Methoden der Datenanalyse

Andreas Handl

Torben Kuhlenkasper

8. Januar 2016

Grundlage des vorliegenden Skripts sind Aufzeichnungen von Andreas Handl, die er bis zum Jahr 2007 an der Universität Bielefeld verfasst und für seine Lehrveranstaltungen verwendet hat. Seit 2012 werden die Skripten von Torben Kuhlenkasper weitergeführt sowie fortlaufend aktualisiert und erweitert.

Anmerkungen und Vorschläge zur Verbesserung und Ergänzung sind jederzeit willkommen und können an `statistik@kuhlenkasper.de` *gesendet werden. Weitere Skripten sind unter* `www.skripten.kuhlenkasper.de` *zu finden.*

Inhaltsverzeichnis

Kapitel 1

Quantile

1.1 Was sind Quantile und wozu benötigt man sie?

Wir betrachten im Folgenden eine stetige Zufallsvariable X mit Verteilungsfunktion $F_X(x)$. Es gilt

$$F_X(x) = P(X \leq x)$$

Mit der Verteilungsfunktion kann man also Wahrscheinlichkeiten bestimmen. Oft ist man aber nicht an Wahrscheinlichkeiten interessiert, sondern man gibt eine Wahrscheinlichkeit p vor und sucht den Wert von X, der mit Wahrscheinlichkeit p nicht überschritten wird. Man spricht vom **Quantil** x_p. Für x_p gilt:

$$F_X(x_p) = p \tag{1.1}$$

Man bestimmt Quantile also über die Verteilungsfunktion. Hierbei muss man bei einer stetigen Zufallsvariablen X zwei Fälle unterscheiden.
Ist die Verteilungsfunktion $F_X(x)$ streng monoton wachsend, so sind alle Quantile eindeutig definiert. Es gilt

$$x_p = F_X^{-1}(p) \tag{1.2}$$

In Abbildung 1.1 wird gezeigt, wie man das 0.8413-Quantil der Standardnormalverteilung bestimmen kann.
Bei der Normalverteilung kann man die inverse Verteilungsfunktion nicht in expliziter Form angeben. Man muss aber nur die Quantile z_p der Standardnormalverteilung tabellieren. Für eine mit den Parametern μ und σ^2 normalverteilte Zufallsvariable gilt

$$x_p = \mu + z_p\,\sigma \tag{1.3}$$

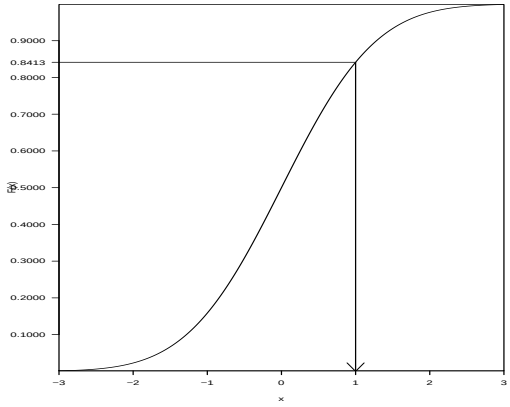


Abbildung 1.1: Bestimmung des 0.8413-Quantils der Standardnormalverteilung

Bei der Exponentialverteilung mit Verteilungsfunktion

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$$

kann man die Quantile in Abhängigkeit von p explizit angeben:

$$x_p = -\frac{1}{\lambda} \ln(1-p) \tag{1.4}$$

Dies sieht man folgendermaßen:

$$\begin{aligned} 1 - e^{-\lambda x_p} = p &\iff e^{-\lambda x_p} = 1 - p \\ &\iff -\lambda x_p = \ln(1-p) \\ &\iff x_p = -\frac{1}{\lambda} \ln(1-p) \end{aligned}$$

Ist die Verteilungsfunktion $F_X(x)$ einer stetigen Zufallsvariablen X nicht streng monoton wachsend, so ist x_p für einen oder mehrere Werte von p nicht eindeutig definiert, da für alle Punkte aus einem Intervall die Verteilungsfunktion den Wert p annimmt. In diesem Fall wählt man den kleinsten Wert von X, für den die Verteilungsfunktion gleich p ist.
Abbildung 1.2 zeigt dies.
Die folgende Definition umfasst beide Fälle

$$x_p = F^{\leftarrow}(p) = \inf\{x | F_X(x) \geq p\} \tag{1.5}$$

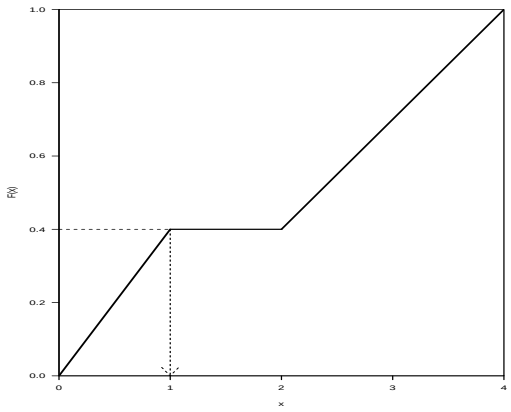


Abbildung 1.2: Bestimmung eines Quantils, falls die Verteilungsfunktion nicht streng monoton wachsend ist

Schauen wir uns einige Beispiele an, bei denen nach Quantilen gefragt wird.

Beispiel 1 40 *Prozent der Fläche Hollands liegt unter dem Meeresspiegel. Um sich gegen Überschwemmungen zu schützen, werden Deiche gebaut. Diese sollten natürlich hoch genug sein, um jeder Sturmflut zu trotzen. Es wird gefordert, dass die Deiche so hoch sind, dass die Wahrscheinlichkeit einer Überschwemmung 0.0001 beträgt. Ist X die Höhe des Wasserspiegels, so wird* $x_{0.9999}$ *gesucht. Der Artikel von de Haan (siehe dazu de Haan (1990)) beschäftigt sich mit diesem Problem.*

Beispiel 2 *Bei Finanzwerten ist der maximale Verlust von Interesse. Der Value at risk (Var) ist derjenige Verlust aus dem Halten eines Finanzwertes, der mit einer hohen Wahrscheinlichkeit nicht überschritten wird. Ist X also der Verlust, und p die Wahrscheinlichkeit, so ist* x_p *gesucht.*

Beispiel 3 *Meister Meister (1984) beschäftigt sich in seiner Arbeit mit unterschiedlichen Aspekten der Quantilschätzung. Als ein Beispiel wählt er die Bestimmung der Obergrenze für den Anteil an Bindegewebe in Wurst. Gesucht ist der Anteil an Bindegewebe, der mit Wahrscheinlichkeit p nicht überschritten wird.*

Quantile werden auch benutzt, um **Charakteristika von Verteilungen** zu beschreiben.
Die **Lage** wird durch den **Median** $x_{0.5}$ und die **Variabilität** durch den **Quartilsabstand**

$$IQR = x_{0.75} - x_{0.25} \tag{1.6}$$

beschrieben. Eine Maßzahl für die **Schiefe** ist der **Koeffizient von Bowley** (siehe dazu Bowley (1920)):

$$Q_S = \frac{(x_{0.75} - x_{0.5}) - (x_{0.5} - x_{0.25})}{x_{0.75} - x_{0.25}} \tag{1.7}$$

und eine Maßzahl für das Verhalten der Verteilung im Zentrum und an den Rändern ist das **Kurtosis-Maß von Moors** (siehe dazu Moors (1988)):

$$KM = \frac{(x_{0.875} - x_{0.625}) - (x_{0.375} - x_{0.125})}{x_{0.75} - x_{0.25}} \tag{1.8}$$

Mit der Maßzahl für die Schiefe werden wir uns im Kapitel über Symmetrie beschäftigen.

1.2 Schätzung von Quantilen

Da die Verteilung der Grundgesamtheit in der Regel nicht bekannt ist, muss man Quantile schätzen. Wir ziehen eine Zufallsstichprobe $x_1, \ldots, x_n$ aus der Grundgesamtheit. Die Beobachtungen $x_1, \ldots, x_n$ sind also Realisationen der unabhängigen, identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$.
Bei der Schätzung der Quantile geht man in Abhängigkeit von den Annahmen, die man über die Grundgesamtheit machen kann, unterschiedlich vor. Wir gehen zunächst davon aus, dass die Verteilungsklasse der Grundgesamtheit bekannt ist.

1.2.1 Schätzung von Quantilen bei bekannter Verteilungsklasse

Ist das parametrische Modell bis auf die Werte der Parameter bekannt, so ist die Quantilschätzung besonders einfach, wenn es sich um eine **Lage-Skalen-Familie** von Verteilungen handelt.

Definition 1.2.1 *Die Verteilung einer Zufallsvariablen X gehört zu einer Lage-Skalen-Familie von Verteilungen, wenn eine Verteilungsfunktion $F(x)$ und Parameter θ und λ existieren, sodass für die Verteilungsfunktion von X gilt*

$$F_X(x) = F\left(\frac{x - \theta}{\lambda}\right) \tag{1.9}$$

Dabei ist θ ein Lage- und λ ein Skalenparameter.

Beispiel 4 *Die Verteilungsfunktion $F_X(x)$ einer mit den Parametern μ und σ^2 normalverteilten Zufallsvariablen gehört zu einer Lage-Skalen-Familie von Verteilungen, da gilt*

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Dabei ist $\Phi(z)$ die Verteilungsfunktion der Standardnormalverteilung, bei der μ gleich 0 und σ gleich 1 ist.

Ist z_p das p-Quantil von $F(z)$, so ist bei einer Lage-Skalen-Familie von Verteilungen das p-Quantil von X gleich

$$x_p = \theta + z_p\,\lambda \tag{1.10}$$

Dies sieht man folgendermaßen:

$$\begin{aligned} F_X(x_p) = F\left(\frac{x_p-\theta}{\lambda}\right) &\iff p = F\left(\frac{x_p-\theta}{\lambda}\right) \\ &\iff F^{\leftarrow}(p) = \frac{x_p-\theta}{\lambda} \\ &\iff z_p = \frac{x_p-\theta}{\lambda} \\ &\iff x_p = \theta + z_p\,\lambda \end{aligned}$$

Beispiel 4 (fortgesetzt) *Für eine mit den Parametern μ und σ^2 normalverteilte Zufallsvariable gilt:*

$$x_p = \mu + z_p\,\sigma \tag{1.11}$$

Dabei ist z_p das p-Quantil der Standardnormalverteilung.

In einer Lage-Skalen-Familie von Verteilungen erhält man einen Schätzer $\hat{x}_p$ von x_p, indem man die Parameter θ und λ schätzt und in Gleichung (1.10) einsetzt:

$$\hat{x}_p = \hat{\theta} + z_p\,\hat{\lambda} \tag{1.12}$$

Beispiel 4 (fortgesetzt) *Die Maximum-Likelihood-Schätzer von μ und σ bei Normalverteilung sind*

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.13}$$

und

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.14}$$

Der Schätzer $\hat{\sigma}$ ist nicht erwartungstreu. Dividiert man $\hat{\sigma}$ durch

$$a_n = \frac{\Gamma[0.5(n-1)]}{\sqrt{2/n}\,\Gamma[0.5n]}$$

so erhält man eine erwartungstreue Schätzfunktion von σ (siehe dazu Johnson et al. (1994)). Dabei ist

$$\Gamma(n) = \int_0^\infty x^{n-1}\, e^{-x}\, dx \tag{1.15}$$

die Gammafunktion.
Für $n > 10$ kann man a_n approximieren durch

$$a_n = 1 + \frac{3}{4(n-1)}$$

siehe dazu Johnson et al. (1994).
Tabelle 1.1 zeigt die Werte von a_n für $n = 6, \ldots, 12$.

Tabelle 1.1: exakte und approximative Werte von a_n

n	6	7	8	9	10	11	12
a_n exakt	1.1512	1.1259	1.1078	1.0942	1.0837	1.0753	1.0684
a_n app.	1.1500	1.1250	1.1071	1.0938	1.0833	1.0750	1.0682

Beispiel 4 (fortgesetzt) *In einer Vorlesung wurde unter anderem nach der Körpergröße der männlichen Studierenden gefragt. Die Daten von 179 Personen sind im Anhang auf Seite 118 zu finden.*
Wir wollen $x_{0.99}$ schätzen und unterstellen Normalverteilung. Es gilt $\bar{x} = 182.2$ und $\hat{\sigma}^2 = 42.58$. Mit $z_{0.99} = 2.3263$ gilt

$$\hat{x}_p = 182.2 + 2.3263 \cdot 6.53 = 197.39$$

Für den erwartungstreuen Schätzer benötigen wir a_{179}*. Es gilt*

$$a_{179} = 1 + \frac{3}{4(179-1)} = 1.0042$$

Somit erhalten wir den erwartungstreuen Schätzer

$$\hat{x}_p = 182.2 + 2.3263 \cdot 6.53/1.0042 = 197.33$$

Meister (1984) schlägt in seiner Arbeit noch andere Schätzer von μ und σ bei Normalverteilung vor, setzt sie in Gleichung 1.11 auf Seite 8 ein und vergleicht die resultierenden Quantilschätzer in einer Simulationsstudie.

1.2.2 Schätzung von Quantilen bei klassierten Daten

Oft liegen die Daten in Form von Klassen vor.

Beispiel 4 (fortgesetzt) *Wir betrachten wieder die Körpergröße der* 179 *Studierenden und bilden* 8 *äquidistante Klassen der Breite* 5*. Die Untergrenze der ersten Klasse ist* 160*. Die Häufigkeitsverteilung ist in Tabelle 1.2 zu finden.*

Tabelle 1.2: Die Häufigkeitstabelle des Merkmals Körpergröße

k	x^*_{k-1}	x^*_k	n_k	h_k	$\hat{F}\left(x^*_{k-1}\right)$	$\hat{F}\left(x^*_k\right)$
1	160	165	1	0.0056	0.0000	0.0056
2	165	170	3	0.0168	0.0056	0.0224
3	170	175	17	0.0950	0.0224	0.1174
4	175	180	31	0.1732	0.1174	0.2906
5	180	185	68	0.3799	0.2906	0.6705
6	185	190	34	0.1899	0.6705	0.8604
7	190	195	19	0.1061	0.8604	0.9665
8	195	200	6	0.0335	0.9665	1.0000

Die Werte der empirischen Verteilungsfunktion ist nur an den Klassengrenzen bekannt. Man unterstellt, dass die Werte innerhalb der Klassen gleichverteilt sind, sodass die empirische Verteilungsfunktion innerhalb der Klassen linear ist.

Beispiel 4 (fortgesetzt) *Abbildung 1.3 zeigt die empirische Verteilungsfunktion.*

Um Quantile zu bestimmen, gehen wir wie bei einer theoretischen Verteilung vor. Sind x^*_{k-1} und x^*_k die Grenzen der i-ten Klasse und $\hat{F}\left(x^*_{k-1}\right)$ und $\hat{F}\left(x^*_k\right)$ der Wert der empirischen Verteilungsfunktion an diesen Klassengrenzen, so bestimmt man zunächst die Klasse k, für die gilt

$$\hat{F}\left(x^*_{k-1}\right) \leq p \leq \hat{F}\left(x^*_k\right)$$

Der Schätzer von x_p ist

$$\hat{x}_p = x^*_{k-1} + \frac{p - \hat{F}(x^*_{k-1})}{h_k} \cdot \Delta_k \tag{1.16}$$

Dabei ist h_k die relative Häufigkeit der k-ten Klasse und Δ_k die Breite der k-ten Klasse.

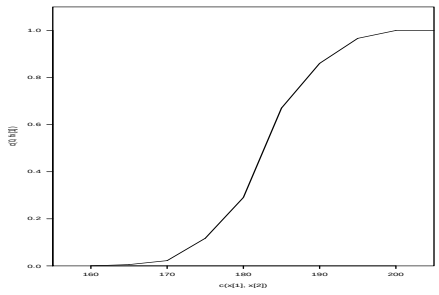


Abbildung 1.3: Empirische Verteilungsfunktion bei klassierten Daten

Beispiel 4 (fortgesetzt) *Wir bestimmen den Median $x_{0.5}$. Dieser liegt in der fünften Klasse. Es gilt*

$$\hat{x}_{0.5} = 180 + \frac{0.5 - 0.2906}{0.3799} \cdot 5 = 182.76$$

Wir bestimmen auch noch $x_{0.99}$. Es liegt in der achten Klasse. Es gilt

$$\hat{x}_{0.99} = 195 + \frac{0.99 - 0.9665}{0.0335} \cdot 5 = 198.51$$

1.2.3 Schätzung der Quantile aus der Urliste

Wir wollen nun Quantile aus Daten schätzen, bei denen keine Klassen gebildet wurden. Ausgangspunkt der Quantilschätzung ist die **empirische Verteilungsfunktion** $F_n(x)$. Die empirische Verteilungsfunktion an der Stelle x ist also gleich der Anzahl der Beobachtungen, die x nicht übertreffen. Sie ist eine **Treppenfunktion**.

Beispiel 5 *Betrachten wir hierzu folgenden Datensatz vom Umfang $n = 10$:*

```
47 48 49 51 52 53 54 57 65 70
```

Abbildung 1.4 zeigt die empirische Verteilungsfunktion.

Die empirische Verteilungsfunktion ist eine Treppenfunktion, sodass ihre Inverse nicht eindeutig definiert ist. Es ist naheliegend, Quantile dadurch zu schätzen, dass man in Gleichung 1.5 auf Seite 5 $F_X(x)$ durch $F_n(x)$ ersetzt:

$$\hat{x}_p = \inf\{x | F_n(x) \geq p\} \tag{1.17}$$

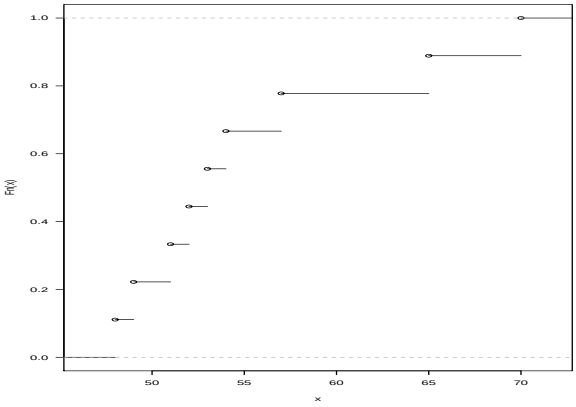


Abbildung 1.4: empirische Verteilungsfunktion

Da die empirische Verteilungsfunktion stückweise konstant ist, erhalten wir folgendes Ergebnis

$$\hat{x}_p = x_{(i)} \quad \text{für } \frac{i-1}{n} < p \leq \frac{i}{n} \tag{1.18}$$

mit $i = 1, \ldots, n$. Dabei sind $x_{(1)}, \ldots, x_{(n)}$ die geordneten Beobachtungen.

Beispiel 5 (fortgesetzt) *Es gilt*

$$\hat{x}_p = \begin{cases} 47 & \textit{für } 0 < p \leq 0.1 \\ 48 & \textit{für } 0.1 < p \leq 0.2 \\ 49 & \textit{für } 0.2 < p \leq 0.3 \\ 51 & \textit{für } 0.3 < p \leq 0.4 \\ 52 & \textit{für } 0.4 < p \leq 0.5 \\ 53 & \textit{für } 0.5 < p \leq 0.6 \\ 54 & \textit{für } 0.6 < p \leq 0.7 \\ 57 & \textit{für } 0.7 < p \leq 0.8 \\ 65 & \textit{für } 0.8 < p \leq 0.9 \\ 70 & \textit{für } 0.9 < p \leq 1 \end{cases}$$

Durch ein $x_{(i)}$ werden also unendlich viele Quantile geschätzt. Um zu eindeutigen Quantilschätzern zu gelangen, wird die empirische Verteilungsfunktion geglättet, indem man sie durch eine stetige stückweise lineare Funktion $\tilde{F}(x)$ ersetzt. Hierbei muss man festlegen, welchen Wert die Funktion $\tilde{F}(x)$ an den geordneten Beobachtungen $x_{(1)}, \ldots, x_{(n)}$ annimmt.

Es ist naheliegend, den Wert der empirischen Verteilungsfunktion in $x_{(i)}$ zu wählen:

$$\tilde{F}(x_{(i)}) = F_n(x_{(i)}) = \frac{i}{n}$$

für $i = 1, \ldots, n$ zu wählen und linear zu interpolieren.
Wie können wir $\hat{x}_p$ in Abhängigkeit von p ausdrücken? Für $p < 1/n$ gilt

$$\hat{x}_p = x_{(1)}$$

Nun gelte

$$\frac{i}{n} \leq p < \frac{i+1}{n}$$

für $i = 1, \ldots, n-1$.
Gilt

$$p = \frac{i}{n}$$

so ist

$$\hat{x}_p = x_{(i)} = x_{(np)}$$

Gilt

$$\frac{i}{n} < p < \frac{i+1}{n}$$

so müssen wir zwischen $x_{(i)}$ und $x_{(i+1)}$ mit $i = \lfloor np \rfloor$ linear interpolieren. Dabei ist $\lfloor a \rfloor$ die größte ganze Zahl, die kleiner oder gleich a ist. Es gilt

$$\frac{\hat{x}_p - x_{(i)}}{x_{(i+1)} - x_{(i)}} = \frac{p - i/n}{(i+1)/n - i/n}$$

Hieraus folgt

$$\hat{x}_p = (1 - (np - i))x_{(i)} + (np - i)x_{(i+1)} \tag{1.19}$$

Mit $g = np - i$ erhalten wir also

$$\hat{x}_p = \begin{cases} x_{(1)} & \text{für} \quad p < \frac{1}{n} \\ (1-g)\, x_{(i)} + g\, x_{(i+1)} & \text{für} \quad \frac{1}{n} \leq p \leq 1 \end{cases} \tag{1.20}$$

Die Grafik links oben in Abbildung 1.5 zeigt die Approximation.

Beispiel 5 (fortgesetzt) *Sei* $p = 0.25$. *Somit ist* $i = \lfloor 10 \cdot 0.25 \rfloor = \lfloor 2.5 \rfloor = 2$ *und* $g = 10 \cdot 0.25 - 2 = 0.5$. *Somit gilt*

$$\hat{x}_{0.25} = (1 - 0.5)\, x_{(2)} + 0.5\, x_{(3)} = 0.5 \cdot 48 + 0.5 \cdot 49 = 48.5$$

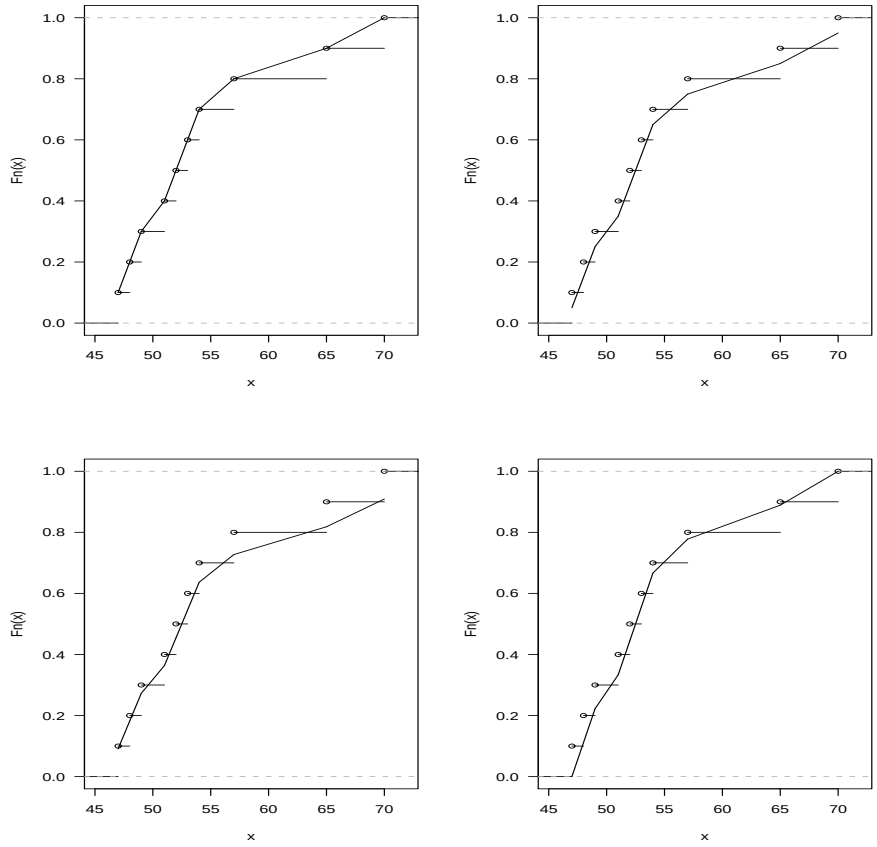


Abbildung 1.5: Vier Möglichkeiten, die empirische Verteilungsfunktion zu approximieren

Sei $p = 0.5$. *Somit ist* $i = \lfloor 10 \cdot 0.5 \rfloor = \lfloor 5 \rfloor = 2$ *und* $g = 10 \cdot 0.5 - 5 = 0$. *Somit gilt*

$$\hat{x}_{0.5} = x_{(5)} = 52$$

Sei $p = 0.99$. *Somit ist* $i = \lfloor 10 \cdot 0.99 \rfloor = \lfloor 9.9 \rfloor = 9$ *und* $g = 10 \cdot 0.99 - 9 = 0.9$. *Somit gilt*

$$\hat{x}_{0.99} = (1 - 0.9)\, x_{(9)} + 0.9\, x_{(10)} = 69.5$$

Der Schätzer besitzt mindestens zwei Mängel.
Als Schätzer für den Median erhalten wir

$$\hat{x}_{0.5} = \begin{cases} x_{(n/2)} & \text{falls } n \text{ gerade ist} \\ \dfrac{x_{((n-1)/2)} + x_{(1+(n-1)/2)}}{2} & \text{falls } n \text{ ungerade ist} \end{cases}$$

Der Median wird aber immer folgendermaßen geschätzt:

$$\hat{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade ist} \\ \frac{x_{(n/2)} + x_{(1+n/2)}}{2} & \text{falls } n \text{ gerade ist} \end{cases} \tag{1.21}$$

Zweitens wird eine Vielzahl von Quantilen durch das Minimum geschätzt. Für $0 < p < \frac{1}{n}$ wird x_p durch das Minimum des Datensatzes geschätzt. Für $p > \frac{n-1}{n}$ wird hingegen jedem p ein anderes x_p zugeordnet. Die beiden Ränder der Verteilung werden also unterschiedlich behandelt. Dieser Nachteil kann dadurch behoben werden, dass die relative Häufigkeit i/n der i-ten Orderstatistik $x_{(i)}$ zu gleichen Teilen auf die Bereiche unterhalb und oberhalb von $x_{(i)}$ aufgeteilt wird.
Somit gilt

$$\tilde{F}(x_{(i)}) = \frac{i - 0.5}{n}$$

Die Grafik rechts oben in Abbildung 1.5 zeigt die Approximation.
Als Quantilschätzer ergibt sich in diesem Fall:

$$\hat{x}_p = \begin{cases} x_{(1)} & \text{für} \quad p < \frac{0.5}{n} \\ (1-g)\, x_{(i)} + g\, x_{(i+1)} & \text{für} \quad \frac{0.5}{n} \leq p \leq \frac{n-0.5}{n} \\ x_{(n)} & \text{für} \quad p > \frac{n-0.5}{n} \end{cases} \tag{1.22}$$

mit $i = \lfloor np + 0.5 \rfloor$ und $g = np + 0.5 - i$. Dieser Schätzer wurde von Hazen (1914) vorgeschlagen.

