Statistical inference for Weibull tail-distributions

Stéphane Girard
INRIA Rhône-Alpes, team Mistis

Joint work with J. Diebolt, L. Gardes and A. Guillou

Outline

1. Weibull tail-distributions.
2. Weighted estimators of the Weibull tail-coefficient.
3. Bias-reduced estimator of the Weibull tail-coefficient.
4. Simulation study.
5. Estimation of extreme quantiles.
6. Nidd river data.

1. Weibull tail-distributions.

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with cumulative distribution function F.

Definition. F represents a Weibull tail-distribution if

$$1 - F(x) = \exp\{-x^{1/\theta} L(x)\},$$

where

- $\theta > 0$ is the Weibull tail-coefficient,
- L is a slowly varying function *i.e.*

$$L(\lambda x)/L(x) \to 1 \text{ as } x \to \infty \text{ for all } \lambda > 0.$$

For instance $L(x) = C$, $L(x) \to C$, $L(x) = \log(x)$, ...

The cumulative hazard function

$$H(x) = -\log(1 - F(x)) = x^{1/\theta} L(x)$$

thus verifies

$$H(\lambda x)/H(x) \to \lambda^{1/\theta} \text{ as } x \to \infty \text{ for all } \lambda > 0.$$

and is said to be regularly varying at infinity with index $1/\theta$.

Examples:

	$H(x)$	θ
Weibull $\mathcal{W}(\alpha, \lambda)$	$\left(\frac{x}{\lambda}\right)^{1/\alpha}$	$1/\alpha$
Gaussian $\mathcal{N}(0,1)$	$x^{1/2}\left(\frac{1}{2} + \frac{\log x}{x^2} + O(1/x^2)\right)$	$1/2$
Gamma $\Gamma(\alpha, \lambda)$	$x\left(\frac{1}{\lambda} + (1-\alpha)\frac{\log x}{x} + O(1/x)\right)$	1

Graphical properties: Let us denote by $q(t)$ the quantile function

$$q(t) = F^{-1}(1-t) = H^{-1}(\log(1/t)).$$

One can show that H^{-1} is still a regularly varying function but with index θ. Consequently,

$$q(t) = (\log(1/t))^{\theta}\ell(\log(1/t)),$$

where ℓ is a slowly varying function. We thus obtain for t and s small:

$$\begin{aligned}\log(q(t)) - \log(q(s)) &= \theta(\log_2(1/t) - \log_2(1/s)) + \log\left(\frac{\ell(\log(1/t))}{\ell(\log(1/s))}\right)\\ &\simeq \theta(\log_2(1/t) - \log_2(1/s)),\end{aligned}$$

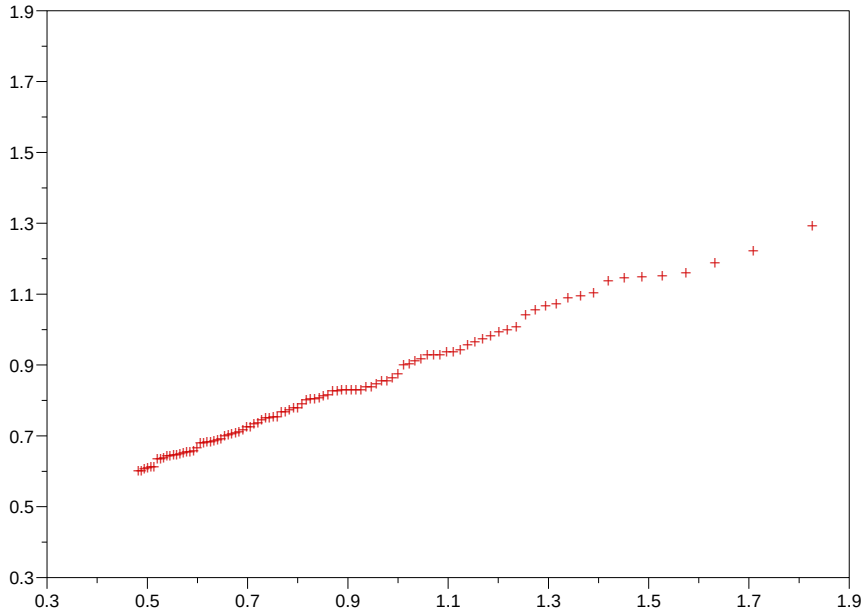
where $\log_2(x) = \log\log(x)$. Considering $t = i/n$, $s = k_n/n$ and replacing F by its empirical counterpart yield

