Angewandte Statistik

Julian Hinz — Universität Bielefeld

Session 12

Bootstrapping

Lernziele

- Wiederholung Stichprobenziehung und
- Einführung in Bootstrapping

Wiederholung: Stichprobenziehung

Stichprobenziehung

- Grundgesamtheit häufig schwierig komplett zu beobachten
- daher: Stichproben
- Beispiel: Zensus, Wahlbefragung, …
- hier: Zufallsstichproben

Beispiel mit R

- Setting: Europawahl 2024
- Wähler mit Attributen und Wahlpräferenzen
- große Grundgesamtheit, daher samplen
- Parameter der Grundgesamtheit: z.B. der wahre Anteil von Studierenden unter den Wählern
- Schätzer für Parameter der Grundgesamtheit: Anteil der Studierenden, berechnet aus der zufälligen Stichprobe
  - da Stichprobe zufällig ist, repräsentativ für Grundgesamtheit und Schätzer unverzerrt

```
> wähler
        person_id alter einkommen         beruf verheiratet kinder
            <int> <int>     <num>        <char>      <char>  <num>
     1:         1    37     48193   Selbständig        Nein      1
     2:         2    78     13735   Arbeiter/in        Nein      0
     3:         3    41     20673 Studierende/r          Ja      1
     4:         4    22     19289   Selbständig        Nein      0
     5:         5    79     19717   Arbeiter/in          Ja      1
    ---
 99996:     99996    27     32280   Arbeiter/in          Ja      0
 99997:     99997    70     22035   Arbeiter/in        Nein      0
 99998:     99998    29     21863   Arbeiter/in        Nein      0
 99999:     99999    24     24526 Angestellte/r        Nein      0
100000:    100000    44     19918   Arbeiter/in        Nein      1
        verkehrsmittel              partei
                <char>              <char>
     1:       Fahrrad Konservative Partei
     2:          Auto     Liberale Partei
     3:          Auto      Soziale Partei
     4:          Auto Konservative Partei
     5:          Auto      Soziale Partei
    ---
 99996:          Auto      Soziale Partei
```

zufällige Stichprobe

- n Beobachtungen aus der Grundgesamtheit zufällig auswählen
- Beobachtungen sollten nicht mehrfach vorkommen
- Stichprobengröße beeinflusst *Standardfehler*
  - je größer desto repräsentativer!

Sampling-Methoden im Überblick

- Einfache Zufallsstichprobe (Simple Random Sampling) mit/ohne Zurücklegen
- Geschichtete Stichprobe (Stratified Sampling): Reduziert Varianz in Teilgruppen
- Klumpenstichprobe (Cluster Sampling): Ganze Cluster (z.B. Gemeinden) auswählen
- Systematische Stichprobe: Auswahl jedes k-ten Elements aus geordneter Liste
- Varianzen der Schätzer hängen vom Stichprobendesign ab

Wiederholung: Jackknife-Resampling

- Leave-One-Out-Ansatz: Für $i = 1, \ldots, n$ entferne Datenpunkt i aus der Stichprobe.
- Schätzer in jeder Teilstichprobe: $\hat{\theta}_{(i)}$.
- Jackknife-Bias:

$$\text{Bias}_{\text{jack}} = (n-1)\left(\bar{\theta}_{(\cdot)} - \hat{\theta}\right), \qquad \bar{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$$

- Jackknife-Varianz:

$$\text{Var}_{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \bar{\theta}_{(\cdot)}\right)^2.$$

- Vorteil: Einfache Bias-Korrektur; Nachteil: konservativ für kleine n.

```
#| code-fold: false
jack_means <- sapply(seq_along(wähler$alter),
```

