

# Bootstrap und Jackknife

Susumu Shikano

20. Dezember 2005

## I. Kurzdarstellung

Bootstrap und Jackknife sind Verfahren, mit denen Konfidenzintervalle für die statistische Inferenz gebildet werden. Bei der Inferenzstatistik interessiert man sich für die Verteilung von statistischen Kennwerten, um von einer Stichprobe auf die Grundgesamtheit zu schließen. Während bei parametrischen Verfahren die interessierende Verteilung mathematisch hergeleitet wird, rekonstruiert das Bootstrap-Verfahren die Verteilung direkt aus einer Stichprobe, indem viele (Bootstrap-)Stichproben aus der untersuchten Stichprobe ‚mit Zurücklegen‘ gezogen werden. Das Bootstrap-Verfahren wurde auf der Basis des Jackknife-Verfahrens entwickelt, das wegen seiner geringeren Rechenintensität schon länger praktiziert wird. Bei diesem Verfahren werden anstelle der Stichproben ‚mit Zurücklegen‘ mehrere Teilmengen aus den Daten gezogen, um dasselbe Ziel zu erreichen.

Der Vorteil beider Verfahren liegt vor allem in ihrer breiten Anwendbarkeit auf unterschiedlichste statistische Kennwerte. Dies wird dadurch ermöglicht, dass für die Verwendung von Bootstrap und Jackknife in Bezug auf Stichprobe und Grundgesamtheit deutlich weniger Voraussetzungen erfüllt sein müssen als bei den üblichen inferenzstatistischen Verfahren. Dies ist gerade für die Politikwissenschaft attraktiv, da ihre Daten sehr oft die für die herkömmlichen Verfahren erforderlichen Voraussetzungen verletzen. Beispielsweise verwendet man in der Vergleichenden Regierungslehre oft Daten, deren Umfang zu gering ist, um mit Hilfe des Zentralen Grenzwertsatzes Schlüsse von der Stichprobe auf die Grundgesamtheit ziehen zu können. Hier kann man die vorgestellten Verfahren einsetzen, ohne eine Entscheidung für eine bestimmte Verteilung der Grundgesamtheit treffen zu müssen. Andererseits sind sowohl das Bootstrap- als auch das Jackknife-Verfahren stärker von der Stichprobe abhängig, sodass man die Qualität der Stichprobe beachten muss, bevor diese Methoden eingesetzt werden können.

## II. Beschreibung des Verfahrens

### 1. Einführung

Bei der Inferenzstatistik interessiert man sich für die Verteilung von statistischen Kennwerten, um von einer *Stichprobe* auf die *Grundgesamtheit* zu schließen. Wäh-

rend bei herkömmlichen parametrischen Verfahren die interessierende Verteilung mathematisch hergeleitet wird, ‚rekonstruieren‘ das *Bootstrap*- und das *Jackknife*-Verfahren die Stichprobenverteilung der interessierenden Kennwerte unmittelbar aus der konkret vorliegenden Stichprobe. Beide Verfahren basieren auf derselben Logik, aber das Jackknife gilt heute als Approximation des Bootstraps. Im Folgenden wird deshalb zunächst das Bootstrap-Verfahren und danach etwas knapper das Jackknife dargestellt.

## 2. Notation und Formalisierung

### 2.1. Bootstrap: Schritt für Schritt

Unser Ausgangspunkt ist eine Forschungsfrage über einen Parameter  $\theta$  (Durchschnitt, Korrelation usw.). Um eine Aussage über  $\theta$  machen zu können, ziehen wir eine Stichprobe des Umfangs  $n$  aus der Grundgesamtheit  $F$ . Anhand dieser Stichprobe können wir einen Schätzer des Parameters ( $\hat{\theta}$ ) berechnen. Neben dem Wert von  $\hat{\theta}$  interessiert uns immer auch die Genauigkeit von  $\hat{\theta}$  in Bezug auf  $\theta$ , da wir letztendlich eine Aussage über  $\theta$  machen wollen. Wenn es sich bei  $\theta$  um die Linearkombination mehrerer unabhängiger Zufallsvariablen handelt, kann bei parametrischer Inferenz mit Hilfe des *Zentralen Grenzwertsatzes* der Standardfehler von  $\hat{\theta}$ , der mit  $\sigma_{\hat{\theta}}$  bezeichnet wird, mathematisch hergeleitet werden. Nach dem Zentralen Grenzwertsatz konvergiert die Verteilung von  $\hat{\theta}$  bei sehr großen Stichprobenumfängen gegen eine Normalverteilung. Dabei konvergiert die Verteilung schneller, d.h. schon bei mittleren Stichprobenumfängen, wenn die Stichprobe aus einer normalverteilten Grundgesamtheit gezogen wird. Kann bei dem Vorliegen der entsprechenden Bedingungen eine Normalverteilung von  $\hat{\theta}$  angenommen werden, dann kann entsprechend ein *Konfidenzintervall* von  $\hat{\theta}$  berechnet werden.

