An introduction to extreme value theory

Anas Mourahib
a.mourahib@tue.nl

Eindhoven University of Technology

Access the slides and code

Scan the QR code below to access all materials:

(Slides, Python code, and additional resources)

https://github.com/AnasMourahib/ASML-presentation-
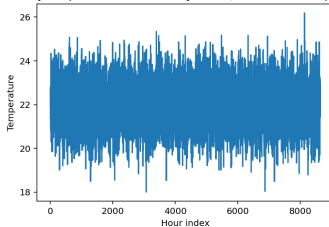
Estimation beyond data

- Daily temperature time series during one year of an ASML component
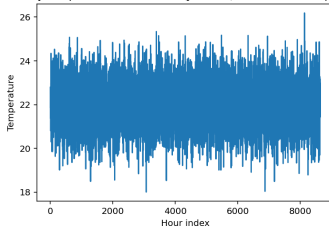

Hourly Temperature Time Series (Jan–Dec) of an ASML Component

Estimation beyond data

- Daily temperature time series during one year of an ASML component
- Assume the component is damaged for a temperature higher than 27 degrees

Hourly Temperature Time Series (Jan–Dec) of an ASML Component

Estimation beyond data

- Daily temperature time series during one year of an ASML component
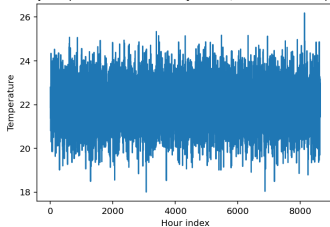- Assume the component is damaged for a temperature higher than 27 degrees
- Question: how to estimate the probability of temperature X exceeding 27 degrees in a period of one year?
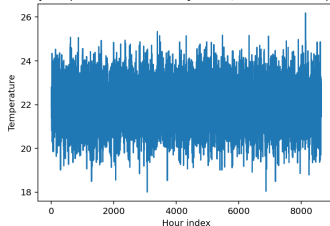

Hourly Temperature Time Series (Jan–Dec) of an ASML Component

Estimation beyond data

- Daily temperature time series during one year of an ASML component

- Assume the component is damaged for a temperature higher than 27 degrees

- Question: how to estimate the probability of temperature X exceeding 27 degrees in a period of one year?

- Challenge: But we have never observed such an event :(:)
  $\rightarrow$ Estimate beyond data, i.e., an event that has never observed

Hourly Temperature Time Series (Jan–Dec) of an ASML Component

To infinity and beyond

Mission

To model rare events, beyond what we have observed so far.

To infinity and beyond

Mission

To model rare events, beyond what we have observed so far.

Guiding principle

To make as little assumptions as possible.

Temperature dataset
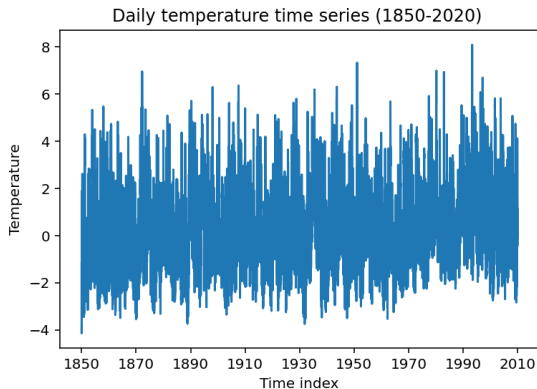
Data

we use daily temperature data between 1850 and 2020

Daily temperature time series (1850-2010)

Table of Contents

Table of Contents

Central limit theorem for averages

- Consider $X_1, \ldots, X_n$ i.i.d copies representing temperature with CDF

$$F(x) = \mathsf{P}(X \leqslant x)$$

Central limit theorem for averages

- Consider $X_1, \ldots, X_n$ i.i.d copies representing temperature with CDF

$$F(x) = \mathsf{P}(X \leqslant x)$$

Sample Average

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Central limit theorem for averages

