Jackknife und Bootstrap: Resampling-Verfahren zur Genauigkeitsschätzung von Parameterschätzungen

Rothe, Günter

Jackknife und Bootstrap: Resampling-Verfahren zur Genauigkeitsschätzung von Parameterschätzungen

Günter Rothe

ZUMA-Arbeitsbericht Nr. 89/04

Zentrum für Umfragen, Methoden und Analysen e.V. (ZUMA)
Postfach 122155
D-6800 Mannheim 1

Seit Juli 1983 sind die ZUMA-Arbeitsberichte in zwei Reihen aufgeteilt:

Die ZUMA-Arbeitsberichte (neue Folge) haben eine hausinterne Begutachtung durchlaufen und werden von Geschäftsführenden Direktor zusammen mit den übrigen Wissenschaftlichen Leitern herausgegeben. Die Berichte dieser Reihe sind zur allgemeinen Weitergabe nach außen bestimmt.

Die ZUMA-Technischen Berichte dienen zur hausinternen Kommunikation bzw. zur Unterrichtung externer Kooperatonspartner. Sie sind nicht zur allgemeinen Weitergabe nach außen bestimmt.

Jackknife und Bootstrap: Resampling-Verfahren zur Genauigkeitsschätzung von Parameterschätzungen

Günter Rothe

1.Vorbemerkungen

In den letzten zehn Jahren hat die "Mathematisierung" der empirischen Sozialwissenschaften in erheblichem Maße zugenommen: Nachdem seit längerem Strukturen nicht nur verbal beschrieben, sondern auch in geeigneten mathematischen Modellen formalisiert werden, hat die Entwicklung auf dem Computersektor es nun ermöglicht gemacht, Schätzverfahren für die in abstrakten Modellen auftretenden unbekannten Parameter nicht nur theoretisch zu untersuchen, sondern auch explizit zu berechnen. Man denke nur an die Iterationsverfahren zur Lösung von Maximum-Likelihood-Gleichungen, deren Rechenaufwand früher die technischen Möglichkeiten der Anwender oft überschritt.

Die Komplexität der zugrundeliegenden Modelle jedoch bewirkt, daß die durch derartige Verfahren gewonnenen Schätzwerte hinsichtlich ihrer Qualität häufig gar nicht hinterfragt werden; z.B. ist weder die Existenz einer Maximum-Likelihood-Lösung stets gesichert, noch steht fest, ob der vom Computer gefundene Schätzer tatsächlich die einzige oder überhaupt eine Lösung des Gleichungssystems darstellt - hier spielen neben der Eindeutigkeit der Lösung auch die Eindeutigkeit der Schätzung etwa in Abhängigkeit vom Startwert der Iteration eine wesentliche Rolle. Selbst wenn jedoch die vom Rechner produzierte Schätzung "die richtige" ist, wird kaum Wert gelegt auf eine Quantifizierung der Genauigkeit der Schätzung: Ein ML-Schätzer ist in der Regel verzerrt und das Ausmaß der Verzerrung im Einzelfall nicht abzuschätzen. Auch die Angabe von Konfidenzbereichen kann nicht explizit erfolgen: Entweder man verzichtet auf sie oder man benutzt Aussagen über die Asymptotik des Verfahrens. Dabei wird oft nicht berücksichtigt, daß der verwendete Stichprobenumfang im Verhältnis zur Zahl der unbekannten Parameter eine derartige Übertragung auf eine finite Situation ggf. gar nicht erlaubt. Asymptotischen Verfahren nutzen in der Regel Aussagen über die asymptotische

Normalität des verwendeten Schätzers. Hierbei ist aber oft noch die asymptotische (bzw. ersatzweise die exakte) Varianz der Schätzers aus den Daten zu schätzen; derartige Schätzer liegen in vielen Fällen jedoch ebenfalls nicht vor.

In den letzten Jahren sind in der mathematischen Statistik Konzepte entwickelt worden, die in dieser Hinsicht eine Verbesserung versprechen und dabei ebenfalls intensiv die Möglichkeiten des Rechners nutzen: Resampling-Verfahren wie Bootstrap und Jackknife wecken die Hoffnung, daß mit ihnen zumindest eine deutliche Erhöhung der Aussagequalität erreicht werden kann: Ihre Begründer preisen sie als Verfahren, mit denen sich die Verzerrungen von Schätzern korrigieren, die mittleren quadratischen Fehler der Schätzer abschätzen und sich Konfidenzbereiche sowie statistische Tests konstruieren lassen, die den "klassischen", auf der Asymptotik beruhenden Verfahren deutlich überlegen sind und Modellüberprüfungen erlauben. In der Tat stimmen die ersten mathematisch-statistischen Arbeiten auf diesem Gebiet ausgesprochen optimistisch. Sie zeigen jedoch auch, daß in sehr vielen Einzelfällen Probleme auftauchen können, sodaß bei jeder Anwendung der Verfahren in einem neuen Bereich erneut Vorsicht angebracht ist.

Das Ziel der vorliegenden Arbeit ist es, die Resampling-Prozeduren "Jackknife" und "Bootstrap" zu beschreiben und zu erläutern, in welcher Weise sie zur Bestimmung von Verzerrungskorrekturen und Varianzschätzern sowie zur Konstruktion von Konfidenzbereichen herangezogen werden können. Die konkrete Durchführung der Verfahren wird anhand eines Beispiels erläutert, das in der Arbeit wiederholt aufgegriffen wird und das bewußt möglichst elementar gehalten ist, um den Leser bei der Einarbeitung auf diesem Gebiet nicht unnötig zu verwirren. In Kapitel 2 wird zunächst das Jackknife-Verfahren beschrieben, Kap.3 befaßt sich mit der Bootstrap-Prozedur. In Kap. 4 werden dann die verschiedenen Verfahren zur Konstruktion von Konfidenzintervallen vorgestellt. Das bis dahin bereits mehrfach angesprochene Beispiel ist schließlich in Kap. 5 Grundlage für einige kleine Simulationsstudien, die dem Leser einen ersten Einblick in die unterschiedliche Wirkungsweise der verschiedenen Verfahren vermitteln sollen.

2.Das Jackknife-Verfahren

2.1.Biaskorrektur

Das Jackknife-Verfahren wurde erstmalig in einer Arbeit von Quenoille (1949) vorgeschlagen (vgl. auch Quenoille (1956)). Es wurde vorgestellt als ein Ansatz zur Korrektur der Verzerrung von Schätzungen. Wir werden diese Überlegungen in diesem Abschnitt skizzieren und dabei immer von folgender Datensituation ausgehen:

Modell: Es seien $X_1,\ldots,X_n$ unabhängige, identisch verteilte Zufallsvariablen, die eine Verteilung mit einer unbekannten Verteilungsfunktion F besitzen. Es sollen nun verschiedene Parameter dieser Verteilung geschätzt werden, die sich als Funktion von F beschreiben lassen, d.h. Parameter der Form $\theta=\theta(F)$. Mögliche Beispiele hierfür wären spezielle Quantile oder Fraktile von F - z.B. der Median - oder Momente von F, also etwa der Erwartungwert von X_1 (das erste Moment von F) oder die Varianz von X_2 (das zweite zentrale Moment von F).

Ferner werden wir uns wiederholt speziell mit folgendem Schätzproblem befassen:

> Beispiel: Wir nehmen zunächst an, daß die Zufallsgrößen gewisse endliche Momente besitzen:
>
> $$E(|X_1^j|) < \infty \quad \text{für } j=1,2,3 \text{ und } 4.$$
>
> Damit existieren automatisch ihr Erwartungswert und ihre Varianz:
>
> $$E(X_1) = \mu \; ; \; m_2 = \sigma^2(X_1) = Var_F(X_1) = \int (x-\mu)^2 \, dF(x) < \infty.$$
>
> Die Größen μ und m_2 lassen sich nun "klassisch" durch das empirische Mittel und die empirische Varianz der beobachteten Daten $X_1,\ldots,X_n$ schätzen:
>
> $$\hat{\mu} = \bar{X}_\bullet = \sum_{1\le i\le n} X_i \quad \text{und}$$

$$\hat{m}_2 = s^2 = n^{-1} \cdot \sum_i (x_i - \bar{x}_\cdot)^2$$

In dieser einfachen Situation rechnet man sofort nach, daß zwar

$$E_F(\hat{\mu}) = \mu,$$

dagegen jedoch

$$E_F(\hat{m}_2) = m_2 \cdot (n-1)/n$$

gilt, d.h. der Varianzschätzer nicht mehr erwartungstreu ist.

Beide beschriebenen Schätzer sind von der Form $\hat{\theta}_n = \theta(\hat{F}_n)$, wobei $\hat{F}_n$ die empirische Verteilungsfunktion der Daten ist. Die Abweichung des Erwartungswertes eines Schätzers von dem Parameter, den er schätzen soll (also die Größe $E(\hat{\theta}_n - \theta)$), wird als Verzerrung bzw. als BIAS des Schätzers bezeichnet. Wenn sich im obigen Beispiel die Verzerrung auch leicht korrigieren läßt (indem man den Schätzer mit $n/(n-1)$ multipliziert), so sind (unkorrigierbare) Abweichungen dieser Art jedoch bei Schätzern der Form $\theta(\hat{F}_n)$ grundsätzlich an der Tagesordnung.

Ähnlich sieht es bei anderen Schätzverfahren aus: Nimmt man als Verteilung der X_i eine Normalverteilung mit unbekanntem Mittelwert und unbekannter Varianz an, so sind $\hat{\mu}$ und $\hat{m}_2$ die Maximum-Likelihood-Schätzer; das Problem ist dabei von ganz allgemeiner Natur: Maximum-Likelihood-Schätzer sind ebenfalls in der Regel verzerrt.

Es stellt sich dabei nun heraus, daß in den meisten Fällen (speziell beim Maximum-Likelihood-Schätzer, aber nicht nur dort) die Verzerrung von der Ordnung $1/n$ ist, d.h.

$$E(\hat{\theta}_n) - \theta + c(\theta)/n \quad \text{bzw.}$$

$$\lim_{n \to \infty} n \cdot E(\hat{\theta}_n - \theta) = c(\theta),$$

wobei $c(\theta)$ eine unbekannte Größe ist. Das Jackknife-Verfahren, so wie es 1949 vorgeschlagen wurde, ist nun ein Verfahren, das "automatisch" zur Verzerrungskorrektur angewandt werden kann, sobald eine Situation vorliegt, wie sie bisher beschrieben wurde. Die Idee des Konzepts läßt sich folgendermaßen einsehen:

Entfernen wir aus $X_1,\ldots,X_n$ eine der Beobachtungen - sagen wir, die i-te -, so können wir eine Schätzung von θ nach dem gleichen Verfahren unter Verwendung von $X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n$ auf der Basis von n-1 Beobachtungen durchführen. Der auf diese Weise gewonnene Schätzer $\hat{\theta}_{n-1\backslash i}$ für θ hat nun die Verzerrung der Größenordnung $c(\theta)/(n-1)$. Damit gilt für die sogenannten "Pseudo-Werte"

$$\hat{\theta}_n^{(i)} := n\cdot\hat{\theta}_n - (n-1)\cdot\hat{\theta}_{n-1\backslash i}$$

die Beziehung

$$E(\hat{\theta}_n^{(i)}) = n\theta + c - (n-1)\theta - c = \theta;$$

der "Jackknife-Schätzer"

$$\hat{\theta}_n^{J} = \frac{1}{n}\cdot\sum_{1\le i\le n}\hat{\theta}_n^{(i)}$$

besitzt somit die Eigenschaften

$$E(\hat{\theta}_n^{J}) = \theta \quad \text{und} \quad \lim_{n\to\infty} n\cdot E(\hat{\theta}_n^{J} - \theta) = 0,$$

die Verzerrung der Ordnung 1/n ist also verschwunden.

Beispiel (Fortsetzung): Führt man die Jackknife-Prozedur für $\hat{\mu}_n$ durch, so zeigt sich schnell, daß hierdurch der Schätzer nicht verändert wird, d.h. es gilt $\hat{\mu}_n = \hat{\mu}_n^{J}$. Beim Schätzer für die Varianz ergibt sich automatisch $\hat{m}_2^{J} = \frac{n}{n-1}\cdot\hat{m}_2$, d.h. es ergibt sich die richtige "Bias-Korrektur".

Das vorgestellte Konzept zur Verzerrungskorrektur läßt sich erweitern. In der Tat läßt sich die Verzerrung meist in der Form

$$b_n(\theta) = E(\hat{\theta}_n - \theta)$$

$$= \sum_{i=1}^{\infty} c_i(\theta) / n^i$$

darstellen. Modifikationen bei der Konstruktion komplexerer "Pseudowerte" erlauben dann auch eine Elimination von Verzerrungskomponenten höherer Ordnung (vgl. Schucany, Gray und Owen, 1971).

2.2. Varianzschätzung

Wir kehren nun wieder zu den "Pseudowerten" zurück, die im vorigen Abschnitt definiert wurden. Durch eine kurze Notiz von Tukey (1958) wurde für das Jackknife-Konzept eine völlig neue Richtung eröffnet. Diese Notiz ist eine Kurzfassung eines Vortrages; sie enthält vier Thesen, die sich für die Forschung auf diesem Gebiet als außerordentlich richtungsweisend herausstellten. Tukey blieb jedoch eine mathematische Rechtfertigung seiner Thesen schuldig. Der Versuch anderer Autoren, festzustellen, was Tukey wirklich gemeint haben könnte (d.h. unter welchen Modellannahmen man den Tukeyschen Vorschlag wirklich übernehmen kann), dauert auch heute, 30 Jahre nach dieser Notiz, noch an. Im vorliegenden Abschnitt zitieren wir kurz Tukeys Behauptungen und fügen jeweils einige kurze Anmerkungen hierzu an.

(1) Man behandle die Pseudowerte wie unabhängige, identisch verteilte Zufallsgrößen! Insbesondere ist dann

$$S_{(J)}^2 := \frac{1}{n-1} \cdot \sum_{1 \le i \le n} \left(\hat{\theta}_n^{(i)} - \hat{\theta}_n^{(J)} \right)^2$$

ein vernünftiger Schätzer für $\mathrm{Var}(\hat{\theta}_n^{(1)})$ bzw. $\hat{\sigma}_{(J)}^2 = S_{(J)}^2 / n$ ein Schätzer für $\mathrm{Var}(\hat{\theta}_n^{(J)})$.

