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

Statistical studies for decision making

- How to normalize the presentation -

by Bernard Beauzamy
Chairman and CEO,
Société de Calcul Mathématique SA
Paris, France

August 2008

To take a decision, for instance connected with investments or with security, requires generally the use of some available data, which are treated by means of statistics. Typical examples are :

- Building a transportation net or an energy net (electricity, gas) in some area ;
- Analysing the effects of pollutions, radiations, and so on.

The decisions which are taken are quite often with heavy consequences and represent important costs. So, they should be sustained by serious studies. Two conditions are absolutely necessary, and they should be imposed by all decidors :

- The study must be verifiable ;
- The study must be reusable.

"Verifiable" means that, if all the elements are given to another expert, he must be able to say if the work that has been done is correct or not.

"Reusable" means that some parts of the study may be used again (for instance, pieces of computer code), for other studies, but also that, five years later, one must be able to compare what the study predicted with what actually happened, in order to see if the reasonings were correct or not.

These two conditions are just common sense. Unfortunately, most of the studies we have to asses are neither verifiable nor reusable. Nobody knows what data were used, neither what reasonings were made. We are left with graphs, maps : here it is red, here it is green ; here it goes up, here it goes down ; all of them obtained using some software. The author, fifteen days after is masterpiece, does not even know what key he has pressed, so we should not expect his science five years later !

Technical recommandations for the presentation of a study

In the vast majority of the studies we have to assess, the authors use statistical tools in an artificial manner ; for instance, they assume that some phenomenon follows a Gauss law, and they try to compute the parameters in cases A and B ; they want to show that these parameters are different between both cases. But this proves absolutely nothing, since the phenomenon has no reason to follow actually a Gauss law.

The same way, they often assume independence between factors, and these assumptions are not satisfied in practice.

We give below some simple rules in order to prepare and present a statistical study. The basic rule (this is obvious !) is that one should never introduce any artificial assumption : one should treat the data as they are !

We will take the case, frequent in practice, where Excel is used ; in most situations, it is quite appropriate.

1. Raw data

The first sheet (sheets(1)) must contain the raw data upon which the study is built. The presence of raw data is an absolute necessity in order that the work can be checked ; it is also necessary if we want to compare what has changed five years later.

2. Working data

On the second sheet (sheets(2)), one will put the working data. They can be raw data, but not necessarily. For instance, sometimes, one needs to normalize the data (put them between 0 and 1), or to consider increments (differences between one year and the next), and so on.

All these transformations are acceptable, provided they are acknowledged as such. We are not in presence of raw data, but of data which have been subject to a first treatment. This treatment has to be justified. One should not confuse working data and raw data.

3. Processing the study

A statistical study often has for ambition to establish some link between some quantity, the "objective", and some "explicative" parameters. One wants to know if the explicative parameters have an influence upon the objective, or not. For instance :

- One wants to analyse the objective "number of cancers in a region", through the explicative parameters "existence of radiations", "number of smokers", and so on.

- One wants to analyse the objective "size of gas stocks", through the explicative parameters "size of imports", "size of consumption", and so on.

The first work to be done is to prepare the histogram of the objective quantity, that is its probability law. This is easily done, using Excel, the following way :

- One divides the set of possible values into intervals of same width ;

- One counts, in the sample, how many values fall into each interval.

For instance, if the intervals have width 10, and if the data are in the first column, first sheet, we will have a VBA code looking like this :

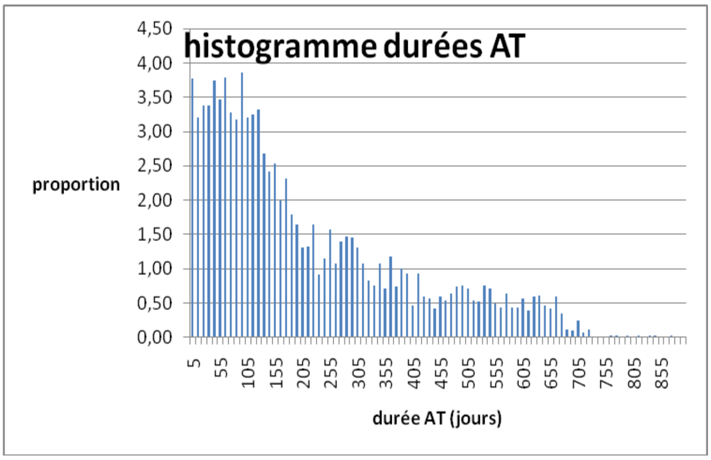
```
for k=1 to nbfinal
j=int( sheets(1).cells(k,1)/10)
sheets(3).cells(j,2)=sheets(3).cells(j,2)+1
next k
```

and the second column of sheet 3 will contain this histogram.