Was können wir aber machen, wenn die Voraussetzungen zur Anwendung des Zentralen Grenzwertsatzes nicht erfüllt sind, d.h. wenn nicht von einer Normalverteilung von  $\hat{\theta}$  ausgegangen werden kann, oder es nicht möglich ist, den Standardfehler  $\sigma_{\hat{\theta}}$  analytisch herzuleiten? Der Bootstrap löst diese Probleme, indem die konkret vorliegende Stichprobe als beste Approximation der Grundgesamtheit  $F$  betrachtet wird. Hier werden mehrere Stichproben aus der konkret vorliegenden Stichprobe gezogen. Die folgenden Schritte werden dabei vorgenommen:

1. Stichproben des Umfangs  $n$  werden  $B$ -mal aus der konkret vorliegenden *Stichprobe ‚mit Zurücklegen‘* gezogen. Wir nennen die hier gezogenen Stichproben ‚Bootstrap-Stichproben‘.
2. Für die einzelnen Bootstrap-Stichproben wird der interessierende Parameter ( $\hat{\theta}_b^*$ ) geschätzt.
3. Von der Verteilung von  $\hat{\theta}^*$  wird das Konfidenzintervall gebildet.

Nehmen wir ein hypothetisches Beispiel. Wir interessieren uns für die durchschnittliche Selbsteinstufung der Studierenden auf einer Links-Rechts-Skala von

null bis zehn. Wir haben zehn zufällig ausgewählte Studierende befragt und einen Datensatz gewonnen, der auf der ersten Zeile von Tabelle 1 zu finden ist. Diese Daten ergeben einen Durchschnitt von 3,2. Aber wie genau ist dieser Mittelwert? Um das abschätzen zu können, ziehen wir mehrere Bootstrap-Stichproben aus den Daten. In der ersten Bootstrap-Stichprobe ( $b = 1$ ) taucht der Wert 6 dreimal auf, obwohl er nur zweimal in der Stichprobe vorkommt. Dies kann durch die Stichprobenziehung ‚mit Zurücklegen‘ möglich sein. Dafür taucht der Wert 0 in der ersten Bootstrap-Stichprobe gar nicht auf. Noch extremer ist die zweite Bootstrap-Stichprobe ( $b = 2$ ), in der nur die Werte 5, 6 und 8 vorkommen. Nachdem wir 100-mal diesen Vorgang wiederholt haben, können wir für diese 100 Bootstrap-Stichproben jeweils den Durchschnitt berechnen. Von diesen  $\hat{\theta}_b^*$ , wobei  $b = 1, \dots, 100$ , gewinnen wir eine Verteilung von  $\hat{\theta}^*$ , um Konfidenzintervall um den oben geschätzten Wert 3,2 zu bilden.

konkret vorliegende Stichprobe		{6,2,1,1,8,1,2,0,6,5}	$\hat{\theta} = 3,2$
Bootstrap-Stichproben	( $b = 1$ )	{2,1,1,1,8,5,2,6,6,6}	$\hat{\theta}_1^* = 3,8$
	( $b = 2$ )	{5,6,8,6,5,6,5,8,6,6}	$\hat{\theta}_2^* = 6,1$
	( $b = 3$ )	{6,2,1,6,1,1,1,6,1,6}	$\hat{\theta}_3^* = 3,1$
		⋮	
	( $b = 100$ )	{2,2,8,6,0,0,1,6,8,1}	$\hat{\theta}_{100}^* = 3,4$

Tabelle 1: Ziehung der Bootstrap-Stichproben

Zur Bildung des Konfidenzintervalls wurden in der Literatur unterschiedliche Methoden vorgeschlagen. Im Folgenden konzentrieren wir uns auf vier Methoden, die in den Standardbefehlen in Stata für Bootstrap (bs, bstrap, bstat) integriert sind:

*Normale Approximationsmethode* (N im Stata-Output): Die Bildung des Konfidenzintervalls nach dieser Methode erfolgt analog zur parametrischen Methode. Anhand der Verteilung von  $\hat{\theta}_b^*$  wird der Standardfehler geschätzt:

$$\hat{\sigma}_{\hat{\theta}}^* = \left( \frac{\sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*)^2}{B-1} \right)^{1/2}, \text{ wobei } \hat{\theta}_{(\cdot)}^* = \frac{\sum_{b=1}^B \hat{\theta}_b^*}{B} \quad (1)$$

Damit kann man eine Normalverteilung um  $\hat{\theta}$  und dementsprechend das Konfidenzintervall ermitteln. Zu beachten ist, dass die Normalverteilung nicht auf  $\hat{\theta}_{(\cdot)}^*$ , sondern auf  $\hat{\theta}$  zentriert wird. Bootstrap ist hauptsächlich für die statistische Inferenz entwickelt worden. Ziel ist dabei nicht die Punktschätzung, sondern die Bildung eines Konfidenzintervalls. Ein Nachteil dieser Methode besteht in der Annahme der Normalverteilung von  $\hat{\theta}$ , was nicht immer angemessen ist.

