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 ein 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:

- Aus Statistik II ist der Momentenschätzer für die Exponentialverteilung bekannt:

$$\hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{\frac{1}{n}\sum_1^n x_i}$$

Damit ergibt sich der Schätzer $\hat{\lambda} =$ `1/mean(x)` ≈ 0.00289 aus den 15 Beobachungen.

- 250 Replikationen des parametrischen Bootstrapschätzer für den Median ergeben sich also mit

```
bootstraps <- replicate(250, median(rexp(15, rate=1/mean(x))))
> mean(bootstraps)
[1] 258.5975
> sd(bootstraps)
[1] 92.2493
```

- Das 99% KI_B für den Median ergibt sich folglich zu

$$KI_{99\%,B} = [\bar{x} \pm z_{0.995} \frac{\hat{sd}}{\sqrt{n}}] = [258.60 \pm 2.58 \cdot 92.25] = [20.595; 496.60].$$

```
mean(bootstraps) +(bzw -) qnorm(0.995)*sd(bootstraps)/sqrt(250)
```

Das Jackknife-Prinzip

- Nach dem Herausziehen an den Stiefelschlaufen nun das Schweizer Taschenmesser.

- Wie gesehen, hat der Bootstrap das prinzipielle Problem, dass die Resampling Stichproben nicht aus der Verteilung zur Dichte f stammen, sondern aus der geschätzten Verteilung zur empirischen Dichte $\hat{f}$, welche aus der einzigen vorhandenen Stichprobe geschätzt wurde.

- Gibt es eine Möglichkeit, Stichproben aus der “wahren” Verteilung f zu erzeugen?

- In der Tat liegen mit der Stichprobe $\mathbf{x} = (x_1, \ldots, x_n)$ bereits eine ganze Reihe derartiger Stichproben aus f vor.

- Anstelle von Stichproben mit Zurücklegen vom Umfang n, wie beim Bootstrap, betrachtet man einfach Stichproben vom Umfang $m < n$ ohne Zurücklegen aus den n vorliegenden Beobachtungen.

- Diese Stichproben sind Stichproben aus f!

- Die Idee geht ursprünglich zurück auf Quenouille (ca. 1955), ist also älter als der Bootstrap!

- 1958 prägte dann Tuckey den Begriff *Jackknife* für das zugrundeliegende Prinzip der Betrachtung von Teilstichproben zur Berechnung von Tests und Konfidenzintervallen.

- Der große Vorteil des Jackknife mit $m = n - 1$ ist der erheblich geringere Rechenaufwand.

- Es wird in diesem Fall tatsächlich überhaupt keine Ziehung von Stichproben benötigt, wenn man alle n Teilstichproben von Umfang $n-1$ betrachtet.
- Statt eines echten Stichprobenziehens, wird reihum jede Beobachtung einmal aus der Ursprungsstichprobe $\mathbf{x}$ entfernt.
- Man spricht deshalb in Falle des Jackknife eher von Subsampling, denn von Resampling.

Die Jackknife Stichproben

- **Definition:** Die Jackknife (Teil-)Stichproben $\mathbf{x}_{(i)}$ erhält man, in dem man für jedes $1 \leq i \leq n$ aus der Stichprobe $\mathbf{x} = (x_1, \ldots, x_n)$ die i. Beobachtung x_i entfernt:

$$\mathbf{x}_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

- Jede Jackknife Stichprobe umfasst $n - 1$ Beobachtungen.
- Offensichtlich gibt es n verschiedene Jackknife Stichproben.
- Man erhält also die Teilstichproben, ohne wirkliches Stichprobenziehen, sondern durch systematisches Durchlaufen aller möglichen Stichproben!

Aufgabe Jackknife Stichproben

- Schreiben Sie eine R Funktion, die zu einem Stichprobenvektor alle zugehörigen Jackknife Stichproben ausgibt!
- (5 Minuten)
- Lösung hier nach der Vorlesung.

Jackknife Replikationen

- **Definition:** Die i. Jackknife Replikation $\hat{\theta}_{(i)}$ einer Teststatistik $\hat{\theta} = s(\mathbf{x})$ ist definiert als:

$$\hat{\theta}_{(i)} := s(\mathbf{x}_{(i)}).$$

- **Beispiel:** Die Jackknife Replikation des Mittelwertes:

$$s(\mathbf{x}_{(i)}) = \frac{1}{n-1} \sum_{j \neq i} x_j = \bar{x}_{(i)}.$$

- Das Ziel des Jackknifing ist dasselbe wie beim Bootstrap: Aus nur einer Stichprobe soll zu einer beliebigen Teststatistik auch eine Schätzung ihrer Varianz (bzw. Standardfehler) und damit eines Konfidenzintervalls angegeben werden können!

Der Standardfehler mittels Jackknifing

- Wenn man die Bootstrap-Idee kennt, ist das Vorgehen beim Jackknife wenig überraschend. Die nötigen Schritte sind wie folgt.