$$\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}) \simeq \theta(\log_2(n/i) - \log_2(n/k_n)),$$

for $i = 1, \ldots, k_n - 1$ and where $k_n/n \to 0$.

Quantile-quantile plot. Drawing the pairs $(\log_2(n/i), \log(X_{n-i+1,n}))$ for $i = 1, \ldots, k_n$ should give a graph which is approximatively linear (with slope θ).

Example : $\mathcal{N}(0,1)$ distribution, $n = 500$, $k_n = 100$.



$\Longrightarrow$ Estimation of θ.

2. Weighted estimators of the Weibull tail-coefficient.

Our method: Estimation via linear combination of upper order statistics.

- Let $\alpha = \{\alpha_{i,n}, i = 1, \ldots, k_n - 1\}$ be a sequence of weights.
- For $i = 1, \ldots, k_n - 1$,

$$\alpha_{i,n} \log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}) \simeq \theta \alpha_{i,n} (\log_2(n/i) - \log_2(n/k_n)),$$

- Summing on $i = 1, \ldots, k_n - 1$ yields the estimator

$$\hat{\theta}_n(\alpha) = \sum_{i=1}^{k_n-1} \alpha_{i,n} (\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})) \Big/ \sum_{i=1}^{k_n-1} \alpha_{i,n} (\log_2(n/i) - \log_2(n/k_n)) .$$

Example 1. Constant weights $\alpha_{i,n} = 1$ for all $i = 1, \ldots, k_n - 1$ yield an existing estimator:

$$\hat{\theta}_n^C = \sum_{i=1}^{k_n-1} (\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})) \Big/ \sum_{i=1}^{k_n-1} (\log_2(n/i) - \log_2(n/k_n)) .$$

Example 2: Least-squares (or Zipf) estimator of θ based on the quantile-quantile plot

$$\hat{\theta}_n^Z = \sum_{i=1}^{k_n-1} (\log_2(n/i) - \tau_n) \log(X_{n-i+1,n}) \Big/ \sum_{i=1}^{k_n-1} (\log_2(n/i) - \tau_n) \log_2(n/i) ,$$

where

$$\tau_n = \frac{1}{k_n - 1} \sum_{i=1}^{k_n-1} \log_2(n/i) .$$

Asymptotic properties.

- **Variance:** Estimator based on the k_n largest observations

  $\Longrightarrow$ Variance proportional to $1/k_n$.

- **Bias:** Consequence of the approximation

$$\begin{aligned}\log(q(t)) - \log(q(s)) &= \theta(\log_2(1/t) - \log_2(1/s)) + \log\left(\frac{\ell(\log(1/t))}{\ell(\log(1/s))}\right)\\ &\simeq \theta(\log_2(1/t) - \log_2(1/s)),\end{aligned}$$

  $\Longrightarrow$ Bias proportional to

$$\sum_{i=1}^{k_n-1} \alpha_{i,n} \log\left(\frac{\ell(\log(n/i))}{\ell(\log(n/k_n))}\right)$$

Second order condition on ℓ.

There exist $\rho \leq 0$ and $b(x) \to 0$ such that, for all $\lambda \geq 1$

$$\log\left(\frac{\ell(\lambda x)}{\ell(x)}\right) \sim b(x) K_\rho(\lambda), \text{ when } x \to \infty,$$

with

$$K_\rho(\lambda) = \int_1^\lambda u^{\rho-1} du.$$

- It can be shown that necessarily b **is regularly varying with index** ρ.
- The second order parameter $\rho \leq 0$ tunes the rate of convergence of $\ell(\lambda x)/\ell(x)$ to 1. The closer ρ is to 0, the slower is the convergence.
- Order of the bias: $b(\log(n/k_n))$.

