Schätzung des Lageparameters einer symmetrischen Verteilung

Andreas Handl

Inhaltsverzeichnis

1 Maßzahlen zur Beschreibung der Lage eines Datensatzes

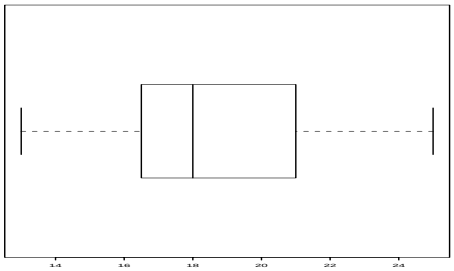
Bei einer symmetrischen Verteilung ist das Symmetriezentrum θ der natürliche Lageparameter. Wir wollen uns im Folgenden mit der Schätzung von θ beschäftigen. Dabei gehen wir von einer Zufallsstichprobe aus einer Grundgesamtheit aus, in der das interessierende Merkmal eine stetige und bezüglich θ symmetrische Verteilung besitzt. Die Beobachtungen $x_1, \ldots, x_n$ sind also Realisationen der unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$.

Beispiel 1
Im Rahmen eines BI-Projektes sollten die Teilnehmer den Lineal-Reaktions-Test durchführen. Die Ergebnisse von Männern, die das Lineal mit der Nicht-Schreibhand fingen, sind:

```
16 19 13 17 19 23 17 25
```

Abbildung 1 zeigt den Boxplot. Dieser deutet auf eine symmetrische Verteilung hin.

Abbildung 1: Boxplot der Reaktionszeit



1.1 Mittelwert und Median

In der Grundausbildung lernt man den Mittelwert $\overline{x}$ und den Median $x_{0.5}$ als Lageschätzer kennen. Für eine Stichprobe $x_1, \ldots, x_n$ ist der Mittelwert folgendermaßen definiert:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

Beim Mittelwert verteilen wir die Summe aller Beobachtungen gleichmäßig auf alle Merkmalsträger.
Beim Median geht man von der geordneten Stichprobe $x_{(1)}, \ldots, x_{(n)}$ aus. Es gilt also

$$x_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade ist} \\ \frac{x_{(n/2)} + x_{(1+n/2)}}{2} & \text{falls } n \text{ gerade ist} \end{cases} \tag{2}$$

Der Median $x_{0.5}$ teilt den geordneten Datensatz $x_{(1)}, \ldots, x_{(n)}$ in zwei gleich große Teile.

Beispiel 1 (fortgesetzt)
Es gilt $\bar{x} = 18.625$ und $x_{0.5} = 18$.

In der Regel werden die Werte des Mittelwertes und Medians bei einer Stichprobe unterschiedlich sein. Will man einen Wert für die Lage der Verteilung angeben, so muss man sich zwischen dem Median und dem Mittelwert entscheiden. Welche dieser beiden Maßzahlen ist besser geeignet, die Lage der konkreten Stichprobe zu beschreiben? Der Mittelwert ist nicht robust. Ein Ausreißer hat einen starken Einfluss auf den Mittelwert. Der Median hingegen ist robust. Liegen also Ausreißer vor, so sollte man den Median wählen. Man kann eine einfache Entscheidungsregel auf Basis des Boxplots aufstellen. Enthält der Boxplot keinen Ausreißer, so sollte man sich für den Mittelwert entscheiden. Ist aber mindestens ein Ausreißer im Boxplot zu erkennen, so sollte man die Lage des Datensatzes durch den Median beschreiben.

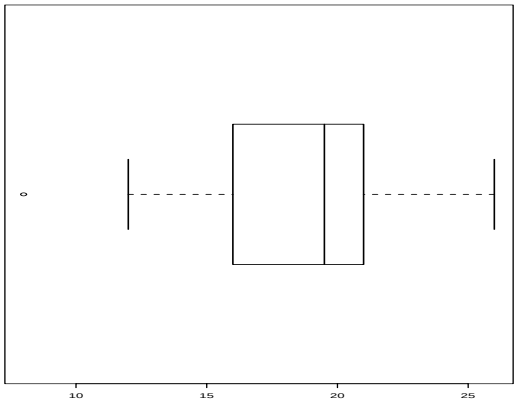
Beispiel 1 (fortgesetzt)
Da kein Ausreißer zu erkennen ist, beschreiben wir die Lage durch den Mittelwert.

Beispiel 2
In dem BI-Projekt wurde der LRT auch bei jungen Männern durchgeführt. Hier sind die Ergebnisse

```
16 19 12 19 21 26 20 23  8 20
```

Es gilt $\bar{x} = 18.4$ und $x_{0.5} = 19.5$. Abbildung 2 zeigt den Boxplot. Dieser deutet auf eine symmetrische Verteilung hin. Es liegt aber ein Ausreißer vor. Wir wählen also den Median.

Abbildung 2: Boxplot der Reaktionszeit



1.2 Getrimmte Mittelwerte und Mittelwerte der getrimmten Beobachtungen

Ist $x_1, \ldots, x_n$ die Stichprobe und $x_{(1)}, \ldots, x_{(n)}$ die geordnete Stichprobe, so ist der Mittelwert

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i=1}^{n} x_{(i)}$$

Eine einzige Beobachtung hat einen starken Einfluss auf den Wert des Mittelwertes.
Ist der Stichprobenumfang n ungerade, so ist der Median gleich

$$x_{0.5} = x_{((n+1)/2)}$$

Ist der Stichprobenumfang n gerade, so ist der Median gleich

$$x_{0.5} = \frac{1}{2} \left(x_{(n/2)} + x_{(1+n/2)} \right)$$

Während beim Mittelwert alle Beobachtungen mit dem gleichen Gewicht berücksichtigt werden, werden beim Median bei einem ungeraden Stichprobenumfang nur die Beobachtung in der Mitte der geordneten Stichprobe und bei einem geraden Stichprobenumfang nur die Beobachtungen in der Mitte der geordneten Stichprobe berücksichtigt. Man kann es auch so sehen, dass bei einem geraden Stichprobenumfang die $n/2 - 1$ kleinsten und $n/2 - 1$ größten Beobachtungen aus der Stichprobe entfernt werden, und der Mittelwert der Beobachtungen bestimmt wird, die nicht aus der Stichprobe entfernt

wurden. Man spricht auch davon, dass Beobachtungen getrimmt werden. Hierdurch ist der Median unempfindlich gegenüber Ausreißern. Der Median ist ein robuster Schätzer.

