Statistique et analyse des données

Bernard Van Cutsem

Simulation des lois de probabilité

Statistique et analyse des données, tome 10, n° 2 (1985), p. 63-87.

<http://www.numdam.org/item?id=SAD_1985__10_2_63_0>

Statistiques et Analyse de Données
Septembre 1985 Vol. 10 n° 2 pp. 63-87

SIMULATION DES LOIS DE PROBABILITE

Bernard VAN CUTSEM

Institut IMAG. Laboratoire TIM3.
Université Scientifique et Médicale de Grenoble.
B.P. 68. 38402 Saint Martin d'Hères Cedex

Résumé : *Ce texte est celui d'une conférence faite dans le cadre des Journées "Statistique et Industrie" organisées à Marseille en janvier 1985. Nous y présentons succintement l'essentiel des connaissances indispensables pour pouvoir utiliser à bon escient un générateur de nombres pseudo-aléatoires, en vérifier la qualité et éventuellement en construire un adapté au matériel informatique dont on dispose.*

Summary : *This text was presented as a lecture given in "Journées Statistiques et Industrie" which were organized in Marseille in january,1985. We briefly present here the main knowledges which are necessary to advisedly use pseudo-random numbers generators, to verify their qualities, and eventually to build a new one specially fitted to the computer at one's disposal.*

Mots clés : *Simulation. Méthodes de Monte-Carlo. Générateurs de nombres pseudo-aléatoires.*

Indices de classification ISI : 15. 100 - 05. 110

Manuscrit reçu le 4 mars 1985
révisé le 21 octobre 1985

I. INTRODUCTION.

En un sens très large, le mot simulation désigne l'ensemble des techniques, physiques, mécaniques, informatiques ou mathématiques qui permettent de fabriquer un système fictif modélisant un système réel et de faire "marcher" ce système fictif pour en tirer des renseignements sur le système réel, soit pour analyser son comportement, soit pour trouver la façon de contrôler son évolution.

On sait, par exemple, que les pilotes des avions de ligne sont entraînés sur des "simulateurs de vol" qui recréent, à terre, toutes les conditions pratiques du pilotage : dans une cellule identique à celle de l'appareil réel, réactions des commandes, bruits, mouvements, sont reproduits le plus fidèlement possible. De même, les essais en soufflerie de maquettes d'avion ou de ponts permettent de simuler sans danger le vol réel ou les effets du vent sur les structures.

De la même façon, on peut dire que l'équation différentielle $x'' + k \sin x = f(t)$ permet de simuler, mathématiquement cette fois, le mouvement d'un pendule simple.

Si le hasard intervient dans le fonctionnement du système réel, il est nécessaire de fabriquer, dans le modèle de simulation, un élément qui reproduise ce hasard. Par exemple, pour étudier la gestion d'un péage d'autoroute (nombre de guichets à ouvrir, longueur des files d'attente,...), il faudra simuler les arrivées de véhicules en distinguant diverses catégories (véhicules légers, avec ou sans remorque, poids lourds,...), simuler la répartition entre conducteurs abonnés, ayant l'appoint, n'ayant pas de monnaie. Il faudra aussi simuler le temps de service de chaque guichetier. Par observation du phénomène réel ou par analogie avec des situations mieux connues, on pourra déterminer les lois de probabilité de ces divers éléments aléatoires, et trouver des espaces probabilisés $(\Omega,\mathcal{A},P)$ qui les modélisent. Il faudra ensuite être capable de choisir un élément de chaque ensemble Ω comme s'il avait été tiré suivant la loi de probabilité P correspondante.

Historiquement, l'ingéniosité des chercheurs a permis de mettre au point des systèmes physiques plus ou moins compliqués, simulant une loi de probabilité donnée. Les Loteries Nationales ont ainsi des systèmes très surveillés, beaucoup plus compliqués que les simples roulettes des casinos, qui simulent la loi uniforme sur des ensembles finis de nombres. On en est très vite arrivé à publier sous forme

de tables des suites de nombres obtenues par de tels systèmes et dits **nombres au hasard**. Puis, avec les développements de l'informatique, des générateurs de **"nombres pseudo-aléatoires"** utilisant des programmes appropriés ont été mis au point. Ces tables ou ces générateurs fournissent, en général, des nombres que l'on peut considérer comme tirés indépendamment les uns des autres suivant la loi uniforme sur un ensemble fini.

Après avoir rappelé ce que sont les tables de nombres pseudo-aléatoires et indiqué brièvement leur utilisation, nous décrirons les principaux types de générateurs actuellement utilisés. Nous présenterons ensuite les principales techniques pour construire une observation d'une loi de probabilité donnée à partir d'observations de la loi uniforme sur le segment [0, 1[. Nous terminerons par une description des outils statistiques les plus courants pour juger de la "qualité" des suites de nombres pseudo-aléatoires fournies par une table ou un générateur.

II. LES TABLES DE NOMBRES AU HASARD.

Une table de nombres au hasard se présente en général comme une suite de digits, soit de 0 et de 1, soit de nombres 0, 1, ..., 9. Ces digits sont considérés comme résultant de tirages indépendants de la loi uniforme sur {0,1} ou sur {0, 1, ... ,9} suivant les cas. Les tables les plus importantes que l'on puisse utiliser sont

- La table de M.G. KENDALL et B. BABINGTON-SMITH publiée en 1937-1938 et qui contient 100 000 digits décimaux,

- La table de la RAND CORPORATION publiée en 1955 qui contient 1 000 000 de digits décimaux.

La plupart des tables statistiques publient, soit des extraits de l'une ou de l'autre de ces deux tables, soit des tables originales, mais en général beaucoup plus courtes.