Examples:

	θ	$\ell(x)$	$b(x)$	ρ
Weibull $\mathcal{W}(\alpha,\lambda)$	$1/\alpha$	λ	0	$-\infty$
Gaussian $\mathcal{N}(\mu,\sigma^2)$	$1/2$	$2^{1/2}\sigma - \frac{\sigma}{2^{3/2}}\frac{\log x}{x} + O(1/x)$	$\frac{1}{4}\frac{\log x}{x}$	-1
Gamma $\Gamma(\alpha,\lambda)$	1	$\frac{1}{\lambda} + \frac{\alpha-1}{\lambda}\frac{\log x}{x} + O(1/x)$	$(1-\alpha)\frac{\log x}{x}$	-1

Asymptotic normality Under the some assumptions on the weights,

$$k_n^{1/2}(\hat{\theta}_n(\alpha) - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2\sigma^2(\alpha)),$$

for any sequence (k_n) satisfying $k_n \to \infty$ and

$$k_n^{1/2} b(\log(n/k_n)) \to 0,$$

where $\sigma^2(\alpha)$ is a positive number depending on the weights.

3. Bias-reduced estimator of the Weibull tail-coefficient.

Principle: The previous estimators are based on the approximation

$$\begin{aligned}\log(q(t)) - \log(q(s)) &= \theta(\log_2(1/t) - \log_2(1/s)) + \log\left(\frac{\ell(\log(1/t))}{\ell(\log(1/s))}\right)\\ &\simeq \theta(\log_2(1/t) - \log_2(1/s)).\end{aligned}$$

The second order condition can be used to precise this approximation:

$$\log(q(t)) - \log(q(s)) = \theta(\log_2(1/t) - \log_2(1/s)) + b(\log(1/s))K_\rho\left(\frac{\log(1/t)}{\log(1/s)}\right)(1+o(1)).$$

- Considering $t = i/n$, $s = k_n/n$ and replacing F by its empirical counterpart yield

$$\begin{aligned}\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}) &\simeq \theta(\log_2(n/i) - \log_2(n/k_n)) + b_n K_\rho\left(\frac{\log(n/i)}{\log(n/k_n)}\right)\\ &\simeq \theta(\log_2(n/i) - \log_2(n/k_n)) + b_n\frac{\log(k_n/i)}{\log(n/k_n)},\end{aligned}$$

  where $b_n = b(\log(n/k_n))$.

- Estimation of θ and b_n by a least-squares method:

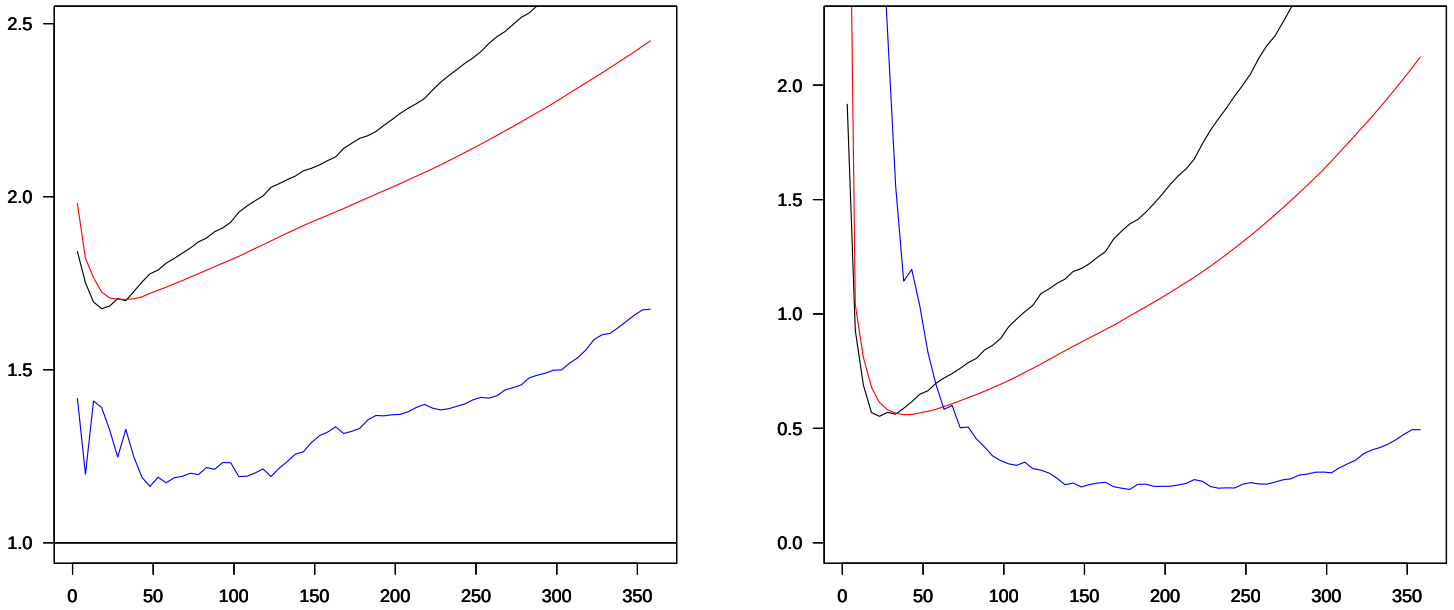
$$(\hat{\theta}_n^R, \hat{b}_n) = \arg\min \sum_{i=1}^{k_n} \left\{ \log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}) - \theta(\log_2(n/i) - \log_2(n/k_n)) - b_n\frac{\log(k_n/i)}{\log(n/k_n)} \right\}^2$$

- Closed-form estimators,
- Asymptotic normality.

4. Simulation study.

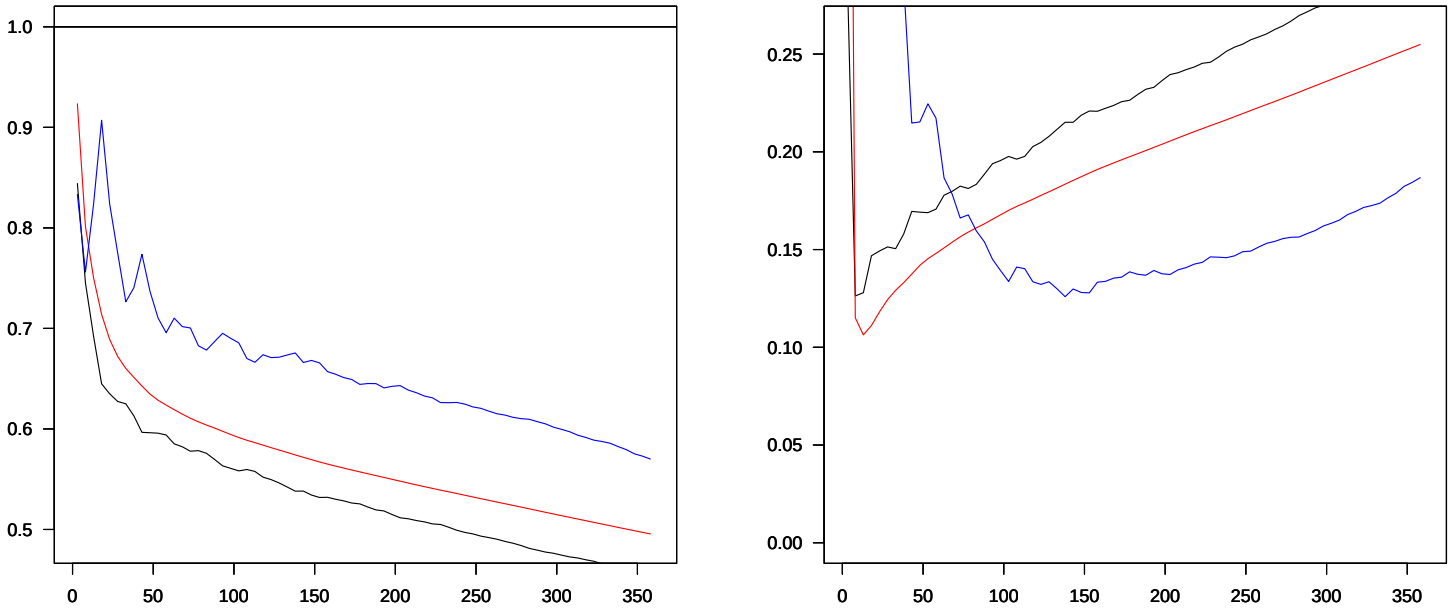
- Comparison of $\hat{\theta}_n^C$ (in black), $\hat{\theta}_n^Z$ (in red) and $\hat{\theta}_n^R$ (in blue) to the true θ (black horizontal line).
- Simulated distributions : Gaussian $\mathcal{N}(0,1)$, Gamma $\Gamma(0.25,1)$, $\Gamma(4,1)$ and Weibull $\mathcal{W}(0.25,0.25)$.
- Sample size $n = 500$, $k_n \in \{2,\ldots,360\}$, 100 replications.
- Computation of the mean estimate (Hill plot, left pannel) and of the Mean Square Error (MSE, right pannel).