Man kann natürlich weniger Beobachtungen aus der Stichprobe entfernen als beim Median. Man spricht dann von einem getrimmten Mittelwert. Beim Trimmen kann man die Anzahl und den Anteil der Beobachtungen vorgeben, die von den Rändern der geordneten Stichprobe $x_{(1)}, \ldots, x_{(n)}$ entfernt werden sollen.

Man spricht von einem k-fach symmetrisch getrimmten Mittelwert $\bar{x}_k$, wenn die k kleinsten und die k größten Bobachtungen aus der Stichprobe entfernt werden und der Mittelwert der Beobachtungen bestimmt wird, die nicht aus der Stichprobe eliminiert wurden. Es gilt

$$\bar{x}_k = \frac{1}{n - 2\,k} \sum_{i=k+1}^{n-k} x_{(i)} \tag{3}$$

Beispiel 3
Wir betrachten die Daten aus Beispiel 2 auf Seite 3. Hier ist die geordnete Stichprobe:

```
8 12 16 19 19 20 20 21 23 26
```

Wir setzen $k = 1$. Es gilt

$$\bar{x}_1 = \frac{1}{8}\,(12 + 16 + 19 + 19 + 20 + 20 + 21 + 23) = 8.75$$

Analog erhalten wir

$$\bar{x}_2 = 19.17 \qquad \bar{x}_3 = 19.5 \qquad \bar{x}_4 = 19.5$$

Gibt man den Anteil α vor, der von jedem Rand der geordneten Stichprobe entfernt werden soll, so spricht man von einem α-getrimmten Mittelwert. In der Regel wird $n\,\alpha$ keine natürliche Zahl sein. Entfernt man jeweils $\lfloor n\,\alpha \rfloor$ Beobachtungen, so erhält man folgenden Schätzer:

$$\bar{x}_\alpha = \frac{1}{n - 2\,g} \sum_{i=g+1}^{n-g} x_{(i)} \tag{4}$$

mit $g = \lfloor n\,\alpha \rfloor$.

Beispiel 3 (fortgesetzt)
Wir wählen $\alpha = 0.05$ und erhalten $g = \lfloor 10 \cdot 0.05 \rfloor = 0$. Als Schätzer erhalten wir $\bar{x}_{0.05} = 18.4$. In Tabelle 1 sind die Werte von $\bar{x}_\alpha$ für ausgewählte Werte von α zu finden.

Tabelle 1: Werte von $\bar{x}_\alpha$für ausgewählte Werte von α

α	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
$\bar{x}_\alpha$	18.75	18.75	19.17	19.17	19.5	19.5	19.5	19.5

Bei kleinen Werten von n schätzt man $\bar{x}_\alpha$ für unterschiedliche Werte von α durch einen Wert. Die Werte von $\bar{x}_\alpha$ sind also nicht stetig in α. Die Abbildungen 10-5 und 10-6 in Hoaglin et al. (1983) verdeutlichen dies. Hoaglin et al. (1983) schlagen vor, auch Anteile von geordneten Beobachtungen zu verwenden. Dies liefert folgenden Schätzer:

$$T(\alpha) = \frac{1}{n(1-2\,\alpha)} \left\{ (1-r)\, \left[x_{(g+1)} + x_{(n-g)} \right] + \sum_{i=g+2}^{n-g-1} x_{(i)} \right\} \tag{5}$$

mit $g = \lfloor n\,\alpha \rfloor$ und $r = n\,\alpha - g$.

Beispiel 3 (fortgesetzt)
Wir wählen $\alpha = 0.05$ und erhalten $g = \lfloor 10 \cdot 0.05 \rfloor = 0$ und $r = 10 \cdot 0.05 - 0 = 0.5$. Als Schätzer erhalten wir

$$\begin{aligned} T(0.05) &= \frac{1}{10(1-2\cdot 0.05)} \left[(1-0.5)(x_{(1)} + x_{(10)}) + \sum_{i=2}^{9} x_{(i)} \right] \\ &= \frac{1}{9} \left[0.5(8+26) + 12 + 16 + 19 + 19 + 20 + 20 + 21 + 23 \right] \\ &= 18.56 \end{aligned}$$

In Tabelle 2 sind die Werte von $T(\alpha)$ für ausgewählte Werte von α zu finden.

Tabelle 2: Werte von $T(\alpha)$für ausgewählte Werte von α

α	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
$T(\alpha)$	18.75	18.93	19.17	19.3	19.5	19.5	19.5	19.5

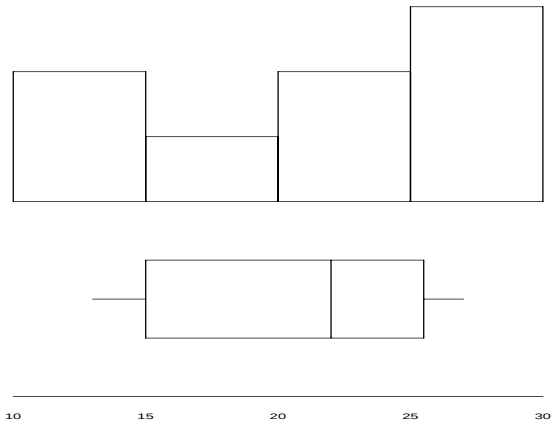
Die Lage einiger Datensätze sollten nicht durch getrimmte Mittelwerte beschrieben werden. Schauen wir uns hierzu ein Beispiel an.

Beispiel 4
Im Rahmen eines BI-Projektes sollten die Teilnehmer den Lineal-Reaktionstest durchführen. Die Ergebnisse von Männern, die während des Versuches abgelenkt wurden, sind:

```
17 26 13 20 27 13 24 25
```

Abbildung 3 zeigt den Boxplot. Dieser deutet auf eine symmetrische Verteilung mit wenig Wahrscheinlichkeitsmasse an den Rändern wie die Gleichverteilung oder auf eine U-förmige Verteilung hin. Das Histogramm bestätigt diese Vermutung.

