Introduction aux probabilités et à la statistique

Jean Bérard

Avertissement

Ces notes sont en cours d'élaboration. Il se peut donc qu'y subsistent un certain nombre d'erreurs, d'incohérences, et/ou de passages inachevés.

Table des matières

Introduction

La théorie des probabilités constitue un cadre mathématique pour la description du hasard et de la variabilité, ainsi que pour le raisonnement en univers incertain. Elle forme un tout cohérent dont les concepts, les méthodes et les résultats interviennent dans de très nombreux domaines des sciences et des technologies, parfois de manière fondamentale. En voici, à titre de motivation pour ce cours, une petite liste non-exhaustive.

En physique, la description de la nature à l'échelle microscopique, donnée par la mécanique quantique, est de nature probabiliste : seule la probabilité pour une particule de se trouver dans tel ou tel état est accessible à la théorie. En physique encore, la description des systèmes constitués d'un très grand nombre de particules (ce qui est le cas de tous les systèmes physiques macroscopiques) s'appuie généralement sur une modélisation probabiliste du comportement individuel des particules (mécanique statistique). En biologie, dans le domaine médical ou environnemental, la prise en compte de la variabilité naturelle des phénomènes étudiés nécessite souvent, et à toute sorte de niveaux, le recours à la modélisation probabiliste (il peut aussi bien s'agir d'étudier des mécanismes moléculaires comme la réplication de l'ADN, le développement morphologique d'un organisme, sa réponse à un traitement médical, ou encore la propagation des épidémies ou des feux de forêt, la croissance et les migrations de populations animales, la diffusion de polluants dans un sol, les phénomènes de crue, etc...). La modélisation probabiliste s'applique aussi au traitement des données et des signaux (codage, compression, débruitage), ou à l'analyse des erreurs de mesure. Elle intervient également dans le domaine économique et industriel (fiabilité et performance des systèmes et des procédés, dont le comportement comme l'environnement de fonctionnement sont variables, gestion des approvisionnements et des stocks, politiques d'assurance, prévisions économiques, décisions d'investissement, et plus généralement évaluation et gestion du risque). L'intelligence artificielle, et notamment les techniques d'apprentissage automatisé et d'extraction de données (reconnaissance de formes, traitement d'image, systèmes experts, fouille de données, réseaux neuronaux...) reposent également, pour une part sur une modélisation probabiliste de l'information qu'ils traitent. Mentionnons enfin l'utilisation devenue in-

contournable du «hasard simulé» par ordinateur, qu'il s'agisse d'étudier *in silico* le comportement d'un système réel que l'on a modélisé, d'employer un algorithme randomisé (d'optimisation, de tri, de vérification,...), ou de résoudre un problème numérique à l'aide d'une méthode de Monte-Carlo.

Un point de vocabulaire

Bien que les frontières délimitant les deux domaines ne puissent pas toujours être très précisément tracées, on distingue en général la **théorie des probabilités** et la **statistique**, en disant que la première a pour objet principal de définir des modèles mathématiques du hasard et de l'incertitude, et d'étudier leurs propriétés, tandis que la seconde a notamment pour but de confronter ces modèles mathématiques à la réalité, en particulier à l'expérience et aux données observées, afin de choisir, d'ajuster et de valider les modèles, et de les exploiter pour effectuer des prévisions, tester des hypothèses, prendre des décisions.

Objectifs du cours

Tous les exemples cités ci-dessus sont d'un niveau assez (voire très) élevé, et se rattachent à des domaines scientifiques spécialisés qu'il est bien entendu impossible d'aborder ou même de résumer dans un cours de base comme celui-ci. L'objectif principal de ce cours, qui requiert idéalement une première familiarisation, à un niveau intuitif avec les notions probabilistes, est de vous fournir des bases solides et correctement formalisées en probabilités. Il s'agira essentiellement d'assimiler les principaux outils conceptuels permettant d'aborder la modélisation mathématique de l'incertitude, du hasard et de la variabilité, ainsi qu'un certain nombre de techniques qui s'y rapportent. Après ce cours, vous devriez être en mesure de comprendre comment s'articulent les différents aspects (formalisation, intégration des données, résolution mathématique et/ou simulation, validation, exploitation, appréciation des limites de validité) de la modélisation de situations simples. Quelques objectifs plus spécifiques :

- dépasser le stade des raisonnements approximatifs et parfois douteux auxquels les étudiants sont bien souvent habitués quand il s'agit de probabilités ;
- aller au-delà des conclusions parfois insuffisantes ou même incohérentes que le simple «bon sens» permet de tirer ;
- être à l'aise vis-à-vis de l'utilisation des probabilités dans des domaines plus spécialisés, lorsque vous les rencontrerez.

Fournir des bases, notamment destinées à permettre un approfondissement et une spécialisation ultérieurs n'exclut pas, bien entendu, de présenter des exemples simples illustrant les applications potentielles dans quelques-uns des domaines plus avancés évoqués précédemment. D'autre part, posséder une connaissance correcte

des notions abordées dans ce cours présente également un intérêt du point de vue de la la formation des citoyens, à l'heure où les arguments fondés sur des modèles et des statistiques de toute nature (économique, sociale, médicale, environnementale,...) sont au cœur des débats, bien que trop peu d'individus possédent un bagage conceptuel suffisant pour soumettre ces arguments à une analyse critique informée et raisonnée.

Le niveau mathématique assez modeste dont nous nous contenterons ne doit pas masquer la véritable difficulté – celle sur laquelle l'effort doit porter principalement – que représente la compréhension en profondeur des notions abordées. Ce cours est entre autres un cours de mathématiques, où s'imposent donc des normes élevées de précision et de rigueur, mais les objets mathématiques qui y sont manipulés sont destinés à modéliser certains aspects de la réalité. Ainsi, toutes les notions abordées présentent un double aspect, formel et concret, ce qui rend leur maîtrise difficile à acquérir.

De nombreux exemples serviront à illustrer le propos, mais il est indispensable de dépasser le stade de la simple compréhension des exemples pour pouvoir utiliser efficacement les notions abordées dans des situations nouvelles.

Dés, cartes, et pièces de monnaie

Les cours de probabilités auxquels vous avez pu être confrontés font souvent la part belle aux exemples issus des jeux de hasard, tirages de carte, roulette, loteries et autres jeux de pile ou face. Quoique l'étude des jeux de hasard ait été l'une des motivations initiales du développement de la théorie des probabilités (principalement à partir du dix-septième siècle), il ne s'agit plus guère aujourd'hui que d'un domaine d'application anecdotique. Les exemples qui sont présentés dans ce cadre ne présentent que peu d'intérêt en tant qu'applications réelles, mais ils permettent facilement d'illustrer des notions ayant une portée beaucoup plus vaste, et peuvent donc servir de représentations conceptuelles simples à des situations réelles complexes. C'est dans cet état d'esprit qu'il est souhaitable d'aborder l'étude de ces exemples, ainsi que des exercices dans lesquelles des hypothèses très simplificatrices sont posées.

Comment travailler ce cours

Le volume de ce document vous affole peut-être... Pas de panique ! Ces notes forment en effet un ensemble d'une longueur certaine, mais le style est généralement peu dense, et une lecture à un rythme soutenu est (en principe) possible. Les définitions et résultats importants sont généralement mis en caractères gras. Des astérisques signalent les parties plus spécialisées et dont la lecture peut être omise sans compromettre sérieusement la compréhension de l'ensemble.

Ces notes sont – en principe – destinées à être lues au moins une fois *dans leur plus grande partie* ; elles servent de référence vis-à-vis du cours magistral, et apportent de nombreux détails et approfondissements par rapport à ce qui est présenté lors des séances de cours. À la fin de chaque chapitre, avant les exercices, se trouvent des questions d'auto-évaluation auxquelles vous devez impérativement savoir répondre, car elles portent sur les notions fondamentales du cours. Si la réponse à l'une de ces questions vous échappe, il est indispensable de relire et de retravailler le chapitre correspondant.

Quant aux nombreux exercices, dont la difficulté est très variable, il est indispensable, pour en tirer profit, d'en chercher d'abord la solution de manière autonome. Une partie importante d'entre eux est destinée à être traitée lors des séances de travaux dirigés. Des commentaires sur les exercices sont également proposés. Rappelons à toutes fins utiles que la solution d'un exercice doit être relue en grand détail de façon à vous assurer que vous en maitrisez toutes les étapes, et que vous en avez assimilé les idées globales. Seul ce travail de fond pourra vous assurer tant l'acquisition durable de connaissances et de méthodes que le succès à l'examen !

Il est important de ne pas vous laisser abuser par le cadre, parfois artificiel ou trivial en apparence, dans lequel certains exercices sont proposés ; il s'agit le plus souvent d'illustrer une question réellement importante, tout en essayant de ne pas vous noyer sous la complexité qu'appelle inévitablement la modélisation de situations plus réalistes.

Par ailleurs, un certain nombre de questions posées ont un caractère ouvert : on ne vous demande pas simplement de prouver tel ou tel résultat, mais de donner un sens précis à une question formulée de manière un peu vague, et de tenter d'y répondre à l'aide d'un modèle que vous aurez vous-même élaboré et justifié. Le but de ces questions n'est pas de vous décontenancer (encore que...) : tout en restant dans un cadre assez simple, elles font bien davantage appel aux capacités d'initiative, d'autonomie et d'esprit critique dont vous aurez à faire preuve dans votre vie professionnelle, et que votre formation est censée vous permettre de développer, que ne le font les questions de type plus traditionnel, et auxquelles vous pouvez être davantage habitués. Elles sont l'occasion de mettre à l'épreuve votre capacité à utiliser vos connaissances, et vous guident également vers une compréhension approfondie des notions et des méthodes abordées.

La manière d'exposer les différentes notions et résultats retenue dans ce cours repose, inévitablement, sur un certain nombre de partis pris pédagogiques. Des variations, légères ou plus significatives, par rapport à d'autres cours ou à des ouvrages cités dans la bibliographie, peuvent donc apparaître, tout-à-fait normalement (le souci de simplicité nous ayant en particulier conduit à ne pas traiter dans toute leur généralité un certain nombre de notions, et à insister sur certains modes de présentation au détriment d'autres, plus classiques). La cohérence avec la plupart des autres

exposés du même sujet est cependant assurée, moyennant éventuellement un petit effort (toujours fructueux) d'adaptation.

Les chapitres 1 et 2 présentent les bases du formalisme de la théorie des probabilités et de sa mise en œuvre pratique, et introduisent l'essentiel des notions utilisées dans la suite. Les chapitres 3 et 4 présentent les deux grandes «lois du hasard» que sont la loi des grands nombres et le théorème de la limite centrale.

Chapitre 1

Le modèle probabiliste

1.1 Introduction

La vie quotidienne, comme la pratique des sciences et des techniques, abondent en situations présentant plusieurs alternatives entre lesquelles il n'est pas possible de trancher a priori avec certitude, que cette incertitude soit attribuée au hasard ou à la chance, au manque d'informations ou de moyens de prévision, ou encore à une variabilité inhérente à la situation considérée. Se borner à constater une telle incapacité à connaître ou prévoir avec certitude ne mène pas très loin, et, fort heureusement, un vaste ensemble de situations peuvent être efficacement décrites à l'aide d'objets mathématiques appelés **modèles probabilistes**, qui permettent de **raisonner de manière cohérente, rigoureuse, et quantitative sur le hasard, la variabilité et l'incertitude**. Le but principal de ce cours est de vous apprendre à construire, manipuler et exploiter ces objets dans des situations simples. Nous aurons ainsi à accomplir plusieurs tâches distinctes :

1. présenter le formalisme mathématique des modèles probabilistes (ou, comme on disait autrefois, du calcul des probabilités), avec les définitions, règles et propriétés importantes qui s'y rattachent ;
2. expliquer le lien entre ce formalisme abstrait et la réalité modélisée ;
3. expliquer comment construire des modèles probabilistes satisfaisants d'une situation donnée ;
4. expliquer comment exploiter les modèles probabilistes une fois ceux-ci construits.

Concernant le point 1, nous procéderons par étapes, afin de ne pas vous noyer sous les définitions. Nous définirons dans ce chapitre le cadre mathématique général des modèles probabilistes (espace des possibles, événements, probabilités), puis les notions fondamentales de probabilité conditionnelle et de dépendance probabiliste. La notion de variable aléatoire, sera abordée dans le chapitre 2, les chapitres 3 et 4

traitant de deux propriétés fondamentales des épreuves aléatoires répétées que sont la loi des grands nombres et le théorème de la limite centrale. Soulignons que le point 1 se situe entièrement dans le champ des mathématiques : on s'y occupe uniquement de définir un formalisme mathématique général pour la modélisation probabiliste, et de démontrer rigoureusement certaines propriétés possédées par les entités qui y interviennent.

Le point 2 se situe, quant à lui, hors du champ exclusif des mathématiques, puisqu'il touche à la réalité concrète : il s'agit de préciser la contrepartie concrète des notions abstraites introduites dans le point 1. La question sera abordée au fur et à mesure que les notions mathématiques abstraites nécessitant des explications seront introduites. Nous verrons que la traduction concrète de la notion de probabilité est bien plus délicate à définir que ce que pourrait laisser supposer le caractère courant de l'utilisation du mot «probabilité». Nous aurons également l'occasion de justifier (par opposition au fait de démontrer) par des arguments concrets la pertinence des règles abstraites du calcul des probabilités.

Le point 3 est probablement le plus difficile. Il pose le problème central de la modélisation : comment, à partir des connaissances et des données disponibles, construire un modèle approprié à la description d'une situation réelle ? Comment juger de la validité d'un modèle ? Il s'agit en général de questions difficiles et complexes, au cœur de la pratique scientifique, et qui n'admettent ni solution systématique ni recette miracle. Nous verrons cependant qu'une bonne compréhension des points 1 et 2, ainsi qu'un minimum de pratique, permettent d'aborder le problème avec un certain succès dans des cas simples.

Le point 4 est pertinent lorsque la complexité des modèles utilisés fait que leur exploitation ne se résume pas à un calcul élémentaire, ce qui ne sera que rarement le cas dans notre contexte. Nous le mentionnons surtout pour souligner la distinction existant entre le fait de construire un modèle d'une situation donnée, et le fait d'exploiter ce modèle. Bien entendu, la construction d'un modèle est souvent, pour partie, orientée par l'exploitation que l'on compte faire de celui-ci.

La séparation entre les points 1 à 4 peut paraître quelque peu artificielle, compte-tenu des nombreux liens qui les unissent. Nous pensons toutefois qu'il n'est pas inutile, afin de bien structurer vos connaissances, de garder systématiquement en tête cette distinction.

Avertissement terminologique

Nous ne chercherons pas, dans ce cours, à définir de manière systématique – si tant est que cela soit possible – les notions de hasard, d'aléa(toire), de variabilité, ou encore d'incertitude. Il nous arrivera souvent d'utiliser ces termes, qui ne sont pourtant pas synonymes, de manière interchangeable, comme des **raccourcis de langage**

commode qui qualifient simplement le fait que, d'une manière générale, le fait que les situations que l'on étudie peuvent se réaliser de plusieurs manières. D'autres fois en revanche, nous les utiliserons en prenant en compte les nuances existant entre eux. De manière très schématique (voir également la discussion sur la traduction concrète de la notion de probabilité dans ce chapitre), on qualifie généralement d'aléatoire ou de produite par le hasard une situation combinant imprévisibilité des situations individuelles, et régularités statistiques lorsque l'on considère des situations répétées un grand nombre de fois (archétype : le lancer d'une pièce de monnaie) ; le terme de variabilité insiste plutôt sur la pluralité des modalités ou des valeurs que peuvent prendre, d'une situation à l'autre, les caractéristiques auxquelles on s'intéresse (archétype : la taille au sein de la population), tandis que l'incertitude désigne, plus généralement, notre incapacité à connaître exactement (archétype : le résultat d'une rencontre sportive avant que celle-ci ait eu lieu). Notons que tous ces termes (et particulièrement celui de hasard) trouvent également d'autres emplois et significations, que nous ne chercherons pas à aborder au risque de nous perdre dans des discussions philosophiques qui ne sont certainement pas l'objet de ce cours !

1.2 Le point de vue formel

Compte tenu du caractère central de la notion de modèle probabiliste dans tout ce qui va suivre, il nous semble préférable d'en donner dès le début une définition exacte, précise et... formelle.

Si vous n'avez jamais rencontré ce formalisme auparavant, tout cela vous paraîtra probablement un peu abstrait. L'objet de ce chapitre (et plus globalement l'un des objectifs de ce cours) est d'expliquer la signification de ce formalisme, la façon dont on le met en œuvre dans les situations concrètes, ainsi que son utilité. Nous adopterons systématiquement ce mode de présentation, consistant à donner d'abord la définition mathématique des objets rencontrés (point 1), puis à étudier leur signification concrète (point 2).

Au sens formel, donc, un **modèle probabiliste** (aussi appelé espace probabilisé, ou encore espace de probabilité) est la donnée d'un couple $(\Omega, \mathbb{P})$ constitué :

- d'un ensemble Ω fini ou dénombrable [1], appelé **espace des possibles**, ou encore **univers**,
- d'une application $\mathbb{P} : \Omega \to [0, 1]$, appelée probabilité sur Ω, et qui vérifie la

1. Il est possible de donner une définition plus générale pouvant faire intervenir des ensembles infinis non-dénombrables. Quoique présentant un grand intérêt, cette généralisation fait appel à des notions mathématiques dont la difficulté dépasse le cadre de ce cours. Nous nous restreignons ici à ce que l'on appelle les modèles probabilistes discrets.

condition suivante, dite de **normalisation** :

$$\sum_{\omega\in\Omega} \mathbb{P}(\omega) = 1.$$

Les éléments ω de l'ensemble Ω sont appelés des **éventualités élémentaires**, et représentent les différentes alternatives possibles, ou encore les issues, de la situation étudiée. La valeur $\mathbb{P}(\omega)$ est appelée **probabilité de l'éventualité élémentaire** ω, ou encore **probabilité de** ω**.**

En termes imagés, la «réalisation du hasard» est représentée, dans un tel modèle, par le choix d'une unique éventualité élémentaire ω dans Ω, qui détermine l'alternative effectivement réalisée : parmi les différentes issues possibles, le hasard «choisit» d'en réaliser une et une seule, chaque issue étant affectée d'une certaine probabilité.

On appellera **événement** (au sens formel) tout sous-ensemble (ou encore toute partie) de Ω. On dira qu'un événement A au sens précédent est réalisé lorsque l'éventualité élémentaire ω correspondant à l'alternative effectivement réalisée est un élément de A, c'est-à-dire lorsque $\omega \in A$. La probabilité d'un événement A est définie par :

$$\mathbb{P}(A) = \sum_{\omega\in A} \mathbb{P}(\omega).$$

Notez bien qu'il s'agit, dans ce cadre abstrait, d'une définition, et non pas d'une propriété que l'on établirait à l'aide d'une démonstration mathématique. Elle permet d'étendre la fonction $\mathbb{P}(\cdot)$, initialement définie sur l'ensemble Ω, à l'ensemble des événements, c'est-à-dire l'ensemble des parties de Ω.

Remarque 1 *lorsque* Ω *est un ensemble fini, c'est-à-dire qu'il n'y a qu'un nombre fini d'éventualités élémentaires, la définition des sommes* $\sum_{\omega\in\Omega} \mathbb{P}(\omega)$ *ou* $\sum_{\omega\in A} \mathbb{P}(\omega)$ *apparaissant ci-dessus ne présente aucune difficulté. Lorsque* Ω *est un ensemble infini, nous supposerons toujours que* Ω *est dénombrable, c'est-à-dire où l'on peut faire la liste de ses éléments sous la forme :* $\Omega = \{u_n : n \in \mathbb{N}\}$*, et les sommes ci-dessus seront comprises comme des séries à termes positifs.*

1.3 Mais que représente exactement ce formalisme ?

1.3.1 Espace des possibles et choix du niveau de description

La caractéristique première des situations que nous cherchons à étudier étant qu'il en existe plusieurs issues possibles, la modélisation d'une telle situation passe évidemment par l'établissement d'une liste de ces différentes issues. Comme nous l'avons dit plus haut, les issues possibles sont représentées par les éléments de Ω, et l'ensemble Ω proprement dit représente donc la liste de toutes les issues possibles. **Etant donnée une telle liste, chaque réalisation possible de la situation**

étudiée doit donc pouvoir être rattachée à une et une seule issue figurant dans la liste. Ainsi, l'espace des possibles doit, d'une part, contenir suffisament d'éléments pour que chaque réalisation possible de la situation puisse se voir rattachée à l'un d'entre eux, et, d'autre part, au plus un élément de l'espace des possibles doit être susceptible de correspondre à une réalisation donnée.

Cette brève description ne suffit malheureusement pas à déterminer de manière unique l'espace des possibles pouvant être utilisé pour décrire une situation donnée. En effet, selon le degré de précision que l'on adopte dans la description de la situation, la notion d'«issue» peut varier du tout au tout, et, pour une même situation, il n'y a donc pas un seul, mais une multitude d'ensembles Ω susceptibles de la décrire, si bien qu'il serait plus correct de dire que l'ensemble Ω représente la liste des issues possibles **au niveau de description adopté.** Par exemple, pour décrire la descente d'une piste de ski par un skieur, on peut se contenter d'un ensemble Ω ne comportant que deux issues :

$$\Omega_1 = \{\text{«chute»}, \text{«pas de chute»}\},$$

selon que le skieur est tombé ou non. Cet ensemble décrit bien toutes les issues possibles, au sens où, lors d'une descente, le skieur peut soit tomber, soit ne pas tomber, sans autre alternative possible : la réalisation de l'expérience correspond au choix d'un et un seul élément de Ω.

Mais on peut également adopter une description plus précise, en prenant pour espace des possibles l'ensemble :

$$\Omega_2 = \{\text{«pas de chute»}, \text{«une chute»}, \text{«deux chutes»} , \text{«trois chutes»} , \ldots\},$$

qui fait correspondre à chaque nombre de chutes une issue différente. Cet ensemble décrit également toutes les issues possibles (un skieur peut soit faire zéro, soit une, soit deux, etc... chutes, sans autre alternative possible), mais avec un niveau de précision plus grand : un élément de Ω_2 comprend plus d'information sur le déroulement de la descente qu'un élément de Ω_1. On notera que l'ensemble Ω_2 contient des éléments qui ne correspondent pas à des issues effectivement réalisables, telles que, par exemple : «2^{150} chutes ». Cela n'est pas gênant, mais signifie simplement que certaines «issues» théoriquement présentes dans la liste que constitue Ω n'apparaîtront jamais. On peut ainsi sans dommage (et cela simplifie souvent la description de celui-ci) inclure dans l'espace des possibles davantage de possibilités qu'il n'en existe réellement. Celles-ci se verront simplement affectées d'une probabilité nulle ou totalement négligeable en pratique. En revanche, les éléments de Ω doivent absolument rendre compte (au niveau de description adopté) de toutes les possibilités réelles, sans en omettre aucune.

Pour décrire encore plus précisément la descente, on peut par exemple ajouter

des informations relatives au nombre de sauts de bosses :

$$\Omega_3 = \mathbb{N} \times \mathbb{N} = \{(i,j) \ : \ i \in \mathbb{N}, \ j \in \mathbb{N}\}.$$

Un élément de Ω_3 est ici un couple de deux nombres entiers, le premier indiquant le nombre de chutes, et le deuxiéme le nombre de bosses sautées par le skieur. Et l'on peut bien entendu continuer à l'infini en ajoutant des informations sur la vitesse de la descente, la forme de la trajectoire, la couleur de la tenue, le nombre de surfeurs croisés, le temps qu'il fait, etc..., en obtenant à chaque fois une description plus précise de la descente effectuée. Chacun des ensembles Ω que l'on obtient décrit les différentes issues du phénomène (la descente de la piste), mais avec un degré de précision et selon une grille de lecture qui lui est propre. Il y a donc une infinité de choix envisageables pour l'espace des possibles Ω, suivant la précision que l'on adopte dans la description du phénomène. Comme il est bien entendu impossible de tenir compte de **tous** les facteurs susceptibles de varier d'une réalisation du phénomène à une autre, il est nécessaire d'en sélectionner un certain nombre, qui figureront dans Ω, les autres n'étant pas pris en considération explicitement. En général, la détermination du niveau de description approprié pour une situation donnée est une question difficile, sur laquelle nous aurons l'occasion de revenir, et il n'existe pas de méthode systématique qu'il suffirait d'appliquer pour la traiter en toute généralité. Mentionnons simplement que le choix de Ω repose en général sur un compromis entre la nature des informations dont on peut disposer, les éléments qu'il semble pertinent de prendre en compte pour décrire la situation, la complexité du modèle obtenu, et l'usage que l'on compte en faire.

Soulignons dès maintenant qu'il s'agit là d'une problématique générale en sciences : la mécanique newtonienne décrit la réalité physique en termes de points matériels, de forces, de vitesses ; la biologie, elle, donne de la réalité qu'elle étudie une description en termes d'organismes, de cellules, d'interactions biochimiques (et non pas d'atomes ou d'interactions physiques fondamentales) ; l'économie de son côté, décrit des agents, qui produisent et échangent des biens et des services (et non pas de gigantesques assemblages de molécules biologiques) ; on décrit le fonctionnement des logiciels informatiques en termes d'instructions exécutées et de tâches accomplies, et pas (en général) en termes d'impulsions électriques dans les matériaux qui constituent le support physique de l'ordinateur... Chaque science donne de la réalité qu'elle étudie une description fondée sur une grille de lecture qui lui est propre,– prenant en compte un certain niveau de détails, en ignorant d'autres – et qui rend cette réalité intelligible. Préciser Ω est donc la premiére étape de la modélisation probabiliste d'une situation, puisque cet ensemble indique le niveau de détail choisi pour aborder l'étude de celle-ci.

On retient de cette discussion que l'espace des possibles Ω représente l'ensemble des issues possibles au niveau de description choisi : l'espace

des possibles n'est pas déterminé uniquement par le phénomène que l'on étudie, mais de manière essentielle par le choix que nous faisons du degré de finesse avec lequel le phénomène doit être décrit.

1.3.2 Sens concret – sens formel

De manière concrète, il est nécessaire, en plus de l'ensemble Ω, de fournir un **dictionnaire** permettant de déterminer la signification concréte de ses éléments dans le contexte étudié, car Ω apparaît souvent comme un codage du phénomène considéré, et non pas comme une description du phénomène lui-même. Selon le contexte, un même ensemble, par exemple l'ensemble $\mathbb{N} \times \mathbb{N}$ des couples d'entiers pourra représenter les coordonnées d'un point mobile sur une surface, les températures minimale et maximale au cours d'une saison, l'âge du capitaine et celui de sa femme, ou encore le nombre de chutes et le nombre de bosses sautées, comme dans l'exemple précédent. L'ensemble Ω est donc souvent un ensemble abstrait, dont la forme exacte des éléments n'a aucune importance, pourvu que la manière dont ceux-ci représentent des réalisations concrètes de la situation étudiée soit précisée. De la même façon, pour modéliser le résultat du lancer d'une pièce de monnaie, en ne tenant compte que du résultat final, on pourra aussi bien utiliser :

$$\Omega = \{\text{«pile»} \ , \ \text{«face»}\},$$

que

$$\Omega = \{\text{P} \ , \ \text{F}\}, \ \Omega = \{\text{bouc} \ , \ \text{chêvre}\}, \ \Omega = \{\text{campanule, myosotis}\},$$

du moment que la signification de chacun des éléments de Ω est précisée (mais cette précision est indispensable, sans quoi il est en général impossible de comprendre ce que représentent les éléments de Ω).

C'est dans ce contexte que la notion d'événement formel, défini comme partie de Ω, trouve sa signification. En français, un événement désigne «quelque chose» qui peut ou non se produire, en rapport avec la situation considérée. De manière générale, **à tout événement concret, défini en français, par sa relation au phénomène considéré, nous associerons le sous-ensemble de Ω (événement formel) constitué par les éventualités élémentaires décrivant les issues pour lesquelles cet événement est effectivement réalisé**. Ainsi, le «choix par le hasard» d'une éventualité élémentaire réalisant un événement (au sens formel) signifie que l'événement (au sens concret) correspondant est réalisé.

Pour reprendre l'exemple du skieur, $A =$«le skieur tombe au moins deux fois», et $B =$ «le skieur ne saute aucune bosse», constituent des événements (au sens concret du terme). Lorsque l'espace des possibles est l'ensemble $\Omega_3 = \mathbb{N} \times \mathbb{N}$, l'événement concret A est associé à l'événement formel (sous-ensemble de Ω_3, qu'avec un abus de

notation courant, on notera également A)

$$A = \{(i,j) \in \mathbb{N} \times \mathbb{N} \ : \ i \geq 2\},$$

et, de même,

$$B = \{(i,j) \in \mathbb{N} \times \mathbb{N} \ : \ j = 0\}.$$

Première remarque : selon l'espace des possibles choisi, la traduction formelle d'un événement concret varie. Ainsi, dans Ω_2, A est associé à l'événement formel

$$A = \{\text{« 2 chutes »}, \text{« 3 chutes »} \ , \ldots\},$$

qui ne correspond en aucun cas à l'événement formel de Ω_3 pourtant associé au même événement concret. (Lorsque l'espace des possibles peut varier, vous remarquerez qu'il n'est pas très raisonnable de noter de la même manière un événement concret, qui reste fixé, et sa traduction formelle, qui varie en fonction de l'espace des possibles. Attention !)

Deuxième remarque : la finesse avec laquelle l'ensemble Ω décrit les réalisations du phénomène doit être compatible avec la définition «en français» pour que celle-ci définisse effectivement un événement au sens formel du terme. Par exemple, l'événement B ci-dessus, qui a un sens (concret) parfaitement défini relativement à l'expérience, ne définit pas un événement au sens formel si l'on adopte l'ensemble Ω_2 pour décrire le phénomène, car la description par Ω_2 des réalisations de l'expérience ne contient aucune information relative au nombre de bosses. : cet événément n'a pas de sens dans la description donnée par Ω_2. C'est en ce sens que les éléments de l'espace des possibles choisi constituent des éventualités «élémentaires» : aucune information sur la manière dont le phénomène se réalise, plus fine que celle contenue dans les éléments de Ω, n'a de sens dans le cadre du modèle, et tout événement ayant un sens dans le cadre du modèle (c'est-à-dire tout sous-ensemble de l'espace des possibles) est constitué par un assemblage d'éventualités élémentaires. Celles-ci constituent donc, en quelque sorte, les «atomes» de la description du phénomène par le modèle. Bien entendu, il ne s'agit d'éventualités élémentaires que relativement au modèle choisi, et, par exemple, l'événement «le skieur chute», qui constitue une éventualité élémentaire dans la description par Ω_1, apparaît comme constitué de plusieurs éventualités élémentaires dans la description par Ω_2 ou Ω_3. En revanche, et ceci peut constituer un premier guide pour choisir l'espace des possibles, nous constatons que **l'ensemble Ω doit décrire le phénomène d'une façon suffisament fine pour que les événements (au sens concret) auxquels on s'intéresse aient un sens dans le cadre du modèle**.

De manière générale, on dira qu'un espace des possibles Ω_a est **plus fin** qu'un autre espace des possibles Ω_b décrivant la même situation lorsque, pour toute éventualité élémentaire de Ω_b, l'événement concret qui lui est associé possède une traduction formelle (au moyen d'une ou plusieurs éventualités élémentaires) dans Ω_b.

Notons que **les opérations logiques usuelles sur les événements concrets** (conjonction, disjonction, négation), **correspondent à des opérations ensemblistes** (intersection, union, complémentaire) **sur les événements formels** (sous-ensembles de Ω) **qui leur sont associés**.

Partant de deux événements A et B (on notera de la même façon les événements décrits en français et les sous-ensembles de Ω qui leur correspondent, ce petit abus de notation ne soulevant pas d'ambiguité lorsque Ω est fixé), on peut en particulier considérer :

- l'événement défini (en français) par «A ou B», qui correspond dans Ω à la réunion de A et B, notée $A \cup B$, et qui désigne l'ensemble des éventualités élémentaires qui réalisent A **ou** B (éventuellement les deux à la fois),
- l'événement défini (en français) par «A et B», qui correspond dans Ω à l' intersection de A et B, notée $A \cap B$, qui désigne l'ensemble des éventualités élémentaires qui réalisent A **et** B.
- l'événement défini (en français) par «A n'a pas lieu», qui correspond dans Ω au complémentaire de A, noté A^c ou $\bar{A}$, et qui désigne l'ensemble des éventualités élémentaires qui ne réalisent **pas** A.

Mise en garde 1 *par convention, le «ou» que nous utilisons est toujours inclusif, c'est-à-dire qu'il n'exclut pas la réalisation simultanée des deux événements. C'est le «ou» de la petite annonce : «secrétaire parlant allemand ou anglais» (éventuellement les deux à la fois). Lorsque nous considérerons le «ou» exclusif (celui du menu : «fromage ou dessert»), qui correspond à la réalisation de l'une ou l'autre des deux éventualités, mais pas des deux à la fois, nous le spécifierons en utilisant l'expression «ou bien».*

Deux événements A et B sont dits **incompatibles** s'ils ne peuvent se réaliser simultanément, ou, autrement dit, si aucune éventualité élémentaire ne peut réaliser à la fois A et B, ou encore, si $A \cap B = \emptyset$. On notera que deux éventualités élémentaires distinctes sont toujours incompatibles, ce qui correspond au fait que la «réalisation du hasard» correspond au choix d'une unique éventualité élémentaire parmi les éléments de Ω.

Par ailleurs, on dira qu'un événement A **implique**, ou **entraîne** un événement B, lorsque A est inclus dans B (notation $A \subset B$), autrement dit, lorsque toute éventualité élémentaire qui réalise A réalise également B (ainsi, on est certain que lorsque A est réalisé, B l'est également). Dans ce cas, la réalisation de A s'accompagne automatiquement de celle de B.

Exemples :

Commençons par un exemple très simple, qui peut, par exemple, servir pour modéliser le résultat de deux lancers successifs d'un dé à six faces.

$$\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\},$$

$$A = \{(1,2); (2,3); (5,4)\} \ , \ B = \{(2,3); (2,6)\}$$

$$A \cup B = \{(1,2); (2,3); (2,6); (5,4)\} \text{ et } A \cap B = \{(2,3)\}.$$

Revenons à l'exemple du skieur, avec $\Omega_3 = \mathbb{N} \times \mathbb{N}$, et définissons trois événements concrets : $A =$«le skieur saute moins de trois bosses», $B =$«le skieur tombe 4 fois ou plus» et $C =$«le skieur saute 5 ou 6 bosses». Avec Ω_3 pour espace des possibles, on a, au sens formel :

$$A = \{(i,j) \in \mathbb{N} \times \mathbb{N} \ : \ j \leq 3\},$$

$$B = \{(i,j) \in \mathbb{N} \times \mathbb{N} \ : \ i \geq 4\},$$

$$C = \{(i,j) \in \mathbb{N} \times \mathbb{N} \ : \ j \in \{5,6\}\}.$$

L'événement «A et B» signifie concrètement que le skieur saute au moins trois bosses et tombe 4 fois ou plus, et s'écrit formellement :

$$A \cap B = \{(i,j) \in \mathbb{N} \times \mathbb{N} \ : \ i \geq 4, \ j \geq 3\},$$

ou encore :

$$A \cap B = \{4, 5, \ldots\} \times \{3, 4, \ldots\}.$$

L'événement «B ou C» signifie que le skieur tombe 4 fois ou plus ou saute 5 ou 6 bosses (éventuellement les deux à la fois), et correspond à l'ensemble

$$\{(i,5) : \ i \geq 4\} \cup \{(i,6) : i \geq 4\}.$$

On note que A et C sont incompatibles , car, au sens concret , le skieur ne, peut bien entendu pas sauter à la fois moins de trois bosses et cinq ou six bosses, et, au sens formel, on observe bien que $A \cap C$ est l'ensemble vide.

Rappelons rapidement quelques propriétés élémentaires satisfaites par les opérations sur les ensembles. étant donnés trois sous-ensembles A, B et C d'un ensemble Ω, (ou encore trois événements d'un espace des possibles) les propriétés suivantes sont vérifiées :

- $A \cup B = B \cup A$ (commutativité de la réunion)
- $A \cap B = B \cap A$ (commutativité de l'intersection)
- $A \cup (B \cup C) = (A \cup B) \cup C$ (associativité de la réunion)
- $A \cap (B \cap C) = (A \cap B) \cap C$ (associativité de l'intersection)

- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ (distributivité de l'intersection par rapport à la réunion)
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (distributivité de la réunion par rapport à l'intersection)
- $(A \cup B)^c = A^c \cap B^c$
- $(A \cap B)^c = A^c \cup B^c$

On définit également deux événements particuliers : l'événement certain, qui est formellement associé à Ω tout entier. quelle que soit l'éventualité élémentaire «choisie par le hasard», celle-ci réalise toujours l'événement Ω (le hasard choisit toujours un élément de Ω) ; et l'événement impossible, associé à l'ensemble vide $\emptyset$. Comme aucune éventualité élémentaire n'appartient jamais à $\emptyset$, cet événement ne se produit jamais.

1.3.3 Signification concrète de la probabilité

La donnée de l'espace des possibles Ω ne suffit pas, à elle seule, à décrire de manière satisfaisante une situation incorporant de l'incertitude, puisqu'elle indique simplement un certain niveau de détail avec lequel on choisit de décrire les différentes issues de cette situation. La probabilité $\mathbb{P}$ sur Ω constitue le second ingrédient fondamental d'un modèle probabiliste, et contient les informations quantitatives susceptibles d'être exploitées concrètement. Du point de vue formel, une probabilité est simplement une application qui associe à chaque élément de l'espace des possibles un nombre compris entre 0 et 1, de telle façon que la somme des probabilités de toutes les éventualités élémentaires soit égale à 1. N'importe quelle application vérifiant cette propriété est une probabilité sur Ω, et il existe donc une infinité de probabilités différentes pour un même espace des possibles. Déterminer, parmi toutes ces probabilités, laquelle (ou lesquelles) sont susceptibles de décrire de manière satisfaisante une situation donnée est, avec la détermination de l'espace des possibles, le problème principal de la modélisation probabiliste (point 3). Avant de pouvoir aborder celui-ci, il nous faut d'abord nous interroger sur ce que représente concrètement la probabilité, dans le cadre de situations réelles (point 2).

Voici donc quelques exemples d'affirmations probabilistes (à replacer dans leur contexte). Avant de lire la discussion qui suit, nous vous invitons à réfléchir par vous-même à la signification concrète du terme «probabilité» dans chacun des cas.

1. «La probabilité pour que le candidat A soit élu lors de la prochaine élection présidentielle est de 60%.»
2. «La probabilité pour que la pièce de monnaie tombe sur face est de 50%.»
3. «La probabilité pour que l'équipe de football du Brésil l'emporte demain face à l'Allemagne est de 1/4.»
4. «La probabilité pour qu'il pleuve demain à Lyon est de 1/3.»

5. «La probabilité pour qu'il ait plu il y a exactement 3000 ans sur le site aujourd'hui occupé par Lyon est de 1/3.»
6. «La probabilité pour qu'une météorite de plus de 500m de diamètre de circonférence percute la terre au cours du prochain millénaire est de moins de 2%.»
7. «La probabilité pour que la fusée explose au décollage est de moins de 2%.»
8. «La probabilité, pour un individu né en France en 1920, de vivre plus de 80 ans est de 75%.»
9. «La probabilité pour un individu né en France en 1954, de vivre plus de 80 ans est de 85%.»
10. «La probabilité pour un individu né en France en 1954 de posséder un chien est de 60%.»
11. «La probabilité pour que D*** (qui est né en France en 1954) possède un chien est de 70%.»
12. «La probabilité pour qu'un atome de carbone 14 subisse une désintégration au cours des 5730 prochaines années est de 50%.»
13. «La probabilité pour qu'un photon incident émis par la source S soit absorbé par le détecteur D est de 1/3.»
14. «La probabilité pour que l'épidémie se propage est de 5%.»
15. «La probabilité pour qu'un paquet de données mette plus de 0,1 seconde pour être transmis dans le réseau est de 10%.»
16. «La probabilité pour que l'enfant à naître soit une petite fille est de 1/2.»
17. «La probabilité pour que la croissance du PIB soit cette année supérieure à 2%, est de 70%.»

La probabilité comme fréquence

Un premier lien, fondamental, entre la notion abstraite de probabilité et la réalité concrète, est l'interprétation de la probabilité d'un événement comme la fréquence avec laquelle cet événement se produit au cours de longues séries d'expériences. Dans l'exemple 13, lorsque l'on affirme que la probabilité pour un photon émis d'être absorbé est de 1/3, cela signifie simplement que l'on s'attend à ce que, systématiquement, sur un grand nombre de photons émis par la source S, la proportion de ceux qui sont absorbés par le détecteur soit de l'ordre de 1/3, et d'autant plus proche de cette valeur que le nombre de photons étudiés est grand. Cette attente se fonde notamment sur l'expérience passée, qui a pu par exemple établir que, chaque fois que l'on étudie l'absorption des photons issus d'une source de même type que S par un détecteur du même type que D, la proportion de photons absorbés est systématiquement de

l'ordre de 1/3, quand on prend en compte un grand nombre de photons successifs [2]. La probabilité apparaît ainsi comme une caractéristique physique objective des photons et du dispositif utilisé, susceptible d'être mesurée expérimentalement. Quoique le comportement individuel (absorption ou non) des photons paraisse imprévisible, on observe une régularité statistique à long terme dans les résultats des expériences. L'exemple 12 relève *a priori* du même type d'interprétation de la probabilité : sur un grand nombre d'atomes de C14, on s'attend toujours à ce qu'environ la moitié d'entre eux subissent une désintégration au cours des 5730 prochaines années. On notera cependant que, dans ce cas, ce n'est pas l'observation directe qui permet d'établir la valeur et le caractère reproductible de cette proportion, mais nécessairement un raisonnement s'appuyant sur un certain nombre de données et d'hypothèses relatives au phénomène considéré (en extrapolant le comportement observé du C14 sur des périodes brèves à son comportement sur plusieurs milliers d'années). Définir et évaluer la probabilité ne sont pas une seule et même chose !

L'exemple 2 paraît encore se rattacher à ce type de définition : au cours d'une longue série de lancers, on s'attend à ce que la pièce tombe sur face dans environ la moitié des cas, soit que l'on ait déjà mené des expériences de lancer avec cette pièce ayant permis d'observer ce comportement, soit que l'on raisonne sur la symétrie de la pièce, rien ne semblant *a priori* favoriser davantage une retombée sur pile qu'une retombée sur face. Ce dernier cas illustre, peut-être plus clairement que les deux précédents, un certain nombre de difficultés en rapport avec l'interprétation de la probabilité comme fréquence. Tout d'abord, on ne peut pas définir ainsi la valeur exacte d'une probabilité : d'une longue série d'expériences à l'autre, la fréquence de pile et de face va légèrement varier, laissant planer une certaine incertitude quant à la valeur exacte à attribuer à la probabilité, et ce n'est que dans l'idéalisation d'une série infinie d'expériences (et rien de tel n'existe concrètement) que l'on pourrait espérer déterminer une valeur unique pour celle-ci. Cette situation n'est en tout cas pas propre à la probabilité, et, de fait, la plupart des grandeurs physiques (la masse ou la longueur d'un objet, par exemple) ne sont pas véritablement définies à mieux qu'une certaine incertitude près. Un modèle abstrait n'entend de toute façon jamais décrire la réalité mieux qu'à une certaine approximation près et dans la limite d'un certain domaine de validité. Une difficulté plus sérieuse est de déterminer précisément comment sont répétées les expériences auxquelles on se réfère : en effet, si l'on répète des lancers dans des conditions exactement identiques, on obtiendra en principe des résultats exactement identiques, et non pas une alternance imprévisible de pile et de face. Par ailleurs, avec de l'entraînement et un peu d'habileté, il est possible d'effec-

2. On peut également imaginer que l'on s'attend à observer ce résultat simplement parce qu'il est une conséquence de la description que fait la mécanique quantique de l'expérience menée, la théorie quantique étant à l'heure actuelle acceptée comme une description correcte – et amplement vérifiée expérimentalement – de ce type de phénomènes.

tuer le lancer de manière à faire retomber la pièce du côté que l'on souhaite. Ainsi, le caractère stable de la fréquence au cours d'un grand nombre d'expériences répétées n'est en aucun cas automatique, et dépend crucialement de la manière dont les expériences sont effectuées. Dans le cas de lancers «honnêtes» [3] d'une pièce symétrique, c'est l'extrême sensibilité du résultat d'un lancer à de très faibles variations – inévitables et imprévisibles – des conditions dans lesquelles celui-ci est effectué, qui est à l'origine de cette propriété (et c'est l'interprétation que l'on peut donner au raisonnement *a priori* sur la symétrie de la pièce pour évaluer les probabilités). (Pour une étude approfondie des lancers répétés de pièces de monnaie, vous pouvez consulter les deux articles sur le sujet cités dans la bibliographie.) Formaliser précisément ce type d'idée, afin d'expliquer comment des systèmes entièrement déterministes peuvent produire des comportements en apparence aléatoires, mais présentant des régularités statistiques, est l'un des buts de la branche de la théorie des systèmes dynamiques appelée théorie ergodique. En pratique, il est difficile de s'assurer que les conditions dans lesquelles on effectue une expérience garantissent la stabilité des fréquences à long terme lorsque celle-ci est répétée dans des conditions comparables (il faudrait préciser exactement quelles conditions expérimentales sont fixées d'une répétition à l'autre, et s'assurer que la variation d'une expérience à l'autre des conditions expérimentales qui ne sont pas fixées a bien toujours pour effet de stabiliser les fréquences autour d'une même valeur), et l'on doit se contenter d'arguments et d'indications partiels allant dans ce sens, dont des vérifications expérimentales de la stabilité des fréquences sont l'un des éléments.

L'exemple 8 semble poser bien moins de problèmes : au sens courant, la probabilité représente simplement la proportion des individus nés en France en 1920 ayant survécu au moins jusqu'à la fin de l'année 2000, et l'examen des registres de l'état-civil doit permettre de déterminer cette proportion avec une précision satisfaisante : la probabilité est définie de manière objective, et peut être évaluée de manière non moins objective, sans hypothèses supplémentaires compliquées sur la nature des phénomènes mis en jeu. L'exemple 10 est totalement similaire.

L'exemple 9 est déjà moins évident : la proportion d'individus nés en France en 1954 et qui vivront au-delà de l'âge de 80 ans est, certes, une quantité définie objectivement, qui permet donc de donner un sens objectif à la probabilité dans ce contexte ; cependant, il n'est pas possible à l'heure actuelle (en 2005) de déterminer quelle sera en définitive la valeur de cette proportion. Par conséquent, comme dans le cas de l'exemple 12, nous ne pouvons en proposer que des estimations, en nous basant sur un raisonnement plus ou moins élaboré, incluant données (par exemple sur ce qui est connu à l'heure actuelle de l'état de santé de la population des individus

3. Les tirages au sort effectués par jet de pièce de monnaie lors de rencontres sportives sont parfois règlementés : on y impose par exemple une hauteur minimale à laquelle la pièce doit s'élever avant de retomber.

nés en 1954), et hypothèses diverses. Encore une fois, définir et évaluer sont deux choses bien distinctes.

Dans la mesure où la probabilité y est définie comme une proportion au sein d'une population, ces trois derniers exemples présentent une analogie formelle avec les trois étudiés plus haut, où la probabilité apparaît comme une fréquence au cours de séries d'expériences. Il y a plus : on interprétera souvent la probabilité, disons de l'exemple 8, comme la probabilité qu'une personne prise au hasard dans la liste des individus nés en France en 1920 ait vécu au moins jusqu'à la fin de l'année 2000. Dans ce cas, on fait référence, non plus seulement à une population d'individus au sein de laquelle on calcule une proportion, mais à une expérience de tirage. Au cours d'une longue série de tirages, on s'attend à ce que la proportion observée d'individus ayant vécu au moins jusqu'à la fin de l'année 2000 soit voisine de la proportion que représentent ces individus dans la population. En termes des exemples précédents, on suppose premièrement que le processus de tirage donne lieu à des fréquences stables lors de tirages répétés, et que de plus ces fréquences sont données par les proportions correspondantes dans la population. C'est le principe même du sondage. Comme précédemment, on notera qu'il est difficile de garantir absolument que ces deux propriétés ont bien lieu.

La probabilité comme mesure raisonnable de plausibilité

Les exemples 1 et 11 font référence à des situations uniques (élection présidentielle, match de football), qui ne peuvent être replacés de manière évidente dans un contexte d'expériences répétées. La probabilité y apparait comme une mesure de plausibilité attachée aux événements, ou encore comme un degré de confiance dans la réalisation de ceux-ci.

Plus la valeur que nous attribuons à la probabilité d'un événement est élevée (c'est-à-dire proche de 1), plus nous serions surpris de ne pas voir l'événement en question se réaliser ; inversement, plus cette valeur est faible (proche de 0), plus nous serions surpris de voir l'événement ne pas se réaliser ; enfin, une plausibilité de 50% signifie au contraire que nous sommes également indécis vis-à-vis de la réalisation de l'événement et de sa non-réalisation.

De manière générale, la définition et l'estimation de la probabilité se fondent dans ce type de situations sur ce que nous appellerons un raisonnement en situation d'incertitude, dont ce cours illustrera un certain nombre de principes généraux. Par exemple, dans le cas de l'élection présidentielle, on s'appuiera sur des considérations concernant l'économie, l'état de l'opinion, les relations internationales, les alliances électorales, etc... en prenant en compte des éléments plus ou moins objectifs. La question qui se pose alors est bien évidemment : comment intégrer de manière cohérente informations, hypothèses, voire opinions, dans un raisonnement, de façon à en dé-

duire une évaluation de la probabilité ? Il faut noter en tout cas que, dans ce type de situations, **la probabilité n'apparaît que comme le reflet du raisonnement et des hypothèses, informations et opinions, sur lesquelles celui-ci est basé**. Même une fois levée l'incertitude concernant l'issue de la situation (par exemple, après que l'élection a eu lieu), on ne dispose pas d'un moyen définitif de confirmer ou d'infirmer telle ou telle valeur initialement proposée de la probabilité (si le candidat A est élu, quelle était *a priori* la bonne estimation de probabilité initiale : 65%, 70%, 80% ?). Et, bien entendu, des raisonnements différents donnent lieu en général à des estimations différentes de la probabilité d'un même événement... La probabilité perd donc, dans ce contexte, le caractère objectif qu'elle possédait, en tant que fréquence, dans les exemples du paragraphe précédent ; dans ce genre de situations, on peut simplement tenter d'évaluer la pertinence des arguments employés pour estimer la probabilité, à la lueur des connaissances et des données disponibles. Notons que la simple exigence de cohérence dans le raisonnement impose, comme nous le verrons plus loin, un certain nombre de règles, qui font que l'on ne peut pas manipuler les plausibilités de manière totalement arbitraire. On peut ainsi s'attendre à ce que, dans une certaine mesure, des individus rationnels aboutissent à des estimations de probabilité comparables s'ils s'appuient sur des informations, hypothèses et opinions comparables.

Les relations entre la probabilité «fréquentielle» du paragraphe précédent et la probabilité «plausible» étudiée ici sont d'une importance fondamentale. Dans les situations étudiées dans les exemples 2, 12 et 13, et en l'absence d'informations supplémentaires, il est naturel d'interpréter la fréquence de long terme avec laquelle un événement se produit comme une mesure de sa plausibilité : on attribuera par exemple une plausibilité de $1/2$ au fait que la pièce retombe côté face lors du prochain lancer. En revanche, en présence d'informations – par exemple, de données cinématiques précises sur la pièce de monnaie quelques instants après le lancer – portant sur les conditions expérimentales non-spécifiées dans la définition de la fréquence, la prise en compte de ces informations peut conduire à une estimation différente de la probabilité, même entendue en un sens purement fréquentiel, comme nous le verrons plus loin. De manière générale, lorsque les informations dont nous disposons sur une situation unique nous permettent seulement de replacer celle-ci au sein d'une certaine collection (population, ou ensemble d'expériences répétées), sans pouvoir la situer plus précisément, il paraît raisonnable d'évaluer la plausibilité des événements relatifs à cette situation à partir des fréquences calculées au sein de cette collection, lorsque celles-ci sont accessibles.

Par exemple, on pourrait évaluer la plausibilité du fait que D*** possède un chien en déterminant la proportion de possesseurs de chiens parmi les individus nés en France en 1954. Si l'on ignorait l'année de naissance de D***, on pourrait évaluer cette plausibilité en comptant la proportion de possesseurs de chiens dans la popula-

tion totale. A l'inverse, si l'on savait qu'en plus d'être né en 1954, D*** vit en zone rurale, on choisirait de considérer la proportion de propriétaires de chiens parmi les individus nés en France en 1954 vivant en zone rurale. Notre degré d'information sur D*** détermine ainsi une collection d'individus, d'autant plus restreinte que ce degré d'information est élevé, et grâce à laquelle on peut tenter d'évaluer la plausibilité d'un événement relatif à D*** et à nos informations à son sujet, en mesurant la fréquence d'apparition de l'événement dans la collection.

La mise en œuvre de cette idée se heurte cependant à toutes sortes de difficultés. Très souvent, l'ensemble des informations dont on dispose sur une situation déterminent complètement celle-ci (par exemple, lorsque l'on connaît exactement l'identité de D***, sans pour autant savoir s'il ou elle possède un chien), et l'on ne peut donc inscrire de manière naturelle cette situation dans une collection plus vaste, sans négliger un certain nombre d'informations pourtant disponibles en ne conservant que celles qui semblent pertinentes. Un délicat problème de choix apparaît donc : comment replacer de manière pertinente une situation unique dans une collection plus vaste de situations à partir des informations disponibles ? Qui plus est, même en ne conservant que les informations qui semblent pertinentes vis-à-vis de la situation étudiée, on peut être conduit à des collections de situations pour lesquelles on ne dispose pas de données suffisantes relatives aux fréquences. Bien souvent, on devra faire appel simultanément à plusieurs collections, correspondant chacune à une partie des informations disponibles (par exemple, relative chacune à tel ou tel élément particulier de la situation considérée), pour tenter d'évaluer les plausibilités intervenant dans le raisonnement. Dans ce contexte, le recours à des hypothèses ou à des estimations subjectives peut s'avérer incontournable afin d'intégrer les différentes données disponibles et de parvenir à un résultat. Bien entendu, plus les informations et les données dont on dispose sont précises et nombreuses, plus on peut s'attendre à obtenir une estimation de plausibilité satisfaisante. Inversement, notre ignorance quant à une situation peut être telle qu'il s'avère impossible de proposer une estimation pertinente de la plausibilité d'un événement. Il faut alors reconnaître les limites de notre capacité à modéliser la situation. Eventuellement, des approches alternatives ou complémentaires à la modélisation probabiliste classique (telles que logique floue, fonctions de croyance, etc...) peuvent être envisagées.

La plupart du temps, le raisonnement probabiliste mêle entre eux les différents aspects (fréquence et plausibilité) de la notion de probabilité. Nous vous laissons, à titre d'exercice (Exercice 3), le soin de réfléchir à la signification de la probabilité dans les exemples dont nous n'avons pas traité.

Les règles abstraites du calcul des probabilités s'appliquent, quant à elles, indépendamment de la signification concrète des quantités manipulées, et nous tenterons dans la suite de justifier leur utilisation à partir des différents points de vue. Nous retiendrons notamment de la discussion précédente que la signification concrète de

la notion de probabilité peut varier considérablement d'une situation à l'autre, que celle-ci dépend très souvent des informations, hypothèses, ou opinions dont nous disposons sur la situation étudiée, et qu'il est à peu près dépourvu de sens de parler de LA probabilité de tel ou tel événement concret, sans préciser le contexte dans lequel celle-ci intervient et peut être évaluée. Lors de l'élaboration ou de l'exploitation d'un modèle probabiliste, il est donc indispensable de préciser systématiquement, face à une probabilité, le sens que possède celle-ci et le contexte dans lequelle elle peut être évaluée, car le sens et donc la validité des hypothèses que l'on peut formuler, ou des conclusions que l'on peut tirer, sont conditionnées par la signification des probabilités utilisées.

Bien comprendre le sens concret de la notion de probabilité n'est pas (qu') une tâche philosophique : c'est une étape indispensable (point 2) pour créer ou exploiter une modélisation probabiliste d'une situation donnée, et en saisir correctement la portée.

Notons que le problème de l'interprétation concrète de la notion de probabilité est, aujourd'hui encore, l'objet de controverses et de recherches, souvent, mais pas toujours, davantage de la part de philosophes que de mathématiciens ou de physiciens. Pour en apprendre davantage sur cette question (en particulier, sur son histoire, sur les descriptions précises des différentes interprétations possibles, ainsi que sur les interprétations alternatives que nous n'avons pas présentées, vous pouvez par exemple consulter l'ouvrage de Ian Hacking et Michel Dufour cité dans la bibliographie, ainsi que la bibliographie de cet ouvrage.) Enfin, une question différente, – fort intéressante, mais trop éloignée du sujet de ces notes pour que nous l'étudiions de manière systématique – est celle de la perception psychologique de la probabilité, et des nombreux biais qui affectent celles-ci. Notons simplement que, dans de nombreuses situations, cette perception a tendance à ne pas coïncider avec les estimations auxquelles conduit une évaluation rationnelle et scientifique de la probabilité. Quelques exemples de telles situations sont donnés dans les exercices. Pour un exposé systématique de ces questions, nous vous invitons à consulter les ouvrages de Kahneman et Tversky cités dans la bibliographie.

1.4 Probabilité et événements

1.4.1 Probabilité d'un événement

Nous avons, au début de ce chapitre, défini la probabilité d'un événement formel comme la somme des probabilités des éventualités élémentaires qui le constituent, seules les éventualités élémentaires, qui correspondent aux différentes issues,

se voyant attribuer une probabilité, d'où la formule :

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega).$$

Cette définition ne vous surprend vraisemblablement pas, si vous avez déjà un tant soit peu manipulé ce formalisme auparavant. Cependant, il importe de garder à l'esprit qu'il ne s'agit là que d'une définition, formulée dans le cadre abstrait des modèles probabilistes (au passage, notez bien que la probabilité d'un même événement varie lorsque la probabilité sur l'espace des possibles change : la probabilité d'un sous-ensemble A de Ω ne sera en général pas la même dans le modèle $(\Omega, \mathbb{P}_1)$ et dans le modèle $(\Omega, \mathbb{P}_2)$). A défaut de la démontrer (une fois encore, il s'agit d'une définition, il n'y a rien à démontrer), notons tout de même que l'on peut (et doit, dans la mesure où un modèle probabiliste n'est pas destiné à rester un objet mathématique abstrait, mais à modéliser des situations réelles) justifier la cohérence de cette définition abstraite vis-à-vis des différentes interprétations possibles de la probabilité. Lorsque la probabilité correspond à un comptage, qu'il s'agisse d'une proportion au sein d'une population, ou d'une fréquence observée au cours d'un (grand) nombre d'expériences, cette définition est tout-à-fait naturelle, car elle traduit une propriété très simple d'additivité du comptage. Illustrons ceci sur un exemple : pour étudier les différents objets issus d'un chapeau de magicien, on fait appel à l'espace des possibles suivant

$$\Omega := \{\text{foulard bleu, foulard vert, foulard rouge, lapin, colombe, bouquet, alligator}\}.$$

Si la probabilité $\mathbb{P}$ que l'on choisit de définir sur Ω représente, par exemple, la fréquence relative avec laquelle chacun des objets est sorti au cours des N premiers tours de magie effectués avec le chapeau, on a, en utilisant la notation

$$N(\omega) = \text{nombre de fois où l'objet } \omega \text{ est sorti lors des } N \text{ premiers tours,}$$

$$\mathbb{P}(\omega) = \frac{N(\omega)}{N},$$

pour tout $\omega \in \Omega$. Considérons à présent l'événement (concret) «c'est un foulard qui sort», qui correspond à l'événement (formel)

$$\{\text{foulard bleu, foulard vert, foulard rouge}\} \subset \Omega.$$

Il est bien évident que le nombre de fois au cours des N tours où un foulard sort est la somme des nombres de fois où un foulard bleu, un foulard vert, ou un foulard rouge sort (nous ne considérons que des foulards unis). Si la probabilité de l'événement formel {foulard bleu, foulard vert, foulard rouge} doit représenter la fréquence avec laquelle un foulard est sorti, on a donc évidemment égalité entre

$$\mathbb{P}(\{\text{foulard bleu, foulard vert, foulard rouge}\})$$

et

$$\mathbb{P}(\text{foulard bleu}) + \mathbb{P}(\text{foulard vert}) + \mathbb{P}(\text{foulard rouge}),$$

ce qui justifie la formule générale définissant la probabilité d'un événement comme somme des probabilités des éventualités élémentaires qui le constituent :

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega).$$

Cette discussion est générale, la seule propriété que nous ayons utilisée étant qu'à un élément compté (un objet issu du chapeau) correspond un et un seul élément de Ω, ce qu'impose naturellement la définition de Ω comme liste des issues possibles, au niveau de description adopté, toute issue étant associée à un et un seul élément de Ω. Dans le cas où la probabilité est plutôt considérée comme une mesure de plausibilité, construite à partir de jugements et d'informations partielles, il est encore possible de justifier cette définition additive de la probabilité, en montrant qu'elle est en un certain sens la seule cohérente du point de vue du raisonnement en univers incertain (ceci fait partie de ce que l'on appelle le théorème de Cox, voir par exemple l'ouvrage de Howson et Urbach, ou l'article de Van Horn cités dans la bibliographie). A notre modeste niveau, disons simplement qu'il ne semble pas déraisonnable d'ajouter entre elles les plausibilités des différentes éventualités incompatibles produisant un événement pour estimer la plausibilité de cet événement.

Au passage, remarquons que la condition de normalisation

$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1,$$

résulte essentiellement d'une convention, qui consiste à attribuer à l'événement certain une probabilité de 1. Du point de vue des fréquences (ou du comptage), cette condition revient simplement à exprimer les fréquences relativement à l'effectif total (par exemple en pourcentages) plutôt qu'en termes absolus, ce qui rend leurs valeurs beaucoup plus faciles à interpréter et à comparer. Du point de vue des plausibilités, seul compte le rapport entre deux probabilités : le fait que telle éventualité soit deux fois plus probable que telle autre, par exemple, et il n'y a donc aucun inconvénient à supposer que la somme totale soit égale à 1, quitte à multiplier toutes les probabilités par un même nombre positif, ce qui ne change pas leurs rapports.

1.4.2 Probabilité et opérations sur les événements

Nous avons vu que les opérations logiques (et, ou, non) portant sur les événements concrets, étaient associées, du point de vue formel, aux opérations ensemblistes (union, intersection, complémentaire). Une question légitime est donc : comment la

probabilité se comporte-t-elle vis-à-vis de ces opérations ? Toutes les réponses générales (c'est-à-dire, sans formuler d'hypothèse supplémentaire) que l'on peut donner à ces questions résultent directement de la définition

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega),$$

et leur preuve est laissée en exercice (essentiellement, il suffit de faire une figure). En fait, nous vous invitons à systématiquement représenter par une figure l'espace Ω et les événements que vous étudiez, ce qui rend évidentes la plupart des formules ci-dessous, inutile leur mémorisation, et bien plus claire l'utilisation qu'il convient d'en faire dans votre contexte. Bien entendu, les idées que ces formules véhiculent sont importantes et il est nécessaire de les retenir ; nous aurons l'occasion de les utiliser abondamment dans la suite.

Tout d'abord, si A et B sont deux événements,

$$A \subset B \text{ entraîne que } \mathbb{P}(A) \leq \mathbb{P}(B).$$

Cette propriété est très importante, et l'on s'en servira, par exemple, pour montrer que la probabilité d'un événement A est petite en la comparant à celle d'un événement B dont la probabilité est elle-même petite, et plus facile à calculer que celle de A.

D'autre part, on a l'égalité

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Nous utiliserons rarement cette propriété telle quelle. Notez bien qu'en général,

$$\mathbb{P}(A \cup B) \neq \mathbb{P}(A) + \mathbb{P}(B),$$

mais que l'on a toujours l'inégalité :

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

Lorsque A et B sont incompatibles, c'est-à-dire lorsque $A \cap B = \emptyset$, $\mathbb{P}(A \cap B) = 0$ et l'on a égalité dans l'inégalité précédente. On déduit de ceci, par exemple, que

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

Plus généralement, si $A_1 \ldots, A_n$ est une famille d'événements deux-à-deux incompatibles, c'est-à-dire que, pour tout $1 \leq i \neq j \leq n$, $A_i \cap A_j = \emptyset$, on a :

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n). \tag{1.1}$$

Sans aucune hypothèse sur les A_i, on a simplement l'inégalité suivante (borne de la réunion) :

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n).$$

Rappelons, à toutes fins utiles, que le fait que les événements $A_1, \ldots, A_n$ soient deux-à-deux incompatibles ne se résume pas à la condition : $A_1 \cap \ldots \cap A_n = \emptyset$.

On utilisera souvent l'égalité 1.1 ci-dessus pour évaluer la probabilité d'un événement en termes de son «découpage» par une famille d'autres événements : si $A_1 \ldots, A_n$ est une famille d'événements deux-à-deux incompatibles recouvrant B, c'est-à-dire, si $B \subset A_1 \cup \ldots \cup A_n$, alors

$$\mathbb{P}(B) = \mathbb{P}(B \cap A_1) + \cdots + \mathbb{P}(B \cap A_n).$$

On définit également la notion de système complet d'événement (que nous utiliserons peu en tant que telle dans ce cours) : une famille d'événements $A_1, \ldots, A_n$ forme un **système complet d'événements** (ou encore une partition de Ω) si les deux conditions suivantes sont vérifiées :

1. $A_1 \ldots, A_n$ est une famille d'événements deux-à-deux incompatibles ;
2. la famille $A_1 \ldots, A_n$ recouvre Ω, autrement dit $\Omega = A_1 \cup \cdots \cup A_n$.

D'après ce qui précède, la probabilité de tout événement B peut alors s'écrire

$$\mathbb{P}(B) = \mathbb{P}(B \cap A_1) + \cdots + \mathbb{P}(B \cap A_n).$$

Dans le cas d'une famille d'événements quelconques, la borne de la réunion peut être raffinée en inégalités plus compliquées (mais en général plus précises), ou en égalité (principe d'inclusion-exclusion). Précisément, en posant, pour $1 \leq k \leq n$,

$$C_k = \sum_{i_1 < \cdots < i_k} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}),$$

on a

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) \leq \sum_{k=1}^{m} (-1)^{k-1} C_k$$

lorsque m est impair,

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) \geq \sum_{k=1}^{m} (-1)^{k-1} C_k$$

lorsque m est pair, et

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) = \sum_{k=1}^{n} (-1)^{k-1} C_k.$$

La preuve de ces résultats constitue l'exercice 10

En passant, rappelons que les événements qui apparaissent dans les expressions ci-dessus sont implicitement supposés former des sous-ensembles d'un même espace des possibles Ω. Des opérations ensemblistes pratiquées sur des sous-ensembles d'espaces des possibles différents n'ont pas de sens !

1.4.3 Quelques exemples de modèles probabilistes

Un exemple horticole :

$$\Omega = \{\text{chou, carotte, navet, potiron, courge, cerfeuil, fenouil}\},$$

$$\mathbb{P}(\text{chou}) = 2/157,\ \mathbb{P}(\text{carotte}) = 30/157,\ \mathbb{P}(\text{navet}) = 24/157,\ \mathbb{P}(\text{potiron}) = 53/157,$$

$$\mathbb{P}(\text{courge}) = 21/157,\ \mathbb{P}(\text{cerfeuil}) = 8/157,\ \mathbb{P}(\text{fenouil}) = 9/157.$$

Un exemple brassicole :

$$\Omega = \{\text{Heineken, Kronembourg, Stella, Mützig, Guiness, Jenlain, Duvel}\},$$

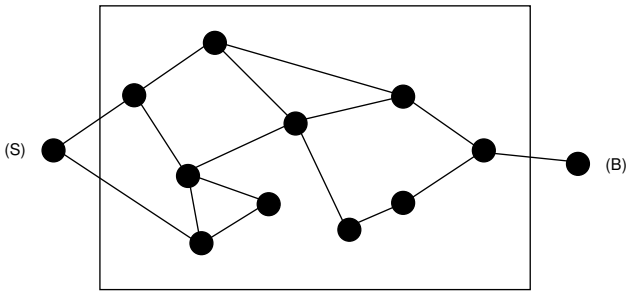
$$\mathbb{P}(\text{Heineken}) = 2/28,\ \mathbb{P}(\text{Kronembourg}) = 1/28,\ \mathbb{P}(\text{Stella}) = 1/28,\ \mathbb{P}(\text{Mützig}) = 6/28,$$

$$\mathbb{P}(\text{Guiness}) = 5/28,\ \mathbb{P}(\text{Jenlain}) = 7/28,\ \mathbb{P}(\text{Duvel}) = 6/28.$$

Ces deux exemples un peu farfelus dont destinés à illustrer le fait qu'un modèle probabiliste abstrait (un modèle probabiliste vu simplement comme un objet mathématique) peut prendre absolument n'importe quelle forme, et que la seule contrainte est que la somme des probabilités de toutes les éventualités élémentaires soit égale à 1. Bien entendu, sans dictionnaire permettant de relier ces modèles abstraits avec une quelconque réalité, ils restent totalement... abstraits. On peut néanmoins s'intéresser à l'étude de leurs propriétés mathématiques.

Un exemple paramétrique :

La figure ci-dessous est le schéma d'un réseau électronique de communication reliant un point source (S) à un point but (B), et comportant un certain nombre de noeuds intermédiaires reliés entre eux par des connexions. A un instant donné, chaque connexion peut éventuellement se trouver coupée, à la suite d'incidents techniques.



Nous décrirons le fonctionnement en termes de fonctionnement/panne de chacune des connexions. Une issue de l'expérience est donc la donnée, pour chacune des connexions, du fait qu'elle fonctionne ou non. Numérotant les connexions de 1 à 16, nous prendrons donc pour espace des possibles l'ensemble

$$\Omega = \{0,1\}^{16} = \{(x_1, \ldots, x_{16}) \ : \ x_i \in \{0,1\}\},$$

avec la convention que $x_i = 1$ traduit le fait que la connexion numéro i fonctionne, et que $x_i = 0$ traduit une panne de la connexion numéro i. Notons que cette modélisation est beaucoup plus riche que celle qui consisterait simplement à coder la circulation ou la non-circulation de l'information de (S) à (B). Nous définirons la probabilité $\mathbb{P}$ sur Ω par :

$$\mathbb{P}\left[(x_1, \ldots, x_{16})\right] = \prod_{i=1}^{16} p^{x_i}(1-p)^{1-x_i},$$

où $p \in [0,1]$ est un paramètre. (Il faudra vérifier que la formule ci-dessus définit bien une probabilité, c'est-à-dire qu'elle donne toujours lieu à des nombres compris entre 0 et 1 et dont la somme sur tous les éléments de Ω est égale à 1 : Exercice 5 !) Autrement dit, la probabilité d'une configuration de fonctionnement/panne des connexions est obtenue en effectuant le produit de 16 facteurs, un par connexion, égal à p lorsque la connexion correspondante fonctionne, et à $(1-p)$ lorsque celle-ci est coupée. La forme de $\mathbb{P}$ est donc fixée (un produit de 16 facteurs), et seule manque la valeur du paramètre p, qu'il faudrait pouvoir évaluer, pour déterminer les valeurs numériques de $\mathbb{P}$. En fonction de p, on peut notamment calculer la valeur de la quantité qui nous intéresse, c'est-à-dire la probabilité de l'événement A =«l'information circule entre (S) et (B)». Par définition,

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega),$$

et, par définition, $\omega \in A$ si et seulement s'il existe une suite de connexions en fonctionnement reliant (S) à (B). Il suffit donc, pour calculer $\mathbb{P}(A)$, de dresser la liste de toutes les configurations fonctionnement/panne telles que (S) et (B) communiquent, puis de calculer la somme ci-dessus, portant sur toutes les configurations de cette liste, en prenant garde au fait que la valeur de $\mathbb{P}(\omega)$ n'est pas la même suivant les configurations. Nous retrouverons souvent cette situation où $\mathbb{P}$ possède une forme fixée et s'exprime en fonction d'un petit nombre de paramètres. Notez bien qu'il ne s'agit ici que d'un exemple, et que cette probabilité n'a aucune raison *a priori* de convenir à la description du réseau étudié.

Un exemple important : la probabilité uniforme

Si l'espace des possibles Ω est un ensemble fini, on peut définir une probabilité qui attribue à chaque éventualité élémentaire la même probabilité, appelée **probabilité**

uniforme. Notons $|\Omega|$ le cardinal de Ω, c'est-à-dire le nombre d'éléments de Ω. Pour satisfaire la condition de normalisation $\sum_{\omega\in\Omega}\mathbb{P}(\omega)=1$, on voit que la valeur de $\mathbb{P}(\omega)$ doit nécessairement satisfaire l'égalité

$$\mathbb{P}(\omega)=\frac{1}{|\Omega|}.$$

Le plus souvent, ce sont des considérations de symétrie (nous en reparlerons plus bas) qui amènent à attacher *a priori* à Ω la probabilité uniforme : si les différentes issues présentent une certaine symétrie, de telle sorte qu'aucune ne semble favorisée par rapport à une autre, ce choix s'impose comme une première suggestion, qui doit naturellement être validée, par l'expérience et/ou par d'autres arguments, suivant le contexte. Attention : ce n'est **que dans le cas très particulier de la probabilité uniforme** que l'on peut appliquer la célèbre formule :

$$\mathbb{P}(A)=\frac{\text{nombre de cas favorables}}{\text{nombre de cas total}}.$$

Celle-ci n'est donc pas une règle générale, mais simplement une conséquence de l'hypothèse de modélisation qui affirme que toutes les éventualités élémentaires sont également probables.

En effet, si Ω est muni de la probabilité uniforme, la probabilité d'un événement A est égale à :

$$\mathbb{P}(A)=\sum_{\omega\in A}\mathbb{P}(\omega)=\sum_{\omega\in A}\frac{1}{|\Omega|}=\frac{|A|}{|\Omega|}.$$

Le «nombre de cas favorables» à la réalisation de A désigne simplement le nombre d'éventualités élémentaires qui réalisent A, et le «nombre de cas total» le nombre total d'éventualités élémentaires présentes dans Ω. Dans les exemples précédents, on voit bien que cette formule ne s'applique absolument pas : le nombre de cas favorable à l'événement «chou ou carotte» est de 2, mais sa probabilité est égale à

$$\mathbb{P}(\{\text{chou, carotte}\})=\mathbb{P}(\text{chou})+\mathbb{P}(\text{carotte})=32/157,$$

qui diffère sensiblement de la valeur $2/7$ que donnerait la formule «nombre de cas favorables»/ «nombre de cas total».

Pour vous convaincre encore davantage, voici un exemple historique destiné à vous mettre en garde contre l'utilisation incontrôlée de cette formule, et à justifier l'utilisation d'un formalisme précis.

Exemple : le problème du Chevalier de Méré

Le Chevalier de Méré soumit, dit-on, à la sagacité de Pascal [4] le paradoxe suivant : on constate en pratique que l'on obtient plus souvent 11 que 12 en lançant trois dés

4. Blaise Pascal, 1623–1662.

et en effectuant la somme de leurs chiffres. Pourtant, le nombre de combinaisons dont la somme fait 12 est le même que le nombre de combinaisons dont la somme fait 11... En effet, les combinaisons donnant lieu à un total de 11 sont les suivantes :

$$\{1;4;6\}, \{1;5;5\}, \{2;3;6\}, \{2;4;5\}, \{3;3;5\}, \{3;4;4\}$$

tandis que les combinaisons donnant lieu à un total de 12 sont les suivantes :

$$\{1;5;6\}, \{2;4;6\}, \{2;5;5\}, \{3;3;6\}, \{3;4;5\}, \{4;4;4\}$$

soit six combinaisons dans les deux cas. La solution de ce paradoxe apparent réside dans une description précise du modèle probabiliste de l'expérience consistant à lancer trois dés. On peut notamment envisager deux espaces des possibles pour décrire l'expérience. Le premier, Ω_1, dans lequel on attribue à chacun des trois dés un numéro, et qui exprime le résultat du lancer sous la forme d'un triplet ordonné (a, b, c) donnant, dans l'ordre, le résultat du dé numéroté 1, le résultat du dé numéroté 2, et le résultat du dé numéroté 3. Comme la seule quantité à laquelle nous nous intéressons est la valeur de la somme des trois chiffres obtenus, il n'est pas indispensable d'ordonner les résultats des dés, et il suffit de décrire l'expérience en donnant les trois chiffres obtenus sans préciser leur ordre d'apparition, sous la forme d'un triplet non-ordonné $\{a, b, c\}$, dont l'ensemble forme l'espace des possibles Ω_2. Dans le premier cas, des considérations classiques de modélisation des lancers (sur lesquelles nous reviendrons : indépendance des lancers successifs, et description du résultat de chaque lancer par la probabilité uniforme) suggèrent que tous les triplets ordonnés (a, b, c) devraient être supposés équiprobables, et que l'expérience doit donc être décrite, au moins en première approximation, par la probabilité uniforme $\mathbb{P}_1$ sur Ω_1, chacun des 216 triplets ordonnés ayant donc une probabilité de $1/216$. Au contraire, sur Ω_2, les mêmes considérations de modélisation entraînent que tous les triplets non-ordonnés ne **devraient pas** être équiprobables, et donc que ce n'est **pas** la probabilité uniforme $\mathbb{P}_2$ sur Ω_2 qui décrit convenablement l'expérience. Par exemple, le triplet non-ordonné $\{2;5;5\}$, correspond, dans la description obtenue à l'aide de Ω_1, à la réunion des trois triplets $(2;5;5)$, $(5;2;5)$, $(5;5;2)$, et on doit donc, pour être cohérent avec la description précédente, lui attribuer la probabilité $3 \times 1/216$. Au contraire, le triplet non-ordonné $\{2;4;6\}$ correspond à la réunion des six triplets $(2;4;6)$, $(2;6;4)$, $(4;2;6)$, $(4;6;2)$, $(6;2;4)$, $(6;4;2)$, et on doit donc lui attribuer la probabilité $6 \times 1/216$. La probabilité sur Ω_2 n'étant pas uniforme, le raisonnement qui consiste à compter le nombre de cas favorables pour calculer la probabilité d'un événement n'est pas valable, puisque les différents «cas» favorables n'ont pas tous la même probabilité, ce qui lève le paradoxe. On vérifie que la probabilité d'obtenir 11 est de $27/216$ tandis que la probabilité d'obtenir 12 est de $25/216$, ce qui rend compte de la différence observée dans les fréquences d'apparition. Il n'y a donc pas

compatibilité entre la description de l'expérience par le modèle $(\Omega_1, \mathbb{P}_1)$ et sa description par le modèle $(\Omega_2, \mathbb{P}_2)$, et, en l'occurrence, l'expérience courante ainsi que des considérations classiques de modélisation conduisent à choisir le premier modèle plutôt que le second. Nous reviendrons sur la notion de compatibilité entre plusieurs modèles pour décrire le même phénomène. Retenons au passage que, même dans des cas extrêmement simples, où l'on ne s'attend pas à rencontrer la moindre difficulté, il est indispensable de bien préciser le modèle utilisé et les hypothèses que l'on formule à son sujet.

Le prestige de l'uniforme

Nous sommes demeurés quelque peu vagues sur les considérations de modélisation justifiant l'utilisation de la loi uniforme dans ce problème. De fait, suivant l'interprétation concrète que l'on donne à la notion de probabilité dans le contexte envisagé, la nature des arguments susceptibles de justifier raisonnablement la description d'une situation au moyen d'un espace des possibles muni de la probabilité uniforme – tout au moins en première approximation –, varie considérablement.

Le « principe de raison insuffisante », ainsi qu'il est parfois appelé, stipule que la probabilité uniforme doit être employée dès lors que l'ensemble des informations dont on dispose sur la situation étudiée sont symétriques vis-à-vis des différentes éventualités élémentaires, c'est-à-dire n'établissent pas de différence entre elles. L'utilisation de ce «principe» appelle au moins trois précautions importantes. D'une part, il n'est quasiment jamais vrai que la totalité des informations disponibles soient totalement symétriques vis-à-vis des différentes éventualités élémentaires. En général, on élimine un certain nombre d'informations dont l'importance est jugée négligeable, et l'on s'accomode d'une symétrie approximative.

D'autre part, ce «principe» est nécessairement cantonné à l'utilisation de la probabilité comme mesure de plausibilité au vu des informations disponibles, et ne saurait certainement pas s'appliquer à la probabilité entendue comme fréquence, ou comme proportion, sans quoi, nous serions en train de déduire de notre propre ignorance au sujet d'une situation des affirmations objectives quant à celle-ci, ce qui est fortement déraisonnable ! Il faut garder ceci en tête lorsque l'on utilise cet argument pour attribuer a priori des probabilités.

Enfin, l'utilisation de ce «principe» suppose que l'on souhaite effectivement attribuer des probabilités aux différentes éventualités élémentaires. Si l'on dispose d'un ensemble d'informations trop limité, on peut décider de ne pas affecter de probabilités, soit qu'on les laisse inattribuées (sous la forme de paramètres) dans le raisonnement, soit même que l'on renonce à décrire la situation dans le cadre de la modélisation probabiliste si l'on juge que l'on dispose vraiment de trop peu d'information et qu'une telle description ne peut être menée à bien. Quoiqu'il en soit, la

pertinence de l'attribution des probabilités par le «principe de raison insuffisante» est clairement conditionnée par la quantité et la qualité des informations dont on dispose (une chose est de poser des probabilités égales entre plusieurs alternatives car on estime ne posséder aucune information autre que l'existence de ces différentes alternatives, une autre est de vérifier qu'un vaste ensemble d'informations relatives à cette situation ne fait apparaître aucune différence entre ces alternatives). Des généralisations de ce principe à des situations plus complexes (maximum d'entropie conditionnelle aux informations disponibles) existent et jouent un rôle important dans les méthodes bayésiennes, voir par exemple l'ouvrage de Howson et Urbach cité dans la bibliographie.

Bien entendu, les arguments de symétrie peuvent également être appelés à jouer un rôle dans le cadre des autres interprétations de la probabilité. Par exemple, dans le cas du problème du Chevalier de Méré, le caractère symétrique des dés employés, joint à des hypothèses supplémentaires sur le processus produisant les lancers (voir par exemple les articles sur les lancers de pièce de monnaie cités dans la bibliographie, et les références qui s'y trouvent), pourra conduire à supposer que les probabilités, entendues au sens fréquentiel, doivent être les mêmes pour chacun des résultats possibles des lancers (tels que décrits par Ω_1). Mais il s'agit alors d'un rôle explicatif positif, qui ne se résume pas à constater que les informations dont nous disposons ne font pas apparaître de différence entre les différentes issues possibles. Insistons sur le fait qu'il est de toute façon indispensable d'expérimenter pour valider un tel modèle : on ne peut modéliser exclusivement à partir d'arguments théoriques *a priori*, et une confrontation à la réalité modélisée est indispensable. Dans ce contexte, disposer de longues listes de résultats expérimentaux sur les lancers de dés ne faisant pas apparaître de différences significatives entre les divers résultats possibles (tels que décrits par Ω_1), constitue une information symétrique vis-à-vis des éléments de Ω_1, et joue un rôle explicatif positif dans le choix d'une probabilité (au sens fréquentiel) uniforme sur Ω_1 pour décrire la situation.

1.5 Probabilités conditionnelles

La notion de probabilité conditionnelle est à peu près aussi fondamentale, dans le cadre de la modélisation probabiliste, que la notion de probabilité elle-même. D'ailleurs, en un certain sens, toute probabilité est une probabilité conditionnelle qui s'ignore, et il est indispensable de bien maîtriser cette notion, qui sera pour nous l'une des briques de base dans la construction de modèles probabilistes.

Le point de vue formel

Commençons par donner une définition formelle. Etant donné un espace de probabilité $(\Omega, \mathbb{P})$ et un événement A de probabilité non-nulle ($\mathbb{P}(A) > 0$), on appelle **probabilité $\mathbb{P}$ conditionnelle à A** (ou encore probabilité $\mathbb{P}$ conditionnée par la réalisation de A, probabilité $\mathbb{P}$ sachant A) la probabilité définie sur Ω par :

$$\begin{cases} \mathbb{P}(\omega|A) = \frac{\mathbb{P}(\omega)}{\mathbb{P}(A)} \text{ si } \omega \in A, \\ \mathbb{P}(\omega|A) = 0 \text{ si } \omega \notin A. \end{cases}$$

Avant tout commentaire, vérifions que la définition ci-dessus donne effectivement lieu à une probabilité sur Ω. Pour toute éventualité élémentaire ω élément de Ω, $\mathbb{P}(\omega|A)$ est bien un nombre positif ou nul car il est égal, soit à zéro, soit au quotient d'un nombre positif ou nul par un nombre positif. Il reste à vérifier la condition de normalisation, autrement dit, à vérifier que

$$\sum_{\omega\in\Omega} \mathbb{P}(\omega|A) = 1.$$

Pour cela, décomposons cette somme en deux parties :

$$\sum_{\omega\in\Omega} \mathbb{P}(\omega|A) = \sum_{\omega\in A} \mathbb{P}(\omega|A) + \sum_{\omega\notin A} \mathbb{P}(\omega|A).$$

Lorsque ω n'est pas dans A, $\mathbb{P}(\omega|A)$ est nul, et la somme la plus à droite dans l'égalité ci-dessus est donc égale à zéro. On obtient donc que :

$$\sum_{\omega\in\Omega} \mathbb{P}(\omega|A) = \sum_{\omega\in A} \mathbb{P}(\omega|A) = \sum_{\omega\in A} \frac{\mathbb{P}(\omega)}{\mathbb{P}(A)} = \frac{1}{\mathbb{P}(A)} \sum_{\omega\in A} \mathbb{P}(\omega) = \frac{1}{\mathbb{P}(A)} \times \mathbb{P}(A) = 1,$$

et la condition de normalisation est donc bien vérifiée. La probabilité $\mathbb{P}$ conditionnelle à un événement est donc une probabilité sur l'espace des possibles Ω, **au même titre que la probabilité initiale** $\mathbb{P}$ à partir de laquelle elle est définie. En particulier, on peut parler de la probabilité conditionnelle d'un événement : si A est un événement de probabilité non-nulle sur l'espace de probabilité $(\Omega, \mathbb{P})$, et B un événement, la probabilité de B vis-à-vis de la probabilité $\mathbb{P}$ sachant A est donnée, d'après la définition de $\mathbb{P}(\cdot|A)$, par :

$$\mathbb{P}(B|A) = \sum_{\omega\in B} \mathbb{P}(\omega|A),$$

conformément à la définition de la probabilité d'un événement comme somme des probabilités des éventualités élémentaires qui le constituent. Insistons lourdement : $\mathbb{P}(B|A)$ représente la probabilité de l'événement B, calculée par rapport à une probabilité définie sur Ω, mais différente de $\mathbb{P}$, à savoir $\mathbb{P}(\cdot|A)$. Il ne s'agit pas de la

probabilité sous $\mathbb{P}$ d'un hypothétique événement «B sachant A». L'usage, un peu ambigu à cet égard, est de lire $\mathbb{P}(B|\ A)$ comme «$\mathbb{P}$ de B sachant A», ou, lorsque l'on omet la référence (pourtant importante !) à $\mathbb{P}$, «probabilité de B sachant A».

En tenant compte du fait que, pour tout ω dans B, $\mathbb{P}(\omega|A)$ est nul si ω n'est pas élément de A, et $\mathbb{P}(\omega|A) = \mathbb{P}(\omega)/\mathbb{P}(A)$ si ω est élément de A, on peut réarranger l'expression ci-dessus :

$$\mathbb{P}(B|A) = \sum_{\omega \in A \cap B} \frac{\mathbb{P}(\omega)}{\mathbb{P}(A)} = \frac{1}{\mathbb{P}(A)} \sum_{\omega \in A \cap B} \mathbb{P}(\omega) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}.$$

On retient le résultat de cette suite d'égalités, connu sous le nom de **formule de Bayes** [5] :

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}.$$

Une conséquence immédiate est ce que l'on désigne parfois sous le nom de **formule des probabilités totales** : si $A_1, \ldots, A_n$ forme un système complet d'événements, la probabilité de tout événement B s'écrit :

$$\mathbb{P}(B) = \mathbb{P}(B|A_1) \times \mathbb{P}(A_1) + \cdots + \mathbb{P}(B|A_n) \times \mathbb{P}(A_n).$$

Contexte

La notion de probabilité conditionnelle à un événement trouve une interprétation concrète qui varie selon le sens donné à la probabilité dans le modèle auquel on se réfère. Commençons par discuter le cas de la probabilité vue comme mesure de plausibilité en situation d'incertitude. Dans ce contexte, **les probabilités conditionnelles telles que nous les avons définies indiquent de quelle manière modifier le modèle probabiliste initial pour tenir compte d'un apport d'information relatif à la réalisation de la situation considérée**. La notion d'information s'entend ici dans un sens bien précis : il s'agit du fait qu'un certain événement $A \subset \Omega$ est réalisé. A priori, la multiplicité des éléments de Ω traduit notre ignorance de la manière exacte dont l'expérience va effectivement se réaliser ; parmi toutes les éventualités élémentaires présentes dans notre modélisation, une seule correspond à réalisation effective de la situation, mais nous ne savons pas laquelle. Le fait de savoir que l'événement A est réalisé nous permet de réduire notre ignorance en restreignant la liste des éventualités élémentaires qui sont effectivement susceptibles de se réaliser (celles qui correspondent à la réalisation de A sont encore possibles, les autres sont éliminées). Le principe de l'utilisation des probabilités conditionnelles, que nous devrons justifier, est alors le suivant : si l'on décrit *a priori* (c'est-à-dire, avant d'incorporer l'information selon laquelle A s'est réalisé) la situation à l'aide

5. Thomas Bayes, 1702–1761.

du modèle $(\Omega, \mathbb{P})$, il est indispensable, pour tenir compte du fait que A est réalisé, de remplacer le modèle $(\Omega, \mathbb{P})$ par le modèle $(\Omega, \mathbb{P}(\cdot|A))$ dans la description de la situation étudiée. Dans le cadre de l'interprétation «fréquentielle» de la probabilité (qui, rappelons-le, est très souvent utilisé de pair avec l'interprétation en termes de plausibilité), la probabilité conditionnelle apparaît lorsque l'on cherche à décrire, non pas la population (resp. la série d'expériences) d'origine, décrite par le modèle $(\Omega, \mathbb{P})$, mais la sous-population (resp. la sous-série d'expériences) obtenue en sélectionnant les éléments de la population (resp. les expériences) conduisant à la réalisation de l'événement A. Le principe de l'utilisation des probabilités conditionnelles est alors le suivant : **si la population (resp. la série d'expériences) d'origine est décrite par** $(\Omega, \mathbb{P})$**, la sous-population (resp. la sous-série d'expériences) formée en sélectionnant les éléments de la population (resp. la série d'expériences) pour lesquels** A **est réalisé est décrite par le modèle** $(\Omega, \mathbb{P}(\cdot|A))$**.**

Insistons bien sur le point suivant : quel que soit le contexte retenu pour l'interprétation de la probabilité, il est **indispensable**, dans les situations décrites ci-dessus, de remplacer le modèle d'origine $(\Omega, \mathbb{P})$ par le modèle modifié $(\Omega, \mathbb{P}(\cdots|A))$. Sinon, on est conduit à raisonner de manière incohérente (dans l'interprétation «plausible») ou à évaluer des fréquences de manière erronnée (dans l'interprétation «fréquentielle»).

Commençons par préciser ceci sur deux exemples.

Dans ce premier exemple, nous discuterons de l'interprétation de la probabilité en termes de plausibilité (mais l'interprétation fréquentielle aurait également toute sa place ici). Intéressons-nous donc à la composition en filles et garçons des familles de deux enfants, en choisissant pour espace des possibles :

$$\Omega = \{\mathrm{GG},\ \mathrm{FG},\ \mathrm{GF},\ \mathrm{FF}\},$$

où G représente une fille, F un garçon, la première lettre codant pour l'aîné, la seconde pour le cadet.

Une manière très simple, et grossièrement en accord avec les données démographiques, est de munir Ω de la probabilité uniforme :

$$\mathbb{P}(\mathrm{GG}) = \mathbb{P}(\mathrm{FG}) = \mathbb{P}(\mathrm{GF}) = \mathbb{P}(\mathrm{FF}) = 1/4.$$

La probabilité qu'une famille soit formée de deux filles est donc, dans cette description, égale à $1/4$. Supposons maintenant que l'on **sache** qu'une famille donnée comporte au moins une fille, sans connaître pour autant le détail de sa composition. Comment la probabilité que la famille soit constituée de deux filles est-elle modifiée par cette information supplémentaire ? Intuitivement, il semble clair que le fait de savoir que la famille comporte déjà une fille doit accroître la probabilité pour qu'elle en comporte deux, puisque l'on sait déjà qu'une «partie» de l'événement «avoir deux filles» est effectivement réalisée. Examinons l'espace des possibles. L'information dont

nous disposons nous permet d'affirmer que l'éventualité élémentaire GG (deux garçons) n'est pas réalisée, et est même équivalente à cette affirmation. Rien ne nous permet en effet de trancher plus particulièrement en faveur de l'une ou l'autre des trois éventalités élémentaires restantes FG,GF et FF, qui avaient *a priori* (à cause du choix initial de la probabilité uniforme pour décrire l'expérience) des probabilités égales de se réaliser. Le seul choix de probabilité cohérent avec l'information supplémentaire dont nous disposons est donc :

$$\mathbb{P}(\mathrm{FG}) = \mathbb{P}(\mathrm{GF}) = \mathbb{P}(\mathrm{FF}) = 1/3 \ , \text{ et } \mathbb{P}(\mathrm{GG}) = 0,$$

et il coïncide bien avec la définition générale de la probabilité conditionnelle que nous avons donnée. Notez, au passage que l'on n'obtient pas la même probabilité conditionnelle en supposant que, par exemple, l'aîné des deux enfant est une fille : dans notre modèle, la probabilité que le deuxième enfant soit une fille sachant que le premier l'est est égale à $1/2$. En revanche, savoir que l'un des deux enfants est une fille sans pour autant savoir s'il s'agit de l'aînée ou de la cadette nous conduit à une probabilité conditionnelle de $1/3$: ces deux informations ne donnent pas lieu à la même localisation de ω dans Ω. Quoiqu'il en soit, nous avons vu comment un apport d'information nous conduit nécessairement à modifier la probabilité sur l'espace des possibles. Pour des raisons évidentes de clarté, il est préférable de nommer différemment la probabilité conditionnelle à un événement obtenue à partir de $\mathbb{P}$ et la probabilité initiale $\mathbb{P}$, pour marquer le fait qu'il s'agit d'une nouvelle probabilité, conditionnelle à une information supplémentaire, qui résulte d'une modification de la probabilité initiale définie sur Ω.

Dans le second exemple, nous discuterons de l'interprétation fréquentielle de la probabilité (mais l'interprétation en termes de plausibilité aurait également toute sa place ici). Intéressons-nous donc à l'incidence de la consommation de tabac sur la survenue éventuelle d'un cancer. Un modèle probabiliste très simple utilisera l'espace des possibles :

$$\Omega = \{\mathrm{FS}\,;\, \mathrm{FC}\,;\, \mathrm{NS}\,;\, \mathrm{NC}\},$$

où F désigne le fait d'être un gros fumeur, N celui de ne pas l'être, C le fait d'être atteint un jour par un cancer et S le fait de ne pas l'être (S pour sain), et la probabilité comme la proportion de chacune des quatre éventualités (FS, FC, NS, NC) au sein de la population (bien entendu, il conviendrait de préciser exactement ce que l'on entend par gros fumeur, à quel instant on considère la population, etc..., mais nous nous affranchirons de ces détails pour conserver à notre exemple sa simplicité). Notez bien qu'il faut attendre la fin de la vie d'un individu pour savoir s'il va ou non développer un cancer, tandis que le fait d'être un gros fumeur ou non est observable bien avant, et que l'on est typiquement dans le cas où une observation partielle de ω est possible. En désignant par N la taille de la population, et par $N(\omega)$ le nombre

d'individus correspondant à l'éventualité élémentaire ω, on a, avec notre choix de définir les probabilités comme des proportions :

$$\mathbb{P}(\text{FS}) = \frac{N(\text{FS})}{N},\ \mathbb{P}(\text{FC}) = \frac{N(\text{FC})}{N},\ \mathbb{P}(\text{NS}) = \frac{N(\text{NS})}{N},\ \mathbb{P}(\text{NC}) = \frac{N(\text{NC})}{N}.$$

Toujours dans l'idée de définir les probabilités comme proportion au sein d'une population, la probabilité de développer un cancer sachant que l'on est fumeur doit donc être définie comme la proportion d'individus amenés à développer un cancer **parmi la population de fumeurs**, et non pas parmi la population totale soit :

$$\frac{N(\text{FC})}{N(\text{«fumeur»})},$$

où $N(\text{«fumeur»})$ désigne le nombre de (gros) fumeurs parmi la population totale. En notant que :

$$\frac{N(\text{FC})}{N(\text{«fumeur»})} = \frac{N(\text{FC})/N}{N(\text{«fumeur»})/N} = \frac{\mathbb{P}(\text{FC})}{\mathbb{P}(\text{«fumeur»})},$$

on constate que cette définition coïncide avec la définition générale des probabilités conditionnelles que nous avons donnée.

Les raisonnements construits dans le cadre des deux exemples précédents s'étendent facilement pour justifier la définition générale des probabilités conditionnelles, dans l'interprétation fréquentielle comme dans l'interprétation en termes de plausibilité.

Du point de vue des plausibilités, la question est de déterminer comment la plausibilité attribuée à chaque éventualité élémentaire doit être modifiée en tenant compte de l'information nouvelle que A s'est réalisé, d'une manière cohérente avec l'attribution initiale des plausibilités aux différents éléments de Ω. Si ω n'est pas un élément de A, autrement dit, si la réalisation de ω n'est pas compatible avec celle de A, nous sommes naturellement conduits à attribuer à ω une plausibilité nulle, puisque nous somme certains que ω n'est pas réalisé, la réalisation de A excluant celle de ω. Par ailleurs, le fait de savoir que A est réalisé ne nous apporte pas d'information particulière sur la façon dont A s'est réalisé, c'est-à-dire sur celle des éventualités élémentaires réalisant A qui est effectivement choisie par le hasard. Autrement dit, si nous estimions, avant de savoir que A était réalisé, qu'une éventualité élémentaire ω_1 réalisant A était deux fois plus plausible qu'une autre éventualité élémentaire ω_2 réalisant également A (autrement dit $\mathbb{P}(\omega_1) = 2 \times \mathbb{P}(\omega_2)$), le simple fait de savoir que A s'est réalisé ne fournit aucune raison de modifier cette estimation, et la probabilité conditionnelle doit donc vérifier : $\mathbb{P}(\omega_1|A) = 2 \times \mathbb{P}(\omega_2|A)$. Cependant, nous ne pouvons pas directement poser, comme il serait tentant de le faire, $\mathbb{P}(\omega|A) = \mathbb{P}(\omega)$ pour ω dans A, car, étant donné le fait que nous devons nécessairement poser $\mathbb{P}(\omega|A) = 0$ pour tout ω qui ne réalise pas A, cette définition ne conduirait pas à une probabilité,

la condition de normalisation n'étant pas satisfaite. Il est facile de voir que la condition selon laquelle les rapports entre les plausibilités des éléments de A doivent être conservés nous oblige à poser $\mathbb{P}(\omega|A) = c \times \mathbb{P}(\omega)$ pour tout ω dans A (et toujours $\mathbb{P}(\omega|A) = 0$ pour tout ω qui ne réalise pas A), où c est une constante. Il existe alors un unique choix de c qui garantit le fait que $\mathbb{P}(\cdots|A)$ définisse bien une probabilité sur Ω, à savoir $c = 1/\mathbb{P}(A)$, comme le montre le calcul effectué plus haut. De ce point de vue, la définition que nous avons donnée d'une probabilité conditionnelle est donc la seule cohérente (pour une justification basée sur des considérations qualitatives beaucoup plus générales, voir l'ouvrage de Howson et Urbach, ou l'article de Van Horn cités dans la bibliographie).

Du point de vue des fréquences, il suffit de raisonner exactement comme dans le cas du deuxième exemple : qu'il s'agisse de proportion au sein d'une population ou de fréquence observée au cours d'une longue série d'expériences répétées, cette manière de définir la probabilité conduit automatiquement à la définition que nous avons donnée d'une probabilité conditionnelle.

Mise en garde 2 *Il est important de ne pas confondre la probabilité d'un événement A sachant qu'un événement B est réalisé, et la probabilité de survenue de l'événement A et B. Dans le cas de l'exemple précédent, la probabilité de développer un cancer sachant que l'on est fumeur, que l'on pourrait encore baptiser «probabilité de développer un cancer* **si** *l'on fume», et la probabilité de développer un cancer* **et** *d'être fumeur sont deux probabilités définies de manière complètement différente. Dans les deux cas, on s'intéresse au nombre de fumeurs qui seront atteints d'un cancer, mais, dans le premier cas, ce nombre est rapporté au nombre de fumeurs, tandis que dans le deuxième cas, il est rapporté à l'effectif total de la population. Dans l'autre exemple, la probabilité d'avoir deux filles sachant que l'on en a au moins une était égale à 1/3, tandis que la probabilité d'avoir deux filles et d'en avoir au moins une est simplement la probabilité d'avoir deux filles, et se trouve, dans le modèle précédent, égale à 1/4.*

1.5.1 Notions de dépendance et d'indépendance entre événements

Le point de vue formel

Etant donnés un modèle probabiliste $(\Omega, \mathbb{P})$ et deux événements A et B tels que $\mathbb{P}(A) > 0$, nous dirons que :

- A favorise la survenue de B si $\mathbb{P}(B|A) > \mathbb{P}(B)$;
- A défavorise la survenue de B si $\mathbb{P}(B|A) < \mathbb{P}(B)$;
- A n'influe pas sur la survenue de B si $\mathbb{P}(B|A) = \mathbb{P}(B)$.

La définition ci-dessus ne fait pas jouer un rôle symétrique aux événements A et B. On vérifie pourtant facilement en appliquant cette même définition que, pour des événements A et B tels que $\mathbb{P}(A) > 0$ et $\mathbb{P}(B) > 0$, le fait que A favorise (resp.

défavorise, resp. n'influe pas sur) la survenue de B est équivalent au fait que B favorise (resp. défavorise, resp. n'influe pas sur) la survenue de A, qui est encore équivalent au fait que $\mathbb{P}(A \cap B)$ soit supérieur (resp. inférieur, resp. égal) à $\mathbb{P}(A) \times \mathbb{P}(B)$. On peut donc donner une forme symétrique à la définition précédente (qui a également l'avantage de s'appliquer à des événements de probabilité nulle), et l'on préférera utiliser une terminologie également symétrique vis-à-vis de A et de B. Ainsi, plutôt que de dire que A favorise B ou que B favorise A, on dira simplement que A et B sont positivement associés. De la même façon, on définira l'association négative de A et de B, et l'indépendance de A et de B.

Précisément, la définition que nous utiliserons est la suivante :

– A et B sont positivement associés si $\mathbb{P}(A \cap B) > \mathbb{P}(A) \times \mathbb{P}(B)$;
– A et B sont négativement associés si $\mathbb{P}(A \cap B) < \mathbb{P}(A) \times \mathbb{P}(B)$;
– A et B sont indépendants si $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$.

Dans les exemples précédents, l'événement «avoir au moins une fille» favorise l'événement «avoir deux filles», car la probabilité (inconditionnelle) d'avoir deux filles est de $1/4$, tandis que la probabilité d'avoir deux filles sachant que l'on en a au moins une est égale à $1/3$. En revanche, toujours dans notre modèle, le fait que le second enfant soit une fille est indépendant du fait que le premier le soit : la probabilité que le second enfant soit une fille est de $1/2$, et la probabilité que le second enfant soit une fille sachant que le premier enfant est une fille est également de $1/2$. Dans le cas de la relation tabac/cancer, les évaluations statistiques des probabilités du modèle montrent que la consommation de tabac favorise la survenue d'un cancer au sens précédent (on notera que le problème de l'évaluation des probabilités telles que nous les avons définies n'est pas si évident, puisqu'il n'est pas possible de déterminer pour les individus vivant actuellement s'ils vont ou non développer plus tard un cancer : une extrapolation à partir des données disponibles actuellement est incontournable).

Même si la notion de dépendance de deux événements entre eux est symétrique, on présente souvent les choses sous forme dissymétrique en comparant $\mathbb{P}(B|A)$ à $\mathbb{P}(B)$, ou $\mathbb{P}(A|B)$ à $\mathbb{P}(A)$, ce qui ne pose aucun problème dans l'absolu, mais donne souvent lieu à des confusions : on a vite fait de comparer $\mathbb{P}(A|B)$ à $\mathbb{P}(B)$ ou $\mathbb{P}(B|A)$ à $\mathbb{P}(A)$, ce qui perd toute signification. Par exemple, pour étudier l'incidence du tabagisme sur la santé, il est loisible de comparer la probabilité pour un fumeur d'être atteint d'un cancer à la probabilité d'être atteint d'un cancer tout court, ou, inversement, de comparer la probabilité pour un individu atteint d'un cancer d'être fumeur à la probabilité d'être fumeur. Dans un registre plus polémique, on pourra s'étonner que la probabilité pour un enfant d'ouvrier d'entrer à l'Ecole Polytechnique soit de moins de un sur mille alors que les enfants d'ouvriers représentent plus de 10% de la population. Pourtant, on compare ici ce qui n'est pas comparable : il faudrait, pour se faire une idée du rôle joué par l'origine sociale dans la poursuite d'études prestigieuses, soit comparer la probabilité pour un enfant d'ouvrier d'entrer à l'Ecole Polytechnique à la

probabilité pour un individu quelconque de devenir polytechnicien, soit la probabilité pour un polytechnicien d'avoir des parents ouvriers (actuellement, moins de 2%) au poids démographique des enfants d'ouvriers (plus de 10%). Dans un cas, on compare deux probabilités de l'ordre de quelques millièmes, dans l'autre cas, deux probabilités de l'ordre du dixième, et échanger les quantités que l'on doit comparer ne conduit qu'à une remarque vide de signification. Petit exercice (Exercice 16) : près de soixante pour cent des accidents de voiture graves impliquant de jeunes enfants se produisent dans des véhicules où les enfants ne sont pas correctement attachés (source : la brochure d'information de ma mutuelle). Soixante pour cent, cela fait beaucoup... A quoi faudrait-il comparer ce chiffre ?

