Lecture 11. 100 years events - extreme loads

Igor Rychlik

Chalmers
Department of Mathematical Sciences

Probability, Statistics and Risk, MVE300 • Chalmers • May 2011

Example:

Consider a stream of events A, for example times between earthquakes worldwide or accidents in mines in UK. Times for events S_i form PPP with intensity λ year^{-1}. If $\lambda = 1/100$ then A is called 100 years event[1]. (Earthquakes, or accidents in mines, were not 100-years events!)

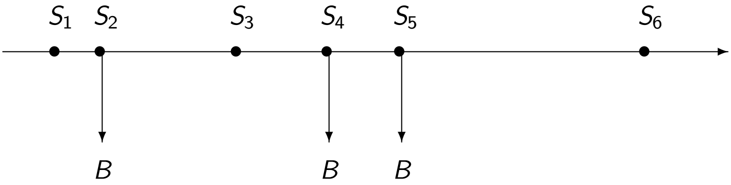
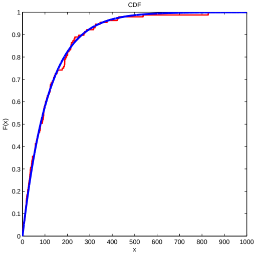
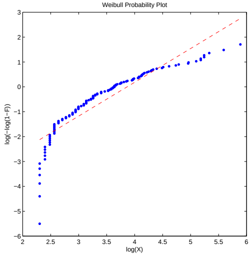


Figure: B that can follow A is 100 years event if $\lambda_{A\cap B} = \lambda\,\mathrm{P}(B) = \frac{1}{100}$, i.e. $\mathrm{P}(B) = \frac{1}{\lambda\,100}$.

[1] An alternative definition is $\mathrm{P}_t(A) = 1/T$ where t is one year. Since $\mathrm{P}_t(A) = 1 - \exp(-\lambda\, t)$ the both definitions are equivalent.





Left figure: the empirical distribution for times between accidents is compared with exponential cdf $\exp(a)$, $a^* = 0.316$ year.[2]

Right figure: observed values of X - the number of perished in the accidents plotted on Weibull probability paper. The fitted parameters are $a^* = 47.7$ and $c^* = 1.36$.

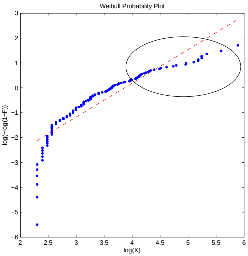
If $B = "X > 150"$ then $\mathrm{P}(B) \approx \exp(-(150/47.1)^{1.36}) = 0.009$.[3]

[2] The intensity of A is $\lambda = 1/0.316$ year^{-1}.
[3] The observed probability is $\mathrm{P}(B) \approx 0.065$.

100-years accident:

Find x_{100} such that for B= "$X > x_{100}$" is a 100 years event.

Solution: $\lambda \mathrm{P}(B) = 0.01, \quad \frac{1}{0.316}\exp(-(x_{100}/47.1)^{1.36}) = 0.01.$



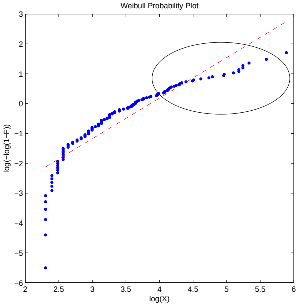
The model gives
$$x^*_{100} = -47.1(-\ln(0.00316))^{1/1.36} = 170.6.$$

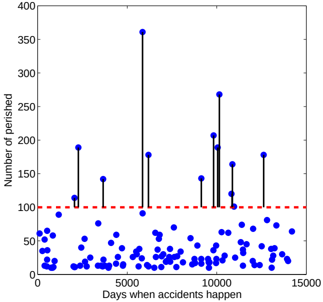
It is too small value. There were 7 accidents during 40 years exceeding 171 perished. The problem is that the central part of data is dominating the fit.

Why not use only the "extreme" observations?

