Mathematische Semesterberichte manuscript No.
(will be inserted by the editor)

Joachim Engel · Rudolf Grübel

Bootstrap — oder die Kunst, sich selbst aus dem Sumpf zu ziehen

Eingegangen: xxx / Angenommen: xxxx

Zusammenfassung Computerbasierte Methoden haben die Anwendungsbreite der Statistik enorm erweitert. Simulationstechniken erlauben neue Zugänge zu komplexen Fragestellungen, die traditionell nur unter sehr restriktiven Annahmen möglich waren. Implementierung und Anwendung von rechenintensiven Algorithmen bieten neue Möglichkeiten, das für die Inferenzstatistik zentrale Konzept der Stichprobenverteilung transparenter und besser greifbar zu machen. Der Aufsatz diskutiert didaktischen Nutzen und mathematische Aspekte des Bootstrap-Verfahrens. Wir illustrieren das Verfahren mit einem Beispiel aus der Publikationsgeschichte der Semesterberichte.

Schlüsselwörter Resampling · Verteilung · Konfidenzbereiche

Mathematics Subject Classification (2000) 62-01 · 62F40

> *Gleichwohl sprang ich auch zum zweiten Male noch zu kurz und fiel nicht weit vom andern Ufer bis an den Hals in den Morast. Hier hätte ich unfehlbar umkommen müssen, wenn nicht die Stärke meines eigenen Armes mich an meinem eigenen Haarzopfe, samt dem Pferde, welches ich fest zwischen meine Knie schloss, wieder herausgezogen hätte.*
> aus: *Baron Münchhausen* von Gottfried August Bürger (1747 –1794)

1 Einleitung

Dank der Entwicklung flexibler Software und preisgünstiger Hardware haben Simulationstechniken in den letzten Jahrzehnten stark an Bedeutung für Sta-

J. Engel: Pädagogische Hochschule Ludwigsburg, 71634 Ludwigsburg, Germany
R. Grübel: Institut für Mathematische Stochastik, Leibniz Universität Hannover, 30167 Hannover, Germany
E-Mail: engel@ph-ludwigsburg.de, rgrubel@stochastik.uni-hannover.de

tistik und Wahrscheinlichkeitsrechnung gewonnen. Viele komplexe Probleme und Prozeduren der Stochastik sind analytisch nur sehr schwer zugänglich, während sich simulativ leicht Approximationen erhalten lassen.

Die meisten Standardverfahren der Statistik stammen aus der Zeit zwischen 1890 und 1940. Damals waren umfangreiche Berechnungen sehr zeitaufwändig und teuer. Durch den Einsatz von Computern ist das Rechnen unvergleichbar schneller und billiger geworden. Neue statistische Verfahren können also mit der Rechenkapazität geradezu verschwenderisch umgehen. Daher lassen sich zwei wesentliche Einschränkungen früherer Verfahren aufheben:

1. Früher mussten einige kaum nachprüfbare Annahmen über die Daten gemacht werden. Man nahm z.B. an, dass sich zufällige Schwankungen oder Fehler bei den Messwerten symmetrisch um den wahren Wert streuen und große Fehler unwahrscheinlicher sind als kleine. Für annäherungsweise normalverteilte Daten sind die Berechnungen dann sehr zuverlässig. Bei nicht normalverteilten Daten sind die Ergebnisse aber sehr ungenau. Viele computer-basierte Verfahren kommen ohne die strikte Annahme der Normalverteilung oder einer sonstigen parametrischen Annahme aus. Dies ist ein beachtlicher Fortschritt.
2. Computer können eine Fülle von Zahlen numerisch verarbeiten. Früher mussten die arithmetischen Operationen mit Papier und Bleistift oder später mit dem Taschenrechner ausgeführt werden. Man war notgedrungen bestrebt, kurze und übersichtliche analytische Formeln zu benutzen, da nur dies den Aufwand bewältigbar machte.

Früher konnte man nur solche statistische Maße berücksichtigen, deren theoretische Eigenschaften man auch analytisch-mathematisch untersuchen konnte. Die Untersuchungen von Stichproben konzentrierten sich also auf wenige einfach zu berechnende Größen, wie z.B. Mittelwert, Standardabweichung, Korrelationskoeffizient. Viele andere statistisch wichtige Eigenschaften einer Stichprobe, gerade auch solche, die in einer (robusten) explorativen Datenanalyse an Bedeutung gewonnen haben, entziehen sich oft einer exakten analytischen Beschreibung. Mit Hilfe neuer computergestützter Verfahren lassen sich aber auch Eigenschaften komplexerer bzw. analytisch kaum zugänglicher Stichprobenfunktionen untersuchen. Hintergrund dieser Methoden sind Zufallsgenerator-basierte Simulationen. Die hergeleiteten Schlussfolgerungen gelten dann in einem asymptotischen Sinne und sind mathematisch mit dem Gesetz der großen Zahl oder anderer Grenzwertsätze der Stochastik begründet. Ein wichtiges derartiges Verfahren wurde 1977 von Bradley Efron[1] von der Stanford-University eingeführt: der Bootstrap [8], [9], [10].

Der scheinbar große Vorteil des Bootstrap-Ansatzes ist, dass er das mathematische Know-how durch schiere Rechenleistung ersetzt. Dies begründet zum Teil die Euphorie, die mit dem Aufkommen des Bootstrap-Verfahrens einsetzte. Allerdings stellen erst intensive mathematische Untersuchungen sicher, wann und wie der Bootstrap mit Aussicht auf Erfolg angewendet werden kann. Der mathematische Hintergrund dieses statistischen Werkzeuges sind Konvergenzbetrachtungen empirischer Prozesse.

[1] "Akademischer Großvater" eines der Autoren dieses Aufsatzes

Heutzutage zählt der Bootstrap ebenso wie andere Resampling-Methoden zum alltäglichen Werkzeug des professionellen Statistikers. Die Methoden werden aber bisher kaum in in der Schule oder in einführenden Vorlesungen gelehrt, obwohl die Grundidee plausibel und einfach zu vermitteln ist. Der Bootstrap löst uns von der Einschränkung, schlussfolgernde Statistik nur in den Situationen zu lehren, wo wir analytische Formeln für Stichprobenstatistiken herleiten können. Verschiedene Konzepte für Lageparameter sind auch schon Schülern bekannt. In der schlussfolgernden Statistik wird dann aber nur noch das arithmetische Mittel betrachtet, weil seine Eigenschaften – etwa im Gegensatz zum Median – mathematisch exakt hergeleitet werden können. Robustere Alternativen wie etwa der Median oder das getrimmte Mittel – wichtige Begriffe in einer daten-orientierten, explorativen Statistik – werden ignoriert, weil sie mathematisch schwer zu bewältigen sind. Der Bootstrap macht statistisches Schließen jedoch für mathematisch komplexere Statistiken genauso einfach wie für das arithmetische Mittel.

Neben seiner Nützlichkeit als flexible Methode für viele Anwendungssituationen hat der Bootstrap auch ein erhebliches didaktisches Potenzial. Das zentrale Konzept der schlussfolgernden Statistik ist die Idee der Stichprobenverteilung. Resampling-Methoden erlauben dem Lernenden – wie viele andere Simulationsmethoden – zu erfahren, wie sich eine Statistik von Stichprobe zu Stichprobe unterscheidet und wie sich – mit zunehmender Zahl von Simulationen – eine empirische Stichprobenverteilung allmählich aufbaut. Daher ist der Bootstrap aus didaktischer Perspektive auch weitaus mehr als ein Werkzeug für anspruchsvolle Anwendungen. Er ist ein Instrument, um grundlegende Ideen der schlussfolgernden Statistik zu erkunden, zu visualisieren und konkreter fassbar zu machen.

Ähnlich wie sich einst Baron von Münchhausen an seinem eigenen Schopf aus dem Sumpf zog, lässt sich mit dem Bootstrap-Verfahren (bootstrap = englisch für Stiefelschlaufe, sinngemäß: “sich an den eigenen Stiefeln [aus dem Sumpf] herausziehen”, englisches Gegenstück für “an den eigenen Haaren aus dem Sumpf ziehen”) das zunächst unmöglich Erscheinende erreichen. Im Gegensatz zur Münchhausen-Geschichte handelt es sich beim Bootstrap jedoch um ein sehr seriöses Verfahren, das auf ein mathematisch solides Fundament gestellt werden kann.

