Collection

Statistique et probabilités appliquées

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Yadolah Dodge, Giuseppe Melfi

Premiers pas en simulation

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{$$

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2}{n_1 - 1} \quad \text{et}$$

$$S_1 = \sqrt{S_1^2} \quad \text{et}$$

$$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1} \quad \text{et}$$

Springer

Premiers pas en Simulation

Yadolah Dodge et Giuseppe Melfi

15 mai 2008

Préface

La simulation a une grande importance dans le monde d'aujourd'hui. Le développement technologique dans les domaines les plus disparates demande souvent des simulations à grande échelle qui se révèlent essentielles pour la conception de projets ou la mise en place de stratégies d'action. C'est pourquoi elle est enseignée dans les plus prestigieuses universités et écoles polytechniques de la planète.

Toutefois, nous nous sommes aperçus que la plupart des ouvrages disponibles, que ce soit en anglais ou en français, s'adressent à un public de spécialistes, et supposent que le lecteur possède un bagage de connaissances en statistique mathématique déjà bien développé. Bref, il n'est pas difficile d'imaginer qu'un lecteur peu habitué à un langage mathématique trouve la plupart de ces excellents ouvrages de lecture difficile, et qu'après avoir passablement peiné dans leur lecture il se décourage pour ne pas dire qu'il abandonne, en jugeant la matière et son exposition trop technique, voire incompréhensible.

Ce livre est le fruit d'années d'enseignement dans le deuxième cycle à la faculté de sciences économiques de l'université de Neuchâtel. Il s'adresse à un public de non-spécialistes, et, pour un étudiant ou un doctorant qui veut s'initier aux techniques de simulation, ce livre avec son langage simple et par son contenu de base qui le rend quasi autosuffisant peut être une introduction à la simulation et aux méthodes de Monte Carlo.

Une brève introduction à la probabilité peut être utile pour ceux qui auraient besoin de quelques rappels. Avec une présentation de concepts fondamentaux on permet aussi la lecture à un public d'informaticiens, d'ingénieurs, ou de mathématiciens qui n'ont pas tous nécessairement des bases en statistique. Le lecteur est ensuite guidé à travers des exemples où les différentes techniques sont appliquées.

Chaque chapitre est suivi d'un certain nombre d'exercices, dont certains demandent l'utilisation d'un logiciel pour le calcul statistique, ce qui pourra être apprécié par la majorité du public visé.

La bibliographie à la fin de l'ouvrage énumère les travaux qui, au cours du XX^{e} siècle, ont permis l'essor de la simulation. Elle répertorie aussi d'autres contributions scientifiques consacrées à des aspects plus détaillés de la théorie.

Neuchâtel, le 2 mars 2008

Yadolah Dodge
Giuseppe Melfi

Table des matières

Chapitre 1

Introduction

1.1 Pourquoi des techniques de simulation ?

Les méthodes de simulation, conçues pour être utilisées en statistique et en recherche opérationnelle, ont connu et connaissent encore un développement rapide dû à l'extraordinaire évolution des ordinateurs. Des applications se rencontrent tant dans l'industrie qu'en économie, ou encore en sciences sociales, en physique des particules, en astronomie et dans de nombreux autres domaines.

Dans beaucoup de situations, que ce soit de la vie courante ou dans la recherche scientifique, le chercheur est confronté à des problèmes dont il recherche des solutions sur la base de certaines hypothèses et contraintes de départ. Pour résoudre ce type de problème, il existe des méthodes analytiques applicables à des situations où le modèle permet de traiter les différentes variables par des équations mathématiquement maniables, et des méthodes numériques où la complexité du modèle impose un morcellement du problème, notamment par l'identification des différentes variables qui entrent en jeu et l'étude de leurs interactions. Cette dernière approche s'accompagne souvent d'une importante masse de calculs. Les techniques de simulation sont des techniques numériques : simuler un phénomène signifie essentiellement reconstituer de façon fictive son évolution.

Après un bref aperçu historique sur l'évolution du concept de hasard et de son usage à travers les siècles, le Paragraphe 1.3 présente la problématique de la simulation est présentée dans toute sa généralité sous ses différents aspects. Les paragraphes suivants montrent des exemples où des techniques de simulation peuvent être appliquées.

La simulation utilise des nombres aléatoires, et les méthodes pour en générer seront abordées au Chapitre 3. Cependant, nous utiliserons déjà des nombres aléatoires dans ce chapitre. Disons simplement pour l'instant que, en tirant un billet d'un chapeau contenant des billets numérotés de 0 à 9, en le remettant dans le chapeau, en mélangeant les 10 billets puis en répétant n fois cette procédure, on génère n nombres aléatoires. En faisant la liste des n nombres

tirés, arrangés par groupes de taille fixée pour en faciliter la lecture, on obtient une table de nombres aléatoires. Notons néanmoins que pour construire des tables contenant des milliers de nombres aléatoires, on utilise des systèmes mécaniques ou des algorithmes bien plus élaborés que cela (voir Chapitre 3).

1.2 Une brève histoire de la notion de « hasard »

Le hasard et la chance jouent un rôle central dans la vie de tous les jours et dans une large palette de domaines. La production agricole est une fonction des conditions météorologiques ; le statut économique d'un individu peut être le résultat de son habileté dans le commerce et de son habileté à tisser des bons réseaux sociaux ainsi que de facteurs extérieurs que l'on qualifie généralement de conjoncture économique. Le hasard, comme on l'entend aujourd'hui, a été pendant longtemps une idée abstraite et, pendant des siècles, dans le sens commun largement liée à une explication métaphysique, le destin, ou à une volonté divine.

Historiquement, la première fois que l'humanité a eu à faire avec des nombres aléatoires fut probablement pour des activités divinatoires. Des dés rituels étaient en usage déjà à l'âge du bronze. Le plus ancien dé de forme cubique avec des faces numérotées de 1 à 6 remonte à 2000 ans avant Jésus-Christ et a été trouvé en Égypte. Des dés du VII^{e} et du VI^{e} siècle avant Jésus-Christ ont été trouvés en Italie centrale et en Chine.

Il n'y a pas de doute que le jeu du lancer des dés est l'un des plus anciens jeux de l'humanité. L'empereur Claude écrivit un livre sur l'art de gagner aux jeux du lancer des dés et, dans le courant du XVII^{e} siècle, Chevalier de Méré, un riche parieur français, était en correspondance avec Fermat et Pascal au sujet de nombre de problèmes concernant des combinaisons gagnantes et les paris correspondants dans le jeu des dés.

C'est seulement au cours du XX^{e} siècle, avec le développement parallèle des sciences statistiques et de la technologie, que les nombres aléatoires ont trouvé une large application. La célèbre phrase d'Einstein « *Dieu ne joue pas aux dés* » est contemporaine à l'élaboration du principe d'incertitude de Heisenberg. Selon l'idée classique de causalité, pour prédire le futur avec un certain degré de précision, il suffisait de connaître le présent avec suffisamment de précision. Heisenberg démontra la non-prédicibilité des événements en physique quantique. La non-prédicibilité est présente partout. Dans des instituts de prévision météorologique les quantités d'intérêt sont estimées à partir d'un grand nombre d'états initiaux, et l'évolution du phénomène est artificiellement simulée avec des modèles numériques utilisant des nombres aléatoires. Et bien que les capacités de calcul des ordinateurs des centres de prévision météorologique soient très élevées, les prévisions ne vont, hélas, jamais au-delà de quelques jours.

De telles techniques constituent ce que l'on appelle généralement la *méthode de Monte Carlo*, et aujourd'hui une vaste littérature sur les applications les plus variées est disponible.

À la base de toute simulation, il y a l'utilisation de nombres aléatoires en grande quantité. De plus, pour qu'une simulation soit fiable il faut que les nombres aléatoires utilisés aient toutes les propriétés que l'on attend. Ainsi, il ne suffit pas de disposer d'une liste finie de 100 ou même d'un million de nombres aléatoires et de l'utiliser en boucle pour des simulations. En bref, la production de nombres aléatoires en grandes quantités n'est pas une simple affaire.

L'exigence d'utiliser des nombres aléatoires dans la science s'est manifestée au début du siècle passé. En 1927, une liste de 41 600 nombres aléatoires pour usage scientifique, produite par Leonard Tippett, a été publiée par Cambridge University Press. Ensuite, La fondation RAND, en 1955, publia *A Million Random Digits with 100,000 Normal Deviates*, sur la base d'une simulation par ordinateur avec un algorithme qui aujourd'hui est considéré comme dépassé. Déjà à l'époque, certains soulevaient de sérieux doutes sur la possibilité de produire des nombres vraiment aléatoires de façon automatique. Neumann, en 1951, remarquait justement que par leur propre nature il ne peut pas exister une méthode algébrique capable de produire des nombres aléatoires. Cela montre bien que la production automatique de nombres aléatoires a été un sujet controversé. Des suites décimales de nombres normaux spéciaux comme π vus comme suite aléatoire de chiffres décimaux ont aussi été proposées (Dodge, 1996). En effet, Borel en 1909 démontra qu'un nombre réel pris au hasard sur l'intervalle $[0, 1]$ est normal avec probabilité 1, c'est-à-dire que toutes les différentes séquences finies de chiffres (sous une base fixée) apparaissent selon une distribution de fréquences uniforme, en faisant d'un nombre normal un bon candidat pour fournir ainsi une suite de nombres aléatoires.

L'histoire de la génération des nombres aléatoires commence avec des machines plus ou moins complexes dont le but était de piocher des boules numérotées d'une urne. Encore aujourd'hui, en dépit d'algorithmes performants et de qualité élevée pour la génération de nombres aléatoires en grandes quantités, de telles machines sont utilisées pour les loteries à numéros et les ordinateurs ne les remplaceront-ils probablement jamais. Il y a une raison philosophique à cela, celle qui est au fond soulevée par Neumann : un algorithme implémenté produira une suite de nombres dont la nature est déterministe, et donc d'une certaine manière prévisible, et la suite aura seulement l'apparence d'être aléatoire.

1.3 Systèmes, modèles et méthodes de résolution

Un *système* est un ensemble d'éléments que l'on peut appeler *composantes*. Chacun de ces éléments possède plusieurs caractéristiques ou attributs qui peuvent prendre des valeurs numériques ou logiques. Par exemple, une installation industrielle peut être considérée comme un système, dont les composantes sont les machines et les ouvriers ; le fait qu'une machine fonctionne ou non est une caractéristique de ce système. Une économie nationale, composée de

ses consommateurs et de ses producteurs, est également un système ; l'un des attributs d'un consommateur peut être l'importance de sa demande pour un produit particulier.

Les composantes d'un système sont interactives. Par exemple, sans opérateur, la machine ne peut pas fonctionner. À côté de ces relations, appelées internes, figurent des relations dites externes. Ces dernières relient les éléments du système avec l'environnement, c'est-à-dire le monde en dehors du système.

Un *modèle* peut être défini comme une architecture mathématique qui représente un certain système. Il existe plusieurs types de modèles.

Un modèle est constitué de symboles mathématiques représentant des systèmes réels. Autrement dit, le modèle mathématique d'un système est l'ensemble des relations mathématiques caractérisant les états possibles du système.

Les problèmes de simulation peuvent être classés en deux grandes catégories : les problèmes *déterministes* et les problèmes *probabilistes*. Les problèmes déterministes sont ceux pour lesquels l'incertitude est soit négligeable, soit entièrement absente et qui comprennent des phénomènes physiques simples comme la chute libre d'un objet ou un mouvement uniforme. Les problèmes probabilistes comprennent, par exemple, le calcul du nombre optimal de distributeurs de billets, du nombre optimal de guichets et, en général, tout autre phénomène dont le déroulement dépend d'une part de hasard. Dans la réalité, des problèmes complètement déterministes sont assez rares. La plupart du temps, des petites erreurs ou incertitudes sont négligées ou délibérément ignorées : les coûts manufacturiers sont habituellement estimés plutôt que connus ; les instruments de mesure, qui ont une précision de 99,5 %, sont considérés comme parfaits, etc. Dans de telles circonstances, l'emploi d'un modèle déterministe se justifie seulement si l'on s'attend à ce que les écarts dans la pratique soient à la fois rares et petits.

Les problèmes de simulation probabilistes ou stochastiques comprennent eux un degré d'incertitude trop important pour être ignoré. Par exemple, sous certaines hypothèses au sujet de l'arrivée aléatoire de passagers à un distributeur de billets, le nombre de personnes faisant la queue peut être en moyenne de 5, mais il peut monter jusqu'à 50, une ou deux fois par jour. Il serait vraiment erroné, dans ce cas, de calculer le nombre optimal de distributeurs de billets sur la base d'un modèle déterministe qui utilise seulement la moyenne de 5 et néglige les cas occasionnels. Pour représenter une incertitude de cette sorte, des modèles d'optimisation stochastiques utilisent des variables aléatoires dont les valeurs sont données par des distributions de probabilité plutôt que par de simples nombres ou équations. Les problèmes stochastiques sont généralement beaucoup plus complexes et plus difficiles à résoudre que les problèmes déterministes.

En économie, on classe habituellement les variables d'un modèle en *variables exogènes* et *variables endogènes*. Les valeurs des variables exogènes ne sont pas déterminées par le modèle. On les appelle aussi variables indépendantes. Dans la terminologie des systèmes, les variables indépendantes peuvent encore

être divisées en variables incontrôlables et variables contrôlables. Les variables incontrôlables, par exemple la demande étrangère, sont les inputs du système. Les variables contrôlables, par exemple les dépenses du gouvernement, sont des variables manipulées par certaines composantes du système.

Les valeurs des variables endogènes dépendent du modèle. Dans la terminologie des systèmes, on peut classer ces variables dépendantes en *variables d'état* ou *intermédiaires* décrivant l'état du système, et en *variables de sortie*.

À côté des variables, nous distinguons, dans le modèle, des *paramètres*. Les paramètres sont des quantités qui influencent les variables endogènes. Cependant, contrairement aux variables, ils sont constants. Par exemple un paramètre peut être lié au critère selon lequel un gestionnaire de stock cadence les nouvelles commandes. Les *relations* indiquent comment les variables et les paramètres sont reliés entre eux.

Pour répondre aux questions relatives à un phénomène étudié, il faut souvent résoudre les équations du modèle. Il existe des méthodes de résolution *analytiques* et des méthodes de résolution *numériques*.

La méthode analytique fait appel au calcul différentiel et intégral. Elle fournit une solution générale sous la forme d'une équation ou d'une formule valable pour différentes valeurs possibles des variables indépendantes et des paramètres. Toutefois, le champ des problèmes qui peuvent être résolus mathématiquement est limité. En effet, les problèmes que l'on rencontre dans la pratique nécessitent que le modèle utilisé soit exprimé sous une forme particulière d'un système d'équations algébriques ou différentielles pouvant être très complexes suivant la nature du phénomène ou du système à étudier. Or, le degré de complexité de ce phénomène peut exiger de l'analyste une simplification parfois abusive du modèle, dans le but de s'adapter aux techniques mathématiques disponibles.

La résolution numérique remplace les variables indépendantes et les paramètres du modèle par des nombres qu'elle manipule. Beaucoup de techniques numériques sont itératives, c'est-à-dire que chaque étape de la résolution donne une meilleure solution que la précédente en utilisant les résultats des étapes antérieures. La *programmation mathématique* est une technique numérique de ce genre.

La *simulation* et la *méthode de Monte Carlo* sont des techniques numériques spécifiques. La méthode de Monte Carlo est une technique de résolution d'un problème qui utilise des échantillons de nombres aléatoires dans le modèle qui le décrit. Nous reviendrons sur cette méthode dans le Chapitre 6. Quant à la simulation, elle représente une expérience dans le temps faite sur un modèle abstrait et impliquant la présence de variables aléatoires. Dans la plupart des cas, les études de simulation utilisent des nombres aléatoires.

Dans le contexte décrit ci-dessus, la simulation est une expérience qui suppose la construction d'un modèle de travail mathématique présentant une similitude de propriétés ou de relations avec le système naturel faisant l'objet de l'étude. De cette façon, nous pouvons prévoir les caractéristiques de fonctionnement de ce système sans avoir à travailler avec des dispositifs physiques. Il s'agit d'effectuer, à l'aide du modèle désigné, des *expériences artificielles* per-

mettant de restituer des valeurs pour certaines variables qui soient conformes aux lois de probabilité observées dans un cas réel. Ces valeurs constituent un *échantillon artificiel*.

Prenons l'exemple de la simulation d'un phénomène de file d'attente aux caisses d'un grand magasin. Supposons que le but de l'étude soit de connaître le meilleur « système de caisse », c'est-à-dire celui pour lequel la somme des coûts d'inactivité des caisses et des coûts d'attente des clients soit la plus faible. L'évolution du phénomène dépend essentiellement de la loi des arrivées des clients aux caisses et de la loi des temps de service. La connaissance de ces lois permet de décrire cette évolution et de calculer ainsi le coût d'un système. Il suffit alors de répéter l'expérience pour différents systèmes et de choisir le meilleur. La connaissance des lois est en général empirique : elle résulte d'une étude statistique à partir de laquelle on peut déterminer les lois de probabilité des variables aléatoires caractérisant le phénomène. Le problème fondamental de la simulation sera par conséquent de construire ces échantillons artificiels relatifs à des variables de lois connues statistiquement. Dans notre exemple, il s'agit de simuler un échantillon d'arrivée de clients et un échantillon des temps de service. Cette construction se fait à l'aide de *générateurs de nombres aléatoires*, sujet faisant l'objet du Chapitre 3.

L'inconvénient de la simulation par rapport à une solution analytique est qu'elle ne fournit que des solutions spécifiques à un problème donné, et non pas des solutions générales. Dans un but pratique, il convient, avant d'en venir aux techniques de simulation, de considérer l'ensemble des techniques mathématiques disponibles pour résoudre le problème. Dans ce sens, l'emploi de la simulation peut se révéler utile et justifié s'il propose une extension ou un complément aux solutions analytiques obtenues au prix d'une trop grande simplification.

Nous allons voir maintenant trois exemples de problèmes qui peuvent être résolus par la simulation.

1.4 Un phénomène de file d'attente

Faire la queue ! Voilà une chose dont quiconque a malheureusement déjà fait l'expérience. Pour acheter un timbre à la poste, pour se faire enregistrer à l'aéroport, il faut attendre son tour.

De manière générale, nous pouvons définir un phénomène de file d'attente par les faits suivants : chaque fois qu'un certain nombre d'unités que nous appellerons *clients* se présente de manière aléatoire, afin de recevoir un *service* d'une *durée aléatoire* de la part d'autres unités que nous appellerons *stations*, on est en présence d'une file d'attente.

Considérons l'exemple d'une file d'attente aux caisses d'un grand magasin. Le client pourra alors se poser plusieurs questions :

(a) combien de temps va-t-il attendre en moyenne dans la queue ?

(b) quelle probabilité a-t-il d'attendre plus d'un temps t ?

(c) combien de clients va-t-il trouver devant lui ?

Si nous essayions de répondre aux questions ci-dessus de manière analytique, il nous serait impossible de tenir compte du fait que les clients arrivent de façon aléatoire au magasin. Nous serions obligés de considérer que les clients arrivent de manière régulière, ou au moins à des moments connus, et que les temps de service soient eux aussi connus.

Or il est certain qu'il y a beaucoup plus de clients aux heures de pointe qu'aux heures creuses de la journée. Les moments exacts où les clients arrivent et les temps de service sont tout sauf connus à l'avance. La prise en compte des différents aléas n'entre guère dans un modèle déterministe.

Pour pouvoir résoudre un tel problème, nous avons besoin de différentes notions de statistique et de probabilités que nous verrons au Chapitre 2. Un exemple complet de file d'attente sera traité au Chapitre 6.

1.5 Un problème de gestion

Dans l'exemple qui suit, on veut simuler l'état, minute par minute, du nombre d'avions supplémentaires (ou d'avions en moins) au sol à l'aéroport de Genève-Cointrin entre 17 h 00 et 17 h 18 par rapport à leur nombre à 17 h 00. Pour cela, il faut connaître les lois de probabilités du nombre d'arrivées et du nombre de départs par minute. Cela pour estimer toute une série de variables utiles à la gestion d'un aéroport comme le nombre maximal d'avions présents au sol. Supposons que les registres de l'aéroport permettent d'affirmer que ces deux lois sont identiques. Elles sont présentées ci-dessous :

Nombre d'arrivées (départs)	0	1	2	3	4	5
Probabilité du nombre d'arrivées (départs)	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{3}{125}$	$\frac{1}{1\,000}$

Tab. 1.1. *Distribution des probabilités du nombre d'arrivées, égale à celle pour les départs.*

À partir de ces lois, nous allons simuler les mouvements d'avions minute par minute.

Si l'on ne possède pas de générateurs de nombres issus des lois de probabilités des arrivées et des départs, on peut utiliser une table de nombres aléatoires comme celle de la Table 1.2.

7	7	1	0	0	4	9	1	4	2	4	3
7	9	6	8	9	2	0	8	6	3	8	8
9	5	1	9	0	4	8	4	7	0	2	5
4	7	5	9	7	3	4	2	1	8	2	5
7	9	9	7	9	8	6	9	1	8	4	1
9	0	2	4	4	1	6	2	9	3	5	5
0	7	0	4	3	6	0	6	1	5	4	2
1	9	3	6	0	8	5	6	6	1	9	9
8	5	1	7	6	4	7	6	1	8	0	5

Tab. 1.2. *Une table de nombres aléatoires.*

Une table de nombres aléatoires se trouve aussi à l'Appendice à la fin de cet ouvrage. On divise la suite des nombres de 000 à 999 proportionnellement aux lois de probabilités des arrivées et des départs :

Nombre aléatoire compris entre :	Nombre d'arrivées	Nombre de départs
000-124	0	0
125-624	1	1
625-874	2	2
875-974	3	3
975-998	4	4
999-999	5	5

Tab. 1.3. *Intervalles de nombres assignés à des valeurs d'une variable aléatoire.*

On procède ensuite au tirage dans la table des nombres aléatoires. Pour cela, on adopte un critère mis au point à l'avance et qui soit apte à bien choisir des nombres aléatoires entre 000 et 999. Dans notre cas, les nombres aléatoires (de 0 à 9) sont utilisés par groupes de trois et placés dans deux colonnes d'un tableau de simulation (voir ci-dessous). Le nombre d'arrivées ou de départs est déterminé par l'intervalle auquel le nombre aléatoire choisi appartient et cela dans l'ordre établi par le critère retenu. Notre critère sera le suivant : le premier numéro tiré correspond au nombre d'arrivées entre 17 h 00 et 17 h 01, le deuxième au nombre de départs durant la même période, et ainsi de suite. Cette technique est décrite dans toute sa généralité au Paragraphe 4.4.

Ainsi, si le premier nombre aléatoire sélectionné est compris entre 625 et 874, le nombre d'arrivées entre 17 h 00 et 17 h 01 sera égal à 2. Il nous reste alors à tirer 35 autres nombres aléatoires pour avoir le nombre d'arrivées et de départs, minute par minute, jusqu'à 17 h 18.

Minutes	Nombres aléatoires		Nombre d'arrivées	Nombre de départs	Nombre d'avions supplémentaires au sol
17 h 00 - 01	771	004	2	0	2
17 h 01 - 02	914	243	3	1	4
17 h 02 - 03	796	892	2	3	3
17 h 03 - 04	086	388	0	1	2
17 h 04 - 05	951	904	3	3	2
17 h 05 - 06	847	025	2	0	4
17 h 06 - 07	475	973	1	3	2
17 h 07 - 08	421	825	1	2	1
17 h 08 - 09	799	798	2	2	1
17 h 09 - 10	691	841	2	2	1
17 h 10 - 11	902	441	3	1	3
17 h 11 - 12	629	355	2	1	4
17 h 12 - 13	070	436	0	1	3
17 h 13 - 14	061	542	0	1	2
17 h 14 - 15	193	608	1	1	2
17 h 15 - 16	566	199	1	1	2
17 h 16 - 17	851	764	2	2	2
17 h 17 - 18	761	805	2	2	2
Total			29	27	

Tab. 1.4. *Un tableau de simulation pour un problème de gestion des départs et des arrivées d'avions.*

Cette simulation, quoique extrêmement simplifée, a donné des résultats d'une certaine utilité pour la gestion d'un aéroport. Ceux-ci pourraient être utilisés pour l'optimisation de la gestion du personnel au sol et pour l'optimisation de l'utilisation des infrastructures aéroportuaires.

1.6 Exemple d'une surface à calculer

La méthode de Monte Carlo peut être très utile aussi dans des problèmes d'estimation de surface. Le principe est simple : supposons que nous voulions estimer la surface d'un carré de côté 0,6. Cette surface est naturellement 0,36. Nous allons montrer comment trouver une estimation de ce résultat par simulation.

Nous reportons le carré sur un système d'axes perpendiculaires dont l'origine est l'angle inférieur gauche du carré (voir Fig. 1.1). Sur le carré $[0, 1] \times [0, 1]$ qui le contient nous allons disposer 40 points aléatoires : certains seront à l'intérieur du carré de côté 0,6. Tous les points que nous allons disposer à l'intérieur du carré auront donc des coordonnées comprises entre 0 et 1. Il suffit donc de prendre deux variables aléatoires uniformes pour obtenir un point. En effet, le premier nombre aléatoire sera l'abscisse et le second l'ordonnée. Nous avons besoin de 2 échantillons de 40 nombres aléatoires pour disposer nos 40 points

à l'intérieur du carré. Il ne nous reste plus qu'à compter le nombre de points qui se trouvent dans S et à calculer le quotient $N^{'}/N$ qui sera une estimation de la surface.

Dans notre exemple, $N' = 15$, donc la surface du carré de côté 0,6 est approximativement égale à $15/40 = 0{,}375$. Cela reste évidemment une approximation puisque dans notre cas la surface réelle est 0,36. L'intérêt de cette méthode réside dans le fait qu'elle peut être appliquée au calcul de surfaces pour lesquelles une formule mathématique n'existe pas, comme c'est le cas pour une surface non régulière en général.

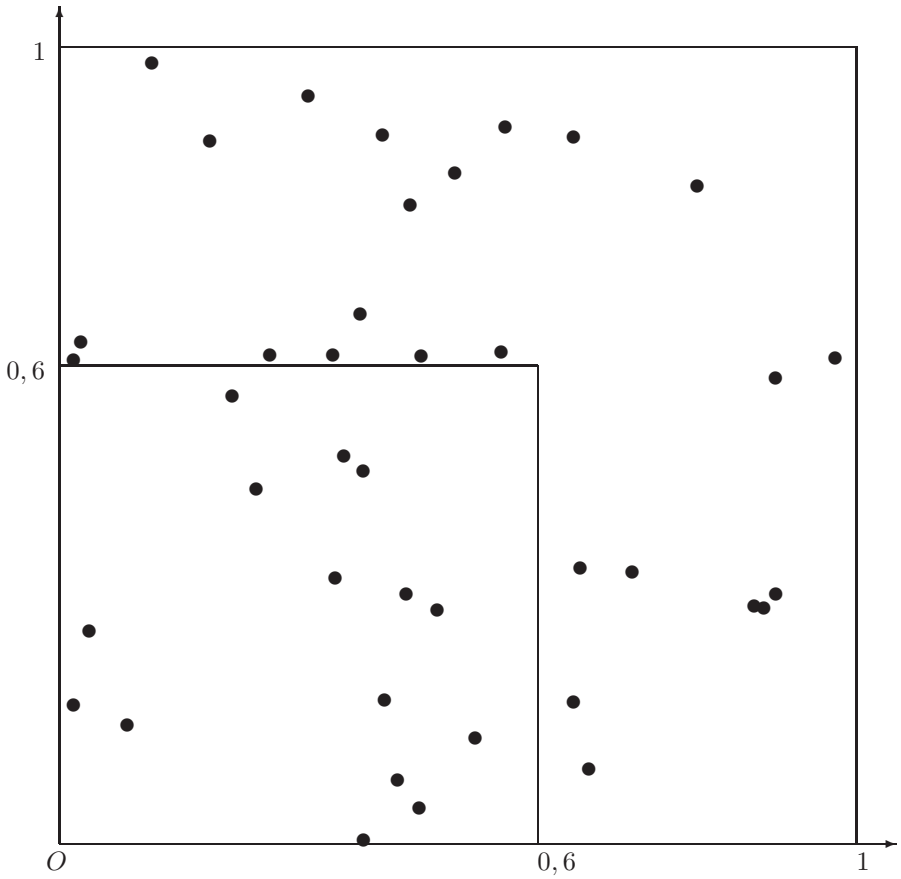

Fig. 1.1. *Estimation de la surface d'un carré.*

On peut également améliorer l'approximation en répétant l'expérience un certain nombre de fois et en prenant la moyenne des surfaces obtenues. Nous avons répété l'expérience 50 fois et le nombre de points aléatoires compris dans S étaient les suivants : 10, 19, 18, 19, 23, 19, 20, 12, 10, 16, 17, 13, 10, 14, 15,

15, 9, 14, 18, 12, 11, 11, 13, 9, 14, 18, 10, 12, 13, 16, 10, 17, 15, 22, 16, 20, 14, 10, 16, 17, 13, 16, 21, 14, 14, 15, 19, 17, 12, 20.

Nous avons à chaque fois calculé le quotient $N^{'}/N$. La moyenne des quotients nous a donné la surface approximative de 0,365 5. Ce nouveau résultat est plus proche de la vraie valeur de 0,36 que l'estimation précédente qui était de 0,375.

Pour mieux comprendre la méthodologie de résolution des problèmes à l'aide de nombres aléatoires, nous allons prendre un autre exemple où une solution par simple formule géométrique n'existe pas.

Supposons que nous devons calculer la surface d'une région S que l'on a mise à l'intérieur d'un carré de côté égal à 1 (voir Fig. 1.2). Il est clair que la valeur se situera entre 0 et 1. Pour l'estimer, nous disposons de façon aléatoire N points à l'intérieur du carré. Nous notons par N' le nombre de ces points se trouvant à l'intérieur de S. Encore une fois, la surface S est estimée par le quotient N'/N.

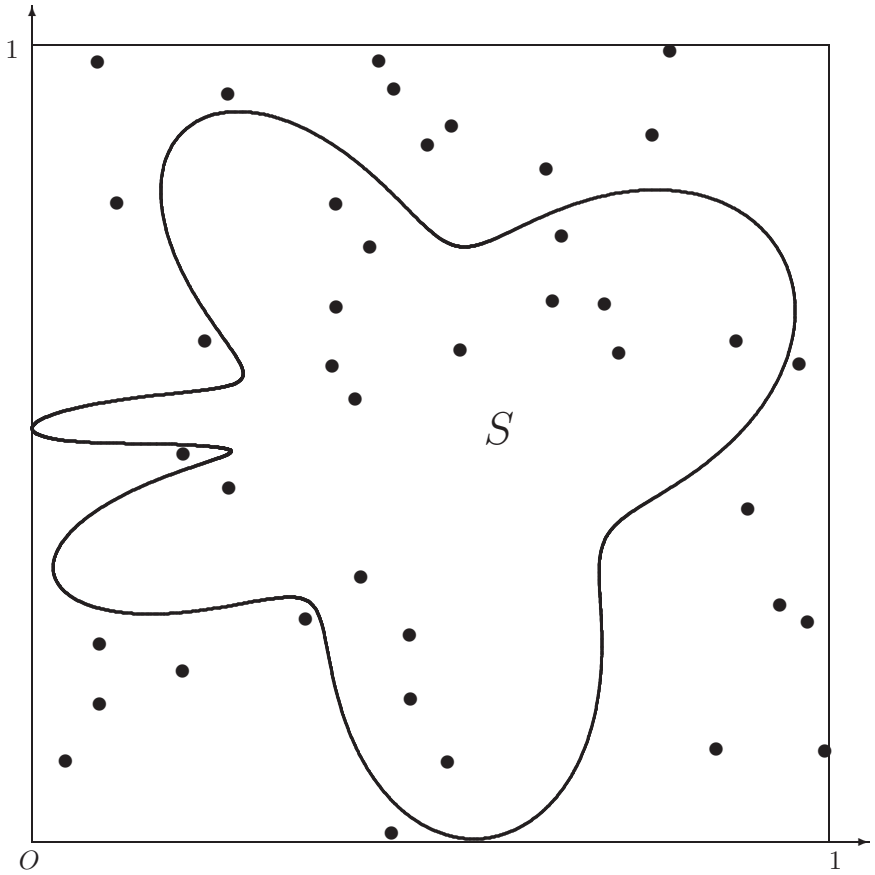

Fig. 1.2. *Estimation d'une surface.*

Dans la Figure 1.2, on dispose 40 points aléatoires à l'intérieur du carré. De ces 40 points, 16 se trouvent à l'intérieur de S. Le quotient N'/N est égal à $16/40 = 0{,}4$, ce qui signifie que la région délimitée par S a approximativement une surface égale à 0,4.

En augmentant la taille de l'échantillon, on peut améliorer cette estimation. En effet, plus N est grand, plus le quotient sera proche de la surface réelle, l'idéal étant de couvrir l'ensemble de l'échantillon et de compter la répartition des points. Mais ce n'est justement pas le but de la simulation qui est de simplifier les calculs et de gagner du temps. Nous reviendrons sur cet exemple au Chapitre 6 où nous analyserons notamment le problème de la fiabilité d'une telle estimation : nous montrerons que l'écart-type des valeurs calculées est proportionnel à $1/\sqrt{N}$.

Exercices

1.1 Considérer la région $D = \{(x, y) \in \mathbb{R}^2 | x \geq 0, x \geq 0, x^2 + y^2 \leq 1\}$. De cette région la surface peut être calculée exactement.

- (a) Comparer le résultat exact avec celui obtenu en utilisant 20 points opportunément choisis dans $[0, 1] \times [0, 1]$ à l'aide de nombres aléatoires.
- (b) Refaire l'opération en utilisant 100 points.
- (c) Quels sont vos commentaires éventuels ?

1.2 On veut simuler 10 issues d'un lancement de dé non biaisé à 6 faces (dé classique). Utilisez des nombres aléatoires de votre choix en expliquant la procédure que vous suivez.

1.3 Estimez la surface de la région

$$D = \{(x, y) \in \mathbb{R}^2 \mid \sqrt{x} + \sqrt{y} \leq 1, x, y \geq 0\}$$

en utilisant 200 nombres aléatoires de votre choix.

1.4 Estimez l'intégrale

$$\int_0^1 |\sin(1/x)| dx$$

en utilisant 100 nombres aléatoires de votre choix.

1.5 Dans le canton de Genève 37 % des ménages sont composés d'une seule personne ; 32 % de 2 personnes ; 12 % de 3 personnes ; 12 % de 4 personnes ; 5 % de 5 personnes et 2 % de 6 personnes ou plus (Recensement fédéral suisse 2000, OFS, Neuchâtel, 2004). Simulez les tailles d'un échantillon de 10 familles habitant le canton en utilisant des nombres aléatoires de votre choix et en expliquant la procédure suivie.

Chapitre 2

Éléments de probabilités

2.1 Introduction

Dans ce chapitre nous présentons quelques éléments de base du calcul des probabilités et de la statistique. Nous parlerons de variables aléatoires et de leurs propriétés, ainsi que de quelques cas spécifiques de variables aléatoires fréquemment utilisées. Ces cas spécifiques de variables aléatoires constituent des lois qui s'adaptent souvent très bien à décrire des séries de phénomènes réels. Le but n'est donc pas de faire un traitement exhaustif de la matière, mais de donner des éléments pour que le lecteur puisse disposer des outils nécessaires pour la suite. Les différentes techniques de simulation d'échantillons utiliseront des transformations de variables aléatoires opportunément choisies et seront exposées dans le Chapitre 4.

Un traitement complet de la théorie des variables aléatoires n'est pas le sujet de cet ouvrage, et le lecteur peut se référer par exemple à l'excellent livre de Lejeune (2004). D'autres références sont le livre de Ross (1996), devenu un classique, et le tout récent ouvrage de Cantoni, Huber et Ronchetti (2006).

Considérons une expérience dont le résultat est incertain et supposons que l'ensemble des résultats possibles, lui, soit connu. On appelle cet ensemble *ensemble fondamental de l'expérience* et on le note par Ω.

On peut s'intéresser à une fonction du résultat plutôt qu'au résultat lui-même. Les événements auxquels on s'intéresse sont liés à des fonctions réelles définies sur l'ensemble fondamental et qui sont appelées *variables aléatoires*.

Une variable aléatoire est une fonction X définie sur un ensemble fondamental à valeurs réelles. Les notations utilisées ici sont celles qui sont en usage dans la littérature : on utilise d'habitude des lettres majuscules de la fin de l'alphabet (U, V, X, Y, Z, ...) pour dénoter une variable aléatoire dans sa généralité. Par Ω on entend l'ensemble fondamental, c'est-à-dire un ensemble d'événements lié à l'expérience que l'on veut représenter et qui couvre tous les cas de figure de l'expérience en question. Chaque événement est donc associé à une valeur numérique qui en général est réelle. L'image d'une variable aléatoire

X peut toutefois être un sous-ensemble quelconque, fini ou infini de $\mathbb{R}$.

Exemple 2.1 *On jette simultanément* 2 *pièces de monnaie et l'on s'intéresse au nombre de piles qui apparaissent. L'ensemble fondamental est alors :*

$$\Omega = \{FF, FP, PF, PP\}$$

où F représente le côté face et P le côté pile.

On obtient une variable aléatoire X qui peut prendre les valeurs 0, 1 *ou* 2 *avec les probabilités suivantes :*

$$\begin{aligned} \Pr(X = 0) &= \Pr(FF) = 0,25 \\ \Pr(X = 1) &= \Pr(FP) + \Pr(PF) = 0,5 \\ \Pr(X = 2) &= \Pr(PP) = 0,25. \end{aligned}$$

Étant donné une variable aléatoire X, il est possible d'y associer des probabilités pour chaque événement donné. Ainsi dans l'exemple précédent l'événement « obtenir au moins une pile » a probabilité $0,75$. Ces probabilités permettent ensuite de définir, quand elles existent et sont finies, des mesures de tendance centrale comme l'espérance ou la valeur médiane, et de dispersion comme la variance.

2.2 Variables aléatoires discrètes

Une variable aléatoire *discrète* X prend ses valeurs sur un ensemble fini $x_1, x_2, \ldots, x_n$ ou sur un ensemble infini dénombrable.

Soit la variable aléatoire discrète X définie par les probabilités suivantes :

X	x_1	x_2	$\ldots$	x_n
$\Pr(X = x_i)$	p_1	p_2	$\ldots$	p_n

Tab. 2.1. *Tableau qui résume une variable aléatoire discrète ayant un nombre fini de valeurs.*

Les $x_1, x_2, \ldots, x_n$ sont les valeurs possibles de la variable X et les $p_1, p_2, \ldots, p_n$ sont les probabilités associées aux valeurs correspondantes. La fonction $f(x)$ qui associe à chaque x la valeur $\Pr(X = x)$ s'appelle aussi *loi de probabilité* ou *densité*. Par conséquent, la probabilité que la variable aléatoire X prenne la valeur x_i est égale à p_i et on écrit :

$$\Pr(X = x_i) = p_i.$$

Les probabilités $p_1, p_2, \ldots, p_n$ doivent satisfaire deux conditions :

(a) tous les p_i sont positifs : $p_i > 0$;

(b) la somme de tous les p_i est égale à 1. Si la variable aléatoire prend ses valeurs sur un ensemble fini alors :

$$p_1 + p_2 + \cdots + p_n = \sum_{i=1}^{n} p_i = 1.$$

Dans le cas où l'ensemble est infini :

$$\sum_{i=1}^{\infty} p_i = 1.$$

La *fonction de répartition*, ou *fonction cumulée de densité* d'une variable aléatoire discrète X est une fonction définie par :

$$F_X(x) = \Pr(X \leq x) = \sum_{x_i \leq x} \Pr(X = x_i).$$

La fonction de répartition de X évaluée en x correspond donc à la probabilité que X soit inférieure ou égale à x.

Si tous les p_i sont égaux, la variable aléatoire X peut prendre chacune des valeurs x_i avec la même probabilité. On parle alors d'*équiprobabilité*. Si les n valeurs possibles $x_1, x_2, \ldots, x_n$ sont équiprobables, alors :

$$p_i = \frac{1}{n}, \qquad \text{pour} \quad i = 1, 2, \ldots, n.$$

Dans ce cas, on dit aussi que la variable aléatoire X est distribuée selon une loi discrète uniforme.

Si X est une variable aléatoire discrète, alors l'*espérance mathématique* de X, notée $E(X)$, est définie par :

$$E(X) = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i.$$

En d'autres termes, $E(X)$ est la moyenne pondérée des valeurs que X peut prendre, les poids étant les probabilités associées à ces différentes valeurs.

La *valeur médiane* d'une variable aléatoire est la valeur x_m telle que $F_X(x_m) = 0,5$. Dans le cas discret s'il n'existe pas de valeur x de X pour lesquelles $F_X(x) = 0,5$, elle est définie par interpolation entre les deux valeurs x_1 et x_2 de X et adjacentes à la valeur médiane (c'est-à-dire $x_1 = \max\{x \mid F_X(x) < 0,5\}$ et $x_2 = \min\{x \mid F_X(x) > 0,5\}$) selon la formule :

$$x_m = x_1 + \frac{0,5 - F_X(x_1)}{F_X(x_2) - F_X(x_1)}(x_2 - x_1).$$

La *variance* de X est une mesure de la dispersion. Sa valeur, toujours non négative, a tendance à être d'autant plus grande que les valeurs de X se dispersent sur un intervalle plus étendu. Elle est notée var (X) et est définie par :

$$\text{var } (X) = E(X - E(X))^2.$$

Le *moment d'ordre* r d'une variable aléatoire discrète est défini par :

$$E(X^r) = \sum_{i=1}^{n} x_i^r p_i.$$

Les moments permettent entre autres choses de définir des indices caractérisant l'ensemble des probabilités associées à une variable aléatoire. Par exemple la variance de X est égale à la différence entre le moment d'ordre 2 de X et le carré du moment d'ordre 1 de X.

Exemple 2.2 *Si l'on reprend l'Exemple* 2.1*, on peut calculer l'espérance mathématique de la variable aléatoire* X*. Rappelons que* X *représente le nombre de piles obtenu en lançant* 2 *pièces de monnaie. Il s'agit d'une variable aléatoire discrète. Ses valeurs et les probabilités associées sont :*

$X = x_i$	0	1	2
$\Pr(X = x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Tab. 2.2. *Variable aléatoire décrivant le nombre de piles obtenu en lançant* 2 *pièces de monnaie.*

L'espérance mathématique de X*, ou le nombre moyen d'apparitions de piles lorsqu'on lance* 2 *pièces de monnaie, est égale à :*

$$E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

La variance de X *est ici égale à :*

$$var\ (X) = E((X-1)^2) = (0-1)^2 \cdot \frac{1}{4} + (1-1)^2 \cdot \frac{1}{2} + (2-1)^2 \cdot \frac{1}{4} = \frac{1}{2}.$$

Exemple 2.3 *En se référant à l'exemple du nombre d'avions au sol dans un aéroport (Paragraphe* 1.5*), on peut calculer l'espérance mathématique de la variable* X *« nombre d'arrivées par minute ».*

$$E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{10} + 4 \cdot \frac{3}{125} + 5 \cdot \frac{1}{1\ 000} = 1,401.$$

Ainsi, entre 17 *h* 00 *et* 17 *h* 18*, le nombre moyen théorique d'arrivées est égal à* $18 \cdot 1,401 = 25,218$.

Remarquons que la moyenne simulée d'arrivées par minute est égale à :

$$\frac{2+3+2+0+3+2+\ldots+1+2+2}{18} = \frac{29}{18} = 1,61.$$

Ainsi, le nombre moyen simulé d'arrivées entre 17 *h* 00 *et* 17 *h* 18 *est égal à* 29.

Nous allons voir maintenant quelques lois de probabilité discrètes.

2.2.1 La loi de Bernoulli

Soit X une variable aléatoire définie par la distribution de probabilité

X	0	1
$\Pr(X = x)$	$1-p$	p

Tab. 2.3. *Schéma d'une variable de Bernoulli.*

Une variable aléatoire ayant un caractère dichotomique et dont la loi de probabilité $\Pr(X = x)$ est définie par un tableau comme celui ci-dessus est dite variable aléatoire de Bernoulli de paramètre p. On a :

$$E(X) = p \qquad \text{var}\,(X) = p(1-p).$$

La loi de Bernoulli est utilisée lorsqu'une expérience aléatoire n'a que deux résultats possibles : le *succès* avec une probabilité de p, et l'*échec* avec une probabilité $q = 1 - p$. Les applications de cette loi sont nombreuses. On peut citer comme exemple l'étude de la composition d'une population (masculin-féminin) ou le contrôle de la qualité de certaines marchandises (bonne ou défectueuse). À noter que dans la littérature, lorsqu'on parle de variables dichotomiques il est d'usage de dénoter $1 - p$ plus simplement par la lettre q.

2.2.2 La loi binomiale

Considérons un ensemble de n variables aléatoires indépendantes, suivant une loi de Bernoulli de même paramètre p. Soit $S_n = \sum_{i=1}^{n} X_i$, la somme de ces n variables de Bernoulli. La loi de probabilité de la variable aléatoire S_n est appelée loi binomiale, notée $\text{bin}(n, p)$ et est donnée par :

$$\Pr(S_n = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

où

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}.$$

L'espérance mathématique d'une variable aléatoire binomiale S_n est égale à np et sa variance vaut $np(1-p)$.

Exemple 2.4 *On lance une pièce de monnaie* 8 *fois. La probabilité d'obtenir exactement* 3 *piles est :*

$$\Pr(S_8 = 3) = \left(\begin{array}{c} 8 \\ 3 \end{array} \right) \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{8-3} \simeq 0,219 = 21,9\ \%.$$

2.2.3 La loi géométrique

Considérons n répétitions indépendantes d'une variable de Bernoulli. La loi géométrique est la loi de la variable aléatoire X : « nombre d'observations jusque (et y compris) au premier succès ». Les probabilités associées à une loi géométrique sont de la forme :

$$\Pr(X = i) = pq^{i-1}$$

où p est la probabilité d'observer un succès, $q = 1 - p$ la probabilité d'échouer et i est le nombre d'essais jusqu'à ce que le premier succès (inclus) se réalise. La loi géométrique est utilisée entre autres choses dans le cadre des études de processus stochastiques et de la théorie des files d'attente.

2.2.4 La loi binomiale négative

La loi de la variable aléatoire Y_k : « temps d'attente jusqu'au $k^{\text{ième}}$ succès » est la loi binomiale négative. Si p est la probabilité de succès et $q = 1 - p$, les probabilités sont données par :

$$\begin{aligned} \Pr(Y_k = n) &= \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) p^k q^{n-k} \\ &= \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^k q^{n-k} \\ &= \frac{(n-1)!}{(k-1)!(n-k)!} p^k q^{n-k} \end{aligned}$$

où k est le nombre de succès et n le nombre d'essais jusqu'à ce que le $k^{\text{ième}}$ succès (inclus) se réalise.

