Universität Dortmund
Fachbereich Statistik

Diplomarbeit

Der Kaigh–Lachenbruch–Quantilschätzer im unzensierten und im zensierten Fall

vorgelegt von
Viviane Grunert

angefertigt unter Anleitung von
Prof. Dr. Ursula Gather

Dortmund, im Oktober 1993

Inhaltsverzeichnis

Symbolverzeichnis

$\mathbb{N}$	Menge der natürlichen Zahlen.
$\mathbb{N}_0$	Menge der natürlichen Zahlen, einschließlich der Null.
$\mathbb{R}$	Menge der reellen Zahlen.
$\mathbb{R}^+$	Menge der positiven reellen Zahlen.
$(\Omega, \mathcal{A}, \mathcal{P})$	Wahrscheinlichkeitsraum.
E(.)	Erwartungswert.
Var(.)	Varianz.
Cov(.,.)	Kovarianz.
MSE(.)	mittlerer quadratischer Fehler.
eff(.,.)	Effizienz.
ξ_p	p–tes Quantil.
ε_n^*	finiter Stichprobenbruchpunkt einer Stichprobe vom Umfang n.
Bin(n,p)	Binomialverteilung mit Parametern n und p.
Beta(a,b)	Betaverteilung mit Parametern a und b.
Exp(λ)	Exponentialverteilung mit Parameter λ.
$m_{a,b}$, $M_{a,b}$	Dichte und Verteilungsfunktion der Betaverteilung mit Parametern a und b.
Hyp(n,m,M)	Hypergeometrische Verteilung mit Parametern n, m und M.
N(μ,σ^2)	Normalverteilung mit Erwartungswert μ und Varianz σ^2.
φ, Φ	Dichte und Verteilungsfunktion der Standardnormalverteilung.
u_α	α–Quantil der Standardnormalverteilung.
R$(0,1)$	Rechteckverteilung auf dem Intervall $[0,1]$.
$t_{k,\alpha}$	α–Quantil der t–Verteilung mit k Freiheitsgraden.

n	Stichprobenumfang, $n \in \mathbb{N}$.
k	Unterstichprobenumfang, $k \in \mathbb{N}$ und $k \leq n$.
k_n	von n abhängiger Unterstichprobenumfang, $k \in \mathbb{N}$ und $k \leq n$.
k_{opt}	optimaler Unterstichprobenumfang des KL–Schätzers im Sinne von minimalem mittleren quadratischen Fehler.
X	Zufallsvariable, die die interessierende Größe repräsentiert.
Y	Zufallsvariable, die die Zensierungen repräsentiert.
Z	Zufallsvariable, die die tatsächlichen Beobachtungen repräsentiert, $Z := \min\{X, Y\}$.
Δ	Bernoulli–verteilte Zufallsvariable, $\Delta := \mathbb{1}_{[0,Y]}(X)$.
H_{Hyp}	Negativ–Hypergeometrisch verteilte Zufallsvariable.
$X_1, \ldots, X_n$	Stichprobe vom Umfang n mit unabhängig und identisch gemäß X verteilten Zufallsvariablen.
$X_1^i, \ldots, X_k^i$	i–te Unterstichprobe vom Umfang k mit unabhängig und identisch gemäß X verteilten Zufallsvariablen.
$X_1^{*b}, \ldots, X_n^{*b}$	b–te Bootstrap–Stichprobe vom Umfang n aus der Stichprobe $X_1, \ldots, X_n$.
$X_{1:n}, \ldots, X_{n:n}$	Ordnungsstatistiken von $X_1, \ldots, X_n$.
$(Z_{1:n}, \Delta_{1:n}), \ldots, (Z_{n:n}, \Delta_{n:n})$	nach den Zufallsvariablen Z_i geordnete Paare mit zugehörigem Δ_i, $i = 1, \ldots, n$.
$X_{r:k}^i$	r–te Ordnungsstatistik der i–ten Unterstichprobe vom Umfang k.
$\hat{X}_{r:k}$	Projektion der Ordnungsstatistik $X_{r:k}$.
f, F	Dichte–und Verteilungsfunktion der Zufallsvariablen X.
g, G	Dichte– und Verteilungsfunktion der Zufallsvariable Y.
H	Verteilungsfunktion der Zufallsvariablen Z.
$f_{(X,Y)}$	gemeinsame Dichtefunktion der Zufallsvariablen X und Y.
F^{-1}	Umkehrfunktion von F.
F_n	empirische Verteilungsfunktion in einer Stichprobe vom Umfang n.

$\hat{F}_n$	Kaplan–Meier–Schätzer in einer Stichprobe vom Umfang n.
$\tilde{F}_n^{(1)}$, $\tilde{F}_n^{(2)}$	Versionen des ACL–Schätzers in einer Stichprobe vom Umfang n (nach Csörgö, 1988).
$\tilde{F}_n$	erweiterter ACL–Schätzer in einer Stichprobe vom Umfang n (modifizierte Form des ACL–Schätzers).
Q	Quantilfunktion zu F.
Q_n	empirische Quantilfunktion in einer Stichprobe vom Umfang n.
$\hat{Q}_n$	empirische Quantilfunktion zum Kaplan–Meier–Schätzer in einer Stichprobe vom Umfang n.
$\tilde{Q}_n$	empirische Quantilfunktion zum erweiterten ACL–Schätzer in einer Stichprobe vom Umfang n.
$K_{r:k,n}$	Kaigh–Lachenbruch–Schätzer für ξ_p im unzensierten Fall mit $r = [(k+1)p] > 0$, nach der Originaldefinition von KAIGH und LACHENBRUCH (1982).
$\overset{emp}{KL}\,(p,k,n)$	Kaigh–Lachenbruch–Schätzer für ξ_p im unzensierten Fall, basierend auf der empirischen Quantilfunktion.
$\overset{Kap}{KL}\,(p,k,n)$	Kaigh–Lachenbruch–Schätzer für ξ_p basierend auf dem Kaplan–Meier–Schätzer.
$\overset{ACL}{KL}\,(p,k,n)$	Kaigh–Lachenbruch–Schätzer für ξ_p basierend auf dem erweiterten ACL–Schätzer.
C_n	zufällige Anzahl der unzensierten Beobachtungen in der Stichprobe vom Umfang n.
S_n	zufälliger Anteil der unzensierten Beobachtungen in der Stichprobe vom Umfang n.
s	erwarteter Anteil der unzensierten Beobachtungen in einer Stichprobe.
$\lambda(x)$	Hazardfunktion an der Stelle x.
Λ_F, Λ_G	kumulierte Hazardfunktionen zu F und G.
$S(.)$	Überlebensfunktion.

$Jack(T_n)$	Jackknife–Schätzer des Schätzers T_n.
$T_{(-i)}$	Schätzer basierend auf der reduzierten Stichprobe $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ zum Schätzer T_n.
T^i_{pseu}	i–ter Pseudowert zum Schätzer T_n.
$\widehat{\mathrm{Var}}_{Jack}(.)$	Jackknife–Schätzer der Varianz.
T^{*b}	b–te Bootstrap–Wiederholung des Schätzers T_n, $b = 1, \ldots, B$.
$\widehat{\mathrm{Var}}_{Boot}(.)$	Bootstrap–Schätzer der Varianz.
$\widetilde{\mathrm{Var}}_{Boot}(.)$	Approximation des Bootstrap–Schätzers der Varianz durch Monte–Carlo–Algorithmus.
U_n	U–Statistik.
h	Kern einer U–Statistik.
$B(.,.)$	Betafunktion.
$\mathbb{1}_A(x)$	Indikatorfunktion an der Stelle x.
Q', Q''	erste bzw. zweite Ableitung der Funktion Q.
$o(.)$, $O(.)$	Landau–Symbole (siehe z.B. SERFLING, 1980, S. 1).
$[x]$	$\max\{z \in \mathbb{N}_0 \mid z \leq x\}$.
$]x[$	$\max\{z \in \mathbb{N}_0 \mid z < x\}$.
lim	Limes.
lim inf	Limes–Inferior.
$F(x-)$	$\lim_{z \to 0} F(x - z)$.
$\longrightarrow 0$	Zahlenfolge konvergiert gegen 0.
f.s.	fast sicher.
$\xrightarrow{f.s.}$	fast sichere Konvergenz.
$\xrightarrow{V}$	Konvergenz in Verteilung.
$\xrightarrow{W}$	Konvergenz in Wahrscheinlichkeit.
$\lvert\{.\}\rvert$	Mächtigkeit einer Menge.
$\lvert X\rvert$	Betrag der Zufallsvariablen X.
$\sim$	gleichverteilt wie.
$\approx$	ungefähr gleich.
■	Ende eines Beweises.

Kapitel 1

Einleitung

Diese Arbeit betrachtet einen von KAIGH und LACHENBRUCH (1982) vorgestellten nichtparametrischen Quantilschätzer im unzensierten wie auch im zensierten Fall.

Es liegt eine Grundgesamtheit mit metrisch skalierten Ausprägungen des interessierenden Merkmals zugrunde. Das zu schätzende p–te Quantil, $p \in (0,1)$, gibt die Merkmalsausprägung an, welche die Grundgesamtheit in einen Anteil von mindestens $p \cdot 100\%$ unterhalb dieser Ausprägung und in einen Anteil von mindestens $(1-p) \cdot 100\%$ oberhalb dieser Ausprägung aufteilt. Eine präzisere Definition des p–ten Quantils wird in **Kapitel 2** gegeben.

Quantile dienen zur Beschreibung von Verteilungen wie z.B. der Median (0,5–Quantil) als Lagemaß und der Quartilsabstand (0,75–Quantil – 0,25–Quantil) als Streuungsmaß (HARTUNG, 1987, S. 114, 118).

Insbesondere bei vorliegenden schiefen und mehrgipfligen Verteilungen wird die Betrachtung von Quantilen gegenüber Mittelwertsberechnungen bevorzugt, weil letztere sich hier oft als ungeeignet erweisen.

Folgende Situation aus der Augenheilkunde ist dafür ein Beispiel.

Beispiel 1.1 *In einer Arbeit von* GARSD *et al. (1983) ist man an der Größe der Blutgefäßwandzellen in der Hornhaut des menschlichen Auges interessiert. Da hier keine Zellteilungen stattfinden, wachsen die Zellen, wenn Nachbarzellen absterben. Dieses hat zur Folge, daß die Verteilung der Zellgröße mit fortschreitendem Alter des Auges schief und mehrgipflig wird. Die Bestimmung der mittleren Zellgröße ist*

daher wenig geeignet für die Spezifizierung der Zellgrößen–Verteilung, und es wird die Betrachtung von Quantilen vorgezogen.

Neben der direkten Beschreibung von Verteilungen können Quantile auch für die Bestimmung von Parametern in einem Verteilungsmodell verwendet werden, wie im folgendem Beispiel aus LAMPKIN und OGAWA (1976).

Beispiel 1.2 *Von Interesse ist der Einfluß eines Insektizids auf Käfer. Daher soll die Überlebenszeit der Käfer untersucht werden, welche einer festen Dosis des Insektizids ausgesetzt sind. Die Käfer stehen solange unter Beobachtung, bis alle gestorben sind. Die Schätzung von 4 oder 5 ausgewählten Quantilen der Überlebenszeit dient dann zur Bestimmung der Parameter des zugrundeliegenden Modells, welches hier nicht erläutert wird.*

Quantile werden in weiteren Bereichen der angewandten Statistik eingesetzt, worauf hier jedoch nicht eingegangen wird.

Die aufgeführten Beispiele beschreiben Situationen im unzensierten Fall. Ebenso wäre es denkbar, daß in Beispiel 1.2 die angesetzte Untersuchungsdauer zu kurz ist, um bei allen Käfern den Tod feststellen zu können. Als Ausweg kann den noch lebenden Käfern eine zensierte Überlebenszeit zugewiesen werden, womit eine zensierte Datensituation vorläge.

Der von KAIGH und LACHENBRUCH (1982) vorgestellte Quantilschätzer wird in dieser Arbeit für den unzensierten wie auch für den zensierten Fall untersucht. Da es sich um einen nichtparametrischen Schätzer handelt, wird dabei von einer speziellen Verteilungsannahme für die Grundgesamtheit abgesehen. Die Motivation für den Kaigh–Lachenbruch–Schätzer (KL–Schätzer) ist das Ziel, einen Quantilschätzer mit verbesserter Schätzgenauigkeit gegenüber herkömmlichen Quantilschätzern zu konstruieren, der auch bei kleinen Stichprobenumfängen eine möglichst „gute“ Schätzung liefert. Die Variabilität der Schätzung soll hierbei durch Betrachtung von Unterstichproben verringert werden. Der KL–Schätzer basiert somit auf Unterstichproben vom gleichen und festen Umfang. Erst die Festlegung dieses Unterstichprobenumfangs erlaubt eine eindeutige Definition des Schätzers.

Zur Einführung in die Thematik beinhaltet das **Kapitel 2** grundlegende Definitionen und Darstellungsformen und gibt einen kurzen Überblick über bekannte nicht-

parametrische Quantilschätzer, um dem Leser eine Einordnung des KL–Schätzers zu ermöglichen.

In **Kapitel 3** wird der KL–Schätzer im unzensierten Fall betrachtet. Nach der Definition des Schätzers wird seine Einordnung in die Klasse der L– und U–Statistiken erläutert. Anschließend erfolgt eine Diskussion seiner Eigenschaften. Diese werden überwiegend mit bekannten Aussagen über U–Statistiken hergeleitet und sind teilweise aus der Literatur übernommen.

Exakte Ausdrücke für den Erwartungswert und die Varianz des KL–Schätzers werden angegeben. Diese basieren auf der unbekannten Verteilungsfunktion F und können nicht explizit berechnet werden. Für die Varianz werden daher Schätzer mit Hilfe von Jackknife–bzw. Bootstrapverfahren weitgehend neu entwickelt.

Im Anschluß erfolgt eine Diskussion zur Konsistenz des KL–Schätzers und zum Konvergenzverhalten in Verteilung. Die Ergebnisse zu gewöhnlichen U–Statistiken reichen dabei noch nicht aus, die starke Konsistenz für den KL–Schätzer zu zeigen. Gesichert ist dagegen die schon bekannte asymptotische Normalität des KL–Schätzers. Der Beweis aus der Literatur wird hierzu aufgearbeitet. Anschließend werden bekannte asymptotische Konfidenzintervalle, die aus der asymptotischen Normalität resultieren, kritisch betrachtet.

Eine weitere Eigenschaft des KL–Schätzers ergibt sich aus der Bestimmung des empirischen Bruchpunktes. Es wird deutlich, daß die Robustheit des KL–Schätzers vom zu schätzenden Quantil abhängt. Je extremer das Quantil, desto weniger robust ist der KL–Schätzer gegenüber Ausreißern. Auch werden Möglichkeiten aufgezeigt, einen möglichst optimalen Unterstichprobenumfang zu finden. Die aus der Literatur genannten Vorschläge sind jedoch meist bedingt durch fehlendes Vorwissen schlecht realisierbar.

Im letzten Abschnitt des dritten Kapitels wird eine alternative Definition des KL–Schätzers basierend auf der empirischen Quantilfunktion vorgeschlagen, welches die Übertragung des Schätzers auf den zensierten Fall vorbereitet. Zusätzlich werden für diese Definition Überlegungen für den Nachweis der starken Konsistenz ausgearbeitet, die auf Ergebnissen zur speziellen Klasse der U–Statistiken von infiniter Ordnung beruhen.

In **Kapitel 4** wird der KL–Schätzer im zensierten Fall betrachtet. Die durch LIO und PADGETT (1987) eingeführte Definition des KL–Schätzers im Modell zufälliger

Rechtszensierung wird angegeben und kurz diskutiert. Es zeigt sich, daß die Einordnung in die Klasse der L–Statistiken in diesem Modell nicht gegeben ist, wohl aber eine U–Statistik vorliegt.

Der Schwerpunkt des 4. Kapitels liegt in der neu entwickelten Definition des KL–Schätzers im sogenannten Koziol–Green–Modell, ein Spezialfall des Modells zufälliger Rechtszensierung. In diesem Modell dient der sogenannte ACL–Schätzer als Schätzer für die unbekannte Verteilungsfunktion. Der KL–Schätzer basiert hier jedoch auf einer modifizierten Form des ACL–Schätzers, da letzterer auf Grund von Unstimmigkeiten in der Literatur neu definiert werden muß. Im Koziol–Green–Modell wird für den KL–Schätzer eine Darstellung als Linearkombination von Ordnungsstatistiken mit zufälligen Koeffizienten gefunden. Damit ist aber keine Einordnung in die Klasse der L–Statistiken gelungen. Der KL–Schätzer im Koziol–Green–Modell ist eine U–Statistik von infiniter Ordnung und gehört so zu einer verallgemeinerten Klasse der gewöhnlichen U–Statistiken. Die Untersuchung der im Koziol–Green–Modell geltenden Eigenschaften des KL–Schätzers wird abschließend motiviert.

Die Arbeit endet mit **Kapitel 5**, wo Ergebnisse zusammengefaßt und offene Fragen und Vorschläge für weitere Untersuchungen angeführt werden.

Kapitel 2

Quantilschätzung im nichtparametrischen Modell

In diesem Kapitel werden in Abschnitt 2.1 elementare Definitionen und Darstellungsformen eingeführt. Abschnitt 2.2 stellt anschließend einige bekannte nichtparametrische Quantilschätzer vor, um eine Einordnung des Kaigh–Lachenbruch–Quantilschätzers zu ermöglichen.

2.1 Elementare Definitionen und Darstellungsformen

Seien $X_1, \ldots, X_n$, $n \in \mathbb{N}$, reelle, unabhängige, auf demselben Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ definierte und identisch wie eine Zufallsvariable X verteilte Zufallsvariablen mit stetiger Verteilungsfunktion F und Lebesgue–Dichte f.

Weiter sei für $p \in (0,1)$ und F das p–te Quantil als die Stelle $\xi_p \in \mathbb{R}$ bezeichnet, für die gilt:

$$P(X \leq \xi_p) \geq p \quad \text{und} \quad P(X \geq \xi_p) \geq 1 - p.$$

Eine solche Stelle existiert stets, ist aber nicht eindeutig bestimmt, wenn die stetige Verteilungsfunktion F konstante Bereiche aufweist. Um die Eindeutigkeit zu gewährleisten, wird üblicherweise die kleinste der betreffenden Stellen betrachtet (WITTING, 1985, S. 18). Dazu wird hier die Quantilfunktion Q definiert.

Definition 2.1 *Die Quantilfunktion (Pseudoinverse) Q von F ist definiert als die Abbildung $Q : [0,1] \longrightarrow \mathbb{R}$ mit*

$$Q(p) := \begin{cases} \sup\{x \in \mathbb{R} : F(x) = 0\} & \text{für} \quad p = 0 \\ \inf\{x \in \mathbb{R} : F(x) \geq p\} & \text{für} \quad 0 < p < 1 \\ \inf\{x \in \mathbb{R} : F(x) = 1\} & \text{für} \quad p = 1 \end{cases}$$

(vgl. z.B. WITTING, *1985, S. 20).*

Im weiteren sei das p–te Quantil durch $\xi_p := Q(p)$, $p \in (0,1)$, definiert und damit eindeutig bestimmt.

Ist F stetig und zusätzlich isoton, so besitzt F eine Umkehrfunktion F^{-1}, und es gilt $\xi_p = Q(p) = F^{-1}(p)$ für alle $p \in (0,1)$ (WITTING, 1985, S. 18).

Ziel ist es, das unbekannte p–te Quantil ξ_p, $p \in (0,1)$, zu schätzen. Dabei wird sich hier auf die Schätzmethoden im nichtparametrischen Modell beschränkt, d.h. es werden keine speziellen Verteilungsannahmen getroffen.

Es werden nun die empirische Verteilungsfunktion, die Ordnungsstatistiken und die empirische Quantilfunktion definiert und weitere Darstellungsformen der empirischen Quantilfunktion angefügt.

Definition 2.2 *Für die Stichprobe $X_1, \ldots, X_n$, $n \in \mathbb{N}$, ist die empirische Verteilungsfunktion F_n definiert als die Abbildung $F_n : \mathbb{R} \longrightarrow [0,1]$ mit*

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i), \quad x \in \mathbb{R}.$$

Dabei wird mit $\mathbb{1}(\cdot)$ die Indikatorfunktion bezeichnet mit

$$\mathbb{1}_A(x) := \begin{cases} 1 & \text{für} \quad x \in A \\ 0 & \text{für} \quad x \notin A, \end{cases} \qquad A \subseteq \mathbb{R}, x \in \mathbb{R},$$

(siehe z.B. SERFLING, *1980, S. 56).*

Definition 2.3 *Wird die Stichprobe $X_1, \ldots, X_n$, $n \in \mathbb{N}$, geordnet, so heißen die Zufallsvariablen*

$$X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$$

Ordnungsstatistiken dieser Stichprobe. Dabei bezeichnet $X_{j:n}$, $j \in \mathbb{N}$ *und* $1 \leq j \leq n$, *die* j*–te Ordnungsstatistik der Stichprobe vom Umfang* n, *und* $(X_{1:n}, \ldots, X_{n:n})$ *heißt der Vektor der Ordnungsstatistiken von* $X_1, \ldots, X_n$.

Weiter ist $X^i_{j:k}$ *die* j*–te Ordnungsstatistik für die* i*–te Unterstichprobe* $X^i_1, \ldots, X^i_k$, $i \in \mathbb{N}$, *die aus der Stichprobe vom Umfang* n *gezogen wird,* $k \in \mathbb{N}$ *und* $1 \leq j \leq k \leq n$ *(siehe z.B.* SERFLING, *1980, S. 87).*

Definition 2.4 *Für die Stichprobe* $X_1, \ldots, X_n$, $n \in \mathbb{N}$, *ist die empirische Quantilfunktion* Q_n *definiert als die Abbildung* $Q_n : [0,1] \longrightarrow \mathbb{R}$ *mit*

$$Q_n(p) := \begin{cases} X_{1:n} & \text{für} \quad p = 0 \\ \sum_{i=1}^{n} X_{i:n} \, \mathbb{1}_{(\frac{i-1}{n}, \frac{i}{n}]}(p) & \text{für} \quad p \in (0,1] \end{cases} \tag{2.1}$$

(siehe z.B. WITTING, *1985, S. 28).*

Bemerkung 2.5

(i) Für die j*–te Ordnungsstatistik aus einer Stichprobe* $X_1, \ldots, X_n$, $j \in \mathbb{N}$ *und* $1 \leq j \leq n$, *gilt:* $X_{j:n} = Q_n\left(\frac{j}{n}\right)$ *(*SERFLING, *1980, S. 88).*

(ii) Häufig wird in der Literatur die Bezeichnung F_n^{-1} *für die empirische Quantilfunktion benutzt. Da jedoch keine Umkehrfunktion für die empirische Verteilungsfunktion existiert, wird hier auf diese Notation verzichtet.*

Weiter ist in der Literatur auch folgende Darstellung der empirischen Quantilfunktion Q_n, $n \in \mathbb{N}$, zu finden:

$$Q_n(p) = \begin{cases} X_{1:n} & \text{für} \quad p = 0 \\ X_{np:n} & \text{für} \quad p \in (0,1] \text{ und } np \in \mathbb{N} \\ X_{[np]+1:n} & \text{für} \quad p \in (0,1] \text{ und } np \notin \mathbb{N}, \end{cases} \tag{2.2}$$

wobei $[x] := \max\{z \in \mathbb{N}_0 | z \leq x\}$ ist (siehe z.B. SERFLING, 1980, S. 88).

Mit der Bezeichnung $]x[:= \max\{z \in \mathbb{N}_0 | z < x\}$ sei hier zusätzlich folgende einfachere Darstellung der empirischen Quantilfunktion eingeführt, welche bei der Übertragung des KL–Schätzers auf den zensierten Fall von Nutzen sein wird (vgl. Kapitel 4).

$$Q_n(p) = \begin{cases} X_{1:n} & \text{für} \quad p = 0 \\ X_{]np[+1:n} & \text{für} \quad p \in (0,1]. \end{cases} \tag{2.3}$$

Der Nachweis für die Identität der Darstellungen (2.1), (2.2), (2.3) für die empirische Quantilfunktion ist im Anhang A.1 aufgeführt.

2.2 Stichprobenquantile als Quantilschätzer

Es folgt eine kurze Vorstellung verschiedener nichtparametrischer Quantilschätzer für ξ_p, $p \in (0,1)$, welche allgemein als Stichprobenquantile bezeichnet werden. Sie sind eine direkte, parameterfreie Approximation der Quantilfunktion an der Stelle p und bieten sich damit als natürliche Quantilschätzer für ξ_p an (MEISTER, 1984, S. 13). Stichprobenquantile basieren entweder auf einer Ordnungsstatistik oder auf mehreren Ordnungsstatistiken, denen entsprechende Gewichte zugewiesen werden.

Auf einer Ordnungsstatistik basierende Stichprobenquantile

Ein wichtiger Vertreter der Stichprobenquantile, welcher auf einer Ordnungsstatistik basiert, ist die empirische Quantilfunktion Q_n an der Stelle p, $p \in (0,1)$, (vgl. Darstellungen (2.1), (2.2) und (2.3)). Der Schätzer $Q_n(p)$ wird auch als empirisches Quantil bezeichnet (WITTING, 1985, S. 29). Unter milden Voraussetzungen an die Verteilungsfunktion F in der Umgebung von ξ_p ist $Q_n(p)$ ein stark konsistenter Schätzer für ξ_p (SERFLING, 1980, S. 74f). Weiterhin ist das empirische Quantil approximativ normalverteilt, wie das folgende Korollar zeigt.

Korollar 2.6 *Sei die Dichtefunktion f von F stetig in einer Umgebung von ξ_p, $p \in (0,1)$, und sei $f(\xi_p) > 0$. Dann gilt:*

$$\sqrt{n}(Q_n(p) - \xi_p) \xrightarrow{V} \mathrm{N}\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right) \quad \textit{für } n \to \infty$$

(siehe z.B. SERFLING, 1980, S. 77, Korollar B).

Von KAIGH und LACHENBRUCH (1982, S. 2218) wird ein weiterer Quantilschätzer angegeben, welcher definiert ist als

$$\hat{Q}_1(p) := X_{[(n+1)p]:n} \tag{2.4}$$

für $p \in (0,1)$ mit $[(n+1)p] > 0$, d.h. $n \geq \frac{1-p}{p}$, $n \in \mathbb{N}$. Auch dieser Schätzer ist approximativ normalverteilt. Es gilt:

$$\sqrt{n}(\hat{Q}_1(p) - \xi_p) \xrightarrow{V} \mathrm{N}\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right) \quad \text{für } n \to \infty \tag{2.5}$$

(vgl. Korollar 2.6) (MEISTER, 1984).

Auf zwei Ordnungsstatistiken basierende Stichprobenquantile

Auf zwei aufeinanderfolgenden Ordnungsstatistiken basiert der Quantilschätzer $\hat{Q}_2(p)$, der definiert ist als

$$\hat{Q}_2(p) := (1-g) \cdot X_{j:n} + g \cdot X_{j+1:n}, \tag{2.6}$$

wobei $j = [(n+1)p]$
und $g = (n+1)p - j$
$= (n+1)p - [(n+1)p], \quad n \in \mathbb{N}$ und $p \in (0,1)$

(HARRELL, DAVIS, 1982, S. 635f).

Für $p = 0.5$ ergibt sich hierbei der bekannte Stichprobenmedian

$$\hat{Q}_2(0.5) = \begin{cases} X_{j:n} & \text{für } \frac{(n+1)}{2} \in \mathbb{N} \\ \frac{1}{2}(X_{j:n} + X_{j+1:n}) & \text{für } \frac{(n+1)}{2} \notin \mathbb{N} \end{cases}$$

(HARRELL, DAVIS, 1982, S. 636).

Vom Typ $\hat{Q}_2(p)$ lassen sich weitere Quantilschätzer konstruieren, indem z.B. $j = [np]$ oder $j = [np + 0.5]$ jeweils mit entsprechendem g gewählt wird (PARRISH, 1990, S. 249). Darüberhinaus werden z.B. in PARRISH (1990, S. 249) und PARZEN (1979, S. 108f) weitere Quantilschätzer angegeben, die auf einer Ordnungsstatistik oder auf zwei Ordnungsstatistiken basieren.

Alle bis hierher aufgeführten nichtparametrischen Quantilschätzer sind Linearkombinationen von Ordnungsstatistiken und gehören somit zur Klasse der L–Schätzer. Den relevanten Ordnungsstatistiken sind entsprechend große Gewichte aus dem Intervall $(0,1]$ zugeordnet, während sich die zugehörigen Gewichtsfunktionen für alle übrigen Ordnungsstatistiken zu Null ergeben. Da die Schätzer jeweils nur auf einer Ordnungsstatistik oder auf zwei Ordnungsstatistiken basieren, können sie insbesondere bei kleinen Stichprobenumfängen eine hohe Variabilität aufweisen.

