KAPITEL 6

Statistik der Extremwertverteilungen

In diesem Kapitel beschäftigen wir uns mit statistischen Anwendungen der Extremwertverteilungen. Wir werden zwei verschiedene Zugänge zur Modellierung von Extremwerten betrachten.

- Der erste Zugang basiert auf der Modellierung von *Blockmaxima* durch die bereits bekannten Extremwertverteilungen, die hier GEV–Verteilungen (Generalized Extreme–Value Distributions) genannt werden.
- Der zweite Zugang (*Peaks Over Threshold Method*) benutzt die verallgemeinerten Pareto–Verteilungen (GPD, Generalized Pareto Distributions).

Wir werden hier nur auf einige grundlegende Ideen der statistischen Modellierung von Extremwerten eingehen. Für mehr Einzelheiten verweisen wir auf die Bücher von S. Coles "*An introduction to statistical modeling of extreme values*", E. Gumbel "*Statistics of extremes*", J. Beirlant, Y. Goegebeur, J. Teugels, J. Segers "*Statistics of extremes*".

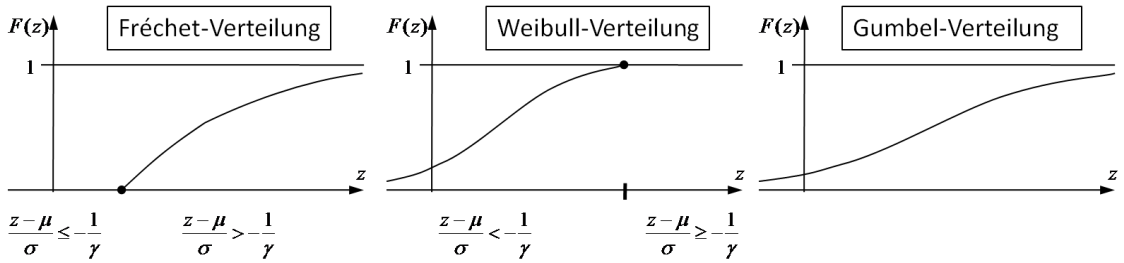
6.1. Statistik der Blockmaxima: GEV–Verteilungen

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 dreiparametrige Familie: $\gamma \in \mathbb{R}$ ist der formgebende Parameter, $\mu \in \mathbb{R}$ ist der Lageparameter und $\sigma > 0$ ist der Skalenparameter. Für γ 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 distributions (GEV–Verteilungen) genannt.

Beispiel 6.1.1 (Wasserstände an einem Deich)**.** Am Tag $j \in \{1, \ldots, 365\}$ im Jahr $i \in \{1, \ldots, n\}$ wurde an einem Deich der Wasserstand x_{ij} gemessen. Wir betrachten die jährlichen Maxima ("Blockmaxima")

$$x_i = \max_{j=1,\ldots,365} x_{ij}$$

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: $x_1, \ldots, x_n$ sind Realisierungen von $X_1, \ldots, X_n$, die u.i.v. Zuvallsvariablen mit einer GEV–Verteilung $G_{\gamma,\mu,\sigma}$ mit Parametervektor $\theta = (\gamma, \mu, \sigma) \in \mathbb{R}^2 \times \mathbb{R}_+$ sind.

Bemerkung 6.1.2. 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 μ und σ enthalten sind.

Bemerkung 6.1.3. Da wir im obigen Modell voraussetzen, dass die X_i identisch verteilt sind, kann das Modell nur auf stationäre Daten angewendet werden, d.h. Daten, die keinen Trend aufweisen. Werden die jährlichen Maxima mit der Zeit immer größer (kleiner), muss ein anderes Modell verwendet werden, siehe unten.

Unser Problem besteht nun darin, den Parametervektor θ zu schätzen. Wir werden die *Maximum–Likelihood–Methode* (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 $\gamma \neq 0$

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

während für $\gamma = 0$

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

Mit Hilfe der Dichten kann man die Log–Likelihoodfunktion aufstellen:

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

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 $i=1,...,n$, und $l(\theta)=-\infty$ sonst. Für $\gamma=0$ gilt:

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

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 werden, sondern muss numerisch ermittelt werden.