Pour utiliser une table de nombres au hasard, on choisit un des nombres de la table et on lit les suivants, soit en ligne soit en colonne. On obtient ainsi une suite de nombres que l'on considère comme tirés indépendamment les uns des autres suivant une loi uniforme sur l'ensemble des digits utilisés par la table.

En prenant comme exemple une table de nombres au hasard utilisant les

digits 0,1, ... ,9, nous allons montrer briévement comment on peut simuler des tirages de la loi uniforme sur des ensembles autres que {0,1, ... , 9}.

1) pour fabriquer un tirage suivant la loi uniforme sur l'ensemble {1,2,...,n}, on procéde ainsi :

• si $n < 9$, on ne garde que les nombres au hasard de la table inférieurs à n,

• si $10^{k-1} \leq n < 10^{k} - 1$, on prend les nombres de la table par paquets de k nombres. A chaque paquet $u_1, u_2, \ldots, u_k$, on associe l'entier $u_1 + 10\, u_2 + \ldots + 10^{k-1}\, u_k$, et on ne garde de ces entiers que ceux qui sont inférieurs à n.

2) pour fabriquer un tirage suivant la loi uniforme sur le segment [0,1[, on réalise en fait un tirage sur l'ensemble des multiples de 10^{-k}, pour un k suffisamment grand en fonction de l'utilisation envisagée. On utilise encore les nombres de la table par paquets de k nombres, en associant cette fois au paquet $u_1, u_2, \ldots, u_k$ le décimal $u_1\, 10^{-1} + u_2\, 10^{-2} + \ldots, + u_k\, 10^{-k}$.

III. LES GENERATEURS DE NOMBRES PSEUDO-ALEATOIRES.

Les tables étant peu utilisables dans les systèmes informatiques à cause de l'encombrement mémoire et de la lenteur d'accés qu'elles entraînent, on préfère les remplacer par un sous-programme qui fournit, à chaque appel, un nombre appartenant au segment [0,1[(qui sera, en fait, toujours un décimal ou un nombre diadique). Comme pour les tables, les nombres obtenus doivent pouvoir être considérés comme indépendants et de loi uniforme sur [0,1[.

Les sous-programmes utilisent des algorithmes de deux types :

- les **algorithmes directs**, qui calculent le $n^{ième}$ nombre en fonction de n,
- les **algorithmes itératifs**, qui calculent le $n^{ième}$ nombre en fonction du ou des précédents.

Les suites de nombres ainsi construites, étant en fait déterministes, ne présentent a priori pas le caractère aléatoire que l'on attend. Seul le choix judicieux de l'algorithme permet d' "imiter" le comportement des tirages

indépendants d'une loi uniforme sur [0,1[. On indique ce caractère déterministe en disant que les nombres construits par de tels algorithmes sont des **nombres pseudo-aléatoires**. Les algorithmes sont alors appelés **générateurs de nombres pseudo-aléatoires**.

Les algorithmes directs sont peu utilisés et nous nous limiterons ici à la description des principaux algorithmes itératifs.

III.1. LES CONGRUENCES LINEAIRES.

Soit a, b, m trois entiers vérifiant $1 \leq a < m$, $0 \leq b < m$. Posons $T(x) = (ax + b) \bmod m$. On choisit x_0 appartenant à l'ensemble $E = \{ 0,1,2,\ldots, m-1 \}$ et on pose

$$x_n = T^n(x_0)$$

et

$$u_n = \frac{1}{m} x_n$$

où $T^n = T \cdot T^{n-1}$. Par un choix convenable des entiers a, b, m, x_0, on peut obtenir une suite u_n de nombres pseudo-aléatoires qui possèdent les propriétés qu'on souhaite.

Puisque l'ensemble E est un ensemble fini, la suite des itérés de x_0 par T ne prend qu'un nombre fini de valeurs et il existe deux entiers p et q tels que

- si $x_1 \neq x_0$:
  1) $0 \leq p, 1 \leq q$
  2) $x_0, \ldots, x_{p+q-1}$ sont distincts deux à deux
  3) $x_{p+q} = x_p$,

- si $x_1 = x_0$:

  $p = 0,\ q = 1$

Si $p > 0$, l'ensemble $\{x_0,\ldots,x_{p-1}\}$ est appelé **partie transitoire** ou **transitoire** de la suite $\{x_n\}$. L'ensemble $\{x_p,\ldots,x_{p+q-1}\}$ est appelé **cycle limite** ou **cycle** de la suite $\{x_n\}$. L'entier p est la **longueur du transitoire** (trapping time), l'entier q est la **longueur du cycle** ou **période**.

On voit que le transitoire peut être vide et dans ce cas sa longueur est nulle, que le cycle limite peut être réduit à un élément. Un point fixe de T est ainsi un cycle de longueur 1. Si PGCD (a,m) = 1, l'application T est inversible et dans ce cas, pour tout $x_0 \in E$, la suite $\{T^n(x_0)\}$ a un transitoire vide.

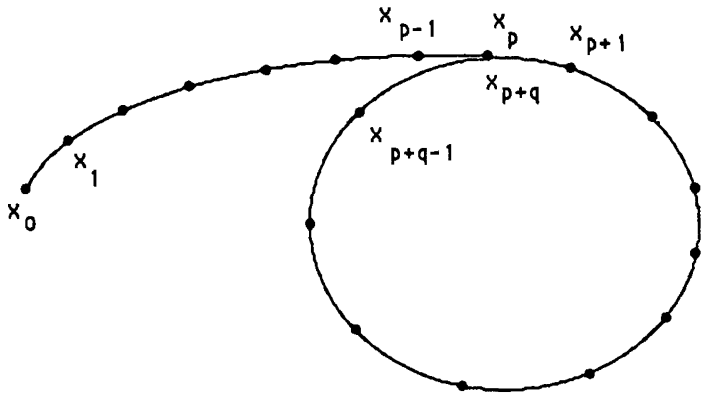

Figure 1

On cherchera alors évidemment à construire des suites sans transitoire et de cycle le plus long possible, en espérant ainsi que, lors de l'utilisation de la suite $\{u_n\}$, le nombre de tirages de nombres pseudo-aléatoires effectués sera inférieur à la longueur du cycle ; il est évident que si on parcourt plusieurs fois un cycle, on ne pourra plus considérer les nombres tirés comme indépendants. Par ailleurs, si le cycle est strictement contenu dans E, on ne peut pas garantir a priori avoir des tirages de loi uniforme sur le segment [0,1[. L'étude complète des transitoires et des cycles des générateurs est donc indispensable.

