Mathematik II für Biologen

Die Bootstrap-Methode

Stefan Keppeler

2. Juli 2008

Eberhard Karls Universität Tübingen

- Stichprobe $x_1, \ldots, x_n$ (~ 10 Werte)
- Annahme: Realisierung von $X_1, \ldots, X_n$ iid
- Mittelwert $\overline{X} = \frac{1}{n}(x_1 + \ldots + x_n)$
- Genauigkeit / Fehler?
- Bestimme z.B. das 95%-Vertrauensintervall für den Erwartungswert
- Im Sinne eines t- oder z-Tests:

$$\left[\overline{X} - 2\frac{s}{\sqrt{n}}\,,\ \overline{X} + 2\frac{s}{\sqrt{n}}\right]$$

wobei $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{X})^2$ (empirische Varianz)

genauer: z-Test: $2 \mapsto 1,96$
t-Test: $2 \mapsto 2,57..2,23..1,98$ (für $5..10..100$ FHGe)

Also bestimmbar aus Mittelwert und empirischer Varianz!

Wie für andere Größen? ...z.B. für den Median?

- vielleicht Vertrauensintervall aus Vorzeichen- oder Wilcoxon-Test
- Ohne Faustregeln (falls n nicht groß genug) muss man Test evt. für viele Nullhypothesen wiederholen.

Und für noch andere Größen?

...z.B. für einen Korrelationskoeffizienten?

- Wie ist der unter einer bestimmten Nullhypothese verteilt?
- Wird schnell schwierig...

bootstrap, wörtlich: *Stiefelschlaufe/-riemen*
engl. Wendung: *to pull oneself up by one's bootstrap*
deutsch: *sich an den eigenen Haaren / am eigenen Schopf aus dem Sumpf ziehen*

Theodor Hosemann (1807-1875)

Münchhausen erzählt: *Bei der Verfolgung eines Hasen wollte ich mit meinem Pferd über einen Morast setzen. Mitten im Sprung musste ich erkennen, dass der Morast viel breiter war, als ich anfänglich eingeschätzt hatte. Schwebend in der Luft wendete ich daher wieder um, wo ich hergekommen war, um einen größeren Anlauf zu nehmen. Gleichwohl sprang ich zum zweiten Mal noch zu kurz und fiel nicht weit vom anderen Ufer bis an den Hals in den Morast. Hier hätte ich unfehlbar umkommen müssen, wenn nicht die Stärke meines Armes mich an meinem eigenen Haarzopf, samt dem Pferd, welches ich fest zwischen meine Knie schloss, wieder herausgezogen hätte.*

Idee des Bootstrap:

- erzeuge künstlich viele "ähnliche" Stichproben
- berechne gewünschte Größe für diese, und bestimme Mittelwert und empirische Varianz s^2 der Werte*
- 95%-Vertrauensintervall wieder als $\pm 2s$-Intervall um den Mittelwert (ohne $\sqrt{n}$!)

Fragen:

- Was heißt "ähnlich"?
- Wie erzeugt man die Daten?

*oder besser: Bestimme VI gleich aus Histogramm der Bootstrap-Daten

- "ähnlich": Gleiche Verteilung(sfunktion) wie die Werte der Ausgangsstichprobe
- **aber:** Verteilung ist nicht bekannt!
- approximiere durch empirische Verteilung(sfunktion)

$$F(x) = \frac{\#\{x_i : x_i \leq x\}}{n} \qquad \text{(aus Stichprobe)}$$

- Ziehe neue Stichprobe:
  - $a_1, \ldots a_n$ gleichverteilt aus [0,1]
  - $b_j = F^{-1}(a_j)$ neue Stichprobe

Praktisch heißt das:

- ziehe (mit Zurücklegen!) n Werte aus der Originalstichprobe
- ergibt eine "neue" Stichprobe / eine Bootstrap-Stichprobe
- wiederhole oft (N mal, N groß)

Berechne dann die gewünschte Größe für jede Bootstrap-Stichprobe, und bestimme daraus die gesuchte Schwankung.

Mittelwerte: Beispiel*

- 7 Mäuse erhalten einen neue Behandlung – Gruppe **B**
- 9 Mäuse erhalten keine Behandlung – (Kontroll-)Gruppe **K**
- dann: Testoperation für alle
- danach: Messung der Überlebensdauer (in Tagen)
  - **B:** 94 197 16 38 99 141 23
    Mittelwert: 86,86
  - **K:** 52 104 146 10 51 30 40 27 46
    Mittelwert: 56,22
- Ist das ein signifikanter Unterschied?