The first column contains the intervals which are used for the construction ; in our example :

sheets(3).cells(k,1)= 10(k-1)

Here is an example :



Graph 1 : histogram

This histogram deals with interruptions of work, for workers, in some companies. The x axis contains the length of the interruption (in days), and the y axis the percentage of workers with this duration of interruption. The size of the interval is 10 days (we count the proportion between 0 and 10 days, between 10 and 20, and so on).

The comparison between two histograms is never easy, because the important information is the area below the function. So, we will prefer working with partial sums, that is with the repartition function. For practical reasons, people often prefer to represent, for a given threshold, the proportion above this threshold (and not below); so we work with $G = 1 - F$ rather than with F, the repartition function.

In the column 3, in cell k, we will put the sum of all columns from column 2, from k till the end :

```
for k=1 to nbfinal
for j=k to nbfinal
sum=sum+sheets(3).cells(j,2)
next j
sheets(3).cells(k,3)=sum
```

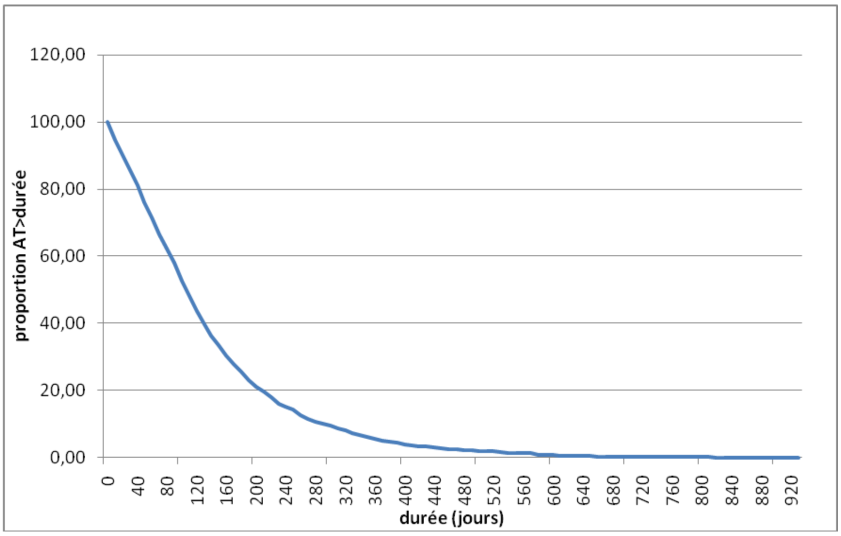
```
sum=0
next k
```

Of course, the first cell of column 3 contains the total number of the sample, so we have an easy way to check that we forgot no data.

And finally, in column 4, we put the cumulated percentages : sum of percentages after k :

```
for k = 2 to nbfinal
sheets(3).cells(4,k)=sheets(3).cells(k,3)/sheets(3).cells(2,3)*100
next k
```

So, this construction is quite elementary, as everyone sees. What we get is a function, starting at 100 and finishing at 0 ; it looks like this :

