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

The information associated with a sample

Bernard Beauzamy
Société de Calcul Mathématique SA

May, 2009

Abstract

We collect a sample : distinct values $x_1,...,x_K$, repeated respectively $n_1,...,n_K$ times. Let $N = n_1 + \cdots + n_K$ be the size of the sample. Let $(p_1,...,p_K)$ be the (unknown) probability of each value ; one usually uses the estimate $p_k \sim \frac{n_k}{N}$ which is correct only asymptotically. Here, we introduce a probability law on the the $K-$uple $(p_1,...,p_K)$ (not just an estimate, but a whole probability law). We study the marginal laws for each p_k and we show that the global variance of this probability law is a good indicator whether the sample is sufficient or not.

Acknowledgements : We thank Igor Carron for his help in the presentation of the paper.

Table of contents

I. Introduction

We consider the following question: we collect a sample (from a physical experiment, or a computational code, or from polls), and we want to know if this sample is "sufficient" or not. The word "sufficient" requires clarification. Roughly speaking, it means: does the sample give a correct idea of the "real" probability law behind the experiment? Or, more precisely, if we collected a sample which would be 10 times bigger, or 100 times bigger, and so on, would it give significantly different information?

In practice, this question is of considerable importance, for two reasons:

- If the sample is recognized as "insufficient", not to take wrong decisions ;
- Reduce, if possible, the time needed to acquire a "sufficient" sample, which is usually costly.

As an example of the first type, we may mention all questions connected to the warranty on industrial products. For instance, the warranty on a car (say 3 years) may be too long, or too short, if the sample upon which it is based is insufficient.

As an example of the second type, we may mention a contract we had with the French "Institut de Radioprotection et de Sûreté Nucléaire" : it dealt with the improvement of nuclear measurements. But these measurements require neutron counting, which takes time ; so the question is when to stop ? (see [IRSN-SCM] for a published description of this work).
To decide whether a sample is sufficient or not is not the same as reconstructing missing data, as we did in [BBOZ]. In the reconstruction of missing data, we have some knowledge of the underlying probability law, and, when a sample is insufficient, precisely we do not have this knowledge.

II. Two different situations : static and dynamical

There are obviously two different situations : either the sample is given to you at once, as a big database, with no information on how it was collected, or you see it growing (or, which is the same, you have an information about the date when each data was collected). We call the first situation "static" and the second one "dynamical". Note that there may be intermediate situations : one sample, made for example of a static database and of a dynamical database. But in this case we consider the whole sample as static.

III. Some terminology

Before we enter the discussion, let us give some terminology. We call a "class" a possible value for the results. For instance, if you measure the heights of people between 1.5 and 2 meters, with precision 0.1 m, you have 5 classes, namely 1.5 - 1.6, 1.6 - 1.7, 1.7 - 1.8, 1.8 - 1.9, 1.9 - 2.

So, the number of classes depends :

- On the range of the measurement (here the range is 1.5 - 2), which usually depends on what you want to measure and for what reason (this is "expert knowledge") ;

- On the precision of the measurement. This precision depends itself on the precision of the measurement device, but also on the objectives of your measurement. For instance, if you simply want to establish correlations between food and height, a precision of 0.1 m might be enough, but if you are a tailor, you might want a precision of 1 cm or even below.

Mathematically, and in terms of computer implementation, a "class" is usually an interval of the type $a \le x < b$, but this has no importance here. To a class we attribute a "value", which is usually the center of the interval. This choice, also, has no importance here.

Simply speaking, we may consider our sample as a list of distinct values $x_1,...,x_K$ (so K is the number of observed classes) with $n_1,...,n_K$ being the number of repetitions for each value $x_1,...,x_K$ respectively. Let $N = n_1 + \cdots + n_K$: this is the size of the sample.

Let us observe that there are situations where the number of possible classes is unknown, or has only a useless upper bound. For instance, if you sell a new device, and want to know the number of failures in a given year, the only bound you know is the total number of devices sold (but of course you expect that not all of them will break down!).

If you go to a new region, where you have never been, and ask the question : what is the law of probability for the number of cows in a farm, you have no a priori idea of the possible range (can it be 1 000 ? or 10 000 ? or more ?).

In the static situation, all you know is a list of values ; you know nothing about the way they have been obtained.

The tool we now describe allows us to build a probability law from the sample, giving in particular a confidence interval for the probability of each value : if this confidence interval is narrow for most values, this is satisfactory.

This gives a very complete description of the information contained in the sample, including the uncertainties about it.

IV. The static situation

We want to evaluate the "true" probability p_k of each value x_k, $k = 1,...,K$. This "true" probability is unknown: we have only the estimate observed on the sample. So p_k should itself be treated as a random variable, of which we want to find the law.

A classical estimate is:

$$p_k \approx \frac{n_k}{N}$$

However, this estimate is correct only asymptotically that is when $N \to \infty$: this is the empirical law of large numbers. A correct estimate, valid for all N, is:

$$p_k = \frac{n_k + 1}{N + K},$$

as we will see below.

But in fact, the general theory presented in [BB1], Chapter 14, §9 (multinomial case) gives more than estimates for each p_k. Indeed, one can build a density function for the $K-$uple $p_1,...,p_K$. This density function is :

$$f(p_1,...,p_K) = c\ p_1^{n_1} \cdots p_K^{n_K} \tag{1}$$

where $p_1,...,p_K$ satisfy $p_k \geq 0$ for all k, $p_1 + \cdots + p_K = 1$, and c is a normalization constant (so that the integral of f is 1).