Abbildung 3: Histogramm und Boxplot der Reaktionszeit



Die Daten im Beispiel 4 deuten auf eine Gleichverteilung auf (a, b) hin. Die Dichtefunktion ist

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{für } a < x < b \\ 0 & \text{sonst} \end{cases}$$

Diese ist symmetrisch bezüglich

$$E(X) = \frac{a+b}{2}$$

Die M-L-Schätzer von $E(X)$ ist

$$\frac{X_{(1)} + X_{(n)}}{2} \tag{6}$$

Der Beweis ist bei Mood et al. (1974) auf den Seiten 282-283 zu finden.
Man nennt den Ausdruck in Gleichung (6) auch den Midrange oder die Spannweitenmitte. Er ist ein Beispiel für einen Mittelwert der getrimmten Werte, der folgendermaßen definiert ist:

$$T(\alpha)^c = \begin{cases} \frac{x_{(1)} + x_{(1)}}{2} & \text{für } n\,\alpha < 1 \\ \frac{1}{2\,n\,\alpha}\left\{ r\,\left[x_{(g+1)} + x_{(n-g)}\right] + \sum_{i=1}^{g}\left(x_{(i)} + x_{(n+1-i)}\right)\right\} & \text{für } n\,\alpha \geq 1 \end{cases} \tag{7}$$

mit $g = \lfloor n\,\alpha \rfloor$ und $r = n\,\alpha - g$.
$T(0.25)^c$ heißt auch Outmean. Bestimmt man $T(0.25)$ wie in Gleichung 4, so gilt

$$T(0.25)^c + T(0.25) = 2\,\bar{X}$$

Beispiel 4 (fortgesetzt)
Wir wählen $\alpha = 0.05$ und erhalten $g = \lfloor 10 \cdot 0.05 \rfloor = 0$ und $r = 10 \cdot 0.05 - 0 = 0.5$. Als Schätzer erhalten wir

$$T(0.05)^c = \frac{x_{(1)} + x_{(10)}}{2} = 20$$

In Tabelle 3 sind die Werte von $T(\alpha)$ für ausgewählte Werte von α zu finden.

Tabelle 3: Werte von $T(\alpha)$ für ausgewählte Werte von α

α	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
$T(\alpha)$	20.0	19.92	19.81	19.75	19.96	20.11	20.28	20.47

2 Die Auswahl einer geeigneten Schätzfunktion zur Beschreibung der Lage einer symmetrischen Verteilung

Wir wollen im Folgenden Verfahren angeben, mit denen man sich auf Basis der Daten zwischen dem Mittelwert und dem Median entscheiden kann. Die auf Seite 3 beschriebene Entscheidungsregel, die auf dem Boxplot beruht, berücksichtigt nur Ausreißer. Die Effizienz der Schätzfunktion wird nicht in Betracht gezogen. Beim Schätzen haben wir das Problem, dass in der Regel nur ein Schätzwert vorliegt, von dem wir nicht wissen, ob er in der Nähe des wahren Wertes des Parameters liegt. Ist die Schätzfunktion aber erwartungstreu und besitzt sie dazu noch eine kleine Varianz, so können wir uns ziemlich sicher sein, dass der Wert der Schätzfunktion in der Nähe des wahren Wertes des Parameters liegt. Schauen wir uns also noch einmal Gütekriterien von Schätzfunktionen an.

2.1 Effiziente Schätzfunktionen

Definition 2.1
Eine Schätzfunktion T heißt **erwartungstreu** für den Parameter θ, wenn für alle Werte von θ gilt:

$$E(T) = \theta$$

Man nennt eine erwartungstreue Schätzfunktion auch **unverzerrt**. Im Englischen spricht man von einem **unbiased estimator**.

Beispiel 5
Sind $X_1, \ldots, X_n$ unabhängige, identisch mit $E(X_i) = \mu$ verteilte Zufallsvariablen, dann ist $\bar{X}$ eine erwartungstreue Schätzfunktion für μ.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\,E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\,n\,\mu = \mu$$

Beispiel 6
Sind $X_1, \ldots, X_n$ unabhängige, identisch mit stetiger und bezüglich θ symmetrischer Verteilungsfunktion verteilte Zufallsvariablen, dann ist $X_{0.5}$ eine erwartungstreue Schätzfunktion für θ.
Wir zeigen, dass für ungerades n die Dichtefunktion des Medians symmetrisch bezüglich θ ist, wenn die Dichtefunktion von X symmetrisch θ ist.
Wir unterstellen also

$$f_X(\theta - x) = f_X(\theta + x) \tag{8}$$

und

$$F_X(\theta - x) = 1 - F_X(\theta + x) \tag{9}$$

Die Dichtefunktion der k-ten Orderstatistik $X_{(k)}$ ist gegeben durch

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,(n-k)!}\, F_X(x)^{k-1}\, [1 - F_X(x)]^{n-k}\, f_X(x) \tag{10}$$

Ein sehr anschaulicher Beweis ist bei David (1981) zu finden.
Ist der Stichprobenumfang n ungerade, so ist der Median gleich $X_{((n+1)/2)}$. Somit gilt

$$\begin{aligned}
f_{X_{(\frac{n+1}{2})}}(x) &= \frac{n!}{(\frac{n+1}{2}-1)!\,(n-\frac{n+1}{2})!}\, F_X(x)^{\frac{n+1}{2}-1}\, [1 - F_X(x)]^{n-\frac{n+1}{2}}\, f_X(x) \\
&= \frac{n!}{(\frac{n-1}{2})!\,(\frac{n-1}{2})!}\, F_X(x)^{\frac{n-1}{2}}\, [1 - F_X(x)]^{\frac{n-1}{2}}\, f_X(x) \\
&= \frac{n!}{\left[(\frac{n-1}{2})!\right]^2}\, F_X(x)^{\frac{n-1}{2}}\, [1 - F_X(x)]^{\frac{n-1}{2}}\, f_X(x)
\end{aligned}$$

Nun gilt

$$\begin{aligned}
f_{X_{(\frac{n+1}{2})}}(\theta - x) &= \frac{n!}{\left[(\frac{n-1}{2})!\right]^2}\, F_X(\theta - x)^{\frac{n-1}{2}}\, [1 - F_X(\theta - x)]^{\frac{n-1}{2}}\, f_X(\theta - x) \\
&\overset{(8)(9)}{=} \frac{n!}{\left[(\frac{n-1}{2})!\right]^2}\, [1 - F_X(\theta + x)]^{\frac{n-1}{2}}\, F_X(\theta + x)^{\frac{n-1}{2}}\, f_X(\theta + x) \\
&= \frac{n!}{\left[(\frac{n-1}{2})!\right]^2}\, F_X(\theta + x)^{\frac{n-1}{2}}\, [1 - F_X(\theta + x)]^{\frac{n-1}{2}}\, f_X(\theta + x) \\
&= f_{X_{(\frac{n+1}{2})}}(\theta + x)
\end{aligned}$$

Somit ist die Dichtefunktion von $X_{(\frac{n+1}{2})}$ symmetrisch bezüglich θ. Existiert der Erwartungswert $X_{(\frac{n+1}{2})}$, so ist er gleich θ.

