



The use of the EPH in order
to extend a truncated probability law

Bernard Beuzamy
Société de Calcul Mathématique SA

July 2009

I. Truncated probability law

It happens quite often in practice that a probability law is "truncated", meaning that the high values (or the low ones) are not correctly recorded. Here are some examples :

- The age of a population is recorded, in 10 years-width intervals, but the upper interval is simply ≥ 100 ;
- The speed of the wind is recorded, in 10 km/h intervals, but the upper interval is simply ≥ 150 km/h ;
- A temperature is recorded, but the thermometer stops working at 50°C : everything above 50 is recorded as 50, or not recorded at all.

In practice, such situations are not exceptional at all. Of course, if nothing was recorded, we cannot know that something happened, but if, for instance, the instrument was broken (as it can be the case of an anemometer in strong wind), then some information remains.

II. Mathematical description of the problem

For us, the mathematical description is as follows :

We have an histogram with intervals $[x_1, x_2[, \dots, [x_k, x_{k+1}[, \dots, [x_{k_0-1}, x_{k_0}[, [x_{k_0}, +\infty[$: the last interval is different from the others, and is of the type $x \geq x_{k_0}$. All previous intervals have same width. We have the probabilities $p_1, \dots, p_k, \dots, p_{k_0}$ of all these intervals. We want to complete the probability law above x_{k_0} : we want to introduce intervals $[x_{k_0}, x_{k_0+1}[, [x_{k_0+1}, x_{k_0+2}[, \dots$ and estimate their probabilities. In other words, we want to "distribute" the infinite interval $[x_{k_0}, +\infty[$ in a reunion of finite intervals of same width and estimate their probabilities.

There is no way of doing this, from usual tools in probability theory. The usual "tricks" would be :

- To fit some specific law using the existing data (for instance normal law, or exponential, or Weibull, or Gumbel, and so on), and use this law for the extension. The problem is that, in usual situations, the data do not fit with any academic law.
- To extend the most right data, using for instance a linear regression. This is, in some sense, a special case of the above, using only some of the data. The drawbacks are : there is no reason to admit a linear model, and this linear model will depend on the data which are selected ; if you keep only the last 20 data, or last 40 data, and so on, you will not have the same model.

The Experimental Probabilistic Hypersurface (EPH) will allow a construction with no such fictitious assumptions. We refer to [EPH1] for the construction of the EPH and its main properties.

Since the intervals $[x_1, x_2[, \dots, [x_{k_0-1}, x_{k_0}[,$ are all of same width (let us call it w), the points $x_{k_0+1}, x_{k_0+2}, \dots$ are known. The only question is : what probability should we put on each of these new intervals ?

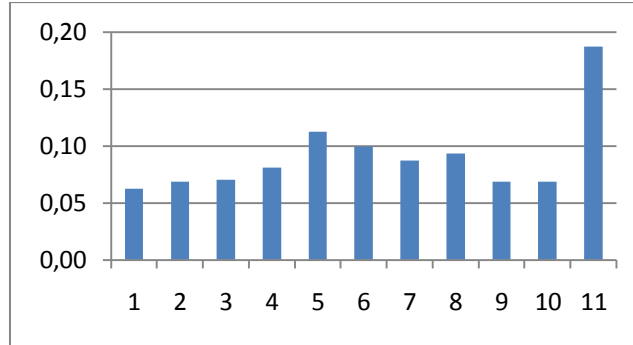
The EPH is a way of sending a probabilistic information to new places, that is to places where nothing is known. Here, the context is simple ; we have the information "for the interval $[x_1, x_2[$ the probability is p_1 ", up to "for the interval $[x_{k_0-1}, x_{k_0}[,$ the probability is p_{k_0-1} ", and we want to know what probability to put above $[x_{k_0}, x_{k_0+1}[, [x_{k_0+1}, x_{k_0+2}[, \dots$

Let us call c_k the center of the interval $I_k = [x_k, x_{k+1}[,$ $k \geq 1$. The problem may be viewed as follows : we have a measurement (namely p_k) above each point c_k , $k < k_0$, and we wish to estimate the value of the p_k , $k \geq k_0$.

Let us fix some terminology. Let us denote by $p_{k_0}^0$ the initial probability of the half-infinite interval $[x_{k_0}, +\infty[$ and let p_k , $k \geq k_0$, the probabilities of the intervals we want to introduce instead. Then the original probability $p_{k_0}^0$ should be redistributed among these new intervals, which means :

$$p_{k_0}^0 = \sum_{k \geq k_0} p_k \quad (1)$$

Typically, we have the following disposition :



where the 11th interval stands for "all above x_{11} " (the data are given below).