Mise en garde 3 *Soulignons que la notion d'indépendance de deux événements dépend de la probabilité associée au modèle, et non pas simplement de la définition des événements considérés, qui, elle, ne se réfère qu'à l'espace des possibles* Ω*. Cela n'a rien d'étonnant, puisque* Ω *ne fait que représenter le degré de précision choisi pour décrire la situation étudiée, tout le reste de l'information sur la situation accessible au modèle étant contenue dans la probabilité* $\mathbb{P}$*. Lançons deux dés, et considérons les deux événements* $A =$*«la somme des deux chiffres obtenus est paire»* *et* $B =$*«le 1 ou le 2 sort au moins une fois». Les événements* A *et* B *sont-ils indépendants ? La question n'a pas de sens indépendamment des probabilités décrivant l'expérience. Si l'on munit l'espace des possibles*

$$\Omega = \{1; 2; 3; 4; 5; 6\} \times \{1; 2; 3; 4; 5; 6\}$$

de la probabilité uniforme, c'est effectivement le cas : A *et* B *sont indépendants. Si la probabilité n'est pas uniforme, ce n'est plus nécessairement le cas. Voir l'exercice 12*
L'indépendance de deux événements dépend de la probabilité sur l'espace des possibles qui décrit l'expérience.

Mise en garde 4 *Il importe de ne pas confondre la notion d'événements indépendants avec celle d'événements incompatibles. Ces deux notions n'ont rien à voir. Rappelons que deux événements* A *et* B *sont incompatibles s'ils ne peuvent se réaliser en même temps, autrement dit, si* $A \cap B = \emptyset$*. Deux événements incompatibles peuvent-ils être indépendants ? Si c'était le cas, on aurait, par indépendance,*

$$0 = \mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B),$$

et donc l'un des deux événements au moins devrait avoir une probabilité nulle. Dans tous les autres cas, deux événements ne peuvent pas être à la fois incompatibles et indépendants, ce qui est également évident intuitivement : si A *et* B *sont incompatibles, le fait de savoir que* A *est réalisé entraîne automatiquement que* B *n'est*

pas réalisé, autrement dit, apporte une information importante sur B. En particulier, le fait que A et B soient indépendants n'entraîne en aucune manière le fait que $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Dépendance et causalité

Comme vous l'avez certainement remarqué, un modèle probabiliste d'une situation se contente de dresser une liste des différentes issues possibles d'une situation, et d'attribuer à chacune d'entre elles une probabilité. De manière générale, il n'identifie pas, ni ne cherche à décrire, un quelconque mécanisme causal sous-jacent à cette situation, qui expliquerait comment s'effectue le choix de l'issue qui est effectivement réalisée parmi l'ensemble des éventualités élémentaires possibles. Bien souvent, c'est justement parce qu'un tel mécanisme est inconnu, ou trop complexe, ou impossible à décrire avec suffisament de précision, que l'on a recours à une modélisation de type probabiliste. Quoiqu'il en soit, la notion de mécanisme causal, ou de relation de cause à effet, qui est au centre des modèles déterministes en sciences (telle cause produit telle conséquence, qui à son tour devient la cause d'une autre conséquence, etc...), est en grande partie remplacée, dans le contexte des modèles probabilistes, par la notion de dépendance définie ci-dessus. Malgré les apparences, la notion de dépendance probabiliste est tout à fait distincte de la notion de relation de cause à effet, et nous allons chercher, dans cette partie, à en délimiter un certain nombre de points communs et de différences. Indiquons dès maintenant que **confondre dépendance probabiliste et relation de cause à effet constitue une très grave erreur de raisonnement**, que cette erreur est malheureusement très répandue, et que l'on est très facilement conduit à la commettre.

Notons d'abord que la notion de causalité est elle-même fort complexe et délicate, et l'on se heurte rapidement à des questions philosophiques si on cherche à l'analyser avec un minimum de détail, ce dont nous nous garderons bien. Très grossièrement, on peut tenter de définir l'existence d'un lien de cause à effet entre un événement A et un événement B lorsque la réalisation de B suit celle de A et lorsque, en l'absence de A, mais toutes choses égales d'ailleurs, on peut conclure que B ne se serait pas réalisé.

Remarquons, cette esquisse de définition étant posée, que de nombreuses explications *a posteriori*, proposées par des experts ou des profanes pour rendre compte des phénomènes les plus variés (par exemple, le prix de tel type de bien, le taux de chômage, les chiffres de la délinquance,...), prétendent distinguer des causes en se basant simplement sur le fait que tel facteur était présent avant l'effet constaté, sans jamais faire allusion à ce qui se serait passé en l'absence de ce facteur (aurait-on ou non observé l'effet en question ?). En fonction de ses préjugés ou de ses intérêts, chacun pourra donc invoquer à sa guise l'explication qui lui sied le mieux, sans tenir

compte de la possibilité que plusieurs facteurs entrent en cause simultanément, ou que le facteur présenté comme la cause n'ait peut-être eu aucun effet réel. Erreur grossière de logique, mais tellement répandue, qui porte le doux nom de «cum hoc ergo propter hoc»...

Dans un contexte statistique, nous ne nous laisserons bien entendu jamais aller à commettre une telle erreur, et nous aurons toujours, au moins schématiquement, comparaison entre une **population témoin**, dans laquelle le facteur causal présumé est absent, et une **population test** dans laquelle celui-ci est présent.

Voici quelques exemples concrets de dépendances (ou encore, d'associations) observées (provenant de divers pays) dans un tel contexte. Voir l'exercice 29.

- en Italie, on a constaté que les régions dans lesquelles les taux d'achat d'ordinateur personnels sont les plus importants sont également celles où les taux de divorce sont les plus élevés ;
- une étude japonaise portant sur 40000 quadragénaires montre que ceux qui se brossent les dents après chaque repas parviennent mieux que les autres à garder la ligne ;
- il existe une association positive entre utilisation de crème solaire et cancer de la peau ;
- le nombre de noyades est positivement associé à la consommation de crèmes glacées ;
- le prix des cigarettes est négativement associé au nombre des agriculteurs en Lozère ;
- en Ecosse, les achats de whisky sont positivement associés au montant des dons reçus par les églises ;
- la carte du vote Le Pen lors des élections présidentielles de 2002 se superpose avec celle de l'irradiation due au nuage de Tchernobyl ;
- dans les communes qui abritent des cigognes, la natalité est plus élevée que dans le reste du pays ;
- la confiance des investisseurs est positivement associée à la croissance économique ;
- la consommation régulière d'alcool pendant la grossesse est corrélée à des retards de QI et des difficultés d'apprentissage chez les enfants ;
- la hausse des recettes publiques allemandes est positivement associée à la hausse des dépenses des ménages espagnols ;
- la proportion de fonctionnaires dans une ville est négativement associée au dynamisme économique ;
- les enfants P*** acceptent plus volontiers les repas lorsqu'ils sont préparés par leur père que par leur mère ;
- la présence d'un médecin obstétricien lors d'un accouchement accroît la probabilité de complications ;

- le fait d'avoir recours à la péridurale diminue la mortalité lors des accouchements ;
- le nombre d'écoles maternelles dans une ville est positivement associé au nombre de crimes et délits ;
- les entreprises réalisant le plus de bénéfices sont celles qui ont les budgets publicitaires les plus importants ;
- un viticulteur diffuse de la musique classique dans son vignoble, et l'on constate que le vin obtenu est meilleur que celui produit par ses voisins, qui disposent pourtant de parcelles comparables pour l'ensoleillement et la nature du sol ;
- une faible cholestérolémie favorise l'apparition du cancer ;
- le fait de consommer régulièrement des moules accroît le risque d'attraper la grippe.

Certains exemples ci-dessus paraissent loufoques, d'autres plus recevables. Il est important de comprendre qu'aucune des associations mentionnées ci-dessus ne constitue un argument suffisant ou même sérieux pour affirmer l'existence d'une relation de cause à effet entre les variables qui sont mentionnées. Bien entendu, dans le cas des exemples franchement loufoques, personne ne peut sérieusement penser qu'il existe une telle relation. Cependant, les exemples d'apparence plus sérieuse sont de même nature, même s'il est beaucoup plus difficile dans leur cas de se défendre contre le penchant naturel consistant à interpréter une dépendance comme un rapport de cause à effet entre événements, par exemple, parce que nous sommes déjà convaincus de l'existence d'un tel rapport, et que nous sommes tentés de voir dans l'association observée une confirmation expérimentale de notre opinion, en omettant d'envisager sérieusement les autres explications possibles.

Cependant, on peut envisager plusieurs types d'explications à une association observée entre deux événements, et l'existence d'un lien de cause à effet ne constitue que l'une de ces explications. Mentionnons donc :

- un véritable lien de cause à effet, éventuellement complexe, entre événements : quand on tourne la clé de contact, le moteur se met en marche (du moins en l'absence de panne) ; les dépenses de publicité d'une entreprise jouent certainement un rôle sur ses bénéfices, mais ses bénéfices jouent certainement également un rôle sur ses dépenses de publicité ;
- un facteur dit de confusion, présentant une dépendance vis-à-vis de l'un des deux événements, mais sans lien de cause à effet avec celui-ci, et présentant en revanche un lien causal avec l'autre : le père des enfants P*** leur propose systématiquement des pâtes, tandis que leur mère leur propose souvent des légumes (les épinards, beurk !).
- une cause commune aux deux événements, mais cachée lorsque l'on fait état de l'association observée : la consommation de crèmes glacées et le nombre de noyades augmentent avec la température extérieure ;

– une coïncidence fortuite (celles-ci étant normalement bannies par la prise en compte d'échantillons de données de taille suffisante, plus à ce sujet dans le chapitre «Statistique»).

Dans la plupart des exemples ci-dessus, on peut facilement imaginer que des causes cachées ou des facteurs de confusion sont à l'origine des associations mentionnées.

Par exemple, les régions d'Italie dans lesquelles les achats d'ordinateurs personnels sont les plus élevés sont davantage les régions du nord, à l'économie prospère et au mode de vie moderne, que les régions du sud, moins développées économiquement, et où la tradition catholique est plus présente. Le mode de vie semble donc une cause susceptible d'expliquer la différence des taux de divorce, naturellement liée aux achats d'équipement informatique. Quoiqu'il en soit, on ne peut pas déduire de cette dépendance que l'utilisation intensive de l'ordinateur a tendance à isoler les époux et détruit les couples.

De même, on peut facilement imaginer qu'une bonne hygiène de vie s'accompagne à la fois d'un brossage de dents réguliers et d'une absence de surcharge pondérale. Le simple fait de se brosser les dents n'est probablement pas à lui seul responsable du maintien de la ligne !

Attention : nous n'avons pas prouvé que, dans ces deux exemples, l'association observée n'était pas due à un lien de cause à effet. Simplement, d'autres explications sont également possibles, et rien ne permet au vu de ce qui est mentionné, de privilégier l'une des explications plutôt que l'autre. Plus généralement, nous ne sommes certainement pas en train d'expliquer qu'une corrélation observée n'est jamais le signe d'une relation de cause à effet entre événements. Par exemple, dans le cas d'une consommation d'alcool au cours d'une grossesse, le risque lié de manière causale à l'alcool est considéré par les médecins comme parfaitement établi. Simplement, la mention de la dépendance statistique entre événements ne suffit pas à prouver l'existence d'un rapport de cause à effet, et ne constitue pas un argument solide pour l'établir, même si elle peut en constituer un indice. Dans l'exemple de l'alcool, on peut également imaginer qu'un facteur de confusion peut jouer un rôle, par exemple, le fait que la consommation régulière d'alcool chez les futures mères soit liée à des difficultés sociales ou relationnelles, qui, à leur tour, peuvent retentir sur les performances scolaires de l'enfant.

On voit ainsi comment **des informations stastistiques parfaitement correctes peuvent conduire à des interprétations qui semble s'imposer naturellement, mais qui sont en réalité totalement infondées.** [6].

6. Ceci est bien connu des débatteurs qui, face à un adversaire qui conteste leur argumentation, lancent le classique et intimidant «Contestez-vous ces chiffres ?». Penaud, l'adversaire est en général obligé d'admettre qu'il est d'accord avec les chiffres avancés, si bien que ce qui aurait dû être le point central de la discussion, à savoir que ce ne sont pas les chiffres qui sont contestés, mais la

Deux questions au moins se posent alors. Premièrement, comment peut-on faire la différence entre une dépendance traduisant réellement une relation de cause à effet, et une dépendance due à une cause commune cachée ou un facteur de confusion ? Deuxièmement, si l'on ne peut faire cette différence, le constat d'une dépendance peut-il néanmoins servir à quelque chose ?

Concernant la première question, notons qu'il est en principe possible d'évaluer le rôle d'une possible cause cachée ou d'un facteur de confusion éventuel en vérifiant si la dépendance observée entre événements continue d'exister lorsque l'on fixe la cause ou le facteur en question.

Tout d'abord, face à une dépendance constatée entre événements, il est nécessaire d'envisager les causes cachées ou les facteurs de confusion pouvant, de manière plausible, expliquer cette dépendance. Si l'on parvient à une suggestion raisonnable de cause cachée ou de facteur de confusion, on peut tenter d'évaluer le rôle de cette cause possible ou de ce facteur en vérifiant si la dépendance persiste lorsque l'on tient compte explicitement de la cause cachée ou du facteur de confusion suggéré en fixant Pour reprendre l'un des exemples précédents : les enfants P*** acceptent-ils toujours plus facilement de manger les repas préparés par leur père plutôt que par leur mère lorsque l'on se restreint aux situations ou le type de nourriture proposé par les parents est fixé (tous les deux des pâtes, ou tous les deux des épinards) ? Si c'est le cas, on ne peut mettre la dépendance observée seulement sur le compte de la cause ou du facteur envisagé. Dans les faits, la situation n'est pas si simple, car on ne dispose pas toujours des informations qui seraient nécessaires pour effectuer une telle vérification. D'autre part, il peut se révéler impossible en pratique de prendre en compte simultanément non pas une cause ou un facteur, mais un ensemble de causes et de facteurs susceptibles d'intervenir simultanément (ce qui supposerait, par exemple, de séparer les individus d'une population en groupes d'individus de même sexe, même âge, même type de lieu de résidence, même catégorie socio-professionnelle, mêmes antécédents de santé, etc...), car on ne disposera pas forcément des informations nécessaires ou de données en quantité suffisante ; des approches statistiques plus sophistiquées ont été développées pour tenter de traiter ce type de problème, généralement au prix d'hypothèses de modélisation supplémentaires, mais leur présentation dépasserait largement le cadre de ce cours. Enfin, on ne peut de cette manière tenir compte que des causes ou des facteurs explicitement suggérés ; or, notre perspicacité, ou notre compréhension du problème, peut parfaitement s'avérer insuffisante pour que nous puissions proposer une explication fondée sur une cause cachée ou un facteur de confusion, même si une telle explication existe. Une solution élégante à ce problème, qui n'est pas toujours praticable (par exemple, s'il s'agit de juger du caractère nocif d'un certain comportement, par exemple, on

manière de les interpréter, a de bonnes chance d'être totalement occulté.

ne peut évidemment pas forcer des individus à adopter ce comportement) est de pratiquer une expérimentation contrôlée randomisée (voir exercice 31).

Pour une introduction plus détaillée, mais non-technique, à ces questions, illustrée d'exemples issus du domaine médical, nous vous recommandons la lecture de l'excellent ouvrage de Schwartz cité dans la bibliographie.

Concernant la deuxième question, une dépendance probabiliste avérée constitue une information utile, qu'elle résulte ou non d'un lien de cause à effet, car elle suffit à définir ce que l'on nomme en épidémiologie des **facteurs de risque**, et peut ainsi servir de base à des décisions rationnelles à l'échelle de populations. Par exemple, même si aucun lien de cause à effet n'est mis en évidence entre le fait pour un individu d'avoir séjourné dans le pays U et de développer ultérieurement la maladie V, mais qu'une association positive est mise en évidence, on considérera le séjour dans le pays U comme un facteur de risque pour la maladie V, et, par exemple, on choisira d'administrer plus systématiquement un traitement préventif de la maladie V aux individus ayant séjourné dans le pays U, ou, tout au moins, ce facteur de risque interviendra de manière importante dans le calcul coût/bénéfice attendu d'un tel traitement. De plus, l'observation d'une telle association peut être l'indice d'un lien de cause à effet lié, au moins en partie, au fait de séjourner dans le pays U, et conduira a rechercher systématiquement l'origine de cette association, et éventuellement à la découverte d'une cause de la maladie V.

Une dernière remarque : nous avons discuté ci-dessus le fait qu'une dépendance entre événements pouvait ou non traduire une relation de cause à effet, mais il ne faut pas pour autant croire qu'une indépendance entre événements soit automatiquement le signe d'un rapport de cause à effet.

1.5.2 Effet de loupe et biais de sélection

On parle parfois d'**«effet de loupe» probabiliste**, pour insister sur le fait que le conditionnement par un événement A nous fait observer «à la loupe» cet événement (en particulier si celui-ci est de faible probabilité), puisque l'on ramène à 1 la probabilité de celui-ci en grossissant proportionnellement les probabilités des éventualités élémentaires qui le constituent, si bien que A joue en quelque sorte le rôle d'espace des possibles à lui tout seul. Il se peut donc que $\mathbb{P}(\cdot|A)$ se révèle très différente de $\mathbb{P}$, au moins pour le calcul de la probabilité d'un certain nombre d'événements.

Cet effet est connu en statistique sous le nom de biais de sélection. Il se manifeste par exemple lorsque l'on cherche à construire un modèle $(\Omega, \mathbb{P})$ décrivant une certaine population, mais que la population réellement atteinte par notre étude est une sous-population de $\mathcal{P}$ obtenue par une certaine forme de sélection, si bien que celle-ci serait adéquatement décrite par le modèle $(\Omega, \mathbb{P}(\cdot|A))$ et non pas $(\Omega, \mathbb{P})$. Si l'on n'est

pas conscient de cette différence entre la population que l'on cherche à étudier et celle que l'on étudie réellement, on sera amené à attribuer à $\mathbb{P}$ des propriétés qui sont en fait celles de $\mathbb{P}(\cdot|A)$, ce qui n'est pas vraiment souhaitable, en particulier si ces probabilités sont fortement distinctes ! Un exemple très simple de ce phénomène est constitué par les enquêtes statistiques dont les réponses sont obtenues sur la base du volontariat. Par exemple, un magazine adresse à ses lecteurs un questionnaire, mais seuls répondent ceux qui le souhaitent. Dans ce cas, la population réellement touchée par l'étude est constituée par les individus ayant souhaité et trouvé le temps d'y répondre, et, dans certains cas, il est parfaitement possible qu'il existe une dépendance entre les réponses aux questions posées et le fait de souhaiter et d'avoir le temps de répondre au questionnaire (par exemple, seuls les lecteurs se sentant particulièrement concernés par les questions posées répondront, et la répartition de leurs réponses peut donc différer de celle des réponses que fourniraient l'ensemble des lecteurs du magazine). De la même manière, la population des lecteurs du magazine forme une sous-population bien particulière de la population totale sont distinctes, et extrapoler les réponses de celle-ci à celle-là revient à ignorer la présence de la sélection. Un exemple historique de biais de sélection est le sondage du magazine Literary Digest qui, à l'occasion de l'élection présidentielle américaine de 1936, avait prévu la victoire du candidat républicain (Landon) contre le candidat démocrate (Roosevelt), sur la base d'une enquête postale portant sur plus de deux millions de personnes. C'est en fait Roosevelt qui fut élu. Pour ce qui nous intéresse de cette histoire, il faut noter que la liste des personnes sondées par le magazine avait été établie à partir d'une liste de ses lecteurs, de détenteurs d'automobiles, et d'usagers du téléphone, ce qui, à l'époque, représentait une forte sélection en faveur des couches aisées de la population, d'où évidemment un biais de sélection. Avec la confusion entre dépendance et causalité, la non-prise en compte d'un possible biais de sélection dans un argument statistique constitue l'une des pires erreurs qui se puissent commettre. La présence d'un biais de ce type n'est cependant pas toujours facile à déceler, celui-ci pouvant se manifester en amont (par exemple au moment de la collecte des données), ou en aval (après que celles-ci ont été collectées). Voir à ce sujet l'exercice 30.

Dans ce qui suit, nous donnons plusieurs exemples simples d'effet de loupe probabiliste.

Exemple : pourquoi votre alarme anti-intrusion se déclenche-t-elle la plupart du temps pour rien ?

Décrivons la situation à l'aide de l'espace des possibles suivant :

$$\Omega = \{\text{CA, CN, TA, TN}\},$$

où A signifie que l'alarme s'est déclenchée au moins une fois pendant vos vacances estivales, N qu'elle ne s'est pas déclenchée, C que des cambrioleurs ont effectivement tenté de s'introduire dans votre domicile, et T que personne n'a rien tenté de semblable (T pour tranquillité). Choisissons les probabilités de la façon suivante : la probabilité d'être victime d'un cambriolage pendant vos vacances est de 1% (nous négligerons la possibilité que deux cambriolages puissent se produire), la probabilité pour que l'alarme se déclenche sachant que des cambrioleurs sont présents (sensibilité) est de 99%, et la probabilité pour que l'alarme ne se déclenche pas en l'absence de cambrioleurs sans raison (spécificité) est de 95%. Ces informations nous permettent de spécifier complètement les probabilités affectées à chaque éventualité élémentaire, grâce à la formule de Bayes. Ainsi, la probabilité $\mathbb{P}(\mathrm{CA})$ n'est autre que la probabilité de l'intersection des deux événements A : «l'alarme se déclenche» et C : «les cambrioleurs sont là», qui, d'après la formule de Bayes, est égale à :

$$\mathbb{P}(A \cap C) = \mathbb{P}(A|C) \times \mathbb{P}(C) = 0,99 \times 0,01 = 0,0099.$$

De même,

$$\mathbb{P}(\mathrm{CN}) = \mathbb{P}(\bar{A} \cap C) = \mathbb{P}(\bar{A}|C) \times \mathbb{P}(C) = (1 - \mathbb{P}(A|C)) \times \mathbb{P}(C) = 0,01 \times 0,01 = 0,0001,$$

$$\mathbb{P}(\mathrm{TA}) = \mathbb{P}(A \cap \bar{C}) = \mathbb{P}(A|\bar{C}) \times \mathbb{P}(\bar{C}) = \mathbb{P}(A|\bar{C}) \times (1 - \mathbb{P}(C)) = 0,05 \times 0,99 = 0,0495,$$

$$\mathbb{P}(\mathrm{TN}) = \mathbb{P}(\bar{A} \cap \bar{C}) = \mathbb{P}(\bar{A}|\bar{C}) \times \mathbb{P}(\bar{C}) = (1 - \mathbb{P}(A|\bar{C})) \times (1 - \mathbb{P}(C)) = 0,95 \times 0,99 = 0,9505.$$

Votre alarme se déclenche... Quelle est la probabilité que ce soit pour rien ? Autrement dit, quelle est la probabilité conditionnelle de l'événement $\bar{C}$ sachant A ? Réponse :

$$\mathbb{P}(\bar{C}|A) = \frac{\mathbb{P}(\bar{C} \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(\mathrm{TA})}{\mathbb{P}(A)} = \frac{0,0495}{\mathbb{P}(A)}.$$

Or A est la réunion des deux éventualités élémentaires TA et CA, d'où :

$$\mathbb{P}(A) = \mathbb{P}(\mathrm{TA}) + \mathbb{P}(\mathrm{CA}) = 0,0495 + 0,0099 = 0,0594.$$

D'où :

$$\mathbb{P}(\bar{C}|A) = \frac{0,0495}{0,0594} = 0,8333...$$

Ainsi, avec une probabilité supérieure à 80%, un déclenchement de l'alarme ne correspond pas à une intrusion de cambrioleurs. La fiabilité du système d'alarme n'est pourtant pas en cause : malgré les apparences, qui pourraient nous faire conclure à sa piètre qualité, il fonctionne avec une probabilité de 99% en présence de cambrioleurs, et les déclenchements erronés ne surviennent qu'avec une probabilité de 5% en l'absence de cambrioleurs. C'est en fait la probabilité relativement faible, 1%, de subir un cambriolage, qui est responsable de la surreprésentation des fausses alarmes parmi les situations où l'alarme se déclenche.

On note l'effet de loupe : en conditionnant par le fait que l'alarme se déclenche, la probabilité d'un fonctionnement incorrect du système est fortement accrue par rapport à ce qu'elle est dans l'absolu (probabilité d'une fausse alarme ou d'un non-déclenchement en présence de cambrioleurs).

Exemple : pourquoi la file d'attente dans laquelle vous vous trouvez au supermarché avance-t-elle très souvent plus lentement que la file voisine ?

Construisons encore un modèle probabiliste très simple, dont l'espace des possibles est

$$\Omega = \{ \text{ S , L } \} \times \{ \text{ V , N } \} \times \{ \text{ S2 , L2 } \},$$

où S signifie que la file dans laquelle vous vous trouvez avance à vitesse satisfaisante, L qu'elle avance anormalement lentement (parce qu'un article a été mal étiqueté, parce que l'imprimante à tickets de caisse tombe en panne...), V signifie que vous vérifiez la vitesse de la file voisine pour confirmer votre infortune, N que vous ne vous intéressez pas à la file voisine, S2 que ladite file voisine avance à une vitesse que vous jugez satisfaisante, et R2 que celle-ci avance anormalement lentement. Choisissons les probabilités de la façon suivante : la probabilité pour que votre file avance lentement est égale à 20% ; si votre file avance rapidement, la probabilité pour que vous vous intéressiez à la vitesse de la file voisine est de 4% (vous n'avez aucune raison de vous y intéresser, et, en plus, vous n'en avez pas le temps car votre file avance rapidement...), mais elle est de 95% si votre file avance lentement (vous avez le temps de regarder autour de vous, et, en plus, vous cherchez une preuve du fait que, décidément, le sort s'acharne sur vous...). Par ailleurs, supposons que, sachant que votre file avance rapidement, ou pas, et que vous vous intéressiez à la file voisine, ou pas, la probabilité que la file voisine avance lentement est, indifféremment, égale à 20%, comme pour la vôtre. La question que nous posons est la suivante : sachant que vous observez la file voisine, quelle est la probabilité que celle-ci avance rapidement et la vôtre lentement ?

Ici encore, nous pourrions facilement calculer les probabilités associées à chacune des éventualités élémentaires. Nous n'en avons cependant pas besoin pour répondre à la question que nous nous posons. Appelons V l'événement «vous observez la file voisine», L l'événement «votre file avance lentement» et $S2$ l'événement «la file voisine avance rapidement». D'après la formule de Bayes,

$$\mathbb{P}(L \cap S2|V) = \frac{\mathbb{P}(L \cap S2 \cap V)}{\mathbb{P}(V)}.$$

D'après la formule de Bayes toujours,

$$\mathbb{P}(L \cap S2 \cap V) = \mathbb{P}(S2|L \cap V) \times \mathbb{P}(L \cap V) = \mathbb{P}(S2|L \cap V) \times \mathbb{P}(V|L) \times \mathbb{P}(L).$$

D'où, avec nos choix de probabilité :

$$\mathbb{P}(L \cap S2 \cap V) = 0,8 \times 0,95 \times 0,2 = 0,152.$$

D'autre part, en constatant que l'événement V s'écrit comme la réunion disjointe des deux événements $V \cap L$ et $V \cap \bar{L}$, nous obtenons que :

$$\mathbb{P}(V) = \mathbb{P}(V \cap L) + \mathbb{P}(V \cap \bar{L}) = \mathbb{P}(V|L) \times \mathbb{P}(L) + \mathbb{P}(V|\bar{L}) \times \mathbb{P}(\bar{L}).$$

Avec nos choix de probabilité :

$$\mathbb{P}(V) = 0,95 \times 0,2 + 0,04 \times 0,8 = 0,222.$$

Finalement, la probabilité conditionnelle recherchée $\mathbb{P}(L \cap S2|V)$ est égale à :

$$\mathbb{P}(L \cap S2|V) = \frac{0,152}{0,222} = 0,684...$$

Autrement dit, avec une probabilité de près de 70%, lorsque vous observez la file voisine, c'est pour constater (avec rage) qu'elle avance nettement plus vite que la vôtre. Ce résultat est à mettre au compte du fait que l'on se retourne très rarement quand sa file avance normalement, et très souvent quand ce n'est pas le cas. Les observations sont ici biaisées en faveur d'un mauvais fonctionnement des caisses, et l'on ne peut s'appuyer sur elles pour affirmer le mauvais fonctionnement global du système : encore un exemple de l'effet de loupe.

Exemple : pourquoi faut-il prendre avec précaution les résultats de tests de dépistage alarmants ?

Pour dépister une maladie, on effectue un test sanguin. Si le patient est effectivement atteint, le test donne un résultat positif avec une probabilité de 99% (sensibilité). Si le patient est sain, le test donne un résultat négatif (spécificité) avec une probabilité de 98%, mais peut donc malheureusement donner un résultat positif avec une probabilité de 2%. Nous supposerons que la probabilité d'être frappé par la maladie est de 0,1% pour un patient se présentant au dépistage (on peut imaginer qu'il s'agit d'un dépistage assez systématique, touchant une large fraction de la population). Sachant que le test donne un résultat positif, quelle est la probabilité que le patient soit effectivement malade ?

Comme précédemment, on construit un modèle probabiliste dont l'espace des possibles est

$$\Omega = \{\text{MP, MN, SP, SN}\},$$

où M désigne le fait que le patient soit malade, S le fait qu'il ne le soit pas, N le fait que le test soit négatif et P le fait qu'il soit positif. Appelons M l'événement «le patient est malade» et P l'événement «le test est positif». Nous cherchons donc la probabilité conditionnelle $\mathbb{P}(\bar{M}|P)$. Grâce à la formule de Bayes, on a :

$$\mathbb{P}(\bar{M}|P) = \frac{\mathbb{P}(\bar{M} \cap P)}{\mathbb{P}(P)} = \frac{\mathbb{P}(P|\bar{M}) \times \mathbb{P}(\bar{M})}{\mathbb{P}(P)} = \frac{0,02 \times 0,999}{\mathbb{P}(P)}.$$

En notant que P est la réunion des deux événements disjoints $P \cap M$ et $P \cap \bar{M}$, on obtient que :

$$\mathbb{P}(P) = \mathbb{P}(P \cap M) + \mathbb{P}(P \cap \bar{M}).$$

D'où, grâce à la formule de Bayes :

$$\mathbb{P}(P) = \mathbb{P}(P|M) \times \mathbb{P}(M) + \mathbb{P}(P|\bar{M}) \times \mathbb{P}(\bar{M}) = 0,99 \times 0,001 + 0,02 \times 0,999 = 0,02097.$$

Finalement, la probabilité conditionnelle recherchée est égale à :

$$\mathbb{P}(\bar{M}|P) = \frac{0,02 \times 0,999}{0,02097} = 0,95278...$$

Autrement dit, lorsque le test donne lieu à un résultat positif, il s'agit d'un «faux positif» avec une probabilité supérieure à 95%... Là encore, c'est la très faible incidence de la maladie dans la population subissant le dépistage qui fait que, malgré les performances apparemment honorables du test, celui-ci se révèle en pratique d'une fiabilité extrêmement réduite... Si seuls se présentaient au dépistage des patients probablement atteints de la maladie (par exemple, s'il s'agissait d'un test servant surtout à confirmer des soupçons bien étayés), la situation serait toute autre... On note que les faux positifs demeurent fort rares dans l'absolu (c'est-à-dire, non rapportés au nombre de positifs, vrais ou faux, mais à la totalité des tests effectués) : la plupart du temps, le test est négatif. De plus, lorsqu'il l'est, c'est la plupart du temps à juste titre, car la probabilité pour que le patient soit malade si le résultat du test est négatif, c'est-à-dire $\mathbb{P}(N|M)$ est de l'ordre de 10^{-5}.

L'effet de loupe entraîne encore ici une modification de la probabilité de fonctionnement correct du test.

Raisonnement bayésien

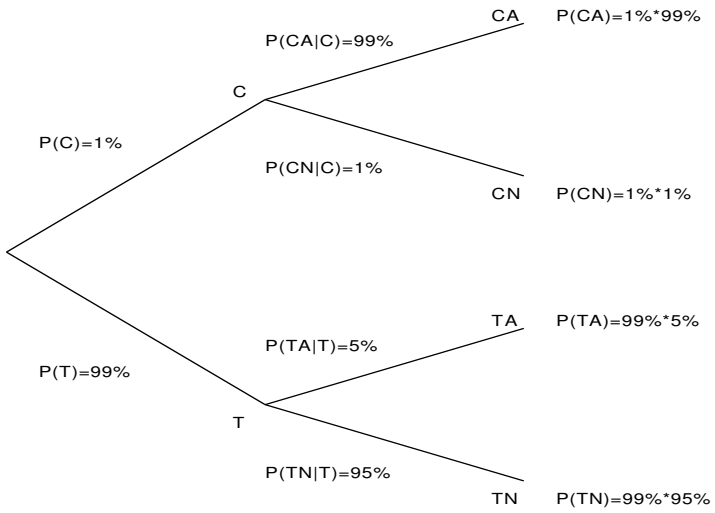
Les exemples précédents illustrent dans des situations très simples ce que l'on appelle communément le raisonnement bayésien, dans lequel on cherche à évaluer les probabilités d'événements pouvant apparaître comme des causes (la présence de cambrioleurs, le fait d'être malade) à partir de l'observation d'événements pouvant apparaître comme des effets (déclenchement de l'alarme, test positif). Il s'agit d'une démarche courante dans la pratique scientifique : évaluer à partir d'observations les probabilités de différentes hypothèses pouvant expliquer celles-ci. On notera qu'il est dans ce contexte nécessaire de disposer d'estimations *a priori* des probabilités relatives aux causes (fréquence des cambriolages, incidence de la maladie) et des probabilités des effets conditionnellement aux causes, sans quoi, le modèle ne peut être complètement spécifié, et l'on ne peut mener à bien ce type de raisonnement. On retient la démarche qui consiste à considérer un modèle général (qu'il est possible, dans nos exemples, de formuler facilement) puis à le conditionner par les événements

observés, afin d'évaluer les probabilités recherchées, qui sont donc des probabilités conditionnelles, plutôt que de chercher à évaluer directement celles-ci. Pour en apprendre beaucoup plus sur le raisonnement bayésien, vous pouvez consulter l'ouvrage de Howson et Urbach cité dans la bibliographie.

1.5.3 Représentation en arbre des modèles probabilistes

Dans cette partie, nous décrivons la struture commune à **tous** les modèles probabilistes qui apparaissent dans le cadre de ce cours, et qui est, en fait, commune à la plupart des modèles probabilistes discrets effectivement employés. Les probabilités conditionnelles y jouent un rôle fondamental, et il est indispensable de maîtriser complètement cette notion, ainsi que ce qui suit.

Les trois exemples (alarme, caisse, dépistage) qui précèdent illustrent l'utilisation des probabilités conditionnelles de deux manières au moins : d'abord pour tirer des conclusions dans le cadre d'un modèle probabiliste déjà construit, en tenant compte d'une information sur le déroulement de l'expérience, mais également, et de façon fondamentale, pour **construire** les modèles probabilistes employés. En effet, la plupart des modèles probabilistes (pour ne pas dire tous) que nous considérerons font intervenir, et de façon prépondérante, les probabilités conditionnelles dans leur construction, et les exemples qui précèdent illustrent cette règle : relisez-les, et vous constaterez qu'ils sont entièrement **formulés** en termes de probabilités conditionnelles. Les quantités pertinentes (probabilité pour que l'alarme se déclenche en présence d'un cambrioleur, probabilité pour que le test de dépistage échoue sur un individu malade,...) qui nous apparaissent naturellement comme les paramètres du modèle, susceptibles d'être évalués expérimentalement, sont des probabilités conditionnelles, et c'est elles qui nous permettent de définir la probabilité sur Ω ! En fait, tous les modèles probabilistes que nous considérerons sont construits à partir d'une structure séquentielle de choix (explicitement présente dans la situation considérée, ou posée par le modélisateur), qui sous-tend la représentation de la situation par les éléments de Ω. Sur cette structure séquentielle se greffent les probabilités conditionnelles qui permettent la spécification de la probabilité $\mathbb{P}$. Nous sommes ainsi amenés naturellement à représenter Ω à l'aide d'un arbre, dont les feuilles correspondent aux éléments de Ω, et aux arêtes duquel sont attachées des probabilités conditionnelles permettant d'obtenir la probabilité de n'importe quelle feuille en effectuant le produit des probabilités conditionnelles le long de la branche de l'arbre menant à cette feuille. C'est en particulier le cas des trois exemples donnés précédemment (relisez-les !), comme l'illustre pour le premier exemple le schéma ci-dessous, et nous allons dans ce qui suit donner une version générale de cette construction.



Notez que, dans notre traitement de cet exemple dans un précédent paragraphe, nous avons donné une description exhaustive de Ω, en fournissant simplement la liste de ses éléments :

$$\Omega = \{\text{CA, CN, TA, TN}\},$$

ce que l'on aurait pu écrire de manière équivalente

$$\Omega = \{\text{C, T}\} \times \{\text{A, N}\}.$$

Une autre possibilité aurait été de représenter Ω sous forme d'un tableau à double entrée (en utilisant qu'il n'y a en présence que deux éléments variables pris en compte dans le modèle : déclenchement ou non-déclenchement de sonnerie, présence ou absence de cambrioleurs), comme suit, chaque case du tableau représentant une éventualité élémentaire.

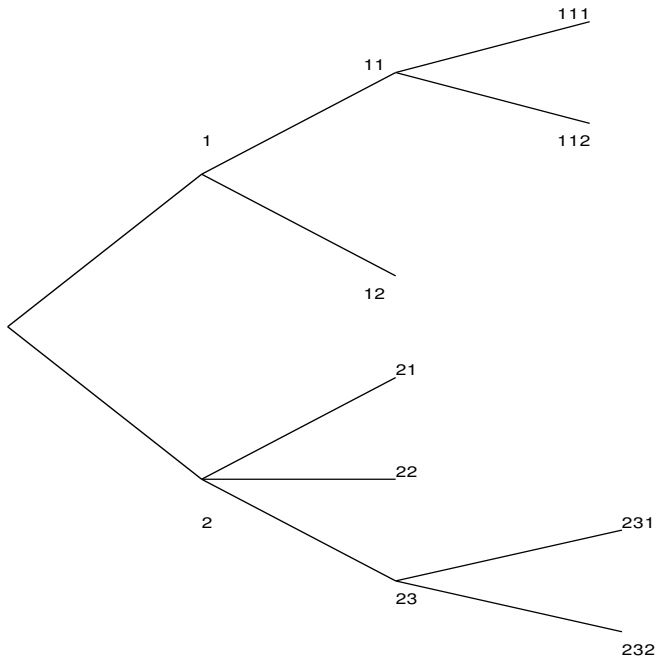
	C	T
A		
N		

Ces deux possibilités de représentation présentent un certain intérêt, mais nous leur préférerons souvent la représentation en arbre, qui s'impose naturellement dans de nombreuses situations.

De manière générale, lorsqu'une situation est décrite en termes de choix successifs qui déterminent progressivement l'issue réalisée (chaque choix comportant un nombre

fini ou dénombrable de possibilités), il est naturel de représenter Ω à l'aide d'un arbre enraciné. Le premier choix à effectuer donne lieu à une première ramification au niveau de la racine, et chaque nouveau choix à effectuer donne lieu à une ramification supplémentaire se greffant sur les précédentes. Chaque ramification comporte autant d'arêtes qu'il y a de possibilités différentes pour le choix correspondant, si bien que chaque arête de l'arbre s'identifie à la spécification d'un choix. Les k premières arêtes d'un chemin déterminent les décisions prises lors des k premiers choix, et un chemin complet menant de la racine à une feuille correspond à une spécification complète des différents choix, autrement dit, à une éventualité élémentaire du modèle.

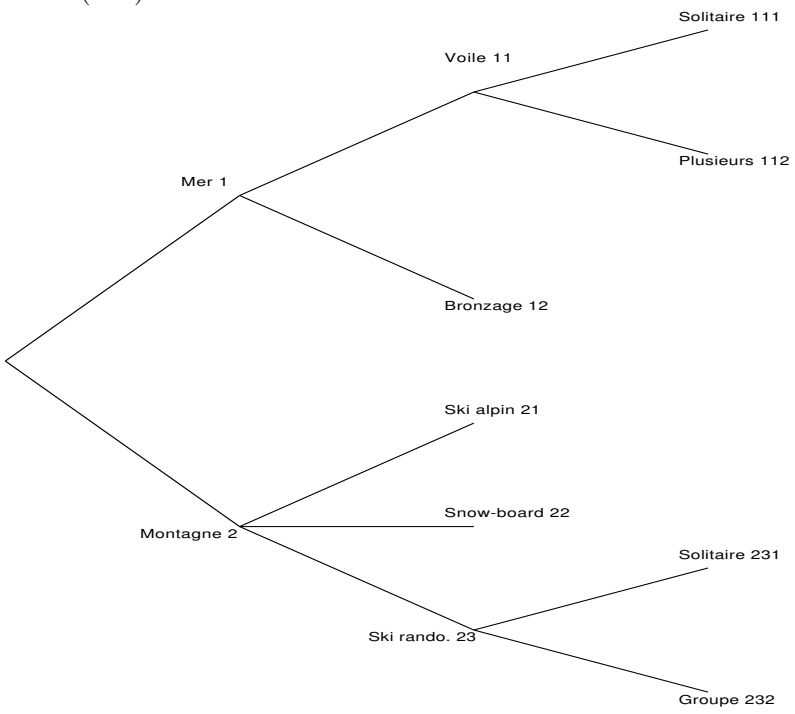
Par commodité, on choisira souvent de repérer les nœuds d'un arbre par des suites de symboles (par exemple des entiers). Chaque sommet du $k-$ème niveau de l'arbre (c'est-à-dire à distance k de la racine) sera numéroté par une suite de k nombres entiers permettant de la repérer, comme décrit sur la figure ci-dessous (la racine de l'arbre est notée r, et représente le niveau 0 de l'arbre).



Les éléments de Ω, c'est-à-dire les éventualités élémentaires, s'identifient aux feuilles de l'arbre (les sommets terminaux), ou, de manière équivalente, aux rayons reliant la racine aux feuilles. Chaque sommet intermédiaire (c'est-à-dire non-terminal) s'identifie à l'événement formé par toutes les éventualités élémentaires qui en descendent, ou encore au sous-arbre formé par ses descendants. Concrètement, s'il s'agit d'un sommet situé au $k-$ème niveau, cet événement s'identifie à la spécification de

l'information relative à la réalisation des k premières étapes de la séquence des choix par laquelle on décrit le phénomène. En d'autres termes, à un sommet de l'arbre est associé le sous-arbre formé par ses descendants, et l'ensemble des feuilles de ce constitue l'événement associé à ce sommet. On notera que, si tout sommet de l'arbre définit ainsi un événement (la racine étant, avec notre représentation, associée à Ω tout entier), tous les événements ne sont pas nécessairement associés à un sommet.

Par exemple, l'arbre ci-dessus pourra représenter l'espace des possibles associé à la description des activités de vacances d'un individu, structuré de la manière suivante : on peut avoir choisi soit la mer (1), soit la montagne. Si la mer a été choisie, on peut soit faire de la voile (11), soit passer son temps à bronzer sur la plage (12). Si l'on choisit la voile, on peut soit faire de la voile à plusieurs (111) soit en solitaire (112). Si l'on a plutôt choisi la montagne (1), on peut, soit faire du ski alpin (21), soit du snowboard (22), soit du ski de randonnée (23). Si l'on choisit de faire du ski de randonnée, on peut soit partir en randonnée seul (231), soit partir à plusieurs (232).



Au niveau de description que nous avons choisi (et qui n'est bien entendu pas le seul possible, il ne s'agit ici que d'un exemple assez rudimentaire, et pas né-

cessairement pertinent), les éventualités élémentaires sont 111,112,12,21,22,231,232. Dans notre description, 2 n'est pas une éventualité élémentaire, mais un événement : «avoir choisi la montagne», qui correspond formellement à toutes les éventualités élémentaires qui en descendent, soit 231 et 232. Bien entendu, des événements tels que $\{112, 231\}$ («voile en solitaire ou randonnée en solitaire») ne sont pas définis simplement par un sommet.

Abordons à présent la manière de spécifier la probabilité pour un modèle représenté par un arbre. Nous associerons à chaque arête $a_1 \dots a_{k-1} \to a_1 \dots a_{k-1}a_k$ la probabilité conditionnelle

$$\mathbb{P}(a_1 \dots a_k | a_1 \dots a_{k-1}).$$

(Si l'événement $a_1 \dots a_{k-1}$ est de probabilité nulle, on peut tout aussi bien l'éliminer du modèle, c'est-à-dire supprimer le sommet qui lui correspond ainsi que tous ses descendants. Aussi, nous supposerons que les événements associés aux différents sommet de l'arbre sont tous de probabilité non-nulle.) La connaissance de ces probabilités conditionnelles permet de calculer la probabilité de n'importe quelle éventualité élémentaire (c'est-à-dire de n'importe quelle feuille de l'arbre), en effectuant le produit des probabilités conditionnelles associées aux arêtes du chemin menant de la racine à la feuille en question. Plus formellement, ceci s'exprime à l'aide de l'égalité :

$$\mathbb{P}(a_1 \dots a_k) = \mathbb{P}(a_1 \dots a_k | a_1 \dots a_{k-1}) \times \mathbb{P}(a_1 \dots a_{k-1} | a_1 \dots a_{k-2}) \times \dots \times \mathbb{P}(a_1 a_2 | a_1) \times \mathbb{P}(a_1).$$

Avant tout commentaire, donnons la preuve (facile) de cette égalité : on vérifie tout d'abord que l'intersection de l'événement représenté par $(a_1 \dots a_{k-1})$ et de l'événement représenté par $(a_1 \dots a_{k-1}a_k)$ est égale à l'événement représenté par $(a_1 \dots a_{k-1}a_k)$ (ou, autrement dit, l'événement représenté par $(a_1 \dots a_{k-1}a_k)$ est inclus dans l'événement représenté par $(a_1 \dots a_{k-1}a_k)$). Ensuite, on se contente d'appliquer la définition des probabilités conditionnelles en tenant compte de la remarque que nous venons de faire : le produit ci-dessus se réécrit sous la forme

$$\frac{\mathbb{P}(a_1 \dots a_{k-1}a_k)}{\mathbb{P}(a_1 \dots a_{k-1})} \times \frac{\mathbb{P}(a_1 \dots a_{k-1})}{\mathbb{P}(a_1 \dots a_{k-2})} \times \dots \times \frac{\mathbb{P}(a_1 a_2)}{\mathbb{P}(a_1)} \times \mathbb{P}(a_1),$$

et tous les termes se simplifient deux-à-deux sauf le premier, d'où l'égalité souhaitée.

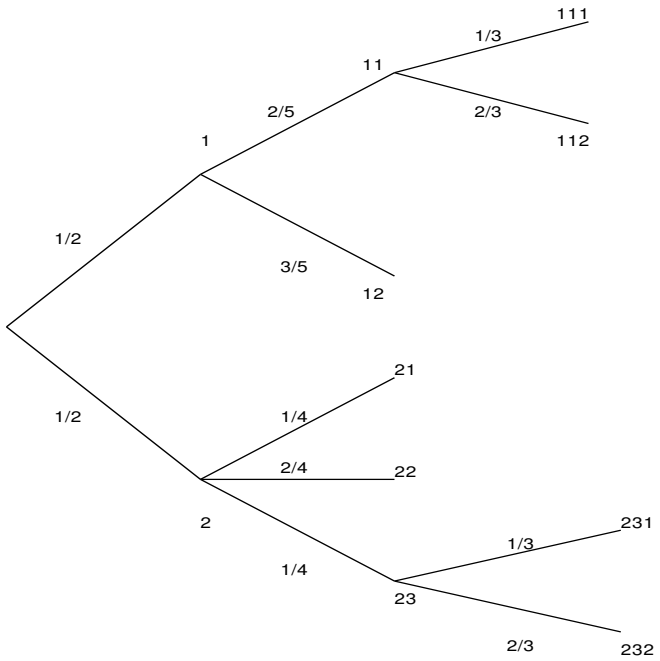
Conclusion : pour définir la probabilité $\mathbb{P}$ sur un espace des possibles représenté par un arbre, on peut soit donner la liste des probabilités associées à chacune des feuilles de l'arbre, soit donner, pour chaque arête de l'arbre, la probabilité conditionnelle qui lui est associée comme nous l'avons indiqué précédemment. Ces deux descriptions sont formellement équivalentes, mais le grand intérêt de la seconde réside dans le fait que les probabilités conditionnelles en question apparaissent souvent comme des quantités pertinentes, ayant un sens en tant que telles dans l'étude du phénomène que l'on cherche à modéliser, et que l'on peut facilement spécifier, au

contraire des probabilités finales associées directement aux éléments de Ω. Plutôt que de spécifier directement la valeur de $\mathbb{P}$ pour chacune des éventualités élémentaires, on spécifiera donc plutôt, pour chaque arête de l'arbre, la probabilité conditionnelle qui lui est associée, la probabilité d'une éventualité élémentaire se déduisant de ces probabilités conditionnelles par produit le long des branches de l'arbre, de la racine à l'extrémité représentant l'éventualité élémentaire en question. Les probabilités conditionnelles de la forme $\mathbb{P}(a_1 \dots a_\ell | a_1 \dots a_k)$, pour $\ell \geq k$, s'expriment également très facilement sous forme de produit des probabilités conditionnelles le long de la portion de chemin reliant dans l'arbre le sommet $a_1 \dots a_k$ au sommet $a_1 \dots a_\ell$. Plus exactement, la relation suivante est vérifiée, pour $k \leq \ell$:

$$\begin{array}{rcl} \mathbb{P}(a_1 \dots a_\ell | a_1 \dots a_k) & = & \\ \mathbb{P}(a_1 \dots a_{k+1} | a_1 \dots a_k) & \times & \mathbb{P}(a_1 \dots a_{k+2} | a_1 \dots a_{k+1}) \times \dots \times \mathbb{P}(a_1 \dots a_\ell | a_1 \dots a_{\ell-1}), \end{array}$$

ce que l'on vérifie aisément. De manière plus imagée, conditionner Ω par un événement de la forme $a_1 \dots a_k$ revient à considérer le modèle formé par le sous-arbre issu de $a_1 \dots a_k$ et dont les probabilités conditionnelles associées aux arêtes sont celles du modèle initial.

Les trois exemples présentés dans la section précédente s'insèrent naturellement dans ce cadre. Pour l'arbre que nous avons décrit ci-dessus, on pourrait par exemple spécifier la probabilité sur Ω de la façon suivante : on a une chance sur deux de choisir la mer, une chance sur deux de choisir la montagne. Si l'on a choisi la mer, on a deux chances sur cinq de choisir la voile, et trois chances sur cinq de choisir le bronzage. Si l'on choisit la voile, on a alors deux chances sur trois d'aller en solitaire, et une chance sur trois d'aller en groupe. Si c'est la montagne qui est choisie, on a alors une chance sur quatre de choisir le ski alpin, deux chances sur quatre de choisir le snow-board, et une chance sur quatre de choisir le ski de randonnée. Enfin, si le ski de randonnée est choisi, on a une chance sur trois de randonner en solitaire, et deux chances sur trois de le faire accompagné. En représentant les probabilités conditionnelles associées à chaque arête, on obtiendrait le schéma suivant :



Pour calculer la probabilité d'une éventualité élémentaire, qui est donc représentée par un sommet terminal, il suffit de calculer le produit des probabilités conditionnelles associées aux arêtes reliant la racine à ce sommet. Par exemple, $\mathbb{P}(111) = 1/2 \times 2/5 \times 1/3$, $\mathbb{P}(231) = 1/2 \times 1/4 \times 1/3$, $\mathbb{P}(12) = 1/2 \times 3/5$. La probabilité d'un événement peut ensuite être obtenue, conformément à la définition générale, en effectuant la somme des probabilités des éventualités élémentaires qui le constituent.

De manière imagée, cette manière de décrire un modèle probabiliste revient à le fabriquer en attachant à chaque sommet non-terminal $a = (a_1 \dots a_k)$ de l'arbre (y compris la racine) un modèle probabiliste $(\Omega_a, \mathbb{P}_a)$, dont les éléments correspondent aux sommets de l'arbre issus de a, la probabilité $\mathbb{P}_a$ associée à chaque élément de Ω_a donnant la probabilité conditionnelle sachant a des différentes sommets issus de a :

$$\mathbb{P}_{a_1 \dots a_k}(a_{k+1}) = \mathbb{P}(a_1 \dots a_k a_{k+1} | a_1 \dots a_k).$$

Nous disposons ainsi d'un moyen d'assembler entre eux des modèles probabilistes simples (par exemple, un tirage uniforme parmi un nombre fini d'éléments) pour en fabriquer de plus élaborés, et c'est toujours ainsi que nous fabriquerons nos modèles. Ainsi, on peut voir l'exemple de modèle décrit ci-dessus comme l'assemblage des

modèles suivants :

$$\begin{cases} \Omega_r = \{1,2\},\ \mathbb{P}_r(1) = 1/2,\ \mathbb{P}_r(2) = 1/2 \\ \Omega_1 = \{1,2\},\ \mathbb{P}_1(1) = 2/5,\ \mathbb{P}_1(2) = 3/5 \\ \Omega_2 = \{1,2,3\},\ \mathbb{P}_2(1) = 1/4,\ \mathbb{P}_2(2) = 2/4,\ \mathbb{P}_2(3) = 1/4 \\ \Omega_{11} = \{1,2\},\ \mathbb{P}_{11}(1) = 1/3,\ \mathbb{P}_{11}(2) = 2/3 \\ \Omega_{23} = \{1,2\},\ \mathbb{P}_{23}(1) = 1/3,\ \mathbb{P}_{11}(2) = 2/3 \end{cases}$$

Bien entendu, rien ne nous oblige à définir les Ω_a à l'aide d'entiers : nous ne les avons présentés ainsi que pour être disposer d'une indexation des sommets de l'arbre par des suites d'entiers $a_1 \dots a_k$. On pourrait aussi bien avoir, par exemple (et de manière plus explicite)

$$\begin{cases} \Omega_r = \{\text{mer,montagne}\}, \mathbb{P}_r(\text{mer}) = 1/2,\ \mathbb{P}_r(\text{montagne}) = 1/2, \\ \Omega_{\text{mer}} = \{\text{voile,bronzage}\},\ \mathbb{P}_{\text{mer}}(\text{voile}) = 2/5,\ \mathbb{P}_{\text{mer}}(\text{bronzage}) = 3/5, \\ \Omega_{\text{montagne}} = \{\text{ski alpin, snow-board, ski rando.}\}, \\ \mathbb{P}_{\text{montagne}}(\text{ski alpin}) = 1/4,\ \mathbb{P}_{\text{montagne}}(\text{snowboard}) = 2/4,\ \mathbb{P}_{\text{montagne}}(\text{ski rando.}) = 1/4, \\ \Omega_{\text{mer voile}} = \{\text{solitaire, plusieurs}\},\ \mathbb{P}_{\text{mer voile}}(\text{solitaire}) = 1/3,\ \mathbb{P}_{\text{mer voile}}(\text{plusieurs}) = 2/3, \\ \Omega_{\text{montagne ski rando.}} = \{\text{solitaire, groupe}\}, \\ \mathbb{P}_{\text{montagne ski rando.}}(\text{solitaire}) = 1/3,\ \mathbb{P}_{\text{montagne ski rando.}}(\text{groupe}) = 2/3, \end{cases}$$

Ainsi, un sommet de l'arbre situé à la profondeur k pourra être repéré par une suite de symboles de la forme $b_1 \dots b_k$, chaque élément b_i étant un élément de l'espace des possibles $\Omega_{b_1 \dots b_{i-1}}$ associé au sommet $b_1 \dots b_{i-1}$ (avec toujours la convention selon laquelle $b_1 \dots b_{i-1}$ désigne la racine r lorsque $i = 1$).

Représentation en arbre et systèmes complets d'événements

Nous aurons parfois à considérer des systèmes complets d'événements qui sont naturellement associés à la représentation en arbre des modèles probabilistes : ceux constitués par des événements associés à des nœuds de l'arbre ou encore aux sous-arbres issus de ces nœuds. Pour un tel système d'événements, chaque rayon issu de la racine de l'arbre rencontre nécessairement un et un seul des sommets associés au système complet.

Donnons maintenant quelques exemples simples de modèles en arbre.

Exemple : répétitions indépendantes d'un tirage uniforme

Considérons un modèle probabiliste décrivant le tirage uniforme d'un objet parmi m :

$$\Omega_1 = \{h_1, \dots, h_m\},\ \mathbb{P}_1(h_1) = \mathbb{P}_1(h_m) = 1/m,$$

chaque objet a la même probabilité $1/m$ d'être choisi. On fabrique un modèle probabiliste $(\Omega_1^n, \mathbb{P}_1^{\otimes n})$ décrivant la répétition indépendante de n tirages uniformes en

associant à la racine, selon le procédé que nous venons de décrire, un exemplaire du modèle $(\Omega_1, \mathbb{P}_1)$, puis, récursivement, en associant à chaque sommet de niveau k un nouvel exemplaire du modèle $(\Omega_1, \mathbb{P}_1)$. Autrement dit, pour chaque sommet $a_1 \dots a_k$, $\Omega_{a_1 \dots a_{k-1}} = \Omega$, et les probabilités conditionnelles sur les arêtes sont définies par

$$\mathbb{P}_1^{\otimes n}(a_1 \dots a_k | a_1 \dots a_{k-1}) = \mathbb{P}_1^{\otimes n}(a_k) = 1/n,$$

moyennant l'identification des éléments $h_1, \dots, h_n$ de Ω aux sommets de l'arbre issus de $a_1 \dots a_{k-1}$. On obtient ainsi un arbre régulier $m-$aire de profondeur n, dont toutes les feuilles possèdent la même probabilité $1/m^n$, c'est-à-dire que $\mathbb{P}_1^{\otimes n}$ est la probabilité uniforme sur Ω_1^n : chaque $n-$uplet de tirages $(z_1, \dots, z_n) \in \Omega^n$ a la même probabilité d'être obtenu.

Ce modèle rend bien compte d'une succession **indépendante** de tirages, car on le définit en posant que, conditionnellement aux résultats des k premiers tirages (c'est-à-dire conditionnellement à $a_1 \dots a_k$), la probabilité d'obtenir l'un quelconque des éléments de Ω au $k+1-$ème tirage est encore uniforme. Attention : il ne s'agit pas de répéter n fois le **même** tirage, au sens où l'on obtiendrait n fois le même objet. Ce que l'on répète, c'est l'expérience consistant à effectuer le tirage, et l'on obtient en général des résultats différents d'un tirage à l'autre.

Bien entendu, cette construction ne se limite pas au cas d'un tirage uniforme, et n'importe quel modèle probabiliste $(\Omega, \mathbb{P})$ pourrait être «répété» de la sorte. Il n'est même pas nécessaire que ce soit le même modèle qui apparaisse à chaque tirage, et nous définirons la notion de succession indépendante d'expériences aléatoires décrites par des modèles distincts $(\Omega_1, \mathbb{P}_1), \dots, (\Omega_n, \mathbb{P}_n)$. Cette notion de succession indépendante d'expériences aléatoires est si importante que nous y reviendrons en grand détail ultérieurement, et en particulier sur le sens précis qu'il faut donner à l'indépendance des n expériences décrites par le modèle. Au passage, notez que, même si le modèle $(\Omega^n, \mathbb{P}^{\otimes n})$ rend compte de la répétition de n tirages, il y a dans l'arbre $1 + m + \dots + m^{n-1}$, et non pas n exemplaires du modèle $(\Omega, \mathbb{P})$.

Exemple : tirages uniformes successifs sans remise

On peut également modéliser par un arbre des tirages successifs mais non-indépendants cette fois. Un exemple simple est la situation où chaque tirage supprime l'objet qui vient d'être tiré des possibilités de tirages ultérieurs (d'où le nom), chaque objet étant tiré uniformément parmi les objets restants. Cette fois, l'arbre définissant Ω est un chouia plus difficile à décrire que dans le cas précédent. Numérotons les objets susceptibles d'êtres tirés par les entiers de 1 à m. On fabrique le modèle probabiliste $(\Omega_{\text{SR}}^n, \mathbb{P}_{\text{SR}}^n)$ décrivant m tirages uniformes sans remises successifs ($n \geq m$) en associant d'abord à la racine le modèle $\Omega_r = \{1, \dots, n\}$ muni de la probabilité uniforme $\mathbb{P}_r$, chaque entier $1, \dots, n$ représentant le numéro de l'objet choisi, puis, récursi-

vement, en associant à chaque sommet numéroté $a_1 \dots a_k$ de niveau k le modèle $\Omega_{a_1 \dots a_k} = \{1, \dots, n\} - \{a_1, \dots, a_k\}$ muni de la probabilité uniforme : ce modèle repose sur l'hypothèse selon laquelle, une fois les k premiers objets tirés, l'objet choisi au $k+1-$ème tirage est tiré uniformément parmi les objets restant. L'espace $\Omega_{a_1 \dots a_k}$ comporte $n - k$ éléments, et, finalement, en effectuant le produit des probabilités conditionnelles le long des branches de l'arbre, on constate que feuille de l'arbre se voit attribuer une probabilité égale à

$$\frac{1}{m} \times \frac{1}{m-1} \times \cdots \times \frac{1}{m-n+1}.$$

Comme dans le modèle précédent, la probabilité sur Ω_{sr}^n est donc la probabilité uniforme, le nombre d'éléments de l'espace des possibles étant cette fois égal à $m(m-1) \times \cdots \times (m-n+1)$. Ici encore, la spécification de la probabilité à l'aide de la structure d'arbre est très naturelle : conditionnellement à la liste d'objets déjà tirés, la probabilité de tirer l'un quelconque des objets restants est uniforme parmi l'ensemble des objets restants.

Une propriété intéressante de ce modèle est son **échangeabilité**. Celle-ci signifie que, si σ est une permutation quelconque des entiers de 1 à n, l'arbre obtenu en indiquant au $i-$ème niveau le tirage du $\sigma(i)-$ème objet (dans la présentation ci-dessus, nous avions $\sigma(i) = i$ car le i-ème niveau de l'arbre représentait le $i-$ème tirage), est le même que celui décrit ci-dessus, avec les mêmes ramifications et, surtout, les mêmes probabilités conditionnelles associées aux arêtes. Voir l'exercice 74. Une telle propriété est également valable pour le modèle de tirages uniformes répétés indépendamment, décrit précédemment (voir plus bas la discussion sur la succession d'épreuves indépendantes).

Retour sur un exemple précédent

Nous avons décrit plus haut l'exemple :

$$\Omega = \{0,1\}^{16} = \{(x_1, \dots, x_{16}) \ : \ x_i \in \{0,1\}\},$$

la probabilité $\mathbb{P}$ sur Ω étant définie par :

$$\mathbb{P}\left[(x_1, \dots, x_{16})\right] = \prod_{i=1}^{16} p^{x_i}(1-p)^{1-x_i},$$

où $p \in [0,1]$ est un paramètre. Celui-ci peut naturellement se réécrire comme un modèle en arbre fabriqué à partir de copies du modèle

$$\Omega^* = \{0,1\}, \mathbb{P}^*(1) = p, \ \mathbb{P}^*(0) = 1-p.$$

A la racine, on associe le modèle $(\Omega^*, \mathbb{P}^*)$, et, récursivement, à tout sommet non-terminal $a_1 \dots a_k$ on associe encore le modèle $(\Omega^*, \mathbb{P}^*)$. On vérifie bien que la probabilité $\mathbb{P}$ sur Ω définie plus haut coïncide avec celle que l'on obtient en effectuant les produits de probabilités conditionnelles le long des arêtes. Nous verrons un peu plus bas que ce modèle traduit l'hypothèse selon laquelle les événements correspondant au fonctionnement des différentes connexions sont globalement indépendants, ou, en termes plus imagés, selon laquelle les différentes connexions fonctionnent (ou tombent en panne) indépendamment les unes des autres, et ont individuellement chacune une probabilité p de fonctionner.

1.6 Construire un modèle approprié

1.6.1 Quelques pistes

La modélisation est, en général, un processus évolutif, résultant d'un dialogue complexe entre connaissances acquises, données recueillies, et hypothèses plus ou moins bien étayées. Notre discussion précédente sur la traduction concrète de la probabilité devrait vous permettre de saisir, en gros, ce que signifie pour une situation le fait d'être décrite de manière satisfaisante par un modèle probabiliste donné. Rappelons seulement que l'utilisation de la notion de probabilité ne va pas sans de multiples hypothèses, souvent implicites, sur la nature de la situation considérée et le contexte dans lequel elle se situe, et que le sens que prend la notion de probabilité dans un modèle affecte les conclusions qui en sont tirées.

De manière générale, le choix d'un modèle probabiliste pour décrire une situation doit au moins obéir aux deux contraintes antagonistes suivantes :

- l'espace des possibles Ω doit donner du phénomène une description suffisament fine pour que les événements concrets intéressants correspondent à des événements formels du modèle,
- il doit être possible d'identifier $\mathbb{P}$, et d'évaluer la probabilité des événements intéressants (et par conséquent l'espace des possibles ne doit pas être trop complexe),

mais cela ne suffit pas en général à déterminer $(\Omega, \mathbb{P})$ de manière unique, loin de là. Dans la plupart des exemples que nous envisagerons, cependant, la structure des phénomènes abordés sera assez simple, et fera assez clairement apparaître à la fois les quantités pertinentes dans la description du phénomène, et les hypothèses de modélisation qu'il est raisonnable de formuler, en première approximation. Même dans ce cadre limité, il n'est pas toujours évident de déterminer $(\Omega, \mathbb{P})$, et les remarques qui suivent ont pour but de vous guider dans cette direction. La représentation en arbre fait apparaître le problème de la détermination du modèle sous une forme assez satisfaisante conceptuellement :

- la détermination des éléments de variabilité pertinents, ceux que l'on choisit de décrire explicitement dans le modèle, aindi que d'un ordre de succession de ces éléments, fournit la structure de l'arbre ; chaque élément de variabilité explicitement pris en compte dans le modèle donne lieu à des ramifications correspondant aux différentes valeurs qu'il peut prendre,
- la structure de l'arbre étant fixée, il faut déterminer les probabilités conditionnelles spécifiant $\mathbb{P}$, la plupart du temps en émettant des hypothèses simplificatrices (indépendance, ou forme simple de dépendance), qui déterminent la forme de $\mathbb{P}$.

Enfin, et nous n'aborderons pas cet aspect en détail pour l'instant, quoiqu'il soit absolument crucial, il est nécessaire d'évaluer les différents paramètres (la plupart du temps au moyen de données expérimentales, ou, en leur absence, en formulant (encore) des hypothèses plausibles à leur sujet), et de tester la validité du modèle et des hypothèses sur lesquelles il repose, en le confrontant à des données expérimentales ou à toute information dont on dispose sur la situation étudiée. Notons que l'on cherchera systématiquement à limiter le nombre de paramètres mis en jeu dans le modèle, afin de lui conserver une certaine simplicité, mais surtout pour nous donner la possibilité d'évaluer correctement ces paramètres sur la base des données dont nous disposerons.

Donnons maintenant une recommandation générale concernant le choix de Ω (que nous représenterons toujours sous forme d'arbre) : les hypothèses que l'on formule sur le modèle doivent permettre la détermination **directe** des probabilités conditionnelles associées aux arêtes de l'arbre. Si l'arbre que vous choisissez pour décrire l'espace des possibles ne fait pas apparaître explicitement les éléments de variabilité de la situation relativement auxquels les hypothèses de modélisation sont formulées, la détermination de $\mathbb{P}$ risque de se transformer en un exercice long et périlleux. En ce sens, il est difficile de dissocier complètement la détermination de Ω de celle de $\mathbb{P}$, puisqu'il doivent tous deux refléter les hypothèses que nous souhaitons formuler.

Par exemple, pour aborder la modélisation du fonctionnement du réseau de communication que nous avons décrit plus haut, il serait *a priori* tout aussi pertinent, puisque tout ce qui nous intéresse en définitive est le fait que l'information puisse circuler de (S) vers (B), d'employer l'espace des possibles à deux éléments

$$\Omega_1 = \{\text{l'information circule entre (S) et (B)}, \text{l'information ne circule pas entre (S) et (B)}\},$$

plutôt que l'espace des possibles

$$\Omega_2 = \{0,1\}^{16} = \{(x_1, \ldots, x_{16}) \ : \ x_i \in \{0,1\}\},$$

que nous avons déjà décrit. Cependant, puisque nous connaissons la structure du réseau, il semble plus efficace de décrire le fonctionnement du réseau en termes du

fonctionnement de chacune des 16 connexions, ce qui nous permet de formuler l'hypothèse selon laquelle les différentes connexions fonctionnent (ou tombent en panne) indépendamment les unes des autres et ont individuellement chacune une probabilité p de fonctionner, et de déduire directement la forme déjà indiquée pour $\mathbb{P}_2$:

$$\mathbb{P}\left[(x_1, \ldots, x_{16})\right] = \prod_{i=1}^{16} p^{x_i}(1-p)^{1-x_i}.$$

Si cette hypothèse est vérifiée, (il s'agira alors d'une information !) et si nous pouvons évaluer p (par exemple, à l'aide de données concernant d'autres connexions du même type), nous pourrons déduire de $(\Omega_2, \mathbb{P}_2)$ la probabilité de fonctionnement du système, c'est-à-dire la probabilité pour que l'information puisse circuler de (S) vers (B). En revanche, Ω_1 ne permet pas directement d'utiliser ces informations pour calculer la probabilité $\mathbb{P}_1$, et le détour par $(\Omega_2, \mathbb{P}_2)$, même implicite, est indispensable. Notre recommandation est alors de choisir directement et sans hésiter $(\Omega_2, \mathbb{P}_2)$.

1.6.2 Compatibilité de deux modèles

Si vous êtes abîmé de perplexité à l'idée que le choix de $(\Omega, \mathbb{P})$ n'est pas automatique, et rongé d'inquiétude en pensant que vous ne parviendrez pas à trouver le «bon» modèle $(\Omega, \mathbb{P})$, il n'y a cependant pas lieu de vous inquiéter : cette recommandation est simplement destinée à vous guider, et son intérêt est d'autoriser la détermination directe de Ω, de forcer les hypothèses de modélisation à apparaître explicitement (ce qui permet souvent de constater que certaines d'entre elles sont incorrectes, et de les corriger), et donc de minimiser les risques d'erreur lors de cette étape fondamentale qu'est la détermination du modèle (et qui précède son exploitation). Plusieurs modèles différents peuvent parfaitement résulter des mêmes hypothèses de modélisation, et donner lieu (heureusement) aux mêmes conclusions concernant les événements intéressants.

D'une manière générale, nous définirons la compatibilité entre deux modèles probabilistes d'une même situation de la façon suivante : $(\Omega_1, \mathbb{P}_1)$ et $(\Omega_2, \mathbb{P}_2)$ sont compatibles lorsque, pour tout événement concret A associé dans chacun des deux modèles à un événement formel, disons $A_1 \subset \Omega_1$ et $A_2 \subset \Omega_2$, on a $\mathbb{P}_1(A_1) = \mathbb{P}_2(A_2)$.

Bien entendu, en général, un événement concret A peut ne pas définir d'événement formel dans l'un ou l'autre des modèles, voire dans les deux, s'ils ne sont pas d'une finesse suffisante. Cependant, si, comme il est évidemment nécessaire, les événements intéressants relatifs au phénomène étudié correspondent toujours à des événements formels (sinon, à quoi le modèle sert-il ?), la compatibilité de deux modèles d'une même situation entraîne le fait qu'ils attribuent la même probabilité aux événements intéressants [7].

7. Notre définition de la compatibilité entre modèles n'est pas la seule possible. En effet, un

Notre recommandation conduit souvent à choisir un modèle probabiliste plus fin qu'il n'est *a priori* nécessaire pour que les événements auxquels on s'intéresse apparaissent formellement dans le modèle, en contrepartie des multiples avantages que nous avons cités.

Souvent, pour une même situation, nous serons en présence d'un modèle de référence $(\Omega_1, \mathbb{P}_1)$, correspondant à la traduction directe des hypothèses formulées sur le modèle, mais nous serons amenés à raisonner de manière ponctuelle sur un modèle $(\Omega_2, \mathbb{P}_2)$ compatible avec $(\Omega_1, \mathbb{P}_1)$ et dans lequel Ω_2 est moins fin que Ω_1, nous permettant de nous concentrer sur un aspect spécifique de la situation considérée, et/ou de mener les calculs de manière plus simple et plus directe que ne le permettrait l'utilisation de $(\Omega_1, \mathbb{P}_1)$.

Dans le cas des modèles décrits par des arbres, une manière naturelle de procéder est d'élaguer l'arbre en éliminant tous les descendants d'un sommet donné, ce sommet devenant alors une feuille de l'arbre élagué, ou, en d'autres termes, une éventualité élémentaire d'un nouveau modèle. Si l'on conserve les même probabilités le long des arêtes, on obtient un modèle moins fin et compatible avec le précédent.

1.6.3 De l'importance de décrire explicitement le modèle

L'un des objectifs principaux de ce cours est de vous rendre entièrement naturelle la démarche consistant, face à une situation incorporant de l'incertitude, à tenter de l'aborder systématiquement au moyen d'une modélisation probabiliste. Dans le cadre limité de ce cours, cet objectif se traduira par le fait que la première étape de l'abord d'un problème consistera toujours à préciser la forme de l'espace des possibles Ω et de la probabilité $\mathbb{P}$ sur Ω. Afin de dissiper les derniers doutes qui pourraient subsister quant à la pertinence de cette démarche, qui peut souvent apparaître, au premier abord, comme inutilement lourde et contraignante, en particulier au vu de la relative simplicité des exemples traités, voici une petite liste d'arguments (solidement) étayés en sa faveur.

La démarche que nous vous proposons d'adopter a l'avantage d'être **systématique**, et de s'adapter aussi bien à des situations simples qu'à d'autres plus complexes, dont le bon sens seul ne suffit pas pour appréhender correctement la structure. Même dans le cas des exemples relativement simples abordés en TD, les limites d'une approche intuitive et non-formalisée des problèmes apparaissent. L'approche systématique des problèmes, que nous vous incitons à pratiquer, permet de préciser clairement les données objectives relatives au phénomène considéré, les hypothèses

modèle probabiliste ne permet pas seulement de calculer les probabilités des événements ayant une traduction formelle dans le modèle, mais peut également fournir des inégalités portant sur des événements concrets qui, sans posséder de traduction dans le modèle, impliquent, ou sont impliqués par, de tels événements concrets. Tenir compte de ce fait conduit par exemple à donner une définition différente de la compatibilité entre modèles, que nous n'aurons pas l'occasion d'utiliser.

de modélisation qu'il est possible ou souhaitable de formuler, la nature des questions qu'il est possible d'aborder dans le cadre de cette modélisation ainsi que la manière de les résoudre. Il devient ainsi possible de critiquer la modélisation effectuée et de mieux en cerner **la portée et les limites de validité**.

L'exigence de **précision** que suppose une telle démarche, outre le fait qu'elle est indispensable pour garantir la validité de votre approche, permet également de la **communiquer** à d'autres, et elle peut ainsi être évaluée, critiquée, confrontée à d'autres approches et finalement exploitée. L'idéal vers lequel il faut tendre dans la présentation de la modélisation d'un phénomène aléatoire est à rapprocher d'un code informatique correct, commenté et documenté. L'expression s'y plie à une norme stricte, les différents éléments qui interviennent sont explicitement définis, chaque étape est justifiée, et le fonctionnement global est également décrit. D'ailleurs, nous verrons ultérieurement qu'un modèle convenablement décrit doit permettre, au moins en principe, une transcription facile sous forme de code informatique permettant de le simuler. Comme un code informatique, la description d'un modèle ou son utilisation peut être entachée d'incohérences (erreurs de syntaxe, qui, dans notre contexte, ne peuvent malheureusement pas être débusquées par le compilateur), et un modèle formellement correct peut fournir des résultats erronés s'il se fonde sur une analyse incorrecte de la situation (un programme qui s'éxécute ne fait malheureusement pas toujours ce que le cahier des charges lui imposait de faire, si des erreurs de conception ou d'implémentation ont été commises.) Bien entendu, toute analogie a ses limites...

Enfin, satisfaire ces (multiples) exigences ne demandera pas, la plupart du temps, un effort surhumain de votre part, car la plupart des modèles que nous utiliserons seront construits de manière presque automatique à partir d'hypothèses standards sur les situations modélisées.

1.7 Un exemple fondamental : la succession d'épreuves indépendantes

La représentation en arbre nous permet facilement d'assembler entre eux des modèles probabilistes simples pour en fabriquer de plus complexes. Nous allons étudier plus en détail l'une des manières d'assembler entre eux des modèles, qui présente une importance fondamentale dans le cadre de la modélisation, et dont nous avons déjà rencontré quelques exemples auparavant : la succession d'épreuves indépendantes. La problématique est la suivante : nous disposons de n modèles probabilistes $(\Omega_i, \mathbb{P}_i),\ i = 1, \ldots, n$, chacun décrivant un phénomène (une «épreuve») particulière, et nous souhaitons fabriquer un modèle probabiliste rendant compte de la succession **indépendante** des épreuves décrites par chacun des modèles $(\Omega_i, \mathbb{P}_i)$. Notez bien qu'il peut s'agir d'une succession au sens chronologique du terme, chaque épreuve

ayant concrètement lieu l'une après l'autre, que d'une succession aussi bien que d'une succession supposée, les épreuves pouvant aussi bien avoir lieu simultanément que dans un ordre chronologique complètement différent de celui suggéré par la numérotation $1, \ldots, n$.

Comme nous l'avons déjà suggéré sur des exemples dans les parties précédentes, on peut décrire cette succession à l'aide du modèle en arbre suivant, défini récursivement : à la racine, on associe le modèle $(\Omega_1, \mathbb{P}_1)$, et, récursivement, au sommet $(a_1 \ldots a_k),\ k \leq n-1$, on associe le modèle $(\Omega_k, \mathbb{P}_k)$. Autrement dit, les probabilités conditionnelles sont définies par :

$$\mathbb{P}(a_1 \ldots a_{k+1} | a_1 \ldots a_k) = \mathbb{P}_k(a_k).$$

Dans l'égalité ci-dessus, a_{k+1} représente une issue de l'épreuve numéro $k+1$, c'est-à-dire un élément de Ω_{k+1}, et la dénomination de succession **indépendante** est justifiée par le fait que, conditionnellement aux réalisations des k premières expériences (représentées par $(a_1 \ldots a_k)$), la probabilité d'obtenir a_{k+1} lors de la $k+1-$ème expérience est égale à $\mathbb{P}_{k+1}(a_{k+1})$, c'est-à-dire la probabilité d'obtenir a_{k+1} dans le modèle $(\Omega_{k+1}, \mathbb{P}_{k+1})$ qui décrit individuellement l'épreuve numéro $k+1$. Autrement dit, la connaissance des réalisations des k premières épreuves ne modifie pas la probabilité $\mathbb{P}_{k+1}$ décrivant individuellement la réalisation de la $k+1-$ème. Notez que cette définition des probabilités conditionnelles est la seule possible si l'on veut traduire l'indépendance des expériences les unes vis-à-vis des autres. Nous allons à présent décrire quelques propriétés de ce modèle, qui, quoiqu'assez évidentes intuitivement, méritent tout de même d'être formulées précisément et prouvées. Nous pourrons ainsi préciser la notion d'indépdance mutuelle sous-jacente au modèle (et ce sera également l'occasion de nous entraîner un peu à la manipulation de ce type de modèle en arbre).

On note tout d'abord que, pour utiliser la notation mathématique courante, l'espace des possibles défini précédemment par sa représentation en arbre s'identifie au produit cartésien :

$$\Omega_1 \times \cdots \times \Omega_n.$$

De plus, la probabilité sur Ω définie par la réprésentation en arbre ci-dessus, que nous noterons $\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$, peut s'exprimer explicitement sous la forme (qui justifie la notation) :

$$\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n(a_1 \ldots a_n) = \mathbb{P}_1(a_1) \times \cdots \times \mathbb{P}_n(a_n),$$

$a_1 \ldots a_n$ correspondant à l'éventualité élémentaire de $\Omega_1 \times \cdots \times \Omega_n$ dans laquelle l'issue de l'expérience numéro i est donnée par a_i, pour tout $i = 1, \ldots n$.

Lorsque tous les $(\Omega_i, \mathbb{P}_i)$ sont égaux à un seul et même $(\Omega_1, \mathbb{P}_1)$, on note

$$\Omega_1^n = \Omega_1 \times \cdots \times \Omega_1$$

et

$$\mathbb{P}_1^n = \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_1,$$

et l'on parle de **répétition indépendante** plutôt que de succession.

Au passage, notez que l'ordre (réel ou supposé) dans lequel la succession des épreuves a lieu n'influe pas sur les conséquences concrètes que l'on tire du modèle : si l'on permutait l'ordre dans lequel les expériences sont indexées, pour fabriquer le modèle décrivant la succession indépendante des mêmes épreuves, mais dans un ordre différent, on aboutirait, pour un même événement concret, à la même probabilité dans chacun deux modèles. On retrouve une propriété d'échangeabilité comparable à celle déjà mentionnée pour le modèle de tirages uniformes successifs sans remise.

L'identité écrite ci-dessus pour $\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$ est encore valable pour des événements dont la définition comporte des «jokers», du type :

$$a_1 * a_3 a_4 * * * a_8 a_9 \ldots a_n,$$

ou, plus généralement :

$$* \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *,$$

correspondant au fait que l'issue de l'épreuve numéro i_1 est donnée par a_{i_1}, celle de l'épreuve numéro i_2 par a_{i_2},..., celle de l'épreuve numéro i_p par a_{i_p}, les issues des autres épreuves n'étant pas spécifiées. Plus précisément, en notant (pour économiser un peu de place) $\Omega := \Omega_1 \times \cdots \times \Omega_n$ et $\mathbb{P} := \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$, on a :

$$\mathbb{P}(* \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *) = \mathbb{P}_{i_1}(a_{i_1}) \times \mathbb{P}_{i_2}(a_{i_2}) \times \cdots \times \mathbb{P}_{i_p}(a_{i_p}).$$

Vérifions cette propriété.Notons A l'événement $* \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *$. Par définition de la probabilité d'un événement comme somme des probabilités des éventualités élémentaires qui le constituent et par définition de $\mathbb{P}$ à partir des $\mathbb{P}_i$, on a :

$$\begin{aligned}
\mathbb{P}(A) &= \sum_{(d_1,\ldots,d_n)\in A} \mathbb{P}(d_1,\ldots,d_n) \\
&= \sum_{(d_1,\ldots,d_n)\in\Omega \,:\, d_{i_1}=a_{i_1},\ldots,d_{i_p}=a_{i_p}} \mathbb{P}(d_1,\ldots,d_n) \\
&= \sum_{(d_1,\ldots,d_n)\in\Omega^n \,:\, d_{i_1}=a_{i_1},\ldots,d_{i_p}=a_{i_p}} \mathbb{P}_1(d_1) \times \cdots \times \mathbb{P}_n(d_n).
\end{aligned}$$

Définissons $J = \{1,\ldots,n\} - \{i_1,\ldots,i_p\}$, et notons $j_1,\ldots,j_{n-p}$ les éléments de

J. D'après ce qui précède,

$$\begin{aligned}
\mathbb{P}(A) &= \sum_{(d_{j_1},\dots,d_{j_{n-p}})\in\Omega_{j_1}\times\cdots\times\Omega_{j_{n-p}}} \mathbb{P}_{j_1}(d_{j_1})\times\cdots\times\mathbb{P}_{j_{n-p}}(d_{j_{n-p}})\times\mathbb{P}_{i_1}(a_{i_1})\times\cdots\times\mathbb{P}_{i_p}(a_{i_p}) \\
&= \mathbb{P}_{i_1}(a_{i_1})\times\cdots\times\mathbb{P}_{i_p}(a_{i_p})\times\left(\sum_{(d_{j_1},\dots,d_{j_{n-p}})\in\Omega_{j_1}\times\cdots\times\Omega_{j_{n-p}}} \mathbb{P}_{j_1}(d_{j_1})\times\cdots\times\mathbb{P}_{j_{n-p}}(d_{j_{n-p}})\right) \\
&= \big(\mathbb{P}_{i_1}(a_{i_1})\times\cdots\times\mathbb{P}_{i_p}(a_{i_p})\big)\times\left(\sum_{d_{j_1}\in\Omega_{j_1}}\mathbb{P}_{j_1}(d_{j_1})\right)\times\cdots\times\left(\sum_{d_{j_{n-p}}\in\Omega_{j_{n-p}}}\mathbb{P}_{j_{n-p}}(d_{j_{n-p}})\right).
\end{aligned}$$

En notant que chacune des sommes du type

$$\sum_{d_{j_k}\in\Omega_{j_k}}\mathbb{P}_{j_k}(d_{j_k})$$

est en fait la somme sur toutes les éventualités élémentaires de l'espace Ω_{j_k} des valeurs de la probabilité $\mathbb{P}_{j_k}$, on constate que toutes ces sommes sont en fait égales à 1, d'où finalement l'égalité recherchée :

$$\mathbb{P}(*\cdots*a_{i_1}*\cdots*a_{i_2}*\cdots*a_{i_p}*\cdots*) = \mathbb{P}_{i_1}(a_{i_1})\times\mathbb{P}_{i_2}(a_{i_2})\times\cdots\times\mathbb{P}_{i_p}(a_{i_p}).$$

Vous observerez que seules les notations sont impressionnantes, la démonstration elle-même n'étant qu'une petite manipulation sur les produits de sommes à partir de la définition de $\mathbb{P}$. Cette remarque vaut pour toutes les démonstrations qui suivent, et le premier qui a peur des notations a perdu ! Il est également possible de prouver ce résultat en utilisant judicieusement une représentation arborescente. Voir l'exercice 62.

Posons-nous à présent la question suivante : que signifie pour un événement A le fait de s'exprimer seulement en termes des résultats des expériences numérotées $i_1,\dots,i_p$ et pas des autres ? Quelques instants de réflexion nous conduisent à la réponse suivante : si A se met sous la forme :

$$A = \{(d_1,\dots,d_n)\in\Omega_1\times\cdots\times\Omega_n \ : \ (d_{i_1},\dots,d_{i_p})\in A_{i_1,\dots,i_p}\},$$

où $A_{i_1,\dots,i_p}$ est un sous-ensemble de $\Omega_{i_1}\times\cdots\times\Omega_{i_p}$. En effet, l'expression ci-dessus traduit bien le fait que, pour une éventualité élémentaire $(a_1,\dots a_n)$, le fait d'être un élément de A ne donne lieu à aucune condition sur a_l si $l\notin\{i_1,\dots,i_p\}$, mais simplement à une condition (sous la forme la plus générale possible) sur $a_{i_1},\dots,a_{i_p}$. Une autre manière de présenter les choses est de dire que A s'écrit comme une réunion d'événements deux-à-deux disjoints du type :

$$*\cdots*a_{i_1}*\cdots*a_{i_2}*\cdots*a_{i_p}*\cdots*.$$

En effet, écrire A sous la forme

$$A = \{(d_1, \ldots, d_n) \in \Omega_1 \times \cdots \times \Omega_n \ : \ (d_{i_1}, \ldots, d_{i_p}) \in A_{i_1,\ldots,i_p}\},$$

revient à l'écrire :

$$A = \bigcup_{(a_{i_1},\ldots,a_{i_p}) \in A_{i_1,\ldots,i_p}} * \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *,$$

les événements apparaissant dans la réunion ci-dessus étant par ailleurs deux-à-deux **disjoints** car, si $(a_{i_1}, \ldots, a_{i_p}) \neq (a'_{i_1}, \ldots, a'_{i_p})$, il existe au moins un indice i_l tel que $a_{i_l} \neq a'_{i_l}$, et les événements

$$* \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *$$

et

$$* \cdots * a'_{i_1} * \cdots * a'_{i_2} * \cdots * a'_{i_p} * \cdots *$$

sont donc incompatibles (ils imposent deux valeurs différentes pour la même coordonnée i_l).

Nous pouvons maintenant énoncer la propriété connue sous le nom de **théorème des coalitions** : si la définition de A ne fait intervenir que les résultats des expériences numérotées $i_1, \ldots, i_p$ et si la définition de B ne fait intervenir que les résultats des expériences numérotées $j_1, \ldots, j_q$, et si les deux ensembles d'indices $I = \{i_1, \ldots, i_p\}$ et $J = \{j_1, \ldots, j_q\}$ sont **disjoints**, alors A et B sont indépendants.

Avant tout commentaire, prouvons cette propriété. Tout d'abord, notons que, si les deux ensembles d'indices $I = \{i_1, \ldots, i_p\}$ et $J = \{j_1, \ldots, j_q\}$ sont disjoints, un événement de la forme $A = * \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *$ et un événement de la forme $B = * \cdots * b_{j_1} * \cdots * b_{j_2} * \cdots * b_{j_q} * \cdots *$ sont toujours indépendants. En effet, leur intersection s'écrit :

$$A \cap B = * \cdots * c_{k_1} * \cdots * c_{k_2} * \cdots * c_{k_{p+q}} * \cdots *,$$

où

$$\{k_1, \ldots, k_{p+q}\} = \{i_1, \ldots, i_p\} \cup \{j_1, \ldots, j_q\},$$

et où $c_{k_i} = a_{k_i}$ si $k_i \in I$ et $c_{k_i} = b_{k_i}$ si $k_i \in J$. (comme I et J sont disjoints, $I \cup J$ comporte $|I| + |J| = p + q$ éléments.) Par conséquent :

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{P}_{k_1}(c_{k_1}) \times \cdots \times \mathbb{P}_{k_{p+q}}(c_{k_{p+q}}) \\ &= \mathbb{P}_{i_1}(a_{i_1}) \times \cdots \times \mathbb{P}_{i_p}(a_{i_p}) \times \mathbb{P}_{j_1}(b_{j_1}) \times \cdots \times \mathbb{P}_{j_q}(b_{j_q}) \\ &= \mathbb{P}(A) \times \mathbb{P}(B), \end{aligned}$$

le passage de la première à la deuxième ligne utilisant le fait que I et J sont disjoints. Le théorème des coalitions, que nous souhaitons démontrer, affirme plus généralement que l'on a indépendance entre A et B lorsque A est de la forme :

$$A = \{(d_1, \ldots, d_n) \in \Omega_1 \times \cdots \times \Omega_n \ : \ (d_{i_1}, \ldots, d_{i_p}) \in A_{i_1,\ldots,i_p}\},$$

et B de la forme

$$B = \{(d_1, \ldots, d_n) \in \Omega_1 \times \cdots \times \Omega_n \ : \ (d_{j_1}, \ldots, d_{j_q}) \in B_{j_1,\ldots,j_q}\},$$

où $A_{i_1,\ldots,i_p}$ est un sous-ensemble de $\Omega_{i_1} \times \cdots \times \Omega_{i_p}$, et $B_{j_1,\ldots,j_q}$ est un sous-ensemble de $\Omega_{j_1} \times \cdots \times \Omega_{j_q}$.

Pour deux tels événements A et B, écrivant A sous la forme d'une réunion d'événements deux-à-deux disjoints :

$$A = \bigcup_{(a_{i_1},\ldots,a_{i_p}) \in A_{i_1,\ldots,i_p}} * \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *$$

et B sous la forme d'une réunion d'événements deux-à-deux disjoints :

$$B = \bigcup_{(b_{j_1},\ldots,b_{j_q}) \in B_{j_1,\ldots,j_q}} * \cdots * b_{j_1} * \cdots * b_{j_2} * \cdots * b_{j_q} * \cdots *,$$

on en déduit que

$$\mathbb{P}(A) = \sum_{(a_{i_1},\ldots,a_{i_p}) \in A_{i_1,\ldots,i_p}} \mathbb{P}^{\otimes n}\left[(* \cdot * a_{i_1} * \cdot * a_{i_2} * \cdot * a_{i_p} * \cdot *)\right]$$

et que

$$\mathbb{P}(B) = \sum_{(b_{j_1},\ldots,b_{j_q}) \in B_{j_1,\ldots,j_q}} \mathbb{P}^{\otimes n}\left[(* \cdot * b_{j_1} * \cdot * b_{j_2} * \cdot * b_{j_q} * \cdot *)\right].$$

D'autre part, on obtient, en distribuant l'intersection par rapport aux réunions, que : l'événement $A \cap B$ se réécrit sous la forme

$$\bigcup_{(a_{i_1},\ldots,a_{i_p}) \in A_{i_1,\ldots,i_p}} \ \bigcup_{(b_{j_1},\ldots,b_{j_q}) \in B_{j_1,\ldots,j_q}} (* \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *) \cap (* \cdots * b_{j_1} * \cdots * b_{j_2} * \cdots * b_{j_q} * \cdots *).$$

Les événements

$$* \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *, (a_{i_1}, \ldots, a_{i_p}) \in A_{i_1,\ldots,i_p}$$

étant deux-à-deux disjoints, de même que les événements

$$* \cdots * b_{j_1} * \cdots * b_{j_2} * \cdots * b_{j_q} * \cdots *, (b_{j_1}, \ldots, b_{j_q}) \in B_{j_1,\ldots,j_q},$$

c'est également le cas des événements

$$(* \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *) \cap (* \cdots * b_{j_1} * \cdots * b_{j_2} * \cdots * b_{j_q} * \cdots *),$$

où $(a_{i_1}, \ldots, a_{i_p})$ décrit $A_{i_1,\ldots,i_p}$ et où $(b_{j_1}, \ldots, b_{j_q})$ décrit $B_{j_1,\ldots,j_q}$. (petit exercice ne présentant aucune difficulté, à chercher vous-même.) Par conséquent, $\mathbb{P}(A \cap B)$ est égale à (les $\cdots$ sont remplacés par des $\cdot$ pour limiter la taille des formules) :

$$\begin{aligned} &\sum\sum \mathbb{P}\left[(* \cdot * a_{i_1} * \cdot * a_{i_2} * \cdot * a_{i_p} * \cdot *) \cap (* \cdot * b_{j_1} * \cdot * b_{j_2} * \cdot * b_{j_q} * \cdot *)\right] \\ = &\sum\sum \mathbb{P}\left[(* \cdot * a_{i_1} * \cdot * a_{i_2} * \cdot * a_{i_p} * \cdot *)\right] \times \mathbb{P}\left[(* \cdot * b_{j_1} * \cdot * b_{j_2} * \cdot * b_{j_q} * \cdot *)\right], \end{aligned}$$

où la notation $\sum\sum$ désigne la sommation

$$\sum_{(a_{i_1},\ldots,a_{i_p}) \in A_{i_1,\ldots,i_p}} \sum_{(b_{j_1},\ldots,b_{j_q}) \in B_{j_1,\ldots,j_q}},$$

d'après le résultat précédent selon lequel, I et J étant disjoints, un événement de la forme $* \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *$ et un événement de la forme $* \cdots * b_{j_1} * \cdots * b_{j_2} * \cdots * b_{j_q} * \cdots *$ sont toujours indépendants. On en déduit que $\mathbb{P}(A \cap B)$ est égal à

$$\left(\sum_{(a_{i_1},\ldots,a_{i_p}) \in A_{i_1,\ldots,i_p}} \mathbb{P}\left[(* \cdot * a_{i_1} * \cdot * a_{i_2} * \cdot * a_{i_p} * \cdot *)\right]\right) \times \left(\sum_{(b_{j_1},\ldots,b_{j_q}) \in B_{j_1,\ldots,j_q}} \mathbb{P}\left[(* \cdot * b_{j_1} * \cdot * b_{j_2} * \cdot * b_{j_q} * \cdot *)\right]\right),$$

d'où finalement, d'après les expressions précédentes de $\mathbb{P}(A)$ et $\mathbb{P}(B)$, le fait que

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B).$$

Il est également possible de prouver ce résultat en utilisant judicieusement une représentation arborescente. Voir l'exercice 62.

Ouf ! Ce résultat est important car on l'utilise souvent en pratique pour évaluer la probabilité de divers événements dans le cadre d'une succession indépendante d'épreuves. Nous le retiendrons sous la forme suivante, qui en justifie le nom : dans le cadre d'une succession indépendante d'épreuves, deux événements dont les définitions font intervenir des coalitions disjointes d'épreuves sont indépendants. Peut-être estimez-vous que ce résultat est évident et ne nécessite donc pas de démonstration. Il était cependant indispensable d'en fournir une afin d'illustrer la cohérence de la définition formelle d'une succession indépendante d'épreuves et l'idée intuitive que nous pouvons nous faire des propriétés d'une telle succession. Ce résultat nous permet également de formaliser la notion, délicate, d'indépendance mutuelle, ou encore globale, d'une famille d'événements $A_1, \ldots, A_n$.

Commençons avec trois événements. Partant de la définition de l'indépendance de deux événements A et B exprimée sous la forme $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$, il serait

tentant d'essayer de définir l'indépendance de trois événements A, B, C par le fait que $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A) \times \mathbb{P}(B) \times \mathbb{P}(C)$.

Une telle «définition» n'est cependant pas raisonnable, car, en prenant $C = \emptyset$, on constate qu'elle est vérifiée pour tout couple d'événements A et B, indépendants ou non. Or il semble raisonnable de demander que trois événements indépendants dans leur ensemble le soient au moins deux-à-deux.

Cela suffit-il ? Etudions un petit exemple.

Exemple :

La duchesse d'Aquitaine et la duchesse de Bourgogne attendent chacune l'héritier de leur duché. On décrit la situation à l'aide du modèle probabiliste suivant :

$$\Omega = \{\text{GG,GF,FG,FF}\},$$

où la première lettre indique le sexe de l'héritier d'Aquitaine, et la seconde celui de l'héritier de Bourgogne, et la probabilité sur Ω est la probabilité uniforme. Considérons les trois événements :

- $A =$ «l'héritier d'Aquitaine est un garçon»,
- $B =$ «l'héritier de Bourgogne est une fille»,
- $C =$ «les deux héritiers sont de sexe opposé».

On vérifie facilement que $A = \{\text{GG, GF}\}$, $B = \{\text{GF, FF}\}$, $C = \{\text{GF, FG}\}$, $A \cap B = \{\text{GF}\}$, $A \cap C = \{\text{GF}\}$, $B \cap C = \{\text{GF}\}$. Par conséquent :

- $\mathbb{P}(A \cap B) = 1/4 = \mathbb{P}(A) \times \mathbb{P}(B)$,
- $\mathbb{P}(A \cap C) = 1/4 = \mathbb{P}(A) \times \mathbb{P}(C)$,
- $\mathbb{P}(B \cap C) = 1/4 = \mathbb{P}(B) \times \mathbb{P}(C)$,

et les trois événements A, B, C forment donc une famille d'événements indépendants deux-à-deux. En revanche, la réalisation de A et de B entraîne automatiquement celle de C, et donc,

$$\mathbb{P}(C|A \cap B) = 1 \neq \mathbb{P}(C).$$

En ce sens, C n'est pas indépendant de A et B puisque qu'une information portant sur A et B (en l'occurrence, leur réalisation simultanée), modifie la probabilité de réalisation de C. On ne peut donc pas raisonnablement dire que A,B et C sont globalement indépendants entre eux. L'indépendance deux-à-deux est donc insuffisante.

Pour définir l'**indépendance mutuelle** d'une famille d'événements $\mathcal{A} = (A_1, \ldots, A_n)$ d'un modèle probabiliste $(\Omega, \mathbb{P})$, définissons d'abord un autre modèle probabiliste $(\Omega_{\mathcal{A}}, \mathbb{P}_{\mathcal{A}})$, compatible avec $(\Omega, \mathbb{P})$ mais moins fin, qui ne rend compte que de la réalisation (ou de la non-réalisation) de chacun des événements A_i :

$$\Omega_{\mathcal{A}} = \{A_1, A_1^c\} \times \cdots \times \{A_n, A_n^c\}.$$

Pour que la probabilité $\mathbb{P}_{\mathcal{A}}$ rende $(\Omega_{\mathcal{A}}, \mathbb{P}_{\mathcal{A}})$ compatible avec $(\Omega, \mathbb{P})$, on doit naturellement avoir :

$$\mathbb{P}_{\mathcal{A}}\left[A_1^{f_1}, A_2^{f_2}, \ldots, A_n^{f_n}\right] = \mathbb{P}\left[A_1^{f_1} \cap A_2^{f_2} \cap \cdots \cap A_n^{f_n}\right],$$

où les f_i peuvent prendre la valeur ^{c}, $A_i^{f_i}$ désignant alors A_i^c, ou la valeur $\ $ (blanc), $A_i^{f_i}$ désignant alors tout simplement A_i. Nous dirons alors que les événements $(A_1, \ldots, A_n)$ sont **mutuellement**, ou encore **globalement** indépendants, lorsque le modèle $(\Omega_{\mathcal{A}}, \mathbb{P}_{\mathcal{A}})$ coïncide avec le modèle formé par la succession indépendante des modèles $(\Omega_i, \mathbb{P}_i)$ définis par :

$$\Omega_i = \{A_i, A_i^c\},\ \mathbb{P}_i(A_i) = \mathbb{P}(A_i),\ \mathbb{P}_i(A_i^c) = 1 - \mathbb{P}(A_i).$$

En termes plus prosaïques, nous pouvons définir l'indépendance mutuelle des événements $A_1, \ldots, A_n$ de la façon suivante : pour toute famille $f_1, \ldots, f_n$ d'indices telle que $f_i \in \{\ , ^c\}$, on a l'égalité :

$$\mathbb{P}\left[A_1^{f_1} \cap A_2^{f_2} \cap \cdots \cap A_n^{f_n}\right] = \mathbb{P}\left[A_1^{f_1}\right] \times \cdots \times \mathbb{P}\left[A_n^{f_n}\right].$$

Les deux définitions sont équivalentes, mais la première, quoiqu'un peu plus abstraite au premier abord, a l'avantage de bien expliquer la seconde. De plus, nous pouvons (en vertu de la première définition) utiliser le théorème des coalitions dans notre contexte : si $A_1, \ldots, A_n$ sont des événements mutuellement indépendants, et si $I = \{i_1, \ldots, i_p\}$ et $J = \{j_1, \ldots, j_q\}$ sont deux familles d'indices disjoints de $\{1, \ldots, n\}$, un événement défini seulement à partir des événements $A_{i_1}, \ldots, A_{i_p}$ et un événement défini seulement à partir des événements $A_{j_1}, \ldots, A_{j_q}$ sont indépendants. Nous voyons donc apparaître ce qui constitue le cœur de la notion d'indépendance globale d'une famille d'événements : non seulement les couples d'événements de cette famille doivent être indépendants, mais également les couples de **coalitions** d'événements de cette famille se rapportant à deux groupes d'événements séparés. Par exemple les événements «A_1 se produit et A_4 ne se produit pas» et «A_2, A_3 ou A_5 se produit».

Comment utiliser cette notion ? La plupart du temps, nous l'utiliserons sous la forme suivante : nous saurons *a priori* que les événements $A_1, \ldots, A_n$ sont mutuellement indépendants, et nous utiliserons ce fait pour calculer les probabilités du type

$$\mathbb{P}\left[A_1^{f_1} \cap A_2^{f_2} \cap \cdots \cap A_n^{f_n}\right],$$

qui seront donc égales à

$$\mathbb{P}\left[A_1^{f_1}\right] \times \cdots \times \mathbb{P}\left[A_n^{f_n}\right].$$

Le problème sera donc la plupart du temps le suivant : comment établir qu'une famille d'événements $A_1, \ldots, A_n$ est bien une famille d'événements mutuellement

indépendants ? Sans surprise, nous rencontrerons la plupart du temps les familles d'événements indépendants dans le cadre des successions indépendantes d'épreuves. Plus précisément, nous allons prouver la **proposition** suivante : dans un modèle $(\Omega, \mathbb{P})$ décrivant une succession indépendante de N épreuves, une famille $A_1, \ldots, A_n$ d'événements dont les définitions se réfèrent à des ensembles d'indices deux-à-deux disjoints, est une famille d'événements mutuellement indépendants. Autrement dit, si, pour tout i, la définition de l'événement A_i ne se réfère qu'aux épreuves dont les numéros figurent dans l'ensemble d'indices $I_i \subset \{1, \ldots, n\}$, et si, pour tout $1 \leq i \neq j \leq n$, $I_i \cap I_j = \emptyset$, alors les événements $A_1, \ldots, A_n$ sont mutuellement indépendants.

