Société de Calcul Mathématique, S. A.
Algorithmes et Optimisation

The probability of extreme events:

Explicit computations

Bernard Beauzamy

September 2012

I. Introduction

The occurrence of extreme events (high temperatures, high waves, strong earthquakes, and so on) has become an increasingly important social concern. This is rather easy to understand: our societies have now more than a hundred years of data, for many phenomena, which is enough to establish tables for the ordinary ones, but not enough for the rare ones. Since moreover there is an increasing social concern about "risks" in general, research tends to concentrate now on rare situations.

So far, extreme phenomena have been treated by means of specific and empirical probability law, such as Gumbel, the merit of which is that they depend on very few parameters, and so are easy to tune in cases where data are scarce. But such law have no real value ; they are only of academic merit. They do not describe any real world situation.

In 2009, in the framework of contracts with the "Caisse Centrale de Réassurance" (Paris), we developed probabilistic methods in order to estimate the probability of extreme events. These methods :

- Make full use of all existing records ;
- Make no fictitious assumption at all. There is no "parametric law" behind the construction, the only assumption being that, the more extreme the phenomenon is, the smaller its probability will be.

However, using this method, computations on real life situations were quite hard to perform, since they required complicated multiple integrals.

In 2010, a new approach, based upon a Monte-Carlo type method, was proposed by Peter Robinson [3] ; this method works well if the number of records and the number of classes is not too high. But, in high-dimensional spaces, Monte-Carlo methods do not perform in a satisfactory manner : see our book [4] for a description of this fact.

Here, we present a new approach allowing explicit and fast computation of the integrals which are the basis of our method. This approach does not rely upon any Monte-Carlo procedure ; it is completely explicit and deterministic.

II. General presentation of the problem

Let us describe it on a specific example, namely the temperatures in Paris. We start with a list of records, which may be of the following form:

temp	nb of days	temp	nb of days
35	9	38	1
35,2	7	38,2	1
35,4	8	38,4	1
35,6	11	38,6	2
35,8	6	38,8	0
36	3	39	0
36,2	7	39,2	0
36,4	6	39,4	3
36,6	6	39,6	0
36,8	4	39,8	0
37	3	40	0
37,2	4	40,2	0
37,4	1	40,4	1
37,6	2	40,6	0
37,8	2	40,8	0
		41	0

Table 1: data for extreme temperatures

In this case, we consider the temperature as "extreme" if it is above 35°C ; the second column is the number of occurrences of the phenomenon, namely the number of days where the given temperature was observed (during the total observation period, which is here 140 years).

For each temperature ϑ_k (between 35°C and 41°C), we want to have an estimate of the probability p_k of this temperature; the discretization is made here in steps of 0.2°C. This estimate is given under the form of random variables Z_k and p_k is the expectation of Z_k. In other words, take for instance the value 40°C. The random variable Z_{40} describes the possible values for the probability and the chosen value p_{40} is the average of these possible values, that is the expectation of the random variable.

Let n_k be the number of occurrences of the $k-$ th value: they are the values of the right column in the table above. For instance, $n_{40}=0$. Let K be the total number of classes, here $K=31$.

We need to have a consistent set of probabilities, namely $\sum_k p_k = 1$.

The theory developed in [1] indicates that the joint law of our set $\left(Z_1,...,Z_K\right)$ will be given by :

$$f(\lambda_1,...,\lambda_K) = c\,1_S\left(\lambda_1,...,\lambda_K\right)\lambda_1^{n_1}\cdots\lambda_K^{n_K}$$

where:

- 1_S is the indicator function of the set :

$$S=\left\{(\lambda_1,...,\lambda_K);\ \lambda_1 \geq \cdots \geq \lambda_K \geq 0, \sum_{k=1}^{K}\lambda_k = 1\right\},$$

  that is the function which is 1 inside this set and 0 outside ;

- c is a normalization constant : the multiple integral of $f(\lambda_1,...,\lambda_K)$ should be 1.

The marginal law for each Z_k has density :

$$f_k(\lambda) = \int \ldots \int_S f(\lambda_1,...,\lambda_{k-1},\lambda,\lambda_{k+1},...,\lambda_K)d\lambda_1...d\lambda_{k-1}d\lambda_{k+1}...d\lambda_K$$

and the assigned value for each p_k will be the expectation :

$$p_k = \int_0^1 \lambda\, f_k(\lambda)d\lambda$$

So, we want to compute the integral of a monomial :

$$M(x_1,...,x_K) = x_1^{n_1}\cdots x_K^{n_K}$$

on the simplex S.