Auf mehr als zwei Ordnungsstatistiken basierende Stichprobenquantile

Die Variabilität der Quantilschätzung kann reduziert werden, indem eine glattere Gewichtsfunktion zu Grunde gelegt wird, d.h. indem mehr Ordnungsstatistiken

für die Schätzung von ξ_p echt positiv gewichtet werden. Dies ist der Fall beim Harrell–Davis–Schätzer (HARRELL, DAVIS, 1982), bei dem allen Ordnungsstatistiken $X_{1:n}, \ldots, X_{n:n}$ echt positive Gewichte zugeordnet werden. Er ist definiert durch

$$\hat{Q}_3(p) := \sum_{i=1}^{n} w_{i,n} \cdot X_{i:n}, \quad p \in (0,1),\ n \in \mathbb{N}, \tag{2.7}$$

mit Gewichten

$$w_{i,n} := \int_{\frac{i-1}{n}}^{\frac{i}{n}} m_{a,b}(x)\, dx \ > \ 0,$$

wobei

$$m_{a,b}(x) := \frac{1}{B\big((n+1)p, (n+1)(1-p)\big)} \cdot x^{(n+1)p-1} \cdot (1-x)^{(n+1)(1-p)-1}, \quad 0 < x < 1,$$

die Dichte der Betaverteilung mit Parametern $a = (n+1)p$ und $b = (n+1)(1-p)$ bezeichnet. Die Betafunktion ist dabei definiert als

$$B(x,y) = \int_0^1 t^{x-1} \cdot (1-t)^{y-1}\, dt, \qquad x > 0,\ y > 0.$$

Der Schätzer $\hat{Q}_3(p)$ für ξ_p ist der exakte Bootstrapschätzer für $\mathrm{E}_F(X_{(n+1)p:n})$, wenn $(n+1)p \in \mathbb{N}$ ist (SHEATHER, MARRON, 1990, S. 411). Er ist unter einigen Voraussetzungen an die Verteilungsfunktion F asymptotisch normalverteilt (GRONEN, 1993, S. 23ff), und es existiert ein Jackknife–Schätzer für seine Varianz (HARRELL, DAVIS, 1982, S. 636). In einer Simulationsstudie weist der Harrell–Davis–Schätzer im Vergleich zum Quantilschätzer $\hat{Q}_2(p)$ (Formel (2.6)) einen kleineren mittleren quadratischen Fehler auf, welches eine Reduktion der Variabilität des Schätzers andeutet (HARRELL, DAVIS, 1982, S. 637ff).

Zwei wichtige Methoden, die Gewichtsfunktion von Quantilschätzern zu glätten, liefern Unterstichproben–Betrachtungen und die Methode der Kernschätzung (analog zu Techniken der Dichteschätzung) (KAIGH, CHENG, 1991a, S. 540). Die aus beiden Methoden resultierenden Quantilschätzer heißen verallgemeinerte Stichprobenquantile bzw. Kern–Quantilschätzer und hängen neben n und p zusätzlich von einem sogenannten Glättungsparameter ab, so daß es sich genau genommen jeweils um Schätzerklassen handelt.

Kern–Quantilschätzer

Ein Schätzer für ξ_p, $p \in (0,1)$, aus der Klasse der Kern–Quantilschätzer geht auf PARZEN (1979, S. 113) zurück und wird in SHEATHER, MARRON (1990, S. 410) wie folgt definiert:

$$\begin{aligned} \hat{Q}_4(p) &:= \int_0^1 \frac{Q_n(p)}{h_n} \cdot K\left(\frac{p-x}{h_n}\right) dx \\ &= \sum_{i=1}^{n} \left(\frac{1}{h_n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} K\left(\frac{p-x}{h_n}\right) dx \right) \cdot X_{i:n}, \end{aligned} \tag{2.8}$$

wobei K eine stetige, um Null symmetrische Wahrscheinlichkeitsdichte ist, und $h_n \in \mathbb{R}^+$ ist mit $h_n \to 0$ für $n \to \infty$. Im allgemeinen wird K als Kernfunktion und h_n, $n \in \mathbb{N}$, als Bandbreite (Glättungsparameter) bezeichnet.

Der Schätzer $\hat{Q}_4(p)$ motiviert YANG (1985), einen ähnlichen Kern–Quantilschätzer von folgender Gestalt vorzuschlagen:

$$\hat{Q}_5(p) = \sum_{i=1}^{n} \frac{1}{n \cdot h_n} K\left(\frac{\frac{i}{n}-p}{h_n}\right) \cdot X_{i:n},$$

wobei K und h_n wie zu (2.8) definiert sind.

Beide Kern–Quantilschätzer sind unter gewissen Annahmen asymptotisch normalverteilt und im quadratischen Mittel asymptotisch äquivalent (YANG, 1985, S. 1006). Simulationsstudien in YANG (1985) lassen annehmen, daß beide Kern–Quantilschätzer im Vergleich zu $Q_n(p)$ in vielen Fällen effizienter schätzen.

Weitere Kern–Quantilschätzer sind z.B. in KAPPENMANN (1987), SHEATHER und MARRON (1990), YANG (1985) und ZELTERMAN (1990) zu finden.

Verallgemeinerte Stichprobenquantile

In der Klasse der Quantilschätzer, die durch Unterstichproben–Betrachtungen geglättet werden, wird der Unterstichprobenumfang k, $k \in \mathbb{N}$ und $1 \leq k \leq n$, als der Glättungsparameter interpretiert. Zu dieser Klasse gehört der KL–Schätzer (KAIGH, LACHENBRUCH, 1982), welcher auf Unterstichproben basiert, die ohne Zurücklegen und ohne Berücksichtigung der Anordnung gezogen werden. In den folgenden Kapiteln dieser Arbeit wird der KL–Schätzer noch ausführlich diskutiert.

Daher wird hier das Grundprinzip der verallgemeinerten Stichprobenquantile nicht detaillierter beschrieben.

Werden die Unterstichproben mit Zurücklegen und mit Berücksichtigung der Anordnung gezogen, so ergibt sich ein Quantilschätzer der Form

$$\hat{Q}_6(p) := \sum_{i=1}^{n} \left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} m_{r,k-r+1}(x)\,\mathrm{dx} \right) \cdot X_{i:n},$$

wobei $n \in \mathbb{N}$, $k \in \mathbb{N}$ mit $1 \leq k \leq n$, $r = [(k+1)p] > 0$, und $m_{r,k-r+1}$ die Dichte der Betaverteilung mit Parametern r und $k - r + 1$ ist (KAIGH, CHENG, 1991a, S. 541f). Der Schätzer $\hat{Q}_6(p)$ unterscheidet sich vom KL–Schätzer nur auf Grund der unterschiedlichen Art, die Unterstichproben zu ziehen. Für die spezielle Wahl $k = n$ und mit der Bedingung $(n+1)p \in \mathbb{N}$ ergibt sich aus $\hat{Q}_6(p)$ der Harrell–Davis–Schätzer $\hat{Q}_3(p)$ (KAIGH, 1983, S. 2430).

KAIGH und CHENG (1991a) betrachten einen weiteren Quantilschätzer auf Basis von Unterstichproben, die mit Zurücklegen und ohne Berücksichtigung der Anordnung gezogen werden. Es wird deutlich, daß sich je nach Art der Ziehung der Unterstichproben neue Quantilschätzer konstruieren lassen.

Die erwähnten, durch die Betrachtung von Unterstichproben geglätteten Quantilschätzer sind für feste Werte von $k \in \mathbb{N}$, $k \leq n$, asymptotisch normalverteilt und im quadratischen Mittel asymptotisch äquivalent (KAIGH, CHENG, 1991a, S. 547f). Simulationsstudien von KAIGH und CHENG (1991a, S. 549ff) deuten an, daß die drei genannten Quantilschätzer ξ_p oft effizienter schätzen als das Stichprobenquantil $\hat{Q}_1(p)$ in (2.4).

Die vorgestellten Methoden zur Glättung der Gewichtsfunktionen scheinen damit in der Tat eine Reduktion der Variabilität gegenüber Quantilschätzern zu bewirken, die auf wenigeren Ordnungsstatistiken basieren.

Kapitel 3

Der Kaigh–Lachenbruch–Schätzer im unzensierten Fall

Das vorliegende Kapitel behandelt den KL–Schätzer im unzensierten Fall. Nach der Definition und Klassifikation des Schätzers in 3.1 folgt eine Diskussion seiner Eigenschaften im Abschnitt 3.2. Der Abschnitt 3.3 diskutiert Möglichkeiten zur Bestimmung eines „optimalen" Unterstichprobenumfangs für den KL–Schätzer. Das Kapitel schließt mit einer alternativen Definition des Schätzers in 3.4.

3.1 Definition und Klassifikation

Im folgenden wird der KL–Schätzer definiert, und es werden Einschränkungen bezüglich seiner Definition aufgezeigt. Weiter wird die Zuordnung zu L– und U–Statistiken erläutert. Einige Abbildungen ermöglichen zusätzlich eine graphische Analyse der Gewichte des KL–Schätzers als L–Statistik.

Seien $X_1, X_2, \ldots, X_n$, $n \in \mathbb{N}$, reelle, unabhängig und identisch wie eine Zufallsvariable X verteilte Zufallsvariablen mit stetiger, unbekannter Verteilungsfunktion F und Lebesgue–Dichte f. Mit $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ seien die zugehörigen Ordnungsstatistiken von $X_1, X_2, \ldots, X_n$ bezeichnet (vgl. Definition 2.3).

Ein häufig verwendeter nichtparametrischer Schätzer für das p–te Quantil ξ_p, $p \in (0,1)$, ist das Stichprobenquantil $\hat{Q}_1(p) = X_{[(n+1)p]:n}$ aus einer Stichprobe vom Umfang n (Formel (2.4)). Da dieser Schätzer, wie schon in Kapitel 2 erwähnt, nur auf

einer Ordnungsstatistik basiert, weist er insbesondere in kleinen Stichproben eine hohe Variabilität auf.

Die Idee des von KAIGH und LACHENBRUCH (1982) vorgeschlagenen Schätzers für ξ_p ist, diese Variabilität durch Bildung des Durchschnitts von den aus Unterstichproben berechneten Stichprobenquantilen $\hat{Q}_1(p)$ zu vermindern. Dazu werden alle Unterstichproben vom festen Umfang k $(1 \leq k \leq n)$ aus einer Stichprobe vom Umfang n ohne Zurücklegen und ohne Berücksichtigung der Anordnung gezogen. Der Kaigh–Lachenbruch–Schätzer ist als Durchschnitt über die entsprechenden Stichprobenquantile aller $\binom{n}{k}$ Unterstichproben definiert.

Definition 3.1 *Für $n \in \mathbb{N}$, $p \in (0,1)$ und für ein festes $k \in \mathbb{N}$ mit $1 \leq k \leq n$ und $[(k+1)p] > 0$ ist mit*

$$K_{[(k+1)p]:k,n} = K_{[(k+1)p]:k,n}(X_1, \ldots, X_n) := \frac{1}{\binom{n}{k}} \sum_{i=1}^{\binom{n}{k}} X^i_{r:k},$$

*$r := [(k+1)p] > 0$, der Kaigh–Lachenbruch–Schätzer (KL–Schätzer) für ξ_p definiert. Dabei bezeichnet $X^i_{r:k}$, $i = 1, \ldots, \binom{n}{k}$, die r–te Ordnungsstatistik aus der i–ten Unterstichprobe vom Umfang k (*KAIGH, LACHENBRUCH, *1982, S. 2218f).*

Der KL–Schätzer wird auch als „verallgemeinertes Stichprobenquantil" bezeichnet, weil er sich bei der Wahl $k = n$ zum Stichprobenquantil $X_{[(n+1)p]:n}$ reduziert. Für $k = 1$ ergibt sich dagegen das Stichprobenmittel $\bar{X}$.

Aus der Bedingung $[(k+1)p] > 0$, $p \in (0,1)$, d.h. $k \geq \frac{1-p}{p}$, resultiert eine Einschränkung bei der Wahl vom Unterstichprobenumfang k und damit beim Stichprobenumfang n. Für die Definition des KL–Schätzers darf abhängig vom Wert p, $p \in (0,1)$, der Unterstichprobenumfang k und damit n einen bestimmten Wert nicht unterschreiten, wie Tabelle 1 zeigt.

Tabelle 1 *Einschränkungen bei der Wahl des Unterstichprobenumfangs k für verschiedene Werte von p.*

p	$[0,5;1)$	0,40	0,20	0,10	0,05	0,01
Minimal notwendiger Wert für k	1	2	4	9	19	99

Insbesondere für die Schätzung von kleinen Quantilen verlangt somit die Bedingung $k \geq \frac{1-p}{p}$ große Stichprobenumfänge. Da die sinnvolle Schätzung von extremen Quantilen im allgemeinen auch große Stichprobenumfänge erfordert, ist obige Einschränkung für praktische Anwendungen jedoch von geringer Bedeutung.

Bemerkung 3.2 *In Kapitel 3.4 wird eine alternative Definition des KL–Schätzers eingeführt, die die obige Einschränkung $k \geq \frac{1-p}{p}$ nicht erfordert.*

Bemerkung 3.3 *Der KL–Schätzer gehört für gegebene Werte von p und n einer Schätzerklasse an, deren Elemente durch die Wahl des Unterstichprobenumfangs k mit $1 \leq k \leq n$ und $[(k+1)p] > 0$ eindeutig definiert sind. Weil die „Güte" der Schätzung u.a. auch von k abhängt, bleibt das Problem einer „optimalen" Wahl von k zu lösen (Kapitel 3.3).*

Die nächsten beiden Unterabschnitte erläutern die Zuordnung des KL–Schätzers zu den L– bzw. U–Statistiken.

Der KL–Schätzer als L–Statistik

Eine L–Statistik T_n, $n \in \mathbb{N}$, wird allgemein durch eine lineare Funktion von Ordnungsstatistiken $X_{1:n}, \ldots, X_{n:n}$ repräsentiert in der Form $T_n = \sum_{j=1}^{n} l_j \cdot X_{j:n}$, wobei l_j, $j = 1, \ldots, n$, datenunabhängige Konstanten sind (SERFLING, 1980, S. 262).

Obwohl der KL–Schätzer auf der Basis von Ordnungsstatistiken der Unterstichproben $X_1^i, \ldots, X_k^i$, $i = 1, \ldots, \binom{n}{k}$, definiert ist, ist es auch möglich, diesen als gewichtete Summe von Ordnungsstatistiken der ursprünglichen Stichprobe $X_1, \ldots, X_n$ darzustellen. Der KL–Schätzer gehört somit zur Klasse der L–Schätzer.

Satz 3.4 *Für $n \in \mathbb{N}$, $p \in (0,1)$ und für alle $k \in \mathbb{N}$ mit $1 \leq k \leq n$ und $[(k+1)p] > 0$ gilt:*

$$K_{[(k+1)p]:k,n} = \sum_{j=r}^{r+n-k} \frac{\binom{j-1}{r-1}\binom{n-j}{k-r}}{\binom{n}{k}} \cdot X_{j:n},$$

$r = [(k+1)p] > 0$ *(*KAIGH, LACHENBRUCH, *1982, S. 2219).*

Beweis: Sei $X^i_{r:k}$ die r–te Ordnungsstatistik der i–ten Unterstichprobe vom Umfang k aus einer Stichprobe vom Umfang n, wobei $1 \leq r \leq k \leq n$ und $i = 1, \ldots, \binom{n}{k}$. Werden die Unterstichproben ohne Zurücklegen und ohne Berücksichtigung der Anordnung gezogen, so gilt:

$$P(X^i_{r:k} = X_{j:n}) = \frac{\binom{j-1}{r-1}\binom{n-j}{k-r}}{\binom{n}{k}}, \tag{3.1}$$

wobei $r \leq j \leq r+n-k$ und $i = 1, \ldots, \binom{n}{k}$.

Der Ausdruck (3.1) geht auf folgende Überlegung zurück. Es gibt $\binom{j-1}{r-1}$ Möglichkeiten, daß $X^i_{1:k}, \ldots, X^i_{r-1:k}$ den jeweiligen Statistiken aus der Gruppe der kleinsten $j-1$ Ordnungsstatistiken $X_{1:n}, \ldots, X_{j-1:n}$ entsprechen. Ebenso gibt es $\binom{n-j}{k-r}$ Möglichkeiten, daß $X^i_{r+1:k}, \ldots, X^i_{k:k}$ den jeweiligen Statistiken aus der Gruppe der größten $n-j$ Ordnungsstatistiken $X_{j+1:n}, \ldots, X_{n:n}$ entsprechen. Insgesamt gibt es $\binom{n}{k}$ mögliche Unterstichproben vom Umfang k aus einer Stichprobe vom Umfang n ohne Zurücklegen und ohne Berücksichtigung der Anordnung.

Es folgt, daß

$$K_{[(k+1)p]:k,n} = \frac{1}{\binom{n}{k}} \sum_{i=1}^{\binom{n}{k}} X^i_{r:k} = \sum_{j=1}^{n} P(X^i_{r:k} = X_{j:n}) \cdot X_{j:n} = \sum_{j=r}^{r+n-k} \frac{\binom{j-1}{r-1}\binom{n-j}{k-r}}{\binom{n}{k}} \cdot X_{j:n},$$

wobei $r = [(k+1)p] > 0$.

■

Analyse der Gewichte

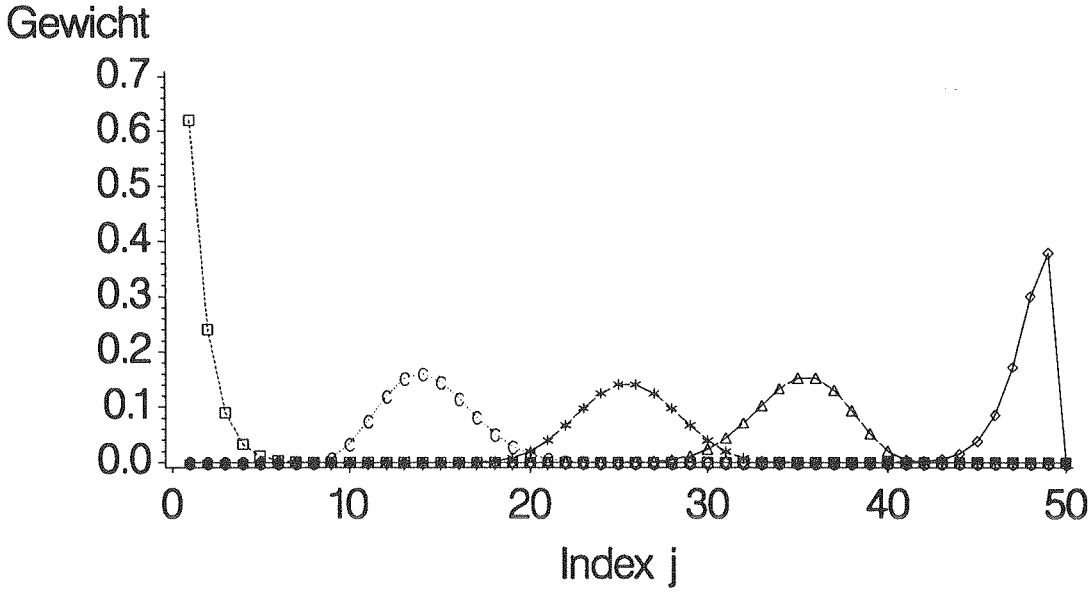
Die Gewichte des KL–Schätzers entsprechen den Ausprägungswahrscheinlichkeiten einer Negativ–Hypergeometrischen Verteilung mit Parametern r, n und k (KAIGH, LACHENBRUCH, 1982, S. 2220). Eine gemäß dieser Verteilung verteilte Zufallsvariable $H_{Hyp}(p,k,n) = H_{Hyp}$ realisiert sich mit Wahrscheinlichkeit eins in der Menge $\{r, r+1, \ldots, r+n-k\}$, $r = [(k+1)p] > 0$. Dabei ergibt sich die Mächtigkeit dieser Menge zu $n-k+1$ und ist somit unabhängig vom Wert p. Es gilt

$$\mathrm{E}(H_{Hyp}) = \frac{r \cdot (n+1)}{k+1} \quad \text{und} \quad \mathrm{Var}(H_{Hyp}) = \frac{r \cdot (n+1)(n-k)(k+1-r)}{(k+1)^2(k+2)}.$$

Abbildung 1 zeigt die Gewichte des KL–Schätzers für verschiedene Werte von p, $p \in (0,1)$, bei einem Stichprobenumfang von $n = 50$ und einem Unterstichprobenumfang

von $k = 31$. Die Gewichte für $p = 0,05$, $p = 0,5$ und $p = 0,95$ können exemplarisch der Tabelle 3 im Anhang B entnommen werden.

Abbildung 1 *Gewichte des KL–Schätzers für* $n = 50$, $k = 31$ *und verschiedene Werte von* p.



$p = 0,05$ ($\square$ – *Linie*), $p = 0,3$ ($\circ$ – *Linie*), $p = 0,5$ ($*$ – *Linie*), $p = 0,7$ ($\triangle$ – *Linie*) *und* $p = 0,95$ ($\diamond$ – *Linie*)

Für $p = 0,5$ sind die Gewichte symmetrisch um $\mathrm{E}\big(H_{Hyp}(0,5;31;50)\big) = 25,5$ (vgl. Tabelle 2). Ein Vergleich der Schätzung von gegensätzlich extremen Quantilen zeigt, daß die Gewichtung der Ordnungsstatistiken nicht analog erfolgt. Zum Beispiel liegt der Modus der Gewichte für $p = 0,05$ bei Index $j = 1$, während sich das Gewicht für $p = 0,95$ bei Index $j = 50$ sogar zu Null ergibt. Beim Vergleich verschiedener Quantile scheint sich die Gewichtung bei der Schätzung von extrem kleinen Quantilen mehr auf wenige Ordnungsstatistiken zu konzentrieren. Die Schätzer $K_{r:k,n}(X_1,\ldots,X_n)$ mit $r = [(k+1)p]$ und $-K_{r:k,n}(-X_1,\ldots,-X_n)$ mit $r = [(k+1)(1-p)]$ sind folglich im allgemeinen nicht gleichzusetzen.

In Tabelle 2 sind zu den Gewichten aus Abbildung 1 einige Kennzahlen der entsprechenden Negativ–Hypergeometrischen Verteilung wie auch die zugehörigen Werte $r = [(k+1)p]$ aufgelistet.

Tabelle 2 *Erwartungswert und Varianz der Zufallsvariablen H_{Hyp} zum KL–Schätzer und der zugehörige Wert $r = [(k+1)p]$ für $n = 50$, $k = 31$ und verschiedene Werte von p.*

p	0,05	0,3	0,5	0,7	0,95
Erwartungswert	1,59	14,34	25,50	35,06	47,81
Varianz	0,89	5,94	7,34	6,31	1,72
$r = [32 \cdot p]$	1	9	16	22	30

Es wird deutlich, daß sich auch die Varianzen der Zufallsvariablen H_{Hyp} nicht symmetrisch um $p = 0,5$ verhalten. Es gilt sogar:

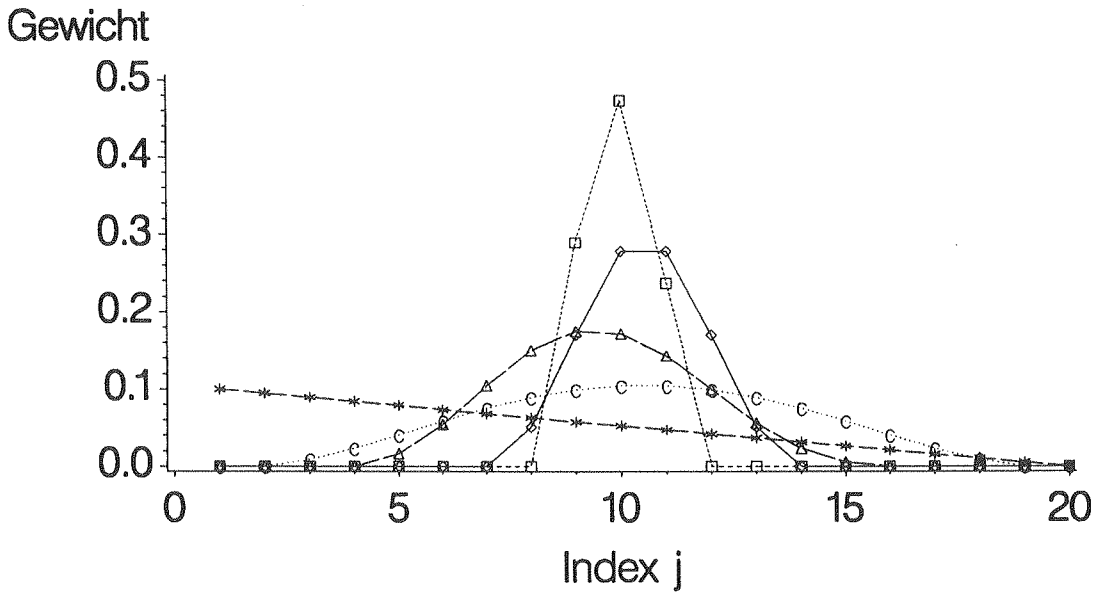
$$\mathrm{Var}\big(H_{Hyp}(0,05;31;50)\big) \approx \frac{1}{2} \cdot \mathrm{Var}\big(H_{Hyp}(0,95;31;50)\big),$$

was obige Feststellung zur Konzentration der Gewichte bestätigt. Für $p = 0,5$ weist $H_{Hyp}(0,5;31;50)$ die maximale Variabilität auf, weil die Gewichte am gleichmäßigsten über die entsprechenden $n - k + 1$ Ordnungsstatistiken verteilt werden.

Abbildung 2 zeigt die Verteilung der Gewichte des KL–Schätzers für $\xi_{0,5}$ bei $n = 20$ und verschiedenen Unterstichprobenumfängen k. Die zugehörigen Werte sind in Tabelle 4 im Anhang B aufgeführt.

Abgesehen für den Fall $k = 2$ wird jeweils das größte Gewicht den Ordnungsstatistiken $X_{9:20}, X_{10:20}$ und/oder $X_{11:20}$ zugeordnet, wobei hier $X_{[(n+1)p]:n} = X_{10:20}$ ist. Die Wahl $k = 2$ bei $n = 20$ scheint für die Schätzung von $\xi_{0,5}$ ungeeignet zu sein, da sich die Gewichtung nach keiner der obigen Ordnungsstatistiken orientiert. Auch für $k = 1$ besteht kein Zusammenhang zwischen den Ordnungsstatistiken und der Größe der einzelnen Gewichte, weil der KL–Schätzer das arithmetische Mittel ergeben würde und jede Ordnungsstatistik mit $\frac{1}{n} = \frac{1}{20}$ gewichtet wäre (nicht abgebildet). Es scheint daher sinnvoll, bei gegebenen p und n die Unterstichprobenumfänge k auszuschließen, die eine wie hier bei $k = 2$ ungeeignete Schätzung liefern.

Abbildung 2 *Gewichte des KL–Schätzers für* $n = 20$, $p = 0,5$ *und verschiedene Werte von* k.



$k = 2$ $(* - \textit{Linie})$, $k = 5$ $(\circ - \textit{Linie})$, $k = 10$ $(\triangle - \textit{Linie})$, $k = 15$ $(\diamond - \textit{Linie})$, *und* $k = 18$ $(\square - \textit{Linie})$

Nach KAIGH, CHENG (1991a, S. 544f) steht die Verteilung der Gewichte des KL–Schätzers im engen Zusammenhang zu der stetigen Betaverteilung mit Parametern r und $k-r+1$. Sei $b_{r,k,n}$ eine Zähldichte, die auf der Menge $\{\frac{1}{n+1}, \ldots, \frac{n}{n+1}\}$ definiert ist, mit zugehöriger Verteilungsfunktion $B_{r,k,n}$ und gelte

$$P\big(H_{Hyp} = j\big) = \frac{\binom{j-1}{r-1}\binom{n-j}{k-r}}{\binom{n}{k}} =: b_{r,k,n}\left(\frac{j}{n+1}\right), \tag{3.2}$$

$k, n \in \mathbb{N}$, $k \leq n$ und $r = [(k+1)p] > 0$, $p \in (0,1)$.

Für festes k und r gilt: Wenn $\frac{j_n}{n+1} \longrightarrow u$ für $n \to \infty$, $j_n \in \{1, \ldots, n\}$ und $u \in (0,1)$, dann ist

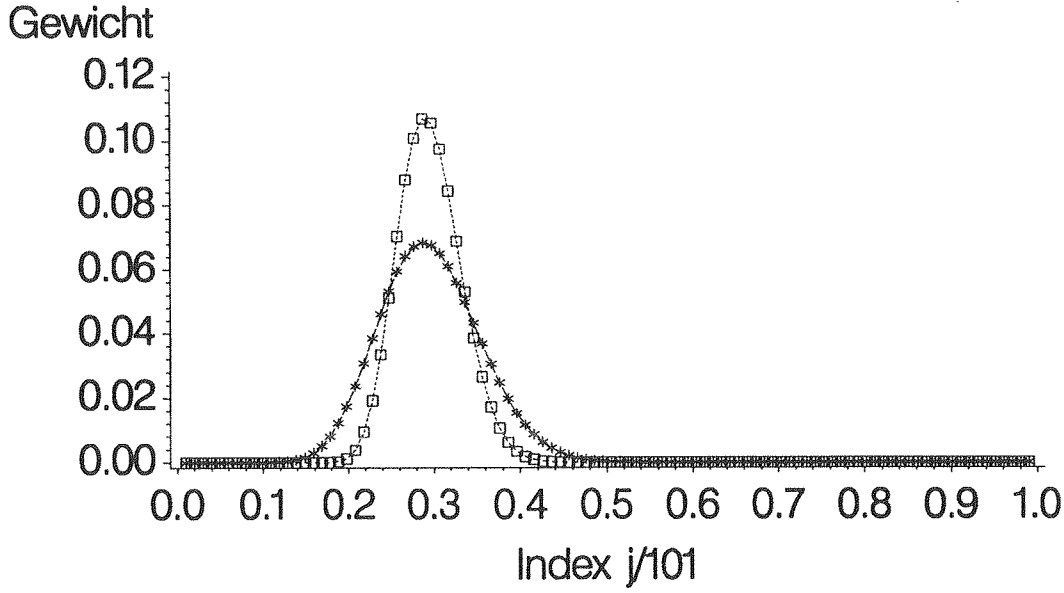
$$n \cdot b_{r,k,n}\left(\frac{j_n}{n+1}\right) \longrightarrow m_{r,k-r+1}(u) \quad \text{für} \quad n \to \infty. \tag{3.3}$$

Dabei ist $m_{r,k-r+1}$ die Dichte der Betaverteilung mit Parametern r und $k-r+1$ (KAIGH, CHENG, 1991a, S. 544).

Abbildung 3 vergleicht das n–fache der Gewichte des KL–Schätzers für $n = 100$, $k = 60$ und $p = 0,3$ über der Menge $\{\frac{1}{n+1}, \ldots, \frac{n}{n+1}\} \subset (0,1)$ mit der entsprechenden

Dichte der Betaverteilung. Vermutlich ist die Konvergenzgeschwindigkeit in (3.3) gering, da die Abbildung für $n = 100$ noch keine gute Approximation der wie oben transformierten Gewichte durch eine Dichte der Betaverteilung zeigt.