La preuve de ce résultat repose sur une application itérée du théorème des coalitions : en effet, considérons des événements $A_1, \ldots, A_n$ satisfaisant les hypothèses ci-dessus, et cherchons à calculer la probabilité d'un événement du type :

$$A_1^{f_1} \cap A_2^{f_2} \cap \cdots \cap A_n^{f_n}.$$

On constate que, d'après nos hypothèses, les deux événements $A_n^{f_n}$, d'une part, et

$$A_1^{f_1} \cap A_2^{f_2} \cap \cdots \cap A_{n-1}^{f_{n-1}}$$

d'autre part, se réfèrent à deux groupes d'épreuves disjoints, donc le théorème des coalitions entraîne que ces deux événements sont indépendants, et, par conséquent, que :

$$\mathbb{P}\left[A_1^{f_1} \cap A_2^{f_2} \cap \cdots \cap A_n^{f_n}\right] = \mathbb{P}\left[A_1^{f_1} \cap A_2^{f_2} \cap \cdots A_{n-1}^{f_{n-1}}\right] \times \mathbb{P}(A_n^{f_n}).$$

En itérant l'argument, on vérifie bien que l'on a finalement :

$$\mathbb{P}\left[A_1^{f_1} \cap A_2^{f_2} \cap \cdots \cap A_n^{f_n}\right] = \mathbb{P}\left[A_1^{f_1}\right] \times \cdots \times \mathbb{P}\left[A_n^{f_n}\right].$$

1.7.1 Une histoire de singe

Décrivons à présent, pour nous distraire un peu avant la fin de ce chapitre et nous récompenser des efforts accomplis jusqu'ici, l'histoire du singe dactylographe. La voici : un singe est placé devant un ordinateur (dans les versions plus anciennes, il s'agissait d'une machine à écrire...) et pianote aléatoirement sur le clavier. Pour simplifier, nous supposerons que le clavier ne comporte que deux touches, P et F, ce qui fait que le singe saisit directement en binaire, et que la succession des touches frappées par le singe peut être modélisée par une succession d'expériences indépendantes consistant à choisir l'une des deux touches, chacune ayant une probabilité de $1/2$ d'être choisie. Posons-nous alors la question : quelle est la probabilité pour que le singe, une fois placé devant l'ordinateur, saisisse directement (codé en binaire caractère par caractère, à l'aide de P et de F) le texte du «Discours de la méthode» de Descartes ? Extrêmement faible, voire nulle, répondrez-vous, et ce avec quelque raison puisque, en estimant que le texte contient environ 130000 caractères alphabétiques et

signes typographiques, et requiert donc environ l'utilisation de $6 \times 130000 = 780000$ caractères binaires, on obtient une probabilité de $2^{-780000}$, et cet événement semble donc pratiquement impossible. En revanche, la probabilité pour que le singe écrive complètement le texte du «Discours de la méthode» **au bout d'un certain temps** devient, si l'on s'autorise à attendre suffisament longtemps, extrêmement proche de 1 ! Précisons ceci. Appelons N le nombre total de touches que l'on autorise le singe à frapper avant d'arrêter l'expérience, et considérons la suite $(a_1, a_2, \ldots, a_L)$ ($L \approx 780000$) formée par le codage binaire du texte de Descartes. L'événement A_1 correspondant au fait que le singe saisisse immédiatement (c'est-à-dire à partir de la première touche frappée) le codage binaire du «Discours de la méthode», s'écrit tout simplement, avec nos notations, sous la forme :

$$A_1 = \underbrace{a_1 a_2 \ldots a_L * \cdots *}_{\text{longueur totale } N},$$

et l'on a, dans notre modèle,

$$\mathbb{P}^{\otimes N}(A_1) = 2^{-L}.$$

Définissons plus généralement, pour $1 \leq i \leq N - L + 1$, l'événement A_i :

$$A_i = \underbrace{* \cdots *}_{i-1 \text{ jokers}} \; a_1 a_2 \ldots a_L \; \underbrace{* \cdots *}_{N-L+1-i \text{ jokers}},$$

qui correspond au fait que le texte du «Discours de la méthode» est saisi à partir du $i-$ème caractère frappé par le singe (comme on arrête l'expérience après la $N-$ème touche frappée, on doit nécessairement avoir $i \leq N-L+1$, sans quoi le texte n'aurait pas la possibilité d'être saisi complètement.) Chaque événement A_i a également une probabilité égale à 2^{-L} dans notre modèle. L'événement

$$B_N = \bigcup_{i=1}^{N-L+1} A_i,$$

correspond, par définition, au fait que, au bout d'un certain temps, le singe saisit entièrement le texte du «Discours de la méthode», et nous allons montrer que, lorsque N tend vers l'infini (c'est-à-dire, lorsque l'on poursuit l'expérience pendant un nombre de touches frappées qui tend vers l'infini), la probabilité $\mathbb{P}^{\otimes N}(B_N)$ tend vers 1, ou, autrement dit, que l'événement B_N devient très probable lorsque N tend vers l'infini.

On ne peut pas calculer la probabilité de B_N en utilisant une relation du type :

$$\mathbb{P}^{\otimes N}(B_N) \text{ « = » } \sum_{i=1}^{N-L+1} \mathbb{P}^{\otimes N}(A_i),$$

car les événements A_i ne sont pas en général deux-à-deux disjoints (si N est assez grand, on pourrait très bien avoir plusieurs versions du «Discours de la méthode» figurant à la suite dans le texte saisi par le singe). (Les guillemets sont là pour rappeler aux amateurs de lecture en diagonale que l'égalité n'est pas valable.)

Pour montrer que la probabilité de B_N est proche de 1 lorsque N tend vers l'infini, nous allons plutôt tenter de montrer que la probabilité de son complémentaire, B_N^c tend vers zéro lorsque N tend vers l'infini. Le complémentaire d'une réunion étant l'intersection des complémentaires, on a :

$$B_N^c = \bigcap_{i=1}^{N-L+1} A_i^c,$$

Chaque événement A_i ayant une probabilité égale à 2^{-L} de se produire, les événements A_i^c ont chacun une probabilité égale à $1-2^{-L}$ de se produire. Si les événements A_i formaient une famille d'événements mutuellement indépendants, la probabilité de B^c serait simplement donnée par :

$$\mathbb{P}^{\otimes N}(B_N^c) \text{ « = » } \prod_{i=1}^{N-L+1} (1 - \mathbb{P}(A_i)).$$

Malheureusement, ce n'est pas le cas, car, par exemple, la réalisation de A_i est incompatible avec celle de A_{i+1} (pourquoi ?). En revanche, A_1 est indépendant de A_{L+1} car ces deux événements font référence à deux groupes disjoints d'expériences : A_1 ne se réfère qu'aux résultats des L premières frappes, alors que A_{L+1} ne se réfère qu'aux résultats des expériences numérotées de $L+1$ à $2L$.

Plus généralement, les événements A_1, A_{L+1}, $A_{2L+1}, \ldots, ..., A_{kL+1}$ sont mutuellement indépendants car ils se rapportent à des groupes d'expériences deux-à-deux disjoints, (k devant bien sûr vérifier l'inégalité $(k+1)L \leq N$, autrement dit, $k \leq \lfloor N/L \rfloor - 1$, $\lfloor u \rfloor$ désignant la partie entière de u). Puisque nous souhaitons simplement obtenir une minoration de la probabilité de l'événement B_N (nous voulons montrer que celle-ci est proche de 1 lorsque N est assez grand), il nous suffit de **majorer** la probabilité de l'événement B_N^c, et il n'est pas nécessaire de la calculer exactement. Or l'événement B_N^c correspond à la réalisation simultanée des $N-L+1$ événements A_1^c, A_2^c,...,A_{N-L+1}^c, et par conséquent :

$$B_N^c = \bigcap_{i=1}^{N-L+1} A_i^c \subset \bigcap_{k=0}^{\lfloor N/L \rfloor - 1} A_{kL+1}^c.$$

Ce dernier événement étant, d'après ce qui précède, une intersection d'événements mutuellement indépendants, sa probabilité peut être facilement calculée :

$$\mathbb{P}^{\otimes N}\left(\bigcap_{k=0}^{\lfloor N/L \rfloor - 1} A_{kL+1}^c\right) = \prod_{k=0}^{\lfloor N/L \rfloor - 1} \mathbb{P}(A_{kL+1}^c) = \left(1 - 2^{-L}\right)^{\lfloor N/L \rfloor}.$$

On en déduit finalement que :

$$\mathbb{P}^{\otimes N}(B_N^c) \leq \left(1 - 2^{-L}\right)^{\lfloor N/L \rfloor}.$$

Lorsque N tend vers l'infini, $\lfloor N/L \rfloor$ tend également vers l'infini (L est fixé, $L \approx 780000$). Comme $(1 - 2^{-L}) < 1$, $\left(1 - 2^{-L}\right)^{\lfloor N/L \rfloor}$ tend vers zéro lorsque N tend vers l'infini, et c'est également le cas de la probabilité $\mathbb{P}^{\otimes N}(B_N^c)$. D'où en définitive le fait que la probabilité de B_N tend effectivement vers 1 lorsque N tend vers l'infini.

Autrement dit, pourvu que N soit assez grand, la probabilité pour que le texte du «Discours de la méthode» figure quelque part dans le texte saisi par le singe peut être rendue arbitrairement proche de 1, et, pour de grandes valeurs de N, la réalisation de cet événement est quasiment certaine. Ceci étant, les valeurs de N nécessaires pour que la probabilité de A soit effectivement proche de 1 sont extrêmement grandes, N devant au moins être de l'ordre de 2^{780000}. En admettant que le singe frappe une touche par seconde, le temps nécessaire pour que la probabilité d'observer effectivement A dépasse la valeur 10^{-3}, par exemple, est très largement supérieur à l'âge estimé de l'univers...

Cette petite anecdote illustre le rôle important joué par les ordres de grandeurs : suivant le nombre de répétitions de l'expérience que l'on réalise, le même événement pourra apparaître comme pratiquement impossible ou au contraire pratiquement certain.

1.7.2 Tout résultat est exceptionnel !

Comme vous n'aurez pas manqué de le noter, l'exemple précédent correspond à un modèle de type «pile ou face», dans lequel des épreuves binaires (à deux issues) sont répétées. En dépit de sa simplicité, et pour toutes sortes de raisons, dont quelques unes apparaîtront dans la suite, ce modèle joue un rôle important dans la théorie des probabilités, ce qui justifie que nous nous attardions quelque peu sur son étude.

L'expérience aléatoire consiste à lancer une pièce de monnaie et à noter le résulat : pile (P), ou face (F) (on suppose que la pièce ne reste jamais sur la tranche). Le modèle probabiliste qui décrit une épreuve est donc $\Omega = \{\mathrm{P},\mathrm{F}\}$, la probabilité sur Ω étant définie par

$$\begin{cases} \mathbb{P}(\mathrm{P}) = p \\ \mathbb{P}(\mathrm{F}) = 1 - p \end{cases}$$

où $p \in [0,1]$ n'est pas nécessairement égal à $1/2$. Le modèle probabiliste correspondant à N successions indépendantes de lancers fait appel à l'espace des possibles $\Omega^N = \{\mathrm{P},\mathrm{F}\}^N$ constitué de toutes les suites de P et de F de longueur N, et la probabilité $\mathbb{P}^{\otimes N}$ est définie par : $\mathbb{P}^{\otimes N}(\omega_1, \ldots, \omega_N) = \mathbb{P}(\omega_1) \times \cdots \times \mathbb{P}(\omega_N)$, chaque ω_i pouvant prendre l'une des deux valeurs P ou F. On peut également représenter

Ω^N par un arbre binaire régulier de profondeur N, à chaque sommet non-terminal étant associée une copie du modèle $(\Omega, \mathbb{P})$. En examinant de plus près l'expression de $\mathbb{P}^{\otimes N}$, on constate que la probabilité d'une suite donnée de P et de F $(\omega_1, \ldots, \omega_N)$ ne dépend que du nombre total de P et de F, et non pas de l'ordre dans lequel ceux-ci surviennent. Ainsi, si $S(\omega_1, \ldots, \omega_N)$ désigne le nombre total de P présents dans la suite $(\omega_1, \ldots, \omega_N)$, la probabilité $\mathbb{P}^{\otimes N}$ se met sous la forme :

$$\mathbb{P}^{\otimes N}(\omega_1, \ldots, \omega_N) = p^{S(\omega_1,\ldots,\omega_N)} \times (1-p)^{N-S(\omega_1,\ldots,\omega_N)},$$

(le nombre de F présents dans $(\omega_1, \ldots, \omega_N)$ étant égal à $N - S(\omega_1, \ldots, \omega_N)$).

Lorsque $p = 1/2$, on a $p = 1 - p$, et l'expression se simplifie :

$$\mathbb{P}^{\otimes N}(\omega_1, \ldots, \omega_N) = (1/2)^{S(\omega_1,\ldots,\omega_N)} \times (1/2)^{N-S(\omega_1,\ldots,\omega_N)} = (1/2)^N,$$

autrement dit, la probabilité $\mathbb{P}^{\otimes N}(\omega_1, \ldots, \omega_N)$ ne dépend pas de $(\omega_1, \ldots, \omega_N)$, et il s'agit donc de la probabilité uniforme sur Ω^N, conformément à une remarque précédente : lorsque $p = 1/2$, l'espace probabilisé décrivant l'expérience d'un seul lancer est muni de la probabilité uniforme, et, par conséquent, la probabilité sur $(\Omega^N, \mathbb{P}^{\otimes N})$ décrivant la succession indépendante de N lancers est la probabilité uniforme sur Ω^N.

Quelle est la suite de P et de F la plus probable dans le modèle précédent ? Si $p = 1/2$, nous venons de voir que la probabilité est uniforme et que, par conséquent, aucune suite n'est plus probable qu'une autre. Si $p > 1/2$, au contraire, la suite la plus probable est celle qui ne comporte que des P (et inversement, si $p < 1/2$, c'est celle qui ne comporte que des F). Dans tous les cas, la suite de P et de F la plus probable a une probabilité de la forme h^N où $0 < h < 1$, et donc, même la probabilité de la suite la plus probable tend extrêmement rapidement vers zéro lorsque N tend vers l'infini. Ainsi, lorsque N est grand, quelle que soit la suite de P et de F que nous observions effectivement, celle-ci n'avait de toute façon *a priori* qu'une probabilité extraordinairement petite de survenir. En ce sens, n'importe quel résultat des N lancers est «exceptionnel» !

1.7.3 Succession indépendante ?

Afin de parfaire votre compréhension de la notion de succession **indépendante** d'expériences, et de vous armer face à quelques difficultés conceptuelles qui apparaissent fréquemment lorsque l'on aborde l'estimation statistique, nous vous invitons à réfléchir à la question suivante. On effectue 101 lancers d'une pièce de monnaie, que l'on modélise par une succession indépendantes de lancers, modélisés chacun individuellement par l'espace de probabilité $\Omega = \{\text{P,F}\}$, la probabilité sur Ω étant définie par

$$\begin{cases} \mathbb{P}(\text{P}) = p \\ \mathbb{P}(\text{F}) = 1 - p \end{cases},$$

la valeur du paramètre p nous étant inconnue. Imaginons qu'après 100 lancers, on constate que l'on a obtenu 80 fois pile, et seulement 20 fois face. Il semble alors raisonnable (nous reviendrons plus en détail sur ce point aux chapitres suivants) d'estimer la valeur de p à environ $80/100 = 0,8$, et donc d'affirmer que le 101-ème lancer a environ 80 pour cent de chances de donner pile. Si l'on avait obtenu 50 pile (et donc 50 face), on aurait de même estimé p à environ $50/100 = 0,5$, et estimé que le 101-ème lancer a environ une chance sur deux de donner pile. Il semble donc que les résultats des 100 premiers lancers influent sur le résultat du 101-ème, puisqu'ils nous permettent de déterminer (en gros) la probabilité pour que celui-ci donne pile. Le modèle $(\Omega^{101}, \mathbb{P}^{\otimes 101})$ est pourtant tel que, quels que soient les résultats des 100 premiers lancers, la probabilité (conditionnelle) d'obtenir pile lors du 101-ème lancer sachant ces résultats est toujours la même, à savoir p, ce qui est la définition même de l'indépendance. Il n'y a là qu'un paradoxe apparent lié à notre ignorance de la valeur de p : l'«information» que nous fournissent les résultats des 100 premiers lancers sur la valeur de p n'est pas de la même nature que l'information fournie sur le déroulement d'une expérience aléatoire par la réalisation d'un certain événement. Dans ce modèle, on considère que la valeur de p est fixée (elle correspond à une caractéristique de la pièce et de la manière dont les lancers sont répétés), même si nous ne la connaissons pas, et qu'elle n'a aucun caractère aléatoire sur la probabilité duquel la réalisation des 100 premiers lancers serait susceptible de nous renseigner. En revanche, si l'on imaginait (on aura alors affaire à un second modèle) que l'on procède à 101 lancers indépendants successifs d'une pièce de monnaie, après avoir choisi la pièce en question au hasard parmi trois pièces présentant des caractéristiques différentes, la probabilité p associée à la pièce utilisée apparaîtrait comme aléatoire, et l'information fournie par les résultats des 100 premiers lancers nous fournirait une information (au sens des probabilités conditionnelles) sur la pièce qui a été choisie. En revanche, sachant celle des pièces qui a été sélectionnée, la succession des lancers est une succession d'expériences indépendantes. En ce sens, et même si l'on ne se trouve pas dans un cadre où p est lui-même l'objet d'un choix aléatoire, on dira parfois que le premier modèle est tel que conditionnellement à la valeur p, la succession des tirages est indépendante. Bien entendu, la succession indépendante d'expériences ne constitue qu'un modèle d'une situation réelle, et pas la réalité elle-même. Même après avoir observé pour les 100 premiers lancers 80 piles et 20 face, le modèle de succession indépendante avec $p = 1/2$ prévoit une probabilité exactement égale à $1/2$ pour pile et $1/2$ pour face lors du 101-ème lancer. Dans ce cas, le modèle est complètement discrédité par les données observées, et, à moins d'avoir de très bonnes raisons de croire par ailleurs à sa validité dans cette situation, il est sans doute plus raisonnable de le jeter aux orties.

1.8 Coïncidences troublantes

1.8.1 C'est vraiment incroyable !

Commençons par citer trois exemples documentés de coïncidences troublantes. (Source : `http://www.csj.org/infoserv_articles/astop_unlikely_events.htm`)

La romancière britannique Rebecca West était en train d'écrire un récit dans lequel une petite fille trouvait un hérisson dans son jardin. Aussitôt le passage écrit, les domestiques l'interrompirent dans son travail pour lui signaler qu'ils venaient de trouver un hérisson dans son jardin.

L'écrivain américain Norman Mailer n'avait pas initialement prévu, lorsqu'il entama la rédaction de son roman Barbary Shore, d'y inclure un espion russe comme personnage. Il le fit pourtant et, au cours de l'écriture du livre, ce personnage passa progressivement d'un rôle secondaire à celui de personnage principal du roman. Après que la rédaction fut achevée, les services américains de l'immigration arrêtèrent le voisin du dessus de Norman Mailer, que l'on présenta comme l'un des principaux espions russes en activité aux Etats-Unis à l'époque.

Plusieurs noms de code utlra-secrets furent utilisées par les forces Alliées dans la préparation du débarquement du 6 juin 1944 en Normandie, parmi eux : Utah, Omaha (désignant les plages où le débarquement devait avoir lieu), Mulberry (pour désigner le port artificiel qui devait être installé une fois le débarquement entamé), Neptune (pour désigner le plan des opérations navales), et Overlord (désignant la totalité de l'opération). Le 3 mai 1944, le mot Utah apparut comme l'une des réponses dans le problème de mots croisés du London Daily Telegraph ; le 23 mai, ce fut au tour d'Omaha ; le 31 mai, celui de Mulberry ; et enfin, le 2 juin, Neptune et Overlord firent leur apparition dans le même contexte ! Après une enquête poussée des services de renseignement britanniques, l'auteur des problèmes de mots croisés apparut comme totalement innocent, sans aucune idée du projet de débarquement, et ayant apparemment choisi au hasard les mots employés.

Plus loufoque : en 1981, le prince Charles s'est marié, Liverpool a été champion d'Europe, et le Pape est décédé. En 2005, également le prince Charles s'est marié, Liverpool a été champion d'Europe, et le Pape est décédé.

Vous avez certainement connaissance d'une foule d'autres anecdotes de ce genre, peut-être issues de votre expérience personnelle («Je pensais justement hier soir à mon ami Jojo que je n'avais pas vu depuis deux ans et... chose incroyable, il m'appelle au téléphone ce matin.» «En visitant le château de Blois lors des dernières vacances, c'est incroyable, je tombe sur mon collègue T*** au beau milieu de la cour.», «C'est vraiment surprenant que tu évoques ce sujet, car justement, nous en parlions hier ma femme et moi.»,...), et les exemples les plus frappants sont parfois rapportés dans les journaux. En rédigeant ce passage, j'ai appris qu'un collègue m'avait aperçu la veille

(un dimanche) à un péage autoroutier, où nous nous trouvions donc simultanément lui et moi.

On justifie souvent son propre étonnement devant ce genre de coïncidence par des arguments basés sur la probabilité extrêmement faible de l'événement en question.

Il paraît en effet assez raisonnable, dans les exemples évoqués plus haut, de n'attribuer qu'une probabilité assez faible aux coïncidences dont il est question. Mais pourquoi au juste devrait-on s'étonner de les avoir observées ?

1.8.2 Ce que l'on observe est presque toujours improbable

Prenons l'exemple le plus simple de modèle probabiliste, c'est-à-dire une succession indépendante de lancers de pile ou face. Lançons une pièce dix fois de suite, et notons la suite de résultats obtenus : P pour pile et F pour face. Nous obtenons donc une suite de P et de F de longueur 10, telle que PPFPFFPPFP. Quelle probabilité une suite $(x_1, \ldots, x_{10}) \in \{\text{P,F}\}^{10}$ a-t-elle de sortir dans notre modèle ? Réponse : $1/2^{10}$ quelle que soit la suite, soit moins d'une chance sur 1000. Autrement dit, quel que soit le résultat produit par nos lancers, nous constaterons toujours qu'il n'avait qu'une très faible probabilité de survenir. Cette constatation ne vaut pas seulement pour ce cas particulier, mais pour la plupart des modèles probabilistes et des situations concrètes, dès que l'on cherche à les décrire autrement que par un très petit nombre d'alternatives différentes. De ce point de vue, tout ce que l'on observe, décrit avec suffisament de détail, possède une probabilité extrêmement faible pour la plupart des définitions raisonnables de la probabilité. La probabilité pour que vous vous trouviez exactement là où vous vous trouvez, et non pas quelques centimètres plus loin, que vous ayez exactement la position que vous avez, que vous ayez rencontré aujourd'hui les personnes que vous avez rencontrées, à l'instant exact où vous les avez rencontrées, est vraisemblablement très faible. De ce point de vue, il n'y a pas lieu de s'étonner de la faible probabilité de l'événement que l'on vient d'observer.

1.8.3 Des coïcidences surprenantes doivent se produire

Une autre manière de raisonner sur les coïncidences frappantes, consiste à les replacer dans un cadre plus général, dans lequel on prend en compte l'ensemble des circonstances susceptibles de nous apparaître comme des coïncidences surprenantes dans un contexte donné (au cours d'une période de temps donnée, parmi un groupe d'individus donnés, etc...). Même si chacune de ces coïncidences possède individuellement une très faible probabilité de survenir, le grand nombre d'événements que nous sommes susceptibles d'interpréter comme des coïncidences étonnantes peut rendre extrêmement probable le fait que nous observions régulièrement – et donc relevions – un certain nombre d'entre elles.

1.8.4 Attention à l'interprétation

Le plus souvent cependant, les coïncidences que nous relevons ne nous frappent pas seulement en raison de leur faible probabilité (la plupart du temps bien réelle, comme nous venons de l'expliquer), mais parce qu'elles semblent suggérer une interprétation qui défie le sens commun – un destin mystérieux conduit des amis s'étant perdus de vue depuis longtemps à se retrouver par hasard lors d'un voyage à l'étranger, un étrange don de prémonition vous a fait deviner les trois premiers chiffres du tirage du loto de ce soir, ou penser à un cousin éloigné juste avant que celui-ci ne vous appelle au téléphone, etc...

L'attitude rationnelle face à ces coïncidences consiste bien entendu à tester d'abord de manière systématique les «conclusions» que leur interprétation suggère, avant de gloser plus avant. Par exemple, le fait de penser à une personne accroît-il réellement la probabilité que celle-ci vous appelle peu après ? Pour en juger, il est nécessaire d'enregistrer systématiquement les occasions où il vous arrive d'évoquer une personne de connaissance en pensée, et de mesurer la fréquence avec laquelle ces pensées sont suivies d'un appel de la personne en question dans un délai raisonnablement bref. Ainsi, on évite le **biais de sélection** (ici, d'origine psychologique), consistant à s'étonner, et donc à retenir, les cas où la personne à laquelle vous venez de penser vous appelle, tout en oubliant de remarquer, et donc en négligeant, tous les cas où l'on pense à une personne sans que celle-ci n'appelle dans les minutes qui suivent, et le problème plus évident, mais parfois ignoré, de l'oubli de variabilité qui consisterait à tirer des conclusions à partir de l'observation d'une unique coïncidence.

Il paraît vraisemblable qu'en procédant de cette manière, aucun accroissement significatif de la probabilité d'être appelé ne sera mis en évidence. Toutefois, cela peut parfaitement être le cas sans que cela soit pour autant le signe que vous possédez un don particulier, tout simplement parce qu'il peut être plus probable d'évoquer en pensée des personnes auxquelles on a eu affaire dernièrement, en particulier ses proches, et qui sont par conséquent plus susceptibles de vous appeler que d'autres.

1.8.5 Quand s'étonner ?

Les observations précédentes sont destinées à vous mettre en garde contre un étonnement infondé ou, pire une interprétation erronnée, face à des coïncidences observées, ou rapportées (à ce propos, se pose toujours le problème de la fiabilité des sources).

Pourtant, si un modèle d'une situation prédit qu'un certain événement ne doit survenir qu'avec une faible probabilité, n'y a-t-il jamais lieu d'être surpris, c'est-à-dire de mettre en doute le modèle, si l'on observe cet événement ? La réponse est positive, mais cela n'est pas incompatible avec les remarques précédentes.

A priori et a posteriori

Dans ce qui précède, nous avons constaté que, la plupart du temps, on pouvait **rétrospectivement** attribuer une très faible probabilité à la manière particulière selon laquelle une situation s'était réalisée. Il est bien évident que, dans ce cas, l'événement dont on examine la probabilité dépend de la manière dont la situation s'est réalisée (c'est complètement évident dans l'exemple des lancers de pile ou face). En revanche, lorsque l'événement de faible probabilité auquel on s'intéresse est fixé indépendamment – par exemple à l'avance – de la réalisation de l'expérience, il y a tout lieu d'être surpris si celui-ci se produit, et cela doit inciter, sinon à rejeter le modèle, du moins à réexaminer les arguments en faveur de celui-ci (de manière systématique, naturellement !).

Quant à savoir à partir de quel niveau de probabilité il convient de s'étonner, tout dépend du contexte, et il n'est pas forcément de bonne politique de fixer une limite *a priori* en-deçà de laquelle les événement sont considérés comme improbables, et au-dessus de laquelle leur apparition doit être considérée comme non-surprenante.

D'autre part, en pratique, il n'est bien entendu pas toujours évident de s'assurer qu'il y a bien indépendance entre l'événement considéré et la réalisation de l'expérience (voir la section «Hypothèses suggérées par les données»).

Familles d'événements

La seconde remarque concernant les événements de faible probabilité consistait à noter qu'un très grand nombre d'événements de faible probabilité susceptibles de survenir seraient remarqués comme des coïncidences, et qu'il était donc plus pertinent de considérer la probabilité de la réunion de la totalité de ces événements, plutôt que la probabilité de l'un d'entre eux (celui qui justement s'est produit) isolément. (On évite ainsi de faire dépendre l'événement que l'on considère de la manière dont l'expérience s'est réalisée.) Dans le cas où l'événement que l'on considère ne dépend pas de la réalisation de l'expérience, il n'y a pas lieu de dresser une telle liste !

Probabilité d'une réunion

Rappelons que, de manière générale, on ne peut pas déduire la probabilité d'une réunion $\mathbb{P}(A_1 \cup \cdots \cup A_n)$ des probabilités individuelles, $\mathbb{P}(A_i)$, et l'on dispose seulement d'inégalités, telle que la borne de la réunion :

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \mathbb{P}(A_i),$$

que l'on peut utiliser en toute généralité, et qui est une égalité lorsque les événements A_i sont deux-à-deux disjoints ; on dispose également, dans le cas général toujours, des inégalités et des égalités provenant du principe d'inclusion-exclusion.

On voit ainsi que, en toute généralité, si l'on dispose de n événements dont toutes les probabilités sont inférieures à une valeur ϵ, tout ce que l'on peut en déduire en général est le fait que $\mathbb{P}(A_1 \cup \ldots \cup A_n) \leq n\epsilon$, et l'on ne peut ainsi affirmer que la réunion de tous ces événements est improbable du fait que chacun des événements l'est, que lorsque $n\epsilon << 1$. Bien entendu, rien ne prouve, et il n'est pas vrai en général, que $n\epsilon$ soit le bon ordre de grandeur pour cette probabilité. Dans le cas particulier d'événements indépendants, on peut néanmoins écrire que

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) = 1 - \prod_{k=1}^{n} (1 - \mathbb{P}(A_i)).$$

En admettant que tous les événements aient pour probabilité exactement ϵ, on en déduit que $\mathbb{P}(A_1 \cup \ldots \cup A_n) = 1 - (1-\epsilon)^n$. On en déduit alors facilement que cette probabilité est voisine de 1 lorsque $n\epsilon >> 1$.

1.8.6 Un magicien doué

Voici un petit exemple imaginaire destiné à illustrer quelques unes des observations précédentes.

Ce soir, au cours d'une émission de télévision à succès, une expérience de parapsychologie de grande ampleur est en train d'avoir lieu. M. M***, magicien de son état, prétend pouvoir, par la seule force de son esprit, être capable de faire griller les ampoules électriques chez les téléspectateurs. Ceux-ci ont donc été invités à allumer chez eux diverses lampes électriques, et, après plusieurs minutes de concentration intense de la part de M. M***, des téléspectateurs appellent par dizaines le standard de l'émission pour témoigner qu'effectivement, une, et même dans certains cas plusieurs ampoules électriques ont rendu l'âme pendant que M. M*** se concentrait.

En admettant que l'émission en question soit regardée par plusieurs millions de foyers, et que la probabilité pour une ampoule électrique de griller au cours d'une minute d'utilisation soit d'environ 1/60000 (ce qui correspond à une durée de vie moyenne d'environ mille heures), on s'attend à ce qu'il y ait plusieurs milliers de téléspectateurs chez qui des ampoules grillent au cours de l'émission, et, par voie de conséquence (que feriez-vous à leur place, hein ?), contactent le standard de l'émission.

Pris individuellement, le fait qu'une ampoule grille au moment précis ou M. M*** se concentre semble très surprenant, car très improbable sous l'hypothèse que M. M*** ne détient aucun pouvoir particulier : environ une chance sur 60000. Pourtant, si notre estimation de 1/60000 est correcte, ainsi que celle de plusieurs millions de téléspectateurs, ainsi que l'hypothèse d'une certaine indépendance entre le grillage des ampoules chez les différents téléspectateurs, ce qui serait surprenant serait plutôt que personne n'appelle pour faire part de sa surprise. Replacés parmi l'ensemble des

grillages d'ampoules susceptibles de survenir chez les téléspectateurs, la multitude d'appels constatée n'a donc rien de surprenant. Bien entendu, les téléspectateurs chez qui rien de particulier n'est survenu, qui forment pourtant l'écrasante majorité (sans doute plus de 99,9%) ne se précipitent pas forcément sur leur combiné pour composer le numéro (peut-être surtaxé) permettant d'appeler l'émission, car ils ne pensent pas avoir observé quoique ce soit de remarquable. Si l'on ne se fie qu'aux appels passés pour estimer la probabilité de succès de M. M***, on commet tout simplement un (atroce) biais de sélection.

Intéressons-nous maintenant à la manière dont peut raisonner un téléspectateur sceptique chez qui une ampoule électrique vient pourtant de rendre l'âme. N'accordant aucun crédit à M. M***, il cherche pourtant à examiner les faits objectivement, et ne peut que constater le succès de M. M*** en ce qui le concerne. L'événement qui vient d'être observé est très improbable sous l'hypothèse que M. M*** ne possède aucun don, et cet événement a bien été défini indépendamment du résultat de l'expérience, avant que celle-ci n'ait lieu (ou du moins, c'est ainsi que M. D*** voit les choses individuellement, nous savons qu'il n'en est rien puisque nous nous intéressons à M. D*** justement à cause du résultat de l'expérience survenu chez lui). M. D*** devrait donc être amené à remettre en question la validité de son hypothèse selon laquelle M. M*** n'est qu'un charlatan ? Eh bien oui ! Cependant, M.D*** doit tenir compte de l'ensemble des éléments dont il dispose, qui, vraisemblablement, l'incitent très fortement à douter de la réalité des pouvoirs de M. M***, et le résultat de l'émission ne constitue donc pas nécessairement un argument très fort en faveur des pouvoirs parapsychologiques. (Le raisonnement bayésien fournirait par exemple un cadre pour quantifier ceci de manière précise.)

Si M. D*** cherche à aborder les choses de manière systématique, il tentera à nouveau l'expérience (en admettant que M. M*** réapparaisse plusieurs fois à la télévision, ou que M. D*** va jusqu'à inviter M. M*** chez lui pour en avoir le cœur net), pour constater que M. M*** ne réussit presque jamais. Ou encore, il s'informera des résultats constatés chez un grand nombre de personnes (pas seulement chez celles ayant contacté l'émission, sous peine de biais de sélection, mais au sein d'un échantillon représentatif), pour constater que M. M*** n'a réussi chez quasiment aucune d'entre elles. Cette situation est bien entendu un peu caricaturale, car peu de gens prennent au sérieux les parapsychologues et autres tordeurs de petites cuillères, mais le même genre de phénomène peut apparaître dans bien d'autres contextes. Imaginons par exemple que 50 équipes scientifiques étudient séparément l'impact d'un nouveau produit, disons la vitamine X, sur la guérison d'une maladie, par exemple le cancer. Chaque équipe conduit son étude dans les règles (essais randomisés en double aveugle, échantillons représentatifs de la population à traiter, constitution de groupes témoins et utilisation de placebos). Sur les 50 équipes, 49 observent des résultats non-concluants quant à l'efficacité du médicament. En revanche l'une des

équipes observe un taux de guérison si élevé chez les patients traités à l'aide de la vitamine X, que, sous l'hypothèse que la vitamine X est sans effet sur le cancer, on ne puisse espérer observer un tel taux qu'avec une probabilité d'environ 2%. L'équipe en question, qui travaille seule, estimera avoir de bonnes raisons de penser que la vitamine X possède un effet réel sur le cancer !

1.9 Auto-évaluation

- Qu'est-ce qu'un modèle probabiliste (en tant qu'objet mathématique) ?
- Que représente concrètement l'espace des possibles ?
- Quelles sont les différentes traductions concrètes de la notion de probabilité ?
- Donnez au moins trois sens nettement différents de la notion de probabilité, assortis d'exemples dans chacun des cas.
- Tout ce qui est *a priori* susceptible de varier dans une expérience aléatoire figure-t-il explicitement dans le modèle ?
- Qu'est-ce qu'un événement formel dans le cadre d'un modèle probabiliste ?
- Quel lien y a-t-il entre événement concret et événement formel ?
- Un événement concret est-il toujours associé à un événement formel ?
- Comment définit-on la probabilité d'un événement à partir de la probabilité associée aux éléments de l'espace des possibles ?
- Y a-t-il en général un ou plusieurs modèles probabilistes susceptibles de décrire la même situation ? Quelles peuvent être les différences ? Que représentent-elles ?
- Que signifie la compatibilité de deux modèles ?
- Y a-t-il toujours compatibilité entre un modèle plus fin et un modèle moins fin d'une même situation ?
- Qu'est-ce qu'un modèle plus fin qu'un autre ?
- A-t-on toujours l'égalité $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ (faire un dessin) ? Sinon, quelle est la relation qui a lieu en général entre ces deux quantités ? Qu'en est-il avec n événements $A_1, \ldots, A_n$ au lieu de deux ?
- Comment est définie (formellement) la probabilité conditionnelle ?
- Que représente-t-elle concrètement ?
- Comment calcule-t-on la probabilité d'un événement conditionnellement à un autre ?
- Qu'est-ce que l'effet de loupe probabiliste (donnez au moins un exemple) ?
- Que signifie la dépendance de deux événements ?
- Que signifie l'indépendance de deux événements ?
- Différence(s) entre relation de cause à effet et dépendance probabiliste entre événements ?
- Dans un modèle en arbre, que représentent les feuilles de l'arbre ?

- Dans un modèle en arbre, que représentent les ramifications de l'arbre ?
- Dans un modèle en arbre, que représentent les nombres associés aux arêtes ?
- Dans un modèle en arbre, comment définit-on la probabilité sur Ω ? Comment se calcule-t-elle ?
- Comment modélise-t-on une succession indépendante d'épreuves ?
- Qu'est-ce que le théorème des coalitions ?
- Que sont n événements mutuellement indépendants ?
- Doit-on être surpris si l'issue effectivement réalisée d'une expérience est très improbable dans le modèle dont on dispose ? Devons-nous alors modifier le modèle ?

1.10 Exercices

Exercice 1 *Chaque matin, au réveil, Jojo peut se livrer (ou non) à chacune des activités suivantes :*

- *se laver*
- *se brosser les dents*
- *boire un café*
- *écouter la radio*
- *se raser*

Décrire un espace des possibles permettant de modéliser les activités matinales de Jojo.

Appelons A l'événement «Jojo se rase», B l'événement «Jojo se brosse les dents», C l'événement «Jojo écoute la radio». Décrire les événements formels correspondants à ces événements dans l'espace des possibles que vous avez choisi.

Exprimer à l'aide des événements A, B, C les événements suivants :

- *ce matin, Jojo se brosse les dents mais n'écoute pas la radio*
- *ce matin, Jojo n'écoute pas la radio mais se brosse les dents*
- *ce matin, Jojo boit un café ou se rase, mais n'écoute pas la radio*
- *ce matin, Jojo, ou bien se rase, ou bien se brosse les dents, et dans tous les cas écoute la radio*
- *ce matin, Jojo, ou bien se rase, ou bien se brosse les dents et écoute la radio*
- *ce matin, Jojo, ou bien se rase et n'écoute pas la radio, ou bien se brosse les dents et écoute la radio*
- *ce matin, Jojo ne se rase pas, ou ne se brosse pas les dents ni n'écoute la radio*
- *ce matin, Jojo se rase ou se brosse les dents ou écoute la radio*
- *ce matin, Jojo ne se rase pas, ou bien il écoute la radio et se brosse les dents*
- *ce matin, Jojo ne se rase pas, ou bien il se rase et écoute la radio*

Exercice 2 *Dans un espace de probabilité $(\Omega, \mathbb{P})$, soient deux événements A, B tels que $A \subset B$. Prouvez que $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$.*

Qu'en est-il si l'on ne suppose plus que $A \subset B$?

Exercice 3 *Pour chacun des exemples ci-dessous, expliquer comment on doit, selon-vous, interpréter la probabilité qui y apparaît, et comment (très grossièrement) on pourrait tenter d'évaluer celle-ci (et donc confirmer ou infirmer les valeurs proposées).*

- *«La probabilité pour que le candidat A soit élu lors de la prochaine élection présidentielle est de 60%.»*
- *«La probabilité pour que la pièce de monnaie tombe sur face est de 50%.»*
- *«La probabilité pour que l'équipe de football du Brésil l'emporte demain face à l'Allemagne est de 1/4.»*
- *«La probabilité pour qu'il pleuve demain à Lyon est de 1/3.»*
- *«La probabilité pour qu'il ait plu il y a exactement 3000 ans sur le site aujourd'hui occupé par Lyon est de 1/3.»*
- *«La probabilité pour qu'une météorite de plus de 500m de diamètre de circonférence percute la terre au cours du prochain millénaire est de moins de 2%.»*
- *«La probabilité pour que la fusée explose au décollage est de moins de 2%.»*
- *«La probabilité, pour un individu né en France en 1920, de vivre plus de 80 ans est de 75%.»*
- *«La probabilité pour un individu né en France en 1954, de vivre plus de 80 ans est de 85%.»*
- *«La probabilité pour un individu né en France en 1954 de posséder un chien est de 60%.»*
- *«La probabilité pour que D*** (qui est né en France en 1954) possède un chien est de 70%.»*
- *«La probabilité pour qu'un atome de carbone 14 subisse une désintégration au cours des 5730 prochaines années est de 50%.»*
- *«La probabilité pour qu'un photon incident émis par la source S soit absorbé par le détecteur D est de 1/3.»*
- *«La probabilité pour que l'épidémie se propage est de 5%.»*
- *«La probabilité pour qu'un paquet de données mette plus de 0,1 seconde pour être transmis dans le réseau est de 10%.»*
- *«La probabilité pour que l'enfant à naître soit une petite fille est de 1/2.»*
- *«La probabilité pour que la croissance du PIB soit cette année supérieure à 2%, est de 70%.»*

Exercice 4 *Trois candidats, appelons-les A, B et C, se présentent à l'élection présidentielle du Jojoïstan. A l'issue du premier tour, le candidat obtenant le moins de*

voix sera éliminé, et le candidat présent au second tour obtenant le plus de voix (et donc la majorité) sera élu président. Un institut de sondage a réalisé une enquête sur l'état de l'opinion publique au Jojoïstan en demandant à 10 000 personnes de classer par ordre de préférence décroissant les trois candidats. Les réponses obtenues (en pourcentages) se répartissent de la façon suivante :

ABC	*ACB*	*CAB*	*CBA*	*BAC*	*BCA*
19%	*16%*	*25%*	*8%*	*10%*	*22%*

Quel est le pourcentage d'individus qui préfèrent A à B ? B à C ? A à C ? En supposant que le sondage reflète fidèlement les intentions de vote des électeurs, quel candidat obtiendra le plus de voix à l'issue du premier tour ? Quel sera le candidat éliminé au premier tour ? Quel sera le candidat finalement élu ?

Fabriquez vous-même un exemple de répartition des préférences pour lequel le pourcentage d'individus qui préfèrent B à A dépasse 50%, le pourcentage d'individus qui préfèrent B à C dépasse 50%, mais pour lequel le candidat finalement élu n'est pas B.

Exercice 5 *Vérifiez que la probabilité $\mathbb{P}$ définie sur*

$$\Omega = \{0,1\}^{16} = \{(x_1, \ldots, x_{16}) \ : \ x_i \in \{0,1\}\},$$

par :

$$\mathbb{P}\left[(x_1, \ldots, x_{16})\right] = \prod_{i=1}^{16} p^{x_i}(1-p)^{1-x_i},$$

en est bien une. Montrez que ce modèle apparaît comme une succession d'épreuves indépendantes.

Exercice 6 *Pour prévoir le temps qu'il fera demain, Alfred se base en partie sur les mouvements de sa grenouille. Béatrice, elle, se fie plutôt aux prévisions de la météorologie nationale. Finalement, Alfred utilise le modèle suivant pour décrire le temps :*

$$\Omega_A = \{\textit{Haut, Milieu, Bas}\} \times \{\textit{Beau, Maussade, Pluvieux}\},$$

avec la probabilité $\mathbb{P}_A$ définie par :

	Haut	*Milieu*	*Bas*
Beau	*1/30*	*2/30*	*7/30*
Maussade	*2/30*	*6/30*	*2/30*
Pluvieux	*7/30*	*2/30*	*1/30*

Béatrice, elle, utilise le modèle :

$\Omega_B = \{$*Beau prévu, Maussade prévu, Pluvieux prévu*$\}\times\{$*Beau, Maussade, Pluvieux*$\}$,

avec la probabilité $\mathbb{P}_B$ *définie par :*

	Beau prévu	*Maussade prévu*	*Pluvieux prévu*
Beau	*3/15*	*1/15*	*1/15*
Maussade	*1/15*	*3/15*	*1/15*
Pluvieux	*1/15*	*1/15*	*3/15*

Enfin, César se contente du modèle plus simple défini par

$$\Omega_C = \{\textit{Beau, Maussade, Pluvieux}\},$$

avec

$$\mathbb{P}_C(\textit{Beau}) = 1/4,\ \mathbb{P}_C(\textit{Maussade}) = 1/2,\ \mathbb{P}_C(\textit{Pluvieux}) = 1/4.$$

Expliquez comment Alfred et Béatrice peuvent exploiter leurs modèles respectifs pour estimer les probabilités relatives au temps qu'il fera demain à partir des informations fournies par la grenouille et la météo nationale respectivement. Donnez des représentations en arbre des modèles $(\Omega_A, \mathbb{P}_A)$, $(\Omega_B, \mathbb{P}_B)$, $(\Omega_C, \mathbb{P}_C)$. *Ces modèles sont-ils compatibles ?*

Exercice 7 *Considérons un modèle probabiliste* $(\Omega, \mathbb{P})$ *et deux événements* A *et* B *tels que* $A \cap B$ *soit de probabilité non-nulle. Notons* $\mathbb{P}_A$ *la probabilité* $\mathbb{P}(\cdot|A)$, *et* $\mathbb{P}_B$ *la probabilité* $\mathbb{P}(\cdot|B)$. *Montrez que*

$$\mathbb{P}_A(\cdot|B) = \mathbb{P}_B(\cdot|A) = \mathbb{P}(\cdot|A \cap B).$$

Que signifie ce résultat ?

Exercice 8 *Jojo fait du ski à la station «Vallées blanches». Il est en haut du téléski des cailloux, et a le choix entre les pistes de Tout-Plat (une bleue), Les-Bosses (une rouge) et Rase-Mottes (une noire). Il va choisir entre ces trois pistes au hasard, de telle façon qu'il choisisse la bleue ou la noire avec probabilité 1/4, et la rouge, qu'il préfère, avec probabilité 1/2. Il descend ensuite la piste choisie. Jojo n'est pas encore très à l'aise cette saison, et il tombe avec une probabilité de 0,1 sur la piste bleue, de 0,15 sur la piste rouge, et de 0,4 sur la piste noire.*
1) Soit A l'événement «Jojo tombe en descendant la piste qu'il a choisie». Calculer $\mathbb{P}(A)$.
2) Bernard, qui attend Jojo en bas des pistes, à la terrasse d'un café, voit arriver Jojo couvert de neige : il est donc tombé. Sachant cela, quelle est la probabilité qu'il ait emprunté la piste noire ?

Exercice 9 *Dans le film «Willow» (Ron Howard, 1988), un sorcier met à l'épreuve trois jeunes gens pour décider lequel sera son apprenti. L'un après l'autre, chacun des trois candidats doit désigner (en public) un doigt de la main du sorcier comme étant le principal dans l'exercice de la magie. Le premier à donner la bonne réponse sera choisi comme apprenti, le ou les suivants étant éliminés, et, si aucun ne fournit la réponse correcte, le sorcier ne prendra aucun apprenti (c'est d'ailleurs ce qui se produit dans le film). Décrire (et justifier) un modèle probabiliste de cette situation, puis proposer une réponse à la question : quel candidat, du premier, du deuxième ou du troisième, a le plus de chances d'être choisi ?*

Exercice 10 *Considérons n événements $A_1, \ldots, A_n$ d'un modèle probabiliste $(\Omega, \mathbb{P})$. Pour $1 \leq k \leq n$, posons*

$$C_k = \sum_{i_1 < \cdots < i_k} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}).$$

Prouvez les relations suivantes :

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) \leq \sum_{k=1}^{m} (-1)^{k-1} C_k$$

lorsque m est impair,

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) \geq \sum_{k=1}^{m} (-1)^{k-1} C_k$$

lorsque m est pair, et

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) = \sum_{k=1}^{n} (-1)^{k-1} C_k.$$

Exercice 11 *Des amis de Jojo lui proposent de participer au jeu suivant : on dispose de trois cartes dont l'une a deux faces blanches, l'autre deux faces rouges, et la troisième une face blanche et une face rouge. Une carte est choisie au hasard parmi les trois et seule l'une des faces, choisie au hasard elle aussi, est exposée au public. Il s'agit de parier sur la couleur de la face cachée. Après quelques secondes de réflexion, Jojo se dit que, si la face exposée est, par exemple, rouge, la carte tirée peut être soit celle dont les deux faces sont rouges, soit celle dont une face est rouge et l'autre blanche, et qu'il y a donc une chance sur deux pour que la face cachée soit rouge, et une chance sur deux pour qu'elle soit blanche. De même lorsque la face exposée est blanche... Il décide donc de répondre de manière aléatoire «rouge» ou «blanc» avec probabilité 1/2 pour chaque réponse. Son raisonnement et sa méthode vous semblent-ils judicieux ? Faites l'expérience par vous-même, en comparant par exemple la stratégie de Jojo à la stratégie qui consiste à parier systématiquement sur la même couleur que celle qui figure sur la face visible de la carte. A vos jeux !*

Exercice 12 *Lançons deux dés à six faces, et considérons les deux événements $A =$ «la somme des deux chiffres obtenus est paire» et $B =$ «le 1 ou le 2 sort au moins une fois». Si l'on munit l'espace des possibles*

$$\Omega = \{1; 2; 3; 4; 5; 6\} \times \{1; 2; 3; 4; 5; 6\}$$

de la probabilité uniforme, montrez que A et B sont indépendants. Si, au contraire, la probabilité sur Ω est donnée par le tableau suivant :

	1	*2*	*3*	*4*	*5*	*6*
1	*2/46*	*1/46*	*2/46*	*1/46*	*2/46*	*1/46*
2	*1/46*	*2/46*	*1/46*	*2/46*	*1/46*	*2/46*
3	*2/46*	*1/46*	*1/46*	*1/46*	*1/46*	*1/46*
4	*1/46*	*2/46*	*1/46*	*1/46*	*1/46*	*1/46*
5	*2/46*	*1/46*	*1/46*	*1/46*	*1/46*	*1/46*
6	*1/46*	*2/46*	*1/46*	*1/46*	*1/46*	*1/46*

montrer que ces mêmes événements A et B ne sont pas indépendants.

Exercice 13 *Un assassinat vient d'être commis, et les suspects se limitent à un ensemble de n personnes. Initialement, chaque suspect possède a priori une probabilité de $1/n$ d'être le coupable. L'enquêteur affirme : «au vu des circonstances, il semble clair que l'assassin doit être gaucher». Après réflexion, il précise sa pensée en affirmant : «la probabilité que l'assassin soit gaucher est de 80%». Construire un modèle probabiliste de la situation et traduire ces affirmations.*

Exercice 14 *Un magasin d'équipement de la personne vend (entre autres choses) des lunettes de soleil et des parapluies. Une étude statistique menée sur plusieurs années concernant les ventes de ces deux produits a révélé que, lors de la saison printemps-été, on peut résumer les résultats des ventes journalières par le modèle probabiliste suivant :*

	moins de 5 paires de lunettes	*entre 5 et 10 paires*	*plus de 10 paires*
moins de 7 parapluies	*9/40*	*9/40*	*18/40*
plus de 7 parapluies	*1/40*	*1/40*	*2/40*

Lors de la saison automne-hiver, en revanche, on peut décrire les résultats des ventes journalières par le modèle probabiliste suivant :

	moins de 5 paires de lunettes	*entre 5 et 10 paires*	*plus de 10 paires*
moins de 7 parapluies	*3/20*	*1/20*	*1/20*
plus de 7 parapluies	*9/20*	*3/20*	*3/20*

En admettant que le nombre de jours d'ouverture du magasin est le même pour les deux saisons, quel serait le modèle probabiliste permettant de décrire les ventes pour toute l'année ? Les ventes de lunettes et de parapluies y apparaissent-elles comme indépendantes ? Qu'en est-il pour chacun des deux modèles pris séparément ? Qu'en conclure ?

Exercice 15 *Pour dépister une maladie, on effectue un test sanguin. Si le patient est effectivement atteint, le test donne un résultat positif avec une probabilité de 99%. Si le patient est sain, le test peut malheureusement donner un résultat positif avec une probabilité de 2%. Nous supposerons que la probabilité d'être frappé par la maladie est de x% pour un patient se présentant au dépistage. Sachant que le test donne un résultat positif, comment se comporte la probabilité que le patient soit effectivement malade en fonction de x ?*

Exercice 16 *Près de soixante pour cent des accidents de voiture graves impliquant de jeunes enfants se produisent dans des véhicules où les enfants ne sont pas correctement attachés. Soixante pour cent, cela fait beaucoup... A quoi faudrait-il comparer ce chiffre ?*

Exercice 17 *Construisez un exemple de modèle probabiliste pour vous convaincre qu'en général*

- $\mathbb{P}(A|B^c) \neq 1 - \mathbb{P}(A|B)$.
- $\mathbb{P}(A|B) + \mathbb{P}(A|B^c) \neq \mathbb{P}(A)$

Expliquez pourquoi ces inégalités sont en général vérifiées.

Exercice 18 *Blanche-Neige passe la serpillière quand la méchante reine se présente, grimée en pauvre vieille, pour lui offrir un panier de cinq pommes bien rouges, dont une empoisonnée et deux véreuses. Blanche-Neige prend les pommes une par une pour les croquer. Si elle tombe sur une pomme véreuse, elle jette le reste des pommes au cochon, sinon elle continue. Evaluer la probabilité pour que :*
a) le porc trépasse
b) Blanche-Neige mange toutes les bonnes pommes.

Exercice 19 *Environ 10% des ouvrages publiés par un éditeur parisien, dont nous tairons le nom, sont des succès de librairie. Afin d'estimer avant sa parution le succès d'un livre, l'éditeur a pour habitude de consulter deux experts différents. Consulté sur un nouvel ouvrage, le premier expert prédit un brillant succès, tandis que le second rend un avis plutôt négatif, et annonce un échec probable.*

On affirme couramment que le jugement du premier expert est fiable à près de 90%, tandis que le jugement du second ne le serait qu'à près de 70%.

Comment de telles affirmations pourraient-elles être traduites, et exploitées pour évaluer la probabilité que le nouvel ouvrage soit un succès ?

L'éditeur dispose en fait, pour une centaine d'ouvrages qu'il a publiés au cours des années précédentes, des avis émis à l'époque par les deux experts, ainsi que des chiffres de vente des ouvrages correspondants.

Que feriez-vous, par exemple, de l'information selon laquelle le premier expert a vu son pronostic confirmé dans 80% des cas, tandis que le second n'a eu raison que dans 60% des cas ?

Et si l'on précise que le premier expert a vu son pronostic confirmé dans 70% des cas lorsqu'il prédisait un succès, et le second également dans 70% des cas, mais lorsqu'il prédisait un échec ?

Quelles autres informations relatives aux prédictions des experts pourrait-on chercher à exploiter en se basant sur les données de l'éditeur, et de quelle manière ?

Exercice 20 *Jojo participe à un jeu télévisé fondé sur le principe suivant. Derrière trois portes fermées se trouvent respectivement une peluche, une barre chocolatée, et un chèque de 5000 euros. Jojo (qui préfère gagner le chèque) doit, pour commencer, désigner l'une des trois portes. Cette porte désignée, le présentateur (qui sait, quant à lui, quels lots se trouvent derrière quelles portes) ouvre l'une des deux autres portes, révélant ainsi le lot qui se trouve derrière. Jojo peut alors choisir, soit de prendre le lot en question, soit de tenter à nouveau sa chance en demandant l'ouverture de l'une des deux portes restantes et gagner le lot situé derrière ladite porte. S'il choisit de retenter sa chance, Jojo est donc placé devant l'alternative suivante : demander l'ouverture de la porte qu'il avait initialement désignée, ou de l'autre porte demeurant fermée (qu'il n'avait pas désignée initialement, et que le présentateur n'a pas ouverte). Il se dit que face à deux portes que rien ne semble distinguer, il a une chance sur deux de trouver le chèque derrière la porte qu'il avait initialement choisie, et une chance sur deux de le trouver derrière l'autre porte. Afin de ne pas regretter d'avoir modifié un choix initial potentiellement gagnant (ce que Jojo estime pire que de perdre en demeurant fidèle à sa première impulsion), Jojo projette donc de maintenir son premier choix. Que pensez-vous du raisonnement de Jojo ?*

Exercice 21 *La duchesse d'Aquitaine et la duchesse de Bourgogne attendent chacune l'héritier de leur duché. On décrit la situation à l'aide du modèle probabiliste suivant :*

$$\Omega = \{GG, GF, FG, FF\},$$

où la première lettre indique le sexe de l'héritier d'Aquitaine, et la seconde celui de l'héritier de Bourgogne, et la probabilité sur Ω est la probabilité uniforme. Considérons les trois événements :

- $A =$ *«l'héritier d'Aquitaine est un garçon»,*
- $B =$ *«l'héritier de Bourgogne est une fille»,*
- $C =$ *«les deux héritiers sont de sexe opposé».*

Les événements A et C sont-ils indépendants ? Et B et C ? et A et C ? Et $A \cap B$ et C ? Est-ce surprenant ?

Exercice 22 *Au casino de Jojo-les-bains, les machines à sous sont scandaleusement truquées. En effet, le mécanisme qui gouverne le fonctionnement des trois rouleaux (comportant chacun 30 signes différents) est le suivant : la position sur laquelle s'arrête le rouleau le plus à gauche est effectivement choisie au hasard, mais le second rouleau et le troisième s'arrêtent automatiquement sur des positions qui présentent un décalage fixé (par exemple 3 positions en plus dans le sens de rotation des rouleaux pour le rouleau du milieu, et 7 pour le rouleau de droite) par rapport au premier rouleau, ce qui fait que personne ne gagne jamais. A la suite de nombreuses plaintes des clients, le patron du casino organise une série d'expériences publiques desquelles il ressort que chaque signe de chacun des trois rouleaux sort approximativement une fois sur 30. Aucun signe n'est donc favorisé par rapport à un autre. Que pensez-vous de cet argument ?*

Exercice 23 *Jojo a mis au point un algorithme randomisé pour tester si un entier est premier. L'algorithme prend en entrée un entier, effectue au cours de son exécution un certain nombre de tirages aléatoires, et donne en sortie une réponse binaire : «premier» ou «composé». Lorsque l'entier testé est effectivement premier, l'algorithme répond toujours «premier». En revanche, lorsque celui-ci est composé, l'algorithme répond «composé» avec une probabilité variable, comprise entre 20% et 100% (pour des exemples de tels algorithmes, par exemple le test de Miller-Rabin, voir par exemple l'ouvrage de Motwani et Raghavan cité dans la bibliographie. Mentionnons simplement ici l'idée générale consistant à exploiter la propriété dite d'«abondance de témoins»* [8] *).*

Comment faire pour déterminer avec un minimum de confiance si un entier est premier en utilisant l'algorithme de Jojo ?

Gégé, moins ingénieux que Jojo, a mis au point un algorithme qui donne une réponse correcte avec une probabilité d'au moins 81% lorsque p est premier, et d'au moins 20% lorsque p est composé. Même question avec l'algorithme de Gégé.

Plus généralement, reprendre la question en supposant que l'on sait que la probabilité de donner une réponse correcte lorsque p est premier est comprise dans une

8. Plus précisément, on supposera définis, pour tout n, un ensemble fini A_n et une application $P : A_n \to \{0,1\}$ tels que, si pour un $i \in A_n$, $P(n,i) = 1$, on peut être certain que n est composé : on dit alors que i témoigne du fait que n est composé. L'idée est que, pour certains choix judicieux des ensembles A_n et de la propriété P, on peut prouver que, pour tout nombre composé n, la proportion d'éléments de A_n qui sont des témoins du fait que n est composé, dépasse une limite inférieure fixée (par exemple 20%). Il suffit alors d'effectuer un tirage aléatoire selon la probabilité uniforme dans A_n pour pouvoir détecter le fait que n est composé avec une probabilité supérieure à cette limite.

certaine fourchette $[\alpha_1, \alpha_2]$ *et que la probabilité de donner une réponse correcte lorsque* p *est composé est comprise dans une certaine fourchette* $[\beta_1, \beta_2]$.

Exercice 24 *Ce soir, Jojo doit se rendre à une soirée très chic, et il hésite quant à la façon de s'habiller. Il a le choix entre le traditionnel smoking (passe-partout, mais qui ne l'enthousiasme guère), son costume hyper-branché à franges luminescentes (qui l'amuse beaucoup plus), et sa tenue de tous les jours (tout de même beaucoup plus confortable, mais pas très présentable). Il sera refoulé à l'entrée avec probabilité 0,1 s'il porte le smoking, 0,3 avec son costume branché, et 0,7 avec sa tenue ordinaire. Ne parvenant pas à choisir, il décide de s'en remettre au hasard en lançant deux dés équilibrés à six faces. Si le maximum des deux dés est égal à 6, il mettra son costume de tous les jours. S'il est égal à 4 ou 5, il mettra son costume branché, et son smoking dans tous les autres cas.*

Les heures passent, et les amis de Jojo, qui l'attendent dans la salle où la soirée se déroule, ne le voient pas arriver : il a donc malheureusement été refoulé à l'entrée. Comment, dans ces conditions, évaluer la probabilité pour que Jojo ait mis son costume branché ? Même question avec la probabilité pour que l'un des deux dés ait donné un 3 ?

Exercice 25 *Un revendeur d'informatique lyonnais reçoit une livraison d'écrans. Le lot peut soit provenir d'un fournisseur japonais, qui produit en moyenne une pièce défectueuse sur 1000, soit d'un fournisseur malais, qui produit en moyenne une pièce défectueuse sur 200. Le fournisseur teste soigneusement l'un des écrans du lot, et ne constate aucun défaut. Comment évaluer la probabilité pour que le lot provienne du fournisseur malais ? Comment évaluer la probabilité pour que le second écran testé ne présente pas non plus de défaut ? Y a-t-il indépendance entre le fait que le premier écran soit défecteux et le fait que le deuxième le soit ?*

Exercice 26 *M. D***, particulièrement inquiet des risques d'attentat à la bombe lors de ses nombreux voyages en avion, a adopté pour se rassurer la solution suivante : il emporte toujours avec lui dans ses bagages une bombe (indétectable). Selon lui, la probabilité pour que deux bombes se trouvent à bord d'un même avion est absolument négligeable. Que pensez-vous de ce raisonnement ? (Justifiez).*

Exercice 27 *Les 52 cartes d'un jeu (sans joker) sont réparties au hasard en tas de 4 cartes, sur 13 emplacements numérotés à l'aide des indices 2, 3 ,..., 10, Valet, Dame, Roi, 1. La répartition effectuée, on procède aux opérations suivantes.*

1. *initialisation : indice-tas-courant* $\leftarrow$ *1 ;*
2. *si le tas numéroté par indice-tas-courant n'est pas vide, enlever du jeu la carte située au sommet de ce tas, sinon STOP ;*

3. *indice-tas-courant ← figure indiquée sur la carte que l'on vient d'enlever ;*
4. *retourner en 2.*

Décrivez un modèle probabiliste simple de la situation, et calculez dans ce modèle la probabilité pour que l'on ne s'arrête qu'une fois que toutes les cartes du jeu ont été examinées.

Exercice 28 *Prenez une grande respiration et... écrivez rapidement le résultat d'une suite de 200 répétitions (imaginaires) de lancers de pile/face indépendants et non-biaisés. Etes-vous satisfait du résultat ?*

Exercice 29 *Pour chacune des affirmations suivantes, commencez par indiquer quelle peuvent être la population témoin et la population test, ainsi qu'une définition précise possible de la dépendance qui est mentionnée. Discutez ensuite la présence possible de liens de cause à effet, de causes cachées et de facteurs de confusion dans chacun des cas.*

- *en Italie, on a constaté que les régions dans lesquelles les taux d'achat d'ordinateur personnels sont les plus importants sont également celles où les taux de divorce sont les plus élevés ;*
- *une étude japonaise portant sur 40000 quadragénaires montre que ceux qui se brossent les dents après chaque repas parviennent mieux que les autres à garder la ligne ;*
- *il existe une association positive entre utilisation de crème solaire et cancer de la peau ;*
- *on constate qu'au cours d'une année, un nombre élevé de noyades enregistrées est positivement associé à une consommation élevée de crèmes glacées ;*
- *sur une longue période, on constate une association négative entre un prix élevé des cigarettes et un nombre élevé d'agriculteurs en Lozère ;*
- *en Ecosse, des achats importants de whisky sont positivement associés à la réception de dons importants par les églises ;*
- *la carte du vote Le Pen lors des élections présidentielles de 2002 se superpose avec celle de l'irradiation due au nuage de Tchernobyl ;*
- *dans les communes qui abritent des cigognes, la natalité est plus élevée que dans le reste du pays ;*
- *une confiance élevée des investisseurs est positivement associée à une forte croissance économique ;*
- *sur une vaste population, on constate que la consommation régulière d'alcool pendant la grossesse est associée à des retards de QI et des difficultés d'apprentissage chez les enfants ;*
- *au cours du temps, un volume élevé des recettes publiques allemandes est positivement associé à un volume élevé de dépenses des ménages espagnols ;*

- *sur un ensemble de villes françaises, on constate qu'une proportion élevée de fonctionnaires est négativement associée à une économie locale dynamique ;*
- *les enfants P*** acceptent plus volontiers les repas lorsqu'ils sont préparés par leur père que par leur mère ;*
- *la présence d'un médecin obstétricien lors d'un accouchement accroît la probabilité de complications ;*
- *le fait d'avoir recours à la péridurale diminue la mortalité lors des accouchements ;*
- *un nombre élevé d'écoles maternelles dans une ville est positivement associé à un nombre élevé de crimes et délits ;*
- *les entreprises réalisant le plus de bénéfices sont celles qui ont les budgets publicitaires les plus importants ;*
- *un viticulteur diffuse de la musique classique dans son vignoble, et l'on constate que le vin obtenu est meilleur que celui produit par ses voisins, qui disposent pourtant de parcelles comparables pour l'ensoleillement et la nature du sol ;*
- *une faible cholestérolémie favorise l'apparition du cancer ;*
- *le fait de consommer régulièrement des moules accroît le risque d'attraper la grippe.*

Exercice 30 *On s'intéresse à la modélisation d'une enquête statistique effectué dans une population (il peut s'agir, par exemple, d'une enquête téléphonique sur les opinions politiques, ou encore d'une enquête sur des traitements médicaux menée en milieu hospitalier, etc...).*

Dans l'idéal, les individus constituant l'échantillon sondé sont choisis uniformément et indépendamment au sein de la population.
1) Qu'entend-on selon vous lorsque l'on dit qu'un tel échantillon est représentatif de la population ?
2) Effectuez une liste des différentes raisons, théoriques et pratiques, qui, selon vous, tendent à faire s'écarter un sondage réel de la situation idéale décrite par ce modèle. Cherchez des exemples concrets pour étayer votre liste.
3) Comment obtenir en pratique des échantillons représentatifs ? Comment tester la représentativité d'un échantillon donné ? Comment corriger (éventuellement) les résultats obtenus à partir d'un échantillon non-représentatif ?
4) Un réseau d'agences immobilières communique régulièrement à la presse un indice des prix obtenu à partir des transactions réalisées par les agences de ce réseau. Dans le document décrivant la méthodologie statistique retenue pour construire cet indice, le nombre important de transactions utilisées pour construire l'indice est souligné, et l'on trouve par ailleurs une brève mention du fait que les transactions sur lesquelles l'indice est basé sont communiquées par les différentes agences sur la base du volontariat. Dans ce exemple, le fait de décrire précisément la méthodologie statistique

retenue vous semble-t-il un gage suffisant de fiabilité des résultats présentés ?

Exercice 31 *Dans le domaine médical, pour tester l'efficacité d'un traitement, on procède idéalement de la manière suivante : un échantillon représentatif de la population sur laquelle on envisage d'utiliser le traitement étant choisi (pour éviter le biais de sélection, voir Exercice 30), on répartit aléatoirement les individus de l'échantillon entre deux groupes : un groupe dans lequel le traitement est administré, et un groupe témoin auquel est administré un placebo. On procède de plus, autant que possible à des essais en «double aveugle», ni les cobayes, ni le personnel encadrant l'étude ne sachant qui a reçu le traitement et qui a reçu un placebo. Expliquez en quoi cette méthode peut en principe éliminer les dépendances dues à une cause cachée ou un facteur de confusion.*

Pour une introduction plus détaillée, mais non-technique, à ces questions, nous vous recommandons la lecture de l'excellent ouvrage de Schwartz cité dans la bibliographie.

Exercice 32 *Antoinette, trente et un ans, est une célibataire élégante qui a son franc-parler. Ce fut une étudiante brillante. A l'époque de ses études, elle milita pour le droit de vote des immigrés et prit part à des manifestations en faveur de la mise en place de crèches dans les administrations. Classez les jugements suivants par ordre de probabilité décroissante (les ex-æquo sont possibles).*

- *Antoinette est une féministe militante.*
- *Antoinette est caissière dans une banque.*
- *Antoinette travaille dans une petite librairie.*
- *Antoinette est caissière dans une banque et féministe militante.*
- *Antoinette est caissière dans une banque, féministe militante, et pratique le yoga.*
- *Antoinette est une féministe militante qui travaille dans une petite librairie et pratique le yoga.*

Exercice 33 *Trois amies, Alice, Bénédicte, et Claire effectuent des stages d'été dans trois pays différents : Alice aux Etats-Unis, Bénédicte au Canada, et Claire en Angleterre. La probabilité de subir un cambriolage l'été est évaluée à 60% aux Etats-Unis, 10% au Canada, et 40% en Angleterre (ces chiffres sont totalement fictifs).*

L'une des trois amies est cambriolée au cours de son stage. Comment évaluer la probabilité qu'il s'agisse de Claire ?

Exercice 34 *Pour deux compagnies aériennes A et B, les tableaux suivant indiquent les nombres totaux de vols effectués à destination de Paris et de Lyon en 2004, ainsi que le nombre de ceux qui sont arrivés sans retard.*

Pour la compagnie A :

	Total	*A l'heure*
Paris	*600*	*534*
Lyon	*250*	*176*

Pour la compagnie B :

	Total	*A l'heure*
Paris	*200*	*188*
Lyon	*900*	*685*

Sur la base de ces données, quelle est la compagnie dont, en 2004, les vols à destination de Paris ont la plus forte probabilité d'arriver à l'heure ? Et pour Lyon ? Et de manière globale ? Le résultat est-il surprenant ?

Exercice 35 *Au détour d'une conversation avec M. D***, celui-ci vous apprend qu'il a deux enfants, dont au moins une fille. Comment évalueriez-vous la probabilité pour que l'autre enfant soit une fille ?*

*Même question si M. D*** ajoute que la fille en question se prénomme Sophie.*

Exercice 36 *Une (longue) liste de N enregistrements $x_1, \ldots, x_N$ nous est communiquée en temps réel, un enregistrement après l'autre. On souhaite en extraire une sous-liste non-ordonnée comportant n enregistrements (avec $n << N$), choisie uniformément au hasard parmi toutes les sous-listes non-ordonnées formées de n enregistrements.*

a) Montrer que si l'on sait résoudre ce problème pour des sous-listes ordonnées, on peut le résoudre pour des sous-listes non-ordonnées.

b) Une première solution pour y parvenir consiste simplement à stocker la totalité de la liste dans un fichier, puis à extraire une sous-liste du fichier en question. Comment peut-on procéder exactement ?

c) Supposons que l'on ne souhaite pas stocker la totalité de la liste, mais simplement décider séquentiellement, lorsqu'un élément de la liste nous est communiqué, de l'inclure ou non dans la sous-liste que l'on cherche à produire. Une approche naïve consisterait à accepter indépendamment chaque élément avec une probabilité égale à n/N, mais cette approche échoue à résoudre exactement le problème initial. Précisez pourquoi. Quelles sont les probabilités d'acceptation permettant de résoudre exactement le problème initial ? Indication : conditionnellement aux décisions d'inclusion/non-inclusion relatives aux éléments $x_1, \ldots, x_k$, comparez les probabilités d'inclusion d'un élément x_i avec $k+1 \leq i \leq N$. Ensuite, que pouvez-vous dire de la valeur de la somme $\sum_{i=k+1}^{N} \mathbf{1}(x_i \text{ est inclus })$?

d) Dans le même contexte que b), lorsque la taille totale N de la liste n'est pas connue à l'avance, montrez que l'on peut procéder comme suit.

On tient à jour une liste d'enregistrements $(y_1, \ldots, y_n)$, qui fournit, lorsque la totalité de la liste a été communiquée, la sous-liste souhaitée. L'indice t désigne le numéro de l'enregistrement courant, qui varie donc de 1 à N. L'algorithme est le suivant.

1. *pour $t = 1, \ldots, n$, affecter $y_t = x_t$;*
2. *pour $t = n+1, \ldots, N$, lorsque le $t-$ème élément de la liste, x_t, est communiqué, tirer un nombre entier M_t au hasard uniformément entre 1 et t ; si $M_t \leq n$, réaffecter $y_{M_t} = x_t$; sinon, passer à l'enregistrement suivant ;*

Exercice 37 *Trois amies mariées depuis peu, Aricie, Brunhilde et Circé, se retrouvent pour une soirée. Au cours de leur discussion est évoqué le fait qu'environ un tiers des mariages se termine par un divorce. Elles en concluent que, «statistiquement», l'une d'entre elles verra son mariage se rompre. Que pensez-vous de cet argument ? En admettant que les trois amies aient, indépendamment les unes des autres (ce qui doit être discuté) une probabilité égale à 1/3 de voir leur mariage se rompre, quelle est la probabilité qu'exactement l'une des trois divorce ? Quelle est la probabilité qu'aucune ne divorce ?*

Quelques années plus tard, Brunhilde divorce effectivement de son mari, ce qui correspond bien à un mariage rompu sur les trois. Aricie et Circé ont-elles lieu d'en être rassurées quant à la longévité de leurs propres mariages ? Au fait, à quoi correspond exactement le chiffre de 1/3 de mariages soldés par un divorce, évoqué au début ? Est-il pertinent de l'appliquer aux trois amies ?

Exercice 38 *(La fin du monde approche)*

Compte-tenu des connaissances actuelles, on estime que, dans quelques milliards d'années, la mort de notre Soleil rendra la vie impossible sur notre planète. Sauf solution encore à imaginer, on peut donc s'attendre à ce que l'espèce humaine finisse par s'éteindre sur notre planète, le nombre total d'êtres humains y ayant vécu à un moment ou à un autre possédant donc une certaine valeur finie N. D'autre part, appelons n le nombre total d'êtres humains ayant vécu au cours de la décennie 1995-2005.

Admettons pour simplifier que l'on n'ait le choix qu'entre les deux hypothèses suivantes :

1. *la fin de l'espèce humaine est pour bientôt (épuisement des ressources, guerres, pollution, cataclysmes,...), et n/N est de l'ordre de 1/10 ;*
2. *la fin de l'espèce n'est pas pour demain (des solutions aux problèmes actuels vont être trouvées, et nous avons encore de beaux jours devant nous), et n/N est bien inférieur à 1/1000 ;*

et que diverses considérations scientifiques nous permettent d'attribuer a priori une probabilité de 1% à l'hypothèse 1 et de 99% à l'hypothèse 2 (soyons optimistes).

Nous observons aujourd'hui l'événement

$E =$ *«Nous sommes des humains ayant vécu au cours de la décennie 1995-2005 »*,

et l'on peut évaluer la probabilité conditionnelle de cet événement relativement à chacune des deux hypothèses 1 et 2 par n/N *(1/10 sous l'hypothèse 1, moins de 1/1000 sous l'hypothèse 2).*

Comment les probabilités des hypothèses 1 et 2 sont-elles modifiées par la prise en compte de E ? Ceci dépend-il du détail des valeurs choisies pour les différentes probabilités ? Que pensez-vous de cet argument ?

Exercice 39 *M. et Mme D*** ont déjà six enfants, dont cinq filles et un garçon. Mme D*** est à nouveau enceinte. Comment évaluez-vous la probabilité que son enfant à naître soit une fille ?*

Exercice 40 *Après quinze jours de vacances bien méritées qu'il a choisi de passer en famille au Jojoïstan, M. D*** doit rentrer en avion. Deux compagnies sont susceptibles d'assurer la liaison : Air-Jojo, et Pigeon-Vole, plus économique. Cependant, un avion de Pigeon-Vole s'est écrasé il y a peu, si bien qu'au moment de faire son choix, M. D*** est partagé entre toute sorte d'arguments, dont voici quelques exemples.*

- *«Pigeon-Vole n'a eu en moyenne qu'un accident sur 10000 vols au cours des dix dernières années. Comme ils viennent d'en avoir un, je peux donc sereinement choisir cette compagnie.»*
- *«Avec une chance sur 10000 d'avoir un accident, il faudrait que je prenne 10000 fois l'avion pour m'inquiéter. J'ai de la marge...»*
- *«Cet accident laisse à penser que Pigeon-Vole n'est pas fiable. Choisir Air-Jojo est peut-être plus prudent.»*
- *«Après un tel accident, Pigeon-Vole va certainement mettre le paquet sur les contrôles et la sécurité pour rassurer ses clients. Aucun risque donc à voyager avec cette compagnie.»*
- *«Un accident sur 10000 vols, cela représente tout de même une excellente fiabilité. Finalement, je ne vois vraiment pas où est le risque de voyager sur Pigeon-Vole.»*
- *«J'ai cru comprendre que, depuis deux ans, Pigeon-Vole avait beaucoup baissé ses tarifs. Peut-être est-ce en mégotant sur l'entretien et le contrôle des avions, ou bien la formation et les conditions de travail du personnel ? C'est bien ce que laisse à penser cet accident... »*
- *«Si je pouvais avoir la garantie que l'avion dans lequel je vais voler n'est pas du même modèle que celui qui s'est écrasé, je choisirais volontier Pigeon-Vole. Mais comment faire si je m'aperçois au moment d'embarquer que l'on ne m'a dit pas la vérité à ce sujet (ce qui n'est pas impossible, après tout, ils cherchent*

probablement avant tout à remplir leurs avions), je ne vais tout de même pas refuser de monter dans l'avion alors que j'aurai déjà pris mon billet...»

- *«Pigeon-Vole a effectué 500 vols cette année. Nous avons donc une probabilité qui monte à présent à 1/500 d'avoir un accident, soit une multiplication du risque par 20 par rapport au chiffre des années précédentes. Il vaut peut-être mieux que je voyage sur Air-Jojo...»*
- *«De toute façon, j'ai toujours eu un peu peur de l'avion. Cette fois, je prendrai le bateau.»*

*Quels sont les modèles susceptibles de traduire les arguments de M. D*** ? Ces arguments sont-ils conciliables ? Comment pourrait-on tenter de les départager ?*

Exercice 41 *Une société commercialise une préparation paramédicale destinée à lutter contre le rhume, et baptisée AtchoumStop. L'un des arguments publicitaires qu'elle emploie est le suivant : au sein d'un groupe de 100 personnes enrhumées (dont la société tient la liste disponible sur simple demande) ayant utilisé le produit, 82 se sont totalement remises de leur rhume dans les trois jours suivant leur première prise d'AtchoumStop. Discutez en détail les raisons qui peuvent vous amener à douter de la portée de cet argument.*

Exercice 42 *M. D*** fait passer des entretiens d'embauche dans une entreprise. Le nombre total de candidats à auditionner est noté N, et le problème qui se pose à M. D*** est qu'il doit indiquer aux candidats s'il sont retenus immédiatement après leur entretien. La stratégie adoptée par M. D*** est la suivante : il choisit d'abord d'auditionner, sans les recruter, un ensemble de M candidats, afin de se former une idée du niveau de qualification auquel il peut s'attendre. Il procède ensuite à l'audition des $N-M$ candidats restants, et recrute le premier candidat dont il estime que le niveau dépasse celui de l'ensemble des M candidats initialement auditionnés. Comment choisir M de façon à maximiser les chances de recruter le meilleur des N candidats ?*

Pour une généralisation considérable de cette question, vous pouvez consulter l'article de Thomas Bruss cité dans la bibliographie.

Exercice 43 *(Paradoxe de Hempel)*

Une idée de base du raisonnement inductif (par opposition au raisonnement déductif) est que plus nous observons la réalisation d'une certaine propriété, plus le degré de confiance que nous lui attribuons est élevée. Par exemple : plus j'observe de corbeaux noirs, plus mon degré de confiance dans le fait que tous les corbeaux sont noirs est élevé. A présent, notons que le fait que tous les corbeaux soient noirs est équivalent au fait que tout ce qui n'est pas noir n'est pas un corbeau. Dans la pièce où je me trouve, et qui ne contient pas de corbeau, tous les objets qui ne sont pas noirs, effectivement, ne sont pas des corbeaux. Un raisonnement inductif me conduit alors

à considérer que cette observation renforce mon degré de confiance dans la propriété selon laquelle tout ce qui n'est pas noir n'est pas un corbeau, c'est-à-dire la propriété selon laquelle tous les corbeaux sont noirs. Comment mon degré de confiance dans une propriété qui ne concerne que les corbeaux peut-il être modifié par cette observation, alors que je n'ai examiné aucun corbeau ?

Que pourrait donner une approche bayésienne du problème (en cherchant à estimer l'augmentation de probabilité de l'affirmation selon laquelle tous les corbeaux sont noirs liée à une observation des objets dans la pièce où je me trouve, et en la comparant avec l'augmentation que l'on obtiendrait en observant des corbeaux dans la nature) ?

Exercice 44 *Voici une définition alternative de l'indépendance de deux événements A et B : $\mathbb{P}(A|B) = \mathbb{P}(A|B^c)$. Vérifier que cette notion est bien équivalente à la notion habituelle.*

Exercice 45 *Pour diverses représentations sous forme d'arbre (autres que des successions indépendantes) de modèles probabilistes rencontrés dans ce chapitre, explicitez les représentations en arbre obtenues en renversant l'ordre des éléments de variabilité employés pour définir la structure de l'arbre.*

Exercice 46 *(Le paradoxe de Berkson)*

On considère un modèle probabiliste $(\Omega, \mathbb{P})$, et deux événements A et B indépendants dans ce modèle. On suppose en outre que $\mathbb{P}(A) > 0$, $\mathbb{P}(B) > 0$, et $\mathbb{P}(A \cup B) < 1$. Prouvez que, dans le modèle probabiliste $(\Omega, \mathbb{P}(\cdot|A \cup B))$, les événements A et B sont négativement associés.

Voici un exemple où ce résultat devrait certainement être pris en compte, expliquez comment. Dans une université américaine (bien entendu, une telle chose est impossible en Europe...), les étudiants admis dans le département d'ingénierie peuvent l'être pour deux raisons : leurs qualités démontrées dans les matières reliées à l'ingénierie, ou leurs performances en base-ball. Admettons (ce qui n'est pas forcément vrai !) qu'il y a indépendance entre ces deux caractéristiques pour la population formée par les jeunes en âge d'étudier dans ce département. Un professeur du département, plein de préjugés, décide néanmoins de démontrer les moindres performances en ingénierie des étudiants doués pour le base-ball. De fait, les chiffres sont accablants, et, en prenant en compte l'ensemble des étudiants du département, le professeur constate de manière très nette une association négative entre les performances au base-ball et les qualités en ingénierie...

De la même manière, si l'on étudie l'ensemble des patients d'un hôpital, on pourra ainsi «montrer» un effet protecteur d'une pathologie donnée vis-à-vis de l'ensemble des autres pathologies possibles alors qu'il y a totale indépendance...

Comment pourrait-on éviter ce problème dans ce(s) contexte(s) ? Plus généralement, comment éviter d'être piégé par ce problème lorsque l'on étudie l'association pouvant exister entre des variables ?

Exercice 47 *Au sujet du référendum de 2005 sur la constitution européenne, deux semaines avant le scrutin, on pouvait lire dans un journal gratuit à propos de ses lecteurs :*
- 73% des gens ont pris leur décision qui est : oui pour 48,45% et non pour 51,55% ;
- 27% n'ont pas pris de décision mais se répartissent en trois catégories : 31% pensent voter oui, 24% pensent voter non et 45% ne savent vraiment pas.

Prenons une personne au hasard dans la population à ce moment. Quelle est selon vous la probabilité pour qu'elle penche pour le oui ? Et quelle est la probabilité pour qu'elle penche pour le non ?

Exercice 48 *Encore à propos des sondages sur la constitution européenne de 2005... Deux semaines avant le scrutin, admettons que l'état de l'opinion est le suivant (on donne à la fois les sympathies politiques des français et leur intention de vote dans le tableau ci-dessous). On note entre parenthèse sur la première ligne le poids relatif de chaque parti dans l'opinion.*

	PC (10%)	*PS (19%)*	*Verts (21%)*	*UDF (20%)*	*UMP (22%)*	*FN/MNR (8%)*
oui	*7%*	*54%*	*57%*	*70%*	*72%*	*12%*
non	*93%*	*46%*	*43%*	*30%*	*28%*	*88%*

a) Montrer que le score p du oui est 53,73%.
b) On tire une personne au hasard dans la population. Elle dit vouloir voter oui. Quelle est selon vous la probabilité pour qu'elle soit sympathisante UDF ?

Exercice 49 *On désire étudier les effets d'une exposition prolongée à un polluant d'origine industrielle sur la survenue ultérieure d'un cancer du poumon. Pour cela, on effectue un suivi de la santé des employés d'un secteur industriel dans lequelle ceux-ci y sont régulièrement exposés. En définitive, on constate chez ces employés un taux de cancer du poumon comparable à celui de la population totale, et l'on est donc tenté de conclure a l'absence d'influence de ce polluant. Cependant, on peut imaginer que le polluant a un réel rôle causal et favorise la survenue d'un cancer, mais que cet effet est compensé par un autre lié (par exemple, tout ceci est fictif) aux habitudes alimentaires plus saines des employés de ce secteur par rapport à la population totale. On ne mettrait alors pas en évidence d'association entre exposition et cancer, bien qu'il y existe un effet causal. Fabriquez un exemple chiffré d'un tel phénomène.*

Exercice 50 *On dit qu'un nouveau-né à terme présente un faible poids de naissance lorsque celui-ci est inférieur à un certain seuil (par exemple 2,5 kg). Il apparaît que, chez les nouveaux-nés à terme de faible poids de naissance, le taux de mortalité est significativement plus élevé que chez les nouveaux-nés présentant un poids de naissance normal. De manière surprenante, on a constaté que, chez les nouveaux-nés à terme de faible poids de naissance, le fait que la mère soit fumeuse a tendance à réduire ce taux de mortalité. Doit-on en déduire que, dans ce cas, la consommation de tabac par la mère a un effet protecteur sur le nouveau-né ?*

Exercice 51 *L'université de Berkeley fut poursuivie pour dicrimination liée au sexe lorsqu'il fut établi que le taux d'admission des jeunes femmes au niveau graduate (plus ou moins équivalent à ce qui est aujourd'hui en France le M2) étaient très significativement inférieur à ceux des hommes. De fait, pour le semestre de printemps de l'année 1973, les résultats (toutes disciplines confondues) étaient les suivants :*

	Nombre de candidats	*Proportion d'admis*
Hommes	*8442*	*44%*
Femmes	*4321*	*35%*

En revanche, en distinguant selon les différents départements (seuls six départements, numérotés de A à F, sont présentés ici) de l'université, les résultats étaient les suivants.

A	*Nombre de candidats*	*Proportion d'admis*
Hommes	*825*	*62%*
Femmes	*108*	*82%*

B	*Nombre de candidats*	*Proportion d'admis*
Hommes	*560*	*63%*
Femmes	*25*	*68%*

C	*Nombre de candidats*	*Proportion d'admis*
Hommes	*325*	*37%*
Femmes	*593*	*34%*

D	*Nombre de candidats*	*Proportion d'admis*
Hommes	*417*	*33%*
Femmes	*375*	*35%*

E	*Nombre de candidats*	*Proportion d'admis*
Hommes	*191*	*28%*
Femmes	*393*	*24%*

F	*Nombre de candidats*	*Proportion d'admis*
Hommes	*272*	*6%*
Femmes	*341*	*7%*

Comment ces données éclairent-elles, selon vous, la question de l'existence d'une discrimination entre hommes et femmes ?

Exercice 52 *Afin de tester les performances d'une nouvelle méthode d'apprentissage de la lecture, on fait passer aux élèves d'une classe d'école primaire un test de lecture, noté sur 20. On sélectionne ensuite les élèves ayant obtenu moins de 7/20 (il s'en trouve 7), et on leur fait suivre pendant plusieurs mois un programme spécial inspiré par la nouvelle méthode d'apprentissage. A l'issue de ce programme, on fait à nouveau passer un test aux élèves de la classe. Bilan : les élèves n'ayant pas suivi le nouveau programme obtiennent, en moyenne, des résultats voisins, voire légèrement inférieurs à ceux qu'ils avaient obtenu lors du premier test, tandis que les 7 élèves ayant suivi le programme spécial voient leur score progresser de plusieurs points en moyenne, jusqu'à atteindre une moyenne proche de 10. Faut-il conclure à la supériorité de la nouvelle méthode sur l'ancienne ?*

Exercice 53 *Comment définiriez-vous la probabilité des coïncidences présentées dans les exemples de la section 1.8.1 ?*

Exercice 54 *Comment définiriez-vous précisément un protocole permettant d'étudier les relations entre le fait de penser à une personne et le fait que celle-ci vous appelle peu après ? Comment comptez-vous procéder pour distinguer un don surnaturel de simples coïncidences ?*

Exercice 55 *Que penser d'une théorie produisant comme résultat le fait que la probabilité d'apparition de la vie sur Terre soit extrêmement faible, et que notre existence doive donc être considérés comme le fruit d'une formidable coincidence ?*

Exercice 56 *Au cours de débats portant sur l'utilité (ou l'inutilité) du redoublement à l'école, l'argument suivant a été employé : «plusieurs études ont montré que, globalement, les résultats scolaires des enfants que l'on fait redoubler ne s'améliorent pas de manière significative à l'issue de ce redoublement.» Cet argument vous semble-t-il constituer, à lui seul, un élément suffisant pour prôner la suppression du redoublement ?*

Exercice 57 *(Le sophisme du procureur)*

*Sur la foi d'un test ADN, M. D*** comparaît devant un tribunal dans le cadre d'une affaire criminelle, et l'expert invité à la barre explique que, à supposer que M.*

*D*** soit innocent, la probabilité pour que son ADN coïncide avec celui trouvé sur les lieux du crime d'après le test effectué est d'environ 1/10000.*
*1) Doit-on en déduire qu'il y ait moins d'une chance sur 10000 pour que M. D*** soit innocent ? Si non, comment évaluer la probabilité pour que M. D*** soit innocent ?*
*2) Comment selon vous peut-on parvenir à des estimations comme celles proposées par l'expert. Quelle fiabilité accorder à celles-ci ? Comment les variations de cette estimation affectent-elles l'estimation de la probabilité pour que M. D*** soit coupable ?*
*3) Deux experts différents proposent deux estimations différentes de la probabilité de coïncidence de l'ADN de M. D*** avec celui trouvé sur les lieux du crime dans l'hypothès où celui-ci est innocent, disons p_1 et p_2, obtenues par deux méthodes différentes. Les propositions suivantes vous semblent-elles raisonnables ? Pour quelles raisons ?*

- *Utiliser comme estimation $\frac{p_1+p_2}{2}$.*
- *Evaluer séparément la probabilité de culpabilité de M. D*** en utilisant p_1 puis p_2, et conserver la plus petite des deux valeurs obtenues.*
- *Réexaminer les deux méthodes employées par les experts pour parvenir à leurs estimations, et ne conserver que la valeur obtenue par la méthode qui semble la plus pertinente.*
- *Multiplier par 10 la plus grande des deux valeurs p_1 ou p_2, multiplier par 1/10 la plus petite, calculer les estimations de la probabilité de culpabilité ainsi obtenues, et considérer qu'une valeur raisonnable doit se trouver dans la fourchette ainsi obtenue.*
- *Décider que si les résultats obtenus à partir des deux méthodes pointent dans la même direction (culpabilité ou innocence), on se satisfait de ce résultat.*
- *Décider que si les résultats obtenus à partir des deux méthodes pointent dans deux directions différentes, on ne peut rien dire.*
- *Analyser les deux méthodes employées et tenter de trouver une troisième méthode qui puisse remédier à leurs défauts potentiels avant de faire quoique ce soit.*
- *Essayer d'estimer, pour chaque méthode, les marges d'erreurs susceptibles d'affecter leurs résultats, et raisonner avec des fourchettes de valeur (comment ?) plutôt qu'avec des valeurs fixées.*

*4) Vous faites partie du jury chargé de statuer sur le sort de M. D***. Êtes-vous plus impressionné par une valeur, tous calculs faits, de la probabilité de culpabilité de 0,9998, que par une valeur de 0,9 ou 0,8 ? Décideriez-vous de déclarer M. D*** coupable en fonction de ce seul calcul (sachant que vous n'êtes censé le faire que lorsque sa culpabilité semble établie au-delà de tout doute raisonnable) ? Si oui, jusqu'à quelle valeur de la probabilité de culpabilité vous décidez-vous pour la culpabilité ? 0,99 ? 0,9 ? 0,8 ? 0,55 ? 0,5000001 ? Comment jugez-vous de la fiabilité de l'estima-*

tion proposée ? Une estimation à 0,99999 par une méthode qui semble douteuse vous convainc-t-elle davantage qu'une estimation de 0,8 par une méthode qui semble plus fiable ?
*5) M. D***, pour expliquer la similarité observée entre son propre ADN et les traces trouvées sur les lieux du crime, prétend avoir été victime d'une machination, des échantillons de ses propres tissus ayant été récupérés sur lui à son insu, puis déposés sur place, par le meurtrier ou l'un de ses complices, dans le but de le faire accuser à tort. Cette machiavélique possibilité a-t-elle été prise en compte dans les évaluations de probabilité ci-dessus ? Si oui, comment, et sinon, comment pourrait-elle l'être ? Même question avec la possibilité pour que de l'ADN de M. D*** se trouve par hasard sur les lieux du crime (c'est-a-dire, sans que celui-ci soit coupable) ?*
6) Comment le principe de la présomption d'innocence est-il, selon vous, pris en compte, ou au contraire ignoré, dans les arguments qui précèdent ?

Exercice 58 *M. H*** joue au Loto, et... gagne le gros lot. Quand il tente de faire valoir ses droits, on refuse de lui verser son gain en lui opposant l'argument suivant. «La probabilité de gagner sans tricher est infime, et vous venez de gagner. Le plus probable est donc que vous n'êtes qu'un tricheur ! Estimez-vous heureux que nous ne vous trainions pas devant les tribunaux, et n'y revenez pas !» Que pensez-vous du bien-fondé de cet argument ?*

Exercice 59 *Voici un extrait du journal Le Monde, daté d'août 2005, dans un article consacré à la sécurité aérienne. «(...) Dans le même temps, les vols irréguliers devenaient plus meurtriers : le nombre de tués voyageant sur des charters représentait environ 20% du total des décès dus à des accidents d'avion à la fin des années 1980, contre 50% aujourd'hui.(...)» Cette phrase vous semble-t-elle convaincante ? Pourquoi ?*

Exercice 60 *Une étude réalisée auprès d'adolescents américains appartenant à des gangs a révélé que 40% de ceux qui se déclaraient athées avaient déjà été condamnés pour des délits accompagnés d'actes violents. Cette proportion est plus de cent fois supérieure à celle des personnes condamnées pour des délits similaires au sein de la population totale. Cette étude montre donc clairement que l'athéisme conduit tout droit à la violence. Que pensez-vous de cet argument ?*

Exercice 61 *(L'affaire Sally Clark : une chance sur 73 millions !)*

En 1997, Mme Sally Clark perdit son premier enfant, alors âgé de 11 semaines, et le décès fut attribué à des causes naturelles. L'année suivante, son deuxième enfant mourut, âgé de huit semaines. Mme Clark fut alors arrêtée et accusée du meurtre de ses deux enfants, puis jugée, reconnue coupable, et condamnée en 1999 à la prison à perpétuité. Pourtant, les éléments de preuve d'ordre médical étaient extrêmement

ténus, voire inexistants, et rien ne laissait à penser a priori que Mme Clark ait pu être une mère négligente ou violente envers ses enfants. En fait, il semble bien que la conviction du jury ait été emportée par un argument de nature statistique, version moderne du dicton selon lequel la foudre ne frappe jamais deux fois au même endroit, et affirmant en substance qu'il faudrait une coïncidence vraiment extraordinaire pour que l'on observe non pas une, mais deux morts subites du nourrisson successives au sein d'une même famille. Sir Meadow, qui témoigna au procès en tant qu'expert médical, affirma que la probabilité d'une telle coïncidence (que surviennent par hasard deux morts subites du nourrisson dans une famille comparable à celle de Mme Clark) était d'environ une chance sur 73 millions, ce qui fut apparemment interprété comme un argument décisif indiquant la culpabilité de Mme Clark, et présenté comme tel par les médias à l'époque.
1) Cette estimation de probabilité vous semble-t-elle constituer un argument décisif ?
2) L'origine de cet estimation de probabilité partait du chiffre d'environ $1/8500$ *pour la probabilité d'une mort subite du nourrisson au sein d'une famille comparable à celle des Clark, estimé d'après des données médico-légales, d'où une estimation de* $1/8500 \times 1/8500 \approx 1/73000000$*. Etes-vous convaincu par cette estimation ?*

Afin de satisfaire la curiosité que n'a sans doute pas manqué de susciter le début de cet exercice, voici quelques éléments sur la suite de l'histoire. Des études statistiques menées ultérieurement par le Professeur Hill, de l'université de Salford, conduisent à proposer que la probabilité d'observer deux morts subites devrait être approximativement 9 fois supérieure à la probabilité pour une mère de causer délibérément la mort de ses deux enfants, d'où une estimation de la probabilité de culpabilité d'environ 1/10. Par ailleurs, au sujet de la méthode d'estimation décrite à la question 2), l'étude menée par le Pr Hill semble indiquer que le risque de mort subite est entre 5 et 10 fois supérieur chez un enfant dont un frère ou une sœur est lui-même décédé de mort subite du nourrisson.

Les Clark firent appel du jugement, s'appuyant en particulier sur des avis de statisticiens dénonçant ces différentes erreurs d'argumentation. L'appel fut rejeté, la conclusion du juge étant que le point essentiel était la rareté de l'apparition de deux morts subites au sein d'une même famille, non remise en question par ces remarques. Devant une telle incompréhension, la Société Royale de Statistique écrivit aux autorités judiciaires pour enfoncer le clou. De plus, on découvrit que des éléments médicaux accréditant largement l'hypothèse d'une mort accidentelle du deuxième enfant avaient été dissimulés lors du procès. Un second procès en appel fut alors organisé, et Mme Clark fut finalement acquittée après avoir passé près de deux ans et demi en prison. Sir Meadow a été radié en 2005 par l'ordre des médecins du Royaume-Uni, pour «serious professional misconduct.»
3) D'après vous, que signifie le fait d'évaluer la probabilité de mort subite du nourrisson dans une famille «comparable» à celle des Clark ? Quels critères peut-on ou

doit-on retenir pour s'assurer de cette «comparabilité» ?
4) Pour citer la lettre de la Société Royale de Statistique, «The fact that two deaths by SIDS [sudden infant death syndrome] is quite unlikely is, taken alone, of little value. Two deaths by murder may well be even more unlikely. What matters is the relative likelihood of the deaths under each explanation, not just how unlikely they are under one explanation.» Pouvez-vous traduire ceci précisément en termes de probabilités conditionnelles ?
5) La «loi de Meadow» citée par les médias lors du procès affirmait : «une mort subite est une tragédie, deux morts subites doivent éveiller les soupçons, trois morts subites : c'est un meurtre.» Etes-vous convaincu ? Comment traduirait-on cette loi en termes de probabilités ?

Exercice 62 *On considère n modèles probabilistes* $(\Omega_1, \mathbb{P}_1), \ldots, (\Omega_n, \mathbb{P}_n)$*, et le modèle décrivant leur succession indépendante :* $\Omega := \Omega_1 \times \cdots \times \Omega_n$, $\mathbb{P} := \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$.
1) Prouvez, en vous appuyant sur la représentation en arbre, le fait que

$$\mathbb{P}(* \cdots * a_{i_1} * \cdots * a_{i_2} * \cdots * a_{i_p} * \cdots *) = \mathbb{P}_{i_1}(a_{i_1}) \times \mathbb{P}_{i_2}(a_{i_2}) \times \cdots \times \mathbb{P}_{i_p}(a_{i_p}).$$

2) De la même manière, prouvez la propriété des coalitions.

Chapitre 2

Variables aléatoires

2.1 Introduction et définition

Dans le chapitre précédent, nous avons présenté le formalisme général des modèles probabilistes, qui permet de représenter mathématiquement des situations incorporant incertitude, variabilité ou hasard. Ce formalisme fait intervenir un espace des possibles Ω, dont les éléments représentent les éventualités élémentaires, c'est-à-dire les différentes issues possibles de la situation considérée, au niveau de précision choisi pour la décrire, et une probabilité $\mathbb{P}$ associant à chaque éventualité élémentaire un nombre représentant la probabilité que la situation soit réalisée *via* cette éventualité particulière. La réalisation de la situation considérée est modélisée comme le choix de l'une des éventualités élémentaires, ω, qui contient donc toute l'information – là encore, au niveau de précision choisi – sur la façon dont la situation s'est réalisée. Comme nous l'avons vu au cours du chapitre précédent, l'espace des possibles Ω est généralement construit en assemblant des «morceaux» d'information portant sur la réalisation de la situation, chacun de ces morceaux représentant une partie de l'information globale contenue dans les éléments de Ω, qui permet de spécifier complètement, au niveau de description choisi, l'issue de la situation étudiée. En particulier, on représente souvent Ω par un arbre, dont chaque ramification correspond à la spécification de l'un des choix dont la liste complète permet de spécifier la manière dont la situation s'est réalisée. Pourtant, même si ce formalisme est suffisant pour donner une description complète de l'incertitude affectant la situation qui est modélisée, il nous sera souvent nécessaire d'extraire du modèle probabiliste $(\Omega, \mathbb{P})$ des informations de nature quantitative, qui ne figurent pas forcément explicitement, en tant que telles, dans le modèle, et permettent d'en résumer numériquement certains des aspects les plus pertinents pour nous. Ceci justifie la définition d'une notion

générale : on appelle **variable aléatoire toute fonction définie sur Ω :**

$$\begin{cases} X : \Omega \to E, \\ \omega \mapsto X(\omega) \end{cases}$$

A chaque valeur particulière de $\omega \in \Omega$ correspond une valeur de X, $X(\omega) \in E$, où E désigne l'ensemble dans lequel la fonction E prend ses valeurs, et sera le plus souvent une partie de $\mathbb{R}$ ou de $\mathbb{R}^n$, mais pourra également représenter la liste des valeurs possibles d'un caractère qualitatif. De manière générale, nous désignerons par S_X l'ensemble des valeurs possibles pour X, c'est-à-dire l'ensemble

$$S_X = \{X(\omega) \ : \ \omega \in \Omega\},$$

que nous appellerons **l'espace image de Ω par** X. En général, S_X est strictement inclus dans E. Lorsque Ω est fini ou dénombrable, S_X est également un ensemble fini ou dénombrable. On parle dans ce cas de **variable aléatoire discrète**. Le cas des **variables aléatoires continues**, pour lesquelles S_X est typiquement un intervalle de $\mathbb{R}$, sera traité dans un cadre séparé.

Exemple

Pour décrire le résultat du lancer de deux dés, on peut faire appel à l'espace

$$\Omega = \{1; 2; 3; 4; 5; 6\}^2,$$

des couples formés par les chiffres du premier et du deuxième dé. Un élément ω de Ω se met donc sous la forme $\omega = (x_1, x_2)$, et la fonction X définie sur Ω par :

$$X((x_1, x_2)) = x_1$$

est une variable aléatoire, qui décrit le résultat du premier dé. La connaissance de la valeur de $X(\omega)$ ne permet pas de reconstituer celle de ω : chacune des six éventualités élémentaires $(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)$ donne la même valeur à X, à savoir 1. $X(\omega)$ ne contient donc pas tout l'information relative à ω, mais seulement une partie. Ainsi, l'événement «$X = 1$» correspond en fait à l'événement formel :

$$\{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\}.$$

Notons que les six événements : «$X = 1$», «$X = 2$», «$X = 3$», «$X = 4$», «$X = 5$», «$X = 6$» sont deux-à-deux incompatibles, et recouvrent Ω, puisque l'un de ces événements est toujours réalisé, X prenant toujours une valeur parmi les entiers de 1 à 6. Ils forment donc une partition de Ω en six événements, chacun des événements

de ce système comprenant six éventualités élémentaires. Si l'on introduit maintenant une autre variable aléatoire Y définie sur Ω par

$$Y((x_1, x_2)) = x_2,$$

qui décrit donc le résultat du deuxième dé, on obtient une autre partition de Ω, associée aux différentes valeurs que peut prendre Y. Les six événements «$Y = 1$», «$Y = 2$», «$Y = 3$», «$Y = 4$», «$Y = 5$», «$Y = 6$» forment également un «découpage» de l'espace des possibles, différent de celui que l'on obtenait avec X. L'événement «$Y = i$» correspond à l'événement : $\{(1, i), (2, i), (3, i), (4, i), (5, i), (6, i)\}$. Si l'on range les éléments de Ω dans un tableau :

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

le découpage relatif aux valeurs de X est formé par les six lignes du tableau, le découpage relatif à Y par les six colonnes. Ces deux découpages coexistent, mais sont distincts. Si l'on introduit maintenant la variable aléatoire Z définie par

$$Z = (X, Y),$$

on obtient un découpage plus fin que les deux précédents, correspondant aux trente-six cases du tableau. En fait, dans ce cas, $Z(\omega) = \omega$.

Si l'on représente Ω sous la forme d'un arbre de profondeur 2 dont la première ramification détermine le résultat du premier dé, et la seconde le résultat du second dé, X et Y sont formellement des fonctions définies sur les feuilles (ou encore, ce qui revient au même, sur les rayons conduisant de la racine à une feuille) de l'arbre, même si X ne fait réellement intervenir que la première ramification, et Y la seconde.

Peut-être cette manière de définir explicitement les variables aléatoires en fonction des éventualités ω vous paraît-elle étrange. En fait, on retrouve la même distinction concret/formel que celle qui s'applique aux événements. Du point de vue formel, une variable aléatoire X est une fonction (ou encore une application), dont on doit définir la valeur $X(\omega)$ pour chaque éventualité élémentaire ω, tandis que, du point de vue concret, une variable aléatoire est simplement une quantité (ou un caractère qualitatif) en rapport avec la situation considéré, et qui, du fait que cette situation est variable, incertaine ou aléatoire, est elle-même variable, incertaine, ou aléatoire. Comme dans le cas des événements, on définira souvent (mais pas toujours) les variables aléatoires de manière concrète, la traduction formelle dans le cadre du

modèle étant implicite (mais nécessitant bien entendu la connaissance du dictionnaire reliant les éléments de Ω à la réalité). Comme dans le cas des événements, une variable aléatoire au sens concret n'est pas nécessairement associée à une variable aléatoire au sens formel : cela dépend de la finesse avec laquelle l'espace des possibles Ω décrit la situation étudiée. Par exemple, la variable aléatoire (au sens concret) correspondant à la durée du lancer du premier dé n'est pas associée à une variable aléatoire au sens formel dans le modèle décrit ci-dessus, car celui-ci n'incorpore aucune information relative à cette durée. Enfin, une même variable aléatoire (au sens concret) pourra correspondre à des variables aléatoires (au sens formel) différentes dans des modèles différents.

Dans l'exemple d'une succession indépendante de N lancers de pile ou face, telle que décrite dans le chapitre précédent par le modèle $(\Omega^N, \mathbb{P}^{\otimes N})$, on pourra définir une variable aléatoire de manière concrète par : «X est le nombre total de face obtenus au cours des N lancers», ou, de manière formelle (et équivalente) par :

$$X(\omega) = \#\{1 \leq i \leq N \ : \ \omega_i = \mathrm{F}\}$$

(où $\#E$ désigne le nombre d'éléments de l'ensemble E, et en se souvenant que les éléments de Ω^N sont de la forme $\omega = (\omega_1, \ldots, \omega_n)$). Définissons les variables aléatoires $X_1, \ldots, X_N$, à valeurs dans {P,F}, représentant les résultats (pile ou face) des lancers successifs, définies de manière formelle par :

$$X_i(\omega) = \omega_i.$$

Ces variables aléatoires figurent explicitement dans le modèle, et c'est en fait à partir de la spécification de leurs valeurs qu'est construit l'espace des possibles Ω^N. Inversement, la variable aléatoire X comptant le nombre de face obtenus ne figure pas explicitement dans le modèle, mais sa valeur se déduit de celle de $(\omega_1, \ldots, \omega_N)$.