We introduce new variables :

$y_K = x_K$

$y_k = x_k - x_{k+1} \quad k = 1,...,K-1.$

So $y_k \geq 0$, $k = 1,...,K.$

Conversely, the x_k's can be computed from the y_k's :

$x_K = y_K$

$x_k = y_k + y_{k+1} + \cdots + y_K$, $k = 1,\ldots,K$.

So the Jacobian of the transformation from the x_k's to the y_k's is 1.

The condition $\sum_{k=1}^{K} x_k = 1$ becomes :

$$y_K + y_K + y_{K-1} + \cdots + y_K + \cdots + y_k + \cdots + y_K + \cdots + y_1 = 1$$

that is :

$$K y_K + (K-1) y_{K-1} + \cdots + k y_k + \cdots + y_1 = 1.$$

With these new variables, the monomial M becomes :

$$P\left(y_1,\ldots,y_K\right) = \left(\sum_{j=1}^{K} y_j\right)^{n_1} \cdots \left(\sum_{j=k}^{K} y_j\right)^{n_k} \cdots \left(\sum_{j=K}^{K} y_j\right)^{n_K}$$

We have to integrate the polynomial $P(y_1,\ldots,y_K)$ on the simplex :

$$S_1 = \left\{\left(y_1,\ldots,y_K\right); y_k \geq 0, \sum_{k=1}^{K} k y_k = 1\right\}$$

Let us make one more change of variables. We set :

$z_k = k y_k$, $k = 1,\ldots,K$

Then the Jacobian of this transformation is $K!$

The polynomial $P(y_1,\ldots,y_K)$ becomes :

$$Q\left(z_1,\ldots,z_K\right) = \left(\sum_{j=1}^{K} \frac{z_j}{j}\right)^{n_1} \cdots \left(\sum_{j=k}^{K} \frac{z_j}{j}\right)^{n_k} \cdots \left(\sum_{j=K}^{K} \frac{z_j}{j}\right)^{n_K}$$

and the set of integration is :

$$S_2 = \left\{ \left(z_1, ..., z_K \right); z_k \geq 0, \sum_{k=1}^{K} z_k = 1 \right\}$$

The integral we want is :

$$I = \frac{1}{K!} \int_{S_2} \left(\sum_{j=1}^{K} \frac{z_j}{j} \right)^{n_1} \cdots \left(\sum_{j=k}^{K} \frac{z_j}{j} \right)^{n_k} \cdots \left(\sum_{j=K}^{K} \frac{z_j}{j} \right)^{n_K} dz_1 ... dz_K \qquad (1)$$

and we recall from [BB1], Chapter 14, §9 that :

$$\int_{S_2} z_1^{n_1} \cdots z_K^{n_K} dz_1 \cdots dz_K = \frac{n_1! ... n_K!}{\left(N + K - 1 \right)!} \qquad (2)$$

with $N = n_1 + \cdots + n_K$.

III. An induction formula

We now show how to compute the integral (1) by induction.

We write:

$$\left(\sum_{j=1}^{K} \frac{z_j}{j} \right)^{n_1} = \left(z_1 + \sum_{j=2}^{K} \frac{z_j}{j} \right)^{n_1} = \sum_{m_1=0}^{n_1} \binom{n_1}{m_1} z_1^{m_1} \left(\sum_{j=2}^{K} \frac{z_j}{j} \right)^{n_1 - m_1}$$

So:

$$Q\left(z_1, ..., z_K \right) = \sum_{m_1=0}^{n_1} \binom{n_1}{m_1} z_1^{m_1} \left(\sum_{j=2}^{K} \frac{z_j}{j} \right)^{n_1 + n_2 - m_1} \left(\sum_{j=3}^{K} \frac{z_j}{j} \right)^{n_3} \cdots \left(\sum_{j=k}^{K} \frac{z_j}{j} \right)^{n_k} \cdots \left(\sum_{j=K}^{K} \frac{z_j}{j} \right)^{n_K} \qquad (3)$$

That is:

$$Q\left(z_1, ..., z_K \right) = \sum_{m_1=0}^{n_1} \binom{n_1}{m_1} z_1^{m_1} Q_2$$

with:

$$Q_2\left(z_2, ..., z_K ; m_1 \right) = \left(\sum_{j=2}^{K} \frac{z_j}{j} \right)^{n_1 + n_2 - m_1} \left(\sum_{j=3}^{K} \frac{z_j}{j} \right)^{n_3} \cdots \left(\sum_{j=k}^{K} \frac{z_j}{j} \right)^{n_k} \cdots \left(\sum_{j=K}^{K} \frac{z_j}{j} \right)^{n_K}$$