Bootstrapping

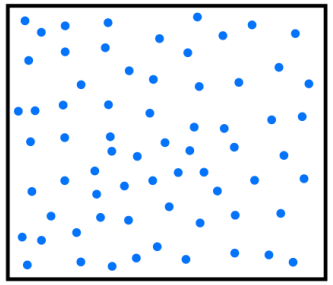
Einführung in Bootstrapping

- Statistiken aus Stichproben sind *Zufallsvariablen*
  - Stichprobenverteilung mit Mittelwert und Standardfehler
- Beispiel: Anteil der Studierenden in Stichprobe
  - Schätzer für den wahren Anteil der Studierenden bei allen Wählern

Einführung in Bootstrapping

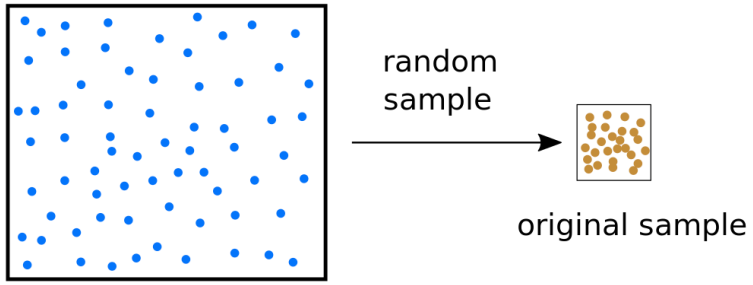
- Problem: In Realität schwierig sehr viele Stichproben zu ziehen
- Lösung: Mit einer einzigen Stichprobe arbeiten!