On démontre que le transitoire est vide si et seulement si a et m sont premiers entre eux. L'étude des cycles est plus délicate. Nous n'en résumons ici que l'essentiel.

Désignons par L (a, b, m, x_0) la longueur du cycle de la suite des entiers x_n définis par $x_n = T^n(x_0)$. On suppose maintenant PGCD (a,m) = 1. On montre en

premier lieu que

$$L(a, b, m, x_0) = L(a, 1, m_1, 0)$$

où

$$m_1 = \frac{m}{PGCD\ ((a-1)\, x_0 + b,\ m)}$$

Un générateur pour lequel $b = 1$ et $x_0 = 0$ est dit **générateur fondamental** et le cycle qu'il engendre est dit **cycle fondamental.**

On montre alors que pour un cycle fondamental de longueur n il existe un entier $t \in \{0, 1, \ldots, m-1\}$ tel que

i) $x_0, x_1, \ldots, x_{t-1}$ sont distincts deux à deux

ii) $\forall\ k \in \mathbb{N},\ \forall\ i \in \{0, \ldots, m-1\},\ x_{k\,t+i} = k\, x_t + x_i \mod m$

iii) il existe un entier r tel que $L(a, 1, m, 0) = n = r\,t$

et on montre que

t est le plus petit entier non nul tel que $a^t = 1 \mod m$

r est le plus petit entier non nul tel que $ar = 0 \mod m$

On voit ainsi la signification de t : tous les éléments du cycle $\{0, x_1, \ldots, x_{n-1}\}$ se déduisent par translation par $x_t \mod m$ de la suite $\{0, x_1, \ldots, x_{t-1}\}$. Pour cette raison, t s'appelle la **période effective du générateur.**

Une étude plus complète permet le calcul de t, r, $L(a, 1, m, 0)$ pour tous les couples (a,m) pour lesquels $PGCD(a,m) = 1$. Pour de plus amples détails, on pourra consulter B. Van Cutsem (1980) et (1985), dans lesquels on trouvera une démonstration des résultats connus et des compléments au travail de G. Marsaglia (1975).

Un des résultats essentiels est le théorème dû à T.E. Hull et A.R. Dobell (1962) :

"Le cycle fondamental est de longueur maximum égale à m si et seulement si les trois conditions suivantes sont vérifiées :

i) PGCD (a,m) = 1, PGCD (b,m) = 1

ii) si p est un nombre premier qui divise m, alors p divise a-1,

iii) si 4 divise m, alors 4 divise a-1 ".

Voici maintenant quelques exemples de générateurs :

$$T(x) = 3^{15} x \mod 2^{32}, \quad x_0 = 2^{31} - 1$$

$$T(x) = (3+2^{16})\, x \mod 2^{31}, \quad x_0 \text{ impair}$$

$$T(x) = 13^{13}\, x \mod 2^{59}, \quad x_0 \text{ impair}$$

$$T(x) = 24\,298\, x + 99\,991 \mod 199\,017$$

dont les périodes respectives sont respectivement:

$2^{30} = 1\,073\,741\,824$

$2^{29} = 536\,870\,912$

$2^{57} = 144\,115\,188\,075\,855\,872$

199 017

III.2. LES CONGRUENCES LINEAIRES EN DIMENSION k.

Des travaux récents ont abordé l'étude des congruences vectorielles $T(x) = A\, x + b \mod {}^{t}(m, m, ..., m)$ où $x \in \mathbb{N}^k$, $b \in \mathbb{N}^k$, $m \in \mathbb{N}$, et où A est une matrice à termes entiers à k lignes et k colonnes.

On sait calculer les périodes et périodes effectives dans les cas les plus simples. On n'a pas, à ma connaissance, de résultats sur les tests de qualité pour

de tels générateurs. On pourra consulter Tahmi El Hadj (1982) sur ces questions.

III.3. LES RECURRENCES D'ORDRE SUPERIEUR.

Divers auteurs se sont intéressés à des générateurs définis par des récurrences du genre :

$$x_{n+1} = T(x_n, \dots, x_{n-k+1})$$

$$= a_0 x_n + \dots + a_{k-1} x_{n-k+1} + b \mod m$$

et ont étudié la suite des itérés de $(x_0, \dots, x_{k-1})$ par l'application T. Les suites de Fibonacci sont des cas simples de telles suites et ont été les premiers exemples abordés. Il est évident aussi que les congruences linéaires en dimension k sont des généralisations des récurrences d'ordre supérieur.