We observe that we cannot apply directly the EPH to the couples (c_k, p_k) , $k < k_0$, since (as the above example shows) the prediction would generally be something above p_{k_0-1} (the last known one). Indeed, at the end, the p_k 's are often decreasing (perhaps with "bumps") and the EPH takes the whole sequence into account. This is not at all what we want : the next p_k 's should be more or less smaller than the existing ones.

So we proceed as follows. We introduce the consecutive differences :

$$q_1 = p_1, q_k = p_k - p_{k-1} \text{ for } 2 \leq k < k_0 \quad (2)$$

and we will work on these differences : we will apply the EPH to them.

III. First step of the construction

We apply the EPH to the sequence $(c_k, q_k)_{k < k_0}$. From it, we deduce a probability law for the parameter q_{k_0} above the point c_{k_0} ; let $f_{k_0}(t)$ be its density. Then, since $p_{k_0} = q_{k_0} + p_{k_0-1}$, and since p_{k_0-1} is known, the probability law of p_{k_0} is known and its density is $f_{k_0}(t + p_{k_0-1})$. But some precautions have to be taken :

- First, the probability law on p_{k_0} has to be between 0 and 1, because p_{k_0} is itself a probability. This does not follow automatically from the construction.
- Second, the probability law on p_{k_0} should not exceed the known value $p_{k_0}^0$.

Both precautions may be combined into 1 : we take the probability law $f_{k_0}(t + p_{k_0-1})$, truncate it at 0 (lower) and $p_{k_0}^0$ (upper) and renormalize, so that the integral will be 1. Let $\varphi_{k_0}(t)$ be the density of probability we obtain this way for p_{k_0} .

Then the value we assign to p_{k_0} is the expectation of this probability law :

$$p_{k_0} = \int_0^{p_{k_0}^0} t \varphi_{k_0}(t) dt \quad (3)$$

By construction, we have :

$$0 \leq p_{k_0} \leq p_{k_0}^0 \quad (4)$$

So we have affected a "portion" of the original probability $p_{k_0}^0$ to the point c_{k_0} , and we still have the rest, namely $p_{k_0}^0 - p_{k_0}$, to affect to further intervals.

So, at the end of the first step, we have affected a probability p_{k_0} to the point c_{k_0} and we have a quantity $p_{k_0+1}^0 = p_{k_0}^0 - p_{k_0}$ for further intervals. We continue inductively.

IV. General step

Assume we know, or have constructed, probabilities p_1, \dots, p_k above the points c_1, \dots, c_k and we are left with a probability p_{k+1}^0 to redistribute for further intervals. We have by construction :

$$p_1 + \dots + p_k + p_{k+1}^0 = 1$$

We form the consecutive differences :

$$q_j = p_j - p_{j-1}, \quad j \leq k$$

Using the information (c_j, q_j) , $j \leq k$, from the EPH we construct a probability law for q_{k+1} above c_{k+1} . From this probability law on q_{k+1} , we deduce (by translation) a probability law on p_{k+1} ; we truncate it at 0 and at p_{k+1}^0 , and renormalize. Let φ_{k+1} be this probability law. We choose as p_{k+1} the expectation :

$$p_{k+1} = \int_0^{p_{k+1}^0} t \varphi_{k+1}(t) dt$$

and define the rest (to be reassigned to further intervals) as :

$$p_{k+2}^0 = p_{k+1}^0 - p_{k+1}$$

Of course, we have :

$$p_1 + \dots + p_k + p_{k+1} + p_{k+2}^0 = 1$$

V. An example

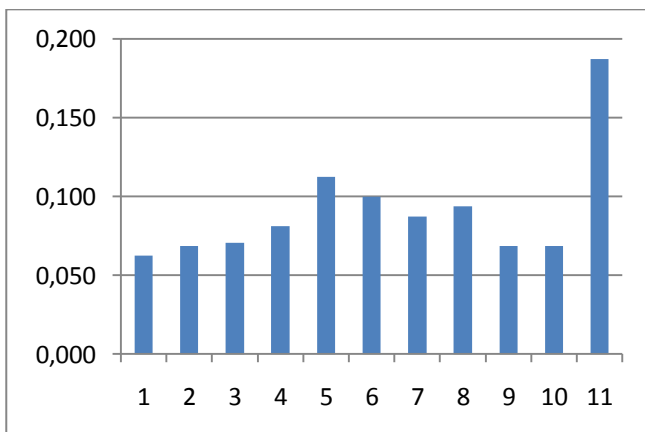
We start with the data given in the illustration above :

