Lösung:

```
replications <- 200
x<- c(6*runif(35), rnorm(65)-1)
median(x)
[1] -0.2982217
sd(replicate(replications,
   median(sample(x,length(x), replace=TRUE))))
[1] 0.2690578
interquartile100 <- function(x){ x<- sort(x); x[75]-x[25]}
interquartile100(x)
[1] 2.26707
sd(replicate(replications,
   interquartile100(sample(x,length(x), replace=TRUE))))
[1] 0.360023
```

Bootstrap-Konfidenzintervalle

- Auch ein Bootstrap-Konfidenzintervall lässt sich mittels der gewonnenen Schätzer definieren.

- Das $(1-\alpha)\%$ Bootstrap-KI für einen Parameter θ ist gegeben durch

$$\mathsf{KI}_{1-\alpha,B}(\theta) := \hat{\theta} \pm z_{1-\frac{\alpha}{2}} \cdot \hat{\mathsf{se}}_B(\hat{\theta}).$$

Bootstrap-Bias

- Genau wie bei jedem klassischen Schätzer kann man auch für einen Bootstrapschätzer $\hat{\theta}$ den Bias definieren:

$$\text{Bias}_{\hat{f}}(\hat{\theta}) = E_{\hat{f}} s(\mathbf{x}^*) - t(\hat{f}) = E(\hat{\theta}^*) - \hat{\theta}.$$

- Achtung: Der Bias wird hier mit Bezug auf den geschätzten Parameter $\hat{\theta}$ der gesamten Stichprobe berechnet.

- Aufgabe: Schreiben Sie ein Programm, das die Abhängigkeit des Bootstrap-Bias von der Zahl der Replikationen grafisch sichtbar macht!

- Erzeugen Sie die erste Stichprobe $\mathbf{x}$ aus einer Verteilung, bei der 10% der Daten aus einer $N(0,1)$ und 90% der Daten aus einer $N(6,9)$ gezogen werden.

- Als einfaches Beispiel soll der Bias der Mittelwertschätzung betrachtet werden.

- Betrachten Sie die in Zehnerschritten die Replikationszahlen von 10 bis 1000.

- Überlegen Sie zunächst, in welche Einzelschritte das Problem für die Programmierung zerlegt werden kann.

- Berechnen Sie zunächst die Werte für den Bias, überlegen Sie erst dann, wie ein geeigneter Plot aussehen könnte.

- Lösung hier!

Konvergenzeigenschaften des Bootstrap

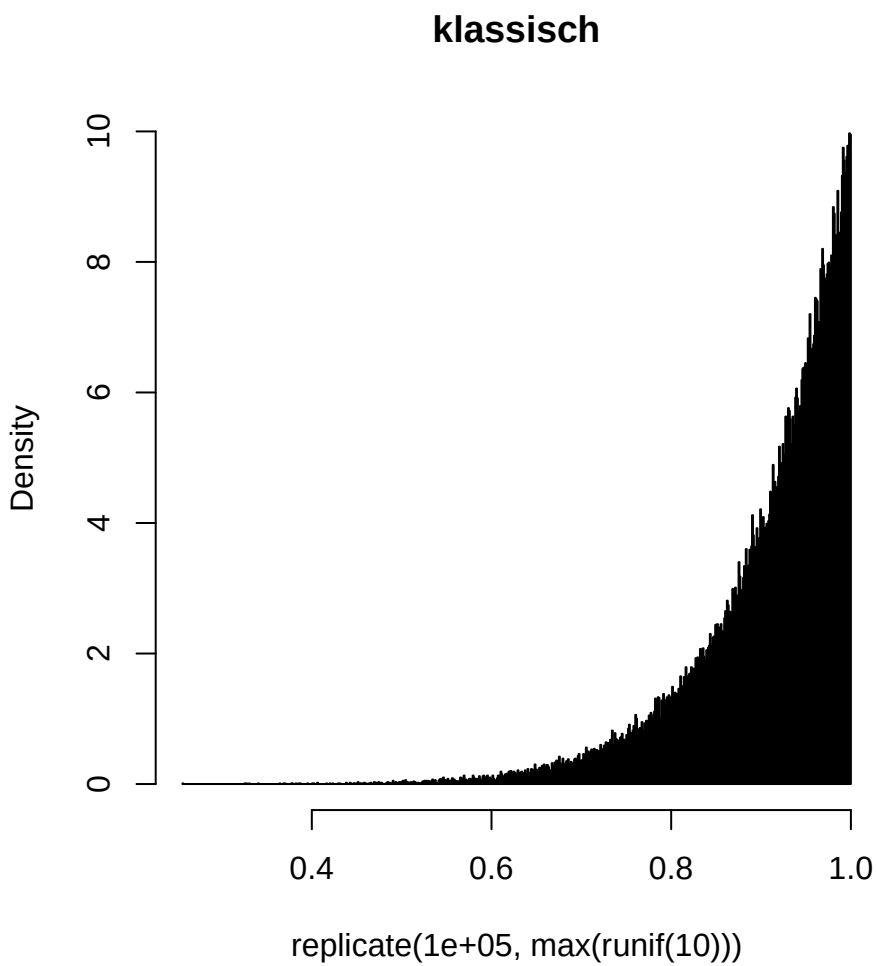
- Es gilt also, dass mit $N \to \infty$: $\text{Bias}_{\hat{f}}(\hat{\theta}) \to 0$!
- Weiterhin gilt, dass mit $n, N \to \infty$: $\hat{\text{se}}_B(\hat{\theta}) \to \text{sd}(\hat{\theta})$ aufgefasst als Zufallsvariable.