Sind zwei Schätzfunktionen erwartungstreu, so wählt man die mit der kleineren Varianz. Je kleiner nämlich die Varianz ist, um so sicherer können wir sein, dass der realisierte Wert der Schätzfunktion in der Nähe des wahren Wertes des Parameters liegt.

Definition 2.2
Seien T_1 und T_2 zwei erwartungstreue Schätzfunktionen des Parameters θ. Die Schätzfunktion T_1 heißt effizienter als die Schätzfunktion T_2, wenn gilt

$$Var(T_1) < Var(T_2)$$

Beispiel 6 (fortgesetzt)
Es gilt

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\, n\, \sigma^2 = \frac{\sigma^2}{n}$$

Die Varianz von vielen Schätzfunktionen kann nicht explizit angegeben werden. In diesem Fall gibt es zwei Möglichkeiten, die Varianz approximativ zu bestimmen:

1. Man kann die asymptotische Varianz bestimmen.
2. Man führt eine Simulation durch.

2.2 Asymptotik

Schauen wir uns zunächst ein Beispiel für Asymptotik an. Wir suchen eine Approximation der Varianz $Var(X_{0.5})$ des Medians bei einer Zufallsstichprobe vom Umfang n aus einer Grundgesamtheit, deren Verteilungsfunktion $F_X(x)$ stetig ist. Dabei sei der Stichprobenumfang n ungerade. Der Median ist somit $X_{((n+1)/2)}$.
Der Median ist eine spezielle Orderstatistik $X_{(k)}$ mit $k = \frac{n+1}{2}$. Die Dichtefunktion von $X_{(k)}$ ist in Gleichung 10 auf Seite 10 zu finden.
Wir bestimmen zunächst die Varianz des Medians für eine Zufallsstichprobe $U_1, \ldots, U_n$ aus der Gleichverteilung auf $(0, 1)$.
Die Zufallsvariable U besitzt eine Gleichverteilung auf $(0, 1)$, wenn die Verteilungsfunktion lautet:

$$F_U(u) = \begin{cases} 0 & \text{für } u \leq 0 \\ u & \text{für } 0 < u < 1 \\ 1 & \text{für } u \geq 1 \end{cases} \tag{11}$$

Die Dichtefunktion der Gleichverteilung auf $(0, 1)$ ist:

$$f_U(u) = \begin{cases} 1 & \text{für } 0 < u < 1 \\ 0 & \text{sonst} \end{cases} \tag{12}$$

Setzen wir diese Gleichungen in die Gleichung 10 auf Seite 10 ein, so erhalten wir für $x \in (0,1)$

$$\begin{aligned} f_{U_{(k)}}(x) &= \frac{n!}{(k-1)!\,(n-k)!}\, F_U(x)^{k-1}\, [1-F_U(x)]^{n-k}\, f_U(x) \\ &= \frac{n!}{(k-1)!\,(n-k)!}\, x^{k-1}\, (1-x)^{n-k} \end{aligned} \tag{13}$$

Ansonsten ist die Dichtefunktion von $U_{(k)}$ gleich 0.
Wir können die Fakultäten über die Gammafunktion

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx$$

ausdrücken. Für $n = 1, 2, \ldots$ gilt

$$\Gamma(n+1) = n!$$

(siehe dazu Rudin (1976), S. 192).
Für $f_{U_{(k)}}(x)$ gilt somit

$$f_{U_{(k)}}(x) = \frac{\Gamma(n+1)}{\Gamma(k)\,\Gamma(n+1-k)}\, x^{k-1}\, (1-x)^{n-k}$$

Die Betafunktion $B(a,b)$ ist definiert durch

$$B(a,b) = \int_0^1 x^{a-1}\,(1-x)^{b-1}\, dx$$

Es gilt

$$B(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)} \tag{14}$$

(siehe dazu Rudin (1976), S. 193).
Für $f_{U_{(k)}}(x)$ gilt somit

$$f_{U_{(k)}}(x) = \frac{1}{B(k, n+1-k)}\, x^{k-1}\, (1-x)^{n-k}$$

Dies ist die Dichtefunktion einer Betaverteilung mit den Parametern $a = k$ und $b = n + 1 - k$.

$$f_X(x) = \begin{cases} \dfrac{1}{B(a,b)}\, x^{a-1}\,(1-x)^{b-1} & \text{für } 0 < x < 1 \\ 0 & \text{sonst} \end{cases} \tag{15}$$

(siehe dazu Mood et al. (1974), S. 116 und S. 534-535)
Für eine mit den Parametern a und b betaverteilte Zufallsvariable X gilt

$$E(X) = \frac{a}{a+b}$$

und

$$Var(X) = \frac{ab}{(a+b+1)(a+b)^2}$$

(siehe dazu Mood et al. (1974), S. 117)
Also gilt

$$E(U_{(k)}) = \frac{k}{k+n+1-k} = \frac{k}{n+1} \tag{16}$$

und

$$\begin{aligned} Var(U_{(k)}) &= \frac{k\,(n+1-k)}{(k+n+1-k+1)(k+n+1-k)^2} = \frac{k\,(n+1-k)}{(n+2)(n+1)^2} \\ &= \left(\frac{1}{n+2}\right)\left(\frac{k}{n+1}\right)\left(1-\frac{k}{n+1}\right) \end{aligned} \tag{17}$$

Ist n ungerade, so ist der Median gleich $U_{((n+1)/2)}$. Also ist $k = (n+1)/2$. Setzen wir $k = (n+1)/2$ in Gleichung (17) ein, so erhalten wir

$$\begin{aligned} Var(U_{((n+1)/2)}) &= \left(\frac{1}{n+2}\right)\left(\frac{(n+1)/2}{n+1}\right)\left(1-\frac{(n+1)/2}{n+1}\right) \\ &= \left(\frac{1}{n+2}\right)\cdot 0.5 \cdot 0.5 = \frac{1}{4(n+2)} \end{aligned}$$

Für großes n gilt also

$$Var(U_{((n+1)/2)}) = \frac{1}{4\,n}$$

Schauen wir uns nun die Varianz des Medians für eine Zufallsvariable X mit stetiger Verteilungsfunktion $F_X(x)$ an. Hierzu benötigen wir den folgenden Satz .

