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

Random sampling of proportions

Bernard Beauzamy
June 2008

From time to time, we find a problem in which we do not deal with values, but with proportions. For instance :

- A total budget has to be spent, and we want to study the repartition between several goods ;
- A certain amount of pollution is detected, and we want to study the repartition between several factors.

In order to find interesting or dangerous situations, one usually tries to simulate various possibilities. Since no information is known a priori about the factors, the law used for such a simulation should be a uniform law.

1. Description of the problem

Mathematically speaking, this means :

Simulate random variables $X_1,...,X_K$ such that :

- each $X_k \geq 0$
- $\sum_{k=1}^{K} X_k = 1$

and the law of the vector $(X_1,...,X_K)$ is the uniform law on the set :

$$C_K = \left\{ (x_1,...,x_K);\ x_k \geq 0\ \forall k, \sum_{k=1}^{K} x_k = 1 \right\}$$

Uniform law means that, for a sample, the number of times we fall into a subset $A \subset C_K$ depends only on the measure of A. It means also that the joint density of the vector $(X_1,...,X_K)$ is constant in the set C_K, that is, does not depend on the particular values of $(x_1,...,x_K)$ in C_K.

The discrete version of the uniform law is much easier to understand. Fix a denominator, say M. We want positive integers $n_1,...,n_K$ such that $n_1+\cdots+n_K=M$, so our proportions are $\frac{n_1}{M},...,\frac{n_K}{M}$, and we want that all repartitions have the same probability, independently of the values of $n_1,...,n_K$.

For instance, take $K=3$, $M=4$. We have the possibilities :

$\left(\frac{4}{4},0,0\right)$,

$\left(\frac{3}{4},\frac{1}{4},0\right),\left(\frac{3}{4},0,\frac{1}{4}\right)$

$\left(\frac{2}{4},\frac{2}{4},0\right),\left(\frac{2}{4},\frac{1}{4},\frac{1}{4}\right),\left(\frac{2}{4},0,\frac{2}{4}\right)$

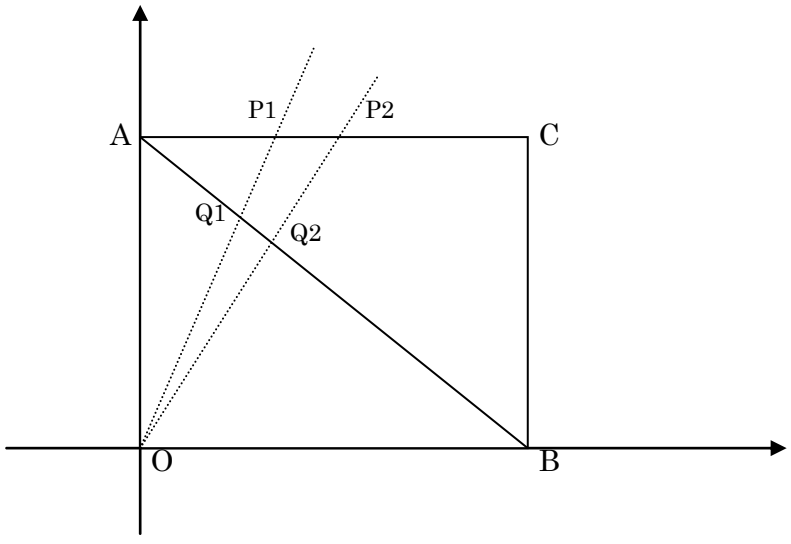
$\left(\frac{1}{4},\frac{3}{4},0\right),\left(\frac{1}{4},\frac{2}{4},\frac{1}{4}\right),\left(\frac{1}{4},\frac{1}{4},\frac{2}{4}\right),\left(\frac{1}{4},0,\frac{3}{4}\right)$

$\left(0,\frac{4}{4},0\right),\left(0,\frac{3}{4},\frac{1}{4}\right),\left(0,\frac{2}{4},\frac{2}{4}\right),\left(0,\frac{1}{4},\frac{3}{4}\right),\left(0,0,\frac{4}{4}\right)$

and each of these 15 possibilities should have probability 1/15.

2. Warning

We have the same warning as previously : to generate uniform variables X_k on $[0,1]$ and then replace them by $\frac{X_k}{\sum_{j=1}^{K} X_j}$ is wrong, because we do not obtain this way a uniform law on C_K. This is clear on the following example (dimension 2) :



Assume we take random variables (X_1, X_2) with uniform law on the segment $[0,1]$. Then, the number of times the point with coordinates (X_1, X_2) falls into the triangle OP_1P_2 is proportional to the area of that triangle. But the normalized point $\frac{X_1}{X_1+X_2}, \frac{X_2}{X_1+X_2}$ is in the segment $Q_1 - Q_2$ if and only if the point (X_1, X_2) is in the triangle OP_1P_2. But the area of the triangle depends upon its orientation : for given length of $Q_1 - Q_2$, we have a larger triangle if it contains the point C, and smaller at the extremities, when the triangle contains A or B. So this construction does not lead to a uniform law on the segment AB.