Beispiel 5 (fortgesetzt) *Sei* $p = 0.25$.
Somit ist $i = \lfloor 10 \cdot 0.25 + 0.5 \rfloor = \lfloor 3 \rfloor = 3$ *und* $g = 10 \cdot 0.25 + 0.5 - 3 = 0$.
Somit gilt

$$\hat{x}_{0.25} = (1 - 0) \cdot x_{(3)} + 0 \cdot x_{(4)} = 49$$

Sei $p = 0.5$.
Somit ist $i = \lfloor 10 \cdot 0.5 + 0.5 \rfloor = \lfloor 5.5 \rfloor = 5$ *und* $g = 10 \cdot 0.5 + 0.5 - 5 = 0.5$.
Somit gilt

$$\hat{x}_{0.5} = (1 - 0.5) \cdot x_{(5)} + 0.5 \cdot x_{(6)} = 52.5$$

Sei $p = 0.99$. *Da* 0.99 *größer als* $(10 - 0.5)/10 = 0.95$ *ist, gilt* $\hat{x}_{0.99} = 70$.

Die beiden bisher betrachteten Schätzer wurden heuristisch entwickelt. Systematische Zugänge nehmen die Gleichung

$$F_X(x_{(i)}) = p_i$$

als Ausgangspunkt. Würden wir $F_X(x)$ kennen, könnten wir sofort p_i angeben. Da $F_X(x)$ aber unbekannt ist, berücksichtigen wir, dass $x_{(i)}$ die Realisation der Zufallsvariablen $X_{(i)}$ ist. Wir betrachten also die Zufallsvariable $F_X(X_{(i)})$ und wählen p_i als Charakteristikum der Verteilung von $F_X(X_{(i)})$. Sinnvolle Charakteristika sind der Erwartungswert, der Modus und der Median. Die Zufallsvariable $F_X(X_{(i)})$ besitzt eine Beta-verteilung mit den Parametern $a = i$ und $b = n - i + 1$, siehe dazu Randles and Wolfe (1979), S. 7. Es gilt also

$$f_{X_{(i)}}(t) = \begin{cases} \dfrac{1}{B(i, n-i+1)}\, t^{i-1}\,(1-t)^{n-i} & \text{für } 0 < t < 1 \\ 0 & \text{sonst} \end{cases} \tag{1.23}$$

mit

$$B(a,b) = \int_0^1 w^{a-1}\,(1-w)^{b-1}\,dw$$

Somit gilt

$$E(X_{(i)}) = \frac{i}{n+1} \tag{1.24}$$

Der Modus ist der Wert, bei dem die Dichtefunktion ihr Maximum annimmt. Da $\frac{1}{B(i,n-i+1)}$ eine multiplikative Konstante ist, müssen wir das Maximum der Funktion

$$g(t) = t^{i-1}\,(1-t)^{n-i}$$

bestimmen. Wir bestimmen das Maximum von

$$\ln g(t) = (i-1)\ln(t) + (n-i)\ln(1-t)$$

Es gilt

$$\frac{d}{dt}\ln g(t) = \frac{i-1}{t} - \frac{n-i}{1-t}$$

Notwendige Bedingung für einen Extremwert in t ist also

$$\frac{i-1}{t} = \frac{n-i}{1-t}$$

Lösen wir diese Gleichung nach t auf, so erhalten wir

$$t = \frac{i-1}{n-1} \tag{1.25}$$

Gleichung 1.24 auf Seite 17 legt folgende Approximation nahe:

$$\tilde{F}(x_{(i)}) = \frac{i}{n+1}$$

Die Grafik links unten in Abbildung 1.5 zeigt die Approximation.
Als Quantilschätzer ergibt sich in diesem Fall:

$$\hat{x}_p = \begin{cases} x_{(1)} & \text{für} \quad p < \frac{1}{n+1} \\ (1-g)\, x_{(i)} + g\, x_{(i+1)} & \text{für} \quad \frac{1}{n+1} \leq p \leq \frac{n}{n+1} \\ x_{(n)} & \text{für} \quad p > \frac{n}{n+1} \end{cases} \tag{1.26}$$

mit $i = \lfloor (n+1)\, p \rfloor$ und $g = (n+1)\, p - i$.

Beispiel 5 (fortgesetzt) *Sei* $p = 0.25$.
Somit ist $i = \lfloor 11 \cdot 0.25 \rfloor = \lfloor 2.75 \rfloor = 2$ *und* $g = 11 \cdot 0.25 - 2 = 0.75$. *Somit gilt*

$$\hat{x}_{0.25} = (1 - 0.75) \cdot x_{(2)} + 0.75 \cdot x_{(3)} = 48.75$$

Sei $p = 0.5$.
Somit ist $i = \lfloor 11 \cdot 0.5 \rfloor = \lfloor 5.5 \rfloor = 5$ *und* $g = 11 \cdot 0.5 - 5 = 0.5$. *Somit gilt*

$$\hat{x}_{0.5} = (1 - 0.5) \cdot x_{(5)} + 0.5 \cdot x_{(6)} = 52.5$$

Sei $p = 0.99$. *Da* 0.99 *größer als* $10/11 = 0.91$ *ist, gilt* $\hat{x}_{0.99} = 70$.

Gleichung 1.25 auf Seite 17 legt folgende Approximation nahe:

$$\tilde{F}(x_{(i)}) = \frac{i-1}{n-1}$$

Hier wird jedem p ein anderer Wert von x_p zugeordnet.
Die Grafik rechts unten in Abbildung 1.5 zeigt die Approximation.
Als Quantilschätzer ergibt sich in diesem Fall:

$$\hat{x}_p = (1-g)\, x_{(i)} + g\, x_{(i+1)} \tag{1.27}$$

mit $i = \lfloor (n-1)\, p + 1 \rfloor$ und $g = (n-1)\, p + 1 - i$.

Beispiel 5 (fortgesetzt) *Sei* $p = 0.25$.
Somit ist $i = \lfloor 9 \cdot 0.25 + 1 \rfloor = \lfloor 3.25 \rfloor = 3$ *und* $g = 9 \cdot 0.25 + 1 - 3 = 0.25$.
Somit gilt

$$\hat{x}_{0.25} = (1 - 0.25) \cdot x_{(3)} + 0.25 \cdot x_{(4)} = 49.5$$

Sei $p = 0.5$.
Somit ist $i = \lfloor 9 \cdot 0.5 + 1 \rfloor = \lfloor 5.5 \rfloor = 5$ *und* $g = 9 \cdot 0.5 + 1 - 5 = 0.5$. *Somit gilt*

$$\hat{x}_{0.5} = (1 - 0.5) \cdot x_{(5)} + 0.5 \cdot x_{(6)} = 52.5$$

Sei $p = 0.99$.
Somit ist $i = \lfloor 9 \cdot 0.99 + 1 \rfloor = \lfloor 9.91 \rfloor = 9$ *und* $g = 9 \cdot 0.99 + 1 - 9 = 0.91$. *Somit gilt*

$$\hat{x}_{0.99} = (1 - 0.91) \cdot x_{(9)} + 0.91 \cdot x_{(10)} = 69.55$$

Dieser Schätzer hat den Vorteil, dass man Quantile, die zu kleinem oder großem p gehören, nicht ausschließlich durch das Minimum oder das Maximum schätzt.
Alle diese Schätzer sind Spezialfälle von:

$$\tilde{F}(x_{(i)}) = \frac{i - \gamma}{n + 1 - \gamma - \delta} \tag{1.28}$$

Die zugehörige Klasse von Quantilschätzern ist:

$$\hat{x}_p = \begin{cases} x_{(1)} & \text{für} \quad p < \frac{1-\gamma}{n+1-\gamma-\delta} \\ (1-g)\, x_{(i)} + g\, x_{(i+1)} & \text{für} \quad \frac{1-\gamma}{n+1-\gamma-\delta} \leq \frac{n-\gamma}{n+1-\gamma-\delta} \\ x_{(n)} & \text{für} \quad p > \frac{n-\gamma}{n+1-\gamma-\delta} \end{cases} \tag{1.29}$$

mit $i = \lfloor (n + 1 - \gamma - \delta)\, p + \gamma \rfloor$ und $g = (n + 1 - \gamma - \delta)\, p + \gamma - i$.
Die bisher betrachteten Schätzer sind Spezialfälle mit:

- Quantilschätzer in Gleichung (1.20) auf Seite 14: $\gamma = 0$, $\delta = 1$
- Quantilschätzer in Gleichung (1.22) auf Seite 16: $\gamma = 0.5$, $\delta = 0.5$
- Quantilschätzer in Gleichung (1.26) auf Seite 18: $\gamma = 0$, $\delta = 0$
- Quantilschätzer in Gleichung (1.27) auf Seite 18: $\gamma = 1$, $\delta = 1$

Hyndman and Fan (1996) betrachten noch zwei weitere Spezialfälle. Wählt man $\gamma = \delta = 1/3$, erhält man eine Approximation des Medians der Verteilung von $F_X(X_{(i)})$. Von Blom (1958) wurde $\gamma = \delta = 3/8$ vorgeschlagen.
Gilt $\gamma = \delta$, so wird der Median nach der Formel in Gleichung 3.2 auf Seite 65 geschätzt. Dies zeigen Hyndman and Fan (1996).
Welchen dieser Schätzer sollte man verwenden? Hyndman and Fan (1996) geben 6 Kriterien an, die Quantilschätzer erfüllen sollten. Nur der Quantilschätzer mit $\gamma = \delta = 0.5$ erfüllt alle 6 Kriterien. Man kann die Schätzer

aber auch hinsichtlich ihrer Effizienz mit einer Simulationsstudie vergleichen. Diese wurde von Dielman et al. (1994) und Handl (1985) durchgeführt. Bevor wir auf das Ergebnis dieser Studie eingehen, schauen wir uns noch einen weiteren Quantilschätzer an.
Alle bisher betrachteten Quantilschätzer verwenden bei der Schätzung eines Quantils höchstens zwei Beobachtungen. Harrell und Davis (siehe Harrell and Davis (1982)) schlagen einen Quantilschätzer vor, der auf allen Beobachtungen beruht. Sie gehen aus von der geordneten Stichprobe $x_{(1)}, \ldots, x_{(n)}$. Die zu diesen Beobachtungen gehörenden Zufallsvariablen heißen Orderstatistiken $X_{(i)}$, $i = 1, \ldots, n$. Die Orderstatistiken sind im Gegensatz zu den Zufallsvariablen $X_1, \ldots, X_n$ nicht unabhängig und auch nicht identisch verteilt. Für die Dichtefunktion $g_j(x)$ von $X_{(j)}$ gilt

$$g_j(x) = \frac{1}{\beta(j, n+1-j)} F(x)^{j-1} (1 - F(x))^{n-j} \tag{1.30}$$

mit

$$B(a, b) = \int_0^1 w^{a-1} (1-w)^{b-1} \, dw$$

(siehe dazu Randles and Wolfe (1979)). Somit ist der Erwartungswert von $X_{(j)}$ gleich

$$E\left(X_{(j)}\right) = \frac{1}{\beta(j, n+1-j)} \int_{-\infty}^{\infty} x \, F(x)^{j-1} (1 - F(x))^{n-j} \, f(x) \, dx$$

Wir substituieren $y = F(x)$ mit $x = F^{-1}(y)$ und $\frac{d}{dx} F(x) = f(x)$. Es gilt

$$E\left(X_{(j)}\right) = \frac{1}{\beta(j, n+1-j)} \int_0^1 F^{-1}(y) \, y^{j-1} (1-y))^{n-j} \, dy \tag{1.31}$$

Blom (1958) zeigt

$$\lim_{n \to \infty} E\left(X_{((n+1)p)}\right) = x_p$$

Harrell und Davis schätzen x_p, indem sie $E\left(X_{((n+1)p)}\right)$ schätzen. Hierzu ersetzen sie $F^{-1}(y)$ durch $F_n^{-1}(y)$ mit

$$F_n^{-1}(p) = x_{(i)}$$

für $(i-1)/n < p \leq i/n$. Wir erhalten

$$\begin{aligned}\widehat{E}\left(X_{((n+1)p)}\right) &= \frac{\int_0^1 F_n^{-1}(y)\, y^{(n+1)p-1}\,(1-y))^{(n+1(1-p)-1}\, dy}{\beta((n+1)p,(n+1)(1-p))} \\ &= \sum_{i=1}^{n} w_{n,i}\, x_{(i)}\end{aligned}$$

mit

$$w_{n,i} = \frac{\int_{(i-1)/n}^{i/n} y^{(n+1)p-1}\,(1-y))^{(n+1)(1-p)-1}\, dy}{\beta((n+1)p,(n+1)(1-p))}$$

Beispiel 5 (fortgesetzt) *Wir erhalten* $\hat{x}_{0.25} = 49.31768$, $\hat{x}_{0.5} = 52.69547$ *und* $\hat{x}_{0.99} = 69.85095$. *Wir sehen, dass der Schätzwert des Medians nicht mit dem aus Gleichung 3.2 auf Seite 65 berechneten Wert identisch ist.*

Welchen der Quantilschätzer soll man anwenden? Von Dielman et al. (1994) und Handl (1985) wurden Simulationsstudien durchgeführt, in denen die Effizienz der Quantilschätzer hinsichtlich des mittleren quadratischen Fehlers für eine Vielzahl von Verteilungen verglichen wurden. Dabei schnitt der Harrell-Davis-Schätzer hervorragend ab, wenn nicht zu extreme Quantile geschätzt wurden. In diesem Fall sollte man aber die Verfahren des nächsten Kapitels anwenden.

Schauen wir uns noch die beiden Quartile $x_{0.25}$ und $x_{0.75}$ an. Tukey hat vorgeschlagen, das untere Quartil $x_{0.25}$ durch den Median der unteren Hälfte des geordneten Datensatzes zu schätzen. Dabei gehört der Median des Datensatzes zur unteren Hälfte des geordneten Datensatzes, wenn der Stichprobenumfang ungerade ist. Entsprechend wird das obere Quartil $x_{0.75}$ durch den Median der oberen Hälfte des geordneten Datensatzes geschätzt.

Beispiel 5 (fortgesetzt) *Der geordnete Datensatz ist*

```
47 48 49 51 52 53 54 57 65 70
```

Die untere Hälfte des geordneten Datensatzes ist

```
47 48 49 51 52
```

Also gilt $\hat{x}_{0.25} = 49$.
Die obere Hälfte des geordneten Datensatzes ist

```
53 54 57 65 70
```

Also gilt $\hat{x}_{0.75} = 57$.

Der Schätzer von Tukey ist nicht unter den bisher betrachteten speziellen Quantilschätzern. Er gehört aber approximativ zur Klasse von Quantilschätzern in Gleichung (1.29) auf Seite 19 mit $\gamma = 1/3$ und $\delta = 1/3$. Der Beweis ist bei Hoaglin et al. (1983) zu finden.

1.3 Schätzung extremer Quantile

Wie die Simulationsstudie von Dielman et al. (1994) zeigt, schneiden die Quantilschätzer, die wir im letzten Abschnitt betrachtet haben, bei der Schätzung extremer Quantile schlecht ab. Dies ist gerade für Werte von p mit $p > 1 - 1/n$ oder $p < 1/n$ der Fall. Es ist also nicht sinnvoll, einen nichtparametrischen Ansatz zu verwenden. Da in der Regel aber die Verteilungsklasse nicht bekannt ist, zu der die Verteilungsfunktion $F_X(x)$ gehört, muss man approximieren. Da extreme Quantile geschätzt werden sollen, sollte man extreme Beobachtungen benutzen. Hierbei kann man entweder die Anzahl k der Beobachtungen oder einen **Schwellenwert** u vorgeben und alle Beobachtungen verwenden, die größer als dieser Schwellenwert sind. An diese Beobachtungen passt man dann eine geeignete Verteilung an.
Die zugrundeliegende Theorie ist ausführlich bei Embrechts et al. (1995) und Coles (2001) beschrieben. Ich werde hier im Folgenden einen Überblick geben.
Man geht von einem Schwellenwert u aus und betrachtet die bedingte Verteilung von $X - u$ unter der Bedingung, dass X größer als u ist:

$$F_U(x) = P(X - u \leq x | X > u) \tag{1.32}$$

Es gilt

$$F_U(x) = \frac{F_X(x+u) - F_X(u)}{1 - F_X(u)} \tag{1.33}$$

Dies sieht man folgendermaßen

$$\begin{aligned} F_U(x) &= P(X - u \leq x | X > u) = \frac{P(u < X \leq x + u)}{1 - P(X \leq u)} \\ &= \frac{F_X(x+u) - F_X(u)}{1 - F_X(u)} \end{aligned} \tag{1.34}$$

Schauen wir uns ein Beispiel an.

Beispiel 6 *X sei exponentialverteilt mit Parameter λ. Es gilt also*

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \textit{für } x > 0 \\ 0 & \textit{sonst} \end{cases} \tag{1.35}$$

Dann gilt

$$\begin{aligned} F_U(x) &= \frac{F_X(x+u) - F_X(u)}{1 - F_X(u)} = \frac{1 - e^{-\lambda(x+u)} - (1 - e^{-\lambda u})}{1 - (1 - e^{-\lambda u})} \\ &= \frac{-e^{-\lambda x} e^{-\lambda u} + e^{-\lambda u}}{e^{-\lambda u}} = \frac{e^{-\lambda u}\left(-e^{-\lambda x} + 1\right)}{e^{-\lambda u}} = 1 - e^{-\lambda x} \end{aligned}$$

Wir sehen, dass die bedingte Verteilung eine Exponentialverteilung mit dem Parameter λ ist. Aus diesem Grund nennt man die Exponentialverteilung auch Verteilung ohne Gedächtnis.

Auf Grund des folgendes Satzes ist es möglich extreme Quantile zu schätzen.

Satz 1.3.1 *Liegt die Verteilungsfunktion $F_X(x)$ der Zufallsvariablen X im Maximum-Anziehungsbereich einer Extremwertverteilung, so ist für große Werte von u die bedingte Verteilung von $X - u$ unter der Bedingung $X > u$ approximativ gleich*

$$H(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-1/\xi} & \textit{für } \xi \neq 0 \\ 1 - e^{-x/\beta} & \textit{für } \xi = 0 \end{cases} \tag{1.36}$$

Eine Beweisskizze ist bei Coles (2001), S.76-77 zu finden.
Was bedeutet der im Satz verwendete Begriff *Anziehungsbereich einer Extremwertverteilung*?
Unter bestimmten Bedingungen besitzt das geeignet standardisierte Maximum einer Zufallsstichprobe $X_1, \ldots, X_n$ aus einer Grundgesamtheit mit Verteilungsfunktion $F_X(x)$ eine der folgenden Grenzverteilungen:

1. Die **Frechet-Verteilung**

$$\Phi(x) = \begin{cases} e^{-x^{-\alpha}} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$$

2. Die **Weibull-Verteilung**

$$\Psi(x) = \begin{cases} e^{-(-x)^{\alpha}} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$$

3. Die **Gumbel-Verteilung**

$$\Lambda(x) = e^{-e^{-x}}$$

Der Beweis ist bei Gnedenko (1943) zu finden.
Verteilungen mit **hoher Wahrscheinlichkeitsmasse an den Rändern** wie die **Cauchy-Verteilung** oder die **t-Verteilung** besitzen die Frechet-Verteilung als Grenzverteilung. Verteilungen mit **finitem rechten Randpunkt** wie die **Gleichverteilung** besitzen als Grenzverteilung die Weibull-Verteilung. Verteilungen wie die **Exponentialverteilung**, **Gammaverteilung**, **Lognormalverteilung** oder **Normalverteilung** besitzen die Gumbel-Verteilung als Grenzverteilung.

Kehren wir zu Satz 1.3.1 auf Seite 24 zurück. Die Verteilung in Gleichung (1.36) heißt **verallgemeinerte Pareto-Verteilung**.
Abbildung 1.6 zeigt die Dichtefunktion der verallgemeinerten Pareto-Verteilung für unterschiedliche Werte von ξ für $\beta = 1$. Wie können wir die Aussagen

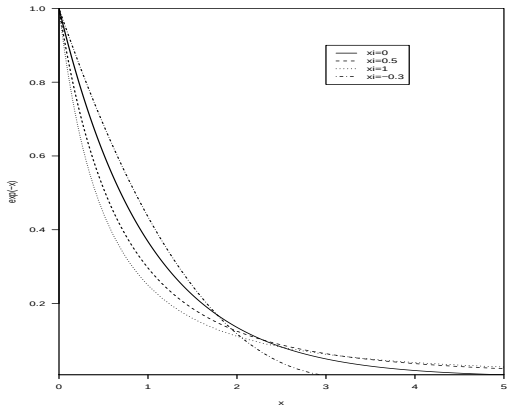


Abbildung 1.6: Dichtefunktion der verallgemeinerten Paretodiagramm

von Satz 1.3.1 auf Seite 24 nutzen, um extreme Quantile zu schätzen?
Wir müssen uns zuerst überlegen, wie wir das p-Quantil x_p von $F_X(x)$ in Abhängigkeit von der verallgemeinerten Pareto-Verteilung ausdrücken können. Hierzu ersetzen wir in Gleichung (1.34) auf Seite 23 x durch $x - u$ und erhalten:

$$F_U(x-u) = \frac{F_X(x) - F_X(u)}{1 - F_X(u)} \tag{1.37}$$

Ersetzen wir $F_U(x-u)$ in Gleichung (1.37) durch $H(x-u)$ aus der ersten Gleichung in Gleichung (1.36) auf Seite 24, so gilt für $\xi \neq 0$:

$$1 - \left(1 + \frac{\xi\,(x-u)}{\beta}\right)^{-1/\xi} = \frac{F_X(x) - F_X(u)}{1 - F_X(u)}$$

Gesucht ist x_p mit $F_X(x_p) = p$. Somit gilt

$$1 - \left(1 + \frac{\xi\,(x_p - u)}{\beta}\right)^{-1/\xi} = \frac{p - F_X(u)}{1 - F_X(u)} \iff$$

$$\left(1 + \frac{\xi\,(x_p - u)}{\beta}\right)^{-1/\xi} = 1 - \frac{p - F_X(u)}{1 - F_X(u)} \iff$$

$$\left(1 + \frac{\xi\,(x_p - u)}{\beta}\right)^{-1/\xi} = \frac{1 - p}{1 - F_X(u)} \iff$$

$$1 + \frac{\xi\,(x_p - u)}{\beta} = \left(\frac{1 - p}{1 - F_X(u)}\right)^{-\xi} \iff$$

$$x_p = u + \frac{\beta}{\xi}\left[\left(\frac{1 - p}{1 - F_X(u)}\right)^{-\xi} - 1\right]$$

Für $\xi = 0$ aus der zweiten Gleichung in Gleichung (1.36) auf Seite 24 erhalten wir:

$$\begin{aligned} 1 - e^{-x/\beta} = \frac{p - F_X(u)}{1 - F_X(u)} &\iff e^{-x/\beta} = 1 - \frac{p - F_X(u)}{1 - F_X(u)} \\ &\iff e^{-x/\beta} = \frac{1 - p}{1 - F_X(u)} \\ &\iff -x/\beta = \ln\left(\frac{1 - p}{1 - F_X(u)}\right) \\ &\iff x_p = -\beta\,\ln\left(\frac{1 - p}{1 - F_X(u)}\right) \end{aligned}$$

Es gilt also

$$x_p = \begin{cases} u + \frac{\beta}{\xi}\left[\left(\frac{1-p}{1-F_X(u)}\right)^{-\xi} - 1\right] & \text{für } \xi \neq 0 \\ -\beta\,\ln\left(\frac{1-p}{1-F_X(u)}\right) & \text{für } \xi = 0 \end{cases} \tag{1.38}$$

Nachdem wir x_p durch die Parameter der verallgemeinerten Pareto-Verteilung ausgedrückt haben, können wir x_p schätzen:

1. Wir geben einen Schwellenwert u vor.

2. Seien $y_{(1)}, \ldots, y_{(k)}$ die geordneten Beobachtungen, die größer als u sind. Wir passen an diese Beobachtungen die verallgemeinerte Paretoverteilung an. Wir bestimmen also die Schätzer $\hat{\beta}$ und $\hat{\xi}$. Hosking and Wallis (1987) zeigen, wie man hierbei vorzugehen hat.

3. Wir setzen die Schätzer in Gleichung 1.38 ein. Außerdem schätzen wir $1 - F_X(u)$ durch den Anteil k/n der Beobachtungen in der Stichprobe die größer als u sind, und erhalten als Quantilschätzer:

$$\hat{x}_p = \begin{cases} u + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{n}{k} (1-p) \right)^{-\hat{\xi}} - 1 \right] & \text{für } \hat{\xi} \neq 0 \\ -\hat{\beta} \ln \left(\frac{n}{k} (1-p) \right) & \text{für } \hat{\xi} = 0 \end{cases} \tag{1.39}$$

Beispiel 6 (fortgesetzt) *Wir wählen* $u = 180$. *Die Beobachtungen, die größer als* 180 *sind, sind*

```
181 181 181 181 181 181 181 182 182 182 182 182 182 182 182
182 182 182 182 182 182 182 183 183 183 183 183 183 183 183
183 183 183 183 183 184 184 184 184 184 184 184 184 184 184
184 184 184 184 185 185 185 185 185 185 185 185 185 185 185
185 185 186 186 186 186 186 186 186 187 187 187 187 187 188
188 188 189 189 189 189 189 189 190 190 190 190 190 190 190
191 191 191 191 192 192 192 192 192 192 193 194 195 196 198
198 198 200
```

Mit Hilfe von `R` *erhalten wir* $\hat{\beta} = 8.475$ *und* $\hat{\xi} = -0.38$.
Wir wollen $x_{0.99}$ *schätzen. Es gilt* $n/k = 179/108 = 1.66$. *Wir setzen diesen Wert und* $\hat{\beta} = 8.475$ *und* $\hat{\xi} = -0.38$ *in Gleichung 1.39 ein:*

$$\begin{aligned} \hat{x}_p &= u + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{n}{k} (1-p) \right)^{-\hat{\xi}} - 1 \right] \\ &= 180 - \frac{8.475}{0.38} \left[(1.66 \cdot 0.01)^{0.38} - 1 \right] = 197.6 \end{aligned}$$

Es stellt sich die Frage, wie man den Schwellenwert u festlegen soll.
Hierzu betrachten wir die **mittlere Exzess-Funktion** (MEF):

$$E(X - u | X > u)$$

Für die generalisierte Pareto-Verteilung gilt:

$$E(X - u|X > u) = \frac{\beta + \xi \cdot u}{1 - \xi}$$

(Siehe dazu Embrechts et al. (1995), S.165-166.) Die mittlere Exzess-Funktion verläuft bei verallgemeinerter Pareto-Verteilung also linear in u.
Als Schätzer der MEF an der Stelle u verwenden wir den um u verminderten Mittelwert der Beobachtungen, die größer als u sind. Wir bezeichnen die zugehörige Funktion als **empirische mittlere Exzess-Funktion** $e(u)$. Als Schätzer von u verwendet man den Wert u_0, ab dem $e(u)$ linear verläuft.

