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

Robust Mathematical Methods

for Extremely Rare Events

by Bernard Beauzamy
SCM SA

August 2009

Executive summary

By means of conditional probabilities, we build robust methods, in order to handle extremely rare events. These methods do not rely upon any pre-defined probability law (such as Gumbel, exponential, Poisson, and so on), since the use of such laws is totally academic and has no reality at all. Our methods give the probability of the rare events, but, which is more, they give a probability law for this probability, with expectation, variance, confidence intervals, etc.

Here is an example of what we obtain, using the historical records of temperature in Paris since 1873, for the value 40.4°C (absolute record, on July 28th, 1947) and for 41°C (which has never been recorded):

	proba per year	proba not in 100 years	proba not in 1000 years	duration for 95 % proba (years)	duration for 50 % proba (years)
40,4	0,003	0,769	0,072	1 140	264
41	0,001	0,913	0,401	3 279	759

It says for instance that we have 50% chances to see again 40.4°C in the next 250 years, and we are completely sure to see it again in the next 1,000 years. So, clearly, this temperature is not of "centennial" type (to be seen every 100 years), but rather of type 500. The temperature 41°C is to be seen in a range 700 to 3,000 years : it is of "millennium" type.

Technically, our method relies upon the evaluation of a very complicated multiple integral, upon a specific volume in a multi-dimensional space. This evaluation is quite delicate, and quite surprisingly, it is done using symbolic exact computation, in Maple.

Acknowledgements

The evaluation of the probability of rare events has been of interest for us for many years ; the chapter XVIII of my book "Méthodes probabilistes pour l'étude des phénomènes reels" [1] is devoted to this topic. However, in this book, we deal with a single value, with no comparison with others, as it is done here.

The need for more complete tools came up from :

- Several discussions with experts from "Snecma Propulsion Solide", Le Haillan : they are interested in better information about the reliability of some components ;

- Several discussions with experts from the "Caisse Centrale de Réassurance", Paris : they are interested by the probability of rare events in climatology.

It is our pleasure to thank these organizations for their interest in these matters.

I. Presentation of the problem

Extremely rare events, as such, are an interesting challenge for a mathematician : since, by definition, very few data are available, they defy the common laws of statistics.

The commonly used method, in order to compensate the lack of data, is to rely upon some conventional probability law. For instance, to say that high temperatures in a given city obey a probability law of Gumbel type, or rare events of seismic type obey a Poisson law, and so on. These laws have the obvious advantage to depend on very few parameters, so they easily fit with rare observations. But, even if they are commonly accepted, such laws are totally fictitious, and do not describe at all the reality. More precisely, some specific contracts we had in 2008 with the French "Commissariat pour l'Energie Atomique" showed that, in epidemiology and seismology, the reason of the poor prediction by the models was precisely the inappropriate choice of such laws.

Our ambition in this article is to build probabilistic methods, especially designed in order to handle rare events. These probabilistic methods are related to "conditional probabilities" ; they answer the following type of question : assuming that events of "magnitudes" $m_1,...,m_k$ have been observed in the past, what is the probability to observe an event of magnitude m? Here, m may be any of the $m_1,...,m_k$ (already observed) or any new value, smaller or bigger.

Such probabilistic methods are totally "robust", in the sense that they require no assumption at all : they do not rely on any parametric law (Poisson, exponential, Gumbel, and so on), and do not require any adjustment (such as linear or polynomial regression).

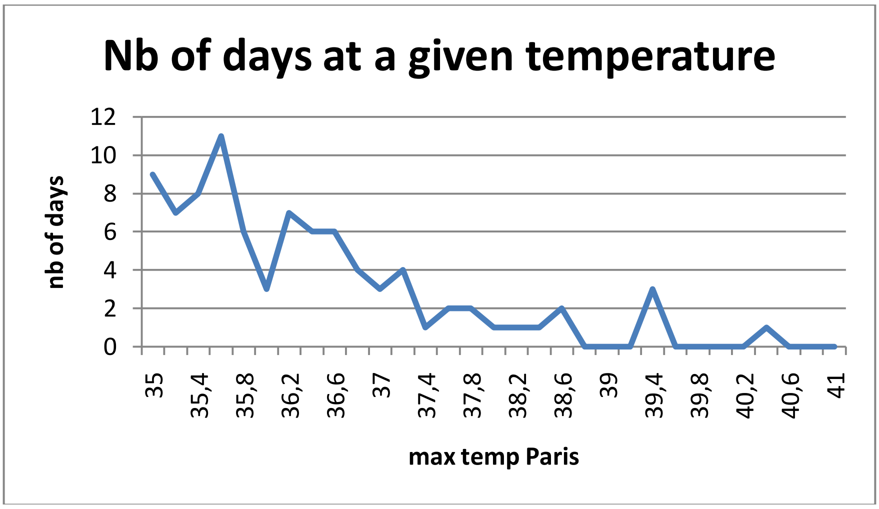
The methods will be built on an example, that of temperatures in Paris, for clarity. Quite obviously, they can be transferred to all rare events.

II. Extreme temperatures in Paris

As it is well-known, the temperature record in Paris is 40.4°C, which was observed on July 28th, 1947. This is based upon meteorological observations in Paris, starting 01/01/1873 and finishing 31/12/2004, that is 48 212 days (courtesy of Meteo-France, more recent data are not easily available).