On possède des résultats fragmentaires sur le calcul des périodes des suites engendrées par de tels générateurs. (Tahmi El hadj (1982) et A. Vince (1981))

Le générateur de ce type le plus connu est sans doute celui proposé par R.C. Tausworthe (1965). Ce générateur utilise la récurrence

$$x_n = a_k x_{n-k} + \dots + a_1 x_{n-1} \mod 2$$

où a_k est égal à 1 et les autres coefficients a_i sont égaux à 0 ou à 1. R.C. Tausworthe a montré que pour un tel générateur, la période maximum est égale à $2^k - 1$ et est atteinte si le polynôme

$$P(x) = 1 + a_1 x + \dots + a_k x^k$$

est un polynôme primitif sur le corps de Galois $K = \{0,1\}$. Les nombres pseudo-aléatoires proposés par Tausworthe sont alors les nombres y_n définis par l'écriture en base 2 suivante :

$$y_n = x_{(n-1)L+1} 2^{L-1} + x_{(n-1)L+2} 2^{L-2} + \dots + x_{(n-1)L+L} 2^0$$

pour un L déterminé. La suite des nombres $(y_n / 2^L)$ peut alors être considérée comme une suite de nombres pseudo-aléatoires sur le segment [0,1[si L et $2^k - 1$ sont premiers entre eux. De plus si hL est premier avec $2^k - 1$ les vecteurs $(u_{nh+1}, \dots, u_{nh+h})$ sont approximativement de loi uniforme sur $[0, 1[^h$.

III. 5. D'AUTRES APPROCHES.

Des méthodes plus ou moins sophistiquées ont également été proposées. Citons, parmi celles-ci les méthodes qui consistent à :

i) permuter les nombres pseudo-aléatoires par paquets ; on permute chaque paquet (de 128 nombres par exemple) en utilisant une permutation fixée une fois pour toutes, et on utilise les nombres pseudo-aléatoires dans l'ordre où ils apparaissent après permutation. On espère ainsi se rapprocher de la situation idéale où les tirages sont indépendants.

ii) utiliser plusieurs générateurs en parallèle et passer de l'un à l'autre. On espère ainsi allonger les périodes.

Ces méthodes ont été contrôlées expérimentalement et il n'y a pas, à ma connaissance, d'étude théorique très poussée les concernant.

IV. SIMULATION DES LOIS DE PROBABILITE.

Nous avons vu, à propos des tables de nombres au hasard, comment simuler une loi uniforme sur un ensemble. On peut, de façon analogue, partant d'une suite de réels $\{ u_n \}$ fournie par un générateur de nombres pseudo-aléatoires, construire une suite $\{ v_n \}$ d'observations pseudo-aléatoires sur l'ensemble $\{ 0,1,\dots,k-1 \}$ en posant, pour tout n :

$$v_n = \text{INT}(ku_n)$$

où INT (a) désigne la partie entière du réel a.

On voit ainsi comment on peut à l'aide d'une suite de nombres pseudo-aléatoires simuler les lois uniformes. Cette technique se généralise de

diverses façons, à quasiment toutes les autres lois de probabilités, et nous allons maintenant en présenter les principales méthodes. On supposera dans chaque cas disposer d'une suite de nombres pseudo-aléatoires source dans [0, 1[, qu'elle soit fournie par une table ou par un générateur. On cherchera, à chaque fois à construire une suite de nombres $\{ x_n \}$ que l'on puisse considérer comme un échantillon de la loi de probabilité que l'on veut simuler.

IV. 1. SIMULATION D'UNE LOI DISCRETE.

Soit $A = \{ a_1, ..., a_n, ... \}$ un ensemble discret et soit P une loi de probabilité sur A. Posons, pour tout entier n, $p_n = P(\{a_n\})$. Pour simuler P, on associe à tout nombre pseudo-aléatoire source u_n, l'entier i défini par :

$$p_1 + ... + p_{i-1} < u_n \leq p_1 + ... + p_i$$

(avec $i = 1$ si $u_n \leq p_1$). Au nombre u_n est ainsi associée l'observation a_i dans E.

On prendra garde au fait que, si on utilise comme générateur une congruence modulo m, les nombres u_n sont des multiples de 1/m , ce qui exclut l'observation de certains événements de probabilité inférieure à 1/m. Par exemple, en simulant par cette méthode la loi de Poisson de paramètre 3 et en utilisant un générateur pour lequel $m = 199117$, la plus grande valeur observable est 12.

IV.2. SIMULATION D'UNE PROBABILITE SUR $\mathbb{R}$ DE FONCTION DE RÉPARTITION CONNUE.

Soit F la fonction de répartition de la probabilité P sur $\mathbb{R}$ que l'on souhaite simuler. Soit G la fonction définie sur [0,1[par

$$G(y) = \text{Inf} \{ x \in \mathbb{R} ; y \leq F(x) \}$$

On montre facilement que si U est une variable aléatoire de loi uniforme sur le segment [0, 1[, alors la variable aléatoire X = G(U) a une loi de probabilité égale à P. On utilise ce résultat pour simuler P en associant au nombre pseudo-aléatoire source u_n, l'observation $x_n = G(u_n)$.

On peut faire pour cette méthode la même remarque que précédemment en ce qui concerne l'inobservation de certains événements de probabilité inférieure à 1/m. (On pourra d'ailleurs vérifier que ces deux méthodes sont identiques).

IV.3.SIMULATION D'UNE LOI IMAGE DE LA LOI UNIFORME SUR $[0,1[^h$.

Chaque fois que l'on peut définir une probabilité P comme image de la loi uniforme sur le cube $[0,1[^h$ par une application X, on peut simuler P en prenant les nombres pseudo-aléatoires source par paquets de h nombres $(u_{nh+1}, \cdots, u_{nh+h})$ et en posant

$$x_n = X(u_{nh+1}, \cdots, u_{nh+h})$$

L'application la plus connue de cette technique est la méthode de Box-Muller pour simuler la loi normale $\mathcal{N}(0, \mathbf{1}_2)$ sur $\mathbb{R}^2$. On utilise le fait que cette loi est l'image de la loi uniforme sur le carré $[0, 1[^2$ par l'application

$$X(u,v) = \begin{bmatrix} \sqrt{-2 \operatorname{Log} u}\ \cos 2\pi v \\ \sqrt{-2 \operatorname{Log} u}\ \sin 2\pi v \end{bmatrix}$$

Cette méthode permet aussi de simuler la plupart des lois usuelles. Ainsi, avec des notations usuelles,

la loi $\mathcal{U}([a,b])$ est l'image de la loi $\mathcal{U}([0,1])$ par $x = a + (b-a)\,u$

la loi $\mathcal{N}(m,\sigma^2)$ est l'image de la loi $\mathcal{N}(0,1)$ par $x = m+\sigma\,u$

la loi $\mathcal{N}_k(m, \Lambda)$ est l'image de la loi $\mathcal{N}_k(0,1_k)$ par $x = m + \Lambda^{1/2}\, u$

la loi $\Gamma(k/2, 2)$ est l'image de la loi $\mathcal{N}_k(0,1_k)$ par $x = ||u||^2$

et on peut prolonger cette liste au cas des lois Bêta, aux lois Gamma et Bêta décentrées, aux lois de Student, ...

