# Extreme Value Theory & Time Series Analysis

Study year 2023/24, Q3/4

Fabian Mies

Delft University of Technology

f.mies@tudelft.nl

# Structure of the course

- Part I: Extreme Value Theory in Q3
- Part II: Time Series Analysis in Q4
- each part has two mandatory assignments and an exam
- Lectures slides are put on Brightspace prior to the lecture

# Structure of the course: assessment

**Important dates:**

- Extreme Value Theory, Assignment 1: hand in by March 05, 2024 (7.5% of overall grade)
- Extreme Value Theory, Assignment 2: hand in by March 26, 2024 (7.5% of overall grade)
- Extreme Value Theory, Exam: April 16, 2024 (30% of overall grade)
- Time Series Analysis, Assignment 1: hand in by May 14, 2024 (7.5% of overall grade)
- Time Series Analysis, Assignment 2: hand in by June 04, 2024 (7.5% of overall grade)
- Time Series Analysis, Exam: June 25, 2024 (30% of overall grade)

**About the assignments**

- Exercise sheet is published two weeks earlier
- Proofs and some programming / data analysis
- Hand-in via Brightspace as a group of 2

**About the exam**

- Written exam, 2h duration
- Contents: Proofs, routine calculations, interpretation
- You may bring 1 handwritten page as cheatsheet (A4, one side blank)

# Extreme Value Theory & Time Series Analysis

Part I: Extreme Value Theory

Study year 2023/24, Q3

Fabian Mies

Delft University of Technology

f.mies@tudelft.nl

## Outline



**Recommended literature:**

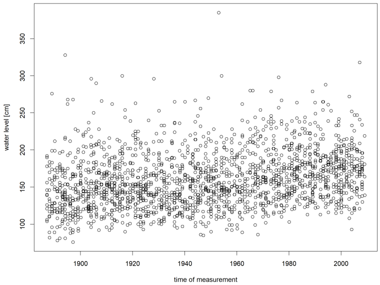
de Haan & Ferreira (2006) Extreme value theory: an introduction (ebook)

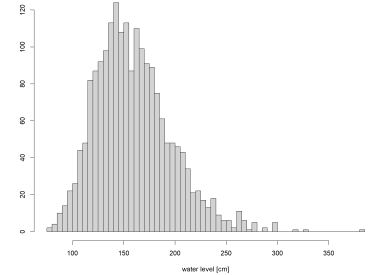
## Outline

1. **Introduction: Estimating the tail of a distribution** — Lecture 1+2
2. Generalized extreme value (GEV) distributions and the extreme value limit theorem — Lecture 2+3
3. Estimation of GEV parameters — Lecture 3+4
4. Assessing m-year returns — Lecture 4
5. Bivariate extremes — Lecture 5+6

**Recommended literature:**
de Haan & Ferreira (2006) Extreme value theory: an introduction (ebook)

# Historical sea levels at Hoek van Holland

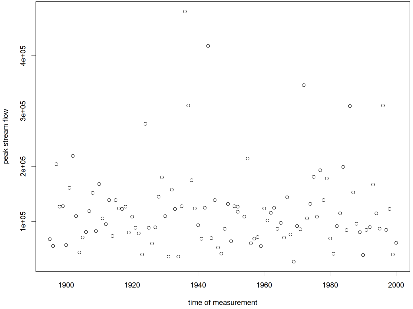


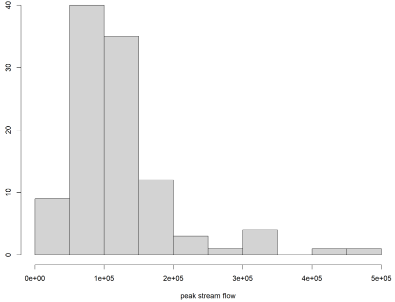




- Records of water levels during storms at Hoek van Holland, 1887-2009
- Measurements $x_1, \ldots, x_{1965}$ (multiple storms per year)

- How often should we expect a sea level above 400 cm, say?
- How high should the dike be to prevent 99.99% of all floods?