2 Simulationen und Resampling

Die Lehr- und Lernforschung betont zunehmend die Bedeutung des Handelns der Lernenden in Situationen des Problemlösens. Lernen wird dabei weniger als Vermittlung und Aneignung von Informationen verstanden, die die Lehrperson vorträgt, sondern ist ein auf individuelle Sinnkonstruktion angelegter Prozess. Das bedeutet, dass auch im Mathematikunterricht Lernende in laborartigen Situationen experimentelle und explorative Arbeitsstile praktizieren, in denen sie zuerst Phänomene erkunden und studieren, bevor aus diesen Erfahrungen tragfähige mentale Modelle für mathematische Begriffe ausgebildet werden.

Viele Anstrengungen didaktischer Forschung der letzten Jahre konzentrierten sich auf die Frage, wie moderne Technologie das Lernen unterstützen

kann. Die Verfügbarkeit moderner Technologien wirkt aber auch zurück auf das, was im technologischen Zeitalter als erstrebenswerter Inhalt des Mathematiklernens angesehen wird. Moore [22] spricht von Synergieeffekten zwischen dem Einsatz neuer Technologien, neuen Inhalten und Fragestellungen und der Umsetzung neuer didaktischer Erkenntnisse. Die Arbeit am Computer kann das Denken von Lernenden über Mathematik qualitativ beeinflussen. Neue Inhalte spiegeln die computer-orientierte und rechenintensive Praxis moderner Statistik wieder.

In der Stochastik sind Simulationstechniken zugleich ein mächtiges Werkzeug zum Problemlösen wie auch ein herausragendes Medium der Visualisierung und Unterstützung von Lernprozessen. Mathematisch basierend auf dem Gesetz der großen Zahl und einem frequentistischen Wahrscheinlichkeitsbegriff lassen sich Wahrscheinlichkeiten durch Simulationen annäherungsweise bestimmen. Simulationen können auch genutzt werden, um das Verstehen von Zufallsprozessen zu fördern, indem sie verdeutlichen, dass Zufallsvariable im Einzelergebnis unvorgesehene Ergebnisse produzieren, langfristig aber vorhersehbare Muster erzeugen. Für den Lernprozess bedeutsam erlauben Simulationen einen experimentellen Arbeitsstil, der neue Zusammenhänge entdecken lässt, indem Lernende neue Szenarien unter der Leitfrage "Was wäre wenn ..." untersuchen. Bei Simulationen ersetzen wir eine reale Situation durch ein Experiment, das ein Modell des Originals ist, das aber leicht manipuliert und analysiert werden kann. Da das Experiment mittels Zufallsgenerator am Computer durchgeführt wird, lassen sich ohne nennenswerten Aufwand Replikationen in hoher Zahl durchführen sowie Auswirkungen von alternativen Modellannahmen und Parameterfestlegungen untersuchen. Simulationen unterstützen valide Vorstellungen bezüglich Zufall und Wahrscheinlichkeit und konfrontieren Fehlvorstellungen. Da allerdings Simulationen lediglich eine (angenäherte) Problemlösung und nicht unmittelbare Gründe für deren Richtigkeit liefern, haben Simulationen per se keine explanative Kraft. Sie erlauben aber einen entdeckenden Arbeitsstil, indem Lernende selbst aktiv Daten produzieren und analysieren und mit Zufallsstichproben einer Population experimentieren, deren Parameter bekannt sind. Auch wenn manche Details zum effizienten Einsatz von Simulationen in der Lehre von Stochastik noch empirisch weiter erforscht werden müssen (Mills [21]), besteht unter Mathematikdidaktikern und Kognitionspsychologen ein breiter Konsensus dahingehend, dass Simulationen herausragende Vorzüge bieten, um bei Lernenden das Verstehen abstrakter Konzepte der Stochastik zu verbessern (Biehler [2], Burrill [5]; Maxara & Biehler [20], Mills [21], Zieffler & Garfield [29]).

Das vielleicht wichtigste Konzept der Inferenzstatistik ist der Begriff der Stichprobenverteilung. Um die Qualität eines Schätzers zu bewerten, gibt uns ein einziger auf Stichprobenbasis errechneter Wert wenig Auskunft. Wir benötigen vielmehr eine Referenzverteilung, d.h. Informationen über das Verhältnis zu anderen möglichen Realisierungen der Stichprobenfunktion. Fragen nach der Qualität von Schätzwerten sind nur über die Stichprobenverteilung, d.h. die Wahrscheinlichkeitsverteilung der Schätzfunktion, zu beantworten. Die analytische Herleitung von Stichprobenverteilungen ist oft sehr aufwändig und verlangt meist den Einsatz fortgeschrittener mathematischer

Methoden. Die Vermittlung komplexer Methoden der Statistik ist jedoch in vielen Ausbildungsstufen nicht realisierbar oder nicht zumutbar. Denn dabei verlieren gerade mathematisch nicht so versierte Lernende im Dschungel mathematischer Umformungen den Zweck aller Bemühungen aus den Augen, nämlich die Herleitung einer Referenzverteilung. Hier sind Simulationen ein äußerst hilfreiches Instrument, weil die formale Mathematik auf ein Minimum reduziert ist und somit der Fokus auf konzeptionelles Verstehen ausgerichtet werden kann. Freilich stellt die Beschränkung auf Simulationen eine didaktische Reduktion dar, die hilft, sich auf das konzeptionell Wesentliche zu konzentrieren. Um die Wirksamkeit des Simulationsansatzes mathematisch exakt zu begründen, muss auf tiefgreifende Konvergenzsätze der Stochastik zurückgegriffen werden.

Insbesondere neuere didaktische Software zum Lernen stochastischer Konzepte wie z.B. FATHOM [14] oder das Trainigsprogramm von Sedlmeier & Köhlers [25] illustrieren durch optionale Animationen, wie sich die Werte einer Stichprobenstatistik von Stichprobe zu Stichprobe unterscheiden und wie mit zunehmender Zahl von Simulationen eine empirische Stichprobenverteilung entsteht. Das ist ein entscheidender didaktischer Vorteil gegenüber klassischen Zugängen, bei denen von vornherein die theoretische Herleitung angestrebt wird.

Während Simulationen in der Wahrscheinlichkeitstheorie in der Regel mit der Vorgabe einer Population oder Vorgabe einer Wahrscheinlichkeitsverteilung starten, um dann Gesetzmäßigkeiten einer Zufallsgröße zu studieren, so ist die typische Denkrichtung der Statistik entgegengesetzt: am Anfang stehen Daten, eine vorliegende Stichprobe, die eine umfassendere Population repräsentiert. Das Ziel aller Bemühungen besteht darin, auf eine Gesetzmäßigkeit oder ein Modell für die Daten zu schließen, das die vorliegenden Beobachtungen erklärt. Sobald Lerndende mit dem simulierten Ziehen von Stichproben von einer bekannten Population vertraut sind, führt ein direkter Weg zur Idee des Resampling. Anstelle aus der nicht verfügbaren Population werden weitere Stichproben – Bootstrap-Stichproben – aus der vorliegenden Stichprobe gezogen. Durch die Wiederverwendung der Daten können Schätzer nicht verbessert werden, da der vorliegende Stichprobenumfang nicht erhöht wird. Es ist jedoch möglich, die Verteilung von Stichprobenstatistiken durch die Bootstrap-Verteilung zu approximieren und damit die Qualität von Schätzern zu beurteilen indem etwa Verzerrung, Perzentile, Vertrauensintervalle oder Standardfehler (approximativ) berechnet werden können.

Der Bootstrap ist ein Paradebeispiel für die Synergie zwischen Technologie und Inhalt: Die Methode ist eine konzeptionell einfache Idee, die allgemein nützlich und lehrreich ist. Der Bootstrap ist aber ohne schnelle und billige Rechenkapazitäten nicht durchführbar.