Dies ist sicherlich der "härteste Brocken". Es kann an dieser Stelle für diese "Idee" nur eine (natürlich von mathematischen Standpunkt her absolut

unakzeptable) Plausibilitätserklärung herhalten: Schaut man sich die Definition des i-ten Pseudowertes an, so könnte man ihn beschreiben als den Teil des Schätzers $\hat{\theta}_n$, der diejenigen Informationen enthält, die man nicht bereits aus der Schätzung durch $\hat{\theta}_{n-1\backslash i}$ gewonnen hätte - also unter Verzicht auf die i-te Beobachtung. Damit wäre aber $\hat{\theta}_n^{(i)}$ eine vorwiegend nur von X_i, nicht aber den anderen Beobachtungen abhängige Größe. Dies läßt einen "hohen Grad an Unabhängigkeit" zwischen den Pseudowerten vermuten (wobei Unabhängigkeit ja eigentlich keine skalierte Größe ist: "Ein bißchen unabhängig" geht genausowenig wie "ein bißchen schwanger"). Daß sie zusätzlich identisch verteilt sind, ist dagegen unproblematisch - aufgrund der symmetrischen Definition der Pseudowerte und der indentischen Verteilungen der X_i.

(2) Der Jackknife-Schätzer hat fast die gleiche Varianz wie der ursprüngliche Schätzer. Damit ist $\hat{\sigma}_{(J)}^2$ auch automatisch ein geeigneter Schätzer für die Varianz von $\hat{\theta}_n$ selbst.

Führt man die Plausibilitätsüberlegung zur letzten "These" konsequent fort, so ist $\hat{\theta}_n^{(J)}$ tatsächlich nicht anderes als die "Wieder-Zusammenführung" der "Teil-Informationen" $\hat{\theta}_n^{(i)}$, $1 \le i \le n$. Tukeys Behauptung "lebt und stirbt" aber natürlich mit der Gültigkeit von These 1. In der Tat läßt sich feststellen, daß in vielen Fällen der Jackknife-Schätzer die Varianz von $\hat{\theta}_n$ eher überschätzt (vgl. Efron und Stein (1981), Efron (1982)); so etwa auch in unserem Beispiel - wir werden dies in einem späteren Abschnitt noch sehen.

(3) Verwende - bei vorgegeben $\alpha>0$ - als zweiseitiges Konfidenzintervall für θ das Intervall

$$I_\alpha = [\, \hat{\theta}_n^{(J)} - t_{\alpha/2,n-1} \cdot \hat{\sigma}_J^2,\ \hat{\theta}_n^{(J)} + t_{\alpha/2,n-1} \cdot \hat{\sigma}_J^2 \,].$$

$t_{\alpha/2,n-1}$ ist hierbei das $\alpha/2$-Fraktil der t-Verteilung mit $(n-1)$ Freiheitsgraden.

Wären die Pseudowerte unabhängig und identisch normalverteilt, hätte das Intervall in der Tat genau das Niveau α. Unter Berücksichtigung der Thesen (1) und (2) ist dann - aufgrund des zentralen Grenzwertsatzes - I_α zumindest als asymptotisches Konfidenzintervall zu rechtfertigen. Dann allerdings besteht kein Grund, warum unbedingt ein t-Fraktil und nicht das (durch die

Asymptotik vorgegebene) Fraktil der Standardnormalverteilung Verwendung finden sollte. Allerdings macht das t-Fraktil das Intervall etwas größer und die Konstruktion des Konfidenzintervalls wird auf diese Weise "konservativer".

(4) Das hier beschriebene Verfahren zur Schätzung der Varianz eines Schätzers und zur Konstruktion von Konfidenzbereichen ist universell einsetzbar, insbesondere auch dann, wenn komplexe Modellparameter zu schätzen sind, für deren Varianzschätzung keine akzeptable Theorie zur Verfügung steht. Das Verfahren stellt also ein Universalwerkzeug dar, es ist quasi das "Taschenmesser" (dt. für "jackknife") des Statistikers.

Eben diese Universalität ist aber derzeit noch nicht vollständig erforscht. Zunächst muß festgestellt werden, daß alle Aussagen vorwiegend asymptotische Aussagen sind, d.h. man befaßt sich vorwiegend mit der Konsistenz des Jackknife-Varianzschätzers. Darüberhinaus muß die Klasse der Verteilungen F und die parametrischen Funktionen $\theta(F)$ identifiziert werden, für die zumindest die Konsistenz des Jackknife-Varianzschätzers gegeben ist. Selbst dann ist der Wert derartiger Konsistenzaussagen für die Qualität der Schätzung bei kleinen Stichproben zweifelhaft, insbeosndere bei einem derartigen "Universalverfahren".

Der Begriff "Jackknife" wurde von Tukey - mit der obigen Begründung - in die Literatur eingeführt. Dieser Name hat der Popularität des Verfahrens sicherlich nicht geschadet.

2.3. Weitere Entwicklungen

In der Literatur finden sich viele Modifikationen des Jackknife-Ansatzes mit zum Teil leicht unterschiedlichen Eigenschaften. Hier sind etwa Ansätze zu nennen derart, daß nicht genau eine, sondern gleichzeitig mehrere Beobachtungen jeweils ausgelassen werden bzw. "Bruchteile" einer Beobachtung; dies führt dann weiter zum infinitesimalen Jackknife von Jaeckel (1972). Eine Übersicht über die wichtigsten Entwicklungen im Bereich des Jackknife in den ersten 25 Jahren findet der interessierte Leser bei Miller (1974).

Beran (1984b) untersucht Modifikationen, bei denen eine oder mehrere Beobachtungen dupliziert werden und leitet hieraus Jackknife-Schätzungen nicht nur für die Varianz, sondern auch für höhere Momente wie etwa die Kurtosis ab. Schließlich sind derzeit auch Möglichkeiten zur Jackknife-Korrektur von Schätzungen in linearen Modellen von großem Interesse (vgl. hierzu Wu (1986)).

3. Bootstrap

3.1.Das Grundprinzip

Das Bootstrap-Verfahren wurde erstmals von Efron (1979) vorgestellt. Selten hat sich eine Prozedur so schnell etabliert wie diese: Die theoretische Untersuchung ihrer Eigenschaften ist aufgrund der Komplexität der hierfür erforderlichen Techniken für Mathematiker bzw. mathematische Statistiker ausgesprochen reizvoll, andererseits gibt sie aber auch dem Anwender Möglichkeiten in die Hand, die noch wesentlich vielversprechender - und universaler - sind als es zwanzig Jahre zuvor Tukey vom Jackknife-Verfahren behauptet hat.

Auch bei der Beschreibung dieses Konzepts werden wir uns auf den "Standardfall" unabhängiger, identisch verteilter Zufallsgrößen $X_1,\ldots,X_n$ konzentrieren. Man beachte, daß $X_1,\ldots,X_n$ unabhängige Zufallsvektoren sein dürfen - damit sind etwa die Modelle der Linearen Strukturgleichungen (LISREL, vgl. Jöreskog und Sörbom (1988)) ebenfalls enthalten. Für andere Situationen, wie etwa für Lineare Modelle, Generalisierte Lineare Modelle, Stichprobenverfahren etc. müssen Modifikationen des Verfahrens gefunden und untersucht werden; dies ist in vielen Fällen bereits gelungen (vgl. z.B. Bickel und Freedman (1984), Rothe(1989), Freedman(1984)).

Sind die Größen X_i unabhängig und identisch nach einer unbekannten Verteilung mit Verteilungsfunktion F verteilt, so stellt F den Parameter dar, der das statistische Modell (zusammen mit der Unabhängigkeitsannahme) vollständig beschreibt. Wieder interessiert uns zunächst eine parametrische Funktion von F, d.h. eine Größe $\theta=\theta(F)$, die geschätzt werden soll, und wiederum, wie bereits im vorigen Abschnitt, möge ein Schätzer $\hat{\theta}_n$ für θ vorgegeben sein. Wir betrachten nun den Schätzfehler

$$B(X_1,\ldots,X_n;F) = \hat{\theta}_n - \theta(F) \ .$$

Diese Größe ist natürlich im konkreten Fall nicht bekannt (sonst bräuchte man θ ja nicht zu schätzen). Grundsätzlich von Interesse wäre aber auch einfach die Verteilung oder die Verteilungsfunktion von $B(X,F)$ unter F, d.h. die Größe

$$G_B(F;x) = P_F \ (\ B(X;F) \leq x \ )$$

Würde man diese Größe kennen, so hätte man eine Vielzahl zusätzlicher Informationen über die Qualität des Schätzers $\hat{\theta}_n$:

- Das erste Moment von $G_B(F;\cdot)$ ist der Erwartungswert des Schätzfehlers und damit die Verzerrung des Schätzers,
- das zweite Moment von $G_B(F,\cdot)$ ist der mittlere quadratische Fehler der Schätzung $\hat{\theta}_n$ von θ,
- das zweite zentrale Moment von $G_B(F,\cdot)$ ist die Varianz von $\hat{\theta}_n$,
- weitere höhere Moment können Aufschluß geben über die Form der Dichte der Verteilung von $\hat{\theta}_n$,
- Fraktile und Quantile der Verteilung würden die Konstruktion exakter Konfidenzintervalle für θ ermöglichen.

In der Tat ist eine solche Verteilung nur in selten bekannt: Wenn wir etwa als sicher annehmen könnten, daß $X_1 \sim N(\mu,\sigma^2)$, also die X_1's normalverteilt sind mit bekannter Varianz σ^2 und unbekanntem Erwartungswert μ, so wäre im Falle $\theta=\mu$ und $\hat{\theta}_n = \bar{X}_\cdot = \sum_{1\leq i\leq n} X_i/n$ nämlich $G_B(N(\mu,\sigma^2);\cdot)$ die Verteilungsfunktion der $N(0,\sigma^2/n)$-Verteilung - also unabhängig vom unbekannten Parameter μ. Leider ist dies aber auch schon fast der einzige Sonderfall.

Die Bootstrap-Idee besteht nun ganz einfach darin, die unbekannte Verteilung - bzw. die Verteilungsfunktion $G_B(F,\cdot)$ - zu schätzen. Prinzipiell ist das hierdurch gestellte Problem gar nicht neu: Auch $G_B(F,\cdot)$ ist nichts anderes als eine parametrische Funktion von F. Die einzige im Moment offensichtliche Schwierigkeit besteht nur darin, daß der Wertebereich dieser parametrischen Funktion die Menge der Verteilungsfunktionen ist und nicht - wie in allen bisher diskutierten Beispielen - die Menge der reellen Zahlen oder bestenfalls ein endlichdimensionaler Vektorraum über den reellen Zahlen.

Hat man sich aber an diesen Gedanken erst einmal gewöhnt, liegt ein Schätzer für $G_B(F,\cdot)$ direkt auf der Hand: Man suche zunächst einen Schätzer für F, nenne diesen etwa $\hat{F}$ und schätze dann $G_B(F,\cdot)$ durch $G_B(\hat{F},\cdot)$. Und auch ein geeigneter Schätzer für F liegt vor - zumindest in der Situation, die wir in diesem Papier ausschließlich behandeln: Man verwende die empirische Verteilungsfunktion der Daten, also

$$\hat{F}_n\,(X_1,\ldots,X_n;t) := \#\{i\leq n \mid X_i \leq t\}\,/n\,,^1$$

die wir im folgenden kurz mit $\hat{F}_n$ bezeichnen werden (obwohl sie stets von $X_1,\ldots,X_n$ abhängt).

Damit ist die wesentliche Idee des Bootstrap bereits beschrieben; nur eine kleine Erweiterung werden wir noch hinzufügen - im Hinblick auf weitere Überlegungen in den nächsten Abschnitten: Wir haben bisher einen Ansatz beschrieben, der dazu dienen soll, die Verteilung des Schätzfehlers B(X,F) zu schätzen. Allerdings sind wir nicht gezwungen, uns auf diesen Fall zu beschränken: Die Idee ist prinzipiell immer anwendbar, solange es um die Schätzung der Verteilung einer Funktion geht, die sowohl von den Beobachtungen $X=(X_1,\ldots,X_n)$ als auch dem unbekannten Parameter F abhängt. Offen ist jedoch immer noch, ob die Vorgehensweise überhaupt sinnvoll ist - auch wenn sie im ersten Moment sehr einleuchtend erscheint. Dies soll in späteren Abschnitten noch diskutiert werden. Zunächst wird das Konzept nochmals allgemein definiert:

Def.: Sei R(X,F) eine Funktion der Daten X und des unbekannten Parameters F sowie

$$G_R(F;\,t) := P_F(\,R(X,F) \leq t)$$

ihre Verteilungsfunktion. Dann ist die Verteilungsfunktion

$$\hat{G}(t) := G_R(\hat{F}_n;\,t)$$

---

1 Das Zeichen # bezeichnet die Anzahl der Elemente der dem Zeichen folgenden Menge.

der Bootstrap-Schätzer von $G_R(F; \cdot)$.

Damit haben wir nun einen Schätzer für die Verteilung von R definiert. Oben haben wir angesprochen, daß - etwa bei der Schätzfehlerverteilung - insbesondere jedoch wieder reellwertige parametrische Funktionen dieser Verteilung zur Beurteilung der Schätzung von Interesse sind. Es ist nun wiederum naheliegend, die interessierende parametrische Funktion der Verteilungsfunktion zu schätzen durch die parametrische Funktion der entsprechenden Bootstrap-Schätzung, d.h. das erste, zweite, zweite zentrale oder dritte Moment bzw. ein α-Fraktil von $G_R(F; \cdot)$ wird schlicht und einfach geschätzt durch das erste, zweite, zweite zentrale oder dritte Moment bzw. ein α-Fraktil von $G_R(\hat{F}_n; \cdot)$.

Man kann sich das Konzept auch an folgender Situation bildlich klarmachen: Man stelle sich vor, daß man die "reale Welt" künstlich generiere - nur mit dem Unterschied, daß die wahre Verteilung der beobachteten X_i nun bekannt sei - nämlich $\hat{F}_n$. Um die "künstlichen" X_i nun von den "realen" X_i zu unterscheiden, werden wir sie im folgenden mit X_i^* bezeichnen. Man stelle sich nun ferner vor, in dieser "künstlichen" Welt müsse ein "künstlicher" Statistiker den ihm unbekannten Parameter θ schätzen. (Wir kennen diesen unbekannten Parameter, denn in der künstlichen Welt ist dieser identisch mit $\theta(\hat{F}_n)$!) Führt dieser Statistiker nun eine Schätzung aufgrund eines Experimentes durch, das ihm die Daten $X_1^*,\ldots,X_n^*$ geliefert hat - genau wie wir mit unseren Daten $X_1,\ldots,X_n$ - so macht er ebenfalls einen Schätzfehler. Im Gegensatz zu ihm kennen wir aber nun seine Schätzfehlerverteilung - denn wir kennen auch den wahren Parameter (nämlich $\hat{F}_n$), der die künstliche Welt vollständig beschreibt. Das Bootstrap-Verfahren macht dann nichts anderes, als daß es die Schätzfehlerverteilung des "künstlichen" Statistikers als Schätzfehlerverteilung unserer realen Situation verwendet[2].