Gamma $\Gamma(0.25, 1) \longrightarrow \theta = 1$, $b(x) > 0$



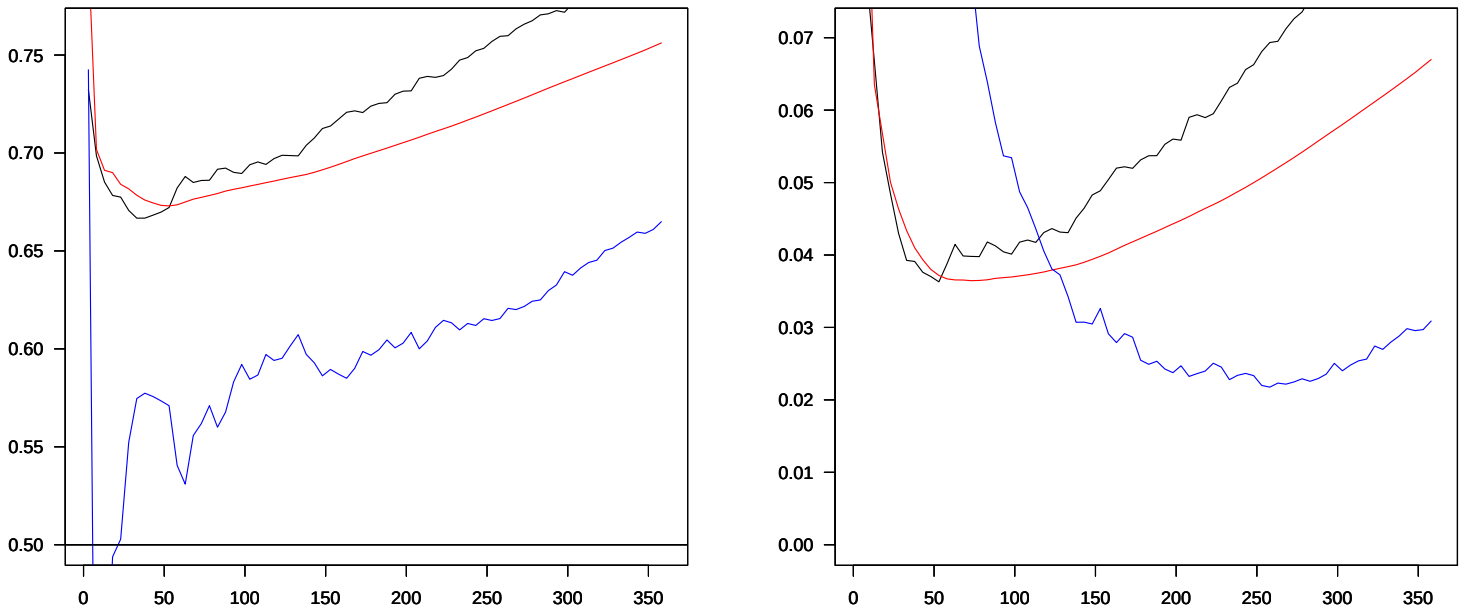
$\hat{\theta}_n^C$ (in black), $\hat{\theta}_n^Z$ (in red) and $\hat{\theta}_n^R$ (in blue)

