

# 4 Statistik der Extremwertverteilungen

In diesem Kapitel beschäftigen wir uns mit statistischen Anwendungen der Extremwerttheorie. Wir werden zwei verschiedene Zugänge zur Modellierung von Extremwerten betrachten. Der erste Zugang basiert auf den bereits bekannten Extremwertverteilungen, die hier GEV-Verteilungen (Generalized Extreme-Value Distributions) genannt werden. Der zweite Zugang (Peaks Over Threshold-Methode) benutzt die verallgemeinerten Pareto-Verteilungen (GPD, Generalized Pareto Distributions).

## 4.1 Statistik der GEV-Verteilungen

Wir haben bisher gesehen, dass Extremwertverteilungen folgende Form haben:

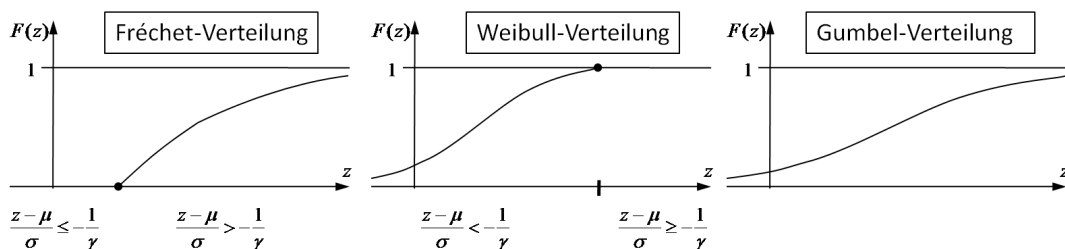
$$G_{\gamma,\mu,\sigma}(z) = \exp \left\{ - \left( 1 + \gamma \frac{z - \mu}{\sigma} \right)^{-\frac{1}{\gamma}} \right\} \text{ für } 1 + \gamma \frac{z - \mu}{\sigma} > 0.$$

Extremwertverteilungen bilden also eine dreiparametrische Familie:  $\gamma \in \mathbb{R}$  ist der formgebende Parameter,  $\mu \in \mathbb{R}$  ist der Lageparameter und  $\sigma > 0$  ist der Skalenparameter. Für  $\gamma$  gilt:

$\gamma > 0$ :  $G$  ist eine Fréchet-Verteilung (definiert für  $\frac{z-\mu}{\sigma} > -\frac{1}{\gamma}$  wie oben, sonst 0).

$\gamma < 0$ :  $G$  ist eine Weibull-Verteilung (definiert für  $\frac{z-\mu}{\sigma} < -\frac{1}{\gamma}$  wie oben, sonst 1).

$\gamma = 0$ :  $G$  ist eine Gumbel-Verteilung (definiert für  $z \in \mathbb{R}$  wie oben).



Extremwertverteilungen, die in der obigen Form dargestellt werden, werden auch General Extreme-Value (GEV) Verteilungen genannt.

**Beispiel 4.1** (Wasserstände an einem Deich). Am Tag  $j \in \{1, \dots, 365\}$  im Jahr  $i \in \{1, \dots, n\}$  wurde an einem Deich der Wasserstand  $X_{ij}$  gemessen. Wir betrachten die jährlichen Maxima  $X_i = \max_{j=1, \dots, 365} X_{ij}$  und wollen aus diesen Daten die Deichhöhe  $Z_p$  bestimmen, bei der eine Überflutung mit einer gegebenen Wahrscheinlichkeit  $p$  in einem Jahr stattfindet. Dabei ist die Wahrscheinlichkeit  $p$  sehr klein (viel kleiner als  $1/n$ , zum Beispiel), so dass alle gemessenen Wasserstände sicherlich viel kleiner als die gesuchte Höhe  $Z_p$  sind.

Dazu betrachten wir folgendes Modell:  $X_1, \dots, X_n$  sind u.i.v. Zufallsvariablen mit einer GEV-Verteilung  $G_{(\gamma, \mu, \sigma)}$  mit Parametervektor  $\theta = (\gamma, \mu, \sigma) \in \mathbb{R}^3$ .