Probability of more than one 100 years events in 40 years period is
$1 - \exp(-0.4) - 0.4\exp(-0.4) = 0.06.$

Peaks over threshold - POT:





Now we will change the definition of initiation event A to major accident:

$A =$ "accident in mines with more than 100 perished"

$\lambda^*_A = 13/40$ years^{-1}. Exceedances over threshold $u = 100$, $H = X - 100$

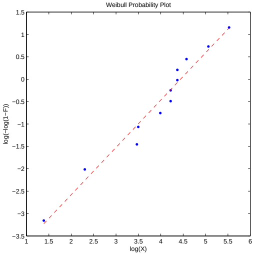
[14, 89, 42, 261, 78, 43, 107, 89, 168, 20, 64, 1, 78]

100-years accident:

Find x_{100} such that B= "$H > x_{100} - 100$" is a 100 years event.

Solution is defined by eq. $\lambda_A \mathrm{P}(B) = 0.01$. The exponential cdf $\exp(a)$ seems to fit well the observed values of H. The estimate a^* is the average 81.1 and the 100 years accident was the one with more than 282 perished:

$$\frac{13}{40}\exp(-(x_{100}-100)/81.1) = 0.01, \qquad x^*_{100} = 100-\ln(\frac{0.4}{13})*81.1 = 282.3.$$



There were one accident in 40 years that could be called 100-years accident. The probability that 100-years accident can happen in 40 years is 0.33.

Probability of more than one is 0.06.

Is the exponential fit accidentally good?. The answer is no!

Tails of a distribution $F_X(x)$.

Some seconds of reflections are needed to see that

$$\mathrm{P}(H > h) = \mathrm{P}(X > u_0 + h | X > u_0), \quad \text{in our example} \quad u_0 = 100.$$

Under suitable conditions on the random variable X, which are always satisfied in our examples, if the threshold u_0 is high, then the conditional probability

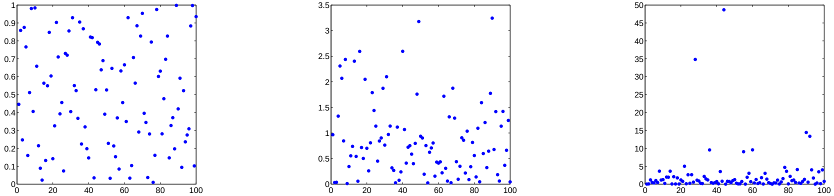
$$\mathrm{P}(X > u_0 + h \,|\, X > u_0) \approx 1 - F(h;\, a, c)$$

where $F(h;\, a, c)$ is a Generalized Pareto distribution (GPD), given by

$$\text{GPD:} \qquad F(h;\, a, c) = \begin{cases} 1 - (1 - ch/a)^{1/c}, & \text{if } c \neq 0, \\ 1 - \exp(-h/a), & \text{if } c = 0, \end{cases} \tag{1}$$

for $0 < h < \infty$ if $c \leq 0$ and for $0 < h < a/c$ if $c > 0$.

In most cases, e.g. when X is normal, Weibul, exponential, log-normal, Gumbel, the tails are exponential. If $c > 0$ there is an upper bound to the tails, e.g. $c = 1$ gives uniform cdf. Generalized Pareto Distribution[4] with $c > 0$ is useful model when there are some physical bounds for X. When $c < 0$ then tails are heavy, i.e. can take very large values,see the following figure where we compare cdf of $c = 1, c = 0$, $c = -1$ and $a = 1$.

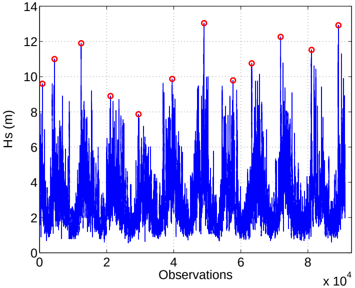


---

[4] Pareto originally used this distribution to describe the allocation of wealth among individuals since it seemed to show rather well the way that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society.

Limitations of standard POT method:

Often the stream of A is not stationary, e.g. storms are more severe in winter than in summer, even parameters in GPD can vary seasonally then more advance methods (based on non-homogeneous Poisson processes) are needed.

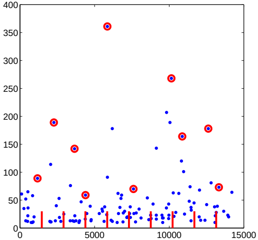


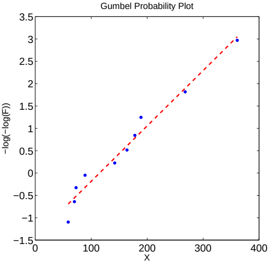
Time series of observations of Hs, 1st July 1993 – 1st July 2003.