*Perzentil-Methode* (P): Diese Methode ist sehr einfach und intuitiv zu verstehen. Das Konfidenzintervall wird nach der Verteilung der Statistik von Bootstrap-Stichproben geschätzt. Für das 95%-Konfidenzintervall werden jeweils das 2,5%-

als untere und das 97,5%-Perzentil als oberen Wert genommen. Diese Schätzung, anders als die normale Approximationsmethode, ist völlig frei von parametrischen Annahmen. Dafür müssen aber mehr Bootstrap-Stichproben gezogen werden. Während 200 Stichproben für die normale Approximationsmethode reichen, werden für diese Methode 1000 Stichproben empfohlen (Efron und Tibshirani 1986: 72). Das größere Problem besteht jedoch in der Voraussetzung der Methode, dass  $\hat{\theta}^*$  nicht schief verteilt ist. Weiterhin soll es keine Verzerrung geben, d.h.  $bias = \hat{\theta}_{(.)}^* - \hat{\theta} = 0$ . Falls diese Voraussetzungen nicht erfüllt sind, muss eine Korrektur der Grenzen des Konfidenzintervalls vorgenommen werden. Zwei Methoden für die Korrektur werden im Folgenden vorgestellt.

*bias-corrected Perzentil-Methode* (BC) und *bias-corrected and accelerated Methode* (BCa): Anstatt von keiner Verzerrung von  $\hat{\theta}^*$  auszugehen, wird bei der BC-Methode ermittelt, wie sich die Abweichungen zwischen  $\hat{\theta}_{(.)}^*$  und  $\hat{\theta}$  verteilen. Hierbei wird eine bestimmte Verteilung angenommen (in der Regel Normalverteilung). Der ermittelten Verteilung entsprechend wird das verzerrte Perzentil von  $\hat{\theta}^*$  auf  $\hat{\theta}$  zentriert. Als weitere Entwicklung dieser Methode gibt es auch die ‚bias-corrected and accelerated Methode (BCa)‘, bei der die Unterschiede in den Schiefen der Verteilungen von  $\hat{\theta}^*$  und  $\hat{\theta}$  berücksichtigt werden. Auf die ausführliche Darstellung der beiden Methoden wird aber aus Platzgründen verzichtet. Interessierte Leser seien auf Mooney und Duval (1993: 37 ff.) und Efron und Tibshirani (1993: Kap. 14) verwiesen. Es ist zu beachten, dass für diese Methode wieder eine parametrische Annahme eingeführt wird.

## 2.2. Bootstrap in der Regressionsanalyse

Die Vorgehensweise, die oben dargestellt wurde, kann für unterschiedliche Parameter eingesetzt werden. Die Struktur des Modells, dem die interessierenden Parameter zugrunde liegen, erfordert jeweils eine spezifische Art und Weise der Anwendung. Im Folgenden wird der Einsatz des Bootstraps bei der Regressionsanalyse, die als eine der am häufigsten eingesetzten statistischen Methoden gilt, vorgestellt.

Im Regressionsmodell:

$$y_i = \beta x_i + \epsilon_i \quad (2)$$

wird der Vektor der Parameter  $\beta$  anhand der abhängigen Variable  $y_i$  und des Vektors der unabhängigen Variable  $x_i$  geschätzt, wobei  $\epsilon_i$  der Fehlerterm ist. Dieses Modell setzt sich aus deterministischen Komponenten ( $\beta x_i$ ) und stochastischen Komponenten ( $\epsilon_i$ ) zusammen. Bei einer herkömmlichen parametrischen Schätzung nach dem Kriterium der kleinsten Quadrate werden die einzelnen  $\epsilon_i$  als Realisierungen mehrerer unabhängiger, normalverteilter und identisch verteilter Zufallsvariablen angenommen. Während man durch die Verwendung des Bootstraps diese Annahme der Normalverteilung umgehen kann, muss man eine weitere Annahme beachten: Nur  $\epsilon_i$  ist stochastisch und die sonstigen Komponenten sind deterministisch. Berücksichtigt man diese Annahme, so darf das Bootstrap-Verfahren nicht

fallweise eingesetzt werden, d.h. die Bootstrap-Stichproben werden nicht aus den in der Stichprobe beobachteten Werten von  $Y$  und  $X$  gebildet, sondern nur der Fehlerterm  $\epsilon_i$  wird wiederholt zufällig ausgewählt.