Abbildung 3 *Das n-fache der Gewichte des KL-Schätzers für* $n = 100$, $k = 60$ *und* $p = 0,3$ *(* $*$ *– Linie) und die Dichte der Betaverteilung mit Parametern* $r = 18$ *und* $k - r + 1 = 43$ *(□ – Linie).*



Aus (3.3) folgern KAIGH und CHENG (1991a, S. 545) weiter, daß gilt:

$$B_{r,k,n}(u) \longrightarrow M_{r,k-r+1}(u) \quad \text{für} \quad n \to \infty, u \in (0,1), \tag{3.4}$$

wobei $B_{r,k,n}$ die Verteilungsfunktion zur Zähldichte $b_{r,k,n}$ und $M_{r,k-r+1}$ die Verteilungsfunktion zur Betaverteilung mit Parametern r und $k - r + 1$ ist.

Für die Zufallsvariable H_{Hyp}, welche die Gewichte des KL–Schätzers repräsentiert, folgt damit aus (3.4):

$$\frac{1}{n+1} \cdot H_{Hyp} \xrightarrow{V} \text{Beta}(r, k - r + 1),$$

denn es gilt:

$$B_{r,k,n}(u) \quad = \sum_{x_i \in \{\frac{1}{n+1}, \ldots, \frac{n}{n+1}\}} b_{r,k,n}(x_i) \cdot \mathbb{1}_{(-\infty,u]}(x_i)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} b_{r,k,n}\left(\frac{i}{n+1}\right) \cdot \mathbb{1}_{(-\infty,u(n+1)]}(i) \\
&\overset{(3.2)}{=} P\big(H_{Hyp} \le u(n+1)\big) \\
&= P\left(\frac{1}{n+1} \cdot H_{Hyp} \le u\right).
\end{aligned}
$$

Abschließend wird bereits eine Eigenschaft des KL–Schätzers angefügt, die direkt aus den Eigenschaften seiner Gewichte resultiert.

Bemerkung 3.5 *Da die Summe der Gewichte des KL–Schätzers 1 beträgt, ist dieser Lage–invariant, das heißt, es gilt:*

$$K_{[(k+1)p]:k,n}(X_1 + c, \ldots, X_n + c) = K_{[(k+1)p]:k,n}(X_1, \ldots, X_n) + c, \quad c \in \mathbb{R},$$

*(*KAIGH, LACHENBRUCH, *1982, S. 2220).*

Der KL–Schätzer als U–Statistik

Eine U–Statistik U_n, $n \in \mathbb{N}$, ist eine Statistik mit der Darstellung

$$U_n = \frac{1}{\binom{n}{k}} \sum_{i=1}^{\binom{n}{k}} h(X_1^i, \ldots, X_k^i). \tag{3.5}$$

Dabei ist $h : \mathbb{R}^k \to \mathbb{R}$ eine in ihren k Argumenten symmetrische und meßbare Funktion, und $X_1^i, \ldots, X_k^i$ ist die i–te Unterstichprobe vom Umfang k. Die Funktion h heißt auch „Kern der Länge k“ (nach SERFLING, 1980, S. 172 und WITTING, 1985, S. 27f).

Aus (3.5) läßt sich die folgende Darstellung herleiten:

$$U_n = \mathrm{E}_F\big(h(X_1, \ldots, X_k) \,|\, (X_{1:n}, \ldots, X_{n:n})\big) =: g(X_{1:n}, \ldots, X_{n:n}) \tag{3.6}$$

mit meßbarer Funktion $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ und $r = [(k+1)p] > 0$ (SERFLING, 1980, S. 176).

Der KL–Schätzer ist damit auch eine U–Statistik mit Kern

$$h(X_1, \ldots, X_k) = X_{r:k}, \quad r = [(k+1)p] > 0, \tag{3.7}$$

(KAIGH, LACHENBRUCH, 1982, S. 2218), und mit (3.6) gilt die Darstellung

$$K_{r:k,n} = \mathrm{E}_F\big(X_{[(k+1)p]:k} \,|\, (X_{1:n}, \ldots, X_{n:n})\big).$$

Aus der Einordnung in die Klasse der L– und U–Statistiken resultieren die Eigenschaften des KL–Schätzers, die im folgendem Abschnitt vorgestellt werden.

3.2 Eigenschaften des Kaigh–Lachenbruch–Schätzers

3.2.1 Erwartungswert

Allgemein ist eine U–Statistik mit Kern h der Länge k ein erwartungstreuer Schätzer für $\mathrm{E}_F\big(h(X_1, \ldots, X_k)\big)$ (SERFLING, 1980, S. 172f). Daher gilt folgendes Korollar.

Korollar 3.6 *Falls der Erwartungswert des KL–Schätzers mit $r = [(k+1)p]$ existiert, so gilt:*

$$\begin{aligned} \mathrm{E}_F\big(K_{r:k,n}\big) &= \mathrm{E}_F\big(X_{r:k}\big) \\ &= \int_0^1 Q(u)\, m_{r:k-r+1}(u)\, du \\ &= \mathrm{E}_F\big(Q(B)\big), \end{aligned} \tag{3.8}$$

wobei die Zufallsvariable B betaverteilt ist mit Parametern r und $k-r+1$ und die Dichte

$$m_{r:k-r+1}(u) = \frac{1}{B(r, k-r+1)} u^{r-1}(1-u)^{k-r}, \quad 0 < u < 1 \tag{3.9}$$

besitzt (KAIGH, LACHENBRUCH, 1982, S. 2220).

Beweis: Als U–Statistik mit Kern h der Länge k wie in (3.7), ist der KL–Schätzer $K_{[(k+1)p]:k,n}$ ein erwartungstreuer Schätzer für

$$\mathrm{E}_F\big(h(X_1, \ldots, X_k)\big) = \mathrm{E}_F\big(X_{r:k}\big)$$

(SERFLING, 1980, S. 172f). Weiter gilt nach DAVID (1981, S. 34):

$$\mathrm{E}_F(X_{r:k}) = k \cdot \binom{k-1}{r-1} \int_0^1 Q(u) u^{r-1}(1-u)^{k-r}\, du.$$

Es folgt:

$$\begin{aligned} \mathrm{E}_F(X_{r:k}) &= \int_0^1 Q(u) \underbrace{k \cdot \binom{k-1}{r-1}}_{\frac{1}{B(r,k-r+1)}} u^{r-1}(1-u)^{k-r}\, du \\ &= \int_0^1 Q(u)\, m_{r:k-r+1}(u)\, du \end{aligned}$$

mit $m_{r:k-r+1}(u)$ gemäß (3.9), $r = [(k+1)p]$.

■

Da im allgemeinen die $[(k+1)p]$-te Ordnungsstatistik einer Stichprobe vom Umfang k keinen erwartungstreuen Schätzer für $\xi_p = Q(p)$ darstellt, ist auch der KL-Schätzer im allgemeinen ein verzerrter Schätzer für ξ_p. Wie jedoch KAIGH und LACHENBRUCH (1982, S. 2225, 2230) bemerken, bildet sowohl die Quantilschätzung für die Rechteckverteilung auf $[0,1]$, wenn $(k+1)p \in \mathbb{N}$, als auch die Schätzung des Medians für eine symmetrische Verteilung, wenn k ungerade ist, eine Ausnahme. Folgende Beispiele zeigen diesen Sachverhalt.

Beispiel 3.7 *Für $F \sim \mathrm{R}(0,1)$ gilt $Q(u) = u$ für alle $u \in [0,1]$, d.h.*

$$\mathrm{E}_F(X_{r:k}) = \mathrm{E}_F(B) = \frac{r}{k+1} = \frac{[(k+1)p]}{k+1}$$

(DAVID, 1981, S. 35). Dabei ist B betaverteilt mit Parametern r und $k-r+1$. Für den Fall $(k+1)p \in \mathbb{N}$ folgt

$$\mathrm{E}_F(K_{r:k,n}) = p = Q(p) = \xi_p,$$

und damit ist hier der KL-Schätzer erwartungstreu für ξ_p.

Beispiel 3.8 *Sei F die Verteilungsfunktion einer symmetrischen Verteilung. Damit gilt*

$$Q(0,5+y) = 2 \cdot \xi_{0,5} - Q(0,5-y) \quad \textit{für alle} \quad y \in [0;0,5). \tag{3.10}$$

Zu schätzen sei $\xi_{0,5}$, d.h. es ist $p = 0,5$ und $r = [(k+1)p] = \frac{k+1}{2}$ für k ungerade. Somit ist $m_{r:k-r+1} = m_{\frac{k+1}{2}:\frac{k+1}{2}} =: m$ eine symmetrische Dichte um $0,5$. Es gilt

$$\mathrm{E}_F\left(K_{\frac{k+1}{2}:k,n}\right) = \mathrm{E}_F\left(X_{\frac{k+1}{2}:k}\right)$$

$$
\begin{aligned}
&\overset{(3.8)}{=} \int_0^1 Q(u) \cdot m(u)\, du \\
&= \int_0^{0,5} Q(u) \cdot m(u)\, du + \int_{0,5}^1 Q(u) \cdot m(u)\, du \\
&\overset{(3.10)}{=} \int_0^{0,5} Q(0,5-y) \cdot m(0,5-y)\, dy \\
&\quad + \int_0^{0,5} (2 \cdot \xi_{0,5} - Q(0,5-y)) \cdot m(0,5-y)\, dy \\
&= \xi_{0,5}.
\end{aligned}
$$

Hiermit ist die Erwartungstreue des KL–Schätzers für $\xi_{0,5}$ in der betrachteten Situation gezeigt.

Da k unabhängig von n festgelegt wird, kann auch keine asymptotische Erwartungstreue für $n \to \infty$ sichergestellt werden, weil der Erwartungswert des KL–Schätzers nur von k, nicht aber von n abhängt. Könnte die Annahme eines festen Unterstichprobenumfangs k in der Definition des KL–Schätzers vernachlässigt werden, und würde k zusammen mit n wachsen, so wäre die asymptotische Erwartungstreue gegeben, wenn $X_{[(k_n+1)p]:k_n}$ gleichmäßig integrierbar ist (nach Theorem A, SERFLING, 1980, S. 14 und mit Verwendung von (2.5)).

3.2.2 Varianz und Varianzschätzer

Folgender allgemeiner Satz gilt für die Varianz von U–Statistiken.

Satz 3.9 *Sei $U_n(X_1, \ldots, X_n)$ eine U–Statistik mit Kern h der Länge k, $n, k \in \mathbb{N}$ und $k \leq n$, und es gelte $\mathrm{E}_F\big(h^2(X_1, \ldots, X_k)\big) < \infty$. Dann ist*

$$
\mathrm{Var}_F\big(U_n(X_1, \ldots, X_n)\big) = \frac{1}{\binom{n}{k}} \sum_{c=1}^{k} \binom{k}{c} \binom{n-k}{k-c} \zeta_c,
$$

wobei $\zeta_c = \mathrm{Var}_F\left(\mathrm{E}_F\big(h(X_1, \ldots, X_k) | (X_1, \ldots, X_c)\big)\right)$ und $0 \leq \zeta_1 \leq \ldots \leq \zeta_k < \infty$ (HOEFFDING, 1948, S. 299, SERFLING, 1980, S. 182f, LEE, 1990, S. 12).

Beweis: siehe LEE, 1990, S. 12f.

Bemerkung 3.10 *Die Darstellung für* ζ_c, $c = 1, \ldots, k$, *ergibt sich aus*

$$\begin{aligned} \zeta_c &= \mathrm{Var}_F\big(h_c(X_1, \ldots, X_c)\big) \quad \textit{mit} \\ h_c(x_1, \ldots, x_c) &= \mathrm{E}_F\big(h(X_1, \ldots, X_k) | X_1 = x_1, \ldots, X_c = x_c\big) \end{aligned}$$

*(*SERFLING, *1980, S. 177, 182).*

Die Ausdrücke ζ_c, $c = 1, \ldots, k$, können auch wie folgt als Kovarianzen interpretiert werden:

$$\zeta_c = \mathrm{Cov}\big(h(X_1^i, \ldots, X_k^i), h(X_1^j, \ldots, X_k^j)\big),$$

wobei $X_1^i, \ldots, X_k^i$ und $X_1^j, \ldots, X_k^j$, $i, j \in \{1, \ldots, \binom{n}{k}\}$, zwei beliebige Unterstichproben vom Umfang k mit c gemeinsamen Elementen sind (LEE, 1990, S. 11). Daraus erklärt sich, daß ζ_c mit größerem c nicht fällt (vgl. Satz 3.9).

Für den KL–Schätzer als U–Statistik mit Kern $h(X_1, \ldots, X_k)$ gemäß (3.7) kann die Varianz mit Anwendung von Satz 3.9 wie folgt exakt formuliert werden.

Korollar 3.11 *Sei* $\mathrm{E}_F\big((X_{r:k})^2\big) < \infty$ *für* $n, k \in \mathbb{N}$, $k \leq n$ *und* $r = [(k+1)p] > 0$ *mit* $0 < p < 1$. *Dann gilt:*

$$\mathrm{Var}_F(K_{r:k,n}) = \frac{1}{\binom{n}{k}} \sum_{c=1}^{k} \binom{k}{c} \binom{n-k}{k-c} \zeta_c,$$

wobei $\zeta_c = \mathrm{Var}_F\big(\mathrm{E}_F(X_{r:k} | (X_1, \ldots, X_c))\big)$.

Bemerkung 3.12

(i) Eine notwendige Bedingung für die Voraussetzung $\mathrm{E}_F\big((X_{r:k})^2\big) < \infty$ *in Korollar 3.11 ist gegeben durch die Bedingung* $\mathrm{E}_F(X^2) < \infty$ *(siehe z.B.* DAVID, *1981, S. 34).*

(ii) Besonders für große Werte von k erweist sich die Varianzformel in Korollar 3.11 als rechenaufwendig, da viele Ausdrücke der Form ζ_c *berechnet werden müssen.*

(iii) Für $c = k$ *ist* $\zeta_k = \mathrm{Var}_F(X_{r:k})$, *und damit ist*

$$\begin{aligned} \mathrm{Var}_F(K_{r:k,n}) &\leq \frac{1}{\binom{n}{k}} \sum_{c=1}^{k} \binom{k}{c} \binom{n-k}{k-c} \mathrm{Var}_F(X_{r:k}) \\ &= \underbrace{\left(1 - \frac{\binom{n-k}{k}}{\binom{n}{k}}\right)}_{(*)} \cdot \mathrm{Var}_F(X_{r:k}). \end{aligned}$$

Dabei ist $(*)$ *gleich eins für* $n < 2k$, *und* $(*)$ *ist Element von* $(0,1)$ *für* $n \geq 2k$. *D.h., zumindest für* $n < 2k$ *ist die Varianz des KL–Schätzers* $K_{[(k+1)p]:k,n}$ *kleiner oder gleich der Varianz von* $X_{[(k+1)p]:k}$ *und damit von* $X_{[(n+1)p]:n}$.

Bei unbekannter Verteilungsfunktion F kann $\mathrm{Var}_F(K_{r:k,n})$ nicht explizit berechnet werden. Daher werden nun Schätzer für die Varianz des KL–Schätzers vorgestellt, die durch die Resampling–Verfahren, dem Jackknife bzw. dem Bootstrap, hergeleitet werden.

Der Jackknife–Schätzer der Varianz

Im allgemeinen ist es möglich, die Varianz einer U–Statistik $U_n(X_1,\ldots,X_n)$, $n \in \mathbb{N}$, mit dem Jackknife–Verfahren zu schätzen. Dieses Verfahren wird in Anhang C.1 kurz erläutert. Mit der dort eingeführten Notation gilt folgender Satz für die Jackknife–Varianz von U–Statistiken.

Satz 3.13 *Sei* $U_n(X_1,\ldots,X_n)$ *eine U–Statistik mit Kern* h *der Länge* k, $n,k \in \mathbb{N}$ *und* $k \leq n$, *dann gilt:*

$$\widehat{\mathrm{Var}}_{Jack}(U_n) = \frac{n-1}{n}\sum_{i=1}^{n}(U_{(-i)} - U_n)^2 \tag{3.11}$$

$$= \frac{n-1}{n^2}\cdot\frac{1}{\binom{n-1}{k}^2}\cdot\sum_{c=0}^{k}(cn-k^2)\cdot\eta_c, \tag{3.12}$$

wobei für $c = 0,1,\ldots,k$

$$\eta_c = \sum_{\mathcal{M}} h(X_1^i,\ldots,X_k^i)\cdot h(X_1^j,\ldots,X_k^j)$$

ist mit

$$\mathcal{M} := \Big\{(\{X_1^i,\ldots,X_k^i\},\{X_1^j,\ldots,X_k^j\}) \;\Big|\; |\{X_1^i,\ldots,X_k^i\}\cap\{X_1^j,\ldots,X_k^j\}| = c,\ i,j \in \{1,\ldots,\tbinom{n}{k}\}\Big\} \tag{3.13}$$

als die Menge aller möglichen Paare von Unterstichproben vom Umfang k, *die genau* c *gemeinsame Elemente aufweisen,* $c = 0,1,\ldots k$.

Beweis: Mit $\bar{U}_{(.)} = \frac{1}{n}\sum_{i=1}^{n} U_{(-i)} = U_n$ (LEE, 1990, S. 219) und der Definition für die Jackknife–Varianz mit Ausdruck (C.1) im Anhang C.1 folgt (3.11). Für (3.12) siehe LEE (1990, S. 218f).

■

Unter Verwendung von Satz 3.13 kann die Jackknife–Varianz des KL–Schätzers als eine mögliche Approximation seiner Varianz wie folgt formuliert werden.

Korollar 3.14 *Für $n, k \in \mathbb{N}$, $k \leq n$ und $r = [(k+1)p] > 0$ mit $p \in (0,1)$ ist die Jackknife–Varianz des KL–Schätzers gegeben durch*

$$\begin{aligned} \widehat{\mathrm{Var}}_{Jack}(K_{r:k,n}) &= \frac{n-1}{n}\sum_{i=1}^{n}\left(K_{r:k,n-1}^{(-i)} - K_{r:k,n}\right)^2 \\ &= \frac{n-1}{n^2}\cdot\frac{1}{\binom{n-1}{k}^2}\cdot\sum_{c=0}^{k}(cn-k^2)\cdot\eta_c, \end{aligned} \tag{3.14}$$

wobei $K_{r:k,n-1}^{(-i)}$ der KL–Schätzer ist, welcher auf der reduzierten Stichprobe $X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n$ definiert ist, und

$$\eta_c = \sum_{\mathcal{M}} X_{r:k}^{i}\cdot X_{r:k}^{j}$$

ist mit $\mathcal{M}$ gemäß (3.13), $c = 0,\ldots,k$.

Bemerkung 3.15

(i) Die Jackknife–Varianz des KL–Schätzers der Form (3.14) wird in KAIGH, CHENG *(1991b, S. 981) vorgestellt.*

(ii) In KAIGH *(1983, S. 2433f) wird die Jackknife–Varianz des KL–Schätzers diskutiert, jedoch nicht explizit formuliert.*

(iii) Nach LEE *(1990, S. 224) soll die Jackknife–Varianz einer U–Statistik ein konsistenter Schätzer für die Varianz dieser U–Statistik sein, und somit wäre $\widehat{\mathrm{Var}}_{Jack}(K_{r:k,n})$ ein konsistenter Schätzer für die Varianz des KL–Schätzers. Gemäß* KAIGH, CHENG *(1991b, S. 981) und* EFRON *(1979, S. 6) ist jedoch keine Konsistenz für den Fall $k = n$ gegeben. Es ist daher noch eine kritische Überprüfung dieses Sachverhalts notwendig.*

Der Bootstrap–Schätzer der Varianz

Für die Anwendung des Bootstrap–Verfahrens zur Bestimmung eines Varianzschätzers sind u.a. zwei grundlegende Vorgehensweisen bekannt (vgl. Anhang C.2).

Zum einen kann der Bootstrap–Varianzschätzer (Bootstrap–Varianz) mit einem Monte–Carlo–Algorithmus approximiert werden, wenn die Varianz nicht explizit als Funktion $C(F,n)$ von F und n, $n \in \mathbb{N}$, angegeben wird. Diese Vorgehensweise wählen KAIGH und CHENG (1991b, S. 980) und geben die Bootstrap–Varianz des KL–Schätzers für $k = n$ an.

Zum anderen kann die Bootstrap–Varianz direkt mit der Funktion $C(F_n,n)$ bestimmt werden, wenn ein expliziter Ausdruck $C(F,n)$ der Varianz als Funktion von F und n bekannt ist. Die Anwendung von Ergebnissen für U–Statistiken ermöglicht einen solchen Ansatz für den KL–Schätzer, der bisher in der Literatur noch nicht beschrieben wurde.

Herleitung der Bootstrap–Varianz durch einen Monte–Carlo–Algorithmus

KAIGH und CHENG (1991b, S. 980) bestimmen die Bootstrap–Varianz des KL–Schätzers für $k = n$. Obwohl durch einen Monte–Carlo–Algorithmus hergeleitet, kann die Bootstrap–Varianz direkt und unabhängig von Bootstrap–Wiederholungen berechnet werden. Es gilt folgendes Korollar.

Korollar 3.16 *Für $n \in \mathbb{N}$, $k = n$ und $r = [(k+1)p] = [(n+1)p] > 0$ lautet die Bootstrap–Varianz des KL–Schätzers*

$$\widehat{\mathrm{Var}}_{Boot}(K_{r:n,n}) = \sum_{i=1}^{n} \left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} m_{r:n-r+1}(x)\,dx \right) \cdot \left(K_{i:n,n} - K_{r:n,n}\right)^2 \quad \text{f.s.}.$$

Dabei ist

$$m_{r:n-r+1}(x) = \frac{1}{B(r,n-r+1)}\, x^{r-1}(1-x)^{n-r},\ 0 < x < 1, \tag{3.15}$$

die Dichte der Betaverteilung mit Parametern r und $n-r+1$ (KAIGH, CHENG, 1991b, S. 980).

Beweis: Aus den B, $B \in \mathbb{N}$ fest, Bootstrap–Wiederholungen $K^{*1}_{r:k,n}, \ldots, K^{*B}_{r:k,n}$, $n, k \in \mathbb{N}$, $k \leq n$ und $r = [(k+1)p] > 0$, ergibt sich folgende Approximation für die Bootstrap–Varianz:

$$\widetilde{\mathrm{Var}}_{Boot}(K_{r:k,n}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(K^{*b}_{r:k,n} - \frac{1}{B} \sum_{b=1}^{B} K^{*b}_{r:k,n} \right)^2 .$$

KAIGH und CHENG (1991b, S. 980) beschränken sich auf den Fall $k = n$, für den gilt:

$$K_{r:n,n} = X_{r:n}, \qquad r = [(n+1)p]. \tag{3.16}$$

Es folgt

$$\widetilde{\mathrm{Var}}_{Boot}(K_{r:n,n}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(X^{*b}_{r:n} - \frac{1}{B} \sum_{b=1}^{B} X^{*b}_{r:n} \right)^2 . \tag{3.17}$$

Auch wenn sich die Bootstrap–Stichproben $X^{*b}_1, \ldots, X^{*b}_n$, $b = 1, \ldots, B$, unabhängig von der Reihenfolge der Ziehung zu denselben Elementen realisieren, werden sie für den Monte–Carlo–Algorithmus verwendet. Daher entsprechen die Bootstrap–Stichproben B Unterstichproben vom Umfang n, die mit Zurücklegen und mit Berücksichtigung der Anordnung aus einer Stichprobe vom Umfang n gezogen werden. Es gilt für alle $b = 1, \ldots, B$:

$$p_i := P\left(X^{*b}_{r:n} = X_{i:n}\right) = \int_{\frac{i-1}{n}}^{\frac{i}{n}} m_{r:n-r+1}(x)\, dx$$

mit $m_{r:n-r+1}$ wie in (3.15) und $i = 1, \ldots, n$ (KAIGH, CHENG, 1991a, S. 541ff).

Weiter sei die zufällige Anzahl der Zufallsvariablen $X^{*1}_{r:n}, \ldots, X^{*B}_{r:n}$, die der Zufallsvariablen $X_{i:n}$ entsprechen, als N_i notiert, wobei gilt: $N_i \sim \mathrm{Bin}(B, p_i)$, $i = 1, \ldots, n$.

Aus (3.17) folgt daher nach Zusammenfassung gleicher Summanden:

$$\begin{aligned} \widetilde{\mathrm{Var}}_{Boot}(K_{r:n,n}) &= \frac{1}{B-1} \sum_{i=1}^{n} N_i \cdot \left(X_{i:n} - \frac{1}{B} \sum_{b=1}^{B} X^{*b}_{r:n} \right)^2 \\ &= \frac{B}{B-1} \sum_{i=1}^{n} \frac{N_i}{B} \cdot \left(X_{i:n} - \frac{1}{B} \sum_{b=1}^{B} X^{*b}_{r:n} \right)^2 . \end{aligned}$$

Für $B \to \infty$ gelten folgende Beziehungen:

$$\lim_{B \to \infty} \frac{B}{B-1} = 1,$$

$$\lim_{B\to\infty} \frac{N_i}{B} = p_i \quad \text{f.s.} \quad \text{(vgl. SERFLING, 1980, S. 9)}$$

und

$$\lim_{B\to\infty} \frac{1}{B} \sum_{b=1}^{B} X_{r:n}^{*b} = \mathrm{E}_F(X_{r:n}^* | X_1, \ldots, X_n) \quad \text{f.s.} \quad \text{(HALL, 1992, S. 287f).}$$

Mit der Existenz obiger Grenzwerte, der Tatsache, daß die Bootstrap–Stichproben wie F_n verteilt sind $(*)$ und der Definition des KL–Schätzers für den Fall $k = n$ $(**)$, ergibt sich der Bootstrap–Schätzer der Varianz von $K_{r:n,n}$ wie folgt:

$$\begin{aligned}
\widehat{\mathrm{Var}}_{Boot}(K_{r:n,n}) &= \lim_{B\to\infty} \widetilde{\mathrm{Var}}_{Boot}(K_{r:n,n}) \\
&= \sum_{i=1}^{n} \left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} m_{r:n-r+1}(x)\, dx \right) \cdot (X_{i:n} - \mathrm{E}_F(X_{r:n}^* | X_1, \ldots, X_n))^2 \quad \text{f.s.} \\
&\overset{(*)}{=} \sum_{i=1}^{n} \left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} m_{r:n-r+1}(x)\, dx \right) \cdot (X_{i:n} - \mathrm{E}_F(X_{r:n} | X_1, \ldots, X_n))^2 \\
&= \sum_{i=1}^{n} \left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} m_{r:n-r+1}(x)\, dx \right) \cdot (X_{i:n} - X_{r:n})^2 \\
&\overset{(**)}{=} \sum_{i=1}^{n} \left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} m_{r:n-r+1}(x)\, dx \right) \cdot (K_{i:n,n} - K_{r:n,n})^2 \,.
\end{aligned}$$

■

Bemerkung 3.17 *Eine Verallgemeinerung der in Korollar 3.16 vorgestellten Bootstrap–Varianz für beliebige $k \in \{2, \ldots, n-1\}$ wäre mit höherem Aufwand verbunden, weil die Beziehung entsprechend zu (3.16) komplizierter ist (vgl. Satz 3.4). Für $k = 1$ ergibt sich die Bootstrap–Varianz des Stichprobenmittels (*LEE, *1990, S. 230).*

Direkte Bestimmung der Bootstrap–Varianz

Für diesen Ansatz ist eine explizite Darstellung der Varianz als Funktion von F und n notwendig, welche für U–Statistiken U_n, $n \in \mathbb{N}$, im folgenden Satz formuliert ist.

Satz 3.18 *Sei $U_n(X_1,\ldots,X_n)$ eine U–Statistik mit Kern h der Länge k, $n,k \in \mathbb{N}$ und $k \leq n$. Dann ist*

$$\begin{aligned}\mathrm{Var}_F\big(U_n(X_1,\ldots,X_n)\big) &= \left(\frac{1}{\binom{n}{k}} \cdot \sum_{c=0}^{k} \binom{k}{c}\binom{n-k}{k-c}\tau_c(F)\right) - \tau_0(F)\\ &=: C(F,n),\end{aligned}$$

wobei

$$\begin{aligned}\tau_c(F) &:= \mathrm{E}_F\big(h_c^2(X_1,\ldots,X_c)\big)\\ &= \mathrm{E}_F\big(\mathrm{E}_F{}^2\big(h(X_1,\ldots,X_k)|(X_1,\ldots,X_c)\big)\big)\\ &= \int_0^1 \ldots \int_0^1 h(x_1,\ldots,x_k)\cdot h(x_1,\ldots,x_c,x_{k+1},\ldots,x_{2k-c}) \prod_{i=1}^{2k-c} dF(x_i)\end{aligned}$$

ein Funktional von F der Ordnung $2k-c$ ist, $c=1,\ldots,k$, und

$$\begin{aligned}\tau_0(F) &:= \mathrm{E}_F{}^2\big(h(X_1,\ldots,X_k)\big)\\ &= \left(\int_0^1 \ldots \int_0^1 h(x_1,\ldots,x_k)\prod_{i=1}^{k} dF(x_i)\right)^2\end{aligned}$$

*ein Funktional von F der Ordnung $2k$ ist (*LEE*, 1990, S. 11f, 231f).*

Beweis:

• In dem Beweis zu Theorem 2 in LEE (1990, S. 11f) wird gezeigt, daß gilt:

$$\begin{aligned}\tau_c(F) &:= \mathrm{E}_F\big(h_c^2(X_1,\ldots,X_c)\big)\\ &= \int_0^1 \ldots \int_0^1 h(x_1,\ldots,x_k)\cdot h(x_1,\ldots,x_c,x_{k+1},\ldots,x_{2k-c}) \prod_{i=1}^{2k-c} dF(x_i).\end{aligned}$$

Weiterhin gilt nach Theorem 1 (ii) in LEE (1990, S. 10f):

$$\begin{aligned}\mathrm{E}_F\big(h(X_1,\ldots,X_k)\big) &= \int_0^1 \ldots \int_0^1 h(x_1,\ldots,x_k)\prod_{i=1}^{k} dF(x_i)\\ &= \mathrm{E}_F\big(h_c(X_1,\ldots,X_c)\big).\end{aligned} \tag{3.18}$$

• Mit (3.18) gilt:

$$\begin{aligned}\mathrm{Var}_F\big(h_c(X_1,\ldots,X_c)\big) &= \mathrm{E}_F\big(h_c^2(X_1,\ldots,X_c)\big) - \mathrm{E}_F{}^2\big(h_c(X_1,\ldots,X_c)\big)\\ &= \mathrm{E}_F\big(h_c^2(X_1,\ldots,X_c)\big) - \mathrm{E}_F{}^2\big(h(X_1,\ldots,X_k)\big)\\ &= \tau_c(F) - \tau_0(F)\end{aligned}$$

und folglich mit der Formulierung der Varianz einer U–Statistik U_n in Satz 3.9 und $\zeta_c = \mathrm{Var}_F\big(h_c(X_1,\ldots,X_c)\big)$ (vgl. Bemerkung 3.10):

$$\begin{aligned}
\mathrm{Var}_F\big(U_n(X_1,\ldots,X_n)\big) &= \frac{1}{\binom{n}{k}} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c}\zeta_c \\
&= \frac{1}{\binom{n}{k}} \cdot \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c} \cdot \big(\tau_c(F) - \tau_0(F)\big) \\
&= \frac{1}{\binom{n}{k}} \cdot \sum_{c=0}^{k} \binom{k}{c}\binom{n-k}{k-c} \cdot \tau_c(F) \\
&\quad - \frac{1}{\binom{n}{k}} \cdot \sum_{c=0}^{k} \binom{k}{c}\binom{n-k}{k-c} \cdot \tau_0(F).
\end{aligned}$$

Mit Berücksichtigung eines Additionstheorems für Binomialkoeffizienten (Formel (2.4) in BRONSTEIN, SEMENDJAJEW, 1987, S. 105) reduziert sich die zweite Summe zu $1 \cdot \tau_0(F)$, und damit ist die Behauptung gezeigt.