An interesting remark, to be made before any theory is developed, is that these high temperatures are very "irregular". For instance, no temperature in the range 39.6 – 40.2 has ever been recorded.

Let us look at the following graph, which will be the key feature for our study :



Graph 1 : high temperatures in Paris

The x axis is the temperature, step 0.2°C. The y axis indicates the number of days (among the $N = 48\,212$ records), during which this temperature was achieved. For instance (first data), there were nine days during which the temperature was between 35°C and 35.2°C (when we speak of "temperature", we mean : maximal temperature of a given day, not average).

We see, as we said before, that this curve is quite irregular. One would expect it to be decreasing, but it is not. The reason is of course that the observations are insufficient : 132 years are not enough to reveal all possible temperatures. We may consider that, if we waited for 1,000 or 10,000 years, the temperatures 39°C or 39.8°C would show up. But how long will this take ? And, beyond the record of 40.4, no higher temperature ever appeared. If we wait for 10,000 years, there is no reason that, for instance, 41°C might not appear. But again how long will this take ? The purpose of this article is to answer these two questions, using probabilistic techniques.

III. Discussion about the season

Let $p = f(\vartheta)$ be the function which, for a given temperature (step 0.2°C) returns its probability. But what does this mean ?

It means that if we take a given temperature, say 36.6°C, and a large number of observations (say a billion days), the proportion of days with this temperature will be $p = f(36.6)$.

This approach of the probability, using the empirical law of large numbers, is correct and allows us to avoid any reference to the situation of the day in the year. Of course, it is very unlikely to meet 36.6°C in January : this has never been met in Paris, but why should we exclude it a priori ? Temperature above 35°C have been met in June, July, August and September, so the "possible season" is not easy to define, and depends upon the temperature we investigate.

Still, we do not want to use a considerable dataset, most of which would be irrelevant. So we will restrict our data to the four months of June, July, August, and September, for the years from 1873 to 2004 : this gives 16,104 data. We will restrict our investigation of high temperatures to $\vartheta \geq 35$; we checked that among the 132 years, no such temperatures occurred during one of the other 8 months. The graph above is the same, but was obtained from 16,104 data, and not 48,212.

The first task we have to perform is to "regularize" the above graph.

IV. Regularizing the high-temperature curve

Let $N = 16{,}104$ be the total number of observations, and K be the number of possible values. Since we investigate the range 35°C to 40.4°C, step 0.2°C, we have 28 possible values v_k : $v_1 = 35,\ v_2 = 35.2, \ldots, v_k = 35 + 0.2 \times (k-1), \ldots, v_{28} = 40.4$, so $K = 28$.

For $k = 1, \ldots, K$, let n_k be the number of times the value v_k has been observed ; this is given by the following table (from which the graph above comes) :

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 2 : number of days for each temperature

So we have a total of $\sum_{k=1}^{K} n_k = 88$ days where the temperature has been $\geq 35°C$.

Let finally p_k ($k = 1, \ldots, K$) be the (unknown) probability of the value v_k. We want to estimate it from the observations n_k. But in fact, we do not just want an estimate : we want a probability law for it.

Recall some elements of the theory developed in [1], chapter 14 : if n accidents have been observed, over N experiences, the probability of an accident (which is itself a random variable) has a density :

$$f_{n,N}(\lambda) = c\lambda^n (1-\lambda)^{N-n}$$

where c is a normalization factor (so the integral is 1).

We consider it as reasonable to admit that, for high temperatures, the probabilities p_k will be decreasing : it is less likely to meet 36.6 than to meet 36.4. Logically, this is not completely clear : there might be two "regimes", for instance high wind and low wind, with different properties in terms of probabilities. Nevertheless, let us take this assumption of decreasing probabilities, at least for temperatures above 35°C.

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

Then the joint law of our set $(p_1,...,p_K)$ will be given by :

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

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 p_k has density :

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

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

$$\overline{p_k} = \int_0^1 \lambda f_k(\lambda)d\lambda$$

In [BB2], we showed how to compute explicitly the marginal laws for a simpler joint density: we did not have the restriction $\lambda_1 \geq \cdots \geq \lambda_K$. Here, the computation is more difficult, both conceptually and technically. We proceed in two steps.

Step 1 : Computing the integral

$$I = \int_V x_1^{n_1}\cdots x_K^{n_K}dx_1...dx_K$$

where :

$$V = \left\{(x_1,...,x_K)\,; x_1 \geq ... \geq x_K \geq 0, \sum_k x_k \leq 1\right\}$$

This step is necessary at the theoretical level, in order to understand step 2, but the computations will not be performed in practice.

For any k, we define $s_k = 1 - x_k - ... - x_K$, with $s_{K+1} = 1$.

First, we set :

$$C_1 = \int_{x_1 = x_2}^{s_2} x_1^{n_1} dx_1$$

Then of course :

$$C_1 = \frac{1}{n_1 + 1}\left(s_2^{\,n_1+1} - x_2^{n_1+1} \right) \qquad (1)$$

Then we set :

$$s_2 = s_3 - x_2$$

and replace in (1).

