

## Statistik II

- GLM (Regression und Varianzanalyse)
- Logistische Regression
- Loglineare Modelle
- Verallgemeinerte Lineare Modelle (GLIM)  
"Generalised Linear Interactive Models"
- Glättungen
- Bootstrapping

## Bootstrap Beispiel (wo wir Bootstrapping nicht brauchen)

Gegeben sei eine Zufallsvariable

$$Y \sim N(\mu, \sigma^2)$$

Der Mittelwert einer unabhängigen Stichprobe der Größe  $n$  hat die Verteilung

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

Für  $\sigma^2$  unbekannt gilt

$$t = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Sei die Stichprobe  $\{y_1, y_2, \dots, y_n\}$  und seien  $B$  Bootstrap Stichproben der Größe  $n$  gezogen:

$$\{y_{j1}, y_{j2}, \dots, y_{jn}\} \quad j = 1, \dots, B$$

mit Mittelwerten und Standardabweichungen

$$(\bar{y}_j, s_j) \quad j = 1, \dots, B$$

## Bootstrapping mit R

> library(boot) *(Library laden)*

> MyData ← read.table("MyData.txt",header=TRUE)

> attach(MyData)

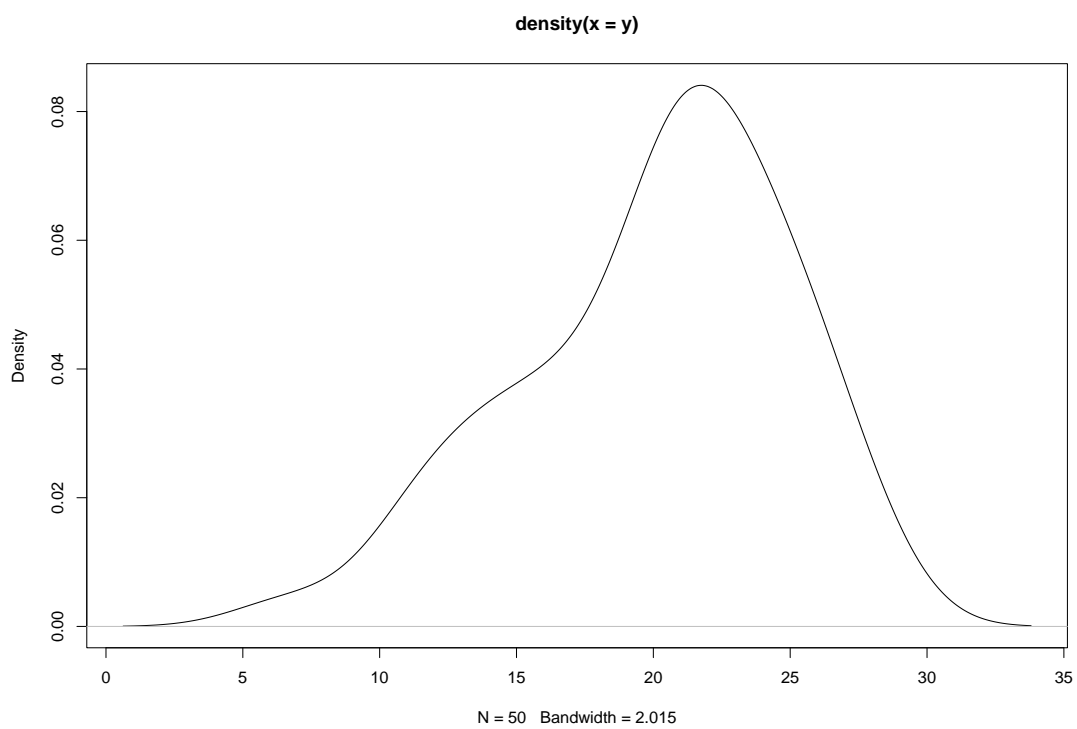
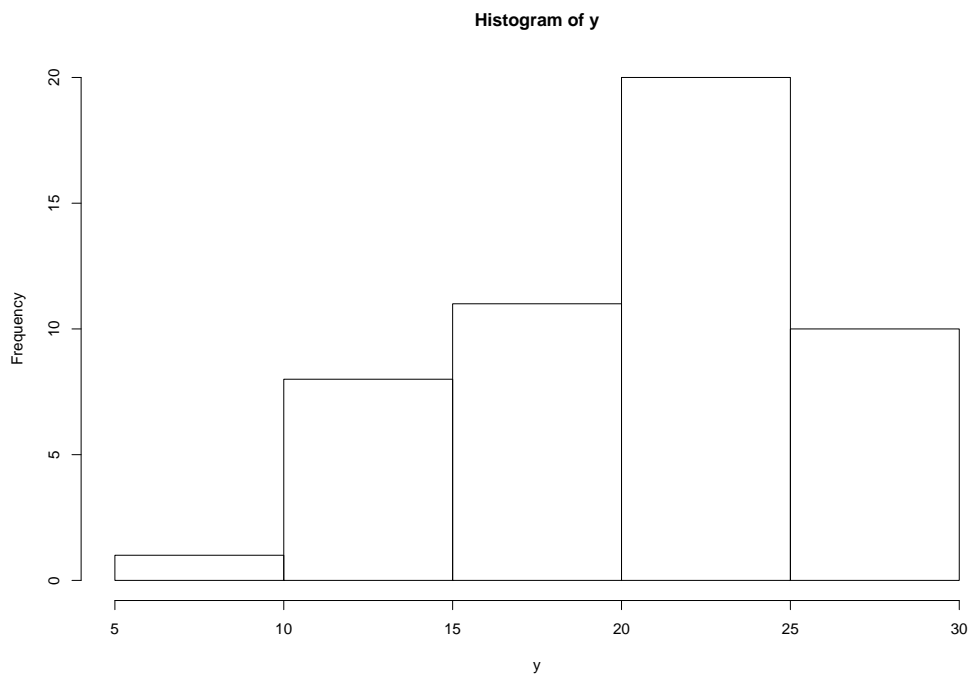
> y.boot←boot(MyData, function(y,i) mean(y[i]), R=999)  
*(Bootstrapsstichproben erzeugen)*