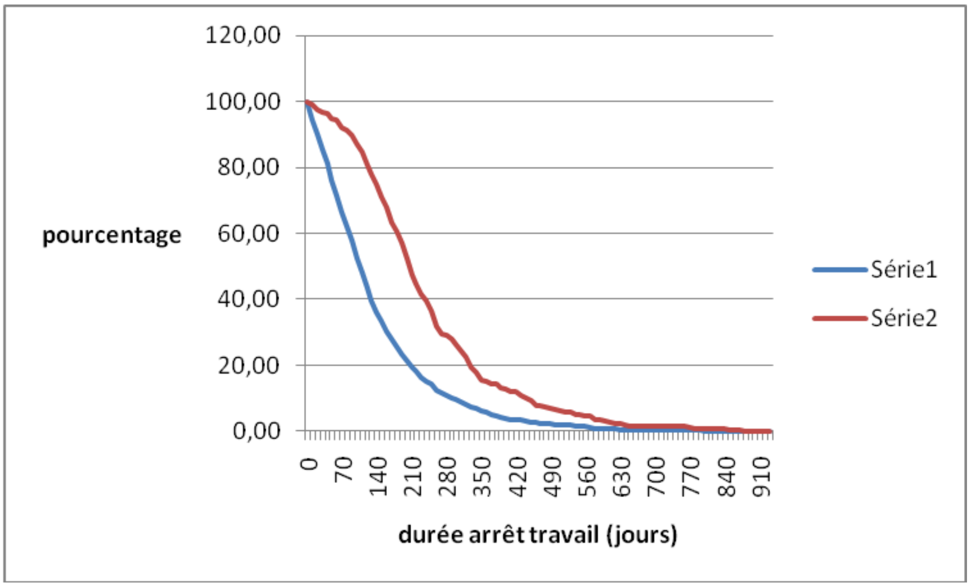


Graph 2 : The function G

Above each x, in our example, we indicate the proportion of workers, the interruption of which is $\geq x$. In mathematical notation, we draw the graph of the function :

$$G(x) = P\left\{X \geq x\right\}$$

The fact that this function always starts at 100 and finishes at 0, as explained earlier, makes comparisons easy. Let us see an example :



Graph 3 : example of comparison

Here, we have two types of workers : the first one in blue, the second one in red. We see that the function G for the blue ones is always below the function G for the red ones.

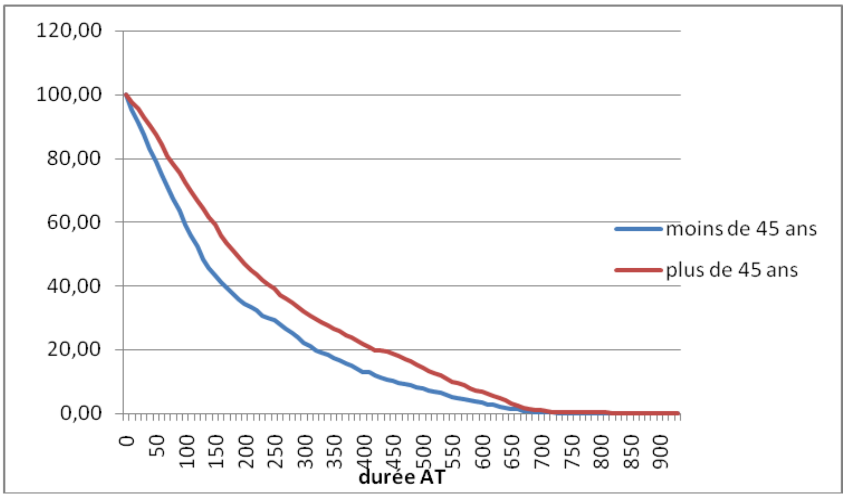
This indicates that for any duration of the interruption, for instance 200 days, the proportion of red workers exceeding this duration is larger than the proportion for the blue workers. In other words, the red workers have longer interruptions than the blue ones. The comparison is straightforward and deals only with factual data.

4. Taking into account the explicative parameters

In order to take into account the explicative parameters, one builds conditional probability laws, the following way. Let us take the first explicative parameter, say that it is the age of the worker, and assume that half of the data concerns ages below 45 and half above. We perform again the previous analysis, when the age is below 45 :

```
for k=1 to nbfinal
for j=k to nbfinal
if param1 <= 45 then
sum=sum+sheets(3).cells(j,2)
end if
next j
sheets(3).cells(k,3)=sum
sum=0
next k
```

and we get a function as previoulsy. We do the same when the age is above 45 ; we can put both on the same graph. If, in a systematic manner, we see that function 1 is below function 2, we deduce that, for every threshold, the probability to be above this threshold is smaller in the case of young workers than in the case of old workers. This shows that old workers have longer interruptions of work.



