Louis Laurencelle

HASARD
NOMBRES ALÉATOIRES ET MÉTHODE MONTE CARLO

Presses de l'Université du Québec

HASARD
NOMBRES ALÉATOIRES
ET MÉTHODE
MONTE CARLO

PRESSES DE L'UNIVERSITÉ DU QUÉBEC
Le Delta I, 2875, boulevard Laurier, bureau 450
Sainte-Foy (Québec) G I V 2M2
Téléphone : (418) 657-4399 • Télécopieur: (418) 657-2096
Courriel : puq@puq.uquebec.ca • Internet : www.puq.uquebec.ca

Distribution :

CANADA et autres pays
DISTRIBUTION DE LIVRES UNIVERS S.E.N.C.
845, rue Marie-Victorin, Saint-Nicolas (Québec) G7A 3S8
Téléphone : (418) 831-7474 / 1-800-859-7474 • Télécopieur : (418) 831-4021

FRANCE
DIFFUSION DE L'ÉDITION QUÉBÉCOISE
30, rue Gay-Lussac, 75005 Paris, France
Téléphone: 33 1 43 54 49 02
Télécopieur: 33 1 43 54 39 15

SUISSE
GM DIFFUSION SA
Rue d'Etraz 2, CH-1027 Lonay, Suisse
Téléphone : 021 803 26 26
Télécopieur: 021 803 26 29

Louis Laurencelle

HASARD
NOMBRES ALÉATOIRES ET MÉTHODE MONTE CARLO

2001

Presses de l'Université du Québec
Le Delta I, 2875, boul. Laurier, bur. 450
Sainte-Foy (Québec) Canada G1V 2M2

Données de catalogage avant publication (Canada)

Laurencelle, Louis, 1946- .
Hasard, nombres aléatoires et méthode Monte Carlo

Comprend des réf. bibliogr. et un index.

ISBN 2-7605-1120-0

1. Nombres aléatoires. 2. Monte Carlo, Méthode de. 3. Variables aléatoires. 4. Suites aléatoires. 5. Nombres aléatoires – Problèmes et exercices. 1. Titre.

QA274.L38 2001 519.2'3 C00-942129-7

Nous reconnaissons l'aide financière du gouvernement du Canada par l'entremise du Programme d'aide au développement de l'industrie de l'édition (PADIÉ) pour nos activités d'édition.

Nous remercions le Conseil des arts du Canada de l'aide accordée à notre programme de publication.

1 2 3 4 5 6 7 8 9 PUQ 2001 9 8 7 6 5 4 3 2 **1**

Dépôt légal — 1er trimestre 2001
Bibliothèque nationale du Québec / Bibliothèque nationale du Canada
Imprimé au Canada

Table des chapitres

Chapitre ***Page***

Chapitre *Page*

Chapitre

1 Nombres aléatoires et méthode Monte Carlo Introduction

1.1 Cet ouvrage traite des séries de nombres aléatoires, de leurs propriétés, leur test, leur génération, leur application dans l'estimation Monte Carlo. L'univers de l'homme moderne, si pétri de rationnalisme, si environné de technologie et d'artéfacts rassurants, est néanmoins imprégné de hasard, plus que ne l'était celui de nos ancêtres. La démographie, la médecine préventive, la météorologie, les assurances, les loteries, les sondages, tout cela a façonné notre appréhension des choses et des événements, de sorte que risque, probabilité, tendance moyenne, fluctuation sont maintenant des concepts bien répandus. Ces concepts reposent essentiellement sur ceux d'événement aléatoire et de série de résultats aléatoires.

Les nombres aléatoires nous confrontent directement dans les jeux de hasard et dans les loteries; les joueurs plus optimistes espèrent le miracle, les plus acharnés prétendent imposer leur martingale. Œuvrant du côté plus sûr, les actuaires équilibrent risques et bénéfices pour les placements boursiers, l'assurance-accident, l'assurance-vie. Certains épidémiologues, prenant le pouls d'une situation, cherchent à mesurer et anticiper le progrès d'une maladie dans la population. Sondeurs, enquêteurs, organismes de planification publique, firmes de marketing, tous font affaire avec une réalité diverse, imprécise et changeante, dont ils prennent connaissance par des collections d'observations et de mesures; le succès de leur travail dépend en bonne partie de leur capacité à bien saisir et décrire l'état actuel du système à l'étude et à prédire son état futur.

1.2 Nous prenons donc les séries de nombres aléatoires comme objet d'étude, mais nous les détachons de leurs applications pratiques afin d'en dégager les caractères universels. C'est ainsi que, sur le plan conceptuel, nous réfléchirons à ce que peuvent signifier les expressions « série de nombres aléatoires », voire « nombre aléatoire ». Par exemple, est-ce que «6,3» est un nombre aléatoire? Est-ce que la série « 4 2 6 4 » est aléatoire? Nous recenserons aussi différents aspects et différentes techniques pour tester le caractère aléatoire d'une série de nombres, d'une source de nombres supposément aléatoires.

En fait, au-delà du fait que les séries de nombres aléatoires imprègnent des portions significatives de notre économie et de notre culture, elles sont aussi, par elles-mêmes, un objet d'intérêt scientifique et un outil, qui ont donné lieu à plusieurs développements en statistique, en mathématiques et en sciences appliquées. Grâce aux plateformes informatiques, à présent accessibles à tous et efficaces de plus en plus, la manipulation de grandes séries de nombres, leur calcul, leur génération automatique même sont devenus monnaie courante et ont permis l'éclosion de méthodes de calcul et de procédés de solution originaux. C'est le cas de la « méthode Monte Carlo », une utilisation massive des capacités de calcul de l'ordinateur pour trouver des solutions numériques approximatives à des problèmes complexes.

1.3 Nous examinons donc les procédés informatiques destinés à « produire du hasard », c'est-à-dire les méthodes de génération par ordinateur de valeurs successives d'une variable aléatoire ayant des propriétés de distribution déterminées. Cet examen fait apparaître une richesse foisonnante d'idées, de principes généraux, d'astuces, et il permet de revoir les principales lois de distribution statistiques, notamment les lois uniforme, normale, binomiale, *t* de Student, pour ne citer que celles-là.

1.4 Le traitement d'une ou de plusieurs séries de nombres aléatoires peut obéir à différents buts. On peut s'interroger, par exemple, sur la présence de facteurs structurants dans les séries, facteurs dont l'opération contrecarre ou modère le caractère purement aléatoire qu'on s'attend d'y trouver. Le diagnostic d'un tel facteur passe par le test des propriétés statistiques de la série. On peut aussi vouloir découvrir, pour une série donnée, à quel modèle de distribution ses propriétés l'apparentent. On peut enfin générer et traiter de nombreuses séries afin d'estimer, par calcul, une propriété, une quantité, difficile à obtenir autrement. Sans négliger les méthodes d'évaluation numérique plus traditionnelles, nous concluons ce livre en mettant l'accent sur l'évaluation Monte Carlo, une approche révolutionnaire, à notre avis, qui a déjà 50 ans d'âge et dont la pénétration et les bénéfices, dans les sciences et l'économie, sont encore à leurs débuts.

Nature et organisation du livre

1.5 Nous n'avons pas voulu écrire un traité, qui eût mené à une compilation et à une présentation encyclopédiques des matières. L'ouvrage apparaît plutôt comme un sub-traité, si l'on nous permet cette expression, c'est-à-dire un traité dans le plan duquel nous aurions opéré des choix. Ce que nous avons retenu correspond, bien sûr, à ce qui nous semblait essentiel et dans quoi nous pouvions prétendre à plus d'aisance; même dans les

matières retenues, nous avons opté pour la simplicité, une certaine parcimonie, et un laconisme qui est notre style. Nous n'avons rien écarté d'important à nos yeux, sinon deux choses : la simulation stochastique (de modèles physiques), que nous mentionnons seulement (au chapitre 7) et qui eût constitué un complément intéressant de cet ouvrage, et la modélisation (de distribution) statistique, qui seyait pourtant à notre propos et que nous n'avons qu'effleurée (au chapitre 4) mais pour laquelle une documentation imposante, ancienne et moderne, est accessible.

1.6 Chaque chapitre de l'ouvrage traite d'un sujet et, à l'exception du présent chapitre, est complété par des exercices et une liste de références. Tout au long des chapitres et sections, nous nous sommes efforcé de conserver un ton convivial et pragmatique, en imaginant que notre lecteur est une personne intelligente, curieuse de connaître différents aspects des matières traitées, capable par elle-même, au besoin, de démontrer formellement ou de documenter les propositions, procédés et résultats que nous avons réunis à son profit. Sans verser entièrement dans le « livre de recettes », un format qui a tout de même le mérite du pragmatisme, nous avons tenu à l'écart les démonstrations rigoureuses et le langage codé; le lecteur désireux d'un supplément d'information rigoureuse en trouvera dans les références fournies. Enfin, malgré la brièveté de l'ouvrage, nombre de théorèmes, procédés, méthodes, propriétés se trouvent mentionnés: la Table des matières restant plutôt globale, nous suggérons au lecteur de consulter l'index, en fin d'ouvrage, pour retrouver l'information cherchée.

1.7 À qui s'adresse ce livre? À toute personne, ayant de bons souvenirs d'algèbre, des connaissances de statistique descriptive et inférentielle, une idée du calcul intégral, et qui a de l'intérêt pour l'étude des phénomènes quantitatifs et des séries manifestant le hasard. Le hasard et les séries au hasard, on l'a dit, sont présents partout, suscitent notre curiosité, nous narguent, mettent au défi notre essentiel besoin d'ordre et de loi: d'une manière réelle quoi qu'indirecte, c'est dans ce débat qu'intervient aussi notre ouvrage. Enfin, tous les férus du traitement de données et du calcul scientifique, dispersés dans toutes les disciplines depuis la physique des particulières jusqu'en sociologie, devraient trouver pâture dans nos chapitres.

1.8 C'est un non-mathématicien qui signe cet ouvrage. Il est l'aboutissement d'une longue pratique de consultation en statistique auprès de chercheurs et d'étudiants dans des disciplines d'application: psychologie expérimentale, psychométrie, physiologie de l'exercice, biomécanique, sciences humaines. C'est aussi le résultat d'un investissement personnel de plusieurs années en programmation informatique de calcul. Le style personnel, le contexte professionnel, les « hasards » de la carrière académique ont laissé des sédiments reconnaissables dans le texte, eu égard

auxquels nous prions le lecteur d'être indulgent. Durant ce cheminement, grâce aux personnes qui sont venues me soumettre leurs problèmes et en étant porté par leur confiance, j'ai appris énormément. En rétrospective, regardant tout ce que j'ai pu réunir et présenter dans le présent ouvrage et que, au départ, j'ignorais, j'ai l'impression d'avoir marché sur un chemin de merveilles. Au vu de cette richesse et de l'intérêt des sujets traités, il était presque impossible pour moi de ne pas en faire quelque chose qui ressemblât à un partage. C'est dans cette perspective modeste, de redonner à d'autres ce que j'ai reçu et qui m'a paru précieux, que j'ai préparé cet ouvrage.

Référence à l'index, conventions, abréviations

1.9 Les concepts, méthodes et vocables retrouvés dans ce livre sont, pour la plupart, définis en quelque part, le plus souvent dans la section de chapitre qui en fait l'usage le plus critique ou le plus important. Ainsi, un concept (e.g. les « moments » d'une distribution) peut être évoqué dans un chapitre (e.g. chap. 3, 4 et 5) et défini dans un chapitre subséquent (e.g. chap. 6), par exemple par une formule ou une courte phrase. Le lecteur pourra s'y repérer en consultant l'index, en fin d'ouvrage.

Chaque section (paragraphe ou groupe de paragraphes), formule importante, exemple, exercice, sont numérotés, par chapitre et par ordre d'apparition. Noter que :

§4.2	désigne, au chapitre 4, la section 4.2 ;
(4.2)	"	au chapitre 4, l'expression ou formule 4.2 ;
exercice 4.2	"	au chapitre 4, l'exercice 4.2, à la fin du chapitre (noté ε4.2 dans l'index) ;
exemple 4.2	"	au chapitre 4, l'exemple numéroté 4.2.

Quant aux abréviations, nous en usons avec parcimonie. Les plus couramment employées sont :

E(x) vs *E(Ø)* E(x) dénote l'espérance (mathématique) de la v.a. x (notée aussi μx ou μ) tandis que E(Ø), en italique, indique la loi (de probabilité) exponentielle de paramètre Ø.

exp(x) dénote la fonction exponentielle, i.e. $e_x = (2{,}71828...)_x$.

f(x), g(x), ... toute fonction de x, rendant un résultat numérique pour chaque valeur x présentée ; aussi (fonction de) densité (de probabilité).

F(x), G(x), ... une fonction de x, habituellement une fonction de répartition f.r.) d'une v.a. continue, telle que $F(x) = pr(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$.

f.r. fonction de répartition d'une v.a. continue ou discrète.

$n!$ factorielle $n\,[= n(n-1)(n-2) \ldots 1\,]$, n entier. Noter que $0! = 1$.

pr(x) (fonction de) masse de probabilité attachée à la valeur «x» d'une v.a. discrète X, ou (fonction de) densité (de probabilité) d'une v.a. continue X au voisinage de x, *i.e.* $pr(x-h \leq X \leq x+h)/2h \rightarrow pr(x)$ *si* $h \rightarrow 0$.

P(x) une fonction de x, habituellement une fonction de répartition (f.r.) d'une v.a. discrète, dénotant un cumul de probabilités tel que $P(x) = \sum_{y=x_0}^{x} pr(y)$; dénote aussi le percentile (ou rang centile) de x dans une population, $0 \leq 100P(x) \leq 100$.

s.o. statistique(s) d'ordre (d'une série statistique)

s.o.u. statistique(s) d'ordre (de v.a.) uniforme(s)

u, u_i v.a. ou, occasionnellement, une valeur quelconque tirée d'une distribution uniforme standard, $U(0,1)$.

v.a. variable aléatoire (discrète ou continue)

v.a.u. variable aléatoire uniforme.

D'autres abréviations et conventions sont disséminées dans l'index.

Chapitre 2 Hasard et irrégularité

2.1 Qu'est-ce que le « hasard »? Avant toute autre chose, le hasard est pour nous une expérience, l'interprétation de ce qui nous est advenu à un moment donné: dans des conditions que nous aurions pu énumérer, tel événement aurait dû se produire et ne s'est pas produit. Quelque chose est intervenu, que par la suite nous avons pu identifier ou non, mais que nous n'avions pas pris en compte, et qui a modifié le cours que nous avions prévu. Nous appelons « fortuit » l'élément interruptif qui a déjoué notre prévision, entendant par là gratuit, sans raison véritable, ne participant pas à l'ordre du monde.

Mais cet aspect « fortuit » n'est qu'une illusion flatteuse, qui masque notre ignorance et excuse l'étroitesse de notre « prévision ». Rien ne se produit vraiment au hasard ou fortuitement; tout a une cause, que nous choisissons de rechercher ou d'ignorer. Depuis longtemps, la science occidentale a adopté cette vérité en tant que dogme et programme. Pierre-Simon Laplace l'exprime admirablement dans son introduction à la Théorie Analytique des Probabilités, parue en 1829:

> *Une intelligence qui pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à 1 'analyse, embrasserait dans la même formule, les mouvements des plus grands corps de l'univers et ceux du plus léger atome: rien ne serait incertain pour elle, et l'avenir comme le passé serait présent à ses yeux.* (p. *ii)*

C'est dans ce contexte épistémologique plus humble que nous tenterons de cerner divers concepts opérationnels liés à la notion de variable aléatoire.

2.2 Lancez une balle en l'air, il est certain qu'elle retombera; fragmentez une galette de 500 grammes en petits morceaux, et la somme des poids des morceaux atteindra 500 grammes; demain matin, le soleil se lèvera (qu'il y ait ou non des nuages): heureusement, la liste est longue des phénomènes

certains et rassurants. Néanmoins, on trouve facilement d'autres sortes de phénomènes, dont la caractéristique est de présenter une part d'imprévisible, une part de hasard. Lancez une pièce de monnaie en l'air: sur quelle face retombera-t-elle? mariez deux jeunes personnes: combien d'années vivront-elles ensemble? prescrivez de l'antibiotique à 100 patients atteints d'une même infection microbienne: combien et lesquels s'en sortiront? Ce sont tous là des phénomènes dits stochastiques, que nous ne pouvons prédire exactement. Les modèles que les savants et chercheurs mettent au point pour étudier, synthétiser ou manipuler ces phénomènes sont aussi des modèles stochastiques. Par contraste avec les modèles et les lois les plus connus en sciences physiques, l'opération d'un modèle stochastique intègre un ou plusieurs éléments de fluctuation, de variation folle, qui représentent notre part d'ignorance dans le phénomène étudié et qui contribuent à rapprocher le comportement global du modèle de ses manifestions naturelles.

2.3 L'objet d'étude principal de la méthode Monte Carlo est les phénomènes stochastiques. Ceci n'exclut pas qu'on s'intéresse à d'autres phénomènes, la grande souplesse de la méthode Monte Carlo permettant parfois d'obtenir des conclusions à propos d'un phénomème très complexe, conclusions qui seraient inabordables par une approche classique.

Dans l'un et l'autre cas, l'étude du phénomène se fait par la mesure de ses manifestions ou par la mesure du comportement du modèle associé. Même s'il présente des aspects aléatoires, « fortuits », il reste qu'on peut tenter de décrire le phénomène en question, d'établir des régularités, de fixer certaines caractéristiques à travers les mesures prises. Pour la plupart des phénomènes qui sont objet d'étude en sciences pures ou appliquées, en biologie, en administration, en économie ou en sciences humaines, leur complexité commande que l'on s'intéresse à plusieurs mesures à la fois, à plusieurs aspects conjoints; dans d'autres cas, un seul aspect, une seule mesure peut suffire. C'est cette mesure (unique ou multiple) d'un phénomène ou d'un modèle stochastique qu'on désigne sous le nom de « variable aléatoire »: c'est une quantité susceptible de présenter des valeurs différentes, et elle est « aléatoire » en ce sens qu'elle redonne les mesures présumément changeantes d'un aspect d'un phénomène stochastique.

2.4 La notion de variable aléatoire constitue le sujet de la statistique et du calcul des probabilités. La variable peut être *discrète,* c'est-à-dire présenter des valeurs parmi un ensemble dénombrable, fini ou infini, de valeurs possibles: le nombre d'accidents de route mortels à chaque mois, au Québec, le numéro sortant d'un dé (à 6 faces), la taille de la population de poissons d'un lac désigné, d'une année à l'autre. La variable peut être *continue;* le plus souvent, toutes les valeurs réelles d'un domaine sont possibles, et le domaine peut être borné, semi-borné ou infini. Le poids

d'un homme constitue une variable continue, de domaine semi-borné selon une borne inférieure de zéro; la position d'un objet sur un axe occupe un domaine infini, et le pourcentage de gras dans un litre de lait est une variable bornée de 0 à 100.

2.5 De même qu'on étudie un phénomène en construisant un modèle (descriptif, structurel, conceptuel) de celui-ci, de même on peut élaborer ou utiliser des modèles de variation aléatoire pour étudier les mesures d'un phénomène stochastique. Ces modèles de variation aléatoire, que nous appelons lois de probabilité, sont constitués simplement par une règle assignant une probabilité à chaque valeur discrète ou à chaque intervalle de valeurs dans le domaine des valeurs possibles. Les modèles de probabilité, par leur énoncé clair et leur formulation mathématique, permettent d'effectuer des calculs et d'obtenir des conclusions qui enrichissent l'étude des phénomènes stochastiques.

2.6 Il y a une querelle épistémologique à propos du concept de probabilité, cette querelle opposant certains mathématiciens (qui favorisent une définition axiomatique et circulaire, fondée sur des micro-événements à probabilités connues) et les scientifiques, incluant R. A. Fisher (qui favorisent une définition intuitive, empiriquement rapportée au concept de fréquence relative). Il n'est pas utile ici d'approfondir cette question, qui a néanmoins de nombreuses implications pratiques. Nous maintenons plutôt notre point de vue pragmatique, en optant à chaque fois qu'il sera nécessaire en faveur de l'opinion régnante dans les applications.

La probabilité est une mesure (de 0 à 1) de la fréquence ordinaire de réalisation (sous des conditions données) d'un sous-ensemble d'états d'un événement aléatoire, ou d'un intervalle de valeurs d'une variable aléatoire. Le modèle de probabilité spécifie cette probabilité pour chaque valeur ou intervalle de valeurs à probabilité non nulle. Par exemple, le modèle du *dé à 6 faces a* comme spécification:

$$\textit{Dé à 6 faces}\text{: } \mathrm{pr}(x) = \tfrac{1}{6}\ , \quad x = 1,2,3,4,5,6$$

et la fameuse loi de Laplace et Gauss, la *loi normale*, s'écrit:

$$\textit{Loi normale}\text{: } \mathrm{pr}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]\ , \text{ tout } x\ .$$

Les chapitres qui suivent présentent plusieurs de ces modèles, et nous offrons en outre des méthodes et procédés permettant de produire des variables aléatoires pour chacun de ceux-là.

2.7 Une fois mis en présence d'une variable aléatoire, il est intéressant et légimitime de reposer la question: qu'est-ce que le hasard? Par exemple, est-ce qu'une séquence de valeurs: $\{ x_1, x_2, x_3, \cdots, x_n, \}$ est « au hasard »?

Exemple 2.1 Un jeu de société

À une soirée mondaine, quelqu'un se présente et vous propose de parier au dé; dans sa main, il tient un dé. Vous êtes consentant à jouer, mais à la condition d'être rassuré sur « l'honnêteté » du dé. Afin de vous rassurer sur le dé, vous l'empruntez à votre interlocuteur, le lancez en l'air 6 fois, et obtenez la séquence de résultats:

{4,2,6,1,3,4}.

Qu'en pensez-vous? Si, au lieu de la séquence précédente, vous aviez obtenu:

{1,4,1,4,1,4},

qu'auriez-vous répondu à votre parieur?

En fait, quelle que soit la séquence de 6 résultats obtenue, sa probabilité d'occurence est théoriquement la même, soit $(\frac{1}{6})(\frac{1}{6})\cdots(\frac{1}{6}) = (\frac{1}{6})^6$. Malgré cela, la seconde séquence vous paraîtra plus inquiétante que la première.

Tout en admettant que, en dernière analyse, le hasard n'existe pas, la question de savoir si une série d'événements, une séquence de nombres, est « au hasard » ou non reste impérieuse et doit être confrontée.

2.8 Le concept anglo-saxon de « randomness » dénote un changement, une variation sans plan défini, sans structure ou finalité. Il ne se trouve pas de traduction directe de ce terme. En réalité, le mot composé « patternlessness », absence de plan ou d'organisation décelable, est un concept descriptif encore plus proche de la réalité que nous voulons cerner. Nous avons privilégié le terme *d'irrégularité* dans son acception descriptive: absence de régularité ou de structure décelable. « Au hasard » signifierait donc « irrégulier », déjouant notre prévision, nous plaçant dans l'incapacité de fixer d'avance le prochain résultat, d'anticiper parfaitement le prochain événement. Peut-on évaluer l'irrégularité?

Le Larousse retrace le mot « hasard » à une origine arabe, « al-zahr » ayant signifié « jeu de dés ». Nonobstant l'étymologie descriptive du mot de même que sa définition essentiellement

descriptive en mathématique, le concept commun de « hasard » désigne généralement un agent, une « cause imprévisible et souvent personnifiée » (Larousse), qui échappe à notre surveillance ou à notre contrôle. Le « hasard » et, en particulier, les « événements de hasard » importants sont, pour nous, ce qu'étaient sans doute les dieux pour les Anciens, des agents qui interviennent à l'improviste (aveuglément ou non) dans notre monde et dans nos vies et qui restent souverainement indifférents aux retombées de leur action sur nous.

En ramenant le concept de « randomness » ou « au hasard » au concept d'« irrégularité », nous sacrifions totalement cette dénotation active, déterminante, du mot hasard. L'irrégularité est ce qui reste après que nous ayons analysé le phénomène à l'étude et que nous en ayons séparé et assigné les variations explicables à leurs causes connues: c'est donc un concept résiduel, négatif et descriptif.

2.9 La probabilité en tant que mesure ne permet pas directement d'évaluer la régularité ou l'irrégularité d'un phénomène, comme l'a montré l'exemple 2.1. Pour démontrer qu'une séquence de mesures d'un phénomème stochastique est « au hasard », il faudra, en général, procéder à la démonstration complémentaire, en prouvant que la séquence de mesures contredit toutes les formes admissibles de régularité.

Cette démonstration complémentaire prend la forme de règles négatives, telles que les suivantes:

- ◊ pour toute séquence de nombres « aléatoires » d'une longueur donnée, le nombre de valeurs plus fortes que la médiane théorique ne devrait pas s'écarter sérieusement du nombre de valeurs moins fortes que la médiane;
- ◊ la longueur (maximale, moyenne) des suites monotones de valeurs croissantes ou décroissantes dans une séquence devrait se conformer aux quantités prescrites pour des variables aléatoires;
- ◊ la corrélation des valeurs consécutives d'une suite devrait être nulle (ou égale à la valeur prescrite, selon la formule de corrélation employée);

etc.

Nous verrons en détail de nombreux *tests d'irrégularité* des variables aléatoires, au chapitre 6. Advenant que tous les tests soient négatifs à propos d'une séquence de valeurs ou d'un ensemble de séquences provenant d'une source commune, la séquence pourra être décrétée aléatoire, « au hasard ».

2.10 Certains auteurs (Martin-Lof 1969; Chaitin 1975) se sont penchés sur la mesure même de l'irrégularité (« randomness ») d'une séquence de nombres; on associe le concept de « complexité » à celui de l'irrégularité de la séquence. Simplifions les propositions de ces auteurs: la complexité de la séquence est proportionnelle à la taille du programme ordinal permettant d'en produire les éléments. Ainsi, pour produire la séquence suivante de *8* coups de dé:

$$A : \{ 1, 2, 1, 2, 1, 2, 1, 2 \} ,$$

le programme pourrait être:

Programme A : Faire 4 fois [Produire «1 »,
Produire «2»].

Dans le programme, chaque opération élémentaire («Faire», «Produire») est dénotée ici par une lettre majuscule. La taille du programme A sera $\aleph(A_8) = 3$, l'indice *8* indiquant la longueur de la série produite. Une autre série :

$$B: \{1,2,3,4,5,6,7,8\}$$

aurait pour programme générateur:

Programme B : Produire «1 »;

Faire 7 fois [Produire « valeur précédente + 1 »].

dont la taille symbolique est également $\aleph(B_8) = 3$.

Est-ce que les séries A et B ci-haut paraissent « irrégulières »? Non, car chaque séquence présente une structure, que le programme a mise à profit. Imaginons une séquence A' de longueur *2n,* telle que:

$$A' : \{ 1, 2, 1, 2, 1, 2, 1, 2, ... \}_{2n}.$$

Le programme pour la produire est essentiellement le même que le programme A (en remplaçant «4» par *« n »)* et il a encore pour taille $\aleph (A^1_{2n}) = 3$. On peut faire de même pour la séquence et le programme B. Le rapport de la longueur du programme à la longueur de la séquence, $\aleph (P_n)/n$, constitue une sorte de mesure de la complexité ou de l'irrégularité de la séquence, évaluée par la longueur du programme nécessaire pour la produire. Pour nos exemples A et B ci-dessus, on observe que $\aleph (P_n) = 3/n \rightarrow 0$ lorsque $n \rightarrow \infty$; ces séquences, dont la mesure d'irrégularité $\aleph (P_n)/n$ approche zéro, apparaissent donc « régulières »; leur comportement n'est pas tout à fait soumis au hasard. La première séquence donnée à l'exemple 2.1,

C: {4,2,6,1,3,4},

représentant 6 faces successives d'un dé lancé, pourrait avoir pour programme:

Programme C : Produire «4»;
Produire «2»;
Produire «6»;
Produire «1 »;
Produire «3»;
Produire «4».

Le programme C a pour longueur $\varsigma(C_6) = 6$, donc une irrégularité de 6/6 = 1. La séquence infinie:

[л]: { 3, 1, 4, 1, 5, 9, 2, 6,···}

correspond aux chiffres successifs de l'expansion décimale de л, la surface d'un cercle de rayon 1: cette séquence a, semble-t-il, une complexité parfaite, soit $\varsigma(\text{л}_n) = n$. S'il était pratique à mesurer et non équivoque, l'indice $\varsigma(P_n)/n$ pourrait refléter vraiment l'irrégularité, le désordre, l'apparence de hasard dans une séquence de nombres[1].

2.11 Ayant démêlé quelque peu les concepts voisins de hasard et d'irrégularité, il nous reste à considérer deux questions très différentes, chacune d'une certaine importance. La première est de savoir si nous, dans le rôle d'utilisateurs de nombres aléatoires et surtout d'interprètes des phénomènes stochastiques, nous jugeons correctement les phénomènes de hasard. Plusieurs études de psychologie cognitive se sont attardées au concept de « probabilité subjective », à la formation du jugement à propos d'événement aléatoires, à la perception d'ordre dans des tableaux formés de points ou de lignes assemblés au hasard, etc.

> Loin de ces raffinements, qu'il nous suffise de demander à un voisin, à un ami, de « se comporter comme un dé qu'on lance » et de fournir, disons, 6 valeurs de faces successives. L'expérience montre que, sur dix personnes sollicitées pour ce faire,

1. Martin-Löf *(op. cit.)* utilise plutôt $\varsigma_0 = \min_j (P_j [x_i]_n)$, c'est-à-dire la taille minimale d'un programme capable de produire la séquence des n valeurs x_i. Cependant, pour le dieu omniscient de Laplace et le canon scientifique (§2.1), l'analyse d'une séquence finie de nombres pourra toujours donner lieu à un programme de taille 1, ou de taille minimale (même s'il faut pour cela examiner des milliers de programmes potentiels), de sorte que, en général, $\varsigma_0 \rightarrow 1$. D'ailleurs, le fait de considérer un seul programme, conçu et formulé naïvement, répond adéquatement aux conditions de mesure que nous cherchons à satisfaire, soit d'évaluer l'apparence (non pas la réalité) de hasard.

environ 5 produiront toutes les faces 1 à 6 dans un ordre quelconque. Or, le calcul (relatif à la *loi d'occupation* d'une variable discrète: voir §6.9) donne une probabilité de $6!\div 6^6 = 0{,}0154$ de produire les 6 faces différentes en 6 coups, un événement rare! Il faut jusqu'à 13 coups pour que la probabilité de produire les 6 faces différentes déborde ½ ($\approx 0{,}5139$), et 36 pour qu'elle équivaille à une quasi certitude ($\approx 0{,}9915$). Sans doute manions-nous bien les concepts de hasard et de séquences aléatoires, mais les réalisations du hasard tissent leur propre brouillard et les intellects les plus éclairés peuvent choir dans un piège du jugement.

2.12 L'autre question fondamentale qu'il reste à considérer est la suivante: une fois rassurés sur l'irrégularité (ou la régularité) d'une séquence de nombres, pouvons-nous vraiment conclure sur le phénomène qui l'a engendrée, sur la source qui l'a produite? Dès 1938, dans un article célèbre, Kendall et Smith faisaient apparaître la distinction opératoire entre le « local randomness » et le « global randomness », l'irrégularité *locale ou globale* d'une source aléatoire. Ainsi, comme on l'a exprimé d'emblée, la séquence de faces d'un dé {1, 4, 1, 4, 1, 4} a ni plus ni moins de chances d'apparaître que la séquence plus « aléatoire » {4, 2, 6, 1, 3, 4}, ce malgré que l'irrégularité locale de la première soit beaucoup moindre que celle de la seconde. On conçoit qu'une source vraiment aléatoire produise à l'occasion des séquences qui n'en ont pas l'apparence, telle une suite de 10 « Faces » au jeu de « Pile ou Face ». En contrepartie, l'inspection de plusieurs séquences ayant chacune une apparence aléatoire pourrait y révéler une surprenante régularité, une redondance: le progrès des sciences expérimentales est tributaire de cette vérité.

Nous sommes donc en présence d'une séquence ou de quelques séquences de nombres aléatoires, et nous tâchons *d'inférer* le degré d'irrégularité globale d'une source à partir de diagnostics sur l'irrégularité locale de ses productions. Sur un plan théorique, cela revient à déterminer si la suite de programmes P_1, P_2, ¨, correspondant à autant de séquences, constitue elle-même une séquence irrégulière. Sur un plan pragmatique, Knuth (1969), qui discute très bien toute cette question, propose la démarche diagnostique suivante:

1) produire (ou mesurer) N séquences d'une longueur donnée;
2) calculer pour chaque séquence une statistique définie selon une règle négative telle que celles énumérées **en §2.9;**
3) former la distribution empirique des N statistiques et l'interpréter (en la comparant, s'il y a lieu, à la distribution échantillonnale de la statistique utilisée).

L'étape 2, on l'a vu, permet déjà un diagnostic sur l'irrégularité locale d'une source aléatoire, par le truchement d'un *test d'hypothèse statistique* sur la séquence de valeurs concernée. Mais la démarche globale, et l'examen de la distribution des N statistiques, nous mettent à même de nous prononcer globalement sur la source aléatoire; si c'est utile, des tests d'hypothèses sont aussi possibles à ce niveau.

Exercices

2.1 Identifier quelques phénomènes comportant des éléments « fortuits » et quelques-uns n'en comportant pas. Scruter le mécanisme ou les conditions de manifestation de ces phénomènes. Montrer qu'il y a des différences de degré, non pas de nature, entre ces deux classes, et que le caractère fortuit ou déterminé dépend du niveau d'analyse (macroscopique vs microscopique, à durée courte ou très longue, etc.).

2.2 Soit x, le nombre de jours dans un mois du calendrier: sous quelles conditions x est-il une variable aléatoire? Quelle est sa loi (ou distribution) de probabilité?

2.3 Un baril contient des milliers de petites billes. On brasse le baril, on y pige une bille et on obtient son poids en la plaçant sur une balance graduée au dixième de gramme, e.g. 28,7 g. En quel sens ce poids est-il une variable aléatoire? Est-ce une variable discrète ou continue (expliciter)? Comment trouver sa loi de probabilité?

2.4 Soit la séquence:

D: {0, 1, 1,0,0, 1, 1,0,0};

construire (au moins) deux programmes permettant de la produire. Soit la séquence prolongée D':

D': { 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, ••• } ;

construire (au moins) un programme pour cette séquence. Trouver et comparer les longueurs $\leq$ de chaque programme.

2.5 Soit E, une séquence virtuelle de 15 éléments et constituée des chiffres 1 à 5, comme suit. Les 5 premières valeurs contiennent 5 chiffres différents, les cinq valeurs suivantes aussi, de même que les 5 dernières: concevoir un programme virtuel pour cette séquence. D'autre part, soit F, une séquence virtuelle de 15 éléments, chacun desquels pouvant être également un chiffre de 1 à 5: concevoir un programme virtuel pour cette séquence. Comparer les longueurs de ces programmes.

Références

CHAITIN, G.J. (1975). Randomness and mathematical proof. *Scientific American,* 232, 47-52.

KENDALL, M.G., SMITH, B.B. (1938). Randomness and random sampling numbers. *Journal of the Royal Statistical Society,* 101, 147-166.

KNUTH, D.E. (1969). *The art of computer programming.* Vol. 2: *Seminumerical algorithms.* Reading (MA), Addison-Wesley.

LAPLACE, P.S. (1829). *Théorie analytique des probabilités* (3e édition), Paris.

MARTIN-LOF, P. (1969). Algorithms and randomness. *Review of the International Statistical Association, 37,* 265-272.

Chapitre 3

Production de nombres pseudo-aléatoires uniformes

3.1 Dans son cas général, la méthode Monte Carlo permet l'étude et l'évaluation des phénomènes quantitatifs en exploitant des séries de nombres aléatoires; ces derniers sont en quelque sorte le combustible de la méthode. Qu'entend-on ici par ces « nombres aléatoires » que la méthode Monte Carlo consomme ?

Considérons une séquence de nombres quelconque, telle que { 4, 2, 6, 1, 3, 4 } . On peut tenter de déterminer l'ordre ou l'irrégularité des éléments de cette séquence, on peut construire un « programme » qui produit cette séquence, etc. Cependant, ce ne sont pas là des nombres aléatoires, mais plutôt des nombres donnés. A la limite, «4» n'est pas un nombre aléatoire, non plus que « 0,426134 ».

Par nombres aléatoires, on désigne des nombres virtuels (c'est-à-dire, non encore produits) provenant d'une source aléatoire. L'épithète « aléatoire » caractérise vraiment la source de nombres plutôt que les nombres eux-mêmes, et elle a deux significations: l'une, à l'effet que la source fonctionne sous l'influence de facteurs divers, inconnus et non contrôlés, c'est-à-dire que sa production fluctue « au hasard », et l'autre, à l'effet qu'il est impossible de prédire les valeurs de la prochaine séquence de nombres produite. La première signification fonde et garantit la seconde, mais c'est à la seconde, à l'imprédictibilité des valeurs successives produites, que font référence les applications de la méthode Monte Carlo.

Nous étudierons d'abord les moyens qu'il y a d'obtenir ou de produire des nombres aléatoires en quantité indéfinie; la source privilégiée de nombres est une fonction linéaire récursive qui fournit des nombres, ou variables, aléatoires à distribution uniforme. Nous identifierons ensuite quelques propriétés des variables aléatoires uniformes. Nous terminerons le chapitre en définissant les statistiques d'ordre de la loi uniforme et leurs fonctions (maximum, minimum, étendue, etc.), avec des techniques permettant de les produire directement. Le traitement de variables aléatoires répondant à différentes lois de distribution fera l'objet du chapitre suivant.

3.2 Il existe au moins trois catégories de sources de nombres aléatoires: les phénomènes stochastiques, les tables de nombres publiées, les fonctions linéaires récursives.

La plupart des phénomènes étudiés, en sciences pures ou appliquées, en administration et ailleurs, comportent des aspects aléatoires, dont la mesure constitue une source possible de nombres aléatoires. C'est ce qui est mis à profit dans les tirages de loterie, pour lesquels l'on s'ingénie même à décanter les aspects aléatoires, à les dégager de contraintes potentiellement biaisantes: on place des boules numérotées dans un baril, boules de formes et de compositions le plus identiques possible, on secoue considérablement le baril (en lui faisant faire plusieurs tours, et le tout s'effectue sans intervention humaine, automatiquement. Des systèmes semblables, basés sur le décompte d'émissions radioactives dans un intervalle de temps fixe ou sur d'autres mécanismes physiques fortement aléatoires, ont permis de constituer des ensembles de nombres aléatoires satisfaisants.

La table de nombres aléatoires[1] la plus connue, celle publiée par la compagnie RAND en 1955, fut bâtie à l'aide d'une machine stochastique. Tous les livres de tables statistiques (p. ex. Laurencelle et Dupuis 2000) présentent des ensembles de nombres aléatoires, constitués pour la plupart des chiffres 0 à 9 arrangés en tableau: citons par exemple les 40 000 chiffres aléatoires dans le fameux *Handbook* de D. B. Owen (1962, p. 519-538). Cette table, tout comme d'autres semblables, a été élaborée par ordinateur à partir d'une fonction linéaire récursive.

Exemple 3.1 Une table de nombres aléatoires

> La table de Owen *(op. cit.)* s'étale sur 20 pages; chaque page contient des blocs disposés en 10 rangées de 10 colonnes, et chaque bloc est un petit tableau haut de 5 chiffres et large de 4. Pour m'y situer « au hasard », je demande à mon voisin de bureau un nombre de 1 à 20; il me donne 17. Sur cette page, je pointe « au hasard » et mon doigt épingle un bloc de rangée 3, colonne 7. Voici un aperçu de la page 17 (= page 535 du livre):

1. Le lecteur objectera à juste titre qu'on ne peut attribuer le qualificatif *d'aléatoireà* quelque ensemble de nombres définis que ce soit, fussent-ils des millions. Pour «fonctionner» comme source de nombres aléatoires, la table de nombres doit être «lue au hasard». Il faut recourir à un procédé stochastique externe (un pointage du doigt à l'aveuglette, des coordonnées de lecture obtenues par des moyens extravagants, etc.) pour amorcer la lecture à un endroit aléatoire, quitte à continuer dans la même rangée (ou colonne, selon un choix préalable) pour obtenir les nombres suivants. On comprend que la méthode de la table peut devenir vite suspecte ou, en tout cas, fastidieuse.

	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10
rangée 2 :	0556	2822	3230	8247	7350	7186	5982	6155	3284	3129
	2098	7991	3518	7761	8583	8441	8702	2517	4957	5450
	7478	7461	3680	6107	6363	7017	8183	5191	1208	2977
	7569	7655	5563	1499	7333	3311	3568	5062	0407	9708
	0982	6840	4171	7387	7059	8947	3896	1428	4075	0262
rangée 3 :	6588	2958	6768	1709	0240	7609	**9906**	1174	7980	9157
	3541	4892	9553	7565	5788	9109	**7127**	6145	7074	0802
	3978	0755	1561	7850	8043	9185	**9273**	8103	3513	0738
	6807	3074	0441	6711	9357	5627	**1918**	8617	6695	5377
	4465	4907	5278	3479	5519	9740	**4684**	5860	6711	9120
rangée 4 :	6086	4619	5233	3980	6986	1871	4643	3638	1176	2387
	3874	3751	2274	5384	2555	5351	8463	6268	0628	3250
	0342	8660	9586	1765	8822	5069	7550	3275	7727	1272
	4767	0418	6234	6324	5946	3686	9023	1787	6578	0545
	8624	1120	4126	3277	8568	8975	4278	9870	3475	5242

Supposons que j'aie besoin de 10 nombres u, à distribution uniforme entre 0 et 1, soit u ~ $U(0,1)$. Je peux les obtenir en « lisant » verticalement les 5 chiffres présentés, chaque bloc fournissant ainsi 4 nombres. Le premier nombre, formé par «9», «7», «9», «1» et «4», serait u_1 = 0,97914; le second, u_2 = 0,91296; et u_3 = 0,02718, u_4 = 0,67384. Le bloc est épuisé; le bloc suivant peut être à gauche, dans la même rangée, et j'obtiens u_5 = 0,79959; u_6 = 0,61167; u_7 = 0,00824; u_8 = 0,99570; puis u_9 = 0,05895 et u_{10} = 0,27035.

Bien entendu, on aurait pu lire horizontalement les nombres (qui auraient eu 4 chiffres), ou combiner deux ou plusieurs blocs pour former le même nombre, ou encore lire verticalement mais *de bas en haut,* etc. Comme on le constate, une table comme celle-ci, qui contient 40 000 chiffres « au hasard », peut satisfaire des besoins ponctuels en nombres aléatoires de façon quasi inépuisable.

La fonction linéaire récursive, parce qu'elle est d'utilisation plus commode et qu'elle s'intègre immédiatement dans un programme d'ordinateur appliquant l'approche Monte Carlo, a presque supplanté les autres sources de nombres aléatoires. Par fonction récursive, on veut dire ici une formule qui produit une valeur nouvelle en exploitant une ou plusieurs valeurs antécédentes. Dans son expression la plus simple, ce serait:

$$x_{n+1} = f(x_n) \; ; \qquad (3.1)$$

dans la mesure où la valeur x_{n+1} n'est pas évidemment prédictible à partir de x_n, x_{n-1}, x_{n-2}, etc., la fonction f indiquée constitue une source possible de nombres aléatoires.

3.3 La suprématie actuelle de la fonction linéaire récursive comme source de nombres aléatoires a consacré en même temps le caractère fondamental de la loi de probabilité uniforme, ou distribution uniforme. En effet, au contraire de certains phénomènes stochastiques dont le résultat obéit à d'autres lois de probabilité (comme la loi exponentielle, la loi de Poisson, la loi Gamma, etc.), la fonction linéaire récursive produit des variables à distribution uniforme. Comme on verra au chapitre 4, il existe maints procédés permettant de produire des variables relevant d'une loi de probabilité quelconque à partir de variables aléatoires uniformes (v.a.u.). De plus, toute v.a. non uniforme peut être convertie en une v.a.u. en faisant l'intégrale de probabilité pour cette variable (cf. §4.2). La distribution uniforme constitue ainsi un point de rencontre de toutes les distributions. Comme, de plus, ses propriétés sont très bien connues, les variables de la distribution uniforme peuvent être soumises à nombre de tests d'hypothèses afin d'en valider l'irrégularité; ce sera l'objet du chapitre 6.

3.4 Il s'agit de trouver une fonction récursive qui produise des nombres x à comportement quasi aléatoire, distribués uniformément entre une borne inférieure A et une borne supérieure B, ou $x \sim U(A, B)$. Dans sa forme standard[2], la distribution uniforme occupe l'intervalle (0..1), avec des variables $u \sim U(0, 1)$ obtenues par la transformation:

$$u = (x - A)/(B - A) . \tag{3.2}$$

Il n'est pas difficile de satisfaire uniquement la contrainte de double borne de cette distribution; reste à trouver une fonction mathématique produisant des nombres qui ont *toutes* les propriétés des variables aléatoires uniformes.

$$u_{n+1} = [\ \text{centre}_k\ (10^k u_n)^2\ ]\ /\ 10^k\ ; \tag{3.3}$$

---

2. Il convient de distinguer la forme *standard* d'une variable aléatoire (ou de sa loi de distribution) et sa forme dite *standardisée*. Par forme standardisée, on entend généralement la réduction d'une v.a. vers une v.a. équivalente de moyenne 0 et de variance 1 grâce à la transformation $z_X = (X-\mu_X)/\sigma_X$. Par ailleurs, chaque v.a. possède (ou peut avoir) une forme standard, souvent sa forme naturelle la plus simple. La loi uniforme générale $X \sim U(A,B)$, qui a pour moyenne $\mu_X = ½(A+B)$ et pour variance $\sigma_X^2 = (A-B)^2/12$, a pour forme standard $U \sim U(0,1)$ par la transformation (3.2) ; la variable standardisée impliquerait la transformation z_X ci-dessus, tandis qu'une variante de la forme générale, utilisant μ_X et σ_X^2, serait $X \sim U(\mu_X-\sqrt{3}\sigma_X,\mu_X+\sqrt{3}\sigma_X)$. Dans le cas de la loi normale (cf. §4.7), forme standard et forme standardisée coïncident, soit $N(\mu=0,\sigma^2=1)$ ou $N(0,1)$.

le produit $10^k\, u_n$, élevé au carré, se développe en *2k* chiffres décimaux avant la virgule; il s'agit d'extraire les *k* chiffres du centre (en escamotant ½*k* chiffres *à* gauche et ½*k à* droite), pour diviser enfin par 10^k. Cette méthode, qui fonctionne bien sous certaines conditions, est de réalisation plus complexe que la méthode privilégiée, celle de la fonction linéaire. D'autres méthodes à l'avenant sont possibles, telles que:

$$u_{n+1} = y - \lfloor y \rfloor \ , \ y = 100\sqrt{(u_n + \pi)} \ , \tag{3.4}$$

$$u_{n+1} = y - \lfloor y \rfloor \ , \ y = \log_e(10u_n) \ , \tag{3.5}$$

ainsi de suite [l'expression [*x*] dénote la partie entière de *x*, *i.e.* l'entier *n* le plus grand tel que $n \leq x$]. Une fonction de ce type, facile à programmer sur une calculette, pourra dépanner à l'occasion un utilisateur.

Certains auteurs (voir Knuth 1969, p. 30 et suiv.) proposent d'utiliser une *fonction composée,* ce afin de satisfaire des besoins spéciaux (période augmentée de la fonction génératrice, protection contre la dépendance séquentielle des nombres produits). La fonction composée consiste en deux fonctions simples utilisées conjointement; la première fournit des valeurs inscrites dans un grand tableau, la seconde pige au hasard dans ce tableau: la valeur fournie, repérée par la seconde fonction, est à son tour remplacée par une nouvelle valeur issue de la première fonction. Sur ce thème, plusieurs variations et d'autres raffinements sont possibles.

3.5 La fonction génératrice privilégiée est de forme linéaire et présente trois paramètres, soit:

$$u_{n+1} = y_{n+1} / m \ , \ y_{n+1} = (ay_n + c) \bmod m \ , \tag{3.6}$$

les paramètres étant le *multiplicateur a,* le *module m* et la *constante additive c.* L'expression « *a* mod *b* » dénote le reste après division entière, soit *a* mod *b* = *a* - *b* *[a/b]* . Knuth retrace sommairement l'historique des fonctions génératrices, cette fonction linéaire étant attribuable à D. H. Lehmer. Sa simplicité présente un double avantage puisque, d'une part, elle se prête assez facilement à l'analyse et que, d'autre part, elle est de calcul très rapide par ordinateur.

Toute fonction numérique récursive de forme (3.1) vient à se répéter éventuellement, à savoir qu'il existe un nombre *t* tel que $x_{n+t} = x_n$ (ou $u_{n+t} = u_n$), et $x_{n+j} \ \& \ x_n$, $1 \leq j \leq t$; ce nombre *t* est appelé la *période* de la fonction génératrice. Pour la fonction linéaire (3.6), la période maximale est déterminée par le module, soit $t \leq m$.

Voici différents générateurs assez satisfaisants, qu'on retrouve dans la documentation:

$$y_{n+1} = 16807y_n \bmod (2^{31} - 1) \quad \{\text{IMSL 1987, « RNUNF »}\} \tag{3.7a}$$

$$y_{n+1} = 524293y_n \bmod 2^{35} \quad \{\text{Owen 1962}\} \tag{3.7b}$$

$$y_{n+1} = 1220703125y_n \bmod 2^{35} \quad \{\text{Rohlf et Sokal 1981}\} \tag{3.7c}$$

$$y_{n+1} = (3141592653y_n + 2718281829) \bmod 2^{35} \quad \{\text{Knuth 19691} \tag{3.7d}$$

[Pour la facilité des opérations dans un ordinateur binaire, le module *m* représente ordinairement une puissance de 2, sous la forme $m = 2^k$, $2^k - 1$ ou $2^k + 1$, k étant la capacité binaire (positive ou absolue) d'une unité-mémoire de la machine.]

3.6 L'expérience acquise et l'étude formelle des fonctions génératrices linéaires ont permis d'établir certaines règles pour guider la construction d'une fonction donnée; ces règles ont trait au choix des paramètres *a, c* et *m* apparaissant dans la fonction (3.6). Nous les retrouvons discutées en détail dans Rubinstein (1981), Knuth (1969) et d'autres.

Le paramètre m. On préfère un module *m* plutôt grand, puisqu'il détermine la période maximale de la fonction. En effet, puisque y_n est le reste de la division d'un entier par l'entier *m,* les valeurs (théoriquement) possibles de ce reste sont 0, 1, 2, ···, m-1; après avoir produit *m* valeurs y (ou valeurs $u = y/m$), la fonction re-produira certainement une valeur déjà passée, et elle en re-produira en fait la série complète. Tel qu'indiqué plus haut, *m* est ordinairement calqué sur l'unité-mémoire de la machine envisagée, par exemple $m = 2^{32}$ pour un ordinateur de 32 bits possédant un multiplicateur et un diviseur à 32 bits de valeur absolue.

Le paramètre c. Une valeur $c > 0$ peut accroître la période de la fonction; en fait, la valeur c = 0, plus simple et rapide, entraîne une réduction de période et oblige à prévenir que la fonction récursive ne dégénère à zéro, si $y_n = 0$. On recommande d'utiliser des valeurs telles que (c mod 2) = 1 pour une machine binaire (i.e. *où* $m = 2^k$), et (c mod 5) ≠ 0 pour une machine décimale.

Le paramètre a. Le multiplicateur *a* joue le rôle crucial dans la fonction (3.6) en influençant l'ordre ou le désordre de la séquence de nombres produite. La valeur $a = 0$ produirait une constante (= c), des valeurs *a* trop petites engendreraient des séquences en dents de scie, etc. Les auteurs recommandent de respecter les inégalités:

$$\sqrt{m} < a < m - \sqrt{m} \text{ et } a > m/100.$$

De plus, pour une machine binaire, on suggère que (a mod 8) = 5 tandis qu'on suggère (a mod 200) = 21 pour une machine décimale. Enfin, Knuth (1969) met en garde contre un multiplicateur à 2 bits, comme $a = 2^i + 1$.

Exemple 3.2

Nous avons besoin d'une source commode de nombres aléatoires, et nous voulons programmer notre calculette dans ce but. Cet appareil est décimal; pour la simplicité, nous utiliserons des nombres à 4 chiffres (décimaux).

Avec $m = 10^4 = 10000$, prenons c = 1 afin d'assurer une période assez grande. Reste à trouver le multiplicateur a. Pour cela, considérons les recommandations:

$$100 < a < 9900 \text{ , } a > 100 \text{ , } a \bmod 200 = 21 \text{ .}$$

Cette dernière spécification peut s'écrire: a = 200k + 21, et les valeurs *k* admissibles occupent l'intervalle discret (1..49). Soit *k* = 40, et l'on obtient:

$$y_{n+1} = (8021 y_n + 1) \bmod 10000 \text{ .}$$

Prenons la valeur de départ, ou *semence,* $y_0 = 1$; alors:

$y_1 = (8021 \times 1 + 1) \bmod 10000 = 8022 \rightarrow u_1 = 0{,}8022$;

$y_2 = (8021 \times 8022 + 1) \bmod 10000 = 4463 \rightarrow u_2 = 0{,}4463$;

$y_3 = (8021 \times 4463 + 1) \bmod 10000 = 7724 \rightarrow u_3 = 0{,}7724$;

on obtient encore $y_4 = 4205$, $y_5 = 8306$, $y_6 = 2427$, etc. La période *t* de cette fonction est de 10000, soit la période maximale.

La fonction (3.6), même munie de paramètres optimaux, n'est pas une source vraiment aléatoire, c'est pourquoi on dit des nombres produits qu'ils forment une séquence de nombres *pseudo-aléatoires. Ce* caractère aléatoire ou pseudo-aléatoire, cet aspect *irrégulier,concerne* les nombres $u = y/m$, ou encore *la partie supérieure* des nombres y. En effet, l'entier y servant d'ingrédient dans la fonction linéaire (3.6) comporte k chiffres, disons, et seuls les quelques chiffres supérieurs peuvent afficher une variation aléatoire dans la séquence. Par contraste, la variation des chiffres inférieurs de y n'est pas garantie et leur exploitation dans un contexte d'échantillonnage aléatoire est fortement contre-indiquée.

Supposons que nous désirions quelques chiffres au hasard entre 0 et 9, et nous utilisons la fonction développée dans l'exemple 3.2 (en stipulant que cette fonction répond à nos exigences sur les séquences

aléatoires). La méthode indiquée consiste ici à obtenir les chiffres [$10u_i$] ; ainsi la séquence obtenue sera: [10x0,8022] = 8, puis 4, 7, 4, 8, etc. L'autre méthode, contre-indiquée, consisterait à calculer (y mod 10); la série résultante serait ici (8022 mod 10) = 2, puis 3, 4, 5, 6, etc. Cette série ne convient manifestement pas à nos besoins.

3.7 La simplicité de la fonction linéaire récursive (3.6), voire de sa forme dénudée (3.1), n'est pas sans inconvénient. En fait, même si les valeurs successives (y_n) qu'elle produit varient dans **R** de manière apparemment aléatoire, les k-tuplets $(y_{n+1}, Y_{n+2}, \cdots, Y_{n+k})$ successifs n'occupent pas tout l'espace dans $\mathbf{R}^k$ mais se logent plutôt dans un sous-espace structuré (Marsaglia 1968). Cette tare intrinsèque, de peu de conséquence dans la plupart des cas, peut devenir inquiétante, voire fatale, lorsque les nombres sont appliqués à la solution de problèmes multidimensionnels. Fort heureusement, pour ces cas et en général, la documentation propose des formes généralisées du générateur (3.6), soit les générateurs récursifs multiples (i.e. d'ordre 2 ou plus) et les générateurs matriciels (Deng et Lin 2000), pour lesquels l'inconvénient mentionné s'estompe ou disparaît.

3.8 Les systèmes et langages de programmation, ainsi que plusieurs calculettes dotées de fonctions statistiques, fournissent des procédures de production de nombres pseudo-aléatoires à distribution uniforme. Ces procédures, désignées par exemple « RAND », « RANDOM », « RND », « RANF », « RUNF », sont presque toutes des fonctions récursives linéaires de type (3.6) et elles fournissent des variables pseudo-aléatoires uniformes, une ou plusieurs à la fois.

Ces procédures étant intrinsèquement récursives, la valeur produite à l'invocation *« n+l »* dépend de celle fournie à l'invocation *« n »* qui la précède. En début d'exécution, la procédure recèle une valeur de départ, y_0, et les utilisations répétées de la fonction produiront toujours les mêmes valeurs consécutives (dans la plupart des systèmes). La *semence* y_0 peut cependant être re-spécifiée au début ou à tout moment par le programme utilisateur; certains énoncés sont prévus à cette fin, tels que « RANDOMIZE » en Basic et « RANSET » ou « SRAND » en Fortran. Un des moyens commodes de relancer la fonction consiste à fournir comme semence la valeur (entière) de l'horloge interne de l'ordinateur; en Basic, l'énoncé sera simplement:

« RANDOMIZE TIMER »

La banque IMSL (1987) fournit quant à elle tout un assortiment de procédures et fonctions touchant les nombres pseudo-aléatoires, de distributions diverses.

Les variables aléatoires uniformes (v.a.u.)

3.9 Les nombres successifs produits par une fonction telle que celles étudiées plus haut correspondent en principe à des variables aléatoires de distribution uniforme. La distribution uniforme (parfois appelée rectangulaire) a pour loi de probabilité (ou fonction de densité):

$$\mathrm{pr}(u) = 1 \ , \ 0 \leq u \leq 1. \tag{3.8}$$

Une variable uniforme occupant un intervalle différent, par exemple $A \leq x \leq B$, peut être ramenée à la forme standard par la transformation (3.2). Johnson, Kotz et Balakrishnan (1994, 1995) consacrent un chapitre de leur magistral traité à la loi uniforme; la documentation la plus complète, avec plusieurs résultats nouveaux, se trouve dans Laurencelle (1993).

Les moments à l'origine (μ') et les moments centraux (μ) d'ordre r de la variable u sont:

$$\mu'_r(u) = 1 \ / \ (r+1) \tag{3.9}$$

$$\mu_r(u) = [1 + (-1)^r]/[2^r(r+1)] \ . \tag{3.10}$$

Ainsi, l'espérance (ou moyenne μ ou p4) est la variance (6^2) 1/12, les indices d'asymétrie et d'aplatissement sont respectivement $y_1 = a_3 = 0$ et $'y_2 = a_4 - 3 = -1{,}2$.

3.10 Étant donné leur rôle fondamental dans l'application de la méthode Monte Carlo et le fait que les v.a. de distributions diverses en tirent leur provenance, il importe d'étudier avec soin les séquences de nombres pseudo-aléatoires uniformes et d'en sanctionner la validité. Ainsi, la moyenne arithmétique de n variables aléatoires uniformes (v.a.u.), dénotée $\bar{u}$, a pour loi de probabilité:

$$\mathrm{pr}_{\bar{u}}(x) = \frac{n^n}{(n-1)!} \sum_{i=0}^{k} (-1)^i \binom{n}{i} \left(x - \frac{i}{n}\right)^{n-1}, \quad k = \lfloor nx \rfloor \ ; \tag{3.11}$$

3. La (fonction de) densité de probabilité s'applique à une v.a. continue alors qu'on parlera plutôt de probabilité ou de fonction de masse (de probabilité) pour une v.a. discrète. Ainsi, en lançant en l'air un dé à 6 faces, la probabilité (ou masse de probabilité) d'obtenir «1» est de $\frac{1}{6}$. D'autre part, pour une v.a. continue x de distribution (ou fonction de densité) $f(x)$, la probabilité d'observer $x = x_0$ est nulle, la probabilité d'observer $x \in (x_0 - \frac{1}{2}\delta_X, x_0 + \frac{1}{2}\delta_X)$ est d'environ $f(x_0) \times \delta_X$ et la densité de probabilité autour de x_0 est $f(x_0)$. Cela dit, tout en nous efforçant à un langage et à des concepts clairs, nous éviterons autant que possible d'encombrer l'exposé par du dialecte savant.

ses moments sont $\mu = ½$, $o^2 = 6/12n$, $y_1 = 0$ et $y_2 = — 1{,}2/n$. Quant à la moyenne géométrique, dénotée MG, elle a pour loi:

$$\mathrm{pr}_{MG}(x) = \frac{n^n x^{n-1}}{(n-1)!}(-\log_e x)^{n-1} , \qquad (3.12)$$

avec pour moments à l'origine $\mu'_r = [n/(n+r)]^r$; moyenne, médiane et mode tendent vers $e^{-1} \approx 0{,}3679$ pour n croissant. De plus, si x_n est une moyenne géométrique de n variables uniformes satisfaisant (3.12), alors $y = -2n\log_e x$ est une variable Khi-deux ayant $2n$ degrés de liberté; la variable x_p, de percentile P *(i.e.* telle que $\mathrm{pr}(x \leq x_p) = P$), correspond à la variable y_{100-P}, du percentile complémentaire.

Laurencelle (1993) donne plusieurs autres résultats concernant le point-milieu, la moyenne harmonique, la variance et la variance permutative des v.a.u. Certains sont donnés dans les exercices, en fin de chapitre, et la plupart sont repris utilement au chapitre 5 touchant les tests d'hypothèses.

3.11 Les *statistiquesd 'ordre* d'une variable aléatoire sont les n variables d'un échantillon replacées en ordre de valeurs croissantes. Ainsi, l'ensemble échantillonné $\{ u_1, u_2, \cdots, u_n \}$ donne lieu aux statistiques d'ordre $\{ u_{(1)}, u_{(2)}, \cdots, u_{(n)} \}$, qui respectent les inégalités:

$$u_{(i)} \leq u_{(j)},\ i < j .$$

On utilise aussi les notations $x_{(i:n)}$, $x(i:n)$, $x_{i:n}$ qui, toutes, indiquent la i^e statistique d'ordre d'un échantillon de n variables. Les théorèmes touchant les statistiques d'ordre uniformes (s.o.u) sont, à l'instar de ceux sur les v.a.u., d'une grande importance vu que la plupart d'entre eux se transfèrent utilement sur toute autre distribution. Par exemple, si, dans une étude Monte Carlo, on veut produire une variable correspondant au score maximal parmi 20 v.a. d'une loi de distribution donnée, il suffira de produire le score maximal de 20 v.a. uniformes, soit u(20:20), puis de transformer ce dernier en une variable de la loi cible, ce qu'on peut faire de différentes manières (voir chapitre suivant). L'ouvrage de David (1981) est le *nec plus ultra* sur le sujet des statistiques d'ordre (voir aussi Arnold, Balakrishnan et Nagaraja 1992).

Exemple 3.3

Soit l'échantillon suivant de $n = 9$ v.a.u.:

$u_i = \{ 0{,}892; 0{,}541; 0{,}355; 0{,}241; 0{,}427; 0{,}776; 0{,}142; 0{,}183; 0{,}211 \}$.

Une fois replacées en ordre de valeurs croissantes, ces nombres deviennent une réalisation des statistiques d'ordre de l'échantillon, des s.o.u.:

$u_{(i)}$ = { 0,142; 0,183; 0,211; 0,241; 0,355; 0,427; 0,541; 0,776; 0,892 } .

Les valeurs minimale et maximale sont respectivement $u_{(1)} = 0{,}142$ et $u_{(9)} = 0{,}892$; l'étendue est $E = u_{(n)} - u_{(1)} = 0{,}892 - 0{,}142 = 0{,}750$. La *médiane,* qui occupe la position centrale de rang $r = \frac{1}{2}(1 + n) = 5$, coïncide ici avec $u_{(5)}$, d'où $Md = 0{,}355$. Dans le cas d'un échantillon de nombre n pair, la médiane sera conventionnellement le point-milieu des deux s.o. centrales, soit $Md(n \text{ pair}) = \frac{1}{2}(x_{(\frac{1}{2}n)} + x_{(\frac{1}{2}n+1)})$.

3.12 Les s.o.u. ont une distribution *Bêta, ou* $\beta(i, n - i+1)$, la variable x_R occupant elle aussi l'intervalle (0, 1). Cette loi de probabilité[4] est:

$$\mathrm{pr}_{u(i:n)}(x) = \frac{n!}{(i-1)!(n-i)!} x^{i-1}(1-x)^{n-i} , \quad 0 \le x \le 1 \qquad (3.13)$$

Les moments et autres caractéristiques principales des s.o.u. sont:

$$\mu[u_{i:n}] = i/(n+1) \qquad (3.14)$$

$$\sigma^2[u_{i:n}] = i(n-i+1)/[(n+1)^2(n+2)] \qquad (3.15)$$

$$\gamma_1[u_{i:n}] = 2(n-2i+1)\sqrt{\{(n+2)/[i(n-i+1)]\}}/(n-1) \qquad (3.16)$$

$$\gamma_2[u_{i:n}] = \{3(n+2)[2(n+1)^2 + i(n-i+1)(n-5)]\}/[i(n-i+1)(n+3)(n+4)] - 3 \qquad (3.17)$$

$$\rho[u_{i:n}, u_{j:n}] = [i(n-j+1)]/\sqrt{[ij(n-i+1)(n-j+1)]} , \quad i \le j \qquad (3.18)$$

La corrélation (3.18) est toujours positive ($\rho > 0$), mais elle décroît à mesure que l'écart ordinal $|j - i|$ d'une variable à l'autre, s'accroît; la corrélation entre $u_{(1)}$ et $u_{(n)}$ est égale à $1/n$.

La fonction de répartition (f.r.) d'une s.o.u., soit $P(u_{(i)}) = \mathrm{pr}(x \le u_{(i)})$, a une forme simple:

$$P_{u_{(i)}}(x) = \sum_{j=i}^{n} \binom{n}{j} x^j (1-x)^{n-j} , \quad 0 \le x \le 1 \qquad (3.19)$$

4. La forme standard de la loi *Bêta* est 0(a,b); voir Laurencelle et Dupuis (2000) et exercice 4.29. Noter que la loi uniforme standard *U*(0,1) est un cas particulier de la loi *Bêta,* soit β (1,1).

d'où on tire notamment les f.r. du minimum et du maximum de n v.a.u., soit:

$$P_{u(1:n)}(x) = 1 - (1 - x)^n \; ; \tag{3.20}$$

$$P_{u(n:n)}(x) = x^n \; . \tag{3.21}$$

La loi de probabilité de l'étendue de n variables uniformes, $E = u_{(n)} - u_{(1)}$, et

$$\mathrm{pr}_E(x) = n(n-1)x^{n-2}(1-x) \; , \quad 0 \le x \le 1 \; ; \tag{3.22}$$

$$P_E(x) = nx^{n-1}(1-x) + x^n \; ; \tag{3.23}$$

l'espérance est $\mu(E) = (n - 1)/(n + 1)$, la variance $0^2(E) = 2(n-1) / [(n+1)^2(n+2)]$. La documentation fournit plusieurs autres résultats.

3.13 L'approche naïve pour produire les statistiques d'ordre uniformes (s.o.u.) consiste d'abord à produire n v.a.u. puis à les placer en ordre au moyen d'une procédure de tri. En marge de cette approche, il est possible de produire directement une ou plusieurs s.o.u. grâce à la simplicité de leurs fonctions de répartition.

Une méthode de production directe d'une s.o.u. consiste à *inverser* la f.r. (3.19), correspondant à une variable *Béta.* En effet, puisque $P_{u\,(i.n)}(x)$ ou $P(x)$ est une (intégrale de) probabilité répartie entre 0 et 1, il **suffit** en premier lieu de produire une v.a.u. **u,** puis d'inverser la fonction (3.19), selon:

$$x = P^{-1}_{u(i:n)}(u) \; , \tag{3.24}$$

la variable x se distribuant comme $u(i:n)$, C'est ce que permet par exemple la fonction *« Betin »* de IMSL (1987), en utilisant l'invocation « x = Betin($u,i,n-i$+l) ».

Les f.r. (3.20 et (3.21) s'inversent aisément afin de produire directement u(l:n) et $u(n:n)$ (voir aussi l'exercice 3.15).

Rabinowitz et Berenson (1974) passent en revue diverses méthodes permettant de produire successivement les s.o.u., à partir des v.a.u. u_1, u_2, etc. La méthode *croissante* proposée par Lurie et Hartley (1972) démarre avec la valeur minimale et procède comme suit:

$$u_{(1)} = 1 - u_1^{1/n} \; ; \quad u_{(i+1)} = 1 - [1 - u_{(i)}]u_{i+1}^{1/(n-i)} \; . \tag{3.25}$$

On comprend que cette méthode, tout comme la méthode *décroissante* de Schucany (1972, voir exercice 3.14), reste avantageuse si l'on a besoin seulement d'une ou de quelques statistiques d'ordre extrêmes; au cas contraire, on appliquera des méthodes globales, plus performantes.

Exercices

3.1 Utilisant une pièce de monnaie lancée plusieurs fois, comme au jeu de « Pile ou Face », construire une séquence de 5 nombres aléatoires rectangulaires (*i.e.* de loi *rectangulaire discrète) x, x* étant un entier entre 1 et 10. Y a-t-il plus d'une méthode pour ce faire ?

3.2 Utilisant l'extrait de la table de chiffres aléatoires présenté à l'exemple 3.1, constituer un échantillon de 10 nombres *x, x* étant un entier entre 1 et 15. Y a-t-il plus d'une méthode pour ce faire ?

3.3 Soit la fonction récursive linéaire (3.6), dotée des paramètres *a = 17,* c = 1 et *m = 100.* Trouver la ou les périodes de cette fonction. La *semence* y_0 influence-t-elle la période ?

3.4 Une variable aléatoire uniforme *x* est bornée de A à B, ou $x \sim U(A, B)$. Montrer que l'espérance et la variance de *x* sont $\mu(x) = \frac{1}{2}(A + B)$ et $0^2(x) = (A - B)^2/12$.

3.5 Prouver (3.11) (a) pour *n = 2* (la distribution de $\bar{u}_{n=2}$ est dénommée loi *triangulaire);* (b) pour tout *n,* par induction.

3.6 En utilisant les équations (3.14) à (3.18), montrer que, pour *n* impair, l'espérance, la variance et les indices de forme d'une *médiane* de *n* v.a.u. sont $\mu = \frac{1}{2}$, $0^2 = 1 / [4(n+2)]$, $y_1 = 0$ et $y_2 = -6/(n+4)$; que, pour *n* pair, les trois premiers moments sont $\mu = \frac{1}{2}$, $0^2 = n / [4(n+1)(n+2)]$, $y_1 = 0$.

3.7 Le *point-milieu* (M) est la moyenne des s.o. extrêmes; il a pour définition:

$$M = \tfrac{1}{2}(x_{(1)} + x_{(n)}) . \qquad (3.26)$$

L'espérance de M dans le cas des v.a.u. est ½. Les moments centraux *impairs* sont tous nuls (i.e. $\mu_r = 0$ pour $r = 2k+1$). Montrer que les moments centraux *pairs* sont donnés par:

$$\mu_r(M) = \frac{r!}{2^r \prod_{i=1}^{r} (n+i)} ; \qquad (3.27)$$

en particulier, la variance est $0^2(M) = [2(n+1)(n+2)]^{-1}$, ce qui fait du point-milieu un estimateur central super-efficient.

3.8 À l'aide de (3.15) et (3.18) pour des v.a.u., trouver ou confirmer (a) la variance du point-milieu M: cf. exercice précédent; (b) la variance de la médiane paire, $Md(x_i;\ n$ pair$) = \frac{1}{2}[x_{(\frac{1}{2}n)}+x_{(\frac{1}{2}n+1)}]$, soit var $= n/[4(n+l)(n+2)]$; *(c)* la variance de l'étendue E (cf. *§.12).*

3.9 Comparer les caractéristiques de distribution des trois estimateurs centraux: moyenne arithmétique, médiane, point-milieu { noter que $y_2(M)=6(n+1)(n+2)]/[(n+3)(n+4)] - 3\}$.

3.10 Montrer que la f.r. d'une moyenne géométrique de v.a.u., dont la densité est donnée par (3.12), correspond à:

$$P_{MG}(x) = 1 - x^n \sum_{i=n}^{\infty} \frac{(nA)^i}{i!}\ ,\ A = (-\log_e x)\ . \qquad (3.28)$$

3.11 Trouver une valeur $u(10{:}10)$ à partir de $u = 0{,}90$ (a) en utilisant (3.21); (b) en exploitant une table de l'intégrale *Bêta,* selon (3.24).

3.12 La loi *F* de Fisher-Snedecor (cf. §4.10) représente notamment la distribution du quotient de deux variances, ces variances étant calculées sur des échantillons indépendants de variables issues d'une même population normale. Cette loi, de variable x_F, a une parenté avec la loi *Bêta* de variable x_β, selon les correspondances:

$$\{\ Bêta(a,b) \rightarrow F(2a,2b)\ \} \quad \text{et} \quad b\cdot x_\beta \div a\cdot(1-x_\beta) \rightarrow x_F \qquad (3.29a)$$

$$\{\ F(a,b) \rightarrow Bêta(\tfrac{1}{2}a,\tfrac{1}{2}b)\ \} \quad \text{et} \quad a\cdot x_F \div a\cdot x_F + b \rightarrow x_\beta \qquad (3.29b)$$

La procédure IMSL *« Fin »* inverse la f.r. du *F;* la séquence « $w \leftarrow Fin(u,2i,2n-2i+2)\ ;\ x \leftarrow i \cdot w / (\ i \cdot w + n-i+1\ )$ » produit la s.o.u. $u(i:n)$ *à* partir de la v.a.u. *u.* (a) Reprendre l'exercice 3.11 à l'aide d'une table de la loi *F.* (b) Par cette même méthode, établir l'intervalle de confiance à *99 %* pour la *médiane* de *5* variables *uniformes U*(0,1) et (c) pour la médiane de *5* variables *normales N(0,1).*

3.13 Par la méthode de Lurie et Hartley indiquée par *(3.25)* et utilisant les v.a.u. { 0,19; *0,80; 0,45 },* trouver $x_N(1{:}9)$, $x_N(2{:}9)$ et $x_N(3{:}9)$, les trois statistiques d'ordre inférieures d'une loi normale N(0,1).

3.14 Schucany (1972) présente une méthode complémentaire à celle de Lurie et Hartley (1972, voir notre (3.25)), en ce sens que les s.o.u. sont produites une à une en partant de la plus grande, $u_{(n)}$. En utilisant les v.a.u. u_n, u_{n-1}, etc. (les indices sont affichés par souci de

clarté mais n'ont pas d'incidence numérique), la méthode procède comme suit:

$$u_{(n)} = u_n^{1/n} \; ; \; u_{(n-i)} = u_{(n-i+1)} u_{n-i+1}^{1/(n-i)} \; . \tag{3.30}$$

Refaire l'exercice 3.13 en appliquant cette fois la méthode de Schucany. *Note:* dans une distribution symétrique, la variable *x(i:n)* a même distribution que la variable [2μ - *x(n-i+1:n)],* μ dénotant l'axe de symétrie (en même temps que la moyenne et la médiane). Par exemple, pour la loi *U*(0,1), μ = ½et les variables *u*(1:9) et 1 - *u*(9:9) ont même distribution.

3.15 David (1981, p. 10 et suiv.) présente la loi de distribution conjointe de deux ou plusieurs statistiques d'ordre. La distribution conjointe de $x_{(1)}$ et $x_{(n)}$ permet de suggérer la procédure suivante, afin de produire les statistiques extrêmes d'un échantillon de *n v.a.u.:*

$$u_{(n)} = u_1^{1/n} \; ; \; u_{(1)} = u_{(n)}\left[1 - u_2^{1/(n-1)}\right] , \tag{3.31}$$

procédure dans laquelle on exploite seules deux v.a.u., u_1 et u_2. Utilisant cette procédure, les v.a.u. $u_1 = 0{,}36$ et $u_2 = 0{,}54$ et une table de la loi *normale,* trouver une valeur de l'étendue (E) de 20 variables normales de distribution *N*(0,1).

Références

ARNOLD, B.C., BALAKRISHNAN, N., NAGARAJA, H.N. (1992). *A first course in order statistics.* New York, Wiley.

DAVID, H.A. (1981). *Order statistics.* New York, Wiley.

DENG, L.-Y., LIN, D.K.J. (2000). Random number generation for the new century. *The American Statistician,* 54, 145-150.

IMSL (1987). *Stat/Library, User 's Manual* (Version 1.0).

KNUTH, D.E. (1969). *The art of computer programming.* Vol. 2: *Seminumerical algorithms.* Reading (MA), Addison-Wesley.

JOHNSON, N.L., KOTZ, S., BALAKRISHNAN, N. (1994, 1995). *Continuous univariate distributions,* Vols. 1 et 2 (2e édition). New York, Wiley.

LAURENCELLE, L. (1993). La loi uniforme: propriétés et applications. *Lettres Statistiques,* 9, 1-23.

LAURENCELLE, L., DUPUIS, F.A. (2000). *Tables statistiques expliquées et appliquées* (2e édition). Sainte-Foy, Le Griffon d'argile.

LURIE, D., HARTLEY, H.O. (1972). Machine-generation of order statistics for Monte Carlo computations. *The American Statistician,* 26, 26-27.

MARSAGLIA, G. (1968). « Random numbers fall mainly in planes », dans *Proceedings of the National Academy of Sciences,* 61, 25-28.

OWEN, D.B. (1962). *Handbook of statistical tables.* Reading (MA), Addison-Wesley.

RABINOWITZ, M. BERENSON, M.L. (1974). A comparison of various methods of obtaining random order statistics for Monte Carlo computations. *The American Statistician,* 28, 27-29.

RAND CORPORATION (1955). *A million random digits with 100,000 normal deviates.* New York, Free Press.

ROHLF, F.J., SONAL, R.R. (1981). *Statistical tables.* New York, Freeman.

RUBINSTEIN, R.Y. (1981). *Simulation and the Monte Carlo method.* New York, Wiley.

SCHUCANY, W.R. (1972). Order statistics in simulation. *Journal of statistical computation and simulation,* 1, 281-286.

Chapitre

4 Production de nombres pseudo-aléatoires obéissant à diverses lois de distribution

4.1 Il est relativement facile d'obtenir, au besoin, une ou plusieurs séries de variables aléatoires uniformes (v.a.u.) de loi $U(0,1)$, soit à partir d'une table de chiffres au hasard, soit grâce à une fonction programmée, comme « RAND ( ) » ou « RANF( ) ». Toutefois, selon l'application que nous avons en vue, il peut nous falloir des variables aléatoires obéissant à une loi de probabilité quelconque, autre que la loi uniforme. Nous devons donc trouver des moyens pour convertir la série des v.a.u. en une série de variables x ayant la distribution voulue. C'est à cette question que répond ce chapitre.