# Peak stream flow of Potomac river




- Yearly maximum stream flow of Potomac river, from 1895-2000, measured at Point of Rocks, Maryland US
- Higher stream flow implies higher speeds and higher water levels

# Estimating probabilities of rare events

For random variables $X_1, \dots, X_n \overset{iid}{\sim} F$ with cdf $F$, we may estimate $F(x) = P(X \leq x)$ by the empirical cdf (ecdf)

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x).$$

Analogously, the exceedance probability $P(X > x) = 1 - F(x) = \overline{F}(x)$ can be estimated by the empirical survival function

$$\widehat{\overline{F}}_n(x) = 1 - \widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i > x).$$

## Relative error of the ECDF

For any $x \in \mathbb{R}$ such that $F(x) \in (0,1)$, we have

$$\frac{\left|\widehat{\overline{F}}_n(x) - \overline{F}(x)\right|}{\overline{F}(x)} = \mathcal{O}_P\left(\frac{1}{\sqrt{n\overline{F}(x)}}\right).$$

Not reasonable for large $x$, i.e. small $\overline{F}(x)$.

For $\overline{F}(x) \ll \frac{1}{n}$, we probably will not see ANY data point exceeding $x$.

## Example 1.1

**The tail estimate can be better in parametric models:**

- For some mean parameter $\mu \in \mathbb{R}$, we suppose that $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$.
- Estimate parameters by $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ such that

$$|\widehat{\mu}_n - \mu| = \mathcal{O}_P(1/\sqrt{n}).$$

- Plug in the estimator to approximate the tail probability

$$P(X > x) = \overline{\Phi}\,(x - \mu) \approx \overline{\Phi}\,(x - \widehat{\mu}_n)$$

- How does this approximation perform for large $x$?

**Relative error:**

### Lemma 1.2

For any $x > 1$, it holds $\frac{3}{4x} \leq \frac{\overline{\Phi}(x)}{\varphi(x)} \leq \frac{1}{x}$

# Parametric vs Nonparametric approach

**Nonparametric estimate:**

- Estimate $P(X > x)$ by $\widehat{\overline{F}}_n(x)$.
- Relative error for Gaussian observations ($\mu = 0$)

$$\frac{\left|\widehat{\overline{F}}_n(x) - \overline{F}(x)\right|}{\overline{F}(x)} = \mathcal{O}_P\left(\frac{1}{\sqrt{n\overline{\Phi}(x)}}\right)$$
$$= \mathcal{O}\left(\frac{x}{\sqrt{n}}\exp(\tfrac{x^2}{2})\right)$$

**Parametric estimate:**

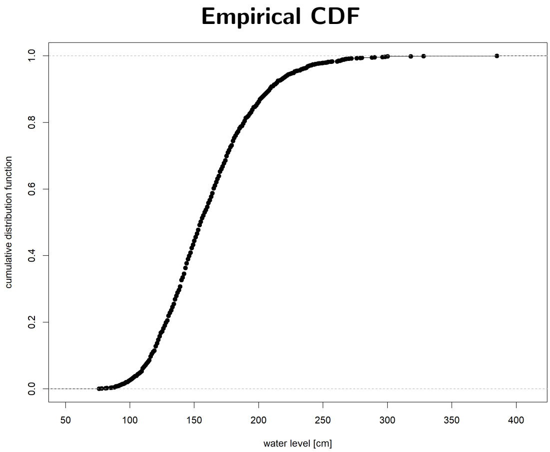
- Estimate $P(X > x)$ by $\overline{\Phi}(x - \widehat{\mu}_n)$.
- Relative error for Gaussian observations ($\mu = 0$)

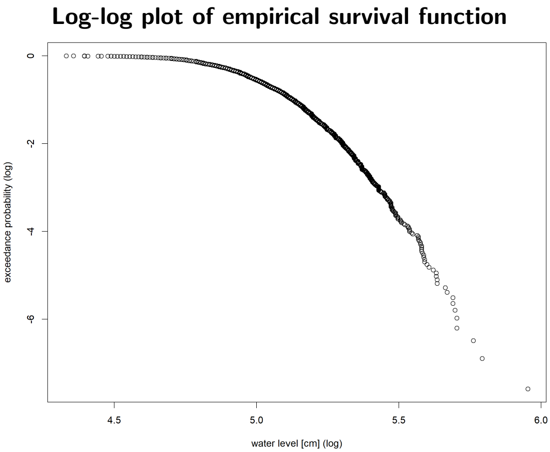
$$\frac{\left|\overline{\Phi}(x - \widehat{\mu}_n) - \overline{F}(x)\right|}{\overline{F}(x)} = \mathcal{O}_P\left(\frac{x}{\sqrt{n}}\right)$$