> plot(y.boot)

*(Histogram und Normal Quantil Plot für die Bootstrap Verteilung der Statistik)*

> boot.ci(y.boot,conf=c(0.95),type=c("stud","basic","perc")  
*(verschiedene KIs für die Statistik)*

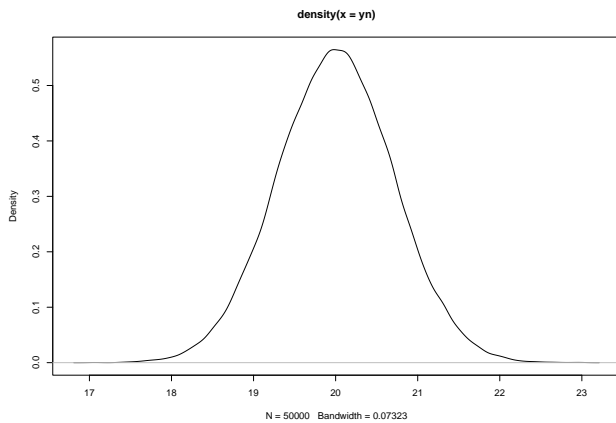
# Histogramm und Dichteschätzer für $n = 50$



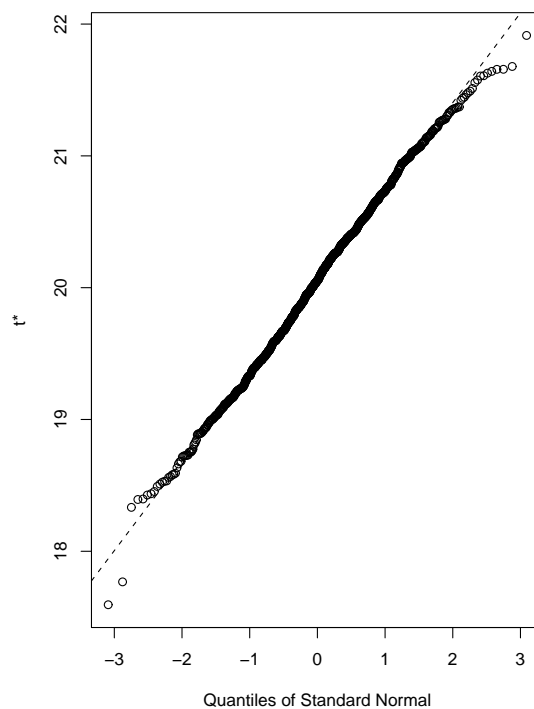
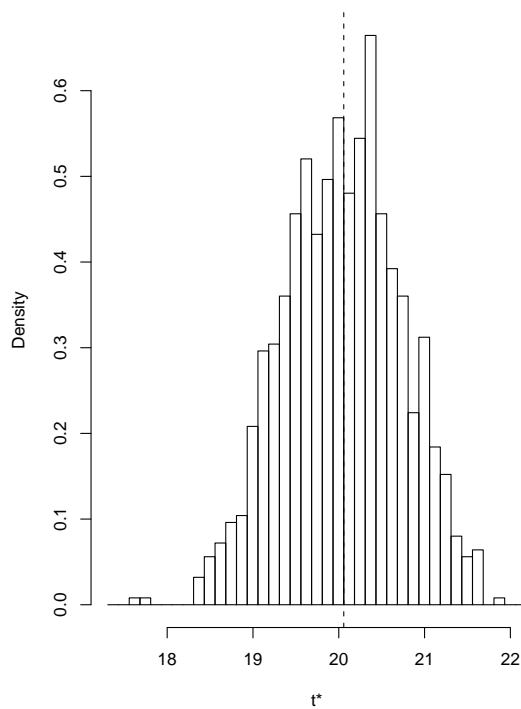
# Theoretische und Empirische Verteilungen

$$\mu = 20 \text{ Standardfehler} = 50/\sqrt{50} = 0.707$$

$$\bar{y} = 20.06 \text{ und } s = 0.68$$



Histogram of t



## **Bootstrap 95% Konfidenzintervalle aus R**

<b>Typ</b>	<b>Untere Grenze</b>	<b>Obere Grenze</b>
Normal	18.74	21.40
Basic	18.79	21.39
Percentile	18.72	21.33
BCa	18.70	21.30

## Variabilität

Wegen der ursprünglichen Stichprobe ( $F_1$  zu  $F_0$ ):  
hängt von  $n$  ab.