Wir weisen an dieser Stelle noch einmal darauf hin, daß zunächst noch überhaupt keine Überlegungen dahingehend angestellt wurden, ob dieses Verfahren

---

2 Science-Fiction-Kenner werden sich hier möglicherweise an den Roman "Welt am Draht" von Daniel F. Galouye erinnern, in den eine solche "künstliche Welt" im Computer ganz naturgetreu simuliert wurde - mit einem ähnlichen Ziel. Man beachte: Die Hauptperson des Romans stellt im Lauf der Zeit fest, daß sie selber nur Teil eines ähnlichen Simulationsprogramm einer noch höheren Ebene ist.

(so intuitiv einleuchtend es im Moment sein mag) überhaupt statistisch sinnvoll ist. Diese Untersuchung ist in der Tat auch nicht trivial: Wir haben es zum Beispiel unter anderem mit dem Problem zu tun, daß der Bootstrap-Schätzer Werte in der Menge der Verteilungsfunktionen annimmt, also eine funktionswertige Zufallsgröße darstellt. Zumindest ist also für eine mathematische Würdigung die Theorie stochastischer Prozesse erforderlich. Da die Bootstrap-Schätzer bestimmter reller Funktionen der Verteilungen ebenfalls über den Umweg dieser Verteilungsschätzer gewonnen wurden, bleibt hierfür das Problem ebenfalls bestehen, auch wenn diese Bootstrap-Schätzer ihrerseits nun wieder "bloß" reellwertig sind.

Zum Abschluß diese Abschnittes sei noch vermerkt, daß natürlich die Verwendung der empirischen Verteilung bei der Konstruktion einer Bootstrapschätzung einer Verteilung kein "Muß" ist: Man könnte sich etwa auch eine "geglättete" Version von $\hat{F}_n$ vorstellen (die dann eine Dichte besitzt), die dann eine andere Schätzung von $G_R(F,\cdot)$ - mit in der Regel auch anderen Eigenschaften - liefern würde. Hat man ein eingeschränktes Modell, d.h. reduziert man die Klasse der grundsätzlich zur Diskussion stehenden Verteilungen F für die Größen X_1, so sind ggf. "vernünftigere" Schätzer sinnvoller. Betrachten wir die weiter oben kurz angesprochene Situation, in den $X_1 \sim N(\mu,\sigma^2)$, so würde als Schätzung für die Verteilung von X_1 sicher $N(\bar{X}_\cdot,\sigma^2)$ sinnvoller sein als die empirische Verteilung, die ja ihrerseits diskret und damit nicht einmal eine Normalverteilung ist.

3.2.Berechnung der Bootstrap-Verteilung

Eigentlich ist die Idee des Bootstrap ja so naheliegend, daß man sich fragt, warum man sich nicht eher mit diesem Verfahren auseinandergesetzt hat. Genaugenommen ist dies auch längst geschehen, wenn auch unbewußt und ohne eine korrekte Würdigung - und zwar in der folgenden Situation:

Gehen wir zunächst wieder davon aus, daß ein Modell mit der Restriktion $X_1 \sim N(\mu,\sigma^2)$ vorliege, wobei nun neben μ auch σ^2 unbekannt sei. Eine geeignete Schätzung für die Verteilung von X_1 ist dann die Verteilung $N(\bar{X}_\cdot, S^2)$ mit

$$S^2 = \sum_{1\le i\le n}(X_i-\bar{X}_\cdot)^2/(n-1)$$

Nun betrachten wir die parametrische Funktion $\theta(F) = \mu$ und interessieren uns für die Verteilung G des Schätzfehlers $\bar{X} - \mu$. Der Bootstrap-Schätzer hierfür ist offensichtlich die Verteilung $N(0,S^2/n)$. Das α-Fraktil dieser Verteilung ist dann $u_\alpha \cdot S/\sqrt{n}$, wobei $u_\alpha = \Phi^{-1}(1-\alpha)$ das α-Fraktil der Standardnormalverteilung N(0,1) ist. Will man nun ein zweiseitiges Konfidenzintervall für μ konstruieren, so wäre dieses gegeben durch

$$I(1-\alpha) = [\ \bar{X}_\bullet - G^{-1}(1-\alpha/2),\ \bar{X}_\bullet - G^{-1}(\alpha/2)\ ],$$

falls diese Quantile $G^{-1}(t)$ bekannt wären. Die Bootstrap-Schätzung dieser Größen wären dann die entsprechenden Quantile der Bootstrap-Schätzung von G, d.h. das zweiseitiges $(1-\alpha)$-Bootstrap-Konfidenzintervall für μ ist dann

$$I_B(1-\alpha) = [\ \bar{X}_\bullet - u_{\alpha/2} \cdot S/\sqrt{n}\ ,\ \bar{X}_\bullet + u_{\alpha/2} \cdot S/\sqrt{n}\ ].$$

Dies ist natürlich kein Konfidenzintervall vom exakten Niveau $(1-\alpha)$ mehr: $u_{\alpha/2}$ müßte durch $t_{\alpha/2,n-1}$ ersetzt werden. Aber genauso hat man für normalverteilte Stichproben Konfidenzintervalle für den Mittelwert im vorigen Jahrhundert berechnet, bis Student (1908) auf diesen Fehler aufmerksam machte.

Dieses einfache Beispiel zeigt bereits, daß durch Bootstrap mit Sicherheit nicht stets exakte Ergebnisse zu erwarten sind, sondern daß es sich dabei um Approximationen an die Wirklichkeit handelt, bei denen die Abweichungen theoretisch analysiert und quantifiziert werden müssen. Dies ist aber, wie wir sofort sehen werden, nicht der Grund für die späte "Wiedergeburt" des Bootstrap. Um den wesentlichen "Teufel im Detail" zu erkennen, betrachten wir nun wieder unser Beispiel der Schätzung des zweiten zentralen Moments von F und versuchen, mittels Bootstrap die Verteilung des Schätzfehlers zu schätzen:

Beispiel (Fortsetzung): Der zu schätzenden unbekannte Parameter ist hier $\theta(F)=m_2(F)$, also das zweite zentrale Moment von F bzw. die Varianz von X_1. Der Schätzfehler ist damit

$$B(X,F) = n^{-1} \cdot \sum_1 (X_1 - \bar{X}_\bullet)^2 - m_2.$$

Wollen wir nun die Bootstrap-Schätzung für dessen Verteilung bestimmen, so betrachten wir die Größe

$$B^*(X^*,\hat{F}_n) = n^{-1}\cdot\sum_i (X_i^* - \bar{X}_\cdot^*)^2 - n^{-1}\cdot\sum_i (X_i - \bar{X}_\cdot)^2 .$$

Deren Verteilung ist nun zu bestimmen unter der Annahme, das die X_i feste Zahlen und die X_i^* Zufallsgrößen mit Verteilung $\hat{F}_n$ sind.

Was das bedeutet, kann man sich wieder am besten vergegenwärtigen am Bild des "künstlichen Statistikers" aus dem vorigen Abschnitt. Dieser Statistiker gewinnt seine Daten X_i^* aus der Verteilung mit Verteilungsfunktion $\hat{F}_n$, d.h. X_i^* nimmt, jeweils mit Wahrscheinkeit 1/n, einen der Werte $X_1,\ldots,X_n$ an. Damit erhält er als Tupel $X^* = (X_1^*,\ldots,X_n^*)$ ein beliebiges n-Tupel mit Komponenten aus $X_1,\ldots,X_n$. Prinzipiell sind für diesen "künstlichen" Statistiker also n^n verschiedene Versuchsausgänge möglich (es sei denn, daß zufällig $X_i = X_j$ für ein $i\neq j$ gilt - diesen Fall wollen wir der Einfachheit halber im folgenden ignorieren); um die Verteilung von B^* explizit zu berechnen, muß somit B^* für quasi jeden dieser n^n Versuchsausgänge ausgerechnet werden [3]. Dies ist bereits bei recht kleinem n auch mit den schnellsten Rechnern noch nicht mit vertretbarem Zeitaufwand zu meistern.

Die Folge ist, daß in den meisten Fällen der Bootstrap-Schätzer der Verteilung gar nicht explizit ausgerechnet werden kann. Damit ist aber auch die Bootstrap-Schätzung von Momenten oder Fraktilen dieser Verteilung nicht mehr möglich.

Es gibt jedoch einen "Ausweg", der allerdings erst in den letzten Jahren in zunehmendem Maße praktikabel wurde: Mit Hilfe einer Monte-Carlo-Studie ist es nämlich möglich, die Bootstrap-Schätzung ihrerseits mit beliebiger Genauigkeit zu schätzen, wobei die Genauigkeit ausschließlich von der Kapazität des verfügbaren Rechners (bzw. die akzeptierte Rechenzeit) vorgegeben ist. Gerade dieses Konzept jedoch, das im folgenden beschrieben werden soll,

---

3 Genaugenommen sind es weniger, da jede Permutation von X^* zum gleichen Wert von B^* führt. Die Zahl der erforderlichen Rechnungen bleibt allerdings nach wie vor unangenehm groß.

ist erst durch die Entwicklung auf dem Computersektor in den letzten Jahren praktikabel geworden und hat auf diese Weise den "Siegeszug" (?!) des Bootstrapverfahrens erst ermöglicht.

Das Monte-Carlo-Verfahren besteht darin, daß tatsächlich das Experiment des "künstlichen Statistikers" wiederholt durchgeführt wird und die auf diese Weise realisierten "Schätzfehler" gesammelt und zur Schätzung der Bootstrap-Verteilung herangezogen werden. Zunächst muß hierzu der Umfang N_{boot} der Simulationsstudie festgelegt werden. Zur Schätzung der Verteilungsfunktion $G_R(\hat{F}_n,x)$ sind dann folgende Schritte erforderlich:

(1) Ziehe eine Stichprobe $(X_1^*,\ldots,X_n^*)$ mit Zurücklegen aus der Menge $\{X_1,\ldots,X_n\}$.
(2) Berechne $R^* = R(X_1^*,\ldots,X_n^*; \hat{F}_n)$ und speichere den Wert.
(3) Wiederhole (1) und (2) N_{boot}-mal; bezeichne die auf diese Weise gewonnen Werte für R^* mit $R_1^*,\ldots,R_{N_{boot}}^*$.
(4) Bilde die empirische Verteilungsfunktion der Resampling-Schätzungen $R_1^*,\ldots,R_{N_{boot}}^*$, also

$$\hat{G}_n := \#\{j \leq N_{boot} \mid R_j^* \leq t\} / N_{boot}$$

und verwende diese als Schätzung $\hat{G}_R(\hat{F}_n,\cdot)$ von $G_R(\hat{F}_n,\cdot)$.

Dieses Konzept der Schätzung eines Schätzers ist übrigens für den Namen "Bootstrap" verantwortlich: Im ersten Moment sieht es so aus, als wolle man dadurch, daß man aus seinen Daten wiederholt Stichproben ohne Zurücklegen zieht, zusätzliche Informationen gewinnen - eine Idee, die an Münchhausen erinnert, der sich ja auch an seinen Haaren aus dem Sumpf ziehen konnte. Die Redensart "sich an den eigenen Haaren aus dem Sumpf ziehen" gibt es im Englischen auch, nur daß man dort nicht an den Haaren, sondern an der Stiefelschlaufe (engl.: bootstrap) zieht. Aus diesem Grunde ist das beschriebene Verfahren in deutschsprachigen Arbeiten gelegentlich mit "Münchhausen-Verfahren" übersetzt worden (vgl. etwa Diaconis und Efron (1983)).

3.3.Einige Anmerkungen zur Qualität

Bei der beschriebenen Schätzung der Verteilung von R überlagern sich zwei Arten von Fehlern: Zunächst erfolgt ein Schätzfehler bei der Schätzung von $G_R(F,\cdot)$ durch $G_R(\hat{F}_n,\cdot)$, danach zusätzlich aber noch ein Fehler bei der Schätzung von $G_R(\hat{F}_n,\cdot)$ durch $\hat{G}_R(\hat{F}_n,\cdot)$. Die zweite Approximation ist schlicht eine Approximation einer Verteilungsfunktion durch eine empirische Verteilungsfunktion, die in der mathematischen Statistik recht eingehend erforscht ist; Probleme entstehen hierbei jedoch bei der Auswahl des Zufallsgenerators, der zur Gewinnung der Sub-Stichproben X^* im obigen Algorithmus unter (1) erforderlich ist: Die in den Computerprogrammen letztlich implementierten Zufallsgeneratoren sind nur "Pseudo-Zufallszahlen-Generatoren", d.h. sie generieren deterministisch Zahlen, aber derart, daß die auf diese Weise gewonnen Zahlenfolgen möglichst vielen Tests auf "Nicht-Zufälligkeit" wiederstehen können. Die Qualität solcher Generatoren ist oft sehr fragwürdig, insbesondere wenn wie beim Bootstrap eine große Zahl von Zufallszahlen gezogen wird und so leicht die Gefahr besteht, daß unerwünschte periodische Effekte auftreten. Wir verweisen an dieser Stelle hierzu nur auf die umfangreiche Monographie von Knuth (1981) über die Konstruktion effektiver Zufallszahlengeneratoren.

