REVUE DE STATISTIQUE APPLIQUÉE

J.-L. CHASSÉ
D. DEBOUZIE

Utilisation des tests de Kiveliovitch et Vialar dans l'étude de quelques générateurs de nombres pseudo-aléatoires

Revue de statistique appliquée, tome 22, n° 3 (1974), p. 83-90
<http://www.numdam.org/item?id=RSA_1974__22_3_83_0>

UTILISATION DES TESTS DE KIVELIOVITCH ET VIALAR DANS L'ÉTUDE DE QUELQUES GÉNÉRATEURS DE NOMBRES PSEUDO-ALÉATOIRES [1]

J.-L. CHASSÉ et D. DEBOUZIE
Laboratoire de biométrie
Université Claude Bernard, Lyon

INTRODUCTION

La simulation sur calculateur de phénomènes ou processus biologiques nécessite l'utilisation de nombres aléatoires. En particulier l'estimation de la taille d'une population et l'étude du déplacement de ses individus font appel à des techniques exigeant des répétitions indépendantes et nombreuses. Par suite, il est nécessaire de disposer d'une quantité importante de nombres aléatoires. Ils peuvent être obtenus par des procédés physiques qui exigent la mise au point du générateur physique et sa connexion au calculateur ou être générés directement sur le calculateur, en se fondant sur un algorithme correctement choisi. Dans ce cas, la séquence obtenue est appelée "pseudo-aléatoire" au sens de Lehmer (15), c'est-à-dire "une séquence dont chaque terme est imprévisible pour un non initié et dont les éléments vérifient certains tests".

La génération d'une série de nombres distribués uniformément sur un intervalle donné est fondamentale car à partir d'elle il est possible de construire des séries distribuées selon d'autres lois. Elle sera donc l'objet de notre étude qui porte sur la comparaison de quelques algorithmes de génération, selon la longueur de la séquence. Ceux-ci sont retenus si les séries qu'ils engendrent satisfont à certains critères parmi lesquels nous avons choisi un test de fréquence et quelques-uns des tests de Kiveliovitch et Vialar (13) relatifs à l'étude de séries chronologiques de nombres entiers.

DONNEES BIBLIOGRAPHIQUES

Les algorithmes de génération les plus connus outre ceux dont l'intérêt n'est plus qu'historique (carré médian par exemple), sont les méthodes de congruence dans lesquelles la série est engendrée par la relation de récurrence suivante :

---

(1) Article remis le 9/1/1973, révisé le 25/1/1973.

$$x_n \equiv a_1 x_{n-1} + a_2 x_{n-2} + \ldots + a_k x_{n-k} + a_{k+1} \pmod{m}$$

Les valeurs des paramètres $a_1, a_2, \ldots, a_{k+1}$ définissent les types suivants :

- $a_1 = a_2 = 1$; $a_i = 0$ $(i = 3, 4, \ldots, k+1)$ méthode additive
- $a_1 \in N - (0, 1)$; $a_i = 0$ $(i = 2, 3, \ldots, k+1)$ méthode multiplicative
- $a_1 \in N - (0, 1)$; $a_{k+1} \neq 0$; $a_i = 0$ $(i = 2, \ldots, k)$ méthode mixte
- $a_{k+1} = 0$; $a_1, a_2, \ldots, a_k \neq 0$ $(k>2)$ méthode linéaire

Notre but n'est pas de faire une bibliographie exhaustive qui ferait suite à celle de Hull et Dobell (12), mais de dégager de l'ensemble des travaux les points qui nous semblent importants. Nous noterons tout d'abord que beaucoup d'auteurs se sont intéressés à la longueur de la période de la séquence selon les types de générateurs et les valeurs des paramètres (a_i, m) (Allard, Dobell et Hull (1), Deltour (7), Gotusso (9), Hull et Dobell (12), Peach (16)). Ces travaux montrent qu'il est possible, pour les différents types de méthode, de trouver des valeurs des paramètres conduisant à des périodes suffisamment longues pour la simulation de processus biologiques. Il est en effet démontré que si les paramètres satisfont à certaines conditions, la période est égale à m pour les méthodes congruentielles mixte et additive et à m/4 pour la multiplicative.

