



## A probabilistic approach for censored data

Bernard Beuzamy  
SCM SA

June 10, 2009

### I. Description of the problem

In medical treatments, the following situation is often met : some people go away, and one does not know whether, for them, the treatment succeeded or not. Another situation is that of industrial tests : some material is subject to a test, but the test is interrupted after some time, and the material is still in good condition ; all we know is that the material lasted as long as the test.

In both cases, we can say that the life expectancy (of the person or of the material) is at least some value, but we do not know what the exact value is. For instance, we can say that the material resisted during 30 days, but we do not have any precise information about the date where it might collapse. This situation where we know only a lower bound for the life duration is called "censored data".

In order to make our vocabulary precise, we will assume that we speak about the life duration of a product, under some test. Let  $X_n$ ,  $n=1, \dots, N$ , be a random variable, which indicates the life duration of the  $n$ -th product tested. The law of  $X_n$  is unknown, but these variables are assumed to be independent (one test does not influence another) and with same law. We can also assume that all variables take integer values (number of years, of days, of seconds,...) ; this is of no importance here.

We realize a sample, that is we perform some tests. Among these tests,  $N$  give precise values, denoted by  $x_1, \dots, x_N$ , and  $M$  give censored values, which we denote by  $c_1, \dots, c_m$ . This means, in this case, that the true value (which is unknown) satisfies  $X \geq c$ .

The question is : how to build the probability law from the sample ? Usually, people restrict themselves, considering that the true value is exactly  $c$ . The result of such a choice is, of course, that higher probabilities for life duration are seldom met.

## II. Construction of a probability law

We present here a method, based upon conditional probabilities, which allows the construction of a probability law for the variable  $X$ , taking censored data into consideration.

### 1. The basic probability law

From the precise values  $x_1, \dots, x_N$ , we build a probability law, the usual way (an histogram) : we divide the interval of values into small intervals (called classes), and count how many times the sample falls in each class. Let  $y_1, \dots, y_K$  be these classes and  $p_1, \dots, p_K$  be their respective probabilities. This probability law does not use the censored data at all. It could be called : empirical probability law from the exact data.

### 2. Adding one censored data

Let now  $c_1$  be the first censored data. It falls into a class  $y_{k_1}$ . We will modify the probabilities  $p_1, \dots, p_K$  the following way : the probabilities  $p_k$ ,  $k < k_1$ , will decrease, and the probabilities  $p_k$ ,  $k \geq k_1$ , will increase. This follows a natural intuition.

How do the probabilities  $p_k$ ,  $k \geq k_1$ , increase ? There is no reason to increase any of them, rather than another one ; the idea to increase only the first one corresponds to a penalty (as we already said). Instead, we take the following idea : the conditional law of  $X \geq c_1$  should not be modified. In other words, the new probabilities should give the same conditional law knowing  $X \geq c_1$ .

Let  $s(k_1) = \sum_{k=1}^{k_1-1} p_k$ ,  $t(k_1) = \sum_{k=k_1}^K p_k$  : these are, respectively  $P\{X < y_{k_1}\}$  and  $P\{X \geq y_{k_1}\}$ .

Among the  $N_1$  exact measurements, the number of points in the  $k$ th class is approximately  $p_k N$ . The conditional law of  $X$ , knowing  $X \geq c_1$ , is :

$$P\{X = y_k | X \geq c_1\} = \frac{p_k}{t(k_1)}$$

We have one more trial, which is censored. We now have  $N_1 + 1$  points, and a new probability law, which we call  $P' = (p'_1, \dots, p'_K)$ . For  $k \geq k_1$ , the conditional law should be the same as before. This means that :

$$p'_k = \lambda p_k, \quad k \geq k_1, \quad \text{for some } \lambda > 0$$

The same way, the conditional law for  $k < k_1$  should be the same as before, which implies :

$$p'_k = \mu p_k, \quad k < k_1, \quad \text{for some } \mu > 0$$

Also, from this new measurement, the probabilities  $s(k_1) = P\{X < y_{k_1}\}$  and  $t(k_1) = P\{X \geq y_{k_1}\}$  are modified. They become respectively :

$$s'(k_1) = \frac{N \cdot s(k_1)}{N+1}, \quad t'(k_1) = \frac{N \cdot t(k_1) + 1}{N+1}$$

From this, we can compute  $\lambda$  and  $\mu$ . Indeed,

$$\sum_{k < k_1} p'_k = \mu \sum_{k < k_1} p_k = \mu s(k_1) = \frac{N s(k_1)}{N+1}$$

which gives :

$$\mu = \frac{N}{N+1}$$

and similarly :

$$\sum_{k \geq k_1} p'_k = \lambda \sum_{k \geq k_1} p_k = \lambda t(k_1) = \frac{N t(k_1) + 1}{N+1}$$

which gives :

$$\lambda = \frac{N t(k_1) + 1}{(N+1)t(k_1)}$$

So we have obtained :

**Proposition 1.** – *The probability law taking into account one censored data and preserving the conditional probabilities is given by :*

$$p'_k = \frac{N}{N+1} p_k, \text{ for } k < k_1,$$

$$p'_k = \frac{N t(k_1) + 1}{(N+1)t(k_1)} p_k, \text{ for } k \geq k_1$$

if the censored data corresponds to the information  $X \geq y_{k_1}$ .

