Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

Génération d'échantillons aléatoires :

applications et mises en garde

par Bernard Beauzamy

novembre 2025

Résumé

Générer un échantillon aléatoire à partir d'une loi de probabilité donnée est une préoccupation très ancienne ; elle se rencontre en particulier auprès de toutes les institutions qui veulent réaliser des simulations ; voici quelques exemples :

- Générer un catalogue de pannes, pour un équipement donné, à partir d'un historique des pannes ;
- Générer un catalogue de cyclones, de séismes, pour une région donnée, à partir d'un historique ;
- Générer un modèle de propagation d'une épidémie, à partir d'un historique.

Bien évidemment, ces simulations ne sont pas des certitudes ; elles n'ont pas de valeur juridique, mais servent à éclairer la décision : ce sont des scénarios. Habituellement, on cherche à se doter de plusieurs scénarios, pour mieux éclairer la question.

Nous montrons ici comment générer un échantillon de taille quelconque, à partir d'un échantillon recueilli, en respectant la loi de l'échantillon recueilli. Nous donnons un exemple concret d'application : à partir de 170 années de températures journalières recueillies, générer un scénario sur 100 000 jours (soit environ 274 ans) en faisant apparaître explicitement les dates où la température dépassera 39°C.

Mais des mises en garde s'imposent : si, dans l'échantillon d'origine, le capteur n'a pas fonctionné la moitié du temps, la valeur 0 aura probabilité 1/2 dans l'échantillon généré. S'il s'agit de données de température, le fait d'avoir plusieurs jours de suite avec canicule sera perdu. En bref, la génération aléatoire suppose que les données sont indépendantes ; si ce n'est pas le cas, danger.

I. Position du problème

Générer un échantillon à partir d'une loi de probabilité donnée a fait l'objet d'une littérature très abondante. Je crois que le meilleur document est le livre :
Luc Devroye : Non Uniform Random Variate Generation,
qui peut être téléchargé librement :
https://luc.devroye.org/LucDevroye-NonUniformRandomVariateGeneration-10.1007_978-1-4613-8643-8-1986-.pdf

Si la loi de probabilité est donnée par une densité continue $f(x)$, avec fonction de répartition $F(x)$, on génère un échantillon selon cette loi de manière très simple :

- On génère un nombre aléatoire y, selon une loi uniforme sur $[0,1]$: c'est l'expression $y = rnd(\ )$ en VBA.
- On cherche le x tel que $F(x) = y$.

Il faut donc inverser la fonction F ; c'est difficile si elle est continue ; encore plus difficile si la fonction f est une loi discrète, car alors F est constante par paliers. On est donc amené à toutes sortes de "bricolages" : soit approximer F par une fonction inversible connue, soit tester des valeurs de x jusqu'à ce qu'on en trouve une telle que $F(x) = y$. La littérature sur ces approches est abondante et peu satisfaisante en pratique : on s'égare aux rives du Gange, comme aurait dit Chateaubriand.

En pratique, précisément, ni la fonction f ni la fonction F ne sont connues : nous ne disposons que d'un échantillon de mesures. Vouloir en déduire une loi de probabilité ne peut se faire sans hypothèses, souvent largement factices. Les économistes, les spécialistes des risques au sein des banques et des compagnies d'assurance, le constatent souvent à leurs dépens, bien entendu sans jamais admettre une mauvaise approche du sujet.

En pratique, donc, il ne faut pas vouloir passer par l'intermédiaire d'une loi de probabilité, il faut générer un échantillon aléatoire à partir de l'échantillon recueilli, en conservant la même loi. Si, dans l'échantillon recueilli, les valeurs sont $x_1 < x_2 < \cdots < x_K$ avec chacune un nombre d'occurrences $n_1, ..., n_K$, on veut que, dans l'échantillon généré, on retrouve les mêmes valeurs, avec la même probabilité, à savoir $p_k = \dfrac{n_k}{n_1 + \cdots + n_K}$.

Ceci se fait très simplement en choisissant au hasard, selon une loi uniforme, un nombre quelconque dans l'échantillon recueilli. Celui-ci doit être laissé tel qu'il est, tel que la Nature nous le fournit. Le programme ci-dessous, en VBA, explique ceci très clairement.

