Graphische Methoden zur Untersuchung des Verteilungsrandes

- Histogramm
- Quantil-Quantil Plots

  $X_1, X_2, \ldots, X_n$ sind i.i.d. ZV mit einer unbekannten Verteilung $\tilde{F}$. Es wird vermutet, dass $\tilde{F}$ am Rand von einer bekannten Verteilung F approximiert wird. Wie kann man diese Vermutung testen?

  Sei $X_{n,n} \le X_{n-1,n} \le \ldots \le X_{1,n}$ eine sortierte Stichprobe aus X_1, X_2,..., X_n.

  qq-plot: $\{(X_{k,n}, F^{\leftarrow}(\frac{n-k+1}{n+1})) \colon k = 1, 2, \ldots, n\}$.

  Bei einer plausiblen Vermutung stellt der qq-plot eine einigermaßen lineare Abhängigkeit dar. Diese Eigenschaft bleibt auch dann erhalten wenn die echte Verteilung und die Referenz-Verteilung nicht übereinstimmen, sondern vom selben Typus sind.

  Faustregel: Je größer das Quantil um so mehr “heavy tailed” ist die Verteilung!

Der Hill Schätzer

Seien $X_1, X_2, \ldots, X_n$ i.i.d. mit Verteilungsfunktion F, sodass $\bar{F} \in RV_{-\alpha}$, $\alpha > 0$, d.h. $\bar{F}(x) = x^{-\alpha} L(x)$ mit $L \in RV_0$.

Ziel: Schätzung von α!

Theorem 14 *(Satz von Karamata)*
Sei L eine regulär variierende und lokal beschränkte Funktion auf $[x_0, +\infty)$ für ein $x_0 \in \mathbb{R}$. Dann gilt:

(a) Für $\kappa > -1$:

$$\int_{x_o}^{x} t^{\kappa} L(t) dt \sim \frac{1}{\kappa + 1} x^{\kappa+1} L(x) \text{ für } x \to \infty$$

(b) Für $\kappa < -1$:

$$\int_{x}^{+\infty} t^{\kappa} L(t) dt \sim -\frac{1}{\kappa + 1} x^{\kappa+1} L(x) \text{ für } x \to \infty$$

Beweis in Bingham et al. 1987.

Annahme: L ist lokal beschränkt in $[u, +\infty)$.

Aus dem Satz von Karamata folgt:

$$E\left(\ln(X) - \ln(u) | \ln(X) > \ln(u)\right) = \lim_{u\to\infty} \frac{1}{\bar{F}(u)} \int_u^\infty (\ln x - \ln u) dF(x) = \alpha^{-1}. \tag{8}$$

Für die empirische Verteilung $F_n(x) = \frac{1}{n}\sum_{k=1}^n I_{[x_k,\infty)}(x)$ und eine hohe Stichprobenabhängige Schwelle $x_{k,n}$, erhält man:

$$E\left(\ln(X) - \ln(x_{k,n}) | \ln(X) > \ln(x_{k,n})\right) \approx$$

$$\frac{1}{\bar{F}_n(x_{k,n})} \int_{X_{k,n}}^\infty (\ln x - \ln x_{k,n}) dF_n(x) = \frac{1}{k-1} \sum_{j=1}^{k-1} (\ln x_{j,n} - \ln x_{k,n}).$$

Wenn $k = k(n) \to \infty$ und $k/n \to 0$, dann $X_{k,n} \to \infty$ für $n \to \infty$, und dann folgt aus (8):

$$\lim_{n\to\infty}\frac{1}{k-1}\sum_{j=1}^{k-1}(\ln x_{j,n}-\ln x_{k,n})=\alpha^{-1}\Longrightarrow \widehat{\alpha}_{k,n}^{(H)}=\left(\frac{1}{k}\sum_{j=1}^{k}(\ln x_{j,n}-\ln x_{k,n})\right)^{-1}$$

Wie wird ein passendes k für eine gegebene Stichprobengröße n gewählt?

k zu klein: hohe Varianz des Schätzers!

k zu gross: Schätzer basiert auf zentrale Werte der Verteilung $\Longrightarrow$ Der Schätzer ist verzerrt!

