Smoluchowski et les tests de fiabilité des générateurs de nombres pseudo-aléatoires

Jean Dayantis ⋆

C.N.R.S., laboratoire de Science des matériaux vitreux
Université Montpellier II, place Eugène Bataillon, 34095 Montpellier cedex

RÉSUMÉ.

On définit le principe d'un test de fiabilité d'un générateur de nombres aléatoires, adapté d'un article de Smoluchowski datant de 1915. Le test est basé sur le dénombrement M_m des séquences de longueur m d'un symbole donné engendré par l'ordinateur. A partir des dénombrements M_m on définit un "temps de récurrence" et un "temps d'attente". Si les séquences de symboles (nombres) engendrées par le générateur sont strictement stochastiques, c'est à dire dénuées de corrélations à courte ou longue distance, alors les deux temps de récurrence et d'attente sont égaux. La généralisation de ce test "de Smoluchowski" primaire conduit à considérer l'ensemble des moments de la distribution des séquences M_m et on montre que ce test généralisé, auquel on adjoint toutefois un test complémentaire de type "run-test", est dans son principe nécessaire et suffisant.

ABSTRACT.

In this article, the principle of a test to check the reliability of random number generators is exposed. The test is based on a work of Smoluchowski going back to 1915. One first enumerates the number M_m of sequences of m successive symbols (numbers) of the same kind. From these enumerations one defines a "recurrence time" and a "waiting time". If the series of symbols is devoid of any short or long range correlations, then the recurrence and waiting times are equal. The generalization of this "Smoluchowski test" leads to the consideration of the entire set of the moments of the m distributions. It is shown that this generalized Smoluchowski test, to which is associated a complementary test of the run-test kind, is in its principle both necessary and sufficient.

⋆ Adresse permanente: Institut Charles Sadron, 6, rue Boussingault, 67083 Strasbourg cedex

1 - Les conditions de stochasticité de Smoluchowski.

Soit un système physique pouvant se trouver dans deux états "0" et "1". Dans un article paru en 1915 et traduit dans les présentes Annales [2], Smoluchowski introduit pour un tel système deux temps moyens, le temps moyen de "récurrence" et le temps moyen "d'attente". Le premier temps moyen, θ_1, est le temps moyen pour que, partant d'un état "0", le système revienne à cet état "0"; le deuxième temps moyen, θ_2, est le temps moyen pour que, partant d'un état "1" choisi au hasard, le système atteigne un état "0". Voici l'expression mathématique de ces deux temps moyens:

$$\theta_1 = \sum_{m=1}^{\infty} mM_m / \sum_{m=1}^{\infty} M_m \tag{1}$$

$$\theta_2 = (1/2) \sum_{m=1}^{\infty} m(m+1)M_m / \sum_{m=1}^{\infty} mM_m \tag{2}$$

où m représente une séquence de m états "0" successifs et M_m le nombre de séquences de m "0″ successifs dans une série temporelle tendant vers l'infini de successions d'états "0" et "1". Remarquons que les deux définitions (1) et (2) sont indépendantes, c'est à dire non déduisibles l'une de l'autre. Ces définitions posées, Smoluchowski, sans démonstration [1], énonce que si les successions des états "0" et "1" sont non-corrélées, en d'autres termes si la suite des états "0" et "1" se fait au hasard, de sorte que la série temporelle forme une suite de Markov [3], alors les deux temps moyens ci-dessus définis sont égaux. Nous dirons qu'une telle série temporelle est "stochastique".

Le problème physique envisagé par Smoluchowski se transpose de suite au problème équivalent d'une succession de nombres pseudoaléatoires "0" et "1" générés à l'ordinateur à l'usage de simulations Monte-Carlo, par exemple, 011101001010001011010111001010..., sauf que maintenant cette succession de "0" et de "1" ne forme plus nécessairement une série temporelle. Dans la série ci-dessus, il n'est pas nécessaire que les deux symboles "0" et "1" apparaissent avec la même probabilité, mais il faut que, si P est la probabilité d'apparition du "0" et $1-P$ celle du "1", P soit constant le long de la série envisagée.

