

# Bootstrapping – ein neuer Standard in Anwendung und Lehre?

Bernd Heinen  
 SAS Institute GmbH  
 In der Neckarhelle 168  
 Heidelberg  
 Bernd.heinen@jmp.com

## Zusammenfassung

Bootstrapping bietet eine Möglichkeit, Verteilungsinformationen auch komplexer Statistiken aus einem Datensatz zu gewinnen. Die Anwendung ist an keine Voraussetzungen geknüpft. Die Ergebnisse sind „nur“ empirische Verteilungen, die aber einen interessanten Einblick in die Gestalt der Verteilung bieten und daher Rückschlüsse auf das Zutreffen von Voraussetzungen wie Symmetrie, Normalverteilung oder ähnlichem zulassen. Da eine Statistik und ihre Verteilung unmittelbar aus einem Datensatz abgeleitet werden, erleichtert Bootstrapping es dem Lernenden, ein Verteilungskonzept zu entwickeln, ohne sich mit theoretischen Verteilungen befassen zu müssen. In der Anwendung erkennt man schnell die Zulässigkeit von Regularitätsbedingungen für Konfidenzintervalle oder Tests, die vielleicht nicht so oft in Frage gestellt werden, wie sie sollten. Die einfache Anwendung mit Desktop Software wie JMP Pro bringt dieses Verfahren auf jeden Schreibtisch.

**Schlüsselwörter:** Bootstrap, Statistik

## 1 Zur Begriffsbildung

Die „Stiefelschlaufe“ (Bootstrap) dient im amerikanischen Sprachgebrauch der Beschreibung der Tatsache, dass man sich aus eigener Kraft aus dem Sumpf zieht. Dem deutschen Leser besser bekannt ist die Analogie bei Münchhausen, der sich „am eigenen Schopfe aus dem Sumpf gezogen hat, mitsamt seinem Pferd“. Im statistischen Umfeld deutet diese Begriffswahl darauf hin, dass man alleine aus den Daten, die man zur Analyse vorliegen hat, die interessierenden Statistiken und deren Standardfehler ableitet, ohne weitere Annahmen über Verteilungen oder asymptotisches Verhalten zu treffen.



## 2 Motivation

Zur Beurteilung der Qualität von Statistiken oder der Durchführung von Tests benötigt man Kenntnisse von deren Verteilung oder zumindest des zugehörigen Standardfehlers. In vielen Fällen der parametrischen Statistik sind beide bekannt. Für den Mittelwert beispielsweise lassen sich Verteilung und Standardfehler exakt herleiten. Schon beim Median geht das nicht. Für viele weitere Statistiken, mit denen wir häufig umgehen und für die wir selbstverständlich Standardfehler und Konfidenzintervalle berechnen, sind aber die Verteilungen auch nicht exakt bekannt. Die Berechnung der Standardfehler und Konfidenzintervalle beruht auf zusätzlichen Annahmen wie Symmetrie, Unabhängigkeit und asymptotischer Konvergenz. Oftmals sind diese Annahmen schwer zu rechtfertigen und dann ist es hilfreich – zumindest aber beruhigend – nichtparametrische Verfahren zur Verfügung zu haben.

