalität stellt

$$\frac{\hat{\theta}_n(Y) - E(\hat{\theta}_n(Y))}{\sqrt{\mathrm{Var}(\hat{\theta}_n(Y))}} \quad \overset{asymp}{\sim} \quad \mathcal{N}(0,1)$$

dar, vergleiche [RI08].

Definition 1.0.5 *Sie $(Y_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und Y Zufallsgröße auf $(\Omega, \mathcal{A}, P)$, dann gilt:*

$$(Y_n)_{n\in\mathbb{N}} \textit{ konvergiert fast sicher gegen } Y \Leftrightarrow P\left(\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}\right) = 1.$$

Für den Fall des arithmetischen Mittelwertes wird mit den zentralen Grenzwertsatz die asymptotische Normalität ersichtlich. Dabei kommt zum Tragen, dass Y eine unabhängig und identisch verteile Zufallsgröße ist.

1.1 ZENSIERUNG

Zur Zusammenstellung einer Charakteristik von Zensur diente als Grundlage [KL97]. Bei der Betrachtung der Lebenszeiten von einer Gruppe von Versuchsobjekten bzw. Individuen, stellt es sich als schwer bis unmöglich dar, diese Daten für jedes Objekt zu beobachten. In solchen Fällen, wenn nur in einem Zeitfenster die Daten bekannt sind, wird von einer Zensierung gesprochen. Dies ist meist durch Kosten oder Zeitgründen in Studien zu beobachten. Es wird hierbei zwischen Rechts-, Links- und Intervallzensierung unterschieden.
Bei einer Rechtzensierung existieren nur bis zu einem bestimmten Zeitpunkt genaue Daten. Danach wird nur festgehalten, ob das Versuchsobjekt zu diesen Zeitpunkt noch nicht ausgefallen ist. Bei einer Linkszensierung liegen keine Daten über die bisherige Lebensdauer vor. Liegen beide Typen von Zensur für die gesamten Daten vor, so wird von einer Intervallzensur gesprochen.
Da in der vorliegenden Arbeit nur Rechtszensur betrachtet wird, wird auf eine genauere Beschreibung der Links- und Intervallzensierung verzichtet. Bei der Rechtzensierung wird zwischen zwei Typen unterschieden. Bei Typ I hängt die Zensurschranke C_r von der Zeit ab. Hingegen ergibt sich bei Typ II der Zeitpunkt der Zensur C_r durch eine vorher bestimmte Anzahl der Versuchsobjekte die ausfallen müssen.

Definition 1.1.1 *Das Paar von Zufallsvariablen (Z, δ) sei der Datensatz eines Experimentes mit der rechten Zensurschranke C_r. Dann gibt δ an, ob eine Zensierung ($\delta = 1$) oder die wirkliche Lebenszeit vorliegt ($\delta = 0$). Z entspricht der Lebenszeit Y im unzensierten Fall und der Zensurschranke C_r im Zensierten. Also gilt $Z = \min(Y, C_r)$.*

1.1.1 TYP I RECHTSZENSUR

Bei einer *Typ I Rechtszensierung* wird mit einer festen Zahl von Versuchsobjekten gestartet und die Observierung dauert nur bis zu einem vorher festgelegten Zeitpunkt an, der durch die Zensurschranke C_r angeben wird.
Als Beispiel kann eine Studie des National Center of Toxicological Research (NCTR)

betrachtet werden. In Ihr wurde eine Anzahl von Mäusen mit einer bestimmten Dosis an Karzinogenen behandelt und ihre Lebensdauer beobachtet. Die Studie endete nach einer gewissen vorher festgelegten Zeit bzw. mit dem Tod aller Mäuse.

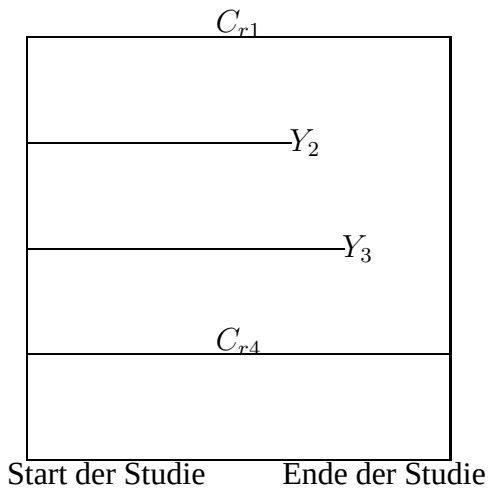


Abbildung 1.1: Beispiel einer Typ I Rechtszensierung

Treten aber mehr als eine Zensurschranke auf spricht man von einer *progressiven Typ I Rechtszensur*. In diesem Fall können die verschieden Objekte unterschiedliche Zensurschranken zugeordnet sein. Diesen Zensurtypus findet insbesondere bei nicht tödlich verlaufenden Krankheiten Anwendung. Hier können die verschiedenen Stadien der Krankheiten über die Zuteilung der verschieden Zensurschranken entscheiden. Um das vorhergehende Beispiel aufzugreifen, wird die Studie mit den Karzinogenen betrachtet. Die Mäuse werden per Zufall in vier Gruppen geteilt, die sich durch die Dosierung unterscheiden. Es werden zwei Zensurzeiten gewählt, um am effektivsten (bzw. ökonomischsten) Informationen zu erhalten. Es wird nur eine geringere Beobachtungszeit für die Gruppen mit den hohen Dosen gewählt und eine lange für die anderen Gruppen.

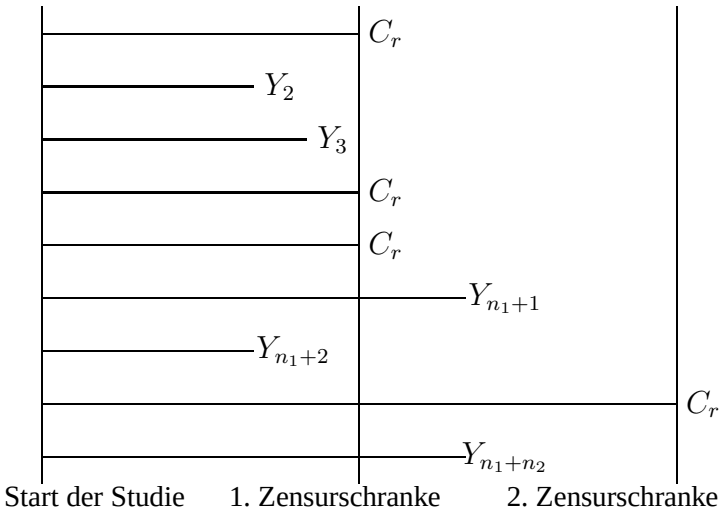


Abbildung 1.2: Beispiel einer progressiven Typ I Rechtszensierung mit zwei verschiedenen Zensurzeitpunkten

Als dritte und letzte Unterscheidung wird die *generalisierte Typ I Rechtszensur* in Augenschein genommen. In diesem Fall treten die Versuchsobjekte zu verschieden Zeitpunkten der Studie bei. Der Zeitpunkt der Zensur ist aber schon vorher festgelegt worden. Somit

besitzt jedes Versuchsobjekt eine eigne Zensurzeit. Um dies darzustellen bieten sich drei Sorten von Diagrammen an (vgl. Abb.1.3). Zuerst betrachten wir ein Diagramm, wie die Vorherigen. Hierbei fällt auf, dass jedes Versuchsobjekt einen eignen Startzeitpunkt besitzt. Verschiebt man nun die Geraden so, dass alle Startzeitpunkte auf 0 fallen, so ergibt sich die übliche Form für generalisierte Typ I Rechtszensuren. Eine weitere Form stellt das Lexis Diagramm dar. Hierfür wird auf der Abszisse die Kalenderzeit eingetragen und auf der Ordinate die Lebenszeit des Individuums. Anders als bei den vorherigen Diagrammen wird hier die Lebenszeit eines Objektes nicht als horizontale Linie, sondern als vertikal dargestellt. Somit besteht ein direkter, im Diagramm einsehbarer, Zusammenhang zwischen Eintritts-, Austrittszeitpunkt in die Studie und Lebensdauer des Versuchsobjektes.

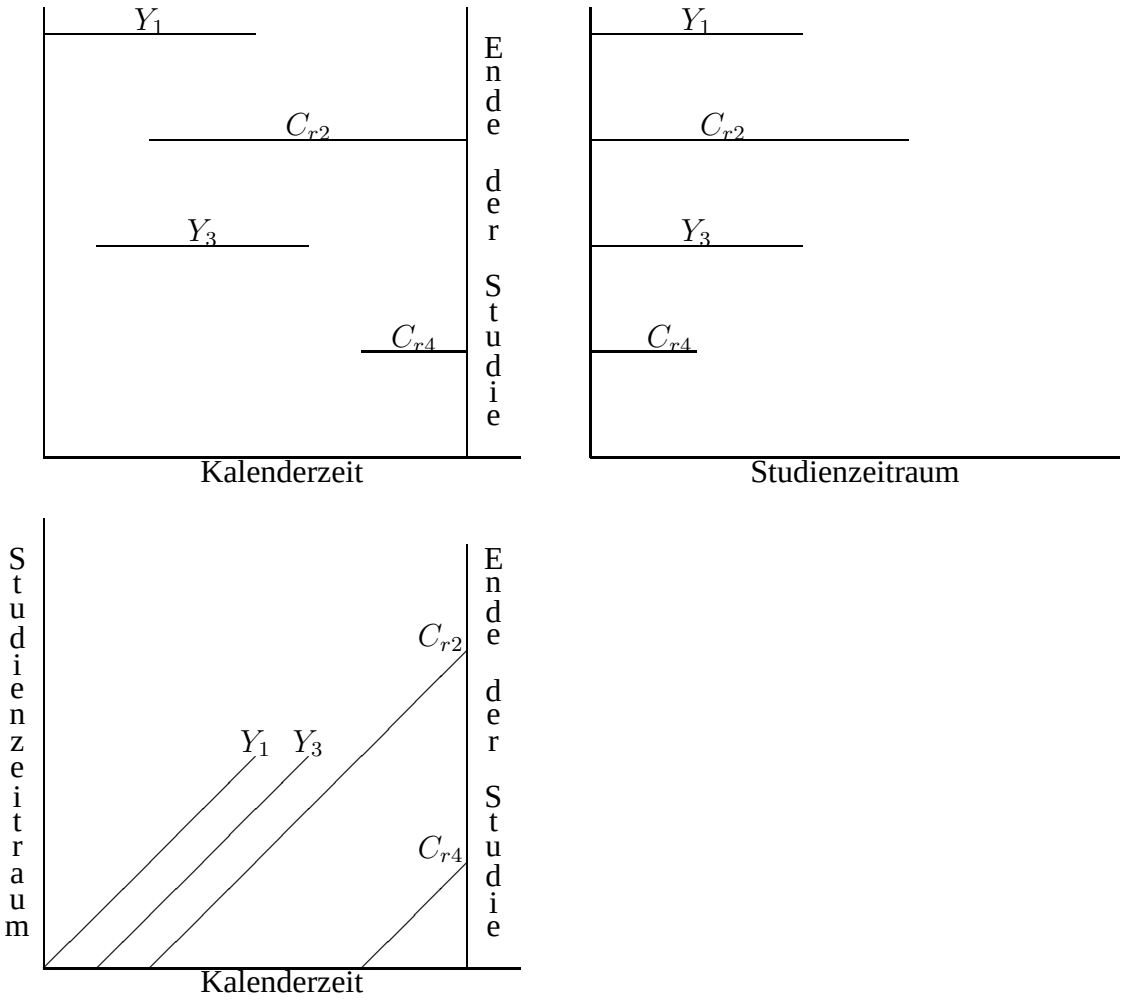


Abbildung 1.3: Beispiel einer generalisierten Typ I Rechtszensur. Diagramme: mit unterschiedlichen (tatsächlichen) Startzeiten(oben links), alle Startzeitpunkte auf 0 verschoben (rechts) und einem Lexis Diagramm (unten links).

1.1.2 TYP II RECHTSZENSUR

Nun wird eine *Typ II Rechtzensierung* betrachtet. Im Gegensatz zu einer Typ I Zensur steht hier der genaue Zeitpunkt der Zensur noch nicht fest. Stattdessen wird die Zensurschranke durch eine vorher bestimmte Anzahl von Ausfällen bestimmt. Diese Form der Zensur wird häufig bei Studien über die Lebenszeiten von Materialien, Teilen oder Maschinen gewählt.
Mit der obigen Erklärung lässt sich die *generalisierte Typ II Rechtszensierung* analog zur Typ I darstellen (mit Ausnahme des Lexis Diagrammes).
Für die *progressive Typ II Rechtszensierung* wird vor Beginn der Studie die Anzahl $(r_1, r_2, ..., r_m)$ der Objekte festgehalten, die ausfallen müssen, damit eine Zensurschranke eintritt. Gibt n die Anzahl der Versuchsobjekte zu Beginn an, so ergeben sich $n_1, n_2, ...n_m$ für die entsprechende Anzahl der Objekte, die bei Eintritt in eine neue Zeniserungsstufe ausscheiden (Folglich: $n \geq n_1 + n_2 + ... + n_m$). Somit sind zum ersten Zensierungszeitpunkt noch $n - r_1$ Objekte nicht ausgefallen. Aber es werden nur $n - n_1$ Objekte für die nächste Stufe zugelassen, also wird die Differenz ($n_1 - r_1$ Objekte) aussortiert.

1.2 GETRIMMTER MITTELWERT (NACH OBEN GETRIMMT)

Die Stichprobe $y = (y_1, ..., y_N)^T$ ist eine Realisierung der Zufallsvariable $Y = (Y_1, ..., Y_N)^T$. Seinen $y_{(1)}, ..., y_{(N)}$ die geordneten Beobachtungen mit $y_{(1)} \leq ... \leq y_{(N)}$ einer Stichprobe. Des Weiteren wird angenommen, dass die Beobachtungen i.i.d. und exponentialverteilt sind: $F_\theta(z) = \int_0^z \frac{1}{\theta} e^{-\frac{y}{\theta}} dy = 1 - e^{-\frac{z}{\theta}}$.

Nun wird eine geordnete Stichprobe y mit $y_{(1)} \leq ... \leq y_{(N)}$ betrachtet. Aus dieser wird einen Anteil $\beta \in [0, 1)$ der größten Beobachtungen herausgestrichen. Danach existieren nur noch $\lceil (1-\beta)N \rceil$ Beobachtungen. Dadurch ergibt sich ein Mittelwert, der in folgender Definition festgehalten wird.

Definition 1.2.1 *Sei $y_{(1)} \leq ... \leq y_{(N)}$ eine geordnete Stichprobe. So ist das (nach oben) getrimmte Mittel durch*

$$\overline{y}_\beta = \frac{1}{\lceil (1-\beta)N \rceil} \sum_{n=1}^{\lceil (1-\beta)N \rceil} y_{(n)}$$

gegeben.

Um diese diskrete Struktur in eine stetige zu verwandeln, wird die Empirische Verteilungsfunktion herangezogen.

1.2.1 FÜR DATEN

Definition 1.2.2 *Für Zufallssgrößen $Y_1, ..., Y_N$ heißt*

$$F_{N,Y(\omega)}(x) := \frac{1}{N}\sum_{n=1}^{N} 1_{(-\infty,x]}(Y_n(\omega))$$

die Empirische Verteilungsfunktion von $Y_1, ...Y_N$.

Die verkürzte Schreibweise

$$F_N(x) = \frac{1}{N}\sum_{n=1}^{N} 1_{(-\infty,x]}(Y_n) = \frac{\#\{Y_n \leq x\}}{N}$$

findet im folgenden immer wieder Anwendung. Da das zugehörige Wahrscheinlichkeits-Maß P_N die Form $P_N = \frac{1}{N}\sum_{n=1}^{N} e_{Y_n}$ besitzt, gilt

$$F_N(x) = \frac{1}{N}\sum_{n=1}^{N} 1_{(-\infty,x]}(Y_n) = \int_0^x y P_N(dy).$$

Hierbei stellt e_{Y_n} das Dirac-Maß dar.

Definition 1.2.3 *Sei F eine Verteilungsfunktion, dann ist*

$$F^{-1}(x) = \inf\{z; F(z) \geq x\}$$

ihr zugehöriges Quantil.

Mit Hilfe dieser zwei Definition wird das Quantil der empirischen Verteilungsfunktion betrachtet.

Satz 1.2.4 *Für eine geordnete Stichprobe $y_{(1)} \leq ... \leq y_{(N)}$ gilt*

$$F^{-1}(1-\beta) = y_{(\lceil(1-\beta)N\rceil)}$$

mit $\beta\ in[0,1)$ und $N \in \mathbb{N}$.

Beweis:

$$\begin{aligned} F^{-1}(1-\beta) &= \inf\{y_n; F_N(y_n) \geq (1-\beta)\} \\ &= \inf\left\{y_n; \frac{\#(Y_n \leq y_n)}{N} \geq (1-\beta)\right\} \\ &= \inf\{y_n; \#(Y_n \leq y_n) \geq (1-\beta)N\} \end{aligned}$$

Da es sich hier um eine geordnete Stichprobe handelt, gilt:

$$\begin{aligned} F^{-1}(1-\beta) &= \inf\{y_n; \text{rang}(y_n) \geq (1-\beta)N\} \\ &= y_{(\lceil(1-\beta)N\rceil)} \end{aligned}$$

□

Lemma 1.2.5 *Sei F_N die Empirische Verteilungsfunktion und P_N die zugehörige Verteilungsfunktion. So gilt für ausreichend großes N ihr getrimmtes Mittel*

$$\overline{y_\beta} = \frac{1}{1-\beta}\int_0^{F_N^{-1}(1-\beta)} yP_N(dy).$$

Beweis: Sei $y_{(1)} \leq ... \leq y_{(N)}$ eine geordnete Stichprobe.

$$\begin{aligned} \overline{y}_\beta &= \frac{1}{\lceil(1-\beta)N\rceil}\sum_{n=1}^{\lceil(1-\beta)N\rceil} y_{(n)} \\ &= \frac{1}{\lceil(1-\beta)N\rceil}\sum_{n=1}^{N} y_{(n)} \cdot 1_{(-\infty, y_{(\lceil(1-\beta)N\rceil)}]}(y_{(n)}) \\ &= \frac{1}{\lceil(1-\beta)N\rceil}\int N \cdot y \cdot 1_{(-\infty, y_{(\lceil(1-\beta)N\rceil)}]}(y) \cdot P_N(dy) \\ &= \frac{N}{\lceil(1-\beta)N\rceil}\int_0^{y_{(\lceil(1-\beta)N\rceil)}} yP_N(dy) \end{aligned}$$

und mit Lemma 1.2.4 folgt

$$\overline{y}_\beta = \frac{N}{\lceil (1-\beta)N \rceil} \int_0^{F_N^{-1}(1-\beta)} y P_N(dy).$$

Mit hinreichend großem N gilt:

$$\begin{aligned} & \lceil (1-\beta)N \rceil \approx (1-\beta)N \\ \Rightarrow \quad & \frac{N}{\lceil (1-\beta)N \rceil} \int_0^{F_N^{-1}(1-\beta)} y P_N(dy) \approx \frac{1}{(1-\beta)} \int_0^{F_N^{-1}(1-\beta)} y P_N(dy) \end{aligned}$$

□

1.2.2 Für Verteilungsfunktionen

Nachdem der getrimmten Mittelwert für Daten untersucht wurde, wird der getrimmten Mittelwert für Verteilungsfunktionen diskutiert.