En admettant la stochasticité, dans une série très longue (tendant vers l'infini), la probabilité p_m de m "0″ successifs suivant un "1" est donnée par

$$p_m = (1-P)P^m \tag{3}$$

On peut alors former les sommes suivantes:

$$\sum_{m=1}^{\infty} p_m = P/(1-P) \tag{4}$$

$$\sum_{m=1}^{\infty} m p_m = P/(1-P)^2 \tag{5}$$

$$(1/2)\sum_{m=1}^{\infty} m(m+1)p_m = (1/P)(\sum_{m=1}^{\infty} m p_m)(\sum_{m=1}^{\infty} p_m) = P/(1-P)^3 \tag{6}$$

La relation (6) entraîne:

$$\sum_{m=1}^{\infty} m^2 p_m = 2P(1-P)^3 - P/(1-P)^2 = P(1+P)/(1-P)^3 \tag{7}$$

A partir des relations (4), (5) et (7), on obtient, pour valeur moyenne $\langle m \rangle$ et carré moyen $\langle m^2 \rangle$ des séquences de zéros:

$$\langle m \rangle = \sum_{m=1}^{\infty} m P^m / \sum_{m=1}^{\infty} P^m = (1-P)^{-1} \tag{8}$$

et

$$\langle m^2 \rangle = \sum_{m=1}^{\infty} m^2 P^m / \sum_{m=1}^{\infty} P^m = (1+P)/(1-P)^2 \tag{9}$$

A partir des relations (5), (6) et (8), les définitions (1) et (2) de θ_1 et θ_2, et de la relation (3), correspondant à l'hypothèse de séquences Markoviennes, on trouve

$$\theta_1 = \theta_2 = (1-P)^{-1} = \langle m \rangle \tag{10}$$

ce qui veut dire, en accord avec la proposition énoncée par Smoluchowski, que si les séquences de "0" sont strictement stochastiques, les deux temps de récurrence et d'attente sont égaux. Observons que les égalités (3) à (10) sont à prendre au sens statistique, ce ne sont pas des égalités stricte au sens algébrique. La même remarque est valable pour nombre d'égalités qui suivent. Des calculs tout à fait analogues sont valables pour les séquences de "1". Ainsi, *la nature strictement stochastique des*

séquences M_m de m zéros successifs se traduit par le fait que la probabilité de celles-ci est donnée par la relation (3), qui implique à son tout que les valeurs de $\langle m \rangle$, et $\langle m^2 \rangle$ de ces séquences sont données respectivement par les relations (8), (9) et que de ce fait les relations (10) sont valables.

Si les séquences M_m ne sont pas strictement stochastiques, alors leurs probabilités respectives n'obéissent plus à la relation (3). De ce fait, et sauf cas exceptionnel de compensation statistique, les relations (10) ne seront pas satisfaites. En somme, les conditions de Smoluchowski reviennent à fixer les deux premières moyennes de la distribution de probabilités des séquences de "0" et des "1". Elles reviennent encore à écrire

$$\langle m \rangle = (1-P)^{-1} \quad ; \quad \langle m^2 \rangle - 2\langle m \rangle^2 + \langle m \rangle = 0 \qquad \text{(8a, 10a')}$$

$$\langle n \rangle = P^{-1} \quad ; \quad \langle n^2 \rangle - 2\langle n \rangle^2 + \langle n \rangle = 0 \qquad \text{(8b, 10b')}$$

où n représente une séquence de n"1" successifs.

2 - Généralisation des conditions de stochasticité de Smoluchowski.