**Bemerkung 4.2.** Den jährlichen Maxima eine GEV-Verteilung zu unterstellen, ist eine natürliche Wahl, da jedes  $X_i$  ein Maximum von vielen u.i.v. Zufallsvariablen ist. Wir haben in früheren Kapiteln gezeigt, dass solche Maxima unter sehr allgemeinen Bedingungen gegen Extremwertverteilungen konvergieren. Natürlich braucht man für die Konvergenz Normierungskonstanten, in unserem Fall kann man aber annehmen, dass die Normierungskonstanten bereits in den Parametern  $\mu$  und  $\sigma$  enthalten sind.

**Bemerkung 4.3.** Da wir im obigen Modell voraussetzen, dass die  $X_i$  identisch verteilt sind, kann das Modell nur auf stationäre Daten angewendet, d.h. Daten, die keinen Trend aufweisen. Werden die jährlichen Maxima mit der Zeit immer größer (kleiner), muss ein anderes Modell verwendet werden.

Unser Problem besteht nun darin, den Parametervektor  $\theta$  zu schätzen. Wir werden die Maximum-Likelihood (ML) Methode benutzen. Dazu benötigt man die Dichte  $f_{(\gamma, \mu, \sigma)}(z)$  der GEV-Verteilung. Durch Ableiten der Verteilungsfunktion  $G_{(\gamma, \mu, \sigma)}$  erhält man, dass für  $\gamma \neq 0$  gilt:

$$f_\theta(z) = f_{(\gamma, \mu, \sigma)}(z) = \begin{cases} \frac{1}{\sigma} (1 + \gamma \frac{z-\mu}{\sigma})^{-\frac{1}{\gamma}-1} \exp\{- (1 + \gamma \frac{z-\mu}{\sigma})^{-\frac{1}{\gamma}}\}, & 1 + \gamma \frac{z-\mu}{\sigma} > 0, \\ 0, & \text{sonst} \end{cases}$$

und für  $\gamma = 0$  gilt:

$$f_\theta(z) = f_{(0, \mu, \sigma)} = \frac{1}{\sigma} e^{-\frac{z-\mu}{\sigma}} \exp\left\{-e^{-\frac{z-\mu}{\sigma}}\right\} \text{ für } z \in \mathbb{R}.$$

Mit Hilfe der Dichten kann man die Log-Likelihoodfunktion

$$l(\theta) := l(\theta|x_1, \dots, x_n) := \sum_{i=1}^n \log f_\theta(x_i)$$

bestimmen. Für  $\gamma \neq 0$  gilt:

$$l(\theta) = -n \log \sigma - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^n \log \left(1 + \gamma \frac{x_i - \mu}{\sigma}\right) - \sum_{i=1}^n \left(1 + \gamma \frac{x_i - \mu}{\sigma}\right)^{-\frac{1}{\gamma}}$$

falls  $1 + \gamma \frac{x_i - \mu}{\sigma} > 0$  für alle  $i = 1, \dots, n$  und  $l(\theta) = -\infty$  sonst. Für  $\gamma = 0$  gilt:

$$l(\theta) = -n \log \sigma - \sum_{i=1}^n \frac{x_i - \mu}{\sigma} - \sum_{i=1}^n e^{-\frac{x_i - \mu}{\sigma}}.$$

Mit der log-Likelihoodfunktion lässt sich der Maximum-Likelihood Schätzer

$$\hat{\theta} = (\hat{\gamma}, \hat{\mu}, \hat{\sigma}) = \operatorname{argmax} l(\gamma, \mu, \sigma)$$

herleiten. Hier kann  $\hat{\theta}$  nicht analytisch bestimmt, sondern muss numerisch ermittelt werden.