The alternative approach is to take yearly maximums.

Extremes:

Let return to the number of perished in mines X and to estimation of the 100 years accident. One way of extracting the extremal events is to take maximums over a period of time usually one year. Then an alternative definition of the 100 years event B can be used. Namely, with $t = 1$ year, B is a 100 years event if $\mathrm{P}_t(B) = 1/100$.[5]





In our case there are in average 3 accidents per year hence not much reduction of data would be achieved by considering yearly maximums. Hence let use maximums over longer period of times, e.g. 4 years.

[5]This definition extends to any T-years event, viz. $\mathrm{P}_t(B) = 1/T$.

Let M_i be maximum number of perished during year i. We assume that M_i are iid. It is easy to see that finding B such that $\mathrm{P}_t(B) = 1/100$ means estimation of x_{100} such that $\mathrm{P}(M_1 > x_{100}) = 1/100$.

Problem: We have data of M, the maximum number of perished during 4 years and not of M_1! Solution:

$$\mathrm{P}(M \leq x) = P(M_1 \leq x, \cdots, M_4 \leq x) = \mathrm{P}(M_1 \leq x)^4.$$

Since $\mathrm{P}(M_1 \leq x) = \mathrm{P}(M \leq x)^{1/4}$ 100-years accident x_{100} is defined by

$$\mathrm{P}(M_1 > x_{100}) = (1 - \mathrm{P}(M \leq x)^{1/4}) = 0.01, \qquad \mathrm{P}(M_1 > x_{100}) \approx \frac{1}{4}\mathrm{P}(M > x),$$

and hence we look for solution of $\mathrm{P}(M > x_{100}) = 0.04.$[6]

For the data the 4-years maximums has Gumbel cdf with $a^* = 67.25$ and $b = 117.8$ giving

$$x^*_{100} = b^* - a^* \ln(-ln(1 - 0.04)) = 332.9.$$

[6] $x^\alpha \approx 1 + \alpha\,(x - 1)$ for $x \approx 1$.

Asymptotic distribution of maximums:

$\mathrm{P}(\max(X_1,\ldots,X_n) \leq x) = F_X(x)^n.$

If there are parameters $a_n > 0$, b_n and non-degenerate probability distribution $G(x)$ such that

$$\mathrm{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = \Big[F(a_n x + b_n)\Big]^n \to G(x)$$

then G is the Generalized Extreme Value distribution

$$\text{GEV:} \quad G(x;\, a, b, c) = \begin{cases} \exp\left(-(1 - c(x-b)/a)_+^{1/c}\right), & \text{if } c \neq 0, \\ \exp\left(-\exp\{-(x-b)/a\}\right), & \text{if } c = 0. \end{cases}$$

[7]

[7] The expression $(1 - c(x-b)/a)_+$ means that $1 - c(x-b)/a \geq 0$ and hence, if $c < 0$, the formula is valid for $x > b + (a/c)$ and if $c > 0$, it is valid for $x < b + (a/c)$. The case $c = 0$ is interpreted as the limit when $c \to 0$ for both distributions.

Gumbel-exponential exceedances:

The extreme value cdf is often used to model variability of demand - load type quantities. Let X be such a variable. Then 100-years demand/load is the value x_{100} such that probability that maximum of X during one year exceeds x_{100} is $1/100$. (Example of X is yearly maximum of the daily rainfalls.) For variable loads GEV are usually good models for the yearly demad/load.

Many real-world maximum loads belong to the GEV cdf with $c = 0$, i.e. Gumbel cd. For instance, if daily loads are normal, log-normal, exponential, Weibull (and some other distributions having the so-called exponential tails) then the yearly (or monthly) maximum loads belong to the Gumbel class of distributions.

Maximum stability:

Recall that a Gumbel distributed r.v. X has the cdf

$$F(x) = \exp(-\mathrm{e}^{-(x-b)/a}), \quad -\infty < x < \infty.$$

Now the maximum $M_n = \max_{1 \le i \le n} X_i$ has distribution

$$\begin{aligned} \mathrm{P}(M_n \le x) &= \left(\exp(-\mathrm{e}^{-(x-b)/a})\right)^n = \exp(-n\mathrm{e}^{-(x-b)/a}) \\ &= \exp(-\mathrm{e}^{-(x-b)/a + \ln n}) = \exp(-\mathrm{e}^{-(x-b-a\ln n)/a}). \quad (2) \end{aligned}$$

Thus, the maximum of n independent Gumbel variables is also Gumbel with b changed to $b + a \ln n$.