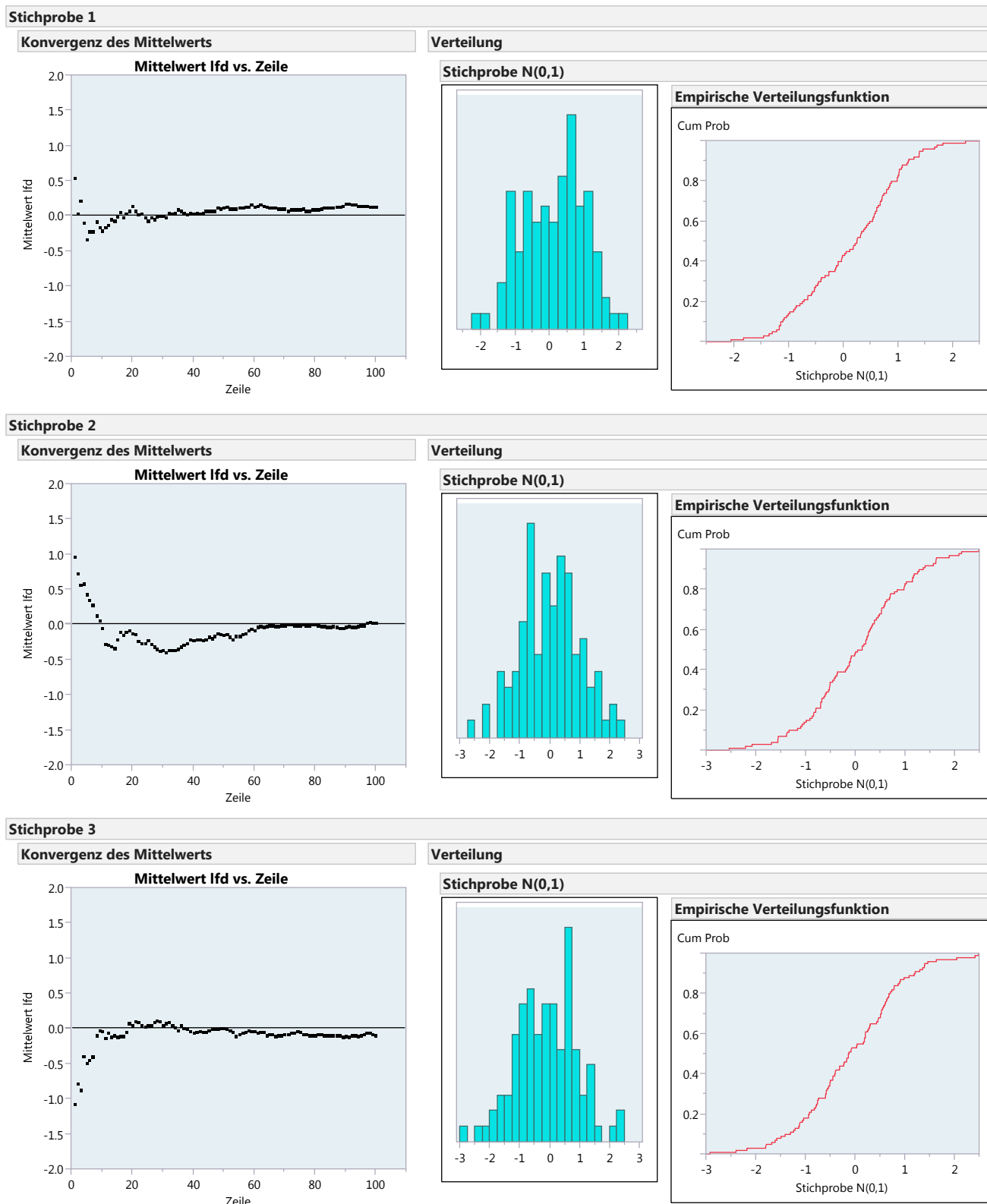
Dementsprechend wächst auch das öffentliche Interesse, hat z.B. die FDA vor 2010 kaum einen Kommentar über Bootstrapping veröffentlicht, steigt die Zahl der Artikel seitdem kontinuierlich an.

Außerdem ist Bootstrapping leicht verständlich, jede neue Stichprobe lässt sich analysieren, es ist intuitiv einleuchtend, dass sich die berechneten Statistiken unterscheiden. Ihre Verteilung lässt sich mit einfachen Mitteln anschaulich darstellen und für Schlussfolgerungen nutzen. Insofern stellt Bootstrapping auch eine pädagogisch interessante Methode dar.

## 3 Das Verfahren

Ziel des Bootstrapping ist es, die Verteilung einer beliebigen Statistik zu erhalten. Dazu werden aus dem zugrunde liegenden Datensatz so viele Stichproben mit Zurücklegen gezogen wie der Datensatz Beobachtungen hat. Für diesen Stichprobendatensatz wird die Statistik berechnet. Diese beiden Schritte, Stichprobenziehung und Statistikberechnung, werden  $k$  mal wiederholt, so dass es  $k$  verschiedene Statistikwerte gibt. Daraus kann man dann die gewünschten Verteilungsinformationen erhalten, um z.B. robuste Schätzer – den Median dieser Verteilung – und Konfidenzintervalle zu erhalten. Gleichzeitig kann man diese Werte mit denjenigen vergleichen, die durch Anwendung asymptotischer parametrischer Verfahren ermittelt wurden.