La loi binomiale négative trouve ses applications en sciences biologiques, en écologie, en recherche médicale et en psychologie.

En simulation, il est souvent nécessaire de générer des suites de nombres entiers compris entre 0 et 9, chacun d'eux ayant la même probabilité de sélection. Au Chapitre 3 nous discuterons de plusieurs manières dont un tel tirage aléatoire peut s'effectuer.

Exemple 2.5 *Nous nous intéressons, dans cet exemple, à la variable aléatoire qui représente le nombre d'entiers tirés avant le premier* 0. *Cette expérience consiste en une suite d'essais identiques et indépendants dont un succès est défini par le tirage d'un* 0. *La probabilité d'un succès est* $\frac{1}{10}$. *Le nombre d'entiers précédent le premier* 0 *est donc une variable aléatoire géométrique avec* $p = \frac{1}{10}$. *Calculons par exemple la probabilité que le premier* 0 *apparaisse au cinquième tirage :*

$$\Pr(X = 5) = \frac{1}{10}\left(\frac{9}{10}\right)^4 = 0{,}065\ 1.$$

La variable aléatoire qui compte le nombre d'entiers jusqu'au quatrième 0 *est évidemment une variable binomiale négative. La probabilité qu'un quatrième* 0 *apparaisse au sixième tirage est :*

$$\Pr(Y_4 = 6) = \begin{pmatrix} 6-1 \\ 4-1 \end{pmatrix}\left(\frac{1}{10}\right)^4\left(\frac{9}{10}\right)^{6-4} = 0{,}000\ 81.$$

La variable aléatoire Z *comptant le nombre de* 0 *dans une suite de* n *nombres est une variable binomiale de paramètres* n *et* $p = \frac{1}{10}$. *La probabilité d'observer quatre* 0 *en* 6 *tirages est :*

$$\Pr(Z = 4) = \begin{pmatrix} 6 \\ 1 \end{pmatrix}\left(\frac{1}{10}\right)^4\left(\frac{9}{10}\right)^{6-4} = 0{,}000\ 486.$$

2.2.5 La loi de Poisson

La variable aléatoire X, comptant le nombre de réalisations d'un certain événement par unité de temps ou par exemple, par unité de surface, dont la loi est donnée par :

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

est dite variable de Poisson de paramètre λ. L'espérance mathématique est égale au paramètre λ. De plus var $(X) = E(X) = \lambda$.

Les applications de la loi de Poisson sont nombreuses et variées. Cette loi peut décrire des phénomènes tels que le nombre d'individus blessés dans un accident de voiture ; le nombre de plantes d'un type donné par unité de surface ; le nombre de défauts sur un écran de télévision ; le nombre de caries par individu ; le nombre de personnes se présentant à un guichet et ainsi de suite.

Exemple 2.6 *Si le nombre moyen d'arrivées de clients à un guichet par heure est égal à* 15, *calculons la probabilité d'observer* 20 *arrivées dans une heure donnée, supposant que les arrivées sont indépendantes les unes des autres. Ici, la valeur de* λ *est égale à* 15 *et la valeur de* k *est égale à* 20. *Nous aurons donc :*

$$\Pr(X = 20) = \frac{e^{-15}15^{20}}{20!} \simeq 0,042 = 4{,}2\ \%.$$

Exemple 2.7 *Supposons qu'à l'aéroport de Londres-Heathrow les mouvements d'avions entre* 17 *h* 00 *et* 17 *h* 15 *soient en moyenne de* 3 *arrivées et de* 3 *départs par minute. En admettant que le nombre d'arrivées soit une variable aléatoire de Poisson, le paramètre de la loi est égal à* 3 *et cette loi est donnée dans la table ci-dessous (au millième près) :*

$$\Pr(X = k) = \frac{e^{-3}3^k}{k!}.$$

Pour la simulation du nombre d'avions au sol, la procédure à utiliser avec la loi de Poisson, connue également sous le nom de méthode de comparaison (voir Paragraphe 4.4), est semblable à celle utilisée avec la loi empirique.

La fonction de répartition d'une loi de Poisson de paramètre λ quelconque peut s'écrire :

$$F_X(x) = \sum_{k \leq x} \Pr(X = k) = \sum_{k \leq x} \frac{e^{-\lambda}\lambda^k}{k!}.$$

Cette fonction peut se représenter graphiquement. La Figure 2.1 représente une fonction de répartition pour la loi de Poisson de l'exemple précédent, avec $\lambda = 3$.

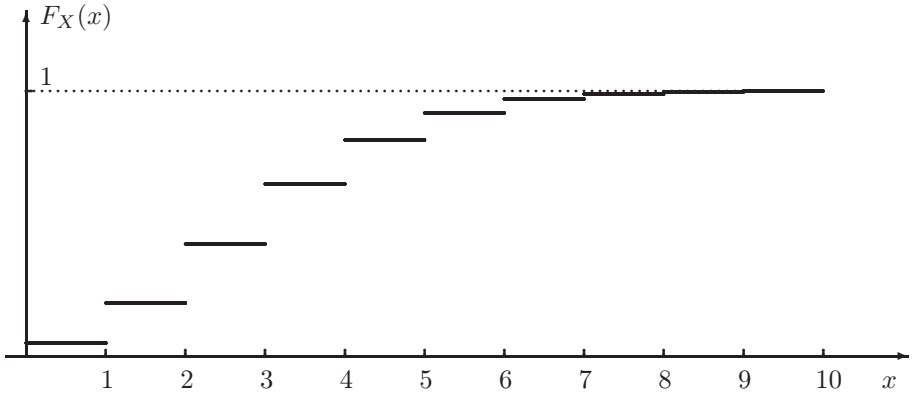

Fig. 2.1. *Fonction de répartition de la loi de Poisson de paramètre* $\lambda = 3$.

Il existe encore d'autres lois de variables aléatoires discrètes, telles que la loi multinomiale et la loi hypergéométrique. Au besoin, nous laissons au lecteur le soin de consulter les ouvrages spécialisés mentionnés dans l'introduction à ce chapitre.

2.3 Variables aléatoires continues

Si une variable aléatoire X peut prendre les valeurs réelles contenues dans un intervalle, elle est dite *continue*. Une variable aléatoire continue X est définie par son intervalle de variation (a, b) et par la fonction $f(x)$ appelée *fonction de densité*.

L'intervalle peut être ouvert ou fermé, même seulement d'un côté, borné ou non. Il est commode de prolonger f à l'intervalle $(-\infty, +\infty)$ en posant $f(x) = 0$ en dehors de l'intervalle de définition originaire. Rappelons finalement que dans la littérature francophone il est courant d'utiliser pour une fonction de densité les synonymes *distribution de probabilité* ou encore *loi de probabilité*.

La fonction de densité $f(x)$ doit satisfaire deux conditions :

(a) la fonction de densité $f(x)$ est non négative : $f(x) \geq 0$;

(b) l'intégrale de la fonction de densité $f(x)$ est égale à 1 ;

$$\int_{\mathbb{R}} f(x)dx = 1.$$

La probabilité que X appartienne à l'intervalle $[c, d]$ est donnée par l'intégrale :

$$\Pr(c < X \leq d) = \int_c^d f(x)dx.$$

Si X est une variable aléatoire continue prenant ses valeurs sur un intervalle (a, b), alors l'espérance mathématique de X, notée $E(X)$, est définie par :

$$E(X) = \int_a^b xf(x)dx.$$

Si l'intégrale définissant l'espérance diverge, on dit que X n'a pas d'espérance. Les notions d'espérance mathématique, de valeur médiane, de variance et de moments se transposent du cas discret au cas continu en substituant au symbole $\sum$ son équivalent infinitésimal $\int$ tout en gardant la même signification. On a donc que la valeur médiane, x_m, de X est telle que

$$\int_{-\infty}^{x_m} f(x)dx = 0{,}5 \ .$$

La variance est définie par :

$$\text{var } (X) = \int_a^b (x - E(X))^2 f(x)dx.$$

Exemple 2.8 *Supposons que la fonction de densité d'une variable aléatoire X soit donnée par :*

$$f(x) = \begin{cases} \frac{2}{3} & si\ 0 < x \leq \frac{1}{2} \\ \frac{1}{3} & si\ \frac{1}{2} < x \leq \frac{5}{2} \\ 0 & sinon \end{cases}$$

X est une variable aléatoire continue et l'on vérifie aisément que $f(x) \geq 0$. De même :

$$\int_{\mathbb{R}} f(x)dx = \int_0^{\frac{1}{2}} \frac{2}{3}dx + \int_{\frac{1}{2}}^{\frac{5}{2}} \frac{1}{3}dx = 1.$$

L'espérance mathématique de X est égale à :

$$\begin{aligned} E(X) &= \int_0^{\frac{5}{2}} xf(x)dx = \int_0^{\frac{1}{2}} \frac{2}{3}xdx + \int_{\frac{1}{2}}^{\frac{5}{2}} \frac{1}{3}xdx \\ &= \frac{x^2}{3}\Big|_0^{\frac{1}{2}} + \frac{x^2}{6}\Big|_{\frac{1}{2}}^{\frac{5}{2}} \\ &= \frac{1}{12} + 1 = \frac{13}{12}. \end{aligned}$$

La *fonction de répartition* d'une variable aléatoire continue X est dénotée par F_X (ou par F ou autre lettre majuscule) et est donnée par :

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(x)dx,$$

où $f(x)$ est la fonction de densité de cette variable.

Étudions à présent quelques lois relatives à des variables aléatoires continues.

2.3.1 La loi uniforme

On dit qu'une variable aléatoire X définie sur l'intervalle $[a, b]$ avec une fonction de densité constante

$$f(x) = \frac{1}{b-a}$$

est distribuée uniformément dans l'intervalle $[a, b]$. On notera $X \sim U(a, b)$.

Par intégration, nous obtenons la fonction de répartition F :

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \dfrac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } x > b. \end{cases}$$

La fonction de densité $f(x)$ et la fonction de répartition $F(X)$ sont représentées dans la Figure 2.2.

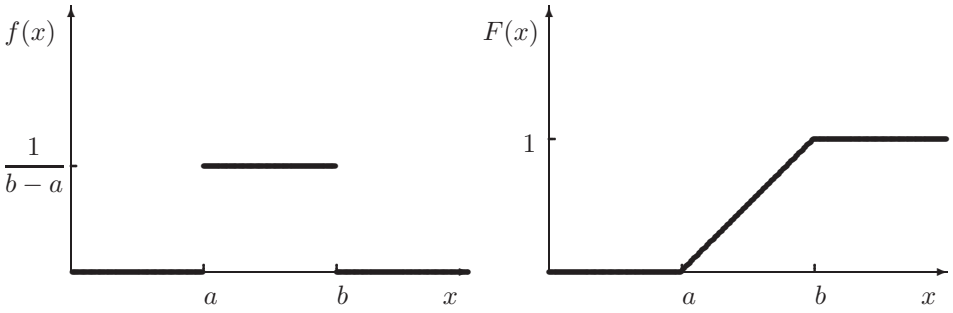

Fig. 2.2. *Fonction de densité et de répartition de la loi uniforme.*

L'espérance mathématique de la loi uniforme est égale à :

$$E(X) = \int_a^b x f(x)dx = \int_a^b x \left(\frac{1}{b-a}\right) dx = \frac{1}{b-a} \left.\frac{x^2}{2}\right|_a^b = \frac{a+b}{2} .$$

Sa variance est égale à :

$$\text{var } (X) = \int_a^b (x - E(X))^2 f(x)dx = \frac{(b-a)^2}{12} .$$

Le moment d'ordre r est égal à :

$$\begin{aligned} E(X^r) &= \int_a^b x^r f(x)dx \\ &= \int_a^b x^r \frac{1}{b-a} dx \\ &= \frac{1}{b-a} \cdot \left.\frac{x^{r+1}}{r+1}\right|_a^b \\ &= \frac{b^{r+1} - a^{r+1}}{(r+1)(b-a)} . \end{aligned}$$

Il est évident que pour $a = 0$, $b = 1$ et $r = 1$ on retrouve l'espérance d'une variable aléatoire uniformément distribuée sur l'intervalle $[0, 1]$.

En simulation, la loi uniforme est utilisée comme base pour générer des nombres aléatoires issus de n'importe quelle loi de probabilité, comme nous le verrons plus loin.

Lorsque les bornes de l'intervalle d'une loi uniforme ne sont pas précisées, il s'agit de l'intervalle $[0, 1]$.

2.3.2 La loi exponentielle

Un système complexe peut tomber en panne. La détermination des probabilités de son fonctionnement est fondamentale pour la gestion de ses relations avec son environnement.

La variable aléatoire « instant où un système complexe tombe en panne » suit souvent une loi exponentielle.

On dit qu'une variable aléatoire X suit une distribution exponentielle de paramètre λ si sa densité f_X est donnée par :

$$f_X(x) = \lambda e^{-\lambda x} \qquad 0 \leq x < \infty,\ \lambda > 0.$$

Son espérance et sa variance sont respectivement données par :

$$E(X) = \frac{1}{\lambda} \qquad \text{et} \qquad \text{var}\ (X) = \frac{1}{\lambda^2}.$$

La fonction $1 - F_X(x) = \Pr(X > x)$ donne la *probabilité de survie* :

$$\Pr(X > x) = \begin{cases} 1 & \text{si } x < 0 \\ e^{-\lambda x} & \text{si } x \geq 0\ . \end{cases}$$

Une caractéristique importante d'une variable aléatoire exponentielle est la propriété d'absence de mémoire donnée par :

$$\Pr(X \leq x_0 + x | X > x_0) = \Pr(X \leq x) \qquad x > 0,\ \ x_0 > 0.$$

Cela signifie que la loi de la durée de vie future d'un objet reste la même quel que soit le temps que l'objet a déjà fonctionné.

La loi exponentielle est fréquemment utilisée pour décrire des événements aléatoires dans le temps (file d'attente, durée de vie d'un composant, etc.).

Exemple 2.9 *Considérons la variable aléatoire X « durée de vie (en heures) d'un composant électronique de type donné ». Supposons que la densité de probabilité de X soit donnée par :*

$$f(x) = \begin{cases} \dfrac{1}{100} e^{-\frac{x}{100}} & \textit{si } x > 0 \\ \\ 0 & \textit{sinon.} \end{cases}$$

Sa fonction de répartition est :

$$F_X(x) = 0 \;\; si \;\; x < 0$$

et pour $x \geq 0$ *:*

$$\begin{aligned} F_X(x) &= \int_0^x \frac{1}{100} e^{\frac{-t}{100}} dt \;\; si \;\; x \geq 0 \\ &= -e^{\frac{-t}{100}} \Big|_0^x \\ &= 1 - e^{\frac{-x}{100}}. \end{aligned}$$

Ainsi la probabilité que la durée de vie d'un composant électronique de ce type soit inférieure à 200 *heures est donnée par :*

$$F_X(200) \simeq 1 - 0{,}135\ 3 = 0{,}864\ 7.$$

L'espérance de vie d'un tel composant est égale à $\int_0^\infty \frac{x}{100} e^{\frac{-x}{100}} dx$*. Par intégration par parties on trouve que la durée moyenne de vie est égale à* 100 *heures.*

2.3.3 La loi normale

Une variable aléatoire X définie sur l'intervalle $(-\infty, \infty)$, caractérisée par la fonction de densité suivante :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

où μ et σ sont deux paramètres, est appelée variable aléatoire normale (ou gaussienne). On écrit $X \sim \mathcal{N}(\mu, \sigma^2)$.

Si X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, il est facile de montrer que $E(X) = \mu$ et var $(X) = \sigma^2$. Le paramètre μ ne change pas la forme du graphe de f. Sa variation déplace uniquement la « cloche » le long de l'axe des x. En revanche, la variation de σ change la forme de la courbe. En effet, ce paramètre représente la dispersion de la variable aléatoire autour de sa moyenne μ. Plus σ est petit, plus la courbe sera « serrée » autour de μ. On montre facilement que :

$$\max f(x) = f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}.$$

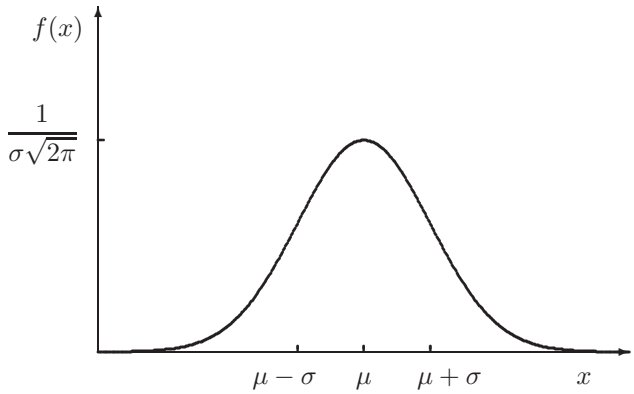

Fig. 2.3. *La fonction de densité d'une variable aléatoire qui suit une loi normale* $\mathcal{N}(\mu, \sigma^2)$.

La loi normale centrée réduite correspond à la loi normale ayant comme paramètres $\mu = 0$ et $\sigma = 1$ et il est d'usage de la dénoter avec la lettre Z. Le passage de la variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$ à la variable aléatoire $Z \sim \mathcal{N}(0, 1)$ s'effectue par la transformation $Z = \frac{X-\mu}{\sigma}$. Ainsi $\Pr(X \leq a) = \Pr\left(Z \leq \frac{a-\mu}{\sigma}\right)$.

Par convention, la fonction de répartition de Z est notée $\Phi(z)$, c'est-à-dire

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt.$$

Les valeurs de $\Phi(z)$ sont données dans la table qui figure à l'Appendice.

Exemple 2.10 *Dans une exploitation agricole la production de fruits passe à travers des machines qui sélectionnent le produit sur la base de leur diamètre. Soit la variable aléatoire X représentant le diamètre d'un abricot. Le diamètre moyen en centimètres est de* 4,25 *et l'écart-type de* 0,5. *Trouvons la probabilité qu'un abricot ait un diamètre compris entre $x_1 = 5$ et $x_2 = 5{,}5$. Pour ce calcul, nous cherchons tout d'abord les valeurs standardisées z_1 et z_2 correspondant à x_1 et x_2 :*

$$z_1 = \frac{5 - 4{,}25}{0,5} = 1{,}5$$

$$z_2 = \frac{5{,}5 - 4{,}25}{0{,}5} = 2{,}5.$$

La probabilité que la variable X soit comprise entre 5 *et* 5,5 *est égale à la probabilité que la variable Z soit entre* 1,5 *et* 2,5 *:*

$$\begin{aligned} \Pr(5 < X \leq 5{,}5) &= \Pr(1{,}5 < Z \leq 2{,}5) \\ &= \Phi(2{,}5) - \Phi(1{,}5) \\ &= 0{,}993\,8 - 0{,}933\,2 \\ &\cong 0{,}06 = 6\ \%\ . \end{aligned}$$

La loi normale est sans aucun doute la loi la plus utilisée en statistique classique. Elle permet très souvent de décrire la loi des erreurs de mesure.

C'est *le théorème central limite* qui justifie l'usage de la loi normale dans certaines situations.

Soient $X_i,\ i = 1, \ldots, n$ des variables aléatoires indépendantes de même loi telles que $E(X_i) = \mu$ et var $(X_i) = \sigma^2$. Notons par S_n la somme de ces variables :

$$S_n = X_1 + X_2 + \cdots + X_n \ .$$

On a :

$$\begin{aligned} E(S_n) &= E\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} E(X_i) \\ &= n\mu \ . \end{aligned}$$

Par l'indépendance des variables X_i, la variance de la somme est égale à la somme des variances. Ainsi :

$$\text{var } (S_n) = n\sigma^2 \ .$$

Théorème 2.1 *Soient $X_i,\ i = 1, \ldots, n$ des variables aléatoires indépendantes de même loi telles que $E(X_i) = \mu$ et var $(X_i) = \sigma^2$. Pour $n \to \infty$,*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0, 1) \ .$$

En d'autres termes, le théorème dit que la somme d'un grand nombre de variables aléatoires indépendantes de même loi, de moyenne μ et de variance σ^2, suit approximativement une loi normale de moyenne $n\mu$ et de variance $n\sigma^2$. La démonstration de ce théorème fait appel à des notions qui dépassent le cadre de cet ouvrage.

2.3.4 La loi gamma

Considérons r variables aléatoires $X_i,\ i = 1, \ldots, r$, indépendantes, identiquement distribuées selon la loi exponentielle de paramètre λ. La variable $Y = \sum_{i=1}^{r} X_i$ est distribuée selon la loi gamma de paramètres (λ, r), dont la densité est :

$$f_Y(y) = \begin{cases} \dfrac{\lambda^r}{\Gamma(r)}\ e^{(-\lambda y)} y^{r-1} & \text{si } y \geq 0 \\ \\ 0 & \text{sinon} \end{cases}$$

où $\Gamma(r)$ est la fonction gamma, définie par

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx.$$

À noter que la loi gamma est définie pour $r \in \mathbb{R}$. On a

$$E(Y) = \frac{r}{\lambda} \quad \text{et} \quad \text{var}\,(X) = \frac{r}{\lambda^2}.$$

Par le changement de variable $z = \lambda y$, la variable aléatoire $Z = \lambda Y$ est distribuée selon la loi *gamma standard de paramètre* r avec :

$$f_Z(z) = \begin{cases} \dfrac{1}{\Gamma(r)}\ z^{r-1}\ e^{-z} & \text{si } z \geq 0 \\ \\ 0 & \text{sinon.} \end{cases}$$

On dénote par $I_r(z)$ sa fonction de répartition. Remarquons que les égalités suivantes :

$$F_Y(y) = \Pr(Y \leq y) = \Pr\left(\frac{Z}{\lambda} \leq y\right) = \Pr(Z \leq \lambda y) = I_r(\lambda y)$$

permettent de déterminer la fonction de répartition d'une variable aléatoire distribuée selon la loi gamma de paramètres (λ, r) à partir de la loi gamma standard de paramètre r.

Exemple 2.11 *Soit la variable aléatoire Y représentant la durée de vie en années d'un système. Lorsque la composante originale de ce système tombe en panne, une des 4 composantes de secours la remplace. Ainsi, la durée de vie du système est $Y = \sum_{i=1}^{5} X_i$, où X_i est la durée de vie de la $i^{ième}$ composante. Nous faisons l'hypothèse que la variable aléatoire X_i suit une loi exponentielle de paramètre $\lambda = 2/5$.*

Calculons la probabilité que ce système ait une durée de vie comprise entre 15 et 20 ans.

La durée de vie Y est une variable aléatoire suivant une loi gamma de paramètre $(\frac{2}{5}, 5)$. Donc :

$$\begin{aligned} \Pr(15 \leq Y \leq 20) &= F_Y(20) - F_Y(15) \\ &= I_5(8) - I_5(6) \\ &\simeq 0{,}900\ 37 - 0{,}714\ 94 \\ &= 0{,}185\ 43 = 18{,}543\ \%\ . \end{aligned}$$

2.3.5 La loi du chi-carré

La loi du chi-carré, noté χ^2, est d'importance fondamentale en statistique. En effet, c'est la densité d'une somme de variables aléatoires normales indépendantes élevées au carré.

Soit ν un entier positif. La densité de probabilité d'une variable aléatoire X distribuée selon la loi du χ^2 de paramètre ν (ou avec ν degrés de liberté) est donnée par :

$$f_\nu(x) = \begin{cases} \dfrac{1}{2^{\frac{\nu}{2}}\Gamma\left(\dfrac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} & \text{si } x > 0 \\ \\ 0 & \text{si } x \leq 0. \end{cases}$$

L'espérance de X est égale à ν et sa variance à 2ν. Remarquons que la loi du χ^2 est une loi gamma particulière.

Pour $X_1, \ldots, X_n$, n variables aléatoires indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$, la variable $\frac{(n-1)s^2}{\sigma^2}$ est une variable du χ^2 avec $(n-1)$ degrés de liberté, où $s^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$.

L'expression d'une variable aléatoire distribuée selon χ^2, en termes d'écarts quadratiques entre effectifs observés et effectifs théoriques (espérés), a permis le développement de tests d'adéquation ainsi que de tests d'indépendance associés aux tables de contingence. Une table de la loi du χ^2 est présentée à l'Appendice, et représente les valeurs de x telles que

$$\chi^2_\nu(x) = \alpha,$$

où $\chi^2_\nu(x)$ est la fonction de répartition de $f_\nu(x)$. La valeur α est appelée *seuil de signification*.

Les utilisations principales de la distribution du χ^2 sont les tests d'indépendance pour les tables de contingence, les tests de comparaison de variances et les tests d'adéquation. Dans le cadre de cet ouvrage nous nous intéresserons essentiellement aux tests d'adéquation. Ces derniers sont utilisés pour faire de l'inférence à propos de la distribution de probabilité des variables aléatoires étudiées dans une population et seront présentés à la fin du Chapitre 5.

2.3.6 La loi de Student

Si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes distribuées selon une loi normale identique $\mathcal{N}(\mu, \sigma^2)$, alors en notant $\overline{X} = \frac{1}{n}\sum X_i$, et $s^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$, la variable aléatoire définie par :

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

a pour densité de probabilité :

$$f(t) = \frac{\Gamma\left(\dfrac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\dfrac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{\left(-\frac{\nu+1}{2}\right)}$$

où $\Gamma(\cdot)$ est la fonction introduite au Paragraphe 2.3.4. Il est : $E(T) = 0$, var $(T) = \frac{\nu}{\nu-2}$, $\nu > 2$ où $\nu = n - 1$ est le nombre de degrés de liberté. La variable T représente celle qui est habituellement appelée *loi de Student*.

Dans le cas de deux populations normales $\mathcal{N}(\mu_i, \sigma^2)$, à variance identique, la variable aléatoire de Student, définie par :

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}},$$

où

$$s^2_{\overline{X}_1 - \overline{X}_2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

est à la base de l'inférence concernant la différence $\mu_1 - \mu_2$.

2.3.7 La loi de Fisher

Le rapport de deux variables indépendantes X_1 et X_2, distribuées selon une loi χ^2, chacune d'elle divisée par ses degrés de liberté, définit une variable aléatoire de Fisher dont la densité de probabilité est donnée par :

$$f_{(\nu_1,\nu_2)}(x) = \frac{\Gamma\left[(\nu_1 + \nu_2)/2\right]}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \cdot \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \cdot \frac{x^{(\nu_1-2)/2}}{\left(1 + \frac{\nu_1}{\nu_2}x\right)^{(\nu_1+\nu_2)/2}}$$

ν_1 et ν_2 étant le nombre de degrés de liberté de X_1 et X_2 respectivement. La distribution de Fisher est utilisée pour la mise en place de certains tests d'hypothèses (voir aussi le Paragraphe 5.4.2).

2.3.8 Autres lois de distributions

Une distribution relativement courante en statistique est la distribution de Pareto. Elle est définie par une fonction de densité du type suivant :

$$f(x) = ak^a x^{-(a+1)}, \qquad \text{pour } x > k > 0.$$

Cette loi est souvent appliquée pour la description de la distribution des revenus.

Une variable aléatoire suit une loi de Cauchy si sa densité de distribution est

$$f(x) = \frac{1}{\pi t}\left(1 + \frac{x - a}{t}\right)^{-1},$$

pour certaines valeurs positives de t et a. Cette distribution n'a pas d'espérance. Sa valeur médiane est a. Son premier et son troisième quartile sont respectivement $a - t$ et $a + t$. Cette loi est utilisée pour décrire par exemple les points d'impact de particules émises en faisceau.

Une variable aléatoire suit une loi bêta si elle est de la forme :

$$f(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{(\alpha+\beta-1)}B(\alpha,\beta)}$$

où $a \leq x \leq b$; $\alpha > 0$; $\beta > 0$, et

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

La loi bêta est utilisée pour ajuster des distributions dont le domaine de variation est connu.

2.4 Les lois bivariées

Soient X et Y deux variables aléatoires. La fonction de répartition conjointe des variables X et Y est définie par :

$$F(x,y) = \Pr(X \leq x \text{ et } Y \leq y)$$

et la paire (X,Y) est appelée aussi variable aléatoire bivariée.

2.4.1 Cas discret

Si X et Y sont deux variables aléatoires discrètes, la fonction de répartition conjointe est déterminée par la loi conjointe $\Pr(x,y)$:

$$\Pr(x,y) = \Pr(X = x \text{ et } Y = y) \ .$$

Les lois marginales $P_X(x) = \Pr(X = x)$ et $P_Y(y) = \Pr(Y = y)$ sont données par :

$$\begin{aligned} P_X(x) &= \sum_y \Pr(x,y) \\ P_Y(y) &= \sum_x \Pr(x,y) \ . \end{aligned}$$

Exemple 2.12 *En programmation, les erreurs les plus fréquentes sont soit des erreurs de syntaxe, soit des erreurs de logique. Notons par X le nombre d'erreurs de syntaxe et par Y le nombre d'erreurs de logique détectées lors de la première compilation d'un programme d'examen à écrire en* `R`*. Une population de* 500 *individus se sont présentés à l'examen, population pour laquelle les probabilités conjointes d'erreurs des deux types ont été calculées lors de la compilation des* 500 *programmes. Le Tableau* 2.4 *contient la loi conjointe et les lois marginales. La fonction de répartition est donnée dans le Tableau* 2.5 *et peut être calculée par la relation :*

$$F(x,y) = \Pr(x,y) + F(x-1,y) + F(x,y-1) - F(x-1,y-1) \ .$$

x	y			$P_X(x)$
	0	1	2	
0	0,3	0,1	0,08	0,48
1	0,2	0,1	0,05	0,35
2	0,05	0,02	0,03	0,10
3	0,02	0,01	0,01	0,04
4	0,01	0,01	0,01	0,03
$P_Y(y)$	0,58	0,24	0,18	1

Tab. 2.4. *Loi conjointe et lois marginales.*

x	y		
	0	1	2
0	0,3	0,4	0,48
1	0,5	0,7	0,83
2	0,55	0,77	0,93
3	0,57	0,80	0,97
4	0,58	0,82	1

Tab. 2.5. *Fonction de répartition.*

2.4.2 Cas continu

Si X et Y sont deux variables aléatoires continues telles que $F(x,y)$, leur fonction de répartition conjointe, soit dérivable en x et en y, la densité conjointe est définie par :

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y) \ .$$

Dans ce cas, $f_X(x)$, la densité marginale de la variable aléatoire X et $f_Y(y)$, celle de Y, sont définies par :

$$\begin{aligned} f_X(x) &= \int_{\mathbb{R}} f(x,y)dy \\ f_Y(y) &= \int_{\mathbb{R}} f(x,y)dx \ . \end{aligned}$$

Les densités conditionnelles $f_{X|Y}$, $f_{Y|X}$ sont définies par les formules suivantes :

$$\begin{aligned} f_{X|Y} &= \frac{f(x,y)}{f_Y(y)}, \qquad f_Y(y) > 0 \\ f_{Y|X} &= \frac{f(x,y)}{f_X(x)}, \qquad f_X(x) > 0 \ . \end{aligned}$$

2.4.3 Cas particulier : la loi normale bivariée

Deux variables aléatoires X et Y sont distribuées selon une loi normale bivariée si leur densité de probabilité conjointe est de la forme :

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}\left(\frac{1}{1-\rho^2}\right)\left[\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(y-\mu_2)^2}{\sigma_2^2} -2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]\right\} .$$

Les paramètres μ_1 et μ_2 sont des paramètres de position et localisent le « centre de la cloche » dans le plan cartésien O_{xy} (voir Fig. 2.4). Les paramètres σ_1 et σ_2 sont des paramètres de dispersion selon l'axe des x pour σ_1 et l'axe des y pour σ_2. Le paramètre ρ, coefficient de corrélation mesurant le lien entre X et Y, détermine la forme et l'orientation de la cloche dans le plan cartésien.

On peut montrer que $f_X(x)$, la densité marginale de la variable aléatoire X, est la densité d'une variable $\mathcal{N}(\mu_1, \sigma_1^2)$. On a aussi $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

La densité conditionnelle $f_{Y|X}$ est donnée par

$$f_{Y|X}(y|X=x) = \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{B}{2\sigma_Y^2(1-\rho^2)}\right\}$$

où

$$B = \left(y-\mu_Y-\frac{\rho\sigma_Y}{\sigma_X}(x-\mu_X)\right)^2 .$$

Et ainsi l'on voit que :

$$E(Y|X=x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$$

ce qui détermine une relation du type régression linéaire simple.

On montre aussi que si X et Y sont distribuées selon une loi normale bivariée, n'importe quelle combinaison linéaire du type $aX+bY+c$ est aussi distribuée normalement avec $aX+bY+c \sim \mathcal{N}(a\mu_1+b\mu_2+c,\ a^2\sigma_1^2+b^2\sigma_2^2+2ab\rho\sigma_1\sigma_2)$.

Remarquons encore que la loi normale bivariée est la seule loi conjointe possédant la propriété ci-dessus. En outre, il existe des distributions bivariées non normales dont les densités marginales sont normales.

On représente souvent la densité conjointe de deux variables aléatoires par des courbes de niveau. La Figure 2.4 illustre les courbes de niveau de la densité conjointe d'une loi normale bivariée.

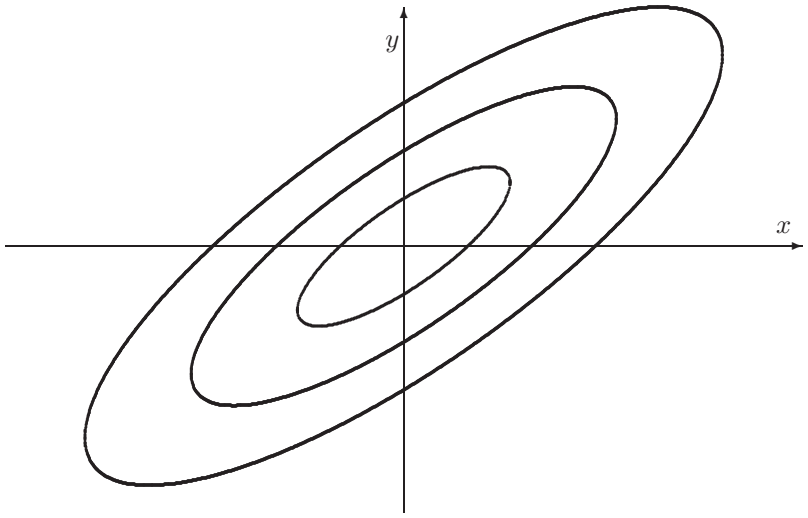

Fig. 2.4. *Courbes de niveau pour la fonction de densité d'une loi normale bivariée avec* $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 0$, $\sigma_2 = 2$, $\rho = 0{,}8$.

Exercices

2.1 Soit X la variable aléatoire discrète définie par

$$\Pr\,(X = n) = \frac{1}{n(n+1)}, \quad n \in \mathbb{N}.$$

(a) Démontrer qu'il s'agit bien d'une distribution de probabilité, c'est-à-dire que

$$\sum_{n=1}^{\infty} \Pr\,(X = n) = 1.$$

(b) Calculer, si elle existe, $E(X)$.

2.2 Soit X la variable aléatoire continue définie pour une constante opportune c par la fonction de densité de probabilité

$$f(x) = \begin{cases} cx^2(x-1) & \text{si } x \in [0,1] \\ 0 & \text{autrement.} \end{cases}$$

(a) Calculer c de sorte que f soit une fonction de densité de probabilité.

(b) Calculer $E(X)$.

(c) Calculer la médiane.

(d) Calculer le mode.

2.3 Dans un magasin, des pommes sont stockées avant d'être mises en vente. Les pommes ont des poids que l'on suppose distribués selon une loi normale de moyenne 200 et écart-type 20 (en grammes).

(a) À l'aide des tables de la loi normale, déterminer la probabilité qu'une pomme ait un poids compris entre 190 et 210 grammes.

(b) Quelle est la probabilité que le poids total de 2 pommes soit compris entre 380 et 420 grammes ? (*Attention : la probabilité n'est pas forcément la même que dans le point précédent !*)

2.4 Au guichet d'une gare le temps de service du bureau des renseignements suit une loi exponentielle de moyenne 5 (en minutes).

(a) Quelle est la probabilité qu'un client soit servi dans un temps compris entre 2 et 3 minutes ?

(b) La durée de travail de l'employé au guichet est de 4 heures. Quelle est la probabilité qu'au moins 50 clients soient servis avant la fermeture ?

2.5 Soit X la variable aléatoire continue qui décrit les temps t séparant deux clients à un guichet. Soit $a(t) = \lambda e^{-\lambda t}$ sa fonction de densité. Démontrer que la variable aléatoire discrète Y qui décrit le nombre d'arrivées par unité de temps suit la loi de Poisson

$$\Pr(Y = n) = \frac{e^{-\lambda}\lambda^n}{n!} \ .$$

2.6 Une marque S d'ampoules standard a les propriétés suivantes : chaque unité coûte 1 euro, sa consommation d'électricité est de 4 centimes par jour et les durées de vie de cette marque sont distribuées selon la loi exponentielle $f_S(t) = e^{-t}$, t ne pouvant qu'être positif.

Une marque E d'ampoules écologiques a en revanche les propriétés suivantes : chaque unité coûte 5 euros, sa consommation d'électricité est de 1 centime par jour et les durées de vie de cette marque sont distribuées selon la loi exponentielle $f_E(t) = 2e^{-2t}$.

Un client ne se souciant pas de l'impact écologique hésite à choisir entre les deux marques. Quelle marque devrait-il choisir afin d'avoir un avantage économique sur le long terme ?

2.7 Une équipe de football marque un nombre de buts par match qui suit une loi de Poisson de moyenne fixée. Quatre équipes, qui marquent en moyenne un but par match, doivent se disputer les deux places pour passer un tour éliminatoire. Les équipes A et B se trouvent en tête du classement provisoire. Le dernier match prévoit A contre B et C contre D.

(a) Estimer au millième près la probabilité que, dans le dernier match, A et B égalisent.

(b) Sachant que pour se qualifier, l'équipe C a besoin de battre l'équipe D sans que A et B égalisent, estimer la probabilité que C se qualifie.

2.8 En France, le salaire minimum garanti est de 1 000 euros par mois. Le revenu médian est de 3 000 euros par mois.

(*a*) Quelle densité de distribution proposeriez-vous pour les salaires en France ?

(*b*) Quel est le pourcentage de Français dont le revenu dépasse 10 000 euros par mois ?

2.9 Une variable aléatoire discrète X est définie par

$$\Pr(X = n) = \frac{k}{n \cdot 2^n} \qquad \text{pour } n \in \mathbb{N}.$$

(*a*) Déterminer la valeur de k.

(*b*) Calculer l'espérance de X.

Chapitre 3

Nombres aléatoires

3.1 Introduction

Comme nous l'avons montré dans le Chapitre 1, pouvoir disposer de nombres aléatoires en grande quantité est un élément central dans toute simulation. Les logiciels statistiques ou non, comme R, SPSS, SAS, Minitab ou Excel, intègrent des fonctions qui permettent d'en obtenir sans grand effort, au point que l'utilisateur ne se pose même pas la question de savoir comment ces nombres sont générés et quels algorithmes se cachent derrière eux.

Ce chapitre tente d'apporter quelques réponses. Différentes techniques sont illustrées, à partir de la méthode de congruence jusqu'au registre à décalage, et à ses évolutions. L'état de l'art de la génération de nombres aléatoires est résumé dans le Paragraphe 3.6.

Le lecteur qui s'intéresse plus spécifiquement aux applications de la simulation peut donc passer aux chapitres suivants qui ne demandent d'ailleurs pas directement une application des techniques illustrées ici.

3.2 Nombres aléatoires et pseudo-aléatoires

Une méthode élémentaire pour générer des nombres aléatoires consiste à tirer d'une urne des billets numérotés de 0 à 9, un billet à la fois, en remettant dans l'urne chaque billet tiré. Une méthode plus sophistiquée consiste à observer 4 fois une variable de Bernoulli dont les résultats équiprobables sont codés par 0 ou 1. Le résultat de cette expérience peut être interprété comme un nombre binaire que l'on traduit (en base 10) dans le système décimal. Si le nombre obtenu est supérieur à 9 on le rejette et on recommence l'expérience ; si le nombre obtenu est compris entre 0 et 9, il sera considéré comme une réalisation d'une variable aléatoire uniforme sur $\{0, 1, 2, \ldots, 9\}$. En procédant plusieurs fois comme il est indiqué ci-dessus on peut construire une table de nombres aléatoires comme celle donnée en annexe.

Formellement on définit un nombre aléatoire comme la réalisation d'une variable aléatoire distribuée uniformément dans l'intervalle $[0, 1]$. Une suite de nombres aléatoires sera dans la suite de cet ouvrage un échantillon issu d'une population distribuée selon la loi $U(0, 1)$ dont les éléments sont indépendants les uns des autres.

Il faut distinguer entre nombres aléatoires dans un intervalle et donc issus d'une variable aléatoire continue, et nombres aléatoires dans un ensemble fini, comme l'ensemble $\{0, 1, \ldots, 9\}$. Dans ce cas les nombres aléatoires sont issus d'une variable aléatoire discrète. L'expérience qui consiste à tirer des billets numérotés de 0 à 9 et ensuite à remettre ces billets dans l'urne qui les contient, produit donc un échantillon issu d'une variable aléatoire discrète, notamment une loi aléatoire uniforme discrète. Toutefois, pour produire des échantillons de nombres aléatoires issus d'une variable aléatoire continue, comme c'est le cas pour la loi uniforme dans un intervalle, on utilisera dans la pratique des nombres choisis dans un ensemble fini où chaque élément représente un nombre de l'intervalle approximé à $1/10^h$ où h est fixé. Cela permettra de produire des nombres aléatoires, approximés au chiffre décimal souhaité, issus d'une variable aléatoire continue à l'aide d'un procédé pour ainsi dire discret.

N'importe quel processus naturel considéré comme aléatoire peut être utilisé comme générateur de nombres aléatoires. Nous pouvons aussi utiliser certaines lois physiques pour générer des nombres aléatoires ; par exemple, le jeu de la roulette. Nous pouvons rassembler les nombres obtenus au sein de tables que l'on appelle tables de nombres aléatoires, très utiles si l'on doit répéter une expérience. Le développement de la technologie informatique a toutefois supplanté ce type de générateurs au profit des générateurs de nombres pseudo-aléatoires.

Toute suite de nombres dont la distribution ne diffère pas significativement d'une distribution uniforme et dont la dépendance n'est pas significative est considérée (et utilisée) comme une suite de nombres aléatoires. Ces conditions n'étant pas réalisables dans la pratique, on se contente des méthodes permettant de générer une suite de nombres vérifiant un certain nombre de propriétés probabilistes. On parle alors de nombres *pseudo-aléatoires*.

Formellement une suite de nombres pseudo-aléatoires est une suite de nombres possédant les mêmes propriétés qu'une suite de nombres aléatoires, mais générée à travers une procédure déterministe. Les nombres pseudo-aléatoires sont donc générés par des fonctions mathématiques ou par des algorithmes. Cette distinction doit être toutefois relativisée. Les méthodes les plus avancées de génération de nombres pseudo-aléatoires produisent en effet des suites qui même quand elles sont soumises à toute sorte de test (voir Chapitre 5) sont pratiquement impossibles à distinguer d'une suite de nombres aléatoires. Dans les applications on parlera de suite de nombres aléatoires, en sachant qu'en effet il pourrait s'agir de nombres pseudo-aléatoires tout à fait utilisables comme des nombres aléatoires.

La plupart des algorithmes pour la génération de nombres pseudo-aléatoires ont pour but de produire des suites uniformément distribuées. Une classe très répandue de générateurs est fondée sur des congruences, et les algorithmes mis au point sont connus sous le nom de méthodes de congruence. D'autres s'inspirent des suites récursives du type de celle de Fibonacci ou font appel à des registres à décalage où une transformation intermédiaire intervient dans la détermination de la suite de nombres.

3.3 La méthode du carré médian

Cette méthode, connue dans la littérature anglophone aussi sous le nom de *middle square*, a été introduite par John von Neumann (1951). Le carré médian est considéré comme la première méthode de génération automatique de nombres pseudo-aléatoires. Bien que dépassée par des techniques plus performantes, elle a aujourd'hui sans doute un intérêt historique. Nous l'illustrerons par un exemple. Soit d_0 un nombre compris entre 0 et 1, par exemple $d_0 = 0{,}917\ 5$. Nous l'élevons au carré en prenant 8 décimales : $d_0^2 = 0{,}841\ 806\ 25$. Parmi ces 8 décimales, nous prenons les 4 du milieu et formons $d_1 = 0{,}180\ 6$. De nouveau, nous élevons d_1 au carré en prenant 8 décimales : $d_1^2 = 0{,}032\ 616\ 36$. Nous sortons les 4 décimales du milieu : $d_2 = 0{,}261\ 6$, et ainsi de suite.

Nous avons alors la suite des nombres pseudo-aléatoires :

$$\begin{aligned} d_0 &= 0{,}917\ 5 \\ d_1 &= 0{,}180\ 6 \\ d_2 &= 0{,}261\ 6 \\ &\vdots \end{aligned}$$

Plus généralement, on génère une suite de nombres ayant chacun m chiffres, où m est un nombre pair. Le successeur d'un nombre de cette suite est obtenu en élevant ce nombre au carré puis en en retenant les m chiffres du milieu. Ce raisonnement donne lieu à l'algorithme suivant pour générer une suite de nombres en base 10 :

(a) **initialisation** : choisir un nombre x_0 à m chiffres ; $i := 0$; $u_0 = x_0/10^m$;

(b) **itération** : $i := i+1$; $x_{i+1} := \left[x_i^2/10^{m/2}\right] \bmod 10^m$; $u_{i+1} = x_{i+1}/10^m$ où $[z]$ signifie partie entière de z et $a \bmod b$ est égale au reste de la division de a par b.

La suite des u_i contient alors les nombres pseudo-aléatoires obtenus par la méthode du carré médian.

Intuitivement cette méthode paraît donner de bons résultats. Cependant ce n'est pas le cas, car très souvent les nombres s'approchent de 0 pour y rester, ou produisent un cycle trop court de chiffres qui se répète indéfiniment.

Exemple 3.1 *Dans l'exemple ci-dessous, où* $x_1 = 1\ 926$ *et* $m = 4$*, la suite se réduit à* 0 *après* 27 *itérations.*

i	x_i	u_i	x_i^2	i	x_i	u_i	x_i^2
1	1 926	0,192 6	03 709 476	15	1 406	0,140 6	01 976 836
2	7 094	0,709 4	50 324 836	16	9 768	0,976 8	95 413 824
3	3 248	0,324 8	10 549 504	17	4 138	0,413 8	17 123 044
4	5 495	0,549 5	30 195 025	18	1 230	0,123 0	01 512 900
5	1 950	0,195 0	03 802 500	19	5 129	0,512 9	26 306 641
6	8 025	0,802 5	64 400 625	20	3 066	0,306 6	09 400 356
7	4 006	0,400 6	16 048 036	21	4 003	0,400 3	16 024 009
8	0 480	0,048 0	00 230 400	22	0 240	0,024 0	00 057 600
9	2 304	0,230 4	05 308 416	23	0 576	0,057 6	00 331 776
10	3 084	0,308 4	09 511 056	24	3 317	0,331 7	11 002 489
11	5 110	0,511 0	26 112 100	25	0 024	0,002 4	00 000 576
12	1 121	0,112 1	01 256 641	26	0 005	0,000 5	00 000 025
13	2 566	0,256 6	06 584 356	27	0 000	0,000 0	00 000 000
14	5 843	0,584 3	34 140 649				

Tab. 3.1. *Tableau pour la génération de nombres aléatoires au moyen de la méthode du carré médian. À noter que la suite converge* à 0.