Einführung in Bootstrapping

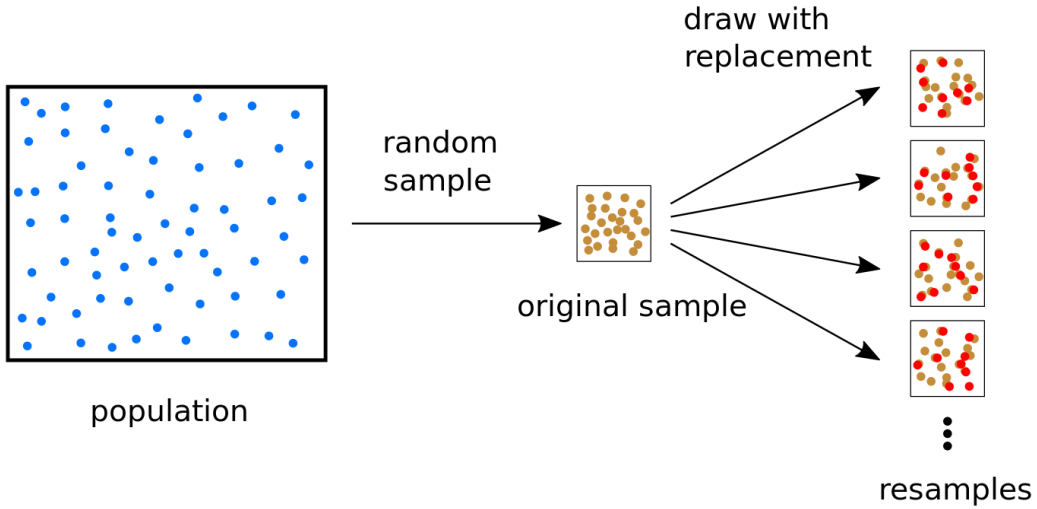


population

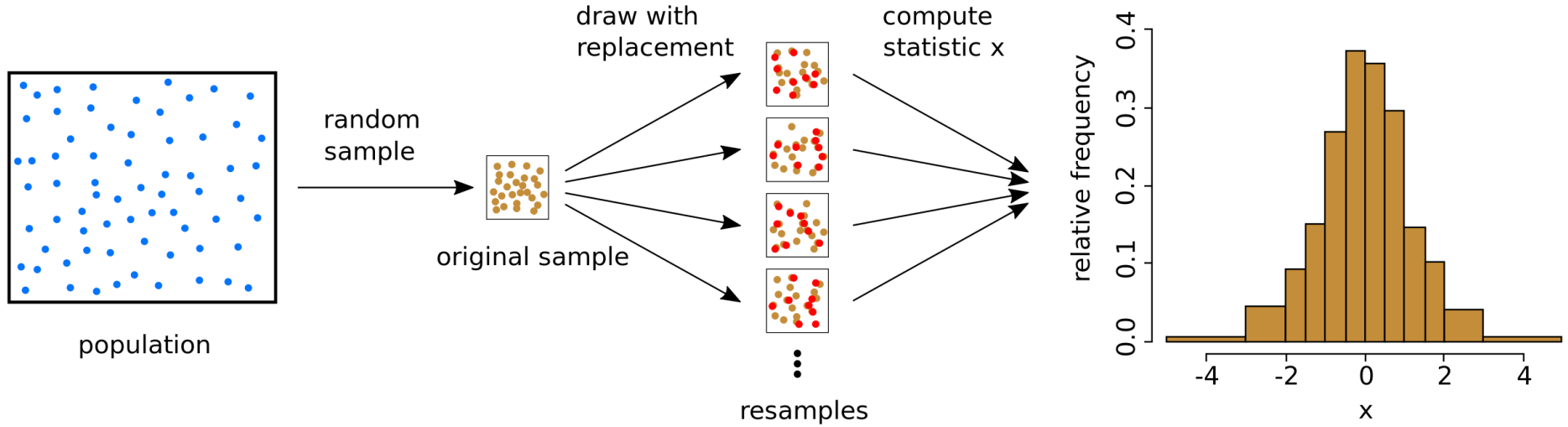
Einführung in Bootstrapping



Einführung in Bootstrapping



Einführung in Bootstrapping

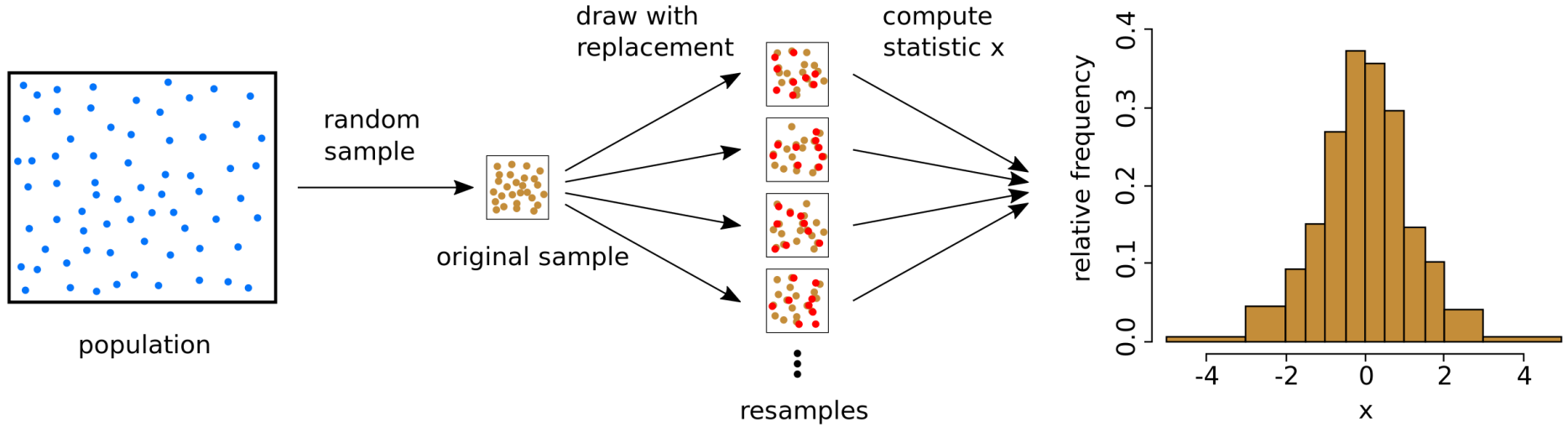


Einführung in Bootstrapping

“To pull oneself up by one’s bootstraps” — sich am eigenen Schopf aus dem Sumpf ziehen