Let us explain the meaning of this formula. In our sample, the value x_1 has been observed n_1 times and we would like to estimate the probability p_1 of this value. This probability is unknown, so, as we said, it should be treated as a random variable. Formula (1) gives the density of this random variable, or, more exactly, the joint density of the $K-$uple $p_1,...,p_K$ of random variables.

The normalization constant c can be easily computed ([BB1], Chapter 14, §9) ; its value is :

$$c = \frac{1}{I(n_1,...,n_K)}$$

with :

$$I(n_1,...,n_K) = \frac{n_1! \cdots n_K!}{(N+K-1)!}, \text{ where } N = n_1 + \cdots + n_K .$$

So we have a very explicit situation : the sample gives birth to a probability law on the probabilities of each value, and the amount of information will be derived from this probability law. For instance, we will derive a probability law for p_1 and the more concentrated this probability law is, the more certain p_1 is. If this probability law for p_1 was a Dirac, say for instance at 0.1, this would mean that p_1 is certainly equal to 0.1.

If this was true for all p_k, we would be sure that our sample characterizes completely the experiment : the sample would be perfect. So we see that, for any sample, the concentration of the law $f(p_1,...,p_K)$ characterizes correctly the amount of information in the sample.

We first compute explicitly the marginal law of any of the p_k's.

1. Computing the marginal law for p_1

We have, by definition :

$$f(p_1,...,p_K)=\frac{p_1^{n_1}\cdots p_K^{n_K}}{I(n_1,...,n_K)},\ \ 0\le p_1,...,p_K\le 1,\ p_1+\cdots+p_K=1$$

with :

$$I(n_1,...,n_K)=\frac{n_1!\cdots n_K!}{(N+K-1)!},\ \text{and}\ N=n_1+\cdots+n_K\ .$$

We define, for any positive integers a,b :

$$f_{a,b}(x)=\frac{x^a(1-x)^b}{I(a,b)}$$

with as before :

$$I(a,b)=\frac{a!b!}{(a+b+1)!}$$

We recall that each function $f_{a,b}$ is a density function : see [BB1], Chapter 14.

Then we have :

Proposition 1. – *The marginal law of each* p_k *is the density* $f_{n_k,N-n_k+K-2}$.

Proof of Proposition 1.

We give it for p_1 in order to simplify the notation ; it is identical for the others. We compute :

$$I_1=\int_0^{1-p_1-...-p_{K-2}} p_1^{n_1}\cdots p_{K-1}^{n_{K-1}}\left(1-p_1-...-p_{K-1}\right)^{n_K}dp_{K-1}$$

(the normalization constants will be considered at the end)

We have :

$$I_1=p_1^{n_1}\cdots p_{K-2}^{n_{K-2}}\int_0^{1-p_1-...-p_{K-2}} p_{K-1}^{n_{K-1}}\left(1-p_1-...-p_{K-1}\right)^{n_K}dp_{K-1}$$

Set $p_{K-1}=t\left(1-p_1-...-p_{K-2}\right)$ with $0\le t\le 1$. Then :

$$I_1 = p_1^{n_1} \cdots p_{K-2}^{n_{K-2}} \left(1 - p_1 - \ldots - p_{K-2}\right)^{n_{K-1}+n_K+1} \int_0^1 t^{n_{K-1}} (1-t)^{n_K} dt$$

and thus, up to some normalization constant :

$$I_1 = p_1^{n_1} \cdots p_{K-2}^{n_{K-2}} \left(1 - p_1 - \ldots - p_{K-2}\right)^{n_{K-1}+n_K+1}$$

Next :

$$I_2 = \int_0^{1-p_1-\ldots-p_{K-3}} p_1^{n_1} \cdots p_{K-2}^{n_{K-2}} \left(1 - p_1 - \ldots - p_{K-2}\right)^{n_{K-1}+n_K+1} dp_{K-3}$$

The same computation gives :

$$I_2 = p_1^{n_1} \cdots p_{K-3}^{n_{K-3}} \left(1 - p_1 - \ldots - p_{K-3}\right)^{n_{K-2}+n_{K-1}+n_K+2}$$

and more generally :

$$I_j = p_1^{n_1} \cdots p_{K-j-1}^{n_{K-j-1}} \left(1 - p_1 - \ldots - p_{K-j-1}\right)^{n_{K-j}+\ldots+n_K+j}$$

The marginal law is obtained for $K - j - 1 = 1$, that is $j = K - 2$. We find the law :

$$f(p_1) = p_1^{n_1} \left(1 - p_1\right)^{n_2+\ldots+n_K+K-2} = p_1^{n_1} \left(1 - p_1\right)^{N-n_1+K-2}$$

Let us now take care about the normalization. Recall that if :

$$I_{a,b} = \int_0^1 x^a (1-x)^b dx,$$

where a, b are positive integers, then :

$$I(a,b) = \frac{a!b!}{(a+b+1)!}.$$

([BB1], Chapter 14, §5, Lemma 2). This gives the final formula :

$$f(p_1) = \frac{p_1^{n_1} \left(1 - p_1\right)^{N-n_1+K-2}}{I(n_1, N - n_1 + K - 2)} \tag{2}$$

We observe that, if the number of classes is strictly above 2, this law does not coincide with the law we would have, dividing into 2 classes: the first one and the rest. This law would have as a density:

$$f_{n_1, N-n_1}(p_1) = \frac{p_1^{n_1} \left(1 - p_1\right)^{N-n_1}}{I(n_1, N - n_1)}$$

([BB1], Chapter 14, §5, Proposition 1).