Definition 1.2.6 *Sei F eine getrimmte Verteilungsfunktion. So bildet das Funktional*

$$T(F) = \frac{1}{1-\beta} \int_0^{F^{-1}(1-\beta)} y P(dy)$$

auf den getrimmten Mittelwert ab.

Zuerst wird das Quantile und seine Konvergenzeigenschaften genauer untersucht, bevor das Funktional betrachtet wird.

Satz 1.2.7 *Sei $n \in \mathbb{N}$, $\alpha \in (0,1)$, zudem sei $F(F^{-1}(\alpha) + \varepsilon) > \alpha$ für alle $\varepsilon > 0$. Dann gilt:*

$$F_{N,\omega}^{-1}(\alpha) \xrightarrow{f.s.} F^{-1}(\alpha).$$

Beweis: Zur besseren Lesbarkeit wird $q := F^{-1}(\alpha)$ und $F_N := F_{N,\omega}$ gesetzt.
Sei $\varepsilon > 0$, dann gibt es ein $\delta > 0$ mit

$$F(F^{-1}(\alpha)-\varepsilon) \leq \alpha-\delta \quad , \quad F(F^{-1}(\alpha)+\varepsilon) \geq \alpha+\delta,$$

bzw.

$$F(q-\varepsilon) \leq \alpha-\delta \quad , \quad F(q+\varepsilon) \geq \alpha+\delta.$$

$$\begin{aligned}
&\lim_{M\to\infty} P\left(\bigcup_{N\geq M}\left\{|F_N(q-\varepsilon)-F(q-\varepsilon)| \geq \frac{\delta}{2} \quad \text{oder} \quad \mid F_N(q+\varepsilon)-F(q+\varepsilon) \mid \geq \frac{\delta}{2}\right\}\right)\\
&\leq \lim_{M\to\infty} P\left(\bigcup_{N\geq M}\left\{\mid F_N(q-\varepsilon)-F(q-\varepsilon) \mid \geq \frac{\delta}{2}\right\}\right) + \lim_{M\to\infty} P\left(\left\{\mid F_N(q+\varepsilon)-F(q+\varepsilon) \mid \geq \frac{\delta}{2}\right\}\right)\\
&= 0
\end{aligned}$$

Wegen dem Satz von Glivenko-Cantelli streben beide Summanden gegen Null. Außerdem folgt, wegen der gezeigten Fast-sicheren-Konvergenz:

$$\begin{aligned}
1 &= \lim_{M\to\infty} P\left(\bigcap_{N\geq M}\left\{\mid F_N(q-\varepsilon)-F(q-\varepsilon) \mid < \frac{\delta}{2} \quad \text{und} \quad \mid F_N(q+\varepsilon)-F(q+\varepsilon) \mid < \frac{\delta}{2}\right\}\right)\\
&\leq \lim_{M\to\infty} P\left(\bigcap_{N\geq M}\left\{F_N(q-\varepsilon) < F(q-\varepsilon)+\frac{\delta}{2} \quad \text{und} \quad F_N(q+\varepsilon) > F(q+\varepsilon)-\frac{\delta}{2}\right\}\right)\\
&\leq \lim_{M\to\infty} P\left(\bigcap_{N\geq M}\left\{F_N(q-\varepsilon) < \alpha-\frac{\delta}{2} \quad \text{und} \quad F_N(q+\varepsilon) > \alpha+\frac{\delta}{2}\right\}\right)\\
&\leq \lim_{M\to\infty} P\left(\bigcap_{N\geq M}\left\{F_N(F^{-1}(\alpha)-\varepsilon) < \alpha \quad \text{und} \quad F_N(F^{-1}(\alpha)+\varepsilon) > \alpha\right\}\right)\\
&\leq \lim_{M\to\infty} P\left(\bigcap_{N\geq M}\left\{\mid F_N^{-1}(\alpha)-F^{-1}(\alpha) \mid < \varepsilon\right\}\right)
\end{aligned}$$

□

Die Bedingung des monotonen Wachstums rechtsseitig von einem $\alpha \in (0,1)$ ($F(F^{-1}(\alpha)+\varepsilon) > \alpha$) ist bei der Verteilungsfunktion für jedes α erfüllt.
Somit kann die fast sichere Konvergenz des getrimmten Mittelwertes hergeleitet werden, indem als Grundlage die vorherigen Erkenntnisse eingesetzt werden.

Satz 1.2.8

$$\frac{1}{1-\beta}\int_0^{F_N^{-1}(1-\beta)} yP_N(dy) \xrightarrow{f.s.} \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yP(dy)$$

Beweis: Nach dem starkem Gesetz der großen Zahlen [GE07, S.120ff.] gilt:

$$\begin{aligned}
\frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP_N(dy) &= \frac{1}{1-\beta}\frac{1}{N}\sum_{n=1}^{N} 1_{(0,F^{-1}(1-\beta)+\frac{1}{l})}(Y_n) \\
&\xrightarrow{f.s.} \frac{1}{1-\beta}E_P(1_{(0,F^{-1}(1-\beta)+\frac{1}{l})}(Y_n)) \\
&= \frac{1}{1-\beta}\int_0^{\infty} 1_{(0,F^{-1}(1-\beta)+\frac{1}{l})}(y)P(dy) \\
&= \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP(dy)
\end{aligned}$$

für alle $l \in \mathbb{N}$. Ebenso gilt

$$\frac{1}{(1-\beta)}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} P_N(dy) \xrightarrow{f.s.} \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} P(dy)$$

für alle $l \in \mathbb{N}$. Sei $B_{k,l} := \left\{\omega; \exists N_0(\omega)\forall N \geq N_0(\omega) :\mid Q_{+l} \mid < \frac{1}{k} \wedge \mid Q_{-l} \mid < \frac{1}{k}\right\}$ mit

$$Q_{+l} := \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP_{N,\omega}(dy) - \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP(dy)$$

und

$$Q_{-l} := \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} yP_{N,\omega}(dy) - \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} yP(dy).$$

Wegen $B_{1,l} \supseteq B_{2,l} \supseteq ... \supseteq B_{k,l}$ gilt

$$\bigcap_{k\in\mathbb{N}} B_{k,l} = \left\{\omega; \lim_{N\to\infty} \int_0^{F^{-1}(1-\beta)\pm\frac{1}{l}} yP_{N,\omega}(dy) = \int_0^{F^{-1}(1-\beta)\pm\frac{1}{l}} yP(dy)\right\}$$

und

$$P\left(\bigcap_{k\in\mathbb{N}} B_{k,l}\right) = 1.$$

Durch die Eigenschaft

$$P\left(\bigcap_{n\in I} A_n\right) \leq P(A_i) \quad \forall i \in I, \quad I \text{ Indexmenge}$$

des Wahrscheinlichkeitsmaßes P ergibt sich

$$P(B_{k,l}) = 1 \quad \forall k, l \in \mathbb{N}.$$

Mit $(1-\beta) \in (0,1)$ und Satz 1.2.7 folgt

$$F_{N,\omega}^{-1}(1-\beta) \xrightarrow{f.s.} F^{-1}(1-\beta).$$

Sei

$$C_l := \left\{\omega; \exists N_0(\omega) \forall N \geq N_0(\omega) :\mid F_{N,\omega}^{-1}(1-\beta) - F^{-1}(1-\beta) \mid < \frac{1}{l}\right\}.$$

Mit den selben Argumenten wie bei $B_{k,l}$ gilt

$$1 = P\left(\bigcap_{l\in\mathbb{N}} C_l\right)$$

und

$$P(C_l) = 1 \quad \forall \in \mathbb{N}.$$

Da P eine stetige Verteilungsfunktion ist, gibt es für alle $k \in \mathbb{N}$ ein $l = l(k) \in \mathbb{N}$ mit

$$\left|\frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP(dy) - \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yP(dy)\right| < \frac{1}{k}$$

und

$$\left|\frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} yP(dy) - \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yP(dy)\right| < \frac{1}{k}.$$

Sei $k \in \mathbb{N}$ beliebig und $l = l(k)$. Dann gilt für alle $\omega \in B_{k,l} \cap C_l$:

$$
\begin{aligned}
&\frac{1}{1-\beta}\frac{1}{N}\sum_{n=1}^{N} 1_{(0,F_{N,\omega}^{-1}(1-\beta))}(Y_n(\omega)) \\
&= \frac{1}{1-\beta}\int_0^{F_{N,\omega}^{-1}(1-\beta)} yP_{N,\omega}(dy) \\
&\leq^{\omega\in C_l} \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP_{N,\omega}(dy) \\
&= \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP_{N,\omega}(dy) - \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP(dy) + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP(dy) \\
&\leq^{\omega\in B_{k,l}} \frac{1}{k} + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP_(dy) \\
&= \frac{1}{k} + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)+\frac{1}{l}} yP(dy) - \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yP(dy) + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yP(dy) \\
&\leq^{l=l(k)} \frac{2}{k} + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yP(dy) \quad ,\forall N \geq N_0(\omega)
\end{aligned}
$$

Analog folgt

$$
\begin{aligned}
&\frac{1}{1-\beta}\frac{1}{N}\sum_{n=1}^{N} 1_{(0,F_{N,\omega}^{-1}(1-\beta))}(Y_n(\omega)) \\
&= \frac{1}{1-\beta}\int_0^{F_{N,\omega}^{-1}(1-\beta)} yP_{N,\omega}(dy) \\
&\geq^{\omega\in C_l} \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} yP_{N,\omega}(dy) \\
&= \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} yP_{N,\omega}(dy) - \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} yP(dy) + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} yP(dy) \\
&\geq^{\omega\in B_{k,l}} -\frac{1}{k} + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} yP(dy) \\
&= -\frac{1}{k} + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)-\frac{1}{l}} yP(dy) - \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yP(dy) + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yP(dy) \\
&\geq -\frac{2}{k} + \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yP(dy) \quad ,\forall N \geq N_0(\omega)
\end{aligned}
$$

Sei

$$A_k := \left\{ \omega; \exists N_0(\omega) \forall N \geq N_0(\omega) : \left| \frac{1}{1-\beta} \int_0^{F_{N,\omega}^{-1}(1-\beta)} y P_{N,\omega}(dy) - \frac{1}{1-\beta} \int_0^{F^{-1}(1-\beta)} y P(dy) \right| \leq \frac{2}{k} \right\}.$$

Also gilt

$$B_{k,l} \cap C_l \subseteq A_k$$

und somit folgt

$$1 = P(B_{k,l} \cap C_l) \leq P(A_k).$$

Mit der σ-Stetigkeit folgt

$$\begin{aligned}
&P\left(\omega; \lim_{N\to\infty} \frac{1}{1-\beta} \int_0^{F_{N,\omega}^{-1}(1-\beta)} y P_{N,\omega}(dy) = \frac{1}{1-\beta} \int_0^{F^{-1}(1-\beta)} y P(dy)\right) \\
&= P\left(\bigcap_{k\in\mathbb{N}} A_k\right) \\
&= P\left(\lim_{k\to\infty} A_k\right) \\
&= \lim_{k\to\infty} P\left(A_k\right) \\
&= 1.
\end{aligned}$$

$\square$

KAPITEL 2

MAXIMUM-LIKELIHOOD-METHODE

2.1 FÜR DATEN UND VERTEILUNGSFUNKTIONEN

An dieser Stelle wird die Maximum-Likelihood-Methode [CZ11, S.83ff] vorgestellt, die ein effizientes Werkzeug zur Bestimmung von Schätzern darstellt. Allerdings ist es ein großer Nachteil, dass eine Annahme über die Dichtefunktion, in Form eines statistischen regulären Modells $\mathcal{P}$, vorausgesetzt werden muss.

Definition 2.1.1 *Ein statistisches Modell ist eine Familie $\mathcal{P}$ von Verteilungsfunktionen. Für ein statistisches Modell $\mathcal{P}$ wird stets die Darstellung*

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$$

verwendet, wobei $\mathbb{P}_\theta$ für alle $\theta \in \Theta$ ein Wahrscheinlichkeitsmaß ist. Θ heißt Parameterraum.

Definition 2.1.2 *Ein statistisches Modell $\mathcal{P}$ heißt regulär, falls eine der folgenden Bedingungen erfüllt ist:*

(i) Alle $\mathbb{P}_\theta$, $\theta \in \Theta$, sind stetig mit Dichtefunktion $p_\theta(y)$.

(ii) Alle $\mathbb{P}_\theta$, $\theta \in \Theta$, sind diskret mit Wahrscheinlichkeitsfunktion $p_\theta(y)$.

Die Idee hinter der Maximum-Likelihood-Methode ist es, den Schätzwert so zu wählen, dass die Wahrscheinlichkeitsdichte für die Beobachtung y maximal ist. Zu diesem Zweck

wird die Likelihood-Funktion eingeführt.

Definition 2.1.3 *Die Funktion $L : \Theta \times \mathbb{R}^n \to \mathbb{R}^+$, gegeben durch*

$$L(\theta, y) := p(y, \theta)$$

mit $\theta \in \Theta$, $y \in \mathbb{R}^n$ heißt Likelihood-Funktion des Parameters θ für die Beobachtung y.

Im diskreten Fall gibt $L(\theta, y)$ an unter welcher Wahrscheinlichkeit die Beobachtung y unter dem Parameter θ eintritt. Das heißt vor allem, dass $L(\theta, y)$ als Maß bzw. Bewertung angesehen werden kann, die angibt wie wahrscheinlich der Parameter θ ist, falls y beobachtet wird.

Im stetigen Fall wird eine ϵ-Umgebung um y angenommen und ϵ gegen Null gehen gelassen, um dann ein Analogon zum stetigen Fall zu erlangen.

Mit der Likelihood-Funktion ist es möglich Schätzer zu bewerten. Im Speziellen kann das Maximum über all diese Bewertungen gesucht werden, um dann, wie oben erwähnt, den maximalen Schätzer zu bestimmen. Dies wird im Folgenden nochmals definiert [CZ11].

Definition 2.1.4 *Gibt es in dem regulären statistischen Modell $\mathcal{P}$ eine meßbare Funktion $\hat{\theta} : \mathbb{R}^n \mapsto \Theta$, so dass*

$$L(\hat{\theta}(y), y) = \max\{L(\theta, y) : \theta \in \Theta\} \qquad \textit{für alle} \quad y \in \mathbb{R}^n,$$

so heißt $\hat{\theta}(Y)$ Maximum-Likelihood-Schätzer (MLS) bzw. Maximum-Likelihood-Estimate (MLE) von θ.

In der Praxis hat es sich bewährt den (natürlichen) Logarithmus der Likelihood-Funktion zu betrachten. Die Maximalität unter der Transformation des Logarithmus bliebt erhalten, da es sich bei dem Logarithmus um eine streng monoton wachsende Funktion handelt.

Definition 2.1.5 *Die Log-Likelihood-Funktion $logL : \Theta \times \mathbb{R}^n \to \mathbb{R}$ ist definiert durch*

$$logL(\theta, y) := \ln(L(\theta, y)).$$

2.2 FÜR DIE NORMALVERTEILUNG

Als erste Anwendung des MLS wird die eindimensionale Normalverteilung mit bekannter Varianz σ^2 [CZ11, S.86f] betrachtet.

Lemma 2.2.1 *Sei* $f_\theta(y_n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$ *die Dichtefunktion der Normalverteilung und die Varianz* σ^2 *bekannt. Dann ist der Maximum-Likelihood-Schätzer gegeben durch*

$$\hat{\theta} = \overline{y}.$$

Beweis: Seien $Y_1, ..., Y_n$ i.i.d. mit $Y \sim N(\theta, \sigma^2)$. In Folge dessen ergibt sich die Likelihood-Funktion:

$$\begin{aligned} L(\theta, y) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}} \\ &= \frac{1}{\sigma\sqrt{2\pi}} \prod_{i=1}^{n} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}}. \end{aligned}$$

Wird auf $L(\theta, y)$ die Log-Likelihood-Funktion angewandt, so ergibt sich die Gleichung

$$logL(\theta, y) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(y_i - \theta)^2}{2\sigma^2}.$$

Durch Differentiation nach θ wird der mögliche MLS bestimmt,

$$\begin{aligned}\frac{\partial logL(\theta,y)}{\partial\theta} &= -\sum_{i=1}^{n} -\frac{2}{2\sigma^2}(y_i-\theta)\\ &= \sum_{i=1}^{n}\frac{1}{\sigma^2}(y_i-\theta)=0\\ \Rightarrow \theta &= \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}.\end{aligned}$$

Mit der zweiten Ableitung wird überprüft, ob es sich um ein Maximum und damit um den gewünschte MLS handelt. Also folgt mit einer weiteren Differentiation nach θ

$$\begin{aligned}\frac{\partial^2 logL(\theta,y)}{\partial\theta^2} &= \sum_{i=1}^{n}\frac{1}{\sigma^2}(-1)\\ &= -\frac{n}{\sigma^2} < 0, \quad \text{da } n\in\mathbb{N}, \sigma^2>0.\end{aligned}$$

Das Ergebnis ist immer kleiner Null und θ ist der gesuchte MLS, also $\hat{\theta}=\overline{y}$.

□

2.3 FÜR DIE EXPONENTIALVERTEILUNG

Zur einführenden Betrachtung dient die Standard-Exponentialverteilung ohne Trimmung und ohne Zensierung [SC03, S.293].

Lemma 2.3.1 *Sei $f(y_n)=\lambda e^{-\lambda y_n}$ die Dichtefunktion der Exponetialverteilung. Dann ist für sie der Maximum-Likelihood-Schätzer gegeben durch*

$$\hat{\lambda}=\frac{1}{\overline{y}}.$$

Außerdem ist für die Dichtefunktion $f(y_n) = \frac{1}{\theta}e^{-\frac{y_n}{\theta}}$ der Exponentialverteilung

$$\hat{\theta} = \overline{y}$$

der Maximum-Likelihood-Schätzer.