Exemple 3.2 *Soit maintenant* $x_1 = 3\ 792$*;* $m = 4$*. Par la méthode du carré médian, puisque* $3\ 792^2 = 14\ 379\ 264$*, il est* $x_2 = x_3 = \cdots = 3\ 792$*. Dans cet autre cas limite la suite ne converge pas à* 0*, sans que pour autant la suite puisse être considérée comme aléatoire.*

Cette méthode ne garantit donc pas la production de suites de nombres aléatoires, et c'est pourquoi elle a été abandonnée depuis longtemps.

3.4 Les méthodes de congruence

3.4.1 La méthode de congruence simple

La méthode de congruence consiste en un algorithme simple pour la génération de nombres pseudo-aléatoires et est définie par la relation de récurrence

$$x_i = (ax_{i-1} + b) \bmod m$$

pour $i = 1, 2, \ldots$ et a, b, m et x_0 des entiers positifs donnés. On dit que a est le multiplicateur, b est l'incrément, m le modulus et x_0 la valeur initiale ou *seed*.

La notation mod m signifie qu'après avoir divisé $ax_{i-1} + b$ par m on ne considère que le reste de la division, qui constitue alors le nouveau nombre x_i.

Le nombre pseudo-aléatoire, u_i, compris entre 0 et 1, est obtenu en divisant x_i par m : $u_i := x_i/m$.

Exemple 3.3 *On pose* $x_0 = 1$ *et*

$$x_i = (5x_{i-1} + 5) \bmod 32.$$

On obtient la suite :

$x_1 = 10$	$x_2 = 23$	$x_3 = 24$	$x_4 = 29$	$x_5 = 22$	$x_6 = 19$
$x_7 = 4$	$x_8 = 25$	$x_9 = 2$	$x_{10} = 15$	$x_{11} = 16$	$x_{12} = 21$
$x_{13} = 14$	$x_{14} = 11$	$x_{15} = 28$	$x_{16} = 17$	$x_{17} = 26$	$x_{18} = 7$
$x_{19} = 8$	$x_{20} = 13$	$x_{21} = 6$	$x_{22} = 3$	$x_{23} = 20$	$x_{24} = 9$
$x_{25} = 18$	$x_{26} = 31$	$x_{27} = 0$	$x_{28} = 5$	$x_{29} = 30$	$x_{30} = 27$
$x_{31} = 12$	$x_{32} = 1$	$x_{33} = 10$	...		

Tab. 3.2. *Le cycle des classes de congruence selon la relation de congruence* $x_i = (5x_{i-1} + 5) \bmod 32$.

La suite ci-dessus se répète après le 32^{e} élément. La longueur du cycle est un indice de qualité de la suite générée et dépend du choix des paramètres a, b et m. Leur influence est illustrée par quelques cas particuliers :

Exemple 3.4 *Soit* $a = 6, b = 0, m = 25$ *et* $x_0 = 1$. *La formule de récurrence*

$$x_i = 6x_{i-1} \bmod 25$$

génère la suite : $x_1 = 6$, $x_2 = 11$, $x_3 = 16$, $x_4 = 21$, $x_5 = 1$, $x_6 = 6$, $x_7 = 11$, ...

Divisons ces nombres par la longueur $m = 25$ de la suite. On obtient des nombres pseudo-aléatoires compris entre 0 et 1 : $u_1 = 0{,}24$, $u_2 = 0{,}44$, $u_3 = 0{,}64$, $u_4 = 0{,}84$, $u_5 = 0{,}04, \ldots$

Il est évident que cette suite n'est pas très satisfaisante en termes probabilistes.

Remarquons que pour les choix de x_0, a et m donnés ci-dessus la suite de nombres obtenue est composée de répétitions d'un cycle de longueur 5. De plus, le cycle généré est une progression arithmétique. La longueur d'un cycle est définie comme étant la quantité de nombres pseudo-aléatoires qui est générée avant d'obtenir la même séquence de nombres.

Il est donc clair que le choix de a, b, m et x_0 influence la qualité de la suite de nombres pseudo-aléatoires. L'indépendance entre les nombres pseudo-aléatoires implique que la « longueur du cycle » soit suffisamment grande. Cette longueur ne peut excéder m puisqu'il y a au plus m nombres différents modulo m.

La longueur d'une suite quelconque générée par une relation récursive du type $(ax_{i-1}+b) \bmod m$ ne peut pas dépasser m puisqu'il y a au plus m nombres différents modulo m.

De manière générale, il faut choisir une grande valeur de m. Pour obtenir une suite de longueur maximale, c'est-à-dire de période m, il faut que les conditions du théorème suivant soient vérifiées :

Théorème 3.1 *Soit $k \in \{0, 1, \ldots, m-1\}$. Soient a, b, m, tels que :*
i) b et m sont premiers entre eux ;
ii) $(a-1)$ est un multiple de chaque nombre premier qui divise m ;
iii) si m est un multiple de 4 alors $(a-1)$ l'est aussi.
Alors la suite définie par la récurrence

$$\begin{cases} x_0 = k \\ x_i = (ax_{i-1} + b) \bmod m \end{cases}$$

a un cycle de longueur m.

Nous devons ce résultat à Hull et Dobell (1962). En particulier pour $m = 2^k$, $a = 4c + 1$ et b un nombre impair la suite de nombres pseudo-aléatoires est cyclique de longueur m ; k et c étant des entiers positifs.

Idéalement les nombres obtenus devraient être distribués uniformément dans l'intervalle $[0, 1]$. Si cette propriété est vérifiée, les paires (x_1, x_2), (x_2, x_3), $\ldots$, (x_{i-1}, x_i), $\ldots$ devraient être uniformément réparties dans un carré. Hélas, comme les Figures 3.1 et 3.2 le montrent, cela n'est pas toujours le cas, et ce constat constitue le point faible principal de la méthode de congruence simple.

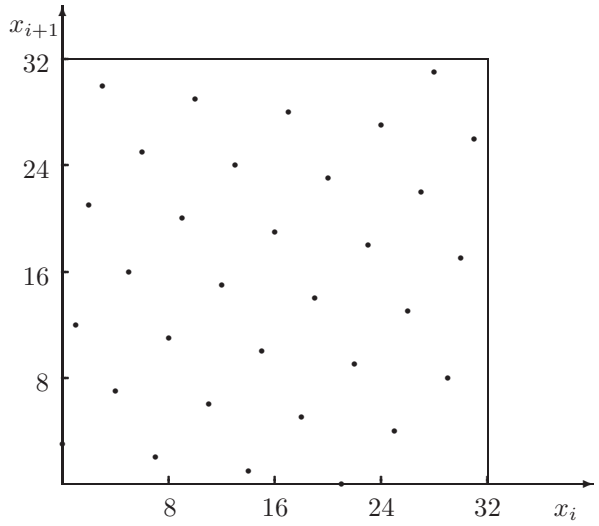

Fig. 3.1. *Les points (x_i, x_{i+1}) pour $i = 0, \ldots, 31$, où $x_0 = 1$ et $x_i = 9x_{i-1} + 3 \bmod 32$.*

Dans la Figure 3.1 on voit que les 32 paires (x_i, x_{i+1}) construites à l'aide du générateur donné par $m = 32, x_0 = 1$, $c = 2, b = 3$ ne sont pas réparties de manière aléatoire dans le carré $[0, 32] \times [0, 32]$, mais réparties plutôt dans un réseau.

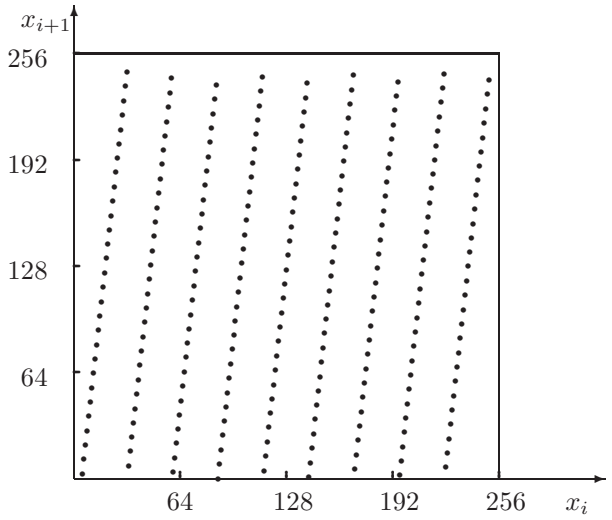

Fig. 3.2. *Les* 256 *points* (x_i, x_{i+1}) *pour* $i = 0, \ldots, 255$, *où* $x_0 = 1$ *et* $x_i = 9x_{i-1} + 3 \bmod 256$.

Dans la Figure 3.2 les 256 paires construites à l'aide du générateur donné par $m = 256, x_0 = 1, c = 2, b = 3$ sont répartis sur 9 segments de droites. Ce constat surprenant s'explique par le fait que la relation de récurrence

$$x_{i+1} = (ax_i + b) \bmod m$$

peut s'écrire $x_{i+1} = ax_i + b - dm$ où $d \in \{0, 1, 2, \ldots, a-1\}$. Or la relation

$$x_{i+1} = ax_i + (b - dm)$$

définit l'équation de a droites.

Pour m suffisamment grand les x_i/m sont distribués approximativement selon une loi uniforme sur $[0, 1]$. Pour une qualité de la suite générée qui soit satisfaisante, il est recommandé de choisir $m > 2^{30}$. Le choix $a = 7^5$, $m = 2^{31} - 1$, $b = 0$, proposé par Park et Miller (1988) est connu sous le nom de *minimal standard*, et si on ne requiert que quelques milliers de nombres aléatoires, ce générateur produit des bonnes suites de nombres pseudo-aléatoires.

3.4.2 La méthode de congruence avec retard

Le point faible de la méthode de congruence réside dans la relation de récurrence qui provoque des irrégularités non désirées. Une modification de la règle de récurrence est donnée par

$$x_i = (ax_{i-r} + b) \bmod m$$

où r indique le retard et où les premiers termes sont calculés avec la relation $x_i = (a'x_{i-1} + b') \bmod m'$, les paramètres a', b' m' pouvant être différents de a, b et m. Pour $r = 1$ on retrouve la méthode de congruence.

Exemple 3.5 *Choisissons les paramètres* $m = 8, x_0 = 1, a = 5, b = 1, r = 2$.
Les deux premiers termes de la suite sont calculés avec $b = 3$ *:*

$$\begin{aligned} x_1 &= (5 \cdot 1 + 3) \bmod 8 = 0 \\ x_2 &= (5 \cdot 0 + 3) \bmod 8 = 3 \end{aligned}$$

ainsi

$$\begin{aligned} x_3 &= (5x_1 + 1) \mod 8 = 1 \\ x_4 &= (5x_2 + 1) \mod 8 = 0 \\ x_5 &= (5x_3 + 1) \mod 8 = 6 \\ x_6 &= (5x_4 + 1) \mod 8 = 1 \\ x_7 &= 7 \\ x_8 &= 6 \\ x_9 &= 4 \\ x_1 &= 7 \\ &\vdots \end{aligned}$$

Les termes suivants sont $5, 4, 2, 5, 3, 2, 0, \ldots$

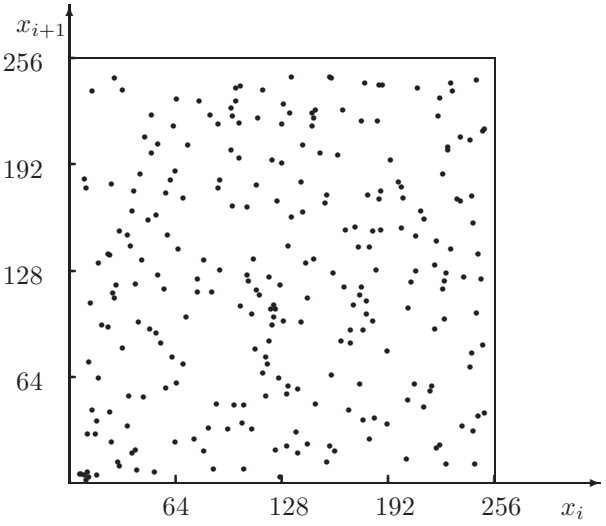

Fig. 3.3. *Les points* (x_i, x_{i+1})*, dans un exemple de congruence avec retard.*

La Figure 3.3 donne la représentation des paires successives formées avec le générateur $m = 256, x_0 = 1, a = 5, b = 1$, $r = 6$. Les premiers termes jusqu'à x_5 ont été calculés avec $m = 8, x_0 = 1, a = 5, b = 1$.

Il n'est pas difficile d'imaginer l'introduction de plusieurs termes de retard pour obtenir des formules de récurrence de la forme

$$x_i = [a(x_{i-r} + x_{i-s}) + b] \bmod m$$

ou

$$x_i = \left[\left(\sum_{j=1}^{r} x_{i-j} \right) + b \right] \bmod m.$$

Ces formules s'inspirent de la célèbre suite de Fibonacci, $1, 1, 2, 3, 5, 8, 13$, où chaque terme est défini comme la somme de deux termes précédents. Le tableau ci-dessous donne la suite construite par le générateur avec $m = 8, x_0 = 1, a = 9, b = 1, r = 6, s = 4$ avec la formule de récurrence

$$x_i = [9\,(x_{i-6} + x_{i-4}) + 1] \bmod (2 \cdot 8)$$

où les termes jusqu'à x_5 ont été déterminés avec la méthode de congruence simple avec $m = 8$, $x_0 = 1$, $a = 5$ et $b = 1$.

i	0	1	2	3	4	5	⋮	6	7	8	9	10	11	...
X_i	1	6	7	4	5	2	⋮	9	11	13	7	15	10	...

Tab. 3.3. *Génération de nombres aléatoires avec la formule de récurrence* $x_i = 9\,(x_{i-6} + x_{i-4}) + 1 \bmod 16$. *Les* 6 *premiers termes sont générés avec la méthode de congruence simple.*

3.4.3 La méthode de congruence avec mélange

Pour éviter certaines régularités, une alternative à la méthode de congruence avec retard consiste à modifier l'ordre des termes de la suite des nombres pseudo-aléatoires. Par exemple pour une suite donnée on forme des groupes de taille t dans lesquels on permute les éléments de manière systématique ou de manière aléatoire en utilisant un autre générateur. Les Figures 3.4 (a), (b) et (c) représentent les paires successives formées par le générateur $m = 2^7$, $x_0 = 1$, $a = 5$, $b = 1$ respectivement sans mélange, et avec mélange systématique par groupes de 2 éléments ; par groupes de 3 éléments : l'ensemble S_3 des permutations de 3 éléments est composé de 6 permutations (voir Tab. 3.4). Le mélange par groupes de 3 éléments est effectué de manière systématique en appliquant les unes après les autres à chaque groupe de 3 éléments les 6 permutations de S_3 en boucle.

1			2			3			4			5			6		
1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	2	3	2	3	1	3	1	2	1	3	2	3	2	1	2	1	3

Tab. 3.4. *L'ensemble de toutes les permutations de* 3 *éléments.*

Remarquons que dans chaque groupe de taille t il y a $t!$ permutations possibles des éléments. Ainsi la méthode de congruence linéaire avec mélange peut donner lieu à $t!$ suites distinctes, qui elles-mêmes dépendent aussi du choix des groupes.

On peut remarquer que plus la taille des groupes qu'on mélange augmente, plus l'apparence des paires successives obéit à une distribution uniforme. La Figure 3.4 (d) représente les paires successives après un mélange aléatoire par groupes de 8 éléments. Le mélange est effectué sur la base des 16 permutations de 8 éléments choisis de manière aléatoire, et dont la liste se trouve à la Table 3.5, qui ont été produites par le générateur intégré dans le logiciel statistique `R`.

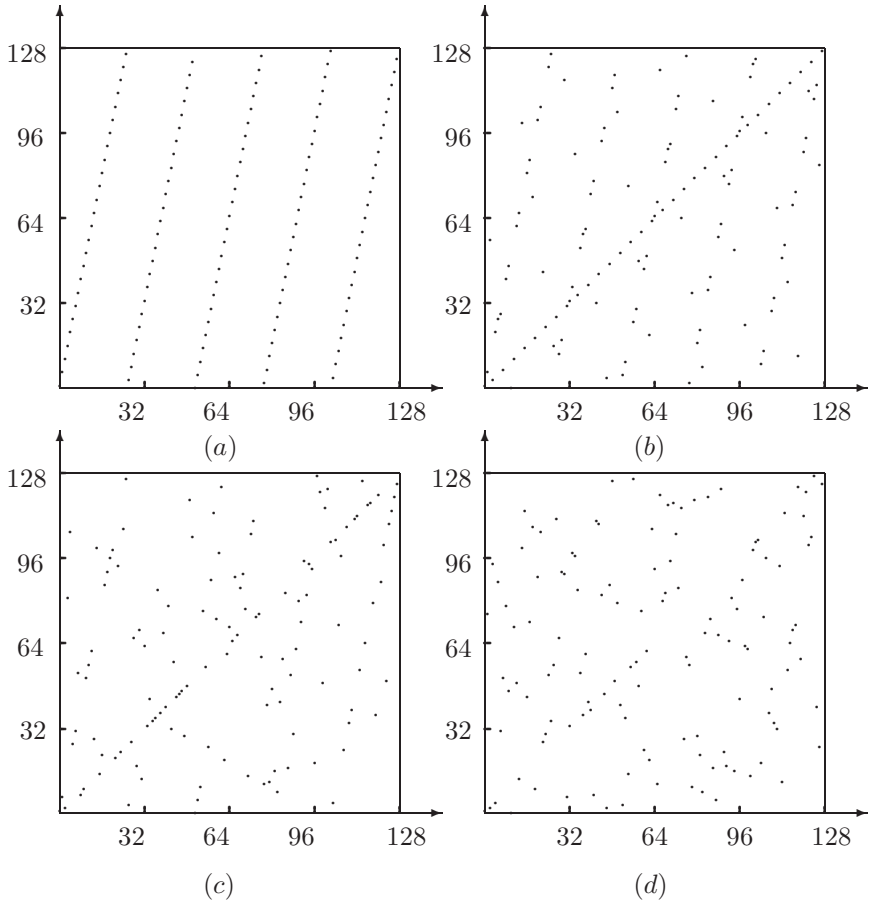

Fig. 3.4. *Les points (x_i, x_{i+1}), dans un exemple de congruence simple (a). La même suite avec mélange systématique par groupes de 2 éléments (b), de 3 éléments (c) et avec mélange aléatoire par groupe de 8 éléments (d).*

	1	2	3	4	5	6	7	8
1	1	4	3	8	6	5	2	7
2	5	8	4	3	7	2	6	1
3	5	4	7	1	6	2	3	8
4	8	2	4	5	3	6	1	7
5	7	3	5	4	8	6	1	2
6	2	4	7	6	5	3	1	8
7	3	4	5	7	1	8	6	2
8	1	5	6	3	4	2	7	8
9	5	8	3	1	4	7	2	6
10	3	1	5	8	6	4	2	7
11	2	7	3	8	1	4	5	6
12	6	7	1	8	2	4	5	3
13	1	8	7	3	6	2	4	5
14	2	8	6	1	4	7	5	3
15	6	2	8	1	4	5	3	7
16	8	1	2	6	5	3	4	7

Tab. 3.5. *Suite de* 16 *permutations de* 8 *éléments produite avec* `R`.

3.4.4 La méthode de l'inverse en congruences

La linéarité de la relation de récurrence de la méthode des congruences peut la rendre inutile dans certains problèmes de simulation. La méthode de l'inverse en congruences (Eichenauer et Lehn, 1986) utilise la notion d'inverse multiplicatif modulo m et supprime ainsi la relation linéaire.

Rappelons que pour p premier, l'inverse de $x \bmod p$ noté $\tilde{x}$ est défini par la relation $x\tilde{x} = 1 \mod p$. Par convention, l'inverse de 0 est 0. La relation de récurrence associée à cette méthode est donnée par :

$$x_{i+1} = (a\tilde{x}_i + b) \bmod p$$

Il est possible de trouver des valeurs de a et b telles qu'on obtient une suite de période p.

Exemple 3.6 *La formule de récurrence* $x_i = (2\tilde{x}_{i-1} + 1) \bmod 7$ *où* $\tilde{x}$ *est l'inverse* mod7 *de* x *(voir ci-dessous) :*

x	1	2	3	4	5	6
$\tilde{x}$	1	4	5	2	3	6

Tab. 3.6. *Les inverses du groupe multiplicatif* $\mathbb{Z}_7^*$.

engendre la suite suivante à partir de $x_0 = 1$ *:*

$$\begin{aligned} x_1 &= (2+1) \mod 7 = 3 \\ x_2 &= (2\cdot 5+1) \mod 7 = 4 \\ x_3 &= (2\cdot 2+1) \mod 7 = 5 \\ x_4 &= 0 \\ x_5 &= 1 \\ x_6 &= 3 \\ &\vdots \end{aligned}$$

La théorie statistique démontre que les nombres pseudo-aléatoires issus de la méthode de l'inverse en congruences ont des propriétés de distribution et d'indépendance proches de celles d'une suite de nombres aléatoires.

3.5 La méthode du registre à décalage avec rétroaction linéaire

Cette méthode, connue dans la littérature anglophone sous le nom de *linear feedback shift register*, est une généralisation de la méthode récursive. Elle se fonde sur une relation du type :

$$x_n = a_t x_{n-1} + a_{t-1} x_{n-2} + \cdots + a_1 x_{n-t-1} \mod 2,$$

où $a_i \in \{0,1\}$. Le polynôme $p(u) = 1 + a_1 u + \cdots + a_t u^t$ s'appelle *polynôme de rétroaction*, et il est appliqué de manière récursive à partir des valeurs $x_0, \ldots, x_{t-1} \in \{0,1\}$, qui constituent l'*initialisation du registre*.

Il est clair que si l'initialisation est $x_i = 0$ pour $i = 0, \ldots, t-1$, ce procédé ne produira qu'une suite de 0. Il y a donc $2^t - 1$ initialisations non banales possibles, et le cycle de nombres pseudo-aléatoires qui est généré a une longueur qui ne dépasse pas $2^t - 1$. Toutefois, si le polynôme p en tant que polynôme de l'anneau $\mathbb{Z}_2[u]$ est un polynôme irréductible et divise $u^{2^t-1} - 1$, le cycle est de longueur maximale $2^t - 1$, et n'importe quelle initialisation non nulle produit ce cycle à partir de positions (*seeds*) différentes. Les nombres aléatoires u_i sont produits alors en utilisant les termes de la suite par groupes non enchevauchés de m éléments.

Notons que l'introduction d'un bit supplémentaire de capacité de mémoire augmente la longueur maximale potentielle du cycle d'un facteur 2. Cette méthode est particulièrement adaptée quand il s'agit d'être utilisée en informatique grâce à l'arithmétique binaire qui la gère et à sa relative simplicité.

3.6 L'évolution des générateurs

De nombreux générateurs de nombres aléatoires utilisés en informatique sont construits à l'aide de la méthode de congruence. Par exemple, le générateur RANDU, utilisé pendant des décennies par IBM, était défini par les paramètres $m = 2^{31}$, $a = 2^{16} + 3$ et $b = 0$. Dans ce cas on peut montrer que $x_{i+1} =$

$(6x_i - 9x_{i-1}) \bmod m$ et que connaissant x_i et x_{i-1} on peut prédire x_{i+1} par une relation linéaire. Pour cette raison RANDU a été remplacé par le générateur de paramètres $m = 2^{31} - 1$, $a = 7^5$ et $b = 0$. Ce choix était utilisé sur la majorité des générateurs implémentés dans les ordinateurs IBM 360-370. Les valeurs $a = 630\,360\,016$, $a = 397\,204\,094$ et $a = 764\,261\,123$, ou encore $a = 314\,159\,269$ et $b = 453\,802\,245$, proposées dans la littérature donnent aussi des générateurs dont les suites possèdent les propriétés statistiques des suites de nombres pseudo-aléatoires. Voir, par exemple, Payne, Rabung, Bogyo (1969), Hoaglin (1976), Fishman (1978), Knuth (1981).

Plus récemment Matsumoto et Nishimura (1998) ont proposé un algorithme fondé sur un type particulier de registre à décalage à rétroaction. Ce générateur, baptisé *Mersenne twister* a des propriétés qui en font aujourd'hui l'un des générateurs de nombres pseudo-aléatoires les plus appréciés : les nombres à 32 bits sont distribués de manière uniforme sur 623 dimensions ; sa complexité en fait l'un des plus rapides et la longueur du cycle généré est de $2^{19\,937} - 1$. Avec une telle longueur de cycle, si l'on avait eu la possibilité de faire tourner tous les ordinateurs de dernière génération depuis la naissance de l'univers (il y a 13,7 milliards d'années), ils ne seraient pas arrivés à compléter un seul cycle !

3.7 Le nombre π comme générateur naturel de nombres aléatoires

Le nombre π (de la lettre grecque π, à prononcer « pi ») suscite l'intérêt des mathématiciens depuis environ 4 000 ans, c'est-à-dire depuis que les Babyloniens et les Égyptiens ont découvert que le rapport entre la circonférence et le diamètre d'un cercle est le même pour tous les cercles. Ce nombre, défini comme étant ce rapport constant, a suscité un intérêt grandissant auprès d'un grand nombre de mathématiciens, qui ont tenté, en vain, de percer certains de ses mystères. Il est fascinant de voir à quel point tous les peuples et toutes les époques se sont intéressés à π et à ses décimales et comment ils s'y sont pris pour l'approcher. Petr Beckmann dans son excellent livre intitulé *A history of pi* (1971) écrit à propos des décimales de π : « Les décimales de π au-delà des quelques premières n'ont aucune valeur pratique ni scientifique. Quatre décimales sont largement suffisantes pour la construction de la machine la plus fine ; dix décimales suffiraient pour obtenir la circonférence de la terre au pouce près si la terre était une sphère parfaite... ».

L'une des raisons qui motiva le calcul des décimales de π fut peut-être l'espoir d'y déceler un cycle, ce qui aurait impliqué que π est le rapport de deux nombres entiers, autrement dit un nombre rationnel. Or, en 1761, le mathématicien suisse Jean Henri Lambert démontra l'irrationalité de π, et ainsi que les décimales de π ne constituent pas une suite périodique. Un siècle plus tard Lindemann (1882) démontra que π est transcendant, c'est-à-dire qu'il n'est pas solution d'une équation polynomiale à coefficients dans $\mathbb{Z}$ de degré quelconque. Tout laisse supposer que la suite des décimales de π et que son comportement

est complètement désordonné. Si tel est le cas, on pourrait proposer d'utiliser la suite de ces décimales comme liste de nombres aléatoires entre 0 et 9.

Les choix optimaux de paramètres appliqués aux générateurs de nombres pseudo-aléatoires présentés plus haut font encore aujourd'hui l'objet de nombreuses recherches (L'Écuyer, 2006). Malgré ces recherches, ce genre de générateurs n'est pas satisfaisant, car les suites de nombres qu'ils produisent ne sont pas aléatoires, mais possèdent des structures qui ne sont pas toujours perceptibles à l'œil nu. On découvrira par exemple que telle suite contient trop de 1 pour être une suite de nombres aléatoires, que dans telle autre, il n'y a jamais de 2 après un 7, ou jamais la séquence 345 après la séquence 678, ou encore, si l'on considère les séquences de 5 chiffres consécutifs, qu'on ne trouve jamais (ou pas assez souvent, ou trop souvent) trois fois le même chiffre, etc. De nombreux tests statistiques ont été mis au point afin de déceler ce genre d'anomalies qui trahissent le manque d'aléatoire. Chaque fois qu'un nouveau générateur de nombres pseudo-aléatoires est introduit sur le marché, il se trouve quelqu'un pour en dénoncer les faiblesses.

Les expériences du passé ont montré qu'un générateur de nombres pseudo-aléatoires peut passer un test ou une série de tests sans que le générateur soit un bon générateur. Marsaglia en 1968 prouva que si l'on représente les résultats de beaucoup de générateurs de nombres aléatoires de l'époque dans un espace euclidien, (par exemple en représentant $(x_1, x_2, x_3), (x_2, x_3, x_4), \dots$) les points tendent à se distribuer sur un réseau plutôt qu'à se distribuer uniformément. Ensuite il développa toute une série de tests, connue sous le nom de Diehard tests, censés contrôler et le cas échéant détecter ces effets. Il publia même un CDROM de nombres aléatoires qui passent le test Diehard.

La suite des décimales de π ne semble pas présenter ce genre d'inconvénients. Citons à ce propos Gardner (1966) : « On a soumis jusqu'ici la suite de décimales de π à tous les tests statistiques qui pouvaient en montrer le caractère aléatoire. C'est un peu déconcertant pour ceux qui pensent qu'il devrait exister un rapport un peu moins irrégulier entre le diamètre et le périmètre d'une courbe aussi belle que le cercle mais la plupart des mathématiciens pensent qu'on ne trouvera jamais la moindre régularité ni aucun ordre dans le développement décimal de π. »

Si les mathématiciens sont un peu désarçonnés par cette constatation, en revanche les statisticiens ne peuvent que s'en réjouir. Cette absence d'ordre que décrit Gardner est justement ce qu'ils recherchent en matière de suites de nombres aléatoires. À ce propos, il est étonnant de constater que, dans ce cas, l'intérêt des mathématiciens, à la recherche de l'ordre, et celui des statisticiens, à la recherche du désordre, divergent radicalement. Si les mathématiciens ne savent que faire des milliards de décimales, les statisticiens pourraient les stocker sur un support numérique et les utiliser comme source naturelle de nombres aléatoires, comme le propose Dodge (1996). Cela aurait le double avantage de proposer une liste de nombres aléatoires apparemment sans défaut et utilisable par tous. On pourrait se demander pourquoi utiliser π au lieu de e ou $\sqrt{2}$. Un argument supplémentaire en faveur de π a été fourni récemment par

Dodge et Melfi (2005) : π ne montre pas seulement un développement décimal (ou aussi binaire, hexadécimal et sous toute autre base) ayant un caractère aléatoire, mais aussi un développement en fraction continue

$$\pi = [3; 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, \ldots]$$

tout aussi aléatoire, les coefficients étant en accord avec la distribution de Khinchin (1964). Selon cette distribution, la probabilité d'avoir un 1 est d'environ 41,5 %, d'avoir un 2 d'environ 17 %, et en général la probabilité d'avoir le coefficient k est

$$p_k = \frac{\log\left(1 + \dfrac{1}{k^2 + 2k}\right)}{\log 2}.$$

Ainsi les efforts des générations de mathématiciens qui ont étudié cette suite de décimales n'auraient pas été vains et, 4 000 ans après son apparition, π pourrait passer du statut de constante mathématique à celui de variable aléatoire.

Exercices

3.1 Soit $x_0 = 17$.

(a) Choisir a, b, m tels que la relation de récurrence

$$x_{i+1} = ax_i + b \bmod m$$

permette de générer des nombres pseudo-aléatoires entre 0 et 999 avec un cycle de longueur maximale 1 000.

(b) Calculer ensuite les 20 premiers éléments.

3.2 Soient $a = 11$, $b = 10$, $x_0 = 1$ et $x_{n+1} = ax_n + b$. La suite des nombres entre $0,00$ et $0,99$ que l'on obtient en prenant la partie décimale de $x_n/100$ vous paraît-elle aléatoire ? Justifiez votre réponse.

Utilisez le théorème de Hull et Dobell pour réaliser une suite de 20 nombres pseudo-aléatoires entre 0,00 et 0,99 en utilisant la méthode de congruence.

3.3 Utilisez la méthode du carré médian pour obtenir 10 nombres pseudo-aléatoires entre 0,000 et 0,999 en utilisant $x_0 = 625$.

(a) La suite ainsi obtenue vous satisfait-elle ? Que remarquez-vous ?

(b) Utilisez maintenant $x_0 = 999$. Que remarquez-vous ?

(a) Utilisez $x_0 = 157$. Avez-vous des critiques à faire à la suite ainsi obtenue ?

(a) Suggérez des conditions minimales pour que la suite générée soit satisfaisante.

3.4 Générer 18 nombres aléatoires entre 0 et 9 :

(*a*) En utilisant la méthode de congruence avec retard où $x_1 = 1$, $x_2 = 3$ et $x_{n+1} = x_n + x_{n-1} \bmod 10$.

(*b*) En utilisant la méthode de congruence où $x_1 = 1$, et $x_{n+1} = x_n + 3 \bmod 10$.

(*c*) En utilisant la méthode de congruence avec mélange sur la suite générée au point précédent, en utilisant les 6 permutations sur 3 éléments dans l'ordre suivant : $(1,2,3) \to (1,2,3)$, $(1,2,3) \to (1,3,2)$, $(1,2,3) \to (2,1,3)$, $(1,2,3) \to (2,3,1)$, $(1,2,3) \to (3,1,2)$, $(1,2,3) \to (3,2,1)$.

(*d*) Établir un classement de qualité entre les suites ainsi générées.

3.5 On veut simuler 12 issues d'un lancement de dé non biaisé à 6 faces (dé classique).

(*a*) Indiquer une manière d'utiliser correctement des nombres aléatoires uniformes entre 000 et 999 comme ceux qui sont donnés en annexe pour simuler les scores du dé.

(*b*) Utiliser la méthode de Hull et Dobell de manière à ce que la période du cycle généré soit 6 000.

Chapitre 4

Transformations de variables et simulation d'échantillons

4.1 Transformations de variables

Dans le chapitre précédent nous avons vu comment simuler des suites de nombres aléatoires distribués de manière uniforme sur un intervalle. Ceux-ci sont à la base de toute simulation. Ces suites sont en effet manipulées afin de produire des suites de nombres susceptibles de simuler des données de situations bien déterminées.

Les transformations de variables nous permettront de simuler des échantillons fictifs d'une variable aléatoire à partir d'un ensemble de nombres aléatoires uniformément distribués sur $[0, 1]$. Dans certains cas, par exemple quand on peut expliciter complètement la fonction de répartition, pour atteindre ce but il suffira d'appliquer simplement une formule de transformation. Dans d'autres cas une formule de transformation sera combinée avec une procédure de sélection opportune.

À la fin du chapitre nous verrons à travers le rééchantillonnage, comment simuler un échantillon d'une variable aléatoire dont on ne connaît pas la loi de distribution mais seulement un petit échantillon.

4.1.1 Variables aléatoires discrètes

Cas univarié

Commençons l'explication de cette notion par un exemple :

Soit X une variable aléatoire de Poisson

$$\Pr(X = x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!} & \text{pour } x = 0, 1, 2, \ldots \\ \\ 0 & \text{sinon.} \end{cases}$$

Notons A l'ensemble des valeurs de X à probabilité non nulle. Soit Y une nouvelle variable définie par $Y = 4X$.

L'ensemble des valeurs possibles de Y, noté B, obtenu par la transformation $y = 4x$, est donné par $B = \{0, 4, 8, 12, \ldots\}$. La transformation $y = 4x$ fait correspondre à chaque point de A un et un seul point de B. Inversement, par la transformation $x = \frac{1}{4}y$, à chaque point de B correspond un et un seul point de A. La transformation $y = 4x$ est donc une bijection. Puisqu'il y a correspondance bijective entre A et B, l'événement $Y = y$ ou $4X = y$ peut survenir uniquement lorsque l'événement $X = \frac{1}{4}y$ a lieu. Ainsi :

$$\Pr(Y = y) = \begin{cases} \Pr\left(X = \dfrac{1}{4}y\right) = \dfrac{\lambda^{y/4} e^{-\lambda}}{(y/4)!} & \text{avec } y = 0, 4, 8, 12, \ldots \\ \\ 0 & \text{sinon.} \end{cases}$$

Passons à présent au cas général. Soit X une variable aléatoire discrète. Notons par A l'ensemble des valeurs possibles de X, c'est-à-dire celles à probabilité non nulle. Soit $y = g(x)$ une transformation bijective qui applique A sur B. Si nous résolvons $y = g(x)$ en termes de y, disons $x = h(y)$, pour chaque $y \in B$ nous avons $x = h(y) \in A$. Par conséquent, les événements $Y = y$ (où $g(X) = y$) et $X = h(y)$ sont équivalents en termes de probabilité. La loi de probabilité de Y est déterminée par :

$$\Pr(Y = y) = \begin{cases} \Pr\left(X = h(y)\right) & y \in B \\ \\ 0 & \text{sinon.} \end{cases}$$

Exemple 4.1 *Soit X une variable aléatoire binomiale de paramètres $\left(n; \frac{1}{3}\right)$*

$$\Pr(X = x) = \begin{cases} \dfrac{n!}{x!(n-x)!} \left(\dfrac{1}{3}\right)^x \left(\dfrac{2}{3}\right)^{n-x} & \textit{avec } x = 0, 1, 2, \ldots n \\ \\ 0 & \textit{sinon.} \end{cases}$$

Cherchons la loi de probabilité de la variable aléatoire $Y = X^2$. La transformation $y = g(x) = x^2$ applique $A = \{x; x = 0, 1, 2, \ldots, n\}$ sur $B = \{y; y = 0, 1, 4, \ldots, n^2\}$. En général, $y = x^2$ ne définit pas une transformation bijective ; ici, cependant, c'en est une, car il n'y a pas de valeurs négatives de x dans $A = \{x; x = 0, 1, 2, \ldots, n\}$. Donc nous avons la fonction inverse unique $x = h(y) = \sqrt{y}$ (et non $-\sqrt{y}$) et ainsi :

$$\begin{aligned}\Pr(Y=y) &= \Pr(X^2=y)=\Pr(X=\sqrt{y})\\ &= \frac{n!}{(\sqrt{y})!(n-\sqrt{y})!}\left(\frac{1}{3}\right)^{\sqrt{y}}\left(\frac{2}{3}\right)^{n-\sqrt{y}}, \quad y=0,1,4,\ldots,n^2.\end{aligned}$$

Cas bivarié

Soit $p(x_1,x_2)$ la loi de probabilité conjointe de deux variables aléatoires discrètes X_1 et X_2 avec A l'ensemble (à deux dimensions) de points pour lesquels $p(x_1,x_2)>0$. Soit g définie par $g(x_1,x_2)=(g_1(x_1,x_2),g_2(x_1,x_2))$ une transformation bijective qui applique A sur B, un autre ensemble à deux dimensions. Il existe donc une fonction $h=(h_1(y_1,y_2),h_2(y_1,y_2))$ de B en A telle que $h(g(x_1,x_2))=(x_1,x_2)$. La loi conjointe des deux nouvelles variables $Y_1=g_1(X_1,X_2)$ et $Y_2=g_2(X_1,X_2)$ est donnée par :

$$p(y_1,y_2)=\begin{cases} p\left(h_1(y_1,y_2),h_2(y_1,y_2)\right) & \text{si} \quad (y_1,y_2)\in B\\ 0 & \text{sinon.}\end{cases}$$

Exemple 4.2 *Soient X_1 et X_2 deux variables aléatoires indépendantes qui suivent la loi de Poisson avec des moyennes μ_1 et μ_2 respectivement. La loi de probabilité conjointe de X_1 et X_2, $p(x_1,x_2)$, est donnée par :*

$$p(x_1,x_2)=\begin{cases} \dfrac{\mu_1^{x_1}\mu_2^{x_2}e^{-\mu_1-\mu_2}}{x_1!x_2!} & x_1=0,1,2,\ldots \quad x_2=0,1,2,\ldots\\ 0 & \textit{sinon.}\end{cases}$$

Par conséquent, l'espace A est l'ensemble des points (x_1,x_2) où chaque x_1,x_2 est un entier non négatif. Nous voulons trouver la loi de probabilité de $Y_1=X_1+X_2$. Si nous utilisons la technique du changement de variables, nous devons définir une seconde variable aléatoire Y_2. Puisque Y_2 est sans intérêt pour nous, choisissons-la de telle sorte que nous ayons une transformation bijective simple. Par exemple, prenons $Y_2=X_2$. Alors $y_1=x_1+x_2$ et $y_2=x_2$ représentent une transformation bijective qui applique A sur $B=\{(y_1,y_2);y_2\leq y_1\}$.

Les fonctions inverses sont données par $x_1=y_1-y_2$ et $x_2=y_2$. Par conséquent, la loi conjointe de Y_1 et Y_2 est :

$$\Pr(Y_1=y_1,Y_2=y_2)=\frac{\mu_1^{y_1-y_2}\mu_2^{y_2}e^{-\mu_1-\mu_2}}{(y_1-y_2)!y_2!}, \qquad (y_1,y_2)\in B.$$

La loi conjointe de Y_1 et Y_2 permet par simple somme d'obtenir la loi de $Y_1=X_1+X_2$.

4.1.2 Variables aléatoires continues

Cas univarié

Soit X une variable aléatoire du type continu, ayant comme fonction de densité :

$$f(x) = \begin{cases} 2x & 0 < x < 1 \\ \\ 0 & \text{sinon.} \end{cases}$$

Ici A, l'espace où $f(x) > 0$ est l'intervalle $]0, 1[$. Définissons la variable aléatoire Y par $Y = 8X^3$. L'ensemble A est appliqué sur l'ensemble $B = \{y; 0 < y < 8\}$ et, de plus, la transformation est bijective. La transformation inverse de $y = 8x^3$ est $x = h(y) = \frac{1}{2}\sqrt[3]{y}$. Nous savons que $\Pr(a \leq Y \leq b)$ avec $0 < a < b < 8$ est égale à $\int_a^b g(y)dy$, où $g(y)$ est la densité de probabilité de Y. Calculons donc $\Pr(a \leq Y \leq b)$:

$$\begin{aligned} \Pr(a \leq Y \leq b) &= \Pr(a \leq 8X^3 \leq b) \\ &= \Pr\left(\frac{1}{2}\sqrt[3]{a} \leq X \leq \frac{1}{2}\sqrt[3]{b}\right) \\ &= \int_{\frac{1}{2}\sqrt[3]{a}}^{\frac{1}{2}\sqrt[3]{b}} f(x)dx = \int_{\frac{1}{2}\sqrt[3]{a}}^{\frac{1}{2}\sqrt[3]{b}} 2xdx. \end{aligned}$$

De plus comme $x = \frac{1}{2}y^{\frac{1}{3}}$, et $dx = \frac{1}{6}y^{-\frac{2}{3}}dy$, on a, par transformation de variable d'une intégrale :

$$\begin{aligned} \Pr(a \leq Y \leq b) &= \int_a^b 2\frac{1}{2}y^{\frac{1}{3}}\frac{1}{6}y^{-\frac{2}{3}}dy \\ &= \int_a^b \frac{1}{6}y^{-\frac{1}{3}}dy. \end{aligned}$$

Comme l'égalité ci-dessus est vraie quel que soit $0 < a < b < 8$, la densité de probabilité $f_Y(y)$ de Y est :

$$f_Y(y) = \begin{cases} \frac{1}{6}y^{-\frac{1}{3}} & \text{si } 0 < y < 8 \\ \\ 0 & \text{sinon.} \end{cases}$$

Soit X une variable aléatoire du type continu avec une fonction de densité $f_X(x)$. Soit A l'espace unidimensionnel où $f_X(x) > 0$. Considérons la variable

aléatoire $Y = g(x)$ où $y = g(x)$ définit une transformation bijective qui applique l'ensemble A sur l'ensemble B. Soit la transformation inverse de $y = g(x)$, notée par $x = h(y)$, et soit $h'(y)$ la dérivée de $h(y)$ par rapport à y continue et non nulle pour tous les points y dans B. Alors la fonction de densité de la variable aléatoire $Y = g(X)$ est donnée par :

$$f_Y(y) = \begin{cases} f_X(h(y)) \cdot |h'(y)| & y \in B \\ \\ 0 & \text{sinon} \end{cases}$$

où $|h'(y)|$ représente la valeur absolue de $h'(y)$. C'est exactement ce que nous avons fait ci-dessus.

Exemple 4.3 *Soit X une variable aléatoire distribuée de manière uniforme sur $[0, 1]$. Sa densité $f(x)$ est :*

$$f(x) = \begin{cases} 1 & si \;\; 0 < x < 1 \\ 0 & sinon. \end{cases}$$

Cherchons la densité de la variable aléatoire $Y = -2 \log X$. Ici la transformation est $y = g(x) = -2 \log x$ est telle que $x = h(y) = e^{-y/2}$. L'espace A est $A = \{x; 0 < x < 1\}$, qui, par l'application de la transformation bijective $y = -2 \log x$, devient $B = \{y; 0 < y < \infty\}$. Comme $h'(y) = -\frac{1}{2} e^{-y/2}$, la fonction de densité $f_Y(y)$ de $Y = -2 \log X$ est :

$$f_Y(y) = \begin{cases} f_X(e^{-y/2})|h'(y)| = \dfrac{1}{2} e^{-y/2} & 0 < y < \infty \\ \\ 0 & sinon. \end{cases}$$

Remarquons que pour la transformation $y = -\frac{1}{\lambda} \log x, \lambda > 0$, de manière analogue au cas ci-dessus, on obtient $f_Y(y)$ la densité de $Y = -\frac{1}{\lambda} \log X$

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y} & 0 < y < \infty \\ \\ 0 & sinon. \end{cases}$$

La variable aléatoire Y est donc distribuée selon une loi exponentielle de paramètre λ.

Cet exemple montre que si l'on possède un générateur de nombres aléatoires uniformes sur $]0, 1[$, on peut, par une transformation logarithmique, générer des nombres aléatoires distribués selon la loi exponentielle de paramètre λ.

Ce résultat peut aussi être obtenu au moyen du théorème suivant :

Théorème 4.1 (Théorème de la transformation inverse) *Soit X une variable aléatoire et $F(x)$ sa fonction de répartition.*

Si $F(x)$ est continue et strictement croissante, alors $Y = F(X)$ est une variable aléatoire dont la distribution est uniforme sur $[0, 1]$.

Dit autrement, si U est une variable aléatoire uniforme sur $[0, 1]$, la variable $F^{-1}(U)$ est une variable aléatoire dont F est sa fonction de répartition.

Preuve :

$$y = F(x) = \Pr(X \leq x) = \Pr(F(X) \leq F(x)) = \Pr(Y \leq y) = F_Y(y),$$

ce qui prouve le théorème. ■

Une application directe de ce théorème permet, à partir d'une loi uniforme, de générer n'importe quelle autre loi. Ainsi, pour générer un échantillon fictif issu d'une variable aléatoire X, il faut connaître $F(x)$ et avoir une suite de nombres $y_1, y_2, \ldots, y_n$ issus d'une variable U uniforme sur $[0, 1]$. L'égalité $x = F^{-1}(y)$ permet d'obtenir l'échantillon

$$x_1 = F^{-1}(y_1),\ x_2 = F^{-1}(y_2),\ \ldots,\ x_n = F^{-1}(y_n),$$

issu d'une population distribuée selon F. La seule difficulté est celle de calculer F^{-1} en connaissant F. Dans certains cas cela n'est pas possible de manière explicite, et dans ces cas d'autres techniques, que l'on verra plus loin, permettent de simuler un échantillon de X.