On peut encore considérer comme un prolongement de cette méthode, la technique qui consiste à simuler une probabilité P en simulant en fait une probabilité Q qui approche P, quand la simulation de Q est aisée. Ainsi une des méthodes les plus courantes pour simuler la loi normale $\mathcal{N}(0,1)$ utilise le théorème de la limite centrale. Si $U_1, \ldots, U_k$ sont k variables aléatoires indépendantes de loi uniforme sur [0,1[, on sait que, pour k suffisamment grand, on peut assimiler la loi de la somme $U_1 + \ldots + U_k$ à une loi normale. Plus précisément, la variable X_k définie par

$$X_k = \frac{U_1 + \ldots + U_k - (k/2)}{\sqrt{(k/12)}}$$

est asymptotiquement normale centrée réduite. On choisit dans la pratique $k = 12$ et on simule une observation x de la loi $\mathcal{N}(0,1)$ en posant $x = u_1 + \ldots + u_{12} - 6$, où les $u_1, \ldots, u_{12}$ sont 12 nombres pseudo-aléatoires source.

IV.4. METHODE DE REJET.

Cette méthode, plus sophistiquée, s'utilise pour simuler les lois de probabilité sur $\mathbb{R}$ de support borné et définies par une densité bornée. Elle repose sur le théorème suivant :

"Soit $I = [a,b[$ un intervalle borné de $\mathbb{R}$ et soit P une probabilité sur I définie par la densité f. On suppose qu'il existe un réel positif M tel que $\sup \{f(x) ; x \in I\} \leq M$. Soit U_n une suite de variables aléatoires indépendantes de loi uniforme sur [0,1[. Posons, pour tout entier n :

$$\begin{bmatrix} V_n \\ W_n \end{bmatrix} = \begin{bmatrix} a + (b-a)\, U_{2n} \\ M\, U_{2n+1} \end{bmatrix}$$

et soit X l'application à valeurs dans I définie par

$$K = \inf \{ n \in \mathbb{N};\ W_n \leq f(V_n) \}$$
$$X = V_K$$

Alors la loi de X est la loi P. "

La méthode de simulation de P se déduit immédiatement de ce théorème. A la suite de nombres pseudo-aléatoires source $\{u_n\}$, on associe la suite $\{ (v_n, w_n) \}$ définie par $v_n = a + (b-a)\, u_{2n}$ et $w_n = M\, u_{2n+1}$, et on détermine le plus petit entier k tel que $w_k \leq f(v_k)$. On pose alors $x = v_k$.

IV.5. METHODE DE DECOMPOSITION.

Cette méthode utilise une décomposition de la loi de probabilité P que l'on veut simuler, par rapport à un noyau de transition. Plus précisément, si P est une probabilité sur $\mathbb{R}^k$ et si on connait une loi de probabilité Q sur $\mathbb{R}^p$ et un noyau de transition K de $\mathbb{R}^p$ vers $\mathbb{R}^k$ tel que

$$P = \int_{\mathbb{R}^p} K(\,.\,,x)\, dQ(x)$$

on simule la probabilité P en deux temps : on fait, par simulation, une observation x de la loi Q dans $\mathbb{R}^p$, puis, par simulation encore, une observation de la loi $K(\,.\,, x)$ dans $\mathbb{R}^k$.

IV.6. METHODE DE DECOMPOSITION - REJET.

Cette méthode est une utilisation alternée des deux méthodes précédentes. On utilise la méthode de rejet lors des simulations des probabilités $K(.,x)$.

Ce résumé des méthodes pratiques de simulation est nécessairement incomplet. Il existe bien des méthodes, toutes plus astucieuses les unes que les autres, pour simuler, efficacement ou rapidement, telle ou telle loi de probabilité particulière. Les lois usuelles de probabilité sont particulièrement riches de telles méthodes qui leur sont spécialement adaptées.

V. COMMENT VERIFIER LES QUALITES D'UN GENERATEUR ?

Comme nous l'avons indiqué dans l'introduction, on souhaite que les nombres pseudo-aléatoires successifs fournis par un générateur ou un table puissent être considérés comme une suite de tirages indépendants d'une loi uniforme sur $[0, 1[$.

Le caractère déterministe du procédé de calcul de u_{n+1} en fonction de u_n ou de $(u_n, u_{n-1}, \ldots)$ va évidemment à l'encontre d'une telle exigence. Les générateurs doivent donc être conçus pour qu'on puisse raisonnablement admettre que les suites de nombres qu'ils engendrent puissent être assimilés à des échantillons de la loi uniforme sur $[0, 1[$. On peut résumer les qualités exigées d'un générateur en deux aspects :

i) les tirages successifs doivent pouvoir être considérés comme tirés suivant la loi uniforme sur $[0, 1[$,
ii) les tirages successifs doivent pouvoir être considérés comme indépendants.