Beweis: Seien die Beobachtungen $y_1, ...y_n$ i.i.d. und die Dichtefunktion durch $f(y_n) = \lambda e^{-\lambda y_n}$ gegeben.

$$f(y_n) = \lambda e^{-\lambda y_n} \cdot 1_{[0,\infty}(y_n)$$

Somit lautet die Likelihood-Funktion:

$$\begin{aligned} L(\lambda, y) &= \prod_{n=1}^{N} \lambda e^{-\lambda y_n} \\ logL(\lambda, y) &= \sum_{n=1}^{N} (\log(\lambda) - \lambda y_n) \\ \frac{\partial logL(\lambda, y)}{\partial \lambda} &= \sum_{n=1}^{N} \left(\frac{1}{\lambda} - y_n\right) = 0 \\ \frac{1}{\lambda} &= \frac{1}{N}\sum_{n=1}^{N} y_n = \overline{y}. \end{aligned}$$

Daher ist wegen $\theta = \frac{1}{\lambda} = \overline{y}$ das arithmetische Mittel der Maximum-Likelihood Schätzer für die Exponentialverteilung, vorausgesetzt es handelt sich wirklich um ein Maximum. Dazu wird die nächste Differentiation betrachtet.

$$\frac{\partial^2 logL(\lambda, y)}{\partial^2 \lambda} = \sum_{n=1}^{N} \left(-\frac{1}{\lambda^2}\right) < 0$$

Die zweite Ableitung ist damit kleiner Null und der Maximum-Likelihood-Schätzer ist $\hat{\theta} = \bar{y}$ $\square$

2.4 FÜR ZENSIERTE EXPONENTIALVERTEILUNGEN

Nachdem der MLS für die Exponentialverteilung bestimmt wurde und bevor auf die Frage der Trimmung von Daten eingegangen wird, wird der MLS für die zensierte Exponentialverteilung entwickelt.

Dafür wird zunächst die Survival Funktion benötigt und diese ist geben durch:

$$S_\theta(y) = P(Y_n \geq y) = 1 - F_\theta(y) = e^{-\frac{y}{\theta}} \quad , \quad 0 < y < \infty$$

Eine weitere wichtige Funktion ist die Hazard Funktion:

$$\lambda_\theta(y) = \frac{f_\theta(y)}{S_\theta(y)} = \frac{\frac{1}{\theta}e^{-y/\theta}}{e^{-y/\theta}} = \frac{1}{\theta}$$

Außerdem wird in einer Indikatorvariable δ_n eine erfolgte Zensierung festgehalten.

$$\delta_n = \begin{cases} 1, & \text{wenn } Y_n \leq C \quad \text{(unzensiert)} \\ 0, & \text{wenn } Y_n > C \quad \text{(zensiert)} \end{cases}$$

Definition 2.4.1 *Sei Y_n die unzensierte Zufallsvariable und C die Zensurschranke, dann ist mit*

$$Z_n := \min(Y_n, C)$$

die erweiterte Zufallsvariable gegeben.

Für die Likelihood-Funktion ergibt sich somit folgendes:

$$\begin{aligned}
L(\theta, y) &= \prod_{n=1}^{N} f_\theta(z_n)^{\delta_n} S_\theta(z_n)^{1-\delta_n} \\
&= \prod_{n=1}^{N} f_\theta(z_n)^{\delta_n} S_\theta(z_n)^{-\delta_n} S_\theta(z_n)^{\delta_n} S_\theta(z_n)^{1-\delta_n} \\
&= \prod_{n=1}^{N} \lambda_\theta(z_n)^{\delta_n} S_\theta(z_n) \\
&= \prod_{n=1}^{N} \left(\frac{1}{\theta}\right)^{\delta_n} e^{-z_n/\theta}.
\end{aligned}$$

Definition 2.4.2 *Die Likelihood-Funktion für die zensierte Exponentialverteilung ist gegeben durch*

$$L(\theta, y) = \prod_{n=1}^{N} \left(\frac{1}{\theta}\right)^{\delta_n} e^{-z_n/\theta}.$$

Mit dieser Funktion kann die Herleitung eines Maximum-Likelihood Schätzer für zensierte Exponentialverteilung in Angriff genommen werden.

Lemma 2.4.3 *Sei $F_\theta(y) = 1 - e^{-\frac{y}{\theta}}$, C die Zensurschranke und n_{uc} die Anzahl der unzensierten Beobachtungen. Dann ist der Maximum-Likelihood-Schätzer gegeben durch*

$$\hat{\theta}_{mle} = \frac{\sum_{n=1}^{N} y_n}{n_{uc}}.$$

Beweis: Sei die erweiterte Zufallsvariable Z_n gegeben.

$$\begin{aligned}
logL(\theta, y) &= -\sum_{n=1}^{N} \delta_n \log(\theta) - \frac{1}{\theta} \sum_{n=1}^{N} z_n \\
\frac{\partial logL(\theta, y)}{\partial \theta} &= -\frac{1}{\theta} \sum_{n=1}^{N} \delta_n + \frac{1}{\theta^2} \sum_{n=1}^{N} z_n = 0
\end{aligned}$$

Somit folgt für den Maximum-Likelihood-Schätzer

$$\begin{aligned}\theta \sum_{n=1}^{N} \delta_n &= \sum_{n=1}^{N} z_n \\ \theta &= \frac{\sum_{n=1}^{N} z_n}{\sum_{n=1}^{N} \delta_n} = \frac{\sum_{n=1}^{N} z_n}{n_{uc}}.\end{aligned}$$

Bleibt nun noch die Frage zu klären, ob es sich wirklich um eine Maximum handelt. Dazu betrachten wir die zweite Ableitung an der Stelle des Schätzers.

$$\begin{aligned}\frac{\partial^2 logL(\theta, y)}{\partial \theta^2} &= \frac{1}{\theta^2} \sum_{n=1}^{N} \delta_n - \frac{2}{\theta^3} \sum_{n=1}^{N} z_n \\ &= \frac{n_{uc}^2}{(\sum_{n=1}^{N} z_n)^2} \cdot n_{uc} - 2 \frac{n_{uc}^3}{(\sum_{n=1}^{N} z_n)^3} \sum_{n=1}^{N} z_n \\ &= \frac{n_{uc}^3}{(\sum_{n=1}^{N} z_n)^2} - 2 \frac{n_{uc}^3}{(\sum_{n=1}^{N} z_n)^2} \\ &= -\frac{n_{uc}^3}{(\sum_{n=1}^{N} z_n)^2} < 0.\end{aligned}$$

Damit ist der gesuchte MLS $\theta = \frac{\sum_{n=1}^{N} z_n}{n_{uc}}$.

□

Lemma 2.4.4 *Sei F_N die Empirische Verteilungsfunktion,die Daten exponentialverteilt, C die Zensurschranke und n_{uc} die Anzahl der unzensierten Beobachtungen. Dann gilt für die Maximum-Likelihood-Schätzer*

$$\hat{\theta}_{mle} = \frac{N \int_0^C y dF_N(y) + (N - NF_N(C))C}{NF_N(C)}.$$

Beweis: Als Grundlage dient der Schätzer aus Lemma 2.4.3.

$$\begin{aligned}\hat{\theta} &= \frac{\sum_{n=1}^{N} y_n}{n_{uc}} \\ &= \frac{\sum_{n=1}^{N} Y_n I(Y_n \leq C) + (N - n_{uc})C}{n_{uc}}.\end{aligned}$$

Mit der Empirschen Verteilungsfunktion F_N und $n_{uc} = \#\{Y_n \leq C\} = N \cdot F_N(C)$ folgt

$$\begin{aligned}\hat{\theta} &= \frac{N \int_0^C y dF_N(y) + (N - n_{uc})C}{n_{uc}} \\ &= \frac{N \int_0^C y dF_N(y) + (N - NF_N(C))C}{NF_N(C)} \\ &= \frac{\int_0^C y dF_N(y) + (1 - F_N(C))C}{F_N(C)}.\end{aligned}$$

□

KAPITEL 3

KONSISTENZ

3.1 KONSISTENZ DES GETRIMMTEN MITTELWERTES

Bevor sich der Konsistenz des getrimmten Mittelwert zugewendet wird, werden zwei Vorüberlegungen gemacht. Als erstes wird das Hauptaugenmerk auf das Integral gelegt. Hierbei wird die obere Grenze gleich einer Konstanten gesetzt. Als zweites wird dann diese Grenze untersucht.

Lemma 3.1.1

$$\int_0^C y\frac{1}{\theta}e^{-\frac{y}{\theta}}dy = \theta - \theta e^{-\frac{C}{\theta}} - Ce^{-\frac{C}{\theta}}$$

Beweis: Mit Partielle Integration ergibt sich:

$$\begin{aligned}\int_0^C y\frac{1}{\theta}e^{-\frac{y}{\theta}}dy &= -ye^{-\frac{y}{\theta}}\,|_0^C + \int_0^C e^{-\frac{y}{\theta}}dy \\ &= -Ce^{-\frac{C}{\theta}} + \theta\int_0^C \frac{1}{\theta}e^{-\frac{y}{\theta}}dy.\end{aligned}$$

Sei $F_\theta(C)$ die Verteilungsfunktion der Exponentialverteilung in den Grenzen 0 und C.

Dann gilt

$$\begin{aligned}\int_0^C y\frac{1}{\theta}e^{-\frac{y}{\theta}}dy &= -Ce^{-\frac{C}{\theta}} + \theta F_\theta(C)\\ &= -Ce^{-\frac{C}{\theta}} + \theta(1-e^{-\frac{C}{\theta}})\\ &= \theta - \theta e^{-\frac{C}{\theta}} - Ce^{-\frac{C}{\theta}}.\end{aligned}$$

□

Als nächstes wird die obere Grenze $F_\theta^{-1}(1-\beta)$ betrachtet.
Zur Vereinfachung wird $q := 1-\beta$ gesetzt.

Lemma 3.1.2

$$F_\theta^{-1}(q) = -\theta\ln(1-q)$$

Beweis:

$$\begin{aligned}& F_\theta(z) = q\\ \Leftrightarrow\quad & 1-e^{-\frac{z}{\theta}} = q\\ \Leftrightarrow\quad & 1-q = e^{-\frac{z}{\theta}}\\ \Leftrightarrow\quad & \ln(1-q) = -\frac{z}{\theta}\\ \Leftrightarrow\quad & \theta\ln(1-q) = -z\\ \Leftrightarrow\quad & z = -\theta\ln(1-q)\end{aligned}$$

□

Mit Hilfe dieser Erkenntnisse, wird sich wieder dem getrimmten Mittelwert für unzensierten Daten zugewandt.

Lemma 3.1.3 *Wird die Exponentialverteilung $F_\theta(x) = 1 - e^{-\frac{x}{\theta}}$ und das Funktional*

$$T_\beta(F) = \frac{1}{1-\beta} \int_0^{F^{-1}(1-\beta)} yF(dy)$$

als getrimmtes Mittel zugrunde gelegt, dann gilt

$$T_\beta(F_\theta) = \frac{1-\beta+\beta\ln(\beta)}{1-\beta}\theta.$$

Beweis: Wegen Lemma 3.1.2 gilt

$$F_\theta^{-1}(1-\beta) = -\theta\ln(\beta) \quad ,$$

und somit folgt:

$$T_\beta(F_\theta) = \frac{1}{1-\beta} \int_0^{-\theta\ln(\beta)} y\frac{1}{\theta}e^{-\frac{y}{\theta}}dy$$

Unter zu Hilfenahme von Lemma 3.1.1 ergibt sich

$$\begin{aligned}
T_\beta(F_\theta) &= \frac{1}{1-\beta}\left[\theta - \theta e^{\frac{\theta\ln(\beta)}{\theta}} + \theta(\ln(\beta))e^{\frac{\theta\ln(\beta)}{\theta}}\right] \\
&= \frac{1}{1-\beta}\left[\theta - \theta e^{\ln(\beta)} + \theta(\ln(\beta))e^{\ln(\beta)}\right] \\
&= \frac{1}{1-\beta}\theta\left[1-\beta+\beta(\ln(\beta))\right] \\
&= \frac{1-\beta+\beta(\ln(\beta))}{1-\beta}\theta.
\end{aligned}$$

□

Das Funktional des getrimmten Mittelwerts ist nicht unverzerrt. Die folgende Definition liefert aber eine recht ähnliche Eigenschaft.

Definition 3.1.4 *Sei $\mathcal{F}$ eine Familie von Verteilungsfunktionen, dann heißt eine Funktional $T : \mathcal{F} \to \mathbb{R}$ Fischer konsistent für die parametrische Familie $\{F_\theta; \theta \in \Theta\}$, falls*

$$T(F_\theta) = \theta$$

für alle $\theta \in \Theta$ gilt.

Mit Hilfe dieser Definition wird die Möglichkeit gegeben, eine Eigenschaft des getrimmten Mittelwertes festzustellen.

Satz 3.1.5 *Für die Exponentialverteilung ist das Funktional*

$$\widetilde{T}_\beta(F) = \frac{1}{1-\beta+\beta\ln(\beta)} \int_0^{F^{-1}(1-\beta)} yF(dy)$$

Fisher konsistent.

Beweis: Wegen Lemma 3.1.3 gilt:

$$\begin{aligned}
\widetilde{T}_\beta(F_\theta) &= \frac{1-\beta}{1-\beta+\beta\ln(\beta)} \cdot \frac{1}{1-\beta} \int_0^{F_\theta^{-1}(1-\beta)} yF_\theta(dy) \\
&= \frac{1-\beta}{1-\beta+\beta\ln(\beta)} \cdot \frac{1-\beta+\beta\ln(\beta)}{1-\beta}\theta \\
&= \theta.
\end{aligned}$$

□

Zur Veranschaulichung wird die unzensierte Empirische Verteilungsfunktion F_N als Funktional in diesem Schätzer betrachtet.

Korollar 3.1.6 *Die Schätzfunktion*

$$\hat{\theta}(Y) = \widetilde{T}_0(F_{N,Y}) = \widetilde{T}(F_N)$$

ist eine konsistente Schätzung für θ.

Beweis: Da der Schätzer ungetrimmt ist, gilt $\beta = 0$:

$$\begin{aligned}\widetilde{T}(F_N) &= \frac{1}{1-0+0\ln(0)}\int_0^\infty yF_n(dy)\\ &= \overline{y_.}\end{aligned}$$

und folglich gilt

$$\begin{aligned}\lim_{N\to\infty}\widetilde{T}(F_N) &= \lim_{N\to\infty}(\overline{y_.})\\ &= E(y)\\ &= \theta.\end{aligned}$$

□

Korollar 3.1.7 *Die Schätzfunktion*

$$\hat{\theta}(Y) = \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yF(dy)$$

ist eine konsistente Schätzung für θ.

Beweis:

$$\begin{aligned}\widetilde{T_\beta}(F_N) &= \frac{1}{1-\beta+\beta\ln(\beta)}\int_0^{F^{-1}(1-\beta)} yF_N(dy)\\ &= \frac{1-\beta}{1-\beta+\beta\ln(\beta)}\frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yF_n(dy)\\ &\overset{Lem 1.2.5}{=} \frac{1-\beta}{1-\beta+\beta\ln(\beta)}\overline{y_\beta}\end{aligned}$$

und folglich gilt

$$\begin{aligned}\lim_{N\to\infty} \widetilde{T_\beta}(F_N) &= \lim_{N\to\infty} \frac{1-\beta}{1-\beta+\beta\ln(\beta)}\overline{y_\beta} \\ &= \frac{1-\beta}{1-\beta+\beta\ln(\beta)} \lim_{N\to\infty} \overline{y_\beta} \\ &= \frac{1-\beta}{1-\beta+\beta\ln(\beta)}\theta.\end{aligned}$$

$$\begin{aligned}\Rightarrow \hat{\theta}(Y) = \frac{1-\beta+\beta\ln(\beta)}{1-\beta}\widetilde{T_\beta}(F_N) &= \frac{1}{1-\beta}\int_0^{F_N^{-1}(1-\beta)} yF_N(dy) \\ &\xrightarrow{f.s.} \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yF(dy)\end{aligned}$$

□

3.2 Konsistenz des getrimmten Mittelwertes bei zensierten Daten

Wie im vorherigen Abschnitt wird hier die Konsistenz des getrimmten Mittels betrachtet, aber es liegen im Gegensatz dazu zensierte Daten vor.
Nun werden rechts zensierte Daten des Typ I betrachtet, die getrimmt werden. Hierbei können zwei Fälle auftreten. Entweder werden die zensierten Daten gar nicht beachtet, da sie herausgetrimmt werden (Lemma 3.2.1), oder die Trimmung beginnt erst innerhalb der Zensierung (Lemma 3.2.2). Der erste Fall tritt für $F_\theta^{-1}(1-\beta) \leq C$ ein. Da hier der getrimmte Mittelwert, trotz vorliegender Zensur, nur auf den unzensierten Daten beruht, kann Lemma 3.1.3 direkt angewendet werden. Im zweiten Fall wird $F_\theta^{-1}(1-\beta) > C$ untersucht. Bei ihm werden die zensierten Daten nicht durch die Trimmung entfernt.

Lemma 3.2.1 *Sei F_θ die Verteilungsfunktion zur Exponentialverteilung. Die Schätzung des getrimmten Mittelwertes sei durch das Funktional*

$$T_\beta(F) = \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yF(dy)$$

mit der Trimmung

$$F_\theta^{-1}(1-\beta) \leq C$$

gegeben. Dann gilt

$$T_\beta(F_\theta) = \frac{1-\beta+\beta\ln(\beta)}{1-\beta}\theta$$

und somit ist

$$\hat{\theta}(y) = \frac{1-\beta}{1-\beta+\beta\ln(\beta)}T_\beta(F_{N,y})$$

konsistenter Schätzer für θ.