Evidemment, les deux premières moyennes d'une distribution sont insuffisantes, dans le cas général, pour définir la distribution elle-même. Les quatre conditions de Smoluchowski ci-dessus sont donc nécessaires mais non suffisantes. Cependant, on peut élargir le travail initial de Smoluchowski, et se donner l'ensemble des moyennes, ou, ce qui revient au-même, des moments de la distribution. Mais pour cela, en vue de la rigueur mathématique, il convient d'abord de montrer l'existence puis de calculer effectivement le moment μ_k d'ordre k, k entier positif quelconque.

Existence et calcul du moment d'ordre k.

Le moment μ_k d'ordre k de la distribution

$$\sum_{m=1}^{\infty} P(x)\delta(x-m) \qquad (11)$$

où δ est la distribution de Dirac, est donné par

$$\mu_k = \sum_{m=1}^{\infty} m^k P^m \quad 0 < P < 1 \qquad (12)$$

Pour montrer que cette somme converge quel que soit k, partons de l'expression

$$\int_0^\infty m^k P^m dm = \int_0^\infty m^k \exp[m \ln P] dm = A_k$$

A_k est si l'on veut le moment k de la distribution continue correspondant à la distribution discrète (11), et existe toujours, puisque P est inférieur à un. On a la majoration suivante:

$$\sum_{m=2}^{\infty} m^k P^m < \int_0^\infty m^k \exp[m \ln P] dm = A_k \tag{13}$$

Il en résulte que

$$\sum_{m=1}^{\infty} m^k P^m < A_k + P \tag{14}$$

et comme $A_k + P$ est toujours fini, il en est de même de μ_k.

Le calcul effectif du moment μ_k d'ordre k quelconque peut alors se faire de la manière suivante: soit la fonction

$$f(P) = \mu_0 = \sum_{m=1}^{\infty} P^m = P/(1-P)$$

On remarque alors que

$$\begin{aligned} \mu_1 &= \sum_{m=1}^{\infty} mP^m = Pf' \\ \mu_2 &= \sum_{m=1}^{\infty} m^2 P^m = P[(d/dP)(Pf')] = P[(d/dP)\mu_1] \end{aligned} \tag{15}$$

et, d'une manière générale, on a la relation de récurrence

$$\mu_k = \sum_{m=1}^{\infty} m^k P^m = P[(d/dp)\mu_{(k-1)}] \tag{16}$$

En exploitant la relation de récurrence (16), on trouve comme expression du moment μ_k

$$\mu_k = \sum_{m=1}^{\infty} m^k P^m = (1-P)^{-(k+1)} \sum_{m=1}^{k} \alpha_r^k P^r \tag{17a}$$

où les coefficients α_r^k sont donnés par la relation de récurrence

$$\alpha_r^k = r\alpha_r^{k-1} + (n-r+1)\alpha_{r-1}^{k-1} \tag{17b}$$

A partir de la valeur des moments ci-dessus indiqués, on obtient facilement les diverses moyennes d'ordre k, $\langle m^k \rangle$.

3 - La fonction caractéristique.

λ étant un réel > 0, la transformée de Fourier de la distribution de probabilités normée

$$p(x) = [(1-P)/P] \sum_{m=1}^{\infty} P^x \delta(x-m) \tag{11}$$

est donnée par

$$\begin{aligned} C(\lambda) =& [(1-P)/P] \int_{-\infty}^{+\infty} \sum_m P^x \delta(x-m) \cdot \exp[-2i\pi\lambda x] dx \\ =& [(1-P)/P] \sum_{m=1}^{\infty} P^m \exp[-2i\pi\lambda m] \end{aligned} \tag{18}$$