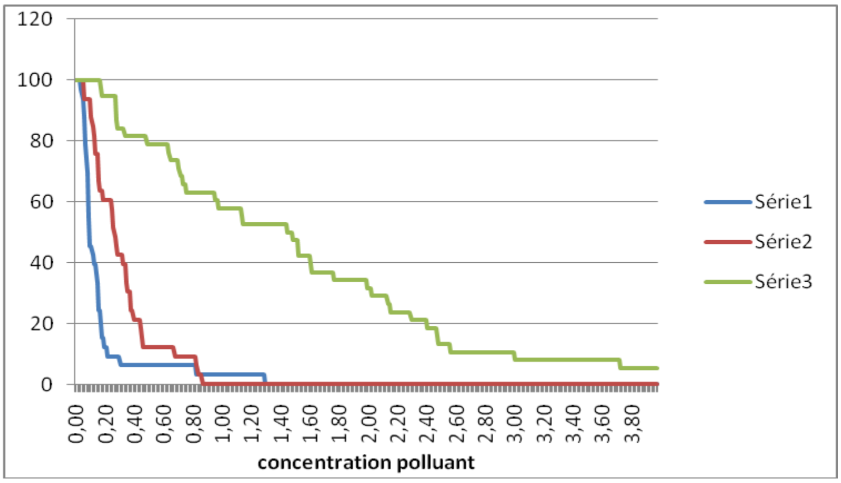
Graph 4 : distinction according to age

In this example, the red curve (old workers) is systematically above the blue one (young workers) ; for a fixed duration of interruption, there will be more old workers than young workers, the interruption of which exceeds this duration : the interruptions are longer for old workers than for young ones.

This way to present the information, by using conditional probabilities, is very satisfactory : we do not use any artificial assumption and do not use any "black box". No statistical test is performed, with disputable validity.

If the objective is a threshold defined in terms of cleanness (for instance, the concentration of some pollutants), we obtain a simple and unquestionable manner to compare different situations.

Here is an example : we deal with the concentration of some pollutant, measured by different stations. The first series, in blue, concerns zones where the density of population is below 80 habitants per square km. The second series, in red, concerns zones where this density is between 80 and 170 ; the last series, in green), zones where the density is above 170.



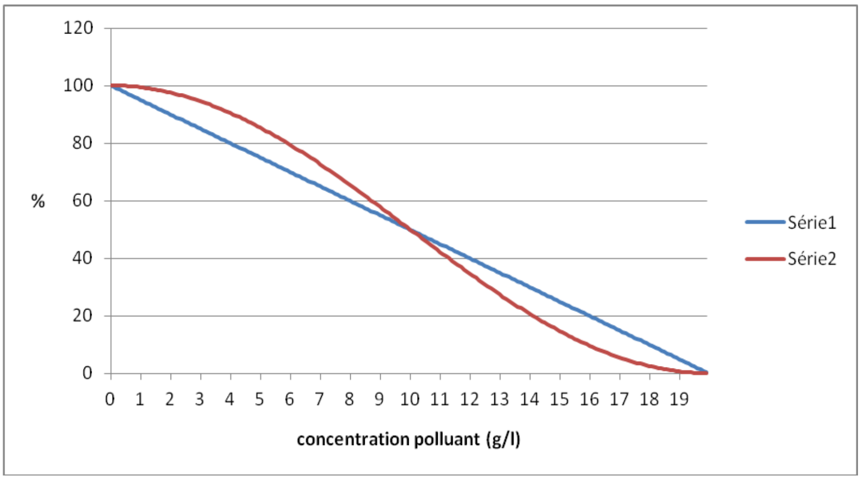
Graph 5 : concentration of polluants, as a function of population density

We see clearly that the function associated to high density is above the two others ; the proportion of stations, in zones with high density, that have a pollution above a given threshold is higher in highly populated zones than in low or medium density zones.

For instance, 40 % of the stations "high density" have a pollution of at least 1.6 g/l, whereas none has such a pollution in the low or medium density zones. So we clearly see that, for this pollutant, high density zones have more pollution.

5. Exploitation of a general example

Let us now see how to interpret a more general example : one of the curves is not always above the other ; they intersect somewhere.



Graph 6 : General situation

In the example above (totally fictitious), both curves intersect at the concentration 10 g/l. The common value for both functions is 50 % at this point. So we will have 50 % of the blue stations and 50 % of red stations above 10 g/l.

If we take a higher concentration, say 15 g/l, we have 25 % of the blue and 14.6 % of the red. We deduce that the blue are less subject to high pollutions.

If we take a lower concentration, say 5 g/l, we have 75 % of the blue and 85.3 % of the red above this value. We deduce that the red are more subject to low pollutions.

Taking differences, in the interval 5 - 10, we have 25 % of the blue and 35.3 % of the red. Here again, the conclusion is very simple : there are more blue than red in the interval 10 - 15, more red than blue in the interval 5 - 10 (in proportion, not necessarily in absolute numbers).

We insist that all these numbers are proportions in each class, not global proportions. In order to make this clear, assume that there are 1000 blue stations and 5000 red stations. Then, in the interval 5 - 10, the number of blue stations is : 25 % of 1000, that is 250, and 35.3 % of 5000, that is 1765. To use the global number of 6000 stations would be a mistake, since the blue stations are 5 times less frequent.

In all cases, the relative position of the functions, on any interval, tells us about the relative importance of the phenomenon.