Société de Calcul Mathématique SA
Outils d'aide à la décision

Incorporating censored data

by Bernard Beauzamy

June 9th, 2013

Quite commonly, the results of some observations are "censored" in the statistical sense, which means that only an upper bound, or a lower bound, is known. This is typically the case in the following situations :

- Environmental concerns, when a threshold is defined (for instance a maximum concentration in some pollutant): then the true data is not kept, only the fact that $observation \leq threshold$.

- Medical situations: people are supposed to follow some treatment, but some of them leave before the end, so we only know that their life expectancy is at least the observed duration.

- Industrial objects which are tested for resistance: in some cases, the test is interrupted before the object breaks, so we only know a lower bound for the life duration of the object.

In practice, censored data are very common; we recommend to avoid this practice when possible (that is, to publish always the rough data when they exist), but in some cases this is unavoidable.

How to treat censored data is explained in my book "Nouvelles méthodes probabilistes pour l'évaluation des risques" [NMP], chapter V, but the presentations and the proofs are too complicated. I will present here an approach which is much simpler in practice (the theory is identical, of course).

Assume we have K "bins" $B_1,...,B_K$ (also called "classes") describing the possible results of some experiment. Let $b_1,...,b_K$ be the associated values: usually the B_k's are intervals and the b_k's are the centers of these intervals. Assume that $b_1 < \cdots < b_K$ (the reverse order is treated the same way).

Assume we have, for each $k=1,...,K$, n_k "normal values" (that is, the experiment shows that they are in the $k-$ th bin) and m_k censored values (that is, satisfying $X \le b_k$, where X is the observation). We treat here only the case $X \le b$; the other situations $X < b, X \ge b, X > b$ are identical.

Let $N = n_1 + \cdots + n_K$: this is the total number of normal values and let $M = m_1 + \cdots + m_K$: this is the total number of censored values.

Let us take an example, in order to explain the theory. We have ten bins, with values from 1 to 10. We have repeated N experiments, which gave some true values $n_1,...,n_{10}$. We now get a value, and we know only that it is ≤ 3 ? What to do with it ?

The answer is very simple. Compared to the information we got with the true values, the new observation ≤ 3 slightly increases the probability to have $X \le 3$ (since we have one more observation with ≤ 3), therefore slightly decreases the probability to have $X > 3$. But this new information says nothing inside the range $1,2,3$. So, if we assume $X \le 3$, the new information should not modify the conditional probability law. More precisely, the probability law of X, conditioned by the information $X \le 3$, should remain exactly the same, before and after the introduction of the censored data.

Using this remark, we will see iteratively how to introduce the censored data, that is in which bin to put them.

To start with, we have m_1 data satisfying $X \le b_1$: these data have to be put in the first bin, because there is no bin lower than this one.

Let us now consider the m_k data satisfying $X \le b_k$; we want to put them in one of the k bins $B_1,...,B_k$. We now describe how to do this. Let $\nu_{j,k}$ be the unknown number of data we will put in the $j-$ th bin, $j=1,...,k.$ Of course, $\nu_{1,k} + \cdots + \nu_{k,k} = m_k$.

We are looking at the conditional law $X \le b_k$. Then X may take only k values, namely $b_1,...,b_k$ and the number of occurrences is respectively $n_1,...,n_k$. Set, for simplicity, $N_k = n_1 + \cdots + n_k$.