Les moyens disponibles pour convertir une v.a. obéissant à une loi de distribution en une v.a. obéissant à une autre loi sont de fait innombrables et d'intérêt inégal. Nous retiendrons ici les moyens que nous jugeons les plus méritoires sur le plan de la simplicité, de l'efficacité de calcul ou de l'intérêt pédagogique. L'ouvrage de Luc Devroye (1986) constitue une véritable somme sur le sujet (voir aussi Gentle 1998 et Knuth 1969). Sans même tenter de traverser la forêt des distributions statistiques connues, nous jetterons tout de même un regard sur plusieurs d'entre elles. Certaines lois concernent des variables *discrètes,* présentant un ensemble dénombrable de valeurs possibles, *e.g.* $x = 1, 2, \cdots$, et d'autres lois touchent des variables *continues,* pour lesquelles tout intervalle admissible de largeur $\Delta x > 0$ dans un domaine donné a une probabilité non nulle. Il existe une documentation excellente sur les distributions statistiques, à commencer par la série de Johnson et Kotz (Johnson, Kotz et Balakrishnan 1994, 1995; Johnson, Kotz et Kemp 1992), qui constitue la bible du sujet. Evans, Hastings et Peacock (2000) ainsi que Patel, Kapadia et Owen (1976) présentent une compilation sommaire qui couvre un grand nombre de distributions différentes et leurs propriétés. En français, Laurencelle et Dupuis (2000) détaillent une quinzaine de distributions, les plus fréquemment utilisées.

Le répertoire des moyens et méthodes applicables pour produire des v.a. de distribution quelconque contient des moyens *ad hoc,* parfois astucieux, pour quelques lois discrètes ou continues, en plus de proposer des méthodes générales de transformation. Notre exposé sera divisé selon ces grandes catégories, mais nous débuterons par le principe de l'inversion de la fonction de répartition (f.r.), qui fonde la plupart des méthodes.

La documentation citée abonde de preuves et de démonstrations quant à la validité des méthodes et techniques proposées. Nous n'alourdirons pas l'exposé en répétant ces preuves, sauf si le cas le justifie. Le lecteur intéressé trouvera de l'information complémentaire dans la documentation et dans les exercices de fin de chapitre.

4.2 *Principe de l'inversion de la fonction de répartition.* Soit la loi de probabilité f_x = pr(x) et la fonction de répartition (f r.) correspondante $F(x)$ = pr(y ≤ x) = $\int_{-\infty}^{x} f_y \, dy$. Bien sûr, quelle que soit la quantité x inscrite, la fonction $F(x)$ aura pour résultat une valeur bornée de 0 à 1. De plus, si la variable inscrite x obéit à la loi de probabilité f_x, la valeur de $F(x)$ se distribue *uniformément* entre les bornes 0 et 1. Pour résumer,

$$F(x) \sim U(0,1) \text{ ssi } x \sim f_x \, . \qquad (4.1)$$

On établit ainsi une correspondance monotone entre la variable x et la variable u. Par conséquent, si l'on dispose d'une source fiable de v.a.u., on peut en principe convertir chaque v.a.u. u en une variable x, en appliquant la transformation inverse:

$$x \leftarrow F^{-1}(u). \qquad (4.2)$$

Cette correspondance et cette transformation permettent évidemment de changer une v.a. x, de loi f_x, en une v.a. y, de loi f_y, en transitant par la variable u, selon la chaîne de transformations:

$$y \leftarrow {}^{F}F^{-1}{}_{y}[F_x(x)]; \qquad (4.3)$$

dans ces transformations, la variable uniforme se voit dotée du rôle d'opérateur neutre, tout comme l'unité en multiplication et le zéro en addition.

Exemple 4.1 Production d'une v.a. U^2 par inversion

Soit une v.a. continue y de loi $f_y = 1/(2\sqrt{y})$, bornée de 0 à 1; par intégration simple, la f r. est $F(y) = \sqrt{y}$. Cette f.r. s'inverse facilement, et l'on obtient:

$$y \leftarrow u^2 \text{ , pour y } \sim 1/(2\sqrt{y}) \text{ et } u \sim U(0,1) ; \quad (4.4)$$

en raison de cette correspondance, cette loi f_y est aussi désignée loi de U^2.

La f.r. $F(x)$ n'admet pas toujours une inverse, ou bien son inversion n'est pas commode. Il reste que ce principe de l'inversion de la fonction de répartition domine et guide les méthodes de transformation des variables aléatoires.

Variables aléatoires discrètes: techniques ad hoc

4.3 *La loi rectangulaire,R(a,b)*[1]. La v.a. rectangulaire x peut occuper les valeurs $x = a$, a+l, $a+2, \cdots, b$, chacune selon une masse de probabilité constante, égale à 1/(b-a+1). A partir d'une v.a.u., la transformation suivante vers une v.a. rectangulaire fait l'affaire (rappelons que la notation [y] désigne la *partie entière* de *y):*

$$x \leftarrow a + [\, u(b-a+1) \,] . \quad (4.5)$$

Il arrive souvent, dans des situations d'échantillonnage, qu'on doive recruter ou amasser k éléments aléatoires *sans remise* parmi n, c'est-à-dire amasser k valeurs distinctes, sans répétition, tirées d'une distribution R(1,n), $k \leq n$. L'algorithme suivant satisfait ces contraintes:

Sélection uniforme sans remise (4.6)

```
[ Initialisation       ] Pour i = 1 à n Faire v[i] ← i;
[ Cycle de sélection ] Pour i = 1 à k Faire
                          Bloc  Obtenir u;
                                j ← i + ⌊ u(n+1−i) ⌋ ;
[   Valeur x retenue ]          x ← v[j];
                                v[j] ← v[i] ; v[i] ← x
                          Fin_Bloc
```

1. La documentation anglosaxonne désigne indifféremment la loi continue *U(a,b)* et la loi discrète *R(a, b)* sous les noms de loi uniforme ou rectangulaire. Nous optons ici de réserver la dénomination « uniforme » pour la loi continue et la dénomination «rectangulaire» pour la loi discrète.

L'énoncé en caractères gras est facultatif; en incluant cet énoncé, le cycle de sélection peut être répété à volonté, sans qu'il soit utile de ré-initialiser le procédé.

4.4 *La loi binomiale, B(n, л).* La loi *binomiale est* l'une des nombreuses lois décrivant une expérience de Bernoulli, dans laquelle chaque essai donne lieu à un « succès », avec probabilité л, ou à un « échec », avec probabilité 1- л. La v.a. binomiale x compte le nombre de succès obtenus en n essais, les essais étant indépendants les uns des autres. Cette loi est:

$$\text{pr}(x = k) = b_k(n,\pi) = \binom{n}{k}\pi^k(1-\pi)^{n-k}\ ; \qquad (4.7)$$

où $\binom{n}{k} = n!/[k!(n-k)!]$. Il n'existe pas de technique directe pour produire x. L'algorithme suivant:

Production d'une variable binomiale *(4.8)*
$x \leftarrow 0$;
Pour $i = 1$ *à* n Faire si $u_i \leq$ л alors $x \leftarrow$ x + 1 ,

donne la variable x en simulant explicitement un processus de Bernoulli. D'autres techniques, beaucoup plus efficaces, de production binomiale seront discutées à l'exemple *4.2.*

4.5 *La loi géométrique G(л).* Autre mesure d'une expérience de Bernoulli, la variable géométrique x indique le numéro d'essai auquel on obtient le premier succès. Pour cette loi:

$$\text{pr}(x = n) = G_n(л) = (1 - л)^{n-1}\ л \qquad (4.9)$$

La variable x~G(л) s'obtient simplement par:

$$x \leftarrow [\log_e u / \log_e(1 - л) \qquad (4.10)$$

où la notation [y] désigne l'entier x le plus petit tel que $x \geq y$.

4.6 *La loi de Poisson, Po(μ).* Si le taux d'occurrence au hasard d'un événement est ≤ et qu'on l'observe durant t unités de temps, le nombre d'événements observés x est une v.a. de Poisson, a pour paramètre μ = ≤ t et obéit à la fonction de masse:

$$\text{pr}(x = k) = Po_k(\mu) = e^{-\mu}\mu^k/k!\ . \qquad (4.11)$$

Cette loi suppose que les événements sont mutuellement indépendants et de taux constant. L'algorithme suivant produit la variable x ~ Po(μ) :

Production d'une variable de Poisson (4.12)

[Initialisation] $C \leftarrow e^{-\mu}$; $x \leftarrow -1$; $q \leftarrow 1$;

[Cycle de comptage] Répéter Obtenir u ; $x \leftarrow x+1$; $q \leftarrow q^{xu}{}_{x+1}$
Jusqu'à q< C .

Le nombre de tours du cycle de comptage correspond à la quantité x+l, d'où le nombre de tours moyen est μ+l, μ correspondant à la fois à l'espérance et à la variance de la loi *Po(μ)*.

Variables aléatoires continues: techniques ad hoc

4.7 *La loi normale,* $N(\mu,0^2)$. La variation d'un phénomène soumis à de très nombreuses influences non cohérentes ou la valeur d'une caractéristique dont la mesure est contaminée par de nombreuses sources d'erreurs tendent vers la distribution dite normale, de loi:

$$\mathrm{pr}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \quad \pi \approx 3{,}1416. \tag{4.13}$$

La forme standard utilise la variable standardisée z (dénommée aussi « écart réduit »), de paramètres $\mu_Z = 0$ et $0^2{}_Z = 1$, soit $x \sim N(0,1)$, en exploitant les transformations:

$$x \leftarrow z\,\sigma + \mu \;\; ; \;\; z \leftarrow (x - \mu) / \sigma \; . \tag{4.14}$$

Les techniques fournissent ordinairement des variables z, qu'on peut convertir ensuite par l'application de la première transformation (4.14). Cette loi normale étant de grande importance en plusieurs domaines, nous présenterons ici trois techniques spécifiques de production.

La technique la plus connue et la plus simple fut proposée par Box et Muller (cf Devroye 1986). Basée sur la transformation de deux v.a. normales en coordonnées polaires, elle utilise deux v.a.u. à la fois et produit deux v.a. normales indépendantes, x_1 et x_2. Soit les deux v.a.u. u_1 et u_2, alors :

$$\begin{aligned} x_1 &\leftarrow \sqrt{(-2\log_e u_1)} \times \sin(2\pi u_2) \\ x_2 &\leftarrow \sqrt{(-2\log_e u_1)} \times \cos(2\pi u_2) \end{aligned} \tag{4.15}$$

sont deux v.a. normales indépendantes, de loi $N(0,1)$.

Une autre technique semblable, légèrement plus efficace du point de vue de la réalisation informatique, est due cette fois à Box, Muller et

Marsaglia (cf Knuth 1969). Elle produit aussi deux variables z~$N(0,1)$ à partir d'un ou plusieurs couples de v.a.u. u_1 et u_2:

Production de deux v.a. normales N(0,1) (4.16)

Répéter Obtenir u_1, u_2 ;
$y_1 \leftarrow 2u_1 - 1$; $y_2 \leftarrow 2u_2 - 1$;
$S \leftarrow y_1^2 + y_2^2$
Jusqu'à $S < 1$;
$Q \leftarrow (-2\log_e S)/S$;
$x_1 \leftarrow y_1 Q$; $x_2 \leftarrow y_2 Q$.

Grâce au théorème central limite, on sait que la somme de k v.a. de même distribution tend généralement vers une loi normale. Soit $S_k = u_1 + u_2 + \cdots + u_k$, la somme de k *v.a.u.;* l'espérance est $\mu(S_k) = \frac{1}{2}k$ et la variance $0^2(S_k) =$ k/12, d'où $x_k = (S_k - \frac{1}{2}k)\sqrt{(k/12)}$ se distribue approximativement selon $N(0,1)$. Utilisant $k = 12$, la formule devient simplement $x_{12} \leftarrow (S_{12} - 6)$ (voir aussi exercice 4.7).

D'autres procédés, des procédés exacts, plus complexes mais aussi plus performants, sont présentés dans une prochaine section (voir aussi les exercices 4.12 et 4.23).

4.8 *La loi Khi-deux,* $x^2(v)$. Une variable y obéit à la loi x^2_v, ou Khi-deux avec v *degrés de liberté,* si elle est strictement positive et a comme loi de probabilité :

$$\text{pr}(y) = \frac{e^{-y/2} y^{(v-2)/2}}{2^{v/2}\,\Gamma(v/2)} \quad . \tag{4.17}$$

[L'expression Γ_x désigne la fonction « Gamma »; si x est entier, alors $\Gamma_x = (x - 1)!$; voir exercice 4.18]. Cette loi a d'abord été proposée par K. Pearson pour un test approximatif sur un tableau de fréquences d'événements; elle représente aussi la distribution échantillonale de la variance s^2 tirée d'une population $N(0,1)$ et sert dans de nombreux contextes.

Si $x_1, x_2, \cdots, x_k$ sont des variables aléatoires normales, de distribution $N(0,1)$, alors la somme de leurs carrés:

$$y \leftarrow x_1^2 + x_2^2 + \cdots + x_k^2 \tag{4.18}$$

se distribue comme une variable x^2 avec k degrés de liberté. Cette définition statistique du Khi-deux, on le voit, constitue aussi un moyen d'en produire des v.a.

Posant v = 2, la loi (4.17) devient simplement une loi exponentielle de paramètre ½, ou *E(½)* (exercice 4.11). La fonction de répartition de cette loi est $F(y) = 1 - \exp(-½y)$, et son inverse est simplement $F^{-1}(u) = -2\log_e(1-u)$. Comme la v.a.u. u a une distribution équivalente à (1 - u), $y = -2\log_e u$ fournit une v.a. x^2 à 2 degrés de liberté, et:

$$y \leftarrow -2\log_e u_1 - 2\log_e u_2 - \cdots - 2\log_e u_k \qquad (4.19)$$
$$-2\log_e(u_1 \cdot u_2 \cdot \cdots \cdot u_k)$$

donne une v.a. de loi x^2 avec $2k$ degrés de liberté.

4.9 *La loi de Student, t(v).* Une autre loi commune dans les applications statistiques est celle du t de Student, dotée elle aussi du paramètre v, ses degrés de liberté. Par sa définition statistique, elle représente le quotient d'une v.a. $N(0,1)$ sur la racine carrée d'une v.a. x^2 standardisée, c'est-à-dire divisée par ses degrés de liberté. Cette définition:

$$z \leftarrow \frac{x_0}{\sqrt{\dfrac{x_1^2 + x_2^2 + \ldots + x_v^2}{v}}}, \qquad (4.20)$$

dans laquelle les v+l variables x, sont toutes des v.a. $N(0,1)$, permet de produire des t.

La loi du t (voir exercices 4.19 et 4.25-4.27) est assez complexe et d'inversion difficile, sauf pour ses deux cas initiaux, $v = 1$ et $v = 2$. Pour $v = 1$, la t_1 est une loi dite *de Cauchy* standard, qui ne possède aucun moment, notamment pas d'espérance (ou moyenne). On l'obtient selon:

$$z \leftarrow \tan[\,\pi(u - ½)] \sim t_1\,,\ \pi \approx 3{,}1416. \qquad (4.21)$$

Pour le second cas, $v = 2$, le t s'obtient aussi grâce à une seule v.a.u., par la transformation:

$$z \leftarrow (2u - 1) / \sqrt{[2u(1-u)]} \sim t_2\,. \qquad (4.22)$$

4.10 *La loi du quotient F, $F(v_1,v_2)$.* Avec les lois normale, du Khi-deux et du t, la loi du F, dite aussi *loi de Fisher-Snedecor,* complète la famille des distributions de Fisher. Le F représente principalement le quotient de deux estimateurs indépendants de variance et il trouve des applications surtout en *analyse de variance* et en *analyse de régression.* Sa définition statistique présente le F comme le quotient de deux v.a. x^2 standardisées, soit:

$$x \leftarrow \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \sim F(\nu_1,\nu_2) \qquad (4.23)$$

Cette définition permet de produire une variable *F à* partir de deux variables de type x^2. Une autre source de production émane du lien qui existe entre la loi *F* et la loi *Bêta:* voir les exercices 3.12 et 4.28-4.32.

Variables aléatoires discrètes: techniques de tableaux

4.11 La plupart des variables discrètes présentent un ensemble relativement réduit de valeurs possibles; à tout le moins, la masse de probabilité, par exemple 0,99, recouvre à peine une dizaine ou quelques dizaines d'unités. Pour ces cas, le recours à un ou à quelques tableaux, dans lequel ou lesquels des calculs préparatoires ont été faits, accélère grandement la production de v.a. Nous présentons trois techniques de tableaux: le tableau simple, qui réalise une inversion de la f.r.; le tableau développé; le tableau bivoque (ou à alias). Les techniques, calquées d'abord pour des v.a. bornées, peuvent être appliquées à des variables discrètes non bornées, par exemple les v.a. géométrique (§4.5) et de Poisson (§4.6), en les hybridant avec une autre technique.

4.12 *Technique du tableau simple.* La technique la plus simple, pour une variable discrète, consiste à préparer un tableau de la fonction de répartition qui présente les probabilités cumulatives des valeurs de la variable. Soit les valeurs successives possibles $x_1, x_2, \cdots, x_n$ et un tableau $P[x_i]$, tel que:

$$\begin{aligned} &P[x_1] = pr(x_1) \ ; \\ &P[x_2] = pr(x_2) + P[x_1] \ ; \\ &P[x_3] = pr(x_3) + P[x_2] \ ; \\ &\cdots \\ &P[x_n] = pr(x_n) + P[x_{n-1}] \ . \end{aligned} \qquad (4.24)$$

Pour produire une v.a. *x* accordée à la distribution choisie, il suffit d'obtenir la v.a. uniforme *u,* puis de repérer la valeur x_i associée telle que $P[x_{i-1}] < u \leq P[x_i]$: c'est la valeur $x = x_i$ produite. Comme on voit, cette technique réalise en quelque sorte l'inversion de la fonction de répartition. L'algorithme suivant illustre le procédé.

Repérage de la v.a. discrète x_i par un tableau simple *(4.25)*

{Le tableau P[1..n] contient les probabilités cumulatives telles que $P[x_i] = pr(x \leq x_i)$ et $P[x_n] = 1$, la v.a. étant bornée par $x_1 \leq x \leq x_n$ }

Obtenir *u; i* ←1 ;
Tant que u > P[x_i] Faire i ← i + 1 ;
Produire x = x_i .

4.13 Il y a différentes manières de mettre en oeuvre la technique du tableau simple, certaines étant plus efficaces que d'autres. Si la distribution de probabilités possède un mode, une manière de repérer x_i consiste à comparer la variable u d'abord à P[*x*(mode)], puis à diriger la comparaison en deçà de *x*(mode) ou au-delà, selon le cas; c'est un repérage par bissection discrète (exercice 4.13). Le point-milieu peut aussi servir de pivot à la place du mode.

Une autre manière d'utiliser la technique du tableau consiste à établir la série des probabilités pr(x_i) par ordre décroissant de probabilité, puis de calculer leurs valeurs cumulatives $P'[x_j]$, *j* dénotant maintenant le rang de probabilité (plutôt que la valeur de la v.a. elle-même); un tableau conjugué des valeurs, disons I[*j*], est alors requis. Posant $P'_o = 0$, le repérage systématique de l'inégalité $P'_{j-1} < u \leq P'_j$, depuis $j = 1$ jusqu'à $j = n$, accélérera en moyenne le repérage de la v.a. à produire, selon I[j] → *x*.

4.14 *Technique du tableau développé.* C'est au cycle de repérage, destiné à satisfaire une double inégalité par rapport à la v.a. u, qu'est imputable le temps d'exécution principal de la technique du tableau simple. La technique du tableau développé élimine ce repérage en découpant la fonction de répartition en petits morceaux, chaque morceau étant assigné à une seule valeur de la variable *x*.

Soit la variable *x,* ses valeurs discrètes *a, b, c,* ¨, et leurs probabilités respectives p(a), p(b), p(c), etc. Prenons un tableau, ou vecteur, T[O...K -1], comportant K cellules, K > 1. La quantité K •*p(a)* est un nombre à partie fractionnaire, de forme *« e,f »;* dans ce nombre, *«e»* est un entier *(e* = 0,1, etc.) et représente le nombre d'unités, ou cellules du tableau T, à assigner à la valeur a^2*;* quant à «*f*», la fraction, c'est un résidu de probabilité, associé lui aussi à la valeur *x* = *a.* En utilisant toutes les valeurs de *x*, c'est-à-dire *a, b, c,* etc., K' = $e_a + e_b + ...$ cellules de T seront nommément assignées, et l'ensemble des (K - K') = $f_a + f_b + ...$ cellules non assignées encaissera la somme des résidus.

2. L'ordre (ou la place) d'assignation des valeurs de *x* dans le tableau T n'importe pas, à la double condition que seules les K' premières cellules soient assignées et que le nombre de cellules assignées à « a », **par exemple, soit** [K•p(a)] , **que ces cellules soient adjacentes ou non.**

Pour produire une variable *x*, on obtient d'abord une v.a. u et on calcule *j* = [K• *u]; si j* < K', alors on se trouve dans la portion des valeurs assignées du vecteur et la correspondance x ← T[*j*] produit immédiatement la variable voulue. D'autre part, l'événement j ≥ K' exige que la variable soit repérée dans l'espace résiduel, par une méthode ou l'autre, par exemple un autre tableau, en utilisant cette fois les probabilités résiduelles p'(a) = f_a / (K-K').

L'efficacité de la technique du tableau développé dépend essentiellement du choix de K, la taille du tableau principal. Pour une distribution de probabilités donnée, certains choix peuvent n'engendrer nuls résidus de sorte que la technique serait alors absolument efficace, en produisant la v.a. voulue en un seul coup. La somme maximale de résidus possibles approche *n/K, n* étant le nombre de valeurs distinctes de *x*; admettant une distribution uniforme des résidus f_x en *U*(0,1), la probabilité d'obtenir tout de suite une v.a. x sans devoir examiner l'espace résiduel, est, en général:

$$\text{pr}\{ \text{ obtention de } x \text{ sans repérage } \} \approx 1 - n/(2K) . \qquad (4.26)$$

Quel que soit le choix de K, on peut encore aider à l'efficacité globale en choisissant une technique optimale pour le repérage de *x* dans l'espace résiduel.

Exemple 4.2 Production d'une v.a. binomiale par tableau

Nous désirons produire x~B(8,⅓), soit une v.a. binomiale de paramètres n = 8 et л = ⅓. La méthode de base, présentée en §4.4, simule un processus de Bernoulli; elle requiert *n* v.a. u_i pour autant de cycles de comparaison et d'addition de 1. Symbolisons le coût de production de chaque v.a. x par une équation indicative,

$$\$C = v_1(U) + v_2(Op) , \qquad (4.27)$$

dans laquelle équation « v_1 » dénote le nombre de variables *u* requises, et « v_2 » indique grossièrement le nombre d'opérations informatiques globales. Le coût de notre méthode de base serait donc simplement $C(base) = *n*(U) + *n(Op),* en particulier 8(U) + 8(Op) pour notre exemple spécifique.

La technique du tableau simple convient parfaitement pour produire une variable binomiale. La distribution de probabilités p(*x*), établie d'après (4.7), et la fonction de répartition *P(x)* sont:

x	0	1	2	3	4	5	6	7	8
$p(x)$	,039	,156	,273	,273	,170	,068	,017	,002	,000
$P(x)$	,039	,195	,468	,741	,912	,980	,997	,999	1,0

Utilisant une seule variable u, l'algorithme de repérage (4.25) examine à tour de rôle P(0), P(1), etc. jusqu'à obtenir $u \leq P(x)$, auquel cas x est la v.a. binomiale produite. Le nombre moyen de valeurs $P(x)$ à examiner est évidemment $1 + E(x) = 1 + n\pi$, ici 3,67. La formule de coût pour cette réalisation-ci sera donc \$C (tableau simple) = 1(U) + 3,67(Op). Comme l'indiquent ces calculs, le coût est proportionnel à n et à π. On peut néanmoins améliorer la performance de cette technique et en réduire le coût: voir les exercices 4.8 et 4.13.

La distribution de probabilités ci-dessus sert encore pour mettre en oeuvre la technique du tableau développé. Prenons un tableau T de taille K = 50 [dans un cas pratique, K pourra être plus grand, par exemple 100, 500 ou 10000]. En effectuant le produit $K \cdot p(x) = (e,f)_x$, comme ci-dessous, on obtient pour chaque valeur x le nombre de cellules assignées *(«e»)* dans le tableau et le résidu associé *(«f»)*; les résidus, normalisés de façon à ce que leur somme soit 1, constituent la distribution des probabilités résiduelles p'(x), qu'on peut traiter à volonté, par une technique de tableau ou une autre.

x	0	1	2	3	4	5	6	7	8
e_x	1	7	13	13	8	3	0	0	0
f_x	,951	,804	,656	,656	,535	,414	,854	,122	,008
$p'(x)$	,190	,161	,131	,131	,107	,083	,171	,024	,002

Des 50 cellules du tableau T, il y aurait donc $\sum e_x = 1 + 7 + \ldots + 0 =$ 45 cellules assignées, qu'on peut représenter comme suit:

Tableau T développé pour la production d'une binomiale $B(8,1/8)$

T[0] =

0	1	1	1	1	1	1	1	2	2
2	2	2	2	2	2	2	2	2	2
2	3	3	3	3	3	3	3	3	3
3	3	3	3	4	4	4	4	4	4
4	4	5	5	5	–	–	–	–	–

= T[49]

L'algorithme de production implique d'abord d'obtenir u, puis $j \leftarrow \lfloor 50u \rfloor$; si $j < 45$, on livre tout de suite $T[j] \rightarrow x$; cet événement

a pour probabilité 45/50 = 0,9. Si ≥ 45, il reste à produire x en exploitant les probabilités résiduelles p'(x). On peut le faire, par exemple, en appliquant la méthode de tableau simple, avec les probabilités résiduelles cumulatives P'(x) et une nouvelle variable u'. Le coût de production, dans ce cas, serait $C (tableau développé) = 1(U) + 0,9[0 (Op)] + 0,1 [1 (U)+3,87(Op)] ≈ 1,1(U)+0,4(Op). Ce coût approximatif est minime; pour l'amener vers la limite générale de 1(U), le moyen le plus bête et le plus efficace est encore d'augmenter K, la taille du tableau développé. On peut aussi éliminer le repérage en appliquant aux probabilités résiduelles la technique du tableau bivoque, présentée à la section suivante.

4.15 *Technique du tableau bivoque (ou de l'alias).* La technique du tableau bivoque, créée par A. J. Walker (voir Devroye 1986; Gentle 1998), permet de produire en deux coups une v.a. discrète obéissant à n'importe quelle distribution bornée. Prenons l'exemple simple d'une variable x à trois états, $P(x_o) = P_o$, $P(x_1) = P_1$, $P(x_2) = P_2$ et $P_o + P_1 + P_2 = 1$. Pour générer la v.a. x = [0. 1. 2] avant la distribution de probabilités P = [⅓, ⅓, ⅓], il suffit de

faire

$$\lfloor 3u \rfloor \rightarrow j \text{ et } x[j] \rightarrow x ,$$

voire:

$$\lfloor 3u \rfloor \rightarrow x ,$$

une solution évidente puisque les probabilités sont réparties également en x. Prenons maintenant une autre distribution, soit x = [0, 1, 2] avec p = [¼, ¼, ½]. La fonction « [3u] », comme ci-dessus, produit un entier dans [0, 1, 2]. Cependant, l'événement « [3u] = 0 », par exemple, advient avec probabilité ⅓, une fréquence trop forte pour x = 0, alors que, pour « [3u] = 2 », la fréquence relative est insuffisante. Pour le cas « [3u] = 0 », la valeur candidate, x = 0, arrive avec probabilité ⅓, la probabilité voulue étant $p(x_o)$ = ¼; il nous faut donc tester l'admissibilité de x_o, ce en recourant à une nouvelle v.a. uniforme, u', et à un seuil de probabilité q_o tel que $q_o x ⅓ = p_0$ ou q_o = ¾. Nous acceptons alors le candidat x = 0 si u' ≤q_o, ce qui se produit avec probabilité finale de ¼ (= ⅓x ¾) ; si, au contraire, u' > q_o, une autre valeur, l'alias de x = 0, est produite, de manière à ce que, au total, chaque valeur de x arrive au hasard avec sa probabilité finale assignée.

Sommairement, pour une distribution x = [x_o, x_1, ..., x_{n-1}] avec probabilités p = [P_o, P_1, ..., P_{n-1}], on effectue d'abord $q_i = n \cdot p_i$, obtenant:

$$[\, q'_0, q'_1, \ldots, q'_{n-1} \,] ;$$

dans ce vecteur, les quantités $q'_i = q_i \leq 1$ constituent les seuils d'admissibilité des résultats x_i correspondants. Pour les quantités $q'_i > 1$ (s'il y en a), elles touchent chaque valeur x_i à probablité $p_i > \frac{1}{n}$; l'excédent de

probabilité, $q'_i - \frac{1}{n}$, doit être réalloué à une ou plusieurs valeurs de x à q_j fractionnaires, constituant ainsi leur alias. Cette réallocation entraîne un remaniement des q'_i et la définition des alias attachés aux $q_i < 1$, ce qui donne lieu aux deux vecteurs:

$$\begin{array}{rl} \text{Seuils :} & [\ q_0 \quad q_1 \quad \ldots \quad q_{n-1}\ ] \\ \text{Alias :} & [\ \text{A}(0) \quad \text{A}(1) \quad \ldots \quad \text{A}(n-1)\ ] \end{array} \tag{4.28}$$

La probabilité de production d'une valeur x résulte alors, soit de sa sélection initiale par le mécanisme « $i \leftarrow [n \bullet u]$ » qui fournit x[i] à condition que $u' \leq q_i$, soit de la sélection d'une ou plusieurs autres valeurs $j \neq i$, *à* condition que $u' > q_j$ et que l'alias de x_j soit x_i. Une fois les deux valeurs *q[i]* et *A(i)* en place, l'algorithme suivant produit notre v.a. x.

Production d 'une v.a. discrète x_i (4.29)
par la technique du tableau bivoque

{ Le vecteur *q[O..n* -1] contient les seuils d'admission des valeurs $x[i]$ et le vecteur *A[O..n* -1] contient leurs alias }

Obtenir u et u' ;
i & *[n x u]* ;
Si $u' \leq q[i]$ alors $x \leftarrow x[i]$ sinon $x \leftarrow A[i]$.

Symboliquement, le coût d'exécution revient à 2(U) + 3(Op), un coût pratiquement fixe, peu élevé, et n'impliquant que deux tableaux ou vecteurs de longueur n. La préparation des vecteurs, qui consiste à remanier les quantités $q'_i = n \bullet p_i > 1$ pour en répartir l'excédent de probabilité et à définir les alias A_i associés aux valeurs $q_i < 1$, peut être faite par l'algorithme élémentaire suivant.

Préparation des vecteurs q[i] et A[i] (4.30)
pour la technique du tableau bivoque

{ Utilise le vecteur de probabilités $p[0..n-1]$ et définit deux autres vecteurs de même longueur $q[\ ]$ et A[ ] }

{ Définir aussi UN = 1+ε, ε > 0, couvrant les erreurs de précision }

Pour $i = 0$ à $n-1$ Faire $q[i] \leftarrow n \times p[i]$, $A[i] \leftarrow -1$;
$q[n] \leftarrow 2$;
$i \leftarrow 0$; $j \leftarrow 0$;

```
                    Répéter
[ Trouver q[i] > 1]  Tant que q[i] ≤ UN  Faire i ← i+1 ;
                       Si i < n  { Cf. sentinelle en q[n] } alors
                       Bloc
[ Répartir l'excès,]   Tant que q[i] > UN  Faire
[        q[i]−1 ]        Bloc
                           Répéter  j ← (j+1) mod n ;
[ Trouver q[j] < 1]        Jusqu'à q[j] < 1 et A[j] < 0
[ Faire de i l'alias de j ]  A[j] ← i ; q[i] ← q[i] − (1−q[j])
[ et remanier q[i] ]       Fin_Bloc
                       Fin_Bloc
                    Jusqu'à i = n .
```

D'autres algorithmes de répartition des *q[i]* excédentaires et des alias sont possibles, d'aucuns un peu plus rapides (et plus complexes): voir les exercices 4.14-4.17.

Exemple 4.2 (suite)

La technique du tableau développé utilise un vecteur de longueur K, T[O..K-1], dont les K' premières cellules sont occupées par des valeurs de x en proportion de leurs probabilités respectives p(x). La technique engendre aussi un espace résiduel, une distribution de probabilités résiduelles p'(x), à laquelle l'algorithme doit avoir recours si j = [K•u] ≥ K'.

Reprenons notre v.a. binomiale x de loi B(8,⅓) et le tableau de taille K = 50, avec K' = 45 et les probabilités résiduelles p'(x) fournies. Appliquant la technique du tableau bivoque à ces p'(x) et grâce à l'algorithme (4.30), nous obtenons les vecteurs q (« seuils ») et A (« alias ») suivants:

x	0	1	2	3	4	5	6	7	8
$q(x)$	,636	,467	,815	,646	,963	,747	1	,216	,018
A(x)	2	3	6	6	0	0	–	0	1

Pour illustrer l'algorithme (4.30), considérons, pour x = 0, p'_0 = 0,190 et $q'_0 = 9 \times p'_0$ = 1,710. Cette « probabilité » excédentaire (> 1) doit être répartie dans la ou les valeurs de x pour lesquelles $q'_x < 1$. La première trouvée, x = 4, a $q = 9 \times p'$ = 0,963. Assignant «0» comme alias de x = 4, nous en déduisons le complément de probabilité

de q'_0, soit q'_0- (1-q_4) = 1,710 - 0,037 = 1,673. Il faut encore soustraire. Nous trouvons ensuite x = 5, q_5 = 0,747; x = 5 prend donc «0» comme alias, et (1- q_5) est déduit de q'_0, *qui* devient 1,42. Encore, et nous trouvons x = *7,* q_7 = 0,216; l'alias «0» est placé aussi en x = 7 et, après déduction, *q;* devient q_o = 0,636. Il faut alors passer à x = *1,* q_i = 9x0,161 = 1,449 > 1, ainsi de suite.

L'algorithme complet, utilisant trois v.a.u. u_1, u_2 et u_3, serait alors:

Production d'une v.a. discrètex = [0, 1, ..., n-1] (4.31)
par une méthode hybride utilisant
les techniques du tableau développé et du tableau bivoque

{ Utilise le vecteur de probabilités *p[0..n* -1], un tableau T[O..K -1], K'& K et les vecteurs *q[0..n* -1] et *A[0..n -1]* } ;

{ Tous les tableaux sont censés préparés }

```
Produire u1 ;
j ← ⌊K·u1⌋ ;
Si j < K' alors  x ← T[j]
         sinon  Produire u2, u3 ;
                i ← ⌊n·u2⌋ ;
                Si u3 ≤ q[i] alors x ← i
                sinon x ← A[i] .
```

L'analyse symbolique indique un coût de (3−2ω)(U) + (6−2ω)(Op), où ω est la probabilité de succès de l'interrogation du tableau T[ ]. Appliquant l'approximation (4.26), nous obtenons un coût de (1 + *n*/K)(U) + (4 + *n*/K)(Op). Selon la distribution $p(x)$ considérée et la taille du tableau développé K, le coût de production par l'algorithme (4.31) se situe dans l'intervalle:

1(U) + 4(Op) ≤ Coût de production < 3(U) + 6(0p) ,

un rendement payant pour qui se donne la peine de mettre sur pied un tel algorithme.

La technique du tableau bivoque peut être appliquée directement pour produire la v.a. voulue, telle notre v.a. binomiale x ~ B(8, ⅓).

Variables aléatoires continues: techniques de pseudo-inversion, rejet et composition

4.16 Les techniques de tableaux, appropriées à la production de v.a. discrètes, sont peu applicables aux v.a. continues, pour lesquelles l'inversion mathématique de la fonction de répartition *F(x)* reste optimale. Il y a toutefois des palliatifs à l'inversion proprement dite de la f.r. Nous considérons d'abord des techniques de pseudo-inversion, soit l'inversion *par repérage* et l'interpolationinverse. La fonction inverse peut elle-même être approchée par un polynôme ou une expansion de Taylor (voir par exemple l'exercice 4.12). Des paragraphes subséquents abordent d'autres techniques spécialisées, notamment le *rejet* et la *composition.*

4.17 *Inversion par repérage.* Soit u_o, une valeur donnée d'une v.a. uniforme telle que $0 < u_o < 1$, et la v.a. *x,* de densité f_x et de fr. F(*x*). Par définition de f.r., $F(-\infty) = 0$ et $F(\infty) = 1$. *Si x* et f_X sont continues, il existe une et une seule valeur x_o telle que $F(x_o) = u_o$; une fois repérée, cette valeur x_o est la v.a. produite. Il s'agit donc de repérer la valeur réelle x_o: c'est en fait la solution de l'équation:

$$\{ x \mid F(x_0) - u_0 = 0 \} , \tag{4.32}$$

et de nombreuses *méthodes numériques y* sont applicables (voir par exemple Gerald et Wheatley 1984).

Parmi les méthodes générales applicables pour solutionner (4.32), c'est-à-dire pour produire la valeur *x* correspondant à la v.a. *u*, la méthode de *Newton-Raphson* procède en formant successivement des valeurs x_1, x_2, ..., soit des approximations récursives, à partir d'une valeur initiale arbitraire x_o comme suit:

$$x_{n+1} \leftarrow x_n - [F(x_n) - u]/f(x_n) . \tag{4.33}$$

Assez rapidement, $x_n \rightarrow x$ telle que $F(x) = u$, donnant la valeur cherchée. La méthode de Newton-Raphson, dite aussi méthode de la tangente, suppose que, en plus de F(*x*), on possède aussi sa dérivée $f_x = dF(x)/dx$ et une valeur initiale x_o; à défaut de fournir une valeur x_o ; approximative, le mode de x ou son point-milieu peuvent faire l'affaire.

Une autre méthode numérique, la plus sûre, est la *bissection.* Elle consiste à fournir d'abord un intervalle (x_1, x_s) le plus étroit possible et tel qu'il englobe certainement la valeur cherchée x, selon $x, < x < x_s$. Utilisant le point-milieu $x_M = \frac{1}{2}(x_1 + x_s)$, on vérifie ensuite si *x* se situe dans le premier demi-intervalle (x_1, x_M) ou dans le second (x_M, xs), selon que $F(x_M)$

& u ou non, ce qui détermine un nouvel intervalle, de moitié plus court. On procède ainsi de suite jusqu'à la précision voulue.

La méthode dite *de la fausse position (ou regula falsi)* fait pendant à la méthode de Newton-Raphson, en remplaçant f_x par une différence finie $[F(x_s)-F(x_I)]/(x_s-x_I)$. Elle débute avec deux valeurs embrassantes x_1 et x_2, telles que $F(x_1) < u < F(x_2)$, puis les valeurs x_n, d'approximations successives, sont données par:

$$x_n \leftarrow x_{n-1} - (x_{n-1}-x_{n-2})[F(x_{n-1})-u]/[F(x_{n-1})-F(x_{n-2})] \ , \quad (4.34)$$

pour $n = 3, 4$, etc., jusqu'à la précision voulue.

Le lecteur aura remarqué que toutes ces méthodes sont itératives et requièrent un re-calcul de la fonction $F(x)$ et, dans le cas de Newton-Raphson, de f_x. Le nombre de cycles d'approximation dépend: (1) de la méthode employée, Newton-Raphson étant la plus efficace; (2) de la précision voulue dans x; (3) de la valeur initiale fournie. La précision de x, c'est-à-dire le nombre de chiffres significatifs exigés, constitue sûrement le facteur le plus pesant parmi les trois mentionnés. En raison de ce facteur, il sera parfois possible d'accélérer énormément la production de v.a. en exigeant seulement une précision raisonnable, dictée par le contexte.

Exemple 4.3 Production d'une v.a. normale N(0,1) par pseudo-inversion

Soit la valeur uniforme u = 0,62. Pour produire $x \sim N(0,1)$ par la méthode de Newton-Raphson, on peut employer la valeur initiale $x_o = 4(u - ½)$, ici $x_o = 0,48$. Le calcul de $F(0,48)$, par une méthode ou l'autre (§8.2 et exemple 8.1), donne 0,6844, avec $f(0,48) = 0,3555$. La première approximation est alors $x_1 = 0,48 - [0,6844-0,62] / 0,3555 \approx 0,2989$, puis $x_2 = 0,2989 - [0,6175-0,62] / 0,3815$ 0,3055. Puisque F(0,3055) 0,6200, le tout s'arrête là, avec 4 chiffres décimaux de précision! L'utilisation de $x_o = 0$ eût rendu la même performance, dans le cas présent.

Illustrons maintenant la méthode de bissection, pour laquelle nous devons fournir un intervalle embrassant x. Pour la loi $N(0,1)$, de domaine doublement infini, un choix s'impose. Comme la loi $N(0,1)$ est symétrique sur 0, avec $F(0) = ½$, $x_i = 0$ constitue une borne inférieure indiquée pour notre exemple avec $u = 0,62$ (ce serait une borne supérieure si on avait observé $u < ½$). Quant à la borne supérieure x_s, on peut employer une inégalité basée sur la loi t_1, ou loi de Cauchy (cf. (4.22)), soit $F_{Cauchy}(x) \leq F_N(x)$ pour $x \geq 0$. Ainsi,

pour une valeur $u \geq ½$ donnée, $F^{-1}_{\text{Cauchy}}(u) \geq F^{-1}_{\text{N}}(u)$, et $\tan[\pi(u - ½)]$ constitue une borne supérieure pour la valeur normale cherchée; ici $x_S = \tan 0{,}377 \approx 0{,}3959$. Le premier intervalle, (0; 0,3959), a pour point-milieu 0,1980, d'intégrale normale 0,5785; la valeur x cherchée occupe donc le demi-intervalle supérieur (0,1980; 0,3959), ayant pour point-milieu 0,2970 d'intégrale 0,6168, etc. Le 12e point-milieu calculé, $x_{12} = 0{,}3055$, a pour intégrale $\approx 0{,}6200$; c'est donc la valeur cherchée.

La méthode *de la fausse position* réclame elle aussi un intervalle de départ et deux évaluations de $F(x)$ à chaque itération[3]; elle est généralement plus efficace que la bissection. A preuve, pour notre valeur $u = 0{,}62$ et les valeurs de départ $x_1 = 0$ et $x_2 = 0{,}3959$, l'expression (4.30) nous fournit $x_3 = 0{,}3959 + (0{,}3959 - 0)\ (0{,}6539 - 0{,}62) \div (0{,}6539 - ½) \approx 0{,}3087$, puis $x_4 = 0{,}3055$, déjà la valeur cherchée.

4.18 *Interpolation inverse.* La fonction de répartition $F(x)$ étant connue et ayant une étendue de 0 à 1, on peut établir d'avance une table de correspondance $F(x) \rightarrow u$, permettant l'interpolation inverse, de type $u \rightarrow x$. Étant donné une valeur u_o quelconque et une table préparée à cet effet, il s'agit de repérer dans la table certaines valeurs de référence, disons $F_1, F_2, \cdots, F_k$ telles que u_o est inscrite dans leur domaine, puis d'utiliser un polynôme d'interpolation, de type:

$$x_0 = \text{Poly}\ \{\ x_1{:}F_1\ ;\ x_2{:}F_2;\ \cdots;\ x_k{:}F_k\ \mid\ u_0\ \}\ ; \qquad (4.35)$$

l'ordre k du polynôme et la fabrication de la table sont dictés par les impératifs de simplicité et de précision.

Étant loin la méthode la plus simple, l'interpolation linéaire à intervalles égaux est aussi la plus rapide. Elle consiste à découper le domaine de la v.a. u en N bandes de largeur 1/N, et à trouver, par précalcul, les valeurs de référence $x_j = F^{-1}(j/N)$, pour $j = 1$ à N-1. La valeur u_o étant donnée, on trouve immédiatement $j = [Nu_{0]}$, x_j . et x_{j+1}, puis la valeur interpolée, selon:

$$x_0 = x_j + (x_{j+1} - x_j)(N \cdot u_0 - j)\ . \qquad (4.36)$$

La précision de ce mode d'interpolation dépend de la taille (N) du tableau de correspondance et inversement de l'amplitude de la densité f_x. Lors du

3. La fonction (4.34) étant récurrente, la valeur calculée de $F(x_{n-1})$ peut être mémorisée puis ré-utilisée en tant que $F(x_{n-2})$ au cycle suivant.

pré-calcul du tableau, l'erreur maximale pour une valeur de N peut facilement être établie empiriquement. Enfin, notons que plusieurs distributions statistiques ne sont pas bornées à gauche ou à droite. Dans un cas pareil, l'interpolation inverse s'appliquerait néanmoins, en la restreignant à un domaine fini, disons (x_i, x_s), et qui contient la masse de probabilité, comme pr($x_I \leq x \leq x_s$) ≈ 0,99 ou 0,999. L'obtention des variables dans le domaine résiduel, un événement de moindre probabilité, peut alors se faire par repérage itératif ou par une technique spécialisée conçue pour les ailes de la distribution.

D'autres modes que l'interpolation linéaire régulière peuvent être mis en oeuvre tout aussi efficacement; l'idée est ici de trouver une fonction interpolante plus souple, rendant une bonne précision tout en gardant le tableau d'interpolation raisonnablement petit.

L'interpolation parabolique inverse, par exemple, utilise la fonction du second degré:

$$x_0 = Au_0^2 + Bu_0 + C \ , \qquad (4.37)$$

les paramètres étant établis sur un triplet $\{ x_{j-1}{:}F_{j-1};\ x_j{:}F_j;\ x_{j+1}{:}F_{j+1} \}$, j satisfaisant $|u_0 - F_j| \leq |u_0 - F_i|$. Si le domaine de u est découpé, comme ci-dessus, en N bandes de largeur 1/N, alors, en utilisant:

$$D_1 = x_{j+1} - x_{j-1} \ ; \ D_2 = x_{j+1} - 2x_j + x_{j-1} \ ,$$

on a:

$$A = \tfrac{1}{2}D_2N^2 \ ;$$

$$B = \tfrac{1}{2}D_1N - D_2F_jN^2 \ ;$$

$$C = x_j - \tfrac{1}{2}D_1F_jN + \tfrac{1}{2}D_2F_j^2N^2 \ .$$

Bien sûr, comme pour l'interpolation linéaire, on peut accommoder l'interpolation parabolique à des intervalles inégaux de l'argument u (voir l'exercice 4.20).

Plus performante encore, en précision comme en vitesse, serait l'interpolation linéaire régulière sur une variable transposée, la transposition ayant pour but d'aplatir quelque peu la densité résultante, $tr(f_x)$. Par exemple, pour la famille des distributions issues de la loi normale, soit la normale elle-même, la t, la x^2, la F, les densités f_x chutent considérablement en queue de distribution; l'interpolation linéaire inverse, suffisamment précise dans la bande u = (0,55; 0,56), devient lourdement inadéquate dans la bande de même largeur (0,98; 0,99). On peut alors faire l'interpolation grâce à une correspondance transformée $x{:}\mathrm{tr}(F_x)$, laquelle ramène les intervalles (x_j; $x_{j\cdot+1}$) à des largeurs plus comparables dans le tableau d'interpolation.

Exemple 4.4 Production d'une v.a. normale N(0,1) par interpolation

Illustrons l'interpolation linéaire (à intervalles égaux) par un petit exemple, appliqué à la loi $N(0,1)$. Nous avons une table de la f.r. normale $F(x)$ découpée en 20 tranches de largeur $\Delta u = 0{,}05$. Comme cette distribution est symétrique, la moitié supérieure suffit. Ainsi, $F_{10}^{-1}(0{,}50) = 0$; $F_{11}^{-1}(0{,}55) = 0{,}12566$; $F_{12}^{-1}(0{,}60) = 0{,}25335$; $F_{13}^{-1}(0{,}65) = 0{,}38532$; $F_{14}^{-1}(0{,}70) = 0{,}52440$; etc. Le tableau s'arrête à $u = 0{,}95$ (et 0,05), le domaine résiduel étant double, (0,95; 1) et (0; 0,05), avec une probabilité d'incidence de 0,1.

Soit $u = 0{,}67$: quelle est la v.a. x correspondante, selon la loi normale $N(0,1)$? La bande concernée, déterminée par le nombre de bandes N et u, est ancrée sur $j = \lfloor Nu \rfloor = 13$ et $j+1 = 14$, soit $F_{13}^{-1}(0{,}65)$ et $F_{14}^{-1}(0{,}70)$. Utilisant (4.36), nous obtenons $x_0 = 0{,}38532 + (0{,}52440 - 0{,}38532)(20\times0{,}67 - 13) = 0{,}44095$. L'abscisse exacte correspondant à $u = 0{,}67$ est $x(0{,}67) = 0{,}43991$, l'erreur de l'estimation linéaire étant de 0,001. Notons, du même souffle, que la production de la valeur interpolée est (en moyenne) presque instantanée.

Pour la même v.a. $u = 0{,}67$, l'interpolation parabolique inverse, selon (4.37), à partir de la même table de correspondance, produit $x_0 = 0{,}440099$; l'erreur est ici de 0,0002, soit 5 fois plus petite et presque négligeable.

Exemple 4.5 Production d'une v.a. t_6 par interpolation

L'interpolation linéaire en intervalles transposés égaux sera appliquée ici pour produire des v.a. de la loi t_6, le t de Student avec 6 degrés de liberté. La f.r. de cette variable, d'après Patel, Kapadia et Owen (1976), peut se calculer par:

$$F_{t_6}(x) = \tfrac{1}{2} + \tfrac{1}{2}A\sqrt{B}\cdot[1 + \tfrac{1}{2}B + \tfrac{3}{8}B^2] ,$$

en exploitant $A = x/\sqrt{6}$ et $B = 6/(6 + x^2)$. Unimodale et symétrique, la forme du t_6 s'intercale entre la loi de Cauchy (ou t_1) et la loi normale. Tentons une transformation logarithmique pour $X(u)$, soit exponentielle sur u, comme suit:

$$z_u = (e^{u-\frac{1}{2}} - 1) / (e^{\frac{1}{2}} - 1),\ u \geq \tfrac{1}{2} ,$$

le cas $u < ½$ étant traité complémentairement [en estimant $-x(u)$]. La variable z, une transformation monotone de u, varie de 0 à 1 pour u allant de ½ à 1, et son domaine peut être découpé en intervalles égaux, disons N = 10 intervalles (ou $\Delta z = 0{,}1$) pour cet exemple. Ainsi, à z = 0 (ou $u = 0{,}5$), correspond $x = 0$; à z = 0,1 (ou $u = 0{,}56286$), correspond $x = 0{,}165095$; etc. Notons que, ici, l'intervalle résiduel est délimité par la borne $z = 0{,}9$, qui correspond à u = 0,959858; la probabilité d'incidence dans l'intervalle résiduel (haut et bas) est donc $2\bullet(1 - 0{,}959858) \approx 0{,}08$.

Prenons $u_o = 0{,}75$. La variable z équivalente est $z_{0\ 75} = 0{,}437823$. Les valeurs de référence embrassantes, z = 0,4 et z = 0,5, ont pour abscisses x respectives 0,651953 et 0,830320. L'estimé par interpolation, utilisant encore (4.36) mais cette fois avec z plutôt que u, est:

$$x_0 = 0{,}651953 + (0{,}830320 - 0{,}651953)(10 \times 0{,}437823 - 4)$$
$$= 0{,}719417 .$$

La valeur correcte, précise à 6 décimales, est $x(0{,}75) = 0{,}717558$. Eussions-nous utilisé l'interpolation linéaire simple, avec 10 bandes de 0,5 à 1 (et $\Delta u = 0{,}05$), le résultat eût été $x_o = 0{,}729542$, l'erreur étant 6 fois plus grosse.

4.19 *La méthode de rejet.* Conçue expressément pour la production de variables aléatoires, la méthode de rejet comporte deux phases: (1) produire une variable x candidate, issue d'une loi parraine g_x, puis (2) tester si cette variable est admissible sous la loi cible f_x. La méthode réclame au moins deux v.a.u., u_1 et u_2, une par phase. Le tout, incluant le test d'admissibilité vérifiant si $x \sim f_x$ sous la condition $x \sim g_x$, peut être symbolisé comme suit:

$$\begin{array}{l} \text{Répéter Produire } x \sim g_x(u_1) \\ \text{Jusqu'à } u_2 \le K[\, f(x)/g(x)] \ ; \end{array} \qquad (4.38)$$

K est une constante de normalisation destinée à ramener le quotient f_x/g_x dans l'intervalle (0,1), pour tout x. La première opération, « Produire $x \sim g_x(u_1)$ », peut être réalisée par inversion, en utilisant la f.r. G_x, « $x \leftarrow G^{-1}(u_1)$ », ou autrement.

Telle que formulée, la méthode suppose que la loi g_x domine f_x, soit $K\bullet f_x \le g_x$; elle est avantageuse à la condition que chaque phase soit promptement menée. D'une part, la variable $x \sim g_x$ doit être « facile » à produire, comparativement à la loi cible f_x; on préférera souvent les lois uniforme, ou triangulaire, ou exponentielle, etc., toutes de production

immédiate. D'autre part, plus proche g_x sera de f_x, (i.e. $K \cdot f_x/g_x \rightarrow 1$), plus souvent le test en (4.38) sera positif, contribuant à l'efficacité de la méthode.

La loi *Bêta,* par exemple, est une loi doublement bornée, soit de 0 à 1 dans sa forme standard (cf. exercice 4.29 et §3.12). Aussi, pour produire une v.a. x de loi $f_x \sim \beta$ *(a,b),* la loi $U(0,1)$ s'impose d'emblée comme loi parraine g_x. Pour a > 1 et b > 1, la distribution β (a,b) est unimodale; on peut alors exploiter une forme ou l'autre de la loi triangulaire, symétrique ou non (rappelons que la somme $u'+u''$ se distribue selon une loi triangulaire symétrique, bornée de 0 à 2). Pour des lois non bornées ou semi bornées, telles la loi normale et ses lois apparentées, le rejet pourra encore être appliqué en utilisant une loi parraine non bornée, ou en utilisant une loi bornée appliquée sur un intervalle déterminé de f_x en complément d'autres principes dans le contexte de la méthode globale de *composition.* L'exemple 4.7 en présente un cas d'espèce.

Exemple 4.6 Production d'une v.a. Bêta ($U_{3:9}$) *par rejet d'une v.a. uniforme*

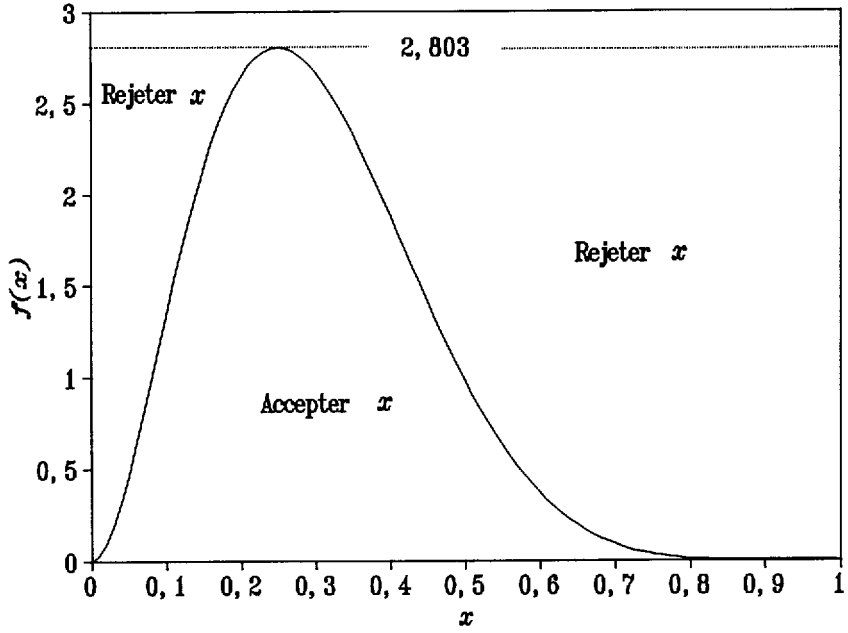
La 3e statistique d'ordre d'un échantillon de 9 variables uniformes, dénotée $u_{3:9}$, a une distribution β (3,7): voir §3.11. Nous avons déjà considéré une méthode de production efficace d'une variable $x = u_{3:9}$ par la fonction récursive (3.25). Nous présentons la version suivante pour sa simplicité.

La figure sur la page suivante montre la distribution de $x \sim \beta\,(3,7) = 252x^2(1-x)^6$. La variable x va de 0 à 1, tout comme u. Son mode, à x = (a -1)/(a+b - 2) = ¼, atteint $f(¼) \approx 2{,}803$, valeur indiquée par la ligne pointillée au haut du graphique.

La fonction parraine $g_x \sim U(0,1)$ peut être utilisée; on a tout simplement $x \leftarrow u_1$. Cette variable x se distribue dans le rectangle correspondant à l'intervalle horizontal (0 ; 1) et la bande verticale (0 ; 2,803); cependant, pour être admissible sous f_x, la variable x doit « apparaître » sous le tracé de f_x, c'est-à-dire que son occurrence est régie par le taux conditionnel $K \cdot f_x/g_x$. Ici, $K = (2{,}803..)^{-1} \approx 0{,}35676$, $g_x \equiv 1$ pour tout x, et le test dans (4.38) devient:

$$u_2 \leq 0{,}35676 \times f(x)\ .$$

Le taux de rejet d'une valeur x candidate dépend du ratio des surfaces $Kf_x : g_x$; en fait, la variable est acceptée selon une probabilité K, et rejetée selon 1 − K. Le nombre de tentatives avant acceptation



est une variable géométrique, d'espérance K^{-1}, ici $0{,}35676^{-1} = 2{,}803$. Somme toute, en comparant avec la fonction (3.25), cette méthode-ci s'avère pratiquement aussi performante.

Les exercices 4.23, 4.28 et 4.33 donnent d'autres exemples commentés de la méthode de rejet.

4.20 *La méthode de composition.* La méthode de composition est basée sur la décomposition de la loi f_x en $r \geq 2$ fonctions élémentaires, soit:

$$f(x) = f_1(x) \oplus f_2(x) \oplus \cdots \oplus f_r(x) , \qquad (4.39)$$

expression dans laquelle le symbole « $\oplus$ » indique une relation quelconque, souvent une simple addition. La composition peut être conçue sur des intervalles disjoints du domaine de x, pour des zones de la surface de f_x recouvrant le même domaine ou, en général, comme l'articulation de deux ou plusieurs méthodes élémentaires exploitées conjointement pour reconstituer f_x. Les techniques de tableaux, pour les variables discrètes (§4.11-4.15), peuvent être vues comme des applications du principe de composition. L'exemple 4.7 illustre ce principe; une v.a. de distribution normale $N(0,1)$ est produite en « composant » deux méthodes de rejet, l'une (rejet d'une v.a. triangulaire) appliquée vers le centre de la distribution,

l'autre (rejet d'une v.a. de Rayleigh) spécialisée pour l'aile d'une distribution normale.

Exemple 4.7 Production d'une v.a. normale N(0,1) *par composition (avec rejet)*

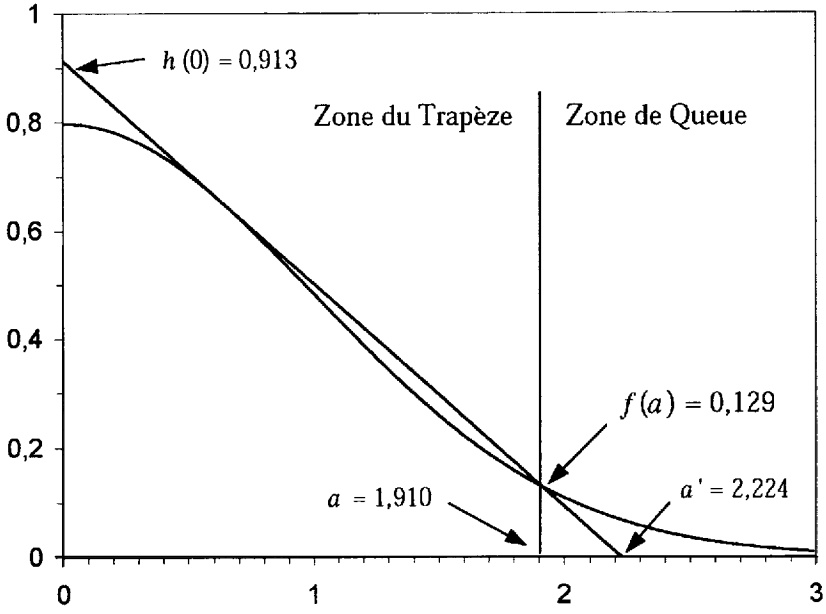
Considérons une variable x à distribution $N(0,1)$; en fait, nous produirons une variable x^+ positive, quitte à lui affecter ensuite un signe aléatoire. La variable x^+ occupe donc le domaine (O, ∞), selon une fonction de densité $f(x^+) = 2xf_{N(0,1)}(x) = \sqrt{(2/\pi)}\cdot\exp(-½x^2)$. La composition consiste ici à diviser le domaine de x^+ en deux parts, (0, a) et (a, ∞). La surface couvrant l'intervalle (0, a) peut être enfermée dans un trapèze; la hauteur de droite (h_a) est $f(a)$, et la hauteur de gauche (h_o) doit être telle que $f(x^+)$, $0 \le x^+ < a$, soit toujours inscrite dans le trapèze, soit:

$$h_0 \ge \frac{f_x - \left(\frac{x}{a}\right) f_a}{1 - \frac{x}{a}} \quad , 0 \le x < a \ .$$

Quant à la surface complémentaire, relative à l'intervalle (a, ∞) dans l'aile (ou queue) de la distribution, elle recevra une solution particulière, plus coûteuse. Ainsi, la v.a. produite x va résulter de la composition suivante:

$$x \leftarrow \text{(signe aléatoire)}[\ x^+_{\text{«Trapèze»}} + x^+_{\text{«Aile»}}\ ].$$

La figure sur la page suivante illustre la solution proposée. Pour être avantageuse, la composition doit favoriser les éléments d'exécution plus rapide, ici la part issue du trapèze. La composante $x_{\text{«Trapèze»}}$ consiste en fait en une variable semi-triangulaire, dont la densité forme un triangle rectangle de côtés proportionnels à a' et h_o; la borne a' s'obtient simplement par $ah_o/[h_o - f(a)]$. L'accroissement de a, qui favorise la composante $x_{\text{«Trapèze»}}$, s'accompagne cependant d'un prix correspondant à la différence entre l'aire du triangle rectangle, soit $A(TR) = ½a'h_o$, et l'aire normale utile, soit $A(N) = \int_o^a f(x)\,dx$; comme, dans cette part, la v.a. x sera produite par *rejet* d'une variable triangulaire, cette différence équivaut au taux de rejet et doit être minimisée. Un critère possible (parmi d'autres) est de maximiser le ratio A(N)/A(TR), ce qu'on obtient avec a = 1,910 (a' = 2,224); la valeur h_o minimale est alors 0,913, et le quotient A(N)/A (TR) ≈ 0,92993 représente le maximum possible. L'aire de la



surface normale positive, de $x^+ = 0$ à 1,910, est de 0,94387. L'algorithme de production par composition peut alors se lire comme suit:

Production d'une v.a. x~ N(0,1) par composition (4.40) *avec rejet d'une v.a. triangulaire*

Obtenir u_1 ;
Si $u_1 < ½$ alors $s \leftarrow -1$ sinon $s \leftarrow +1$, $u_1 \leftarrow u_1 - ½$;
$u_1 \leftarrow 2u_1$;
Si $u_1 \leq 0{,}94387$ alors

[Trapèze] Répéter Obtenir $x \leftarrow 2{,}224\,|u_2+u_3-1|$
Jusqu'à $x \leq 1{,}910$ et $f(x) \geq u_4[0{,}913-0{,}410522x]$
sinon

[Aile] Répéter Obtenir $x \leftarrow (a^2 - 2\log_e u_2)^{½}$ et u_3
Jusqu'à $x \cdot u_3 \leq a$.

La technique de rejet mise en oeuvre pour produire la variable ; x «Aile» est due à Marsaglia (voir Devroye, *op. cit.*, p. 380) (voir aussi les exercices 4.21-4.22).

Statistiques d'ordre

4.21 *Statistiques d'ordre d'une loi f_x donnée.* Pour produire une ou plusieurs statistiques d'ordre (s.o.) $x_{(i:n)}$ d'une variable obéissant à une loi f_x donnée, la solution naïve consiste à produire d'abord les n variables x_i selon f_x, puis à les placer en ordre croissant de telle façon que $x_{(i:n)} \leq x_{(2:n)} \leq$ etc.: les algorithmes de tri fournis en appendice du chapitre 6 sont disponibles à cet effet. Une autre solution est basée sur le fait que, si x est une v.a., $u = F(x)$ l'est aussi, où $F(x)$ est la fonction de répartition de x. Or, on a vu (§3.12) qu'on peut produire directement les s.o. uniformes $u_{(i:n)}$, notamment en commençant par les extrêmes $x_{(i:n)}$ ou $x_{(n:n)}$. *A* partir des s.o. uniformes $u_{(i:n)}$, les s.o. de x découlent ensuite par inversion de la f.r., selon:

$$x_{(i:n)} \leftarrow F^{-1}(u_{(i:n)}). \qquad (4.41)$$

Cette solution peut être très avantageuse si l'on n'a besoin que d'une ou de quelques statistiques d'ordre *et* si la technique d'inversion exploitée est suffisamment efficace (voir §4.16-4.18). Par exemple, pour produire $x_{(20:20)}$ de loi E(1), soit le maximum d'une série de 20 variables exponentielles de paramètre $\lambda = \mu^{-1} = 1$, il suffit d'obtenir d'abord une v.a.u. u, puis d'effectuer «- $\log_e[1 - u^{1/20}] \rightarrow x$ ».

Gerontidis et Smith (1982) présentent enfin une troisième approche, fondée sur la segmentation de l'étendue de x en intervalles de probabilité égale et sur un échantillonnage avec rejet. La méthode, de construction plus difficile, a une efficacité asymptotique généralement meilleure que les méthodes concurrentes. Devroye (1986) recense quelques autres techniques, certaines spécifiques à une loi f_x donnée, pour générer des s.o.

Récapitulation

4.22 Nous avons passé en revue les principales techniques qu'on peut trouver afin de produire des v.a. d'une sorte ou d'une autre. Qu'il s'agisse d'une technique *ad hoc* et d'une véritable méthode appliquant un principe général, chacune peut servir dans plus d'un contexte; le cloisonnement que notre exposé nous a imposé ne reflète pas une limitation correspondante des techniques. Ainsi, la technique du tableau bivoque, conçue pour une v.a. discrète bornée à n états, est en fait une composition de n v.a. binaires, *i.e.* la valeur x et son alias $A(x)$. L'interpolation inverse exploite un tableau. Le principe du rejet convient aussi aux v.a. discrètes (exerice 4.33). Ainsi de suite.

4.23 Malgré que l'intérêt des chercheurs pour les techniques de production de variables aléatoires se soit déployé depuis peu, nombre de techniques et d'éléments théoriques ont vu le jour. L'ouvrage de Devroye (1986; aussi Fishman 1996 et Gentle 1998) mentionne encore les techniques de transformation trigonométrique (« polaire »), la méthode du quotient de variables uniformes, les méthodes de série ou de polynôme, etc. Les exercices en fin de chapitre donnent quelques exemples de ces autres méthodes.

Après l'exposé des méthodes et techniques, Devroye (1986) passe en revue un grand nombre de distributions statistiques et, pour chacune, il dresse un bilan des techniques applicables et de leurs efficacités, fournissant en même temps toutes références utiles.

Exercices

4.1 Soit des variables discrètes de loi rectangulaire (4.5), $x \sim R(1,n)$. Montrer que, dans un échantillonnage de k valeurs *avec remise*, le nombre k' de valeurs distinctes a pour espérance $n[1-(1-1/n)^k]$. Pour $k \ll n$ tels que $k^2/(2n) \to 0$, montrer que $k' \approx k$, signifiant que l'échantillonnage avec remise équivaut en pratique à celui sans remise (voir aussi exercice 8.21).

4.2 Montrer que, pour une variable binomiale $x \sim B(n,\pi)$ dont les probabilités sont définies en (4.7), on a $\mu = E(x) = n\pi$, $\text{var}(x) = \sigma^2 = n\pi(1-\pi)$, $\gamma_1 = \alpha_3 = (1-2\pi)/\sigma$ et $\gamma_2 = \alpha_4 - 3 = [1-6\pi(1-\pi)]/\sigma^2$. [Suggestion: Les *moments factoriels* $f_r(X)$ de la v.a. discrète X sont définis comme $E\{ X(X-1)\cdots(X-r+1) \}$ et peuvent être obtenus par $g(s) = E(s^X) = \sum_x p_x s^x$, selon $f_r(X) = g^{(r)}(1)$. Définissant les moments à l'origine par $\mu'_r(X) = E(X^r)$, nous avons $f_1 = \mu'_1$, $f_2 = \mu'_2 - \mu'_1$, soit $\mu = \mu'_1 = f_1$, $\mu'_2 = f_2 + f_1$ et $\sigma^2 = f_2 + f_1 - f_1^2$, etc. Pour la loi (4.7), $g(s) = E(s^X) = \sum \binom{n}{x} \pi^x (1-\pi)^{n-x} s^x = (\pi s + 1 - \pi)^n$. Les moments f_r suivent.]

4.3 Prouver la validité de la transformation (4.10). [Suggestion: La fr. de la v.a. géométrique indiquée par (4.9) est $F(x) = \Sigma^x_{n=1} (1-\pi)^{n-1} \pi = 1-(1-\pi)^x$ et peut être inversée.]

4.4 Montrer que, pour une variable géométrique répondant à (4.9) et de paramètre π, on a $E(x) = 1/\pi$, $0^2 = (1-\pi)/\pi^2$, $\gamma_1 = (2-\pi)/\sqrt{(1-\pi)}$, $\gamma_2 = 6 + \pi^2/(1-\pi)$. [Suggestion: En conditionnant sur le premier événement, on peut écrire $E(x^r) = E[\, 1^r \times \pi + (1+x)^r \times (1-\pi)\,]$; développant et simplifiant ces expressions, on obtient les moments à l'origine $\mu'_r(x) = E(x^r)$, dont les autres moments et indices peuvent être tirés.]

4.5 Prouver la validité de l'algorithme *(4.12)* [cf. Devroye 1986, *p. 504]*.

4.6 Montrer que, pour une variable de Poisson répondant à (4.11) et de paramètre μ ($= \lambda t$), on a $E(x) = 0^2 = \mu$, $\gamma_1 = 1/\sqrt{\mu}$, $\gamma_2 = 1/\mu$. [Suggestion: Recourant encore aux moments factoriels, définis à l'exercice 4.2, nous avons, pour la loi (4.11), $g(s) = E(s^x) = \Sigma e^{-\mu x}/x! s^x = e^{\mu(s-1)}$, d'où $f_r = g^{(r)}(1) = \mu^r$, et le reste suit.]

4.7 La variable $x = (\sum_{i=1}^{12} u_i - 6)$, basée sur 12 v.a. u, est standardisée à moyenne 0 et variance 1, et sa distribution est proche de la loi $N(0,1)$. D. Teichroew (cf. Knuth 1969) améliore l'approximation en appliquant la transformation polynomiale:

$$z \leftarrow (((((a y^2 + b)\cdot y^2 + c)\cdot y^2 + d)\cdot y^2 + e)\cdot y ,$$

où $y = x/4$, a = 0,029899776, b = 0,008355969, c = 0,076542912, d = 0,252408784, e = 3,949846138. Établir les moments (μ, σ^2, γ_1, γ_2) de z. Montrer que, même si les variables x et z sont bornées (à ±6 et ±8,65 resp.), la surface normale « oubliée » est négligeable, soit $\mathrm{pr}_{N(0,1)}(|x| \geq 6{,}0) \approx 2{,}0\times10^{-9}$ et $\mathrm{pr}_{N(0,1)}(|x| \geq 8{,}65) \approx 5{,}1\times10^{-18}$.

4.8 Reprendre l'exemple 4.2 en appliquant la technique du tableau simple à probabilités décroissantes. Montrer que le nombre moyen de valeurs à examiner est alors de 2,672.

4.9 La *loi de Pascal*, une généralisation de la loi géométrique (§4.5) pour un processus de Bernoulli, indique la probabilité d'obtenir pour la première fois k succès au n^{e} essai; sa fonction de masse est $Pa(n;k,\pi) = \binom{n-1}{k-1}\pi^k(1-\pi)^{n-k}$, $n \geq k$. Concevoir un algorithme basé sur une technique de tableau hybride pour produire une variable de cette loi, en arrêtant le tableau à la valeur $n = n_{\mathrm{T}}$ et en le complétant par une fouille pour $n > n_{\mathrm{T}}$ [noter que $Pa(n+1) = Pa(n)\times(1-\pi)n/(n+1-k)$].

4.10 La *loi des succès consécutifs [W.* Feller, *An introduction to probabilitytheory and its applications, 3*e édition, New York, Wiley 1968 et L. Laurencelle, « La loi des succès consécutifs dans un processus de Bernoulli », *Lettres statistiques,* 1987, *vol. 8,* 23-47], une généralisation de la *loi de Pascal,* décrit la probabilité d'observer pour la première fois k succès consécutifs au n^{e} essai. Chaque succès ayant une probabilité individuelle égale à *Tt,* la f.r. peut être calculée récursivement selon P(n+1) = *P(n)* + [(1-P(n-k)](1- л) л ^{k}, en posant $P_1 = P_2 = \cdots = P_{k-1} = 0$, $P(k) = л^k$ et $n > k$. Concevoir un algorithme basé sur une technique de tableau hybride pour produire une variable de cette loi, *i.e.* en arrêtant le tableau à la valeur $n = n_T$ et en le complétant par une fouille pour $n > n_{\mathrm{T}}$.

4.11 *Loi exponentielle,E(ς).* Une variable $x > 0$, obéissant à la densité $f(x) = \varsigma e^{-\varsigma x}$, est dite de distribution exponentielle. (a) Montrer que l'espérance et la variance sont $?\varsigma^{-1}$ et ς^{-2} respectivement. (b) Soit u ~ U(0,1), une v.a. uniforme: prouver la validité de la relation « x←-$\varsigma^{-1}\log_e(u)$ ». (c) Vérifier qu'une v.a. de loi x^2_2, distribuée selon (4.17), est aussi une v.a. de loi E(½).

4.12 *Inversion de la loi normale par approximation rationnelle.* La fonction $g(u) = F^{-1}(u)$ peut être développée comme un polynôme en t, ou un quotient de polynômes en t, *où* $t = h(u)$. Cette technique a été abondamment étudiée pour l'inversion de la f.r. normale. Soit u ≥½et $t = h(u) = \{-½ л \log e[u(1-u)]\}^{½}$, alors :

$$\hat{x} \leftarrow t\cdot(1+0{,}0078365t^2-0{,}0002881t^4+0{,}0000043728t^6)$$

se distribue approximativement comme une normale positive (cf. B. J. R. Bailey, « Alternatives to Hasting's approximation to the inverse of the normal cumulative distribution », *Applied statistics*, 1981, vol. 30, p. 275-276). Montrer que l'erreur absolue $|\hat{x}(u) - x(u)| \leq 0{,}00029$ pour $|x| \leq 4{,}98$. Odeh et Evans (1974) présentent un quotient de polynômes du 4e degré, dont le résultat $\hat{x}$ a 7 chiffres décimaux de précision.

4.13 Élaborer un algorithme général de production d'une v.a. discrète bornée utilisant la technique du tableau simple et le repérage par bissection. Dans le cas de distributions asymétriques, comparer en efficacité la bissection stricte (amorcée au point-milieu du domaine de *x)* et la bissection modale (amorcée au mode de *x).*

4.14 Soit une v.a. discrète x = [0; 1; 2; 3] et sa distribution de probabilités p = [0,6; 0,3; 0,08; 0,02]. Trouver les trois réalisations différentes des vecteurs q et A dans la technique du tableau bivoque. Prouver l'équivalence résultante des trois réalisations.

4.15 L'algorithme (4.30) repère chaque quantité *q[i]* > 1 et l'épuise en créant des alias et en déversant l'excédent 1-q[i] dans les *q[j]* fractionnaires, rencontrés au hasard. Ce déversement au hasard, qui peut être occasionnellement répété pour le même q[i], constitue une source d'inefficacité possible pour l'algorithme. Concevoir un algorithme plus efficace, qui réduit ou tend à réduire cette phase de l'exécution (voir aussi Devroye 1986, p. 109). Comparer les coûts d'exécution des deux algorithmes.

4.16 Peut-on accélérer l'algorithme (4.29), p. ex. en utilisant une seule v.a. uniforme?

4.17 Définir les vecteurs de seuils q[0..8] et d'alias A[0..8] pour la variable binomiale B(8,⅓) étudiée à l'exemple 2, en employant l'algorithme (4.30).

4.18 La fonction Gamma, dénotée $\Gamma(x)$, est définie par $\int_0^\infty t^{x-1}e^{-x}\,dt$. Par intégration par parties, montrer que $\Gamma(x) = (x-1)\Gamma(x-1)$ et que, pour x entier, $\Gamma(x+1) = x!$. De plus, pour x entier, $\Gamma(x+\frac{1}{2}) = (x-\frac{1}{2})(x-\frac{3}{2}) \cdots (\frac{1}{2})\sqrt{\pi}$.

4.19 Par quel procédé doit-on exploiter k+1 v.a.u. pour produire une variable de loi t_{2k} ?

4.20 L'interpolation non linéaire inverse permet d'optimiser le ratio (précision) / (taille de la table de référence), en plaçant judicieusement quelques valeurs de référence qui garantissent une précision donnée.

Soit $y = f(x) = Ax^2 + Bx + C$, une fonction parabolique dont on connaît trois applications, P_1, P_2 et P_3. En utilisant P_2 comme point central (i.e. $x_1 < x_2 < x_3$) et en établissant $\Delta = (x_2 - x_1)$ et $h = (x_3 - x_2) / (x_2 - x_1)$, on obtient:

$$A = \alpha/\Delta^2 \; ; \; B = \beta/\Delta - 2Ax_2 \; ; \; C = \gamma - Ax_2^2 - Bx_2 \; ,$$

où $a \leftarrow [y_3 - y_2 - h(y_2 - y_1)] / [h(h+1)]$, $\beta \leftarrow [y_3 - y_2 + h^2(y_2 - y_1)] / [h(h+1)]$, $y \leftarrow y_2$. Montrer l'équivalence de ces formules et de celles du paragraphe 4.18 lorsque $h = 1$. Construire deux tables d'interpolation pour l'intégrale normale, l'une linéaire à intervalles réguliers et l'autre parabolique à intervalles irréguliers, les deux observant une précision de 1/1000; comparer la taille des tables et les algorithmes d'interpolation.

4.21 La technique appliquée pour produire une v.a. située dans l'aile d'une normale, dans l'exemple 4.7, consiste sommairement à générer une v.a. de Rayleigh[4], puis à en tester l'admissibilité. La probabilité d'admettre la v.a. produite est $a \cdot e^{a^2/2}(1 - \Phi(a))\sqrt{(2\pi)}$, où $\Phi(x)$ est l'intégrale (ou f.r.) de la normale $N(0,1)$. En moyenne, combien d'itérations cette phase de l'algorithme devra-t-elle accomplir pour admettre une valeur?