The conditions :

$$x_2 \leq x_1 \leq s_2 = s_3 - x_2$$

lead to :

$$x_2 \leq \frac{s_3}{2}$$

Next, we compute :

$$C_2 = \int_{x_2 = x_3}^{s_3/2} C_1(x_2) x_2^{n_2} dx_2$$

we replace :

$$s_3 = s_4 - x_3$$

and the condition

$$x_3 \leq x_2 \leq s_3 / 2 = (s_4 - x_3) / 2$$

leads to :

$$x_3 \leq \frac{s_4}{3}$$

We continue inductively. Each time, the integrand is a polynomial. The general step is :

$$C_k = \int_{x_k = x_{k+1}}^{s_{k+1}/k} C_{k-1}(x_k) x_k^{n_k} dx_k$$

and :

$$s_{k+1} = s_{k+2} - x_{k+1}$$

The final step is for $k = K$:

$$C_K = \int_{x_K=0}^{1/K} C_{K-1}(x_K) x_K^{n_K} dx_K$$

Step 2 : Computing the integral

$$J = \int_S x_1^{n_1} \cdots x_K^{n_K} dx_1 ... dx_K$$

where :

$$S = \left\{ (x_1, ..., x_K) ; x_1 \geq ... \geq x_K \geq 0, \sum_k x_k = 1 \right\}$$

The s_k are defined as before. We have :

$$J = \int_{V_2} (s_3 - x_2)^{n_1} x_2^{n_2} \cdots x_K^{n_K} \ dx_2 ... dx_K$$

with :

$$V_2 = \left\{ (x_2, ..., x_K) ; x_2 \geq ... \geq x_K \geq 0, \sum_{k \geq 2} x_k \leq 1 \right\}$$

So we proceed as in step 1. First, we compute :

$$D_2 = \int_{x_2 = x_3}^{s_3/2} (s_3 - x_2)^{n_1} x_2^{n_2} \ dx_2$$

and we replace :

$$s_3 = s_4 - x_3$$

Next, we compute :

$$D_3 = \int_{x_3 = x_4}^{s_4/3} D_2(x_3) x_3^{n_3} dx_3$$

and we continue inductively as before. The general step is :

$$D_k = \int_{x_k = x_{k+1}}^{s_{k+1}/k} D_{k-1}(x_k) x_k^{n_k} dx_k$$

and :

$$s_{k+1} = s_{k+2} - x_{k+1}$$

The final step is for $k = K$:

$$D_K = \int_{x_K=0}^{1/K} D_{K-1}(x_K) x_K^{n_K} \, dx_K$$

In practice, evaluation of these integrals is quite difficult, because the numerical values are extremely small (the order of magnitude is 10^{-160}).

We proceed inductively, using Maple and its symbolic computation capabilities, which prove to be of remarkable help in this situation. Each step is a polynomial, and each integration is performed in an exact manner (with no truncation or rounding off, except at the final stage). What we keep at each stage is a fraction of the type $\frac{p}{q}$, where p, q are extremely large integers. Moreover, in order to limit the size of q, we multiply at each step by a constant factor 10^6. Here is the Maple program (we use C[k] instead of D[k] in the program) :

```
C[2]:=int((s[3]-x[2])^n[1]*x[2]^n[2],x[2]=x[3]..s[3]/2):
s[3]:=s[4]-x[3]:
C[2]:=expand(10^6*C[2]):
s[30]:=1:
x[29]:=0:
for k from 3 to 28 do
C[k]:=int(C[k-1]*x[k]^n[k],x[k]=x[k+1]..s[k+1]/k) :
s[k+1]:= s[k+2]-x[k+1]:
C[k]:=expand(10^6*C[k]):
od :
```

At the end, C[28] is obtained at the exact quotient of two very large integers :

```
(1337609176957210412554728074928312721362678169202352836773988779108585199978664017255887107666252597062730
01603618580497873778657580043149612712585990182198114943116072922816246609299153656219924086824307176803717
25481675840669186891835839988931014496059780516570434686030716965245374317961147493434947044443035715963894
26544083404941728867288019065048526645485656484855253119065231958453251785297082472977262420376180008423345
73001850202271753906268193814049198229839180518643606359599451158494964010577732215268566772788114016147881
91173675982287511590546680679689065920234548030803127830416864228591411380783349769852430427608540132431270
32687942136598111156114628762193576995323412821873036458501442524121092444621536046557312333020545068738949
22840748479831365551923776351475428239647159470568726820622834401555340455426426894575005703239356765414836
59742017723185575360085731362513)/(5510244194996558663119023105829897769342382461252032027825171462390597555
49722446805239987387002669149154119765354882823959982685716684446828597368958530246636450947024215294889895
03454281252657850873642923145155121113295500525434567528385812943989664074199191441202996268148586012820716
14180747816344216690387132263303871806573958733176941111534937350061353434353307433158214826366553704901574
65210749951738928880165714683129014545423638367100838987376285230867615905384225930130249927652149553891398
97260977777135483225652651882061303913787195377041167932296144128932460334370380674789389761102137402558312
53356968480079208032711379909334110424174290386475027181387714625453365196320461511490262804946782402903244
32355695349005607682694665502736709197336385321407075981948141832962535376248756406687869343687626723682033
0290786030663441355227853850044122994489342033920000000000000000000000000)
```

and is converted to a double-precision real number : 0.2427495279e-8

One checks easily by induction that each D_k is of degree $n_1 + \cdots + n_k + k - 1$ with respect to the variable x_{k+1}. The polynomial of highest degree is therefore D_{K-1}, which has degree 119 in our case.

