Quantile

Andreas Handl

Inhaltsverzeichnis

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:

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

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

$$x_p = F_X^{-1}(p) \tag{2}$$

In Abbildung 1 wird gezeigt, wie man das 0.8413-Quantil der Standardnormalverteilung bestimmen kann.

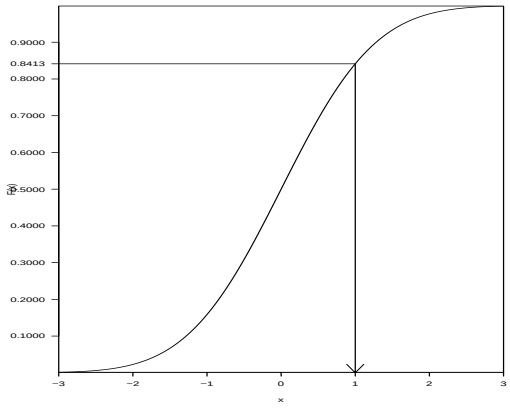

Abbildung 1: Bestimmung des 0.8413-Quantils der Standardnormalverteilung

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

$$x_p = \mu + z_p \, \sigma \tag{3}$$

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:

$$x_p = -\frac{1}{\lambda} \ln (1 - p) \tag{4}$$

Dies sieht man folgendermaßen:

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

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 2 zeigt dies.

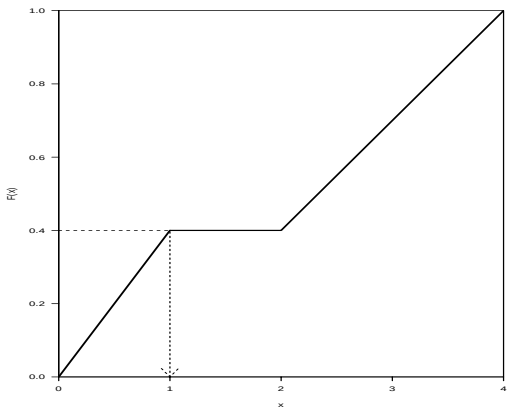

Abbildung 2: Bestimmung eines Quantils, falls die Verteilungsfunktion nicht streng monoton wachsend ist

Die folgende Definition umfasst beide Fälle

$$x_p = F^{\leftarrow}(p) = \inf\{x | F_X(x) \geq p\} \tag{5}$$

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 (4)) 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 (14) 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**

$$IQR = x_{0.75} - x_{0.25} \tag{6}$$

beschrieben. Eine Maßzahl für die **Schiefe** ist der **Koeffizient von Bowley** (siehe dazu (2)):

$$Q_S = \frac{(x_{0.75} - x_{0.5}) - (x_{0.5} - x_{0.25})}{x_{0.75} - x_{0.25}} \tag{7}$$

und eine Maßzahl für das Verhalten der Verteilung im Zentrum und an den Rändern ist das **Kurtosis-Maß von Moors** (siehe dazu (16)):

$$KM = \frac{(x_{0.875} - x_{0.625}) - (x_{0.375} - x_{0.125})}{x_{0.75} - x_{0.25}} \tag{8}$$

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

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.

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 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

$$F_X(x) = F\left(\frac{x-\theta}{\lambda}\right) \tag{9}$$

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

$$x_p = \theta + z_p\,\lambda \tag{10}$$

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:

$$x_p = \mu + z_p\,\sigma \tag{11}$$

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 (10) einsetzt:

$$\hat{x}_p = \hat{\theta} + z_p\,\hat{\lambda} \tag{12}$$

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

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{13}$$

und

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \tag{14}$$

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 (15)). Dabei ist

$$\Gamma(n) = \int_0^{\infty} x^{n-1}\,e^{-x}\,dx \tag{15}$$

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

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

siehe dazu Johnson und Kotz (15).
Tabelle 1 zeigt die Werte von a_n für $n = 6, \ldots, 12$.

Tabelle 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 28 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 (14) schlägt in seiner Arbeit noch andere Schätzer von μ und σ bei Normalverteilung vor, setzt sie in Gleichung 11 auf Seite 6 ein und vergleicht die resultierenden Quantilschätzer in einer Simulationsstudie.