■

Mit der Betrachtung von $C(F_n, n)$ zu der Varianzdarstellung aus Satz 3.18 ergibt sich die Bootstrap–Varianz einer U–Statistik U_n. Damit läßt sich die Bootstrap–Varianz des KL–Schätzers direkt bestimmen, wie folgendes Korollar besagt.

Korollar 3.19 *Es seien Bindungen bei F_n ausgeschlossen. Dann ist für $n, k \in \mathbb{N}$, $k \leq n$ und $r = [(k+1)p] > 0$ mit $p \in (0,1)$ die Bootstrap–Varianz des KL–Schätzers gegeben durch*

$$\widehat{\mathrm{Var}}_{Boot}(K_{r:k,n}) = \left(\frac{1}{\binom{n}{k}} \cdot \sum_{c=0}^{k} \binom{k}{c}\binom{n-k}{k-c}\tau_c(F_n)\right) - \tau_0(F_n)$$

mit

$$\tau_c(F_n) = \frac{1}{n^{2k-c}} \cdot \sum_{\mathcal{M}_1} X_{r:k}^{S_{(1)}} \cdot X_{r:k}^{S_{(2)}}.$$

Dabei bezeichnet

$$\begin{aligned}
\mathcal{M}_1 := \Big\{ &(S_{(1)}, S_{(2)}) = (\{X_{i_1},\ldots,X_{i_k}\}, \{X_{j_1},\ldots,X_{j_k}\}) \;\big|\; \\
&|\{X_{i_1},\ldots,X_{i_k}\} \cap \{X_{j_1},\ldots,X_{j_k}\}| = c,\; i_1,\ldots,i_k, j_1,\ldots,j_k = 1,\ldots,n \Big\},
\end{aligned}$$

c=0,1,... k, die Menge aller möglichen Paare von Stichproben vom Umfang k, die mit Zurücklegen und ohne Berücksichtigung der Anordnung aus der Stichprobe vom Umfang n gezogen werden können und die genau c gemeinsame Elemente aufweisen.

Beweis: Die Bootstrap–Varianz folgt unmittelbar aus der Darstellung der Varianz einer U–Statistik (Satz 3.18) mit Übergang zu den empirischen Funktionalen $\tau_c(F_n)$, $c = 0, 1, \ldots, k$, $n, k \in \mathbb{N}$ und $k \leq n$. Es gilt:

$$\begin{aligned}\tau_c(F_n) &= \int_0^1 \cdots \int_0^1 h(x_1, \ldots, x_k) \cdot h(x_1, \ldots, x_c, x_{k+1}, \ldots, x_{2k-c}) \prod_{i=1}^{2k-c} dF_n(x_i) \\ &= \frac{1}{n^{2k-c}} \cdot \sum_{i_1=1}^{n} \cdots \sum_{i_{2k-c}=1}^{n} h(X_{i_1}, \ldots, X_{i_k}) \cdot h(X_{i_1}, \ldots, X_{i_c}, X_{i_{k+1}}, \ldots, X_{i_{2k-c}}).\end{aligned}$$

Für den KL–Schätzer als eine U–Statistik mit Kern $h(X_1, \ldots, X_k) = X_{r:k}$, $r = [(k+1)p] > 0$, folgt somit die Behauptung.

■

3.2.3 Konsistenz

Eine U–Statistik mit Kern h der Länge k ist stark konsistent für $\mathrm{E}_F\big(h(X_1, \ldots, X_k)\big)$, wenn $\mathrm{E}_F|h(X_1, \ldots, X_k)| < \infty$ (SERFLING, 1980, S. 190, Theorem A). Daher gilt folgendes Korollar für den KL–Schätzer gemäß Definition 3.1.

Korollar 3.20 *Sei* $\mathrm{E}_F|X_{r:k}| < \infty$ *für* $n \in \mathbb{N}$, $k \in \mathbb{N}$ *fest,* $k \leq n$, *und* $r = [(k+1)p] > 0$ *mit* $p \in (0,1)$. *Dann gilt:*

$$K_{r:k,n} \xrightarrow{f.s.} \mathrm{E}_F(X_{r:k}) \quad \textit{für} \quad n \to \infty.$$

Beweis: Nach Theorem A in SERFLING (1980, S. 190) gilt für eine U–Statistik U_n, $n \in \mathbb{N}$, mit Kern h der Länge k: Wenn $\mathrm{E}_F|h(X_1, \ldots, X_k)| < \infty$, dann folgt

$$U_n \xrightarrow{f.s.} \mathrm{E}_F\big(h(X_1, \ldots, X_k)\big) \quad \text{für} \quad n \to \infty.$$

Für den KL–Schätzer mit Kern $h(X_1, \ldots, X_k) = X_{r:k}$, $r = [(k+1)p] > 0$, folgt damit die Behauptung.

■

Mit Korollar 3.20 ist jedoch nicht die starke Konsistenz des KL–Schätzers für ξ_p, $p \in (0,1)$, gewährleistet, da im allgemeinen $\mathrm{E}_F(X_{r:k}) \neq \xi_p$ ist, $r = [(k+1)p]$.

3.2.4 Konvergenz in Verteilung

Es sei k ein fester, aber beliebiger Unterstichprobenumfang. Unter Ausnutzung eines Theorems über das Approximationsverhalten von U-Statistiken kann das folgende Korollar formuliert werden (nach KAIGH, LACHENBRUCH, 1982, S. 2222f, Theorem 3.1).

Korollar 3.21 *Sei* $\mathrm{Var}_F\left(\mathrm{E}_F(X_{r:k}|X_1)\right) > 0$, $\mathrm{E}_F\left((X_{r:k})^2\right) < \infty$, *und die Ableitung* Q' *existiere. Für* $p \in (0,1)$, $k \in \mathbb{N}$ *fest,* $r = [(k+1)p] > 0$ *und* $n \in \mathbb{N}$ *mit* $n \to \infty$ *und* $1 \leq k \leq n$ *gilt:*

$$\sqrt{n}\,\left(K_{r:k,n} - \mathrm{E}_F(X_{r:k})\right) \xrightarrow{V} \mathrm{N}\left(0, \sigma^2_{r:k}(F)\right)$$

mit

$$\begin{aligned} \mathrm{E}_F(X_{r:k}) &= \int_0^1 Q(u)\, m_{r:k-r+1}(u)\, du, \\ \sigma^2_{r:k}(F) &= k^2 \cdot \mathrm{Var}_F\left(\mathrm{E}_F(X_{r:k}|X_1)\right) \\ &= \int_0^1 \int_0^1 \left(\min(u,v) - u\,v\right) Q'(u)Q'(v)\, m_{r:k-r+1}(u) m_{r:k-r+1}(v)\, du dv, \end{aligned} \tag{3.19}$$

wobei

$$m_{r:k-r+1}(x) = \frac{1}{B(r, k-r+1)}\, x^{r-1}(1-x)^{k-r}, \quad 0 < x < 1, \tag{3.20}$$

die Dichte der Betaverteilung mit Parametern r *und* $k-r+1$ *bezeichnet (*KAIGH, LACHENBRUCH, *1982, S. 2222f, Theorem 3.1).*

Beweis: Mit $\mathrm{E}_F\left((X_{r:k})^2\right) < \infty$ und $\zeta_1 = \mathrm{Var}_F\left(\mathrm{E}_F(X_{r:k}|X_1)\right) > 0$ gilt nach einem Theorem für U-Statistiken (siehe SERFLING (1980), S. 192, Theorem A):

$$\sqrt{n}\,\left(K_{r:k,n} - \mathrm{E}_F(X_{r:k})\right) \xrightarrow{V} \mathrm{N}\left(0, k^2\zeta_1\right).$$

• Nach DAVID (1981, S. 34) gilt (vgl. Beweis zu Korollar 3.6):

$$\mathrm{E}_F(X_{r:k}) = \int_0^1 Q(u)\; m_{r:k-r+1}(u)\; du$$

mit $m_{r:k-r+1}(u)$ gemäß (3.20), $r = [(k+1)p]$.

• Sei die Projektion der Ordnungsstatistik $X_{r:k}$ definiert als

$$\hat{X}_{r:k} := \left(\sum_{i=1}^{k} \mathrm{E}_F(X_{r:k}|X_i)\right) - (k-1)\mathrm{E}_F(X_{r:k})$$

(STIGLER, 1969, S.772).

Da die Zufallsvariablen $X_1, \ldots, X_k$ stochastisch unabhängig und identisch verteilt sind, gilt: $\mathrm{E}_F(X_{r:k}|X_1), \ldots, \mathrm{E}_F(X_{r:k}|X_k)$ sind als Funktion von X_1 bzw. X_2 usw. unabhängig und identisch verteilt. Daher ist

$$\mathrm{Var}_F(\hat{X}_{r:k}) \;=\; k \cdot \mathrm{Var}_F\big(\mathrm{E}_F(X_{r:k}|X_1)\big) \;=\; k\zeta_1.$$

Mit der Existenz von Q' gilt nach STIGLER (1969, S. 774, Lemma 2) für die Projektion der Ordnungsstatistik $X_{r:k}$:

$$k \cdot \mathrm{Var}_F(\hat{X}_{r:k}) \;=\; \int_0^1 \int_0^1 \big(\min(u,v) - u\,v\big)\, Q'(u)Q'(v)\, m_{r:k-r+1}(u) m_{r:k-r+1}(v)\, du dv.$$

Daraus folgt:

$$\begin{aligned} k^2\zeta_1 &= k^2 \cdot \mathrm{Var}_F\big(\mathrm{E}_F(X_{r:k}|X_1)\big) \\ &= k \cdot \mathrm{Var}_F(\hat{X}_{r:k}) \\ &= \int_0^1 \int_0^1 \big(\min(u,v) - u\,v\big)\, Q'(u)Q'(v)\, m_{r:k-r+1}(u) m_{r:k-r+1}(v)\, du dv. \end{aligned}$$

■

Bemerkung 3.22

(i) Für den KL–Schätzer als U–Statistik genügen die Bedingungen $\mathrm{E}_F\big((X_{r:k})^2\big) < \infty$ *und* $\mathrm{Var}_F\left(\mathrm{E}_F(X_{r:k}|X_1)\right) > 0$, *um die Konvergenz gegen eine normalverteilte Zufallsvariable zu zeigen. Wenn zusätzlich* $Q' = \frac{1}{f(\xi_p)}$ *(*KAIGH, LACHENBRUCH, *1982, S. 2218) existiert, kann die approximative Varianz durch obigen Integralausdruck formuliert werden.*

(ii) Eine notwendige Bedingung für die Voraussetzung $\mathrm{E}_F\big((X_{r:k})^2\big) < \infty$ *in Korollar 3.21 ist gegeben durch* $\mathrm{E}_F(X^2) < \infty$ *(siehe z.B.* DAVID, *1981, S. 34).*

Bemerkung 3.23 *Es ist denkbar, daß mit Ergebnissen aus* SERFLING *(1980, S. 192, 194) unter den Annahmen, daß* $\mathrm{E}_F\big((X_{r:k})^2\big) < \infty$ *und*

$$\zeta_1 \;=\; \mathrm{Var}_F\left(\mathrm{E}_F(X_{r:k}|X_1)\right) \;=\; 0 \;<\; \zeta_2 \;=\; \mathrm{Var}_F\left(\mathrm{E}_F\big(X_{r:k}|(X_1,X_2)\big)\right)$$

ist, gezeigt werden kann, daß $n\cdot\big(K_{r:k,n} - \mathrm{E}_F(X_{r:k})\big)$ *in Verteilung gegen eine gewichtete Summe von unabhängigen* ${\chi_1}^2$*–verteilten Zufallsvariablen konvergiert.* SERFLING *(1980) zeigt dieses unter gewissen Annahmen für eine U–Statistik.*

Läßt man zu, daß mit n auch der Unterstichprobenumfang $k = k_n$ in Abhängigkeit von n wachsen darf (daher k indiziert mit n), so kann mit einer Zusatzbedingung folgendes Korollar formuliert werden (KAIGH, LACHENBRUCH, 1982, S. 2234f, Theorem 5.1).

Korollar 3.24 *Es existiert ein* $\varepsilon > 0$*, so daß* $\lim_{x\to\infty} |x|^{\varepsilon}\big(1 - F(x) + F(-x)\big) = 0$*. Wenn die Dichte* f *in einer Umgebung um* ξ_p *stetig ist und* Q' *existiert, dann gilt mit* $p \in (0,1)$*,* $k_n \in \mathbb{N}$*,* $r_n = [(k_n+1)p] > 0$ *und* $n \in \mathbb{N}$ *mit* $1 \le k_n \le n$

(i) für $k_n \longrightarrow \infty$*, wenn* $n \longrightarrow \infty$*:*

$$\sqrt{n}\;\left(K_{r_n:k_n,n} \;-\; \mathrm{E}_F(X_{r_n:k_n})\right) \;\xrightarrow{V}\; \mathrm{N}\left(0, \sigma_p^2(F)\right),$$

wobei mit $m_{r_n:k_n-r_n+1}$ *analog zu (3.20)*

$$\begin{aligned}\sigma_p^2(F) \;&=\; p(1-p)\left(Q'(p)\right)^2,\\ \mathrm{E}_F(X_{r_n:k_n}) \;&=\; \int_0^1 Q(u)\, m_{r_n:k_n-r_n+1}(u)\, du.\end{aligned}$$

(ii) für $k_n \longrightarrow \infty$ *und* $\liminf_{n\to\infty} \frac{k_n}{n} > 0$*:*

$$n\,\mathrm{E}_F\left(K_{r_n:k_n,n} \;-\; X_{[(n+1)p]:n}\right)^2 \;\longrightarrow\; 0$$

und

$$\sqrt{n}\;\left(K_{r_n:k_n,n} \;-\; \xi_p\right) \;\xrightarrow{V}\; \mathrm{N}\left(0, \sigma_p^2(F)\right)$$

*(*KAIGH, LACHENBRUCH*, 1982, S. 2234f, Theorem 5.1).*

Bemerkung 3.25 *Korollar 3.24 (ii) zeigt die asymptotische Äquivalenz im quadratischen Mittel der Schätzer* $K_{r_n:k_n,n}$ *und* $X_{[(n+1)p]:n}$ *(*KAIGH, LACHENBRUCH*, 1982, S. 2236).*

3.2.5 Asymptotische Konfidenzintervalle für Quantile

Ein Schätzer kann zur Konstruktion von Konfidenzintervallen für eine interessierende Charakeristik verwendet werden, wenn er sich zu einem sogenannten Pivot standardisieren läßt, einer Statistik, die selbst von der interessierenden Charakteristik abhängt, deren Verteilung jedoch nicht. Im günstigsten Fall ist die Verteilung dieser Statistik bekannt. Ist nur ihre asymptotische Verteilung bekannt, so können asymptotische Konfidenzintervalle konstruiert werden.

KAIGH (1983, S. 2434, Formel (3.1)) schlägt ein solches asymptotisches Konfidenzintervall für ξ_p, $p \in (0,1)$, mit dem KL–Schätzer vor, welches auf einer Jackknife–Varianz basiert. Wie in Bemerkung 3.15 (ii) schon erwähnt worden ist, wird diese jedoch nicht explizit formuliert und sei deshalb hier mit „Jack–Var" bezeichnet, weil die Gleichheit zur Jackknife–Varianz aus Korollar 3.14 nicht sichergestellt werden kann. KAIGH (1983) behauptet, daß der mit $\mathrm{E}_F(X_{r:k})$, $r = [(k+1)p] > 0$, und „Jack–Var" standardisierte KL–Schätzer in Verteilung gegen eine Normalverteilung mit Erwartungswert Null und Varianz $\sigma^2_{r:k}(F)$ (gemäß Korollar 3.21) konvergiert (Kaigh, 1983, S. 2434, Theorem 3.1 (iii)). Er schlägt das folgende approximative $(1-\alpha)$–Konfidenzintervall für ξ_p vor, $\alpha \in (0,1)$:

$$\left[K_{r:k,n} - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\text{Jack–Var}}{n}} \;,\; K_{r:k,n} + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\text{Jack–Var}}{n}} \right],$$

wobei $u_{1-\frac{\alpha}{2}}$ das $(1-\frac{\alpha}{2})$–Quantil der Standard–Normalverteilung bezeichnet.

Durch die oben genannte Standardisierung ist jedoch keine approximative Standard–Normalverteilung gegeben, da im allgemeinen die Varianz $\sigma^2_{r:k}(F)$ der approximierten Normalverteilung ungleich Eins ist. Somit erscheint die Verwendung von $u_{1-\frac{\alpha}{2}}$ nicht unmittelbar gerechtfertigt. Auch hängt der so standardisierte KL–Schätzer von $\mathrm{E}_F(X_{r:k})$ und nicht von ξ_p ab, so daß genaugenommen nur ein asymptotisches Konfidenzintervall für $\mathrm{E}_F(X_{r:k})$ konstruiert werden könnte. Es ist daher fraglich, ob die vorgegebene Überdeckungswahrscheinlichkeit von $(1-\alpha)$ für $n \to \infty$ eingehalten werden kann.

KAIGH und CHENG (1991b, S. 982) schlagen ein weiteres approximatives $(1-\alpha)$–Konfidenzintervall für ξ_p mit der Jackknife–Varianz des KL–Schätzers aus Korollar

3.14 vor. Es ist gegeben durch

$$\left[K_{r:k,n} - t_{n-k,1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathrm{Var}}_{Jack}(K_{r:k,n})} \; , \; K_{r:k,n} + t_{n-k,1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathrm{Var}}_{Jack}(K_{r:k,n})} \right], \tag{3.21}$$

wobei $t_{n-k,1-\frac{\alpha}{2}}$ das $\left(1-\frac{\alpha}{2}\right)$–Quantil der t–Verteilung mit $n-k$ Freiheitgraden ist.

Weiter stellen sie ein approximatives $(1-\alpha)$–Konfidenzintervall basierend auf der Bootstrap–Varianz des KL–Schätzers (vgl. Korollar 3.16) für $k = n$ vor, welches gegeben ist durch

$$\left[K_{r:n,n} - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathrm{Var}}_{Boot}(K_{r:n,n})} \; , \; K_{r:n,n} + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathrm{Var}}_{Boot}(K_{r:n,n})} \right]. \tag{3.22}$$

Für die Konfidenzintervalle (3.21) und (3.22) werden weder die verwendeten Pivots erwähnt, noch die Verwendung der Quantile der Standardnormalverteilung bzw. der t–Verteilung mit $n-k$ Freiheitgraden näher begründet. Auch ist unklar, warum die übliche Division durch $\sqrt{n}$ unterlassen wird.

Die Einhaltung der vorgegebenen Überdeckungswahrscheinlichkeiten für $n \to \infty$ wird daher auch für die von KAIGH und CHENG (1991b) vorgeschlagenen Konfidenzintervalle in Frage gestellt, was eine von KAIGH und CHENG (1991b, S. 983ff) durchgeführte Studie für einen festen Stichprobenumfang von $n = 19$ u.a. andeutet. Auch die Ergebnisse einer Simulationsstudie von STEINBERG und DAVIS (1985, S. 985ff) mit verschiedenen Stichprobenumfängen lassen dieses für das Konfidenzintervall (3.21) vermuten.

Für die Plausibilität der aufgeführten Konfidenzintervalle ist es notwendig, die Konvergenz in Verteilung ähnlich wie in Korollar 3.24 (ii) zu zeigen, wobei jeweils mit der Jackknife–Varianz sowie der Bootstrap–Varianz standardisiert wird. Damit könnten aus dem KL–Schätzer Pivots konstruiert werden, die echt von ξ_p abhängen.

3.2.6 Robustheit

Im Vergleich zu Quantilschätzern, die auf dem Mittel aller Ordnungsstatistiken $X_{1:n}, \ldots, X_{n:n}$ basieren, welche echt positiv gewichtet werden (z.B. der Harrell–Davis–Schätzer gemäß (2.7)), ist der KL–Schätzer als getrimmtes gewichtetes Mittel von Ordnungsstatistiken robuster gegenüber Ausreißern (KAIGH, 1983, S. 2429).

Die Robustheit des KL–Schätzers wird von der Stärke der Trimmung beeinflußt. Diese hängt vom Unterstichprobenumfang k und von p ab, da der Summationsindex des KL–Schätzers Werte zwischen r und $r+n-k$ annimmt, $r = [(k+1)p] > 0$.

Eine Kenngröße für die Robustheit eines Schätzers gegenüber Ausreißern ist der finite Stichprobenbruchpunkt ε_n^* (finite–sample breakdown point) für eine Stichprobe mit Umfang n (vgl. HAMPEL et al., 1986, S. 98). Er gibt den maximalen Anteil von Beobachtungen an, die beliebig abgeändert werden können, ohne daß sich der Schätzwert des Schätzers unbeschränkt verändern kann.

Definition 3.26 *Sei T_n ein Schätzer basierend auf einer Stichprobe $X_1, \ldots, X_n$, $n \in \mathbb{N}$. Dann ist der finite Stichprobenbruchpunkt des Schätzers T_n definiert als*

$$\varepsilon_n^*(T_n; x_1, \ldots, x_n) := \frac{1}{n} \max\left\{m; \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} |T_n(z_1, \ldots, z_n)| < \infty\right\},$$

*wobei sich für die Stichprobe die Realisationen $z_1, \ldots, z_n$ ergeben, wenn die m Realisationen $x_{i_1}, \ldots, x_{i_m}$ durch beliebige Werte $y_1, \ldots, y_m$ ersetzt werden (*HAMPEL *et al., 1986, S. 98).*

Aus der Darstellung des KL–Schätzers als gewichtete Summe von Ordnungsstatistiken (Satz 3.4) wird ersichtlich, daß maximal $\min(r-1, k-r)$ Beobachtungen beliebig abgeändert werden können, ohne daß sich der Schätzwert des KL–Schätzers über alle Schranken hinweg verändern kann.

Korollar 3.27 *Der finite Stichprobenbruchpunkt des KL–Schätzers $K_{[(k+1)p]:k,n}$ ist*

$$\varepsilon_n^*(K_{[(k+1)p]:k,n}; x_1, \ldots, x_n) = \frac{\min(r-1, k-r)}{n},$$

wobei $r := [(k+1)p]$.

Sind z.B. extreme Quantile zu schätzen, d.h. nimmt p Werte nahe 0 bzw. 1 an, so ergeben sich kleine Werte für den finiten Stichprobenbruchpunkt des KL–Schätzers, da auch extreme Ordnungsstatistiken für die Schätzung von ξ_p benötigt werden. Folglich ist der KL–Schätzer in diesem Fall wenig robust gegenüber Ausreißern.

3.2.7 Effizienzvergleich des Kaigh–Lachenbruch–Schätzers mit einem Stichprobenquantil

KAIGH und LACHENBRUCH (1982, S. 2226ff) vergleichen in einer Simulationsstudie den KL–Schätzer mit dem Stichprobenquantil $\hat{Q}_1(p) = X_{[(n+1)p]:n}$. Als Vergleichskriterium wird die relative Effizienz $\mathrm{eff}\big(K_{[(k+1)p]:k,n}, X_{[(n+1)p]:n}\big)$ der beiden Schätzer gewählt. Sie ist als reziprokes Verhältnis der jeweiligen mittleren quadratischen Fehler (MSE) definiert.

Definition 3.28 *Der mittlere quadratische Fehler des KL–Schätzers ist definiert als*

$$\begin{aligned} \mathrm{MSE}_F(K_{r:k,n}) &:= \mathrm{E}_F\left((K_{r:k,n} - \xi_p)^2\right) \\ &= \mathrm{Var}_F(K_{r:k,n}) + \left(\xi_p - \mathrm{E}_F(K_{r:k,n})\right)^2, \end{aligned}$$

wobei $n, k \in \mathbb{N}$, $k \leq n$, $p \in (0,1)$ *und* $r = [(k+1)p]$.

Definition 3.29 *Die relative Effizienz des KL–Schätzers für* $\xi_p, p \in (0,1)$, *zum Stichprobenquantil* $X_{[(n+1)p]:n}$ *ist definiert als:*

$$\mathrm{eff}\big(K_{[(k+1)p]:k,n}, X_{[(n+1)p]:n}\big) := \frac{\mathrm{MSE}\big(X_{[(n+1)p]:n}\big)}{\mathrm{MSE}\big(K_{[(k+1)p]:k,n}\big)}.$$

KAIGH und LACHENBRUCH (1982) führen 10000 Simulationen für $p = 0,5$ zu sieben symmetrischen Verteilungen (Rechteck–, Dreiecks–, Arcussinus–, logistische, Normal–, Doppelexponential– und Cauchyverteilung) durch, wobei $n = 99$ gewählt wird, und k zwischen den Werten 9,19,29,39 und 79 variiert. Da unter diesen Bedingungen die betrachteten Schätzer für den Median erwartungstreu sind (vgl. Beispiel 3.8), reduziert sich die relative Effizienz zu dem reziproken Verhältnis der beiden Varianzen.

Weiter wird das Verhalten der beiden Schätzer sowohl für symmetrische Verteilungen (wie oben) wie auch für nicht–symmetrische Verteilungen (Power– und Exponentialverteilung) mit $n = 99$ und Unterstichprobenumfängen $k = 39$ und $k = 79$ simuliert. Dabei nimmt p Werte aus $\{0,05; 0,10; \ldots; 0,90; 0,95\}$ an.

Es zeigt sich, daß der KL–Schätzer im allgemeinen eine höhere Effizienz gegenüber dem konventionellem Stichprobenquantil $X_{[(n+1)p]:n}$ aufweist. Dies trifft jedoch nicht

für die Schätzung extremer Quantile ($p \leq 0,10$ und $p \geq 0,90$) bei den untersuchten Verteilungen mit dicken Tails (Exponential–, Doppelexponential– und Cauchyverteilung) zu. Ebenso scheint die Schätzung extremer Quantile für die Power–, logistische und Normalverteilung problematisch zu sein.

Weiterhin ergeben die Simulationen, daß größere Unterstichprobenumfänge k zu einer verbesserten Schätzung extremer Quantile durch den KL–Schätzer führen und kleinere Umfänge k sich besser zur Schätzung mittlerer Quantile eignen.

KAIGH und LACHENBRUCH (1982, S. 2224f) berechnen explizit die Varianzausdrücke der jeweils approximierten Normalverteilungen (vgl. (2.5) und (3.19)) für beide Schätzer bei speziellen Rechteck–, logistischen, Exponential– und Powerverteilungen. Dabei ergibt sich folgender ungefährer Zusammenhang, der auch durch obige Simulationsstudie für $0,2 \leq p \leq 0,8$ bestätigt werden kann:

$$\frac{\sigma_p^2(F)}{\sigma_{r:k}^2(F)} \approx \left(1 - \frac{\binom{k+1}{r}^2}{\binom{2k+2}{2r}}\right)^{-1} > 1, \tag{3.23}$$

für $0,2 \leq p \leq 0,8$, $1 \leq k \leq n$ und $r = [(k+1)p] > 0$, wobei $\sigma_p^2(F) = \frac{p(1-p)}{f^2(\xi_p)}$ und $\sigma_{r:k}^2(F)$ gemäß (3.19).

3.3 Zur Wahl eines „optimalen“ Unterstichprobenumfangs

Wie bereits in Bemerkung 3.3 erwähnt, gehört der KL–Schätzer einer Schätzerklasse an, deren Elemente erst durch die Wahl vom Unterstichprobenumfang k eindeutig definiert sind. Ähnlich wie bei Kernschätzern für Quantile, wo die optimale Bandbreite bestimmt werden muß (vgl. Kapitel 2.2), ist für den KL–Schätzer ein „optimaler“ Unterstichprobenumfang k_{opt} als Glättungsparameter zu bestimmen.

Folgende Überlegung macht diese Notwendigkeit deutlich.

