Chapter 1: Motivations for extreme-value statistics

Stéphane Girard

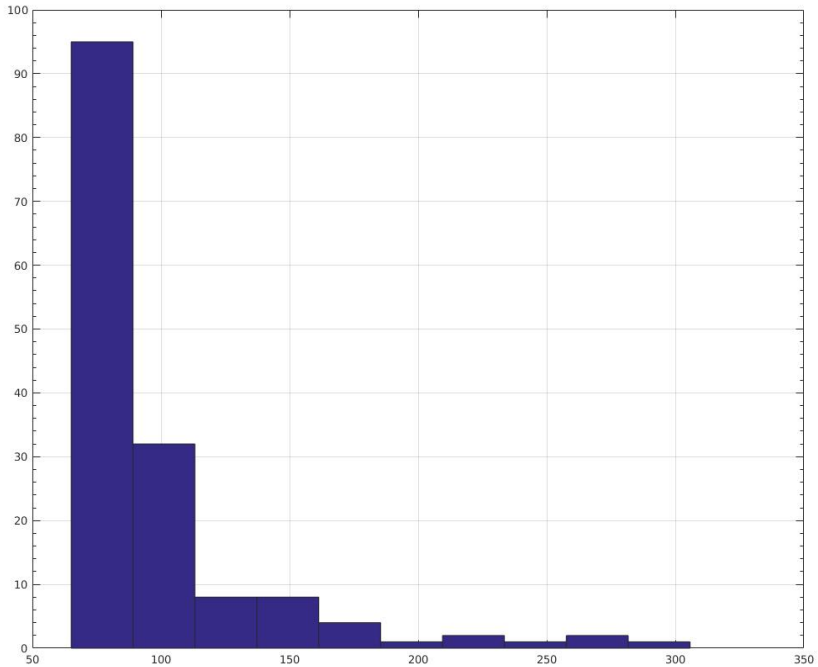
Inria Grenoble Rhône-Alpes
http://mistis.inrialpes.fr/people/girard/

November-december 2018

Extreme rainfall in Montpellier, France, 2014. Three hours of rainfall = 50% of mean annual rainfall (source `http://www.dailymail.co.uk`).

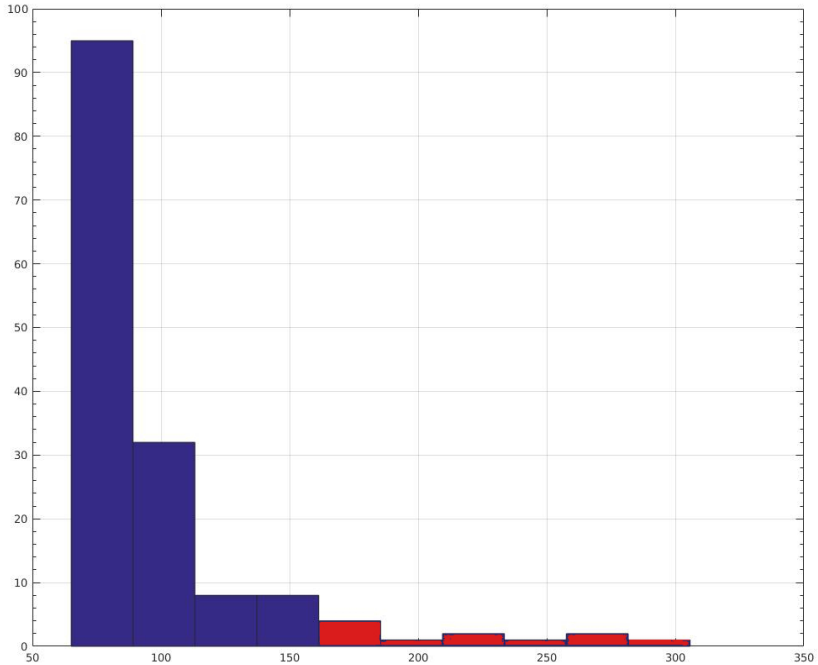
Nidd river (England)

Histogram : $n = 154$ river flows (m^3/s), one measure every three months during 38.5 years.



Nidd river (England)

Probability to observe a flow larger than $160 m^3/s$?