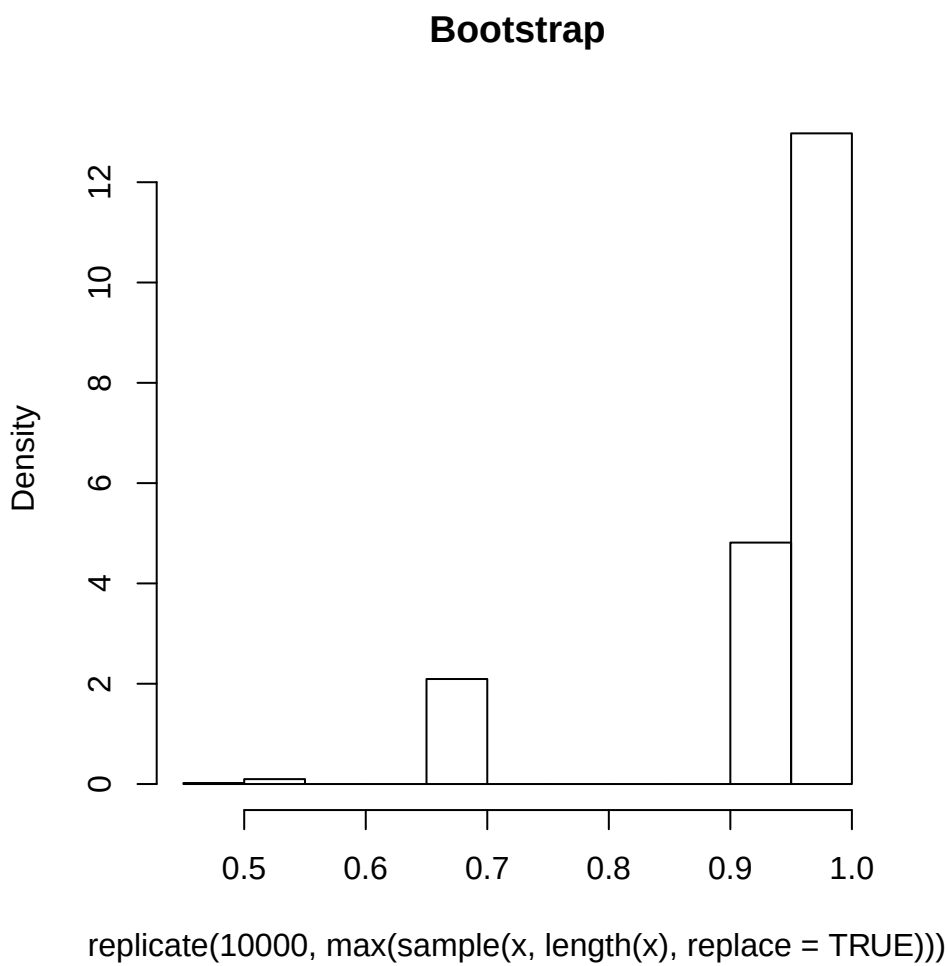
Grenzen des nicht-parametrischen Bootstrap

- Wenn nur eine kleine Stichprobe vorliegt, kann die Verteilung einer Teststatistik nur wenige diskrete Werte annehmen, auch wenn ihre wahre Verteilung stetig ist.