- Consider $X_1, \ldots, X_n$ i.i.d copies representing temperature with CDF

$$F(x) = \mathsf{P}(X \leqslant x)$$

Sample Average

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Central Limit Theorem

Under some assumptions:

$$\sqrt{n}(\bar{X}_n - \mu) \overset{d}{\to} \mathcal{N}(0, \sigma^2)$$

Central limit theorem for maxima

- Consider $X_1, \ldots, X_n$ i.i.d copies representing temperature with CDF

$$F(x) = \mathsf{P}(X \leqslant x)$$

Maximum of the Sample

$$M_n = \max(X_1, \ldots, X_n)$$

Central limit theorem for maxima

- Consider $X_1, \ldots, X_n$ i.i.d copies representing temperature with CDF

$$F(x) = \mathsf{P}(X \leqslant x)$$

Maximum of the Sample

$$M_n = \max(X_1, \ldots, X_n)$$

Assumption: CLT for Maxima

There exist sequences $a_n > 0$ and b_n and a nondegenerate distribution G such that

$$\mathsf{P}\left(\frac{M_n - b_n}{a_n} \leqslant x\right) = F^n(a_n x + b_n) \to G(x) \quad (n \to \infty)$$

We say that F is in the max-domain of attraction of G and note $F \in \mathrm{DA}(G)$

Extreme Value Distributions

Question 1

Which distribution function G can arise in the limit?

Answer

G must be an extreme value distribution, parametrized by three parameters.

Extreme Value Distributions

Question 1

Which distribution function G can arise in the limit?

Answer

G must be an extreme value distribution, parametrized by three parameters.

Question 2

For a given G, which distributions F are attracted by it?

Answer

$\bar{F} = 1 - F$ must be regularly varying with some index $\gamma > 0$, meaning that for every $\lambda > 0$

$$\lim_{x \to \infty} \frac{\bar{F}(x\lambda)}{\bar{F}(x)} = \lambda^{\gamma}.$$

Example: Exponential distribution

- Assume $X_1, \ldots, X_n$ are i.i.d. unit exponential random variables:

$$P(X \leqslant x) = 1 - e^{-x}, \qquad x > 0.$$

Example: Exponential distribution

- Assume $X_1, \ldots, X_n$ are i.i.d. unit exponential random variables:

$$P(X \leqslant x) = 1 - e^{-x}, \qquad x > 0.$$

- Take $a_n = 1$ and $b_n = \log(n)$, for every $x > 0$ we get

$$P(M_n - \log(n) \leqslant x) = \left(1 - \frac{e^{-x}}{n}\right)^n$$
$$\to \exp\left(-e^{-x}\right) := \Lambda(x)$$

- Thus, $G = \Lambda$ is the **Gumbel distribution.**

General form of extreme value distributions

Parameters

An extreme value distribution is parametrized by:

- Location parameter $\mu \in \mathbb{R}$
- Scale parameter $\alpha > 0$
- Shape parameter $\gamma \in \mathbb{R}$

General form of extreme value distributions

Parameters

An extreme value distribution is parametrized by:

- Location parameter $\mu \in \mathbb{R}$
- Scale parameter $\alpha > 0$
- Shape parameter $\gamma \in \mathbb{R}$

Functional Form

$$G(x) = \begin{cases} \exp\left[- \left(1 + \gamma \dfrac{x - \mu}{\alpha} \right)^{-1/\gamma} \right], & \gamma \neq 0, \\[2em] \exp\left[- \exp\left(- \dfrac{x - \mu}{\alpha} \right) \right], & \gamma = 0. \end{cases}$$
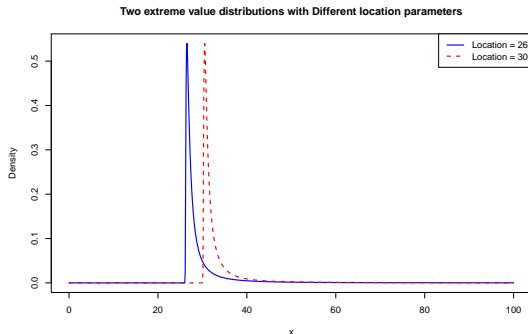
Interpretation: location parameter

Meaning

Location (or shift) parameter μ shifts the entire tail left or right.