4.22 L'exemple 4.7 utilise des paramètres (a et h_o) fixés afin de maximiser le quotient des aires A(N)/A(TR). Un autre critère consisterait à maximiser la fonction $A(N)^2/A(TR)$, dans le but de favoriser le choix de la composante « trapèze » de la solution; les paramètres appropriés sont alors a = 2,11 et h_o = 0,897; l'intégrale normale positive à ce point est 0,96514, et les quotients A(N)/A(TR) ≈ 0,92157 et $A(N)^2/A(TR)$ ≈ 0,88945, ce dernier étant le maximum global. Un troisième critère consiste à tenir compte des coûts d'exécution respectifs des composantes « trapèze » et « queue », le but étant de minimiser C$ = ξ C$(« trapèze ») + (1- ξ)C$(« queue »); ce critère est difficile d'évaluation puisqu'il est récursif, les deux coûts élémentaires dépendant de a, lequel détermine à son tour ξ. Étudier

4. Cette loi de Rayleigh (cf. Evans, Hastings et Peacock 2000) a pour densité $(x/b^2) \cdot \exp[-x^2/(2b^2)]$ et, pour f.r., $1-\exp[-x^2/(2b^2)]$, $x > 0$. Les premiers moments sont $E(x) = b\sqrt{(\frac{1}{2}\pi)}$ et $\text{var}(x) = (2 - \pi/2)b^2$; le mode est b, la médiane 1,17741 b. La f.r. s'inverse aisément.

et comparer les solutions optimales de ces trois critères, en exploitant les données de l'exercice précédent.

4.23 La loi normale positive $N^+(0,1)$, qui correspond à la moitié droite de la loi normale standard $N(0,1)$, a pour densité $f(x) = \sqrt{(2/\pi)} \cdot e^{-x^2/2}$ (voir exemple 4.7). Cette loi s'étale semblablement à la loi exponentielle E(1), de densité $g(x) = e^{-x}$ (cf. exercice 4.11). Pour $x \geq 0$, montrer que la valeur maximale K pour laquelle $Kf(x) \leq g(x)$ est $\sqrt{[\pi/(2e)]} \approx 0{,}760173$. Elaborer un algorithme de production d'une v.a. $N(0,1)$ basée sur la génération d'une v.a. E(1) avec rejet et l'ajout d'un signe positif ou négatif, au hasard. Quel est le taux de rejet des v.a. exponentielles générées? L'algorithme résultant est-il aussi efficace que ceux indiqués par (4.15) et (4.16) ?

4.24 *Production efficace d'une v.a.* x^2_ν . R. C. H. Cheng et G. M. Feast (« Some simple Gamma variate generators », *Applied Statistics,* 1979, vol. 28, p. 290-295; voir aussi Fishman 1996) donnent une méthode pour produire des v.a. x^2 ayant ν degrés de liberté: une v.a. de loi *Gamma* est générée par un quotient de v.a.u., puis on procède par rejet. La méthode, assez efficace pour tout ν, utilise les constantes $A = 1+\sqrt{(2/e)} \approx 1{,}857764$, $B = \sqrt{(\nu/2)}$, $C = (3\nu^2-2)/[3\nu(\nu-2)$, $D = 4/(\nu - 2)$ et $E = \nu - 2$. L'algorithme récurrent suit:

Répéter (4.42)
  Répéter Obtenir u_1 et $u_2 \sim U(0,1)$;
    $z \leftarrow u_1 + (1 - A \times u_2)/B$;
  Jusqu'à $0 < z \leq 1$;
  $w \leftarrow C \times u_1 / z$;
Jusqu'à $[D\times(z-1)+w+1/w < 2]$ ou $[D\times\log_e(z)-\log_e(w)+w < 1]$
$x \leftarrow E \times w$.

À partir de quelle valeur paire ν' la méthode de Cheng et Feast devient-elle préférable à celle indiquée par (4.20)? Quel nombre moyen de v.a. u_i différentes la méthode consomme-t-elle pour cette valeur ν'?

4.25 Une variable x issue de la distribution t de Student avec ν degrés de liberté a pour loi de probabilité [Patel, Kapadia et Owen]:

$$\mathrm{pr}(x) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\nu\pi}} \left(1 + x^2/\nu\right)^{-(\nu+1)/2} \qquad (4.43)$$

(voir aussi Johnson, Kotz et Balakrishnan 1994, 1995 pour différentes méthodes exactes ou approximatives de calcul de la f.r.). Ses moments sont $\mu = 0$, $0^2 = v/(v\text{-}2)$, $y_1 = 0$ et $y_2 = 6/(v\text{-}4)$.

Devroye *(op. cit.)* présente un algorithme simple, basé sur un quotient de deux v.a. uniformes et produisant une v.a. $x \sim t_3$ (i.e. t avec 3 degrés de liberté), soit:

$$\begin{aligned} &\text{Répéter Obtenir } u \sim U(0,1) \text{ et } v \sim U(-\tfrac{1}{2},+\tfrac{1}{2}) \\ &\text{Jusqu'à } u^2 + v^2 \le u \ ; \\ &x \leftarrow \sqrt{3}v/u \ . \end{aligned} \tag{4.44}$$

En moyenne, combien de fois le cycle sera-t-il activé pour satisfaire le critère ?

4.26 (Suite du précédent) Devroye *(op. cit.)* rapporte un algorithme (attribué à D.J. Best) qui produit une variable $x \sim t_v$ par rejet d'une variable t_3 En voici la teneur :

$$\begin{aligned} &\text{Répéter} \quad \text{Obtenir } x \sim t_3 \text{ et } u \sim U(0,1) \ ; \\ &\qquad z \leftarrow x^2 \ ; \ w \leftarrow 1 + z/3 \ ; \ y \leftarrow 2\log_e[\ 9w^2/16u\ ] \ ; \\ &\qquad \text{Ok} \leftarrow (y \ge 1 - z) \ ; \\ &\qquad \text{Si (non Ok) alors Ok} \leftarrow \left[\ y \ge (v+1)\log_e\left(\frac{v+1}{v+z}\right)\right] \\ &\text{Jusqu'à Ok} \ . \end{aligned} \tag{4.45}$$

Comparer (analytiquement ou empiriquement) l'efficacité de cet algorithme et celui indiqué par (4.21) pour t_5, t_{10} et t_{25}.

4.27 *Production efficace d'une v.a. t_v. R. W.* Bailey (« Polar generation of random variates with the t-distribution », *Mathematics of computation,* 1994, vol. 62, p. 779-781), cité par Gentle (1998), conçoit une variante de la méthode « polaire » de Box et Muller afin de générer des v.a. obéissant à la loi t_v de Student. L'algorithme suivant met en oeuvre la technique.

$$\begin{aligned} &\{\ \text{Utilise C} = -2/v\ \} \\ &\text{Répéter} \quad \text{Obtenir } u_1, u_2 \sim U(0,1) \ ; \\ &\qquad t \leftarrow 2u_1 - 1 \ ; \ v \leftarrow 2u_2 - 1 \ ; \ r \leftarrow t^2 + v^2 \\ &\text{Jusqu'à} \quad r < 1 \ ; \\ &x \leftarrow t \times \sqrt{[v \times (r^C - 1)/r]} \ . \end{aligned} \tag{4.46}$$

Montrer que l'algorithme produit des v.a. ayant les moments appropriés (voir exercice 4.25).

4.28 L'exemple 4.6, relatif à la production d'une v.a. β (3,7) par rejet d'une uniforme, peut être reconçu en utilisant comme variable parraine une v.a. triangulaire asymétrique, c'est-à-dire en enfermant $f_X = 252x^2(1-x)^6$ dans un triangle scalène. Le triangle optimal a un sommet de coordonnées x = 0,211 et y = 3,02; l'intégrale de f_x *à* c = 0,211 est 0,29124. Elaborer un algorithme utilisant cette méthode et la variable semi-triangulaire $y \leftarrow 3{,}02\ u_1 + u_2 - 1\ |$. Montrer que le taux global de rejet est d'environ 0,312 et le taux de répétition d'environ 1,51. Comparer cette solution à celle donnée à l'exemple 4.6.

4.29 Soit *x (3(a,b),* une v.a. *Bêta* à paramètres *a, b* positifs quelconques (même fractionnaires) telle que celle gouvernant la variation des s.o.u. (§3.12). Dans son expression standard, la densité est:

$$\mathrm{pr}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1}\ ,\ 0 \le x \le 1\ . \qquad (4.47)$$

Par intégration par parties, prouver que les moments à l'origine (μ'_r) de x sont $\Gamma(a+r)\Gamma(a+b) / \Gamma(a)\Gamma(a+b+r)$ et que, en conséquence, les moments caractéristiques sont:

$$\mu = a / (a+b)\ ; \qquad (4.48a)$$
$$\sigma^2 = ab / [(a+b+1)(a+b)^2]\ ; \qquad (4.48b)$$
$$\gamma_1 = 2(b-a)\sqrt{(a+b+1)} / [(a+b+2)\sqrt{(ab)}]; \qquad (4.48c)$$
$$\gamma_2 = 6[(a+b+1)(a^2+b^2-3ab)-ab] / [ab(a+b+2)(a+b+3)]\ . \qquad (4.48d)$$

4.30 L'algorithme suivant (cf. Knuth 1969; Bratley, Fox et Shrage 1987):

$$\begin{array}{ll} \text{Répéter} & \text{Obtenir } u_1, u_2 \sim U(0,1)\ ; \\ & y_1 \leftarrow u_1^{1/a}\ ;\ y_2 \leftarrow u_2^{1/b} \\ \text{Jusqu'à} & y_1 + y_2 \le 1\ ; \\ x \leftarrow y_1 / (y_1 + y_2) & \end{array} \qquad (4.49)$$

fabrique des v.a. x de distribution β (a, *b).* Comparer empiriquement l'efficacité de cette méthode par rapport à celle présentée à l'exercice 4.28, pour *a = 3 et b = 7.* L'efficacité de cette méthode-ci s'améliore-t-elle ou se détériore-t-elle quand *a* et *b* augmentent ?

4.31 *Loi du F (de Fisher-Snedecor).* La variable x qui obéit à la loi du F(v,,v,) a pour densité:

$$\mathrm{pr}(x) = \frac{(\nu_1/\nu_2)^{\nu_1/2}\,\Gamma([\nu_1+\nu_2]/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\, x^{\nu_1/2-1}\left(1+\frac{\nu_1}{\nu_2}x\right)^{-½(\nu_1+\nu_2)}\ . \qquad (4.50)$$

Montrer que ses premiers moments sont $E(x) = \nu_2/(\nu_2 - 2)$ et $\text{var}(x) = 2\nu_2^2(\nu_1+\nu_2-2) / [\nu_1(\nu_2-2)^2(\nu_2-4)]$ (voir aussi Johnson, Kotz et Balakrishnan 1994, 1995 ou Laurencelle et Dupuis 2000).

4.32 *Production efficace d'une v.a. $F(\nu_1,\nu_2)$.* L'efficacité à produire des v.a. de la loi F, dite de Fisher-Snedecor, dépend généralement des paramètres ν_1 et ν_2. R.C.H. Cheng (« Generating Beta variates with nonintegral shape parameters », *Communications of the ACM,* 1978, vol. 21, p. 317-322; voir aussi Fishman 1996) propose une méthode d'efficacité raisonnable pour produire une v.a. $x \sim \beta(a,b)$, qu'on transforme en une v.a. $F(\nu_1,\nu_2)$ grâce à (3.29a) (voir exercice 3.12). La méthode procède par réduction analytique (§ 10.12) et rejet. L'algorithme correspondant suit.

[Préparation] $a \leftarrow \nu_1/2$; $b \leftarrow \nu_2/2$; $s = a + b$; (4.51)
Si $\min(a,b) \leq 1$ alors $\beta \leftarrow 1/\min(a,b)$
sinon $\beta \leftarrow \sqrt{[(s-2)/(2ab-s)]}$;
$g \leftarrow a + 1/\beta$; $C = \log_e 4 \approx 1{,}3862943611$.

[Production] Répéter Obtenir $u_1, u_2 \sim U(0,1)$;
$v \leftarrow \beta \cdot \log_e[u_1/(1-u_1)]$; $w \leftarrow a \cdot e^v$
Jusqu'à $s \cdot \log_e[s/(b+w)] + g \cdot v - C \geq \log_e(u_1^2 \cdot u_2)$;
$x \leftarrow w / a$. [*]

Pour produire la v.a. de loi $\beta(a, b)$, ou $\beta(\frac{1}{2}\nu_1,\frac{1}{2}\nu_2)$, correspondante, il suffit de remplacer la ligne marquée «*» par « $x \leftarrow w/(b+w)$ ». Comparer la performance du présent algorithme par rapport à un algorithme produisant des v.a. $F(\nu_1,\nu_2)$ par le quotient de deux v.a. x^2 tel que défini en (4.24) et utilisant la méthode de Cheng et Feast (exercice 4.24).

4.33 *Production efficace d'une v.a. de Poisson par rejet d'une v.a. géométrique.* La loi de Poisson *Po(μ)* (§4.6) et la loi géométrique $G(\pi)$ (§4.5) sont deux lois discrètes, la première pour x = 0, 1, 2, etc., et la seconde, pour x = 1, 2, etc. Les deux décroissent régulièrement pour x croissant, à partir d'un mode respectif, de sorte que la loi $G(\pi)$, facile à générer par (4.10), peut servir de parraine à l'autre. L'algorithme suivant génère des v.a. $x \sim Po(\mu)$ [noter que [y] dénote le plus petit entier n tel que $n \geq y$].

[Préparation] $\hat{x} \leftarrow \lceil \mu \rceil$; (4.52)
$K \leftarrow e^\mu \hat{x}! / (\mu+1)^{\hat{x}+1}$; $K' = K/e^\mu$;
$A \leftarrow \log_e[\mu/(\mu+1)]$.

[Production] Répéter Obtenir u_1, $u_2 \sim U(0,1)$;
$x \leftarrow \lfloor \log_e(u_1) / A \rfloor$
Jusqu'à $u_2 \leq K'(\mu+1)^{x+1}/x!$

Prouver la validité de l'algorithme ci-dessus. Quel est le taux de rejet de cet algorithme et comment son efficacité se compare-t-elle à celle de l'algorithme (4.12) ? Noter que, à la fonction factorielle indiquée par « $x!$ », sur la dernière ligne, on peut substituer un tableau de valeurs précalculées « fa[x] », d'une taille à déterminer. [Suggestion: En égalisant les espérances, $E(x_{Po})$ et $E(x_G - 1)$, on obtient le paramètre $\pi = 1/(\mu+1)$, permettant de produire x_G et le candidat $x_{Po} = x_G - 1$. L'écart maximal entre $f_{Po}(x)$ et $g_G(x-1)$ se produit au-dessus de mode de x_{Po}, soit à $\hat{x} = \lceil \mu \rceil$, la valeur K ci-dessus en étant le résultat.]

4.34 Soit une v.a. $y \sim \chi^2_\nu$ définie par la relation (4.18), $y = z_1^2 + z_2^2 + ... + z_\nu^2$, $z_i \sim N(0,1)$. Grâce à cette relation, montrer que:

$$\mu(\chi^2_\nu) = E\{\sum_{i=1}^{\nu} z_i^2\} = \nu \qquad (4.53a)$$

et:
$$\sigma^2 = 2\nu \; . \qquad (4.53b)$$

En utilisant la densité (4.17) et par intégration par parties ou par la relation avec la loi *Gamma* et ses moments (exercice 5.9), prouver que:

$$\gamma_1(\chi^2_\nu) = \sqrt{(8/\nu)} \qquad (4.53c)$$

$$\gamma_2(\chi^2_\nu) = 12/\nu \; . \qquad (4.53d)$$

Références

BRATLEY, P., Fox, B.L., SCHRAGE, L.E. (1987). *A guide to simulation* (2e édition). New York, Springer-Verlag.

DEVROYE, L. (1986). *Non-uniform random variate generation.* New York, Springer-Verlag.

EVANS, M., HASTINGS, N., PEACOCK, B. (2000). *Statistical distributions* (3e édition). New York, Wiley.

FISHMAN, G.S. (1996). *Monte Carlo : concepts, algorithms, and applications.* New York, Springer.

GENTLE, J.E. (1998). *Random number generation and Monte Carlo methods.* New York, Springer.

GERALD, C.F., WHEATLEY, P.O. (1984). *Applied numerical analysis* (3e édition). Reading (MA), Addison-Wesley.

GERONTIDIS, I., SMITH, R.L. (1982). Monte Carlo generation of order statistics from general distributions. *Applied statistics,* 31, 238-243.

JOHNSON, N.L., KoTz, S. BALAKRISHNAN, N. (1994, 1995). *Continuous univariate distributions,* Vols. 1 et 2 (2e édition). New York, Wiley.

JOHNSON, N.L., KoTz, S., KEMP, A.W. (1992). *Univariate discrete distributions* (2e édition). New York, Wiley.

KNUTH, D.E. (1969). *The art of computer programming.* Vol. 2: *Seminumerical algorithms.* Reading (MA), Addison-Wesley.

LAURENCELLE, L., DUPUIS, F.A. (2000). *Tables statistiques expliquées et appliquées (*2e édition). Sainte-Foy, Le Griffon d'argile.

ODEH, R.E., EVANS, J.O. (1974). The percentage points of the normal distribution. *Applied statistics,* 23, 96-97.

PATEL, J.K., KAPADIA, C.H., OwEN, D.B. (1976). *Handbook of statistical distributions.*New York, Marcel Dekker.

Chapitre

5 Production de variables aléatoires corrélées

5.1 Pour l'étude Monte Carlo de certains modèles ou dans diverses situations, il arrive que le chercheur ait besoin de deux ou de plusieurs variables aléatoires qui entretiennent entre elles des liens de corrélation. Rappelons que la corrélation (linéaire) entre les variables X et Y exprime la concordance de leurs variations respectives. Le *coefficientde corrélation* p (« rho »),

$$\rho(X, Y) = \frac{E(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y} , \qquad (5.1)$$

est une mesure du degré de covariation proportionnelle des deux variables. Les méthodes du chapitre précédent nous ont mis à même de produire des v.a. indépendantes x_i obéissant à une loi de distribution $f(x)$: il s'agit ici de produire une v.a. double *(x, y),* ou une v.a. triple *(x, y, z),* etc., de façon telle que la corrélation entre les v.a. individuelles est spécifiée, par exemple $P(x,y) = P$.

> Il existe d'autres formes de *dépendance* entre variables aléatoires que celle exprimée par la corrélation linéaire. En fait, pour des variables X et Y, toute loi conjointe $f(X,Y)$ satisfaisant généralement:
>
> $$f(X,Y) \neq f(X)f(Y) \qquad (5.2)$$
>
> engendre des v.a. mutuellement associées, *dépendantes,* la dépendance ressortissant à une corrélation p non nulle, ou dépendance linéaire, n'est qu'un cas particulier. Certaines formes de dépendance, autres qu'une dépendance corrélationnelle, peuvent apparaître dans le cas de v.a., la fonction conjointe $f(X,Y)$ étant alors spécifiée dans un tableau (voir §5.8). Nous privilégions ici la *dépendance linéaire (ou* corrélationnelle), telle qu'elle s'exprime par le coefficient de corrélation p.

Comme au chapitre précédent, où il s'agissait de produire une ou plusieurs séries de v.a. d'une loi spécifiée $f(X)$, il nous faut ici produire une

ou plusieurs séries de v.a. conjointes (X, Y) telles que X~*f*(X), Y~ *f* (Y) et $p(X, Y) \neq 0$, la corrélation p reflétant la dépendance linéaire entre les séries. Dans la plupart des cas, les lois univariées f et f seront identiques. L'estimateur r du coefficient de corrélation p, basé sur la série statistique elle-même, se calcule par:

$$r(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \; ; \qquad (5.3)$$

c'est le coefficient de corrélation de Galton, aussi attribué à K. Pearson. Notons que le coefficient r est un estimateur biaisé de p; pour des variables (X, Y) normales, Olkin et Pratt (1958) montrent que, approximativement:

$$E(r) \approx \rho[1 - (1-\rho^2)/(2n)] \; , \qquad (5.4)$$

de sorte que le coefficient calculé sous-estime légèrement le paramètre en espérance[1]. Laurencelle (2000) scrute la difficile question de l'estimation ponctuelle de p. Les séries à produire, selon les méthodes présentées plus bas, sont régies par le paramètre p. Ainsi, malgré que la corrélation (r) mesurée pour une série tende vers p (i.e. $r_n \rightarrow p$ pour $n \rightarrow \infty$) selon tous les critères, rien n'assure une égalité même approximative entre r et p pour une série donnée.

5.2 Étant un problème particulier d'application limitée, la production de v.a. corrélées n'a pas gagné autant d'attention des auteurs, et il n'existe que relativement peu de techniques publiées sur ce sujet. La technique *d'inversion* de la f.r. (§4.2), qui inspirait tant d'autres techniques de production de v.a. indépendantes, n'a pas cours ici: ou bien la f.r. bivariée n'a pas d'expression connue, ou bien, si c'est le cas (comme pour la loi normale bivariée ou multivariée), celle-ci est multivoque ou très difficile à inverser. La méthode de rejet (§4.19), qui s'appliquerait peut-être ici, se heurte encore au fait de l'absence de f.r. bivariée, etc.

Tout n'est pas si noir pourtant, et il existe plusieurs cas particuliers intéressants et très simples, basés sur l'addition de v.a. indépendantes; nous en examinons quelques-uns. Une approche générale pour produire des v.a. corrélées (X, Y), de *lois f(X)* et *f*(Y), consiste à produire, d'une manière ou d'une autre, des v.a.u. corrélées (u_x, u_y), toutes deux de loi $U(0,1)$, puis à les inverser selon $X \leftarrow F_x^{-1}(u_x)$ et $Y \leftarrow F_y^{-1}(u_y)$: ce sera l'objet de la

1. Olkin et Pratt (1985) proposent le coefficient corrigé $r^* = r + (1 - r^2)/[2(n-4)]$, tel que $E(r^*) \approx p$.

section suivante. L'utilisation d'un *mélange* de deux distributions fournit aussi des v.a. corrélées d'un type particulier. Nous considérons enfin la production de v.a. corrélées discrètes.

5.3 *Variables uniformes corrélées.* Laurencelle (1993) présente une méthode inspirée de Châtillon (1984), dont le but est de produire des v.a. corrélées. La méthode consiste à construire une ellipse orientée à 45°, ayant un rapport d'axes a/A tel que $a/A = [\,1 - \rho)/(1+\rho)]^{1/2}$, et à y parsemer des points *uniformément:* les coordonnées (X, Y) de ces points ont alors une corrélation proche de p. Cependant, même si la densité bivariée $f(X,Y)$ est uniforme, les densités univariées $f(X)$ ou $f(Y)$ ne le sont pas, sauf pour $\rho=0$ ou $\rho=\pm1$.

La production de v.a.u. corrélées, dont les densités restent strictement uniformes pour toute valeur p, reste un objectif prioritaire puisque, cela fait, il y a moyen de fabriquer des v.a. de lois quelconques f_X et f_Y par l'une ou l'autre technique d'inversion univariée. Devroye (1986) présente quelques méthodes (voir aussi Laurencelle 1993). L'une d'elles, due à K. V. Mardia, procède comme suit afin de produire des paires (X, Y) ~ $U(0,1,0,1,\rho)$:

Soit $u_1, u_2 \sim U(0,1)$, deux v.a.u. indépendantes:

$$\begin{aligned} z &\leftarrow u_1(1-u_1)\ ; \\ x &\leftarrow u_2\ ; \\ y &\leftarrow \{\ 2z(a^2x+1-x) + a(1-2z) \\ &\qquad -(1-2u_1)\sqrt{[a(a+4zx(1-x)(1-a)^2)]}\ \}\ /\ [2a+2z(1-a)^2]\ . \end{aligned} \tag{5.5a}$$

La corrélation $\rho(X,Y)$ résultante dépend du coefficient a ($a \geq 0$), selon:

$$\rho(X,Y) = [\ -(1-a^2) - 2a\log_e a\ ]\ /\ (1-a)^2\ . \tag{5.5b}$$

Une autre approche, par addition (§5.4 et sections suivantes), consiste à construire les variables x et y' à partir de u_1 et u_2, selon:

$$\begin{aligned} x &\leftarrow u_1 \\ y' &\leftarrow a\cdot u_1 + (1-a)u_2, \end{aligned} \tag{5.6a}$$

selon $0 \leq a \leq 1$. La variable x reste uniforme, mais y' devient trapézoïdale et doit être ramenée à la densité uniforme. Soit $f_T(a)$ et $F_T(a)$, la densité et la f.r. de y'; alors,

$$y \leftarrow F_T(y';a) \tag{5.6b}$$

fournit la v.a. uniforme cherchée (voir exercice 5.4). La corrélation $\rho(X,Y;a)$ est approximativement déterminée, de même que la corrélation

correspondante entre les v.a. $G^{-1}(X)$ et $G^{-1}(Y)$ de quelques distributions. Noter que, en utilisant (5.6a), on a $p_a \geq 0$; pour obtenir une corrélation négative correspondante, (5.6a) doit être modifiée selon:

$$y' \leftarrow a(1-u_1) + (1-a)u_2 \,. \tag{5.6c}$$

La transformation (5.6) a été étudiée particulièrement par M. E. Johnson et A. Tenenbein (1981). Pour x et y distribuées selon $U(0,1,0,1,p)$, les auteurs proposent la relation suivante entre p(X,Y) et a(p>O,≤ a≤ 1):

$$\begin{aligned} \rho &= [a(10-13a]/[10(1-a)^2] \quad \{0 \leq a < 0{,}5\} \\ &= [3a^3+16a^2-11a+2]/10a^3 \quad \{0{,}5 \leq a < 1\}, \end{aligned} \tag{5.7}$$

relation qui donne lieu aux équivalences du tableau 5.1

Johnson et Tenenbein indiquent d'autres relations approximatives pour les lois normale, double exponentielle et exponentielle (exercice 5.5), le coefficient de corrélation utilisé étant la « corrélation de rangs » de Spearman, p_s[2]. Ce coefficient s'obtient en calculant le coefficient habituel (5.3) sur les *rangs* des X et Y, chacuns établis dans leur série respective.

Tableau 5.1 Relation entre p(X,Y) et le paramètre a de la fonction (5.6), les v.a. X et Y étant toutes deux de distribution $U(0,1)$[†]

ρ	0,0	0,01	0,05	0,1	0,2	0,3	0,4	0,5
a	0,0	0,00993	0,04832	0,09354	0,17607	0,25000	0,31732	0,37987
ρ	0,6	0,7	0,8	0,9	0,95	0,99	1,0	
a	0,43970	0,50000	0,57143	0,70000	0,74614	0,87280	1,0	

† Pour obtenir une corrélation négative, voir texte.

5.4 Le principe d'addition. Si les variables X et Y sont construites par l'addition de v.a. indépendantes et partagent ainsi certaines de leurs composantes, elles auront entre elles une corrélation d'une valeur déterminée.

2. Wilks (1962) montre qu'il y a une relation approximative et invertible entre $\rho_u = \rho(u_x,u_y)$, basée sur deux v.a. uniformes, et $\rho_N = \rho(z_x,z_y)$, les v.a. normales correspondantes, où $u_x = \Phi(z_x)$. Cette double relation est $\rho_u = (6/\pi)\sin^{-1}(\frac{1}{2}\rho_N)$ et $\rho_N = 2\sin[\pi\rho_u/6]$. Étant donné que r_X/n (où r_X est le *rang* de X, de 1 à n) et u_X (= F_X) ont asymptotiquement même distribution, la relation de Wilks se généralise au coefficient ρ_s de Spearman et justifie les approximations de Johnson et Tenenbein.

Deux recettes additives se présentent, la première basée sur l'addition simple de plusieurs composantes indépendantes:

$$\begin{aligned} X &= v_1 + v_2 + \cdots + v_k + v_{k+1} + \cdots + v_N \\ Y &= v_1 + v_2 + \cdots + v_k + v'_{k+1} + \cdots + v'_N \ . \end{aligned} \tag{5.8}$$

Les variables X et Y partagent k v.a. indépendantes parmi les N qui les constituent; si ces v.a. ont toutes même variance, alors $p(X,Y) = k / N$, simplement. L'autre recette, appliquée plus haut en variante pour la production de deux v.a. corrélées, est un composé linéaire de deux v.a. indépendantes, sous forme:

$$X = v_1 \ ; \ \ Y = a \cdot v_1 + \sqrt{(1-a^2)}\, v_2 \quad \{-1 \le a \le 1\}; \tag{5.9}$$

dans ce cas, si v_1 et v_2 ont même variance, alors p = a. La section suivante (voir aussi les exercices 5.10-5.11) présente une variante, qui se généralise à plus de deux v.a. corrélées.

Dans l'une et l'autre recettes, les variables produites ont une distribution autre que la distribution des variables originales v_1, v_2, etc. (sauf, bien entendu, la variable X dans (5.9)); la loi normale est la seule qui fait exception à la règle, la somme de deux ou plusieurs v.a. normales étant elle aussi normale: on parle alors d'une loi homogène. Nous examinons ce cas particulier au paragraphe suivant. Par ailleurs, pour quelqu'un désireux d'obtenir des v.a. corrélées et de distribution donnée, le principe d'addition fournit un moyen supplémentaire et parfois plus commode que celui consistant à se baser sur des v.a. uniformes corrélées: nous y revenons plus loin.

5.5 *Variables normales corrélées.* La loi multinormale, ou distribution normale multivariée, a pour densité:

$$f_k(\mathbf{X}) = \frac{1}{(2\pi)^{k/2}\sqrt{|\mathbf{V}|}} \exp\left[-\tfrac{1}{2}(\mathbf{X}-\mathbf{M})^t \mathbf{V}^{-1}(\mathbf{X}-\mathbf{M})\right] \ ; \tag{5.10}$$

ici, X est un vecteur-colonne ayant k composantes, tout comme M, le vecteur des espérances, avec $M_i = \mu(X_i)$, et V, d'ordre k x k, est la matrice des covariances, avec $V_{i,i} = 0^2(X_i)$ et $V_{i,j} = p_{i,j}\ 0(X_i)0(X_j)$. Or, la somme ou la différence de deux v.a. normales indépendantes sont encore normales, leurs moyennes étant la somme ou la différence des moyennes originales, leurs variances étant la somme des variances originales (Kendall et Stuart 1977). Ce théorème, doublé de la recette (5.9) basée sur la somme pondérée

Tableau 5.2 Coefficients a_{ii} de pondération de deux ou plusieurs v.a. normales pour produire autant de v.a. normales corrélées, selon $X_i = \Sigma\, a_{ij}\, z_j\ (j \le i)$[†]

	z_1	z_2	z_3	z_4
X_1	1			
X_2	ρ_{12}	$\sqrt{(1-\rho_{12}^2)}$		
X_3	ρ_{13}	$\rho'_{23.1}$	$\sqrt{(1-\rho_{13}^2-\rho'^2_{23.1})}$	
X_4	ρ_{14}	$\rho'_{24.1}$	$\rho'_{34.12}$	$\sqrt{(1-\rho_{14}^2-\rho'^2_{24.1}-\rho'^2_{34.12})}$

[†] Les coefficients ρ' sont des coefficients de corrélation semi-partielle, leur valeur étant obtenue par soustraction afin d'égaliser les espérances $\rho_{ij} = E(X_iX_j)$. Ainsi, $\rho'_{23.1} = (\rho_{23}-\rho_{12}\rho_{13}) / \sqrt{(1-\rho_{12}^2)}$, $\rho'_{24.1} = (\rho_{24}-\rho_{12}\rho_{14}) / \sqrt{(1-\rho_{12}^2)}$ et $\rho'_{34.12} = (\rho_{34}-\rho_{13}\rho_{14} - \rho'_{23.1}\rho'_{24.1}) / \sqrt{(1-\rho_{13}^2-\rho'^2_{23.1})}$. Voir exercice 5.7.

de deux v.a. indépendantes, permet de produire des vecteurs de v.a. normales corrélées satisfaisant à toute matrice de covariances (ou de corrélations) V.

Soit z_1, z_2, z_3, etc. des v.a. normales de distribution N(0,1), le tableau 5.2 donne les coefficients permettant de produire successivement les v.a. normales X_1, X_2, X_3 •••, de corrélations P_{12}, P_{13}, P_{23} ··· spécifiées, où $p_{ij} = p(X_i, X_j)$ (voir aussi les exercices 5.7-5.8).

Une autre méthode, dite symétrique, permet de produire facilement des v.a. normales standard X_1, X_2, ..., X_k à corrélations $p(X_i,X_j) = p$. Soit Z_o, Z_1, ..., Z_k, des v.a. normales standard indépendantes, alors les variables

$$X_i = \sqrt{\rho}Z_0 + \sqrt{(1-\rho)}Z_i \qquad (5.11)$$

ont le comportement voulu. Enfin, Gentle (1998) rapporte une méthode qui fournit deux v.a. X,Y normales standard à corrélation ρ(X,Y), grâce à une variante astucieuse de la technique de Box et Muller (§4.7 et (4.15)). Soit $\omega = \cos^{-1}\rho$ (en radians). Obtenant deux v.a.u u_1 et u_2, alors :

$$X \leftarrow \sqrt{(-2\log_e u_1)} \times \sin(2\pi u_2) \qquad (5.12)$$
$$Y \leftarrow \sqrt{(-2\log_e u_1)} \times \sin(2\pi u_2 + \omega)$$

sont deux v.a. normales standard conjointes, à corrélation ρ.

5.6 *Lois additives.* La loi normale, on l'a vu, est une loi homogène, en ce sens qu'elle conserve sa forme sous l'addition: par exemple, si z_1, z_2, etc. sont des v.a. normales de loi $N(0,1)$, alors $(z_1+z_2+\cdots+z_k)/\sqrt{k}$ est aussi distribuée comme $N(0,1)$. Plusieurs autres lois, sans être homogènes, sont additives: la somme de deux ou plusieurs v.a. issues d'une loi pareille est distribuée selon une loi de la même famille, de gré avec un changement de paramètre.

Quelles sont ces lois additives? Pour en identifier quelques-unes, nommons la loi Gamma[3]: *[Ga(a_1,* B) + Ga(a_2,B) →Ga(a_1+a_2, B)], incluant la loi Khi-deux comme cas particulier, dans lequel B = 2 et a = ½v: [$x^2(v_1)$ + $x^2(v_2) \rightarrow x^2(v_1+v_2)$]; la loi binomiale: *[B(n_1,* л*)* + *B(n_2,* л*)* →*B(n_1+n_2,* л*);* la loi de Poisson: *[Po(λ_1)* + *Po(λ_2)* → *Po(λ_1+λ_2)];* la loi de Pascal: *[Pa(r_1,* л*)* + *Pa(r_2,* л*)* Pa(r_1+r_2, л*)],* etc. Chacune de ces lois reste dans la même famille sous l'opération d'addition; certaines, comme les lois Gamma et Pascal, incluent une autre distribution fondamentale comme cas particulier, soit les lois exponentielle et géométrique, respectivement. On peut ajouter à ce groupe d'autres lois pour lesquelles la somme de variables a une distribution autre mais connue, telles la loi uniforme et la loi multinomiale. Le traité de Johnson et Kotz (1992, 1994, 1995) donne les informations suffisantes.

Supposons par exemple que nous voulions produire deux variables positives, corrélées positivement et ayant une légère asymétrie positive: alors l'addition de quelques v.a. de loi x^2 pourra faire l'affaire.

> Pour mieux illustrer la procédure, posons que nous désirons deux v.a. positives, à corrélation p≈ 0,25 et à asymétrie $\gamma_1 \approx 0{,}5$. La loi du Khi-deux paraît une bonne candidate. Pour cette loi, à paramètre v, $\gamma_1 = \sqrt{(8/v)} = 0{,}5$, d'où $v = 32$. Pour obtenir une corrélation p = 0,25, on peut fractionner les variables résultantes en 2 parties, soit x_1 distribuée selon $x^2(8)$, *et* x_2 et x_3 distribuées chacune selon $x^2(24)$. Dans ce cas, les variables:
>
> $$X \leftarrow x_1 + x_2 \; ; \quad Y \leftarrow x_1 + x_3$$
>
> seront chacune distribuée comme $x^2(32)$, avec asymétrie (γ_1) de 0,5 et en ayant une corrélation de 0,25.

3. Il faut distinguer la *fonction* Gamma, $\Gamma(x)$ (exercice 4.18) de la *loi* (ou distribution) Gamma, *Ga(a,B),* définie à l'exercice 5.9. On obtient la loi Gamma standard, $G_x(a)$, en posant B = 1, et la densité devient simplement $x^{a-1}e^{-x}/\Gamma(a)$. La loi standard $G_x(a)$, pour a positif et entier, peut être vue comme la somme $x_1+x_2+...+x_a$ de a v.a. exponentielles E(1), de densités individuelles e^{-x}.

Nous avons illustré ici la situation symétrique dans laquelle les deux v.a. produites ont même distribution. Le cas général, dans lequel les v.a. sont obtenues par une addition asymétrique (dans l'exemple ci-haut, x_2 et x_3 répondraient à des lois x^2 de paramètres différents), se solutionne aussi facilement et s'adapte à des contraintes de production plus variées.

5.7 Le *mélange de distributions* consiste à produire des v.a. successivement issues d'une loi ou d'une autre, selon des probabilités déterminées: on obtient ainsi une variable *mélangée,* en fait une variable hybride, d'une espèce nettement différente de celle obtenue par addition. Soit les v.a. x_1 et x_2 et leurs distributions $f_1(x_1)$ et $f_2(x_2)$, alors:

$$x_M \leftarrow \{ (x_1)_{\omega \in (0,P)}, (x_2)_{\omega \in (P,1)} \}, \qquad (5.13)$$

expression dans laquelle w est le *sélecteur aléatoire et P* identifie le dosage du mélange: habituellement, w est une v.a.u., de loi $U(0,1)$.

> Ne pas confondre le *mélange* présenté ici avec la *mixture,* qu'on retrouve plus souvent dans la littérature statistique (Johnson *et al.* 1994, 1995; Kendall et Stuart 1977). La mixture désigne habituellement une loi, disons $f_1(p_1 p_2, \cdots)$ dotée de paramètres, dont au moins l'un des paramètres, tel p_i, varie selon une autre loi, disons f_2. La loi *Bêta-Binomiale,* une Binomiale dont le paramètre л varie selon une Bêta, en est un exemple. Le vocable « mixture » a parfois été employé pour désigner également le *mélange de distributions* (voir Blischke, Kruskal et Tanur 1978, p. 174 et suiv.).

Les moments de la variable mélangée sont un mélange proportionnel des moments correspondants des variables constitutives: par exemple, $\mu(x_M) = P\mu(x_1)+(1-P)\mu(x_2)$, $0^2(x_M) = P0^2(x_1)+(1-P)0^2(x_2)$, etc. Il en est ainsi de la covariance, $0(x_M,y) = P0(x_1, y) + (1-P)0(x_2,y)$, ce qui permet de produire facilement des variables corrélées (X_M,Y) ayant la corrélation p et les moments univariés voulus. Pour produire les variables (X,Y) corrélées et de distributions marginales f_x et f_y semblables, il suffit d'assigner les v.a. indépendantes x_1 et x_2 de mêmes lois, comme suit:

$$x_M \leftarrow \{ (x_1)_{\omega \in (0,P)}, (x_2)_{\omega \in (P,1)} \} \qquad (5.14)$$

et:

$$y \leftarrow x_1 ;$$

La loi conjointe $g(X_M,Y)$ des variables définies en (5.14) est un hybride de lois univariées et elle est dite *dégénérée.* En d'autres mots, la distribution du couple (X_M,Y) n'est pas une loi bivariée mais elle en imite les moments. Cette méthode, facile et très générale quant à la forme de

distribution et au degré de corrélation voulus, présente ainsi une anomalie fondamentale quant au comportement des v.a. produites. L'utilisateur devra décider si l'anomalie de la loi bivariée constitue ou non une infirmité rédhibitoire pour son application.

5.8 *Variables discrètes corrélées.* Deux variables discrètes, X et Y, sont dites indépendantes et sont non corrélées si leurs probabilités conjointes $p(X_iY_j)$ sont égales au produit de leurs probabilités propres, $p(X_i)p(Y_j)$, pour tout *i, j;* dans les cas contraires, ces variables sont associées, ou « corrélées » au sens large. Pour produire de telles variables, il est donc possible de construire un tableau *k(X) x k(Y)* contenant les probabilités conjointes $p_{ij} = p(X_i, Y_j)$, puis d'y repérer le couple (x_i, y_j) au hasard proportionnel, selon une méthode ou l'autre (exercices 5.13-5.14).

Pour obtenir des variables discrètes linéairement corrélées, c'est-à-dire en correspondance proportionnelle plus ou moins stricte, les moyens disponibles se restreignent sérieusement. Le moyen privilégié reste l'addition (§5.4 et §5.6), l'autre est le mélange (§5.7) de v.a. indépendantes. L'autre moyen, général mais problématique, consiste à spécifier les probabilités conjointes p_u; les publications se font rares sur cette épineuse question.

Les variables de Bernoulli, ou variables binaires (à valeurs 0 ou 1), constituent un cas particulier pour lequel la solution est simple. Soit les probabilités individuelles p(x=1) = л $_x$, *p(x=0)* = 1p(y=l) = л $_y$, *p(y=0)* = 1- л $_y$. Dans ce cas en effet, la corrélation linéaire p(X,Y) dépend essentiellement de $p_{11} = p(x=1,y=1)$, selon l'équation :

$$p_{11} = \pi_x\pi_y + \rho\sqrt{[\pi_x\pi_y(1-\pi_x)(1-\pi_y)]}; \qquad (5.15)$$

de là, on obtient p_{10} = л $_x$-p_{11}, P_{01} = л $_y$-P_{11}, P_{00} = 1- л $_x$- л $_y$+P_{11} Toutes ces probabilités devant être positives dans l'intervalle (0,1), certaines combinaisons s'avèrent impossibles. Ainsi, p ne peut atteindre 1 qu'à condition que л $_X$ = л $_y$. L'ensemble des probabilités ainsi spécifiées permet alors de générer les v.a. souhaitées: l'exercice 5.17 explicite la solution. Park, Park et Shin (1996, voir aussi Gentle 1998, p. 114) proposent une autre approche.

> Pour produire v.a. discrètes corrélées, on peut modifier le tableau des probabilités conjointes { p_{ij} } de façon simpliste, en accroissant la probabilité associée aux couples à corrélation maximale et en réduisant les autres couples également. Par exemple, si X et Y sont toutes deux de loi rectangulaire R(1,*k*) (§4.3), le tableau bivarié est de dimensions *k x k.* Lorsque X et Y sont mutuellement indépendantes, satisfaisant $p_{ij} = p_i \times p_j$, les probabilités conjointes sont uniformément

égales à $1/k^2$. Si l'on modifie les probabilités diagonales $p_{ii} = p(X_i,Y_i|X_i=Y_i)$ et les autres, $p_{ij} = p(X_i,Y_j|X_i\neq Y_j)$, par:

$$p_{ii} = [1+(k-1)\delta]/k^2 \;\; ; \;\; p_{ij} = (1-\delta)/k^2 \;, \qquad (5.16)$$

alors la série bivariée a une corrélation $\rho = \delta$. Qui plus est, le tableau ainsi élaboré représente en fait un *mélange* de distributions!

5.9 *Corrélation sérielle.* Il existe une autre forme de corrélation que celle concernant deux v.a. distinctes: c'est la *corrélation sérielle*, $\rho(x_i,x_{i+1})$, existant entre les valeurs consécutives d'une série statistique. Le sujet est développé plus loin, en §6.14. La méthode la plus simple pour produire une série à corrélation ρ est d'appliquer une formule de récurrence, soit:

$$x_i \leftarrow \rho x_{i-1} + \sqrt{(1-\rho^2)}\, z_i \;, \qquad (5.17)$$

formule dans laquelle les $\{ z_i \}$ constituent une série de v.a. indépendantes de référence[4], et p est la corrélation appliquée.

La série (5.17) est aussi désignée *fonction auto-régressivede premier ordre;* une série de second ordre impliquerait deux valeurs antécédentes, *i.e.* « $x_i \leftarrow p_1x_{i-1} + p_2x_{i-2} + \sqrt{(1-p^2_1-p^2_2)}z_i$ », ainsi de suite. La documentation spécialisée dans les problèmes reliés à l'analyse des *séries temporelles* présente les techniques d'estimation de l'ordre et des coefficients de régression, etc.; cette analyse est connexe *à l'analyse spectrale* (décomposition en polynômes de Fourier): voir Cox et Lewis (1969), Jenkins et Watts (1968) ou Priestley (1981).

On peut appliquer un calcul semblable à (5.17) pour imposer une corrélation sérielle, ou *dépendance séquentielle,* à des variables de Bernoulli (X = 0 ou X = 1). La formule symbolique:

$$x_i \leftarrow Px_{i-1} + (1-P)\, z_i \qquad (5.18)$$

indique de répéter x_{i-1} (sur x_i) selon une probabilité P ou d'engendrer une nouvelle variable de Bernoulli z_i selon la probabilité complémentaire 1- P; la variable de Bernoulli est elle-même «1» avec probabilité л, et «O» avec 1- л. Cette séquence constitue en fait une *chaîne de Markov* discrète, à deux états (Kemeny et Snell 1976). Moore (1979) et d'autres ont étudié les problèmes spécifiques d'estimation pour ces séries.

4. Le facteur de z_i est arbitraire et il n'a pour but que d'assurer à la série $\{x_i\}$ la même variance que la série $\{z_i\}$.

Exercices

5.1 Montrer que p(X,Y) = ±l implique l'identité de distributions pour X et Y et qu'en conséquence, | p | < 1 si X et Y sont de distributions différentes.

5.2 Trouver le coefficient a dans (5.5a,b) pour obtenir p = 0,5; pour obtenir p = 0,0.

5.3 Soit deux v.a. indépendantes, X et Y, et a (0 a 1) un paramètre de pondération. Développer algébriquement (en espérance) la corrélation p[X,aX+(1-a)Y]. Si X et Y sont toutes deux uniformes, la valeur trouvée indique la corrélation entre une v.a. uniforme et une v.a. trapézoïdale.

5.4 Soit U_1 et U_2, deux v.a. uniformes, et a ($0 \leq a \leq 1$) un paramètre de pondération; $X = aU_l+(1-a)U_2$, un mélange de deux uniformes, observe une loi *trapézoïdale;la* loi uniforme *(a = 0* ou *a* = 1) et la loi triangulaire (a = ½*)* sont deux cas particuliers.

Montrer que la densité est f(X) = X / [a(1- a)], 1 / (1-a), (1-X) /[a(1-a)] respectivement pour $0 \leq X < a$, $a \leq X < 1$-a, $1-a \leq X < 1$. De même, la f.r. correspondante est $X^2/[2a(1-a)]$, *(2X-a)/[2(1-a)],* $1-(1-X)^2/[2a(1-a)]$. Les moments de X sont, selon a: μ 0,5; $0^2 = (2a^2-2a+1)/12$; $\mu_3 = y_1 = 0$; μ_4 = (4a2—6a+3)· (4a² -2a+1)/240 et y_2 = -1,2·*(2a⁴ -4a³ +6a²-4a+1)/(2a²-2a+1)* .

5.5 Johnson et Tenenbein (1981) indiquent une correspondance approximative entre le paramètre a, dans (5.6a), et la corrélation ordinale, ou de Spearman, p_s (voir aussi note infrapaginale 2), pour les lois *normale, exponentielle* et *double exponentielle.* Si, dans (5.6a), *x* et *y'* sont toutes deux transformées en v.a. *normales* x_N, Y_N, alors $p_s(x_N,y_N) \approx (6/л) \sin^{-1}[½a/\sqrt{[a^2+(1-a)^2]}$. Si elles sont transformées en variables de loi *exponentielle (voir* §4.8) x_E, y_E, on a $p_s(x_E y_E) \approx a(9 -18a^2+14a^3 - 3a^4)/[2(2 -a)^2]$. Enfin, la loi *double exponentielle,* ou « première loi de l'erreur » de Laplace (la seconde étant la loi normale), a pour densité $'½e^{-|x|}$, $-\infty < x < x\infty$. Après transformation dans cette loi, les variables x_{DE}, y_{DE} auront corrélation $p_s(x_{DE},y_{DE}) \approx a(3 - 2a)/(2 -a)$·

Établir les valeurs du paramètre a correspondant aux valeurs de p affichées au tableau 5.1, mais cette fois pour les distributions *normale, exponentielle* et *double exponentielle.* Faire une courte étude Monte

Carlo pour déterminer les correspondances p:a et p:p_s, p étant le coefficient de corrélation usuel.

5.6 Utilisant des v.a. indépendantes notées v_1, v_2, ..., comme dans les expressions (5.8), chacune avec une variance notée 0^2_1, 0^2_2, ..., nous pouvons construire de nouvelles v.a. X et Y par l'addition de composantes v_i. Démontrer que, dans ce cas, la corrélation entre X et Y est donnée par:

$$\rho(X,Y) = \frac{\sum_{X,Y} \sigma_i^2}{\sqrt{\left(\sum_{X,Y} \sigma_i^2 + \sum_X \sigma_j^2\right)\left(\sum_{X,Y} \sigma_i^2 + \sum_Y \sigma_j^2\right)}} ,$$

où Σ_X y 0^2_i, Σ_x 0^2_j , Σ_y 0^2_j désignent respectivement la somme des variances des composantes communes à X et Y, celle des composantes de X, enfin celle des composantes de Y. La corrélation entre des variables X et Y ainsi construites apparaît donc comme la moyenne géométrique de la proportion de variance que chacune partage avec l'autre[5]. Lorsque les composantes ont toutes même variance, que les v.a. X et Y en comportent N, parmi lesquelles *k* sont communes, alors p(X,Y) = *k/N*.

5.7 Soit z_1, z_2, etc., des v.a. normales indépendantes, et X_1, X_2, etc. des v.a. normales à corrélations p_{ij} prescrites. La méthode de production de v.a. normales proposée en §5.5 implique l'utilisation de coefficients de corrélation semi-partielle. Prenons l'exemple du coefficient $p'_{23.1}$. Pour ce cas, $X_1 = z_1$; $X_2 = P_{12}z_1 + Q_2z_2$, où $Q_i = (1-\Sigma^{w2}_{ij})$, j < i; $X_3 = P_{13}z_1 + w_{23}z_2 + Q_3z_3$. Le coefficient $p'_{23.1} = w_{23}$ doit être tel que $p(X_2,X_3) = p_{23}$; les variables étant standardisées (avec $0^2 = 1$), on a:

$$\begin{aligned}\rho_{23} = \rho(X_2,X_3) &= \mathrm{cov}(X_2,X_3) \\ &= \mathrm{cov}[\ \rho_{12}z_1+Q_2z_2,\ \rho_{13}z_1+w_{23}z_2+Q_3z_3\ ],\end{aligned}$$

soit, après simplification:

$$\rho_{23} = \rho_{12}\rho_{13} + Q_2w_{23}\ ;$$

de là, on obtient $w_{23} = \rho'_{23.1} = (\rho_{23}-\rho_{12}\rho_{13})/\sqrt{(1-\rho^2_{12})}$. L'obtiention de Q_3 est ensuite facile.

---

5. Cette interprétation découle de l'équivalence symbolique $\rho_{X,Y} = \sigma_{X,Y} / \sqrt{[(\sigma_{X,Y}+\sigma^2_X)(\sigma_{X,Y}+\sigma^2_Y)]} = \sqrt{[\,\sigma_{X,Y}/(\sigma_{X,Y}+\sigma^2_X) \times \sigma_{X,Y}/(\sigma_{X,Y}+\sigma^2_Y)\,]}$, au signe près.

Par ce procédé, trouver les coefficients nécessaires à la production de X_5, à corrélations p_{15}, p_{25}, p_{35} et p_{45} déterminées.

5.8 L'approche suggérée à l'exercice précédent correspond en fait à une suite de deux opérations matricielles. La première consiste à factoriser la matrice de covariances **V** en un produit AA', où A est une matrice triangulaire inférieure (semblable à celle présentée au tableau 5.2). La seconde opération, le produit matriciel X = AZ, fabrique le vecteur multinormal corrélé X à partir du vecteur multinormal indépendant Z. Scheuer et Stoller (1962; voir aussi Rubinstein 1981, p. 65-67) donnent l'algorithme de factorisation, en présentant les règles suivantes:

$$a_{i1} = v_{i1} / \sqrt{v_{11}} \qquad \{ 1 \le i \le k \} \quad (5.19)$$

$$a_{ii} = \sqrt{v_{ii} - \sum_{k=1}^{i-1} a_{ik}^2} \qquad \{ 1 < i \le k \}$$

$$a_{ij} = \left[v_{ij} - \sum_{k=1}^{i-1} a_{ik} a_{jk} \right] / a_{jj} \qquad \{ 1 < j < i \le k \}$$

$$a_{ij} = 0 \; . \qquad \{ i < j \le k \}$$

Depuis une matrice de corrélations 4 x 4, appliquer les règles ci-haut et reconstituer les coefficients du tableau 5.2. Rédiger un programme qui fabrique automatiquement les coefficients a_{ij}.

5.9 *La loi Gamma, Ga(a,B).* Une variable *x* est de loi Gamma, *Ga(a,B),* si sa densité est:

$$\mathrm{pr}(x) = x^{\alpha-1} e^{-x/\beta} / [\beta^{\alpha} \Gamma(\alpha)] \quad \{x \ge 0\}; \qquad (5.20)$$

la fonction Gamma, $\Gamma(a)$, est définie à l'exercice 4.18. Les moments à l'origine de cette variable sont $\mu'_r = B^r \Gamma(a+r)/\Gamma(a)$, d'où on obtient $\mu = aB$, $0^2 = aB^2$, $y_1 = 2/\sqrt{a}$ et $y_2 = 6/a$. Il s'agit donc d'une distribution d'asymétrie positive et leptokurtique, mais qui se rapproche de la loi normale quand $a \to \infty$. Quel est le mode de cette distribution, Mo(a) ?

Posant B = 2 et a = ½*v*, nous obtenons la loi du Khi-deux avec *v* degrés de liberté, $x^2(v)$ (voir §4.8). Quels sont les moments et le mode de cette loi ?

5.10 La méthode symétrique de production de v.a. corrélées, concrétisée par (5.11), donne aux v.a. produites une particularité statistique, révélée par leurs corrélations partielles. Quelle est cette particularité?

5.11 La méthode symétrique, exprimée par les fonctions (5.11), peut être relaxée dans le but d'imposer des corrélations ρ_{ij} diverses entre les v.a. produites. Pour ce faire, chaque v.a. X, est produite selon :

$$X_i \leftarrow \sqrt{c_i}\, Z_0 + \sqrt{(1-c_i)}\, Z_i \ ,$$

en déterminant les coefficients e_t de manière appropriée. Montrer que, pour produire trois v.a. X_1, X_2 et X_3 à corrélations ρ_{12}, ρ_{13} et ρ_{23}, les coefficients requis sont :

$$c_1 = \rho_{12}\rho_{13}/\rho_{23} \ ; \ c_2 = \rho_{12}\rho_{23}/\rho_{13} \ ; \ c_3 = \rho_{13}\rho_{23}/\rho_{12} \ ,$$

ce en respectant les contraintes $0 < c_i \leq 1$. Généraliser cette recette au-delà de trois variables. Montrer que chaque nouvelle variable X_i, après la troisième, ajoute un degré de liberté et permet donc de fixer arbitrairement un et un seul coefficient ρ_{ij}.

5.12 Démontrer la validité de la méthode exprimée par (5.12). [Suggestion: ou bien, montrer que $\iint X(u_1,u_2)\, du_1 du_2 = \iint Y(u_1,u_2)\, du_1 du_2 = 0$, $\iint X^2(u_1,u_2)\, du_1 du_2 = \iint Y^2(u_1,u_2)\, du_1 du_2 = 1$ et $\iint X(u_1,u_2)\times Y(u_1,u_2)\, du_1 du_2 = \rho$, ou bien, admettant la validité de (4.15) et des coefficients du tableau 5.2, montrer l'équivalence de « $\sin(2\pi u+\cos^{-1}\rho)$ » et « $\rho\sin(2\pi u)+\sqrt{(1-\rho^2)}\cos(2\pi u)$» .]

5.13 Soit une distribution bivariée discrète, $p(X_i,Y_j)$, $1 \leq i \leq I$, $1 \leq j \leq J$. Appliquant la technique du tableau simple (§4.11) pour la loi bivariée $p(X_i,Y_j)$, on peut construire deux tableaux conjugués, $T_p(u)$ et $T_{xy}(u)$, $1 \leq u \leq IxJ$, le premier (T_p) contenant des probabilités (conjointes) cumulatives, le second (T_{xy}) les couples de valeurs (X,Y) associées. Elaborer un algorithme de production de v.a. discrètes corrélées selon le principe du tableau simple et ses diverses optimisations.

5.14 La loi bivariée $f(X,Y)$, qu'il s'agisse de variables discrètes ou continues, peut être décomposée en la ***conditionnant*** sur X, selon:

$$f(X,Y) = f(Y \mid X)\, f(X).$$

Cette relation suggère de produire le couple (X,Y) en deux phases: (1) produire d'abord $X \sim f(X)$ par une méthode quelconque; (2) produire ensuite $Y \sim f(Y \mid X)$. Dans le cas de v.a. discrètes, $f(Y \mid X)$ est facile

à spécifier. Si, de plus, $f(X)$ est une loi simple (v.g. loi binomiale ou loi rectangulaire), la première phase peut être grandement accélérée.

Dans le contexte de l'exercice précédent, élaborer un algorithme de production de v.a. (X,Y) en deux phases. Comparer l'efficacité des deux approches et déterminer sous quelle(s) condition(s) l'une est préférable à l'autre.

5.15 *Le coefficient de contingence, Co.* On peut mesurer la corrélation, au sens large, entre les caractères X et Y, que ceux-ci soient numériques ou non, en construisant un « tableau de contingence » T(x,y), qui reflète la covariation des caractères, et en calculant le *coefficient de contingence (Co),*

$$Co = \sqrt{\frac{X^2}{X^2 + n}} \ , \qquad (5.21)$$

proposé par K. Pearson. Chaque co-apparition $(X_i = x_j, Y. = y_k)$ est repérée dans le tableau $T(x,y)$ et inscrite à la cellule (j,k) appropriée. Au besoin, si elles sont numériques, les données (X_i ou Y_i) devront être catégorisées. Les n données de la série $\{ X_i, Y_i \}$ sont ainsi transcrites dans le tableau de contingence, comportant L lignes et C colonnes. Une statisique X^2 (formule (6.1)) est alors calculée, qui compare les fréquences observées f_{jk} du tableau T(x,y) aux fréquences attendues ft_{ik} sous l'hypothèse de l'indépendance statistique, soit $E(ft_{jk}) = nL_jC_k$, L_j dénotant la proportion d'observations dans la j^e ligne du tableau, C_k la proportion dans la k^e colonne. La statistique X^2 obtenue est appelée « test d'interaction » ou « test d'indépendance » et elle se distribue approximativement comme x^2 avec $v = (L-1)x(C-1)$. Le coefficient Co est donc une mesure du degré de dépendance plutôt que de corrélation linéaire entre les caractères X et Y.

Montrer que le coefficient Co a pour maximum $\{ [\min(L,C) - 1] / \min(L,C)\}^{1/2}$ et que, si L = C, $|p(X_i,X_j)| \rightarrow$ Co lorsque les p_{ij} [$= p(X_i,Y_j)$] décroissent autour de la diagonale du tableau, a mesure que $|i-j|$ croît.

5.16 *Le coefficient de corrélation φ.* Pour des variables de Bernoulli X,Y mentionnées en §5.8, le tableau de probabilités bivariées (p_{11}, p_{10}, p_{01}, p_{00}) permet le calcul du coefficient de corrélation linéaire ρ. Montrer qu'alors ρ = φ, où φ (le coefficient « phi ») est:

$$\varphi = \frac{p_{11}p_{00} - p_{10}p_{01}}{\sqrt{p_{1.}(1-p_{1.})\,p_{.1}(1-p_{.1})}} \ ; \qquad (5.22)$$

pour des valeurs observées de X,Y, on utilise les estimateurs correspondants, tels n_{11}, n_{10}, etc. Montrer que la valeur de φ, située en principe entre -1 et 1, est en fait bornée par $-\min\{\sqrt{[p_{.1}/(1-p_{.1})]}\ \sqrt{[p_{1.}/(1-p_{1.})]},\ \sqrt{[(1-p_{.1})/p_{.1}]}\sqrt{[(1-p_{1.})/p_{1.}]}\}$ et $+\sqrt{[p_{.1}/(1-p_{.1})]} \times \sqrt{[(1-p_{1.})/p_{1.}]}$, cette seconde borne selon $p_{1.} \geq p_{.1}$. Aussi, noter que $\varphi^2 = nX^2$, où X^2 est la statistique (ou test) X^2 d'interaction (distribuée approximativement comme χ^2_1) sur le tableau 2 x 2 d'observations n_{ij}.

5.17 Considérons les probabilités de Bernoulli conjointes présentées en §5.8 et selon (5.13) et le tableau cumulatif T[O] = P_{00}, T[1] = T[0] + P_{10}, T[2] = T[1] +P_{01}, T[3] = 1. Prouver que l'algorithme suivant génère des v.a. de Bernoulli (X,Y) respectant la spécification ($л_x$, $л_y$, p), à partir de v.a. uniformes u.

Production de v.a. de Bernoulli X,Y à corrélation ρ (5.23)

Produire u ;
$i \leftarrow 0$;
Tant que $u >$ T[i] faire $i \leftarrow i + 1$;
Selon $i = 0$ produire X ← 0 et Y ← 0,
$i = 1$ produire X ← 1 et Y ← 0,
$i = 2$ produire X ← 0 et Y ← 1,
$i = 3$ produire X ← 1 et Y ← 1.

5.18 Montrer que la modification simpliste du tableau *k x k,* en (5.16), est un mélange. Dissocier les distributions mélangées et en identifier le *dosage (P).*

5.19 Les variables de la série auto-régressive d'ordre 1, en (5.17), ont une corrélation sérielle $p(x_i,x_{i-1}) = p$. Quelle est $p(x_i,x_{i-s})$? Dans le cas d'une fonction auto-régressive d'ordre 2, avec les coefficients de régression p_1 et p_2, quelle est $p(x_i,x_{i-1})$? en général, $p(x_i,x_{i-s})$?

Références

BLISCHICE, W.R., KRUSKAL, W.H., TANUR, J.M. (DIR.). (1978). *International encyclopedia of statistics,* Vol. 1. New York, Free Press.

CHÂTILLON, G. (1984). The balloon rules for a rough estimate of the correlation coefficient. *The American Statistician,* 38, 58-60.

Cox, D.R., LEWIS, P.A.W. (1969). *L'analyse statistique des séries d'événements.* Paris, Dunod [édition originale anglaise en 1966].

DEVROYE, L. (1986). *Non-uniform random variate generation.* New York, Springer-Verlag.

GENTLE, J.E. (1998). *Random number generation and Monte Carlo methods.* New York, Springer-Verlag.

JENKINS, G.M., WATTS, D.G. (1968). *Spectral analysisand its applications.* San Francisco, Holden-Day.

JOHNSON, N.L., KOTZ, S., BALAKRISHNAN, N. (1994,1995). *Continuous univariate distributions, Vols.* 1 et 2. (2^{e} édition). New York, Wiley.

JOHNSON, N.L., KOTZ, S. KEMP, A.W. (1992). *Univariate discrete distributions* (2^{e} édition). New York, Wiley.

JOHNSON, M.E., TENENBEIN, A. (1981). A bivariate distribution family with specified marginals. *Journal of the American Statistical Society,* 76, 198-201.

KENDALL, M.G., STUART, A. (1977). *The advanced theory of statistics.* Vol. 1: *Distribution theory.* New York, Macmillan.

KEMENY, J.G., SNELL, J.L. (1976). *Finite Markov chains.* New York, Springer-Verlag.

LAURENCELLE, L. (1993). La loi uniforme: propriétés et applications. *Lettres statistiques,* 9, 1-23.

LAURENCELLE, L. (2000). Distribution, moments et estimation de la corrélation dans une population normale bivariée. *Lettres statistiques,* 11, 31-45.

MOORE, M. (1979). Alternatives aux estimateurs à vraisemblance maximale dans un modèle de Bernoulli avec dépendance. *Annales des sciences mathématiques du Québec,* 111, 119-133.

OLKIN, I., PRATT, J.W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of mathematical statistics,* 29, 201-211.

PARK, C.G., PARK, T., SHIN, D.W. (1996). A simple method for generating

PRIESTLEY, M.B. (1981). *Spectral analysis and time series.* London, Academic Press.

RUBINSTEIN, R.Y. (1981). *Simulation and the Monte Carlo method.* New York, Wiley.

SCHEUER, E.M., STOLLER, D.S. (1962). On the generation of normal random vectors. *Technometrics,* 4, 278-281.

WILKS, S.S. (1962). *Mathematical statistics.* New York, Wiley.

Chapitre

6 Tests d'hypothèses sur l'irrégularité des séquences de nombres

6.1 La méthode Monte Carlo utilise des nombres aléatoires à la manière d'un combustible, afin de vitaliser un modèle structurel ou de reproduire la variété échantillonnale de données d'observation. Elle s'apparente en cela aux techniques de statistique inférentielle, pour lesquelles les « données » sont réputées être des réalisations d'une variable aléatoire issues d'une loi de probabilité, très souvent la loi normale.

Pour appliquer la méthode Monte Carlo, dans l'une ou l'autre de ses variantes, et pour en obtenir des conclusions valides ou des estimations justes, les nombres utilisés doivent être statistiquement adéquats, c'est-à-dire qu'ils doivent présenter les propriétés d'une variable aléatoire à distribution connue. Cependant, comme on l'a vu au chapitre 4, les diverses distributions s'obtiennent ordinairement à partir de la distribution uniforme; il suffit donc souvent de valider la source de nombres aléatoires uniformes en vérifiant que les v.a. u_I produites constituent des séquences de v.a. indépendantes et de distribution U(0,1).

6.2 Le test des propriétés de distribution pour des séries de nombres déborde largement le seul cadre des applications de la méthode Monte Carlo. Des scientifiques de toutes disciplines se sont demandé, à propos d'observations, si celles-ci étaient apparues en ordre fortuit, si elles évoluaient au hasard ou, généralement, si, d'une façon ou d'une autre, elles juraient avec le comportement attendu de v.a. indépendantes. Certains tests ont été faits pour, ou même par, des chercheurs en psychologie, en biologie, en agronomie. Dans ces cas, la distribution de référence n'est plus uniquement la loi uniforme, voire cette distribution n'est pas même spécifiée, ce qui donne lieu à des tests dits a-distributionnels (Bradley 1968) ou non paramétriques (Lehmann 1975).

Les tests de normalité, enfin, constituent une classe à part: ils aident à décider si une série statistique donnée répond ou non à la forme de la distribution normale. Leur application la plus courante concerne les tests d'hypothèse de type fishérien, qui utilisent l'une ou l'autre des lois normale, *t,* x^2 et F. Dans ce contexte, les tests de normalité servent à valider la

condition de distribution normale imposée à la variable de base. Une excellente revue de ces tests se trouve dans Mardia (1980). Nous nous penchons sur ces tests au paragraphe §6.11.

6.3 Ii s'agit donc, à partir d'une séquence de nombres, d'un groupe de valeurs successivement produites par un générateur mécanique ou issues d'observations, de déterminer si l'hypothèse du hasard est tenable: étant donné telle séquence de *n* nombres, l'hypothèse *nulle* voulant que ces nombres proviennent d'une source purement aléatoire, à loi de distribution spécifiée, est-elle plausible? La question ainsi posée peut donner lieu à une série de tests statistiques, chacun traitant un aspect particulier du comportement attendu de la variable aléatoire. Chaque test porte sur la séquence de nombres et permet de se prononcer sur *l'irrégularité locale* de la source, par opposition à son irrégularité globale (voir §2.12).

Le test, effectué sur une séquence, peut être « significatif » ou non. Il s'agit d'abord de bien identifier la distribution échantillonnale du test, c'est-à-dire la loi probabiliste qui gouverne la variation de son résultat lorsque les données proviennent d'une source purement aléatoire. Puis, on établit la probabilité extrême, $p_{ext'}$ du résultat observé; celle-ci est définie comme la probabilité (sous l'hypothèse d'une variation au hasard pur) d'un résultat égal au résultat observé ou d'un plus éloigné encore de la valeur centrale. Étant donné un seuil de signification a ($0 < a < \%2$), le test est déclaré significatif si $p_{ext} \leq a$, auquel cas l'hypothèse d'une source purement aléatoire pour nos données devra être rejetée.

L'*irrégularité globale* de la source, mentionnée plus haut, peut elle aussi être soumise à des tests; nous y reviendrons bientôt (en §6.26). Entretemps, quelles sont les propriétés, qu'on doit retrouver dans nos séquences de nombres, qui reflètent leur origine dans une source aléatoire à loi de distribution f donnée?

Propriétés des séquences aléatoires

6.4 Les tests répertoriés et les tests possibles sur les séquences de nombres aléatoires sont très nombreux, aussi convient-il d'en tenter un classement. Nous proposons le classement suivant, basé sur les propriétés des v.a. issues d'une loi de distribution *f* quelconque:

Classe 1: tests sur la *forme* et les *moments* de la distribution. Ces tests dépendent spécifiquement de la forme (ou fonction de densité) de la distribution *f* d'origine.

Classe 2: tests sur l'indépendance statistique dans les séquences de nombres.

Classe 3: tests sur *l'équivalence des permutations.*

Classe 4: tests divers, basés sur des propriétés dérivées ou combinées à partir des propriétés de forme, indépendance ou équivalence permutationnelle.

Les tests des classes 2 et 3 s'appliquent à toute séquence de nombres, d'observations ou de mesures à propos de laquelle l'hypothèse du hasard est posée et ce, quelle que soit la distribution *f* d'origine.

Examinons à présent quelques-uns des tests de toutes catégories. Il est hors de question d'énumérer ici tous les tests applicables à notre question, voire tous les tests publiés. Les tests présentés constituent seulement un échantillon, espérons-le représentatif, des familles de tests possibles. Le lecteur trouvera une documentation partielle dans Gentle (1998), Kennedy et Gentle (1980), Knuth (1968), Lewis (1975), Marsaglia (1995) et Rubinstein (1981).

Tests sur la forme et les moments de la distribution

6.5 Des v.a., issues d'une source conforme à une loi de distribution *f* donnée, devraient présenter une distribution empirique et des moments statistiquement conformes à cette loi. Les tests du Khi-deux et de Kolmogorov-Smirnov, les deux principaux tests dits d'ajustement, permettent de vérifier l'hypothèse de la conformité de la distribution empirique. On peut évaluer aussi les moments centraux d'ordres 3 et 4, ayant trait à l'asymétrie et l'aplatissement (ou voussure) de la distribution. Enfin, une v.a. uniforme u, à distribution standard $U(0,1)$, devrait occuper tout intervalle (a, b), où $0 \leq a < b \leq 1$, selon la probabilité égale à $b - a$: la distribution des variables u_i dans un ensemble d'intervalles disjoints donne ainsi lieu à d'intéressants tests d'uniformité.

6.6 *Le test du Khi-deux.* Les n observations ou valeurs successives d'une v.a. $(x_1, x_2, ..., x_n)$ sont classées et réparties dans un ensemble de k intervalles ou catégories. Le test du Khi-deux consiste alors à comparer la série des fréquences (f_j) de valeurs observées dans les k classes de la variable à la série correspondante des *fréquences théoriques* (ft_i), chaque ft_j reflétant la probabilité qu'une observation d'une loi de distribution *f* donnée tombe dans la classe désignée. La statistique calculée :

$$X^2 = \sum_{j=1}^{k} \frac{(f_j - ft_j)^2}{ft_j} \qquad (6.1)$$

se distribue approximativement comme x^2 (4.17), avec v = k-1-p degrés de liberté; *p* représente le nombre de paramètres qu'on a dû estimer à partir des valeurs observées pour fixer la distribution stipulée et déterminer les ft_i. Une valeur significativement forte de X^2 indique que la séquence de nombres analysée déroge à la forme de la distribution invoquée.

L'approximation par la loi x^2 est robuste et assortie de conditions faciles à satisfaire (Kendall et Stuart 1979). La méthode utilisée pour définir les classes, ou intervalles, de la variable importe peu: la théorie, basée sur une structure de classes posée a priori, s'étend aussi à des classes a posteriori, voire à toute espèce de classes. La série des *k* fréquences $\{f_1; f_2; ...; f_k\}$ est comparée via (6.1) à la série théorique $\{ft_1; ft_2 ;...; ft_k\}$, *où* $f_1 = n \cdot p_f$ et $p_j$est la probabilité d'incidence dans la classe *j sous* la loi *f.* Les ft_j doivent avoir une moyenne d'au moins *4,* d'où $k \leq n/4$; de plus, dans chaque classe *j,* on recommande $ft_j \geq 1$. Dans le test d'un générateur de nombres pseudo-aléatoires, ou quantité de nombres sont disponibles à volonté, ces conditions peuvent être généreusement rencontrées.

Quant aux degrés de liberté de la loi x^2 approximative, $v=k- 1 -p,$ la perte de p unités a lieu lorsque les paramètres à estimer le sont en exploitant la série multinomiale $\{f_1; f_2; ...; f_k\}$. Si, comme il arrive, on dispose des *n* valeurs individuelles et que les paramètres sont estimés par la méthode habituelle du maximum de vraisemblance (par ex., $\hat{u} = x$ et $\hat{o} = s_x$), alors moins de p degrés de liberté sont perdus, et l'on $a\ k-1-p \leq v \leq$ k-1 (Kendall et Stuart 1979). Lorsque *k* n'est pas très grand, l'interprétation doit tenir compte de cette incertitude.

À moins que la variable donne lieu à un classement naturel, Kendall et Stuart (1979) recommandent de former généralement des classes *à probabilités égales, à* savoir de délimiter les classes de telle façon que, pour chacune, la probabilité d'incidence soit $p_j = 1/k$. Dans ce cas, la formule (6.1) se calcule plus simplement par :

$$X^2 = \frac{k}{n} \sum_{j=1}^{k} f_j^2 - n \, . \qquad (6.2)$$

Cette méthode entraîne ordinairement la création de classes à largeurs d'intervalles variables. Dans ce cas, le nombre de classes optimal, du point de vue de la puissance statistique, est proportionnel $à\ n^{2/5}$; la règle heuristique, $k \approx 2n^{2/5}$ peut servir de guide.

Les valeurs critiques du Khi-deux se retrouvent partout, notamment dans Laurencelle et Dupuis (2000). L'égalité approximative entre centiles $x^2_{v,p} = ''½[z_p + \sqrt{(2v - 1)}]^2$, où z_p est le centile l 00p de la loi normale standard, peut

***6.7** Le test de Kolmogorov-Smirnov.* Le test de Kolmogorov-Smirnov (K-S), à l'instar du Khi-deux, compare la distribution de la série statistique étudiée $\{x_i,\ i = 1,n\}$ à une distribution théorique *f;* dans ce cas, ce sont les fonctions de répartition (f.r.) correspondantes qui sont comparées, soit:

$$\text{la f.r. empirique } \hat{F}_n(x) = \#(x_i \le x) \tag{6.3}$$

et :

$$\text{la f.r. théorique } F(x) = \int^x F'\,dx\ ; \tag{6.4}$$

l'expression « $\#(x_i \le x)$ » désigne la fréquence cumulative de valeurs inférieures ou égales à x observées dans l'échantillon. La statistique de K-S, dénotée D_n, est alors:

$$D_n = \sqrt{n} \cdot \max_i |\hat{F}_n(x_i) - F(x_i)|\ , \tag{6.5}$$

soit la valeur maximale de la différence absolue entre les deux f.r., multipliée par $\sqrt{n}$.

La probabilité extrême de D_n est, asymptotiquement, $p_{ext} = 1 - 2 \cdot \exp(-2Dn2)$, ce qui fournit les valeurs critiques 1,358 et 1,628 pour les seuils a de 0,05 et 0,01 respectivement. Ces valeurs asymptotiques débordent les valeurs exactes D_n pour *n* fini; Rohlf et Sokal (1981) fournissent ces valeurs exactes pour $n \le 100$.

Au contraire du test du Khi-deux qui s'applique à un regroupement des données en *k* classes, le K-S exploite les *n* données individuelles et, pour cette raison, s'avère généralement plus puissant comme test d'ajustement. Pour l'effectuer, il convient d'abord de *trier* les données en ordre de valeurs croissantes puis, pour chaque i, de comparer $F[x_{(i]}$ *à i/n,* qui représente ici la f.r. empirique. Deux procédures de tri sont proposées en appendice du chapitre, à la suite des exercices.

La distribution de D_n, pour n fini ou infini, est connue et spécifiée, quelle que soit la loi f, dans le cas où la fr. de référence *(F)* est complètement déterminée. Dans plusieurs cas, toutefois, il faut estimer les paramètres de F (ou *f) :* ainsi, pour obtenir la f.r. associée à une loi normale, il nous faut des estimations de la moyenne p, et de la variance 0^2 basées sur nos n données. L'effet de ces estimations sur le test de K-S est de le rendre conservateur, *i.e.* moins sensible aux dérogations de la série observée, étant donné que, suite à l'estimation des paramètres, la f.r. théorique se trouve artificiellement rapprochée de la f.r. empirique.

Lilliefors (1967), pour le cas d'une loi normale où l'on exploite les statistiques $\bar{x}$ et s_x pour estimer μ et σ, fournit les valeurs critiques

asymptotiques D_n = 0,886 et 1,031 pour les seuils de 0,05 et 0,01, respectivement, en plus de quelques valeurs pour petits n (< 30). Le présent auteur a étudié la question de l'ajustement à la loi normale en se basant sur deux estimateurs robustes des paramètres, soit la médiane Md (pour μ) et l'écart-médian EM = Md $|x_i$ - Md$|$ /0,67449 (pour 0). Des valeurs critiques Monte Carlo basées sur 10 000 échantillons normaux de n = 9 à 169 furent obtenues, donnant lieu aux fonctions estimatives suivantes (tous les $r^2 > 0{,}999$):

$$D_n(0{,}95) = 1{,}0428 + 0{,}04240\log_e n$$
$$D_n(0{,}975) = 1{,}1715 + 0{,}04147\log_e n$$
$$D_n(0{,}99) = 1{,}3074 + 0{,}04455\log_e n$$
$$D_n(0{,}995) = 1{,}3903 + 0{,}05001\log_e n$$

Si la série à tester n'est pas normalement distribuée, les estimateurs proposés (Md, EM) semblent préférables aux estimateurs usuels ($\bar{x}$ et s_x), peu robustes, et qui sont optimaux seulement lorsque la distribution sous-jacente est normale.