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 2 zu finden.

Tabelle 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 3 zeigt die empirische Verteilungsfunktion.

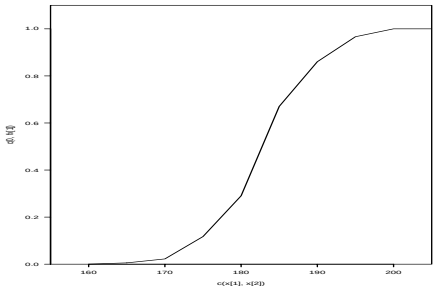

Abbildung 3: Empirische Verteilungsfunktion bei klassierten Daten

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

$$\hat{x}_p = x^*_{k-1} + \frac{p - \hat{F}(x^*_{k-1})}{h_k} \cdot \Delta_k \tag{16}$$

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

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

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

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 4 zeigt die empirische Verteilungsfunktion.

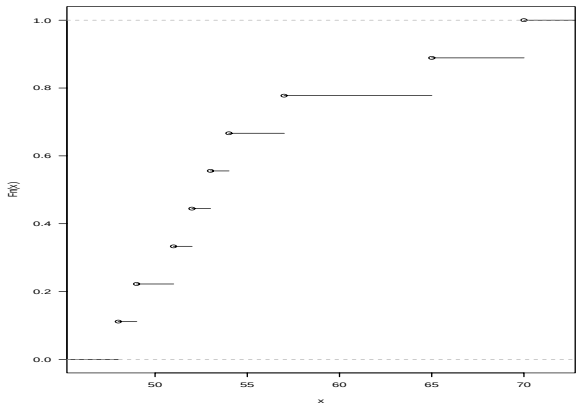

Abbildung 4: empirische Verteilungsfunktion

Die empirische Verteilungsfunktion ist eine Treppenfunktion, sodass ihre Inverse nicht eindeutig definiert ist. Es ist naheliegend, Quantile dadurch zu schätzen, dass man in Gleichung 5 auf Seite 4 $F_X(x)$ durch $F_n(x)$ ersetzt:

$$\hat{x}_p = \inf\{x | F_n(x) \geq p\} \tag{17}$$

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

$$\hat{x}_p = x_{(i)} \quad \text{für } \frac{i-1}{n} < p \leq \frac{i}{n} \tag{18}$$

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

Beispiel 5 (fortgesetzt)
Es gilt

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

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

$$\hat{x}_p = (1 - (np - i))x_{(i)} + (np - i)x_{(i+1)} \tag{19}$$

Mit $g = np - i$ erhalten wir also

$$\hat{x}_p = \begin{cases} x_{(1)} & \text{für} \quad p < \frac{1}{n} \\ (1-g)\, x_{(i)} + g\, x_{(i+1)} & \text{für} \quad \frac{1}{n} \leq p \leq 1 \end{cases} \tag{20}$$

Die Grafik links oben in Abbildung 5 zeigt die Approximation.