Satz 2.1
Sei $F(x)$ eine stetige Verteilungsfunktion. Ist U gleichverteilt auf $(0,1)$, so besitzt $X = F^{-1}(U)$ die Verteilungsfunktion $F(x)$.

Beweis
Es gilt

$$F_X(x) = P(X \le x) = P\left(F^{-1}(U) \le x\right) = P(U \le F(x)) = F_U(F(x)) = F(X)$$

Somit gilt für die k-te Orderstatistik $X_{(k)}$ einer Zufallsstichprobe vom Umfang n aus einer Grundgesamtheit mit Verteilungsfunktion $F_X(x)$:

$$X_{(k)} = F^{-1}\left(U_{(k)}\right) \tag{18}$$

Dabei ist $U_{(k)}$ die k-te Orderstatistik einer Zufallsstichprobe vom Umfang n aus einer Grundgesamtheit mit Gleichverteilung auf $(0,1)$.
Wir approximieren die Varianz von $X_{(k)}$, indem wir $F^{-1}\left(U_{(k)}\right)$ um $E\left(U_{(k)}\right)$ linearisieren:

$$\begin{aligned} F^{-1}\left(U_{(k)}\right) &\approx F^{-1}\left(E\left(U_{(k)}\right)\right) + \left(U_{(k)} - E\left(U_{(k)}\right)\right)\left(F^{-1}\right)'\left(E\left(U_{(k)}\right)\right) \\ &\overset{(17)}{=} F^{-1}\left(\frac{k}{n+1}\right) + \left(U_{(k)} - \frac{k}{n+1}\right)\left(F^{-1}\right)'\left(\frac{k}{n+1}\right) \end{aligned} \tag{19}$$

Es gilt

$$\left(F^{-1}\right)'(u) = \frac{1}{f(F^{-1}(u))}$$

(siehe Heuser (2001)).
Somit gilt

$$F^{-1}\left(U_{(k)}\right) \approx F^{-1}\left(\frac{k}{n+1}\right) + \left(U_{(k)} - \frac{k}{n+1}\right)\frac{1}{f\left(F^{-1}(\frac{k}{n+1})\right)} \tag{20}$$

Also gilt

$$\begin{aligned} Var\left(F^{-1}\left(U_{(k)}\right)\right) &\approx Var\left(F^{-1}\left(\frac{k}{n+1}\right) + \left(U_{(k)} - \frac{k}{n+1}\right)\frac{1}{f(F^{-1}\left(\frac{k}{n+1}\right))}\right) \\ &= Var\left(\left(U_{(k)} - \frac{k}{n+1}\right)\frac{1}{f(F^{-1}\left(\frac{k}{n+1}\right))}\right) \\ &= \frac{1}{\left(f(F^{-1}(\frac{k}{n+1}))\right)^2} Var\left(U_{(k)}\right) \\ &\overset{(17)}{=} \frac{1}{\left(f(F^{-1}(\frac{k}{n+1}))\right)^2}\left(\frac{1}{n+2}\right)\left(\frac{k}{n+1}\right)\left(1 - \frac{k}{n+1}\right) \end{aligned} \tag{21}$$

Mit $k = \frac{n+1}{2}$ gilt also

$$Var\left(X_{(\frac{n+1}{2})}\right) = Var\left(F^{-1}\left(U_{(\frac{n+1}{2})}\right)\right) \approx \frac{1}{\left(f(F^{-1}(0.5))\right)^2}\left(\frac{1}{n+2}\right)0.5 \cdot 0.5$$

Somit gilt für ungerades n für die Varianz des Medians

$$Var(X_{0.5}) = \frac{1}{4\,n\,(f(F^{-1}(0.5)))^2} \tag{22}$$

Die gleiche Beziehung gilt für gerade Stichprobenumfänge.

Beispiel 6 (fortgesetzt)
Sind die Zufallsvariablen $X_1, \ldots, X_n$ standardnormalverteilt, so gilt $\sigma^2 = 1$ und $f(0) = \frac{1}{\sqrt{2\,\pi}}$.
Also gilt bei Standardnormalverteilung

$$Var(\bar{X}) = \frac{1}{n}$$

und

$$Var(X_{0.5}) \approx \frac{2\,\pi}{4\,n} = \frac{\pi}{2\,n} = \frac{1.57}{n}$$

Um zwei Schätzfunktionen T_1 und T_2 für einen Parameter θ zu vergleichen, betrachtet man das Verhältnis der Varianzen

$$\frac{Var(T_1)}{Var(T_2)}$$

Man spricht auch von der relativen Effizienz.

Beispiel 6 (fortgesetzt)
Bei Normalverteilung gilt

$$\frac{Var(\bar{X})}{Var(X_{0.5})} \approx \frac{2}{\pi} = 0.637$$

Tabelle 4 zeigt die exakten Werte von $Var(\bar{X})/Var(X_{0.5})$ bei Normalverteilung. Wir sehen, dass für kleine Werte von n die Asymptotik noch nicht greift.

Tabelle 4: $\frac{Var(\bar{X})}{Var(X_{0.5})}$ bei Normalverteilung

n	1	3	5	7	9	11	13	15	17
$\frac{Var(\bar{X})}{Var(X_{0.5})}$	1	0.743	0.697	0.679	0.669	0.663	0.659	0.656	0.653

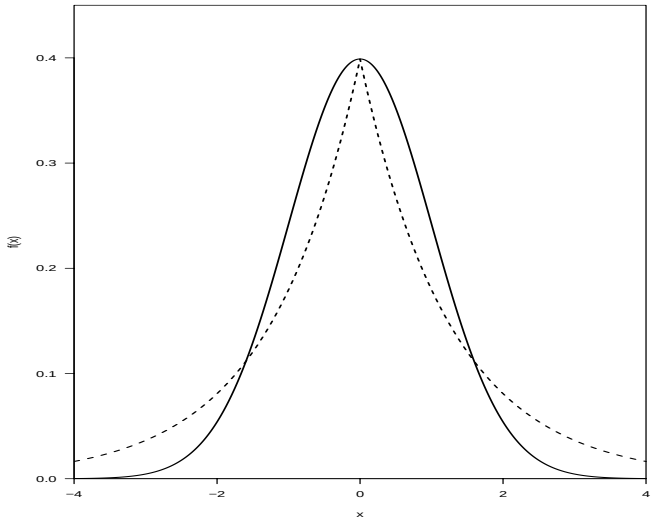
Quelle: Kendall et al. (1991).