Beweis: Sei $F_{\theta,C}$ die Verteilungsfunktion für die zensierten Daten.
Da $F_\theta^{-1}(1-\beta) \leq C$ gilt, werden alle zensierten Daten weggetrimmt. Also liegt allen Daten die unzensierte Verteilungsfunktion F_θ zugrunde.

$$\begin{aligned} T(F_{\theta,C}) &= \frac{1}{1-\beta}\int 1_{[0,F_{\theta,C}^{-1}(1-\beta)]}(y)yF_{\theta,C}(dy) \\ &= \frac{1}{1-\beta}\int 1_{[0,F_\theta^{-1}(1-\beta)]}(y)yF_\theta(dy) \\ &= T_\beta(F_\theta). \end{aligned}$$

Ab hier analog zu Beweis von Lemma 3.1.3. □

Lemma 3.2.2 *Sei F_θ die Verteilungsfunktion zur Exponentialverteilung, die Verteilungsfunktion*

$$F_{\theta,C}(x) = (1 - e^{-\frac{x}{\theta}})1_{[0,C)}(x) + 1_{[C,\infty)}(x)$$

gegeben, und die Trimmung erfüllt

$$F_\theta^{-1}(1-\beta) > C.$$

Wenn das getrimmmte Mittel durch das Funktional

$$T(F) = \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} yF(dy) = \frac{1}{1-\beta}\int 1_{[0,F^{-1}(1-\beta)]}(y)yF(dy)$$

gegeben ist, gilt für

$$T_\beta(F_{\theta,C}) = \frac{1}{1-\beta}[\theta - \theta e^{-\frac{C}{\theta}}] = \frac{1-e^{-\frac{C}{\theta}}}{1-\beta}\theta.$$

Sollte das T_β durch

$$T(F) = \frac{1}{1-\beta}\int 1_{[0,F^{-1}(1-\beta))}(y)yF(dy)$$

definiert sein, dann gilt

$$T(F_{\theta,C}) = \frac{1}{1-\beta}\left[\theta - \theta e^{-\frac{C}{\theta}} - Ce^{-\frac{C}{\theta}}\right].$$

Beweis:

$$\begin{aligned}
& F_\theta^{-1}(1-\beta) > C \\
\Rightarrow \quad & F_\theta(C) < 1-\beta \\
\Rightarrow \quad & \lim_{x\uparrow C} F_{\theta,C}(x) < 1-\beta, F_{\theta,C}(C) = 1 \\
\Rightarrow \quad & F_{\theta,C}^{-1}(1-\beta) = C.
\end{aligned}$$

Somit folgt für den 1.Fall

$$\begin{aligned}
T(F_{\theta,C}) &= \frac{1}{1-\beta}\int 1_{[0,F^{-1}_{\theta,C}(1-\beta)]}(y)yF_{\theta,C}(dy) \\
&= \frac{1}{1-\beta}\int 1_{[0,C]}(y)yP_{\theta,C}(dy) \\
&= \frac{1}{1-\beta}\left(\int_0^C 1_{[0,C]}(y)yP_{\theta,C}dy + \int_C^\infty 1_{\{C\}}(y)yP_{\theta,C}(dy)\right).
\end{aligned}$$

Da $\int P_{\theta,C}(dy) = 1$ gelten muss, folgt

$$\begin{aligned}
\int_C^\infty P_{\theta,C} &= \int_0^\infty P_{\theta,C}(dy) - \int_0^C P_{\theta,C}(dy) \\
&= 1 - \int_0^C \frac{1}{\theta}e^{-\frac{y}{\theta}}dy \\
&= 1 - \left(1 - e^{-\frac{C}{\theta}}\right) \\
&= e^{-\frac{C}{\theta}}.
\end{aligned}$$

Da im zensierten Fall y immer den Wert C annimmt, und $\int_C^\infty 1_C(y)yP_{\theta,C}(dy)$ wegen der Definition von $F_{\theta,C}$ nur Masse am Punkt C hat, gilt somit:

$$T(F_{\theta,C}) = \frac{1}{1-\beta}\left[\int_0^C y\frac{1}{\theta}e^{-\frac{y}{\theta}}dy + Ce^{-\frac{C}{\theta}}\right].$$

Einsetzen von Lemma 3.1.1 liefert

$$\begin{aligned} T(F_{\theta,C}) &= \frac{1}{1-\beta}\left[\theta - \theta e^{-\frac{C}{\theta}} - Ce^{-\frac{C}{\theta}} + Ce^{-\frac{C}{\theta}}\right] \\ &= \frac{1-e^{-\frac{C}{\theta}}}{1-\beta}\theta. \end{aligned}$$

Für den 2.Fall gilt

$$\begin{aligned} T(F_{\theta,C}) &= \frac{1}{1-\beta}\int 1_{[0,F^{-1}_{\theta,C}(1-\beta))}(y)yF(dy) \\ &= \frac{1}{1-\beta}\int 1_{[0,C)}(y)yP_{\theta,C}(dy) \\ &= \frac{1}{1-\beta}\int_0^c y\frac{1}{\theta}e^{-\frac{y}{\theta}}dy. \end{aligned}$$

Einsetzen von Lemma 3.1.1 liefert

$$T(F_{\theta,C}) = \frac{1}{1-\beta}\left[\theta - \theta e^{-\frac{C}{\theta}} - Ce^{-\frac{C}{\theta}}\right].$$

□

Die Untersuchung beider Schätzer ergibt, dass sie nicht Fischer konsistent sind. Auch die Möglichkeit durch multiplizieren mit einer Konstante, die unabhängig von θ ist (vergleiche Lemma3.1.5), scheidet hier offensichtlich in beiden Fällen aus.

3.3 FÜR GETRIMMTE, ZENSIERTE EXPONENTIALVERTEILUNGEN

Die folgenden Aussagen liegt die Quelle [CL09] zugrunde.
Wenn die Bedingung $F_N^{-1}(1-\beta) \leq C$ erfüllt ist, kann mit dem Mittel $\hat{\theta}_\beta$ als Schätzer von θ eine neue Verteilungsfunktion gewonnen werden. Für $\hat{\theta}_\beta$ wird der Schätzer aus Lemma 3.2.1 verwendet:

$$\hat{\theta}_\beta(F_N) = \frac{1}{1-\beta+\beta\log(\beta)} \int_0^{F_N^{-1}(1-\beta)} yF_N(dy).$$

Dieser Parameter fließt in die Verteilungsfunktion $F_{\hat{\theta}_\beta}$ ein und ersetzt die Empirische Verteilungsfunktion in Lemma 2.4.4.

Definition 3.3.1 *Sei $F_N^{-1}(1-\beta) \leq C$ und die Daten exponentialverteilt. Mit ihr ergibt sich der Pseudo Maximum-Likliehood-Schätzer*

$$\hat{\theta}_{pmle} = \frac{N\int_0^C ydF_{\hat{\theta}_\beta}(y) + [N - NF_{\hat{\theta}_\beta}(C)]C}{NF_{\hat{\theta}_\beta}(C)}.$$

Für diese Definition ist die Voraussetzung $F_\theta^{-1}(1-\beta) \leq C$ entscheidend. Ansonsten blieben nach der Trimmung noch zensierte Daten übrig, und das $\hat{\theta}_\beta$ ließe sich nicht berechnen. Was ist aber wenn die Trimmung nicht bis zur Zensurschranke reicht?
Für den Fall $F_\theta^{-1}(1-\beta) > C$ wird der schon bekannten Schätzer $\hat{\theta}_{mle}$ aus Lemma 2.4.4 verwendet. Diese Erkenntnis wird festgehalten in folgender Definition.

Definition 3.3.2 *Für exponentialverteilte Daten erhalten wir als modifiziert getrimmten Maximum-Likelihood-Schätzer*

$$\hat{\theta}_{new} = \begin{cases} \hat{\theta}_{pmle} & , \quad \textit{wenn } 1-\beta \leq F_N(C) \\ \hat{\theta}_{mle} & , \quad \textit{wenn } 1-\beta > F_N(C). \end{cases}$$

Als weitere Möglichkeit wird der Pseudo korrigierte Maximum-Likelihood-Schätzer in [CL09] gegeben. Dazu wird ein Korrekturfaktor eingeführt.
Das Integral über die Empirische Verteilungsfunktion wird noch einmal genauer betrachtet

$$\int_0^C y dF_N(y) = \int_0^{F_N^{-1}(1-\beta)} y dF_N(y) + \int_{F_N^{-1}(1-\beta)}^C y dF_N(y).$$

An dieser Stelle wird das mittlere Integral mit dem Funktional $T_\beta(F)$ aus Lemma 3.2.1 umschrieben. Somit ergibt sich der Korrekturfaktor

$$Correction(F_N, \beta, C) = (1-\beta)T_\beta(F_N) + \int_{F_N^{-1}(1-\beta)}^C y dF_N(y).$$

Dieser wird nun in den Schätzer aus Lemma 2.4.4 eingebunden.

Definition 3.3.3 *Sei F_N die Empirische Verteilungsfunktion für exponentialverteilte Daten. Dann ist der Pseudo korrigierter Maximum-Likelihood-Schätzer gegeben als*

$$\hat{\theta}_{pcmle} = \frac{N \cdot Correction(F_N, \beta, C) + [N - NF_N(C)]C}{NF_N(C)}$$

mit

$$Correction(F_N, \beta, C) = (1-\beta)T_\beta(F_N) + \int_{F_N^{-1}(1-\beta)}^C y dF_N(y)$$

KAPITEL 4

EINFLUSSFUNKTION UND ROBUSTHEIT

Die Ausführungen in diesen Kapitel beruhen auf [SH90].

4.1 ROBUSTHEIT

Mit der Einflussfunktion liegt ein Konzept vor, das Aufschluss über die Robustheit bzw. Fehlerempfindlichkeit von Schätzern gibt. Dabei misst sie den Einfluss der Beobachtung x auf den Schätzer.
Um diese Verfälschung bzw. Kontamination der Daten darstellen zu können, wird die neue kontaminierte Verteilungsfunktion eingeführt. Sie wandelt eine beliebige Verteilungsfunktion entsprechend ab.

Definition 4.1.1 *Sei F eine Verteilungsfunktion und $\varepsilon \in (0,1)$. Dann ist ihre kontaminierte Verteilungsfunktion gegeben durch*

$$F_{x,\varepsilon}(y) = (1-\varepsilon)F(y) + \varepsilon 1_{[x,\infty)}(y).$$

Die kontaminierte Verteilungsfunktion spiegelt die Idee wieder, dass die Beobachtung x mit der Wahrscheinlichkeit ε verfälscht ist und mit $(1-\varepsilon)$ unverfälscht aus der Verteilungsfunktion F stammt. Sie dient als Grundlage für die Einflussfunktion.

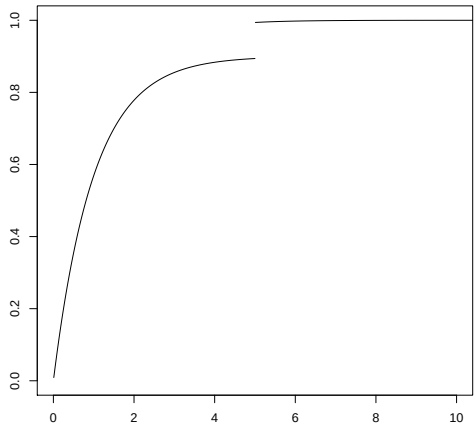


Abbildung 4.1: Kontaminierte Verteilungsfunktion der Exponentialverteilung mit $x = 0,5$ und $\varepsilon = 0,1$.

Definition 4.1.2 *Seien F und G zwei Verteilungsfunktionen und $\varepsilon \in \mathbb{R}$. Gilt*

$$\sup_y |F(y) - G(y)| < \varepsilon,$$

dann befinden sich F und G in ε-Nachbarschaft.

Um einen Schätzer beurteilen zu können, wird zuerst seine ε-Nachbarschaft von $T(F)$ und $T(F_N)$ untersucht. Natürlich sollte bei einem guten Schätzer ein hinreichend kleines ε genügen.

Definition 4.1.3 *Sei T ein Funktional, F eine Verteilungsfunktion und $T(F) \in \mathbb{R}$. Dann heißt die Abbildung $IF(T, F, x) : \mathbb{R} \to \mathbb{R}$, gegeben durch*

$$IF(T, F, x) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)F + \varepsilon 1_{[x,\infty)}) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(F_{x,\varepsilon})\Big|_{\varepsilon=0} \quad ,$$

Einflussfunktion von T auf F.

Korollar 4.1.4 *Sei F ein Funktional und $T(F)$ der zugehörige Mittelwert. Dann gilt für die zugehörige Einflussfunktion*

$$IF(T, F, x) = x - T(F).$$

Beweis: Für den Mittelwert gilt

$$\begin{aligned}
T(F_{x,\varepsilon}) &= T((1-\varepsilon)F + \varepsilon 1_{[x,\infty)}) \\
&= \int x d(1-\varepsilon)F + \int x d\varepsilon 1_{[x,\infty)} \\
&= (1-\varepsilon)\int x dF + \varepsilon \int x d1_{[x,\infty)} \\
&= (1-\varepsilon)T(F) + \varepsilon x.
\end{aligned}$$

Somit ergibt sich für eine Einflussfunktion folgendes

$$\begin{aligned}
IF(T,F,x) &= \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)F + \varepsilon 1_{[x,\infty)}) - T(F)}{\varepsilon} \\
&= \lim_{\varepsilon \downarrow 0} \frac{(1-\varepsilon)T(F) + \varepsilon x - T(F)}{\varepsilon} \\
&= \lim_{\varepsilon \downarrow 0} \frac{\varepsilon x - \varepsilon T(F)}{\varepsilon} \\
&= x - T(F).
\end{aligned}$$

□

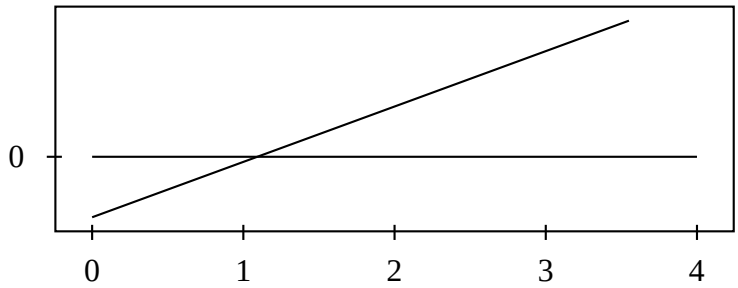


Abbildung 4.2: Einflussfunktion des normalen Mittelwertes

Der Normale Mittelwert ist nicht sehr Robust. Wie an der Einflussfunktion leicht erkannt werden kann, haben Ausreißer einen unbegrenzten Einfluss. Dies wird deutlich in Abb. 4.2, da die Funktion nach oben unbeschränkt ist.

4.2 EINFLUSSFUNKTION VON QUANTILEN

Lemma 4.2.1 *Sei F stetig und streng monoton über x. Dann gilt für das Quantil der kontaminierten Verteilungsfunktion*

$$F_{x,\varepsilon}^{-1}(q) = \begin{cases} F_{x,\varepsilon}^{-1}\left(\frac{q}{1-\varepsilon}\right), & q < (1-\varepsilon)F(x) \\ x, & (1-\varepsilon)F(x) \leq q \leq (1-\varepsilon)F(x) + \varepsilon \\ F_{x,\varepsilon}^{-1}\left(\frac{q-\varepsilon}{1-\varepsilon}\right), & (1-\varepsilon)F(x) + \varepsilon < q. \end{cases}$$

Beweis: 1.Fall: $q < (1-\varepsilon)F(x)$

$$\begin{aligned} F_{x,\varepsilon}^{-1}(q) &= \inf\{y; F_{x,\varepsilon}(y) \geq q\} \\ &= \inf\{y; (1-\varepsilon)F(y) + \varepsilon 1_{[x;\infty)}(y) \geq q\}. \end{aligned}$$

Wegen der Stetigkeit von F und der Anforderung an q, gilt

$$\begin{aligned} F_{x,\varepsilon}^{-1}(q) &= \inf\{y \in (-\infty, x); (1-\varepsilon)F(y) \geq q\} \\ &= \inf\left\{y \in (-\infty, x); F(y) \geq \frac{q}{(1-\varepsilon)}\right\} \\ &= F^{-1}\left(\frac{q}{1-\varepsilon}\right). \end{aligned}$$

2.Fall:$(1-\varepsilon)F(x) \leq q \leq (1-\varepsilon)F(x) + \varepsilon$
Da F monoton ist, gilt für alle $y < x$

$$F_{x,\varepsilon}(y) = (1-\varepsilon)F(y) < (1-\varepsilon)F(x) \leq q.$$

Desweiteren gilt nach Vorraussetzung $F_{x,\varepsilon}(x) = (1-\varepsilon)F(x) + \varepsilon \geq q$.

Somit ergibt sich für das Quantil

$$F_{x,\varepsilon}^{-1}(q) = x.$$

3.Fall:$(1-\varepsilon)F(x) + \varepsilon < q$
Für alle $y < x$ gilt

$$F_{x,\varepsilon}(y) = (1-\varepsilon)F(y) < (1-\varepsilon)F(x) < q.$$

Somit wird in diesen Fall das Quantil im Interval $[x, \infty)$ gesucht:

$$\begin{aligned} F_{x,\varepsilon}^{-1}(q) &= \inf\{y \in [x,\infty); (1-\varepsilon)F(y) + \varepsilon \geq q\} \\ &= \inf\left\{y \in [x,\infty); F(y) \geq \frac{q-\varepsilon}{1-\varepsilon}\right\} \\ &= F^{-1}\left(\frac{q-\varepsilon}{1-\varepsilon}\right). \end{aligned}$$

□

Satz 4.2.2 *Sei F streng monoton über x, f die zugehörige Dichtefunktion, $q \in (0,1)$ und $T(F) = F^{-1}(q)$ das q-Quantile von F. Dann ist*

$$IF(T,F,x) = \begin{cases} \frac{q-1}{f(F^{-1}(q))}, & x < F^{-1}(q) \\ 0, & x = F^{-1}(q) \\ \frac{q}{f(F^{-1}(q))}, & x > F^{-1}(q) \end{cases}$$

die Einflussfunktion des Quantils.

Beweis: Da F stetig und streng monoton über x ist, gilt $F^{-1}(F(x)) = x$. Außerdem gilt $F(F^{-1}(q)) = q$, wenn $F^{-1}(q) = x$ erfüllt ist.
1.Fall: $x < F^{-1}(q)$
Offensichtlich gilt $F(x) < q$. Somit existiert ein ε_0, so dass $q > F(x) + \varepsilon(1 - F(x)) = (1-\varepsilon)F(x) - \varepsilon$ für alle $0 < \varepsilon < \varepsilon_0$ erfüllt ist.

$$\begin{aligned} IF(T,F,x) &= \lim_{\varepsilon \downarrow 0} \frac{F_{x,\varepsilon}^{-1}(q) - F^{-1}(q)}{\varepsilon} \\ &\overset{Lem4.2.1}{=} \lim_{\varepsilon \downarrow 0} \frac{F_{x,\varepsilon}^{-1}\left(\frac{q-\varepsilon}{1-\varepsilon}\right) - F^{-1}(q)}{\varepsilon} \\ &= \frac{\partial}{\partial \varepsilon} F^{-1}\left(\frac{q-\varepsilon}{1-\varepsilon}\right)\Bigg|_{\varepsilon=0} \end{aligned}$$

Mit der Kettenregel ergibt sich sofort

$$IF(T,F,x) = \frac{\partial}{\partial \varepsilon} F^{-1}(y)\Big|_{y=0} \cdot \left(\frac{\partial}{\partial \varepsilon} \frac{q-\varepsilon}{1-\varepsilon}\Big|_{\varepsilon=0}\right).$$

Anwendung der Umkehrregel, mit der Tatsache das F die Dichtefunktion f besitzt, und der Quotientenregel liefert das gesuchte Ergebnis:

$$\begin{aligned} IF(T,F,x) &= \frac{1}{f(F^{-1}(q))}\left[\frac{-1(1-\varepsilon) - (-1(q-\varepsilon))}{(1-\varepsilon)^2}\right]\Bigg|_{\varepsilon=0} \\ &= \frac{1}{f(F^{-1}(q))}\left[\frac{q-1}{(1-\varepsilon)^2}\right]\Bigg|_{\varepsilon=0} \\ &= \frac{q-1}{f(F^{-1}(q))}. \end{aligned}$$

2.Fall: $x = F^{-1}(q)$
Da F stetig und streng monton über x ist, gilt $q = F(x)$.
Somit folgt sofort $q \geq (1-\varepsilon)F(x)$ für alle $\varepsilon > 0$.

