

# Bootstrapping

## Motivation

Viele Annahmen sind notwendig, um Verteilungen von Statistiken zu bestimmen, und sogar dann sind die Resultate oft nur asymptotisch bewiesen. Ohne die Verteilungen kann man keine richtige Statistik machen.

(Für Glättungsverfahren ist die Situation noch schlimmer, da wir keine starke Verteilungstheorie haben.)

Diese Probleme entstehen nicht nur bei komplexen Modellen (was ist die Verteilung eines geschätzten Koeffizienten in einer logistischen Regression?), sondern auch bei einfachen Statistiken:

1. Verteilung des Medians im allgemeinen
2. Mittelwert-Vergleich zweier normalverteilten Zufallsvariablen mit unbekanntem Varianzen
3. Test für den Korrelationskoeffizient zweier nicht normalverteilten Zufallsvariablen

# Konzepte

## Klassische Statistik

Grundgesamtheit von Individuen mit Eigenschaften, die durch Zufallsvariablen,  $X, Y \dots$ , modelliert werden. Eine Zufällige Stichprobe der Größe  $n$  gibt die Daten:

$\{x_1, x_2, \dots, x_n\} \{y_1, y_2, \dots, y_n\} \dots$

und daraus werden univariate (Mittelwert, Median, ...) und multivariate ( $r$ , Modellkoeffizienten, ...) Statistiken berechnet.

Grundgesamtheit

Zufällige Stichprobe

Modellannahmen

Parameterschätzer

Standardfehler

Konfidenzintervalle

z.B. Gauß Test, t-test, Regression

Falls entsprechende Annahmen nicht gemacht werden können, sollten wir im Prinzip viele Stichproben erheben, um die empirische Verteilung der in jeder Stichprobe berechneten Statistik zu schätzen.

Das geht nicht, deshalb simuliert man dieses Verfahren.

## Bootstrapping

Alles was wir über die Stichprobe wissen ist in der Stichprobe. Wir schätzen die unbekannte Verteilung der Zufallsvariablen in der Grundgesamtheit mit der empirischen Verteilungsfunktion:

$$\hat{F}(x) = \frac{\#(\{x_i \leq x\})}{n}$$

Dann erheben wir B zufällige Stichproben der Größe n aus dieser Verteilung:

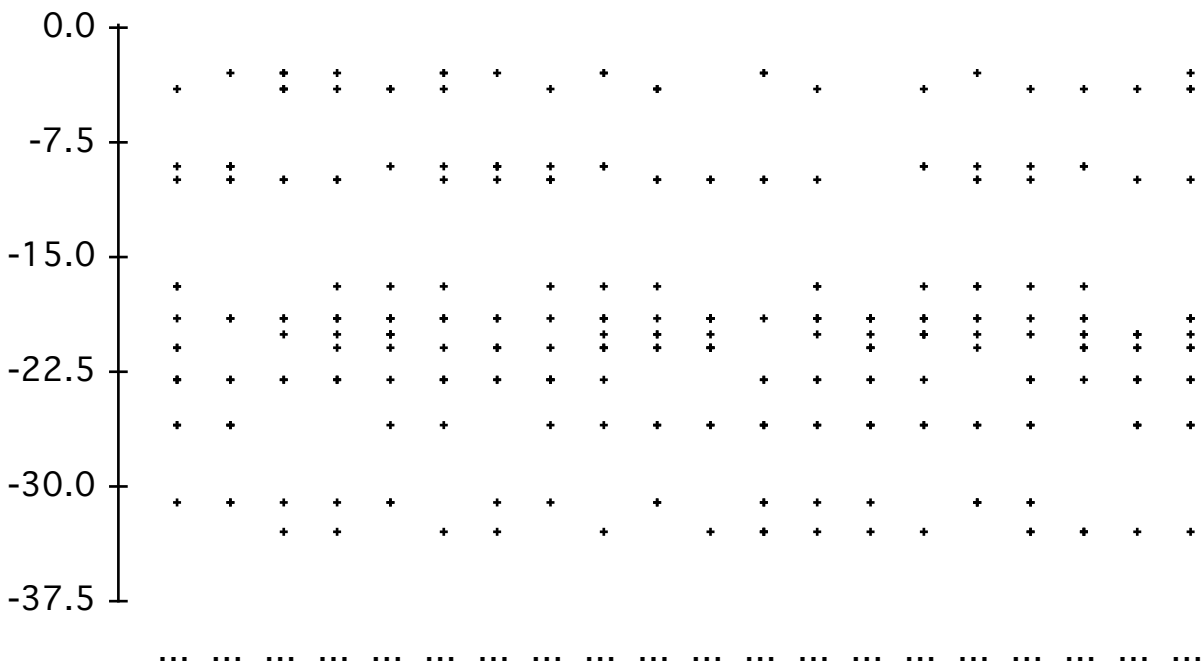
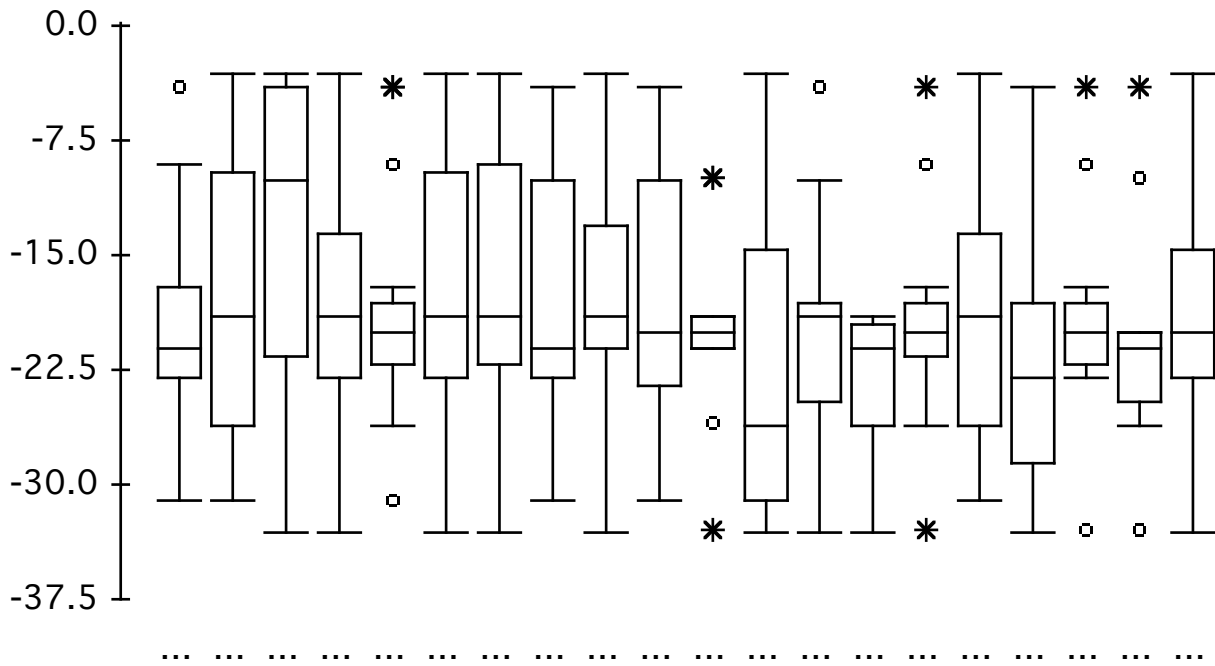
$$\{x_{j1}^*, x_{j2}^*, \dots, x_{jn}^*\}_{j=1, \dots, B}$$

(Stichproben mit Zurücklegen) Es gibt notwendigerweise fast immer multiple Werte aus der echten Stichprobe (zweimal  $x_4$  oder viermal  $x_{21}$  usw). Die für uns interessanten Statistiken (z.B.  $\theta$ ) werden für jede simulierte Stichprobe berechnet. Falls B sehr groß ist, können wir, im Prinzip, die Verteilung von  $\hat{\theta}$  schätzen. Wenigstens den Standardfehler:

$$se_B = \left[ \frac{\sum [s(x_j^*) - s(\cdot)]^2}{B - 1} \right]^{\frac{1}{2}}$$