Step 3 : Computing the expectations

The theory presented above gives an explicit value to each expectation, since we know the marginal law for each density. We find :

$$\overline{p_k} = E(p_k) = \frac{J_k}{J}$$

where :

$$J_k = \int_S x_1^{n_1} \cdots x_{k-1}^{n_{k-1}} x_k^{n_k+1} x_{k+1}^{n_{k+1}} \cdots x_K^{n_K} \, dx_1 ... dx_K$$

(the $k-th$ exponent has been increased by 1)

and J has been defined before :

$$J = \int_S x_1^{n_1} \cdots x_K^{n_K} \, dx_1 ... dx_K$$

with :

$$S = \left\{ (x_1, ..., x_K) \, ; x_1 \geq ... \geq x_K \geq 0, \sum_k x_k = 1 \right\}$$

The computation of the J_k is made using the same Maple program.

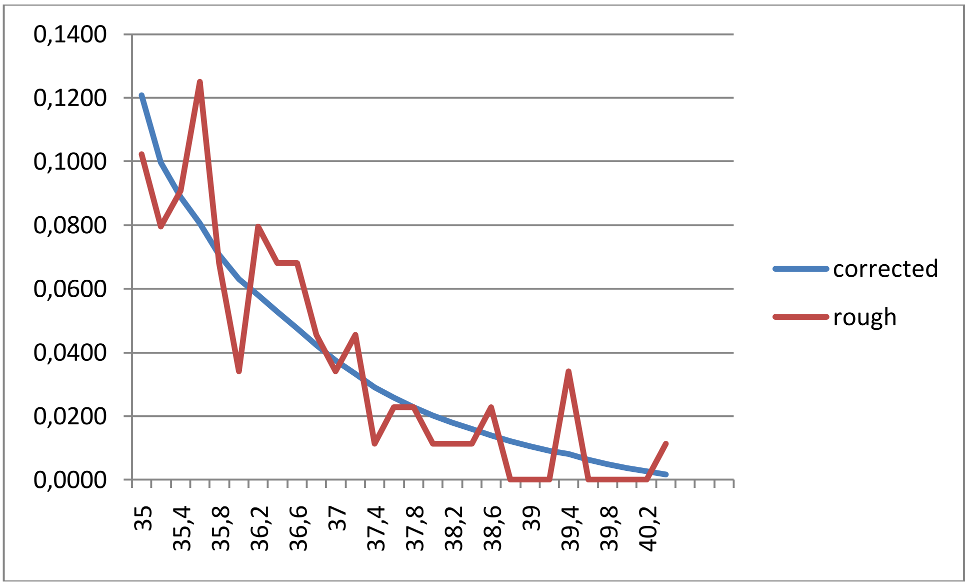
Here are the results :

temperature	35	35,2	35,4	35,6	35,8	36	36,2	36,4	36,6	36,8
proba	0,1209	0,0997	0,0889	0,0805	0,0707	0,0631	0,0580	0,0528	0,0476	0,0423

temperature	37	37,2	37,4	37,6	37,8	38	38,2	38,4	38,6	38,8
proba	0,0374	0,0332	0,0289	0,0257	0,0228	0,0202	0,0179	0,0159	0,0140	0,0121

temperature	39	39,2	39,4	39,6	39,8	40	40,2	40,4
proba	0,0105	0,0092	0,0080	0,0064	0,0050	0,0038	0,0027	0,0017

and here is the comparison with the original data :



Graph 3 : Original and corrected data

So, we have obtained (by purely probabilistic techniques) a "regularized" curve (in the present case, decreasing) of the probabilities for each event. From this theory, follows that, for instance, the probability of the event "temperature equals 40.4°C", which was seen only once in history, is 0.0017.

Here, we restricted ourselves to temperatures $\geq 35°C$, so the precise meaning of this statement is the following : among all days with temperature $\geq 35°C$, the probability to meet a day with temperature 40.4°C is 0.0017.

Since we had 88 days with temperature $\geq 35°C$ in 48,212 days (that is roughly 132 years), the number of days with temperature 40.4°C is on average $88\times 0.0017=0.1496$. So, per year, the probability to meet such an extremely hot day is $\lambda=\dfrac{0.1496}{132}=1.13\times 10^{-3}$.

The probability not to meet it in 1000 years is $(1-\lambda)^{1000}\approx 0.32$, so we have 68 % chances to meet this temperature again in the next millennium. However, $(1-\lambda)^{100}\approx 0.89$, so we have only 11 % chances to see it again during the next 100 years. Our conclusion is here that the record temperature of 40.4°C is of "millennium" type, not of century type (and this record is already more than 60 years old).

But, in all this construction, we made the assumption that the temperatures $\geq 35°C$ were all in the range $35-40.4$; of course, this is not completely correct : even if the temperature 40.4°C has never been surpassed, it may be. So, in the next paragraph, we will see how to compute the probabilities of higher values.

V. The prediction of new records

In this paragraph, we will consider 4 higher temperatures, namely 40.6°C, 40.8°C, 41°C, 41.2°C and compute their probability.