Similarly:

$$\left(\sum_{j=2}^{K}\frac{z_j}{j}\right)^{n_1-m_1+n_2}=\left(\frac{z_2}{2}+\sum_{j=3}^{K}\frac{z_j}{j}\right)^{n_1+n_2-m_1}=\sum_{m_2=0}^{n_1+n_2-m_1}\binom{n_1+n_2-m_1}{m_2}\left(\frac{z_2}{2}\right)^{m_2}\left(\sum_{j=3}^{K}\frac{z_j}{j}\right)^{n_1+n_2-m_1-m_2}$$

which gives:

$$Q_2=\sum_{m_2=0}^{n_1+n_2-m_1}\binom{n_1+n_2-m_1}{m_2}\left(\frac{z_2}{2}\right)^{m_2}\left(\sum_{j=3}^{K}\frac{z_j}{j}\right)^{n_1+n_2+n_3-m_1-m_2}\left(\sum_{j=4}^{K}\frac{z_j}{j}\right)^{n_4}\cdots\left(\sum_{j=k}^{K}\frac{z_j}{j}\right)^{n_k}\cdots\left(\sum_{j=K}^{K}\frac{z_j}{j}\right)^{n_K}$$

Set $\nu_k=n_1+\cdots+n_k$, $\mu_k=m_1+\cdots+m_k$; repeating the procedure, we get:

$$Q_2=\sum_{m_2=0}^{n_1+n_2-m_1}\binom{n_1+n_2-m_1}{m_2}\left(\frac{z_2}{2}\right)^{m_2}Q_3$$

with:

$$Q_3\left(z_3,\dots,z_K;m_1,m_2\right)=\left(\sum_{j=3}^{K}\frac{z_j}{j}\right)^{\nu_3-\mu_2}\left(\sum_{j=4}^{K}\frac{z_j}{j}\right)^{n_4}\cdots\left(\sum_{j=k}^{K}\frac{z_j}{j}\right)^{n_k}\cdots\left(\sum_{j=K}^{K}\frac{z_j}{j}\right)^{n_K}$$

More generally:

$$Q=\sum_{m_1=0}^{n_1}\binom{\nu_1}{m_1}z_1^{m_1}\sum_{m_2=0}^{\nu_2-\mu_1}\binom{\nu_2-\mu_1}{m_2}\left(\frac{z_2}{2}\right)^{m_2}\cdots\sum_{m_{k-1}}^{\nu_{k-1}-\mu_{k-2}}\binom{\nu_{k-1}-\mu_{k-2}}{m_{k-1}}\left(\frac{z_{k-1}}{k-1}\right)^{m_{k-1}}Q_k$$

with:

$$Q_k\left(z_k,\dots,z_k;m_1,\dots,m_{k-1}\right)=\left(\sum_{j=k}^{K}\frac{z_j}{j}\right)^{\nu_k-m_{k-1}}\left(\sum_{j=k+1}^{K}\frac{z_j}{j}\right)^{n_{k+1}}\cdots\left(\sum_{j=K}^{K}\frac{z_j}{j}\right)^{n_K}$$

$$Q_{K-1}=\left(\frac{z_{K-1}}{K-1}+\frac{z_K}{K}\right)^{\nu_{K-1}-\mu_{K-2}}\left(\frac{z_K}{K}\right)^{n_K}$$

which gives:

$$Q_{K-1}=\sum_{m_{K-1}}^{\nu_{K-1}-\mu_{K-2}}\binom{\nu_{K-1}-\mu_{K-2}}{m_{K-1}}\left(\frac{z_{K-1}}{K-1}\right)^{m_{K-1}}Q_K$$

with:

$$Q_K\left(z_K;m_1,\dots,m_{K-1}\right)=\left(\frac{z_K}{K}\right)^{\nu_K-\mu_{K-1}}$$

Since all coefficients are positive, there is no cancellation, and each polynomial gives its own contribution. We denote by EQ_j the contribution of the $j-$ th polynomial. We get :

$$EQ_K\left(m_1,\ldots,m_{K-1}\right)=\frac{\left(\nu_K-\mu_{K-1}\right)!}{K^{\nu_K-\mu_{K-1}}}$$

We observe that this quantity depens on $\mu_{K-1}=m_1+\cdots+m_{K-1}$ only : this will be useful for practical computations.