Example: Assume that the maximum load on a construction during one year is given by a Gumbel distribution with expectation 1000 kg and standard deviation 200 kg. (Show that $a = 156$, $b = 910$.) Suppose the construction will be used for 10 years. Then the maximum load over the period is Gumbel too with mean $1000 + 156 \cdot \ln 10 = 1.4 \cdot 10^3$ kg and standard deviation 200 kg.

Gumbel or GEV?

Since for many standard models for variable daily loads the maximum load supposed to be Gumbel distributed it is a popular model. Having data one can check whether the more general GEV model explains better the variability of maximums than Gumbel model does.

One can use the deviance:

$$\mathrm{DEV} = 2\big(l(a^*, b^*, c^*) - l(\tilde{a}^*, \tilde{b}^*)\big),$$

where $l(a^*, b^*, c^*)$ is the log-likelihood function and a^*, b^*, c^* are ML estimates of parameters in a GEV cdf, while $l(\tilde{a}^*, \tilde{b}^*)$ is the log-likelihood function and $\tilde{a}^*, \tilde{b}^*$ are ML estimates of parameters in a Gumbel cdf. If the deviance DEV$> \chi^2_\alpha(1)$ then the Gumbel model should be rejected.

One can also construct the asymptotic confidence interval for c that with approximate confidence $1-\alpha$

$$c \in \big[c^* - \lambda_{\alpha/2}\sigma^*_{\mathcal{E}},\ c^* + \lambda_{\alpha/2}\sigma^*_{\mathcal{E}}\big],$$

where $\sigma^*_{\mathcal{E}} \approx \mathrm{D}[C^*]$ (one of the outputs of most programs estimating the parameters in a GEV cdf). If $c = 0$ is not in the interval then the Gumbel model should be rejected.

100 years values:

The T-years maximum ($T = 100, 1000$ years) is equal to the level x_T solving the equation

$$\frac{1}{T} = \mathrm{P}(M_1 > x_T),$$

where M_1 is the yearly maximum modelled as GEV distribution then

$$\begin{aligned} x_T &= b - a\ln(-\ln(1-1/T)) \approx b + a\ln(T), \quad \text{if } c = 0, \\ x_T &= b + \frac{a}{c}\big(1 - (-\ln(1-1/T))^c\big), \quad \text{if } c \neq 0. \end{aligned}$$

Next, using the observed yearly maxima a GEV cdf can be fitted to data, i.e. and estimates $\theta^* = (a^*, b^*, c^*)$ found. Then x_T^* is obtained by replacing a, b, c by a^*, b^*, c^*.

Uncertainty analysis of x_T: Gumbel case:

For $T \geq 50$, $-\ln(1-1/T) \approx 1/T$ and hence $x_T^* = b^* + a^* \ln T$, The ML estimators A^*, B^*, are asymptotically normally distributed with variances

$$\mathrm{V}[A^*] \approx 0.61\frac{(a^*)^2}{n}, \quad \mathrm{V}[B^*] \approx 1.11\frac{(a^*)^2}{n}, \quad \mathrm{Cov}[A^*, B^*] \approx 0.26\frac{(a^*)^2}{n}.$$

and hence with[8]

$$\sigma_{\mathcal{E}}^* = a^* \sqrt{\frac{1.11 + 0.61(\ln T)^2 + 0.52\ln T}{n}},$$

we have that with approximately $1-\alpha$ confidence

$$x_T \in [x_T^* - \lambda_{\alpha/2}\sigma_{\mathcal{E}}^*, x_T^* + \lambda_{\alpha/2}\sigma_{\mathcal{E}}^*].$$

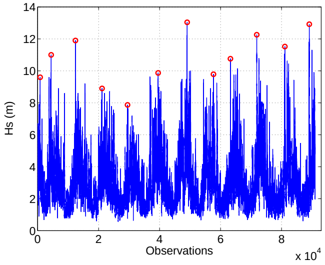
[8]

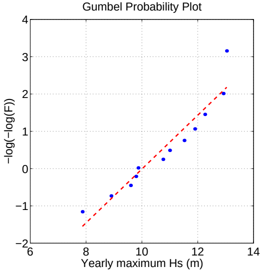
$$\mathrm{V}[X_T^*] \approx 1.11\frac{(a^*)^2}{n} + (\ln T)^2 \cdot 0.61\frac{(a^*)^2}{n} + 2 \cdot 0.26 \cdot \ln T\,\frac{(a^*)^2}{n}$$

Analysis of buoy data

Let study wave data from 1993-2003 given in (left panel). Let extract yearly maxima (marked as circles in left panel). Assume that those are iid.