Die Untersuchungen über theoretische Eigenschaften des Bootstrap-Schätzers in der mathematisch-statistischen Fachliteratur befassen sich hingegen vorwiegend mit der Größenordnung des anderen der beiden oben angesprochenen "Schätzfehler". Ein elementares Konzept ist hierbei der Begriff der Konsistenz. Bevor wir skizzieren, was hierunter zu verstehen ist, betrachten wir die Situation noch einmal für den Fall des Schätzfehlers B(X,F). Ist der Schätzer $\hat{\theta}_n$ für θ halbwegs "vernünftig", so wird er stark (oder zumindest schwach) konsistent sein, d.h. es wird gelten

$$\hat{\theta}_n - \theta \longrightarrow 0 \quad P_F\text{-fast sicher}$$

oder zumindest nach P_F-Wahrscheinlichkeit; ferner wird der Schätzfehler nach geeigneter Standardisierung asymptotisch normalverteilt sein, d.h. es gibt für jedes F ein $\sigma_B^2(F) > 0$ derart, daß

$$\sqrt{n}\cdot(\hat{\theta}_n - \theta) \longrightarrow N(0,\sigma_B^2(F)) \text{ nach } P_F\text{-Verteilung.}$$

Wir nehmen nun grundsätzlich allgemein an, daß die Verteilung von $R(X,F)$ unter F - ggf. nach einer geeigneten Standardisierung - für $n \to \infty$ gegen eine nichtentartete Normalverteilung (d.h. mit positiver Varianz) konvergiere. Es ist sicherlich eine Minimalforderung an die Funktionsfähigkeit des Bootstrap-Verfahrens, daß dann (ggf. nach der gleichen Standardisierung) auch die Verteilung von $R(X^*,\hat{F}_n)$ unter $\hat{F}_n$ den gleichen Grenzwert besitzt. Eine mathematisch präzise Formulierung dieser Forderung sieht etwas verwirrend aus, da es sich bei der Verteilung von $R(X^*,\hat{F}_n)$ unter $\hat{F}_n$ um eine Zufallsgröße handelt. Der Konvergenzbegriff muß sich darauf beziehen, es muß sich z.B. um eine Konvergenz "P_F-fast sicher" oder "nach P_F-Wahrscheinlichkeit" handeln. Im Falle der Verteilung des standardisierten Schätzfehlers bedeutet also die (starke bzw. schwache) Konsistenz ihres Bootstrap-Schätzers:

Es gelte

$$G_B(F,\ \sigma_B^2(F)\cdot t/\sqrt{n}\ ) \to \Phi(t) \text{ für alle } t \in R.$$

Dann ist der Bootstrap-Schätzer schwach bzw. stark konsistent, falls

$$G_B(\hat{F}_n,\ \sigma_B^2(F)\cdot t/\sqrt{n}\ ) \to \Phi(t) \text{ für alle } t \in R$$
nach P_F-Wahrscheinlichkeit bzw. P_F-fast sicher.

Tatsächlich ist das in dieser Arbeit beschriebene statistische Modell das hinsichtlich dieser Fragestellung am umfangreichsten untersuchte; die Klasse der parametrischen Funktionen $\theta(F)$, für die die Bootstrap-Konsistenz für den standardisierten Schätzfehler nachgewiesen werden konnte, ist inzwischen recht groß. Exemplarisch sei hierzu auf die Arbeiten von Singh (1981), Bickel und Freedman (1981), Beran (1984a), Lohse (1984) oder Klenk und Stute (1986) hingewiesen.

Konsistenz allein wäre allerdings kein Kriterium zugunsten des Bootstrap; sie bedeutet ja nur, daß die Bootstrap-Schätzung die wahre Verteilung von R asymptotisch ebensogut approximiert wie deren asymptotische Verteilung. Dann aber gäbe es nur in dem Fall einen Grund für die Anwendung des Bootstrap, wenn sich die asymptotische Verteilung von R einem klassischen Zugang (etwa unter Verwendung einer konsistenten Schätzung der asymptotischen Varianz)

entzieht. In anderen Fällen wie etwa unserem Beispiel wäre Bootstrap wegen des erheblich größeren Rechenaufwandes im Nachteil.

Bemerkenswert beim Bootstrap ist jedoch die Tatsache, daß es in der Regel gegenüber der asymptotischen Verteilung einen erheblichen Vorteil besitzt: Es zeigt sich nämlich, daß "unter schönen Bedingungen" (wir werden uns mit diesen Regularitätsbedingungen hier im Detail nicht befassen) die Approximation der Verteilung von R durch Bootstrap besser ist als durch die asymptotische Verteilung: Die Bootstrap-Schätzung ist "so gut" wie die erste Edgeworth-Approximation (vgl. z.B. Beran (1982)). Dieses im Prinzip zunächst ausschließlich theoretische Ergebnis (auch diese Aussage ist "nur" eine asymptotische Aussage) ist jedoch Hauptursache für die "Euphorie" über Bootstrap in den letzten Jahren; sieht es doch so aus, als würden mit Bootstrap bessere Approximationen z. B. an die Schätzfehlerverteilung und damit statistisch genauere Aussagen über die Genauigkeit von Schätzungen möglich - und dies sogar noch mit einem "automatisierbaren" Konzept, das wenig Einblick in die insgesamt ablaufenden technischen Rechenvorgänge bei der Bestimmung des Schätzers erforderlich macht.

4.Konstruktion von Konfidenzintervallen

4.1. "Klassische" Asymptotik: Normalapproximation

In diesem Abschnitt befassen wir uns mit einer Vielzahl von Ansätzen zur Konstruktion eines Konfidenzbereichs für den zu schätzenden Parameter $\theta=\theta(F)$. Dabei werden wir uns zunächst konzentrieren auf solche Schätzer, die die KAN-Bedingung erfüllen, d.h. die konsistent und asymptotisch normalverteilt sind, also die bereits im vorangegangen Abschnitt angesprochenen Eigenschaften

$$\hat{\theta}_n - \theta \dashrightarrow 0 \quad P_F\text{-fast sicher}$$

bzw. nach P_F-Wahrscheinlichkeit sowie

$$\sqrt{n} \cdot (\hat{\theta}_n - \theta) \dashrightarrow N(0,\sigma_B^2(F)) \text{ nach } P_F\text{-Verteilung}$$

für ein $\sigma_B^2(F) > 0$ besitzen. Wäre $\sigma_B^2(F)$ bekannt, so wäre

$$I_n^{(as)}(1-\alpha) = [\ \hat{\theta}_n - u_{\alpha/2}\cdot\sigma_B(F)/\sqrt{n}\ ,\ \hat{\theta}_n + u_{\alpha/2}\cdot\sigma_B(F)\ /\ \sqrt{n}\ ]$$

ein asymptotisches (1-α)-Konfidenzintervall, für das gilt

$$P_F(\ \theta \in I_n^{(as)}(1-\alpha)\ ) \dashrightarrow 1-\alpha\ .$$

Hierbei bezeichnet $u_t = \Phi^{-1}(1-t)$ das t-Fraktil der Standardnormalverteilung. In der Regel ist aber auch $\sigma_B(F)$ nicht bekannt; es muß also geschätzt werden. Oft liegt die asymptotische Varianz aber in einer funktionalen Form vor, sodaß $\sigma_B(F)$ durch $\sigma_B(\hat{F}_N)$ konsistent geschätzt werden kann.

Beispiel (Fortsetzung): Wir betrachten nun im folgenden gleich die "erwartungstreue" Version des Schätzers für m_2, d.h. die Größe

$$S_n^2 = (n-1)^{-1}\cdot\sum_{1\le i\le n}(X_i-\bar{X}_\cdot)^2$$

und wollen ein Konfidenzintervall für m_2 konstruieren. In diesem speziellen Beispiel sind die oben angesprochenen Voraussetzungen gegeben, d.h. es handelt sich um einen konsistenten Schätzer für m_2, darüberhinaus ist er asymptotisch normal; es gilt konkret

$$\sqrt{n}\ (S_n^2 - m_2) \dashrightarrow N(\ 0,\ m_4 - m_2^2),$$

hierbei ist $m_4 = E_F(\ X_1 - \mu)^4$ das vierte zentrale Moment von F. Eine geeignete Schätzung für die asymptotische Varianz ist dann der klassische Momentenschätzer

$$\hat{v}_{aa} = \frac{1}{n}\cdot\sum_i\ (X_i-\bar{X}_\cdot)^4 - S_n^4.$$

Damit ergibt sich als "klassisches" asymptotisches Konfidenzintervall für m_2

$$\hat{I}_n^{(as)}(1-\alpha) = [\ S_n^2 - u_{\alpha/2}\cdot\hat{v}_{aa}\ /\sqrt{n},\ S_n^2 + u_{\alpha/2}\cdot\hat{v}_{aa}\ /\sqrt{n}\ ].$$

Meist läßt sich die asymptotische Normalität auch darstellen in der Form

$$(\hat{\theta}_n - \theta) \,/\, \sigma_F(\hat{\theta}_n) \dashrightarrow N(0,1),$$

wobei $\sigma_F^2(\hat{\theta}_n) = Var_F(\hat{\theta}_n)$ die exakte Varianz des Schätzers ist. Besitzt man einen Schätzer für die Varianz von $\hat{\theta}_n$, so ist in der Regel

$$\hat{I}_n^{(fin)}(1-\alpha) = [\, \hat{\theta}_n - u_{\alpha/2} \cdot \hat{v},\ \hat{\theta}_n + u_{\alpha/2} \cdot \hat{v} \,]$$

ebenfalls ein asymptotisches Konfidenzintervall, hier auf der Basis eines Schätzers für die exakte Varianz.

Beispiel (Fortsetzung): Im Falle des Schätzers für m_2 gibt es einen geschlossenen Ausdruck für die Varianz von S_n^2 direkt, nämlich (vgl. z.B. Serfling (1980))

$$Var\ S_n^2 = (m_4 - m_2^2)/n + 2 \cdot m_2^2 / (n(n-1)).$$

Damit unterscheiden sich die wahre und die asymptotische Varianz von $\sqrt{n} \cdot S_n^2$ um einen additiven Term der Ordnung 1/n. Bei der Verwendung der Momentenschätzung von $\sqrt{n} \cdot S_n^2$ wird diese Abweichung korrigiert:

$$\hat{v}_{mf} = \frac{1}{n} \cdot \sum_1 (X_1 - \bar{X}_\cdot)^4 - S_n^4 + 2 \cdot S_n^4/(n-1)$$

ist etwas größer als $\hat{v}_{ma}$; damit deckt das Konfidenzintervall

$$\hat{I}_n^{(fin)}(1-\alpha) = [\, S_n^2 - u_{\alpha/2} \cdot \hat{v}_{mf} / \sqrt{n},\ S_n^2 + u_{\alpha/2} \cdot \hat{v}_{mf} / \sqrt{n} \,]$$

einen größeren Bereich ab, das tatsächliche Niveau ist also höher als bei $I_n^{(as)}$. Wir werden später sehen, daß das Niveau in der Regel immer noch wesentlich kleiner ist als das nominale Niveau 1-α.

In vielen Fällen kann weder die exakte noch die asymptotische Varianz von $\hat{\theta}_n$ in einer Form dargestellt werden, die ad hoc eine "gute" Schätzung dieser Größe aus den vorliegenden Daten ermöglicht. Für solche Fälle ist der Jackknife-Varianzschätzer konzipiert; hier wird in $\hat{I}_n^{(fin)}$ der Schätzer $\hat{\sigma}_{(J)}$ für

die Varianz von $\hat{\theta}_n$ verwendet, so wie es ja bereits im Abschnitt 2.2 angesprochen wurde. Analog kann auch das Bootstrap-Verfahren zur Schätzung der Varianz herangezogen werden: Man bestimme die Bootstrap-Schätzung der Verteilung von $\hat{\theta}_n - \theta$ und berechne deren zweites zentrales Moment $\hat{\sigma}^2_{boot}$.

> Beispiel (Fortsetzung): Aus den ersten Blick scheint eine Jackknife- oder Bootstrap-Schätzung der Varianz grundsätzlich eigentlich nicht erforderlich, da ja eine gute Schätzung für die Varianz mit der Momentenmethode zur Verfügung steht. In der Tat ist aber die oben beschriebene Varianzschätzung $\hat{v}_{mf}$ genau der Bootstrap-Schätzer für die Varianz; eine Simulationsstudie wäre hier also überhaupt nicht erforderlich[4]! Der Jackknife-Schätzer ist ein wenig anders, prinzipiell läßt er sich jedenfalls auch berechnen.

4.2. Die Perzentil-Methode

Der Ansatz zur Konstruktion eines Konfidenzintervalls im letzten Abschnitt zieht seine Berechtigung vorwiegend aus der asymptotischen Normalität der Schätzer. Ein derartiges Intervall ist nur dann effektiv, wenn die tatsächliche Verteilung des Schätzfehlers bestimmte Eigenschaften der Normalverteilung erfüllt, wie etwas Unimodalität, Symmetrie etc.; insbesondere wird jedoch die Anpassung umso schlechter sein, je größer die "Schwänze" der Verteilung sind. Je weniger "normal" die Verteilung der Verzerrung ist, desto stärker wird ihr $\alpha/2$-Fraktil von dem der Grenzverteilung abweichen.

Der "Idealfall" wäre der, daß die wahre Verteilung des Schätzfehlers bekannt wäre. In diesem Fall ließe sich ein zweiseitiges Konfidenzintervall für $\hat{\theta}_n$ exakt angeben unter Verwendung der Fraktile $c(B,F;\alpha) = G_B^{-1}(F, 1-\alpha)$: Es ist

$$P_F \left(c(B,F, 1- \alpha/2) < \hat{\theta}_n - \theta < c(B,F, \alpha/2)) \geq 1-\alpha \right.$$

---

4 Ein Beispiel dafür, daß vor Inbetriebnahme des Rechners immer noch das Gehirn eingeschaltet werden sollte.

und damit

$$I_n^{(B)}(1-\alpha) = [\ \hat{\theta}_n - G_B^{-1}(F,\ \alpha/2)\ ,\ \hat{\theta}_n - G_B^{-1}(F,\ 1-\alpha/2)\ ]$$

ein Konfidenzbereich mit exaktem Niveau $(1-\alpha)$. Die Perzentilmethode schätzt die unbekannten Fraktile in diesem Ausdruck unter Verwendung der Bootstrap-Methode; formal also durch

$$\hat{I}_n^{(B)}(1-\alpha) = [\ \hat{\theta}_n - G_B^{-1}(\hat{F}_n,\ \alpha/2)\ ,\ \hat{\theta}_n - G_B^{-1}(\hat{F}_n,\ 1-\alpha/2)\ ]$$

Beispiel (Fortsetzung): Bei der konkreten Durchführung der Bootstrap-Schätzung des Fraktils einer Verteilung ist zu berücksichtigen, daß die Bootstrap-Verteilung in der Regel ebenfalls nicht zur Verfügung steht, sondern durch eine Monte-Carlo-Studie geschätzt wird. Die Schätzung des Fraktils der Bootstrap-Verteilung ist dann das Fraktil der empirischen Verteilungsfunktion, durch die die Bootstrap-Verteilung approximiert wird. Will man im unserem Beispiel z.B. ein zweiseitiges 90%-Konfidenzintevall für $m_2(F)$ auf diese Weise konstruieren, ist etwa folgende Vorgehensweise denkbar: Man führe 1000 Bootstrap-Replikationen durch so, wie sie in Abschnitt 3.2 zur Konstruktion der Schätzung $\hat{G}_n$ beschrieben wurden. Das $\alpha/2$-Fraktil ist dann die Stelle, an der die Verteilungsfunktion $\hat{G}_n$ den Wert $1-\alpha/2$ überschreitet. Dies ist genau der $N_{boot} \cdot \alpha/2$ - größte Wert (also $1000 \cdot 0.05 = 50$) der durch die Bootstrap-Replikationen gewonnene Wert. Bezeichnen wir diesen Wert mit $\underline{\underline{c}}_B$ und den 50-kleinsten Wert mit $\underline{c}_B$, so ist das Intervall [$\underline{\underline{c}}_B$, $\underline{c}_B$] ein Monte-Carlo-Schätzer für

$$[\ c(B,F,\ 1-\alpha/2\ )\ ,\ c(B,F,\ \alpha/2\ )\ ],$$

d.h. das entsprechende Konfidenzintervall für $m_2(F)$ würde dann lauten

$$\hat{I}_n^{(B)}(1-\alpha) = [\ S_n^2 - \underline{\underline{c}}_B,\ S_n^2 - \underline{c}_B]$$

Simulationsstudien haben gezeigt, daß die Konstruktion von Konfidenzintervallen mittels Bootstrap die am Ende von Abschnitt 3 angesprochene Euphorie nicht unbedingt gerechtfertigt hat (Schenker, 1986). Dies hat zu einer Vielzahl von Versuchen geführt, die Bootstrap-Schätzung der Fraktile zu modifizieren und zu korrigieren (Efron 1979,1981,1982, 1985, 1987). Einen Überblick über diese Ansätze und eine auch theoretisch zufriedenstellende Erklärung der Zusammenhänge zwischen diesen Ansätzen wurde erst in jüngster Zeit präsentiert (Hall (1988)), obwohl bereits früher viele Anzeichen dafür sprachen, daß insbesondere eine der "Alternativen" vielversprechend ist (vgl. etwa Abramovitch und Singh (1985)). Dieses "Alternativ-Konzept" soll im folgenden Abschnitt beschrieben werden.