II. Programmation informatique

Voici le programme en VBA sous Excel :

```
Option Explicit
Sub macro1()
Dim DL As Long
DL = Sheets(2).Cells(Rows.Count, 1).End(xlUp).Row
'la dernière cellule occupée par l'échantillon recueilli
'(mis sur la seconde feuille, colonne 1)
Sheets(1).Cells(11, 4) = DL-1
'la taille de l'échantillon recueilli
Dim TEV As Long 'taille échantillon voulu
TEV = Sheets(1).Cells(5, 4)
'lecture à partir de l'indication donnée par l'utilisateur, qui met ce qu'il veut
Dim k As Long
Randomize
Dim n As Long
Dim data As Variant 'mise en mémoire de l'échantillon recueilli
'cela va beaucoup plus vite que l'accès aux cellules
data = Sheets(2).Cells(2, 1).Resize(DL, 1)
For k = 1 To TEV
n = Int((DL - 1) * Rnd()) + 1
Sheets(2).Cells(k + 1, 3) = data(n, 1)
Next k
Dim FF As Long
FF = Sheets(2).Cells(Rows.Count, 3).End(xlUp).Row
Sheets(1).Cells(12, 4) = FF - 1
End Sub
```

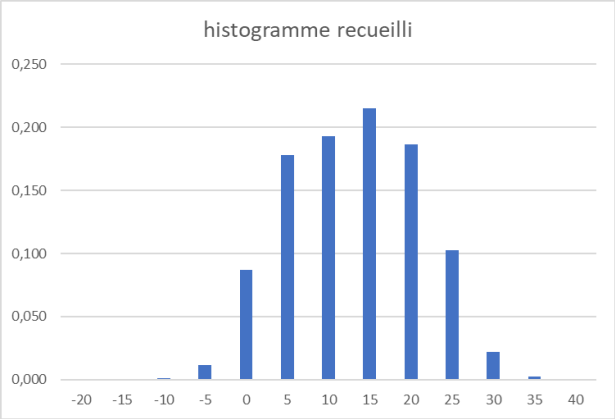
Revenons sur l'instruction n = Int((DL - 1) * Rnd()) + 1
qui est l'élément central du programme.

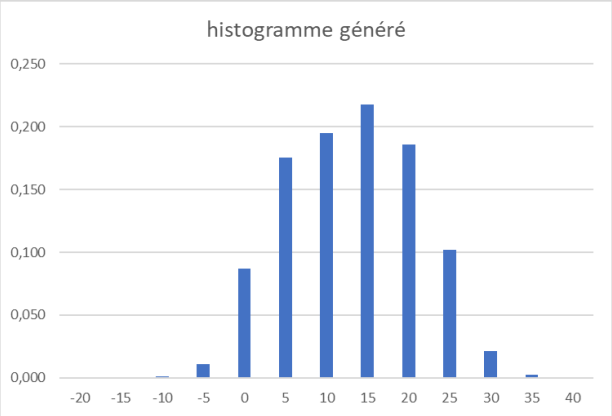
DL-1 est la taille de l'échantillon recueilli. Rnd() prend des valeurs arbitraires, selon une loi uniforme, entre 0 et 1. Par conséquent, (DL - 1) * Rnd() prend des valeurs arbitraires, selon une loi uniforme, entre 0 et DL-1 : mais ce sont des valeurs réelles, et non des entiers.

L'instruction Int(x) renvoie le plus grand entier inférieur ou égal à x. Par conséquent,
Int((DL - 1) * Rnd()) prend des valeurs arbitraires, entières, avec loi uniforme, entre 0 et DL-1.
Enfin, Int((DL - 1) * Rnd()) + 1 prend des valeurs arbitraires, entières, avec loi uniforme, entre 1 et DL : c'est ce que nous voulons. Vérifions : si rnd()=0, on trouve 1, si rnd()=1, on trouve DL.

III. Analyse d'un exemple

On peut évidemment construire les histogrammes, pour l'échantillon recueilli et l'échantillon généré. Voici un exemple ; il concerne des températures (en °C) ; on a recueilli 36 214 données et on en a généré 100 000, suivant la même loi.





Les deux ont bien même apparence.