Nachdem der Parameter  $\theta$  geschätzt wurde, können wir die Deichhöhe  $Z_p$  schätzen. Wir erinnern, dass  $Z_p$  die Deichhöhe ist, bei der eine Überflutung mit Wahrscheinlichkeit  $p$  in einem Jahr stattfindet. Das Problem besteht also darin, dass  $p$ -Quantil des jährlichen Maximums zu schätzen. Wir schätzen  $Z_p$  indem wir die Gleichung

$$G_{(\hat{\gamma}, \hat{\mu}, \hat{\sigma})}(\hat{Z}_p) = 1 - p$$

lösen (falls es mehrere Lösungen gibt, betrachten wir die kleinste):

$$\hat{Z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\gamma}} \{1 - (-\log(1 - p))\}^{\frac{1}{\hat{\gamma}}}, & \hat{\gamma} \neq 0, \\ \hat{\mu} - \hat{\sigma} \log(-\log(1 - p)), & \hat{\gamma} = 0. \end{cases}$$

Für  $\hat{\gamma} < 0$  (im Fall der Weibull-Verteilung) besitzt die Verteilung  $G_{(\hat{\gamma}, \hat{\mu}, \hat{\sigma})}$  einen endlichen rechten Endpunkt, der übrigens per Definition  $Z_0$  ist. In diesem Fall gehen wir davon aus, dass es einen absolut höchsten Wasserstand gibt, der niemals überschritten wird. Der Schätzer für  $Z_0$  ist dann gegeben durch:

$$\hat{Z}_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\gamma}}.$$

Nachdem nun das Problem gelöst wurde, stellt sich die Frage, wie wir die Lösung verifizieren können. Wie können wir überprüfen, ob die Daten  $X_1, \dots, X_n$  durch die Verteilungsfunktion  $\hat{G} = G_{(\hat{\gamma}, \hat{\mu}, \hat{\sigma})}$  tatsächlich gut beschrieben werden? Zur Verifikation des Modells gibt es mehrere Methoden, die wir im Folgenden betrachten.

Ordnen wir die Stichprobe  $X_1, \dots, X_n$  monoton aufsteigend an, so erhalten wir die Ordnungsstatistiken

$$X_{(1)} \leq \dots \leq X_{(n)}.$$

**Definition 4.4** (Probability Plot). *Der PP-Plot ist die Menge*

$$\left\{ \left( \hat{G}(X_{(i)}), \frac{i}{n+1} \right) : i = 1, \dots, n \right\} \subset [0, 1]^2.$$

Trifft die Annahme, dass die  $X_i$  gemäß  $\hat{G}$  verteilt sind, zu, so sollte  $\hat{G}(X_{(i)}) \approx \frac{i}{n+1}$  gelten bzw. sollten die Punkte in einem Probability Plot auf der Winkelhalbierenden liegen (etwa wie in Grafik 4.1).

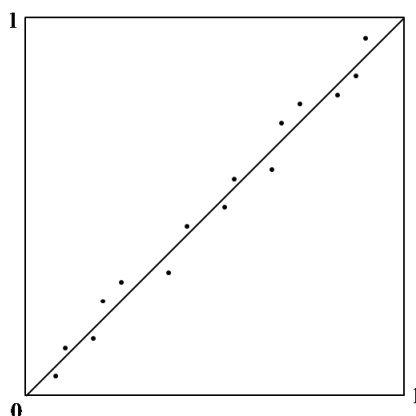


Abbildung 4.1: PP-Plot

Der PP-Plot besitzt einen Nachteil: Für  $i \approx n$  gilt  $\hat{G}(X_{(i)}) \approx 1$  und  $\frac{i}{n+1} \approx 1$  egal ob  $\hat{G}$  die Daten gut beschreibt oder nicht. Mit anderen Worten, auch wenn  $\hat{G}$  die Daten im Bereich der großen Werte nicht gut beschreibt, sieht man das in einem Probability Plot möglicherweise nicht. Dabei sind gerade die großen Werte besonders interessant für uns. Wir betrachten deshalb eine andere Methode, die dieser Überlegung Rechnung trägt. Das  $q$ -Quantil  $\hat{G}^{\leftarrow}(q)$  der Verteilung  $\hat{G}$  ist definiert als (die kleinste) Lösung  $z$  der Gleichung

$$\hat{G}(z) = q.$$

**Definition 4.5** (Quantil Plot). *Der QQ-Plot ist die Menge*

$$\left\{ \left( \hat{G}^{\leftarrow} \left( \frac{i}{n+1} \right), X_{(i)} \right), i = 1, \dots, n \right\} \in \mathbb{R}^2.$$

Beim QQ-Plot werden auf der  $x$ -Achse also die  $\frac{1}{n+1}, \frac{2}{n+1}, \dots, \frac{n}{n+1}$ -Quantile der Verteilung  $\hat{G}$  abgetragen und auf der  $y$ -Achse die Ordnungsstatistiken  $X_{(1)}, \dots, X_{(n)}$ . Wenn die  $\hat{G}$  die Daten gut beschreibt, sollte  $\hat{G}^{\leftarrow} \left( \frac{i}{n+1} \right) \approx X_{(i)}$  gelten bzw. sollten die Punkte in Grafik 4.2 auf der Winkelhalbierenden liegen.