Es sei ξ_p, $p \in (0,1)$, für eine Stichprobe $X_1, \ldots, X_n$ zu schätzen. Für $k = 1$ ergibt sich das arithmetische Mittel als Schätzer für ξ_p (unabhängig davon, welches p aus $[0,5;1)$ betrachtet wird; vgl. Tabelle 1); die Wahl von $k = n$ führt dagegen zu dem Schätzer $X_{[(n+1)p]:n}$.

Wäre z.B. $\xi_{0,5}$ für eine Normalverteilung zu schätzen, ist das arithmetische Mittel als Schätzer für $\xi_{0,5}$ dem Schätzer $X_{[(n+1)p]:n}$ bezüglich der asymptotischen relativen Effizienz beider Schätzer überlegen (SERFLING, 1980, S. 50f, 85f). Hier wäre daher $k = 1$ geeignet. Dagegen würde für eine Doppelexponentialverteilung die Wahl $k = n$ zu einer asymptotisch effizienteren Schätzung für ξ_p führen (KAIGH, 1988, S. 2199). Offensichtlich hängt die Wahl von k_{opt} bei gegebenem n von p und der zu Grunde liegenden Verteilung F ab, welche jedoch unbekannt ist. Diese Feststellung wird durch die Simulationsstudie von KAIGH und LACHENBRUCH (1982) (Kapitel 3.2.7) bestätigt.

Zur Bestimmung von k_{opt} bietet sich an, die Minimierung des mittleren quadratischen Fehlers $\mathrm{Var}_F(K_{r:k,n}) + \left(\xi_p - \mathrm{E}_F(K_{r:k,n})\right)^2$ in Abhängigkeit von k für gegebenes p und n als Optimalitätskriterium zu vereinbaren (KAIGH, LACHENBRUCH, 1982, S. 2230).

KAIGH und LACHENBRUCH (1982) ersetzen die Varianz des KL–Schätzers durch den Ausdruck $\frac{1}{n}\sigma^2_{r:k}(F)$ aus Korollar 3.21. Weiter verwenden sie den exakten Ausdruck des Erwartungswerts $\mathrm{E}_F(K_{r:k,n}) = \mathrm{E}_F(X_{r:k})$ aus Korollar 3.6. So ergibt sich der folgende asymptotische Ausdruck für $\mathrm{MSE}_F(K_{r:k,n})$ (KAIGH, LACHENBRUCH, 1982, S. 2230):

Approximation 1

$$\begin{aligned}\mathrm{MSE}_F(K_{r:k,n}) &\approx \frac{1}{n}\sigma^2_{r:k}(F) + \left(\xi_p - \mathrm{E}_F(X_{r:k})\right)^2 \\ &= \frac{1}{n}\int_0^1\int_0^1 \left(\min(u,v) - u\,v\right) Q'(u)\,Q'(v)\,m_{r:k-r+1}(u)\cdot \\ &\quad \cdot\, m_{r:k-r+1}(v)\,du dv + \left(\xi_p - \int_0^1 Q(u)\,m_{r:k-r+1}(u)\,du\right)^2.\end{aligned}$$

Eine Vereinfachung dieser Approximation ist möglich, wenn $\sigma^2_{r:k}(F)$ und $\mathrm{E}_F(X_{r:k})$ weiter approximiert werden.

Unter der Annahme, daß $Q' = \frac{1}{f(\xi_p)}$ existiert, und die Dichte f für alle x, die in einer Umgebung von ξ_p liegen, annähernd konstant ist, d.h. $Q'(x) \approx Q'(p)$, läßt sich für $\sigma^2_{r:k}(F)$ der Zusammenhang in (3.23) herleiten (KAIGH, LACHENBRUCH, 1982, S. 2226, Formel (4.6)):

$$\sigma^2_{r:k}(F) \approx p(1-p)\left(Q'(p)\right)^2\left(1 - \frac{\binom{k+1}{r}^2}{\binom{2k+2}{2r}}\right). \tag{3.24}$$

Sei ebenso vorausgesetzt, daß die Quantilfunktion Q in einer Umgebung um den Erwartungswert $\frac{r}{k+1} \approx p$ der Betaverteilung mit Parametern r und $k-r+1$ dreimal differenzierbar ist. Dann kann Q durch die ersten drei Taylorpolynome approximiert werden, und man erhält den folgenden approximativen Ausdruck für $\mathrm{E}_F(X_{r:k})$ (KAIGH, LACHENBRUCH, 1982, S. 2230):

$$\mathrm{E}_F(X_{r:k}) \approx \xi_p + \underbrace{\frac{p(1-p)\,Q''(p)}{2(k+2)}}_{appr.\ Bias}. \tag{3.25}$$

Mit (3.24) und (3.25) resultiert dann die einfachere Approximation für $\mathrm{MSE}_F(K_{r:k,n})$ (KAIGH, LACHENBRUCH, 1982, S. 2230):

Approximation 2

$$\mathrm{MSE}_F(K_{r:k,n}) \approx \frac{1}{n} p(1-p)\left(Q'(p)\right)^2 \cdot \left(1 - \frac{\binom{k+1}{r}^2}{\binom{2k+2}{2r}}\right) + \left(\frac{p(1-p)\,Q''(p)}{2(k+2)}\right)^2.$$

Prinzipiell ist es möglich, den mittleren quadratischen Fehler des KL–Schätzers mit Hilfe einer der obigen Approximationen annähernd zu bestimmen und bezüglich k zu minimieren. Folgende Probleme ergeben sich dabei.

- Da die Verteilungsfunktion F der interessierenden Grundgesamtheit unbekannt ist, müssen die Quantilfunktion Q und ihre Ableitungen zuvor aus vorliegenen Daten geschätzt werden. Hierbei erweist sich Q_n' bzw. Q_n'' als ungeeigneter Schätzer für Q' bzw. Q'', da Q_n als Treppenfunktion in ihren Sprungstellen nicht differenzierbar ist und in den konstanten Bereichen eine Ableitung von Null besitzt.
- Durch die Approximation von $\mathrm{Var}_F(K_{r:k,n})$ und $\mathrm{E}_F(K_{r:k,n})$ kann der optimale Unterstichprobenumfang k nur mit einer gewissen Ungenauigkeit bestimmt werden. Insbesondere bei kleinem Stichprobenumfang n ist daher zu befürchten, daß der Optimierungsvorgang durch die Schätzung von Q und durch die durchgeführten Approximationen k_{opt} nur stark verzerrt schätzen könnte.

Wünschenswert ist daher eine Darstellung des mittleren quadratischen Fehlers des KL–Schätzers, die unabhängig von der Quantilfunktion Q ist und nur von n, k und p abhängt, um damit k_{opt} für gegebenes n und p bestimmen zu können.

Bemerkung 3.30

(i) Der approximative Ausdruck des Erwartungswertes (3.25) deutet darauf hin, daß der Bias umso kleiner wird, je größer der Unterstichprobenumfang k gewählt wird.

(ii) Unabhängig von Optimalitätseigenschaften des Schätzers schlagen KAIGH *und* LACHENBRUCH *(1982, S. 2221) vor, k so zu wählen, daß gilt:* $(k+1)p \in \mathbb{N}$. *Für* $p = 0,05$ *müßte k z.B. gleich 19 oder 29 (usw.) sein. Möglicherweise wird damit* k_{opt} *systematisch verfehlt.*

(iii) Für den Fall, daß k_{opt} *nicht eindeutig bestimmt werden kann, scheint es sinnvoll, den maximalen Wert aller optimalen k als* k_{opt} *zu wählen. Somit kann mit größerem Unterstichprobenumfang k und damit mit einer kleineren Anzahl von möglichen Unterstichproben der Rechenaufwand bei der Schätzung von* ξ_p *reduziert werden.*

Neben der von KAIGH und LACHENBRUCH (1982) vorgeschlagenen Methode k_{opt} zu bestimmen, sind auch noch andere Methoden in der Literatur zu finden, die im weiteren kurz vorgestellt werden.

Bestimmung von k_{opt} durch die asymptotische Äquivalenz zu einem Kern–Quantilschätzer

In SHEATHER und MARRON (1990) wird eine allgemeine Methode zur Schätzung der optimalen Bandbreite von verschiedenen Kern–Quantilschätzern (Kapitel 2.2) vorgestellt. Es wird nachgewiesen, daß der KL–Schätzer asymptotisch einem bestimmten Typ von Kern–Quantilschätzern entspricht. Damit ist es möglich, die Bestimmung der optimalen Bandbreite auf die Bestimmung eines asymptotisch optimalen Unterstichprobenumfangs k_{opt} zu übertragen.

Korollar 3.31 *Sei* Q'' *in einer Umgebung von p stetig. Für alle* $p \in (0,1)$ *(ausgenommen* $p = 0,5$*, wenn F Verteilungsfunktion einer symmetrischen Verteilung ist) ist der asymptotisch optimale Unterstichprobenumfang k für den KL–Schätzer* $K_{[(k+1)p]:k,n}$ *gegeben durch*

$$k_{opt} = \frac{p(1-p)}{\left(\alpha(\varphi)\,\beta(Q)\,n^{-\frac{1}{3}}\right)^2},$$

wobei φ die Dichte der Standard–Normalverteilung ist und

$$\alpha(\varphi) = \left(\frac{2 \int_{-\infty}^{\infty} u\,\varphi(u)\,\varphi^{-1}(u)\,du}{\left(\int_{-\infty}^{\infty} u^2\,\varphi(u)\,du\right)^2} \right)^{\frac{1}{3}} = \left(2 \int_{-\infty}^{\infty} u\,\varphi(u)\,\varphi^{-1}(u)\,du \right)^{\frac{1}{3}}$$

$$\beta(Q) = \left(\frac{Q'(p)}{Q''(p)} \right)^{\frac{2}{3}}$$

(nach SHEATHER, MARRON, *1990, S. 411f).*

Bemerkung 3.32

*(i) Ist $p = 0{,}5$ und F Verteilungsfunktion einer symmetrischen Verteilung, so kann k_{opt} mit obiger Methode nicht eindeutig bestimmt werden (*SHEATHER, MARRON, *1990, S. 411f).*

(ii) Wie bei der Approximation des mittleren quadratischen Fehlers des KL–Schätzers nach KAIGH *und* LACHENBRUCH *(1982) muß die Quantilfunktion Q mit ihren Ableitungen Q' und Q'' geschätzt werden. Die dort aufgezeigten Probleme sind daher auch hier von Bedeutung.*

Bestimmung von k_{opt} mit der Kreuz–Validierungsmethode

ZELTERMAN (1990, S. 345ff) schlägt für die Bestimmung von k_{opt} die Kreuz–Validierungsmethode vor, ein Resampling–Verfahren, welches im Anhang C.3 kurz beschrieben wird. Er erklärt das Vorgehen am Beispiel eines Kern–Quantilschätzers.

Als k_{opt} wird der Wert ausgewählt, welcher den kreuz–validierten Vorhersagefehler von $K_{r:k,n}$, $r = [(k+1)p]$, minimiert. Letzterer ist bei ZELTERMAN (1990, S. 346) mit Übertragung auf den KL–Schätzer wie folgt definiert:

$$\begin{aligned} \bar{M}(k) &:= \sum_{i=1}^{n} i(n-i+1)\left(K^{(-i)}_{r_i:k,n-1} - X_{i:n}\right)^2 \\ &\quad - \sum_{i=1}^{n} i(n-i+1)\left(K_{r_i:k,n} - X_{i:n}\right)^2 \end{aligned}$$

mit $r_i = [(k+1)\frac{i}{n+1}]$.

Auffällig ist, daß k_{opt} bei bekanntem n und p trotzdem unabhängig von p bestimmt wird. Es wird folglich weniger Information verwertet als zur Verfügung steht. Vermutlich könnte das Verfahren verbessert werden, wenn auch die Information über p berücksichtigt werden würde.

3.4 Alternative Definition des Kaigh–Lachenbruch–Schätzers

Im Hinblick auf die Übertragung des KL–Schätzers auf den zensierten Fall (Kapitel 4) wird in diesem Kapitel auch eine alternative Definition des KL–Schätzers eingeführt, die auf der empirischen Quantilfunktion Q_n basiert.

Definition 3.33 *Für feste Werte von $n \in \mathbb{N}$, $p \in (0,1)$ und $k \in \mathbb{N}$ mit $1 \leq k \leq n$ ist mit*

$$\overset{emp}{KL}(p,k,n) := \frac{1}{\binom{n}{k}} \sum_{i=1}^{\binom{n}{k}} Q_k^i(p)$$

der KL–Schätzer basierend auf der empirischen Quantilfunktion definiert. Dabei bezeichnet $Q_k^i(p)$ die empirische Quantilfunktion in der i–ten Unterstichprobe vom Umfang k an der Stelle p, $i = 1, \ldots, \binom{n}{k}$.

Mit der Darstellung (2.3) für die empirische Quantilfuntion (Kapitel 2.1) ergibt sich folgende Darstellung für die alternative Definition des KL–Schätzers:

$$\overset{emp}{KL}(p,k,n) = \frac{1}{\binom{n}{k}} \sum_{i=1}^{\binom{n}{k}} X^i_{]kp[+1:k} \quad .$$

Lemma 3.34 *Es gilt $]np[+1 = [(n+1)p]$ für alle $(n,p) \in \big(\mathbb{N} \times (0,1)\big) \backslash \mathbf{A}$, wobei $\mathbf{A} = \{(n,p) \in \mathbb{N} \times (0,1) : np + \varepsilon \in \mathbb{N}$ für ein ε mit $0 < p < \varepsilon < 1\}$.*

Beweis: siehe Anhang A.2.

Bemerkung 3.35

(i) Die Definitionen 3.1 und 3.33 des KL–Schätzers sind nicht gleich, da nach Lemma 3.34 Werte $n \in \mathbb{N}$ und $p \in (0,1)$ existieren, so daß gilt:

$$Q_n(p) = X_{]np[+1:n} \neq X_{[(n+1)p]:n}.$$

(ii) In Definition 3.33 ist keine Einschränkung für $p \in (0,1)$ und k der Form $[(k+1)p] > 0$ notwendig (vgl. Tabelle 1).

Für die alternative Definition des KL–Schätzers gilt der Satz 3.4 aus Kapitel 3.1 analog für $r =]kp[+1$. Folglich handelt es sich auch bei dem KL–Schätzer gemäß Definition 3.33 um einen L–Schätzer.

Ebenso liegt eine U–Statistik mit Kern $h(X_1, \ldots, X_k) = X_{]kp[+1:k}$ vor. Mit $r =]kp[+1$ können demnach prinzipiell die in Kapitel 3.2 aufgeführten Eigenschaften auf den Schätzer $\overset{emp}{KL}\,(p,k,n)$ übertragen werden, soweit nicht extra der Ausdruck $r = [(k+1)p]$ verwendet wurde. Letzteres betrifft die Beispiele 3.7 und 3.8. Die dort gezeigte Erwartungstreue gilt nicht für den alternativen KL–Schätzer.

Mit der fast sicheren Konvergenz des Stichprobenquantils $X_{]np[+1:n}$, $p \in (0,1)$, gegen ξ_p und mit der Annahme, daß mit dem Stichprobenumfang n auch der Unterstichprobenumfang k wächst, könnte ein Nachweis der starken Konsistenz gegen ξ_p für $\overset{emp}{KL}\,(p,k,n)$ versucht werden. Die Überlegungen basieren dabei auf dem Theorem 2.1 in FREES (1989, S. 31f), welches unter gewissen Annahmen die starke Konsistenz für U–Statistiken infiniter Ordnung gegen ξ_p sichert. Es lautet:

Sei U_n eine U–Statistik mit Kern h_n der Ordnung k_n und gelte:

(i) Die Funktion h_n ist gleichmäßig integrierbar.

(ii) $h_n(X_1, \ldots, X_{k_n}) \xrightarrow{V} h(X_1, X_2, \ldots)$ für $n \to \infty$
(Formel (1.3) in FREES, 1989, S. 29).

(iii) $\sum_{n=1}^{\infty} \mathrm{E}_F\big|h_n(X_1, \ldots, X_n) - h(X_1, X_2, \ldots)\big| < \infty$
(Bemerkung in FREES, 1989, S. 32).

Dann folgt: $U_n \xrightarrow{f.s.} \mathrm{E}_F\big(h(X_1, X_2, \ldots)\big)$.

Folgendes sei für den KL–Schätzer in Bezug auf die Annahmen (i)-(iii) angemerkt:

zu (i): Es bleibt zu untersuchen, ob und für welche F die gleichmäßige Integrierbarkeit von $h_n(X_1,\ldots,X_{k_n}) = X_{]k_np[+1:k_n}$ gewährleistet ist.

zu (ii): Mit der Voraussetzung, daß ξ_p, $p \in (0,1)$, die eindeutige Lösung von $F(x-) \leq p \leq F(x)$ ist, gilt nach SERFLING (1980, S. 75):

$$X_{]np[+1:n} \xrightarrow{f.s.} \xi_p, \quad n \to \infty.$$

Damit folgt auch die fast sichere Konvergenz von $X_{]k_np[+1:k_n}$ gegen ξ_p. Weiter folgt mit SERFLING (1980, S.10) die Konvergenz in Wahrscheinlichkeit

$$X_{]k_np[+1:k_n} \xrightarrow{W} \xi_p, \quad k_n \to \infty,$$

und damit die Konvergenz in Verteilung (Korollar A in SERFLING, 1980, S. 19). D.h. es gilt:

$$X_{]k_np[+1:k_n} \xrightarrow{V} \xi_p, \quad k_n \to \infty. \tag{3.26}$$

Setze $h(X_1, X_2, \ldots) = \xi_p$. Dann folgt (ii) mit $h_n(X_1,\ldots,X_{k_n}) = X_{]k_np[+1:k_n}$.

zu (iii): Mit (3.26) folgt die Konvergenz in Verteilung auch für $X_{]np[+1:n}$ gegen ξ_p, und mit (i) würde die gleichmäßige Integrierbarkeit von $X_{]np[+1:n}$ gelten. Damit würde mit Theorem A aus SERFLING (1980, S. 14) folgen: $\mathrm{E}_F|X_{]np[+1:n}| < \infty$. Mit der Darstellung

$$\mathrm{E}_F\left|X_{]np[+1:n} - \xi_p\right| = \frac{\sqrt{p(1-p)}}{\sqrt{n}\cdot f(\xi_p)} \cdot \int |x|\, d\Phi(x) + O\left(\frac{1}{n}\right)$$

(REISS, 1989, S. 208, Formel (6.1.4)) wäre dann

$$\lim_{n\to\infty} \mathrm{E}_F\left|X_{]np[+1:n} - \xi_p\right| = 0. \tag{3.27}$$

Das Konvergenzverhalten in (3.27) ist jedoch nur eine notwendige Bedingung für die in (iii) geforderte Konvergenz der Reihe

$$\sum_{n=1}^{\infty} \mathrm{E}_F\left|h_n(X_1,\ldots,X_n) - h(X_1,X_2,\ldots)\right| = \sum_{n=1}^{\infty} \mathrm{E}_F\left|X_{]np[+1:n} - \xi_p\right|.$$

Zum korrekten Nachweis von (iii) müßten zusätzlich geeignete Darstellungen oder Abschätzungen für $\mathrm{E}_F\left|X_{]np[+1:n} - \xi_p\right|$ gefunden werden, um eines der bekannten Konvergenzkriterien für Reihen anwenden zu können.

Kapitel 4

Der Kaigh–Lachenbruch–Schätzer im zensierten Fall

In Kapitel 3 wurde der KL–Schätzer für den unzensierten Fall betrachtet. Dieses Kapitel untersucht nun den KL–Schätzer für zensierte Daten. Abschnitt 4.1 erläutert kurz die Datensituation im zensierten Fall. Anschließend wird der KL–Schätzer jeweils für das Modell zufälliger Rechtszensierung (Abschnitt 4.2) und für das Koziol–Green–Modell (Abschnitt 4.3) definiert, und seine Eigenschaften werden im Ansatz diskutiert. Dabei liegt der Schwerpunkt der Betrachtungen auf dem Koziol–Green–Modell.

4.1 Die Datensituation im zensierten Fall

Bei der Analyse von Überlebenszeiten liegt in der Regel die zensierte Datensituation vor. Interessierende Größen sind die Zeit bis zum Ausfall einer physikalischen Komponente oder die Überlebenszeit (Zeitspanne bis zum Tod) einer biologischen Einheit, z.B. eines Patienten in einer klinischen Studie. Hierbei können Stichproben häufig nicht vollständig beobachtet werden, da z.B. Patienten aus unterschiedlichen Gründen die Studie vorzeitig verlassen haben. Von diesen Patienten ist nur bekannt, daß sie einen bestimmten Zeitpunkt überlebt haben, nicht aber ihr eigentlicher Todeszeitpunkt. Die entsprechenden Beobachtungen liegen dann als sogenannte Zensierungen (Zensierungszeiten) vor. Zensierungen sind im allgemeinen dadurch ge-

kennzeichnet, daß sie die Information über die interessierende Zufallsvariable nur teilweise oder gar nicht beinhalten.

Von den bekannten Zensierungsarten, der Intervall-, Links- und Rechtszensierung, wird hier nur letztere betrachtet. Rechtszensierung tritt auf, wenn zu große Realisationen nicht mehr beobachtet werden können. Drei verschiedene Typen von Rechtszensierungen seien an dieser Stelle erwähnt:

- *Typ-I–Zensierung:* Sei z als feste Zensierungszeit gegeben. Dann können nur solche Elemente beobachtet werden, deren Überlebenszeiten den Wert z nicht überschreiten. Die Anzahl der zensierten Beobachtungen ist hierbei zufällig, die Zensierungszeit nicht.

- *Typ-II–Zensierung:* Sei $n \in \mathbb{N}$ der Stichprobenumfang und $r \in \mathbb{N}$ mit $r \leq n$ gegeben. Es können nur die Elemente mit den r kleinsten Überlebenszeiten beobachtet werden. Die Anzahl der zensierten Beobachtungen ist hierbei fest, die Zensierungszeit ist zufällig.

- *Zufällige Zensierung:* Sowohl die Anzahl der zensierten Beobachtungen als auch die Zensierungszeiten sind zufällig (weitere Ausführungen in Kapitel 4.2.1)

(vgl. MILLER, 1981, S. 1ff).

In klinischen Studien veranlassen Zeit, Kosten und ethische Gründe den Forscher häufig, die Studie nach einer gewissen Zeitspanne (Typ-I–Zensierung) oder nach dem Tod des r–ten Patienten (Typ-II–Zensierung) abzubrechen. Beginnen und verlassen die Patienten eine Studie zufällig und zu unterschiedlichen Zeitpunkten, so ist das Modell zufälliger Rechtszensierung anwendbar (Kapitel 4.2.1).

Es sind verschiedene Veröffentlichungen bekannt, die speziell das Konzept der Quantilschätzung auf den zensierten Fall übertragen haben. Einige werden hier genannt.

SANDER (1975) stellt eine Übertragung des empirischen Quantils $Q_n(p)$ auf das Modell zufälliger Rechtszensierung vor. Ebenso führen dieses PADGETT (1986) für den Parzen–Quantil–Schätzer $\hat{Q}_4(p)$, GRONEN (1993) für den Harrell–Davis–Schätzer $\hat{Q}_3(p)$ und LIO und PADGETT (1987) für den KL–Schätzer durch. Weiter wird das empirische Quantil in GIJBELS und VERAVERBEKE (1989) und in GHORAI (1991)

für das Koziol–Green–Modell untersucht. Dieses ist ein Spezialfall des Modells zufälliger Rechtszensierung. Auch GRONEN (1993) betrachtet den Harrell–Davis–Schätzer im Koziol–Green–Modell sowie GHORAI (1991) den Parzen–Quantil–Schätzer.

Der nächste Abschnitt behandelt den KL–Schätzer im Modell zufälliger Rechtszensierung.

4.2 Der Kaigh–Lachenbruch–Schätzer im Modell zufälliger Rechtszensierung

In 4.2.1 wird das Modell zufälliger Rechtszensierung zusammen mit dem Kaplan–Meier–Schätzer vorgestellt. Anschließend erfolgt in 4.2.2 die Definition des KL–Schätzers in diesem Modell mit einer kurzen Diskussion seiner Eigenschaften.

4.2.1 Das Modell zufälliger Rechtszensierung

Seien $X_1, \ldots, X_n, n \in \mathbb{N}$, reelle, nicht–negative, unabhängig und identisch verteilte Zufallsvariablen, die wie eine Zufallsvariable X verteilt sind mit stetiger Verteilungsfunktion F und Lebesgue–Dichte f. Weiter seien $Y_1, \ldots, Y_n$ reelle, nicht–negative, unabhängig und identisch wie eine Zufallsvariable Y verteilte Zufallsvariablen mit stetiger Verteilungsfunktion G und Lebesgue–Dichte g.

Dabei repräsentiert X die Überlebenszeit und Y die Zensierungszeit. Die zugehörigen Verteilungsfunktionen F und G sind unbekannt, und somit kommen nichtparametrische Verfahren zur Anwendung.

Die Paare $(Z_1, \Delta_1), \ldots, (Z_n, \Delta_n)$ bezeichnen die Zufallsvektoren der tatsächlich beobachteten Zeiten (zensiert oder unzensiert). Dabei sind $Z_1, \ldots, Z_n, n \in \mathbb{N}$, reelle, nicht–negative, unabhängig und identisch wie eine Zufallsvariable Z verteilte Zufallsvariablen mit Verteilungsfunktion H, und $\Delta_1, \ldots, \Delta_n$ sind unabhängig, identisch wie eine Zufallsvariable Δ Bernoulli–verteilte Zufallsvariablen. Es ist

$$\begin{aligned} Z_i &:= \min\{X_i, Y_i\} \quad \text{und} \\ \Delta_i &:= \mathbb{1}_{[0,Y_i]}(X_i) = \begin{cases} 1, & Z_i \text{ ist nicht zensiert} \\ 0, & Z_i \text{ ist zensiert} \end{cases} \end{aligned}$$

mit $i = 1, \ldots, n$ und $n \in \mathbb{N}$. Mit der Bedingung

„Die Zufallsvariablen X und Y sind stochastisch unabhängig." (4.1)

und obigen Bezeichnungen ist das *Modell zufälliger Rechtszensierung* (random censorship model) definiert (siehe z.B. MILLER, 1981, S. 5ff).

Im Hinblick auf praktische Anwendungen des Modells bedeutet Bedingung (4.1) eine starke Einschränkung, da häufig die oben geforderte Unabhängigkeit nicht gegeben ist. Verlassen z.B. Patienten eine klinische Studie vorzeitig, weil sie auf Grund ihres kritischen Gesundheitszustandes in eine Spezialklinik eingeliefert werden müssen, so sind Überlebenszeit und zensierte Zeit nicht stochastisch unabhängig voneinander.

Weitere Definitionen werden an dieser Stelle eingeführt. Es sei $x \geq 0$.

- Mit $S(x) := 1 - F(x) = P(X > x)$ ist die Überlebensfunktion definiert.
- Die Hazardfunktion ist definiert als $\lambda(x) := \frac{f(x)}{S(x)}$ und charakterisiert das Risiko einer Untersuchungseinheit unmittelbar auszufallen, wenn sie bis zum Zeitpunkt x noch nicht ausgefallen ist.
- Die kumulierten Hazardfunktionen zu den Verteilungsfunktionen F und G sind definiert als $\Lambda_F := -\ln(1 - F(x))$ bzw. $\Lambda_G := -\ln(1 - G(x))$

(vgl. MILLER, 1981, S. 2f und CSÖRGŐ, 1988, S. 438).

Folgende Beziehungen gelten im Modell zufälliger Rechtszensierung.

Lemma 4.1 *Im Modell zufälliger Rechtszensierung gilt*

(i) für alle $x \geq 0$: $1 - H(x) = (1 - F(x))(1 - G(x))$
*(*LIO, PADGETT, *1987, S. 3304);*

(ii) für den erwarteten Anteil s *von unzensierten Beobachtungen:*

$$s := P(\Delta = 1) = \int_0^\infty (1 - G(x))\, dF(x), \quad x \geq 0$$

*(*CSÖRGŐ, *1988, S. 437).*

Beweis:

zu (i): Die Behauptung folgt aus der stochastischen Unabhängigkeit der Zufallsvariablen X und Y und da gilt: $1 - H(x) = P\big(\min\{X,Y\} > x\big)$, $x \geq 0$.

zu (ii): Mit der stochastischen Unabhängigkeit der Zufallsvariablen X und Y gilt:

$$\begin{aligned} s &= P(\Delta = 1) = P(X \leq Y) \\ &= \int_0^\infty \int_x^\infty f_{(X,Y)}(x,y)\, dy\, dx = \int_0^\infty \int_x^\infty f(x)g(y)\, dy\, dx \\ &= \int_0^\infty \int_x^\infty g(y)\, dy\, dF(x) = \int_0^\infty \underbrace{(1 - G(x))}_{P(Y>X)}\, dF(x), \end{aligned}$$

wobei $f_{(X,Y)}$ die gemeinsame Dichte von X und Y bezeichnet.

■

Der Kaplan–Meier–Schätzer

Der Maximum–Likelihood–Schätzer für die Verteilungsfunktion F im Modell zufälliger Rechtszensierung ist der in der Literatur oft verwendete Kaplan–Meier–Schätzer $\hat{F}_n$ (KAPLAN, MEIER, 1958). Er ist das Analogon zur empirischen Verteilungsfunktion F_n für das Modell zufälliger Rechtszensierung und ist wie folgt definiert.

Definition 4.2 *Für $n \in \mathbb{N}$, $x \geq 0$ und $i = 1, \ldots, n-1$ ist der Kaplan–Meier–Schätzer (Product–Limit–Schätzer) für F im Modell zufälliger Rechtszensierung definiert als*

$$\hat{F}_n(x) := \begin{cases} 0 & \text{für} \quad 0 \leq x < Z_{1:n} \\ 1 - \prod\limits_{j=1}^{i} \left(\frac{n-j}{n-j+1}\right)^{\Delta_{j:n}} & \text{für} \quad Z_{i:n} \leq x < Z_{i+1:n} \\ 1 & \text{für} \quad x \geq Z_{n:n}, \end{cases}$$

wobei $(Z_{1:n}, \Delta_{1:n}), \ldots, (Z_{n:n}, \Delta_{n:n})$ die nach den Zufallsvariablen $Z_1, \ldots, Z_n$ geordneten Zufallsvektoren mit zugehörigen Δ_i bezeichnen (siehe z.B. LIO, PADGETT, 1987, S. 3304).