4.3.Studentisierter Bootstrap

Die Schwäche der Perzentil-Methode besteht darin, daß die Schätzungen der Perzentile bzw. der Fraktile in der Regel relativ schlecht sind. Den wesentlichen Faktor bei dieser Ungenauigkeit stellt dabei nicht die Schätzung der Bootstrap-Verteilung durch die Monte-Carlo-Studie dar, sondern die Abweichung der Bootstrap-Verteilung von der wahren Verteilung. Im folgenden befassen wir uns mit einem Ansatz, mit dem man dieses unangenehme Phänomen abmildern kann. Hierzu sei folgende "Plausibilitätsüberlegung" vorangestellt:

Warum weicht die Bootstrap-Schätzung einer Verteilung überhaupt von der wahren Verteilung ab? In der Tat ist die Abbildung

$$F \dashrightarrow G_R(F;\cdot)$$

schlicht eine Funktion mit der Menge der Verteilungsfunktionen (F) als Definitionsbereich und den Verteilungsfunktionen als Wertebereich. Da diese Abbildung als Funktion von F nicht konstant ist, ist selbstverständlich, daß in der Regel die Funktionswerte $G_R(F;\cdot)$ und $G_R(\hat{F}_n;\cdot)$ verschieden sind. Geht

man davon aus, daß G_R in der Umgebung von F hinreichend "glatt"[5] ist, so hängt der Grad der Abweichung insbesondere davon ab, wie gut die Schätzung $\hat{F}_n$ von F ist (also wie "nahe" $\hat{F}_n$ an F liegt); ändert sich aber G_R relativ wenig, so könnte diese Schätzung auch eine etwas größere Schätzungenauigkeit "vertragen".

Ferner ist leicht einzusehen, daß nicht allein die Kenntnis der Schätzfehlerverteilung die Konstruktion eines Konfidenzintervalls für $\theta(F)$ ermöglicht: Ist $\tau(X) > 0$, so wäre ebenso die Verteilung von

$$R(F;X) = B(F;X) / \tau(X)$$

hierzu geeignet;

$$I_n^{(S)}(1-\alpha)=[\ \hat{\theta}_n-G_R^{-1}(F;1-\alpha/2)\cdot\tau(X),\ \hat{\theta}_n-G_R^{-1}(F;\alpha/2)\cdot\tau(X)]$$

ist ebenfalls ein $(1-\alpha)$-Konfidenzintervall - allerdings in der Regel ein "schlechteres": Seine Länge ist zufällig (nämlich $\tau(X)\cdot(G_R^{-1}(F;1-\alpha/2) - G_R^{-1}(F;\alpha/2))$, also von $\tau(X)$ abhängig) und in der Regel länger als I_α aus Abschnitt 4.2.

Allerdings kann man sich vorstellen, daß eine geeignete Wahl von τ die Verteilung $G_R(F;\cdot)$ "glatter" macht als die von G_B: Bei beiden Verteilungen liegt (zumindest bei "vernünftigen" Schätzern für θ) ihr erstes Moment in der Nähe von 0. Bei G_B schwankt das zweite Moment so, wie sich $Var_F(\hat{\theta}_n)$ als Funktion von F verhält. Hat man nun jedoch einen "vernünftigen" Schätzer für $Var_F(\hat{\theta}_n)$ und verwenden dessen Wurzel als Größe $\tau(X)$, so liegt das zweite Moment von G_R stets in der Größenordnung von 1. Damit sind die ersten beiden Momente von G_R relativ stabil gegen Abweichungen von $\hat{F}_n$ von F. Zumindest kann dann vermutet werden, daß sich dies auch auf die Genauigkeit auswirkt, mit der sich G_R vom Bootstrap schätzen läßt.

---

5 Diese "Glattheit" läßt sich quantifizieren; hierzu ist ein Differentationskalkül für Funktionale erforderlich, wie es mit den Begriffen der Gateau- oder Frechet-Differenzierbarkeit zur Verfügung steht (vgl. z.B. Serfling (1980, sec. 6.2)) und wie es darüberhinaus auch bei der Untersuchung der oben angesprochenen Bootstrap-Konsistenz verwendet wird (Parr (1983), Lohse (1984)).

Das soeben beschriebene Verfahren ist natürlich bereits seit Jahrzehnten bekannt: Haben wir eine Stichproben von unabhängigen, identisch normalverteilten Zufallsgrößen mit unbekannten Mittelwert m und unbekannter Varianz σ^2, so läßt sich ein exaktes Konfidenzintervall für m unter Verwendung der Tatsache konstruieren, daß

$$R(m,\sigma^2;\ X) = \sqrt{n}\cdot(\bar{X}_\bullet - m) \ /\ S \ \sim\ t_{n-1}$$

also t-verteilt mit n-1 Freiheitsgraden ist. In diesem Fall ist der "studentisierte Schätzfehler" sogar völlig unabhängig von den unbekannten Parametern (und damit einer der Spezialfälle, in denen G_R tatsächlich als Funktion des unbekannten Parameters konstant ist). Eine derartige Statistik wird in der Regel als "Pivot" bezeichnet.

Im der allgemeineren Situation, mit der wir uns in dieser Arbeit befassen, ist die studentisierte Größe zwar in der Regel kein Pivot, aber eine Größe, die "näher an einem Pivot" ist als der Schätzfehler selbst.

Die Konstruktion eines Konfidenzintervalls mittels Bootstrap unter Verwendung des studentisierten Schätzfehlers ist dann technisch nicht mehr problematisch: Man führe auf die inzwischen bekannte Weise eine Bootstrap-Schätzung $G_R(\hat{F}_n;\bullet)$ von $G_R(F;\bullet)$ durch und verwende das Intervall

$$\hat{I}_n^{(S)}(1-\alpha) = [\hat{\theta}_n - G_R^{-1}(\hat{F}_n;1-\alpha/2)\cdot\tau(X),\ \hat{\theta}_n - G_R^{-1}(\hat{F}_n;\alpha/2)\cdot\tau(X)].$$

Hierbei ist $\tau^2(X)$ ein Schätzer für $Var_F(\hat{\theta}_n)$. Problematisch wird dieses Verfahren allerdings dann, wenn es einen Schätzer für diese Varianz in "natürlicher" und damit leicht berechenbarer Form nicht gibt: Muß man diese erst selbst mit einem Jackknife- oder gar einem Bootstrap-Verfahren bestimmen, so wird der Rechenaufwand "erheblich" (Bei Bootstrap führt dies zu "ineinandergeschachtelten" Monte-Carlo-Studien: In jedem MC-Schritt zur Bestimmung von $R(\hat{F}_n;X^*)$ ist eine MC-Studie zur Bestimmung von $\hat{Var}_F(\hat{\theta}_n)$ erforderlich!). Der "studentisierte" Bootstrap kann in solchen Fällen in der Tat nicht mehr praktikabel sein (wobei dies natürlich von der Rechengeschwindigkeit der zu Verfügungen stehenden Rechner und den dafür erforderlichen Kosten abhängen mag).

Beispiel (Fotsetzung): In Abschnitt 4.1 haben wir einen Schätzer für die Varianz von S^2 angegeben:

$$\hat{v}_{mf} = \frac{1}{n} \cdot \sum_i (X_i - \bar{X}_\bullet)^4 - S_n^4 + 2 \cdot S_n^4/(n-1)$$

Damit ist ein "studentisierter Schätzfehler" die Größe

$$R(F,X) = (S_n^2 - \sigma_F^2) \; / \; \hat{v}_{mf}.$$

Will man nun die Bootstrap-Verteilung bestimmen, so muß sie wieder durch eine Monte-Carlo-Studie geschätzt werden (in unserem konkreten Beispiel sind wir oben jeweils von einem Umfang von N_{boot}=1000 ausgegangen). In jedem MC-Schritt wird wiederum eine Stichprobe $X_1^*,\ldots,X_{n_*}^*$ mit Zurücklegen aus $X_1,\ldots,X_n$ gezogen, hieraus die Größen S_n^{*2} und $\hat{v}_{mf}^*$ berechnet und schließlich der Wert

$$R(\hat{F}_n, X^*) = (S_n^{*2} - S_n^2) \; / \; \hat{v}_{mf}^*$$

abgespeichert. Die auf diese Weise gewonnen 1000 Werte werden geordnet und der 50-kleinste ($\underline{c}_S$) und der 50-größte ($\bar{c}_S$) bestimmt. Das studentisierte zweiseitige 90%-Bootstrap-Konfidenzintervall ist dann

$$\hat{I}_n^{(S)}(1-\alpha) = [\, S_n^2 - \bar{c}_S \cdot \hat{v}_{mf},\; S_n^2 - \underline{c}_S \cdot \hat{v}_{mf}].$$

4.4. Kürzeste Bootstrap-Konfidenzintervalle

Bei der Konstruktion des zweiseitigen Konfidenzintervalls in Abschnitt 4.2 wie auch in 4.3 wurden das (1- α/2)- und das α/2-Fraktil der Bootstrap-Verteilung herangezogen. Dies ist noch ein "Relikt" aus der Normalverteilungstheorie: Die Normalverteilung ist symmetrisch und unimodal, damit ist das auf diesen beiden Fraktilen basierende Intervall das kürzeste mit Niveau 1-α. Bei der Bootstrap-Verteilung, die ja in der Regel eine bessere Anpassung an die tatsächliche Fehlerverteilung darstellen soll als die Grenzverteilung, muß dies nicht mehr gelten. Damit besteht aber auch kein Grund

mehr, auf beiden Seiten des Konfidenzintervalls möglichst genau die "Verteilungsmasse" $\alpha/2$ abzuschneiden, sondern es wäre sinnvoll, auf der einen Seite α_1 und auf der anderen Seite α_2 so "abzuschneiden", daß noch $\alpha_1+\alpha_2 \leq \alpha$ gilt und das daraus resultierende Intervall möglichst kurz ist.

Beispiel (Fortsetzung): Das Prinzip der Konstruktion läßt sich am einfachsten anhand der konkreten Situation erläutern, wie sie in 4.2 beschrieben wurde. Wir betrachten hierzu nur die Monte-Carlo-Schätzung der Bootstrap-Verteilung: Seien $b_1, b_2, \ldots, b_{N_{boot}}$ die Werte, die bei der Bestimmung von $\hat{S}_n^{*2} - S_n^2$ in der Simulationstudie entstanden sind. Wir ordnen nun diese Werte zunächst der Größe nach an und erhalten Werte

$$b_{(1)}, \ldots, b_{(N_{boot})} \text{ mit } b_{(1)} \leq b_{(2)} \leq \ldots$$

In unserem Fall war $N_{boot}=1000$. Die empirische Verteilungsfunktion $\hat{G}_n$ hat dann an den Stellen $b_{(i)}$ jeweils einen Sprung der Höhe 1/1000. (Prinzipiell kann es Ausnahmen geben: Wenn ein Wert unter $b_1, \ldots, b_{1000}$ mehrfach auftritt, ist die Sprunghöhe das entsprechende Mehrfache von 1/1000. Dieser "Ausnahmefall" macht jedoch wenig Schwierigkeiten, wir werden ihn i.f. ignorieren.) Damit hat die zu $\hat{G}_n$ gehörige Verteilung jeweils an den Punkten b_i die "Masse" 1/1000. 900 Punkte haben damit die "Masse" $1-\alpha$ (wenn $\alpha=0.1$ vorgegeben ist). Ein Intervall $I_i = [b_{(i)}, b_{(i+899)}]$ hat dann die $\hat{G}_n$-Wahrscheinlichkeit $1-\alpha$ - vorausgesetzt, daß I_i ein "vernünftiger" Ausdruck ist (es muß $1 \leq i \leq 101$ gelten). Damit hat das Intervall I_i die Länge $b_{(i+899)} - b_{(i)}$.
Schließlich suchen wir das i_0, für das die Länge von I_i minimal ist. Die Bootstrap-Schätzung für das kürzeste 90%-Konfidenzintervall ist dann

$$\hat{I}_n^{(B;min)}(1-\alpha) = [\, S_n^2 - b_{(i_0+899)},\ S_n^2 - b_{(i_0)}].$$

Führe wir das gleiche Konzept für den studentisierten Bootstrap aus Abschnitt 4.3 durch, so sind $b_1, \ldots, b_{N_{boot}}$ die Wer-

te, die sich als Werte von $R(\hat{F}_n,X^*) = (\overset{*}{S}_n^2 - S_n^2) / \hat{v}_{mf}^*$ ergeben. Analog wie oben ist ein b_{i_0} zu bestimmen, für das die Differenz $b_{(i+899)} - b_{(i)}$ minimal ist. Das kürzeste studentisierte Bootstrap-Konfidenzintervall ist dann

$$\hat{I}_n^{(S;min)}(1-\alpha) = [\, S_n^2 - \hat{v}_{mf} \cdot b_{(i_0+899)},\ S_n^2 - \hat{v}_{mf} \cdot b_{(i_0)}\,].$$

4.5.Vorpivotisierung

Ein Ansatz von Beran (1985, 1987 und 1988) treibt das Konzept der "Studentisierung" auf die Spitze. An dieser Stelle soll nur die Idee kurz skizziert werden, die vollständigen Eigenschaften dieses "Tricks" sind derzeit noch nicht vollständig erforscht.