- Beispiel: Gegeben eine Stichprobe $\mathbf{x}$ vom Umfang 10 aus $U_{[0,1]}$. Die gesuchte Statistik ist das Maximum x_{max} für Stichproben vom Umfang 10 aus $U_{[0,1]}$.

- Im folgenden Bild werden einmal die "Dichteschätzer" für das x_{max} von 10000 Stichproben aus $U_{[0,1]}$ und zum anderen die geschätzte Dichte des entsprechende Bootstrapschätzers gegenübergestellt.


klassisch
Density
0
2
4
6
8
10
0.4
0.6
0.8
1.0
replicate(1e+05, max(runif(10)))


Bootstrap
Density
0
2
4
6
8
10
12
0.5
0.6
0.7
0.8
0.9
1.0
replicate(10000, max(sample(x, length(x), replace = TRUE)))

- Das Programm für diese Bilder:

```
par(mfrow=c(1,2))
hist(replicate(10000, max(runif(10))), prob=TRUE, nclass=100)
x <- runif(10)
hist(replicate(10000, max(sample(x,length(x),replace=TRUE))),
     prob=TRUE)
```

- Es ist klar zu erkennen, dass in dieser Situation der Bootstrap völlig versagt. Die kleine Stichprobe enthält einfach zu wenig Information über die zugrundeliegende Verteilung!

- Als Ausweg bleibt nur entweder eine "glatter" Schätzung von $\hat{f}$ durchzuführen, bzw. über f weiter Annahmen zu treffen. Eine typische Annahme ist die Zugehörigkeit von f zu einer bestimmten Verteilungsklasse. Beispielsweise $f = \varphi(\mu, \sigma^2)$, eine Normalverteilungsdichte.

Der parametrische Bootstrap

- Der parametrische Bootstrap liegt relativ nah an den klassischen Methoden und soll die Schwächen des klassischen Bootstrap beheben.

- Er kommt zum Einsatz, wenn die klassischen Methoden dabei versagen, die Genauigkeit der Schätzer zu bestimmen.

- Zum parametrischen Bootstrap gehört immer eine Annahme über die Verteilung f_Θ, aus der die ursprüngliche Stichprobe $\mathbf{x}$ gezogen wurde. Dabei ist $\Theta \in \mathcal{R}^m$, m die Dimension des Parameterraums von f.

- Beispiel: Gleichverteilung auf $[\vartheta_1, \vartheta_2]$ mit $\Theta = (\theta_1, \theta_2)$.

- Aus dieser Stichprobe wird zum einen die gesuchte Teststatistik $\hat{\theta}$ (Maximum, 0.75 Quantil ...) berechnet, zum anderen Schätzer $\hat{\Theta} = (\hat{\vartheta_1}, \ldots, \hat{\vartheta_m})$ für die Parameter der angenommenen Verteilung hergeleitet.

- Der eigentliche Bootstrap-Schritt arbeitet dann nicht mit einem Resampling aus $\mathbf{x}$, sondern mit einem Sampling neuer Stichproben $\mathbf{x}_i^*,\ 1 \leq i \leq N$ des Umfangs n aus $\hat{F}_{\hat{\Theta}}$.

- Die Berechnungen der Bootstrap-Schätzer geschehen dann mit denselben Formeln, wie beim nichtparametrischen Bootstrap.

Beispiel: Parametrischer Bootstrap

- Angenommen unsere Stichprobe vom Umfang $n = 10$ stammt aus einer Gleichverteilung auf $U_{[0,\vartheta]}$, ϑ unbekannt.

```
x<- runif(10)
```

- Gesucht ist die Verteilung des Maximums x_{max}.

