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

La génération de nombres aléatoires :

est-elle utile ? est-elle possible ?

par Bernard Beauzamy
PDG, SCM SA

septembre 2013

On rencontre assez souvent, dans la vie courante comme dans la recherche scientifique, des situations où la génération de nombres dits "aléatoires" est présentée comme nécessaire. Nous allons voir ce que cela signifie, quel est le but recherché et quelles sont les difficultés. Je prendrai cet exemple, parce qu'il me paraît représentatif : on est au confluent d'un besoin social, mal défini et en définitive assez simple, d'un besoin formulé par d'autres disciplines scientifiques (la physique notamment) et de la recherche mathématique. Cette situation se rencontre souvent.

1. Les jeux

La situation la plus connue est celle des jeux, comme par exemple le loto. Le joueur désigne 6 nombres parmi 49, mise une certaine somme sur ces choix et la Française des Jeux (ou l'un de ses concurrents) annonce les numéros qui sortent. Le joueur gagne s'il a pronostiqué correctement. Les résultats sont, en principe, issus d'un "générateur de nombres aléatoires", qui tire des numéros parmi 49.

Que réclame le peuple des joueurs ? Tout d'abord l'équité, qui veut que tous les nombres aient la même chance de sortir à un tirage donné, ce que l'on nomme "équiprobabilité". Mais personne ne peut vérifier cela, pour un tirage donné. Ensuite, l'égalité des proportions, sur tout l'historique des tirages : personne n'accepterait par exemple que le chiffre 3 sorte moins que les autres. Enfin, il ne doit pas exister de "martingale" (ce mot étant pris dans le sens usuel) qui permette de pronostiquer un résultat ; la connaissance complète de l'historique n'apporte aucune information sur le tirage suivant.

Remarquons bien que les tirages peuvent individuellement être faussés (non équiprobables) sans que cela se voie sur l'historique : il suffit que le défaut "tourne". La première condition disparaît ; seules les deux dernières demeurent.

Pour satisfaire le joueur, j'aurai besoin d'un générateur de nombres aléatoires assez primitif : il doit me donner six nombres parmi 49, plus ou moins avec équiprobabilité, et oublier d'un jour sur l'autre ce qu'il a fait la veille. En général, on partira d'un générateur d'entiers (de 0 à 9) et on en déduira les six nombres voulus, par une méthode quelconque, que j'appellerai le "lien".

Voici un exemple quant à la manière de procéder : j'écris les décimales de π et je range ces décimales par groupes de 2 ; chaque groupe est donc un nombre entre 00 et 99. A un groupe j'associe un nombre de 1 à 49 d'une manière quelconque ; par exemple : si 00 je ne prends pas, de 01 à 49 je prends tel quel, si 50, je ne prends pas, de 51 à 99, je remplace par 100 – x. Si par malchance un même nombre apparaît deux fois parmi les 6, j'en tire un de plus. Je puis évidemment commencer n'importe où dans la liste des décimales de π et je puis définir le lien de multiples manières. Je puis aussi utiliser $1/\pi$, $\sqrt{\pi}$, etc.

Les joueurs n'ont accès qu'aux résultats des tirages (des listes de nombres entre 1 et 49) ; même si, à cause d'une fuite, ils venaient à savoir que les décimales de π sont utilisées, ils n'auraient aucun moyen de pronostiquer les résultats, sauf à connaître le lien. Nous avons donc ainsi un procédé simple, honnête et fiable, répondant aux attentes. Il ne nécessite aucune recherche complémentaire.

Les séries n'étant pas très longues, les contraintes des jeux, en ce qui concerne les générateurs de nombres aléatoires, ne sont pas très fortes.