Desweitern gilt

$$(1-\varepsilon)F(x)+\varepsilon=(1-\varepsilon)q+\varepsilon=q+\varepsilon(1-q)\geq q$$

für alle $\varepsilon>0$. Mit diesen Eigenschaften von q kann nun Lemma 4.2.1 angewendet werden, es ergibt sich $F_{x,\varepsilon}^{-1}(q)=x$ für alle $\varepsilon>0$.

$$IF(T,F,x)=\lim_{\varepsilon\downarrow 0}\frac{F_{x,\varepsilon}^{-1}(q)-F^{-1}(q)}{\varepsilon}=\lim_{\varepsilon\downarrow 0}\frac{x-x}{\varepsilon}=0$$

3.Fall: $x>F^{-1}(q)$
Da F stetig und streng monton über x ist, gilt $q<F(x)$. Also gibt es ein $\varepsilon_0>0$, so dass $(1-\varepsilon)F(x)>q$ für alle $0<\varepsilon<\varepsilon_0$. Mit diesen Eigenschaften von q kann nun Lemma 4.2.1 angewendet werden. So ergibt sich $F_{x,\varepsilon}^{-1}(q)=F^{-1}(\frac{q}{1-\varepsilon})$ für alle $0<\varepsilon<\varepsilon_0$:

$$\begin{aligned}IF(T,F,x)&=\lim_{\varepsilon\downarrow 0}\frac{F_{x,\varepsilon}^{-1}(q)-F^{-1}(q)}{\varepsilon}=\lim_{\varepsilon\downarrow 0}\frac{F^{-1}(\frac{q}{1-\varepsilon})-F^{-1}(q)}{\varepsilon}\\&=\frac{\partial}{\partial\varepsilon}F^{-1}(\frac{q}{1-\varepsilon})\bigg|_{\varepsilon=0}=\frac{1}{f(F^{-1}(q))}\left[\frac{\partial}{\partial\varepsilon}\frac{q}{1-\varepsilon}\bigg|_{\varepsilon=0}\right]\\&=\frac{1}{f(F^{-1}(q))}\cdot\frac{q}{(1-\varepsilon)^2}\bigg|_{\varepsilon=0}=\frac{q}{f(F^{-1}(q))}.\end{aligned}$$

□

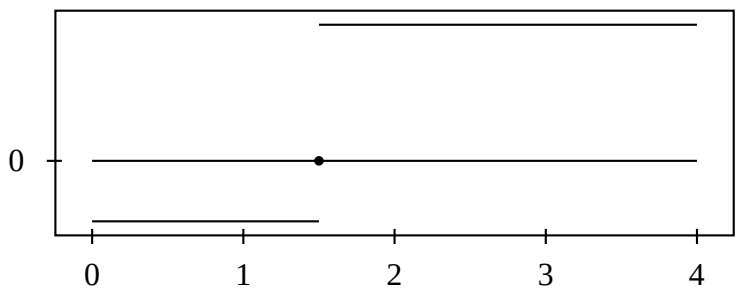


Abbildung 4.3: Einflussfunktion des q-Quantils

4.3 EINFLUSSFUNKTION DES GETRIMMTEN MITTELWERTES

Der folgende Abschnitt beruht auf den Ausarbeitungen von [SH90, S.55ff].

Satz 4.3.1 *Sei F streng monoton über x und $T_\beta(F)$ das getrimmte Mittel von F. Dann ist*

$$IF(T_\beta, F, x) = \begin{cases} \frac{x-\beta x_{1-\beta}}{1-\beta} - \mu_\beta, & 0 \leq x < x_{1-\beta} \\ \frac{x}{1-\beta} - \mu_\beta, & x = x_{1-\beta} \\ x_{1-\beta} - \mu_\beta, & x_{1-\beta} < x \end{cases}$$

die Einflussfunktion des getrimmten Mittels mit

$$x_{1-\beta} := F^{-1}(1-\beta) \quad , \quad \mu_\beta := T_\beta(F) = \int_0^{x_{1-\beta}} \frac{y}{1-\beta} dF(y).$$

Beweis:

Es wird das getrimmte Mittel der kontaminierten Verteilungsfunktion betrachtet:

$$\begin{aligned} T_\beta(F_{x,\varepsilon}) &= \int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \frac{y}{1-\beta} dF_{x,\varepsilon}(y) \\ &= \int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \frac{y}{1-\beta} dF(y) - \varepsilon \int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \frac{y}{1-\beta} dF(y) \\ &\quad + \varepsilon \int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \frac{y}{1-\beta} d1_{[x,\infty)}(y) \\ &= \int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \frac{y}{1-\beta} dF(y) + \varepsilon \int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \frac{y}{1-\beta} d(1_{[x,\infty)} - F)(y). \end{aligned}$$

Danach kommt die Ableitung zum Tragen. Auf den ersten Summanden wird die Kettenregel angewendet, auf den zweiten die Produktregel:

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}T_\beta(F_{x,\varepsilon}) = &\frac{F_{x,\varepsilon}^{-1}(1-\beta)}{1-\beta}f(F_{x,\varepsilon}^{-1}(1-\beta))\frac{\partial}{\partial\varepsilon}[F_{x,\varepsilon}^{-1}(1-\beta)]\\ &+1\int_0^{F_{x,\varepsilon}^{-1}(1-\beta)}\frac{y}{1-\beta}d(1_{[x,\infty)}-F)(y)\\ &+\varepsilon\frac{\partial}{\partial\varepsilon}\int_0^{F_{x,\varepsilon}^{-1}(1-\beta)}\frac{y}{1-\beta}d(1_{[x,\infty)}-F)(y).\end{aligned}$$

Als nächstes wird sich der Einflussfunktion gewidmet. $IF(F^{-1}(q),F,x)$ beschreibtn die Einflussfunktion des q-Quantils der kontaminierten Verteilungsfunktion aus Lemma 4.2.1:

$$\begin{aligned}IF(T_\beta,F,x) &= \frac{\partial}{\partial\varepsilon}F_{x,\varepsilon}\Big|_{\varepsilon=0}\\ &= \frac{F^{-1}(1-\beta)}{1-\beta}f(x_{1-\beta})IF(F^{-1}(q),F,x)\\ &\quad+\int_0^{F^{-1}(1-\beta)}\frac{y}{1-\beta}d(1_{[x,\infty)}-F)(y)\\ &= \frac{x_{1-\beta}}{1-\beta}f(x_{1-\beta})IF(F^{-1}(q),F,x)-\mu_\beta+\frac{x}{1-\beta}1_{(0,x_{1-\beta}]}(x).\end{aligned}$$

Einsetzen der Ergebnisse aus Lemma 4.2.1 liefert die drei Fälle. Dabei ist zu beachten $q=1-\beta$.
1.Fall: $x<x_{1-\beta}=F^{-1}(1-\beta)\Rightarrow IF(F^{-1}(q),F,x)=\frac{-\beta}{f(x_{1-\beta})}$

$$IF(T_\beta,F,x)=\frac{x_{1-\beta}}{1-\beta}f(x_{1-\beta})\frac{-\beta}{f(x_{1-\beta})}-\mu_\beta+\frac{x}{1-\beta}=\frac{x-\beta(x_{1-\beta})}{1-\beta}-\mu_\beta.$$

2.Fall: $x = x_{1-\beta} = F^{-1}(1-\beta) \Rightarrow IF(F^{-1}(q), F, x) = 0$

$$IF(T_\beta, F, x) = \frac{x_{1-\beta}}{1-\beta} f(x_{1-\beta}) \cdot 0 - \mu_\beta + \frac{x}{1-\beta} = \frac{x}{1-\beta} - \mu_\beta.$$

3.Fall: $x > x_{1-\beta} = F^{-1}(1-\beta) \Rightarrow IF(F^{-1}(q), F, x) = \frac{1-\beta}{f(x_{1-\beta})}$

$$IF(T_\beta, F, x) = \frac{x_{1-\beta}}{1-\beta} f(x_{1-\beta}) \frac{1-\beta}{f(x_{1-\beta})} - \mu_\beta + \frac{x}{1-\beta} \cdot 0 = x_{1-\beta} - \mu_\beta.$$

□

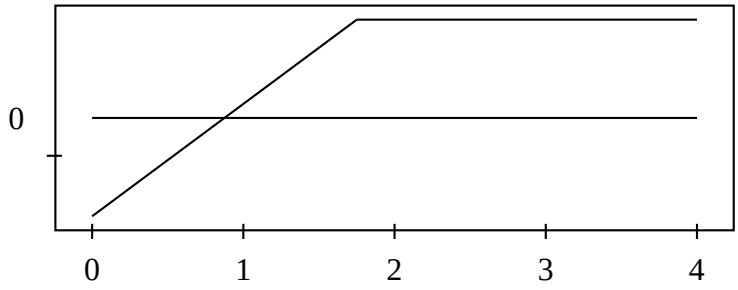


Abbildung 4.4: Einflussfunktion des getrimmten Mittelwertes

KAPITEL 5

SEGMENTIERUNG

In der Bildverarbeitung stellt die Segmentierung einen wichtigen Schritt dar. Mit ihrer Hilfe werden verschiedene Pixel des Ausgangsbildes klassifiziert. Im Fall von Rissen in schwarz-weiß Bildern bietet es sich an, zwei disjunkte Klassen zu wählen. Für diese Klassifizierung wird ein Schwellenwert benötigt, der die Graustufe angibt, ab der ein Pixel als Riss (schwarz) erkannt wird und nicht als Hintergrund (weiß). Dazu werden 3 Verfahren näher untersucht. Bevor der Schwellenwert mit Hilfe dieser Verfahren bestimmt wird, werden die Bilder mediangefiltert [ST08]. Dadurch verschwinden Schatten und Belichtungsunterschiede, oder werden zumindest abgemildert.

Außerdem stellt die Segmentierung für die spätere Lebenszeitanalyse eine wichtige Vorbearbeitung dar. Es muss darauf geachtet werden, dass die Schwellwerte richtig gewählt werden, damit die nachfolgende Berechnung und Bestimmung der Risse korrekt ist. Ansonsten kann es zu Unregelmäßigkeiten in der Lebenszeitanalyse kommen und die Maximum-Likelihood-Schätzer ergeben keine korrekten Ergebnisse.

5.1 VERFAHREN DES STEILSTEN ANSTIEGES

Die Idee des Verfahrens des steilsten Anstiegs führt darauf zurück, dass in einem Histogramm bzw. Versuchsanordnung von zweier aufeinander folgenden Punkten, mit Hilfe eines geeigneten Verfahrens, die Steigung ermittelt wird. Das ideengebende Verfahren dafür wurde von Tsai (1995) veröffentlicht in [TS95]. An der Stelle mit der größten Steigung ist nun der Zuwachs am größten und im Fall der Risserkennung kann angenommen

werden, dass dort die Schwelle von *hellen* und *dunklen* Farbpunkten ist.
Als Grundlage dient hier das Histogramm der Graustufen. Da es in der Natur der Sache liegt, ergibt sich eine diskrete Struktur. Aber um den Anstieg bestimmen zu können, wird eine stetige Verteilungsfunktion benötigt, deshalb wird die Kerndichteschätzung zu Hilfe genommen. Die Schätzfunktion ist gegeben durch:

$$f(z) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\lambda} \kappa \left(\frac{y_n - z}{\lambda} \right).$$

Dabei ist λ die Bandweite, die angibt, wie viele der umliegenden Punkte des Histogramms beim Bilden der stetigen Verteilungsfunktion an jeder Stelle mit einfließen. Wird λ zu klein gewählt, so fällt jeder Ausreißer stark ins Gewicht. Wird dagegen die Bandweite zu groß gewählt, so wirkt das Ergebnis verwässert und hat nur noch wenig mit dem Histogramm zu tun. Die Abbildung 5.1 verdeutlich diese Auswirkungen der Bandweite auf die Kerndichte Schätzung und zeigt außerdem noch eine gute Approximation an das Histogramm.

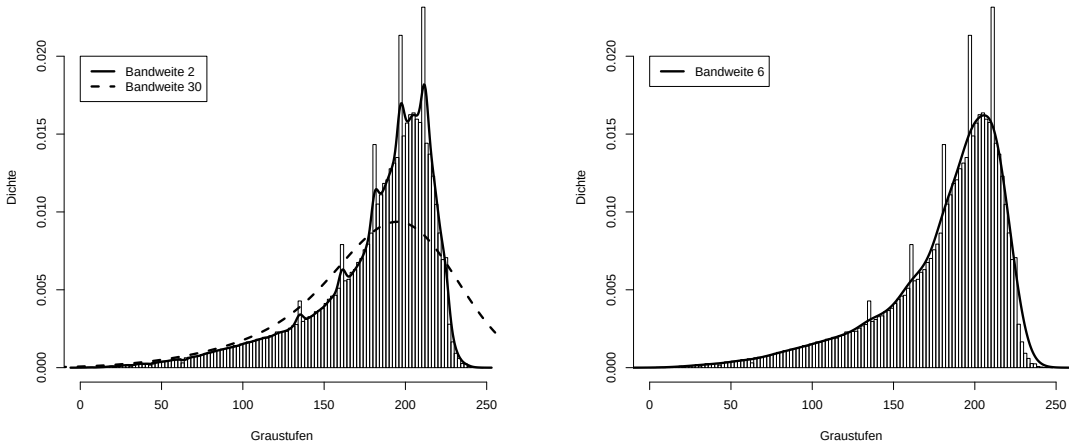


Abbildung 5.1: In das Histogramm eingetragene Kerndichteschätzungen

κ ist der Kern, in unserem Fall der Gaußkern, der wie folgt definiert ist:

$$\kappa(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}t^2}.$$

Zusammengesetzt ergibt sich

$$
\begin{aligned}
f(z) &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\lambda}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\frac{y_n-z}{\lambda})^2} \\
&= \frac{1}{N\lambda\sqrt{2\pi}}\sum_{n=1}^{N}\mathrm{e}^{-\frac{1}{2}(\frac{y_n-z}{\lambda})^2}
\end{aligned}
$$

Da es sich bei $\frac{1}{\sqrt{2\pi}}$ um eine positive Konstante handelt, also eine Streckung bzw. Stauchung in Richtung der y-Achse, kann diese entfallen, ohne die Stellen der Extrema der Funktion bzw. ihrer Ableitungen zu beeinflussen. Da nur die Steigung von Interesse ist, wird die erste Ableitung gebildet:

$$
\begin{aligned}
\frac{\partial}{\partial z}f(z) &= \frac{1}{N\lambda}\sum_{n=1}^{N}-\frac{1}{2}\left(-\frac{1}{\lambda}\right)2\left(\frac{y_n-z}{\lambda}\right)\mathrm{e}^{-\frac{1}{2}(\frac{y_n-z}{\lambda})^2} \\
&= \frac{1}{N\lambda^2}\sum_{n=1}^{N}\frac{y_n-z}{\lambda}\mathrm{e}^{-\frac{1}{2}(\frac{y_n-z}{\lambda})^2}.
\end{aligned}
$$

Bleibt nur noch der Wert z zu suchen für den die 1.Ableitung das Maximum annimmt. Dazu wird noch einmal in Erinnerung gerufen, dass mit einer diskreten Struktur begonnen wurde, und ein Schwellenwert aus eben jener abzählbaren Menge gesucht wurde. Da diese Menge überschaubar klein ist (sie umfasst die Natürlichen Zahlen von 0 bis 255), wird die gewöhnlichen Herangehensweisen bei einer stetigen Funktion vernachlässigt. Es werden nur die Werte für 0 bis 255 betrachtet, und daraus das Maximum gewählt. Der Wert, für den das Maximum angenommen wird, ist der gesuchte Schwellenwert.

Das R-Programm sieht wie folgt aus:

Listing 5.1: R Source Code Verfahren des steilsten Anstieges

```
`threshold.msi` <- function(mat,bw=30)
    {          z <- seq(0,255,1)
               N <- length(mat)
              fE <- seq(from = 0, to = 0,length.out = 256)
        for (i in 1:256)
         {
          fE[i] <- sum(((mat-z[i])/bw)*exp(-0.5*
                        (((mat-z[i])/bw)^2))/N*bw^2)
         }
        return(which.max(fE))
      }
```

Hierbei ist z nicht eine bestimmte Zahl, sondern ein Vektor, der alle Zahlen von 0 bis 255 enthält. fE ist ein Vektor der Länge 256 und beinhaltet die Ergebnisse der ersten Ableitung der Kerndichteschätzung für 0 bis 255. Mit which.max(fE) wird die Stelle des Maximums herausgesucht und somit auch den Schwellenwert.
Es sei erwähnt, dass bw für die Bandweite steht, und hier auf 30 voreingestellt ist. Die Gründe, wieso sich gerade eine Bandweite von 30 für die Risserkennung eignen folgen im nächsten Abschnitt.

5.1.1 Untersuchung der Bandweite des Verfahren des steilsten Anstieges

Der nächste wird die Bandweite untersucht und welche Auswirkung sie auf das Verfahren des steilsten Anstieges hat. Darüber hinaus wird ermittelt, welche Bandweiten für die Risserkennung am besten geeignet sind. Wie im vorherigen Anschnitt zu diesem Verfahren zu lesen ist, spielt die Bandweite eine zentrale Rolle. Die Abbildung 5.2 demonstriert dies gut:
Bandweite 2 ist zu klein gewählt und ist sehr empfindlich für Ausreißer. Bandweite 30 dagegen wirkt etwas abgeflacht im Vergleich zum Histogramm. Eine sehr gute Approximation liefert dagegen die Bandweite 6.

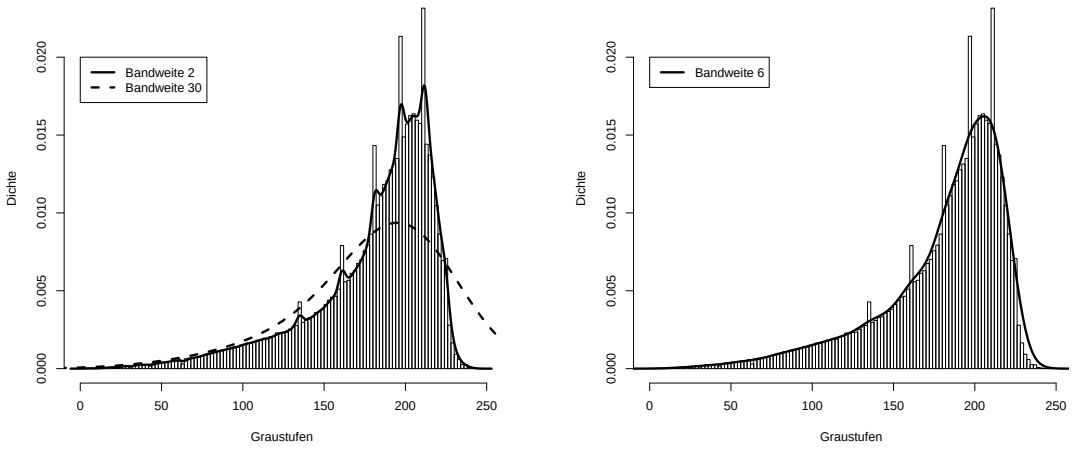


Abbildung 5.2: In das Histogramm eingetragene Kerndichteschätzungen

Intuitiv bietet sich somit eine Bandweite an, deren Kurve des Histogramms gut approximiert. Leider sind die Schwellenwerte, die so erhalten werden, nicht geeignet für die Risserkennung. Durch Untersuchung verschiedener Bandweiten an vielen Bildern der Risse, stellte sich heraus, dass sich eine Bandweite zwischen 25 und 35 empfiehlt. Dies gilt in erster Linie für mediangefilterte Bilder. Für ungefilterte Bilder passen Bandweiten, die zwischen 30 und 40 liegen. Dies wird hier aber nicht näher aufgezeigt bzw. demonstriert, da Mediangefilterte Bilder sich besser zur Risserkennung eignen.
Zur Demonstration der Auswirkung der verschiedenen Bandweiten wird der letzten Zeitpunkt S18 gewählt und das Bild mit dem markanten Zickzack-Riss betrachtet (S18_B40_X=61.950_Y=211.500).