Outre une période convenable, la séquence doit présenter un caractère pseudo-aléatoire qui est défini par certains critères choisis par l'utilisateur. L'étude et surtout la mise en oeuvre de ces critères constituent la deuxième catégorie de travaux. Dans la quasi totalité des cas la première condition réclamée pour la séquence est l'équirépartition sur (0, m) des nombres engendrés, l'intervalle étant divisé en parties égales, généralement 10 à 100. Le test utilisé est le test χ^2 ou parfois le test de Kolmogorov-Smirnov (test des fréquences). Remarquons que si la séquence satisfait à ce test, l'utilisation du test des moments (Taussky et Todd (19)) ou du test binomial (Kendall, Badington-Smith in Shreider (18)) n'apporte en général aucun renseignement supplémentaire. La condition d'équirépartition permet déjà de rejeter le plus souvent un algorithme de génération fondé sur la méthode congruentielle additive. Par contre, si la condition est remplie, il faut alors tester l'indépendance des termes de la séquence. Pour ce faire, les tests généralement proposés appartiennent à plusieurs catégories :

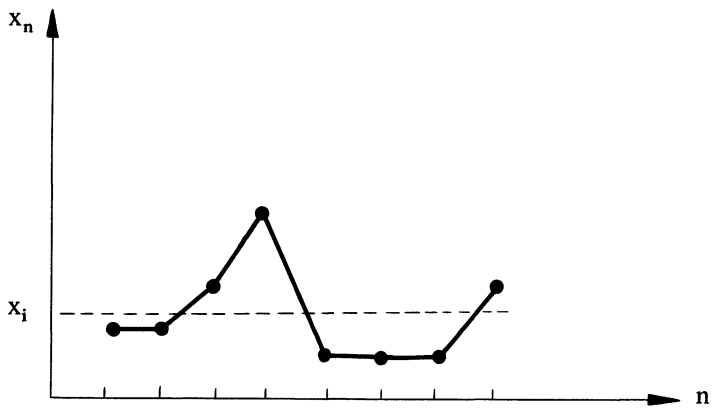
- "gap test", "poker test", suites croissantes et décroissantes, suites au-dessus et en-dessous de la médiane (Cameron et Handy(3), Downham et Roberts (8), Horton et Smith (11), Tootill, Robinson et Adams (20)).
- test de corrélation sériale (Coveyou (4)).
- analyse de Fourier de la séquence générée (Coveyou et Macpherson (5)).

Nous remarquons que les tests de la première catégorie sont soit une partie des tests de Kiveliovitch-Vialar, soit une application de ceux-ci au cas d'une série composée d'éléments égaux à 0 ou 1. Les tests de Kiveliovitch-Vialar permettant de déceler, dans une série chronologique, l'existence de périodes et (ou) tendances, recouvrent les tests de corrélation sériale. C'est pourquoi, nous consi-

dérerons qu'une séquence est pseudo-aléatoire si, outre le test des fréquences, elle satisfait aux tests de Kiveliovitch-Vialar suivants :

- test des croisements
- test des phases
- test des paliers
- test du nombre par palier

Nous en rappelons brièvement le principe en considérant le graphe de la séquence $x_n = f(n)$.



Test des croisements (CR)

Un croisement est défini comme l'intersection du graphe et d'une droite d'ordonnée α_i. Pour chaque valeur de i on compte le nombre de croisements.

Test des phases (PH)

Une phase est une portion limitée par deux extremums consécutifs ; elle est caractérisée par le nombre de segments reliant les deux extremums.

Test des paliers (P)

Un palier est une suite de valeurs égales. Ce test implique le regroupement en classes des x_n. Pour chaque classe, on compte le nombre de paliers.

Test du nombre dans les paliers (NP)

Pour chaque classe on compte le nombre d'éléments appartenant à un palier.

Pour ces quatre tests, les valeurs observées sont comparées aux valeurs théoriques calculées sous l'hypothèse d'indépendance (les intervalles de confiance sont calculés au risque de 0,05).