Beispiel 6 (fortgesetzt)
Schauen wir noch ein weiteres Verteilungsmodell an. Die Dichtefunktion der Laplace-Verteilung ist gegeben durch:

$$f(x) = \frac{1}{2\,\beta}\, e^{-|x-\mu|/\beta}$$

Abbildung 4 zeigt die Dichtefunktion der Laplace-Verteilung mit $\mu = 0$ und $\beta = 1$ und die Dichtefunktion der Standardnormalverteilung. Die Dichtefunktionen sind so skaliert, dass sie im Nullpunkt die gleiche Höhe haben.Wir sehen, dass die Laplace-Verteilung im Zentrum steiler als die Standardnormalverteilung ist und an den Rändern mehr Wahrscheinlichkeitsmasse als die Standardnormalverteilung besitzt. Somit treten extreme Werte bei der Laplace-Verteilung häufiger auf als bei der Standardnormalverteilung.

Abbildung 4: Dichtefunktionen der Laplace-Verteilung und der Standardnormalverteilung



Es gilt $E(X) = \mu$ und $Var(X) = 2\,\beta^2$. (siehe dazu Mood et al. (1974), S. 117)
Somit gilt

$$Var(\bar{X}) = \frac{2\,\beta^2}{n}$$

Mit

$$f(0) = \frac{1}{2\,\beta}$$

gilt also

$$Var(X_{0.5}) = \frac{1}{4\,n\,(1/(2\,\beta))^2} = \frac{\beta^2}{n}$$

Somit gilt

$$\frac{Var(\bar{X})}{Var(X_{0.5})} = 2$$

Wir sehen, dass der Median bei Laplace-Verteilung viel effizienter ist als der Mittelwert.

2.3 Simulation

Wir haben gesehen, dass die Asymptotik für kleine Werte von n noch nicht greift. Um auch hier vergleichen zu können, sollte man eine Simulation durchführen. Simulieren bedeutet: 'so tun als ob'. In der Statistik verwendet man Simulationen, um die Verteilung einer Stichprobenfunktion $S = g(X_1, \ldots, X_n)$ zu schätzen, wenn in der Grundgesamtheit eine Verteilung mit Verteilungsfunktion $F(x)$ vorliegt. Hierbei erzeugt man B Stichproben $x_1, \ldots, x_n$ aus der Verteilung und bestimmt für jede den Wert der Stichprobenfunktion. Man schätzt die Verteilung der Stichprobenfunktion durch die empirische Verteilung der realisierten Stichproben.
Um mit dem Computer Zufallszahlen aus speziellen Verteilungen ziehen zu können, benötigt man einen Generator für Zufallszahlen, die aus einer Gleichverteilung auf $(0, 1)$ stammen. Die Verteilungsfunktion und Dichtefunktion einer auf $(0, 1)$ gleichverteilten Zufallsvariablen ist in den Gleichungen (11) und (12) auf Seite 11 zu finden.
Naeve (1995) stellt eine Vielzahl von Verfahren zur Erzeugung auf $(0, 1)$ gleichverteilter Zufallszahlen an. Außerdem zeigt er, wie man testen kann, ob ein Generator unabhängige, auf $(0, 1)$ gleichverteilte Zufallszahlen erzeugt. Wir gehen im Folgenden davon aus, dass ein Zufallszahlengenerator für auf $(0, 1)$ gleichverteilte Zufallszahlen vorliegt.
Schauen wir uns zunächst an, wie man mit Hilfe von auf $(0, 1)$ gleichverteilten Zufallszahlen eine Stichprobe aus einer Grundgesamtheit mit einem diskreten Merkmal X zieht. Seien $x_1, \ldots, x_k$ die Merkmalsausprägungen der diskreten Zufallsvariablen X. Die Wahrscheinlichkeitsfunktion $P(X = x_i) = p_i$ sei für $i = 1, 2, \ldots, k$ bekannt.

Beispiel 7
Wir werfen einen fairen Würfel einmal. Sei X die Augenzahl. Dann gilt

$$P(X = i) = \frac{1}{6}$$

für $i = 1, 2, 3, 4, 5, 6$.

Um nun nun Zufallszahlen zu erzeugen, die die Verteilung von X besitzen, benötigt man nur auf $(0, 1)$ gleichverteilte Zufallszahlen. Diese liefert jedes statistische Programmpaket. Wir erzeugen eine auf $(0, 1)$ gleichverteilte Zufallszahl u und bilden

$$x = \begin{cases} x_1 & \text{falls } 0 < u \leq p_1 \\ x_2 & \text{falls } p_1 < u \leq p_1 + p_2 \\ x_3 & \text{falls } p_1 + p_2 < u \leq p_1 + p_2 + p_3 \\ \vdots & \vdots \\ x_k & \text{falls } p_1 + p_2 + \ldots + p_{k-1} < u < 1 \end{cases}$$

Mit $p_0 = 0$ können wir dies auch schreiben als:
Wähle

$$x = x_k$$

wenn gilt

$$\sum_{i=0}^{k-1} p_i < u \leq \sum_{i=0}^{k} p_i$$

Somit gilt

$$\begin{aligned} P(X = x_k) &= P\left(\sum_{i=0}^{k-1} p_i < U \leq \sum_{i=0}^{k} p_i\right) = F_U\left(\sum_{i=0}^{k} p_i\right) - F_U\left(\sum_{i=0}^{k-1} p_i\right) \\ &= \sum_{i=0}^{k} p_i - \sum_{i=0}^{k-1} p_i = p_k \end{aligned}$$

Beispiel 7 (fortgesetzt)
Um einmal zu würfeln, erzeugen wir eine gleichverteilte Zufallszahl u und bilden

$$x = \begin{cases} 1 & \text{falls } 0 < u \leq \frac{1}{6} \\ 2 & \text{falls } \frac{1}{6} < u \leq \frac{2}{6} \\ 3 & \text{falls } \frac{2}{6} < u \leq \frac{3}{6} \\ 4 & \text{falls } \frac{3}{6} < u \leq \frac{4}{6} \\ 5 & \text{falls } \frac{4}{6} < u \leq \frac{5}{6} \\ 6 & \text{falls } \frac{5}{6} < u \leq 1 \end{cases}$$

Ist die gezogene gleichverteilte Zufallszahl u gleich 0.5841235, so wird eine 4 gewürfelt.

Bei der Erzeugung von Zufallszahlen aus einer Grundgesamtheit mit stetiger Verteilungsfunktion $F(x)$ greifen wir auf die Aussag von Satz 2.1 auf Seite 13 zurück. Wir erzeugen eine auf $(0, 1)$ gleichverteilte Zufallszahl u und erhalten die Zufallszahl aus $F(x)$ durch $x = F^{-1}(u)$.