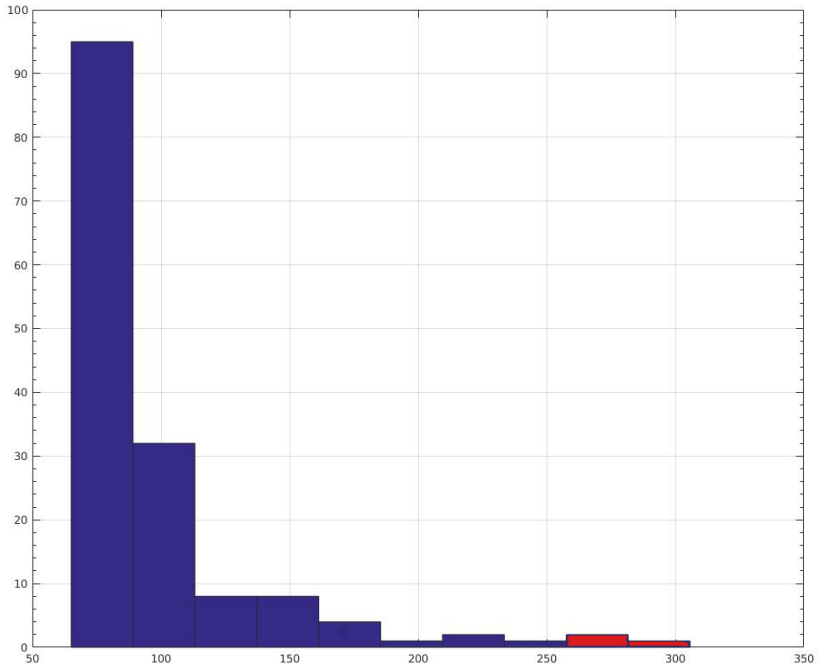


Histogram: $\mathbb{P}(X \geq 160) \simeq \text{nb}(X_i \geq 160)/n = 11/154 = 1/14$.

There is such a flow every 3.5 years.

Nidd river (England)

Probability to observe a flow larger than $255 m^3/s$?

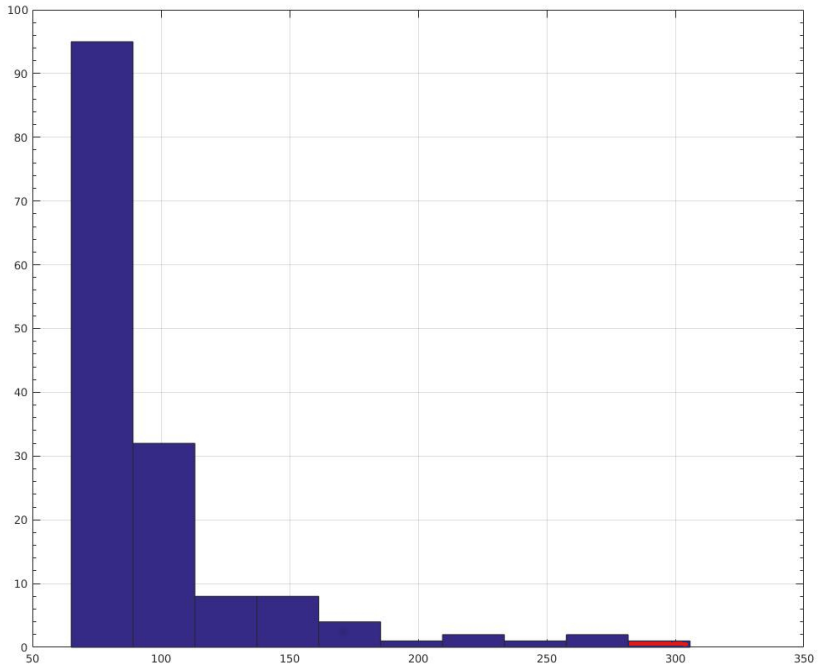


Histogram: $\mathbb{P}(X \geq 255) \simeq \text{nb}(X_i \geq 255)/n = 3/154$.

There is such a flow every 12.8 years.

Nidd river (England)

Probability to observe a flow larger than $280 m^3/s$?