War der Bootstrap in den ersten Jahrzehnten seiner Konzeption eine interessante Methode für den Spezialisten in Datenanalyse und angewandter Statistik, so wurde in den letzten Jahren auch sein erhebliches didaktisches Potenzial für das Erlernen von Konzepten der Inferenzstatistik entdeckt (Hesterberg [16], Johnson [17], Wood [28]). Bootstrap-Methoden finden allmählich

Eingang in einführende Lehrbücher zur Statistik für angehende Lehrer (A. Engel [11]) und für Anwender (siehe Chance & Rossman [6], Moore [23]).

3 Eine erste illustrative Anwendung

Wir stellen zunächst den Bootstrap in einem didaktisch gewählten Kontext vor, der sich in handlungsorientierten Lehrkonzepten als sehr nützlich erwiesen hat (siehe Scheaffer et al. [24] oder Engel [12], [13]). In seiner paradigmatischen Version geht es um das Schätzen der unbekannten Anzahl N von Fischen in einem Teich basierend auf der Capture-Recapture-Methode. Dazu wird eine Zahl m von Fischen eingefangen, markiert ("Capture"), wieder im See freigelassen und nach einer Weile wird eine Stichprobe vom Umfang n gezogen. Es werden also ein zweites Mal Fische gefangen ("Recapture"), und die Stichprobe besteht darin, die Markierung bei jedem gefangenen Fisch zu notieren. Die Zahl der markierten Fische in der Stichprobe – eine Zufallsgröße – sei mit K bezeichnet. Ein plausibler Schätzer des Populationsumfangs N ist dann

$$\widehat{N} = \frac{m \cdot n}{K}.$$

Auf der Basis einer einzelnen realisierten Stichprobe ist es schwer, die Qualität dieser Schätzung zu beurteilen. Mit K ist auch $\widehat{N}$ eine Zufallsgrße. Eine Wiederholung des Experiments wird daher mit sehr hoher Wahrscheinlichkeit zu einer anderen Schätzung des Populationsumfanges führen. Bei vorliegender Population könnten wir dieses Experiment – in natura oder simulativ – viele Male wiederholen um so eine (empirische) Stichprobenverteilung von $\hat{N}$ zu erhalten. Die Population ist aber nicht verfügbar, sondern es liegen lediglich die gefangenen Fische, eine Stichprobe vom Umfang n ("Recapture"), vor. Diese Daten – falls durch einen Zufallsmechanismus als Zufallsstichprobe aus der Population gezogen – können jedoch als eine gute Repräsentation der gesamten Fischpopulation angesehen werden. Die Bootstrap-Idee besteht nun darin, die vorliegende Stichprobe als Ersatz für die nicht verfügbare Population zu nehmen. Es werden dann weitere Stichproben (Resamples) aus der vorliegenden Stichprobe gezogen:

1. Ziehe eine Stichprobe (Resample oder Bootstrap-Stichprobe) vom Umfang n mit Zurücklegen aus der Stichprobe und notiere die Zahl $k^\star$ der markierten Elemente.
2. Berechne $\widehat{N^\star} = \frac{mn}{k^\star}$ (Bootstrap-Schätzer)
3. Wiederhole Schritt 1 und 2 sehr oft (z.B. 500 Mal), um die (empirische) Bootstrap-Verteilung von $\widehat{N}^\star$ zu erhalten. Diese Bootstrap-Verteilung dient als Annäherung an die unbekannte Stichprobenverteilung von $\widehat{N}$.

Angesichts der nicht verfügbaren Verteilung von $\widehat{N}$ basieren jetzt alle Schlussfolgerungen auf der Bootstrap-Verteilung, d.h. der Verteilung von $\widehat{N^\star}$. Jeder interessierende Parameter des Schätzers $\widehat{N}$ wie z.B. Verzerrung, Standardfehler, Konfidenzintervalle, etc. kann von der gerade generierten Bootstrap-Verteilung ermittelt werden, und dieser Wert dient als Schätzer des entsprechenden Parameters der Stichprobenverteilung von $\hat{N}$. Um diese

Überlegungen weiter zu präzisieren, geben wir nun ein konkretes Zahlenbeispiel.

Beispiel: Wir haben $m = 80$ Fische markiert und eine Stichprobe von $n = 60$ Tieren wiedereingefangen, von denen $k = 13$ eine Markierung hatten, was zu einer Schätzung des Populationsumfangs von $80 \cdot 60/13 \approx 369$ führt. Basierend auf der Stichprobe vom Umfang 60 wurden 1000 Bootstrap-Stichproben vom Umfang 60 gezogen, um die empirische Bootstrap-Verteilung von $\widehat{N}$, wie in Abbildung 1 dargestellt, zu erhalten, woraus sich ein Vertrauensintervall von $[240; 685]$ für $\widehat{N^\star}$ errechnet.[2] Darüber hinaus ist es leicht, den Standardfehler des Schätzers $\widehat{N}$ mittels der Standardabweichung von $\widehat{N^\star}$ zu schätzen. Ebenso einfach ist es jetzt, die Verzerrung von $\widehat{N}$, definiert als Differenz aus Erwartungswert und wahrem Populationsumfang von $E(\widehat{N}) - N$, zu schätzen: dazu betrachten wir die Differenz zwischen dem arithmetischen Mittel der Bootstrap-Verteilung und dem Schätzer von $\widehat{N}$: $392.663 - 369.231 = 21.052$.[3]

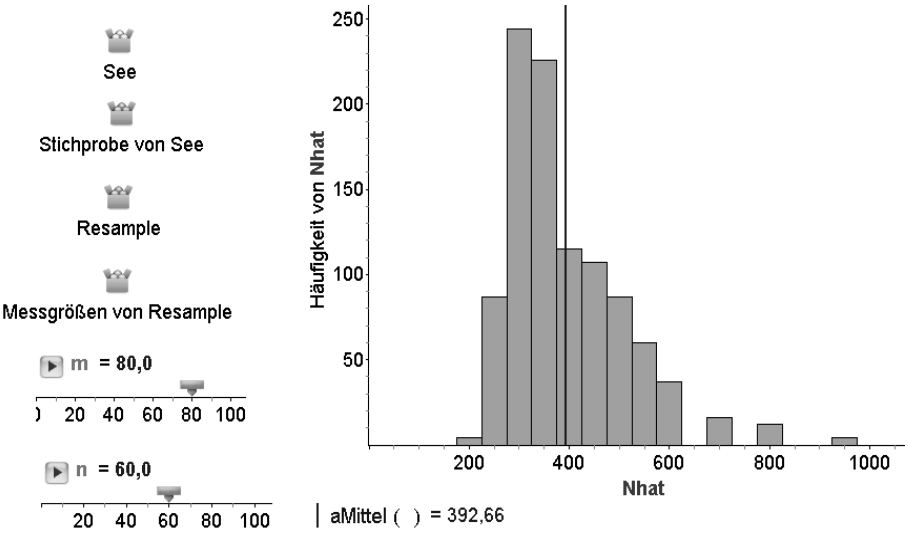

Abb. 1 Bootstrap-Verteilung der geschätzten Populationsumfangs: Implementierung in Fathom (links) und Histogramm (rechts)

Das Histogramm in Abbildung 1 wurde mit der Software FATHOM [14] erstellt, mit der es sehr natürlich ist, Resampling-Methoden wie den Bootstrap zu implementieren. Nachdem ein Datensatz (der "See") mit einer vorher spezifizierten Zahl von Elementen definiert wurde, von denen m Elemente

[2] Basierend auf der hypergeometrischen Verteilung der Zahl K markierter Elemente ist es hier möglich, das klassische 95% Konfidenzintervall als $[245; 578]$ zu berechnen, siehe Johnson et al. [18].

[3] Zum Vergleich: der wahre Populationsumfang in der Simulation ist $N = 310$.

markiert sind, ziehen wir eine Stichprobe vom Umfang n. Weitere Stichproben werden jetzt nicht vom Datensatz "See" gezogen (weil ja die Population für weitere Betrachtungen nicht verfügbar ist), sondern aus "Stichprobe vom See" gezogen. In diesen Bootstrap-Stichproben wird jeweils die Zahl der markierten Elemente gezählt und daraus ein Schätzwert für den Populationsumfang errechnet. Die durch 500-fache Wiederholung erhaltenen Bootstrap-Schätzungen werden schließlich als Histogramm dargestellt.