Les défauts des générateurs peuvent donc être envisagés comme des écarts avec l'une et l'autre de ces hypothèses. Ces défauts d'uniformité et d'indépendance peuvent en effet avoir des effets désastreux pour certaines applications. Deux soucis majeurs apparaissent ainsi

i) on souhaite pouvoir contrôler la "qualité" des suites fournies par un générateur ou une table,

ii) on souhaite pouvoir concevoir des générateurs dont on puisse garantir une "qualité", éventuellement en précisant qu'elle n'est assurée que pour des suites de longueur inférieure à une borne donnée.

On mesure tout de suite la difficulté de "quantifier" ces défauts. La première idée qui vient à l'esprit est d'utiliser les tests statistiques usuels d'adéquation d'un échantillon à la loi uniforme sur $[0, 1[$, ou des tests d'indépendance. En fait, on peut inventer de multiples tests statistiques et les statisticiens ont fait preuve d'une grande imagination dans ce domaine. On verra plus loin une liste de quelques uns de ces tests.

Toutes ces techniques d'évaluation font usage de la suite $\{u_0, ..., u_{n-1}\}$ comme d'un échantillon pour tester la loi $\mathcal{U}([0, 1[^n)$ sur $\mathbb{R}^n$ contre l'une ou l'autre famille de lois sur $\mathbb{R}^n$. Elles sont regroupées depuis D. E. Knuth (1969) sous le nom de **tests empiriques**. On leur oppose sous le nom de **tests théoriques**, les techniques qui, pour un critère de qualité donné (par exemple par un test empirique), cherchent les paramètres de l'application T qui définit le générateur, de telle façon qu'une suite $\{u_n\}$ fournie par ce générateur conduise à accepter la loi uniforme $\mathcal{U}([0, 1[^n)$ pour ce critère.

Les résultats connus sur les tests théoriques sont beaucoup plus fragmentaires et se limitent, pour une congruence linéaire T, à l'étude des suites dont la longueur du cycle est égale à la longueur maximum m. Nous en citerons quelques exemples.

Terminons l'introduction de ce paragraphe par une remarque importante. On utilise souvent, on l'a vu au paragraphe IV, les nombres pseudo-aléatoires par paquets de k nombres, chacun de ces paquets servant à calculer une observation d'une loi donnée. Les deux hypothèses suivantes sont alors primordiales :

i) chacun des vecteurs $(u_{nk+1}, ..., u_{nk+k})$ doit pouvoir être considéré comme tiré suivant la loi $\mathcal{U}([0, 1[^k)$,

ii) les vecteurs successifs doivent pouvoir être considérés comme indépendants.

Ces deux hypothèses sont évidemment équivalentes à celles exprimées plus haut avec k = 1. Mais il faut savoir que, pour les congruences linéaires, on peut avoir une bonne adéquation des échantillons aux lois $\mathcal{U}([0, 1[^k)$ pour les petites valeurs de k (k = 1, 2, 3, ...), mais que les choses se dégradent très vite quand k augmente. On vérifie, en effet, que les points M_n de $\mathbb{R}^k$ de coordonnées $(u_{nk+1}, .. , u_{nk+k})$ sont situés dans des hyperplans parallèles de plus en plus éloignés les uns des autres quand k augmente.

On désignera, dans la suite, par hypothèse nulle, l'hypothèse que toute suite $\{ u_0, ..., u_n\}$ est un échantillon de la loi uniforme sur [0, 1[.

V. 1. TESTS EMPIRIQUES.

Il y a une liste impressionnante de tels tests. Nous n'en citerons que quelques uns.

V. 1. a. TEST DU CHI - DEUX.

a) On partage [0, 1[en r sous ensembles disjoints, et on vérifie l'adéquation de la suite $\{ u_0, ..., u_n\}$ à la loi $\mathcal{U}([0, 1[)$ par un test classique du Chi-deux.

b) On partage $[0, 1[^k$ en r sous ensembles disjoints, et on vérifie l'adéquation de la suite $\{ (u_{ik+1} ,..., u_{ik+k}) ; 0 \le i \le n \}$ à la loi $\mathcal{U}([0, 1[^k)$ par le test du Chi-deux.

V. 1. b. TESTS DE KOLMOGOROV - SMIRNOV, CRAMER - VON MISES, ...

On teste l'adéquation de la suite $\{ u_0, ..., u_n\}$ à la loi $\mathcal{U}([0, 1[)$ par l'un des tests classiques utilisant la fonction de répartition empirique.

V. 1. c. TESTS UTILISANT LA "CONFIGURATION" DE $(u_{nk+1}, \ldots, u_{nk+k})$.

Tous les tests que nous regroupons sous ce nom, utilisent l'idée que si les points M_n de $\mathbb{R}^k$ de coordonnées $u_{nk+1}, \ldots, u_{nk+k}$, sont uniformément répartis dans $\mathbb{R}^k$, la fréquence d'appartenance de ces points à certains sous ensembles particuliers de $[0, 1[^k$, doit être proche de leur probabilité pour la loi uniforme.

V.1.c.1. Test du poker (M. G. Kendall et B. Babington-Smith (1938-1939))

On partage $[0, 1[$ en r sous ensembles disjoints $A_1, \ldots, A_r$ de même longueur. A la suite finie $u_{nk+1}, \ldots, u_{nk+k}$, on associe le mot de k lettres $A_{i_1}, \ldots, A_{i_k}$ où A_{i_j} désigne le sous ensemble $A_1, \ldots, A_r$ qui contient u_{nk+j}.
On dénombre alors les mots où

- toutes les lettres sont distictes,

- il existe une paire de lettres identiques,

- il existe deux paires de lettres identiques,

- ...

(on voit l'analogie avec le jeu du poker), et on compare ces fréquences à leurs probabilités sous l'hypothèse nulle.

V. 1.c.2. Test des permutations.

A chaque terme u_{nk+i} de la suite finie $u_{nk+1}, \ldots, u_{nk+k}$, on associe son rang R_i dans la suite des u_{nk+j} ordonnée par ordre croissant. En supposant que tous les u_n sont distincts, ces rangs définissent une permutation σ des entiers 1,...,k (en posant $R_i = \sigma(i)$).