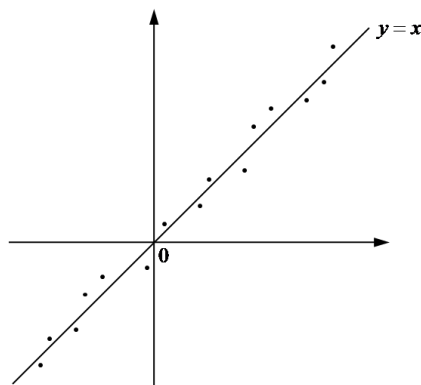


Abbildung 4.2: QQ-Plot

Es kann vorkommen, dass die Daten  $X_1, \dots, X_n$  einen Trend aufweisen (z.B. werden die jährlichen Maxima höher). Wir betrachten nun ein Modell, dass der Nichtstationarität der Daten Rechnung trägt. Es seien  $X_1, \dots, X_n$  unabhängig aber nicht identisch verteilt mit

$$X_i \sim G_{(\gamma(i), \sigma(i), \mu(i))}, \quad i = 1, \dots, n.$$

Dabei ist der Parametervektor  $(\gamma(i), \sigma(i), \mu(i))$  eine Funktion der Zeit  $i$ . Für diese Funktion kann man z.B. den folgenden Ansatz verwenden:

$$\gamma(i) = \gamma, \quad \sigma(i) = \sigma, \quad \mu(i) = \beta_0 + \beta_1 \cdot i.$$

Wir gehen also von einem konstanten Formparameter  $\gamma$ , einem konstanten Skalenparameter  $\sigma$  und einem linearen Trend, der im Lageparameter  $\mu$  berücksichtigt wird, aus. Die Parameter  $(\gamma, \sigma, \beta_0, \beta_1)$  lassen sich wieder mit der ML-Methode schätzen und somit lässt sich das Problem mit den bereits im Fall von stationären Daten betrachteten Methoden lösen. Möchte man das Modell verifizieren, so kann man PP- oder QQ-Plots erstellen. Davor muss man aber die Stichprobe  $X_1, \dots, X_n$  von dem Trend bereinigen:

$$X'_i := (X_i - \hat{\beta}_0) / (\hat{\beta}_1 i).$$

Die bereinigte Stichprobe  $X'_1, \dots, X'_n$  sollte man dann mit der Verteilungsfunktion  $G_{\hat{\gamma}, 0, \hat{\sigma}}$  vergleichen.

**Bemerkung 4.6.** Der Ansatz kann verallgemeinert werden, ohne dass sich das Modell grundsätzlich ändert. So ist es zum Beispiel problemlos möglich, einen exponentiellen Trend zu modellieren:

$$\gamma(i) = \gamma, \quad \sigma(i) = \sigma, \quad \mu(i) = e^{\beta_0 + \beta_1 i}.$$

## 4.2 Statistik der GP-Verteilungen

Die oben beschriebene Methode basiert auf der Betrachtung von Block-Maxima (z.B. von jährlichen Maxima). Es gibt eine andere Methode (Peaks over Threshold), bei der man nur Beobachtungen berücksichtigt, die einen Schwellenwert überschreiten. Im Folgenden beschäftigen wir uns mit dieser Methode.

Wir fangen damit an, dass wir die verallgemeinerten Pareto-Verteilungen definieren. Es sei  $X$  eine Zufallsvariable, die man sich z.B. als eine Schadenhöhe vorstellen kann. Wir interessieren nur für die großen Werte von  $X$ . Wie ist der sogenannte Exzess  $X - u$  verteilt, gegeben dass  $X > u$ ? Dabei geht  $u \rightarrow \infty$ . Wir betrachten drei Beispiele.