2. Les méthodes de Monte-Carlo

Une autre application des nombres aléatoires se rencontre dans les méthodes dites de "Monte-Carlo". Rappelons que si f est une fonction, par exemple continue entre 0 et 1, si x_n sont des points tirés selon une loi uniforme sur l'intervalle [0,1], la moyenne $\frac{1}{N}\sum_{n=1}^{N} f\left(x_n\right)$ converge vers $\int_0^1 f\left(t\right)dt$ lorsque $N \to +\infty$. Ceci est vrai en n'importe quelle dimension, et la vitesse de convergence est indépendante de la dimension. En dimension 1, il vaut mieux utiliser des points équirépartis, du type k/N, pour évaluer l'intégrale, mais cela devient impossible en dimension élevée. Si on travaille en dimension 40, et que l'on découpe chaque intervalle $\left[0,1\right]$ en dix segments, cela conduit à 10^{40} points : hors de portée de n'importe quel ordinateur. L'idée selon laquelle quelques milliards de points, issus du hasard, vont suffire est donc tentante et même fascinante : nous ne savons pas calculer, laissons faire le hasard.

Il y a tout de même de sérieuses restrictions. Les méthodes de Monte-Carlo peuvent être utilisées pour évaluer une intégrale, mais non pour rechercher des zones à risque. Prenons encore la situation d'un code de calcul dépendant de 40 paramètres, et imaginons que la zone à risque (celle qui va faire exploser la centrale !) soit celle où tous les paramètres sont $<1/2$. La me-

sure de cette zone est donc $\frac{1}{2^{40}}$; en lançant quelques milliards de points, vous n'avez aucune chance de la détecter.

Même pour calculer une intégrale, certaines précautions doivent être prises. Il faut que les points utilisés soient "bien répartis" et la définition de ceci n'est pas claire. En théorie, cela signifie que chaque sous-ensemble de l'espace reçoit un nombre de points qui ne dépend que de la mesure de l'ensemble. Par exemple, tous les intervalles de $[0,1]$ de taille 1/10 doivent recevoir le même nombre de points. Mais une telle propriété ne peut être satisfaite par un nombre fini de points : elle ne peut être qu'asymptotique, lorsque le nombre de tirages augmente indéfiniment. Comment donc définir le fait qu'une suite finie est bien répartie ? Il n'y a pas de définition précise. Par contre, il existe des algorithmes, plus ou moins performants, qui permettent d'obtenir de telles suites (voir en particulier le livre [Devroye]).

La question de la construction de telles suites est loin d'être simple si l'on travaille par exemple sur une sphère, ou même sur un simplexe. Dans nos travaux récents portant sur l'évaluation de probabilités de phénomènes extrêmes, nous avons utilisé une suite de points bien répartis sur le simplexe :

$$\left\{x_n \geq 0\,;\, x_1 \geq x_2 \geq \cdots, \sum x_n = 1\right\}$$

(Peter Robinson [Robinson], Luc Devroye [Devroye]). Voir le livre [NMP] pour les erreurs à ne pas commettre dans la construction de telles suites.

L'outil de base est toujours, ici, une suite de nombres bien répartis dans l'intervalle $[0,1]$; c'est cette propriété qui est exploitée pour le calcul de l'intégrale. Peu importe dans quel ordre ils ont été obtenus. Quant à la convergence vers l'intégrale, on la vérifie de manière empirique : on tire des points et on évalue la moyenne ; on s'arrête lorsque la moyenne ne varie presque plus. Evidemment, le procédé n'est pas absolument répétable : une autre suite donnera un autre résultat, un peu différent. Ce n'est pas très important en pratique : les incertitudes sur les données d'entrée sont généralement telles que ces petites différences n'ont aucune importance.

Ici, à la différence des jeux, nous ne sommes tenus par aucun secret, et une suite (x_n) avec de bonnes propriétés peut être connue de tous et utilisée par tous dans tous les calculs.

Là encore, on peut partir de la suite des décimales de π ; si on veut des nombres avec 16 décimales, on groupe les décimales par blocs de 16 et on met 0. devant chaque bloc. On obtiendra simplement et à peu de frais une suite de nombre équirépartis dans l'intervalle $[0,1]$.

On voit, sur ces deux exemples, que l'on n'utilise pas réellement toutes les propriétés que l'on est en droit d'attendre d'un générateur de nombres aléatoires. La suite des décimales de π suffit, bien qu'elle n'ait rien d'aléatoire !