Before we introduce any censored data, the conditional law of each bin, knowing $X \le b_k$, that is $P\{X = b_j \mid X \le b_k\}$ is simply:

$$p_{j,k} = \frac{n_j + 1}{N_k + k}$$

Indeed, recall from [NMP], chapter II, page 34, that if we have k classes, with n_j values in each of them, the probability of the $j-$ th class is:

$$p_{j,k} = \frac{n_j + 1}{\sum_{l=1}^{k} n_l + k}$$

Now, if we introduce $\nu_{j,k}$ values in the $j-$ th bin, this probability becomes:

$$q_{j,k} = \frac{n_j + \nu_{j,k} + 1}{N_k + m_k + k}$$

The theory says that, for all j, $q_{j,k} = p_{j,k}$, that is :

$$\frac{n_j + \nu_{j,k} + 1}{N_k + m_k + k} = \frac{n_j + 1}{N_k + k}$$

This gives:

$$\nu_{j,k} = \frac{n_j + 1}{N_k + k}\left(N_k + m_k + k\right) - \left(n_j + 1\right)$$

and finally:

$$\nu_{j,k} = \frac{m_k}{N_k + k}\left(n_j + 1\right)$$

So, the final answer is very simple: the new values will be distributed proportionally to the existing ones (plus one). Since the probability law is preserved at each step, one can start with $k = 1$, or with $k = K$, or in any order. The order of operations does not matter.

At the end, the $j-$th bin gets a number of elements equal to what it had, plus all incorporations, for $k \geq j$, that is:

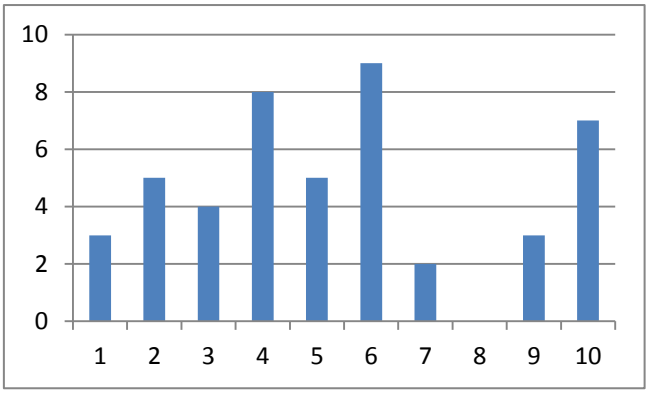
$$n_{j,final} = n_j + \left(n_j + 1\right)\sum_{k=j}^{K} \frac{m_k}{N_k + k}$$

An example

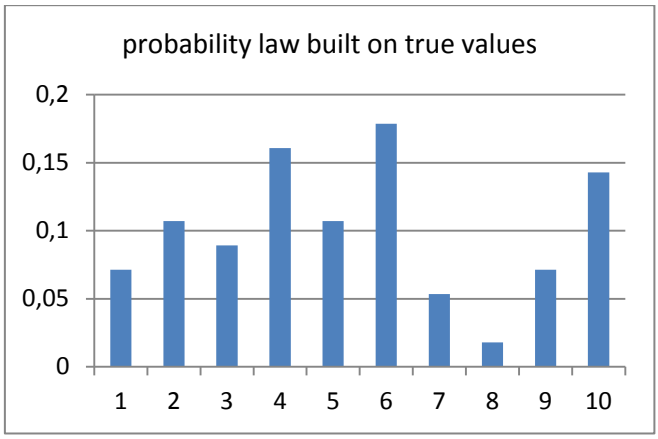
Let us treat an example, with 10 bins:

bin	nb of true values	nb of censored values
1	3	1
2	5	3
3	4	5
4	8	3
5	5	6
6	9	1
7	2	0
8	0	3
9	3	2
10	7	3

Here is the number of occurrences of true values:



Here is the associated probability law, built on true values:



Now, let us incorporate the censored values. The first one must of course be incorporated into the first bin. So, the composition of the bins after incorporation of the first becomes:

bin	nb of values
1	4
2	5
3	4
4	8
5	5
6	9
7	2
8	0
9	3
10	7

Now, we wish to incorporate the next 3 censored values, and they must be put either in the first or in the second.