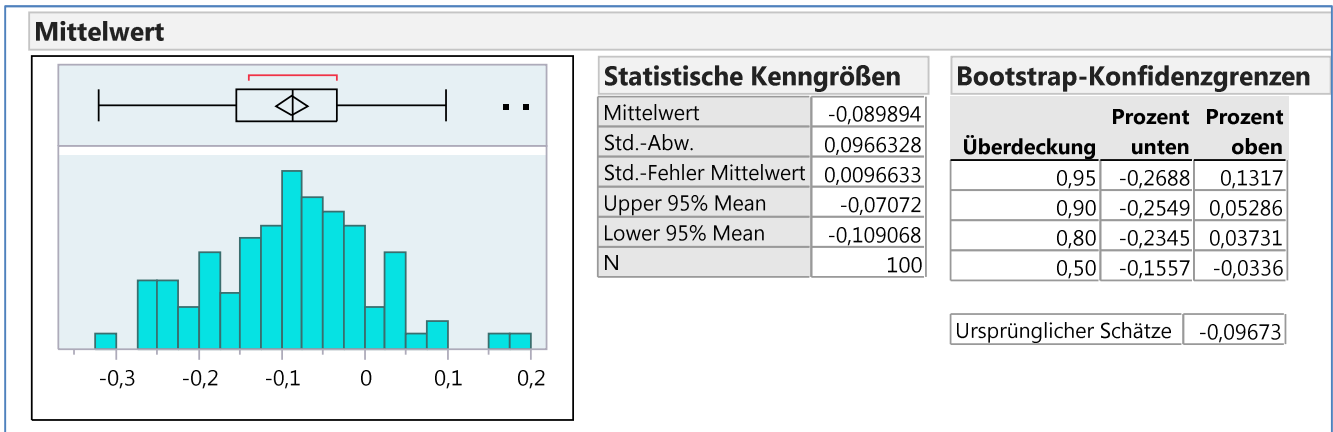
Um die Qualität des Verfahrens zu beurteilen, ist es naheliegend, eine bekannte Verteilung zugrunde zu legen, und dann zu sehen, inwieweit diese sich durch das Bootstrapverfahren reproduzieren lässt. Nehmen wir zur Illustration einen Datensatz mit 100 Zufallszahlen, die aus einer  $N(0,1)$  Verteilung stammen. Zusätzlich habe ich gleitende Mittelwerte berechnet, d.h. in der  $i$ -ten Zeile steht der Mittelwert aus den Zufallszahlen der Zeilen 1 bis  $i$ . Die Verteilung der Zufallszahlen und die Konvergenz des Mittelwertes kann man grafisch darstellen. Für verschiedene Sets von 100 Zufallszahlen können die Verteilung so aussehen:



**Abbildung 1:** Verteilungen von 100 Zufallszahlen

Die empirischen Verteilungen zeigen eine gute Annäherung an die zugrunde liegende Standardnormalverteilung, aber es werden auch die zufallsbedingten Schwankungen deutlich. Die Daten stammen aus einer Grundgesamtheit mit bekannter Verteilung, ebenso sind die parametrischen Statistiken für den Mittelwert gut bekannt und leicht zu

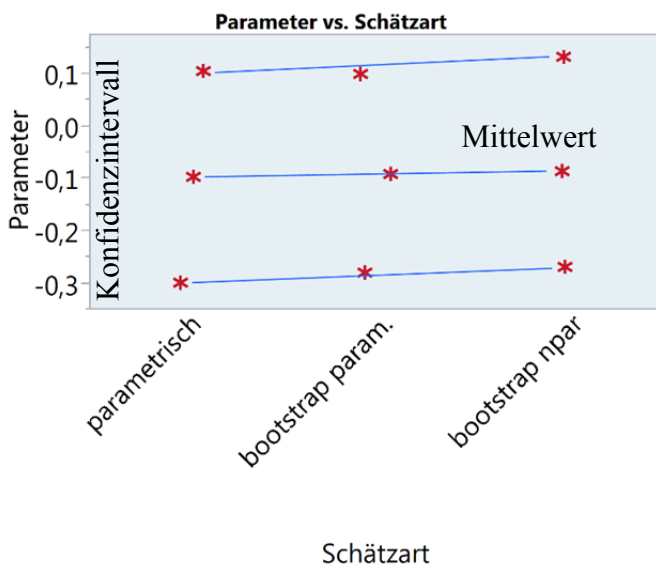
berechnen. Daher bietet sich diese Situation dazu an, zu überprüfen, wie genau das Bootstrapverfahren die entsprechenden Schätzer reproduziert. Für einen dieser simulierten Datensätze beträgt der Mittelwert  $-0,097$  mit einer Standardabweichung  $1,015$  und einem Konfidenzintervall von  $-0,3$  bis  $0,11$  (in Abbildung 3 die Werte für parametrisch). Wendet man darauf das Bootstrapverfahren mit 100 Wiederholungen an, ergibt sich die Verteilung für den Mittelwert aus Abbildung 2.