3. Que peut-on attendre ?

En théorie, et voici enfin la définition, on voudrait générer une suite de 0 et de 1 de telle sorte que :

– à chaque étape, 0 et 1 aient la même probabilité de sortir ;

– les étapes sont indépendantes, au sens probabiliste du mot ; plus précisément, la connaissance de toutes les étapes précédentes ne donne aucune information sur la suivante. Nous dirons que le générateur est "sans mémoire".

Si on parvenait à réaliser ceci avec deux valeurs, 0 et 1, on parviendrait à donner des nombres quelconques dans n'importe quel intervalle. Concentrons-nous donc sur ce problème, en apparence très simple.

La question est d'abord dans la génération de deux nombres (0 ou 1) avec égale probabilité : ceci a un sens mathématique clair, mais n'a aucun sens physique ou expérimental. On ne constate la probabilité que par la répétition, et en fait de manière asymptotique, lorsque le nombre de répétitions tend vers l'infini ! Il est donc tout à fait impossible de démontrer qu'un générateur, à un instant donné, agit de manière équiprobable.

La question est aussi de l'absence de mémoire : les résultats des tirages précédents ne doivent apporter aucune information sur le tirage suivant. Ceci est possible dans un univers axiomatique, mathématiquement idéal, mais est indémontrable en pratique.

En pratique, pour un générateur, on va utiliser un phénomène physique quelconque : le lancer d'une pièce, la date dans un ordinateur, un signal électromagnétique, ou tout ce que l'on voudra. Nous avons donc des observations de l'univers à des instants $1,2,...,N$ et, même si le phénomène oscille très vite, il est logiquement impossible de dire que cela ne donne aucune information sur l'état de l'univers à l'instant $N+1$; le second axiome demandé, à savoir l'absence de mémoire, ne sera jamais réalisé complètement : il peut l'être de manière approximative, et cela suffit généralement en pratique.

On peut espérer faire appel à la mécanique quantique, qui est fondamentalement probabiliste. Un essai en ce sens est fait dans l'article [Quantis], mais ce n'est absolument pas convaincant. On veut bien croire que, avec probabilité 1/2, un photon puisse traverser un miroir ou être réfracté, mais comment être assuré que :

– Cette probabilité 1/2 sera constante dans le temps ? En mécanique quantique, chaque observation modifie la probabilité du phénomène, et ici nous avons une suite d'expériences.

– Que les tirages successifs seront indépendants ? A force de recevoir des photons, le miroir peut prendre l'habitude de les laisser passer, ou l'habitude de les réfracter.

Nous n'en savons rien, mais nous notons que ces deux questions ne sont même pas abordées dans l'article en question.

Il existe bien sûr des tests pour vérifier si une suite est aléatoire ; nous les mentionnons pour les suites de -1 et +1 (l'énoncé est plus facile) :

- Les moyennes $\frac{1}{N}\sum_{1}^{N} x_n$ doivent suivre une loi normale centrée (théorème central limite) ;

- Les sommes partielles $\sum_{1}^{N} x_n$ doivent "osciller", c'est-à-dire prendre des valeurs négatives et des valeurs positives dont l'amplitude va croissant.

(Voir par exemple le livre [MPPR] pour des explications plus détaillées)

Mais ces conditions ne sont ni nécessaires ni suffisantes pour caractériser un "bon" générateur de nombres aléatoires. Elles ne sont satisfaites qu'asymptotiquement, et elles portent sur les sommes partielles : toute permutation des éléments d'une suite possédera les mêmes propriétés. Enfin, la suite des décimales de π y satisfait parfaitement, sans être le moins du monde aléatoire.

Laurent Schwartz expliquait que Hardy et Littlewood, deux mathématiciens britanniques du début du 20ème siècle, savaient reconnaître une suite réellement aléatoire d'une suite fabriquée par leurs collègues, mais il ajoutait qu'ils parvenaient à se tromper l'un l'autre.