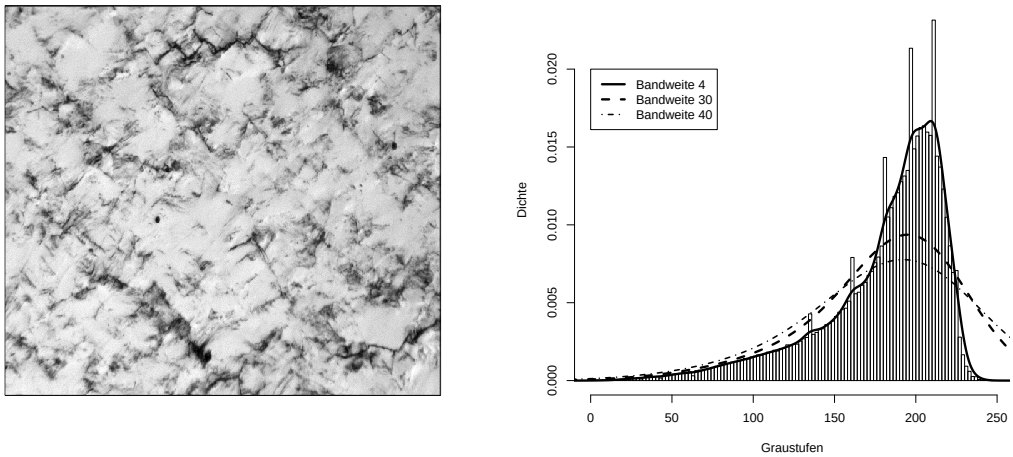


Abbildung 5.3: Bild und zugehörige, in das Histogramm eingetragene, Kerndichteschätzungen

Wird die Bandweite 4 (Abb.5.4) gewählt, so ergibt sich mit 173 einen viel zu großen Schwellenwert und die Risse verschmelzen zu einigen wenigen, viel zu langen Risse. Das Hauptaugenmerk richten sich nun auf den Zickzack-Riss am oberen Rand des Bildes. Dieser wird erst ab einer Bandweite von 25 relativ gut erkannt. Da die Bandweite zwischen 25 und 35 liegen soll, wird die Bandweite 30 betrachtet. Bei ihr wird ein Schwellenwert von 143 berechnet, der den Zickzack-Riss sehr gut erkennt. Wird hingegen eine zu große Bandweite gewählt, so wird der Schwellenwert zu klein und einige hellere Teile des Risses werden nicht mehr erkannt. Folglich zerfällt der Riss in zwei oder auch mehrere einzelne Risse. Diese Gegebenheit wird sehr gut ab einer Bandweite von 40 und einem Schwellenwert von 130 erkannt, wie es auf der nächsten Seite illustriert wird.

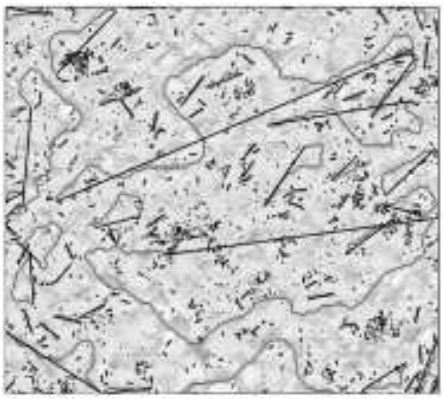
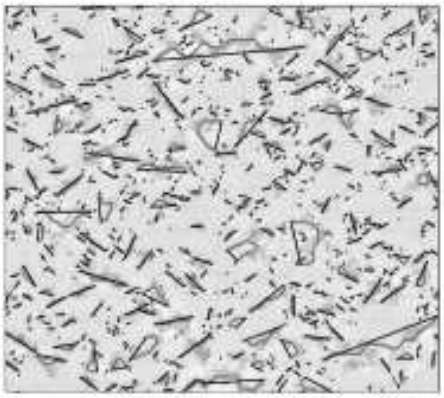


Abbildung 5.4: Risse (Bandweite 4, Schwellenwert 173)



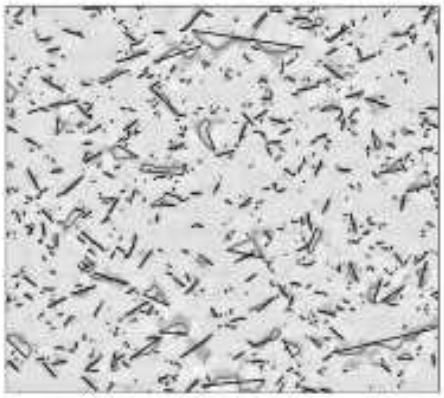


Abbildung 5.5: Risse mit Bandweite 30 bzw.40 und Schwellenwert 143 bzw. 130

Diese Untersuchung wurde für viele Bilder und Zeitpunkte getätigt. Dabei wurden verschiedene Bandweiten und daraus resultierende Schwellenwerte untersucht und miteinander verglichen. Die Ergebnisse wurden in folgender Tabelle festgehalten. Dabei ist zu beachten das mediangefilterte und ungefilterte Bilder betrachtete wurden.

Vergleich und Bewertung von Bandweiten:

Bild/ Bandweite	**1 -10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**
S18_61.950_211.500	0	0	0	0	0	0	1	1
zugehörige Schwellenwerte	bw= 8, sw=185	181	174	168	161	155	148	141
S07_61.950_211.500	0	0	0	0	0	0,5	0,5	0
zugehörige Schwellenwerte	bw= 5, sw=195	189	183	178	172	167	161	156
S00_61.950_211.500	0	0	0	0	0	0	0,5	0,5
zugehörige Schwellenwerte	bw= 7, sw=197	193	189	185	180	175	170	166
S18_61.000_212.250	0	0	0	0	0	0,5	0,5	0,5
zugehörige Schwellenwerte	bw= 9, sw=159	68	68	70	74	83	92	93
S07_61.000_212.250	0	0	0	0	0	0,5	0,5	0
zugehörige Schwellenwerte	bw= 7, sw=195	188	176	166	155	144	133	122
S00_61.000_212.250	0	0	0	0	0,5	0,5	0,5	0
zugehörige Schwellenwerte	bw= 7, sw=203	193	186	178	170	160	149	137
S18_61.950_211.500_Median	0	0	0	1	0,5	0,5	0	0
zugehörige Schwellenwerte	bw= 4, sw=173	161	155	149	143	136	130	124
S07_61.950_211.500_Median	0	0,5	0,5	1	1	0	0	0
zugehörige Schwellenwerte	bw= 2, sw=202	189	183	177	172	166	161	156
S00_61.950_211.500_Median	0	0	0	0,5	0,5	0,5	0,5	0
zugehörige Schwellenwerte	bw= 1, sw=229	217	213	208	203	198	193	188
S18_61.000_212.250_Median	0	0	0	0	1	1	0	0
zugehörige Schwellenwerte	bw= 4, sw=157	144	138	132	126	121	115	110
S07_61.000_212.250_Median	0	0	0,5	0,5	0,5	0	0	0
zugehörige Schwellenwerte	bw= 3, sw=199	186	180	175	170	164	159	154
S00_61.000_212.250_Median	0	0	0,5	0,5	0,5	0	0	0
zugehörige Schwellenwerte	bw= 2, sw=225	214	209	204	200	195	190	185
Summe (ohne Median):	0	0	0	0	0,5	2	3,5	2
Summe (nur Median):	0	0,5	1,5	3,5	4	2	0,5	0
Summe:	0	0,5	1,5	3,5	4,5	4	4	2

Legende: bw: Bandweite, 0: Schlechte Risserkennung,
0,5: Akzeptable Risserkennung, 1: Gute Risserkennung

Abschließend lässt sich das Fazit ziehen, dass für mediangefilterte Bilder eine Bandweite von 25-35 empfehlenswert ist

5.2 VERFAHREN VON RIDLER UND CALVARD

Das Verfahren geht darauf zurück, dass davon ausgegangen wird zwei große Gebiete zu haben, in denen sich die Farbwerte häufen. Das klassische Anwendungsgebiet war damals das Erkennen von Handschriften oder andere Schwarz-Weiß-Bilder, die verunreinigt waren bzw. einen schlechten Kontrast besaßen [RI78]. Dann wurde versucht diese Gebiete zu gewichten, um daraus einen Schwellenwert zu ermitteln. Das ganze Verfahren erfolgt iterativ durch folgende Vorschrift:

$$T_{k+1} = \frac{\sum_{b=0}^{T_k} b \cdot n(b)}{2\sum_{b=0}^{T_k} n(b)} + \frac{\sum_{b=T_k+1}^{N} b \cdot n(b)}{2\sum_{b=T_k+1}^{N} n(b)}.$$

Wobei T_k der Schwellenwert im k-ten Iterationsschritt darstellt. Die Graustufe wird mit b bezeichnet, und $n(b)$ die dazugehörige relative bzw. absolute Häufigkeit ist.
Als Abbruchkriterium gilt $T_k = T_{k+1}$ [TR79].
Das Schöne an diesen Verfahren ist, neben seiner einfachen Berechnung, auch die schnelle Konvergenz.

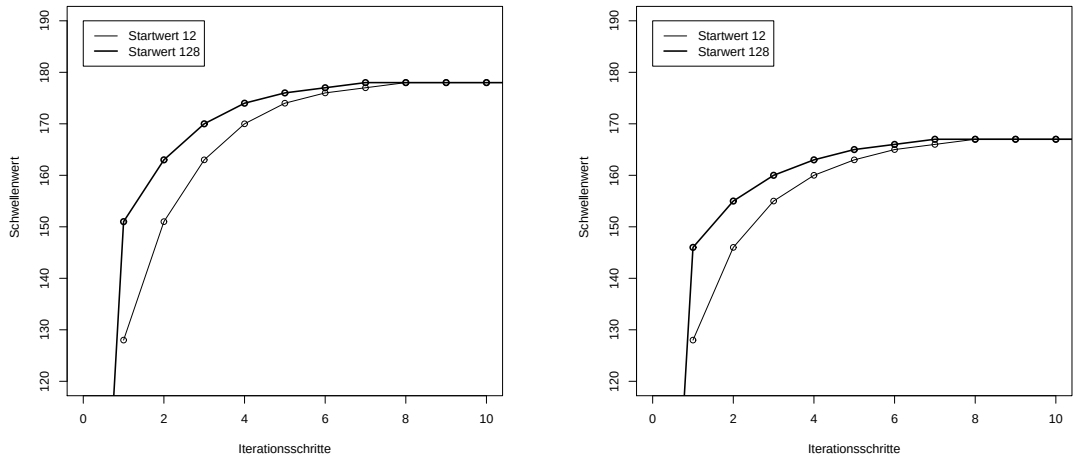


Abbildung 5.6: Ergebnisse des Verfahren von Ridler und Calvard an zwei verschiedenen Bildern mit zwei verschiedenen Startwerten

An Abbildung 5.6 wird sehr gut dargestellt, wie schnell das Verfahren konvergiert. Beim betrachten dieser Bilder, scheint das Verfahren nicht vom Startwert abzuhängen. Leider ist dies ein Trugschluss. Das Verfahren hängt von der Wahl des Startwertes ab.

Die Konvergenz lässt sich leicht mit Hilfe von zwei Lemmatas beweisen.
Dabei wird benutzt, dass in diesem Fall 256 Graustufen vorliegen. Die Beweise laufen natürlich auch mit anderen Werte als Schranken analog durch [VE80].

Lemma 5.2.1 *Seien $\mu_{1,k}$ und $\mu_{2,k}$ die Mittelwerte der ersten bzw. der zweiten Klasse, und T_k der Schwellenwert im k-ten Iterationsschritt. Wenn bei $T_0 = \lfloor(\mu_{1,0} + \mu_{2,0} \cdot \frac{1}{2})\rfloor$ gestartete wird, dann gilt $0 \leq \mu_{1,k} \leq T_k \leq \mu_{2,k} \leq 255$.*

Beweis: Durch Induktion über k.
Induktionsanfang k=0:
Offensichtlich gilt $0 \leq \mu_{1,0} \leq \mu_{2,0} \leq 255$. Da $T_0 = \lfloor(\mu_{1,0} + \mu_{2,0}) \cdot \frac{1}{2}\rfloor$ gilt, muss T_0 zwischen den zwei Mittelwerten eingebettet werden. Also $0 \leq \mu_{1,0} \leq T_0 \leq \mu_{2,0} \leq 255$.
Induktionsvoraussetzung:
Für ein beliebiges aber festes k $\in \mathbb{N}$ gilt $0 \leq \mu_{1,k} \leq T_k \leq \mu_{2,k} \leq 255$.
Induktionsschritt:
Offensichtlich gilt $\mu_{1,k+1} \in I_{1,k} = [0, T_k]$ und $\mu_{2,k+1} \in I_{2,k} = [T_k, 255]$. Folglich besteht auch hier die Beziehung $\mu_{1,k+1} \leq \mu_{2,k+1}$, und mit $T_{k+1} = \lfloor(\mu_{1,k+1}, \mu_{2,k+1}) \cdot \frac{1}{2}\rfloor$ wurde $0 \leq \mu_{1,k+1} \leq T_{k+1} \leq \mu_{2,k+1} \leq 255$ gezeigt. □

Lemma 5.2.2 *Gilt $T_k \leq T_{k+1}$ dann folgt $T_{k+1} \leq T_{k+2}$. Analog impliziert $T_k \geq T_{k+1}$ die Ungleichung $T_{k+1} \geq T_{k+2}$.*

Beweis: Sei $T_k = T_{k+1}$, dann verändern sich die Mittelwerte auch nicht. Da $\mu_{1,k+1} = \mu_{1,k+2}$ und $\mu_{2,k+1} = \mu_{2,k+2}$ gilt, folgt $T_{k+1} = T_{k+2}$.
Sei $T_k \leq T_{k+1}$ dann wird der Mittelwert $\mu_{1,k+2}$ über das alte Intervall $[0, T_k]$ und den neu hinzugekommen Werten für die erste Klasse gebildet (Repräsentiert durch das Intervall $[T_k + 1, T_{k+1}]$). Da nur größere Werte hinzugekommen, wird auch der Mittelwert angehoben, d.h. $\mu_{1,k+1} \leq \mu_{1,k+2}$. Analog folgt : $\mu_{2,k+2} \in I_{2,k+1} = I_{2,k}/[T_k + 1, T_{k+1}]$.Da hier die kleinsten Werte wegfallen, steigt der Mittelwert ebenfalls. Somit folgt: $\mu_{2,k+1} \leq \mu_{2,k+2}$. Offensichtlich folgt $T_{k+1} \leq T_{k+2}$. Analog folgt die zweite Aussage. □

Lemma 5.2.1 zeigt, dass der Schwellenwert von den Mittelwerten eingerahmt ist, und Lemma 5.2.2 zeigt, dass der Schwellenwert nur wachsen oder nur schrumpfen kann. Sollte er einmal stagnieren, so wird er immer gleich bleiben. Somit ist gezeigt, dass das Verfahren konvergiert.

Das R-Programm sieht wie folgt aus:

Listing 5.2: R Source Code Verfahren von Ridler und Calvard

```
`threshold.rc` <- function (mat,tstart=128)
  {
          vec <-as.vector(mat)
        meanl <-mean(vec[vec<tstart])
        meanu <-mean(vec[vec>=tstart])
         tnew <-(meanl+meanu)/2
    while(tstart != tnew)
      {
        tstart <-tnew
         meanl <-mean(vec[vec<tstart])
         meanu <-mean(vec[vec>=tstart])
          tnew <-round((meanl+meanu)/2)
      }
    return(tnew)
  }
```

5.3 Verfahren von Otsu

Die Grundlagen für die Untersuchung der Varianzanalyse beruht auf den Definition aus [KR03]. Das eigentlich Verfahren beruht auf [OT79].

Das Verfahren von Otsu geht auf die Varianzanalyse (ANOVA) zurück. Dabei ist zu beachten, dass bei einer Varianzanalyse die Gleichheit von Messgrößen getestet wird, und für die Schwellenwertwahl von Otsu genau das Gegenteil erwünscht ist. Es wird bei diesen Verfahren anhand einer Varianzanalyse versucht den F-Wert möglichst zu maximieren. Der F-Wert ist dabei der Quotient der Varianz zwischen den Gruppen und der Varianz innerhalb der Gruppen.

Wird nun die eigentliche Varianzanalyse in Betracht gezogen, ergibt sich, dass statt der Varianz die Summe der quadratischen Abweichungen benutzt wird. Also gilt für die Quadratische Abweichung:

Definition 5.3.1 *Sei M die Anzahl der zu untersuchenden Gruppen, N_i die Anzahl der Beobachtungen in der i-ten Gruppe und $y_{i,j}$ die j-te Beobachtung der i-ten Gruppe. Außerdem gilt $\sum_{i=1}^{M} N_i = N$. Dann ist*

$$Q_{gesamt} = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(y_{i,j} - \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} y_{i,j}\right)^2$$

die gesamte quadratische Abweichung.

Lemma 5.3.2 *Sei $\overline{y_{..}}$ das Arithmetische Mittel über alle Beobachtungen aller Gruppen. Im Fall von zwei Gruppen gilt*

$$Q_{gesamt} = \sum_{j=1}^{N_1}(y_{1,j} - \overline{y_{..}})^2 + \sum_{j=1}^{N_2}(y_{2,j} - \overline{y_{..}})^2$$

Beweis:

Sei $M = 2$

$$\begin{aligned}
Q_{gesamt} &= \sum_{i=1}^{2}\sum_{j=1}^{N_i}\left(y_{i,j} - \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{N_i} y_{i,j}\right)^2 \\
&= \sum_{j=1}^{N_1}\left(y_{1,j} - \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{N_i} y_{i,j}\right)^2 + \sum_{j=1}^{N_2}\left(y_{2,j} - \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{N_i} y_{i,j}\right)^2 \\
&= \sum_{j=1}^{N_1}\left(y_{1,j} - \frac{1}{N}\left(\sum_{j=1}^{N_1} y_{1,j} + \sum_{j=1}^{N_2} y_{2,j}\right)\right)^2 \\
&\quad + \sum_{j=1}^{N_2}\left(y_{2,j} - \frac{1}{N}\left(\sum_{j=1}^{N_1} y_{1,j} + \sum_{j=1}^{N_2} y_{2,j}\right)\right)^2 \\
&= \sum_{j=1}^{N_1}(y_{1,j} - \overline{y_{..}})^2 + \sum_{j=1}^{N_2}(y_{2,j} - \overline{y_{..}})^2 .
\end{aligned}$$

□

Wird nun die quadratische Abweichung innerhalb der Gruppen betrachtet, so ergibt sich für die i-te Gruppe:

$$Q_i = \sum_{j=1}^{N_i} \left(y_{i,j} - \frac{1}{N_i} \sum_{j=1}^{N_i} y_{i,j} \right)^2 .$$

Aufsummieren aller Q_i zeigt, dass dies ein Maß für die quadratische Abweichung innerhalb der Gruppen ist.