Example

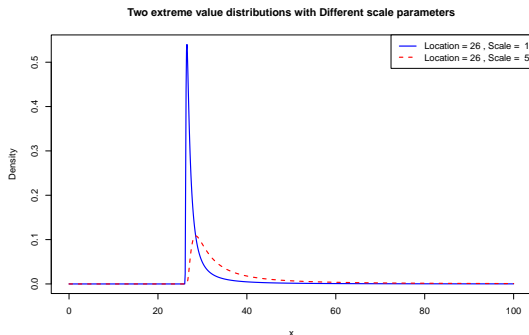If $\mu = 26$, then temperatures around $26°$ are the "average" extremes.

Two extreme value distributions with Different location parameters

Interpretation: scale parameter

Meaning

The scale parameter σ controls the spread of extremes around μ.

Example

If $\sigma = 1$, extremes cluster close to μ; if $\sigma = 5$, extremes can deviate far from μ.

Two extreme value distributions with Different scale parameters

Interpretation: shape parameter

Meaning

Shape parameter γ controls the tail behavior of the distribution.

Cases

- $\gamma = 0$ (Gumbel): light tail — extremes grow slowly (e.g., temperature).
- $\gamma > 0$ (Fréchet): heavy tail — very large extremes possible (e.g., floods).
- $\gamma < 0$ (Weibull): bounded tail — hard upper limit (e.g., wind speed).
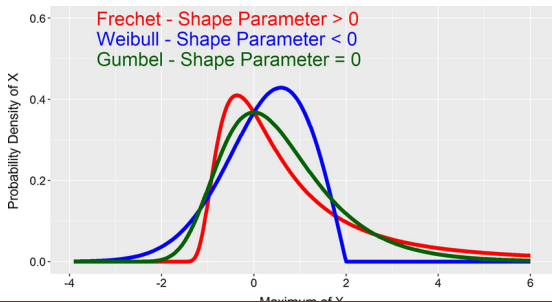
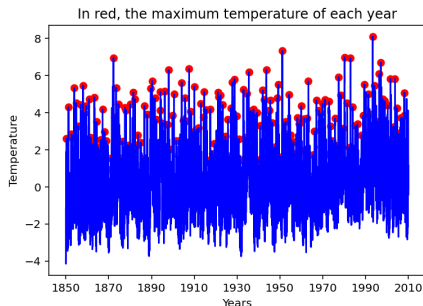# Table of Contents

Block maxima: divide data

- Recall daily observations of temperature between 1850 and 2020
- We divide data by years, meaning $t = 1, \ldots, 171$ years

Block maxima: divide data

- Recall daily observations of temperature between 1850 and 2020
- We divide data by years, meaning $t = 1, \ldots, 171$ years
- We denote the observations of temperature during a year t as $X_{1,t}, \ldots, X_{n,t}$
  (Here $n = 92$ since we consider only the summer months)

Block maxima: divide data

- Recall daily observations of temperature between 1850 and 2020
- We divide data by years, meaning $t = 1, \ldots, 171$ years
- We denote the observations of temperature during a year t as $X_{1,t}, \ldots, X_{n,t}$ (Here $n = 92$ since we consider only the summer months)
- $M_{n,t} = \max(X_{1,t}, \ldots, X_{n,t})$ the maximum of temperature during a year t