Einführung in Bootstrapping

Einführung in Bootstrapping



Ursprung des Bootstrap

- Bradley Efron: “Bootstrap methods: another look at the jackknife” (1979)
- baut auf dem “jackknife” auf
- Bayesianische Erweiterung: Samplen mit Gewichten zwischen 0 und 1
- andere vorgeschlagene Namen für den “bootstrap”: Swiss Army Knife, Meat Axe, Shotgun, …

Vorteil gegenüber traditionellen Methoden

- Häufig Annahmen über Verteilung oder Momente (Mittelwert, Varianz, …)
- Beispiel: Eine Regression mit normalverteilter Zielvariable
- Annahme bei Bootstrap: empirische Verteilungsfunktion kann tatsächliche Verteilungsfunktion hinreichend gut approximieren
  - Stichprobe nicht zu klein
  - Stichprobe repräsentativ

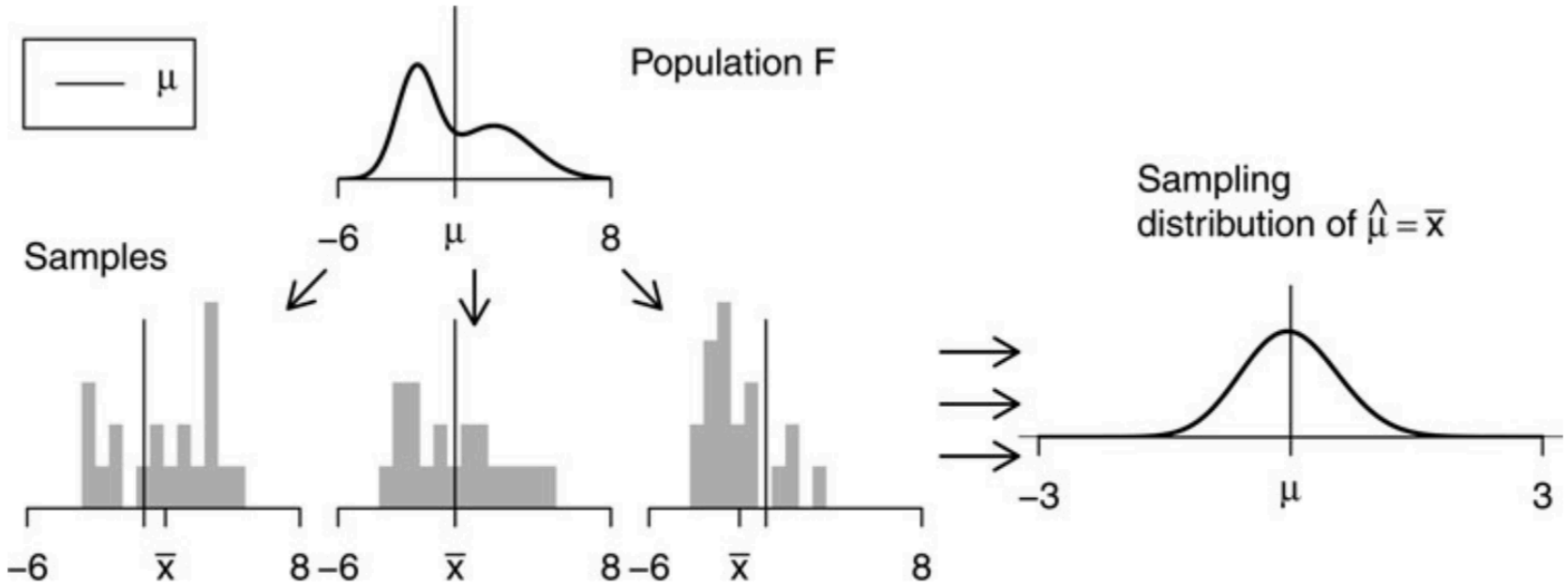
Unsicherheit des Schätzers

- Parameter der Grundgesamtheit aus nur einer Stichprobe schätzen
  - z.B. Anteil der Studierenden unter Wählern
- jede Stichprobe zufällig und Schätzer somit eine Zufallsvariable
  - wie gut ist der Schätzer?
- wenn man theoretische Verteilung kennt, mittels Standardfehler Konfidenzintervall berechnen