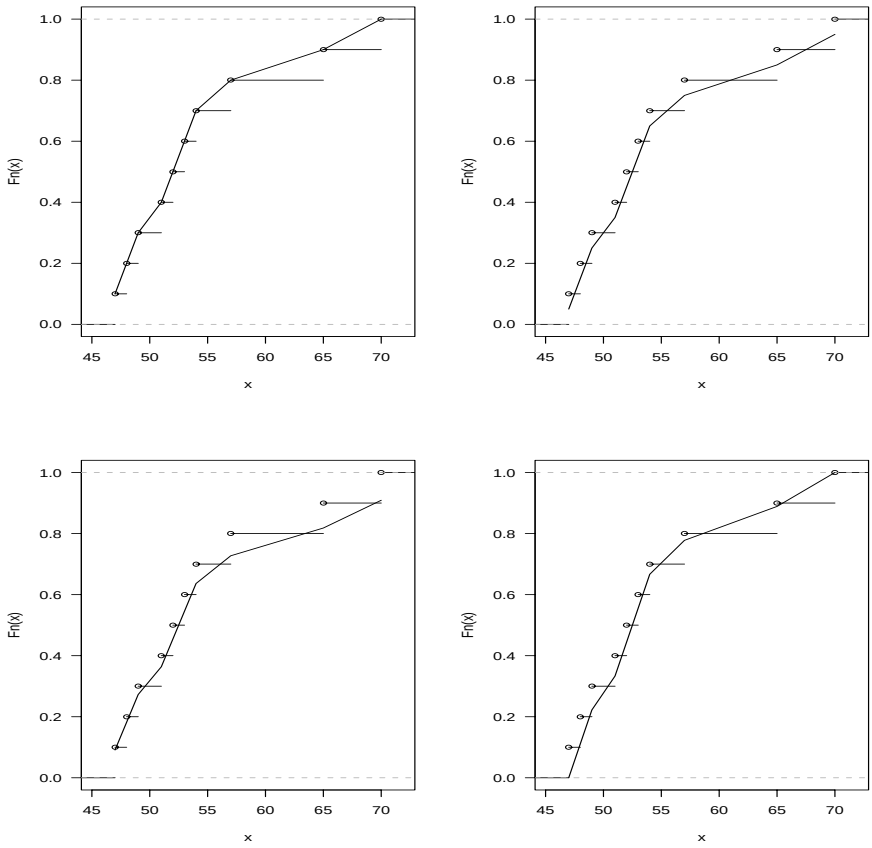

Abbildung 5: Vier Möglichkeiten, die empirische Verteilungsfunktion zu approximieren

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

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

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

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

Der Median wird aber immer folgendermaßen geschätzt:

$$\hat{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade ist} \\ \frac{x_{(n/2)} + x_{(1+n/2)}}{2} & \text{falls } n \text{ gerade ist} \end{cases} \tag{21}$$

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 5 zeigt die Approximation.
Als Quantilschätzer ergibt sich in diesem Fall:

$$\hat{x}_p = \begin{cases} x_{(1)} & \text{für} \quad p < \frac{0.5}{n} \\ (1-g)\, x_{(i)} + g\, x_{(i+1)} & \text{für} \quad \frac{0.5}{n} \leq p \leq \frac{n-0.5}{n} \\ x_{(n)} & \text{für} \quad p > \frac{n-0.5}{n} \end{cases} \tag{22}$$

mit $i = \lfloor np + 0.5 \rfloor$ und $g = np + 0.5 - i$. Dieser Schätzer wurde von Hazen (9) 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 und Wolfe (17), S. 7. Es gilt also

$$f_{X_{(i)}}(t) = \begin{cases} \dfrac{1}{B(i, n-i+1)}\, t^{i-1}\,(1-t)^{n-i} & \text{für } 0 < t < 1 \\ 0 & \text{sonst} \end{cases} \tag{23}$$

mit

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

Somit gilt

$$E(X_{(i)}) = \frac{i}{n+1} \tag{24}$$

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

$$t = \frac{i-1}{n-1} \tag{25}$$

Gleichung 24 auf Seite 14 legt folgende Approximation nahe:

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

Die Grafik links unten in Abbildung 5 zeigt die Approximation.
Als Quantilschätzer ergibt sich in diesem Fall:

$$\hat{x}_p = \begin{cases} x_{(1)} & \text{für} \quad p < \frac{1}{n+1} \\ (1-g)\, x_{(i)} + g\, x_{(i+1)} & \text{für} \quad \frac{1}{n+1} \leq p \leq \frac{n}{n+1} \\ x_{(n)} & \text{für} \quad p > \frac{n}{n+1} \end{cases} \tag{26}$$

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 25 auf Seite 15 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 5 zeigt die Approximation.
Als Quantilschätzer ergibt sich in diesem Fall:

$$\hat{x}_p = (1-g)\, x_{(i)} + g\, x_{(i+1)} \tag{27}$$

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:

$$\tilde{F}(x_{(i)}) = \frac{i - \gamma}{n + 1 - \gamma - \delta} \tag{28}$$

Die zugehörige Klasse von Quantilschätzern ist:

$$\hat{x}_p = \begin{cases} x_{(1)} & \text{für} \quad p < \frac{1-\gamma}{n+1-\gamma-\delta} \\ (1-g)\, x_{(i)} + g\, x_{(i+1)} & \text{für} \quad \frac{1-\gamma}{n+1-\gamma-\delta} \leq \frac{n-\gamma}{n+1-\gamma-\delta} \\ x_{(n)} & \text{für} \quad p > \frac{n-\gamma}{n+1-\gamma-\delta} \end{cases} \tag{29}$$