Quelques confusions à éviter

Comme pour toute fonction, il importe de bien différencier la valeur ponctuelle prise par une variable aléatoire pour un certain ω, $X(\omega)$, qui représente la valeur de X que l'on observe lorsque la situation étudiée se réalise selon l'éventualité élémentaire ω, et la variable aléatoire elle-même, qui est une fonction sur Ω, et décrit la totalité des valeurs possibles de cette variable. Quand on parle de variable aléatoire, on considère donc implicitement toutes les valeurs possibles que celle-ci peut prendre, et non pas simplement celle qui s'est effectivement réalisée. De même, il importe de bien distinguer l'espace Ω sur lequel la variable est définie, de l'espace S_X dans lequel celle-ci prend ses valeurs. Notons qu'une variable aléatoire ne prend pas nécessairement des valeurs numériques, (tout dépend de l'ensemble S_X) mais peut également représenter un caractère qualitatif comme une couleur, en prenant par exemple des «valeurs»

telles que «rose», «bleu», «vert». Cependant, la plupart des variables aléatoires que nous considérerons prendront des valeurs numériques.

Notons que l'on est en général amené à considérer plusieurs variables aléatoires définies sur un même espace de probabilité – le modèle de référence utilisé pour décrire la situation. Dans l'exemple de pile ou face ci-dessus, les variables X, $X_1, \ldots, X_N$ n'ont rien à voir entre elles, mais sont toutes définies sur le même espace Ω^N.

Un exemple simple et fondamental de variable aléatoire : la fonction indicatrice d'un événement

Étant donné un événement formel A (c'est-à-dire un sous-ensemble de Ω), on définit la **fonction indicatrice** de l'événement A, notée $\mathbf{1}_A$, par :

$$\begin{cases} \mathbf{1}_A(\omega) = 1 \quad \text{si } \omega \in A, \\ \mathbf{1}_A(\omega) = 0 \quad \text{si } \omega \notin A. \end{cases}$$

La fonction $\mathbf{1}_A$ est donc une variable aléatoire, puisqu'il s'agit d'une fonction définie sur Ω, et sa valeur indique la réalisation ou la non-réalisation de l'événement A : $\mathbf{1}_A$ prend la valeur 1 lorsque A est réalisé, et 0 lorsqu'il n'est pas réalisé. C'est probablement l'exemple le plus simple de variable aléatoire.

2.2 Loi d'une variable aléatoire

2.2.1 Le point de vue formel pour les variables aléatoires discrètes

De manière générale, comment la probabilité $\mathbb{P}$ définie sur l'espace des possibles Ω affecte-t-elle les variables aléatoires définies sur Ω ? Remarquons en passant que la notion de variable aléatoire est définie indépendamment de la probabilité sur Ω : il s'agit d'une notion relative seulement à la structure de Ω (ou encore de la situation considérée), sans référence à $\mathbb{P}$. Pour chaque valeur que peut prendre une variable aléatoire, il convient donc de préciser quelle est la probabilité pour que cette valeur soit effectivement prise. Considérons une variable aléatoire X définie sur Ω, dont nous notons S_X l'espace image, c'est-à-dire l'ensemble des valeurs possibles :

$$X : \Omega \to S_X.$$

À chaque éventualité élémentaire $\omega \in \Omega$ est attaché un élément $X(\omega)$ de S_X. Inversement, à chaque valeur possible que peut prendre X, c'est-à-dire à chaque élément $s \in S_X$, est attaché l'événement (concret) «$X = s$», ou encore «X prend la valeur s», dont le correspondant formel dans Ω est l'ensemble des éventualités élémentaires ω de Ω telles que $X(\omega) = s$. Cet événement possède une probabilité, définie

comme la somme des probabilités des éventualités élémentaires qui le constituent :

$$\mathbb{P}(X = s) = \sum_{\omega \,:\, X(\omega)=s} \mathbb{P}(\omega).$$

Il est important de comprendre parfaitement la signification de cette formule : pour une valeur fixée s que peut prendre la variable aléatoire X, un certain nombre d'éventualités élémentaires $\omega \in \Omega$ donnent effectivement à X la valeur s, c'est-à-dire sont telles que $X(\omega) = s$. La probabilité pour que X prenne la valeur s est donc la somme des probabilités de toutes ces éventualités élémentaires. Les deux difficultés présentes ici sont que :

1. plusieurs éventualités élémentaires ω peuvent fournir la même valeur pour X,
2. les éléments de S_X, c'est-à-dire les valeurs que peut prendre X (souvent des nombres), ne sont pas de la même nature que les éléments de Ω.

Reprenons l'exemple précédent des deux dés, en supposant que la probabilité définie sur Ω est la probabilité uniforme. L'événement «X=2» correspond à l'événement

$$\{(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\},$$

et la probabilité pour que $X = 2$ est donc donnée par :

$$\mathbb{P}(X = 2) = \mathbb{P}(2,1) + \mathbb{P}(2,2) + \mathbb{P}(2,3) + \mathbb{P}(2,4) + \mathbb{P}(2,5) + \mathbb{P}(2,6) = 6 \times \frac{1}{36} = \frac{1}{6}.$$

Les valeurs que peut prendre X forment l'ensemble des entiers entre 1 et 6, ici $S_X = \{1;2;3;4;5;6\}$, et ne sont pas de la même nature que les éléments de Ω, qui, lui, est constitué par des couples d'entiers.

Nous l'avons vu précédemment sur un exemple, et la généralisation de cette observation est immédiate, la liste des événements «$X = s$», s décrivant l'ensemble des éléments de S_X (les valeurs possibles de X) forme une partition de l'espace Ω, c'est-à-dire un système complet d'événements. La liste des probabilités associées à ce découpage, c'est-à-dire **la liste des probabilités $\mathbb{P}(X = s)$, s décrivant S_X, est appelée loi de la variable aléatoire X, (ou loi de probabilité de X, ou encore distribution de probabilité de X), relativement à la probabilité $\mathbb{P}$.** (Lorsqu'il n'y a pas d'ambiguïté, on parle simplement de la loi de X sans préciser la probabilité de référence sur Ω).

Dans l'exemple des deux dés, on vérifie facilement que la loi de X est donnée par :

$$\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = \mathbb{P}(X = 3) = \mathbb{P}(X = 4) = \mathbb{P}(X = 5) = \mathbb{P}(X = 6) = \frac{1}{6}.$$

La liste des événements «$X = s$» formant une partition de Ω, on a la relation

$$\sum_{s \in S} \mathbb{P}(X = s) = 1.$$

La fonction $s \mapsto \mathbb{P}(X = s)$ définie sur l'ensemble S_X apparaît dont comme une probabilité sur l'ensemble S_X. Ceci nous donne une autre manière de présenter la loi d'une variable aléatoire X définie sur un modèle probabiliste $(\Omega, \mathbb{P})$. Si l'on considère un modèle moins fin que $(\Omega, \mathbb{P})$ pour décrire la même situation dans lequel on ne prend en compte que l'information relative à la valeur prise par X, on obtient un modèle (S_X, p_X), dont les éventualités élémentaires sont simplement les différentes valeurs possibles pour X, c'est-à-dire les éléments de l'espace image S_X, sur lequel, (pour assurer la compatibilité avec la description de la situation fournie par $(\Omega, \mathbb{P})$), on doit nécessairement avoir :

$$p_X(s) = \mathbb{P}(X = s).$$

La loi de X apparaît alors simplement comme la probabilité p_X sur l'espace des possibles S_X de ce modèle : à chaque élément de S_X est associée la probabilité pour que X prenne effectivement cette valeur. Nous appellerons (S_X, p_X) le **modèle image de** $(\Omega, \mathbb{P})$ **par** X. (La probabilité p_X sur S_X dépend bien entendu de $\mathbb{P}$ et pas seulement de X.)

Notons que, et, si H désigne un sous-ensemble de S_X, on a

$$\mathbb{P}(X \in H) = \sum_{s \in H} \mathbb{P}(X = s).$$

On comparera cette expression à la suivante, tout aussi valable :

$$\mathbb{P}(X \in H) = \sum_{\omega \in \Omega \,:\, X(\omega) \in H} \mathbb{P}(\omega),$$

en notant que l'une repose sur une représentation au moyen du modèle $(\Omega, \mathbb{P})$, et l'autre sur l'utilisation du modèle (S_X, p_X).

Le fait que la loi de probabilité d'une variable aléatoire apparaisse simplement comme une probabilité sur l'ensemble des valeurs de cette variable justifie le fait que l'on parle souvent d'une loi de probabilité sur un ensemble S **sans référence particulière à une variable aléatoire susceptible de posséder cette loi**. Une loi de probabilité sur un ensemble S (fini ou dénombrable), indépendamment de la notion de variable aléatoire, désigne simplement une probabilité sur l'ensemble S, vu comme un espace des possibles [1]. Au risque d'insister inutilement, **une loi de probabilité sur** S (sans référence à une variable aléatoire) est donc la donnée d'une probabilité sur l'ensemble S, c'est-à-dire d'une fonction $p : S \to [0, 1]$, vérifiant la condition de normalisation $\sum_{s \in S_X} p(s) = 1$. Dire qu'une variable aléatoire suit la loi p sur S_X, c'est simplement dire que, pour tout $s \in S_X$, on a $\mathbb{P}(X = s) = p(s)$,

1. On peut noter, que, étant donné un espace de probabilité (S, p), la variable aléatoire X définie sur S par $X(s) = s$ suit la loi p.

autrement dit, que la loi de la variable aléatoire X (en tant que probabilité sur l'espace des possibles S_X) coïncide avec la probabilité p.

Dans le même ordre d'idées, on spécifie implicitement un modèle $(\Omega, \mathbb{P})$ d'une situation en définissant de manière concrète une variable aléatoire X et en spécifiant sa loi. Ce modèle correspond alors à $\Omega = S_X$ et $\mathbb{P} =$ loi de X.

Il est important de bien comprendre que plusieurs variables aléatoires définies sur le même espace de probabilité, mais bien distinctes, peuvent parfaitement partager la même loi. Par exemple, dans le modèle $(\Omega^N, \mathbb{P}^{\otimes N})$ décrivant une répétition indépendante de N lancers de pile ou face, $(\Omega = \{\mathrm{P}, \mathrm{F}\},\ \mathbb{P}(\mathrm{P}) = p,\ \mathbb{P}(\mathrm{F}) = 1 - p)$, chacune des variables aléatoires X_i représentant le résultat du $i-$ème lancer possède la même loi, à savoir : $\mathbb{P}^{\otimes N}(X_i = \mathrm{P}) = p,\ \mathbb{P}^{\otimes N}(X_i = \mathrm{F}) = 1 - p$. Ces variables aléatoires ne sont pourtant pas en général égales entre elles ! De même, dans le cas des deux dés, lorsque la probabilité décrivant les lancers est uniforme, les trois variables aléatoires X, Y et $7 - X$ ont la même loi (exercice facile), et ne sont pas égales en général. De même, des variables aléatoires définies sur des espaces de probabilité différents, et intervenant dans la modélisation de situations concrètes complètement différentes, pourront également posséder la même loi. Mieux : parfois, la loi des variables aléatoires auxquelles on est confronté est une loi «classique», dont les propriétés sont bien connues, et qui apparaît systématiquement lorsque certaines propriétés générales sont présentes dans le modèle.

En ce sens, la notion de loi est portable, les calculs menés à partir de la loi ne faisant intervenir que le modèle «portable» (S_X, p_X), et non pas les détails du modèle $(\Omega, \mathbb{P})$ sous-jacent sur lequel X est définie – qui sont susceptibles de varier considérablement d'une situation à l'autre –, et une même loi est donc susceptible d'intervenir dans de très nombreux modèles, indépendamment des détails de ceux-ci.

Dans la suite, nous donnons une liste (non-exhaustive) de lois classiques, ainsi que les hypothèses qui permettent d'identifier immédiatement une variable aléatoire comme possédant une telle loi. La loi apparaîtra donc comme une notion portable, qui pourra souvent être manipulée sans autre référence au modèle probabiliste sous-jacent $(\Omega, \mathbb{P})$ que quelques propriétés générales, essentiellement d'indépendance, le plus souvent sans rapport avec la structure détaillée du modèle et de la situation que l'on modélise.

2.2.2 La loi dans l'interprétation fréquentielle de la probabilité – notion de loi empirique

Etant donné un échantillon de valeurs $\hat{x} = (x_1, \ldots, x_N)$, et un ensemble S (fini ou dénombrable) contenant $x_1, \ldots, x_N$, la **loi empirique** sur S associée à l'échantillon $\hat{x}$ est celle qui attribue à chaque valeur $s \in S$ une probabilité égale à sa fréquence

relative d'apparition dans l'échantillon :

$$p_{emp.,\hat{x}}(s) = \frac{1}{N} \times \text{ nombre d'indices } i \text{ pour lesquels } x_i = s.$$

(Lorsqu'un élément de S n'apparaît pas dans l'échantillon, il est affecté d'une probabilité nulle.) Lorsqu'il n'y a pas d'ambiguïté concernant l'échantillon utilisé, on note parfois simplement $p_{emp.}$ la loi correspondante (mais cette loi dépend néanmoins de l'échantillon $\hat{x}$ utilisé pour la définir !).

La loi empirique associée à un échantillon n'est donc rien d'autre qu'une description de cet échantillon, au moyen des fréquences d'apparition des différentes valeurs dans cet échantillon (c'est la probabilité au sens des fréquences dans la population constituée exclusivement par les valeurs de l'échantillon) [2].

Une autre manière de présenter les choses est de dire que la loi empirique associée à $\hat{x}$ est la loi de probabilité d'un élément choisi au hasard selon la probabilité uniforme dans l'échantillon [3] $\hat{x}$.

La notion de loi empirique est fondamentale dans l'interprétation fréquentielle de la probabilité. **En effet, dans ce contexte, la loi (tout court) d'une variable aléatoire n'est autre que la loi empirique, dans la limite d'un grand nombre de répétitions de la situation considérée** (la manière dont les répétitions sont effectuées devant bien entendu être définie avec précision, – cela fait partie de la définition de la probabilité dans ce contexte – et assurer la stabilisation des fréquences à long terme).

Dans le cadre de cette interprétation, et de la même manière que l'on distingue un terme d'une suite de la limite de celle-ci, on sera amener à distinguer ce que l'on appelle la **loi théorique** d'une variable aléatoire, qui correspond à la limite de la loi empirique dans l'idéalisation d'un nombre infini de répétitions, de la loi empirique associée à un échantillon donné. Comme nous l'avons déjà noté au chapitre précédent, cette notion de loi théorique n'est qu'une idéalisation, et les probabilités qui lui sont associées ne sauraient en réalité être définies à mieux qu'un certain degré d'imprécision près. Cependant, cette idéalisation est très utile en tant qu'outil conceptuel.

Juste pour fixer les idées, voici les résultats obtenus avec quelques simulations menées à l'aide du logiciel `R` et censées simuler des répétitions indépendantes de lancers de pile et face équiprobables. La loi théorique est donc ici la loi sur l'ensemble {P,F} attribuant à P et à F une probabilité de 1/2. En effectuant 100 simulations

2. Lorsque la probabilité utilisée dans le modèle désigne simplement la fréquence au sein d'une certaine population, il importe de ne pas confondre la loi empirique associé à un échantillon tiré de cette population, et la loi globale, qui correspond aux fréquences au sein de la population totale.

3. Rappelons ici que la probabilité au sens des fréquences au sein d'une population peut également se voir comme la probabilité fréquentielle associée à des tirages aléatoires uniformes répétés au sein de cette population.

de lancer, on a trouvé 44 fois pile, et 56 fois face. La loi empirique associée à cet échantillon simulé correspond donc à une probabilité empirique de $0,44$ pour pile, et de $0,56$ pour face. Bien entendu, 100 nouvelles simulations de lancers donneront en général lieu à une loi empirique différente. Hop ! Un nouveau tirage nous donne 52 pile et 48 face, la probabilité empirique décrivant ce tirage est donc $0,52$ pour pile et $0,48$ pour face. Plus la taille de l'échantillon est grande, plus on s'attend à ce que la loi empirique soit proche de la loi théorique. Avec 10000 simulations, on a trouvé 4934 fois pile, et 5067 fois face. En lançant une nouvelle simulation, nous obtenons 5042 fois pile et 4958 fois face. Avec 1000000 simulations, 500290 fois pile, et 499710 fois face. Au premier abord, ceci ne semble pas en contradiction flagrante avec notre idéalisation d'une loi théorique de $1/2$ pour pile et $1/2$ pour face. Des exemples plus trépidants sont présentés sous forme de graphiques dans la suite. Nous reviendrons beaucoup plus en détail sur ces questions dans les chapitres suivants (loi des grands nombres et courbe en cloche).

Concluons par une remarque terminologique.

Remarque 2 *Dans le langage courant, le terme de loi, appliqué à un phénomène naturel, ou, plutôt, à une classe de phénomènes naturels, désigne une propriété censée être vérifiée par l'ensemble des phénomènes de cette classe. En physique, on parle ainsi de la loi de la gravitation de Newton censée décrire les phénomènes d'attraction entre les corps pesants, de la loi des gaz parfaits, censée décrire la relation entre pression, volume et température pour une certaine catégorie de gaz, de la loi d'Archimède sur la force exercée par un liquide sur un corps solide immergé, ou encore la loi d'Ohm censée décrire le lien entre intensité et tension électriques dans certains types de matériaux conducteurs. Dans d'autres domaines, on parle par exemple de la loi de Moore (la puissance de calcul permise par les ordinateurs double environ tous les 18 mois), de la loi des rendements décroissants en économie, etc... (sans oublier la loi de la jungle ou la loi de Murphy). Bien entendu, ces différentes lois ne sont pas toutes de même nature et n'ont pas toutes le même statut. Nous vous renvoyons à un ouvrage d'épistémologie et/ou d'histoire des sciences pour une discussion de la notion de loi dans ce contexte. Retenons simplement le rôle fondamental joué par la vérification empirique des lois, c'est-à-dire la confrontation de leurs prédictions à la réalité observée. En général, lorsqu'une loi est qualifiée d'«empirique», c'est pour souligner qu'elle correspond effectivement aux observations, mais que l'on ne dispose pas d'arguments théoriques permettant de la justifier.*

La terminologie employée en probabilités recoupe plus ou moins ces usages du mot «loi». Ainsi, dans l'interprétation fréquentielle de la probabilité, la loi (au sens usuel, non-probabiliste du terme) attachée à une quantité aléatoire est que, au cours d'un grand nombre d'expériences répétées fournissant des mesures de cette quantité (les conditions de répétition devant naturellement être précisées), les fréquences d'appari-

tion des différentes valeurs que cette quantité peut prendre se stabilisent au voisinage de limites qui sont justement décrites par la loi de probabilité de la variable aléatoire modélisant cette quantité. Comme toute loi (au sens non-probabiliste), celle-ci peut-être fausse ou approximative, plus ou moins bien vérifiée en pratique, et plus ou moins bien étayée par des arguments théoriques.

Dans ce contexte, la loi de probabilité empirique attachée à un échantillon de valeurs mesurées de cette quantité est simplement le résumé de l'information obtenue expérimentalement sur la répartition des valeurs de celle-ci. Lors de l'élaboration d'un modèle probabiliste d'une situation, et notamment de la spécification des lois de probabilité des variables aléatoires intervenant dans le modèle, les lois empiriques associées à des valeurs mesurées sont l'un des éléments fondamentaux (parfois le seul – auquel cas le modèle peut être considéré comme complètement empirique – mais souvent accompagné de considérations théoriques, connaissances ou hypothèses sur le phénomène étudié) de la démarche.

2.2.3 Fonction de répartition d'une loi discrète

Si p est une probabilité sur un sous-ensemble fini ou dénombrable S de $\mathbb{R}$, on définit la **fonction de répartition** de p comme la fonction définie sur $\mathbb{R}$ par

$$F_p(x) = p\left(\{s \in S \ : \ s \leq x\}\right).$$

Par définition,

$$F_p(x) = \sum_{s\in \ : \ s\leq x} p(s).$$

On définit la fonction de répartition d'une variable aléatoire X à valeurs réelles et définie sur $(\Omega, \mathbb{P})$ par $F_X = F_{p_X}$, où, plus explicitement,

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{s\in S_X \ : \ s\leq x} \mathbb{P}(X = x).$$

On vérifie que la fonction F_p est croissante, et qu'il s'agit en fait d'une fonction en escalier continue à droite (si on lit le graphe de F_p dans le sens des abscisses croissantes, la fonction effectue des sauts vers le haut aux points dont les abscisses correspondent aux éléments de S.

On vérifie que $\lim_{x\to-\infty} F_p(x) = 0$, et que $\lim_{x\to-\infty} F_p(x) = 1$.

La connaissance de F_p est équivalente à celle de la loi de p, car, pour $x, y \in S$ tels que $x < y$ et $]x, y[\cap S = \emptyset$, on a $p(y) = F_p(y) - F_p(x)$.

2.2.4 Représentations graphiques

La loi d'une variable aléatoire discrète est la donnée, pour chaque valeur que peut prendre cette variable aléatoire, de la probabilité attachée à cette valeur. Il

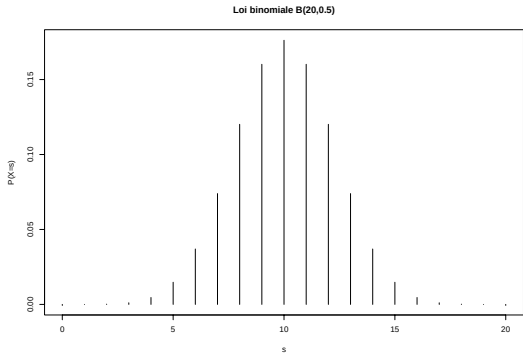
est très utile de disposer de diverses représentations graphiques d'une loi, car cela permet de saisir visuellement un certain nombre de propriétés qu'il serait parfois difficile de dégager directement d'une liste ou d'un tableau de nombres, d'une formule, ou d'indicateurs numériques synthétiques (tels qu'espérance, médiane, écart-type,..., nous en discuterons dans la suite).

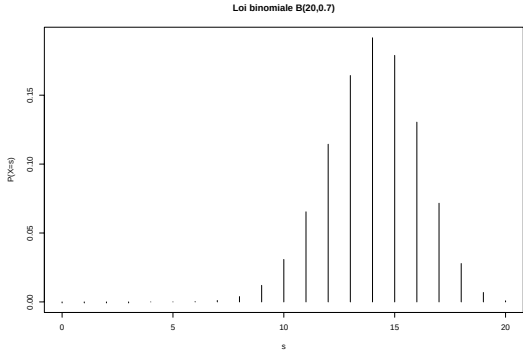
Nous nous limiterons essentiellement au cas le plus simple : la représentation graphique des lois de variables aléatoires à valeurs dans $\mathbb{R}$ – on parle de lois univariées (avec une brève excursion au cas bi-varié dans le chapitre sur la régression). Il existe par ailleurs de très nombreux outils destinés à traiter le cas de variables aléatoires qualitatives, ou de variables multi-dimensionnelles.

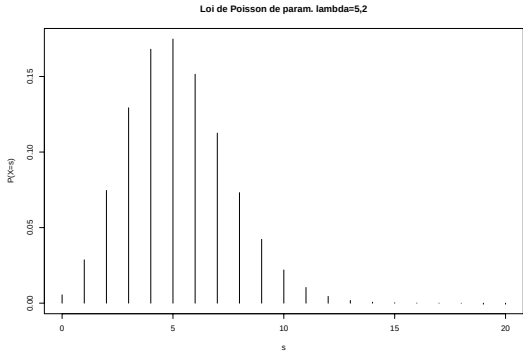
Nous vous renvoyons à un cours de statistique descriptive pour une description détaillée des divers types de représentation graphique, dont nous ne présentons dans la suite que quelques exemples.

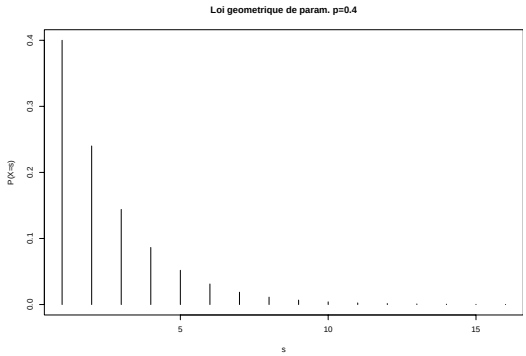
Diagramme en bâtons

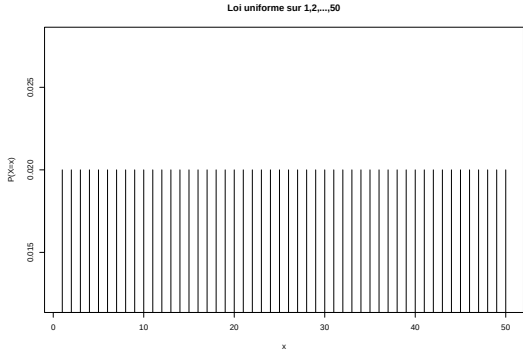
La représentation la plus simple, ou diagramme en bâtons, consiste à tracer, en regard de chaque valeur possible de la variable, un trait dont la hauteur représente la probabilité associée à cette valeur. Voici quelques exemples de tels diagrammes (nous donnerons plus bas les définitions exactes des lois qui sont représentées).



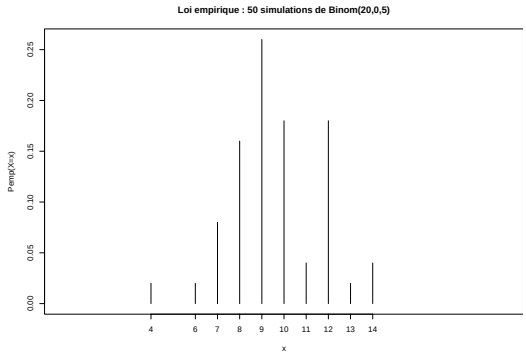

Loi binomiale B(20,0.7)
P(X<=s)
0.00
0.05
0.10
0.15
0
5
10
15
20
s

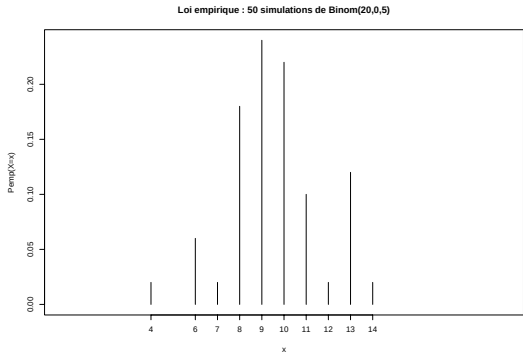

Loi de Poisson de param. lambda=5,2
P(X<=s)
0.00
0.05
0.10
0.15
0
5
10
15
20
s

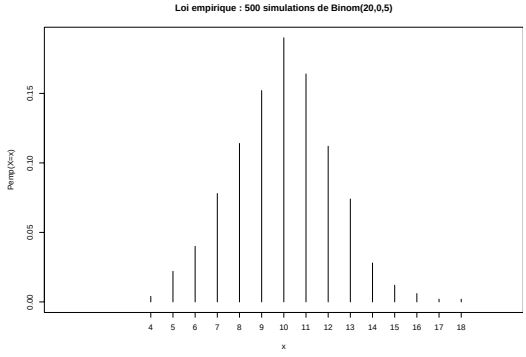

Loi geometrique de param. p=0.4
P(X<=s)
0.0
0.1
0.2
0.3
0.4
5
10
15
s

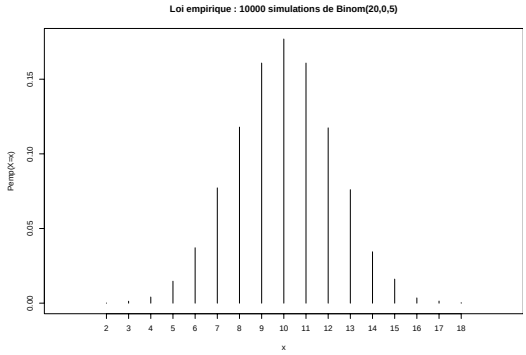


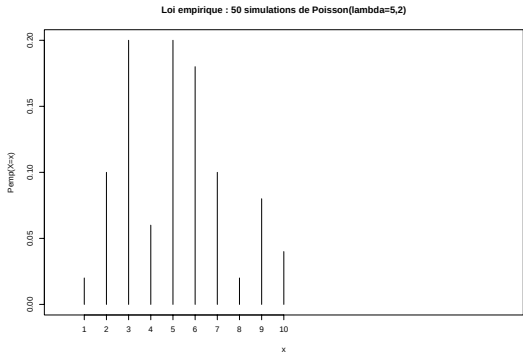
Voici à présent les diagrammes en bâtons obtenus à partir des lois empiriques associées à des échantillons simulés (simulations menées sous `R`) de variables aléatoires.

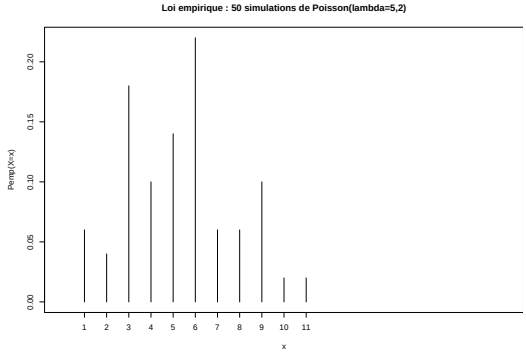


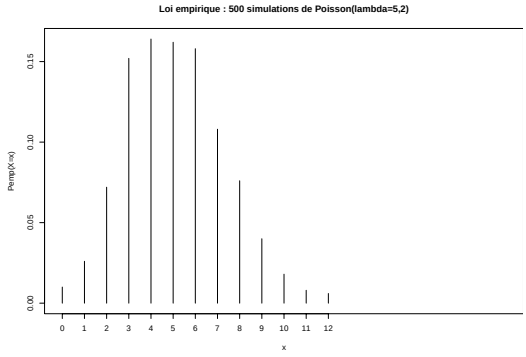


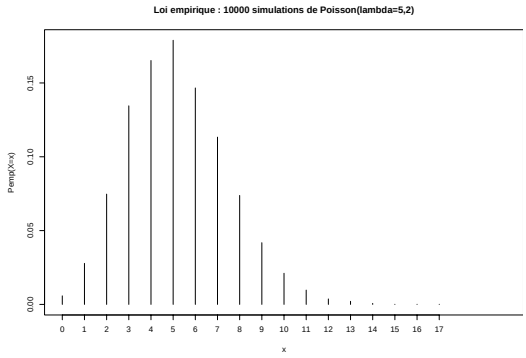

Loi empirique : 500 simulations de Binom(20,0,5)
Pemp(X<=x)
0.00
0.05
0.10
0.15
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
x

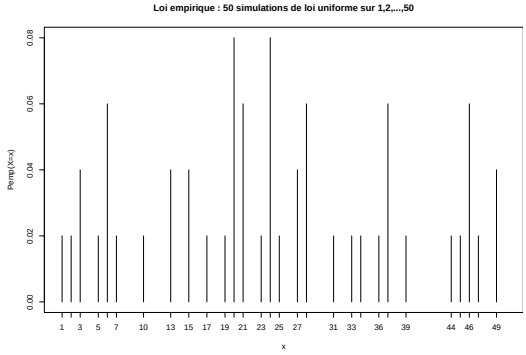

Loi empirique : 10000 simulations de Binom(20,0,5)
Pemp(X<=x)
0.00
0.05
0.10
0.15
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
x

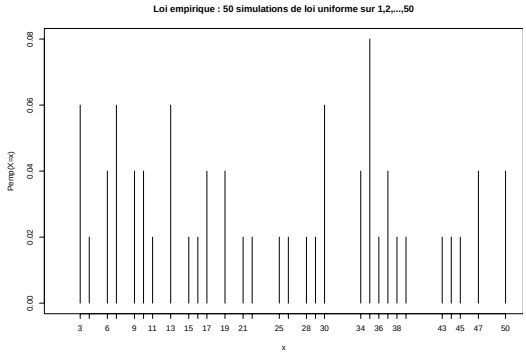

Loi empirique : 50 simulations de Poisson(lambda=5,2)
Pemp(X<=x)
0.00
0.05
0.10
0.15
0.20
1 2 3 4 5 6 7 8 9 10
x

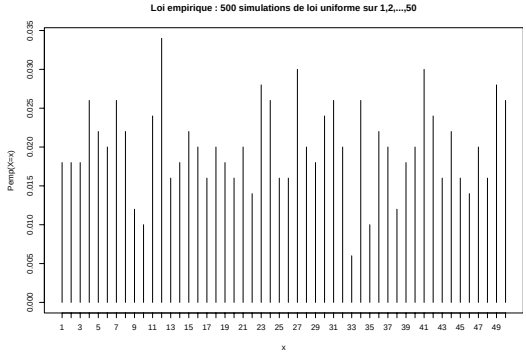

Loi empirique : 50 simulations de Poisson(lambda=5,2)
Pemp(X=x)
0.00
0.05
0.10
0.15
0.20
1 2 3 4 5 6 7 8 9 10 11
x


Loi empirique : 500 simulations de Poisson(lambda=5,2)
Pemp(X=x)
0.00
0.05
0.10
0.15
0 1 2 3 4 5 6 7 8 9 10 11 12
x


Loi empirique : 10000 simulations de Poisson(lambda=5,2)
Pemp(X=x)
0.00
0.05
0.10
0.15
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
x


Loi empirique : 50 simulations de loi uniforme sur 1,2,...,50
Pemp(X=x)
0.00
0.02
0.04
0.06
0.08
1 3 5 7 10 13 15 17 19 21 23 25 27 31 33 36 39 44 46 49
x

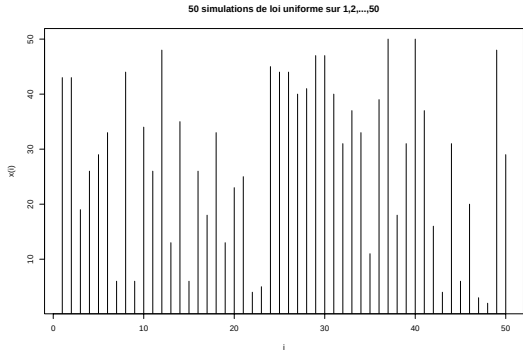

Loi empirique : 50 simulations de loi uniforme sur 1,2,...,50
Pemp(X=x)
0.00
0.02
0.04
0.06
0.08
3 6 9 11 13 15 17 19 21 25 28 30 34 36 38 43 45 47 50
x


Loi empirique : 500 simulations de loi uniforme sur 1,2,...,50
Pemp(X=x)
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
x

On observe bien la différence existant entre loi théorique d'une part, et, d'autre part loi empirique associée à un échantillon produit par simulation, et censé être modélisé par cette loi théorique.

Mise en garde 5 *Il importe de ne pas confondre les diagrammes ci-dessus, qui représentent la loi empirique associée à des échantillons de la forme $x_1, \ldots, x_N$, avec le tracé de x_i en fonction de i (qui peut avoir un intérêt, mais n'a pas de rapport avec ce qui est représenté ci-dessus).*

Pour bien saisir la différence, voici à quoi peut ressembler un tel tracé.

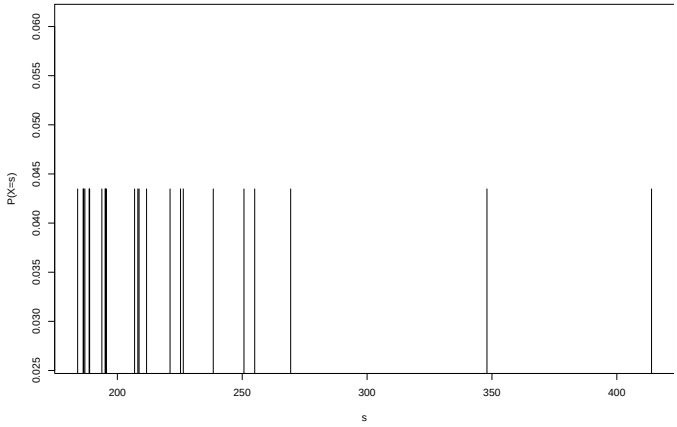


On remarque au passage que la loi empirique associée à un échantillon ne dépend pas de l'ordre dans lequel les valeurs apparaissent dans l'échantillon. Si celui-ci reflète, par exemple, un ordre chronologique entre des mesures effectuées à des dates différentes, ou, plus généralement, si celui-ci présente un rapport avec la situation étudiée, ne retenir d'un échantillon que sa loi empirique peut donc conduire à totalement ignorer certaines structures présentes dans l'échantillon et potentiellement importantes dans l'étude de la situation considérée. Voir le chapitre «Statistique» à ce sujet.

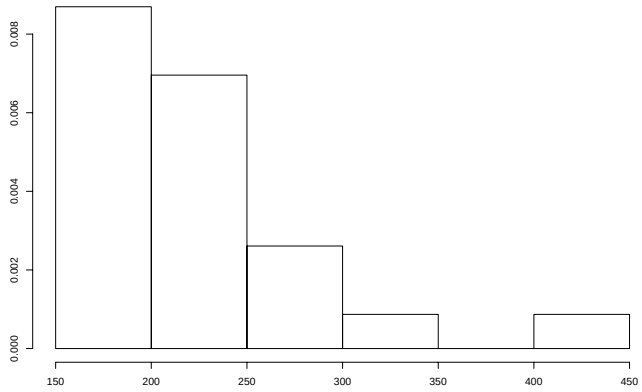
Histogramme

Lorsque la variable aléatoire à laquelle on s'intéresse prend un grand nombre de valeurs distinctes, (cela sera en particulier le cas des lois empiriques associées aux lois continues, que nous étudierons ultérieurement) il est souvent plus commode de regrouper entre elles les valeurs proches, et de représenter la loi sous la forme d'un histogramme constitué d'un nombre limité de barres, la surface des barres représentant la probabilité qu'une valeur se trouve dans l'intervalle formant la base de cette barre. L'information contenue dans l'histogramme est donc moins détaillée que celle que fournit un diagramme en bâtons, – l'histogramme ne permet pas en général de retrouver la loi, et il y a en ce sens perte d'information lorsque l'on utilise cette représentation – mais elle en fournit un «résumé» souvent plus lisible. Plus précisément, pour construire un histogramme décrivant la loi d'une variable aléatoire X, on fixe une largeur Δ, et l'on découpe l'ensemble des valeurs que peut prendre X en **classes** deux-à-deux disjointes de la forme $[a_i, a_{i+1}[$ (ou $]a_i, a_{i+1}]$, cela dépend des définitions). Au-dessus de chaque intervalle $[a_i, a_{i+1}[$, on trace une barre dont la surface est proportionnelle à $\mathbb{P}(X \in [a_i, a_{i+1}[)$. La hauteur de la barre située au-dessus de l'intervalle $[a_i, a_{i+1}[$ est donc proportionnelle à $\mathbb{P}(X \in [a_i, a_{i+1}[)/(a_{i+1} - a_i)$. On choisit la plupart du temps des classes de même largeur, c'est-à-dire pour lesquelles $a_{i+1} = a_i + \Delta$ pour tout i. Le choix des paramètres de l'histogramme (largeur des classes, points de borne inférieure et supérieure, échelles) contient une part d'arbitraire, et différentes règles automatiques de choix de ces paramètres sont utilisées par les logiciels de statistique tels que `R`.

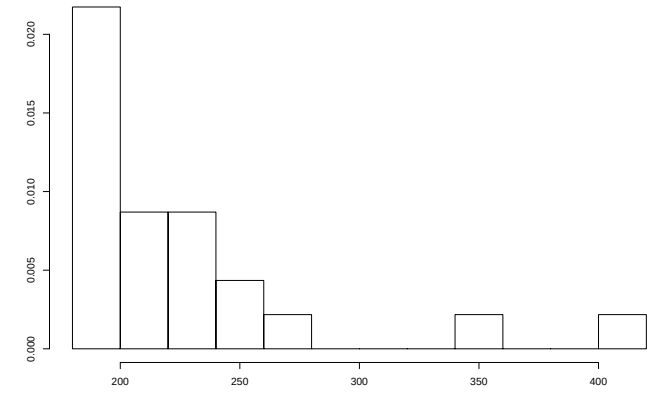
Exemple 1 *Jojo mesure le temps de transmission d'un message de son ordinateur à un autre par le réseau internet, à divers moments. Il recueille les données suivantes (temps exprimé en millisecondes) : 188,9 ; 188,7 ; 184,1 ; 348,0 ; 187,0 ; 195,3 ; 255,0 ; 413,9 ; 225,3 ; 221,1 ; 269,4 ; 208,7 ; 211,7 ; 206,9 ; 226,4 ; 186,5 ; 193,8 ; 208,2 ; 238,4 ; 250,7 ; 195,1 ; 186,3 ; 195,6. Si l'on effectue un diagramme en bâtons de la loi empirique associée à cet échantillon, on obtient, toutes les valeurs de l'échantillons étant distinctes (et donc affectées chacune d'une probabilité égale à 1/23), le diagramme suivant, dont l'aspect abscons devrait vous convaincre de l'intérêt d'utiliser des histogrammes :*



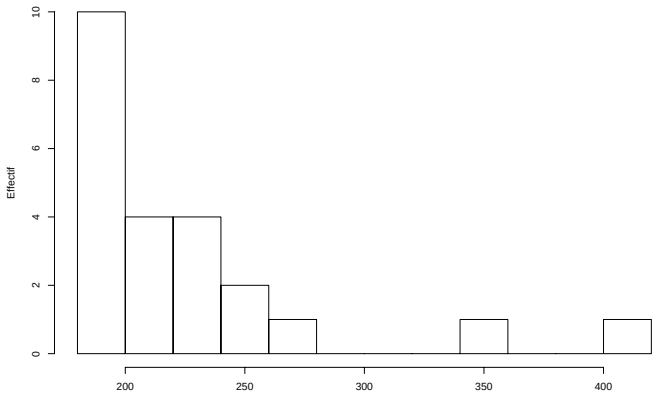
La représentation par un histogramme de la loi empirique associée à cet échantillon de valeurs donne le résultat (plus parlant) suivant, en choisissant une largeur de classes égale à 50ms :



En réduisant la largeur d'une classe à 20ms, on obtient le résultat (plus précis, mais moins lisible) suivant :



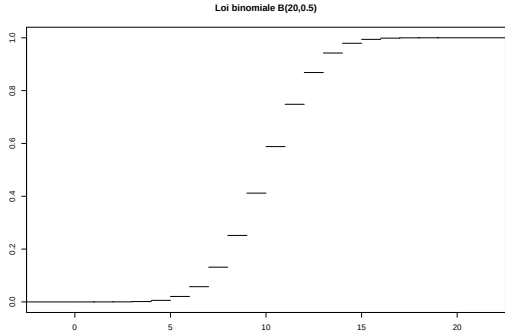
Lorsqu'on représente, comme nous venons de la faire, une loi empirique associée à un échantillon de données, on indique parfois en ordonnée **l'effectif** correspondant à la barre, c'est-à-dire le nombre de valeurs de l'échantillon qui se trouvent dans l'intervalle délimité par la base de la barre. Par exemple :

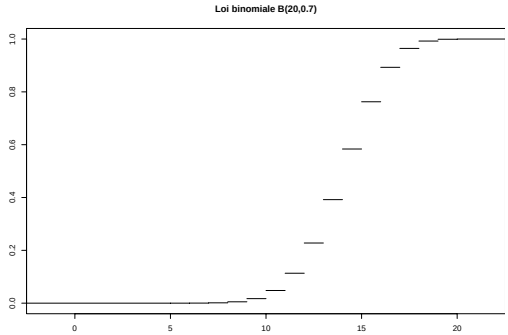


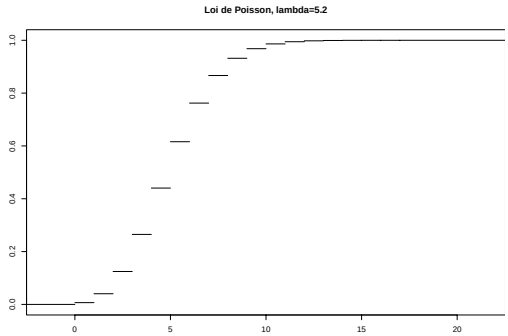
On note que, dans ce cas, il n'est pas forcément pertinent de comparer des histogrammes associés à des échantillons de tailles différentes en les superposant, pour d'évidentes raisons d'échelle.

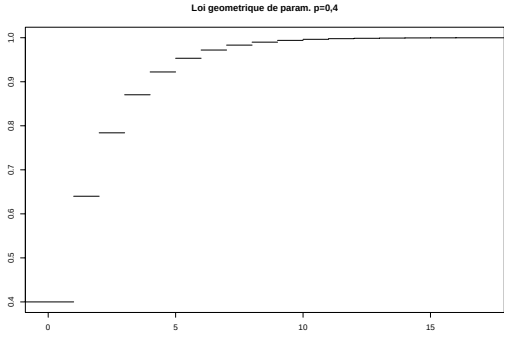
Représentation graphique de la fonction de répartition

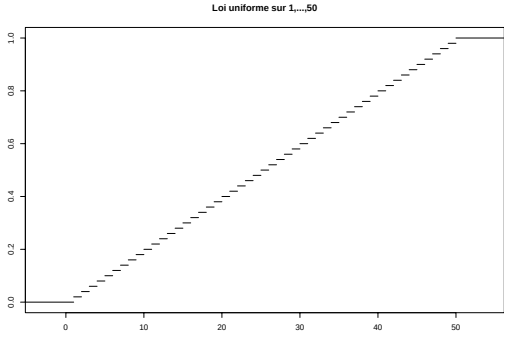
Il s'agit simplement de représenter le graphe de la fonction de répartition. Voici quelques exemples reprenant les lois représentées précédemment.

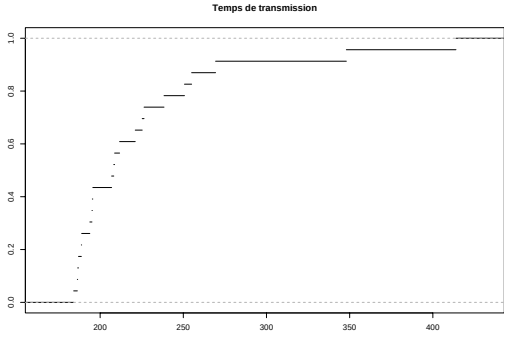

Loi binomiale B(20,0.5)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


Loi binomiale B(20,0.7)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20

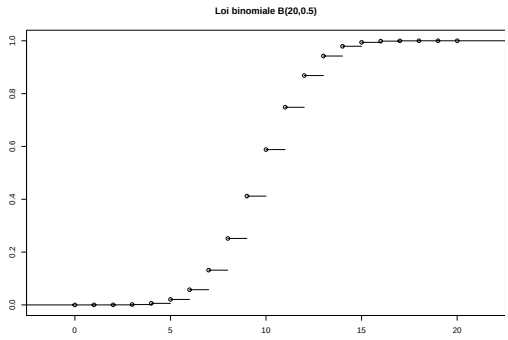

Loi de Poisson, lambda=5.2
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20







Pour marquer le fait que la fonction de répartition est continue à droite, on note parfois les points situés aux bords gauches des sauts, comme ceci.



Autres représentations graphiques

Il existe de nombreux autres types de représentations graphiques, pouvant servir à résumer avec plus ou moins de précision une loi de probabilité unidimensionnelle. Mentionnons, sans prétendre à l'exhaustivité, la représentation de la fonction de répartition et le boxplot, qui sont décrits plus bas. De nombreux raffinements (tels que l'emploi de procédés de lissage) de ces méthodes de base existent. (voir n'importe quel ouvrage comportant le mot «Statistique descriptive» dans son titre pour plus de détails).

Comparaison graphique de deux lois au moyen du tracé quantile-quantile *

Cette partie nécessite la connaissance de la notion de quantile d'une loi de probabilité, définie plus bas.

Le tracé quantile-quantile (appelé quantile-quantile plot, ou encore qq-plot en anglais) est une méthode de représentation graphique visant à comparer deux distributions de probabilité unidimensionnelles (la plupart du temps, au moins l'une des deux distributions est la loi empirique associée à un échantillon, l'autre pouvant soit être une loi théorique à laquelle on souhaite comparer la distribution empirique de l'échantillon, soit la loi empirique d'un autre échantillon, si l'on souhaite comparer entre elles les distributions des deux échantillons). En gros, le principe est le suivant : partant de deux distributions de probabilité μ_X et μ_Y, on représente les couples de la forme (x_r, y_r), où x_r (resp. y_r) désigne le fractile d'ordre r de la loi μ_X (resp. μ_Y).

De fait, il existe une certaine latitude dans la définition exacte du tracé (ce que nous venons d'en dire ne suffit pas à le spécifier complètement), et l'on rencontre différentes versions du tracé quantile-quantile suivant les logiciels que l'on utilise. Nous n'entrerons pas dans ces détails, et nous nous contenterons d'illustrer ce type de tracé au moyen de quelques exemples.

Des exemples...

2.2.5 Quelques lois discrètes classiques

Loi de Bernoulli

On dira qu'une variable aléatoire X définie sur un espace probabilisé $(\Omega, \mathbb{P})$ suit une loi de Bernoulli de paramètre $p \in [0,1]$ si elle prend la valeur 1 avec probabilité p et la valeur 0 avec la probabilité $1-p$. N'importe quelle variable aléatoire ne pouvant prendre que les valeurs 0 et 1, par exemple, n'importe quelle fonction indicatrice d'un événements, suit donc une loi de Bernoulli.

Loi binomiale

On dira qu'une variable aléatoire X définie sur un espace probabilisé $(\Omega, \mathbb{P})$ suit une loi binomiale de paramètres $n \geq 0$ et $p \in [0,1]$ si elle ne prend que des valeurs entières entre 0 et n, avec les probabilités :

$$\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k} \ , \ 0 \leq k \leq n.$$

Autrement dit, la loi binomiale (sans référence à une variable aléatoire) est la probabilité $p_{binom(n,p)}$ sur l'ensemble $\{0, \ldots, n\}$ définie par :

$$p_{binom(n,p)}(k) = C_n^k p^k (1-p)^{n-k}.$$

Expliquons dans quel contexte cette loi intervient : supposons que, sur un espace probabilisé $(\Omega, \mathbb{P})$, nous nous intéressions à la réalisation de n événements $A_1, \ldots, A_n$, mutuellement indépendants, et survenant chacun avec une probabilité commune égale à p :

$$\mathbb{P}(A_1) = \cdots = \mathbb{P}(A_n) = p,$$

et définissons X comme le nombre total (aléatoire) d'événements A_i qui se réalisent effectivement. Par exemple, nous effectuons n répétitions indépendantes d'une même expérience, qui peut à chaque répétition donner lieu à un succès (avec probabilité p) ou à un échec (avec probabilité $(1-p)$), et nous comptons le nombre total (aléatoire) de succès obtenus au cours des n expériences. Alors, X suit une loi binomiale de paramètres n et p. Pour le voir, il suffit de vérifier que l'événement : réalisation d'exactement k événements parmi les A_i, c'est-à-dire la réalisation de k d'entre eux, et la non-réalisation des $n-k$ restants, peut s'écrire comme la réunion de C_n^k événements deux-à-deux disjoints, chacun de probabilité $p^k(1-p)^{n-k}$. En effet, il y a C_n^k sous-ensembles d'indices I inclus dans $\{1; 2; \ldots; n\}$ comportant k éléments, et, pour tout tel sous-ensemble d'indices I, l'événement E_I : réalisation des k événements A_i dont les indices se trouvent dans I, et non-réalisation de ceux des A_i dont les indices

ne figurent pas dans I, possède, du fait de l'indépendance mutuelle des A_i, une probabilité égale à $p^k(1-p)^{n-k}$. Pour deux sous-ensembles distincts I_1 et I_2 d'indices, les deux événements E_{I_1} et E_{I_2} sont incompatibles (leur réalisation simultanée exige en même temps la réalisation et la non-réalisation d'au moins l'un des A_i), d'où la formule ci-dessus. Il conviendrait de vérifier que l'on a bien

$$\sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} = 1,$$

afin de prouver que l'on a bien défini une loi de probabilité. Cette égalité résulte de la formule du binôme de Newton appliquée à $(p + (1-p))^n$. Cependant, elle résulte également du fait que nous avons prouvé que $p_{binom(n,p)}(k)$ apparaît effectivement comme la loi d'une variable aléatoire dans un contexte particulier.

Mise en garde 6 *Un raisonnement erronné donne comme résultat $\mathbb{P}(X = k) = p^k$ au lieu de $C_n^k p^k (1-p)^{n-k}$: on demande qu'il y ait k succès, tous indépendants de probabilité p, d'où p^k. Autre raisonnement erronné : on demande qu'il y ait k succès, tous indépendants de probabilité p, d'où p^k, et, indépendamment, $n-k$ échecs tous indépendants de probabilité $1-p$, d'où $p^k(1-p)^{n-k}$.*

Exemple 2 *On administre un traitement à 200 malades, et, pour chaque malade, la probabilité que le traitement soit efficace est de 90%. Si l'on suppose que les guérisons des différents malades forment une famille d'événements mutuellement indépendants, ou, ce qui revient au même, que les guérisons des malades peuvent être modélisées par une succession indépendante d'épreuves (guérison/non-guérison), le nombre total de malades qui guérissent suit une loi binomiale de paramètres 200 et 90%.*

Exemple 3 *On effectue 50 lancers successifs d'un dé. En supposant que les résultats des lancers peuvent être modélisés par une succession indépendante, et que le lancer d'un dé est toujours décrit par la probabilité uniforme, le nombre total de 5 que l'on obtient après les 50 lancers suit une loi binomiale de paramètres 50 et 1/6.*

Remarque 3 *On constate que l'on peut être en présence de variables aléatoires suivant la loi binomiale même (et surtout) lorsque le modèle ne se résume pas à une succession d'épreuves indépendantes de Bernoulli (c'est-à-dire ne possédant que deux issues). Par exemple, le modèle de succession indépendante de 50 lancers de dé contient plus d'information que le simple fait que le 5 sorte ou ne sorte pas, pour chaque lancer. Globalement, il suffira qu'un modèle moins fin que $(\Omega, \mathbb{P})$ mais compatible avec celui-ci soit effectivement constitué par une succession indépendante d'épreuves de Bernoulli (par exemple, dans le cas précédent, le modèle qui ne tient compte, pour chaque lancer, que du fait d'obtenir ou non un 5).*

Loi uniforme

On dira qu'une variable aléatoire X définie sur un espace probabilisé $(\Omega, \mathbb{P})$ suit la loi uniforme si elle ne peut prendre qu'un nombre fini de valeurs, chaque valeur étant affectée de la même probabilité. Autrement dit, si l'ensemble des valeurs que peut prendre X est $S_X = \{s_1, \ldots, s_p\}$,

$$\mathbb{P}(X = s_i) = \frac{1}{p}, \text{ pour tout } 1 \leq i \leq p.$$

La loi uniforme sur S_X (sans référence à une variable aléatoire) est la probabilité $p_{unif(S)}$ sur l'ensemble S_X est donc définie par :

$$p_{unif(S)}(s_i) = \frac{1}{p} \text{ pour tout } 1 \leq i \leq p.$$

Mise en garde 7 *Le fait que X suive la loi uniforme n'implique pas que l'espace de probabilité sous-jacent $(\Omega, \mathbb{P})$ soit muni de la probabilité uniforme. Par exemple, si $\Omega = \{0,1\} \times \{0,1\}$ et $\mathbb{P}(0,0) = 1/3$, $\mathbb{P}(0,1) = 1/6$, $\mathbb{P}(1,0) = 1/6$, $\mathbb{P}(1,1) = 1/3$, la variable aléatoire définie par $X(x_1, x_2) = x_1 + x_2$ suit la loi uniforme sur $\{0,1,2\}$, mais $\mathbb{P}$ n'est manifestement pas la probabilité uniforme sur Ω. Inversement, le fait que $\mathbb{P}$ soit la probabilité uniforme n'entraîne pas que X suive la loi uniforme : si $\Omega = \{0,1\} \times \{0,1\}$ et si $\mathbb{P}$ est la probabilité uniforme sur Ω, $\mathbb{P}(X = 1) = 1/2$ alors que $\mathbb{P}(X = 0) = 1/4$, et X ne suit donc pas la loi uniforme.*

Loi de Poisson

Cette loi tire son nom du mathématicien Poisson [4], et ne présente donc a priori pas de rapport avec la pêche en mer ou l'aquariophilie.

On dira qu'une variable aléatoire X définie sur un espace probabilisé $(\Omega, \mathbb{P})$ suit une loi de Poisson de paramètre $\lambda > 0$ si elle ne prend que des valeurs entières positives ou nulles, avec les probabilités :

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \ , \ k \geq 0.$$

Autrement dit, la loi de Poisson (sans référence à une variable aléatoire) est la probabilité $p_{Poiss(\lambda)}$ sur l'ensemble $\mathbb{N}$ définie par :

$$p_{Poiss(\lambda)}(k) = \frac{\lambda^k}{k!} e^{-\lambda} \ , \ k \geq 0.$$

Cette loi intervient dans le même contexte général que la loi binomiale : celle d'un comptage du nombre de succès enregistrés au cours d'une succession indépendante

4. Siméon Denis Poisson (1781–1840)

d'expériences ayant chacune la même probabilité de succès, mais dans un régime asymptotique particulier : celui où le nombre de répétitions est très grand, la probabilité de succès étant elle-même très petite. Considérons une variable aléatoire X_n de loi binomiale de paramètres n et $p = \lambda/n$, définie sur un espace de probabilité $(\Omega_n, \mathbb{P}_n)$, n étant suffisament grand pour que $\lambda/n \leq 1$. Pour tout entier k fixé et tout entier $n \geq k$, on a, par définition,

$$\mathbb{P}_n(X_n = k) = C_n^k \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Lorsque n tend vers l'infini, k étant fixé, on a

$$C_n^k = \frac{n \times (n-1) \times \cdots \times (n-k+1)}{k!} \sim \frac{n^k}{k!},$$

et

$$\left(1 - \frac{\lambda}{n}\right)^{n-k} \to e^{-\lambda}.$$

D'où, lorsque n tend vers l'infini,

$$\mathbb{P}_n(X_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda},$$

et la loi de Poisson apparaît comme un cas limite de la loi binomiale de paramètres n et λ/n lorsque n est grand. Bien que la probabilité de succès tende vers 0 lorsque n tend vers l'infini, ceci est compensé par le grand nombre d'expériences, de telle sorte que le nombre de succès observé prend des valeurs finies et non-nulles avec une probabilité qui ne tend pas vers zéro lorsque n tend vers l'infini. Pour cette raison, on appelle parfois le résultat ci-dessus le théorème des événements rares.

Remarque 4 *Plus généralement, on peut vérifier, exactement de la même manière, que la loi binomiale de paramètres n et p_n tend vers une loi de Poisson de paramètre λ pourvu que $\lim_{n\to+\infty} np_n = \lambda$. Voir l'exercice 114.*

Remarque 5 *Concernant la quantification de l'approximation de la loi binomiale par la loi de Poisson, on peut par exemple prouver l'inégalité suivante (voir l'ouvrage de Shiryaev cité dans la bibliographie) : pour tout λ tel que $0 < \lambda/n < 1$,*

$$\sum_{k=0}^{+\infty} \left| p_{Poiss(\lambda)}(k) - p_{binom(n,\lambda/n)}(k) \right| \leq \frac{2\lambda}{n} \min(2, \lambda).$$

Voir également à ce sujet l'exercice 115.

Pour vérifier que l'on a bien affaire à une loi de probabilité, il faut vérifier que l'on a :

$$\sum_{k=0}^{+\infty} p_{Poiss(\lambda)}(k) = 1,$$

ce qui est une conséquence de la formule :

$$e^{\lambda} = \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!}.$$

On peut aussi vérifier cette relation en passant à la limite dans la relation analogue valable pour la loi binomiale (mais le passage à la limite est un peu délicat).

Exemple 4 *Un fabricant d'écrans d'ordinateur s'intéresse au nombre de défauts présents sur la surface de ses écrans. Un écran est partagé en petites zones de contrôle deux-à-deux disjointes de surfaces égales, et l'on fait l'hypothèse que la présence ou l'absence de défauts de fabrication dans chacune de ces zones forment des événements mutuellement indépendants. De plus, on suppose que les zones de contrôle choisies sont suffisament petites pour qu'il puisse y avoir au plus un défaut par zone, et que la probabilité de trouver un défaut dans l'une de ces petites zones est proportionnelle à sa surface : plus la surface de la (petite) zone est grande, plus la probabilité d'y trouver un défaut est élevée.*

Quelle est la loi du nombre total de défauts présents sur l'écran ? Appelons n le nombre de zones de contrôle. Le nombre total de défauts est le nombre total d'événements «un défaut est présent» qui se réalisent parmi les n événements associés chacun à une zone de contrôle. Ces n événements étant mutuellement indépendants, le nombre total de défauts suit donc une loi binomiale de paramètres n et p, où p est la probabilité d'apparition d'un défaut sur une zone de contrôle. Cette probabilité étant, d'après notre hypothèse, proportionnelle à la surface de la zone, elle est de la forme

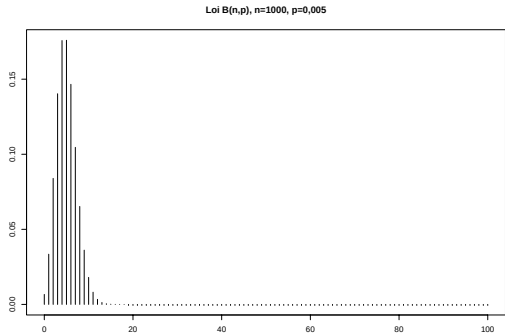
$$p = \alpha \frac{S}{n},$$

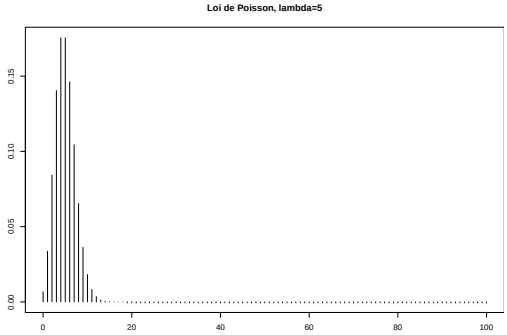
où S est la surface totale de l'écran, et α le coefficient de proportionnalité (ces deux quantités étant exprimées avec la même unité de surface). En posant $\lambda = \alpha S$, et en supposant n «grand», on constate que le nombre total de défauts présents sur l'écran suit (approximativement) une loi de Poisson de paramètre λ. Il est important de noter la différence entre le résultat que l'on obtient ici, et celui que l'on obtiendrait avec une loi binomiale $B(n,p)$, avec n grand et p de l'ordre de $1/3$ (par exemple). Dans le cas de la loi de Poisson, même si n est grand, les valeurs typiques prise par la variable aléatoire sont très grossièrement de l'ordre de λ, aussi grand n soit-il : pour un grand nombre d'expériences, le nombre de succès prend essentiellement des valeurs de l'ordre de quelques unités. À l'inverse, dans le cas de la loi binomiale

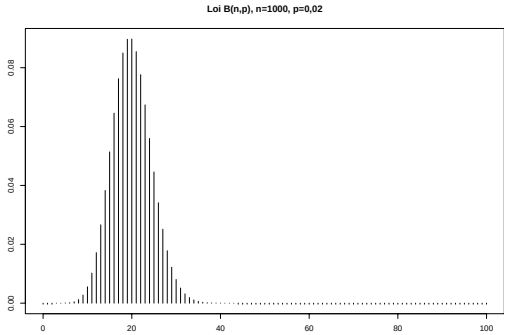
classique, le nombre de succès croît proportionnellement à n : avec n tentatives et une probabilité de succès de $1/3$, on s'attend à obtenir de l'ordre de $n/3$ succès.

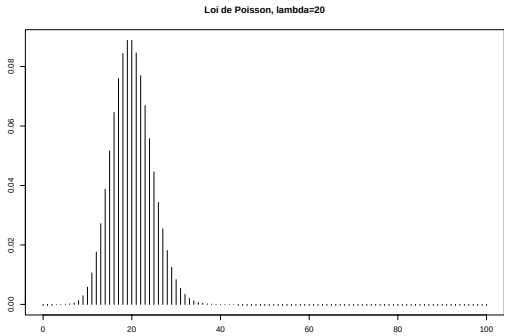
Exemple 5 *Des chasseurs à l'affût s'intéressent au nombre total de galinettes cendrées passant à leur portée en une journée. On admet qu'il peut passer au plus une galinette chaque minute, et que la probabilité de passage d'une galinette pendant un intervalle de temps (inférieur à une minute) donné est proportionnelle à la durée de cet intervalle, le coefficient de proportionnalité (pour une durée exprimée en secondes) étant noté β. De plus, on suppose que les passages de galinettes au cours d'intervalles de temps deux-à-deux disjoints sont mutuellement indépendants (ou, ce qui revient au même, que la succession des passages de galinettes, second après seconde, peut être modélisée par une succession indépendante d'épreuves de même probabilité de succès). De même que dans l'exemple précédent, le nombre total de galinettes qui passent en une journée est le nombre total d'événements «une galinette passe» qui se réalisent, parmi les n événements associés au découpage d'une journée en n «petits» intervalles de durée égale. Ici encore, on voit, en choisissant n «grand», que le nombre de galinettes qui passent en une journée suit approximativement une loi de Poisson de paramètre $\lambda = \beta D$, où D est la durée d'une journée en secondes et β le coefficient de proportionnalité.*

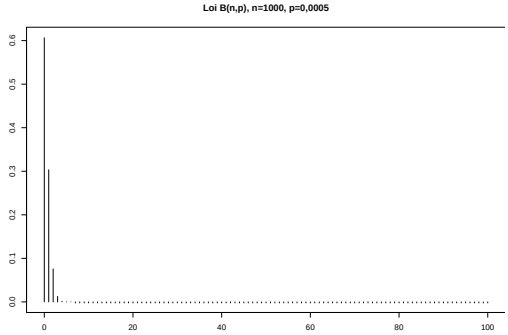
Voici quelques illustrations de l'approximation de la loi binomiale par la loi de Poisson.

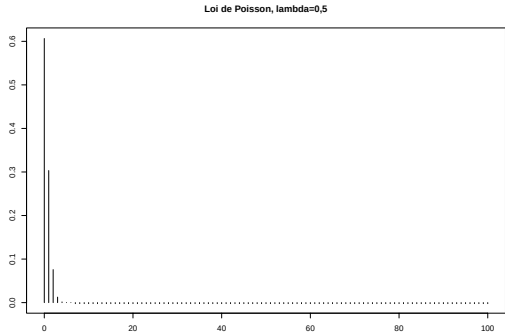



Loi de Poisson, lambda=5
0.00
0.05
0.10
0.15
0
20
40
60
80
100


Loi B(n,p), n=1000, p=0,02
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100


Loi de Poisson, lambda=20
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100





Loi géométrique

On dira qu'une variable aléatoire X définie sur un espace probabilisé $(\Omega, \mathbb{P})$ suit une loi géométrique de paramètre $p \in [0,1]$ si elle ne prend que des valeurs entières strictement positives, avec les probabilités :

$$\mathbb{P}(X = k) = (1-p)^{k-1}p \ , \ k \geq 1.$$

Autrement dit, la loi géométrique (sans référence à une variable aléatoire) est la probabilité $p_{geom(p)}$ sur l'ensemble $\mathbb{N}^*$ définie par :

$$p_{geom(p)}(k) = (1-p)^{k-1}p.$$

Expliquons dans quel contexte cette loi intervient. Supposons que nous répétions indépendamment une expérience aléatoire, chaque expérience étant susceptible de donner lieu à un certain événement appelé «succès» avec une probabilité p, jusqu'à obtenir un succès pour la première fois. Alors, le numéro de la première expérience se soldant par un succès suit une loi géométrique de paramètre p. Exercice : construisez un modèle en arbre rendant compte de cette situation, et prouvez la validité de la formule ci-dessus. Le seul point délicat est qu'il n'est pas évident *a priori* que le

nombre de tentatives nécessaires pour obtenir un succès est nécessairement fini. Après tout, il serait imaginable que l'on soit confronté à une succession infinie d'échecs...

Mise en garde 8 *On emploie parfois le terme de loi géométrique pour désigner la loi décalée définie par* $p(k) = p_{geom(p)}(k+1)$ *pour tout* $k \geq 0$*, et la terminologie est donc légèrement ambiguë.*

Exemple 6 *On suppose qu'à chaque seconde, la probabilité pour qu'un piéton traverse la rue est égale à* $1/10$*, et que les événements «un piéton traverse à la seconde numéro i» forment une succession indépendante. Alors, (en supposant que l'on observe le passage des piétons à partir de la seconde numéro 1) le numéro de la première seconde pendant laquelle un piéton va traverser suit une loi géométrique de paramètre* $1/10$*.*

Exemple 7 *Jojo lance des fléchettes sur une cible. On suppose que la probabilité pour qu'il atteigne le mille est de 25%, et que les succès de chaque tentative forment une succession indépendante. Combien de lancers Jojo doit-il effectuer avant d'atteindre le mille ? Un nombre aléatoire qui suit une loi géométrique de paramètre 25%.*

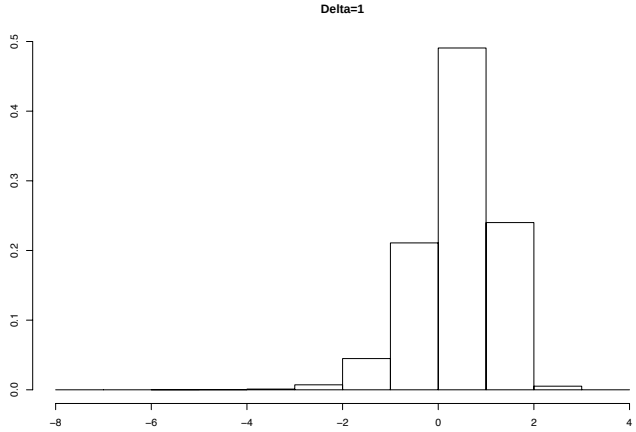
2.2.6 Variables aléatoires et lois continues

Introduction

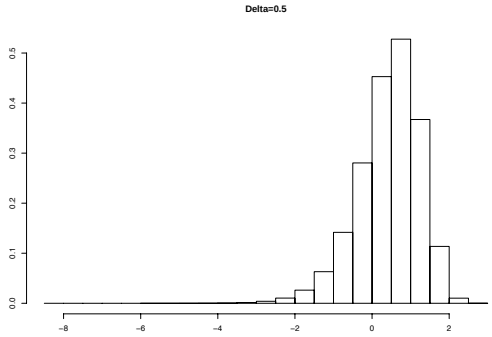
Les variables aléatoires, telles que les avons définies au début de ce chapitre, se rattachent à la catégorie dite des variables aléatoires discrètes. Celles-ci apparaissent comme des fonctions définies sur un ensemble Ω fini ou dénombrable, et, par conséquent, ne peuvent prendre qu'un nombre fini ou dénombrable de valeurs distinctes. Pour modéliser des quantités pouvant prendre un continuum de valeurs (par exemple l'ensemble des valeurs comprises dans un intervalle), il est donc nécessaire de faire appel à une définition plus générale des modèles probabilistes, englobant des espaces des possibles non-dénombrables. Il faut pour cela se placer dans le cadre de la théorie mathématique de la mesure abstraite, dont le niveau technique dépasse largement celui de ce cours (vous pouvez consulter les ouvrages classiques d'introduction à la théorie mathématique des probabilités cités en bibliographie pour en avoir un exposé). Par conséquent, nous nous contenterons, ce qui n'est pas absurde d'un point de vue pratique, de présenter les variables aléatoires continues comme un cas limite de variables aléatoires discrètes à une échelle microscopique, mais pouvant être considérées comme continues à une échelle plus macroscopique.

Commençons par un exemple de telle situation limite.

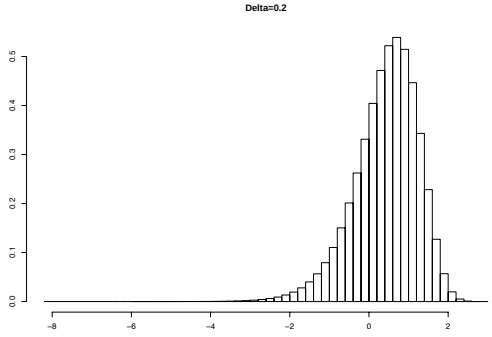
Voici l'histogramme d'une loi de probabilité, dans lequel la largeur des barres est fixée à $\Delta = 1$ (la graduation verticale représente la **surface** de chaque barre.)



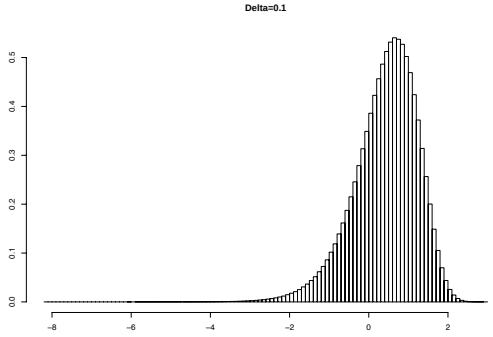
Pour la même loi de probabilité, réduisons progressivement la largeur des barres. Voici l'histogramme obtenu avec $\Delta = 0,5$.



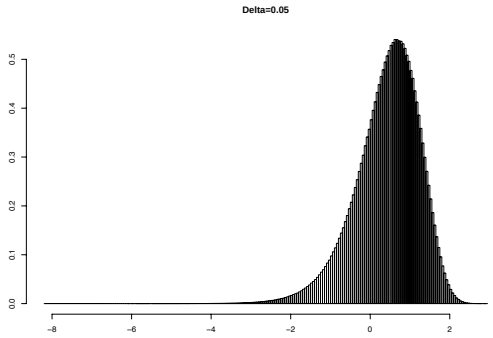
Réduisons encore : voici l'histogramme obtenu pour $\Delta = 0,2$.



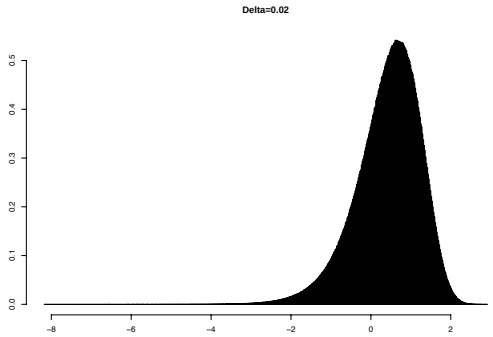
De plus en plus fort ! Voici l'histogramme obtenu pour $\Delta = 0,1$.



Voici encore l'histogramme obtenu pour $\Delta = 0,05$



Et enfin, l'histogramme obtenu pour $\Delta = 0,02$.



Que constate-t-on ? Si l'on examine le comportement d'une barre dont l'extrémité gauche prend une valeur fixée a, on constate que la hauteur de cette barre se rapproche, à mesure que la largeur Δ des barres diminue, d'une valeur $f(a)$, qui semble définir une fonction continue de a. Autrement dit, en se rappelant que c'est la surface d'une barre qui représente la probabilité de trouver une valeur dans l'intervalle

correspondant à la base de cette barre :

$$\lim_{\Delta \to 0} \frac{\mathbb{P}(X \in [a, a+\Delta[)}{\Delta} = f(a). \tag{2.1}$$

Nous trichons bien entendu en écrivant cette limite, car, en toute rigueur, la valeur de Δ diminue, mais Δ ne tend pas réellement vers zéro.

Une autre conséquence de nos observations est l'identité

$$\mathbb{P}(a \leq X < b) = \int_a^b f(u)du. \tag{2.2}$$

En effet, pour un intervalle de la forme $[a, a+n\Delta[$, on a

$$\mathbb{P}(a \leq X < a+n\Delta) = \sum_{i=1}^{n} \mathbb{P}(X \in [a(i-1)\Delta, a+i\Delta[).$$

En se rappelant que $\mathbb{P}(X \in [a(i-1)\Delta, a+i\Delta[)$ n'est autre que la surface de la barre de l'histogramme ayant pour base l'intervalle $[a(i-1)\Delta, a+i\Delta[$, et que l'intégrale d'une fonction continue représente la surface sous une portion de courbe, on se convainc aisément en observant les histogrammes précédent de la validité de l'identité (2.2), au moins en tant qu'égalité approchée valable lorsque Δ est petit.

Contrairement aux apparences, la loi représentée par les histogrammes précédents est une loi discrète, en fait simplement la loi empirique associée à un échantillon comportant un très grand nombre (10^7) de valeurs toutes distinctes, mais très proches les unes des autres. Si l'on regarde à la loupe la zone située autour de 1, par exemple, on pourra observer la répartition suivante (il s'agit d'un diagramme en bâtons, chaque bâton représente une valeur possible de la variable).

Pour des valeurs de Δ de l'ordre de 10^{-6}, la répartition, observée avec un pas de discrétisation de Δ, n'est donc plus du tout régulière. En revanche, lorsque Δ prend des valeurs de l'ordre de 10^{-1} ou 10^{-2}, comme nous avons pu le constater, les deux équations (2.1) et (2.2) précédentes constituent d'**excellentes approximations** de la situation réelle, et nous pouvons raisonner, à cette échelle, comme si les

valeurs de X formaient un continuum et non pas un ensemble discret. En revanche, sur une échelle plus fine, cette approximation n'est plus du tout pertinente. Ceci est tout-à-fait courant en physique, où l'on modélise en général les quantités macroscopiques telles que la masse ou le volume par des quantités continues, même si, à l'échelle atomique, ce type de description perd complètement sa pertinence. Notons également que, lorsque l'on effectue des simulations, du fait de la précision finie avec laquelle sont codés les nombres réels sur ordinateur (par exemple, une trentaine ou une cinquantaine de décimales), on ne manipule nécessairement que des variables aléatoires discrètes, même si celles-ci peuvent être considérées comme continues à l'échelle macroscopique.

Définition

Ceci nous conduit à poser **la définition générale d'une variable aléatoire continue à valeurs réelles : on dit que X est une variable aléatoire continue de densité f, où $f : \mathbb{R} \to \mathbb{R}$ est une fonction à valeurs positives ou nulles, que nous supposerons toujours continue par morceaux, lorsque, pour tout intervalle $]a, b[$, les bornes pouvant éventuellement être $-\infty$ ou $+\infty$, on a :**

$$\mathbb{P}(X \in]a, b[) = \int_a^b f(u)du.$$

(Pour des raisons techniques, il est plus commode d'utiliser la propriété (2.2) plutôt que la propriété (2.1) pour donner une définition générale, même si ces deux définitions sont essentiellement équivalentes dans la plupart des situations.) Comme nous l'avons mentionné plus haut, la définition ci-dessus n'est pas compatible avec le cadre des espaces de probabilités discrets auquel nous nous sommes confinés jusqu'à présent, et donc... ce n'est pas une véritable définition dans le cadre de ce cours, puisqu'aucune variable aléatoire telle que définie précédemment ne peut la satisfaire exactement.

Par exemple, en toute rigueur, la définition (2.1) ci-dessus entraîne que la probabilité $\mathbb{P}(X = b)$ est nulle pour tout b, donc que X ne peut prendre aucune valeur ! En effet, d'après la définition :

$$\mathbb{P}(X = b) \leq \mathbb{P}(X \in]b - \delta, b + \delta[) = \int_{b-\delta}^{b+\delta} f(u)du,$$

et, en faisant tendre δ vers zéro, on constate bel et bien que $\mathbb{P}(X = b) = 0$. Il faut donc – et, en tout cas, on peut sans aucune difficulté dans le cadre de ce cours – voir cette définition comme caractérisant correctement une situation limite, la variable aléatoire X pouvant en réalité être considérée comme une variable aléatoire discrète, pour laquelle l'équation (2.2) ci-dessus caractérise à une bonne approximation près

la loi de X, à une échelle pouvant être très petite (par rapport à 1), mais demeurant grande devant l'échelle microscopique des valeurs de X.

Dans cette interprétation, le paradoxe apparent décrit ci-dessus, provient simplement du fait qu'il existe une échelle en-deçà de laquelle l'équation (2.2) cesse d'être valable.

Cependant, la plupart du temps, on utilise directement l'équation (2.2) ci-dessus et ses conséquences, comme si celle-ci était valable sans restriction, c'est-à-dire sans repréciser systématiquement que l'on ne manipule en réalité que des approximations de la validité desquelles il faudrait s'assurer systématiquement. Il est possible de le faire de manière cohérente, comme nous l'expliquons dans ce qui suit. Simplement, les sommes qui interviennent dans les manipulations usuelles concernant les variables aléatoires discrètes doivent être remplacées par des intégrales (qui en sont en réalité des approximations).

Soulignons que, pour prix de ces (légères) complications, nous gagnons la possibilité d'utiliser un puissant outil de modélisation et de calcul. La notion de variable aléatoire continue permet de traiter de manière unifiée un grand nombre de problèmes, discrets à une échelle microscopique, mais pouvant être considérés comme continus à l'échelle envisagée, et, surtout, nous autorise à utiliser le puissant arsenal de techniques provenant du calcul différentiel et intégral.

Propriétés

Étudions maintenant de plus près la manière de manipuler les variables aléatoires continues.

On note que la relation (2.2) ne caractérise pas complètement la densité f, car, par exemple, si l'on modifie la valeur de f en un nombre fini de points, cela ne modifie pas la valeur des intégrales de la forme $\int_a^b f(u)du$.

Par ailleurs, comme nous l'avons vu plus haut, la probabilité pour qu'une variable aléatoire continue prenne une valeur fixée est toujours nulle, et, par conséquent, on a $\mathbb{P}(X \in [a,b]) = \mathbb{P}(X \in]a,b[) = \mathbb{P}(X \in [a,b[) = \mathbb{P}(X \in]a,b])$, ainsi que $\mathbb{P}(X \in]-\infty,b[) = \mathbb{P}(X \in]-\infty,b])$ et $\mathbb{P}(X \in]a,+\infty[) = \mathbb{P}(X \in [a,+\infty[$, pour tous $a,b \in \mathbb{R}$,

De plus, le fait que $\mathbb{P}(X \in \mathbb{R}) = 1$ entraîne la relation

$$\int_{-\infty}^{+\infty} f(u)du = 1.$$

De manière générale, nous appellerons **densité de probabilité** (sans référence à une variable aléatoire particulière) toute fonction à valeurs positives continue par morceaux et qui vérifie $\int_{-\infty}^{+\infty} f(u)du = 1$, de même qu'une loi de probabilité (sans référence à une variable aléatoire particulière) sur un ensemble S fini ou dénombrable est la donnée d'une fonction p sur S vérifiant $\sum_{s\in S} p(s) = 1$.

La probabilité d'observer la valeur d'une telle variable aléatoire dans un petit intervalle $[a, a+\delta[$ est égale à $\int_a^{a+\delta} f(u)du$, et, par conséquent, si f est continue en a,

$$\mathbb{P}(X \in [a, a+\delta[) = \delta f(a)(1+o(1)).$$

En d'autre termes, cette probabilité est (au premier ordre en δ) proportionnelle à δ, et le coefficient de proportionnalité est $f(a)$, d'où le nom de densité de probabilité pour f (penser à la définition de la densité locale d'un fluide comme le coefficient de proportionnalité entre la masse et le volume d'un petit élément de fluide). On note parfois symboliquement cette relation par : $\mathbb{P}(X \in [a, a+da[) = f(a)da$. Plutôt que de manipuler les probabilités pour que X prenne telle ou telle valeur, nous aurons donc à considérer les **probabilités pour que X se trouve dans un intervalle** (ou une réunion d'intervalles).

La fonction de répartition de X est définie, comme dans le cas discret, par la relation $F_X(x) = \mathbb{P}(X \leq x)$. Par définition, on a donc

$$F_X(x) = \int_{-\infty}^{x} f(u)du.$$

Comme dans le cas discret, la fonction F_X est croissante, mais cette fois F_X est une fonction continue. Si f est continue au point x, F_X est dérivable en x et l'on a $F'_X(x) = f(x)$.

On vérifie que, dans le cas continu comme dans le cas discret, $\lim_{x\to-\infty} F_X(x) = 0$ et $\lim_{x\to-\infty} F_X(x) = 1$.

Comparons plus précisément ces résultats à ceux qui prévalent pour les variables aléatoires discrètes. Dans ce cas, pour a et b fixés, nous pourrions faire la liste des valeurs possibles de la variable comprises entre a et b, et faire la somme :

$$\mathbb{P}(a < X < b) = \sum_{a<x<b, x\in S} \mathbb{P}(X = x).$$

Pour les variables continues, la somme Σ portant sur les éléments de S compris entre a et b est remplacée par l'intégrale $\int_a^b$, et la probabilité $\mathbb{P}(X = x)$ par la probabilité «infinitésimale» $f(x)dx$. C'est systématiquement ainsi que nous passerons des identités portant sur les variables aléatoires discrètes à leurs analogues continus. On retient donc le tableau suivant :

$$\begin{cases} \sum_{a<x<b} \leftrightarrow \int_a^b \\ \mathbb{P}(X = x) \leftrightarrow f(x)dx \end{cases}$$

Rappelons que l'on peut toujours retrouver ces relations en considérant une loi continue comme limite de lois discrètes.

Loi continue et loi empirique

Dans l'interprétation fréquentielle de la probabilité, la loi empirique associée à un grand nombre de répétitions de l'expérience donnant lieu à la variable aléatoire X fournit une approximation de la loi théorique de X, la fréquence avec laquelle une valeur x_k apparaît dans un échantillon fournissant une approximation de la probabilité $\mathbb{P}(X = x_k)$. Dans le cas où la loi théorique en question est une loi continue, on ne peut s'attendre à ce que l'approximation ait lieu exactement en ce sens.

L'analogue de cette propriété dans le cas continu est que, pour tous les intervalles de la forme $]a, b[$ la probabilité empirique d'observer une valeur dans $]a, b[$ fournit une approximation de la probabilité théorique $\mathbb{P}(X \in]a, b[)$. Deux manières, parmi d'autres, de visualiser graphiquement une telle approximation, sont :

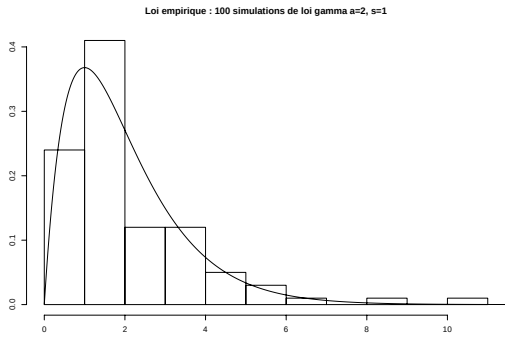
- une largeur $\Delta > 0$ étant fixée, comparer l'histogramme de la loi empirique à celui de la loi théorique (ou au graphe de la densité de la loi théorique) ;
- comparer le graphe de la fonction de répartition de la loi empirique à celui de la fonction de répartition de la loi théorique.

Même dans le cas où la loi continue considérée n'est qu'une approximation d'une loi discrète à l'échelle microscopique, les échantillons que l'on peut s'attendre à manipuler en pratique ont la plupart du temps une taille bien trop faible pour faire apparaître ce caractère discret, et il n'est pas raisonnable de s'attendre à ce que la loi empirique associée à un tel échantillon fournisse une approximation de la loi théorique dans le sens qui a cours pour les lois discrètes.

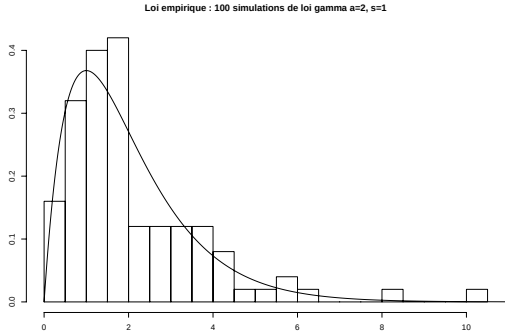
Nous reviendrons sur cette question dans le chapitre sur la loi des grands nombres. Présentons rapidement quelques exemples.

Les graphiques qui suivent représentent les histogrammes associés à des échantillon simulés de la loi gamma de paramètres $a = 2$ et $s = 1$, sur lesquels on a superposé la densité de ladite loi.

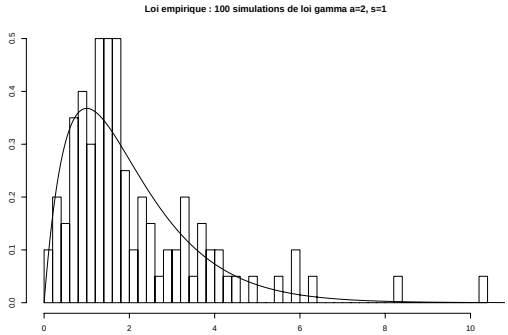
Avec 100 valeurs et $\Delta = 1$.



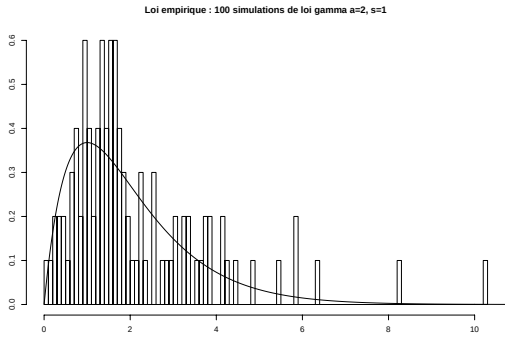
Avec les mêmes 100 valeurs et $\Delta = 0,5$.



Avec les mêmes 100 valeurs et $\Delta = 0,2$.



Avec les mêmes 100 valeurs et $\Delta = 0,1$.

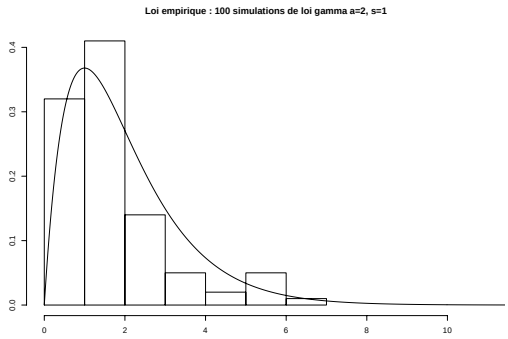


Quelques remarques. Lorsque la largeur des barres est suffisament petite pour que la densité soit à peu près constante sur l'intervalle correspondant, on s'attend à ce que la hauteur de la barre soit voisine de la densité. Lorsque la densité fluctue sur l'intervalle $[a, a+\Delta[$ formant la base d'une barre, c'est la **valeur moyenne** de la densité sur l'intervalle $[a, a+\Delta[$, soit $\frac{1}{\Delta}\int_a^{a+\Delta} f(u)du$ qui doit être voisine de la hauteur de la barre. Par conséquent, il est normal que la densité ne «colle» pas au

plus près de l'histogramme lorsque les barres de celui-ci ne sont pas suffisament fines.

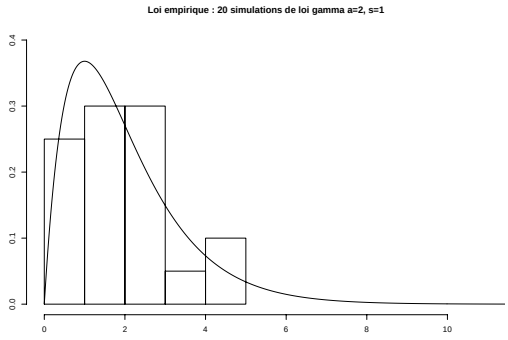
D'autre part, (pour enfoncer le clou), insistons sur le fait que la loi empirique varie d'un échantillon à l'autre, comme l'illustre le graphique suivant.

Avec un autre échantillon de 100 valeurs et $\Delta = 1$.



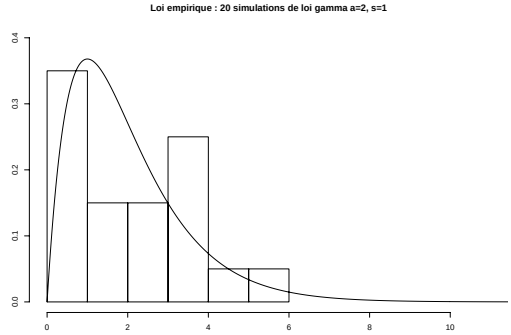
Il est important de comprendre que, plus la taille de l'échantillon est importante, plus l'on peut s'attendre à ce que l'adéquation entre histogramme et densité soit précise et valable jusqu'à de petites échelles (tout ceci pouvant être quantifié de manière précise, comme nous le verrons par la suite).

Avec cette fois un échantillon de 20 valeurs, et $\Delta = 1$.

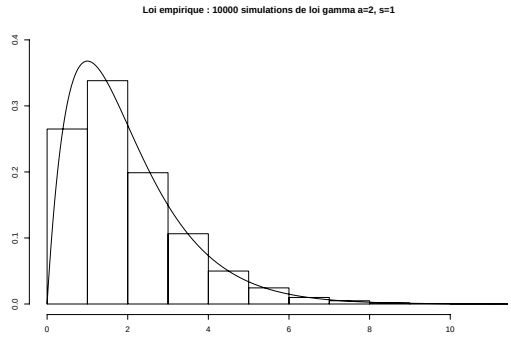


L'adéquation n'est pas excellente, mais c'est normal : il n'y a tout simplement pas assez de valeurs dans l'échantillon pour que l'on puisse s'attendre à mieux.

Avec un autre échantillon de 20 valeurs, et $\Delta = 1$.

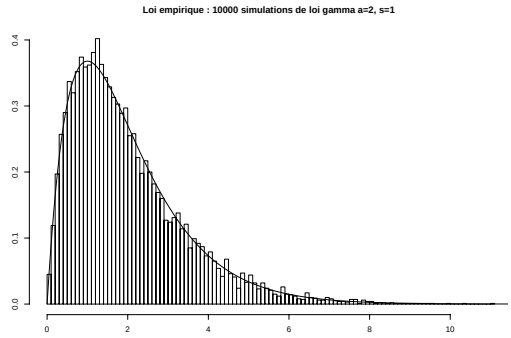


Avec un échantillon de 10000 valeurs et $\Delta = 1$.

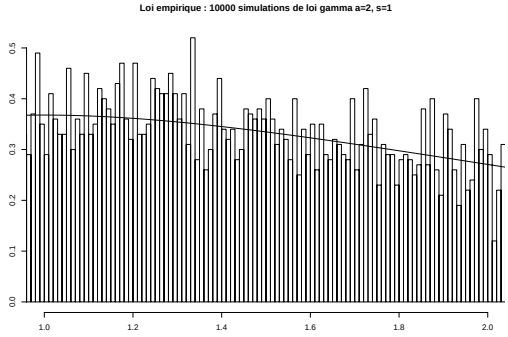


Cette fois, l'adéquation est bien meilleure (c'est la valeur moyenne de la densité sur un intervalle qui doit être comparée à la largeur d'une barre).

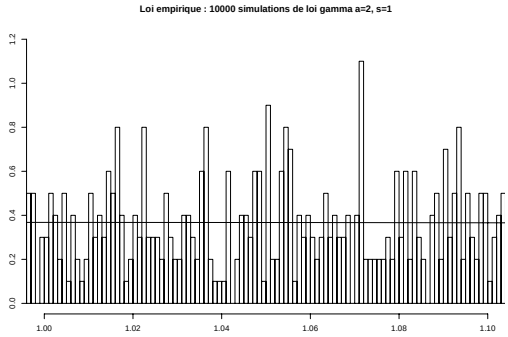
Avec le même échantillon de 10000 valeurs et $\Delta = 0,1$.



Avec le même échantillon de 10000 valeurs et $\Delta = 0,01$ (en restreignant l'intervalle des valeurs présentées de manière à pouvoir voir quelque chose).

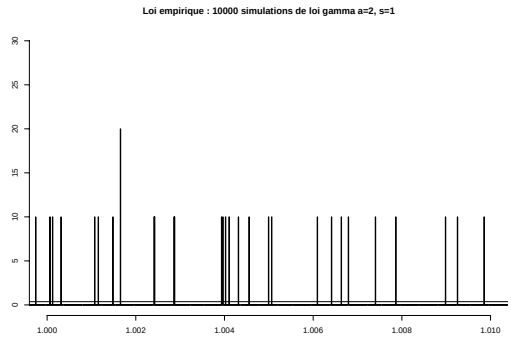


Avec le même échantillon de 10000 valeurs et $\Delta = 0,001$ (en restreignant encore plus l'intervalle des valeurs présentées).



L'adéquation entre densité et histogramme est valable jusqu'à une échelle nettement plus fine que lorsque l'on ne disposait que de 100 ou 20 valeurs. Cependant, en affinant suffisament l'échelle, on parvient toujours à mettre en évidence une inadéquation criante. Sur les deux histogrammes précédents, on constate que la densité fournit encore une approximation satisfaisante lorsque l'on regroupe plusieurs intervalles consécutifs.

Avec le même échantillon de 10000 valeurs, en diminuant encore l'échelle, on obtient l'histogramme suivant, pour lequel les choses se gâtent vraiment.



Une discussion quantitative sur la manière de juger quand un écart entre histogramme et densité est ou non raisonnable, et comment les valeurs de Δ peuvent être choisies par rapport à la taille de l'échantillon, est manifestement nécessaire. Elle sera présentée dans le chapitre «Statistique.»

Transformations affines d'une variable aléatoire de loi continue

De nombreuses familles paramétriques (c'est-à-dire, indexées par des paramètres) de lois continues (mais pas toutes, cependant) vérifient le fait que, si X suit une loi appartenant à cette famille, c'est aussi le cas de la variable aléatoire $aX + b$, tout au moins pour certaines valeurs de a et de b. Etudions la façon dont la densité se transforme sous l'effet d'une telle opération. Le résultat est le suivant : **si f désigne la densité de X, alors, pour tout $a \neq 0$ et tout b, $aX + b$ possède la densité**

$$x \mapsto \tfrac{1}{|a|} f\left(\tfrac{x}{a} - b\right).$$

Cette formule est une simple conséquence de l'égalité, due à la formule de changement de variables pour les intégrales : $\int_x^y \frac{1}{|a|} f\left(\frac{u}{a} - b\right) du = \int_{\frac{x}{a} - b}^{\frac{y}{a} - b} f(u)du$.

On peut encore la vérifier en étudiant la façon dont se transforme un histogramme sous l'effet d'une telle transformation : la probabilité pour que $aX + b$ soit compris entre x et $x + \Delta$ n'est autre que la probabilité pour que X soit compris entre $x' = \frac{x}{a} - b$ et $x' + \frac{\Delta}{a}$. La surface de la barre B correspondant à l'intervalle $[x, x + \Delta[$ dans l'histogramme de la loi de $aX + b$ est donc la même que celle de la barre B' correspondant à l'intervalle $[x', x' + \frac{\Delta}{a}]$ dans l'histogramme de la loi de X. La hauteur de B doit donc être égale $\frac{1}{|a|}$ fois la hauteur de B', puisque la largeur de B est $|a|$ fois celle de B. D'où la formule, en se rappelant le lien entre densité et histogramme associé à un découpage en barres de bases très fines.

Le cas d'une transformation plus générale qu'une transformation affine est discuté dans la partie 2.5.

2.2.7 Exemples de lois continues

Loi uniforme sur un intervalle

La loi uniforme sur un intervalle $[a, b]$ est la plus simple des lois continues. Conformément à la définition donnée dans le cas discret – à savoir, la loi qui attribue à chaque élément de S la même probabilité, – il s'agit de la loi qui attribue à chaque élément de $[a, b]$ la même **densité** de probabilité. La probabilité attribuée par cette loi aux valeurs extérieures à l'intervalle $[a, b]$ devant être nulle, la densité est donc nulle hors de $[a, b]$. La densité de la loi uniforme sur $[a, b]$ doit donc valoir : une constante c sur $[a, b]$, et zéro hors de $[a, b]$. Pour que la condition $\int_{\mathbb{R}} f(x)dx = 1$ soit vérifiée, on constate que la seule valeur possible pour c est $1/(b-a)$, d'où finalement :

$$\begin{cases} f(x) = \frac{1}{b-a} \text{ si } x \in [a, b], \\ f(x) = 0 \text{ si } x \notin [a, b]. \end{cases}$$

Loi exponentielle

Il s'agit en quelque sorte d'une version en temps continu de la loi géométrique, et qui apparaît dans le même contexte de modélisation : le premier instant de survenue d'un événement. Le lien résulte du même passage à la limite que celui qui fournit l'approximation de la loi binomiale par la loi de Poisson.

Divisons chaque intervalle de temps de 1 seconde en n intervalles de taille égale, et intéressons-nous au premier instant de survenue d'un succès, mesuré en secondes, lors de la répétition d'expériences indépendantes associées chacune à un petit intervalle de temps de durée $1/n$, et de même probabilité de succès p. La loi de cet instant aléatoire **mesuré en nombre d'expériences**, est une loi géométrique de paramètre p. Par conséquent, la probabilité qu'il faille attendre moins de t **secondes** pour voir survenir l'événement est (comme il y a n expériences par seconde) :

$$\mathbb{P}(Y < t) = 1 - \mathbb{P}(Y \geq t) = 1 - \sum_{k=\lfloor nt \rfloor}^{+\infty} p \times (1-p)^{k-1} = 1 - (1-p)^{\lfloor nt \rfloor - 1}.$$

En supposant que $p \sim \lambda/n$, lorsque n tend vers l'infini, où λ est une constante, on constate que

$$\mathbb{P}(Y < t) = 1 - (1 - \lambda/n)^{\lfloor nt \rfloor - 1} \to_{n \to +\infty} \exp(-\lambda \times t).$$

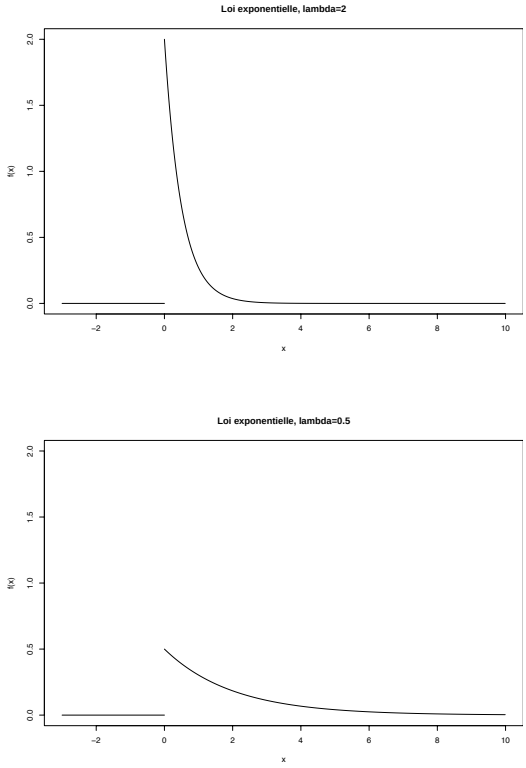
En appelant Y le premier instant (mesuré en secondes) au bout duquel le premier succès survient, on constate donc que, à la limite, on doit poser :

$$\mathbb{P}(Y < t) = \exp(-\lambda \times t),$$

d'où l'on déduit, par dérivation, la densité $f_{exp(\lambda)}(t) = \lambda e^{-\lambda t}$. Pour $t < 0$, cette densité est bien entendu nulle (l'instant que l'on étudie prend toujours une valeur positive), et l'on a donc :

$$\begin{cases} f_{exp(\lambda)}(t) = \lambda e^{-\lambda t} \text{ si } t \geq 0, \\ f(t) = 0 \text{ si } t < 0. \end{cases}$$

Voici le graphe de la densité de la loi exponentielle pour deux valeurs de λ.



Le paramètre λ^{-1} joue pour la loi exponentielle le rôle de **paramètre d'échelle**. Plus précisément, si X suit la loi exponentielle de paramètre 1, alors $\lambda^{-1}X$ suit la loi exponentielle de paramètre λ. Ceci se vérifie facilement à partir de la formule sur les transformations affines donnée précédemment.

Voir l'exercice 128.

La loi gaussienne (ou loi normale)

Cette loi tire son nom de celui du mathématicien Gauss [5], et on lui attache souvent également le nom de Laplace [6].

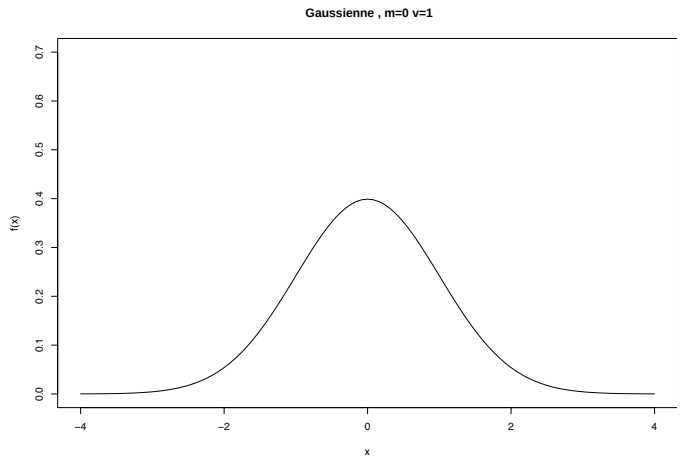
5. Carl-Friedrich Gauss (1777–1855)
6. Pierre-Simon Laplace (1749–1827).

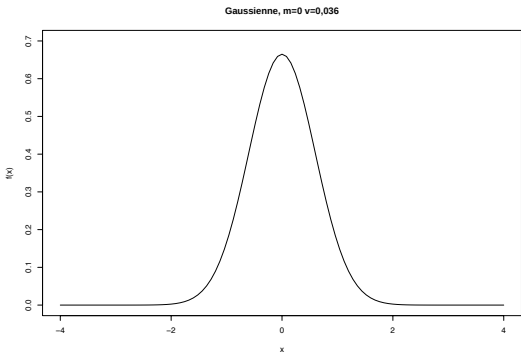
Il s'agit d'une loi continue intervenant dans un très grand nombre de situations, et dont l'étude fait à elle seule l'objet du chapitre «courbe en cloche», dans lequel seront entre autres présentées des explications à son apparition fréquente. Contentons-nous d'en rappeler la définition : sa densité est donnée par la fonction définie sur $\mathbb{R}$

$$\phi_{m,v}(x) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(x-m)^2}{2v}\right).$$

Le paramètre m joue le rôle d'un paramètre de position, et v celui d'un paramètre d'échelle. En effet, si X suit la loi gaussienne de paramètres $m = 0$ et $v = 1$, $\alpha X + \beta$ suit la loi gaussienne de paramètres $m = \beta$ et $v = \alpha^2$.