The method is the same as above, but this time $K = 32$ and the last four n_j's are 0, since such temperatures have never been recorded before. Each probability is the expectation of the corresponding multiple integral :

$$J_k = \int_S x_1^{n_1} \cdots x_{k-1}^{n_{k-1}} x_k^{n_k+1} x_{k+1}^{n_{k+1}} \cdots x_K^{n_K}\, dx_1 ... dx_K$$

with $k = 1,...,32$.

The program for computation is the same as before (same Maple program), but it takes longer. Evaluation of each J_k requires 30 explicit integrations, and takes about 20 minutes (each time with exact symbolic computation).

Here are the results :

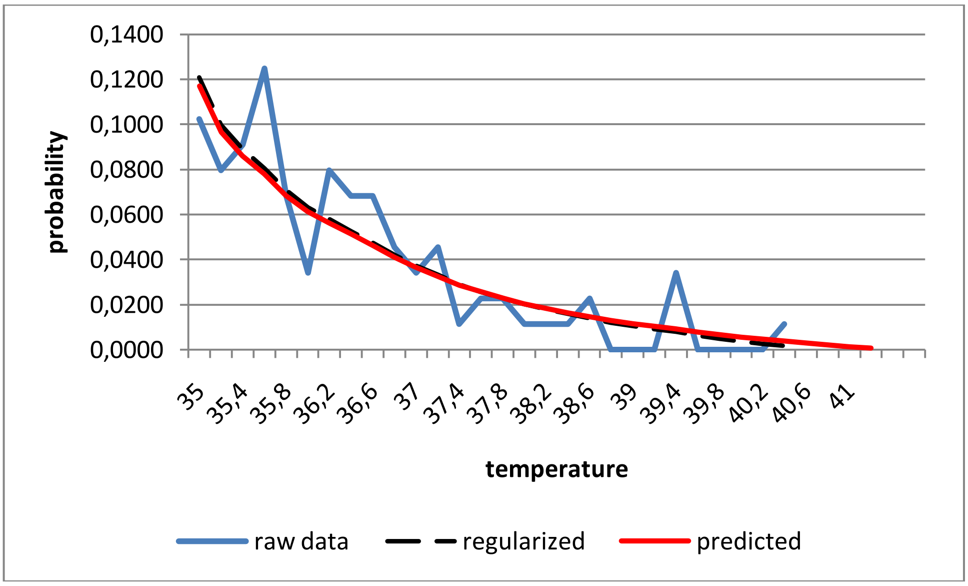
temp	raw data	regularized	predicted
35	0,1023	0,1209	0,1170
35,2	0,0795	0,0997	0,0965
35,4	0,0909	0,0889	0,0861
35,6	0,1250	0,0805	0,0780
35,8	0,0682	0,0707	0,0686
36	0,0341	0,0631	0,0612
36,2	0,0795	0,0580	0,0563
36,4	0,0682	0,0528	0,0513
36,6	0,0682	0,0476	0,0464
36,8	0,0455	0,0423	0,0413
37	0,0341	0,0374	0,0367
37,2	0,0455	0,0332	0,0327
37,4	0,0114	0,0289	0,0287
37,6	0,0227	0,0257	0,0257
37,8	0,0227	0,0228	0,0230
38	0,0114	0,0202	0,0205
38,2	0,0114	0,0179	0,0184
38,4	0,0114	0,0159	0,0165
38,6	0,0227	0,0140	0,0148
38,8	0,0000	0,0121	0,0131
39	0,0000	0,0105	0,0116
39,2	0,0000	0,0092	0,0104
39,4	0,0341	0,0080	0,0093
39,6	0,0000	0,0064	0,0079
39,8	0,0000	0,0050	0,0067
40	0,0000	0,0038	0,0057
40,2	0,0000	0,0027	0,0048
40,4	0,0114	0,0017	0,0039

40,6	0,0030
40,8	0,0021
41	0,0014
41,2	0,0007

Table 4 : Raw data, regularized and predicted probabilities

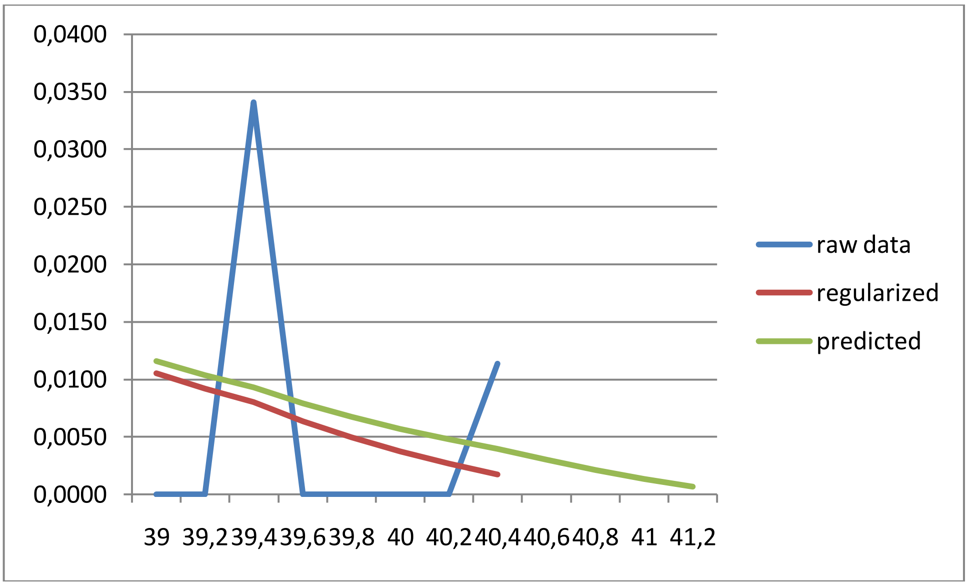
In this table, "raw data" are just the recordings : number of days, divided by 88. "Regularized" is the set obtained in the previous paragraph, and "predicted" are the data we just obtained. Of course, for each column, the sum is 1.