Beispiel 6 (fortgesetzt) *Wir wählen* $u = 190$*. Die Anzahl der Beobachtungen, die größer als* 190 *sind, sind:*

```
191 191 191 191 192 192 192 192 192 192 193 194 195 196 198
198 198 200
```

Der Mittelwert dieser Beobachtungen ist 193.78*. Also nimmt die empirische mittlere Exzess-Funktion an der Stelle* 190 *den Wert* $193.78 - 190 = 3.78$ *an. Abbildung 1.7 zeigt die empirische mittlere Exzess-Funktion* $e(u)$*. Die Entscheidung ist hier nicht einfach. Aber man kann sagen, dass die empirische MEF ab dem Wert 180 linear verläuft.*

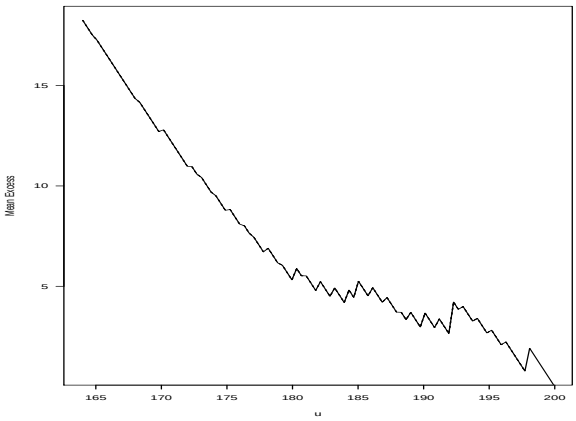


Abbildung 1.7: ME-Plot

Neben der empirische mittlere Exzess-Funktion sollte man die geschätzten Werte der Parameter in Abhängigkeit vom Schwellenwert u zeichnen. Ist

nämlich die verallgemeinerte Paretoverteilung ein geeignetes Modell für Beobachtungen, die größer als ein Schwellenwert u_0 sind, so ist sie auch ein geeignetes Modell für Schwellenwerte größer als u_0. Der Parameter ξ ändert sich nicht, während sich β ändert. Coles (2001) schlägt auf Seite 83 eine Reparametrisierung von β vor, die zu einem konstanten Wert führt. Man wird also den Wert u_0 wählen, ab dem die Parameterschätzer nahezu konstant sind.

Beispiel 6 (fortgesetzt) *Abbildung 1.8 zeigt die Parameterschätzer gegen den Schwellenwert.*

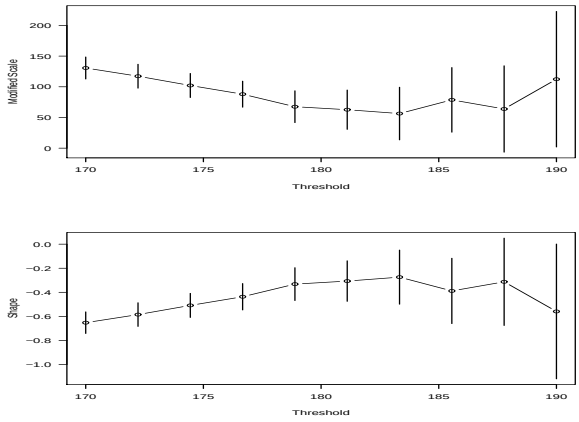


Abbildung 1.8: Parameterschätzer gegen den Schwellenwert

Wir sehen, dass die Schätzwerte ab dem Schwellenwert 180 *konstant sind.*

Hat man sich für einen Schwellenwert entschieden, so sollte man überprüfen, wie gut die Anpassung ist. Coles (2001) schlägt auf Seite 84 den **Wahrscheinlichkeits-Plot** und den **Quantil-Plot** vor. Sind $y_{(1)}, \ldots, y_{(k)}$ die geordneten Beobachtungen, die größer als u sind, so zeichnet man beim Wahrscheinlichkeits-Plot $\hat{H}(y_{(i)}$ gegen $i/(k+1)$. Dabei ist

$$\hat{H}(y) = 1 - \left(1 + \frac{\hat{\xi}\, y}{\hat{\beta}}\right)^{-1/\hat{\xi}}$$

Beim Quantil-Plot zeichnet man $y_{(i)}$ gegen $\hat{H}^{-1}(i/(k+1))$. Dabei ist

$$\hat{H}(y)^{-1} = u + \frac{\hat{\beta}}{\hat{\xi}}\left(y^{-\hat{\xi}} - 1\right)$$

Ist die verallgemeinerte Paretoverteilung ein geeignetes Modell zur Beschreibung der Beobachtungen, die größer als u sind, so sollte die Punktewolke auf einen linearen Zusammenhang hindeuten.
Außerdem sollte man noch das Histogramm der Exzesse mit der geschätzten Dichtefunktion zeichnen.

Beispiel 6 (fortgesetzt) *Abbildung 1.9 zeigt die Diagnostika.*

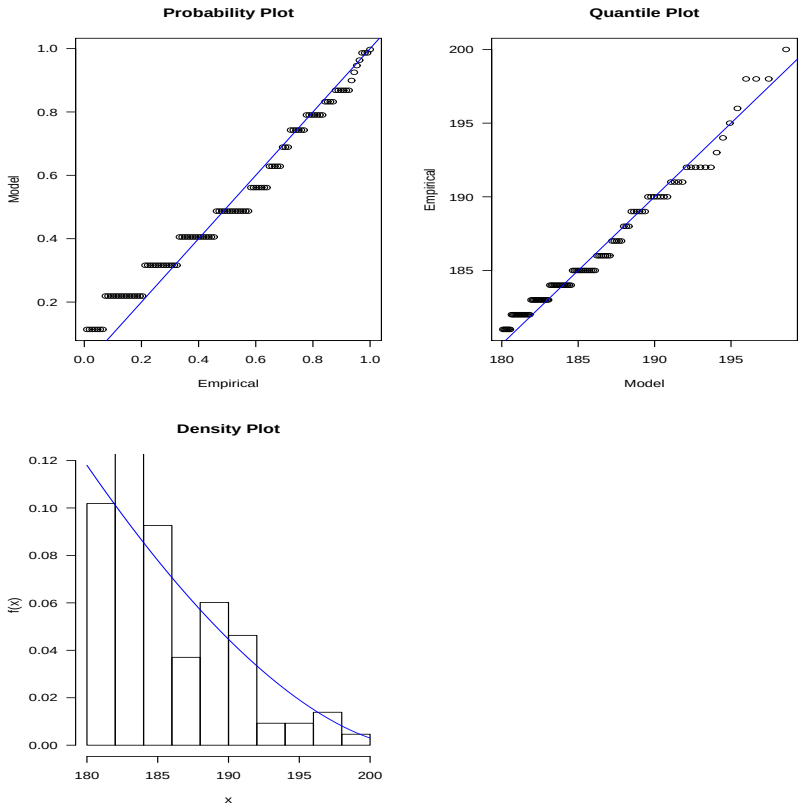


Abbildung 1.9: Diagnostika

Alle Zeichnungen deuten auf eine sehr gute Anpassung hin.

Kapitel 2

Symmetrie und Schiefe

2.1 Was ist Symmetrie?

Wir gehen aus von einer stetigen Zufallsvariablen X mit Dichtefunktion $f_X(x)$ und Verteilungsfunktion $F_X(X)$. Schauen wir uns exemplarisch die Dichtefunktion einer mit den Parametern μ und σ^2 normalverteilten Zufallsvariablen an. Es gilt

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.1}$$

Die Dichtefunktion nimmt für $(x-\mu)^2 = t^2$ mit $t \geq 0$ identische Werte an. Es gilt

$$(x-\mu)^2 = t^2 \quad \Longleftrightarrow \quad |x-\mu| = t \quad \Longleftrightarrow \quad x-\mu = \pm t$$

Somit sind die Werte der Dichtefunktion in $x = \mu - t$ und $x = \mu + t$ für $t \geq 0$ identisch. Abbildung 2.1 verdeutlicht dies exemplarisch für $t = 1$.

Definition 2.1.1 *Sei X eine stetige Zufallsvariable mit Dichtefunktion $f_X(x)$. Die Verteilung von X heißt* **symmetrisch** *bezüglich θ, wenn für alle $t \in \mathbb{R}$ gilt*

$$f_X(\theta - t) = f_X(\theta + t) \tag{2.2}$$

Die Gleichverteilung, die logistische Verteilung, die Laplaceverteilung und die Cauchyverteilung sind symmetrische Verteilungen. Abbildung 2.2 zeigt die Dichtefunktionen dieser Verteilungen.
Wir können Symmetrie auch über die Verteilungsfunktion definieren. Dies kann man sich an Abbildung 2.1 klarmachen. Die Fläche unterhalb von $\theta - t$ ist gleich der Fläche oberhalb von $\theta + t$. Dies führt zu folgender

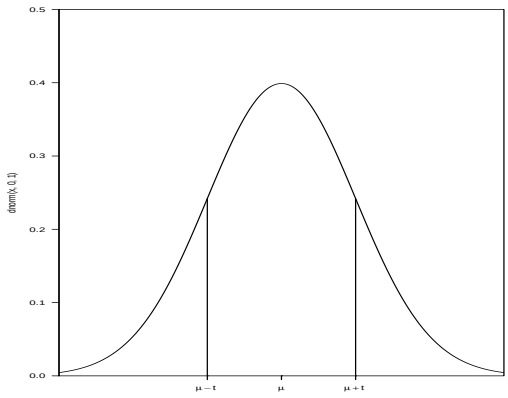


Abbildung 2.1: Die Dichtefunktion der Normalverteilung

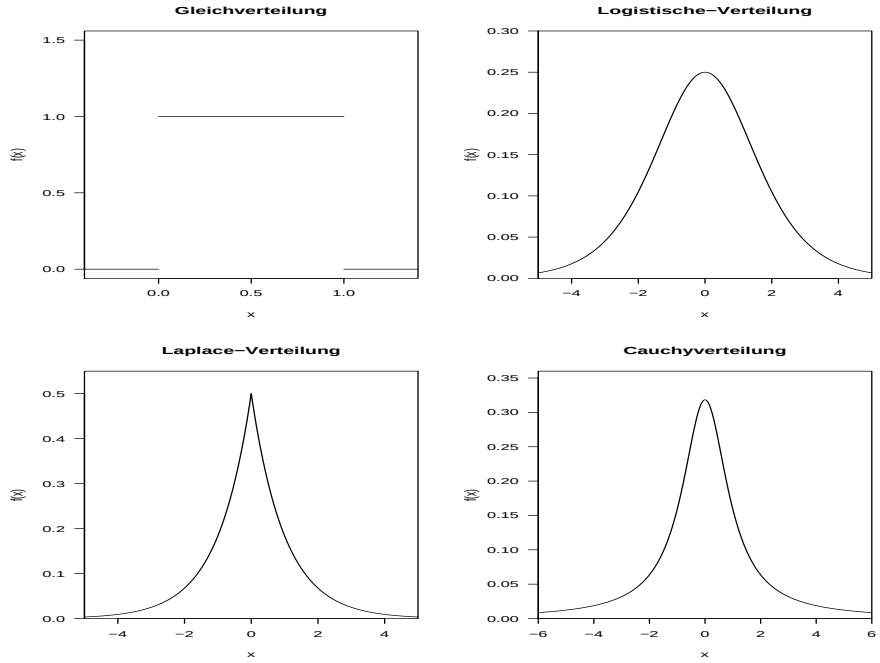


Abbildung 2.2: Symmetrische Verteilungen

Definition 2.1.2 *Sei X eine stetige Zufallsvariable mit Verteilungsfunktion $F_X(x)$. Die Verteilung von X heißt symmetrisch bezüglich θ, wenn für alle*

$t \in \mathbb{R}$ *gilt*

$$F_X(\theta - t) = 1 - F_X(\theta + t) \tag{2.3}$$

Da die Fläche unter der Dichtefunktion durch die Gerade, die durch θ parallel zur Ordinate verläuft, halbiert wird, gilt $\theta = x_{0.5}$. Abbildung 2.3 verdeutlicht, dass wir die Symmetrie über Quantile auch folgendermaßen definieren können:

Definition 2.1.3 *Sei X eine stetige Zufallsvariable mit Verteilungsfunktion $F_X(x)$. Die Verteilung von X heißt symmetrisch bezüglich θ, wenn für alle $p \in (0, 0.5)$ gilt*

$$x_{0.5} - x_p = x_{1-p} - x_{0.5} \tag{2.4}$$

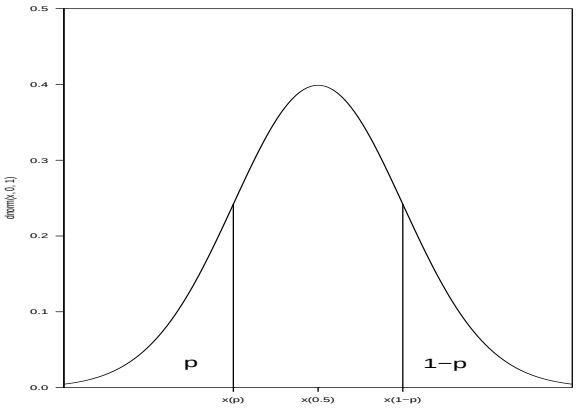


Abbildung 2.3: Definition von Symmetrie über Quantile

Eine Verteilung, die nicht symmetrisch ist, heißt **schief**. Ein Beispiel für eine schiefe Verteilung ist die Exponentialverteilung mit Parameter λ, deren Dichtefunktion folgendermaßen definiert ist:

$$f(x) = \begin{cases} \lambda\, e^{-\lambda\, x} & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

Die Grafik links oben in Abbildung 2.4 zeigt die Dichtefunktion der Exponentialverteilung mit $\lambda = 1$.
Man unterscheidet rechtsschiefe und linksschiefe Verteilungen.

Eine Verteilung heißt **rechtsschief** bzw. **linkssteil**, wenn für alle $p \in (0, 1)$ gilt

$$x_{0.5} - x_p < x_{1-p} - x_{0.5} \tag{2.5}$$

Entsprechend heißt eine Verteilung **linksschief** bzw. **rechtssteil**, wenn für alle $p \in (0, 1)$ gilt

$$x_{0.5} - x_p > x_{1-p} - x_{0.5} \tag{2.6}$$

Die Exponentialverteilung ist rechtsschief. Abbildung 2.4 zeigt neben der Dichtefunktion der Exponentialverteilung die Dichtefunktion der Gammaverteilung, der Lognormal-Verteilung und der Weibull-Verteilung.

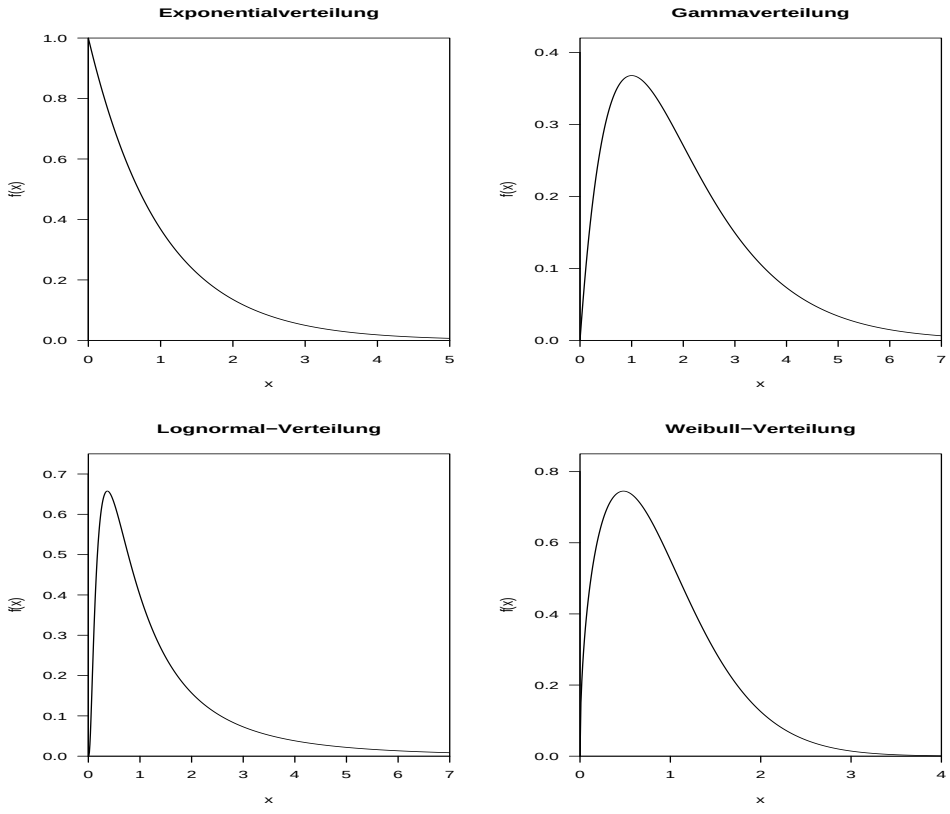


Abbildung 2.4: Schiefe Verteilungen

2.2 Wozu benötigt man Symmetrie?

Bei einer symmetrischen Verteilung ist der Lageparameter eindeutig. Man wählt das Symmetriezentrum. Existiert der Erwartungswert $E(X)$, so gilt bei einer symmetrischen Verteilung

$$E(X) = x_{0.5}$$

Bei einer schiefen Verteilung unterscheiden sich Erwartungswert und Median. Bei einer rechtsschiefen Verteilung gilt $x_{0.5} < E(X)$. Schauen wir uns exemplarisch die Exponentialverteilung an. Es gilt $x_{0.5} = \frac{1}{\lambda} \ln 2 = \frac{0.693}{\lambda}$. Wegen $E(X) = \frac{1}{\lambda}$ gilt $x_{0.5} < E(X)$.
Bei einer linksschiefen Verteilung gilt $x_{0.5} > E(X)$.
Will man also die Lage einer schiefen Verteilung beschreiben, so muss man angeben, welchen Lageparameter man benutzt.
Bei einigen statistischen Tests wird unterstellt, dass die Verteilung der Grundgesamtheit symmetrisch ist. Schauen wir uns exemplarisch den Wilcoxon-Vorzeichen-Rangtest an. Dieser ist ein Test auf einen Lageparameter θ im Einstichprobenproblem. Wir gehen aus von einer Zufallsstichprobe vom Umfang n aus einer Grundgesamtheit, deren Verteilung stetig und symmetrisch bezüglich θ ist. Es soll getestet werden

$$H_0 : \theta = \theta_0 \quad \text{gegen} \quad H_1 : \theta > \theta_0 \,. \tag{2.7}$$

Wir setzen im Folgenden θ gleich 0. Wollen wir auf einen Wert $\theta_0 \neq 0$ testen, so betrachten wir die Beobachtungen $x_1 - \theta_0, \ldots, x_n - \theta_0$.

Beispiel 7 *Gegeben seien die Beobachtungen*

$$x_1 = -0.8 \quad x_2 = -0.5 \quad x_3 = 0.4 \quad x_4 = 0.9 \quad x_5 = 1.2 \quad x_6 = 1.7$$

Der Wilcoxon-Vorzeichen-Rangtest verwendet zwei Informationen:

1. die Vorzeichen s_i der Beobachtungen
2. die Abstände $|x_i|$ der Beobachtungen vom Nullpunkt

Dabei ist

$$s_i = \begin{cases} 1 & \text{falls } x_i > 0 \\ 0 & \text{sonst} \end{cases}$$

Beispiel 7 (fortgesetzt) *Es gilt*

$$s_1 = 0 \quad s_2 = 0 \quad s_3 = 1 \quad s_4 = 1 \quad s_5 = 1 \quad s_6 = 1$$

und

$$|x_1| = 0.8 \quad |x_2| = 0.5 \quad |x_3| = 0.4 \quad |x_4| = 0.9 \quad |x_5| = 1.2 \quad |x_6| = 1.7$$

Beim Wilcoxon-Vorzeichen-Rangtest werden die Ränge R_i der $|x_i|$ betrachtet.

$$R_1 = 3 \quad R_2 = 2 \quad R_3 = 1 \quad R_4 = 4 \quad R_5 = 5 \quad R_6 = 6$$

Die Teststatistik des Wilcoxon-Vorzeichen-Rangtests ist gleich der Summe der Ränge der $|x_i|$.

$$W^+ = \sum_{i=1}^{n} s_i \, R_i$$

Unter welchen Bedingungen ist der Wilcoxon-Vorzeichen-Rangtest ein geeigneter Test für die Hypothesen in Gleichung (2.7)? Zur Beantwortung dieser Frage schauen wir uns eine Abbildung der Daten an.

Diese Stichprobe kann aus unterschiedlichen Grundgesamtheiten stammen. Zum einen kann die Verteilung der Grundgesamtheit symmetrisch sein. Dies ist in Abbildung 2.5 der Fall.

Kommen die Daten aus der Gleichverteilung, so ist der Wert des Lageparameters gleich 0.5. Somit ist H_1 erfüllt. Der Wilcoxon-Vorzeichen-Rangtest ist geeignet, diese Lagealternative aufzudecken, da die Mehrzahl der positiven Beobachtungen weiter vom Nullpunkt entfernt sind als die negativen Beobachtungen. Also sind auch die Ränge der positiven Bobachtungen größer. Dies führt zu einem großen Wert von W^+.

Ist die Verteilung der Grundgesamtheit aber schief, so ändert sich die Interpretation. Abbildung 2.6 zeigt eine schiefe Verteilung, die ein für die Daten angemessenes Modell darstellt.

Wir sehen, dass der Median dieser Verteilung gleich 0 ist. Auf Grund der Konstellation der Beobachtungen würde der Wilcoxon-Vorzeichen-Rangtest die Hypothese H_0 aber ablehnen. Ist die Verteilung der Grundgesamtheit also schief, so muss eine andere Hypothese betrachtet werden, für die der

Abbildung 2.5: Daten mit Gleichverteilung

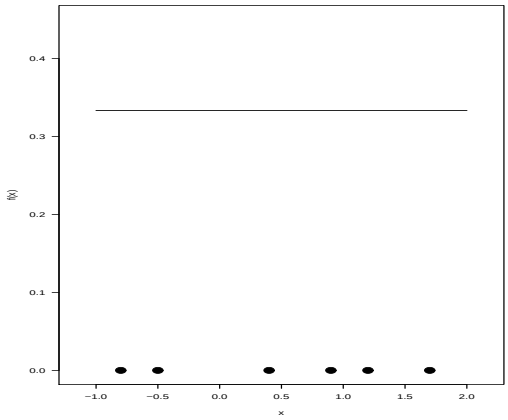
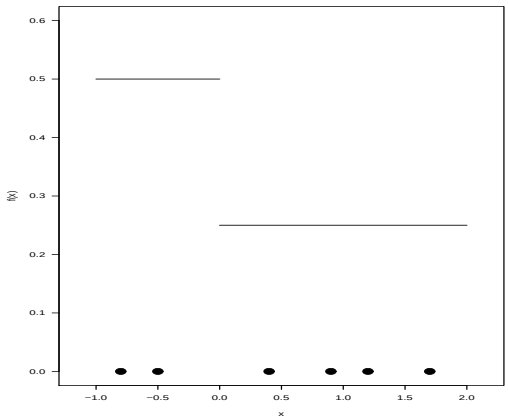


Abbildung 2.6: Daten mit schiefer Verteilung



Wilcoxon-Vorzeichen-Rangtest geeignet ist. Er ist in diesem Fall ein Test auf Symmetrie bezüglich eines bekannten Symmetriezentrums.
Um den Wilcoxon-Vorzeichen-Rangtest als Test auf einen Lageparameter auffassen zu können, benötigen wir also die Symmetrie der Verteilung der Grundgesamtheit.

2.3 Maßzahlen für die Schiefe einer Verteilung

Die klassischen Maßzahlen für die Lage und die Variabilität einer Verteilung sind der Erwartungswert $E(X)$ und die Varianz $Var(X) = E(X^2) - E(X)^2$. Die dabei auftretenden Größen $E(X)$ und $E(X^2)$ sind spezielle Momente der Zufallsvariablen X.

Definition 2.3.1 *Sei X eine Zufallsvariable. Dann heißt*

$$\mu_r = E(X^r) \tag{2.8}$$

r-tes Moment von X.

Neben den Momenten sind noch die zentralen Momente von Interesse.

Definition 2.3.2 *Sei X eine Zufallsvariable mit Erwartungswert μ. Dann heißt*

$$\mu'_r = E\left[(X - E(X))^r\right] \tag{2.9}$$

r-tes zentrales Moment von X.

Offensichtlich ist das erste zentrale Moment gleich Null und das zweite zentrale Moment gleich der Varianz.
Höhere zentrale Moment beschreiben bestimmte Charakteristika einer Verteilung. So ist μ'_3 bei einer symmetrischen Verteilung gleich 0. Für eine bezüglich μ symmetrische Verteilung gilt nämlich für alle $t \in \mathbb{R}$:

$$f_X(\mu - t) = f_X(\mu + t) \tag{2.10}$$

Somit gilt

$$\begin{aligned}
\mu'_3 &= \int_{-\infty}^{\infty} (x-\mu)^3 f_X(x)\, dx \\
&= \int_{-\infty}^{\mu} (x-\mu)^3 f_X(x)\, dx + \int_{\mu}^{\infty} (x-\mu)^3 f_X(x)\, dx \\
&\overset{y=x-\mu}{=} \int_{-\infty}^{0} y^3 f_X(\mu+x)\, dy + \int_{0}^{\infty} y^3 f_X(\mu+x)\, dx
\end{aligned}$$

Wir substituieren $y = -t$ im ersten Summanden und erhalten

$$\begin{aligned}
\mu_3' &= -\int_0^\infty t^3 \, f_X(\mu - t)\, dt + \int_0^\infty y^3 \, f_X(x + \mu)\, dx \\
&\overset{(2.10)}{=} -\int_0^\infty y^3 \, f_X(\mu + y)\, dy + \int_0^\infty y^3 \, f_X(x + \mu)\, dx \\
&= 0
\end{aligned}$$

Ist also $\mu_3' \neq 0$, so ist die Verteilung schief. Auf Grund dieser Eigenschaft kann man μ_3' als Maßzahl für die Schiefe auffassen. Diese hat jedoch den Nachteil, dass sie von der Skalierung der Daten abhängt. Es gilt nämlich

$$\begin{aligned}
E[(a\,X - E(a\,X))^3] &= E[(a\,X - a\,E(X))^3] = E[(a(X - E(X)))^3] \\
&= E[a^3(X - E(X))^3] = a^3\, E[(X - E(X))^3]
\end{aligned}$$

Die folgende Definition gibt eine auf μ_3' basierende Maßzahl für die Schiefe an, die skalenunabhängig ist.