In red, the maximum temperature of each year

Block maxima: fit a generalized extreme value distribution

- Recall that the idea is to fit the maximum values $M_{n,t}, t = 1850, \ldots, 2020$ to a GEV $G_{\mu,\sigma,\gamma}$
- Question: How to do that ?
  $\rightarrow$ Using Maximum Likelihood, we estimate the parameters μ, σ, γ of the GEV G
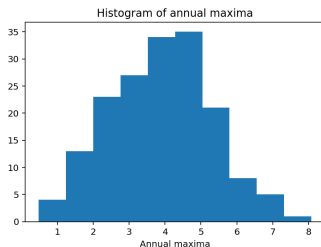
Table of Contents

Histogram of temperature data

- Consider the histogram of annual maxima temperature data
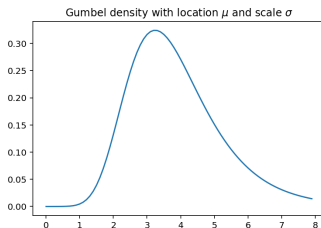

Histogram of annual maxima

Histogram of temperature data

- Consider the histogram of
  annual maxima temperature
  data
- More data are on the left
  which suggests right skewed
  data structure
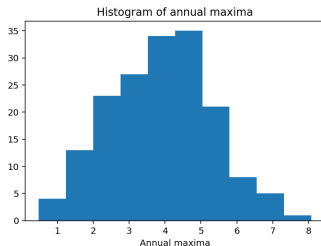

Histogram of annual maxima

Histogram of temperature data

- Consider the histogram of annual maxima temperature data

- More data are on the left which suggests right skewed data structure

- Consider the Gumbel distribution: this is the case $\gamma = 0$

Histogram of annual maxima

Gumbel density with location μ and scale σ

Quantile Gumbel function

- The quantile function of F:

$$q(u) = \inf\{x \in \mathbb{R} : F(x) > u\}$$

Quantile Gumbel function

- The quantile function of F:

$$q(u) = \inf\{x \in \mathbb{R} : F(x) > u\}$$

- For the standard case $\mu = 0$ and $\sigma = 1$, we get

$$q_{0,1}(u) = \log(-\log(u))$$

Quantile Gumbel function

- The quantile function of F:

$$q(u) = \inf\{x \in \mathbb{R} : F(x) > u\}$$

- For the standard case $\mu = 0$ and $\sigma = 1$, we get

$$q_{0,1}(u) = \log(-\log(u))$$

- The quantile function of a $\mathrm{Gumbel}(\mu, \sigma)$ is given by

$$q_{\mu,\sigma}(u) = -\sigma q_{0,1}(u) + \mu$$

Quantile Gumbel function

- The quantile function of F:

$$q(u) = \inf\{x \in \mathbb{R} : \ F(x) > u\}$$

- For the standard case $\mu = 0$ and $\sigma = 1$, we get

$$q_{0,1}(u) = \log(-\log(u))$$

- The quantile function of a $\mathrm{Gumbel}(\mu, \sigma)$ is given by

$$q_{\mu,\sigma}(u) = -\sigma q_{0,1}(u) + \mu$$

- location μ and scale σ can be seen as an intercept and a slope, respectively.

Q-Q plot: standard Gumbel fit

- Recall the annual maximum values M_t, $t = 1850, \ldots, 2020$

Q-Q plot: standard Gumbel fit

- Recall the annual maximum values M_t, $t = 1850, \ldots, 2020$
- Consider the order

Q-Q plot: standard Gumbel fit

- Recall the annual maximum values M_t, $t = 1850, \ldots, 2020$
- Consider the order

$$M_{1,T} \leqslant M_{2,T} \leqslant \ldots \leqslant M_{T,T}$$

- The Q-Q plot for a standard Gumbel fit ($\mu = 0, \sigma = 1$) is given by

$$\left(-\log\left(-\log \frac{t}{T+1} \right), M_{t,T} \right), \qquad t = 1, \ldots, T$$
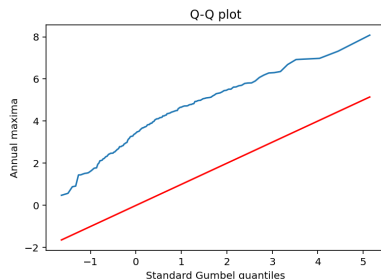
Q-Q plot: standard Gumbel fit