  1. Erzeuge die n Jackknife Stichproben $X_{(1)}, \ldots, x_{(n)}$ aus $\mathbf{x}$.
  2. Berechnen die n Jackknife Replikationen der gesuchten Teststatistik $\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)})$.
  3. Der Jackknife Schätzer für den Standardfehler von $\hat{\theta}$ ist dann definiert als

$$\hat{\mathsf{se}}_J(\hat{\theta}) = \left[\frac{n-1}{n} \sum_1^n (\bar{\hat{\theta}}_{(\cdot)} - \hat{\theta}_{(i)})^2\right]^{\frac{1}{2}}.$$

- Der Inflationsfaktor $\frac{n-1}{n}$ bedarf einer gesonderten Betrachtung.

Der Jackknife Inflationsfaktor

- Es ist zunächst zu sehen, dass der Faktor $\frac{n-1}{n}$ viel größer ist, als $\frac{1}{N}$, dem Faktor der beim Bootstrapping auftritt.

- Intuitiv ist klar, dass die Varianz der Jackknife-Statistiken kleiner ist, als die der Bootstrap-Statistiken, da die Teilstichproben sich "weniger" von der Urstichprobe unterscheiden, als die Resampling-Stichproben im Bootstrap-Fall.

- Tatsächlich kann man zeigen, dass für den Spezialfall $\hat{\theta} = \bar{x}$ gilt:

$$\hat{\text{se}}_J(\hat{\theta}) = \frac{s}{\sqrt{n}},$$

wobei s der e-treue Schätzer für die Varianz ist. Für andere Statistiken ist der Faktor einfach übernommen worden.

Aufgabe: Standardfehler Jackknife

- Schreiben Sie eine R-Funktion, die den Jackknife Standardfehler der zufälligen Abstände zweier Punkte im n-dimensionalen Einheitswürfel berechnet. Der Stichprobenumfang betrage wie im Bootstrap-Fall 100 bzw. 1000..
- Simulieren Sie für die Dimensionen $1 \leq n \leq 10$ die Schätzer für die Interquartilsabstände und ihre Standardfehler mittels Jackknife.
- (30 min)
- Lösung hier.

Die Beziehung zwischen Jackknife und Bootstrap

- Wenn n nicht zu groß ist, ist es mit weniger Rechenaufwand verbunden, die n Jackknife Replikationen zu berechnen, als eine entsprechend hinreichend große Bootstrapstichprobe zu erzeugen.

- Im Jackknife wird die vollständige Menge der Teilstichproben betrachtet! Ansonsten stimmen die hergeleiteten Eigenschaften der Schätzer natürlich nicht!

- Durch den Übergang zum Jackknife verliert man keine Information gegenüber dem Bootstrap.

- Tatsächlich ist es so, dass für lineare Statistiken die Schätzer aus dem Jackknife und aus einem vollständigen Bootstrap übereinstimmen.

- Der Jackknife hat immer dann Probleme, wenn das nicht-glatte Statistiken betrachtet werden. Ein Beispiel für eine solche nicht-glatte Statistik ist der Median. Der Median als Statistik ist nicht bezüglich der Stichprobe differenzierbar!

- Durch das Auslassen von $d > 1$ Beobachtungen in jedem Schritt, dem sogenannten *delete-d Jackknife*, lässt sich diese Unzulänglichkeit des Verfahrens etwas abmildern,

delete-d Jackknife

- **Definition:** Die *delete-d* Jackknife Stichproben ergeben sich aus einer Urstichprobe durch entfernen von jeweils d Beobachtungen.
- Jede Teilstichprobe hat den Umfang $n - d$.
- Es gibt folglich $\binom{n}{d}$ delete-d Jackknife Stichproben!
- Da diese Zahl exponentiell steigt und der Jackknife auf der Erhebung aller Teilstichproben beruht, ist ein großes d nicht sinnvoll!
- Für den Standardfehlerschätzer des delete-d Jackknife gilt $(n = r \cdot d)$:

$$\hat{\text{se}}_{J,d}(\hat{\theta}) = \left\{ \frac{r}{\binom{n}{d}} \sum_{\text{alle Teilstichproben}} (\bar{\hat{\theta}} - \hat{\theta}_i) \right\}.$$

- Insgesamt gilt für den Jackknife, dass er in manchen Fällen weniger Rechenzeit erfordert als der Bootstrap.

- Die (Teil-)Stichproben des Jackknife sind aus f, nicht aus $\hat{f}$, wie beim Bootstrap.

- Der Vorteil des Bootstrap bei großen Stichproben ist, dass man die Zahl der Stichproben im Griff hat.

Aufgabe *delete-d* Jackknife

- Programmieren Sie eine Funktion, die für ein wählbares d alle *delete-d* Jackknife Stichproben erzeugt.
- Vermutlich schwierig! Leiten Sie eine rekursive Prozedur für die Erzeugung der Stichproben her!