$$\widehat{\mu} \pm 1.96 \cdot SE$$

—

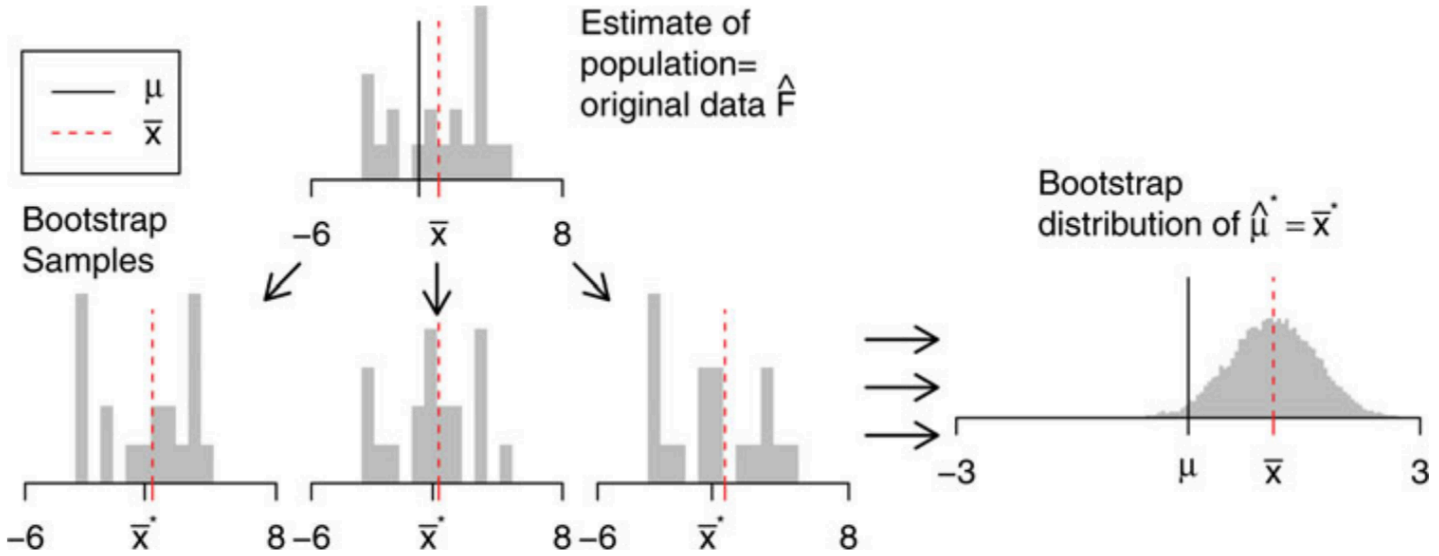
Normalerweise



Quelle: Hesterberg (2015)

- mehrere echte Stichproben aus einer Grundgesamtheit
- Stichprobenverteilung und Schätzer für Parameter der Grundgesamtheit
- 95% Konfidenzintervalle aus $\widehat{\mu} \pm 1.96 \cdot SE$

Bootstrap



Quelle: Hesterberg (2015)

- Idee: Stichprobe eine (repräsentative) Miniatur der Grundgesamtheit
- Bootstrap-Stichproben durch ziehen von neuen Stichprobe *mit Zurücklegen*
- 95% Konfidenzintervall aus 2.5% und das 97.5% Quantil der Bootstrap-Stichproben

Zurück zur Anwendung

- *eine* Stichprobe ziehen und mit dieser arbeiten

```
bootstrap_umfrage <- rep_sample_n(wähler,
                                  size = stichprobe_n,
                                  replace = TRUE, # wichtig!
                                  reps = anzahl_umfrage)
```

Algorithmus des Bootstrap-Verfahrens

1. Ziehe aus der Stichprobe n Beobachtungen mit Zurücklegen.
2. Berechne den gewünschten Schätzer $\hat{\theta}_b^*$ in jeder Bootstrap-Stichprobe ($b = 1, \ldots, B$).
3. Wiederhole Schritt 1–2 B-mal, um die Bootstrap-Verteilung von $\hat{\theta}$ zu erhalten.