Les auteurs ont généralement mis en oeuvre les tests qu'ils proposaient sur des séquences relativement courtes : 5000 nombres (Lehmer), séries de 128 à 1024 nombres (Cameron et Handy) par exemple. D'autre part les travaux systématiques sur ces méthodes sont peu nombreux ; on peut citer ceux de Allard,

Dobell et Hull qui pour la méthode mixte ont testé plus de mille multiplicateurs (paramètre a_1).

Enfin nous signalons la tendance qu'ont eue certains auteurs à vouloir mettre au point des algorithmes fondés sur une complication des méthodes congruentielles : ce sont les "shuffling methods" (Král (14)), la "compound randomization" (Horton (10), Beasley et Wilson (2)) ou des méthodes plus complexes (Cugiani et Liverani (6)).

Il nous semble plutôt qu'il faille chercher une solution en tenant compte – ce qui n'est pas toujours noté dans les travaux – des capacités du calculateur qui sera employé, du but pour lequel la séquence est générée et en outre utiliser une méthode congruentielle mixte, linéaire ou multiplicative. C'est cette dernière d'ailleurs qui a été à la fois la plus étudiée d'un point de vue théorique et la plus utilisée. Il est en outre intéressant de noter, que Coveyou et Macpherson abordant l'étude de la séquence d'une façon nouvelle (analyse de Fourier), tirent comme première conclusion "qu'il n'y a pas jusqu'à présent de méthode de génération de séquences pseudo-aléatoires meilleure que la simple méthode congruentielle multiplicative avec un multiplicateur soigneusement choisi".

TRAVAUX

Nous avons testé les deux types de méthodes :

$x_n \equiv a x_{n-1} \pmod m$ Type I

$x_n \equiv a x_{n-1} + x_{n-2} \pmod m$ Type II

pour les valeurs suivantes :

$(a, m) = (23, 10^5 + 1)$; $(17, 10^5 + 1)$; $(5717, 10^5 + 1)$; $(23, 10^8 + 1)$;

$(5^5, 2^{18})$; $(5^{13}, 2^{36})$

Le premier couple de paramètres est celui proposé par Lehmer. Pour les trois suivants nous avons fait varier soit le multiplicateur soit le modulo, le multiplicateur étant choisi de façon à répondre aux critères de Hull. Le dernier couple est celui qui a été utilisé par Rohlf et Davenport (17) pour des simulations de comportement locomoteur proches de celles que nous étudions. Le couple $(5^5, 2^{18})$ qui ne diffère du précédent que par la réduction des paramètres a été choisi pour être éventuellement utilisé sur des calculateurs de petite ou moyenne capacité.

La génération des nombres pseudo-aléatoires et les tests ont été effectués sur un calculateur Wang 700 qui permet de disposer de 24 chiffres significatifs pendant le cours du calcul.

Dans chaque cas, nous avons procédé au tirage de 10 séries de 10000 nombres et 10 séries de 100000 nombres. Chaque nombre x_i appartient à l'intervalle (0, 1) mais pour procéder aux tests nous avons considéré la séquence des nombres u_i appartenant à l'intervalle (0, 9) tels que $u_i = E(10.x_i)$.

Les résultats sont regroupés dans le tableau 1. Pour simplifier la lecture de ce tableau nous avons seulement indiqué si les hypothèses d'équirépartition et d'indépendance des termes pouvaient être acceptées ou non, selon les différents tests.

	Méthode congruentielle multiplicative (Type I)					Méthode congruentielle linéaire (Type II)				
	F	CR	PH	P	NP	F	CR	PH	P	NP
$(17,10^5+1)$	+	+	+	+	+	o	o	o	o	+
	+	+	+	+	+	+	+	+	+	+
$(23,10^5+1)$	+	+	+	+	+	o	o	+	o	+
	+	+	+	+	+	o	o	+	o	+
$(5717,10^5+1)$	o	o	+	o	+	+	+	+	+	+
	o	+	+	+	+	+	+	+	+	+
$(23,10^8+1)$	o	+	+	+	+	o	+	+	o	+
	o	+	+	+	+	o	o	+	o	+
$(5^5, 2^{18})$	o	o	+	o	+	o	o	+	o	+
	o	o	o	o	o	o	o	+	o	+
$(5^{13}, 2^{36})$	o	+	+	+	+	o	o	+	o	+
	o	o	o	o	+	o	+	+	o	+

TABLEAU 1

Résultats des tests concernant les hypothèses d'équirépartition et d'indépendance des termes.