Man beachte, dass die Verteilung von $\widehat{N}$ verzerrt und rechtssschief ist. Falls $n+m < N$, so wird mit positiver Wahrscheinlichkeit gar kein markierter Fisch in der Stichprobe sein, weshalb dann $\mathrm{E}(\widehat{N}) = \infty$. Damit sind auch Verzerrung und Standardfehler von $\widehat{N}$ unendlich, weshalb robustere Maße wie etwa die robuste Verzerrung, definiert als Differenz zwischen Median und Populationsumfang, $\mathrm{Median}(\widehat{N})-N$, interessant sind. Die robuste Verzerrung ist analytisch kaum handhabbar, der Bootstrap-Algorithmus lässt sich hier jedoch ebenso leicht implementieren wie bei der (nicht-robusten) Verzerrung.

4 Bootstrap-Algorithmus: Plug-In und Monte-Carlo

In allgemeiner Notation kann der Bootstrap wie folgt charakterisiert werden: Unser Interesse gilt einem Parameter $\theta = \theta(F)$ einer uns unbekannten Verteilung F. Es liegen n Beobachtungen $x_1, \ldots, x_n$ vor, Realisierungen von gemäß F identisch verteilten Zufallsgrößen $X_1, \ldots, X_n$. Basierend auf den Beobachtungen schätzen wir θ mittels $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$. Mit F ist auch die Verteilung der Zufallsgröße $\widehat{\theta}(X_1, \ldots, X_n)$ nicht bekannt, was der Grund aller Schwierigkeiten ist. Etwas präziser notieren wir diese Abhängigkeit mit $\widehat{\theta}(X_1, \ldots, X_n|F)$. In dieser Situation hilft uns die Bootstrap-Methode, die aus zwei Schritten besteht: einem statistischen Teil, der auf einer Plug-In Idee basiert, und einem numerischen Teil, der aus einer Monte-Carlo-Simulation besteht.

1. *Plug-In-Schritt*
   Die unbekannte Verteilung von $\widehat{\theta}(X_1, \ldots, X_n)$ wird geschätzt indem F einfach durch die empirische Verteilungsfunktion $\widehat{F_n}$ ersetzt wird, die gegeben ist durch
   $$\widehat{F_n}(x) = \frac{1}{n}\#\{x_i | x_i \leq x\}.$$
   Somit wird die Verteilung von $\widehat{\theta}(X_1, \ldots, X_n|F)$ durch $\widehat{\theta}(X_1, \ldots, X_n|\widehat{F_n})$ geschätzt: die *Bootstrap-Verteilung*.
2. *Monte-Carlo-Schritt*
   Bei n ursprünglichen Beobachtungen ist der Stichprobenraum für die Bootstrap-Replikationen zwar endlich, im Allgemeinen mit n^n Elementen aber sehr groß. Dies ist die Zahl der unterschiedlichen Resamples (Ziehen mit Zurücklegen) einer Stichprobe vom Umfang n. Während in diesem Stichprobenraum die Laplaceannahme der Gleichwahrscheinlichkeit aller Elemente gilt, ist die Verteilung von $\widehat{\theta}(X_1, \ldots, X_n|\widehat{F_n})$ – von speziellen Ausnahmen abgesehen, siehe unten – recht kompliziert und

erlaubt oft kaum eine analytische Berechnung. Um die Verteilung von $\widehat{\theta}(X_1, \ldots, X_n|\widehat{F_n})$ zu erhalten, greifen wir auf Simulationen zurück: wir simulieren indem wir eine Stichprobe vom Umfang n gemäß $\widehat{F_n}$ ziehen, die wir mit $x_1^\star, \ldots, x_n^\star$ notieren (*die Bootstrap-Stichprobe*). Dies ist nichts anderes, als ein Resample der ursprünglichen Stichprobe. Dann berechnen wir $\widehat{\theta}(x_1^\star, \ldots, x_n^\star)$. Dieser Vorgang wird oft, sagen wir B mal, wiederholt. Das Resultat ist eine empirische Approximation an die Bootstrap-Verteilung.

Abbildung 2 illustriert die Vorgehensweise der Bootstrap-Methode.

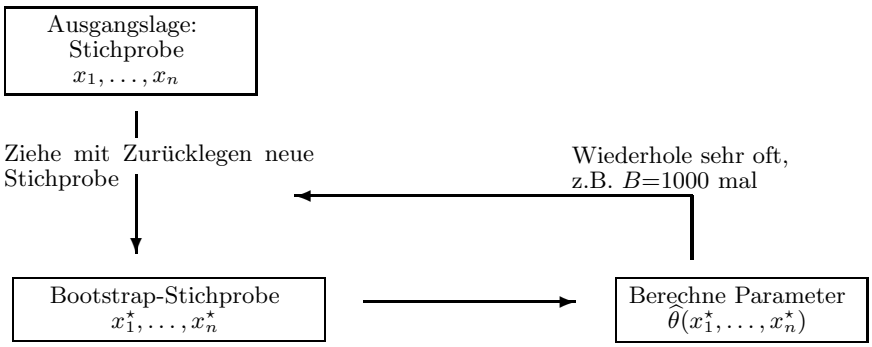

Abb. 2 Darstellung des Vorgehens beim Bootstrap-Verfahren für die Verteilung eines Stichprobenparameters

Man beachte, dass wir es beim Bootstrap-Algorithmus mit zwei unterschiedlichen Approximationen zu tun haben: die (exakte) Bootstrap-Verteilung ist eine Annäherung an die Verteilung der ursprünglichen Stichprobenstatistik. Die Qualität dieser Approximation ist von der ursprünglichen Stichprobengröße n abhängig, die zu erhöhen in der Regel sehr aufwändig und teuer ist (mehr Daten sammeln). Die Annäherung der empirischen Bootstrap-Verteilung an die exakte Bootstrap-Verteilung ist eine numerische Approximation. Der Approximationsfehler kann durch vergrößern der Anzahl der Replikationen B beliebig klein gemacht werden. Ein Zuwachs der Anzahl der Bootstrap-Stichproben ist sehr billig in Zeiten fallender Hardwarekosten und ständig leistungsfähigerer Software.

Der Simulationsschritt führt auf das wiederholte Ziehen mit Zurücklegen, das bei der Beschreibung von Resampling-Verfahren eine solch prominente Rolle spielt. Für die statistische Bedeutung ist er irrelevant, und in der Tat gibt es eine Reihe von Situationen, in denen er überflüssig ist, weil die exakte Bootstrap-Verteilung analytisch berechnet oder durch eine numerisch effizientere Methode ersetzt werden kann.

Hierzu betrachten wir ein illustratives Beispiel: Der Datenmittelwert $\overline{x}$ (von Realisierungen unabhängiger, identischer verteilter Zufallsvariabler) ist der übliche Schätzer des Erwartungswertes μ. Der mittlere quadratische Fehler (MQF) ist definiert durch

$$\mathrm{MQF} = E(\bar{X} - \mu)^2 \;\text{ mit } \bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \;\text{ und }\; \mu := E(X_i).$$

Der MQF hängt somit von der unbekannten Verteilung der X-Variablen ab. Eine einfache Rechnung zeigt, dass der MQF hier gerade der Quotient aus der Varianz der zugrunde liegenden Verteilung und dem Stichprobenumfang n ist. Als Varianz zu $\widehat{F}_n$ erhält man $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$, insgesamt also $\frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ als Bootstrap-Schätzer für den MQF von $\bar{x}_n$. Numerische Approximationen sind in diesem einfachen Beispiel nicht nötig.