- Recall the annual maximum values M_t, $t = 1850, \ldots, 2020$
- Consider the order

$$M_{1,T} \leqslant M_{2,T} \leqslant \ldots \leqslant M_{T,T}$$

- The Q-Q plot for a standard Gumbel fit $(\mu = 0, \sigma = 1)$ is given by

$$\left(-\log\left(-\log \frac{t}{T+1} \right), M_{t,T} \right), \qquad t = 1, \ldots, T$$

Figure: QQ plot (blue) with reference line $x = y$ (red).

Q-Q plot: $\mathrm{Gumbel}(\mu, \sigma)$ fit

- Recall that

$$q_{\mu,\sigma}(u) = -\sigma q_{0,1}(u) + \mu$$

- To estimate the parameters μ and σ, we regress standard Gumbel quantiles on the empirical quantiles , we get

$$\mu \approx 3.14, \qquad \sigma \approx 1.24$$

Q-Q plot: $\mathrm{Gumbel}(\mu, \sigma)$ fit

- Recall that

$$q_{\mu,\sigma}(u) = -\sigma q_{0,1}(u) + \mu$$

- To estimate the parameters μ and σ, we regress standard Gumbel quantiles on the empirical quantiles , we get

$$\mu \approx 3.14, \qquad \sigma \approx 1.24$$
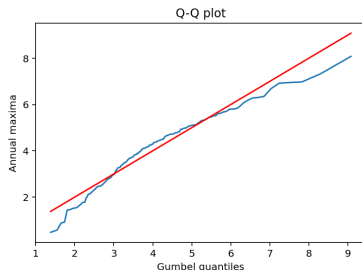
- We get the following Q-Q plot

Figure: QQ plot (blue) with reference line $x = y$ (red).

Table of Contents

Maximum likelihood

- Recall that the idea is to fit a generalized extreme value distribution $\mathrm{GEV}(\mu, \sigma, \gamma)$ to the annual maxima $M_t,, \ t = 1850, \ldots, 2020$
- The log likelihood function of $\mathrm{GEV}(\mu, \sigma, \gamma)$ for $\gamma > 0$ is

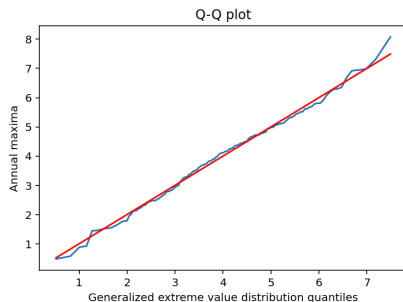$$\ell_{\mu,\sigma,\gamma}(x) = \ell_\gamma \left(\frac{x - \mu}{\sigma} \right) - \log(\sigma), \quad \text{where}$$

$$\ell_\gamma(x) = -(1 + 1/\gamma) \log(1 + \gamma x) - (1 + \gamma x)^{-1/\gamma}$$

- The MLE is consistent and asymptotically normal[1] if $\gamma > -0.5$.

---

[1] see Bücher, Axel, and Johan Segers. "On the maximum likelihood estimator for the generalized extreme-value distribution." Extremes 20.4 (2017): 839-872.