Gamma $\Gamma(4,1) \longrightarrow \theta = 1,\ b(x) < 0$



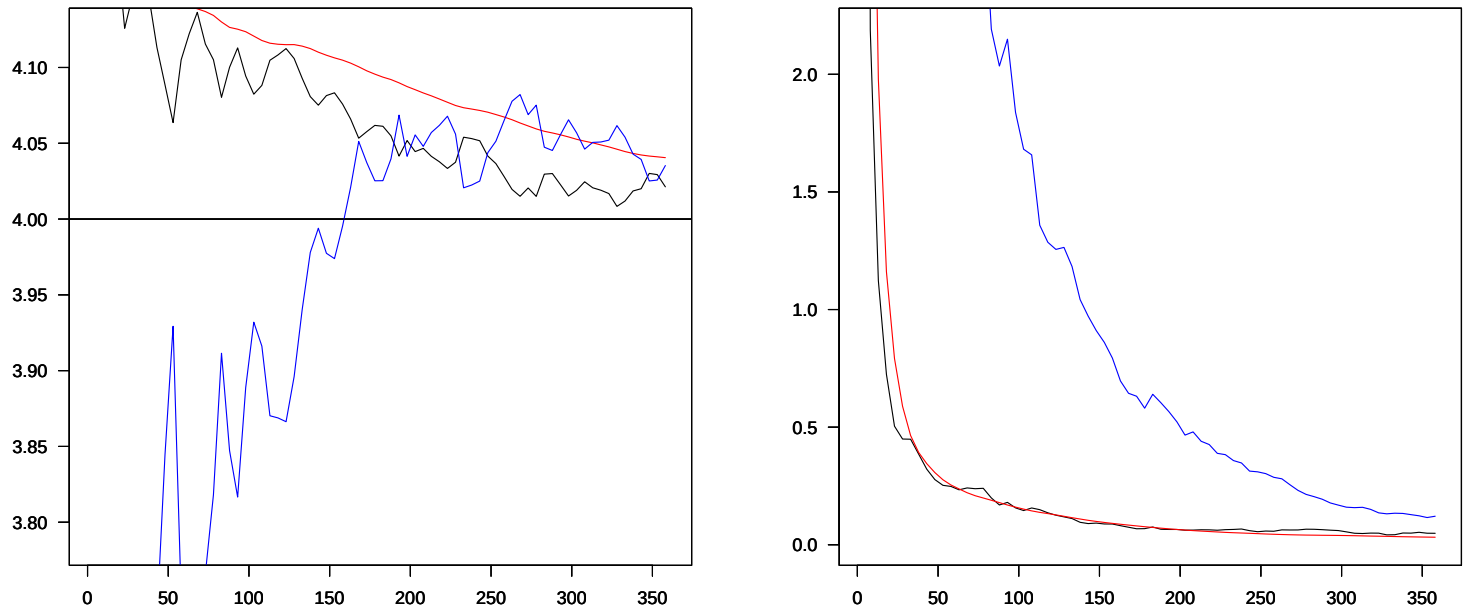
$\hat{\theta}_n^C$ (in black), $\hat{\theta}_n^Z$ (in red) and $\hat{\theta}_n^R$ (in blue)