**Abbildung 2:** Verteilung und Parameter aus 100 Bootstrap Stichproben

Der Mittelwert aller Mittelwerte liegt mit  $-0,09$  nahe an dem Originalwert, das aus den wiederholten Stichproben resultierende Konfidenzintervall reicht von  $-0,28$  bis  $0,1$  (alle Werte in Abbildung 3 unter „bootstrap param.“) und ist damit symmetrisch zum parametrischen Intervall, aber es ist ein bisschen kleiner. Nimmt man die entsprechenden Perzentile (2,5% - 50% - 97,5%) als Schätzer, ergibt sich  $-0,086$  als Median mit einem Konfidenzbereich von  $-0,27$  bis  $0,13$  (in Abbildung 3 unter „bootstrap npar“), der gegenüber beiden anderen Konfidenzintervallen leicht nach oben verschoben ist.

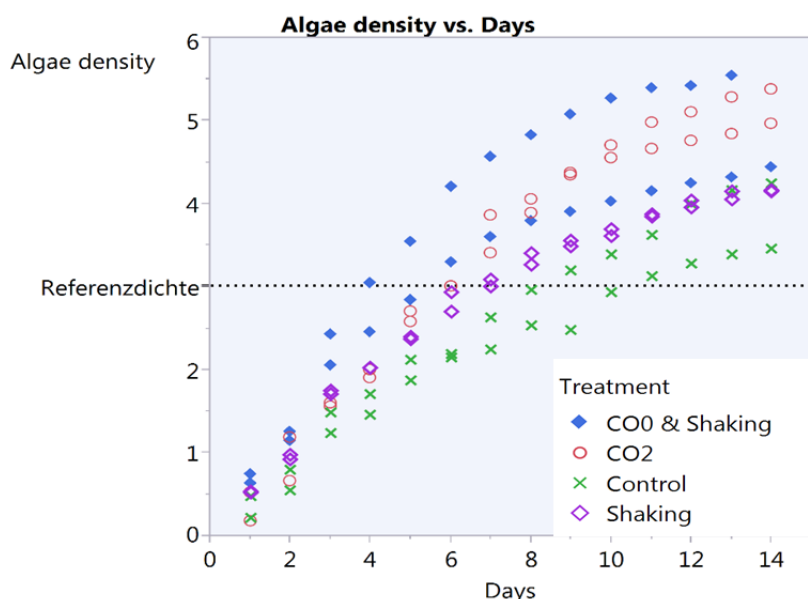
Alle Bootstrap Resultate entsprechen der zugrundeliegenden theoretischen Verteilung und schätzen Parameter und deren Konfidenzintervalle korrekt. Durch die beliebig wählbare Zahl der Stichprobenziehungen bilden auch die Perzentile hinreichend differenzierte Werte und liefern Konfidenzintervalle, die nicht einmal konservativer als ihre parametrischen Entsprechungen sind. Das Beispiel veranschaulicht diesen Sachverhalt.



**Abbildung 3:** Mittelwerte und KIs

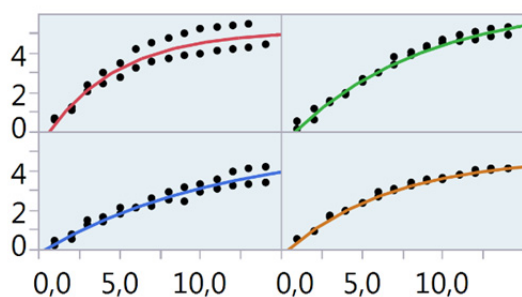
## 4 Anwendung

Als Beispiel für die Anwendung dient ein Datensatz über Algenwachstum, das mit Hilfe verschiedener Maßnahmen beschleunigt werden soll. Der zeitliche Verlauf zeigt deutlich das nichtlineare Wachstum.