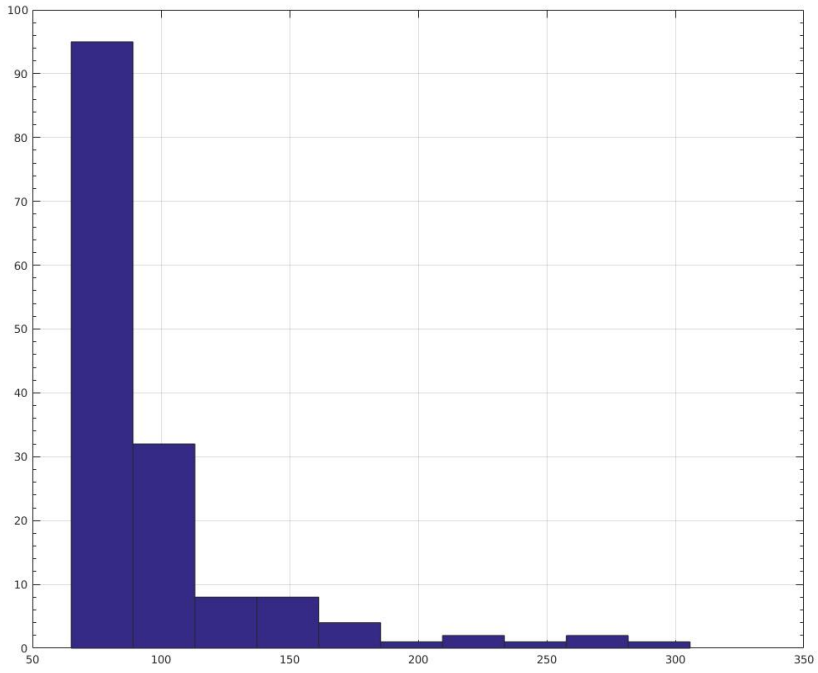


Histogram: $\mathbb{P}(X \geq 280) \simeq \text{nb}(X_i \geq 280)/n = 1/154$.

There is such a flow every 38.5 years.

Nidd river (England)

Probability to observe a flow larger than $500 m^3/s$?



Histogram: $\mathbb{P}(X \geq 500) \simeq \text{nb}(X_i \geq 500)/n = 0$.
It is not possible to estimate the probability of such a flow from the histogram.
The same problem occurs to estimate the 100-year river flow *i.e.* the flow t which is reached once every 100 years: $\mathbb{P}(X \geq t) = 1/400$.

Statistical problem

Definition (Survival function)

$\bar{F}(x) = \mathbb{P}(X \geq x) = 1 - F(x)$ where F is the cumulative distribution function (cdf) associated with X.

Two problems: From $\{X_1, \ldots, X_n\}$ iid with cdf F,

1. **Estimation of small tail probabilites**.
   Given $x_n > X_{n,n} = \max(X_1, \ldots, X_n)$, estimate $p_n = \bar{F}(x_n)$.
2. **Estimation of extreme quantiles**.
   Given p_n such that $np_n \to c < \infty$ as $n \to \infty$, estimate x_{p_n} such that $p_n = \bar{F}(x_{p_n})$.

Common difficulty: The survival function $\bar{F}(x)$ is unknown and difficult to estimate for $x > X_{n,n}$.

Remark. If $np_n \to c < \infty$ then

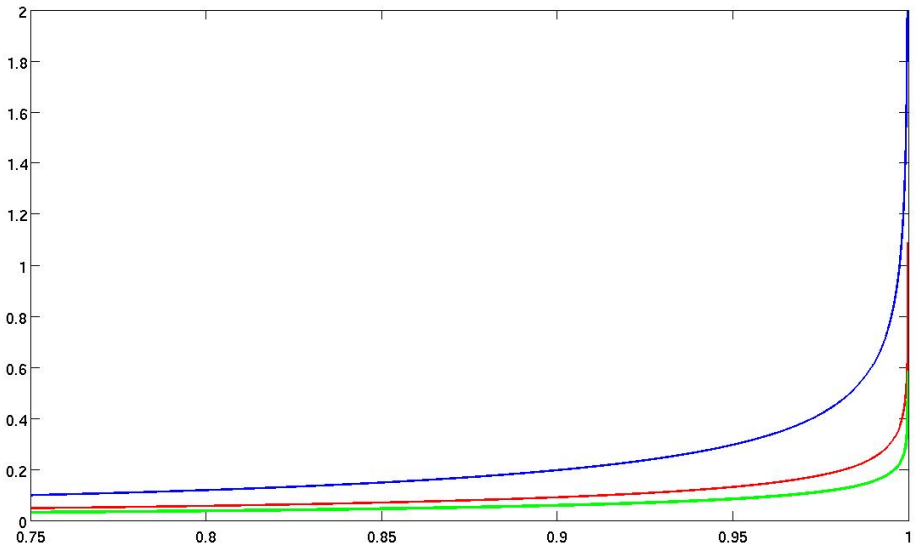
$$\mathbb{P}(x_{p_n} > X_{n,n}) = \mathbb{P}^n(X_1 < x_{p_n}) = F^n(x_{p_n}) = (1-p_n)^n = \exp(n \log(1-p_n)) \to e^{-c}$$