Grafische Inspektion des Hill Plots: $\left\{\left(k,\widehat{\alpha}_{k,n}^{(H)}\right):k=2,\ldots,n\right\}$

Für einen gegebenen Schätzer $\widehat{\alpha}_{k,n}^{(H)}$ von α erhält man folgenden Schätzer für die Randverteilung $\bar{F}$:

$$\widehat{\bar{F}}(x)=\frac{k}{n}\left(\frac{x}{x_{k,n}}\right)^{-\widehat{\alpha}_{k,n}^{(H)}}.$$

und folgenden Quantil–Schätzer:

$$\widehat{q}_p=\widehat{F}^{\leftarrow}(p)=\left(\frac{n}{k}(1-p)\right)^{-1/\widehat{\alpha}_{k,n}^{(H)}}x_{k,n}.$$

Die POT Methode (Peaks over Threshold)

Definition 18 *(Die verallgemeinerte Pareto Verteilung (GPD))*
Die standard GPD G_γ:

$$G_\gamma(x) = \begin{cases} 1 - (1+\gamma x)^{-1/\gamma} & \text{für } \gamma \neq 0 \\ 1 - \exp\{-x\} & \text{für } \gamma = 0 \end{cases}$$

wobei $x \in D(\gamma)$

$$D(\gamma) = \begin{cases} 0 \leq x < \infty & \text{für } \gamma \geq 0 \\ 0 \leq x \leq -1/\gamma & \text{für } \gamma < 0 \end{cases}$$

oder $G_\gamma(x) = 1 - (1+\gamma x)^{-1/\gamma}$, $x \in D(\gamma)$ und $G_0 = \lim_{\gamma\to 0} G_\gamma$.

Sei $\nu \in \mathbb{R}$ und $\beta > 0$. Eine GPD ist durch die untenstehende Veretilungsfunktion gegeben

$$G_{\gamma,\nu,\beta} = 1 - (1+\gamma\frac{x-\nu}{\beta})^{-1/\gamma}$$

wobei $x \in D(\gamma,\nu,\beta)$ und

$$D(\gamma,\nu,\beta) = \begin{cases} \nu \leq x < \infty & \text{für } \gamma \geq 0 \\ \nu \leq x \leq \nu - \beta/\gamma & \text{für } \gamma < 0 \end{cases}$$

Theorem 15 *(Charakterisierung von $MDA(H_\gamma)$)*
Sein $\gamma \in \mathbb{R}$. Die untenstehenden Aussagen sind äquivalent:

(i) $F \in MDA(H_\gamma)$

(ii) *Es existiert eine positive messbare Funktion $a(\cdot)$, sodass für $x \in D(\gamma)$*

$$\lim_{u \uparrow x_F} \frac{\bar{F}(u + xa(u))}{\bar{F}(u)} = \bar{G}_\gamma(x).$$

Definition 19 *(Exzess-Verteilung)*
Sei X eine ZV mit Verteilungsfunktion F und rechtem Endpunkt x_F. Für $u < x_F$

$$F_u(x) = P(X - u \le x | X > u), x \ge 0$$

heißt Exzess-Verteilungsfunktion über die Schwelle u.

Theorem 16 *(Eine weitere Charakterisierung von $MDA(H_\gamma)$)*
Sei $\gamma \in \mathbb{R}$. Folgende Aussagen sind äquivalent:

(i) $F \in MDA(H_\gamma)$

(ii) *Es existiert eine positive messbare Funktion $\beta(\cdot)$, sodass*

$$\lim_{u \uparrow x_F} \sup_{x \in (0, x_F - u)} |F_u(x) - G_{\gamma, 0, \beta(u)}(x)| = 0$$