**Discuss**

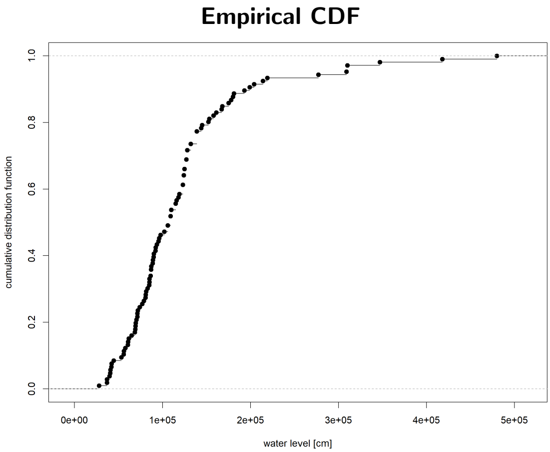
Which approach to use?

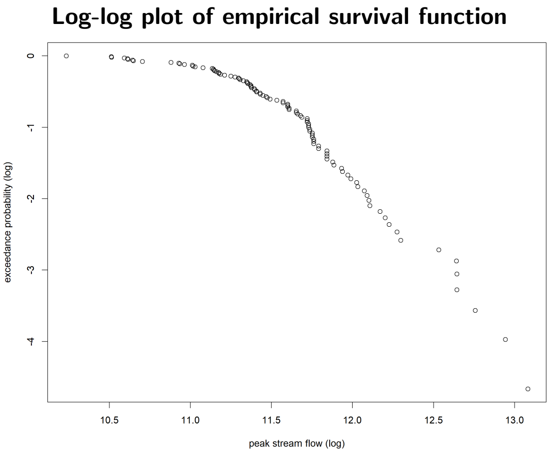
# Hoek van Holland water levels (again)



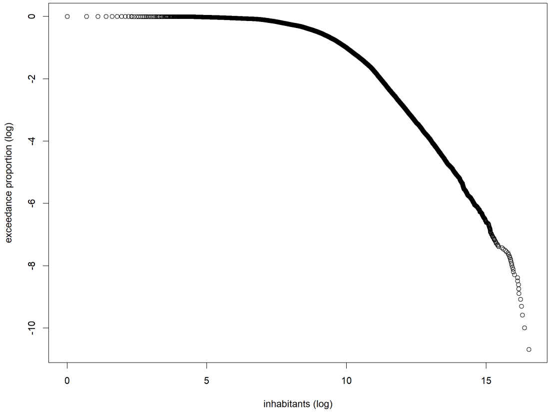
# Potomac river stream flow (again)



# Global size distribution of cities

**Log-log plot of empirical survival function**




## Findings

- log-log plot is approximately linear in the tail
- this suggests

$$\boxed{\overline{F}(x) \approx a x^{-\alpha}, \quad x \to \infty}$$

  for some $a > 0$ and $\alpha > 0$.
- The exponent $\gamma = 1/\alpha$ is referred to as the extreme value index of $F$
- **Central idea:** Use the data to find $a$ and $\gamma$ to extrapolate $\widehat{\overline{F}}_n(x)$ for large values of $x$

**Extreme value theory** makes this extrapolation mathematically **rigorous**.

# Making the log-log plot rigorous

## Definition 1.3 (Regularly varying functions)

A function $f : (0, \infty) \to \mathbb{R}$ is regularly varying with index $\alpha \in \mathbb{R}$ if

$$\lim_{t \to \infty} \frac{f(tx)}{f(t)} = x^\alpha, \quad \forall x > 0.$$

If $\alpha = 0$, we say that $f$ is slowly varying.

## Exercise

Which functions are regularly varying?

- $f(x) = x^\alpha$
- $f(x) = \log(x)$
- $f(x) = x^\alpha \log(\log(x))$
- $f(x) = \exp(x)$

## Lemma 1.4

Suppose that a function $f : (0, \infty) \to \mathbb{R}$ satisfies

$$\lim_{t \to \infty} \frac{f(tx)}{f(t)} = g(x), \quad \forall x > 0.$$

Then there exists some $\alpha \in \mathbb{R}$ such that $g(x) = x^\alpha$.