6.8 *Les tests sur les moments.* Toute loi de probabilité est complètement caractérisée par ses moments; au-delà de la moyenne (μ = μ'1) et de la variance ($0^2 = \mu_2$), on recourt volontiers aux moments centraux μ_3 et μ_4 et, en particulier, à leurs valeurs standardisées:

$$\gamma_1 = \mu_3 / \sigma^3 \; ; \qquad (6.6a)$$
$$\gamma_2 = \mu_4 / \sigma^4 - 3 \; . \qquad (6.6b)$$

Ces deux statistiques, aussi désignées « indices de forme », servent à caractériser sommairement la forme de la distribution; y_1 (« gamma 1 ») est un indice d'asymétrie, y_2 (« gamma 2 ») un indice d'aplatissement, ou de voussure. Les estimateurs empiriques des moments μ, 0^2, y_1 et y_2 sont:

$$\bar{x} = \sum x_i / n \qquad (6.7)$$
$$s^2 = \sum(x_i - \bar{x})^2 / (n-1) \qquad (6.8)$$
$$g_1 = \sqrt{n}\sum(x_i - \bar{x})^3 / [\sum(x_i - \bar{x})^2]^{3/2} \qquad (6.9)$$
$$g_2 = n \sum(x_i - \bar{x})^4 / [\sum(x_i - \bar{x})^2]^2 - 3 \; . \qquad (6.10)$$

Les estimateurs $\bar{x}$ et s^2 sont généralement sans biais, *i.e.* $E(\bar{x}) = \mu$ et $E(s^2) = \sigma^2$, ce qui n'est pas le cas pour les estimateurs g_1 et g_2 (Kendall et Stuart 1977).

Toute distribution symétrique a des moments centraux impairs nuls, en particulier $\gamma_1 = 0$; c'est le cas des lois uniforme et normale. La loi

normale a, par définition, un indice d'aplatissement nul, $y_2 = 0$, alors que la loi uniforme est platykurtique, avec $y_2 = -1{,}2$.

Pour des v.a. normales x, de distribution $N(\mu,\sigma^2)$, il est bien connu que la distribution de la moyenne $\bar{x}$ est une $N(\mu,\sigma^2/n)$, celle de $(n-1)s^2/\sigma^2$, une χ^2_{n-1}. Quant aux v.a. uniforme, de distribution $U(0,1)$, leur moyenne arithmétique $\bar{u}$ a pour moments $\mu = ½$, $\sigma^2 = 1/12n$, $\gamma_1 = 0$ et $\gamma_2 = -6/5n$ (Laurencelle 1993); la distribution de $\bar{u}$, assez complexe, peut s'approcher par une normale dès que γ_2 s'amenuise, disons pour $n > 15$. On connaît aussi (Laurencelle 1993) les moments de la variance de données uniformes, ce qui permet d'approcher sa distribution par une loi χ^2, de paramètre ν approprié.

Les distributions échantillonnales de g_1 et g_2 ne sont connues pour aucune population statistique. Différents efforts ont été publiés pour tenter d'approcher ces distributions pour une population normale (Kendall et Stuart 1977; Mardia 1980); les résultats les plus sûrs, du moins pour petits échantillons, proviennent néanmoins de données Monte Carlo.

Pour une population normale, les moments des statistiques g_1 et g_2 sont connus. Notamment, g_1 est symétrique, d'où $\mu(g_1) = \gamma_1(g_1) = 0$; sa variance est $\sigma^2(g_1) = 6(n-2)/[(n+1)(n+3)]$, et sa voussure $\gamma_2(g_1) = 36(n^3-5n^2-19n+35)/[(n+9)(n+7)(n+5)(n-2)]$. Ces données suggèrent entre autres d'approcher la distribution de g_1 par un t_ν, dont la variance est $\sigma^2(t_\nu) = \nu/(\nu-2)$ et la voussure $\gamma_2(t_\nu) = 6/(\nu-4)$. Quant à g_2, il a pour espérance $\mu(g_2) = -6/(n+1)$, pour variance $\sigma^2(g_2) = 24n(n-2)(n-3)/[(n+3)(n+5)(n+1)^2]$; il est positivement asymétrique, avec $\gamma_1(g_2) = \{216/n\,[1-29/n+519/n^2-7637/n^3+\ldots]\}^{½}$, et leptokurtique, avec $\gamma_2(g_2) = 540/n-20196/n^2+470412/n^3-\ldots$

Prenons le cas d'un échantillon de n = 50 données normales. La voussure de g_1 est alors ≈ 0,452; pour ce degré de voussure, le t correspondant aurait $\nu = 17{,}3 \approx 17$, avec un écart-type de $\sqrt{(17/15)} \approx 1{,}0646$. Les percentiles 95 et 99 de ce t_{17} sont 1,740 et 2,567. La conversion par $t0_{g1}/0$ propose donc 0,533 et 0,787 pour valeurs critiques. Des données de simulation, basées sur 99 999 échantillons normaux, indiquent les valeurs 0,532 et 0,788; les tables *Biometrika* (Pearson et Hartley 1970) fournissent quant à elles 0,534 et 0,787.

Des valeurs critiques approximatives pour g_1 et g_2, relatives à des populations normales, apparaissent dans quelques publications (D'Agostino et Pearson 1973; D'Agostino et Tietjen 1973; Pearson et Hartley 1970).

L'absence de toute donnée pour une population uniforme a motivé une étude des indices g_1 et g_2 pour des échantillons de cette population. Les centiles 1, 5, 95 et 99 ont été estimés à partir de la distribution de 99999 échantillons Monte Carlo, pour des tailles n allant de 10 à 1000. Outre g_1

et g_2, les estimateurs $\hat{Y}_1$ et $\hat{Y}_z$, exploitant les valeurs connues des paramètres μ et 0^2, sont obtenus selon:

$$\hat{\gamma}_1 = \Sigma(x_i - \mu)^3 / n\sigma^3 \qquad (6.11)$$

$$\hat{\gamma}_2 = \Sigma(x_i - \mu)^4 / n\sigma^4 - 3 \; . \qquad (6.12)$$

Des fonctions prédictives, établies sur la base des estimations de centiles Monte Carlo, ont permis d'établir les valeurs critiques présentées aux tableaux 6.1 et 6.2.

6.9 *Les tests* d'occupation. Il existe une autre famille de tests, celle des tests *d'occupation.* Ces tests sont applicables pour des variables à distribution uniforme; on peut les adapter aussi à toute distribution. Dans un premier temps, on découpe l'étendue de la distribution en *k* intervalles consécutifs et de probabilité égale; dans le cas d'une loi non bornée, il faudra garder un ou deux intervalles ouverts. Dans un deuxième temps, chaque nouvelle donnée est placée dans l'intervalle approprié: la statistique résultante x est le nombre d'intervalles, ou cases, occupés, ou bien le nombre v de cases vides, selon x + $v = k$, ou encore le nombre n de données requis pour occuper x cases. La distribution de v_n, le nombre de cases encore vides après l'inscription de *n* données, est bien connue (Johnson, Kotz et Kemp 1992), soit:

$$\mathrm{pr}(v_n) = \binom{k}{v_n} \sum_{0}^{k-v_n} (-1)^i \binom{k-v_n}{i} \left(\frac{k-v_n-i}{k}\right)^n \; ; \qquad (6.13)$$

pour *n* fort, une bonne approximation de $\mathrm{pr}(v_n)$ s'obtient par:

$$\mathrm{pr}(v_n) \approx \binom{k}{v_n} (1-e^{-n/k})^k \left(\frac{e^{-n/k}}{1-e^{-n/k}}\right)^{v_n} \; . \qquad (6.14)$$

Cette distribution a plusieurs applications; elle permet notamment de détecter des anomalies locales dans le domaine du générateur, par exemple une paresse du générateur à fournir des valeurs à la limite supérieure de l'étendue.

L'espérance et la variance de v_n, le nombre de cases vides après *n* inscriptions parmi les k cases de l'étendue, sont:

$$\mu(v_n) = k(1-1/k)^n \qquad (6.15a)$$

$$\sigma^2(v_n) = \mu(1-\mu) + k(k-1)(1-2/k)^n \; . \qquad (6.15b)$$

Tableau 6.1 Valeurs critiques des indices $\hat{\gamma}_1$ et $\hat{\gamma}_2$ pour des v.a. uniformes

	$\hat{\gamma}_1$				$\hat{\gamma}_2$			
n	$P = 0{,}01$	0,05	0,95	0,99	$P = 0{,}01$	0,05	0,95	0,99
10	−1,445	−1,022	1,022	1,445	−2,678	−2,351	0,150	0,790
20	−1,022	−0,722	0,722	1,022	−2,310	−2,037	−0,269	0,160
30	−0,834	−0,590	0,590	0,834	−2,132	−1,892	−0,448	−0,107
40	−0,723	−0,511	0,511	0,723	−2,020	−1,804	−0,553	−0,262
50	−0,646	−0,457	0,457	0,646	−1,942	−1,743	−0,623	−0,366
75	−0,528	−0,373	0,373	0,528	−1,816	−1,646	−0,732	−0,526
100	−0,457	−0,323	0,323	0,457	−1,738	−1,588	−0,796	−0,620
150	−0,373	−0,264	0,264	0,373	−1,644	−1,518	−0,872	−0,729
200	−0,323	−0,228	0,228	0,323	−1,586	−1,476	−0,916	−0,794
250	−0,289	−0,204	0,204	0,289	−1,546	−1,447	−0,947	−0,838
300	−0,264	−0,187	0,187	0,264	−1,517	−1,426	−0,969	−0,870
400	−0,229	−0,162	0,162	0,229	−1,475	−1,396	−1,000	−0,915
500	−0,204	−0,144	0,144	0,204	−1,446	−1,375	−1,022	−0,946
750	−0,167	−0,118	0,118	0,167	−1,401	−1,343	−1,055	−0,993
1000	−0,145	−0,102	0,102	0,145	−1,375	−1,324	−1,074	−1,021
∞[1]	−4,570	−3,231	3,231	4,570	−5,524	−3,923	3,960	5,585

(1) Les valeurs critiques C_n s'obtiennent à partir des valeurs asymptotiques C_∞ selon $C_n \approx C_\infty/\sqrt{n}$ pour $\hat{\gamma}_1$ et $C_n \approx C_\infty/\sqrt{n - 1{,}2}$ pour $\hat{\gamma}_2$.

Tableau 6.2 Valeurs critiques des indices g_1 et g_2 pour des v.a. uniformes

	g_1				g_2			
n	$P = 0{,}01$	0,05	0,95	0,99	$P = 0{,}01$	0,05	0,95	0,99
10	−1,135	−0,768	0,769	1,136	−1,747	−1,625	0,039	1,065
20	−0,778	−0,539	0,538	0,775	−1,623	−1,518	−0,464	0,023
30	−0,628	−0,438	0,437	0,625	−1,562	−1,468	−0,650	−0,325
40	−0,541	−0,379	0,377	0,537	−1,523	−1,437	−0,749	−0,502
50	−0,482	−0,338	0,336	0,479	−1,495	−1,415	−0,813	−0,611
75	−0,391	−0,275	0,274	0,389	−1,450	−1,381	−0,904	−0,761
100	−0,338	−0,238	0,237	0,335	−1,422	−1,359	−0,954	−0,840
150	−0,275	−0,194	0,193	0,273	−1,387	−1,333	−1,009	−0,925
200	−0,238	−0,167	0,167	0,236	−1,365	−1,317	−1,040	−0,970
250	−0,212	−0,150	0,149	0,211	−1,349	−1,306	−1,060	−1,000
300	−0,194	−0,136	0,136	0,192	−1,337	−1,298	−1,075	−1,021
400	−0,168	−0,118	0,117	0,167	−1,321	−1,286	−1,094	−1,049
500	−0,150	−0,105	0,105	0,149	−1,309	−1,278	−1,107	−1,067
750	−0,122	−0,086	0,086	0,121	−1,290	−1,265	−1,126	−1,094
1000	−0,106	−0,074	0,074	0,105	−1,279	−1,257	−1,137	−1,110
∞[1]	−3,343	−2,312	2,332	3,322	−2,568	−1,962	1,782	2,660

(1) Les valeurs critiques C_n s'obtiennent à partir des valeurs asymptotiques C_∞ selon $C_n \approx C_\infty/\sqrt{n}$ pour g_1 et $C_n \approx C_\infty/\sqrt{n - 1{,}2}$ pour g_2.

Quant au nombre *n* d'inscriptions (ou de données) qu'il faut pour occuper les k cases, c'est une variable géométrique composée Çà paramètre décroissant), et sa fonction de répartition correspond à (6.13) avec v_n= O. L'espérance de *n* est:

$$\mu(n) = kH_k\,, \quad H_k = 1 + \tfrac{1}{2} + \ldots + \tfrac{1}{k}\,, \tag{6.16}$$

et sa variance:

$$\sigma^2(n) = k\sum_{r=1}^{k-1} \frac{r}{(k-r)^2}\ . \tag{6.17}$$

On peut approcher l'espérance par k[$\log_e k$ + *0,5772 + 1/(2k)]* et la variance par *k(k-3)* л ^{2}*/6.* Quant au nombre de données requis pour qu'il ne reste que v cases vides, son espérance est:

$$\mu(n\,|\,v) = k(H_k - H_v) \approx k[\log_e(k/v) - (k-v)/(2kv)]\ . \tag{6.18}$$

6.10 Le premier test issu de la distribution d'occupation (6.13) concerne le nombre de données nécessaire pour occuper tous les intervalles du domaine, par exemple les *k* intervalles de longueur *1/k* pour une variable uniforme $U(0,1)^2$. Le tableau 6.3 fournit les nombres critiques *(n*)* de données aux seuils approximatifs a = 0,05 et 0,01, pour différentes segmentations (k) du domaine; si, pour une série de v.a. donnée, le nombre *n* requis pour combler les *k* intervalles est tel que $n \geq n^*(a)$, l'hypothèse d'une distribution statistique uniforme peut être mise en doute. Noter que, pour *k* fort, $n^* \approx k\log_e(k/a)$.

À défaut d'exiger l'occupation des *k* cellules, la paresse sélective du générateur de nombres (ou de la variable) peut être vérifiée après l'occupation de seulement *r* des *k* cases, où $r \leq k$-1. Le nombre d'essais requis pour occuper l'une des (k-r) cases restantes, disons n_r, est une variable géométrique simple (cf. §4.5), de moyenne *k/(k-r)* et de variance $rk/(k-r)^2$; le nombre critique au seuil *a* est simplement $n_r(a) = [\log_e a/\log_e(r/k)]$.

Un autre test de l'uniformité de la distribution consiste à repérer le numéro (N_r) du r^e intervalle occupé, quel que soit le nombre d'inscriptions

---

1. La loi de probabilité, quant à elle, est $\mathrm{pr}(n) = \sum_{i=0}^{k-1} (-1)^i \binom{k-1}{i} \left(\frac{k-1-i}{k}\right)^{n-1}$.

2. Le « test du collectionneur » est une généralisation de ce test, appliqué à plusieurs séries aléatoires et destiné à juger l'irrégularité globale de la source de données: voir §6.26.

Tableau 6.3 Nombre critique de données pour occuper les k intervalles équiprobables d'un domaine

k	$p \le 0{,}05$	$p \le 0{,}01$	k	$p \le 0{,}05$	$p \le 0{,}01$
2	6	8	20	117	149
3	11	15	25	152	192
4	16	21	30	189	237
5	21	28	35	226	282
6	27	36	40	264	328
7	33	43	45	302	375
8	38	51	50	341	422
9	44	58	60	421	518
10	51	66	70	502	616
12	63	82	80	585	715
14	76	98	90	669	815
16	90	115	100	754	916
18	103	132	> 100	$< k\log_e(20k)$	$< k\log_e(100k)$

requis pour combler les k intervalles est tel que $n \ge n^*(a)$, l'hypothèse d'une distribution statistique uniforme peut être mise en doute. Noter que, pour k fort, $n^* \approx k\log_e(k/a)$.

A défaut d'exiger l'occupation des k cellules, la paresse sélective du générateur de nombres (ou de la variable) peut être vérifiée après l'occupation de seulement r des k cases, *où* $r \le k$-1. Le nombre d'essais requis pour occuper l'une des (k-r) cases restantes, disons n_r, est une variable géométrique simple (cf §4.5), de moyenne $k/(k-r)$ et de variance $rk/(k-r)^2$; le nombre critique au seuil a est simplement $n_r(a) = [\log_e a/\log_e(r/k)$.

Un autre test de l'uniformité de la distribution consiste à repérer le numéro (N_r) du r^e intervalle occupé, quel que soit le nombre d'inscriptions requis pour y parvenir. Un générateur ou une variable sans biais tomberont équitablement dans tous les intervalles. D'un autre côté, pour autant que r approche k, une paresse sélective de la variable devrait se manifester par une hésitation à occuper certains intervalles. Répéter l'expérience T fois et noter N_r à chaque fois nous mettent en mesure d'appliquer le test du Khi- deux (formule (6.2)) sur l'équipartition des k numéros de cellules. De nombreux autres tests, exploitant la théorie donnée en §6.9, peuvent aussi être mis en œuvre.

La tombée des n données d'une série dans les k intervalles équiprobables constitue une réalisation de la loi multinomiale en regard de laquelle des tests locaux et globaux sont aussi possibles. Les exercices 6.5 à 6.8 parcourent la question.

6.11 *Les tests de normalité.* Les procédures générales de test examinées plus haut conviennent aussi pour vérifier l'hypothèse d'une distribution normale ou de toute autre forme de distribution. Étant donné l'importance de la distribution normale en statistique appliquée, les auteurs ont confectionné plusieurs tests spécifiques à celle-ci, les plus célèbres étant le W de Shapiro et Wilk (1965) et de D de D'Agostino (1971).

Les études comparatives de puissance (Shapiro, Wilk et Chen 1968; Mardia 1980) accordent un certain mérite au test W. Ce test est basé sur les statistiques d'ordre $x_{(i)}$ de l'échantillon, et on le calcule par:

$$W = \frac{\left\{ \sum_{i=1}^{\lfloor n/2 \rfloor} a_i \left(x_{(n+1-i)} - x_{(i)} \right) \right\}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} ; \qquad (6.19)$$

comme on le voit, il s'agit d'un quotient d'estimateurs de la variance. Les cœfficients a_i peuvent être approchés de diverses manières (Shapiro et Wilk 1965). Les tables de *Biometrika* (Pearson et Hartley 1970) les fournissent, de même que des valeurs critiques de W (voir aussi Royston 1982 et l'exercice 6.9).

Exemple 6.1 Tests de normalité de 60 longueurs de ficelle

Les données. L'exemple suivant illustre quelques tests sur la forme distributionnelle, parmi ceux proposés. Le département d'expédition d'une manufacture utilise de la ficelle pour emballer les colis. Afin d'étudier la consommation de ficelle par les emballeurs, le gérant échantillonne et défait 60 colis apprêtés et il en mesure les bouts de ficelle, en centimètres. Voici les données:

45 65 68 55 84 51 57 52 50 55 74 48 94 48 58 48 51 60 56 81
53 64 67 48 54 48 62 53 63 54 65 54 53 60 56 68 61 56 57 66
47 82 57 70 46 51 60 58 61 58 64 66 52 65 56 52 56 77 55 79

L'hypothèse. Le gérant suppose que les longueurs de ficelle se distribuent normalement. Il veut donc tester cette hypothèse, en plus d'obtenir la moyenne et l'écart-type.

Test de g_1 *et* g_2. Les quatre premiers moments à l'origine, $m_k = \sum x^k/n$, sont ici $m_1 = 59{,}5\overline{6}$, $m_2 = 3652{,}4\overline{6}$, $m_3 = 231236{,}2\overline{6}$ et $m_4 = 15153233{,}4\overline{6}$. Le moment central μ_2 s'estime par $m_2 - m_1^2 \approx 104{,}279$, μ_3 par $m_3 - 3m_2m_1 + 2m_1^3 \approx 1247{,}910$, et μ_4 par $m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^3 \approx 46254{,}991$. Les estimations de moyenne ($\hat{\mu}$) et d'écart-type ($\hat{\sigma}$) sont donc $\bar{x} = m_1 \approx 59{,}567$ et $s_x = \sqrt{\hat{\mu}_2} \times \sqrt{[n/(n-1)]} \approx 10{,}298$. Les formules (6.9) et (6.10) peuvent s'écrire aussi:

$$g_1 = \hat{\mu}_3 / \hat{\mu}_2^{3/2} \quad ; \quad g_2 = \hat{\mu}_4 / \hat{\mu}_2^2 - 3 \quad , \qquad (6.20)$$

et nous obtenons $g_1 \approx 1{,}172$ et $g_2 \approx 1{,}254$: la forme de la distribution serait allongée vers la droite et plus pointue que la normale (ou leptokurtique). Les tables de *Biometrika* (Pearson et Hartley 1970), pour $n = 60$ et au percentile 0,99, indiquent les valeurs critiques 0,723 pour g_1 et environ 1,73 pour g_2: l'asymétrie apparaît trop forte pour croire que les données proviennent d'une loi normale. Des tests approximatifs, qui exploitent les estimateurs grossiers $0^2(g_1) \approx 6/n$ et $0^2(g_2) \approx 24/n$ afin de fabriquer des écarts-réduits pour g_1 et g_2, interprétés comme des v.a. N(0,1), aboutissent aux mêmes conclusions.

Test du Khi-deux. Pour $n = 60$ données, la règle suggérée pour fixer le nombre de classes en propose $k = 11$ ($k \approx 2n^{2/5} = 10{,}29$). Les bornes centiles de ces classes sont de $P = \frac{1}{2} \pm (2j - 1)/(2k)$; ainsi, la classe centrale (j=1) occupe la zone d'intégrale (0,4545; 0,5455), correspondant aux écarts-réduits normaux $z = \pm 0{,}1142$ et à l'intervalle en x (58,4; 60,7), selon $\hat{u} = 59{,}567$ et $\hat{o} = 10{,}298$. La classe supérieure (j = 5) est bornée inférieurement par le percentile 0,9091, d'où $z(0{,}9091) \approx 1{,}3352$ et $x \approx 73{,}3$. La répartition des 60 données dans ces 11 classes produit les fréquences (de bas en haut): 1, 8, 9, 6, 11, 3, 4, 5, 5, 1, 7, toutes associées à une fréquence théorique commune, $n/k = 5{,}45$. Par la formule (6.2), $X^2 = 18{,}467$; enfin les degrés de liberté, $v =$ k-1 -p, se situent entre 8 et 10. Au seuil a = 0,01, $x^2(8;0{,}99) = 20{,}09$ et $x^2(10;0{,}99) = 23{,}21$; ainsi, quel que soit v, le test du X^2 ne rejette pas l'hypothèse spécifiant que les données se distribuent normalement.

Test de Kolmogorov-Smirnov. Afin d'appliquer le test d'ajustement de Kolmogorov-Smirnov, les estimations $\hat{\mu} = 59{,}567$ et $\hat{\sigma} = 10{,}298$ sont utilisées pour convertir chaque donnée x_i en écart-réduit z_i, puis trouver l'intégrale normale correspondante, selon $F(x_i) = \Phi(z_i)$. La plus faible donnée, 45, a pour écart-réduit $-1{,}4145$ et $F(45) = \Phi(-1{,}4145) \approx 0{,}079$; l'intégrale empirique correspondante est $\hat{F}_{60}(45) = 1/60 \approx 0{,}017$; etc. La différence absolue maximale observée est 0,1438, et $D = 0{,}1438\sqrt{60} = 1{,}1139$. Selon le critère de Lilliefors,

cette valeur est significative à a = 0,01; elle ne l'est pas selon le critère habituel du K-S. En utilisant les estimateurs Md=57 et EM ≈ 8,896 pour μet O, la procédure du présent auteur s'applique, et nous obtenons D = 0,6907, la valeur critique à a = 0,05 étant 1,0428+0,04240log$_e$60 ≈1,2164; ce critère rejoint le critère habituel, la conclusion étant la conservation de l'hypothèse normale.

Test W. Notre calcul de la statistique de Shapiro et Wilk suit les indications de Royston (1982); l'exercice 6.9 en donne le détail. Le dénominateur de *W* est simplement $nû_2 = (n-1)s^2 \approx$ 6256,73; pour le numérateur, de longs calculs donnent 5669,19. Le quotient *W*, toujours inférieur à l'unité, est de 0,9061. Le test normal est approché en utilisant ξ = 0,213685, μ = 0,452, O = 0,04041, et z = [(1 - W)^{ξ} - μ]/O = [(1 - 0,9061)$^{0,2\ 3685}$ — 0,452]/0,04041≈ 3,742, un résultat fortement significatif (a < 0,001). Ce test, à l'instar du test de l'indice g_1 sur l'asymétrie, découvre lui aussi un aspect non normal significatif de la distribution étudiée.

L'écart à la forme de distribution normale est-il sérieux et réel? Oui, nous le croyons. Des tests positifs obtenus, soit la forte asymétrie positive, le *W* de Shapiro et Wilk, voire le test de Kolmogorov-Smirnov au critère de Lillifors, chacun suffit à nous convaincre.

Tests sur l'indépendance des valeurs successives

6.12 Par-delà la forme de distribution des variables aléatoires, l'idée de « nombres au hasard » fait d'abord référence à l'irrégularité plus ou moins grande de leurs valeurs successives. à leur imprévisibilité. Cette « irrégularité » ne peut pas être définie: c'est une qualité négative, un caractère dénotant l'absence de qualités observables, et qui résulte de la propriété *d'indépendance* d'une variable aléatoire.

L'indépendance statistique, une notion d'abord probabiliste, peut présenter des degrés. L'indépendance est totale, et l'association corrélative nulle, lorsque la probabilité d'apparition conjointe de deux ou plusieurs valeurs de la variable égale le produit de leurs probabilités respectives, soit:

$$p(X_1, X_2, \cdots, X_n) = p(X_1)p(X_2) \cdots p(X_n)\ ; \qquad (6.21)$$

dans ce cas, en effet, chaque variable a une distribution ne dépendant d'aucune autre, puisque par exemple, $p(X_1 \mid X_2, X_3, \cdots, X_n) = p(X_1)$. De plus, il suit que toutes les permutations de l'ensemble $\{x_1, x_2, \cdots, x_n\}$ sont

équivalentes en probabilité. Dans les cas où l'égalité (6.21) n'est pas respectée, l'indépendance est imparfaite, et certaines régularités s'installent. A la limite, si par exemple $p(X_1=x_1 \mid X_2=x_2) \rightarrow 1$ ou $p(X_1=x_1 \mid X_2=x_2) \rightarrow 0$, on observera des séries à forte corrélation sérielle, positive ou négative selon le cas. Les tests d'hypothèses qu'on applique aux séquences de valeurs, dans le présent cadre d'étude de la méthodologie Monte Carlo, reposent tous sur le postulat de l'indépendance statistique et en vérifient l'actualité.

Il existe différentes stratégies afin de vérifier jusqu'à quel point une ou plusieurs de nos séries de nombres contreviennent au postulat d'indépendance. Nous avons arbitrairement classé les tests en trois groupes, ceux qui traitent explicitement d'indépendance (des valeurs successives), ceux qui relèvent de l'équiprobabilité des permutations et, enfin, quelques autres tests mélangés; c'est dans ce dernier groupe qu'on retrouvera les tests sur l'irrégularité globale (plutôt que locale) d'une source de nombres aléatoires.

6.13 *Le test sériel.* Sans doute le plus connu des tests d'indépendance séquentielle, le test sériel exploite directement l'équation (6.21). Dans le cas d'une v.a. uniforme u ~ $U(0,1)$, le domaine (0,1) est divisé en r intervalles égaux, chacun de capacité 1/r. Soit, alors, un t-tuplet $\{u_1, u_2, ..., u_t\}$ formé de t v.a.u. successives. Chaque composante u_i est repérée dans l'intervalle approprié $c_i.(u_i)$; on localise la cellule appropriée $(c_1, c_2,..., c_t)$ dans l'hypercube de dimension, ou tableau $[r]^t$, et on l'y enregistre. On procède ainsi répétitivement, en utilisant à chaque fois un nouveau t-tuplet, pour un total de N inscriptions. Sous l'hypothèse d'indépendance, l'égalité (6.21) s'applique, et les N t-tuplets inscrits devraient se répartir de manière homogène dans le tableau $[r]^t$. Le test du Khi-deux permet de vérifier ce résultat, en utilisant une fréquence théorique uniforme de N/r^t. Par la formule (6.2), il y a $k = r^t$ cellules à sommer et le x^2 présente $v = k$—1 degrés de liberté. Ce test peut s'appliquer aussi à des v.a. d'une quelconque forme distributionnelle ou en utilisant des intervalles de capacités inégales avec la formule générale (6.1) du Khi-deux.

> L'indépendance statistique stipulée par (6.21) désigne des multiplets de toutes grandeurs, en fait des t-tuplets pour $t \rightarrow \infty$. En pratique, la faisabilité du test X^2 impose un plafond sévère à la taille t du multiplet. Il est recommandé de viser une *fréquence théorique* $ft \geq 5$. Comme il faut t variables pour constituer un t-tuplet et qu'il y a r^t cellules à emplir, le nombre total de v.a. à produire est de l'ordre de $n = 5tr^r$: par exemple, pour r = 10 intervalles et des séquences de longueur $t = 4$, $n \sim 200\,000$. Ce « coût » très élevé, voire prohibitif, du test sériel a motivé la recherche d'autres tests, à consommation plus raisonnable.

Le lecteur astucieux pourra être tenté d'envisager l'enchaînement des multiplets successifs, selon le modèle: $\{x_1,x_2,...,x_t\}$, $\{x_2,x_3, ..., x_t,x_{t+1}\}$, ainsi de suite, en prolongeant le tout N fois. Le coût, en nombre de variables requis, s'en trouve réduit d'un facteur t, toutefois le test résultant n'a plus la distribution x^2 avec $v = r^t-1$, ce qui grève sérieusement l'avantage de cet expédient.

6.14 *Les corrélations sérielles.* Si les variables sont statistiquement indépendantes, la corrélation paramétrique d'une valeur à l'autre est nulle, et les corrélations mutuelles des valeurs d'une série devraient être nulles, en particulier la corrélation entre les valeurs successives. Soit $r = corr(x_i,x_j)$, un coefficient établi selon un procédé ou l'autre à partir de n données, l'espérance est zéro et la variance, $1/(n-1)$; ainsi, pour n modéré ($n \geq 30$, disons), $r\sqrt{(n-1)}$ se distribue à peu près comme une normale $N(0,1)$[3].

Il existe au moins trois procédés de calcul du coefficient de corrélation sérielle. Le plus conventionnel consiste à regrouper n' données en $n = \frac{1}{2}n''$ paires, soit (x_1,x_2), (x_3,x_4), ..., $(x_{n'-1},x_{n'})$, puis à appliquer un calcul de corrélation classique (r) à ces paires de données. Sous l'hypothèse d'indépendance séquentielle (impliquant $p = 0$) et si les valeurs x_i sont normales, la quantité $r\sqrt{(n-2)}/\sqrt{(1-r^2)}$ se distribue exactement comme t_{n-2}, une v.a. t de Student avec $v = n - 2$ degrés de liberté; pour la plupart des distributions, ces prescriptions donnent une approximation satisfaisante.

Une autre procédure de corrélation consiste à enchaîner les paires successives, en bouclant sur la n^e valeur, selon le modèle: (x_1,x_2), (x_2,x_3), ..., (x_{n-1},x_n), (x_n,x_1); cette procédure établit un coefficient R basé sur n paires formées de n données. Knuth (1969) rapporte que, pour des v.a. normales indépendantes ($p = 0$), l'espérance de R est $-1/(n-1)$ et sa variance, $[n(n-3)]/[(n+1)(n-1)^2]$. L'auteur indique que ces quantités s'appliquent aussi, au moins approximativement, à des v.a. uniformes et que la distribution de R est asymptotiquement normale. Anderson (1942) donne la distribution exacte pour x normal: l'exercice 6.13 fournit des valeurs critiques. Johnson, Kotz et Balakrishnan (1994, 1995) indiquent que la variable transformée $z \leftarrow \sin^{-1}(R)$, en radians, a une distribution approximativement normale, d'espérance 0 et de variance $(n-1)/n^2$.

Enfin, le troisième procédé de calcul consiste à enchaîner les paires successives, cette fois sans boucler: à partir des n données, on forme les $n-1$ paires (x_1,x_2), (x_2,x_3), ..., (x_{n-1},x_n), grâce auxquelles on évalue un coefficient R'. Pour $p = 0$, l'espérance de zéro s'associe à une variance approximative de $(n-1)/n^2$, plus précisément $[n(n-1)+5]/n^3$. Johnson et

3. Pour toutes v.a. X, Y indépendantes et possédant des variances, $E(r_{x,y}) = 0$ et $var(r_{x\,y}) = 1/(n-1)$ (Denis Allaire, Université de Sherbrooke, 1999, communication personnelle).

coll. (1994, 1995) rapportent une légère platykurtose, de $n = 10$ ($\gamma_2 \approx$-0,21) jusqu'à n = 500 ($\gamma_2 \approx$-0,01); la distribution étant symétrique, l'utilisation d'une loi *Bêta* symétrique comme modèle de densité se suggère d'elle-même.

Les études Monte Carlo montrent que l'approximation normale suggérée pour les deux formes de corrélations sérielles enchaînées fonctionne bien dans le cas de v.a. uniformes; les données indiquent cependant une tendance à l'asymétrie négative dans les distributions échantillonnales de R et R' pour $n < 50$.

6.15 *La variance d'une moyenne.* Soit la moyenne arithmétique $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$. La théorie statistique pose, d'une part, que $E(\bar{x}) = E(x_i) = \mu$ et, d'autre part, que $\text{var}(\bar{x}) = \sigma^2/n$ si $\text{var}(x_i) = \sigma^2$ *et* si les v.a. x_i sont mutuellement indépendantes. Le comportement de var($\bar{x}$), ou de sa forme standardisée, $n \cdot \text{var}(\bar{x})$, nous renseigne donc sur l'indépendance des v.a. utilisées. Pour le cas de v.a. normales à distribution $N(\mu,\sigma^2)$, il est évident que la quantité $n(k-1)s^2(\bar{x}_1, \bar{x}_2, .., \bar{x}_k)/\sigma^2$ est une v.a. du Khi-deux à $\nu = k-1$ degrés de liberté; diverses formes de test découlent de ce fait. Dans le cas de v.a. uniformes, seul Laurencelle (1993) rapporte quelques résultats.

> Laurencelle (1993) étudie la « variance centrée » d'une moyenne de n v.a. uniformes; il s'agit en fait de la statistique $v = (\bar{u}-½)^2$, analogue à l'erreur quadratique moyenne (EQM). Il établit la densité et la f.r. pour $n = 1, 2, 3$ et 4. Il étudie aussi une statistique $w = v, + v$, pour $n = 1$ et 2. Les statistiques v et w tendent vers x^2_i et x^2_2respectivement, à des constantes de proportionnalité près.

Pour effectuer un test inspiré de ce principe, il faut constituer, disons, k séries contenant chacune n variables, trouver ensuite la moyenne ($\bar{x}_j$) dans chaque série, puis calculer $\text{var}(\bar{x}_j) = \sum(\bar{x}_j - \bar{x}_G)^2 \div (k-1)$, où $\bar{x}_G = \sum \bar{x}_j / k$; le test résultant constitue *ipso facto* un test d'irrégularité globale. Toutefois, la nécessité de connaître la variance paramétrique σ^2 pourra en restreindre l'application.

6.16 *La variance permutative.* La variance permutative (Laurencelle 1983) est une statistique de variance basée sur la différence entre valeurs consécutives d'une série. Sa formule est:

$$s_p^2 = \frac{\sum_{i=1}^{n-1} (x_i - x_{i+1})^2}{2(n-1)} . \qquad (6.22)$$

L'espérance de s_p^2 est σ^2, où σ^2 est la variance paramétrique des x_i. Quant à la variance de s_p^2, elle est donnée généralement par $[(2n-3)\mu_4+\sigma^4]$ /

[2(n -1)^2]; en standardisant sur O^2, nous obtenons var(s^2_p) / O^2 = *(9n* -11) / *[5(n* -1)^2] pour des v.a. uniformes et *(3n* -4) / *(n* —1)^2 pour des v.a. normales. La présence de corrélation positive (p>0) dans la série tendra à amoindrir s^2_p, et vice versa. Cette statistique, qui correspond étroitement à la corrélation sérielle (R') examinée plus haut, est dotée d'une distribution un peu plus simple. Les auteurs (J. Von Neumann, L. C. Young et d'autres, voir Laurencelle 1983) ont aussi considéré le quotient de variance permutative,

$$\text{QVP} = s_p^2 / s^2 \; ; \qquad (6.23)$$

ce quotient a une espérance de 1 sous l'hypothèse d'indépendance des v.a. considérées.

Pour tester l'hypothèse d'indépendance d'une source de v.a.u., il suffit en principe d'utiliser *n* valeurs successives de la variable, de calculer la statistiques s^2_p selon (6.22), puis de la comparer à une valeur critique. Le tableau *6.4* présente un jeu de valeurs critiques à cette fin; noter que la distribution de s^2_p est positivement asymétrique, avec $\gamma_1 \approx 2{,}2/\sqrt{n}$ [Laurencelle 1993 fournit les moments exacts]. Une valeur significativement basse de s^2_p indique la présence de corrélation positive des valeurs successives.

> L'approche de s^2_p vers la forme normale est assez lente, et les valeurs critiques approximatives $v_a = 1 + z_a O$, où z_a est une v.a. $N(0,1)$ d'intégrale a, sont valables à partir de $n = 1000$.

D'autre part, pour une série de v.a. supposées normales, on ne connaît pas d'ordinaire la variance O^2 (ni l'espérance t) de la distribution, de sorte que le quotient de variance permutative constitue un test plus pratique. Hart (1942) donne une tabulation des f.r. de QVP pour quelques valeurs de n. Young (1941), d'autre part, étudie la distribution de QVP à travers toutes les permutations des *n* variables et il fournit les moments génériques de même que ceux spécifiques à une loi normale.

> Pour des v.a. de loi normale, μ(QVP) = 1, O^2 = (n-2)/(n^2-1), $\gamma_1 = 0$ et $\gamma_2 = [3(n^2+2n-12)]/[(n^2-1)(n+3)(n+5)] - 3$, soit $\gamma_2 < 0$. Il s'agit donc d'une distribution symétrique et platykurtique, tout comme la loi *Bêta* symétrique (3(p,p) déjà mentionnée (en §6.14). La variable $y = \beta(p,p)$ varie de 0 à 1, et $x = K_n(y-0{,}5)$ varie autour de 0 avec une variance contrôlée par K_n. D'abord, le paramètre p est fixé afin de mimer l'aplatissement de la distribution de QVP: la loi β *(p,p)* ayant un indice de voussure $\gamma_2 = -6/(2p+3)$, on estime $p = -3/\gamma_2(n) - 1{,}5$. Ensuite, puisque la variance de y est 1/(8p+4), le paramètre K,, est fixé par $2\sqrt{[(n^2-1)(2p+1)]}/(n-1)$. De plus, il appert ici que l'approche vers la forme normale est assez rapide de sorte que, par exemple, la quantité (QVP -1)/a est à peu près de loi $N(0,1)$ dès $n \geq 100$.

Tableau 6.4 Centiles de $12s_p^2(u)$, la variance permutative standardisée de n v.a.uniformes†

$n \setminus P$	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
2	$,0^437$	$,0^315$	$,0^395$	$,0^238$	3,62	4,25	4,86	5,18
3	,005	,010	,026	,053	2,81	3,32	3,93	4,31
4	,024	,039	,074	,122	2,46	2,84	3,32	3,65
5	,053	,076	,125	,185	2,25	2,58	2,97	3,24
6	,083	,113	,171	,236	2,12	2,40	2,75	2,99
7	,115	,148	,211	,279	2,00	2,25	2,56	2,77
8	,144	,180	,248	,319	1,93	2,16	2,45	2,67
9	,170	,209	,278	,351	1,87	2,08	2,35	2,52
10	,196	,235	,306	,380	1,81	2,02	2,25	2,43
11	,217	,257	,334	,405	1,77	1,96	2,19	2,35
12	,237	,280	,354	,425	1,73	1,90	2,12	2,28
13	,257	,301	,374	,445	1,70	1,86	2,06	2,20
14	,277	,322	,390	,464	1,67	1,82	2,01	2,15
15	,293	,335	,407	,481	1,64	1,79	1,98	2,10
20	,369	,410	,480	,544	1,55	1,67	1,82	1,92
25	,422	,463	,530	,591	1,48	1,59	1,72	1,81
30	,463	,501	,563	,622	1,44	1,53	1,65	1,73
40	,521	,560	,618	,670	1,37	1,45	1,56	1,62
50	,566	,601	,654	,704	1,33	1,40	1,48	1,55
75	,636	,667	,714	,756	1,27	1,32	1,39	1,44
100	,681	,708	,750	,787	1,23	1,28	1,34	1,38
150	,736	,759	,794	,825	1,19	1,22	1,27	1,30
200	,769	,790	,821	,848	1,16	1,19	1,23	1,26
250	,792	,811	,839	,864	1,14	1,17	1,21	1,23
300	,809	,827	,853	,875	1,13	1,16	1,19	1,21
400	,834	,849	,872	,892	1,11	1,14	1,16	1,18
500	,851	,865	,885	,903	1,10	1,12	1,14	1,16
1000	,893	,903	,918	,931	1,07	1,09	1,10	1,11

† Pour $n = 2$, les valeurs données sont exactes; pour $n = 3$ à 7, elles ont été obtenues par un quadrillage systématique de l'hypercube $[0,1]^n$; pour $n = 8$ à 50, elles proviennent de 99 999 échantillons Monte Carlo; au-delà de 50, nous les avons produites grâce au modèle du x^2 dans Laurencelle (1993, p. 15).

Tableau 6.5 Centiles de QVP, le quotient de variance permutative, pour des v.a. normales†

$n \setminus P$	0,005	0,01	0,025	0,05	$n \setminus P$	0,005	0,01	0,025	0,05
4	,303	,313	,342	390	20	,476	,520	,588	,650
5	,240	,269	,332	,410	25	,523	,564	,627	,684
6	,238	,281	,360	,445	30	,559	,598	,657	,709
7	,259	,307	,389	,468	35	,588	,624	,680	,729
8	,284	,331	,412	,491	40	,612	,647	,699	,746
9	,305	,354	,435	,512	45	,632	,665	,716	,760
10	,327	,376	,456	,531	50	,649	,681	,730	,772
11	,347	,396	,474	,548	60	,678	,708	,752	,791
12	,319	,385	,482	,565	70	,700	,728	,770	,806
13	,383	,431	,507	,578	80	,719	,745	,784	,818
14	,399	,447	,522	,591	90	,734	,759	,796	,828
15	,414	,461	,535	,603	100	,747	,771	,806	,837

† Les centiles pour $n = 4$ à 9 sont obtenus par interpolation dans les tables de Hart (1942). Pour $n = 10$ et au delà, les valeurs proviennent d'une loi *Bêta* symétrique (voir texte). L'approximation normale $v_\alpha = 1 + z_\alpha\sigma$, où $\sigma^2 = (n-2)/(n^2-1)$ et z_α est le $100\alpha^e$ centile d'une v.a. $N(0,1)$, devient sûre à partir de $n = 100$. Noter que, la distribution de QVP étant symétrique autour de l'espérance 1, les centiles supérieurs sont obtenus simplement par $C_{1-\alpha} = 2 - C_\alpha$.

Le tableau 6.5, ci-dessus, présente des valeurs critiques pour la statistique QVP dans le cas de données normales.

Tests sur l'équivalence des permutations

6.17 *Le test des permutations.* Si des séries différentes de n valeurs chacune sont vraiment au hasard, les séries des n rangs correspondants le sont aussi et constituent autant de permutations équiprobables des entiers 1 à n, parmi les $n!$ permutations possibles. Cette interprétation de l'hypothèse d'indépendance est a-distributionnelle et elle procède comme suit. On constitue S séries de n valeurs, telle la série $\{x_1, x_2, \cdots, x_n\}$, puis, pour chacune, on identifie son numéro de permutation d'après une convention de numérotation quelconque et on enregistre ce numéro dans un tableau présentant les $n!$ numéros possibles. Sous l'hypothèse d'indépendance

(6.21), chaque numéro a une probabilité égale à 1 */n* ! *:* la répartition des S numéros enregistrés est alors jugée par un test d'ajustement, comme le Khi-deux (6.2).

> Si puissant qu'il soit pour détecter une violation de l'hypothèse d'indépendance, ce test est de réalisation ardue. D'abord le nombre N de données *x*; requises peut être très élevé. Posons une *fréquence théorique* moyenne de 5 pour le test du Khi-deux; il y a *n*! cases à remplir, chaque enregistrement représentant *n* données x_i, d'où $N = 5n^2(n-1)!$, soit N = 3000 pour tester des quintuplets *(n = 5),* N = 21 600 pour des sextuplets *(n* = 6), etc.
>
> La difficulté du test s'accroît de deux autres aspects: le premier, de ce qu'il faut convertir la série des données $\{x_1, x_2, ..., x_n\}$ en une série de rangs $\{r_1, \mathbf{r}_2, ..., r_n\}$, la seconde, de ce qu'il faut « numéroter » la série de rangs, c'est-à-dire identifier et classer la permutation obtenue parmi les *n*! permutations possibles afin d'en enregistrer l'apparition. Les exercices 6.21 à 6.24 proposent des algorithmes pour ces divers problèmes.

Malgré les difficultés de mise en œuvre que ce test présente, il constitue, conjointement avec le test sériel (§6.13), la preuve la plus directe de l'indépendance des séries de v.a. et ce, quelle qu'en soit la distribution.

6.18 *Les suites monotones. Au* lieu du test exhaustif consistant à étudier l'ensemble des permutations d'une série de *n* nombres, il est possible et intéressant d'en considérer certains aspects révélateurs: les variations monotones dans la série, croissantes ou décroissantes, en sont un. Cet aspect désigne globalement, dans une série, les suites de nombres qui varient dans la même direction, soit en valeurs croissantes, soit au contraire. On utilise couramment trois définitions de suites monotones: les suites alternées, les suites simples, les suites isolées. Sous l'hypothèse de hasard pur et d'indépendance séquentielle, chacune des *n*! permutations des *n* nombres d'une série est équiprobable, ce qui donne lieu à une distribution statistique du nombre de suites monotones d'une sorte ou l'autre, de leurs longueurs et de leur longueur maximale. Des données mathématiques partielles sont disponibles sur ces distributions (Bradley 1968; Knuth 1975; Olmstead 1946).

Les *suites monotones alternées,* ou mixtes, dénotent une fragmentation de la série de *n* nombres en S suites monotones comportant deux ou plusieurs éléments chacune, la variation d'une suite à l'autre alternant d'orientation. Par exemple, dans la série:

{22 6 14 39 18 12 8 11 1 },

nous observons les suites alternées (22 6)↓, (6 14 39)↑, (39 18 12 8)~, (8 11)↓, la flèche indiquant l'orientation de variation. La probabilité $pr_n(S)$, *i.e.* celle d'observer S suites alternées de longueurs quelconques dans une série de n données, admet une définition récursive:

$$pr_n(S) = \frac{1}{n}\{\, S\cdot pr_{n-1}(S)+2\cdot pr_{n-1}(S-1)+(n-S)\cdot pr_{n-1}(S-2)\,\}\ ; \quad (6.24)$$

en particulier $pr_n(0) = pr_n(n) = 0$ et $pr_n(1) = 2/n!$. Les moments de cette distribution sont:

$$\mu = (2n-1)/3\ ; \quad (6.25a)$$

$$\sigma^2 = (16n-29)/90\ ; \quad (6.25b)$$

$$\gamma_1 = \frac{-\sqrt{10240}\,(n+1)}{7(16n-29)^{3/2}}\ ; \quad (6.25c)$$

$$\gamma_2 = \frac{-3(1408n-3317)}{7(16n-29)^2}\ . \quad (6.25d)$$

Bradley (1968) fournit des valeurs critiques basées sur la distribution (6.24), pour $n \leq 25$. Pour n assez fort $(n \geq 50)$, les indices y_1 et y_2 approchent zéro et l'approximation normale, utilisant μ et O^2, fait très bien l'affaire. Le nombre de suites alternées de longueurs $L \geq 1$ a pour espérance:

$$\mu[\#L \geq l] = \frac{2 + 2(n-l)(l+1)}{(l+2)!}\ ; \quad (6.26)$$

pour $l \geq \frac{1}{2}n$, l'espérance *(6.26)* représente en même temps la probabilité d'obtenir une suite monotone de longueur 1 ou plus. Olmstead (1946) fournit des valeurs critiques de la longueur de la suite la plus longue.

Quant aux *suites monotones simples,* par exemple, les suites ascendantes, elles dénotent tout groupe d'un à plusieurs éléments successifs en variation croissante (les suites descendantes, de nombre S -, sont définies pareillement). La série {22 6 14 39 18 12 8 11} contient $S^+ = 5$ suites simples, en groupes soulignés. La distribution de probabilités du nombre de suites simples (S^+) est connue (exercice 6.26). Espérance et variance sont respectivement (n+1)/2 et (n+1)/12, puis $y_i = 0$ et $y2 = -6/[5(n+1)]$: l'approximation par une *Bêta* symétrique est donc indiquée. Notons la remarquable parenté de cette statistique avec la somme de n+l variables de loi $U(0,1)$, leurs quatre premiers moments étant identiques. La longueur moyenne (L) des suites simples d'une série a pour espérance $2n/(n+1)$.

Knuth (1975) donne aussi l'espérance de la longueur de la j^e suite d'une série, les espérances approchant rapidement la constante 2. Enfin, Knuth (1969; Grafton 1981) présente un test du Khi-deux du nombre de suites simples des longueurs 1 à 6 et plus; la taille de série *(n)* recommandée pour ce test est d'au moins 4000. Le calcul est complexe, compte tenu du fait que ces suites ne sont pas statistiquement indépendantes et qu'une suite longue a tendance à être suivie d'une suite plus courte.

Même dans une série de nombres au hasard, les suites simples ne sont pas statistiquement indépendantes l'une de l'autre. Pour obvier à cet inconvénient, on peut définir des *suites isolées,* en fait des suites simples qu'on débarrasse de leur élément butoir. Après recodage et en considérant les suites ascendantes, notre exemple devient {22 (6) 14 39 (18) 12 (8) 11}, comportant $S^{(+)} = 4$ suites isolées (ascendantes), les nombres enlevés étant mis entre parenthèses. On montre aisément que, parmi les suites trouvées, la probabilité d'observer une suite isolée de longueur L est égale 1/L! -1/(L+1)!. Les moments de la variable $S^{(+)}$ ne sont pas connus.

6.19 *Le nombre d'inversions.* Ayant considéré le nombre de suites monotones, on peut aussi regarder le nombre d'inversions dans la série par rapport au modèle d'une série strictement ascendante: par exemple, la série { 22 6 14 39 18 12 8 11 } comporte q = 18 inversions, soit 6 inversions avec «22» (i.e. 22>6, 22>14, ..., 22>11), 0 avec «6», 3 avec «14» (14>12, 14>8, 14>11), etc. La variable q varie de 0 à $\frac{1}{2}n(n-1)$. Dans une série aléatoire de n nombres, q a pour moments:

$$\mu = n(n-1)/4 \; ; \tag{6.27a}$$

$$\sigma^2 = n(2n+5)(n-1)/72 \; ; \tag{6.27b}$$

$$\gamma_1 = 0 \; ; \tag{6.27c}$$

$$\gamma_2 = -\left(\frac{6}{5}\right)\left[\frac{6n^3+21n^2+31n+31}{n(n-1)(2n+5)^2}\right] . \tag{6.27d}$$

Le coefficient τ (tau) de Kendall, une mesure dite non paramétrique de corrélation, est en quelque sorte une re-expression « normalisée » du nombre d'inversions q. La relation entre les deux est:

$$\tau = 1 - 4q/[n(n-1)] \; ; \; q = n(n-1)(1-\tau)/4 \; ; \tag{6.28}$$

on a $-1 \le \tau \le 1$. Kendall et Stuart (1979) fournissent les premiers moments du coefficient τ.

Pour modéliser cette variable (ou la statistique associée τ), la loi *Bêta* symétrique, $\beta(p,p)$, fera encore bon office, en fixant p selon l'indice de voussure γ_2.

6.20 *Le débordementdu r^e maximum.* Supposons que nous avons obtenu, ou produit, quelques valeurs de x, disons r valeurs. La moyenne de celles-ci, leur étendue, leur valeur maximale dépendent toutes de la loi f dont ces valeurs proviennent (voir David 1981, p. 31, exercice 2.7.6). Cependant, la probabilité que la valeur suivante, x_{r+1}, déborde l'ensemble des r valeurs antérieures est indépendante de la loi d'origine, ce qui fonde le test du débordement du r^e maximum.

La distribution de probabilité de n, le nombre de nouvelles variables requises pour déborder le maximum de r variables, est (voir exercice 6.30):

$$g(n) = \int_0^1 u^{n-1}(1-u)r\cdot u^{r-1}\,du = [(n+r)(n+r-1)]^{-1}; \qquad (6.29)$$

pour la f.r. de n, on a simplement $n/(n+r)$. Cette distribution n'a pas de moments, l'espérance de n étant infinie. Pour atteindre un seuil de probabilité α, on détermine la valeur critique $n(\alpha)$ selon $n(\alpha) \geq r(1 - \alpha)/\alpha$.

Le processus de débordement du maximum peut être envisagé de manière *cumulative.* Soit une première valeur observée $M_1 = x_1$. Il faudra n_1 nouvelles v.a. pour déborder x_1 (de telle sorte que $x_2 < M_1, x_3 < M_1, ..., x_{n1+1} > M_1$). Nous avons alors un nouveau maximum cumulatif, $M_2 = x_{n1+1}$, qu'il s'agit à présent de déborder, etc. Notons que, à la différence de tantôt, il faut déborder ici M_r, c'est-à-dire le maximum d'une série de $1 + n_1 + ... + n_r$ variables; la distribution de la statistique n_1 coïncide avec celle vue plus haut lorsque $r = 1$. La résolution du cas $r = 2$ est plus complexe (voir l'exercice 6.31 et David 1981, p. 31, exercice 2.7.6); notons tout de même les valeurs critiques $P(n_2 \geq 104) < 0{,}05$ et $P(n_2 \geq 715) < 0{,}01$.

La distribution et les moments de la *valeur* du r^e maximum issu d'une séquence $x_1, x_2, x_3, ...$, où x obéit à une densité de probabilité donnée, sont étudiés dans Johnson, Kotz et Balakrishnan (1994, 1995), sous le vocable « *record value* ». Les exercices *6.32* et *6.33* présentent le cas pour des v.a. uniformes.

6.21 *Les suites d'éléments catégorisés: taxonomie.* Des suites d'une autre espèce, plus connues celles-là, sont obtenues en catégorisant les données d'une série en deux ou en plusieurs types. Soit la série $(x'_i, x_2, ..., x_n)$, comportant n valeurs; la série convertie, $(x'_i, x'_2, ..., x'_n)$, où $x_i = t_j$, contient n_j de chaque type t_i, $1 \leq j \leq k$. La série $(x'_i, x'_2, ..., x'_n)$ se présente comme une succession de types d'éléments, chaque séquence d'éléments d'un même type constituant une suite. L'étude des séries catégorielles est ancienne et très riche (Mood 1940), cependant la théorie distributionnelle n'en est pas encore achevée. Laurencelle (2000) fait le tour de la question.

La caractéristique la plus couramment mesurée des séries catégorielles concerne le nombre de suites, particulièrement le nombre total de suites (S), tous types confondus. On étudie aussi la longueur des suites, particulièrement la longueur de suite maximale (L_{max}), tous types confondus. Le nombre de répétitions de types (R), le nombre de chaînes (ou suites enchaînées), etc., constituent d'autres objets statistiques.

Il est loisible d'étudier la structure d'une série d'éléments donnés, telle la série ($x_1, x_2,..., x_n$), pour laquelle la composition des valeurs et des types est fixée d'emblée, ou bien la structure d'une série d'éléments libres, qu'on pourrait noter ($X_1, X_2, ..., X_n$), chaque v.a. X_i provenant d'une loi multinomiale $M(n; л_1, л_2, ..., л_k)$ déterminée (cf. exercice 6.5). L'examen et la validation d'une source de v.a., d'un générateur, ressortissent plutôt à cette seconde option en vertu de laquelle on cherche à prédire le comportement de la source. La première option, utilisant une série d'éléments donnés, concerne surtout les applications statistiques ponctuelles, l'enquêteur voulant vérifier le caractère aléatoire de sa série d'observations.

Enfin, les séries sont catégorisées en *k* classes, des classes naturelles ou imposées, la catégorisation binaire *(k=2)* constituant un cas particulier, privilégié dans la documentation.

> La catégorisation de la série doit se faire en utilisant un critère global: par exemple un critère binaire naturel (le « sexe » des personnes, la parité des nombres); un critère externe (l'inclusion d'une v.a. uniforme dans l'intervalle *[(j -1)/k, j/k] ou* la sanction Succès/Echec d'un test scolaire selon X ≥60 % et X < 60 %); enfin, un critère interne *et* global ($X_i \geq$ Md ou X. < Md, où Md est la médiane de la série). Un critère interne, tels l'accroissement (+) ou la diminution (—) d'un nombre par rapport au nombre précédent dans la série, n'aboutit pas à une suite binaire, mais constitue une codification d'une suite monotone.

6.22 *Les séries binairesd 'éléments donnés. La* série suivante, représentant une file de 34 personnes, est codée selon le sexe:

{ ♀♂♀♀♂♀♂♂♂♀♂♂♂♀♀♂♂♀♂♀♂♂♂♂♀♂♂♀♂♀♀♂♂ } .

On compte $n_1 = 21$ mâles, $n_2 = 13$ femelles, 10 suites de type 1 (♂) et 10 de type 2 (♀), soit un total de S = 20 suites; la suite-1 la plus longue est de 5, la suite-2 est de 2, la suite la plus longue (soulignée) étant donc de 5.

La distribution de S, le nombre total de suites binaires, est connue depuis longtemps (Mood 1940). Sa loi de probabilité est:

$$\text{pr}(S=2u) = \frac{2\binom{n_1-1}{u-1}\binom{n_2-1}{u-1}}{\binom{n}{n_1}} ; \qquad (6.30)$$

$$\text{pr}(S=2u+1) = \frac{\binom{n_1-1}{u-1}\binom{n_2-1}{u}+\binom{n_1-1}{u}\binom{n_2-1}{u-1}}{\binom{n}{n_1}} ,$$

et les moments de S sont (Laurencelle 1995):

$$\mu(S) = \frac{2n_1n_2}{n} + 1 \qquad (6.31a)$$

$$\sigma^2(S) = \frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)} = \frac{(\mu-1)(\mu-2)}{n-1} \qquad (6.31b)$$

$$\gamma_1(S) = \sqrt{\frac{n-1}{2(n-2)^2}} \times \frac{16n_1^2n_2^2 - 4n_1n_2n(n+3) + 3n^3}{\sqrt{n_1n_2(2n_1n_2-n)^3}} ; \qquad (6.31c)$$

$$\gamma_2(S) = \qquad (6.31d)$$
$$\frac{(n-1)\left[n^4(8n_1n_2-13) - 7n^5 + 2n^3(52n_1n_2+3) - 120n^2n_1^2n_2^2 + 24nn_1^2n_2^2(n_1n_2-6) + 144n_1^3n_2^3\right]}{2n_1n_2(2n_1n_2-n)^2(n-2)(n-3)} - 3$$

Le troisième moment central (μ_3 ou γ_1) a comme facteur —(n_1—n_2)^2; l'asymétrie, négative en général, s'annule lorsque $n_1 = n_2$. Des valeurs critiques pour petits n se trouvent dans Owen (1962), pour n_1,n_2 20, *et* Laurencelle et Dupuis (2000), pour n_1,$n_2 \leq 50$. Pour grands n, l'approximation normale est boiteuse si $n_1 \neq n_2$, même en incluant la correction pour la continuité. Laurencelle (1995) suggère d'utiliser l'approximation par expansion de Edgeworth (Johnson, Kotz et Balakrishnan 1994, 1995), comme suit. Soit S, le nombre total de suites binaires, alors la f.r. P(S) est approchée selon:

$$z = (S \pm ½ - \mu)/\sigma \qquad (6.32)$$

et:

$$P(S) = \Phi(z) - \varphi(z)\left[\frac{\gamma_1}{6}(z^2-1) + \frac{\gamma_2}{24}(z^3-3z) + \frac{\gamma_1^2}{72}(z^5-10z^3+15z)\right], \quad (6.33)$$

où φ et Φ sont les fonctions de densité et de répartition de la loi $N(0,1)$. Pour la série de 34 observations reproduite ci-dessus, avec $n_1 = 21$ et $n_2 = 13$, moyenne et variance de S sont 17,06 et 7,33, $\gamma_1 \approx -0{,}042$ et $\gamma_2 \approx -0{,}088$. Au seuil de signification bilatéral de 5%, les valeurs critiques sont 11 et 23; de plus, l'évaluation par (6.33), avec $z \approx 0{,}902$ et $\Phi(0{,}902) \approx 0{,}816$, donne l'approximation $P(20) \approx 0{,}814$: par conséquent, notre résultat, S = 20, ne dénote pas un regroupement significatif des personnes selon le sexe.

Laurencelle (2000) donne aussi la distribution et les moments de S_1, le nombre de suites d'éléments du type 1, de même que la distribution et un jeu de valeurs critiques pour L_{max} ($n_1 \leq 100$, $n_2 \leq 110$). Les exercices 6.34 et 6.35 donnent d'autres informations sur L_{max}.

6.23 *Les séries binaires d'éléments libres (ou séries binomiales)* .Une série dite binomiale est une séquence de v.a. issues d'une loi *B(n, л)* (§4.4) ou d'un processus équivalent; dans ce cas, les nombres n_1 et n_2 ne sont pas fixés d'avance, si ce n'est que leur somme, n_1+n_2, est connue. La distribution de probabilités de S, le nombre total de suites, résulte alors de la convolution (ou addition croisée) de la distribution de S (6.27) et de n_1 (4.7). soit:

$$\text{pr}(S \mid n,\pi) = \sum_{n_1=S}^{n} \text{pr}(S \mid n_1, n-n_1) \times b_{n_1}(n,\pi) . \quad (6.34)$$

Le calcul de (6.34) permet à Laurencelle (2000) de produire des jeux de valeurs critiques pour différentes valeurs de n ($2 \leq n \leq 50$) et de π ($\pi = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{10}$). Les deux premiers moments de S, pour π quelconque, sont:

$$\mu(S \mid n,\pi) = 1 + 2\pi(1-\pi)(n-1) \quad (6.35a)$$

$$\sigma^2(S \mid n,\pi) = \pi(1-\pi)[4(n-1)(1-3\pi(1-\pi)) - 2(1-2\pi)^2] . \quad (6.35b)$$

Pour le cas égalitaire, *i.e.* л =½,la distribution et les moments sont très simples: voir §6.25.

Laurencelle (2000) donne aussi des valeurs critiques de S_1 de même que des indications et un exemple pour juger L_{max} et $L_{max:1}$ dans le présent contexte.

Exemple 6.2 Tests sur les suites monotones et les suites binaires d'une série

La série. Nous obtenons n = 20 valeurs successives d'une v.a., constituant la série suivante:

{ 65,5 51,3 46,8 40,0 39,3 33,8 57,3 47,7 48,1 36,0
51.7 57.8 63.8 55.8 38.8 68.9 60.7 59.5 57.4 48.2 }

La série comporte S=9 suites alternées, la plus longue (L_{max}) étant de 6 éléments; S^{+}=14 suites (simples) ascendantes, S^{-}=7 descendantes (avec $S^{+}+S^{-} = n+1 = 21$); $S^{(+)}$=6 suites isolées ascendantes et $S^{(-)}$=9 descendantes, la plus longue comportant (L_{max}=) 6 éléments descendants.

L'analyse. Le nombre de suites alternées (S) a pour moments (6.25) $\mu = 13{,}000$, $\sigma^2 \approx 3{,}233$, $\gamma_1 \approx -0{,}061$ et $\gamma_2 \approx -0{,}126$. D'après (6.24), la probabilité d'observer S=9 suites ou moins, pr(S≤9), égale 0,025483, indiquant une tendance à de la corrélation positive dans la série[4]. Quant aux suites simples (p.ex. S^{+}), leurs moments sont μ=10,5, σ^2=1,75, γ_1= 0 et $\gamma_2 = -6/105 \approx -0{,}057$. L'indice γ_2 étant très faible, l'approximation normale peut convenir. Transformant la statistique S^{+} en écart-réduit, $z = (S^{+} - \mu \pm ½)/\sigma \approx 2{,}268$ [incorporant une réduction de ½ pour le recours à une loi continue], nous trouvons $\Phi(2{,}268) \approx 0{,}988335$, soit une probabilité extrême d'environ 0,012, qui dénote un nombre peut-être surprenant de suites simples (les suites ascendantes apparaissant trop nombreuses, et les descendantes, trop peu nombreuses). La loi *Bêta* appropriée, $\beta(p,p)$, à indice $\gamma_2 = -6/(2p+3)$, a un double coefficient p = 51 pour émuler la distribution avec $\gamma_2 = -6/105$. Après conversions et réduction de continuité, la valeur de la f.r. d'une loi *Bêta* ajustée est de 0,988870, pratiquement la même valeur qu'avec le modèle normal. Enfin, la suite isolée (descendante) la plus longue a L=6 éléments, la probabilité d'observer une suite égale ou plus longue étant de 1/6! = 0,00138. Nous proposons le test approximatif suivant. Chaque suite isolée étant bornée, la série comporte à peu près $t = n/(L+1)$ occasions de manifester une suite de longueur L, et la probabilité d'observer au moins une apparition d'une telle suite est $p = 1-(1-1/L!)^t$; ici, $t \approx 2{,}857$ et $p \approx 0{,}00396$; nous pouvons encore

4. Par comparaison à des séries sans corrélation interne, la corrélation positive (p>0) se manifeste par une tendance à la viscosité, aux changements «lents », dans la variation d'une donnée à la suivante, alors que la corrélation négative (p<0) se manifeste par de la turbulence, des oscillations plus rapides dans les données successives.

conclure à la présence d'une variation monotone significative dans notre série.

La série recodée en types binaires. Supposons que la série ci-dessus provienne d'un processus générateur dont la médiane est de 50. Nous pouvons alors recoder la série en convertissant chaque valeur selon qu'elle est inférieure (—) ou bien égale ou supérieure (+) à 50. La série recodée:

{ + + — — — — + — — — + + + + — + + + + — }

contient n_1=11 valeurs de type 1 (+), n_2=*9* de type 2 (—), S-8 suites (ou séquences homogènes) au total, la plus longue ($L_{max.}$) ayant 4 éléments. Notons aussi que, pour chaque type, L_{max}:1 = 4 et L_{max}:2 = 4.

L 'analyse. Considérant les $n_1 + n_2 = 20$ éléments comme donnés, les premiers moments du nombre de suites S sont $\mu = 10{,}9$ et $O^2 \approx 4{,}6374$. Les valeurs critiques, au seuil global de 10%, sont $S_{0,05}$=6 et $S_{0,95}$=15: pas de regroupement significatif des codes. La valeur L_{max}= 4 n'atteint jamais le seuil de signification de 0,05, quels que soient les n_j (Laurencelle et Dupuis 2000). D'un autre côté, si M = 50 est vraiment la médiane du processus parent de la série, alors la série codée d'après les relations « X < M » et « X ≥ M » constitue un processus de Bernoulli à probabilité л = ½. Dans ce cas, les moments de S (6.35) deviennent $\mu = (n+1)/2 = 10{,}5$ et $O^2 = (n-1)/4 = 4{,}75$. Le calcul de (6.34) fournit la distribution de probabilités, puis la f.r. de S; les tables de Laurencelle et Dupuis donnent$S_{0,05}$=6 et $S_{0,95}$=15 comme valeurs critiques pour n_1=11, n_2 =9 et л =½. Par (6.34), nous obtenons pr(S≤8) ≈ 0,17964 et pr(S≥8) ≈ 0,82036: rien de surprenant dans le regroupement des données. Quant à la longueur maximale, il n'y a pas de méthode établie pour la distribution de Lmax en situation de Bernoulli. Nous pouvons cependant trouver la distribution de $_{Lmax:1}$ en convoluant la distribution propre de $L_{max:1}$ (exercice 6.34) pour (n_1, n_2) avec la distribution binomiale de n_1 (4.7), obtenant:

$$\mathrm{pr}(L_{max:1} \geq L \mid n,\pi) = \sum_{n_1=L}^{n} \pi^{n_1}(1-\pi)^{n-n_1}\left[\sum_{i=1}^{n_1/L}(-1)^{i+1}\binom{n-n_1+1}{i}\binom{n-iL}{n-n_1}\right] . \quad (6.36)$$

Le calcul est quelque peu facilité dans notre cas, pour lequel л = 1— л = ½. L'évaluation donne ici pr($L_{max:1}$≥4) ≈*0,478019:* encore, pas de quoi fouetter un chat! Des résultats similaires découlent d'un autre recodage, celui dans lequel nous aurions employé la médiane empirique (Md) pour classifier les nombres: ici, Md = 51,5;

toutefois, nous n'aurions pas été autorisé alors à traiter la série codée comme provenant d'un processus de Bernoulli.

S'il fallait conclure, nous pourrions dire que l'analyse en variation monotone s'est montrée plus puissante et plus expressive, puisqu'elle a permis de faire apparaître une tendance anormale aux lentes fluctuations. L'analyse après recodage binaire, forcément moins puissante après la perte d'information encourue, eût nécessité une plus longue série pour s'avérer révélatrice.

6.24 *Les séries multiplesd 'éléments donnés.* La série d'éléments convertis en k types différents donne lieu à n ! permutations parmi lesquelles seules $M_k(n;\ n_j)$ permutations sont mutuellement discernables, $M_k(n;\ n_j)$ étant le coefficient multinomial (6.46). Pour le cas général d'une composition $\{n_1, n_2, \ldots, n_k\}$ quelconque de la série, l'espérance et la variance du nombre total de suites (S) sont (David et Barton 1962):

$$\mu(S) = n + 1 - \frac{\sum n_j^2}{n} \tag{6.37a}$$

$$\sigma^2(S) = \frac{\left(\sum n_j^2\right)^2 + n(n+1)\sum n_j^2 - n\left(n^2 + 2\sum n_j^3\right)}{n^2(n-1)} . \tag{6.37b}$$

Les moments se simplifient pour des compositions égalitaires, avec $n_j = n/k$, et nous avons alors (Laurencelle 2000):

$$\mu(S) = n + 1 - n/k \tag{6.38a}$$

$$\sigma^2(S) = \frac{n(n-k)(k-1)}{(n-1)k^2} = \frac{(\mu-1)(n-\mu)}{n-1} \tag{6.38b}$$

$$\gamma_1(S) = -\sqrt{\frac{n(n-1)}{(n-k)(k-1)}} \times \frac{k-2}{n-2} \tag{6.38c}$$

$$\gamma_2(S) = \frac{(n^3-n+6)k^2 - 6nk(n^2-n+2) + 6n^3}{n(n-2)(n-3)(n-k)(k-1)} . \tag{6.38d}$$

La distribution de probabilités de S, malcommode à calculer, peut être construite à partir d'une fonction récursive de Saughnessy (1981; voir l'exercice 6.36). Pour k fixe, l'accroissement de n fait tendre cette distri-

bution vers la normale alors que, pour n fixe, elle développe une asymétrie négative et une leptokurtose à mesure que k croît.

6.25 *Les séries multiples d'éléments libres (ou séries multinomiales). Si,* maintenant, nous voulons caractériser une source de v.a. à types multiples, plutôt qu'une série d'observations donnée, nous devons traiter avec un processus multinomial, répondant à la loi multinomiale $M(n;\ л_1, л_2, ..., л_k)$; dans un tel cas, la composition $(n_1, n_2, ..., n_k)$ des séries étudiées n'est pas fixée, puisqu'elle est elle-même une réalisation de la loi multinomiale, une v.a. Deux cas se présentent, selon que les probabilités $\{ л_j \}$ soient égales ou inégales.

Pour le cas d'un vecteur de probabilités $\{ л_i, л_2, ..., л_k \}$ quelconque, seuls les deux premiers moments de S sont connus, soit:

$$\mu(S\,|\,n, \pi_1\ \pi_2 \cdots \pi_k) = 1 + (n-1)[1 - \textstyle\sum_{j=1}^{k} \pi_j^2] \qquad (6.39a)$$

$$\sigma^2(S\,|\,n, \pi_1\ \pi_2 \cdots \pi_k) = (n-1)\textstyle\sum \pi_j^2 + 2(n-2)\sum \pi_j^3 - (3n-5)[\sum \pi_j^2]^2. \qquad (6.39b)$$

Pour le cas égalitaire, avec $\pi_j = 1/k$ pour tout j, la distribution de S, plus précisément celle de $S' = S-1$, est binomiale, avec les paramètres $n' = n-1$ et $\pi = (k-1)/k$. À l'image d'une v.a. binomiale (exercice 4.2), les moments de S sont ici:

$$\mu(S\,|\,n, \pi_j=1/k) = 1 + (n-1)(k-1)\,/\,k \qquad (6.40a)$$

$$\sigma^2(S\,|\,n, \pi_j=1/k) = (n-1)(k-1)\,/\,k^2 = (\mu-1)\,/\,k \qquad (6.40b)$$

$$\gamma_1(S) = (2-k)\,/\,\sqrt{[(n-1)(k-1)]} \qquad (6.40c)$$

$$\gamma_2(S) = (k^2-6k+6)\,/\,[(n-1)(k-1)]\,. \qquad (6.40d)$$

De plus, il existe de nombreuses publications offrant des valeurs critiques de la loi binomiale $B(n;\ л)$ ainsi que des méthodes d'approximation de la f.r. binomiale, ce qui contourne la nécessité de préparer des tables expresses pour ce cas-ci (voir cependant Laurencelle 2000, qui présente des tables pour $n \_\leq 50$, $k = 2$ à 6, 10, $\alpha = 0{,}05$ et 0,01 en mode uni- et bilatéral).

> Pour obtenir les valeurs critiques de $S(n; k)$ dans un processus multinomial à k types équiprobables, on doit donc utiliser $x = S-1$, qui suit une distribution $B(n-1; 1-1/k)$. Plusieurs tables binomiales, toutefois, donnent des valeurs critiques relatives à un paramètre $\pi = 1/k$ (plutôt que $1-1/k$), ce qui complique une peu la transposition. Dans ces cas, on peut utiliser la relation «$S^*_\alpha(n;k) = n - x^*_{1-\alpha}(n-1;\ 1/k)$». Par exemple, pour $k = 5$ et $n = 30$, les tables binomiales (*e.g.* Laurencelle et Dupuis 2000) indiquent $x^*_{0,95}(29; \frac{1}{5}) = 10$ et $x^*_{0,05}(29; \frac{1}{5}) = 1$, ce qui

fournit immédiatement$S^*_{0,05}(30;5) = 20$ et $S^*_{0,95}(30;5) = 29$. Dans le cas général et faute de tables, l'approximation directe par l'expansion de Edgeworth (6.33), avec les moments (6.40), convient très bien.

Le nombre (R) de répétitions de types dans une série de même que le nombre de chaînes (ou suites enchaînées) de longueur 2 (C_2) se distribuent comme le complément du nombre de suites, soit:

$$R\ ,\ C_2 \sim n\ -\ S \sim B(n-1, \tfrac{1}{k})\ . \qquad (6.41)$$

Les tables de f.r. et de valeurs critiques de la loi binomiale conviennent immédiatement. Les moments de R (ou de C_2) sont d'obtention facile (exercice 6.38).

Rien ne se trouve dans la documentation à propos des distributions des longueurs ou de la longueur maximale ($L_{max.}$) de suites pour les séries à types multiples.

Tests globaux d'irrégularité et divers tests

6.26 *Les tests de tests.* La plupart des tests que nous avons considérés jusqu'à présent permettent de se prononcer sur l'irrégularité d'une série statistique, la série étudiée, ou sur ce qu'il est convenu d'appeler l'irrégularité *locale* de la source de nombres (ou du domaine d'observations). Or, il est parfois possible d'exploiter à volonté cette source, comme c'est le cas d'un générateur programmé de nombres pseudoaléatoires. Dans ce contexte, les auteurs suggèrent de considérer aussi l'irrégularité *globale* de la source en étudiant le comportement d'un certain nombre de séries différentes, pour une statistique donnée. Ainsi, si l'hypothèse nulle d'une source homogène de nombres aléatoires et indépendants est valide, il peut et il doit advenir des séries « significatives », c'est-à-dire des séries pour lesquelles l'hypothèse d'irrégularité *locale* mesurée par une statistique sera rejetée; l'on s'attend cependant à ce que la répartition des statistiques de test suive globalement la loi caractéristique de cette statistique. D'un autre côté, les séries analysées pourraient ne jamais apparaître significatives tout en donnant systématiquement des statistiques de percentiles élevés, par exemple 75 à 80, ce qui dénoterait un biais global de la source aléatoire. Différentes concrétisations de cette idée sont possibles.

Par addition de la valeur des tests. Plusieurs tests d'irrégularité donnent lieu à une statistique simple, le plus souvent approximative, par

exemple un écart-réduit normal (z) ou un Khi-deux (X^2). On peut alors répéter ces tests *k* fois sur autant de séries différentes, puis additionner simplement les valeurs des statistiques. Pour les écarts-réduits normaux ou quasi normaux, leur somme est normale aussi, avec moyenne égale à 0 et variance égake à k sous l'hypothèse nulle, de sorte que la statistique $\Sigma^{k} z_i / \sqrt{k} \sim N(0,1)$ constitue un test global; la somme de leurs carrés, $X^2 = \Sigma^{k} z^2_j$ constitue cette fois un test du Khi-deux, avec k degrés de liberté. Pour les tests X^2 eux-mêmes, leur somme est encore un X , les degrés de liberté effectifs étant la somme des degrés de liberté des tests additionnés.

Par combinaison des probabilités. Sous l'hypothèse d'une source vraiment aléatoire, la f.r. d'un test, son percentile P, a une distribution *U*(0,1) sous l'hypothèse nulle, et $-21\log_e P$ se distribue alors comme x^2_2 La réalisation répétée du test sur k séries différentes donne lieu à autant de probabilités P_j, et l'on peut soit additionner leurs x^2_2équivalents, soit calculer directement $-21\log_e(P_1 \bullet P_2 \bullet \cdots \bullet P_k)$, les deux expressions se distribuant comme x^2 avec $v = 2k$.[5]

Par étude de la distributiondes valeurs des tests. Prenons l'exemple du *test sériel* (§6.13), dans lequel un calcul de Khi-deux (X^2) permet de statuer sur l'indépendance séquentielle dans la série statistique étudiée. En effectuant *n* tests pareils sur autant de séries différentes, nous obtenons *n* valeurs X^2, chacune avec v degrés de liberté. La valeur moyenne de ces tests devrait égaler à peu près v, leur variance 2v, leur médiane v —⅔+1 /(9v). Grâce à cette dernière, par exemple, on peut mettre sur pied un test d'équipartition des X^2 obtenus (selon qu'ils sont inférieurs ou supérieurs à la médiane) basé sur une loi binomiale B(*n*,½), voire un test d'ajustement par Khi-deux avec k=2 cellules: ces tests permettent de vérifier si les statistiques X^2 obtenues sont ou trop faibles ou trop fortes par rapport à leur répartition idéale. Pour étudier la distribution des *n* valeurs des tests, Knuth (1969) recommande spécialement le test de Kolmogorov-Smirnov *(§6.7)* grâce auquel la fonction de répartition empirique, correspondant aux *k* valeurs obtenues de la statistique, est comparée à la f.r. théorique.

