

Bootstrap-Methode

Informatik
Hauptcampus

H O C H
S C H U L E
T R I E R

- Methode zur Approximation der Verteilung eines Parameters auf Basis von Stichprobendaten
- Das Verfahren wurde 1977 von Bradley Efron in dem Paper „Another Look at the Jackknife“ vorgestellt und stellt eine Weiterentwicklung der Jackknife-Methode dar
- Bootstrapping bietet ein breiteres Anwendungsspektrum als die Jackknife-Methode
- Die Präsentation basiert hauptsächlich auf den Inhalten des Papers „Another Look at the Jackknife“

- Schätzung von Konfidenzintervallen
- Durchführung von Hypothesentests ohne starke Annahmen über die zugrunde liegende Verteilung der Daten
- Schätzung von Standardfehlern
- Schätzen des Bias in statistischen Schätzern
- Viele Anwendungen in Bereich Machine-Learning wie z.B. Cross-Validation und Performance-Bewertung

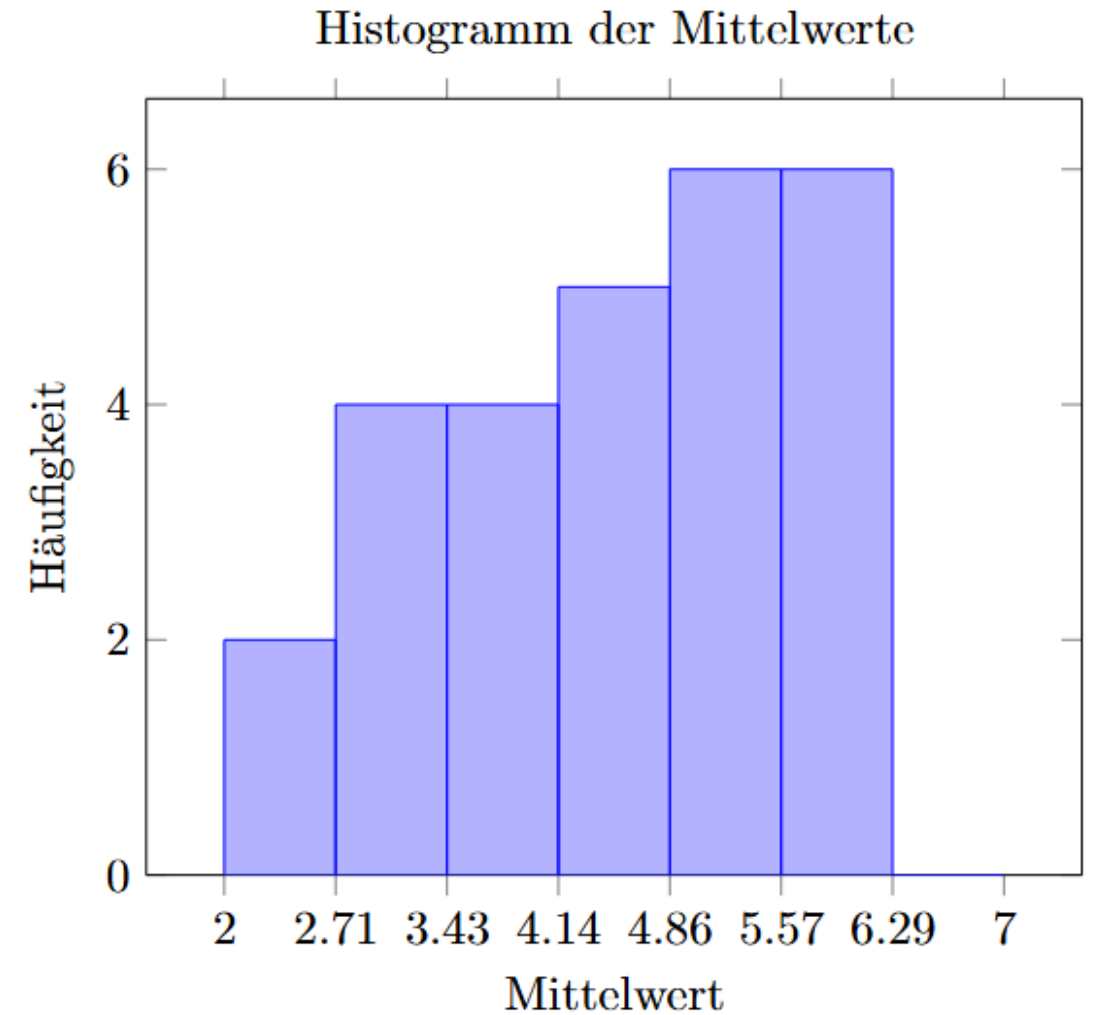
- Gegeben: Stichprobe $X=(X_1,\dots,X_n)$ aus einer unbekannten Verteilung F mit Realisierung $x=(x_1,\dots,x_n)$
- Definition einer Zufallsvariablen $R(X,F)$ die sowohl von der Stichprobe X als auch von der Verteilung F abhängt
- Ziel: Stichprobenverteilung von R approximieren
- Die Stichprobenverteilung ist die Verteilung von R die entsteht durch wiederholtes Ziehen von Stichproben der Länge n aus F