Fitting a GEV in R and Python

- **Python**: using the method
  `genextrem.fit()` from the
  package `scipy.stats`

Fitting a GEV in R and Python

- **Python**: using the method `genextrem.fit()` from the package `scipy.stats`
- **R**: using the function `gev()` from the package `evir`

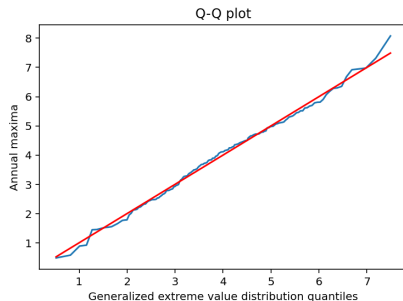 $$\mu \approx 3.35,\ \sigma \approx 1.40,\ \gamma \approx -0.24$$

Q-Q plot

Annual maxima (y-axis) vs Generalized extreme value distribution quantiles (x-axis)

Table of Contents

Table of Contents

Threshold exceedances: set the threshold

- Recall daily observations of temperature between 1850 and 2020
- We set a threshold u and consider observations that are higher than u

Exceedances of the 0.95 quantile in red

- The choice of the threshold u can be challenging, it's a bias-variance trade off

Table of Contents

Generalized Pareto distributions

Assumption:

The excess-over-threshold distribution must converge to a non-degenerate distribution H:

$$X - u \mid X > u \sim H$$

Generalized Pareto distributions

Assumption:

The excess-over-threshold distribution must converge to a non-degenerate distribution H:

$$X - u \mid X > u \sim H$$

Question

Which distribution function H can arise in the limit?

Answer

H must be an generalized Pareto distribution, parametrized by two parameters.

General form of generalized Pareto distributions

Parameters

A generalized pareto distribution is parametrized by:

- Scale parameter $\alpha > 0$
- Shape parameter $\gamma \in \mathbb{R}$

General form of generalized Pareto distributions

Parameters

A generalized pareto distribution is parametrized by:

- Scale parameter $\alpha > 0$
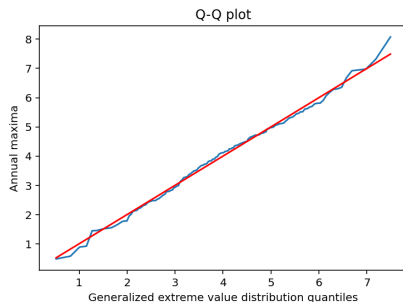- Shape parameter $\gamma \in \mathbb{R}$

Functional Form

$$H_{\sigma,\gamma}(x) = \begin{cases} 1 - \left(1 + \dfrac{\gamma x}{\sigma}\right)^{-1/\gamma}, & \gamma \neq 0, \\[2mm] 1 - \exp\left(-\dfrac{x}{\sigma}\right), & \gamma = 0, \end{cases} \qquad x \geqslant 0, \ 1 + \frac{\gamma x}{\sigma} > 0.$$

Relation with generalized extreme value distributions

- The location parameter μ disappears when considering a threshold u in the Generalized Pareto Distribution (GPD).
- $F \in \mathsf{DA}(G_{\mu,\sigma,\gamma})$ if and only if the exceedances of temperatures follow a GPD $H_{\sigma,\gamma}$.
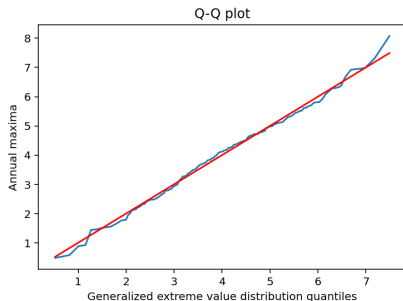
Fitting a GPD in R and Python

- **Python**: using the method `genpareto.fit()` from the package `scipy.stats`

Q-Q plot

Fitting a GPD in R and Python

- **Python**: using the method `genpareto.fit()` from the package `scipy.stats`
- **R**: using the function `fitgpd()` from the package `evir`

$$\sigma \approx 1.40, \ \gamma \approx -0.24$$

Q-Q plot

Books on extreme value theory

De Haan, Laurens, and Ana Ferreira. *Extreme Value Theory: An Introduction.*
Springer New York, 2006.

Beirlant, Jan, et al. *Statistics of Extremes: Theory and Applications.* John
Wiley Sons, 2006.

Thank You for Your Attention!