Also:

$$EQ_{K-1}\left(\mu_{K-2}\right)=\sum_{m_{K-1}=0}^{\nu_{K-1}-\mu_{K-2}}\binom{\nu_{K-1}-\mu_{K-2}}{m_{K-1}}\frac{m_{K-1}!}{\left(K-1\right)^{m_{K-1}}}EQ_K\left(\mu_{K-1}\right)$$

More generally,

$$EQ_{k-1}\left(\mu_{k-2}\right)=\sum_{m_{k-1}=0}^{\nu_{k-1}-\mu_{k-2}}\binom{\nu_{k-1}-\mu_{k-2}}{m_{k-1}}\frac{m_{k-1}!}{\left(k-1\right)^{m_{k-1}}}EQ_k\left(\mu_{k-1}\right)$$

and:

$$EQ=EQ_1=\sum_{m_1=0}^{n_1}\binom{\nu_1}{m_1}\frac{m_1!}{1^{m_1}}EQ_2\left(\mu_1\right)$$

Finally, by formulas (1) and (2), we get:

$$I=\frac{EQ}{\left(N+K-1\right)!K!}$$

We can write the explicit estimate:

$$I=\frac{1}{K!\left(N+K-1\right)!}\sum_{m_1=0}^{n_1}\binom{n_1}{m_1}\frac{m_1!}{1^{m_1}}\sum_{m_2=0}^{\nu_2-m_1}\binom{\nu_2-m_1}{m_2}\frac{m_2!}{2^{m_2}}\cdots\sum_{m_k=0}^{\nu_k-\mu_{k-1}}\binom{\nu_k-\mu_{k-1}}{m_k}\frac{m_k!}{k^{m_k}}\cdots$$
$$\times\sum_{m_{K-1}=0}^{\nu_{K-1}-\mu_{K-2}}\binom{\nu_{K-1}-\mu_{K-2}}{m_{K-1}}\frac{m_{K-1}!}{(K-1)^{m_{K-1}}}\frac{(\nu_K-\mu_{K-1})!}{K^{\nu_K-\mu_{K-1}}}$$

IV. A first example

We treat the situation of temperatures, taken from [1]. The data are given in the table 1 above.

Here are the results of the evaluation:

temperature	nb of days	corrected proba	temperature	nb of days	corrected proba
35	9	0,118	38	1	0,020
35,2	7	0,097	38,2	1	0,018
35,4	8	0,087	38,4	1	0,016
35,6	11	0,079	38,6	2	0,015
35,8	6	0,069	38,8	0	0,013
36	3	0,062	39	0	0,011
36,2	7	0,057	39,2	0	0,010
36,4	6	0,052	39,4	3	0,009
36,6	6	0,047	39,6	0	0,008
36,8	4	0,042	39,8	0	0,006
37	3	0,037	40	0	0,005
37,2	4	0,033	40,2	0	0,004
37,4	1	0,029	40,4	1	0,003
37,6	2	0,026	40,6	0	0,002
37,8	2	0,023	40,8	0	0,002
			41	0	0,001

Table 2: results of the evaluation

Here are the two sets of data on the same graph. In red, the "rough" probability, estimated simply from the number of occurrences. In blue, the data corrected using our method. The x axis represents the temperature and the y axis the probability of that temperature.

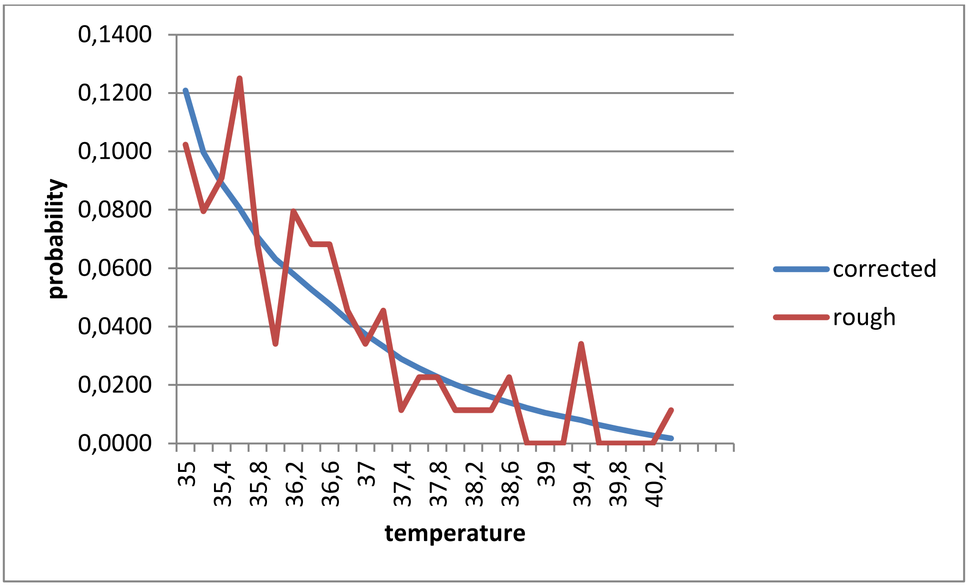

Figure 3: graph of corrected values

V. A new example

We treat here the situation of 3 classes, with the numbers $n_1 = 19, n_2 = 19, n_3 = 1$. This situation arises naturally, in the following context : some accident may occur with extremely small probability. If it occurs, we consider the costs of the accident, and we have the following costs :