> Pour le cas des tests à distribution normale, la f.r. normale requise, dénotée Φ(x), se retrouve dans plusieurs tables d'intégrale; elle peut aussi être calculée (cf. exemple 8.1). On peut utiliser aussi l'une des nombreuses approximations, notamment celle de Hastings (1955):

5. On serait tenté, par ce moyen des probabilités extrêmes, de combiner les résultats de k tests différents *de la même série statistique.* Cette méthode est généralement à proscrire puisque, les tests étant basés sur la même information, leurs résultats seront corrélés, ce qui invaliderait le théorème d'indépendance (des événements aléatoires) sur lequel l'interprétation du test Khi-deux est fondée.

$$\Phi(x) \approx 1 - \varphi(x)t(b_1 + t(b_2 + t(b_3 + t(b_4 + tb_5)))) \quad \{x \geq 0\} \qquad (6.42)$$

où $\varphi(x) = \exp(-\frac{1}{2}x^2)/\sqrt{(2\pi)}$ est la loi de densité normale $N(0,1)$, $t = (1 + 0{,}2316419x)^{-1}$, $b_1 = 0{,}31938153$, $b_2 = -0{,}356563782$, $b_3 = 1{,}781477937$, $b_4 = -1{,}821255978$, $b_5 = 1{,}330274429$; l'erreur d'approximation est partout inférieure à 10^{-7}.

Pour le cas des tests à distribution χ^2, de densité (4.17), la f.r. peut s'obtenir par intégration par parties. Selon que les degrés de liberté (ν) sont pairs ou impairs, nous avons:

$$F_{\chi^2_\nu}(x) = 1 - e^{-x/2} \sum_{i=0}^{(\nu-2)/2} (x/2)^i/i! \qquad \{\nu \text{ pair}\} \quad (6.43a)$$

$$= 2\Phi(\sqrt{x}) - 1 - e^{-x/2} \sum_{i=0}^{(\nu-3)/2} (x/2)^{i+\frac{1}{2}}/\Gamma(i+3/2) \qquad \{\nu \text{ impair}\} \quad (6.43a)$$

ces expressions deviennent lourdes pour v forts; diverses approximations sont aussi disponibles (Johnson, Kotz et Balakrishnan 1994, 1995; Laurencelle et Dupuis 2000).

D'autres tests parus dans la littérature s'inspirent de ce principe, tel le test du collectionneur décrit dans Knuth (1969), qui globalise un test d'occupation.

6.27 *L 'étendue de n v.a. uniformes.* Une source de v.a. uniformes devrait projeter des échantillons uniformément dans le domaine [0,1), l'une des caractéristiques importantes de ces variables étant leur *étendue.* Soit $E = u_{(n:n)} - u_{(1:n)}$, l'étendue d'une série de *n v.a.u.*. Obtenant *k* séries différentes et les estimations d'étendue correspondantes $E_l, E_2, ..., E_k$, on peut comparer la distribution de ces estimations à leur f.r. théorique donnée par (3.23), soit $F_E(x) = nx^{n-1}(1 - x) + x^n$. Ce test, que nous recommandons, a « débusqué » pour nous un générateur de nombres pseudo-aléatoires qui avait passé avec succès la plupart des autres tests.

6.28 *Le maximum de n v.a. uniformes.* Le maximum de *n v.a.u.,* $\max(u_1)$ ou $u_{n.n)}$, a une fr. théorique particulièrement simple (éq. (3.21)). Soit $y = \max(u_1, u_2, ..., u_n)$, alors $x = y^n$ est elle-même une v.a.u.. Obtenant *k* séries différentes, chacune avec son maximum *y* et sa statistique $x = y^n$, un simple test de Kolmogorov-Smirnov (§6.7) basé sur la loi uniforme, soit $D_k = \sqrt{k}\, x \max \left| j/k - x_{(j)} \right|$, permet de se prononcer sur le comportement aléatoire du générateur.

6.29 *Le test spectral.* La plupart du temps, pour tester l'irrégularité d'une série de nombres ou d'observations, nous n'avons en mains que ces nombres eux-mêmes, soit en une série, soit quelques séries; nous pouvons procéder alors à des tests empiriques de l'hypothèse d'irrégularité sur ces séries. Dans certains cas cependant, nous disposons aussi de la source aléatoire elle-même, l'étude formelle de cette source pouvant nous

renseigner théoriquement sur ses propriétés aléatoires. Les générateurs de nombres pseudo-aléatoires utilisés dans les programmes informatiques constituent de telles sources et s'expriment par une fonction ou un groupe de fonctions mathématiques. L'une d'elles, le générateur linéaire récursif (cf §3.5):

$$y_{n+1} = (ay_n + c) \bmod m \,, \tag{6.44}$$

a inspiré un test théorique d'irrégularité, le test spectral.

Le test spectral (Coveyou et MacPherson 1965, voir Knuth 1969) consiste à analyser un générateur de type (6.44), avec des valeurs spécifiques des paramètres *{a, c, m }* . Le but est de se prononcer sur l'indépendance statistique des nombres successifs, en prouvant que tous les t-tuplets *{u1, u2, ..., ut}, t 2,* sont équiprobables; on reconnaît là un pendant théorique du test sériel (§6.13). Basée sur une application de la transformée de Fourier finie, la mathématique de ce test est assez touffue: Knuth (1969, p. 93-96) élabore les détails et Golder (1976) fournit un programme de calcul clef-en-main.

Considérons l'analyse des t-tuplets d'un générateur spécifique tel que (6.44), avec *t* fixé. Il s'agit de trouver les quantités entières $s_1, s_2, \ldots, s_t$ satisfaisant:

$$s_1 + s_2 a + s_3 a^2 + \cdots + s_t a^{t-1} = 0 \bmod m \,,$$

dont la somme quadratique $Q = s^2_1 + s^2_2 + + s^2_t$ est minimale; sont exclues les valeurs triviales $s_1 = 0$, $s_2 = 0$, etc. Alors $V_t = \sqrt{Q}$ indique la « puissance » du t-tuplet. On procède ainsi pour t = 2, 3, 4, ... Pour nous faire une idée de ce test, reprenons un exemple dans Knuth (1969, p. 86 et suiv.). Le générateur analysé est:

$$y_{n+1} = (3141592621 y_n + 1) \bmod 10^{10}$$

avec les paramètres $a = 3141592621$, c = 1 et $m = 10^{10}$. Pour les 2-tuplets (u_1, u_2), il faut $s_1 + 3141592621\, s_2 = 0 \bmod 10^{10}$. Knuth trouve $s_1 = 67654$ et $s_2 = 226$, $Q = s^2_1 + s^2_2$ est un minimum, et $V_2 = \sqrt{Q} \approx 67654{,}4$, la puissance des 2-tuplets. L'auteur interprète ce résultat V_t en termes de chiffres significatifs, à savoir que, pour un peu moins que 5 chiffres significatifs (ou $\log_2 V_2 \approx 16$ bits), les paires de nombres (u_1, u_2) sont statistiquement indépendants. De même, il trouve $V_3 \approx 1017{,}2$, $V_4 \approx 249{,}9$, $V_5 \approx 42{,}2$ et $V_6 \approx 23{,}3$. Dans ce dernier cas, par exemple, le test sériel (§6.13), appliqué aux 6-tuplets, détecterait la non-indépendance séquentielle si le domaine de la v.a.u. était découpé en 30 ou 50 intervalles; toutefois, pour l'appliquer avec 30 intervalles en respectant la norme *de ft* ≥ 5, le test sériel requerrait plus de 2×10^{10} v.a.u., si l'utilisateur est assez patient!

Knuth recommande fortement ce test là où il peut être appliqué, en particulier pour les générateurs pseudo-aléatoires servant à alimenter des études Monte Carlo intensives. Pour le citer (p. 82), « ce test est spécialement important puisque, non seulement tous les bons générateurs le passent avec succès mais aussi tous ceux qu'on sait maintenant être défectueux l'échouent. »

6.30 Quelques tests d'irrégularité parus dans la littérature n'ont pas été recensés dans le présent chapitre, notamment les « classiques » tests d'intervalle *(Gap test,* voir l'exercice 6.44) et de *poker* (Knuth 1969, p. 56-58). Certains de ces tests, qu'on retrace dès avant l'ère des premiers ordinateurs, sont devenus caducs et ont été transformés et généralisés dans les tests présentés ici. Marsaglia (1995) propose une batterie de tests clef-en-main, déjà programmés et disponibles sur CD-ROM. De plus, chacun peut avancer d'autres tests, des tests d'intérêt général et à saveur théorique, comme L'Écuyer (1998) qui en propose quelques-uns, ou élaborer un test particulier, conçu pour un contexte spécifique et approprié aux objectifs de son application: les quelques tests que nous avons décrits pourront servir d'inspiration pour cette entreprise.

En fin de compte, qu'est-ce qu'une série aléatoire? Que l'on examine une série spécifique de nombres ou d'observations, comme $\{x_1, x_2, ..., x_n\}$, ou bien qu'il s'agisse d'une source de nombres dont on veut tester la validité, les différents tests qu'on lui applique peuvent accroître ou détruire la confiance que l'on a dans une série donnée, confiance quant à la forme de distribution des nombres, à leur non-dépendance ou non-corrélation sérielle, à l'équiprobabilité des 2-tuplets ou 3-tuplets, etc. D'une part, on ne peut pas *prouver* qu'une source de nombres, encore moins une série particulière, est aléatoire (ou irrégulière) puisqu'il est possible d'inventer encore un test qui débusquera l'idiosyncrasie de la source, de la série. D'autre part, même si elle ne satisfait pas *tous* les tests (réalisés et réalisables), une source peut rester utile et valide pour certains usages: l'analyse rigoureuse du contexte mathématique de l'application envisagée nous permettra seule d'en décider.

Exercices

6.1 Utilisant les données de l'exemple 6.1, ré-appliquer un test du Khi-deux (X^2) à intervalles égaux (hormis les semi-intervalles aux extrémités), avec $k = 11$ intervalles (centrés sur $\bar{x}$) de largeur $0{,}4s$. Calculer les fréquences théoriques (a) en se basant sur les estimateurs $\bar{x}$ et s dérivés des données brutes; (b) en exploitant $\bar{x}_G$ et s_G, établis à partir du tableau des k fréquences de classes [les degrés de liberté sont alors diminués de $p = 2$ unités]. Comparer et interpréter les résultats.

6.2 La transformation $F(x_i)$, appliquée pour le test K-S (voir §6.7), convertit en quelque sorte la série des n variables aléatoires de loi f en une série de n v.a.u. Utilisant les données de l'exemple 6.1 et leurs valeurs $u_i = F(x_i)$, appliquer un test X^2 à intervalles (et probabilités) égaux, en k = 10 intervalles, pour tester l'hypothèse d'une distribution uniforme des $F(x_i)$.

6.3 Sur les mêmes 60 données u_i exploitées à l'exercice précédent, appliquer le test des moments 3 et 4, en l'interprétant à partir du tableau 6.1 ou du tableau 6.2.

6.4 Démontrer les expressions des moments (6.16) et (6.17).

6.5 *Loi multinomiale* $M(n;\pi_1,\pi_2,\cdots,\pi_k)$. Une variable qui, échantillonnée n fois, peut à chaque fois prendre l'une de k valeurs distinctes c_j avec probabilité π_j, $j = 1$ à k, est dite multinomiale, et l'ensemble des fréquences d'apparition des valeurs c_j, $(n_1, n_2, \ldots, n_k)$, est une réalisation de la loi multinomiale $M(n; \pi_1,\pi_2,\cdots,\pi_k)$ (cf. Johnson, Kotz et Balakrishnan 1997), avec probabililité:

$$\text{pr}(n_1, n_2, \ldots, n_k) = \text{M}_k(n;n_j)\ \Pi_{j=1}^{k}\ \pi_j^{n_j}\ , \tag{6.45}$$

où :
$$\text{M}_k(n;n_j) = \frac{n!}{n_1!n_2!\cdots n_k!} \tag{6.46}$$

est le coefficient multinomial.

Soit une série d'observations $(x_1, x_2, \ldots, x_n)$. Utilisant les variables indicatrices y telles que $y_i = 1$ si $x_i = c_j$ et $y_i = 0$ si $x_i \neq c_j$, $n_j = \sum_{i=1}^{n} y_i$, montrer que les moments simples et conjoints des n_j sont $\text{E}(n_j) = n\pi_j$, $\text{var}(n_j) = n\pi_j(1-\pi_j)$, $\text{cov}(n_j,n_l) = -n\pi_j\pi_l$ et $\rho(n_j,n_l) = -\{\pi_j\pi_l / [(1-\pi_j)(1-\pi_l)]\}^{1/2}$. Ainsi, chaque quantité n_j se comporte comme

une binomiale $B(n, л_j)$ autonome (cf. §4.4), si ce n'est que la somme des n_j est contrainte et égale à n, ce qui entraîne leurs corrélations mutuelles négatives.

6.6 Soit une multinomiale égalitaire $M(n;\frac{1}{k},\frac{1}{k},\cdots,\frac{1}{k})$, notée $M^*(n,k)$, issue par exemple de la répartition de n v.a. uniformes en k intervalles égaux, et $n_{max} = \max\{n_1, n_2, \ldots, n_k\}$, la fréquence maximale enregistrée. Montrer que (1) $\lceil n/k \rceil \le n_{max} \le n$; (2) $\text{pr}(n_{max}=1) = (k-1)!/[(k-n)!k^{n-1}]$, pour $0 < n \le k$; (3) $\text{pr}(n_{max}=r) = k \cdot b_r(n,\frac{1}{k})$, pour $r > \frac{1}{2}n$ et $b_r(\ )$ est une probabilité binomiale définie par (4.7). Noter que la distribution complète de n_{max} peut être obtenue par énumération, une méthode à cet effet étant indiquée en appendice du chapitre 11.

6.7 (Suite du précédent). L'étude empirique de la distribution et des moments de n_{max}, fait voir une v.a. plutôt lourde, qui reste tassée dans les basses valeurs. Soit une multinomiale de n composantes égalitaires échantillonnée *n fois, M*(n,n).* L'espérance de n_{max}. pour une loi *M*(n,n)* évolue approximativement selon :

$$E(n_{max}) \approx 1{,}369 + 0{,}6267 \log_e(n - \tfrac{3}{4})$$

(avec $r^2 > 0{,}9997$). La variance, $\sigma^2(n_{max})$, ne déborde pas 0,6 et les moments γ_1 et γ_2 indiquent une loi leptokurtique et d'asymétrie positive. Le tableau suivant indique des valeurs critiques (trouvées pour $2 \le n \le 50$):

P	(n)	n_{max}	(n)	n_{max}	(n)	n_{max}	(n)
0,95	(4)	3	(6)	4	(20)	5	(→50)
0,99	(5)	4	(9)	5	(27)	6	(→50)
0,999	(6)	5	(9)	6	(25)	7	(→50)

[Par exemple, avec $n = 30$ données, il faut obtenir $n_{max} \ge 5$ pour atteindre la significativité de *0,05,* $n_{max} \ge 6$ pour 0,01 et $n_{max} \ge 7$ pour 0,001.] Elaborer une procédure de test d'irrégularité locale basée sur ces informations.

6.8 (Suite des précédents). La distribution de probabilités du maximum n_{max} dans une multinomiale égalitaire peut être obtenue par énumération, ce qui permet de réaliser un test d'irrégularité globale utilisant le Khi-deux (§6.6). Le tableau en page suivante présente les distributions (partielles) de n_{max} pour trois configurations *M*(n,n),* ainsi que ses moments principaux. Élaborer une procédure de test d'irrégularité globale basée sur ces informations.

Distribution de probabilités de n_{max} (cf. exercice 6.8)

$n_{\max}$	pr[$n_{\max}$|M*(10,10)]	pr[$n_{\max}$|M*(20,20)]	pr[$n_{\max}$|M*(30,30)]
1	0,00036	~ 0	~ 0
2	0,39545	0,12135	0,03710
3	0,47724	0,58450	0,53244
4	0,11060	0,24303	0,34461
5	0,01488	0,04453	0,07362
6 +	0,00147	0,00659	0,01223
μ	2,74869	3,23124	3,49299
σ^2	0,51776	0,55360	0,55450
γ_1	0,754	0,700	0,845
γ_2	0,612	1,261	1,237

6.9 *Procédure du test de normalité (W) de Shapiro et Wilk.* La procédure suivante est indiquée dans Royston (1982) pour des séries statistiques comportant $n > 20$ éléments.

Étape 1. Pour obtenir les coefficients a_i, obtenir d'abord $b_i = 2\Phi^{-1}[(i-3/8)/(n+1/4)]$, pour $i = 2, 3, \cdots, n-1$, puis $\sum_{n-2} = b_2^2 + b_3^2 + \cdots + b_{n-1}^2$. Calculer $g(n)$, selon:

$$g(n) = \left[\frac{6n+7}{6n+13}\right]\left(\frac{2{,}71828}{n+2}\left[\frac{n+1}{n+2}\right]^{n-2}\right)^{1/2} ;$$

$b_n = -b_1 = \{g(n)/[1 - 2g(n)]\sum_{n-2}\}^{1/2}$. Les b_i ($i = 2$ à $n-1$) représentent en fait le double des espérances des statistiques d'ordre normales, estimées par l'inversion suggérée.

Étape 2. Calculer $\sum_n = \sum_{n-2} + 2b_n^2$, puis obtenir les coefficients a_i ($i = 1$ à n) normalisés par $a_i = b_i / \sqrt{\sum_n}$.

Étape 3. Appliquer la formule (6.19).

Étape 4. L'interprétation statistique exploite un écart-réduit d'une variable normalisée $y = (1 - W)^{\lambda}$, de moyenne μ_y et d'écart-type σ_y. Les paramètres λ, μ_y et σ_y sont donnés par:

Soit $q = \log_e n - 5$,

$$\lambda = 0{,}480385+0{,}318828q-0{,}0241665q^3+0{,}00879701q^4+0{,}002989646q^5$$

$$\log_e(\mu_y) = -1{,}91487-1{,}378880q-0{,}04183209q^2+0{,}1066339q^3 -0{,}03513666q^4-0{,}01504614q^5$$

$$\log_e(\sigma_y) = -3{,}73538 - 1{,}015807q - 0{,}331885q^2 + 0{,}1773538q^3 - 0{,}01638782q^4 - 0{,}03215018q^5 + 0{,}003852646q^6$$

La statistique résultante, $z = [(1 - W)^2 - \mu_y]/O_y$, s'interprète alors comme un écart-réduit normal.

À l'étape 1, l'inversion de la f.r. normale, $\Phi^{-1}(p)$, peut se faire précisément, au moyen d'une inversion par repérage (voir §4.17 et exemple 4.3) ou par une approximation rationnelle (voir exercice 4.12).

Utilisant les fonctions reproduites ci-haut, vérifier les valeurs de X, μ_y et O_y indiquées dans l'exemple 6.1.

6.10 La statistique D fut proposée par D'Agostino (1971) comme substitut plus simple du W de Shapiro et Wilk. A l'instar de $\sqrt{W}$, c'est un quotient d'estimateurs de O, soit:

$$D = T / (n^2 S) ,$$

où $S = s\sqrt{[(n-1)/n]}$ est l'estimateur biaisé commun, et T/n^2 est un estimateur ordonné de σ, où $T = \sum_{i=1}^{n} [i - \frac{1}{2}(n+1)]x_{(i)}$. La variable $z = \sqrt{n}(33{,}34892D - 9{,}407556)$ est asymptotiquement normale. L'auteur présente des valeurs critiques approximatives, reproduites en partie en tableau, en contrebas.

Montrer que, pour les données de l'exemple 6.1, $D \approx 0{,}26761$ et $z \approx -3{,}742$, une valeur diablement proche de la valeur normalisée du W (= 3,742). Pour $n = 60$, à $\alpha = 0{,}01$ bilatéral, la valeur critique interpolée à $P = 0{,}005$ est d'environ $-3{,}876$, amenant l'acceptation (au seuil α choisi) de l'hypothèse d'une distribution normale pour ces données.

Centiles approximatifs de la statistique D (D'Agostino)

$n \setminus P$	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
50	−3,949	−3,442	−2,757	−2,220	0,923	1,038	1,140	1,192
100	−3,584	−3,150	−2,552	−2,075	1,137	1,303	1,470	1,569
150	−3,409	−3,009	−2,452	−2,004	1,223	1,423	1,623	1,746
200	−3,302	−2,922	−2,391	−1,960	1,290	1,496	1,715	1,853
250	−3,227	−2,861	−2,348	−1,926	1,328	1,545	1,779	1,927
300	−3,172	−2,816	−2,316	−1,906	1,357	1,528	1,826	1,983
400	−3,094	−2,753	−2,270	−1,873	1,396	1,633	1,893	2,061
500	−3,040	−2,709	−2,239	−1,850	1,423	1,668	1,938	2,114
750	−2,956	−2,640	−2,189	−1,814	1,465	1,722	2,010	2,199
1000	−2,906	−2,599	−2,159	−1,792	1,489	1,754	2,052	2,249

6.11 Montrer que l'égalité $p(X_1, X_2, \cdots, X_n) = p(X_1)p(X_2) \cdots p(X_n)$ implique l'indépendance de distribution, soit par exemple $p(X_1 \mid X_2) = p(X_1)$. Noter cependant que l'inverse n'est pas vrai.

6.12 Soit le coefficient de corrélation sérielle R_n, calculé à partir de n variables enchaînées et bouclées: montrer que, généralement, $R_2 = -1$ et $R_3 = -\frac{1}{2}$, à la seule condition que les 2 ou 3 variables impliquées soient distinctes.

6.13 *Valeurs critiques de corrélation sérielle.* Dans le calcul du coefficient de corrélation sérielle R, selon la procédure de calcul enchaîné et bouclé, l'utilisation égale des n variables au numérateur permet d'en simplifier la formule. Soit $\bar{x}$, la moyenne des n variables x_i, et $y_i = x_i - \bar{x}$, alors:

$$R = \frac{y_1y_2 + y_2y_3 + \cdots + y_{n-1}y_n + y_ny_1}{y_1^2 + y_2^2 + \cdots + y_n^2} .$$

La distribution exacte de R est connue dans le cas de v.a. normales (Anderson 1942). Exploitant les valeurs critiques présentées en tableau, sur la page suivante, vérifier si la série de données suivante, réputées normales, est vraiment « au hasard » (utiliser un seuil α = 5 % bilatéral):

{ 21,8; 21,5; 21,8; 20,9; 20,0; 19,7; 22,0; 24,7; 26,4; 27,3; 33,8; 39,0 }.

6.14 Utilisant la série de données de l'exercice précédent, calculer R′ et en vérifier la significativité par le recours à une approximation normale. Supposant que $\gamma_2(R') \approx -0{,}24$ pour $n = 12$, construire le modèle approprié basé sur une loi *Bêta* symétrique (voir §6.16) et l'utiliser pour tester encore la significativité.

6.15 Soit $\bar{u}$, la moyenne arithmétique de n v.a. uniformes $U(0,1)$. La f.r. de $y = (\bar{u} - \frac{1}{2})^2$, pour $n = 4$, est $F_1(y) = (128/3)[\frac{1}{8}\sqrt{y} - y^{3/2} + 1{,}5y^2]$, $0 \le y < 1/16$, et $F_2(y) = (128/3)[\frac{1}{4}\sqrt{y} - 0{,}75y + y^{3/2} - \frac{1}{2}y^2] - \frac{1}{3}$, $1/16 \le y < \frac{1}{4}$ (Laurencelle 1993). L'espérance de $12n \cdot y$ est 1, la variance 2,125 et la médiane, obtenue par $F^{-1}(\frac{1}{2})$, est 0,48657. Les indices de forme sont $\gamma_1 \approx 2{,}178$ et $\gamma_2 \approx 5{,}749$, reflétant une distribution pointue et très asymétrique. Utilisant la f.r. ci-dessus, composer un programme informatique qui réalise un test de Kolmogorov-Smirnov sur la variance centrée de 4 v.a. uniformes.

6.16 Établir l'équivalence algébrique entre les statistiques R′ et QVP.

Centiles du coefficient de corrélation sérielle R (données normales) (voir exercice 6.13)†

$n \setminus P$	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
5	−,803	−,798	−,781	−,753	,253	,281	,298	,303
6	−,902	−,863	−,788	−,708	,345	,402	,447	,467
7	−,829	−,799	−,740	−,674	,370	,444	,510	,543
8	−,809	−,764	−,691	−,625	,371	,453	,531	,573
9	−,779	−,737	−,664	−,593	,366	,449	,532	,580
10	−,747	−,705	−,634	−,565	,360	,440	,525	,576
11	−,724	−,679	−,607	−,539	,354	,431	,515	,567
12	−,700	−,656	−,584	−,516	,348	,423	,505	,556
13	−,678	−,634	−,563	−,497	,341	,415	,494	,545
14	−,659	−,615	−,544	−,478	,335	,406	,485	,534
15	−,641	−,597	−,527	−,462	,328	,398	,475	,524
20	−,567	−,525	−,459	−,399	,299	,362	,432	,477
25	−,513	−,473	−,411	−,356	,276	,333	,398	,440
30	−,472	−,433	−,375	−,324	,257	,310	,370	,410
40	−,412	−,377	−,325	−,279	,229	,276	,329	,365
50	−,369	−,338	−,290	−,248	,208	,251	,299	,332
60	−,338	−,308	−,264	−,226	,192	,231	,276	,306
70	−,313	−,286	−,244	−,208	,180	,216	,258	,286

† Valeurs établies d'après les formules d'Anderson (1942). Les présentes valeurs corrigent plusieurs petites inexactitudes dans le tableau 1 d'Anderson (p. 8). Au delà de $n = 70$, le calcul selon la méthode d'Anderson devient aléatoire; utiliser plutôt l'approximation normale $R_P = -1/(n-1) + z_P\sigma$, où $\sigma^2 = [n(n-3)] / [(n+1)(n-1)^2]$ et z_P est l'écart-réduit normal d'intégrale P.

6.17 Utilisant les données de l'exercice 6.13, calculer QVP et en tester la significativité.

6.18 *QVP sur les rangs.* La variance permutative s^2_p (6.22), au lieu d'être mesurée sur la série des valeurs x_i, peut l'être aussi sur les rangs de x, r_i = rang(x_i): le test devient a-distributionnel puisque, alors, la valeur de s^2_p dépend uniquement de la permutation des x_i. De plus, la variance s^2_p des rangs, toujours les mêmes (i.e. 1, 2, ..., n), est constante et égale à $n(n+1)/12$: le quotient QVP_{rangs} a donc la même distribution que s^2_p.

Ces statistiques n'ont, semble-t-il, jamais été étudiées. L'espérance de QVP est encore 1. La variance tend vers $1/n$, plus précisément vers $n/(n+1)^2$, et la distribution approche la forme normale, de sorte que, dès $n \geq 50$, la quantitié $1 + z_\alpha\sqrt{n/(n+1)}$tient lieu du centile de QVP à l'intégrale α, z_α étant le centile $N(0,1)$ correspondant. Noter une platykurtose ($\gamma_2 < 0$) assez forte pour petits n, de même qu'une asymétrie positive légère ($\gamma_1 > 0$) dès $n \geq 5$ et s'atténuant au delà.

Le minimum de QVP_{rangs} étant $6/[n(n+1)]$ et son espérance 1, montrer que le maximum tend vers *2* et est approximativement égal à $[2n^3 - 3n^2+25n - 84] / [n(n^2 -1)]$, et que la variance tend approximativement vers $[991\, n^3+733n^2 - 3798n - 3960] / [1000n^2(n+1)^2]$.

6.19 *Conversion en rangs.* Plusieurs tests statistiques requièrent que les données x_i d'une série soient d'abord converties en *rangs,* c'est-à-dire en une permutation des entiers 1, *2,...,* n manifestant les mêmes relations d'ordre que les valeurs x_i. Par exemple, la série réelle *{45,6; -9,5;* 11,8; 19,1} sera convertie en la série de rangs {4; 1; 2; 3}. Cette conversion en rangs n'est pas une opération triviale. L'algorithme le plus direct est le suivant:

Conversion en une série de rangs: méthode directe

{ Soit un vecteur donné { X_i } et le vecteur à produire { R_i }, $i = 1, n$ }

Pour $i = 1$ à n Faire $R_i \leftarrow 1$;
Pour $i = 1$ à $n - 1$ Faire
  $k \leftarrow 0$;
  Pour $j = i+1$ à n Faire Si $X_i > X_j$ alors $k \leftarrow k+1$
    sinon $R_j \leftarrow R_j+1$
  $R_i \leftarrow R_i+k$

Les cas d'égalité des données x_i, s'il y a lieu, sont traditionnellement traités en ajoutant ½ à chaque valeur de rang. Cette solution est contre-indiquée pour les fins d'établir une permutation des n premiers entiers naturels: on ajoutera plutôt +1 à R_i ou R_j selon une décision aléatoire, prise comme au jeu de Pile ou Face.

Montrer que le coût d'exécution de cet algorithme est proportionnel à n^2.

6.20 *Conversion en rangs.* Outre la méthode directe présentée à l'exercice précédent, la conversion en rangs peut s'effectuer en deux phases, une première consistant à mettre en ordre la série des données x_i accompagnées d'un vecteur-jumeau contenant les entiers 1 à *n,* puis une seconde, consistant à inverser la série-jumelle des *n* entiers, produisant ainsi le vecteur-rangs. En voici l'algorithme:

Conversion en une série de rangs: méthode en deux phases

{ Soit un vecteur donné { X_i }, son jumeau { J_i } et le vecteur à produire { R_i }, $i = 1$ à n }

Pour $i = 1$ à n Faire $J_i \leftarrow i$;
Trier ensemble les vecteurs { X_i } et { J_i } selon les valeurs X_i;
Pour $i = 1$ à n Faire $R[J_i] \leftarrow i$

Prouver la validité de cet algorithme. L'efficacité de l'algorithme dépend de la technique de tri employée, du code programmé lui-même, et elle varie aussi selon *n.*

Fabriquer trois programmes, l'un utilisant la conversion directe de l'exercice 6.19, un autre la méthode en deux phases avec un tri par *insertion avec sentinelle,* un dernier avec un tri de mode *Quicksort:* ces deux procédures de tri sont détaillées plus loin, en appendice du chapitre. Déterminer pour quelles valeurs de *n* chaque programme est optimal.

6.21 *Le numéro d'une permutation.* Comme il y a *n!* permutations de *n* objets, il doit exister *n*! numéros différents pour les identifier: plusieurs techniques existent à cet effet. L'une d'elles consiste à attribuer à une permutation le numéro de son rang dans l'ordre d'énumération des permutations. Par exemple, selon l'ordre dit lexicographique (dans lequel les éléments de rang moins élevé se déplacent moins vite), nous aurions, pour $n = 4$: 1=(1234), 2=(1243), 3=(1324), ..., 23=(4312), 24=(4321). Cette numérotation, basée sur la structure récursive de l'énumération, implique un coût de calcul proportionnel à *n* pour chaque permutation, coût souvent excessif pour les applications répétitives. Élaborer un algorithme qui calcule directement le numéro lexicographique d'une permutation quelconque.

6.22 *Le numéro d 'une permutation* (suite). La méthode de numérotation la plus expéditive est celle d'un codage à base B fixe, B n: ainsi, utilisant $n = 4$ et B = 10, la permutation (1234) aurait le numéro 1234, etc. jusqu'au numéro 4321 pour la permutation (4321); le coût

temporel d'exécution est ici proportionnel à 1. Un tel système, peut-être satisfaisant dans certains contextes, n'est cependant pas compact (ce qui peut représenter un inconvénient dans le contexte d'un calcul par tableau indicé): il utilise un intervalle de (4321 — 1234 + 1) = 3088 nombres pour en identifier 24, un ratio d'environ 128:1 pour n = 4. Construire un programme de numérotation par base et comparer l'efficacité temporelle de celui-ci avec le programme de numérotation lexicographique, à l'exercice précédent. Montrer qu'en utilisant B = n, le numérateur du ratio de compacité approche n^{n-1}/n!

6.23 *Le numéro d'une permutation (suite).* Dans une numérotation à base B ≥ n pour les permutations de n objets, établir la table des correspondances entre le numéro de permutation de base B et le numéro lexicographique. Compléter le programme élaboré à l'exercice précédent afin qu'il délivre (indirectement) le numéro lexicographique; comparer l'efficacité temporelle de ce programme à celle d'un programme de numérotation lexicographique directe.

6.24 *Le numéro d'une permutation (Gilles Brassard).* Peut-on numéroter une permutation de n objets en temps 1, mais avec un système de numérotation compact? Gilles Brassard, professeur à l'Université de Montréal, propose l'algorithme suivant, basé sur l'exploitation de la permutation inverse (Knuth 1969):

Numérotation compacte d'une permutation

{ Soit un vecteur { X_i }, i = 1 à n, contenant une permutation des entiers 1 à n, un vecteur de travail { h_i }, i = 1 à n, et R = R{X}, le numéro à trouver }

Pour i = 1 à n Faire $h_i[X_i] \leftarrow i$;
R ← 0 ; p ← 1 ;
Pour i = 1 à n − 1 Faire R ← R + p × (h_i − i) ;
p ← p × (n+1−i) ;
$X[h_i] \leftarrow X_i$; $h[X_i] \leftarrow h_i$

Montrer que le numéro de permutation R est unique et compact, *i.e.* qu'il donne lieu à un ratio 1/1. Comparer l'efficacité temporelle du programme appliquant cette numérotation au programme le plus performant de l'exercice précédent.

6.25 Il existe plusieurs mesures possibles liées aux propriétés des suites monotones d'une série de n observations distinctes. Le nombre (S) de suites alternées (ascendantes et descendantes) s'évalue simplement:

Dénombrement des suites monotones alternées

{ Soit un vecteur { X_i }, i = 1 *à n* (contenant des valeurs distinctes) et S, le nombre à trouver }

S ← 1 ;
D ← signe[$X_2 - X_1$] ;
Pour i = 2 à n Faire D1 ← signe[$X_i - X_{i-1}$] ;
Si D1 ≠ D alors S ← S+1; D ← D1

Modifier l'algorithme ci-dessus afin de déterminer L_{max}, la longueur de suite maximale rencontrée dans la série.

6.26 Utilisant la définition déjà donnée de S^+ pour les suites ascendantes et une définition pareille de S^- pour les suites descendantes, montrer que $S^+ + S^- = n+1$ et, par symétrie, $\mu(S^+) = \mu(S^-) = (n+1)/2$. Knuth (1975) présente une fonction d'Euler, $K_n(S^+)$, qui calcule le nombre de permutations (parmi les $n!$ permutations possibles) contenant x (= S^+) suites ascendantes, soit:

$$K_n(x) = \sum_{j=0}^{x-1} (-1)^j (x-j)^n \binom{n+1}{j} , \qquad (6.47)$$

une fonction symétrique telle que $K_n(x) = K_n(n+1-x)$ et $\sum_{x=1}^{n} K_n(x) = n!$. Montrer que les moments pairs de S^+ (ou x) sont: $\sigma^2 = (n+1)/2$ et $\gamma_2 = -1{,}2/(n+1)$ pour $n \geq 4$.

6.27 Montrer que la longueur (L_1) de la première suite ascendante d'une série de n v.a. aléatoires a pour distribution de probabilité $pr(x) = 1/x! - 1/(x+1)!$, $1 \leq x < n$, et $pr(n) = 1/n!$. De là, montrer que les moments sont, approximativement pour n fort, $\mu \approx e-1 \approx 1{,}718282$, $\sigma^2 \approx e(3-e) \approx 0{,}765789$, $\gamma_1 \approx 1{,}271958$ et $\gamma_2 \approx 1{,}705391$.

6.28 Un autre algorithme utile à l'étude des suites monotones consiste à dénombrer les suites isolées (p.ex. ascendantes), selon leurs longueurs 1, 2, etc., afin d'en comparer la fréquence d'apparition à la fréquence attendue, au moyen d'un test du Khi-deux. Limitant à 6 la longueur maximale retenue, nous aurions par exemple:

Fréquence des suites isolées (ascendantes) selon leur longueur

{ Soit un vecteur X_i, i = 1 à n+1

et H[1] à H[6], l'histogramme des longueurs à mettre à jour, toutes les séries de longueurs L ≥ 6 étant inscrites en H[6] }

```
Pour L = 1 à 6  Faire H[L] ← 0; i ← 1;
Répéter  L ← 0 ; X[n+1] ← X[i] ;
         Répéter  L ← L + 1 ; i ← i + 1
         Jusqu'à  X[i] ≤ X[i−1] ;
         Si L < 6  alors  H[L] ← H[L] + 1
                   sinon H[6] ← H[6] + 1
Jusqu'à i > n .
```

Déterminer d'après §6.18 les probabilités relatives à chaque longueur L : ici pr(6) devra inclure la somme des probabilités pour L = 6 et plus. Quelle longueur minimale de série *(n)* faut-il pour s'assurer que $S^{(+)}$ x pr(6) ≥ 1 (« 1» est la fréquence théorique minimale recommandée pour le test Khi-deux), $S^{(+)}$ étant le nombre total de suites isolées. Rédiger un programme complet qui effectue le test du X^2 sur une série quelconque de *n* observations.

6.29 Montrer que les moments de la longueur L des suites isolées, dans une série aléatoire à *n* élevé, sont les mêmes que ceux de L_1, la longueur de la première suite simple (cf exercice 6.27).

6.30 *La distribution de débordement.* La probabilité qu'une r+1^e variable déborde, *i.e.* excède le maximum d'une série de *r* variables, soit $\mathrm{pr}[x_{r+1} > M_r(x)]$, où $M_r(x) = \max(x_1, x_2, \cdots, x_r)$, est la même que $\mathrm{pr}[u_{r+1} > M_r(u)]$, $u \sim U(0,1)$, étant donné la correspondance $u = F(x)$ où F est la f.r. de la variable x. De plus, le nombre n de nouvelles v.a. requises [*e.g.* $x_r+1 \le M_r(x), \cdots, x_{r+n-1} \le M_r(x), x_{r+n} > M_r(x)$] pour déborder $M_r(x)$ suit une loi géométrique (§4.5), soit $n \sim G(1-y) = y^{n-1}(1-y)$, où $y = F\{M_r(x)\}$; or la variable y, représentant le maximum de *r* v.a. uniformes, a pour densité ry^{r-1} (cf. (3.21)). Par composition, la fonction de masse de n devient alors:

$$\dot{g}(n) = \int_0^1 G(n) r y^{r-1}\, dy \qquad (6.48)$$

$$= \int_0^1 r \cdot y^{n+r-2}(1-y)\, dy = r \,/\, [(n+r)(n+r-1)]\,,$$

en intégrant par parties. Montrer que la f.r. correspondante est $n/(n+r)$ et que cette distribution n'a pas de moments.

6.31 *La distribution de débordement* (suite). Soit $u^{(1)} = u_1 = F(x_1)$, une première variable, et $u^{(2)}$, la n_1+1 ^{e} variable et la première *à déborder* $u^{(1)}$. Pour des valeurs données de n_1 et $u^{(2)}$, chaque nouvelle v.a.u. tombe dans l'intervalle $(0, u^{(2)}]$ selon la probabilité égale à $u^{(2)}$, ou bien elle devient un nouveau maximum $u^{(3)}$ selon la probabilité 1-$u^{(2)}$, en tombant dans l'intervalle $(u^{(2)},1]$. Intégrant sur $u^{(2)}$, qui est

le maximum de n_1+1 v.a.u. et additionnant enfin pour toutes les valeurs de n_1, nous obtenons la probabilité de déborder le second maximum cumulatif à la *n;* variable:

$$p(n_2) = \rho_{n_1 \geq 1} \int_0^1 u_1^{n_1-1} \int_0^1 u_2^{n_2-1}(1-u_2)\, du_1\, du_2 \qquad (6.49)$$

$$= \sum_{n_1 \geq 1} \frac{1}{n_1(n_1+n_2)(n_1+n_2+1)} .$$

ainsi, p(l) = 1/4, *p(2)* = 5/36, p(3) = 13/144, etc. Trouver une expression pour établir directement $P(n_2)$, *la* probabilité cumulative associée à cette distribution.

6.32 *La distribution du r^e maximum (« record value »).* Soit une v.a. continue x, de densité $f(x)$ et de f.r. $F(x)$, que nous échantillonnons jusqu'à l'obtention de y_r, le r^e maximum tel que $y_1 = x_1$, $y_2 = x_{1+n_2} > \{x_1, ..., x_{n_2}\}$, ... $y_r = x_{1+n_2+...+n_r}$. La densité de y_r est donnée par:

$$g_{y_r}(x) = [(r-1)]^{-1}\{-\log(1-F(x)\}^{r-1} f(x) . \qquad (6.50)$$

Pour une v.a. uniforme $u \sim U(0,1)$ telle que $f(u) = 1$ *et* $F(u) = u$, les fonctions de densité et de répartition sont respectivement:

$$g_{y_r}(u) = [(r-1)]^{-1}\{-\log(1-u)\}^{r-1} \qquad (6.51a)$$

$$G_{y_r}(u) = u - (1-u) \sum_{i=1}^{r-1} [-\log_e(1-u)]^i / i! . \qquad (6.51b)$$

Montrer que, pour des v.a. uniformes $U(0,1)$, l'espérance (μ) de la valeur du r^e maximum, y_r, est égale à 1— $(½)^r$, et que les moments, pour $r = 1$ à 6, sont tels qu'au tableau suivant [noter que $y_r = u$] :

r	μ	σ^2	γ_1	γ_2
1	0,50000	0,08333	0,00000	−1,20000
2	0,75000	0,04861	−0,97191	0,15184
3	0,87500	0,02141	−1,80083	3,38486
4	0,93750	$0,0^28439$	−2,68247	9,17305
5	0,96875	$0,0^23139$	−3,70676	19,24918
6	0,98438	$0,0^21128$	−4,95105	36,77379

6.33 *La distribution du r^e* maximum (suite). Exploitant la f.r. (6.5 lb) donnée plus haut, il est facile de déterminer des centiles servant de valeurs critiques pour tester la valeur du r^e maximum cumulatif d'une v.a. uniforme et, partant, d'une variable x de distribution quelconque, en utilisant la transformation inverse $x_{(r)} = F^{-1}(u_{(r)})$. Vérifier les centiles suivants d'une variable $u_{(r)}$, pour $r = 2$ *à* 8:

$r \setminus P$	,001	,005	,01	,025	,05	,5	,95	,975	,99	,995	,999
2	,04439	,09832	,13805	,21511	,29908	,81332	,99130	,99620	,99869	$,9^3407$	$,9^4023$
3	,17348	,28671	,35341	,46134	,55855	,93103	,99816	$,9^3272$	$,9^3776$	$,9^4061$	$,9^4867$
4	,34855	,48942	,56100	,66374	,74496	,97458	$,9^3571$	$,9^3844$	$,9^4566$	$,9^4829$	$,9^5788$
5	,52259	,65970	,72171	,80279	,86056	,99064	$,9^3894$	$,9^4643$	$,9^5088$	$,9^5661$	$,9^6624$
6	,66949	,78496	,83225	,88941	,92669	,99655	$,9^4728$	$,9^5144$	$,9^5797$	$,9^6284$	$,9^7286$
7	,78136	,86962	,90272	,94006	,96257	,99873	$,9^5281$	$,9^5787$	$,9^6530$	$,9^6842$	$,9^7857$
8	,86066	,92355	,94531	,96838	,98133	$,9^3533$	$,9^5805$	$,9^6455$	$,9^6887$	$,9^7638$	$,9^8700$

6.34 *La longueur maximale des suites de type 1, $L_{max:1}$.* Soit une série de n éléments donnés comportant n_1 éléments d'un type et n_2 d'un autre ou de plusieurs autres types, avec $n_1 + n_2 = n$. Bradley (1968) montre que la probabilité $P_1(L)$ d'observer au moins une suite de type «1» de longueur L ou plus est:

$$P_1(\mathrm{L}) = \binom{n}{n_1}^{-1} \sum_{i=1}^{n_1/L} (-1)^{i+1} \binom{n_2+1}{i} \binom{n - iL}{n_2} . \qquad (6.52)$$

À l'aide de l'expression (6.52), reprendre l'exemple donné en §6.22 et vérifier si l'obtention d'une suite de 5 mâles, dans une série comportant 21 mâles et 13 femelles, constitue un événement de hasard exceptionnel.

6.35 *La longueur maximale d'une suite binaire,* L_{max}. Bradley (1968) propose un calcul ingénieux de la probabilité d'observer une suite d'au moins L éléments de même type dans une série binaire comportant n_1 et n_2 éléments. La formule globale:

$$P(\mathrm{L}) = P_1(\mathrm{L}) + P_2(\mathrm{L}) - P_{1,2}(\mathrm{L}) \qquad (6.53)$$

totalise les probabilités d'observer (au moins) une suite-L de types 1 ou 2 et en soustrait la probabilité que les deux types en présentent conjointement. La quantité $P_1(L)$, définie par (6.52), se transpose en $P_2(L)$, *mutatis mutandis.* Quant à $P_{1,2}(L)$, elle est:

$$P_{1,2} = \binom{n}{n_1}^{-1} \sum_{s_1=1}^{n_1-\mathrm{L}+1} \left\{ A(s_1)\left[B(s_1) + 2C(s_1) + D(s_1) \right] \right\} , \qquad (6.54)$$

où :

$$A(s_1) = \sum_{i=1}^{(n_1-s_1)/(\mathrm{L}-1)} (-1)^{i+1} \binom{s_1}{i} \binom{n_1 - 1 - i(\mathrm{L}-1)}{s_1 - 1}$$

$$B(s_1) = \sum_{i=1}^{(n_2-s_1+1)/(\mathrm{L}-1)} (-1)^{i+1} \binom{s_1 - 1}{i} \binom{n_2 - 1 - i(\mathrm{L}-1)}{s_1 - 2}$$

$$C(s_1) = \sum_{i=1}^{(n_2-s_1)/(\mathrm{L}-1)} (-1)^{i+1} \binom{s_1}{i} \binom{n_2 - 1 - i(\mathrm{L}-1)}{s_1 - 1}$$

$$D(s_1) = \sum_{i=1}^{(n_2-s_1-1)/(\mathrm{L}-1)} (-1)^{i+1} \binom{s_1 + 1}{i} \binom{n_2 - 1 - i(\mathrm{L}-1)}{s_1} .$$

Considérant que $\mathrm{pr}(\mathrm{L} \mid n_1, n_2) = P(\mathrm{L}) - P(\mathrm{L}+1)$, montrer que, pour $n_1 = n_2$, l'espérance de L augmente logarithmiquement, selon $\mu(\mathrm{L}) \approx 0{,}039 + 1{,}7 \log_e(n_1 + 0{,}7)$.

6.36 Soit une série d'éléments donnés à k types et à composition (n_1, n_2, ···, n_k) et $\mathrm{T}_k(\mathrm{S} \mid n_j) = \mathrm{T}(\mathrm{S} \mid n_1, n_2, \cdots, n_k)$, le nombre d'arrangements différents de la série présentant S suites. La quantité $\mathrm{T}_k(\mathrm{S} \mid n_j)$ satisfait la récurrence (Saughnessy 1981):

$$\mathrm{T}(v+r \mid n_1, n_2, \cdots, n_v) = \sum_{t=1}^{r+1} \left\{ \mathrm{T}(v+r-t \mid n_1, n_2, \cdots, n_{v-1}) \right. \qquad (6.55)$$

$$\left. \times \sum_{j=0}^{n-n_v-v+1} \binom{n-n_v-v-r+t}{j} \binom{n_v - 1}{t-1-j} \binom{v+r-t+1}{t-2j} \right\} .$$

Afin de trouver $\mathrm{T}_k(\mathrm{S} \mid n_j)$ depuis le niveau $v = k$, c-à-d. avec l'invocation $\mathrm{T}(\mathrm{S} \mid n_1, n_2, \cdots, n_k)$ et $n = \sum_{j=1}^{k} n_j$, la descente récursive opère en réduisant v à $k-1$, puis $k-2$ etc., chaque fois avec l'invocation $\mathrm{T}(\mathrm{S} \mid n_1, n_2, \cdots, n_v)$ et $n = \sum_{j=1}^{v} n_j$. Lorsque enfin $v = 1$, $\mathrm{T}(\mathrm{S} \mid n_1)$, qui dénote le nombre d'arrangements de n_1 éléments produisant S suites, vaut 1 pour S = 1 et vaut 0 dans tous les autres cas. Laurencelle (2000, p. 174) présente un programme en Pascal qui réalise ces opérations.

Étant donné qu'une série de n éléments à composition $(n_1, n_2, ..., n_k)$ donne lieu *à* $M_k(n;n_j)$ arrangements distincts, la quantité $pr(S) = T_k(S \mid n_j) / M_k(n ; n_j)$ désigne la probabilité d'observer S suites dans un arrangement aléatoire de cette série. Utiliser ces informations pour composer un programme qui repère les valeurs critiques à quelques seuils a pour une série à composition $(n_1, n_2, ..., n_k)$ donnée.

6.37 Soit une série d'éléments donnés à k types et à composition $(n_1, n_2,...,n_k)$, cette dernière étant arrangée de telle sorte que $n_1 \geq n_2 \geq ... \geq n_k$. Montrer que le nombre d'arrangements de la série produisant $L_{max.} = n_1$, la valeur la plus haute de L_{max} , est (Laurencelle 2000):

$$C(n_1) = \sum_{i=1}^{t} (-1)^{i+1}\, i! \binom{t}{i} \binom{n - \sum_{j=1}^{i} n_j + i}{i} \frac{\left(n - \sum_{j=1}^{i} n_j\right)!}{\prod_{j=i+1}^{k} n_j!} , \qquad (6.56)$$

expression dans laquelle t dénote le nombre de valeurs n_j égales à n_1, *i.e.* tel que $n_1 = n_2 = ... = n_t$, $t \geq 1$. De plus, admettant $n_1 = \max[n_j, j\neq]$ + d, $d>0$, on obtient facilement le nombre $C(n_1 - u)$, pour $u \leq \min(d-1, n_1/2+1)$, en utilisant:

$$C'(L \geq n_1 - u) = \frac{n - n_1 + 1}{u!} \times \frac{(n - n_1 + u)!}{\prod_{j=2}^{k} n_j!} , \qquad (6.57)$$

puis en calculant $C(n_1—u)$ = C'(L ≥ n_1-u)—C'(L ≥ n_1—u+l).

6.38 *Moments de R (ou C_2) dans une série multinomiale égalitaire.* Le nombre de répétitions (R) et le nombre de chaînes de longueur *2 (C_2)* contrevarient exactement avec le nombre total de suites (S), selon R = C_2 = n — S. Utilisant les moments (6.40) de S pour une série multinomiale égalitaire ou les moments binomiaux (exercice 4.2) avec n' = n -1 et л = 1/k, montrer que les moments de R et C_2 sont: μ = (n-1)/k, O^2 = (n—1)(k—1)1k^2,y_1 = (k—2)/ √ (n—1)(k—1)],y_2 = (k^2 — 6k+6)/[(n -1)(k—1)].

6.39 *Nombre de chaînes de longueur r, C_r.* Montrer par induction que, dans une série de n éléments k-nomiaux équiprobables, l'espérance de C_r, le nombre de chaînes (ou suites enchaînées) de longueur r, est:

$$E\{ C_r(n,k) \} = (n - r + 1) / k^{r-1} . \qquad (6.58)$$

6.40 *Nombre de suites de longueur r, S_r.* Une chaîne de longueur r contient 1 suite de longueur r, 2 de longueur $r-1$, etc. Considérant les relations suivantes:

$$C_r = \sum_{t\geq 1} t \times S_{r-1+t} \; ; \; S_r = C_r - 2C_{r+1} + C_{r+2} \qquad (6.59)$$

et utilisant (6.58), montrer que l'espérance de S_r dans une série multinomiale égalitaire obéit à:

$$\begin{aligned} E\{S_r(n,k)\} &= [(n-r)(k-1)^2+(k^2-1)] \,/\, k^{r+1} \,, \quad \text{pour } 1\leq r\leq n-2 \qquad (6.60) \\ &= 2(k-1) \,/\, k^{n-1} \qquad\qquad r = n-1 \\ &= 1 \,/\, k^{n-1} \qquad\qquad r = n \end{aligned}$$

6.41 En s'inspirant de l'exemple en §6.16, établir un modèle *Bêta* symétrique, ou $\beta(p,p)$, pour la distribution du nombre d'inversions dans une série de $n = 15$ variables (incorporer une correction de continuité, *i.e.* $q \pm \frac{1}{2}$). Reformuler le modèle *Bêta* pour la distribution correspondante du τ de Kendall.

6.42 La distribution χ^2 tend lentement vers la loi normale quand les degrés de liberté ν croissent. Voici trois fonctions (T_i) qui estiment approximativement la f.r. d'une variable x, de loi χ^2_ν, à partir de la f.r. normale, Φ. (1) $T_1 = x$, $\mu = \nu$, $\sigma^2 = 2\nu$; (2) $T_2 = \sqrt{(2x)}$, $\mu \approx \sqrt{(2\nu-1)}$, $\sigma^2 \approx 1$; (3) $T_3 = (x/\nu)^{1/3}$, $\mu \approx 1 - 2/(9\nu)$, $\sigma^2 = 2/(9\nu)$. En évaluant $F(x)$ par $\Phi[(T-\mu)/\sigma]$, montrer que pour des degrés de liberté ν subasymptotiques, l'erreur des fonctions approximatives observe l'inégalité $\varepsilon(T_1) > \varepsilon(T_2) > \varepsilon(T_3)$, et que la fonction T_3 est satisfaisante à toutes fins pratiques.

6.43 Appliquer le *test spectral* au petit générateur linéaire récursif de l'exemple 3.2, à savoir $y_{n+1} = (8021y_n + 1) \bmod 10000$. Montrer par une méthode ou l'autre que $V_2 \approx 97{,}3$ ($s_1 = -16$; $s_2 = 96$), $V_3 \approx 18{,}3$ ($s_1 = -1$, $s_2 = 3$, $s_3 = 18$), $V_4 \approx 10{,}9$ ($s_1 = 2$, $s_2 = 4$, $s_3 = s_4 = 7$).

6.44 *Le test d'intervalle.* Le test d'intervalle (*Gap test*) vérifie le taux d'incidence des v.a. x_i consécutives dans un intervalle (B_1, B_2) de capacité γ [$\gamma = F(B_2) - F(B_1)$, $B_1 < B_2$, F étant la f.r. de x]. Montrer que le nombre de valeurs consécutives requises pour atteindre l'intervalle (B_1, B_2) suit une loi géométrique de paramètre $G(\gamma)$ (voir §4.5). Construire un algorithme de test basé sur cette idée.

Appendice

Les procédures de tri Insertion avec sentinelle *et* Quicksort

La procédure de tri a pour but de placer les *n* données d'un vecteur X en ordre de valeurs croissantes (ou non décroissantes). Pour gagner en détail tout en gardant la simplicité, les deux procédures sont présentées ici en langage BASIC élémentaire.

Insertion avec sentinelle

{Le vecteur est logiquement divisé en deux segments, un segment inférieur trié et un segment supérieur contenant des données à trier, en les adjoignant une à une au segment inférieur. Chaque nouvelle donnée du segment supérieur est « glissée » à sa place dans le segment inférieur, en tassant au besoin les données déjà triées. La donnée à trier est recopiée comme *sentinelle à* l'origine du vecteur afin de bloquer la glissade. }

```
DIM X(N)
FOR I = 2 TO N
    C = X(I):  X(0) = C: J = I
    WHILE C < X(J−1): X(J) = X(J−1): J = J−1: WEND
    X(J) = C
NEXT.
```

Quicksort (de C. A. R. Hoare)

{Une valeur arbitraire C est trouvée, et le vecteur est réorganisé de façon à ce que les données plus fortes que C et celles plus faibles soient respectivement placées de part et d'autre, constituant ainsi deux nouveaux vecteurs; ces derniers sont soumis à la même procédure, *et cætera* jusqu'à épuisement. Le principe récursif est réalisé ici par l'inscription d'un des deux sous-vecteurs à trier dans une « pile » et par le traitement immédiat de l'autre. Différentes optimisations sont possibles. La grandeur de la pile peut être fixée prudemment à $1{,}51og_2n$ *ou* plus.}

```
      DIM X(N), PILEG(50), PILED(50)
      P = 0: G = 1: D = N
10    C = X((G+D)/2): I = G: J = D
20    WHILE X(I) < C: I = I+1: WEND
      WHILE X(J) > C: J = J-1: WEND
      IF I < J THEN SWAP X(I), X(J)
      IF I <=J THEN I = I+1: J = J-1
      IF I < J THEN GOTO 20
      IF I < D THEN P = P+1: PILEG(P) = I: PILED(P) = D
      IF G < J THEN D = J: GOTO 10
      IF P > 0 THEN G = PILEG(P): D = PILED(P): P = P-1: GOTO 10
```

Remarque 1. Pour de petits vecteurs (par exemple, $n \leq 50$), la procédure d'insertion, plus simple, s'avérera sans doute plus efficace que Quicksort. Toutefois le coût d'exécution d'une telle procédure, d'ordre n^2, excédera bientôt celui de Quicksort, d'ordre inférieur, soit $n\log n$ (Knuth 1975).

Remarque 2. Pour effectuer une conversion en rangs en deux phases (voir exercice 6.20), un vecteur-jumeau, disons Y(1..N), doit être ajouté aux programmes ci-haut, et on doit l'initialiser selon la phrase « FOR I = 1 TO N : Y(I) = I : NEXT ». Par la suite, chaque mouvement de X(I) doit être accompagné d'un mouvement concomitant de Y(I).

Remarque 3. Il existe plusieurs types de procédures de tri, soit six ou plus selon les situations. Le lecteur intéressé trouvera une documentation supplémentaire dans Knuth (1973) et Wirth (1975).

Références

ANDERSON, R.L. (1942). Distribution of the serial correlation coefficient. *Annals of mathematical statistics,* 13, 1-13.

BRADLEY, J.V. (1968). *Distribution-free statistical tests.* Englewood-Cliffs (NJ), Prentice-Hall.

D'AGOSTINO, R. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika,* 58, 341-348.

D'AGOSTINO, R, PEARSON, E.S. (1973). Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika,* 60, 613-622

D'AGOSTINO, R., TIETJEN, G.L. (1973). Approaches to the null distribution of $\sqrt{b_1}$,. *Biometrika,* 60, 169-173.

DAVID, H.A. (1981). *Order statistics.* New York, Wiley.

DAVID, F.N., BARTON, D.E. (1962). *Combinatorial chance.* London, Charles Griffin.

EVANS, M., HASTINGS, N., PEACOCK, B. (2000). *Statistical distributions* (3e édition). New York, Wiley.

GENTLE, J.E. (1998). *Random number generation and Monte Carlo methods.* New York, Springer.

GOLDER, E.R. (1976). The spectral test for the evaluation *of* congruential pseudo-random generators (algorithm AS 98). *Applied statistics,* 25, 173-180.

GRAFTON, R.G.T. (1981). The runs-up and runs-down tests. Algorithm AS 157. *Applied statistics,* 30, 81-85.

HART, B.I. (1942). Tabulation of the probabilities for the ratio of the mean square successive difference to the variance. *Annals of mathematical statistics,* 13, 207-214.

HASTINGS, C. (1955). *Approximations for digital computers.* Princeton, Princeton University Press.

JOHNSON, N.L., KOTZ, S., BALAKRISHNAN, N. (1994, 1995). *Continuous univariate distributions,* Vols. 1 et 2 (2e édition). New York, Wiley.

JJOHNSON, N.L., KOTZ, S., BALAKRISHNAN , N. (1997). *Discrete multivariate distributions.New* York, Wiley.

JOHNSON, N.L., KOTZ, S., KEMP, A.W. (1992). *Univariate discrete distributions* (2e édition). New York, Wiley.

KENDALL, M.G., STUART, A. (1977). *The advanced theory of statistics.* Vol. 1*: Distribution theory* (4*e* édition); (1979). Vol. 2*: Inference and relationship* (4e édition). New York, Macmillan.

KENNEDY, W.J., GENTLE, J.E. (1998) *Statistical computing.* New York, Springer.

KNUTH, D.E. (1969). *The art of computer programming.* Vol. 2*: Seminumerical algorithms.* (1975). Vol. 3*: Sorting and searching.* Reading (Mass.), Addison-Wesley.

LAURENCELLE, L. (1983). La variance permutative. *Lettres statistiques, 7,* chap. 2, 22 p.

LAURENCELLE, L. (1993). La loi uniforme: propriétés et applications. *Lettres statistiques, 9,* 1-23.

LAURENCELLE, L. (1995). Le nombre total de suites: moments, approximations et valeurs critiques. Actes du colloque « *Méthodes et domaines d'application de la statistique* 1995 » (p. 73-88). Québec, Bureau de la Statistique du Québec.

LAURENCELLE, L. (2000). L'étude statistique des séries de deux ou plusieurs types d'événements. *Lettres statistiques,* 11, 139-174.

LAURENCELLE, L., Du**p**uis, F.A. (2000). *Tables statistiques expliquées et appliquées* (2^{e} édition). Sainte-Foy, Le Griffon d'argile.

L'ÉCUYER, P. (1998). « Random number generators and empirical tests », dans H. Niederreiter, P. Hellekalek, G. Larcher et P. Zinterhof (dir.): *Monte Carlo and quasi-Monte Carlo methods* 1996 (p. 124-138). New York, Springer.

LEHMANN, E.L. (1975). *Nonparametrics: Statistical methos based on ranks.* San Francisco, Holden-Day.

LEwIS, T.G. (1975). *Distribution sampling for computer simulation.* Toronto, Lexington Books.

LILLIEFORS, H.W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association,* 62, 399-402.

MARDIA, K.V. (1980). Tests of univariate and multivariate normality, dans P. R. Krishnaiah (dir.): *Handbook of statistics,* Vol. I (p. 279-320). North-Holland.

MARSAGLIA, G. (1995). *The Marsaglia random number CD-ROM, including the DIEHARD battery of tests of randomness.* Tallahassee (FA), Florida State University (Department of statistics).

MOOD, A.M. (1940). The distribution theory of runs. *Annals of mathematical statistics, 11,* 367-392.

OLMSTEAD, P.S. (1946). Distribution of sample arrangements for runs up and down. *Annals of mathematical statistics, 17,* 24-33.

OWEN, D.B. (1962). *Handbook of statistical tables.* Reading (MA), Addison-Wesley.

PEARSON, E.S., HARTLEY, H.O. (1970). *Biometrika tables for statisticians,* Vols. 1 et 2. Cambridge, Cambridge University Press.

ROHLF, F.J., SoKAL, R.R. (1981). *Statistical tables* (2^{e} édition). New York, Freeman.

ROYSTON, J.P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied statistics, 31,* 115-124.

RUBINSTEIN, R.Y. (1981). *Simulation and the Monte Carlo method.* New York, Addison-Wesley.

SAUGHNESSY, P.W. (1981). Multiple runs distributions: recurrences and critical values. *Journal of the American Statistical Association, 76,* 732-736.

SHAPIRO, S.S., WILK, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52,* 591-611.

SHAPIRO, S.S., WILK, M.B., CHEN, H.J. (1968). Comparative study of various tests for normality. *Journal of the American Statistical Association, 63,* 1343-1372.

WIRTH, N. (1976). *Algorithms + Data structures = Programs.* Englewood-Cliffs (NJ), Prentice-Hall.

YOUNG, L.C. (1941). On randomness in ordered sequences. *Annals of mathematical statistics, 12,* 293-300.

Chapitre
7 L'étude des phénomènes quantitatifs par évaluation numérique

7.1 Les phénomènes de la nature, d'ordre physique ou biologique, ne sont pas essentiellement quantitatifs; ils se définissent en eux-mêmes et nous y sommes confrontés globalement, à la fois comme êtres participant à cette nature et comme scientifiques. Néanmoins, l'approche quantitative est un outil privilégié de l'étude scientifique des phénomènes. Elle consiste à repérer les dimensions d'un phénomène (extension spatiale, chaleur, durée, fréquence, co-apparitions, etc.) et à les convertir en grandeurs, intensités ou nombres. Grâce à elle, nous pouvons transcrire « sur papier » le phénomène étudié, le traiter symboliquement et parvenir peut-être à un modèle, une explication, voire un procédé nous permettant de l'assujettir et le contrôler à volonté.

Personne ne s'étonnera que les nombres aléatoires, artificiellement produits, que nous avons étudiés servent à la solution de problèmes d'évaluation en statistique, par exemple pour décider si les données de deux groupes expérimentaux reflètent une différence systématique ou hautement improbable entre leurs niveaux respectifs. C'est par une méthode semblable, mais après une démarche plus compliquée, que la méthode Monte Carlo est appliquée aussi à l'étude des phénomènes quantitatifs ou à l'étude des modèles quantitatifs des phénomènes. Nous tâcherons d'en donner une idée dans les paragraphes suivants.

La *méthode Monte Carlo* est une méthode basée sur l'utilisation de nombres aléatoires par informatique dont le but est de trouver, mesurer ou vérifier une solution d'un modèle quantitatif ou de décrire son comportement simulé et ses états transitoires. Aussi baptisée « méthode des essais statistiques », Kleijnen (1974, p. 6) la définit comme toute technique appliquée à la solution d'un modèle quantitatif et utilisant des nombres aléatoires. Il existe plusieurs champs d'application de la méthode Monte Carlo, que nous examinons à présent.

7.2 Intégrales et autres problèmes déterministes. Il peut paraître surprenant que l'évaluation d'intégrales constitue la principale application de la méthode Monte Carlo, surprenant en ce qu'on utilise du hasard afin

de mesurer une grandeur fixe. C'est aussi l'avis de Hammersley et Handscomb (1964), selon qui « *every Monte Carlo computation that leads to quantitative results may be regarded as estimating the value of a multiple integral* » (p. *50).*

Il existe différentes sortes d'intégrales, et mieux vaut une solution analytique et exacte lorsqu'elle se trouve. L'estimation Monte Carlo qui, on s'en doute, donnera une solution seulement approximative s'applique lorsque la fonction à intégrer n'est pas connue mais se manifeste uniquement par une quantité observable, une variable: on parle alors d'évaluation (ou intégration) Monte Carlo par variable. Même lorsque la fonction est connue, celle-ci peut être très complexe et défier une approche exacte; les méthodes d'intégration numérique (voir par exemple Gerald et Wheatley 1984) sont alors disponibles, dans le rang desquelles il est légitime de placer la méthode Monte Carlo: on retrouve ici certaines techniques d'intégration Monte Carlo par manipulation de fonction. Enfin, pour des intégrales d'ordre élevé (en $\mathbf{R}^n$, $n > 1$), l'échantillonnage Monte Carlo peut représenter la seule solution pratique.

Les autres applications de la méthode Monte Carlo, que nous passerons maintenant en revue, se ramènent d'une manière ou l'autre à un problème d'intégration.

7.3 *Études d'échantillonnage statistique.* La méthode Monte Carlo est l'outil naturel pour étudier les objets statistiques, tels les moments, les centiles ou les extrêmes de fonctions de variables aléatoires. Soit les variables $\{x_i\}$ d'une distribution $f(x)$ donnée, et:

$$y = h(x_1, x_2, \cdots, x_n) , \qquad (7.1)$$

une fonction d'un échantillon de n v.a. En se basant sur m valeurs échantillonnées de y, on utilisera $moy(y^k)$, $y_{(r:m)}$ où $r = [P(m+1)+½]$, et $y_{(man)}$ pour estimer respectivement le moment à l'origine $\mu'_k(y)$, le centile $y100p$ et le maximum $max\ y$. Fondamentalement, il s'agit d'intégrales multiples; la plupart sont analytiquement intraitables ou très difficiles, a fortiori lorsqu'il existe une structure de dépendance (ou une corrélation) dans les données échantillonnées, comme il arrive dans l'étude des séries chronologiques en économétrie ou en météorologie. L'approche Monte Carlo donne tout de suite une estimation valable de la réponse.

Les études sur la performance des estimateurs statistiques (en statistique théorique, biologie quantitative, économétrie, psychologie, etc.), sur la robustesse ou sur la puissance des tests statistiques, les réalisations des tests dits de ré-échantillonnage (bootstrap, tests permutationnels ou de

combinatoire approximative, etc.), tombent toutes dans cette catégorie de l'échantillonnage statistique. Nous y reviendrons à la faveur d'exemples.

7.4 La *simulation stochastique* est une autre classe d'applications de l'approche Monte Carlo; il s'agit ici d'évaluer la réponse globale d'un modèle soumis à des conditions ou à des valeurs paramétriques données.

Un modèle est un simulacre d'un système, construit afin d'imiter certaines propriétés de structure et de comportement de ce dernier. Il y a des modèles physiques, concrets, comme dans l'étude des turbulences induites par une soufflerie autour d'une aile d'avion, ou l'étude de la communication verbale dans un groupe social mais en laboratoire. Les modèles qui nous intéressent ici sont abstraits, formels, en ce sens que les propriétés à imiter sont construites comme des qualités ou des règles de fonctionnement du modèle cible. Dans certains cas, un tel modèle abstrait pourra se ramener à quelques équations mathématiques; dans d'autres cas, il s'agira d'un organigramme, d'un schéma composé de différents blocs reliés les uns aux autres et assortis de règles de fonctionnement.

C'est dans un programme informatique que s'opérationnalise le plus souvent un modèle formel: par l'exécution séquentielle des instructions du programme, le modèle se trouve animé comme le système qu'il imite. Il existe même des langages spécialisés permettant la simulation de modèles formels par ordinateur: *Simscript, GPSS, Simula* et d'autres; on trouvera un avant-goût de ces langages dans Bratley, Fox et Schrage (1987). Les langages numériques habituels tels que *Fortran, Basic* ou *Pascal* permettent aussi la simulation, moyennant une part de créativité du programmeur, et leur adéquation pour calculer des statistiques du modèle est bien établie.

La simulation d'un système au moyen d'un modèle, voire d'un modèle formel opérationnalisé dans un programme, ne relève pas forcément du domaine des applications Monte Carlo. Pour en relever, il est généralement requis que la simulation fasse usage de nombres aléatoires représentant un aspect ou l'autre du modèle étudié; il s'agit donc d'un modèle stochastique, c'est-à-dire un modèle comportant au moins un aspect soumis à une influence aléatoire, et c'est pourquoi l'on parle alors de simulation stochastique. La simulation stochastique devient spécifiquement une application Monte Carlo si son but premier est d'évaluer une quantité, un paramètre, un ensemble d'états caractéristiques du modèle à l'étude.

Dans certains cas très simples, l'évaluation de quantités descriptives d'un modèle stochastique peut aussi se faire analytiquement: en fait foi l'exemple d'une file d'attente à temps de service constant et à temps d'arrivées obéissant à une même loi *Gamma.* Toutefois, la plupart du temps, il faut recourir à l'évaluation statistique de ces quantités, ce qui définit

précisément l'objet d'une étude Monte Carlo. Par cette étude, on peut aussi détecter ou dénombrer des états transitoires du modèle, des séquences ou interactions séquentielles dont la réponse globale ne garde plus trace, la présence possible d'effets rares, etc. Ce n'est pas le lieu d'exposer ici les nombreux mérites de la modélisation et des études de simulation[1]; il suffit de considérer que la méthode Monte Carlo y joue un rôle parfois capital.

7.5 La méthode Monte Carlo connaît deux autres champs d'applications, moins spécifiques que ceux mentionnés plus haut. Il s'agit de l'animation d'un modèle formel et de la pédagogie des mathématiques.

Modélisation et étude Monte Carlo sont deux activités totalement distinctes. Une fois qu'il est construit, la fonction d'un modèle n'est pas toujours de permettre le calcul de quantités caractéristiques. Le modèle peut servir aussi d'instrument d'observation. Ainsi, pour un modèle formel transcrit dans un programme informatique, l'étude sera grandement facilitée si le modèle est mis en branle, animé, à l'instar du système qu'on cherche à reproduire. L'animation consiste à faire passer le modèle d'un état ou d'une phase à l'autre, grâce à un mécanisme quelconque d'horlogerie. Elle comporte habituellement des paramètres de contexte ou des conditions globales variables, des mécanismes de perturbation susceptibles de produire des effets et des interactions peu prévisibles dans le modèle, etc. Cette simulation stochastique est effectuée dans un contexte où tous les états et les transitions d'états du modèle sont explicites et observables. Dans la mesure où le modèle est une *boîte noire* représentant le système à l'étude, la simulation stochastique et son observation peuvent donner lieu à une induction réciproque sur le système et à des intuitions touchant ses caractéristiques subtiles ou imprévues.

Que dire enfin de l'utilité pédagogique de la méthode Monte Carlo, considérant la multiplicité et la richesse de ses ingrédients? Pour l'enseignant et l'étudiant en statistique, la méthode Monte Carlo les met en présence d'un foisonnement de variables aléatoires dont ils peuvent observer les propriétés *in vivo:* par exemple, un graphique progressif de la distribution des statistiques d'un modèle de probabilité, l'effet du théorème limite central sur la somme de v. a. rendu visible, la transformation d'une v.a. d'une forme distributionnelle vers une autre, etc. Il en va de même pour l'étudiant en informatique, qui prend ainsi contact avec la statistique et la modélisation. Pour l'étudiant en sciences, en physique par exemple, la modélisation est habituelle et l'approche Monte Carlo rendra possible l'étude dynamique des modèles soumis à toutes conditions qu'il lui plaira

1 Voir pour cela Naylor (1971) ou l'introduction de l'ouvrage déjà cité de Bratley, Fox et Schrage (1987).

d'imaginer et mettre en œuvre. En fait, l'usage universel de l'ordinateur personnel a fait de la méthode Monte Carlo un outil accessible et bientôt aussi banal que le traitement de texte ou le chiffrier électronique.

7.6 Historique de la méthode Monte Carlo. La première mention publiée de la méthode Monte Carlo se trouve dans un article de N. Metropolis et S. Ulam, « *The Monte Carlo method »,* paru en 1949; l'expression et les premières techniques sont attribuées à J. von Neumann et S. Ulam, qui les développèrent à l'occasion de leur travail sur la bombe atomique durant la deuxième guerre mondiale (voir Rubinstein 1981). Certaines techniques d'échantillonnage ou de ré-échantillonnage intensif avaient été expérimentées bien avant, dès le XVIII^{e} siècle. Un exemple illustre est celui de *Student* (1908), l'alias de W. S. Gosset, qui s'en servit pour valider sa loi de la distribution d'une moyenne de deux à quatre observations. En 1936, Fisher, s'inspirant de son principe de « randomisation » (ou mélange aléatoire) des conditions d'une expérimentation, décrivait une procédure d'énumération exhaustive de la combinatoire d'un problème de test statistique, procédure connue sous le nom de « randomisation test ». Hoeffding (1952), Barnard (1963) et Edgington (1969, 1980) généralisèrent cette procédure en la limitant à une énumération incomplète, obtenue par échantillonnage aléatoire, de la combinatoire du problème de test et en démontrant sa validité: la documentation désigne ces procédures sous les vocables de tests permutationnels ou tests par combinatoire exhaustive ou approximative (§8.6-8.8).

La disponibilité de l'ordinateur et sa vitesse ont favorisé l'usage accru de la méthode Monte Carlo pour attaquer et solutionner toutes sortes de problèmes, parfois même au détriment d'une analyse préalable de ceux-ci. La vogue de la méthode Monte Carlo, la multiplication des applications, la taille des problèmes traités ont fait apparaître la question de l'efficacité de la méthode et des moyens d'améliorer cette efficacité. Kahn et Marshall publièrent en 1953 le premier article touchant la question de la taille des expérimentations Monte Carlo; ils furent suivis de Hammersley et Handscomb en 1964, puis d'une foule d'autres auteurs s'intéressant aux techniques destinées à réduire la variance des estimations Monte Carlo. Devant cette profusion de techniques et en confondant la forêt pour les arbres, certains auteurs en sont venus à parler « des » méthodes Monte Carlo. Ne voilà-t-il pas un signe de santé florissante pour cette approche statistico-informatique de la solution de problèmes quantitatifs?

7.7 Les buts spécifiques de la méthode Monte Carlo sont, comme les applications que nous avons évoquées plus haut, multiples et diversifiés. Toutefois, afin d'unifier notre exposé, nous mettrons l'accent sur le but général et principal de la méthode, soit d'obtenir une estimation $\hat{Q}$ d'une quantité Q à évaluer. Cette quantité pourra être simple (scalaire) ou

multiple, correspondre à une intégrale ou mesurer la performance d'un modèle, etc. Les différents exemples présentés ré-introduiront les aspects spécifiques que peut gommer le principe d'un exposé unifié.

Le chapitre 9 présente la méthode Monte Carlo dans sa formule fondamentale. Au chapitre 10, sont discutées des techniques destinées spécialement à améliorer l'efficacité de la méthode, son ratio qualité: coût. Enfin, le chapitre 11 présente quelques exemples plus élaborés d'études Monte Carlo: ces exemples servent en même temps de synthèse pédagogique afin d'aider à mieux intégrer les nombreux éléments répertoriés.

Cependant, avant d'aborder la méthode Monte Carlo dont le but, rappelons-le, est de produire une solution approximative d'un problème quantitatif, nous passerons rapidement en revue, au chapitre 8 qui suit, quelques méthodes d'évaluation déterministes, des méthodes qui donnent à chaque coup la solution exacte (ou virtuellement exacte). Le lecteur, une fois averti de ces méthodes et de leurs applications, pourra opter ou non pour la méthode approximative Monte Carlo et ce, en connaissance de cause.

Exercices

7.1 Définir le plus rigoureusement possible les notions de système (ou phénomène), théorie et modèle. Montrer que la raison d'être spécifique d'un modèle est son utilité et qu'il constitue le support heuristique de la relation entre système et théorie.

7.2 Montrer de quelle façon chacun des problèmes d'évaluation suivants revient à un calcul d'intégrale: (a) L'espérance et la variance de $M_G(x_i, i = 1, n)$, la moyenne géométrique de n v.a. issues d'une loi $f(x)$ donnée, où $M_G = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{1/n}$. (b) Le temps d'attente moyen et la distribution des temps d'attente dans une queue de service à temps de service constant et temps d'arrivées obéissant à une loi exponentielle. (c) L'exactitude (ou taux de rejet effectif, α_{eff}) du test t de la différence de deux moyennes quand les séries statistiques proviennent de populations à corrélations sérielles non nulles. (d) Le mérite relatif de deux ou trois méthodes de détermination d'un seuil physiologique à partir de quelques échantillons infra- et supra-liminaires. (e) Le volume maximal d'une chambre à particules sphérique tel que, malgré la dissipation spontanée d'énergie cinétique et en raison des collisions inter-particules, la chaleur à la surface cesse d'être proportionnelle au taux de bombardement.

7.3 Dans chacun des domaines suivants, concevoir un problème d'évaluation comportant ou non des composantes aléatoires, pour lequel un plan de solution par la méthode Monte Carlo est ébauché: (a) En démographie. (b) En physique. (c) En économie. (d) En médecine (épidémiologie). (e) En psychologie sociale.

Références

BARNARD, G.A. (1963). (Commentaire). *Journal of the Royal Statistical Society B, 25,* 294.

BRATLEY, P., Fox, B.L., SCHRAGE, L.E. (1987). *A guide to simulation.* New York, Springer-Verlag.