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 (20) auf Seite 11: $\gamma = 0$, $\delta = 1$
- Quantilschätzer in Gleichung (22) auf Seite 13: $\gamma = 0.5$, $\delta = 0.5$
- Quantilschätzer in Gleichung (26) auf Seite 15: $\gamma = 0$, $\delta = 0$
- Quantilschätzer in Gleichung (27) auf Seite 15: $\gamma = 1$, $\delta = 1$

Hyndman und Fan (13) 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 (1) wurde $\gamma = \delta = 3/8$.
Gilt $\gamma = \delta$, so wird der Median nach der Formel in Gleichung 21 auf Seite 13 geschätzt. Dies zeigen Hyndman und Fan (13).
Welchen dieser Schätzer sollte man verwenden? Hyndman und Fan (13) 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, Lowry und Pfaffenberger (5) und Handl (8) 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 (10)) 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

$$g_j(x) = \frac{1}{\beta(j, n+1-j)} F(x)^{j-1} (1 - F(x))^{n-j} \tag{30}$$

mit

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

(siehe dazu (17)). 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

$$E\left(X_{(j)}\right) = \frac{1}{\beta(j, n+1-j)} \int_0^1 F^{-1}(y) \, y^{j-1} (1-y))^{n-j} \, dy \tag{31}$$

Blom (1) 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 21 auf Seite 13 berechneten Wert identisch ist.

Welchen der Quantilschätzer soll man anwenden? Von Dielman, Lowry und Pfaffenberger (5) und Handl (8) 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 (29) auf Seite 16 mit $\gamma = 1/3$ und $\delta = 1/3$. Der Beweis ist bei Hoaglin et al (11) zu finden.

3 Schätzung extremer Quantile

Wie die Simulationsstudie von Dielman (5) 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, Klüppelberg und Mikosch (6) und Coles (3) 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:

$$F_U(x) = P(X - u \leq x | X > u) \tag{32}$$

Es gilt

$$F_U(x) = \frac{F_X(x+u) - F_X(u)}{1 - F_X(u)} \tag{33}$$

Dies sieht man folgendermaßen

$$\begin{aligned} F_U(x) &= P(X - u \leq x | X > u) = \frac{P(u < X \leq x + u)}{1 - P(X \leq u)} \\ &= \frac{F_X(x+u) - F_X(u)}{1 - F_X(u)} \end{aligned} \tag{34}$$

Schauen wir uns ein Beispiel an.

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

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases} \tag{35}$$

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 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

$$H(x) = \begin{cases} 1 - \left(1 + \frac{\xi\,x}{\beta}\right)^{-1/\xi} & \text{für } \xi \neq 0 \\ 1 - e^{-x/\beta} & \text{für } \xi = 0 \end{cases} \tag{36}$$

Eine Beweisskizze ist bei Coles (3), 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 (7) 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 3.1 auf Seite 21 zurück. Die Verteilung in Gleichung (36) heißt **verallgemeinerte Pareto-Verteilung**.
Abbildung 6 zeigt die Dichtefunktion der verallgemeinerten Pareto-Verteilung für unterschiedliche Werte von ξ für $\beta = 1$.