Si X suit une loi exponentielle négative de moyenne égale à 1, alors sa fonction de répartition est :

$$F(x) = 1 - e^{-x}.$$

Pour appliquer le théorème de la transformation inverse, nous devons chercher la fonction inverse de $u = 1 - e^{-x}$. Cela donne $x = -\log(1 - u)$.

Remarquons que si U est une variable aléatoire uniforme, $(1 - U)$ est aussi une variable aléatoire uniforme. Donc, si U est une variable aléatoire uniforme, alors

$$X = -\log U$$

est une variable aléatoire de loi exponentielle négative (de moyenne 1), conformément à ce que l'on avait trouvé dans l'Exemple 4.3. En appliquant ce qui précède à une suite de nombres aléatoires u, on trouve un échantillon issu de X :

u	x
0,213 175	1,545 64
0,135 927	1,995 64
0,908 940	0,095 48
0,961 277	0,039 49
0,043 313	3,139 30
0,737 556	0,304 41
0,977 600	0,022 65
0,183 427	1,695 94
0,785 606	0,241 30
0,506 813	0,679 61
$\vdots$	$\vdots$

Tab. 4.1. *Des nombres aléatoires distribués de manière uniforme* $U(0,1)$ *génèrent des nombres aléatoires distribués selon une loi exponentielle négative de paramètre* 1 *selon la transformation* $x_i = -\log(u_i)$.

D'autres applications du théorème de la fonction inverse sont présentées plus loin dans cet ouvrage.

Cas bivarié

Seules des fonctions qui définissent une transformation bijective seront considérées pour l'instant. Soit $g(x_1, x_2) = (g_1(x_1, x_2), g_2(x_1, x_2))$ avec g_1 et g_2 différentiables, une transformation bijective qui applique un ensemble mesurable A (à deux dimensions) dans le plan $O_{x_1x_2}$ sur un ensemble mesurable B (à deux dimensions) dans le plan $O_{y_1y_2}$. Si nous exprimons chaque x_1 et x_2 en termes de y_1 et y_2, nous pouvons écrire $x_1 = h_1(y_1, y_2)$, $x_2 = h_2(y_1, y_2)$, de manière unique pour deux opportunes fonctions h_1 et h_2. Si g est suffisamment régulière, h_1 et h_2 satisfont aussi à des conditions de régularité comme la différentiabilité, et le déterminant du *jacobien* associé à la transformation :

$$\begin{vmatrix} \dfrac{\partial h_1}{\partial y_1} & \dfrac{\partial h_1}{\partial y_2} \\ \dfrac{\partial h_2}{\partial y_1} & \dfrac{\partial h_2}{\partial y_2} \end{vmatrix}$$

ne sera pas 0, et sera noté par le symbole J. Par la suite on supposera que les dérivées partielles de premier ordre sont continues et que le jacobien J n'est pas égal à 0 dans B.

Exemple 4.4 *Soit A le carré unitaire, $A = \{(x_1, x_2) \mid 0 < x_1 < 1,\ 0 < x_2 < 1\}$ représenté sur la Figure 4.1. Déterminons B, l'image de A dans le plan $O_{y_1 y_2}$, par la transformation bijective*

$$\begin{cases} y_1 = g_1(x_1, x_2) = x_1 + x_2 \\ y_2 = g_2(x_1, x_2) = x_1 - x_2 \ . \end{cases}$$

On a donc :

$$\begin{cases} x_1 = h_1(y_1, y_2) = \frac{1}{2}(y_1 + y_2) \\ x_2 = h_2(y_1, y_2) = \frac{1}{2}(y_1 - y_2) \end{cases}$$

et l'image de la droite $x_1 = 1$ qui contient un des segments de A (voir Fig. 4.1) est donnée par $1 = \frac{1}{2}(y_1 + y_2)$, c'est-à-dire la droite d'équation $y_2 = 2 - y_1$ dans le plan $O_{y_1 y_2}$.

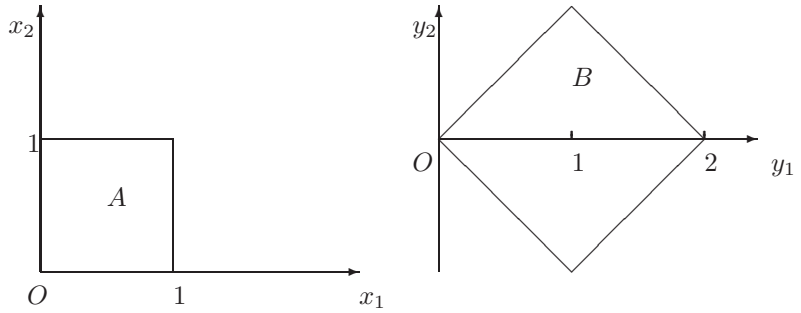

Fig. 4.1. *La région A est transformée dans la région B par la bijection $y_1 = g_1(x_1, x_2)$, $y_2 = g_2(x_1, x_2)$.*

De manière analogue on trouve les images des droites $x_1 = 0$, $x_2 = 0$, et $x_2 = 1$, supports des trois autres côtés du carré A. Les équations des droites support des côtés de B sont donc $y_2 = -y_1$, $y_2 = y_1$, $y_2 = y_1 - 2$, et $y_2 = 2 - y_1$ calculée ci-dessus. Par conséquent, B est le carré représenté dans la Figure 4.1.

Le jacobien de la transformation ci-dessus est égal à :

$$J = \begin{vmatrix} \dfrac{\partial h_1}{\partial y_1} & \dfrac{\partial h_1}{\partial y_2} \\ \dfrac{\partial h_2}{\partial y_1} & \dfrac{\partial h_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} \dfrac{1}{2} & \dfrac{1}{2} \\ \dfrac{1}{2} & -\dfrac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Soient X_1 et X_2 deux variables aléatoires du type continu ayant une fonction de densité conjointe $f_{X_1 X_2}(x_1, x_2)$. Soit A l'ensemble à deux dimensions dans le plan $O_{x_1 x_2}$ où $f_{X_1 X_2}(x_1, x_2) > 0$. Soit $Y_1 = g_1(X_1, X_2)$ et $Y_2 = g_2(X_1, X_2)$ deux variables aléatoires dont on cherche la fonction de densité. Si $y_1 = g_1(x_1, x_2)$ et

$y_2 = g_2(x_1, x_2)$ définissent une transformation bijective régulière de A sur un ensemble B dans le plan $O_{y_1 y_2}$, on cherche la fonction de densité conjointe de Y_1 et Y_2.

Soit A' un sous-ensemble de A et B' l'image de A' par cette transformation bijective. Les événements $(X_1, X_2) \in A'$ et $(Y_1, Y_2) \in B'$ ayant la même probabilité de réalisation, on a :

$$\begin{aligned} \Pr\left((Y_1, Y_2) \in B'\right) &= \Pr\left((X_1, X_2) \in A'\right) \\ &= \int\int_{A'} f_{X_1 X_2}(x_1, x_2) dx_1 dx_2 \ . \end{aligned}$$

Par un changement de variables d'intégration

$$\left\{ \begin{array}{l} y_1 = g_1(x_1, x_2) \\ y_2 = g_2(x_1, x_2) \end{array} \right.$$

où

$$\left\{ \begin{array}{l} x_1 = h_1(y_1, y_2) \\ x_2 = h_2(y_1, y_2) \end{array} \right.$$

on obtient :

$$\int\int_{A'} f_{X_1 X_2}(x_1, x_2) dx_1 dx_2 = \int\int_{B'} f_{X_1 X_2}(h_1(y_1, y_2), h_2(y_1, y_2)) |J| dy_1 dy_2 \ .$$

Par conséquent, pour chaque ensemble B' dans B, on a l'égalité :

$$\Pr\left((Y_1, Y_2) \in B'\right) = \int\int_{B'} f_{X_1 X_2}(h_1(y_1, y_2), h_2(y_1, y_2)) |J| dy_1 dy_2.$$

Donc la fonction de densité conjointe $f_{Y_1 Y_2}(y_1, y_2)$ de Y_1 et Y_2 est :

$$f_{Y_1 Y_2}(y_1, y_2) = \left\{ \begin{array}{ll} f_{X_1 X_2}(h_1(y_1, y_2), h_2(y_1, y_2)) |J| & (y_1, y_2) \in B \\ \\ 0 & \text{sinon.} \end{array} \right.$$

Exemple 4.5 *Soient X_1 et X_2 deux variables aléatoires indépendantes distribuées uniformément sur $]0,1[$. La densité conjointe de X_1 et X_2 est alors :*

$$f(x_1,x_2)=\begin{cases}1 & si \quad 0<x_1<1, \quad 0<x_2<1\\ 0 & sinon.\end{cases}$$

Considérons les deux variables aléatoires $Y_1=X_1+X_2$ et $Y_2=X_1-X_2$. Nous cherchons la fonction de densité conjointe de Y_1 et de Y_2. Ici, l'espace à deux dimensions dans le plan $O_{x_1x_2}$ est celui de l'Exemple 4.4. La transformation bijective $y_1=x_1+x_2$, $y_2=x_1-x_2$ applique A sur l'espace B du même exemple. De plus, le jacobien de cette transformation est $J=-1/2$. Par conséquent :

$$g(y_1,y_2)=f\left(\frac{1}{2}(y_1+y_2),\frac{1}{2}(y_1-y_2)\right)|J|$$

qui est équivalent à

$$g(y_1,y_2)=\begin{cases}\dfrac{1}{2} & si \quad (y_1,y_2)\in B\\ \\ 0 & sinon.\end{cases}$$

Exemple 4.6 *Soient $X_1,\ X_2$ deux variables aléatoires indépendantes ayant comme fonction de densité :*

$$f_{X_i}(x)=\begin{cases}e^{-x} & 0<x<\infty\\ 0 & sinon.\end{cases}$$

Soient $Y_1=X_1+X_2$ et $Y_2=X_1/(X_1+X_2)$. Nous allons montrer que Y_1 et Y_2 sont indépendantes. Puisque la fonction de densité conjointe de X_1 et X_2 est :

$$f_{X_1X_2}(x_1,x_2)=\begin{cases}f_{X_1}(x_1)f_{X_2}(x_2)=e^{-x_1-x_2} & 0<x_{1<}\infty,\ 0<x_2<\infty\\ \\ 0 & sinon\end{cases}$$

l'espace A est le premier quadrant du plan $O_{x_1x_2}$, non compris les points sur les axes. Maintenant

$$y_1=g_1(x_1,x_2)=x_1+x_2$$

$$y_2=g_2(x_1,x_2)=\frac{x_1}{(x_1+x_2)}$$

peut s'écrire $x_1=y_1y_2$, $x_2=y_1(1-y_2)$ de sorte que :

$$J=\begin{vmatrix}y_2 & y_1\\ 1-y_2 & -y_1\end{vmatrix}=-y_1\ .$$

La transformation est bijective et elle applique A sur $B = \{(y_1, y_2); 0 < y_1 < \infty,\ 0 < y_2 < 1\}$ dans le plan $O_{y_1 y_2}$. La fonction de densité conjointe de Y_1 et Y_2 est alors :

$$f_{Y_1 Y_2}(y_1, y_2) = \begin{cases} y_1 e^{-y_1} & 0 < y_1 < \infty,\ 0 < y_2 < 1 \\ 0 & \textit{sinon.} \end{cases}$$

À noter que Y_2 est uniforme sur l'intervalle $[0, 1]$.

Le théorème de la transformation inverse donne la justification des méthodes qui génèrent une suite de nombres pseudo-aléatoires selon une fonction de répartition F donnée. On a vu que quand la fonction de répartition F est inversible, un échantillon fictif compatible avec la fonction de répartition F se simule à travers une suite de nombres pseudo-aléatoires $u_1, \ldots, u_n$ avec la simple transformation $x_i = F^{-1}(u_i)$, pour $i = 1, \ldots, n$. Dans de nombreux cas, quand pour une raison ou pour une autre cette procédure n'est pas applicable, il existe toutefois des alternatives fondées, elles aussi, sur l'existence de nombres pseudo-aléatoires. Ces méthodes sont notamment adaptées pour simuler des échantillons d'une distribution dont la fonction de répartition n'est pas inversible avec l'aide des fonctions classiques.

4.2 Génération de nombres aléatoires suivant une loi normale

Dans ce paragraphe nous allons voir comment générer des variables aléatoires qui suivent une loi normale. Nous verrons aussi comment générer des paires de variables normales corrélées.

Rappelons que pour une variable aléatoire U uniformément distribuée dans l'intervalle $[0, 1]$, on a $E(U) = \frac{1}{2}$ et $V(U) = \frac{1}{12}$. Pour un échantillon aléatoire $\{u_1, \ldots, u_n\}$ de réalisations d'une loi uniforme, chaque élément u_i suit une loi $U_i \sim U(0, 1)$, et l'espérance de la somme de n variables aléatoires uniformes $U(0, 1)$ est l'espérance de $T = \sum_{i=1}^{n} U_i$ qui est égale à $\frac{n}{2}$. Sa variance est égale à $\frac{n}{12}$. D'après le théorème central limite la distribution asymptotique de $T = \sum_{i=1}^{n} U_i$ suit une loi normale. Pour des valeurs finies de n on obtient de bonnes approximations de la loi normale. Le choix de n dépend évidemment de la qualité de l'approximation désirée. En prenant $n = 12$, la variable

$$Z^* = \sum_{i=1}^{12} \left(U_i - \frac{1}{2} \right)$$

a espérance 0 et variance 1 et donne une bonne approximation de la variable aléatoire normale centrée réduite.

En littérature, il existe plusieurs méthodes pour produire une loi normale à partir de nombres uniformément distribués. Bray et Marsaglia (1964) ont proposé une méthode qui s'adapte à la programmation sur ordinateur et demande un minimum de mémoire.

La méthode de Box et Muller (1958) est une autre méthode pour obtenir des réalisations de variables dont la distribution est normale. Cette méthode est fondée sur la transformation des coordonnées polaires en coordonnées cartésiennes.

Théorème 4.2 (Théorème de Box et Muller) *Soient U_1 et U_2 deux variables aléatoires uniformes et indépendantes sur l'intervalle $[0,1]$. Les variables*

$$\begin{aligned} Z_1 &= \sqrt{-2\log(U_1)}\cos(2\pi U_2) \\ Z_2 &= \sqrt{-2\log(U_1)}\sin(2\pi U_2) \end{aligned}$$

sont alors deux variables aléatoires normales centrées réduites indépendantes.

Preuve : La démonstration est une application instructive de ce que nous avons vu à propos des transformations de variables continues. Soient Z_1 et Z_2 deux variables aléatoires indépendantes avec $Z_i \sim \mathcal{N}(0,1), \quad i = 1,2$. Leur densité conjointe est donnée par :

$$f(z_1, z_2) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right).$$

Dans le plan $\mathbb{R}^2$, le point (z_1, z_2) est situé à une distance $r = \sqrt{z_1^2 + z_2^2}$ de l'origine, et la droite passant par (0,0) et (z_1, z_2) forme un angle θ avec l'axe horizontal, tel que $\tan(\theta) = z_2/z_1$. Le passage des coordonnées cartésiennes (z_1, z_2) aux coordonnées polaires (r, θ) s'effectue par la transformation :

$$\begin{aligned} z_1 &= r\cos\theta \\ z_2 &= r\sin\theta \end{aligned}$$

dont le jacobien J est égal à :

$$J = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r \ .$$

Ainsi :

$$\begin{aligned} f(r,\theta) &= |J| f(z_1, z_2) \\ &= r \cdot \frac{1}{2\pi}\exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right) \\ &= \frac{1}{2\pi} \cdot r\exp\left(-\frac{r^2}{2}\right) \ . \end{aligned}$$

Cela montre que les variables aléatoires $R = \sqrt{Z_1^2 + Z_2^2}$ et Θ (la variable aléatoire définie à partir de Z_1 et Z_2 comme la fonction qui associe à Z_1 et Z_2 la seule valeur dans $[0, 2\pi[$ dont $\cos\Theta$ a le même signe que Z_1 ; $\sin\Theta$ a le même signe de Z_2 et $Z_1 \sin\Theta = Z_2 \cos\Theta$) sont des variables aléatoires indépendantes dont les densités respectives sont :

$$f_R(r) = r \exp\left(-\frac{r^2}{2}\right) \quad \text{pour } -\infty < r < +\infty,$$

$$f_\Theta(\theta) = \frac{1}{2\pi} \quad \text{pour } 0 \leq \theta < 2\pi.$$

En posant $U = \exp(-R^2/2)$, on a, pour $u > 0$:

$$\begin{aligned}
\Pr(U \leq u) &= \Pr(-R^2/2 \leq \log u) \\
&= \Pr(R \geq \sqrt{-2\log u}) \\
&= \int_{\sqrt{-2\log u}}^{\infty} r \exp(-r^2/2) dr \\
&= -\exp(-r^2/2)\big|_{\sqrt{-2\log u}}^{\infty} \\
&= \exp(-(\sqrt{-2\log u})^2/2) \\
&= \exp(\log u) \\
&= u.
\end{aligned}$$

On a ainsi transformé (Z_1, Z_2), deux variables aléatoires normales indépendantes, en une paire de variables uniformes indépendantes. Les transformations inverses permettent donc de simuler deux échantillons indépendants d'une loi normale à partir de deux échantillons indépendants issus d'une loi uniforme. ■

Cette méthode a l'avantage de ne pas nécessiter l'inverse de la fonction de répartition de la loi normale centrée réduite Φ ni la fonction de répartition elle-même. Dans le cas de la simulation d'une loi normale, la méthode de l'inversion est d'application difficile, car l'inverse de Φ ne s'explicite pas par des fonctions élémentaires, et la fonction de répartition de la loi normale centrée réduite Φ s'exprime seulement à travers des fonctions élémentaires.

La qualité des générateurs des nombres pseudo-aléatoires pour simuler U_1 et U_2 influence naturellement la qualité de la simulation de loi normale. En utilisant la méthode de congruence $x_i = (x_{i-1} + b) \bmod m$ dans le pire des cas par exemple avec $m = 8$, $x_0 = 1$, $a = 5$, $b = 1$ et $m = 256$, $x_0 = 3$,

$a = 5$, $b = 2$, la représentation graphique des paires (z_1, z_2) simulées pour $(Z_1; Z_2)$ et données dans la Figure 4.2 (a) montre clairement que la propriété d'indépendance du théorème de Box et Muller n'est pas satisfaite, en dépit du fait que, par le théorème de Hull et Dobell, les deux générateurs pour U_1 et U_2 ont un cycle de longueur maximale ! La même remarque s'impose pour le choix $m = 512$, $x_0 = 1$, $a = 5$, $b = 1$ et $m = 512$, $x_0 = 1$, $a = 9$, $b = 1$ illustré dans la Figure 4.2 (b). Cette régularité disparaît dès que la longueur du cycle est plus élevée que le nombre de points à simuler. C'est le cas de la Figure 4.2 (c) où les points sont simulés en utilisant un générateur pour U_1 et U_2 fondé sur la méthode de congruence avec $m = 65\,536$, $a = 5$, $b = 1$, $x_0 = 10\,000$ pour U_1 et $x_0 = 20\,000$ pour U_2. Dans la Figure 4.2 (d), les deux suites U_1 et U_2 sont déterminées avec la méthode de congruence avec retard : la suite U_1 est définie par la relation de récurrence $x_n = x_{n-1} + x_{n-2} + x_{n-3} + 1 \bmod 512$ avec $x_0 = 10$, $x_1 = 20$, $x_2 = 30$. Pour la suite U_2 la même relation de récurrence est utilisée, mais avec $x_0 = 5$, $x_1 = 15$, $x_2 = 25$.

La génération d'une loi normale bivariée

La simple transformation $Y = \sigma X + \mu$ permet de générer une variable $\mathcal{N}(\mu, \sigma^2)$ à partir d'une variable $X = \mathcal{N}(0, 1)$. En suivant le même principe, l'application

$$\begin{cases} Y_1 = & \sigma_1 X_1 + \mu_1 \\ \\ Y_2 = & \sigma_2 \varrho X_1 + \sigma_2 \sqrt{1 - \varrho^2} X_2 + \mu_2 \end{cases}$$

transforme une loi normale bivariée dont les composantes X_1 et X_2 sont normales centrées réduites indépendantes en une loi bivariée où les composantes Y_1 et Y_2 sont distribuées selon une loi normale d'espérance respectivement μ_1 et μ_2, de variance respectivement σ_1^2 et σ_2^2, et dont le coefficient de corrélation est ϱ. Le même principe peut être généralisé au cas d'une loi normale multidimensionnelle.

4.3 La méthode du rejet

Cette méthode de simulation d'échantillons s'applique aux variables aléatoires continues X non nulles sur un intervalle de $\mathbb{R}$. Dans le cas où la densité f est définie sur un intervalle $[a, b]$, il est possible de représenter la surface entre le graphe de la densité et l'axe des x à l'intérieur d'un rectangle, de choisir au hasard un point dans ce rectangle et de déterminer s'il est situé au-dessus ou au-dessous de la courbe $y = f(x)$. Cette idée, banale en soi, est la clé de la méthode du rejet décrite ci-après.

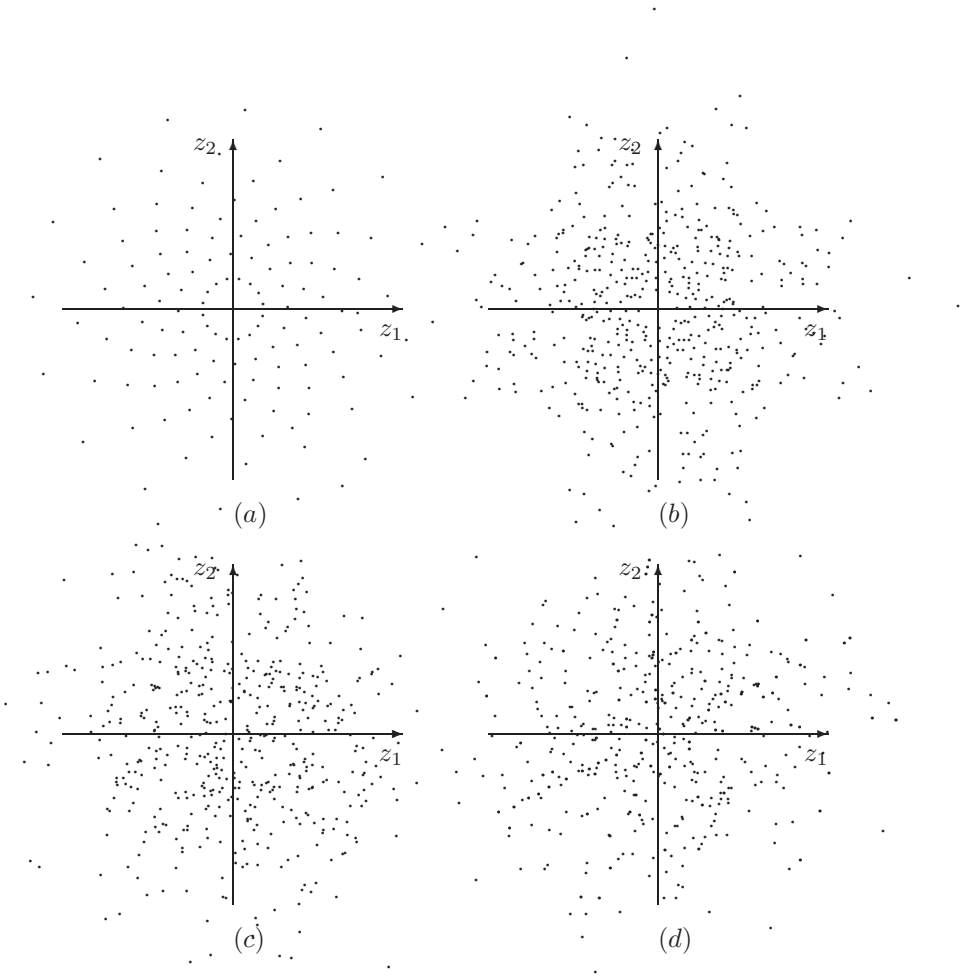

Fig. 4.2. *Simulation de* 4 *paires de lois normales. La méthode de Box et Muller peut produire des distributions trop régulières (a) et (b). Cette méthode produit des distributions aléatoires normales indépendantes seulement si les nombres pseudo-aléatoires utilisés sont de qualité suffisamment bonne, comme dans (c) et (d). Pour le détail des générateurs utilisés, voir le Paragraphe* 4.2.

Soit X une variable aléatoire dont la densité $f(x)$ a un support contenu dans l'intervalle $[a, b]$ et dont la valeur maximale est égale à m. Considérons le rectangle $R = [a, b] \times [0, m]$ et U_1 et U_2 deux variables aléatoires indépendantes et uniformes sur $[0, 1]$. Les coordonnées $(a + (b - a)U_1; mU_2)$ sont réparties de manière uniforme dans le rectangle R. Les points situés au-dessus de $f(x)$ sont rejetés et ceux situés au-dessous de $f(x)$ sont acceptés. Pour chaque paire $(u_1; u_2)$ tirée de (U_1, U_2) les valeurs simulées de X seront donc données par les valeurs de $a + (b - a)u_1$ pour lesquelles $f(a + (b - a)u_1) \leq mu_2$. La probabilité qu'un point arbitraire soit situé sous la courbe $y = f(x)$ est donnée par le rapport de l'aire comprise sous la courbe et de l'aire du rectangle, c'est-à-dire $1/m(b - a)$. On voit que pour a et b fixés, plus m est petit plus le nombre de points rejetés sera petit. L'algorithme est donc d'autant plus performant que m est petit.

Ce principe peut se généraliser, en remplaçant le rectangle par une surface délimitée par le graphe d'une fonction non négative opportunément choisie. Pour les densités à longues queues il est préférable d'utiliser une *courbe enveloppante* $g(x)$ de forme semblable à $f(x)$ et par conséquent distincte d'un rectangle. On choisira dans la mesure du possible une densité $g(x)$ facile à simuler, de forme semblable à $f(x)$ définie sur le même intervalle et on définira la courbe enveloppante par $k \cdot g(x)$, avec $kg(x) \geq f(x)$ pour tout x où f (et g) est définie. Notons qu'il n'est plus nécessaire que f soit définie sur un rectangle. Puisque $g(x) \neq f(x)$ et les deux fonctions sont des densités, il existe des points où $g(x) < f(x)$. Si

$$k = \sup \frac{f(x)}{g(x)},$$

on a $k > 1$ et $kg(x) \geq f(x)$. Cette valeur de k sera aussi la plus petite possible.

Les points à rejeter seront ceux compris entre $f(x)$ et $kg(x)$, nettement moins nombreux que ceux situés dans le rectangle au-dessus de $f(x)$ (si un tel rectangle existe).

L'algorithme est donc le suivant :

Algorithme du rejet avec courbe enveloppante

(*a*) Initialisation : $n := 1$.

(*b*) Générer une issue u_n tirée d'une distribution ayant densité $g(x)$ et une issue v_n uniforme dans $[0, kg(u_n)]$.

(*c*) Si $v_n < f(u_n)$ alors $x := u_n$, et x est une valeur retenue.

(*d*) $n := n + 1$ et retourner au pas *(b)*.

La probabilité d'obtenir un point situé sous la courbe $f(x)$ est

$$p = \frac{\int_{-\infty}^{\infty} f(x)dx}{\int_{-\infty}^{\infty} kg(x)dx} = \frac{1}{k}.$$

Le nombre N d'itérations nécessaires pour simuler n issues de X suit une loi binomiale négative d'espérance $E(N) = kn$.

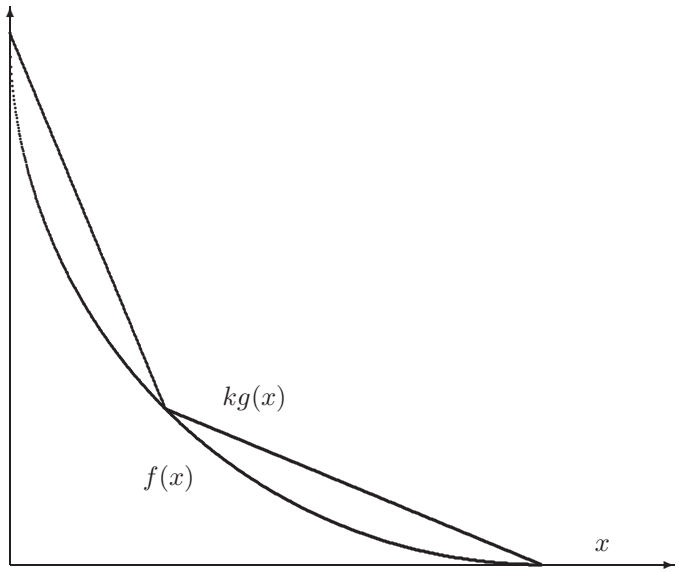

Fig. 4.3. *La courbe enveloppante $kg(x)$ est composé ici de 2 segments. Les points simulés entre le graphe de $f(x)$ et la courbe enveloppante sont rejetés.*

La méthode du rejet avec courbe enveloppante permet de simuler des variables aléatoires continues définies sur $\mathbb{R}$ en évitant que la probabilité de rejet soit proche de 1.

Par exemple, pour simuler une variable aléatoire normale centrée réduite Z par la méthode du rejet, on choisira comme courbe enveloppante $g(x) = e^{-x}$ pour $x > 0$ et $k = \sqrt{e/2\pi} \simeq 0{,}657\,7$. À noter qu'ici $k < 1$. Puisque f, la fonction de densité de Z, est symétrique, on simulera seulement des valeurs positives d'une distribution normale, pour ensuite les étaler de manière aléatoire de part et d'autre de l'axe de symétrie. Cette variante de la méthode entraîne seulement $k > 1/2$.

On génère u_1 et u_2 issues de U_1 et U_2 deux variables aléatoires indépendantes uniformes dans $[0, 1]$. Par la transformation $X = -\log U_1$ on obtient une variable aléatoire distribuée selon la loi exponentielle de densité $g(x)$; donc $x = -\log u_1$ est issu de X. Ensuite on obtient une variable aléatoire par la transformation $Y = U_2 \cdot kg(X)$. Si $y = u_2 \cdot kg(x) < f(x)$, la valeur x de la variable X est retenue et on pose $w = x$, sinon elle est rejetée et l'algorithme

recommence la boucle par la génération d'une nouvelle paire issue de (U_1, U_2). L'ensemble des valeurs w retenues est issu d'une variable instrumentale W. On définira Z à l'aide d'une variable uniforme $U_3 \sim U(0,1)$ selon la règle :

$$Z = \begin{cases} W & \text{si } U_3 \leq \frac{1}{2} \\ -W & \text{si } U_3 > \frac{1}{2} \ . \end{cases}$$

Une simulation complète est détaillée dans l'Exemple 4.7.

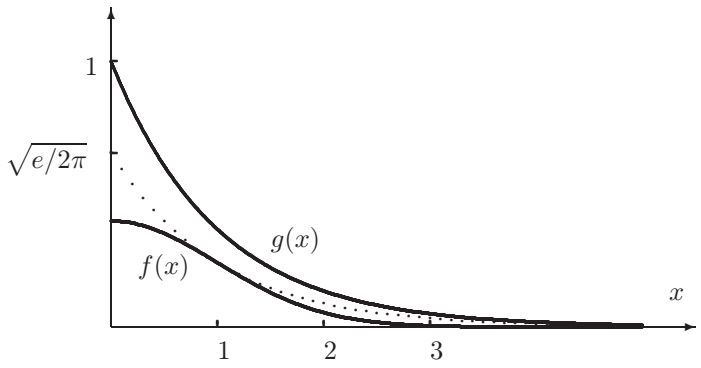

Fig. 4.4. *Ici la fonction f représente la distribution normale centrée réduite pour $x > 0$ et $g(x) = e^{-x}$. La courbe enveloppante est pointillée.*

Exemple 4.7 *Soit $f(x)$ la densité d'une distribution normale centrée réduite, et $g(x) = e^{-x}$ pour $x > 0$. Une application de la méthode du rejet est utilisée pour simuler des réalisations de la loi normale en accord avec la procédure décrite plus haut. Dans la simulation montrée dans la Table* 4.2*, sur* 30 *paires $(u_1; u_2)$,* 26 *valeurs de x correspondantes ont été acceptées :*

U_1	U_2	U_3	X	Y	$f(X)$	acc./rej.	Z
0,593	0,290	0,805	0,522	0,113	0,348	1	−0,522
0,533	0,770	0,036	0,627	0,270	0,327	1	0,627
0,797	0,569	0,268	0,226	0,298	0,388	1	0,226
0,546	0,146	0,891	0,603	0,052	0,332	1	−0,603
0,197	0,406	0,862	1,622	0,052	0,106	1	−1,622
0,283	0,177	0,483	1,260	0,033	0,180	1	1,260
0,895	0,480	0,084	0,110	0,283	0,396	1	0,110
0,086	0,519	0,949	2,445	0,029	0,020	0	–
0,475	0,625	0,734	0,743	0,195	0,302	1	−0,743
0,039	0,527	0,075	3,240	0,013	0,002	0	–
0,191	0,818	0,838	1,652	0,103	0,101	0	–
0,874	0,395	0,596	0,134	0,227	0,395	1	−0,134
0,214	0,111	0,459	1,537	0,015	0,122	1	1,537
0,965	0,515	0,995	0,035	0,327	0,398	1	−0,035
0,112	0,150	0,610	2,185	0,011	0,036	1	−2,185
0,809	0,169	0,560	0,210	0,090	0,390	1	−0,210
0,288	0,754	0,786	1,243	0,143	0,184	1	−1,243
0,252	0,153	0,260	1,377	0,025	0,154	1	1,377
0,744	0,473	0,135	0,295	0,231	0,381	1	0,295
0,358	0,932	0,752	1,024	0,220	0,236	1	−1,024
0,535	0,412	0,806	0,624	0,145	0,328	1	−0,624
0,237	0,875	0,640	1,439	0,136	0,141	1	−1,439
0,645	0,810	0,676	0,438	0,344	0,362	1	−0,438
0,184	0,500	0,544	1,688	0,060	0,095	1	−1,688
0,836	0,124	0,500	0,178	0,068	0,392	1	−0,178
0,814	0,272	0,471	0,205	0,146	0,390	1	0,205
0,220	0,915	0,246	1,512	0,132	0,127	0	–
0,858	0,534	0,283	0,152	0,301	0,394	1	0,152
0,519	0,360	0,567	0,654	0,123	0,322	1	−0,654
0,126	0,004	0,561	2,070	0,000	0,046	1	−2,070

Tab. 4.2. *Tableau de simulation pour la génération d'un échantillon en accord avec une loi normale centrée réduite à l'aide de l'algorithme d'acceptation-rejet.*

4.4 La méthode de comparaison

La méthode décrite dans ce paragraphe et que l'on appelle méthode de comparaison permet de simuler des variables aléatoires discrètes à valeurs dans $\mathbb{N}$ ou un sous-ensemble de $\mathbb{N}$.

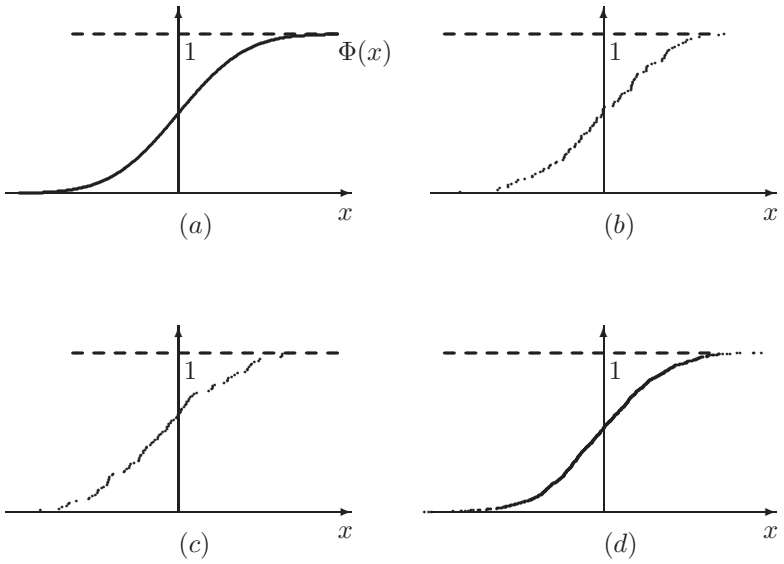

Fig. 4.5. *Représentation graphique des fonctions de répartition des variables aléatoires normales exactes* (a) *; simulées avec* 100 *points par la méthode de Box et Muller* (b) *; simulées avec* 100 *points par la méthode du rejet* (c)*. Quand des simulations plus massives sont effectuées, l'approximation devient meilleure, par exemple en utilisant* 1 000 *points par la méthode de Box et Muller* (d)*.*

Cette méthode est fondée sur la définition des variables aléatoires uniformes. Considérons une variable U uniforme sur $[0,1]$; on a $\Pr(a < U \leq b) = b - a$ pour $0 \leq a < b \leq 1$. Pour simuler une variable aléatoire discrète X satisfaisant $p_i = \Pr(X = i)$, on utilise le fait que

$$\Pr(X = 0) = p_0 = \Pr(0 < U \leq p_0)$$

et

$$\Pr(X = i) = p_i = \Pr\left(\sum_{j=0}^{i-1} p_j < U \leq \sum_{j=0}^{i} p_j\right).$$

La méthode de comparaison est définie de la manière suivante à l'aide d'un échantillon issu d'une variable aléatoire uniforme $U(0,1)$.

$$\begin{aligned} X = 0 \quad &\text{si} \quad 0 < U \leq p_0 \\ X = i \quad &\text{si} \quad \sum_{j=0}^{i-1} p_j < U \leq \sum_{j=0}^{i} p_j \ . \end{aligned}$$

On compare donc la valeur de U avec la loi cumulative de X. Cette méthode est l'équivalent pour les variables discrètes de la méthode de la transformation inverse. Remarquons que i peut varier dans un ensemble fini ($i = 1, \ldots, n$) ou dans un ensemble infini ($i \in \mathbb{N}$).

Exemple 4.8 (La loi de Bernoulli) *Considérons une variable aléatoire de Bernoulli de paramètre p. Puisque $\Pr(X = 0) = 1-p$ et $\Pr(X = 1) = p$, la méthode de comparaison donne dans ce cas :*

$$\begin{array}{lll} X = 0 & si & 0 \leq U \leq 1 - p \\ X = 1 & si & 1 - p < U \leq 1 \ . \end{array}$$

Le jet d'une pièce de monnaie, lancée 10 fois, est simulé dans le tableau ci-dessous en posant $X = 0$ (pile) et $X = 1$ (face) avec $p = \frac{1}{2}$.

i	U	X
1	0,17	0
2	0,46	0
3	0,83	1
4	0,20	0
5	0,98	1
6	0,47	0
7	0,10	0
8	0,43	0
9	0,75	1
10	0,64	1

Tab. 4.3. *Simulation du lancer d'une pièce de monnaie.*

Remarquons que par comptage du nombre de 1 on obtient une réalisation d'une variable binomiale.

Exemple 4.9 (La loi binomiale) *Considérons une variable aléatoire binomiale de paramètre $p = \frac{1}{4}$ et $n = 5$. La loi cumulative est donnée dans le tableau :*

i	0	1	2	3	4	5
$\Pr(X \leq i)$	0,237 3	0,632 8	0,896 5	0,984 4	0,999 0	1

Tab. 4.4. *Probabilités associées à une variable binomiale de paramètres $p =$ 0,25 et $n = 5$.*

Par comparaison des réalisations d'une variable uniforme (simulée ou non) avec le tableau ci-dessus on génère des réalisations d'une variable binomiale avec $n = 5$ *et* $p = \frac{1}{4}$ *:*

U	X
0,463 2	1
0,671 2	2
0,820 1	2
$\vdots$	$\vdots$
0,174 1	0
0,999 2	5

Tab. 4.5. *Simulation d'un échantillon compatible avec une loi binomiale où* $p = 0{,}25$ *et* $n = 5$.

4.5 L'échantillonneur de Gibbs

L'échantillonneur de Gibbs est une manière de générer des distributions de deux (ou plusieurs) variables à partir d'un modèle qui définit les distributions de probabilité conditionnelles. Dans le cas d'un modèle à deux variables, la méthode consiste à prendre un élément de départ (x_0, y_0) et à générer à l'aide de nombres aléatoires les éléments d'un échantillon fictif (x_n, y_n) par itération en choisissant dans l'ordre x_n d'une variable aléatoire de densité $f_{X|Y=y_{n-1}}$, et y_n d'une variable aléatoire de densité $g_{Y|X=x_n}$, où $f_{X|Y=y_{n-1}}$ et $g_{Y|X=x_n}$ sont les distributions de probabilité conditionnelles supposées connues ou modélisées.

Pour le cas d'un modèle à plusieurs variables où les distributions conditionnelles seraient connues, l'élément au départ serait $(x_1^0, x_2^0, \ldots, x_k^0)$ et l'échantillon est construit en simulant $(x_1^n, x_2^n, \ldots, x_k^n)$ en choisissant dans l'ordre

x_1^n d'une variable aléatoire de densité $f_{X_1|X_2=x_2^{n-1},\ldots,X_k=x_k^{n-1}}$;

x_2^n d'une variable aléatoire de densité $f_{X_2|X_1=x_1^n,X_3=x_3^{n-1},\ldots,X_k=x_k^{n-1}}$;

x_3^n d'une variable aléatoire de densité $f_{X_3|X_1=x_1^n,X_2=x_2^n,X_4=x_4^{n-1},\ldots,X_k=x_k^{n-1}}$

et ainsi de suite jusqu'à

x_k^n d'une variable aléatoire de densité $f_{X_k|X_1=x_1^n,\ldots,X_{k-1}=x_{k-1}^n}$.

L'échantillon ainsi généré sera une simulation conforme aux probabilités conditionnelles imposées et la distribution se conformera d'autant plus aux probabilités conditionnelles que le nombre d'itérations sera élevé.

Un point est donc généré à partir du point précédent, sans utiliser l'information sur les autres points générés : il s'agit donc d'une chaîne de Markov. L'échantillonneur de Gibbs est un cas particulier de l'algorithme de Metropolis-Hastings (1953), qui génère une suite de réalisations et applique le principe de l'acceptation-rejet sous une forme plus générale que celle montrée au Paragraphe 4.3, fondé sur une densité de distributions donnée.

Exemple 4.10 *Une distribution de points sur la portion de plan*

$$Q = \{(x, y),\ x > 0, y > 0\}$$

est telle que la distribution des x*, pour* y *fixé suit une loi exponentielle de paramètre* y *et que la distribution des* y *pour* x *fixé suit une loi exponentielle de paramètre* x*. Un échantillonneur de Gibbs peut être utilisé pour simuler un échantillon de* 10 *points de* Q *conforme aux distributions conditionnelles données.*

Les distributions de répartition conditionnelles peuvent s'écrire comme suit :

$$F(x \mid y) = \int_0^x ye^{-ty}dt; \qquad G(y \mid x) = \int_0^y xe^{-tx}dt.$$

Donc à l'aide d'un échantillonneur de Gibbs si u_n *et* v_n *sont des nombres aléatoires issus d'une loi uniforme* $U(0, 1)$*, l'échantillon peut être calculé à l'aide des relations*

$$x_{n+1} = -\frac{1}{y_n}\log u_n; \qquad y_{n+1} = -\frac{1}{x_{n+1}}\log v_n.$$

Par exemple en utilisant les nombres aléatoires

u_n	v_n
0,864 289	0,550 899
0,004 736	0,406 955
0,975 364	0,976 381
0,082 672	0,695 312
0,253 519	0,212 330
0,159 700	0,140 474
0,593 694	0,440 392
0,900 385	0,931 119
0,941 360	0,586 161
0,360 221	0,972 795

Tab. 4.6. *Nombres aléatoires utilisés pour simuler la distribution bivariée avec un échantillonneur de Gibbs.*

à partir du point $(1, 1)$ *on obtient les* 10 *points* $(0,145; 4,087)$*,* $(1,309; 0,686)$*,* $(0,036; 0,657)$*,* $(3,788; 0,095)$*,* $(14,308; 0,108)$*,* $(16,938; 0,115)$*,* $(4,499; 0,182)$*,* $(0,575; 0,123)$*,* $(0,487; 1,095)$*,* $(0,931; 0,029)$*, représentés dans la Figure* 4.6*.*

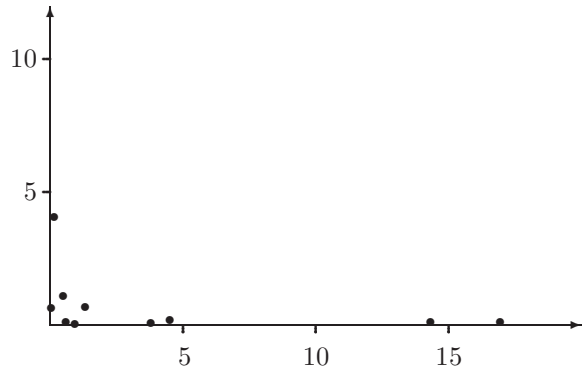

Fig. 4.6. *Simulation d'une distribution bivariée avec un échantillonneur de Gibbs.*

4.6 L'algorithme de Metropolis-Hastings

Introduit dans sa forme la plus simple par Metropolis en 1953, et généralisé par Hastings en 1970, l'algorithme qui porte aujourd'hui leur nom est une procédure permettant de simuler un échantillon d'une distribution, univariée ou multidimensionnelle, de variables aléatoires. Il demande que la fonction de densité, p, soit connue, même seulement à une constante près. La procédure génère un échantillon sous la forme d'une chaîne de Markov ergodique, dont la loi stationnaire est p. C'est une application d'une méthode générale connue sous le nom de *Monte Carlo Markov Chain* (abrégé en MCMC dans la littérature). Il est utilisé notamment pour l'estimation d'intégrales particulièrement complexes (voir aussi le Chapitre 7).

L'algorithme s'appuie sur une distribution de proposition, donnée sous la forme d'une distribution conditionnelle $q(\cdot|x)$, dont le support doit inclure celui de p, et depuis laquelle pour tout x il est facile de simuler des échantillons fictifs de $q(\cdot|x)$. En voici l'énumération en toutes ses étapes :

Étape 1 : Poser $n = 0$ et choisir $x^{(0)}$ dans le domaine de p.

Étape 2 : Générer y de la distribution dont la loi est $q(\cdot|x^{(n)})$.

Étape 3 : Générer u de $U(0,1)$.

Étape 4 : On définit $x^{(n+1)}$ comme suit :

$$x^{(n+1)} = \begin{cases} y & \text{si } u < \dfrac{p(y)}{p(x^{(n)})} \cdot \dfrac{q(x^{(n)}|y)}{q(y|x^{(n)})} \\ x^{(n)} & \text{autrement.} \end{cases}$$

Étape 5 : Poser $n = n + 1$ et revenir à l'Étape 2.