Definition 2.3.3 *Die Schiefe einer Zufallsvariablen X ist definiert durch*

$$\gamma_1 = \frac{E((X - \mu)^3)}{\sigma^3}. \tag{2.11}$$

In Tabelle 2.1 sind die Werte von γ_1 für ausgewählte schiefe Verteilungen zu finden. Positive Werte von γ_1 sprechen für eine rechtsschiefe Verteilung,

Verteilung	γ_1	QS	τ_3
Exponentialverteilung	2.00	0.26	0.33
Gammaverteilung (r=2)	0.71	0.17	0.16
Lognormalverteilung	6.19	0.33	0.46

Tabelle 2.1: Werte von γ_1 für ausgewählte schiefe Verteilungen

während negative Werte auf eine linksschiefe Verteilung hindeuten. Ist γ_1 gleich 0, so ist die Verteilung nicht notwendigerweise symmetrisch.
Die Maßzahl γ_1 hat einige Nachteile. Sie ist schwer zu interpretieren. Außerdem muss sie nicht existieren. Dies ist bei der Cauchyverteilung der Fall. Da der Erwartungswert der Cauchyverteilung nicht existiert, existiert auch nicht γ_1.

Aus einer Stichprobe $x_1, \ldots, x_n$ schätzt man γ_1 durch

$$\hat{\gamma}_1 = \frac{\hat{\mu}'_3}{\hat{\mu}'^{1.5}_2}. \tag{2.12}$$

Dabei ist

$$\hat{\mu}'_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^r \tag{2.13}$$

Beispiel 8 *Studenten wurden in einer Vorlesung gefragt, wie viele CDs sie besitzen. Hier sind die Daten von* 10 *Studierenden:*

```
10 20 30 40 60 70 90 150 200 300
```

Es gilt $\hat{\gamma}_1 = 1.15$.

Wie das folgende Beispiel zeigt, ist $\hat{\gamma}_1$ nicht robust. Ein Ausreißer hat einen starken Einfluss auf den Wert von $\hat{\gamma}_1$.

Beispiel 9 *Für den Datensatz*

```
-3 -1 0 1 3
```

ist $\hat{\gamma}_1$ *gleich* 0.
Nimmt die fünfte Beobachtung den Wert 8 *an, so erhalten wir* $\hat{\gamma}_1 = 1.03$.

Royston (1992) zeigt mit Simulationsstudien, dass $\hat{\gamma}_1$ für kleine Stichprobenumfänge einen großen Bias besitzt.
Da $\hat{\gamma}_1$ viele Nachteile besitzt, sollte man eine andere Maßzahl für die Schiefe berechnen. Einige dieser Maßzahlen basieren auf Quantilen. Ausgangspunkt ist Gleichung (2.4) auf Seite 33. Subtrahieren wir bei dieser Gleichung auf beiden Seiten $x_{0.5} - x_p$, so erhalten wir eine Größe, die als Maßzahl für die Schiefe aufgefasst werden kann:

$$(x_{1-p} - x_{0.5}) - (x_{0.5} - x_p) \tag{2.14}$$

Bowley (1920) setzt $p = 0.25$ und betrachtet

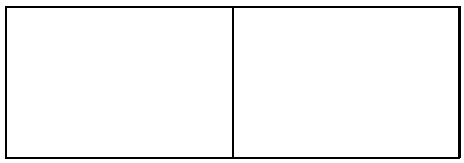
$$QS = \frac{x_{0.75} - x_{0.5} - (x_{0.5} - x_{0.25})}{x_{0.75} - x_{0.25}} = \frac{x_{0.75} + x_{0.25} - 2x_{0.5}}{x_{0.75} - x_{0.25}} \tag{2.15}$$

Die Division durch den Interquartilsabstand bewirkt, dass QS unabhängig von der Skalierung der Daten ist. Offensichtlich liegt QS zwischen -1 und 1. Werte von QS für ausgewählte Verteilungen sind in Tabelle 2.1 auf Seite 39

zu finden. Man erhält Schätzer für QS und die anderen quantilbasierten Schiefemaße, indem man die Quantile schätzt und in die Formeln einsetzt. Man schätzt QS also, indem man $\hat{x}_{0.25}$, $\hat{x}_{0.5}$ und $\hat{x}_{0.75}$ schätzt und in die Formel (2.15) einsetzt:

$$\widehat{QS} = \frac{\hat{x}_{0.75} - \hat{x}_{0.5} - (\hat{x}_{0.5} - \hat{x}_{0.25})}{\hat{x}_{0.75} - \hat{x}_{0.25}} = \frac{\hat{x}_{0.75} + \hat{x}_{0.25} - 2\hat{x}_{0.5}}{\hat{x}_{0.75} - \hat{x}_{0.25}} \tag{2.16}$$

Beim Boxplot zeichnet man ein Rechteck, das vom unteren Quartil $x_{0.25}$ bis zum oberen Quartil $x_{0.75}$ verläuft. Den Median markiert man im Rechteck als vertikale Linie. Der Median teilt das Rechteck in zwei kleinere Rechtecke. Der Nenner der Maßzahl von Bowley ist gleich der Länge des großen Rechtecks, während der Zähler gleich der Differenz aus der Länge des rechten Rechtecks und der Länge des linken Rechtecks ist. Bei einer symmetrischen Verteilung teilt der Median das Rechteck in zwei gleich große Rechtecke.



In diesem Fall ist QS gleich 0.
Bei einer rechtsschiefen Verteilung ist das linke Rechteck kleiner als das rechte.

In diesem Fall ist QS größer als 0.
Bei einer linksschiefen Verteilung ist das linke Rechteck größer als das rechte.

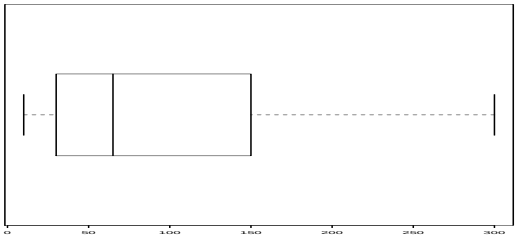
In diesem Fall ist QS kleiner als 0.

Beispiel 8 (fortgesetzt) *Wir bestimmen $x_{0.25}$ und $x_{0.75}$ so, wie es Tukey vorgeschlagen hat. Es gilt $x_{0.25} = 30$ und $x_{0.75} = 150$. Außerdem gilt $x_{0.5} = 65$. Somit gilt*

$$QS = \frac{150 - 65 - (65 - 30)}{150 - 30} = 0.42$$

Abbildung 2.7 zeigt den Boxplot.

Abbildung 2.7: Boxplot der Anzahl Cds



Hinley (1975) hat diesen Schätzer verallgemeinert

$$H = \frac{x_{1-p} - x_{0.5} - (x_{0.5} - x_p)}{x_{1-p} - x_p} \tag{2.17}$$

Brys, Hubert, and Struyf (2003) setzen $p = 0.125$.

$$H = \frac{x_{0.875} - x_{0.5} - (x_{0.5} - x_{0.125})}{x_{0.875} - x_{0.125}} \tag{2.18}$$

Handl (1985) betrachtet nicht die Differenz sondern den Quotienten aus $x_{1-p} - x_{0.5}$ und $x_{0.5} - x_p$:

$$Q = \frac{x_{1-p} - x_{0.5}}{x_{0.5} - x_p} \tag{2.19}$$

Sinnvolle Werte für p sind hier 0.25 oder 0.125.
Hosking (1990) wählte einen anderen Zugang zu einer Maßzahl für die Schiefe. Er geht aus von den geordneten Beobachtungen einer Stichprobe vom Umfang n, die er mit $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$ bezeichnet. Er definiert für $r = 1, 2, \ldots$ sogenannte L-Momente

$$\lambda_r = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E\left(X_{r-k:r}\right) \tag{2.20}$$

Für $r = 1, 2, 3$ gilt

$$\lambda_1 = E(X_{1:1}) \tag{2.21}$$

$$\lambda_2 = \frac{1}{2}\left[E(X_{2:2}) - E(X_{1:2})\right] \tag{2.22}$$

$$\lambda_3 = \frac{1}{3}\left[E(X_{3:3}) - 2\,E(X_{2:3}) + E(X_{1:3})\right] \tag{2.23}$$

Man kann λ_3 folgendermaßen umformen:

$$\begin{aligned}
\lambda_3 &= \frac{1}{3}\left[E(X_{3:3}) - 2\,E(X_{2:3}) + E(X_{1:3})\right] \\
&= \frac{1}{3}\left[E(X_{3:3}) - E(X_{2:3}) - E(X_{2:3}) + E(X_{1:3})\right] \\
&= \frac{1}{3}\left[E(X_{3:3}) - E(X_{2:3}) - (E(X_{2:3}) - E(X_{1:3}))\right])
\end{aligned}$$

Somit ist λ_3 bei einem Stichprobenumfang von $n = 3$ gleich der Differenz aus der erwarteten Distanz zwischen Maximum und Median und der erwarteten Distanz aus Median und Minimum. Bei einer symmetrischen Verteilung nimmt diese Differenz gleich den Wert 0 an, während sie bei einer schiefen Verteilung ungleich 0 ist. Ein Schätzer von λ_3 ist

$$l_3 = \frac{1}{3}\left[x_{3:3} - 2x_{2:3} + x_{1:3}\right] \tag{2.24}$$

Um λ_3 auf Basis einer Zufallsstichprobe $x_1, \ldots, x_n$ zu schätzen, bestimmt er für jede geordnete Teilstichprobe $x_{(i)}, x_{(j)}, x_{(k)}$ aus $x_1, \ldots, x_n$ den Schätzwert

$$\frac{1}{3}\left[x_{(3)} - 2x_{(2)} + x_{(1)}\right] \tag{2.25}$$

Als Schätzer l_3 für λ_3 auf Basis der n Beobachtungen dient dann der Mittelwert der Schätzwerte der Teilstichproben:

$$l_3 = \frac{1}{3}\binom{n}{3}^{-1} \sum_{i<j<k} (x_{(k)} - 2x_{(j)} + x_{(i)}) \tag{2.26}$$

Beispiel 8 (fortgesetzt) *Wir betrachten aus Gründen der Übersichtlichkeit die letzten* 5 *Beobachtungen*

```
70  90 150 200 300
```

Mit $\binom{5}{3} = 10$ *gilt*

$$\begin{aligned} l_3 &= \frac{1}{30}[(150 - 2\cdot 90 + 70) + (200 - 2\cdot 90 + 70) + (300 - 2\cdot 90 + 70) \\ &+ (200 - 2\cdot 150 + 70) + (300 - 2\cdot 150 + 70) + (300 - 2\cdot 200 + 70) \\ &+ (200 - 2\cdot 150 + 90) + (300 - 2\cdot 150 + 90) + (300 - 2\cdot 200 + 90) \\ &+ (300 - 2\cdot 200 + 150)] \\ &= 15 \end{aligned}$$

Hosking (1990) betrachtet folgende Maßzahl für die Schiefe

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \tag{2.27}$$

Einen Schätzer für τ_3 erhält man, indem man λ_3 und λ_2 schätzt:

$$\hat{\tau}_3 = \frac{l_3}{l_2} \tag{2.28}$$

Den Schätzer von λ_3 kennen wir bereits. Wenn wir beim Schätzer l_2 von λ_2 das gleiche Prinzip anwenden, erhalten wir

$$l_2 = \frac{1}{2}\binom{n}{2}^{-1} \sum_{i<j} (x_{(j)} - x_{(i)}) \tag{2.29}$$

Beispiel 8 (fortgesetzt) *Mit* $\binom{5}{2} = 10$ *gilt*

$$\begin{aligned} l_2 &= \frac{1}{20}[90 - 70 + 150 - 70 + 200 - 70 + 300 - 70 + 150 - 90 \\ &+ 200 - 90 + 300 - 90 + 200 - 150 + 300 - 150 + 300 - 200] \\ &= 57 \end{aligned}$$

Somit gilt

$$\hat{\tau}_3 = \frac{l_3}{l_2} = \frac{15}{57} = 0.26$$

Hosking (1990) zeigt, dass man l_2 und l_3 folgendermaßen in Abhängigkeit von den geordneten Beobachtungen $x_{(1)}, \ldots, x_{(n)}$ und $\bar{x}$ darstellen kann:

$$\begin{aligned} l_2 &= 2w_2 - \bar{x} \\ l_3 &= 6w_3 - 6w_2 + \bar{x} \end{aligned}$$

mit

$$w_2 = \frac{1}{n(n-1)} \sum_{i=2}^{n} (i-1)\, x_{(i)}$$

und

$$w_3 = \frac{1}{n(n-1)(n-2)} \sum_{i=2}^{n} (i-1)(i-2)\, x_{(i)}$$

Beispiel 8 (fortgesetzt) *Wir betrachten wieder die letzten* 5 *Beobachtungen*

```
70  90 150 200 300
```

Es gilt $\bar{x} = 162$. *Außerdem gilt*

$$w_2 = \frac{1}{5 \cdot 4} \left[90 + 2 \cdot 150 + 3 \cdot 200 + 4 \cdot 300\right] = 109.5$$

und

$$w_3 = \frac{1}{5 \cdot 4 \cdot 3} \left[2 \cdot 150 + 6 \cdot 200 + 12 \cdot 300\right] = 85$$

Also gilt

$$\begin{aligned} l_2 &= 2 \cdot 109.5 - 162 = 57 \\ l_3 &= 6 \cdot 85 - 6 \cdot 109.5 + 162 = 15 \end{aligned}$$

2.4 Ein Test auf Symmetrie

Die von Hosking (1990) vorgeschlagene Maßzahl t_3 für die Schiefe einer Verteilung beruht für eine geordnete Stichprobe $x_{(1)}, x_{(2)}, x_{(3)}$ vom Umfang 3 auf folgender Größe

$$x_{(1)} + x_{(3)} - 2\,x_{(2)}$$

Bereits 1980 wurde diese Größe von Randles, Fligner, Policello, and Wolfe (1980) benutzt, um einen Test auf Symmetrie zu konstruieren. Sie betrachten also folgende Hypothesen

H_0 : Die Verteilung der Grundgesamtheit ist symmetrisch
H_1 : Die Verteilung der Grundgesamtheit ist schief

Randles, Fligner, Policello, and Wolfe (1980) nennen das Tripel $x_{(1)}, x_{(2)}, x_{(3)}$ ein **rechtes Tripel**, wenn gilt

$$x_{(1)} + x_{(3)} - 2\,x_{(2)} > 0$$

Für ein rechtes Tripel gilt also

$$x_{(3)} - x_{(2)} > x_{(2)} - x_{(1)}$$

Die folgende Abbildung veranschaulicht den Sachverhalt:

Ein Tripel $x_{(1)}, x_{(2)}, x_{(3)}$ mit

$$x_{(1)} + x_{(3)} - 2\,x_{(2)} < 0$$

heißt **linkes Tripel**. Für ein linkes Tripel gilt also

$$x_{(3)} - x_{(2)} < x_{(2)} - x_{(1)}$$

Die folgende Abbildung veranschaulicht den Sachverhalt:

Um dieses Konzept auf eine Stichprobe $x_1, \ldots, x_n$ zu übertragen, betrachten Randles, Fligner, Policello, and Wolfe (1980) für $1 \leq i < j < k \leq n$ alle geordneten Stichproben $x_{(i)}, x_{(j)}, x_{(k)}$ aus der Stichprobe $x_1, \ldots, x_n$ und bestimmen für jede dieser Stichproben den Wert der folgenden Funktion:

$$f^*(x_{(i)}, x_{(j)}, x_{(k)}) = \begin{cases} 1 & \text{für } x_{(i)} + x_{(k)} - 2\,x_{(j)} > 0 \\ -1 & \text{für } x_{(i)} + x_{(k)} - 2\,x_{(j)} < 0 \\ 0 & \text{sonst} \end{cases}$$

Beispiel 9 *Studierende wurden in einer Vorlesung gefragt, wie viele CDs sie besitzen. Hier sind die Daten von* 5 *Studierenden:*

```
30  40  60  100 150
```

Es gilt

$$\begin{aligned} f^*(x_{(1)}, x_{(2)}, x_{(3)}) &= 1 & f^*(x_{(1)}, x_{(2)}, x_{(4)}) &= 1 \\ f^*(x_{(1)}, x_{(2)}, x_{(5)}) &= 1 & f^*(x_{(1)}, x_{(3)}, x_{(4)}) &= 1 \\ f^*(x_{(1)}, x_{(3)}, x_{(5)}) &= 1 & f^*(x_{(1)}, x_{(4)}, x_{(5)}) &= -1 \\ f^*(x_{(2)}, x_{(3)}, x_{(4)}) &= 1 & f^*(x_{(2)}, x_{(3)}, x_{(5)}) &= 1 \\ f^*(x_{(2)}, x_{(4)}, x_{(5)}) &= -1 & f^*(x_{(3)}, x_{(4)}, x_{(5)}) &= 1 \end{aligned}$$

Die Teststatistik des Tests von Randles, Fligner, Policello, and Wolfe (1980) ist

$$T = \sum_{1 \leq i < j < k \leq n} f^*(x_{(i)}, x_{(j)}, x_{(k)}) \tag{2.30}$$

Dies ist die Differenz aus der Anzahl der rechten Tripel und der Anzahl der linken Tripel in allen geordneten Stichproben vom Umfang 3 aus der Stichprobe $x_1, \ldots, x_n$.

Beispiel 9 (fortgesetzt) *Es gilt* $T = 6$.

Ist die Anzahl der rechten Tripel und linken Tripel ungefähr gleich, so deutet dies auf Symmetrie hin. In diesem Fall nimmt T einen Wert in der Nähe von 0 an. Gibt es in der Stichprobe aber viel mehr rechte als linke Tripel, so spricht dies für eine rechtsschiefe Verteilung. In diesem Fall nimmt T einen großen Wert an. Ein kleiner Wert von T ist ein Indikator für eine linksschiefe Verteilung. Wir lehnen also H_0 ab, wenn T zu groß oder zu klein ist.

Zur Testentscheidung benötigen wir die Verteilung von T, wenn die Nullhypothese der Symmetrie der Verteilung zutrifft. Randles, Fligner, Policello, and Wolfe (1980) zeigen, dass

$$\frac{T - E(T)}{\sqrt{Var(T)}}$$

approximativ standardnormalverteilt ist, wenn H_0 zutrifft. Wir lehnen H_0 zum Signifikanzniveau α ab, wenn gilt

$$\left| \frac{T - E(T)}{\sqrt{Var(T)}} \right| \geq z_{1-\alpha/2}$$

Dabei ist $z_{1-\alpha/2}$ das $1 - \alpha/2$-Quantil der Standardnormalverteilung.
Trifft H_0 zu, so gilt

$$E(T) = 0$$

und

$$\begin{aligned} Var(T) &= \frac{(n-3)(n-4)}{(n-1)(n-2)} \sum_{t=1}^{n} B_t^2 + \frac{n-3}{n-4} \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} B_{s,t}^2 \\ &\quad + \frac{n(n-1)(n-2)}{6} - \left[1 - \frac{(n-3)(n-4)(n-5)}{n(n-1)(n-2)} \right] T^2 \end{aligned}$$

Der Beweis ist bei Randles, Fligner, Policello, and Wolfe (1980) zu finden. Dabei ist B_t gleich der Differenz aus der Anzahl der rechten Tripel und der Anzahl der linken Tripel, in denen $x_{(t)}$ vorkommt, und $B_{s,t}$ gleich der Differenz aus der Anzahl der rechten Tripel und der Anzahl der linken Tripel, in denen $x_{(s)}$ und $x_{(t)}$ vorkommen.

Beispiel 9 (fortgesetzt) *Es gilt*

$$B_1 = 4 \quad B_2 = 4 \quad B_3 = 6 \quad B_4 = 2 \quad B_5 = 2\,.$$

Außerdem gilt

$$\begin{aligned} &B_{1,2} = 3 \quad B_{1,3} = 3 \quad B_{1,4} = 1 \quad B_{1,5} = 1 \quad B_{2,3} = 3 \\ &B_{2,4} = 1 \quad B_{2,5} = 1 \quad B_{3,4} = 3 \quad B_{3,5} = 3 \quad B_{4,5} = -1 \end{aligned}$$

Also gilt

$$\sum_{t=1}^{5} B_t^2 = 4^2 + 4^2 + 6^2 + 2^2 + 2^2 = 76$$

und

$$\sum_{s=1}^{4}\sum_{t=s+1}^{5} B_{s,t}^2 = 3^2+3^2+1^2+1^2+3^2+1^2+1^2+3^2+3^2+(-1)^2 = 50$$

Somit gilt

$$Var(T) \;=\; \frac{2\cdot 1}{4\cdot 3}\cdot 76 + \frac{2}{1}\cdot 50 + \frac{5\cdot 4\cdot 3}{6} - \left[1-\frac{2\cdot 1\cdot 0}{4\cdot 3\cdot 2}\right]6^2 = 86.67$$

Wir bestimme den Wert der Teststatistik

$$T = \frac{6}{\sqrt{86.67}} = 0.64$$

Wegen $z_{0.975} = 1.96$ *lehnen wir* H_0 *zum Niveau* 0.05 *nicht ab.*

Beispiel 10 *Studierende wurden in einer Vorlesung gefragt, wie viele CDs sie besitzen. Hier sind die Daten von* 10 *Studierenden:*

```
10 20 30 40 60 70 90 150 200 300
```

Wir testen auf Symmetrie. Es gilt $T = 58$ *und* $Var(T) = 786.83$*. Somit gilt*

$$T = \frac{58}{\sqrt{786.83}} = 2.07$$

Wegen $z_{0.975} = 1.96$ *lehnen wir* H_0 *zum Niveau* 0.05 *ab.*

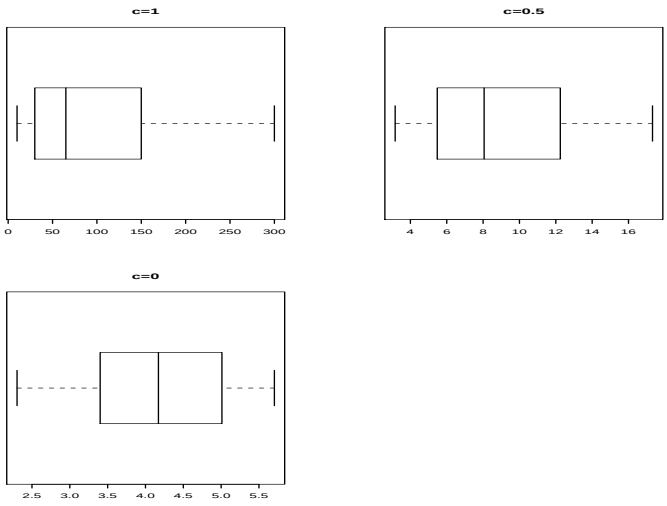
2.5 Transformation auf Symmetrie

Ist die Verteilung der Grundgesamtheit symmetrisch, so können Parameter leichter interpretiert werden. Deshalb versucht man die Daten so zu transformieren, dass die Verteilung symmetrisch ist. Da lineare Transformationen die Schiefe einer Verteilung nicht verändern, betrachtet man nichtlineare Transformationen. Am einfachsten zu interpretieren ist die Power-Transformation:

$$g(x) = \begin{cases} x^c & \text{für } c > 0 \\ \ln x & \text{für } c = 0 \\ -x^c & \text{für } c < 0 \end{cases} \tag{2.31}$$

Das negative Vorzeichen für $c < 0$ ist notwendig, da sonst die Ordnung der Daten zerstört würde.

Abbildung 2.8: Boxplots der Originaldaten und der transformierten Daten



Beispiel 10 (fortgesetzt) *Abbildung 2.8 zeigt die Boxplots für* $c = 1, 0.5, 0$. *Wir sehen, dass die Verteilung durch Logarithmieren der Daten nahezu symmetrisch ist. Aber auch Verteilung der Quadratwurzel der Daten sieht symmetrisch aus.*

Emerson and Stoto (1982) zeigen, wie man den Transformationsparameter systematisch schätzen kann. Sie gehen von der bekannten Gleichung

$$x_{0.5} - x_p = x_{1-p} - x_{0.5}$$

aus, die für alle $p \in (0, 0.5)$ gilt, wenn die Verteilung symmetrisch ist. Man kann diese Gleichung umformen zu

$$\frac{x_p + x_{1-p}}{2} = x_{0.5}$$

Transformiert man die Beobachtungen $x_1, \ldots, x_n$ mit einem Wert von c , so muss bei Symmetrie für alle $p \in (0, 0.5)$ gelten:

$$\frac{x_p^c + x_{1-p}^c}{2} = x_{0.5}^c \tag{2.32}$$

bzw.

$$\frac{\ln x_p + \ln x_{1-p}}{2} = \ln x_{0.5} \tag{2.33}$$

Man kann c schätzen, indem man einen Wert für p vorgibt und den Wert von c wählt, für den die Gleichung (2.32) bzw. (2.33) erfüllt ist.