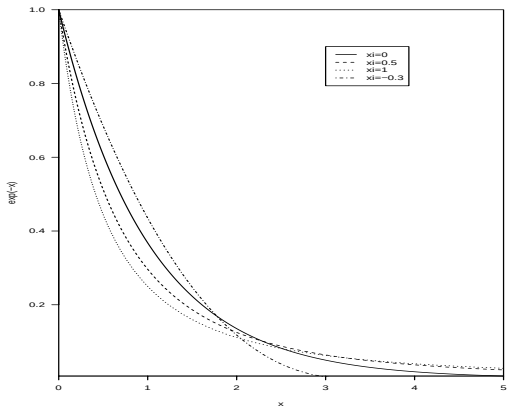

Abbildung 6: Dichtefunktion der verallgemeinerten Paretodiagramm

Wie können wir die Aussagen von Satz 3.1 auf Seite 21 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 (34) auf Seite 20 x durch $x-u$ und erhalten:

$$F_U(x-u) = \frac{F_X(x) - F_X(u)}{1 - F_X(u)} \tag{37}$$

Ersetzen wir $F_U(x-u)$ in Gleichung (37) durch $H(x-u)$ aus der ersten Gleichung in Gleichung (36) auf Seite 21, 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 (36) auf Seite 21 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

$$x_p = \begin{cases} u + \frac{\beta}{\xi}\left[\left(\frac{1-p}{1-F_X(u)}\right)^{-\xi} - 1\right] & \text{für } \xi \neq 0 \\ -\beta\,\ln\left(\frac{1-p}{1-F_X(u)}\right) & \text{für } \xi = 0 \end{cases} \tag{38}$$

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 und Wallis (12) zeigen, wie man hierbei vorzugehen hat.

3. Wir setzen die Schätzer in Gleichung 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:

$$\hat{x}_p = \begin{cases} u + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{n}{k} (1-p) \right)^{-\hat{\xi}} - 1 \right] & \text{für } \hat{\xi} \neq 0 \\ -\xi \ln \left(\frac{n}{k} (1-p) \right) & \text{für } \hat{\xi} = 0 \end{cases} \tag{39}$$

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 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 (6), 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 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.

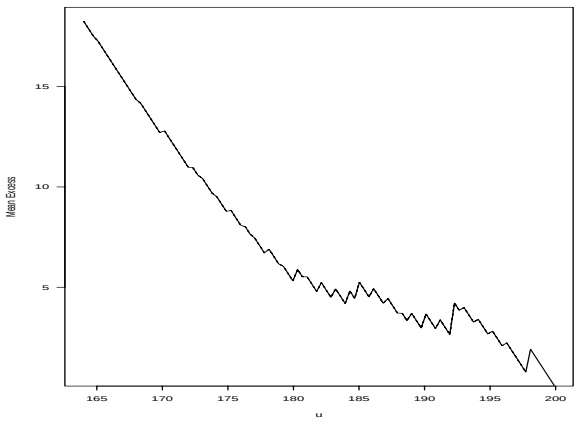

Abbildung 7: ME-Plot

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 (3) 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 8 zeigt die Parameterschätzer gegen den Schwellenwert.

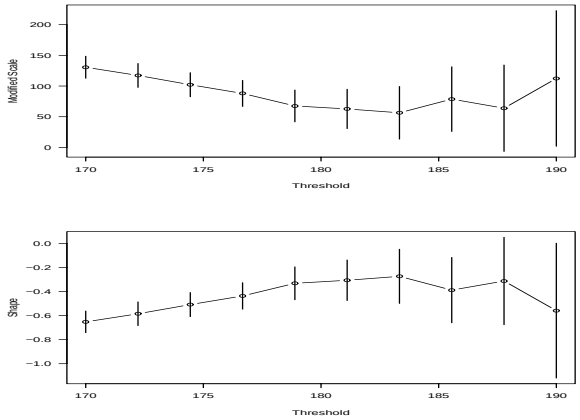

Abbildung 8: Parameterschätzer gegen den Schwellenwert

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 (3) 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 9 zeigt die Diagnostika.