Ein anderer interessanter Fall, in dem die Bootstrap-Verteilung direkt, d.h. ohne Monte-Carlo-Schritt, errechnet werden kann, ist die Verteilung des Stichprobenmedians bei einem ungeraden Stichprobenumfang, also $n = 2m+1$ mit $m \in \mathbb{N}_0$ und paarweise verschiedenen Daten $x_1, \ldots, x_n$. Bringt man die Werte in aufsteigende Reihenfolge,

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)},$$

so ist $\widehat{\theta} = x_{(m+1)}$ der Datenmedian. Jedes Resample $x_1^\star, \ldots, x_n^\star$ hat als Median $\widehat{\theta}^\star$ (n ist immer noch ungerade) einen der Werte des Resamples, also einen der Ausgangswerte. Offensichtlich gilt $\widehat{\theta}^\star \le x_{(i)}$ genau dann, wenn mindestens $(m+1)$-mal einer der i Datenwerte gezogen wurde, die kleiner oder gleich $x_{(i)}$ sind, d.h. es gilt für $i = 1, \ldots, n$

$$\mathbb{P}(\widehat{\theta}^\star \le x_{(i)} \mid X_1 = x_1, \ldots, X_n = x_n) \;=\; \sum_{k=m+1}^{n} \binom{n}{k} \Big(\frac{i}{n}\Big)^k \Big(1 - \frac{i}{n}\Big)^{n-k}.$$

Dies zeigt, dass sich die bedingte Verteilung von $\widehat{\theta}^\star$ unter $X_1 = x_1, \ldots, X_n = x_n$ simulationsfrei bestimmen lässt.

Auch in unserem illustrativen Beispiel in Abschnitt 3 lässt sich die Bootstrap-Verteilung direkt ausrechnen. Hat man k markierte Fische in der Stichprobe, so ist, da die Bootstrap-Stichprobe aus der Ausgangsstichprobe durch Ziehen mit Zurücklegen entsteht, die Anzahl $K^\star$ der markierten Tiere in einem Resample biomialverteilt mit den Parametern n und k/n. Wegen $\widehat{N^\star} = mn/K^\star$ folgt somit für die Verteilung des Bootstrap-Schätzers $\widehat{N^\star}$

$$\mathbb{P}(\widehat{N^\star} = x) \;=\; \mathbb{P}\Big(K^\star = \frac{mn}{x}\Big) \;=\; \binom{m}{\frac{mn}{x}} \Big(\frac{k}{n}\Big)^{\frac{mn}{x}} \Big(1 - \frac{k}{n}\Big)^{n - \frac{mn}{x}},$$

vorausgesetzt es gilt $\dfrac{mn}{x} \in \{0, \ldots, n\}$ (formal gehört dabei zum Quotienten 0 der Schätzwert $\widehat{N^\star} = \infty$).

5 Ein Beispiel: Länge der Aufsätze in *Mathematische Semesterberichte*

War das bisherige kurze Beispiel eher elementar, so illustrieren wir nun im Kontext eines Regressionsproblems an einem realen Datenbeispiel zunächst den klassischen parametrischen Zugang und dann das praktische Vorgehen beim Bootstrap. Konkret geht es um die Frage, ob sich die Länge der Aufsätze in den *Mathematischen Semesterberichten* im Laufe der Zeit geändert hat. Seit einigen Jahren werden die Artikel hauptsächlich in die zwei Kategorien 'Mathematik in Forschung, Lehre und Anwendung' (FLA) und 'Mathematik in historischer und philosophischer Sicht' (PHS) eingeteilt. Sind im Verlauf der Jahre die Beiträge in den jeweiligen Kategorien kürzer oder länger geworden? Es gibt verschiedene Ansätze, diese Frage zu mathematisieren. Eine plausible Vorgehensweise geht von der Annahme eines linearen Trend in der mittleren Seitenzahl der Beiträge aus. Zu Testen ist dann die Hypothese, ob es keinen Unterschied in den beiden Geradensteigungen gibt. Abbildung 3 zeigt für den Zeitraum der letzten Jahre (von 1996, Heft 1 bis 2005, Heft 2) für jeden dieser Artikel die Länge (Anzahl der Seiten) sowie das Erscheinungsjahr. Die Artikeltypen sind durch entsprechende Symbole kenntlich gemacht.

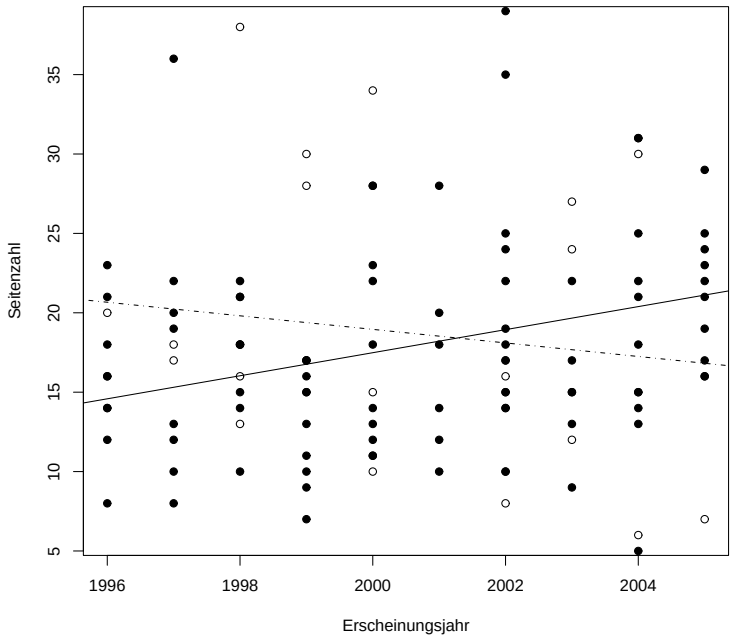

Abb. 3 Seitenzahl der Aufsätze in den Rubriken PHS (○) und FLA (●) sowie zugehörige Regressionsgeraden für PHS (gestrichelt) und FLA (glatt)

In der Abbildung sind auch die jeweiligen Regressionsgeraden für beide Kategorien eingezeichnet; die Gerade mit der negativen Steigung gehört zum Typ PHS. Ist die Differenz der Steigungen statistisch signifikant?

Sowohl im klassischen Rahmen als auch bei dem modernen, computergestützten Zugang, um den es in diesem Aufsatz geht, ist der Ausgangspunkt ein bestimmtes Modell: Die Seitenzahl eines Artikels besteht aus einem systematischen Teil, der (affin-)linear vom Erscheinungsjahr abhängt, und einem zufälligen Teil, über den nun unterschiedliche Annahmen gemacht werden. In beiden Fällen betrachten wir den kleinste-Quadrate-Schätzer $\widehat{\theta}$ für die Differenz θ der Steigungen, und in beiden Fällen wird die Verteilung $\mathcal{L}(\widehat{\theta} - \theta)$ geschätzt. Der Schätzer für diese Verteilung kann dann wiederum als Ausgangspunkt für einen formalen Test (ist die Differenz auf einem bestimmten Niveau α signifikant?) verwendet werden.

Bezeichnet $y_{1,i}$ die Anzahl der Seiten des i-ten Artikels der Rubrik FLA und $y_{2,j}$ die des j-ten Artikels der Rubrik PHS, $1 \le i \le m$, $1 \le j \le n$, so liegt somit eine Struktur der Form

$$y_{1,i} = \beta_{1,0} + \beta_{1,1} x_{1,i} + \varepsilon_{1,i}, \quad y_{2,i} = \beta_{2,0} + \beta_{2,1} x_{2,i} + \varepsilon_{2,i},$$

vor, wobei die jeweilige x-Variable für das Erscheinungsjahr steht und die β-Parameter die Regressionsgeraden spezifizieren. Fasst man dies in Matrizenschreibweise zusammen, mit $\beta = (\beta_{1,0}, \beta_{1,1}, \beta_{2,0}, \beta_{2,1})'$ und der Design-Matrix

$$X = \begin{pmatrix} 1 & x_{1,1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,m} & 0 & 0 \\ 0 & 0 & 1 & x_{2,1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2,n} \end{pmatrix},$$

so wird der kleinste-Quadrate-Schätzer $\widehat{\beta}$ für den Parametervektor β gegeben durch $\widehat{\beta} = (X'X)^{-1}X'y$ und der kleinste-Quadrate-Schätzer für die Parameterfunktion

$$\theta := \beta_{1,1} - \beta_{2,1} = c'\beta \quad \text{mit } c = (0, 1, 0, -1)'$$

ist $\widehat{\theta} = c'\widehat{\beta}$. Sind die ε-Variablen unkorreliert mit Erwartungswert 0 und derselben Varianz $\sigma^2 > 0$, so haben diese Schätzer unter allen linearen erwartungstreuen Schätzern den kleinsten mittleren quadratischen Fehler (Satz von Gauß-Markov, siehe z.B. Bickel & Doksum [1]).