On teste alors, par un test du Chi-deux par exemple, que les permutations observées sont uniformément réparties dans l'ensemble des permutations de {1,...,k}.

V. 1. c.3. Test des figures fondamentales (J.R. Barra (1972)).

On appelle figure d'ordre k un sous ensemble A de $[0, 1[^k$ tel que

- A soit non vide
- il n'existe pas de partie stricte B de $[0, 1[^{k-1}$ telle que

$$A = B \times [0, 1[\quad \text{ou} \quad A = [0, 1[\times B$$

On dénombre alors les suites

$$(u_0, \dots, u_{k-1}), (u_1, \dots, u_k), \dots, (u_{n+1}, \dots, u_{n+k})$$

qui appartiennent à A.

Le test, assez sophistiqué, utilise la probabilité de A sous l'hypothèse nulle. De plus, si k est une puissance de 2, on peut définir un système fondamental de figures A dont les fréquences sont indépendantes sous l'hypothèse nulle.

V.1.d. TESTS UTILISANT DES TEMPS D'ARRET DE LA SUITE $\{u_n\}$

V.1.d.1. Gap test (M. G. Kendall et B. Babington-Smith (1938-1939))

Soit A un sous ensemble strict de [0, 1[. On appelle X le nombre de termes de la suite $\{u_n\}$ qui séparent deux apparitions successives de l'événement $\{u_i \in A\}$. On compare alors les observations de X à leur loi, facilement calculable sous l'hypothèse nulle (par un test du Chi-deux par exemple). On peut, de plus, étudier la loi de X sous une hypothèse markovienne (pour les états A et $[0, 1[- A_j$) et définir un test de l'hypothèse nulle contre une hypothèse

markovienne. On sait en outre calculer la puissance de ce test.

V.1.d.2. Test du collectionneur (Greenwood (1955))

Soit $A_1, \ldots, A_r$ une partition de $[0, 1[$ en r sous ensembles non vides. Soit X le plus petit entier n tel que dans la suite $(u_0, \ldots, u_{n-1})$, on observe chacun des événements $A_1, \ldots, A_r$. On compare alors les observations de X à leur loi sous l'hypothèse nulle, par un test du Chi-deux, par exemple.

V.1.d.3. Test de la plus longue suite (J.R. Barra (1966)).

Soit A un sous ensemble strict de $[0, 1[$. On calcule les "durées de séjour" de la suite $\{u_0, \ldots, u_n\}$ dans l'ensemble A, et on appelle X la durée de séjour maximum observée. On compare les observations de X à leur loi de probabilité sous l'hypothèse nulle. On peut aussi calculer la loi de X sous une hypothèse markovienne et déterminer un test de l'hypothèse nulle contre cette hypothèse markovienne. Un calcul de la puissance est possible.

V. 1.e. TEST SPECTRAL (Coveyou et Mac Pherson (1967)).

Ce test est conçu pour des digits aléatoires sur un ensemble fini $E = \{0, 1, 2, \ldots, m-1\}$. Il utilise la transformée de Fourier de la suite $\{(u_{ik+1}, \ldots, u_{ik+k}) ; 0 \leq i \leq n\}$ pour évaluer son adéquation à la loi uniforme sur E^k.

Cette transformée de Fourier est une fonction sur E égale à 1 à l'origine et à 0 partout ailleurs, sous l'hypothèse nulle. Les seuils de ce test actuellement proposés dans la littérature sont choisis expérimentalement sans référence à des calculs précis ou approchés.

V.1.f. TESTS DIVERS.

Il existe encore d'autres tests, qui n'entrent pas dans les catéories précédentes. Les plus connus sont sans doute le test des séquences, qui utilise le nombre de sous suites monotones croissantes qu'on peut extraire de $\{u_0, \ldots, u_n\}$, et le test du coefficient de corrélation sérielle qui utilise le coefficient de corrélation entre $(u_0, \ldots, u_n)$ et $(u_1, \ldots, u_{n+1})$.

V.2. TESTS THEORIQUES.

Ces tests ne concernent, pour l'essentiel, que les congruences linéaires en dimension un. Quelques applications plus récentes concernent le générateur de Tausworthe.

Soit $T(x) = a\,x + b \mod m$, une congruence linéaire et soit $x_0 \in E = \{0, 1, \ldots, m-1\}$. On souhaite déterminer les ensembles de paramètres (a, b, m, x_0) tels que la suite de nombres pseudo-aléatoires engendrés entraîne l'acceptation de l'hypothèse nulle par l'un des tests empiriques précédents. Le projet est ambitieux et on en connait que quelques résultats fragmentaires. On sait ainsi calculer en fonction de a, b, m, x_0

i) Card $\{ (u_i, u_{i+1}) ; u_i \leq u_{i+1}$ et $0 \leq i \leq m-1 \}$

ii) le coefficent de corrélation entre entre $(u_0, \ldots, u_{m-1})$ et $(u_1, \ldots, u_{m+1})$

iii) la transformée de Fourier de la suite $(x_0, \ldots, x_{k-1})$, $(x_k, \ldots, x_{2k-1})$, ...

si T est un générateur n'ayant qu'un seul cycle de longueur maximum m.

Ces résultats permettent bien sûr de choisir a et b en fonction de m par exemple. Mais ils sont en plus très importants car il apportent des renseignements précieux sur la structure profonde des suites engendrées par les congruences linéaires.

Des études récentes (R. Blacher, 1983) ont, par exemple, mis en évidence le rôle joué par la décomposition en fractions continues de a / m dans l'étude de la "corrélation" entre les tirages successifs fournis par un générateur linéaire.