$cst > 0.1$ U with probability 0.975
$cst > 0.4$ U with probability 0.500
$cst > 1$ U with probability 0.025

The maximal cost is 1.9 U. The letter U indicates the unit of cost, whatever it may be.

So we have three intervals :

Interval of cost	Probability
0.1-0.4	0.475
0.4-1	0.475
1-1.9	0.025

Table 4: the data with three intervals

The total probability is not 1, since there is a 0.025 probability that the cost may be ≤ 0.1.

If we restrict ourselves to the situation of extreme accidents, that is of costs > 0.1, we have the following table :

Interval of cost	Probability
0.1-0.4	19/39
0.4-1	19/39
1-1.9	1/39

Table 5: the extreme accidents

In order to describe this situation, we assume that a total of 39 accidents occurred, with 19 in the first interval, 19 in the second and 1 in the third.

We take 0.3 as the step of subdivision. So the first interval contains just one class, the second contains two classes, namely $0.4-0.7$ and $0.7-1$ and the third contains three classes, namely $1-1.3$, $1.3-1.6$, $1.6-1.9$. So, altogether, we now have 6 classes of equal width; we want to estimate the probabilities $p_1, ..., p_6$ of these classes. These probabilities are evaluations of random variables $Z_1, ..., Z_6$ satisfying, for the previous paragraph :

$$Z_1 \geq Z_2 \geq \cdots \geq Z_6$$

$$Z_1 + \cdots + Z_6 = 1$$

From the table above, we can assume the following number of accidents in each class:

class	number of accidents
1	57
2	28
3	29
5	1
6	1
7	1

Table 6: the number of accidents per class

Indeed, we have three classes in the last interval, and if 39 accidents occur, only 1 is in this interval. So we have to assume that $39 \times 3 = 117$ accidents occur, with 1 in each class of the final interval. The 57 accidents of the original second class are put evenly into the two new classes 2 and 3.

Here are the results:

class	number of accidents	estimated probability	cost
1	57	0,472	0,1-0,4
2	28	0,263	0,4-0,7
3	29	0,216	0,7-1
4	1	0,026	1-1,3
5	1	0,015	1,3-1,6
6	1	0,008	1,6-1,9

Table 7 : the estimated probabilities and costs

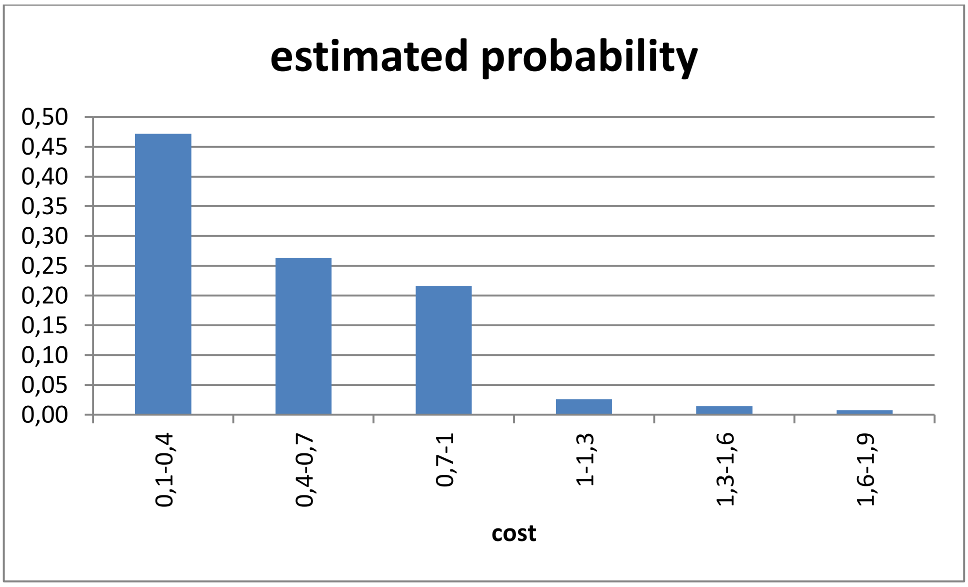

Figure 8: the corresponding graph