Will man in diesem Modell Konfidenzintervalle für θ erhalten oder beispielsweise die Hypothese $\theta = 0$ einem formalen statistischen Test unterziehen, so benötigt man einen Schätzer für die Verteilung von $\widehat{\theta} - \theta$. Unter der (zusätzlichen) Annahme, dass die ε-Variablen normalverteilt sind, hat die Variable

$$\eta := \frac{\widehat{\theta} - \theta}{\sqrt{c'(X'X)^{-1}c\,\widehat{\sigma^2}}},$$

mit

$$\widehat{\sigma^2} := \frac{\text{SSE}}{m+n-p} \qquad \text{und} \qquad \text{SSE} := (y - X\widehat{\beta})'(y - X\widehat{\beta}),$$

eine t-Verteilung mit $n + m - p$ Freiheitsgraden. Dabei ist p die Dimension des Parametervektors β. Im vorliegenden Beispiel gilt $m = 30$, $n = 89$ und $p = 4$, und es ergibt sich $c'(X'X)^{-1}c\,\text{SSE} = 0.4064$. Das klassische Modell liefert somit eine mit dem Wert 0.6375 skalierte t-Verteilung mit 115 Freiheitsgraden als Schätzer für $\mathcal{L}(\widehat{\theta} - \theta)$.

Wie erhält man nun den Bootstrap-Schätzer für die Verteilung von $\widehat{\theta} - \theta$? Es gilt auch hier – mutatis mutandis – das Schema aus Abbildung 2. Ausgangspunkt sind hier allerdings nicht die Daten, sondern die Residuen

$$r = \begin{pmatrix} r_1 \\ \vdots \\ r_{m+n} \end{pmatrix} := y - X\widehat{\beta},$$

aus denen 'künstliche Stichproben' $r^\star = (r_1^\star, \ldots, r_{m+n}^\star)'$ durch Ziehen mit Zurücklegen gewonnen werden. Jedes solche $r^\star$ definiert einen Vektor $y^\star$ von abhängigen Variablen durch $y^\star := X\widehat{\beta} + r^\star$. Zu diesem $y^\star$ wiederum erhält man einen kleinste-Quadrate-Schätzer $\widehat{\beta}^\star$ durch $\widehat{\beta}^\star = (X'X)^{-1}X'y^\star$ und einen zugehörigen Schätzwert $\widehat{\theta}^\star = c'\widehat{\beta}^\star$. Bei insgesamt B Wiederholungen ergeben sich somit B Werte für $\widehat{\theta}^\star - \widehat{\theta}$, deren empirische Verteilung der gewünschte Bootstrap-Schätzer ist. Abbildung 4 enthält das Resultat eines Durchlaufs mit $B = 1000$, zusammen mit der Verteilungsfunktion zum oben beschriebenen klassischen (parametrischen) Schätzer (glatte Linie). Die beiden Funktionen liegen bemerkenswert nahe beieinander.

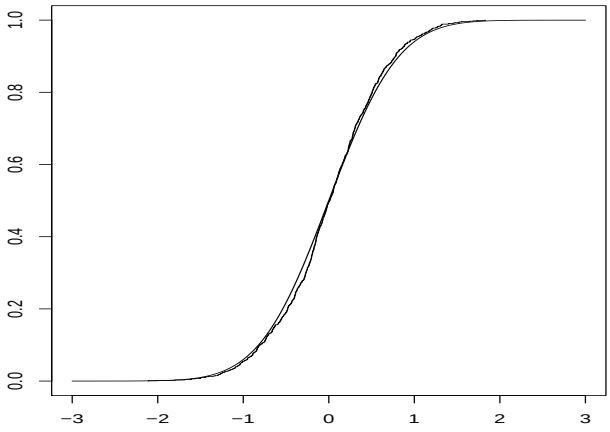

Abb. 4 Klassischer und Bootstrap-Schätzer für die Verteilungsfunktion von $\widehat{\theta} - \theta$

Im statistischen Programmsystem `R`, das sich im akademischen Bereich weitgehend durchgesetzt hat und unter `http://cran.r-project.org` frei erhältlich ist, reduziert sich die Gesamtprozedur zur Implementierung des Bootstraps auf wenige Zeilen. Details hierzu sind in den Anhang ausgelagert worden.

Die Schätzer für $\mathcal{L}(\widehat{\theta}-\theta)$ lassen sich nun für die Konstruktion von Konfidenzintervallen oder die Bestimmung von kritischen Größen zu Tests verwenden. Will man beispielsweise die Hypothese testen, dass die Steigung zu den PHS-Artikeln gleich groß ist wie die zu den FLA-Artikeln, also

$$H:\ \theta\ \big(=\beta_{1,2}-\beta_{2,2}\big)\ =\ 0,$$

so würde man auf dem Niveau α ablehnen, wenn der Schätzwert $\widehat{\theta}$ kleiner ist als das α-Quantil der geschätzten Verteilung. Alternativ kann der p-Wert angegeben werden, also das kleinste α, bei dem noch eine Ablehnung erfolgt. Dies ist einfach der Wert des jeweiligen Schätzers für die Verteilungsfunktion zu $\mathcal{L}(\widehat{\theta}-\theta)$ im Punkt $\widehat{\theta}$; in unserem Beispiel ergeben sich in $\widehat{\theta}=-1.153$ die Werte 0.066 bei der klassischen Prozedur und 0.073 beim Bootstrap. Bei dem kanonischen Signifikanzniveau $\alpha=0.05$ wird also in beiden Fällen nicht abgelehnt. Empörte Leserbriefe der PHS-Fraktion an die Herausgeber können sich also nicht auf unsere Analyse berufen, es sei denn sie basieren ihr Argument auf einen einseitigen Test mit der Hypothese

$$H:\ \theta\ \big(=\beta_{1,2}-\beta_{2,2}\big)\ >\ 0,$$

der auf die p-Werte 0.0335 (klassisch) bzw. 0.038 (Bootstrap) führt. Im Fall des einseitigen Tests wäre allerdings zu kritisieren, dass die Hypothese wohl erst nach Betrachtung der Daten aufgestellt wurde und damit eines der zentralen Gebote der Neyman-Pearson Lehre verletzt wäre, siehe etwa Krengel [19, Abschnitt 6.9].

Mit diesem Beipiel wurden die verschiedenen Vorgehensweisen an einem realen Datensatz illustriert. Es ging nicht um eine profunde statistische Analyse, die beispielsweise andere funktionale Zusammenhänge zwischen Artikellänge und Erscheinungsjahr untersucht hätte und bei der sicher auch Robustheitsaspekte einbezogen worden wären.

Das Beispiel kann auch verwendet werden, um zwei weitere Aspekte zu illustrieren: Zum einen ist bei großen Datensätzen und kleiner Zahl von Parametern in der Regel die klassische Analyse zumindest näherungsweise korrekt, auch wenn die Verteilung der ε-Variablen nicht normal ist. Dies ist natürlich bereits bei den hier vorliegenden Daten ein wichtiger Punkt, denn die Seitenzahlen sind offensichtlich nicht normalverteilt. Übrigens ist die hier verwendete t-Verteilung aufgrund der großen Zahl von Freiheitsgraden für die meisten praktischen Zwecke von der Standardnormalverteilung ununterscheidbar. Zum anderen bietet sich hier die Verwendung einer 'studentisierten' Version des Bootstrap-Verfahrens an: Bei dieser würde man auch die Schätzer $\eta^\star$ der Standardabweichung von $\widehat{\theta}^\star$ in jedem Schleifendurchgang ausrechnen lassen und dann die empirische Verteilung zu den B erhaltenen Werten $(\widehat{\theta}^\star-\widehat{\theta})/\eta^\star$ als Schätzer für die Verteilung von $(\widehat{\theta}-\theta)/\eta$ verwenden.

Es ist bekannt, dass die Überdeckungswahrscheinlichkeiten bei Konfidenzintervallen, die eine solche studentisierte Version verwenden, in vielen Fällen näher an den nominalen Konfidenzwahrscheinlichkeiten liegen als bei der einfachen Version, die den Streuungsparameter nicht berücksichtigt.