2. Computing the expectations and the variances

We recall that the expectation of $f_{n,N-n}(x) = c\, x^n (1-x)^{N-n}$ is $\frac{n+1}{N+2}$ ([BB1], Chapter 14, §5.B, Proposition 3).

The expectation of $f(p_1) = \frac{p_1^{n_1}(1-p_1)^{N-n_1+K-2}}{I(n_1, N-n_1+K-2)}$ is therefore :

$$\mu_1 = E(p_1) = \frac{n_1+1}{N+K},$$

and the same for the other p_k :

$$\mu_k = E(p_k) = \frac{n_k+1}{N+K} \tag{3}$$

We see that this quantity is different from n_k / N ; both coincide only asymptotically, when $N \to +\infty$. We observe also that μ_k is never equal to 0, even if $n_k = 0$: a value which has never been observed may be seen in the future.

The sum of expectations is of course equal to 1 :

$$\sum_{k=1}^{K} E(p_k) = \sum_{k=1}^{K} \frac{n_k+1}{N+K} = \frac{1}{N+K}\left(\sum_{k=1}^{K}(n_k+1)\right) = \frac{1}{N+K}(N+K) = 1$$

By [BB1], Chapter 14, §5.E, Proposition 5, the variance of $f_{n,N-n}$ is :

$$\sigma_{n,N}^2 = \frac{(n+1)(N-n+1)}{(N+2)^2(N+3)}$$

with N replaced by $N+K-2$. So, for each $k = 1,...,K$

$$\sigma_k^2 = \frac{(n_k+1)(N+K-1-n_k)}{(N+K)^2(N+K+1)} \tag{4}$$

Let us compute the sum of all variances. We find :

$$\sigma^2 = \sum_{k=1}^{K} \sigma_k^2 = \frac{\sum_{k=1}^{K}(n_k+1)(N+K-1-n_k)}{(N+K)^2(N+K+1)}$$

that is :

$$\sigma^2 = \frac{\left(\sum_{k=1}^{K}(n_k+1)\right)^2 - \sum_{k=1}^{K}(n_k+1)^2}{(N+K)^2(N+K-1)}$$

or :

$$\sigma^2 = \frac{\left(\sum_{k=1}^{K}(n_k+1)\right)^2 - \sum_{k=1}^{K}(n_k+1)^2}{\left(\sum_{k=1}^{K}(n_k+1)\right)^2\left(\sum_{k=1}^{K}(n_k+1)-1\right)}$$

Set $\nu_k = n_k + 1$. We get a simple formula :

$$\sigma^2 = \frac{\left(\sum_{k=1}^{K}\nu_k\right)^2 - \sum_{k=1}^{K}\nu_k^2}{\left(\sum_{k=1}^{K}\nu_k\right)^2\left(\sum_{k=1}^{K}\nu_k-1\right)}$$

and, with $S = \sum_{k=1}^{K}\nu_k$ and $S_2 = \sum_{k=1}^{K}\nu_k^2$, we can write :

$$\sigma^2 = \frac{S^2 - S_2}{S^2(S-1)}$$

However, this global variance is not a good indicator ; it gives the same weight to all elementary terms, and, moreover, we are interested in separate estimates for all k 's. So, let us come back to elementary variances.

We have the following elementary proposition :

Proposition 2. – *For each* k *, for fixed* N *and* K *, the variance* σ_k^2 *, as a function of* n_k *, is increasing when* $0 \le n_k \le \frac{N+K}{2} - 1$ *and decreasing when* $\frac{N+K}{2} - 1 \le n_k \le N$ *. The maximum value, obtained for* $n_k = \frac{N+K}{2} - 1$*, is* $\sigma^2_{k,\max} = \frac{1}{4(N+K+1)}$*. The minimum value, obtained for* $n_k = 0$*, is* $\sigma^2_{k,\min} = \frac{N+K-1}{(N+K)^2(N+K+1)} \approx \frac{1}{(N+K)^2}$.

The proof of this proposition follows immediately from formula (4). We observe that the value for $n_k = N$ is $\frac{(N+1)(K-1)}{(N+K)^2(N+K+1)}$, which is usually bigger than $\frac{1}{(N+K)^2}$.

Here is the graph of σ_k^2 for $N = 1000$, $K = 10$:

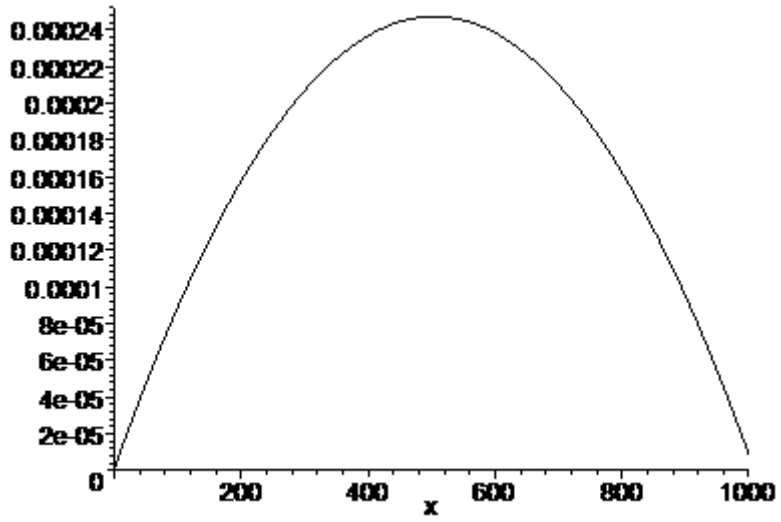

Figure 1 : Graph of the variance of each term

As a corollary, we see that the variance, for each k, tends to zero when the sample increases.