Im Abschnitt 4.3 haben wir festgestellt, daß die Konstruktion eines exakten Konfidenzintervalles nicht nur bei der Kenntnis der Verteilung des Schätzfehlers möglich ist. Die Studentisierung stellt hierbei eine (zufällige, d.h. von den Beobachtungen X abhängige) Transformation des Schätzfehlers dar, die dafür sorgt, daß die Verteilung der transformierten Größe mittels Bootstrap genauer zu schätzen ist. Ein Konstruktion von Konfidenzintervallen ist möglich, da es sich bei der Transformation um die Multiplikation des Schätzfehlers mit einer positiven (zufälligen) Konstante handelt. Grundsätzlich wäre eine solche Konstruktion auch immer stets dann möglich, wenn der Schätzfehler auf eine andere, nachvollziehbare Weise streng monoton transformiert würde, d.h. wenn

$$R(F;X) = g(X;\hat{\theta}_n - \theta\ ),$$

wobei $g(X;\cdot)$ für (P_F-fast sicher) jede Realisation von X eine monotone Funktion ist. Im Fall der Studentisierung ist

$$g(X,t) := t/\tau(X)$$

wobei $\tau^2(X)$ ein Schätzer für die Varianz von $\hat{\theta}_n$ ist. Wie ist nun $g(X,\cdot)$ zu wählen, damit die Verteilung von R(F;X) in den Umgebung von F möglichst wenig schwankt?

Beran's Ansatz basiert auf den bekannten Tatsache, daß für eine Zufallsvariable Z mit stetiger Verteilungsfunktion H die Größe H(Z) stets eine Gleichverteilung auf dem Einheitsintervall besitzt. Eine Transformation von

$$B(F;X) = \hat{\theta}_n - \theta$$

zu

$$R_0(F;X) = G_B(F;\ B(F;X))$$

würde - vorausgesetzt, G_B wäre stetig - eine Gleichverteilung von R_0 (unabhängig von F) zur Folge haben, d.h. würde B(F;X) "pivotisieren". Aufgrund dieser Überlegung liegt es nun nahe, als Transformation g(X;•) die Bootstrap-Schätzung der Verteilungsfunktion von B(F;X) zu verwenden, d.h. die Größe

$$R(F;X) = G_B(\ \hat{F}_n;\ B(F;X)\ ).$$

Damit ist R(F;X) "ungefähr" rechteckverteilt, die exakte Verteilung (und damit die wesentlichen Abweichungen von eben dieser Rechteckverteilung) lassen sich dann wieder mittels Bootstrap schätzen.

Diese Plausibilitätserklärung hat noch einen kleinen Haken: die Transformation mit $G_B(\hat{F}_n,\bullet)$ ist nur schwach, nicht aber streng monoton; dies ist jedoch nicht allzu problematisch: Entweder man verwendet statt der hier beschriebenen Bootstrap-Schätzung von G_B eine "geglättete" Version, oder man gibt sich mit der Tatsache zufrieden, daß bei nicht allzu kleinem Stichprobenumfang bereits $G_B(\hat{F}_n,\bullet)$ "nahezu" stetig ist (d.h. nur noch Sprünge von sehr geringer Höhe hat) und gibt sich mit einer "konservativen" Abschätzung des Konfidenzintervalls, die dieser Tatsache Rechnung trägt, zufrieden.

Bei der Berechnung der Bootstrap-Verteilung von R(F;X) tritt dann allerdings ein gravierendes Problem auf, das beim Konzept des studentisierten Bootstraps am Rande bereits erwähnt wurde: Hier müssen zwei ineinandergeschachtelte Monte-Carlo-Schleifen durchgeführt werden, der Rechenaufwand potenziert sich also!

Was diese "Vorpivotisierung" bedeutet, läßt sich wiederum an dem in Abschnitt 3.1 zur Motivation herangezogenen "künstlichen Statistiker" erläutern: Das Verfahren ist eine Operationalisierung der Überlegung, was passiert, wenn man den "künstlichen Statistiker" selbst einen Bootstrap durchführen läßt, damit der seine Schätzung verbessert.

Solange Rechenzeitprobleme dem nicht im Wege stehen, kann dieses Konzept der "Vorpivotisierung" noch iteriert werden: Man führe auch noch eine Pivotisierung von R(F;X) durch usw... Es ist derzeit nicht klar, ob ein solcher iterierter Bootstrap überhaupt noch eine Verbesserung der Genauigkeit zur Folge haben kann - außer in einigen künstlich konstruierten Spezialfällen (vgl. Beran (1988)).

4.6.Qualität der Konstruktionsverfahren für Konfidenzintervalle

Bisher wurden in diesem Kapitel nur verschieden Verfahren beschrieben, mit denen sich Konfidenzintervalle konstruieren lassen und bestenfalls Plausibilitätsargumente geliefert, warum das eine Verfahren besser als das andere ist oder nicht. Präzise Qualitätsaussagen wurden bisher nicht präsentiert.

Zunächst ist festzustellen, daß unter relativ schwachen Regularitätsbedingungen alle in Kap. 4 beschriebenen Konstruktionsverfahren tatsächlich asymptotisch das angestrebte Niveau einhalten, d.h. die tatsächliche Überdeckungswahrscheinlichkeit konvergiert gegen das nominale Niveau für $n \longrightarrow \infty$ (vgl. Beran (1984a)). Damit ist aber zunächst nur sichergestellt, daß die Bootstrap-Verfahren asymptotisch nicht schlechter sind als die klassische Methode, der Ausnutzung der asymptotischen Normalität.

Es würde den Rahmen dieser Arbeit überschreiten, die verschiedenen Ansätze hierzu zu diskutieren. Einen guten Überblick stellt hier die Arbeit von Hall (1988) dar, auf die wir an dieser Stelle verweisen wollen und die wir im folgenden zumindest skizzieren. Dennoch sind die dort präsentierten Ergebnisse vom praktischen Standpunkt her recht unbefriedigend, da sie - wie alle mathematischen Qualitätsaussagen im Zusammenhang mit Bootstrap - grundsätzlich asymptotischer Natur sind: Sie sagen etwas darüber aus, wie schnell die

Bootstrap-Schätzung den wahren Wert approximiert bei wachsendem Stichprobenumfang, aber nichts darüber, wie groß die Abweichungen beim "Start", also kleinen Umfängen, bereits sind.

Die Aussagen beziehen sich in der Regel auf die Ordnung der Approximationsgeschwindigkeit in Abhängigkeit vom Stichprobenumfang n. Schätzt man etwa ein Fraktil eines studentisierten Schätzfehlers durch das Fraktil der Standardnormalverteilung, so ist die Abweichung von der Ordnung $1/\sqrt{n}$.[6] Bei der Verwendung des Bootstraps zur Schätzung der Verteilung und damit auch des Fraktils von $B(F;\cdot)$ (vgl. Hall (1988), dieses Verfahren wird dort als "Hybridmethode" bezeichnet) ergibt sich eine Abweichung von der Ordnung $1/n$, also in diesem Sinne eine Verbesserung. Bei der Verwendung des studentisierten Bootstraps schließlich liegt sogar nur eine Abweichung zwischen "theoretischem" und "geschätztem" Fraktil der Größenordnung $n^{-3/2}$ vor.

Bemerkenswerterweise wirken sich diese Approximationsgeschwindigkeiten bei der Überdeckungswahrscheinlichkeit zweiseitiger Konfidenzintervalle unterschiedlich aus (im Gegensatz zu einseitigen Konfidenzintervallen): Hier eliminieren sich Terme zweiter Ordnung, sodaß alle beschriebenen Verfahren das Niveau bis auf eine Größenordnung von n^{-1} approximieren; dies gilt sowohl für die "equal-tailed"-Ansätze, die in den Abschnitten 4.2 und 4.3 beschrieben wurden als auch für den Fall des "kürzesten" Intervalls (vgl. Abschn. 4.4). In den Simulationsstudien des nächsten Kapitels wird sich allerdings zeigen, daß dennoch erhebliche Unterschiede vorliegen[7].

---

6 Eine Folge $\{a_n\}$ von reellen Zahlen bzw. von Zufallsvariablen ist von der Ordnung $c(n)$, wenn die Folge $\{a_n/c(n)\}$ beschränkt bzw. stochastisch beschränkt ist.

7 Die Aussage "$\{a_n\}$ ist von der Ordnung $c(n)$" sagt eben noch nichts über die beschränkenden Konstante $\max\{a_n/c(n)\}$ aus.

5. Numerische Beispiele

5.1. Einfache Zahlenbeispiele

Im letzen Kapitel dieser Arbeit werden wir nun ausschließlich das bisher wiederholt als Beispiel herangezogene Problem der Schätzung des zweiten Moments einer Verteilung betrachten. Zunächst erläutern wir noch einmal anhand konkreter Zahlenbeispiele die Vorgehensweise bei der Durchführung der bisher beschriebenen Prozeduren.

Mit Hilfe der Programmiersprache GAUSS wurden einige Prozeduren erstellt, die einfach die Berechnung von Jackknife-Biaskorrekturen und -Varianzschätzern von Schätzern sowie Bootstrap-Verteilungsschätzern reellwertiger parametrischer Statistiken der Form R(F;X) auf dem PC erlauben.

Wir beginnen mit folgendem Zahlenbeispiel: Gegeben sei die Stichprobe vom Umfang n=10, wie sie in Tab.1 vorgegeben sei. Als Schätzung für das zweite zentrale Moment der diesen Daten zugrundeliegende Verteilung ergibt sich

$$S_n^2 = 1.160$$

und als finiter Momentenschätzer für die Varianz dieser Schätzung

$$\hat{v}_{mf}^2 = 0.323^2 = 0.104$$

Da S_n^2 unverzerrt ist, ist der Jackknife-Schätzer für m_2 identisch mit S_n^2. Als Schätzung für die Jackknife-Varianz ergibt sich

$$v_J^2 = 0.416^2,$$

die Bootstrap-Schätzung der Varianz aufgrund einer Monte-Carlo-Studie vom Umfang N_{boot}=1000 war

$$v_B^2 = 0.337.$$

Zur Berechnung dieser Werte wurde ebenfalls ein kurzes Gauss-Programm erstellt. Man beachte, daß die tatsächliche Bootstrap-Schätzung der Varianz im vorliegenden Beispiel mit $\hat{v}_{mf}^2$ identisch ist und die Abweichung zwischen $\hat{v}_B^2$

und $\hat{v}_{mf}^2$ ausschließlich auf den "Zufall" der Monte-Carlo-Studie zurückzuführen ist.

Die Ergebnisse bei der Verwendung der verschiedenen Ansätze zur Konstruktion von Konfidenzintervallen zum Niveau 90% (also α=0.05) sind in Tab. 2 wiedergegeben. Wo erforderlich, wurde grundsätzlich die Größe $\hat{v}_{mf}^2$ zur Varianzschätzung herangezogen.

Bei der vorliegenden Stichprobe handelt es sich um 10 mit einem Zufallszahlengenerator erzeugte unabhängige Realisationen aus einer N(0,1)-Verteilung (d.h. es gilt m_2=1). Ein zweites Zahlenbeispiel, wieder mit Zufallszahlen aus der gleichen Verteilung, findet sich in Tab. 3.

Man sieht sofort, daß die auf der Basis des studentisierten Bootstraps gewonnenen Konfidenzintervalle erheblich größer sind als die anderen. Tatsächlich wird auch die Varianz von S^2 erheblich unterschätzt (diese ist im Beispiel $2/9 = 0.2222 = (0.471)^2$). Für eine konkrete Aussage über die Qualität des Verfahrens liefern solche Rechenbeispiele allerdings wenig[8].

5.2.Simulationsstudien

Um einen Einblick in die Wirkungsweise der Verfahren zu gewinnen, sind grundsätzlich theoretische Aussagen über ihre Eigenschaften erforderlich. Stehen diese in der erforderlichen Allgemeinheit nicht zur Verfügung, so gibt es keinen anderen Weg, als zu versuchen, sich anhand einiger konkreter Beispiele zu vergewissern, mit welchen Problemen man gegebenenfalls zu rechnen hat. Die Resultate einer Simualtionsstudie ersetzen natürlich nie eine "theoretische" Analyse, man kann mit ihnen bestenfalls "Erfahrungen" über die Wirkungsweise der Prozeduren sammeln. Aus diesem Grund sind natürlich bei einer Simulationsstudie grundsätzlich die verschiedenen denkbaren "Parametersituationen" möglichst breit einzubeziehen.

---

8 Dies gilt natürlich insbesondere, wenn man die Verfahren "anhand konkreter Daten" vergleicht: Es fehlt das Validierungskriterium!

In der in dieser Arbeit beschriebenenen Problematik ist dies besonders "schwierig": Grundsätzlich müßten die verschiedensten Verteilungsparameter F, parametrische Funktionen $\theta(F)$, Schätzer $\hat{\theta}_n$, Stichprobenumfänge n und Niveaus α berücksichtigt werden. Eine solche umfangreiche Untersuchung würde - sofern sie überhaupt denkbar ist - sicherlich den Rahmen dieser Arbeit sprengen. Wir werden uns daher ausschließlich mit dem bereits mehrfach angesprochenen Beispiel auseinandersetzen: Schätzung des zweiten zentralen Moments einer Verteilung durch die empirische Varianz einer Stichprobe.

Die Wahl dieses Beispiels hat seine Ursache in einer Arbeit von Schenker (1986), der es dazu heranzieht, um aufzuzeigen, daß das Bootstrap-Verfahren nicht unbedingt die hervorragenden Eigenschaften besitzt, die es aufgrund der theoretischen Resultate verspricht. Allerdings wird "Bootstrap" dort relativ "unfair" behandelt. Der "naive" Bootstrap, also die Perzentilmethode, wird gegenübergestellt dem tatsächlichen Niveau, das sich erreichen ließe, wenn die tatsächliche Verteilung wirklich bekannt wäre: es werden standardnormalverteilte Daten generiert und die Konfidenzintervalle mit denen durch "klassische" Verfahren (die nur unter Kenntnis der Normalität exakt sind) verglichen. Dies war der Anlaß einer etwas extensiveren Untersuchung (Rothe 1986), die wir hier kurz beschreiben.