EDGINGTON, E.S. (1969). Approximate randomization tests. *Journal of psychology, 72,* 143-149.

EDGINGTON, E.S. (1980). *Randomization tests.* New York, Marcel Dekker.

FFISHER, R.A. (1936). « The coefficient of racial likeness » and the future of craniometry. *Journal of the Anthropological Institute, 66,* 57-63.

GERALD, C.F., WHEATLEY, P.O. (1984). *Applied numerical analysis.* Reading (MA), Addison-Wesley.

HAMMERSLEY, J.M., HANDSCOMB, D.C. (1964). *Monte Carlo methods.* London, Chapman and Hall.

HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of mathematical statistics, 23,* 169-192.

KAHN, H., MARSHALL, A.W. (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of operations research of the Society of America, 1,* 263-278.

KLEIJNEN, J.P.C. (1974). *Statistical techniques in simulation.* New York, Marcel Dekker.

METROPOLIS, N., ULAM, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association, 14,* 335-341.

NAYLOR, H. (dir.) (1971). *Computer simulation experiments with models of economic systems,* New York, Wiley.

RUBINSTEIN, R.Y. (1981). *Simulation and the Monte Carlo method,* New York, Wiley.

STUDENT (1908). The probable error of a mean. *Biometrika, 6,* 1-25.

Chapitre 8

Méthodes déterministes d'évaluation

8.1 La méthode Monte Carlo a comme fonction spécifique l'évaluation d'une quantité caractéristique par une procédure utilisant des nombres aléatoires. Comme on l'a vu au chapitre précédent, cette évaluation équivaut très souvent au calcul d'une intégrale, même alors qu'on ne dispose pas d'une fonction mathématique qu'on puisse intégrer. Dans certains cas, l'évaluation Monte Carlo produit une approximation, une estimation statistique d'une quantité que d'autres méthodes permettraient d'obtenir exactement. Cependant, les méthodes exactes sont parfois impraticables, parce que trop longues ou trop coûteuses, ou parce que la personne confrontée au problème n'a pas l'expertise nécessaire pour les mettre en œuvre. Il existe enfin d'autres situations dans lesquelles la méthode Monte Carlo semble être la seule voie de solution, comme dans la simulation de modèles stochastiques complexes ou le calcul d'intégrales de dimension élevée.

Il reste que, assez souvent, l'évaluation Monte Carlo est conçue et élaborée pour un problème qui admettrait peut-être un traitement exact, Or la mathématique appliquée recèle de nombreux outils qui permettent l'évaluation précise d'une quantité donnée; au lieu d'être *statistique,* c'est-à-dire sujette à une fluctuation intrinsèque, l'évaluation est ici *déterministe* et elle produit à chaque fois le même résultat quand elle est appliquée au même problème, un résultat exact quand la méthode est exacte et un résultat quasi exact quand la méthode est approximative.

Il serait déplacé de ré-introduire toutes les méthodes de calcul applicables et de les présenter en détail. Nous nous contentons plutôt de mentionner les méthodes qui nous semblent les plus utiles — notamment l'intégration directe, l'intégration numérique et l'énumération combinatoire —, en renvoyant le lecteur aux ouvrages spécialisés. C'est notre expérience que, une fois familiarisés avec la méthode Monte Carlo et l'ayant maîtrisée dans ses aspects principaux, les utilisateurs s'en obnubilent volontiers et veulent l'appliquer à tous les problèmes (car elle est d'application universelle) même lorsqu'une solution exacte est disponible. La présentation de méthodes d'évaluation déterministes, dans ce chapitre, a pour but

principal de situer la méthode Monte Carlo dans l'ensemble des méthodes généralement disponibles pour l'évaluation quantitative.

8.2** Intégration directe. La* méthode d'évaluation la plus facile à présenter, et qui contient aussi les subtilités et les difficultés les plus grandes, est l'intégration directe, ou analytique. Dénotons ici par une lettre minuscule (e.g. « ***f ») une fonction d'une variable (x), et par la lettre majuscule correspondante (e.g. « F ») sa fonction *primitive,* i.e. une fonction telle que sa dérivée redonne la première, soit:

$$d\,F(x) = f(x)\,dx\ .$$

Dans cette notation, *l'intégrale* de $f(x)$ entre les bornes $x = a$ et $x = b$ est exprimée et donnée par:

$$\int_a^b f(x)\,dx = F(b) - F(a)\ . \qquad (8.1)$$

Cette intégrale, soit la valeur $Q = \int_a^b f(x)\,dx$, peut être vue comme l'aire de la surface délimitée d'une part par l'axe horizontal, à $y = 0$, et le tracé de la fonction $y = f(x)$, et d'autre part par les bornes $x = a$ et $x = b$.

Un exemple des plus simples est fourni par le calcul des moments à l'origine (μ'_r) de la v.a. uniforme u, bornée de 0 à 1, soit $u \sim U(0,1)$ (cf. §3.9). Par définition, le moment à l'origine d'ordre r pour une v.a. continue x est:

$$\mu'_r(x) = \int x^r f(x)\,dx \quad \{\, f(x) \geq 0 \,\}\ . \qquad (8.2)$$

Ici, comme la densité de u est $f(u) = 1$ et la fonction à intégrer, $u^r \bullet f(u) = u^r$ dans l'intervalle de 0 à 1, sa «primitive» $F(u)$ est donc $u^{r+1}+C$. Appliquant (8.1), on obtient :

$$\mu'_r(u) = \int_0^1 u^r\,du = \frac{1^{r+1}}{r+1} - \frac{0^{r+1}}{r+1}\ , \qquad (8.3)$$

c'est-à-dire le résultat bien connu, $\mu'_r(u) = 1/(r+1)$.

La solution d'un problème de calcul intégral est ordinairement plus complexe, et différentes techniques peuvent être mises à contribution: l'intégration par parties, le développement en fractions partielles, le développement en série de Taylor, le changement de variables, les substitutions trigonométriques, etc. Des ouvrages comme Swokowski (1993) ou Ayres (1972) en français, ou Schwartz (1967) en anglais, restent précieux comme livres de chevet. L'exemple suivant illustre l'exploitation d'une série de Taylor pour calculer l'intégrale de la loi normale. Les exercices en fin de chapitre mettent en œuvre d'autres techniques.

Exemple 8.1 Intégration directe de la loi normale

Soit $\Phi(x)$, l'intégrale, ou fonction de répartition, de la loi normale, telle que $\Phi(x)$ = f.r.(x) = pr$(z \leq x)$; à toutes fins pratiques, nous supposons que la variable x est standardisée, avec moyenne (μ) 0 et variance (O^2) 1. Par la symétrie de cette loi autour de μ=0, nous avons ($\Phi(-x)$ = $1-\Phi(x)$, de sorte que l'évaluation peut se restreindre au domaine $x \geq O$.

Nous devons donc évaluer:

$$\Phi(x) = \int_{-\infty}^{x} f(z)\, dz = \tfrac{1}{2} + \int_{0}^{x} f(z)\, dz \quad \{ x \geq 0 \} \tag{8.4}$$

en utilisant la densité normale standard:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} . \tag{8.5}$$

La fonction e^z a pour primitive $e^z + C$, mais $e^{f(z)}$ n'a pas de primitive si $f(z)$ n'est pas linéaire sur z. D'un autre côté, l'expansion de e^z en série de Tavlor (Schwartz 1967) est:

$$e^z = 1 + z + z^2/2! + z^3/3! + z^4/4! + \cdots , \tag{8.6}$$

le terme général étant $z^n/n!$ et la somme à droite convergeant sur la valeur cherchée e^z. Cette expansion a pour avantage d'exprimer toute fonction comme une somme des puissances successives de la variable, i.e. z^n, dont les primitives sont faciles à obtenir. Or l'intégrale (ou la dérivée) d'une somme est égale à la somme des intégrales (ou des dérivées) de ses composantes, ce qui permet de solutionner (8.4).

Remplaçant z par $-\frac{1}{2}z^2$ dans (8.6), l'expansion de $\exp(z^2/2)$ a pour terme général $(-z^2/2)^n/n!$ ou $(-1)^n z^{2n}/(2^n n!)$. Les quelques premiers termes en sont $1 - z^2/2 + z^4/[4\cdot2] - z^6/[8\cdot6]$ + etc. La fonction primitive de $\exp(-z^2/2)$ s'obtient alors en sommant les primitives de chaque terme, chacune étant trouvée selon la variable d'intégration z. Ainsi, le terme général étant $(-1)^n z^{2n}/(2^n n!)$, le facteur-clef z^{2n} a pour primitive $z^{2n+1}/(2n+1)$, d'où le terme général de la fonction primitive devient $(-1)^n z^{2n+1}/[2^n n!(2n+1)]$. La solution complète de (8.4), en appliquant (8.1) à $z = x$ et $z = 0$, devient alors:

$$\Phi(x) = \tfrac{1}{2} + \frac{1}{\sqrt{2\pi}} \left[x - \frac{x^3}{6} + \frac{x^5}{40} - \frac{x^7}{336} + \cdots + \frac{(-1)^n x^{2n+1}}{2^n n!(2n+1)} + \cdots \right] . \tag{8.7}$$

La somme dans (8.7) converge rapidement pour de petites valeurs de x telles que x<1,5, et plus lentement par la suite; des difficultés de convergence peuvent se produire pour x > 3, selon la précision de l'arithmétique employée (voir aussi l'exercice 8.9).

8.3 *Intégration numérique. À* défaut de connaître ou d'obtenir la primitive F de la fonction f à intégrer, ou si cette primitive n'existe pas, on peut recourir à d'autres méthodes pour mesurer l'aire à trouver, $Q = \int_a^b f(x)\,dx$: il s'agit des méthodes dites numériques. Imaginons l'intégrale comme l'aire de la surface sous le tracé de $f(x)$, comprise entre $x = a$ et $x = b$. L'approche globale exploitée par ces méthodes revient à découper la surface à intégrer en bandes de largeur constante Δx; on ajuste alors une figure géométrique simple sur chaque bande, on détermine son aire et on la somme, le total des aires approchant l'aire de la surface entière. La précision de l'aire trouvée dépend d'abord du degré d'adéquation entre la figure exploitée pour chaque bande et la forme de celle-ci; elle varie aussi, bien entendu, avec le nombre de bandes utilisées.

La documentation présente différentes techniques pour l'intégration numérique d'une fonction (Gerald et Wheatley 1984; Kincaid et Cheney 1991; Rougier 1987): technique des rectangles, des trapèzes, de Simpson, extrapolation de Romberg, intégration de Gauss, adaptative, etc. Les quelques premières techniques, dites de Newton-Cotes, correspondent à des polynômes d'interpolation et peuvent être dénotées symboliquement comme suit:

Technique	Degré du polynôme	Formule symbolique	
Rectangle	0	$\sum \Delta x\ f(x_i)$	(8.8a)
Trapèze	1	$\sum \Delta x\ \tfrac{1}{2}[f(x_{i+1})+f(x_i)]$	(8.8b)
Simpson ⅓	2	$\sum \Delta x\ \tfrac{1}{3}[f(x_{i+2})+4f(x_{i+1})+f(x_i)]$	(8.8c)
Simpson ⅜	3	$\sum \Delta x\ \tfrac{3}{8}[f(x_{i+3})+3f(x_{i+2})+3f(x_{i+1})+f(x_i)]$	(8.8d)

Nous présentons plus bas la technique la plus simple, celle des rectangles, et la plus utile, la règle ⅓ de Simpson, dite aussi règle parabolique. Des données complémentaires sont fournies dans les exercices.

8.4 *Intégration par rectangles.* La méthode la plus simple, sans contredit, est celle dite des rectangles, basée sur (8.8a). Le domaine d'intégration, disons de $x = a$ *à* $x = b$, est divisé en m bandes de largeur individuelle $\Delta x = (b-a)/m$; la première bande a pour intervalle $(a,\ a+\Delta x)$, la seconde $(a+\Delta x,\ a+2\Delta x)$, etc. jusqu'à la dernière bande, $(a+(m-1)\Delta x,\ b)$. Chaque bande

découpée dans la fonction $f(x)$ est remplacée par un rectangle. L'aire totale Q est alors estimée par la somme des aires des rectangles,

$$\hat{Q} = \hat{Q}(R_1) + \hat{Q}(R_2) + \cdots + \hat{Q}(R_m) . \qquad (8.9)$$

L'aire de chaque rectangle correspond au produit de sa base Δx par sa hauteur $f(x_i)$, $x_i \in (a+(i-1)\Delta x,\ a+i \cdot \Delta x)$; n'importe quelle valeur dans l'intervalle fait l'affaire, le point-milieu $x_i = a + (i - \frac{1}{2})\Delta x$ étant un choix naturel.

L'intégration par rectangles est en fait la méthode qui, par sa structure, s'approche le plus d'une méthode d'estimation statistique. Dénotons par $x_1, x_2, \ldots, x_m$ les abscisses jalonnant les intervalles consécutifs du domaine d'intégration (a, b), et par $y_1 = f(x_1)$, $y_2 = f(x_2)$, etc. les ordonnées ou valeurs correspondantes de la fonction à intégrer. Alors, l'estimateur de l'intégrale Q est:

$$\hat{Q} = \frac{(b-a)}{m} y_1 + \frac{(b-a)}{m} y_2 + \cdots + \frac{(b-a)}{m} y_m , \qquad (8.10)$$

soit, équivalemment:

$$\hat{Q} = (b-a)\,\bar{y} . \qquad (8.11)$$

Cette dernière formule, utilisant la moyenne y des m ordonnées $f(x_i)$, est la même qui sera mise en œuvre dans l'évaluation Monte Carlo d'une intégrale, du moins par la méthode de base. Bien sûr, dans ce dernier cas, les points d'abscisses x_i sont aléatoires, tout au contraire de leur espacement régulier dans la technique des rectangles.

La technique dite « Monte Carlo *pondérée* » (« weighted Monte Carlo »), ou par échantillonnage systématique (§ 10.11), utilise une suite d'intervalles à largeurs (Δx) variables de même que des points d'abscisse à positions aléatoires dans l'intervalle; cette curieuse méthode est comme une hybridation de la technique des rectangles et du Monte Carlo de base.

8.5 *Règle parabolique de Simpson.* À partir d'une seule coordonnée $\{x_i, f(x_i)\}$, on peut fixer un rectangle. En ajoutant une seconde coordonnée, $\{x_j, f(x_j)\}$, un trapèze est défini. Avec une troisième coordonnée $\{x_k, f(x_k)\}$, enfin, on obtient une surface qu'on peut délimiter par un arc parabolique, surface dont l'aire se calcule très simplement. En fait, si les trois abscisses requises sont équidistantes, avec un écart individuel Δx, la formule (8.8c) donne exactement l'aire sous la parabole traversant les trois coordonnées.

Comme pour la technique des rectangles, il est d'usage de découper le domaine d'intégration en plusieurs bandes, ici en m bandes égales, bornées par m+1 coordonnées, où m est pair. Pour chaque couple de bandes

consécutives, un polynôme du 2e degré en x est exactement ajusté sur trois coordonnées $\{x_i, f(x_i)\}$, l'intégrale correspondante est obtenue, puis le tout est sommé, donnant enfin:

$$\hat{Q} = \frac{\Delta x}{3}\{f(a)+4f(a+\Delta x)+2f(a+2\Delta x)+4f(a+3\Delta x)+ \cdots +4f(b-\Delta x)+f(b)\} ;$$

c'est la justement fameuse « règle de Simpson généralisée ». Cette règle est, parmi les techniques générales applicables à l'intégration approximative d'une fonction, la plus performante (voir cependant les exercices 8.14-8.15). Les ouvrages spécialisés rapportent de façon détaillée l'erreur encourue par chaque technique (voir Gerald et Wheatley, *op. cit.,* et l'exercice 8.13).

Exemple 8.2 Intégration numérique de la fonction e^{-X}

Supposons qu'on veuille connaître $Q = \int e^{-X} dx$, de $x = 0$ à $x = 1$. Par intégration directe, la primitive de $f(x) = e^{-X}$ est $F(x) = - e^{-X} + C$, d'où $Q = F(1)-F(0) = -e^{-1}-(-1) = 1-e^{-1} \approx 0{,}63212$. Faisons comme si cette solution exacte n'était pas connue (cette situation se produit plus souvent qu'autrement dans les applications statistiques) et appliquons un procédé d'intégration numérique.

Par la technique des rectangles, nous pouvons utiliser d'abord un seul rectangle, de base $\Delta x = 1$, la valeur x est fixée au centre (=½), et $f(½) = 0{,}60653$. L'estimation avec 1 rectangle centré est alors $Q(R_1) = 1 \times 0{,}60653 = 0{,}60653$, impliquant une erreur de -4,0 %. Avec deux rectangles, centrés aux abscisses 1/4 et 3/4, et $\Delta x = ½$, nous avons à peu près $Q(R_2) = 0{,}38940 + 0{,}23618 = 0{,}62558$, pour une erreur de -1,0 %; l'erreur absolue tombe sous la barre de 0,1 % en utilisant 7 rectangles, avec $Q(R_7) = 0{,}63158$.

Par la règle de Simpson, appliquée en un seul coup, il nous faut ici $\Delta x = ½$ et les coordonnées $\{0, f(0)\}$, $\{½, f(½)\}$ et $\{1, f(1)\}$; d'où, avec un seul arc parabolique, $Q(A_1) \approx ⅓ \cdot ½ \{ 1 + 4 \times 0{,}60653 + 0{,}36788\} = 0{,}63233$, avec une erreur de +0,03 %, un excellent résultat déjà, utilisant seulement 3 coordonnées. Avec $m+1 = 7$ coordonnées, $½m = 3$ arcs peuvent être formés, utilisant des intervalles de largeur $\Delta x/m = 1/6$. L'application de la règle généralisée donne ici:

$$Q(A_3) = ⅓(1/6)\{f(0)+4f(1/6)+2f(2/6)+4f(3/6)+2f(4/6)+4f(5/6)+f(6/6)\}$$
$$= 0{,}63212 ,$$

soit un résultat exact à 5 décimales.

La règle de Simpson peut donner une très bonne approximation de l'intégrale normale (8.4) en un seul coup, en l'ancrant sur le point d'inflexion de la densité normale *f(x)* à l'abscisse $x = 1$. Si $0 < x < 1$, on aura ($\Phi(x) = ½ + Q(A_1)$, le calcul étant fait dans l'intervalle $(0, x)$. Si $x \geq 1$, il suffit de faire ($\Phi(x) = 0{,}84134 + Q(A_1)$, l'intervalle utilisé étant ici $(1, x)$. L'exercice 8.11 poursuit cette question.

8.6 *Combinatoire énumérative.* Un lieu privilégié d'application de la méthode Monte Carlo est la statistique inférentielle, où l'on s'occupe d'estimation et du test de la probabilité d'écart (ou test d'écart significatif) pour des échantillons de variables aléatoires. Dans plusieurs cas, la logique de l'opération d'estimation ou de test peut être traduite dans le langage de la combinatoire et l'opération être réalisée par une énumération. L'estimation et le test par la méthode Monte Carlo produisent alors des approximations des solutions de combinatoire; le plus souvent, seules les approximations sont praticables puisque la taille des opérations de combinatoire interdit habituellement leur réalisation complète.

L'histoire moderne des tests statistiques par combinatoire remonte, semble-t-il, à Fisher, en 1936: ce fondateur de la statistique moderne accréditait les divers tests dits paramétriques, tel le test *t* de Student sur la différence entre deux moyennes de groupes, par le fait que leurs résultats approchaient ceux des tests de combinatoire correspondants.

> Cette proposition de Fisher dérive de son principe de *« randomisation »,* ou répartition aléatoire: c'est pourquoi les tests par combinatoire sont aussi appelés, en anglais, « *randomisation tests* ». La statistique étudiée émane d'une configuration de résultats observée. Or, si le hasard seul joue et si nulle influence systématique ne biaise les données, la configuration observée peut être vue comme l'une quelconque des configurations constituant la combinatoire du test considéré. Soit T, la taille de cette combinatoire et f_{ext}, le nombre de configurations donnant lieu à une valeur aussi extrême ou plus extrême que la valeur observée. Alors, la probabilité extrême:
>
> $$P_{ext} — f_{ext} / T \qquad (8.13)$$
>
> est une mesure de la rareté, du taux d'exceptionnalité de la configuration observée. Si cette rareté est suffisante et atteint un seuil, i.e si $p_{ext} \leq \alpha$, alors la configuration sera déclarée significative; en d'autres mots, si la configuration observée est avérée rare, l'on conclura à la présence d'un biais systématique dans les données, ou à l'influence d'un source d'influence externe.

La praticabilité de ces tests, dits aussi de combinatoire exhaustive, dépend principalement de T, la taille de l'ensemble combinatoire. Lorsque la taille est démesurée, l'on recourra plutôt à l'approche Monte Carlo, consistant ici

à former T' configurations au hasard (avec remise) parmi les T disponibles, à déterminer f_{ext} tel que mentionné, puis:

$$p'_{ext} = f'_{ext} / T' , \qquad (8.14)$$

un estimateur qui approche raisonnablement bien la quantité (8.13) ci-dessus, même pour des tailles T' peu élevées (mais au moins égales à $1/\alpha$-1): il s'agit alors de combinatoire approximative. Laurencelle (1987) compare systématiquement ces deux approches.

Conley (1984) décrit une autre application de combinatoire énumérative associée cette fois à la solution de problèmes d'optimisation.

8.7 *Le bootstrap.* La méthode du « bootstrap » fut proposée par B. Efron (1982, 1983) pour résoudre des problèmes d'estimation et de testing par combinatoire, dans le contexte d'un seul échantillon de données. Soit une statistique S calculée sur les données d'un échantillon, ou $S = S(x_1, x_2, ..., x_n)$. La méthode consiste à bâtir une *distribution de bootstrap,* ou auto-distribution, de la statistique S en formant des échantillons possibles; chaque pseudo-échantillon est composé de n valeurs, $(x^*1, x^*2, ..., x^*n)$ pigées *avec remise* dans l'échantillon $\{x_1, x_2, ..., x_n\}$, produisant chacun une mesure S^*. L'ensemble combinatoire ainsi formé représente une sorte de distribution échantillonnale de la statistique S, qui permet la détermination de centiles ou le calcul d'intervalles de confiance ou de tests.

> Le but général de la procédure est d'estimer $S = S(F; n)$ à partir de $S(\hat{F}; n)$, où F est la fonction de répartition réelle mais inconnue de la variable x, et $\hat{F}$ est la f.r. empirique, définie uniquement aux points x_1, x_2, etc., chacun ayant une probabilité de $1/n$. Un échantillonnage exhaustif de la « population » $\{F; n\}$ permettrait d'établir la « vraie » distribution échantillonnale de S, son espérance, sa variance, etc. Par analogie, l'échantillonnage complet ou approximatif de $\{\hat{F}; n\}$ fournit une estimation, une estimation non paramétrique optimale, des mêmes caractéristiques.

La taille de base de l'énumération de type bootstrap étant n^n (voir cependant les exercices 8.16-8.18), on ne s'étonne guère que les estimations bootstrap soient presque toujours approximatives et basées sur une énumération incomplète. Dès $n = 8$, on affronte $T = 8^8 \approx 17\ 000\ 000$ échantillons, une tâche suffisante pour faire ahaner nos ordinateurs les plus performants !

8.8 *Les « randomisation tests ».* Les tests d'hypothèses en statistique reposent sur une hypothèse nulle selon laquelle, en général, nulle influence autre que le hasard ne gouverne les données échantillonnées et n'en vient influencer le niveau ou le comportement. Dans les tests de combinatoire

inspirés du principe de « *randomisation»* de Fisher, cette hypothèse nulle est ré-interprétée et formulée comme suit:

Hypothèse nulle d'un test de combinatoire

> Si seul le hasard a influencé la valeur des données dans la configuration observée, cette configuration est ordinaire et la statistique calculée à partir de celle-ci constitue un résultat ordinaire, qui devrait apparaître avec une forte probabilité et vers le centre de la distribution des résultats issus de toutes les configurations possibles.

En regard de cette nouvelle hypothèse[1], le test indiqué en (8.13) ou (8.14) permet de décider contre l'hypothèse nulle ou en sa faveur.

Le principe des tests de combinatoire peut être appliqué de façon presque universelle dès qu'il s'agit de comparer deux ou plus groupes de sujets ou deux ou plusieurs occasions de mesure du même groupe. Edgington (1980) fait le tour des situations (voir aussi Laurencelle, *op. cit.,* et Siegel 1956).

> Le test le plus courant ou, du moins, le plus connu, concerne la comparaison de deux groupes d'observations, par exemple un groupe de sujets expérimentaux *versus* un groupe de sujets témoins. Soit n_1 et n_2, les tailles respectives des groupes, et soit S_o, la valeur observée de la statistique. La statistique « S » peut être la différence entre les moyennes des groupes, ou bien le quotient ou la différence de leurs variances, écarts-types ou étendues, etc. Il y a $T = \binom{n_1+n_2}{n_1}$ configurations possibles des résultats, c'est-à-dire T répartitions distinctes des n_1+n_2 sujets en deux groupes, et T valeurs correspondantes de S. Sous l'hypothèse nulle, la valeur observée S_o est une valeur ordinaire, logée centralement dans la distribution des T valeurs possibles. Si, en vertu du calcul (8.13), la valeur S_o apparaît exceptionnelle, l'on rejette l'hypothèse nulle et l'on conclut à une différence réelle, irréductible au hasard, entre les deux groupes.

1. Sur les plans du modèle et de la règle d'inférence appliqués, les deux hypothèses sont disparates. En effet, le test paramétrique classique «généralise» à partir de tous les échantillons virtuels issus du même univers que l'échantillon observé, alors que le test de combinatoire «généralise» à partir de l'ensemble (fini) des configurations déterminées par la combinatoire du test. Dans le premier cas, on se demande si le type de résultats observé (par exemple, une différence de moyennes) se retrouvera en général dans les autres échantillons; dans le second cas, si d'autres échantillons, placés dans des conditions comparables, donneront naissance au même type de configuration exceptionnelle. W. Hoeffding (1952) démontre l'équivalence asymptotique de ces deux approches.

La comparaison des résultats mesurés chez les mêmes sujets à deux occasions différentes, par exemple « Avant » et « Après » une intervention expérimentale, donne lieu à une combinatoire contenant 2^n configurations distinctes, qui proviennent de l'interversion des données de chaque sujet. L'étude de la corrélation entre deux séries de n données, disons { X. } et { Y, }, engendre $n!$ configurations, et statistiques correspondantes, selon les $n!$ permutations de l'une des séries. L'énumération combinatoire s'applique aussi, en particulier, aux plans d'expérience et leurs analyses de variance, la solution pratique restant contrainte par l'habileté du programmeur et, surtout, par la puissance des moyens informatiques mis en œuvre.

Advenant le cas fréquent où la taille de la combinatoire (T) interdise une solution exacte, par énumération complète, des solutions approximatives *de même nature* restent possibles grâce à l'échantillonnage Monte Carlo. Hope (1968) et surtout Laurencelle (1987) rendent compte de la puissance comparative de ces solutions.

Exercices

AVERTISSEMENT: Les exercices 8.1 à 8.9 constituent tout au plus un rappel et une occasion de rafraîchissement des règles d'intégration analytique, et encore nous n'y mettons en jeu que les règles d'application les plus courantes. Le lecteur novice en ces matières fera bien de consulter les textes spécialisés pour une explication plus complète et un ensemble d'exercices suffisants.

8.1 *Intégrale d'une somme.* Appliquant la règle:

$$\int u(x) + v(x) + w(x)\, dx = \int u(x)\, dx + \int v(x)\, dx + \int w(x)\, dx,$$

montrer que $\int x^2 - 2x\, dx = \frac{1}{3}x^3 - x^2 + C$, C étant une constante à déterminer.

8.2 *Intégrale d'une fonction u_x de x.* Appliquant la règle:

$$\int u_x^n\, du = u_x^{n+1} / (n+1) + C\ ,$$

montrer que $\int (1 + 4x)^3\, dx = (1 + 4x)^4/16 + C$. [Suggestion: Puisque, ici, $u = (1+4x)$, on a alors $du = 4\, dx$. On peut donc faire $(1+4x)^3\, du = \frac{1}{4}(1+4x)^3 \cdot 4\, dx$, la solution étant conséquente.]

8.3 *Intégration par parties.* La règle:

$$\int v\, du = uv - \int u\, dv$$

provient de l'autre règle sur la dérivée d'un produit de fonctions, soit $d(u \cdot v) = v \cdot du + u \cdot dv$, *du* et *dv* dénotant les dérivées de *u* et v respectivement. Appliquant cette règle, montrer que $\int x^2 \log_e x\, dx = \frac{1}{3}x^3 \log_e x - x^3/9 + C$. [Suggestion: substituer $u = \log_e x$ et $dv = x^2\, dx$; alors $du = x^{-1}\, dx$ et $v = \frac{1}{3}x^3$; l'application de la règle donne la solution.]

8.4 *Réduction en fractions partielles.* Soit une fonction apparaissant comme un quotient de polynômes, $f(x)/g(x)$, le quotient étant réduit de sorte que le numérateur soit de degré inférieur au dénominateur. L'intégration est facilitée par la réduction du quotient en fractions partielles correspondant aux différents facteurs (et puissances) du dénominateur $g(x)$, selon l'équivalence:

$$\frac{f(x)}{g(x)} = \frac{A}{g_1(x)} + \frac{B}{g_2(x)} + \cdots$$

Par cette technique, montrer que $\int (x^4-x^3-x-1) / (x^3-x^2)\, dx = \frac{1}{2}x^2 - 1/x + 2\log_e[x/(x-1)] + C$. [Suggestion: D'abord, le quotient se réduit à $x-(x+1)/(x^3-x^2)$. Puis, considérant que le dénominateur (x^3-x^2) a comme facteurs x^2, x et $(x-1)$, il s'agit d'identifier l'équation $(x+1)/(x^3-x^2) = A/x + B/x^2 + C/(x-1)$, obtenant $A = -2$, $B = -1$ et $C = 2$. Le calcul final suit, en se rappelant que $\int 1/x = \log_e |x| + C$.]

8.5 *Changement de variable.* Si une fonction u_x s'avère d'intégration difficile, on peut la remplacer par une autre, selon $y_x = f(u_x)$, et en calculant:

$$\int_a^b u_x\, du = \int_{f^{-1}(a)}^{f^{-1}(b)} y_x\, d[f^{-1}(y)] .$$

Appliquant cette règle (en plus de la réduction en fractions partielles), montrer que $\int x^{-1}(1-x)^{-\frac{1}{2}}\, dx = \log_e \left| \frac{1-\sqrt{1-x}}{1+\sqrt{1-x}} \right| + C$.

[Suggestion: Utiliser l'équivalence $z^2 = 1-x$, de sorte que $\int x^{-1}(1-x)^{-\frac{1}{2}}\, dx = -2\int 1/(1-z^2)\, dz = -2\int A/(1+z) + B/(1-z)\, dz$, etc.]

8.6 Montrer que: (a) $\int (x-2)^{3/2}\, dx = \frac{2}{5}(x-2)^{5/2}+C$; (b) $\int x/(x^2-1)\, dx = \frac{1}{2}\log_e|x^2-1|+C$; (c) $\int \sin^2 x \cos x\, dx = \frac{1}{3}\sin^3 x+C$; (d) $\int x^2/\sqrt{(1-x^6)}\, dx = \frac{1}{3}\sin^{-1}x^3+C$; (e) $\int \sqrt{(x^2+4x)}\, dx = \frac{1}{2}(x+2)\sqrt{(x^2+4x)} - 2\log_e|x+2+\sqrt{(x^2+4x)}|+C$; (f) $\int x^2 e^{-3x}\, dx = -\frac{1}{3}e^{-3x}(x^2+\frac{2}{3}x+\frac{2}{9})+C$; (g) $\int 2x^3/(x^2+1)^2\, dx = \log_e(x^2+1)+1/(x^2+1)+C$.

8.7 *Loi Gamma, Ga(k,β).* Une v.a. positive ($x \geq 0$) obéit à la loi *Gamma,* de paramètres k et β, si sa fonction de densité est telle qu'en (5.20). La loi Gamma standard est obtenue en fixant $\beta = 1$; noter aussi qu'une v.a. $Ga(k,\beta)$ peut être vue comme l'addition de k v.a. exponentielles $E(\lambda)$ (exercice 4.11), à paramètre égal à $\lambda = 1/\beta$. Pour la loi standard ($\beta = 1$) et k entier, montrer que la f.r. $F(x)$ est donnée par:

$$F(x) = 1 - e^{-x}[1 + x + x^2/2! + x^3/3! + \cdots + x^{k-1}/(k-1)!] . \quad (8.15)$$

[Suggestion: utiliser l'intégration par parties et procéder par induction, à partir de k = 1, 2, etc. Noter que la loi du Khi-deux avec v degrés de liberté (cf. §4.8) est en fait une $Ga(\frac{1}{2}v,\frac{1}{2})$.]

8.8 (Suite du précédent) Montrer que les moments d'une v.a. obéissant à la loi *Ga(k,(3)* sont:

$$E(x)=k\beta \ ; \ var(x)=k\beta^2 \ ; \ \gamma_1=2/\sqrt{k} \ ; \ \gamma_2=6/k. \qquad (8.16)$$

[Suggestion: en procédant par intégration par parties, montrer d'abord que les moments à l'origine μ'_r sont $\beta T(a+r)/\Gamma(\alpha).$]

8.9 *Le quotient de Mill.* La série de Taylor exploitée à l'exemple 8.1 pour trouver $\Phi(x)$, la f.r. normale, converge plus ou moins efficacement selon la grandeur absolue de x: pour x « fort » (e.g. $|x| > 3$, ou 5, ou 8), la convergence en utilisant des nombres à précision finie est douteuse. L'intégrale de la queue de la densité normale admet néanmoins une solution spéciale, à partir du « quotient de Mill », soit $R(x) = [1-\Phi(x)]/\varphi(x)$. Montrer que, pour x fort, le recours au quotient $R(x)$ équivaut au calcul de:

$$\phi(x) \simeq 1 - \frac{e^{\frac{1}{2}x^2}}{\sqrt{2\pi}}\left[\frac{1}{x} - \frac{1}{x^3} + \frac{1\cdot 3}{x^5} - \frac{1\cdot 3\cdot 5}{x^7} + \cdots\right], \qquad (8.17)$$

où x est pris en valeur absolue. Montrer qu'en utilisant seul le premier terme de la série, soit $1/x$, la précision atteinte est de 10^{-6} dès $|x| \geq 4{,}2$, ou de 10^{-9} dès $|x| \geq 5{,}5$.

8.10 *Intégration par trapèzes.* L'aire d'un trapèze de base b et de côtés c_1 et c_2 est $\frac{1}{2}b\bullet(c_1+c_2)$, en unités appropriées, conformément à la formule (8.8b). L'estimateur d'intégrale par la règle généralisée des trapèzes est alors:

$$\hat{Q} = \tfrac{1}{2}\Delta x \ \{ f(a)+2f(a+\Delta x)+2f(a+2\Delta x)+...+f(b) \} \ . \qquad (8.18)$$

Déterminer la différence algébrique entre cet estimateur et celui de la règle (8.10) utilisant des rectangles. Pour l'exemple 8.1, montrer que l'aire approchée avec 7 points et 6 trapèzes est moins précise que par la technique des rectangles.

8.11 Conformément à l'indication donnée en §8.5, étudier l'erreur encourue par l'approximation de l'intégrale normale '(x) au moyen d'un seul arc de cercle, $\frac{1}{2}+Q_x(A_1)$. Montrer que, dans l'intervalle (0; 1), ε_{max} = *max[Φ(x)* — $\frac{1}{2}$—$Q_x(A_1)] < 0{,}00018$; pour $x > 1$, on observe ε_{max}= 0,0001 dans le sous-intervalle (1; 1,836); 0,001 dans (1; 2,403); 0,01 dans (1; 4,815).

8.12 *Règle ⅜ de Simpson.* Appliquer la règle indiquée en (8.8d) au problème de l'exemple 8.2 et comparer sa précision à celle des autres règles d'intégration.

8.13 La documentation (voir Gerald et Wheatley, *op. cit.,* p. *245* et suiv.) rapporte les marges d'erreur suivantes pour (a) la règle des rectangles, l'abscisse étant prise sur la borne *a ou b: ½(b–a) Δxf' (x); (b) idem,* l'abscisse étant prise au point-milieu (on parle aussi de la *règle du point-milieu): (b–a)/24xΔx²f"*(x); (c) la règle des trapèzes: –(b–a)/12Δx²f"(x) ; (d) la règle ⅓ de Simpson: –(b–a)/180x $\Delta x^4 f^{iv}(x)$; (e) la règle ⅜ de Simpson: *–(b–a)/80 x* $\Delta x^4 f^{iv}(x)$, *où* $f^r(x)$ indique la valeur de la r^e dérivée de la fonction f à intégrer évaluée au point *x,* selon $a \leq x \leq b$. Vérifier que les écarts d'approximation des divers exemples du chapitre s'inscrivent dans les bornes indiquées. Noter que la règle ⅜ de Simpson (avec 4 valeurs de référence) est légèrement moins performante que la règle ⅓ (avec 3 valeurs).

8.14 *Technique d'extrapolation de Romberg.* La technique de Romberg est en fait un procédé d'accélération; elle consiste à extrapoler la valeur d'une approximation d'intégrale Δx *à* la valeur attendue en portant mathématiquement Δx *à* zéro.

Supposant pour la règle des trapèzes que l'erreur est d'ordre Δx^2, on peut écrire l'équation approximative:

$$Q = \hat{Q}(T_k) + C(\Delta x)^2 ,$$

l'estimateur $\hat{Q}$ étant basée sur k trapèzes. Fixant deux valeurs de k, donc deux valeurs de l'intervalle Δx (puisque $k \cdot \Delta x = b-a$), on calcule deux valeurs $\hat{Q}_1$ et $\hat{Q}_2$. Ces données permettent d'écrire un système à deux équations grâce auquel on peut estimer C, puis enfin $Q = \hat{Q}(T_\infty)$.

Par exemple, l'aire de $\int e^{-x}\, dx$, entre 0 et 1, approchée par un trapèze et $\Delta x = 1$ est $\hat{Q} \approx 0{,}68394$; avec 2 trapèzes, on a $\Delta x = ½$ et $\hat{Q} \approx 0{,}64524$. Posant $Q = 0{,}68394 + C(1)^2 = 0{,}64524 + C(½)^2$, on obtient $C \approx -0{,}0516$ et $\hat{Q}(T_\infty) \approx 0{,}63234$, encourant une erreur d'à peine 0,03 %. Appliquer cette technique à partir de la règle ⅓ de Simpson en considérant une erreur de $C(\Delta x)^4$. Montrer comment cette technique peut s'appliquer en chaîne.

8.15 *Intégration de Gauss-Legendre.* Plutôt que d'utiliser des abscisses prédéfinies, voire équidistantes, pour estimer la valeur de l'intégrale $\int_a^b f(x)\, dx$, la méthode de Gauss-Legendre pose celles-ci comme autant de paramètres supplémentaires. De cette façon, une fonction

d'estimation $\hat{Q} = \sum c_i f(x_i)$, exploitant n abscisses, $a \le x_1 < x_2 < \cdots < x_n \le b$, n ordonnées $f(x_i)$ et n coefficients c_i, correspond à un polynôme de degré $2n-1$. L'erreur d'estimation est ici proportionnelle à $f^{2n}(x)$: voir Davis et Rabinowitz (1967). Une liste partielle des abscisses et coefficients utiles (tirée de Scheid 1988) apparaît au tableau 1, en contrebas: les valeurs sont standardisées pour l'intervalle d'intégration $y \in (-1, 1)$.

Appliquer la méthode d'intégration de Gauss-Legendre au problème de l'exemple 8.2 et montrer que, dès $n = 2$, l'erreur relative est à peine plus forte que $-0{,}02$ %.

Tableau 1 Abscisses relatives de l'intervalle $(-1, 1)$ et coefficients de l'approximation Gauss-Legendre à n points†

n	y_i	c_i	n	y_i	c_i
2	±0,57735027	1,00000000	10	±0,97390653	0,06667134
4	±0,86113631	0,34785485		±0,86506337	0,14945135
	±0,33998104	0,65214515		±0,67940957	0,21908636
6	±0,96246951	0,17132449		±0,43339539	0,26926672
	±0,66120939	0,36076157		±0,14887434	0,29552422
	±0,23861919	0,46791393	12	±0,98156063	0,04717534
8	±0,96028986	0,10122854		±0,90411725	0,10693933
	±0,79666648	0,22381034		±0,76990267	0,16007833
	±0,52553241	0,31370665		±0,58731795	0,20316743
	±0,18343464	0,36268378		±0,36783150	0,23349254
				±0,12533341	0,24914705

† Pour l'intégration dans $x \in (a, b)$, trouver d'abord $x = \frac{1}{2}[y(b-a) + (b+a)]$, puis calculer $\hat{Q} = \frac{1}{2}(b-a)\sum c_i f(x_i)$.

8.16 Soit une collection de n symboles (ou nombres) à partir desquels on forme des ensembles de n éléments; chaque élément peut apparaître de zéro à plusieurs fois. Montrer que le nombre T d'ensembles distincts (sans considération d'ordre) est:

$$T(n) = (2n-1)! \,/\, [n!(n-1)!]\,. \tag{8.19}$$

[Suggestion: considérer que T(1)=1, T(2)=3, T(3)=10, ... soit $T(n) = 1 \times 3 \times (3+\frac{1}{3}) \times \cdots \times (3+(n-2)/n)$. Par l'approximation de Sterling, i.e. $n! \approx \sqrt{(2\pi n)}\cdot(n/e)^n$, montrer que $T(n) \approx 2^{2n-1}/\sqrt{(n\pi)}$.]

8.17 (Suite du précédent) Montrer qu'en général le nombre G d'ensembles distincts en *p* positions qu'on peut former à partir de s symboles est:

$$G(s,p) = (s + p - 1)! / [(s-1)!\, p!]. \qquad (8.20)$$

8.18 Utilisant (8.20), trouver une expression indiquant le taux de répétition échantillonnale impliquée dans une étude de bootstrap.

8.19 Dans une population de taille N, on échantillonne avec remise n éléments: on calcule une moyenne simple, $m_r = \sum x/n$, et une autre basée sur les n_d éléments *distincts* (en rejetant les éléments *repris*), $m_d = \sum x/n_d$; noter que $n_d \leq n$. Soit le quotient de variances $R(n) = var(m_d)/var(m_r)$; montrer que $R(2) = 1$, $R(3) = 1-1/(2N)$; $R(4) = 1-1/N$; $R(5) = 1-3/(2N)+1/(6N^2)+1/(6N^3)$ et que, en général, $var(m_d) < var(m_r)$ pour $n \geq 3$ (Raj et Khamis 1958; Gabler 1985). Se rappeler que la variance d'une moyenne de n_d éléments pigés sans remise est $\frac{\sigma^2}{n_d}[(N-n_d)/(N-1)]$.

8.20 (Suite du précédent) Dans une population de taille N, nous échantillonnons jusqu'à obtenir n_d éléments distincts, amassant en ce faisant un total de n éléments. Montrer encore que $var(m_d) < var(m_r)$ pour $n_d \geq 2$. Pour $n_d = 2$ par exemple, montrer que $var(m_r) = O^2$ (N -1) $\{N\log_e[N/(N-1)] - 1\}$ et $var(m_d)/var(m_r) < 1$ et $\to 1$ si $N \to \infty$ utiliser $N^{-1} + 2N^{-2} + 3N^{-3} + ... = \log e[N/(N-1)]$.

8.21 Utilisant la distribution *d'occupation* (6.13) et ses moments (6.15), trouver l'espérance et la variance de n_d, le nombre d'éléments distincts, pour une population de taille N et un échantillon avec remise de n éléments. Montrer sous quelle condition de relation n:N l'effort requis pour réaliser un échantillonnage *sans remise* est superflu.

8.22 Composer un algorithme général du bootsrap et rédiger un schéma de programme informatique pour estimer l'espérance, la variance et les quantiles d'une statistique $S = S(x^*_1, x^*_2, ..., x^*_n)$ à partir de T' auto-échantillons, les valeurs x^*_i étant pigées avec remise dans la série observée $(x_1, x_2, ..., x_n)$.

8.23 Composer un algorithme général de test de différence entre deux groupes de données par combinatoire exhaustive, utilisant une statistique $S = S(G_1, G_2)$, à définir.

8.24 (Suite du précédent) Rédiger un programme de test de différence par combinatoire exhaustive pour le cas où la statistique est la différence des moyennes, $S = \bar{x}_1 - \bar{x}_2$. Montrer que S est linéairement équivalent et a même distribution que $\sum_1$ ou $-\sum_2$, où $\sum_k$ est la somme des données d'un groupe. Étudier d'autres procédés d'accélération de cette comparaison, par exemple en faisant préalablement le tri des n_1+n_2 observations en ordre décroissant: voir P. Ferland et L. Laurencelle: « Un algorithme efficace pour la comparaison de deux moyennes indépendantes par combinatoire exhaustive », *Lettres Statistiques*, 1993, vol. 9, p. 93-114.

8.25 Composer un algorithme général d'étude des *n!* permutations d'une série de *n* observations. Appliquer cet algorithme dans un programme qui calcule une statistique $S = S(x_1, x_2, ..., x_n)_p$, qui reflète une propriété d'ordre de chaque permutation p, par exemple la corrélation entre une série $\{y_i\}$ rigide et une série $\{x_i\}$ permutée, ou le nombre de suites monotones (§6.17) dans la série $\{x_i\}$.

8.26 Composer un algorithme général de test des différences entre k séries *(k = 2,* 3, etc.) de *n* données jumelées (ou « blocs aléatoires »), à travers une statistique S basée sur les moyennes respectives (ou totaux) des *k* séries.

Références

AYRES, F. JR. (1972). *Calcul différentiel et intégral: théorieet applications.* New York, McGraw-Hill (Série Schaum).

CONLEY, W. (1984). *Computer optimization techniques.* New York, Petrocelli.

DAVIS, P.J., RABINOWITZ, P. (1967). *Numerical integration.* Waltham (MA), Blaisdell Publishing.

EDGINGTON, E.S. (1980). *Randomization tests.* New York, Marcel Dekker.

EFRON, B. (1982). *The Jacknife, the Bootstrap and other resampling plans.* Philadelphia (PA), SIAM Monograph (no 38).

EFRON, B., GONG, G. (1983). A leisurely look at the Bootstrap, the Jacknife and cross-validation. *The American Statistician, 37,* 36-48.

FISHER, R.A. (1936). « The coefficient of racial likeness » and the future of craniometry. *Journal of the Anthropological Institute, 66,* 57-63.

GABLER, S. (1985). On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement. *Biometrika, 71,* 171-175.

GERALD, C.F., WHEATLEY, P.O. (1984). *Applied numerical analysis* (3e édition). Reading (MA), Addison-Wesley.

HOEFFDING, W. (1952). The large sample power of tests based on permutations of observations. *Annals of mathematical statistics, 23,* 169-192.

HOPE, A.C.A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal StatisticalSociety B, 30,* 582-598.

KINCAID, D., CHENEY, W. (1991). *Numerical analysis.* New York, Brooks/Cole.

LAURENCELLE, L. (1987). Le nombre de permutations dans les tests permutationnels. *Lettres Statistiques, 8,* 49-80.

RAJ, D., KHAMIS, S.H. (1958). Some remarks on sampling with replacement. *Annals of mathematical statistics, 29,* 550-557.

ROUGIER, J.-P. (1987). *Méthodes de calcul numérique* (3e édition). Paris, Masson.

SCHEID, F.J. (1988). *Schaum 's outline of theory and problems of numerical analysis* (2e édition). New York, McGraw-Hill.

SCHWARTZ, A. (1967). *Calculus and analytic geometry* (2e édition). New York, Holt, Rinehart and Winston.

SIEGEL, S. (1956). *Nonparametricstatistics.* New York, McGraw-Hill.

SWOKOWSKI (1993). *Analyse* (Trad. M. Cetta, 5e édition). Bruxelles, De Boeck Université.

Chapitre
9 La méthode Monte Carlo: les bases

9.1 La méthode Monte Carlo a des applications très diverses, comme on l'a vu au chapitre 7, mais ces applications ont cependant des points en commun. La méthode, appuyée le plus souvent sur un support informatique, utilise des nombres aléatoires; elle est appliquée afin d'évaluer une quantité approximativement ou estimer les caractéristiques dynamiques d'un phénomène, cela au moyen d'un calcul, d'une évaluation numérique. L'objet d'étude peut être un *modèle* représentant un système complexe ou supportant un phénomène; il s'agit ou d'un modèle quantitatif ou d'un modèle qualitatif dont on veut caractériser numériquement certains aspects. La méthode sert aussi à estimer une *moyenne,* une *intégrale* définie, qui seraient inabordables par une méthode exacte. La définition suivante englobe tous ces aspects :

> *La méthode Monte Carlo est un calcul de fonctions de nombres aléatoires pour l'évaluation approximative d'une quantité définie ou pour l'étude dynamique d'un phénomène quantitatif ou qualitatif.*

Comme l'indique la définition ci-dessus, la *modélisation* et la méthode Monte Carlo, malgré leur air de parenté, n'ont en commun que leur utilité réciproque. En effet, avant qu'un système, un phénomène puisse être étudié par la méthode Monte Carlo ou par un autre moyen, il doit être analysé, puis réduit à sa forme essentielle, formulé dans un réseau de paramètres et de relations paramétriques qui constitue son *modèle.* La modélisation, en fait la modélisation mathématique, forme un domaine en soi (Dym et Ivey 1980). L'obtention du modèle peut, dans certains cas, terminer l'étude du système ou du phénomène. On peut aussi obtenir satisfaction en élaborant les propriétés formelles du modèle, par exemple par calcul (équations différentielles, calcul différentiel et intégral) ou au moyen des techniques dites de recherche opérationnelle (Moder et Elmaghraby 1978) : ces dernières approches constituent une *simulation déterministe du* modèle. La simulations tochastique, qui est une application de type Monte Carlo, consiste à évaluer le modèle en le faisant vivre ou fonctionner d'une façon réaliste et comportant un ou plusieurs aspects

aléatoires. Naylor (1971), Bratley, Fox et Schrage (1987) et d'autres détaillent le processus et les étapes de l'étude d'un système par simulation.

9.2 *L'espérance d'une variable aléatoire.* L'application la plus simple de la méthode Monte Carlo consiste à étudier une fonction d'une variable aléatoire afin d'en établir les caractéristiques, notamment l'*espérance*, ou « valeur attendue ». L'estimation Monte Carlo étant basée sur l'accumulation d'évidence à partir d'échantillons, l'espérance mathématique $E(x)$ est approchée par la moyenne arithmétique $\bar{x}$, où $\bar{x} = \sum x_i / T$. La moyenne est un estimateur sans biais de l'espérance mathématique, c'est-à-dire $E(\bar{x}) = E(x) = \mu$, quel que soit le nombre $(T)^1$ d'échantillons moyennés. De plus, l'erreur-type de cette moyenne est:

$$\sigma_\varepsilon = \sigma(\bar{x} - \mu) = \sigma(\bar{x}) = \sigma_x / \sqrt{T}, \qquad (9.1a)$$

où »O_x dénote l'écart-type de la statistique analysée. À défaut de connaître O_x, l'écart-type s_x, basé sur les valeurs x_i des T échantillons, est employé dans :

$$\hat{\sigma}_\varepsilon = s / \sqrt{T}. \qquad (9.1b)$$

L'erreur-type nous indique la précision de la valeur estimée obtenue. L'exemple suivant clarifie ces idées.

Exemple 9.1 L'espérance de la moyenne harmonique de deux v.a. uniformes

La moyenne harmonique $\bar{x}_h$, définie généralement comme l'inverse de la moyenne arithmétique des inverses, a comme formule générale

$$\bar{x}_h = k / [x_1^{-1} + x_2^{-1} + \cdots + x_k^{-1}] .$$

Le comportement de cette statistique ne paraît pas avoir été étudié, sauf par Laurencelle (1993) qui l'établit pour $k = 2$ et $k = 3$ *v.a.* uniformes.

1. Nous utilisons la lettre «T» (majuscule) pour indiquer la taille d'une expérimentation Monte Carlo, au lieu de la lettre habituelle «n» ou «N», cette dernière étant souvent réservée pour caractériser la quantité étudiée elle-même (e.g. détermination de l'espérance de la moyenne harmonique de n v.a. uniformes). Quant au symbole d'écart-type «O» (ou de son estimateur «s»), il sera nécessaire de l'identifier spécifiquement à chaque occasion particulière.

Supposons qu'on dispose d'une suite infinie de v.a. de distribution $U(0,1)$, dénotées u_i, u_{i+1}, u_{i+2}, ···. Le schéma de programme suivant permet d'estimer l'espérance de $\bar{x}_h$, la moyenne harmonique de deux v.a. uniformes, de même que l'erreur-type de cette estimation.

Estimer l'espérance $Q = E[\bar{x}_h]$ *par la moyenne de T échantillons Monte Carlo*

[Initialisation] $n \leftarrow 0$; $\Sigma_1 \leftarrow 0$; $\Sigma_2 \leftarrow 0$
[Cycle] Exécuter T fois
Obtenir deux v.a. u' et u''
$\bar{x}_h \leftarrow 2/(1/u' + 1/u'')$
$\Sigma_1 \leftarrow \Sigma_1 + \bar{x}_h$; $\Sigma_2 \leftarrow \Sigma_2 + \bar{x}_h^2$
[Calculs finals] $\hat{Q} = \Sigma_1 / T$
$V_\varepsilon(\hat{Q}) = (\Sigma_2 - \Sigma_1^2/T)/[T(T-1)]$.

À titre illustratif, nous avons rédigé un court programme calculant $\hat{Q}$ selon l'algorithme esquissé ci-dessus. Utilisant T = 10000 échantillons, nous obtenons une première fois $\hat{Q} = 0{,}40980$ et $\hat{\sigma}_\varepsilon = \sqrt{V_\varepsilon} = 0{,}00246$. Les valeurs $(\hat{Q} \pm 2\hat{\sigma}_\varepsilon)$ fournissent un intervalle de confiance approximatif de 95% pour la valeur Q à estimer, ici (0,40488; 0,41472). Une deuxième estimation, pareille à la première, fournit $\hat{Q}$=0,40578 et $\hat{\sigma}_\varepsilon$=0,00245, d'où l'intervalle (0,40088; 0,41068).

Dans le cas du présent exemple, Laurencelle (1993) montre que $E(\bar{x}_h) = 0{,}409137$ et $\sigma^2(\bar{x}_h) = 0{,}060018$, d'où $\sigma_\varepsilon(\hat{Q}) = \sqrt{(0{,}060018/T)}$. Avec T = 10000, $\sigma_\varepsilon(\hat{Q}) \approx 0{,}00245$, valeur dont s'approchent nos estimations d'erreur-type ci-dessus. Pour obtenir une estimation $\hat{Q}$ telle qu'elle ait une précision dénotée ε au degré de confiance de 95%[2], soit:

$$\text{pr}\{\ |\hat{Q} - Q| \leq \varepsilon\} = 0{,}95\ ,$$

il faut que l'intervalle $(\hat{Q} \pm 2\hat{\sigma}_\varepsilon)$ ait une étendue égale à ε, d'où:

$$T \approx 16\sigma^2/\varepsilon^2\ ,$$

2. L'intervalle serait rigoureux en probabilité si la statistique étudiée (ici, $\bar{x}_h$) était de distribution normale *et* la taille échantillonnale (T) était infinie. Toutefois, pour des valeurs raisonnablement fortes de T, soit T > 30, la moyenne $\hat{Q}$ de toute statistique est approximativement normale et la distribution échantillonnale t_{T-1} (de *Student*), qui conviendrait ici, se confond avec la loi normale, de sorte que l'intervalle proposé est approximativement juste.

expression dans laquelle O^2 est la variance de la statistique étudiée. Ici, utilisant $O^2 \approx 0{,}0600$, il faudrait calculer une moyenne basée sur 960 000 échantillons pour espérer une précision de $\varepsilon = 0{,}001$.

9.3 *L'intégrale d'une fonction.* L'évaluation approximative d'une intégrale constitue une autre application majeure de la méthode Monte Carlo. Dans le cas le plus simple, celui d'une fonction d'une variable $f(x)$, il s'agit d'estimer l'aire sous le tracé de la fonction dans l'intervalle borné par $x = a$ et $x = b$. Dans les cas plus complexes, la fonction à intégrer est multivariée et l'aire à estimer correspond à un volume k-dimensionnel : le lecteur peut imaginer facilement des situations où l'intégration analytique pose un défi herculéen, situations dans lesquelles l'estimation Monte Carlo est plus que bienvenue!

La technique dite de *sondage de fonction* est l'une des premières mises en oeuvre pour estimer une intégrale simple. Elle consiste *grosso modo* à définir une surface dans laquelle s'inscrit la fonction étudiée, à lancer des « coups de sonde » à l'aveuglette dans cette surface puis à calculer la proportion P de coups de sonde qui tombent sous la fonction. Cette proportion P, qui constitue en même temps une estimation de probabilité, reflète proportionnellement l'aire cherchée. Soit Q, l'intégrale définie à déterminer, S l'aire de la surface englobant la fonction et P la proportion de coups de sonde admissibles, alors :

$$\hat{Q} = S \cdot P \tag{9.2}$$

donne l'estimation d'aire voulue.

L'estimation d'intégrale par la technique de sondage de fonction exploite une variable de loi binomiale, $t \sim B(T, Q/S)$, où T représente le nombre total de coups de sonde, le quotient Q/S est la proportion juste de l'aire de la fonction à déterminer par rapport à la surface qui l'englobe, et t le nombre de coups de sonde tombant sous la fonction à intégrer. D'après les caractéristiques d'une v.a. binomiale (exercice 4.2), la statistique $\hat{Q} = S \cdot P = S \cdot t/T$ estime sans biais l'aire Q cherchée, avec l'erreur-type :

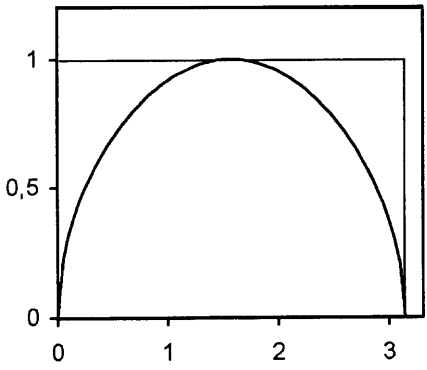
$$\sigma_\varepsilon(\hat{Q}) = \sqrt{\frac{Q(S-Q)}{T}} \; ; \tag{9.3a}$$

plus réalistement, en utilisant la proportion $P = t / T$ mesurée, on a:

$$\sigma_\varepsilon(\hat{Q}) = \sqrt{\frac{\hat{Q}(S-\hat{Q})}{T-1}} = S\sqrt{\frac{P(1-P)}{T-1}} \; . \tag{9.3b}$$

Exemple 9.2 L'intégrale de (sin x) ^{iz} *de 0 à* Tc.

La fonction étudiée, $f(x) = \sqrt{\sin x}$, est illustrée ci-contre, dans l'intervalle $(0,\pi)$. On peut montrer que l'aire ainsi définie est $Q = \int (\sin x)^{1/2}\, dx \approx 2{,}39628$.



Pour estimer Q, considérons une v.a. uniforme $x \sim U(0,\pi)$: cette variable occupe tout l'espace cadré dans le rectangle $((0,\pi) \times (0,1))$, d'aire $S = \pi$: appelons cet espace E_S. La technique consiste alors à « lancer » une variable x dans E_S puis à vérifier si x tombe ou non sous $f(x)$: la proportion de fois où x s'inscrira sous $f(x)$ permettra d'obtenir $\hat{Q}$ et son erreur-type, en appliquant (9.2) et (9.3). Le schéma de programme suivant indique la façon de faire.

Estimer l'intégrale définie $\int_a^b f(x)\, dx$ *par sondage de fonction*

{ Soit une fonction englobante $G(x)$ de domaine (a, b) et sous laquelle $f(x)$ est inscrite, selon $f(x) \le G(x)$, et $S = \int_a^b G(x)\, dx$ est connue et $Q = \int_a^b f(x)\, dx$ est à estimer }

[Initialisation] $t \leftarrow 0$
[Cycle] Exécuter T fois
Obtenir une v.a. $x \in (a,b)$ et $u \sim U(0,1)$
Si $f(x)/G(x) \ge u$ alors $t \leftarrow t + 1$
[Calculs finals] $\hat{Q} = S \times t/T$.

Pour notre exemple, avec $f(x) = (\sin x)^{1/2}$, $G(x) = 1$, $a = 0$ et $b = \pi$ ($\approx 3{,}14159$), nous avons $S = \pi$, $Q \approx 2{,}39628$ et, par (9.3a), $O_\varepsilon \approx 1{,}33640/\sqrt{T}$. Afin d'obtenir une précision de ε à probabilité de 0,95, il faudrait mettre en oeuvre $T = 16\, O^2 / \varepsilon^2 \approx 28{,}5756 / \varepsilon^2$ coups de sonde, soit, par exemple, autant que $T \approx 2858$ pour mériter une précision d'à peine 0,1!

La technique du sondage de fonction, comme on verra plus loin, ne se classe pas avantageusement dans le palmarès des techniques efficaces d'estimation d'intégrale. Outre sa relative ancienneté, son mérite premier

tient à ce qu'elle représente souvent, pour les intégrales multidimensionnelles, la seule solution pratiquement abordable.

9.4 Il existe maintes applications de la méthode Monte Carlo, au delà des deux exemples juste présentés; les exercices de ce chapitre et des suivants en donneront des illustrations. Ces applications sont néanmoins toutes des variantes des deux premières, lesquelles sont à leur tour des variantes l'une de l'autre.

Le calcul de l'*espérance* E(x) d'une variable aléatoire x est abordé en produisant des exemplaires aléatoires x_i de cette variable puis en calculant leur moyenne arithmétique; ce calcul correspond aussi à l'intégration $\int_a^b x \cdot f(x)\ dx = \int_a^b g(x)\ dx$, soit une intégrale simple dans laquelle $g(x) = x \cdot f(x)$. Noter que chacune des bornes a et b peut être infinie. De même, réciproquement, l'intégration d'une fonction peut être ramenée à un calcul d'espérance, ou de *moyenne*, grâce notamment au théorème sur la *valeur moyenne* d'une intégrale (Swokowski 1993), soit:

$$\int_a^b f(x)\ dx = \int_a^b x \cdot f'(x)\ dx$$
$$= \mathrm{m}(x) \cdot \int_a^b f'(x)\ dx\ ,$$

expression dans lesquelles on a factorisé $f(x)$ en $x \cdot f'(x)$. Ici, m(x) est la « valeur moyenne », et dans le cas d'une fonction de probabilité $f'(x) \geq 0$ et $\int f'(x)\ dx = 1$, m(x) est en même temps l'*espérance* de la v.a. x.

9.5 La méthode Monte Carlo, comme on a vu, est un procédé d'estimation statistique soumis, comme plusieurs procédés semblables, à la *loi de l'inverse de la racine carrée*: l'imprécision (ou « erreur ») dans la valeur estimée $\hat{Q}$ est, par exemple, coupée de moitié si l'on quadruple le nombre d'échantillons. Les équations d'erreur-type (9.1) et (9.3) précisent cette relation. Quant à la *variance d'erreur*, $\sigma_\varepsilon^2(\hat{Q}) = V_\varepsilon(\hat{Q})$, elle est inversément proportionnelle à la taille T, soit:

$$V_\varepsilon(\hat{Q}) = \sigma^2 / T\ , \qquad (9.4)$$

où O^2 dénote la variance de l'estimateur d'un échantillon à l'autre, appelée aussi *variance unitaire* dans le présent contexte.

La précision des réponses, dans une étude Monte Carlo, dépend donc à la fois de la taille échantillonnale T et de la variance (O^2) de la variable ou de la fonction étudiée. Cette précision, mesurée par l'erreur-type de l'estimateur Monte Carlo, restera souvent insatisfaisante à moins d'accroître la taille T énormément, rendant ainsi l'étude Monte Carlo trop longue ou trop coûteuse. L'efficacité du procédé Monte Carlo se présente donc comme

une question significative, et l'effort qu'on y consacrera nous ouvrira de nombreuses avenues d'amélioration.

Notons ici que, dans la mesure où elle réfère d'abord au coût d'exécution ou au nombre d'échantillons consommés, l'efficacité d'estimation varie de façon inversement proportionnelle à ce nombre. D'autre part, la précision d'estimation renvoie à l'intervalle d'incertitude de l'estimateur autour de la quantité visée et elle est proportionnelle à l'erreur type, donc à l'inverse de la racine carrée du nombre d'échantillons. Ainsi, toutes choses étant égales par ailleurs, un procédé d'estimation qui serait k fois plus efficace qu'un autre sera donc $\sqrt{k}$ fois plus précis.

Augmenter l'efficacité d'un procédé Monte Carlo vise en premier lieu l'obtention de résultats plus précis, ou moins coûteux en temps d'exécution informatique; le calcul à partir de nombres aléatoires *est imprécis* et l'estimation Monte Carlo ne sera parfois exploitable qu'après que sa précision ait été drastiquement augmentée. Il est par ailleurs normal, du moins dans une perspective scientifique, qu'on se préoccupe de la précision des procédés et des résultats. On peut utiliser aveuglément l'estimation Monte Carlo à coups de milliers et de milliers d'échantillons et en accepter le résultat naïvement : telles sont la simplicité et la force de cette méthode. Cependant, l'analyse du procédé, de son contexte statistique détaillé et l'étude de ses conditions de convergence nous renseigneront directement sur la valeur du résultat obtenu ou attendu en même temps qu'elles nous indiqueront les manières, les astuces susceptibles d'améliorer l'efficacité de l'estimation. Enfin, ce souci d'efficacité et les considérations fouillées qui s'en inspirent peuvent nous donner une compréhension différente, plus profonde, du modèle étudié.

9.6 *Indices d'efficacité Monte Carlo.* Dans le chapitre subséquent, nous passerons en revue des variantes du procédé Monte Carlo, variantes dont le but consiste à améliorer d'une façon ou d'une autre l'efficacité d'estimation. Ce faisant, il sera avantageux de disposer d'une mesure de cette efficacité afin de quantifier le mérite d'un procédé par rapport à un autre. Le mérite ainsi attribué reste circonstanciel, car il dépend : a) de l'indice d'efficacité retenu, b) du modèle Monte Carlo étudié, c) du programme informatique et de la machine exploités.

Le concept d'efficacité se présente d'emblée comme un ratio, confrontant les entités Valeur et Coût. La fonction du procédé Monte Carlo, telle qu'étudiée ici, est de produire une estimation $\hat{Q}$ d'une quantité Q. La « valeur » de l'estimation est évidemment liée à sa proximité de la quantité cible : l'erreur quadratique moyenne (EQM) :

$$EQM(\hat{Q}) = E[\ (\hat{Q} - Q)^2\ ] \qquad (9.5)$$

est la mesure la plus répandue de proximité. Si, comme c'est le cas usuel, l'estimation est de *biais nul*, à savoir:

$$B(\hat{Q}) = E(\hat{Q}) - Q = 0 , \tag{9.6}$$

alors l'EQM se ramène simplement à une variance d'erreur:

$$EQM[\hat{Q} \mid B(\hat{Q})=0] = V_\varepsilon(\hat{Q}) . \tag{9.7}$$

Quant à l'aspect « coût », sa quantification est plus délicate. Il y a d'abord le « coût préparatoire », disons $C_P(\hat{Q})$, reflétant l'investissement de temps ou d'argent nécessaire pour concevoir et mettre en place le procédé d'estimation. Se rajoute à cela le « coût d'exécution », $C_E(\hat{Q};T)$, mesuré en temps ou en argent, qui dépend évidemment du programme informatique et de la machine exploités, mais varie avant tout selon T, le nombre d'échantillons à partir desquels l'estimation est construite.

Retenons comme « valeur » de l'estimation la précision du résultat obtenu, spécifiquement l'inverse de sa variance d'erreur (9.4), qu'on multiplie par un coefficent arbitraire V$ afin de l'exprimer en valeur monétaire, soit $V\$/\sigma_\varepsilon^2(\hat{Q})$, et, comme « coût », la somme pondérée des deux coûts impliqués, soit $c_1\$C_P(\hat{Q})+c_2\$C_E(\hat{Q};T)$. Les deux quantités étant maintenant commensurables, on peut construire un indice basé sur leur différence ou bien sur leur quotient. L'indice habituel, un indice de quotient, met de côté le coût préparatoire et, stipulant $V\$/c_2\$ = 1$, il correspond à:

$$1 / [\sigma_\varepsilon^2(\hat{Q})\cdot C_E(\hat{Q};T)] .$$

Dans le contexte d'une étude Monte Carlo, le coût d'exécution, toujours relatif à la même machine et au même langage de programmation, est équitablement mesuré par le « temps d'exécution » $t(\hat{Q};T)$. La variance d'erreur (9.4) étant une fonction simple de la variance unitaire d'estimation, disons σ_1^2, nous pouvons définir l'*indice empirique d'efficacité*:

$$IE(\hat{Q}) = T / [\sigma_1^2\cdot t(\hat{Q};T)] ; \tag{9.8}$$

la variance unitaire σ_1^2 est soit connue, soit estimée ($\hat{\sigma}_1^2$) à partir des statistiques unitaires Q_1, Q_2, ···, Q_T qui servent à obtenir $\hat{Q}$. Le temps $t(\hat{Q})$ correspond à la durée totale d'exécution du programme. Cette durée totale $t(\hat{Q})$ découle du re-calcul itératif de $\hat{Q}$ à partir des T valeurs individuelles Q_i. Ainsi, cette durée totale est quasi proportionnelle à T, soit $t(\hat{Q}) \approx t_1\times T$.

> En fait, la durée totale t d'un programme Monte Carlo, comme de tout autre programme informatique, se décompose en trois phases

naturelles, $t = t_I + t_R + t_C$: t_I pour la phase *d'initialisation* (ou préparation des calculs); t_R pour la phase *itérative,* et *tc* pour la phase de *conclusion.* Seule la phase *itérative* est strictement proportionnelle à T, le nombre d'échantillons retenus, qui correspond presque toujours au nombre d'itérations dans cette phase. Néanmoins, et pour refléter plus fidèlement l'efficacité *globale* du procédé d'estimation, c'est la durée totale qui est retenue. Lorsque le nombre T est élevé, comme il est coutume, le quotient t_R/t 1, c'est-à-dire que la durée totale reflète très bien les itérations de calcul.

En conséquence, *l'indice unitaire d'efficacité,* qui ignore encore le coût d'analyse et de préparation du procédé Monte Carlo, est donné par :

$$IE_u(\hat{Q}) = 1 / [\sigma_1^2 \cdot t_1] , \qquad (9.9)$$

c'est-à-dire la réciproque du produit de la variance unitaire d'estimation (O^2_1) et du temps unitaire de production (t_1). Enfin, pour comparer l'efficacité de divers procédés, disons les procédés A et B pour estimer la quantité Q. le quotient de leurs indices d'efficacité respectifs :

$$ER(A,B) = IE(\hat{Q}_A) / IE(\hat{Q}_B) \qquad (9.10)$$

définit l'*efficacité relative* de l'un par rapport à l'autre. C'est cette mesure d'efficacité relative qui guidera notre quête d'améliorations dans l'estimation Monte Carlo.

9.7 *Analyse du coût temporel d'un algorithme.* Le temps total d'exécution d'un algorithme, dénoté $t(\hat{Q})$, peut être mesuré à l'aide d'un chronomètre externe ou encore dans le programme même, en exploitant une fonction de chronométrage si elle est disponible dans le système informatique utilisé. Une tout autre manière d'aborder l'évaluation du coût temporel passe par l'analyse formelle, ou symbolique, de l'algorithme : au lieu de compter les secondes ou millisecondes d'exécution réelle du programme, les coûts des phases, sous-phases et opérations majeures sont estimés en unités arbitraires, en tenant compte des paramètres-clés de l'algorithme étudié (Knuth 1969). Cette analyse agit comme une sorte de révélateur sur le texte de l'algorithme, en faisant ressortir les phases les plus dispendieuses de l'exécution, et elle indique assez précisément l'importance relative des paramètres. Son résultat prend la forme d'une équation de coût, de la forme :

$$t=c_0+c_1A+c_2B+\text{etc.},$$

A et B étant les paramètres globaux de l'algorithme ou des fonctions de ceux-ci, et c_0, c_1, c_2 des constantes.

L'analyse du coût temporel, que d'aucuns désignent « analyse d'efficacité », s'applique symboliquement et correspond aux opérations formelles d'une exécution informatique : en d'autres mots, l'analyse concerne davantage l'algorithme plutôt que le programme. Ce caractère plus abstrait a deux inconvénients : les mesures ne peuvent pas être aussi précises que si l'on énumérait et quantifiait la séquence des instructions-machine supportant les opérations, et les coûts sont exprimés en unités arbitraires. Par contre, l'analyse symbolique est de réalisation plus commode et, comme elle caractérise un algorithme, sa conclusion s'applique généralement à tous les programmes qui en découlent.

9.8 Mettant en pratique les directives précédentes à propos de l'analyse du coût d'un algorithme, établissons la convention suivante. L'exécution d'une opération « élémentaire » coûte 1 unité (de temps), les opérations un peu plus complexes engendrant 2 unités, et le reste à l'avenant. Le tableau ci-dessous présente une convention de coûts possible. Au besoin, la convention

Coût temporel	Opérations
0	Redirections, branchements («goto», «call», etc.)
1	« X ← », +, −, ×, /, abs, $()^2$, etc.
2	$\sqrt{}$, sin, atn, log, exp, etc.
≥ 3	RND, etc.

peut-être raffinée en recourant à une estimation particulière et empirique des coûts d'opération relatifs à un contexte d'exécution précis : par exemple, si les opérations «+», «— », «x» coûtent chacune 5 unités de temps, la division «/» peut en coûter 6 ou 7, la valeur absolue « abs » 2, et l'affectation, « X F← ()» 1 seule unité.

L'algorithme présenté à l'Exemple 9.1, plus haut, se prête bien à l'analyse symbolique. La première ligne comporte trois affectations, d'où $L_1 = 3$; les lignes du « Cycle » itératif correspondent respectivement à $L_2 = 1$, $L_3 = 6$, $L_4 = 5$ et $L_5 = 7$; les lignes finales ont quant à elles $L_6 = 2$ et $L_7 = 10$. Par exemple, la ligne 4, « $\overline{x}_h \leftarrow 2/(1/u' + 1/u'')$ », comporte deux divisions et une addition dans la parenthèse, une division de «2» par le résultat de la parenthèse, puis l'affectation du quotient à la variable x, d'où un coût de 5 unités. De plus, les lignes L_2 à L_5 sont effectuées T fois, la ligne L_2 étant effectuée une fois de plus, au moment où l'inégalité « $n <$ T » est invalidée. Le calcul de coût s'exprime donc comme suit:

$$t = L_1 + T \times [L_2, L_3, L_4, L_5] + L_2 + [L_6, L_7]$$

$$= 16 + 19T .$$

La première équation identifie les éléments de coût en même temps que la structure en phases de l'algorithme, la seconde donne la formule symbolique de coût. Le coût de l'algorithme est ici d'à peu près 19(T+l) : qui dit mieux?

9.9 *L'optimisation Monte Carlo*. Pour estimer une quantité Q dans un contexte donné, on peut généralement mettre au point plus d'un procédé d'estimation stochastique, disons les procédés P_1, P_2, etc. Soit les estimateurs correspondants $\hat{Q}(P_1)$, $\hat{Q}(P_2)$, etc.: chaque procédé est caractérisé par une variance (unitaire) σ_1^2 et une durée totale (t) ou unitaire (t_1) de production t. *Le procédé sera d'autant plus efficace que le produit* $t_1 \cdot \sigma_1^2$ *sera petit.* L'optimisation Monte Carlo désigne l'ensemble des interventions faites dans ce but.

Pour accroître l'efficacité d'un procédé d'estimation Monte Carlo, les équations d'efficacité (9.8) et (9.9) suggèrent de réduire la variance unitaire O^2_1, de réduire la durée unitaire t_1, enfin de réduire le nombre d'échantillons T. Bien entendu, l'utilisation d'un nombre réduit d'échantillons entraînera une perte de précision, à moins que cette réduction du nombre soit compensée par une réduction équivalente de la variance d'estimation. Il en va de même pour les autres voies d'optimisation possibles. Or, nonobstant sa simplicité globale, le contexte réel d'un procédé Monte Carlo est souvent riche et fertile en possibilités statistiques, et il peut donner prise à des améliorations significatives en vue d'une plus grande précision des estimations ou à un calcul accéléré. Ces améliorations composent le menu du prochain chapitre.

Exercices

9.1 Soit $x \sim D(\mu,O^2)$, une v.a. issue d'une distribution quelconque possédant l'espérance $E(x) = \mu$ et la variance $var(x) = O^2$. Montrer que la moyenne de T v.a. mutuellement indépendantes issues de cette distribution a pour espérance μ et pour variance O^2/T.

9.2 Trouver l'équation reliant Erreur quadratique moyenne (9.5), Biais (9.6) et Variance (9.4).

9.3 *Loi Khi,* x_1 Une v.a. normale de loi $N(0,1)$, une fois transformée en valeur absolue, a une distribution Khi à v=1 degré de liberté, dénotée x_1 [J. K. Patel, C. H. Kapadia et D. B. Owen, « Handbook of statistical distributions », New York, Marcel Dekker 1976, caractérisent la variable x_v, la racine carrée d'une x^2_v]. Le schéma de programme suivant estime l'espérance de cette variable.

Estimer l'espérance $Q = E(y)$, $y = |x|$, $x \sim N(0,1)$,
par la moyenne de T échantillons Monte Carlo

[Initialisation] $n \leftarrow 0$; $\sum \leftarrow 0$
[Cycle] Tant que $n < T$ faire
Produire u_1 et u_2
$x \leftarrow \sqrt{(-2\log_e u_1)} \times \sin(2\pi u_2)$
$y \leftarrow abs(x)$
$\sum \leftarrow \sum + y$
$n \leftarrow n + 1$
[Calculs finals] $\hat{Q} = \sum / T$.

L'obtention d'une v.a. normale $N(0,1)$ depuis deux v.a. uniformes exploite la transformation Box-Muller (4.15).

Faire l'analyse du coût temporel de l'algorithme ci-haut. Sachant que l'espérance et la variance du x_1 sont $\sqrt{(2/л)}$ et $1-2/л$, fabriquer un indice d'efficacité, de type (9.8) ou (9.9), pour apprécier l'algorithme.

9.4 (Suite du précédent). La transformation Box-Müller utilisée dans l'algorithme ci-haut se base sur deux v.a. uniformes pour produire deux v.a. normales : l'usage ici fait d'une seule v.a. normale est

gaspilleur. Un changement de cet ordre ainsi que d'autres améliorations plus pointilleuses donnent lieu à la version suivante :

Estimer l'espérance Q = E(y), y =x ~ N(0,1),
par la moyenne de T échantillonsMonte Carlo - Version accélérée

[Initialisation] $n \leftarrow 0$; $\Sigma \leftarrow 0$
$K \leftarrow 2\pi$

[Cycle] Tant que $n < T$ faire
Produire u_1 et u_2
$V \leftarrow \sqrt{(-2\log_e u_1)}$
$x_1 \leftarrow V\sin(\pi u_2)$; $x_2 \leftarrow V\cos(Ku_2)$
$\Sigma \leftarrow \Sigma + x_1 + \text{abs}(x_2)$
$n \leftarrow n + 2$

[Calculs finals] $\hat{Q} = \Sigma/T$.