**Proof:**

**Claim:**

$$\text{If } \lim_{t\to\infty} \frac{f(tx)}{f(t)} = g(x)$$

$$\text{then } g(x) = x^{\alpha}$$

## Proposition 1.5

Let $X$ be a random variable with cdf $F$ such that $\overline{F}$ is regularly varying with exponent $-\frac{1}{\gamma}$ for some $\gamma > 0$. Then

$$P\left(X > tx \,|\, X > t\right) \xrightarrow{t\to\infty} \begin{cases} x^{-1/\gamma}, & x > 1, \\ 1, & x \leq 1. \end{cases}$$

That is, the conditional distribution of $X/t$ given $X > t$ converges towards a Pareto distribution as $t \to \infty$.
We refer to $\gamma$ as the **extreme value index** of $F$.

## Definition 1.6 (Pareto distribution)

A random variable $X$ has a Pareto distribution $X \sim \mathrm{Par}(\alpha)$ with parameter $\alpha > 0$, if

$$P(X \leq x) = \max\left(0,\, 1 - x^{-\alpha}\right).$$

# Different distributions with the same tail behavior

**Definition 1.6 (Pareto distribution)**

A random variable $X$ has a Pareto distribution $X \sim \text{Par}(\alpha)$ with parameter $\alpha > 0$, if

$$P(X \leq x) = \max\left(0, 1 - x^{-\alpha}\right).$$

**Definition 1.7 (Frechet distribution)**

A random variable $X$ has a Frechet distribution $X \sim \text{Frechet}(\alpha)$ with parameter $\alpha > 0$, if

$$P(X \leq x) = \exp(-x^{-\alpha}), \quad x > 0.$$

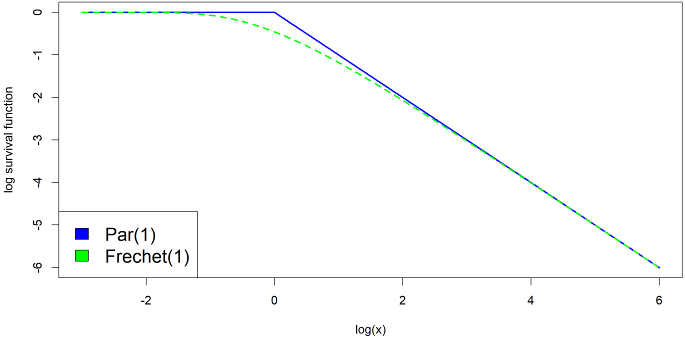



**Figure 1:** Theoretical log-log plot of the Pareto and Frechet distribution, with shape parameter $\gamma = 1$.

**Exercise**

Show that the Frechet distribution is regularly varying.

# MLE for pareto distribution

How to estimate the extreme value index $\gamma$ of a distribution?

### Example 1 (MLE for the Pareto distribution)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Par}(1/\gamma)$ with unknown parameter $\gamma$.

- The Pareto pdf is

$$f_\gamma(x) = \frac{1}{\gamma} x^{-\frac{1}{\gamma}-1} \mathbb{1}(x > 1).$$

- Maximize the log-likelihood

$$l_n(X_1, \ldots, X_n) = -n \log(\gamma) - \left(\tfrac{1}{\gamma} + 1\right) \sum_{i=1}^n \log(X_i)$$

$$\overset{\text{max}}{\Longrightarrow} \quad \widehat{\gamma}_n = \frac{1}{n} \sum_{i=1}^n \log(X_i)$$

### Exercise

Show that $\widehat{\gamma}_n$ is consistent.

# Semiparametric tail index estimation

Now let $X_1, \ldots, X_n \overset{iid}{\sim} F$ for some cdf $F$ with extreme value index $\gamma > 0$.

- Choose a threshold value $t > 0$ and only use data larger than $t$
- Treat those observations as if they were Pareto distributed with parameter $\frac{1}{\gamma}$:

$$\widehat{\gamma}_n = \widehat{\gamma}_n(t) = \frac{1}{n(t)} \sum_{i=1}^{n} \mathbb{1}_{X_i > t} \log\left(\frac{X_i}{t}\right),$$

$$n(t) = \sum_{i=1}^{n} \mathbb{1}_{X_i > t}.$$

## Theorem 1.8

Suppose that $\overline{F}$ has extreme value index $\gamma > 0$, and choose the threshold $t = t_n$ such that $t_n \ll n^{\frac{1}{\gamma}}$. Then

$$\widehat{\gamma}_n = \widehat{\gamma}_n(t_n) \xrightarrow{P} \gamma, \quad n \to \infty.$$

# Proof of Theorem 1.8 (1/3)

- For any $\epsilon > 0$, we have

$$
\begin{aligned}
&P(|\widehat{\gamma}_n - \gamma| > \epsilon) \\
&= \sum_{k=0}^{\infty} P\left(|\widehat{\gamma}_n - \gamma| > \epsilon \,\middle|\, n(t) = k\right) \cdot P(n(t) = k)
\end{aligned}
$$

- It thus suffices to show that
  1. $n(t_n) \xrightarrow{P} \infty$ as $n \to \infty$, and
  2. conditionally on $n(t) = k$, $\widehat{\gamma}_n \xrightarrow{P} \gamma$ as $k \to \infty$.