6 Ergänzungen und Schlussbemerkungen

Es gibt zahlreiche Varianten des Bootstrap-Verfahrens. Beispielsweise kann man anstelle der empirischen Verteilungsfunktion $\widehat{F_n}$ andere Schätzer von F verwenden wie z.B. einen glatteren Kernschätzer, was zum *geglätteten Bootstrap* führt. Ein anderer Fall liegt vor, wenn wir Grund zu der Annahme haben, dass F zu einer bestimmten parametrischen Familie von Verteilungen gehört. Dann kann auch eine parametrische Schätung von F benutzt werden, was zum *parametrischen Bootstrap* führt. Einer der größten Vorteile des Bootstraps gegenüber klassischen Methoden der Inferenzstatistik ist jedoch gerade seine Robustheit und sein Verzicht auf parametrische Verteilungsannahmen.

In den vorangegangenen Abschnitten haben wir den Bootstrap vor allem als ein *Verfahren* betrachtet: Was ist der Input, welche Rechenschritte sind auszuführen, was soll mit dem Output erreicht werden? Vom Standpunkt des Anwenders sind dies die zentralen Fragen – für den Mathematiker fängt erst jetzt die Arbeit an. Statistische Inferenz basiert auf der Stichprobenverteilung von

$$\widehat{\theta}(X_1, \ldots, X_n | F),$$

aber mit dem Bootstrap beziehen wir uns auf die Verteilung von

$$\widehat{\theta}(X_1, \ldots, X_n | \widehat{F_n}).$$

Um gültige Schlüsse zu ziehen, muss nachgewiesen werden, dass diese beiden Verteilungen nahe beieinander sind, zumindest in einem geeigneten asymptotischen Sinne. Wir brauchen also ein Stetigkeitsargument, um sicher zu stellen, dass der Bootstrap mehr ist als ein Stechen im Dunkeln. Allgemein geht es dabei um die Frage, ob Resampling 'funktioniert'; konkret ist zu untersuchen, ob das Bootstrap-Verfahren zu einem konsistenten Schätzer für die Verteilung einer Statistik oder den mittleren quadratischen Fehler eines Schätzers führt, der also mit wachsendem Stichprobenumfang fast sicher oder in Wahrscheinlichkeit gegen den zu schätzenden Wert konvergiert.

Das in Abschnitt 3 zur Erläuterung des Verfahrens betrachtete Beispiel kann auch verwendet werden, um diese Problematik in einer elementaren Situation zu illustrieren. Hierzu erinnern wir zunächst an die bereits verwendeten Verteilungsaussagen: Die Anzahl K der markierten Fische in der Stichprobe ist hypergeometrisch verteilt mit Parametern N (die unbekannte Gesamtzahl der Fische im See), m (die Anzahl der markierten Fische) und n (der Umfang der Stichprobe), kurz: $K \sim \mathrm{HypGeo}(N, m, n)$. Als bedingte Verteilung der Anzahl $K^\star$ der markierten Tiere in einem Resample erhält man die Binomialverteilung mit Parametern n und K/n, für die wir

$\mathrm{Bin}(n, K/n)$ schreiben. Der Schätzer $\widehat{N}$ ist eine (deterministische) Funktion von K, und die Verteilung von $\widehat{N^\star}$ ist das Bild der Verteilung von $K^\star$ unter dieser Funktion. Wir werden somit auf die Frage geführt, wie nahe die (zufällige!) Verteilung $\mathrm{Bin}(n, K/n)$, der Schätzer für die Verteilung von K, an der tatsächlichen Verteilung von K, also $\mathrm{HypGeo}(N, m, n)$, liegt. Nun ist wohlbekannt, dass es wenig Unterschied macht, ob man eine Stichprobe vom Umfang n mit oder ohne Zurücklegen aus einer Grundgesamtheit vom Umfang N entnimmt, wenn nur n klein ist im Vergleich zu N (unter Verwendung der Steinschen Methode wird in [26] gezeigt, dass der sog. Totalvariationsabstand zwischen $\mathrm{HypGeo}(N, m, n)$ und $\mathrm{Bin}(n, m/n)$ kleiner als $(n-1)/(N-1)$ ist). Dies führt auf den Vergleich zweier Binomialverteilungen, beide mit erstem Parameter n, die eine mit festem zweiten Parameter m/N und die andere mit zufälligem zweiten Parameter K/n. Zumindest im Mittel fällt dieser Vergleich hervorragend aus, da sich als Erwartungswert des zufälligen zweiten Parameters aufgrund der bekannten Formel für den Erwartungswert zur hypergeometrischen Verteilung der Wert m/N ergibt.

Es reicht natürlich nicht, im Mittel richtig zu liegen (ein Sachverhalt, der Gegenstand eines bekannten Spottgedichtes ist, siehe S.1 in [15]). Als Maß dafür, wie stark K/n um seinen Erwartungswert herum konzentriert ist, bietet sich die Variationskoeffizient dieser Variablen an, also der Quotient aus der Standardabweichung und dem Erwartungswert. Unter abermaliger Verwendung einer bekannten Formel für hypergeometrische Verteilungen erhält man

$$\frac{\sqrt{\mathrm{var}\,(K/n)}}{E\,(K/n)} \;=\; \sqrt{\frac{(N-m)(N-n)}{nm(N-1)}}.$$

Betrachten wir den konkreten Fall $m \approx \alpha N$ mit einem $\alpha > 0$, so ergibt sich für den Variationskoeffizienten bei großem N und im Vergleich hierzu kleinem n näherungsweise der Wert

$$\frac{\sqrt{\mathrm{var}\,(K/n)}}{E\,(K/n)} \;\approx\; \sqrt{\frac{1-\alpha}{n\alpha}}.$$

Es darf also α nicht zu klein werden – man muss schon eine substantielle Anzahl von Fischen markieren, um der Gesamtprozedur vertrauen zu können. Dies sind natürlich exakt die Bedingungen, die die Qualität von K als Schätzer für $(nm)/N$ gewährleisten: Dieser Schätzer ist erwartungstreu und hat denselben Variationskoeffizienten. Ist m zu klein, so ist nach dem Gesetz der seltenen Ereignisse eine Poisson-Approximation angebracht, und sowohl der Schätzer K als auch der Bootstrap-Schätzer für die Verteilung von K sind zu variabel, um noch sinnvoll eingesetzt werden zu können. In der Tat ist bereits die dann große Wahrscheinlichkeit für eine Stichprobe völlig ohne markierte Elemente ein Indiz dafür, dass Capture-Recapture nicht funktioniert.

Überlegungen dieser Art können mit Grenzwertsätzen mathematisch konkretisiert werden. Im vorliegenden Fall bedeutet dies, dass man die Parameter N, m und n durch Folgen $(N_j)_{j\in\mathbb{N}}$, $(m_j)_{j\in\mathbb{N}}$ und $(n_j)_{j\in\mathbb{N}}$ ersetzt und dann j

gegen ∞ gehen lässt. Um auch dies zu illustrieren, betrachten wir das Verhalten von Konfidenzschranken für N_j bei bekanntem m_j und nehmen dabei

$$\lim_{j\to\infty} n_j = \infty, \quad \lim_{j\to\infty} \frac{n_j}{N_j} = 0, \quad \lim_{j\to\infty} \frac{m_j}{N_j} = \alpha$$

mit $0 < \alpha < 1$ an. Solche Schranken basieren auf der Verteilung von K_j, für deren Asymptotik aufgrund des erwähnten Resultats von Soon der Satz von de Moivre-Laplace herangezogen werden kann:

$$\frac{K_j - n_j\alpha}{\sqrt{n_j\alpha(1-\alpha)}}$$

konvergiert mit $j \to \infty$ in Verteilung gegen die Standardnormalverteilung. Das Bootstrap-Gegenstück hierzu ist

$$\frac{K_j^\star - n_j(K_j/n_j)}{\sqrt{n_j(K_j/n_j)\,(1-(K_j/n_j))}}\,.$$

Nach dem starken Gesetz der großen Zahlen konvergiert K_j/n_j mit Wahrscheinlichkeit 1 gegen α. Da wir $0 < \alpha < 1$ angenommen haben, erhalten wir hieraus mit einer leichten Erweiterung des Satzes von de Moivre-Laplace, dass auch dieses Bootstrap-Gegenstück, und zwar mit Wahrscheinlichkeit 1, die Standardnormalverteilung als Verteilungsgrenzwert hat. Da die klassische Prozedur auf asymptotisch korrekte Konfidenzschranken führt, gilt dies somit auch für die Bootstrap-Variante. Kurz: der Bootstrap funktioniert.