On en arrive à ce premier paradoxe :

Paradoxe 1. - Prenons N quelconque fixé, mettons un milliard. On ne sait pas définir une suite "bien aléatoire" de longueur N. On est incapable de donner une définition qui puisse être utilisée par l'industrie, la justice, etc. (sur les implications judiciaires, on pourra consulter l'article [Erreurs]).

4. Quantité d'information

On admet généralement que plus une chose est laissée au hasard et moins elle contient d'information. Celle-ci, mathématiquement, se mesure au moyen d'une entropie ; l'entropie associée à une loi de probabilité p_i est $E = -\sum p_i Log\left(p_i\right)$. Plus l'entropie est grande, moins il y a d'information. Ainsi, si nous avons deux choix possibles avec probabilité 1/2, cela nous donne une entropie $E = Log\left(2\right)$; si l'une des valeurs est certaine, cela nous donne une entropie nulle.

Dans son roman "La voix du Maître" Stanislaw Lem ([Lem]) décrit une situation où un signal est reçu en provenance des étoiles ; on se rend compte qu'il émane d'une intelligence parce qu'il est périodique (quelques semaines). Cet argument n'est pas convaincant. Le signal pourrait être émis, de manière constante et naturelle, par une source en rotation autour d'un astre, lui-même en rotation, etc. ; du fait de l'occultation par d'autres sources elles-mêmes en rotation, on peut ainsi obtenir, sur Terre, un signal en apparence périodique.

Imaginons que des extraterrestres, au lieu d'un tel signal, nous envoient une suite réellement aléatoire. Nous aurions alors un second paradoxe :

Paradoxe 2. – Une suite totalement aléatoire ne contient aucune information, et pourtant elle caractériserait une civilisation plus avancée que la nôtre (au moins au sens mathématique !).

5. En conclusion

Cette génération de nombres aléatoires a un côté amusant et aussi un côté mystique : "Dieu joue-t-il aux dés ?". Mais je recommanderais une attitude simple et pragmatique :

- Faire la liste des différentes utilisations, présentes ou futures, et définir correctement dans chaque cas les propriétés nécessaires (ce qu'on appelle un "cahier des charges") ; nous l'avons fait grossièrement au début de cet article.

- Pour chaque situation, pour chaque cahier des charges, chercher à normaliser, à certifier, les outils proposés, de manière à garantir leur efficacité.

Par exemple, pour une fonction sur l'intervalle $[0,1]$, il s'agit de savoir combien de points seront nécessaires pour approximer la fonction avec telle précision : cela dépend de sa forme.

Cette attitude de vérification systématique donnerait de bien meilleurs résultats qu'une recherche dans l'abstrait, consistant à dire "nous allons chercher de nouveaux générateurs de nombres aléatoires". Le soin, quant à lui, ne devrait pas être laissé au hasard.

Références

[Devroye] Luc Devroye : Non-Uniform Random Variate Generation, Springer-Verlag, New York, 1986.

[Robinson] Peter Robinson : Efficient Calculation of Certain Integrals For Modelling Extremely Rare Events, 2009.
http://www.scmsa.eu/RMM/ART_2010_Peter_Robinson_Efficient_Integration.pdf

[MPPR] Bernard Beauzamy : Méthodes Probabilistes pour l'étude des phénomènes réels. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA, ISBN 2-9521458-0-6, ISSN 1767-1175. Mars 2004.

[NMP] Bernard Beauzamy : Nouvelles Méthodes Probabilistes pour l'évaluation des risques. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA. ISBN 978-2-9521458-4-8. ISSN 1767-1175, avril 2010.

[Quantis] Quantis : True random number generator exploiting quantum physics. http://www.idquantique.com/images/stories/PDF/quantis-random-generator/quantis-whitepaper.pdf

[Erreurs] Bernard Beauzamy : Erreurs judiciaires, erreurs mathématiques, 2012. http://scmsa.eu/Grand_public/2012_Erreurs_judiciaires_erreurs_mathematiques.pdf

[Lem] Stanislas Lem : La voix du Maître. Présence du Futur, 1976.