Definition 5.3.3 *Sei M die Anzahl der zu untersuchenden Gruppen, N_i die Anzahl der Beobachtungen in der i-ten Gruppe und $y_{i,j}$ die j-te Beobachtung der i-ten Gruppe. So ist die quadratische Abweichung innerhalb der Gruppe gegeben durch*

$$Q_{innerhalb} = \sum_{i=1}^{M} Q_i = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left(y_{i,j} - \frac{1}{N_i} \sum_{j=1}^{N_i} y_{i,j} \right)^2 .$$

Lemma 5.3.4 *Sei $\overline{y_{i.}}$ das Arithmetische Mittel über alle Beobachtungen der i-ten Gruppe. Im Fall von zwei Gruppen gilt*

$$Q_{innerhalb} = \sum_{j=1}^{N_1} (y_{1,j} - \overline{y_{1.}})^2 + \sum_{j=1}^{N_2} (y_{2,j} - \overline{y_{2.}})^2 .$$

Beweis: Sei $M = 2$

$$\begin{aligned} Q_{innerhalb} &= \sum_{i=1}^{2} \sum_{j=1}^{N_i} \left(y_{i,j} - \frac{1}{N_i} \sum_{j=1}^{N_i} y_{i,j} \right)^2 \\ &= \sum_{i=1}^{2} \sum_{j=1}^{N_i} (y_{i,j} - \overline{y_{i.}})^2 \\ &= \sum_{j=1}^{N_1} (y_{1,j} - \overline{y_{1.}})^2 + \sum_{j=1}^{N_2} (y_{2,j} - \overline{y_{2.}})^2 \end{aligned}$$

□

Schließlich wird noch die quadratischen Abweichung zwischen den Gruppen untersucht. Sie wird berechnet, indem man die quadratische Abweichung über die Gruppenmittelwerte bestimmt. Dabei wird jeder Eintrag durch $\overline{y_{i.}}$ das arithmetische Mittel der i-ten Gruppe

identifiziert. Da dieser N_i-mal auftritt, wird die entsprechend Summe durch $N_i \cdot \overline{y_{i.}}$ ersetzt. Vergleiche dazu Definition 5.3.1.

Definition 5.3.5 *Sei M die Anzahl der zu untersuchenden Gruppen,N die Anzahl aller Beobachtungen, N_i die Anzahl der Beobachtungen in der i-ten Gruppe und $y_{i,j}$ die j-te Beobachtung der i-ten Gruppe. So ist die quadratische Abweichung zwischen den Gruppe gegeben durch*

$$Q_{zwischen} = \sum_{i=1}^{M} N_i \left(\frac{1}{N_i} \sum_{j=1}^{N_i} y_{i,j} - \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{i,j} \right)^2 .$$

Lemma 5.3.6 *Sei $\overline{y_{..}}$ das Arithmetische Mittel über alle Beobachtungen und $\overline{y_{i.}}$ das Arithmetische Mittel über alle Beobachtungen der i-ten Gruppe. Im Fall von zwei Gruppen gilt*

$$Q_{zwischen} = N_1(\overline{y_{1.}} - \overline{y_{..}})^2 + N_2(\overline{y_{2.}} - \overline{y_{..}})^2.$$

Beweis: Sei $M = 2$

$$\begin{aligned}
Q_{zwischen} &= \sum_{i=1}^{2} N_i \left(\frac{1}{N_i} \sum_{j=1}^{N_i} y_{i,j} - \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{N_i} y_{i,j} \right)^2 \\
&= N_1 \left(\frac{1}{N_1} \sum_{j=1}^{N_1} y_{1,j} - \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{N_i} y_{i,j} \right)^2 + N_2 \left(\frac{1}{N_2} \sum_{j=1}^{N_2} y_{2,j} - \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{N_i} y_{i,j} \right)^2 \\
&= N_1 \left(\frac{1}{N_1} \sum_{j=1}^{N_1} y_{1,j} - \overline{y_{..}} \right)^2 + N_2 \left(\frac{1}{N_2} \sum_{j=1}^{N_2} y_{2,j} - \overline{y_{..}} \right)^2 \\
&= N_1(\overline{y_{1.}} - \overline{y_{..}})^2 + N_2(\overline{y_{2.}} - \overline{y_{..}})^2
\end{aligned}$$

□

Zur leichteren Beweisführung des nächsten Satzes wird ein kurzes Lemma gezeigt.

Lemma 5.3.7

$$\sum_{i=1}^{M} \sum_{j=1}^{N_i} \left(y_{i,j} - \frac{1}{N_i} \sum_{k=1}^{N_i} y_{i,k} \right) \left(\frac{1}{N_i} \sum_{l=1}^{N_i} y_{i,l} - \frac{1}{N} \sum_{m=1}^{M} \sum_{n=1}^{N_m} y_{m,n} \right) = 0$$

Beweis:

$$
\begin{aligned}
&\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(y_{i,j}-\frac{1}{N_i}\sum_{k=1}^{N_i}y_{i,k}\right)\left(\frac{1}{N_i}\sum_{l=1}^{N_i}y_{i,l}-\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N_m}y_{m,n}\right)\\
&=\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(\frac{1}{N_i}\sum_{l=1}^{N_i}y_{i,l}-\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N_m}y_{m,n}\right)\left(y_{i,j}-\frac{1}{N_i}\sum_{k=1}^{N_i}y_{i,k}\right)\\
&=\sum_{i=1}^{M}\left(\frac{1}{N_i}\sum_{l=1}^{N_i}y_{i,l}-\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N_m}y_{m,n}\right)\sum_{j=1}^{N_i}\left(y_{i,j}-\frac{1}{N_i}\sum_{k=1}^{N_i}y_{i,k}\right)\\
&=\sum_{i=1}^{M}\left(\frac{1}{N_i}\sum_{l=1}^{N_i}y_{i,l}-\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N_m}y_{m,n}\right)\left(\sum_{j=1}^{N_i}y_{i,j}-\sum_{j=1}^{N_i}\frac{1}{N_i}\sum_{k=1}^{N_i}y_{i,k}\right)\\
&=\sum_{i=1}^{M}\left(\frac{1}{N_i}\sum_{l=1}^{N_i}y_{i,l}-\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N_m}y_{m,n}\right)\underbrace{\left(\sum_{j=1}^{N_i}y_{i,j}-N_i\frac{1}{N_i}\sum_{k=1}^{N_i}y_{i,k}\right)}_{=0}\\
&=0
\end{aligned}
$$

□

Satz 5.3.8 *Mit den Definitionen 5.3.1, 5.3.3 und 5.3.5 gilt die Beziehung*

$$Q_{gesamt}=Q_{innerhalb}+Q_{zwischen}.$$

Beweis:

$$
\begin{aligned}
&Q_{gesamt} \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(y_{i,j} - \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} y_{i,j}\right)^2 \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(y_{i,j} - \frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j} + \frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j} - \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} y_{i,j}\right)^2 \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(y_{i,j} - \frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j}\right)^2 + 2\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(y_{i,j} - \frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j}\right)\left(\frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j} - \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} y_{i,j}\right) \\
&\qquad + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(\frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j} - \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} y_{i,j}\right)^2
\end{aligned}
$$

Wegen Lemma 5.3.7 kann der gemischte Term weggelassen werden.

$$
\begin{aligned}
&Q_{gesamt} \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(y_{i,j} - \frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j}\right)^2 + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(\frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j} - \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} y_{i,j}\right)^2 \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(y_{i,j} - \frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j}\right)^2 + \sum_{i=1}^{M} N_i\left(\frac{1}{N_i}\sum_{j=1}^{N_i} y_{i,j} - \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} y_{i,j}\right)^2 \\
&= Q_{innerhalb} + Q_{zwischen}
\end{aligned}
$$

□

Der zu bestimmende F-Wert für die Varianzanalyse ist geben durch

$$F = \frac{\text{Varianz zwischen den Gruppen}}{\text{Varianz innerhalb der Gruppe}}.$$

Dabei ist zu beachten, dass die beiden Varianzen wie folgt berechnet werden:

$$V_{zwischen} = \frac{Q_{zwischen}}{M-1}$$

und

$$V_{innerhalb} = \frac{Q_{innerhalb}}{N-M}.$$

Die Konstanten $\frac{1}{M-1}$ und $\frac{1}{N-M}$ sollen die Erwartungstreue erhalten. Somit ergibt sich der F-wert für die Varianzanalyse durch:

$$F = \frac{N-M}{M-1} \cdot \frac{Q_{zwischen}}{Q_{innerhalb}}.$$

Im Fall $M = 2$ ergibt sich:

$$\begin{aligned}
F &= \frac{N-M}{M-1} \cdot \frac{Q_{zwischen}}{Q_{innerhalb}} \\
&= \frac{N-2}{2-1} \cdot \frac{N_1 \left(\frac{1}{N_1} \sum_{j=1}^{N_1} y_{1,j} - \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{N_i} y_{i,j} \right)^2 + N_2 \left(\frac{1}{N_2} \sum_{j=1}^{N_2} y_{2,j} - \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{N_i} y_{i,j} \right)^2}{\sum_{j=1}^{N_1} \left(y_{1,j} - \frac{1}{N_1} \sum_{j=1}^{N_1} y_{1,j} \right)^2 + \sum_{j=1}^{N_2} \left(y_{2,j} - \frac{1}{N_2} \sum_{j=1}^{N_2} y_{2,j} \right)^2} \\
&= \frac{N-2}{2-1} \cdot \frac{N_1 (\overline{y_{1.}} - \overline{y_{.}})^2 + N_2 (\overline{y_{2.}} - \overline{y_{.}})^2}{\sum_{j=1}^{N_1} (y_{1,j} - \overline{y_{1.}})^2 + \sum_{j=1}^{N_2} (y_{2,j} - \overline{y_{2.}})^2}
\end{aligned}$$

für die Varianzanalyse.

Definition 5.3.9 *Bei dem Verfahren von Otsu soll der Quotient*

$$F_{Otsu} = \frac{\frac{N_1}{N} (\overline{y_{1.}} - \overline{y_{..}})^2 + \frac{N_2}{N} (\overline{y_{2.}} - \overline{y_{..}})^2}{\frac{N_1}{N(N_1-1)} \sum_{j=1}^{N_1} (y_{1,j} - \overline{y_{1.}})^2 + \frac{N_2}{N(N_2-1)} \sum_{j=1}^{N_2} (y_{2,j} - \overline{y_{2.}})^2}$$

maximal sein.

Hieran wird sehr schön deutlich, dass der F-Wert der Varianzanalyse und der Quotient von Otsu bis auf Gewichte identisch sind.

Das R-Programm sieht wie folgt aus:

Listing 5.3: R Source Code Verfahren von Otsu

```
vec<-as.vector(mat)
Q<-NULL
l<-length(vec)
for(t in 2:254){
    l.l<-length(vec[vec<t])
    l.u<-length(vec[vec>=t])
    if(l.l>1 & l.u>1){
        mean.l<-mean(vec[vec<t])
        mean.u<-mean(vec[vec>=t])
        var.l<-var(vec[vec<t])
        var.u<-var(vec[vec>=t])
        mean.all<-mean(vec)
        sigma.between<-(l.l/l)*(mean.l-mean.all)^2
                        +(l.u/l)*(mean.u-mean.all)^2
        sigma.within<-(l.l/l)*var.l + (l.u/l)*var.u
        Q<-rbind(Q,c(t,sigma.between/sigma.within))
    }
}
Q[which.max(Q[,2]),1]
}
```

5.4 Vergleich der verschiedenen Verfahren

Als Nächstes werden die 3 Verfahren und der intuitive Schwellenwert des arithmetischen Mittels untereinander verglichen. Dabei stellt sich das Verfahren des steilsten Anstieges mit einer Bandweite von 30 als Bestes heraus. Bei der Betrachtung sollten im Idealfall die Bilder immer mediangefiltert vorliegen. Es mag zwar vereinzelt Bilder geben, die durch ein anderes Verfahren besser erkannt werden, aber die große Masse wird am zuverlässigsten mit diesen Verfahren bearbeitet. Sollte einmal ein passenderer Schwellenwert durch z.b.: Ridler und Calvard gegeben sein, so liegt das Verfahren des steilsten Anstieges selten

weit weg. Somit empfiehlt sich das Verfahren des steilsten Anstieges als Segmentierung.

Zur Veranschaulichung wird das Bild S06_61.475_212.250 (Abb.5.7) betrachtet. Dem ersten Verfahren, dass untersucht wurde, liegt die naive Vorstellung des mittleren Grauwertes zu Grunde. Es wird das Arithmetische Mittel über die Graustufen aller Pixel gebildet. Als nächstes wird das Verfahren von Ridler und Calvard gefolgt von Otsu betrachtet. Als letztes wird das Verfahren des steilsten Anstieges gewählt. Hierbei wird eine Bandweite von 30 gesetzt. Die genauen Gründe, wieso diese Bandweite gewählt wird, wurden in Abschnitt 5.1.1 aufgeführt.

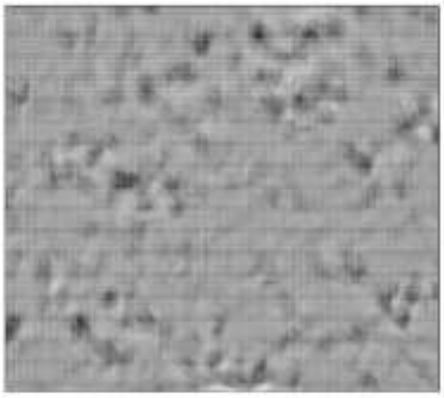
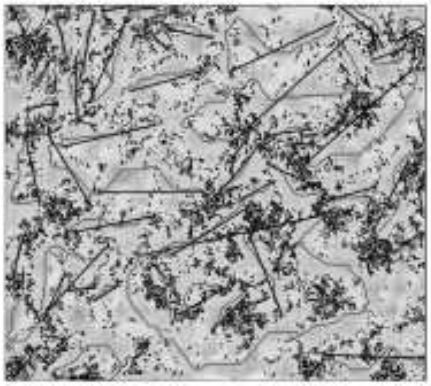
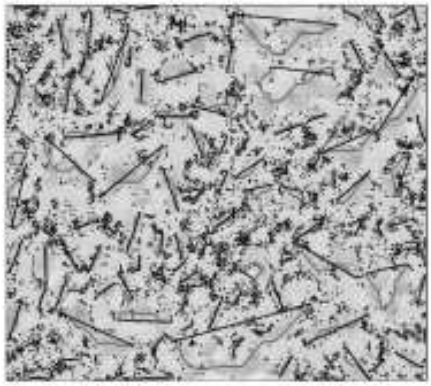


Abbildung 5.7: Originalbild

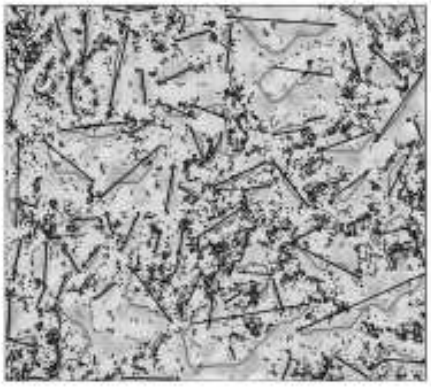
Für diese 4 Verfahren und ihre zugehörigen Schwellenwerte sind die Risse eingezeichnet. Es fällt sofort ins Auge, dass die ersten drei Verfahren viel zu hohe Werte liefern. Nur das Verfahren des steilsten Anstieges erkennt die Risse gut.



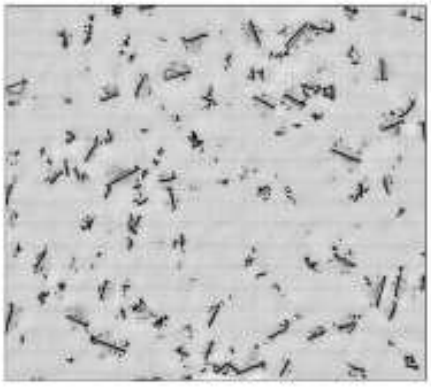
Mittlerer Grauwert t=135



Verfahren von Ridler und Calvard t=128

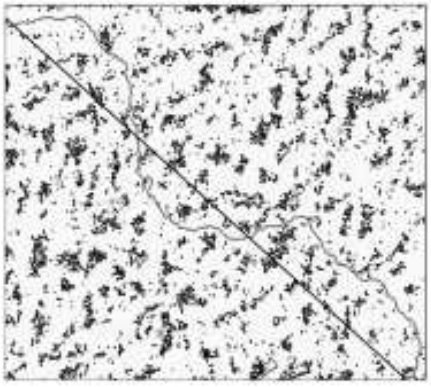


Verfahren von Otsu t=130

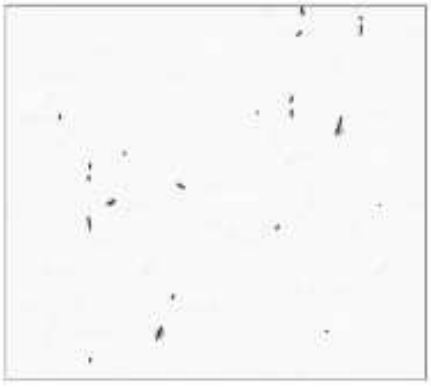


Verfahren des steilsten Anstieges bw=30, t=105

Auch für frühe Zeitpunkte arbeitet das Verfahren des steilsten Anstieges zuverlässig. Es sei aber erwähnt, dass hier die zwei anderen Verfahren (das von Ridler und Calvard sowie Otsu) ebenfalls brauchbare Ergebnisse liefern. Nur der mittlere Grauwert lieferte auch hier schlechte Schwellenwerte. Als Beispiel dient das Bild S00_61.950_211.500 .



mittlerer Grauwert t=130



Verfahren des steilsten Anstieges bw=30, t=203

Somit ergibt sich, dass über sämtliche Zeitpunkte, das Verfahren des steilsten Anstieges die erste Wahl ist. Es liefert zuverlässig über alle Bilder und Zeitpunkte brauchbare Schwellenwerte für die Segmentierung.