We observe that the variance of each term does not need to decrease when the sample increases (that is, N is replaced by $N+1$). Let us see this on an example, when n_1 is replaced by n_1+1. The condition:

$$\sigma_1^2(n_1+1)=\frac{(n_1+2)(N+K-1-n_1)}{(N+K+1)^2(N+K+2)}\le\sigma_1^2(n_1)=\frac{(n_1+1)(N+K-1-n_1)}{(N+K)^2(N+K+1)}$$

is equivalent to :

$$\frac{n_1+2}{(N+K+1)(N+K+2)}\le\frac{n_1+1}{(N+K)^2}$$

or :

$$\frac{n_1+2}{n_1+1}\le\frac{(N+K+1)(N+K+2)}{(N+K)^2}$$

$$1+\frac{1}{n_1+1}\le\frac{(N+K)^2+3(N+K)+2}{(N+K)^2}=1+\frac{3}{N+K}+\frac{2}{(N+K)^2}$$

$$\frac{1}{n_1+1}\le\frac{3}{N+K}+\frac{2}{(N+K)^2}$$

which is not true in general.

Asymptotically, when $N\to+\infty$, the behavior of σ_k is easy to obtain :

Proposition 3. – *When* $N \to +\infty$,

$$\sigma_k^2 \sim \frac{p_k(1-p_k)}{N},$$

where $p_k = \lim_{N\to+\infty} \frac{n_k}{N}$.

3. Dependence upon the number of classes

We observe that the number of classes K appears in all these estimates. If we restrict ourselves to the values which appeared, this is perfectly clear and legitimate. However, as the examples below will show, we often want to take into account values which have not appeared but might, if the sample was bigger. For instance, if in some river we observed, for some pollutant, a concentration of 0.3 g/l and in another a concentration of 0.5 g/l, there is no reason that a concentration of 0.4 g/l might not exist somewhere. Also, there is a question about the upper limit and the lower limit : should we stop at 1 g/l, 10 g/l, and so on ?

Another example is about temperatures. If our experiment deals with the temperatures that one can observe in Paris, what extremes should we take ? Of course, we should not restrict ourselves to the temperatures which have already been observed, since warmer ones and colder ones are also possible.

There is no standard, theoretical answer to this question. The list of possible classes depends on the problem ; it has to be set both looking at the existing (historical) data, and the physics of the problem.

4. Confidence interval for each p_k

Since we have a probability law for p_k, and since we know its expectation and variance, we can deduce a confidence interval, using Bienaymé-Chebycheff's inequality, for each k, under the form :

$$P\{|X-\mu|>\varepsilon\} \le \frac{\sigma^2}{\varepsilon^2}$$

that is, with our present notation :

$$P\{\mu_k-\varepsilon \le p_k \le \mu_k+\varepsilon\} \ge 1-\frac{\sigma_k^2}{\varepsilon^2}$$

We observe that, usually, when the law of the variable is explicitly given (as it is the case here), Bienaymé-Chebycheff's inequality is not the best way to obtain confidence intervals : explicit computations of the integrals give better results in general. But here, it turns out that this inequality gives more precise intervals than the ones which were derived (by explicit, but approximate computations) in [BB1], Chapter 14, §11.

We want a confidence interval for p_k, with confidence $\alpha = 0.95$ (this is an arbitrary choice). This means that :

$$1 - \frac{\sigma_k^2}{\varepsilon^2} = \alpha$$

Therefore, we have to take :

$$\varepsilon = \frac{\sigma_k}{\sqrt{1-\alpha}}$$

So the interval :

$$I_k = \left\{ \mu_k - \frac{\sigma_k}{\sqrt{1-\alpha}} \le p_k \le \mu_k + \frac{\sigma_k}{\sqrt{1-\alpha}} \right\}$$

is a α - confidence interval for p_k. This is true in general ; however, this interval may be quite large. We will say that the sample is satisfactory is most of these intervals are small enough, namely if they fit within a given precision.

Fix a precision $\eta = 1/10$ (this choice is arbitrary). We will say that the interval I_k above fits within the precision η if it is of the form $\left[(1-\eta)x,(1+\eta)x\right]$. This is the case if the condition :

$$\frac{\sigma_k}{\sqrt{1-\alpha}} \le \mu_k \eta$$

that is :

$$\frac{\sigma_k}{\mu_k} \le \eta\sqrt{1-\alpha}$$

is satisfied. We define :

$$w_k = \frac{\sigma_k}{\mu_k}$$

(the letter w stands for "width" of the interval). Using the above definitions of μ_k and σ_k we get :

$$w_k = \frac{N + K - 1 - n_k}{(N + K + 1)(n_k + 1)}$$

and we want :

$$w_k \le \eta^2 (1-\alpha) \qquad \text{(C)}$$

We have obtained :

Proposition 4. - *We have :*

$$P\left\{\frac{n_k+1}{N+K}(1-\eta) \le p_k \le \frac{n_k+1}{N+K}(1+\eta)\right\} \ge \alpha$$

as soon as :

$$n_k \ge \frac{N+K}{1+\eta^2(1-\alpha)(N+K+1)} - 1 \qquad (C_1)$$

This condition is a fortiori satisfied as soon as :

$$n_k \ge \frac{1}{\eta^2(1-\alpha)} \qquad (C_2)$$

Proof of Proposition 4 : the condition (C_1) is simply a rephrasing of condition (C), using the definition of w_k ; the fact that (C_2) is stronger than (C_1) is obvious.