+ Rejet de l'hypothèse

o Acceptation de l'hypothèse

F : test d'équirépartition (χ^2)

CR, PH, P, NP : tests d'indépendance (Kiveliovitch-Vialar)

La ligne supérieure concerne les séquences de 10000 nombres et la ligne inférieure celles de 100000 nombres.

1) L'examen du tableau I conduit à rejeter les générateurs suivants qui ne satisfont soit à aucun test soit seulement au test χ^2 :

– Séquences de 10000 ou 100000 nombres

Type I : $(17,10^5 + 1)$; $(23,10^5 + 1)$; $(23,10^8 + 1)$

Type II : $(5717,10^5 + 1)$

– Séquences de 10000 nombres

Type I : $(5^{13}, 2^{36})$

– Séquences de 100000 nombres

Type I : $(5717, 10^5 + 1)$

Type II : $(17, 10^5 + 1)$

Par contre un seul générateur satisfait à tous les tests :

Type I : $(5^5, 2^{18})$ séquence de 100000

Les autres générateurs sont défaillants au niveau d'au moins un test de Kiveliovitch-Vialar. Il faut cependant remarquer qu'un examen détaillé des résultats des tests de Kiveliovitch-Vialar montre que pour les générateurs de Type II les différences significatives sont sans tendance, les valeurs observées n'étant pas, selon les répétitions, systématiquement inférieures ou supérieures aux valeurs théoriques.

2) Pour les générateurs I (a, $10^\alpha+1$) nous constatons que l'augmentation du multiplicateur conduit à une amélioration de l'équirépartition, mais ce qui semble important est le comportement totalement différent des deux multiplicateurs faibles. En effet, avec le multiplicateur 17 nous obtenons, dans tous les cas, des valeurs de χ^2 très fortes ($p<0.001$) alors que pour 23 l'équirépartition est "trop bonne", les valeurs de χ^2 étant systématiquement trop proches de zéro ($\chi^2<0.002$). L'existence d'une période très inférieure à 10000 avec équirépartition à l'intérieur de celle-ci pourrait en être la cause. En fait, cette période n'existe pas et nous ne pouvons pour l'instant expliquer ce résultat.

De plus l'utilisation du multiplicateur 23 conduit à un nombre de paliers plus grand que prévu pour les valeurs 0, 4, 5, 9 indépendamment de la valeur du diviseur. Un calcul de probabilité montre que cet excès est dû pour les paliers extrêmes (0 et 9) à la faible valeur du multiplicateur et que pour les paliers de 4 et 5 il est lié en outre au découpage en dix intervalles.

3) L'augmentation de la longueur de la séquence conduit évidemment à augmenter la puissance des tests. On pourrait alors s'attendre, en plus des chances d'apparition d'une période, à mettre en évidence des différences significatives qui n'apparaissaient pas dans les séquences de 10000 nombres. Or, on constate que tous les algorithmes ne réagissent pas de la même façon. Pour la méthode II(a, $10^\alpha+1$) ce passage conduit, dans un cas, à juger comme inacceptable un algorithme $(17, 10^5+1)$ jugé précédemment satisfaisant. Par contre, avec les paramètres $(5^\beta, 2^\delta)$ l'augmentation de la longueur de la séquence conduit à une amélioration sensible de la qualité pour le Type I alors que pour le Type II il n'y a pas de modification notable.

CONCLUSION

Parmi les méthodes de génération de séquences pseudo-aléatoires que nous avons mises en oeuvre, les critères d'équirépartition (test χ^2) et d'indépendance des termes de la séquence (tests de Kiveliovitch-Vialar) nous ont conduits à ne retenir que la méthode suivante :

$$x_n \equiv 5^5 \, x_{n-1} \pmod{2^{18}}$$

L'étude comparée des méthodes montre une sensibilité très différente aux tests utilisés selon le type de la méthode, la forme et la valeur des paramètres ainsi que la longueur de la séquence. Ceci nous confirme dans l'idée qu'il n'est pas possible de trouver un générateur idéal. Chaque fois qu'il sera nécessaire d'utiliser une séquence pseudo-aléatoire, il faudra tenir compte de la capacité du calculateur, de la période désirée et de l'emploi qui sera fait de la séquence.