VI. CONCLUSIONS.

Si les générateurs implantés sur les gros matériels des centres de Calcul, éprouvés par l'expérience, et construits très sérieusement (ils sont souvent en multiprécision) peuvent être utilisés sans crainte, les petits nombres de bits des micro contraignent à des générateurs de période très courte (j'en ai vu de période 32 768 !), voire avec des périodes effectives encore plus courtes, et aussi à des corrélations relativement forte entre termes consécutifs.

La prétention de cet exposé est de mettre en évidence le fait qu'il ne faut absolument pas faire confiance a priori aux générateurs de nombres pseudo-aléatoires implantés par les constructeurs sur les micro-ordinateurs.

Il ne faut pas non plus négliger les phénomènes de troncature et d'arrondi. Les suites déduites des suites $\{u_n\}$, par troncature ou arrondi peuvent présenter des périodes notablement plus courtes que la suite initiale.

On doit aussi réaliser que la copie idéale du hasard n'existe pas. Les techniques de simulation fournissent des modèles approchés. Une excellente approximation est théoriquement possible : elle risque de coûter très cher.

VII. BIBLIOGRAPHIE.

Nous citons d'abord plusieurs ouvrages généraux sur la simulation et les générateurs de nombres pseudo aléatoires qui ont paru récemment :

P. BRATLEY, B.L. FOX, L.E. SCHRAGE, 1983. A guide to simulation. Springer Verlag.

E.J. DUDEWICZ, T.G. RALLEY, 1981. The handbook of random number generation and testing with TESTRAND computer code. American Sciences Press.

W.J. KENNEDY Jr, J.E. GENTLE, 1980. Statistical computing.
Marcel Dekker.

J.P.C. KLEIJNEN, 1974-1975. Statistical techniques in simulation.
Marcel Dekker . Part 1 : 1974, Part : 1975.

B.J.T. MORGAN, 1984. Elements of simulation.
Chapman and Hall.

et la référence plus ancienne mais toujours actuelle :

D.E. KNUTH, 1969. The art of computer. Vol. 2, chapitre 3,
Addison Wesley.

Ces ouvrages contiennent une bibliographie très riche. On trouvera, de plus, une bibliographie très complète dans les articles :

B.D. RIPLEY, 1983. Computer generation of random variables : a tutorial.
International Statistical Review, Vol 51, pp 301-319.

E.R. SOWEY, 1972-1978. A chronological and classified bibliography on random number generation and testing.
International Statistical Review. Vol. 40, pp. 355-371,1972.

A second classified bibliography on random number generation and testing,
International Statistical Review, Vol. 46, pp. 89-102, 1978 .

Nous complétons par les quelques références spécialisées que nous avons citées dans le texte :

J.R. BARRA, 1967. Contrôle statistique d'une suite de digits aléatoires.
Revue de Statistique Appliquée, Vol. XV, n°3, pp 31-42, 1967.

J.R. BARRA, 1972. Contrôle statistique des suites de digits aléatoires.
Revue de Statisque Appliquée, Vol. XIX, n°3, pp 19-26, 1971.

R. BLACHER, 1983. Quelques propriétés des congruences linéaires considérées comme générateurs de nombres pseudo-aléatoires.
Rapport de recherche n°345, Laboratoire TIM 3, Université Scientifique et Médicale de Grenoble.

R. BLACHER, 1983. Indicateurs de dépendance entre deux variables aléatoires fournies par le développement en série de la densité de probabilité.
Thèse de 3ème cycle. Université scientifique et Médicale de Grenoble. 30 mars 1983.

R.R. COVEYOU, R.D. MAC PHERSON, 1967. Fourier analysis of uniform random number generators.
J.A.C.M., Vol.14, pp 100-119.

R.E. GREENWOOD, 1955. Coupon collector's test for random digits.
M.T.A.C. Vol.9, pp 1-5.
(Errata in M.T.A.C. (1955), Vol.9, pp 224 and 229)

T.E. HULL, A.R. DOBELL, 1962. Random number generators.
Siam Review, Vol. 4, n°3, pp 230-254.

M.G. KENDALL, B. BABINGTON - SMITH, 1938.
Randomness and random sampling numbers.
J.R.S.S., A, Vol.101, pp 147-166.

M.G. KENDALL, B. BABINGTON - SMITH, 1939.
Second paper on random sampling numbers.
J.R.S.S., B, Vol 6, suppl. pp 51-61.

G. MARSAGLIA, 1972. The structure of linear congruential sequences.
in Application of numbers theory to numerical analysis,
Edité par S.K. ZAREMBA. Academic Press, , pp 249-285.

TAHMI EL HADJ, 1982. Contribution aux générateurs de vecteurs pseudo-aléatoires.
Thèse 3ème cycle, Université Houari Boumedienne, Alger.
8 juin 1982.

R.C. TAUSWORTHE, 1965. Random numbers generated by linear recurrence modulo 2.
Math. Comp. Vol.19, pp 201-209.

B. VAN CUTSEM, 1980. Etude de la période de la suite de nombres pseudo-aléatoires engendrée par une congruence linéaire.
Bulletin de l'Association des Professeurs de Mathématique de l'Enseignement Public, n°324.

B. VAN CUTSEM, 1985. Itérations des congruences linéaires et génération des nombres pseudo-aléatoires.
Rapport de recherche n°568, Laboratoire TIM 3, Université Scientifique et Médicale de Grenoble.

A. VINCE, 1981. Periode of a linear recurrence.
Acta Mathématica, Vol. XXXIX, pp 303-311.

Les deux tables de nombres pseudo-aléatoires que nous avons signalées ont pour références :

M.G. KENDALL, B. BABINGTON - SMITH, 1939.
A table of random sampling numbers.
Tracts for Computers n° 24. C.U.P.

RAND CORPORATION, 1955. A million random digits with 100,000 normal deviates. Glencoe Free Press.