**Abbildung 4:** Algenwachstum

Prognosen der Zeiten bis zum Erreichen der Referenzdichte aus?



**Abbildung 5:** Kurvenanpassungen

JMP liefert die entsprechende Tabelle und eine Grafik für alle vier Gruppen und führt darin die zu erwartende Zeitdauer auf sowie die asymptotisch berechneten Konfidenzintervalle. Für die vorhergesagte Zeitdauer soll jetzt die Verteilung über das Bootstrapverfahren bestimmt werden, um daraus die Konfidenzintervalle abzuleiten. Ein Rechtsklick in der entsprechenden Spalte öffnet das Kontextmenü, aus dem der Punkt „Bootstrap“ ausgewählt wird. Ein Fenster gibt noch die Wahl für die Zahl der zu bildenden Stichproben, in diesem Beispiel wird 1000 gewählt. Nun werden 1000-mal Stichproben mit Zurückziehen gebildet, die Kurven werden neu angepasst und jeweils die mittlere Zeit bis zum Erreichen der Algendichte 3 berechnet. Die Ergebnisse werden in einer Tabelle abgespeichert, die mit den üblichen JMP Funktionen analysiert werden kann. Die Verteilung zeigt die Ergebnisse übersichtlich an. Am Beispiel der Behandlung „CO0 & Shaking“ sieht man, dass die Verteilung des Prognosewertes durch eine Normalverteilung einigermaßen angenähert wird, mit kleinen Abweichungen, speziell am oberen Rand. Interessant an der Verteilung der Statistik ist auch, dass man die Einflüsse der leicht asymmetrischen Verteilung deutlich sieht. Sowohl Median als auch Mittelwert der Stichproben liegen auf Grund der vielen Ausreißer nach oben knapp oberhalb des ursprünglichen Schätzers.

Das exponentielle Wachstum wird über die Funktion

$$w = a (1 - b e^{-c \text{ days}})$$

mit den Parametern  $a$ ,  $b$  und  $c$  beschrieben, die für jede der Behandlungen unterschiedlich geschätzt werden.

Die Anpassungen für die einzelnen Gruppen sind zutreffend, wie sieht es nun für die

Vorhersagewerte					
Treatment	Algae density	Days	Std.-Fehler	Untere	Obere
	angegeben	vorhergesagt		Grenze 95%	Grenze 95%
CO2 & Shaking	3,0	4,35	0,19	3,97	4,73
CO2	3,0	5,68	0,20	5,28	6,08
Control	3,0	9,10	0,34	8,43	9,78
Shaking	3,0	6,77	0,32	6,16	7,39

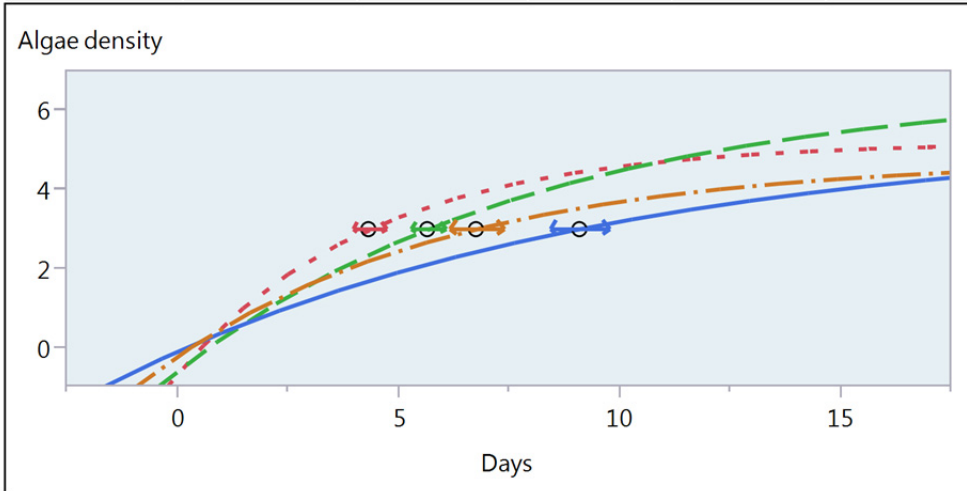


Abbildung 6: Schätzungen für Zeitpunkte

Das nichtparametrische Konfidenzintervall, das durch Perzentile der Verteilung berechnet wird, ist deutlich asymmetrisch und weiter, als das mit  $[3,9744, 4,7347]$  berechnete parametrische Intervall des Schätzers. Interessant ist, dass für die anderen Gruppen die nichtparametrischen Konfidenzintervalle keineswegs größer sind als die asymptotisch berechneten.