Voici quelques exemples de graphes de la densité gaussienne.





Voir l'exercice 130.

Loi de Cauchy

Cette loi tire son nom de celui du mathématicien Cauchy [7].

7. Augustin-Louis Cauchy (1789–1857)

Il s'agit de la loi sur $\mathbb{R}_+$ dont la densité est définie pour $\ell \in \mathbb{R}$ et $s > 0$ par

$$f_{Cauchy(\ell,s)}(x) = \frac{1}{\pi s(1 + \left[\frac{x-\ell}{s}\right]^2)}.$$

Le paramètre ℓ est un paramètre de position, et le paramètre $s > 0$ un paramètre d'échelle : si X suit la loi de Cauchy de paramètres $\ell = 0$ et $s = 1$, $\alpha X + \beta$ suit la loi de Cauchy de paramètres $\ell = \beta$ et $s = \alpha$.

Voir l'exercice 131.

Loi gamma

Il s'agit de la loi sur $\mathbb{R}_+$ dont la densité est définie de la manière suivante

$$f_{gamma(a,s)}(x) = \frac{1}{s^a \Gamma(a)} x^{a-1} \exp(-x/s)$$

pour $x \geq 0$, et $f_{gamma(a,s)}(x) = 0$ si $x < 0$. Les deux paramètres $a > 0$ et $s > 0$ sont respectivement appelés paramètre de forme et d'échelle.

Si X suit la loi gamma de paramètres a et 1, αX suit la loi gamma de paramètres a et α.

Rappelons que la fonction gamma d'Euler est définie pour $a > 0$ par

$$\Gamma(a) = \int_0^{+\infty} x^{a-1} \exp(-x) dx.$$

Voir l'exercice 128 pour un exemple de contexte dans lesquels cette loi intervient.

Loi beta

Il s'agit de la loi sur $[0, 1]$ dont la densité est définie de la manière suivante

$$f_{beta(a,b)}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1},$$

pour $x \in [0, 1]$, et $f_{beta(a,b)}(x) = 0$ si $x \notin [0, 1]$, les deux paramètres réels a et b devant être tels que $a > 0$ et $b > 0$. Voir l'exercice 134 pour un exemple de contexte dans lequel cette loi intervient.

Loi du chi-deux

Pour n entier ≥ 1, la loi du chi-deux à n degrés de liberté est la loi sur $\mathbb{R}_+$ définie par la densité

$$f_{\chi^2(n)}(x) = 2^{-n/2} \Gamma(n/2) x^{n/2-1} \exp(-x/2)$$

pour $x \geq 0$, et $f_{\chi^2(n)}(x) = 0$ pour $x < 0$.

Il s'agit donc d'un cas particulier de la loi gamma avec $s = 2$ et $a = n/2$. Elle intervient de manière fondamentale dans le test statistique du chi-deux, que nous étudierons dans la partie «Statistique». Voir l'exercice 129.

Remarque 6 *Vous noterez que, comme annoncé, les hypothèses qui nous permettent d'affirmer automatiquement que l'on est en présence de telle ou telle loi classique sont très générales, et ne font qu'assez peu intervenir la forme détaillée des modèles. Nous avons seulement besoin de propriétés assez générales d'indépendance sous-jacentes au modèle sur lequels les variables aléatoires étudiées sont définies. La facilité que procure l'utilisation de ces lois classiques ne doit pas pour autant nous faire abandonner la méthode systématique qui convient pour aborder la modélisation d'une situation : il est indispensable de préciser les données, connaissances, et hypothèses de modélisation sur lesquelles on s'appuie. C'est seulement une fois cette étape accomplie que l'on peut se demander si telle ou telle variable aléatoire suit une loi classique, en vérifiant que le modèle possède bien les propriétés correspondantes. Il faut donc voir les résultats ci-dessus sur l'identification des lois classiques comme des moyens de gagner du temps en évitant de refaire des raisonnements ou des calculs qui ont déjà été menés (il est inutile de refaire à chaque fois le raisonnement qui mène à la loi binomiale, il suffit de retenir dans quelles conditions générales de modélisation celle-ci apparaît). En revanche, commencer l'étude d'une situation en tentant de plaquer dessus telle ou telle loi classique, sans s'interroger sur la forme du modèle, les données et les connaissances disponibles, et les hypothèses de modélisation qu'il est pertinent de formuler, n'est pas une démarche acceptable, et conduit le plus souvent à des résultats erronés. Il est important de noter que des lois distinctes des lois classiques apparaissent dans de nombreuses situations.*

2.3 Loi jointe de plusieurs variables aléatoires, vecteurs aléatoires

Lorsque l'on dispose de plusieurs variables aléatoires $X_1, \ldots, X_m$ définies sur un même Ω et à valeurs respectivement dans des ensembles $S_{X_1}, \ldots, S_{X_m}$, on appelle **loi jointe de** $X_1, \ldots, X_m$ la loi de la variable aléatoire $X(\omega) = (X_1(\omega), \ldots, X_m(\omega))$, définie sur Ω et à valeurs dans l'ensemble produit $S_{X_1} \times \cdots \times S_{X_m}$.

Mise en garde 9 *Les S_{X_i} peuvent être totalement différents les uns des autres, mais, pour que la notion de loi jointe ait un sens, il est nécessaire que toutes les variables aléatoires considérées soient définies sur le même espace de probabilité $(\Omega, \mathbb{P})$.*

Souvent, on spécifie implicitement un modèle probabiliste d'une situation en spécifiant simplement la loi jointe d'un certain nombre de variables aléatoires qui interviennent dans celui-ci.

Mise en garde 10 *Il est important de bien noter que la connaissance de la loi individuelle de X_i pour tout $1 \leq i \leq n$ ne suffit pas en général à déterminer la loi jointe de ces variables. Prenons par exemple le modèle $(\Omega^N, \mathbb{P}^{\otimes N})$ décrivant une répétition indépendante de N lancers de pile ou face, la variable aléatoire X_i représentant le résultat du $i-$ème lancer. Dans le $N-$uplet $(X_1, \ldots, X_N)$, chaque variable possède individuellement une loi de Bernoulli de paramètre p. C'est aussi le cas si l'on constituant le $N-$uplet $(X_1, X_1, \ldots, X_1)$. Pourtant, il est bien évident que ces deux $N-$uplets n'ont pas la même loi !*

2.3.1 Indépendance de variables aléatoires, cas discret

Etant donnés un modèle probabiliste $(\Omega, \mathbb{P})$, et m variables aléatoires $X_1, \ldots, X_m$ définies sur Ω et dont les ensembles de valeurs possibles sont $S_1, \ldots, S_m$, considérons le modèle image $\left(S_{(X_1,\ldots,X_m)}, p_{(X_1,\ldots,X_m)}\right)$ compatible avec $(\Omega, \mathbb{P})$ et décrivant les valeurs prises par le $N-$uplet $(X_1, \ldots, X_m)$.

On dira que les variables aléatoires $X_1, X_2, \ldots, X_m$ sont globalement, ou encore mutuellement, indépendantes si

$$p_{(X_1,\ldots,X_m)} = p_{X_1} \otimes \cdots \otimes p_{X_m},$$

ou, autrement dit, si le modèle image $\left(S_{(X_1,\ldots,X_m)}, p_{(X_1,\ldots,X_m)}\right)$ associé au $m-$uplet $(X_1, \ldots, X_m)$ s'identifie à la succession indépendante des modèles $(S_{X_1}, p_{X_1}), \ldots, (S_{X_m}, p_{X_m})$ associés individuellement aux variables aléatoires $X_1, \ldots, X_m$.

En termes plus élémentaires, $X_1, \ldots, X_m$ sont mutuellement indépendantes lorsque, pour tout $s_1 \in S_1, \ldots, s_m \in S_m$, on a

$$\mathbb{P}(X_1 = s_1, \ldots, X_m = s_m) = \mathbb{P}(X_1 = s_1) \times \cdots \times \mathbb{P}(X_m = s_m)$$

Dans le cas où les variables aléatoires $X_1, \ldots, X_m$ sont indépendantes, la donnée de leurs lois individuelles permet donc de reconstituer la loi jointe de $(X_1, \ldots, X_m)$. D'une manière générale, il est nécessaire de connaître la structure de dépendance des X_i pour pouvoir reconstituer la loi jointe, comme on le voit simplement dans le cas $m = 2$ en écrivant

$$\mathbb{P}((X_1, X_2) = (s_1, s_2)) = \mathbb{P}(X_1 = s_1) \times \mathbb{P}(X_2 = s_2 | X_1 = s_1).$$

2.3.2 Vecteur aléatoire continu

La généralisation à $\mathbb{R}^m$ de la notion de variable aléatoire de loi continue à valeurs dans $\mathbb{R}$, est connue sous le nom de vecteur aléatoire de loi continue. Exactement comme dans le cas d'une variable aléatoire réelle, on peut voir cette situation comme un cas limite de variables aléatoires discrètes à valeurs dans $\mathbb{R}^m$. La généralisation de la relation (2.2) est la suivante : pour tout sous-ensemble A suffisament régulier (par exemple un pavé de la forme $[a_1, b_1] \times \cdots \times [a_m, b_m]$, ou une réunion finie de pavés de ce type) de $\mathbb{R}^m$, on a

$$\mathbb{P}((X_1, \ldots, X_m) \in A) = \int_A f(x_1, \ldots, x_m) dx_1 \cdots dx_m,$$

où $f : \mathbb{R}^m \to \mathbb{R}_+$ est une fonction positive que nous supposerons suffisament régulière (par exemple continue), appelée la densité du vecteur aléatoire $(X_1, \ldots, X_m)$. Nous resterons quelque peu évasif sur la notion de régularité (pour f comme pour A) dont il est question ici, le bon cadre pour développer cette théorie étant celui de la théorie mathématique de la mesure.

En général, le $m-$uplet formé par m variables aléatoires de loi continue ne forme pas un vecteur aléatoire de loi continue dans le sens défini précédemment. Par exemple, on peut facilement se convaincre que, si X_1 est une variable aléatoire réelle de loi continue, ce n'est pas le cas de (X_1, X_1) (mais ce type de difficulté disparaît dans le formalisme de la théorie de la mesure).

On dira que m variables aléatoires continues $X_1, \ldots, X_m$ sont mutuellement (ou encore globalement) indépendantes lorsque $(X_1, \ldots, X_m)$ possède une loi continue dont la densité $f_{(X_1,\ldots,X_m)}$ vérifie

$$f_{(X_1,\ldots,X_m)}(x_1, \ldots, x_m) = f_{X_1}(x_1) \times \cdots \times f_{X_m}(x_m).$$

On vérifie facilement que cette notion est le passage à la limite naturel de la définition donnée dans le cas discret.

2.3.3 Somme de variables aléatoires indépendantes

Cas discret

Partant d'une famille de variables aléatoires, par exemple, deux variables X_1 et X_2 définies sur le même espace des possibles Ω, et à valeurs dans des ensembles S_1 et S_2 respectivement, on peut en fabriquer une troisième, également définie sur Ω, et à valeurs dans $S_1 \times S_2$, définie par :

$$Y(\omega) = (X_1(\omega), X_2(\omega)),$$

dont les valeurs décrivent les valeurs **simultanées** de X_1 et de X_2. Connaissant la loi de X_1 et de X_2, est-il possible d'en déduire la loi de Y ?

Sans plus d'information, la réponse est : NON, car nous ne pouvons pas déterminer la loi jointe de X_1 et X_2. Illustrons ceci à l'aide d'exemples.

Considérons un espace de probabilité $(\Omega, \mathbb{P})$, sur lequel sont définies deux variables aléatoires X_1 et X_2, chacune suivant la loi uniforme sur $\{1; 2; \ldots; 10\}$.

Par exemple, partons d'une variable aléatoire X de loi uniforme sur $\{1; 2; \ldots; 10\}$, et définissons $X_1 = X$, $X_2 = 11 - X$. X_1 suit donc la loi uniforme sur $\{1; 2; \ldots; 10\}$, et X_2 également. On constate, par exemple, que la probabilité $\mathbb{P}((X_1, X_2) = (1, 1))$ est égale à zéro. À présent, choisissons X_1 et X_2 égales à X toutes les deux. X_1 et X_2 suivent encore chacune la loi uniforme sur $\{1; 2; \ldots; 10\}$, mais cette fois, la probabilité $\mathbb{P}((X_1, X_2) = (1, 1))$ est égale à $1/10$ (c'est la probabilité pour que $X = 1$). On constate donc que la connaissance de la loi de X_1 et de la loi de X_2 prises séparément ne permet pas d'en déduire la loi du couple $Y = (X_1, X_2)$. En revanche, si l'on suppose que X_1 et X_2 sont **indépendantes**, on a nécessairement :

$$\mathbb{P}((X_1, X_2) = (s_1, s_2)) = \mathbb{P}(X_1 = s_1) \times \mathbb{P}(X_2 = s_2),$$

et la loi du couple peut donc être déduite des lois individuelles de X_1 et X_2.

Reprenons les exemples précédents pour aborder la question, importante en pratique, de la loi d'une somme de deux variables aléatoires. Dans le premier exemple, la somme de X_1 et X_2 est égale à $X_1 + X_2 = X + 11 - X = 11$, et prend donc la valeur constante 11. Dans le deuxième exemple, X_1 et X_2 sont égales à X toutes les deux et leur somme est égale à $2X$, dont la loi est la loi uniforme sur les entiers pairs compris entre 2 et 20. Partant de deux variables aléatoires possédant chacune la loi uniforme sur $\{1; 2; \ldots; 10\}$, nous obtenons donc, en en prenant la somme, deux variables aléatoires de lois complètement différentes, et la seule connaissance des lois respectives de X_1 et de X_2 ne suffit donc pas pour déterminer la loi de $X_1 + X_2$. Examinons la situation de plus près. Comment détermine-t-on la probabilité pour que $X_1 + X_2 = 12$ (par exemple) ? Il nous faut d'abord déterminer toutes les éventualités élémentaires ω telles que $X_1(\omega) + X_2(\omega) = 12$, et calculer la somme des probabilités de toutes ces éventualités élémentaires. Une autre manière de procéder, puique nous nous intéressons seulement aux valeurs prises par X_1 et X_2, consiste à «découper» Ω suivant les valeurs prises par la variable aléatoire $Z = (X_1, X_2)$ (nous avons déjà vu qu'une variable aléatoire fournissait un découpage de l'espace des possibles en considérant les événements associés à chaque valeur que peut prendre la variable aléatoire). Ce découpage de Ω est *a priori* moins fin, plus grossier que le découpage de Ω par les éventualités élémentaires, puisqu'il est obtenu en regroupant toutes les éventualités élémentaires qui donnent la même valeur au couple (X_1, X_2) (il peut y en avoir plusieurs, car la description de l'expérience que fournit Ω ne se résume pas forcément aux valeurs prises par X_1 et X_2, mais peut, par exemple, décrire également les valeurs prises par d'autres variables aléatoires $X_3, X_4, \ldots$ auxquelles nous ne nous intéressons pas ici).

Considérons donc le découpage de Ω formé par les événements formé par les 100 événements :

$$A_{(a,b)} = \{X_1 = a \ , \ X_2 = b\} \ , \ 1 \leq a, b \leq 10.$$

L'événement $X_1 + X_2 = 12$ est formé par la réunion des 11 événements deux-à-deux disjoints :

$$A_{(1,11)}, A_{(2,10)}, A_{(3,9)}, A_{(4,8)}, A_{(5,7)}, A_{(6,6)}, A_{(7,5)}, A_{(8,4)}, A_{(9,2)}, A_{(10,2)}, A_{(11,1)},$$

et, par conséquent, la probabilité que nous cherchons est égale à :

$$\mathbb{P}(X_1 + X_2 = 12) = \sum_{i=1}^{11} \mathbb{P}(A_{(i,12-i)}) = \sum_{i=1}^{11} \mathbb{P}(X_1 = i \ , \ X_2 = 12 - i).$$

Le problème, si nous ne connaissons que les lois de X_1 et X_2 prises séparément, est que nous ne sommes pas en mesure de déterminer les probabilités du type : $\mathbb{P}(X_1 = i \ , \ X_2 = 12 - i)$, qui font intervenir la réalisation simultanée des deux variables. Bien entendu, si l'on suppose que X_1 et X_2 sont **indépendantes** (ce qui n'est le cas ni lorsque $X_1 = 11 - X_2$, ni lorsque $X_1 = X_2$), ces probabilités s'expriment simplement en termes des lois respectives de X_1 et X_2 :

$$\mathbb{P}(X_1 = i \ , \ X_2 = 12 - i) = \mathbb{P}(X_1 = i) \times \mathbb{P}(X_2 = 12 - i).$$

Nous retiendrons que, si X_1 et X_2 sont deux variables aléatoires indépendantes à valeurs réelles, dont les ensembles de valeurs possibles sont respectivement notés S_{X_1} et S_{X_2}, on peut calculer la loi de $X_1 + X_2$ à l'aide de la formule :

$$\begin{aligned} \mathbb{P}(X_1 + X_2 = z) &= \sum_{x \in S_{X_1}} \mathbb{P}(X_1 = x) \times \mathbb{P}(X_2 = z - x) \\ &= \sum_{y \in S_{X_2}} \mathbb{P}(X_2 = y) \times \mathbb{P}(X_1 = z - y). \end{aligned}$$

Remarquons que bien entendu, cette formule ne fait pas intervenir explicitement l'espace des possibles Ω, mais simplement les lois des variables aléatoires définies sur Ω. L'indépendance supposée de X_1 et X_2 nous permet de déduire directement la loi jointe de X_1 et X_2 des lois individuelles.

Cas continu

La généralisation des formules précédentes au cas continu est facile : si X et Y sont deux variables aléatoires indépendantes et de loi continue, de densité f et g respectivement, $X + Y$ est encore une variable aléatoire continue, dont la densité est donnée par

$$h(z) = \int_{-\infty}^{+\infty} f(x)g(z - x)dx = \int_{-\infty}^{+\infty} f(z - y)g(y)dy.$$

2.4 Opérations sur les lois de probabilité

Nous avons vu, dans ce qui précède, plusieurs définitions pouvant être présentées comme celles d'opérations sur des probabilités (ou des lois de probabilités), même si nous n'utiliserons pas beaucoup ce point de vue abstrait.

L'une d'entre elles est **le produit tensoriel**. Etant donnés $(\Omega_1, \mathbb{P}_1)$ et $(\Omega_2, \mathbb{P}_2)$ deux modèles probabilistes, il s'agit de la probabilité $\mathbb{P}_1 \otimes \mathbb{P}_2$ définie sur $\Omega_1 \times \Omega_2$ par $\mathbb{P}_1 \otimes \mathbb{P}_2(\omega_1, \omega_1) = \mathbb{P}_1(\omega_1) \times \mathbb{P}_2(\omega_2)$, et qui permet de modéliser la succession indépendante de la situation décrite par $(\Omega_1, \mathbb{P}_1)$ par la situation décrite par $(\Omega_2, \mathbb{P}_2)$. On vérifie que ce produit est associatif (à condition de faire les identifications qui s'imposent), mais certainement pas commutatif.

Une autre est le **produit de convolution**. Si $\mathbb{P}_1$ et $\mathbb{P}_2$ sont les lois de deux variables aléatoires à valeurs réelles X et Y (Ω_1 est donc l'ensemble des valeurs possibles de X, et Ω_2 l'ensemble des valeurs possibles de Y), le produit de convolution de $\mathbb{P}_1$ par $\mathbb{P}_2$, noté $\mathbb{P}_1 \star \mathbb{P}_2$, est simplement la loi de $X + Y$ dans le modèle $(\Omega_1 \times \Omega_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$, où $X(\omega_1, \omega_2) = X(\omega_1)$ et $Y(\omega_1, \omega_2) = Y(\omega_2)$. Autrement dit, c'est la loi de $X+Y$ en supposant que X et Y sont indépendantes. D'après la partie précédente,

$$\mathbb{P}_1 \star \mathbb{P}_2(z) = \sum_{x \in \Omega_1} \mathbb{P}_1(x) \times \mathbb{P}_2(z - x) = \sum_{y \in \Omega_2} \mathbb{P}_1(z - y) \times \mathbb{P}_2(y).$$

On vérifie immédiatement que ce produit est associatif et commutatif (car l'addition dans $\mathbb{R}$ l'est !)

Une troisième opération, que nous n'avons pas encore formellement définie, est le **mélange** de probabilités.

Si $(a_1, \ldots, a_n)$ définit une probabilité sur l'ensemble $\{1, \ldots, n\}$, et si $\mathbb{P}_1, \ldots, \mathbb{P}_n$ sont des probabilités sur un ensemble Ω donné, l'application définie sur Ω par $\omega \mapsto a_1 P_1(\omega) + \cdots + a_n P_n(\omega)$ est appelée le mélange de $\mathbb{P}_1, \ldots, \mathbb{P}_n$ par rapport aux poids $a_1, \ldots, a_n$. Considérons le modèle $\{1, \ldots, n\} \times \Omega$, dans lequel on choisit d'abord un entier entre 1 et n selon les probabilités $a_1, \ldots, a_n$, puis conditionnellement au choix de cet entier, un élément de Ω selon la probabilité numérotée par le choix de cet entier. Si l'on s'intéresse seulement au modèle image associé à l'élément de Ω qui est choisi, la probabilité associée est $a_1 P_1(\omega) + \cdots + a_n P_n$.

Un exemple simple est celui où l'on considère une probabilité décrivant une population constituée de plusieurs sous-populations. Supposons par exemple que l'ensemble Ω décrive les différentes valeurs que peut prendre un caractère (quantitatif ou qualitatif) associé à un individu d'une certaine population, et que la population étudiée est partitionnée en deux sous-populations numérotées 1 et 2.

Si $\mathbb{P}_1$ et $\mathbb{P}_2$ décrivent la répartition de ce caractère dans chacune des deux sous-populations, la probabilité sur Ω décrivant la répartition associée à la population totale est $p_1\mathbb{P}_1 + p_2\mathbb{P}_2$, où p_1 et p_2 désignent les proportion d'individus figurant respectivement dans les sous-population numérotées 1 et 2 (on a donc $p_1 + p_2 = 1$).

2.5 Loi d'une fonction d'une variable aléatoire

On rencontre souvent le problème suivant : étant donnée une variable aléatoire $X : \Omega \to S$, et une fonction $h : S \to V$, trouver la loi de la variable aléatoire $h(X)$. Clairement, l'ensemble $S_{h(X)}$ des valeurs possibles de $h(X)$ est l'ensemble $\{h(s) : s \in S_X\}$, où S_X désigne l'ensemble des valeurs possibles pour X.

Dans le cas discret, il suffit d'écrire que, pour $v \in S_{h(X)}$,

$$\mathbb{P}(h(X) = v) = \sum_{s \in \{x \in S_X \,:\, h(x) = v\}} \mathbb{P}(X = s).$$

On notera la composition $\Omega \xrightarrow{X} S_X \xrightarrow{h} S_{h(X)}$, et l'utilisation dans la formule ci-dessus d'un découpage selon les valeurs de l'ensemble «intermédiaire» S_X.

Dans le cas continu, le calcul précédent prend une forme spécifique lorsque, comme il est courant, la fonction h possède de bonnes propriétés de régularité. Appelons f la densité de X, et supposons par exemple, que h est un C^1-difféomorphisme de $\mathbb{R}$ sur lui-même. On vérifie alors que $h(X)$ possède la densité

$$x \mapsto \frac{1}{|h'(h^{-1}(x))|} f(h^{-1}(x)).$$

Ceci se vérifie (au moins formellement) en écrivant que, lorsque h est croissante (et h est nécessairement croissante ou décroissante avec nos hypothèses), $\mathbb{P}(a \leq h(X) \leq a + da) = \mathbb{P}(h^{-1}(a) \leq X \leq h^{-1}(a + da))$. En négligeant les termes d'ordre supérieur à 1 en da, on peut alors écrire que $h^{-1}(a + da) \approx h^{-1}(a) + (h^{-1})'(a)da$, et que $\mathbb{P}(h^{-1}(a) \leq X \leq h^{-1}(a) + (h^{-1})'(a)da) = f(h^{-1}(a)) \times (h^{-1})'(a)da$, d'où le résultat. Le cas où h est décroissante se traite de la même façon. Une manière de procéder plus correcte mathématiquement est de considérer des intervalles de taille finie (et non pas infinitésimale) et d'appliquer la formule de changement de variables pour les intégrales, qui conduit au même résultat. Lorsque h n'est pas un C^1-difféomorphisme, rien ne nous empêche de tenter une approche similaire en tenant compte des propriétés spécifiques de la fonction h considérée.

Dans le cas multidimensionnel, ce qui précède se généralise sous la forme suivante. Si $(X_1, \ldots, X_m)$ est un vecteur aléatoire continu de densité f sur $\mathbb{R}^m$ et si $h = (h_1, \ldots, h_m)$ est un C^1-difféomorphisme de $\mathbb{R}^m$ sur lui-même, $h(X_1, \ldots, X_m)$ est encore un vecteur aléatoire, de densité

$$(x_1, \ldots, x_m) \mapsto \frac{1}{\left| \text{dét} \left[\left(\frac{\partial h_i}{\partial x_j} \right)_{1 \leq i,j \leq m} \right] (h^{-1}(x_1, \ldots, x_m)) \right|} f(h^{-1}(x_1, \ldots, x_m)).$$

C'est une simple conséquence de la formule de changement de variables pour les intégrales sur $\mathbb{R}^m$.

2.6 Espérance et variance

2.6.1 Définition

Intéressons-nous maintenant spécifiquement aux variables aléatoires à valeurs réelles, qui représentent donc des quantités numériques telles que : la taille, le poids d'un individu, le temps de transfert d'un paquet de données sur internet ou encore la valeur du patrimoine d'un ménage... **L'espérance d'une variable aléatoire X à valeurs réelles, définie sur un espace de probabilité $(\Omega, \mathbb{P})$ est définie par :**

$$\mathbb{E}_{\mathbb{P}}(X) = \sum_{\omega \in \Omega} X(\omega) \times \mathbb{P}(\omega).$$

L'espérance d'une variable aléatoire est donc un nombre réel, non-aléatoire (il ne dépend pas de ω) obtenu en effectuant la somme sur toutes les éventualités élémentaires, – c'est-à-dire sur toutes les issues de la situation étudiée, au niveau de description adopté –, de la valeur que prend la variable aléatoire dans le cas où c'est cette issue qui est réalisée, multipliée par la probabilité que cette éventualité se réalise. Plus une éventualité élémentaire est probable, plus le poids attribué dans cette somme à la valeur que prend X lorsque cette éventualité est réalisée est grand, et plus celle-ci «contribue» à la valeur totale de la somme. En d'autres termes, l'espérance est une moyenne pondérée des valeurs que prend la variable X sur les différentes éventualités élémentaires, la pondération étant fournie par les probabilités de ces éventualités élémentaires.

Lorsqu'il n'y a aucune ambiguité concernant la probabilité sur Ω à laquelle on se réfère, on note simplement l'espérance sous la forme $\mathbb{E}(X)$. Inversement, lorsque plusieurs probabilités sur Ω peuvent être envisagées, il convient de préciser !

Exemple 8

$$\begin{cases} \Omega = \{a, b, c\}; \\ X(a) = 2 \ , \ X(b) = 4 \ , \ X(c) = 2; \\ \mathbb{P}_1(a) = 2/8 \ , \ \mathbb{P}_1(b) = 1/8 \ , \ \mathbb{P}_1(c) = 5/8. \end{cases}$$

L'espérance de X est donnée par :

$$\mathbb{E}_{\mathbb{P}_1}(X) = X(a)\times\mathbb{P}_1(a)+X(b)\times\mathbb{P}_1(b)+X(c)\times\mathbb{P}_1(c) = 2\times2/8+4\times1/8+2\times5/8 = 2,25.$$

Si l'on modifie la probabilité en accordant un poids plus important à b, cette valeur se rapproche de $X(b) = 4$. Par exemple si

$$\mathbb{P}_2(a) = 2/8 \ , \ \mathbb{P}_2(b) = 5/8 \ , \ \mathbb{P}_2(c) = 1/8,$$

l'espérance de X est donnée par :

$$\mathbb{E}_{\mathbb{P}_2}(X) = X(a)\times\mathbb{P}_2(a)+X(b)\times\mathbb{P}_2(b)+X(c)\times\mathbb{P}_2(c) = 2\times2/8+4\times5/8+2\times1/8 = 3,75.$$

Une autre manière de définir l'espérance consiste à l'écrire, non plus comme une somme sur toutes les éventualités élémentaires, pondérées chacune par leur probabilité de réalisation, mais comme une somme sur toutes les valeurs que peut prendre X, pondérées chacune par leur probabilité d'apparition. (Ou, autrement dit, en se plaçant sur l'espace de probabilité image (S_X, p_X) associé à X et compatible avec le modèle $(\Omega, \mathbb{P})$ pour le calcul de l'espérance). En effet, si, dans la formule qui définit l'espérance de X, nous regroupons toutes les éventualités élémentaires ω qui donnent à X la valeur s, leur contribution totale dans la somme est :

$$\sum_{\omega \,:\, X(\omega)=s} X(\omega)\times\mathbb{P}(\omega) = \sum_{\omega \,:\, X(\omega)=s} s\times\mathbb{P}(\omega) = s\times\left(\sum_{\omega \,:\, X(\omega)=s} \mathbb{P}(\omega)\right) = s\times\mathbb{P}(X=s).$$

En considérant l'ensemble S_X de toutes les valeurs possibles que peut prendre la variable aléatoire X, on constate que **l'espérance de X s'exprime donc également par la formule :**

$$\mathbb{E}(X) = \sum_{s\in S_X} s \times \mathbb{P}(X=s).$$

Cette dernière formule montre en particulier que l'espérance de X **ne dépend que de la loi de** X, puisqu'elle ne fait intervenir X qu'au travers des probabilités $\mathbb{P}(X = s)$. La somme ne porte plus ici sur les éléments de Ω, mais directement sur les valeurs que peut prendre X, chaque valeur étant multipliée par sa probabilité d'apparition, et contribuant donc d'autant plus à la somme totale que cette probabilité est importante.

Dans le cas d'une variable aléatoire continue, cette définition de l'espérance (qui est la seule que nous puissions donner, la première définition ne pouvant être généralisée sans appel à la théorie mathématique de la mesure) se transcrit en

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \times f(x)dx.$$

Dans l'exemple simple précédent, cette nouvelle manière d'exprimer l'espérance revient à regrouper dans le calcul les éventualités a et c, qui donnent la même valeur à X :

$$\mathbb{E}(X) = 2 \times \mathbb{P}(\{a, c\}) + 4 \times \mathbb{P}(b) = 2 \times (\mathbb{P}(a) + \mathbb{P}(c)) + 4 \times \mathbb{P}(b).$$

Remarque 7 *dans les deux définitions ci-dessus, nous ne nous sommes pas préoccupés de l'existence des sommes de la forme : $\sum_{\omega\in\Omega}$ ou $\sum_{s\in S}$. Lorsque les variables aléatoires considérées ne prennent qu'un nombre fini de valeurs, et, a fortiori, lorsque Ω ne comporte qu'un nombre fini d'éléments, cette écriture ne soulève aucune difficulté. En revanche, si Ω est infini (nous le supposerons toujours dénombrable), il faut s'assurer que les sommes que l'on manipule sont bien définies, et* **ce n'est pas**

toujours le cas. *Par exemple, considérons une variable aléatoire dont l'ensemble des valeurs est $\mathbb{N}^*$, et dont la loi est définie par :*

$$\mathbb{P}(X=n)=\frac{1}{n(n+1)}=\frac{1}{n}-\frac{1}{n+1}.$$

On définit bien ainsi une loi de probabilité car $\sum_{n=1}^{+\infty}\frac{1}{n(n+1)}=1$. Cependant, la somme qui définit l'espérance de X ne converge pas :

$$\sum_{n=1}^{+\infty} n\times\mathbb{P}(X=n)=\sum_{n=1}^{+\infty}\frac{n}{n(n+1)}=\sum_{n=1}^{+\infty}\frac{1}{(n+1)},$$

et l'espérance n'est donc pas définie. On pourrait cependant convenir que, dans ce cas, que l'espérance de X est égale à $+\infty$. L'exemple d'une variable aléatoire à valeurs dans $\mathbb{Z}^*$ et qui vérifie $\mathbb{P}(X=n)=\frac{1}{2|n|(|n|+1)}$ devrait vous convaincre que l'on ne peut vraiment pas toujours définir l'espérance.

Dans tous les cas, nous ne parlerons de l'espérance de X que lorsque la série :

$$\sum_{\omega\in\Omega}|X(\omega)|\mathbb{P}(\omega)$$

converge, ou, ce qui revient au même, lorsque la série

$$\sum_{s\in S}|s|\times\mathbb{P}(X=s)$$

converge, et nous dirons alors que l'espérance de X existe, ou est bien définie, ou encore que X possède une espérance. Dans ce cas, la série définissant l'espérance converge, et le résultat ne dépend pas de l'ordre dans lequel la sommation est effectuée. En tout cas, il faut à chaque fois s'assurer, lorsque l'on manipule une espérance, que ces convergences ont bien lieu.

Lorsque h est une fonction à valeur réelles, l'espérance de $h(X)$ (à condition qu'elle soit définie), peut se mettre dans le cas discret sous la forme

$$\mathbb{E}(h(X))=\sum_{v\in S_{h(X)}} v\times\mathbb{P}(h(X)=v)=\sum_{s\in S_X} h(s)\mathbb{P}(X=s).$$

Dans le cas continu :

$$\mathbb{E}(h(X))=\int_{-\infty}^{+\infty} h(u)f(u)du.$$

(On déduit facilement cette formule de celle qui prévaut dans le cas discret).

Mise en garde 11 *Attention à ne pas écrire d'absurdité du genre*

$$\mathbb{E}(h(X)) = \sum_{s \in S_X} h(s) h(\mathbb{P}(X = s))$$

ou encore $\mathbb{E}(h(X)) = \int_{-\infty}^{+\infty} h(u)h(f(u))du$*, malheureusement fréquentes, en particulier lorsque* $h(t) = t^2$ *et que l'on cherche à calculer la variance, qui sera définie plus bas.*

On note bien que c'est le caractère numérique, quantitatif, d'une variable aléatoire, qui permet de donner un sens à son espérance, définie comme une somme pondérée de valeurs. Que serait l'espérance d'une variable aléatoire dont la valeur serait un prénom ou une couleur ?

2.6.2 Espérance et moyenne, loi empirique

Expliquons à présent le lien entre la notion d'espérance que nous venons de définir, et la notion de moyenne d'un échantillon de valeurs au sens usuel.

Partons d'un échantillon de N valeurs numériques $x_1, \ldots, x_N$. La moyenne arithmétique, au sens usuel, est définie par :

$$\frac{1}{N} \sum_{i=1}^{N} x_i.$$

En appelant S l'ensemble des valeurs *distinctes* présentes dans cet échantillon, et en regroupant les dans la somme ci-dessus les x_i possédant la même valeur, la moyenne se réécrit :

$$\sum_{s \in S} s \times \frac{1}{N}(\text{ nombre d'indices } i \text{ pour lesquels } x_i = s).$$

Autrement dit, la moyenne (au sens usuel) des valeurs d'un échantillon s'écrit également comme la somme des valeurs présentes dans cet échantillon pondérées par leurs fréquences relatives d'apparition dans l'échantillon. La formule donnant l'espérance sous la forme :

$$\mathbb{E}(X) = \sum_{s \in S} s \times \mathbb{P}(X = s),$$

apparaît donc comme une extension de cette définition, dans laquelle les fréquences d'apparition des différentes valeurs sont remplacées par leurs probabilités. Remarquons que, lorsque la loi de X est la loi empirique décrivant un échantillon de valeurs $x_1, \ldots, x_N$, c'est-à-dire lorsque les probabilités affectées aux différentes valeurs de X sont prises égales aux fréquences d'apparition de ces valeurs dans l'échantillon, l'espérance de cette loi empirique est égale à la moyenne (au sens usuel) des valeurs de l'échantillon.

Interprétation fréquentielle

Dans le cadre de l'interprétation fréquentielle de la probabilité – nous reviendrons sur ce point dans le chapitre traitant de la loi des grands nombres –, l'espérance apparaît donc comme la valeur limite de la moyenne (au sens usuel) des valeurs obtenues en répétant un grand nombre de fois l'expérience donnant lieu à la variable aléatoire considérée (voir ce que nous avons dit précédemment sur le lien entre loi et loi empirique).

Cette propriété fondamentale de l'espérance est l'une des raisons pour laquelle cette quantité joue un rôle essentiel en probabilités.

2.6.3 Le raisonnement de Huygens *

Une justification de l'utilisation de l'espérance dans le contexte des paris, indépendante de l'interprétation fréquentielle (qui a bien entendu une grande importance si l'on effectue des paris répétés) a été proposée notamment par Huygens. Nous vous renvoyons à l'exercice 112 pour une description de ce raisonnement.

2.6.4 L'utilité espérée *

Supposons que nous ayons à choisir entre deux situations (aléatoires, variables, incertaines), modélisées respectivement par $(\Omega, \mathbb{P}_1)$ et $(\Omega, \mathbb{P}_2)$, à chaque élément ω de Ω étant associé un nombre réel $U(\omega)$ mesurant quantitativement notre degré de satisfaction lorsque ω est l'issue effectivement réalisée, et appelé **l'utilité** que nous attachons à ω. (Nous supposons ici pour simplifier que Ω est un ensemble fini.) La **règle de maximisation de l'utilité espérée** stipule qu'un individu rationnel sélectionnera celle des deux situations qui attribue à U l'espérance la plus élevée, ou, autrement dit, que les préférences peuvent s'exprimer simplement à partir de l'espérance de l'utilité attachée à une situation.

Bien entendu, il n'est pas toujours facile de définir en pratique une telle mesure d'utilité, et celle-ci dépend de toute façon de nos propres choix et préférences, qui peuvent fort bien varier d'un individu à l'autre. En particulier, même dans le cas simple où les différentes issues de ω sont associées à des gains quantifiables de manière naturelle (par exemple financiers, ou en termes de performances d'un dispositif), l'utilité qu'un individu peut attacher à ω ne s'identifie pas forcément à ce gain, en raison, par exemple, de différences de risque entre les situations envisagées (voir à ce sujet l'exercice 80, et l'exercice 121). La fonction d'utilité doit refléter ce risque, par exemple en pénalisant les gains associés aux situations les plus risquées.

Dans un contexte de choix répétés où la probabilité est interprétée de manière fréquentielle, la règle de la maximisation de l'utilité espérée est assez naturelle, puisque l'espérance représente la moyenne à long terme. Bien entendu, encore faut-il avoir

la possibilité d'effectuer ces choix à long terme (par exemple, ne pas risquer d'être ruiné après quelques échecs et donc dans l'impossibilité de participer aux choix ultérieurs), ce qui n'est pas toujours garanti, et limite la portée de cette règle même dans ce contexte. Quantifier précisément le risque correspondant et ce que signifie un «long» terme en pratique, est, dans ce contexte, une question délicate, mais importante (voir par exemple le chapitre «Loi des grands nombres» pour en apprendre davantage), que nous n'aborderons pas de manière systématique.

Par ailleurs, Von Neumann et Morgenstern ont prouvé que, sous des hypothèses générales censées être vérifiées par un individu rationnel, les préférences entre différentes situations peuvent toujours être exprimées en termes d'utilité espérée.

Plus précisément, supposons donnée une relation de préférence $\prec$ entre les différentes probabilités sur Ω, $\mathbb{P}_1 \prec \mathbb{P}_2$ signifiant que l'on préfère (au sens large) $\mathbb{P}_2$ à $\mathbb{P}_1$, l'indifférence étant traduite par le fait que $\mathbb{P}_1 \prec \mathbb{P}_2$ et $\mathbb{P}_2 \prec \mathbb{P}_1$. Nous noterons $\mathbb{P}_1 \prec_s \mathbb{P}_2$ le fait que l'on préfère $\mathbb{P}_2$ à $\mathbb{P}_1$ au sens strict, c'est-à-dire que $\mathbb{P}_1 \prec \mathbb{P}_2$ et que l'on n'a pas $\mathbb{P}_2 \prec \mathbb{P}_1$.

Le résultat de Von Neumann et Morgenstern est que, si la relation $\prec$ vérifie les quatre propriétés présentées ci-après, il existe nécessairement une fonction U définie sur Ω et à valeurs réelles telles que $\mathbb{E}_{\mathbb{P}_1}(U) \leq \mathbb{E}_{\mathbb{P}_2}(U)$ si et seulement si $\mathbb{P}_1 \prec \mathbb{P}_2$.

Voici ces quatre propriétés :

- la relation de préférence $\prec$ est totale, ce qui signifie que l'on a toujours une préférence (qui peut éventuellement être l'indifférence) entre deux probabilités $\mathbb{P}_1$ et $\mathbb{P}_2$;
- la relation de préférence est transitive, ce qui signifie que si l'on préfère $\mathbb{P}_2$ à $\mathbb{P}_1$ et $\mathbb{P}_3$ à $\mathbb{P}_2$, on doit préférer $\mathbb{P}_3$ à $\mathbb{P}_1$;
- si $\mathbb{P}_1 \prec_s \mathbb{P}_2 \prec_s \mathbb{P}_3$, il existe $a, b \in]0,1[$ tels que $\mathbb{P}_2 \prec_s (a\mathbb{P}_1 + (1-a)\mathbb{P}_3)$ et $(b\mathbb{P}_1 + (1-b)\mathbb{P}_3) \prec_s \mathbb{P}_2$;
- si $\mathbb{P}_1 \prec \mathbb{P}_2$, alors, pour tout $a \in [0,1]$ et toute probabilité $\mathbb{P}_3$, $a\mathbb{P}_1 + (1-a)\mathbb{P}_3 \prec a\mathbb{P}_2 + (1-a)\mathbb{P}_3$.

Nous vous laissons le soin de réfléchir à la signification de ces propriétés en termes de rationalité des choix. Voir l'exercice 81 pour des exemples illustrant le fait que les relations de préférence réelles des individus ne satisfont pas toujours ces axiomes.

2.6.5 L'espérance comme indicateur de position

La donnée de la loi d'une variable aléatoire est une information complexe, pouvant comprendre un grand nombre de valeurs différentes associées à des probabilités variées, et peut donc s'avérer difficile à exploiter directement, par exemple pour effectuer des comparaisons. Il est donc très utile de disposer d'indicateurs numériques qui, sous la forme d'un nombre unique, «résument» une loi en en dégageant des caractéristiques importantes. Dans ce contexte, on utilise souvent l'espérance comme

un résumé numérique synthétique (un seul nombre) susceptible de donner une idée de la localisation des valeurs de la variable aléatoire considérée, qui, rappelons-le, est une fonction, dont les valeurs sont affectées de probabilités variées.

Deux caractérisations de l'espérance

Plaçons-nous, pour simplifier, dans le cas où Ω est un ensemble fini, et supposons que nous cherchions, pour toute probabilité $\mathbb{P}$ sur Ω et toute variable aléatoire X définie sur Ω à valeurs réelles, à définir un nombre unique $h(X, \mathbb{P})$ censé résumer la localisation des valeurs de cette variable. Il semble naturel de demander que h vérifie les conditions suivantes :

- si X et Y sont deux variables aléatoires sur $(\Omega, \mathbb{P})$ vérifiant $\mathbb{P}(X \leq Y) = 1$, alors $h(X, \mathbb{P}) \leq h(Y, \mathbb{P})$ (positivité)
- si $\lambda \in \mathbb{R}$ est un réel fixé, $h(\lambda X, \mathbb{P}) = \lambda h(X, \mathbb{P})$ (invariance par changement d'échelle) ;
- si $c \in \mathbb{R}$ est un réel fixé, $h(X + c, \mathbb{P}) = h(X, \mathbb{P}) + c$ (invariance par translation) ;
- $h(X, \mathbb{P})$ ne dépend que de la loi de X.

Si l'on ajoute la condition supplémentaire suivante (qui peut également sembler naturelle) :

- si X et Y sont deux variables aléatoires sur $(\Omega, \mathbb{P})$, $h(X + Y, \mathbb{P}) = h(X, \mathbb{P}) + h(Y, \mathbb{P})$;

on montre alors facilement (c'est l'exercice 93) que nécessairement $h(X, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}(X)$.

Une autre caractérisation (voir exercice 94) de l'espérance est la suivante : $\mathbb{E}(X)$ est l'unique nombre qui minimise la fonction $a \mapsto \mathbb{E}(X - a)^2$, autrement dit, si l'on cherche à approcher X par une constante, et que l'on mesure l'erreur d'approximation par $\mathbb{E}(X - a)^2$, – on parle alors d'approximation au sens des moindres carrés – l'espérance constitue la meilleure approximation de X par une constante. (Bien entendu, cette propriété ne peut servir à définir l'espérance puisqu'elle suppose dans sa formulation que la notion est déjà définie).

Mentionnons également le rôle joué par l'espérance comme **paramètre de position** dans la définition de certaines lois de probabilité (voir plus haut).

Espérance et valeur typique

Une première confusion, qui vaut tant pour la notion usuelle de moyenne que pour la notion d'espérance, est de croire que celle-ci fournit en général une valeur «typique», ou encore «représentative» des valeurs prises par la variable aléatoire considérée. Pour au moins deux raisons distinctes, ce n'est pas le cas en général.

Une première raison est la compensation pouvant exister entre valeurs supérieures et inférieures à $\mathbb{E}(X)$. Un exemple caricatural est une variable aléatoire prenant la

valeur $a - b$ avec probabilité $1/2$ et $a + b$ avec probabilité $1/2$. L'espérance de cette variable est toujours égale à a, quelle que soit la valeur de b. Si b est effectivement faible devant a (et si $a \neq 0$), on peut raisonnablement considérer que a représente une valeur typique, ou tout au moins, fournit un bon ordre de grandeur, pour la variable aléatoire en question. Si b est au contraire grand devant a, l'espérance ne donne aucune idée des valeurs typiquement prises par la variable aléatoire considérée.

Par exemple, une entreprise dont la moitié des salariés gagne 1000 euros par mois tandis que l'autre moitié gagne 4000 euros par mois fournit un salaire moyen de 2500 euros par mois, qui ne représente en aucun cas une valeur typique du salaire des personnels de cette entreprise.

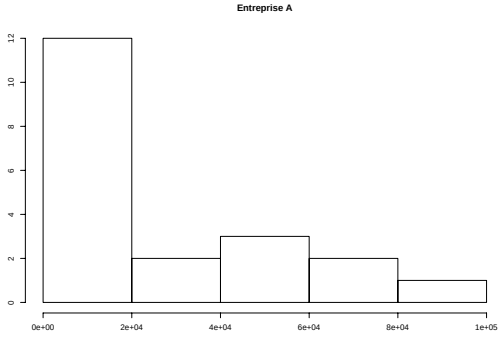
Ou encore : une personne ayant une probabilité de $0,6$ de faire un pas de 1 mètre vers la droite, et $0,4$ de faire un pas de un mètre vers la gauche, se déplace, en espérance, de 20 centimètres vers la droite, mais cette espérance ne représente en rien les possibilités de déplacement réelles. Pire : avec une probabilité de $0,5$ d'aller à gauche et $0,5$ d'aller à droite, le déplacement espéré est nul. Pourtant, la personne se déplace systématiquement d'un mètre par rapport à sa position initiale !

Voici d'autres exemples.

L'entreprise A emploie 20 salariés, dont les rémunérations annuelles nettes en 2005 (classées par ordre décroissant) sont données (en euros) dans le tableau suivant :

Directeur	99123
Cadre 1	66244
Cadre 2	65908
Cadre 3	58163
Cadre 4	52284
Cadre 5	45928
Cadre 6	33354
Cadre 7	25736
Employé 1	15262
Employé 2	14634
Employé 3	13253
Employé 4	13078
Employé 5	12044
Employé 6	12027
Employé 7	12010
Employé 8	11773
Employé 9	11602
Employé 10	11244
Employé 11	10640
Employé 12	10283

L'histogramme correspondant est le suivant.

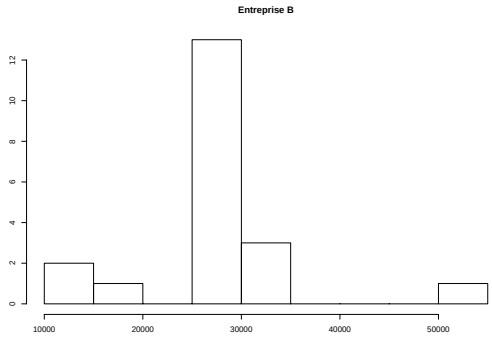


La moyenne de la rémunération des salariés de l'entreprise A s'élève à 29729,5 euros. Manifestement, cette valeur ne représente pas une valeur typique de la rémunération : tous les cadres sauf deux gagnent nettement plus, tandis que tous les employés gagnent nettement moins. De toute façon, il ne saurait exister une unique valeur typique dans ce cas, puisque les rémunérations sont clairement découpées en deux groupes bien distincts, et d'importance numérique comparable.

Considérons maintenant le même tableau pour l'entreprise B, qui opère dans un secteur d'activité totalement différent.

Directeur	50123
Cadre 1	33244
Cadre 2	32908
Cadre 3	31163
Cadre 4	29284
Cadre 5	29128
Cadre 6	29054
Cadre 7	28736
Cadre 8	28363
Cadre 9	28284
Cadre 10	27928
Cadre 11	27854
Cadre 12	27736
Cadre 13	27654
Cadre 14	26936
Cadre 15	26854
Cadre 16	25732
Employé 1	19262
Employé 2	13634
Employé 3	12253

L'histogramme correspondant est le suivant.



Pour l'entreprise B, la rémunération moyenne s'élève à 27806,5 euros, soit une valeur relativement proche de la précédente. Mais cette fois, la distribution de ces valeurs est totalement différente de celle de l'entreprise A, et l'espérance fournit une idée raisonnable de la rémunération typique du personnel de l'entreprise.

Rien ne permet pourtant, à partir de la seule valeur de la rémunération moyenne, de distinguer entre ces deux situations.

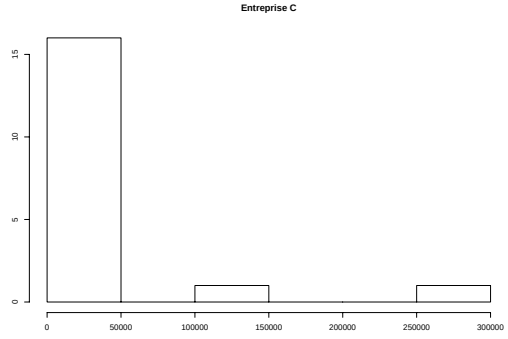
Une deuxième «raison» pour laquelle l'espérance ne représente pas en général

une valeur typique est ce que l'on appelle sa sensibilité aux valeurs extrêmes.

Prenons l'exemple de l'entreprise C.

Directeur	250123
Cadre 1	132244
Employé 1	11262
Employé 2	11189
Employé 3	11124
Employé 1	10982
Employé 2	10903
Employé 3	10884
Employé 1	10880
Employé 2	10875
Employé 3	10864
Employé 1	10859
Employé 2	10841
Employé 3	10838
Employé 1	10832
Employé 2	10822
Employé 3	10818

L'histogramme correspondant est le suivant.



La moyenne des rémunérations est de 30962,33 euros environ. Pourtant, tous les salariés sauf deux gagnent moins de 1000 euros par mois ! Les deux rémunérations du directeur et du cadre sont tellement importantes que leur faible poids dans la moyenne (10%) est compensé par leur valeur élevée. On parle parfois de phénomène du loto pour désigner cette situation : l'existence d'un gain très élevé mais très rare, et donc nullement représentatif, contribue de manière déterminante à la valeur de l'espérance. Voir l'exercice 71. Le même problème peut également se poser lorsqu'un

échantillon de valeurs contient une valeur «aberrante» anormalement élevée, provenant par exemple d'un mauvais fonctionnement de l'appareil de mesure, ou d'une erreur de saisie ou de transmission de la valeur mesurée.

Insistons : les exemples précédents n'ont rien d'exceptionnel ou d'inhabituel. En général, la valeur de l'espérance d'une variable aléatoire est le fruit de compensations entre des valeurs supérieures et des valeurs inférieures à celle-ci, – qui peuvent être très différentes entre elles, et très différentes de l'espérance –, ainsi que de compensations entre valeurs de la variable et probabilités attachées à ces valeurs. L'espérance ne peut être considérée à elle seule comme indiquant en général ne serait-ce qu'un ordre de grandeur des valeurs de la variable.

D'autres indicateurs de position

Pour tenter de pallier les limitations les plus flagrantes de l'espérance en tant qu'indicateur de position, on a souvent recours à d'autres indicateurs numériques, qui ont leurs défauts et limitations propres, mais permettent d'affiner la description de la loi d'une variable aléatoire par rapport à la seule donnée de l'espérance.

L'une d'entre elles est la médiane, ou encore, l'intervalle médian, dont voici les définitions.

On pose $x_{1/2,-}(X) = \sup\{x \in \mathbb{R} \ : \ \mathbb{P}(X \geq x) > 1/2\}$ et $x_{1/2,+}(X) = \inf\{x \in \mathbb{R} \ : \ \mathbb{P}(X \leq x) > 1/2\}$.

On vérifie que ces deux nombres sont toujours bien définis et finis, du fait que $\mathbb{P}(X \leq x)$ tend vers zéro (resp. 1) lorsque x tend vers $-\infty$ (resp. $+\infty$). Qui plus est, par croissance de la fonction de répartition F_X, on vérifie que $x_{1/2,-} \leq x_{1/2,+}$. On vérifie également le fait que $\mathbb{P}(X \leq x_{1/2,+}) \geq 1/2$, et $\mathbb{P}(X \geq x_{1/2,-}) \geq 1/2$.

L'intervalle médian de X est l'intervalle $[x_{1/2,-}; x_{1/2,+}]$. Lorsque $x_{1/2,+} = x_{1/2,+}$, cette valeur commune est appelée **la médiane de** X. Lorsque $x_{1/2,+} \neq x_{1/2,-}$, on prend souvent pour médiane le milieu de l'intervalle médian, soit $\frac{x_{1/2,+}+x_{1/2,-}}{2}$, ce qui permet de définir la médiane de manière systématique. Clairement, la médiane ne présente pas le même phénomène de sensibilité aux valeurs extrêmes que l'espérance.

Considérons le cas particulier d'une loi empirique associée à un échantillon de valeurs $x_1, \ldots, x_N$, on a, en notant $x_{i_1} \leq x_{i_2} \leq \ldots \leq x_{i_N}$. Si $N = 2k+1$, la médiane est simplement égale à $x_{i_{k+1}}$. Si $N = 2k$, $x_{1/2,-} = x_{i_k}$ randis que $x_{1/2,+} = x_{i_{k+1}}$.

Dans les exemples précédents, la médiane associée aux rémunérations est de 13943,5 euros pour l'entreprise A, de 28106 euros pour l'entreprise B, et de 10882 euros pour l'entreprise C. Pour l'entreprise A, la valeur de la médiane est nettement inférieure à l'espérance, et traduit le fait qu'une majorité de salariés sont des employés gagnant moins de 13943 euros, cette valeur ne représentant, pas plus que l'espérance, une valeur typique de la rémunération (il ne peut de toute façon pas y avoir de valeur typique unique dans ce cas). Notamment, cette valeur ne donne

aucune idée précise de la rémunération, relativement homogène, des cadres de l'entreprise. Pour l'entreprise B, médiane et moyenne sont relativement proches, ce qui est cohérent avec la grande homogénéité des rémunérations. Quant à l'entreprise C, on observe bien l'insensibilité de la médiane aux revenus extrêmes.

Le **milieu du domaine** est simplement défini comme $\frac{1}{2}(\sup X - \inf X)$, et il n'est bien défini que lorsque le domaine de X est borné. Cet indicateur ne tient aucun compte des probabilités affectant les différentes valeurs possibles de X, et sa portée est donc assez limitée. Dans les exemples précédents, le milieu du domaine des rémunération de l'entreprise A est 54703 euros, 33688 euros pour l'entreprise B, et 130470,5 euros pour l'entreprise C.

Le **mode** est une notion surtout appropriée aux variables prenant un petit nombre de valeurs distinctes : c'est simplement la valeur la plus probable de X (le mode n'est pas toujours défini car plusieurs valeurs peuvent être ex-æquo). Le fait que, lorsqu'il est défini, le mode soit plus probable que n'importe quelle autre valeur prise individuellement ne signifie pas qu'il soit affecté d'une probabilité importante. Même la valeur la plus probable peut n'avoir qu'une probabilité très faible et ne pas représenter grand-chose de pertinent. Dans les exemples précédents des trois entreprises A, B et C, chaque valeur apparaît exactement une fois, si bien que le mode n'est pas correctement défini.

On peut également définir le deuxième mode, comme la deuxième valeur la plus probable, le troisième mode comme la troisième valeur la plus probable, etc...

Dans le cadre des lois continues, les modes seront plutôt définis comme les pics de la densité.

Voici un extrait du «World Almanac and Book of Facts» (1975), dans lequel se trouve une estimation du nombre des grandes inventions mises au point chaque année entre 1860 et 1959, soit

5 3 0 2 0 3 2 3 6 1 2 1 2 1 3 3 3 5 2 4 4 0 2 3 7 12 3 10 9 2 3 7 7
2 3 3 6 2 4 3 5 2 2 4 0 4 2 5 2 3 3 6 5 8 3 6 6 0 5 2 2 2 6 3 4 4
2 2 4 7 5 3 3 0 2 2 2 1 3 4 2 2 1 1 1 2 1 4 4 3 2 1 4 1 1 1 0 0 2 0

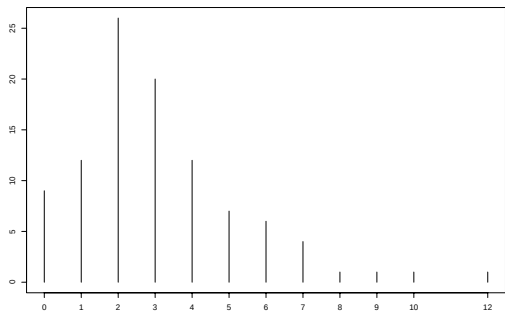
(source : base de données du logiciel `R`)

Voici le tableau des effectifs et des fréquences associé à cette liste.

x	effectif	fréquence
0	9	0.09
1	12	0.12
2	26	0.26
3	20	0.20
4	12	0.12
5	7	0.07
6	6	0.06
7	4	0.04
8	1	0.01
9	1	0.01
10	1	0.01
12	1	0.01

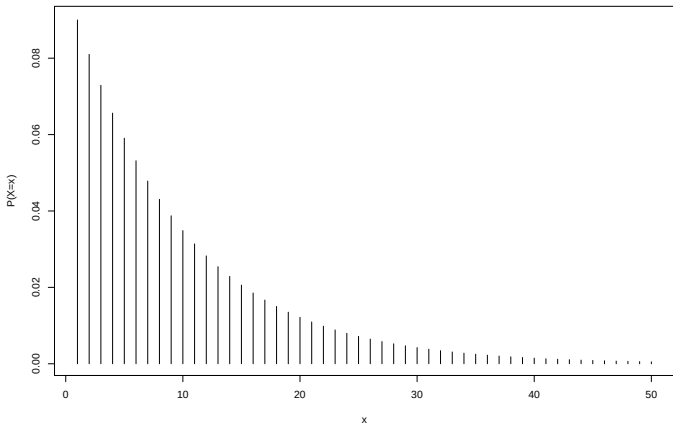
Le mode associé à la loi empirique correspondant à l'échantillon de valeurs proposées est donc 2 (avec une probabilité de 25%), suivi de près par 3 (avec une probabilité de 20%).

Voici le diagramme en bâtons correspondant.



On note en passant son absence de symétrie. L'espérance est ici égale à 3,1, et la médiane à 3. Il y donc coïncidence entre ces trois indicateurs. Comme les exemples précédents le prouvent, ce n'est en général pas le cas.

Pour insister, encore un exemple : la loi géométrique de paramètre $p = 0,1$. Le mode est en 1, la médiane vaut 7, et l'espérance vaut 10. Voici le diagramme en bâtons correspondant.



Des confusions malheureuses mais fréquentes

Voici quelques exemples d'erreurs communément commises à propos de l'espérance.

- « Je raisonne en remplaçant la variable aléatoire par son espérance. » Confusion entre espérance et valeur typique, aux conséquences souvent dévastatrices. Résulte parfois d'une mauvaise compréhension de la signification de l'espérance dans le cadre fréquentiel.
- « Il doit y avoir environ une chance sur deux pour que X soit supérieur à $\mathbb{E}(X)$, et environ une chance sur deux pour que X lui soit inférieur.» C'est totalement faux en général. Cela revient (en gros) à confondre espérance et médiane. En prenant pour X la rémunération d'un salarié, et pour loi la loi empirique associée aux tableaux fournis, la probabilité pour que X dépasse $\mathbb{E}(X)$ est de $0,35$ pour l'entreprise A, de $0,5$ pour l'entreprise B, et de $0,1$ pour l'entreprise C.
- « L'espérance de X est (ou est proche de) la valeur la plus probable de X». C'est également totalement faux. Cela revient (encore en gros) à confondre l'espérance et le mode, qui n'est en général pas une valeur particulièrement probable. Même dans le sens vague du mot, cette affirmation est fausse, comme l'exemple de l'entreprise A le montre bien.
- « Il doit y avoir à peu près autant de chances pour que $X = \mathbb{E}(X) + t$ et $X = \mathbb{E}(X) - t$.» Ou encore «Autant de chances pour que $X \leq \mathbb{E}(X)$ et $X \geq \mathbb{E}(X)$.» Cela revient à supposer que la loi de probabilité de X est symétrique par rapport à son espérance. Lorsque l'on a effectivement symétrie par rapport à une valeur, celle-ci est effectivement égale à l'espérance. En revanche, la plupart des lois ne sont symétriques par rapport à aucune valeur, et en particulier pas

symétriques par rapport à leur espérance. Voir les exemples précédents.

Par ailleurs, croire qu'il existerait un «bon» indicateur dont l'utilisation systématique s'imposerait est une erreur : chaque indicateur présente des avantages et des défauts, peut apporter une information pertinente dans certains cas, ou au contraire se révéler trompeur dans d'autres. L'information contenue dans la loi d'une variable aléatoire est trop riche pour pouvoir, en toute généralité, être résumée par un ou même plusieurs indicateurs numériques synthétiques.

Centrage d'une variable aléatoire

Il s'agit simplement de l'opération consistant à écrire X sous la forme $X = (X - \mathbb{E}(X)) + \mathbb{E}(X)$. D'après la propriété de linéarité de l'espérance (que nous verrons plus bas), on écrit ainsi X comme la somme d'un terme constant égal à son espérance, et d'une variable aléatoire d'espérance égale à zéro.

2.6.6 Variance

La **variance d'une variable aléatoire X est définie comme l'espérance des écarts quadratiques de la variable à son espérance**, c'est-à-dire :

$$\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right],$$

lorsque les espérances X et $X - \mathbb{E}(X)$ possèdent une espérance.

Dans le cas d'une variable aléatoire continue, l'espérance d'une variable aléatoire de densité X est définie de la manière suivante (conformément aux règles de passage du cas discret au cas continu) :

$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} (s - \mathbb{E}(X))^2 f(s) ds.$$

La variable aléatoire $(X - \mathbb{E}(X))^2$ mesure l'écart entre la variable aléatoire X et la valeur constante $\mathbb{E}(X)$. La variance est l'espérance de cet écart.

On introduit également **l'écart-type**, défini comme la racine carrée de la variance :

$$\sigma(X) = \sqrt{\mathbb{V}(X)},$$

et qui a l'avantage de s'exprimer dans la même unité que X.

La variance d'une variable aléatoire joue un rôle prépondérant dans le théorème de la limite centrale, que nous étudierons dans un chapitre ultérieur. Nous nous contenterons d'étudier ici son rôle comme indicateur de dispersion d'une variable aléatoire. Mentionnons également le rôle joué par l'écart-type comme **paramètre d'échelle** dans la définition de certaines lois de probabilité (voir plus haut).

La variance comme indicateur de dispersion

Les indicateurs de dispersion viennent en complément des indicateurs de position, dont l'espérance (ainsi que la médiane) fournit un exemple important. Le but de ces indicateurs est de quantifier la dispersion de la loi de la variable (par exemple, par rapport à un indicateur de position donné).

Dans ce contexte, la variance apparaît comme l'espérance d'une quantité mesurant l'écart entre X et $\mathbb{E}(X)$. Elle mesure donc, avec toutes limitations inhérentes à l'utilisation de l'espérance pour résumer la loi d'une variable aléatoire, la dispersion des valeurs prises par X par rapport à $\mathbb{E}(X)$.

L'écart-type permet de comparer directement la mesure de dispersion fournie par la variance aux valeurs prises par X. Le nom d'écart-type est trompeur, puisque l'espérance de $(X - \mathbb{E}(X))^2$ ne correspond pas en général à une valeur typique de $(X - \mathbb{E}(X))^2$.

Une caractérisation alternative de la variance, qui ne fait pas apparaître explicitement l'espérance de X, est donnée dans l'exercice 95.

Notons que, si l'on se trouve dans une situation où $\mathbb{E}(X)$ n'est pas un indicateur de position pertinent pour X, la pertinence de la variance en tant qu'indicateur de dispersion est d'emblée remise en question. Enfin, même dans les cas où l'espérance fournit une indication satisfaisante de position pour X, la variance peut très bien ne pas fournir une indication de dispersion satisfaisante.

Pour prendre un exemple extrême, une variable aléatoire prenant la valeurs a avec probabilité $99,9\%$, $a + b$ avec probabilité $0,5\%$ et $a - b$ avec probabilité $0,5\%$ possède une espérance égale à a, que l'on peut raisonnablement considérer comme une valeur typique. Pourtant, l'écart-type est égal à $(b^2 \times 0,1\%)^{1/2}$, et, si l'on choisit par exemple $b = 10000$, on obtient un écart-type d'environ 316, qui ne représente certainement pas une valeur typique de l'écart ! (Voir également l'exercice 71)

Remarquons par ailleurs que, dans le cas limite d'une variable aléatoire X vérifiant $\mathbb{V}(X) = 0$, X est nécessairement égale à une constante avec probabilité 1 : $\mathbb{P}(X = \mathbb{E}(X)) = 1$ (X n'est pas forcément constante *stricto sensu*, car elle peut prendre des valeurs arbitraires pour des ω de probabilité nulle sans que sa loi en soit modifiée).

Reprenant les exemples des trois entreprises A, B et C, on obtient des écarts-types pour la loi empirique de la rémunération égaux respectivement à : 25882 euros environ pour l'entreprise A, 7586 euros environ pour l'entreprise B, et 61695 euros environ pour l'entreprise C.

Pour l'entreprise A, on note que l'écart-type surestime significativement (nous ne donnons pas pour l'instant à ce terme de signification plus précise que sa signification courante) l'écart entre la rémunération moyenne (29729,5 euros) et la rémunération des employés. En effet, cet écart s'échelonne entre 14467,5 et 19446,5 euros. Concer-

nant les cadres, cet écart n'est quasiment jamais proche de l'écart réel, la liste des (valeurs absolues des) écarts étant la suivante : 69393,5 ; 36514,5 ; 36178,5 ; 28433,5 ; 22554,5 ; 16198,5 ; 3624,5 ; 3993,5. L'écart-type fournit néanmoins, de manière très grossière, une mesure de l'écart, et une indication de l'ordre de grandeur de la dispersion des rémunérations.

Pour l'entreprise B, l'écart-type surestime globalement l'écart à la rémunération moyenne, qui a par exemple plus de 70% de chances d'être inférieur à la moitié de l'écart-type, deux autres valeurs étant voisines de celui-ci, et deux autres encore très éloignées. Ici encore, on n'obtient qu'une estimation très grossière de l'écart, et de l'ordre de grandeur de la dispersion des rémunérations.

Quant à l'entreprise C, l'écart-type ne représente à peu près rien, l'espérance étant elle-même affectée par les deux valeurs extrêmes. L'écart entre la rémunération et sa valeur moyenne est, avec une probabilité de 90%, de l'ordre de 20000 euros, et, pour les deux valeurs extrêmes, de 219160,67 et 101281,67 euros.

Autres indicateurs de dispersion

Bien d'autres types d'indicateurs de dispersion peuvent être utilisés.

Un indicateur très grossier est par exemple la largeur de l'intervalle des valeurs de X.

Un autre, très utilisé, est la distance interquartile.

De manière générale, pour $r \in]0,1[$, on définit l'intervalle fractile d'ordre r comme l'intervalle $[x_{r,-}, x_{r,+}]$, où $x_{r,+}(X) = \sup\{x \in \mathbb{R} \ : \ \mathbb{P}(X \geq x) > r\}$; et $x_{r,-}(X) = \inf\{x \in \mathbb{R} \ : \ \mathbb{P}(X \leq x) > 1 - r\}$.

On vérifie, comme dans le cas de la médiane, que ces quantités sont toujours définies, et vérifient le fait que $\mathbb{P}(X \geq x_{r,+}(X)) \leq r$ et $\mathbb{P}(X \leq x_{r,-}(X)) \leq 1 - r$.

Lorsque cet intervalle est réduit à un point, on l'appelle le **fractile d'ordre** r **de** X**,** et on le note $x_r(X)$. Lorsque ce n'est pas le cas, on considère souvent le point $\frac{x_{r,+}(X)+x_{r,-}(X)}{2}$ afin que les fractiles soient toujours définis.

On utilise le terme de «quantile» de manière interchangeable avec celui de «fractile».

(La définition des fractiles n'est pas totalement fixée, voir par exemple l'aide en ligne de `R` : `help(quantile)` pour une liste de définitions possibles, différentes, quoique voisines. ou l'article Hyndman, R. J. and Fan, Y. (1996) Sample quantiles in statistical packages, American Statistician, 50, 361–365.)

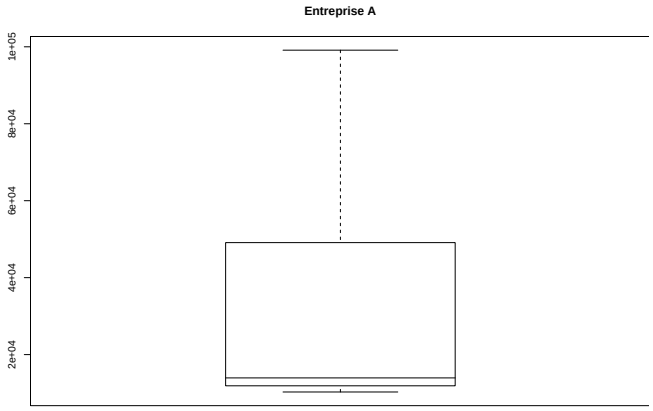
La médiane (ou l'intervalle médian) correspond au fractile d'ordre $1/2$. On appelle **quartiles** les fractiles d'ordre $1/4, 2/4, 3/4$, **déciles** les fractiles d'ordre $1/10, 2/10, \ldots, 9/10$.

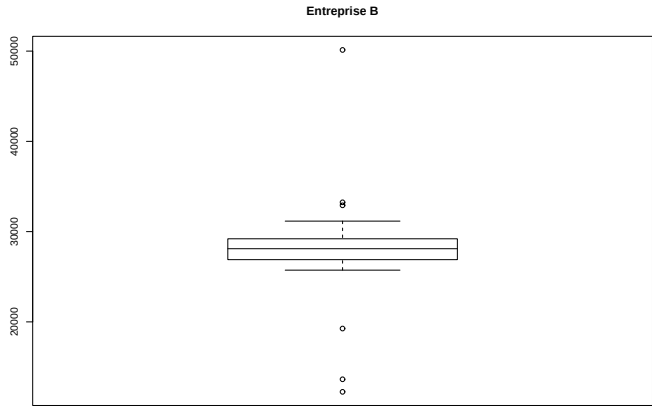
La **distance interquartile** $d_{iq.}$ est alors définie comme l'écart entre le fractile d'ordre $3/4$ et le fractile d'ordre $1/4$, et fournit un indicateur numérique de dispersion, qui a l'avantage d'être insensible aux valeurs extrêmes.

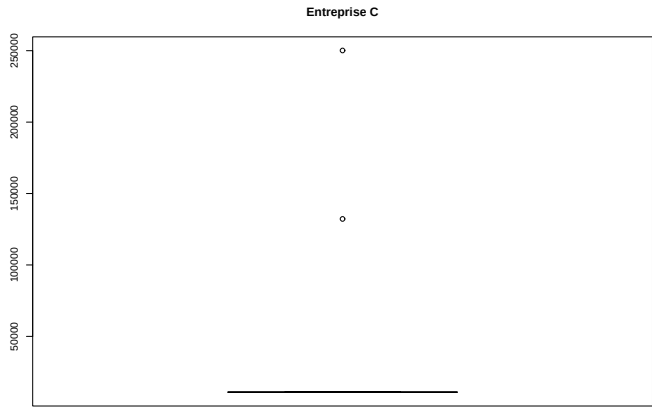
Cet indicateur est à la base de la représentation graphique dite «**diagramme à moustaches**», aussi connu sous le nom de **diagramme en boîte** ou **boxplot**, qui consiste à résumer la distribution de X par un graphique sur lequel on représente sur un axe vertical gradué :

- la médiane, et le premier et le dernier quartile (ceux d'ordre $1/4$ et $3/4$), sous forme de traits horizontaux qui délimitent la boîte ;
- les deux intervalles (les moustaches) reliant respectivement $x_{1/4}$ à la plus grande valeur de la variable qui soit inférieure à $x_{1/4} + 1,5 \times d_{iq.}$, et $x_{3/4}$ à la plus petite valeur de la variable qui soit supérieure à $x_{3/4} - 1,5 \times d_{iq.}$; les moustaches recouvrent donc l'ensemble des valeurs considérées comme non extrêmes ;
- les valeurs qui se trouvent soit au-dessus de $x_{1/4} + 1,5 \times d_{iq.}$, soit en-dessous de $x_{3/4} - 1,5 \times d_{iq.}$ (et qui sont considérées comme des valeurs extrêmes), sous forme de points.

Voici par exemple les diagrammes à moustache associés à la distribution empirique des rémunérations dans les entreprises A,B et C.







On constate que ces graphiques rendent bien compte des différences qualitatives existant entre les trois distributions : deux groupes de rémunérations pour l'entreprise A, l'un assez resserré (les employés), l'autre plus étalé (les cadres) ; une répartition assez concentrée des revenus pour l'entreprise B ; une répartition comportant deux extrêmes très éloignés du reste de la distribution pour l'entreprise C.

La même remarque générale que celle faite à propos des indicateurs de position s'applique : chaque indicateur possède des avantages et des défauts, qui rendent leur valeur plus ou moins pertinente ou trompeuse selon le contexte. L'information contenue dans la loi d'une variable aléatoire est trop riche pour pouvoir, en toute généralité, être résumée par un ou même plusieurs indicateurs numériques synthétiques.

2.6.7 L'inégalité de Markov

L'inégalité de Markov permet d'extraire des informations quantitatives sur la localisation d'une variable aléatoire à partir de la connaissance de l'espérance de celle-ci et d'une borne sur le domaine de ces valeurs. Sans perte de généralité (quitte à changer le signe et à ajouter une constante de manière à notre variable aléatoire), le problème qui se pose est le suivant : supposons que nous ayons affaire à une variable aléatoire dont les valeurs ne peuvent être que positives ou nulles, et que nous ne connaissions de cette variable aléatoire que la valeur de son espérance. Que pouvons-nous en déduire sur la localisation des valeurs de cette variable ?

Un résultat simple mais d'une grande importance, dans cette direction, est l'inégalité suivante, appelée **inégalité de Markov** [8]. Si X est une variable aléatoire positive, alors, pour tout $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

La preuve de cette inégalité est très simple. Partons de la définition de l'espérance. $\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)$. Appelons A l'ensemble des éventualités élémentaires ω telles que : $X(\omega) \geq a$ (autrement dit, A est l'événement : «$X \geq a$» dont nous cherchons à majorer la probabilité). et ne conservons, dans la somme définissant l'espérance ci-dessus, que les termes associés aux éléments de A. Comme tous les $X(\omega)$ sont positifs ou nuls, on obtient ainsi un nombre inférieur ou égal $\mathbb{E}(X)$. Autrement dit : $\sum_{\omega \in A} X(\omega)\mathbb{P}(\omega) \leq \mathbb{E}(X)$. Comme, pour tout $\omega \in A$, $X(\omega)$ est supérieur ou égal à a, nous obtenons encore que : $a\mathbb{P}(A) = \sum_{\omega \in A} a\mathbb{P}(\omega) \leq \sum_{\omega \in A} X(\omega)\mathbb{P}(\omega) \leq \mathbb{E}(X)$, ce qui prouve l'inégalité de Markov.

Connaissant l'espérance d'une variable aléatoire positive (si cette espérance est définie), on dispose donc d'une borne sur la probabilité que cette variable prenne des valeurs plus grandes qu'une valeur seuil a. Le terme par lequel l'inégalité majore $\mathbb{P}(X \geq a)$ est d'autant plus petit que a est grand en comparaison de $\mathbb{E}(X)$. La probabilité pour que X prenne des valeurs dont l'ordre de grandeur dépasse beaucoup son espérance est donc faible, et nous disposons d'une expression quantitative de ce fait. On note que l'inégalité de Markov ne nous renseigne réellement sur $P(X \geq a)$ que si $a > \mathbb{E}(X)$ (une probabilité est toujours inférieure ou égale à 1 !).

A quel point cette inégalité peut-elle être considérée comme précise ? Une réponse possible est que, très souvent, cette inégalité est assez grossière, c'est-à-dire que $\mathbb{P}(X \geq a)$ est bien plus petite que $\frac{\mathbb{E}(X)}{a}$. Qui plus est, cette inégalité ne fournit une information non-triviale que lorsque $a > \mathbb{E}(X)$. Prenons l'exemple d'une variable aléatoire X de loi de Poisson de paramètre 2, pour laquelle on a donc $\mathbb{E}(X) = 2$. (Les valeurs présentées sont arrondies au plus proche à partir de la deuxième décimale significative.)

8. Du nom de A. A. Markov (1856–1922).

a	1	2	3	4	5	6	7	8
$\mathbb{P}(X \geq a)$	0,86	0,59	0,32	0,14	0,053	0,016	0,0045	0,0010
$\mathbb{E}(X)/a$	2	1	0,67	0,50	0,40	0,33	0,29	0,25

A présent, voici l'exemple d'une variable aléatoire de loi binomiale de paramètres $n = 100$ et $p = 0,4$ pour laquelle on a donc $\mathbb{E}(X) = 40$. (Les valeurs présentées sont arrondies au plus proche à partir de la deuxième décimale significative.)

a	40	42	44	46	48	50	52	54
$\mathbb{P}(X \geq a)$	0,54	0,38	0,24	0,13	0,064	0,027	0,010	0,003
$\mathbb{E}(X)/a$	1	0,95	0,91	0,86	0,83	0,80	0,77	0,74

Ces deux exemples illustrent le fait que, dans certains cas (en fait, souvent), la fonction $a \mapsto \mathbb{P}(X \geq a)$ décroît bien plus rapidement avec a que $\mathbb{E}(X)/a$, ce qui fait que l'inégalité de Markov, quoique valable (nous l'avons prouvée ! ! !), n'est pas précise. Autre exemple : la loi exponentielle, pour laquelle on a $\mathbb{P}(X \geq a) \leq \exp(-a/\mathbb{E}(X))$, ce qui met encore en évidence ce phénomène.

Pour autant, on ne peut pas en toute généralité espérer (c'est-à-dire pour toute variable aléatoire positive dont l'espérance est définie) obtenir mieux que l'inégalité de Markov, car il est facile (voir l'exercice 119) de construire des exemples de variables aléatoires positives pour lesquels $\mathbb{P}(X \geq a)$ est aussi proche de $E(X)/a$ qu'on le souhaite, au moins pour certaines valeurs de a. Des hypothèses supplémentaires sur la loi de X (comme par exemple, le fait que la loi de X appartienne à une famille de lois paramétriques particulière, comme les lois de Poisson, ou exponentielle, par exemple) sont donc nécessaires pour que l'on puisse espérer déduire de la seule connaissance de l'espérance de X des informations sur la localisation des valeurs de X plus précises que celles fournies par l'inégalité de Markov.

L'inégalité de Markov fournit une borne supérieure sur les probabilités du type $\mathbb{P}(X \geq a)$, c'est-à-dire sur la probabilité pour que X dépasse une certaine valeur a, cette inégalité ayant un réel contenu lorsque $a > \mathbb{E}(X)$.

La connaissance de $\mathbb{E}(X)$ nous permet-elle de déduire des informations non-triviales sur d'autres probabilités relatives à la localisation des valeurs de X ?

On pourrait chercher à obtenir des bornes inférieures sur des probabilités du type $\mathbb{P}(X \geq a)$ lorsque $a > \mathbb{E}(X)$ (ce qui est un peu contradictoire avec l'utilisation de l'espérance comme indicateur de position, mais bon...), on voit facilement que cette probabilité peut être rendue égale à zéro dans certains cas, et que l'on ne peut donc rien dire à ce sujet au seul vu de l'espérance.

Pour $a = \mathbb{E}(X)$, on note que l'on a nécessairement $\mathbb{P}(X \geq \mathbb{E}(X)) > 0$ et $\mathbb{P}(X \leq \mathbb{E}(X)) > 0$. Il est facile de construire des exemples où l'une ou l'autre de ces probabilités sont aussi petites qu'on le souhaite (elles ne peuvent évidemment pas être petites simultanément, du fait que leur somme est supérieure ou égale à 1),

et l'on ne peut donc pas dire quoique ce soit de plus en toute généralité (c'est-à-dire sans hypothèses supplémentaires sur la loi de X).

Pour $a < \mathbb{E}(X)$, étant donnés deux nombres $a, b > 0$ vérifiant $a < b$ et $0 < p < 1$, on peut toujours construire une variable aléatoire positive X vérifiant $\mathbb{P}(X \leq a) = p$ et $\mathbb{E}(X) = b$. Il suffit de choisir X prenant la valeur a avec probabilité p et $(b - ap)/(1 - p)$ avec probabilité $1 - p$.

On constate donc que l'on ne peut rien dire sans hypothèse supplémentaire sur la probabilité $\mathbb{P}(X \geq a)$ ou, en passant au complémentaire, $\mathbb{P}(X < a)$.

Notons par ailleurs que l'hypothèse selon laquelle la variable aléatoire X considérée ne prend que des valeurs positives est essentielle. Sans hypothèse de ce type, la seule connaissance de l'espérance $\mathbb{E}(X)$ ne permet pas de dire quoique ce soit de quantitatif sur les probabilités du type $\mathbb{P}(X \geq a)$ ou $\mathbb{P}(X \leq a)$ sans hypothèses supplémentaires, hormis le fait trivial que $\mathbb{P}(X \geq \mathbb{E}(X)) > 0$ et $\mathbb{P}(X \leq \mathbb{E}(X)) > 0$. Ceci en raison des compensations entre valeurs positives et négatives qui peuvent survenir dans le calcul de $\mathbb{E}(X)$.

Par exemple, une variable aléatoire d'espérance égale à zéro peut prendre des valeurs positives et négatives arbitrairement grandes en valeur absolue (penser à une v.a. prenant la valeur a avec probabilité $1/2$ et $-a$ avec probabilité $1/2$).

On peut néanmoins obtenir des estimations sur des variables de signe quelconque, mais en considérant les espérances de fonctions positives de ces variables aléatoires, telles que $|X|^p$ ou $\exp(tX)$.

Un exemple célèbre et important est **l'inégalité de Bienaymé-Tchebychev** [9], que l'on obtient en appliquant l'inégalité de Markov à la variable aléatoire positive $[X - \mathbb{E}(X)]^2$, soit

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq t\right) \leq \frac{\mathbb{V}(X)}{t^2} = \left[\frac{\sigma(X)}{t}\right]^2,$$

valable pour tout $t > 0$.

L'inégalité de Bienaymé-Tchebychev fournit donc une majoration de la probabilité pour que la valeur prise par X s'écarte de $\mathbb{E}(X)$ d'une distance supérieure (ou égale) à t. Cette inégalité fait intervenir le rapport entre l'écart-type de X, $\sigma(X)$, et t, c'est-à-dire l'écart à l'espérance, et la majoration fournit une valeur d'autant plus petite que l'écart entre X et son espérance est supposé grand devant l'écart-type $\sigma(X)$. En ce sens, observer une valeur de X dont l'écart par rapport à $\mathbb{E}(X)$ dépasse de beaucoup l'écart-type est donc très improbable. Ceci entraîne que, lorsque l'écart-type est lui-même petit devant l'espérance, la valeur de $\mathbb{E}(X)$ représente la valeur typique de X.

Conformément à la discussion précédente sur l'inégalité de Markov, cette inégalité est très souvent imprécise (la majoration est exagérément pessimiste), mais on ne

9. I.-J. Bienaymé (1796–1878), P. L. Tchebychev (1821–1894).

peut pas l'améliorer en toute généralité, car il existe des cas où celle-ci est peut-être rendue arbitrairement précise. Enfin, on ne peut rien déduire, en l'absence d'informations ou d'hypothèses supplémentaires au sujet de la variable aléatoire considérée, sur la probabilité pour que l'écart soit effectivement plus grand qu'une fraction donnée de l'écart-type : un écart beaucoup plus grand que l'écart-type est, d'après ce qui précède, très improbable, mais rien ne prouve que les écarts ne sont pas typiquement beaucoup plus petits que l'écart-type (voir l'exemple de l'entreprise C dans ce qui précède, ou l'exercice 71).

Exemple des entreprises A, B, C.

Pour illustrer cette inégalité, considérons une variable aléatoire de loi binomiale de paramètres $n = 50$ et $p = 0,6$. (Les valeurs présentées sont arrondies au plus proche à partir de la deuxième décimale significative.)

a	2	3	4	5	6	7	8	9
$\mathbb{P}(\lvert X - \mathbb{E}(X)\rvert \geq a)$	0,67	0,47	0,31	0,19	0,11	0,059	0,029	0,0013
$V(X)/a^2$	3	1,33	0,75	0,48	0,33	0,24	0,19	0.15

Considérons à présent une variable aléatoire de loi de Poisson de paramètre $\lambda = 15$. (Les valeurs présentées sont arrondies au plus proche à partir de la deuxième décimale significative.)

a	4	5	6	7	8	9	10
$\mathbb{P}(\lvert X - \mathbb{E}(X)\rvert \geq a)$	0,37	0,24	0,15	0,09	0,050	0,027	0,014
$V(X)/a^2$	0,94	0,60	0,41	0,30	0,23	0,19	0,12

2.6.8 Opérations algébriques : linéarité de l'espérance

Étant données deux variables aléatoires X et Y à valeurs réelles définies sur un même espace de probabilité $(\Omega, \mathbb{P})$, on peut leur asocier diverses variables aléatoires en combinant X et Y à l'aide d'opérations algébriques telles que somme et produit : la variable aléatoire somme, définie, pour tout $\omega \in \Omega$, par $(X+Y)(\omega) = X(\omega)+Y(\omega)$, la variable aléatoire produit, définie, pour tout $\omega \in \Omega$, par $(X \times Y)(\omega) = X(\omega) \times Y(\omega)$. Il est important, en théorie et en pratique, de savoir comment l'espérance et la variance se comportent vis-à-vis de ces opérations, car les sommes ou les produits de variables aléatoires interviennent dans de nombreuses situations.

En ce qui concerne l'espérance, on regroupe sous le nom de **linéarité de l'espérance** les deux propriétés fondamentales suivantes, valables pour tout couple de variables aléatoires X et Y dont l'espérance est bien définie, et tout nombre réel λ :

$$\begin{cases} \mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y), \\ \mathbb{E}(\lambda \times X) = \lambda \times \mathbb{E}(X). \end{cases}$$

La démonstration de ces propriétés est presque immédiate. Partant de la définition de l'espérance, on vérifie que :

$$\mathbb{E}(X)+\mathbb{E}(Y) = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega)+\sum_{\omega\in\Omega} Y(\omega)\mathbb{P}(\omega) = \sum_{\omega\in\Omega}(X(\omega)+Y(\omega))\mathbb{P}(\omega) = \mathbb{E}(X+Y),$$

la somme définissant l'espérance de $X+Y$ étant bien définie dès lors que les sommes définissant l'espérance de X et l'espérance de Y le sont en vertu de l'inégalité $|X(\omega)+Y(\omega)| \leq |X(\omega)| + |Y(\omega)|$. De même,

$$\lambda \times \mathbb{E}(X) = \lambda \times \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega) = \sum_{\omega\in\Omega} \lambda \times X(\omega)\mathbb{P}(\omega) = \mathbb{E}(\lambda \times X).$$

Cette propriété de linéarité de l'espérance est fondamentale, en particulier parce qu'elle fournit la possibilité d'évaluer l'espérance d'une somme de variables aléatoires à partir des espérances individuelles de ces variables, même lorsqu'il existe entre celles-ci des relations de dépendance éventuellement complexes.

Cette propriété a bien entendu des conséquences sur le comportement de la variance.

Ainsi, la propriété de linéarité de l'espérance permet d'en donner une **nouvelle expression :**

$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

Pour vérifier cette formule, partons de la définition :

$$\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right].$$

En développant, on obtient que :

$$(X - \mathbb{E}(X))^2 = X^2 - 2\mathbb{E}(X) \times X + (\mathbb{E}(X))^2.$$

La propriété de linéarité de l'espérance entraîne donc que :

$$\mathbb{V}(X) = \mathbb{E}(X^2) - 2\mathbb{E}(\mathbb{E}(X) \times X) + \mathbb{E}((\mathbb{E}(X))^2).$$

L'espérance $\mathbb{E}(X)$ étant un nombre déterministe (non-aléatoire, constant), la linéarité de l'espérance, toujours, ainsi que le fait que l'espérance d'une variable aléatoire constante est égale à cette constante, entraîne que :

$$\mathbb{V}(X) = \mathbb{E}(X^2) - 2\mathbb{E}(X) \times \mathbb{E}(X) + (\mathbb{E}(X))^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

Étudions maintenant l'effet produit sur la variance et l'écart-type par la multiplication par un nombre réel fixé λ. À l'aide, par exemple, de la nouvelle expression pour la variance, et grâce à la linéarité de l'espérance, on constate que :

$$\begin{aligned}
\mathbb{V}(\lambda \times X) &= \mathbb{E}((\lambda \times X)^2) - (\mathbb{E}(\lambda \times X))^2 \\
&= \mathbb{E}(\lambda^2 \times X^2) - (\lambda \times \mathbb{E}(X))^2 \\
&= \lambda^2 \times \mathbb{E}(X^2) - \lambda^2 \times (\mathbb{E}(X))^2 \\
&= \lambda^2 \times \mathbb{V}(X).
\end{aligned}$$

Multiplier une variable aléatoire par λ a donc pour effet de multiplier sa variance par λ^2, et l'écart-type se trouve donc, lui, multiplié par $|\lambda|$.

$$\mathbb{V}(\lambda \times X) = \lambda^2 \times \mathbb{V}(X) \ , \ \sigma(\lambda \times X) = |\lambda| \times \sigma(X).$$

Le comportement de la variance vis-à-vis de la somme sera étudié ultérieurement.

Normalisation par l'écart-type

Etant donnée une variable aléatoire X de variance non-nulle, on obtient une variable aléatoire de variance 1 en considérant $X(\sigma(X))^{-1}$. Cette dernière quantité est alors une variable aléatoire sans dimension, et de variance fixée à 1. On dit parfois que $X(\sigma(X))^{-1}$ est la variable X réduite.

Un exemple de raisonnement basé sur la linéarité de l'espérance

Le chef du protocole doit organiser les invitations au bal de l'ambassade. Le prince héritier (il s'agit d'une monarchie) donne pour instruction, soit d'inviter le Pérou, soit d'exclure le Quatar. La reine, de son côté, réclame que soient invités le Quatar ou la Roumanie (ou les deux à la fois). Le roi, par esprit de contradiction, exige que soient exclus du bal soit la Roumanie, soit le Pérou, soit les deux. Y a-t-il un moyen de satisfaire simultanément toutes les exigences de la famille royale ?

On peut coder le problème à l'aide de variables booléennes : p prendra la valeur VRAI si l'ambassadeur du Pérou est invité, et FAUX sinon, et, de même, q (resp. r) prendront les valeurs VRAI ou FAUX suivant que l'ambassadeur du Quatar (resp. de Roumanie) est invité ou non.

Les trois contraintes imposées par la famille royale peuvent se traduire de la façon suivante :

$$(p \vee \bar{q}) \ , \ (q \vee r) \ , \ (\bar{r} \vee q)$$

(le symbole $\vee$ désignant le «ou».) Ce problème est un exemple (simple) du problème général de la satisfaisabilité des clauses logiques, défini de la façon suivante : on considère n variables booléennes $x_1, \ldots, x_n$ pouvant prendre chacune la valeur «vrai» ou «faux», et des «clauses logiques» de la forme :

$$y_{l_1} \vee y_{l_2} \vee \ldots \vee y_{l_k},$$

chaque y_i pouvant être égal à x_i ou à sa négation $\bar{x}_i$. Par exemple : $x_1 \vee \bar{x}_3 \vee x_4$, ou $\bar{x}_2 \vee x_3 \vee \bar{x}_4 \vee x_5$ sont de telles clauses. Une clause est dite satisfaite si l'une au moins des variables qui y figurent prend la valeur «vrai». Le problème de la satisfaisabilité est de déterminer si, étant donné un ensemble de clauses, il est possible de trouver une assignation des variables $x_1, \ldots, x_n$ qui satisfasse simultanément toutes ces

clauses. (Ce n'est bien entendu pas toujours possible.) Il s'agit d'un problème NP-complet, jouant un rôle important en informatique théorique, et personne ne sait à l'heure actuelle s'il existe une méthode pour le résoudre essentiellement meilleure que d'essayer une par une toutes les combinaisons de valeurs possibles pour les y_i.

Nous allons prouver, à l'aide d'un raisonnement de probabilités, le résultat suivant : pour tout ensemble de m clauses, il existe une assignation des variables y_i telle qu'au moins $m/2$ clauses sont satisfaites.

Supposons que les valeurs des n variables booléennes sont tirées au hasard, c'est-à-dire données par n variables aléatoires mutuellement indépendantes $X_1, \ldots, X_n$ prenant chacune la valeur «vrai» avec probabilité $1/2$, et «faux» avec probabilité $1/2$. Notons $C_1, \ldots, C_m$ les différentes clauses, et intéressons-nous à la probabilité qu'une clause donnée C_i soit satisfaite. Appelons k le nombre de variables apparaissant dans C_i. Par définition, C_i est satisfaite dès que l'une au moins des k variables qui y figurent prend la valeur «vrai». En conséquence, la probabilité pour que C_i ne soit pas satisfaite est la probabilité pour que chacune de ces k variables prenne la valeur «faux», et vaut donc, les variables étant mutuellement indépendantes, $(1/2)^k$. La probabilité pour que C_i soit satisfaite est donc égale à $1 - (1/2)^k$, et se trouve donc toujours supérieure ou égale à $1/2$.

À présent, intéressons-nous au nombre total X de clauses satisfaites simultanément. Par définition, ce nombre s'écrit :

$$X = \sum_{i=1}^{m} \mathbf{1}_{C_i \text{ est satisfaite}}.$$

Grâce à la propriété d'additivité de l'espérance, on a :

$$\mathbb{E}(X) = \sum_{i=1}^{m} \mathbb{E}\left(\mathbf{1}\{C_i \text{ est satisfaite}\}\right) = \sum_{i=1}^{m} \mathbb{P}(C_i \text{ est satisfaite}) \geq m \times \frac{1}{2}.$$

L'espérance du nombre total de clauses satisfaites en attribuant les valeurs des n variables booléennes aléatoirement est donc supérieur ou égal à $m/2$. En particulier, il existe obligatoirement au moins une assignation des variables telle qu'au moins $m/2$ clauses soient satisfaites, ce qui constitue le résultat que nous souhaitions démontrer. Ce petit argument illustre, sur un exemple très simple, la puissance de ce que l'on appelle la *méthode probabiliste*, qui consiste à introduire artificiellement le hasard dans une situation où il n'intervient pas initialement, de façon à résoudre plus simplement le problème posé. Ici, la difficulté fondamentale réside dans le fait que plusieurs clauses peuvent faire intervenir les mêmes variables, ce qui se traduit par le fait que les satisfactions des différentes clauses ne forment pas des événements indépendants. Pour autant, la propriété d'additivité de l'espérance, valable sans aucune hypothèse d'indépendance, permet de conclure très simplement. Essayez-donc

de prouver le résultat par une autre méthode ! Pour en savoir (beaucoup) plus sur ce type d'approche, et en particulier sur l'utilisation de l'aléatoire pour concevoir des algorithmes simples et performants dans de nombreuses situations, vous pouvez consulter l'ouvrage de Motwani et Raghavan cité dans la bibliographie.

2.6.9 Opérations algébriques : espérance d'un produit

Contrairement à ce qui a lieu pour la somme, **l'espérance du produit $X \times Y$ n'est pas en général le produit des espérances**, comme le montre l'exemple suivant : si X suit une loi de Bernoulli de paramètre $p \in]0,1[$, $X = X \times X$ car X ne prend que les valeurs 0 et 1. Par conséquent, l'espérance de X est égale à l'espérance de X^2, et vaut $p \times 1 + (1-p) \times 0 = p$, et diffère par conséquent du carré de l'espérance de X, égal à $p \times p = p^2$.

Une propriété fondamentale de l'espérance est la suivante : **si X et Y sont indépendantes, l'espérance du produit de X par Y est le produit de leurs espérances :**

$$\mathbb{E}(X \times Y) = \mathbb{E}(X) \times \mathbb{E}(Y).$$

Le fait que l'espérance du produit XY existe fait également partie du résultat. Avant de démontrer cette propriété, signalons qu'elle ne suffit pas à caractériser l'indépendance de X et Y. En effet, sur l'espace des possibles $\Omega = \{1,2,3,4\}$ muni de la probabilité uniforme, définissons X et Y par :

$$X(1) = -1 \ , \ X(2) = 0 \ , \ X(3) = 0 \ , \ X(4) = 1,$$

et

$$Y(1) = 0 \ , \ Y(2) = -1 \ , \ Y(3) = 1 \ , \ Y(4) = 0.$$

On vérifie que $X \times Y(\omega) = 0$ pour tout $\omega \in \Omega$, et par conséquent, $\mathbb{E}(X \times Y) = 0$. D'autre part, $\mathbb{E}(X) = \frac{1}{4} \times (-1) + \frac{1}{4} \times 0 + \frac{1}{4} \times 0 + \frac{1}{4} \times 1 = 0 = \mathbb{E}(Y)$. On a donc bien $\mathbb{E}(X \times Y) = \mathbb{E}(X) \times \mathbb{E}(Y)$. En revanche, X et Y ne sont pas indépendantes, car, par exemple, $\mathbb{P}(X = 0, Y = 0) = 0$ alors que $\mathbb{P}(X = 0) = \mathbb{P}(Y = 0) = \frac{1}{2}$, d'où le fait que $\mathbb{P}(X = 0, Y = 0) \neq \mathbb{P}(X = 0) \times \mathbb{P}(Y = 0)$. Nous reviendrons dans la partie suivante sur cette question.

Réinsistons sur le fait que, sans hypothèses supplémentaires (telles que l'indépendance), l'espérance d'un produit n'a aucune chance d'être le produit des espérances. Considérons encore un exemple.

Jojo réclame une augmentation de salaire à son employeur. Celui-ci, apparemment convaincu par les arguments de Jojo, lui propose l'arrangement suivant : si les résultats de l'entreprise continuent d'être satisfaisants, Jojo verra son salaire augmenter de 20% dès cette année. En revanche, son salaire n'augmentera pas l'année suivante. Si au contraire les résultats sont inférieurs à ceux qui étaient attendus,

le salaire de Jojo ne sera pas augmenté cette année, mais sera de toute façon accru l'année suivante de 20%. Au vu de la situation économique incertaine, Jojo, qui n'est guère optimiste, estime à 1/2 la probabilité pour que l'entreprise atteigne ses objectifs cette année.

Appelons A_1 l'augmentation relative de salaire (aléatoire) reçue par Jojo cette année, et A_2 l'augmentation relative de l'année suivante.

L'espérance de A_1 est :

$$\mathbb{E}(A_1) = \frac{1}{2} \times 1,2 + \frac{1}{2} \times 1 = 1,1.$$

Celle de A_2 se calcule de la même manière :

$$\mathbb{E}(A_2) = \frac{1}{2} \times 1,2 + \frac{1}{2} \times 1 = 1,1,$$

l'augmentation de Jojo ayant une chance sur deux de se produire cette année, et une sur deux de se produire l'année suivante. Que dire de l'augmentation totale $A_1 \times A_2$ perçue par Jojo sur les deux années ? Jojo étant certain d'être augmenté de 20% cette année ou bien l'année suivante, $A_1 \times A_2$ est toujours égal à $1,2$. En particulier, $\mathbb{E}(A_1 \times A_2) = 1,2$. En revanche, le produit des espérances $\mathbb{E}(A_1) \times \mathbb{E}(A_2)$ est égal à $1,1 \times 1,1 = 1,21$. L'espérance de $a_1 \times A_2$ n'est donc pas égale au produit des espérances de A_1 et de A_2. Ces deux variables ne sont bien entendu pas indépendantes, puisqu'une augmentation cette année entraîne une absence d'augmentation l'année suivante, et inversement.

Démontrons à présent la propriété. Considérons donc deux variables aléatoires indépendantes X et Y, dont les espérances sont bien définies. Notons S_X et S_Y les ensembles de valeurs possibles pour X et Y respectivement.

Par définition :

$$\mathbb{E}(X) \times \mathbb{E}(Y) = \left(\sum_{s \in S_X} s \times \mathbb{P}(X = s) \right) \times \left(\sum_{t \in S_Y} t \times \mathbb{P}(Y = t) \right).$$

En utilisant la distributivité de la multiplication par rapport à l'addition, nous obtenons donc que :

$$\mathbb{E}(X) \times \mathbb{E}(Y) = \sum_{s \in S_X, t \in S_Y} (s \times t) \times \mathbb{P}(X = s) \times \mathbb{P}(Y = t).$$

X et Y étant deux variables aléatoires indépendantes, cette égalité se réécrit :

$$\mathbb{E}(X) \times \mathbb{E}(Y) = \sum_{s \in S_X, t \in S_Y} (s \times t) \times \mathbb{P}(X = s \ , \ Y = t).$$

Regroupons dans la somme ci-dessus tous les couples (s,t) tels que $s \times t = u$. Leur contribution totale dans la somme ci-dessus est donc :

$$\sum_{(s,t)\ :\ s\times t=u} u \times \mathbb{P}(X = s\ ,\ Y = t) = u \times \sum_{(s,t)\ :\ s\times t=u} \mathbb{P}(X = s\ ,\ Y = t).$$

La famille d'événements «$X = s\ ,\ Y = t$», (s,t) décrivant l'ensemble des couples tels que $s \times t = u$, forme une famille d'événements deux-à-deux incompatibles, dont la réunion est l'événement «$X \times Y = u$», ou, autrement dit, une partition de cet événement. On en déduit que

$$\sum_{(s,t)\ :\ s\times t=u} \mathbb{P}(X = s\ ,\ Y = t) = \mathbb{P}(X \times Y = u).$$

Finalement, on en déduit, en considérant toutes les valeurs possibles u, que :

$$\mathbb{E}(X) \times \mathbb{E}(Y) = \sum_{u\in S_{XY}} u \times \mathbb{P}(X \times Y = u) = \mathbb{E}(X \times Y),$$

où S_{XY} désigne l'ensemble des valeurs possibles pour le produit d'un élément de S_X par un élément de S_Y.

Remarque 8 *L'argument ci-dessus ne pose aucun problème lorsque S_X et S_Y sont des ensembles finis. Lorsque ce n'est plus le cas, il est nécessaire de travailler d'abord avec $|X|$ et $|Y|$ de façon à ne manipuler que des nombres positifs, pour lesquels on est certain que l'argument ci-dessus fonctionne. On peut ensuite reprendre l'argument pour X et Y, l'étape précédente ayant établi que les séries qui interviennent sont absolument convergentes. En particulier, l'énoncé que nous venons de prouver contient l'affirmation que, si les espérances de X et Y sont bien et que X et Y sont indépendantes, l'espérance de $X \times Y$ est également définie.*

Covariance et corrélation entre deux variables aléatoires

Le comportement de la variance vis-à-vis de la somme est plus complexe que celui de l'espérance, puisque **en général, la variance d'une somme n'est pas la somme des variances**, et il en va de même pour les écarts-types, comme le montre l'exemple très simple suivant. Si X suit une loi de Bernoulli de paramètre $p \in]0,1[$, $\mathbb{V}(X + X) = \mathbb{V}(2 \times X) = 4 \times \mathbb{V}(X)$, alors que $\mathbb{V}(X) + \mathbb{V}(X) = 2 \times \mathbb{V}(X)$. $\mathbb{V}(X)$ étant égale à $\mathbb{E}(X^2) - (\mathbb{E}(X))^2 = p - p^2 = p(1-p)$, et donc différente de zéro, on a donc $\mathbb{V}(X + X) \neq \mathbb{V}(X) + \mathbb{V}(X)$. Quant aux écarts-types, considérons les deux variables aléatoires X et $-X$. $X + (-X) = 0$ et son écart-type est donc égal à zéro. En revanche, la somme des deux écart-types est égale à $2\sqrt{p(1-p)} > 0$.

Etant données deux variables aléatoires X et Y définies sur un même espace de probabilité $(\Omega, \mathbb{P})$, pour lesquelles $\mathbb{V}(X)$ et $\mathbb{V}(Y)$ sont définies, la **covariance** de X et de Y est définie par

$$cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right].$$

(L'identité entre les deux définitions s'obtient en développant les produits et en utilisant la linéarité de l'espérance. L'inégalité $|XY| \leq |X^2| + |Y^2|$ entraîne le fait que l'espérance de XY est bien définie avec nos hypothèses.)

On a la **propriété fondamentale suivante :**

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2cov(X,Y). \tag{2.3}$$

(L'inégalité $|X+Y|^2 \leq 2(|X|^2 + |Y|^2)$ entraîne le fait que la variance de $X+Y$ est bien définie avec nos hypothèses.)