We check that  $\sum p'_k = 1$ . Indeed,

$$\begin{aligned} \sum p'_k &= \sum_{k < k_1} p'_k + \sum_{k \geq k_1} p'_k \\ &= \frac{N s(k_1)}{N+1} + \frac{N t(k_1) + 1}{(N+1)t(k_1)} t(k_1) \\ &= \frac{N s(k_1) + N t(k_1) + 1}{N+1} = 1 \end{aligned}$$

### 3. General case

We now treat the case of any number of censored data. The theory is not an iteration of what was done for one point (since in such an approach the order of the points would come into consideration) : one takes into account all censored data at once.

Recall that we have classes  $y_1, \dots, y_K$  (possible values), with probabilities  $p_1, \dots, p_K$ . We have  $M$  censored data  $c_1, \dots, c_M$ . Let  $m_k$  be the number of censored data which fall into the  $k$ -th class ( $0 \leq m_k \leq M$ ,  $\sum_{k=1}^K m_k = M$ ).

Let us first look at the first class  $y_1$ . Among the  $N$  points of the precise sample, it received (on average)  $Np_1$  such points. Now,  $M$  censored data appear, among which  $m_1$  fall into this first class. We consider that their true value is distributed according to the original law, that is the proportion of these  $m_1$  data which have a life duration really in the first class is

$\frac{p_1}{p_1 + \dots + p_K}$ . The same way, the proportion of these  $m_1$  data which have a life duration really

in the  $j$ -th class is  $\frac{p_j}{p_1 + \dots + p_K}$ ,  $j = 1, \dots, K$ .

So, in  $N + M$  trials, the number of points which really fall into the first class is :

$$v_1 = Np_1 + \frac{p_1 m_1}{p_1 + \dots + p_K}$$

and therefore the probability of the first class, after the censored data have been incorporated, is :

$$p'_1 = \frac{Np_1 + \frac{p_1 m_1}{p_1 + \dots + p_K}}{N + M}$$

Let us turn to the second class. It received  $Np_2$  points from the exact sample. It receives

$\frac{p_2 m_1}{p_1 + \dots + p_K}$  points among the censored points which apparently fell into the first class, and it

keeps  $\frac{p_2 m_2}{p_2 + \dots + p_K}$  among the censored points which fell into the second class.

Note a change in the denominator : the censored points which fell into the second class can distribute only between  $K - 1$  classes.

So, the total number of points which are really expected to fall in the second class is :

$$v_2 = Np_2 + \frac{p_2 m_1}{p_1 + \dots + p_K} + \frac{p_2 m_2}{p_2 + \dots + p_K}$$

and its final probability, after the censored data have been incorporated, is :

$$p'_2 = \frac{Np_2 + \frac{p_2 m_1}{p_1 + \dots + p_K} + \frac{p_2 m_2}{p_2 + \dots + p_K}}{N + M}$$

Let us now turn to the  $k$ -th class,  $1 \leq k \leq K$ . It received originally  $Np_k$  points. From the censored points falling apparently in the first class, the  $k$ -th class will receive a number  $\frac{p_k m_1}{p_1 + \dots + p_K}$ .

From the censored points falling apparently in the  $j$ -th class, the  $k$ -th class ( $j \leq k$ ) will receive a number  $\frac{p_k m_j}{p_j + \dots + p_K}$ . So, the total number of points which are really expected to fall in the  $k$ -th class is :

$$v_k = Np_k + p_k \sum_{j=1}^k \frac{m_j}{p_j + \dots + p_K}$$

and the probability of the  $k$ -th class after all censored data have been introduced becomes :

$$p'_k = \frac{1}{N + M} \left( Np_k + p_k \sum_{j=1}^k \frac{m_j}{p_j + \dots + p_K} \right)$$

We observe that the sum  $p'_k$  is 1 ; indeed the sum of the terms  $Np_k$  is  $N$ , and then the sum of all contributions to all classes, coming from the censored data apparently falling in the first class is  $m_1$ , and so on : the contributions coming from the  $j$ -th class give  $m_j$  and therefore the total number of points to be incorporated is  $N + \sum m_k = N + M$ .