**Proof of (1):**

# Proof of Theorem 1.8 (2/3)

$$\widehat{\gamma}_n = \widehat{\gamma}_n(t) = \frac{1}{n(t)} \sum_{i=1}^{n} \mathbb{1}_{X_i > t} \log\left(\frac{X_i}{t}\right)$$

1. $n(t_n) \xrightarrow{P} \infty$ as $n \to \infty$, and
2. Conditionally on $n(t) = k$, $\widehat{\gamma}_n \xrightarrow{P} \gamma$ as $k \to \infty$.

**Proof of (2):**

# Proof of Theorem 1.8 (3/3)

$$E\left(\log \tfrac{X}{t} \mid X > t\right) = \int_0^\infty \frac{\overline{F}(t\exp(s))}{\overline{F}(t)}\, ds$$
$$\to \int_0^\infty \exp(s)^{-1/\gamma}\, ds = \gamma$$

## Theorem 1.9 (Karamata's representation, dHF B.1.6)

Let $f : [0,\infty) \to \mathbb{R}$ be regularly varying with index $\alpha$, then there exist funtions $a(t)$ and $c(t)$ with

$$c(t) \xrightarrow{t\to\infty} c_0 \in (0,\infty), \qquad a(t) \xrightarrow{t\to\infty} \alpha$$

and a $t_0 > 0$ such that

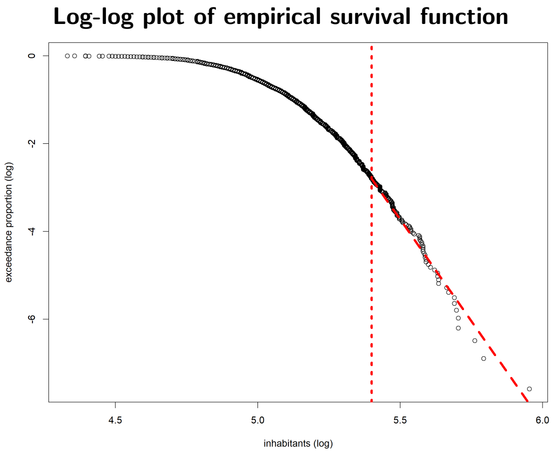
$$f(t) = c(t)\exp\left(\int_{t_0}^t \frac{a(s)}{s}\, ds\right), \quad t > t_0.$$

## Corollary 1.10

Let $f$ be regularly varying with index $\alpha$. Then for any $\epsilon > 0$, there exists some $C = C(\epsilon)$ such that

$$f(t) \leq C(\epsilon) t^{\alpha+\epsilon}.$$

# Application to the water levels at Hoek van Holland




- if $F$ has extreme value index $\gamma > 0$, then

$$\boxed{\overline{F}(tx) \approx \overline{F}(t) x^{-1/\gamma}, \quad x \to \infty}$$

- Estimate the extreme value index by $\widehat{\gamma}_n$ as above, with some suitable threshold $t = t_n$.
- Extrapolate the tail of the distribution as

$$\widetilde{\overline{F}}(y) = \widehat{\overline{F}}(t) \left(\frac{y}{t}\right)^{-1/\widehat{\gamma}_n}$$

**Exercise**

Find a corresponding estimate for a large quantile $F^{-1}(1-\alpha)$, for $\alpha \approx 0$.

## Exercise (optional)

1. Choose a continuous distribution $F$ that has a regularly varying tail with extreme value index $\gamma > 0$. In particular, prove this regular variation for your choice of $F$.
2. Simulate data from this distribution.
3. Pretend that you don't know the real value of $\gamma$, and use $\widehat{\gamma}_n$ to estimate $\gamma$. Choose a suitable value of $t$. Compute the error $e = \hat{\gamma}^H - \gamma$.
4. Repeat steps (2) and (3) for 500 times. You collect $\{e_1, \ldots, e_{500}\}$. Make a boxplot and histogram on the errors. Comment.