Démontrons cette propriété, en partant de la formule :

$$\mathbb{V}(X+Y) = \mathbb{E}((X+Y)^2) - (\mathbb{E}(X+Y))^2.$$

Étudions séparément chaque terme. En développant le carré, et grâce à la linéarité de l'espérance, on obtient que :

$$\mathbb{E}((X+Y)^2) = \mathbb{E}(X^2 + 2XY + Y^2) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2).$$

Par ailleurs,

$$(\mathbb{E}(X+Y))^2 = \mathbb{E}(X) + \mathbb{E}(Y))^2 = \mathbb{E}(X)^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y)^2.$$

En faisant la différence entre ces deux expressions, on obtient que :

$$\begin{aligned} \mathbb{V}(X+Y) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 + \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 + 2\mathbb{E}(XY) - 2\mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathbb{E}(XY) - 2\mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

On en retient le fait que $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$ si et seulement si $cov(X,Y) = 0$. En particulier, **si X et Y sont indépendantes, la variance de leur somme est égale à la somme de leurs variances :**

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y).$$

On dit que X et Y sont corrélées lorsque $cov(X,Y) \neq 0$, et non-corrélées sinon. Comme nous l'avons déjà mentionné dans la partie précédente, le fait que X et Y ne sont pas corrélées n'entraîne pas le fait que X et Y sont indépendantes.

Lorsque $\sigma(X)$ et $\sigma(Y)$ sont non-nuls, c'est-à-dire lorsque X et Y ne sont pas égales à des constantes avec probabilité 1, on définit le **coefficient de corrélation linéaire** (souvent appelé simplement coefficient de corrélation) de X et de Y par

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)},$$

lorsque $\sigma(X)$ et $\sigma(Y)$ sont non-nuls.

On vérifie que $-1 \leq corr(X,Y) \leq 1$. Lorsque X et Y sont indépendantes, $corr(X,Y) = 0$ d'après ce qui précède.

Lorsque $corr(X,Y) = 1$ ou $corr(X,Y) = -1$, X et Y sont proportionnelles, c'est-à-dire qu'il existe un nombre réel $\lambda \neq 0$ tel que $X = \lambda Y$ avec probabilité 1, le signe de λ étant celui de $corr(X,Y)$. La dépendance entre X et Y est donc maximale.

Pour cette raison, on présente parfois le coefficient de corrélation comme une mesure normalisée (résumée par un nombre entre -1 et $+1$) de la dépendance pouvant exister entre X et Y. Cette terminologie est toutefois abusive (sauf dans le cas très particulier des vecteure gaussiens que nous étudierons ultérieurement), puisque X et Y peuvent parfaitement ne pas être indépendantes tout en possédant un coefficient de corrélation égal à zéro. En revanche, un coefficient de corrélation non-nul entre deux variables aléatoires est le signe d'une dépendance entre celles-ci.

Bien entendu, si l'on ne dispose que d'un échantillon de valeurs, il se peut que la loi empirique présente une corrélation non-nulle entre X et Y alors même que X et Y sont indépendantes sous la loi théorique de (X,Y).

Par exemple, le coefficient de corrélation associé à l'échantillon suivant de couples de valeurs :

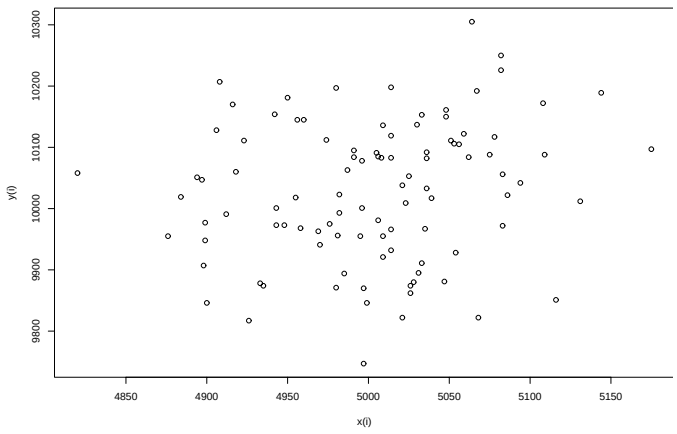
	x	y
1	0,08	0,09
2	0,93	0,58
3	5,36	0,69
4	1,02	0,53
5	0,42	0,53
6	1,00	0,31
7	1,28	0,26
8	2,86	0,95
9	3,91	0,10
10	0,01	0,44

est égal à environ $0,28$. Pourtant, ces valeurs ont été simulées à partir d'un modèle dans lequel X et Y sont indépendantes, X étant une variable aléatoire de loi exponentielle de paramètre 1 arrondie à deux décimales, et Y une variable aléatoire indépendante de X, de loi uniforme sur $[0,1]$ également arrondie à deux décimales.

La question de savoir à partir de quelle valeur un coefficient de corrélation non-nul calculé sur une loi empirique peut être considéré comme accréditant une non-indépendance dans la loi théorique n'a rien d'évident, et sera abordée dans la partie «Statistique».

Un exemple de «spurious correlation»

Intéressons-nous à l'influence possible du nombre de cheminées que compte une ville sur la natalité (il faut bien que les cigognes puissent travailler !). On pourrait imaginer de quantifier ce lien en étudiant la corrélation linéaire existant, pour un ensemble de villes, entre le nombre de naissances annuelles, et le nombre de cheminées. Cependant, on s'expose ainsi à mettre en évidence une corrélation due simplement au fait que des villes vastes et peuplées comporteront simultanément plus de cheminées et plus de naissances que des villes d'importance moindre. Simulons par exemple indépendamment 100 villes selon le modèle suivant : le nombre d'habitants Z d'une ville suit une loi de Poisson de paramètre 50000, et, conditionnellement à ce nombre d'habitants Z, le nombre de naissances X dans cette ville suit une loi de Poisson de paramètre $Z/5$, tandis que le nombre de cheminées suit, quant à lui, une loi de Poisson de paramètre $Z/10$.



Le coefficient de corrélation entre le nombre de naissances et le nombre de cheminées associé ce modèle est d'environ 0,12 (et le coefficient associé à la loi empirique de notre échantillon de 100 villes est, quant à lui, de 0,18).

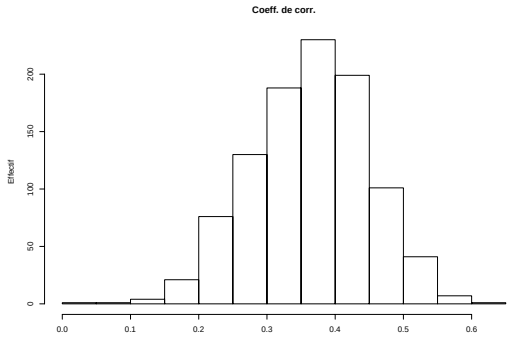
Il semble donc plus pertinent de rapporter aussi bien le nombre de naissances que le nombre de cheminées à une mesure de l'importance de la ville, telle que, par exemple, le nombre de femmes en âge de procréer.

Admettant le modèle (purement fictif, sans aucune prétention à une quelconque pertinence) suivant : pour une ville donnée, le nombre de femmes en âge de procréer

suit une loi de Poisson de paramètre $\lambda = 10000$, le nombre de cheminées suit une loi de Poisson de paramètre $\mu = 10000$, et le nombre de naissances suit une loi de Poisson de paramètre $\nu = 20000$, ces variables étant indépendantes entre elles, et indépendantes d'une ville à l'autre.

Répétons 1000 fois l'expérience consistant à simuler, pour 100 villes, le nombre de naissances, le nombre de cheminées, et le nombre de femmes en âge de procréer, puis à calculer le coefficient de corrélation linéaire entre les deux rapports : (nombre de cheminées)/(nombre de femmes en âge de procréer) et (nombre de naissances)/(nombre de femmes en âge de procréer).

Les 1000 valeurs obtenues pour le coefficient de corrélation fournissent l'histogramme suivant :



Avec ce procédé, il est donc très probable d'observer un coefficient de corrélation linéaire important entre ces deux variables. Un esprit non-averti pourrait en déduire l'existence d'une influence du nombre de cheminées sur la natalité, alors que, dans notre modèle, ces deux variables sont indépendantes ! En revanche, les deux rapports que nous avons calculé ne le sont pas ceci étant, dans notre exemple, expliqué par le fait qu'ils font intervenir la même variable au dénominateur. Rappelons que, de manière générale, l'existence d'un coefficient de corrélation non-nul n'est que le signe d'une dépendance entre variables, et ne signifie donc absolument pas qu'il y ait nécessairement une relation de cause à effet entre ces variables (nous vous renvoyons en particulier à l'article de D. Freedman cité dans la bibliographie).

2.6.10 Espérance et variance des lois usuelles

Loi de Bernoulli

Nous avons déjà calculé l'espérance et la variance d'une loi de Bernoulli pour donner des contre-exemples dans ce qui précède. Récapitulons : si X suit une loi de

Bernoulli de paramètre p,

$$\begin{cases} \mathbb{E}(X) = p, \\ \mathbb{V}(X) = p(1-p). \end{cases}$$

Loi binomiale

Nous l'avons vu, la loi binomiale de paramètres n et p intervient lorsque l'on compte le nombre aléatoire d'événements réalisés au sein d'une famille de n événements mutuellement indépendants ayant chacun la probabilité p de de produire. Considérons donc le modèle standard $(\Omega^n, \mathbb{P}^{\otimes n})$ décrivant une succession indépendante d'épreuves de Bernoulli de probabilité de succès p, et appelons $A_1, \ldots, A_n$ les événements définis par A_i =«succès à l'épreuve numéro i». Considérons leurs fonctions indicatrices

$$\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}.$$

Ces variables aléatoires prennent chacune la valeur 1 avec probabilité p, et 0 avec probabilité $1-p$. Elles suivent donc chacune une loi de Bernoulli de paramètre p. De plus, la variable X, qui compte le nombre d'événements A_i qui sont réalisés, s'exprime à l'aide de ces fonctions indicatrices :

$$X = \mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n}.$$

Grâce à la linéarité de l'espérance, on en déduit que :

$$\mathbb{E}(X) = \mathbb{E}(\mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n}) = \mathbb{E}(\mathbf{1}_{A_1}) + \cdots + \mathbb{E}(\mathbf{1}_{A_n}) = p + \cdots + p = np.$$

De plus, l'indépendance mutuelle des événements A_i entraîne que les variables aléatoires $\mathbf{1}_{A_i}$ sont mutuellement indépendantes, et par conséquent :

$$\mathbb{V}(X) = \mathbb{V}(\mathbf{1}_{A_1}+\cdots+\mathbf{1}_{A_n}) = \mathbb{V}(\mathbf{1}_{A_1})+\cdots+\mathbb{V}(\mathbf{1}_{A_n}) = p(1-p)+\cdots+p(1-p) = np(1-p).$$

Cette déduction est un peu rapide, car nous n'avons prouvé l'additivité des variances que pour une somme de deux variables aléatoires indépendantes. Pour passer à n variables, il suffit de remarquer que, par exemple, les deux variables $\mathbf{1}_{A_n}$ et $\mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_{n-1}}$ sont indépendantes, et d'itérer l'argument.

Récapitulons : si X suit une loi binomiale de paramètres n et p :

$$\begin{cases} \mathbb{E}(X) = np, \\ \mathbb{V}(X) = np(1-p). \end{cases}$$

Remarquons que nous aurions également pu, pour calculer $\mathbb{E}(X)$ et $\mathbb{V}(X)$, partir de la définition de la loi binomiale

$$\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k} \ , \ 0 \leq k \leq n,$$

et calculer

$$\mathbb{E}(X) = \sum_{k=0}^{n} k C_n^k p^k (1-p)^{n-k}$$

et

$$\mathbb{V}(X) = \sum_{k=0}^{n} k^2 C_n^k p^k (1-p)^{n-k} - (\mathbb{E}(X))^2$$

à l'aide d'identités portant sur les coefficients binomiaux.

Loi de Poisson

Nous l'avons vu, la loi de Poisson de paramètre λ apparaît comme limite de la loi binomiale de paramètres n et λ/n lorsque n tend vers l'infini. Il est donc tentant d'affirmer que l'espérance et la variance de cette loi s'obtiennent comme limites de l'espérance et de la variance associées à la loi binomiale :

$$\mathbb{E}(X) = \lim_{n \to +\infty} n \left(\frac{\lambda}{n}\right) = \lambda,$$

et

$$\mathbb{V}(X) = \lim_{n \to +\infty} n \left(\frac{\lambda}{n}\right) \left(1 - \frac{\lambda}{n}\right) = \lambda,$$

et ce raisonnement peut être rendu rigoureux, au prix d'un peu de travail supplémentaire.

Pour obtenir ces valeurs, il serait également possible de partir de la définition de la loi de Poisson

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \ , \ k \geq 0.$$

et de calculer

$$\mathbb{E}(X) = \sum_{k=0}^{+\infty} k \frac{\lambda^k}{k!} e^{-\lambda}$$

et

$$\mathbb{V}(X) = \sum_{k=0}^{n} k^2 \frac{\lambda^k}{k!} e^{-\lambda} - (\mathbb{E}(X))^2$$

à l'aide d'identités sur les séries entières.

Récapitulons : si X suit une loi de Poisson de paramètre λ :

$$\begin{cases} \mathbb{E}(X) = \lambda, \\ \mathbb{V}(X) = \lambda. \end{cases}$$

Loi géométrique

Rappelons la définition de la loi géométrique :

$$\mathbb{P}(X = k) = (1-p)^{k-1}p \ , \ k \geq 1.$$

Son espérance et sa variance se calculent à l'aide d'identités sur les séries entières :

$$\mathbb{E}(X) = \sum_{k=1}^{+\infty} k(1-p)^{k-1}p = \frac{1}{p},$$

$$\mathbb{V}(X) = \sum_{k=1}^{+\infty} k^2(1-p)^{k-1}p - (\mathbb{E}(X))^2 = \frac{1-p}{p^2}.$$

Un moyen de trouver la valeur de $\mathbb{E}(X)$ et $\mathbb{V}(X)$ est de s'appuyer sur le raisonnement heuristique suivant (ce raisonnement peut en fait être rendu complètement rigoureux, mais cela nécessite sensiblement plus d'arguments et de détails que ce qui figure ci-après, en particulier pour le calcul de la variance).

Considérons n répétitions indépendantes d'une variable aléatoire de loi Bernoulli de paramètre p, notées $Z_1, \ldots, Z_n$. Découpons à présent la séquence de 0 et de 1 constituée par $Z_1 \ldots Z_n$ en tronçons successifs de la forme $0 \ldots 01$ (le nombre de 0 précédant le 1 dans un tel tronçon pouvant être égal à 0), auquel s'ajoute éventuellement un dernier tronçon incomplet constitué uniquement de 0. Appelons N_n le nombre de tronçons dans le découpage (en n'incluant pas l'éventuel tronçon final incomplet), et $L_1, \ldots, L_{N_n}$ les longueurs respectives de de ces tronçons. Intuitivement, il est clair que, lorsque $n >> 1$, on doit avoir $L_1 + \cdots + L_{N_n} \sim n$, car la longueur de l'éventuel dernier tronçon incomplet est négligeable devant n. Par ailleurs, en négligeant le fait que notre découpage doit s'arrêter à la fin de la séquence, on peut approcher la suite $L_1, \ldots, L_{N_n}$ par une répétition indépendante de variables aléatoires de loi géométrique de paramètre p, du fait que les Z_i constituent des répétitions indépendantes d'une même variable de Bernoulli, chaque L_i comptant le nombre de répétitions à effectuer pour obtenir un 1 en partant du tronçon précédent. On s'attend par conséquent (voire à ce sujet le chapitre suivant «Loi des grands nombres») à ce que, lorsque $n >> 1$, on ait $L_1 + \cdots + L_{N_n} \sim \mathbb{E}(X) \times N_n$. En notant que N_n n'est autre que le nombre de 1 présents dans la séquence $Z_1, \ldots, Z_n$, soit $N_n = Z_1 + \cdots + Z_n$, on s'attend par ailleurs à ce que $N_n \sim np$. On en déduit que $L_1 + \cdots + L_{N_n} \sim \mathbb{E}(X) \times np$, et, en utilisant l'observation précédente selon laquelle $L_1 + \cdots + L_{N_n} \sim n$, on en déduit que l'on doit avoir $\mathbb{E}(X) = 1/p$, ce qui correspond bien à la véritable valeur.

Pour le calcul de la variance, on note que, au fait près de négliger l'éventuel tronçon final incomplet, on doit avoir $N_n - np \sim \sum_{j=1}^{N_n}(1 - L_j p)$. En traitant les variables L_j comme des variables indépendantes de loi géométrique de paramètre p,

et en remplaçant N_n par np, on obtient que $\mathbb{V}(N_n - np) \sim np \times \mathbb{V}(1 - Xp)$. En utilisant le fait que N_n suit une loi binomiale de paramètres n et p, d'où le fait que $\mathbb{V}(N_n - np) = np(1-p)$, et que $\mathbb{V}(1 - Xp) = p^2\mathbb{V}(X)$, on en déduit que l'on doit avoir $np(1-p) \sim np \times p^2\mathbb{V}(X)$, d'où le fait que $\mathbb{V}(X) = \frac{1-p}{p^2}$.

Loi exponentielle

Si X suit une loi exponentielle de paramètre λ,

$$\begin{cases} \mathbb{E}(X) = \frac{1}{\lambda}, \\ \mathbb{V}(X) = \frac{1}{\lambda^2}. \end{cases}$$

Ceci peut se voir en calculant les intégrales $\int_0^{+\infty} t\exp(-\lambda t)dt$ et $\int_0^{+\infty} t^2\exp(-\lambda t)dt$, ou en utilisant l'approximation par une loi géométrique de paramètre λ/n renormalisée par n.

On note que l'espérance s'identifie ici au paramètre d'échelle.

Loi gaussienne

Si X suit une loi gaussienne de paramètre m et v,

$$\begin{cases} \mathbb{E}(X) = m, \\ \mathbb{V}(X) = v. \end{cases}$$

Ceci peut se voir en calculant les intégrales correspondantes (voir le chapitre «Courbe en cloche»). L'espérance s'identifie donc au paramètre de position, et la variance au paramètre d'échelle.

Loi gamma

En vertu de l'exercice 128, si X suit une loi gamma de paramètres a et s avec a entier, l'espérance de X doit être égale à as et la variance à as^2. C'est également vrai si a n'est pas un nombre entier, comme on peut le voir en calculant les intégrales correspondantes.

$$\begin{cases} \mathbb{E}(X) = a, \\ \mathbb{V}(X) = as^2. \end{cases}$$

Loi beta

Si X suit une loi beta de paramètres a et b,

$$\begin{cases} \mathbb{E}(X) = \frac{a}{a+b}, \\ \mathbb{V}(X) = \frac{ab}{(a+b)^2(a+b+1)}. \end{cases}$$

comme on peut le vérifier en calculant les intégrales correspondantes. (Voir également l'exercice 134).

Loi de Cauchy

La loi de Cauchy est l'exemple le plus classique de loi pour lesquelles l'espérance n'est pas définie (et, par voie de conséquence, la variance ne l'est pas non plus).

En effet, on voit facilement que $\int_{-\infty}^{+\infty} \frac{x}{1+x^2} dx = +\infty$.

Cette loi intervient pourtant dans diverses situations de modélisation. Le fait que l'espérance d'une variable aléatoire puisse ne pas exister n'est pas qu'une vue de l'esprit !

2.6.11 Régression linéaire

De manière générale, le problème de la régression se pose de la manière suivante : à partir de la connaissance de la valeur prise par une variable aléatoire X, proposer une approximation de la valeur prise par une autre variable aléatoire Y. En d'autres termes, on cherche une fonction h telle que $h(X)$ représente une approximation de la valeur de Y. La différence $Y - h(X)$ est généralement appelée le résidu de la régression.

On cherche naturellement à ce que le résidu soit le plus faible possible, au sens d'un critère qui doit être précisé.

Ce type de problème intervient dans de très nombreuses applications. Par exemple, prédire de la meilleure façon possible la taille d'un garçon à l'âge adulte (Y) en fonction de la taille de son père (X), ou encore, estimer la valeur d'une quantité physique (Y) à partir d'une mesure indirecte et bruitée de cette quantité (X). Autre exemple, X pourra représenter une mesure de la concentration de certains marqueurs biologiques dans le sang d'un patient, tandis que Y représente le degré de gravité de l'atteinte de celui-ci (à estimer au mieux sur la base des mesures). Encore un exemple : X représente l'image numérisée (sous forme d'une grille de pixels) d'une lettre manuscrite, et Y représente ladite lettre (A,B,C,...), et l'on cherche à automatiquement retrouver Y à partir de X. De fait, d'innombrables autres problèmes concrets peuvent se mettre sous la forme de problèmes de régression. Nous ne discuterons ici que le cas très particulier où X et Y sont deux variables aléatoires à valeurs réelles.

Une première étape indispensable est de définir précisément la manière dont on mesure l'écart entre l'approximation proposée $h(X)$, et la véritable valeur Y, différentes manières de mesurer cet écart menant en général à différentes notions de ce qu'est la «meilleure» approximation de Y par une fonction de X.

Un choix fréquent est l'écart quadratique moyen : $\mathbb{E}\left[(Y - h(X))^2\right]$. Bien entendu, ce choix n'est pas le seul possible, et présente un certain nombre d'avantages et d'inconvénients – la mesure de l'erreur par ce critère est donc discutable, et cette discussion rejoint celle sur la pertinence de l'espérance en tant qu'indicateur de position (voir ce qui a été dit précédemment à ce sujet). Dans le cadre fréquentiel, ce

critère fournit un contrôle sur la somme des erreurs quadratiques commises. L'inégalité de Markov assure au moins qu'une faible valeur de l'écart en ce sens conduit à un écart typiquement faible.

Ce choix étant fixé, le problème de la régression est donc de trouver une fonction h qui minimise la quantité $\mathbb{E}\left[(Y-h(X))^2\right]$. On parle alors de régression au sens des moindres carrés. Une solution théorique à ce problème de minimisation est fournie par le raisonnement suivant.

Dans notre contexte, notons que l'on peut écrire, dans le cas d'une variable aléatoire X discrète dont S_X est l'ensemble des valeurs :

$$\mathbb{E}\left[(Y-h(X))^2\right] = \sum_{s\in S_X} \mathbb{E}\left[(Y-h(x))^2|X=x\right]\mathbb{P}(X=x),$$

où $\mathbb{E}(\cdots|X=x)$ désigne l'espérance par rapport à la probabilité $\mathbb{P}(\cdots|X=x)$ (Voir la partie sur l'espérance conditionnelle pour plus de détails).

Dans le cas d'une variable aléatoire X continue et possédant la densité f, on peut encore écrire

$$\mathbb{E}\left[(Y-h(X))^2\right] = \int_{-\infty}^{+\infty} \mathbb{E}\left[(Y-h(x))^2|X=x\right] f(x)dx,$$

et nous vous renvoyons aux remarques sur le conditionnement par une variable aléatoire continue effectuées plus bas pour une discussion des problèmes techniques soulevés par cette situation.

S'il n'existe aucune contrainte liant entre elles les valeurs de $h(x)$ pour différentes valeurs de x (telles que, par exemple, des contraintes de continuité) – ce qui est le cas lorsque l'on cherche une régression sous la forme $h(X)$, où h est la fonction la plus générale possible, il suffit de minimiser séparément pour chaque valeur de x la quantité $\mathbb{E}\left[(Y-f(x))^2|X=x\right]$. L'exercice 94 entraîne que le minimum est atteint en choisissant

$$h(x) := \mathbb{E}\left[Y|X=x\right].$$

Notons que la variable aléatoire $h(X)$ n'est autre que l'espérance conditionnelle $\mathbb{E}(Y|X)$, notion étudiée en tant que telle dans une autre partie.

Pour être simple à définir, cette solution au problème de la régression n'est en général que théorique, car, entre autres, de redoutables problèmes d'estimation se posent lorsque l'on cherche concrètement, à partir de listes de valeurs mesurées $(x_i, y_i)_{i=1,\ldots,n}$ du couple de variables (X,Y), à estimer la fonction h définie ci-dessus.

Nous allons dans cette partie nous intéresser à une version restreinte du problème : rechercher la meilleure approximation de Y non pas par une variable aléatoire de la forme $h(X)$, où h peut-être une fonction quelconque (ou presque), mais en nous restreignant aux fonctions affines, c'est-à-dire de la forme $h(x)=ax+b$. Nous serons donc amenés à chercher les réels a et b qui minimisent la quantité $\mathbb{E}([Y-(aX+b)]^2)$.

On parle dans ce cas de **régression linéaire**, pour insister sur le fait que les fonctions h considérées sont linéaires (en fait, affines).

Un problème de ce type d'approche est qu'en général, même en choisissant a et b de manière optimale, l'approximation $aX + b$ de Y est différente de $\mathbb{E}(Y|X)$. Autrement dit, notre approximation n'est pas la meilleure au sens des moindres carrés. En revanche, ce choix conduit à des problèmes d'estimation faciles à résoudre, et résulte donc d'un compromis entre précision de l'approximation fournie par la régression, et possibilité de calculer concrètement (et pas seulement de manière théorique) celle-ci. Le rôle privilégié de la régression linéaire dans les modèles gaussiens (où elle coïncide effectivement avec la régression optimale au sens des moindres carrés $\mathbb{E}(Y|X)$, nous en reparlerons dans le chapitre sur la courbe en cloche) est une autre raison de l'importance de ce type de régression.

De nombreuses méthodes plus élaborées que la régression linéaire (tels que splines, réseaux de neurones, arbres de décision,...), et réalisant des compromis différents – et plus ou moins bien adaptés aux différents contextes – existent, et sont devenus utilisables ces dernières années notamment grâce à l'accroissement de la puissance de calcul des ordinateurs. Pour en apprendre (beaucoup) plus sur le sujet, vous pouvez consulter par exemple l'ouvrage de Hastie, Tibshirani et Friedman cité dans la bibliographie.

Expliquons maintenant comment calculer les coefficients de la régression linéaire de Y sur X, c'est-à-dire les réels a et b qui minimisent la quantité $\mathbb{E}([Y-(aX+b)]^2)$.

On vérifie qu'une manière équivalente de poser le problème consiste à chercher une écriture de Y sous la forme $Y = aX+b+W$, où W vérifie $\mathbb{E}(W) = 0$ et $cov(W, X) = 0$. Ou encore, à chercher à écrire Y sous la forme $\alpha(X - \mathbb{E}(X)) + \mathbb{E}(Y) + W$, où W vérifie $\mathbb{E}(W) = 0$ et $cov(W, X) = 0$, soit une somme d'un terme constant ($\mathbb{E}(Y)$), un terme proportionnel à l'écart entre X et $\mathbb{E}(X)$, et un terme résiduel centré et non-corrélé à X.

Remarque 9 *Dans le cas où la loi du couple (X, Y) est la loi empirique associée à un échantillon de valeurs $(x_1, y_1), \ldots, (x_N, y_N)$, on vérifie que le problème revient à chercher la droite d'approximation des moindres carrés du nuage de points du plan formé par $(x_1, y_1), \ldots, (x_N, y_N)$, donnée par son équation $y = ax + b$.*

Si l'on suppose que $\mathbb{V}(X) \neq 0$, on obtient que la meilleure approximation est obtenue avec $a = a_{X,Y} := cov(X, Y)(\mathbb{V}(X))^{-1}$ et $b = b_{X,Y} := \mathbb{E}(Y) - a\mathbb{E}(X)$. En d'autres termes, l'approximation obtenue est $\left(cov(X, Y)(\mathbb{V}(X))^{-1}\right)(X - \mathbb{E}(X)) + \mathbb{E}(Y)$. Cette variable aléatoire est appelée la régression linéaire de Y sur X. La différence $\epsilon_{X,Y} := a_{X,Y}X - b_{X,Y}$ est appelée le résidu de la régression, et l'on vérifie que $cov(\epsilon_{X,Y}, X) = 0$.

On a alors $\mathbb{V}(Y) = \mathbb{V}(\epsilon_{X,Y}) + \mathbb{V}(a_{X,Y}X + b_{X,Y})$. La quantité $\mathbb{V}(\epsilon_{X,Y})$ est appelée la variance résiduelle de la régression. La quantité $\mathbb{V}(a_{X,Y}X + b_{X,Y})$ est, quant à

elle, souvent appelée la variance expliquée par la régression, car elle apparaît dans l'expression ci-dessus comme la part de la variance de Y qui est «expliquée» par la variance de X dans le modèle de régression linéaire ; toutefois, cette terminologie peut prêter à confusion, et il faut se garder (comme de la peste) de confondre régression (linéaire ou non) et explication (voir par exemple l'article de D. Freedman cité dans la bibliographie), de même qu'une simple association entre événements ne permet pas de conclure à l'existence d'un lien de cause à effet entre ceux-ci.

On vérifie que $\mathbb{V}(\epsilon_{X,Y}) = (1 - corr(X,Y)^2)\mathbb{V}(Y)$ tandis que $\mathbb{V}(a_{X,Y}X + b_{X,Y}) = corr(X,Y)^2\mathbb{V}(Y)$, et l'on en déduit donc que le coefficient de corrélation r fournit une mesure de la précision de la régression linéaire de Y sur X.

Voici à présent une version «normalisée» de la régression de X sur Y, dans laquelle ces variables sont ramenées sur une échelle où leur espérance est nulle et leur écart-type égale à 1 [10].

Lorsque $\sigma(X)$ et $\sigma(Y)$ sont non-nuls, on définit $\tilde{X} = (X - \mathbb{E}(X))(\sigma(X))^{-1}$ et $\tilde{Y} = (Y - \mathbb{E}(Y))(\sigma(Y))^{-1}$.

Le coefficient de corrélation $corr(X,Y)$ est alors égal au coefficient a de la régression linéaire de la variable aléatoire $\tilde{Y}$ sur $\tilde{X}$. On vérifie que l'erreur d'approximation

$$\mathbb{E}\left[\left(\tilde{Y} - corr(X,Y)\tilde{X}\right)^2\right],$$

est alors égale à $1 - corr(X,Y)^2$, et le coefficient de corrélation fournit donc une mesure de la qualité de la régression linéaire de $\tilde{Y}$ sur $\tilde{X}$.

Les résultats mentionnés ci-dessus sans preuve peuvent être soient prouvés de manière élémentaire (en développant tous les carrés et en analysant les variations des fonctions d'une ou de deux variables obtenues), ou à partir de l'interprétation géométrique présentée plus bas. Voir l'exercice 96).

Comme nous l'avons mentionné, un exemple de situation dans laquelle le régression linéaire intervient est celui où l'on dispose d'une mesure de la variable X, et où l'on cherche à prédire le mieux possible la valeur de Y par une fonction affine de X, à partir d'une connaissance de la loi jointe de (X,Y), qui peut par exemple être obtenue à partir d'un échantillon de valeurs mesurées du couple (X,Y), de la forme $(x_1,y_1),\ldots,(x_N,y_N)$ [11].

Mise en garde 12 *Il importe de ne pas confondre la question de la régression linéaire, dans lequel la prédiction est effectuée avec une fonction affine, avec la question*

10. Nous aurons l'occasion de discuter à nouveau de cette normalisation dans le chapitre «Courbe en cloche»

11. Au passage, insistons sur le fait que la problème de la régression porte sur le couple de variables aléatoires (X,Y). Les échantillons de données mesurées que l'on utilisera seront donc de la forme $(x_i,y_i)_{i=1,\ldots,N}$, dans lesquels l'appariement entre la valeur x_i et la valeur y_i (et non pas avec une autre valeur y_j de la liste) est extrêmement important.

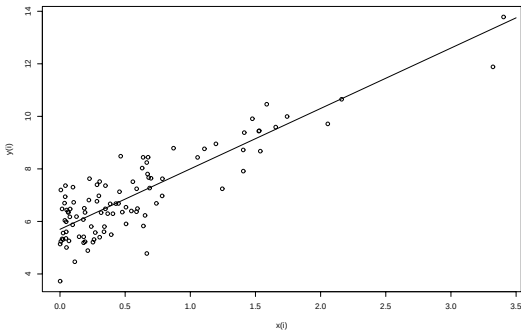
plus générale consistant à chercher la meilleure approximation de Y sous la forme d'une fonction quelconque de X (pas nécessairement affine), et dont la réponse est fournie par l'espérance conditionnelle, étudiée dans une autre partie.

Comme nous l'avons mentionné, l'un des intérêts de la régression linéaire est que les paramètres a et b peuvent être facilement (c'est-à-dire au moyen de calculs relativement peu coûteux) et en général efficacement (c'est-à-dire avec une précision raisonnable lorsque l'on suppose de données en nombre raisonnable, voir le chapitre «Statistique» pour une introduction à ce type de questions) estimés à partir d'un échantillon $(x_1, y_1), \ldots, (x_N, y_N)$ de valeurs mesurées du couple (X, Y).

D'autre part, la régression linéaire joue un rôle privilégié dans les modèles gaussiens, sur lesquels nous reviendrons dans le chapitre sur la courbe en cloche.

Une situation particulièrement confortable pour la régression linéaire est celle où Y peut effectivement se mettre sous la forme $Y = aX + b + W$, où W est centrée, possède une variance, et est indépendante de X. En effet, dans ce cas, les coefficients a et b sont nécessairement ceux de la régression linéaire de Y sur X, et $aX + b$ constitue la meilleure estimation possible de Y par une fonction quelconque de X, au sens des moindres carrés, autrement dit, en anticipant quelque peu, $aX + b$ est l'espérance conditionnelle de Y sachant X.

Afin d'illustrer un peu cette situation, voici un exemple du nuage de points obtenus en générant un échantillon de 100 valeurs (x_i, y_i) selon le modèle $Y = 2,3 \times X + 5,7 + W$, où X suit une loi exponentielle de paramètre 1, et W est indépendante de X et suit une loi gaussienne de paramètres $m = 0$ et $v = 0,64$. Sur le nuage de points, nous avons également tracé la droite d'équation $y = 2,3x + 5,7$ (dans ce cas, nous connaissons à l'avance ces coefficients, et nous ne nous posons pas, pour l'instant, la question de leur estimation à partir des données).

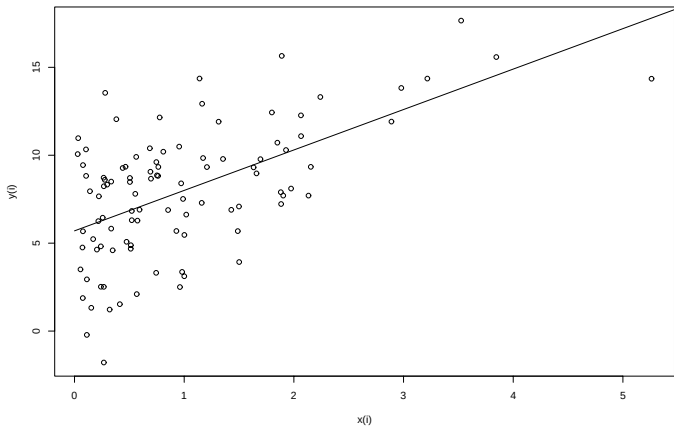


On constate que les abscisses des points ne sont pas uniformément réparties, ce qui est normal et ne fait que refléter le fait que la distribution des abscisses n'est pas de loi uniforme, mais de loi exponentielle de paramètre 1. On constate des écarts

aléatoires entre la droite tracée (appelée droite de régression) et les ordonnées des points, et l'on observe grossièrement le caractère symétrique de leur loi de probabilité.

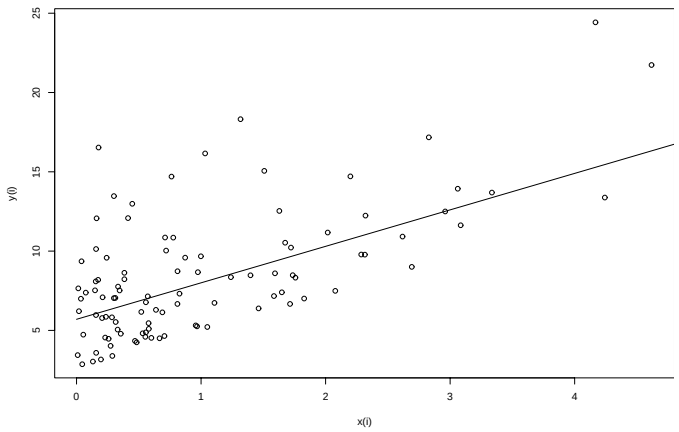
Modifions la valeur de v en la portant à $v = 9$.

Nous obtenons le graphique suivant. On observe que la prédiction de $y(i)$ fournie par $ax(i)+b$ est en général moins précise que dans le cas précédent. Cette prédiction est néanmoins la meilleure possible au sens des moindres carrés, et c'est la dispersion plus importante des valeurs de W qui limite la qualité des prédictions qu'il est possible d'effectuer à partir de la seule valeur de X.



A présent, choisissons W de la forme $W = V - 3$, où V suit une loi exponentielle de paramètre $1/3$ et est indépendante de X.

Nous obtenons le graphique suivant.

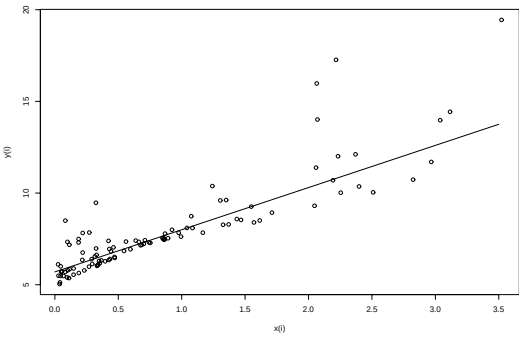


Les écarts ne sont plus symétriques, mais demeurent centrés, les valeurs positives plus rares et plus grandes compensant les valeurs négatives plus fréquentes et plus

faibles en valeur absolue.

La situation se corse à peine si l'on autorise W à dépendre de X, mais en restant centré conditionnellement à la valeur de X, autrement dit : pour tout x, l'espérance de W sachant que $X = x$ est encore égale à zéro. Dans ce cas, les coefficients a et b sont encore ceux de la régression linéaire de Y sur X, et $aX + b$ constitue encore la meilleure estimation possible de Y par une fonction quelconque de X, au sens des moindres carrés, autrement dit, en anticipant quelque peu, $aX + b$ est l'espérance conditionnelle de Y sachant X.

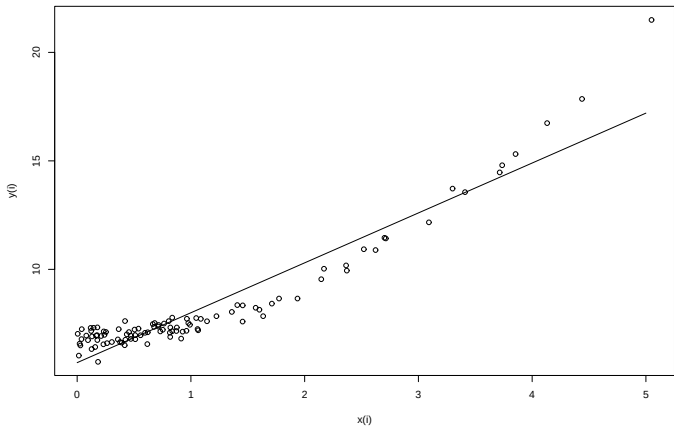
Dans l'exemple ci-dessous, la loi de W sachant que $X = x$ est une loi exponentielle de paramètre $\lambda = (|x - 0,7| + 0,1)^{-1}$ translatée par son espérance, de manière à vérifier le fait que l'espérance de W sachant que $X = x$ est nulle.



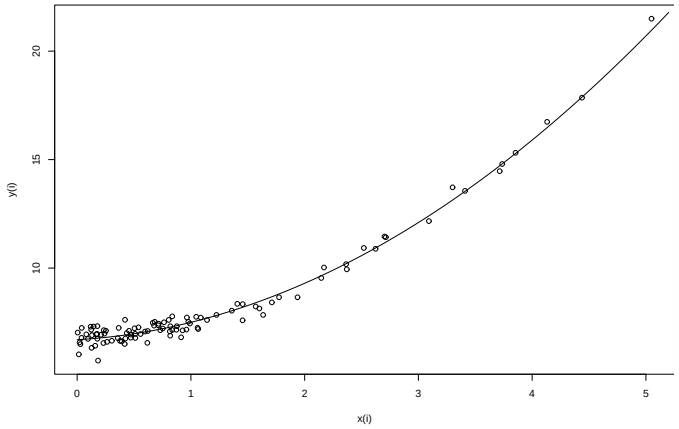
On observe effectivement des variations de la dispersion des écarts en fonction de la valeur de X.

A présent, donnons un exemple où la régression linéaire n'est plus une méthode aussi satisfaisante.

On définit $W = X^2/2 - 2X + 1 + \epsilon$, où ϵ est une variable aléatoire indépendante de X de loi gaussienne de paramètres $m = 0$ et $v = 0,55$. Clairement, W n'est pas indépendante de X, et, pire, l'espérance de W sachant que $X = x$ est égale à $x^2/2 - 2x + 1$ et n'est donc pas nulle en général, si bien que $aX + b$ n'est pas la meilleure estimation possible de Y par une fonction de X au sens des moindres carrés. Pourtant, a et b sont encore les coefficients de la régression linéaire de Y sur X. Voici le gaphique obtenu.



En fait, on constate facilement que la meilleure estimation possible de Y par X au sens des moindres carrés est donnée par la fonction quadratique $g(x) = 2,3x + 5,7 + x^2/2 - 2x + 1$. Voici la superposition de cette courbe au nuage de points précédent.



Sur cet exemple, les choses sont relativement claires, et une simple observation des données suffit à suggérer qu'un procédé de régression quadratique (ou tout au moins autre que linéaire) est plus approprié. Dans des cas plus complexes, soit que l'on ne dispose pas de suffisament de données pour se faire une idée précise de la loi jointe de (X, Y), soit que les objets manipulés ne soient pas simplement des variables réelles unidimensionnelles, mais des objets de nature plus élaborée, il est difficile, voire impossible, de déterminer la manière optimale de prédire Y à partir de X, et l'on se restreint souvent à l'utilisation de certains types de procédés de régression, dont la régression linéaire est certainement le plus simple à tout point de

vue.

Pour en revenir à la régression linéaire, tout en sachant qu'elle ne constitue pas en général le moyen d'obtenir la meilleure estimation, on note que le coefficient de corrélation fournit une estimation de l'erreur commise (plus précisément, de sa variance). Toutefois, une même erreur d'estimation peut recouvrir des situations très différentes. De même que la démarche consistant simplement à calculer l'espérance et l'écart-type d'une variable aléatoire à valeurs réelles et à considérer que l'on obtient ainsi l'essentiel des informations sur la loi de cette variable aléatoire est une démarche catastrophique (sauf lorsque l'on dispose d'informations spécifiques sur la loi en question, telle que, par exemple, son appartenance à une famille paramétrique de lois telles les gaussiennes), la démarche consistant, en présence d'un couple de variables aléatoires à valeurs réelles (X, Y), à calculer simplement l'espérance et l'écart-type de X et de Y, ainsi que le coefficient de corrélation, est elle aussi catastrophique (sauf, là encore, lorsque l'on dispose d'informations spécifiques sur la loi du couple). Il est indispensable de procéder à une analyse plus détaillée, par exemple au moyen d'autres indicateurs et d'outils de visualisation.

En voici une illustration classique, due à Anscombe (Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21).

Supposons donc que l'on dispose d'un échantillon de 11 mesures portant simultanément sur huit caractères quantitatifs $X_1, \ldots, X_4$ et $Y_1, \ldots, Y_4$.

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8,04	9,14	7,46	6,58
2	8	8	8	8	6,95	8,14	6,77	5,76
3	13	13	13	8	7,58	8,74	12,74	7,71
4	9	9	9	8	8,81	8,77	7,11	8,84
5	11	11	11	8	8,33	9,26	7,81	8,47
6	14	14	14	8	9,96	8,10	8,84	7,04
7	6	6	6	8	7,24	6,13	6,08	5,25
8	4	4	4	19	4,26	3,10	5,39	12,50
9	12	12	12	8	10,84	9,13	8,15	5,56
10	7	7	7	8	4,82	7,26	6,42	7,91
11	5	5	5	8	5,68	4,74	5,73	6,89

En considérant la loi empirique de la variable aléatoire $(X_1, \ldots, X_4, Y_1, \ldots, Y_4)$, on obtient les valeurs numériques suivantes (la deuxième décimale significative est arrondie).

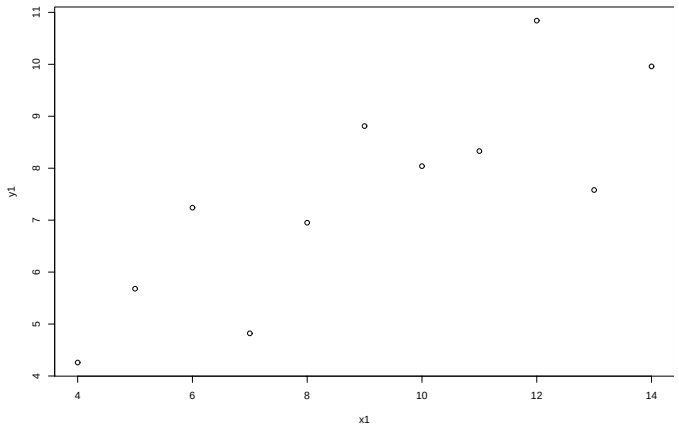
1. $\mathbb{E}_{emp.}(X_1) = 9, \mathbb{E}_{emp.}(Y_1) = 7,5, \mathbb{V}_{emp.}(X_1) = 11, \mathbb{V}_{emp.}(Y_1) = 4,13, corr_{emp.}(X_1, Y_1) = 0,82$;

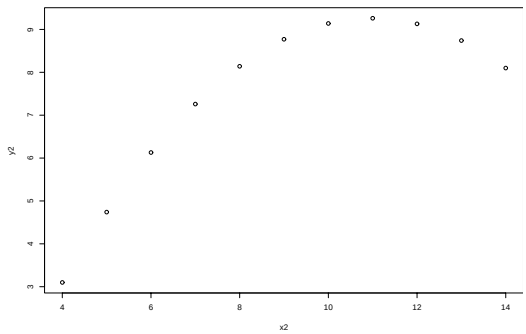
2. $\mathbb{E}_{emp.}(X_2) = 9, \mathbb{E}_{emp.}(Y_2) = 7,5, \mathbb{V}_{emp.}(X_2) = 11, \mathbb{V}_{emp.}(Y_2) = 4,13, corr_{emp.}(X_2, Y_2) = 0,82$;

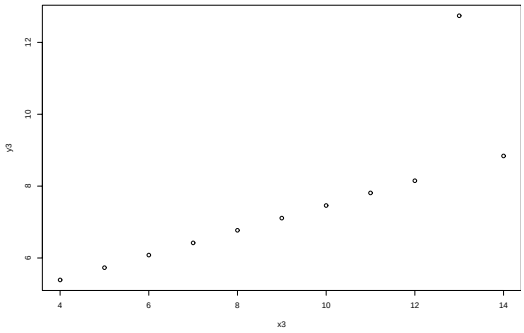
3. $\mathbb{E}_{emp.}(X_3) = 9, \mathbb{E}_{emp.}(Y_3) = 7,5, \mathbb{V}_{emp.}(X_3) = 11, \mathbb{V}_{emp.}(Y_3) = 4,12, corr_{emp.}(X_3, Y_3) = 0,82$;

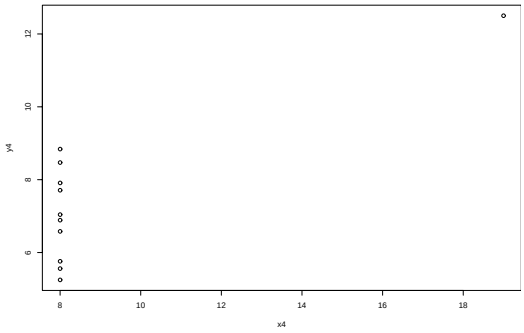
4. $\mathbb{E}_{emp.}(X_4) = 9, \mathbb{E}_{emp.}(Y_4) = 7,5, \mathbb{V}_{emp.}(X_4) = 11, \mathbb{V}_{emp.}(Y_4) = 4,12, corr_{emp.}(X_4, Y_4) = 0,82$;

Ces indicateurs ne font donc pas apparaître de différence entre les quatre paires de variables (X_i, Y_i) pour i, pour lesquelles une corrélation élevée. En revanche, les graphiques suivants, qui représentent y_i en fonction de x_i pour $i = 1, 2, 3, 4$, font clairement apparaître des différences qualitatives fondamentales entre ces variables, en particulier (mais pas seulement) dans la dépendance pouvant exister entre deux membres d'une même paire.









Le premier graphique évoque ceux obtenus dans le cas où Y s'écrit sous la forme d'une fonction linéaire de X à laquelle s'ajoute un bruit aléatoire centré indépendant de X.

Le deuxième suggère très fortement que Y_2 doit s'exprimer de manière déterministe en fonction de Y_2, mais comme une fonction non-linéaire. La dépendance est donc bien plus forte que ce que le coefficient de corrélation laisse supposer, et n'a rien de linéaire.

Le troisième suggère également très fortement le fait que Y_3 doit s'exprimer exactement comme une fonction linéaire de X_3 différente de celle calculée par régression linéaire, et qu'un point aberrant affecte les mesures (la mesure à l'origine de ce point devant certainement être réexaminée de manière critique [12]).

Le dernier nous rappelle que X_4 est constante, à l'exception d'une unique valeur, la distribution de X_4 semblant relativement uniforme entre 5 et 9.

Nous nous contentons ici d'une discussion très rapide et informelle, davantage devant être dit sur ce type de question dans le chapitre «Statistique». Le point essentiel est de noter les très fortes différences entre ces situations, qui donnent pourtant lieu à des indicateurs d'espérance/variance/covariance totalement identiques.

12. La question des points aberrants sera reprise dans le chapitre «Statistique»

Interprétation géométrique : théorie $\mathcal{L}^2$ *

Cette partie est principalement destinée aux lecteurs ayant déjà une connaissance au moins rudimentaire de la théorie des espaces euclidiens, voire hilbertiens.

Les résultats qui précèdent ont une interprétation géométrique très simple.

En appelant $\mathcal{L}^2(\Omega, \mathbb{P})$ l'ensemble des variables aléatoires définies sur $(\Omega, \mathbb{P})$ à valeurs réelles possédant une variance, on vérifie que $\mathcal{L}^2(\Omega, \mathbb{P})$ est un espace vectoriel vis-à-vis des opérations d'addition des variables aléatoires et de leur multiplication par un scalaire, et que l'application $X \mapsto \mathbb{E}(X^2)$, définit (le carré d') une norme euclidienne $|| \cdot ||$ sur $\mathcal{L}^2(\Omega, \mathbb{P})$, dont le produit scalaire est donné par $< X, Y >= \mathbb{E}(XY)$.

Appelons $\mathcal{C}$ le sous-espace vectoriel de $\mathcal{L}^2(\Omega, \mathbb{P})$ formé par les fonctions constantes, et $\mathcal{L}^2_0(\Omega, \mathbb{P})$ le sous-espace vectoriel de $\mathcal{L}^2(\Omega, \mathbb{P})$ formé par les variables aléatoires dont l'espérance est nulle.

On vérifie immédiatement que $\mathcal{L}^2(\Omega, \mathbb{P}) = \mathcal{L}^2_0(\Omega, \mathbb{P}) \overset{\perp}{\oplus} \mathcal{C}$.

On vérifie que l'espérance de X n'est autre (voir l'exercice 94) que la projection orthogonale de X sur $\mathcal{C}$.

Par conséquent, le centrage de X, c'est-à-dire la transformation $\Phi : X \mapsto X - \mathbb{E}(X)$ n'est autre que la projection de X sur $\mathcal{L}^2_0(\Omega, \mathbb{P})$. On voit ainsi que $\mathbb{V}(X) = ||\Phi(X)||^2$ et $cov(X, Y) =< \Phi(X), \Phi(Y) >$.

Dans ce cadre, l'équation 2.3 n'est autre que la reformulation de la formule bien connue sur la norme euclidienne d'une somme de deux vecteurs : $||\Phi(X + Y)||^2 = ||\Phi(X)||^2 + ||\Phi(Y)||^2 + 2 < \Phi(X), \Phi(Y) >$.

L'indépendance de X et de Y entraîne le fait que $\Phi(X)$ et $\Phi(Y)$ sont orthogonales, la réciproque étant fausse en général.

Le problème de la régression linéaire de Y sur X s'interprète alors simplement comme celui de la recherche de la projection orthogonale de Y sur le sous-espace constitué par les vecteurs de la forme $aX + b,\ (a, b) \in \mathbb{R}^2$.

La normalisation de X par $\tilde{X}$ revient simplement à normaliser le vecteur $\Phi(X)$, c'est-à-dire à le diviser par sa norme.

Dans ce contexte, le coefficient de corrélation de X et de Y se présente comme le cosinus de l'angle entre les deux vecteurs $\Phi(X)$ et $\Phi(Y)$.

2.7 Probabilité, loi et espérance conditionnelles

Dans cette partie, nous nous restreindrons au cas discret pour des raisons de simplicité d'exposition. Lorsqu'elles peuvent être définies, les notions analogues dans le cas continu se déduisent facilement de ce qui est présenté ici, à partir des transformations usuelles permettant de passer du cas discret au cas continu (voir également la partie suivante).

Considérons un modèle probabiliste $(\Omega, \mathbb{P})$. Etant donné un événement $A \subset \Omega$ tel que $\mathbb{P}(A) > 0$, nous avons défini au chapitre précédent la probabilité $\mathbb{P}$ conditionnelle à A, notée $\mathbb{P}(\cdot|A)$.

Nous utiliserons également la notion d'espérance conditionnelle à un événement : si X est une variable aléatoire définie sur $(\Omega, \mathbb{P})$ à valeurs réelles et possédant une espérance, on notera $\mathbb{E}_{\mathbb{P}}(X|A)$, ou encore $\mathbb{E}(X|A)$ quand il n'y a pas d'ambiguïté, l'espérance de X calculée non pas à partir de la probabilité $\mathbb{P}$, mais à partir de la probabilité $\mathbb{P}(\cdot|A)$. Autrement dit, $\mathbb{E}(X|A) := \mathbb{E}_{\mathbb{P}(\cdot|A)}(X)$. Plus explicitement, dans le cas discret, $\mathbb{E}(X|A) = \sum_{\omega\in\Omega} X(\omega) \times \mathbb{P}(\omega|A) = \sum_{s\in S_X} s \times \mathbb{P}(X = s|A)$.

On déduit facilement de cette définition que

$$\mathbb{E}(X|A) = \frac{\mathbb{E}(X\mathbf{1}(A))}{\mathbb{P}(A)}.$$

Supposons maintenant que nous disposions d'un système complet d'événements $\mathcal{A} = (A_1, \ldots, A_p)$, et que nous soyons capable de déterminer non pas par lequel des éléments ω de Ω la situation se réalise, mais simplement lequel des événements de $\mathcal{A}$ est réalisé. (On ne localise donc pas parfaitement ω, mais l'unique événement A_i auquel il appartient).

Lorsque l'on sait que A_i est réalisé, nous avons vu qu'il convient de décrire la situation à l'aide du modèle modifié $(\Omega, \mathbb{P}(\cdot|A_i))$.

Ainsi, suivant l'événement de $\mathcal{A}$ qui se réalise, on est amené à décrire la situation à l'aide d'une probabilité différente de $\mathbb{P}$, et qui dépend de A_i. En ce sens, la probabilité avec laquelle il convient de décrire la situation étudiée en tenant compte de l'événement de $\mathcal{A}$ qui s'est réalisé est elle-même une variable aléatoire, puisqu'elle varie en fonction de celui des événements de $\mathcal{A}$ qui s'est effectivement réalisé.

Ceci justifie la définition générale suivante.

On appelle **probabilité $\mathbb{P}$ conditionnelle à $\mathcal{A}$** la variable aléatoire définie sur Ω et à valeurs dans l'ensemble des probabilités sur Ω définie par l'équation suivante :

$$\text{pour tout } 1 \leq i \leq p \text{ et tout } \omega \in A_i,\ \mathbb{P}(\cdot|\mathcal{A}) = \mathbb{P}(\cdot|A_i).$$

Vous ne rêvez donc pas, il s'agit bel et bien d'une probabilité aléatoire sur Ω. A tout ω est associé une probabilité sur Ω, égale à la probabilité conditionnelle à l'unique événement de $\mathcal{A}$ contenant ω (l'existence et l'unicité d'un tel événement provient du fait que $\mathcal{A}$ forme un système complet d'événements).

Un (léger) problème de définition provient du fait que $\mathbb{P}(\cdot|A_i)$ n'est pas définie lorsque $\mathbb{P}(A_i) = 0$. Un choix arbitraire d'une probabilité sur Ω (par exemple $\mathbb{P}$) dans le cas où $\mathbb{P}(A_i) = 0$ permet de définir complètement $\mathbb{P}(\cdot|\mathcal{A})$. Ce choix arbitraire n'a aucune importance en pratique, puisque, si $\mathbb{P}(A_i) = 0$, les éléments $\omega \in A_i$ ne correspondent jamais à une issue réalisée de la situation considérée.

Dans le cas où l'on dispose d'une représentation en arbre de Ω et où $\mathcal{A}$ est formé par des événements associés à des nœuds de l'arbre, la probabilité conditionnelle consiste simplement à associer à toute feuille de l'arbre sa probabilité conditionnelle à l'unique nœud de $\mathcal{A}$ dont elle descend. Comme cette probabilité dépend du nœud considéré, elle apparaît naturellement comme une variable aléatoire.

On observe que la relation

$$\sum_{i=1}^{p} \mathbb{P}(B) = \sum_{i=1}^{p} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

entraîne le fait que, pour tout événement B :

$$\mathbb{E}_{\mathbb{P}}(\mathbb{P}(B|\mathcal{A})) = \mathbb{P}(B).$$

Etant donnée une variable aléatoire X à valeurs réelles définie sur Ω, on parlera de la **loi conditionnelle de X par rapport à $\mathcal{A}$ et $\mathbb{P}$** pour désigner la variable aléatoire constituée par la loi de X par rapport à la probabilité aléatoire $\mathbb{P}(\cdot|\mathcal{A})$ sur Ω.

Si X est à valeurs réelles et possède une espérance, on pourra définir de la même manière **l'espérance conditionnelle de X par rapport à $\mathcal{A}$ et $\mathbb{P}$**, notée $\mathbb{E}_{\mathbb{P}}(X|\mathcal{A})$ (en utilisant le fait que si X possède une espérance sous $\mathbb{P}$, il en va de même de $\mathbb{P}(\cdot|A_i)$).

(Comme toujours, lorsqu'il n'y a pas d'ambiguité, nous écrirons simplement $\mathbb{E}(X|\mathcal{A})$.) On vérifie facilement que, si X possède une espérance, c'est également le cas de $\mathbb{E}(X|\mathcal{A})$, et que

$$\mathbb{E}_{\mathbb{P}}\mathbb{E}_{\mathbb{P}}(X|\mathcal{A}) = \mathbb{E}_{\mathbb{P}}(X).$$

Pour le voir, il suffit d'écrire que $1 = \sum_{i=1}^{p} \mathbf{1}(A_i)$, et donc que $X = \sum_{i=1}^{p} X\mathbf{1}(A_i)$, d'où le fait que $\mathbb{E}(X) = \sum_{i=1}^{p} \mathbb{E}(X\mathbf{1}(A_i)) = \sum_{i=1}^{p} \mathbb{E}(X|A_i) \times \mathbb{P}(A_i) = \mathbb{E}\left(\mathbb{E}(X|A_i)\right)$.

Une caractérisation importante de l'espérance conditionnelle, qui découle de la caractérisation de l'espérance donnée dans l'exercice 94 est la suivante : si X possède une variance, c'est également le cas de $\mathbb{E}(X|\mathcal{A})$, et $\mathbb{E}(X|\mathcal{A})$ est la meilleure approximation au sens des moindres carrés de X par une variable aléatoire qui ne dépend que de la réalisation des événements de $\mathcal{A}$ (c'est-à-dire une fonction de la variable aléatoire $(\mathbf{1}(A_1), \ldots, \mathbf{1}(A_p))$, ou encore, pour réutiliser une définition du chapitre précédent, une variable aléatoire possédant une traduction formelle dans l'espace des possibles $\Omega_{\mathcal{A}} = \{A_1, A_1^c\} \times \cdots \times \{A_n, A_n^c\}$ décrivant la réalisation des événements A_i).

En particulier, si l'on prend comme système complet d'événements la liste des événements «$X = s$ », que nous noterons $\mathbb{A}(X)$, et si Y est une variable aléatoire définie sur $(\Omega, \mathbb{P})$ et possédant une variance, $\mathbb{E}(Y|\mathcal{A}(X))$, que nous noterons parfois simplement $\mathbb{E}(Y|X)$ est la meilleure approximation au sens des moindres carrés de

Y par une variable aléatoire qui s'écrit comme une fonction de X. (Voir à ce sujet la partie sur la régression).

En termes plus abstraits, et pour reprendre l'interprétation géométrique présentée précédemment à propos de la régression linéaire, on vérifie que l'ensemble $\mathcal{L}^2(\Omega, \mathcal{A}(X), \mathbb{P})$ des variables aléatoires possédant une variance et s'écrivant comme une fonction de X forme un sous-espace vectoriel fermé de $\mathcal{L}^2(\Omega, \mathbb{P})$. L'application $Y \mapsto \mathbb{E}(Y|\mathcal{A}(X))$ définie sur $\mathcal{L}^2(\Omega, \mathbb{P})$ s'identifie alors, d'après ce qui précède, à la projection orthogonale sur $\mathcal{L}^2(\Omega, \mathcal{A}(X), \mathbb{P})$.

Pour plus de détails sur cette notion importante, nous vous invitons à consulter les ouvrages d'introduction à la théorie mathématique des probabilités cités dans la bibliographie.

2.8 Conditionnement par une variable aléatoire de loi continue

Donner une présentation mathématiquement rigoureuse du conditionnement par des variables aléatoires continues est possible, mais nécessite le recours à la théorie mathématique de la mesure. Nous nous contenterons donc d'une approche plus intuitive, conforme à celle adoptée pour la définition des variables aléatoires continues, et basée sur l'idée selon laquelle l'utilisation de variables aléatoires continues peut se voir comme un procédé d'approximation de modèles discrets, mais dont le caractère discret n'apparaît qu'à une échelle microscopique.

Si X est une variable aléatoire de loi continue, on ne peut pas définir directement les probabilités conditionnelles telles que $\mathbb{P}(A|X = x)$, du fait que $\mathbb{P}(X = x) = 0$. Revenant à notre interprétation des variables aléatoires continues comme cas limite de variables aléatoires discrètes, considérons une variable aléatoire discrète à l'échelle microscopique, mais dont le comportement macroscopique peut être décrit (avec une bonne approximation) par celui d'une variable continue de densité f.

Les probabilités de la forme $\mathbb{P}(A|X = x)$ sont alors bien définies lorsque $x \in S_X$, mais certainement pas pour tout $x \in \mathbb{R}$. De plus, la valeur d'une telle probabilité est a priori très sensible à la valeur exacte de X. Nous nous intéresserons plutôt aux probabilités de la forme $\mathbb{P}(A|x \leq X \leq x + dx)$, pour des valeurs de dx petites à l'échelle macroscopique, mais grandes devant l'échelle microscopique à laquelle le caractère discret de la variable X apparaît.

Si l'événement A dépend d'une manière suffisament régulière des valeurs prises par la variable X (en particulier, on suppose que l'événement A n'est pas sensible à la valeur exacte prise par X), on peut alors s'attendre à ce que, pour des valeurs de dx suffisament petites (mais pas trop, comme expliqué ci-dessus), la valeur de $\mathbb{P}(A|x \leq X \leq x + dx)$ soit proche d'une valeur fixée $h(A, x)$ ne dépendant que de x

et de A, mais pas de la valeur précise de dx.

(Illustration sur un exemple et graphique.)

Partant de l'identité $\mathbb{P}(A) = \sum_{i=-\infty}^{+\infty} \mathbb{P}(A|idx \leq X < idx + dx)\mathbb{P}(idx \leq X < idx + dx)$, on en déduit que $\mathbb{P}(A)$ doit alors être proche de la valeur

$$\int_{-\infty}^{+\infty} h(A, x) f(x) dx$$

.

Voyant le cas d'une variable continue comme un procédé d'approximation de la situation discrète que nous venons de décrire, nous serons donc amenés à utiliser la définition suivante : pour une variable continue, $\mathbb{P}(A|X = x) = \lim_{dx \to 0} \mathbb{P}(A|x \leq X \leq x + dx)$, lorsque cette limite existe.

On peut alors écrire que

$$\mathbb{P}(A) = \int_{-\infty}^{+\infty} \mathbb{P}(A|X = x) f(x) dx. \tag{2.4}$$

Pour une variable discrète, cela revient donc à utiliser $h(A, x)$ en lieu et place de $\mathbb{P}(A|X = x)$, même si cette dernière expression est définie. C'est à cette condition que l'on peut utiliser la densité de X pour faire les calculs dans des expressions telles que 2.4.

Cette définition étant acquise, on peut utiliser l'expression $\mathbb{P}(A|X = x)$ essentiellement comme on le ferait dans le cas discret, en se rappelant les règles de passage usuelles du cas discret au cas continu ($\sum \leftrightarrow \int$, et $\mathbb{P}(X = x) \leftrightarrow f(x)dx$).

2.9 Transformées de Laplace et de Fourier d'une loi de probabilité *

Nous décrivons très brièvement dans cette partie des outils mathématiques très utiles dans l'étude des lois de probabilité. Le niveau mathématique nécessaire à une présentation rigoureuse de leurs propriétés dépasse nettement celui de ce cours, et nous vous renvoyons aux ouvrages classiques d'introduction à la théorie mathématique des probabilités qui sont cités dans la bibliographie pour plus de détails. Dans notre contexte, nous aurons surtout l'occasion de les utiliser comme des intermédiaires commodes pour calculer explicitement des lois de sommes de variables aléatoires indépendantes, la portée de ces outils dépassant cependant de très loin ce cadre d'application restreint.

2.9.1 Fonction génératrice

Considérons une variable aléatoire X à valeurs dans $\mathbb{N}$.

Du fait que $\mathbb{P}(X = k) \geq 0$ pour tout k et que $\sum_{k=0}^{+\infty} \mathbb{P}(X = k) = 1$, la série entière

$$G_X(z) = \sum_{k=0}^{+\infty} \mathbb{P}(X = k) z^k,$$

converge pour tout $z \in \mathbb{C}$ tel que $|z| \leq 1$. On l'appelle la série génératrice de la variable aléatoire X. On vérifie immédiatement que G_X ne dépend que de la loi de X, et que l'on peut en fait écrire $G_X(z) = \mathbb{E}(z^X)$.

Les théorèmes habituels sur les séries entières montrent que la fonction G_X (c'est-à-dire l'ensemble des valeurs $G_X(z)$, et non pas seulement la valeur en un point donné) caractérise entièrement la suite $\mathbb{P}(X = k)$, et donc la loi de X.

L'un des intérêts de cette notion est qu'elle se comporte particulièrement bien vis-à-vis de l'addition des variables aléatoires indépendantes.

Ainsi, si X et Y sont indépendantes, on a, pour tout $z \in \mathbb{C}$ tel que $|z| < 1$, l'identité

$$G_{X+Y}(z) = G_X(z) \times G_Y(z).$$

Cette identité est une conséquence immédiate du fait que z^X et z^Y sont indépendantes.

2.9.2 Transformée de Laplace

Cette notion est utilisée pour les variables aléatoires à valeurs réelles positives (discrète ou continues). Il s'agit de la fonction définie sur $\mathbb{R}_+$ par $\mathcal{L}_X(t) = \mathbb{E}(\exp(-tX))$. (Cette espérance est toujours définie car $\exp(-tX)$ est compris entre 0 et 1 du fait que X est à valeurs positives). Bien entendu, cette fonction ne dépend que de la loi de X.

Le lien avec les fonctions génératrices définies dans le paragraphe précédent est le suivant : pour une variable aléatoire à valeurs entières positives, on a $G_X(\exp(-t)) = \mathcal{L}_X(t)$, comme on le vérifie immédiatement à partir des définitions.

Ici encore, on vérifie que, si X et Y sont indépendantes (et positives), on a, pour tout $t \in \mathbb{R}_+$, l'identité

$$\mathcal{L}_{X+Y}(z) = \mathcal{L}_X(z) \times \mathcal{L}_Y(z).$$

Cette identité est une conséquence immédiate du fait que $\exp(-tX)$ et $\exp(-tY)$ sont indépendantes.

Nous admettrons que la donnée de la fonction $\mathcal{L}_X$ caractérise entièrement la loi de X.

Autrement dit, si deux variables aléatoires positives X et Y sont telles que $\mathcal{L}_X = \mathcal{L}_Y$, alors X et Y ont même loi.

2.9.3 Transformée de Fourier

Cette notion est utilisée pour les variables aléatoires à valeurs réelles générales (discrètes ou continues, sans hypothèse de positivité comme pour la transformée de Laplace).

Il s'agit de la fonction définie sur $\mathbb{R}$ par $\mathcal{F}_X(t) = \mathbb{E}(\exp(itX))$. (Cette espérance est toujours définie car $|\exp(itX)| = 1$ pour tout t).

Bien entendu, cette fonction ne dépend que de la loi de X.

Nous admettrons que la donnée de la fonction $\mathcal{F}_X$ caractérise entièrement la loi de X.

Autrement dit, si deux variables aléatoires X et Y sont telles que $\mathcal{F}_X = \mathcal{F}_Y$, alors X et Y ont même loi.

Ici encore, on vérifie que, si X et Y sont indépendantes, on a, pour tout $t \in \mathbb{R}_+$, l'identité

$$\mathcal{F}_{X+Y}(z) = \mathcal{F}_X(z) \times \mathcal{F}_Y(z).$$

Cette identité est une conséquence immédiate du fait que $\exp(itX)$ et $\exp(itY)$ sont indépendantes.

2.9.4 Transformées des lois classiques

Loi de Bernoulli

On voit immédiatement que, si X suit une loi de Bernoulli de paramètre p, on a

$$\mathcal{F}_X(t) = p\exp(it) + 1 - p \text{ et } \mathcal{L}_X(t) = p\exp(-t) + 1 - p.$$

Loi binomiale

Une conséquence de ce qui précède est que, si X suit une loi binomiale de paramètres n et p, on a

$$\mathcal{F}_X(t) = [p\exp(it) + 1 - p]^n \text{ et } \mathcal{L}_X(t) = [p\exp(-t) + 1 - p]^n.$$

Loi de Poisson

Si X suit une loi de Poisson de paramètre λ, on a

$$\mathcal{F}_X(t) = \exp[\lambda(\exp(it) - 1)] \text{ et } \mathcal{L}_X(t) = \exp[\lambda(\exp(-t) - 1)].$$

(Calcul, ou approximation par une loi binomiale).

Loi géométrique

Si X suit une loi géométrique de paramètre p, on a

$$\mathcal{F}_X(t) = \frac{p\exp(it)}{1-(1-p)\exp(it)} \text{ et } \mathcal{L}_X(t) = \frac{p\exp(-t)}{1-(1-p)\exp(-t)}.$$

Par le calcul, ou en observant que la loi de X est le mélange de la loi de $X+1$ (avec probabilité $(1-p)$) et de la loi concentrée sur la valeur constante 1 (avec probabilité p).

Loi exponentielle

Si X suit une loi exponentielle de paramètre λ, on a

$$\mathcal{F}_X(t) = \frac{i\lambda}{t+i\lambda} \text{ et } \mathcal{L}_X(t) = \frac{\lambda}{-t+\lambda}.$$

Par le calcul, ou en approchant par une loi géométrique renormalisée.

Loi gaussienne

Si X suit une loi gaussienne de paramètres m et v, on a

$$\mathcal{F}_X(t) = \exp\left(itm - \frac{t^2}{2v}\right)$$

.

Par le calcul de l'intégrale correspondante.

Loi de Cauchy

Si X suit une loi gaussienne de paramètres ℓ et s, on a $\mathcal{F}_X(t) = \exp\left(it\ell - \frac{s|t|}{4}\right)$.
Par le calcul de l'intégrale correspondante.

2.10 Quelques mots de théorie de l'information *

Ce qui suit est fortement inspiré de la présentation donnée dans l'ouvrage de P. Brémaud cité en bibliographie. Les ouvrages traitant de la théorie de l'information sont très nombreux. Nous citons dans la bibliographie celui de Cover et Thomas.

2.10.1 Entropie

Etant donné un ensemble fini S, et une probabilité $\mathbb{P}$ sur S, on définit **l'entropie en base 2 de** $\mathbb{P}$ par la formule

$$H_2(\mathbb{P}) = -\sum_{x\in S} \mathbb{P}(x)\log_2(\mathbb{P}(x)),$$

avec la convention $0 \log_2(0) = 0$.

On vérifie aisément les propriétés suivantes :

- l'entropie est un nombre positif ou nul ;
- si $(S_1, \mathbb{P}_1)$ et $(S_2, \mathbb{P}_2)$ sont deux modèles probabilistes, $H_2(\mathbb{P}_1 \otimes \mathbb{P}_2) = H_2(\mathbb{P}_1) + H_2(\mathbb{P}_2)$;
- la probabilité uniforme sur S maximise l'entropie parmi les distributions de probabilité possibles ;
- à l'autre extrême, si la probabilité $\mathbb{P}$ est concentrée sur un seul élément $x \in S$, (i.e. $\mathbb{P}(x) = 1$), l'entropie est nulle.

2.10.2 Questionnaires

Définition

Formellement, un **questionnaire binaire** Q permettant d'identifier les éléments d'un ensemble fini S est la donnée d'un arbre enraciné T_Q dans lequel tout sommet non-terminal possède un ou deux fils, et dont les feuilles sont en bijection avec les éléments de S (on dit que les sommets de l'abre sont étiquetés par les éléments de S). Le questionnaire est dit **efficace** lorsque tous les sommets non-terminaux possèdent exactement deux fils (ce dernier terme n'est pas standard).

Pour tout sommet v de T_Q, nous noterons $S(v)$ l'ensemble des éléments de S étiquetant les feuilles du sous-arbre de T_Q issu de v.

Si v est un sommet non-terminal de T_Q, on note $d(v)$ le nombre de ses fils ; $d(v)$ est donc égal à 1 ou 2. Si $d(v) = 1$, on note v_1 l'unique fils de v. Si $d(v) = 2$, on note v_1 et v_2 ces deux fils, (en choisissant arbitrairement celui qui est numéroté 1 et celui qui est numéroté 2).

Si $d(v) = 2$, on constate que $S(v) = S(v_1) \cup S(v_2)$ et $S(v_1) \cap S(v_2) = \emptyset$.

Si $d(v) = 1$, on a $S(v) = S(v_1)$.

On obtient un questionnaire (au sens usuel) en associant à chaque sommet non-terminal v de l'arbre, la question : « x appartient-il à $S(v_1)$? ».

Partant d'un élément x de S inconnu, on pose d'abord la question associée à la racine, notée pour la circonstance $w(x, 0)$: « x appartient-il à $S(w(x, 0)_1)$? » Si ce n'est pas le cas, x appartient nécessairement à $S(w(x, 0)_2)$. En notant $w(x, 1)$ l'unique fils de $w(x, 0)$ tel que $x \in S(w(x, 1))$, on pose ensuite la question : « x appartient-il à $S(w(x, 1)_1)$? » Si ce n'est pas le cas, x appartient nécessairement à $S(w(x, 1)_2)$. On définit ensuite $w(x, 2)$ comme l'unique fils de $w(x, 1)$ tel que $x \in S(w(x, 2))$, et... on itère le procédé en définissant successivement $w(x, 3), w(x, 4), \ldots$, jusqu'à avoir défini un $w(x, i)$ dont x est un fils. On a alors identifié x, au moyen de questions successives relatives à sa localisation dans S, qui permettent de le localiser de plus en plus finement jusqu'à l'identifier complètement. De manière imagée, on remonte l'arbre, en partant de la racine jusqu'à la feuille étiquetée par x, en posant à chaque

fois la question qui permet de déterminer quelle est la bifurcation dirigée vers x (lorsqu'il y a possibilité de bifurcation, c'est-à-dire deux fils).

On note qu'une question associée à un sommet v qui vérifie $d(v) = 1$ est inutile, car sa réponse est identique à celle de la question associée au père de v. C'est en ce sens qu'un questionnaire possédant des sommets vérifiant $d(v) = 1$ est dit inefficace.

Partant d'un questionnaire qui n'est pas efficace, il suffit de contracter toutes les arêtes reliant un sommet à son fils unique, c'est-à-dire toutes les questions inutiles, pour le transformer en un questionnaire efficace.

(Dessin.)

La profondeur du sommet de l'arbre étiqueté par un élément donné x de S correspond donc au nombre de questions qu'il est nécessaire de poser avec ce questionnaire pour identifier x (en tenant compte des questions inutiles dans le cas de questionnaires inefficaces). Etant donné un questionnaire Q relatif à S et un élément x de S, nous noterons $\ell_Q(x)$ cette profondeur.

Questionnaires et codes préfixes

Il y a **correspondance bijective entre questionnaires binaires et codes binaires p ossédant la propriété du préfixe** (c'est-à-dire qu'aucun mot de code n'est le préfixe d'un autre, ce qui évite les ambiguités de décodage lorsque plusieurs mots sont transmis à la suite).

Un questionnaire fournit un tel code en associant à chaque élément de S un mot de code formé par les réponses successives aux questions posées (de la racine jusqu'à la feuille) pour localiser cet élément.

Inversement, partons d'un code binaire possédant la propriété du préfixe. Appelons M la longueur du plus long mot de code, et considérons $\mathbb{T}$, l'arbre binaire complet de profondeur M. Pour tout sommet non-terminal, numérotons par 0 et 1 les deux fils de ce sommet, de manière à pouvoir repérer chaque sommet de $\mathbb{T}$ par une suite $(a_1, \ldots, a_k)$ de 0 et et de 1 indiquant les choix successifs de fils menant de la racine à cette feuille. On peut ainsi associer à tout mot de code binaire le sommet de $\mathbb{T}$ repéré par ce mot de code dans l'indexation de $\mathbb{T}$ que nous venons de décrire. Notons C l'ensemble des sommets associés aux mots du code par ce procédé. En élaguant l'arbre $\mathbb{T}$ par suppression de tous les descendants des éléments de C, si bien que ceux-ci constituent l'ensemble des feuilles de l'arbre ainsi élagué, et en étiquetant ces feuilles par les éléments de S associés aux mots de code correspondant, on obtient un questionnaire binaire permettant d'identifier les éléments de S (le fait que cette construction fonctionne utilise le fait que le code possède la propriété du préfixe ; où donc ?).

Le nombre de questions à poser pour identifier un élément x de S au moyen d'un questionnaire est égal, dans cette correspondance, à la longueur du mot de code

associé à x.

Inégalité de Kraft

Nous allons prouver que tout questionnaire binaire vérifie l'inégalité suivante, appelée **inégalité de Kraft** :

$$\sum_{x\in S} 2^{-\ell_Q(x)} \leq 1.$$

Posons $M = \max\{\ell_Q(x) \ : \ x \in S\}$. Complétons l'arbre associé à Q en un arbre binaire complet de profondeur M, noté $\mathbb{T}$, et qui comporte donc 2^M feuilles.

Le sous-arbre de $\mathbb{T}$ formé par les descendants d'un sommet situé à une profondeur $\ell_Q(x)$ comporte $2^{M-\ell_Q(x)}$ feuilles. Clairement, les sous-arbres ainsi obtenus en partant de sommets distincts de l'arbre associé à Q sont disjoints, d'où le fait que $\sum_{x\in S} 2^{M-\ell_Q(x)} \leq 2^M$, ce qui entraîne l'inégalité annoncée.

Pour un questionnaire efficace, le même argument prouve que l'on a exactement $\sum_{x\in S} 2^{-\ell_Q(x)} = 1$.

Nous allons maintenant prouver la **réciproque de l'inégalité de Kraft** : à toute famille d'entiers $d_x,\ x \in S$, supérieurs ou égaux à 1, et vérifiant

$$\sum_{x\in S} 2^{-d_x} \leq 1,$$

on peut associer un questionnaire Q relatif à S et tel que $\ell_Q(x) = d_x$.

Ecrivons la liste des éléments de S sous la forme $\{x_1, \ldots, x_n\}$, de telle sorte que $d_{x_1} \leq d_{x_2} \leq \cdots \leq d_{x_n}$. Posons $M = \max\{d_x \ : \ x \in S\} = d_{x_n}$, et considérons une fois encore $\mathbb{T}$, l'arbre binaire complet de profondeur M, en repérant les sommets par des suites de 0 et de 1 grâce à l'indexation décrite précédemment. A chaque feuille v de $\mathbb{T}$, on associe ensuite un entier $\phi(v)$ définie par $\phi(a_1, \ldots, a_M) = \sum_{i=0}^{M-1} a_i 2^i$, où $(a_1, \ldots, a_M)$ est l'indexation de cette feuille.

Ensuite, les feuilles de $\mathbb{T}$ numérotées de 1 à 2^{M-d_1} sont étiquetées par x_1, les feuilles numérotées de $2^{M-d_1}+1$ à $2^{M-d_1}+2^{M-d_2}$ sont étiquetées par x_2, etc... L'inégalité que nous avons supposée sur les d_x garantit que l'on peut poursuivre ce procédé jusqu'à avoir étiqueté les $\sum_{x\in S} 2^{M-d_x}$ premières feuilles, les $2^M - \sum_{x\in S} 2^{M-d_x}$ dernières feuilles de $\mathbb{T}$ restant sans étiquettes. Qui plus est, le fait que la suite $(M-d_{x_i})_{1\leq i\leq n}$ soit décroissante entraîne le fait que, pour tout i, il existe un sommet h_i de $\mathbb{T}$ situé à une profondeur égale à d_{x_i}, tel que les feuilles qui en descendent sont exactement les feuilles étiquetées par x_i dans l'étiquetage que nous venons de définir.

On construit alors un questionnaire de la manière suivante : on étiquette chaque h_i par x_i, puis on élague $\mathbb{T}$ en supprimant tous les descendants des h_i, si bien que $\{h_1, \ldots, h_n\}$ forme l'ensemble des feuilles de l'arbre élagué.

Questionnaires optimaux et entropie : borne de Shannon

Une conséquence simple de l'inégalité de Kraft est que, dans tout questionnaire Q, il existe au moins un élément x de S tel que $\ell_Q(x) \geq \lceil \log_2(|S|) \rceil$. Réciproquement, il est clair que l'on peut toujours construire un questionnaire dans lequel tous les éléments ont une profondeur inférieure ou égale à $\lceil \log_2(|S|) \rceil$. Si la performance d'un questionnaire est mesurée par sa profondeur maximale (le nombre de questions qu'il est nécessaire de poser pour identifier un élément dans la pire des cas), la question de trouver un questionnaire optimal n'est donc pas très intéressante. En revanche, lorsque S est muni d'une probabilité $\mathbb{P}$, et que l'on étudie le nombre de questions qu'il est nécessaire de poser pour identifier un élément de S **choisi selon la probabilité** $\mathbb{P}$, on obtient une variable aléatoire dont la loi peut différer très fortement d'un questionnaire à l'autre. Intuitivement, on peut tirer parti de différences de probabilité entre les différents éléments de S en associant aux plus probables les nombres de questions les plus faibles.

Nous supposerons dans la suite que $\mathbb{P}(x) > 0$ pour tout $x \in S$ (si ce n'est pas le cas, il suffit d'éliminer de S les éléments de probabilité nulle, qui, de toute façon, ne peuvent jamais apparaître).

On s'intéressera spécifiquement à l'espérance du nombre de questions à poser pour identifier un élément de S choisi selon la probabilité $\mathbb{P}$, soit

$$L_{\mathbb{P}}(Q) = \sum_{x \in S} \ell_Q(x) \mathbb{P}(x).$$

Si l'on doit utiliser un questionnaire de manière répétée pour localiser des éléments de x choisis selon la probabilité $\mathbb{P}$ (ou, dans l'interprétation en termes de codage, si l'on doit coder de manière répétée des éléments de x), $L_{\mathbb{P}}(Q)$ est *a priori* une quantité plus pertinente que la longueur de codage dans le pire des cas, puisqu'elle représente, à long terme, le nombre moyen de questions par élément de x qu'il nous faudra poser.

Nous montrerons dans la suite que, $(S, \mathbb{P})$ étant donné, il existe un questionnaire Q qui minimise la valeur de $L_{\mathbb{P}}(Q)$ parmi l'ensemble des questionnaires possibles, et que l'on dispose d'un algorithme efficace pour construire un tel questionnaire.

Ce que nous allons prouver pour l'instant est le résultat suivant, connu sous le nom de **borne de Shannon** [13] : $(S, \mathbb{P})$ étant donné,

$$H_2(\mathbb{P}) \leq \min L_{\mathbb{P}}(Q) \leq H_2(\mathbb{P}) + 1,$$

le minimum étant pris sur la totalité des questionnaires binaires permettant d'identifier les éléments de S.

13. Claude Elwood Shannon (1916–2001).

Par conséquent, **l'entropie de $\mathbb{P}$ apparaît comme une mesure** (au moins approchée) **du minimum du nombre moyen de questions à poser pour identifier les éléments de S lorsque ceux-ci sont générés selon la probabilité $\mathbb{P}$.**

Lorsque $\mathbb{P}$ est la loi uniforme, on constate que, d'après l'inégalité ci-dessus, on ne peut guère faire mieux que le questionnaire dans lequel tous les éléments sont associés au même nombre de questions, le cas moyen et le pire cas étant essentiellement équivalents.

Remarque 10 *Si l'on considère des suites indépendantes d'éléments de S de longueur n, générées selon la probabilité $\mathbb{P}^{\otimes n}$, on obtient donc que le nombre moyen minimal de questions à poser est compris entre $nH_2(\mathbb{P})$ et $nH_2(\mathbb{P})+1$, et, par conséquent, ce nombre moyen rapporté à la longueur de la suite tend vers $H_2(\mathbb{P})$ lorsque n tend vers l'infini, ce qui confère un caractère naturel à $H_2(\mathbb{P})$ (qui n'intervenait qu'à un terme d'erreur pouvant aller jusqu'à 1 dans l'inégalité précédente).*

Remarque 11 *L'entropie, telle que nous l'avons introduite dans cette partie, intervient dans bien d'autres contextes (par exemple en physique statistique, en statistique bayésienne, en intelligence artificielle,...), où elle joue un rôle important, avec des interprétations parfois très différentes.*

Prouvons maintenant la borne de Shannon. Dans la suite, nous noterons $\Xi((d_x)_{x\in S}) = \sum_{x\in S} d_x\mathbb{P}(x)$ et

Appelons $\mathcal{D}$ l'ensemble des familles d'entiers $(d_x)_{x\in S}$, supérieurs ou égaux à 1, et vérifiant $\sum_{x\in S} 2^{-d_x} \leq 1$.

D'après l'inégalité de Kraft et sa réciproque, le problème de minimisation que nous étudions se ramène au suivant :

$$\text{minimiser } \Xi \text{ sur l'ensemble } \mathcal{D}.$$

On vérifie que, lorsque $(d_x)_{x\in S}$ tend vers l'infini, c'est également le cas de $\Xi((d_x)_{x\in S})$ par positivité des $\mathbb{P}(x)$. Par conséquent, Ξ possède bien un minimum absolu sur $\mathcal{D}$.

Le fait que l'ensemble $\mathcal{D}$ soit constitué de nombres entiers nous complique la vie car nous ne pouvons pas utiliser les outils du calcul différentiel pour résoudre ce problème de minimisation.

Appelons $\mathcal{D}'$ des familles de nombres réels $(d_x)_{x\in S}$, supérieurs ou égaux à 0, et vérifiant $\sum_{x\in S} 2^{-d_x} \leq 1$, et considérons le problème de minimisation suivant :

$$\text{minimiser } \Xi \text{ sur l'ensemble } \mathcal{D}'.$$

On note qu'en réalité, un élément de $\mathcal{D}'$ vérifie toujours que $d_x > 0$ pour tout $x \in S$, sans quoi l'inégalité $\sum_{x\in S} 2^{-d_x} \leq 1$ serait contredite.

Par continuité de Ξ, et toujours du fait que $\Xi((d_x)_{x\in S})$ tend vers l'infini lorsque $(d_x)_{x\in S}$ tend vers l'infini, Ξ possède bien un minimum absolu sur $\mathcal{D}'$. De plus, si la contrainte $\sum_{x\in S} 2^{-d_x} \leq 1$ est satisfaite, c'est encore le cas si l'on augmente certains des d_x. Or cette opération fait croître strictement la fonctionnelle $\sum_{x\in S} d_x \mathbb{P}(x)$. On en déduit que le minimum ne peut être atteint que pour une famille d_x vérifiant $\sum_{x\in S} 2^{-d_x} = 1$.

Calculons la différentielle de Φ en $(d_x)_{x\in S}$: $D\Phi((d_x)_{x\in S}) = \sum_{x\in S} \mathbb{P}(x) Dd_x$. Par ailleurs, $D(\sum_{x\in S} 2^{-d_x}) = \sum_{x\in S} \log(2) 2^{-d_x} Dd_x$. En écrivant, comme nous y autorise le théorème des extrema liés, que $D\Phi$ doit être proportionnelle à $D(\sum_{x\in S} 2^{-d_x})$ en un extremum local sous contrainte, on en déduit que le minimum est atteint pour $d_x = -\log_2(\mathbb{P}(x))$.

Par conséquent, $\min_{\mathcal{D}'} \Xi = H_2(\mathbb{P})$.

En notant que $\min_{\mathcal{D}} \Xi \geq \min_{\mathcal{D}'} \Xi$ puisque $\mathcal{D} \subset \mathcal{D}'$, on en déduit que $\min_{\mathcal{D}} \Xi \geq H_2(\mathbb{P})$, ce qui fournit une moitité de l'inégalité annoncée. Quant à l'autre moitié, on vérifie que la famille d'entiers $(\lceil \log_2(\mathbb{P}(x)) \rceil)_{x\in S}$ est dans $\mathcal{D}$, et l'on vérifie facilement que $\Xi(\lceil \log_2(\mathbb{P}(x)) \rceil_{x\in S}) \leq \Xi(\log_2(\mathbb{P}(x))_{x\in S}) + 1$. L'inégalité affirmant que $\min_{\mathcal{D}} \Xi \leq H_2(\mathbb{P}) + 1$ en résulte.

L'algorithme de Huffmann

L'algorithme de Huffmann permet de construire de manière récursive un questionnaire optimal.

Pour décrire le principe de l'algorithme, écrivons $S = \{x_1, \ldots, x_n\}$, où l'indexation est choisie de telle sorte que $\mathbb{P}(x_1) \geq \mathbb{P}(x_2) \geq \cdots \geq \mathbb{P}(x_n)$. Si S ne comporte que deux éléments, le questionnaire optimal consiste tout simplement à poser une seule question pour déterminer auquel des deux éléments on a affaire. Si S comporte $n \geq 3$ éléments, un questionnaire optimal s'obtient en appelant récursivement l'algorithme afin de construire un questionnaire optimal pour l'ensemble $S' = \{x_1, x_2, \ldots, x_{n-2}, y\}$ muni de la probabilité $\mathbb{P}'$ définie par $\mathbb{P}'(x_i) = \mathbb{P}(x_i)$ pour $i = 1, 2, \ldots, n-2$, et $\mathbb{P}'(y) = \mathbb{P}(x_{n-1}) + \mathbb{P}(x_n)$, cet ensemble comportant un élément de moins que S. Ensuite, dans l'arbre associé au questionnaire ainsi obtenu, on greffe deux enfants sur la feuille étiquetée par y, les deux feuilles ainsi obtenues étant étiquetées par x_{n-1} et x_n respectivement. On obtient ainsi un questionnaire optimal pour $(S, \mathbb{P})$.

Prouvons l'optimalité de cet algorithme.

Appelons Q^* un questionnaire optimal pour $(S, \mathbb{P})$. Nous allons montrer que l'on peut le transformer en un autre questionnaire optimal pour lequel la feuille sœur de celle étiquetée par x_n est étiquetée par x_{n-1}.

Commençons par observer que, si $\mathbb{P}(a) < \mathbb{P}(b)$, on a nécessairement $\ell_{Q^*}(a) \leq \ell_{Q^*}(b)$. (Cette propriété est intuitivement claire : dans un questionnaire optimal,

les éléments les plus probables doivent avoir la profondeur la plus faible possible.) En effet, notons que, si $\ell_{Q^*}(a) > \ell_{Q^*}(b)$, il suffit d'échanger les étiquetages des feuilles associées à a et à b dans l'arbre du questionnaire pour obtenir un nouveau questionnaire Q_2^* tel que $L_{\mathbb{P}}(Q_2^*) < L_{\mathbb{P}}(Q^*)$, ce qui est impossible par optimalité de Q^*.

Considérons la feuille sœur de celle étiquetée par x_n dans Q^*, et appelons z son étiquette. Si $\mathbb{P}(z) < \mathbb{P}(x_{n-1})$, on a nécessairement $\ell_{Q^*}(z) \leq \ell_{Q^*}(x_{n-1})$. Si $\ell_{Q^*}(x_{n-1}) = \ell_{Q^*}(z)$, il suffit d'échanger les étiquetages des feuilles associées à z et à x_{n-1}, pour obtenir un questionnaire dans lequel les feuilles étiquetées par x_n et x_{n-1} sont sœurs, tout en conservant la valeur de $L_{\mathbb{P}}(Q^*)$, et donc l'optimalité. Si $\ell_{Q^*}(x_{n-1}) > \ell_{Q^*}(z) = \ell_{Q^*}(x_n)$, en appelant w l'étiquette de la feuille sœur de x_{n-1}, on doit nécessairement avoir $\mathbb{P}(w) = \mathbb{P}(x_n)$. En échangeant les étiquetages des feuilles associées à w et à x_n, on obtient encore un questionnaire où les feuilles étiquetées par x_n et x_{n-1} sont sœurs, tout en conservant la valeur de $L_{\mathbb{P}}(Q^*)$, et donc l'optimalité. Enfin, si $\mathbb{P}(z) = \mathbb{P}(x_{n-1})$, en échangeant les étiquetages des feuilles associées à z et à x_{n-1}, on obtient encore un questionnaire où les feuilles étiquetées par x_n et x_{n-1} sont sœurs, tout en conservant la valeur de $L_{\mathbb{P}}(Q^*)$, et donc l'optimalité.

(Dans le cas où l'on a $Q(x_1) > \cdots > Q(x_n)$, on note que le raisonnement ci-dessus peut être grandement simplifié, car on a nécessairement que $\ell_Q(x_n) = \ell_Q(x_{n-1})$.)

Considérons donc un questionnaire optimal Q^* tel que les feuilles étiquetées par x_n et x_{n-1} sont sœurs.

Partant de Q^*, nous pouvons construire un questionnaire $\beta(Q^*)$ sur S' en transformant le père de x_n et x_{n-1} en une feuille, étiquetée par y, et qui vérifie donc que $L_{\mathbb{P}}(Q^*) = L_{\mathbb{P}'}(\beta(Q^*)) + \mathbb{P}(x_{n-1}) + \mathbb{P}(x_n)$.

Inversement, partant d'un questionnaire Q sur S', nous appellerons $\theta(Q)$ le questionnaire fabriqué par l'algorithme de Huffmann sur S à partir de Q. Clairement, $L_{\mathbb{P}}(\theta(Q)) = L_{\mathbb{P}'}(Q) + \mathbb{P}(x_{n-1}) + \mathbb{P}(x_n)$. De plus, $\theta(\beta(Q^*)) = Q^*$.

Nous voyons à présent que $\beta(Q^*)$ doit être optimal pour $(S', \mathbb{P}')$. Sinon, on pourrait trouver un questionnaire Q sur S' tel que $L_{\mathbb{P}'}(Q) < L_{\mathbb{P}'}(\beta(Q^*))$, d'où le fait que $L_{\mathbb{P}}(\theta(Q)) < L_{\mathbb{P}}(Q^*))$, ce qui contredirait l'optimalité de Q^*.

Nécessairement, il existe donc un questionnaire optimal pour $(S', \mathbb{P}')$ que θ transforme en Q^*.

Comme la différence entre $L_{\mathbb{P}}(\theta(Q)) - L_{\mathbb{P}'}(Q)$ est toujours égale à $\mathbb{P}(x_{n-1}) + \mathbb{P}(x_n)$, $\theta(Q)$ est optimal pour $(S, \mathbb{P})$ pour tout Q questionnaire optimal pour $(S', \mathbb{P}')$, ce qui prouve le bon fonctionnement de l'algorithme de Huffmann.

Nous n'avons présenté ici que quelques mots extraits du vaste corpus connu sous le nom de théorie de l'information, qui aborde toutes sortes de problèmes et de questions (tels que codage et décodage rapides, ou encore compression de données avec perte d'information, transmission d'information dans des canaux bruités,...).

2.11 Quelques mots sur le hasard simulé

Pour plus tard...

2.12 Les lois de Benford et de Zipf

2.12.1 La loi de Benford

Pour un entier $k \geq 1$ donné, on appelle loi de Benford sur k chiffres la loi de probabilité sur $\{0, \ldots, 9\}^k$ définie par

$$p_{Benford}(d_1, \ldots, d_k) = \log_{10}\left(1 + \sum_{i=1}^{k} d_i 10^{k-i}\right).$$

De manière surprenante, au moins pour de faibles valeurs de k (au moins $k = 1$), lorsque l'on regroupe des valeurs numériques de provenances variées, la distribution empirique des k premiers chiffres significatifs de l'écriture en base 10 des valeurs ainsi obtenues obéït avec une assez bonne approximation à la loi de Benford. Ce phénomène fut formellement constaté par Simon Newcomb en 1881, puis à nouveau mentionné et étudié par Frank Benford en 1938 (un indice amusant en faveur de cette loi étant l'usure inégale des premières et des dernières pages des tables de logarithme : les nombres commençant par 1 sont plus fréquemment manipulés que ceux commençant par 2, eux-mêmes plus fréquents que ceux commençant par 3, etc...). Des tests basés sur la loi de Benford sont par exemple utilisés par les services fiscaux de certains pays pour tenter de détecter les ensembles de chiffres fabriqués (par exemple, de faux bilans comptables), par opposition à de vraies valeurs, dont on pourrait s'attendre à ce qu'elles suivent la loi de Benford.

Diverses explications mathématiques à l'apparition, non pas systématique, mais au moins fréquente, de cette loi, ont été avancées.

Pour plus d'informations, nous vous renvoyons à l'article de Hill cité dans la bibliographie et aux références qui s'y trouvent.

Notez au passage que l'on joue ici sur les deux sens du mot «loi» : le terme «loi de Benford» désigne à la fois la loi de probabilité définie ci-dessus, et l'affirmation selon laquelle la répartition empirique des premiers chiffres significatifs dans des listes de valeurs de provenances diverses a souvent (mais pas toujours, insistons) tendance à suivre cette loi de probabilité.

Des exemples...

2.12.2 Lois de Zipf-Mandelbrot et de Pareto

On appelle lois de Zipf-Mandelbrot les lois de probabilité sur $\{1, 2, \ldots\}$ de la forme $p(n) = K(a + bn)^{-c}$, où $a \geq 0$, $b > 0$ et $c > 1$. (K est alors entièrement

déterminé par a, b et c).

Dans le cas continu, un analogue est constitué par les lois de Pareto : ce sont les lois dont la densité est de la forme $f(x) = Kx^{-c}$ pour $x \geq b$, avec $b > 0$ et $c > 1$.

Ces lois sont qualifiées de «lois de puissance» («power-law» en anglais), et de nombreux phénomènes (biologiques, physiques, sociaux) font apparaître des lois de probabilité qui, sans s'identifier parfaitement aux lois précédentes, reproduisent approximativement le comportement en x^{-c} ou n^{-c} pour une vaste gamme de valeurs de x ou n. Notez que ce comportement tranche avec celui des lois classiques (exponentielle, Poisson, gamma, gaussienne) pour lesquelles la décroissance de la probabilité ou de la densité pour les grandes valeurs est beaucoup plus rapide. Une propriété remarquable de ces lois est leur propriété d'invariance d'échelle : observée à des échelles différentes, une variable aléatoire possédant une loi de ce type présente toujours la même loi (voir l'exercice 150). De multiples explications, parfois très spéculatives, parfois moins, ont été proposées pour tenter de rendre compte de l'apparition de ce type de loi dans des situations réelles.

Exemple 9 *Un exemple relativement bien établi est que, dans de nombreux textes, si l'on classe par ordre décroissant les fréquences d'apparition des différents mots du texte, soient $f(1) \geq f(2) \geq \cdots$, (ce qui signifie donc que $f(1)$ est la fréquence du mot le plus représenté, $f(2)$ la fréquence du second mot le plus représenté, etc...) $f(n)$ correspond souvent approximativement à une loi de Zipf-Mandelbrot, tout au moins pour un certain domaine de valeurs de n. Vous pouvez vous-même tester expérimentalement la validité de la loi de Zipf sur les textes de votre choix en vous rendant sur le site maintenu par Emmanuel Giguet :* `http://users.info.unicaen.fr/~giguet/java/zipf.html`.

Autres exemples...

2.13 Auto-évaluation

- Qu'est-ce qu'une variable aléatoire (concrètement, en français) ?
- Qu'est-ce qu'une variable aléatoire (en tant qu'objet mathématique) ?
- La définition d'une variable aléatoire dépend-elle de la probabilité sur Ω ?
- Qu'est-ce que la loi d'une variable aléatoire ?
- Deux variables aléatoires possédant la même loi sont-elles nécessairement égales ? Deux variables aléatoires définies sur des espaces de probabilité différents peuvent-elles néanmoins posséder la même loi ?
- Qu'est-ce qu'une loi de probabilité en général (sans référence à une variable aléatoire) ?
- Donnez la définition des lois de Bernoulli, Binomiale, de Poisson, uniforme, géométrique, leurs paramètres, le contexte exact (hypothèses sur le modèles, exemples concrets) dans lequel on sait qu'elles interviennent.

- Qu'est-ce que la loi empirique associée à un échantillon de valeurs ?
- Pour une variable aléatoire donnée, quelle différence y a-t-il entre sa loi, et la loi empirique associée à un échantillon de valeurs collectées de cette variable aléatoire ?
- Donnez la définition de deux, et plus généralement de n variables aléatoires indépendantes.
- Les variables aléatoires qui interviennent dans la modélisation d'un phénomène doivent-elles nécessairement être définies sur le même espace de départ ? Avoir le même espace d'arrivée ?
- Donnez les deux formules définissant l'espérance. Quelle est la différence entre ces deux formules ?
- Deux variables aléatoires ayant la même loi peuvent-elles avoir des espérances différentes ? Et des variances différentes ?
- Quel est le lien entre l'espérance et la moyenne d'une liste de valeurs au sens usuel ?
- Donnez la définition de la variance et de l'écart-type d'une variable aléatoire.
- Que représente la variance ? Et l'écart-type ? En quoi la définition traduit-elle précisément cette intuition ?
- Rappelez l'inégalité de Markov. Comment utilise-t-on cette inégalité ? Est-elle toujours précise ?
- Rappelez l'inégalité de Bienaymé-Tchebychev. Comment utilise-t-on cette inégalité ? Est-elle toujours précise ?
- Qu'est-ce que le «phénomène du loto» ?
- Quel lien précis peut-il exister entre espérance et moyenne des valeurs mesurées d'une variable aléatoire ?
- L'espérance est-elle toujours une valeur typique de la variable considérée ?
- Quels sont les principales limites à l'utilisation de l'espérance en tant qu'indicateur de position ?
- L'écart-type représente-t-il effectivement un écart typique de la variable ? (Un écart entre quoi et quoi, au fait ?)
- Rappelez l'expression de l'espérance et de la variance des lois usuelles en fonction de leurs paramètres.
- Rappelez les liens entre opérations algébriques (somme, produit) sur les variables aléatoires, et opérations algébriques sur leur espérance et leur variance.
- Comment sont définies les variables aléatoires continues ? Quel lien possible avec les variables aléatoires de loi discrète ?
- Comment passe-t-on des définitions et propriétés relatives aux v.a. discrètes à leurs analogues dans le cas continu ?
- Qu'est-ce que la densité d'une variable aléatoire continue ?
- Qu'est-ce que la densité d'une variable aléatoire continue permet de calculer ?

- Comment définit-on l'espérance et la variance d'une variable aléatoire de loi continue ?
- Donnez la définition des lois gaussienne, exponentielle, et uniforme sur un intervalle, précisez leurs paramètres, le contexte exact (hypothèses sur le modèles, exemples concrets) dans lequel on sait qu'elles interviennent.
- Qu'appelle-t-on régression d'une variable sur une autre ?
- Qu'appelle-t-on régression linéaire ?
- Comment sont définis la covariance et le coefficient de corrélation de deux variables aléatoires ?
- Définissez le diagramme en bâtons, l'histogramme, le tracé de la fonction de répartition, le box-plot, associés à la loi d'une variable aléatoire.
- Connaissant la loi de X, comment calculer la loi de $h(X)$?
- Connaissant la loi de X et la loi de Y, peut-on calculer la loi de $X + Y$?