**Beispiel 4.7.** Sei  $X$  exponentialverteilt mit Parameter  $\lambda > 0$ , d.h.  $\bar{F}(t) = e^{-\lambda t}$ ,  $t > 0$ . Dann gilt

$$\mathbb{P}[X - u > t | X > u] = \frac{\mathbb{P}[X > u + t, X > u]}{\mathbb{P}[X > u]} = \frac{\mathbb{P}[X > u + t]}{\mathbb{P}[X > u]} = \frac{e^{-\lambda(u+t)}}{e^{-\lambda u}} = e^{-\lambda t}.$$

Es gilt also: die bedingte Verteilung von  $X - u$  gegeben, dass  $X > u$ , ist die Exponentialverteilung mit Parameter  $\lambda$ . Dies ist die Gedächtnislosigkeit der Exponentialverteilung.

**Beispiel 4.8.** Sei  $X$  aus dem Max-Anziehungsbereich der Frechet-Verteilung  $\Phi_\alpha$ ,  $\alpha > 0$ . D.h.,  $\bar{F} \in RV_{-\alpha}$ . Dann gilt für alle  $t > 0$ :

$$\mathbb{P}\left[\frac{X - u}{u} > t | X > u\right] = \frac{\mathbb{P}[X > u + ut]}{\mathbb{P}[X > u]} = \frac{\bar{F}(u(t+1))}{\bar{F}(u)} \rightarrow (1+t)^{-\alpha}$$

für  $u \rightarrow \infty$ . Es gilt also: gegeben, dass  $X > u$ , konvergiert die Verteilung von  $(X - u)/u$  gegen die Verteilungsfunktion  $1 - (1+t)^{-\alpha}$ ,  $t > 0$ .

**Beispiel 4.9.** Sei  $X$  standardnormalverteilt. Folgende Relation kann mit der Regel von L'Hospital bewiesen werden:

$$\mathbb{P}[X > u] \sim \frac{1}{\sqrt{2\pi}u} e^{-u^2/2} \text{ für } u \rightarrow \infty.$$

Unter Verwendung dieser Relation erhalten wir für jedes  $t > 0$ :

$$\mathbb{P}[u(X - u) > t | X > u] = \frac{\mathbb{P}[X > u + \frac{t}{u}]}{\mathbb{P}[X > u]} \sim \frac{\exp\{-\frac{-u^2}{2} - t - \frac{t^2}{2u^2}\}}{\exp\{-\frac{u^2}{2}\}} \rightarrow e^{-t}$$

für  $u \rightarrow \infty$ . Es gilt also: gegeben, dass  $X > u$ , konvergiert die Verteilung von  $u(X - u)$  gegen die Exponentialverteilung mit Parameter 1.

In allen drei Beispielen konnte die bedingte Verteilung von  $X - u$  gegeben, dass  $X > u$ , durch eine Verteilung approximiert werden. Wir werden nun ein allgemeines Resultat

formulieren, das die drei Beispiele als Spezialfälle beinhaltet. Die Verteilungsfunktion

$$P_{\gamma,\sigma}(t) = 1 - \left(1 + \frac{\gamma t}{\sigma}\right)^{-\frac{1}{\gamma}}$$

heißt die verallgemeinerte Pareto-Verteilung (GPD, Generalized Pareto Distribution) mit Index  $\gamma \in \mathbb{R}$  und Skalenparameter  $\sigma > 0$ . In dieser Formel ist  $t > 0$  falls  $\gamma > 0$  und  $t \in [0, -\frac{\sigma}{\gamma}]$  falls  $\gamma < 0$ . Für  $\gamma = 0$  interpretieren wir die Formel als Grenzwert:

$$P_{0,\sigma}(t) = \lim_{\gamma \rightarrow 0} \left(1 - \left(1 + \frac{\gamma t}{\sigma}\right)^{-\frac{1}{\gamma}}\right) = 1 - e^{-t/\sigma}, \quad t > 0.$$

Somit stimmt  $P_{0,\sigma}$  mit der Exponentialverteilung mit Parameter  $1/\sigma$  überein.