We choose to model the yearly maxima using a Gumbel distribution. Since only 12 yearly maxima are available it is hard to make a proper validation of the model and we only present the values on a Gumbel probability plot (right panel).





The ML estimates of the parameters are $a^* = 1.5$ and $b^* = 10.0$, which gives the estimate of the 100-year significant wave height

$$x^*_{100} = b^* - a^* \ln(1/100) = 16.9 \quad [\mathrm{m}].$$

Next the standard deviation of the estimation error

$$\sigma^*_{\mathcal{E}} = 1.5\sqrt{\frac{1.11 + 0.52\ln(100) + 0.61(\ln(100))^2}{12}} = 1.756$$

and hence, with approximately 95% confidence, x_{100} is bounded by $16.9 + 1.64 \cdot 1.756 = 19.8$ m.

Rain data at Maiquetia international airport, Venezuela

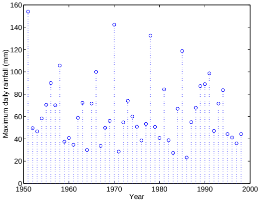
The maximal daily rainfall during the years $1951, \dots, 1998$ was recorded.

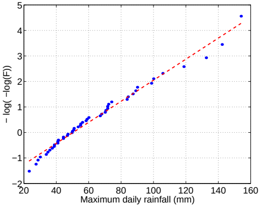
Let choose the GEV class of distributions to model the data. ML estimates are found as $a^* = 19.9$, $b^* = 49.2$ and $c^* = -0.16$ and the standard deviation $\mathsf{D}[C^*] \approx 0.14$.

With approximately 95% confidence, c lies in

$$[-0.16 - 1.96 \cdot 0.14,\ -0.16 + 1.96 \cdot 0.14] = [-0.43, 0.11].$$

We conclude that c^* does not significantly differ from zero[9] .





[9] DEV=1.67 is smaller than $\chi^2_{0.05}(1) = 3.84$ which confirms our conclusions that three-parameter GEV-distribution does not explain the variability of data significantly better than the Gumbel distribution does.

Rain data at Maiquetia international airport, Venezuela

Suppose that one wishes to design a system that takes care of the large amounts of rain water in the tropical climate. Recommendations indicates the safety index $\beta_{\mathsf{HL}} = 3.7$ which corresponds to a risk for failure during one year to be 1 per 10 000. Hence one wishes to estimate x_{10000}.

For a Gumbel cdf with parameters $a^* = 21.5$ and $b^* = 50.9$ the design criterion is that the system should manage $x^*_{10000} = 249$ mm rain fall during one day.

Using formulas shown above we find that, with approximately 95% confidence, $x_{10000} \le 249 + 1.64 \cdot 23.6 = 295$ mm.[10]

In 1999 a catastrophe happened with an accumulated rain during one day of 410 mm, causing around 50 000 deaths. The conclusion was that "the impossible had happened".

[10]The confidence level is achieved under the assumption that the Gumbel distribution is the correct model for maximal daily rain during one year.

The model error

Before the 1999 maximum was observed, there were no indications that the Gumbel model was not correct and a natural question is why not always use the GEV model to describe the variability of yearly maxima, instead of assuming that $c = 0$?[11].

In the case studied here, including one more parameter c to the model would not explain better the variability of data but made the design value more uncertain causing additional costs to meet the required safety level.

Let compute x^*_{10000} using the GEV model estimated for the data from the years 1951-1998, i.e. $a^* = 19.9$, $b^* = 49.2$ and $c^* = -0.16$. The design load $x^*_{10000} = 468$ mm and, with approximately 95% confidence, it is smaller than 1030 mm.

Clearly, using the design load 468 mm, one could be better prepared for the cathastrophe that occurred 1999.

[11] Often in statistical practice, it is not recommended to use more complicated models than needed to describe data adequately