L'exécution correcte suppose que T est un nombre pair. Noter que la mise en valeur absolue de x_1, tel que calculé, n'est pas nécessaire étant donné que la fonction antisymétrique sin θ est positive pour $0 \le \theta \le \pi$.

Faire l'analyse du coût temporel et comparer l'équation de coût à celle obtenue à l'exercice précédent. Calculer le mérite relatif de la présente version en appliquant (9.10).

9.5 Faire l'analyse du coût temporel de la section de programme, « Insertion avec sentinelle », en appendice du chapitre 6. Que remarque-t-on dans l'équation de coût?

9.6 Par sondage de fonction, estimer l'espérance de $y = \sqrt{u}$, la racine carrée d'une v.a. uniforme de loi $U(0,1)$. [Noter que l'espérance et la variance de y sont $\frac{2}{3}$ et $\frac{1}{18}$]. Comparer la précision relative de cette technique (utilisant T coups de sonde) à celle d'une simple moyenne basée sur T v.a. y.

9.7 *Variance de la médiane normale.* La médiane (Md) de n v.a. normales, de distribution $N(\mu,O^2)$, a pour espérance et pour variance K_nO^2, n étant impair. F. N. David et N. L. Johnson [« Statistical treatment of censored data. Part I: Fundamental formule », *Biometrika,* vol. 41, p. 228—240] fournissent une approximation pour K_n, soit :

$$K_n \approx 1{,}5708/(n+2) + 2{,}4674/(n+2)^2 + 3{,}4627/(n+2)^3 \; . \qquad (9.11)$$

Soit K*n*, la variance (approximative) d'une médiane normale Md basée sur *n* données de loi *N*(0,1). Est-ce que la variance de la médiane *groupée* (Md_G), *i.e.* une médiane basée sur un regroupement des *n* valeurs en classes d'intervalle commun *I,* est plus petite, comparable ou plus grande [voir G. Châtillon, R. Gélinas, L. Martin et L. Laurencelle: « When is it preferable to estimate population percentiles from a set of classes rather than from the raw data? », *Journal of Educational Statistics,* 1987, vol. 12, p. 395-409] *?*

Concevoir une étude Monte Carlo destinée à comparer var(Md) et var(Md_G) selon diverses valeurs de *n (n = 3, 5,* 11, 21, 51, 101) et selon différentes formes de regroupement, un regroupement basé: soit sur des bornes de classe fixes, *i.e.* (0, I], (1,21], (21,31] etc.; soit sur des intervalles à bornes flottantes, *i.e.* (ε,ε+I], (ε+I,ε+2I] etc.; (C) soit sur les *k* classes réparties dans l'étendue des valeurs observées et marquant un intervalle ajusté, par exemple I $(x_{(n)} - x_{(1)})/k$.

9.8 (Suite du précédent). Le calcul de Md_G, la médiane groupée, dépend essentiellement des *fréquences* d'observations réparties dans les classes et des bornes de classe, plutôt que des valeurs observées elles-mêmes. Concevoir un algorithme permettant l'étude Monte Carlo de la médiane groupée basé sur la loi trinomiale, l'échantillonnage concernant la série de fréquences f_{j-1}, f_j ,f_j +1, $\Sigma f = n$. Identifier les restrictions et précautions à prendre [voir aussi D. Allaire, L. Laurencelle et G. Châtillon: « La méthodologie des études sur la médiane groupée: l'approche Monte Carlo, l'approche multinomiale et l'approche trinomiale », *in* L. Laurencelle (dir.): *Trois essais de méthodologie quantitative*(p. 63-122), 1994, Sainte-Foy, P.U.Q.].

Comparer le coût formel de l'algorithme élaboré à l'algorithme élémentaire de l'exercice précédent, opérant dans un contexte comparable. Un algorithme a-t-il toujours l'avantage sur l'autre, quels que soient *n* et la loi de distribution concernée?

9.9 *Loi Gamma restreinte,* $Ga_R(k,\beta,L)$. Dans une loi Gamma (définie aux exercices 5.9 et 8.7), l'intervalle de réalisation des *k* événements (T_k), chacun de loi exponentielle $E(\beta^{-1})$, varie librement et il coïncide ainsi avec le temps total (T) de l'expérience aléatoire. Dans la loi restreinte, on exige que l'intervalle du premier jusqu'au le événement ne déborde pas une limite *L, i.e.* $T_k = \Sigma_k \, {}^{t}{}_{i} < L$; si cette limite est débordée par la suite d'événements (1, 2*, ..., k*), l'expérience continue en recrutant un nouvel événement, et le test se répète avec la suite

d'événements (2, 3, k+l), ainsi de suite (il s'agit ici de la version totale de la loi Gamma restreinte, alors qu'en version individuelle, chaque événement constitutif doit respecter la limite prescrite, *i.e.* $t_i < L$). La v.a. est bien entendu le temps T requis pour obtenir le succès; le nombre *(n)* d'événements requis pour y parvenir est une v.a. auxiliaire (voir L. Laurencelle, « Les lois Gamma restreintes », *Lettres statistiques,* 1998, vol. 10, p. 67-84).

Utilisant $\beta = 1$ en forme standard, mettre au point un programme d'estimation Monte Carlo afin d'évaluer les caractéristiques d'une v.a. $Ga_R(k,l,L)$, notamment l'espérance, la variance et les centiles 50, 90, 95 et 99, pour diverses valeurs de *k* et *L.* [Suggestion : utiliser un vecteur cyclique de *k* positions, V[0.. k—1], constitué de v.a. exponentielles standard, en faisant correspondre la position *j* du vecteur et le numéro d'événement *n* par : $j \leftarrow n$ MOD *k,* et en renouvelant à chaque fois la valeur de cette j^e position.]

9.10 *Loi Gamma restreinte individuelle,* $Ga_{Ri}(k,\beta,L)$ (Suite du précédent). Pour obtenir le succès, on exige de produire une suite d'événements exponentiels, de loi $E(\beta^{-1})$, chacun d'intervalle borné par *L,* soit $t_i < L$: la v.a. de cette loi est le temps total (T) requis pour y parvenir. Quant à la variable auxiliaire *n,* le nombre d'événements requis, elle obéit à la loi de Pascal restreinte, dite aussi « loi des succès consécutifs » (voir L. Laurencelle : « La loi des succès consécutifs dans un processus de Bernoulli », *Lettres Statistiques,* 1987, vol. 8, p. 25 -47), de paramètres *P* et *k, où* $P = 1 - e^{-L\beta}$. L'espérance et la variance de cette loi sont:

$$E(n)=\frac{1-P^k}{(1-P)P^k} \quad ; \quad \mathrm{var}(n)=\frac{1-P^{2k+1}-(2k+1)(1-P)P^k}{\left[(1-P)P^k\right]^2} \quad . \quad (9.12)$$

Reprendre l'exercice précédent pour réaliser des estimations Monte Carlo basées sur la loi $G_{Ri}(k,1,L)$. Montrer que $\mathrm{moy}(T) \rightarrow \beta E(n)$ et $\mathrm{var}(T) \approx \beta^2 [k + \mathrm{var}(n)]$.

9.11 Soit un groupe composé de n individus, dans lequel chacun doit indiquer par vote le candidat de son choix (autre que lui-même). Soit c_i , le nombre de votes que reçoit l'individu *i,* $0 \le c_i \le n-1$, $\Sigma c_i = n$ et $c_{max} = \max(c_i)$: la quantité c_{max}, est une v.a. relative au « choix exceptionnel » (voir L. Laurencelle : « L'interprétation stochastique du sociogramme et le problème des choix exceptionnels », *Lettres Statistiques,* 1993, vol. 9, p. 115 -133).

Stipulant un modèle de choix équiprobable entre les n individus, déterminer par un calcul Monte Carlo à partir de quelles valeurs de c_{max} le vote peut être jugé significatif d'une préférence réelle dans le groupe. Estimer en même temps l'espérance, la variance et la distribution de probabilités de c_{max}.

Références

BRATLEY, P., Fox, B.L., SCHRAGE, L.E. (1987). *A guide to simulation.* New York, Springer-Verlag.

DYM, C.L., IVEY, E.S. (1980). *Principles of mathematical modeling.* New York, Academic Press.

KNUTH, D.E. (1969). *The art of computer programming.* Vol. 2: *Seminumerical algorithms.* Reading (MA), Addison-Wesley.

LAURENCELLE, L. (1993). La loi uniforme: propriétés et applications. *Lettres statistiques, 9,* 1 -23.

MODER, J.J., ELMAGHRABY, S.E. (dir.) (1978). *Handbook of operations research.* New York, Van Nostrand.

NAYLOR, T.H. (dir.) (1971). *Computer simulation experiments with models of economic systems.* New York, Wiley.

SWOKOWSKI (1993). *Analyse* (Trad. M. Cette, 5e édition). Bruxelles, De Bceck Université.

Chapitre
10 Techniques d'optimisation de l'intégration Monte Carlo

10.1 L'exécution d'un programme informatique exploitant des nombres aléatoires afin d'estimer la grandeur, la force, la longueur, les dimensions d'un phénomène quantitatif modélisé, — l'intégration Monte Carlo, en un mot —, implique à la fois un coût : le nombre et la durée des opérations mises en oeuvre, et un bénéfice : la valeur, ou précision, de l'estimation de grandeur obtenue. Reprenant les conventions du chapitre précédent, posons une quantité Q à estimer, et l'estimateur Q résultant d'un calcul basé sur T itérations, ou répétitions de mesure indépendantes, selon :

$$\hat{Q} = \tfrac{1}{T}[Q_1 + Q_2 + \ldots + Q_T] \; . \qquad (10.1)$$

La variance d'erreur de notre valeur estimée, $\sigma_\varepsilon^2(\hat{Q})$, peut être approchée par:

$$\hat{\sigma}_\varepsilon^2(\hat{Q}) = s_1^2 / T = \textstyle\sum_i (Q_i - \hat{Q})^2 / [T(T-1)] \qquad (10.2a)$$

où:

$$s_1^2 = \textstyle\sum_i (Q_i - \hat{Q})^2 / (T-1) \qquad (10.2b)$$

est la variance (dite unitaire) des T estimations individuelles $Q_1, Q_2, \cdots, Q_T$, notée aussi $\hat{\sigma}_1^2$. Ainsi, la précision, soit l'inverse de la variance d'erreur, est directement proportionnelle au nombre T d'estimations effectuées. D'autre part, dénotant par $t(\hat{Q},T)$ la durée totale d'exécution du programme basé sur T itérations et par t_1 la durée d'une itération simple, l'efficacité d'un programme d'estimation peut se mesurer approximativement par :

$$IE(\hat{Q}) \approx T \; / \; [s_1^2 \cdot t(\hat{Q},T)] \approx 1 \; / \; (s_1^2 \cdot t_1) \; , \qquad (10.3)$$

ces formules reproduisant les indices d'efficacité (9.8) et (9.9) déjà présentés.

L'optimisation Monte Carlo consiste à rendre le programme d'estimation en cours plus efficace en accroissant la valeur d'un indice tel que (10.3). La simplicité superficielle de cet indice peut décourager d'abord tout effort d'analyse et de remaniement du problème. Or, le système d'estimation Monte Carlo, représenté dans un programme informatique, est

relativement complexe : il comporte un appareil informatique de support, des fonctions mathématiques et statistiques, des séries de nombres aléatoires, etc. Cette complexité même facilite d'une certaine façon l'analyse et le remaniement – puisque plusieurs ingrédients doivent être agencés, souvent de manière arbitraire –, de sorte que ces manœuvres sont, presque à chaque fois, utiles et productives d'amélioration. En vue d'accroître l'efficacité d'un programme d'estimation Monte Carlo et en gardant les autres conditions égales, nous verrons d'abord comment rendre le programme plus performant en réduisant la durée d'exécution $t(\hat{Q})$, puis comment rendre la technique d'estimation plus précise en réduisant la variance unitaire σ_1^2 (ou, équivalemment, s_1^2).

Dans ce chapitre, nous passerons en revue plusieurs techniques et façons de faire afin d'optimiser d'une manière ou d'une autre le procédé d'estimation Monte Carlo. Nous le ferons *in abstracto,* en énonçant le principe ou la technique d'optimisation, et nous en fournirons, à l'occasion, un petit exemple. Des exemples plus réalistes, l'un d'eux comparant différentes techniques d'optimisation et leurs efficacités, sont présentés au chapitre suivant.

Réduction de la durée d'estimation, $t(\hat{Q})$

10.2 La préparation et l'exécution d'une étude Monte Carlo impliquent ordinairement —un modèle mathématique, —un algorithme et —un support informatique. Considérons d'abord le modèle mathématique tel qu'il apparaît à l'orée de l'étude. Ce modèle représente, voire constitue le système dans lequel l'évaluation a lieu. L'analyse détaillée du modèle mathématique lui-même peut donner lieu à sa simplification ou à sa reformulation en un modèle équivalent plus efficace : interviennent ici la culture mathématique et la créativité du chercheur. Une technique d'intérêt consiste à décomposer le modèle à l'étude en un agrégat de modèles partiels, disons $M = \{ M_1, M_2, , ..., M_k \}$, et à définir la solution de M par l'agrégat des solutions partielles. Cette technique de décomposition est spécifiquement avantageuse lorsque la solution d'une ou de plusieurs parties du modèle global est immédiate (i.e. connue) ou qu'elle peut être obtenue par un procédé déterministe. La portion de l'évaluation qui incombe au procédé Monte Carlo est réduite d'autant, de sorte que la rapidité *et* la précision du calcul final, partiellement estimatif, s'en trouvent accrues.

10.3 Une fois le modèle établi, l'algorithme de l'estimation Monte Carlo doit être considéré. Par cette expression, nous désignons le schéma séquentiel des principales opérations du programme servant à estimer la quantité Q cherchée : ce schéma reste parfois conceptuel et, parfois, il prend

la forme d'un programme en pseudo-langage (tel qu'illustré dans les schémas de programmes apparaissant ailleurs dans ce livre), un jargon à moitié explicite et dont le but est de *faire comprendre* l'action détaillée du programme. L'algorithme lui-même peut être optimisé et, lorsqu'il l'est, tous les programmes qui en découlent bénéficient *ipso facto* de cette accélération. Répétons que nous nous intéressons ici à l'optimisation en temps d'exécution; nous nous tournerons plus loin vers l'optimisation en précision.

À l'instar de maints schémas de programmes informatiques, le schéma général d'un programme d'évaluation Monte Carlo comporte trois phases : une phase d'initialisation, ou de préparation des calculs, la phase itérative ou de calcul proprement dit, et la phase de conclusion. Comme nous l'avons souligné plus haut (§9.6), c'est la phase itérative qui pèse le plus lourdement sur le temps total d'exécution puisque chaque itération, chaque cycle de calcul de la phase itérative, est multiplié par le nombre d'estimations requises. C'est donc à l'allégement de cette partie de l'algorithme de calcul que doit viser l'effort d'optimisation. Ainsi, certaines parties constantes, qui interviennent répétitivement dans le cycle itératif du programme, peuvent être anticipées et déplacées vers la phase d'initialisation, où elles sont calculées une fois pour toutes. D'autres manoeuvres extraordinaires d'accélération méritent d'être tentées pour les opérations du cycle itératif: même si ces manoeuvres semblent de peu de portée lorsque considérées isolément, le facteur multiplicatif qui les affecte en justifie souvent l'effort.

10.4 Enfin, le support informatique exploité pour opérer l'algorithme de calcul en détermine essentiellement l'efficacité temporelle. Une machine informatique (calculateur, microprocesseur, etc.) plus rapide est évidemment plus efficace. De plus, en raison de leurs modes d'exécution interne, certains langages ou dialectes de langage informatiques sont plus rapides que d'autres. Cela étant dit et dussions-nous disposer d'une machine informatique super-efficace, il reste important et intéressant de nous soucier d'optimisation car, dans ce cas, l'horizon des études pratiquement réalisables aura été repoussé et nous devrons affronter encore de nouvelles limitations.

Cependant, comme c'était le cas pour l'algorithme ou schéma de calcul, c'est encore sur la phase itérative du programme informatique (on dit aussi « code de programme ») que l'examen doit porter. Par exemple, une variable aléatoire, d'une distribution donnée, peut généralement être produite par plus d'une méthode : laquelle, dans les conditions spécifiques du programme, s'avère-t-elle la plus performante? Autre exemple : dans certains langages, toute élévation à une puissance (e.g. x^k) s'effectue par une transformation logarithmique (e.g. $\exp[k \bullet \log_e x]$), une opération lourde et

arithmétiquement peu précise, qui serait contre-indiquée pour un petit exposant entier (e.g. remplacer x^3 par $x \bullet x \bullet x$). Dernier exemple : si le produit ou la somme de deux quantités ou plus apparaît deux fois dans la boucle itérative, il peut être avantageux d'effectuer à part et mémoriser l'opération, en ré-injectant son résultat aux places appropriées. On désigne sous le nom de *dégraissage de code* ces manoeuvres de simplification et de réduction des opérations d'un programme. L'optimisation des programmes informatiques est un domaine truffé d'idées et de moyens raffinés : nous référons le lecteur intéressé à la documentation pertinente (Abrash 1996 ; Knuth 1968, 1973 ; Shaffer 1997 ; Weiss 1992).

Réduction de la variance d'estimation, s_1^2

10.5 L'estimation Monte Carlo, même si elle procède par des calculs programmés sur ordinateur, est en elle-même un procédé statistique. En fait, le travail consiste presque toujours à estimer une moyenne, une intégrale (ou une somme), une espérance, ou bien il peut être reformulé dans ce contexte. L'estimateur global $\hat{Q}$, défini par (10.1), est lui-même une moyenne d'estimations individuelles Q_i, et ses propriétés globales de biais et de variance sont solidaires d'un ensemble de théorèmes statistiques qu'on peut mettre à profit. Les quelques techniques exposées dans les paragraphes suivants en sont des exemples.

La documentation livresque disponible sur les techniques de réduction de la variance Monte Carlo est abondante (Bratley, Fox et Schrage 1987 ; Evans et Swartz 2000; Fishman 1996; Gentle 1998; Hammersley et Handscomb 1964; Kalos et Whitlock 1986; Kleijnen 1974; Rubinstein 1981; etc.) et elle est l'écho d'une non moins abondante documentation, parfois ancienne, dans les périodiques. L'un des tout premiers articles portant sur la question, celui de Kahn et Marshall, a paru en 1953.

Quoi qu'il en soit de l'abondance et de l'ancienneté des publications sur le sujet, le traitement qui est fait des différentes techniques manque souvent d'unité; parfois, ce que deux auteurs présentent sous deux noms différents se ramène à la même technique, au même théorème statistique. Le fait est que les chercheurs qui se sont intéressés à la question n'ont pas ré-inventé la statistique. Ils ont plutôt tenté d'introduire dans le procédé d'estimation Monte Carlo l'un ou l'autre des théorèmes statistiques disponibles. Aussi, vu le foisonnement actuel et la diversité des études Monte Carlo, il est à prévoir que d'autres théorèmes statistiques, des techniques à dénomination nouvelle, verront aussi le jour.

Enfin, dans notre itinéraire parmi les techniques de réduction de variance, nous ignorons généralement le biais d'estimation (éq. (9.6)), habituellement nul. En fait, la présence d'un biais non nul dans l'estimateur $\hat{Q}$ peut ressortir à trois facteurs, agissant seuls ou de concert: 1) le mécanisme d'échantillonnage des estimations individuelles (Q_i) est gauchi ou fautif; 2) la structure de l'estimateur global $\hat{Q}$ est incorrecte; 3) la valeur à estimer Q, ou estimande, ne possède pas d'estimateur statistique sans biais. Le cas (3), quand même peu fréquent, n'est pas sans exemples (*e.g.* le coefficient de corrélation r, éq. (5.4)). Rubinstein (1981) discute le cas (2), en le dédramatisant. La confection soignée d'échantillons de variables aléatoires (cf. chap. 3 à 5) devrait nous garantir contre le cas (1). Toujours est-il qu'il faut rester vigilant sur la question du biais et de ses sources possibles, comme nous tenterons de le faire dans la suite de l'ouvrage.

Techniques aveugles, par manipulation de la variable

10.6 *La technique de l'antivariable.* Soit deux v.a. x et x', chacune ayant pour espérance la même valeur, $E(x) = E(x') = Q$, et pour variance, $\text{var}(x) = \text{var}(x') = \sigma_x^2$. Leur moyenne,

$$\hat{x} = \tfrac{1}{2}(x + x') , \tag{10.4}$$

a la même espérance Q, et sa variance est donnée par:

$$\begin{aligned} \text{var}(\hat{x}) &= \text{var}\,[\ \tfrac{1}{2}(x + x')\ ] \\ &= \tfrac{1}{4}[\sigma_x^2 + \sigma_{x'}^2 + 2\sigma_x\sigma_{x'}\rho(x,x')] \\ &= \tfrac{1}{2}\sigma_x^2[1 + \rho(x,x')] . \end{aligned} \tag{10.5}$$

Lorsque la corrélation entre x et x' est négative, alors x' représente une antivariable de x. Dans ce cas, nous observons que $2\times\text{var}(\hat{x}) < \sigma_x^2$ et, pour cette raison, l'exploitation de cette méthode dans (10.4) réduit l'imprécision de $\hat{x}$ avantageusement.

Pour certains estimateurs $\hat{Q}$, l'échantillonnage Monte Carlo est une fonction monotone assez simple d'une seule variable aléatoire, par exemple une v.a. uniforme u de loi $U(0, 1)$. Pour de tels estimateurs, on peut écrire:

$$\begin{aligned} x &= f(\,u\,) \\ x' &= f(\,1-u\,) . \end{aligned}$$

Comme, d'une part, les v.a. u et $(1-u)$ ont la même distribution requise tout en ayant une corrélation mutuelle de -1 et, d'autre part, l'estimateur

est une fonction monotone de u, alors x' sera négativement corrélée avec x, permettant l'application de la technique de l'antivariable. La plupart des situations et des estimateurs, cependant, ne présentent pas un lien de conséquence aussi immédiat par rapport *à une seule v.a.,* uniforme ou non, et l'ingéniosité du chercheur doit être mise à contribution. Notons enfin que, parce qu'elle revient globalement à contrôler l'espace de variation de notre estimateur, la technique de l'antivariable peut être vue comme un cas particulier de la technique de l'échantillonnage stratifié, qu'on examine plus bas.

Exemple 10.1 Estimation de $E\{u^2\}$ par antivariable

Soit la quantité à estimer $Q = E\{u^2\}$, $u \sim U(0,1)$. Dans ce cas très simple, on sait que $Q = E\{u^2\} = \int_0^1 u^2\, du = ⅓$ et $\text{var}\{u^2\} = 4/45$. Le procédé d'estimation brut, qui consiste à obtenir une valeur aléatoire u, la mettre au carré et l'additionner pour en obtenir une moyenne, a pour variance unitaire $\sigma_1^2 = 4/45$ et, pour durée unitaire, $t_1 \approx 3+1+1 = 5$, d'après les conventions du chapitre précédent, à la section §9.8. La technique de l'antivariable s'applique bien ici, puisque l'estimateur Q est une fonction monotone de u: nous utilisons donc l'antivariable $(1-u)^2$ et l'estimateur $\hat{x} = ½[\, u^2 + (1-u)^2 \,]$. L'algorithme d'estimation suit.

Estimer $E(u^2)$ *par la technique de l'antivariable*

[Initialisation] $\Sigma \leftarrow 0$;
[Cycle] Exécuter ½T fois
$u \leftarrow$ RND ;
$\Sigma \leftarrow \Sigma + u^2 + (1-u)^2$
[Calculs finals] $\hat{Q} \leftarrow \Sigma / T$.

Le nombre de cycles itératifs exécutés est divisé par deux, chaque cycle s'appuyant sur deux quantités, u et 1—u, ce afin de garder les algorithmes comparables d'une technique à l'autre. Remarquer que la division par deux de l'expression « $u^2+(1—u)^2$ », semblable à chaque cycle, est différée implicitement jusqu'au calcul final.

L'analyse de l'algorithme ci-dessus procède ainsi. Chaque cycle utilise le générateur $U(0,1)$ une seule fois, ce pour un usage double. Le coût des opérations du cycle est, évidemment, de 3 pour la première ligne et de 4 pour la seconde, d'où $2\times t_1 \approx 8$ et $t_1 \approx 4$. Quant à la variance unitaire, nous avons $2\sigma_1^2 = \text{var}(\hat{x}) = ¼[\sigma^2\{u^2\} + \sigma^2\{(1-u)^2\} + 2\rho\{u^2,(1-u)^2\} \times \sigma\{u^2\}\sigma\{(1-u)^2\}$. Or, puisque u et

$1-u$ ont même distribution, il en va de même pour u^2 et $(1-u)^2$, avec variance $\sigma^2\{u^2\} = 4/45$. Quant à $\rho\{u^2,(1-u)^2\} = \text{cov}\{u^2,(1-u)^2\} / \sigma\{u^2\}$, on obtient facilement $\text{cov}\{u^2,(1-u)^2\} = 1/30$ et $\rho\{u^2,(1-u)^2\} = -0{,}875$. Nous calculons donc $\text{var}(\hat{x}) = \frac{1}{4}[2\times 4/45 + 2\times(-0{,}875)\times 4/45] = 2/45[1 - 0{,}875]$ et $\sigma_1^2 = 1/45[0{,}125] = 1/45\times 1/8 = 1/360$, soit une variance unitaire 8 fois plus petite que par l'estimation brute. L'efficacité du procédé par antivariable, d'après (10.3), serait donc $(1/360\times 4)^{-1} \approx 90$, comparativement à $(4/45\times 5)^{-1} \approx 2{,}25$ pour le procédé brut.

Il est à noter que, lorsqu'elle s'applique, la technique de l'antivariable ne requiert aucune sorte d'analyse de coût préalable, si ce n'est la présomption d'une corrélation négative entre la variable originale (x) et l'antivariable (x'). On peut donc la qualifier de technique aveugle et ayant une portée universelle.

10.7 *La technique de la covariable à espérance connue.* Soit une variable x dont on veut déterminer l'espérance, $E(x) = Q$, et une autre variable positivement corrélée avec x, la *covariable* y, d'espérance connue μ_Y. Alors la fonction:

$$\hat{x} = x - b\,(y - \mu_Y) \quad \{\, b > 0 \,\} \tag{10.6}$$

constitue un estimateur efficace de Q. En effet, $\hat{x}$ estime Q sans biais puisque:

$$\begin{aligned} E\{\,\hat{x}\,\} &= E\{\,x \;\; - b\,(y - \mu_Y)\,\} \\ &= E\{\,x\,\} - b\,E\{\,y-\mu_Y\,\} \\ &= \quad Q \quad - b\times 0 \;\; = Q\,. \end{aligned} \tag{10.7}$$

La précision de $\hat{x}$ est mesurée inversement par sa variance, soit:

$$\begin{aligned} \text{var}\{\,\hat{x}\,\} &= \text{var}\{\,x - b\,(y - \mu_Y)\,\} \\ &= \text{var}\{\,x\,\} + b^2\cdot\text{var}\{y-\mu_Y\} - 2b\cdot\text{cov}\{\,x, y-\mu_Y\,\} \\ &= \sigma_X^2 + b^2\sigma_Y^2 - 2b\sigma_X\sigma_Y\rho_{X,Y}\,, \end{aligned} \tag{10.8}$$

où σ_X^2 et σ_Y^2 dénotent les variances des variables x et y, et $\rho_{X,Y}$, leur corrélation. Si cette corrélation $\rho_{X,Y}$ entre variable et covariable est positive, on peut espérer réduire la variance d'estimation, laquelle dépend alors du choix du facteur b. La valeur optimale de b, donnée par $d\,\text{var}\{\,\hat{x}\,\} / db = 0$, serait égale à:

$$\hat{b} = \rho_{X,Y}\sigma_X/\sigma_Y\,. \tag{10.9}$$

Toutefois, cette détermination de la valeur optimale de b n'est pas aveugle, puisqu'elle utilise O_X et $P_{X,\,y}$, de valeurs habituellement inconnues. En principe, comme le montre (10.9), toute valeur positive de b, pas trop forte, comme $b \leq O_X / O_Y$ *ou* $b \approx 1$ (si les variables X et Y semblent être d'amplitudes comparables), peut faire l'affaire.

Notons enfin que l'estimateur Monte Carlo $\hat{Q} = \text{moy}\{\,\hat{x}\,\}$, basé sur T itérations, peut être ramené à une forme simple, de calcul parcimonieux, en considérant:

$$\begin{aligned} \hat{Q} &= \text{moy}\{\,\hat{x}\,\} && (10.10)\\ &= \text{moy}\{\,x - b\,(y - \mu_Y)\,\} \\ &= \bar{x} - b \times (\,\bar{y} - \mu_Y\,)\,; && (10.10a) \end{aligned}$$

il suffit donc de faire itérativement le calcul de $\bar{x}$ et $\bar{y}$, plutôt que de calculer (10.6) à chaque tour, économisant ainsi plusieurs opérations.

Exemple 10.2 L'espérance de la moyenne harmonique de deux v.a. uniformes par la technique de la covariable à espérance connue

Nous reprenons l'estimation de la moyenne harmonique $\bar{x}_h = 2/(1/u' + 1/u'')$, abordée dans l'exemple 9.1. Cette estimation, basée sur deux v.a. uniformes, admet clairement comme covariable leur moyenne arithmétique, $y = \frac{1}{2}[u' + u'']$, d'espérance $\mu_Y = \frac{1}{2}$. Utilisant l'estimateur $\hat{x} = \bar{x}_h - (y - \mu_Y)$, avec le facteur $b = 1$ implicite, nous pouvons programmer l'estimation de $Q = E\{\,\bar{x}_h\,\}$ comme suit:

Estimer l'espérance $E[\bar{x}_h]$ par la technique de la covariable à espérance connue

[Initialisation] $\Sigma_1 \leftarrow 0$; $\Sigma_2 \leftarrow 0$
[Cycle] Exécuter T fois
Obtenir deux v.a. u' et u''
$x \leftarrow 2/(\,1/u' + 1/u''\,) - (\,u' + u''\,)/2$
$\Sigma_1 \leftarrow \Sigma_1 + x$; $\Sigma_2 \leftarrow \Sigma_2 + x^2$
[Calculs finals] $\hat{Q} = \Sigma_1 / T + \frac{1}{2}$
$V_\varepsilon(\hat{Q}) = (\Sigma_2 - \Sigma_1^2/T)/[T(T-1)]$.

Dans cet algorithme, nous avons sacrifié l'économie de calcul suggérée par l'expression (10.10a), ce dans le but d'obtenir des valeurs individuelles de l'estimateur $\hat{x}$ et d'en calculer la variance.

Illustrons, encore ici, l'exécution de l'algorithme. Appliquant T = 10 000, nous obtenons $\hat{Q} = 0{,}40951$ et $\hat{\sigma}_\varepsilon = \sqrt{V_\varepsilon} = 0{,}00104$ une première fois, puis $\hat{Q} = 0{,}40869$ et encore $\hat{\sigma}_\varepsilon = 0{,}00104$ la seconde fois. Rappelons que l'erreur-type de l'estimateur brut (cf. exemple 9.1) était égale à $\sigma_\varepsilon = \sqrt{(0{,}0600/T)} \approx 0{,}00245$; ainsi, le présent estimateur apparaît environ 2,4 fois plus précis.

10.8 *La technique de l'échantillonnage stratifié.* La précision d'une estimation dépend du nombre d'éléments échantillonnés au hasard qu'elle utilise : la variance de la moyenne obtenue est inversément proportionnelle à ce nombre, comme l'indique l'expression (10.2). Nonobstant cette loi, les traités sur l'échantillonnage (p. ex. Cochran 1963 ; Kish 1967) indiquent un moyen efficace d'augmenter cette précision sans accroître en même temps le nombre d'éléments : c'est l'échantillonnage stratifié. La technique consiste d'abord à définir le domaine à échantillonner, à le segmenter en k parties, puis à échantillonner des éléments dans chaque partie, en combinant les parties au prorata de leur importance dans la population.

Dans une étude Monte Carlo comme dans les sondages et enquêtes démographiques, l'unité échantillonnale (disons X_i) ne coïncide que rarement avec la quantité à évaluer, ou variable cible (disons Y_i). La population est définie comme une grande collection peut-être infinie de X_i, la mesure est une fonction d'information, comme $Y_i = f(X_i)$: on désigne aussi X sous le nom de variable de stratification ou, parfois, « covariable ». La relation statistique entre Y et X est quantifiée par le coefficient de corrélation linéaire «$P_{x,\ y}$, si cette relation est monotone et linéaire. Pour une relation de dépendance générale, le rapport de corrélation $\eta_{y,\ x}$ (« êta », voir Martin et Baillargeon 1989) peut servir de mesure. En fait, considérant les espérances conditionnelles (Mood, Graybill et Boes 1974, p. 159) :

$$\mathrm{var}(Y) = E_X\{\ \mathrm{var}(Y|X)\ \} + \mathrm{var}_X\{\ E(Y|X)\ \}\ , \qquad (10.11)$$

nous pouvons définir :

$$\eta^2_{Y,X} = \frac{\mathrm{var}_X\{\ E(Y|X)\ \}}{\mathrm{var}(Y)} = 1 - \frac{E_X\{\ \mathrm{var}(Y|X)\ \}}{\mathrm{var}(Y)}\ , \qquad (10.12)$$

en rappelant aussi l'inégalité $\rho^2_{X,Y} \le \eta^2_{Y,X}$.

Nous cherchons donc à estimer Q = E(Y), la moyenne de population pour Y, en stratifiant le domaine échantillonnal par la variable X et en obtenant une moyenne stratifiée de Y, soit $\bar{Y}_{st}$.

Nous divisons d'abord le domaine de X en *k* intervalles contigus; dans un sondage démographique, la population est découpée en *k* strates, en fonction d'une variable X choisie (e.g. l'âge, le lieu d'habitation, le niveau de revenu familial). La strate j contient une portion $p_j(X)$ de la population totale, et $\Sigma_{j=1}^{k}\ p_j = 1$; l'espérance $\mu_j(Y)$ et la variance $O^2_j\ (Y)$ de la variable cible dans la strate *j* sont, bien entendu, inconnues.

Il s'agit alors d'échantillonner les valeurs Y_i, par strate de X. À partir du nombre total (n) d'éléments convenu, le nombre (n_j) d'éléments à piger au hasard, dans chaque strate, peut être déterminé selon des objectifs et des stratégies divers. L'échantillon $\{ Y_1, Y_2, \cdots, Y_{n_j} \}$, dans la strate *j*, permet d'obtenir $\bar{Y}_j = (\sum Y_i)/n_j$, une estimation de la moyenne de strate $\mu_j(Y)$, laquelle doit être combinée à celles des autres strates pour fournir l'estimateur global.

L'estimateur par moyenne stratifiée global est alors:

$$\hat{Q} = \bar{Y}_{st} = \sum_{j=1}^{k} p_j(X)\ \bar{Y}_j \quad . \tag{10.13}$$

La variance de l'estimateur (10.13) n'a pas d'expression simple qui soit en même temps exacte et universelle: sa valeur dépend du lien existant entre X et Y, de la stratégie de stratification, du comportement de var(Y | X) selon X, etc. La formule suivante, surtout indicative, fait voir le jeu des principaux paramètres:[1,2]

$$\sigma_\varepsilon^2(\hat{Q}) \approx \frac{\sigma_Y^2}{n}\left[1 - \eta_{Y,X}^2 + \frac{\eta_{Y,X}^2}{k^2}\right]. \tag{10.14}$$

Les traités sur l'échantillonnage discutent ces questions. Les exercices 10.4 à 10.6 examinent une autre formulation du problème.

1. La formule donnée (10.14), inspirée de Cochran (1963) et Kish (1967), est qualitativement juste et constitue une première approximation quantitative pour la plupart des situations; elle n'est exacte qu'en de rares cas. Remarquer, entre autres, que la formule ne reconnaît aucun rôle au régime de stratification (qui détermine les portions p), ni à la stratégie d'allocation des échantillons (n_j) d'une strate à l'autre. De plus, 1 indice « universel» $v_{x,Y}$ ne peut convenir qu'approximativement à l' ensemble des cas.

La formule donnée, qui pivote sur le quantificateur η, convient non seulement à l'estimation d'une intégrale ou d'une espérance mathématique, tel que privilégié dans cet ouvrage, mais aussi aux sondages pour lesquels la variable de stratification X est souvent catégorielle (e.g. « langue parlée à la maison », type d'emploi, région géographique). La statistique $\eta_{x,y}$ coïncide à peu près, à la « corrélation intra-classe », s'exprime dans un même registre conceptuel que le coefficient de corrélation $P_{x\ y}$ et remplace, à notre avis, avantageusement les lourdes constructions algébriques de l'analyse de variance, habituellement mentionnées.

Soit $_jY$, la variable cible dans la strate j, avec son espérance μ_j et sa variance O^2_j. L'estimateur de moyenne stratifiée a alors pour variance

$$\mathrm{var}(\hat{Q}) = \Sigma_{j=1}^{k} [p_j \sigma_j^2] / n_j , \qquad (10.15)$$

où $\Sigma n_j = n$, variance qui est minimisée pour n_j lorsque :

$$n_j = n \times p_j \sigma_j / \Sigma p_j \sigma_j \qquad (10.16)$$

(Rubinstein 1981). Cependant, pour une valeur donnée du nombre de strates (k), la recette pour minimiser absolument var($\hat{Q}$) dépend aussi des quantités p_j et σ_j, donc du régime de segmentation du domaine d'échantillonnage et de la connaissance explicite de la distribution de Y. Ces exigences ainsi que les considérations raffinées nécessaires pour atteindre la variance minimale semblent être hors de mesure avec le contexte présent, celui d'une estimation Monte Carlo à l'aveugle, et nous n'y entrerons pas.

L'extrémum d'efficacité de l'échantillonnage stratifié se produit lorsque la variable de stratification est celle-là même dont on veut estimer l'espérance : dans ce cas, $\eta = \rho = 1$, et $\sigma^2_\varepsilon(\hat{Q}) \to \sigma^2_Y / nk^2$. La limite est atteinte si Y (=X) a une distribution uniforme et la stratification est faite par tranches d'importance égale, selon $p_j(X) = 1/k$.

Dalenius et Hodges (1959) et Bethel (1989) donnent des précisions quant à l'opérationnalisation et l'efficacité comparative de différentes stratégies de stratification.

Pour appliquer ce principe à l'estimation Monte Carlo et en retirer des bénéfices de précision, il faut donc, en premier lieu, fixer son choix sur une covariable (ou variable de stratification), stratifier le domaine d'estimation en fonction de cette covariable et construire enfin un estimateur $\bar{Y}_{st}$ (10.13) de la variable cible basé sur cette stratification.

Exemple 10.3 L 'espérance de la moyenne harmonique de deux v.a. uniformes par échantillonnage stratifié

Revenons au chantier de l'exemple 9.1, touchant l'estimation de l'espérance de la moyenne harmonique de deux uniformes, $\bar{x}_h = 2 / (1/u' + 1/u'')$, $u', u'' \sim U(0,1)$. En fait, on conçoit aisément que $\bar{x}_h$ covarie positivement avec [u', u''], en ce sens que, si [u', u''] augmentent ou diminuent conjointement, $\bar{x}_h$ fait de même. En considérant cette relation (non linéaire) conjointe, il s'annonce avantageux de stratifier à la fois sur u' et u''. C'est ce que propose l'algorithme suivant.

Estimer l'espérance $E[\bar{x}_h]$ par échantillonnage doublement stratifié

{ Le domaine ((0..1)×(0..1)) de u' et u'' est segmenté en $k' = k^2$ strates; la moyenne globale, accumulée dans les k^2 cellules, est réitérée $t = \lfloor T/k' \rfloor$ fois }

```
[Initialisation ]  Σ1 ← 0 ; Σ2 ← 0
[Cycle          ]  Exécuter t fois  { t = ⌊T/k'⌋ }
[Stratification ]      m_strat ← 0
                       Pour j1 = 0 à k−1 faire
                         Pour j2 = 0 à k−1 faire
                           Obtenir deux v.a. u' et u''
                           w' ← (j1 + u')/k ; w'' ← (j2 + u'')/k
                           x ← 2/(1/w' + 1/w'')
                           m_strat ← m_strat + x          (*)
                     m_strat ← m_strat / k'
                     Σ1 ← Σ1 + m_strat ; Σ2 ← Σ2 + m_strat²
[Calculs finals ]  Q̂ = Σ1 / t
                   Vε(Q̂) = (Σ2 − Σ1²/t)/[(t−1)T] .
```

La ligne marquée d'un astérisque complète le calcul d'un exemplaire de « m_strat », la moyenne stratifiée; tel que rédigé, l'algorithme obtient et moyenne T/k' exemplaires, de sorte que l'estimateur résultant $\hat{Q}$, basé sur T « morceaux » individuels, est commensurable à l'estimateur brut de l'exemple 9.1 et il lui est équivalent lorsque $k' = k = 1$.

Utilisant $k' = k^2 = 4$ strates (*i.e.* 2 strates pour chaque v.a. uniforme) et T = 10 000, nous avons obtenu $\hat{Q} = 0{,}40941$ et $\sqrt{V_\varepsilon} = 0{,}00737$ une première fois, puis 0,41041 et 0,00708 une seconde fois. En fait, une brève étude montre que, pour cet exemple:

$$V_\varepsilon(\hat{Q}) = \frac{\sigma_Y^2}{T} \times \frac{f(k')}{k'^2} \quad , \quad 1 \le f(k') < 2 \; ;$$

la précision obtenue par échantillonnage stratifié est ainsi de $k\sqrt{2}$ à k fois meilleure qu'avec l'échantillonnage brut, ce qui confirme notre choix opportun d'une double stratification pour cet exemple. L'exercice 10.7 examine l'efficacité globale de l'algorithme utilisé.

10.9 *Autres techniques aveugles de réduction de la variance.* On qualifie d'« aveugles » les précédentes techniques de réduction de variance parce qu'elles n'exigent pas la connaissance ni l'exploitation de la fonction de densité de la variable étudiée, disons $f(X)$: pour les mettre en oeuvre, il suffit de disposer d'une source ou d'un générateur qui produisent à volonté les valeurs requises. D'autres procédés aveugles sont évidemment possibles.

La *mise en commun* table sur le fait que deux ou plusieurs estimateurs basés sur le même lot de v.a. générées ont tendance à corréler positivement entre eux. Si, par exemple, nous voulons obtenir $Q = E(T_1 - T_2)$, où T_1 et T_2 sont deux fonctions des mêmes v.a. $\{ X_i \}$, alors, généralement:

$$\begin{aligned} \operatorname{var}(\hat{Q}) &= \operatorname{var}(T_1 - T_2) \\ &= \operatorname{var}(T_1) + \operatorname{var}(T_2) - 2\rho(T_1,T_2)\sigma(T_1)\sigma(T_2) \ , \end{aligned} \tag{10.17}$$

ce théorème représentant l'autre versant du théorème (10.5). Si, comme à l'accoutumée, T_1 et T_2 sont estimés indépendamment l'un de l'autre, alors $\rho(T_1,T_2) = 0$. Si, au contraire, T_1 et T_2 peuvent être calculés à chaque fois sur les mêmes v.a. $\{ X_i \}$, alors $\rho(T_1,T_2) > 0$, entraînant une réduction proportionnelle de la variance globale d'estimation.

L'*estimation indirecte* permet aussi, dans certains cas, d'obtenir une estimation $\hat{Q}'$ de la quantité Q en recourant à un estimateur Q′ dont l'espérance coïncide avec Q et dont la variance d'erreur est moindre. Citons deux exemples. Pour une v.a. à distribution symétrique autour de θ, la moyenne ($\bar{X}$), la médiane (Md) et le point-milieu (M) sont des estimateurs concurrents de θ, la moyenne étant plus précise, *i.e.* à variance moindre, pour la distribution normale, la médiane pour la distribution double-exponentielle de Laplace, le point-milieu pour la distribution uniforme (Johnson, Kotz et Balakrishnan 1994, 1995). Pour estimer θ, autant choisir l'estimateur le plus précis dans chaque contexte.

> La variance de la moyenne $\bar{X}$ est toujours égale à σ^2/n, où $\sigma^2 = \operatorname{var}(X)$. La variance de la médiane Md, proche de celle de la moyenne pour de petits n, tend asymptotiquement vers $\frac{1}{2}\pi^2\sigma^2/n \approx 1{,}571\sigma^2/n$ pour les v.a. normales, vers $3\sigma^2/n$ pour des v.a. uniformes et environ vers $0{,}6\sigma^2/n$ pour celles à distributions plus pointues comme la double-exponentielle de Laplace. Quant au point-milieu, il n'a d'intérêt (et de signification statistique) que dans un domaine doublement borné, comme celui des distributions *Bêta* symétriques : dans le cas particulier de la distribution uniforme, sa variance tend rapidement vers $\sigma^2/(6n^2)$.

L'autre exemple d'estimation indirecte concerne les statistiques d'ordre, en général, tels le maximum d'une série statistique, un quantile quelconque, la

médiane, etc. Étant donné que les statistiques d'ordre en X, soit { $X_{(1:n)}$, $X_{(2:n)}$, ..., $X_{(n:n)}$ }, équivalent aux statistiques d'ordre uniformes { $U_{(1)}$, $U_{(2)}$, ..., $U_{(n)}$ } par la transformation $F_X(X) \rightarrow U$, $0 \leq U \leq 1$, on peut estimer la valeur d'une statistique d'ordre sous la forme de distribution la plus avantageuse, notamment celle ayant la variance la plus petite. Par exemple, on pourra établir un intervalle de confiance de la médiane de 5 données normales directement alors que, pour estimer le maximum de 5 données normales, *i.e.* $Q = X_{(5:5)}$, il sera profitable d'obtenir d'abord $Q' = U_{(5:5)}$, puis $Q = \Phi^{-1}_{\mu,\sigma^2}(Q')$, où $\Phi(\ )$ est la fonction de répartition d'une loi normale $N(\mu,\sigma^2)$.

Une dernière approche que nous rangeons parmi les techniques aveugles est l'*estimation de densité* nonparamétrique (voir p. ex. Parzen 1979 et Izenman 1991). Le principe de cette approche, dans notre contexte, est d'abord d'accumuler des données et acquérir ainsi une fonction estimative de densité, disons $\hat{f}(X)$. Cette fonction palliative peut ensuite être exploitée, en lieu de la fonction réelle, en appliquant l'une ou l'autre des techniques dites fonctionnelles d'optimisation Monte Carlo.

Techniques fonctionnelles, ou par manipulation de la fonction

10.10 Il arrive que nous ayons en main la fonction à intégrer sans toutefois être à même de l'évaluer exactement, par calcul intégral ou en appliquant une méthode déterministe (cf. chapitre 8). Ou bien l'expression mathématique de la fonction est complexe et ne se prête pas à la réduction à une forme élémentaire, ou bien la dimension ou l'anatomie de la fonction rendent sa quadrature impossible. Reste donc l'estimation approximative par la méthode Monte Carlo. Cependant, plutôt que d'effectuer une simple somme de variables aléatoires, nous pouvons tabler ici sur la fonction mathématique elle-même et, qu'elle soit complexe ou non, nous escomptons un avantage de précision en y recourant.

La quantité Q à estimer prend normalement la forme d'une intégrale définie ou bien l'on peut l'y ramener. Cette quantité s'exprime ordinairement par :

$$Q = \int_a^b f(x)\, dx\,, \tag{10.18}$$

où chacune des bornes a et b peut être infinie. Mais on rencontre souvent, aussi, une intégrale de type similaire à l'espérance mathématique en statistique, qui a la forme :

$$Q = \int_a^b h(x)\, f(x)\, dx\,, \tag{10.19}$$

où :

$$f(x) \geq 0 \quad \text{et} \quad \int_a^b f(x)\,dx = 1 \ ,$$

f(x) étant la fonction de densité de probabilité pour la variable x. Évidemment, en définissant g(*x*) = *h(x)f(x),* nous ré-exprimons la forme (10.19) dans la forme générale (10.18). Notons, pour rappel, que, lorsque *x* est une v.a. de densité *f*(*x*), son espérance, E(X) = $m_l(X) = \mu_x$, est obtenue par (10.19) avec *h(x) = x;* quant à sa variance, var(X) = O^2_x , on obtient d'abord $m_2(X)$ par (10.19) avec *h(x)* = x^2, puis var(X) = $m_2(X)$ — $[m_1(X)]^2$.

10.11 *La technique de l'échantillonnage de fonction et ses variantes.* Si le domaine indiqué par les bornes d'intégration *a* et *b* est fini, l'évaluation de l'intégrale peut se ramener au calcul d'une moyenne. Utilisant *y = f(x)* et (10.18), nous avons en effet :

$$Q = \int_a^b y\,dx = (b-a)\, m_1(Y) \ , \qquad (10.20)$$

$m_1(Y)$ étant la valeur moyenne de la fonction. On estime alors cette valeur moyenne par :

$$\hat{m}_1(Y) = (Y_1 + Y_2 + \ldots + Y_T) \,/\, T \ , \qquad (10.21)$$

où les valeurs $Y_i = f(X_i)$ sont des échantillons simples de la fonction *f*(*x*), afin de produire l'estimation par échantillonnage de fonction simple,

$$\hat{Q} = (b-a)\, \hat{m}_1(Y) \ . \qquad (10.22)$$

L'estimateur $\hat{m}_1(Y)$ a pour variance σ^2_Y/T ; on peut approcher σ^2_Y en calculant $\{\sum Y_i^2 - T\times[\hat{m}_1(Y)]^2\}/(T-1)$. Finalement, la variance d'erreur de cet estimateur, donc sa précision, est donnée par:

$$V_\varepsilon(\hat{Q}) = (b-a)^2\sigma^2_Y \,/\, T \ . \qquad (10.23)$$

La méthode décrite ci-dessus, pour qui dispose de la fonction à intégrer, est le pendant de l'estimation Monte Carlo dite naïve, décrite en §9.2, utilisée à partir de séries de v.a.. La méthode est appliquée sans difficulté puisque, avec les formes (10.18) ou (10.19), l'argument d'intégration « *dx*» indique une valeur uniformément étalée entre les bornes d'intégration. Ainsi, pour générer les échantillons Y. exploités dans la moyenne (10.21), il suffit d'exécuter répétitivement la séquence d'opérations:

Obtenir une v.a. $u \sim U(0,1)$;
$x \leftarrow a + (b-a)\times u$;
$y \leftarrow f(x)$.

La méthode exige que les bornes a et b soient finies: voir cependant l'exercice 10.8.

On peut apporter à cette méthode générale des variations avantageuses. Ainsi, l'échantillonnage stratifié, applicable au domaine de la variable (cf. §10.8), peut être appliqué aussi au domaine de la fonction à intégrer. On peut: (1) découper le domaine d'intégration ($a..b$) de X en strates égales puis échantillonner au hasard un élément Y_i dans chaque strate; (2) séparer le domaine en deux ou quelques groupes, de grandeurs inégales (*group sampling*: Buslenko *et al.* 1966); (3) stratifier *au hasard* en découpant le domaine en intervalles de bornes et de grandeurs aléatoires (*weighted Monte Carlo*: Yakowitz *et al.* 1978; *randomized quadrature*: Evans et Swartz 2000); (4) stratifier par un découpage du domaine de la fonction en strates égales puis échantillonner dans chaque strate selon un même décalage aléatoire (symétrisation de fonction: Buslenko *et al.* 1966). En concentrant l'échantillonnage dans un intervalle $(x, x+\Delta x)$ restreint, toutes ces méthodes visent à obtenir des fragments de $m_1(Y)$ à variance réduite afin de réduire globalement $V_\varepsilon(\hat{Q})$.

Exemple 10.4 L'intégrale de $(\sin x)^{1/2}$ de 0 à π par échantillonnage de fonction

Nous reprenons ici le calcul approché de $Q = \int_0^\pi (\sin x)^{1/2}\, dx$, déjà abordé en §9.3 par la méthode du sondage de fonction. En utilisant (10.20), la moyenne $m_1(Y)$ est d'abord estimée à partir de T échantillons de $f(x)$ répartis dans le domaine $(0..\pi)$ de X.

Estimer $\int_0^\pi (\sin x)^{1/2}\, dx$ par échantillonnage de fonction

[Initialisation] $\Sigma \leftarrow 0$;
[Cycle] Exécuter T fois
$x \leftarrow \text{RND} \times \pi$;
$\Sigma \leftarrow \Sigma + (\sin x)^{1/2}$
[Calculs finals] $\hat{Q} = \pi \times \Sigma / T$.

Un calcul indépendant montre que $\text{var}(Y) \approx 0{,}05482$, d'où il ressort que l'estimateur $\hat{Q}$ par échantillonnage de fonction a comme variance d'erreur $V_\varepsilon(\hat{Q}) \approx 0{,}54105/T$, comme erreur-type $\sigma_\varepsilon \approx 0{,}73556/\sqrt{T}$. Cet estimateur s'avère donc 3,3 plus efficace, ou $\sqrt{3{,}3} \approx 1{,}82$ plus précis, que l'estimateur par sondage de fonction.

Les variantes de l'échantillonnage de fonction, notamment celles qui stratifient d'une manière ou d'une autre le domaine d'intégration, atteignent une efficacité relative au moins aussi grande que celle

obtenue par échantillonnage stratifié sur une covariable (cf. §10.8), soit de l'ordre de $V_\varepsilon(\hat{Q}_{STRAT}) \approx V_\varepsilon(\hat{Q}) / k^2$, k dénotant le nombre de strates. Pour rejoindre une telle efficacité, le procédé de stratification mis en œuvre doit être lucide et il doit tenir compte de l'anatomie globale de la fonction. La fonction $\sqrt{\sin x}$ dans le domaine $(0..\pi)$ est un cas d'espèce. Étant donné que cette fonction est symétrique autour de l'axe $x = \pi/2$, tout le calcul de $m_1(x)$ peut et devrait se faire dans le sous-domaine $(0..\pi/2)$, et l'établissement des k strates, s'il y a lieu, ne devrait pas déborder non plus ce sous-domaine.

10.12 *La technique de réduction analytique.* Approche « intelligente » de l'optimisation Monte Carlo, la réduction analytique[3] suppose une connaissance explicite ou descriptive[4] de la fonction à intégrer. Globalement, il s'agit de ré-exprimer mathématiquement la fonction en une forme nouvelle, $f'(x)$, dont la variation est moindre que celle de $f(x)$ dans le domaine d'intégration. Un expédient de la technique consiste à trouver ou à définir une co-fonction $g(x)$, d'intégrale connue, co-fonction qui est corrélée avec la fonction $f(x)$ et qu'on lui soustrait, similairement à la technique de la covariable à espérance connue (§ 10.7). Dans ce cas, il suffit de trouver une fonction imitative g(x), c'est-à-dire une fonction ayant des variations grossièrement semblables *à f(x) et* d'intégrale connue, A_g. Dans ce contexte, (10.18) devient :

$$Q = \int_a^b f(x)\, dx$$

$$= \int_a^b g(x)\, dx + \int_a^b [f(x) - g(x)]\, dx \qquad (10.24a)$$

$$= \quad A_g \quad + \quad Q_{f-g} \; ; \qquad (10.24b)$$

la variance d'erreur de l'estimateur $\hat{Q}$, laquelle porte uniquement sur la quantité résiduelle Q_{f-g}, se trouve ainsi réduite. L'exemple suivant illustre la technique par le recours à une co-fonction d'intégrale simple.

3. Buslenko *et al.* (1966) parlent de cette technique comme « l'extraction de la partie régulière » de la fonction.

4. Par l'adjectif « explicite », nous entendons bien sûr l'expression mathématique complète de la fonction, alors que la seule connaissance « descriptive » provient de l'observation du comportement et des variations globales de la fonction *f(x)* selon *x,* même si l'expression mathématique *f(x)* elle-même est hermétique et constitue une « boîte noire ».

Exemple 10.5 L'intégrale de $(\sin x)^{1/2}$ de 0 à π par réduction analytique

La fonction à intégrer pour $x = 0$ à $x = \pi$, $Y = \sqrt{\sin x}$, varie symétriquement dans les intervalles $(0, \pi/2)$ et $(\pi/2, \pi)$, de sorte qu'il suffit d'évaluer $m_1(Y)$ dans le premier intervalle, $(0\,\pi/2)$. Dans cet intervalle, la fonction se développe de 0 (pour $x = 0$) jusqu'à 1 (pour $x = \pi/2$), et son comportement est grossièrement semblable à celui de la fonction $\sqrt{x}$, tout comme x imite $\sin x$. [En fait, le développement en série de $\sqrt{\sin x}$ a, pour premiers termes, $\sqrt{x} - x^{2,5}/12 + x^{4,5}/1440$.] Ainsi, $\sqrt{x}$ représente une fonction imitative, et sa primitive, $\frac{2}{3}x^{1,5}$, permet d'obtenir facilement l'intégrale $\int_a^{\pi/2} \sqrt{x}\,dx = \frac{2}{3}(\pi/2)^{1,5} \approx 1{,}312467$, puis la moyenne de $\sqrt{x}$, $m_1(\sqrt{x}) \approx 0{,}835543$. Il reste à évaluer la moyenne résiduelle $m_1(\text{Rés}) = \left[\int_a^{\pi/2} \sqrt{\sin x} - \sqrt{x}\,dx\right] / (\pi/2)$, ce que nous faisons par échantillonnage de fonction, puis à compléter les calculs. L'algorithme approprié suit.

Estimer $\int_0^{\pi} (\sin x)^{1/2}\,dx$ par réduction analytique (soustraction de $\sqrt{x}$)

[Initialisation] $\Sigma \leftarrow 0$;
[Cycle] Exécuter T fois
$x \leftarrow \text{RND} \times \pi/2$;
$y \leftarrow \sqrt{\sin x} - \sqrt{x}$;
$\Sigma \leftarrow \Sigma + y$
[Calculs finals] $\hat{Q} \leftarrow (\Sigma/T + 0{,}835543) \times \pi$.

La valeur attendue de $m_1(\text{Rés})$, estimée par Σ/T à la dernière ligne, est $-0{,}072783$ environ. Un calcul Monte Carlo indique que cet estimateur, dont la variation est confinée à la bande résiduelle $\sqrt{\sin x} - \sqrt{x}$, a une variance d'erreur $V_\varepsilon(\hat{Q}) \approx 0{,}0541/T$, soit environ 10 fois plus petite que l'estimateur par échantillonnage de fonction naïf vu à l'exemple 10.4. Le léger coût supplémentaire de calcul, engendré par l'opération de racine carrée (*i.e.* $\sqrt{x}$) et de soustraction, n'entame pas sérieusement cet avantage.

10.13 *La technique du changement de variable.* La variabilité de l'estimateur $\hat{Q}$, en (10.21-10.22), provient essentiellement de la variation des valeurs $Y_i = f(X_i)$ de la fonction $f(\,)$ dans le domaine d'intégration en X. Une façon de réduire cette variabilité consiste donc à trouver une fonction équivalente $h(\,)$, pour laquelle la variation $Y_i = h(X_i)$ soit amortie. C'est la technique dite du changement de variable, connue dans le monde anglo-saxon sous l'appellation de « importance sampling » (Buslenko *et al.* 1966; Evans et Swartz 2000; Fishman 1996; Rubinstein 1981).

Considérant la variation de la fonction f(X) à moyenner, il faut trouver une fonction imitative g(X), c'est-à-dire une fonction dont la variation imite plus ou moins grossièrement la variation de la fonction originale. Cette fonction g(X) doit présenter deux propriétés complémentaires: (1) g(X) est une fonction de densité, telle que g(X) ≥ 0 et $\int g(X)\, dX = 1$ dans le domaine d'intégration, et (2) la variable W = g(X) peut être échantillonnée, générée. En d'autres mots, l'argument *dX,* de répartition uniforme dans le domaine d'intégration (a..b), est remplacé par l'argument *dW* = g(X) *dX* dans la nouvelle expression. La valeur de l'intégrale est alors la même, puisque:

$$Q = \int_a^b f(X)\, dX$$

$$= \int_a^b \frac{f(X)}{g(X)}\, g(X)\, dX \qquad (10.25a)$$

$$= \int_a^b h(W)\, dW\ . \qquad (10.25b)$$

La difficulté, s'il y a lieu, tient essentiellement à la génération de la nouvelle variable W.

Exemple 10.6 L'intégrale de exp(x) de 0 à 1 par changement de variable

L'intégrale $Q = \int_0^1 e^x\, dx$ permet d'illustrer simplement la technique du changement de variable. Quant à sa solution analytique, la primitive de e^x étant aussi e^x, nous avons $Q = e^x\ ]_0^1 = e-1 \approx 1{,}718282$.

La fonction e^x part de 1 (pour $x = 0$) jusqu'à $e \approx 2{,}718282$ (pour $x = 1$) dans une progression monotone ; la variable $y = 1 + e{\cdot}x$, suit un chemin semblable. Pour en faire une v.a., nous devons trouver sa fonction de densité $g(y)$ puis sa fonction de répartition $G(y)$, laquelle nous permettra (peut-être) de générer la variable y par l'inversion $G^{-1}(u) \rightarrow y$, où $u \sim U(0,1)$.

D'abord, fixons $g(y) = K(1 + e{\cdot}y)$ telle que $\int_0^1 g(y)\, dy = 1$. On obtient facilement $\int K(1+e{\cdot}y)\, dy = K(y + e{\cdot}y^2/2$, d'où $K = 1 + e/2 = 2/(2 + e)$. De là, on obtient aussi facilement la f.r. $G(y) = (2y + e{\cdot}y^2) / (2 + e)$. Égalisant $G(y) = u$, nous obtenons l'équation quadratique $e{\cdot}y^2 + 2y - (2+e){\cdot}u = 0$, qui a pour solution:

$$y \leftarrow \{\ \sqrt{[1+e(2+e){\cdot}u]} - 1\ \}\ /\ e\ ;$$

cette expression nous donne une fonction génératrice de la v.a. y a partir d'une v.a. u.

En disposant de la variable imitative y et de sa fonction de densité, la recette du changement de variable peut être programmée.

Estimer $\int_0^1 exp(x)\,dx$ *par changement de variable*

```
{ Noter que e ≈ 2,718282 }
[Initialisation ] ∑ ← 0 ;
[Cycle          ] Exécuter T fois
                      u ← RND ;
                      y ← [√(1+e(2+e)u) − 1] / e ;
                      z ← exp(y) / (1 + e·y) ;          (*)
                      ∑ ← ∑ + z
[Calculs finals ] Q̂ ← [∑ × (2 + e)/e] / T .
```

À la ligne du Cycle marquée d'un astérisque, nous avons simplifié le quotient de fonctions $f(X)/g(X)$ en supprimant le facteur constant $K = 2/(2+e)$, lequel est restitué d'un coup dans la ligne des Calculs finals.

L'estimateur $\hat{Q}$ présente ici une variance d'environ $V_\varepsilon(\hat{Q}) \approx 0{,}0177/T$. L'estimateur brut, basé sur la variable $z \leftarrow \exp(u)$, a pour variance $(4e - e^2 - 3)/2T \approx 0{,}242036/T$, donnant un rapport d'environ 13,5 : 1 en faveur de notre estimateur.

Buslenko *et al.* (1966) traitent la même intégrale en proposant la variable $y = 1 + x$, en fait une v.a. de densité $\frac{2}{3}(1 + x)$, inspirée de l'expansion de Taylor $e^X := 1 + x + x^2/2! + \ldots$ La variance d'erreur résultante est de 0,0269/T. D'autres solutions, utilisant sans doute des variables plus difficiles a générer, sont possibles.

Les auteurs cités font grand cas de la technique du changement de variable, sous le nom de « importance sampling », une technique qui fait appel a l'ingéniosité de l'usager et lui promet un gain de performance intéressant. Nous nous rallions en partie dans cet engouement, ce a la condition que la fonction a intégrer puisse être imitée par une v.a. simple et que celle-ci a la fois se génère facilement et possède une densité connue ou facile a construire. Ce bouquet de conditions favorables se rencontre hélas rarement. Les exemples un peu avancés qu'on retrouve dans la documentation spécialisée sont en fait des tours de force dont le niveau d'expertise mathématique détonne par rapport au contexte fondamentalement rustique de l'évaluation Monte Carlo.

10.14 *Autres techniques fonctionnelles.* D'autres noms de techniques visant a réduire la variance de l'estimateur Monte Carlo apparaissent dans la documentation. Certaines techniques sont des variantes de celles exposées plus haut ; d'autres, comme la technique de rejet (ou d'acceptation, voir §4.19), récupèrent les techniques appliquées pour la production de v.a. a distributions diverses, revues au chapitre 4. Quelques-unes sont franchement originales et s'adressent a des cas spécifiques, telles la simplification dimensionnelle (Evans et Swartz 2000) pour attaquer une intégrale multiple ou 1'orthogonalisation mise en oeuvre pour traiter une intégrale multinormale a matrice de corrélation quelconque (Escoufier 1967; voir aussi Kotz, Balakrishnan et Johnson 2000 pour la réduction des intégrales de probabilité multiples).

10.15 Nonobstant l'intérêt des paragraphes précédents, il reste que la réduction de la variance d'estimation Monte Carlo par manipulation de la fonction est un art, tout comme le calcul intégral et les équations différentielles; pour cette raison, elle présuppose chez l'usager une longue pratique et une dose certaine d'érudition mathématique. L'usager est confronté, d'une part, a un problème d'efficacité dans l'estimation approximative d'une quantité Q et, d'autre part, il peut connaître certaines recettes générales d'optimisation Monte Carlo; l'application de ces dernières a son problème n'est pas chose aisée et risque souvent de le laisser perplexe. Y a-t-il une stratégie globale qu'on puisse adopter?

Si la fonction a intégrer est connue, il y a peut-être mieux a faire que recourir a l'approximation Monte Carlo, comme effectuer analytiquement l'intégration ou se tourner vers l'une des méthodes d'intégration numérique disponibles (cf. chap 8). Rappelons a ce sujet les recommandations de Kahn et Marshall (1953) dans leur article fondateur sur l'optimisation Monte Carlo:

> *(a) Ne jamais utiliser l'échantillonnage aléatoire au lieu d'une méthode exacte facile pour quelque partie du problème que ce soit, et (b) ne jamais faire comme si le problème déterminait la méthode d'évaluation (échantillonnale) — il détermine seulement la valeur numérique de la réponse à trouver.* (p. 278)

Ensuite, il appert que l'échantillonnage de fonction simple (§10.11) est généralement plus efficace que l'estimation Monte Carlo de base par variable (§9.2), laquelle domine encore l'estimation par sondage de fonction (§9.3). Quant au choix d'une technique de réduction de variance de l'estimateur Monte Carlo parmi l'éventail de techniques disponible, la structure du problème et la commodité de la technique constituent des critères et devraient se convenir mutuellement; ce choix exige bien sûr une

étude minimalement raffinée du problème de la part de l'usager. De ce point de vue, les techniques les plus « accommodantes » sont peut-être celles de l'antivariable (§ 10.6), de l'échantillonnage stratifié (§ 10.8) et de la réduction analytique (§ 10.12), lorsqu'elles s'appliquent. Dans chaque cas spécifique, cependant, ce sera a l'usager de déterminer le degré de précision, le temps de calcul maximum alloué et l'effort de pénétration mathématique qui conviennent a sa situation.

Exercices

10.1 *La transformée de Fourier discrète et son accélération.* La transformée de Fourier discrète (TFD) consiste, comme son nom l'indique, à ré-exprimer une série statistique $\{ X_i, i = 0, n-1 \}$ en une autre, $\{ A_f, B_f, f = 0, n-1 \}$, selon la relation:

$$X_i = \sum_{f=0}^{n-1} [A_f \cdot \sin(2\pi \cdot f \cdot i/n) + B_f \cdot \cos(2\pi \cdot f \cdot i/n)] .$$

Les coefficients A_f et B_f s'obtiennent par:

$$A_f = \frac{1}{n} \sum_{i=0}^{n-1} X_i \cdot \sin(2\pi \cdot f \cdot i/n)$$
$$B_f = \frac{1}{n} \sum_{i=0}^{n-1} X_i \cdot \cos(2\pi \cdot f \cdot i/n) .$$

M. B. Priestley (*Spectral analysis and time series*, New York, Academic Press, 1981) couvre très bien le sujet[5].

Le programme suivant, en langage BASIC, réalise les calculs.

```
REM transformée de Fourier discrète (TFD)
INPUT "n = "; N                                          [1]
DIM X(N), A(N), B(N)                                     [2]
FOR I = 1 TO N: INPUT X(I): NEXT                         [3]
PI = 3.141592654#                                        [4]
FOR F = 0 TO N - 1                                       [5]
A1 = 0: B1 = 0                                           [6]
FOR I = 1 TO N                                           [7]
A1 = A1 + COS(2 * PI * F * (I - 1) / N) * X(I)           [8]
B1 = B1 + SIN(2 * PI * F * (I - 1) / N) * X(I)           [9]
NEXT                                                     [10]
A(F) = A1 / N: B(F) = B1 / N                             [11]
NEXT                                                     [12]
FOR F = 0 TO N - 1: PRINT F, A(F), B(F): NEXT            [13]
```

Étudier l'algorithme et scruter le programme ci-dessus afin d'en accélérer le résultat. Élaborer (au moins) une autre version. Comparer l'efficacité temporelle des versions disponibles, pour des séries de tailles $n = 20, 100, 500, 2500$.

5. Il reste intéressant de rappeler les incidences immédiates de la TFD dans le domaine de la statistique. Notons d'abord que $B_0 = \bar{X}$ et $[\sum_{f=1}^{n-1} A_f^2 + B_f^2] \times n/(n-1) = s_X^2$. De plus, pour $X_i \sim N(\mu, \sigma^2)$, $B_0 \sim N(\mu, \sigma^2/n)$ et $A_f, B_f \sim N(0, \sigma^2/(2n))$, $f = 1$ à $n/2-1$ (n pair) ou $(n-1)/2$ (n impair).

Afin d'illustrer le processus d'analyse et de dégraissage de code, nous soumettons au lecteur les quelques considérations suivantes.

Révision d'ordre algorithmique. L'analyse mathématique fine de la TFD ou, plus simplement, la fouille des documentations font bientôt apparaître un principe de calcul et un algorithme beaucoup plus efficaces que celui, naïf, indiqué ci-dessus : c'est le principe dit du FFT (Fast Fourier Transform, ou transformée de Fourier rapide), imaginé par J. W. Cooley et J. W. Tukey en 1965 (voir Priestley, *op. cit.).* Alors que le temps d'exécution de l'algorithme naïf est d'ordre $0(n^2)$, celui du FFT suit plutôt $0(n \log n)$, devenant rapidement avantageux pour n forts. W. H. Press, B. P. Flannery, S. A. Tenkolsky et W. T. Vetterling *(Numerical recipes,* Cambridge, Cambridge University Press, 1986) donnent le détail de l'algorithme et ses variantes. Le principe du FFT est en partie basé sur une décomposition de n en ses facteurs ; la version la plus répandue, et la plus performante, mise sur le nombre premier 2, exigeant alors des séries de tailles $n = 2^k$. Pour de petites valeurs de n, l'algorithme ci-dessus peut donc rester avantageux.

Révision résultant de la structure de la solution. L'examen du vecteur résultant $\{A_f, B_f\}$, aussi appelé *spectre* du vecteur temporel $\{X_t, t = 0, n—1\}$, montre que ses valeurs sont doublées et obéissent aux équations: $A_f = — A_{n-f}$, $B_f = B_{n-f}$. Ces relations, de même que $A_o = 0$ et $B_o = X$, permettent d'escamoter la moitié des calculs.

Révision (dégraissage) du code du programme. Les lignes de programme [8] et [9], dans la boucle itérative la plus profonde, sont exécutées n^2 fois. Dans ces lignes, l'expression correspondant *a 2лfn* a une valeur constante et peut être remplacée par une variable, a définir a l'entrée de la boucle. D'autres améliorations possibles concernent l'origine du vecteur X, qu'il serait avantageux de déplacer a zéro, la division préalable de chaque valeur X. par n, etc.

10.2 *Espérance d'une variable x_2 par la technique de 1'antivariable.* Une variable x_v, (dite Khi, avec v degrés de liberté) est la racine carrée d'une variable x^2_v (Khi-deux) et s'obtient en conséquence (cf. §4.8). Ainsi, pour produire un exemplaire selon $v = 2$, il suffit d'effectuer « x x← √[—21og$u$] » a partir d'une v.a. $u \sim U(0,1)$. Composer un algorithme qui permette d'estimer la moyenne (μ), ou espérance, d'une variable x_2 par la technique de l'antivariable. Notant que $\mu_x = \sqrt{/[л/2]}$, $O^2_x = (4—л)/2$, p{√[—21og$u$], .√[—21og(1—u)]}≈ 0,947094, comparer la précision relative de l'estimation par antivariable a l'estimation Monte Carlo simple.

10.3 (Suite du précédent). L'efficacité relative (ER) de la technique de l'antivariable est généralement bornée entre 1 /(1 +p) et 2/(1 +p), où p est la corrélation (négative) entre variable et antivariable. Dans le cas de la variable x_2 de l'exercice précédent, p ≈ -0,947094 et nous avons ER € (18,9; 37,8). A partir de leurs algorithmes (ou des programmes) appropriés, analyser le coût temporel approximatif de la technique de l'antivariable (t_{ANTI}) et du moyennage simple (t_{SIMPLE}) et trouver l'efficacité relative réelle de la première par ER = 2/(1 +p) *x* t_{SIMPLE} / t_{ANTI}.

10.4 *Échantillonnage stratifié uniforme (sur soi-même).* L'évaluation statistique de Q = E(U) = ½, U ~ *U*(0,1), par échantillonnage stratifié sert d'abord a poser et clarifier les concepts en jeu et définit aussi une limite optimale de précision de ce mode d'évaluation. La stratification, c'est-à-dire l'échantillonnage par strates, opère sur la variable dite de stratification (ou covariable) X, et l'on veut estimer la valeur Y de chaque élément échantillonné, soit Y =*f*(X), Y étant la variable cible. On suppose qu'il y a une dépendance statistique quelconque entre X et Y, mesurée généralement par le rapport de corrélation $\eta_{x,y}$ (10.12) ou par le coefficient de corrélation $P_{x,Y}$ (5.1). Dans le présent exemple, la variable cible Y coïncide avec la variable de stratification X, de sorte que $\eta_{x\ y} = P_{x,Y} = 1$.

Le domaine d'échantillonnage, ou l'étendue de X, est découpé en *k* strates, chaque strate captant une portion p_j, *j = 1 à k,* de la « population ». Nous découpons ici *k* strates égales dans (0,1), la strate *j* ayant pour bornes *(j -1)/k* et *j/k* et pour portion $p_j = 1/k$. Ainsi, chaque élément $_jX_i$ pigé dans la strate *j* donne lieu a une valeur $_jY_i$ correspondante: ici, $_jY_i = {}_jY_i$ (puisque $\eta_{x,y} = 1$).

Dans le contexte défini ci-dessus, montrer d'abord que, dans la strate *j*, l'espérance $\mu(_jY)$ égale $(j-1)/k$ et la variance, $1/(12k^2)$ ou σ_Y^2/k^2, où $\sigma_Y^2 = 1/12$ est la variance de U ~ *U*(0,1). Quant à la moyenne stratifiée $\bar{Y}_{st}$ (10.13), elle est basée sur les moyennes $\bar{Y}_j$ estimées dans chaque strate à partir de n_j données. Supposant un échantillonnage réparti également d'une strate à l'autre, soit $n_j = n/k$, montrer que $Q = E(\bar{Y}_{st}) = ½ = \mu_Y$, μ_Y étant l'espérance de U ~ *U*(0,1), et que $var(\bar{Y}_{st}) = \sigma_Y^2/(n \cdot k^2) = var(\bar{Y})/k^2$, $var(\bar{Y})$ représentant la variance de l'estimateur par échantillonnage simple.

En conclusion, si la variable cible est parfaitement contrôlée par le procédé de stratification (*i.e.* si $\eta_{X,Y} = 1$) et si le découpage de population est optimal selon *f*(X), l'efficacité de l'échantillonnage en *k* strates peut être jusqu'à k^2 plus grande que celle de l'échantillonnage au hasard simple.

10.5 (Suite du précédent). Soit $X \sim U(0,1)$, la variable de stratification échantillonnée en k strates égales de portions $p_j = 1/k$, et $Y = X^Q$ ($Q > 0$), la variable cible. La corrélation $\rho(X,Y|Q)$ est égale à $\sqrt{(6Q+3)/(Q+2)}$. Pour $Q = 1, 2, 3, \cdots$, montrer que la variance de la moyenne stratifiée dépend de Q, selon la formule $\text{var}(\bar{Y}_{st}) = \text{var}(\bar{Y}) \times h(Q)/k^2$, où $h(Q) = Q^2/(2Q-1)$, ce qui quantifie une atténuation de la sur-efficacité de la moyenne stratifiée à mesure que la corrélation $\rho(X,Y)$ diminue.

10.6 *Échantillonnage stratifié normal (sur soi-même).* Reprenons les conventions de l'exercice 10.4 mais cette fois pour des variables X et Y normales plutôt qu'uniformes. Ici encore, la variable cible (Y) coïncide avec la variable d'échantillonnage (X). Supposons, comme alors, que le domaine est segmenté en k strates de portions égales, $p_j = 1/k$. Pour ce faire, nous appliquons à la strate j les bornes (B_{j-1}, B_j), où $B_j = \Phi^{-1}(j/k)$, et $\Phi()$ est la f.r. de la loi normale $N(0,1)$; noter que les strates extrêmes 1 et k sont ouvertes. On peut alors évaluer μ_j et σ_j^2, l'espérance et la variance propres à la strate j.

> Soit $x_j \in \{B_{j-1}, B_j\}$, où $B_j = \Phi^{-1}(j/k)$. Posons $\varphi(x) = C \cdot \exp(-x^2/2)$, la densité normale au point x. Alors, $\varphi'(x) = -x\varphi(x)$ et $\varphi''(x) = x^2\varphi(x) - \varphi(x)$. Ces dérivées font apparaître les expressions à trouver dans $\mu_j = E(x) = \int x\varphi(x)\,dx$ et dans $E(x^2) = \int x^2\varphi(x)\,dx$, d'où on peut évaluer directement μ_j tout comme σ_j^2. Prenons un exemple, avec $k = 6$ strates, pour lequel nous évaluerons $x_5 \in \{B_4, B_5\} = \{0{,}4307; 0{,}9674\}$. Pour μ_5, nous obtenons tout de go $[\varphi(B_4) - \varphi(B_5)] / \frac{1}{6} \approx 0{,}6825$. Nous calculons ensuite $E(x_5^2)$ par $[1 + B_4\varphi(B_4) - B_5\varphi(B_5)] / \frac{1}{6} \approx 0{,}4893$, d'où $\sigma^2 = 0{,}4893 - 0{,}6825^2 \approx 0{,}0235$.