Beispiel 10 (fortgesetzt) *Setzt man zum Beispiel $p = 0.25$, so erhält man folgende Schätzer*

$$x_{0.25} = 30 \qquad x_{0.5} = 65 \qquad x_{0.25} = 150$$

Wir suchen also den Wert von c, für die folgende Gleichung erfüllt

$$\frac{30^c + 150^c}{2} = 65^c$$

Da die Bestimmung der Nullstelle der Gleichung (2.32) nicht einfach ist, wählen Emerson and Stoto (1982) ein approximatives Verfahren. Sie entwickeln die Funktionen x_p^c und x_{1-p}^c in eine Taylorreihe um den Median $x_{0.5}$ und brechen sie nach dem quadratischen Glied ab. Sie bilden also

$$x_p^c \approx x_{0.5}^c + c\, x_{0.5}^{c-1} (x_p - x_{0.5}) + \frac{c\,(c-1)}{2}\, x_{0.5}^{c-2} (x_p - x_{0.5})^2 \tag{2.34}$$

und

$$x_{1-p}^c \approx x_{0.5}^c + c\, x_{0.5}^{c-1} (x_{1-p} - x_{0.5}) + \frac{c\,(c-1)}{2}\, x_{0.5}^{c-2} (x_{1-p} - x_{0.5})^2 \tag{2.35}$$

Setzen wir die Gleichungen (2.34) und (2.35) in die die Gleichung (2.32) ein, so erhalten wir

$$\begin{aligned} x_{0.5}^c &= x_{0.5}^c + \frac{c}{2}\, x_{0.5}^{c-1}\, (x_p + x_{1-p} - 2x_{0.5}) \\ &\quad + \frac{c\,(c-1)}{4}\, x_{0.5}^{c-2} \left[(x_p - x_{0.5})^2 + (x_{1-p} - x_{0.5})^2 \right] \end{aligned}$$

Subtrahieren wir $x_{0.5}^c$ von beiden seiten dieser Gleichung, so erhalten wir die äquivalente Gleichung

$$\frac{c}{2}\, x_{0.5}^{c-1}\, (x_p + x_{1-p} - 2x_{0.5}) + \frac{c\,(c-1)}{4}\, x_{0.5}^{c-2} \left[(x_p - x_{0.5})^2 + (x_{1-p} - x_{0.5})^2 \right] = 0$$

Wir klammern aus der Summe auf der linken Seite der Gleichung den Ausdruck $\frac{c}{2}\, x_{0.5}^{c-2}$ aus

$$\frac{c}{2}\, x_{0.5}^{c-2} \left(x_{0.5}\, (x_p + x_{1-p} - 2x_{0.5}) + \frac{c-1}{2}\, \left[(x_p - x_{0.5})^2 + (x_{1-p} - x_{0.5})^2 \right] \right) = 0$$

Dividieren wir beide Seiten dieser Gleichung durch $\frac{c}{2}\, x_{0.5}^{c-2}$, so erhalten wir

$$x_{0.5}\, (x_p + x_{1-p} - 2x_{0.5}) + \frac{c-1}{2}\, \left[(x_p - x_{0.5})^2 + (x_{1-p} - x_{0.5})^2 \right] = 0$$

Wir dividieren beide Seiten dieser Gleichung durch $2\,x_{0.5}$, so erhalten wir

$$\frac{x_p + x_{1-p}}{2} - x_{0.5} + (c-1)\,\frac{\left[(x_p - x_{0.5})^2 + (x_{1-p} - x_{0.5})^2\right]}{4\,x_{0.5}} = 0$$

Nun lösen wir diese Gleichung nach $\frac{x_p+x_{1-p}}{2} - x_{0.5}$ auf:

$$\frac{x_p + x_{1-p}}{2} - x_{0.5} = (1-c)\,\frac{(x_p - x_{0.5})^2 + (x_{1-p} - x_{0.5})^2}{4\,x_{0.5}} \tag{2.36}$$

Es gibt nun zwei Möglichkeiten, mit Gleichung (2.36) c zu bestimmen. Man kann die Gleichung nach c auflösen:

$$c = 1 - \frac{2\,x_{0.5}\,(x_p + x_{1-p} - 2\,x_{0.5})}{(x_p - x_{0.5})^2 + (x_{1-p} - x_{0.5})^2} \tag{2.37}$$

Dann gibt man einen Wert von p vor, schätzt die Quantile x_p, $x_{0.5}$ und x_{1-p} und setzt diese in die Gleichung (2.37) ein.

Beispiel 10 (fortgesetzt) *Wir setzen $p = 0.25$ und erhalten die Schätzer*

$$x_{0.25} = 30 \qquad x_{0.5} = 65 \qquad x_{0.25} = 150$$

Setzen wir dies in Gleichung (2.37) *ein, so erhalten wir*

$$c = 1 - \frac{2 \cdot 65\,(30 + 150 - 2 \cdot 65)}{(30 - 65)^2 + (150 - 65)^2} = 0.23$$

Dies spricht dafür, dass man die Daten logarithmieren sollte.

Emerson and Stoto (1982) schätzen für $p = 1/2, 1/4, 1/8, 1/16, \ldots$ die Quantile x_p und x_{1-p}. Für diese geschätzten Werte kann man die Gleichung (2.36) als Regressionsgleichung in der erklärenden Variablen $\frac{(x_p-x_{0.5})^2+(x_{1-p}-x_{0.5})^2}{4\,x_{0.5}}$ und der zu erklärenden Variablen $\frac{x_p+x_{1-p}}{2} - x_{0.5}$ auffassen. Man kann durch die Punktewolke die Regressionsgerade legen und erhält so einen Schätzer für $1-c$ und auch für c. Hierbei ist zu beachten, dass die Regressionsgerade durch den Ursprung gilt, da für $p = 0.5$

$$\frac{(x_{0.5} - x_{0.5})^2 + (x_{0.5} - x_{0.5})^2}{4\,x_{0.5}} = 0$$

und

$$\frac{x_p + x_{1-p}}{2} - x_{0.5} = 0$$

Emerson and Stoto (1982) schlagen vor, den Steigungsparameter robust zu schätzen. Hierzu bestimmen sie die Steigungen aller Geraden durch den Nullpunkt und die Punkte

$$\left(\frac{(x_p - x_{0.5})^2 + (x_{1-p} - x_{0.5})^2}{4\, x_{0.5}}, \frac{x_p + x_{1-p}}{2} - x_{0.5} \right)$$

für $p = 1/2, 1/4, 1/8, 1/16, \ldots$ und den Punkt $\left(x_{(1)}, x_{(n)}\right)$. Der Schätzer des Steigungsparameters ist der Median der Steigungen.

Beispiel 10 (fortgesetzt) *Emerson and Stoto (1982) schlagen vor, x_p für $p = 1/2, 1/4, 1/8, 1/16, \ldots$ nach der Methode von Tukey zu schätzen. Somit ist $x_{0.25}$ der Median der unteren Hälfte des geordneten Datensatzes, also von*

```
10 20 30 40 60
```

und $x_{0.75}$ der Median der oberen Hälfte des geordneten Datensatzes, also von

```
70 90 150 200 300
```

Als Schätzer für $x_{0.125}$ wählen wir den Median der unteren Hälfte der unteren Hälfte des geordneten Datensatzes, also von

```
10 20 30
```

Als Schätzer für $x_{0.875}$ wählen wir den Median der oberen Hälfte der oberen Hälfte des geordneten Datensatzes, also von

```
150 200 300
```

Der geordnete Datensatz wird so oft geteilt, bis nur noch eine Beobachtung übrig bleibt. Für das Beispiel erhalten wir die Schätzer

$$\begin{aligned} x_{0.25} &= 30 & x_{0.75} &= 150 \\ x_{0.125} &= 20 & x_{0.875} &= 200 \\ x_{0.0625} &= 15 & x_{0.9375} &= 250 \end{aligned}$$

In Tabelle 2.2 werden die relevanten Größen bestimmt, wobei zu $p = 0$ das Minimum bzw. Maximum der Daten gehört. In der letzten Spalte der Tabelle stehen die Steigungen der Geraden. Der Schätzer $\hat{c}$ ist gleich dem Median der Werte in der letzten Spalte, also $\hat{c} = 0.47$. Man sollte also die Quadratwurzel der Daten bilden.

Tabelle 2.2: Hilfstabelle zur Berechnung der Schätzer

p	x_p	x_{1-p}	$\frac{x_p+x_{1-p}}{2}$	$\frac{x_p+x_{1-p}}{2} - x_{0.5}$	$\frac{(x_p-x_{0.5})^2+(x_{1-p}-x_{0.5})^2}{4\,x_{0.5}}$	$\hat{c}$
0.2500	30	150	90.0	25.0	32.50	0.23
0.1250	20	200	110.0	45.0	77.88	0.42
0.0625	15	250	132.5	67.5	141.25	0.52
	10	300	155.0	90.0	224.04	0.60

2.6 Wie man eine Funktion durch eine lineare bzw. quadratische Funktion approximiert

Wir gehen aus von einer Funktion $f : D \subset \mathbb{R} \to \mathbb{R}$, die in $x = a$ existiert. Außerdem mögen die erste Ableitung $f'(x)$ und zweite Ableitung $f''(x)$ in einer Umgebung von $x = a$ existieren.

Beispiel 11 *$f : \mathbb{R} \to \mathbb{R}$ mit $x \mapsto y = f(x) = e^x$. Unser Ziel ist es, die Funktion f durch eine einfache Funktion zu approximieren.*

Es liegt nahe, die Funktion f durch eine Gerade zu approximieren. Wir setzen an

$$g(x) = a_0 + a_1\,(x - a) \tag{2.38}$$

Wie sollen wir a_0 und a_1 wählen? Wir fordern, dass g mit f an der Stelle $x = a$ sehr gut übereinstimmt. Es sollte also gelten

$$g(a) = f(a) \tag{2.39}$$

Außerdem fordern wir noch, dass die erste Ableitung von g mit der ersten Ableitung von f an der Stelle $x = a$ übereinstimmt:

$$g'(a) = f'(a) \tag{2.40}$$

gilt. Unter diesen Bedingungen können wir die Konstanten a_0 und a_1 bestimmen. Es gilt

$$g(a) = a_0 + a_1\,(a - a) = a_0$$

Mit Gleichung (2.39) auf Seite 54 folgt also

$$a_0 = f(a)$$

Außerdem gilt

$$g'(a) = a_1$$

Mit Gleichung (2.40) auf Seite 54 folgt also

$$a_1 = f'(a)$$

Also ist die gesucht Lösung gleich

$$g(x) = f(a) + f'(a)\,(x - a) \tag{2.41}$$

Beispiel 11 (fortgesetzt) *Wir betrachten* $f(x) = e^x$ *und* $a = 0$*. Es gilt*

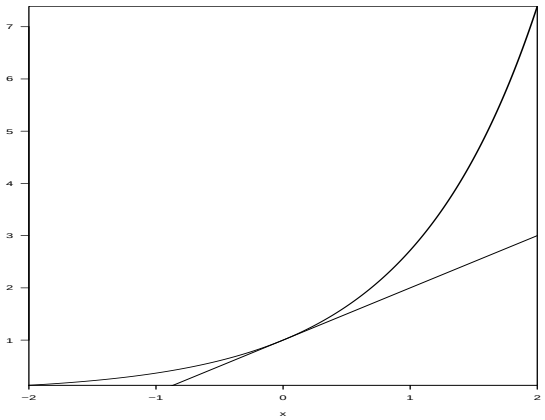
$$f'(x) = e^x$$

Aus $f(0) = 1$ *und* $f'(0) = 1$ *folgt*

$$g(x) = 1 + x \tag{2.42}$$

Abbildung 2.9 zeigt die Approximation.

Abbildung 2.9: Approximation von $f(x) = e^x$ durch eine Gerade



Wir können f aber auch durch eine quadratische Funktion approximieren. Gesucht ist eine Funktion

$$g(x) = a_0 + a_1\,(x - a) + a_2\,(x - a)^2 \tag{2.43}$$

In Analogie zur linearen Approximation fordern wir zusätzlich zur Gültigkeit der Gleichungen (2.39) und (2.40) auf Seite 54 noch die Gültigkeit von

$$g''(a) = f''(a) \tag{2.44}$$

Unter diesen Bedingungen können wir die Konstanten a_0, a_1 und a_2 bestimmen. Es gilt

$$g(a) = a_0 + a_1 (a - a) + a_2 (a - a)^2 = a_0$$

Mit Gleichung (2.39) auf Seite 54 folgt also

$$a_0 = f(a)$$

Außerdem gilt

$$g'(x) = a_1 + 2\, a_2 (x - a)$$

Somit gilt

$$g'(a) = a_1$$

Mit Gleichung (2.40) auf Seite 54 folgt also

$$a_1 = f'(a)$$

Außerdem gilt

$$g''(x) = 2\, a_2$$

Also gilt speziell

$$g''(a) = 2\, a_2$$

Mit Gleichung (2.44) auf Seite 56 folgt also

$$a_2 = \frac{1}{2} f''(a)$$

Somit gilt

$$g(x) = f(a) + f'(a) (x - a) + \frac{1}{2} f''(a) (x - a)^2 \tag{2.45}$$

Beispiel 11 (fortgesetzt) *Es gilt*

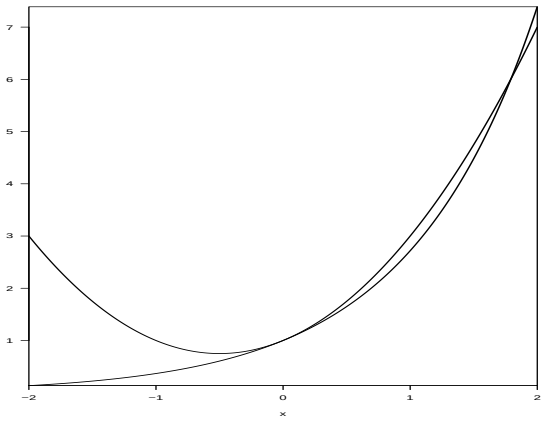
$$f''(x) = e^x$$

Mit $f''(0) = 1$ *folgt*

$$g(x) = 1 + x + x^2$$

Abbildung 2.10 zeigt die Approximation.

Abbildung 2.10: Approximation von $f(x) = e^x$ durch eine quadratische Funktion



2.7 Eine Anwendung der Approximation einer Funktion durch eine quadratische Funktion

Wir gehen im Folgenden von einer Zufallsstichprobe $x_1, \ldots, x_n$ aus einer Grundgesamtheit mit Verteilungsfunktion F aus. Somit sind die Beobachtungen $x_1, \ldots, x_n$ also Realisationen der unabhängigen und identisch mit Verteilungsfunktion F verteilten Zufallsvariablen $X_1, \ldots, X_n$. Die Verteilungsfunktion hänge von einem Parameter θ ab. Diesen wollen wir schätzen. Beim **Maximum-Likelihood-Verfahren** stellt man die gemeinsame Dichtefunktion $f(x_1, \ldots, x_n)$ bzw. Wahrscheinlichkeitsfunktion $P(X_1 = x_1, \ldots, X_n = x_n)$ auf und fasst diese als Funktion des Parameters θ gegeben die Daten $x_1, \ldots, x_n$ auf. Man bezeichnet sie als Likelihoodfunktion $L(\theta)$. Der **Maximum-Likelihood-Schätzer** ist nun der Wert von θ, für den die Likelihoodfunktion ihr Maximum annimmt. Ist die Likelihoodfunktion differenzierbar, so können wir die Verfahren der klassischen Analysis anwenden, um das Maximum zu bestimmen. Da die Likelihoodfunktion gleich dem Produkt der Randdichten bzw. Randwahrscheinlichkeitsfunktionen ist, erleichtert Logarithmieren die Bestimmung des Maximums beträchtlich. Man erhält also die sogenannte **Loglikelihoodfunktion**:

$$l(\theta) = \ln L(\theta)$$

Da der Logarithmus eine monotone Transformation ist, nimmt die Loglikelihoodfunktion ihr Maximum an der gleichen Stelle an wie die Likelihood-

funktion. Manchmal wird das Maximum schnell gefunden.

Beispiel 12 *Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig und identisch mit Parameter λ poissonverteilt. Somit gilt*

$$P(X_i = x_i) = \frac{\lambda^{x_i}}{x_i!}\, e^{-\lambda}$$

Die Likelihoodfunktion ist somit

$$L(\lambda) = \frac{\lambda^{x_1}}{x_1!}\, e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!}\, e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}\, e^{-n\,\lambda} = \frac{\lambda^{n\,\bar{x}}}{\prod_{i=1}^n x_i!}\, e^{-n\,\lambda}$$

Die Loglikelihoodfunktion ist

$$l(\lambda) = n\,\bar{x}\,\ln\lambda - \sum_{i=1}^{n} \ln x_i! - n\,\lambda$$

Eine notwendige Bedingung für einen Extremwert ist, dass die erste Ableitung $l'(\theta)$ gleich 0 ist.
Es gilt

$$l'(\lambda) = \frac{n\,\bar{x}}{\lambda} - n$$

Offensichtlich erfüllt $\hat{\lambda} = \bar{x}$ die Gleichung

$$l'(\hat{\lambda}) = 0$$

Weiterhin gilt

$$l''(\lambda) = -\frac{n\,\bar{x}}{\lambda^2}$$

Es gilt

$$l''(\bar{x}) = -\frac{n}{\bar{x}}$$

Sind nicht alls x_i gleich 0, so handelt es sich um ein Maximum.

Wie das folgende Beispiel zeigt, kann man das Maximum nicht immer so leicht bestimmen.

Tabelle 2.3: Häufigkeitstabelle des Merkmals Anzahl Kinder

Anzahl Kinder	absolute Häufigkeit	relative Häufigkeit
1	15	0.170
2	48	0.545
3	17	0.193
4	7	0.080
5	1	0.011

Beispiel 13 *In einer Vorlesung wurden die Studierenden gefragt, wie viele Kinder ihre Eltern haben. Da die befragte Person ein Kind seiner Eltern ist, war die bei der Antwort gegebene Zahl mindestens* 1. *Tabelle 2.3 zeigt die Häufigkeitsverteilung.*
Hier kann man die Poissonverteilung nicht als Modell verwenden, da der Wert 0 *nicht annehmen wird. Bei der Poissonverteilung ist diese Wahrscheinlichkit aber* $e^{-\lambda}$ *und somit größer als* 0. *Ein Modell für eine Zählvariable, die bei* 1 *zu zählen beginnt, ist die abgeschnittene Poissonverteilung. Deren Wahrscheinlichkeitsfunktion ist für* $x = 1, 2, \ldots$ *gegeben durch*

$$P(X = x) = \frac{\lambda^x}{x!} \frac{e^{-\lambda}}{1 - e^{-\lambda}}$$

Um den Maximum-Likelihood-Schätzer aus einer Stichprobe $x_1, \ldots, x_n$ *vom Umfang* n *zu bestimmen, stellen wir die Likelihood auf:*

$$L(\lambda) = \frac{\lambda^{n\,\bar{x}}}{\prod_{i=1}^{n} x_i!} \frac{e^{-n\,\lambda}}{(1 - e^{-\lambda})^n}$$

Die Loglikelihoodfunktion ist

$$l(\lambda) = n\,\bar{x}\,\ln\lambda - \sum_{i=1}^{n} \ln x_i! - n\,\lambda - n\,\ln\left(1 - e^{-\lambda}\right) \tag{2.46}$$

Eine notwendige Bedingung für einen Extremwert ist, dass die erste Ableitung $l'(\theta)$ *gleich* 0 *ist.*
Es gilt

$$l'(\lambda) = \frac{n\,\bar{x}}{\lambda} - n - \frac{n\,e^{-\lambda}}{(1 - e^{-\lambda})} \tag{2.47}$$

Die Gleichung

$$\frac{n\,\bar{x}}{\hat{\lambda}} - n - \frac{n\,e^{-\hat{\lambda}}}{\left(1 - e^{-\hat{\lambda}}\right)}$$

kann man nicht explizit nach $\hat{\lambda}$ auflösen. Hat man aber den Schätzwert gefunden, so kann man schnell nachprüfen, ob es sich um ein Maximum handelt, da die zweite Ableitung der Loglikelihoodfunktion lautet:

$$l''(\lambda) = -\frac{n\,\bar{x}}{\lambda^2} + \frac{n\,e^{-\lambda}}{(1 - e^{-\lambda})^2} \tag{2.48}$$

Kann man das Maximum der Loglikelihoodfunktion nicht explizit bestimmen, so sollte man ein iteratives Verfahren verwenden. Wir schauen uns das **Newton-Raphson-Verfahren** an, das von Everitt (1987) sehr anschaulich beschrieben wird.
Beim Newton-Raphson-Verfahren gibt man einen Startwert θ_0 für θ vor und approximiert die Loglikelihood $l(\theta)$ durch die quadratische Funktion $g(\theta)$, für die gilt

$$\begin{aligned} g(\theta_0) &= l(\theta_0) \\ g'(\theta_0) &= l'(\theta_0) \\ g''(\theta_0) &= l''(\theta_0) \end{aligned}$$

Wie wir in Kapitel 2.6 auf Seite 56 gesehen haben, ist die Funktion $g(x)$ gegeben durch

$$g(\theta) = l(\theta_0) + l'(\theta_0)\,(\theta - \theta_0) + \frac{1}{2}\,l''(\theta_0)\,(\theta - \theta_0)^2 \tag{2.49}$$

Das Maximum der Funktion $g(x)$ in Gleichung (2.49) ist leicht zu bestimmen. Wir bilden die erste Ableitung

$$g'(\theta) = l'(\theta_0) + l''(\theta_0)\,(\theta - \theta_0)$$

setzen sie gleich 0

$$l'(\theta_0) + l''(\theta_0)\,(\theta - \theta_0) = 0 \tag{2.50}$$

und lösen die Gleichung nach θ auf

$$\theta = \theta_0 - \frac{l'(\theta_0)}{l''(\theta_0)} \tag{2.51}$$

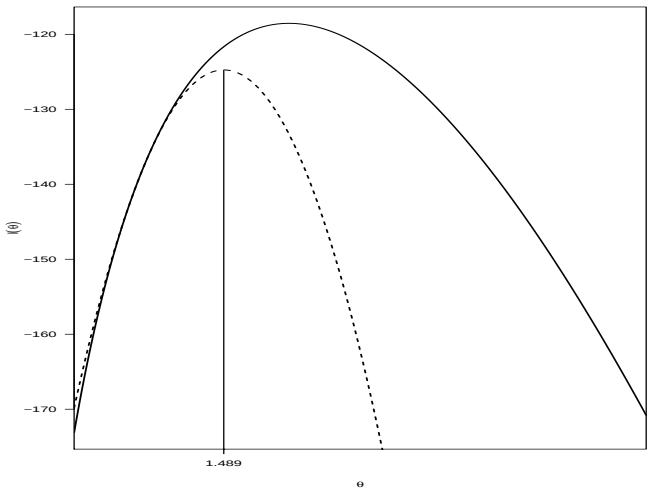
Hierdurch erhalten wir einen neuen Wert für das Maximum, der oft näher am Maximum der Loglikelihoodfunktion liegt als der Startwert θ_0.

Beispiel 13 (fortgesetzt) *Es gilt* $n = 88$ *und* $\bar{x} = 2.22$. *Wir wählen* $\lambda_0 = 1$. *Es gilt* $l(1) = -138.04$, $l'(1) = 55.79$ *und* $l''(1) = -113.98$. *Somit approximieren wir* $l(\theta)$ *durch*

$$g(x) = -138.04 + 55.79\,(\lambda - 1) - \frac{1}{2}\,113.98\,(\lambda - 1)^2 \tag{2.52}$$

Abbildung 2.11 zeigt die Likelihoodfunktion zusammen mit der approximierenden quadratischen Funktion und dem Maximum der approximierenden Funktion. Als neuen Schätzwert für das Maximum erhalten wir

Abbildung 2.11: Likelihoodfunktion zusammen mit der approximierenden quadratischen Funktion



$$\lambda = 1 - \frac{55.79}{-113.98} = 1.489$$

Wir sehen in Abbildung 2.11, dass das Maximum der approximierenden quadratischen Funktion näher am Maximum der Loglikelihood liegt als der Startwert. Wir approximieren nun die Loglikelihood durch die quadratische Funktion, die am neuen Maximum mit der Loglikelihood übereinstimmt und bestimmen das Maximum dieser Funktion. Dieses Verfahren iterieren wir so lange, bis sich der Wert nicht mehr ändert. Diese Vorgehensweise nennt man das Newton-Raphson-Verfahren. Hier ist der Algorithmus:

1. Wähle einen Startwert θ_0, eine vorgegeben Genauigkeit ϵ und setze i auf den Wert 0.

2. Berechne $\theta_{i+1} = \theta_i - l'(\theta_i)/l''(\theta_i)$.

3. Gehe zu Schritt 4, wenn $|\theta_i - \theta_{i+1}| < \epsilon$ gilt. Gilt aber $|\theta_i - \theta_{i+1}| \geq \epsilon$, so erhöhe i um 1 und gehe zu Schritt 2.

4. Wähle θ_i als Wert des Maximums.

Man kann im Schritt 2. des Algorithmus auch die $f$nderung der Loglikelihoodfunktion als Kriterium wählen. In diesem Fall bestimmt man also $|l(\theta_i) - l(\theta_{i+1})|$

Beispiel 13 (fortgesetzt) *Wir wählen* $\epsilon = 0.01$.
Mit $\lambda_1 = 1.489$, $l'(1.489) = 17.32$ *und* $l''(1.489) = -54.85$ *erhalten wir*

$$\lambda_2 = 1 - \frac{17.32}{-54.85} = 1.805$$

Wegen $\lambda_2 - \lambda_1| = 0.316$ *sind wir noch nicht am Ende angelangt.*
Mit $\lambda_2 = 1.805$, $l'(1.805) = 2.71$ *und* $l''(1.805) = -39.12$ *erhalten wir*

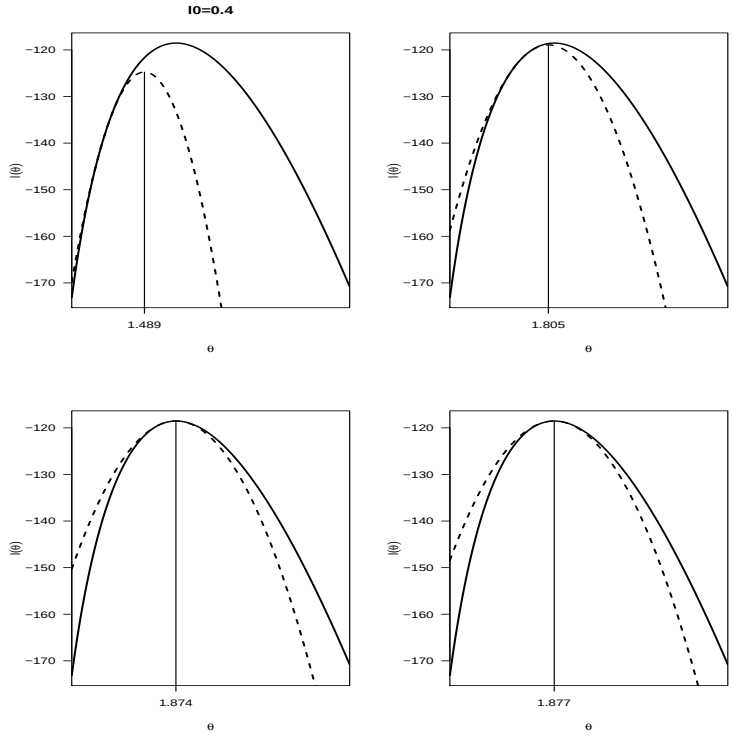
$$\lambda_3 = 1.805 - \frac{2.71}{-39.12} = 1.874$$

Wegen $|\lambda_3 - \lambda_2| = 0.069$ *sind wir noch nicht am Ende angelangt.*
Mit $\lambda_3 = 1.874$, $l'(1.874) = 2.71$ *und* $l''(1.874) = -39.12$ *erhalten wir*

$$\lambda_4 = 1.874 - \frac{0.097}{-36.67} = 1.877$$

Wegen $|\lambda_4 - \lambda_3| = 0.003$ *sind wir am Ende angelangt. Als M-L-Schätzer erhalten wir* $\hat{\lambda} = 1.877$.
Abbildung 2.12 zeigt die Iteration.