Wir untersuchen wieder das in dieser Arbeit durchgehend als Beispiel herangezogenen Problem: Gegeben sei eine Stichprobe vom Umfang n aus einer unbekannten Verteilung mit Verteilungsfunktion F; geschätzt wird das zweite zentrale Moment dieser Verteilung unter Verwendung der empirischen Varianz der Stichprobe. Ziel sei die Angabe eines Konfidenzintervalls für dieses Moment. Verglichen werden

(1) die "klassische" asymptotische Methode unter Verwendung eines Momentenschätzers für die Varianz des Schätzers (Konfidenzintervall $\hat{I}_n^{(fin)}(1-\alpha)$ in Abschnitt 4.1);
(2) der "naive" Bootstrap, also die Verwendung der Perzentilmethode (Konfidenzintervall $\hat{I}_n^{(B)}(1-\alpha)$ in Abschnitt 4.2);
(3) das Konzept des "studentisierten" Bootstraps (Konfidenzintervall $\hat{I}_n^{(S)}(1-\alpha)$ in Abschnitt 4.3).

In (2) und (3) wurde der "equal-tailes-Ansatz" verwendet. Als nominales Ni-

veau für das Konfidenzintervall wurde grundsätzlich 90%, also α=0.1 angesetzt. Die Bootstrap-Schätzungen der jeweiligen Verteilungen basierten jeweils auf einer Monte-Carlo-Studie vom Umfang N_{boot}=1000.

Als datengenerierende Verteilung mit Verteilungsfunktion F wurden drei verschiedene Verteilungen herangezogen: Zunächst die (bereits von Schenker verwendete) N(0,1)-Verteilung, darüberhinaus die Gleichverteilung auf dem Intervall [0,$\sqrt{12}$], U(0,$\sqrt{12}$),sowie schließlich die Standard-Exponentialverteilung E(1) mit der Dichte

$$f(t) = t\, e^{-t} \text{ fürt } t>0,\ 0 \text{ sonst.}$$

Alle diese Verteilungen haben das zweite zentrale Moment 1. Für jede Verteilung wurde das Verhalten der Schätzungen unter Stichprobenumfängen von n=10,20,35 sowie 100 untersucht.

Für jede Kombination Verteilung/Stichprobenumfang wurde eine Monte-Carlo-Studie zur Schätzung der wahren Überdeckungswahrscheinlichkeit des wahren Parameters durchgeführt. Der Umfang dieser Monte-Carlo-Studie betrug jeweils (wie bei Schenker) N_{MC}=1600. Damit besitzt die MC-Schätzung der Überdeckungswahrscheinlichkeit eine Standardabweichung von unter
0.012.

Die Simulationsstudie wurde Anfang 1986 am Rechner des Hochschulrechenzentrums der Universität Dortmund unter Verwendung diverser Fortran-Programme durchgeführt.

Eine Untersuchung der Vorpivotisierungsmethode ist nun in der Tat sehr aufwendig: Waren bereits bei den soeben beschriebenen Studien jeweils Monte-Carlo-Studien von Monte-Carlo-Studien erforderlich, so werden hier drei ineinandergeschachtelte MC-Studien benötigt. Bei der damals durchgeführten Studie wurde dieses Verfahren daher nur für den Fall der N(0,1)-Verteilung und für den Stichprobenumfang n=10 untersucht. Ferner wurden die beiden MC-Studien zur Bestimmung des Bootstrap-Schätzers auf 300 reduziert; N_{MC}=1600 dagegen wurde beibehalten. Weitere Berechnungen wurden nicht durchgeführt; allein die Bestimmung des einen Schätzwertes für das exakte Niveau des vorpivotisierten Bootstrap-Konfidenzintervall machte die Erzeugung und Verrechnung von ca. $1{,}44 \cdot 10^9$ Zufallszahlen erforderlich; auf dem Großrechner

der Universität Dortmund bedeutete dies bereits eine CPU-Rechenzeit von über sechs Stunden!

Die Ergebnisse der MC-Studie in Tab.4 zeigen, daß das klassische asymptotische Verfahren und die Perzentil-Methode in der Tat insbesondere bei kleinen Stichproben sehr schlecht sind, d.h. die nominale Überdeckung von 90% bei weitem nicht erreichen; beide Verfahren unterscheiden sich in ihren realisierten Überdeckungen nur geringfügig. Anders ist die Situation beim studentisierten Bootstrap. Hier ist gegenüber den anderen Methoden eine deutliche Verbesserung der Überdeckung festzustellen; die Verwendung des studentisierten Bootstraps und der damit verbundene Zusatzaufwand scheint also insbesondere bei kleinen Stichproben durchaus gerechtfertigt! Bemerkenswert ist bei diesem Vergleich die schlechte Überdeckung im Falle des klassischen Ansatzes bei einer zugrundeliegenden Exponentialverteilung; hier ist die Verbesserung durch den studentisierten Bootstrap besonders auffällig.

Die Vorpivotisiertung führt in dem einen untersuchten Fall zu einem geschätzten Niveau von 0.874, ist also dem studentisierten Bootstraps ebenbürtig. Der Vorteil der "Vorpivotisierung" scheint sich wohl ohnehin vorwiegend in Situationen zu zeigen, bei der eine Studentisierung nicht direkt gerechtfertigt ist, etwa wenn die Schätzfehlerverteilung nicht asymptotisch normalverteilt ist. In der vorliegenden Situation ist Vorpivotisieren tatsächlich überflüssiger Aufwand.

In einer zweiten, bisher unveröffentlichten Simulationsstudie wurden ferner für den Stichprobenumfang n=10 Untersuchungen über die Lage der Konfidenzintervalle angestellt. Die Berechnungen hierfür wurden an einem PC unter Verwendung von GAUSS durchgeführt; um den Rechenaufwand zeitlich in akzeptablen Grenzen zu halten, haben wir uns auf MC-Stichprobenumfänge von 500 sowohl für N_{boot} als auch für N_{MC} eingeschränkt. Verglichen wurden die klassische Asymptotik, die Verwendung des Jackknife-Konfidenzintervalls sowie vom studentisierten Bootstrap sowohl der "equal-tailed-Ansatz" als auch der Ansatz des kleinsten Konfidenzintervalls; wiederum wurden die bereits oben verwendeten Verteilungen simuliert.

Berechnet wurden die durchschnittlichen linken und rechten Endpunkte der Konfidenzintervalle sowie ihre durchschnittlichen Längen. Die Ergebnisse

sind in Tab. 5 wiedergegeben. Man beachte, daß die Schätzungen für die exakten Niveaus aufgrund des geringeren MC-Umfanges (NMC=500 statt wie oben NMC=1600) ungenauer sind als die Schätzungen in Tab. 4.

Zunächst fällt auf, daß das Jackknife-Verfahren offenbar zumindest etwas besser zu sein scheint als der "naive" Ansatz. Tatsächlich ist dies aber eine "Scheinverbesserung": Wir haben weiter oben bereits vermerkt, daß im vorliegenden Fall Jackknife die Varianz des Schätzers überschätzt; hier überlagern sich also zwei Fehler: Schätzung des Schätzfehlers mit der Normalverteilung (schlecht) unter Verwendung des Jackknife-Varianzschätzers (auch schlecht); diese "Fehler" schwächen sich hier gegenseitig ab; dennoch kann dies nicht als geeignetes Verfahren empfohlen werden. Beim Vergleich der Intervallgrenzen und -längen zeigt sich auch, wo die Schwäche der klassischen Asymptotik liegt: Durch sie werden die Intervallängen gravierend unterschätzt. Dies kann in praktischen Situationen zu ganz erheblich falschen Einschätzungen der Sachlage, ja möglicherweise zu schweren Fehlentscheidungen führen (indem Hypothesen über Parameterkonstellationen "signifikant" zurückgewiesen werden, die tatsächlich nicht hätten abgelehnt werden dürfen). Besonders auffällig ist dies im Falle der Exponentialverteilung, die ja in praktischen Anwendungen keine unbedeutende Rolle spielt (in der Ereignisanalyse ist sie zum Beispiel der elementarste Verteilungsansatz für eine Wartezeit - entsprechend einem konstanten Risiko).

Der Ansatz des "kürzesten Konfidenzintervalls" liefert dagegen ohne Genauigkeitsverlust ein deutlich kleineres Intervall, in den hier analysierten Situationen ergeben sich um 20-30% kürzere Intervalle, ohne daß sich dies in irgendeiner Weise auf die Qualität des Intervalls, also sein exaktes Niveau, auswirkt!

6. Abschließende Bemerkungen

Es ist sicherlich illusorisch, aus einer so kleinen Untersuchung wie der aus Kap. 5 Rückschlüsse über Eigenschaften der verschiedenen Resampling-Verfahren zur Konstruktion von Konfidenzintervallen ziehen zu wollen. Dennoch sind die Unterschiede in der Wirkungsweise dieser Verfahren in den untersuchten Modellen beeindruckend.

Einige sehr allgemeine Ratschläge, die vor der Verwendung der Verfahren beachtet werden sollten, lassen sich aber wohl formulieren:

Die klassische Normalapproximation ist - sofern theoretisch abgesichert - immer noch das Mittel der Wahl, solange der zugrundeliegende Stichprobenumfang hinreichend groß ist. In diesen Situationen sind Bootstrap ohnehin, aber auch Jackknife zu aufwendig im Verhältnis zu dem dabei erzielten Gewinn an Genauigkeit. Problematisch ist dabei allerdings sicherlich die Frage, ab wann ein Stichprobenumfang in dieser Hinsicht als "hinreichend groß" angesehen werden kann; dies hängt ja nicht zuletzt auch von der (nicht bekannten) zugrundeliegenden Verteilung ab.

Das Jackknife-Verfahren sollte als ein Verfahren gesehen werden, das die Verzerrung eines Schätzers schätzt. Als Varianzschätzung ist es eher eine ad-hoc-Prozedur und sollte nach dem derzeitigen Stand der Forschung nur herangezogen werden, wenn andere Verfahren nicht zu Verfügung stehen.

Besteht Anlaß zur Befürchtung, daß asymptotische Verfahren - etwa aufgrund zu geringen Stichprobenumfangs - bei der Konstruktion von Konfidenzbereichen keine Verwendung finden können, so bietet sich eine Bootstrap-Schätzung an. Hier hat sich - sowohl aus theoretischen Überlegungen als auch aufgrund obiger Simulationsstudien - das Verfahren des studentisierten Bootstraps (und des darauf basierenden kürzesten Konfidenzintervalls) als das mit Abstand effektivste herausgestellt. Erfahrungen mit dem vorpivotisierten Bootstrap liegen dabei allerdings bisher kaum vor, da seine Berechnung derzeit noch zu zeitaufwendig ist. Es ist jedoch nach wie vor zu berücksichtigen, daß die Verwendung des studentisierten Bootstraps die Verfügbarkeit eines geeigneten Varianzschätzers für die Schätzstatistik voraussetzt. Aufgrund der Konzeption und der oben diskutierten Plausibilitätsüberlegungen zur Motivierung des Verfahrens kann an dieser Stelle jedoch durchaus ein Jackknife-Schätzer herangezogen werden. Die Verwendung eines "naiven" Bootstraps (also der Perzentilmethode) erscheint nur dann sinnvoll, wenn man auf eine Varianzschätzung verzichten will (bei einer Jackknife-Schätzung der Varianz würde ein studentisierter Bootstrap im Vergleich zur einfachen Perzentilmethode ein Vielfaches an Zeit benötigen, der Faktor liegt etwa in der Größenordnung des vorliegenden Stichprobenumfangs).

Literaturhinweise.

Abramovitch,L. und Singh,K. (1985).Edgeworth corrected pivotal statitics and the bootstrap.
Ann. Statist. 13, 116-132.

Beran,R.J. (1982). Estimated sampling distributions: the bootstrap and competitors.
Ann Statist. 10, 212-225

Beran,R.J. (1984a) Bootstrap methods in statistics.
Jber. der Dt. Math. Verein. 86, 14-30.

Beran,R.J. (1984b). Jackknife approximations to bootstrap estimates.
Ann. Statist. 12, 101-118.

Beran,R.J. (1985): Stochastic procedures: bootstrap and random search methods in statistics.
Bull. Int. Statist. Inst.: Proc. of the 45th Session, Amsterdam 1985.

Beran, R.J. (1987). Prepivoting to reduce level error of confidence sets.
Biometrika 74, 457-468.

Beran, R.J. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refindements.
J. Amer. Statis. Assoc. 83, 687-697.

Bickel,P.J. and Freedman,D.A. (1981). Some asymptotic theory for the bootstrap.
Ann. Statist. 9, 196-217.

Bickel,P.J. and Freedman,D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling.
Ann. Statist. 12, 470-482.

Diaconis,P. und Efron B. (1983). Statistik per Computer: Der Münchhausen-Trick.
Spektrum der Wissenschaften, Juli 1983, 56-71.
(übersetzt aus: Scientific American, Mai 1983.)

Efron,B. (1979) Bootstrap Methods: Another look at the jackknife.
Ann Statist. 7 (1979), 1-26.

Efron,B. (1981). Nonparametric standard errors and confidence intervals.
Canadian Journal of Statistics 9,139-172.

Efron,B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans.
SIAM, Philadelphia.

Efron,B. (1985). Bootstrap confidence intervals for a class of parametric problems.
Biometrika 72, 45-58.

Efron,B. (1987). Better bootstrap confidence intervals.
J. Amer. Statist. Assoc. 82, 139-172.

Efron,B. und Stein,C (1981). The jackknife estimate of variance.
Ann Statist. 9 (1981), 586-596.

Freedman,D.A.(1984). Bootstrapping regression models.
Ann. Statist. 9, 1218-1228.

Hall,P. (1988). Theoretical comparison of bootstrap confidence intervals.
Ann. Statist. 16, 927-954.

Jaeckel,L. (1972). The infinitesimal Jackknife.
Memorandum MM 72-1215-11,Bell Laboratories, Murray Hill.

Jöreskog,K.G. und Sörbom,D. (1988). LISREL VII.
Wiley, New York.

Knuth, D.E. (1981). The Art of Computer Programming -Seminumerical Algorithms.
Reading, Massachusetts, Addison-Wesley.

Klenk,A. und Stute,W. (1986). Bootstrapping of L-statistics.
Statistics and Decisions 5, 77-78.