$$s(\cdot) = \frac{1}{B} \sum s(x_j^*)$$

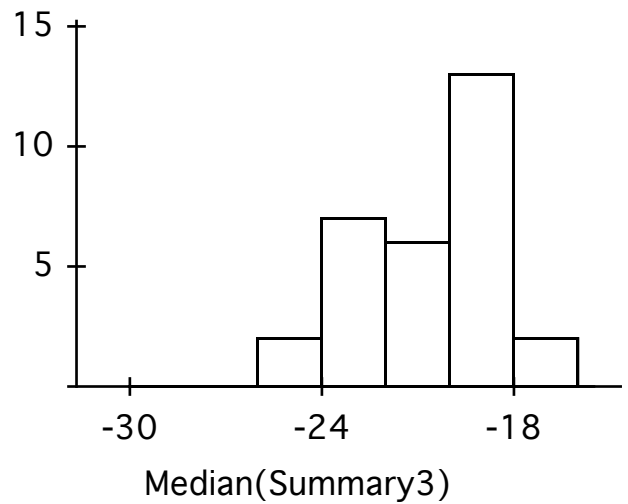
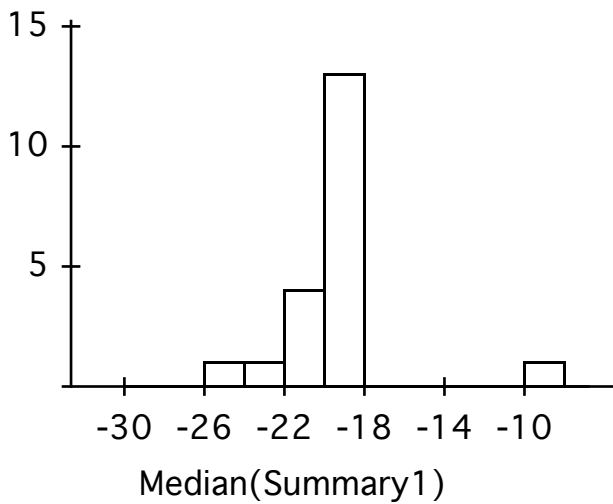
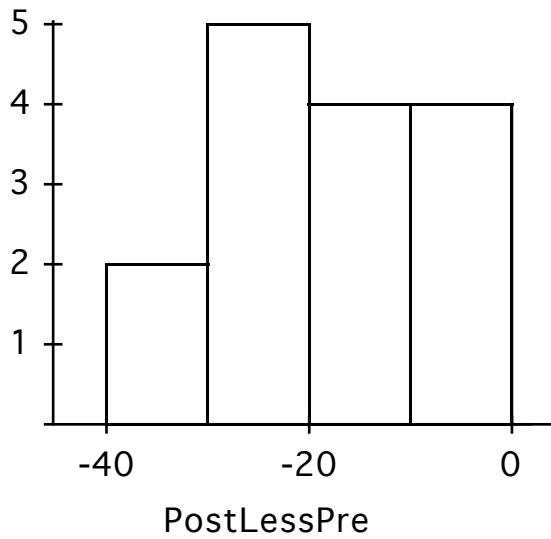
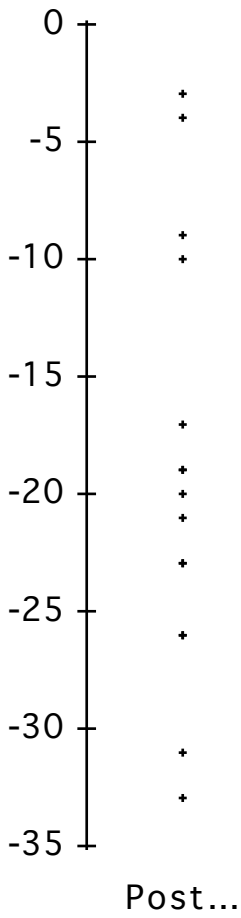
# Bootstrap Stichproben



# Captopril Daten

Summary of PostLessPre  
No Selector

Total Cases	15
Mean	-18.9333
Median	-20
StdDev	9.02747
Min	-33
Max	-3
Range	30



## Multinomialmodell für die Stichproben

$$P\{j_1 \text{ mal } x_1, j_2 \text{ mal } x_2, \dots\} = \frac{n!}{\prod j_i!} \left(\frac{1}{n}\right)^n$$

Es gibt  $\binom{2n-1}{n}$  verschiedene Bootstrapstichproben aus  $n$  Werten. ( $n$  identische Murmeln und  $n$  unterscheidbare Kästchen.)