Abbildung 2.12: Likelihoodfunktion zusammen mit der approximierenden quadratischen Funktion



Kapitel 3

Schätzung des Lageparameters einer symmetrischen Verteilung

3.1 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 14 *Im Rahmen eines BI-Projektes sollten die Teilnehmer den Lineal-ReaktionsTest durchführen. Die Ergebnisse von Männern, die das Lineal mit der Nicht-Schreibhand fingen, sind:*

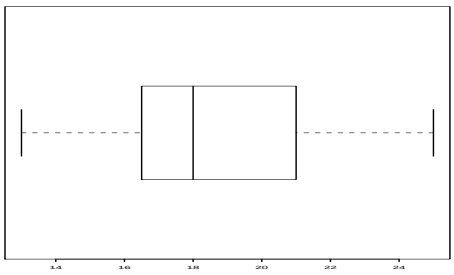
```
16 19 13 17 19 23 17 25
```

Abbildung 3.1 zeigt den Boxplot. Dieser deutet auf eine symmetrische Verteilung hin.

3.1.1 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

Abbildung 3.1: Boxplot der Reaktionszeit



folgendermaßen definiert:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.1}$$

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

$$x_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade ist} \\ \dfrac{x_{(n/2)} + x_{(1+n/2)}}{2} & \text{falls } n \text{ gerade ist} \end{cases} \tag{3.2}$$

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

Beispiel 14 (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.

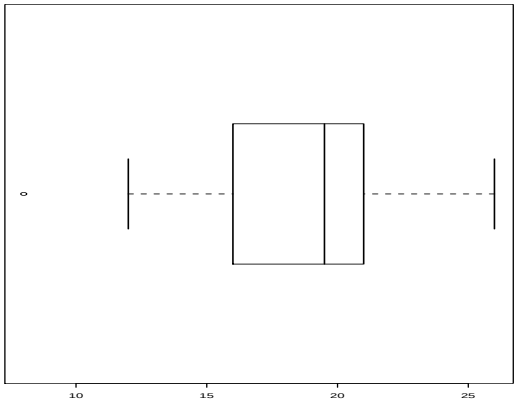
Beispiel 14 (fortgesetzt) *Da kein Ausreißer zu erkennen ist, beschreiben wir die Lage durch den Mittelwert.*

Beispiel 15 *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 3.2 zeigt den Boxplot. Dieser deutet auf eine symmetrische Verteilung hin. Es liegt aber ein Ausreißer vor. Wir wählen also den Median.*

Abbildung 3.2: Boxplot der Reaktionszeit



3.1.2 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

$$\bar{x}_k = \frac{1}{n-2\,k}\sum_{i=k+1}^{n-k} x_{(i)} \tag{3.3}$$

Beispiel 16 *Wir betrachten die Daten aus Beispiel 15 auf Seite 66. Hier ist die geordnete Stichprobe:*

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

Wir setzen $k = 1$. Es gilt

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

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:

$$\bar{x}_\alpha = \frac{1}{n - 2\,g} \sum_{i=g+1}^{n-g} x_{(i)} \tag{3.4}$$

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

Beispiel 16 (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 3.1 sind die Werte von* $\bar{x}_\alpha$ *für ausgewählte Werte von* α *zu finden.*

Tabelle 3.1: 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, Mosteller, and Tukey (1983) verdeutlichen dies. Hoaglin, Mosteller, and Tukey (1983) schlagen vor, auch Anteile von geordneten Beobachtungen zu verwenden. Dies liefert folgenden Schätzer:

$$T(\alpha) = \frac{1}{n(1 - 2\,\alpha)} \left\{ (1 - r)\,\left[x_{(g+1)} + x_{(n-g)} \right] + \sum_{i=g+2}^{n-g-1} x_{(i)} \right\} \tag{3.5}$$

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

Beispiel 16 (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*

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

Tabelle 3.2: 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

In Tabelle 3.2 sind die Werte von $T(\alpha)$ für ausgewählte Werte von α zu finden.

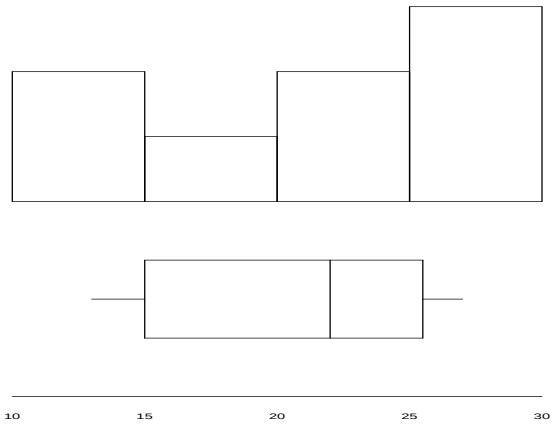
Die Lage einiger Datensätze sollten nicht durch getrimmte Mittelwerte beschrieben werden. Schauen wir uns hierzu ein Beispiel an.

Beispiel 17 *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 3.3 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 3.3: Histogramm und Boxplot der Reaktionszeit



Die Daten im Beispiel 17 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

$$\frac{X_{(1)} + X_{(n)}}{2} \tag{3.6}$$

Der Beweis ist bei Mood, Graybill, and Boes (1974) auf den Seiten 282-283 zu finden.
Man nennt den Ausdruck in Gleichung (3.6) auch den Midrange oder die Spannweitenmitte. Er ist ein Beispiel für einen Mittelwert der getrimmten Werte, der folgendermaßen definiert ist:

$$T(\alpha)^c = \begin{cases} \frac{x_{(1)} + x_{(1)}}{2} & \text{für } n\,\alpha < 1 \\ \frac{1}{2\,n\,\alpha} \left\{ r \left[x_{(g+1)} + x_{(n-g)} \right] + \sum_{i=1}^{g} \left(x_{(i)} + x_{(n+1-i)} \right) \right\} & \text{für } n\,\alpha \geq 1 \end{cases} \tag{3.7}$$

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 3.4, so gilt

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

Beispiel 17 (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 3.3 sind die Werte von $T(\alpha)$ *für ausgewählte Werte von* α *zu finden.*

Tabelle 3.3: 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

3.2 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 65 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.

3.2.1 Effiziente Schätzfunktionen

Definition 3.2.1 *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 18 *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 19 *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

$$f_X(\theta - x) = f_X(\theta + x) \tag{3.8}$$

und

$$F_X(\theta - x) = 1 - F_X(\theta + x) \tag{3.9}$$

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

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,(n-k)!}\, F_X(x)^{k-1}\, [1 - F_X(x)]^{n-k}\, f_X(x) \tag{3.10}$$

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

$$\begin{aligned}
f_{X_{(\frac{n+1}{2})}}(\theta - x) &= \frac{n!}{\left[(\frac{n-1}{2})!\right]^2}\, F_X(\theta - x)^{\frac{n-1}{2}}\, [1 - F_X(\theta - x)]^{\frac{n-1}{2}}\, f_X(\theta - x) \\
&\overset{(3.8)(3.9)}{=} \frac{n!}{\left[(\frac{n-1}{2})!\right]^2}\, [1 - F_X(\theta + x)]^{\frac{n-1}{2}}\, F_X(\theta + x)^{\frac{n-1}{2}}\, f_X(\theta + x) \\
&= \frac{n!}{\left[(\frac{n-1}{2})!\right]^2}\, F_X(\theta + x)^{\frac{n-1}{2}}\, [1 - F_X(\theta + x)]^{\frac{n-1}{2}}\, f_X(\theta + x) \\
&= f_{X_{(\frac{n+1}{2})}}(\theta + x)
\end{aligned}$$

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 3.2.2 *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 19 (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.

3.2.2 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 3.10 auf Seite 72 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:

$$F_U(u) = \begin{cases} 0 & \text{für } u \leq 0 \\ u & \text{für } 0 < u < 1 \\ 1 & \text{für } u \geq 1 \end{cases} \tag{3.11}$$

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

$$f_U(u) = \begin{cases} 1 & \text{für } 0 < u < 1 \\ 0 & \text{sonst} \end{cases} \tag{3.12}$$

Setzen wir diese Gleichungen in die Gleichung 3.10 auf Seite 72 ein, so erhalten wir für $x \in (0, 1)$

$$\begin{aligned} f_{U_{(k)}}(x) &= \frac{n!}{(k-1)!\,(n-k)!}\, F_U(x)^{k-1}\, [1 - F_U(x)]^{n-k}\, f_U(x) \\ &= \frac{n!}{(k-1)!\,(n-k)!}\, x^{k-1}\, (1-x)^{n-k} \end{aligned} \tag{3.13}$$

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

$$B(a, b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)} \tag{3.14}$$

(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 $b = n + 1 - k$.

$$f_X(x) = \begin{cases} \frac{1}{B(a,b)}\, x^{a-1}\,(1-x)^{b-1} & \text{für } 0 < x < 1 \\ 0 & \text{sonst} \end{cases} \tag{3.15}$$

(siehe dazu Mood, Graybill, and Boes (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, Graybill, and Boes (1974), S. 117)
Also gilt

$$E(U_{(k)}) = \frac{k}{k+n+1-k} = \frac{k}{n+1} \tag{3.16}$$

und

$$\begin{aligned} Var(U_{(k)}) &= \frac{k\,(n+1-k)}{(k+n+1-k+1)(k+n+1-k)^2} = \frac{k\,(n+1-k)}{(n+2)(n+1)^2} \\ &= \left(\frac{1}{n+2}\right)\left(\frac{k}{n+1}\right)\left(1-\frac{k}{n+1}\right) \end{aligned} \tag{3.17}$$

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 (3.17) 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 3.2.1 *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

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

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

$$X_{(k)} = F^{-1}\left(U_{(k)}\right) \tag{3.18}$$

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:

$$\begin{aligned} F^{-1}\left(U_{(k)}\right) &\approx F^{-1}\left(E\left(U_{(k)}\right)\right) + \left(U_{(k)} - E\left(U_{(k)}\right)\right)\left(F^{-1}\right)'\left(E\left(U_{(k)}\right)\right) \\ &\overset{(3.17)}{=} F^{-1}\left(\frac{k}{n+1}\right) + \left(U_{(k)} - \frac{k}{n+1}\right)\left(F^{-1}\right)'\left(\frac{k}{n+1}\right) \end{aligned} \tag{3.19}$$

Es gilt

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

(siehe Heuser (2001)).
Somit gilt

$$F^{-1}\left(U_{(k)}\right) \approx F^{-1}\left(\frac{k}{n+1}\right) + \left(U_{(k)} - \frac{k}{n+1}\right)\frac{1}{f\left(F^{-1}(\frac{k}{n+1})\right)} \tag{3.20}$$

Also gilt

$$\begin{aligned} Var\left(F^{-1}\left(U_{(k)}\right)\right) &\approx Var\left(F^{-1}\left(\frac{k}{n+1}\right)+\left(U_{(k)}-\frac{k}{n+1}\right)\frac{1}{f(F^{-1}\left(\frac{k}{n+1}\right))}\right) \\ &= Var\left(\left(U_{(k)}-\frac{k}{n+1}\right)\frac{1}{f(F^{-1}\left(\frac{k}{n+1}\right))}\right) \\ &= \frac{1}{\left(f(F^{-1}(\frac{k}{n+1}))\right)^2}Var\left(U_{(k)}\right) \\ &\overset{(3.17)}{=} \frac{1}{\left(f(F^{-1}(\frac{k}{n+1}))\right)^2}\left(\frac{1}{n+2}\right)\left(\frac{k}{n+1}\right)\left(1-\frac{k}{n+1}\right) \quad (3.21) \end{aligned}$$

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

$$Var(X_{0.5})=\frac{1}{4\,n\,(f(F^{-1}(0.5)))^2} \quad (3.22)$$

Die gleiche Beziehung gilt für gerade Stichprobenumfänge.

Beispiel 19 (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 19 (fortgesetzt) *Bei Normalverteilung gilt*

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

Tabelle 3.4 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. Quelle: Kendall, Stuart, and Ord (1991).

Tabelle 3.4: $\frac{Var(\bar{X})}{Var(X_{0.5})}$ 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

Beispiel 19 (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 3.4 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.
Es gilt $E(X) = \mu$ und $Var(X) = 2\,\beta^2$. (siehe dazu Mood, Graybill, and Boes (1974), S. 117)
Somit gilt

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

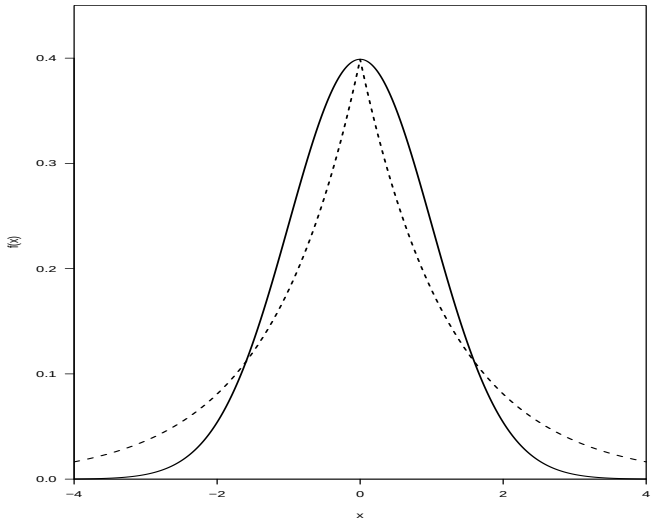
Mit

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

gilt also

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

Abbildung 3.4: Dichtefunktionen der Laplace-Verteilung und der Standardnormalverteilung



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.

3.2.3 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 (3.11) und (3.12) auf Seite 73 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 20 *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 20 (fortgesetzt) *Um einmal zu würfeln, erzeugen wir eine gleichverteilte Zufallszahl u und bilden*

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

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 3.2.1 auf Seite 76 zurück. Wir erzeugen eine auf $(0, 1)$ gleichverteilte Zufallszahl u und erhalten die Zufallszahl aus $F(x)$ durch $x = F^{-1}(u)$.

Beispiel 21 *Die Zufallsvariable X besitze eine Laplace-Verteilung mit den Parametern $\mu = 0$ und $\beta = 1$. Die Dichtefunktion von X lautet also*

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

Die Verteilungsfunktion lautet

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

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

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

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 20 (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 3.5 sind die Ergebnisse zu finden. Für* $i = 1$ *und* $i = 6$ *können wir*

Tabelle 3.5: 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

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 20 (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 3.6 sind die Ergebnisse der Simulation zu finden. Wir sehen, dass

Tabelle 3.6: 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

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.

3.2.4 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 20 (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 3.7 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. 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

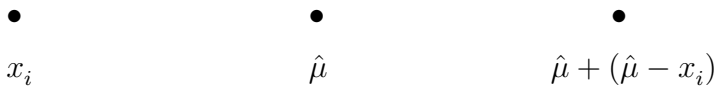
Tabelle 3.7: Bootstrap-Stichproben

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

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 20 (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 3.8 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 3.8: 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

Kapitel 4

Statistische Intervalle

4.1 Einführung

Bei der Veröffentlichung der Ergebnisse statistischer Erhebungen werden vielfach Punktschätzer angegeben. So ist beim Umwelt- und Prognose-Institut am 16.01.2004 folgende Aussage zu finden:

> Die durchschnittliche Fahrleistung des Autofahrers liegt seit Jahren stabil bei 12000 Kilometern im Jahr.

Dieser Wert ist eine Schätzung. Dies merkt man schon daran, dass es sich um eine runde Zahl handelt. Es fehlt aber eine Genauigkeitsangabe in Form der Varianz. Oft wird ein Intervall angegeben. Dies trägt dem Umstand Rechnung, dass die Schätzung fehlerbehaftet ist. Wir wollen uns im Folgenden Intervalle anschauen.
Dabei gehen wir von einer Zufallsstichprobe $x_1, \ldots, x_n$ aus einer Grundgesamtheit mit stetiger Verteilungsfunktion $F_X(x)$ aus. Die Beobachtungen $x_1, \ldots, x_n$ sind also Realisationen der unabhängigen, identisch mit Verteilungsfunktion $F_X(x)$ verteilten Zufallsvariablen $X_1, \ldots, X_n$.
Wir betrachten im Folgenden das Intervall

$$\left[x_{(1)}, x_{(n)}\right] \tag{4.1}$$

Dabei ist $x_{(1)}$ die kleinste und $x_{(n)}$ die größte Beobachtung in der Stichprobe.

Beispiel 21 *Ein Arbeitnehmer hat eine neue Stelle angenommen. Er will wissen, welche Charakteristika die Fahrzeit zur Arbeitsstelle besitzt. Deshalb notiert er die Fahrzeit an* 9 *aufeinander folgenden Tagen. Hier sind die Werte in Sekunden:*

```
1670 1775 1600 1700 2000 1890 1740 1880 1945
```

Es gilt $x_{(1)} = 1600$ *und* $x_{(9)} = 2000$. *Wir betrachten also das Intervall*

$$[1600, 2000] \tag{4.2}$$

In Abhängigkeit von der Fragestellung kann man das Intervall in Gleichung (4.1) unterschiedlich interpretieren.
Ist man an einem Parameter wie dem Erwartungswert $E(X)$ oder dem Median $X_{0.5}$ interessiert, so wird man ein Konfidenzintervall aufstellen.

Definition 4.1.1 *Seien* $X_1, \ldots, X_n$ *unabhängige, identisch verteilte Zufallsvariablen, deren Verteilungsfunktion* $F_X(x)$ *von einem Parameter* θ *abhängt. Außerdem seien* $T_1 = g_1(X_1, \ldots, X_n)$ *und* $T_2 = g_2(X_1, \ldots, X_n)$ *zwei Stichprobenfunktionen. Dann heißt das Intervall*

$$[T_1, T_2] \tag{4.3}$$

mit

$$P\left(T_1 \leq \theta \leq T_2\right) = 1 - \alpha \tag{4.4}$$

zweiseitiges Konfidenzintervall *für* θ *zum* **Konfidenzniveau** $1 - \alpha$.

Ist der Parameter θ gleich dem Median $X_{0.5}$, so können wir das Intervall in Gleichung (4.1) auf Seite 87 als Konfidenzintervall für den Median $X_{0.5}$ auffassen. Es ist einfach, das Konfidenzniveau des Intervalls $\left[x_{(1)}, x_{(n)}\right]$ bestimmen. Es gilt

$$P\left(X_{(1)} \leq X_{0.5} \leq X_{(n)}\right) = 1 - \left[P\left(X_{(1)} > X_{0.5}\right) + P\left(X_{(n)} < X_{0.5}\right)\right]$$

Nun gilt

$$\begin{aligned} P\left(X_{(1)} > X_{0.5}\right) &= P\left(\text{alle } X_i > X_{0.5}\right) = P\left(X_1 > X_{0.5}, \ldots, X_n > X_{0.5}\right) \\ &= P\left(X_1 > X_{0.5}\right) \cdots P\left(X_n > X_{0.5}\right) = 0.5^n \end{aligned}$$

Analog erhalten wir

$$P\left(X_{(n)} < X_{0.5}\right) = 0.5^n$$

Somit gilt

$$P\left(X_{(1)} \leq X_{0.5} \leq X_{(n)}\right) = 1 - 2 \cdot 0.5^n = 1 - 0.5^{n-1}$$

Beispiel 21 (fortgesetzt von Seite 87) *Der Arbeitnehmer ist an der mittleren Fahrzeit interessiert, die er mit dem Median misst. Das Konfidenzintervall für den Median ist in Gleichung* (4.2) *auf Seite 88 zu finden. Das Konfidenzniveau beträgt*

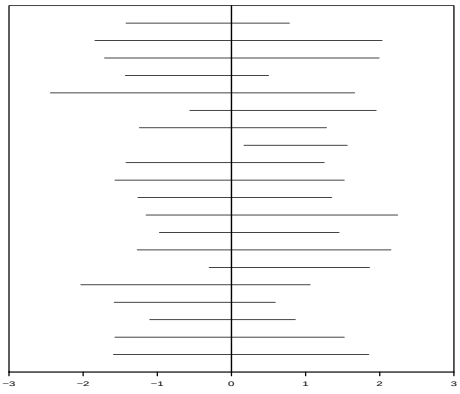
$$1 - \alpha = 1 - 0.5^8 = 0.996$$

Wir wollen verdeutlichen, wie das Konfidenzniveau zu interpretieren ist. Für ein konkretes Intervall gibt es nur zwei Möglichkeiten. Der unbekannte Wert des Parameters liegt in dem Intervall oder er liegt nicht in dem Intervall. Wir wissen nicht, welche der beiden Möglichkeiten zutrifft. Wir können aber die Wahrscheinlichkeit bestimmen, dass wir ein Intervall gefunden haben, das den Wert des Parameters überdeckt. Hierzu führen wir eine Simulation durch. Wir unterstellen Standardnormalverteilung und ziehen 20 Stichproben vom Umfang $n = 9$. Für jede dieser Stichproben stellen wir das Konfidenzintervall aus Gleichung (4.1) auf Seite 87 für $X_{0.5}$ auf. Die Daten sind in Tabelle A.1 auf 110 zu finden.
Die erste simulierte Stichprobe lautet:

```
1.05  1.07 -0.12  0.03 -1.59  0.35  1.85  0.40  0.36
```

Das Intervall ist $[-1.59, 1.85]$. Der Wert von $X_{0.5}$ ist uns bekannt. Er ist 0. Das Intervall enthält den Wert von $X_{0.5}$. Abbildung 4.1 verdeutlicht dies und zeigt die anderen 19 Konfidenzintervalle.

Abbildung 4.1: 20 Konfidenzintervalle



Wir sehen, dass 19 Konfidenzintervalle den Wert 0 enthalten.

Das Konfidenzniveau bezieht sich auf den Prozess und nicht das konkrete Intervall. Stellen wir sehr viele Konfidenzintervalle zum Konfidenzniveau $1-\alpha$ auf, so erwarten wir, dass $100 \cdot (1-\alpha)$ Prozent den wahren Wert des Parameters überdecken. Ist $1-\alpha$ nahe bei 1, so können wir uns bei einem konkreten Intervall also ziemlich sein, dass es den wahren Wert des Parameters enthält. Im Beispiel haben wir den Stichprobenumfang vorgegeben. Wollen wir das Intervall $\left[x_{(1)}, x_{(n)}\right]$ als Konfidenzintervall zum vorgegebenen Konfidenzniveau $1-\alpha$ aufstellen, so bestimmen wir den Stichprobenumfang n folgendermaßen:

$$n \geq 1 + \frac{\ln(\alpha)}{\ln 0.5}$$

Dies sieht man folgendermaßen:

$$1 - 0.5^{n-1} \geq 1 - \alpha \iff 0.5^{n-1} \leq \alpha \iff (n-1)\ln 0.5 \leq \ln(\alpha)$$

$$\iff (n-1) \geq \frac{\ln(\alpha)}{\ln 0.5} \iff n \geq 1 + \frac{\ln(\alpha)}{\ln 0.5}$$

Beispiel 22 *Wollen wir das Intervall* $\left[x_{(1)}, x_{(n)}\right]$ *als Konfidenzintervall zum Konfidenzniveau* 0.95*, so benötigen wir eine Stichprobe vom Umfang* 6*, denn*

$$n \geq 1 + \frac{\ln(1-0.95)}{\ln 0.5} = 5.32$$

Wir können auch einseitige Konfidenzintervalle aufstellen.

Definition 4.1.2 *Seien* $X_1, \ldots, X_n$ *unabhängige, identisch verteilte Zufallsvariablen, deren Verteilungsfunktion* $F_X(x)$ *von einem Parameter* θ *abhängt. Außerdem seien* $T_1 = g_1(X_1, \ldots, X_n)$ *und* $T_2 = g_2(X_1, \ldots, X_n)$ *zwei Stichprobenfunktionen. Dann heißen die Intervalle*

$$(-\infty, T_2] \tag{4.5}$$

mit

$$P(\theta \leq T_2) = 1 - \alpha \tag{4.6}$$

und

$$[T_1, \infty) \tag{4.7}$$

mit

$$P(T_1 \leq \theta) = 1 - \alpha \tag{4.8}$$

einseitige Konfidenzintervalle *für* θ *zum Konfidenzniveau* $1-\alpha$*.*

Einseitige Konfidenzintervalle werden verwendet, wenn man angeben will, welchen Wert ein Parameter mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ mindestens oder höchstens annehmen kann.
Die Intervalle

$$\left[x_{(1)}, \infty\right) \tag{4.9}$$

und

$$\left(-\infty, x_{(n)}\right] \tag{4.10}$$

kann man als einseitige Konfidenzintervalle für den Median auffassen. Für beide Intervalle gilt

$$1 - \alpha = 1 - 0.5^n$$

Ein Konfidenzintervall ist ein Intervall für einen Parameter. Will man aber einen zukünftigen Wert x_{n+1} einer Zufallsvariablen auf Basis einer Stichprobe $x_1, \ldots, x_n$ vorhersagen, so spricht man von Prognose. Wie bei der Schätzung kann man auch bei der Prognose Intervalle für den zukünftigen Wert angeben.

Definition 4.1.3 *Seien $X_1, \ldots, X_n, X_{n+1}$ unabhängige, identisch mit Verteilungsfunktion $F_X(x)$ verteilte Zufallsvariablen. Außerdem seien $T_1 = g_1(X_1, \ldots, X_n)$ und $T_2 = g_2(X_1, \ldots, X_n)$ zwei Stichprobenfunktionen.*
Dann heißt das Intervall

$$[T_1, T_2] \tag{4.11}$$

mit

$$P\left(T_1 \leq X_{n+1} \leq T_2\right) = 1 - \alpha \tag{4.12}$$

Prognoseintervall *für X_{n+1} zur* **Sicherheit** $1 - \alpha$.