3. Generation of uniform variables on C_K

The correct construction is as follows. We follow the book "Non Uniform Random Variate Generation", by Luc Devroye, chapter 5, theorem 2.2.

Theorem. - *Let $X_1, ..., X_K$ be independent random variables with exponential law. Then, the variables $Y_1, ..., Y_K$ defined by :*

$$Y_1 = \frac{X_1}{X_1 + \cdots + X_K}, \ldots, Y_K = \frac{X_K}{X_1 + \cdots + X_K}$$

are positive, have sum 1, and follow a uniform law on C_K.

Note : in order to generate random variables with exponential law, on should generate random variables with uniform law Y_k and then take $X_k = Log\frac{1}{Y_k}$.

So, assume you want to generate a sample of $N = 1000$ proportions for $K = 40$ goods, do the following :

for $n = 1$ to N
for $k = 1$ to K
generate $Y_1, ..., Y_K$ independent variables with uniform law on $[0,1]$
take $X_k = Log\frac{1}{Y_k}$, for $k = 1, ..., K$
compute $\sum_{j=1}^{K} X_j$
replace each X_k by $\frac{X_k}{\sum_{j=1}^{K} X_j}$
next k
next n

Proof of the Theorem. Let $S = X_1 + \cdots + X_K$. Let $Z = (X_1, ..., X_{K-1})$, and $z = (x_1, ..., x_{K-1})$, to simplify the notation.

Let $f_{Z,S}(z,s)$ be the density of the joint law of the $K-$uple (Z,S). This density can be computed as a conditional probability, knowing Z, that is :

$$f_{Z,S}(z,s) = f_{S|Z}(s) \times f_Z(z) \tag{1}$$

where $f_{S|Z}(s)$ denotes the conditional probability density of S knowing Z and $f_Z(z)$ is the density of Z.

Since the variables $X_1, ..., X_{K-1}$ are independent and follow an exponential law, we have :

$$f_Z(z) = e^{-(x_1 + \cdots + x_{K-1})} \tag{2}$$

The law of S knowing $X_1 = x_1, ..., X_{K-1} = x_{K-1}$ is easy to find. Indeed,

$$P\{S \le a \mid X_1 = x_1, ..., X_{K-1} = x_{K-1}\} = P\{X_K \le a - x_1 - ... - x_{K-1} \mid X_1 = x_1, ..., X_{K-1} = x_{K-1}\}$$

and therefore :

$$f_{S|Z}(s) = e^{-s + x_1 + \cdots + x_{K-1}} \tag{3}$$

for $s \ge x_1 + \cdots + x_{K-1}$, 0 otherwise.

We deduce from (1), (2), (3) that :

$$f_{Z,S}(z,s) = e^{-s} \tag{4}$$

for $s \ge x_1 + \cdots + x_{K-1}$, 0 otherwise.

Now, we compute the joint density of $\left(\frac{X_1}{S}, ..., \frac{X_{K-1}}{S}, S\right)$. This is obtained by a change of variable in the density (4). We set :

$$y_1 = \frac{x_1}{s}, ..., y_{K-1} = \frac{x_{K-1}}{s}, s = s .$$

Let $U = \left(\frac{X_1}{S}, ..., \frac{X_{K-1}}{S}\right)$; we get :

$$f_{U,S}(u,s) = s^K e^{-s} \tag{5}$$

The density of U can be obtained from the above formula, integrating with respect to s. We get :

$$f_U(u) = \int_0^{+\infty} f_{U,S}(u,s)ds = \int_0^{+\infty} s^K e^{-s} ds = K! \tag{6}$$

and we see that this value is constant on the whole set $x_1 \geq 0, ..., x_{K-1} \geq 0, x_1 + \cdots + x_{K-1} \leq 1$.

But now, since the density of $\frac{X_1}{S}, ..., \frac{X_{K-1}}{S}$ is constant, so is the density of

$$\left(\frac{X_1}{S}, ..., \frac{X_{K-1}}{S}, 1 - \frac{X_1}{S} - ... - \frac{X_{K-1}}{S} \right) = \left(\frac{X_1}{S}, ..., \frac{X_{K-1}}{S}, \frac{X_K}{S} \right) \tag{7}$$

This concludes the proof of the Theorem.

Remark

A package in Matlab, due to Roger Stafford, generates random numbers with uniform law, and fixed sum : it can be found on the Interned under the name of randfixedsum.m.

This package uses another approach, more complicated than the approach described here (using exponential variables), but it can solve more general problems : find x_i with $a \leq x_i \leq b$ and $\sum_{i=1}^{n} x_i = S$.

Acknowledgements

We thank Paul Deheuvels for indicating the book by Luc Devroye, Luc Devroye for his comments about the proofs, and Roger Stafford for his comments about the Matlab package.