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

Random sampling according to a given discrete law

Bernard Beauzamy

November 19th, 2008

In this note, we indicate how to construct a sample from a given probability law : this is quite simple. But conversely, how well does this sample approximate the probability law ? The answer is : much less than most people think !

I. Constructing a sample from a given law

Assume we are given some values $x_1,...,x_N$ with a discrete probability law $p_1,...,p_N$. We want to create a sample, on the computer, from these values using this probability law. The simplest is to use the inverse repartition function, as follows (our example takes 5 points into account, just to have small tables).

values	proba	cumul. proba
x1	p1	p1
x2	p2	p1+p2
x3	p3	p1+p2+p3
x4	p4	p1+p2+p3+p4
x5	p5	1

Now, you choose a number z at random, with uniform law between 0 and 1 (there is a function in Excel for that). If :

$0 \le z < p_1$ choose x_1
$p_1 \le z < p_1 + p_2$ choose x_2
$p_1 + p_2 \le z < p_1 + p_2 + p_3$ choose x_3
$p_1 + p_2 + p_3 \le z < p_1 + p_2 + p_3 + p_4$ choose x_4
$p_1 + p_2 + p_3 + p_4 \le z \le 1$ choose x_5

Here is a simple VBA code which does this. Let $q(j) = \sum_{i=1}^{j} p_i$ be the cumulated probabilities of column 3 in the table above. Then :

```
dim z as double
For j = 1 To Ntot 'number of points in your sample
```

```
If q(j) > z Then
GoTo finfin
End If
Next j
finfin:
ech(j) = ech(j) + 1
'the array ech() will contain the number of times each point has been drawn
Next n
```

II. Reconstructing the probability law from the sample

A. Numerical evidences

This sampling process (Monte-Carlo method) will allow us to reconstruct the original law: for each j, if n_j is the number of times where j appears, then $\frac{n_j}{N} \to p_j$. However, this convergence is very slow, as the following example shows. We have only 5 classes, the probabilities of which are given in the first column. Then the reconstructions using 10^4 to 10^7 points are given in the columns 3, 4, 5, 6 :

	proba	cumul	rec /10^4	rec /10^5	rec /10^6	rec /10^7
1	0,30962318	0,30962318	0,3007	0,30594	0,310159	0,3098193
2	0,26192854	0,57155172	0,2305	0,2446	0,261723	0,2619889
3	0,33920459	0,9107563	0,3711	0,35176	0,33935	0,3394784
4	0,02310692	0,93386323	0,0273	0,02475	0,023248	0,0229469
5	0,06613677	1	0,0704	0,07295	0,06552	0,0657665

The theory will indicate why the convergence is slow. Assume for instance that (as here), we have 5 classes and that the probability of the first is 0.3. Assume we want to reconstruct this first probability with an accuracy of 1/100 (that is, we want to be in the interval $\frac{99p_1}{100}, \frac{101p_1}{100}$). What size of sample do we need ?

The probability to choose the first class k times among N is given by a binomial law :

$$P\{X_1 = k\} = \binom{N}{k} p_1^k (1-p_1)^{N-k}$$

We want to compute the probability that $(1-\varepsilon)p_1 \le \frac{k}{N} \le (1+\varepsilon)p_1$, with $\varepsilon = 1/100$. This probability is simply given by the sum :

$$S = \sum_{k=(1-\varepsilon)p_1 N}^{(1+\varepsilon)p_1 N} \binom{N}{k} p_1^k (1-p_1)^{N-k}$$

Direct computation of this sum can be made with Maple, for various values of N. We find :

N	S
1 000	0.19
10 000	0.49
100 000	0.96

So we see that it will take a sample of size 100 000 to reconstruct correctly the interval of probability 0.3.

If we now take p_1 smaller, the results are of course worse. For instance, for $p_1 = 0.1$, we get :

N	S
1 000	0.13
10 000	0.27
100 000	0.71

This last result means the following : take $p_1 = 0.1$ and draw a sample of size 100 000. Then you have only 71 chances /100 that your sample will satisfy the property $(1-\varepsilon)\, p_1 \le \frac{n_1}{N} \le (1+\varepsilon)\, p_1$, with $\varepsilon = 0.01$. With $N = 10\,000$, you have only 27 chances /100, which means that you should not expect that your sample will satisfy the precision property $0.099 \le \frac{n_1}{N} \le 0.101$. In other words, to satisfy a precision property of 1/100, a size of sample of 10 000 is clearly insufficient.

B. Theory behind these results

A theorem proved by Upensky (Introduction to Mathematical Probability, New York, 1937) states that :

$$P\left\{ p(1-\varepsilon) \le \frac{X}{N} \le p(1+\varepsilon) \right\} \ge 1 - 2\exp\left\{ -N\left(\varepsilon p\right)^2 / 2 \right\}$$

For $p = 0.1$ and $\varepsilon = 0.01$, as above, the right-hand side will be ≥ 0.95 if :

$$-N(\varepsilon p)^2 / 2 \le Log\left(0.025\right)$$

which gives $N \ge 7\,400\,000$.

Slight improvements were given by Masashi Okamoto "Some inequalities relating to the partial sum of binomial probabilities", Annals of the Institute of Statistical Mathematics, vol. 10, Number 1, March 1959, pp. 29-35.