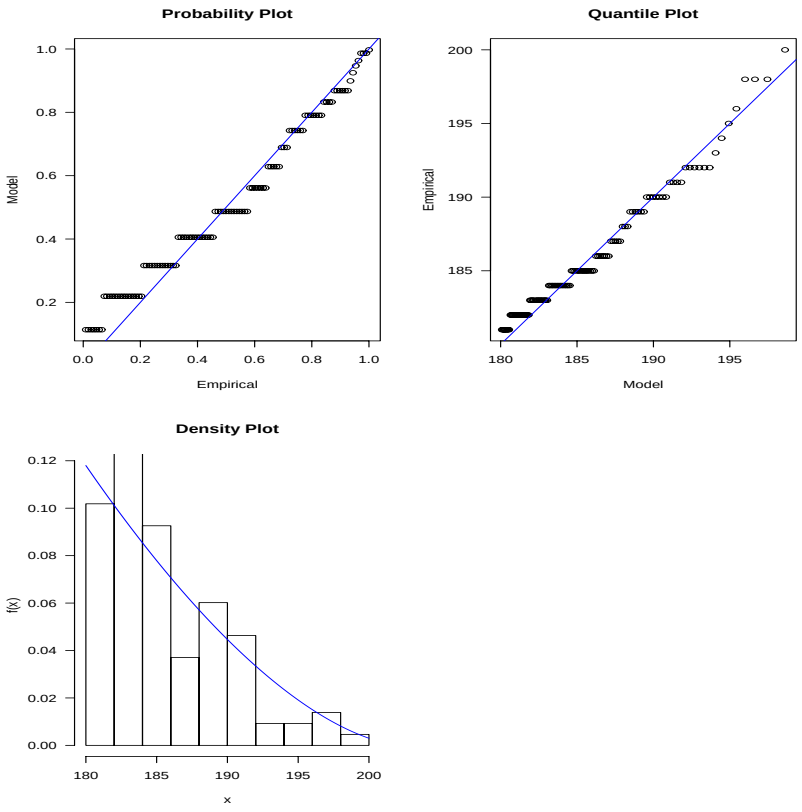

Abbildung 9: Diagnostika

Alle Zeichnungen deuten auf eine sehr gute Anpassung hin.

4 Anhang

4.1 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

[1] Blom, G. (1958): Statistical Estimates and Transformed Beta Variables. New York.

[2] Bowley, A.L. (1920): Elements of statistics. New York.

[3] Coles, S. (2001): An introduction to statistical modeling of extreme values. London.

[4] de Haan, L. (1990): Fighting the arch-enemy with mathmatics. Statistica neerlandica, 45-68

[5] Dielman, T.E. C. Lowry, R. Pfaffenberger (1994): A comparison of quantile estimators, Communications in Statistics - Simulation and Computation 23, 355-371.

[6] Embrechts, P. ,C. Klüppelberg, Th. Mikosch (1995): Modelling extremal events for insurance and finance . Berlin.

[7] Gnedenko, B.V. (1943), Sur la distribution limite du terme maximum d'une serie aleatoire. Annals of Mathematics, 44, 423-453.

[8] Handl, A. (1985): Verteilungsfreie Quantilschätzer - ein Vergleich. Diskussionsarbeit Nr. 6/85, Institut für Quantitative Ökonomik und Statistik, Freie Universität Berlin

[9] Hazen, A.(1914), Storage to be provided in impounding reservoirs for municipal water supply, Transactions of the American Society Civil Engineers, 77, 1539 -1640.

[10] Harrell F.E., C.E. Davis (1982): A new distribution-free quantile estimator. Biometrika 69:635-640.

[11] Hoaglin, D.C., F. Mosteller, J.W. Tukey (1983): Understanding robust and exploratory data analysis. New York.

[12] Hosking J.R.M., J.R. Wallis (1987); Parameter and quantile estimation for the generalized Pareto distribution, Technometrics 29, 339-349.

[13] Hyndman, R.J., Y. Fan (1996): Sample quantiles in statistical packages. American Statistician, 50 361-365.

[14] Meister, R. (1984): Ansätze zur Quantilschätzung. Dissertation, FU Berlin.

[15] Johnson, N.L. , S. Kotz , N. Balakrishnan (1994): Continuous univariate distributions. New York.

[16] Moors, J.J.A. (1988): A quantile alternative for kurtosis. The Statistician, 37, 25-32

[17] Randles, R.H., D.A. Wolfe (1979): Introduction to the theory of nonparametric statistics. New York.

[18] Reiss, R.D., M. Thomas (2001): Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. 2. Auflage