1. Nach dem Kriterium der kleinsten Quadrate wird  $\hat{\beta}$  anhand der untersuchten Stichprobe geschätzt, womit auch  $\hat{\epsilon}_i = y_i - \hat{\beta}x_i$  berechnet werden kann.
2. Aus  $\hat{\epsilon}$  wird die Bootstrap-Stichprobe  $B$ -mal gezogen, so dass eine Verteilung  $\epsilon_{bi}^*$  entsteht.
3. Für jede einzelne Bootstrap-Stichprobe wird die durch die Bootstrap-Methode berechnete abhängige Variable  $y_{bi}^* = \hat{\beta}x_i + \epsilon_{bi}^*$  geschätzt.
4. Anhand von  $y_{bi}^*$  und  $x_i$  wird  $B$ -mal eine Regressionsanalyse durchgeführt, sodass  $\hat{\beta}_b^*$  berechnet werden kann.
5. Anhand der Verteilung von  $\hat{\beta}_b^*$  wird das Konfidenzintervall von  $\hat{\beta}$  berechnet.

In der sozialwissenschaftlichen Praxis verwendet man aber oft  $x_i$  aus Umfragedaten, die stochastische Komponenten enthalten können (Mooney und Duval 1993: 17). In diesem Fall sollte man eher fallweise und nicht aus den Residuen die Stichproben ziehen (Efron und Tibshirani 1993: 113).

### 2.3. Jackknife

Das Bootstrap-Verfahren wurde auf der Basis des Jackknife-Verfahrens entwickelt, das wegen seiner geringeren Rechenintensität schon länger praktiziert wird. Anstelle von Stichproben ‚mit Zurücklegen‘ werden mehrere Teilmengen aus den Daten gezogen, um dasselbe Ziel zu erreichen. Umgekehrt formuliert werden für jeden Versuch bestimmte Teilmengen von den Daten ausgelassen; deshalb heißt das Verfahren ‚Jackknife‘ (Klappmesser). Dabei bleibt die Grundidee die Gleiche wie beim Bootstrap: Die empirisch untersuchte Stichprobe ist die beste Approximation der Grundgesamtheit und daraus sollen mehrere Stichproben gezogen werden. Im Folgenden werden die einzelnen Schritte dargestellt:

1. Die konkret vorliegende Stichprobe des Umfangs  $n$  wird in  $g$  Gruppen mit der Größe von  $h$  zerlegt. Es gilt:  $n = gh$ . Meistens wird  $h$  auf eins gesetzt, sodass  $g = n$ . So ist es auch beim Stata-Befehl ‚jknife‘ in der Standardeinstellung.
2. Von den insgesamt  $g$  Gruppen wird die  $i$ -te Gruppe,  $i = 1, \dots, g$ , aus der Stichprobe entfernt und der interessierende Parameter  $\hat{\theta}_{-i}$  wird für die verbliebene Teilmenge der Stichprobe mit dem Umfang  $n - h = (g - 1)h$  berechnet.
3. Schritt 2 wird für alle  $g$  Gruppen wiederholt.

4. Der folgende Pseudo-Wert wird für jeden Versuch berechnet:

$$\tilde{\theta}_i = g\hat{\theta} - (g-1)\hat{\theta}_{-i}$$

5. Der Jackknife-Schätzer  $\tilde{\theta}$  ist der Durchschnitt der  $\tilde{\theta}_i$ , d.h.  $\tilde{\theta} = \frac{1}{g} \sum_{i=1}^g \tilde{\theta}_i$

Anhand der Standardfehler von  $\tilde{\theta}_i$  kann man eine t-Verteilung um  $\tilde{\theta}$  berechnen, auf deren Basis das Konfidenzintervall ermittelt werden kann. Während dies der normalen Approximationsmethode beim Bootstrap ähnelt, gilt das Konfidenzintervall dieses Verfahrens als weniger genau als das des Bootstraps. Tatsächlich benutzt das Jackknife-Verfahren nur Teile der Informationen, die das Bootstrap berücksichtigen kann, weshalb dieses Verfahren als Approximation zum Bootstrap gilt. Diese Approximation ist linearer Natur, daher kann das Jackknife für die Bildung eines Konfidenzintervalls für eine nichtlineare Statistik (z.B. den Korrelationskoeffizienten) ziemlich ineffizient sein (Efron und Tibshirani 1993: 146). Aus diesem Grund wird das Verfahren seltener eingesetzt, seitdem das Bootstrap-Verfahren durch leistungsstärkere Rechner handhabbar wurde. Dieses Verfahren gibt aber noch Informationen an, die das Bootstrap nicht liefern kann. Beim Jackknife kann man bei jeder Wiederholung deutlich identifizieren, welches Element der Stichprobe ausgelassen wurde. Aus diesem Grund ist das Jackknife auch heute noch weiterhin sehr nützlich bei der Ermittlung und Beurteilung von Verzerrungen, die durch einzelne Ausreißer in der Stichprobe verursacht werden. Hierfür wird der Pseudo-Wert  $\tilde{\theta}_i$  für jede Beobachtung betrachtet.