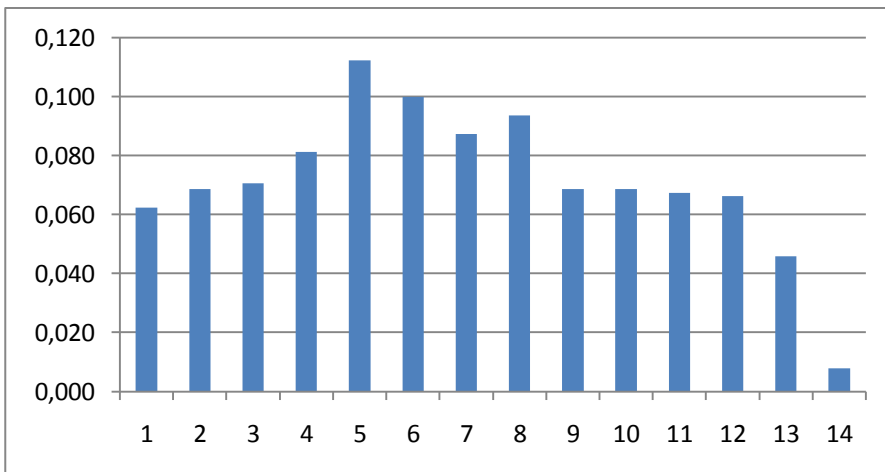
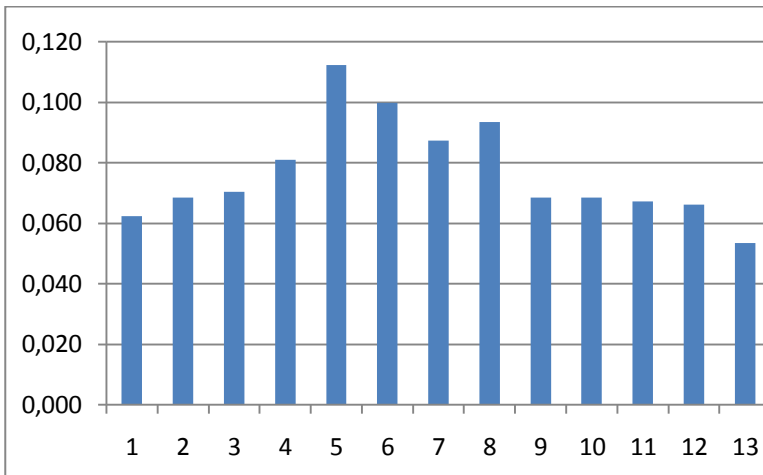
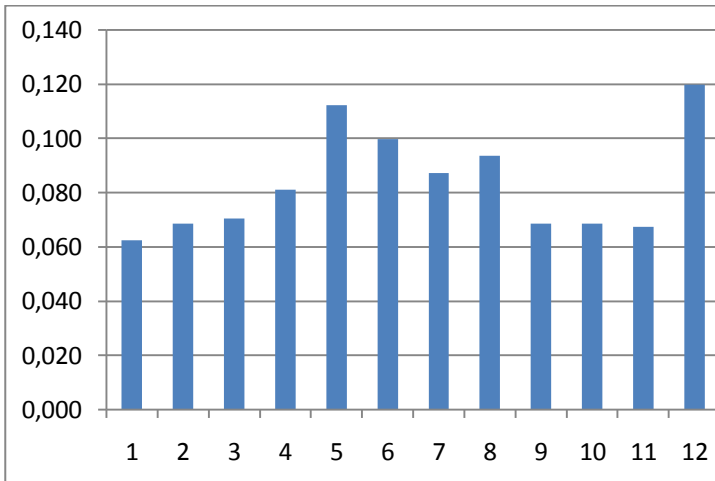
point	proba
1	0,062
2	0,069
3	0,070
4	0,081
5	0,112
6	0,100
7	0,087
8	0,094
9	0,069
10	0,069
>10	0,187

The repeated application of the EPH gives the following probabilities :

point	proba	point	proba	point	proba
1	0,062	1	0,062	1	0,062
2	0,069	2	0,069	2	0,069
3	0,070	3	0,070	3	0,070
4	0,081	4	0,081	4	0,081
5	0,112	5	0,112	5	0,112
6	0,100	6	0,100	6	0,100
7	0,087	7	0,087	7	0,087
8	0,094	8	0,094	8	0,094
9	0,069	9	0,069	9	0,069
10	0,069	10	0,069	10	0,069
11	0,067	11	0,067	11	0,067
12	0,120	12	0,066	12	0,066
		13	0,054	13	0,046
				14	0,008

This leads to the following histograms :





References

[EPH1] http://pagesperso-orange.fr/scmsa/RMM/Rapport1_SCM_IRSN_2008_06.pdf