We have obtained :

**Theorem.** – *Let a sample be made of  $N$  "exact" points, and  $M$  "censored" points (one knows only a lower bound for their value). Let  $y_1, \dots, y_K$  be the possible distinct values of the results (which we call "classes") and let  $p_1, \dots, p_K$  be their respective probabilities built from the exact sample only. Then, the probability law which respects conditional probabilities, after the censored data is incorporated, is given by the formula :*

$$p'_k = \frac{1}{N + M} \left( Np_k + p_k \sum_{j=1}^k \frac{m_j}{p_j + \dots + p_K} \right) \quad k = 1, \dots, K$$

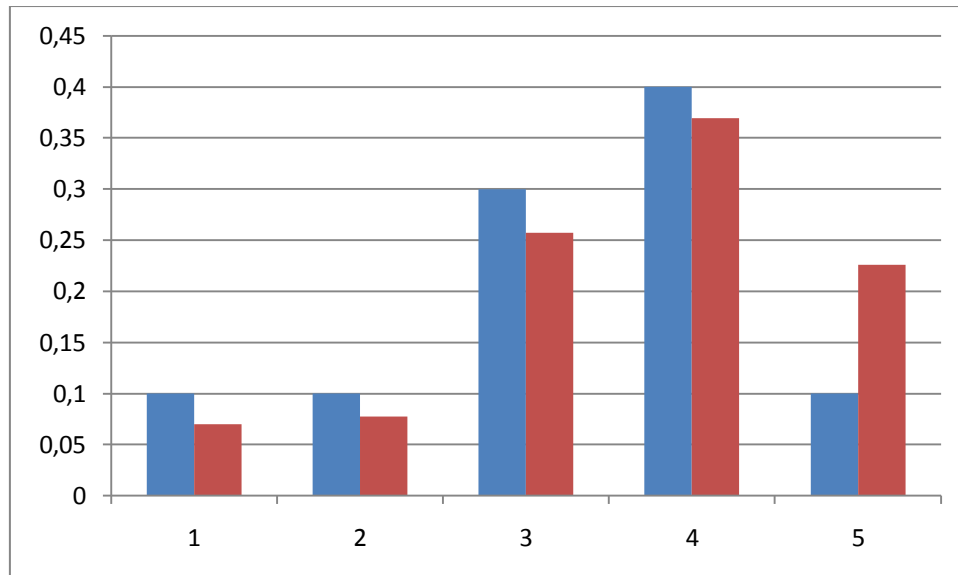
where  $m_k$  is the number of censored data which apparently fall into the  $k$ -th class.

### III. A simple example

Let us see the construction on a simple example. The data are the following :

months	number of deaths, uncensored	proba without censored	number, censored	proba with censored
1	2	0,1	1	0,07
2	2	0,1	2	0,08
3	6	0,3	2	0,26
4	8	0,4	1	0,37
5	2	0,1	4	0,23

the last column is obtained as an application of the Theorem above. The probability laws are given on the following graph (in blue : without censored data, in red : with censored data) :



Graph 1 : comparison of the probability laws

Here is a simple VBA code in order to apply the above formula :

```
Option Explicit
Const Ntot = 20
Const Mtot = 10
Const Ktot = 5

Sub macro1()

Dim k As Integer
Dim p(1 To Ktot) As Double
For k = 1 To Ktot
p(k) = Sheets(1).Cells(k + 1, 3)
Next k

Dim m(1 To Ktot) As Integer
For k = 1 To Ktot
```

```
m(k) = Sheets(1).Cells(k + 1, 4)
Next k
```

```
Dim cumul(1 To Ktot) As Double
cumul(Ktot) = p(Ktot)
For k = Ktot - 1 To 1 Step -1
cumul(k) = cumul(k + 1) + p(k)
Next k
```

```
Dim sum1(0 To Ktot) As Double
sum1(0) = 0
For k = 1 To Ktot
sum1(k) = sum1(k - 1) + m(k) / cumul(k)
Next k
```

```
Dim q(1 To Ktot) As Double
For k = 1 To Ktot
q(k) = 1 / (Ntot + Mtot) * (Ntot * p(k) + p(k) * sum1(k))
Next k
```

```
For k = 1 To Ktot
Sheets(1).Cells(k + 1, 5) = q(k)
Next k
Sheets(1).Cells(1, 5) = "proba with censored"
```

```
End Sub
```

#### **IV. Further remarks**

If we have a law which is truncated at some  $k_0$ , meaning that all values satisfying  $x \geq k_0$  are put at the place  $k_0$ , our method coincides with the usual one. No new class is added.

More generally, our method does not build any new class. This is satisfactory if  $N$  is large and  $M$  small, but if the converse happens, one may expect that the censored data would introduce new classes. This can be decided only by using some physical interpretation ; probabilistic methods by themselves never allow to introduce new values.