2.14 Exercices

Exercice 63 *Étant donnés deux événements A et B sur un espace des possibles Ω, comment exprimer à l'aide des fonctions indicatrices de A et B :*

- *la fonction indicatrice de $A \cup B$?*
- *la fonction indicatrice de $A \cap B$?*
- *la fonction indicatrice de A^c ?*
- *la fonction indicatrice de l'événement «A ou bien B» (ou exclusif) ?*

Exercice 64 *La Jojomobile dans laquelle roule Jojo est une véritable antiquité, et les déboires de Jojo avec son véhicule sont un sujet d'amusement permanent pour ses collègues. En particulier, les portières ont une fâcheuse tendance à s'ouvrir inopinément lorsque la voiture est en marche. Conducteur peu scrupuleux, Jojo ne s'arrête que lorsque la moitié au moins des portières s'ouvre. En admettant qu'au cours d'un trajet, le mécanisme de fermeture de chaque portière a une probabilité p de s'ouvrir, indépendamment des autres, précisez quelle est la loi du nombre de portières qui s'ouvrent au cours d'un trajet avec la Jojomobile. Quelle est la probabilité pour que Jojo s'interrompe au cours d'un trajet ? Lassé de s'arrêter aussi souvent, Jojo décide d'attacher avec du ruban adhésif la portière avant gauche à la portière arrière gauche, et la portière avant droite à la portière arrière droite. Grâce à cet ingénieux dispositif, une portière ne s'ouvre plus que lorsque les mécanismes de fermeture de cette portière et de celle à laquelle elle est attachée s'ouvrent simultanément. Quelle est à présent la probabilité pour que Jojo s'interrompe au cours d'un trajet ?*

Exercice 65 *Une information très secrète concernant le fonctionnement d'une nouvelle technologie militaire circule à travers une chaîne d'agents secrets et d'espions.*

En fait, chaque membre du réseau est plus ou moins un agent double et, pour des raisons qui lui sont propres, transmet l'information opposée à celle qu'il vient de recevoir avec une probabilité $p \in [0,1]$. *Seul le premier maillon de la chaîne sait effectivement si la technologie concernée est au point, et l'on suppose que les décisions prises par chacun des agents – transmettre correctement ou non l'information qu'ils ont reçue – sont mutuellement indépendantes. Calculez la probabilité* p_n *pour que l'information fournie par le* $n-$*ème maillon de la chaîne soit correcte. Trouver une relation de récurrence entre* p_n *et* p_{n+1}. *Que se passe-t-il lorsque* n *tend vers l'infini ?*

Exercice 66 *Pour agrémenter d'un peu de fantaisie son morne quotidien, un marchand de confiseries décide de piéger quelques unes des boîtes de chocolats de son étalage en y plaçant des pétards qui exploseront à l'ouverture de la boîte. Aujourd'hui, sur les 52 boîtes de chocolats disposées sur l'étalage, 4 sont piégées. Un client entre dans la boutique, choisit une boîte de chocolat au hasard (uniformément parmi les boîtes présentées), l'achète, et s'en va. Quelle est la probabilité qu'il emporte, sans le savoir, l'une des boîtes spécialement «arrangées» par notre facétieux confiseur ? Peu après, un deuxième client pénètre dans la boutique, choisit à son tour une boîte de chocolats, et l'emporte, après, naturellement, l'avoir payée. Quelle est la probabilité pour qu'il ait choisi une boîte de chocolats piégée ? Même question pour le troisième client, le quatrième, etc...*

Exercice 67 *Le jeune Dirichlet a placé sa paire de chaussettes bleues à pois mauves dans l'un des tiroirs de la commode mais, distrait, il a oublié de quel tiroir il s'agissait. Déterminé à retrouver coûte que coûte sa précieuse paire de chaussettes, il ouvre les tiroirs uniformément au hasard, les uns après les autres, en prenant bien soin de ne pas refermer les tiroirs déjà ouverts, jusqu'à remettre la main sur ses chaussettes. Décrivez un modèle probabiliste simple de la situation. Quelle est la loi du nombre de tiroirs qu'il lui faut ouvrir avant de retrouver ses chaussettes ? Quelle est son espérance ?*

Exercice 68 *Les quarante marins qui forment l'équipage du «Jojo des mers» descendent au port pour une nuit de beuverie. Au petit matin, complètement ivres, ils retournent sur le bateau, et chacun choisit une cabine au hasard parmi les quarante possibles, indépendamment de ses camarades. Quelle est la loi du nombre de marins qui dorment dans leur propre cabine ? Quelle est son espérance ? Et sa variance ?*

Exercice 69 *Chaque matin, Jojo consomme un grand bol de céréales avant de se rendre sur son lieu de travail. Dans chaque paquet de cérérales de la marque favorite de Jojo se trouve une vignette sur laquelle est représentée, au choix, l'une des sept images suivantes : un épi de maïs, une abeille, un vélo, un ours, une chouette, un mulot ou un bouc. Une fois en possession d'un exemplaire de chaque image, Jojo*

aura gagné un T-shirt portant le logo de la marque. Combien de paquets en moyenne Jojo devra-t-il acheter avant de pouvoir bénéficier de cette alléchante proposition ? (On admettra qu'il y a la même probabilité de trouver chacune des sept vignettes dans un paquet, et que les images présentes dans les différents paquets qu'achète Jojo sont mutuellement indépendantes.) Quelle est la variance de ce nombre de paquets ? Numérotons les vignettes de 1 à 7. En introduisant les événements $E(n,i)$ définis par : « Jojo n'a trouvé la vignette numéro i dans aucun des n premiers paquets », proposez une formule exacte pour la loi du nombre de paquets étudié, ainsi qu'une majoration de la probabilité pour que celui-ci dépasse une valeur fixée n.

Exercice 70 *À partir d'une suite de nombres aléatoires indépendants suivant la loi uniforme sur $[0,1]$, proposez une méthode pour générer :*

- *une suite de variables aléatoires suivant la loi uniforme sur $\{0,\dots,M\}$ où M est un entier positif,*
- *une suite de variables aléatoires de Bernoulli de probabilité $p \in [0,1]$,*
- *une variable aléatoire de loi binomiale de paramètres n et p,*
- *une variable aléatoire de loi de Poisson de paramètre λ,*
- *une variable aléatoire de loi géométrique de paramètre p.*

Exercice 71 *Considérons une loterie simplifiée, à laquelle participent un million d'individus. Chaque participant verse un euro de participation, et, après un tirage aléatoire uniforme parmi l'ensemble des participants, un seul des participants emporte le million d'euros ainsi collecté. Calculez l'espérance et l'écart-type du gain d'un individu donné participant à cette loterie. Sont-ils des indicateurs pertinents de de la loi de probabilité du gain ?*

Exercice 72 *(Dés de Sicherman)*

On considère deux dés cubiques dont les faces sont numérotées de 1 à 6. Quelle est, en supposant que, pour chaque dé, chaque face possède la même probabilité que les autres de sortir, et qu'il y a indépendance entre les deux dés, la loi de probabilité de la somme des chiffres obtenus après un lancer de ces deux dés ?

Même question avec deux dés cubiques dont les faces sont numérotées ainsi : $(1,2,2,3,3,4)$ *pour le premier dé, et* $(1,3,4,5,6,8)$ *pour le deuxième.*

Exercice 73 *(Dés d'Efron)*

On considère quatre dés cubiques notés A, B, C, D.

Chaque face de chaque dé porte un nombre entier. Voici précisément, pour chaque dé, la liste de ces nombres.

- *A :* $(0,0,4,4,4,4)$
- *B :* $(3,3,3,3,3,3)$
- *C :* $(2,2,2,6,6,6)$

– D : $(1, 1, 1, 5, 5, 5)$

On considère à présent le jeu à deux joueurs suivant. Un premier joueur choisit l'un des quatre dés ci-dessus, et le deuxième joueur choisit un dé parmi les trois restants. Chacun lance ensuite son dé, et le joueur ayant obtenu le plus grand chiffre gagne la partie. Comment conseilleriez-vous aux deux joueurs de choisir leurs dés ?

Exercice 74 *On considère une urne contenant m objets distincts, numérotés de 1 à m, et l'on suppose que l'on effectue n tirages successifs sans remise parmi ces m objets (on suppose donc que $n \leq m$), chaque objet étant tiré uniformément au hasard parmi les objets restants dans l'urne au moment du tirage.*

Pour tout $1 \leq i \leq n$, appelons X_i le numéro de l'objet tiré de l'urne lors du $i-$ème tirage.

Prouvez que, pour toute permutation σ de l'ensemble $\{1, \ldots, n\}$, la loi de

$$(X_{\sigma(1)}, X_\sigma(2), \ldots, X_\sigma(n))$$

est la même que celle de $(X_1, \ldots, X_n)$. Quelle est cette loi ? Déduisez-en le fait que, si $I = \{i_1, \ldots, i_k\}$ et $J = \{j_1, \ldots, j_k\}$ sont deux sous-ensembles de $\{1, \ldots, n\}$ comportant chacun k éléments, $(X_{i_1}, \ldots, X_{i_k})$ et $(X_{j_1}, \ldots, X_{j_k})$ possèdent la même loi. Quel résultat obtient-on en spécialisant ce résultat au cas d'ensembles comportant un seul élément ?

Comment ce résultat peut-il s'appliquer dans les exercices 9, 66, 86, 67, 124, 36 ?

Exercice 75 *Jojo souhaite générer une suite aléatoire de 0 et de 1, indépendants, et distribués selon la loi uniforme sur $\{0, 1\}$. En fouillant dans sa poche, il n'a trouvé qu'une pièce de monnaie très usée dont la symétrie lui paraît douteuse, si bien qu'il pense que la probabilité p d'obtenir «pile» en lançant sa pièce n'est pas égale à $1/2$. Comment peut-il s'y prendre, sans connaître la valeur de p, pour générer tout de même la suite de valeurs dont il a besoin. Indication : en effectuant deux lancers successifs, quelle est la probabilité d'obtenir «pile» suivi de face. Et «face» suivi de «pile» ?*

Exercice 76 *Albéric est un cambrioleur malheureux, spécialisé dans les bijouteries. À chaque tentative de cambriolage d'une bijouterie, il échoue avec une probabilité de 80%. Appelons X_n le nombre de cambriolages réussis après n tentatives (avec par convention $X_0 = 0$). Quelle est la loi de X_n ? Quelle est son espérance ? Appelons T le nombre de tentatives nécessaires avant de réussir un cambriolage. Quelle est la loi de T ? Quelle est son espérance ? Quelle est la loi de X_T ? Son espérance ? Quelle est la loi de X_{T-1} ? Son espérance ?*

Exercice 77 *Comparez l'espérance et la variance des deux variables aléatoires X et Y dont les deux lois sont définies par :*

$$\mathbb{P}(X=0)=14/30\ ,\ \mathbb{P}(X=1)=3/30\ ,\ \mathbb{P}(X=2)=12/30\ ,\ \mathbb{P}(X=3)=1/30,$$

et

$$\mathbb{P}(Y=0)=13/30\ ,\ \mathbb{P}(Y=1)=6/30\ ,\ \mathbb{P}(Y=2)=9/30\ ,\ \mathbb{P}(Y=3)=2/30.$$

Exercice 78 *(Violation de l'inégalité de Bell)*

On considère quatre variables aléatoires X_1, X_2, Y_1, Y_2 définies sur un même espace de probabilité $(\Omega, \mathbb{P})$, chaque variable ne pouvant prendre que les valeurs 0 ou 1. Prouvez que :

$$(X_1 \neq Y_1 \text{ et } Y_1 \neq X_2 \text{ et } X_2 \neq Y_2) \Rightarrow X_1 \neq Y_2.$$

En déduire l'inégalité de Bell :

$$\mathbb{P}(X_1 = Y_2) \leq \mathbb{P}(X_1 = Y_1) + \mathbb{P}(Y_1 = X_2) + \mathbb{P}(X_2 = Y_2).$$

Application. On s'intéresse à l'expérience de physique suivante : un atome de calcium est excité par un faisceau laser, et, en retournant à son état non-excité, émet une paire de photons qui partent dans deux directions opposées $\vec{x}$ et $\vec{y}$. Chacun des deux photons est intercepté par un filtre de polarisation, qu'il peut ou non traverser, un photo-détecteur étant placé au-delà de chaque filtre pour déterminer si le photon a effectivement traversé. L'événement auquel on s'intéresse est la coïncidence des comportements des deux photons : traversent les filtres tous les deux, ou restent bloqués tous les deux. Chacun des deux filtres peut être réglé selon deux positions différentes notées 1 et 2, ce qui donne lieu à quatre dispositifs expérimentaux différents, notés (i, j), i désignant la position sur laquelle est réglée le filtre situé dans la direction $\vec{x}$, j la position du filtre situé dans la direction $\vec{y}$. La mécanique quantique prédit, et ceci est confirmé par l'expérience, que, pour un choix conven able des différentes positions que peuvent prendre les filtres, les probabilités d'observer un comportement identique des deux photons sont données, pour chacun des dispositifs expérimentaux, par :

$$p_{(1,2)} = 0,85\ ;\ p_{(2,2)} = 0,15\ ;\ p_{(1,1)} = 0,15\ ;\ p_{(2,1)} = 0,15.$$

On désire modéliser l'état du système formé par la paire de photons à l'aide d'un espace de probabilité $(\Omega, \mathbb{P})$. Sur cet espace de probabilité sont définies quatre variables aléatoires X_1, X_2, Y_1, Y_2, qui indiquent si l'état du système autorise ou non les photons à traverser les filtres, suivant les positions de ceux-ci. Ainsi, X_1 prend la valeur 1 lorsque l'état du système permet au photon parti dans la direction $\vec{x}$ de traverser le filtre si celui-ci est placé en position 1, et prend la valeur 0 dans le cas contraire. De même, X_2 prend la valeur 1 lorsque l'état du système permet au photon parti dans la direction $\vec{x}$ de traverser le filtre si celui-ci est placé en position 2, et

prend la valeur 0 *dans le cas contraire. On définit de même* Y_1 *et* Y_2 *pour le photon parti dans la direction* $\vec{y}$.

Comment les probabilités $p_{(i,j)}$ *s'expriment-elles à l'aide de* X_1, X_2, Y_1, Y_2 *? Qu'en concluez-vous ?*

Exercice 79 *(Tri rapide randomisé)*

On désire trier par ordre croissant une liste de n *nombres* $x_1, \ldots, x_n$. *On utilise pour cela l'algorithme de tri rapide randomisé suivant : on choisit au hasard (uniformément) un indice* I *parmi* $1, 2, \ldots, n$, *et l'on compare* x_I *(le pivot) aux autres nombres de la liste, de façon à former deux sous-listes, contenant respectivement les éléments de la liste inférieurs (ou égaux) à* x_I, *et les éléments strictement supérieurs à* x_I. *On applique ensuite récursivement la méthode à ces deux sous-listes, et l'algorithme s'arrête lorsque les sous-listes à trier ne contiennent plus chacune qu'un seul élément. On peut représenter le résultat de l'exécution de l'algorithme par un arbre binaire étiqueté, sur les noeuds duquel figurent les éléments de la liste. La racine est étiquetée par* x_I, *son descendant gauche est étiqueté par l'élément choisi pour le tri de la sous-liste des éléments supérieurs à* x_I, *son descendant droit par l'élément choisi pour le tri de la sous-liste des éléments inférieurs à* x_I, *et ainsi de suite (il n'y a pas de descendant associé à une sous-liste vide). Comment utiliser cet arbre pour produire la liste* $x_1, \ldots, x_n$ *ordonnée ? Effectuez vous-même (à la main) le tri de la liste* $2; 5; 4; 7; 1; 6; 3; 9; 8$ *à deux reprises selon l'algorithme indiqué. Obtenez-vous le même arbre ? On s'intéresse au temps d'exécution (aléatoire) de l'algorithme, tel que mesuré par le nombre total de comparaisons effectuées. Quel est le pire cas possible ? Montrez que, pour une paire d'indices* $\{i, j\}$ *donnée, l'algorithme effectue au plus une comparaison entre* x_i *et* x_j *lors de son exécution. Le temps d'exécution de l'algorithme est donc défini par :*

$$T = \sum_{1 \leq i < j \leq n} \mathbf{1}_{\{x_i \text{ et } x_j \text{ sont comparés}\}}.$$

Notons $y_1, \ldots, y_n$ *la liste* $x_1, \ldots, x_n$ *ordonnée de façon croissante. À quelle condition l'algorithme effectue-t-il une comparaison entre* y_i *et* y_j *lors de son déroulement ? En déduire la probabilité :*

$$\mathbb{P}(y_i \text{ et } y_j \text{ sont comparés}).$$

En déduire finalement la valeur de l'espérance de T, $\mathbb{E}(T)$. *Cette espérance dépend-elle de la liste initiale ? Comment se comporte-t-elle lorsque la taille* n *de la liste à trier tend vers l'infini ? L'appellation de «tri rapide» vous semble-t-elle justifiée ?*

Exercice 80 *(Aversion au risque)*

On vous propose de participer à une loterie, en vous laissant le choix entre deux modalités :

– *modalité 1 : gain de 1000 euros avec probabilité 1/2, gain de 0 euro avec probabilité 1/2*
– *modalité 2 : gain de x euros avec probabilité 1.*

A partir de quelle valeur de x préférez-vous l'option 2 à l'option 1 ? Quelle valeur accepteriez-vous de payer un billet de loterie vous donnant droit à l'option 2 ?

Exercice 81 *(Paradoxe d'Allais)*

Un généreux membre de votre famille ayant fait fortune grâce aux jeux de hasard, et désireux de vous faire un petit cadeau, vous propose le choix (choix 1) entre les deux options suivantes :

– *option 1 : il vous offre 1000 euros ;*
– *option 2 : il vous offre un billet de loterie donnant 10% de chances de gagner 5000 euros, 89% de chances de gagner 1000 euros, et 1% de chances de ne rien gagner du tout.*

Il souligne par ailleurs que le billet de loterie qui vous est proposé est vendu plus de 1000 euros.

Laquelle des deux options choisiriez-vous ?

Même question avec le choix (choix 2) suivant :

– *option 1 : un billet de loterie qui vous donne 11% de chances de gagner 1000 euros et 89% de chances de ne rien gagner du tout ;*
– *option 2 : un billet de loterie qui vous donne 10% de chances de gagner 5000 euros et 90% de chances de ne rien gagner du tout.*

Même question avec le choix (choix 3) suivant :

– *option 1 : il vous offre 1000 euros ;*
– *option 2 : un billet de loterie qui vous donne* $10/11$ *chances de gagner 5000 euros et* $1 - 10/11$ *de chances de ne rien gagner du tout.*

Comment vos réactions se confrontent-elles aux arguments suivants ?

– *si l'on est cohérent, on doit choisir la même option (1 ou 2) lors du deuxième et du troisième choix, car l'option 1 du deuxième choix correspond à 11% de chances de gagner l'option 1 du troisième choix, et de ne rien gagner sinon, tandis que l'option 2 du deuxième choix correspond à 11% de chances de gagner l'option 2 du troisième choix, et de ne rien gagner sinon ;*
– *si l'on est cohérent, on doit choisir la même option (1 ou 2) lors du premier et du troisième choix, car l'option 1 du choix 1 correspond à 11% de chances de gagner l'option 1 du choix 3 avec un lot de consolation en cas de perte de 1000 euros (tiré par les cheveux, hein ?), tandis que l'option 2 du choix du choix 1 correspond à 11% de chances de gagner l'option 2 du choix 3 avec un lot de consolation en cas de perte de 1000 euros.*

Exercice 82 *Vous jouez au jeu des devinettes avec votre petit cousin. Le principe du jeu consiste à choisir un nombre entre 1 et 6, et à le faire deviner par l'autre*

en répondant les unes après les autres à ses questions, qui sont du type «le nombre figure-t-il dans A ?», où A est un sous-ensemble de $\{1,2,3,4,5,6\}$.

A force de jouer, vous avez fini par attribuer à chaque entier i entre 1 et 6 une probabilité p_i d'être choisi par votre petit cousin :

$$(p_1, p_2, \ldots, p_6) = (6/30, 7/30, 4/30, 5/30, 3/30, 5/30)$$

On cherche à deviner le nombre choisi en posant, en moyenne, le moins de questions possibles.

- *première méthode : on ne pose que des questions de la forme «le nombre est-il égal à i» en épuisant les différentes possibilités les unes après les autres. Quel est l'ordre qui minimise le nombre moyen de questions à poser ? Quelle est la valeur de ce nombre moyen ?*
- *deuxième méthode : on pose la question «le nombre est-il égal à 1 ou à 2». Si la réponse est oui, on demande s'il est égal à 1. Si la réponse est non, on teste les possibilités restantes comme dans la méthode précédente. Quel est le nombre moyen de questions requis par cette stratégie ?*
- *en utilisant l'algorithme de Huffman, vous pouvez déterminer une méthode optimale. Quelle est le nombre moyen de questions qu'elle requiert ? Comment ce nombre se compare-t-il à l'entropie de la probabilité $(p_i)_{1\leq i\leq 6}$?*

Exercice 83 *Jojo vit dans une région où les risques sismiques, quoique faibles, sont loin d'être négligeables. L'immeuble dans lequel vit Jojo a été conçu pour résister à des secousses dont la magnitude ne dépasse pas 20 (sur l'échelle de Jojo). Des études expérimentales menées sur plusieurs années ont permis de conclure que des séismes de faible amplitude survenaient une fois par an, avec une magnitude moyenne de 10 et une variance de 4 (sur l'échelle de Jojo). Jojo peut-il dormir tranquille ? Pour combien de temps ? (Justifiez.)*

Exercice 84 *Une séquence génétique se présente comme une suite de lettres de l'alphabet $\{A, C, G, T\}$. On modélise de façon très simplifiée une séquence génomique de longueur 1000 issue du génôme d'un individu comme une suite $X_1, X_2, \ldots, X_{1000}$ de variables aléatoires indépendantes, les probabilités attribuées à chaque lettre étant données par :*

$$\mathbb{P}(X_i = A) = 0,4\ ;\ \mathbb{P}(X_i = C) = 0,2\ ;\mathbb{P}(X_i = G) = 0,1;\ \mathbb{P}(X_i = T) = 0,3.$$

On appelle N_A le nombre de A apparaissant dans la séquence $X_1, \ldots, X_{1000}$. Pouvez-vous déterminer la loi de N_A ? De même, on appelle N_C le nombre de C apparaissant dans la séquence. Quelle est la loi de N_C ? Comment le nombre total de A et de C apparaissant dans la séquence s'exprime-t-il en fonction de N_A et de N_C ? Pouvez-vous calculer son espérance ? Sa variance ? Appelons S l'indice du premier A

présent dans la séquence (le plus petit indice i tel que $X_i = A$). (Si la séquence ne comporte aucun A, on pose par convention $S = 1001$.) Quelle est la loi de S ? Lorsque la séquence comporte au moins un A, le A situé en position S est la première lettre d'une suite ininterrompue de A (éventuellement réduite à un seul A). Par exemple, dans la séquence :

$$CGCTGTAAAGCTC...$$

on a $S = 7$, et le A situé en position S est la première lettre d'une suite de A ininterrompue de longeur 3. Appelons U la position du A situé le plus à droite dans la séquence ininterrompue de A commençant en S (par convention, on pose $U = 1001$ lorsque la séquence ne comporte pas de A). Sur l'exemple ci-dessus, on a donc $U = 9$. Quelle est (de manière générale) la probabilité pour que $U = 30$?

Exercice 85 *Chouette ! En râclant le fond de ses poches, Jojo vient de trouver 100 euros. Heureuse coïncidence, son ami Pierrot vient de lui proposer d'investir de l'argent dans une affaire commerciale qui s'annonce, affirme-t-il, très lucrative. Jojo sera rémunéré à hauteur de son investissement : s'il investit x euros, il recevra au bout d'un an $x \times (1 + L)$ euros, L désignant le taux (aléatoire) de rentabilité de l'affaire. Jojo hésite alors entre deux stratégies. La première consiste simplement à investir ses 100 euros dans l'affaire proposée. La seconde, plus complexe, consiste d'abord à emprunter 10000 euros à la banque, qu'il devra rembourser au bout d'un an, en donnant immédiatement ses 100 euros à titre d'intérêts, et à investir dans l'affaire les 10000 euros empruntés. Discutez les avantages et les inconvénients de ces deux stratégies, notamment les risques de perte et les perspectives de gain des deux stratégies, en fonction des propriétés de L (avec des arguments précis, bien entendu).*

Exercice 86 *Un certain soir, Jojo reçoit dix de ses amis chez lui. En fin de soirée, après un repas bien arrosé, ceux-ci ne sont plus en état de retrouver leur chapeau parmi ceux des autres, et s'en retournent donc chez eux (en taxi) après avoir choisi au hasard l'un des dix chapeaux en présence. On s'intéresse au nombre X des amis de Jojo ayant effectivement retrouvé leur propre chapeau. Décrivez précisément la modélisation du tirage aléatoire des chapeaux par les invités que vous allez adopter (indication : une affectation des chapeaux aux invités peut, par exemple, se représenter par une permutation des entiers de 1 à 10). On définit les variables aléatoires $X_1, X_2, \ldots, X_{10}$ par :*

$$X_i = \mathbf{1}_{\{\textit{l'invité numéro } i \textit{ a retrouvé son chapeau}\}}.$$

- *exprimez X en fonction des X_i ;*
- *calculez l'espérance de X ;*
- *calculez la variance de X ;*
- *donnez une formule explicite pour la loi de X (commencer par $\mathbb{P}(X = 0)$).*

Exercice 87 *Lorsqu'il télécharge des documents sur internet, Jojo a pour habitude d'interrompre le chargement lorsque la durée de celui-ci dépasse une minute. Son idée est qu'une durée de téléchargement anormalement longue (supérieure à une minute) est le signe probable d'un problème technique ralentissant considérablement le téléchargement, et rendant donc inutile le fait d'attendre une ou deux minutes supplémentaires. Il préfère donc, dans le but de gagner du temps, abandonner le chargement en cours, et retenter un nouveau téléchargement quelques dizaines de minutes plus tard. Cette idée est-elle compatible avec la modélisation de la durée totale de téléchargement d'un fichier sans interruption ni nouvelle tentative (en secondes par exemple) à l'aide d'une loi géométrique de paramètre fixé p ? (Argumentez votre réponse, en comparant, par exemple, la méthode de Jojo avec celle qui consisterait simplement à attendre le chargement complet d'un fichier, sans interrompre celui-ci au bout d'une minute.)*

Exercice 88 *Ce soir, Jojo joue aux échecs avec son ami Horace. Du moins le croît-il. En effet, Horace est un joueur de niveau assez moyen, mais il demande parfois à son frère jumeau, Hyacinthe, excellent joueur, de le remplacer, sans que personne ne puisse s'apercevoir de l'imposture. Lorsqu'il joue contre Horace, Jojo a une probabilité de 0,5 de l'emporter. En revanche, lorsqu'il joue contre Hyacinthe, cette probabilité chute à 0,2. Après cinq parties jouées, Jojo en a déjà perdu trois, et, de mauvaise humeur, commence à maugréer qu'il se trouve probablement en face de Hyacinthe et non de son frère. Pouvez-vous lui donner raison ?*

La raison de la mauvaise humeur de Jojo est que, un peu présomptueux, celui-ci a parié un repas au restaurant avec Horace qu'il remporterait au moins quatre parties sur les sept que ceux-ci projetaient de jouer ce soir (dont les cinq premières ont donc déjà été jouées). Jojo propose un arrangement : les deux parties restantes ne seront pas jouées, et les deux amis se répartiront la note du restaurant équitablement au vu du résultat des cinq premières parties. Quelle est selon vous cette répartition «équitable» ?

Exercice 89 *(Le jeu de la Belle au bois dormant)*

Belle dort en permanence, sauf peut-être le mardi et le mercredi, où l'équipe organisant le jeu a la possibilité de la réveiller pour quelques instants. Plus précisément, chaque lundi, l'équipe procède au lancer d'une pièce de monnaie équilibrée. Si pile est obtenu, Belle est réveillée le mardi (et on la laisse dormir dans le cas contraire). De plus, Belle est réveillée chaque mercredi quel que soit le résultat du lancer. À chaque réveil, on demande à Belle de parier sur le résultat (pile ou face) du lancer de la semaine en cours. Belle ignore totalement la date courante, et en particulier le jour de la semaine, car elle perd tous ses souvenirs à chaque fois qu'elle s'endort. Comment conseilleriez-vous à Belle de parier ? Sachant que Belle gagne 50 euros à

chaque pari gagné, de quelle somme devrait-elle disposer après 6 mois si elle applique votre méthode ?

Exercice 90 *Combien obtient-on en moyenne de 6 en lançant 4 fois un dé ? Et de paires de 6 en lançant 24 fois deux dés ? Est-il plus probable d'obtenir un 6 en lançant 4 fois un dé que d'obtenir une paire de 6 en lançant 24 fois deux dés ? Ces résultats sont-ils cohérents ?*

Exercice 91 *Montrez que toute variable aléatoire ne pouvant prendre qu'un nombre fini de valeurs distinctes peut se mettre sous la forme d'une combinaison linéaire de fonctions indicatrices.*

Exercice 92 *On dispose de $n+1$ urnes $U_0, \ldots, U_n$ contenant chacune N boules dont certaines sont rouges et d'autres blanches. Pour tout $0 \leq i \leq n$, l'urne U_i contient une proportion i/n de boules rouges, et $1 - i/n$ de boules blanches. On choisit une urne uniformément au hasard parmi $U_0, \ldots, U_n$, et l'on effectue des tirages successifs avec remise (uniformes et indépendants) dans l'urne choisie. Si l'on n'a obtenu que des boules rouges au cours des p premiers tirages, comment évaluer la probabilité d'obtenir une boule rouge au $p+1-$ème tirage ? Que se passe-t-il lorsque n tend vers l'infini ? Le résultat est appelé loi de succession de Laplace*[14] *Application (quelque peu tirée par les cheveux) : sachant que le Soleil s'est levé chaque matin au cours des 2000 dernières années, quelle est la probabilité pour que celui-ci se lève demain ?*

Exercice 93 *Montrez que, si Ω est un ensemble fini, il existe une seule fonction h associant à toute probabilité $\mathbb{P}$ sur Ω et toute variable aléatoire X à valeurs réelles et définie sur Ω, un nombre réel $h(X, \mathbb{P})$ vérifiant les conditions suivantes :*

- *si X et Y sont deux variables aléatoires sur $(\Omega, \mathbb{P})$ vérifiant $\mathbb{P}(X \geq Y) = 1$, alors $h(X, \mathbb{P}) \leq h(Y, \mathbb{P})$ (positivité)*
- *si $\lambda \in \mathbb{R}$ est un réel fixé, $h(\lambda X, \mathbb{P}) = \lambda h(X, \mathbb{P})$ (invariance par changement d'échelle) ;*
- *si $c \in \mathbb{R}$ est un réel fixé, $h(X+c, \mathbb{P}) = h(X, \mathbb{P}) + c$ (invariance par translation) ;*
- *si X et Y sont deux variables aléatoires sur $(\Omega, \mathbb{P})$, $h(X+Y, \mathbb{P}) = h(X, \mathbb{P}) + h(Y, \mathbb{P})$;*
- *$h(X, \mathbb{P})$ ne dépend que de la loi de X.*

En conclure que cette fonction vérifie nécessairement $h(X, \mathbb{P}) = \mathbb{E}(X)$.

Quelles sont les propriétés ci-dessus que vérifient ou ne vérifient pas la médiane, le mode, et le milieu du domaine ?

Exercice 94 *On se donne une variable aléatoire X possédant une espérance et une variance. Prouvez que $\mathbb{E}(X)$ est l'unique minimum de la fonction définie sur $\mathbb{R}$ par $a \mapsto \mathbb{E}(X-a)^2$.*

14. Pierre-Simon Laplace (1749–1827).

De même, montrez que l'ensemble des points où la fonction définie sur $\mathbb{R}$ *par* $a \mapsto \mathbb{E}\,|X-a|$ *atteint son minimum est l'intervalle médian. Enfin, montrez que, si* X *est bornée, l'ensemble des points où la fonction définie sur* $\mathbb{R}$ *par* $a \mapsto \sup|X-a|$ *atteint son minimum est le milieu du domaine de* X*.*

Exercice 95 *Considérons une variable aléatoire* X *à valeurs réelles, définie sur un modèle probabiliste* $(\Omega, \mathbb{P})$*. Introduisons le modèle* $(\Omega^2, \mathbb{P}^{\otimes 2})$ *correspondant à deux répétitions indépendantes de* $(\Omega, \mathbb{P})$*. Définissons* $X_1(\omega_1, \omega_2) = X(\omega_1)$ *et* $X_2(\omega_1, \omega_2) = X(\omega_2)$*. On vérifie ainsi que* X_1 *et* X_2 *sont deux variables aléatoires indépendantes, et que* X_1 *et* X_2 *suivent chacune individuellement la loi de* X*. Prouvez que* $2\mathbb{V}(X) = \mathbb{E}\left([X_1 - X_2]^2\right)$*. En d'autres termes, sans référence explicite à l'espérance de* X*, la variance de* X *peut être vue comme une mesure de la variation existant entre deux variables aléatoires indépendantes de même loi que* X*.*

Exercice 96 *Etant données deux variables aléatoires* X *et* Y *possédant une espérance et une variance, vérifiez que le coefficient de corrélation de* X *et de* Y *est toujours compris entre* -1 *et* 1*. Caractérisez l'égalité à* 1 *et à* -1*. Calculez, après avoir prouvé leur existence, les réels* a *et* b *qui minimisent* $\mathbb{E}\left([Y - aX - b]^2\right)$*.*

Exercice 97 *Deux amis, Amédée et Basile jouent au jeu suivant. Amédée pense à deux nombres réels distincts, choisit à pile ou face l'un de ces deux nombres et le communique à Basile. Basile, de son côté, doit tenter de deviner si le nombre qui lui a été communiqué est le plus grand ou le plus petit des deux auxquels Amédée a pensé. Il ne semble guère possible de faire mieux en toute généralité que de répondre en tirant à pile ou face, avec exactement une chance sur deux de gagner. Et pourtant... Supposons que Basile s'aide en générant une variable aléatoire réelle* X*, de loi continue, possédant une densité strictement positive sur* $\mathbb{R}$ *tout entier, et réponde à Amédée de la manière suivante. Lorsque le nombre communiqué par Amédée est inférieur à la valeur de* X*, Basile parie sur le fait que ce nombre est le plus petit des deux, et, réciproquement, lorsque le nombre communiqué est supérieur à la valeur de* X*, Basile parie sur le fait que ce nombre est le plus grand des deux. Montrez qu'ainsi Basile possède strictement plus d'une chance sur deux de gagner. Discutez ce résultat.*

Exercice 98 *L'entreprise «Jojo16i» propose des services de saisie informatisée de documents manuscrits. Chaque document est saisi par deux dactylographes différentes, et les deux versions sont ensuite comparées automatiquement afin de déceler d'éventuelles discordances. Quel est l'avantage de cette méthode par rapport à une saisie simple ? Précisez ceci en admettant que, par exemple, chaque dactylographe a une probabilité d'environ 0,3% de se tromper lors de la saisie d'une entrée, et que les documents saisis comportent en général de l'ordre de 10000 entrées.*

Exercice 99 *On appelle équation de Drake l'égalité*

$$N = R^* \times f_p \times n_e \times f_l \times f_i \times f_c \times L,$$

où :

- *N est le nombre de civilisations extra-terrestres présentes dans notre galaxie et avec lesquelles nous pourrions nous attendre à pouvoir communiquer ;*
- *R^* est le taux de formation d'étoiles dans notre galaxie ;*
- *f_p est la proportion de ces étoiles possédant des planètes ;*
- *n_e est le nombre moyen de planètes susceptibles d'abriter la vie rapporté au nombre d'étoiles possédant des planètes ;*
- *f_l est la fraction des planètes ci-dessus qui vont réellement voir la vie se développer ;*
- *f_i est la proportion d'entre elles qui vont voir une civilisation intelligente se développer ;*
- *f_c est la fraction des civilisations ci-dessus qui sont désireuses de communiquer et capables de le faire ;*
- *L est la durée moyenne d'existence d'une telle civilisation.*

Sur quels présupposés et approximations cette équation repose-t-elle ? Comment pourrait-on tenter d'évaluer les différents termes apparaissant dans l'équation ? Pourquoi cette équation comporte-t-elle un produit de 7 termes et non pas 8 ou 6 ? Peut-on imaginer d'autres équations visant à estimer N ?

Exercice 100 *Un arbre de jeu est un arbre fini enraciné, dont les noeuds à distance paire de la racine sont étiquetés MIN et les noeuds à distance impaire sont étiquetés MAX. A chaque feuille de l'arbre est associée la valeur 0 ou 1. L'évaluation de l'arbre consiste à attribuer itérativement une valeur à chaque noeud de l'arbre, en partant des feuilles, de la manière suivante : la valeur associée à un noeud étiqueté MIN est le minimum des valeurs associées à ses enfants, et la valeur associée à un noeud étiqueté MAX en est le maximum.*
1) Concrètement, un tel arbre représente le déroulement d'un jeu à deux joueurs, dans lequel chacun des deux joueurs joue à son tour, les ramifications de l'arbre représentant, à chaque étape, les différentes possibilités offertes au joueur dont c'est le tour de jouer. Les feuilles de l'arbre correspondent aux fins de partie, et sont étiquetées 0 lorsque la partie s'est soldée par une victoire du joueur ayant joué le premier coup, et 1 dans le cas d'une victoire du joueur ayant joué le deuxième coup (on suppose qu'il n'y a pas de nul possible, et qu'une partie doit toujours se terminer). Que traduit l'évaluation de l'arbre, et en particulier la valeur attribuée à la racine ? Comment modifier ce modèle pour prendre en compte la possibilité d'un match nul ?

Dans la suite, on se place dans le cas particulier d'un arbre binaire régulier de profondeur $n \geq 2$ fixée.

2) Est-il toujours nécessaire de prendre en compte la valeur de toutes les feuilles pour calculer la valeur de la racine ou peut-on parfois en ignorer certaines ?
3) On considère maintenant des algorithmes déterministes (i.e. non-randomisés) permettant de calculer l'étiquette attachée à la racine à partir de la lecture de tout ou partie des étiquettes attachées aux feuilles. Plus précisément, un algorithme déterministe d'évaluation de l'arbre fonctionne de la manière suivante. Il commence par spécifier une feuille de l'arbre, dont la valeur est lue. Ensuite, à chaque étape, une nouvelle feuille est spécifiée en fonction des résultats obtenus au cours des étapes précédentes, et sa valeur est lue à son tour. L'algorithme s'arrête lorsque les valeurs qu'il a lues lui permettent de déterminer l'étiquette attachée à la racine.

Montrez (par exemple par récurrence) qu'il est toujours possible de trouver une affectation de 0 et de 1 aux feuilles de l'arbre qui force un tel algorithme à lire toutes les feuilles de l'arbre avant de pouvoir déterminer la valeur de la racine.
4) On considère un algorithme randomisé fonctionnant de la manière suivante : pour évaluer un noeud MIN, l'algorithme choisit au hasard avec probabilité 1/2 l'un de ses deux descendants, qui est lui-même évalué en faisant appel à l'algorithme de manière récursive. Si celui-ci a pour valeur 0, la valeur du noeud MIN est donc déterminée et est égale à 0. Si le descendant a pour valeur 1, on évalue l'autre descendant de la même manière. Dans le cas d'un noeud MAX, on procède suivant le même principe, à ceci près que la valeur du noeud est déterminée par son premier desccendant lorsque celui-ci a pour valeur 1. Prouvez que, pour toute affectation des valeurs des feuilles, le nombre moyen de feuilles lues par cet algorithme est inférieur ou égal à 3^k. Comment ce temps moyen se compare-t-il au pire cas ?

Exercice 101 *Afin de déterminer le nombre moyen d'enfants par famille, on sonde un grand nombre d'enfants en leur demandant combien ils possèdent de frères et de sœurs (y compris eux-mêmes). En faisant la moyenne des valeurs obtenues, on obtient un nombre bien supérieur à 2, qui est pourtant approximativement la valeur correcte. Que s'est-il passé ?*

En admettant que le nombre moyen d'enfants par famille soit égal à 2,2, peut-on en déduire que la population devrait augmenter au cours des prochaines années ?

Exercice 102 *(Froeppel)*

Soient A et B deux points à la même distance l'un de l'autre. Comment déplacer B sans que A s'en aperçoive ?

Exercice 103 *Soit n un nombre premier, et $\mathbb{Z}/n\mathbb{Z}$ l'ensemble des (classes de congruence d') entiers modulo n. On part de deux variables aléatoires A et B à valeurs dans $\mathbb{Z}/n\mathbb{Z}$, indépendantes et de loi uniforme. Pour tout $1 \leq i \leq n$, on définit $Y_i = Ai + B \bmod n$. Montrez que Y_i suit la loi uniforme sur $\mathbb{Z}/n\mathbb{Z}$, et que, pour tout*

couple $1 \leq i \neq j \leq n$, Y_i *et* Y_j *sont indépendantes. Les variables aléatoires* $Y_1, \ldots, Y_n$ *sont-elles mutuellement indépendantes ?*

Exercice 104 *Pour tester une certaine propriété* P *pouvant ou non être vérifiée par un objet* x*, on suppose que l'on dispose d'un algorithme randomisé prenant en entrée* x *ainsi qu'un entier uniformément choisi entre* 1 *et* n*,* n *étant un entier premier. Si* x *vérifie effectivement la propriété* P*, l'algorithme répond toujours que* P *est vérifiée. En revanche, si* P *n'est pas vérifiée, tout ce que l'on sait est que la probabilité pour que l'algorithme réponde que* P *n'est pas vérifiée est supérieure ou égale à* $1/2$*. On suppose que* n *est trop grand pour qu'il soit rentable de tester la totalité des entiers compris entre* 1 *et* n *(ce qui permettrait de décider de manière certaine si* x *possède ou non la propriété). En utilisant* r *répétitions indépendantes de son algorithme, Jojo parvient à diminuer la probabilité d'erreur à* 2^{-r} *au pire (voir l'exercice 23). Combien de bits aléatoires (i.e. de v.a. de Bernoulli indépendantes symétriques) faut-il pour générer* r *exécutions de l'algorithme ? Si l'on utilise à la place la méthode de l'exercice 103 pour générer les* r *(supposé inférieur à* n*) nombres aléatoires de loi uniforme sur* $\{1, 2, \ldots, n\}$ *nécessaires aux* r *exécutions successives de l'algorithme, à combien ce nombre passe-t-il ? Que peut-on dire alors de la probabilité d'erreur ?*

Exercice 105 *Dans un pays dont nous tairons le nom, les préjugés sexistes sont tels que la plupart des femmes planifient ainsi les naissances de leurs enfants : donner naissance à des enfants jusqu'à obtenir un garçon ou quatre enfants. D'après vous, cette attitude a-t-elle plutôt tendance à augmenter ou à diminuer la proportion de filles parmi les naissances ? Montrez qu'il en est de même de toute stratégie de planification des naissances dans lesquelles la décision d'arrêter ou de continuer d'avoir des enfants est prise en fonction des naissances précédentes, et pour lesquelles le nombre maximum d'enfants ne peut pas dépasser une certaine limite.*

Qu'en est-il de la stratégie suivante : continuer d'avoir des enfants jusqu'à ce que le nombre de garçons dépasse d'au moins un le nombre de filles (sans restriction sur le nombre total d'enfants) ?

Exercice 106 *On désire envoyer un message* A *à travers un système de communication qui ne peut acheminer qu'un seul message à la fois. A chaque seconde, le système peut être occupé par la transmission d'un autre message que* A*, et ceci indépendamment chaque seconde, avec une probabilité* p*.*
1) Le message A *que l'on souhaite envoyer nécessite une seconde de transmission. Quelle est la loi de la variable aléatoire* T_1 *donnant le temps d'attente nécessaire avant que le message* A *ait fini d'être transmis ?*
2) Cette fois, le message A *nécessite deux secondes consécutives pour être correctement transmis. Appelons* T_2 *le temps d'attente nécessaire avant que* A *ait fini d'être transmis. Proposez une borne supérieure simple sur* $\mathbb{P}(T_2 > n)$*.*

Montrez que, pour tout $n \geq 2$,

$$\mathbb{P}(T_2 > n) = p\mathbb{P}(T_2 > n-1) + (1-p)p\mathbb{P}(T_2 > n-2).$$

En déduire la loi de T_2.
3) Prouver une relation similaire pour le temps T_k *correspondant à un message nécessitant* k *secondes de transmission.*
4) Reprenez les questions précédentes en supposant que le message puisse être divisé en fragments d'une seconde pouvant être transmis de manière non-consécutive.

Exercice 107 *Deux amis, appelons-les Jojo et Gégé, décident de jouer au jeu suivant. Deux enveloppes indiscernables contiennent l'une un montant de* m *euros, et l'autre un montant de* $2m$ *euros (où* m *est un montant non-nul, inconnu des deux joueurs, mais fixé.) On répartit aléatoirement les deux enveloppes entre Jojo et Gégé. Jojo ouvre l'enveloppe qui lui a été attribuée, et y découvre une somme de* X *euros. On lui propose ensuite d'échanger le montant de son enveloppe avec celui de l'enveloppe de Gégé (qu'il n'a pas pu observer). Le raisonnement de Jojo est alors le suivant : «il y a une chance sur deux pour que mon enveloppe contienne le montant le plus élevé (*$2m$ *euros), et une chance sur deux pour qu'elle contienne le montant le plus bas (*m *euros). Par conséquent, il y a une chance sur deux pour que le montant de l'enveloppe de Gégé soit égal au double du montant contenu dans mon enveloppe, et une chance sur deux pour que le montant de l'enveloppe de Gégé soit égal à la moitié du montant contenu dans mon enveloppe. En moyenne, l'enveloppe de Gégé doit donc contenir* $1/2(1/2 \times X) + 1/2(2 \times X) = 5/4 \times X$ *euros. Or* $(5/4)X > X$, *et, par conséquent, j'ai intérêt à accepter l'échange qui m'est proposé.» Le problème est que Gégé, de son côté, peut se livrer exactement au même raisonnement et parvenir à la conclusion que lui aussi a intérêt à procéder à l'échange. Comment Jojo et Gégé peuvent-ils avoir intérêt simultanément à procéder à l'échange des montants contenus dans leurs enveloppes respectives ? En vous appuyant sur une modélisation probabiliste détaillée du problème, pouvez-vous confirmer ou infirmer le raisonnement de Jojo, et présenter une solution à ce paradoxe apparent ?*

Exercice 108 *Supposons que l'on tire un nombre aléatoire* U *de loi uniforme sur l'intervalle* $[0, 1]$, *puis que l'on effectue* n *lancers indépendants d'une pièce de monnaie ayant une probabilité de* U *de tomber sur pile, et* $1 - U$ *de tomber sur face. Quelle est la loi de probabilité du nombre de pile obtenus ?*

Exercice 109 *Supposons que l'on tire un nombre aléatoire* A *selon une loi exponentielle de paramètre* $\lambda > 0$, *puis, ce tirage effectué, un nombre aléatoire* X *selon une loi exponentielle de paramètre* A. *Quelle est la loi de probabilité de* X ?

Exercice 110 *On remplit une urne avec N boules selon la procédure suivante. Partant d'une urne vide, on effectue successivement N lancers indépendants d'une pièce de monnaie (pas nécessairement symétrique). A chaque lancer, on ajoute une boule dans l'urne, de couleur rouge si la pièce a donné pile, de couleur noire si la pièce a donné face. Une fois l'urne remplie, on tire uniformément au hasard, et sans remise, des boules dans l'urne, jusqu'à avoir vidé l'urne. Montrez que la couleur de la boule tirée à l'étape i (avec $1 \leq i \leq N$) est indépendante des couleurs des boules tirées précédemment. A présent, considérons le raisonnement suivant. «Une fois l'urne remplie, celle-ci contient N boules, dont un nombre aléatoire R de boules rouges. Au premier tirage, la probabilité d'obtenir une boule rouge est alors de R/N. Si c'est effectivement une boule rouge que j'obtiens, le deuxième tirage s'effectue avec une boule rouge de moins dans l'urne, et donc la proportion des boules rouges par rapport aux boules noires est moindre que lors du premier tirage. La probabilité d'obtenir une boule rouge au deuxième tirage doit donc être inférieure à ce qu'elle était lors du premier tirage.» Comment concilier ceci avec le fait que, d'après ce qui précède, la probabilité d'obtenir une boule rouge au i-ème tirage ne dépend pas des couleurs des boules tirées précédemment ? Reprendre la totalité de la question en supposant que l'on effectue des tirages répétés avec remise.*

Exercice 111 *(Méthode du second moment)*

Considérons une variable aléatoire positive X telle que $\mathbb{E}(X)$ et $\mathbb{E}(X^2)$ sont définies, et $\mathbb{P}(X = 0) < 1$. Prouvez l'inégalité suivante :

$$\mathbb{P}(X \geq \tfrac{1}{2}\mathbb{E}(X)) \geq \tfrac{1}{4}\frac{\mathbb{E}(X^2)}{[\mathbb{E}(X)]^2}$$

En quoi cette inégalité fourni-elle un complément à l'inégalité de Markov ?

Exercice 112 *(Le raisonnement de Huygens[15])*

1) Pour tout entier $q \geq 1$, montrez qu'il existe une fonction $f : \{1, \ldots, q\} \times \{1, \ldots, q\} \to \{1, \ldots, q\}$ vérifiant les deux conditions suivantes :

- *pour tout $i \in \{1, \ldots, q\}$, la fonction $f(i, \cdot) : \{1, \ldots, q\} \to \{1, \ldots, q\}$ est une bijection ;*
- *pour tout $j \in \{1, \ldots, q\}$, la fonction $f(\cdot, j) : \{1, \ldots, q\} \to \{1, \ldots, q\}$ est une bijection.*

2) Considérons X une variable aléatoire définie sur $(\Omega, \mathbb{P})$ satisfaisant la condition suivante : il existe une suite de nombres réels $x_1, \ldots, x_q$ telle que X suit la loi empirique associée à l'échantillon $(x_1, \ldots, x_q)$. Quelles sont les variables aléatoires X vérifiant une telle condition de manière exacte ? Et en s'autorisant une approximation arbitrairement petite ?

15. Christiaan Huygens (1629–1695).

3) Supposons à présent que q individus $I_1, \ldots, I_q$ participent à un pari basé sur la règle suivante. Chacun des q individus apporte une mise égale à $\frac{x_1+\cdots+x_q}{q}$. Ensuite, un entier L est choisi selon la loi uniforme sur l'ensemble $\{1, \ldots, q\}$, et la mise totale est répartie entre les joueurs de telle façon que, pour tout $1 \leq i \leq q$, l'individu I_i reçoit une somme égale à $x_{f(i,L)}$, où f est une fonction satisfaisant les propriétés mentionnées à la question 1). Montrez que la totalité de la mise $x_1 + \cdots + x_q$ est redistribuée entre les joueurs. Pour $1 \leq i \leq q$, quelle est la loi de probabilité du gain de l'individu i ?

4) Supposons qu'avant que le tirage ait pu avoir lieu, le pari soit interrompu. Quelle serait, selon vous, la répartition équitable de la mise totale $x_1 + \cdots + x_q$ entre les individus ?

5) Existe-t-il d'autres types de pari pour lesquels cette répartition devrait, selon vous, être différente ? Comment ce qui précède se compare-t-il à la règle de l'utilité espérée ?

Exercice 113 *(Aiguille de Buffon)*

On jette de manière répétée une aiguille au hasard sur un parquet constitué de lattes rectangulaires identiques, et parallèles. On s'intéresse à la moyenne du nombre d'intersections de l'aiguille avec les rainures délimitant les lattes. Pour simplifier, on choisit de négliger les effets de bord en considérant un parquet infini recouvrant la totalité du plan identifié à $\mathbb{R}^2$, et dont les rainures sont identifiées aux ensembles de la forme $\{i\} \times \mathbb{R}$, où i décrit $\mathbb{Z}$. De son côté, l'aiguille est assimilée à un segment de droite de longueur $L > 0$.

1) Décrivez précisément un modèle probabiliste du lancer, vérifiant, en termes informels, le fait que toutes les positions possibles relatives de l'aiguille par rapport au réseau formé par les lattes sont équiprobables. En appelant I notre aiguille, on note $N(I)$ la variable aléatoire indiquant le nombre de points d'intersection de l'aiguille avec les rainures.

Dans le cadre de votre modèle, comment l'espérance $\mathbb{E}(N(I))$ s'exprime-t-elle ?

2) Montrez que si, au lieu d'un segment, on jette sur le parquet (selon le même modèle) un objet formé de plusieurs aiguilles $I_1, I_2, \ldots, I_p$ mises bout-à-bout dans un même plan (autrement dit, une logne polygonale inscrite dans un plan), et rigidement attachées les unes aux autres, (deux aiguilles mises bout-à-bout peuvent former un angle quelconque, mais cet angle est fixé une fois pour toutes et ne varie pas au cours du mouvement de l'objet ainsi formé), le nombre total de points d'intersection de cette ligne polygonale avec les rainures vaut $\sum_{k=1}^{p} \mathbb{E}(N(I_k))$.

3) En déduire que $\mathbb{E}(N(I))$ est de la forme cL, où c est une constante.

4) En approchant un cercle de diamètre 1 par des lignes polygonales, montrez que la constante c est égale à $2/\pi$.

Exercice 114 *Prouvez que, si $(p_n)_{n\geq 1}$ est une suite de nombres compris entre 0 et 1 telle que $\lim_{n\to+\infty} np_n = \lambda > 0$, la loi binomiale de paramètres n et p_n converge vers une loi de Poisson de paramètre λ.*

Exercice 115 *On considère un entier positif n et un nombre réel p compris entre 0 et 1. On considère ensuite n variables aléatoires $U_1, \ldots, U_n$ indépendantes et de loi uniforme sur l'intervalle $[0,1]$. On définit alors, pour tout $1 \leq i \leq n$, les variables aléatoires X_i et Y_i par*

$$X_i = 0 \text{ si } U_i \leq 1-p \text{ et } X_i = 1 \text{ si } U_i > 1-p,$$

$$Y_i = k \text{ si } \sum_{j=0}^{k-1} e^{-p}\frac{p^j}{j!} < U_i \leq \sum_{j=0}^{k} e^{-p}\frac{p^j}{j!}$$

(avec la convention $\sum_{j=0}^{-1} \cdots = 0$). Enfin, on définit $S_n = \sum_{i=1}^n X_i$ et $T_n = \sum_{i=1}^n Y_i$.
1) Déterminer la loi de S_n et de T_n.
2) En utilisant l'inégalité $1-p \leq e^{-p}$, prouver que, pour tout i, $\mathbb{P}(X_i = Y_i) \geq 1-p^2$.
3) En déduire que $\mathbb{P}(S_n = T_n) \geq (1-p^2)^n$.
4) En écrivant $\mathbb{P}(S_n \in A) = \mathbb{E}(\mathbf{1}(S_n \in A))$ et $\mathbb{P}(T_n \in A) = \mathbb{E}(\mathbf{1}(T_n \in A))$, et en utilisant l'inégalité de l'exercice 116, prouvez que $|\mathbb{P}(S_n \in A) - \mathbb{P}(T_n \in A)| \leq (1-(1-p^2)^n)$.
5) En déduire l'inégalité suivante :

$$\sum_{k\in\mathbb{N}} \left| p_{Poiss(np)}(k) - p_{binom(n,p)}(k) \right| \leq 2\left(1-(1-p^2)^n\right),$$

et la comparer à celle donnée dans le cours.

Exercice 116 *Prouvez que, si X est une variable aléatoire possédant une espérance, et si $|X|$ possède une espérance, on a l'inégalité $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$.*

Exercice 117 *Si X et Y sont deux variables aléatoires indépendantes, X suivant une loi de Poisson de paramètre λ et Y suivant une loi de Poisson de paramètre μ, prouver que $X+Y$ suit une loi de Poisson de paramètre $\lambda+\mu$ par trois méthodes différentes :*

- *en calculant directement la loi de $X+Y$ à partir de la loi de X et de la loi de Y ;*
- *en utilisant les fonctions génératrices ;*
- *en utilisant la représentation d'une loi de Poisson comme limite d'une loi binomiale.*

Exercice 118 *Si X et Y sont deux variables aléatoires indépendantes, X suivant une loi binomiale de paramètres n et p et Y suivant une loi binomiale de paramètres*

m et q, prouver que, si $p = q$, $X + Y$ suit une loi binomiale de paramètres $n + m$ et p, par le calcul, et à partir du contexte dans lequel intervient la loi binomiale.

Qu'en est-il lorsque $p \neq q$?

Exercice 119 *Etant donnés $a > 0$ et $\epsilon > 0$, construire une variable aléatoire positive X possédant une espérance et vérifiant $\mathbb{P}(X \geq a) \geq (1 - \epsilon)\mathbb{E}(X)/a$.*

Exercice 120 *Considérons une variable aléatoire X de loi continue sur $\mathbb{R}$, donnée par une densité f. Montrez que, lorsque n tend vers l'infini, la loi de $nX \mod 1$ tend vers une loi uniforme sur l'intervalle $[0, 1]$.*

Exercice 121 *(Le paradoxe de Saint-Petersbourg)*

On se propose de jouer au jeu suivant. Une mise initiale de M euros ayant été versée, on lance une pièce de monnaie de manière répétée, et le jeu s'arrête lorsque la pièce retombe sur pile pour la première fois. Une somme de 2^T euros est alors versée au joueur, T désignant le nombre total de lancers effectués. Quelle est l'espérance de gain de ce jeu ? Quelle mise seriez-vous prêt à investir au maximum dans ce jeu ?

Exercice 122 *(Loi multinomiale)*

Considérons un modèle probabiliste $(\Omega, \mathbb{P})$, et m événements $A_1, A_2, \ldots, A_m \subset \Omega$ formant un système complet d'événements. Posons $p_i = \mathbb{P}(A_i)$ pour $1 \leq i \leq m$. Considérons maintenant le modèle $(\Omega^N, \mathbb{P}^{\otimes N})$ correspondant à N répétitions indépendantes de $(\Omega, \mathbb{P})$, et, pour tout $1 \leq i \leq m$, définissons sur ce modèle la variable aléatoires N_i comme le nombre de fois où l'événement A_i se réalise. La loi de N_i est donc une loi binomiale de paramètres N et p_i. Par ailleurs, on a $N_1 + \cdots + N_m = N$.

Montrez que la loi jointe de $(N_1, \ldots, N_m)$ est donnée par la formule suivante. Pour tout $p-$uplet d'entiers $(d_1, \ldots, d_m)$ compris entre 0 et N et vérifiant $d_1 + \cdots + d_m = N$,

$$\mathbb{P}^{\otimes N}(N_1 = d_1, \ldots, N_m = d_m) = \frac{N!}{d_1! \times \cdots \times d_m!} p_1^{d_1} \times \cdots \times p_m^{d_m}.$$

Cette loi est appelée loi multinomiale de paramètres N et $(p_1, \ldots, p_m)$. Pour $m = 2$, on retrouve la loi binomiale habituelle.

Si $i_1, \ldots, i_s$ est un sous-ensemble d'indices de $\{1, \ldots, m\}$, que pouvez-vous dire de la loi de $N_{i_1} + \cdots + N_{i_s}$? Et de la loi de $(N_{i_1}, \ldots, N_{i_s})$ conditionnellement à l'événement $N_{i_1} + \cdots + N_{i_s} = k$, où $0 \leq k \leq N$ est un nombre fixé ? Et de la loi jointe des deux variables aléatoires $(N_{i_1}, \ldots, N_{i_s})$ et $(N_j)_{j \notin \{i_1, \ldots, i_s\}}$ conditionnellement à ce même événement ?

Exercice 123 *On considère un modèle probabiliste $(\Omega, \mathbb{P})$ sur lequel est définie une variable aléatoire X de loi binomiale de paramètres n et p. Est-il toujours vrai que l'on peut définir sur Ω une famille de n variables aléatoires mutuellement indépendantes, toutes de loi de Bernoulli de paramètre p ?*

Exercice 124 *On considère une urne contenant m boules dont a sont rouges et $m-a$ sont blanches. On effectue un nombre $n \leq m$ de tirages sans remise dans l'urne, en supposant que chaque tirage est effectué uniformément au hasard dans l'ensemble des boules restantes au moment où celui-ci a lieu. Appelons N_a le nombre total de boules rouges figurant parmi les n boules tirées. La loi de N_a est appelée loi hypergéométrique de paramètres n, a et m.*
1) Prouvez que l'on a, pour tout $0 \leq k \leq \min(a,m)$, $\mathbb{P}(N_a = k) = C_a^k C_{m-a}^{n-k}/C_m^n$. (Proposez au moins trois arguments de dénombrement différents !)
2) Pouvez-vous calculer, à partir de la formule précédente, $\mathbb{E}(N_a)$ et $\mathbb{V}(N_a)$?
3) On définit, pour $1 \leq i \leq m$, la variable X_i comme l'indicatrice de l'événement : tirer une boule rouge lors du $i-$ème tirage. Quelle relation y a-t-il entre N_a et les variables X_i ? Pouvez-vous en déduire l'espérance et la variance de N_a ?
4) Comment la loi hypergéométrique se différencie-t-elle de la loi binomiale de paramètres n et a/m ? Prouvez que, si n est fixé, et si m et a tendent vers l'infini de telle sorte que a/m tend vers une valeur limite p, on obtient à la limite la loi binomiale de paramètres n et p.

Exercice 125 *On considère des répétitions indépendantes de tirages de Bernoulli (succès ou échec) avec probabilité de succès p.*

Pour tout $k \geq 1$, on appelle N_k le nombre d'essais qu'il est nécessaire d'effectuer jusqu'à parvenir à un total de k succès. La loi de N_k est appelée loi binomiale négative de paramètres k et p.
1) Quelle est la loi de N_1 ?
2) Montrez que l'on a, pour tout $n \geq 1$, $\mathbb{P}(N_k = n) = C_{n-1}^{k-1} p^k (1-p)^{n-k}$.
3) Pouvez-vous calculer, à partir de la formule précédente, $\mathbb{E}(N_k)$ et $\mathbb{V}(N_k)$?
4) Que peut-on dire des variables aléatoires $N_{k+1} - N_k$ pour $k \geq 1$? En déduire l'espérance et la variance de N_k.

Exercice 126 *(Analyse en moyenne de l'algorithme de tri rapide)*

Pas encore écrit en détail...

Exercice 127 *Appelons $\mathfrak{S}_n$ l'ensemble des permutations de l'ensemble des entiers de 1 à n. On cherche à générer un élément σ de $\mathfrak{S}_n$ de loi uniforme, à partir de variables aléatoires indépendantes de loi uniforme sur $[0,1]$.*
1) Proposez (et prouvez la validité d') une méthode très simple, basée sur la génération progressive de $\sigma(1)$, suivie de $\sigma(2)$, et ainsi de suite jusqu'à $\sigma(n)$. Evaluez le coût de cette méthode en terme de nombre d'opérations effectuées.
2) Voici une alternative. On part de n variables aléatoires $U_1, \ldots, U_n$ indépendantes et de loi uniforme sur $[0,1]$, que l'on trie par ordre croissant. En notant $i_1, \ldots, i_n$ l'unique famille d'indices vérifiant $U_{i_1} < U_{i_2} < \ldots < U_{i_n}$, la permutation σ renvoyée par l'algorithme est définie par $\sigma(k) = i_k$ pour tout $1 \leq k \leq n$. Prouvez la validité

de cette méthode. Evaluez le coût de cette méthode en termes de nombre d'opérations effectuées. Comment gérer les égalités entre différentes variables en tenant compte du fait que l'on ne manipule les réels qu'avec un nombre fini de décimales ?
3) Considérons la méthode suivante : Partant de $\sigma = Id$, on effectue la boucle suivante : pour j décroissant de n à 1, tirer un entier J entre 1 et i selon la loi uniforme (indépendamment des tirages précédemment effectués), et échanger les valeurs de $\sigma(i)$ et de $\sigma(J)$. Quel est le coût de cette méthode en termes de nombre d'opérations effectuées ? Prouvez que la permutation qui en résulte suit effectivement la loi uniforme sur $\mathfrak{S}_n$, en interprétant cette méthode comme une implémentation (efficace !) de la méthode proposée en 1).

Exercice 128 *Soient $X_1, \ldots, X_n$ n variables aléatoires globalement indépendantes de loi exponentielle de paramètre λ, et soit $S_n = X_1 + \cdots + X_n$.*

Montrez que la loi de S_n est une loi gamma de paramètres $a = n$ et $s = \lambda$.

Indication pour prouver le résultat (presque) sans calculs : montrer que l'événement $\mathbb{P}(X_1 + \cdots + X_n \leq t) = \mathbb{P}(N_t \geq n)$, où N_t est une variable aléatoire de loi de Poisson de paramètre λt. Autre approche : passer par l'approximation discrète des variables de loi exponentielle par des variables de loi géométrique.

Exercice 129 *Soient $X_1, \ldots, X_n$ n variables aléatoires globalement indépendantes de loi gaussienne de paramètre $m = 0$ et $v = 1$, et soit $S_n = X_1^2 + \cdots + X_n^2$.*

Montrez que la loi de S_n est une loi du chi-deux à n degrés de liberté.

Exercice 130 *Soit X et Y deux variables aléatoires indépendantes, X suivant une loi gaussienne de paramètres m et v, Y une loi gaussienne de paramètres m' et v'. Montrez que la loi de $X + Y$ est une loi gaussienne. Quels sont ses paramètres ?*

Exercice 131 *Soient $X_1, \ldots, X_n$ n variables aléatoires globalement indépendantes de loi de Cauchy de paramètres ℓ et s, et soit $S_n = X_1 + \cdots + X_n$.*

Montrez que la loi de S_n est une loi de Cauchy de paramètre $n\ell s$.

Exercice 132 *Soient X et Y deux variables aléatoires indépendantes de loi gassienne de paramètres $m = 0$ et $v = 1$.*

Montrez que la loi de X/Y est une loi de Cauchy.

Exercice 133 *(Vérification rapide randomisée d'un produit matriciel)*

Considérons une matrice $n \times n$ A à coefficients réels, et dont au moins un coefficient n'est pas nul. Considérons un vecteur aléatoire $v = (e_1, \ldots, e_n)$ dont les coordonnées sont des variables aléatoires indépendantes et de loi de Bernoulli de paramètre $1/2$. Prouvez que $\mathbb{P}(Av \neq 0) \geq 1/2$.

Application : considérons trois matrices $n \times n$, X, Y et Z, à coefficients réels, et supposons que $XY \neq Z$. D'après ce qui précède, $\mathbb{P}(XYv \neq Zv) \geq 1/2$.

Quel est le coût du calcul de XYv et de Zv ? Déduisez de ce résultat une méthode permettant de détecter la différence entre XY et Z avec une probabilité supérieure à $99,9\%$. Comment le coût de cette méthode se compare-t-il au calcul direct du produit XY ?

(Rappel : l'algorithme le plus simple de multiplication des matrices a un coût en $O(n^3)$, les meilleurs algorithmes de multiplication rapide connus ont un coût inférieur à $O(n^{2,4})$.)

Pour en savoir plus, sur ce type de technique, qui s'étendent à des questions bien plus générales de vérification rapide d'identités entre objets, vous pouvez consulter l'ouvrage de Motwani et Raghavan cité dans la bibliographie.

Exercice 134 *(Urne de Pólya[16])*

On dispose d'une urne contenant initialement $a \geq 1$ boules rouges et $b \geq 1$ boules noires. On répète ensuite le petit jeu suivant : on tire une boule uniformément dans l'urne, et, une fois la couleur de cette boule observée, on la remet dans l'urne accompagnée de Δ boules de la même couleur. Appelons $X_1, \ldots, X_n$ la suite des couleurs obtenues au cours n premiers tirages effectués dans l'urne, avec comme codage $X_i = 1$ si la boule obtenue au $i-$ème tirage est rouge et $X_i = 0$ si celle-ci est noire.

Par ailleurs, effectuons l'expérience suivante. On tire un nombre q entre 0 et 1 selon la loi beta de paramètres a/Δ et b/Δ, puis l'on tire n variables aléatoires indépendantes de Bernoulli de paramètre q. Montrez que les lois jointes de $X_1, \ldots, X_n$ d'une part, et de $Y_1, \ldots, Y_n$ d'autre part, sont identiques.

Encore une représentation du même modèle : on part d'un jeu de cartes comportant $a+b-1$ cartes, les a cartes du dessus étant considérées comme «rouges», et les $b-1$ cartes du dessous étant considérées comme «noires». On répète ensuite l'insertion de nouvelles cartes dans le jeu, chaque carte étant insérée en une position choisie uniformément au hasard parmi les emplacements possibles dans le paquet (dans un paquet de k cartes, il y a donc $k+1$ emplacements possibles). Si une carte est insérée au-dessus de la carte rouge la plus basse, elle est elle-même considérée comme «rouge», et elle est considérée comme noire sinon. Montrez que la loi de la suite des couleurs des cartes s'identifie aux modèles décrits ci-dessus. Montrez comment cette représentation permet de calculer facilement la loi du nombre de cartes rouges après n insertions.

Exercice 135 *(Jeux à somme nulle)*

Deux amis, Anselme et Barnabé, jouent au jeu suivant. Anselme doit choisir une option parmi n possibles, numérotées $1, 2, \ldots, n$, tandis que Barnabé doit choisir une option parmi m possibles, numérotées $1, 2, \ldots, m$. Si Anselme a choisi l'option i et

16. George Pólya (1887–1985).

Barnabé l'option j, Barnabé doit à Anselme une somme de a_{ij} euros, cette somme pouvant être soit positive (Anselme a vraiment gagné, et Barnabé lui doit de l'argent), soit négative (auquel cas, c'est en fait Barnabé qui a gagné, et Anselme qui lui doit de l'argent, puisque la somme «due» à Anselme par Barnabé est négative.)
1) Supposons par exemple que $n = 2$, $m = 3$, et que la matrice (a_{ij}) soit la suivante

a_{ij}	$j=1$	$j=2$	$j=3$
$i=1$	15	-10	20
$i=2$	10	-20	-20

Comment Anselme et Barnabé devraient-ils jouer, selon vous, dans le but de maximiser leur bénéfice ?
2) Même question avec la matrice suivante.

a_{ij}	$j=1$	$j=2$	$j=3$
$i=1$	12	10	15
$i=2$	17	5	-20

3) Même question avec la matrice suivante.

a_{ij}	$j=1$	$j=2$	$j=3$
$i=1$	30	-10	21
$i=2$	10	20	-20

4) Pour tout entier k, notons $P_k = \{(r_1, \ldots, r_k) \in \mathbb{R}^k \ : \ r_i \geq 0$ pour tout i et $\sum_{i=1}^k r_i = 1 \ \}$.

Pour $p = (p_1, \ldots, p_n) \in P_n$ et $q = (q_1, \ldots, q_m) \in P_m$ que représente vis-à-vis du jeu l'expression $S(p,q) = \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} a_{ij} p_i q_j$?

Le théorème du minimax de Von Neumann[17] *affirme que* $\max_{p \in P_n} \min_{q \in P_m} S(p,q) = \min_{q \in P_m} \max_{p \in P_n} S(p,q)$.

Quelle conséquence ce résultat peut-il avoir, selon vous, sur la manière dont devraient jouer Anselme et Barnabé ?
5) (Football) Supposons qu'Anselme soit gardien de but, tandis que Barnabé tente de marquer des penalties.

Pour simplifier, mettons que les options de Barnabé soient de tirer vers la droite ou vers la gauche, tandis que celles d'Anselme sont de plonger vers la droite ou vers la gauche (il doit de toute façon décider avant que Barnabé n'ait tiré, compte-tenu de la vitesse du ballon). En admettant qu'Anselme soit aussi habile pour arrêter les ballons arrivant à sa gauche que ceux arrivant à sa droite, comment Anselme et Barnabé devraient-ils jouer selon vous ? Même question en supposant qu'Anselme ait 70% de

17. John Von Neumann (1903 – 1957).

chances d'intercepter un ballon arrivant à sa droite, et seulement 50% de chances d'intercepter un ballon arrivant à sa gauche.
6) Supposons qu'à présent le résultat du jeu se traduise par une dette (éventuellement négative) d'Anselme et Barnabé non pas directement l'un vis-à-vis de l'autre, mais vis-à-vis d'une banque. Plus précisément, si Anselme a choisi l'option i et Barnabé l'option j, la banque doit à Anselme une somme de a_{ij} euros, cette somme pouvant être soit positive soit négative, et à Barnabé une somme de b_{ij} euros. Si l'on suppose que $a_{ij} = -b_{ij}$, la situation se ramène à la précédente, et la banque n'est qu'un intermédiaire sans effet sur le jeu. Si, en revanche, on ne suppose plus une telle relation, que devient la validité des raisonnements précédents ?

Exercice 136 *Si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes de même loi, dont la fonction de répartition est notée F, calculez les fonctions de répartition des variables aléatoires suivantes :* $\max(X_1, \ldots, X_n)$ *et* $\min(X_1, \ldots, X_n)$.

Exercice 137 *On choisit un angle selon la loi uniforme dans l'intervalle* $[0, 2\pi]$. *Quelle est la loi de la tangente de cet angle ?*

Exercice 138 *Pourquoi peut-on affirmer, sans même effectuer de calcul de probabilités, que la plupart des loteries (la loterie nationale, l'euro-million) présentent une espérance de gain négative ? Le fait que de très nombreux individus participent à ces jeux est-il compatible avec la règle de l'utilité espérée ? Estimez-vous, selon les termes de Flaubert, que le loto est « un impôt volontaire sur la bêtise » ?*

Exercice 139 *Un groupe de 20 personnes a été fait prisonnier par une troupe de bandits aux regards cruels et aux cœurs insensibles. Après plusieurs jours de captivité, le chef des bandits expose aux prisonniers le (triste) sort qui les attend. Ceux-ci seront numérotés de 1 à 20, puis, l'un après l'autre, amenés dans une salle où se trouvent 20 coffrets, disposés de gauche à droite sur le sol. Les coffrets sont également numérotés de 1 à 20, mais le numéro attribué à chaque coffret est inscrit à l'intérieur de celui-ci, et il faut donc ouvrir un coffret pour connaître son numéro. Bien entendu, la disposition extérieure des coffrets ne renseigne en rien sur les numéros qui leur sont attachés.*

Une fois admis dans la salle, un prisonnier devra tenter d'ouvrir le coffret portant son propre numéro, mais n'aura le droit, pour essayer d'atteindre cet objectif, que d'ouvrir 10 coffrets au plus.

Ensuite, ledit prisonnier sera évacué, sans avoir la possibilité de communiquer avec les prisonniers suivants, et donc sans pouvoir leur fournir aucune indication sur les numéros des différents coffrets qu'il a pu observer.

Si, à l'issue de l'expérience, chaque prisonnier est parvenu à ouvrir le coffret portant son propre numéro, les prisonniers seront libérés. Si un seul d'entre eux

échoue, ils seront impitoyablement exécutés. Telle est la décision du chef des bandits, qui, souligne-t-il, a tenu à ménager aux prisonniers une infime chance de s'en tirer.
1) En admettant que chaque prisonnier choisisse au hasard les coffres qu'il peut ouvrir, quelle devrait être la probabilité de succès d'un prisonnier ? Qu'en est-il alors, de la probabilité de survie du groupe ?

Après avoir mené ce petit calcul, les prisonniers sont bien désemparés, mais... l'un d'entre eux les invite à ne pas totalement perdre espoir, et leur affirme qu'il détient une méthode leur permettant d'augmenter considérablement leurs chances de succès.

Sa méthode est la suivante : le prisonnier titulaire du numéro i devra ouvrir en premier le $i-$ème coffret en partant de la droite. En appelant j le numéro inscrit à l'intérieur de ce coffret, il devra ensuite ouvrir le $j-$ème coffret en partant de la droite. En appelant k le numéro inscrit à l'intérieur de ce nouveau coffret, il devra ensuite ouvrir le $k-$ème coffret, et ainsi de suite jusqu'à avoir découvert le coffret portant le numéro i, ou, malheureusement, épuisé les dix coffrets qu'il était en droit d'ouvrir.
2) En appelant $\sigma(i)$ le numéro contenu dans le coffret placé en $i-$ème position en partant de la droite, et en admettant que σ est une permutation aléatoire de loi uniforme sur l'ensemble des permutations des entiers de 1 à 20, calculez la probabilité de succès de l'ensemble des prisonniers. (Indication : caractérisez l'événement correspondant au succès des prisonniers en termes d'existence de cycles de longueur supérieure à 10 pour la permutation σ. Ensuite, pour $k \geq 11$, comptez le nombre de permutations des entiers de 1 à 20 possédant un cycle de longueur k.)
3) Au courant du stratagème imaginé par les prisonniers, et afin de les désespérer plus encore, le chef laisse filtrer l'information selon laquelle il permutera les coffrets de telle façon qu'il existe au moins un cycle de longueur supérieure à 10. Comment, en se mettant d'accord à l'avance sur une permutation aléatoire τ des entiers de 1 à 20, dont ils garderont le secret, les prisonniers peuvent-ils contourner cet obstacle ?
3) Le nombre i étant fixé, quelle est la probabilité pour que le prisonnier numéro i réussisse à ouvrir le coffret portant son propre numéro ?
4) Appelons X le nombre total de prisonniers parvenant à ouvrir le coffret portant leur numéro. Quelle sont l'espérance et la variance de X ? Si les succès des différents prisonniers étaient mutuellement indépendants, quelle serait la loi de X ? En étudiant ce qui advient lorsqu'il existe un cycle de longueur supérieure à 10 dans la permutation appliquée par les prisonniers, et en reprenant les calculs de la question 2), calculez la loi de X.

Exercice 140 *Considérons un jeu de loto dans lequel N personnes achètent des bulletins coûtant chacun 1 euro. Chaque personne indique sur son bulletin une combinaison de chiffres, m combinaisons différentes étant disponibles, puis fait valider*

son bulletin. Un tirage est ensuite effectué, au cours duquel l'une des combinaisons est choisie aléatoirement, selon la loi uniforme. On répartit ensuite un pourcentage fixé (disons α) des N euros collectés entre les personnes dont les bulletins portent la combinaison qui a été tirée.

Supposons qu'il existe un numéro particulier que personne ne pense jamais à jouer. Quelle serait l'espérance de gain d'une personne qui choisirait justement de miser sur ce numéro ?

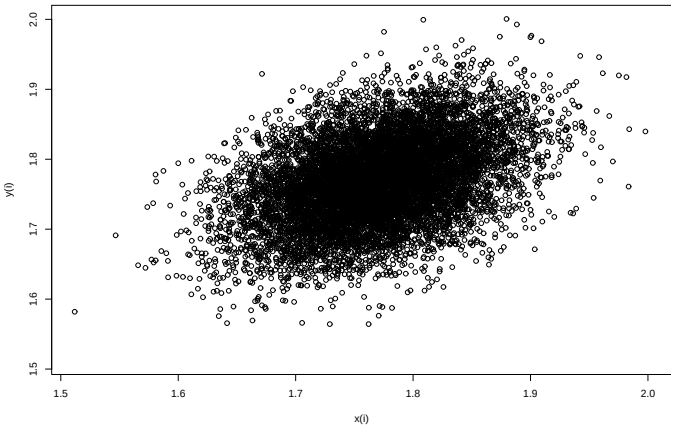
Ceci vous suggère-t-il une stratégie vous permettant de gagner de l'argent en jouant au loto ?

Exercice 141 *(Le problème de Galton[18])*

On a mesuré dans 10000 familles différentes les deux quantités suivantes : taille du père, et taille du fils aîné (à l'âge adulte). (En fait, nous utilisons dans cet exercice des données simulées, et non pas de véritables données mesurées, mais le modèle employé pour la simulation est inspiré par les données réelles étudiées par Galton.) Pour $1 \leq i \leq 10000$, nous noterons (x_i, y_i) le couple formé des deux valeurs (taille du père, taille du fils) dans la $i-$ème famille étudiée.

L'une des principales questions auxquelles s'intéressait Galton était la suivante : quelle est l'influence de la taille du père sur la taille du fils ?

Voici le nuage de points formé par les paires (x_i, y_i) pour $1 \leq i \leq 10000$.

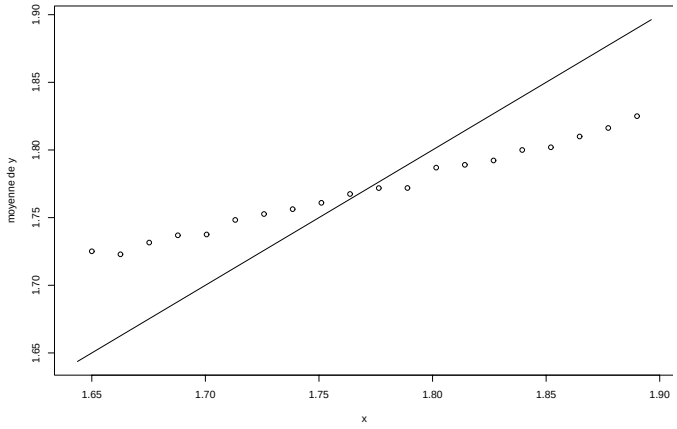


1) Quelle observation très grossière sur l'association entre taille du père et taille du fils peut-on faire, simplement à partir de l'observation de ce nuage de points ?
2) Les tailles moyennes calculées à partir des données présentées sont très voisines chez les pères et chez les fils $1,770m$ pour les pères, et $1,771m$ pour les fils (en arrondissant au millimètre). Le graphique suivant représente, pour différentes tranches

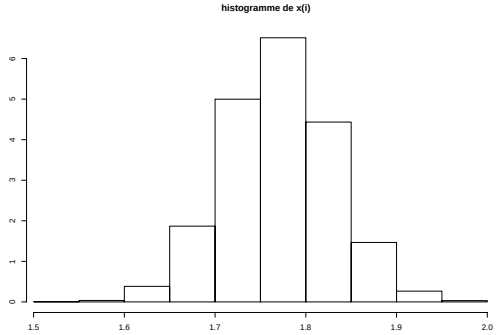
18. Sir Francis Galton (1822–1911).

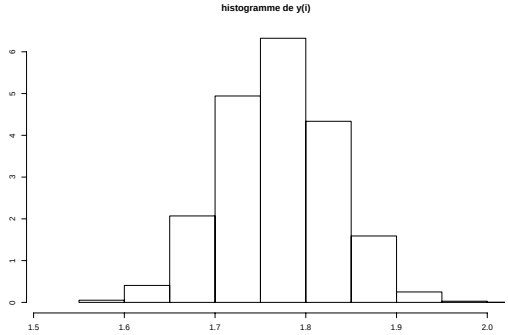
de valeurs de la taille du père, la valeur moyenne de la taille du fils dans les familles correspondantes, et en surimpression la droite d'équation $y = x$.

On a regroupé entre elles les observations de la taille du père par tranches de longueur d'environ 1,2 cm afin de disposer de données assez nombreuses dans chaque tranche, les ordonnées des points représentant les milieux de chaque tranche.

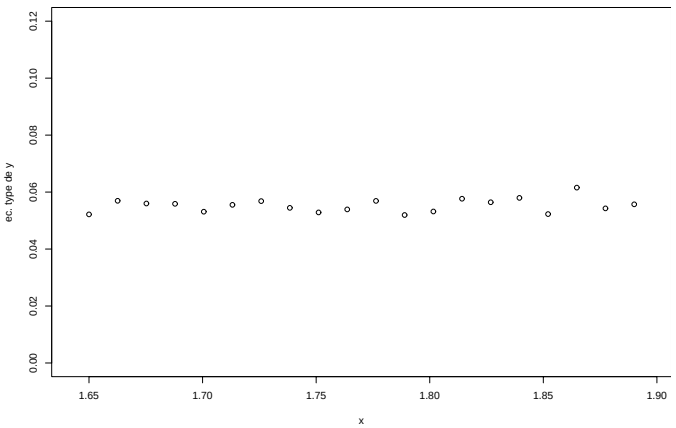


L'observation qui avait frappé Galton était la suivante : la courbe obtenue est approximativement une droite, mais dont la pente est nettement inférieure à 1, coupant la droite d'équation $y = x$ au niveau de la taille moyenne de la population, ce qui signifie que les enfants nés d'un père plus grand que la moyenne, sont, également, en moyenne, plus grands que la moyenne de la population, mais que leur taille moyenne est plus proche de la moyenne que ne l'est celle de leur père. La même observation peut être faite, en sens inverse, pour les enfants issus d'un père de taille inférieure à la moyenne. On observe donc un phénomène de retour vers la moyenne, chaque individu donnant naissance à des enfants en moyenne plus proches qu'eux mêmes de la taille moyenne de la population. On note donc que la taille d'un fils n'est pas en moyenne égale à celle de son père, mais présente un décalage dans la direction de la moyenne de la population. La conclusion en apparence logique de cette observation serait que, au fur et à mesure des générations, la taille des individus a tendance à converger vers la valeur moyenne (1,77 m dans notre exemple). Pourtant, si l'on examine les deux distributions de taille, chez les pères et chez les fils, on n'observe aucun phénomène de «resserrement» des tailles autour de la moyenne dans la population des fils par rapport à celle des pères, et les deux distributions des tailles semblent très voisines. Les écarts-types, quant à eux, sont tous les deux égaux à 0,060 (en arrondissant au millimètre).





Comment pourrait-on alors expliquer une telle situation ? Comment votre explication s'accomode-t-elle du graphique suivant, qui représente non plus la moyenne, mais l'écart-type, calculé dans chacune des tranches de taille des pères présentées ci-dessus, et qui suggère également que la variabilité de la taille des fils – telle que mesurée par l'écart-type – ne varie pas ou peu avec la taille des pères ?



Pour préciser la question et l'explication, notons que, en termes mathématiques, on cherche à comprendre comment on peut effectivement disposer de deux variables aléatoires X et Y telles que :

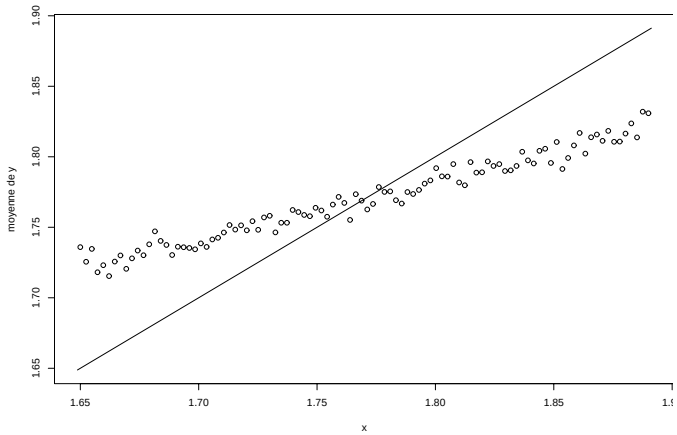
- $\mathbb{E}(X) = \mathbb{E}(Y) = m$;
- $\mathbb{V}(X) = \mathbb{V}(Y) = v$;
- *pour tout x, $\mathbb{E}(Y|X = x)$ est plus proche de m que x ;*
- *pour tous x_1, x_2, $\mathbb{V}(Y|X = x_1) = \mathbb{V}(Y|X = x_2)$.*

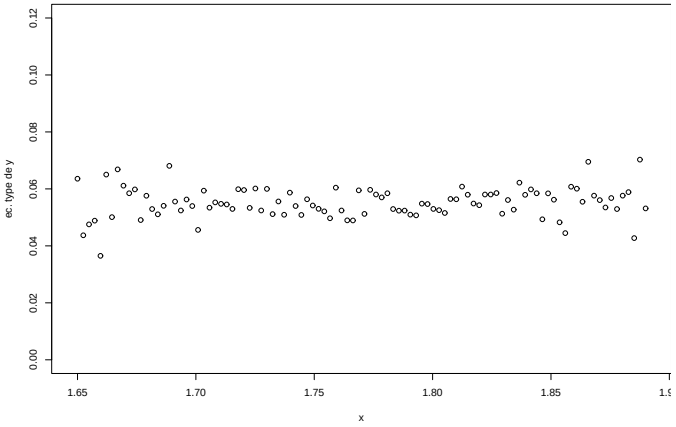
Suggestion : étudier les exemples de la forme $Y = m + \alpha(X - m) + W$, où m et α sont des constantes, X est une variable aléatoire quelconque (possédant une espérance égale à m et une variance égale à v), et W une variable aléatoire indépendante de X. (En fait, les simulations présentées appartiennent effectivement à cette catégorie d'exemples. Pour comprendre comment des paires (taille du père, taille du fils) véritablement mesurées dans une population humaine peuvent effectivement entrer dans ce cadre, nous vous renvoyons à la suite de cet exercice présentée dans le chapitre «Courbe en cloche».)

3) Nous n'avons pas du tout abordé les questions liées à l'estimation de quantités à partir de lois empiriques. Bien entendu, un traitement statistique correct (voir la partie «Statistique») peut et doit prendre en compte ces questions – très importantes en théorie comme en pratique –, de manière quantitative.

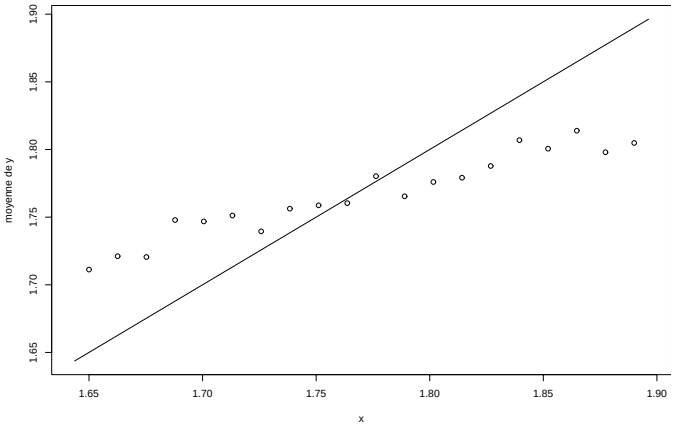
Les graphiques suivant fournissent juste une petite illustration qualitative des phénomènes liés à la taille de la population étudiée, et au choix de la taille des tranches qui permettent de découper le domaine des valeurs de la taille du père.

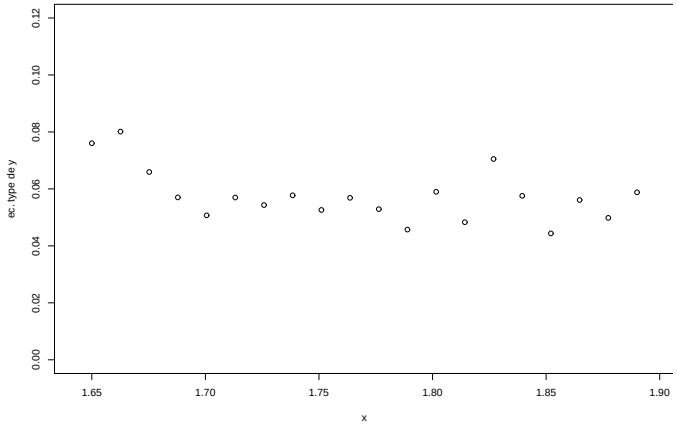
En conservant la même population (10000 mesures), et en considérant des tranches de taille 2,5mm environ.



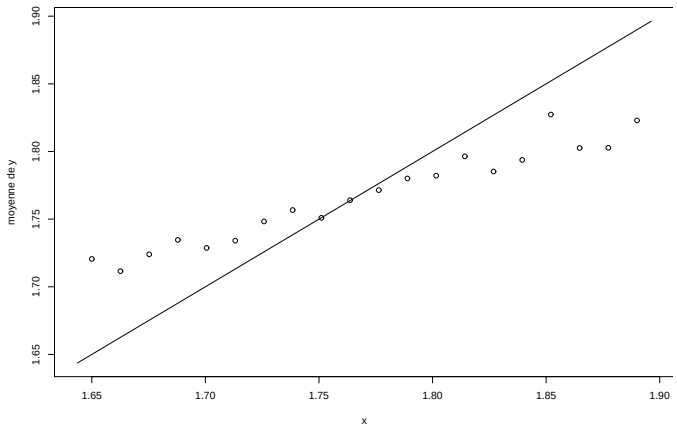


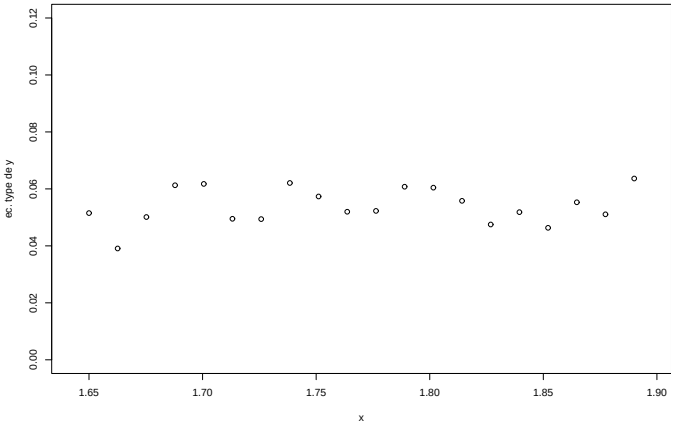
En considérant une population plus petite (1000 mesures),





En considérant une nouvelle population de 1000 mesures indépendante de la précédente.





Exercice 142 *(Mariages stables) On appelle problème des mariages stables la question suivante. On dispose de deux populations A et B comportant chacune n individus (disons, les hommes et les femmes). Chaque individu possède une liste de préférence personnelle, dans laquelle les n individus de la population du sexe opposé sont classés par ordre de préférence. Un mariage entre ces deux populations est simplement la donnée de n couples $(a_1, b_1), \ldots, (a_n, b_n)$ tels que chaque entier entre 1 et n figure une et une seule fois dans chacune des deux listes $(a_1, \ldots, a_n)$ et $(b_1, \ldots, b_n)$. Si l'on voit les éléments a_i comme numérotant des hommes, et les b_i comme numérotant des femmes, un mariage est donc simplement un appariement entre tous les hommes et toutes les femmes de la population. On dit qu'un tel mariage est stable lorsqu'il ne comporte aucune paire de couples (a_i, b_i) et (a_j, b_j) tels que a_i classe b_j avant b_i dans sa liste de préférence, tandis que b_j classe a_i avant a_j dans sa liste de préférences (a_i et b_j auraient alors tendance à rompre leurs couples pour se regrouper tous les deux).*

Un résultat non-trivial est que, quelles que soient les listes de préférences, il existe toujours au moins un mariage stable. La question est ensuite : comment trouver algorithmiquement un tel mariage stable. Compte-tenu du nombre de mariages possibles ($n!$), il n'est pas question d'énumérer tous les mariages possibles. La méthode naïve consistant à partir d'un mariage arbitraire pour essayer de le corriger progressivement en éliminant les mariages instables ne fonctionne pas, mais l'algorithme suivant («les hommes proposent, les femmes disposent») répond à cette question. Cet algorithme fonctionne de la manière suivante. A tout moment de son déroulement, un mariage partiel (certains couples mariés sont formés, tandis que d'autres individus peuvent être célibataires) entre les deux populations est défini, et chaque homme a déjà enregistré un certain nombre de refus de mariage de la part de certaines femmes. Initialement, aucun individu n'est marié. Ensuite, l'un des hommes non mariés (par exemple celui possédant le plus petit indice) propose de se marier à la femme qui se

trouve le plus haut placée dans sa liste de préférence, et qui ne l'a pas déjà refusé. Si cette femme n'est pas mariée, elle accepte le mariage avec cet homme. Si elle est déjà mariée, mais que son mari actuel se trouve moins bien placé dans sa liste de préférences que le nouveau prétendant, le mariage précédent est défait, et la femme est remariée avec le prétendant. Dans le cas contraire, la femme repousse la proposition qui lui est faite.

Lorsque tous les hommes (et donc toutes les femmes) sont mariés, l'algorithme s'arrête.
1) Prouver que l'algorithme s'arrête après n^2 étapes au pire, et que le mariage constitué lorsqu'il s'arrête est un mariage stable.
2) On s'intéresse à la distribution de probabilité du temps d'exécution (compté en nombre d'étapes) de l'algorithme lorsque les listes de préférences des hommes sont obtenues en effectuant une permutation aléatoire de loi uniforme sur l'ensemble des permutations des entiers de 1 à n, et ce, indépendamment d'un homme à l'autre, les listes de préférence des femmes pouvant, quant à elles, être totalement arbitraires (on ne fait aucune hypothèse de modélisation à leur sujet). Appelons T ce temps d'exécution, et introduisons le temps T' obtenu en modifiant l'algorithme de la manière suivante : au lieu de suivre sa liste de préférences, chaque homme tire à chaque fois uniformément au hasard la femme à laquelle il va proposer de se marier (il se peut donc qu'il repropose le mariage à une femme qui l'a déjà rejeté, et ne pourra donc que refuser à nouveau). Montrer que pour tout $k \geq 0$, $\mathbb{P}(T \geq k) \leq \mathbb{P}(T' \geq k)$. Montrez ensuite que le temps T' peut-être analysé comme dans le problème du collectionneur de vignettes à n vignettes (exercice 69). Que peut-on en déduire sur la distribution de probabilité du temps T ?

Exercice 143 *(Coupure minimale dans un graphe) Un multigraphe est la donné d'un ensemble fini de sommets V et d'un ensemble fini d'arêtes V reliant des sommets entre eux (les arêtes que nous considérons sont non-orientées, deux sommets peuvent être reliés par plus d'une arête, il n'y a pas de boucles). On dit qu'un tel graphe est connexe lorsque l'on peut toujours passer de tout sommet à tout autre en suivant un chemin constitué d'arêtes. Une coupure d'un graphe connexe est un ensemble d'arêtes tel que, si l'on supprime les arêtes figurant dans cet ensemble, le graphe perd la propriété de connexité. Le problème de la coupure minimale consiste à rechercher une coupure comportant le plus petit nombre d'arêtes possibles.*

Voici un algorithme randomisé destiné à résoudre ce problème : on choisit une arête uniformément au hasard dans l'ensemble des arêtes, et l'on contracte celle-ci, c'est-à-dire que l'on identifie les deux sommets que cette arête relie, tout en supprimant toutes les arêtes pouvant exister entre ces deux sommets. On obtient alors un nouveau graphe, auquel on réapplique l'opération précédente, et l'on continue jusqu'à obtenir un graphe ne comportant plus que deux sommets. L'ensemble des arêtes

reliant ces sommets est ensuite renvoyé par l'algorithme.
1) Prouvez que toute coupure de l'un des graphes intermédiaires manipulés par l'algorithme est une coupure du graphe original, et que l'algorithme renvoie donc toujours une coupure du graphe (pas forcément minimale).
2) Considérons une coupure minimale du graphe. Prouver que, si aucune des arêtes de cette coupure n'est contractée par l'algorithme, alors l'ensemble d'arêtes renvoyé par l'algorithme est exactement constitué par les arêtes de cette coupure.
3) Considérons une coupure minimale du graphe, et k le nombre d'arêtes qu'elle contient. Soit n le nombre de sommets du graphe. Prouver que le graphe comporte au moins $kn/2$ arêtes.
4) A l'aide des deux questions précédentes, prouvez que la probabilité pour que l'algorithme renvoie une coupure minimale est supérieure ou égale à $\prod_{i=1}^{n-2}\left(1-\frac{2}{n-i+1}\right)$, et donc à $2/n^2$.
5) Comment faire pour obtenir une probabilité élevée de succès (disons $99,9\%$) ? Combien de pas sont alors nécessités par l'algorithme ? (Quid de la manipulation des structures de données qui interviennent ?)

Exercice 144 *Ce soir, Jojo reçoit ses beaux-parents chez lui pour la première fois. Soucieux que tout se passe pour le mieux, il va jusqu'à s'interroger sur le bon fonctionnement des ampoules électriques installées à son domicile. En particulier, l'ampoule éclairant la salle à manger n'a pas été changée depuis plus de deux ans, et Jojo redoute que celle-ci ne claque pendant le repas. Il préfère donc changer ladite ampoule en la remplaçant par une ampoule neuve, du même modèle que la précédente, en espérant diminuer la probabilité d'un claquage au cours du repas. En admettant que la durée de vie (en secondes) d'une ampoule après son installation puisse être modélisée à l'aide d'une loi géométrique, ce que vient de faire Jojo est-il judicieux ?*

Exercice 145 *Jojo désire coder un long message à l'aide d'un code binaire. Spécifiquement, il cherche à associer à chaque mot du message un mot de code binaire, constitué d'une suite finie de 0 et de 1, et, pour des raisons de facilité de décodage, il souhaite que son code possède la propriété du préfixe : aucun mot du code binaire ne doit être le début d'un autre mot du code. Supposons que le message soit écrit dans un langage très primaire qui ne comporte que 6 mots différents, notés $A_1, \ldots, A_6$, et que, dans le message que Jojo cherche à transmettre, les fréquences de chacun des mots soient les suivantes : A_1 représente 12% des mots du message, A_2 25%, A_3 8%, A_4 11%, A_5 14% , et A_6 30%.*

Quel code pouvez-vous proposer à Jojo afin de minimiser la longueur du message une fois codé ? Quel est le nombre moyen de signes binaires utilisés par votre code pour coder le message de Jojo ? Comment se compare-t-il à l'entropie associée aux fréquences des différents mots dans le message ?

Exercice 146 *Au cours d'une émission, on invite une vingtaine de médiums censés deviner des informations sur des membres du public choisis au hasard (par exemple, leur nombre d'enfants, s'ils sont ou non célibataires, etc...). A chaque étape, les médiums ayant deviné juste restent sur scène, tandis que les autres sont éliminés. Après cinq étapes, M. H*** est le seul à rester en lice, et couronné comme possédant un don vraiment exceptionnel. Pensez-vous que cela soit justifié ? En quoi l'élimination progressive peut-elle tendre à accréditer indûment, – auprès des spectateurs non-avertis, bien entendu – M. H*** ?*

Exercice 147 *(L'affaire du testament Howland)*

Lorsque la riche Mme Sylvia Howland mourut en 1865, il apparut que son testament, daté de 1863, stipulait qu'environ la moitié de sa fortune devait être répartie entre des légataires variés, tandis que l'autre moitié (soit plus d'un million de dollars de l'époque) serait placée, et les intérêts ainsi produits versés à sa nièce, Mme Henrietta Howland Green, à la mort de laquelle le principal serait redistribué entre d'autres légataires.

Mme Howland Green, qui comptait bien hériter de la totalité de la somme, et non pas seulement des intérêts, produisit alors un exemplaire plus ancien du testament, daté de 1862 (donc antérieur à celui effectivement exécuté lors de la succession), qui lui attribuait la quasi-totalité des biens de sa tante Sylvia, accompagné d'une page supplémentaire censée annuler «tout testament rédigé avant ou après celui-ci.» Si l'authenticité du testament de 1862 ne semblait pas devoir être mise en doute (il avait été signé de la défunte Mme Howland et de trois témoins), celle de la page supplémentaire était plus suspecte, celle-ci ne portant la signature que de la défunte et de sa nièce. L'exécuteur testamentaire de Mme Howland refusant d'accorder foi à la seconde partie du document, l'affaire fut portée devant les tribunaux, et plusieurs experts furent convoqués.

Un examen attentif[19] *d'un échantillon de 42 signatures réalisées par la défunte Mme Howland lors de ses dernières années fut mené. Celui-ci révéla d'une part, que chaque signature comportait systématiquement trente traits dirigés vers le bas, et, d'autre part, qu'entre deux signatures quelconques, en moyenne six traits dirigés vers le bas homologues (c'est-à-dire correspondant à un même élément d'une même lettre de la signature) étaient exactement superposables.*

En revanche, en comparant la signature présente sur le testament de 1862 avec celle figurant sur la page supplémentaire de celui-ci, ce fut une coïncidence complète

19. Réalisé pour le compte de l'exécuteur testamentaire de Sylvia Howland, Thomas Mandell, par le célèbre mathématicien et astronome américain Benjamin Peirce, assisté de son non moins célèbre fils Charles Peirce. Voir par exemple `http://www-groups.dcs.st-and.ac.uk/ history/Mathematicians/Peirce_Benjamin.html` et `http://www-groups.dcs.st-and.ac.uk/ history/Mathematicians/Peirce_Charles.html`.

des trente traits qui fut observée, suggérant la possibilité que la signature inscrite sur la page supplémentaire du testament ait été recopiée à partir de l'autre.

Les Peirce affirmèrent qu'au vu de leur étude, on pouvait évaluer la probabilité qu'une telle coïncidence survienne de manière accidentelle à $1/5^{30}$*, soit, d'après les Peirce toujours, environ* $1/2,666... \times 10^{-21}$*, La conclusion était qu'une probabilité si faible indiquait que, selon toute raison, la page supplémentaire du testament était un faux.*

1) Vérifiez que $1/5^{30} \approx 1/2,666... \times 10^{-21}$.

2) Expliquez en quoi cet argument apparaît comme un (bel) exemple du sophisme du procureur. Quelles probabilités aurait-il également fallu évaluer pour tenter de conclure de manière correcte ? Dans quelles conditions pourrait-on néanmoins considérer que les probabilités mettent sérieusement en cause l'authenticité du document produit par Mme Howland Green ?

3) Tentez d'expliquer comment les Peirce ont pu parvenir, à partir de leur étude, à la valeur de $1/5^{30}$*. Sur quelles hypothèses ont-ils pu s'appuyer ? Comment jugez-vous la pertinence et la fiabilité de leur argument ?*

4) Dans le cadre du procès, un échantillon de 110 signatures tracées par l'ancien président des Etats-Unis John Quincy Adams fut analysé, révélant que les douze signatures de l'échantillon les plus proches entre elles présentaient des similarités supérieures à celles observées entre les deux signatures figurant sur le testament de 1862. L'argument fut employé par les avocats de Mme Howland Green pour affirmer qu'une telle similitude pouvait survenir de manière naturelle. Les avocats de la partie adverse rétorquèrent que le président Adams était connu pour posséder une écriture particulièrement uniforme. D'autres exemples de signatures très voisines produites par une même personne furent donnés (entre autres, à partir de chèques bancaires). Quelle est, selon-vous, la portée de ces arguments ?

5) Il fut également proposé qu'une similitude importante pouvait exister entre des signatures réalisées par une même personne à peu de temps d'intervalle, à la même place et sur le même bureau, par exemple. Que pensez-vous de cet argument ?

6) En définitive, si vous deviez étudier vous-même la question, de quelles données chercheriez-vous à disposer, et comment procéderiez-vous ?

Sophisme du procureur ou pas, l'affaire fut tranchée en définitive sur la base d'arguments purement juridiques et complètement indépendants des considérations présentées ci-dessus, qui donnèrent tort à Mme Howland Green. La question de savoir si la cour aurait tranché en sa faveur si la seconde signature avait éét considérée comme authentique, reste ouverte...

Exercice 148 *Considérons une variable aléatoire* X *telle que* $\mathbb{E}(X) > 0$ *et* $\sqrt{\mathbb{V}(X)} << \mathbb{E}(X)$*. Montrez que* $\frac{X}{\mathbb{E}(X)}$ *est typiquement proche de* 1*. Peut-on en déduire que* $|X - \mathbb{E}(X)| << 1$ *? Si inversement* $\sqrt{\mathbb{V}(X)} >> \mathbb{E}(X)$*, peut-on en déduire que* $\frac{X}{\mathbb{E}(X)}$ *a une*

probabilité significative d'être éloignée de 1 *?*

Voir également l'exercice 111.