#### 2.4. Unterschiede zu weiteren Methoden

Das Bootstrap wird oft mit dem Monte-Carlo(MC)-Verfahren verglichen, welches die Idee der Simulation von neuen Stichproben angeht, die erst durch die gesteigerte Leistungsfähigkeit des Computers ermöglicht wurde (zum MC-Verfahren siehe „Bayesianische Datenanalyse“). Grundsätzlich sind aber die beiden Verfahren als völlig unterschiedlich zu verstehen. Der größte Unterschied besteht im Ausgangspunkt der Analyse. Während das MC-Verfahren vor der Analyse eine bestimmte Verteilung annimmt, ist eine solche Entscheidung beim Bootstrap nicht notwendig. Denn das Bootstrap verwendet nur die Information der empirisch untersuchten Stichprobe. Andererseits ist das Bootstrap bei seinem Ergebnis stärker von der Stichprobe abhängig, während das Ergebnis beim MC-Verfahren unabhängig von den Daten generalisierbar ist. Die Überprüfung der Leistung des Bootstrap-Verfahrens durch das MC-Verfahren verdeutlicht dies. Dort wird zunächst eine bestimmte Verteilung für  $F$  angenommen, aus der durch das MC-Verfahren mehrere Stichproben gezogen werden. Das direkt von den Stichproben gebildete Konfidenzintervall dient hier als Norm. Von jeder Stichprobe werden dann mehrere Bootstrap-Stichproben gezogen, womit das Konfidenzintervall für jede Stichprobe gebildet wird. Diese Bootstrap-Konfidenzintervalle werden nun mit der bereits berechneten Norm verglichen, um die Wahrscheinlichkeit der Fehlentscheidung beim Bootstrap zu ermitteln (Mooney und Duval 1993: 27 ff.).

### 3. Modellannahmen und Diagnostik

Die beiden Verfahren müssen weniger Voraussetzungen erfüllen als die meisten herkömmlichen statistischen Verfahren. Trotzdem gibt es einige Voraussetzungen für ihren Einsatz.

In beiden Fällen geht man davon aus, dass die untersuchten Daten die Grundgesamtheit gut abbilden. Die wichtigste Voraussetzung ist deshalb eine gute Stichprobe. Sie sollte zufällig gezogen werden. Außerdem sollten die Daten unabhängig und identisch verteilt sein. Ist dies nicht der Fall, approximiert die Verteilung von Bootstrap-Stichproben nicht  $F$ . Zeitreihendaten sind typische Beispiele hierfür. Für solche Daten wird als Lösung, obwohl nicht unumstritten (z.B. Hall und Horowitz 1996), das ‚Block‘-Bootstrap vorgeschlagen, bei dem Bootstrap-Stichproben nicht direkt aus der gesamten Stichprobe, sondern blockweise gezogen werden. Für Zeitreihendaten werden die angrenzenden Beobachtungen gruppiert und von dort die Stichproben gezogen (vgl. ‚moving blocks bootstrap‘ in Efron und Tibshirani 1993: 99ff.).

Bei der Größe der ursprünglichen Stichprobe muss man nicht den Zentralen Grenzwertsatz ( $n \geq 30$ ) beachten, da für die beiden Methoden keine Normalverteilung vorausgesetzt werden muss. Wie Mooney und Duval (1993: 21) bei mehreren Studien zeigen, kann sogar mit einem  $n$  von 10 bis 20 eine zuverlässige Inferenz möglich sein. Man sollte aber dabei auch die Variablen, die für die Parameterschätzung verwendet werden, berücksichtigen. Da das Bootstrap- und das Jackknife-Verfahren nur die Informationen aus der Stichprobe verwenden, müssen die Daten genügend Variation in jeder Variable besitzen, um die Grundgesamtheit adäquat abzubilden.

Im Zusammenhang mit den einzelnen Methoden für die Bildung eines Konfidenzintervalls beim Bootstrap-Verfahren wurden bereits oben die Annahmen erörtert. Während es generelle Empfehlungen für die BCa gibt (z.B. Mooney und Krause 1997: 108), haben die anderen Methoden (Normale Approximationsmethode, Perzentil-Methode) vor allem in ihrer Handhabbarkeit einen Vorzug. Deshalb sollte die angemessene Methode abhängig vom Untersuchungskontext gewählt werden, wobei die folgenden Kriterien beachtet werden sollen: 1. die Normalverteilung von  $\hat{\theta}^*$  und 2.  $bias = 0$ . Eine Auswahlstrategie kann wie folgt aussehen: Man wählt zunächst die Normale Approximationsmethode. Wenn man bei der Diagnostik feststellt, dass das erste Kriterium nicht erfüllt ist, so wählt man die Perzentil-Methode. Falls die weitere Diagnose feststellt, dass auch das zweite Kriterium nicht erfüllt ist, wählt man die BC bzw. BCa-Methode. Ein Histogramm der Bootstrap-Verteilung zusammen mit einer Normalverteilungskurve als Referenz ist als Diagnoseinstrument dieser beiden Kriterien sehr nützlich. Abbildung 1 (c) im Anwendungsbeispiel dieses Kapitels zeigt z.B., dass sogar beide Kriterien nicht immer erfüllt sein müssen. In so einem Fall soll man sich daher auf die BC bzw. BCa-Methode zur Konstruktion von Konfidenzintervallen stützen.