Donc dans un échantillon construit avec l'algorithme de Metropolis-Hastings les éléments de la suite générée ne sont pas indépendants, et peuvent contenir des valeurs qui se répètent : cela arrive si $x^{(n+1)} = x^{(n)}$ dans l'Étape 4.

À noter aussi que la quantité $\frac{p(y)}{p(x^{(n)})} \cdot \frac{q(x^{(n)}|y)}{q(y|x^{(n)})}$ peut être ≥ 1, ce qui implique l'acceptation de la valeur y proposée dans l'Étape 2.

La distribution de proposition peut être indépendante, c'est-à-dire non conditionnelle. Dans ce cas $q(\cdot|x) = q(\cdot)$. Dans des nombreuses applications elle peut être une marche aléatoire, c'est-à-dire de la forme $q(\cdot|x) = x + \varepsilon$ avec ε indépendant de x, comme $\varepsilon \sim \mathcal{N}(0, 1)$. La distribution de proposition q peut avoir une forme symétrique ($q(y|x) = q(x|y)$), comme dans le cas d'une marche aléatoire fondée sur une loi normale. Dans ce cas l'expression de l'Étape 4 se simplifie en $u < p(y)/p(x^{(n)})$.

La quantité $p(y)/p(x^{(n)})$ n'est rien d'autre que le rapport de vraisemblance, et la généralisation $\frac{p(y)}{p(x^{(n)})} \cdot \frac{q(x^{(n)}|y)}{q(y|x^{(n)})}$ est appelée rapport de Hastings. Le fait que la densité p apparaît dans le rapport de Hastings au numérateur et au dénominateur facilite l'usage de l'algorithme, notamment quand p est connue seulement à une constante près.

À noter finalement que puisque $x^{(0)}$ est choisi de manière arbitraire, les premiers éléments de la suite générée constituent l'allumage (*burn in* en anglais) de la chaîne et en général ne doivent pas être considérés dans la suite finale. Comme pour l'échantillonneur de Gibbs, l'algorithme de Metropolis-Hastings produit seulement asymptotiquement des échantillons avec les qualités désirées.

Exemple 4.11 *Soit* $p(x, y) = 1 - x^2 - y^2$ *si* $1 - x^2 - y^2 \geq 0$ *et* 0 *autrement. Voici comment générer un échantillon de* 20 *unités distribuées selon une fonction de densité dont on sait qu'elle est proportionnelle à p, en utilisant l'algorithme de Metropolis-Hastings : on choisit* $x^{(0)} = (0, 0)$ *; on choisit une loi uniforme sur* $(-1, 1) \times (-1, 1)$ *comme distribution de proposition. Les coordonnées du point* $y = (y_1, y_2)$ *sur* $(-1, 1) \times (-1, 1)$ *se trouvent sur les colonnes* y_1 *et* y_2 *du Tableau* 4.7*. Le rapport de Hastings est donc simplement le rapport de vraisemblance qui figure sur la dernière colonne. Ainsi, par exemple pour déterminer* $x^{(1)}$*, puisque* $u > p(y)/p(x^{(0)})$*, on a* $x^{(1)} = x^{(0)}$*. Dans cette simulation on a* $u > p(y)/p(x^{(1)})$*,* $u > p(y)/p(x^{(2)})$*,* $u > p(y)/p(x^{(3)})$*, et donc* $x^{(2)} = x^{(3)} = x^{(4)} = (0, 0)$*. À l'itération suivante on a* $u < p(y)/p(x^{(4)})$ *et donc* $x^{(5)} = (y_1^{(4)}, y_2^{(4)})$*. En suivant l'algorithme, l'échantillon final est composé des points dont les coordonnées figurent sur les colonnes* x_1 *et* x_2 *du tableau. À noter que des points sont répétés même* 4 *fois.*

	y_1	y_2	$p(y)$	x_1	x_2	$p(x)$	u	$p(y)/p(x)$
0	0,418 0	0,788 1	0,204 0	0,000 0	0,000 0	1,000 0	0,731	0,204 0
1	0,441 6	−0,852 0	0,079 0	0,000 0	0,000 0	1,000 0	0,526	0,079 0
2	0,993 8	0,910 3	0,000 0	0,000 0	0,000 0	1,000 0	0,075	0,000 0
3	0,355 8	−0,898 1	0,066 7	0,000 0	0,000 0	1,000 0	0,945	0,066 7
4	0,217 1	−0,671 2	0,502 3	0,000 0	0,000 0	1,000 0	0,331	0,502 3
5	0,169 5	−0,417 8	0,796 6	0,217 1	−0,671 2	0,502 3	0,464	1,585 9
6	−0,730 2	−0,365 3	0,333 2	0,169 5	−0,417 8	0,796 6	0,598	0,418 3
7	0,627 0	0,414 1	0,435 3	0,169 5	−0,417 8	0,796 6	0,034	0,546 4
8	0,429 1	−0,790 3	0,191 1	0,627 0	0,414 1	0,435 3	0,728	0,439 1
9	0,203 5	0,215 4	0,912 1	0,627 0	0,414 1	0,435 3	0,778	2,095 4
10	0,477 8	−0,541 2	0,478 7	0,203 5	0,215 4	0,912 1	0,810	0,524 8
11	−0,407 6	−0,220 9	0,785 0	0,203 5	0,215 4	0,912 1	0,811	0,860 6
12	−0,576 1	0,487 9	0,429 9	−0,407 6	−0,220 9	0,785 0	0,953	0,547 7
13	0,728 6	0,100 3	0,459 0	−0,407 6	−0,220 9	0,785 0	0,051	0,584 7
14	0,757 4	0,995 8	0,000 0	0,728 6	0,100 3	0,459 0	0,913	0,000 0
15	−0,290 3	0,156 9	0,891 0	0,728 6	0,100 3	0,459 0	0,067	1,941 1
16	0,302 7	0,823 5	0,230 0	−0,290 3	0,156 9	0,891 0	0,419	0,258 1
17	−0,551 2	−0,670 8	0,246 1	−0,290 3	0,156 9	0,891 0	0,575	0,276 2
18	−0,548 9	−0,601 3	0,337 0	−0,290 3	0,156 9	0,891 0	0,056	0,378 2
19	0,852 4	−0,725 4	0,000 0	−0,548 9	−0,601 3	0,337 0	0,685	0,000 0
20				−0,548 9	−0,601 3	0,337 0		

Tab. 4.7. *Tableau de simulation pour un échantillon fictif dont la loi conjointe est proportionnelle à* $1-x^2-y^2$*, construit à l'aide de l'algorithme de Metropolis-Hastings.*

Dans les Chapitres 6 et 7 nous montrons d'autres applications de l'algorithme de Metropolis-Hastings, permettant en particulier une approche par simulation pour le calcul de certaines intégrales elliptiques.

4.7 Échantillonnage

L'échantillonnage traite, définit et étudie les méthodes de sélection de sous-ensembles de populations dont on désire estimer un certain paramètre, que l'on ne peut pas mesurer sur la population entière.

La population est composée d'un nombre N d'unités définies sans ambiguité, identifiables et indexables (numérotables) de 1 à N.

À chaque unité est associée une valeur d'une variable X. Les données observées sont les valeurs prises par la variable X sur chaque individu d'un échantillon de taille n de la population. La procédure de sélection de n unités de la population, formant un échantillon, est appelée *plan d'échantillonnage* (en anglais *sample design*). Un plan est déterminé par la donnée d'une loi de probabilité définie sur l'ensemble de tous les échantillons. Dans la pratique, un plan d'échantillonnage peut aussi être décrit par une règle de sélection des n unités et non par la loi de probabilité définie sur tous les échantillons.

Plus généralement, le principe de l'échantillonnage est donc d'utiliser les informations d'un échantillon pour en tirer des informations sur la population.

Dans ce paragraphe nous passons en revue les notions de base de l'échantillonnage, pour voir dans le paragraphe suivant le rééchantillonnage, une procédure qui permet d'étudier des propriétés structurelles complexes à partir d'un échantillon même de taille modeste.

4.7.1 Échantillonnage aléatoire sans remise

Soit N la taille d'une population finie à échantillonner. Le plan d'échantillonnage sans remise consiste à sélectionner un échantillon de n unités parmi tous les $\begin{pmatrix} N \\ n \end{pmatrix}$ échantillons, de manière à ce que tous les échantillons aient la même probabilité d'être sélectionnés. Cela équivaut à choisir les n unités les unes après les autres au hasard parmi les unités indexées par $\{1, 2, \ldots, N\}$, et en les sélectionnant pour l'échantillon final seulement si elles n'ont pas encore été choisies.

Ainsi la probabilité que la $k^{\text{ième}}$ unité de la population appartienne à l'échantillon est égale à n/N, et est la même pour chaque unité. La probabilité de sélection d'un échantillon particulier de n unités distinctes est, quant à elle, égale à $\begin{pmatrix} N \\ n \end{pmatrix}^{-1}$. La sélection est généralement faite à l'aide de nombres aléatoires uniformément distribués sur $\{1, \ldots, N\}$.

4.7.2 Échantillonnage aléatoire avec remise

Le plan d'échantillonnage avec remise consiste à former un échantillon comme une liste de n unités de telle sorte que les unités soient choisies les unes après les autres parmi toutes les unités possibles et de manière aléatoire, donc avec probabilité $1/N$, tout en gardant la possibilité de sélectionner éventuellement plusieurs fois la même unité dans l'échantillon. La probabilité qu'une unité appartienne à l'échantillon sélectionné est donc : $p = 1 - (1 - 1/N)^n$. À noter que dans le cas d'un échantillon avec remise il convient de garder les éléments de l'échantillon dans l'ordre dans lequel ils ont été choisis (Särndal, 1992) : un échantillon issu d'un plan d'échantillonnage avec remise *n'est pas* un sous-ensemble de la population. Toutefois il peut déterminer par réduction un sous-ensemble de $m \leq n$ unités distinctes.

Exemple 4.12 *On dispose d'une population de taille* $N = 100$ *dont les unités sont indexées avec les nombres entiers de* 1 *à* 100. *Comment générer un échantillon de taille* 25 *issu d'un plan d'échantillonnage avec remise ? Et sans remise ?*

Pour $N = 100,\ n = 25$, la suite de nombres pseudo-aléatoires de la Table 4.8 détermine l'échantillon qui sera composé des unités indexées dans l'ordre par les nombres 40, 14, 15, 92, 63, 98, 22, 69, 16, 87, 90, 98, 83, 38, 79, 49, 61, 4, 90, 83, 12, 70, 44, 91, 35. Le nombre 0,395 326 simule l'élément 40 ; le nombre 0,132 467 l'élément 14 et ainsi de suite. À noter que certaines unités sont sélectionnées plusieurs fois.

0,395 326	0,892 030	0,113 373
0,132 467	0,976 945	0,697 203
0,148 880	0,829 858	0,433 949
0,916 417	0,370 981	0,905 937
0,622 784	0,788 620	0,346 952
0,976 908	0,486 565	0,301 489
0,218 581	0,603 469	0,819 012
0,687 608	0,036 195	0,660 974
0,158 074	0,894 422	0,867 319
0,867 026	0,821 761	0,091 507

Tab. 4.8. *Nombres aléatoires uniformes en* $[0, 1]$. *Dans l'Exemple* 4.12 *ils sont utilisés pour un plan de sondage sans remise et pour un plan de sondage avec remise.*

Dans le but de générer un échantillon de 25 nombres aléatoires entre 1 et 100 selon un plan d'échantillonnage sans remise, la même suite va générer l'échantillon des unités dont l'indice est donné par les nombres 4, 12, 14, 15, 16, 22, 31, 35, 38, 40, 44, 49, 61, 63, 67, 69, 70, 79, 82, 83, 87, 90, 91, 92. À noter que dans les 25 premiers nombres aléatoires il y avait 2 nombres aléatoires entre $0,82$ et $0,83$; entre $0,89$ et $0,90$ et entre $0,97$ et $0,98$. Cela nous a obligé d'ignorer les doublons pour considérer les 26^{e}, 27^{e} et 28^{e} nombres aléatoires pour compléter l'échantillon.

Si l'on veut obtenir un échantillon de taille 25 d'une population de taille 100, tiré d'un plan d'échantillonnage sans remise en utilisant le logiciel `R`, l'instruction à donner est :

```
sample(c(1 :100),25)
```

Pour un plan d'échantillonnage avec remise :

```
sample(c(1 :100),25,replace=TRUE)
```

4.7.3 La distribution d'échantillonnage d'un estimateur

Pour estimer un paramètre θ associé à une population de taille N, il faut procéder en sélectionnant un échantillon de taille $n \leq N$ d'unités de la population, $(y_1, \ldots, y_n)$. On calcule une approximation du vrai paramètre à partir des individus composant l'échantillon. Cette approximation se calculant sur la base d'un sous-ensemble particulier, il est très probable qu'en sélectionnant un autre échantillon, on obtient une estimation différente. Il s'ensuit que l'estimateur du paramètre inconnu est une variable aléatoire distribuée selon une certaine loi. En tirant beaucoup d'échantillons de la population globale, on obtient des réalisations d'une variable aléatoire, représentées par l'estimateur, que l'on note $\hat{\theta}(y_1, \ldots, y_n)$. Le symbole « $\hat{}$ » sur le θ indique qu'il s'agit d'un *estimateur* de θ. Selon l'aspect auquel on s'intéresse, un estimateur peut être vu comme variable aléatoire ou comme une fonction de n variables.

Distribution d'échantillonnage d'une moyenne

À titre d'exemple, soit $(y_1, \ldots, y_n)$ un échantillon de taille n où chaque élément (ou unité) est indépendant et identiquement distribué selon une loi normale $\mathcal{N}(\mu, \sigma^2)$. Le rapport entre la taille de l'échantillon n et la taille de la population est 0 ou négligeable. Dans cette approche les plans d'échantillonnage avec remise et sans remise coïncident. Rappelons qu'un estimateur est dit *sans biais* lorsque son espérance est égale à la vraie valeur du paramètre pour lequel l'estimateur a été construit. Considérons par exemple la moyenne arithmétique d'un échantillon

$$\overline{y} = \frac{1}{n}(y_1 + \cdots + y_n).$$

Pour les différents échantillons, la fonction $\overline{Y}$ qui associe à chaque échantillon $(y_1, \ldots, y_n)$ la valeur $\overline{y}$, est un estimateur sans biais de la moyenne de la population. En fait, la somme de variables aléatoires normales suit aussi une loi normale : si $y_i \sim \mathcal{N}(\mu, \sigma^2)$,

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}(n\mu, n\sigma^2) \ .$$

En divisant la somme par n, on obtient :

$$\overline{y} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$$

et donc, en tant que variable aléatoire, $E(\overline{Y}) = \mu$. On a ainsi déterminé la loi d'échantillonnage de la moyenne arithmétique pour le cas où l'échantillon provient d'une loi normale.

Supposons que la distribution de l'échantillon soit inconnue. Le théorème central limite (voir Théorème 2.1) permet d'affirmer que, pour n suffisamment grand, la distribution échantillonnale de la moyenne arithmétique est approximativement normale, quelle que soit la distribution de départ. La moyenne de cette distribution échantillonnale sera égale à la moyenne de la population, et sa variance sera la variance de la population divisée par n, σ^2/n.

En général, on a les propriétés suivantes pour la distribution d'échantillonnage des moyennes et pour n suffisamment grand :

(a) soit μ la moyenne de la population globale, et $\mu_{\overline{X}}$ la moyenne de la distribution d'échantillonnage des moyennes. Alors $\mu_{\overline{X}} = \mu$, quelle que soit la distribution de l'échantillon ;

(b) soit σ l'écart-type de la population globale, et désignons par $\sigma_{\overline{X}}$ l'écart-type de la distribution d'échantillonnage des moyennes. Alors, on peut montrer que (Särndal, 1992) :

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \ \frac{N-n}{N-1}.$$

Le terme $(N-n)/(N-1)$ est le facteur correctif utilisé pour une population de taille finie, N. Dans le cas d'une population infinie, le facteur correctif tend vers 1, car

$$\lim_{N \longrightarrow \infty} \frac{N-n}{N-1} = 1$$

d'où $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.

Distribution d'échantillonnage d'une proportion

Si l'on veut estimer la proportion π d'une population ayant une certaine caractéristique, l'estimateur utilisé est le suivant :

$$p = \frac{\text{nombre d'individus ayant la caractéristique}}{\text{nombre d'individus dans l'échantillon}} \ .$$

Pour n assez grand, la distribution d'échantillonnage de p est une loi normale de moyenne μ_p et de variance σ^2_p. On a alors :

(a) $\mu_p = \pi$;

(b) $\sigma^2_p = \dfrac{\sigma^2}{n}\dfrac{N-n}{N-1}$, où $\sigma^2 = \pi(1-\pi)$.

Une question importante associée aux méthodes d'échantillonnage est celle de la taille de l'échantillon. Lorsque l'on estime un paramètre θ d'une population par un estimateur $\hat{\theta}$ on désire que $\hat{\theta}$ soit proche de θ avec une probabilité élevée. Formellement, cela peut s'écrire, pour un nombre réel positif d petit :

$$\Pr(|\hat{\theta} - \theta| > d) < \alpha.$$

Pour tous les cas où $\hat{\theta}$ est un estimateur sans biais de θ, normalement distribué, alors :

$$\frac{\hat{\theta} - \theta}{\sqrt{\text{var } (\hat{\theta})}} \sim \mathcal{N}(0,1)$$

et

$$\Pr\left(|\hat{\theta} - \theta| > z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\text{var } (\hat{\theta})}\right) = \alpha \ ,$$

où $z_{(1-\frac{\alpha}{2})}$ est le $1-\alpha/2$ quantile associé à la loi normale centrée réduite, c'est-à-dire $\Phi^{-1}(1-\alpha/2)$.

Ainsi l'inégalité ci-dessus est satisfaite pour

$$z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\text{var } (\hat{\theta})} \leq d.$$

Donc, lorsque l'erreur d'estimation d est fixée et que var $(\hat{\theta})$ est fonction de n, on résoudra l'inéquation ci-dessus par rapport à n pour déterminer la taille de l'échantillon.

Cela permet de définir des *intervalles de confiance* pour θ :

$$\hat{\theta} - z_{(1-\frac{\alpha}{2})}\sqrt{\text{var }(\hat{\theta})} \leq \theta \leq \hat{\theta} + z_{(1-\frac{\alpha}{2})}\sqrt{\text{var }(\hat{\theta})},$$

où $1-\alpha$ est le *niveau de confiance*, qui dans la littérature est souvent fixé à 95 %.

Exemple 4.13 *Soit $(y_1, \dots y_n)$ un échantillon aléatoire indépendant, identiquement distribué avec $y_i \sim \mathcal{N}(\mu, \sigma^2)$ pour chaque i. La moyenne arithmétique des y_i est une estimation non biaisée de μ :*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

L'estimateur $\hat{\mu}$ est normalement distribué avec $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/n)$ donc

$$\Pr\left(|\hat{\mu} - \mu| > z_{(1-\frac{\alpha}{2})}\ \frac{\sigma}{\sqrt{n}}\right) = \alpha$$

en supposant que le paramètre σ est connu.

Si σ^2 n'est pas connu, il peut être estimé en utilisant un échantillon de taille n par la formule

$$\hat{\sigma}^2 = \frac{\sum(y_i - \overline{y})^2}{n-1},$$

et alors la distribution d'échantillonnage ne suit plus une loi normale, mais une loi de Student avec $n-1$ degrés de liberté (Dodge, 1999). Cette distribution est comparable à une distribution normale pour n grand.

Exemple 4.14 *On veut estimer la proportion de la population ayant une certaine caractéristique, par exemple la proportion de gens chez qui on a décelé certains symptômes d'une maladie. On désigne par « 1 » la présence des symptômes et par « 0 » l'absence. On considère alors l'estimateur :*

$$\hat{p} = \frac{\text{nombre de « 1 » dans l'échantillon}}{\text{taille de l'échantillon}}\ .$$

Lorsque la taille de l'échantillon est suffisamment grande, la distribution d'échantillonnage de $\hat{p}$ suit une loi normale de moyenne μ_p et de variance σ^2 ; p désigne la proportion des individus, dans la population globale, ayant la caractéristique voulue. On a :

$$\sigma_p^2 = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

où σ^2 est la variance correspondant à la population globale, $\sigma^2 = \mu_p(1-\mu_p)$. Ainsi on a :

$$\Pr\left(|\hat{p} - p| > z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}}\right) = \alpha\ .$$

Pour les cas où certaines unités de la population ont des probabilités différentes d'être sélectionnées, le choix se fait par simulation de lois discrètes.

Exemple 4.15 *Supposons que l'échantillonnage s'effectue avec remplacement et qu'à chaque tirage la probabilité p_i de sélectionner la $i^{ième}$ unité de la population est donnée par :*

$$p_i = \begin{cases} \dfrac{3}{4N} & i = 1, \ldots, \dfrac{N}{3} \text{ ou } i = \dfrac{2N}{3} + 1 \ldots N \\ \dfrac{3}{2N} & i = \dfrac{N}{3} + 1, \ldots, \dfrac{2N}{3}. \end{cases}$$

La taille de la population est $N = 120$. On veut simuler un échantillon de 10 unités. Pour un plan d'échantillonnage avec remplacement, on utilise la méthode d'inversion pour les lois discrètes. Ainsi on génère d'abord 10 nombres aléatoires distribués selon $U(0, 1)$. On sélectionne l'unité 1 de la population si le nombre aléatoire est compris entre 0 et $1/160$; l'unité 2 de la population si le nombre aléatoire est compris entre $1/160$ et $1/80$ et ainsi de suite jusqu'à l'unité $N/3 = 40$. Pour i entre 41 et 80, la probabilité est double. Ainsi l'unité 41 est choisie si le nombre aléatoire se situe entre $1/4$ et $1/4 + 1/80$.

Supposons que les nombres aléatoires uniformes dans $[0, 1]$ soient : 0,512 4, 0,862 9, 0,809 0, 0,022 0, 0,392 9, 0,606 8, 0,257 6, 0,511 3, 0,128 5, 0,844 2. Les unités sélectionnées sont alors respectivement les unités : 61, 99, 90, 4, 52, 69, 41, 61, 21 et la 96. À noter que l'unité 61 a été choisie deux fois, ce qui est tout à fait permis, s'agissant d'un échantillonnage avec remise.

4.8 Rééchantillonnage

4.8.1 Le principe

À l'origine les méthodes de rééchantillonnage ont été développées pour estimer le biais et la variabilité d'un estimateur sans faire d'hypothèses sur la distribution de probabilité de la population dont on estime un paramètre.

Le principe de ces méthodes est fondé sur l'utilisation répétée de l'échantillon de départ en le divisant, en supprimant, ou encore en répétant des observations pour obtenir plusieurs valeurs de l'estimateur permettant le calcul empirique d'une moyenne et d'une variance.

Formellement, considérons n variables aléatoires X_i indépendantes identiquement distribuées selon la fonction de répartition $F(x; \theta)$ dépendant du paramètre θ. Les méthodes de rééchantillonnage permettent une étude des propriétés et caractéristiques d'un estimateur $T_n = T_n(X_1, \ldots, X_n)$ de θ. Dans les cas simples, il est possible d'expliciter l'espérance et la variance de T_n et d'en proposer des estimateurs comme il a été fait dans le paragraphe précédent. Dans le cas général, il est impossible d'obtenir la distribution exacte de

T_n si la fonction de répartition F n'est pas connue. On doit donc utiliser des approximations ; les approximations asymptotiques ne sont pas toujours performantes pour des échantillons de taille raisonnable ; des approximations plus satisfaisantes, connues comme approximations d'Edgeworth (Hall, 1992), sont généralement difficiles à calculer ou à utiliser.

Les méthodes de rééchantillonnage offrent une solution alternative dans la recherche d'approximation de la distribution de T_n.

4.8.2 Le bootstrap

Le principe de la méthode bootstrap, introduite par Efron (1979) et rendue populaire par Efron et Tibshirani (1993), est le suivant : à partir d'un échantillon de n variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon la fonction de répartition $F(x, \theta)$, on construit une fonction $\hat{F}_n$ estimant F. Si $\hat{F}_n$ est proche, dans un certain sens, de F on étudie les propriétés de $T_n(X_1, \ldots, X_n)$ estimateur de θ sous la loi $\hat{F}_n$. Plusieurs choix sont possibles pour $\hat{F}_n$, le plus simple étant la fonction de répartition empirique des observations notée F_n. Le choix $\hat{F}_n = F_n$ donne la version du bootstrap la plus étudiée, appelée parfois *bootstrap naïf*.

Le calcul explicite de la distribution de $T_n(X_1, \ldots, X_n)$ sous $\hat{F}_n$ n'étant pas toujours possible, on recourt à la simulation en générant des échantillons indépendants de loi $\hat{F}_n$. Pour le bootstrap naïf, un tirage avec remise de n valeurs dans l'ensemble des observations donne un *échantillon bootstrap* noté $X^* = \{x_1^* \ldots, x_n^*\}$ pour indiquer que ces valeurs ont été obtenues par un tirage avec remise de l'échantillon de départ $X = \{x_1, \ldots, x_n\}$. Les N échantillons bootstrap permettent de calculer N valeurs de la statistique T_n utilisées pour construire une approximation de la distribution de $T_n(X_1, \ldots, X_n)$.

La popularité des méthodes bootstrap est liée aux propriétés de la moyenne bootstrap, généralisées à des statistiques s'exprimant comme des fonctions des différents moments. Il est clair que la moyenne bootstrap, en elle-même, n'est pas utile, mais les résultats obtenus montrent l'intérêt du bootstrap. En effet, il est possible de montrer, sous quelques hypothèses concernant l'existence des moments des variables considérées, que la distribution bootstrap de la moyenne a un comportement comparable à l'approximation asymptotique.

En utilisant les quantiles de la distribution bootstrap, il est possible de construire des intervalles de confiance pour un paramètre θ. En notant F_n^* la distribution bootstrap empirique de $T_n(X_1, \ldots, X_n)$, c'est-à-dire :

$$F_n^*(t) = \frac{1}{N} \sum_{i=1}^{N} h_{T_n(X_i^*)}(t),$$

où $X_i^* = \{x_{i1}^* \ldots, x_{in}^*\}$ est le $i^{\text{ième}}$ échantillon bootstrap et

$$h_{T_n(X^*)}(t) = \begin{cases} 1 & \text{si } t > T_n(X^*) \\ \\ 0 & \text{sinon.} \end{cases}$$

Le α-quantile de cette distribution est défini par :

$$q_n^*(\alpha) = F_n^{*-1}(\alpha)$$

et l'on approximera l'intervalle recherché par $[q_n^*(\alpha/2); q_n^*(1-\alpha/2)]$.

Notons encore que l'on peut aussi construire des tests bootstrap en déterminant les valeurs critiques selon le principe ci-dessus.

Exemple 4.16 *La population de laquelle l'échantillon de taille* 20 *ci-dessous :*

$$51, 45, 49, 66, 53, 41, 58, 56, 60, 63, 75, 89, 73, 84, 66, 85, 73, 71, 78, 65,$$

a été tiré a une distribution inconnue. On veut trouver un intervalle de confiance à 95 *% des moyennes échantillonales (pour échantillons de taille* 20*).*

Pour cela, nous allons *bootstraper* cet échantillon. Cela veut dire que nous allons rééchantillonner cet échantillon en un certain nombre N d'échantillons avec remise de taille 20, à partir desquels nous calculons la variance des moyennes et les intervalles de confiance empiriques à 95 %. Nous allons considérer les cas $N = 50, 200, 1\,000$, et 5 000. Les variances $S^2(\bar{x}^N)$ sont calculées selon la formule usuelle de l'estimateur sans biais d'une variance :

$$S^2(\bar{x}^N) = \frac{\sum_{i=1}^{N} \left(\bar{x}_i^* - \bar{x}^N\right)^2}{N-1},$$

où $\bar{x}_i^*$ est la moyenne de l'$i^{\text{ième}}$ échantillon bootstrap X_i^* et $\bar{x}^N$ est la moyenne de ces moyennes : $\bar{x}^N = \frac{1}{N}\sum \bar{x}_i^*$.

Pour l'estimation des variances on obtient :

N	50	200	1 000	5 000
$S^2(\bar{x}^N)$	9,761 49	9,146 84	8,757 24	8,618 86

Tab. 4.9. *Estimation bootstrap d'une variance selon le nombre N d'échantillons.*

Pour l'estimation des intervalles de confiance $[a, b]$:

N	a	moyenne	b
50	59,35	64,812	70,95
200	59,37	65,043	70,70
1 000	59,40	65,001	70,95
5 000	59,25	65,052	70,80

Tab. 4.10. *Estimation bootstrap des intervalles de confiance à* 95 *% selon le nombre N d'échantillons.*

Ces résultats peuvent être comparés à ceux que l'on obtient par des méthodes de statistique inférentielle classique. Ceux-ci nous donnent une moyenne de l'échantillon de base de $65,05$, et une estimation non biaisée de la variance de la population qui est donnée par

$$\begin{aligned} S^2 = \quad & ((51-65{,}05)^2+(45-65{,}05)^2+(49-65{,}05)^2+(66-65{,}05)^2+ \\ & +(53-65{,}05)^2+(41-65{,}05)^2+(58-65{,}05)^2+(56-65{,}05)^2+ \\ & +(60-65{,}05)^2+(63-65{,}05)^2+(75-65{,}05)^2+(89-65{,}05)^2+ \\ & +(73-65{,}05)^2+(84-65{,}05)^2+(66-65{,}05)^2+(85-65{,}05)^2+ \\ & +(73-65{,}05)^2+(71-65{,}05)^2+(78-65{,}05)^2+(65-65{,}05)^2)/19 \\ = \quad & 184,366. \end{aligned}$$

L'intervalle de confiance pour la moyenne est alors à la troisième décimale près :

$$\left(65{,}05-2{,}086\frac{\sqrt{184{,}366}}{\sqrt{20}}, 65{,}05+2{,}086\frac{\sqrt{184{,}366}}{\sqrt{20}}\right) \simeq (58{,}716\,6; 71{,}383\,4).$$

Comme on le voit, l'intervalle de confiance bootstrap est meilleur.

Exercices

4.1 Comment peut-on générer, à partir de nombres uniformes sur l'intervalle $[0,1]$, 20 réalisations des lois de probabilité suivantes :

- (a) exponentielle de moyenne 30 ;
- (b) uniforme sur l'intervalle $[-1,4]$;
- (c) binomiale avec paramètres $n=4$ et $p=0{,}6$;
- (d) Poisson de paramètre 4.

4.2 La cible d'un jeu de tir est composée de cercles ayant le même centre, de rayon 1, 2 et 3. Si un tireur marque un tir à l'intérieur du cercle central, il marque 10 points ; entre le premier et le deuxième il en marque 5 ; entre le deuxième et le troisième il en marque 2 ; il ne marque pas de points si le tir est à l'extérieur du troisième cercle.

Imaginons un système de coordonnées centré sur la cible, où $(0,0)$ est le point au centre de la cible. Un tireur a à sa disposition 5 tirs pour faire un maximum de points, et les tirs ont des coordonnées (x,y) avec x et y variables aléatoires normales de moyenne 0 et de variance 4. Simuler à l'aide d'une table de nombres aléatoires uniformes 6 jeux de tirs au moyen de la transformation de Box et Muller. Quel est le score moyen ainsi obtenu ?

4.3 Une distribution de points sur la région $R=\{(x,y),\ x>0, y>0\}$ est telle que les variables aléatoires X et Y représentant les deux coordonnées x,y satisfont aux conditions suivantes :

$$f_{X|Y=y}(x) = y^2 e^{-xy^2} \qquad \text{pour } x>0$$

et

$$f_{Y|X=x}(y) = x^2 e^{-x^2 y} \qquad \text{pour } y > 0.$$

À partir d'un point de votre choix (par exemple le point $(x_0, y_0) = (1, 1)$) et à l'aide de nombres aléatoires de votre choix, utilisez un échantillonneur de Gibbs pour simuler un échantillon de 10 points de R conforme aux distributions conditionnelles données.

4.4 Un générateur de nombres binaires aléatoires a donné la suite suivante :

$$1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0.$$

(a) Comment peut-on transformer cette suite en une suite de nombres aléatoires ayant une distribution uniforme dans l'ensemble discret $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$?

(b) Sur la base de ces 20 nombres binaires aléatoires, combien de nombres aléatoires entre 0 et 9 peut-on ainsi obtenir ?

4.5 L'équipe de football A marque en moyenne 2 buts par match. L'équipe B en marque en moyenne 1,5.

(a) Simuler 5 matchs entre A et B.

(b) Exprimer la probabilité que B gagne et calculer une approximation au millième.

4.6 Un générateur de nombres pseudo-aléatoires censé donner des nombres en accord avec la loi uniforme $U(0, 1)$ donne 30 nombres entre 0 et 1 dont la moyenne est 0,439 294. La valeur de la moyenne est-elle suffisamment éloignée de 0,5 pour dire que c'est un mauvais générateur ? Motivez votre réponse.

Chapitre 5

Tests d'hypothèses et nombres aléatoires

5.1 Introduction

En simulation, il est souvent question de créer des échantillons fictifs pour étudier le comportement de certaines variables qu'il serait difficile d'étudier avec une approche déterministe. Les échantillons ainsi créés, chacun suivant une distribution opportune, sont supposés dans un certain sens reproduire la réalité.

Il est donc important pour le chercheur de savoir au préalable à quelles distributions obéissent les variables qui composent un phénomène à simuler. Si, par exemple, on veut simuler une file d'attente à un guichet de banque entre 8 h et 12 h, peut-on vraiment supposer que les clients arrivent de manière aléatoire de sorte que les temps entre une arrivée et la suivante soient distribués selon une loi exponentielle négative ? Ou cette loi, plutôt complexe, demande-t-elle la prise en compte de facteurs contingents comme la cadence des transports publics proches de la banque ? Quoi qu'il en soit, le chercheur, avant d'effectuer une simulation, doit être certain d'utiliser la bonne distribution pour simuler des échantillons fictifs.

Pour décider quelle distribution il va utiliser, le chercheur va se fonder sur des registres ou des fichiers électroniques répertoriant des événements passés, comme les moments exacts des arrivées des clients à un guichet. Ces données sont ensuite utilisées pour en faire un test. Le chercheur va par exemple tester si dans le passé l'ensemble des temps d'interarrivée des clients s'est distribué selon une loi exponentielle négative ou non.

La procédure de test est donc, à l'inverse d'une simulation, un moyen qui part d'échantillons déjà constitués pour arriver aux lois qui en sont à la base. Ces lois sont ensuite utilisées pour des simulations sur un plus long terme, ou sous des conditions que le commanditaire de l'enquête voudrait explorer.

Ce chapitre est consacré aux tests d'hypothèses. Dans un premier temps, nous nous intéresserons à des tests d'équidistribution. Ces tests ont joué un rôle dans la traque aux défauts des premiers générateurs de nombres pseudo-aléatoires. Ensuite nous allons passer en revue des tests plus complexes, comme celui de Kolmogorov-Smirnov ou d'Anderson-Darling qui s'appliquent à des distributions arbitraires.

5.2 Tests d'hypothèses

Dans beaucoup d'investigations statistiques, nous sommes amenés à fixer une valeur préalable d'une caractéristique de la population et à confirmer ou à infirmer cette valeur à l'aide des résultats obtenus à partir d'un échantillon. Un candidat aux élections qui emploie un sondage pour connaître les chances de sa réussite veut, en effet, savoir si la proportion de la population qui votera pour lui dépassera ou non la barre des 50 %. Les résultats obtenus sur l'échantillon lui permettront le cas échéant de confirmer son idée (il bénéficiera de plus de 50 % des voix).

Lorsque les résultats de l'échantillon diffèrent considérablement de la valeur de référence (50 % dans le cas mentionné ci-dessus) il est facile de tirer des conclusions dans un sens ou dans l'autre. Toutefois, il arrive fréquemment que la différence entre les résultats statistiques et la valeur de référence ne soit pas très grande. Dans ce cas, la bonne décision ne s'impose pas d'elle-même. Dans l'exemple précédent, si l'étude sur un échantillon représentatif donne un pourcentage de 70 %, le candidat pourrait, sans plus d'information, conclure qu'il a de fortes chances de bénéficier de plus de 50 % des voix. En revanche, si le pourcentage échantillonnal n'est que de 52 %, la conclusion n'est plus aussi évidente.

Les caractéristiques d'une population sont souvent exprimées en termes de moyenne, de variance ou de pourcentage. Ces paramètres sont du type quantitatif. Les tests d'hypothèses vont nous permettre soit d'accepter l'hypothèse de départ concernant la valeur du paramètre en question, soit de la rejeter. Dans les paragraphes qui suivent nous allons étudier les tests d'hypothèses sur la moyenne d'une population et sur le pourcentage d'individus possédant une caractéristique donnée dans une population.

À titre d'exemple, prenons le cas suivant : nous savons, d'après des études pédagogiques, que, pour une bonne compréhension des matières enseignées, les étudiants de l'université devraient consacrer environ 45 heures de travail par semaine, avec un écart-type de 9 heures, selon les disciplines. La valeur « 45 heures » représente notre hypothèse de départ afin d'examiner si la situation actuelle diffère sensiblement ou non de cette opinion. Nous prenons un échantillon aléatoire de 36 étudiants inscrits l'année considérée à l'université, auxquels nous posons la question : « Combien d'heures par semaine consacrez-vous à vos études ? (cours universitaires et travaux personnels inclus) ». Pour

des questions méthodologiques, la question doit naturellement être la même que celle qui a été posée lors de l'étude pédagogique de référence.

Nous comparons la moyenne de cet échantillon avec l'hypothèse précédente de 45 heures. Si la moyenne d'échantillonnage obtenue est beaucoup plus élevée que 45 heures, nous pourrons être amenés à croire que le nombre d'heures de travail des étudiants est supérieur à 45. Cependant, si la moyenne de l'échantillon n'est que faiblement plus grande, nous ne pourrons pas conclure que le travail des étudiants de cette année est significativement supérieur à la norme, le résultat de l'échantillon pouvant être dû au simple hasard.

5.3 Définitions et rappels

Les hypothèses à tester concernent en général un paramètre de la distribution de la variable aléatoire étudiée. On rencontre deux types d'hypothèses que dans la littérature on dénote par :

H_0 : l'hypothèse nulle
H_1 : l'hypothèse alternative.

Ces deux hypothèses s'excluent l'une l'autre. En général, l'hypothèse nulle, H_0, énonce une propriété supposée être établie sous des conditions ordinaires (par exemple, elle a été vérifiée longtemps dans le passé, ou elle s'est vérifiée beaucoup plus souvent), ou supposée représenter des conditions d'équilibre (par exemple, un pourcentage de 50 % pour l'acceptation d'une réforme par referendum populaire). L'hypothèse alternative représente un événement nouveau, ou un changement qui déséquilibre les acquis de l'hypothèse nulle (par exemple, un rendement accru par rapport au passé pour un produit financier, ou un pourcentage supérieur à 50 % lors d'un referendum populaire ou encore de barrage).

Une procédure de décision (à définir) aboutit à une décision qui peut être correcte ou non. Il existe deux types d'erreurs auxquelles on peut être confrontés lors d'un test d'hypothèses : l'*erreur du type I*, dite aussi *erreur* α ou encore *erreur de première espèce*, est commise quand H_0 est rejetée alors que H_0 est vraie ; l'*erreur du type II*, ou *erreur* β ou *de deuxième espèce*, est commise quand H_0 n'est pas rejetée alors que H_0 est fausse.

L'erreur du type I peut être aisément contrôlée, car si l'on veut que la probabilité de commettre une erreur du type I soit, disons, de 0,05, il est toujours possible de construire un test pour lequel la probabilité d'erreur du type I est 0,05. L'erreur du type II ne peut pas être si facilement contrôlée, car sa probabilité dépend de la nature réelle de l'hypothèse H_1 qui se vérifie. Une erreur du type II est commise quand on accepte H_0 alors que H_1 est vraie. Mais si selon H_1, un certain paramètre, disons μ, est tel que $\mu > 0$, la probabilité d'erreur du type II change selon si $\mu = 1$, $\mu = 2$ et ainsi de suite.

	H_0 est vraie	H_0 est fausse
Non rejet de H_0	Décision correcte : $1-\alpha$	Erreur type II : β
Rejet de H_0	Erreur type I : α	Décision correcte : $1-\beta$

Tab. 5.1. *Véridicité d'une hypothèse et décisions. Les probabilités respectives figurent avec les notations standard utilisées.*

La probabilité de l'erreur du type I est d'habitude dénotée avec α. Celle de l'erreur du type II avec β.

La probabilité de l'erreur du type I est appelée *seuil de signification* du test.

La quantité $1-\beta$, c'est-à-dire le complément à 1 de la probabilité de l'erreur du type II, $1-\beta$, est appelée *puissance du test*.

On peut donc écrire :

$$\Pr(\text{ne pas rejeter } H_0 \mid H_0 \text{ est vraie}) = 1-\alpha.$$

$$\Pr(\text{rejeter } H_0 \mid H_0 \text{ est fausse}) = 1-\beta.$$

On peut montrer que si α augmente, β diminue.

La puissance d'un test est une fonction des différentes valeurs possibles du paramètre selon l'hypothèse H_1 qui se vérifie.

La *statistique du test* est une variable aléatoire, fonction de l'échantillon, et utilisée pour la décision. Sa distribution doit être connue supposant H_0 vraie.

La *région critique* d'un test d'hypothèses est l'ensemble des valeurs de la statistique du test pour lesquelles on rejette H_0. Très souvent on fixe α et ainsi on détermine la région critique. À noter que si α augmente, la région critique est plus étendue.

Si la statistique du test appartient à la région critique, on rejette H_0 ; dans le cas contraire on accepte H_0. Remarquons que la décision dépend du seuil de signification.

Exemple 5.1 *On considère une population normale dont on connaît l'écart-type $\sigma = 20$. Pour effectuer un test d'hypothèses sur la moyenne μ de cette population, on prélève un échantillon de taille $n = 16$.*

Notons $\overline{x}_c$ la moyenne de l'échantillon. Les données à considérer sont donc : $X \sim \mathcal{N}(\mu; \sigma^2)$; $n = 16$; $\sigma = 20$.

(a) Les hypothèses :

$$H_0 : \mu = 750$$

$$H_1 : \mu \neq 750.$$

(b) Statistique du test : $\overline{X}$. Sous l'hypothèse H_0,

$$\overline{X} \sim \mathcal{N}(750; \frac{20^2}{16}) = \mathcal{N}(750; 5^2)$$

et

$$\frac{\overline{X} - 750}{5} \sim \mathcal{N}(0;1).$$

(c) Région (valeur) critique :

$$\begin{aligned} \alpha &= \Pr(\textit{rejeter } H_0 \mid H_0 \textit{ est vraie}) \\ &= \Pr(\textit{rejeter } H_0 \mid \mu = 750) \end{aligned}$$

donc :

$$1 - \alpha = \Pr\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{x}_c < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) .$$

(d) Valeurs critiques pour $\alpha = 0{,}05$:

$$\begin{aligned} 750 + 1{,}96 \cdot 5 &= 759{,}8 \\ 750 - 1{,}96 \cdot 5 &= 740{,}2. \end{aligned}$$

(e) Règle de décision : Si $740{,}2 < \overline{x}_c < 759{,}8$ *alors on accepte* H_0 *; sinon on rejette* H_0.

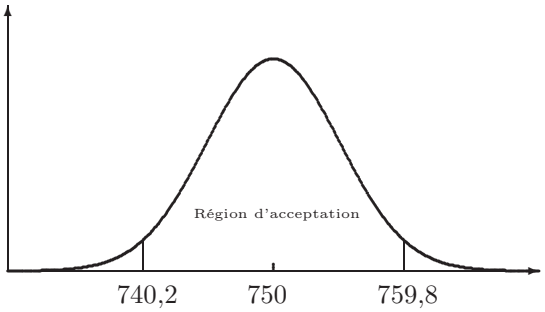

Fig. 5.1. *La région d'acceptation (au seuil de signification* $\alpha = 0{,}05$*) est délimitée par les deux valeurs critiques au-delà desquelles sous l'hypothèse* H_0 *se situe de part et d'autre une probabilité de* $\alpha/2$ *pour la moyenne calculée. La fonction représente la densité de probabilité des valeurs de* $\overline{x}$ *pour échantillons aléatoires de taille* n *sous l'hypothèse* H_0.

5.3.1 Puissance d'un test

Nous avons vu précédemment que la puissance d'un test est définie par la probabilité de rejeter H_0 lorsque H_0 est fausse. Si $\beta = \beta(H_1)$ est la probabilité de l'erreur du type II et qui dépend de la nature de H_1, la puissance est $1 - \beta$. Souvent l'hypothèse H_1 ne représente pas une seule valeur possibile d'un paramètre, mais un ensemble de valeurs alternatives (par exemple $H_1 : \mu > 0$).

La détermination de $1 - \beta$, ou de β, est difficile. Cependant, pour les hypothèses alternatives simples considérées dans ce chapitre, la puissance du test est facilement calculable.

Exemple 5.2 *Suite de l'Exemple* 5.1.

f) Erreur du type II :

$$\begin{aligned}\beta &= \Pr(\textit{accepter } H_0 \mid H_0 \textit{ est fausse}) \\ &= \Pr(740{,}2 < \overline{x} < 759{,}8 \mid H_0 \textit{ est fausse}) \ .\end{aligned}$$

Si par exemple $\mu = 755$:

$$\begin{aligned}\beta &= \Pr(740{,}2 < \overline{x} < 759{,}8 \mid \mu = 755) \\ &= \Pr\left(\frac{740{,}2 - 755}{5} < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < \frac{759{,}8 - 755}{5}\right) \\ &= \Pr\left(-2{,}96 < Z < 0{,}96\right) \\ &= 0{,}998\ 5 - 0{,}168\ 5 \\ &= 0{,}830\ 0 \ .\end{aligned}$$

Quand $\mu = 755$ *on accepte, dans* 83 % *des cas,* H_0 *comme vraie alors qu'elle ne l'est pas.*

g) Puissance du test :

$$1 - \beta = \Pr(\textit{rejeter } H_0 \mid H_0 \textit{ est fausse}) \ .$$

Pour $\mu = 755$ *nous avons* $1 - \beta = 1 - 0{,}830\ 0 = 0{,}170\ 0$. *Cela signifie que ce test reconnaît dans* 17 % *des cas que l'hypothèse* H_0 *est fausse, si la vraie valeur de* μ *est égale à* 755.
De manière analogue, on peut calculer la puissance du test pour d'autres valeurs de μ. *Dans le Tableau* 5.2, *on constate que plus* μ *s'éloigne de* $\mu = 750$, *plus la puissance augmente. Il est évident que plus la différence est grande entre la vraie valeur de* μ *et celle donnée par* H_0, *plus on a de chances de s'en apercevoir.*

α	μ	β	$1 - \beta$
0,05	730	0,020 8	0,979 2
0,05	740	0,484 0	0,516 0
0,05	745	0,829 7	0,170 3
0,05	750		
0,05	755	0,829 7	0,170 3
0,05	760	0,484 0	0,516 0
0,05	770	0,020 8	0,979 2

Tab. 5.2. *Seuil de confiance* (α), *vraie valeur du paramètre testé* (μ), *erreur du type II* (β), *et puissance du test* ($1 - \beta$).