We note that condition (C_2) does not depend on N or K ; this is an absolute condition, which says that if there are enough observations in the k -th class, then there is a high probability to have a sharp estimate on p_k .

For instance, if $\eta = 0.1$ (10% precision) and $\alpha = 0.95$ (95% confidence), we find $n_k = 2\,000$. It means that, no matter how large the total sample is, no matter how many classes there are, if a class has more than 2 000 observations, a 10 % precision will be obtained upon p_k with 95% probability.

Such an estimate, valid for all N and K, is very conservative (this is due to the use of the inequality of Bienaymé-Tchebycheff). So later we will find more precise means to determine whether the information is sufficient.

In fact, we do not need condition (C) to hold for all k 's. We will be satisfied if it holds for most of them. Fix a "satisfaction index" $\beta = 0.96$ (this is arbitrary ; we took it different from α). Then we will be satisfied with satisfaction index β if condition (C) holds for a set $p_{k_1},...,p_{k_n}$ with $p_{k_1}+\cdots+p_{k_n} \ge \beta$.

In practice, one will check the converse, namely that the sum of all p_k for which condition (C) is not satisfied is at most $1-\beta$.

So, finally, we need three concepts in order to define global satisfaction :

- The number β (here 0.96), which indicates the proportion of classes which are correctly handled from the sample. Here, we want 96 % of the classes.
- The number α which denotes the confidence interval. We are sure, with probability $\alpha = 0.95$ that each p_k will be in the chosen interval.

– The precision η which relates the size of the interval to a precision of the measurement. It says that the interval will be of the form $\left[(1-\eta)p_k,(1+\eta)p_k\right]$.

5. Asymptotic estimates for the confidence interval for each p_k

We now determine explicitly the set where the density of each p_k is small ; these estimates improve upon the ones given in [BB1], Chapter 14. They are only asymptotic estimates, valid when $N \to +\infty$.

In order to simplify the notation, let p be any of the p_k's and let n be the corresponding n_k. Let $f(x)$ be the density of p. We have :

Proposition 5. – *The maximum value of f is taken for $x = p$; its value is :*

$$\max_x f(x) = \sqrt{\frac{N}{2\pi p(1-p)}}$$

For any $\varepsilon > 0$, outside the interval $I = \left[p-\eta, p+\eta\right]$ with :

$$\eta = \left(\frac{1}{N} Log \frac{\sqrt{N}}{\varepsilon\sqrt{2\pi p(1-p)}} \right)^{1/2}$$

the function f satisfies $f(x) < \varepsilon$.

Finally, for any $\varepsilon > 0$, on the interval $I' = \left[p-\eta', p+\eta'\right]$ with :

$$\eta' = \sqrt{\frac{p(1-p)}{N\varepsilon}}$$

the function f satisfies $\int_{I'} f(x)dx \geq 1-\varepsilon$ .Moreover, at the endpoints :

$$f(p \pm \eta') \sim \sqrt{\frac{N}{2\pi p(1-p)}} - \frac{1}{\varepsilon}\sqrt{\frac{p(1-p)}{2\pi N}}$$

Proof of Proposition 5.

When $N \to +\infty$, $n \sim pN$, and the density $f(x)$ given by Proposition 1, simplifies to :

$$f(x) \sim \varphi(x)$$

with :

$$\varphi(x) = \frac{\left(x^p(1-x)^{1-p}\right)^N}{I(pN,(1-p)N)}$$

Using Stirling's formula, we have :

$$\begin{aligned} I(pN,(1-p)N) &= \frac{(pN)!((1-p)N)!}{(N+1)!} \\ &\sim \frac{p^{pN}(1-p)^{(1-p)N}}{N+1}\sqrt{2\pi p(1-p)N} \\ &\sim \left(p^{p}(1-p)^{(1-p)}\right)^{N}\sqrt{\frac{2\pi p(1-p)}{N}} \end{aligned}$$

and therefore :

$$\varphi(x) \sim \left(\frac{x^{p}(1-x)^{1-p}}{p^{p}(1-p)^{(1-p)}}\right)^{N}\sqrt{\frac{N}{2\pi p(1-p)}}$$

Since the maximum of the function $h(x) = x^{p}(1-x)^{1-p}$ is obtained for $x = p$, we see that the maximum is $\varphi(x)$ is also obtained for $x = p$ and the maximum value is :

$$\varphi(p) = \max_{x}\varphi(x) = \sqrt{\frac{N}{2\pi p(1-p)}}\,.$$

When $N \to +\infty$, the function $\varphi(x)$ is more and more concentrated around its mean, which is p (since $\sigma \to 0$). Therefore, let $x = p-\eta$, with $\eta \to 0$ when $N \to +\infty$, and let us evaluate $\varphi(x)$. We have :

$$\begin{aligned} \varphi(p-\eta) &\sim \left(\frac{(p-\eta)^{p}(1-p+\eta)^{1-p}}{p^{p}(1-p)^{(1-p)}}\right)^{N}\sqrt{\frac{N}{2\pi p(1-p)}} \\ &\sim \left(\left(1-\frac{\eta}{p}\right)^{p}\left(1+\frac{\eta}{1-p}\right)^{1-p}\right)^{N}\sqrt{\frac{N}{2\pi p(1-p)}} \\ &\sim \left((1-\eta)(1+\eta)\right)^{N}\sqrt{\frac{N}{2\pi p(1-p)}} \\ &\sim \left(1-\eta^{2}\right)^{N}\sqrt{\frac{N}{2\pi p(1-p)}} \end{aligned}$$

and the same way we obtain :

$$\varphi(p+\eta) \sim \left(1-\eta^{2}\right)^{N}\sqrt{\frac{N}{2\pi p(1-p)}}$$

So finally :

$$\varphi(p\pm\eta) \sim \left(1-\eta^{2}\right)^{N}\sqrt{\frac{N}{2\pi p(1-p)}} \qquad (1)$$