#### **4. Datenstruktur und praktische Hinweise**

Beiden Verfahren sind in Bezug auf die Datenstruktur wegen der geringen Anzahl der Voraussetzungen vielseitig anwendbar. Sie können auch auf Daten einer Vollerhebung angewandt werden, die bei vielen vergleichenden Untersuchungsdesigns mit kleiner Fallzahl typisch sind. In diesem Zusammenhang müssen die Leser – wegen des Charakters des Bandes als Lehrbuch – ausdrücklich darauf hingewiesen werden, dass die Durchführung von Signifikanztests mit Daten aus einer Vollerhebung nicht unumstritten sind (Behnke 2005; Broscheid und Gschwend 2005). Daten aus einer Vollerhebung enthalten keine Stochastizität aufgrund der Stichprobenziehung. Trotzdem wird das Bootstrap im Anwendungsbeispiel dieses Kapitels auf Lijphart'sche Daten angewandt, die als Vollerhebung gelten. Dies wird nach der Ansicht des Autors wegen des Messfehlers für notwendig gehalten. Bei sozialwissenschaftlichen Untersuchungen kann meistens nicht von einer fehlerfreien Messung ausgegangen werden, was auch bei den Lijphart'schen Daten der Fall zu sein scheint. Deshalb sollten Schätzergebnisse aus Daten mit potenziellen Messfehlern mit Hilfe von Konfidenzintervalle vorsichtiger interpretiert werden.

#### **5. Innermethodische Kritik**

Das größte Hindernis für eine weite Verbreitung der beiden Verfahren war die hohe Rechenintensität; dies trifft in besonderer Weise auf das Bootstrap-Verfahren zu. Dies scheint aber aufgrund der gesteigerten Leistungsfähigkeit moderner Computer kein Problem mehr zu sein. Andererseits sind das Bootstrap und das Jackknife bei ihrem Ergebnis stärker von der Stichprobe abhängig, sodass man vor der Verwendung der Verfahren die Qualität der Stichprobe beachten muss. Dies ergibt eine gewisse Einschränkung des Vorteils dieser Verfahren bezüglich des Stichprobenumfangs. Während der Zentrale Grenzwertsatz hier nicht gelten muss, muss die Stichprobe gut die Grundgesamtheit abbilden. Dies ist bei einer Stichprobe mit ziemlich kleinem Umfang schwer realisierbar.

#### **6. Anwendungsgebiete**

Der Vorteil beider Verfahren liegt vor allem in ihrer Anwendbarkeit auf die unterschiedlichsten statistischen Kennwerte, die durch deutlich weniger notwendige Voraussetzungen für die Stichprobe und die Grundgesamtheit ermöglicht wird. Dies ist gerade für die Politikwissenschaft attraktiv, da entsprechende Daten sehr oft die für die herkömmlichen Verfahren erforderlichen Voraussetzungen verletzen. Beispielsweise wird man in der Vergleichenden Regierungslehre oft mit Daten eines kleinen Stichprobenumfangs konfrontiert, die der Voraussetzung des Zentralen Grenzwertsatzes nicht entsprechen. Hier kann man die vorgestellten Methoden einsetzen, ohne vorher eine Entscheidung für eine bestimmte Verteilung der Grundgesamtheit treffen zu müssen. Auch bei der Untersuchung mit den Umfragedaten eines großen Stichprobenumfangs können die beiden Methoden zum Einsatz

kommen, wenn es dabei um eine komplexe Stichprobe handelt (siehe „Schätzer für komplexe Stichproben“).

Zudem sind beide Verfahren auch in Situationen einsetzbar, in denen keine statistische Theorie für den interessierenden Parameter bekannt ist, oder in denen die Schätzung eines Konfidenzintervalls empirisch, z.B. wegen der nichtinvertierbaren Hesse-Matrix, nicht möglich ist.