Nachdem der Parameter θ 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 $(1-p)$–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 $x_1,\ldots,x_n$ 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 $x_1,\ldots,x_n$ monoton aufsteigend an, so erhalten wir die Ordnungsstatistiken

$$x_{(1)}\leq ...\leq x_{(n)}.$$

Definition 6.1.4 (Probability–Plot)**.** Der **PP–Plot** ist die Menge

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

Trifft die Annahme, dass die Daten $x_1,\ldots,x_n$ 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 1).

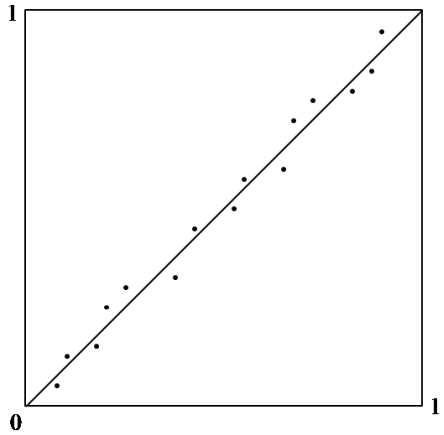


Abbildung 1. PP–Plot

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 $\hat{G}^{\leftarrow}(q)$, wobei $q \in (0,1)$, einer Verteilungsfunktion $\hat{G}$ ist definiert als (die kleinste) Lösung z der Gleichung

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

Definition 6.1.5 (Quantil–Plot)**.** Der **QQ–Plot** ist die Menge

$$\left\{ \left(\hat{G}^{\leftarrow}\left(\frac{i}{n+1} \right), x_{(i)} \right) : i = 1, \ldots, n \right\} \subset \mathbb{R}^2.$$

Beim QQ–Plot werden auf der horizontalen Achse die $\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$-Quantile der Verteilung $\hat{G}$ abgetragen und auf der vertikalen Achse die Ordnungsstatistiken $x_{(1)}, \ldots, x_{(n)}$. Wenn $\hat{G}$ die Daten gut beschreibt, sollte

$$\hat{G}^{\leftarrow}\left(\frac{i}{n+1} \right) \approx x_{(i)}$$

gelten bzw. sollten die Punkte in Abbildung 2 auf der Winkelhalbierenden liegen.

Es kann vorkommen, dass die Daten $x_1, \ldots, x_n$ einen Trend aufweisen (z.B. werden die jährlichen Maxima höher). Wir betrachten nun ein Modell, das der Nichtstationarität der Daten Rechnung trägt. Die beobachteten Blockmaxima $x_1, \ldots, x_n$ seien Realisierungen von

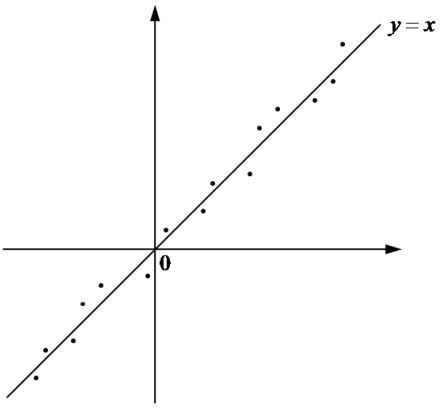


ABBILDUNG 2. QQ–Plot

Zufallsvariablen $X_1, \ldots, X_n$, die unabhängig aber nicht identisch verteilt seien mit

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

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:

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

Wir gehen also von einem konstanten Formparamter γ, einem konstanten Skalenparameter σ und einem linearen Trend, der im Lageparameter μ 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 $x_1, \ldots, x_n$ von dem Trend bereinigen:

$$x_i' := x_i - \hat{\beta}_0 - \hat{\beta}_1 i.$$

Die bereinigte Stichprobe $x_1', \ldots, x_n'$ sollte man dann mit der Verteilungsfunktion $G_{\hat{\gamma},0,\hat{\sigma}}$ vergleichen.