The graph below represents all data :



Graph 5 : Raw, regularized and predicted data

We observe that, for high temperatures, there are significant differences between regularized and predicted data. For instance, temperature 40.4°C had probability 0.0017 ; it has now probability 0.0039 : more than double ! This is due to the fact that, in the previous case, it was considered as the "high end" of the scale (thus very rare), whereas now it is only one of the highest, thus not so rare. We see this on the "zoom" below :



Graph 6 : A zoom on the region above 39°C

Let us now compute the probabilities of all these rare events, in this setting. The same reasoning as before provides the following results :

	proba among hot days	proba per year	proba not in 100 years	proba not in 1000 years	duration for 95 % proba (years)	duration for 50 % proba (years)
40,4	0,004	0,003	0,769	0,072	1 140	264
40,6	0,003	0,002	0,819	0,135	1 498	347
40,8	0,002	0,001	0,867	0,239	2 093	484
41	0,001	0,001	0,913	0,401	3 279	759
41,2	0,001	0,000	0,957	0,645	6 831	1 580

Table 7 : Probabilities for extreme temperatures

So we see that, for the temperature 40.4°C, we can reasonably expect (proba 0.5) it to come back after 260 years, and we can be quite sure in 1 100 years. For the higher temperature 41.2°C (which has never been observed), we can reasonably expect it in 1,580 years and be quite sure in 6,800 years.

VI. Computation of variance

The same methods allow to compute the variance of each p_k. We need the three integrals :

$$J = \int_S x_1^{n_1} \cdots x_K^{n_K} \, dx_1 ... dx_K$$

$$J_k = \int_S x_1^{n_1} \cdots x_{k-1}^{n_{k-1}} x_k^{n_k+1} x_{k+1}^{n_{k+1}} \cdots x_K^{n_K} \, dx_1 ... dx_K$$

$$V_k = \int_S x_1^{n_1} \cdots x_{k-1}^{n_{k-1}} x_k^{n_k+1} x_{k+1}^{n_{k+2}} \cdots x_K^{n_K} \, dx_1 ... dx_K$$

and then :

$$\operatorname{var}(p_k) = \frac{V_k}{J} - \left(\frac{J_k}{J} \right)^2$$

For instance, for $k = 32$, temperature 41.2°C, the estimated value for the probability, computed above, was $p_{41.2} = 6.58 \times 10^{-4}$ and $\sigma = 6.30 \times 10^{-4}$.

VII. A complete probability law for each p_k

In the computations presented above, we gave the expectation of each p_k and we explained how to compute its variance. But the question is : can we give the whole probability law for each p_k ?

The answer is yes in theory, no in practice, except for the last one, as we will see.

In theory, the repartition function for each p_k is given by the following formula :

$$F_k(z) = \frac{J_k(z)}{J}$$

with, as before :

$$J = \int_S x_1^{n_1} \cdots x_K^{n_K} \, dx_1 ... dx_K$$

$$S = \left\{ (x_1, ..., x_K) \, ; x_1 \geq ... \geq x_K \geq 0, \sum_k x_k = 1 \right\}$$

and :

$$J_k(z) = \int_{S_k(z)} x_1^{n_1} \cdots x_K^{n_K} \, dx_1 ... dx_K$$

$$S_k(z) = \left\{ (x_1, ..., x_K) \, ; x_1 \geq ... \geq x_K \geq 0, \sum_k x_k = 1 \, ; x_k \leq z \right\}$$

In practice, the integration over $S_k(z)$ is quite complicated : one has to distinguish various cases, depending on the position of the previous bounds with respect to z.

The only case which is simple is that of the last k ; in our case, $k=32$. We compute directly the repartition function, integrating from 0 to z in the last integral (in the previous computation, it was between 0 and $\frac{1}{32}$). The result is a polynomial of degree 119, with extremely large coefficients. Here is the leading term :

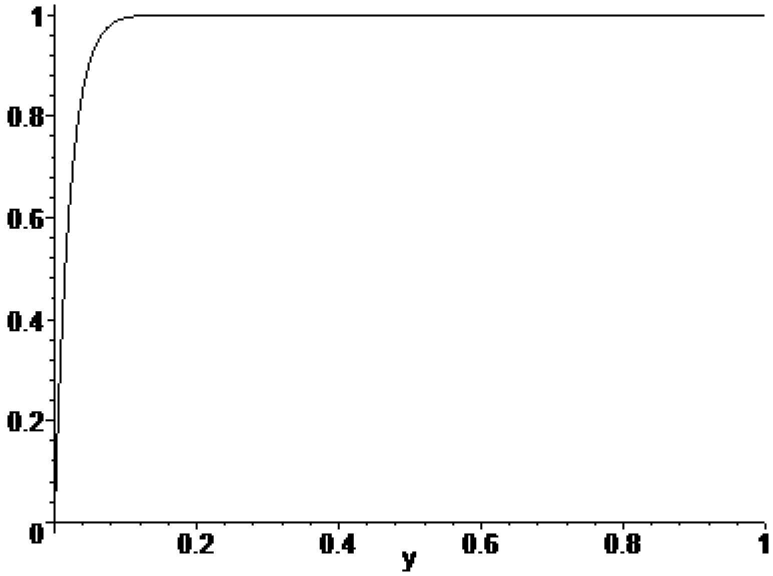
$$-0.105022562293228555795032874957 \times 10^{168} \times z^{119}$$