The number in the first is:

$$\nu_{1,2} = \frac{m_2}{N_2 + 2}(n_1 + 1) = \frac{3}{8+2}(3+1) = 1.2$$

and the number in the second is:

$$\nu_{2,2} = \frac{m_2}{N_2 + 2}(n_2 + 1) = \frac{3}{8+2}(5+1) = 1.8$$

so the sum of both is 3. If we want integers, we will round up, and take $\nu_{1,2} = 1$, $\nu_{2,2} = 2$.

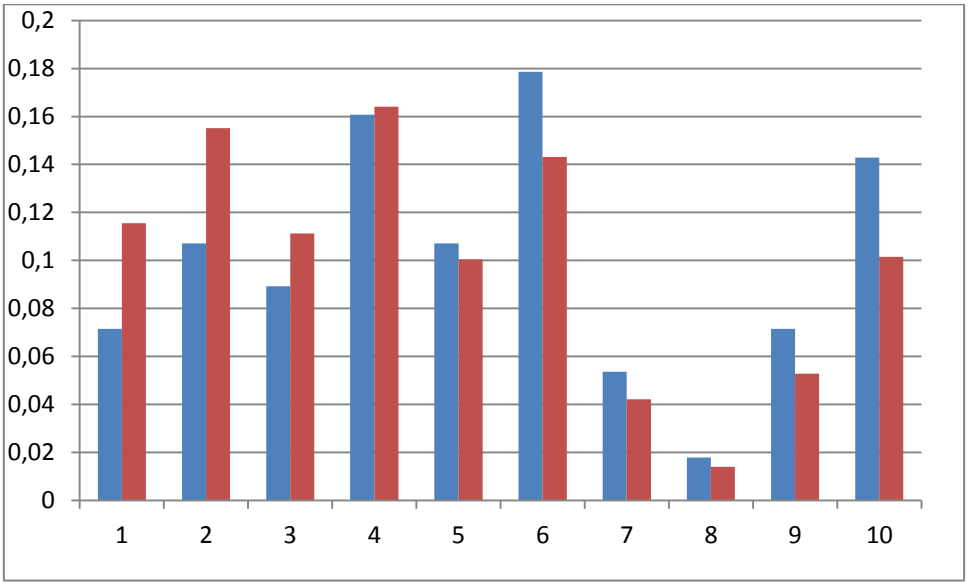
Here is the repartition at each step:

j\k	1	2	3	4	5	6	7	8	9	10	total
1	1,00	1,20	1,33	0,50	0,80	0,10	0,00	0,27	0,17	0,21	5,59
2		1,80	2,00	0,75	1,20	0,15	0,00	0,41	0,25	0,32	6,88
3			1,67	0,63	1,00	0,13	0,00	0,34	0,21	0,27	4,23
4				1,13	1,80	0,23	0,00	0,61	0,38	0,48	4,62
5					1,20	0,15	0,00	0,41	0,25	0,32	2,33
6						0,25	0,00	0,68	0,42	0,54	1,88
7							0,00	0,20	0,13	0,16	0,49
8								0,07	0,04	0,05	0,16
9									0,17	0,21	0,38
10										0,43	0,43
total	1	3	5	3	6	1	0	3	2	3	27

Here is the final probability law, after incorporation:

bin	nb of true values	nb of added values	total	probability
1	3	5,59	8,59	0,12
2	5	6,88	11,88	0,16
3	4	4,23	8,23	0,11
4	8	4,62	12,62	0,16
5	5	2,33	7,33	0,10
6	9	1,88	10,88	0,14
7	2	0,49	2,49	0,04
8	0	0,16	0,16	0,01
9	3	0,38	3,38	0,05
10	7	0,43	7,43	0,10

And this is the comparison between the original law (true values only) and the final law (censored data being incorporated):



In blue : original (true values only), in red: after incorporation of censored data.

We observe that the final law may be very different from the original one: only the conditional laws are kept at each step. For instance, if we have a large amount of data which fall into the first two bins, the final law will have a much larger contribution on these two bins.