Bemerkung 4.3 *Im Gegensatz zur empirischen Verteilungsfunktion F_n werden beim Kaplan–Meier–Schätzer $\hat{F}_n$ nicht allen Beobachtungen positive Gewichte zugeordnet. Nur bei den unzensierten Beobachtungen weist der Kaplan–Meier–Schätzer Sprungstellen auf. Die Sprünge können wie bei der empirischen Verteilungsfunktion von ungleicher Höhe sein (*LIO, PADGETT, *1987, S. 3303).*

Ausgehend von der alternativen Definition des KL–Schätzers, basierend auf der empirischen Quantilfunktion im unzensierten Fall (Kapitel 3.4), wird im folgenden der KL–Schätzer auf das Modell zufälliger Rechtszensierung übertragen.

4.2.2 Der Kaigh–Lachenbruch–Schätzer basierend auf dem Kaplan–Meier–Schätzer

LIO und PADGETT (1987) übertragen den KL–Schätzer auf das Modell zufälliger Rechtszensierung, indem sie in Definition 3.33 die empirische Quantilfunktion durch die Quantilfunktion des Kaplan–Meier–Schätzers ersetzen.

Definition 4.4 *Für $n \in \mathbb{N}$, $p \in (0,1)$ und für $k \in \mathbb{N}$ mit $1 \leq k \leq n$ ist mit*

$$\overset{Kap}{KL}(p,k,n) := \frac{1}{\binom{n}{k}} \sum_{i=1}^{\binom{n}{k}} \hat{Q}_k{}^i(p)$$

*der KL–Schätzer basierend auf dem Kaplan–Meier–Schätzer definiert, wobei $i = 1, \ldots, \binom{n}{k}$ und $\hat{Q}_k{}^i(p)$ die Quantilfunktion zur Kaplan–Meier–Schätzfunktion an der Stelle p in der i–ten Unterstichprobe vom Umfang k ist (*LIO, PADGETT, *1987, S. 3304).*

Da die Kaplan–Meier–Schätzfunktion nur Sprünge bei unzensierten Beobachtungen aufweist (Bemerkung 4.3), ist es nicht möglich, $\hat{Q}_n$ in Intervallabschnitten durch Ordnungsstatistiken darzustellen (vgl. LIO, PADGETT, 1987, S. 3303), weil die Intervallgrenzen davon abhängig wären, ob die entsprechenen Beobachtungen zensiert sind oder nicht. So kann der KL–Schätzer im Modell zufälliger Rechtszensierung nicht durch eine Linearkombination von Ordnungsstatistiken formuliert werden (LIO, PADGETT, 1987, S. 3305) und gehört daher nicht zur Klasse der L–Statistiken.

Unter einigen nicht–restriktiven Bedingungen ist der KL–Schätzer im Modell zufälliger Rechtszensierung eine U–Statistik mit Kern $\hat{Q}_k(p)$, der symmetrisch ist bezüglich $(Z_1, \Delta_1), \ldots, (Z_k, \Delta_k)$ (LIO, PADGETT, 1987, S. 3302f und S. 3305).

Analog zu den Korollaren 3.6 und 3.11 ergeben sich daher folgende Ausdrücke für Erwartungswert und Varianz des KL–Schätzers in diesem Modell:

$$\begin{aligned} \mathrm{E}_F\left(\overset{Kap}{KL}\,(p,k,n)\right) &= \mathrm{E}_F\left(h((Z_1,\Delta_1),\ldots,(Z_k,\Delta_k))\right) \\ &= \mathrm{E}_F(\hat{Q}_k(p)) \end{aligned}$$

und

$$\mathrm{Var}_F\left(\overset{Kap}{KL}\,(p,k,n)\right) = \frac{1}{\binom{n}{k}} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c} \cdot \zeta_c,$$

wobei $\zeta_c = \mathrm{Var}_F(\mathrm{E}_F(\hat{Q}_k(p)|(X_1,\ldots,X_c)))$, $p \in (0,1)$.

Als U–Statistik kann für den KL–Schätzer im Modell zufälliger Rechtszensierung auch die asymptotische Normalität für $n \to \infty$ und k fest gezeigt werden (LIO, PADGETT, 1987, S. 3306, Theorem 3.1). Da der KL–Schätzer hier keine L–Statistik ist, ist der Ausdruck für die Varianz der approximierten Normalverteilung im unzensierten Fall (in Korollar 3.21) jedoch nicht auf dieses Modell übertragbar. Zur Bestimmung des optimalen Unterstichprobenumfangs k schlagen LIO und PADGETT (1987) die Untersuchung des asymptotischen mittleren quadratischen Fehlers (mean integrated squared error) vor (LIO, PADGETT, 1987, S. 3306).

Um den Schwerpunkt auf die nun folgende Situation im Koziol–Green–Modell zu legen, wird der KL–Schätzer im Modell zufälliger Rechtszensierung nicht detaillierter diskutiert.

4.3 Der Kaigh–Lachenbruch–Schätzer im Koziol–Green–Modell

Nach der Einführung des Koziol–Green–Modells in 4.3.1 werden Unklarheiten zur Definition des ACL–Schätzers in CSÖRGŐ (1988) aufgezeigt und als Konsequenz daraus eine neue Definition des ACL–Schätzers entwickelt. Es folgt der Beweis für die Darstellung der ACL–Quantilfunktion mit Ordnungsstatistiken. Unter Verwendung der neuen Definition des ACL–Schätzers wird anschließend in 4.3.2 der

KL-Schätzer im Koziol-Green-Modell definiert. Die Zugehörigkeit zur Klasse der L- bzw. U-Statistiken wird diskutiert und die Untersuchung seiner Eigenschaften motiviert.

4.3.1 Das Koziol-Green-Modell

Das Koziol-Green-Modell (proportionales Hazards-Modell) ist ein Submodell des Modells zufälliger Rechtszensierung (Kapitel 4.2.1). Es wird definiert durch die zusätzliche Forderung:

$$1 - G(x) = (1 - F(x))^{\alpha}, \quad x \geq 0, \tag{4.2}$$

wobei $\alpha > 0$ eine feste Konstante ist (CSÖRGŐ, 1988, S. 438).

Bemerkung 4.5 *Das Koziol-Green-Modell, das auch als proportionales Hazards-Modell bezeichnet wird, ist nicht zu verwechseln mit dem proportionalen Hazards-Modell von* COX *(1972), welches die Einflüsse von Kovariablen berücksichtigt.*

Benannt ist das Modell nach J.A. Koziol und S.B. Green, die 1976 unter obiger Modellannahme eine Cramér-von Mises Statistik (Teststatistik für Tests auf Güte der Anpassung) basierend auf dem Kaplan-Meier-Schätzer untersuchen. KOZIOL und GREEN (1976) führen das Modell ein, weil der Kaplan-Meier-Schätzer unter der betreffenden Bedingung konsistent ist, und der daraus hergeleitete stochastische Prozess schwach gegen einen Gauss-Prozess konvergiert (BRESLOW, CROWLEY, 1974 und CSÖRGŐ, HORVÁTH, 1981 mit einer Korrektur des Beweises).

Auch vor 1976 werden schon die Annahmen des Koziol-Green-Modells oder Spezialfälle davon für Probleme mit zensierten Daten benutzt. Zum Beispiel verwendet EFRON (1967, S. 848ff) einen Fall des Koziol-Green-Modells, indem er sowohl die unzensierten als auch die zensierten Größen als exponentialverteilt annimmt, um damit Tests im Zweistichproben-Problem zu vergleichen.

Nach 1976 wird das Koziol-Green-Modell namentlich bei HOLLANDER und PROSCHAN (1979) erwähnt, die verschiedene Anpassungstests für zensierte Daten vergleichen, wie auch bei CSÖRGŐ und HORVÁTH (1981), die im Koziol-Green-Modell

unter anderem Konfidenzbänder für die Verteilungsfunktion der Zensierungen herleiten. In CHENG, HOLLANDER und LANGBERG (1982) werden unter anderem exakte Ausdrücke für Momente des Kaplan–Meier–Schätzers im Koziol–Green–Modell hergeleitet. Zu den jüngeren Veröffentlichungen gehören HERBST (1992), der einen Schätzer für $E_F(X^r)$ im Koziol–Green–Modell einführt und untersucht und STUTE (1992), welcher die starke Konsistenz von Schätzern für $\int \phi\, dF$ (ϕ eine F–integrierbare Funktion) zeigt.

Die aufgeführte Literatur zum Koziol–Green–Modell ist unvollständig, soll aber auf das Interesse an diesem Modell in der Forschung hinweisen.

Nachfolgende Beziehungen gelten im Koziol–Green–Modell.

Lemma 4.6 *Im Koziol–Green–Modell gilt*

(i) für die Wahrscheinlichkeit s, daß sich die Zufallsvariable Z zu einer unzensierten Beobachtung realisiert:

$$s = P(\Delta = 1) = \frac{1}{1+\alpha} > 0;$$

(ii) für alle $x \geq 0$:

$$1 - F(x) = \big(1 - H(x)\big)^s, \quad s > 0 \tag{4.3}$$

(CSÖRGŐ, 1988, S. 438).

Beweis:

zu (i): Die Behauptung folgt aus Lemma 4.1(ii) und der Modellvoraussetzung (4.2).

zu (ii): Die Behauptung folgt mit Einsetzen der Modellvoraussetzung (4.2) in die Formel von Lemma 4.1(i).

■

Bemerkung 4.7

*(i) Mit größeren Werten von α steigt die erwartete Anzahl von zensierten Beobachtungen in einer Stichprobe (*KOZIOL, GREEN, *1976, S. 466).*

*(ii) Da α mit $P(\Delta = 0) = \frac{\alpha}{1+\alpha}$ den erwarteten Zensierungsanteil bestimmt, wird α auch als Zensierungsparameter bezeichnet (*KOZIOL, GREEN, *1976, S. 466).*

(iii) Sei $S_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_i$ *der zufällige Anteil der unzensierten Beobachtungen in einer Stichprobe vom Umfang* n, $n \in \mathbb{N}$. *Dann ist* S_n *ein stark konsistenter Schätzer für* $\frac{1}{1+\alpha}$ *(*CSÖRGŐ *und* HORVÁTH, *1981, S. 395).*

Das nachfolgende Korollar beschreibt eine Charakterisierung des Koziol–Green–Modells.

Korollar 4.8 *Im Modell zufälliger Rechtszensierung existiert ein* $\alpha > 0$, *so daß die Bedingung (4.2) erfüllt ist, genau dann, wenn die Zufallsvariablen* Z *und* Δ *stochastisch unabhängig sind (siehe z.B.* CSÖRGŐ, *1988, S. 438).*

Ebenso wie Bedingung (4.1) für das Modell zufälliger Rechtszensierung impliziert Korollar 4.8 eine starke Einschränkung für die praktische Anwendung des Koziol–Green–Modells. Häufig ist z.B. bei klinischen Studien erst im späteren Verlauf der Untersuchung mit Abgängen zu rechnen, die zu zensierten Beobachtungen führen. So ist hier die Tatsache, ob eine Beobachtung zensiert ist oder nicht, abhängig von den tatsächlich beobachteten Überlebenszeiten, und die in Korollar 4.8 geforderte Beziehung trifft nicht zu.

Weitere Eigenschaften des Koziol–Green–Modells sind in folgender Bemerkung aufgeführt.

Bemerkung 4.9
Im Koziol–Green–Modell gilt:

(i) $\sup\{x : F(x) < 1\} = \sup\{x : G(x) < 1\}$ *für* $x \geq 0$ *(*CHENG, HOLLANDER *und* LANGBERG, *1982, S. 142).*

(ii) Die kumulierten Hazardfunktionen der zensierten und unzensierten Größen sind proportional zueinander, d.h. es gilt: $\Lambda_G(x) = \alpha\Lambda_F(x)$ *(*CSÖRGŐ, *1988, S. 438). Daraus erklärt sich die alternative Bezeichnung „Proportionales Hazards–Modell".*

Das nachstehende Beispiel beschreibt eine Datensituation, die den Modellannahmen des Koziol–Green–Modells entspricht.

Beispiel 4.10 *Die unzensierten Größen sind wie eine Zufallsvariable X verteilt mit $X \sim \text{Exp}(\lambda)$, $\lambda > 0$. Werden die Zensierungen durch die Zufallsvariable Y repräsentiert mit $Y \sim \text{Exp}(\alpha\lambda)$, $\alpha, \lambda > 0$, dann ist die Modellvoraussetzung (4.2) erfüllt.*

Der ACL–Schätzer

Unter Verwendung von (4.3) mit Übergang zur empirischen Verteilungsfunktion formulieren ABDUSHUKUROV (1984) und CHENG, LIN (1984) unabhängig voneinander den nach ihren Initialen benannten ACL–Schätzer. Er liefert als Maximum–Likelihood–Schätzer für die unbekannte Verteilungsfunktion F das Analogon zur empirischen Verteilungsfunktion unter dem Koziol–Green–Modell (CSÖRGŐ, 1988, der die bekannten Eigenschaften des ACL–Schätzers zusammenfassend beschreibt).

In Beziehung (4.3) wird $s > 0$ vorausgesetzt; bei Übergang zur empirischen Verteilungsfunktion ist jedoch $P(S_n = 0) > 0$ mit $S_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_i$. Genaugenommen lassen sich aus CSÖRGŐ (1988) zwei unterschiedliche Definitionen des ACL–Schätzers entnehmen, wobei für die erste Variante der Definition die obige Tatsache nicht berücksichtigt wurde.

Definition 4.11 *(1. Variante) Für $n \in \mathbb{N}$ und $x \geq 0$ ist der ACL–Schätzer für F im Koziol–Green–Modell definiert als*

$$\tilde{F}_n^{(1)}(x) := \begin{cases} 1 - \left(1 - H_n(x)\right)^{S_n} & \text{für } (H_n(x), S_n) \in \{0, \frac{1}{n}, \ldots, 1\} \times \{\frac{1}{n}, \frac{2}{n}, \ldots, 1\} \\ \text{nicht definiert} & \text{für } (H_n(x), S_n) \in \{0, \frac{1}{n}, \ldots, 1\} \times \{0\}, \end{cases}$$

*wobei H_n die empirische Verteilungsfunktion zu H und $S_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_i$ der zufällige Anteil der unzensierten Beobachtungen in einer Stichprobe vom Umfang n ist (*CSÖRGŐ*, 1988, S. 439, Formel (1.5)).*

Definition 4.12 *(2. Variante) Für $n \in \mathbb{N}$ und $x \geq 0$ ist der ACL–Schätzer für F im Koziol–Green–Modell definiert als*

$$\tilde{F}_n^{(2)}(x) := \begin{cases} 1 - \left(1 - H_n(x)\right)^{S_n}, & (H_n(x), S_n) \in \{0, \frac{1}{n}, \ldots, 1\} \times \{\frac{1}{n}, \frac{2}{n}, \ldots, 1\} \\ 0, & (H_n(x), S_n) \in \{0, \frac{1}{n}, \ldots, 1\} \times \{0\},\ x < Z_{n:n} \\ 1, & (H_n(x), S_n) \in \{0, \frac{1}{n}, \ldots, 1\} \times \{0\},\ x \geq Z_{n:n} \end{cases}$$

*mit H_n und S_n wie in Definition 4.11 (*CSÖRGŐ*, 1988, S. 440).*

Für $S_n = 0$ ist der ACL–Schätzer in CSÖRGŐ (1988) entweder nicht definiert (1. Variante) oder die Information in den zensierten Daten wird bei der Schätzung ignoriert (2. Variante), weil auf eine Einpunktverteilung in $Z_{n:n}$ geschlossen wird.

Daher wird hier nach der Idee von WITTING und NÖLLE (1970, S. 76f) eine erweiterte Definition des ACL–Schätzers eingeführt, welche die Definition des ACL–Schätzers über dem gesamten Stichprobenraum sicherstellt und die Information in den Daten auch für $S_n = 0$ berücksichtigt. WITTING und NÖLLE (1970, S. 77, Definition 2.31) setzen für eine erweiterte Definition des Maximum–Likelihood–Schätzers folgendes voraus: Die Wahrscheinlichkeit, daß sich eine Folge von Zufallsvariablen $(X_n)_{n\in\mathbb{N}}$ in der Menge realisiert, für die der konventionelle Maximum–Likelihood–Schätzer nicht definiert werden kann, konvergiert für $n \to \infty$ gegen Null.

Definition 4.13 *Für $n \in \mathbb{N}$ und $x \geq 0$ ist der erweiterte ACL–Schätzer für F im Koziol–Green–Modell definiert als*

$$\tilde{F}_n(x) := \begin{cases} 1 - \big(1 - H_n(x)\big)^{S_n} & \text{für } (H_n(x), S_n) \in \{0, \frac{1}{n}, \ldots, 1\} \times \{\frac{1}{n}, \frac{2}{n}, \ldots, 1\} \\ 1 - \big(1 - H_n(x)\big)^{\frac{1}{n+1}} & \text{für } (H_n(x), S_n) \in \{0, \frac{1}{n}, \ldots, 1\} \times \{0\}, \end{cases}$$

wobei H_n die empirische Verteilungsfunktion zu H und $S_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_i$ der zufällige Anteil der unzensierten Beobachtungen in einer Stichprobe vom Umfang n ist.

Die Definition 4.13 ist sinnvoll, da analog zur Voraussetzung von Definition 2.31 in WITTING, NÖLLE (1970, S. 77) der folgende Satz gilt.

Satz 4.14 *Unter dem Koziol–Green–Modell gilt für $n \in \mathbb{N}$ und $x \geq 0$:*

$$P\left((H_n(x), S_n) \in \left\{0, \frac{1}{n}, \ldots, 1\right\} \times \{0\}\right) \longrightarrow 0 \quad \text{für} \quad n \to \infty.$$

Beweis: Sei $C_n := \sum_{i=1}^{n} \Delta_i$ die zufällige Anzahl der unzensierten Beobachtungen in der Stichprobe vom Umfang n. Es gilt:

$$\begin{aligned} P\left((H_n(x), S_n) \in \left\{0, \frac{1}{n}, \ldots, 1\right\} \times \{0\}\right) &= P(S_n = 0) \\ &= P(C_n = 0) \\ &= (1 - s)^n, \end{aligned} \tag{4.4}$$

da mit

$$\Delta_i \sim \text{Bin}(1,s),\ i=1,\ldots,n, \quad \text{gilt:} \quad C_n = \sum_{i=1}^{n} \Delta_i \sim \text{Bin}(n,s),$$

wobei $s = P(\Delta = 1) = \frac{1}{1+\alpha}$ (Bemerkung 4.6 (i)).

Mit $0 < s \leq 1$ ist $0 \leq (1-s) < 1$ und daher $(1-s)^n \longrightarrow 0$ für $n \to \infty$.

■

Bemerkung 4.15

(i) In der erweiterten Definition des ACL–Schätzers (Definition 4.13) wird die Folge $\frac{n}{n}, \frac{n-1}{n}, \ldots, \frac{1}{n}$ *der möglichen Realisationen von* S_n *durch den Term* $\frac{1}{n+1}$ *nach unten fortgesetzt, um den Wert Null in der Potenz zu vermeiden. Ebenso wären andere Fortsetzungen aus dem Intervall* $(0, \frac{1}{n})$ *denkbar, wie z.B.* $\frac{1}{2n}$.

(ii) Die Versionen der ACL–Schätzer–Definitionen nach CSÖRGŐ *(1988) (Definition 4.11 und 4.12) und der erweiterte ACL–Schätzer aus Definition 4.13 sind für* $n \to \infty$ *asymptotisch äquivalent (nach* LEHMANN, *1983, S. 40) in dem Sinne, daß für* $n \in \mathbb{N}$ *und* $x \geq 0$ *gilt:*

$$\tilde{F}_n(x) = \tilde{F}_n^{(1)}(x) = \tilde{F}_n^{(2)}(x) \quad \text{für} \quad (H_n(x), S_n) \in \{0, \tfrac{1}{n}, \ldots, 1\} \times \{\tfrac{1}{n}, \ldots, 1\}$$

und

$$\tilde{F}_n(x) \neq \tilde{F}_n^{(1)}(x) \neq \tilde{F}_n^{(2)}(x) \quad \text{für} \quad (H_n(x), S_n) \in \{0, \tfrac{1}{n}, \ldots, 1\} \times \{0\}$$

mit $P\left((H_n(x), S_n) \in \{0, \frac{1}{n}, \ldots, 1\} \times \{0\}\right) \longrightarrow 0$ *für* $n \to \infty$.

(iii) Im Gegensatz zum Kaplan–Meier–Schätzer liegt beim erweiterten ACL–Schätzer $\tilde{F}_n$ *bei jeder Beobachtung (zensiert oder unzensiert) eine Sprungstelle vor. Die Sprunghöhen sind dabei zufällig (nach* CSÖRGŐ, *1988).*

Im folgenden wird nur noch der erweiterte ACL–Schätzer aus Definition 4.13 betrachtet, und dieser wird der Einfachheit halber als der ACL–Schätzer bezeichnet.

Die ACL–Quantilfunktion

Mit Bemerkung 4.15 (iii) ist es möglich, die ACL–Quantilfunktion zu $\tilde{F}_n$ ähnlich wie die empirische Quantilfunktion im unzensierten Fall (vgl. Kapitel 2.1, Darstellung (2.3)) mit Hilfe von Ordnungsstatistiken darzustellen. Es gilt das folgende Korollar.

Korollar 4.16 *Für die ACL–Quantilfunktion* $\tilde{Q}_n : [0,1] \longrightarrow \mathbb{R}$, $n \in \mathbb{N}$, *gilt:*

$$\tilde{Q}_n(p) = \begin{cases} Z_{1:n} & \text{für } p = 0 \\ Z_{R:n} & \text{für } p \in (0,1] \end{cases}$$

mit

$$R := \begin{cases} \left]n\left(1-(1-p)^{\frac{1}{S_n}}\right)\right[+ 1 & \text{für } S_n \in \{\frac{1}{n}, \ldots, 1\} \\ \left]n\left(1-(1-p)^{n+1}\right)\right[+ 1 & \text{für } S_n = 0. \end{cases} \tag{4.5}$$

Dabei ist $]x[= \max\{z \in \mathbb{N}_0 | z < x\}$ *und* $S_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_i$ *der zufällige Anteil der unzensierten Beobachtungen in der Stichprobe vom Umfang* n.

Beweis: Setze

$$D_n = \begin{cases} S_n & \text{für } S_n \in \{\frac{1}{n}, \ldots, 1\} \\ \frac{1}{n+1} & \text{für } S_n = 0. \end{cases}$$

Nach Definition 4.13 und CSÖRGŐ (1988, S. 445) gilt:

$$\tilde{Q}_n(p) = \begin{cases} Z_{1:n} & \text{für } p = 0 \\ Z_{j:n} & \text{für } 1 - \left(1 - \frac{j-1}{n}\right)^{D_n} < p \leq 1 - \left(1 - \frac{j}{n}\right)^{D_n}, \end{cases} \tag{4.6}$$

$j \in \{1, \ldots, n\}$.

Die Bedingung für $Z_{j:n}$ in (4.6) fordert, daß p in einem halboffenen Intervall mit von j abhängigen Intervallgrenzen liegt. Durch äquivalente Umformungen dieser Ungleichung ergibt sich im folgenden eine äquivalente Bedingung für j:

$$\begin{aligned} & 1 - \left(1 - \tfrac{j-1}{n}\right)^{D_n} < p \leq 1 - \left(1 - \tfrac{j}{n}\right)^{D_n} \\ \iff \quad & \left(1 - \tfrac{j}{n}\right)^{D_n} \leq 1 - p < \left(1 - \tfrac{j-1}{n}\right)^{D_n} \\ \iff \quad & 1 - \tfrac{j}{n} \leq (1-p)^{\frac{1}{D_n}} < 1 - \tfrac{j-1}{n} \\ \iff \quad & j < n\left(1 - (1-p)^{\frac{1}{D_n}}\right) + 1 \quad \wedge \quad n\left(1 - (1-p)^{\frac{1}{D_n}}\right) \leq j \\ \iff \quad & n\left(1 - (1-p)^{\frac{1}{D_n}}\right) \leq j < n\left(1 - (1-p)^{\frac{1}{D_n}}\right) + 1. \end{aligned}$$

Daraus folgt

$$\tilde{Q}_n(p) = \begin{cases} Z_{1:n} & \text{für } p = 0 \\ Z_{j:n} & \text{für } n\left(1 - (1-p)^{\frac{1}{D_n}}\right) \leq j < n\left(1 - (1-p)^{\frac{1}{D_n}}\right) + 1, \end{cases} \tag{4.7}$$

$j \in \{1, \ldots, n\}$, und damit die Behauptung (vgl. Vorgehen im Anhang A.1).

■

Bemerkung 4.17 *Im Unterschied zur empirischen Verteilungsfunktion im unzensierten Fall ist der Index R gemäß (4.5) der Ordnungsstatistik $Z_{R:n}$, $n \in \mathbb{N}$, bei der Darstellung der ACL–Quantilfunktion eine zufällige Größe, da S_n zufällig ist.*

Im nächsten Unterabschnitt wird ausgehend von Definition 3.33 der KL–Schätzer im Koziol–Green–Modell definiert. Dieses erfolgt analog zum Vorgehen für das Modell zufälliger Rechtszensierung (vgl. Kapitel 4.2.2).

4.3.2 Der Kaigh–Lachenbruch–Schätzer basierend auf dem ACL–Schätzer

Der KL–Schätzer kann auf das Koziol–Green–Modell übertragen werden, indem die empirische Quantilfunktion Q_n in Definition 3.33 durch die Quantilfunktion des ACL–Schätzers $\tilde{Q}_n$ ersetzt wird.

Definition 4.18 *Sei $n_0 \in \mathbb{N}$ und $d \in \{0, 1, \ldots, n_0 - 1\}$ fest. Für festes $n \in \mathbb{N}$, $n \geq n_0$, $p \in (0,1)$ und für ein $k_n \in \mathbb{N}$ mit $k_n := n - d$ ist mit*

$$\overset{ACL}{KL}(p, k_n, n) := \frac{1}{\binom{n}{k_n}} \sum_{i=1}^{\binom{n}{k_n}} \tilde{Q}^i_{k_n}(p)$$

der KL–Schätzer basierend auf dem ACL–Schätzer im Koziol–Green–Modell definiert. Dabei ist $\tilde{Q}^i_{k_n}(p)$ die Quantilfunktion an der Stelle p zur ACL–Schätzfunktion aus Definition 4.13 in der i–ten Unterstichprobe vom Umfang k_n.

Gemäß WITTING und NÖLLE (1970, S. 76f) ist obige Definition des KL–Schätzers im Koziol–Green–Modell sinnvoll, weil die Wahrscheinlichkeit zur Realisierung der Menge M_0, bei der auf die Erweiterung der ACL–Schätzer–Definition zurückgegriffen werden muß, für $n \to \infty$ gegen Null konvergiert. Dabei beinhaltet M_0 alle Realisationen von C_n (Anzahl der unzensierten Beobachtungen in einer Stichprobe vom Umfang n), welche die Existenz von komplett zensierten Unterstichproben mit Umfang k_n ermöglichen. Es gilt der folgende Satz.

Satz 4.19 *Sei $n_0 \in \mathbb{N}$ und $d \in \{0, 1, \ldots, n_0 - 1\}$ fest. Für $n \in \mathbb{N}$, $n \geq n_0$, und $k_n := n - d$ gilt:*

$$P(C_n \in M_0) \longrightarrow 0 \quad \textit{für} \quad n \to \infty,$$

mit $M_0 = \{0, 1, \ldots, n - k_n\}$. Dabei ist C_n die zufällige Anzahl unzensierter Beobachtungen in der Stichprobe vom Umfang n.

Beweis:
Mit $C_n \sim \text{Bin}(n, s)$ und $k_n = n - d$ gilt:

$$\begin{aligned}
P(C_n \in M_0) &= P(C_n \leq n - k_n) \\
&= \sum_{j=0}^{n-k_n} \binom{n}{j} s^j (1-s)^{n-j} \\
&= \sum_{j=0}^{d} \binom{n}{j} s^j (1-s)^{n-j} \\
&= (1-s)^n + n \cdot s(1-s)^{n-1} + \sum_{j=2}^{d} \frac{(n-j+1) \cdot \ldots \cdot n}{j!} s^j (1-s)^{n-j} \\
&\leq \underbrace{(1-s)^n}_{\to 0, \text{ vgl. (4.4)}} + \underbrace{n \cdot s(1-s)^{n-1}}_{\to 0} + \underbrace{\sum_{j=2}^{d} \frac{n^j}{j!} \cdot s^j (1-s)^{n-j}}_{\to 0} \\
&\longrightarrow 0 \quad \text{für } n \to \infty,
\end{aligned}$$

da mit $(1-s)^{n-i} = o\left(\frac{1}{n^i}\right)$, $n \to \infty$, für $i = 1, \ldots, d$, $0 < s < 1$, eine Summe mit fester Anzahl von Summanden vorliegt, die jeweils gegen Null konvergieren. Für $s = 1$ ergibt sich obige Summe zu Null.

Es folgt: $P(C_n \leq n - k_n) \longrightarrow 0 \quad \text{für} \quad n \to \infty$.

■

Mit der Darstellung der ACL–Quantilfunktion durch Ordnungsstatistiken (Korollar 4.16) läßt sich im Koziol–Green-Modell ähnlich wie im unzensierten Fall eine Formulierung des KL–Schätzers durch Ordnungsstatistiken finden. Dabei ist zu beachten, daß der Unterstichprobenumfang in kritischer Weise von n abhängt. Es gilt folgender Satz.