$$\text{z.B. } n = 10 \quad \binom{19}{10} = 92378$$

und  $n = 20$  ergibt  $\simeq 6.9 \times 10^{10}$

$P(\text{ein bestimmter Wert } x_k \text{ nicht in einer Bootstrapstichprobe erscheint}) = \left(\frac{n-1}{n}\right)^n$

$$\text{z.B. } n = 10 \quad P(x_k \neq x_{j_i}^*) = 0.35$$

$P(\text{alle Werte in einer bestimmten Bootstrapstichprobe enthalten sind}) = P(x^* = x) = n!n^{-n}$

$P(\text{mindestens ein } x_k \text{ fehlt in einer bestimmten Bootstrapstichprobe}) = 1 - \frac{n!}{n^n}$  (0.99964 for  $n = 10$ )

## Nichtparametrische und Parametrische Bootstrapping

Im nichtparametrischen Fall werden keine Annahmen gemacht (außer Unabhängigkeit und Belanglosigkeit der Reihenfolge der Daten). Nur Werte aus der ursprünglichen Stichprobe können erhoben werden. Der Ereignisraum ist diskret.

Im parametrischen Fall wird  $F$  als bekannt angenommen, aber die Parameter nicht.  $\theta$  wird aus der Stichprobe geschätzt  $\hat{\theta}(x_1, x_2, \dots, x_n)$  und die Bootstrapstichproben werden aus der Verteilung

$$F(x, \hat{\theta})$$

erhoben.

Vorteil: Unser Ereignisraum ist nicht mehr diskret.

Nachteil: Die Strukturannahme.

## Bias (Verzerrung)

$$Bias_F = E[\hat{\theta}] - \theta(F)$$

Mit dem Bootstrapverfahren schätzen wir diesen Bias mit

$$Bias_B = \hat{\theta}^*(.) - \theta(\hat{F})$$

Eine Idee wäre dann unseren ursprünglichen Schätzer gegen Bias zu korrigieren:

$$\begin{aligned}\hat{\theta}_{adjustiert} &= \hat{\theta} - Bias_B \\ &= 2\hat{\theta} - \hat{\theta}^*(.)\end{aligned}$$

$\hat{\theta}$  (echte Welt) gleicht  $\theta(\hat{F})$  in der Bootstrap Welt

In der Praxis scheint dieses Verfahren nicht zu klappen, weil soviel Variabilität enthalten ist. "Wenn die Bias klein ist, macht es nichts aus. Wenn die Bias groß ist, ist der Schätzer wahrscheinlich schlecht." (Efron und Tibshirani).

## Warum funktioniert der Bootstrap?

Für  $n$  groß wird  $\hat{F}$  eine gute Approximation für  $F$  sein und man kann sich auf die Resultate verlassen (wohlgemerkt, numerische Resultate, nicht analytische Resultate). Wenn  $n$  klein ist, weißt man, das  $\hat{F}$  mag eine schlechte Approximation sein, aber sie ist alles was man hat. (c.f. das Problem beim t-test. Die Daten können wohl aus einer Normalverteilung stammen, aber wir können Pech mit unserer Stichprobe haben.)

In einer anderen Schreibweise setzen wir

$$\begin{aligned} F_0 &= F \\ F_1 &= \hat{F} \\ F_{2j} &= F_j^* \end{aligned}$$

## Zwei Ansichten

### Efron (der Vater vom Bootstrap)

Wir wollen  $\hat{\theta}$  als Schätzer für  $\theta$  gegeben  $F_0$

d.h. die Verteilung von  $\theta(F_1)$ , um  $\theta(F_0)$  zu schätzen.

Wir brauchen  $\theta^*$  als Schätzer für  $\hat{\theta}$  gegeben  $F_1$ .

d.h. die Verteilung von  $\theta(F_2)$  um  $\theta(F_1)$  und dadurch  $\theta(F_0)$  zu schätzen.

### Hall

Wir wollen  $\hat{\theta} - \theta$  gegeben  $F_0$

d.h. die Verteilung von  $\theta(F_1) - \theta(F_0)$

Wir haben die Verteilung von  $\theta^* - \hat{\theta}$  gegeben  $F_1$ .

d.h. die Verteilung von  $\theta(F_2) - \theta(F_1)$ .

Russische Puppen (Hall): das Verhältnis zwischen den Puppen 2 und 1 sollte uns etwas über das Verhältnis zwischen den Puppen 1 und 0 sagen.