Graphiquement, on a :

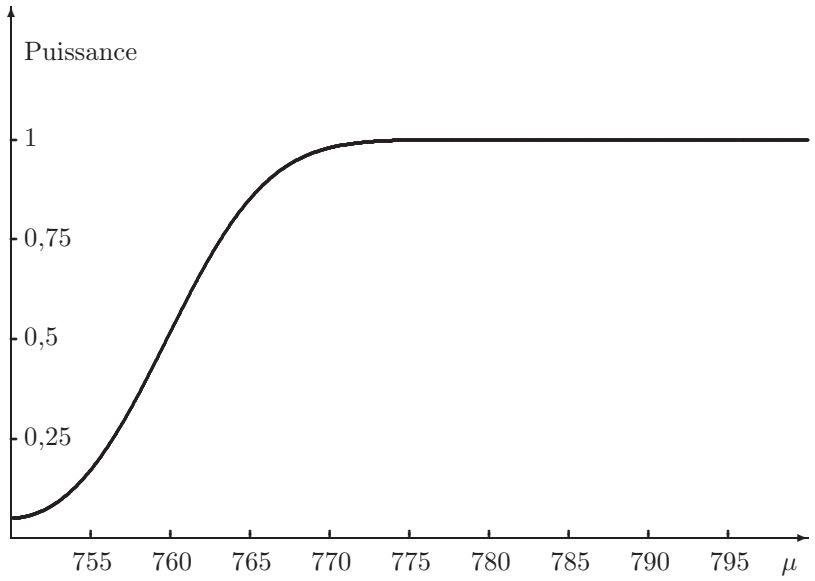

Fig. 5.2. *Puissance du test : la fonction puissance ici représente la probabilité que, pour la vraie valeur de μ se situant sur l'axe horizontal, le test n'accepte pas l'hypothèse $H_0 : \mu = 750$. Les données sont celles de l'Exemple* 5.1.

5.4 Tests statistiques

Nous avons vu que la loi normale a une importance centrale pour la mise en place d'un test d'hypothèse. D'autres lois sont aussi importantes que la loi normale. Nous verrons ici quelques autres tests et les lois de probabilité utilisées.

5.4.1 Le test du χ^2

Quand un résultat d'une simulation ou d'une expérience se présentant comme un ensemble d'observations classées sous différentes catégories doit être testé, notamment en le comparant avec un modèle théorique, le test du χ^2 est souvent le premier test qui est envisagé. C'est un test qui donne une réponse à la question de savoir si oui ou non un certain échantillon (simulé ou réel) est distribué selon une certaine loi. Quand la taille de l'échantillon est grande, souvent la réponse (en général en termes de valeur p) tend à être plus nette.

Considérons une variable aléatoire X dont les valeurs sont réparties en k classes. La variable X peut être une variable qualitative ou quantitative dont la distribution de probabilité est supposée connue.

Le test du χ^2 se fonde sur la quantité χ_c^2 calculée à partir de l'échantillon dont on dispose selon la formule

$$\chi_c^2 = \sum \frac{(\text{fréquence observée} - \text{fréquence théorique})^2}{\text{fréquence théorique}}.$$

Cette statistique est une mesure de la distance entre la distribution observée et la distribution théorique et permet d'évaluer la qualité de l'adéquation.

Considérons une expérience statistique dont le résultat de chaque essai peut être naturellement classifié dans une des k catégories possibles e_1, e_2, ..., e_k avec probabilités supposées être p_1, p_2, ..., p_k respectivement. On a $p_1 + p_2 + \cdots + p_k = 1$.

Le degré d'adéquation de la distribution théorique comparé à celui de la distribution observée se mesure par la quantité χ_c^2,

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i},$$

où x_i dénote la fréquence observée de la catégorie e_i, pour $i = 1, \ldots, k$, et np_i dénote la fréquence théorique ou attendue. Soit $p_i^{(\text{obs})} = x_i/n$.

Le test d'adéquation, au seuil de signification α, par rapport à la distribution de probabilité définissant les p_i teste les hypothèses :

$$\begin{aligned} H_0 &: \; p_i^{(\text{obs})} = p_i \\ H_1 &: \; \text{il existe } i \text{ tel que } p_i^{(\text{obs})} \neq p_i \end{aligned}$$

et rejette H_0 si $\chi_c^2 > \chi^2(k-1, 1-\alpha)$ ou accepte H_0 sinon. Puisque χ_c^2 est sensible à des faibles valeurs des fréquences attendues np_i, il est d'usage d'imposer, le cas échéant moyennant un regroupement de classes contiguës, que $np_i \geq 5$ pour tout $i = 1, \ldots, k$.

La quantité χ_c^2 suit une distribution χ^2 avec $k-1$ degrés de liberté. Donc ici $\chi^2(k-1, 1-\alpha)$ est la quantité critique que l'on peut lire sur une table du χ^2 (voir les Tables à la fin de cet ouvrage), et qui n'est rien d'autre que la fonction de répartition calculée en $1-\alpha$ de la loi du χ^2 à $k-1$ degrés de liberté introduite dans le Paragraphe 2.3.5.

Si la loi n'est pas complètement connue, mais est connue à 1, 2, ou m paramètres près, estimés à partir des observations, le nombre de degrés de liberté, et donc le premier paramètre à utiliser dans la distribution du χ^2, sera respectivement $k-2$, $k-3$, $k-m-1$.

Exemple 5.3 *La Table* 5.3 *montre le nombre de naissances par jour à Genève en* 1988 *durant* 52 *semaines complètes (*1^{er} *janvier – 29 décembre). Si les naissances étaient distribuées d'une façon strictement uniforme durant la semaine, on devrait s'attendre en moyenne que* $1/7$ *des naissances se produise les lundis,* $1/7$ *les mardis et ainsi de suite pour chaque jour de la semaine. Les valeurs*

observées montrent un certain écart par rapport à ce ratio. Le problème est de tester si les valeurs observées respectent la loi uniforme discrète. Si ce n'est pas le cas, la différence sera alors significative et l'on pourra dire que certains jours de la semaine sont plus favorables que d'autres sur le plan des naissances.

i	Jour de la semaine	Fréquences observées x_i	Fréquences théoriques np_i
1	lundi	598	$604{,}4 = 4\,231 \cdot 1/7$
2	mardi	636	$604{,}4 = 4\,231 \cdot 1/7$
3	mercredi	635	$604{,}4 = 4\,231 \cdot 1/7$
4	jeudi	662	$604{,}4 = 4\,231 \cdot 1/7$
5	vendredi	563	$604{,}4 = 4\,231 \cdot 1/7$
6	samedi	607	$604{,}4 = 4\,231 \cdot 1/7$
7	dimanche	530	$604{,}4 = 4\,231 \cdot 1/7$
	total	4 231	4 231

Tab. 5.3. *Nombre de naissances par jour de semaine. Source : Office cantonal de statistique, Genève.*

Le test du χ^2 donne la valeur suivante pour la mesure d'adéquation de la distribution théorique à la distribution observée :

$$\begin{aligned}
\chi_c^2 &= \sum_{k=1}^{7} \frac{(x_i - np_i)^2}{np_i} \\
&= \frac{(598 - 604{,}4)^2}{604{,}4} + \frac{(636 - 604{,}4)^2}{604{,}4} + \cdots + \frac{(530 - 604{,}4)^2}{604{,}4} \\
&= 20{,}76.
\end{aligned}$$

Ce calcul aurait pu être simplifié en notant qu'en général

$$\sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i} = \left(\sum_{i=1}^{k} \frac{x_i^2}{np_i} \right) - n$$

et dans le cas particulier où l'on a $k = 7$ et $p_i = 1/7$:

$$\begin{aligned}
\sum_{i=1}^{7} \frac{(x_i - np_i)^2}{np_i} &= \left(\sum_{i=1}^{7} \frac{x_i^2}{n \cdot \frac{1}{7}} \right) - n \\
&= 7 \cdot \frac{598^2 + 636^2 + \cdots + 530^2}{4\,231} - 4\,231 \\
&= 20{,}76.
\end{aligned}$$

La comparaison de la valeur calculée avec celle de la table du χ^2 correspondant au seuil de signification de 5 % et à $7-1=6$ degrés de liberté donne :

$$\chi^2_c = 20{,}76 \quad > \quad \chi^2_{(0{,}05,6)} = 12{,}60.$$

Cela indique que les fréquences observées sont significativement différentes de l'hypothèse nulle qui présume que les naissances journalières sont uniformes pour chaque jour de la semaine.

En effet, il semble que le nombre des naissances le dimanche est nettement plus bas que celui des autres jours de la semaine, en particulier les mardi, mercredi et jeudi.

Il est important de noter que le test du χ^2 est sensible au groupement des catégories de la variable à étudier. En se référant à l'exemple précédent, le groupement des journées de la semaine en *jours ouvrables* et *week-end* aboutirait à des résultats différents pour la valeur du test du χ^2, ($\chi^2 = 8{,}03$ avec $1 = 2-1$ degré de liberté) mais identiques quant à la conclusion. On trouve en effet une différence significative entre les naissances qui se produisent le week-end et celles des jours ouvrables. Mais d'une façon plus générale, pour quelques raisons ou par un pur hasard (dans ce dernier cas seulement avec probabilité α), il se pourrait que la conclusion du test du χ^2 après groupement soit contraire à celle fondée sur le test du χ^2 sans groupement.

Voici un autre exemple où la distribution théorique de la variable aléatoire X n'est pas une distribution uniforme.

Exemple 5.4 *Des clients arrivent à un guichet selon un modèle où la fréquence horaire des clients X est la suivante :*

$X = i$	0	1	2	3
$p_i = \Pr(X = i)$	1/8	1/2	1/4	1/8

Tab. 5.4. *Modèle de variable aléatoire qui décrit les arrivées à un guichet.*

Sur une semaine de test pendant laquelle le guichet est resté ouvert 48 *heures, les fréquences observées ont été les suivantes :*

7 fois	0 client
20 fois	1 client
12 fois	2 clients
9 fois	3 clients

Tab. 5.5. *Les fréquences observées dans la période de test sont comparées avec le modèle assumé (voir Tab.* 5.4.*)*

Peut-on dire que les clients sont arrivés selon le modèle supposé ?

Nous avons ici 4 classes. L'effectif théorique de chaque classe est $\phi_i = np_i$ où $n = 48$ est le nombre total de cas appartenant à chaque classe. Pour évaluer la qualité de l'adéquation de la distribution observée, on calcule la statistique $\chi_c^2 = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i}$ où $k = 4$ est le nombre de classes. On teste l'hypothèse nulle

$$H_0 : x_i = np_i$$

contre l'alternative

$$H_1 : \text{il existe } i \text{ tel que } x_i \neq np_i \ .$$

On a :

$$np_1 = 48 \cdot \tfrac{1}{8} = 6 \qquad np_3 = 48 \cdot \tfrac{1}{4} = \tfrac{1}{2}$$

$$np_2 = 48 \cdot \tfrac{1}{2} = 24 \qquad np_4 = 48 \cdot \tfrac{1}{8} = 6$$

d'où

$$\begin{aligned} \chi_c^2 &= \frac{(7-6)^2}{6} + \frac{(20-24)^2}{24} + \frac{(12-12)^2}{12} + \frac{(9-6)^2}{6} \\ &\simeq 2{,}333. \end{aligned}$$

On compare cette valeur à la valeur critique d'une loi du chi-carré à $4-1 = 3$ degrés de liberté, au seuil de signification $\alpha = 0{,}05$ ce qui donne $\chi_c^2 = 2{,}333 \leq 7{,}81$.

On accepte l'hypothèse H_0 et ainsi nous pouvons conclure que la distribution de la population dont est issu l'échantillon observé ne diffère pas de la distribution théorique modélisée.

Nous avons vu que la statistique

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i}$$

permet de mesurer une distance entre l'ensemble des fréquences observées x_i et l'ensemble des fréquences attendues np_i d'une variable aléatoire dont les valeurs sont réparties en k classes.

Une application importante du test du χ^2 concerne les générateurs de nombres pseudo-aléatoires. Les nombres générés peuvent facilement être assignés à des classes dont l'effectif théorique est en général facilement calculable.

Exemple 5.5 *Une variable aléatoire binomiale de paramètres* $(6; 0{,}23)$ *a été simulée par un ordinateur. Les résultats d'une simulation de* 250 *observations sont groupés dans le Tableau* 5.6 *:*

bin(6; 0,23)	0	1	2	3	4	5	6
x_i	40	90	77	30	10	3	0

Tab. 5.6. *Les fréquences observées d'un générateur de variable binomiale à tester.*

Par construction des données un choix naturel de la probabilité p est 0,23*, valeur qui ici n'est pas estimée à partir des observations.*

Ces données peuvent-elles être considérées comme engendrées par une loi binomiale de paramètres $n = 6$ *et* $p = 0{,}23$ *?*

Un test d'adéquation permet d'apporter une réponse à la question. Le test du χ^2, au seuil de signification $\alpha = 5$ %, donne une valeur espérée pour les classes 5 et 6 inférieure à 5. Donc il est judicieux de former une seule classe composée des valeurs $4, 5$ et 6, pour l'ensemble desquelles la valeur espérée est environ 7. Le nombre de degrés de liberté est 4, et le χ^2_c calculé est égal à 9,007 3, qui est inférieur à $\chi^2_{(0,05;4)} = 9{,}49$.

Donc l'hypothèse selon laquelle le générateur a correctement généré les nombres aléatoires permettant de simuler la loi bin(6; 0,23) n'est pas rejetée.

5.4.2 Le test de Student et le test de Fisher

Le test de Student est un test classique en statistique inférentielle, et il est utilisé pour tester si la moyenne d'un échantillon de taille modeste ($n \leq 30$) est égale à une valeur prédéfinie μ. Ce test est utilisé aussi pour tester l'égalité des moyennes de deux populations d'où deux échantillons de taille modeste sont tirés. La statistique du test est celle qui est définie au Paragraphe 2.3.6.

Exemple 5.6 *Considérons un problème de file d'attente. En supposant que le temps d'attente dans une file donnée est une variable aléatoire distribuée selon une loi normale, on aimerait déterminer si le temps moyen d'attente est égal à* 8,5 *minutes ou à plus de* 8,5 *minutes.*

On est donc en présence d'un test unilatéral pour les hypothèses

$$H_0 : \mu = 8{,}5$$
$$H_1 : \mu > 8{,}5 \ .$$

Afin de prendre une décision statistique en faveur d'une des deux hypothèses, on est amené à prendre un échantillon de taille n dans la population concernée. Différentes contraintes nous ont obligés à prendre un échantillon de taille $n = 16$, pour lequel nous avons observé une moyenne égale à 9,2 et un écart-type égal à 1,4. La valeur observée de la statistique de Student est ainsi égale à 2, valeur qui conduit au rejet de H_0 au niveau de signification $\alpha = 0{,}05$.

Un autre test très utilisé en statistique inférentielle est le test de Fisher, qui se fonde sur la statistique définie au Paragraphe 2.3.7. Il est utilisé par exemple en analyse de variance quand l'égalité des moyennes de plusieurs populations doit être testée à l'aide d'échantillons tirés de chaque population. Dans une telle situation la moyenne globale est comparée à la moyenne de chaque population, donnant lieu à une statistique appelée *moyenne des carrés entre les groupes* (MC_{ent}) ; les moyennes des différentes populations sont comparées aux observations des populations respectives, produisant une statistique appelée *moyenne des carrés à l'intérieur des groupes* (MC_{int}). Sous l'hypothèse nulle d'égalité des moyennes, ces deux statistiques, par construction, suivent des lois du χ^2. Le test sur l'égalité des moyennes se fonde sur la quantité $MC_{\text{ent}}/MC_{\text{int}}$ qui suit donc une loi de Fisher.

Une situation semblable se produit dans d'autres contextes, par exemple en régression quand on veut tester le degré de signification des coefficients d'une régression (Dodge, 1999).

5.4.3 Le test de Kolmogorov-Smirnov

Les tests d'adéquation permettent de déterminer si la distribution de probabilité de la population dont on connaît un échantillon est d'un certain type.

Le test de Kolmogorov-Smirnov est un test d'adéquation pour des variables aléatoires continues. On considère donc n variables aléatoires indépendantes identiquement distribuées dont la fonction de répartition F, inconnue, est continue. Notons F_n la fonction de répartition empirique définie par l'échantillon $(x_1, \ldots, x_n)$ où les x_i représentent des réalisations de la variable aléatoire en question. La fonction F_n est utilisée comme estimateur de F. Nous avons :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} h_i(x),$$

où

$$h_i(x) = \begin{cases} 1 & \text{si } x > x_i \\ \\ 0 & \text{sinon.} \end{cases}$$

On désire effectuer un test d'hypothèses concernant F. Les hypothèses à tester sont :

$$\begin{array}{lcl} H_0 & : & F = F_0, \text{ une fonction de répartition continue spécifiée} \\ H_1 & : & F \neq F_0 \end{array}$$

La statistique D_n de Kolmogorov-Smirnov est définie par

$$D_n = \sup_x |F_n(x) - F_0(x)| \ .$$

Sous H_0 pour $n \longrightarrow \infty$, $D_n \longrightarrow 0$. Ainsi on rejette H_0 si $D_n > c$ et l'on accepte H_0 sinon. La valeur de c est déterminée par l'équation $\Pr(D_n > c|H_0) = \alpha$ où α est le seuil de signification du test. Pour $n \leq 100$ et quelques valeurs de α les valeurs critiques c sont tabulées.

De manière équivalente on utilise la statistique de test D_n définie comme $D_n = \max(D_n^+, D_n^-)$, où

$$\begin{aligned} D_n^+ &= \sup_x (F_n(x) - F_0(x)) \\ D_n^- &= \sup_x (F_0(x) - F_n(x)) . \end{aligned}$$

Exemple 5.7 *Supposons que nous voulions tester la normalité d'une suite x_1, x_2, ..., x_n de nombres aléatoires. En ordonnant les nombres de la suite par ordre croissant et en les mettant sous forme standard, nous obtenons une suite ordonnée :*

$$z_{(i)} = \frac{x_{(i)} - \overline{x}}{s}, \qquad i = 1, \ldots, n,$$

où

$$s^2 = \frac{1}{n-1} \left(\sum_i (x_i - \overline{x})^2 \right).$$

Si l'hypothèse nulle est vraie, c'est-à-dire si les données sont normales, alors la fonction de répartition définie par

$$F_n(z) = \begin{cases} 0 & si -\infty < z < z_{(1)} \\ \dfrac{i}{n} & si\ z_{(i)} \leq z_{(i+1)}, i = 1, \ldots, n \\ 1 & si\ z_{(n)} \leq z < \infty, \end{cases}$$

ne devrait pas être significativement différente de la fonction de répartition Φ d'une loi normale centrée réduite. La statistique de test est donnée par $D_n = \max(D_n^+, D_n^-)$, où

$$\begin{aligned} D_n^+ &= \max_{1 \leq i \leq n} \left(\frac{i}{n} - \Phi(z_{(i)}) \right) \\ D_n^- &= \max_{1 \leq i \leq n} \left(\Phi(z_{(i)}) - \frac{i-1}{n} \right). \end{aligned}$$

À noter que ce test existe aussi sous une autre forme, permettant de tester si oui ou non deux échantillons proviennent de la même distribution.

5.4.4 Le test d'Anderson-Darling

Le test d'Anderson-Darling est un test d'adéquation entre la fonction de distribution théorique d'une variable aléatoire connue même à des paramètres

près, continue et la fonction de distribution empirique observée sur un échantillon issu de cette variable aléatoire. Par exemple, il permet de tester si la distribution empirique obtenue est conforme ou non à une loi normale.

Sous cet aspect, il a donc une finalité similaire à celle du test de Kolmogorov-Smirnov. La différence principale consiste dans le fait que la statistique du test de Kolmogorov-Smirnov est sensible davantage dans les valeurs autour de la médiane de la distribution, tandis que le test d'Anderson-Darling est uniforme sur toute la distribution.

Anderson et Darling ont utilisé la statistique connue aujourd'hui sous le nom de statistique d'Anderson-Darling, notée A^2, tout d'abord pour le test de conformité à une loi de distribution dont les paramètres sont parfaitement spécifiés (1952 et 1954). Ultérieurement, au cours des années 1960 et surtout 1970, Stephens (1974) essentiellement et quelques autres auteurs ont adapté le test à une gamme plus large de lois de distributions dont les paramètres sont ou ne sont pas connus.

Considérons la variable aléatoire X suivant une certaine loi ayant comme fonction de répartition $F_X(x;\theta)$ où θ est un paramètre (ou un ensemble de paramètres).

Une observation d'un échantillon de taille n issu de la variable X fournit une fonction de répartition empirique $F_n(x)$. La statistique A^2 d'Anderson-Darling est alors construite comme une somme pondérée des carrés des écarts $F_X(x;\theta) - F_n(x)$. Partant du fait que A^2 est une variable aléatoire positive qui suit une certaine loi de distribution sur l'intervalle $[0;+\infty[$, il est possible de tester, pour un seuil de signification fixé *a priori*, si $F_n(x)$ est ou non la réalisation de la variable aléatoire $F_X(X;\theta)$, c'est-à-dire si X est bien distribuée selon une loi de probabilité de fonction de répartition $F_X(x;\theta)$.

Nous organisons les observations $x_1, x_2, \ldots, x_n$ de l'échantillon issu de X de manière croissante, de sorte que $x_1 < x_2 < \cdots < x_n$. Notons alors $z_i = F_X(x_i;\theta)$, $(i = 1, 2, \ldots, n)$. Compte tenu des définitions précédentes, A^2 est définie formellement par :

$$A^2 = -\frac{1}{n}\left(\sum_{i=1}^{n}(2i-1)(\log(z_i) + \log(1 - z_{n+1-i}))\right) - n.$$

Des valeurs critiques de A^2 ont été tabulées et varient selon la loi de X. Ces valeurs sont normalement intégrées dans tous les logiciels pour le calcul statistique.

Nous traitons ici seulement le cas où X suit une loi normale d'expérance μ et variance σ^2. Ces paramètres peuvent être connus ou à estimer. Dans cette situation nous pouvons dénombrer quatre cas de figure, relatifs à la connaissance des paramètres μ et σ^2 (on dénote par Φ la fonction de répartition de la loi normale centrée réduite) :

(a) μ et σ^2 sont connues et $F_X(x;(\mu,\sigma^2))$ est alors parfaitement spécifiée. Nous avons alors naturellement $z_i = \Phi(w_i)$ où $w_i = (x_i - \mu)/\sigma$;

(b) σ^2 est connue et μ est inconnue et estimée par $\overline{x} = \frac{1}{n}\sum_i x_i$, moyenne de l'échantillon. Posons alors $z_i = \Phi(w_i)$, où $w_i = (x_i - \overline{x})/\sigma$;

(c) μ est connue mais σ^2 est inconnue et estimée par $s'^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2$. Dans ce cas, posons $z_i = \Phi(w_i)$, où $w_i = (x_i - \mu)/s'$;

(d) μ et σ^2 sont inconnues et estimées respectivement par $\overline{x}$ et

$$s^2 = \frac{1}{n-1}\left(\sum_i (x_i - \overline{x})^2\right).$$

Posons alors $z_i = \Phi(w_i)$, où $w_i = (x_i - \overline{x})/s$.

Les lois de distribution asymptotiques de la statistique A^2 ont été explicitées par Anderson et Darling pour le cas (a) et par Stephens pour les cas (b) et (c). Pour le cas (d), Stephens a déterminé la distribution asymptotique de la transformée $A^* = A^2(1,0 + \frac{0,75}{n} + \frac{2,25}{n^2})$ de A^2. Ainsi, nous disposons d'une table donnant, selon les cas (a) à (d), pour des seuils de signification de 10 %, 5 %, 2,5 %, et 1 %, les valeurs limites de A^2 (et A^* pour le cas (d)) au-delà desquelles l'hypothèse de normalité est rejetée (Tab. 5.7) :

Cas	Statistique	Seuil de signification			
		0,1	0,05	0,025	0,01
(a)	A^2	1,933	2,492	3,070	3,857
(b)	A^2	0,894	1,087	1,285	1,551
(c)	A^2	1,743	2,308	2,898	3,702
(d)	A^*	0,631	0,752	0,873	1,035

Tab. 5.7. *Valeurs critiques de la statistique d'Anderson-Darling dans le cas d'une loi normale à paramètres μ et σ^2 connus (a) ; μ inconnu et σ^2 connu (b) ; μ connu et σ^2 inconnu (c) ; paramètres inconnus (d).*

La loi de distribution de A^2 n'étant explicitée qu'asymptotiquement, le test requiert que la taille n de l'échantillon soit relativement importante. En toute rigueur, lorsque nous sommes confrontés aux cas 1 et 2, la distribution de A^2 n'est pas connue et il faudrait procéder à une transformation du type de celle de $A^2 \rightarrow A^*$, où A^* serait susceptible d'être déterminée. Toutefois, lorsque $n > 20$, nous pouvons éviter une telle transformation et alors les données de la table ci-dessus sont valides.

Nous avons insisté sur l'application de la statistique d'Anderson-Darling à un test de conformité à la loi normale. Cependant, le test d'Anderson-Darling présente l'avantage de s'appliquer avec une notable puissance sur une grande variété de lois de probabilité (non seulement loi normale, mais aussi lois exponentielle, logistique, gamma, etc.). Cela permet, lorsque, dans un premier temps, le test a rejeté l'hypothèse selon laquelle la variable aléatoire étudiée

suit une loi de distribution particulière, de le réitérer par rapport à une large gamme de distributions alternatives.

Exemple 5.8 *Les données suivantes permettent d'illustrer le test d'Anderson-Darling pour les hypothèses de normalité.*

Soit un échantillon relatif à la taille (en cm) de 25 *étudiants de sexe masculin : le tableau suivant récapitule les observations ordonnées ainsi que les* w_i *et* z_i*.*

Par ailleurs, calculons $\overline{x}$ *et* s *à partir de ces données :* $\overline{x} = 177{,}36$ *et* $s = 4{,}98$*.*

Soit Φ *la fonction de répartition normale centrée réduite. Un tableau pour la statistique d'Anderson-Darling est compilé ci-après (Tab.* 5.8*).*

Obs.	x_i	$w_i = \frac{x_i - \overline{x}}{s}$	$z_i = \Phi(w_i)$
1	169	−1,678	0,047
2	169	−1,678	0,047
3	170	−1,477	0,070
4	171	−1,277	0,100
5	173	−0,875	0,191
6	173	−0,875	0,191
7	174	−0,674	0,250
8	175	−0,474	0,318
9	175	−0,474	0,318
10	175	−0,474	0,318
11	176	−0,273	0,392
12	176	−0,273	0,392
13	176	−0,273	0,392
14	179	0,329	0,629
15	180	0,530	0,702
16	180	0,530	0,702
17	180	0,530	0,702
18	181	0,731	0,767
19	181	0,731	0,767
20	182	0,931	0,824
21	182	0,931	0,824
22	182	0,931	0,824
23	185	1,533	0,937
24	185	1,533	0,937
25	185	1,533	0,937

Tab. 5.8. *Un tableau pour le calcul de la statistique d'Anderson-Darling.*

Nous obtenons alors $A^2 \simeq 0{,}436$*, d'où* $A^* = A^2 \cdot (1{,}0 + \frac{0{,}75}{25} + \frac{0{,}25}{625}) = A^2 \cdot (1{,}033\ 6) = 0{,}451$.

Comme nous nous trouvons dans le cas (d)*, pour un seuil de signification fixé* a priori *à* 1 *%, la valeur calculée de* A^* *est très inférieure à la valeur tabulée* (1,035)*. Au seuil de* 1 *%, l'hypothèse de normalité ne peut donc pas être rejetée.*

5.4.5 Le test des permutations

Le test des permutations est un test d'adéquation qui est sensible à la structure d'ordre des suites d'éléments à tester. Ainsi appliqué à des suites de nombres réels censés provenir d'une variable aléatoire continue ou de la simuler, il peut tester si la suite apparaît suffisamment « dans le désordre » comme il se doit.

Soit $x_1, x_2, \ldots, x_n$ une suite de nombres pseudo-aléatoires censés simuler une distribution continue (pas forcément une distribution uniforme !), et soit r un entier. On considère les r-uples

$$\{x_{mr+i}, i = 1, \ldots, r\}, \qquad 0 \leq mr < n$$

où m est un entier. À chaque élément du r-uple on associe son rang. On obtient ainsi la permutation de ces r nombres qui consiste à les ranger dans l'ordre croissant. Il y a $r!$ permutations possibles, et la probabilité d'obtenir une de ces permutations est de $1/r!$. On compte le nombre de fois qu'apparaît chaque permutation possible. Le test des permutations est un test d'indépendance du χ^2 où le nombre de classes d'appartenance est $r!$, et la probabilité d'appartenir à une classe quelconque est $p_j = \frac{1}{r!}, j = 1, \ldots, r$.

Il est clair que la longueur de la suite doit être suffisamment grande. Puisqu'un test du χ^2 n'est fiable que si dans chaque classe le nombre espéré d'éléments est d'au moins 5, la longueur de la suite doit être d'au moins $5(r!)$.

5.5 Tests et qualité des générateurs

Dans le chapitre consacré aux nombres aléatoires et pseudo-aléatoires nous avons mis en évidence quelques faiblesses des générateurs. Que l'on utilise des tables de nombres aléatoires ou des suites de nombres pseudo-aléatoires obtenues par des méthodes numériques, et selon l'usage que l'on se propose de faire, il est nécessaire de déterminer si les propriétés d'indépendance et de distribution sont respectées. Cette dernière remarque s'applique aussi pour des suites de nombres pseudo-aléatoires non uniformes.

La meilleure manière de parer à l'éventualité qu'un générateur ait un défaut de quelconque nature est de le soumettre à une batterie de tests avant d'être utilisé.

Il existe une grande quantité de tests auxquels un générateur peut être soumis. Déjà au début des années 1980 Knuth en proposait une liste, en les divisant en tests empiriques, tests théoriques et tests spectraux. Marsaglia dans les années 1990 a ensuite repris le flambeau, en proposant une batterie de tests, nommée Diehard, qu'un générateur devrait passer, et en publiant un CDROM de nombres aléatoires, qui passe ces tests. Un handicap pour une telle liste de nombres aléatoires disponible sur un support numérique est que son utilisation entraîne une procédure de simulation lente et mobilise des grandes quantités de mémoire vive d'un ordinateur. Parmi les tests les plus courants dans la

littérature on retiendra encore le test du poker, le test de série sur les paires, sur les triplets, le test du collectionneur de coupons et le test du maximum du T.

Le nombre π se compose dans son développement décimal d'une suite de chiffres qui passe tout test statistique d'adéquation (voir Tab. 5.9).

Aujourd'hui, on dispose de plus de mille milliards de chiffres décimaux, bien plus que ceux qui entrent dans un CDROM. Cette suite garantit donc une qualité certainement non inférieure à celle proposée par une liste de provenance inconnue.

Test	χ^2_c	d	p
Test d'équidistribution	4,13	9	0,903
Test de série sur les paires	92,05	99	0,677
Test de série sur les triplets	1 029,43	999	0,245
Test du poker	10,54	6	0,104
Test du collectionneur de coupons	42,42	66	0,989
Test des permutations	6,69	5	0,245
Test du maximum du T	5,45	9	0,793

Tab. 5.9. *Résultats de* 7 *différents tests d'adéquation fondés sur des distributions de fréquences, appliqués sur les* 200 000 000 *premiers décimaux de* π*, d'après Murier et Rousson (1998). Une valeur du paramètre p supérieure à* 0,05 *garantit le passage du test correspondant.*

Toutefois d'autres générateurs, tout aussi robustes et fiables et bien plus rapides, comme le *Mersenne twister* (Matsumoto et Nishimura, 1998 ; Nishimura, 2000) sont devenus plus populaires. Une période très longue, une qualité en termes d'expertise avec toutes sortes de tests statistiques, et une rapidité dans l'exécution sont les qualités qui font aujourd'hui d'un générateur de nombres aléatoires un bon générateur.

Si le lecteur désire en savoir plus, Pierre L'Écuyer (2006) est l'auteur d'une discussion complète avec un état de l'art détaillé de toutes les techniques de génération de nombres aléatoires et une discussion de leurs qualités et de leurs défauts.

Exercices

5.1 On a demandé à un échantillon de 1 000 personnes de dire un nombre au hasard entre 1 et 100. La distribution des nombres ainsi obtenus peut se résumer dans le tableau suivant :

Intervalle	Quantité
(1-10)	130
(11-20)	100
(21-30)	107
(31-40)	98
(41-50)	102
(51-60)	105
(61-70)	90
(71-80)	85
(81-90)	90
(91-100)	93

(*a*) Peut-on dire que les nombres ainsi obtenus sont aléatoires ?

(*b*) Supposons qu'après avoir demandé à 10 000 personnes un nombre au hasard entre 1 et 100 les résultats soient distribués selon le tableau suivant :

Intervalle	Quantité
(1-10)	1 195
(11-20)	1 057
(21-30)	1 068
(31-40)	980
(41-50)	968
(51-60)	1 005
(61-70)	906
(71-80)	902
(81-90)	895
(91-100)	1 024

Peut-on encore dire que les nombres ainsi obtenus sont aléatoires ? Qu'en concluez-vous ?

5.2 Pour étudier les arrivées des clients à une station de service, nous procédons comme suit : 100 fois de suite, pendant un intervalle de temps de 20 minutes, nous comptons le nombre de clients qui prennent de l'essence. Le nombre d'arrivées par intervalle et les fréquences observées ont été reproduits dans le tableau ci-dessous :

No. d'arrivées	0	1	2	3	4	5	6	7	8	9	10	11
Fréq. observées	2	8	14	19	20	15	11	6	3	1	0	1

(*a*) Calculer la fréquence moyenne des arrivées par intervalle de 20 minutes ainsi que le nombre d'arrivées par minute (λ).

(*b*) Peut-on assimiler la distribution des fréquences observées à une loi de Poisson ? Utiliser le test du chi-carré.

5.3 Un générateur de nombres aléatoires $U(0,1)$ fournit 100 nombres selon le tableau suivant :

Classe	Quantité
0,000-0,099	17
0,100-0,199	19
0,200-0,299	15
0,300-0,399	18
0,400-0,499	23
0,500-0,599	22
0,600-0,699	24
0,700-0,799	21
0,800-0,899	19
0,900-0,999	22

Dites s'il est un bon générateur en effectuant un test opportun, en justifiant votre réponse.

5.4 Soit la suite de nombres :

$$4, 7, 9, 10, 1, 2, 4, 7, 3, 8, 4, 7, 8, 1, 1, 6, 5, 3, 7, 7.$$

(a) Tester si ces 20 nombres sont distribués uniformément sur l'intervalle $[0, 10]$.

(b) Déterminer la valeur minimale de α pour laquelle on peut rejeter l'hypothèse nulle que l'échantillon provient d'une loi uniforme $[0, 10]$. Discuter.

5.5 Huit générateurs de nombres aléatoires ont été utilisés pour générer 1 000 nombres entiers uniformément répartis entre 0 et 9. Les fréquences obtenues sont données dans le tableau suivant :

Nombre	0	1	2	3	4	5	6	7	8	9	Total
Gén. 1	99	101	99	101	99	101	99	101	99	101	1 000
Gén. 2	89	110	105	88	104	96	109	101	93	105	1 000
Gén. 3	100	100	100	100	100	100	100	100	100	100	1 000
Gén. 4	84	88	92	96	98	102	104	108	112	116	1 000
Gén. 5	100	100	101	100	100	100	99	100	100	100	1 000
Gén. 6	76	120	98	65	135	88	114	68	98	138	1 000
Gén. 7	89	104	91	121	115	76	93	104	109	98	1 000
Gén. 8	100	108	92	100	108	92	100	108	92	100	1 000

(a) Commenter les résultats à l'aide d'un test du chi-carré.

(b) Quel(s) générateur(s) de nombres aléatoires est (sont) le(s) plus efficace(s) ?

(c) À l'aide de la méthode de congruence, générer 10 nombres aléatoires avec les paramètres $m = 100$, $a = 31$ et $x_0 = 17$, puis commenter.

5.6 Un statisticien veut tester la bonté du générateur de nombres aléatoires d'un logiciel statistique. Il génère 1 000 nombres censés être aléatoires uniformes dans l'intervalle $[0, 1]$. Les nombres sont ainsi répartis :

Classe	Fréquence	Classe	Fréquence
0-0,1	106	0,5-0,6	102
0,1-0,2	80	0,6-0,7	103
0,2-0,3	94	0,7-0,8	109
0,3-0,4	100	0,8-0,9	100
0,4-0,5	93	0,9-1	113

(a) Dire si le générateur passe le test d'équidistribution du χ^2.

(b) Reproduire l'expérience à l'aide d'un logiciel de statistique.

5.7 Un statisticien veut tester la qualité d'un générateur de nombres aléatoires. Il génère 1 000 nombres dans l'intervalle $[0, 1]$. Les nombres sont ainsi repartis :

Classe	Fréquence	Classe	Fréquence
0-0,1	107	0,5-0,6	103
0,1-0,2	81	0,6-0,7	104
0,2-0,3	95	0,7-0,8	108
0,3-0,4	105	0,8-0,9	93
0,4-0,5	92	0,9-1	112

(a) Dire si le générateur passe le test d'équidistribution pour des nombres uniformément distribués sur l'intervalle $[0, 1]$.

(b) Fournir une méthode pour générer des nombres pseudo-aléatoires et déterminer des paramètres suffisants pour que l'on produise un cycle de longueur 100 000.

Chapitre 6

La méthode de Monte Carlo et ses applications

6.1 Introduction

La méthode de Monte Carlo peut être définie comme toute technique numérique de résolution de problèmes au moyen d'un modèle stochastique dans lequel on utilise des nombres aléatoires. On attribue la méthode de Monte Carlo, développée vers 1949, aux mathématiciens américains John von Neumann et Stanislav Ulam. Ce n'est toutefois qu'avec l'avènement des ordinateurs que l'on a pu réellement utiliser cette méthode. Quant au nom de Monte Carlo, on le doit bien sûr à la capitale de la principauté de Monaco, célèbre pour son casino. En effet, la roulette est l'un des mécanismes les plus simples pour générer des nombres aléatoires.

Dans ce chapitre, nous présentons quelques applications en simulation. L'estimation d'une surface, le problème des files d'attente, l'ajustement de l'offre d'un bien en fonction des conditions climatiques, l'estimation d'une intégrale, la gestion de stocks et le rendement d'un investissement sont des problèmes qui peuvent être résolus par ces méthodes.

6.2 Estimation d'une surface

Nous avons vu au Chapitre 1 comment on peut estimer une surface S à l'aide de points aléatoires disposés autour de S selon une loi uniforme. Nous allons revenir sur l'exemple de la région S du Chapitre 1 pour mieux saisir cette technique. Combien de points sont nécessaires pour une estimation ? Avec quelle marge d'erreur ? Nous allons répondre à la question et donner aussi une estimation avec un intervalle de confiance, dont la largeur dépend du nombre de points choisis.

Nous devons calculer la surface de la région S du Chapitre 1 à l'aide de points aléatoires placés à l'intérieur du carré de côté égal à 1. Nous allons voir maintenant comment trouver et disposer ces différents points.

Nous reportons le carré sur un système d'axes perpendiculaires dont l'origine est l'angle inférieur gauche du carré.

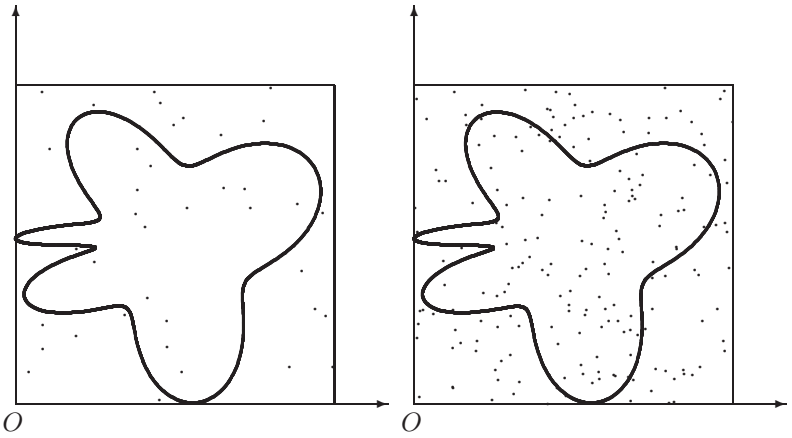

Fig. 6.1. *Estimation d'une surface au moyen de points aléatoires.*

Tous les N points que l'on va disposer à l'intérieur du carré auront donc des coordonnées comprises entre 0 et 1. Il suffit par conséquent de prendre deux variables aléatoires uniformes pour obtenir un point. En effet, le premier nombre aléatoire sera l'abscisse et le second l'ordonnée. Pour simuler, comme il a été fait au Chapitre 1, 40 points aléatoires sur le carré, nous avons besoin de deux échantillons de 40 nombres aléatoires. Il ne nous reste plus qu'à compter le nombre N' de points qui se trouvent dans S et à calculer le quotient N'/N.

Dans l'exemple discuté au Chapitre 1, résumé ici dans la Figure 6.1 (a), $N' = 16$, et $N = 40$, ce qui donne une estimation de la surface de S égale à $16/40 = 0{,}4$.

Il est clair que si le nombre de points est plus élevé, l'estimation sera plus fiable et plus précise. Dans la Figure 6.1 (b), 200 points sont simulés. Parmi ceux-ci, 95 se trouvent à l'intérieur, ce qui donne une estimation pour la surface de S de 0,475.

Ces deux estimations peuvent paraître contradictoires. Il faut toutefois considérer chaque point aléatoire comme une variable de Bernoulli X_i. Le point i sera à l'intérieur $(X_i = 1)$ avec probabilité $p =$ Aire (S) et à l'extérieur $(X_i = 0)$ avec probabilité $1 - p$. Si l'on considère la variable $Y = \frac{1}{N}\sum_{i=1}^{N} X_i$, une réalisation de Y est le nombre total de points à l'intérieur de S divisé par le nombre total de points. Il va de soi que $E(Y) = p =$ Aire (S). Mais on peut dire plus : en effet

$$\text{var } (Y) = \frac{1}{N^2} Np(1-p) = \frac{1}{N} p(1-p).$$

Il est vrai que la valeur de p n'est pas connue, mais cela nous dit comment la variance diminue avec l'augmentation du nombre N de points choisis.

Par la même occasion on peut aussi construire des intervalles de confiance. Pour les deux situations ($N = 40$ et $N = 200$), on a ainsi

$$p_1 = 0{,}4, \qquad p_2 = 0{,}475.$$

À défaut de connaître la variance, celle-ci est estimée dans les deux situations par S_1^2 et S_2^2 comme suit :

$$S_1^2 = \frac{0{,}4 \cdot 0{,}6}{40} = 0{,}006, \qquad S_2^2 = \frac{0{,}475 \cdot 0{,}525}{200} \simeq 0{,}001\ 25.$$

Les deux intervalles de confiance à 95 % sont (voir aussi le Paragraphe 4.7) :

$$0{,}4 - 1{,}96\sqrt{0{,}006} \simeq 0{,}248 \leq p \leq 0{,}552 \simeq 0{,}4 + 1{,}96\sqrt{0{,}006}$$

$$0{,}475 - 1{,}96\sqrt{0{,}001\ 2} \simeq 0{,}406 \leq p \leq 0{,}544 \simeq 0{,}475 + 1{,}96\sqrt{0{,}001\ 2}.$$

Les intervalles de confiance montrent que les deux résultats sont tout à fait compatibles. À noter que le deuxième intervalle, fondé sur un nombre de points plus que 4 fois supérieur, est large environ 0,138, soit moins de la moitié que le premier. Cela donne une relation entre le nombre de points N et la largeur de l'intervalle. Ainsi pour une largeur de 0,046 il faudra environ 1 800 points. Pour une largeur de 0,002, il faudra environ $1\ 800 \times 23^2 = 952\ 200$ points, ce qui donne une estimation du type $p = p_0 \pm 0{,}001$.

En augmentant la taille de l'échantillon de points choisis, on peut améliorer cette estimation. En effet, plus N est grand, plus le quotient sera proche de la surface réelle, l'idéal étant de considérer l'ensemble de la population de points et de compter la répartition asymptotique des points. Mais ce n'est justement pas le but de la simulation qui est de simplifier les calculs et de gagner du temps.

L'intérêt de cette méthode réside dans le fait qu'elle peut être appliquée au calcul de surfaces pour lesquelles une formule mathématique n'existe pas, comme c'est le cas pour une surface non régulière en général.

6.3 Problèmes de files d'attente

Un phénomène de file d'attente se caractérise par le fait que plusieurs clients en même temps demandent à être servis. Les clients arrivent de manière aléatoire et nous supposons que les arrivées sont indépendantes les unes des autres. Le temps qui sépare deux arrivées consécutives sera appelé *interarrivée*. Il se peut que certaines interarrivées soient corrélées entre elles (par exemple, de petits groupes de clients arrivent en même temps, selon les horaires des trains). Dans l'exemple qui va suivre, nous supposerons cependant que les interarrivées sont indépendantes. Nous savons qu'il existe des heures ou des périodes de

pointe. Tout commerçant sait qu'il y a davantage de clients juste avant Noël par exemple. Pour étudier un phénomène d'attente, on peut donc décomposer la période concernée en plusieurs périodes élémentaires, telles que dans chacune d'entre elles la loi des interarrivées soit considérée comme constante.

Le temps nécessaire pour fournir au client le service qu'il demande, la *durée* ou *temps de service*, doit être pris en compte dans une simulation. D'autres éléments essentiels dans un problème de file d'attente est le nombre de guichets et la distribution des temps de service pour chaque opérateur de guichet.

Dans la plupart des cas, on peut considérer que la loi des interarrivées est une loi exponentielle. C'est en effet la loi qui correspond à une fréquence uniforme, et à des arrivées non corrélées entre elles. La densité de probabilité qu'un temps t sépare l'arrivée de deux clients est alors égale à :

$$a(t) = \lambda e^{-\lambda t}$$

où λ représente le nombre moyen d'arrivées de clients par unité de temps. Cela revient à dire que la probabilité p_n d'observer n arrivées pendant l'unité de temps suit la loi de Poisson :

$$p_n = \frac{e^{-\lambda}\lambda^n}{n!}.$$

Des hypothèses peuvent être faites pour modéliser le phénomène en question. Nous énonçons les hypothèses adoptées par la suite à propos de la durée de service :

(a) la durée de service suit la même loi de probabilité pour tous les clients, c'est-à-dire qu'il n'existe pas plusieurs catégories de clients ;

(b) la durée de service suit la même loi à chaque guichet, c'est-à-dire qu'il n'y a pas de serveur plus rapide que d'autres ;

(c) les durées de service sont indépendantes les unes des autres.