Wir können das Intervall $\left[x_{(1)}, x_{(n)}\right]$ als Prognoseintervall auffassen. Es gilt für jede stetige Verteilung

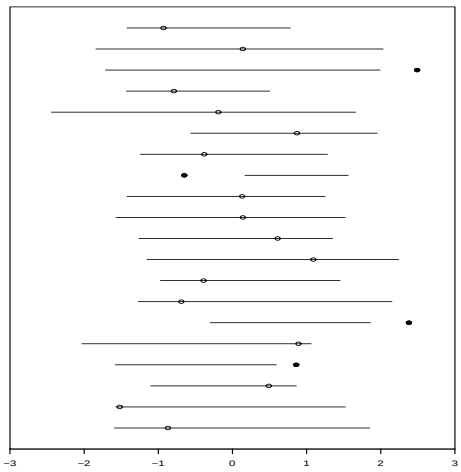
$$1 - \alpha = \frac{n-1}{n+1} \tag{4.13}$$

Bevor wir uns überlegen, warum $1 - \alpha$ diesen Wert annimmt, wollen wir zunächst $1 - \alpha$ interpretieren.
Wie auch bei Konfidenzintervallen gibt es bei Prognoseintervallen für ein konkretes Intervall $[t_1, t_2]$ zwei Möglichkeiten. Die Realisation x_{n+1} liegt im Intervall $[t_1, t_2]$ oder sie liegt nicht im Intervall $[t_1, t_2]$. Die Wahrscheinlichkeitsaussage bezieht sich auch hier nicht auf das Intervall sondern auf die Prozedur. Wenn wir sehr viele Stichproben $x_1, \ldots, x_n, x_{n+1}$ vom Umfang $n+1$ ziehen und für $x_1, \ldots, x_n$ das Prognoseintervall bestimmen, so erwarten wir, dass in $100 \cdot 1 - \alpha$ Prozent der Prognoseintervalle der Wert x_{n+1} liegt.
Wir wollen dies für das Intervall $\left[x_{(1)}, x_{(n)}\right]$ mit einer Simulation veranschaulichen. Wir unterstellen Standardnormalverteilung und ziehen 20 Stichproben vom Umfang $n = 9$. Für jede dieser Stichproben stellen wir das Prognoseintervall $\left[x_{(1)}, x_{(n)}\right]$ auf. Die Daten sind in Tabelle A.1 auf 110 zu finden. Die erste simulierte Stichprobe lautet:

```
1.05  1.07 -0.12  0.03 -1.59  0.35  1.85  0.40  0.36
```

Das Intervall ist $[-1.59, 1.85]$. Der Wert von x_{10} ist -0.87. Das Intervall enthält den Wert von x_{10}. Abbildung 4.2 verdeutlicht dies. Außerdem sind noch die anderen 19 Prognoseintervalle zu finden. Wir sehen, dass 16 Prognoseintervalle den Wert x_{n+1} enthalten.

Abbildung 4.2: 20 Prognoseintervalle



Beispiel 21 (fortgesetzt von Seite 89) *Will er nur wissen, wie lange er bei der nächsten Fahrt unterwegs ist, so wird er eine Prognose bestimmen. Ein Prognoseintervall sagt ihm, zwischen welchen Grenzen die nächste Fahrzeit liegen wird. Verwendet er das Intervall* $[1600, 2000]$ *als Prognoseintervall, so gilt*

$$1 - \alpha = \frac{9-1}{9+1} = 0.8$$

Wenden wir uns jetzt der Frage zu, warum Gleichung (4.13) auf Seite 92 gilt. Wenn die Verteilungsfunktion $F_X(x)$ bekannt ist, gilt

$$P\left(X_{(1)} \leq X \leq X_{(n)}\right) = P\left(F_X\left(X_{(1)}\right) \leq F_X(X) \leq F_X\left(X_{(n)}\right)\right) \qquad (4.14)$$

Besitzt X die Verteilungsfunktion $F_X(x)$, so ist die Zufallsvariable $F_X(X)$ gleichverteilt auf $(0, 1)$. Der Beweis ist bei Randles and Wolfe (1979) auf der Seite 6 zu finden. $F_X(X_{(i)})$ ist also die i-te Orderstatistik einer Zufallsstichprobe vom Umfang n aus einer Gleichverteilung auf $(0, 1)$. Diese bezeichnen wir mit $U_{(i)}$. Bickel and Doksum (2001) zeigen auf Seite 254:

$$P\left(X_{(1)} \leq X \leq X_{(n)}\right) = E\left(U_{(n)}\right) - E\left(U_{(1)}\right)$$

Es gilt

$$E\left(U_{(i)}\right) = \frac{i}{n+1}$$

(siehe dazu Randles and Wolfe (1979), S. 7). Hieraus folgt

$$E\left(U_{(n)} - U_{(1)}\right) = \frac{n}{n+1} - \frac{1}{n+1} = \frac{n-1}{n+1}$$

Im Beispiel haben wir den Stichprobenumfang vorgegeben. Wollen wir das Intervall in Gleichung (4.1) auf Seite 87 als Prognoseintervall zum vorgegebenen Prognoseniveau $1-\alpha$ aufstellen, so bestimmen wir den Stichprobenumfang n folgendermaßen:

$$n \geq \frac{2-\alpha}{\alpha}$$

Dies sieht man folgendermaßen:

$$\frac{n-1}{n+1} \geq 1-\alpha \iff n-1 \geq (n+1)(1-\alpha) \Longleftrightarrow n-1 \geq n(1-\alpha)+1-\alpha$$

$$\iff n\,\alpha \geq 2-\alpha \Longleftrightarrow n \geq \frac{2-\alpha}{\alpha}$$

Beispiel 23 *Verwenden wir das Intervall* $\left[x_{(1)}, x_{(n)}\right]$ *als Prognoseintervall zur Sicherheit* 0.95*, so benötigen wir eine Stichprobe vom Umfang* 39*, denn*

$$n \geq \frac{2-0.05}{0.05} = 39$$

Wir können auch einseitige Prognoseintervalle aufstellen.

Definition 4.1.4 *Seien* $X_1, \ldots, X_n, X_{n+1}$ *unabhängige, identisch mit Verteilungsfunktion* $F_X(x)$ *verteilte Zufallsvariablen. Außerdem seien* $T_1 = g_1(X_1, \ldots, X_n)$ *und* $T_2 = g_2(X_1, \ldots, X_n)$ *zwei Stichprobenfunktionen. Dann heißen die Intervalle*

$$(-\infty, T_2] \tag{4.15}$$

mit

$$P\left(X_{n+1} \leq T_2\right) = 1-\alpha \tag{4.16}$$

und

$$[T_1, \infty) \tag{4.17}$$

mit

$$P\left(T_1 \leq X_{n+1}\right) = 1-\alpha \tag{4.18}$$

einseitige Prognoseintervalle *für* x_{n+1} *zur Sicherheit* $1-\alpha$.

Einseitige Prognoseintervalle werden verwendet, wenn man angeben will, welchen Wert eine zukünftige Beobachtung mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ mindestens oder höchstens annehmen kann. Die Intervalle

$$\left[x_{(1)}, \infty\right) \tag{4.19}$$

und

$$\left(-\infty, x_{(n)}\right] \tag{4.20}$$

kann man als einseitige Prognoseintervalle auffassen. Für beide Intervalle gilt

$$1-\alpha = \frac{n}{n+1}$$

Hahn (1970) nennt das Prognoseintervall auch das Astronauten-Intervall. Ein Astronaut interessiert sich nur für den nächsten Flug und will wissen, welche Werte der interessierenden Merkmale er erwarten kann. Der Hersteller eines Produktes ist aber nicht nur an einer Beobachtung interessiert, sondern an der gesamten zukünftigen Produktion. Er will ein Intervall angeben, in dem sich mindestens der Anteil p aller zukünftigen Beobachtungen befindet.

Definition 4.1.5 *Seien $X_1, \ldots, X_n$ unabhängige, identisch mit Verteilungsfunktion $F_X(x)$ verteilte Zufallsvariablen.*
Außerdem seien $T_1 = g_1(X_1, \ldots, X_n)$ und $T_2 = g_2(X_1, \ldots, X_n)$ zwei Stichprobenfunktionen.
Dann heißt das Intervall

$$[T_1, T_2] \tag{4.21}$$

mit

$$P\left(P(T_1 \leq X_{n+1} \leq T_2) \geq p\right) = 1-\alpha \tag{4.22}$$

Toleranzintervall für den Mindestanteil p zur Sicherheit $1-\alpha$.

Wie können wir diese Wahrscheinlichkeit interpretieren?
Das Toleranzintervall soll mindestens $100 \cdot p$ Prozent aller Realisationen der Zufallsvariablen enthalten. Für ein konkretes Intervall gibt es zwei Möglichkeiten. Es enthält mindestens die Wahrscheinlichkeitsmasse p oder es enthält sie nicht.
Wir wollen dies für das Intervall $\left[x_{(1)}, x_{(n)}\right]$ mit einer Simulation veranschaulichen. Wir unterstellen Standardnormalverteilung und ziehen 20 Stichproben vom Umfang $n = 9$. Für jede dieser Stichproben stellen wir das Konfidenzintervall aus Gleichung (4.1) auf Seite 87 für $X_{0.5}$ auf. Die Daten sind in Tabelle A.1 auf 110 zu finden. Nehmen wir an, die erste simulierte Stichprobe lautet

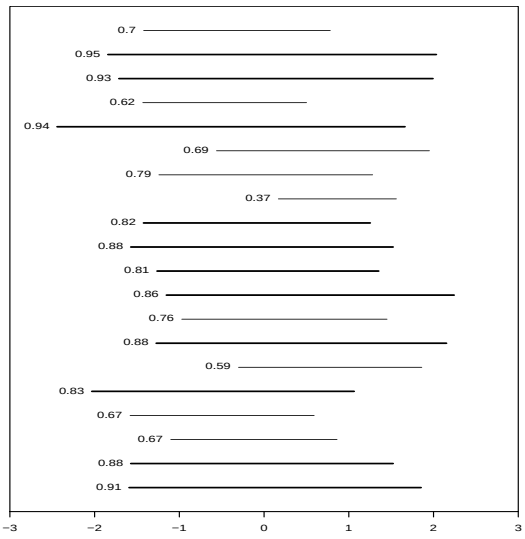
```
1.05  1.07 -0.12  0.03 -1.59  0.35  1.85  0.40  0.36
```

Das Intervall ist $[-1.59, 1.85]$. Sei $p = 0.9$. Ist X standardnormalverteilt, so gilt

$$P(-1.59 \leq X \leq 1.85) = \Phi(1.85) - \Phi(-1.59) = 0.9119.$$

Das Intervall $[-1.59, 1.85]$ enthält $100 \cdot 0.9119$ Prozent aller Realisationen einer standardnormalverteilten Zufallsvariablen. Abbildung 4.3 verdeutlicht dies und zeigt die 19 anderen Intervalle. Wir sehen, dass 11 Toleranzintervalle, also 55 Prozent mindestens die Wahrscheinlichkeitsmasse 0.8 enthalten.

Abbildung 4.3: 20 Toleranzzintervalle



Bei einem Toleranzintervall bezieht sich die innere Wahrscheinlichkeitsaussage in Gleichung (4.22) auf Seite 95 auf das Intervall, während sich die äußere Wahrscheinlichkeitsaussage auf die Prozedur bezieht. Wenn wir sehr viele Stichproben $x_1, \ldots, x_n$ und für jede das Toleranzintervall $[t_1, t_2]$ bestimmen, so erwarten wir, dass $100 \cdot (1 - \alpha)$ Prozent der Intervalle mindestens die Wahrscheinlichkeitsmasse p enthalten.

Fassen wir das Intervall $\left[x_{(1)}, x_{(n)}\right]$ als Toleranzintervall mit Mindestanteil p auf, so gilt für jede stetige Verteilung

$$1 - \alpha = 1 - p^n - n(1 - p)p^{n-1} \tag{4.23}$$

Diese Beziehung wird in Mood, Graybill, and Boes (1974) auf den Seiten 516-517 hergeleitet.

Beispiel 21 (fortgesetzt von Seite 93) *Der Arbeitnehmer ist an einem Toleranzintervall mit Mindestanteil* 0.8 *interessiert. Verwendet er das Intervall* $[1600, 2000]$ *als Toleranzintervall mit Mindestanteil* 0.8*, so beträgt die Sicherheit*

$$1 - \alpha = 1 - 0.8^9 - 9 \cdot 0.2 \cdot 0.8^9 = 0.624$$

Im Beispiel haben wir den Stichprobenumfang vorgegeben. Wollen wir das Intervall $\left[x_{(1)}, x_{(n)}\right]$ als Toleranzintervall mit Mindestanteil p zur vorgegebenen Sicherheit $1 - \alpha$ aufstellen, so müssen wir die Gleichung (4.23) nach n auflösen. Dies ist nicht explizit möglich. Eine gute Approximation ist bei Conover (1999) auf Seite 151 zu finden. Sie lautet

$$n \geq \frac{1+p}{4 \cdot (1-p)} \cdot \chi^2_{1-\alpha,4} + 0.5 \tag{4.24}$$

Dabei ist $\chi^2_{1-\alpha,4}$ das $1 - \alpha$-Quantil der χ^2-Verteilung mit 4 Freiheitsgraden.

Beispiel 24 *Wollen wir das Intervall* $\left[x_{(1)}, x_{(n)}\right]$ *als Toleranzintervall für den Mindesanteil* 0.9 *zur Sicherheit* 0.95 *verwenden, so benötigen wir eine Stichprobe vom Umfang* 46*, denn*

$$n \geq \frac{1+0.9}{4 \cdot (1-0.9)} \cdot 9.49 + 0.5 = 45.6$$

Setzen wir $n = 46$ *und* $p = 0.9$ *in Gleichung* (4.23) *ein, so erhalten wir*

$$1 - \alpha = 1 - 0.9^{46} - 46 \cdot 0.1 \cdot 0.9^{45} = 0.952$$

Für $n = 45$ *und* $p = 0.9$ *gilt*

$$1 - \alpha = 1 - 0.9^{45} - 45 \cdot 0.1 \cdot 0.9^{44} = 0.948$$

Wir können auch einseitige Toleranzintervalle aufstellen.

Definition 4.1.6 *Seien* $X_1, \ldots, X_n$ *unabhängige, identisch mit Verteilungsfunktion* $F_X(x)$ *verteilte Zufallsvariablen.*
Außerdem seien $T_1 = g_1(X_1, \ldots, X_n)$ *und* $T_2 = g_2(X_1, \ldots, X_n)$ *zwei Stichprobenfunktionen. Dann heißen die Intervalle*

$$(-\infty, T_2] \tag{4.25}$$

mit

$$P(P\left(X \leq T_2\right) \geq p) = 1 - \alpha \tag{4.26}$$

und

$$[T_1, \infty) \tag{4.27}$$

mit

$$P(P\,(X \geq T_1) \geq p) = 1 - \alpha \tag{4.28}$$

einseitige Toleranzintervalle *für den Mindestanteil p zur Sicherheit* $1-\alpha$.

Die Intervalle

$$\big[x_{(1)}, \infty\big) \tag{4.29}$$

und

$$\big(-\infty, x_{(n)}\big] \tag{4.30}$$

kann man als einseitige Toleranzintervalle auffassen.
Für beide Intervalle gilt

$$1 - \alpha = 1 - p^n$$

4.2 Intervalle bei Normalverteilung

Im letzten Kapitel haben wir uns mit verteilungsfreien Intervallen beschäftigt. Ist die Verteilung der Grundgesamtheit bekannt, so kann man Intervalle aufstellen, die bei gleichem Niveau wesentlich schmaler sind. Wir wollen im Folgenden nur Normalverteilung unterstellen. Wir gehen also davon aus, dass die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und mit den Parametern μ und σ^2 normalverteilt sind.

4.2.1 Konfidenzintervalle

Die Dichtefunktion der Normalverteilung hängt von den Parametern μ und σ^2 ab. Wir wollen im Folgenden Konfidenzintervalle für jeden der beiden Parameter herleiten. Wie man ein simultanes Konfidenzintervall für beide Parameter erhält, wird von Mood, Graybill, and Boes (1974) auf den Seiten 384-385 beschrieben.

4.2.1.1 Konfidenzintervall für μ

Ist die Varianz der Grundgesamtheit bekannt, so ist das Konfidenzintervall für μ bei Normalverteilung gegeben durch:

$$\left[\bar{X} - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{4.31}$$

Dabei ist $z_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Standardnormalverteilung.
Es gilt nämlich

$$\begin{aligned}
&P\left(\bar{X} - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right)\\
&= P\left(-z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} \le \mu - \bar{X} \le z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right)\\
&= P\left(-z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} \le \bar{X} - \mu \le z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right)\\
&= P\left(-z_{1-\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}\right)\\
&= \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2})\\
&= 1-\alpha
\end{aligned}$$

Schauen wir uns das Konfidenzintervall in Gleichung (4.31) unter praxisrelevanten Gesichtspunkten an. Bei einer Datenanalyse wird σ^2 in der Regel unbekannt sein. Es liegt nahe, σ^2 durch

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

zu schätzen und diesen Schätzer in Gleichung (4.31) für σ^2 einzusetzen. Das Intervall sieht also folgendermaßen aus:

$$\left[\bar{X} - z_{1-\alpha/2}\,\frac{S}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]. \tag{4.32}$$

Für kleine Stichprobenumfänge gilt aber

$$P\left(\bar{X} - z_{1-\alpha/2}\,\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + z_{1-\alpha/2}\,\frac{S}{\sqrt{n}}\right) \neq 1-\alpha\,. \tag{4.33}$$

Eine kleine Simulation zeigt dies. Wir erzeugen 5000 Stichproben vom Umfang 4 aus der Standardnormalverteilung, stellen für jede das Konfidenzintervall in Gleichung (4.32) auf und zählen, wie viele der Konfidenzintervalle den Wert 0 überdecken. Das geschätze Konfidenzniveau beträgt 0.8757. Die Konfidenzintervalle sind also im Mittel zu schmal. Um zu sehen, woran dies liegt, formen wir den Ausdruck in der Klammer auf der linken Seite von Gleichung (4.33) um:

$$\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}$$

$$\Longleftrightarrow -z_{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu - \bar{X} \leq z_{1-\alpha/2}\frac{S}{\sqrt{n}}$$

$$\Longleftrightarrow -z_{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{1-\alpha/2}\frac{S}{\sqrt{n}}$$

$$\Longleftrightarrow -z_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq z_{1-\alpha/2}$$

Wir müssen also folgende Wahrscheinlichkeit bestimmen:

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq z_{1-\alpha/2}\right)$$

Die Zufallsvariable

$$t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

ist nicht standardnormalverteilt, wenn die $X_1, \ldots, X_n$ unabhängig und mit den Parametern μ und σ^2 normalverteilt sind. Die Schätzung von σ^2 führt dazu, dass t stärker streut als die Standardnormalverteilung. Von Gossett (1908) wurde gezeigt, dass t eine t-Verteilung mit $n-1$ Freiheitsgraden besitzt. Ist $t_{n-1;1-\alpha/2}$ das $1-\alpha/2$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden, so gilt

$$P\left(-t_{n-1;1-\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{n-1;1-\alpha/2}\right) = 1-\alpha \tag{4.34}$$

Wir formen den Ausdruck in der Klammer auf der linken Seite von Gleichung (4.34) so um, dass zwischen den Ungleichheitszeichen nur noch μ steht. Es gilt

$$P(\bar{X} - t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

Durch diese Umformung haben wir ein Konfidenzintervall für μ bei Normalverteilung mit unbekannten σ^2 gefunden. Es lautet:

$$\left[\bar{X} - t_{n-1;1-\alpha/2}\,\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right] . \tag{4.35}$$

Dabei ist $t_{n-1;1-\alpha/2}$ das $1-\alpha/2$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.

Beispiel 21 (fortgesetzt von Seite 97) *Wir unterstellen Normalverteilung und stellen das Konfidenzintervall für μ zum Konfidenzniveau* 0.95 *auf. Es gilt $\bar{x} = 1800$ und $s = 135.4$. Mit $n = 9$ gilt also $s/\sqrt{n} = 45.13$. Der Tabelle der t-Verteilung entnehmen wir $t_{8;0.975} = 2.306$. Mit*

$$t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}} = 104.1$$

Wir erhalten also folgendes Konfidenzintervall

$$[1695.9, 1904.1]$$

Das verteilungsfreie Konfidenzintervall $\left[x_{(1)}, x_{(n)}\right]$ lautet $[1600, 2000]$. *Das Konfidenzniveau ist* 0.996. *Zum Vergleich stellen wir noch das Konfidenzintervall für μ zum Konfidenzniveau* 0.996 *auf. Es gilt $t_{8;0.998} = 3.99$. Also erhalten wir folgendes Konfidenzintervall*

$$[1619.9, 1980.1]$$

Dieses ist schmaler als das verteilungsfreie Konfidenzintervall.

Schauen wir die Länge L des Konfidenzintervalls an. Es gilt

$$L = 2\,t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}} \tag{4.36}$$

Dies sieht man folgendermaßen

$$\begin{aligned} L &= \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}} - \left(\bar{x} - t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}}\right) \\ &= 2\,t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}} \end{aligned}$$

Das Intervall wird breiter, wenn wir das Konfidenzniveau $1-\alpha$ vergrößern. Es wird aber nicht notwendigerweise schmaler, wenn wir den Stichprobenumfang erhöhen. Für eine neue Stichprobe werden wir auch einen anderen Wert

von s erhalten, sodass das Intervall größer werden kann. Es ist auch nicht möglich den Mindeststichprobenumfang zu bestimmen, um eine vorgegebene Länge des Intervalls nicht zu überschreiten, da der Stichprobenumfang von s abhängt. Aus $L \leq l$ folgt nämlich

$$n \geq \frac{4\, t^2_{n-1;1-\alpha/2}\, s^2}{l^2}\,.$$

Die einseitigen Konfidenzintervalle für μ bei Normalverteilung sind.

$$\left(-\infty, \bar{X} + t_{n-1;1-\alpha}\frac{S}{\sqrt{n}}\right]\,. \tag{4.37}$$

und

$$\left[\bar{X} - t_{n-1;1-\alpha}\,\frac{S}{\sqrt{n}}, \infty\right)\,. \tag{4.38}$$

Dabei ist $t_{n-1;1-\alpha}$ das $1-\alpha$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.

Beispiel 21 (fortgesetzt von Seite 101) *Wir suchen den Wert, den μ mit Wahrscheinlichkeit* 0.95 *mindestens annimmt.*
Es gilt $t_{0.95,8} = 1.86$. Mit $\bar{x} = 1800$, $s = 135.4$ und $s/\sqrt{n} = 45.13$ ist dieser Wert also gleich

$$1800 - 1.86 \cdot 45.13 = 1716.1$$

4.2.1.2 Konfidenzintervall für σ^2

Wir gehen wiederum davon aus, dass Normalverteilung vorliegt und wollen ein Konfidenzintervall für σ^2 aufstellen. Um zu verstehen, wie man dies erhält, schauen wir uns noch einmal das Konfidenzintervall für μ an. Bei diesem sind wir von

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

ausgegangen. Diese Zufallsvariable ist mit $n-1$ Freiheitsgraden t-verteilt, wenn die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und mit den Parametern μ und σ^2 normalverteilt sind. Somit gilt

$$P\left(-t_{n-1,1-\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

Formen wir den Ausdruck in der Klammer so um, dass μ isoliert ist, so haben wir das Konfidenzintervall für μ gefunden.
Die Vorgehensweise zeigt, wie man ein Konfidenzintervall für einen Parameter finden kann. Man benötigt eine Stichprobenfunktion,

1. die von dem unbekannten Parameter abhängt
2. deren Verteilung bekannt ist

Um ein Konfidenzintervall für σ^2 aufzustellen, liegt es nahe, von

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

auszugehen, da S^2 eine erwartungstreue Schätzfunktion von σ^2 ist. Multipliziert man S^2 mit $n-1$ und dividiert die so gewonnene Größe durch σ^2, so erhält man

$$\frac{(n-1)\,S^2}{\sigma^2} \tag{4.39}$$

Die Zufallsvariable in Gleichung (4.39) genügt den Anfoderungen 1 und 2. Sie hängt von σ^2 ab und ist χ^2-verteilt mit $n-1$ Freiheitsgraden. Dies zeigen Mood, Graybill, and Boes (1974) auf den Seiten 243-245. Somit gilt

$$P\left(\chi^2_{n-1,\alpha/2} \leq \frac{(n-1)\,S^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha \tag{4.40}$$

Dabei sind $\chi^2_{n-1,\alpha/2}$ das $\alpha/2$-Quantil und $\chi^2_{n-1,1-\alpha/2}$ das $1-\alpha/2$-Quantil der χ^2-Verteilung mit $n-1$ Freiheitsgraden.
Wir formen den Ausdruck in der Klammer in Gleichung (4.40) so um, dass σ^2 isoliert ist.

$$\chi^2_{n-1,\alpha/2} \leq \frac{(n-1)\,S^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}$$

$$\iff \frac{1}{\chi^2_{n-1,\alpha/2}} \geq \frac{\sigma^2}{(n-1)\,S^2} \geq \frac{1}{\chi^2_{n-1,1-\alpha/2}}$$

$$\iff \frac{(n-1)\,S^2}{\chi^2_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha/2}}$$

Das Konfidenzintervall für σ^2 bei Normalverteilung ist also:

$$\left[\frac{(n-1)\,S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha/2}}\right] \tag{4.41}$$

Beispiel 21 (fortgesetzt von Seite 102) *Wir unterstellen Normalverteilung und stellen das Konfidenzintervall für σ^2 zum Konfidenzniveau* 0.95 *auf. Es gilt $s^2 = 18331.25$. Der Tabelle der χ^2-Verteilung entnehmen wir $\chi^2_{8;0.025} = 2.18$ und $\chi^2_{8;0.975} = 17.53$. Wir erhalten also folgendes Konfidenzintervall*

$$[8363.49, 67278.95]$$

4.2.2 Prognoseintervalle

Sind die Parameter μ und σ^2 bekannt, so ist das zentrale Schwankungsintervall

$$[\mu - z_{1-\alpha/2}\sigma, \mu + z_{1-\alpha/2}\sigma] \tag{4.42}$$

ein Prognoseintervall, denn es gilt

$$P(\mu - z_{1-\alpha/2}\sigma \leq X \leq \mu + z_{1-\alpha/2}\sigma) = 1 - \alpha$$

In der Regel sind μ und σ^2 aber unbekannt. Es liegt nahe, diese zu schätzen und die Schätzwerte in die Gleichung (4.42) einzusetzen. Dies ist aber nur für sehr große Stichprobenumfänge sinnvoll. Für kleine Stichprobenumfänge hingegen ist das exakte Prognoseintervall bei Normalverteilung gegeben durch

$$[\bar{x} - t_{n-1;1-\alpha/2}\, s\, \sqrt{1+1/n}, \bar{x} + t_{n-1;1-\alpha/2}\, s\, \sqrt{1+1/n}] \tag{4.43}$$

Dabei ist $t_{n-1;1-\alpha/2}$ das $1-\alpha/2$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.
Schauen wir uns ein Beispiel an, bevor wir zeigen, wie man dieses Intervall gewinnt.