as $n \to \infty$.

Parametric approach

- Suppose *a priori* a parametric model for the survival function: $\bar{F} \in \{\bar{F}_\theta,\ \theta \in \Theta\}$.
- Estimate θ by $\hat{\theta}_n$.

Problem: A good fit on the sample does not necessarily lead to a good modelling above the maximum.



Horizontally: p. Vertically: relative error between the quantile of order p estimated from $\mathcal{N}(0,1)$ and t_4, t_8, t_{12} distributions.

Nonparametric approach

Empirical survival function: $\mathbb{P}(X \geq x)$ is estimated by the proportion of observations above x:

$$\hat{\bar{F}}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \geq x\}$$

Problem: As illustrated on the Nidd river: $\hat{\bar{F}}_n(x) = 0$ if $x > X_{n,n}$.

From an asymptotic point of view: If $x_n \to \infty$, $\mathbb{E}(\hat{\bar{F}}_n(x_n)) = \bar{F}(x_n)$ and thus $\mathbb{E}(\hat{\bar{F}}_n(x_n)/\bar{F}(x_n)) = 1$ the estimator is still unbiased but

$$var(\hat{\bar{F}}_n(x_n)/\bar{F}(x_n)) = \frac{\bar{F}(x_n)F(x_n)}{n\bar{F}^2(x_n)} \sim \frac{1}{n\bar{F}(x_n)}.$$

The estimator is not convergent when $n\bar{F}(x_n) \to c > 0$.

Extreme-value approach

Theorem of the Central Limit. Let $X_1, \ldots, X_n$ be iid random variables with expectation μ and finite variance σ^2. Then,

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Assume we have a similar result for the maxima, *i.e.*

$$\frac{1}{a_n}(\max(X_1, \ldots, X_n) - b_n) \xrightarrow{d} Z,$$

where Z is a random variable with cdf H.

It would follow

$$\mathbb{P}\left(\frac{1}{a_n}(\max(X_1,\ldots,X_n)-b_n)\leq x\right)\simeq\mathbb{P}(Z\leq x),$$

or equivalently,

$$\mathbb{P}\left(\max(X_1,\ldots,X_n)\leq a_n x+b_n\right)\simeq\mathbb{P}(Z\leq x).$$

Letting $t=a_n x+b_n$ we obtain

$$\mathbb{P}\left(\max(X_1,\ldots,X_n)\leq t\right)\simeq\mathbb{P}\left(Z\leq(t-b_n)/a_n\right).$$

Since $\mathbb{P}(\max(X_1,\ldots,X_n)\leq t)=\mathbb{P}^n(X\leq t)$, it follows that

$$\mathbb{P}\left(X\leq t\right)\simeq\mathbb{P}^{1/n}\left(Z\leq(t-b_n)/a_n\right),$$

which can be rewritten as

$$1-\mathbb{P}\left(X\geq t\right)\simeq H^{1/n}\left((t-b_n)/a_n\right).$$

Taking the logarithm yields

$$\log(1 - \mathbb{P}(X \geq t)) \simeq \frac{1}{n} \log H((t - b_n)/a_n).$$

Since t is large, $\mathbb{P}(X \geq t)$ is small, a first order Taylor expansion of $\log(1 + u)$ thus yields

$$\mathbb{P}(X \geq t) \simeq -\frac{1}{n} \log H((t - b_n)/a_n)$$

for n large. Such an approximation would yield estimators for small tail probabilites and for extreme quantiles:

$$x_{p_n} \simeq b_n + a_n H^{-1}(\exp(-np_n)).$$

The goal of the next chapter is to identify the limit distribution H and the normalizing constants a_n and b_n.