*nach Efron & Tibshirani *An Introduction to the Bootstrap*

Antworten:

- ohne Bootstrap: 95%-VI für Mittelwerte: $\overline{X} \pm 2s/\sqrt{n}$

  B: $86{,}86 \pm 50{,}47\,,$ **K:** $56{,}22 \pm 28{,}32$

  VIe überlappen stark, also kein signifikanter Unterschied

- mit Bootstrap:
  - Originalstichprobe
    B: 94 197 16 38 99 141 23
  - ziehe Bootstrap-Stichproben, z.B.
    B$'_1$: 23 99 38 94 197 94 38
    B$'_2$: 99 99 38 94 38 16 94 etc.
  - Mittelwerte dazu: $83{,}29$ $68{,}29$ etc.
  - $s' := \sqrt{\text{emp. Varianz dieser Mittelwerte}} \approx 23{,}01$
    (hier mit $N = 10000$ Bootstrap-Stichproben)
  - Vertrauensintervall für **B**: $\overline{X} \pm 2s' = 86{,}86 \pm 46{,}02$
    (analog für **K**)

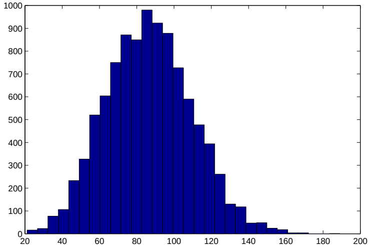
MATLAB-Code

- zur Berechnung des Vertrauensintervalls
- des Mittelwerts
- mittels Bootstrap

```
>> B=[94 197 16 38 99 141 23];
>> n=10000;
>> mittelwerte=bootstrp(n,'mean',B);
>> sqrt(var(mittelwerte))

ans = 23.0122


>> hist(mittelwerte,30)
```



Beispiel: Durchschnittsnoten eines Hochschuleingangstests (LSAT) sowie einer Grundstudiumsprüfung (GPA) von 15 Hochschulen:*

Schule	1	2	3	4	5	6	7	8
LSAT	576	635	558	578	666	580	555	661
GPA	3,39	3,30	2,81	3,03	3,44	3,07	3,00	3,43

Schule	9	10	11	12	13	14	15
LSAT	651	605	653	575	545	572	594
GPA	3,36	3,13	3,12	2,74	2,76	2,88	2,96

- Stichprobe: $x_i = (\text{LSAT}_i, \text{GPA}_i)$
- Korrelation (Pearson): $r_{\text{LSAT},\text{GPA}} \approx 0{,}776$
- Aber mit welcher Genauigkeit?
- Bootstrap!

*nach Efron & Tibshirani *An Introduction to the Bootstrap*

Vertrauensintervall für die Korrelation mittels Bootstrap

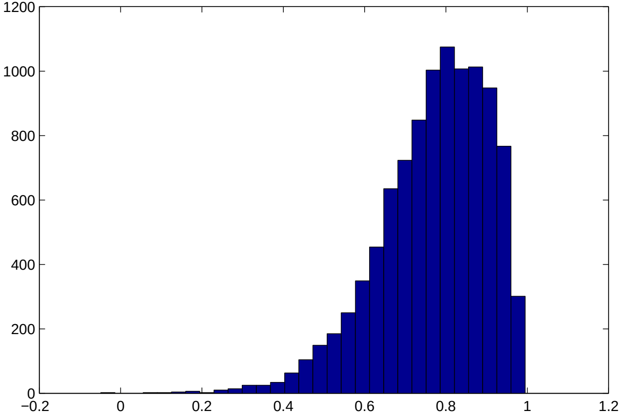
- Ziehe viele Bootstrap-Stichproben,
  Umfang 15, gezogen wird jeweils ein Paar x_i
- berechne deren Korrelationskoeffizienten...
- ...sowie die Varianz ${s'}^2$ derselben.
- MATLAB-Code

```
>> n=10000;
>> korrelationen=bootstrp(n,'corrcoef',lsat,gpa);
>> sqrt(var(korrelationen(:,2)))

ans = 0.1347
```

- 95%-Vertrauensintervall für $r_{\text{LSAT,GPA}}$: $0{,}776 \pm 0{,}269$
- > 1 ? Vielleicht noch besser Histogramm betrachten...

```
>> hist(korrelationen(:,2),30)
```



95%-Vertrauensintervall für $r_{\mathrm{LSAT,GPA}}$: $[0{,}45\,,\ 0{,}97]$