Satz 4.20 *Sei $n_0 \in \mathbb{N}$ und $d \in \{0, 1, \ldots, n_0 - 1\}$ fest. Für $n \in \mathbb{N}$, $n \geq n_0$, $p \in (0,1)$ und für ein $k_n \in \mathbb{N}$ mit $k_n := n - d$ gilt:*

$$\overset{ACL}{KL}(p, k_n, n) = \frac{1}{\binom{n}{k_n}} \sum_{i=1}^{\binom{n}{k_n}} Z^i_{R_i:k_n} \tag{4.8}$$

$$= \sum_{j=1}^{n} l_j(n, k_n, \Delta_{1:n}, \ldots, \Delta_{n:n}) \cdot Z_{j:n}, \tag{4.9}$$

wobei

$$l_j(n, k_n, \Delta_{1:n}, \ldots, \Delta_{n:n}) := \sum_{x=\max\{0, C_n - n + k_n\}}^{\min\{C_n, k_n\}} \sum_{a=\max\{0, x-1-k_n+r_x\}}^{\min\{x, r_x - 1\}} \sum_{b=\max\{0, x-a-k_n+r_x\}}^{\min\{1, x, x-a\}}$$
$$\frac{1}{\binom{n}{k_n}} \cdot \binom{\sum_{i=1}^{j-1} \Delta_{i:n}}{a} \binom{\sum_{i=1}^{j-1}(1 - \Delta_{i:n})}{r_x - 1 - a} \binom{\Delta_{j:n}}{b} \binom{1 - \Delta_{j:n}}{1 - b}$$
$$\binom{\sum_{i=j+1}^{n} \Delta_{i:n}}{x - a - b} \binom{\sum_{i=j+1}^{n}(1 - \Delta_{i:n})}{k_n + a + b - x - r_x},$$

$$R_i := \begin{cases} \left]k_n\left(1 - (1-p)^{(S^i_{k_n})^{-1}}\right)\right[+ 1 & \text{für } S^i_{k_n} \in \{\frac{1}{k_n}, \ldots, 1\} \\ \left]k_n\left(1 - (1-p)^{k_n+1}\right)\right[+ 1 & \text{für } S^i_{k_n} = 0 \end{cases} \tag{4.10}$$

und

$$r_x := \begin{cases} \left]k_n\left(1 - (1-p)^{\frac{k_n}{x}}\right)\right[+ 1 & \text{für } x > 0 \\ \left]k_n\left(1 - (1-p)^{k_n+1}\right)\right[+ 1 & \text{für } x = 0. \end{cases}$$

Dabei ist $Z^i_{R_i:k_n}$ die R_i–te Ordnungsstatistik in der i–ten Unterstichprobe vom Umfang k_n, $S^i_{k_n} = \frac{1}{k_n} \sum_{j=1}^{k_n} \Delta^i_{j:k_n}$ der zufällige Anteil der unzensierten Beobachtungen in der i–ten Unterstichprobe vom Umfang k_n und $C_n = \sum_{i=1}^{n} \Delta_{i:n}$ die zufällige Anzahl der unzensierten Beobachtungen in der Stichprobe vom Umfang n.

Beweis:

(i)

Mit der Definition des KL–Schätzers, basierend auf dem ACL–Schätzer im Koziol–Green–Modell (Definition 4.18) und der Darstellung der ACL–Quantilfunktion in Korollar 4.16, ergibt sich der Ausdruck (4.8), denn es gilt für die i–te Unterstichprobe, $i = 1, \ldots, \binom{n}{k_n}$, und für $p \in (0,1)$:

$$\tilde{Q}^i_{k_n}(p) = Z^i_{R_i:k_n}$$

mit R_i gemäß (4.10). Dabei ist $S^i_{k_n}$ als Anteil der unzensierten Beobachtungen in der i–ten Unterstichprobe vom Umfang k_n, $i = 1, \ldots, \binom{n}{k_n}$, eine zufällige Größe und damit auch R_i.

(ii)
Sei zunächst angenommen, daß sich $C_n = \sum_{i=1}^{n} \Delta_i$, die zufällige Anzahl unzensierter Beobachtungen in der Stichprobe vom Umfang n, zu c_n, $c_n \in \{0, 1, \ldots, n\}$, realisiert hat. Unter dieser Annahme bleibt der Anteil S_{k_n} unzensierter Beobachtungen in einer Unterstichprobe vom Umfang k_n zufällig. Die Formulierung gemäß (4.9) wird zuerst für beliebiges realisiertes $c_n \in \{0, 1, \ldots, n\}$ hergeleitet. Da diese Herleitung für alle c_n aus $\{0, 1, \ldots, n\}$ gültig ist, ist anschließend der Schluß auf (4.9) mit zufälligem C_n möglich.

Die Herleitung erfolgt in zwei Schritten. Ausgehend von (4.8) werden im ersten Schritt Summanden entsprechend der Realisationen von R_i, $i = 1, \ldots, \binom{n}{k_n}$, zusammengefaßt, und im zweiten Schritt wird mit Hilfe von kombinatorischen Überlegungen von Ordnungsstatistiken aus den Unterstichproben auf Ordnungsstatistiken aus der ursprünglichen Stichprobe vom Umfang n übergegangen.

• 1. Schritt: a) Unter der Bedingung $C_n = c_n$ ist der zufällige Anteil $S^i_{k_n}$ von unzensierten Beobachtungen in der i–ten Unterstichprobe vom Umfang k_n, welche ohne Zurücklegen und ohne Berücksichtigung der Anordnung gezogen wird, für alle i, $i = 1, \ldots, \binom{n}{k_n}$, identisch verteilt wie eine Zufallsvariable $\left(S_{k_n} \middle| C_n = c_n\right)$, und es gilt:

$$\left(k_n \cdot S_{k_n} \middle| C_n = c_n\right) \sim \mathrm{Hyp}(n, c_n, k_n).$$

Es folgt

$$\begin{aligned} P\left(k_n \cdot S_{k_n} = x \middle| C_n = c_n\right) &= P\Big(\sum_{i=1}^{k_n} \Delta_{i:k_n} = x \Big| C_n = c_n\Big) \\ &= P\left(\begin{array}{l}\text{Eine Unterstichprobe hat genau} \\ x \text{ unzensierte Beobachtungen}\end{array} \middle| C_n = c_n\right) \\ &= \frac{\binom{c_n}{x}\binom{n-c_n}{k_n-x}}{\binom{n}{k_n}}, \end{aligned}$$

wobei $x = \max\{0, c_n - n + k_n\}, \ldots, \min\{c_n, k_n\}$.

b) Für Unterstichproben mit der realisierten Anzahl von genau x unzensierten Beobachtungen realisiert sich die entsprechende Zufallsvariable $S^i_{k_n}$ zu $\frac{x}{k_n}$ und damit

R_i gemäß (4.10) zu

$$r_x = \begin{cases}]k_n(1-(1-p)^{\frac{k_n}{x}})[+ 1 & \text{für } x > 0 \\ ]k_n(1-(1-p)^{k_n+1})[+ 1 & \text{für } x = 0. \end{cases} \tag{4.11}$$

Mit Überlegungen a) und b) können demnach Unterstichproben mit gleicher realisierter Anzahl von unzensierten Beobachtungen in Ausdruck (4.8) wie folgt zusammengefaßt werden:

$$\left(\overset{ACL}{KL}\ (p,k_n,n)\middle|C_n=c_n\right) = \frac{1}{\binom{n}{k_n}} \cdot \sum_{x=\max\{0,c_n-n+k_n\}}^{\min\{c_n,k_n\}} \sum_{i=1}^{\binom{c_n}{x}\binom{n-c_n}{k_n-x}} Z^{x^i}_{r_x:k_n} \tag{4.12}$$

mit r_x gemäß (4.11). Mit $Z^{x^i}_{r_x:k_n}$ wird hierbei die r_x–te Ordnungsstatistik aus der i–ten Unterstichprobe vom Umfang k_n mit genau x unzensierten Beobachtungen bezeichnet.

• 2. Schritt: Der Ausdruck (4.12) läßt sich als gewichtete Summe der Ordnungsstatistiken $Z_{1:n},\ldots,Z_{n:n}$ formulieren. Es ist

$$\begin{aligned}
&\left(\overset{ACL}{KL}\ (p,k_n,n)\middle|C_n=c_n\right) \\
&= \frac{1}{\binom{n}{k_n}} \sum_{x=\max\{0,c_n-n+k_n\}}^{\min\{c_n,k_n\}} \sum_{j=1}^{n} \left|\left\{Z_{r_x:k_n} = Z_{j:n} \wedge \sum_{i=1}^{k_n}\Delta_{i:k_n} = x \,\middle|\, C_n = c_n\right\}\right| \cdot Z_{j:n} \\
&= \frac{1}{\binom{n}{k_n}} \sum_{x=\max\{0,c_n-n+k_n\}}^{\min\{c_n,k_n\}} \sum_{a=\max\{0,x-1-k_n+r_x)\}}^{\min\{x,r_x-1\}} \sum_{b=\max\{0,x-a-k_n+r_x\}}^{\min\{1,x,x-a\}} \sum_{j=1}^{n} \Big|\Big\{Z_{r_x:k_n} = Z_{j:n} \wedge \\
&\quad \wedge \sum_{i=1}^{k_n}\Delta_{i:k_n} = x \wedge \sum_{i=1}^{r_x-1}\Delta_{i:k_n} = a \wedge \Delta_{r_x:k_n} = b \,\Big|\, C_n = c_n\Big\}\Big| \cdot Z_{j:n}.
\end{aligned} \tag{4.13}$$

Dabei ergeben sich die Grenzen für die Laufindizes a und b durch die vorgegebenen Größen k_n, x und r_x mit den folgenden Bedingungen (vgl. Abbildung 4):
für a:

$$\begin{aligned} 0 \le a \le x \quad &\wedge \quad 0 \le a \le r_x - 1 \\ &\wedge \quad a = x - (b+c), \end{aligned} \tag{4.14}$$

wobei $(b+c)$ beliebig sind aus $\{0,1,\ldots,1+k_n-r_x\}$, d.h. (4.14) fordert

$$x - (1 + k_n - r_x) \le a \le x\,;$$

für b:

$$\begin{aligned} 0 \le b \le x \quad &\wedge \quad 0 \le b \le 1 \\ &\wedge \quad b = x - (a + c), \end{aligned} \tag{4.15}$$

wobei c beliebig ist aus $\{0, 1, \ldots, k_n - r_x\}$, d.h. (4.15) fordert

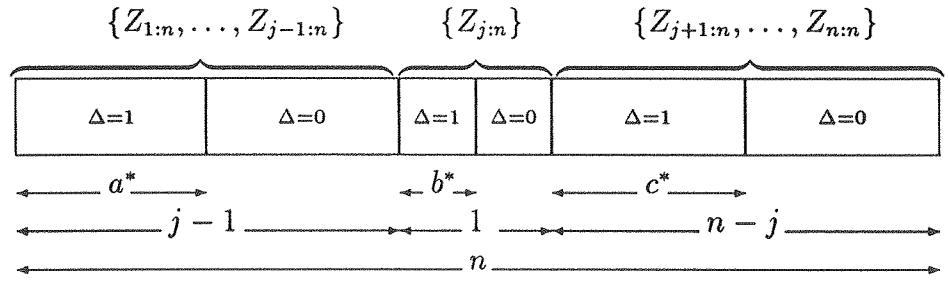
$$x - (a + k_n - r_x) \le b \le x - a.$$

Durch Äquivalenzüberlegungen ergeben sich daher für a und b die folgenden scharfen Grenzen:

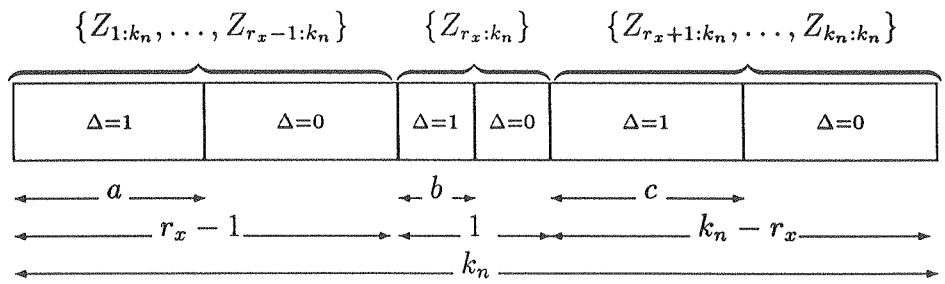
$$\begin{aligned} \max\{0, x - (1 + k_n - r_x)\} &\le a \le \min\{x, r_x - 1\} \\ \max\{0, x - (a + k_n - r_x)\} &\le b \le \min\{1, x, x - a\}. \end{aligned}$$

Abbildung 4 *Veranschaulichung der Strukturen in der Stichprobe vom Umfang n mit c_n unzensierten Beobachtungen und in einer Unterstichprobe vom Umfang k_n mit x unzensierten Beobachtungen.*

Stichprobe vom Umfang n mit $a^* = \sum_{i=1}^{j-1} \Delta_{i:n}$, $b^* = \Delta_{j:n}$ und $c^* = c_n - a^* - b^*$:



Unterstichprobe vom Umfang k_n mit $c = x - a - b$:



Mit kombinatorischen Überlegungen ergibt sich für den Ausdruck $\left|(.)\right|$ in (4.13) (vgl. Abbildung 4)

$$\begin{aligned}
&\left|\left\{Z_{r_x:k_n} = Z_{j:n} \wedge \sum_{i=1}^{k_n} \Delta_{i:k_n} = x \wedge \sum_{i=1}^{r_x-1} \Delta_{i:k_n} = a \wedge \Delta_{r_x:k_n} = b \,\middle|\, C_n = c_n\right\}\right| \\
= &\binom{\sum_{i=1}^{j-1} \Delta_{i:n}}{a}\binom{\sum_{i=1}^{j-1}(1-\Delta_{i:n})}{r_x-1-a}\binom{\Delta_{j:n}}{b}\binom{1-\Delta_{j:n}}{1-b}\binom{\sum_{i=j+1}^{n} \Delta_{i:n}}{x-a-b} \cdot \\
&\cdot\binom{\sum_{i=j+1}^{n}(1-\Delta_{i:n})}{(k_n-x)-(r_x-1-a)-(1-b)}.
\end{aligned}$$

Damit folgt für $\left(\overset{ACL}{KL}\,(p,k_n,n)|C_n = c_n\right)$ der Ausdruck gemäß (4.9) für beliebige Realisationen c_n aus $\{0,1,\ldots,n\}$. Mit Übergang zur Zufallsvariablen C_n ist für $\overset{ACL}{KL}\,(p,k_n,n)$ der Satz bewiesen.

■

Liegen keine Zensierungen vor, so läßt sich der KL–Schätzer im Koziol–Green–Modell weitgehend auf den KL–Schätzer aus Definition 3.33 (basierend auf der empirischen Quantilfunktion) im unzensierten Fall zurückführen. Ein Unterschied besteht nur in der im Koziol–Green–Modell geforderten Abhängigkeit des Unterstichprobenumfangs k von n. Es gilt folgendes Lemma.

Lemma 4.21 *Sei $n_0 \in \mathbb{N}$, $d \in \{0,1,\ldots,n_0-1\}$ fest und $s = P(\Delta = 1) = 1$. Für $n \in \mathbb{N}$, $n \geq n_0$, $p \in (0,1)$ und für ein $k_n \in \mathbb{N}$ mit $k_n := n - d$ reduziert sich der KL–Schätzer zu folgenden Ausdrücken:*

$$\overset{ACL}{KL}\,(p,k_n,n) = \frac{1}{\binom{n}{k_n}} \sum_{i=1}^{\binom{n}{k_n}} Q^i_{k_n}(p) \tag{4.16}$$

$$= \frac{1}{\binom{n}{k_n}} \sum_{i=1}^{\binom{n}{k_n}} Z^i_{]k_np[+1:k_n} \tag{4.17}$$

$$= \sum_{j=r}^{r+n-k_n} \frac{\binom{j-1}{r-1}\binom{n-j}{k_n-r}}{\binom{n}{k_n}} Z_{j:n}\,, \tag{4.18}$$

wobei $r =]k_np[+1$ und $Q^i_{k_n}(p)$ die empirische Quantilfunktion in der i–ten Unterstichprobe vom Umfang k_n an der Stelle p bezeichnet.

Beweis:

- Mit der Voraussetzung $s = 1$ ist $S^i_{k_n} = 1$ f.s. für alle $i = 1, \ldots, \binom{n}{k_n}$, und es folgt für R_i aus Satz 4.20:

$$R_i = \left]k_n\left(1-(1-p)^1\right)\right[+ 1 =]k_n p[+ 1 \quad \text{f.s.}.$$

Damit ergibt sich für den Ausdruck (4.8) aus Satz 4.20:

$$\overset{ACL}{KL}\,(p, k_n, n) = \frac{1}{\binom{n}{k_n}} \sum_{i=1}^{\binom{n}{k_n}} Z^i_{R_i:k_n} = \frac{1}{\binom{n}{k_n}} \sum_{i=1}^{\binom{n}{k_n}} Z^i_{]k_n p[+1:k_n}.$$

Es folgt die Gültigkeit von (4.17).

- Unter Verwendung der Darstellung (2.3) der empirischen Quantilfunktion für $p \in (0,1)$ in Kapitel 2.1 folgt:

$$\overset{ACL}{KL}\,(p, k_n, n) = \frac{1}{\binom{n}{k_n}} \sum_{i=1}^{\binom{n}{k_n}} Q^i_{k_n}(p),$$

d.h. (4.16) ist gezeigt.

- Es reduzieren sich mit $s = 1$ die Größen in (4.9) aus Satz 4.20 wie folgt:

$C_n = n$ f.s., $x = k_n$, $r_x = r =]k_n p[+1$, $a =]k_n p[$, $b = 1$ und

$\sum_{i=1}^{j-1} \Delta_{i:n} = j-1$ f.s., $\sum_{i=1}^{j-1}(1-\Delta_{i:n}) = 0$ f.s., $\Delta_{j:n} = 1$ f.s.,

$1 - \Delta_{j:n} = 0$ f.s., $\sum_{i=j+1}^{n} \Delta_{i:n} = n-j$ f.s., $\sum_{i=j+1}^{n}(1-\Delta_{i:n}) = 0$ f.s..

Damit ergibt sich aus dem Ausdruck (4.9) in Satz 4.20

$$\begin{aligned}
\overset{ACL}{KL}\,(p, k_n, n) &= \frac{1}{\binom{n}{k_n}} \sum_{j=1}^{n} \binom{j-1}{]k_n p[} \binom{0}{0} \binom{1}{1} \binom{0}{0} \binom{n-j}{k_n -]k_n p[- 1} \cdot \\
&\qquad \binom{0}{0} \cdot Z_{j:n} \\
&= \frac{1}{\binom{n}{k_n}} \sum_{j=1}^{n} \binom{j-1}{]k_n p[} \binom{n-j}{k_n -]k_n p[- 1} \cdot Z_{j:n} \\
&= \sum_{j=r}^{r+n-k_n} \frac{\binom{j-1}{r-1}\binom{n-j}{k_n - r}}{\binom{n}{k_n}} \cdot Z_{j:n},
\end{aligned}$$

womit (4.18) gezeigt ist. ■

Für die Unterstichprobenumfänge $k_n = 1$ bzw. $k_n = n$ reduziert sich der KL–Schätzer im Koziol–Green-Modell zum arithmetischen Mittel aller Beobachtungen bzw. zu einer einzigen Ordnungsstatistik als Schätzer für ξ_p, $p \in (0,1)$. Es gilt folgendes Lemma.

Lemma 4.22 *Sei $n_0 \in \mathbb{N}$ und $d \in \{0, 1, \ldots, n_0 - 1\}$. Für $n \in \mathbb{N}$, $n \geq n_0$, $p \in (0,1)$ und für ein $k_n \in \mathbb{N}$ mit $k_n := n - d$ reduziert sich der KL–Schätzer im Koziol–Green–Modell zu folgenden Ausdrücken:*

(i) Für $k_n = 1$ (d.h. $d = n - 1$):

$$\overset{ACL}{KL}(p, 1, n) = \frac{1}{n} \sum_{j=1}^{n} Z_{j:n}.$$

(ii) Für $k_n = n$ (d.h. $d = 0$):

$$\overset{ACL}{KL}(p, n, n) = Z_{r:n},$$

wobei

$$r = \begin{cases} \left]n\left(1 - (1-p)^{\frac{1}{S_n}}\right)\right[& \text{für } S_n \in \{\frac{1}{n}, \ldots, 1\} \\ \left]n\left(1 - (1-p)^{n+1}\right)\right[& \text{für } S_n = 0. \end{cases} \tag{4.19}$$

Beweis:

zu (i): Mit $k_n = 1$ ergibt sich aus (4.8):

$$\overset{ACL}{KL}(p, 1, n) = \frac{1}{\binom{n}{1}} \cdot \sum_{i=1}^{\binom{n}{1}} Z^i_{R_i:1} = \frac{1}{n} \sum_{i=1}^{n} Z_{i:n}.$$

zu (ii): Mit $k_n = n$ ergibt sich aus (4.8):

$$\overset{ACL}{KL}(p, n, n) = \frac{1}{\binom{n}{n}} \cdot \sum_{i=1}^{\binom{n}{n}} Z^i_{R_i:n} = Z_{r:n}$$

mit r gemäß (4.19).

■

Klassifikation des KL–Schätzers im Koziol–Green–Modell

In Satz 4.20 wurde gezeigt, daß sich der KL–Schätzer im Koziol–Green–Modell in der Form

$$\sum_{j=1}^{n} l_j\,(n, k_n, \Delta_{1:n}, \ldots, \Delta_{n:n}) \cdot Z_{j:n} \tag{4.20}$$

darstellen läßt. Der Ausdruck (4.20) ist eine Linearkombination von Ordnungsstatistiken mit zufälligen Koeffizienten und entspricht demnach nicht den Voraussetzungen für eine L–Statistik. So ist mit den Ergebnissen dieser Arbeit noch keine Darstellung für den KL–Schätzer als L–Statistik im Koziol–Green–Modell gefunden worden.

Zum Beispiel wäre die verallgemeinerte Form

$$\sum_{j=1}^{n} c_j(n, k_n) \cdot g\big(Z_{j:n}, \Delta_{j:n}\big)$$

mit Konstanten $c_j(n, k_n)$, $j = 1, \ldots, n$, $\sum_{j=1}^{n} c_j(n, k_n) = 1$ und mit einer Funktion $g : \mathbb{R}^+ \times \{0, 1\} \longrightarrow \mathbb{R}^+$ für eine Darstellung als L–Statistik denkbar.

Mit der Darstellung (4.8) ist der KL–Schätzer im Koziol–Green–Modell eine U–Statistik mit Kern

$$h_n : (\mathbb{R}^+ \times \{0, 1\})^{k_n} \longrightarrow \mathbb{R},$$

wobei

$$h_n\left((Z_1, \Delta_1), \ldots, (Z_{k_n}, \Delta_{k_n})\right) = Z_{R:k_n} = \tilde{Q}_{k_n}(p)$$

mit

$$R = \begin{cases} \left]k_n\big(1 - (1-p)^{(S_{k_n})^{-1}}\big)\right[+ 1 & \text{für} \quad S_{k_n} \in \{\frac{1}{k_n}, \ldots, 1\} \\ \left]k_n\big(1 - (1-p)^{k_n+1}\big)\right[+ 1 & \text{für} \quad S_{k_n} = 0 \end{cases} \tag{4.21}$$

und $S_{k_n} = \frac{1}{k_n} \cdot \sum_{i=1}^{k_n} \Delta_{i:k_n}$, $k_n = n - d$ mit $k_n \leq n$, $d \in \mathbb{N}_0$. Dabei ist $\tilde{Q}_{k_n}(p)$ die ACL–Quantilfunktion einer Stichprobe vom Umfang k_n an der Stelle p, $p \in (0, 1)$.

Da der Unterstichprobenumfang $k_n = n - d$, $d \in \mathbb{N}_0$ und $d \leq n$, von n abhängt und auch zusammen mit n wächst, gehört $\overset{ACL}{KL}\,(p, k_n, n)$ zur Klasse der U–Statistiken mit infiniter Ordnung (vgl. FREES, 1989). Eigenschaften dieser Klasse von U–Statistiken werden in FREES (1989) diskutiert.

Eigenschaften des KL–Schätzers im Koziol–Green–Modell

Mit Anwendung der Ergebnisse aus FREES (1989) ergeben sich die folgenden Überlegungen für den KL–Schätzer im Koziol–Green–Modell:

- Für den Erwartungswert gilt:

$$\mathrm{E}_F\big(\overset{ACL}{KL}(p, k_n, n)\big) = \mathrm{E}_F\big(\tilde{Q}_{k_n}(p)\big) = \mathrm{E}_F\big(Z_{R:k_n}\big)$$

  mit R gemäß (4.21). Für detailliertere Aussagen müßte die Verteilung von $\tilde{Q}_{k_n}(p)$ bzw. $Z_{R:k_n}$ untersucht werden.

- Ein Konsistenznachweis für $\overset{ACL}{KL}(p, k_n, n)$ könnte analog zu den Überlegungen für $\overset{emp}{KL}(p, k, n)$ im unzensierten Fall (Kapitel 3.4) versucht werden. Hierfür wäre u.a. auch der Nachweis der fast sicheren Konvergenz von $Z_{R:k_n}$ mit R wie in (4.21) gegen ξ_p von Nutzen. Vermutlich kann letzteres analog zum Vorgehen von GIJBELS und VERAVERBEKE (1989, S. 1650ff) erfolgen. Dort wird die starke Konsistenz zu ξ_p für die ACL–Quantilfunktion an der Stelle p, $p \in (0,1)$, gezeigt, welche auf der 1. Variante des ACL–Schätzers nach CSÖRGŐ (1988) basiert (siehe Definition 4.11).

- Mit Theorem 2.2 in FREES (1989, S. 32) könnte der Nachweis der asymptotischen Normalität des KL–Schätzers im Koziol–Green–Modell versucht werden. Dieses wäre erreicht, wenn (i)–(v) gelten:

  (i)

$$h_n\big((Z_1,\Delta_1),\ldots,(Z_{k_n},\Delta_{k_n})\big) = Z_{R:k_n} \xrightarrow{V} \xi_p =: h\big((Z_1,\Delta_1),(Z_2,\Delta_2),\ldots\big)$$

  (Formel (1.3) in Frees, 1989, S. 29).

  (ii)

$$\Big(h_n\big((Z_1,\Delta_1),\ldots,(Z_{k_n},\Delta_{k_n})\big)\Big)^2 = \big(Z_{R:k_n}\big)^2$$

  mit R gemäß (4.21) ist gleichmäßig integrierbar.

  (iii)

$$\begin{aligned} & n\cdot \mathrm{E}_F\Big|h_n\big((Z_1,\Delta_1),\ldots,(Z_n,\Delta_n)\big) - h\big((Z_1,\Delta_1),(Z_2,\Delta_2),\ldots\big)\Big| \\ = \; & n\cdot \mathrm{E}_F\Big|Z_{R:n} - \xi_p\Big| \longrightarrow 0 \quad \text{für} \quad n\to\infty \end{aligned}$$

  (Bemerkung in FREES, 1989, S. 32).

(iv) Für die Existenz der Varianz σ^2 der approximierten Normalverteilung muß das Produkt von zwei Reihen, die jeweils die Summe über die Differenzen von dem Funktionswert des Kerns und Grenzwert repräsentieren, für $n \to \infty$ konvergieren. Dabei enthält der Kern in der jeweiligen Reihe als Argumente bestimmte Teilfolgen von $(Z_1, \Delta_1), \ldots, (Z_n, \Delta_n)$ (Formel (2.16) in FREES, 1989, S. 35).

(v) Es muß gelten: $n \cdot \mathrm{Var}_F \left(\overset{ACL}{KL}\ (p, k_n, n) \right) \longrightarrow \sigma^2$.

Ein weiterer Ansatz für den Nachweis der asymptotischen Normalität könnte über die asymptotische Äquivalenz zu Kern–Quantilschätzern (vgl. Kapitel 2.2 und 3.3) versucht werden. Hierfür müßte ein geeigneter asymptotisch normalverteilter Kern–Quantilschätzer im Koziol–Green–Modell gefunden werden.

Bemerkung 4.23 STUTE *(1992) gibt einen Konsistenznachweis zu* $\int \phi\, dF$ *für Statistiken im Koziol–Green–Modell der Form*

$$\sum_{i=1}^{n} c_{ni}\phi(Z_{i:n}) = \int \phi\, d\tilde{F}_n^{(1)} \quad mit \quad c_{ni} = \left(1 - \frac{i-1}{n}\right)^{S_n} - \left(1 - \frac{i}{n}\right)^{S_n}, \qquad (4.22)$$

wobei ϕ *eine* F*–integrierbare Funktion ist und schwachen Annahmen unterliegt. Die Darstellung (4.9) entspricht jedoch nicht (4.22), und so können die Ergebnisse in* STUTE *(1992) nicht für den KL–Schätzer verwendet werden.*

Vermutlich können weitere Eigenschaften von U–Statistiken finiter Ordnung, wie z.B. der exakte Ausdruck für die Varianz, auf U–Statistiken infiniter Ordnung übertragen werden. Da jedoch auf keine Literatur verwiesen werden kann, wird diese Überlegung im Hinblick auf den KL–Schätzer im Koziol–Green–Modell unterlassen. Hier könnten zukünftige Forschungen ansetzen.

Kapitel 5

Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurde der nichtparametrische KL–Schätzer im unzensierten wie auch im zensierten Fall betrachtet. KAIGH und LACHENBRUCH (1982) führten diesen Quantilschätzer zuerst für unzensierte Daten ein. Fünf Jahre später übertrugen LIO und PADGETT (1987) ihn auf das Modell zufälliger Rechtszensierung. In dieser Arbeit erfolgte zusätzlich die Übertragung des KL–Schätzers auf das Koziol–Green–Modell. Letzteres ist ein Spezialfall des Modells zufälliger Rechtszensierung.