Let $\varepsilon > 0$; the condition $\varphi(p-\eta) \leq \varepsilon$ is equivalent to :

$$1-\eta^2 \leq \varepsilon^{1/N} \left(\frac{2\pi p(1-p)}{N} \right)^{1/2N}$$

or $\eta \geq \eta_1$, with :

$$\eta_1 = \left(1-\varepsilon^{1/N} \left(\frac{2\pi p(1-p)}{N} \right)^{1/2N} \right)^{1/2}$$

We observe that :

$$\eta_1 \sim \left(\frac{1}{N} Log \frac{\sqrt{N}}{\varepsilon\sqrt{2\pi p(1-p)}} \right)^{1/2}$$

This proves the second part of our Proposition. The third part follows immediately from Bienaymé-Tchebycheff's inequality, since :

$$\int_{m-\frac{\sigma}{\sqrt{\varepsilon}}}^{m+\frac{\sigma}{\sqrt{\varepsilon}}} f(x)dx = P\left\{ m-\frac{\sigma}{\sqrt{\varepsilon}} \leq X \leq m+\frac{\sigma}{\sqrt{\varepsilon}} \right\} \geq 1-\varepsilon$$

with $m = p$ and $\sigma = \sqrt{\frac{p(1-p)}{N}}$. The estimates at the endpoints come from (1), with η replaced by $\eta' = \sqrt{\frac{p(1-p)}{N\varepsilon}}$.

6. A first example : pollution in rivers

The following example comes from a contract we had in 2008 with the Water Agency "Artois-Picardie" in France. It deals with the concentration in the pollutant NH4.

There are 104 measure points, and at each point the concentration in NH4 is measured. Among these 104 measure points, there is only one where the concentration is 0.03 g/l (to be understood as in the class 0.025 - 0.035).

So, with our previous notation, we have $n_1 = 1,\ N = 104$. The number of possible values is : $K = 2000$ (basically, from 0 to 20 g/l). We are interested in the probability density of p_1, to be in the class 0.025 - 0.035 g/l.

This density is :

$$f(p_1) = \frac{p_1(1-p_1)^{2101}}{I(1,2101)}$$

with :

$$I(1,2101) = \frac{1!\,2101!}{2103!} = \frac{1}{2102 \times 2103}$$

The expectation is :

$$\mu_1 = E(p_1) = \frac{n_1 + 1}{N + K} = \frac{2}{2104} = 0.95 \times 10^{-3}$$

and the variance :

$$\sigma_1^2 = \frac{(n_1 + 1)(N + K - 1 - n_1)}{(N + K)^2 (N + K + 1)} = \frac{2 \times 2102}{(2104)^2 \times 2105} = 0.45 \times 10^{-6}$$

Now, we use Bienaymé-Chebycheff's inequality, under the form :

$$P\{\mu_1 - \varepsilon \le p_1 \le \mu_1 + \varepsilon\} \ge 1 - \frac{\sigma^2}{\varepsilon^2}$$

We want this probability to be 0.95 (this is an arbitrary choice). So we need to take ε such that :

$$1 - \frac{\sigma^2}{\varepsilon^2} = 0.95$$

which gives :

$$\varepsilon = \frac{\sigma}{\sqrt{0.05}} = 0.3 \times 10^{-2}$$

The quantity $\mu_1 - \varepsilon$ is negative, which gives no information, since a concentration must be positive. So, using only the bound $\mu_1 + \varepsilon$, we get that :

$$0 \le p_1 \le 0.405 \times 10^{-2}$$

with probability 0.95.

So we obtain the following result : the probability, for a given measure station, to be in the class $\mathrm{NH4} = 0.03$ is ≤ 0.004 with a 95% confidence. This amount of pollution (very low concentration) is expected to be extremely rare.

Similar results may be obtained for all classes.

7. A second example : Trains being late

During the year 2008, we had a contract with the "Réseau Ferré de France" (French Railways) ; it dealt with the statistics of the delays for the trains. Data are as follows : for a given delay x_k, in minutes, between 1 and $K = 600$ maximum, we know the number n_k of trains having this delay, among all French trains in a given year. The total $N = \sum_{k=1}^{600} n_k$ is therefore the number of trains which, during that year, experienced a delay between 1 and 600 minutes.

With the above notation, we fix $\alpha = 0.90$ (90% confidence interval) $\eta = 0.1$ (precision 10%). We find that condition (C), that is $w_k \leq \eta^2 (1-\alpha)$, is satisfied for 84 % of the trains ; more precisely those which have delay ≤ 26 min, and is not satisfied for those which have longer delays.

What does this mean in practice ? We consider here the delays of the trains as a random variable (this delay depends on many factors, which are not perfectly known). Our statement means that, if we want to know the law of this random variable, the sample coming from the year 2006 is satisfactory (up to the precision announced) for trains with delays ≤ 26 min and is insufficient for trains with longer delays. In other words, for trains with longer delays, the estimate we get from one year of observation is too vague, because such trains are not numerous enough ; in order to improve the sample, more years are necessary.