Lohse,K. (1984). Zur Konsistenz des Bootstrap-Verfahrens.
Dissertation, Hamburg.

Miller,R.G. (1974). The jackknife - a review.
Biometrika 61, 1-17.

Parr,W.C. (1983). A note on the jackknife, the bootstrap and the delta method estimators of bias and variance.
Biometrika 70, 719-722.

Quenoille,M. (1949). Approximate tests of correlation in time series.
J. Roy. Statis. Coc. Ser. B, 11, 18-84.

Quenoille,M. (1956). Notes on bias in estimation.
Biometrika 43, 353-360.

Rothe,G. (1986). Some remarks on bootstrap techniques for constructing confidence intervals.
Statistische Hefte 27, 165-172.

Rothe,G. (1989) Bootstrap in Generalized Linear Models.
Statistische Hefte (ersch. demnächst).

Schenker,N. (1985). Qualms about bootstrap confidence intervals.
J. Amer. Statist. Assoc. 80, 360-361.

Schucany,W., Gray,H. und Owen,O. (1971). On bias reduction in estimation.
J. Amer. Statist. Assoc. 66, 524-533.

Serfling,R.J. (1980). Approximation Theorems of Mathematical Statistics.
Wiley, New York.

Singh,K. (1981). On the asymptotic accuracy of Efron's bootstrap.
Ann. Statist 9, 1187-1195.

Student (1908). The probable error of a mean.
Biometrika 6, 1-25.

Tukey,J. (1958). Bias and confidence sets in not quite large samples.
Ann. Math. Statist. 29, 614 (abstract).

Wu,C.F.J. (1986). Jackknife, Bootstrap and other resampling methods in regression analysis.
Ann Statist. 14, 1261-1294.

Tab.1: Zahlenbeispiel zu Abschn.5.1

i	X_i
1	0.203
2	0.811
3	0.758
4	1.214
5	-1.131
6	1.032
7	0.956
8	-0.502
9	-1.882
10	-0.828

Tab.2: zweiseitige Konfidenzintervalle für $m_2(F)$ zum nominalen Niveau von $1-\alpha=90\%$ zu Datensatz gem. Tab.1

Konstruktions-prinzip	Intervall-Endpunkte links	rechts	Länge
$\hat{I}_n^{(fin)}$	0.629	1.691	1.062
$\hat{I}_n^{(J)}$	0.475	1.845	1.370
$\hat{I}_n^{(B)}$	0.726	1.833	1.106
$\hat{I}_n^{(B;min)}$	0.747	1.835	1.088
$\hat{I}_n^{(S)}$	0.685	2.596	1.911
$\hat{I}_n^{(S;min)}$	0.534	2.333	1.799

Tab.3: Zweites Zahlenbeispiel:

i	x_i
1	0.336
2	0.828
3	0.512
4	-0.232
5	1.451
6	0.509
7	-0.504
8	0.945
9	-1.371
10	0.669

$\hat{m}_2(F)=S^2=0.663$

$\hat{v}_{mf}^2=0.267^2=0.071$

$\hat{v}_J^2=0.348^2$

$\hat{v}_B^2=0.348^2$

Konstruktions-prinzip	Intervall-Endpunkte links	rechts	Länge
$\hat{I}_n(fin)$	0.224	1.101	0.877
$\hat{I}_n(J)$	0.091	1.235	1.144
$\hat{I}_n(B)$	0.246	1.140	0.893
$\hat{I}_n(B;min)$	0.337	1.217	0.880
$\hat{I}_n(S)$	0.229	2.670	2.440
$\hat{I}_n(S;min)$	0.151	2.123	1.972

Tab.4: MC-Schätzungen des exakten Niveaus einiger asymptotischer Konfidenzintervalle

n	$\hat{I}_n$(fin)	$\hat{I}_n$(B)	$\hat{I}_n$(S)	
10	0.713	0.736	0.871	
20	0.799	0.803	0.882	
				N(0,1)
35	0.854	0.835	0.898	
100	0.847	0.881	0.889	
10	0.769	0.783	0.889	
20	0.859	0.847	0.903	
				U(0,√12)
35	0.876	0.883	0.905	
100	0.895	0.885	0.907	
10	0.557	0.554	0.805	
20	0.648	0.625	0.819	
				E(1)
35	0.740	0.709	0.851	
100	0.814	0.791	0.867	

Tab.5: MC-Schätzungen durchschnittlicher Intervallendpunkte und -längen verschiedener Konstruktionskonzepte für Konfidenzintervalle

($n=10$, $N_{BOOT} = N_{MC} = 500$, $\alpha = 0.1$)

Konstruktions-konzept	mittl. linker Endpunkt	mittl. rechter Endpunkt	mittl. Länge	exaktes Niveau	Verteilung
$\hat{I}_n^{(fin)}$	0.4239	1.4796	1.0557	0.702	N(0,1)
$\hat{I}_n^{(J)}$	0.2669	1.6367	1.3699	0.778	
$\hat{I}_n^{(S)}$	0.5007	3.8031	3.3024	0.850	
$\hat{I}_n^{(S;min)}$	0.2958	3.0129	2.7171	0.870	
$\hat{I}_n^{(fin)}$	0.5442	1.4330	0.8888	0.792	U(0,√12)
$\hat{I}_n^{(J)}$	0.4167	1.5605	1.1438	0.882	
$\hat{I}_n^{(S)}$	0.6466	2.5619	1.9153	0.860	
$\hat{I}_n^{(S;min)}$	0.5589	2.0582	1.4993	0.874	
$\hat{I}_n^{(fin)}$	0.2557	1.7447	1.4889	0.570	E(1)
$\hat{I}_n^{(J)}$	0.0281	1.9723	1.9443	0.628	
$\hat{I}_n^{(S)}$	0.1806	16.9565	16.7759	0.800	
$\hat{I}_n^{(S;min)}$	-0.3097	12.2521	12.5618	0.788	

ZUMA-Arbeitsberichte

80/15 Gerhard Arminger, Willibald Nagl, Karl F. Schuessler
Methoden der Analyse zeitbezogener Daten. Vortragsskripten der ZUMA-Arbeitstagung vom 25.09. - 05.10.79

81/07 Erika Brückner, Hans-Peter Kirschner, Rolf Porst, Peter Prüfer, Peter Schmidt
Methodenbericht zum "ALLBUS 1980"

81/19 Manfred Küchler, Thomas P. Wilson, Don H. Zimmerman
Integration von qualitativen und quantitativen Forschungsansätzen

82/03 Gerhard Arminger, Horst Busse, Manfred Küchler
Verallgemeinerte Lineare Modelle in der empirischen Sozialforschung

82/08 Glenn R. Carroll
Dynamic analysis of discrete dependent variables: A didactic essay

82/09 Manfred Küchler
Zur Messung der Stabilität von Wählerpotentialen

82/10 Manfred Küchler
Zur Konstanz der Recallfrage

82/12 Rolf Porst
"ALLBUS 1982" - Systematische Variablenübersicht und erste Ansätze zu einer Kritik des Fragenprogramms

82/13 Peter Ph. Mohler
SAR - Simple AND Retrieval mit dem Siemens-EDT-Textmanipulationsprogramm

82/14 Cornelia Krauth
Vergleichsstudien zum "ALLBUS 1980"

82/21 Werner Hagstotz, Hans-Peter Kirschner, Rolf Porst, Peter Prüfer
Methodenbericht zum "ALLBUS 1982"

83/09 Bernd Wegener
Two approaches to the analysis of judgments of prestige: Interindividual differences and the general scale

83/11 Rolf Porst
Synopse der ALLBUS-Variablen. Die Systematik des ALLBUS-Fragenprogramms und ihre inhaltliche Ausgestaltung im ALLBUS 1980 und ALLBUS 1982

84/01 Manfred Küchler, Peter Ph. Mohler
Qualshop (ZUMA-Arbeitstagung zum "Datenmanagement bei qualitativen Erhebungsverfahren") - Sammlung von Arbeitspapieren und -berichten, Teil I + II

84/02 Bernd Wegener
Gibt es Sozialprestige? Konstruktion und Validität der Magnitude-Prestige-Skala

84/03 Peter Prüfer, Margrit Rexroth
Erfahrungen mit einer Technik zur Bewertung von Interviewerverhalten

84/04 Frank Faulbaum
Ergebnisse der Methodenstudie zur internationalen Vergleichbarkeit von Einstellungsskalen in der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) 1982

84/05 Jürgen Hoffmeyer-Zlotnik
Wohnquartiersbeschreibung. Ein Instrument zur Bestimmung des sozialen Status von Zielhaushalten

84/07 Gabriele Hippler, Hans-Jürgen Hippler
Reducing Refusal Rates in the Case of Threatening Questions: The "Door-in-the-Face" Technique

85/01 Hartmut Esser
Befragtenverhalten als "rationales Handeln" - Zur Erklärung von Antwortverzerrungen in Interviews

85/03 Rolf Porst, Peter Prüfer, Michael Wiedenbeck, Klaus Zeifang
Methodenbericht zum "ALLBUS 1984"

86/01 Dagmar Krebs
Zur Konstruktion von Einstellungsskalen im interkulturellen Vergleich

86/02 Hartmut Esser
Können Befragte lügen? Zum Konzept des "wahren Wertes" im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung

86/03 Bernd Wegener
Prestige and Status as Function of Unit Size

86/04 Frank Faulbaum
Very Soft Modeling: The Logical Specification and Analysis of Complex Process Explanations with Arbitrary Degrees of Underidentification and Variables of Arbitrary Aggregation and Measurement Levels

86/05 Peter Prüfer, Margrit Rexroth (Übersetzung: Dorothy Duncan)
On the Use of the Interaction Coding Technique

86/06 Hans-Peter Kirschner
Zur Kessler-Greenberg-Zerlegung der Varianz der Meßdifferenz zwischen zwei Meßzeitpunkten einer Panel-Befragung

86/07 Georg Erdmann
Ansätze zur Abbildung sozialer Systeme mittels nicht-linearer dynamischer Modelle

86/09 Heiner Ritter
Einige Ergebnisse von Vergleichstests zwischen den PC- und Mainframe-Versionen von SPSS und SAS

86/10 Hans-Peter Kirschner
Der Stichprobenplan zum Projekt ISSP 1985 und seine Realisierung

86/11 Günter Rothe
Bootstrap in generalisierten linearen Modellen

87/01 Klaus Zeifang
Die Test-Retest-Studie zum ALLBUS 1984 - Tabellenband

87/02 Klaus Zeifang
Die Test-Retest-Studie zum ALLBUS 1984 - Abschlußbericht

87/03 Michael Braun
ALLBUS-Bibliographie (6. Fassung, Stand: 30.06.87)

87/04 Barbara Erbslöh, Michael Wiedenbeck
Methodenbericht zum "ALLBUS 1986"

87/05 Norbert Schwarz, Julia Bienias
What Mediates the Impact of Response Alternatives on Behavioral Reports?

87/06 Norbert Schwarz, Fritz Strack, Gesine Müller, Brigitte Chassein
The Range of Response Alternatives May Determine the Meaning of the Question: Further Evidence on Informative Functions of Response Alternatives

87/07 Fritz Strack, Leonard L. Martin, Norbert Schwarz
The Context Paradox in Attitude Surveys: Assimilation or Contrast?

87/08 Gudmund R. Iversen
Introduction to Contextual Analysis

87/09 Seymour Sudman, Norbert Schwarz
Contributions of Cognitive Psychology to Data Collection in Marketing Research

87/10 Norbert Schwarz, Fritz Strack, Denis Hilton, Gabi Naderer
Base-Rates, Representativeness, and the Logic of Conversation

87/11 George F. Bishop, Hans-Jürgen Hippler, Norbert Schwarz, Fritz Strack
A Comparison of Response Effects in Self-Administered and Telephone Surveys

87/12 Norbert Schwarz
Stimmung als Information. Zum Einfluß von Stimmungen und Emotionen auf evaluative Urteile

88/01 Antje Nebel, Fritz Strack, Norbert Schwarz
Tests als Treatment: Wie die psychologische Messung ihren Gegenstand verändert

88/02 Gerd Bohner, Herbert Bless, Norbert Schwarz, Fritz Strack
What Triggers Causal Attributions? The Impact of Valence and Subjective Probability

88/03 Norbert Schwarz, Fritz Strack
The Survey Interview and the Logic of Conversation: Implications for Questionnaire Construction

88/04 Hans-Jürgen Hippler, Norbert Schwarz
"No Opinion"-Filters: A Cognitive Perspective

88/05 Norbert Schwarz, Fritz Strack
Evaluating One's Life: A Judgment of Subjective Well-Being

88/06 Norbert Schwarz, Herbert Bless, Gerd Bohner, Uwe Harlacher, Margit Kellenbenz
Response Scales as Frames of Reference:
The Impact of Frequency Range on Diagnostic Judgments

88/07 Michael Braun
Allbus-Bibliographie
(7. Fassung, Stand: 30.6.88)

88/08 Günter Rothe
Ein Ansatz zur Konstruktion inferenzstatistisch verwertbarer Indices

88/09 Ute Hauck, Reiner Trometer
Methodenbericht
International Social Survey Program - ISSP 1987

88/10 Norbert Schwarz
Assessing frequency reports of mundane behaviors:
Contributions of cognitive psychology to questionnaire construction

88/11 Norbert Schwarz, B. Scheuring (sub.).
Judgments of relationship satisfaction: Inter- and intraindividual comparison strategies as a function of questionnaire strcuture

88/12 Rolf Porst, Michael Schneid
Ausfälle und Verweigerungen bei Panelbefragungen
- Ein Beispiel -

88/13 Cornelia Züll
SPSS-X. Anmerkungen zur Siemens BS2000 Version.

88/14 Michael Schneid
Datenerhebung am PC - Vergleich der Interviewprogramme "interv^{+}" und "THIS"

88/15 Norbert Schwarz, Bettina Scheuring
Die Vergleichsrichtung bestimmt das Ergebnis von Vergleichsprozessen:
Ist - Idealdiskrepanzen in der Partnerwahrnehmung

89/01 Norbert Schwarz, George F. Bishop, Hans-J. Hippler, Fritz Strack
Psychological Sources Of Response Effects In Self-Administered And Telephone Surveys

89/02 Michael Braun, Reiner Trometer, Michael Wiedenbeck
Methodenbericht. Allgemeine Bevölkerungsumfrage der Sozialwissenschaften - ALLBUS 1988 -

89/03 Norbert Schwarz
Feelings as Information:
Informational and Motivational Functions of Affective States