### III. Anwendungsbeispiel

Im Folgenden werden die Daten aus der Studie von Lijphart (1999) über die Typologie der Demokratie herangezogen. Lijphart sammelte seine Daten für 36 Länder, die im Jahr 1996 mindestens seit 19 Jahren als Demokratie gelten. Um verschiedene Muster der Demokratie zu untersuchen, erhob er zehn Variablen. Mit Hilfe der Hauptkomponentenanalyse stellt er fest, dass es zwei Dimensionen hinter den zehn Merkmalen gibt (S. 245 ff.). Taagepera (2003) kritisiert dieses Ergebnis von Lijphart mit dem Argument, dass die beiden Dimensionen inhaltlich nicht konsistent seien. Taagepera bezeichnet vor allem die beiden Messungen, ‚Interessengruppen-Pluralismus‘ und ‚Unabhängigkeit der Zentralbank‘, als logisch fremd in der jeweiligen Dimension. Aus dieser Kritik stellt sich die Frage, ob die beiden Dimensionen tatsächlich ausreichend sind, um die möglichen Modelle der Demokratie ausreichend abzubilden. Die Hypothese von Lijphart wäre hier, dass beide Dimensionen genügen ( $H_A$ ). Die Gegen-Hypothese wäre hingegen, dass mindestens noch eine Dimension notwendig ist ( $H_B$ ).

Das von Lijphart verwendete Kriterium zur Bestimmung der Zahl der Dimensionen, was hier nicht weiter diskutiert wird, obwohl es nicht unumstritten ist, ist die Größe der Eigenwerte ( $\lambda$ ): Die Dimensionen, deren Eigenwert größer als eins ist, sollen extrahiert werden. Die Gegen-Hypothese kann dementsprechend umformuliert werden:  $H_B : \lambda_3 > 1$ . Abbildung 1 (a) zeigt die Eigenwerte aller Dimensionen. Während nur die ersten beiden Dimensionen den Wert eins überschreiten, verfehlt die dritte Dimension knapp diesen Schwellenwert. Hier würde uns interessieren, ob der Eigenwert der dritten Dimension tatsächlich unter eins liegt. Ist dies nicht der Fall, so muss die dreidimensionale Lösung gewählt werden. Leider schweigt das Modell der Hauptkomponentenanalyse über die Verteilung der Eigenwerte. Zudem erfüllen die Daten von Lijphart die Voraussetzung für die Faktorenanalyse nicht, dass die Variablen normal verteilt sein sollen. Diese Probleme sollen durch das Bootstrap-Verfahren gelöst werden.

Nun werden 2000 Bootstrap-Stichproben von den Daten mit 36 Ländern gezogen und für jede Stichprobe werden die Eigenwerte der Korrelationsmatrix berechnet. Abbildung 1 (b) zeigt die Verteilung der Eigenwerte der Dimensionen 1 bis 4 mittels eines Boxplots. Während die Eigenwerte der 1. und 2. Dimension deutlich über eins liegen, sind die Eigenwerte der 3. Dimension der Bootstrap-Stichproben ober- und unterhalb des Grenzwerts 1. Sogar der Medianwert liegt über eins. Abbildung 1 (c) zeigt noch detaillierter diese Verteilung. Die gepunktete Linie

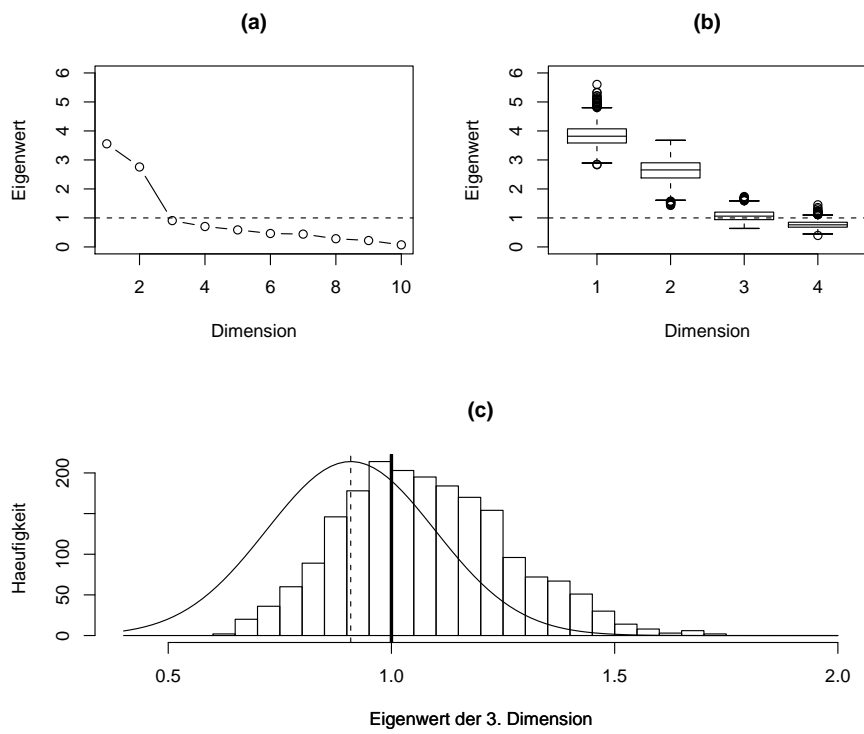


Abbildung 1: Überprüfung der Eigenwerte

Reps	Observed	Bias	Std. Err.	90% Conf. Interval		
2000	.9088015	.169852	.1880732	.5993052	1.218298	(N)
				.7910792	1.408162	(P)
				.6189165	1.040865	(BC)
				.6424747	1.043485	(BCa)