A partir de (18) on retrouve la distribution de départ par une transformation de Fourier inverse:

$$\begin{aligned} p(x) =& [(1-P)/P] \int_{-\infty}^{+\infty} C(\lambda) \exp[2i\pi\lambda x] d\lambda \\ =& [(1-P)/P] \int_{-\infty}^{+\infty} \sum_m P^m \exp[2i\pi\lambda(x-m)] d\lambda \\ =& [(1-P)/P] \sum_{m=1}^{\infty} P^m \delta(x-m) \end{aligned} \tag{11'}$$

où on a fait usage de la relation [4]

$$\delta(x) = \int_{-\infty}^{+\infty} \exp[-2i\pi\lambda x] d\lambda \tag{19}$$

La distribution de Dirac étant "tempérée" au sens de Schwartz [4], il s'ensuit que la somme (11') ci-dessus est aussi une distribution tempérée,

ce qui entraîne à la fois l'existence et l'unicité de la transformation de Fourier. Il en est de même pour la transformation inverse. Par ailleurs, la transformée de Fourier (ou fonction caractéristique) donnée par (18), est aussi la fonction génératrice des moments ou moyennes de la distribution de probabilités de départ (11') ci-dessus (la distribution étant normée, moments et moyennes se confondent):

$$C(\lambda) = 1 - 2i\pi\langle m\rangle\lambda + (2\pi)^2\langle m^2\rangle\lambda^2/2! - i(2\pi)^3\langle m^3\rangle\lambda^3/3! + \ldots \quad (20)$$

Ces moyennes existent en vertu de la majoration (14) ci-dessus. Il en résulte que la donnée des moments μ_k (ou moyennes $\langle m^k\rangle$), définit sans ambiguïté la distribution $C(\lambda)$, et vu l'unicité de la transformation $p(x) \to C(\lambda)$ et de la transformation inverse $C(\lambda) \to p(x)$, la distribution de départ $p(x)$ elle-même. En conséquence, le test généralisé de Smoluchowski (ou test des moments) est, dans son principe, **nécessaire et suffisant**, pour s'assurer que la distribution $p_{ord}(x)$ donnée par l'ordinateur se confond au sens statistique avec la distribution attendue (11'). Si il en est ainsi, les longueurs des séquences des symboles "0" et "1" sont stochastiques.

Ceci cependant ne suffit pas pour s'assurer que la série elle-même (et non plus ses séquences) est stochastique. Car, on peut imaginer un démon de Maxwell (en fait, un programme informatique annexe associé au programme initial du générateur) qui redistribuerait de manière artificielle et arbitraire les séquences à l'intérieur d'une longue série de symboles pseudoaléatoires. Par exemple, ce programme complémentaire pourrait redistribuer les séquences de "0" et de "1" par ordre de longueur croissant, de sorte que les séquences les plus courtes se trouveraient en début de série, et les séquences les plus longues en fin de série. On peut ainsi imaginer une infinité de reclassements arbitraires de ce genre, qui de toute évidence introduisent des corrélations et sont indétectables par le test des moments. En bref, le test généralisé de Smoluchowski détecte, dans son principe, tout biais dans la longueur attendue des séquences (au sens statistique, comme toujours), mais ne détecte nullement une redistribution artificielle de ces séquences le long de la série selon quelque loi arbitraire, prenant en considération leurs longueurs comparées.

4 - Le test annexe.

Vu ce qui précède, il convient de disposer d'un test complémentaire, qui mettrait en évidence toute corrélation entre position des séquences

selon leur longueur. Ce test complémentaire peut être le "run-test", décrit dans l'ouvrage de Knuth [5], mais appliqué non pas aux symboles individuels de la série, mais à leurs séquences. Pour fixer les idées, supposons que ces symboles soient les chiffres usuels de la numération décimale $0, 1, 2 \ldots$. Le test originel des "runs-up" et "runs-down", évalue le nombre de séquences "montantes" et "descendantes" c'est à dire celles dont les valeurs des symboles vont en augmentant et celles où elles vont en diminuant. Par exemple, dans la série $1, 3, 7, 2, 5, 6, 5, 4, 0$, il y a une séquence montante $1, 3, 7$ de trois, suivie d'une séquence descendante de deux, $7, 2$, puis une séquence montante de trois, $2, 5, 6$, et enfin une séquence descendante de quatre, $6, 5, 4, 0$.

On démontre que, statistiquement, on doit avoir autant de séquences montantes que descendantes, si la série dans son ensemble est stochastique.