Gaussian $\mathcal{N}(0,1) \longrightarrow \theta = 1/2,\ b(x) > 0$



$\hat{\theta}_n^C$ (in black), $\hat{\theta}_n^Z$ (in red) and $\hat{\theta}_n^R$ (in blue)

Weibull $\mathcal{W}(0.25, 0.25) \longrightarrow \theta = 4,\ b(x) = 0$



$\hat{\theta}_n^C$ (in black), $\hat{\theta}_n^Z$ (in red) and $\hat{\theta}_n^R$ (in blue)

5. Estimation of extreme quantiles.

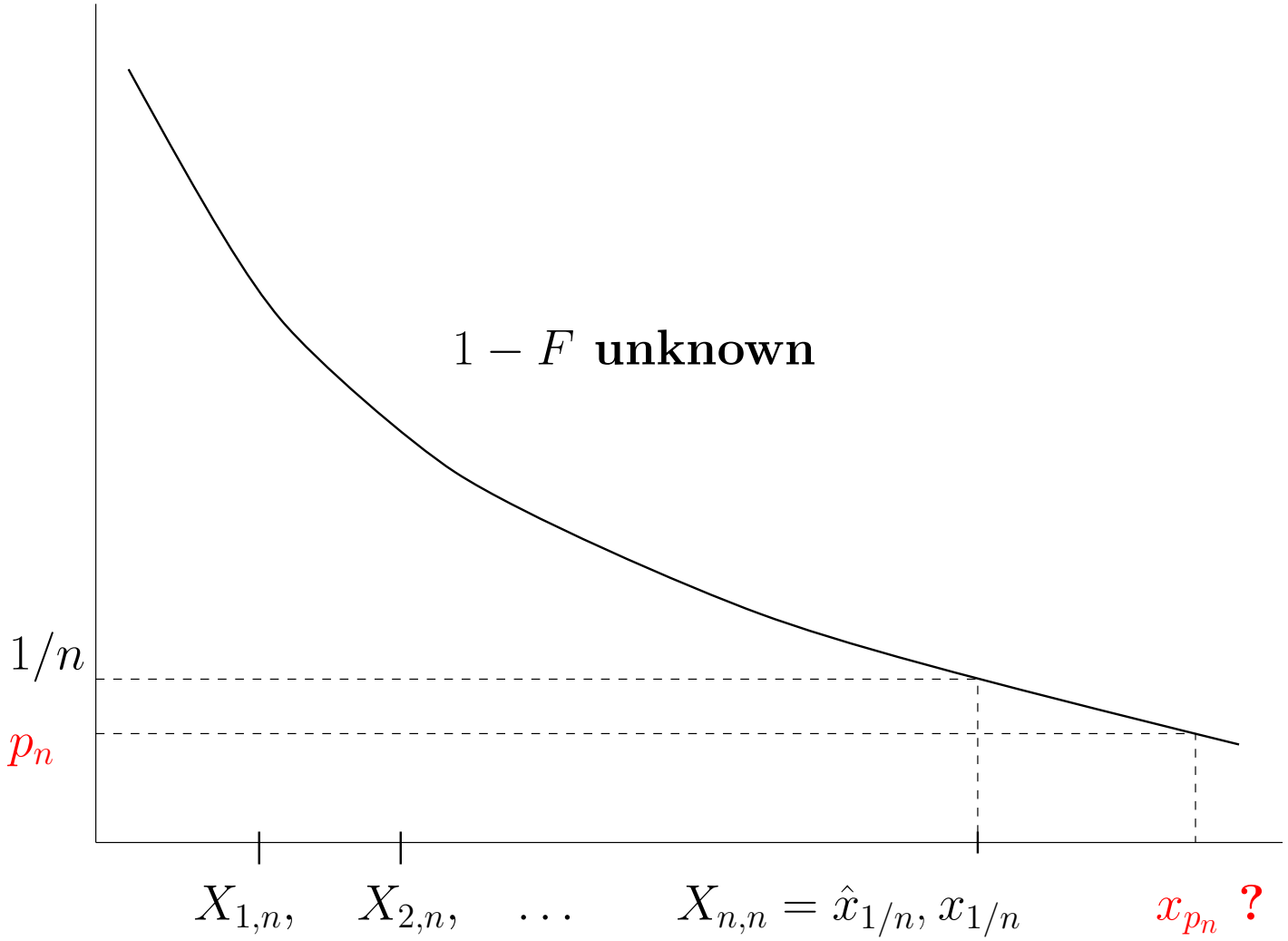
Definition: An extreme quantile x_{p_n} of order $p_n < 1/n$ is defined by:

$$1 - F(x_{p_n}) = p_n.$$

Problems:

- F unknown,
- x_{p_n} is "usually" larger than the maximal observation

$$P(X_{n,n} \leq x_{p_n}) = F^n(x_{p_n}) = (1 - p_n)^n \to 1 \text{ when } n \to \infty \text{ and } np_n \to 0.$$


$1 - F$ **unknown**
$1/n$
p_n
$X_{1,n},$
$X_{2,n},$
$\ldots$
$X_{n,n} = \hat{x}_{1/n}, x_{1/n}$
x_{p_n} **?**

Principle: Recall that, for small t and s

$$\frac{q(t)}{q(s)} = \frac{H^{-1}(\log(1/t))}{H^{-1}(\log(1/s))} = \left(\frac{\log(1/t)}{\log(1/s)}\right)^{\theta} \frac{\ell(\log(1/t))}{\ell(\log(1/s))} \simeq \left(\frac{\log(1/t)}{\log(1/s)}\right)^{\theta}.$$

Considering $t = p_n$, $s = k_n/n$, replacing F by its empirical counterpart and θ by an estimator $\hat{\theta}_n$ yield the following estimator

$$\hat{x}_{p_n}(\hat{\theta}_n) = X_{n-k_n+1,n} \left(\frac{\log(1/p_n)}{\log(n/k_n)}\right)^{\hat{\theta}_n}.$$

Asymptotic normality.

Bias-reduced estimator: Basing on the refined approximation

$$\log(q(t)) - \log(q(s)) \simeq \theta(\log_2(1/t) - \log_2(1/s)) + b(\log(1/s))K_\rho\left(\frac{\log(1/t)}{\log(1/s)}\right).$$

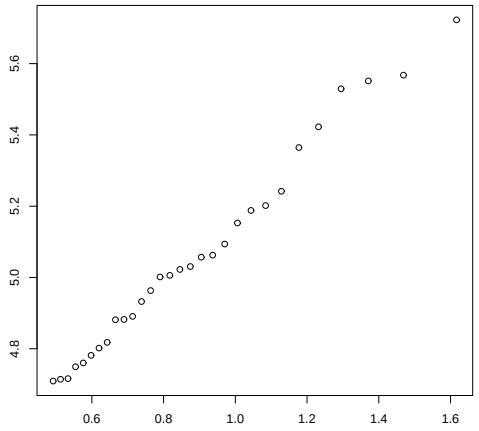
it is natural to introduce

$$\hat{x}_{p_n}^R = X_{n-k_n+1,n}\left(\frac{\log(1/p_n)}{\log(n/k_n)}\right)^{\hat{\theta}_n^R} \exp\left\{\hat{b}_n K_\rho\left(\frac{\log(1/p_n)}{\log(n/k_n)}\right)\right\}.$$

An asymptotic normality theorem is still available.

6. Nidd river data.

- 154 exceedances of the level 65 m^3s^{-1} by the river Nidd (Yorkshire, England) during the period 1934-1969 (35 years).
- The N-year return level is the water level which is exceeded on average once in N years.



Quantile-quantile plot

$k_n = 29,$
$\hat{\theta}_n^R \simeq 0.89,$

Estimation of the 100-year return level:
$\hat{x}_{p_n}(\hat{\theta}_n^R) = 366 m^3 s^{-1}$