L'échantillon généré possède les mêmes propriétés statistiques que l'échantillon recueilli, mais les nombres ne sont pas dans le même ordre. Voici un exemple (même données que précédemment) :

échantillon recueilli		échantillon généré	
17,0		12,50	
22,6		16,70	
24,6		21,10	
24,6		10,40	
24,6		12,20	
27,2		23,40	
26,3		6,13	
27,2		18,19	
26,3		17,90	

IV. Un exemple de scénario

Nous allons voir sur un exemple comment réaliser un scénario : nous partons des données précédentes. Ce sont des données journalières de température, au nombre de 36 214, soit environ 137 ans. On veut construire un scénario de températures extrêmes sur 100 000 jours, soit environ 274 ans. On s'intéresse au retour de températures très élevées, mettons 39°C. Le scénario généré donne la liste suivante, pour les jours où cette température sera atteinte ou dépassée :

3838, 7342, 9878, 13568, 20992, 39618, 43538, 49652, 52767, 65555, 72271, 72569, 80844, 82148, 96399.

La première occurrence se manifestera dans 3 838 jours, soit approximativement 10 ans et 188 jours, la seconde dans 7 342 jours, soit approximativement 20 ans et 42 jours, etc.

V. Mise en garde

Il faut prendre garde à la nature des données générées. L'outil décrit ici traite les données comme si elles étaient indépendantes et génère un nouvel échantillon à partir du précédent.

Nous allons voir deux exemples illustrant cette difficulté.

1. Un exemple simple

Il s'agit de températures ; on dispose de 20 données, mais le capteur n'a pas fonctionné la moitié du temps :

échantillon recueilli		échantillon généré	
0		27,20	
0		26,27	
0		0,00	
0		0,00	
0		21,77	
0		0,00	
0		24,58	
0		0,00	
0		22,59	
17,0		0,00	
22,6		17,02	
24,6		24,58	
24,6		17,02	
24,6		26,27	
27,2		0,00	
26,3		24,58	
27,2		24,58	
26,3		27,20	
18,9		0,00	
21,8		24,58	

Il y a bien des zéros dans l'échantillon généré, mais ils sont répartis n'importe où et l'information "le capteur n'a pas fonctionné la moitié du temps" est perdue.

2. L'exemple des températures extrêmes

De même, pour les températures extrêmes, les données ne sont pas indépendantes : il est fréquent qu'il fasse très chaud plusieurs jours de suite. Les données ci-dessous sont celles vues précédemment ; elles sont extraites d'un recueil de températures journalières pour une ville en France, tous les jours entre 1921 et 2021. On s'intéresse à la situation où la température dépasse 39°C. Les données extraites sont :

29492	06/08/2003	39,5
29493	07/08/2003	39,8
29494	08/08/2003	39,9
29498	12/08/2003	40,6
32793	19/08/2012	39,7
35323	24/07/2019	40,4
35324	25/07/2019	41,8
35696	31/07/2020	39,9

On a deux manières de compter :

- On compte tous les jours, sur les 36 214 données disponibles, où la température dépasse le seuil. La probabilité sera donc approximativement 8/36 214 =0,00022. On se dit que c'est extrêmement faible : seulement 8 jours en 100 ans !

- On compte par années : il y a 4 années où il a fait très chaud, à savoir 2003 (4 jours), 2012 (un jour), 2019 (4 jours) et 2020 (un jour). Là, on se dit qu'il y a 4 années chaudes en cent ans, donc la probabilité est 4/100=0,04, ce qui est bien supérieur au précédent. De plus, la durée de retour moyenne d'une année chaude est de 20 ans (répartir 4 points sur un intervalle de longueur 100, tous les intervalles ayant même longueur). Là, on commence à s'inquiéter !

En réalité, "année chaude" devrait être traité différemment selon qu'il y a un, deux, trois, quatre, jours dépassant le seuil. En définitive, il n'y a qu'une seule occurrence, en cent ans, où l'on ait vu 4 jours dépassant le seuil (2003).

Il faut faire attention, si on génère des scénarios de température à partir des données recueillies : le fait d'avoir plusieurs jours consécutifs est perdu.

VI. En conclusion

La méthode présentée ci-dessus permet très facilement de générer des scénarios à partir d'un échantillon recueilli, en respectant la loi de cet échantillon : on ne triche en aucune manière et on ne fait aucune hypothèse factice de loi.

Par contre, et ceci est très important, la méthode suppose que les relevés sont indépendants, ce qui se répercute sur l'échantillon généré. Si les données ne sont pas indépendantes à l'origine (la même donnée répétée 300 fois), cette information est perdue lors de la génération, ce qui peut donner lieu à des erreurs graves.

Il est donc recommandé de s'interroger sur ce que représentent les données recueillies et de ne pas les traiter à l'aveugle.