Beispiel: Bootstrap des Mittelwerts

```r
#| code-fold: false
B <- 2000
boot_means <- replicate(B, mean(sample(wähler$alter, size = stichprobe_n, replace = TRUE)))
hist(boot_means, breaks = 30,
     main = "Bootstrap-Verteilung des Mittelwerts",
     xlab = "Mittelwert (Alter)")
abline(v = mean(wähler$alter), col = "red", lwd = 2)
quantile(boot_means, probs = c(0.025, 0.975))
```

Bootstrap-Konfidenzintervalle

- Percentile-Methode: Quantile der Bootstrap-Verteilung (2.5 %, 97.5 %).
- Standard-Error-Methode: $\hat{\theta} \pm z_{1-\alpha/2}\, \mathrm{SE}^*_\theta$.
- Bias-korrigierte und beschleunigte Methode (BCa).
- Studentized Bootstrap: Resampling der standardisierten Statistik.

Parametrischer Bootstrap

- Annahme: Daten folgen einer param. Verteilung F_θ.
- Schätze Parameter $\hat{\theta}$ aus der Stichprobe.
- Simuliere B Datensätze aus $F_{\hat{\theta}}$.
- Berechne Bootstrap-Statistik in jedem simulierten Datensatz.

Block-Bootstrap: Motivation

- Zeitreihen oder räumliche Daten → **Autokorrelation** verletzt i.i.d-Annahme
- Idee: *Blöcke* fester Länge (l) mit Zurücklegen resamplen statt Einzelpunkte
- Bewahrt Korrelation **innerhalb** eines Blocks, bricht sie **zwischen** Blöcken

Block-Bootstrap: Umsetzung in R

```r
l  <- 12          # Blocklänge (z. B. Monate)
B  <- 2000
ts_boot <- replicate(B, {
  idx <- sample(seq_len(n - l + 1), ceiling(n / l), replace = TRUE)
  series_star <- unlist(lapply(idx, \(s) ts_data[s:(s + l - 1)]))[1:n]
  mean(series_star)                       # gewünschte Statistik
})
hist(ts_boot, breaks = 30,
     main = "Block-Bootstrap-Verteilung",
     xlab = "Mittelwert")
```

Bootstrap bei Regressionsmodellen

- Residual Bootstrap:
  1. Passe das Modell an, verwende die Residuen für das Resampling.
  2. Erstelle neue Zielvariable: $\hat{y}_i + e_i^*$.
  3. Passe das Modell auf die neuen Daten an und ermittle die Bootstrap-Schätzer.
- Paarweiser Bootstrap:
  1. Ziehe Beobachtungen mit Zurücklegen.
  2. Passe Modell an jedem Resample an.

Zusammenfassung

- Bootstrap zieht **mit Zurücklegen** aus *einer* Stichprobe, um unbekannte Stichprobenverteilung eines Schätzers zu approximieren
- Aus **Bootstrap-Verteilung** erhalten wir Standardfehler, Bias-Schätzung und Konfidenzintervalle (Percentile, SE- oder BCa-Methode) **ohne starke Verteilungsannahmen**
- Funktioniert gut bei ausreichend großer, repräsentativer Stichprobe
- Vorsicht bei Abhängigkeiten, extremen Ausreißern oder sehr kleinem (n)

Weiterführende Literatur

- Bradley Efron & Robert J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman & Hall.
- Antony C. Davison & David V. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge University Press.
- Tim Hesterberg (2015). What Teachers Should Know About the Bootstrap. *The American Statistician*.