Beispiel 8
Die Zufallsvariable X besitze eine Laplace-Verteilung mit den Parametern $\mu = 0$ und $\beta = 1$. Die Dichtefunktion von X lautet also

$$f_X(x) = \frac{1}{2}\, e^{-|x|} = \begin{cases} \frac{1}{2}\, e^{x} & \text{für } x < 0 \\ \frac{1}{2}\, e^{-x} & \text{für } x \geq 0 \end{cases}$$

Die Verteilungsfunktion lautet

$$F_X(x) = \begin{cases} \frac{1}{2}\, e^{x} & \text{für } x < 0 \\ 1 - \frac{1}{2}\, e^{-x} & \text{für } x \geq 0 \end{cases}$$

Wir erhalten eine Zufallszahl aus der Laplace-Verteilung, indem wir eine auf $(0, 1)$ gleichverteilte Zufallszahl u erzeugen. Die Zufallszahl x aus der Laplace-Verteilung ist dann

$$x = \begin{cases} \ln 2\, u & \text{für } u < 0.5 \\ -\ln(2 - 2\, u) & \text{für } u \geq 0.5 \end{cases}$$

Schauen wir uns an, wie man mit einer Simulation die Verteilung einer Stichprobenfunktion $S = g(X_1, \ldots, X_n)$ schätzen kann, wenn man für die Verteilung der Grundgesamtheit ein spezielles Verteilungsmodell $F_X(x)$ unterstellt. Hierbei geht man folgendermaßen vor:

1. Gib die Anzahl B der Stichproben vor.
2. Setze i auf den Wert 1.
3. Erzeuge eine Zufallsstichprobe $x_1, \ldots, x_n$ aus der Verteilung $F_X(x)$.
4. Bestimme den Wert s_i der Statistik $S = g(X_1, \ldots, X_n)$ für diese Stichprobe.
5. Erhöhe die Zählvariable i um 1.
6. Gehe nach 3., wenn gilt $i \leq B$.

7. Schätze die Verteilung von $S = g(X_1, \ldots, X_n)$ durch die Verteilung von $s_1, \ldots, s_B$.

Beispiel 7 (fortgesetzt)
Uns interessiert die Verteilung des Minimums $M = \min\{X_1, X_2, X_3, X_4\}$, also $P(M = i)$ für $i = 1, 2, 3, 4, 5, 6$. Wir schätzen diese Verteilung durch Simulation. Dabei werfen wir 4 Würfel 10000-mal. In Tabelle 5 sind die Ergebnisse zu finden.

Tabelle 5: Durch Simulation geschätzte Verteilung des Minimums beim Wurf von vier Würfeln

i	1	2	3	4	5	6
$\widehat{P(M = i)}$	0.5196	0.2848	0.1346	0.0474	0.0131	0.0005

Für $i = 1$ und $i = 6$ können wir diese Werte leicht mit den wahren Werten $P(X = i)$. Das Minimum nimmt den Wert 1 an, wenn mindestens eine 1 bei den 4 Würfen aufgetreten ist. Also gilt

$$P(X = 1) = 1 - \left(\frac{5}{6}\right)^4 = 0.5177$$

Das Minimum ist gleich 6, wenn bei allen 4 Würfen die 6 auftritt:

$$P(X = 6) = \left(\frac{1}{6}\right)^4 = 0.00077$$

Beispiel 7 (fortgesetzt)
Wir schätzen die Varianz von $\bar{X}$ und $X_{0.5}$ für Stichproben vom Umfang $n = 5$, $n = 10$ und $n = 20$ aus der Laplace-Verteilung mit Parametern $\mu = 0$ und $\beta = 1$ mit 10000 Wiederholungen.
In Tabelle 6 sind die Ergebnisse der Simulation zu finden.

Tabelle 6: Durch Simulation geschätzte Varianz von $\bar{X}$ und $X_{0.5}$ für Stichproben vom Umfang $n = 5$, $n = 10$ und $n = 20$ aus der Laplace-Verteilung mit Parametern $\mu = 0$ und $\beta = 1$ mit 10000 Wiederholungen

n	$\widehat{Var}(\bar{X})$	$\widehat{Var}(X_{0.5})$
5	0.390	0.342
10	0.198	0.144
20	0.100	0.066

Wir sehen, dass die Varianz von $\bar{X}$ bei der Laplace-Verteilung größer ist als die Varianz von $X_{0.5}$. Somit ist der Median bei der Laplace-Verteilung ein effizienterer Schätzer als der Mittelwert.

2.4 Der Bootstrap

Wir haben oben gesehen, dass man die Verteilung einer Stichprobenfunktion $g(X_1, \ldots, X_n)$ durch Simulation schätzen kann. Hierzu erzeugt man Stichproben aus der Verteilung und bestimmt für jede Stichprobe den Wert der Stichprobenfunktion. Die empirische Verteilung der Stichprobenfunktion approximiert dann die theoretische Verteilung.
Nun ist in der Regel die Verteilung die Verteilungsfunktion $F_X(x)$ der Grundgesamtheit unbekannt. Man kann somit die Verteilung der Stichprobenfunktion nicht mit einer Simulation dadurch schätzen, dass man Stichproben aus $F_X(x)$ zieht. Efron (1979) hat vorgeschlagen, die Stichproben nicht aus der unbekannten Verteilungsfunktion $F_X(x)$ sondern aus der empirischen Verteilungsfunktion $F_n(x)$ zu ziehen. Das bedeutet, dass man aus der Stichprobe $x_1, \ldots, x_n$ mit Zurücklegen B Stichproben $x_1^*, \ldots, x_N^*$ ziehen. Efron nannte diesen Verfahren den Bootstrap. Man spricht auch von der Bootstrap-Stichprobe $x_1^*, \ldots, x_N^*$. Dabei muss nicht notwendigerweise N gleich n sein.
Ist man also an der Verteilung einer Stichprobenfunktion $S = g(X_1, \ldots, X_n)$ interessiert, wenn gilt $X_i \sim F_X(x)$, so approximiert der Bootstrap diese Verteilung durch die Verteilung von $S^* = g(X_1^*, \ldots, X_N^*)$, wobei gilt $X_i^* \sim F_n(x)$. Die Bootstrap-Verteilung kann man nun mit Hilfe einer Simulation bestimmen:

1. Gib die Anzahl B der Stichproben vor, die gezogen werden sollen.

2. Setze i auf den Wert 1.

3. Erzeuge eine Bootstrap-Stichprobe $x_1^*, \ldots, x_N^*$ aus der empirischen Verteilungsfunktion $F_n(x)$, d. h. ziehe mit Zurücklegen eine Stichprobe $x_1^*, \ldots, x_N^*$ aus der Stichprobe $x_1, \ldots, x_n$.