Les calculs indiqués permettent de déterminer $\text{var}(\bar{Y}_{st})$, la variance d'une moyenne stratifiée, pour des v.a. normales. Montrer que, en exprimant cette variance sous la forme $\text{var}(\bar{Y}_{st}) = \text{var}(\bar{Y})/f(k)$, f étant une fonction numérique du nombre de strates, nous obtenons que $k < f(k) < k^2$; ainsi, la stratification (égalitaire) se révéle un peu moins efficace pour des v.a. normales plutôt qu'uniformes. En fait, le facteur d'efficacité obéit presque exactement à l'équation de régression « $f(k) \approx k[0{,}242848 + 0{,}897539 \log_e(k+1{,}59)]$ », avec un coefficient de détermination (r^2) supérieur à 0,99995.

10.7 Faire l'analyse de coût des divers algorithmes d'estimation de $E(\bar{x}_h)$, la moyenne harmonique de deux v.a. uniformes, tels que décrits aux exemples 9.1, 10.2 et 10.3. Utilisant les variances d'erreur V_ε présentées, bâtir les indices d'efficacité de chacun et les comparer.

10.8 *Manipulation d'une fonction semi-bornée par changement de variable.* Les techniques de base qui procèdent par manipulation de fonction, soit le sondage de fonction (§9.3) et l'échantillonnage simple ou stratifié (10.11), supposent une variable d'intégration a la fois uniforme *et* bornée. Dans le cas d'une fonction semi-bornée, soit, par exemple :

$$Q = \int_a^{\infty} f(x)\, dx \ ;$$

le changement de la variable x en la variable $y = (x-a)/(1+x-a)$ réinscrit la fonction dans l'intervalle (0, 1). Puisque $x = y/(1-y) + a$, alors :

$$Q = \int_0^1 f[y/(1-y)+a]\ (dx/dy)\ dy$$

$$= \int_0^1 f[y/(1-y)+a]\ (1-y)^{-2}\ dy \ ;$$

cette forme, qui comporte un élément de variation uniforme dans l'intervalle (0, 1), convient donc aux méthodes d'échantillonnage.

La loi de probabilité du Khi avec $\nu = 1$ (χ_1 ; voir exercices 9.3 et 9.4), a pour fonction de densité $f(x) = \sqrt{(2/\pi)}e^{-x^2/2}$ et pour domaine (0.. ∞). Ré-exprimer cette loi dans une fonction $g(x)$ de domaine (0..1). Utiliser cette fonction pour estimer l'intégrale partielle (ou probabilité) $\Pr\{x \le 1{,}5\} = \int_0^{1,5} f(x)\, dx \approx 0{,}866396$ par échantillonnage de fonction.

10.9 (Suite du précédent). Utiliser directement la fonction de densité $f(x)$ donnée à l'exercice précédent pour estimer l'intégrale partielle $\int_0^{1,5} f(x)\, dx$, en échantillonnant dans l'intervalle (0, 1,5). Définissant l'intégrale partielle $Q(t) = \int_0^t f(x)\, dx$, comparer l'efficacité des deux méthodes pour différentes valeurs de t. Montrer que la méthode directe, préférable à l'autre pour $t \le 1{,}40$, le cède en efficacité pour $t > 1{,}40$, en particulier si l'on intègre le domaine complémentaire (y..1), où $y = t/(1+t)$.

10.10 *L'intégrale* $\int_0^{\pi} \sin x\, dx$ *par changement de variable.* Dans l'intervalle de 0 à π, la fonction « sin x » évolue symétriquement autour de π/2, d'où l'intégrale peut être calculée par $Q = 2\int_0^{\pi/2} \sin x\, dx$ (= 2), en utilisant l'argument uniforme x. Appliquer la technique du changement de variable en utilisant l'argument $\sqrt{[\pi x/2]} = \pi/2 \cdot \sqrt{u} \rightarrow w$, $h(w) = (\sin w) / (8w/\pi^2)$. Par rapport à l'échantillonnage de fonction simple (dont la variance unitaire égale $\pi^2/2 - 4$), montrer que le quotient des variances est d'environ 56 à l'avantage de la présente procédure.

10.11 (Suite du précédent). Utilisant les données de l'exercice précédent, élaborer l'algorithme (ou le programme) le plus performant pour l'échantillonnage simple et pour l'échantillonnage avec changement de variable afin d'estimer $\int_0^\pi \sin x \, dx = 2$. Comparer les indices d'efficacité des deux solutions.

10.12 *L'intégrale* $\int_0^\pi \sqrt{\sin x} \, dx$ *par sondage de fonction et stratification.* Reprendre le calcul de l'aire sous $\sqrt{\sin x}$, entre $x = 0$ et $x = \pi$, par sondage de fonction (cf. exemple 9.2), cette fois en stratifiant le domaine de la fonction en segments égaux. Exploiter la symétrie de la fonction autour de l'axe $x = \pi/2$. Montrer que, pour cette fonction, l'efficacité relative (ER) de ce procédé par rapport au sondage de fonction simple est bornée supérieurement par 1,43.

10.13 Expliquer pourquoi, de manière générale, la stratification ne réduit pas ou guère la variance de l'estimateur par sondage de fonction.

10.14 *L'intégrale* $\int_0^1 e^x \, dx$ *par réduction analytique.* L'inspection visuelle de la fonction e^x dans le domaine $x \in (0..1)$ montre une croissance monotone presque droite. L'algorithme suivant exploite la co-fonction $g(x) = x$, d'intégrale $\int_0^1 x \, dx = ½$, pour estimer $Q = \int_0^1 e^x \, dx$.

Estimer $\int_0^1 exp(x) \, dx$ *par réduction analytique (soustraction de x)*

[Initialisation] $\Sigma \leftarrow 0$;
[Cycle] Exécuter T fois
$u \leftarrow$ RND
$\Sigma \leftarrow \Sigma + \exp(u) - u$
[Calculs finals] $\hat{Q} \leftarrow \Sigma/T + 0{,}5$.

La corrélation entre x et e^x, dans l'intervalle $(0,1)$, est de $(3-e)/\sqrt{[(4e-e^2-3)/6]} \approx 0{,}991827$; par conséquent, la variance de l'estimateur $\hat{Q}$, ci-dessus, calculée par $V_\varepsilon(\hat{Q}) = \sigma^2(e^x) + \sigma^2(x) - 2\rho(x, e^x)\sigma(e^x)\sigma(x)$, égale à peu près 0,043651, comparativement à $\sigma^2(e^x) = (4e-e^2-3)/2 \approx 0{,}242036$ pour l'échantillonnage de fonction simple, soit un quotient de 5,54 à l'avantage de notre estimateur.

Déterminer l'indice d'efficacité de l'algorithme ci-dessus. Comparer les efficacités de cet algorithme et de celui de l'exemple 10.6 mettant en œuvre la technique du changement de variable.

10.15 (Suite du précédent). La fonction « exp(*x*) » varie de 1 a *e* ≈ 1,718283 pour *x* allant de 0 a 1. Afin d'éponger cette variation, le procédé le plus simple de réduction analytique est de soustraire *x*, selon « exp(*x*) — *x* ». Cette solution simple peut être étendue (et rendue un peu plus complexe) en utilisant une expression comme « exp(*x)* —b•x^c», quitte a rajouter la constante « b/(c+1) » au lieu de *0,5* dans les calculs finals. Pour c = 1 (Kalos et Whitlock 1986), montrer par calcul intégral que la valeur optimale de b égale 18—16^e ≈ 1,690309, produisant une variance d'estimation d'environ *0,003940.* Par un procédé quelconque, montrer que la solution (b≈1,641; *c≈1,441)* est optimale et produit une variance d'estimation d'environ *0,000170.*

Références

ABRASH, M. (1996). *Le zen de l'optimisationdu code.* Paris, Sybex.

BETHEL, J. (1989). Minimum variance estimation in stratified sampling. *Journal of the American Statistical Association, 84,* 260-265.

BRATLEY, P., Fox, B.L., SCHRAGE, L.E. (1987). *A guide to simulation* (2^e édition). New York, Springer-Verlag.

BUSLENKO, N.P., GOLENKO, D.I., SHREIDER, Y.A., SOBOL, I.M., SRAGOVICH, V.G. (1966). *The Monte Carlo method.* Oxford, Pergamon.

COCHRAN, W.G. (1963). *Sampling techniques.* New York, Wiley.

DALENIUS, T., HODGES, J.L. (1959). Minimum variance stratification. *Journal of the American StatisticalAssociation, 54,* 88-101.

ESCOUFIER, Y. (1967). Calculs de probabilités par une méthode de Monte Carlo pour une variable p-normale. *Revue de statistique appliquée,* 15, 5-15.

EVANS, M., SWARTZ, T. (2000). *Approximatingintegrals via Monte Carlo and deterministic methods.* Oxford, Oxford University Press.

FISHMAN, G.S. (1996). *Monte Carlo : Concepts, Algorithms, and Applications.* New York, Springer.

GENTLE, J.E. (1998). *Random number generation and Monte Carlo methods.* New York, Springer.

HAMMERSLEY, J.M., HANDSCOMB, D.C. (1964). *Monte Carlo methods.* London, Chapman and Hall.

IZENMAN, A.J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association, 86,* 205-224.

JOHNSON, N.L., KOTZ, S., BALAKRISHNAN, N. (1994, 1995). *Continuous univariate distributions,* Vols. *1* et *2* (2e édition). New York, Wiley.

KAHN, M., MARSHALL, A.W. (1953). Methods of reducing sample size in Monte Carlo computations. *Operations research, 1,* 263-278.

KALOS, M.H., WHITLOCK, P.A. (1986). *Monte Carlo methods. Vol.* 1: *Basics.* New York, Wiley.

KISH, L. (1967). *Survey sampling.* New York, Wiley.

KLEIJNEN, J.P.C. (1974). *Statistical techniques in simulation.* New York, Marcel Dekker.

KNUTH, D.E. (1968). *The art of computer programming.* Vol. *2: Seminumerical algorithms;* (1973) Vol. 3*: Sorting and searching.* Reading (MA), Addison-Wesley.

KOTZ, S., BALAKRISHNAN, N., JOHNSON, N.L. (2000). *Continous multivariate* distributions.Vol. 1: *Models and applications* (2e édition). New York, Wiley.

MARTIN, L., BAILLARGEON, G. (1989). *Statistique appliquée à la psychologie* (2*e* édition). Trois-Rivières (Québec), SMG.

MOOD, A.M., GRAYBILL, F.A., BoEs, D.C. (1974). *Introduction to the theory of statistics* (3e édition). New York, McGraw-Hill.

PARZEN, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association, 74,* 105-121.

ROBERT, C.P., *CASELLA, G.* (1999). *Monte Carlo statistical methods.* New York, Springer.

RUBINSTEIN, R.Y. (1981). *Simulation and the Monte Carlo method.* New York, Wiley.

SHAFFER, C.A. (1997). *A practical introduction to data structures and algorithm analysis.* Upper Saddle River (NJ), Prentice-Hall.

YAKOWITZ, S. KRIMMEL, J.E., SZIDAROVSKY, F. (1978). Weighted Monte Carlo integration. *SIAM journal of numerical analysis, 15,* 1289-1300.

WEISS, M.A. (1992). *Data structures and algorithm analysis.* Redwood City (CA), Benjamin/Cummings.

Chapitre

11 Études illustratives de la méthode Monte Carlo

11.1 En épilogue de ce livre sur les nombres aléatoires, nous présentons quelques exemples d'applications de la méthode Monte Carlo. Ainsi que nous l'avons esquissé plus haut, la méthode Monte Carlo peut servir a l'étude de modèles en simulant leur activité et en accumulant des données de rendement dans un contexte réaliste. C'est ainsi qu'on l'applique a des modèles de marchés de valeurs, de chambres de particules en physique, de systèmes de communication ou de production, etc. (p. ex. Fishman 1996; Gordon 1969; Kalos et Whitlock 1986; Naylor 1971 ; Sobol 1974). Les modèles statistiques, eux, ne sont pas en reste, grâce sans doute a l'affinité qu'ils ont avec les ensembles de nombres aléatoires (Diaconis et Efron 1983 ; Gentle 1998 ; Robert et Casella 1999). La méthode Monte Carlo leur est appliquée pour étudier les propriétés formelles des modèles ou, simplement, pour trouver une réponse numérique approximative d'un problème autrement insoluble.

En dernière analyse, cependant, la méthode Monte Carlo sert à estimer la valeur d'une fonction, — espérance mathématique ou intégrale définie —, en échantillonnant une variable aléatoire sur laquelle cette fonction est basée. Les exemples élaborés, que nous examinons dans ce chapitre, illustrent différents aspects de ce calcul.

Le premier exemple, soit l'estimation de E (y), $y = U_{(3:9)}$, c'est-a-dire l'espérance de la troisième statistique d'ordre d'un échantillon de 9 v.a. uniformes de distribution $U(0,1)$, sert de canevas pédagogique pour illustrer les techniques d'analyse et de réduction de variance survolées dans les chapitres précédents. Le second exemple, soit l'exécution par combinatoire approximative d'une analyse de variance de plan A x B_R, *i.e.* un plan à deux dimensions et avec des mesures répétées sur la seconde dimension, montre les contraintes, l'utilité et la puissance de la méthode Monte Carlo lorsqu'appliquée au test d'hypothèses statistiques. Le dernier exemple, de plus petite envergure, donne au lecteur une idée du degré de raffinement rendu accessible par la méthode et, parfois, de sa capacité quasi exclusive à fournir une réponse.

L 'estimation de l 'espérance E($U_{(3:9)}$)

11.2 L'estimation de la valeur moyenne, ou espérance, d'une fonction correspond à un besoin fréquent dans les applications statistiques. Nous nous attardons ici à une fonction pas trop complexe, soit la troisième statistique d'ordre d'un échantillon de 9 v.a. uniformes standard, ou la s.o.u. $U_{(3:9)}$. Toutes les propriétés des s.o.u. sont connues, bien sûr; nous les avons passées en revue au chapitre 3, en sections §3.10-3.15. Ainsi, l'espérance cherchée est égale à 0,3, et la variance de $U_{(3:9)}$, d'un échantillon à l'autre, est de $0,01\overline{90}$. Pour les fins de l'illustration, nous feindrons ignorer ces propriétés, lesquelles sont inconnues dans le cas de la plupart des populations statistiques.

Dans un premier groupe d'efforts, nous procédons par des estimations a l'aveuglette de l'espérance de $U_{(3:9)}$, sans recours aucun a la loi de densité de cette variable, mais en tentant néanmoins d'y aller intelligemment, en mettant en œuvre les recettes et astuces a notre disposition. Dans un second temps, nous exploitons cette loi, en mettant en œuvre des techniques Monte Carlo plus sophistiquées. Dans chaque cas, la précision, le coût d'exécution et l'efficacité seront mesurés. Enfin, nous obtenons l'espérance par intégration. Quant au coût d'exécution de l'estimation Monte Carlo, sa détermination peut ou bien procéder formellement, en appliquant une convention de coût a l'analyse de l'algorithme, ou bien être empiriquement approchée en chronométrant un grand nombre d'exécutions programmées, généralement 10 000, lesquelles varient quelque peu en raison du caractère aléatoire des séries de v.a. utilisées.

11.3 *Estimation par production d'échantillons et calcul d'une moyenne simple.* La méthode Monte Carlo de base, rappelons-le, consiste à produire des échantillons, a trouver pour chacun la valeur de la fonction a estimer, puis a faire la moyenne des valeurs ainsi engendrées. Dans le cas présent, cette méthode donne lieu a la procédure suivante:

1) Produire un échantillon de $n = 9$ v.a. uniformes, de loi $U(0,1)$;

2a) Trier les *9* valeurs en ordre (croissant), et

2b) Extraire la 3^{e} valeur la plus petite, soit x~$U(3:9)$

3) Sommer les valeurs x, puis en donner la moyenne .

L'étape 1, qui recourt a un générateur de nombres pseudo-alélatoires (p. ex. « RND ») de distribution $U(0,1)$, reste élémentaire, tout comme l'étape 3. L'étape 2a, le tri, risque d'être la plus coûteuse en temps d'exécution. Le schéma de programme suivant peut faire l'affaire.

Estimer $E(U_{(3:9)})$ par tri d'échantillons et moyennage simple

{ Utiliser un vecteur X[0]..X[9], trié sur place par la méthode d'insertion avec sentinelle (cf. chapitre 6, appendice) }

```
[Initialisation ] Σ ← 0 ;
[Cycle          ] Exécuter T fois
[ Production    ]   Pour i = 1 jusqu'à 9  Faire X[i] ← RND ; [1]
[ Tri           ]   Pour i = 2 jusqu'à 9  Faire              [2]
                        C ← X[i] ; X[0] ← C ; j ← i ;        [3]
                        Tantque C<X[j−1] Faire
                                X[j] ← X[j−1] ; j ← j−1      [4]
                        X[j] ← C                             [5]
[ Sommation     ]   Σ ← Σ + X[3]                             [6]
[Calculs finals ] Q̂ ← Σ / T .
```

Analyse sommaire de l'algorithme. La durée d'exécution étant évidemment proportionnelle au nombre T d'itérations, nous nous contentons ici d'estimer t_1, la durée d'un cycle, et σ_1^2, la variance unitaire de la variable concernée, ici X[3] ou $U_{(3:9)}$. La variance unitaire, nous l'avons mentionné, est égale à $\sigma_1^2 = \mathrm{var}(U_{(3:9)}) = 0{,}01\overline{90}$. L'analyse formelle des durées intra-Cycle, suivant nos conventions du chapitre 10, procède ainsi. La ligne [1] coûte 4×9 unités, la ligne [6], 2 unités. Les lignes [3] et [5], effectuées 8 fois, coûtent 4 unités à chaque fois. Quant à la ligne [4], le test « C<X[j−1] », coûtant 1 unité, est exécuté $\binom{9}{2} = 36$ fois au total, et il peut s'avérer positif, coûtant alors 2 autres unités, ou non. Ce test, qui dépend de la permutation des données observées, tourne positif une fois sur deux, en moyenne. Le coût cumulatif est de 142 unités (ou n^2+7n-2, pour produire $U_{(r:n)}$, tout r, n). L'indice d'efficacité (10.3) formel serait donc de $1 / (0{,}01\overline{90}\times 142) \approx 0{,}369$. Nous avons programmé cet algorithme et mesuré son temps d'exécution sur une machine peu performante, obtenant $t_1 \approx 7{,}26$ ms[1], et un indice empirique (IE) de $1 / (0{,}01\overline{90}\times 7{,}26) \approx 7{,}22$. Nous utilisons ces indices empiriques pour dresser, plus loin, un bilan d'efficacité de nos différentes tentatives d'estimation.

1. L'unité de temps symbolique, appliquée dans l'analyse formelle de l'algorithme, équivaut donc à une durée approximative de 7,26 ms / 142 ≈ 51 µs, pour notre contexte. L'examen systématique des correspondances entre durées formelles et durées empiriques permettrait d'aboutir à un calibrage réel des coûts d'opérations symboliques *pour un contexte (ordinateur + logiciel) donné*. Ces considérations minutieuses sur le temps d'exécution sont peu utiles à notre propos et nous n'y entrerons pas.

11.4 La méthode de base présentée ci-dessus, qui inclut la production de *9* v.a. uniformes et leur tri par la méthode d'insertion, peut être bonifiée de différentes manières. Le tri par insertion n'est relativement pas efficace, le nombre d'opérations requises pour l'effectuer étant de l'ordre de $½n^2$ (cf. exercice 9.5). La méthode de tri appelée Quicksort (voir chap. *6,* appendice), dont le coût d'exécution serait plutôt d'ordre n log n (Knuth 1973), pourrait aider a la situation. Hélas! ce n'est pas le cas: le temps moyen t_1 mesuré pour une version de programme utilisant la méthode Quicksort égale 9,74 ms (plutôt que 7,26 ms), pour un IE≈ 5,38. En fait, dans nos tests, l'insertion avec sentinelle domine le tri par Quicksort jusqu'à des vecteurs de taille $n = 1$; pour $n \geq 20$, Quicksort prédomine, et il devient deux fois plus rapide pour $n = 50$. Le désavantage de Quicksort pour de petits vecteurs émane des opérations de séparation des données et de gestion de pile, dont le coût n'est compensé que par le traitement d'un grand nombre de données.

Si le principe Quicksort (ou tout autre principe de tri sophistiqué) n'est pas avantageux pour notre petit vecteur de 9 données, il peut en être autrement d'une autre démarche. En effet, le tri par sélection[2], une méthode de tri plus simple encore que l'insertion quoique légèrement plus coûteuse, procède en fixant l'une après l'autre les positions basses du vecteur. La méthode semble avantageuse pour notre cas, car il suffit de stopper le tri a la position 3, qui fixe X[3], la valeur cherchée[3]. La durée d'exécution t_1 pour cette version est d'environ 4,28 ms, donnant lieu a un IE de 12,22, la meilleure efficacité obtenue jusqu'a présent.

11.5 *Estimation utilisant une antivariable.* La variable d'intérêt, $U_{(3:9)}$, est issue d'une population uniforme $U(1,0)$, de distribution symétrique (avec $y_1 = 0$). Dans toute distribution symétrique ayant pour axe de symétrie la valeur μ, la fonction de densité de la variable X observe l'égalité $f(X) = f(2\mu—X)$. De même, les statistiques d'ordre $X_{(r:n)}$ héritent du comportement symétrique de X, grâce a la relation de densité $h(X_{(r:n)}) = h(2\mu—X_{(n+1-r,n)})$, et ont entre elles une corrélation $p(X_{(r:n)},X_{(s.n)})$ positive. Par conséquent, on peut voir que $X_{(r:n)}$ et $2\mu—X_{(n+1-r:n)}$ sont deux variables ayant même densité, donc mêmes espérances et variances, et qu'elles sont en corrélation négative. Dans notre cas, a la variable $U_{(3:9)}$, correspond $2\mu—U_{(9+1-3:9)}$ =

2. L'algorithme, qui consiste *à* placer la plus petite valeur du sous-vecteur *X [i]..X [n]* en position X[i], comporte deux boucles itératives imbriquées, la première avec l'indice *i* variant de 1 à *n*-1, le seconde avec l'indice *j* variant de *i* (ou i+l) *à n.* Le coût d'exécution est proportionnel *à* $½n^2$, tout comme celui de l'insertion et d'un autre, basé sur le tri répété des positions voisines, connu sous le nom de *« Bubblesort ».*

3. C'est-à-dire, en référence à la note précédente, faire varier ide 1 à 3 (dans la première boucle).

1-$U_{(7:9)}$), qui peut donc servir d'antivariable. Pour ces s.o. uniformes, il appert que $p(U_{(3:9)},U_{(7:9)}) = 3/7$ (éq. 3.18).

La variable a construire, disons Y, est simplement :

$$Y = \tfrac{1}{2}[\, U_{(3:9)} + 1 - U_{(7:9)} \,] ,$$

ayant pour variance unitaire $\sigma_1^2 = 0{,}00\overline{54}$, soit 2/7 de la variance de $U_{(3:9)}$. Pour incorporer l'antivariable « $1 - U_{(7:9)}$ » dans l'estimation Monte Carlo, le seul changement requis dans l'algorithme en §11.3 touche la ligne [6] qui, au lieu de « $\Sigma \leftarrow \Sigma + X[3]$ », contient maintenant « $\Sigma \leftarrow \Sigma + (X[3] + 1 - X[7]) / 2$ » et coûte 6 unités conventionnelles; le coût total d'estimation passerait de 142 à 146. Pour le programme correspondant, nous obtenons un temps chronométré moyen de 7,38 ms, lequel, combiné à la variance unitaire, donne une efficacité IE d'environ 24,85.

L'approche avec antivariable peut encore être raffinée, par exemple en appliquant un tri par sélection qui stopperait a la variable X[7], distillant peut-être un petit gain d'efficacité[4], ou encore en réduisant l'arithmétique dans la ligne [6] (e.g. $\Sigma = \Sigma + X[3] - X[7]$), quitte a rectifier le calcul en fin de boucle . Notre propos étant d'abord didactique, nous n'insistons pas sur ces « améliorations » même si, selon nous, elles sont généralement souhaitables et devraient préoccuper le programmeur au moment de finaliser son travail.

11.6 *Estimation utilisant une covariable à espérance connue.* L'utilisation d'une covariable fournit une autre technique aveugle permettant de réduire la variance d'estimation. Dans notre exemple qui exploite des échantillons de 9 v.a. uniformes, deux covariables à espérance connue s'imposent, soit la médiane $U_{(5:9)}$ et la moyenne $\bar{U}$, qui ont toutes deux une espérance égale à ½. Une fois le tri complété, l'exploitation de $U_{(5:9)}$ est immédiate tandis que, pour $\bar{U}$, il faut, à chaque itération, effectuer la sommation des 9 v.a.u. générées. Par (éq. 3.18), nous trouvons $\rho = \sqrt{(3/7)} \approx 0{,}655$ entre $U_{(3:9)}$ et $U_{(5:9)}$ et, après quelque algèbre, $\rho = \sqrt{(7/11)} \approx 0{,}798$ entre $U_{(3:9)}$ et $\bar{U}$. Nous examinons d'abord la solution naïve, qui consiste à forcer « $b = 1$ » dans (10.6), puis la solution « éclairée » en appliquant la valeur optimale de b.

Appliquant les expressions idoines de variances et de covariance, la variable compensée $\hat{X} = U_{(3:9)} - (U_{(5:9)} - \tfrac{1}{2})$ a pour variance unitaire $\sigma_1^2 =$

4. Pour cette version, le temps t, avoisine 6,98 ms et une efficacité de 26,28.
5. Pour cette solution, nous obtenons t, 6,29 ms et IE 25,13.

$0{,}01\overline{45}$; dans l'algorithme en §11.3, seule la ligne [6] est changée[6], pour un coût formel total de 143, un temps t_1 estimé de 7,32 ms et un IE d'environ 9,39. Pour la variable compensée $\hat{X} = U_{(3:9)} - (\bar{U} - ½)$, elle a pour variance $\sigma_1^2 \approx 0{,}007138$. Si la réduction de variance est plus grande qu'avec $U_{(5:9)}$, les ajouts à l'algorithme d'estimation s'avèrent aussi plus importants, notamment à la ligne [1], de sorte que le coût formel total s'élève à 163[7]. Le temps unitaire chronométré est de 7,89 ms pour notre version, avec un IE ≈ 17,76 ; l'alourdissement du calcul, nous le voyons, est plus que compensé par la réduction de variance.

Notre exemple, quelque peu artificiel, nous met à même de trouver la valeur optimale de b dans (10.6) ; cet avantage est exceptionnel et a peu de chances de se présenter dans un cas d'estimation pratique. Utilisant donc $U_{(3:9)}$ et l'équation (10.9), nous obtenons $\hat{b} = 0{,}75$. Avec (10.8), nous calculons $\sigma_1^2 \approx 0{,}011420$, puis estimons un t_1 moyen de 7,31[8] et un IE d'environ 11,99. Pour $\bar{U}$, nous obtenons $\hat{b} = 1{,}1\overline{45}$ et $\sigma_1^2 \approx 0{,}006942$, puis, par exécution programmée, $t_1 \approx 7{,}88$ ms et IE ≈ 18,29. Les indices de performance sont légèrement améliorés par le recours aux valeurs optimales de b (plutôt que $b = 1$).

Y a-t-il d'autres améliorations possibles de cette approche? Oui, car, dans le cas présent, le recours a la covariable $\bar{U}$ n'exclut pas l'accélération déjà éprouvée et qui consiste a faire un tri par sélection jusqu'à X[3] seulement. En effet, le calcul de $\bar{U}$ peut s'effectuer en amont du tri (voir note 6), ce qui permet la solution sommaire d'un tri partiel par sélection. Utilisant donc la covariable $\bar{U}$ en solution naïve (avec b = 1) et ce tri partiel, le temps d'exécution est ramené a $t_1 \approx 4{,}89$ ms, donnant un IE d'environ 28,62, notre meilleure performance jusqu'ici.

11.7 *Estimation par échantillonnage de fonction.* La fonction mathématique a laquelle obéit la variable aléatoire étudiée nous est parfois connue, ce qui ouvre de nouvelles avenues d'estimation. Lorsqu'elle est

6. En fait, la ligne [6] devient « Σ ← Σ + X[3] − X[5] », tandis que la ligne de Conclusion, dont le coût ponctuel est ignoré, devient « $\hat{Q}$ ← Σ / T + ½ ».

7. Pour établir la covariable $\bar{U}$, dénotée XM, la ligne [1] de l'algorithme, en §11.3, devient « XM ← 0 ; Pour i = 1 jusqu'à 9 Faire X[i] ← RND ; XM ← XM + X[i] ». Les autres changements sont, à la ligne [6], « Σ ← Σ + X[3] − XM / 9 » et, à la ligne [9] et sans imputation de coût, « $\hat{Q}$ ← Σ / T + ½ ».

8. La multiplication supplémentaire par b, requise à la ligne [6], entraîne un supplément de coût. Le temps moyen obtenu de 7,31 ms est apparu un peu plus faible que celui de 7,32 ms dans la version n'impliquant pas le recours au multiplicateur b. Cette minime différence, basée néanmoins sur 10000 itérations, est imputable à la fluctuation des temps d'exécution, notamment pour le tri des valeurs aléatoires.

suffisamment simple, on peut manipuler la fonction directement pour obtenir une expression analytique de son intégrale ou, sinon, l'intégrer par une méthode numérique. Dans le cas présent, d'après l'expression générale (3.13), la densité de la s.o.u. $U_{(3:9)}$ est:

$$\mathrm{pr}_{U_{(3:9)}}(x) = 252x^2(1-x)^6\ ,\quad 0 \le x \le 1\ . \tag{11.1}$$

Cette expression (polynomiale) se prête facilement a différents traitements, que nous considérons plus loin.

Soit pr(x), la fonction de densité d'une v.a. X, et $f(x) = x\bullet\mathrm{pr}(x)$, sa « fonction d'espérance » telle que $Q = \int_a^b f(x)\,dx = E(x)$, l'espérance cherchée. L'échantillonnage de fonction (§10.11) revient ici a estimer l'espérance
$E(U_{(3:9)})$ en échantillonnant au hasard la fonction « $x\bullet\mathrm{pr}(x)$ » = « $252x^3(1-x)^6$ », pour x variant *uniformément* entre 0 et 1, et en en faisant la moyenne. L'algorithme suivant illustre le procédé.

Estimer $E\{U_{(3:9)}\}$ par échantillonnage de fonction

```
[Initialisation ] ∑ ← 0 ;
[Cycle          ] Exécuter T fois
                      Produire u ;                          [1]
                      ∑ ← ∑ + 252×u^3×(1−u)^6               [2]
[Calculs finals ] Q̂ ← ∑ / T .
```

Analyse sommaire de l'algorithme. Seules, les lignes de code [1] et [2] sous-tendent le coût formel de cet algorithme ; la première vaut 3 unités, la seconde, 13 (en comptant chaque élévation a une puissance comme 4 unités), d'où un total de 16 unités (comparativement a 142 pour la méthode de référence, en §11.3). La durée chronométrée, quant à elle, est d'environ 0,58 ms (vs 7,26 ms en référence). Ainsi, le ratio des coûts formels, soit 142/13≈10,9, se compare assez bien au ratio des coûts de programmes réels, 7,26/0,58≈12,5. Quant a la variance unitaire de l'échantillonnage de fonction, nous avons par ailleurs obtenu $E\{[x\bullet\mathrm{pr}(x)]^2\}\approx 0{,}180043$, puis $O^2_1\approx 0{,}180043 - 0{,}3^2 = 0{,}090043$. La méthode affiche donc une efficacité résultante de $1 / (0{,}58\text{x}0{,}090043)\approx 19{,}15$.

11.8 *Estimation par sondage de fonction.* Semblable a la précédente, la méthode de sondage de fonction exploite elle aussi la « fonction d'espérance » de notre variable x, étant donné que l'intégrale Q a obtenir ici est le produit de x par sa densité pr(x), soit $f(x) = 252x^3(1-x)^6$.

La variable d'intégration x, qui varie de 0 a 1, a pour mode $x_M = ⅓$ selon $f'(x_M) = 0$, d'où f culmine a $M = f(⅓)\approx 0{,}819387$. Le procédé de

sondage consiste alors à produire x (= u') et u'', puis à comparer $f(x)$ à $M \cdot u''$: la proportion de fois pour lesquelles $f(x) > M \cdot u''$ correspond à Q/M. Dans l'algorithme en §11.7, la ligne [1] devient « Produire u' et u'' », et la [2], « Si $252 \times u^3 \times (1-u)^6 > M \times u''$ alors $\sum \leftarrow \sum + 1$ ». Quant aux Calculs finals, « $\hat{Q} \leftarrow \sum / T \times M$ ». La valeur t_1 estimée pour cette version est de 0,65 ms, la variance, $\sigma_1^2 = Q \times (M-Q) \approx 0{,}155816$, enfin l'IE ≈ 9,90.

11.9 *Estimation par la méthode de rejet.* Pour estimer $Q = E(x)$, on peut soit générer la v.a. $x = U_{(3:9)}$ indirectement (cf. § 11.3) ou directement (cf. §1.11), soit générer directement une v.a. plus simple, disons y, et l'accepter seulement si elle est admissible ou, sinon, la rejeter: c'est la technique du rejet (§4.19). Soit la v.a. x, de domaine (0..1) ; sa densité pr(x), en éq. (11.1), culmine au point x_M = ¼, selon pr' (x_M) = 0, a la valeur M = pr(¼)≈ 2,803162. Il suffit ainsi de produire une v.a. candidate y, de densité uniforme $h(y) = 1$, et u, toutes deux y, $u \sim U(0,1)$. Appliquant alors la règle (4.34), on « accepte » y ($\rightarrow x$) si et seulement si $u \leq M^{-1}$ [pr(y) / h(y)] ou $M \bullet u \leq$ pr(y). Voici l'algorithme d'estimation.

Estimer E{ $U_{(3:9)}$ } par rejet d'une v.a. uniforme

{ Noter que $f(x) \leq M = f(¼) \approx 2{,}803162$ }
[Initialisation] $\sum \leftarrow 0$;
[Cycle] Exécuter T fois
Répéter Produire u' et u'' (= y) [1]
Jusqu'à $M \times u' \leq 252 \times y^2 \times (1-y)^6$; [2]
$\sum \leftarrow \sum + y$ [3]
[Calculs finals] $\hat{Q} \leftarrow \sum / T$.

Analyse sommaire de l'algorithme. Les lignes [1] et [2] valent, formellement, 19 unités, la ligne [3], 2 unités. Toutefois, la satisfaction du test en ligne [2] obéit à une loi géométrique (cf. §4.5) de paramètre π = 1/M et d'espérance $1/\pi = M$. Ainsi, le coût attendu de l'algorithme serait de $M \times 19 + 2 \approx 55{,}26$. La variable y admise ayant alors une distribution comme $U_{(3:9)}$, sa variance unitaire est la même, soit $0{,}01\overline{90}$. Le temps t_1 d'exécution chronométrée étant de 1,83 ms, nous calculons IE ≈ 28,65, une efficacité excellente.

11.10 *Estimation par changement de variable.* Soit $Q = \int_0^1 f(x)\,dx$, *où f(x)* est la « fonction d'espérance » de $U_{(3:9)}$ selon (11.1), soit $f(x) = x \bullet$ pr(x) = $252x^3(1-x)^6$. L'échantillonnage de fonction, en §11.7, a consisté a estimer $Q = E\{f(x)\}$ en échantillonnant la v.a. uniforme x dans le domaine (0..1) la variance effective, var{ $f(x)$ }, était de 0,090043. Pour réduire cette variance, il faudrait échantillonner, dans le même domaine, une v.a. dont la

densité, disons g(y), approcherait $f(x)$ de sorte que la variation de la fonction restructurée $h(y) = f(y)/g(y)$ soit moindre.

Nous optons pour une variable Khi-deux (Laurencelle et Dupuis 2000) à titre de variable imitatrice. La v.a. y de loi $\chi^2(\nu)$ est positive, de domaine semi-borné (0..∞) et d'asymétrie positive (4.53c), comme $f(x)$. Sa fonction de densité (4.17), soit $K_1 e^{-y/2} y^{\nu/2-1}$, contient une puissance de y, tout comme la fonction d'espérance $f(x)$, soit $K_2(1-x)^6 x^3$. Enfin, les v.a. y de loi χ^2_ν se génèrent facilement. Dans $f(x)$, l'exposant de x égale 3 ; pour obtenir un tel exposant dans (4.17), il nous faut $\nu = 8$. De plus, le mode de x étant ⅓, d'après $f'(x_M) = 0$, et celui de $\chi^2(\nu)$ étant $\nu-2 = 6$, nous échantillonnerons la variable $w = y/18$, faisant ainsi coïncider les modes, en enfermant cependant la v.a. w dans l'intervalle $0 \le w \le 1$, dont l'intégrale de probabilité est $C_{18} = P_y(18) = \Pr\{\, \chi^2_8 \le 18 \,\} \approx 0{,}978773514$. La densité résultante de notre $y/18 = w$ est donc $g(w) = 1093{,}5 e^{-9w} w^3 / C_{18}$, et la fonction d'espérance restructurée, $f(w)/g(w) \approx 0{,}225561 e^{9w}(1-w)^6$. Le schéma de programme suivant illustre le procédé.

Estimer $E\{U_{(3:9)}\}$ *par changement de variable* $(\chi^2_8/18)$

$\{ M = 252 \times 96 \times 18^{-4} \times C_{18} \approx 0{,}225561 \; ; \; K = -2 / 18 \}$
[Initialisation] $\Sigma \leftarrow 0$;
[Cycle] Exécuter T fois
Répéter [1]
Produire u_1, u_2, u_3, u_4 ; [2]
$w \leftarrow K \times \log(u_1 \times u_2 \times u_3 \times u_4)$ [3]
Jusqu'à $w \le 1$; [4]
$\Sigma \leftarrow \Sigma + M \times \exp(9 \times w) \times (1-w)^6$ [5]
[Calculs finals] $\hat{Q} \leftarrow \Sigma / T$.

Analyse sommaire de l'algorithme. L'analyse formelle est laissée en exercice 11.2. Nous avons estimé empiriquement la variance de notre fonction restructurée, obtenant environ $\hat{\sigma}^2_1 = 0{,}0142$; cela représente une réduction par un facteur de 6,3 par rapport à la fonction d'espérance originelle. Au chronomètre, chaque cycle d'estimation coûte $t_1 \approx 0{,}82$ ms. L'efficacité, finalement, est d'à peu près 85,97.

11.11 *Estimation par production directe de* $U_{(3,9)}$. Notre variable, $x = U_{(3:9)}$, est une statistique d'ordre uniforme, et il existe quelques méthodes afin de générer une séquence $U_{(r:n)}$, pour $r = 1, 2, \ldots$, ou pour $r = n, n-1, \ldots$ (voir §3.12). En appliquant la méthode récursive de Lurie et Hartley (1972, voir notre éq. (3.25)) et utilisant les trois v.a. $u_1, u_2, u_3 \sim U(0,1)$, nous pouvons effectuer:

Tableau 1 Performance de différents procédés d'estimation Monte Carlo pour évaluer l'espérance E{ $U_{(3:9)}$ }

Méthode d'estimation	σ^2_1	t_1 (ms)	IE	Réf.
Production d'échantillons, tri (insertion) et moyennage de X[3]	0,0191	7,26	7,22	§11.3
Production d'échantillons, tri (sélection jusqu'à X[3]) et moyennage de X[3]	0,0191	4,28	12,22	§11.4
Production d'échantillons, tri (insertion) et utilisation d'une antivariable, selon ½(X[3] + 1-X[7])	0,0055	7,38	24,85	§11.5
Production d'échantillons, tri (insertion) et utilisation de la covariable X[5], selon X[3] - (X[5] -½)	0,0145	7,32	9,39	§11.6
Production d'échantillons, tri (insertion) et utilisation de la covariable Ū, selon X[3]- (Ū-'/2)	0,0071	7,89	17,76	§11.6
Production d'échantillons, tri (sélection jusqu'à X[3]) et utilisation de la covariable Ū, selon X[3] - (Ū-'/2)	0,0071	4,89	28,62	§11.6
Échantillonnage de la fonction d'espérance *«x• v f(x)»*	0,0900	0,58	19,15	§11.7
Sondage (de l'aire) de fonction « x• *f*(x) »	0,1558	0,65	9,90	§11.8
Production de $U_{(3:9)}$ par rejet de U	0,0191	1,83	28,65	§11.9
Changement de variable, avec échantillonnage d'une x^2_8	0,0142	0,82	85,97	§11.10
Production directe de $U_{(3:9)}$ selon Lurie et Hartley 1972	0,0191	0,73	71,67	§11.11

$$x \leftarrow 1 - u_1^{1/9} \times u_2^{1/8} \times u_3^{1/7} ,$$

obtenant immédiatement $x \sim f(U_{(3:9)})$. Cette ligne de production remplace à elle seule les lignes [1] à [5] dans l'algorithme Monte Carlo de base, en §11.3 ; le coût formel avoisine 41 unités. Le temps d'exécution unitaire s'avère de t_1 = 0,73 ms. Avec la variance unitaire σ_1^2 = 0,0190, nous obtenons comme efficacité la valeur IE ≈ 71,67.

11.12 Nous n'avons pas, dans les paragraphes précédents, épuisé le magasin de moyens et d'astuces permettant d'accélérer l'estimation Monte

Carlo. Notre but était seulement de faire voir que, a partir de « l'approche de base », que ce soit par échantillonnage de la variable (§ 11.3) ou de la fonction (§ 11.7), il est intéressant et peut être productif d'explorer d'autres procédés. Le tableau 1, sur la page suivante, résume l'essentiel de cette exploration.

La technique de changement de variable *(« importance sampling»,* en anglais), c'est bien connu, donne souvent d'excellents résultats. Toutefois, les ressources et l'ingéniosité habituellement requises pour la mettre en œuvre en font une curiosité plutôt qu'une solution pratique, au contraire des techniques de l'antivariable ou de rejet, par exemple.

11.13 Nous ne pouvons pas mettre le point final au traitement de cet exemple sans mentionner la solution exacte du problème abordé. La variable $x = U_{(3:9)}$ est une variable *Bêta* (§3.11) ayant comme fonction de densité l'expression (11.1). Si elle n'était pas déjà connue, nous pourrions obtenir Q, l'espérance cherchée, en calculant l'intégrale $\int_0^1 x \cdot 252x^2(1-x)^6\, dx$. Opérant par intégration directe (cf. §8.2), nous pouvons développer l'expression polynomiale, puis, en l'intégrant terme à terme, obtenir $252\times[1/4-6/5+15/6-20/7+15/8-6/9+1/10] = 0{,}3$. En procédant par intégration par parties (exercice 8.3), avec $v = 252x^3$ et $du = (1-x)^6\, dx$, nous aboutissons à $252\times6/(7\times8\times9\times10) = 0{,}3$. Les moments à l'origine μ'_k des s.o.u. $U_{(r:n)}$ étant $r^{(k)}/(n+1)^{(k)}$ (Laurencelle 1993), où $x^{(k)}$ dénote une factorielle ascendante, nous avons encore et toujours $\mu'_1[U_{(3:9)}] = Q = 3^{(1)}/(9+1)^{(1)} = 3/10 = 0{,}3$. Les méthodes d'intégration numérique (§8.3) donneraient, en approximation, la même réponse.

L'analyse de variance de plan A x B_R en solution Monte Carlo

11.14 La statistique et les nombres aléatoires se nourrissant aux mêmes sources, il n'est pas surprenant que tous les secteurs de la statistique appliquée témoignent de la présence de la méthode Monte Carlo. Ceci est vrai en particulier des tests d'hypothèses qui, dans leur version Monte Carlo, sont aussi nommés « tests permutationnels », « tests par randomisation » ou « tests par combinatoire approximative » (Edgington 1980; Laurencelle 1987; Sokal et Rohlf 1981).

L'analyse de variance (ANOVA) a été créée par R. A. Fisher pour assurer le traitement statistique de données issues de protocoles expérimentaux complexes, enregistrées dans des tableaux structurés; c'est une méthode générique de tests d'hypothèses. L'ANOVA a deux dimensions, avec mesures répétées sur la seconde dimension, est dénotée AxB_R. Cette application, un peu complexe, est néanmoins d'usage très

fréquent et elle nous permet d'illustrer les subtilités et la force de l'approche Monte Carlo dans l'inférence statistique. Pour les fondements algébriques et statistiques de cette technique et des illustrations pratiques, nous renvoyons le lecteur a la documentation classique (Kirk 1982; Laurencelle 1998a; Winer 1991).

Le principe général du test permutationnel et de sa solution Monte Carlo pour les tests d'hypothèses est le suivant. On veut décider si des différences observées entre des blocs de données sont réelles, sérieuses, significatives. Selon la stipulation théorique désignée « hypothèse nulle », ou Ho, seul le hasard joue, et les différences constatées résultent de fluctuations de hasard, sans cause assignée. Prenons une mesure M_{obs} de ces différences (« Obs » pour « valeur observée ») : la fonction « M » dépend du contexte. Sous H_o, les données disponibles ont un arrangement observé réputé aléatoire, cet arrangement faisant partie de l'ensemble des T arrangements différents, tous également possibles. Pour chacun de ces arrangements virtuellement équivalents sous H_o, on peut établir une mesure M_{var} (« Var » pour « valeur d'arrangement variable »); l'ensemble des mesures M_{var} constitue la « distribution nulle » de M, celle à laquelle notre mesure $_{Mobs}$ devrait se conformer si nul facteur n'a eu d'influence réelle sur nos données. Si, par contre, une cause assignée est réellement à l'œuvre dans les données, les différences dans l'arrangement observé tendront a être fortes et a ressortir, et la mesure M_{obs} tendra a occuper un rang centile exceptionnel dans la distribution nulle. L'hypothèse nulle pourra être rejetée si, parmi l'ensemble des T mesures M_{var} ou, en solution Monte Carlo, parmi un échantillon aléatoire de T' mesures, la valeur observée M_{obs} apparaît suffisamment exceptionnelle, par exemple selon $Pr\{M_{va} \geq M_{obs}\} \leq \alpha$; la probabilité « Pr » est indiquée par le rang centile de M_{obs} dans l'ensemble des T ou T' mesures ^{M}var.

Avant d'exposer la solution Monte Carlo détaillée, demandons-nous d'abord pour quelles raisons ou a la suite de quelles prémisses nous en viendrions a utiliser l'ANOVA a critère Monte Carlo plutôt que d'appliquer le test usuel, qui exploite la distribution F (§4.10)? La première considération a faire est que, sauf pour des analyses impliquant très peu de sujets (ou lignes de données), la solution Monte Carlo équivaut en puissance[9] a la solution classique, c'est-à-dire qu'elle a presque la même puissance sous les conditions optimales de l'approche classique et qu'elle

9. La « puissance statistique » est techniquement définie comme la probabilité de rejeter l'hypothèse nulle (H_o) sous des conditions déterminées, notamment que H_o est fausse. Elle indique la capacité qu'a une procédure de test de détecter des différences réelles dans les données, sa sensibilité aux variations systématiques (plutôt qu'aléatoires).

Tableau 2 Arrangement des données d'un plan d'analyse $A \times B_R$

Groupe	Code de groupe	B_1	B_2	B_j	B_J
A_1	1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,J}$
	1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,J}$
	...				
A_2	2	$X_{2,1}$	$X_{2,2}$	...	$X_{2,J}$
	...				
A_i	i	$X_{i,1}$	$X_{i,2}$	...	$X_{i,J}$
	...				
A_I	I	$X_{I,1}$	$X_{I,2}$	...	$X_{I,J}$
	I	$X_{I,1}$	$X_{I,2}$	...	$X_{I,J}$

n'en perd aucunement sous un changement de conditions. Advenant des violations patentes des conditions de validité de la solution classique, ou après une manipulation spéciale des données qui répond a un besoin justifié du chercheur, ou lorsque les degrés de liberté du quotient *F* sont trop faibles, la solution Monte Carlo fournira souvent la seule réponse possible.

11.15 Le plan de données AxB_R est constitué de S sujets répartis en I groupes, dénotés A_1, A_2, ..., A_I, chaque sujet étant mesuré J fois, soit aux moments B_1, B_2, ..., B_J. Chaque groupe comporte n_i sujets, où $n_1+n_2+ ... + n_I = S$. Le tableau 2 montre une disposition possible des données.

Trois tests d'hypothèses sont en jeu dans cette analyse : un test sur A (F_A), indiquant si les groupes diffèrent vraiment entre eux ; sur B *(F_B),* indiquant si, tous groupes confondus, les sujets varient d'un moment B_j a l'autre; sur l'interaction AxB (F_{AB}), indiquant si les variations d'un moment B_j a l'autre diffèrent d'un groupe A. a l'autre. Un sommaire des calculs, auquel nous revenons en appendice, apparaît au tableau 3.

L'analyse des données produit donc trois mesures, ou tests : F_A, F_B et F_{AB} ; ce sont la les valeurs observées. Le nombre d'arrangements différents possibles (T) est, comme c'est le cas habituel, extrêmement élevé[10]. Plutôt que de traiter la combinatoire complète du problème, nous

10. Pour le plan d'analyse AxB_R, $T = (J!)^S M_I(n ; n_1, n_2, ... , n_I)$, où $M_k( )$ est un nombre multinomial (6.46), égal *à* $(n_1+n_2+...+n_I)!/[n_1!n_2!...n_I!)$. Le nombre de valeurs différentes des statistiques Mv_{ar} résultant de ces arrangements est, bien sûr, légèrement moindre. Laurencelle (1987) donne d'autres indications.

Tableau 3. Sommaire des calculs de l'analyse de variance à plan $A \times B_R$

Source de variation	Somme de carrés	Degrés de liberté	Carré moyen	Quotient F
Totale	(SC_{Totale}	$SJ-1$)		
Inter-sujets	SC_1	$S-1$		
Groupes (A)	SC_A	$I-1$	CM_A	$CM_A / CM_{I\text{-}G}$
Intragroupe	$SC_{I\text{-}G}$	$S-I$	$CM_{I\text{-}G}$	
Intrasujet	SC_2	$S(J-1)$		
Moments (B)	SC_B	$J-1$	CM_B	CM_B / CM_{BS}
A × B	SC_{AB}	$(I-1)(J-1)$	CM_{AB}	CM_{AB} / CM_{BS}
B × Sujets	SC_{BS}	$(S-I)(J-1)$	CM_{BS}	

procédons plutôt par échantillonnage au hasard, ou combinatoire approximative: c'est la solution Monte Carlo.

Il s'agit donc de générer un grand nombre (T') de ré-arrangements aléatoires du tableau de données et, pour chaque ré-arrangement, de re-calculer les trois mesures (F'_A, F'_B, F'_{AB}), puis d'établir progressivement le rang des mesures F_A, F_B et F_{AB} en comptant: $N_A = \#(F'_A \geq F_A)$, $N_B = \#(F'_B \geq F_B)$, $N_{AB} = \#(F'_{AB} \geq F_{AB})$. Finalement, les estimations de probabilité sont obtenues par les quotients:

$$Pr\{A\} \approx N_A / T' \; ; \; Pr\{B\} \approx N_B / T' \; ; \; Pr\{A \times B\} \approx N_{AB} / T' \; ;$$

La validité (Hoeffding 1952) et la précision (Laurencelle 1987) de ces estimations de probabilité sont tout a fait rassurantes.

11.16 Pour réaliser la procédure de test Monte Carlo, il reste a préciser ce qu'on entend spécifiquement par « générer un ré-arrangement aléatoire ». L'opération de ré-arrangement des données doit tenir compte de la structure du problème a l'étude. Dans notre plan AxB_R, l'hypothèse nulle générale se particularise en trois hypothèses particulières, l'une sur les effets des groupes l'autre sur les effets des moments B_j, enfin la troisième sur les effets conjoints (ou différentiels) AB_{ij} ; de plus, le même sujet est mesuré sous les différents moments (ou conditions) B_j. Le premier ordre de réarrangements touche donc l'assignation des sujets aux I groupes. Si l'appartenance des données d'un sujet dans un groupe est le fait du hasard, ce sujet, avec ses J mesures, peut être ré-affecté au hasard, parmi les

I groupes, en respectant cependant les tailles n_t initiales. Le second ordre de ré-arrangements touche la série des J mesures d'un sujet a travers les moments B_i: si les moments (ou conditions expérimentales) B_j sont sans effet réel ou systématique sur les données, chaque série peut être réarrangée, *i.e.* permutée au hasard et ce, indépendamment pour les S sujets.

Une méthode commode pour réaliser ce travail de ré-arrangement par informatique suppose que les données, tel qu'au tableau 2, sont inscrites dans une matrice d'ordre SxJ, disons DONNÉES (S,J), et que l'appartenance de groupe est indiquée dans un vecteur d'ordre S, disons GROUPE(S), contenant 1 pour chaque sujet du groupe 1, 2 pour le groupe 2, etc. Avec ces structures, le travail de ré-arrangements des sujets a travers les groupes se réduira a permuter au hasard les S données du vecteur GROUPE() ; quant aux ré-arrangements des séries de J données de chaque sujet, il faudra y procéder ligne par ligne dans la matrice DONNÉES.

11.17 Nous fournissons, en appendice du chapitre, une version programmée de cette analyse, en langage BASIC. Notre solution, auquel le lecteur pourra se référer, mérite l'examen en ce qu'elle présente l'opérationnalisation détaillée des idées présentées ci-dessus et qu'elle contient plusieurs optimisations, notamment pour le re-calcul de l'ANOVA a chaque ré-arrangement. Noter enfin que, avec un nombre de réarrangements aléatoires (T') de 10000, la décision statistique encourue atteint une précision et une puissance relative plus que satisfaisantes (Laurencelle 1987).

Le « profil 4-8 » au MMPI de 32 délinquants sexuels est-il exceptionnel ?

11.18 L'évaluation psychologique, faite au moyen de tests standardisés, pose quelquefois des défis d'ordre statistique. Plusieurs tests utilisés sont de structure multidimensionnelle, qu'il s'agisse d'*inventaires de personnalité* comme le « Minnesota Multiphasic Personality Inventory » (MMPI) (Hathaway et McKinley 1996) ou le « 16 Personality Factors » (16PF) de R.B. Cattell (Cattell *et al.* 1993), ou bien d'épreuves d'aptitudes, tels les célèbres tests de Quotient Intellectuel. Le chercheur ou le praticien qui les exploitent obtiennent donc, pour un test donné, plusieurs scores censés caractériser le client, ou le patient, les propriétés statistiques de ces vecteurs de scores n'étant pas vraiment documentées dans les manuels (voir aussi Laurencelle 1998b). L'exemple du « profil 4-8 » au MMPI (Parisien et Laurencelle 1990) illustre ce point.

Le MMPI, un test classique pour dépister des troubles névrotiques ou psychotiques chez des patients, est basé sur un questionnaire de 566 questions, a répondre par Vrai ou Faux, et il produit dix scores de base (ou échelles), numérotés de 1 a 10, en même temps que maints scores auxiliaires. La série des dix scores obtenus par un patient, ou la représentation graphique de cette série, constitue son profil de base, dont le spécialiste peut s'inspirer pour le diagnostic de la personnalité troublée.

À l'occasion d'une recherche portant sur des délinquants sexuels juvéniles, dans le région de Montréal, l'examen des profils MMPI a fait apparaître une double élévation de scores[11] plus fréquente, pour la paire de scores 4-8[12], ce dans 8 cas sur 32. La question se posait de savoir s'il s'agissait d'un trait particulier de ce sous-groupe de personnes ou si le hasard seul suffisait a justifier ce résultat.

11.19 La réponse finale a la question posée dépasse largement le cadre d'un résultat statistique, si sophistiqué soit-il. Néanmoins, dans le présent exemple ainsi que dans les innombrables applications où la statistique est interpellée, sa contribution est importante car elle permet de clarifier les avenues de solution possibles, allant parfois jusqu'à montrer l'horizon où gît la réponse.

Quelle question nous est posée dans l'exemple du profil 4-8 ? Sur 32 dossiers MMPI, constitués chacun d'une série de 10 scores, on observe 8 fois le double maximum « 4-8 ». Supposons maintenant que (a) chacune des dix échelles, exprimée dans des unités standard, est susceptible des mêmes variations, et (b) chaque échelle varie de façon indépendante (i.e. sans corrélation) par rapport aux autres. Ces suppositions[13], tout irréalistes soient-elles, sont souvent invoquées a toutes fins pratiques et à défaut

11. Les critères appliqués pour décréter une «élévation» de score ne sont pas universels. Selon une convention, dès que le score maximal d'une série atteint ou dépasse la valeur 70, on décrète l'élévation, qui s'étend alors au second maximum (pour une double élévation), voire au troisième (pour une triple élévation). Dans une convention plus restrictive, les 1-, 2- ou 3-élévations sont composées obligatoirement de scores débordant 70; dans une approche moins restrictive, les deux scores maximaux forment une double élévation, qu'ils débordent ou non le seuil.

12. Les intitulés d'échelle sont « Déviance psychopathique » pour le numéro 4 et « Schizophrénie » pour le numéro 8.

13. Les scores du MMPI classique sont exprimés sur l'échelle standard dite du « T linéaire » et présentent donc tous une moyenne de 50 et un écart-type de 10. La forme de distribution n'est pas contrôlée, varie vraisemblablement d'une échelle à l'autre et affecte certainement le comportement des maxima. De plus, comme dans la plupart des inventaires de personnalité, les échelles sont mutuellement corrélées à des degrés divers.

d'informations contraires. Dans ce contexte, si la variation se fait au hasard, chacune des 10 échelles peut se classer maximale indépendamment des autres, comme chacune des $\binom{10}{2} = 45$ paires d'échelles, chacun des $\binom{10}{3} = 120$ triplets, etc. Au hasard seul, donc, la paire maximale 4-8 a 1 chance sur 45 d'être trouvée dans chaque dossier. En appliquant la loi binomiale §4.4; aussi, Laurencelle et Dupuis 2000), on peut déterminer la probabilité que le hasard seul produise 8 dossiers ou davantage contenant la paire maximale 4-8, cette probabilité[14] étant d'environ 0,00000039. D'après ce résultat hautement exceptionnel, la présence de la paire 4-8 dans 8 de nos 32 dossiers en constituerait une caractéristique marquante.

11.20 Toutefois, la probabilité trouvée ci-dessus n'est pas une réponse juste: elle indique que, si dans chaque dossier n'importe laquelle des 45 paires de scores était également possible, l'obtention d'au moins 8 paires 4-8 sur 32 serait un résultat remarquable. Pour rendre la réponse plus juste et convaincante, il faut d'abord préciser la question : revenons donc a nos 32 dossiers. Nous en compilons les résultats, nous notons dans chaque cas la paire d'échelles maximales et, en fin de course, nous constatons que la paire maximale *qui revient le plus souvent est* la paire 4-8, comptée 8 fois. Ainsi, ce a quoi nous nous intéressons vraiment, l'objet statistique de notre enquête, n'est pas la paire 4-8 comme telle mais, spécifiquement, la paire maximale *la plus fréquente.* Reprenant les suppositions simplificatrices énoncées plus haut, nous devons maintenant trouver une nouvelle probabilité, celle que la paire maximale la plus fréquemment produite au hasard dans 32 dossiers se produise 8 fois ou davantage.

La réponse exacte a cette question n'est pas explicitée dans les références classiques (David et Barton 1962 ; Johnson, Kotz et Balakrishnan 1997). La variable concernée, le maximum d'une loi multinomiale égalitaire, n'a pas une distribution simple, et son obtention reste très laborieuse: nous présentons une solution possible en appendice (voir aussi les exercices 6.6-6.8). Par contre, l'estimation Monte Carlo est a la portée de tous. L'algorithme esquissé suivant, qui approche la distribution d'une k-nomiale égalitaire, peut s'appliquer.

> *Estimer la distribution h[1..n] du maximum* m *d'une loi multinomiale en* k *cellules égalitaires (ou équiprobables) sollicitée n fois, où* $[n/k] \leq m \leq n$
>
> { Le vecteur Cell[1..k] contient les k cellules de la distribution multinomiale, et *m* = max(Cell[j]) est la variable tabulée dans *h* }

14. C'est une binomiale B(32,1/45), pour laquelle on cherche la probabilité extrême $\Pr\{ x \geq 8 \} = \Sigma^{32}_{k=8} b_k(32,1/45)$, où l'élément de probabilité $b_k(n,\pi)$ réfère à (4.7).

```
[ Initialisation ]  Vider h[1..n] ;
[ Cycle          ]  Exécuter T fois
                        Vider Cell[1..k] ;
[  Réaliser la   ]      Pour i = 1 jusqu'à n  Faire
[  k-nomiale     ]        Bloc  j ← 1 + ⌊RND × k⌋ ;
                                 Cell[j] ← Cell[j] + 1
                          Fin_Bloc
[  Repérer m     ]      m ← 1 ;
                        Pour j = 1 jusqu'à k  Faire
                             Si Cell[j] > m alors m ← Cell[j]
[  Compter m     ]      h[m] ← h[m] + 1
[Calculs finals  ]  Pour m = 1 jusqu'à n  Faire p(m) ≈ h[m]/T .
```

Appliquant l'algorithme ci-dessus à notre exemple, donc avec $k = 45$ (paires possibles), $n = 32$ (dossiers MMPI, ou patients) et T = 10 000 itérations, nous obtenons les valeurs approchées suivantes :

$$p[1] \approx 0{,}000 \; ; \; p[2] \approx 0{,}156 \; ; \; p[3] \approx 0{,}618 \; ;$$
$$p[4] \approx 0{,}193 \; ; \; p[5] \approx 0{,}029 \; ; \; p[6] \approx 0{,}002 \; ,$$

ainsi que $p[m] \approx 0{,}000$ pour $m \geq 7$. Les valeurs Monte Carlo obtenues, précises a $\pm 0{,}005$[15] ou mieux, sont très proches des valeurs exactes, données en appendice, et restent beaucoup plus faciles a obtenir. Ainsi, l'obtention au hasard de 8 paires maximales ou plus, sur 45 paires possibles, a une probabilité infime (< 0,001). Sous réserve des suppositions invoquées plus haut, le « profil 4-8 » apparaît comme une caractéristique spécifique des délinquants sexuels juvéniles[16]. L'exercice 9.11 offre une illustration tant soit peu différente d'un problème de maximum.

15. Il s'agit d'estimations p_l de probabilités π_l, où $0 \leq p_l, \pi_l \leq 1$, dont l'erreur-type (O) est $\sqrt{[\pi(1-\pi)/T]}$, l'erreur maximale, de $0{,}5/\sqrt{T}$, advenant pour $\pi = \frac{1}{2}$.

16. M. Michel Parisien, Ph.D., psychologue clinicien, a retrouvé un profil caractéristique semblable chez les délinquants juvéniles de tout acabit, de sorte que le « profil 4-8 » ne peut pas être retenu comme un signe distinctif associé à la délinquance sexuelle (Communication personnelle, 1992).

Exercices

11.1 Estimer l'espérance de $U_{(3:9)}$ en appliquant la technique de la covariable à espérance connue, avec la covariable ½ ($U_{(3:9)}$+$U_{(7:9)}$), d'espérance Trouver par (10.9) la valeur b optimale, et montrer qu'en l'utilisant, l'estimateur ainsi construit est identique à celui basé sur l'antivariable $U_{(7:9)}$ en § 11.5.

11.2 Faire l'analyse formelle du coût d'exécution de l'algorithme par changement de variable en §11.10.

11.3 Réunir et compléter les analyses formelles de coût d'exécution des différents algorithmes présentés au tableau 1. Comparer les indices d'efficacité relative (9.10) établis par analyse formelle et par chronométrage.

11.4 Concevoir et rédiger un programme d'évaluation Monte Carlo du test t de différence entre deux moyennes indépendantes (Howell 1998; Laurencelle 1998a). Déterminer à partir de quelles tailles (n_1, n_2) de séries la solution Monte Carlo (ou par combinatoire approximative), qui consiste à échantillonner au hasard T' arrangements dans l'ensemble des T arrangements différents des données, est à préférer à la solution par combinatoire exhaustive (exercice 8.24), qui traverse cet ensemble.

11.5 Concevoir et rédiger un programme d'évaluation Monte Carlo du test t de différence entre deux moyennes jumelées (Howell 1998; Laurencelle 1998a). Comme à l'exercice précédent, déterminer la taille d'échantillon n à partir de laquelle la solution par combinatoire approximative est préférable à l'autre.

11.6 Soit un problème de corrélation, dans lequel on observe une série statistique bivariée, { (x_1,y_1) (x_2,y_2) (x_n,y_n) } et on calcule le coefficient de corrélation r (cf. (5.3)). Concevoir et programmer deux algorithmes de test par Monte Carlo afin de décider de l'hypothèse nulle « H_0 : E(r) = 0 », le premier par combinatoire (ou permutation de la série y_i), le second par la méthode du Bootstrap (cf. §8.7). Clarifier et comparer les principes logiques des deux algorithmes ; contextualiser leur validité respective.

Appendice

Analyse de variance A x B_R par Monte Carlo : programme en langage QBASIC

L'analyse de variance consiste sommairement à partager la somme de variation totale d'un tableau de données (sa variance) en autant de parties qu'en permet la structure du tableau. Or, le critère permutationnel consiste, en conservant les mêmes données, à les permuter diversement selon l'hypothèse nulle envisagée. Par conséquent, dans le plan de données A x B_R indiqué au tableau 3, la variation totale (SC_{Totale}), la variation Inter-sujets (SC_1) et, par soustraction, la variation Intrasujet (SC_2, ou SC_{Totale}—SC_1), sont constantes. Seules, les sommes indiquées SC_A, SC_B et SC_{AB} doivent être recalculées à chaque cycle de permutations, les autres (SC_{I-G}, SC_{BS}) pouvant être obtenues par soustraction.

Noter enfin que, dans la solution Monte Carlo, il est arbitraire d'exploiter le quotient *F* (F_A, F_B, F_{AB}) ou son seul numérateur (CM_A, CM_B, CM_{AB}), ces deux critères ayant pratiquement la même distribution ordinale et des jeux de centiles correspondants.

```
REM  Analyse de variance A x B(R) en solution Monte Carlo, avec
REM   T ré-arrangements aléatoires
REM  IM groupes de N(i) sujets chacun, chaque sujet étant mesuré JM fois.
REM  Les données dans les structures DONNEES et GROUPE sont
REM   supposées présentes (cf. §11.16).
REM  De plus, IM (<=20), JM (<=20) sont supposées connues,
REM   de même que SM (<= 500), le nombre total de sujets.
REM
DIM DONNEES(500, 20), GROUPE(500), N(20), AXB(20, 20), A(20), B(20)
' Nombre de ré-arrangements aléatoires requis
T = 10000
'...
' Lire les données, tel qu'en §11.16
' DONNEES(·, ·) contient SM lignes et JM colonnes,
' GROUPE(·) contient SM cellules
'...
REM
REM  Étape 1 : Calculs préliminaires (C, SC1, SC2)
ST1 = 0: ST2 = 0: SS2 = 0
FOR i = 1 TO SM
SS = 0
FOR j = 1 TO JM: X = DONNEES(i, j)
```

```
ST1 = ST1 + X: ST2 = ST2 + X * X: SS = SS + X
NEXT: SS2 = SS2 + SS * SS
NEXT
C = ST1 * ST1 / (JM * SM)
SC1 = SS2 / JM - C
SC2 = ST2 - SS2 / JM
REM
REM  Étape 2 : Analyse de l'arrangement observé et obtention des mesures
REM            à tester
GOSUB ANOVA
FAobs = FA: FBobs = FB: FABobs = FAB
' Mettre les compteurs de rangs centiles à 1 (incluent les valeurs F observées)
NA = 1: NB = 1: NAB = 1
REM
REM  Étape 3 : Cycle de ré-arrangements aléatoires, avec analyse
REM            et obtention des mesures de la « distribution nulle »
' « Brouiller » le générateur de nombres aléatoires uniformes
RANDOMIZE TIMER
FOR iii = 1 TO T - 1
' Répartir les sujets au hasard à travers les IM groupes
FOR i = 1 TO SM - N(IM)
g = i + INT((SM + 1 - i) * RND): SWAP GROUPE(i), GROUPE(g)
NEXT
' Pour chaque sujet, permuter au hasard ses JM données
FOR S = 1 TO SM
FOR j = 1 TO JM - 1
h = j + INT((JM+1-j)*RND): SWAP DONNEES(S,j), DONNEES(S,h)
NEXT: NEXT
' Recalculer les mesures de l'analyse de variance et comparer
GOSUB ANOVA
IF FAobs < FA THEN NA = NA + 1
IF FBobs < FB THEN NB = NB + 1
IF FABobs < FAB THEN NAB = NAB + 1
NEXT
REM
REM  Étape 4 : Calculs finals et impression des résultats
PRINT " ANOVA AxBr basée sur"; T; " ré-arrangements aléatoires."
PRINT USING " F(A) = ###.##  dl = ## et ##  Pr = #.###"; FAobs; IM - 1;
  SM - IM; NA / T
PRINT USING " F(B) = ###.##  dl = ## et ##  Pr = #.###"; FBobs; JM - 1;
  (SM - IM) * (JM - 1); NB / T
PRINT USING "F(AxB)= ###.##  dl = ## et ##  Pr = #.###"; FABobs; (IM - 1) *
  (JM - 1); (SM - IM) * (JM - 1); NAB / T
END
'
```

```
ANOVA:
' Compléter l'analyse de la configuration actuelle des données.
' Le numéro du groupe auquel appartiennent les données de chaque sujet
'  (dans une ligne) est indiqué par le vecteur GROUPE.
'
FOR i = 1 TO IM: FOR j = 1 TO JM: AXB(i,j) = 0: NEXT: NEXT
FOR i = 1 TO SM: g = GROUPE(i)
FOR j = 1 TO JM: AXB(g,j) = AXB(g,j) + DONNEES(i,j): NEXT
NEXT
SAB2 = 0: SA2 = 0: SB2 = 0
FOR i = 1 TO IM: SA1 = 0: FOR j = 1 TO JM: X = AXB(i,j)
SAB2 = SAB2+X*X / N(i): SA1 = SA1 + X: NEXT: SA2 = SA2+SA1*SA1 / N(i)
NEXT
FOR j = 1 TO JM: SB1 = 0: FOR i = 1 TO IM: SB1 = SB1 + AXB(i,j): NEXT
SB2 = SB2 + SB1 * SB1: NEXT
SCA = SA2 / JM – C
SCB = SB2 / SM – C
SCAB = SAB2 – C – SCA – SCB
FA = SCA / (SC1 – SCA) * (SM – IM) / (IM – 1)
FB = SCB / (SC2 – SCB – SCAB) * (SM – IM)
FAB = SCAB / (SC2 – SCB – SCAB) * (SM – IM) / (IM – 1)
RETURN
```

Distribution du maximum d'une multinomiale égalitaire

L'auteur a élaboré une méthode de calcul basée sur les *partitions* d'un entier. Posons n objets à répartir au hasard dans k cellules, k^n le nombre total d'arrangements possibles des objets dans les cellules, et $f(m)$ le nombre d'arrangements tels que le nombre maximal dans les cellules est m, $[n/k] \leq m \leq n$. La distribution de probabilité de m, $p(m)$, est donnée par:

$$f(m) = k^n \cdot p(m) = \sum_{p \in P} \left(A_{k,L_p} \times R_{p,L_p} \times M_{n,p} \right) ,$$

la sommation se faisant pour l'ensemble P des partitions $p = p(n\backslash k,m)$ idoines.

> Une partition $p(n\backslash k,m)$ est une décomposition unique, non croissante, du nombre n en k parties ou moins, ayant comme premier terme la valeur m (voir L. Laurencelle, « L'interprétation stochastique du sociogramme et le problème des choix exceptionnels », dans *Lettres statistiques,* vol. 9, 1993, p. 115-133, *ou R.C.* Bose et B. Manvel, *Introduction to combinatorial theory,* Wiley, 1984) *;* dans notre notation, P dénote l'ensemble (la liste) des partitions adéquates. Ainsi, pour $n = 10$, $k = 4$, $m = 6$, les différentes partitions sont (6 4), (6 3

1), (6 2 2), (6 2 1 1), d'où card(P) = 4 ; enfin, L_P désigne la longueur effective de la partition, où $L_p \leq k$.

La sommation globale, indiquée ci-dessus, concerne donc la liste des partitions p = (c_1 c_2 ...) comprises dans P.

Le facteur A_{k,L_p} égale $C(k,L_p) = k!/[L_p!(k-L_p)!]$ et dénote le nombre d'arrangements différents des L_p cellules occupées parmi les k cellules. Le facteur R_{p,L_p} égale $L_p!/\Pi u_i!$, où u_i est le nombre de cellules contenant i éléments, où $i \geq 1$ et $\sum i \times u_i = n$; R indique le nombre de répartitions distinctes de la partition p dans les L_p cellules désignées. Le facteur multinomial $M_{n,p}$ égale $n!/\Pi c_i!$ et dénote le nombre d'arrangements des n éléments dans la partition.

Pour illustrer la méthode, prenons $k = 45$ et $n = 32$, comme dans l'exemple des paires maximales du MMPI, et m = 29, une valeur qui permet de garder le calcul simple. Les k-partitions de n avec maximum m sont ici (29 3), de longueur $L = 2$; (29 2 1), $L = 3$, et (29 1 1 1), $L = 4$. Pour la première partition, A = C(45,2) = 990, R = 2!/(1 ! 1 !) = 2, M = 32!/(29!3!) = 4960, d'où AxRxM = 9820800 pour la seconde, nous avons AxRxM = 14190x6x14880 = 1266883200 ; enfin, pour la dernière, AxRxM = 148995x4x29760 = 17736364800. Le total, soit $f(29)$ = 19013068800, doit être divisé par $k^n \approx$ 7,995x 10^{52} pour déterminer $p(m) \approx$ 2,378x 10^{-43}, la probabilité que la cellule la plus fréquentée parmi les 45 affiche un nombre de 29 occupants.

Le calcul manuel de ces quantités est fastidieux, voire prohibitif. Quant à sa programmation informatique, elle n'est pas simple. De meilleures approximations que celles parues (p. ex. R.M. Koselka, « Approximate upper percentage points for extreme values in multinomial sampling)>, *Annals of mathematical statistics,* vol. 27, 1956, p. 507-512) seraient bienvenues, et la solution Monte Carlo semble toujours de mise.

L'exercice 6.6 donne des expressions plus simples pour les cas particuliers suivants, soit $p(1)$ et $p(m)$, $m > n/2$.

Voici, à des fins de comparaison, les probabilités exactes, arrondies à 5 décimales, qui s'appliquent à notre exemple d'une multinomiale à (k =) 45 catégories égales, échantillonnée (n =) 32 fois:

$p(1) = 0{,}00000$; $p(2) = 0{,}15640$; $p(3) = 0{,}62130$; $p(4) = 0{,}19275$

$p(5) = 0{,}02656$; $p(6) = 0{,}00274$; $p(7) = 0{,}00023$; $p(8) = 0{,}00002$

et $\sum p(m) < 0{,}000005$ pour $m \geq 9$.

Références

CATTELL, R.B., CATTELL, A.K., CATTELL, H. (1993). *Sixteen personality factors questionnaire* (5e édition). Champaign (IL), Institute for Personality and Ability Testing.

DAVID, F.N., BARTON, D.E. (1962). *Combinatorial chance.* London, Charles Griffin.

DAVID, H.A. (1981). *Order statistics.* New York, Wiley.

DIACONIS, P., EFRON, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248,* 116-130.

EDGINGTON, E.S. (1980). *Randomization tests.* New York, Marcel Dekker.

FISHMAN, G.S. (1996). *Monte Carlo: concepts, algorithms, and applications.* New York, Springer.

GENTLE, J.E. (1998). *Random number generation and Monte Carlo methods.* New York, Springer-Verlag.

GORDON, G. (1969). *System simulation.* Englewood Cliffs (NJ), Prentice-Hall.

HATHAWAY, S.R., MCKINLEY, J.C. (1996). *Inventaire multiphasique de personnalité du Minnesota-2.* Paris, Les éditions du Centre de Psychologie Appliquée.

HOEFFDING, W. (1952). The large-sample power *of* tests based on permutations *of* observations. *Annals of mathematical statistics,* 23, 169-192.

HOWELL, D.C. (1998). *Méthodes statistiques en sciences humaines.* Bruxelles, De Boeck.

JOHNSON, N.L., KOTZ, S., BALAKRISHNAN , N. (1997). *Discrete multivariate distributions.New* York, Wiley.

KALOS, M.A., WHITLOCK, P.A. (1986). *Monte Carlo methods. Vol.* 1: *Basics.* New York, Wiley.

KIRK, R.E. (1982). *Experimental design: Procedures for the behavioral sciences* (2e édition). Belmont (CA), Brooks/Cole.

KNUTH, D.E. (1973). *The art of computer programming.* Vol. 3*: Sorting and searching.* Reading (MA), Addison-Wesley.

LAURENCELLE, L. (1987). Le nombre de permutations dans les tests permutationnels. *Lettres Statistiques, 8,* 49-80.

LAURENCELLE, L. (1993). La loi uniforme: propriétés et applications. *Lettres statistiques, 9,* 1-23.

LAURENCELLE, L. (1998a). Les analyses statistiques, dans S. Bouchard et C. Cyr (dir.): *Recherche psychosociale: pour harmoniser recherche et pratique* (chap. 9, p. 345-388). Sainte-Foy, Presses de l'Université du Québec.

LAURENCELLE, L. (1998b). Perspectives d'une approche multivariée en psychométrie, les normes multivariées et le différentiel de sélection multiple. Communication au *66' Congrès de l 'ACFAS,* Mai, Université Laval, Québec.

LAURENCELLE, L., DUPUIS, F.A. (2000). *Tables statistiques expliquées et appliquées* (2e édition). Sainte-Foy, Le Griffon d'argile.

LURIE, D., HARTLEY, H.O. (1972). Machine-generation of order statistics for Monte Carlo computations. *The American Statistician, 26,* 26-27.

NAYLOR, T.J. (1971). *Computer simulation experiments with models of economic systems.* New York, Wiley.

PARISIEN, M., LAURENCELLE, L. (1990). L'évaluation des agresseurs sexuels juvéniles. 1: L'utilisation du MMPI. Communication à la *Société de Criminologie du Québec,* Mars, Institut Philippe Pinel de Montréal.

ROBERT, C.P., CASELLA, G. (1999). *Statistical Monte Carlo methods.* New York, Springer.

SOBOL, I.M. (1974). *The Monte Carlo method.* Chicago, The University of Chicago Press.

SOKAL, R.R., ROHLF, F.J. (1981). *Biometry* (2e édition). San Francisco, Freeman.

WINER, B.J., BROWN, D.R., MICHELS, K.M. (1991). *Statistical principles in experimental design* (3e édition). New York, McGraw-Hill.

Index

L'indication

§4.2 renvoie au paragraphe (ou section) 4.2 du chapitre 4;
(4.2) renvoie à l'expression mathématique (4.2) du chapitre 4;
ϵ4.2 renvoie à l'exercice 4.2, aussi du chapitre 4.

MEMBRE DU GROUPE SCABRINI

Québec, Canada
2001