Bemerkung 6.1.6. Der Ansatz kann verallgemeinert werden, ohne dass sich das Modell grundsätzlich ändert. So ist es zum Beispiel problemlos möglich, einen exponentiellen Trend zu modellieren:

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

6.2. Peaks over Threshold: Statistik der GP–Verteilungen

Die oben beschriebene Methode basiert auf der Betrachtung von Blockmaxima (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 uns nur für die großen Werte von X und stellen die folgende Frage:

Wie ist der sogenannte **Exzess** $X-u$ asymptotisch verteilt, gegeben dass $X > u$? Dabei geht $u \to \infty$.

Wir betrachten drei Beispiele.

Beispiel 6.2.1. Sei X exponentialverteilt mit Parameter $\lambda > 0$, d.h. $\bar{F}(t) = \mathrm{e}^{-\lambda t}$, $t > 0$. Dann gilt

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

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

Beispiel 6.2.2. Sei X aus dem Max–Anziehungsbereich der Fréchet–Verteilung Φ_α, $\alpha > 0$. D.h., $\bar{F} \in RV_{-\alpha}$. Dann gilt für alle $t > 0$:

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

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

Beispiel 6.2.3. Sei X standardnormalverteilt. Folgende Relation wurde in Lemma **??** mit der Regel von L'Hôspital bewiesen:

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

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

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

für $u \to \infty$. 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.

Definition 6.2.4. Die **verallgemeinerte Pareto–Verteilung** (GPD, Generalized Pareto Distribution) mit Index $\gamma \in \mathbb{R}$ und Skalenparameter $\sigma > 0$ ist definiert durch die Verteilungsfunktion

$$P_{\gamma,\sigma}(t) = 1 - \left(1 + \frac{\gamma t}{\sigma}\right)^{-\frac{1}{\gamma}} \text{ mit } \begin{cases} t > 0, & \text{falls } \gamma > 0, \\ t \in [0, -\frac{\sigma}{\gamma}], & \text{falls } \gamma < 0. \end{cases}$$

Bemerkung 6.2.5. Für $\gamma = 0$ interpretieren wir die Formel als Grenzwert:

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

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

Satz 6.2.6 (Pickands–Balkema–de Haan, 1974)**.** Sei X eine Zufallsvariable mit Verteilungsfunktion F, die im rechten Endpunkt x^* stetig ist. Dann liegt F im Max–Anziehungsbereich von G_γ 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 gilt die Approximation

$$\mathbb{P}[X - u \leq t | X > u] \approx P_{\gamma,\beta(u)}(t),$$

falls X im Max–Anziehungsbereich von G_γ liegt.

Nun werden wir die GP–Verteilungen in der Statistik anwenden. Es seien $x_1, \ldots, x_n$ 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 $x_{i_1}, \ldots, x_{i_k}$, die u überschreiten. Wir definieren die Exzesse

$$y_1 = x_{i_1} - u, \ldots, y_k = x_{i_k} - u$$

und ignorieren alle anderen Daten. Der Satz von Pickands–Balkema–de Haan macht folgendes Modell plausibel: Die Exzesse $y_1, \ldots, y_k$ sind Realisierungen von unabhängigen und identisch verteilten Zufallsvariablen $Y_1, \ldots, Y_k$, 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

$$f_{\gamma,\sigma}(t) = \frac{1}{\sigma}\left(1 + \frac{\gamma t}{\sigma}\right)^{-\frac{1}{\gamma}-1} \quad \text{mit} \quad \begin{cases} t > 0, & \text{falls } \gamma > 0, \\ t \in [0, -\frac{\sigma}{\gamma}], & \text{falls } \gamma < 0, \end{cases}$$

und $f_{0,\sigma}(t) = \sigma \mathrm{e}^{-t/\sigma}$, $t > 0$, für $\gamma = 0$. 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 die Gleichung nach Z_p umgestellt wird, erhält man schließlich

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

Dabei haben wir Schätzer für σ und γ bereits hergeleitet. Die Wahrscheinlichkeit $\mathbb{P}[X > u]$ kann durch $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i > u}$ geschätzt werden. Es ergibt sich der Schätzer

$$\hat{Z}_p = u + \frac{\hat{\sigma}}{\hat{\gamma}}\left[\left(\frac{1}{np}\sum_{i=1}^{n} \mathbb{1}_{x_i > u}\right)^{\hat{\gamma}} - 1\right].$$