Il est clair que la loi des arrivées des clients et la loi des durées de service étant données, plus le nombre de guichets sera élevé, plus le temps d'attente sera court. Cependant, il faudra contrebalancer le coût de guichets supplémentaires avec le coût de l'attente des clients. Une attente plus grande entraîne une attractivité amoindrie qui se traduit par un manque à gagner. Par conséquent, le problème est de trouver le nombre optimal de guichets à ouvrir de sorte que les clients attendent le moins possible.

Nous allons maintenant étudier un exemple complet et voir comment appliquer la théorie que nous venons de présenter dans ses grandes lignes.

Exemple 6.1 *Le gérant d'une agence de voyages a observé que le temps de service des employés qui reçoivent les clients est distribué selon la fonction de densité*

$$f(t) = \begin{cases} \dfrac{1}{10} e^{-\frac{1}{10}(t-2)} & \textit{si } t \geq 2 \\ \\ 0 & \textit{sinon.} \end{cases}$$

Le nombre moyen de clients est de 15. *Trois guichets sont ouverts et organisés pour que les clients (qui font une queue unique) soient servis au premier guichet qui se libère. Le gérant désirerait savoir si l'attente des clients au guichet est acceptable ou non (le critère étant le suivant : l'attente des clients est acceptable si la somme des attentes de tous les clients ne dépasse pas le temps durant lequel ces mêmes clients ont été servis).*

Si les temps d'interarrivée sont régis par une loi exponentielle, aider le gérant à prendre sa décision en simulant le phénomène depuis l'heure d'ouverture à l'heure de fermeture, une heure après.

Pour simuler l'arrivée (ou plutôt l'interarrivée) des clients, en sachant que les clients arrivent avec une moyenne de 15 clients par heure (1 client toutes les 4 minutes), nous utiliserons des nombres aléatoires u_i distribués uniformément sur $[0,1]$, et la transformation $t_i = -4\log(u_i)$ (voir le Paragraphe 4.1.2). Ensuite, il faut calculer la fonction de répartition des temps de service :

$$F(t) = \int_2^t \frac{1}{10} e^{-\frac{1}{10}(x-2)} dx = 1 - e^{-\frac{1}{10}(t-2)}, \text{ pour } t \geq 2.$$

À l'aide d'un autre ensemble, indépendant du premier, de nombres aléatoires distribués uniformément sur $[0,1]$, v_i, cette fonction permet de simuler les temps de service $t_{s,i}$. La formule

$$1 - e^{-\frac{1}{10}(t_{s,i}-2)} = 1 - v_i$$

donne

$$t_{s,i} = 2 - 10\log v_i.$$

Voici le résultat d'une simulation :

u_i	v_i	t_i	t_i cumulé	$t_{s,i}$	gui.	fin 1	fin 2	fin 3	attente
0,141	0,549	7,819	7,819	7,991	1	15,810	0,000	0,000	0,000
0,897	0,442	0,430	8,249	10,144	2	15,810	18,394	0,000	0,000
0,264	0,665	5,322	13,571	6,065	3	15,810	18,394	19,637	0,000
0,502	0,284	2,749	16,321	14,561	1	30,882	18,394	19,637	0,000
0,693	0,378	1,461	17,782	11,710	2	30,882	30,105	19,637	0,611
0,582	0,271	2,164	19,947	15,048	3	30,882	30,105	34,996	0,000
0,592	0,456	2,094	22,042	9,842	2	30,882	39,947	34,996	8,062
0,062	0,346	11,067	33,109	12,612	1	45,721	39,947	34,996	0,000
0,862	0,454	0,590	33,699	9,889	3	45,721	39,947	44,885	1,296
0,342	0,133	4,290	37,990	22,103	2	45,721	62,050	44,885	1,957
0,821	0,024	0,786	38,776	38,923	3	45,721	62,050	83,808	6,108
0,328	0,724	4,456	43,232	5,221	1	50,943	62,050	83,808	2,488
0,093	0,063	9,464	52,697	29,621	1	82,318	62,050	83,808	0,000
0,505	0,488	2,726	55,423	9,171	2	82,318	71,221	83,808	6,627
0,535	0,962	2,494	57,918	2,378	2	82,318	73,599	83,808	13,303
0,481	0,917	2,926		2,864					

Tab. 6.1. *Résultat de la simulation : dans cette simulation* 15 *clients arrivent pendant l'heure d'ouverture.*

Dans le tableau de simulation ci-dessus, les deux premières colonnes représentent les nombres aléatoires utilisés pour simuler respectivement les temps des

interarrivées des clients figurant sur la troisième colonne, et les temps de service figurant sur la cinquième colonne. La quatrième colonne représente le temps écoulé depuis l'ouverture du guichet ; la sixième colonne représente le guichet qui sert le client en question ; les colonnes 7 à 9 représentent le temps écoulé de l'ouverture des guichets et après avoir servi le client respectif, pour les guichets 1, 2 et 3. La dernière colonne représente l'attente des clients.

Dans cette simulation, 15 clients arrivent avant l'heure. Seulement 8 clients attendent quelques minutes, donc l'attente des clients est acceptable. On pourrait suggérer de faire 1 000 simulations et de calculer le pourcentage de cas où l'attente des clients n'est pas acceptable (voir aussi le Paragraphe 7.2).

Le modèle M/M/1 est une simplification de la situation précédente : il suppose des temps d'interarrivée et des temps de service exponentiels et un guichet unique. Si les interarrivées sont gérées par une loi exponentielle de paramètre λ et les temps de service par une loi exponentielle de paramètre μ avec $\lambda < \mu$, le système est stable sur le long terme et il existe une solution analytique pour le nombre moyen de clients dans la queue, $N_{\text{moy}} = \lambda/(\mu - \lambda)$, et le temps moyen d'attente (incluant le service) des clients $T_{\text{moy}} = 1/(\mu - \lambda)$.

Dans l'Exemple 6.1 l'espérance de la variable temps d'interarrivée est de 4 minutes et celle du temps de service de chaque guichet est de 12 minutes. Puisqu'il y a trois guichets ouverts, un client toutes les 4 minutes peut être servi. Il y a un équilibre instable entre clients qui arrivent et clients servis. L'attente moyenne dans ces cas dépend de la durée d'ouverture des guichets : l'attente moyenne est différente selon que les guichets sont ouverts pendant une heure ou pendant deux heures. L'approche par simulation montre ici toute son utilité.

L'exemple illustre aussi la marche à suivre dans le cas général. Si les clients n'arrivent pas de manière aléatoire, mais en suivant une distribution de leurs arrivées plus complexe, ou encore si les temps de service de chaque guichet suivent une loi de distribution propre à l'opérateur qui en est chargé, le tableau de simulation contiendra des colonnes supplémentaires de nombres aléatoires. D'autres colonnes supplémentaires représenteront les échantillons fictifs des nouvelles variables considérées.

6.4 Ajustement de l'offre d'un bien en fonction des conditions climatiques

Dans ce paragraphe, nous allons étudier l'évolution de l'offre d'un bien en fonction des conditions climatiques comme l'ensoleillement, la quantité de pluie, les jours de canicule ou de froid glacial, comme pourrait l'être l'offre de blé, de vin ou d'huile d'olive.

Nous prenons l'année comme base temporelle, et dans le modèle que nous adoptons il existe trois cas possibles :

(a) l'année est mauvaise ;

(b) l'année est moyenne ;

(c) l'année est bonne.

Ces trois possibilités vont donc influer sur l'offre du bien en question et ainsi sur son prix. L'offre va également dépendre du prix de l'exercice précédent, car, en l'absence d'ententes cartellaires, plus le prix de l'exercice précédent est élevé, plus les producteurs tendront à offrir le bien en question l'année suivante. La tendance de l'offre en fonction du prix de l'année précédente est donc une fonction croissante. Il y aura par conséquent deux offres différentes : la tendance de l'offre qui sera fonction du prix de l'exercice précédent et l'offre effective qui sera déterminée par le prix de l'exercice précédent ainsi que par le climat.

La demande dépend du prix courant selon une fonction décroissante : plus le prix sera élevé, plus faible sera la demande.

Le but de cet exemple est de simuler les conditions climatiques sur 20 ans et de comparer la tendance de l'offre avec l'offre effective.

6.4.1 Le modèle

Nous supposons que l'échange se fait à un prix tel que la demande est égale à l'offre. La demande dépend du prix de l'année en cours. Le prix de l'année 0 est donné. Nous avons donc le modèle suivant :

(a) le prix de départ p_0 ;

(b) la tendance de l'offre au temps t : $T_t = f(p_{t-1})$;

(c) l'offre effective au temps t : $O_t = C_t \cdot f(p_{t-1})$;

(d) la demande au temps t : $D_t = g(p_t)$.

Nous définissons les différentes fonctions d'offre (f) et de demande (g) comme suit :

$$\begin{aligned} f(p) &= e^p - 1 \\ g(p) &= \frac{1}{p^2} \, . \end{aligned}$$

Ce qui donne finalement :

$$\begin{aligned} T_t &= e^{p_{t-1}} - 1 \\ O_t &= C_t(e^{p_{t-1}} - 1) \\ D_t &= \frac{1}{p_t^2} \, . \end{aligned}$$

Le prix au temps t est tel que la demande soit égale à l'offre, et donc tel que $D_t = O_t$, ce qui donne

$$p_t = \frac{1}{\sqrt{C_t(e^{p_{t-1}} - 1)}}.$$

Il nous reste à définir le facteur climatique C_t. Remarquons que C_t est un facteur multiplicatif de la tendance de l'offre. Nous choisissons donc les valeurs suivantes pour C_t, selon que l'année est bonne, moyenne ou mauvaise :

(a) année bonne : $C_t = 1{,}2$;

(b) année moyenne : $C_t = 1{,}0$;

(c) année mauvaise : $C_t = 0{,}9$.

Nous associons à ces différentes valeurs de C_t les probabilités de 0,2 en cas d'année bonne, de 0,4 en cas d'année moyenne et de 0,4 en cas d'année mauvaise. Nous considérons donc que C_t est une variable aléatoire d'espérance égale à 1.

6.4.2 La simulation

Nous allons maintenant simuler les conditions climatiques sur 20 ans. Pour cela, il est possible de procéder de multiples manières.

On place dans une urne 10 billets numérotés de 0 à 9. On tire un billet, on note le chiffre qui est inscrit dessus et on le remet dans l'urne. On répète le tirage 20 fois. On peut aussi prendre les 20 premiers chiffres de la table de nombres aléatoires de l'Appendice.

En tout cas, la probabilité d'apparition d'un numéro est égale à :

$$p = \frac{1}{10} = 0{,}1.$$

Pour construire la valeur de C_t, nous nous référons aux probabilités théoriques que nous avons associées aux différentes valeurs de C_t, selon la méthode de comparaison illustrée au Paragraphe 4.4.

Nous définissons que C_t prend la valeur 1,2 quand le chiffre inscrit sur les billets est 0 ou 1. En effet, la probabilité d'avoir 0 ou 1 est égale à $0{,}1+0{,}1 = 0{,}2$, ce qui correspond bien à notre probabilité théorique d'avoir une année bonne.

De manière similaire C_t prend la valeur de 1,0 lorsque les billets portent les chiffres 2, 3, 4 ou 5. En effet, la probabilité de tirer l'un de ces numéros est égale à $0{,}1 + 0{,}1 + 0{,}1 + 0{,}1 = 0{,}4$. Cette probabilité est la même que celle associée à une année moyenne.

Finalement, C_t prend la valeur de 0,9 lorsque les chiffres tirés sont 6, 7, 8 ou 9. Là aussi, la probabilité de tirer l'un de ces numéros est égale à la probabilité d'avoir une année mauvaise, c'est-à-dire 0,4.

Nous pouvons donc construire un tableau de simulation (Tab. 6.2) avec, dans les différentes colonnes, l'année, le chiffre inscrit sur le billet, la valeur de C_t, le prix, la tendance de l'offre, et l'offre effective. Supposons que $p_0 = 0{,}86$ est donné.

Examinons par exemple les calculs pour l'année 2 : le chiffre tiré est le 5, par conséquent la valeur associée de C_t est 1,0.

Année t	Nombre	Valeur C_t	Prix p_t	Tendance de l'offre	Offre effective
0	-	-	0,86	-	-
1	1	1,2	0,78	1,36	1,64
2	5	1,0	0,92	1,18	1,18
3	0	1,2	0,74	1,51	1,81
4	1	1,2	0,87	1,10	1,32
5	1	1,2	0,77	1,39	1,66
6	8	0,9	0,97	1,17	1,05
7	7	0,9	0,82	1,65	1,48
8	4	1,0	0,89	1,27	1,27
9	9	0,9	0,88	1,43	1,28
10	3	1,0	0,84	1,42	1,42
11	5	1,0	0,87	1,32	1,32
12	9	0,9	0,89	1,39	1,25
13	8	0,9	0,88	1,44	1,30
14	4	1,0	0,84	1,40	1,40
15	2	1,0	0,87	1,33	1,33
16	9	0,9	0,90	1,38	1,24
17	7	0,9	0,87	1,45	1,30
18	4	1,0	0,84	1,40	1,40
19	2	1,0	0,87	1,33	1,33
20	5	1,0	0,85	1,38	1,38

Tab. 6.2. *Résultats de la simulation de l'offre d'un bien en fonction de la demande.*

La tendance de l'offre est $T_2 = e^{p_1} - 1 = e^{0,78} - 1 = 1,18$.
L'offre effective est $O_2 = C_t(e^{p_1 - 1}) = 1,0 \cdot (e^{0,78} - 1) = 1,18$.
Le prix est $p_2 = \frac{1}{\sqrt{D_2}} = \frac{1}{\sqrt{O_2}} = \frac{1}{\sqrt{1,18}} = 0,92$.

Calculons maintenant les fréquences absolues et les probabilités observées des conditions climatiques et comparons-les aux probabilités théoriques :

	Fréquence absolue	Probabilité observée	Probabilité théorique
bonne année	4	$4/20 = 0,2$	0,2
année moyenne	9	$9/20 = 0,45$	0,4
mauvaise année	7	$7/20 = 0,35$	0,4

Tab. 6.3. *Comparaison de la simulation obtenue avec les fréquences théoriques.*

Le résultat qui nous intéresse est la différence entre la tendance de l'offre et l'offre effective. Nous pouvons calculer pour chaque année la différence entre les deux, mais pour avoir une vision d'ensemble, nous allons reporter les deux fonctions sur un seul graphe. Nous plaçons le temps sur l'axe horizontal, T_t et O_t sur l'axe vertical.

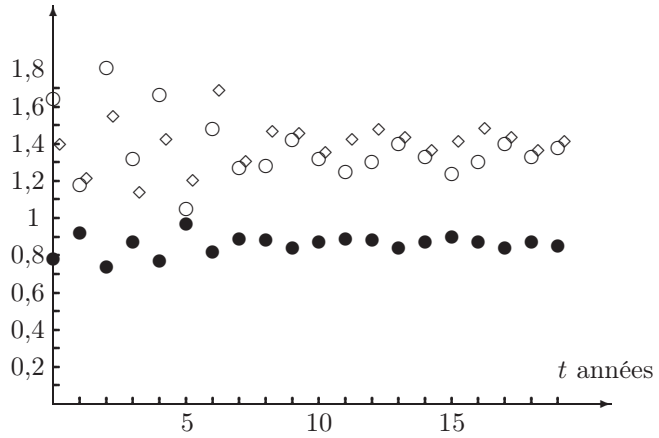

Fig. 6.2. *Évolution temporelle de la tendance de l'offre (◇) et de l'offre effective (○). Sur le même graphe apparaît aussi le prix (●).*

Le prix est une autre variable dont la simulation donne des informations importantes. Son évolution, et en particulier l' écart-type constaté dans le tableau de simulation, donne une mesure du risque que les investisseurs (opérateurs de marché, acheteurs, et vendeurs) prennent. Dans un modèle plus complexe, les jours de soleil, de pluie et autres facteurs météorologiques peuvent être simulés séparément avec des différents jeux de nombres aléatoires. Cela donne un tableau de simulation plus complexe, tout en gardant le même type de structure. Un exemple comme celui que l'on vient d'illustrer montre toute l'importance de la simulation dans le domaine économique.

6.5 Estimation d'une valeur d'intégrale

Nous présentons dans ce paragraphe une application de la méthode de Monte Carlo pour résoudre un problème déterministe comme celui de l'estimation de la valeur d'une intégrale. Nous commençons par le cas d'une intégrale unidimensionnelle, pour généraliser ensuite le résultat au cas d'une intégrale multiple, plus courant dans la pratique de la simulation.

De manière générale, le procédé de Monte Carlo peut être utilisé pour l'estimation d'une intégrale de la forme

$$I = \int_a^b h(x)dx,$$

où $h(x) = h_1(x)h_2(x)$ avec h_2 une fonction de probabilité.

Si $x_i \sim X$ est un échantillon distribué selon la densité $h_2(x)$, en définissant

$$g(x) = \begin{cases} h_1(x) & \text{si } x \in (a,b) \\ \\ 0 & \text{autrement} \end{cases}$$

nous avons que

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)h_2(x)dx = \int_a^b h_1(x)h_2(x)dx = \int_a^b h(x)dx.$$

Un estimateur non biaisé de I est donc :

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} g(x_i).$$

Nous allons maintenant illustrer cette procédure à l'aide d'un exemple simple.

Exemple 6.2 *Estimons l'intégrale suivante par la méthode de Monte Carlo :*

$$\eta = \int_0^{\infty} xe^{-x}dx.$$

Nous savons que la valeur exacte de cette intégrale est égale à 1. En effet, nous pouvons la résoudre par parties :

$$\begin{aligned} \eta &= \int_0^{\infty} xe^{-x}dx = \left[-xe^{-x}\right]_0^{\infty} + \int_0^{\infty} e^{-x}dx \\ &= \left[-xe^{-x} - e^{-x}\right]_0^{\infty} \\ &= \left[-e^{-x}(x+1)\right]_0^{\infty} \\ &= 1 \end{aligned}$$

$$\begin{aligned} \text{avec } u = x, &\quad dv = e^{-x}, \\ du = 1, &\quad v = -e^{-x}. \end{aligned}$$

Nous allons maintenant l'estimer par la méthode de Monte Carlo. Prenons un échantillon de nombres aléatoires u_1, u_2, ..., u_{10}. Nous calculons les x_i correspondant aux u_i par la formule $x_i = -\log u_i$. Cela donne :

i	u_i	$x_i = -\log u_i$
1	0,23	1,47
2	0,13	2,04
3	0,63	0,46
4	0,88	0,13
5	0,28	1,27
6	0,12	2,12
7	0,68	0,38
8	0,92	0,08
9	0,70	0,36
10	0,30	1,20

Tab. 6.4. *Simulation pour une distribution exponentielle de paramètre* 1.

Nous estimons alors η par $\hat{\eta}$ de la façon suivante (ici $g(x_i) = x_i$) :

$$\hat{\eta} = \frac{1}{10} \sum_{i=1}^{10} g(x_i) = 0,951.$$

Ce résultat est assez proche de 1. Si nous augmentons la quantité de nombres aléatoires, nous pouvons accroître la précision.

Exemple 6.3 *Supposons maintenant que nous voulions déterminer la valeur de l'intégrale ci-dessous dénotée par $\xi(\lambda, v)$ ou plus brièvement par ξ.*

$$\xi(\lambda, v) = \int_v^{\infty} \frac{1}{x} \lambda e^{-\lambda x} dx \qquad (\lambda, v > 0).$$

Cette intégrale semble très simple, mais ne peut néanmoins pas être résolue par l'intégration par parties.

On remarque que pour $v = 0$ l'intégrale est divergente. Évaluons cette intégrale pour $v = 1$ et $\lambda = 1$.

Prenons le même échantillon de nombres aléatoires et définissons une autre fonction $g(x)$:

$$g(x) = \begin{cases} 0 & \text{si } x < v \\ \\ \dfrac{1}{x} & \text{si } x \geq v \ . \end{cases}$$

Pour les 5 valeurs supérieures à 1, nous calculons $g(x_i) = \frac{1}{x_i}$, les autres correspondant à des valeurs de $g(x_i) = 0$:

i	x_i	$g(x_i) = \frac{1}{x_i}$
1	1,47	0,68
2	2,04	0,49
3	0,46	0
4	0,13	0
5	1,27	0,79
6	2,12	0,47
7	0,38	0
8	0,08	0
9	0,36	0
10	1,20	0,83

Tab. 6.5. *Transformation des valeurs simulées d'une distribution pour l'estimation d'une valeur d'intégrale.*

Nous déterminons ξ :

$$\hat{\xi} = \frac{1}{10} \sum_{i=1}^{10} g(x_i) = 0,326.$$

En résumé, nous constatons que pour intégrer une fonction par la méthode de Monte Carlo, il faut pouvoir l'écrire sous la forme d'un produit de deux autres fonctions, dont l'une est une fonction de densité d'une distribution.

Remarque 6.1 *La fonction de densité n'apparaît pas toujours aussi explicitement que dans les deux exemples susmentionnés. Par exemple, si l'on doit intégrer :*

$$h(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-b}{c}\right)^2}}{\log x},$$

il faut poser :

$$h_2(x) = \frac{1}{\sqrt{2\pi}c} e^{-\frac{1}{2}\left(\frac{x-b}{c}\right)^2} .$$

D'où $h_1(x) = \frac{\sqrt{2\pi}c}{\log x}$. *On a ainsi* $h(x) = h_1(x)h_2(x)$.

Le principe de l'estimation d'une intégrale peut être appliqué au cas d'intégrales multiples.

Soit $h_2(x_1, \ldots, x_n)$ une fonction de n variables définie sur un domaine $D \subseteq \mathbb{R}^n$, et qui sur ce domaine est une densité conjointe de n variables aléatoires $X_1, \ldots, X_n$. Soit h_1 une autre fonction de n variables quelconque. Alors l'intégrale multiple

$$I = \int_D h_1(x_1, \ldots, x_n) h_2(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

peut être estimée à travers la somme

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} h_1(x_1^{(i)}, \ldots, x_n^{(i)})$$

où pour $i = 1, \ldots, N$ la suite $(x_1^{(i)}, \ldots, x_n^{(i)})$ est un échantillon issu en accord avec la distribution multivariée h_2, par exemple en utilisant l'algorithme de Metropolis-Hastings.

Exemple 6.4 *Soit* $f(x, y)$ *une fonction de deux variables proportionnelle à* $1 - x^2 - y^2$ *définie sur le cercle* $C = \{(x, y),\ x^2 + y^2 \leq 1\}$, *et dont la constante de proportionnalité est telle que* $f(x, y)$ *est une fonction de densité conjointe. Nous voulons estimer l'intégrale :*

$$J = \int\int_C e^{-(x^2+y^2)} f(x, y) dx dy.$$

En utilisant l'algorithme de Metropolis-Hastings, il faut un échantillon fictif $(x_1, y_1), \ldots, (x_N, y_N)$ de variables aléatoires (X, Y) dont la loi conjointe est f. Si l'on reprend $N = 20$, et l'échantillon simulé dans l'Exemple 4.11, nous avons l'estimation :

$$\hat{J} = \frac{1}{20} \left(4 + e^{(-0{,}217^2 - 0{,}671^2)} + \cdots + 2e^{(-0{,}548^2 - 0{,}601^2)}\right) \simeq 0{,}785\ 4.$$

Dans le chapitre suivant nous verrons comment implémenter un algorithme pour des calculs d'intégrales comme celui ci-dessus.

6.6 Gestion de stocks

L'utilisation de la simulation rencontre de nombreuses applications dans le domaine du management, de la gestion ou du marketing. En matière de gestion de stocks, la demande pour un article soit est suivie de sa vente au client, soit doit faire face à une pénurie causée par son absence dans le stock. La vente d'un article est elle-même suivie de la mise à jour du stock et, s'il y a lieu, du déclenchement d'une commande auprès du fournisseur. Aujourd'hui il est courant de mettre à jour en temps réel un inventaire, et cela de manière automatique, par exemple par lecture optique des codes à barres des articles qui abandonnent un stock.

Nous allons illustrer ici, par un exemple simple, la solution d'un problème d'inventaire.

Exemple 6.5 *Les statistiques d'un concessionnaire d'une grande marque de voitures sur la demande journalière d'un certain modèle présentent la distribution suivante :*

Quantité demandée	0	1	2	3	4	5
Probabilité	0,10	0,10	0,20	0,30	0,20	0,10

Tab. 6.6. *Statistiques sur la demande journalière.*

De même, le délai de livraison peut varier entre 2 *et* 5 *jours avec les probabilités suivantes :*

Délai de livraison (en jours)	2	3	4	5
Probabilité	0,20	0,50	0,20	0,10

Tab. 6.7. *Statistiques sur les délais de livraison.*

L'objectif du gestionnaire du stock sera d'avoir constamment l'objet en stock, tout en respectant certaines contraintes élémentaires de coût, à savoir éviter d'avoir un stock trop important et des commandes en trop faibles quantités (économies d'échelle).

Deux éléments sont à déterminer, à savoir : la quantité en stock T critique qui déclenche une nouvelle commande et la quantité à commander Q. En supposant que les valeurs Q et T soient fixées par le gestionnaire, il s'agit de simuler le processus sur une période donnée et de répondre aux questions suivantes : quel a été le nombre moyen d'objets en stock par jour ? Quel a été le nombre de jours durant lesquels l'objet manquait en stock ? Quel a été le nombre de clients n'ayant pas pu être servis ?

On génère deux séries de nombres aléatoires compris par exemple entre 0 et 99 afin de simuler les variables aléatoires représentant les quantités demandées

et les délais de livraison. Les nombres aléatoires sont répartis selon la procédure suivante :

Quantité demandée	Probabilité	Nombre aléatoire
0	1/10	0 - 09
1	1/10	10 - 19
2	1/5	20 - 39
3	3/10	40 - 69
4	1/5	70 - 89
5	1/10	90 - 99

Tab. 6.8. *Schéma pour la simulation des quantités demandées.*

Délai livraison	Probabilité	Nombre aléatoire
2	1/5	0 - 19
3	1/2	20 - 69
4	1/5	70 - 89
5	1/10	90 - 99

Tab. 6.9. *Schéma pour la simulation des délais de livraison.*

Le stock au début de l'exercice est de 10 objets. Le gestionnaire effectue une simulation pour $T = 4$ (on passe commande dès que le nombre d'objets en fin de journée est égal ou inférieur à 4) et $Q = 12$ (on commande 12 objets à la fois). La simulation effectuée sur 40 jours ouvrables est présentée dans la Table 6.10.

La simulation fondée sur l'option du gestionnaire de passer une commande de 12 objets à la fois, dès qu'il n'en reste plus que 4 ou moins en stock en fin de journée amène aux conclusions suivantes :

(a) nombre moyen d'objets en stock : $187/40 = 4{,}67$;

(b) nombre de jours durant lesquels l'objet manque en stock : 10 jours ;

(c) nombre de clients n'ayant pas pu être servis : 22.

Compte tenu du délai de livraison et de la demande journalière, on s'aperçoit que le gestionnaire passe commande beaucoup trop tard. En prenant par exemple $T = 6$ et en utilisant les mêmes nombres aléatoires, le nombre moyen d'objets par jour serait de 5,45, le nombre de jours durant lesquels l'objet manque en stock 7, et le nombre de clients n'ayant pas pu être servis 11.

Il s'agit dès lors pour le gestionnaire de refaire la simulation en modifiant les valeurs T et Q jusqu'à ce qu'il obtienne des résultats compatibles avec ses

attentes en tenant compte, d'une part, des frais de transport et, d'autre part, des coûts de stockage.

Jour	Livr.	Stock matin	Nombre aléatoire	Vente	Stock soir	Déficit stock	Nombre aléatoire	Délais jours
1		10	60	3	7			
2		7	16	1	6			
3		6	24	2	4		46	3
4		4	60	3	1			
5		1	88	4	0	–3		
6	12	12	38	2	10			
7		10	02	0	10			
8		10	37	2	8			
9		8	45	3	5			
10		5	55	3	2		33	3
11		2	84	4	0	–2		
12		0	34	2	0	–2		
13	12	12	06	0	12			
14		12	21	2	10			
15		10	82	4	6			
16		6	39	2	4		86	4
17		4	48	3	1			
18		1	34	2	0	–1		
19		0	90	5	0	–5		
20	12	12	62	3	9			
21		9	35	2	7			
22		7	06	0	7			
23		7	55	3	4		85	4
24		4	46	3	1			
25		1	39	2	0	–1		
26		0	59	3	0	–3		
27	12	12	36	2	10			
28		10	38	2	8			
29		8	00	0	8			
30		8	70	4	4		07	2
31		4	14	1	3			
32	12	15	77	4	11			
33		11	90	5	6			
34		6	02	0	6			
35		6	51	3	3		90	4
36		3	47	3	0			
37		0	43	3	0	–3		
38		0	24	2	0	–2		
39	12	12	79	4	8			
40		8	24	2	6			

Tab. 6.10. *Résultats d'une simulation en gestion de stocks.*

6.7 Analyse de la rentabilité d'un investissement

La mise en place de projets d'investissement doit tenir compte d'un environnement économique et social particulièrement instable. Dans un monde régi

par l'incertitude, il est difficile de prévoir à plus ou moins long terme le prix des matières premières, le prix de vente, le risque de change, le taux de croissance du marché considéré ou la durée de vie d'un produit pour ne citer que quelques variables. Le plus souvent, on considère individuellement chaque facteur sujet à fluctuation et on calcule une estimation supposée être la meilleure. Il n'existe toutefois aucune garantie pour que cette estimation donne le rendement réel de l'investissement. Une alternative est d'utiliser les techniques de simulation en associant une probabilité à chaque facteur indéterminé.

Exemple 6.6 *Disposant d'un budget annuel de* 12 000 *euros, correspondant aux frais fixes pour faire tourner l'entreprise, on sait que le prix de vente de l'objet produit variera entre deux limites, selon une distribution de probabilité donnée et que la quantité produite sera fonction de différents facteurs sujets à fluctuation, tout comme les charges variables.*

Dans un environnement de forte concurrence, le prix est fixé par le marché et non par le producteur. Sur la base de l'évolution passée, on sait toutefois qu'il peut fluctuer selon les probabilités suivantes :

Prix	Probabilité	Nombre aléatoire assigné
14,50	1/10	0 - 09
15,00	1/10	10 - 19
15,50	1/5	20 - 39
16,00	1/2	40 - 89
16,50	1/10	90 - 99

Tab. 6.11. *Schéma pour la simulation des prix de vente.*

Les coûts de production sont également soumis à certaines variations environnementales et peuvent prendre les valeurs associées aux probabilités suivantes :

Coût de production par unité	Probabilité	Nombre aléatoire assigné
9,60	1/5	0 - 19
9,80	3/10	20 - 49
10,00	1/5	50 - 69
10,20	1/5	70 - 89
10,40	1/10	90 - 99

Tab. 6.12. *Schéma pour la simulation des coûts de production.*

La quantité produite va également être soumise à quelques aléas tels que l'absence du personnel et la qualité de la matière première :

Quantité produite (pièces)	Probabilité	Nombre aléatoire assigné
1 800	1/20	0 - 4
1 900	1/20	5 - 9
2 000	1/10	10 - 19
2 100	1/5	20 - 39
2 200	2/5	40 - 79
2 300	1/10	80 - 89
2 400	1/20	90 - 94
2 500	1/20	95 - 99

Tab. 6.13. *Schéma pour la simulation des quantités produites.*

Par souci de simplification, on suppose que les variables aléatoires sont statistiquement indépendantes (dans le cas contraire, il faudrait inclure cette dépendance dans le modèle). Le profit de l'entreprise sous sa forme la plus simple est calculé selon la relation suivante :

	(Prix de vente unitaire) · (Quantité produite)
−	(Coût variable unitaire) · (Quantité produite)
−	Frais fixes
=	Profit de l'entreprise.

La fonction de profit de l'entreprise peut évidemment être étendue en introduisant de nouvelles variables aléatoires ou en lui assignant des lois de probabilité bien définies. Une possibilité serait de construire un arbre de décision et de calculer les probabilités pour chacune des situations. Dans cet exemple, si la variable *Prix* peut prendre 4 valeurs, la variable *Coûts* 5 et la variable *Quantité produite* 8, on dénombre $4 \cdot 5 \cdot 8 = 160$ états possibles. Il est par conséquent peu recommandé de construire un tel arbre. En simulant 25 fois le prix de vente unitaire, le coût variable unitaire et la quantité produite selon les distributions de probabilités indiquées ci-dessus et en admettant des frais fixes d'un montant de 12 000 euros, pour la période considérée, on obtient un tableau de simulation illustré dans le Tableau 6.14.

On obtient, pour la première itération, les valeurs simulées suivantes : un prix de vente de 16 euros, un coût unitaire de 9,80 euros pour une quantité produite de 2 500 unités, entraînant un résultat positif de 3 500 euros ($= 6{,}2 \times 2\,500 - 12\,000$). En exécutant la procédure 20 fois, on obtient un prix moyen de 15,80 euros, un coût de production moyen de 9,95 euros, pour une quantité produite moyenne de 2 150, ce qui entraîne un résultat positif moyen de 578 euros.

Il va de soi que répéter 20 fois cette opération ne suffit pas à estimer précisément le résultat attendu. La capacité de l'outil informatique permet aujourd'hui des simulations massives qui produisent des estimations bien plus précises.

Expérience	Nombre aléatoire	Prix	Nombre aléatoire	Coût variable	Nombre aléatoire	Quantité produite	Résultat de bilan
1	51	16,0	31	9,8	98	2 500	3 500
2	95	16,5	03	9,6	28	2 100	2 490
3	63	16,0	05	9,6	15	2 000	800
4	43	16,0	94	10,4	61	2 200	320
5	60	16,0	38	9,8	69	2 200	1 640
6	73	16,0	30	9,8	71	2 200	1 640
7	54	16,0	14	9,6	32	2 100	1 440
8	05	14,5	23	9,8	37	2 100	–2 130
9	46	16,0	82	10,2	00	1 800	–1 560
10	48	16,0	99	10,4	88	2 300	880
11	16	15,0	71	10,2	43	2 200	–1 440
12	17	15,0	77	10,2	50	2 200	–1 440
13	24	15,5	65	10,0	45	2 200	100
14	86	16,0	15	9,6	37	2 100	1 440
15	51	16,0	31	9,8	75	2 200	1 640
16	85	16,0	20	9,8	82	2 300	2 260
17	75	16,0	94	10,4	13	2 000	–800
18	96	16,5	61	10,0	63	2 200	2 300
19	56	16,0	35	9,8	13	2 000	400
20	11	15,0	79	10,2	38	2 100	–1 920

Tab. 6.14. *Simulation de* 20 *exercices.*

En répétant la procédure 5 000 fois, on obtient un ensemble de simulations qui fournit un résultat positif moyen de 517,50 euros, ce qui correspond à un rendement annuel attendu de 4,312 5 %.

Exercices

6.1 Supposons que le temps pour acheminer un SMS suit une loi dont la densité de distribution est

$$f(t) = \begin{cases} \dfrac{1}{t^2} & t > 1 \\ \\ 0 & t \leq 1 \end{cases}$$

où t représente des secondes.

(a) Utiliser les nombres aléatoires de la table annexe pour simuler un échantillon de 10 temps d'acheminement.

(b) Supposons que les SMS soient envoyés avec un accusé de réception : un SMS de confirmation est automatiquement envoyé, dès que le SMS a été reçu. Simuler 10×2 temps entre l'envoi d'un SMS, son acheminement à destination, et la réception de l'accusé de réception.

6.2 Estimer et calculer l'intégrale suivante à l'aide de la méthode de Monte Carlo :

$$\int_{-2}^{2} \sqrt{4 - x^2} dx \ .$$

Que remarquez-vous dans l'expression de l'intégrale ?

6.3 Le gérant d'un magasin a observé que le temps de service (X) de sa caissière est distribué selon les probabilités suivantes :

Temps de service	Probabilité
3 minutes	0,6
4 minutes	0,2
5 minutes	0,1
6 minutes	0,1

Le nombre moyen de clients arrivant en 1 heure est de 6.
Le gérant désirerait savoir si l'attente actuelle est acceptable ou non. Dans le cas où celle-ci ne l'est pas, le gérant, pour ne pas perdre une partie de sa clientèle, décidera d'engager une seconde caissière.
Si le temps d'interarrivée (Y) est régi par une loi exponentielle, aider le gérant à prendre sa décision en simulant le phénomène sur 2 heures.

6.4 L'administrateur d'une gare dispose des statistiques suivantes concernant le temps de service à un guichet :

Temps de service	Probabilité
9 minutes	0,6
10 minutes	0,2
13 minutes	0,2

Il sait aussi qu'en moyenne il y a 10 clients par heure. Supposant que les clients arrivent indépendamment les uns des autres, qu'ils font une queue unique et qu'ils soient servis au premier guichet libre, simuler un phénomène de file d'attente pendant 1 heure, où les clients ont à leur disposition 2 guichets.

(a) Quelle est l'attente totale des clients ?

(b) Combien de clients ont dû attendre avant d'être servis ?

6.5 Dans un championnat, trois voitures, A, B et C, ont respectivement une vitesse moyenne de 210 km/h, de 205 km/h et de 200 km/h, mais la voiture A a une panne qui l'empêche de continuer une course avec probabilité 0,3, la voiture B a une panne avec probabilité 0,15, la voiture C a une panne avec probabilité 0,1. Utiliser les nombres aléatoires de la table annexe pour simuler un championnat de 10 courses où 10 points sont assignés à la voiture qui gagne ; 6 points à la deuxième arrivée ; 4 points à la troisième et aucun point à une voiture qui tombe en panne.

6.6 Estimer l'intégrale suivante :

$$\int_0^1 \frac{e^{-x}}{1+x^2} dx$$

en utilisant un échantillon de 10 nombres aléatoires.

6.7 Supposons que le jour $t_0 = 0$ l'euro vaut $p_0 = 1{,}54$ francs suisses. Chaque jour le taux de change de l'euro face au franc varie de façon aléatoire selon les probabilités suivantes :

Si $p_t < 1{,}53$:

$$p_{t+1} = \begin{cases} p_t + 0{,}01 & \text{avec probabilité } \frac{1}{2} \\ p_t & \text{avec probabilité } \frac{1}{4} \\ p_t - 0{,}01 & \text{avec probabilité } \frac{1}{4}. \end{cases}$$

Si $1{,}53 \leq p_t \leq 1{,}55$:

$$p_{t+1} = \begin{cases} p_t + 0{,}01 & \text{avec probabilité } \frac{2}{5} \\ p_t & \text{avec probabilité } \frac{1}{5} \\ p_t - 0{,}01 & \text{avec probabilité } \frac{2}{5}. \end{cases}$$

Si $p_t > 1{,}55$:

$$p_{t+1} = \begin{cases} p_t + 0{,}01 & \text{avec probabilité } \frac{1}{4} \\ p_t & \text{avec probabilité } \frac{1}{4} \\ p_t - 0{,}01 & \text{avec probabilité } \frac{1}{2}. \end{cases}$$

Au départ, un investisseur dispose de 10 000 euros et de 15 000 francs. Il décide de changer ses euros contre des francs à chaque fois que un euro vaut $p_t \geq 1{,}55$, et de changer ses francs contre des euros à chaque fois que un euro vaut $p_t \leq 1{,}53$.

Simuler le comportement de son investissement après 20 jours (c'est-à-dire au temps $t = t_{20}$) à l'aide d'une table de nombres aléatoires.

6.8 Dans la discipline sportive de 100 m, huit coureurs (A, B, C, D, ..., H) doivent se qualifier pour les jeux Olympiques. Trois coureurs se qualifient. Le coureur A court le 100 m en 10 secondes en moyenne, le coureur B en 10"01, le coureur C en 10"02, et ainsi de suite jusqu'au coureur H qui court le 100 m en 10"07, tous avec des prestations distribuées selon des lois normales avec un écart-type de 10 centièmes de seconde. Simuler à l'aide d'un ordinateur 200 courses de 100 m et estimer la probabilité que le coureur D se qualifie aux jeux Olympiques.

6.9 Le gérant d'un magasin a observé que le temps de service (X) de l'employé à la caisse (en minutes) est distribué selon la fonction de densité

$$f(x) = \begin{cases} \frac{1}{10} e^{-\frac{1}{10}x} & \text{si } x \geq 0 \\ \\ 0 & \text{sinon.} \end{cases}$$

Le nombre moyen de clients arrivant en 1 heure est de 5. Le gérant désirerait savoir si l'attente des clients à la caisse est acceptable ou non. Dans le cas où celle-ci ne l'est pas, le gérant, pour ne pas perdre une partie de sa clientèle, décidera d'ouvrir une autre caisse. Si le temps d'interarrivée (Y) est régi par une loi exponentielle, aider le gérant à prendre sa décision en simulant le phénomène sur 2 heures.

6.10 Estimez à l'aide de la méthode de Monte Carlo l'intégrale :

$$I = \int_1^\infty \frac{|\sin x|}{x^2} dx$$

en utilisant 20 nombres aléatoires de votre choix.

6.11 Le prix d'un litre de vin de table « standard » est de 1,90 euros. Ce prix est déterminé par la récolte de l'année à grande échelle. L'historique des années montre des statistiques selon lesquelles on peut avoir des années bonnes où la production atteint $C = 1{,}1$ fois la production moyenne ; des années moyennes $C = 1$; et des années mauvaises où la production n'est que $C = 0{,}9$ fois la production moyenne ; selon les mêmes statistiques, une année bonne arrive avec probabilité 0,25 ; une année moyenne avec probabilité 0,5 et une année mauvaise avec probabilité 0,25.

Le prix du vin à l'année t est déterminé par :

(a) l'offre effective $O_t = C \cdot T_t$ où T_t représente la tendance de l'offre $T_t = f(p_{t-1})$;

(b) la demande $D_t = g(p_t)$;

(c) et l'équation $O_t = D_t$.

La fonction d'offre est

$$f(p) = \frac{1}{2} p.$$

La fonction de demande est

$$g(p) = \frac{4}{p^2}.$$

À l'aide d'une table de nombres aléatoires, simuler le comportement du prix du vin sur les 10 ans à venir et calculer la moyenne et l'écart-type constatés.

6.12 Une station de ski des Alpes désire étudier sur 15 ans l'évolution des prix de l'offre d'hébergement selon les conditions climatiques. L'offre à la saison d'hiver t dépend du prix des chambres de la saison d'hiver précédente selon la formule $O_t = e^{p_{t-1}} - 1$, où p_t est le prix à l'année t exprimé en centaines d'euros. La demande dépend du prix de la saison en cours et des conditions d'enneigement selon la formule $D_t = C_t \cdot 1/p_t^3$. Le coefficient C_t peut avoir une valeur comprise entre 0,5 (absence de neige) et 1,5 (enneigement optimal), et cela de manière uniforme et équiprobable sur l'intervalle $[0{,}5; 1{,}5]$. En supposant que le prix (en centaines d'euros) à l'année 0 soit 0,81, que le marché évolue à un prix tel que la demande soit égale à l'offre, et en utilisant une méthode de votre choix, éventuellement à l'aide des nombres aléatoires en annexe, décrivez un critère adéquat pour effectuer la simulation (comment utilisez-vous les nombres aléatoires, comment associez-vous les probabilités aux événements et aux nombres aléatoires, etc.) et illustrez-le.

Chapitre 7

Simulation assistée par ordinateur

Le but d'une simulation est de contourner au moyen de calculs répétés la difficulté d'un problème due à l'interaction aléatoire de plusieurs facteurs.

Dans le chapitre précédent nous avons vu comment les estimations en simulation peuvent être incertaines si des nombres aléatoires sont utilisés en quantité insuffisante. Une simulation demande souvent des calculs en grande quantité. C'est pourquoi les techniques de Monte Carlo se sont affirmées de plus en plus comme un instrument précieux depuis la seconde moitié du siècle passé, grâce notamment à la facilité avec laquelle les ordinateurs accomplissent ce type de travail.

Nous verrons dans ce chapitre quelques exemples de simulation assistée par ordinateur. Le but sera de familiariser le lecteur avec ses techniques, en lui donnant les moyens de pouvoir faire par soi-même toute sorte de simulation.

7.1 Un cas d'estimation d'une surface

Nous avons vu comment estimer la surface d'une région délimitée par une courbe dans le plan à l'aide d'une simulation. Nous avons aussi discuté de la précision d'une telle estimation. En particulier quand on veut une bonne approximation d'une telle estimation, il faut considérer un grand nombre de points distribués de manière aléatoire sur une région qui contient la surface que l'on veut estimer.

Or il est impensable d'effectuer des milliers de calculs à la main. Nous allons illustrer ici comment effectuer ces calculs à l'aide d'un exemple.

Pour venir à bout de ce problème nous allons utiliser un logiciel statistique très populaire comme Minitab. De ce logiciel nous utiliserons seulement des fonctions élémentaires manipulant des lignes et des colonnes d'une table. Donc nous donnerons les instructions pour ce logiciel, tout en rappelant que d'autres

logiciels tout aussi populaires comme SPSS ou Excel permettent d'effectuer les mêmes procédures avec quelques différences dans la syntaxe.

Exemple 7.1 *On considère la région D à l'intérieur de la courbe définie par l'équation*

$$x^4 + y^4 - x - y - 1 = 0.$$

Il est clair, comme on peut le voir dans la Figure 7.1*, que $D \subseteq [-2, 2] \times [-2, 2]$. Mais comment évaluer la surface de D par la méthode de Monte Carlo avec trois chiffres décimaux significatifs ?*

Nous remarquons que la région D correspond aux points (x, y) tels que $f(x, y) = x^4 + y^4 - x - y - 1$ est négative. Dans le chapitre précédent nous avons vu que pour une précision de trois décimales il est nécessaire de considérer environ un million de points.