**Theorem 4.10** (Pickands–Balkema–de Haan). *Sei  $X$  eine Zufallsvariable mit Verteilungsfunktion  $F$ , die im rechten Endpunkt  $x^*$  stetig ist. Dann liegt  $F$  im Max-Anziehungsbereich von  $G_\gamma$  genau dann, wenn es eine positive messbare Funktion  $\beta(u)$  gibt mit*

$$\lim_{u \uparrow x^*} \sup_{t \in [0, x^* - u]} |\mathbb{P}[X - u \leq t | X > u] - P_{\gamma, \beta(u)}(t)| = 0.$$

Grob gesagt, es gilt die Approximation  $\mathbb{P}[X - u \leq t | X > u] \approx P_{\gamma, \beta(u)}(t)$  falls  $X$  im Max-Anziehungsbereich von  $G_\gamma$  liegt.

Nun werden wir die GP-Verteilungen in der Statistik anwenden. Es seien  $X_1, \dots, X_n$  unabhängige identisch verteilte Beobachtungen, z.B. Wasserstände an einem Deich an  $n$  Tagen. Wir interessieren uns nur für die extrem großen Beobachtungen. Das heißt, wir wählen einen Schwellenwert  $u$  und betrachten nur die Beobachtungen  $X_{i_1}, \dots, X_{i_k}$ , die  $u$  überschreiten. Wir definieren die Exzesse

$$Y_1 = X_{i_1} - u, \dots, Y_k = X_{i_k} - u$$

und ignorieren alle anderen Daten. Der Satz von Pickands–Balkema–de Haan macht folgendes Modell plausibel: die Exzesse  $Y_1, \dots, Y_k$  sind unabhängige identisch verteilte Zufallsvariablen, die gemäß einer verallgemeinerten Pareto-Verteilung  $P_{\gamma,\sigma}$  verteilt sind. Dabei sind  $\gamma \in \mathbb{R}$  und  $\sigma > 0$  unbekannte Parameter. Die Dichte von  $P_{\gamma,\sigma}$  ist gegeben durch

$$f_{\gamma,\sigma}(t) = \frac{1}{\sigma} \left(1 + \frac{\gamma t}{\sigma}\right)^{-\frac{1}{\gamma}-1}.$$

Damit ergibt sich für die Log-Likelihoodfunktion

$$l(\gamma, \sigma) := \sum_{i=1}^k \log f_{\gamma, \sigma}(y_i) = -k \log \sigma - \left(1 + \frac{1}{\gamma}\right) \sum_{i=1}^k \log \left(1 + \frac{\gamma y_i}{\sigma}\right),$$

zumindest für  $\gamma \neq 0$ . Der ML-Schätzer:  $(\hat{\gamma}, \hat{\sigma}) = \operatorname{argmax}_{\gamma, \sigma} l(\gamma, \sigma)$  muss numerisch berechnet werden.

Nun werden wir für ein gegebenes kleines  $p$  die Deichhöhe  $Z_p$  schätzen, die an einem Tag mit Wahrscheinlichkeit  $p$  überflutet wird. Es sei  $X$  die Zufallsvariable, die den Wasserstand an einem Tag beschreibt. Mit Berücksichtigung des Satzes von Pickands–Balkema–de Haan gehen wir davon aus, dass für großes  $u$ :

$$\mathbb{P}[X - u > t | X > u] \approx \left(1 + \frac{\gamma t}{\sigma}\right)^{-\frac{1}{\gamma}}.$$

Mit  $t = Z_p - u$  folgt also:

$$\mathbb{P}[X > Z_p] \approx \mathbb{P}[X > u] \left(1 + \gamma \frac{Z_p - u}{\sigma}\right)^{-\frac{1}{\gamma}}.$$

Nun setzen wir die rechte Seite gleich  $p$ . Wenn man die Gleichung nach  $Z_p$  umstellt, erhält man schließlich

$$Z_p = u + \frac{\sigma}{\gamma} \left[ \left( \frac{\mathbb{P}[X > u]}{p} \right)^{\gamma} - 1 \right].$$

Dabei haben wir Schätzer für  $\sigma$  und  $\gamma$  bereits hergeleitet. Die Wahrscheinlichkeit  $\mathbb{P}[X > u]$  kann durch  $\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i > u}$  geschätzt werden.