1. Gegeben ist eine Stichprobe $X=(X_1, \dots, X_n)$ mit Realisierung $x=(x_1, \dots, x_n)$
2. Konstruktion der Verteilung $\hat{F}$ basierend auf Stichprobe wobei jeden Datenpunkt $x_1, \dots, x_n$ die gleiche Wahrscheinlichkeit zugewiesen wird
3. Ziehen von Bootstrap-Stichproben $X^*=(X^*_1, \dots, X^*_n)$ aus $\hat{F}$ mit Zurücklegen
4. Approximation der Stichprobenverteilung von $R(X, F)$ durch die Stichprobenverteilung von $R(X^*, \hat{F})$

- Sei die „unbekannte Verteilung“ F mit der Zähldichte p wie folgt definiert:
 $p(1)=1/16$, $p(2)=4/16$, $p(4)=5/16$,
 $p(5)=1/16$, $p(7)=4/16$, $p(9)=1/16$, der Mittelwert $\mu=5,6$
- Es werden basierend auf F folgende Stichproben gezogen
- $R(X,F)$ ist eine Zufallsvariable auf den Mittelwert der Stichprobe

Stichprobe	Zug 1	Zug 2	Zug 3	Mittelwert $R(X, F)$
1	4	7	4	5.00
2	2	7	7	5.33
3	2	5	7	4.67
4	2	4	7	4.33
5	7	2	4	4.33
6	5	2	5	4.00
7	9	2	7	6.00
8	1	2	4	2.33
9	9	2	1	4.00
10	9	2	1	4.00
11	4	4	7	5.00
12	9	7	2	6.00
13	4	2	7	4.33
14	2	7	2	3.67
15	9	5	4	6.00
16	7	7	2	5.33
17	4	7	4	5.00
18	2	7	5	4.67
19	2	1	7	3.33
20	4	4	2	3.33
21	9	7	2	6.00
22	4	4	2	3.33
23	4	4	1	3.00
24	4	7	7	6.00
25	2	4	1	2.33
26	7	7	1	5.00
27	9	7	2	6.00