4. Bestimme den Wert s_i^* der Statistik $S^* = g(X_1^*, \ldots, X_N^*)$ für diese Stichprobe.

5. Erhöhe die Zählvariable i um 1.

6. Gehe nach 3., wenn gilt $i \leq B$.

7. Schätze die Verteilung von $S = g(X_1, \ldots, X_n)$ durch die Verteilung von $s_1^*, \ldots, s_B^*$.

Beispiel 7 (fortgesetzt)
Wir wollen die Varianz des Medians für die folgende Stichprobe schätzen:

```
16 19 13 17 19 23 17 25
```

Wir ziehen 10 Bootstrap-Stichproben und bestimmen für jede den Median. Tabelle 7 zeigt die Stichproben mit den zugehörigen Werten des Medians $\tilde{x}_i^*$. Wir bezeichnen den Median hier mit $\tilde{x}$ und nicht mit $x_{0.5}$, um nicht einen doppelten Index benutzen zu müssen.

Tabelle 7: Bootstrap-Stichproben

Stichprobe	x_1^*	x_2^*	x_3^*	x_4^*	x_5^*	x_6^*	x_7^*	x_8^*	$\tilde{x}_i^*$
1	19	19	13	23	19	19	19	23	19
2	23	19	25	23	19	19	25	16	19
3	17	19	17	13	23	17	17	23	17
4	25	19	17	25	16	17	23	25	19
5	19	19	19	17	19	23	19	17	19
6	16	25	17	25	13	17	16	13	16
7	19	19	17	23	13	23	19	19	19
8	19	17	17	23	19	25	17	19	19
9	19	16	25	17	23	17	17	17	17
10	17	19	19	17	17	19	17	19	17

Wir schätzen die Varianz des Medians durch

$$\widehat{Var}(X_{0.5}) = \frac{1}{B-1} \sum_{i=1}^{B} \left(\tilde{x}_i^* - \overline{\tilde{x}_i^*}\right)^2$$

mit

$$\overline{\tilde{x}_i^*} = \frac{1}{B} \sum_{i=1}^{B} \tilde{x}_i^*$$

Es gilt $\overline{\tilde{x}_i^*} = 18.1$ und $\widehat{Var}(X_{0.5}) = 1.43$.

Wir können den Bootstrap auch benutzen, um uns auf Basis der Stichprobe $x_1, \ldots, x_n$ für einen der beiden Schätzer zu entscheiden. Wählen wir die Varianzen der Schätzfunktionen als Kriterium, so müssen wir unterstellen, dass die Verteilung der Grundgesamtheit symmetrisch ist. Die Verteilung der Stichprobe, aus der wir die Bootstrap-Stichproben ziehen, sollte die Annahmen erfüllen, die wir an die Grundgesamtheit stellen. Wir müssen die Stichprobe symmetrisieren. Ist ein Wert x_i in der Stichprobe, so muss auch der bezüglich des Symmetriezentrums $\hat{\mu}$ symmetrische Wert in der Stichprobe sein. Die folgende Abbildung zeigt, dass der zu x_i bezüglich $\hat{\mu}$ symmetrische Punkt gleich

$$\hat{\mu} + (\hat{\mu} - x_i) = 2\,\hat{\mu} - x_i)$$

ist.

Die symmetrisierte Stichprobe ist also

$$x_1, \ldots, x_n, \hat{\mu} - (x_1 - \hat{\mu}), \ldots, \hat{\mu} - (x_n - \hat{\mu})$$

In der Regel wählt man $\hat{\mu} = x_{0.5}$.

Beispiel 7 (fortgesetzt)
Wir wählen $\hat{\mu} = x_{0.5} = 18$ und erhalten die symmetrisierte Stichprobe

16 19 13 17 19 23 17 25 **20 17 23 19 17 13 19 11**

Die durch die Symmetrisierung gewonnenen Werte sind fett gedruckt. So erhält man den wert 20 aus dem Wert 16 durch

$$2 \cdot 18 - 16 = 20$$

Wir ziehen mit Zurücklegen 10 Stichproben vom Umfang $n = 8$ aus der symmetrisierten Stichprobe. Die Stichproben und die Werte des Mittelwertes und des Medians jeder Stichprobe sind in Tabelle 8 zu finden.

Es gilt $\widehat{Var}(\bar{X}) = 1.524$ und $\widehat{Var}(X_{0.5}) = 2.62$. Somit sollte man die Lage des Datensatzes durch den Mittelwert beschreiben.

Tabelle 8: Bootstrap-Stichproben aus symmetrisierter Stichprobe

Stichprobe	x_1^*	x_2^*	x_3^*	x_4^*	x_5^*	x_6^*	x_7^*	x_8^*	$\bar{x}_i^*$	$\tilde{x}_i^*$
1	19	13	23	13	20	19	11	17	16.875	18
2	25	17	19	13	19	17	16	17	17.875	17
3	19	17	19	11	23	13	23	23	18.500	19
4	23	13	11	20	17	17	19	17	17.125	17
5	25	25	13	17	11	11	19	11	16.500	15
6	13	19	19	19	17	19	23	19	18.500	19
7	19	17	16	23	16	23	17	19	18.750	18
8	13	17	19	17	17	17	17	19	17.000	17
9	23	19	25	17	17	17	13	17	18.500	17
10	25	11	25	17	19	16	25	23	20.125	21

Literatur

David, H. A. (1981). *Order statistics*. Wiley, New York, 2 edition.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7:1–26.

Heuser, H. (2001). *Lehrbuch der Analysis Teil 1*. Teubner, Stuttgart, 14 edition.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. Wiley, New York.

Kendall, M. G., Stuart, A., and Ord, J. K. (1991). *The advanced theory of statistics.*, volume 2 Classical inference and relationship. Arnold, London, 5 edition.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the theory of statistics*. McGraw-Hill, New York.

Naeve, P. (1995). *Stochastik für Informatik*. Oldenbourg, München, 1 edition.

Rudin, W. (1976). *Principles of mathematical analysis*. McGraw-Hill, New York, 3 edition.