Pour appliquer ce test non plus aux symboles eux-mêmes mais à leurs séquences, on peut procéder comme suit: soit P la probabilité globale dans la série du symbole "0". Si on considère uniquement la succession des séquences de "0", et qu'on fait donc abstraction des séquences intercalées de "1", il faut, vu ce changement de support, normer les probabilités initiales par $\sum_{m=1}^{\infty} P^m = P/(1-P)$. Dans ces conditions, la probabilité pour obtenir une séquence de "0" comportant de 1 à m symboles est donnée par

$$[(1-P)/P] \sum_{q=1}^{m} P^q = 1 - P^m \tag{21}$$

et celle d'avoir une séquence de plus de m "0″ est en conséquence P^m. Notons aussi que la probabilité d'avoir une séquence égale (comportant aussi m "0″) est $P^{m-1} - P^m = P^{m-1}(1-P)$. Si donc la dernière séquence construite de la série comporte m "0″ successifs, les probabilités pour que la séquence suivante soit plus longue, égale, ou plus courte sont respectivement

$$\text{Prob} \begin{matrix} \nearrow \\ \rightarrow \\ \searrow \end{matrix} \begin{pmatrix} P^m \\ P^{m-1}(1-P) \\ 1 - P^{m-1} \end{pmatrix} \tag{22a, b, c}$$

Il faut maintenant calculer la probabilité sur toutes les valeurs de m pour que la séquence qui suit la dernière générée soit plus longue, égale, ou plus courte. Dans le premier cas, celle-ci sera la somme sur tous les m

du produit des probabilités d'une séquence égale à m et d'une séquence supérieure à m, soit:

$$(\text{Prob} \nearrow) = \sum_{m=1}^{\infty} [(1-P)/P]P^m \cdot P^m = P/(1+P) \qquad (23a)$$

De manière analogue, les deux autres probabilités seront données par

$$(\text{Prob} \rightarrow) \sum_{m=1}^{\infty} [(1-P)/P]P^m \cdot (P^{m-1} - P^m) = (1-P)/(1+P) \quad (23b)$$

$$(\text{Prob} \searrow) = \sum_{m=1}^{\infty} [(1-P)/P]P^m \cdot (1 - P^{m-1}) = P/(1+P) \qquad (23c)$$

Il ne peut se faire que l'on obtienne, par hasard, les relations (23), suite à un effet de compensation statistique (sommation compensatrice) entre valeurs erronées (au sens statistique) P'^m, de telle sorte que l'on ait $\sum_m P'^m = \sum_m P^m$. En effet, des valeurs statistiquement incorrectes P'^m pour les probabilités des séquences de m"0", auraient été au préalable mises en évidence par le test des moments. Des relations (23) découle que dans une série strictement stochastique, si on considère les séquences successives du symbole "0" deux-à-deux, il y a, en moyenne statistique, autant de couples de séquences croissantes que de couples de séquences décroissantes. Leur nombre total dépend de la probabilité globale dans la série du symbole "0".

Un calcul tout analogue est valable pour les séquences de "1", en remplaçant partout P par $1-P$. Si la série est strictement stochastique, le calcul ci-dessus effectué en considérant deux séquences successives de "0" et conduisant aux relations (23), reste valable si les deux séquences ne sont plus successives, mais distantes de Q, où Q est un entier positif quelconque. En d'autres termes, les relations (23) en posant $Q = 1$ impliquent la non-corrélation entre les longueurs de deux séquences successives de "0", et ces mêmes relations où on pose Q entier > 1, impliquent la non-corrélation entre la séquence de "0" de référence et la Qième séquence de "0" le long de la série. Si donc les relations (23) sont vérifiées quel que soit Q entier, positif, il s'ensuit qu'il ne peut exister aucune corrélation de longueur entre séquences de "0", quelle que soit la distance de ces séquences le long de la série de symboles pseudoaléatoires. Des considérations tout analogues sont valables pour les séquences de "1".