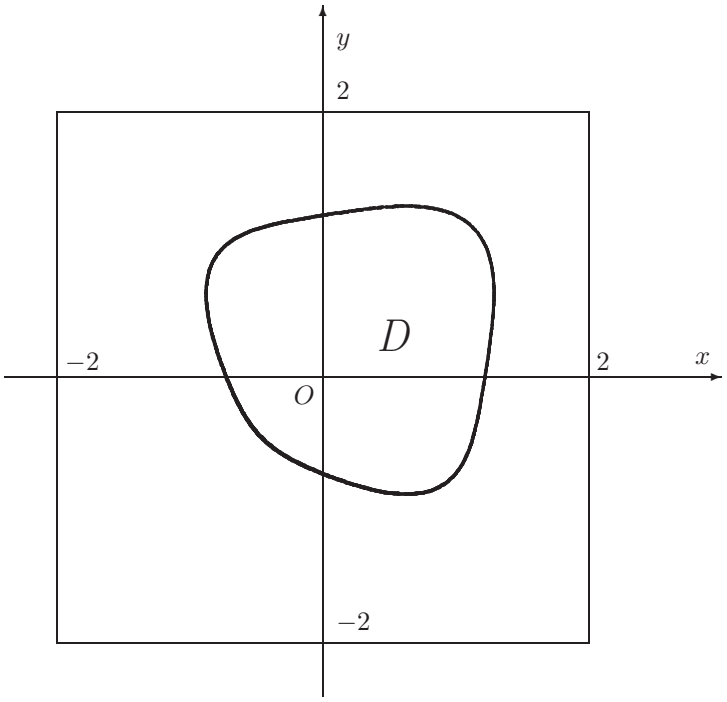

Fig. 7.1. *La région D entourée du carré $[-2, 2] \times [-2, 2]$.*

Il faut donc simuler un million de points aléatoires dans le rectangle $[-2, 2] \times [-2, 2]$. Pour cela il suffit d'assigner à la première et à la deuxième colonne d'un tableur un million de valeurs aléatoires entre -2 et 2. Avec Minitab :

```
MTB> Random 1000000 c1;
SUBC>  Uniform -2 2.
MTB> Random 1000000 c2;
SUBC>  Uniform -2 2.
```

Ensuite on assigne à la troisième colonne la valeur $f(x, y)$ où x et y sont les coordonnées figurant sur les deux premières colonnes.

```
MTB > Let c3 = c1**4+c2**4-c1-c2-1
```

Ensuite on assigne à la quatrième colonne la valeur 1 si $f(x, y) \leq 0$ et 0 autrement :

```
MTB > Let c4 = 0.5*(1-sign(c3))
```

Pour savoir combien parmi le million de points tombent à l'intérieur de D nous allons compter les « 1 » de la quatrième colonne :

```
MTB > Let c5 = sum(c4)
```

Nous obtenons une valeur n qui divisée par 1 000 000 représente la probabilité qu'un point au hasard de $[-2, 2] \times [-2, 2]$ tombe à l'intérieur de D. Puisque la surface de $[-2, 2] \times [-2, 2]$ est 16, la surface de D est estimée par

$$\text{Aire } (D) = 16\frac{n}{1\ 000\ 000}.$$

Si par exemple $n = 235\ 128$, l'estimation de la surface de D est 3,762.

7.2 Une simulation d'une file d'attente

Souvent, lors d'une simulation d'un phénomène, il est nécessaire de faire en sorte que plusieurs variables simulées interagissent au fur et à mesure que le phénomène évolue.

Dans ce genre de situation, une approche assistée par ordinateur demande l'implémentation d'un algorithme.

L'un des logiciels le plus répandu aujourd'hui est sans doute le logiciel gratuit `R`. Au moment où ce livre est édité il peut être téléchargé à l'adresse `www.r-project.org` et offre toute une série de fonctions très utiles en statistique.

Sa relative simplicité permet au néophyte de se mesurer aussi dans la programmation.

Nous allons voir ici un exemple de programmation en `R`. Nous reprenons le problème considéré dans le Paragraphe 6.3 où des clients doivent choisir le premier guichet qui se libère parmi 3 guichets qui sont ouverts au public.

Exemple 7.2 *Supposons que l'agence de voyages de l'Exemple* 6.1 *soit ouverte* 8 *heures par jour. Comment évaluer l'attente moyenne des clients et l'opportunité ou non de garder* 3 *guichets ?*

Dans un problème comme celui-là nous sommes confrontés à deux types de difficultés. D'abord, effectuer une simulation manuelle apparaît extrêmement difficile, car pour une seule simulation il faut s'attendre à l'arrivée d'environ 120 clients, ce qui entraîne la compilation d'un tableau de simulation contenant autant de lignes. Et deuxièmement, contrairement au cas de l'évaluation d'une surface, où le nombre de points était directement lié à la précision de l'estimation, une simulation seulement ne dit rien sur la fiabilité de l'estimation qui est faite. Ainsi, si l'on veut construire un intervalle de confiance pour l'estimation de l'attente moyenne, il faudrait répéter la simulation un grand nombre de fois, un peu comme on le fait en rééchantillonnage.

Le code ci-dessous définit une fonction appelée « `fdat` » qui permet de calculer la moyenne des temps d'attente et la moyenne des temps de service. À noter que, selon la définition donnée, la moyenne sur le long terme des temps de service a une valeur espérée de 12 minutes.

```
fdat <- function(n) {
#
# l'instruction fdat(n) effectue une simulation d'une
# file d'attente à trois guichets pendant n minutes,
# où les clients arrivent selon une poissonienne en moyenne
# toutes les 4 minutes et les guichets nécessitent
# en moyenne 12 minutes pour servir un client
#
truen <-numeric(n)      # il est 1 quand z ne dépasse pas
                        # n (minutes), 0 autrement
x <- numeric(n)         # RND pour interarrivée
tt <- numeric(n)        # RND pour service
z <- numeric(n)         # minute arrivée
m <- numeric(n)         # minutes "interarrivée"
d1 <- numeric(n)        # minute départ guichet 1
d2 <- numeric(n)        # minute départ guichet 2
d3 <- numeric(n)        # minute départ guichet 3
d1new <- numeric(n)     # minute départ guichet 1 (auxiliaire)
d2new <- numeric(n)     # minute départ guichet 2 (auxiliaire)
d3new <- numeric(n)     # minute départ guichet 3 (auxiliaire)
g <- numeric(n)         # numéro du guichet
v <- numeric(n)         # minute à laquelle il sera servi
att <- numeric(n)       # attente
R <- matrix(0,n,10)     # matrice résumé
t <- numeric(n)         # minute de service
x[1] <- runif(1,0,1)
tt[1] <- runif(1,0,1)
t[1] <- 2-log(tt[1])*10
m[1] <- -4*log(x[1])
z[1] <- m[1]
v[1] <- 0
d1[1] <- z[1] + t[1]
d2[1] <- 0
```

```
d3[1] <- 0
g[1] <- 1
att[1] <- 0
for(i in 2:n){
x[i] <- runif(1,0,1)
tt[i] <- runif(1,0,1)
t[i] <- 2-log(tt[i])*10  # temps de service pour i
m[i] <- -4*log(x[i])     # minute interarrivée de i
z[i] <- z[i-1] + m[i]    # mn arrivée
att[i] <- max(0,min((d1[i-1]-z[i]),(d2[i-1]-z[i]),(d3[i-1]-z[i])))
v[i] <- max(z[i],min(d1[i-1],d2[i-1],d3[i-1]))
# définition du guichet et départ provisoires dinew
if (d3[i-1] == min(d1[i-1],d2[i-1],d3[i-1]))
{
g[i] <- 3
d3new[i] <-v[i]+t[i]
d2new[i] <- d2[i-1]
d1new[i] <- d1[i-1]
}
if (d2[i-1] == min(d1[i-1],d2[i-1],d3[i-1]))
{
g[i] <- 2
d3new[i] <-d3[i-1]
d2new[i] <- v[i]+t[i]
d1new[i] <- d1[i-1]
}
if (d1[i-1] == min(d1[i-1],d2[i-1],d3[i-1]))
{
g[i] <- 1
d3new[i] <-d3[i-1]
d2new[i] <- d2[i-1]
d1new[i] <- v[i]+t[i]
}
# fin définition du guichet

# définition des minutes de départ dans les guichets

d1[i] <- d1new[i]
d2[i] <- d2new[i]
d3[i] <- d3new[i]

# fin de définition des minutes de départ dans les guichets

#truen[i] <- max(floor(1-(max(z[i],n)-n)),1)
truen[i]<- max(floor(-max(0,z[i]-n)+1),0)
} # fin du cycle "for"

# définition des différentes colonnes du tableau de simulation
# les clients arrivant aprês la minute n n'entrent pas en compte
```

```
w <- 1+sum(truen) # nombre d'arrivées avant minute n
att<-att[1:w]
t<-t[1:w]
x<-x[1:w]
tt<-tt[1:w]
m<-m[1:w]
z<-z[1:w]
g<-g[1:w]
d1<-d1[1:w]
d2<-d2[1:w]
d3<-d3[1:w]
v<-v[1:w]

matt <- mean(att); ttt <- mean(t)

#mettre les dièses à partir d'ici
##R <- round(matrix(cbind(x,tt,m,z,t,g,d1,d2,d3,att),ncol=10, nrow=w),2)
##     # dessine le graphe de l'attente ...
##plot(att,type="l",main = "attente", ylab="min")
##return(list(R = R))
S <- matrix(cbind(matt,ttt))
return(list(S=S))
}
```

L'interprétation des différentes lignes de ce code n'est pas difficile et des commentaires (le texte qui suit le symbole « # ») aident à la compréhension.

Ainsi, la commande

```
> fdat(480)
```

peut donner comme résultat :

```
          [,1]
[1,]  8,21143
[2,] 11,87713
```

ce qui signifie que la moyenne des temps d'attente pour une simulation de clients arrivés avant les 8 heures (480 minutes) pendant lesquelles l'agence a été ouverte au public est d'environ 8 minutes, et que la moyenne des temps de service a été de presque 12 minutes.

L'exécution de la procédure ne demande qu'une fraction de seconde. Donc cela vaut la peine de la répéter un certain nombre de fois pour avoir une idée de la stabilité des valeurs ainsi estimées. Après quelques dizaines de simulations on se rend compte que, en dépit de la première simulation dont le résultat suggérait le contraire, l'attente moyenne des clients est presque toujours supérieure au temps de service. La décision qui s'impose au gérant est celle d'ouvrir au moins un autre guichet.

À noter que le code qui est présenté ci-dessus permet aussi de produire un tableau de simulation similaire à ceux que l'on a vus au Chapitre 6 et un graphe montrant l'évolution de l'attente des clients tout au long d'une simulation. Pour cela il suffit d'enlever le symbole `##` dans les quatre lignes qui précèdent les deux dernières lignes du code (le symbole `#` indique un commentaire dans un code en `R`) et de mettre ces symboles dans les deux dernières lignes.

7.3 Échantillonnage d'une surface non plane

En sciences de la terre ou en science de l'environnement il est parfois nécessaire d'échantillonner une parcelle de terrain plus ou moins homogène. Si la parcelle est plane, cela est relativement facile, car cela se réduit à une simulation de points aléatoires sur un rectangle. Mais si la parcelle n'est pas plane, ou si l'on désire échantillonner davantage certains endroits plutôt que d'autres, le problème est plus complexe.

Un problème typique est celui de choisir des points distribués de manière aléatoire sur une surface non plane, par exemple pour simuler une distribution d'arbres sur une forêt, ou les quantités de neige cumulées sur un relief alpin.

Exemple 7.3 *Supposons d'avoir une surface S définie comme une fonction différentiable f sur un ensemble compact $D \subset \mathbb{R}^2$, c'est-à-dire :*

$$S = \left\{(x, y, f(x, y)) \in \mathbb{R}^3 | (x, y) \in D\right\}.$$

Choisir un certain nombre de points, disons n, aléatoirement répartis sur S, uniformément par unité de surface.

Supposons d'avoir à disposition un générateur de nombres aléatoires. Celui-ci fournit une suite $\{u_h\}_{h\in\mathbb{N}}$ avec $u_h \in [0, 1]$ tiré de $U(0, 1)$. Les étapes de l'algorithme sont (Melfi et Schoier, 2004) :

Étape 1 : Générer une distribution uniforme de N points dans D. Vu que D est un ensemble compact dans $\mathbb{R}^2$, il est borné et fermé et peut être contenu dans un rectangle $(a, b) \times (c, d)$. Par une transformation affine appropriée, des points aléatoires distribués uniformément (u_{2k-1}, u_{2k}) dans $[0, 1] \times [0, 1]$ peuvent être transformés en des points distribués uniformément dans D, sans considérer les points qui tombent éventuellement en dehors de D. Cette procédure permet de simuler le nombre souhaité de points aléatoires distribués uniformément sur D :

$$(x_i, y_i) \qquad \text{pour } i = 1, \ldots, N.$$

Étape 2 : Assigner à chaque point généré en D un nombre aléatoire, en utilisant encore le générateur de nombres pseudo-aléatoires distribués uniformément dans $[0, 1]$. Cette opération peut être considérée comme une fonction :

$$\omega : \{1, \ldots, N\} \longrightarrow [0, 1].$$

Étape 3 : En considérant la fonction

$$m(x,y) = \left(1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right)^{\frac{1}{2}}$$

définie sur D, calculer :

$$M = \max_D\{m(x,y)\} \ .$$

Vu que D est compact et f est différentiable, le maximum de $m(x,y)$ existe. La raison pour considérer cette fonction est qu'un élément de surface $\Delta x \Delta y$ correspondant à un point (x,y) en D est projeté à travers la fonction f dans un élément de S dont l'aire peut être approximée (Kaplan, 1992) par :

$$\left(1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right)^{\frac{1}{2}} \Delta x \Delta y \ .$$

Évidemment, $1 \leq M < \infty$.

Étape 4 : Sélectionner le point $(x_i, y_i, f(x_i, y_i))$ dans l'échantillon final de points aléatoires sur S si :

$$\omega(i) < \frac{m(x_i, y_i)}{M}.$$

Cette procédure du type acceptation-rejet permet de sélectionner en moyenne le même nombre de points par unité de surface sur S. La fonction $m(x,y)$ est utilisée pour corriger l'effet de la projection du plan D sur la surface S. Les nombres aléatoires $\omega \in [0,1]$ associés à chaque point simulé dans D sont comparés à la probabilité de sélection des points, qui est

$$p = \frac{m(x,y)}{M}.$$

Les points simulés dans D subissent une « distorsion » quand ils sont projetés sur la surface S : plus la surface est « différente » (en termes de pente) de la surface plane D et moins il y aura de points sur cette projection. La fonction $m(x,y)$ équilibre en quelque sorte cette distorsion. En effet, la fonction varie positivement avec la pente de la surface S, dans les directions de x et de y. En d'autres mots, les points en correspondance de parties « raides » de S ont plus de probabilité d'être sélectionnés que ceux qui se trouvent en correspondance de régions de « plaine ».

Ici nous présentons le code en `R` complet d'une fonction créé pour l'implémentation de l'algorithme ci-dessus. Ce code, avec toutes les options annexes, nous a été mis à disposition par Sandro Petrillo, à qui l'on doit aussi toutes les explications qui accompagnent le reste de ce paragraphe.

Avec `R` il y a deux librairies qui permettent de faire des graphiques en trois dimensions : `scatterplot3d` et `lattice`. Pour pouvoir utiliser ces librairies, il est nécessaire de les avoir installées. Les librairies `R` sont librement disponibles sur le site officiel du logiciel. Elles peuvent aussi être installées directement à partir de `R` avec les commandes `install.packages("scatterplot3d")` et `install.packages("lattice")`.

Pour pouvoir utiliser les fonctions dont le code `R` est présenté ci-dessous, il suffit de copier le tout dans la ligne de commande du logiciel, après quoi les fonctions seront disponibles.

```
#automated function for algorithm
urda<-function(N=1000,aa=-3,bb=3,cc=-3,dd=3,fxy=expression(6*
exp(-x^2-y^2))){
#creation of the compact set DD (aa,bb)x(cc,dd)
#Step 1: generation of N points in DD (uniformly distributed)
u12<-runif(2*N);i<-1:N; i1<-2*i; i2<-2*i - 1;
xi<-(u12[i1]*(bb-aa))+aa;
yi<-(u12[i2]*(dd-cc))+cc;
DD<-data.frame(xi,yi)
names(DD)<-c("x","y")

#Step 2: assignement of a uniform random number (0,1) to each
#random point generated in DD
w<-runif(N)

#Step 3: consider the function
#m1(x,y)=(1 + (d fxy/ d x)^2 + (d fxy/ d y)^2)^0.5 defined on DD
dfdx<-D(fxy,"x");dfdy<-D(fxy,"y");
m1<-sqrt(1+eval(dfdx,envir=DD)^2+eval(dfdy,envir=DD)^2)
#Compute M1=max_d(m1)
M1<-max(m1)

#Step 4: select the point (xi, yi, f(xi,yi)) in the final sample of
#random points on S if w_i<m1(x,y)/M1
i3<-numeric(N)
for(i in 1:N){
if(w[i]<m1[i]/M1) i3[i]<-1 else i3[i]<-0
}
f<-eval(fxy, envir=DD);frange<-range(f);
DD<-data.frame(DD,f);DDtot<-DD;DD<-DD[i3==1, ];

require(scatterplot3d)
s3d<-scatterplot3d(DD,highlight.3d=TRUE,pch=20,xlim=c(aa,bb),
ylim=c(cc,dd),
zlim=frange,main=fxy,sub=paste("Selected points: ",nrow(DD),
"over ",N))
list(DD=DD,selected=nrow(DD), prop.selected=nrow(DD)/N)
}
```

La fonction `urda()` que l'on vient de présenter prend les arguments suivants :

(a) `N` : le nombre de points que l'on veut simuler. Les points faisant partie de l'échantillon final seront seulement une partie (variable) de N. La valeur de défaut est $N = 1\,000$;

(b) `aa, bb, cc, dd` : ce sont les coordonnées du rectangle $(a, b) \times (c, d)$ de $D \subset \mathbb{R}^2$. Les valeurs de défaut sont `aa=-3, bb=3, cc=-3, dd=3` qui correspondent au rectangle $(-3, 3) \times (-3, 3)$;

(c) `fxy` : la fonction $f(x, y)$ qui définit la surface S sur laquelle on veut simuler des points aléatoires (uniformes). La valeur de défaut est :

```
expression(6*exp(-x^2-y^2))
```

correspondant à la fonction $f(x, y) = 6e^{-(x^2+y^2)}$. Cet argument doit être introduit sous la forme d'un objet du type `expression`, pour permettre le calcul des dérivées partielles.

En utilisant la fonction `urda()` sans arguments, l'algorithme est exécuté avec les valeurs de défaut. Donc, la commande :

```
urda()
```

simule $N = 1\,000$ points dans l'ensemble compact $D \subset \mathbb{R}^2$ (dans le carré $(-3, 3) \times (-3, 3)$), sur la surface S définie par la fonction $f(x, y) = 6e^{-(x^2+y^2)}$. Seulement une partie de ces 1 000 points est sélectionnée dans l'échantillon final de points sur S.

On présente ci-dessous quelques exemples de la fonction `urda()`, en utilisant différentes fonctions f, domaines D et nombre de points que l'on simule. Par exemple, si l'on veut simuler des points distribués de manière uniforme sur la surface définie par la fonction :

$$f(x, y) = x^2 + y^2$$

dans le domaine $(-1, 1) \times (-1, 1)$, il faut utiliser la commande suivante :

```
urda(N=1500, aa=-1, bb=1, cc=-1, dd=1, fxy=expression(x^2+y^2))
```

On peut aussi déclarer les arguments que l'on veut passer à la fonction séparément. Par exemple, les commandes :

```
l.bound<- -1; u.bound<- 1;
f<- expression(x^2-y^2);
urda(aa=l.bound,bb=u.bound,cc=l.bound,dd=u.bound,fxy=f)
```

vont simuler des points aléatoires uniformes sur la surface $f(x, y) = x^2 - y^2$ définie dans le domaine $D = (-1, 1) \times (-1, 1)$, en partant de la simulation de $N = 1\,000$ points dans D (argument de défaut).

Si l'on a installé la librairie `scatterplot3d`, la fonction crée automatiquement le graphique en trois dimensions.

Une dernière chose à remarquer est la possibilité de mémoriser le résultat d'une simulation dans un objet `R`. Par exemple, la commande :

```
sim.atan<-urda(N=500, fxy=expression(6*atan(x)))
```

va stocker les résultats de la simulation sur la surface $f(x,y) = 6 \arctan x$ (dans $D = (-3,3) \times (-3,3)$, le domaine de défaut) dans l'objet `sim.atan`. Ce dernier est un objet de type `list`, contenant les informations suivantes :

(a) `DDtot`, contenant les coordonnées (x,y) des points aléatoires simulés initialement dans D, les valeurs de la fonction $f(x,y)$ correspondantes, les nombres $\omega \sim U(0,1)$ assignés, les probabilités de sélection $m(x,y)/M$ et l'information `TRUE/FALSE` indiquant si le point est ou non sélectionné dans l'échantillon final de points sur la surface S ;

(b) `DD`, contenant les coordonnées $(x,y,f(x,y))$ des points qui ont été sélectionnés dans l'échantillon final de points sur la surface S ;

(c) `selected`, indiquant le nombre de points qui ont été sélectionnés ;

(d) `prop.selected`, indiquant la proportion de points sélectionnés dans l'échantillon final.

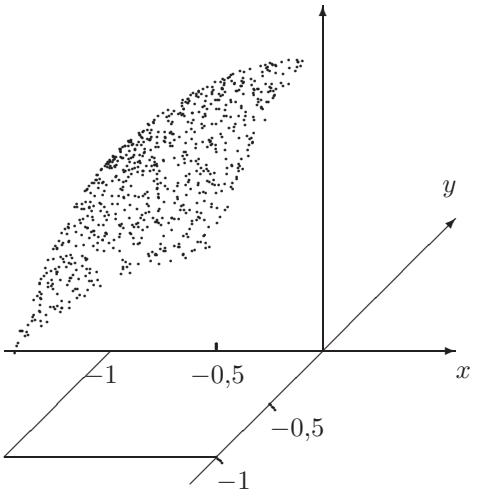

Fig. 7.2. *Simulation de* 753 *points aléatoires distribués de manière uniforme par unité de surface. Ici* $f(x,y) = 3 - x^2 - y^2$ *et la parcelle considérée est* $[-1,0] \times [-1,0]$ *sur le plan* xOy.

L'un des avantages de mémoriser le résultat d'une simulation dans un objet est de pouvoir utiliser les résultats par exemple pour changer les coordonnées du graphique, ou d'exporter les résultats dans un fichier texte permettant l'utilisation des données avec d'autres logiciels.

Les composantes de l'objet créé peuvent être invoquées avec le symbole `$`. Dans notre exemple, on peut voir les coordonnées des points sélectionnés dans l'échantillon final avec la commande :

```
sim.atan$DD
```

On peut par exemple faire le graphique en trois dimensions sous un autre point de vue avec la commande :

```
scatterplot3d(sim.atan$DD, angle=120, highlight.3d=TRUE,pch=20)
```

7.4 Intégrales multiples

Nous avons vu dans le Chapitre 6 comment, à l'aide de l'algorithme de Metropolis-Hastings, on peut estimer la valeur d'une intégrale multiple d'une fonction positive définie sur un domaine qui souvent peut être assez difficile d'expliciter autrement que par l'équation d'une courbe.

L'estimation se fondait sur la génération d'un échantillon fictif issu d'une certaine distribution conjointe de variables aléatoires.

L'estimation qui en découle peut donc varier d'une simulation à l'autre, et si l'on est intéressé par une estimation précise, disons à trois ou quatre décimales, il faut disposer d'un échantillon d'une certaine taille, et la procédure manuelle présentée au Paragraphe 6.5 ne peut plus s'envisager.

Nous présentons ici un exemple de calcul d'intégrale, estimé à travers l'algorithme de Metropolis-Hastings implémenté dans `R`.

La fonction à intégrer est la même que celle de l'Exemple 6.4, et, comme avant, l'intégrale sera dénotée par J :

$$J = \iint_C ke^{-(x^2+y^2)}(1 - x^2 - y^2)dxdy,$$

où k est tel que

$$\iint_C k(1 - x^2 - y^2)dxdy = 1$$

et où $C = \{(x, y) \ |x^2 + y^2 \leq 1\}$. Dans le code ci-dessous on dispose d'une fonction, « `int` », qui estime cette intégrale avec un échantillon de taille n.

```
#
#  Ce code calcule l'intégrale de
#  k*e^[-(x^2+y^2)]*(1-x^2-y^2)
#  dans le domaine où (1-x^2-y^2)>0,
#  où k est la constante telle que k*(1-x^2-y^2)
#  est une densité conjointe.
#  Il faut introduire le nombre d'itérations, n.
#  Par exemple int(n)
#
```

```
#
int <- function(n) {

# n  # nombre d'itérations souhaité

y1 <- numeric(n+1) # RND pour première coordonnée U(-1,1)
y2 <- numeric(n+1)  # RND pour deuxième coordonnée U(-1,1)
py <- numeric(n+1)  # p(y)
x1 <- numeric(n+1)
x2 <- numeric(n+1)
px <- numeric(n+1)
u  <- numeric(n+1) # RND U(0,1)
pp <- numeric(n+1)
h  <- numeric(n+1)

x1[1]=0 # Le burn in part de (0,0)
x2[1]=0 #

y1=runif(n+1,-1,1)
y2=runif(n+1,-1,1)

y1i=y1[1]
y2i=y2[1]
x1i=x1[1]
x2i=x2[1]

py[1]=max(0,1-y1i^2-y2i^2)
px[1]=1-x1i^2-x2i^2

u[1]<- runif(1,0,1)
pp[1]=py[1]/px[1]

if( u[1]<pp[1]) x1[2]=y1i else x1[2]=x1[1]
if( u[1]<pp[1]) x2[2]=y2i else x2[2]=x2[1]
  h[1]=exp(-(x1[1]^2+x2[1]^2))


for (i in 2:n){
y1i=y1[i]
y2i=y2[i]
x1i=x1[i]
x2i=x2[i]

py[i]=max(0,1-y1i^2-y2i^2)
px[i]=1-x1i^2-x2i^2

u[i]<- runif(1,0,1)
pp[i]=py[i]/px[i]
```

```
if( u[i]<pp[i]) x1[i+1]=y1[i] else x1[i+1]=x1[i]
if( u[i]<pp[i]) x2[i+1]=y2[i] else x2[i+1]=x2[i]
  h[i]=exp(-(x1[i]^2+x2[i]^2))
}


#R <- round(cbind(y1,y2,py,x1,x2,px,u,pp,h),3)
#return(list(R = R))
I=mean(h[-(n+1)])
return(list(I=I))
}
```

Ainsi, la commande `int(50000)` peut donner $\hat{J} = 0{,}734\,8$, une valeur correcte à un millième près.

Si l'on désire disposer d'un tableau de simulation dans le style du Tableau 4.7, avec une colonne supplémentaire pour les valeurs de la fonction h (ici $h(x, y) = e^{-(x^2+y^2)}$), il suffit d'enlever les dièses dans les dernières lignes du code ci-dessus et de les mettre devant les deux lignes suivantes.

Exercices

7.1 Un groupe de 20 touristes suisses doivent passer le contrôle des passeports à la douane française de la gare de Genève. Ils arrivent à la douane de la gare de Genève à 11 h et leur train pour la Côte d'Azur part à 11 h 15. Supposant que le contrôle des passeports prends en moyenne 30 secondes, que les temps de contrôle suivent une loi exponentielle, que les trains pour la France se trouvent à 1 minute de la douane et partent à l'heure, que le groupe monte dans le train seulement si tout le monde est sur le quai, estimer la probabilité que le groupe ne rate pas le train en simulant 100 fois le passage du groupe à la douane.

7.2 Utiliser 1 000 nombres aléatoires pour estimer l'intégrale

$$I = \int_0^1 \frac{|\sin x|}{x} dx.$$

7.3 En utilisant l'algorithme de Metropolis-Hastings, estimer l'intégrale

$$K = \int\int_{x^2+y^2 \leq \pi} \frac{\sin(x^2 + y^2)}{1 + x^2 + y^2} dxdy.$$

Appendice : Tables

Bien que les tables numériques soient intégrées dans tous les logiciels de calcul statistique, ces tables ont une fonction didactique importante.

Dans cette appendice nous présentons les tables numériques les plus utilisées. On pourra trouver donc la table du χ^2 ; la table de Fisher ; la table de Gauss sur les valeurs de la loi normale $\mathcal{N}(0,1)$; la table de Student et une table de nombres aléatoires, qui a été tirée des décimales de π.

Chaque table est précédée d'une brève explication.

Table du chi-carré

Pour

$$\chi^2_\nu(x) = \int_0^x \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} t^{\frac{\nu}{2}-1} e^{-\frac{t}{2}} dt,$$

les valeurs de cette table représentent des valeurs de x telles que

$$\chi^2_\nu(x) = \alpha,$$

avec α égal à une valeur correspondante au seuil de signification. Dans la théorie des tests d'hypothèses, ces valeurs de x sont dénotées par $\chi^2_{(\alpha,\nu)}$ et représentent des valeurs critiques.

	Seuil de signification α						
ν	0,990	0,975	0,950	0,10	0,05	0,025	0,01
1	0,00	0,00	0,00	2,71	3,84	5,02	6,63
2	0,02	0,05	0,10	4,61	5,99	7,38	9,21
3	0,11	0,22	0,35	6,25	7,81	9,35	11,34
4	0,29	0,48	0,71	7,78	9,49	11,14	13,23
5	0,55	0,83	1,14	9,24	11,07	12,83	15,09
6	0,87	1,24	1,63	10,64	12,53	14,45	16,81
7	1,23	1,69	2,16	12,02	14,07	16,01	18,48
8	1,64	2,18	2,73	13,36	15,51	17,53	20,09
9	2,08	2,70	3,32	14,68	16,92	19,02	21,67
10	2,55	3,25	3,94	15,99	18,31	20,48	23,21
11	3,05	3,82	4,58	17,29	19,68	21,92	24,72
12	3,57	4,40	5,23	18,55	21,03	23,34	26,22
13	4,11	5,01	5,89	19,81	22,36	24,74	27,69
14	4,66	5,63	6,57	21,06	23,68	26,12	29,14
15	5,23	6,26	7,26	22,31	25,00	27,49	30,58
16	5,81	6,91	7,96	23,54	26,30	28,85	32,00
17	6,41	7,56	8,67	24,77	27,59	30,19	33,41
18	7,02	8,23	9,39	25,99	28,87	31,53	34,81
19	7,63	8,91	10,12	27,20	30,14	32,85	36,19
20	8,26	9,59	10,85	28,41	31,41	34,17	37,57
21	8,90	10,28	11,59	29,62	39,67	35,48	38,93
22	9,54	10,98	12,34	30,81	33,92	36,78	40,29
23	10,20	11,69	13,09	32,01	35,17	38,08	41,64
24	10,86	12,40	13,85	33,20	36,42	39,36	42,98
25	11,52	13,12	14,61	34,38	37,65	40,65	44,31
26	12,20	13,84	15,38	35,56	38,89	41,92	45,64
27	12,88	14,57	16,15	36,74	40,11	43,19	46,96
28	13,57	15,31	16,93	37,92	41,34	44,46	48,28
29	14,26	16,05	17,71	39,09	42,56	45,72	49,59
30	14,95	16,79	18,49	40,26	43,77	46,98	50,89

Table de Fisher

Les valeurs de cette table représentent les x tels que

$$F_{(\nu_1,\nu_2)}(x) = 0{,}95,$$

où $F_{(\nu_1,\nu_2)}$ est la fonction de répartition de la loi de Fisher de densité $f_{(\nu_1,\nu_2)}$ (voir Paragraphe 2.3.7).

ν_2	ν_1									
	1	2	3	4	5	6	7	8	9	10
1	161	199	215	224	230	234	236	238	240	241
2	18,5	19,0	19,1	19,2	19,3	19,3	19,3	19,3	19,3	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

Table de Fisher (*Suite*) :

ν_2	ν_1								
	12	15	20	24	30	40	60	120	∞
1	243	245	248	249	250	251	252	253	254
2	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Loi normale (Table de Gauss)

Les valeurs x de cette table sont relatives à l'α pour lequel $\Pr(Z < \alpha) = x$, où Z est une variable aléatoire centrée réduite. Par exemple :

$$\Pr(Z < 1{,}25) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1{,}25} e^{-\frac{t^2}{2}} dt = \Phi(1{,}25) \simeq 0{,}894.$$

α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500	0,504	0,508	0,512	0,516	0,519	0,523	0,527	0,531	0,535
0,1	0,539	0,543	0,547	0,551	0,555	0,559	0,563	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,594	0,598	0,602	0,606	0,610	0,614
0,3	0,617	0,621	0,625	0,629	0,633	0,636	0,640	0,644	0,648	0,651
0,4	0,655	0,659	0,662	0,666	0,670	0,673	0,677	0,680	0,684	0,687
0,5	0,691	0,695	0,698	0,701	0,705	0,708	0,712	0,715	0,719	0,722
0,6	0,725	0,729	0,732	0,735	0,738	0,742	0,745	0,748	0,751	0,754
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,793	0,796	0,799	0,802	0,805	0,807	0,810	0,813
0,9	0,815	0,818	0,821	0,823	0,826	0,828	0,831	0,834	0,836	0,838
1,0	0,841	0,843	0,846	0,848	0,850	0,853	0,855	0,857	0,859	0,862
1,1	0,864	0,866	0,868	0,870	0,872	0,874	0,877	0,879	0,881	0,883
1,2	0,884	0,886	0,888	0,890	0,892	0,894	0,896	0,898	0,899	0,901
1,3	0,903	0,904	0,906	0,908	0,909	0,911	0,913	0,914	0,916	0,917
1,4	0,919	0,920	0,922	0,923	0,925	0,926	0,927	0,929	0,930	0,931
1,5	0,933	0,934	0,935	0,937	0,938	0,939	0,940	0,941	0,942	0,944
1,6	0,945	0,946	0,947	0,948	0,949	0,950	0,951	0,952	0,953	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,959	0,960	0,961	0,962	0,963
1,8	0,964	0,964	0,965	0,966	0,967	0,967	0,968	0,969	0,969	0,970
1,9	0,971	0,971	0,972	0,973	0,973	0,974	0,975	0,975	0,976	0,976
2,0	0,977	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981
2,1	0,982	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985
2,2	0,986	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989
2,3	0,989	0,989	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991
2,4	0,991	0,992	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993
2,5	0,993	0,994	0,994	0,994	0,994	0,994	0,994	0,994	0,995	0,995
2,6	0,995	0,995	0,995	0,995	0,995	0,996	0,996	0,996	0,996	0,996
2,7	0,996	0,996	0,996	0,996	0,996	0,997	0,997	0,997	0,997	0,997
2,8	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,998	0,998
2,9	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
3,0	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,999	0,999

Table de Student

Cette table représente des valeurs x telles que :

$$\alpha = \int_{-\infty}^{x} \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)} \left(1+\frac{t^2}{\upsilon}\right)^{\left(-\frac{\upsilon+1}{2}\right)} dt.$$

ν	α					
	0,900	0,950	0,975	0,990	0,995	0,999
1	3,078	6,314	12,706	31,821	63,656	318,289
2	1,886	2,920	4,303	6,965	9,925	22,328
3	1,638	2,353	3,182	4,541	5,841	10,214
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,894
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,106	4,025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,492	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
40	1,303	1,684	2,021	2,423	2,704	3,307
50	1,299	1,676	2,009	2,403	2,678	3,261
60	1,296	1,671	2,000	2,390	2,660	3,232
70	1,294	1,667	1,994	2,381	2,648	3,211
80	1,292	1,664	1,990	2,374	2,639	3,195
90	1,291	1,662	1,987	2,368	2,632	3,183
100	1,290	1,660	1,984	2,364	2,626	3,174
∞	1,282	1,645	1,960	2,326	2,576	3,090

Table de nombres aléatoires

Les nombres présents sur cette table sont générés à partir des décimales de $\pi = 3{,}141\ 592\ 653\ 589\ 793\ldots$.

1 415 926 535	3 305 727 036	5 024 459 455	8 583 616 035	8 164 706 001
8 979 323 846	5 759 591 953	3 469 083 026	6 370 766 010	6 145 249 192
2 643 383 279	0 921 861 173	4 252 230 825	4 710 181 942	1 732 172 147
5 028 841 971	8 193 261 179	3 344 685 035	9 555 961 989	7 235 014 144
6 939 937 510	3 105 118 548	2 619 311 881	4 676 783 744	1 973 568 548
5 820 974 944	0 744 623 799	7 101 000 313	9 448 255 379	1 613 611 573
5 923 078 164	6 274 956 735	7 838 752 886	7 747 268 471	5 255 213 347
0 628 620 899	1 885 752 724	5 875 332 083	0 404 753 464	5 741 849 468
8 628 034 825	8 912 279 381	8 142 061 717	6 208 046 684	4 385 233 239
3 421 170 679	8 301 194 912	7 669 147 303	2 590 694 912	0 739 414 333
8 214 808 651	9 833 673 362	5 982 534 904	0 331 367 702	4 547 762 416
3 282 306 647	4 406 566 430	2 875 546 873	8 989 152 104	8 625 189 835
0 938 446 095	8 602 139 494	1 159 562 863	7 521 620 569	6 948 556 209
5 058 223 172	6 395 224 737	8 823 537 875	6 602 405 803	9 219 222 184
5 359 408 128	1 907 021 798	9 375 195 778	8 150 193 511	2 725 502 542
4 811 174 502	6 094 370 277	1 857 780 532	2 533 824 300	5 688 767 179
8 410 270 193	0 539 217 176	1 712 268 066	3 558 764 024	0 494 601 653
8 521 105 559	2 931 767 523	1 300 192 787	7 496 473 263	4 668 049 886
6 446 229 489	8 467 481 846	6 611 195 909	9 141 992 726	2 723 279 178
5 493 038 196	7 669 405 132	2 164 201 989	0 426 992 279	6 085 784 383
4 428 810 975	0 005 681 271	3 809 525 720	6 782 354 781	8 279 679 766
6 659 334 461	4 526 356 082	1 065 485 863	6 360 093 417	8 145 410 095
2 847 564 823	7 785 771 342	2 788 659 361	2 164 121 992	3 883 786 360
3 786 783 165	7 577 896 091	5 338 182 796	4 586 315 030	9 506 800 642
2 712 019 091	7 363 717 872	8 230 301 952	2 861 829 745	2 512 520 511
4 564 856 692	1 468 440 901	0 353 018 529	5 570 674 983	7 392 984 896
3 460 348 610	2 249 534 301	6 899 577 362	8 505 494 588	0 841 284 886
4 543 266 482	4 654 958 537	2 599 413 891	5 869 269 956	2 694 560 424
1 339 360 726	1 050 792 279	2 497 217 752	9 092 721 079	1 965 285 022
0 249 141 273	6 892 589 235	8 347 913 151	7 509 302 955	2 106 611 863
7 245 870 066	4 201 995 611	5 574 857 242	3 211 653 449	0 674 427 862
0 631 558 817	2 129 021 960	4 541 506 959	8 720 275 596	2 039 194 945
4 881 520 920	8 640 344 181	5 082 953 311	0 236 480 665	0 471 237 137
9 628 292 540	5 981 362 977	6 861 727 855	4 991 198 818	8 696 095 636
9 171 536 436	4 771 309 960	8 890 750 983	3 479 775 356	4 371 917 287
7 892 590 360	5 187 072 113	8 175 463 746	6 369 807 426	4 677 646 575
0 113 305 305	4 999 999 837	4 939 319 255	5 425 278 625	7 396 241 389
4 882 046 652	2 978 049 951	0 604 009 277	5 181 841 757	0 865 832 645
1 384 146 951	0 597 317 328	0 167 113 900	4 672 890 977	9 958 133 904
9 415 116 094	1 609 631 859	9 848 824 012	7 727 938 000	7 802 759 009

Références

[1] Anderson, T.W., Darling, D.A. (1952), Asymptotic theory of certain goodness of fit criteria based on stochastic processes, *Annals of Mathematical Statistics*, 23 : 193-212

[2] Anderson, T.W., Darling, D.A. (1954), A test of goodness of fit, *Journal of the American Statistical Association*, 49 : 765-9

[3] Beckmann, P. (1971), *A History of* π, Golem Press, Boulder CO

[4] Borel, E. (1909), Les probabilités dénombrables et leurs applications arithmétiques, *Rendiconti del Circolo Matematico di Palermo*, 27 : 247-71

[5] Box, G.E.P., Muller, M.E. (1958), A note on the generation of random normal deviates, *Annals of Mathematical Statistics*, 29 : 610-1

[6] Bray, T.A., Marsaglia, G. (1964), A convenient method for generating normal variables, *SIAM Review*, 6 : 260-4

[7] Cantoni, E., Huber Ph., Ronchetti, E. (2006), *Maîtriser l'aléatoire : Exercices résolus de probabilités et statistique*, Springer, Paris

[8] Davison, A.C., Hinkley, D.V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press

[9] Dodge Y. (1996), A natural random number generator, *International Statistical Review*, 64 : 329-44

[10] Dodge, Y. (1999), *Premiers pas en statistique*, Springer, Paris

[11] Dodge, Y., Melfi, G. (2005), Random number generators and rare events in the continued fraction of π, *Journal of Statistical Computation and Simulation*, 75 : 189-97

[12] Durbin, J., Knott, M., Taylor, C.C. (1975), Components of Cramer-Von Mises statistics, II, *Journal of the Royal Statistical Society, Series B,* 37 : 216-37

[13] Efron, B. (1979), Bootstrap methods : another look at the jackknife, *Annals of Statistics*, 7 : 1-26

[14] Efron, B., Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York

[15] Eichenauer J., Lehn, J. (1986), A non-linear conguential pseudo random number generator, *Statististical Papers*, 27 : 315-26

[16] Fishman, G.S. (1978), *Principles of Discrete Event Simulation*, John Wiley & Sons, New York

[17] Fisz, M. (1980), *Probability Theory and Mathematical Statistics*, Huntington, Robert E. Krieger Publishing Company, New York

[18] Gardner, M. (1966), The transcendental number pi, dans *New Mathematical Diversions from Scientific American*, Martin Gardner, Simon and Schuster, New York : 91-102

[19] Gentle, J. (2003), *Random Number Generation and Monte Carlo Methods*, Second Edition, Springer, New York

[20] Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer, New York

[21] Hardy G. H., Wright E. M. (1979), *An Introduction to the Theory of Numbers*, Clarendon Press, Oxford

[22] Hastings, W.K. (1970), Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, 57 : 97-109.

[23] Hoaglin, D. (1976), *Theoretical Properties of Congruential Random Number Generators : an Empirical View*, Memorandum NS-340, Harvard University, Department of Statistics

[24] Hogg, R.V., Craig, A.T., McKean, J. (2004), *Introduction to Mathematical Statistics*, Prentice Hall, Englewood Cliffs NJ

[25] Hull, W.E., Dobell, A.R. (1962), Random number generators, *SIAM Review*, 4 : 230-54

[26] Kaplan, W. (1992), *Advanced Calculus*, Addison-Wesley, Reading MS

[27] Khinchin, A. (1964), *Continued Fractions*, University of Chicago Press

[28] Knuth, D. (1981), *The Art of Computer Programming, Seminumerical Algorithms*, Addison-Wesley, Reading MS

[29] L'Écuyer, P. (1990), Random numbers for simulation, *Communications of the ACM*, 33 (10) : 85-97

[30] L'Écuyer, P. (2006), Random number generation, Chapitre 3 de *Elsevier Handbooks in Operations Research and Management Science : Simulation*, S.G. Henderson, B.L. Nelson, eds., Elsevier Science, Amsterdam : 55-81

[31] Lejeune, M. (2004), *Statistique : la théorie et ses applications*, Springer, Paris

[32] Lindemann, F. (1882), Über die Zahl π, *Mathematische Annalen*, 20 : 213-25

[33] Marsaglia, G. (1968), Random numbers fall mainly in the planes, *Proceedings of the National Academy of Sciences*, 61 : 25-8

[34] Matsumoto, M., Nishimura, T. (1998), Mersenne twister : A 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Transactions on Modeling and Computer Simulation*, 8 : 3-30

[35] Melfi, G., Schoier, G. (2004), Simulation of random distributions on surfaces, dans *Atti della XLII Riunione della Società Italiana di Statistica*, CLEUP, Padova : 173-6

[36] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., et al. (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, 21 : 1087-92

[37] Murier, T., Rousson, V. (1998), On the randomness of decimals of π, *Student*, 2 : 237-46

[38] Neumann, J., von (1951), Various techniques used in connection with random digits, dans *Monte Carlo Method*, A.S. Householder, G.E. Forsythe, H.H. Germond, eds., National Bureau of Standards Applied Mathematics Series, 12, U.S. Government Printing Office, Washington DC : 36-8

[39] Nishimura, T. (2000), Tables of 64-bit Mersenne twisters, *ACM Transactions on Modeling and Computer Simulation*, 10 : 348-57

[40] Park, S.T., Miller, K.W. (1988), Random number generators : good ones are hard to find, *Communications of the ACM*, 31 (10) : 1192-201

[41] Payne, W.H., Rabung, J.R., Bogyo T.P. (1969), Coding the Lehmer pseudo-random number generator, *Communications of the ACM*, 12 (2) : 85-6

[42] RAND Corporation (1955), *A Million Random Digits with* 100,000 *Normal Deviates*, Free Press, Glencoe IL

[43] Robert, C., Casella, G. (2004), *Monte Carlo Statistical Methods*, Series : Springer Texts in Statistics, 2nd edition, New York

[44] Ross, S.M. (1996), *Initiation au probabilités*, Presses polytechniques et universitaires romandes, Lausanne

[45] Särndal, C.E., Swensson, B., Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York

[46] Sowey, E.R. (1986), A third classified bibliography on random number generation and testing, *Journal of the Royal Statistical Society, Series A*, 149 : 83-107

[47] Stephens, M.A. (1974), EDF statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association*, 69 : 730-7

[48] Tippett, L.H.C., Pearson, K. (1927), Random sampling numbers : *Tracts for computers*, 15, Cambridge University Press

Index

Collection
Statistique
et probabilités
appliquées

Dirigée par
Yadolah Dodge

Yadolah Dodge, Giuseppe Melfi

Premiers pas en simulation

Cette collection met à la disposition du public intéressé par la statistique (étudiants, enseignants, chercheurs) des ouvrages qui concilient effort pédagogique et travail permanent de mise à jour. Cette démarche implique de prendre en compte de façon sélective et critique les renouvellements des concepts, des champs d'application et des outils de traitement. Seules une compréhension profonde et une appropriation des connaissances permettront de s'adapter aux évolutions qui n'ont pas fini de bouleverser cette discipline.

ISBN :

Ce livre est une introduction aux techniques de simulation. Cette méthode numérique permet de reconstituer fictivement l'évolution d'un phénomène et offre un complément aux méthodes analytiques parfois incapables de traiter des problèmes aux multiples variables. La simulation trouve de nombreuses applications dans l'industrie, en économie, en sciences sociales, en physique des particules, en astronomie.

Après un bref rappel des techniques fondamentales du calcul des probabilités, le livre expose diverses méthodes pour générer en grande quantités des nombres aléatoires. Ceux-ci sont un élément central de toute simulation. Les transformations de variables utilisées pour simuler des échantillons fictifs d'une variable aléatoire, et les tests d'hypothèses, qui permettent de valider un modèle pour des simulations à plus long terme, font l'objet des chapitres suivants. Enfin, la dernière partie porte sur la méthode de Monte Carlo et ses applications.

Tout au long de l'ouvrage, le lecteur est guidé par de nombreux exemples qui illustrent des applications très concrètes de ces diverses méthodes. En fin de chapitre, des exercices permettent à chacun de développer son savoir-faire.

Écrit dans un langage simple, ce livre s'adresse à un public de non-spécialistes : non seulement les mathématiciens n'ayant pas de formation approfondie en statistique, mais aussi les ingénieurs et les informaticiens confrontés quotidiennement à l'analyse de phénomènes complexes trouveront des indications utiles pour leur travail.

› springer.com