8. Properties of the precision width w_k

Recall that :

$$w_k = \frac{N + K - 1 - n_k}{(N + K + 1)(n_k + 1)}$$

defines the size of the interval around each p_k. We omit the subscript k and we write $w(N,K,n)$ to indicate that this number depends on the total size of the sample N, on the number of classes K and on the number of realizations n in that particular class.

We have the following results :

Proposition 6. - *The number $w(N,K,n)$ decreases (that is, we have a smaller interval) if a new measurement appears in that particular class, that is :*

$$w(N+1,K,n+1) \leq w(N,K,n)$$

The proof of this statement is obvious, since the numerator does not change, whereas the denominator increases.

Proposition 7. - *The number* $w(N,K,n)$ *increases if a new measurement appears in another class, that is :*

$$w(N+1,K,n) \geq w(N,K,n)$$

Indeed, this statement means :

$$\frac{N+K-n}{(N+K+2)(n+1)} \geq \frac{N+K-n-1}{(N+K+1)(n+1)}$$

or :

$$\frac{N+K-n}{N+K+2} \geq \frac{N+K-n-1}{N+K+1}$$

which is equivalent to :

$$1-\frac{n+2}{N+K+2} \geq 1-\frac{n+2}{N+K+1}$$

which is satisfied.

Proposition 8. - *The number* $w(N,K,n)$ *increases if a new (empty) class appears, that is :*

$$w(N,K+1,n) \geq w(N,K,n) .$$

The proof is the same as for Proposition 7 above.

These properties are easy to understand : a new measurement in the first class improves the estimate upon p_1 ; all other situations make it worse.

Our conclusion in this paragraph is that the probability law $f(p_1,...,p_K)$ gives a satisfactory information about the sample : we may conclude if the sample is sufficient or not, or, more specifically, if some classes are sufficient or not.

We observe that the precision of the measurement devices appears in this definition, but in a hidden manner : it appears in the number of classes. If the measurement is very precise, the number of classes is high ; if the precision is low, so is the number of classes. In our example regarding trains, the precision is one minute ; therefore we have 600 classes for delays between 1 minute and 10 hours. If the range was the same, but with precision 10 seconds, we would have 6 times more classes.

V. The dynamical situation

The dynamical situation seems much more satisfactory: we see the sample growing. So we may decide to stop when the total sample contains hardly more information than the samples we collected earlier.

Let, as before, K be the number of classes. Let $\nu = 1,...,N$ be the successive sizes of the sample. Let $X_{i,k}$ be the random variable which indicates whether, at the i-th trial, we fell into the k-th class or not : so $X_{i,k} = 1$ if we fell in the k-th class, or 0 otherwise. We have :

$\sum_{k=1}^{K} X_{i,k} = 1$, for all i (this means that we have only one result for each trial)

$\sum_{i=1}^{N} X_{i,k} = n_k$, for all k , where n_k is, as before, the number of occurrences of the k-th class among N trials.

The $X_{i,k}$ are independent Bernoulli random variables. The probability $p_k = P\{X_{i,k} = 1\}$ is unknown, it is the same for all i : this is what we want to estimate. We will show how to do it for the first class and it will be the same for all k. Therefore, in the sequel, we drop the subscript k and speak simply of p instead of p_k. The same way, we write simply X_i instead of $X_{i,k}$.

The X_i are independent random variables with same law. Each variable X_i has mean p. We set $\sigma_0 = \sqrt{p(1-p)}$, so σ_0^2 is the variance of all the X_i's.

At the ν-th step, we look at the quantity $Y_\nu = \frac{X_1 + \cdots + X_\nu}{\nu}$; it indicates the proportion of successes of the first class among the first ν trials. By the empirical law of large numbers, $Y_\nu \to p$ when $\nu \to +\infty$.

Let y_ν be the realization of Y_ν on our sample (the observed values). Then the quantity :

$$\sigma_N^2 = \frac{1}{N}\sum_{\nu=1}^{N} y_\nu^2 - \left(\frac{1}{N}\sum_{\nu=1}^{N} y_\nu\right)^2$$

can be called the variance of the sample of the $y_1,...,y_N$. This is a realization of the random variable :

$$S_N^2 = \frac{1}{N}\sum_{\nu=1}^{N} Y_\nu^2 - \left(\frac{1}{N}\sum_{\nu=1}^{N} Y_\nu\right)^2$$

Of course, when $N \to +\infty$, $S_N \to 0$, both almost everywhere and in law ; this is simply due to the fact that $Y_\nu \to p$ for the a.e. convergence and follows from the Central Limit Theorem for the convergence in law.

Then one may take the following decision rule : when the variance of the sample $y_1,...,y_N$ is small enough, we consider that we have reached the limit with proper precision, and we take the estimate :

$$p = \frac{y_1 + \cdots + y_N}{N}.$$

The question is : is this decision rule correct, and what uncertainty does it give upon the estimate for p ? Or, in other words, how do we estimate the difference $\left| p - \frac{y_1 + \cdots + y_N}{N} \right|$?

Asymptotically, when $N \to +\infty$, this decision rule is correct, since each Y_ν follows a normal law with parameters $p, \frac{\sigma_0}{\sqrt{\nu}}$. Then the law of $\frac{Y_1 + \cdots + Y_N}{N}$ is explicit ; its expectation is p and the probability :

$$P\left\{ \left| \frac{Y_1 + \cdots + Y_N}{N} - p \right| > \varepsilon \right\}$$

can be computed explicitly. But all this is true only asymptotically !