Les résultats que nous avons obtenus nous incitent à insister sur le très grand soin qui doit être apporté au choix du générateur pour la mise en oeuvre d'une méthode de simulation.

BIBLIOGRAPHIE

[1] ALLARD, J.L., DOBELL, A.R. et HULL, T.E. (1963) "Mixed congruential random number generators for decimal machines" *J. Ass. computg. Machin,* USA, 10, 131–141.

[2] BEASLEY, J.D. et WILSON, K. (1969) "Design and testing of the System 4 random number generator" *Computer J.,* GB, 12, 368–372.

[3] CAMERON, J.M. et HANDY, B.F. Jr (1952) "Results on some tests of randomness on pseudo-random numbers" Abstract in *Annals Math. Stat.,* 23, 138.

[4] COVEYOU, R.R. (1960) "Serial correlation in the generation of pseudo-random numbers" *J. Ass, computg. Machin,* USA, 7, 72–74.

[5] COVEYOU, R.R. et MACPHERSON, R.D. (1967) "Fourier analysis of uniform random number generators" *J. Ass. computg. Machin,* USA 14, 100–119.

[6] CUGIANI, M. et LIVERANI, A. (1966) "Un nuovo metodo per la generazione di numeri pseudo-casuali" *Atti Seminar, mat. fis. Univ. Modena,* 15, 194–197.

[7] DELTOUR, J. (1967) "Etude d'une distribution de nombres pseudo-aléatoires". *Bull. Rech. agron. Gembloux,* 2, 450–460.

[8] DOWNHAM, D.Y. et ROBERTS, F.D.K. (1967) "Multiplicative congruential pseudo-random number generators" *Computer J.,* G.B., 10, 74–77.

[9] GOTUSSO, L. (1966) "Su alcune sequenze pseudo casuali di tipo congruenziale misto" *Atti seminar, mat. fis. Univ. Modena,* 15, 188–193.

[10] HORTON, H.B. (1948) "A method for obtaining random numbers" *Ann Math. Statistics,* 19, 81–85

[11] HORTON, H.B. et SMITH, R.T. (1949) III "A direct method for producing random digits in any number system" *Ann. Math. Statistics,* 20, 82–90.

[12] HULL, T.E. et DOBELL, A.R. (1962) "Random numbers generators" *SIAM Review,* 4, 230–254.

[13] KIVELIOVITCH, M. et VIALAR, J. (1957) "Les séries chronologiques et la théorie du hasard" Publ. sci. tech, Minist. Air, *Notes tech.,* Fr, n° 65.

[14] KRAL, J. (1971) "Experimental properties of some additive pseudo-random number generators with random shuffling" *Apl. Nat., Ceskolsl.,* 16, 395–401.

[15] LEHMER, D.R. (1949) "Mathematical methods in large scale computing units" Proc. sympos. on large-scale digital calcul. machinery, Harvard Univ. press, 141–146.

[16] PEACH, P. (1961) "Bias in pseudo-random numbers" *J. amer. statist. Ass.* 56, 610–618.

[17] ROHLF, F.J. et DAVENPORT, D. (1969) "Simulation of Simple Models of Animal Behavior with a Digital Computer" *J. Theoret, Biol.,* 23, 400–424.

[18] SHREIDER, Y.A. (1966) The Monte Carlo Method, Pergamon Press.

[19] TAUSSKY, O. et TODD, J. (1954) "Generation and testing of pseudo-random numbers" Symposium on Monte Carlo Methods, Herbert, A. Meyer Ed., 15–28.

[20] TOOTILL, J.P.R., ROBINSON, W.D. et ADAMS, A.G. (1971) "The runs up-and-down performance of Tausworthe pseudo-random number generators". *J. Ass. comptg. Machin,* USA, 18, 381–399.