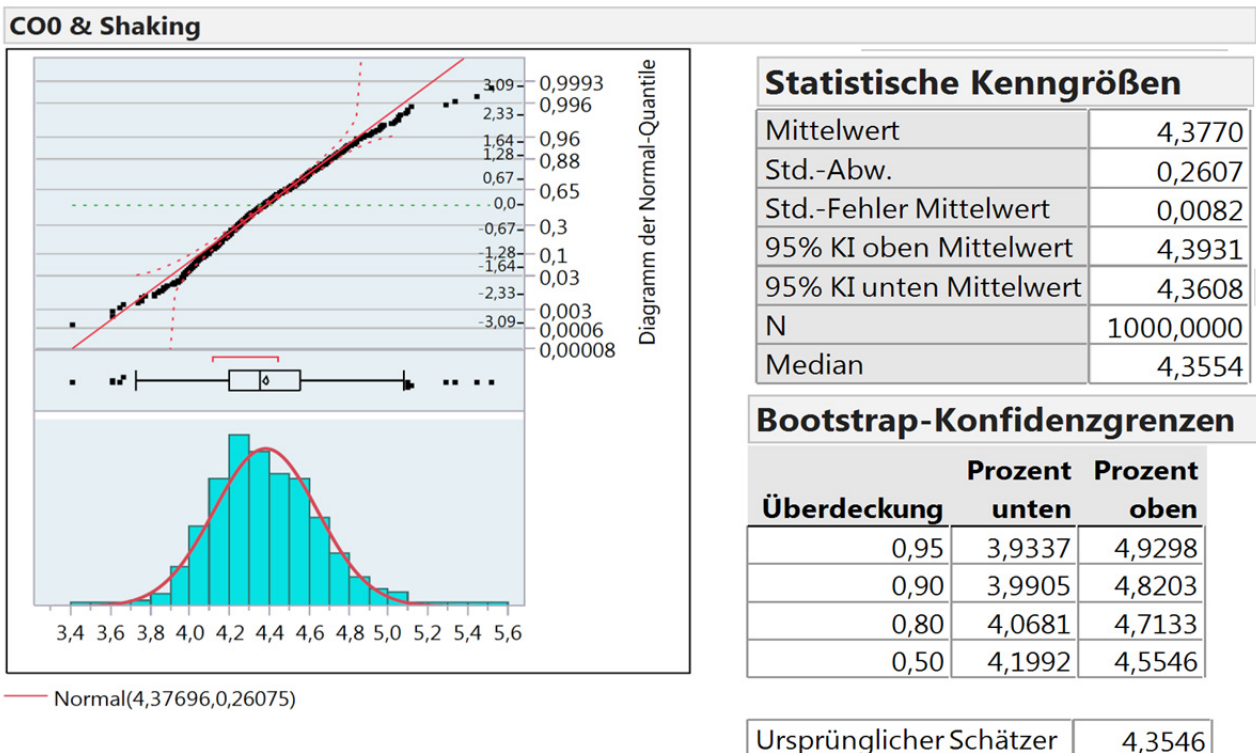
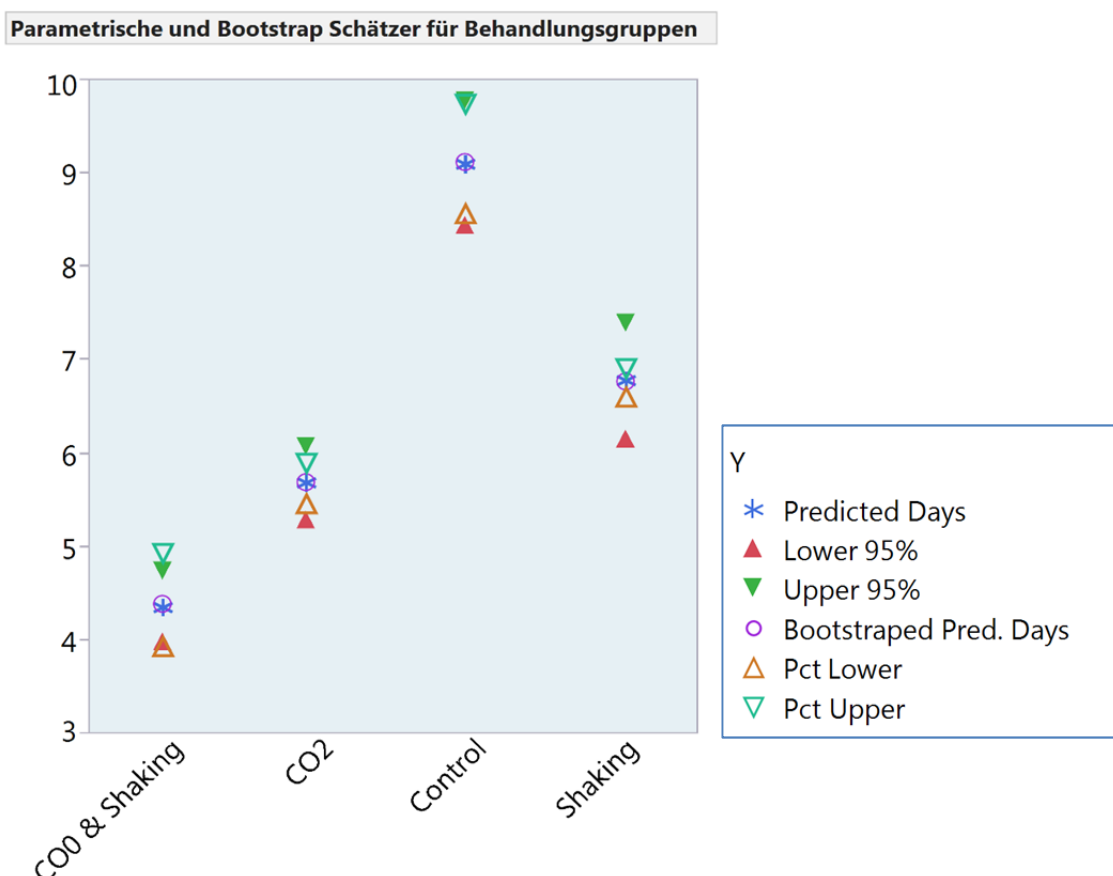