- Stichprobenverteilung der Zufallsvariablen $R(X, F)$

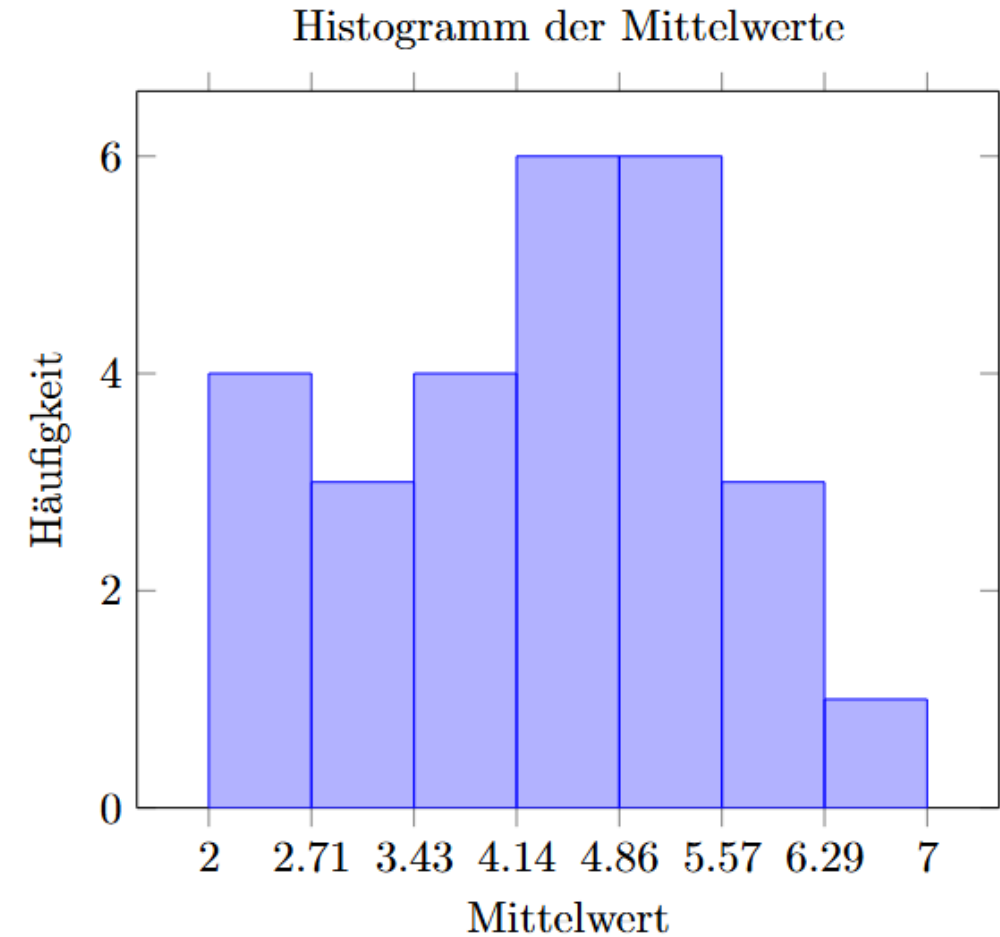


1. Gegeben ist eine Stichprobe $X=(X_1, \dots, X_3)$ mit Realisierung $x=(2, 4, 7)$ die aus einer unbekannten Verteilung F stammt
2. Konstruktion der Verteilung $F^\wedge$ mit der Zähldichte $p^\wedge$ für die gilt: $p^\wedge(2)=1/3$, $p^\wedge(4)=1/3$, $p^\wedge(7)=1/3$

3. Durch ziehen aus der Verteilung $F^{\wedge}$ entstehen folgende Bootstrap-Stichproben und deren Mittelwerte

Bootstrap-Stichprobe	Zug 1	Zug 2	Zug 3	Mittelwert $R(X^*, \hat{F})$
1	2	2	2	2.00
2	2	2	4	2.67
3	2	2	7	3.67
4	2	4	2	2.67
5	2	4	4	3.33
6	2	4	7	4.33
7	2	7	2	3.67
8	2	7	4	4.33
9	2	7	7	5.33
10	4	2	2	2.67
11	4	2	4	3.33
12	4	2	7	4.33
13	4	4	2	3.33
14	4	4	4	4.00
15	4	4	7	5.00
16	4	7	2	4.33
17	4	7	4	5.00
18	4	7	7	6.00
19	7	2	2	3.67
20	7	2	4	4.33
21	7	2	7	5.33
22	7	4	2	4.33
23	7	4	4	5.00
24	7	4	7	6.00
25	7	7	2	5.33
26	7	7	4	6.00
27	7	7	7	7.00

4. Approximation der
Stichprobenverteilung von $R(X, F)$
durch die Bootstrap-Stichproben
Verteilung von $R(X^*, \hat{F})$



- Wenn n die Stichprobengröße ist gibt es n^n verschiedene Möglichkeiten aus $F^\wedge$ eine Stichprobe zu ziehen
- Deshalb wird häufig die Monte-Carlo-Methode eingesetzt
- Es werden Stichproben $X^c = (X^{1c}, \dots, X^{Nc})$ gezogen mit Realisierungen $x^{1c}, \dots, x^{Nc}$ aus $F^\wedge$
- Es kann gezeigt werden das die Stichprobenverteilung von $R(X^*, F^\wedge)$ sehr schnell angenähert werden kann durch die Stichprobenverteilung von $R(X^c, F^\wedge)$
- Dadurch wird der Rechenaufwand extrem reduziert

- Sei P die Wahre Stichprobenverteilung von $R(X, F)$ und P^* die Bootstrap-Stichproben Verteilung von $R(X^*, \hat{F})$
- Der Wert von R kann unter Umständen auch über eine Funktion f und die zugehörige kumulierte Wahrscheinlichkeitsfunktion bestimmt werden, ausgedrückt als $f(P_{\text{cdf}}(X, F)_n)$ beziehungsweise $f(P_{\text{cdf}}(X^*, \hat{F})_n)$, wobei n die Stichprobengröße ist
- Dies geht für gewisse Statistische Größen wie den Mittelwert, die Varianz und die Kovarianz

- Zu zeigen ist, dass die $P_{\text{cdf}}(X^*, F^{\wedge})_n$ Verteilung die Verteilung $P_{\text{cdf}}(X, F)_n$ annähert bzw. schätzt
- Die Verteilungen $P_{\text{cdf}}(X^*, F^{\wedge})_n$ und $P_{\text{cdf}}(X, F)_n$ hängen von der Stichprobengröße n ab, daher werden die Grenzverteilungen (limiting distributions) $P_{\text{Lcdf}}(X^*, F^{\wedge})$ und $P_{\text{Lcdf}}(X, F)$ mit den Zähldichten $p_{\text{Lcdf}}(X^*, F^{\wedge})$ und $p_{\text{Lcdf}}(X, F)$ betrachtet

- Es wird gezeigt das $p_{\text{Lcdf}}(X^*, F^{\wedge})$ die Funktion $p_{\text{Lcdf}}(X, F)$ annähert
- Hierfür wird folgendes gezeigt: $\sup_x | p_{\text{Lcdf}}(X^*, F^{\wedge})(x) - p_{\text{Lcdf}}(X, F)(x) | \rightarrow 0$, dies wird auch uniform convergence genannt

- B. Efron. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7(1):1– 26, 1979
- Formaler Beweis der: <https://www.stat.purdue.edu/~dasgupta/bootstrap.pdf>
- Einfach Erläuterung des bootstrapping Beweises:
<https://www.youtube.com/watch?v=BiNcdYbyiWw&t=369s>
- Einfache Erläuterung des Bootstrapping Verfahrens:
<https://www.youtube.com/watch?v=Xz0x-8-cgaQ&t=404>