Beispiel 21 (fortgesetzt von Seite 103) *Der Arbeitnehmer sucht ein Prognoseintervall für die Fahrzeit des nächsten Tages zur sicherheit* 0.95*. Es gilt* $n = 9$*,* $\bar{x} = 1800$ *und* $s = 135.4$*. Mit* $t_{8;0.975} = 2.306$ *erhalten wir folgendes Prognoseintervall*

$$[1470.9, 2129.1]$$

Um das Intervall herzuleiten, fragen wir uns zunächst, wie wir den Wert von X_{n+1} prognostizieren sollen. Da er möglichst nahe an allen Beobachtungen liegen sollte, bieten sich zwei Kriterien an. Wählt man als Maß für die Nähe, die euklidische Distanz, so erhält man folgendes Kriterium

$$\min \sum_{i=1}^{n} |x_i - x_{n+1}| \tag{4.44}$$

Die quadrierte euklidische Distanz liefert

$$\min \sum_{i=1}^{n} (x_i - x_{n+1})^2 \tag{4.45}$$

Im ersten Fall prognostiziert man x_{n+1} durch den Median, im zweiten Fall durch den Mittelwert der Beobachtungen. Da die Verteilung des Mittelwerts angegeben werden kann, verwenden wir diesen. Als Ausgangspunkt der Konstruktion des Prognoseintervalls wählen wir $X_{n+1} - \bar{X}$. Wir gehen im Folgenden davon aus, dass die Zufallsvariablen $X_1, \ldots, X_n, X_{n+1}$ unabhängig und identisch mit den Parametern μ und σ^2 normalverteilt sind. Unter diesen Annahmen gilt

$$E(X_{n+1} - \bar{X}) = E(X_{n+1}) - E(\bar{X}) = \mu - \mu = 0$$

und

$$Var(X_{n+1} - \bar{X}) = Var(X_{n+1}) + Var(\bar{X}) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2(1 + 1/n)$$

Außerdem ist $X_{n+1} - \bar{X}$ normalverteilt. Also ist

$$\frac{X_{n+1} - \bar{X}}{\sigma\sqrt{1 + 1/n}} \tag{4.46}$$

standardnormalverteilt. Schätzen wir σ durch S und setzen es in Gleichung (4.46) ein, so erhalten wir folgende mit $n - 1$ Freiheitsgraden t-verteilte Zufallsvariable

$$\frac{X_{n+1} - \bar{X}}{S\sqrt{1 + 1/n}}$$

Es gilt also

$$P\left[-t_{n-1;1-\alpha/2} \leq \frac{X_{n+1} - \bar{X}}{S\sqrt{1 + 1/n}} \leq t_{n-1;1-\alpha/2}\right] = 1 - \alpha \tag{4.47}$$

Formen wir diesen Ausdruck so um, dass zwischen den Ungleichungen X_{n+1} steht, so erhalten wir das Prognoseintervall in Gleichung (4.43):

$$-t_{n-1;1-\alpha/2} \leq \frac{X_{n+1} - \bar{X}}{S\sqrt{1 + 1/n}} \leq t_{n-1;1-\alpha/2}$$

$$\iff t_{n-1;1-\alpha/2}\, S\sqrt{1 + 1/n} \leq X_{n+1} - \bar{X} \leq t_{n-1;1-\alpha/2}\, S\sqrt{1 + 1/n}$$

$$\iff \bar{X} - t_{n-1;1-\alpha/2}\, S\sqrt{1 + 1/n} \leq X_{n+1} \leq \bar{X} + t_{n-1;1-\alpha/2}\, S\sqrt{1 + 1/n}$$

Da ein Konfidenzintervall ein Intervall für den Wert eines Parameters und ein Prognoseintervall ein Intervall für eine Realisation einer Zufallsvariablen ist,

ist die Aussage beim Prognoseintervall mit größerer Unsicherheit behaftet. Dies zeigt sich in der größeren Länge des Prognoseintervalls. Die Länge des Konfidenzintervalls für μ ist nämlich

$$L = 2\, t_{n-1;1-\alpha/2}\, s\, \sqrt{1/n}$$

Die Länge des Prognoseintervalls beträgt

$$L = 2\, t_{n-1;1-\alpha/2}\, s\, \sqrt{1+1/n}$$

Da gilt

$$1 + \frac{1}{n} > \frac{1}{n}$$

gilt auch

$$\sqrt{1+\frac{1}{n}} > \sqrt{\frac{1}{n}}$$

Die Länge des Konfidenzintervalls konvergiert gegen 0, während die Länge des Prognoseintervalls gegen die Länge des zentralen Schwankungsintervalls konvergiert.

Oft sucht man einseitige Prognoseintervalle. Man will also wissen, welchen Wert die nächste Beobachtung mindestens oder höchsten annimmt.

Bei Normalverteilung gibt es folgende einseitige Prognoseintervalle

$$[\bar{x} - t_{n-1;1-\alpha}\, s\, \sqrt{1+1/n}, \infty) \tag{4.48}$$

$$(-\infty, \bar{x} + t_{n-1;1-\alpha}\, s\, \sqrt{1+1/n}] \tag{4.49}$$

Dabei ist $t_{n-1;1-\alpha}$ das $1-\alpha$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.

Beispiel 21 (fortgesetzt von Seite 104) *Der Arbeitnehmer will wissen, welchen Wert seine Fahrzeit am nächsten Tag nicht überschreiten wird. Er stellt ein einseitiges Prognoseintervall zur Sicherheit* 0.95 *auf. Es gilt* $n = 9$, $\bar{x} = 1800$ *und* $s = 135.4$. *Mit* $t_{8;0.95} = 1.86$ *erhalten wir folgendes Prognoseintervall*

$$(-\infty, 2065.4]$$

Will er wissen, welchen Wert seine Fahrzeit am nächsten Tag überschreiten wird, so erhält er folgendes einseitige Prognoseintervall zur Sicherheit 0.95*:*

$$[1534.6, \infty)$$

4.2.3 Toleranzintervalle

Sind die Parameter μ und σ^2 unbekannt, so ist das zweiseitige Toleranzintervall bei Normalverteilung gegeben durch

$$[\bar{x} - q_{1-\alpha,p,n}\, s, \bar{x} + q_{1-\alpha,p,n}\, s] \tag{4.50}$$

Dabei gilt

$$q_{1-\alpha,p,n} = r \sqrt{\frac{n-1}{\chi^2_{\alpha,n-1}}} \tag{4.51}$$

mit

$$\Phi\left(\frac{1}{\sqrt{n}} + r\right) - \Phi\left(\frac{1}{\sqrt{n}} - r\right) = p \tag{4.52}$$

Dabei ist $\chi^2_{\alpha,n-1}$ das α-Quantil der χ^2-Verteilung mit $n-1$ Freiheitsgraden und $\Phi(z)$ der Wert der Verteilungsfunktion der Standardnormalverteilung an der Stelle z.

Die Gleichungen (4.51) und (4.52) werden von Kendall, Stuart, and Ord (1991) auf den Seiten 774-776 hergeleitet.

Tabellen für $q_{1-\alpha,p,n}$ sind bei Hahn and Meeker (1991), Rinne (1997) und auf den Seiten 112 und 113 zu finden.

Wir können $q_{1-\alpha,p,n}$ aber auch numerisch bestimmen. Hierzu müssen wir zuerst den Wert von r bestimmen, der die Gleichung (4.52) erfüllt. Lösen wir die Gleichung (4.52) nach 0 auf, so erhalten wir

$$\Phi\left(\frac{1}{\sqrt{n}} + r\right) - \Phi\left(\frac{1}{\sqrt{n}} - r\right) - p = 0$$

Wir suchen also die Nullstelle von

$$f(r) = \Phi\left(\frac{1}{\sqrt{n}} + r\right) - \Phi\left(\frac{1}{\sqrt{n}} - r\right) - p$$

Ein numerisches Verfahren ist das Newton-Raphson-Verfahren. Ist die Nullstelle der Funktion $f(r)$ gesucht, so iteriert man hier beginnend mit dem Startwert r_0

$$r_n = r_{n-1} - \frac{f(r_{n-1})}{f'(r_{n-1})} \tag{4.53}$$

für $n = 1, 1, \ldots$ so lange, bis r_n sich stabilisiert.

Als Startwert r_0 kann man den Wert von r wählen, der für $n \to \infty$ die Gleichung (4.52) erfüllt.
Für gegebenes p wird der Wert r_0 gesucht mit

$$\Phi\left(r_0\right) - \Phi\left(-r_0\right) = p$$

Wegen

$$\Phi\left(-r_0\right) = 1 - \Phi\left(r_0\right)$$

muss also gelten

$$2\,\Phi\left(r_0\right) - 1 = p$$

Somit folgt

$$r_0 = \Phi^{-1}\left(\frac{p+1}{2}\right)$$

Mit

$$f'(r) = \phi\left(\frac{1}{\sqrt{n}} + r\right) + \phi\left(\frac{1}{\sqrt{n}} - r\right)$$

können wir die Iteration aus Gleichung (4.53) auf Seite 107 also durchführen.
Dabei gilt

$$\phi(z) = \frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,z^2}$$

Beispiel 21 (fortgesetzt von Seit 106) *Der Arbeitnehmer will wissen, in welchem Intervall mindestens* 90 *Prozent der Fahrzeiten mit einer Wahrscheinlichkeit von* 0.95 *liegen.*
Es gilt also $p = 0.9$ *und* $1 - \alpha = 0.95$. *Wir bestimmen zunächst* r. *Es gilt*

$$r_0 = \Phi^{-1}\left(\frac{0.9+1}{2}\right) = \Phi^{-1}\left(0.95\right) = 1.645$$

Somit gilt

$$\begin{aligned} r_1 &= 1.645 - \frac{\Phi\left(\frac{1}{\sqrt{9}} + 1.645\right) - \Phi\left(\frac{1}{\sqrt{9}} - 1.645\right) - 0.9}{\frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,(1/3+1.645)^2} + \frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,(1/3-1.645)^2}} \\ &= 1.645 - \frac{\Phi(1.98) - \Phi(1.31) - 0.9}{\frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,(1/3+1.645)^2} + \frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,(1/3-1.645)^2}} \\ &= 1.645 - \frac{0.9761 - 0.0948 - 0.9}{0.0564 + 0.1688} = 1.645 - \frac{-0.0187}{0.2252} = 1.728 \end{aligned}$$

Ausgehend von r_1 *bestimmen wir* r_2 *und erhalten* $r_2 = 1.734$. *Auch* r_3 *ist gleich* 1.734. *Wir können also* $q_{0.95,0.9,9}$ *bestimmen. Mit* $\chi^2_{0.05,8} = 2.7326$ *gilt*

$$q_{0.95,0.9,9} = 1.734 \cdot \sqrt{\frac{8}{2.7326}} = 2.967$$

Dieser Wert stimmt mit den Werten in den Tabellen in Hahn and Meeker (1991) und Rinne (1997) überein. Mit $\bar{x} = 1800$ *und* $s = 135.4$ *erhalten wir folgendes Toleranzintervall*

$$[1398.3, 2201.7]$$

Wir können auch einseitige Toleranzintervalle bestimmen:

$$[\bar{x} - w_{1-\alpha,p,n}\, s, \infty) \tag{4.54}$$

$$(-\infty, \bar{x} + w_{1-\alpha,p,n}\, s] \tag{4.55}$$

Die Werte von $w^{ue}_{1-\alpha,p,n}$ sind bei Hahn and Meeker (1991) und Rinne (1997) tabelliert.

Anhang A

Die simulierten Daten

Tabelle A.1: 20 Stichproben vom Umfang $n = 9$ aus der Standardnormalverteilung

Stichprobe									
1	1.05	1.07	-0.12	0.03	-1.59	0.35	1.85	0.40	0.36
2	-0.58	0.12	1.27	0.28	0.68	0.08	0.01	-1.57	1.52
3	0.82	-1.10	-0.79	0.48	0.86	0.22	-0.15	-0.61	0.36
4	0.59	-0.24	-0.64	-0.76	0.51	-0.60	-1.51	-1.58	-0.31
5	-2.03	0.60	1.06	-0.14	-0.29	0.11	0.50	-0.19	-0.71
6	1.86	1.71	0.17	0.87	0.67	-0.17	0.75	1.13	-0.30
7	-0.75	-1.12	2.15	-0.88	0.18	-1.27	-0.49	-0.47	-0.01
8	0.39	-0.17	-0.97	-0.74	0.91	-0.55	-0.61	-0.68	1.45
9	2.24	1.73	-1.15	0.20	1.54	-0.57	-0.90	0.24	-0.30
10	-1.26	-0.13	-0.23	0.89	1.35	-0.18	0.95	-0.73	0.62
11	0.53	1.52	0.84	-0.43	0.38	-0.54	-1.57	-0.85	0.71
12	1.25	-1.42	0.29	0.45	-0.55	0.48	-0.17	1.07	-0.51
13	0.41	0.94	1.56	0.17	1.10	0.76	0.20	1.46	0.51
14	-1.24	0.29	-1.07	1.28	0.13	0.82	-0.98	-0.27	-0.06
15	0.27	-0.25	0.44	-0.56	0.79	-0.25	0.10	1.95	0.82
16	1.03	-0.01	-2.44	1.66	-2.17	1.44	0.04	0.20	1.06
17	-0.76	0.06	-1.25	0.25	0.19	-1.43	0.50	-0.18	-0.04
18	-0.42	-1.71	0.62	-0.91	0.48	1.29	1.99	0.39	1.24
19	2.03	0.17	-0.40	-1.84	1.50	0.16	-1.15	-0.64	0.97
20	0.78	-0.02	-1.25	-1.42	-0.85	0.38	0.74	-0.99	0.44

20 standardnormalverteilte Zufallszahlenn

```
-0.87 -1.52  0.49  0.86 0.89  2.38 -0.69 -0.39 1.09  0.61
 0.14  0.13 -0.65 -0.38 0.87 -0.19 -0.79  2.49 0.14 -0.93
```

Anhang B

Tabellen

Tabelle B.1: Werte von $q_{0.95,p,n}$ für $p = 0.9, 0.95, 0.99$ und $n = 2, 3, \ldots, 20$

n \ p	0.90	0.95	0.99
2	32.019	37.674	48.430
3	8.380	9.916	12.861
4	5.369	6.370	8.299
5	4.275	5.079	6.634
6	3.712	4.414	5.775
7	3.369	4.007	5.248
8	3.136	3.732	4.891
9	2.967	3.532	4.631
10	2.839	3.379	4.433
11	2.737	3.259	4.277
12	2.655	3.162	4.150
13	2.587	3.081	4.044
14	2.529	3.012	3.955
15	2.480	2.954	3.878
16	2.437	2.903	3.812
17	2.400	2.858	3.754
18	2.366	2.819	3.702
19	2.337	2.784	3.656
20	2.310	2.752	3.615

Tabelle B.2: Werte von $q_{0.99,p,n}$ für $p = 0.9, 0.95, 0.99$ und $n = 2, 3, \ldots, 20$

n \ p	0.90	0.95	0.99
2	160.194	188.491	242.301
3	18.930	22.401	29.055
4	9.398	11.150	14.527
5	6.612	7.855	10.260
6	5.337	6.345	8.301
7	4.613	5.488	7.187
8	4.147	4.936	6.468
9	3.822	4.550	5.966
10	3.582	4.265	5.594
11	3.397	4.045	5.308
12	3.250	3.870	5.079
13	3.130	3.727	4.893
14	3.029	3.608	4.737
15	2.945	3.507	4.605
16	2.872	3.421	4.492
17	2.808	3.345	4.393
18	2.753	3.279	4.307
19	2.703	3.221	4.230
20	2.659	3.168	4.161

Tabelle B.3: Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2.0	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999
3.0	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Tabelle B.4: Quantil z_p der Standardnormalverteilung

p	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
0.50	0.000	0.002	0.005	0.008	0.010	0.012	0.015	0.018	0.020	0.023
0.51	0.025	0.028	0.030	0.033	0.035	0.038	0.040	0.043	0.045	0.048
0.52	0.050	0.053	0.055	0.058	0.060	0.063	0.065	0.068	0.070	0.073
0.53	0.075	0.078	0.080	0.083	0.085	0.088	0.090	0.093	0.095	0.098
0.54	0.100	0.103	0.106	0.108	0.110	0.113	0.116	0.118	0.121	0.123
0.55	0.126	0.128	0.131	0.133	0.136	0.138	0.141	0.143	0.146	0.148
0.56	0.151	0.154	0.156	0.159	0.161	0.164	0.166	0.169	0.171	0.174
0.57	0.176	0.179	0.182	0.184	0.187	0.189	0.192	0.194	0.197	0.199
0.58	0.202	0.204	0.207	0.210	0.212	0.215	0.217	0.220	0.222	0.225
0.59	0.228	0.230	0.233	0.235	0.238	0.240	0.243	0.246	0.248	0.251
0.60	0.253	0.256	0.258	0.261	0.264	0.266	0.269	0.272	0.274	0.277
0.61	0.279	0.282	0.284	0.287	0.290	0.292	0.295	0.298	0.300	0.303
0.62	0.306	0.308	0.311	0.313	0.316	0.319	0.321	0.324	0.327	0.329
0.63	0.332	0.334	0.337	0.340	0.342	0.345	0.348	0.350	0.353	0.356
0.64	0.358	0.361	0.364	0.366	0.369	0.372	0.374	0.377	0.380	0.383
0.65	0.385	0.388	0.391	0.393	0.396	0.399	0.402	0.404	0.407	0.410
0.66	0.412	0.415	0.418	0.421	0.423	0.426	0.429	0.432	0.434	0.437
0.67	0.440	0.443	0.445	0.448	0.451	0.454	0.456	0.459	0.462	0.465
0.68	0.468	0.470	0.473	0.476	0.479	0.482	0.484	0.487	0.490	0.493
0.69	0.496	0.499	0.501	0.504	0.507	0.510	0.513	0.516	0.519	0.522
0.70	0.524	0.527	0.530	0.533	0.536	0.539	0.542	0.545	0.548	0.550
0.71	0.553	0.556	0.559	0.562	0.565	0.568	0.571	0.574	0.577	0.580
0.72	0.583	0.586	0.589	0.592	0.595	0.598	0.601	0.604	0.607	0.610
0.73	0.613	0.616	0.619	0.622	0.625	0.628	0.631	0.634	0.637	0.640
0.74	0.643	0.646	0.650	0.653	0.656	0.659	0.662	0.665	0.668	0.671

Tabelle B.5: Quantil z_p der Standardnormalverteilung

p	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
0.75	0.674	0.678	0.681	0.684	0.687	0.690	0.694	0.697	0.700	0.703
0.76	0.706	0.710	0.713	0.716	0.719	0.722	0.726	0.729	0.732	0.736
0.77	0.739	0.742	0.745	0.749	0.752	0.755	0.759	0.762	0.766	0.769
0.78	0.772	0.776	0.779	0.782	0.786	0.789	0.793	0.796	0.800	0.803
0.79	0.806	0.810	0.813	0.817	0.820	0.824	0.827	0.831	0.834	0.838
0.80	0.842	0.845	0.849	0.852	0.856	0.860	0.863	0.867	0.870	0.874
0.81	0.878	0.882	0.885	0.889	0.893	0.896	0.900	0.904	0.908	0.912
0.82	0.915	0.919	0.923	0.927	0.931	0.935	0.938	0.942	0.946	0.950
0.83	0.954	0.958	0.962	0.966	0.970	0.974	0.978	0.982	0.986	0.990
0.84	0.994	0.999	1.003	1.007	1.011	1.015	1.019	1.024	1.028	1.032
0.85	1.036	1.041	1.045	1.049	1.054	1.058	1.062	1.067	1.071	1.076
0.86	1.080	1.085	1.089	1.094	1.098	1.103	1.108	1.112	1.117	1.122
0.87	1.126	1.131	1.136	1.141	1.146	1.150	1.155	1.160	1.165	1.170
0.88	1.175	1.180	1.185	1.190	1.195	1.200	1.206	1.211	1.216	1.221
0.89	1.226	1.232	1.237	1.243	1.248	1.254	1.259	1.265	1.270	1.276
0.90	1.282	1.287	1.293	1.299	1.305	1.311	1.316	1.322	1.328	1.335
0.91	1.341	1.347	1.353	1.360	1.366	1.372	1.379	1.385	1.392	1.398
0.92	1.405	1.412	1.419	1.426	1.432	1.440	1.447	1.454	1.461	1.468
0.93	1.476	1.483	1.491	1.498	1.506	1.514	1.522	1.530	1.538	1.546
0.94	1.555	1.563	1.572	1.580	1.589	1.598	1.607	1.616	1.626	1.635
0.95	1.645	1.655	1.665	1.675	1.685	1.695	1.706	1.717	1.728	1.739
0.96	1.751	1.762	1.774	1.787	1.799	1.812	1.825	1.838	1.852	1.866
0.97	1.881	1.896	1.911	1.927	1.943	1.960	1.977	1.995	2.014	2.034
0.98	2.054	2.075	2.097	2.120	2.144	2.170	2.197	2.226	2.257	2.290
0.99	2.326	2.366	2.409	2.457	2.512	2.576	2.652	2.748	2.878	3.090

Tabelle B.6: Quantile der χ^2-Verteilung mit k Freiheitsgraden

k	$\chi^2_{k;0.025}$	$\chi^2_{k;0.05}$	$\chi^2_{k;0.1}$	$\chi^2_{k;0.9}$	$\chi^2_{k;0.95}$	$\chi^2_{k;0.975}$
1	0.001	0.004	0.016	2.706	3.841	5.024
2	0.051	0.103	0.211	4.605	5.991	7.378
3	0.216	0.352	0.584	6.251	7.815	9.348
4	0.484	0.711	1.064	7.779	9.488	11.143
5	0.831	1.145	1.610	9.236	11.070	12.833
6	1.237	1.635	2.204	10.645	12.592	14.449
7	1.690	2.167	2.833	12.017	14.067	16.013
8	2.180	2.733	3.490	13.362	15.507	17.535
9	2.700	3.325	4.168	14.684	16.919	19.023
10	3.247	3.940	4.865	15.987	18.307	20.483
11	3.816	4.575	5.578	17.275	19.675	21.920
12	4.404	5.226	6.304	18.549	21.026	23.337
13	5.009	5.892	7.042	19.812	22.362	24.736
14	5.629	6.571	7.790	21.064	23.685	26.119
15	6.262	7.261	8.547	22.307	24.996	27.488
16	6.908	7.962	9.312	23.542	26.296	28.845
17	7.564	8.672	10.085	24.769	27.587	30.191
18	8.231	9.390	10.865	25.989	28.869	31.526
19	8.907	10.117	11.651	27.204	30.144	32.852
20	9.591	10.851	12.443	28.412	31.410	34.170
21	10.283	11.591	13.240	29.615	32.671	35.479
22	10.982	12.338	14.041	30.813	33.924	36.781
23	11.689	13.091	14.848	32.007	35.172	38.076
24	12.401	13.848	15.659	33.196	36.415	39.364
25	13.120	14.611	16.473	34.382	37.652	40.646

Anhang C

Daten

Die Körpergröße von 179 Studierenden.

```
164 165 168 168 170 170 170 170 170 170 172 172 172 172 172
173 173 173 174 174 174 175 175 175 175 175 175 175 176 176
176 176 176 177 177 177 178 178 178 178 178 178 178 178 178
178 178 179 179 179 179 179 180 180 180 180 180 180 180 180
180 180 180 180 180 180 180 180 180 180 180 181 181 181 181
181 181 181 182 182 182 182 182 182 182 182 182 182 182 182
182 182 182 183 183 183 183 183 183 183 183 183 183 183 183
183 184 184 184 184 184 184 184 184 184 184 184 184 184 184
185 185 185 185 185 185 185 185 185 185 185 185 185 186 186
186 186 186 186 186 187 187 187 187 187 188 188 188 189 189
189 189 189 189 190 190 190 190 190 190 190 191 191 191 191
192 192 192 192 192 192 193 194 195 196 198 198 198 200
```

´

Literatur

Bickel, P. J. and K. A. Doksum (2001). *Mathematical statistics* (2 ed.). Upper Saddle River, NJ: Prentice Hall.

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variables*. New York: Wiley.

Bowley, A. L. (1920). *Elements of statistics*. New York: Charles Scribner's sons.

Brys, G., M. Hubert, and A. Struyf (2003). A comparison of some new measures of skewness. In R. Dutter, P. Filzmoser, U. Gather, and P. Rousseeuw (Eds.), *Developments in Robust Statistics (ICORS 2001)*, Heidelberg, pp. 98–113. Physika.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. London: Springer.

Conover, W. J. (1999). *Practical nonparametric statistics* (3 ed.). New York: Wiley.

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

de Haan, L. (1990). Fighting the arch-enemy with mathmatics. *Statistica neerlandica 44*, 45–68.

Dielman, T., C. Lowry, and R. Pfaffenberger (1994). A comparison of quantile estimators, communications in statistics. *Simulation and Computation 23*, 355–371.

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

Embrechts, P., C. Kloppelberg, and T. Mikosh (1995). *Modelling extremal events for insurance and finance*. Berlin: Springer.

Emerson, J. D. and M. A. Stoto (1982). Exploratory methods for choosing power transformation. *Journal of the American Statistical Association 77*, 103–108.

Everitt, B. S. (1987). *Introduction to optimization methods and their application in statistics*. Chapman and Hall.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematiks 44*, 423–453.

Gossett, W. S. (1908). The probable error of a mean. *Biometrika 6*, 1–25.

Hahn, G. J. (1970). Statistical intervals for a normal population: part 1, tables, examples, and applications. *Journal of Quality Technology 2*, 115–125.

Hahn, G. J. and W. Q. Meeker (1991). *Statistical intervals : a guide for practitioners* (1 ed.). New York: Wiley.

Handl, A. (1985). *Maßzahlen zur Klassifizierung von Verteilungen bei der Konstruktion adaptiver Tests im unverbundenen Zweistichproben-Problem*. Ph. D. thesis, Freie Universtität Berlin.

Harrell, F. and C. Davis (1982). A new distribution-free quantile estimator. *Biometrika 69*, 635–640.

Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society Civil Engineers 77*, 1539–1640.

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

Hinley, D. V. (1975). Pm power transformations to symmetry. *Bimoetrika 62*, 101–111.

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

Hosking, J. and J. Wallis (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics 29*, 339–349.

Hosking, J. R. M. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B 52*, 105–124.

Hyndman, R. and Y. Fan (1996). Sample quantiles in statistical packages. *The American Statistician 50*, 361–365.

Johnson, N., S. Kotz, and N. Balakrishnan (1994). *Continuous univariate distributions*. New York: Wiley.

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

Meister, R. (1984). *Ansätze zur Quantilschätzung*. Ph. D. thesis, Freie Universtität Berlin.

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

Moors, J. (1988). A quantile alternative for kurtosis. *The Statistician 37*, 25–32.

Naeve, P. (1995). *Stochastik f¸r Informatik* (1 ed.). M¸nchen: Oldenbourg.

Randles, R. H., M. A. Fligner, G. E. Policello, and D. A. Wolfe (1980). An asymptotically distribution-free test for symmetry vs. asymmertry. *Journal of the American Statistical Association 75*, 168–172.

Randles, R. H. and D. A. Wolfe (1979). *Introduction to the theory of nonparametric statistics* (1 ed.). New York: Wiley.

Rinne, H. (1997). *Taschenbuch der Statistik* (2 ed.). Thun: Deutsch.

Royston, P. (1992). Which measures of skewness and kurtosis are best? *Statistics in Medicine 11*, 333–343.

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