The value of z is between 0 and $\frac{1}{32}$, so the monomials z^j are extremely small, which poses considerable difficulties, in terms of numerical computation. As before, all computations are made with exact (rational) values, converted to decimal at the end only. One checks that with the value $z=1/32$ one has $F_{32}(z)=1$.

Letting $z=\frac{y}{32}$, we convert our repartition function $F_{32}(z)$ into a repartition function $G(y)=F_{32}\left(\frac{z}{32}\right)$, defined on the interval $[0,1]$, with more reasonable coefficients. The low degree terms are :

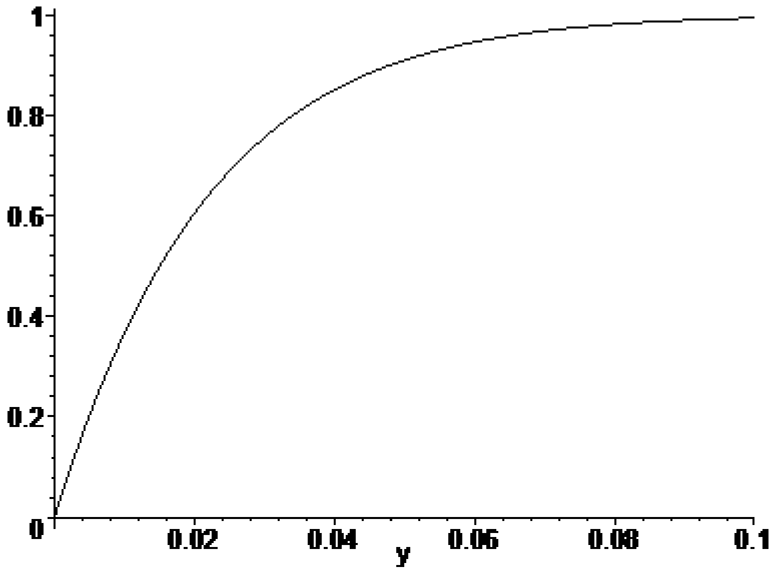
$$45.15-1910y+36915y^2-410566y^3+2467674y^4+....$$

Here is the graph of this repartition function :



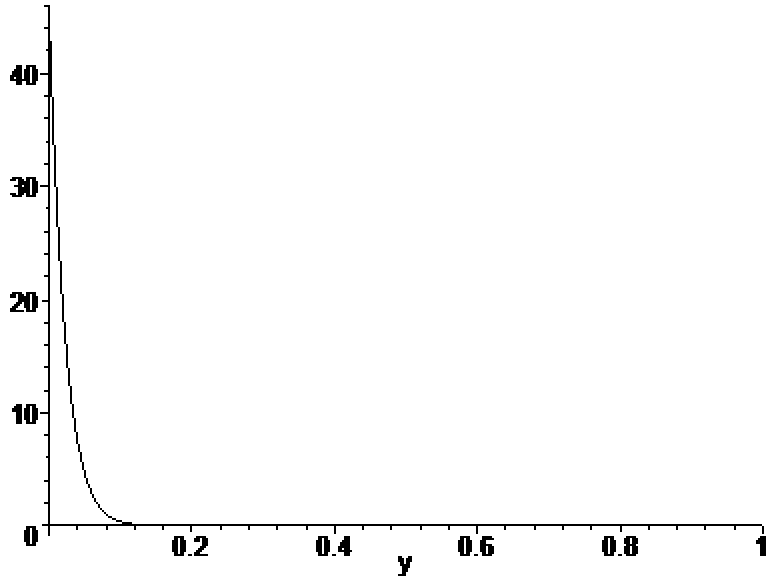
Graph 8 : The repartition function of the 32-th probability

and here is a zoom on the part between 0 and 0.1:



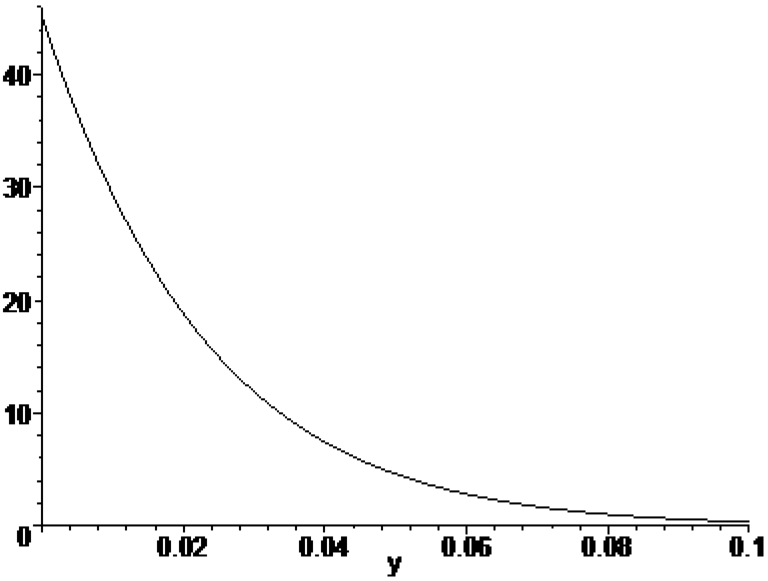
Graph 9 : Zoom on the interval 0 -0.1 of the repartition function of the 32-th probability

Here is the graph of the associated density :



Graph 10 : The density of the 32-th probability

and a zoom on the interval 0 - 0.1:



Graph 11 : A zoom on the interval 0 - 0.1 of the density of the 32-th probability

We observe that, quite naturally, this density is decreasing. Of course, it is defined on the interval 0 - 1 (a probability is always in this interval), but the low values are more likely than the high ones.

VIII. References

[1]. Bernard Beauzamy : Méthodes Probabilistes pour l'étude des phénomènes réels (in French). Ouvrage édité et commercialisé par la *Société de Calcul Mathématique SA*, ISBN 2-9521458-0-6. Mars 2004.

[2]. Bernard Beauzamy : The information associated with a sample. Published in the "Robust Mathematical Modeling" program, May 2009.
http://www.scmsa.com/RMM/BB_information_sample_2009_05.pdf

Annex : a simple case

Let us see the application of the method in a very simple case : only two situations are possible.

The first one was met $n_1 = 1$ time, and the second one $n_2 = 2$ times. We are looking for the law of p_1 (or the law of $p_2 = 1 - p_1$). As previously, we assume $p_1 \geq p_2$: the first situation is more likely than the second one.

This can be met, for instance, if we study the life expectation of a product : the situation 1 is the case where the product lasts ten years at most, and the situation two is the case where the product lasts more than ten years.

The law of p_2 is proportional to the function :

$$f(x) = c1_S(x)(1-x)x^2$$

where c is a normalization constant, and S is the set :

$$S = \{x\,; 0 \leq x \leq 1-x\}$$

that is :

$$S = \left\{x\,; 0 \leq x \leq \frac{1}{2}\right\}$$

Let us first compute the coefficient c. We have :

$$\int_0^{1/2} (1-x)x^2 dx = \frac{5}{192},$$

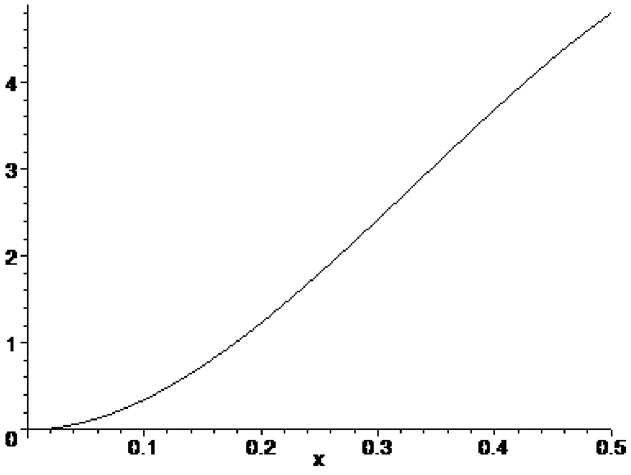
so :

$$c = \frac{192}{5}.$$

The density of p_2 is the function :

$$f_2(x_2) = \frac{192}{5}\left(1-x_2\right)x_2^2,\ 0 \leq x_2 \leq \frac{1}{2}, \text{ 0 otherwise.}$$

This is the graph of this function :

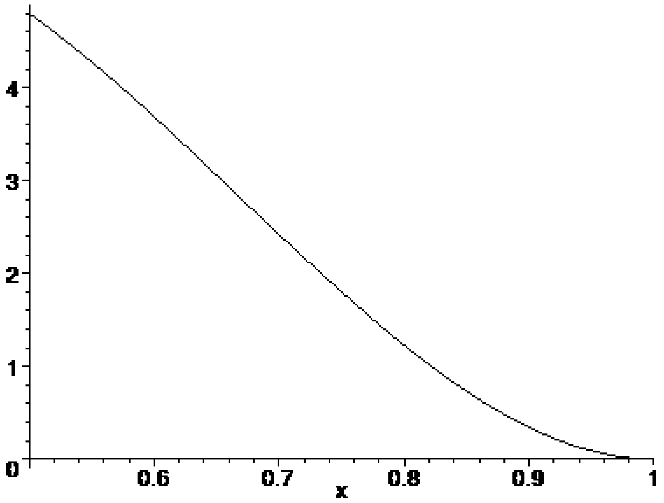


Graph of the density of p_2

and the density of p_1 is the function :

$$f_1(x_1) = \frac{192}{5} x_1 (1 - x_1)^2 \text{ , } \frac{1}{2} \le x_1 \le 1, \text{ 0 otherwise.}$$

Here is the graph of this function :



Graph of the density of p_1