Tabelle 2: Stata-Output

ist der geschätzte Eigenwert aufgrund der ursprünglichen Daten von Lijphart. Darauf wird die Normalverteilung zentriert eingezeichnet, denn das Konfidenzintervall nach der normalen Approximationsmethode beruht darauf. Tabelle 2 zeigt einen Ausschnitt des Stata-Outputs, in dem Konfidenzintervalle nach drei verschiedenen Methoden zu finden sind. Da hier eine ‚einseitige‘ Überprüfung auf dem 95%-Niveau beabsichtigt wird, wird das Signifikanzniveau auf 90% gesetzt, wobei nur die Obergrenze des Intervalls interessiert. Das Histogramm in 1 (c) zeigt, dass der Bootstrap-Schätzer stark verzerrt ist. Daher sollte man sich bei der Entscheidung,  $H_B$  zu behalten oder abzulehnen, nicht auf die ersten beiden Methoden, die normale Approximations- bzw. die Perzentil-Methode, stützen. Die Konfidenzintervalle nach BC bzw. BCa zeigen, dass wir nicht die Hypothese ablehnen können, dass der Eigenwert größer als eins ist. Lijphart hätte demnach eine weitere Dimension berücksichtigen müssen. Wenn man aus den Daten drei Dimensionen extrahiert und nach der Varimax-Methode rotiert, korrelieren die beiden Variablen ‚Disproportionalität‘ und ‚Pluralismus‘ stärker mit der neu extrahierten dritten Dimension als mit der ersten Dimension. Dies unterstützt teilweise die oben genannte Kritik von Taagepera, wonach keine strukturelle Verbindung des ‚Interessengruppen-Pluralismus‘ mit den weiteren Indikatoren der Lijphart’schen ‚Executive-Parties‘-Dimensionen bestehen soll.

## IV. Kommentierte Literaturempfehlungen

Als leichter Einführungstext zum Thema ist Mooney und Duval (1993) geeignet. Mooney und Krause (1997) geben einen Überblick über Weiterentwicklungen des Verfahrens und einen Vergleich mit anderen Methoden. Als ausführlicher Text zum Bootstrap gilt Efron und Tibshirani (1993). Für konkrete politikwissenschaftliche Anwendungen des Bootstraps wird Mooney (1996) und Mooney und Krause (1997) empfohlen sowie Gschwend und Norpoth (2000) als Anwendungsbeispiel aus der deutschsprachigen Literatur. Einen guten Überblick über das Jackknife findet sich in Miller (1974).

## Literatur

*Behnke, Joachim*, 2005: Lassen sich Signifikanztests auf Vollerhebungen anwenden? Einige essayistische Anmerkungen. Politische Vierteljahresschrift 46:01–

- Broscheid, Andreas*, und *Thomas Gschwend*, 2005: Zur statistischen Analyse von Vollerhebungen. *Politische Vierteljahresschrift* 46:O16–26.
- Efron, Bradley*, und *Robert Tibshirani*, 1986: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1:54–75.
- Efron, Bradley*, und *Robert Tibshirani*, 1993: An introduction to the bootstrap. New York: Chapman and Hall.
- Gschwend, Thomas*, und *Helmut Norpoth*, 2000: Soll und Haben: Die deutsche Wählerschaft rechnen mit den Parteien ab. In: *Markus Klein, Wolfgang Jagodzinski, Ekkehard Mochmann, und Dieter Ohr* (Hg.), 50 Jahre empirische Wahlforschung in Deutschland: Entwicklung, Befunde, Perspektiven, Daten, 389–409. Wiesbaden: Westdeutscher Verlag.
- Hall, Peter*, und *Joel L. Horowitz*, 1996: Bootstrap Critical Values for Tests Based on Generalized-Method-of-Moments Estimators. *Econometrica* 64:891–916.
- Lijphart, Arend*, 1999: Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries. New Haven: Yale University Press.
- Miller, Rupert G.*, 1974: The Jackknife - a review. *Biometrika* 61:1–15.
- Mooney, Christopher Z.*, 1996: Bootstrap Statistical Inference: Examples and Evaluations for Political Science. *American Journal of Political Science* 40:570–602.
- Mooney, Christopher Z.*, und *Robert D. Duval*, 1993: Bootstrapping: A Nonparametric Approach to Statistical Inference. Newbury Park, Calif.: Sage.
- Mooney, Christopher Z.*, und *George A. Krause*, 1997: Of silicon and political science - Computationally intensive techniques of statistical estimation and inference. *British Journal of Political Science* 27:83–110.
- Taagepera, Rein*, 2003: Arend Lijphart's Dimensions of Democracy: Logical Connections and Institutional Design. *Political Studies* 51:1–19.