KAPITEL 6

ANGEWANDTE RESULTATE

6.1 DIE VERSCHIEDENEN MAXIMUM-LIKELIHOOD-SCHÄTZER

In den kommenden Kapitel werden die verschiedenen MLS miteinander verglichen. Die Daten für den Vergleich stammen aus einer Forschungsgruppe der Universität Kassel. Bis zu diesem Zeitpunkt wurden alle Bilder einer Segmentierung unterzogen und der *Dijkstra's Shortest Path Algorithm* aus [MÜ09] auf sie angewandt. Die daraus entstandenen Risslängen sind in Tabelle 1 und Tabelle 2 zu entnehmen und werden im Folgenden für die verschiedenen MLS verwendet. Dabei sind horizontal die verschiedenen relevanten Bilder aufgelistet und vertikal die Zeitpunkte hinterlegt. Die Werte in der Tabelle geben die Lebenszeit an, wobei eine reelle Zahl die erwartete Lebenszeit ist und ein C für zensierte Daten steht. Zudem wurden die Daten unter zwei Bandweiten, 20 und 30, erhoben. Der Unterschied zwischen Tabelle 1 und Tabelle 2 liegt darin, dass bei Tabelle 1 der effektive Weg des Risses gemessen wurde und bei Tabelle 2 die Projektion des Risses auf die x-Achse.

An dieser Stelle wird nochmal auf die Entstehung der Daten eingegangen. Die oben erwähnte Forschungsgruppe beschäftigte sich mit der Bestimmung von künstlichen Lebenszeiten von graduierten Material. Dafür wurden entsprechende Proben unter dem Mikroskop fotografiert. Die dadurch ersichtliche Rissentwicklung wurde dann auf den einzelnen Proben dokumentiert und ausgewertet. Mit Hilfe dieser Daten und aus der Praxis bekannten Werten, wann eine vorhandene Probe anhand eines Risses bricht, wurde eine künstliche Lebenszeit für jede Probe berechnet. Diese Daten sind in Tabelle 1 und Tabelle

2 zu finden.

Künstliche Lebenszeiten werden vor allem in Bereichen benötigt, in denen das Wissen über eine ungefähre Lebensdauer nützlich, wichtig oder sogar überlebensnotwendig ist. Um bei der Forschungsgruppe zu bleiben und damit die Ermüdung von graduierten Material aufzugreifen, wird die Abnutzung von Eisenbahnrädern betrachtet. In diesem Fall ist es zum einen für den Betreiber von Vorteil zu wissen, wie viele Kilometer ein solches Rad im Betrieb bleiben kann, was wirtschaftliche Vorteile mit sich bringt. Zum anderen bietet die künstliche Lebenszeit eines Rades auch eine gewisse Sicherheit, indem sie Anhaltspunkte für Kontrolltermine liefert.

Bis jetzt wurden nur die Lebenszeiten auf jedem einzelnen Bild betrachtet. Da die Einzelbilder ein Gesamtbild ergeben, muss nun eine Verteilungsfunktion darüber bestimmt werden. Damit aber Ausreißer nicht so stark ins Gewicht fallen, wird eine Zensurschranke eingeführt. Die gewählte Zensur ist eine Typ-I-Rechtszensur, da mit einer festen Menge von Testobjekten gestartet und eine feste rechte Zensurschranke gewählt wurde.

Bandweite 20

Y/X	61000	61475	61950	62425	62900	63375
210750	C	C	C	79313.71	C	C
211125	C	81313.71	C	79313.71	C	81313.71
211500	C	77313.71	72313.71	79313.71	79313.71	79313.71
211875	C	79313.71	79313.71	79313.71	79313.71	81313.71
212250	C	C	72313.71	79313.71	75313.71	79313.71
212625	79313.71	71313.71	75313.71	77313.71	70313.71	C
213000	77313.71	72313.71	75313.71	79313.71	79313.71	73313.71
213375	81313.71	79313.71	77313.71	72313.71	C	C
213750	C	79313.71	C	C	C	C

Bandweite 30

Y/X	61000	61475	61950	62425	62900	63375
210750	C	C	C	C	C	C
211125	C	81313.71	C	C	C	81313.71
211500	C	79313.71	79313.71	79313.71	C	C
211875	C	79313.71	81313.71	79313.71	79313.71	C
212250	C	C	77313.71	81313.71	79313.71	81313.71
212625	C	C	81313.71	C	73313.71	C
213000	79313.71	C	81313.71	81313.71	81313.71	C
213375	C	81313.71	C	77313.71	C	C
213750	C	C	C	C	C	C

Tabelle 6.1: Künstliche Lebenszeiten des graduierten Materials: Anfangspunkt und Endpunktes des Risses > 160 Pixel

Bandweite 20

Y/X	61000	61475	61950	62425	62900	63375
210750	C	C	C	79313.71	C	C
211125	C	81313.71	C	79313.71	C	C
211500	C	77313.71	72313.71	79313.71	79313.71	C
211875	C	79313.71	79313.71	79313.71	79313.71	81313.71
212250	C	C	73313.71	79313.71	77313.71	79313.71
212625	C	73313.71	75313.71	79313.71	72313.71	C
213000	77313.71	73313.71	75313.71	79313.71	79313.71	C
213375	C	79313.71	77313.71	75313.71	C	C
213750	C	79313.71	C	C	C	C

Bandweite 30

Y/X	61000	61475	61950	62425	62900	63375
210750	C	C	C	C	C	C
211125	C	81313.71	C	C	C	81313.71
211500	C	79313.71	79313.71	79313.71	C	C
211875	C	79313.71	81313.71	81313.71	79313.71	C
212250	C	C	81313.71	81313.71	81313.71	81313.71
212625	C	C	81313.71	C	79313.71	C
213000	79313.71	C	81313.71	81313.71	81313.71	C
213375	C	81313.71	C	77313.71	C	C
213750	C	C	C	C	C	C

Tabelle 6.2: Künstliche Lebenszeiten des graduierten Materials: X-Achsen-Projektion $>$ 160 Pixel

Als erstes wird der Median als intuitiven Schätzer angeschaut. Die Ergebnisse sind für die künstlichen Lebenszeiten in Tabelle 6.3 festgehalten. Wird davon ausgegangen, dass die künstlichen Lebenszeiten exponentialverteilt mit $F_\theta(z) = \int_o^z \frac{1}{\theta} e^{-\frac{y}{\theta}} dy$ sind, und die Zensurschranke als Lebenszeit interpretiert wird, dann kann der normalen Maximum-Likelihood-Schätzer ausgerechnet werden. Als mögliche Zensurschranke kann in diesen Fall C$= 81314$ gewählt werden, da davon ausgegangen werden muss, dass sie noch mindesten einen Lastwechsel länger halten müssen als die unzensierten Beobachtungen.

Die Werte für den zensierten Maximum-Likelihood-Schätzer aus Lemma 2.4.4 wird als drittes in Tabelle 6.3 angegeben unter ML(ohne Trimmung). Diesen Schätzer kann mit Hilfe des getrimmten Mittels zu einen Pseudo Maximum-Likelihood-Schätzer [CL09] erweitert werden (vergleiche Definition 3.3.1). Dieser ist als zweite Gruppe in den Tabellen festgehalten. Als nächstes folgt der Pseudo korrigierte Maximum-Likelihood-Schätzer aus Definition 3.3.3.

	Bw 30 projiziert	Bw 30 direkt	Bw 20 projiziert	Bw 20 direkt
Median	81314.71	81313.71	80313.71	79313.71
MLE = arithmetisches Mittel	80980.99	80721.12	79350.75	78887.78
ML (ohne Trimmung) Clarke	80980.99	80721.12	79350.75	78887.78
Pseudo-MLE Clarke, beta=10^{-10}	80980.99	80721.12	79350.75	78887.78
Pseudo-MLE Clarke, beta=0.01	85789.64	85514.34	84062.60	83572.15
Pseudo-MLE Clarke, beta=0.1	120913.80	120525.80	118479.70	117788.40
Pseudo-MLE Clarke, beta=0.25	200733.00	200088.80	196692.00	195544.40
Pseudo-MLE Clarke, beta=0.5	527816.50	526122.70	252198.40	278624.50
Pseudo-MLE Clarke, beta=0.75	913846.00	2357616.00	1130130.00	562175.00
PCMLE Clarke, beta=10^{-10}	80980.99	80721.49	79351.11	78888.15
PCMLE Clarke, beta=0.01	80980.99	80721.48	79351.11	78888.15
PCMLE Clarke, beta=0.1	80980.99	80721.46	79351.10	78888.13
PCMLE Clarke, beta=0.25	80980.99	80721.39	79351.02	78888.06
PCMLE Clarke, beta=0.5	80980.99	80721.25	38926.71	43180.70
PCMLE Clarke, beta=0.75	31288.75	80721.15	38826.34	19678.09

Tabelle 6.3: Schätzer für die Künstliche Lebenszeiten des graduierten Materials

Die intuitiven Schätzer ergeben durchgängig konstante Schätzungen. Ein Vergleich mit dem Pseudo-MLE von Clarke liefert die Erkenntnis, dass für geringe Trimmung ähnliche Werte erzielt werden. Bei steigendem Anteil getrimmter Daten nimmt die Differenz zu den intuitiven Schätzern zu. Im Gegenteil dazu steht der korrigierte Pseudo-MLE (PCMLE Clarke), der selbst bei einer Trimmung von 25% kaum Abweichungen zeigt. Speziell bei einer Bandweite von 30 liegt dieses Phänomen noch ausgeprägter vor.

6.2 R IMPLEMENTIERUNG DER SCHÄTZFUNKTIONEN

Zuerst werden hier ein paar Variablen, die im Source Code immer wieder kehren, erläutert. l beinhaltet den Vektor mit den Beobachtungen, folglich gibt seine Länge auch die Anzahl der Beobachtungen an. Für den Anteil der Trimmung wird stets beta verwendet und ist als Standard auf 0,25 festgesetzt. Da die Zensurschranke C bei den graduierten Material der höchsten künstlichen Lebenszeit entspricht, ist sie auf 81314,71 voreingestellt. L ist ein boolsche Vektor, indem festgehalten wird, ob der entsprechende Eintrag kleiner gleich der Zensurschranke bzw. des Quantile $F_n^{-1}(1-\beta)$ ist. Die Länge des Vektors L

ist identisch mit $N \cdot F_N(C)$. Sollten also $\sum Y_i I(Y_i \leq C) = N \int_0^C y dF_N(y)$ berechnet werden, wird einfach nur l mit L multipliziert, und es wird einen Vektor, der alle Einträge kleiner gleich der Zensurschranke enthält, bestimmt. Wird nun noch die Summe über die einzelnen Einträge dieses neuen Vektors gebildet, so ergibt sich die gesuchte Summe.

Listing 6.1: R Source Code ML-Schätzer

```
'ml_Clarke' <-
function(l,C=81314.71)
{
L <- l <= C
(sum(boolean*vec))/sum(boolean)
}
```

Dieser Code berechnet den Maximum-Likelihood-Schätzer $\hat{\theta} = \frac{\sum_{n=1}^{N} y_n}{n_{uc}}$ aus Lemma 2.4.4.

Listing 6.2: R Source Code getrimmtes Mittel

```
'mean_trim_Clarke' <-
function(l,beta=0.25,C=81314.71)
{
  if (quantile(l,(1-beta)) <= C)
  {
     L <- l <= quantile(l,(1-beta))
    mt <- (1/((1-beta+beta*log(beta))*length(l)))*sum(L*l)
    return(mt)
  } else
    print(c("Fehler : Zu groSSes Quantile",quantile(l,(1-beta))))
}
```

Hier ist der Source Code für den Schätzer aus Lemma 3.2.1 angegebn. Die if-Abfrage dient dazu, dass die Voraussetzung $F_n^{-1}(1-\beta) \leq C$ nicht verletzt wird. Das getrimmte Mittel wird wie folgt berechnet:

$$\begin{aligned}\hat{\theta} &= \frac{1}{1-\beta+\beta\log(\beta)} \int_0^{F_n^{-1}(1-\beta)} y dF_n(y) \\ &= \frac{1}{1-\beta+\beta\log(\beta)} \cdot N^{-1} \sum Y_i I(Y_i \leq C)\end{aligned}$$

Listing 6.3: R Source Code Pseudo ML-Schätzer

```
'pml_Clarke' <-
function(l,beta=0.25,C=81314.71)
{
  theta <- mean_trim_Clarke(l,beta)
  ((length(l)*(theta-theta*exp(-C/theta)-C*exp(-C/theta)))
  +(length(l)-length(l)*(1-exp(-C/theta)))*C)
  /(length(l)*(1-exp(-C/theta)))
}
```

Hier wird der Pseudo Maximum-Likelihood-Schätzer aus Lemma 3.3.1 berechnet. Dazu wird sich dem getrimmten Mittel aus dem vorherigen Source Code bzw. Lemma 3.2.1 bedient. Mit Hilfe von Lemma 3.1.1 lässt sich folgende Umformung vornehmen:

$$\begin{aligned}
\hat{\theta}_{pmle} &= \frac{N\int_0^C y dF_{\hat{\theta}_\beta}(y) + [N - NF_{\hat{\theta}_\beta}(C)](C)}{NF_{\hat{\theta}_\beta}(C)} \\
&= \frac{N\int_0^C y\frac{1}{\hat{\theta}_\beta}e^{-y/\hat{\theta}_\beta}dy + [N - N(1-e^{-C/\hat{\theta}_\beta})]C}{N(1-e^{-C/\hat{\theta}_\beta})} \\
&= \frac{N[\hat{\theta}_\beta - \hat{\theta}_\beta e^{-C/\hat{\theta}_\beta} - Ce^{-C/\hat{\theta}_\beta}] + [N - N(1_e^{-C/\hat{\theta}_\beta})]C}{N(1-e^{-C/\hat{\theta}_\beta})}
\end{aligned}$$

Listing 6.4: R Source Code Korrektur für PCMLE

```
'Correction' <-
function(l,beta=0.25,C=81314.71)
{
 theta <- mean_trim_Clarke(l,beta,C)
 Fn <- quantile(l,(1-beta))
 cor <- ((1-beta+beta*log(beta))*theta)-theta*exp(-C/theta)
          +theta*exp(-Fn/theta)-C*exp(-C/theta)+Fn*exp(-Fn/theta)
 return(cor)
}
```

Da mit dem Funktional $T_\beta(F) = \frac{1}{1-\beta}\int_0^{F^{-1}(1-\beta)} ydF(y)$für das getrimmte Mittel $\frac{1-\beta}{1-\beta+\beta\log(\beta)}T_\beta(F)$ gilt, lässt sich der Korrekturfaktor wie folgt ausdrücken:

$$
\begin{aligned}
&Correction(F_N, \beta, C) = (1-\beta)T_\beta(F_N) + \int_{F_N^{-1}(1-\beta)}^{C} y dF_{\hat{\theta}_\beta}(y) \\
&= (1-\beta+\beta\log(\beta))\frac{1-\beta}{1-\beta+\beta\log(\beta)}T_\beta(F_N) + \int_0^C y dF_{\hat{\theta}_\beta}(y) - \int_0^{F_N^{-1}(1-\beta)} y dF_{\hat{\theta}_\beta}(y) \\
&= (1-\beta+\beta\log\beta)\hat{\theta}_\beta \\
&\quad + [\hat{\theta}_\beta - \hat{\theta}_\beta e^{-C/\hat{\theta}_\beta} - Ce^{-C/\hat{\theta}_\beta} - (\hat{\theta}_\beta - \hat{\theta}_\beta e^{-F_N^{-1}(1-\beta)/\hat{\theta}_\beta} - F_N^{-1}(1-\beta)e^{-F_N^{-1}(1-\beta)/\hat{\theta}_\beta})] \\
&= (1-\beta+\beta\log\beta)\hat{\theta}_\beta \\
&\quad + [-\hat{\theta}_\beta e^{-C/\hat{\theta}_\beta} + \hat{\theta}_\beta e^{-F_N^{-1}(1-\beta)/\hat{\theta}_\beta} - Ce^{-C/\hat{\theta}_\beta} + F_N^{-1}(1-\beta)e^{-F_N^{-1}(1-\beta)/\hat{\theta}_\beta}]
\end{aligned}
$$

Listing 6.5: R Source CodePseudo korrigerter ML-Schätzer

```
'pcml_Clarke' <-
function(l,beta=0.25,C=81314.71)
{
L <- l <= C
cor <- Correction(l,beta,C)
(length(l)*cor+(length(l)-length(L))*C)/(length(L))
}
```

Der Pseudo korrigierte ML-Schätzer aus Lemma 3.3.3 wird wie folgt berechnet.

$$
\hat{\theta}_{pcmle} = \frac{N \cdot Correction(F_N, \beta, C) + [N - NF_N(C)]C}{NF_N(C)}
$$

LITERATURVERZEICHNIS

[CL09] B. R. Clarke. *Robust Estimation using β-Trimmed Mean and Trimmed Likelihood Approaches when applied to Censored Data and a Life Time Distribution.* Manuskript

[CZ11] Claudia Czado, Thorsten Schmidt. *Mathematische Statistik.* Springer-Verlag, Berlin, Heidelberg, 2011

[GE07] Hans-Otto Georgii. *Stochastik: Einführung in die Wahrscheinlichkeitsrechnung und Statistik 3. Auflage.* de Gruyter, Berlin, 2007

[KL97] J.P. Klein & M.L. Moeschberger. *Survival Analysis - Techniques for Censored and Truncated Data* Springer Science+Business Media, LLC 1997

[KR03] Ulrich Krengel. *Einführung in die Wahrscheinlichkeitstheorie und Statistik.* Friedrich. Vieweg & Sohn Verlag/GWV Fachverlag GMBH, Wiebaden 2003

[MÜ09] C.H. Müller, C. Gunkel, A. Stepper, A.C. Müller. *Micro Crack Detection with Dijkstra's Shortest Path Algorithm.* Manuskript, Dortmund , 2009

[OT79] Nobuyuki Otsu. *A Threshold Selection Method from Gray-Level Histogramms.* IEEE Transaction on Systems, Man, and Cybernetics, vol. SMC-9, no.1, pp. 62-66, January 1979.

[RI78] T. W. Ridler, S. Calvard. *Picture Thresholding Using an Iterative Selection Methode.* IEEE Transaction on Systems, Man, and Cybernetics, vol. SMC-8, no.8, pp. 630-632, August 1978.

[RI08] Horst Rinne. *Taschenbuch der Statistik.* Wissenschaftlicher Verlag Harri Deutsch GMBH, Frankfurt am Main 2008.

[SC03] Rainer Schlittgen. *Einführung in die Statistik 10. Auflage.* Oldenburger Wissenschaftsverlag GmbH, München, 2003

[SH90] Robert G. Staudte & Simon J.Sheater. *Robust Estimation and Testing*. John Wiley & Sons, Inc., New York [u.a.] 1990.

[ST08] A. Stepper. *Lebenszeit-Modelle zur Analyse der Ermüdung von gradierten Materialien*. Diplomarbeit, Kassel, 2008.

[TR79] H. J. Trussell. *Comments on "Picture Thresholding Using an Iterative Selection Method "*. IEEE Transaction on Systems, Man, and Cybernetics, vol. SMC-9, no.5, p. 311, May 1979.

[TS95] Tsai, D.M.: *A fast thresholding selection procedure for multimodal and unimodal histogramms*. Pattern Recogn. Lett. 16, 653-666, Juni 1995

[VE80] F. R. Dias Velasco. *Thresholding using the ISODATA Clustering Algorithm* IEEE Transaction on Systems, Man, and Cybernetics, vol. SMC-10, no.11, pp. 771-774, November 1980.