- Der Momentenschätzer für den Parameterschätzer $\hat{\vartheta}$ ergibt sich aus

$$E_{f\{\vartheta}(X) = \frac{\vartheta}{2}.$$

- Also ist

$$\hat{\vartheta} = 2\bar{X}$$

der Momentenschätzer für die obere Grenze θ der angenommenen Verteilung.

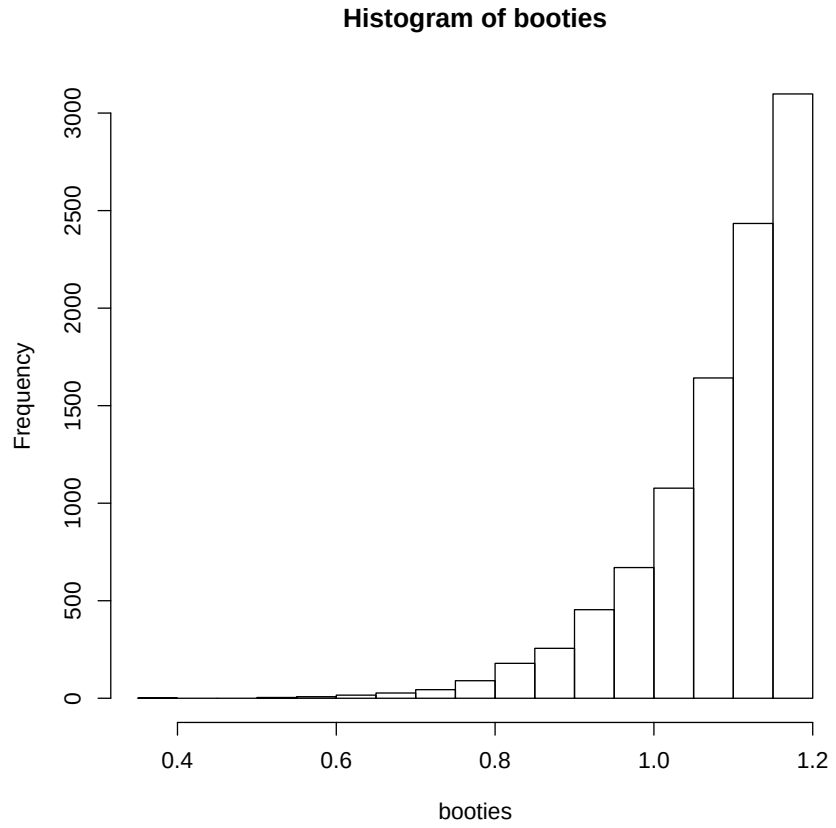
```
hattheta <- 2* mean(x)
```

- Aus dieser geschätzten Verteilung $\hat{f} = f_{\hat{\vartheta}}$ werden nun $N = 10000$ neue Stichproben gezogen und jeweils der ein Bootstrapschätzer für ϑ berechnet.

```
booties <- replicate(10000, max(hattheta*runif(10)))
```

- Das Histogramm kann nun eine erste Schätzung der Verteilung der gesuchten Statistik bieten.

```
hist(booties)
```



- Aus den Bootstrap Samples kann nun natürlich auch ein Standardfehler für die gesuchte Statistik abgeleitet werden.

```
sd(booties)
```

- Mittels Bootstrapping lassen sich auch empirisch p-Wert berechnen, Konfidenzintervalle etc.

Aufgabe zum parametrischen Bootstrap

- Ihnen liegt eine Stichprobe $\mathbf{x}$ vom Umfang 15 von Wartezeiten an einem Schalter vor (gerundet auf Sekunden).
  x=(260, 522, 619, 1433, 417, 121, 105, 438, 227, 402, 41, 6, 102, 225, 259)
  Sie vermuten, da es sich um Wartezeiten handelt, dass es sich um eine Stichprobe aus einer Exponentialverteilung handelt.

- Den Auftraggeber interessiert nun eine Konfidenzintervall für die mediane Wartezeit aus der unbekannten, zugrundeliegenden Exponentialverteilung.

- Schätzen Sie mit der Momentenmethode den Parameter λ der Exponentialverteilung.

- Schätzen Sie mittels parametrischem Bootstrap den Standardfehler des Medianschätzers für $N = 250$ Replikationen.

- Geben Sie ein 99%-Bootstrap-Konfidenzintervall für den Median an.

- Lösung hier