If now we want to obtain a correct result for a given value of N, we may proceed as follows :

Proposition 9. – *If* $N \geq \frac{1}{(1-\alpha)\varepsilon^2 p}$, *within a confidence level* α, *we have the estimate :*

$$\frac{\overline{p}}{1+\varepsilon} \leq p \leq \frac{\overline{p}}{1-\varepsilon}.$$

Proof of Proposition 9.

We compute explicitly :

$$P\left\{ \left| Y_N - p \right| > \varepsilon p \right\}.$$

Since $Y_N = \frac{X_1 + \cdots + X_N}{N}$, and the X_i are independent r.v., we get :

$$\sigma(Y_N) = \frac{\sigma_0}{\sqrt{N}},$$

and therefore, by Bienaymé-Tchebycheff :

$$P\left\{\left|Y_N - p\right| > \varepsilon p\right\} \leq \frac{\sigma_0^2}{N\varepsilon^2 p^2} = \frac{1-p}{N\varepsilon^2 p}.$$

For given $\varepsilon > 0$, this probability is small when N is large enough. Therefore, for most realizations of Y_N, the observed value must be close to p.

Let $\overline{p} = \frac{x_1 + \cdots + x_N}{N}$ be the observed realization.

If $\left|p - \overline{p}\right| > \varepsilon p$, we would have a realization of Y_N far from p (since this realization is on $\overline{p}$), which is very unlikely : this has probability at most $\frac{1-p}{N\varepsilon^2 p}$. So we see that :

$$\frac{\overline{p}}{1+\varepsilon} \leq p \leq \frac{\overline{p}}{1-\varepsilon}$$

with probability at least $1 - \frac{1-p}{N\varepsilon^2 p}$. If we want a confidence level α (say $\alpha = 0.95$), then we get the estimate :

$$N \geq \frac{1}{(1-\alpha)\varepsilon^2 p} \tag{5.1}$$

We observe that p appears in this estimate : one needs to know at least a lower bound for p in order to apply it (for instance, one must know that $p \geq 0.1$). This proves our Proposition.

Now, if we use $Y_1, ..., Y_N$ (and not just Y_N), we will improve upon this proposition, as follows :

Proposition 10. – *If :*

$$\frac{2}{N} - \frac{\left(Log(N)\right)^2}{2N^2} \leq \varepsilon^2 p(1-\alpha)$$

then, with a confidence level α, we have the estimate :

$$\frac{\overline{p}}{1+\varepsilon} \leq p \leq \frac{\overline{p}}{1-\varepsilon}.$$

Proof of Proposition 10.

We compute :

$$P\left\{\left|\frac{Y_1 + \cdots + Y_N}{N} - p\right| > \varepsilon p\right\}$$

First, we need to compute $\sigma^2\left(\frac{Y_1+\cdots+Y_N}{N}\right)$. We have :

$$\begin{aligned} Y_1+\cdots+Y_N &= \sum_{\nu=1}^{N} \frac{1}{\nu}\sum_{j=1}^{\nu} X_j \\ &= \sum_{j=1}^{N}\left(\sum_{\nu=j}^{N}\frac{1}{\nu}\right)X_j \end{aligned}$$

and therefore :

$$\frac{1}{N^2}\sigma^2\left(Y_1+\cdots+Y_N\right) = \frac{\sigma_0^2}{N^2}\sum_{j=1}^{N}\left(\sum_{\nu=j}^{N}\frac{1}{\nu}\right)^2$$

Set $u(N)=\sum_{j=1}^{N}\left(\sum_{\nu=j}^{N}\frac{1}{\nu}\right)^2$. Then, Bienaymé-Tchebycheff's inequality gives :

$$P\left\{\left|\frac{Y_1+\cdots+Y_N}{N}-p\right|>\varepsilon p\right\}\le\frac{u(N)\sigma_0^2}{N^2\varepsilon^2 p^2}$$

Let now

Let $\overline{p}=\frac{y_1+\cdots+y_N}{N}$ be now the observed realization of $\frac{Y_1+\cdots+Y_N}{N}$. The same reasoning as above shows that :

$$\frac{\overline{p}}{1+\varepsilon}\le p\le\frac{\overline{p}}{1-\varepsilon}$$

with probability $1-\frac{u(N)(1-p)}{N^2\varepsilon^2 p}$. If we want a confidence level α, we get the estimate :

$$\frac{u(N)}{N^2}\le\varepsilon^2 p(1-\alpha)$$

Using a Taylor expansion of $u(N)$, we obtain the sufficient condition :

$$\frac{2}{N}-\frac{\left(Log(N)\right)^2}{2N^2}\le\varepsilon^2 p(1-\alpha) \qquad (5.2)$$

which is better than (5.1). Of course, it still contains the parameter p. This concludes the proof.

References

[IRSN-SCM] Anne-Laure Weber, Anne Karcher, Nicolas Pépin, Frédéric Huynh, Pierre Funk (IRSN), Laure Le Brize, Olga Zeydina, Bernard Beauzamy (SCM SA), Implementation of an experimental design to evaluate the codes used to determine the enrichment of uranium samples. Paper presented by IRSN and SCM at the *"European Safeguards Research and Development Association"* meeting, May 2007.

[BB1] Bernard Beauzamy : Méthodes probabilistes pour l'étude des phénomènes réels, ISBN : 2-9521458-0-6, Editions de la SCM, mars 2004.

[BBOZ] Bernard Beauzamy et Olga Zeydina : Méthodes probabilistes pour la reconstruction de données manquantes,ISBN : 2-9521458-2-2, Editions de la SCM, avril 2007.