En conclusion, le run-test appliqué aux séquences de symboles, détecte dans son principe toute redistribution artificielle selon une loi déterministe quelconque des séquences de "0" et de "1" le long de la série.

5 - Suffisance de la somme des deux tests.

On peut résumer ce qui précède comme suit: (a) *le test de Smoluchowski généralisé (ou test des moments) détecte, dans son principe, tout biais dans la distribution des longueurs des séquences; et (b) le run-test appliqué aux séquences détecte, dans son principe, toute corrélation entre les longueurs de celles-ci le long de la série.* Comme les seuls biais possibles d'une série sont soit des biais en longueur, soit des biais en positionnement des séquences (soit évidemment une combinaison des deux), il s'ensuit que le test global qui comporte la somme des deux tests ci-dessus (test de Smoluchowski généralisé + run-test appliqué aux séquences), est un test **nécessaire et suffisant** pour mettre en évidence la stochasticité ou la non-stochasticité d'une série de symboles pseudoaléatoires. Il faut cependant s'empresser d'ajouter, que si, dans son principe, le test global ci-dessus est nécessaire et suffisant, cela ne signifie nullement **qu'en pratique**, d'autres tests, nécessaires mais non-suffisants, perdraient de ce fait leur utilité. Car, intervient ici de façon décisive la **sensibilité** des tests, soit à l'échelon local soit à grande distance. C'est là une question essentielle, qu'il faudrait examiner de près, mais qui sort du cadre du présent travail, qui s'attache uniquement aux aspects de principe.

Signalons pour terminer que nous n'avons pas trouvé décrit dans les ouvrages que nous avons pu consulter [5 - 8] le test qui a fait l'objet du présent travail; cependant, avec sans doute des différences, ce test est parent du "gap-test" décrit dans l'ouvrage de Knuth [5].

Remerciements.

Monsieur Jean-François Palierne, Ecole Normale Supérieure de Lyon, Service Recherche Physique, a activement participé à l'élaboration de ce travail. Les relations (17) lui sont dues, et la remarque est sienne, comme quoi le test de Smoluchowski généralisé n'est pas à lui seul suffisant, car il ne saurait mettre en évidence une redistribution artificielle des séquences le long de la série selon une loi arbitraire.

Nous remercions également Monsieur Michel Karatchentzeff, de la Fondation Louis de Broglie, pour une correspondance utile.

Références

[1] M. Von Smoluchowski - *Molekulartheoretische Studien Uber Umkehr thermodynamisch irreversibler Vorgange und Uber Wiederkehr abnormaler Zustande* - Sitz. Kaiserl. Akad. Wiss. Wien, Mathem. Naturw. Klasse, Abteilung IIa, **124**, 339-368 (1915). Article reproduit dans la série *Ostwald's Klassiker der Exacten Wissenschaften*, ouvrage intituté *Abhandlungen Uber die Brownsche Bewegung und verwandte Erscheinungen* par M. Von Smoluchowski, édité par R. Fürth-Prag, Leipzig 1923.

[2] Annales de la Fondation Louis de Broglie, **19** (1 et 2), 1 (1994).

[3] A.A. Markoff, *Wahrscheinlischkeitsrechnung*, Leipiz 1912; voir S. Chandrasekhar, Rev. Mod. Phys., **15**, 1 (1943).

[4] Laurent Schwartz, *Méthodes Mathématiques pour les Sciences Physiques*, Hermann, Paris 1961.

[5] D.E. Knuth, *The Art of Computer Programming*, Sec. Ed., Addison-Wesley, Reading, Mass. 1981.

[6] R.Y. Rubinstein, *Simulation and the Monte Carlo Method*, Wiley, New-York 1981.

[7] L.P. Devroye, *Non-Uniform Random Variate Generation*, Springer, New-York 1986.

[8] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania 1992.

(Manuscrit reçu le 16 juillet 1996, révisé le 23 mai 1997)