Wegen Bootstrapping ("Resampling") ( $F_2$  zu  $F_1$ ):  
hängt von  $B$  ab.

Gegeben  $n$ , kann man durch  $B$  die Gesamtvariabilität nicht  
viel kleiner machen.

## Bootstrap Konfidenz Intervalle

(a) **Bootstrap-t** ("stud" in R)

(Normalverwandter Ansatz für Lageparameter)

Für jede Bootstrapstichprobe  $b$  berechnet man die standardisierte Variable

$$z(b) = \frac{(\hat{\theta}^*(b) - \hat{\theta})}{se(b)}$$

Dann findet man, z.B., die 2,5% und 97,5% Werte von  $z(b)$  aus allen  $B$ ,  $\hat{t}_\alpha$  und  $\hat{t}_{(1-\alpha)}$ , um ein 95% KI für  $\theta$  zu erstellen:

$$P(\hat{t}_\alpha < \frac{(\hat{\theta} - \theta)}{\hat{se}} < \hat{t}_{(1-\alpha)}) = 1 - 2\alpha$$

führt zu

$$[\hat{\theta} - \hat{t}_{(1-\alpha)}\hat{se}, \hat{\theta} - \hat{t}_\alpha\hat{se}]$$

## Probleme mit Bootstrap-t

1. Schätzen von  $\hat{t}_\alpha$  und  $\hat{t}_{(1-\alpha)}$  ist schlecht. Es geht um Schwanzquantilen.
2. Wie soll  $se(b)$  geschätzt werden? Für  $\theta$  einen Lageparameter, wie den Mittelwert, gibt es eine Formel für  $se(b)$ , aber wir wissen, dass sie für kleine Stichproben schlecht ist. Für andere Parameter haben wir keine Formel. "Double-Bootstrapping" wird vorgeschlagen.
3. Diese Methode ist transformationsabhängig.  
Man bekommt andere Resultate, wenn man mit  $h(\theta)$  direkt, statt indirekt über  $\theta$ , arbeitet.

**(b) KI — direkter Ansatz ("perc"/"basic" in R)**

Wir nehmen die  $\alpha$  und  $1 - \alpha$  Werte von  $\hat{\theta}^*(b)$  aus allen B als  $(1 - 2\alpha)$  KI Endpunkte für  $\hat{\theta}$  und deshalb für  $\theta$  (Efron):

$$[v_{(\alpha)}^*, v_{(1-\alpha)}^*]$$

Oder (Hall) wir betrachten  $(\hat{\theta}^* - \hat{\theta})$  mit  $(1 - 2\alpha)$  KI Endpunkte:

$$[v_{(\alpha)}^* - \hat{\theta}, v_{(1-\alpha)}^* - \hat{\theta}]$$

Daraus ergibt sich ein  $(1 - 2\alpha)$  KI für  $\theta$  von

$$[2\hat{\theta} - v_{(1-\alpha)}^*, 2\hat{\theta} - v_{(\alpha)}^*]$$

Für  $\hat{\theta}^*$  symmetrisch gilt

$$\hat{\theta} - v_{(\alpha)}^* = v_{(1-\alpha)}^* - \hat{\theta}$$

und die KIs sind gleich.

## **Kommentare zum direkten Ansatz**

Diese Methode ist nicht transformationsabhängig.

Aber "Coverage" (Überdeckung) ist immer noch schlecht. Überdeckung eines KIs gibt die Prozente der Intervalle aus Simulationsstudien, die tatsächlich den wahren Parameter enthalten. Ein ideales 95% KI sollte eine Überdeckung von 95% haben. Interessant ist es, Überdeckung gegen Konfidenz zu plotten.

Es gibt viele Berichtigungsvorschläge, diese Intervalle zu verbessern.

## Bootstrapping Lowess

Wie soll man  $(X, Y)$  bootstrappen, um die Variabilität um eine Lowess Glättung zu schätzen?

1. Lowess Modell:  $y = m(x) + \epsilon$

Wir schätzen  $m(x)$  aus dem Datensatz und machen Bootstrapping mit den Residuen

$$\{e_i = y_i - \hat{m}(x_i)\}$$

Vorteil: leicht, kann verallgemeinert werden.

Nachteil:  $X$  könnte mit Fehlern bemessen werden.

2. Bivariate Bootstrapping

Bootstrap mit  $(X, Y)$  direkt machen.

Vorteil: entspricht oft der Datenstruktur

Nachteil: entspricht nicht dem Lowess Modell