Dagegeben erhält man im Falle

$$\lim_{j\to\infty} n_j = \lim_{j\to\infty} m_j = \infty, \quad \lim_{j\to\infty} \frac{m_j n_j}{N_j} = \lambda \in (0,\infty)$$

für K_j selbst im Limes $j \to \infty$ eine Poisson-Verteilung mit Parameter λ, und für die bedingte Verteilung von $K_j^\star$ unter K_j eine Poisson-Verteilung mit zufälligem Parameter Λ, wobei Λ Poisson-verteilt ist mit Parameter λ. In dieser Situation ist also selbst bei großem j die Verteilung von $K_j^\star$ kein brauchbarer Schätzer für die Verteilung von K_j.

Wir schließen diesen Einblick in eine der interessantesten Entwicklungen der modernen Statistik mit einer Beobachtung wissenschaftssoziologischer Natur: Viele Anwender haben Resampling-Verfahren als Allheilmittel, auch gegen das mangelnde tiefere Verständnis der verwendeten Prozeduren, betrachtet. In der Tat legen Beispiele wie das im vorangegangenen Abschnitt (siehe Abbildung 4) nahe, dass man die theoretischen Überlegungen, die dort auf eine t-Verteilung führten, durch Anwendung von Resampling-Verfahren ersetzen kann. Es ist in diesem Zusammenhang gerade für Mathematiker interessant zu sehen, wie anspruchsvoll eine theoretische Durchdringung des Bootstraps sein kann, insbesondere wenn diese über die Untersuchung von Einzelfällen hinausgeht. Eine mathematisch solide Fundierung findet man beispielsweise in [27], wo das Verfahren vor dem Hintergrund der modernen Theorie der empirischen Prozesse behandelt wird.

Literatur

1. Bickel, P. & Doksum, K.: Mathematical Statistics: Basic Ideas and Selected Topics Oakland: Holden-Day 1977.
2. Biehler, R. (1991): Computers in probability education. In: Kapadia, R. & Borovcnik, M. (eds.). Chance Encounters — Probability in Education. Dordrecht: Kluwer, 169-211 (1991)
3. Biehler, R.; Hofmann, T.; Maxara, C. & Prömmel, A.: Fathom 2. Eine Einführung. Heidelberg: Springer 2006
4. Billingsley, P.: Probability and Measure. New York: Wiley 1986
5. Burrill, G.: Simulation as a tool to develop statistical understanding. International Statistical Institute (Ed.), Proceedings 6th International Conference on Teaching Statistics, Cape Town (on CD) 2002
6. Chance, B. & Rossman, A.: Investigating Statistical Concepts, Applications, and Methods. Duxbury 2006
7. Davison, A.C. & Hinkley, D.V.: Bootstrap Methods and Their Application. Cambridge University Press 1997
8. Diaconis, P. & Efron, B.: Statistik per Computer: der Münchhausen-Trick. Spektrum der Wissenschaft (7), 56-71 (1983).
9. Efron, B.: The Jackknife, the Bootstrap and other Resampling Plans. Society for Industrial and Applied Mathematics, CBNS-NSF 38, Philadelphia 1982
10. Efron, B.& Tibshirani, R.: An Introduction to the Bootstrap. New York: Chapman and Hall 1993
11. Engel, A.: *Stochastik.* Stuttgart: Klett 1987
12. Engel, J.: Markieren – Einfangen – Schätzen: Wie viele wilde Tiere? Stochastik in der Schule, 2, 17-24 (2000)
13. Engel, J.: On Teaching the Bootstrap. Bulletin of the International Statistical Institute 56th Session, Lisbon 2007
14. FATHOM ^{TM}. Dynamic Data Software. Keycurriculum Press: Emeryville, CA
15. Hartung, J.: Statistik. Lehr- und Handbuch der angewandten Statistik. Oldenbourg: München 1982
16. Hesterberg, T.: Simulation and Bootstrapping for Teaching Statistics, in: American Statistical Association: Proceedings of the Section on Statistical Education, 44 - 52 (1998)
17. Johnson, R.: An Introduction to the Bootstrap. Teaching Statistics, Vol. 23 (2), 49- 54 (2001)
18. Johnson, N. L.; Kotz, S. & Kemp, A. W.: *Univariate Discrete Distributions.* 2nd Edition, New York: Wiley (1992)
19. Krengel, U.: Einführung ind die Wahrscheinlichkeitstheorie und Statistik. Vieweg: Wiesbaden 1988
20. Maxara, C. & Biehler, R.: Students' probabilistic simulation and modeling competence after a computer-intensive elementary course in statistics and probability. International Statistical Institute (Ed.), Proceedings 7th International Conference on Teaching Statistics, Salvador, Brazil (on CD) 2006
21. Mills, J.: Using Computer Simulation Methods to Teach Statistics: A Review of the Literature. J. Statistics Education, Vol. 10 (1). Online: http://www.amstat.org/publications/jse/ (2002)
22. Moore, D.: New pedagogy and new content: the case of statistics. International Statistical Review **65**, 123-166 (1997)
23. Moore, D.: The Practice of Business Statistics. New York: Freeman 2003
24. Scheaffer, R.; Gnanadesikan, M.; Watkins, A. & Witmer, J.: Activity-Based-Statistics. New York: Springer 1996
25. Sedlmeier, P. & Köhlers, D.: Wahrscheinlichkeiten im Alltag. Statistik ohne Formeln. Braunschweig: Westermann 2005
26. Soon, S.Y.T.: Binomial approximation for dependent indicators, Statistica Sinica **6**, 703-714 (1996)
27. van der Vaart, A.W.& Wellner, J.A. Weak Convergence and Empirical Processes. Springer: New York 1998

28. Wood, M.: The Role of Simulation Approaches in Statistics. *Journal of Statistics Education*, vol 13 (3). Online: http://www.amstat.org/publications/jse/ (2005)
29. Zieffler, A. & Garfield, J.: Studying the Role of Simulation in Developing Students' Statistical Reasoning. Bulletin of the International Statistical Institute 56th Session, Lisbon 2007

Anhang: Programmcode in R zum Beispiel in Abschnitt 5

Wir gehen davon aus, dass die Variablen y, X, m, n, p bereits mit den entsprechenden Werten belegt sind:

```
G <- solve(t(X) %*% X)
A <- G %*% t(X)
hatbeta <- A %*% y
c <- matrix(c(0,1,0,-1),4,1)
hattheta <- t(c) %*% hatbeta
res <- y - X %*% hatbeta
SSE <- t(res) %*% res
hatsigma2 <- SSE / (m + n - p)
hatsdvtheta <- sqrt( hatsigma2 * t(c) %*% G %*% c)
testgroesse <- hattheta / hatsdvtheta
haty <- X %*% hatbeta

hatthetastar <- matrix(0,B,1) # fuer Schaetzwerte zu den Resamples
for (b in 1:B){               # Hauptschleife
   resampleindices <- ceiling((n + m) * runif(n+m))
   rstar <- res[resampleindices]
   ystar <- haty + rstar
   hatbetastar <- A %*% ystar
   hatthetastar[b] <- t(c) %*% hatbetastar
}

sorterg <- sort(hatthetastar-hattheta[1,1])
empdfx <- rep(sorterg,each=2) # zeichne emp. Verteilungsfunktion
empdfy <- c(0,rep(1:(B-1),each=2)/B,1)
plot(empdfx, empdfy, type="l",xlim<-c(-3,3), xlab="",ylab="")
x0 <- (-99:99)/33             # der 'klassische' Schaetzer
y0 <- pt(x0/hatsdvtheta, df=117)
points(x0,y0,type="l", lty=1)
```

Natürlich lässt sich dies durch Rückgriff auf passende (und existierende) Makros kürzer abhandeln; in der hier angegebenen Form sind jedoch die einzelnen in Abschnitt 5 angegebenen Schritte leicht nachvollziehbar.