Nach der Einleitung diente das zweite Kapitel in erster Linie als kurzer Überblick über bekannte nichtparametrische Quantilschätzer. Neben Quantilschätzern, die nur auf einer Ordnungsstatistik oder auf zwei Ordnungsstatistiken basieren, gehört der KL–Schätzer zur Gruppe der Quantilschätzer, die mehrere Ordnungsstatistiken mit Hilfe einer glättenden Gewichtsfunktion berücksichtigen. Dabei wird für den KL–Schätzer eine Glättung erreicht, indem Unterstichproben vom festen Umfang betrachtet werden.

Im dritten Kapitel wurde der KL–Schätzer im unzensierten Fall untersucht. Ausgehend von der bekannten Zugehörigkeit zur Klasse der L– und U–Statistiken wurden bekannte Eigenschaften des KL–Schätzers zusammengetragen, teilweise aufgearbeitet, oder es wurden Eigenschaften neu untersucht.

Eine graphische Analyse der Gewichte des KL–Schätzers als L–Statistik ergab folgende Ergebnisse: Die Gewichtung der Ordnungsstatistiken erfolgt bei der Schätzung von jeweils gegensätzlich extremen Quantilen nicht analog. Weiter kann eine ungeeignete Wahl des Unterstichprobenumfangs k für die Schätzung von bestimmten

Quantilen zu völlig ungeeigneten Gewichtungen der Ordnungsstatistiken führen. Schließlich deutete ein Beispiel an, daß die geeignet transformierten Gewichte nur relativ langsam gegen die Dichte einer entsprechenden Betaverteilung konvergieren (Kapitel 3.1).

Weiter ließ sich mit Hilfe der L–Statistik–Darstellung der empirische Bruchpunkt für den KL–Schätzer finden. Dieser zeigt, daß die Robustheit des KL–Schätzers gegenüber Ausreißern wie erwartet abnimmt, wenn extremere Quantile zu schätzen sind (Kapitel 3.2.6).

Die verbleibenden aufgeführten Eigenschaften des KL–Schätzers basieren auf Ergebnissen für U–Statistiken.

Es wurden exakte Ausdrücke für den Erwartungswert und die Varianz des KL–Schätzers angegeben. Dabei zeigte die Formulierung des Erwartungswertes, daß der KL–Schätzer generell weder erwartungstreu noch asymptotisch erwartungstreu ist, wenn $n \to \infty$ wächst und k fest ist (Kapitel 3.2.1). Dürfte der Unterstichprobenumfang k dagegen zusammen mit n wachsen, so würde die asymptotische Erwartungstreue unter gewissen Annahmen an F gelten. Ebenso wie der Ausdruck des Erwartungswertes hängt auch der neu formulierte exakte Ausdruck der Varianz des KL–Schätzers von der unbekannten Verteilungsfunktion F ab. Daher wurden im Anschluß Varianzschätzer angegeben, die mit Resampling–Verfahren hergeleitet wurden (Kapitel 3.2.2). Die Bootstrap–Varianz konnte dabei mit Einführung einer geeigneten Varianzdarstellung des KL–Schätzers auf beliebige Unterstichprobenumfänge $k \leq n$ verallgemeinert werden.

Ein weiterer Abschnitt behandelte das asymptotische Verhalten des KL–Schätzers. Die starke Konsistenz für $n \to \infty$ gegen ξ_p konnte mit den verfügbaren Mitteln noch nicht gezeigt werden (Kapitel 3.2.3). Demgegenüber ist die asmptotische Normalität des KL–Schätzers für $n \to \infty$ und k fest wie auch für n und $k \to \infty$ schon bekannt. Diese wurde insbesondere für $n \to \infty$ und k fest im Beweis vorgeführt (Kapitel 3.2.4). Anschließend wurden asymptotische Konfidenzintervalle, welche auf Bootstrap– bzw. Jackknife–Varianzen basieren, aus der Literatur vorgestellt und im Bezug auf ihre Konstruktion kritisiert (Kapitel 3.2.5).

Kapitel 3.2.7 behandelte die Effizienz des KL–Schätzers im Sinne des minimalen mittleren quadratischen Fehlers im Vergleich zu dem Stichprobenquantil $X_{[(n+1)p]:n}$, $n \in \mathbb{N}$ und $p \in (0,1)$. Zu diesem Zweck wurden Ergebnisse einer Simulationsstudie

von KAIGH und LACHENBRUCH (1982) zusammengefaßt. Es zeigte sich, daß der KL–Schätzer abgesehen von der Schätzung von extremen Quantilen im allgemeinen eine größere Effizienz gegenüber $X_{[(n+1)p]:n}$ aufweist. Ebenso verdeutlichte die Studie den Einfluß des Unterstichprobenumfangs k auf den mittleren quadratischen Fehler und somit auf die Effizienz des KL–Schätzers. Generell ergaben größere Unterstichprobenumfänge verbesserte Schätzungen für extreme Quantile (Kapitel 3.2.7).

Ein weiteres Unterkapitel stellte bekannte Methoden zur Bestimmung des optimalen Unterstichprobenumfangs vor. Die genannten Vorschläge sind jedoch meist aus Mangel an Vorwissen schlecht realisierbar, oder Vorwissen wie die Kenntnis über p wurde nicht ausgenutzt (Kapitel 3.3).

Als Vorbereitung für die Übertragung auf den zensierten Fall wurde im letzten Abschnitt des dritten Kapitels eine alternative Definition des KL–Schätzers basierend auf der empirischen Quantilfunktion eingeführt (Kapitel 3.4). Im wesentlichen gelten auch für diese Definition die aufgeführten Eigenschaften des KL–Schätzers. Darüberhinaus motivierten bekannte Konsistenzeigenschaften der empirischen Quantilfunktion an der Stelle p, $p \in (0,1)$, zu weiteren Überlegungen für den Nachweis der starken Konsistenz des KL–Schätzers für n und $k \to \infty$.

Das vierte Kapitel untersuchte den KL–Schätzer für den zensierten Fall. Das Modell zufälliger Rechtszensierung wurde erläutert und die Übertragung des KL–Schätzers auf dieses Modell kurz beschrieben. Gemäß LIO und PADGETT (1987) kann hier der KL–Schätzer nur den U–Statistiken nicht aber den L–Statistiken zugeordnet werden (Kapitel 4.2).

Der Schwerpunkt des Kapitels lag in der Übertragung des KL–Schätzers auf das Koziol–Green–Modell mit dem ACL–Schätzer als Schätzer für die unbekannte Verteilungsfunktion. Nach Einführung einer modifizierten Definition des ACL–Schätzers wurde der KL–Schätzer basierend auf dem ACL–Schätzer im Koziol–Green–Modell definiert. Für den KL–Schätzer wurde eine Darstellung als Linearkombination von Ordnungsstatistiken der ursprünglichen Stichprobe gefunden. Da jedoch die zugehörigen Koeffizienten zufällig sind, ist damit noch keine Einordnung in die Klasse der L–Statistiken gelungen. Die Einordnung in die Klasse der U–Statistiken infiniter Ordnung gelang in diesem Modell. Abschließend wurden Ansätze für den Nachweis von Eigenschaften des KL–Schätzers im Koziol–Green–Modell vorgeschlagen (Kapitel 4.3).

Ausgehend von den Ergebnissen dieser Arbeit werden im folgenden Anmerkungen und noch offene Fragestellungen, zunächst für den unzensierten Fall, angefügt.

Der KL–Schätzer wurde von KAIGH und LACHENBRUCH (1982) mit dem Ziel eingeführt, einen möglichst effizienten nichtparametrischen Quantilschätzer zu erhalten. Wie schon die Simulationsstudie von KAIGH und LACHENBRUCH (1982) andeutete, hat der Unterstichprobenumfang einen entscheidenden Einfluß auf die Effizienz des KL–Schätzers. Daher ist die Optimierung des Unterstichprobenumfangs im Sinne des minimalen mittleren quadratischen Fehlers von großer Bedeutung. Vielleicht kann zu diesem Zweck unter Verwendung des exakten Ausdrucks für die Varianz des KL–Schätzers eine geeignetere Formulierung bzw. bessere Approximation des mittleren quadratischen Fehlers gefunden werden. Ebenso ist eine Anwendung der Jackknife– oder Bootstrap–Varianz denkbar. Vermutlich würde auch die Entwicklung eines Bootstrap–Schätzers für den mittleren quadratischen Fehler des KL–Schätzers von Nutzen sein.

Im allgemeinen lassen sich Effizienzvergleiche erst mit der Bestimmung eines optimalen Unterstichprobenumfangs k_{opt} für den KL–Schätzer vernünftig interpretieren. Damit könnte auch der Aufwand solcher Simulationsstudien verringert werden.

Generell sollten mehr Effizienzvergleiche für unsymmetrische Verteilungen durchgeführt werden. Es reicht nicht aus, sich wie bei KAIGH und LACHENBRUCH (1982) überwiegend auf den Vergleich der Varianzen bei symmetrischen Verteilungen für $p = 0,5$ zu beschränken, da der KL–Schätzer in den meisten Fällen nicht erwartungstreu ist.

Darüberhinaus sind Effizienzvergleiche mit anderen verallgemeinerten Stichprobenquantilen oder Kern–Quantilschätzern (vgl. Kapitel 2.2) sinnvoll, um das Ausmaß der Variabilitätsverringerung in der Schätzung festzustellen. YANG (1985) z.B. vergleicht in seiner Studie den KL–Schätzer mit dem Yang–Schätzer $\hat{Q}_5(p)$ (vgl. Kapitel 2.2), wobei keiner der beiden Schätzer ξ_p weitgehend effizienter schätzt.

Auch sollte bei Effizienzvergleichen immer nach der Art des zu schätzenden Quantils unterschieden werden, um die Eignung des KL–Schätzers speziell für mittlere bzw. für extreme Quantile untersuchen zu können.

In der Regel scheinen asymptotische Betrachtungen für den KL–Schätzer bessere Ergebnisse zu liefern, wenn k in Abhängigkeit von n wachsen darf. Unter dieser Voraussetzung ist der KL–Schätzer unter gewissen Annahmen asymptotisch erwar-

tungstreu, und wenn die starke Konsistenz von $X_{[(n+1)p]:n}$ gegen ξ_p gezeigt werden könnte, würden die Überlegungen zur Konsistenz für die alternative Definition des KL–Schätzers in analoger Weise auch für den von KAIGH und LACHENBRUCH (1982) definierten Schätzer relevant sein. Auch eignet sich erst die asymptotische Normalität für $k \to \infty$ und $n \to \infty$ zur Konstruktion von asymptotischen Konfidenzintervallen. Mit der Abhängigkeit des optimalen Unterstichprobenumfangs von n, stellt sich daher die Frage, ob nicht bereits bei der Definition des KL–Schätzers k in Abhängigkeit von n festgelegt werden sollte. So wäre zusätzlich die Eindeutigkeit seiner Definition gewährleistet.

In dieser Arbeit wurde prinzipiell die Stetigkeit der Verteilungsfunktion F vorausgesetzt. Es ist dagegen auch interessant zu untersuchen, wie sich der KL–Schätzer für diskrete Verteilungsfunktionen oder bei finiten Grundgesamtheiten verhält. Weiter wurde grundsätzlich angenommen, daß die Zufallsvariablen $X_1, \ldots, X_n$ stochastisch unabhängig sind. Es sind daher auch Robustheitsuntersuchungen für Abweichungen von dieser Unabhängigkeit von Interesse.

Im zensierten Fall und speziell für das Koziol–Green–Modell sind die Kenntnisse über den KL–Schätzer sind noch nicht weit fortgeschritten. Für fast alle Eigenschaften sind noch Untersuchungen notwendig. Hierbei sei auf die Überlegungen am Ende des vierten Kapitels hingewiesen. Weiter sollten Robustheitseigenschaften wie das Verhalten des KL–Schätzers bei Vorliegen von Ausreißern und bei Modellverletzungen im zensierten Fall untersucht werden. Schließlich kann auch die Übertragung auf weitere Zensierungsmodelle versucht werden.

Anhang A

Beweise

A.1 Nachweis der Identität verschiedener Darstellungen der empirischen Quantilfunktion

Die empirische Quantilfunktion ist mit Darstellung (2.1) wie folgt definiert:

$$Q_n(p) := \begin{cases} X_{1:n} & \text{für} \quad p = 0 \\ \sum_{i=1}^{n} X_{i:n} \, \mathbb{1}_{(\frac{i-1}{n}, \frac{i}{n}]}(p) & \text{für} \quad p \in (0,1]. \end{cases}$$

Es folgt:

$$Q_n(p) = \begin{cases} X_{1:n} & \text{für} \quad p = 0 \\ X_{i:n} & \text{für} \quad \frac{i-1}{n} < p \leq \frac{i}{n}, \quad i \in \{1, \ldots, n\}. \end{cases}$$

Obige Bedingung für p läßt sich in eine Bedingung für i äquivalent umformen. Es gilt:

$$Q_n(p) = \begin{cases} X_{1:n} & \text{für} \quad p = 0 \\ X_{i:n} & \text{für} \quad np \leq i < np + 1, \quad i \in \{1, \ldots, n\}, \, p \in (0,1] \end{cases}$$

und damit

$$Q_n(p) = \begin{cases} X_{1:n} & \text{für} \quad p = 0 \\ X_{np:n} & \text{für} \quad p \in (0,1] \text{ und } np \in \mathbb{N} \\ X_{[np]+1:n} & \text{für} \quad p \in (0,1] \text{ und } np \notin \mathbb{N}, \quad n \in \mathbb{N}, \end{cases}$$

(Darstellung (2.2) der empirischen Quantilfunktion). Daraus folgt unmittelbar:

$$Q_n(p) = \begin{cases} X_{1:n} & \text{für} \quad p = 0 \\ X_{]np[+1:n} & \text{für} \quad p \in (0,1] \text{ und } n \in \mathbb{N} \end{cases}$$

mit $]np[= \max\{z \in \mathbb{N} | z < np\}$.

■

A.2 Beweis zu Lemma 3.34

Sei $np + \varepsilon \in \mathbb{N}$ mit $0 \leq \varepsilon < 1$ und $n \in \mathbb{N}$, $p \in (0,1)$. Es folgt:

$$]np[+1 = np + \varepsilon - 1 + 1 = np + \varepsilon. \tag{A.1}$$

Zu zeigen ist:

$$\begin{aligned}]np[+1 &= [(n+1)p] \quad \text{für} \quad 0 \leq \varepsilon \leq p < 1 \qquad \text{und} \\ ]np[+1 &\neq [(n+1)p] \quad \text{für} \quad 0 < p < \varepsilon < 1. \end{aligned}$$

Es gilt:

a) Sei $0 \leq \varepsilon \leq p < 1$: $[(n+1)p] = [np + p] = np + \varepsilon \overset{(A.1)}{=}]np[+1$.

b) Sei $0 < p < \varepsilon < 1$: $[(n+1)p] = [np + p] = np + \varepsilon - 1 \overset{(A.1)}{\neq}]np[+1$.

■

Anhang B

Tabellen

Tabelle 3 *(zu Abbildung 1) Gewichte des KL–Schätzers für* $n = 50$, $k = 31$ *und* $p = 0,05$, $p = 0,5$ *und* $p = 0,95$.

$p = 0,05$	Index j zur Ordnungsstatistik $X_{j:n}$					
	1	2	3	4	5	
Gewicht	**0,62**	0,24	0,09	0,03	0,01	

$p = 0,50$	Index j zur Ordnungsstatistik $X_{j:n}$										
	19	20	21	22	23	24	25	26	27	28	29
Gewicht	0,01	0,02	0,04	0,07	0,10	0,10	**0,14**	**0,14**	0,12	0,10	0,07
	30	31	32								
Gewicht	0,04	0,02	0,01								

$p = 0,95$	Index j zur Ordnungsstatistik $X_{j:n}$							
	43	44	45	46	47	48	49	
Gewicht	0,01	0,02	0,04	0,09	0,17	0,30	**0,38**	

Die Werte der Gewichte sind gerundet. Gewichte von nicht aufgeführten Ordnungsstatistiken ergeben sich zu 0,00. Der Modus der Gewichte ist fett hervorgehoben.

Tabelle 4 *(zu Abbildung 2) Gewichte des KL–Schätzers für* $n = 20$, $p = 0,5$ *und verschiedene Werte von* k.

Index j zur Ordnungsstatistik $X_{j:n}$	Unterstichprobenumfang k				
	2	5	10	15	18
1	0,10	0,00	0,00	0,00	0,00
2	0,09	0,00	0,00	0,00	0,00
3	0,09	0,01	0,00	0,00	0,00
4	0,08	0,02	0,00	0,00	0,00
5	0,08	0,04	0,02	0,00	0,00
6	0,07	0,06	0,05	0,00	0,00
7	0,07	0,08	0,10	0,00	0,00
8	0,06	0,09	0,15	0,05	0,00
9	0,06	0,10	0,18	0,17	0,29
10	0,05	0,10	0,17	0,28	0,47
11	0,05	0,10	0,14	0,28	0,24
12	0,04	0,10	0,10	0,17	0,00
13	0,04	0,09	0,06	0,05	0,00
14	0,03	0,08	0,02	0,00	0,00
15	0,03	0,06	0,01	0,00	0,00
16	0,02	0,04	0,00	0,00	0,00
17	0,02	0,02	0,00	0,00	0,00
18	0,01	0,01	0,00	0,00	0,00
19	0,01	0,00	0,00	0,00	0,00
20	0,00	0,00	0,00	0,00	0,00

Die Werte der Gewichte sind gerundet.

Anhang C

Resampling–Verfahren

C.1 Das Jackknife–Verfahren

Das Jackknife–Verfahren wird zur nichtparametrischen Schätzung von Schätzervarianzen und zur Biasreduktion verwendet. Es stellt eine direkte numerische Approximation dieser Größen zur Verfügung (HINKLEY, 1983, S. 280).

Es werden folgende Bezeichnungen für einen interessierenden Schätzer $T_n(X_1, \ldots, X_n)$, $n \in \mathbb{N}$, verwendet:

- $T_{(-i)}$ ist die entsprechende Schätzfunktion, welche auf der reduzierten Stichprobe $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ vom Umfang $n-1$ definiert ist, $i = 1, \ldots, n$.
- $T^i_{pseu} := n \cdot T_n - (n-1) \cdot T_{(-i)}$ ist der i–te Pseudowert, $i = 1, \ldots, n$.

Damit ist der Jackknife–Schätzer des Schätzers T_n, $n \in \mathbb{N}$,

$$Jack(T_n) := \frac{1}{n} \sum_{i=1}^{n} T^i_{pseu} \quad ,$$

als arithmetisches Mittel über alle Pseudowerte definiert.

Der Jackknife–Varianzschätzer (Jackknife–Varianz) für T_n, $n \in \mathbb{N}$, ist wie folgt definiert:

$$\widehat{\mathrm{Var}}_{Jack}(T_n) \;:=\; \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(T^i_{pseu} - Jack(T_n)\right)^2$$

$$= \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{(-i)} - \bar{T}_{(.)}\right)^2, \qquad \text{(C.1)}$$

wobei $\bar{T}_{(.)} = \frac{1}{n} \sum_{i=1}^{n} T_{(-i)}$ ist (LEE, 1990, S. 217f).

Letzterer kann entweder als Varianzschätzer des Jackknifeschätzers $Jack(T_n)$ oder als Varianzschätzer des ursprünglichen Schätzers T_n verwendet werden (EFRON, STEIN, 1981, S. 586).

C.2 Das Bootstrap–Verfahren

Das Bootstrap–Verfahren ist eine weitere Resampling–Methode, die zur Schätzung verschiedener Charakteristiken wie z.B. der Varianz oder des Bias eines interessierenden Schätzers verwendet werden kann.

Es sei $C(F, n)$ eine Charakteristik eines Schätzers $T_n(X_1, \ldots, X_n)$, $n \in \mathbb{N}$, wobei $C(F, n)$ nicht notwendigerweise explizit als Funktion von F und n angegeben sein muß. Die Idee des Bootstrap ist es, die obige Charakteristik an der Stelle F_n zu betrachten, notiert als $C(F_n, n)$.

Ist es nicht möglich, $C(F, n)$ explizit anzugeben, so kann $C(F_n, n)$ mit einem Monte–Carlo–Algorithmus approximiert werden. Dazu werden B, $B \in \mathbb{N}$, sogenannte Bootstrap–Stichproben $X_1^*, \ldots, X_n^*$ vom Umfang n, $n \in \mathbb{N}$, gezogen, welche unabhängig und identisch wie die empirische Verteilungsfunktion F_n verteilt sind. Aus diesen Stichproben werden die B sogenannten Bootstrap–Wiederholungen $T^{*b}(X_1^{*b}, \ldots, X_n^{*b})$, $b = 1, \ldots, B$, berechnet. Mit letzteren läßt sich dann der approximierte Bootstrap–Schätzer der Charakteristik $C(F, n)$ formulieren (LEE, 1990, S. 230, EFRON und GONG, 1983).

C.3 Die Kreuz–Validierungsmethode

Die Kreuz–Validierungsmethode dient allgemein zur Schätzung von sogenannten Vorhersagefehlern. Ein spezielles Verfahren dieser Methode wird kurz erläutert.

Für einen interessierenden Schätzer $T_n(X_1, \ldots, X_n)$, $n \in (0, 1)$, werden wie beim Jackknife–Verfahren die Schätzer $T_{(-i)}$ betrachtet, welche auf der reduzierten Stichprobe $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ basieren, $i = 1, \ldots, n$. Weiter wird für alle $i =$

$1, \ldots, n$ der Fehler berechnet, der durch die Schätzung mit $T_{(-i)}$ für den ausgelassenen Wert X_i enstanden ist. Das Mittel über alle Fehler liefert den kreuz–validierten Vorhersagefehler des Schätzers T_n (EFRON, 1982, S. 49).

Literaturverzeichnis

[1] **Abdushukurov, A.A. (1984).** On some estimates of the distribution function under random censorship. *In:* Conference of young scientists. Math. Inst. Acad. Sci. Uzbek SSR, Tashkent, VINITI No. 8756-V (in russisch).

[2] **Breslow, N.; Crowley, J. (1974).** A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, **2**(3), 437-453.

[3] **Bronstein, I.N.; Semendjajew, K.A. (1987).** *Taschenbuch der Mathematik,* 23. Auflage. Verlag Harri Deutsch, Thun und Frankfurt am Main.

[4] **Chen, Y.Y.; Hollander, M. & Langberg, N.A. (1982).** Small–sample results for the Kaplan–Meier estimator. *Journal of the American Statistical Association*, **77**(377), 141-144.

[5] **Cheng, P.E.; Lin, G.D. (1984).** Maximum likelihood estimation of a survival–function under the Koziol–Green proportional hazards model. *Technical Report B-84-5.* Institute of Statistics, Academia Sinica, Taipei, Taiwan.

[6] **Cox, D.R. (1972).** Regression models and life–tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187-202.

[7] **Csörgő, S. (1988).** Estimation in the proportional hazards model of random censorship. *Statistics*, **19**(3), 437-463.

[8] **Csörgő, S.; Horváth, L. (1981).** On the Koziol–Green model for random censorship. *Biometrika*, **68**(2), 391-401.

[9] **David, H.A. (1981).** *Order Statistics,* Second Edition. John Wiley & Sons, New York.

[10] **Davis C.E.; Steinberg S.M. (1986).** Quantile estimation. *In:* Encyclopedia of Statistical Sciences 7, 408-412, S. Kotz und N.L. Johnson (Hrsg.). John Wiley & Sons, New York.

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

[12] **Efron, B. (1979).** Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1), 1-26.

[13] **Efron, B. (1982).** *The Jackknife, the Bootstrap and other Resampling Plans.* SIAM, monograph No. 38, CBNS–NSF.

[14] **Efron, B.; Gong, G. (1983).** A leisurely look at the bootstrap, the jackknife, and cross–validation. *The American Statistician*, **37**(1), 36-48.

[15] **Efron, B.; Stein, C. (1981).** The jackknife estimate of variance. *The Annals of Statistics*, **9**(3), 586-596.

[16] **Frees, E.W. (1989).** Infinite order U–statistics. *Scandinavian Journal of Statistics*, **16**, 29-45.

[17] **Garsd, A.; Ford, G.E.; Waring III, G.O. & Rosenblatt, L.S. (1983).** Sample size for estimating the quantiles of endothelial cell–area distribution. *Biometrics*, **39**, 385-394.

[18] **Ghorai, J.K. (1991).** Estimation of a smooth quantile function under the proportional hazards model. *Annals of the Institute of Statistical Mathematics*, **43**(4), 747-760.

[19] **Gijbels, I.; Veraverbeke, N. (1989).** Quantile estimation in the proportional hazards model of random censorship. *Communications in Statistics–Theory and Methods*, **18**(5), 1645-1663.

[20] **Gronen, S. (1993).** *Der Harrell–Davis–Quantilschätzer.* Diplomarbeit, Fachbereich Statistik, Universität Dortmund.

[21] **Hall, P. (1992).** *The Bootstrap and Edgeworth Expansion.* Springer Verlag, New York.

[22] **Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J. & Stahel, W.A. (1986).** *Robust Statistics. The Approach based on Influence Functions.* John Wiley & Sons, New York.

[23] **Harrell, F.E.; Davis, C.E. (1982).** A new distribution-free quantile estimator. *Biometrika*, **69**(3), 635-640.

[24] **Hartung, J. (1987).** *Statistik.* R. Oldenbourg Verlag GmbH, München.

[25] **Herbst, T. (1992).** Estimation of moments under Koziol–Green model. *Communications in Statistics–Theory and Methods*, **21**(3), 613-624.

[26] **Hinkley, D. (1983).** Jackknife methods. *In:* Encyclopedia of Statistical Sciences 4, 280-287, S. Kotz und N.L. Johnson (Hrsg.). John Wiley & Sons, New York.

[27] **Hoeffding, W. (1948).** A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, **19**, 293-325.

[28] **Hollander, M.; Proschan, F. (1979).** Testing to determine the underlying distribution using randomly censored data. *Biometrics*, **35**, 393-401.

[29] **Kaigh, W.D. (1983).** Quantile interval estimation. *Communications in Statistics–Theory and Methods*, **12**(21), 2427-2443.

[30] **Kaigh, W.D. (1988).** O–Statistics and their applications. *Communications in Statistics–Theory and Methods*, **17**(7), 2191-2210.

[31] **Kaigh, W.D.; Cheng C. (1991a).** Subsampling quantile estimators and uniformity criteria. *Communications in Statistics–Theory and Methods*, **20**(2), 539-560.

[32] **Kaigh, W.D.; Cheng C. (1991b).** Subsampling quantile estimator standard errors with applications. *Communications in Statistics–Theory and Methods*, **20**(3), 977-995.

[33] **Kaigh, W.D.; Lachenbruch, P.A. (1982).** A generalized quantile estimator. *Communications in Statistics–Theory and Methods*, **11**(19), 2217-2238.

[34] **Kaplan, E.L.; Meier, P. (1958).** Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457-481.

[35] **Kappenmann, R.F. (1987).** Improved distribution quantile estimation. *Communications in Statistics–Simulation and Computation*, **16**(2), 307-320.

[36] **Koziol, J.A.; Green, S.B. (1976).** A Cramér–von Mises statistic for randomly censored data. *Biometrika*, **63**(3), 465-474.

[37] **Lampkin, H.; Ogawa, J. (1976).** Estimation of distribution parameters in time mortality trials. An example of time mortality analysis. *The Canadian Journal of Statistics*, 4(1), 65-93.

[38] **Lee, A.J. (1990).** *U–Statistics Theory and Practise.* Marcel Dekker, New York.

[39] **Lehmann, E.L. (1983).** *Theory of Point Estimation.* John Wiley & Sons, New York.

[40] **Lio, Y.L.; Padgett, W.J. (1987).** A generalized quantile estimator under censoring. *Communications in Statistics–Theory and Methods*, **16**(11), 3301-3321.

[41] **Meister, R. (1984).** *Ansätze zur Quantilschätzung.* Dissertation, Fachbereich Informatik (20), Technische Universität, Berlin.

[42] **Miller, R.G. (1981).** *Survival Analysis.* John Wiley & Sons, New York.

[43] **Padgett, W.J. (1986).** A kernel–type estimator of a quantile function from right–censored data. *Journal of the American Statistical Association*, **81**(393), 215-222.

[44] **Parrish, R.S. (1990).** Comparison of quantile estimators in normal sampling. *Biometrics*, **46**, 247-257.

[45] **Parzen, E. (1979).** Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74(365), 105-121.

[46] **Reiss, R.-D. (1989).** *Approximate Distributions of Order Statistics.* Springer Verlag, New York.

[47] **Sander, J. (1975).** The weak convergence of quantiles of the product–limit estimator. *Technical Report No.5*, Stanford University, Department of Statistics.

[48] **Serfling, R.J. (1980).** *Approximation Theorems of mathematical Statistics.* John Wiley & Sons, New York.

[49] **Sheather, S.J.; Marron, J.S. (1990).** Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410), 410-416.

[50] **Steinberg, M.; Davis, C.E. (1985).** Distribution–free confidence intervals for quantiles in small samples. *Communications in Statistics–Theory and Methods*, **14**(4), 979-990.

[51] **Stigler, S.M. (1969).** Linear functions of order statistics. *The Annals of Mathematical Statistics*, 40(3), 770-788.

[52] **Stute, W. (1992).** Strong consistency under the Koziol–Green model. *Statistics and Probability Letters*, **14**, 313-320.

[53] **Witting, H. (1985).** *Mathematische Statistik I.* B.G. Teubner, Stuttgart.

[54] **Witting, H.; Nölle, G. (1970).** *Angewandte mathematische Statistik. Optimale finite und asymptotische Verfahren.* B.G. Teubner, Stuttgart.

[55] **Yang, S.-S. (1985).** A smooth nonparametric estimator of a quantile function. *Journal of the American Statistical Association*, **80**(392), 1004-1011.

[56] **Zelterman, D. (1990).** Smooth nonparametric estimation of the quantile function. *Journal of Statistical Planning and Inference*, **26**, 339-352.

Hiermit erkläre ich, daß ich die vorliegende Arbeit selbständig verfaßt habe und keine anderen als die angegebenen Quellen verwendet habe.

Dortmund, im Oktober 1993

Viviane Grunert

(Viviane Grunert)