Abbildung 7: Bootstrapschätzer für Zeit und Konfidenzintervalle

Die Schätzer für die benötigte Zeitspanne stimmen auch in allen Gruppen weitgehend überein. Hier bestätigt die Bootstrapmethode die Ergebnisse der asymptotischen Schätzung, was aber keineswegs immer der Fall sein muss.

## 5 Fazit

Bootstrapverfahren erlauben einen zuverlässigen und detailreichen Blick auf die Verteilung einer beliebigen Statistik. Sie sind leicht zu berechnen und ihre Aussagen sind nicht unbedingt konservativer als ihre parametrischen Entsprechungen. Sie stellen auf jeden Fall ein probates Mittel zur Überprüfung der üblichen, durch asymptotische Abschätzungen gewonnenen, Konfidenzintervalle der Statistiken dar.



**Abbildung 8:** verschiedene Schätzmethoden

Mit Blick auf den Titel dieses Artikels möchte ich annehmen, dass Mathematiker und Statistiker sich in ihrer Ausbildung weiterhin mit Verteilungen und deren asymptotischer Annäherung befassen werden (müssen). Für Studierende anderer Wissenschaften ist aber durchaus vorstellbar, dass die durch Bootstrapverfahren gewonnenen Informationen für die eigene Arbeit ausreichen. Zumindest aber sind sie ein hervorragendes Mittel zur Motivation, sich mit Verteilungen von Statistiken zu befassen.

**Literatur**

- [1] Bradley Efron, Bootstrap Methods: Another Look at the Jackknife, s): B. Efron  
The Annals of Statistics, Vol. 7, No. 1 (Jan., 1979), pp. 1-26  
<http://www.jstor.org/stable/2958830>
  
- [2] Bradley Efron and Gail Gong, A Leisurely Look at the Bootstrap, the Jackknife,  
and Cross-Validation The American Statistician, Vol. 37, No. 1, (Feb., 1983), pp.  
36-48  
<http://www.jstor.org/stable/2685844>