Exercice 149 *(Le paradoxe de Parrondo)*

On considère les trois jeux suivants. Le premier jeu, noté A, consiste simplement à lancer une pièce de monnaie, le joueur gagnant 1 euro lorsque la pièce retombe sur pile, et perdant un euro lorsque celle-ci retombe sur face. Les probabilités de pile et face ne sont pas a priori égales, et valent respectivement p *et* $1-p$*, où* p *est un paramètre du problème. Le deuxième jeu, noté B, consiste également à lancer une pièce de monnaie en gagant 1 euro lorsque celle-ci retombe sur pile, et en perdant 1 euro lorsque celle-ci retombe sur face, à cette différence près que la pièce que l'on lance est choisie parmi deux pièces (numérotées 1 et 2), aux caractéristiques différentes, le choix de la pièce lancée dépendant du capital total (en euros) dont dispose le joueur avant le lancer. La pièce numéro 1 possède une probabilité* p_1 *de retomber sur pile (et donc* $1-p_1$ *de retomber sur face), tandis que la probabilité pour la pièce numéro 2 de retomber sur pile vaut* p_2 *(d'où une probabilité* $1-p_2$ *de retomber sur face),* p_1 *et* p_2 *étant des paramètres du problème vérifiant* $0 < p_1, p_2 < 1$*. La règle fixant le choix des pièces pour le jeu B est la suivante : si le capital disponible avant le lancer est un multiple de 3, la pièce lancée est la pièce numéro 1, tandis que, dans le cas contraire, c'est la pièce numérotée 2 qui est choisie. Enfin, le troisième jeu, noté C, est en réalité une combinaison de A et de B : avec probabilité 1/2, le joueur joue au jeu A, avec probabilité 1/2, il joue au jeu B.*

Nous allons voir qu'il est possible de choisir p, p_1, p_2 *de telle façon que A et B soient tous les deux, en un certain sens, des jeux perdants à long terme, tandis que C est, quant à lui, gagnant à long terme. Ce résultat est connu sous le nom de paradoxe de Parrondo, et a fait, depuis son apparition en 1999, l'objet de commentaires abondants, en particulier, mais pas seulement, du fait qu'il illustre très simplement les propriétés de certains modèles physiques censés décrire des «moteurs moléculaires».*

Nous allons d'abord préciser le comportement à long terme du jeu B. Supposons donc que l'on joue de manière répétée au jeu B, et notons Y_n *le montant total disponible après le* $n-$*ème lancer, réduite modulo 3. Définissons ensuite ensuite, pour* $n \geq 0$*, le vecteur ligne* ν_n *dont les coordonnées donnent la loi de* Y_n*, c'est-à-dire, pour* $i = 0, 1, 2$*,* $\nu_n(i) := \mathbb{P}(Y_n = i)$*.*

2) Montrez que la relation suivante est satisfaite pour tout $n \geq 0$ *:* $\nu_{n+1} = M\nu_n$*, où* ν_n *est le vecteur ligne dont les coordonnées sont* $(\nu_n(0), \nu_n(1), \nu_n(2))$*, et où* M *est la matrice suivante :*

$$M := \begin{pmatrix} 0 & p_1 & 1-p_1 \\ 1-p_2 & 0 & p_2 \\ p_2 & 1-p_2 & 0 \end{pmatrix}.$$

En déduire que, pour tout $n \geq 1$, $\nu_n = M^n \nu_0$.
3) Montrez que les solutions de l'équation $\nu = M\nu$, où ν est un vecteur ligne dont les coordonnées, sont de la forme :

$$\nu = (\nu(0), \nu(1), \nu(2)) = \lambda(1 - p_2 + p_2^2, 1 - p_2 + p_1 p_2, 1 - p_1 + p_1 p_2),\ \lambda \in \mathbb{R}.$$

En déduire qu'il existe une unique solution dont les coordonnées décrivent une loi de probabilité sur l'ensemble $\{0, 1, 2\}$. Soit $\nu_* = (\nu_*(0), \nu_*(1), \nu_*(2))$ cette solution.

Nous admettrons (il s'agit en fait d'un résultat général provenant de la théorie des chaînes de Markov) que, quelle que soit la valeur de ν_0, on a toujours $\lim_{n \to +\infty} \nu_0 M^n = \nu_\infty$.
4) En déduire, en fonction de p_1 et p_2, la valeur limite l'espérance de gain au $n-$ème pas en jouant de manière répétée au jeu B, lorsque n tend vers l'infini. A quelle condition celle-ci est-elle négative ?
5) Même question avec le jeu C : quelle est la valeur limite, en fonction de p, p_1, p_2, de l'espérance de gain au $n-$ème pas en jouant de manière répétée au jeu C, lorsque n tend vers l'infini, et à quelle condition celle-ci est-elle négative ? Indication : reprendre la stratégie employée pour les questions 2) et 3).
6) Donnez un exemple de valeur de p, p_1, p_2 pour lequel les jeux A et B sont perdants à long terme, tandis que C est gagnant à long terme.

Exercice 150 *On considère une variable aléatoire X dont la loi possède une densité de la forme $f(x) = Kx^{-c}$ pour $x \geq b$, avec $b > 0$ et $c > 1$.*
1) Montrez que la valeur de K est entièrement déterminée par la donnée de c et de b.
2) Etant donné un nombre $a \geq b$, on se concentre sur les valeurs de X supérieures ou égales à a, autrement dit, on s'intéresse à la loi de X conditionnelle au fait que $X \geq a$. Pour pouvoir comparer entre elles des lois associées à différentes valeurs de a, on ramène la valeur de a à l'échelle 1, en considérant la loi de X/a conditionnelle au fait que $X \geq a$. Montrez que cette loi ne dépend en fait pas de a. Cette propriété est ce que l'on appelle l'invariance d'échelle de la loi de X : les valeurs de X supérieures à une valeur donnée ont exactement (après mise à l'échelle) la même distribution de probabilité que X.
3) Supposons maintenant que X suive une loi exponentielle de paramètre $\lambda > 0$. Quelle est cette fois la loi de X/a conditionnelle au fait que $X \geq a$?

Exercice 151 *Deux lignes d'autobus, les lignes 1 et 2, effectuent la liaison entre la gare de Jojo-les-Pins et la place du marché, située au centre-ville. Les bus de la ligne 1, sans arrêt ou presque sur ce trajet, effectuent la liaison en 10 minutes en moyenne. En revanche, les bus de la ligne 2 comportent plusieurs arrêts sur le parcours, et effectuent la liaison en 20 minutes, toujours en moyenne. Pour simplifier,*

on suppose on supposera que les durées de parcours sont toujours exactement égales à 10 et à 20 minutes, respectivement pour les lignes 1 et 2. On modélise la durée de l'attente de l'autobus pour un passager venant d'arriver à la gare, et souhaitant prendre la ligne 1, par une variable aléatoire de loi exponentielle de paramètre λ_1. La même loi est employée pour un passager attendant un bus de la ligne 2, mais avec un paramètre λ_2 a priori différent de λ_1.
1) Quelles hypothèses de modélisation sous-jacente pourrait expliquer l'emploi de lois exponentielles dans ce contexte ?
2) Quel est en moyenne le temps total (attente plus trajet) pour un passager arrivant à la gare et souhaitant se rendre place du marché en utilisant un bus de la ligne 1 ? Même question avec un bus de la ligne 2 ? A quelle condition est-il plus avantageux de prendre la ligne 1 que la ligne 2 ?
3) Considérons à présent un passager choisissant de se rendre au marché par le premier autobus (de la ligne 1 ou de la ligne 2) qui arrive. En supposant l'indépendance entre le temps d'attente d'un bus de la ligne 1 et d'un bus de la ligne 2, quelle est la loi du temps d'attente de ce passager avant de pouvoir monter dans un bus ? Quel est le temps total moyen mis par le passager pour se rendre à destination ? Comment ceci se compare-t-il, en fonction de λ_1 et λ_2, au choix le plus avantageux obtenu à la question 2) ? Donnez des exemples de valeurs numériques réalistes pour lesquelles cette comparaison a lieu dans un sens, et dans l'autre.

Exercice 152 *Une girafe cherche (mais pourquoi ?) à traverser une route étroite, la durée nécessaire pour qu'elle effectue sa traverée étant estimée à un nombre a de minutes. On suppose qu'il passe en moyenne 6 véhicules par minute sur cette route à l'endroit où la girafe cherche à traverser, et, plus précisément, que le nombre total de véhicules traversant la route au cours d'une période de temps donnée de a minutes suit une loi de Poisson de paramètre proportionnel à a.*
1) Quelles hypothèses de modélisation sous-jacentes le choix de cette loi de Poisson peut-il traduire ?
2) Pour quelles valeurs de a la girafe a-t-elle moins de 5% de chances d'entrer en collision avec un véhicule ? Pour quelles valeurs de a cette probabilité est-elle supérieur à 95% ?

Exercice 153 *Un laboratoire d'analyses médicales effectue des tests sanguins destinés à détecter la présence d'une certaine substance dans le sang des personnes sur lesquelles l'analyse est pratiquée. Une première manière de procéder pour le laboratoire consiste simplement à effectuer individuellement un test sur chacun des échantillons recueillis. Compte-tenu du coût unitaire élevé des tests, le laboratoire envisage de réduire le nombre de ceux-ci en procéadant de la manière suivante. Deux échantillons, au lieu d'un seul, sont prélevés sur chacune des personnes concernées.*

On divise ensuite l'ensemble des personnes testées en groupes comportant chacun m individus. Pour chaque groupe, on procéde alors de la manière suivante : un échantillon de chaque personne du groupe est utilisé pour être mélangé aux autres, et le test de détection est pratiqué sur le mélange ainsi obtenu. Si le résultat de ce test est négatif, on considère que l'ensemble des personnes faisant partie du groupe obtient un résultat négatif pour le test. Inversement, si le résultat est positif, on teste séparément chacun des échantillons individuels restants pour les personnes de ce groupe. En supposant que la sensibilité du test est suffisante pour qu'un seul échantillon contenant la substance entraîne la détection, même si celui-ci est mélangé à d'autres échantillons qui ne la contiennent pas, on cherche à déterminer si cette méthode est réellement avantageuse par rapport à la première solution consistant à tester individuellement chaque échantillon. En supposant que l'on peut modéliser le présence/absence de la substance dans le sang de l'ensemble des personnes testées par une répétition indépendante de variables de Bernoulli, discutez de la comparaison du nombre moyen de tests effectués entre ces deux méthodes.

Chapitre 3

Loi des grands nombres

3.1 Introduction

La loi des grands nombres constitue le premier des «théorèmes limites» de la théorie des probabilités. Dans sa version la plus simple, elle affirme que la moyenne d'un grand nombre de variables aléatoires à valeurs réelles, indépendantes et de même loi est, typiquement, approximativement égale à l'espérance commune de ces variables aléatoires, lorsque celle-ci existe. Dans ce chapitre, nous présentons et discutons différentes versions de ce résultat, leur interprétation et leur portée pratique.

3.2 Loi faible des grands nombres

3.2.1 Cadre et hypothèses

Considérons un espace de probabilité $(\Omega, \mathbb{P})$, et une variable aléatoire X définie sur Ω et à valeurs dans $\mathbb{R}$. Considérons ensuite l'espace de probabilité $(\Omega^N, \mathbb{P}^{\otimes N})$ décrivant N **répétitions indépendantes** de $(\Omega, \mathbb{P})$, et notons $X_1, \ldots, X_N$ les variables aléatoires correspondant à X dans chacune des N réalisations successives. De manière plus précise (voir le chapitre précédent), les variables aléatoires X_i sont définies par Ω^N par $X_i((\omega_1, \ldots, \omega_N)) = X(\omega_i)$.

On vérifie que les variables aléatoires $X_1, \ldots, X_N$ sont mutuellement indépendantes, et qu'elles possèdent toutes la même loi que X.

On note que, partant de n'importe quel modèle probabiliste $(\mathcal{W}, \mathbb{Q})$ sur lequel est définie une famille de variables aléatoires $Y_1, \ldots, Y_N$ mutuellement indépendantes et possédant chacune la même loi, on peut se ramener à la situation décrite ci-dessus en considérant le modèle-image de $(Y_1, \ldots, Y_N)$. Par conséquent, la loi des grands nombres, que nous énoncerons dans le paragraphe suivant, s'applique dans ce cadre général.

Nous considérerons la variable aléatoire $\frac{1}{N}(X_1 + \cdots + X_N)$. Celle-ci représente la moyenne arithmétique des valeurs de X obtenues au cours des N répétitions indépendantes (il s'agit d'une variable aléatoire, puisque chaque X_i est elle-même une variable aléatoire). En d'autres termes, la variable aléatoire $\frac{1}{N}(X_1 + \cdots + X_N)$ n'est autre que la moyenne empirique associée à l'échantillon de valeurs (aléatoires) $X_1, \ldots, X_N$.

Un cas particulier très important est celui où la variable aléatoire X est la fonction indicatrice d'un événément A de Ω. Dans ce cas, $\frac{1}{N}(X_1 + \cdots + X_N)$ n'est autre que la proportion de fois où l'événement A s'est réalisé au cours des N répétitions, que nous noterons $f_N(A)$.

Une hypothèse fondamentale pour la suite est que **la variable aléatoire X possède une espérance.** Dans le cas d'une indicatrice, cette espérance existe toujours, et n'est autre que la probabilité $\mathbb{P}(A)$.

3.2.2 Enoncé

Dans le cadre et sous les hypothèses décrits dans le paragraphe précédent, c'est-à-dire N variables aléatoires $X_1, \ldots, X_N$ représentant N répétitions indépendantes d'une variable aléatoire X possédant une espérance, la **loi faible des grands nombres** affirme que, pour tout $\epsilon > 0$,

$$\lim_{N \to +\infty} \mathbb{P}^{\otimes N}\left(\left|\frac{1}{N}(X_1 + \cdots + X_N) - \mathbb{E}(X)\right| \geq \epsilon\right) = 0.$$

Dans le cas particulier où X est la fonction indicatrice d'un événément A, la loi des grands nombres se réécrit, en notant $f_N(A)$ la proportion de fois où l'événement A s'est réalisé au cours des N répétitions, sous la forme suivante.

$$\lim_{N \to +\infty} \mathbb{P}^{\otimes N}\left(|f_N(A) - \mathbb{P}(A)| \geq \epsilon\right) = 0.$$

Ainsi, étant donné un $\epsilon > 0$ fixé, mais que l'on peut choisir arbitrairement petit, la probabilité pour que $\frac{1}{N}(X_1 + \cdots + X_N)$ soit éloigné de $\mathbb{E}(X)$ d'un écart supérieur à ϵ, tend vers zéro lorsque N tend vers l'infini. En d'autres termes, lorsque N tend vers l'infini, la loi de la variable aléatoire $\frac{1}{N}(X_1 + \cdots + X_N)$ se concentre autour de la valeur $\mathbb{E}(X)$.

En termes plus imagés, la loi des grands nombres affirme donc que, **lorsque N est suffisament grand**, la variable aléatoire $\frac{1}{N}(X_1 + \cdots + X_N)$ est, **typiquement** (avec une probabilité proche de 1), **approximativement** (à ϵ près) **égale à l'espérance** $\mathbb{E}(X)$.

Dans le cas d'une indicatrice, on obtient que la proportion de fois où A se produit est typiquement approximativement égale à la probabilité de A.

Remarque 12 *Soulignons que l'on ne peut espérer se passer, dans l'énoncé ci-dessus, d'aucun des deux termes «approximativement» et «typiquement».*

Pour s'en convaincre, il suffit de penser à l'exemple du jeu de pile ou face, modélisé par une suite de lancers indépendants donnant lieu à pile ou face de manière équiprobable. Après 10000 lancers, on peut s'attendre à ce que la proportion observée de pile soit proche de 1/2, mais certainement pas à obtenir exactement 5000 fois pile et 5000 fois face. De même, il est physiquement **possible** *que l'on obtienne 10000 fois face au cours des 10000 lancers, et il est donc physiquement possible que la proportion observée de pile soit très différente de 1/2. Simplement, une telle éventualité est extrêmement* **improbable** *(dans le modèle de lancers indépendants avec équiprobabilité de pile et de face), et c'est pourquoi, bien que l'on ne puisse pas exclure le fait qu'elle puisse survenir, on s'attend à ce que typiquement, elle ne se produise pas.*

Soulignons que la loi des grands nombres énoncée ci-dessus est un théorème mathématique, qui nécessite pour que l'on puisse l'appliquer que ses hypothèses (variables aléatoires indépendantes et de même loi possédant une espérance) soient satisfaites (voir également à ce sujet le paragraphe sur la robustesse de la loi des grands nombres) que nous allons démontrer dans le paragraphe suivant.

3.2.3 Preuve

Pour simplifier, nous donnerons une preuve en nous plaçant sous l'hypothèse supplémentaire selon laquelle la variance de X, et non seulement son espérance, est définie.

D'abord, on vérifie que l'espérance de la variable aléatoire $N^{-1}(X_1 + \cdots + X_N)$ est égale à $\mathbb{E}(X)$, grâce à la propriété de linéarité de l'espérance :

$$\mathbb{E}\left(\frac{1}{N}(X_1 + \cdots + X_N)\right) = \frac{1}{N}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_N)) = \frac{1}{N}(N \times \mathbb{E}(X)) = \mathbb{E}(X),$$

chaque X_i possédant individuellement la même loi que X, et donc la même espérance. Nous cherchons à montrer que cette espérance est également la valeur typique de $N^{-1}(X_1 + \cdots + X_N)$. Pour cela, nous pouvons étudier la variance :

$$\mathbb{V}\left(\frac{1}{N}(X_1 + \cdots + X_N)\right) = \frac{1}{N^2}\mathbb{V}(X_1 + \cdots + X_N).$$

Du fait de **l'indépendance mutuelle des variables aléatoires** X_i, que nous avons supposée, les variances des variables aléatoires X_i s'ajoutent, et l'on a :

$$\mathbb{V}(X_1 + \cdots + X_N) = \mathbb{V}(X_1) + \cdots + \mathbb{V}(X_N) = N \times \mathbb{V}(X),$$

en utilisant le fait que toutes les possèdent individuellement la même loi que X, et donc la même variance. En définitive, on obtient que :

$$\mathbb{V}\left(\frac{1}{N}\left(X_1 + \cdots + X_N\right)\right) = \frac{\mathbb{V}(X)}{N}.$$

La variance de la moyenne empirique associée à N réalisations indépendantes de la variable aléatoire X est donc N fois plus petite que la variance de X. Cette égalité traduit donc le fait que la moyenne empirique fluctue d'autant moins autour de son espérance $\mathbb{E}(X)$ que N est grand. Plus précisément, l'inégalité de Bienaymé-Tchebychev (voir le chapitre «Variables aléatoires») entraîne que, pour tout $\epsilon > 0$,

$$\mathbb{P}^{\otimes N}\left(\left|\frac{1}{N}\left(X_1 + \cdots + X_N\right) - \mathbb{E}(X)\right| \geq \epsilon\right) \leq \frac{\mathbb{V}(X)}{N\epsilon^2},$$

ce qui implique la loi des grands nombres énoncée plus haut, en prenant la limite lorsque N tend vers l'infini.

Notons que, malgré son aspect anodin, la propriété d'additivité des variances dans le cas de variables aléatoires indépendantes est la clef de la preuve ci-dessus : *a priori*, on pourrait s'attendre à ce que la variance de $X_1 + \ldots + X_N$ soit une quantité d'ordre N^2, car elle fait intervenir le carré de quantités d'ordre N (somme de N variables aléatoires). Le fait que cette variance s'avère en réalité être d'ordre N (du fait de l'additivité des variances) est donc un résultat non banal (provenant de l'indépendance des variables aléatoires $X_1, \ldots, X_N$) !

3.2.4 Qu'est-ce qu'un grand nombre ?

La loi des grands nombres telle que nous l'avons énoncée ci-dessus est un résultat asymptotique affirmant qu'une certaine probabilité tend vers zéro lorsque N tend vers l'infini. Une question fondamentale, si l'on souhaite tirer des conséquences pratiques de ce résultat, et donc l'extrapoler à des valeurs de N grandes mais finies, consiste donc à se demander à partir de quelle valeur de N on peut considérer l'approximation $\frac{1}{N}\left(X_1 + \cdots + X_N\right) \approx \mathbb{E}(X)$ comme satisfaisante. D'après ce qui précède, une manière précise de poser le problème est de fixer deux nombres $\epsilon > 0$ et $0 < \alpha < 1$ et de demander quelle à partir de quelle valeur de N l'inégalité [1]

$$\mathbb{P}^{\otimes N}\left(\left|\frac{1}{N}\left(X_1 + \cdots + X_N\right) - \mathbb{E}(X)\right| \geq \epsilon\right) \leq \alpha, \tag{3.1}$$

est valable. Comme nous l'avons observé précédemment, mais il n'est peut-être pas inutile d'insister, on ne peut se passer ni du ϵ ni du α pour aborder cette question, ceux-ci permettant de quantifier le «approximativement» (ϵ) et le «typiquement» (α) intervenant dans la loi des grands nombres.

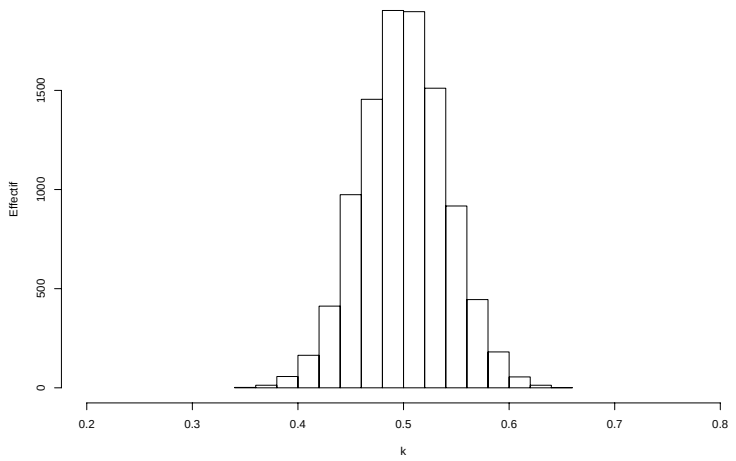
1. Une inégalité telle que (3.1) est souvent appelée **inégalité de déviation**.

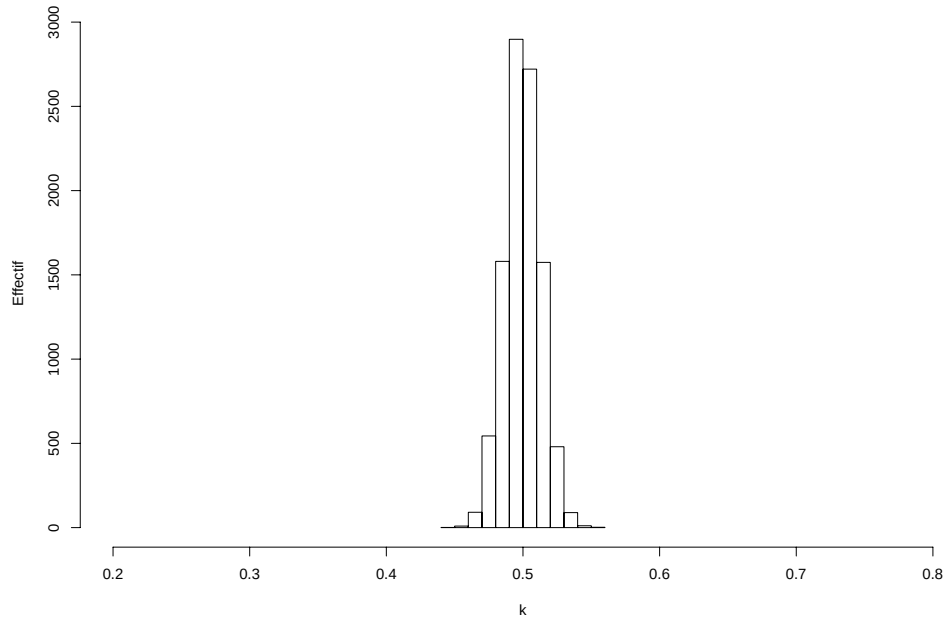
La première chose à retenir à ce sujet est la suivante : la valeur d'un N tel que l'inégalité (3.1) soit valable dépend de ϵ, de α, et de la loi de X. En aucun cas il ne peut exister de nombre N «grand» dans l'absolu, qui permettrait de garantir que l'approximation $\frac{1}{N}(X_1 + \cdots + X_N) \approx \mathbb{E}(X)$ est satisfaisante pour toute valeur de ϵ, ou de α, ou de X.

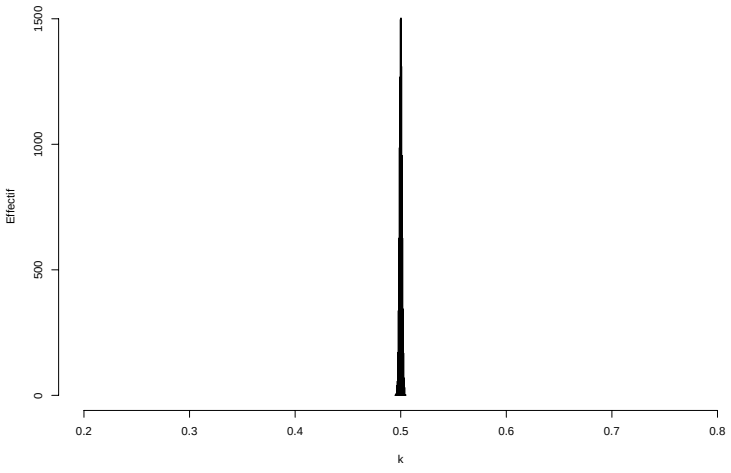
Afin d'illustrer ce point, voici quelques simulations effectuées, comme toujours dans ce cours, à l'aide du logiciel `R`.

Les histogrammes ci-dessous représentent, pour diverses valeurs de N, la répartition empirique obtenue en effectuant 10000 simulations de $\frac{1}{N}(X_1 + \cdots + X_N)$, les X_i étant mutuellement indépendantes et toutes de loi uniforme sur $[0, 1]$.

Voici les histogrammes obtenus respectivement pour $N = 50$, $N = 500$ et $N = 50000$. On constate, conformément à la loi des grands nombres, que ces histogrammes sont de plus en plus concentrés autour de la valeur $1/2$ lorsque N croît. Un tel histogramme de la loi de $\frac{1}{N}(X_1 + \cdots + X_N)$ permet facilement d'estimer, pour un α donné, quelle est la plus petite valeur de ϵ telle que l'inégalité (3.1) soit satisfaite, ou, inversement, étant donné ϵ, de trouver la plus petite valeur de α telle que l'inégalité (3.1) soit satisfaite (le tout pour la valeur de N correspondant à l'histogramme, bien entendu).

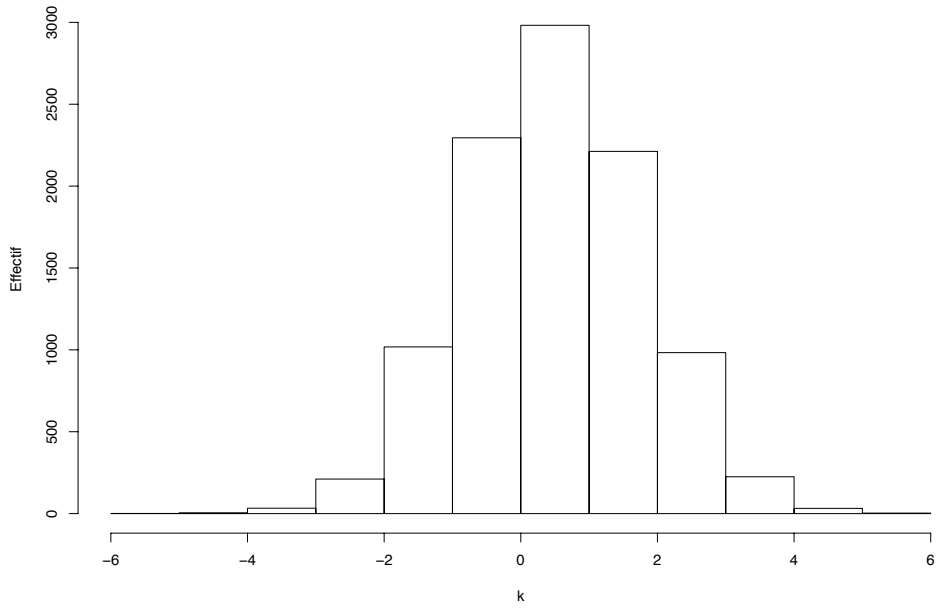


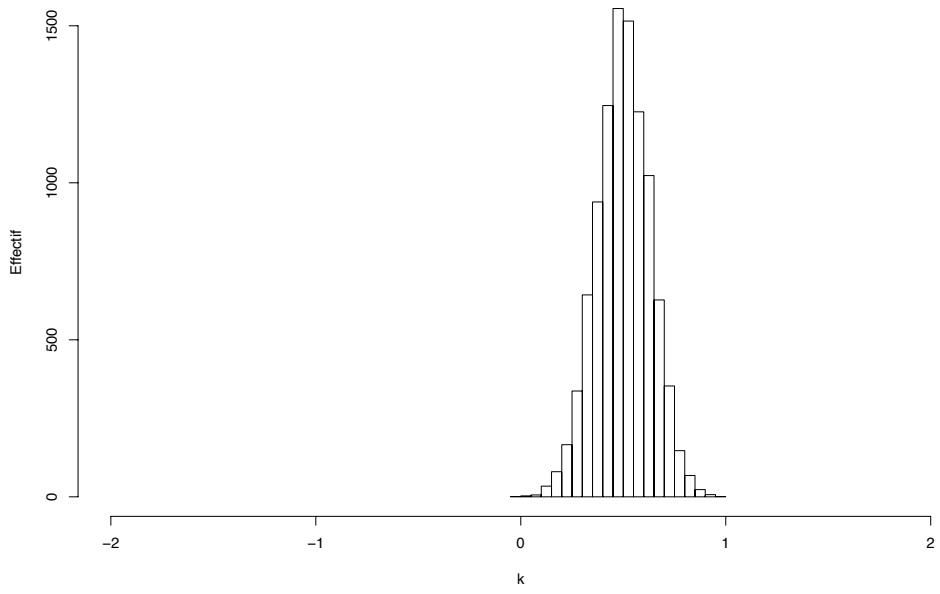




Effectuons à présent des simulations avec des variables aléatoires X_i de loi uniforme sur $[-49,5; 50,5]$, pour lesquelles on a encore $\mathbb{E}(X) = 1/2$, et donc exactement le même énoncé de la loi des grands nombres.

Voici les histogrammes correspondant respectivement à $N = 500$ et $N = 50000$, obtenus, comme précédemment, au cours de 10000 simulations.

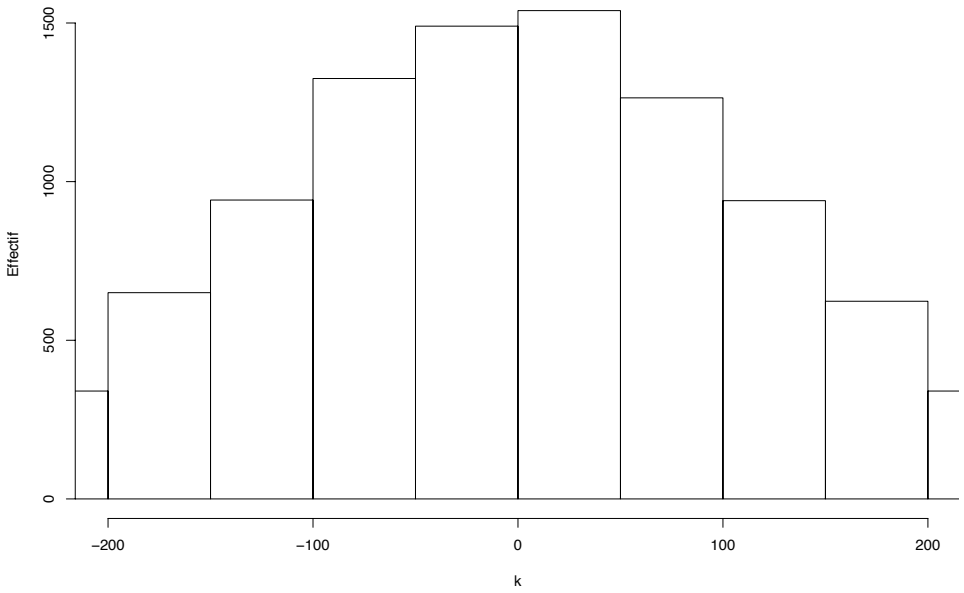


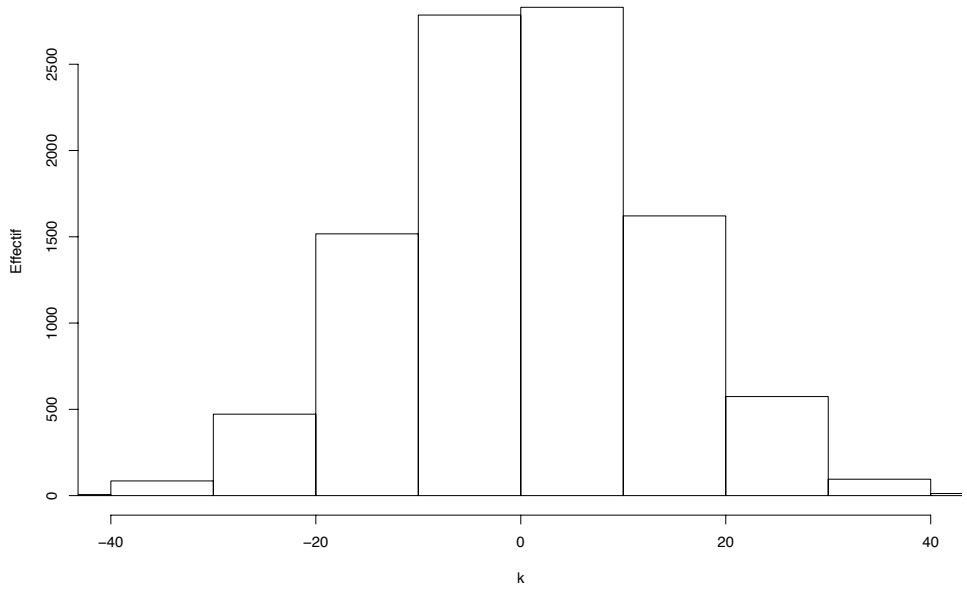


On constate que la concentration de la loi autour de 1/2 est beaucoup moins

nette que dans le cas des variables aléatoires uniformes sur $[0,1]$, les fluctuations aléatoires autour de $1/2$ s'avérant beaucoup plus importantes, de telle sorte que l'approximation $\frac{1}{N}(X_1+\cdots+X_N)\approx 1/2$ est nettement moins bonne (de l'ordre de l'unité pour $N=500$, de l'ordre du dixième pour $N=50000$).

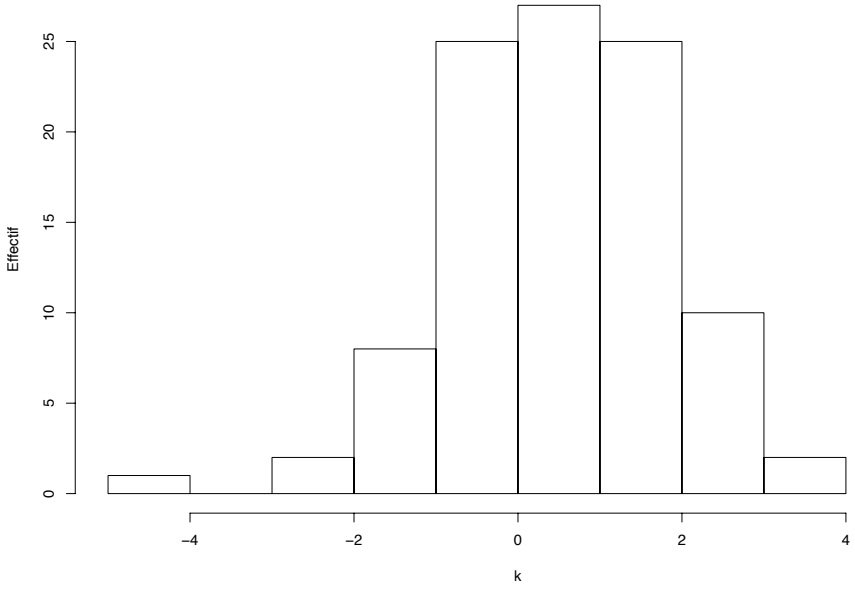
En reprenant les mêmes expériences, avec cette fois des variables aléatoires uniformes sur $[-4999,5;5000,5]$, on obtient les histogrammes suivants pour $N=500$ et $N=50000$ (toujours avec 10000 tirages), les voici.





On constate que la concentration autour de $1/2$ est encore moins nette, et que les flucutations aléatoires autour de $1/2$ sont si importantes que l'approximation $\frac{1}{N}(X_1 + \cdots + X_N) \approx 1/2$ semble perdre sa pertinence : pour $N = 500$, les écarts se mesurent en centaines, et en dizaines pour $N = 50000$.

En allant chercher des valeurs de N plus grandes, l'approximation obtenue s'améliore. Voici par exemple l'histogramme obtenu avec $N = 5000000$ et 100 expériences (on n'effectue que 100 expériences au lieu de 10000 comme précéndemment, car, vu les nombres de variables aléatoires à générer, les simulations commencent à consommer du temps !). On constate que les écarts sont de l'ordre de l'unité.



En augmentant encore les valeurs de N, on parviendrait à obtenir des valeurs typiquement encore plus proches de $1/2$.

Nous retiendrons de ces quelques expériences qu'aucune valeur de N n'est suffisament «grande» dans l'absolu pour que l'on puisse systématiquement considérer que $\frac{1}{N}(X_1 + \cdots + X_N) \approx \mathbb{E}(X)$ avec une probabilité raisonnable. Suivant les cas, une valeur de N égale à 500, 5000, ou même 50 pourra être suffisante pour obtenir une approximation correcte. Dans d'autres cas, même une valeur de 5000000 pourra donner lieu à une approximation très médiocre. Tout ceci dépend de la loi commune des variables aléatoires X_i que l'on ajoute (et naturellement aussi de la manière précise dont on définit l'approximation, par exemple : quelles valeurs de ϵ et de α sont considérées comme satisfaisantes). Nous aborderons cette question de manière plus systématique dans le chapitre suivant, ainsi que dans la partie de ce chapitre consacrée aux inégalités de déviation.

Remarquons simplement que, dans les exemples précédents, plus la variable aléatoire X_i a tendance à fluctuer, plus la variable aléatoire $\frac{1}{N}(X_1 + \cdots + X_N)$ a elle-même tendance à fluctuer, et ceci se retrouve dans le calcul de la variance de $\frac{1}{N}(X_1 + \cdots + X_N)$, qui nous a servi a prouver la loi faible des grands nombres.

3.2.5 Attention à l'approximation

Comme toujours lorsque l'on considère la question de l'approximation d'une quantité par une autre, il convient d'être prudent, et en tout cas précis, quant au type d'approximation utilisé. L'une des erreurs les plus fréquentes consiste à confondre approximation en valeur absolue et approximation en valeur relative, ce qui peut conduire à diverses aberrations.

Par exemple, si l'on choisit un ϵ petit devant 1, mais grand devant $\mathbb{E}(X)$, le fait de savoir qu'avec une forte probabilité on a $\left|\frac{1}{N}(X_1 + \cdots + X_N) - \mathbb{E}(X)\right| < \epsilon$ ne permet en aucun cas d'affirmer que $\frac{1}{N}(X_1 + \cdots + X_N)/\mathbb{E}(X)$ est voisin de 1 avec forte probabilité. Il faudrait pour cela choisir ϵ petit, non seulement devant 1, mais également devant $\mathbb{E}(X)$, ce qui est d'ailleurs impossible si $\mathbb{E}(X) = 0$. Voir à ce sujet l'exercice 161.

Notons également que le fait que $\frac{1}{N}(X_1 + \cdots + X_N) = \mathbb{E}(X) + \epsilon$ avec $\epsilon << 1$ n'entraîne certainement pas que $X_1 + \cdots + X_N = \mathbb{E}(X) + \eta$ avec $\eta << 1$. Tout ce que l'on peut déduire est que $X_1 + \cdots + X_N = \mathbb{E}(X) + N\epsilon$, et, $N\epsilon$ peut aussi bien être $<< 1$, $>> 1$, que de l'ordre de 1, suivant les cas.

3.2.6 Loi forte des grands nombres

Une autre manière d'énoncer la loi des grands nombres, qui peut sembler plus naturelle (mais n'est pas mathématiquement équivalente), est fournie par ce que l'on appelle habituellement la **loi forte des grands nombres**. Précisément, celle-ci affirme que, sous les mêmes hypothèses que la loi faible,

$$\lim_{N\to+\infty} \frac{1}{N}(X_1 + \cdots + X_N) = \mathbb{E}(X) \text{ avec une probabilité égale à 1.}$$

En d'autres termes (c'est la définition même de la notion de limite) pour tout $\epsilon > 0$, on pourra trouver un indice m tel que, pour tout $N \geq m$, on ait

$$\left|\frac{1}{N}(X_1 + \cdots + X_N) - \mathbb{E}(X)\right| \leq \epsilon.$$

Notez que le «typiquement» de la formulation de la loi faible a disparu : pourvu que N soit suffisament grand, on est **certain** que l'écart entre $\frac{1}{N}(X_1 + \cdots + X_N)$ et $\mathbb{E}(X)$ est inférieur à ϵ. En revanche, et c'est là que réside la subtilité de la formulation de la loi forte, le «suffisament grand» est devenu aléatoire : le nombre de répétitions qu'il est nécessaire d'effectuer pour que l'écart devienne inférieur à ϵ est lui-même une variable aléatoire, sur la valeur de laquelle on ne peut imposer aucune borne certaine. En reprenant l'exemple des lancers successifs d'une pièce de monnaie, on voit bien que, pour tout entier k, il est possible (même si c'est très improbable lorsque k est grand) que la pièce retombe sur face lors des k premiers lancers, et qu'il faille donc

effectuer strictement plus de k répétitions avant de parvenir à un écart inférieur, par exemple, à $0,1$. Ainsi, même avec cette formulation de la loi des grands nombres, on ne peut en véritablement échapper au «typiquement» dans l'énoncé de la loi des grands nombres.

Nous ne pousserons pas beaucoup plus cette discussion, si ce n'est pour mentionner une difficulté, à la fois technique et conceptuelle, qui intervient dans la formulation de la loi forte des grands nombres. Plaçons-nous dans le contexte de l'exemple le plus simple, celui d'une répétition de lancers de pile ou face. Pour pouvoir décrire la limite de $\frac{1}{N}(X_1 + \cdots + X_N)$, le modèle $(\Omega, \mathbb{P})$ doit décrire la totalité des lancers donnant lieu à $X_1, X_2, \ldots$, soit une infinité dénombrable de lancers successifs. L'espace Ω peut alors naturellement être représenté par un arbre binaire régulier de profondeur infinie, les probabilités conditionnelles associées aux arêtes étant par exemple prises toutes égales à $1/2$. Mais...cet arbre ne possède pas de feuilles, et il n'y a donc pas d'éventualités élémentaires dans le modèle... Si l'on cherche à effectuer des produits le long des rayons infinis de l'arbre, on obtient des probabilités dont la limite est toujours égale à zéro... On peut certes tronquer l'arbre à une certaine profondeur, ce qui nous permet de calculer les probabilités relatives à un nombre donné de lancers, mais on ne peut alors rien calculer directement concernant la suite infinie des lancers (qui intervient pourtant dans la définition de la limite que l'on cherche à étudier). Il ne s'agit pas de difficultés anecdotiques, que l'on pourrait contourner en étant simplement astucieux, mais de problèmes de fond qui apparaissent dès que l'on s'autorise à considérer des suite infinies d'expériences aléatoires. Une approche satisfaisante de ce type de problème est fournie par la théorie de Kolmogorov, qui replace le calcul des probabilités dans le cadre de la théorie mathématique de la mesure et de l'intégration, mais il s'agit malheureusement d'une théorie beaucoup trop difficile pour que nous puissions ne serait-ce que songer à l'aborder dans le cadre de ce cours.

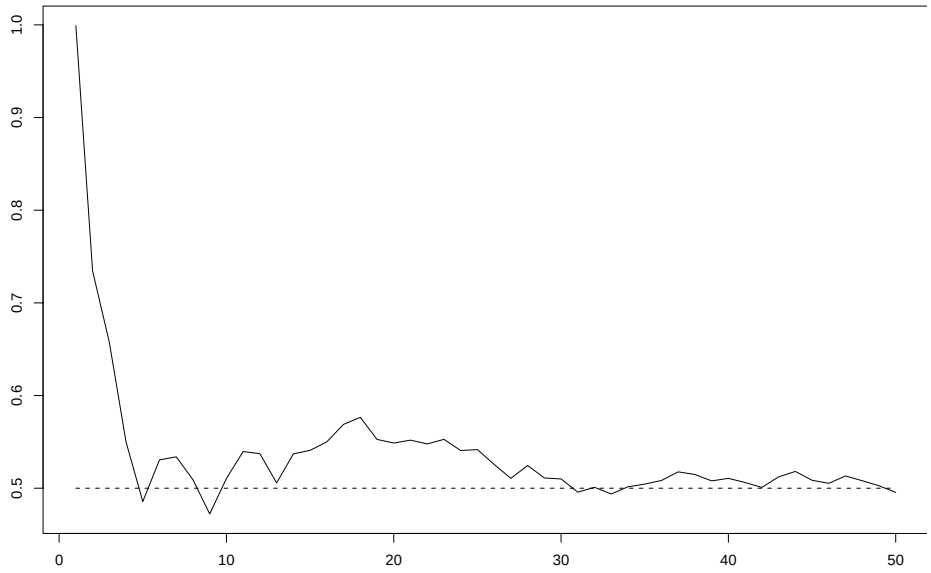
Voici en revanche des illustrations expérimentales reprenant celles effectuées précédemment, mais dans l'esprit de la loi forte des grands nombres.

Cette fois, on effectue une simulation d'un grand nombre de variables aléatoires $X_1, \ldots, X_N$ indépendantes et de même loi, et l'on représente $\frac{1}{i}(X_1 + \cdots + X_i)$ en fonction de i pour $1 \leq i \leq N$. D'après la loi forte des grands nombres, on s'attend à observer la convergence de la suite en question vers $\mathbb{E}(X)$. Répéter plusieurs fois ce type de simulation permet de se convaincre (ou tout au moins de fournir une illustration) du fait que cette convergence a lieu de manière systématique. Nous vous invitons à examiner attentivement la différence existant entre cette représentation et celle donnée précédemment.

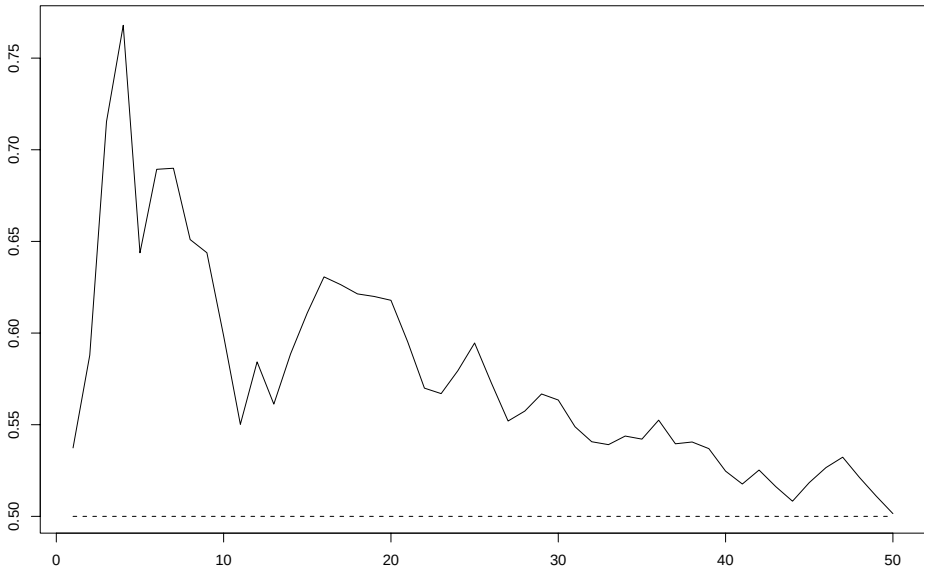
Commençons avec des suites de variables aléatoires indépendantes de loi uniforme sur $[0,1]$, pour lesquelles on a donc $\mathbb{E}(X) = 1/2$.

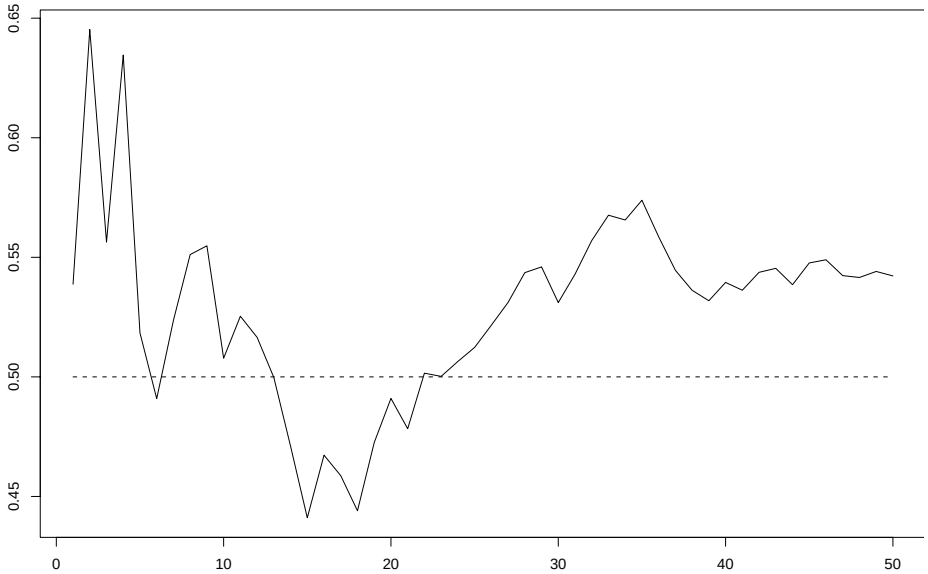
Voici une première représentation graphique de $\frac{1}{i}(X_1 + \cdots + X_i)$ en fonction de i obtenue pour i allant de $i = 1$ à $i = 50$. Les points ont été reliés entre eux par des

portions de droite, mais il faut se rappeler qu'il s'agit en réalité de points dont les coordonnées horizontales sont des nombres entiers.



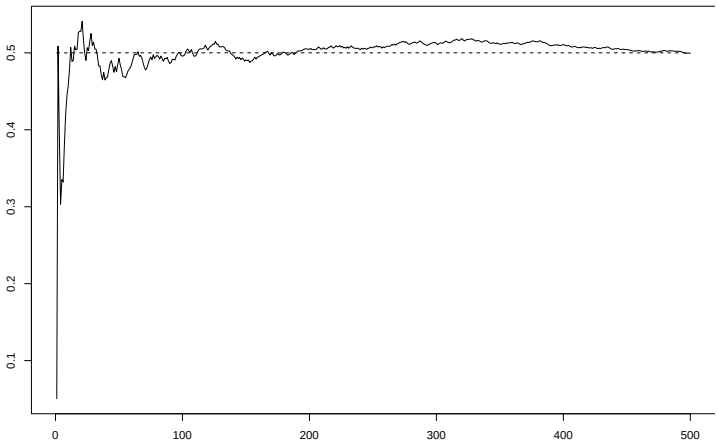
Conformément à nos attentes, la courbe obtenue se rapproche, en gros, de $1/2$ à mesure que i croît. En recommençant l'expérience une deuxième et une troisième fois, on obtient les deux courbes suivantes, qui présentent un comportement en gros comparable (si vous avez l'impression contraire, prenez garde à l'échelle verticale, qui change d'une figure à l'autre !), mais qui ne sont en aucun cas identiques. Les courbes présentent un caractère aléatoire, même si elles suggèrent toutes la convergence vers $1/2$.

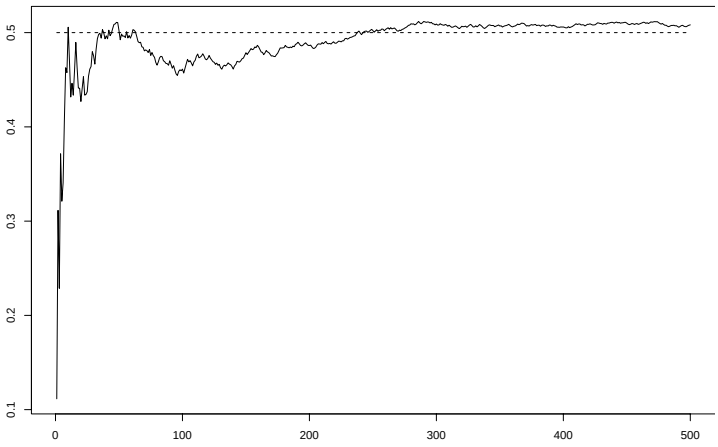




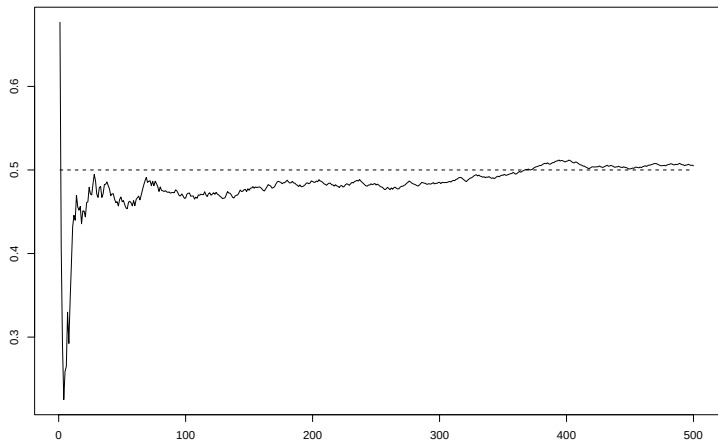
Voici à présent les courbes obtenues en suivant le même principe, mais avec i

allant de 1 à 500.

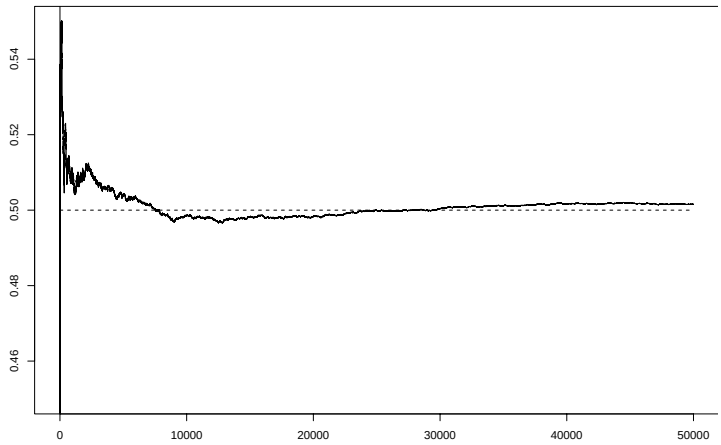


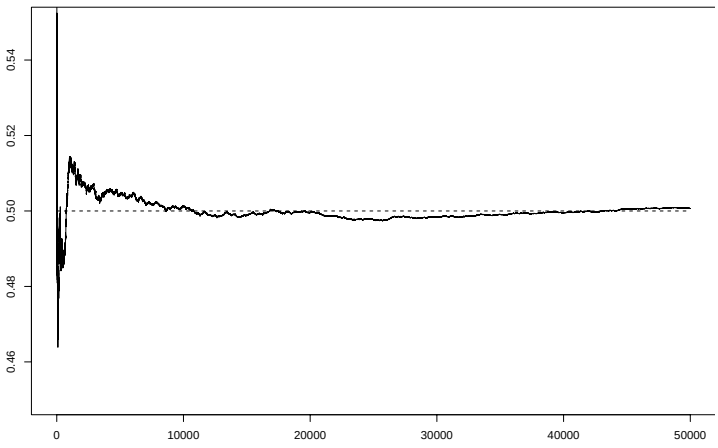


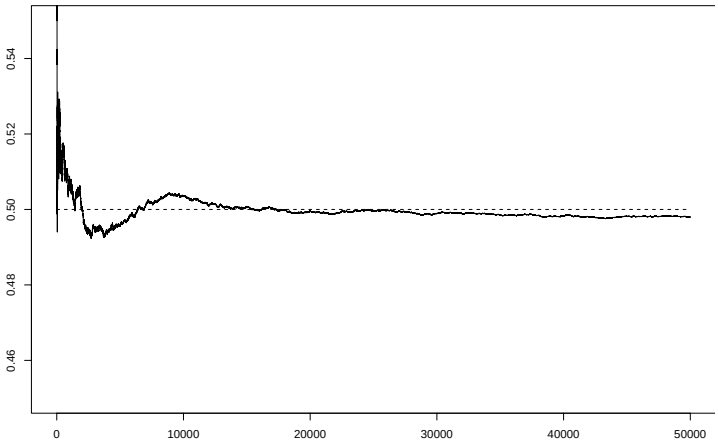
300



En recommençant avec i allant de 1 à 50000, et en restreignant l'échelle verticale (si bien que certaines portions de la courbe dépassent du cadre de la figure et sont donc tronquées, mais que l'on en observe plus précisément la fin), on obtient les courbes suivantes.



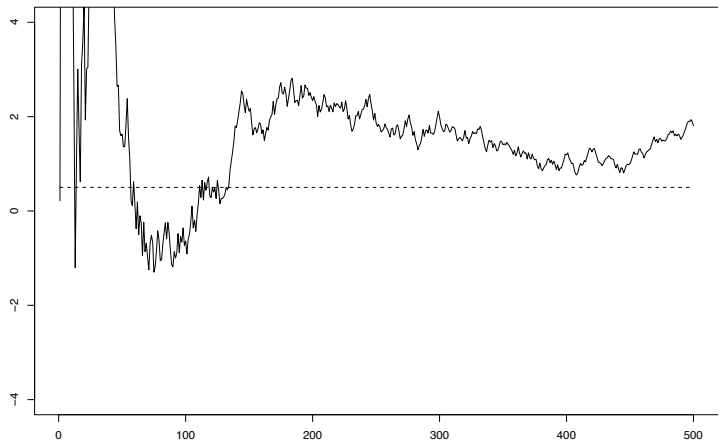


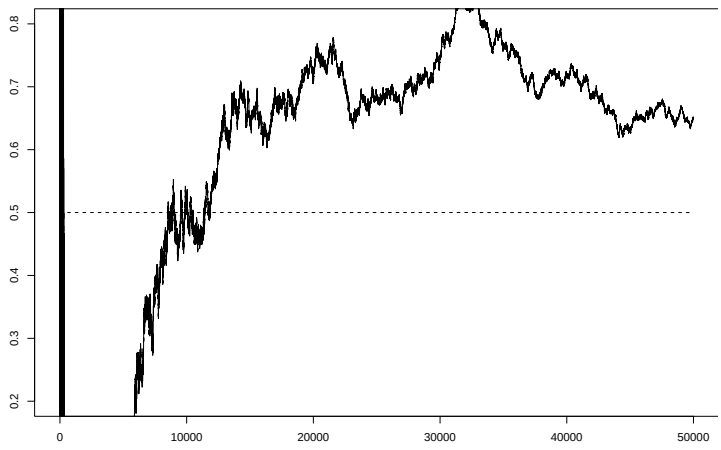


Globalement, on observe que, dans chaque expérience, les courbes obtenues se rapprochent de $1/2$ lorsque i croît, en présentant des fluctuations aléatoires d'amplitude de plus en plus faible autour de cette valeur.

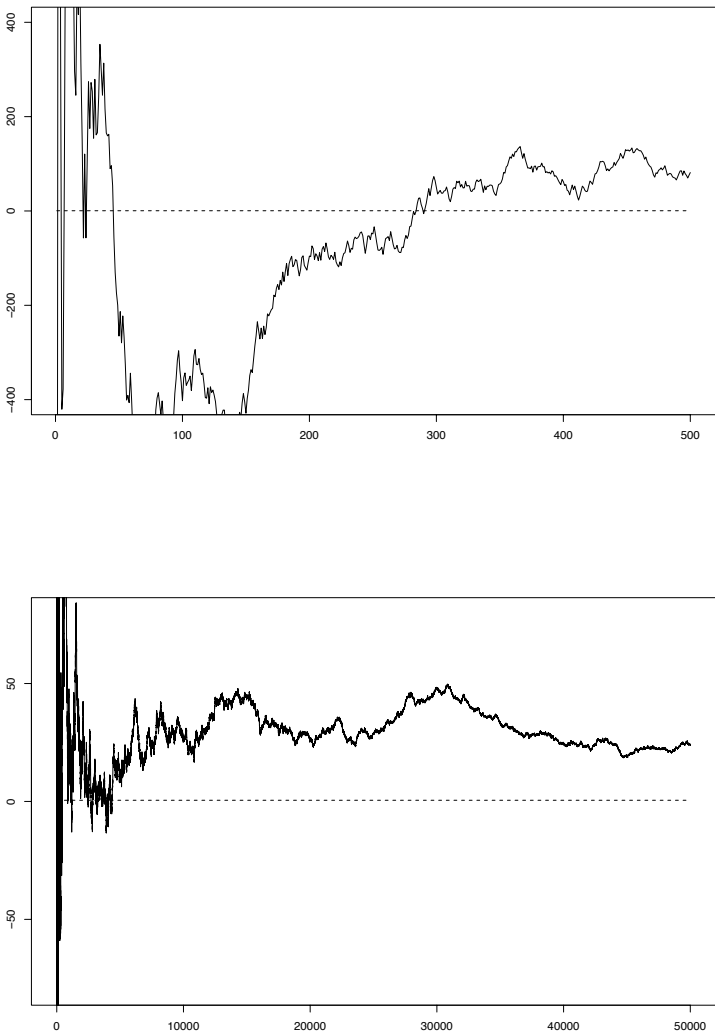
Reprenons l'expérience, mais avec des variables aléatoires de loi uniforme sur $[-49,5; 50,5]$,

Voici le tracé de $\frac{1}{i}(X_1 + \cdots + X_i)$ en fonction de i pour $1 \leq i \leq N$, avec $N = 500$ et $N = 50000$ (les échelles sont tronquées verticalement).





En reprenant les mêmes expériences, avec des variables aléatoires uniformes sur $[-4999,5; 5000,5]$, on obtient les graphiques suivants (tracé de $\frac{1}{i}(X_1 + \cdots + X_i)$ en fonction de i pour $1 \leq i \leq N$, pour $N = 500$ puis $N = 50000$, avec échelles sont tronquées verticalement).



On observe le même phénomène que dans les simulations précédentes, à savoir le ralentissement de la convergence à mesure que l'amplitude des fluctuations de X augmente.

3.2.7 Robustesse

Nous avons énoncé la loi (faible) des grands nombres dans le contexte d'une répétition indépendante de modèles probabilistes, donnant lieu à des variables aléatoires $X_1, \ldots, X_N$ mutuellement indépendantes, de même loi, et pour lesquelles l'espérance est définie. Il est naturel de s'interroger sur la robustesse de la loi des grands nombres

vis-à-vis de ce cadre particulier. Que se passe-t-il lorsque l'on considère des variables aléatoires qui présentent entre elles une certaine dépendance, ne sont plus exactement distribuées de la même façon, ou pour lesquelles l'espérance n'est pas définie ?

De manière générale, il existe un très grand nombre de résultats dont la formulation s'apparente à celle de la loi des grands nombres que nous avons présentée, et qui étendent celle-ci dans diverses directions. Plutôt qu'un résultat unique, le terme de «loi des grands nombres» désigne donc un vaste ensemble de résultats qui diffèrent par la nature exacte de leurs hypothèses et la forme précise de leurs conclusions. Tous ont en commun le fait d'énoncer que la somme d'un grand nombre de variables aléatoires, sous certaines hypothèses qui caractérisent la dépendance existant entre celles-ci, ainsi que l'ordre de grandeur des valeurs que ces variables peuvent prendre, conduit, après une normalisation adéquate (en général le nombre de variables présentes dans la somme), à une valeur essentiellement constante et déterministe (non-aléatoire). Dans l'énoncé que nous avons donné précédemment, la dépendance entre les variables est caractérisée par le fait que celles-ci sont indépendantes, et l'hypothèse concernant l'ordre de grandeur des valeurs prises est que celles-ci possèdent toutes la même loi, dont l'espérance est définie.

Dans la discussion qui suit, nous tenterons simplement d'illustrer sur quelques exemples – principalement par simulation –, la robustesse, ou, au contraire, la non-robustesse, de la loi des grands nombres, vis-à-vis de certaines altérations du contexte simple dans lequel nous l'avons énoncée.

3.2.8 L'hypothèse de répétition indépendante

La loi des grands nombres continue de s'appliquer lorsque les variables aléatoires $X_1, \ldots, X_N$ que l'on étudie sont produites au cours d'une succession d'expériences qui ne sont ni exactement indépendantes, ni décrites individuellement par des modèles exactement semblables, mais satisfont cependant ces hypothèses de manière approchée. Lorsque l'on s'écarte trop de ces hypothèses en revanche, la loi des grands nombres cesse en général d'être valable.

Donner une formulation mathématique précise de ce que peut être une succession approximativement indépendante d'expériences approximativement semblables, et plus encore de prouver la loi des grands nombres dans ce contexte ou tenter de déterminer précisément la frontière à partir de laquelle la loi des grands nombres ne s'applique plus dépasse largement le cadre de ce cours.

Nous décrivons simplement dans ce qui suit trois situations dans lesquelles des suites de variables aléatoires possédant chacune exactement la même loi, mais présentant des degrés de dépendance variés, présentent ou non un comportement du type décrit par la loi des grands nombres.

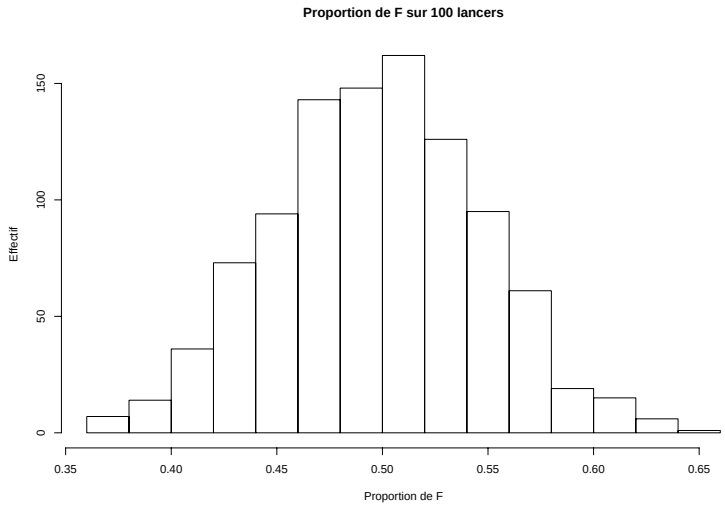
Une pièce normale

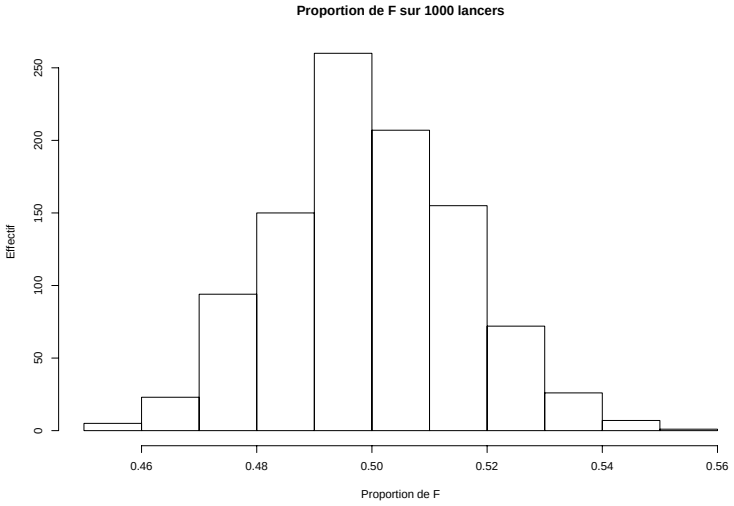
A titre de comparaison, commençons par une pièce dont les lancers successifs sont décrits par une suite de variables aléatoires indépendantes de loi de Bernoulli de paramètre $1/2$: $\mathbb{P}(X_i = \mathrm{P}) = \mathbb{P}(X_i = \mathrm{F}) = 1/2$. La proportion de F obtenue au cours des N premiers lancers peut s'écrire

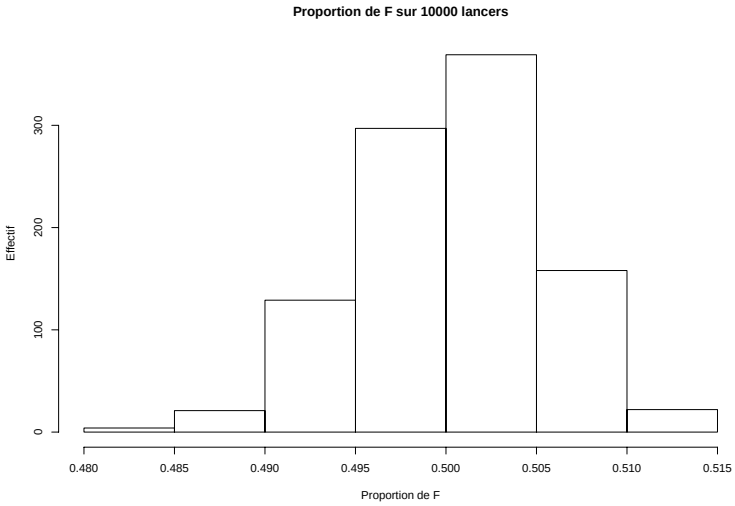
$$T_N = \frac{f(X_1) + \cdots + f(X_N)}{N},$$

en posant $f(\mathrm{F}) = 1$ et $f(\mathrm{P}) = 0$.

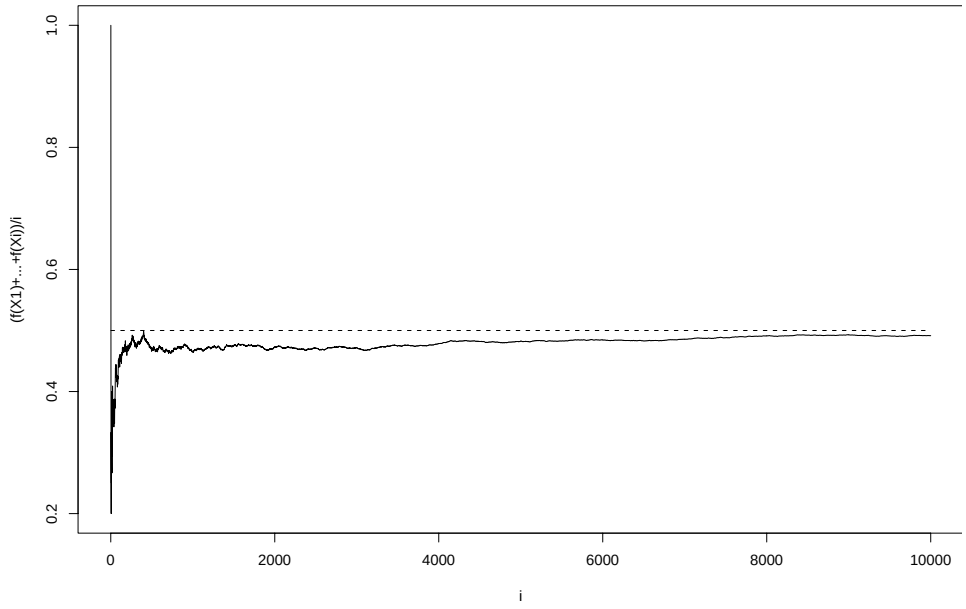
Voici d'abord les histogrammes obtenus en effectuant 1000 simulations de T_N, avec $N = 100, 1000, 10000$.

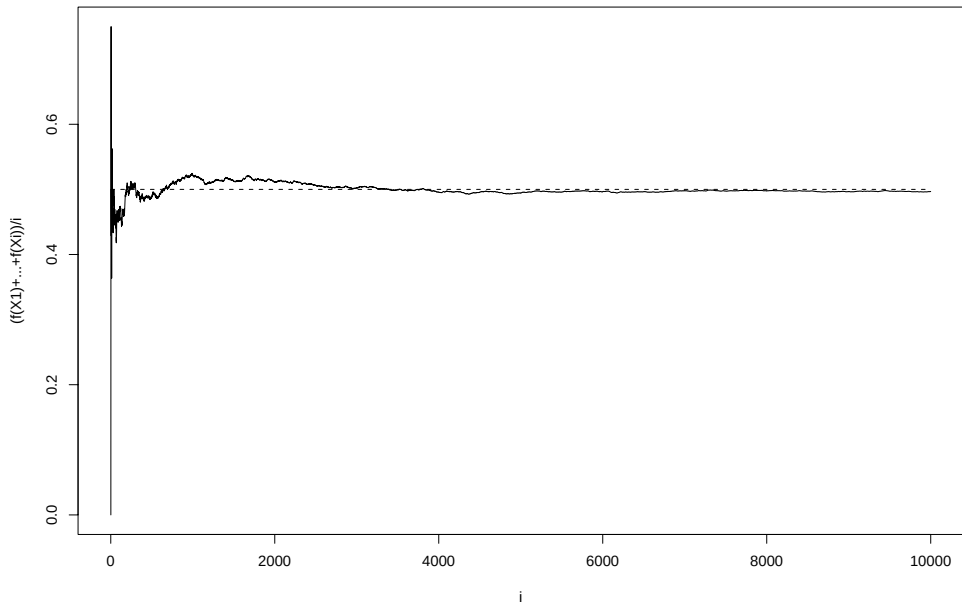


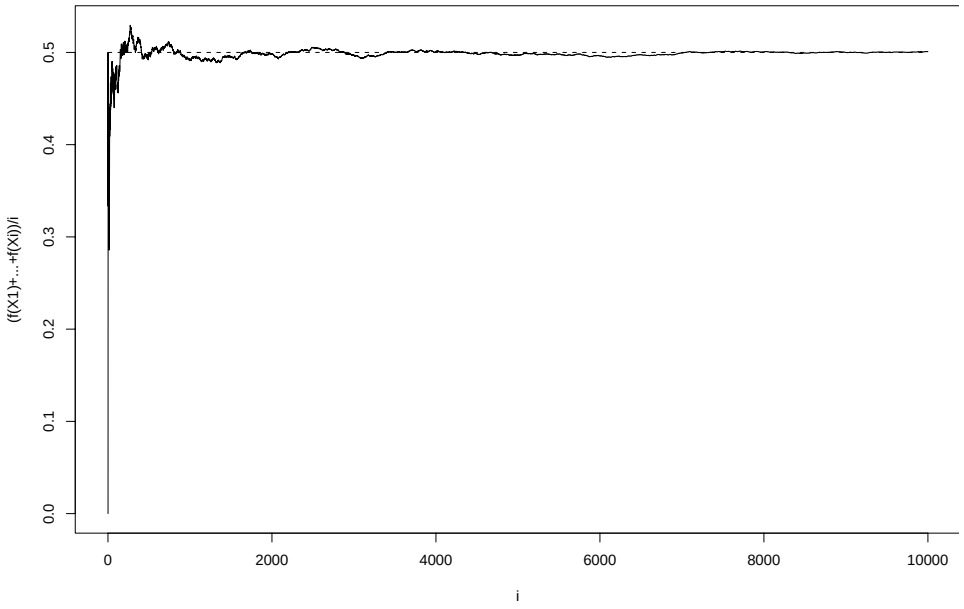


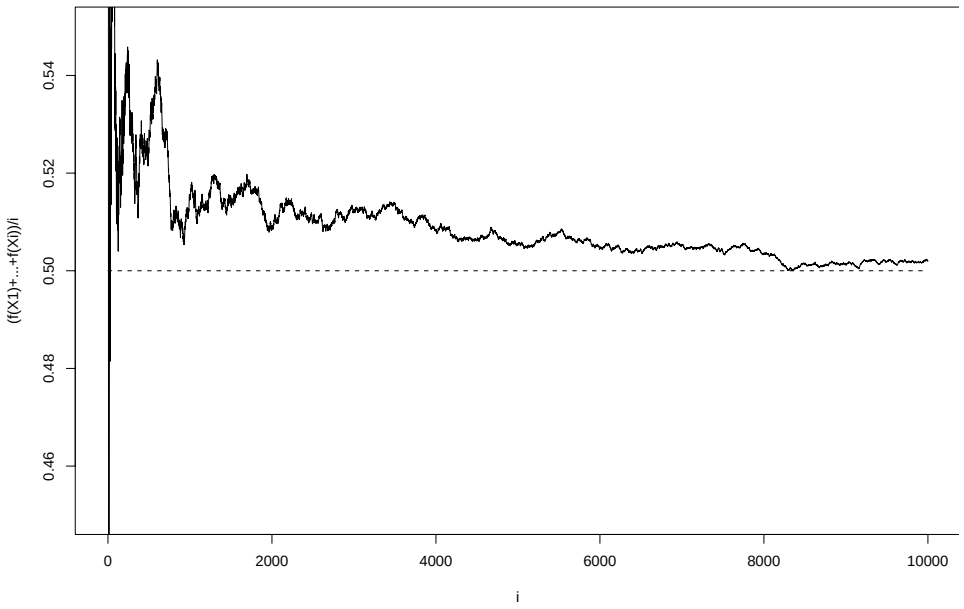


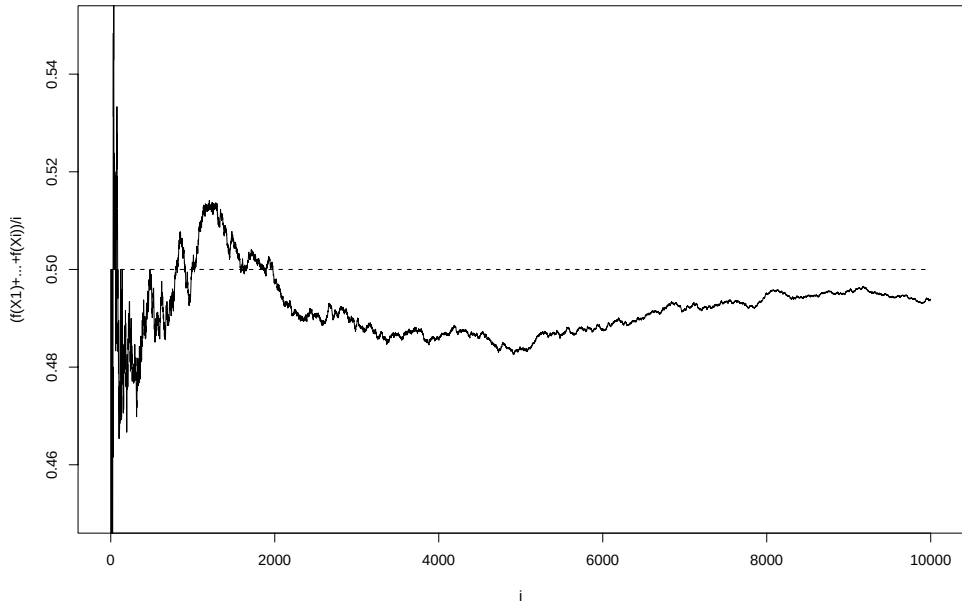
A présent, six exemples de simulation de tracés de $i \mapsto T_i$ (en resserrant l'échelle verticale pour les trois derniers, de façon à pouvoir distinguer précisément l'écart à la valeur limite 1/2).

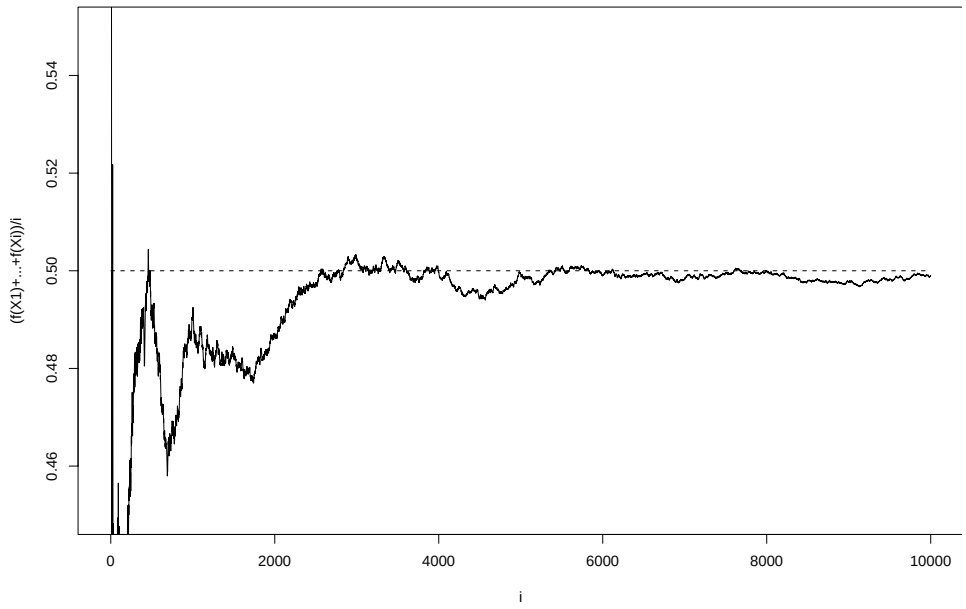

(f(X1)+...+f(Xi))/i
1.0
0.8
0.6
0.4
0.2
0
2000
4000
6000
8000
10000
i


(f(X1)+...+f(Xi))/i
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000
i


(f(X1)+...+f(Xi))/i
0.0
0.1
0.2
0.3
0.4
0.5
0
2000
4000
6000
8000
10000
i


(f(X1)+...+f(Xi))/i
0.46
0.48
0.50
0.52
0.54
0
2000
4000
6000
8000
10000
i


(f(X1)+...+f(Xi))/i
0.46
0.48
0.50
0.52
0.54
0
2000
4000
6000
8000
10000
i


(f(X1)+...+f(Xi))/i
0.46
0.48
0.50
0.52
0.54
0
2000
4000
6000
8000
10000
i

Une pièce de monnaie obstinée

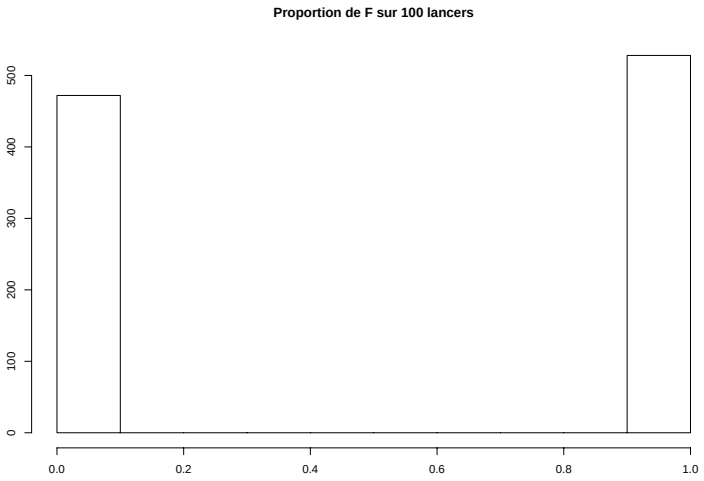
On suppose que l'on a affaire à une pièce de monnaie obstinée possédant la propriété suivante : une fois la pièce sortie de sa boîte, le premier lancer est effectivement aléatoire, pouvant donner pile ou face avec une probabilité égale à $1/2$, mais, au cours de tous les lancers suivants, la pièce se souvient du résultat de son premier lancer, et s'arrange toujours pour retomber exactement du même côté. Si l'on note $X_1, \ldots, X_N$ les résultats des N premiers lancers de la pièce, on se trouve ici dans un cas extrême de non-indépendance : la valeur de X_{i+1} est toujours égale à la valeur de X_i. En revanche, les lancers sont tous décrits individuellement par une loi de Bernoulli de paramètre $1/2$: $\mathbb{P}(X_i = \mathrm{P}) = \mathbb{P}(X_i = \mathrm{F}) = 1/2$. La proportion de F obtenue au cours des N premiers lancers peut s'écrire

$$T_N = \frac{f(X_1) + \cdots + f(X_N)}{N},$$

en posant $f(\mathrm{F}) = 1$ et $f(\mathrm{P}) = 0$.

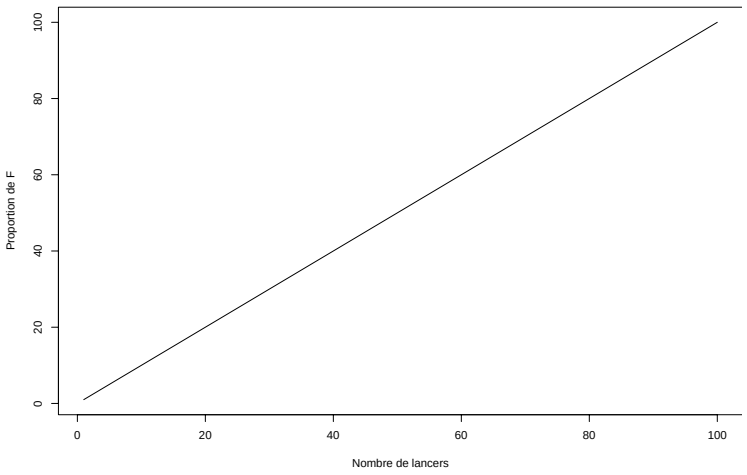
Bien entendu, la loi des grands nombres ne s'applique pas à T_N, puisque la suite des résultats obtenus est soit exclusivement constituée de P, soit exclusivement constituée de F.

Répétons un grand nombre de fois (mettons 1000) l'expérience consitant à sortir la pièce obstinée de sa boîte et à effectuer 100 lancers successifs. L'histogramme obtenu pour T_N est le suivant :

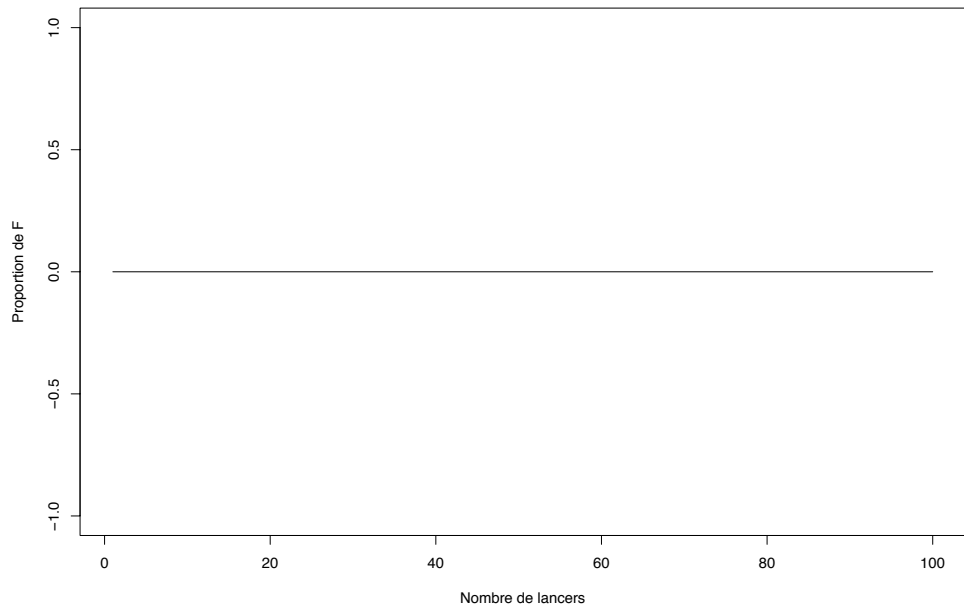


traduisant le fait que l'on obtient soit 0% soit 100% de P, avec une probabilité de $1/2$.

Si l'on trace l'évolution de T_i en fonction de i pour i variant de 1 à N, on obtient environ la moitié du temps le graphe suivant :



et l'autre moitié du temps le graphe suivant :



A comparer aux graphiques que l'on obtenait dans le cadre d'une répétition indépendante !

Une pièce moins obstinée

Considérons à présent une autre pièce obstinée, conservant également la mémoire de ses lancers passés, mais de manière moins stricte que la précédente. Spécifiquement, une fois la pièce sortie de sa boîte, le premier lancer effectué est aléatoire, donnant pile ou face avec une probabilité égale à $1/2$. Ensuite, pour tout $i \geq 1$, étant donnés les résultats des i premier lancers, le $i+1-$ème lancer se déroule de la façon suivante : la pièce reproduit le résultat du $i-$ème lancer avec une probabilité p fixée, et produit le résultat inverse avec une probabilité $1-p$. Si p est égal à 1, la pièce se comporte comme celle étudiée dans le paragraphe précédent. Si $p = 0$, on obtient une alternance stricte de P et de F. Nous supposerons dans la suite que $0 < p < 1$. Si $1/2 < p < 1$, la pièce conserve sa tendance à redonner lors d'un lancer la valeur obtenue à l'issue du lancer précédent, mais de manière moins stricte que dans le cas précédent. Si $p = 1/2$, on retrouve une suite de répétitions indépendantes de lancers de Bernoulli. Enfin, si $0 < p < 1/2$, la pièce a tendance à produire lors d'un lancer un résultat inversé par rapport au lancer précédent.

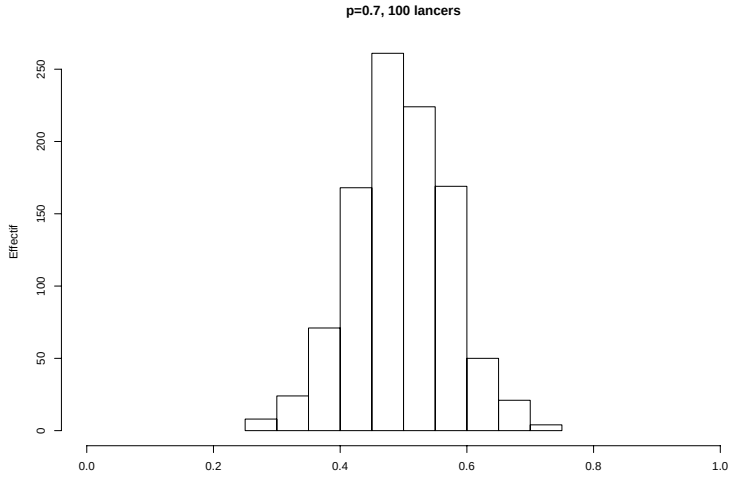
Les variables aléatoires $X_1, \ldots, X_N$ ne sont donc pas indépendantes lorsque $p \neq 1/2$, puisque le résultat obtenu au cours d'un lancer affecte la loi de probabilité attachée au lancer suivant. Cependant, il semble clair que, si k est suffisament grand, le résultat du lancer $i+k$ doit être approximativement indépendant du résultat du lancer i, car la mémoire du résultat du lancer i est de plus en plus brouillée au fur et à mesure que les lancers se répètent (voir à ce sujet l'exercice 65). Il existe donc une certaine forme d'indépendance approchée entre les résultats suffisament éloignés dans la séquence des lancers.

On peut par ailleurs facilement vérifier que, pris de manière individuelle, les lancers sont décrits par une loi de Bernoulli de paramètre $1/2$: $\mathbb{P}(X_i = \mathrm{P}) = \mathbb{P}(X_i = \mathrm{F}) = 1/2$.

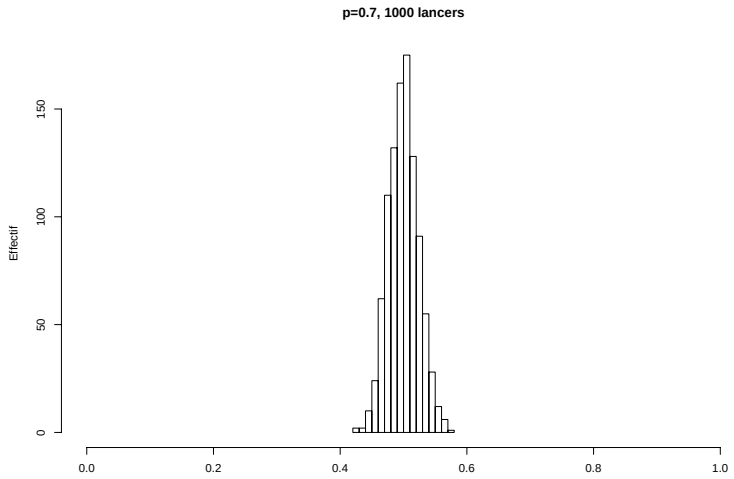
Il se trouve que, quelle que soit la valeur de $p \in]0,1[$ dans ce modèle, la loi des grands nombres est effectivement vérifiée par la proportion de P obtenue après N lancers, que nous notons T_N comme dans le paragraphe précédent.

Prenons par exemple $p = 0,7$.

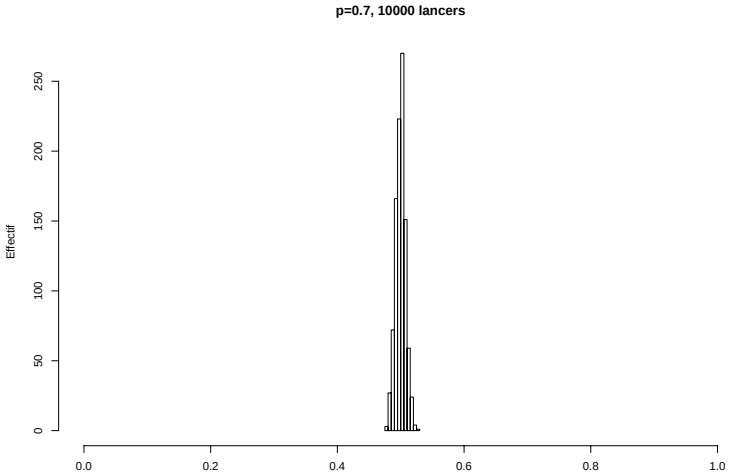
En effectuant 1000 simulations de 100 lancers, on obtient l'histogramme suivant pour la proportion de F.



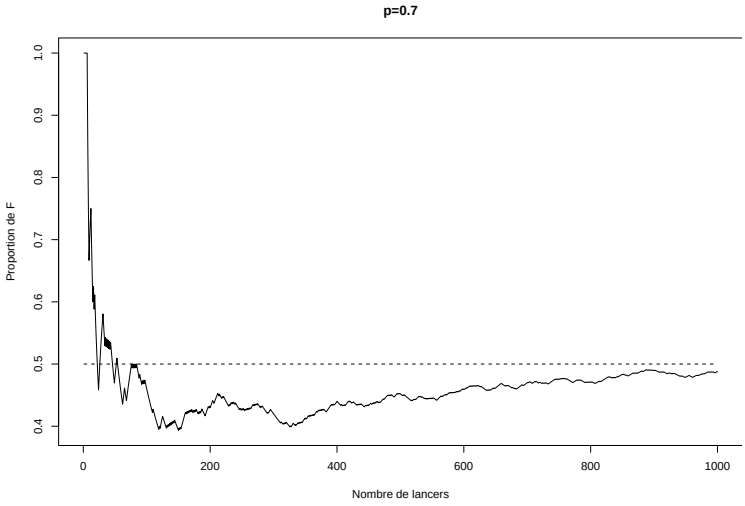
Voici maintenant l'histogramme obtenu en effectuant 1000 simulations de 1000 lancers.



Et enfin l'histogramme obtenu en effectuant 1000 simulations de 10000 lancers.



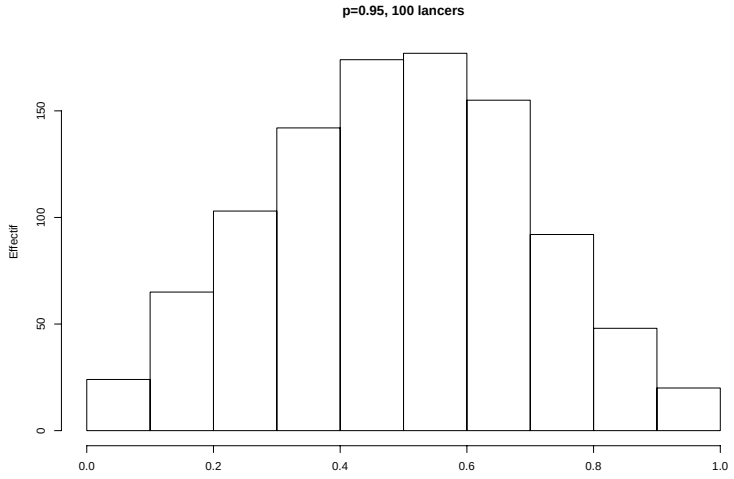
Un exemple de tracé de la proportion de F en fonction du nombre de lancers est donné par le graphique suivant.



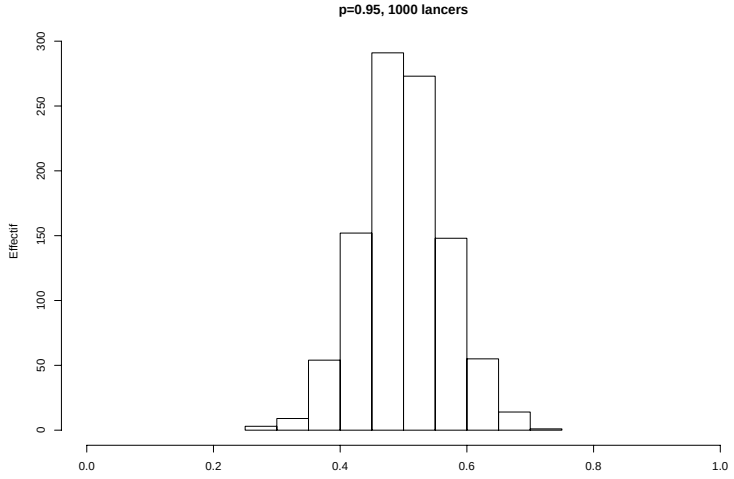
On constate bien sur ces graphiques un comportement de type loi des grands nombres, la proportion de pile se concentrant autour de la valeur 1/2 lorsque l'on effectue un grand nombre de lancers.

En prenant par exemple $p = 0,95$, on constate le même type de phénomène, mais avec une convergence plus lente à se manifester, conséquence de la plus forte similarité entre valeurs successives.

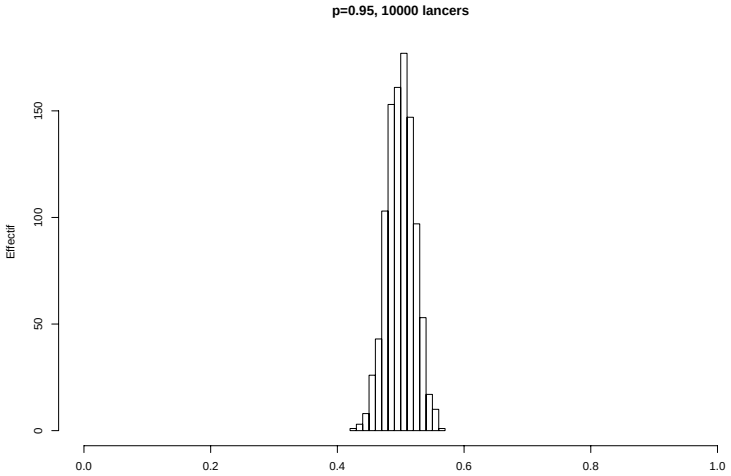
Voici l'histogramme obtenu pour la proportion de F en effectuant 1000 simulations de 100 lancers.



Avec 1000 simulations de 1000 lancers :



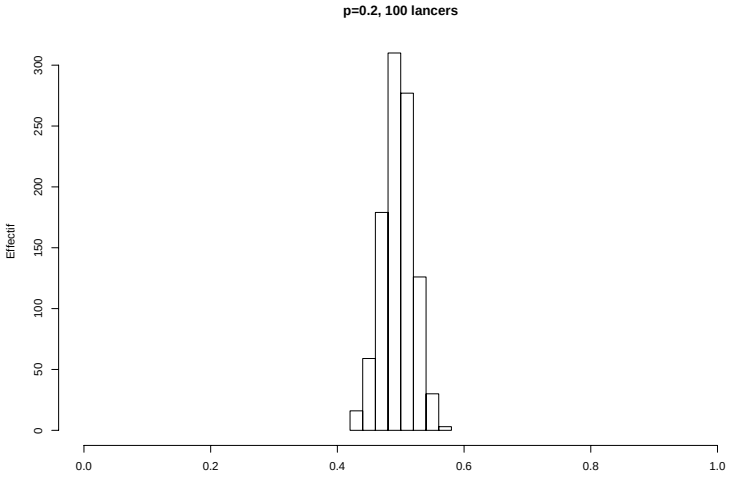
Avec 1000 simulations de 10000 lancers :



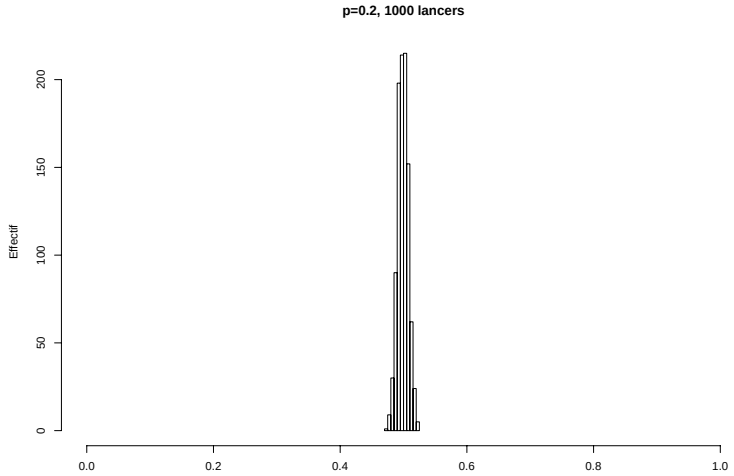
La plus forte dépendance entre valeurs successives se traduit donc ici par une convergence plus lente.

En prenant $p = 0,2$, on observe encore constate le même type de phénomène, mais avec une convergence qui se manifeste de manière plus rapide. En effet, les résultats des lancers successifs ont tendance à alterner plus souvent que dans le cas de lancers indépendants, ce qui stabilise plus rapidement autour de $1/2$ la proportion de F.

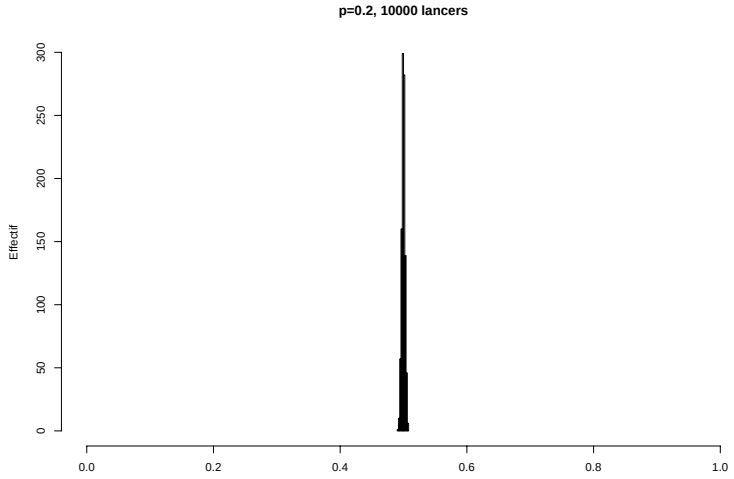
Voici l'histogramme obtenu avec 1000 simulations de 100 lancers.



Puis 1000 simulations de 1000 lancers :



Et enfin 1000 simulations de 10000 lancers :



En conclusion, dans cet exemple, la dépendance entre valeurs successives reste suffisament modérée pour que la loi des grands nombres demeure valable, la vitesse de convergence étant manifestement affectée par cette dépendance.

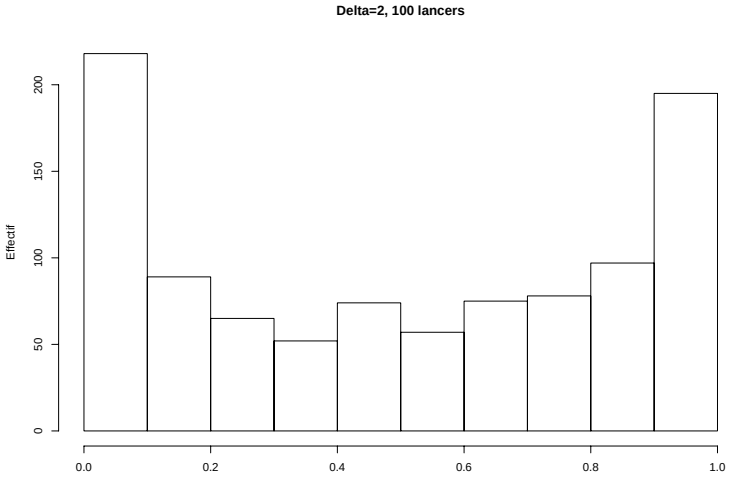
Encore une pièce obstinée

Considérons à présent une pièce dont les lancers successifs sont reliés entre eux de la manière suivante. Une fois la pièce sortie de sa boîte, le premier lancer effectué est aléatoire, donnant pile ou face avec une probabilité égale à 1/2. Ensuite, pour

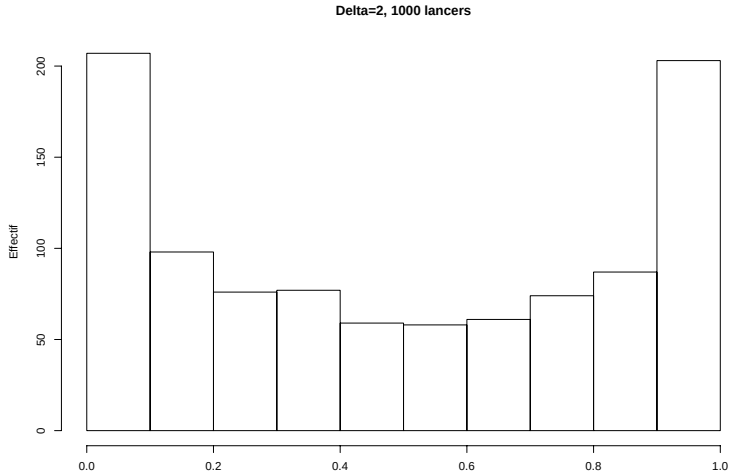
tout $i \geq 1$, étant donnés les résultats des i premier lancers, le $i+1-$ème lancer se déroule de la façon suivante : la pièce accorde à P une probabilité proportionnelle à $1 + \Delta N_i(P)$ et à F une probabilité proportionnelle à $1 + \Delta N_i(F)$, $N_i(P)$ et $N_i(F)$ désignant respectivement les nombres de fois où P et F sont sortis au cours des i premiers lancers, et $\Delta > 0$ désignant un paramètre. En d'autres termes, chaque nouveau lancer donnant lieu à un F renforce d'une valeur égale à Δ le poids accordé à F dans les futurs lancers, et il en va de même pour P. On peut vérifier facilement que, pris de manière individuelle, les lancers sont décrits par une loi de Bernoulli de paramètre $1/2$: $\mathbb{P}(X_i = \mathrm{P}) = \mathbb{P}(X_i = \mathrm{F}) = 1/2$.

Voici quelques exemples de simulations effectuées avec $\Delta = 2$.

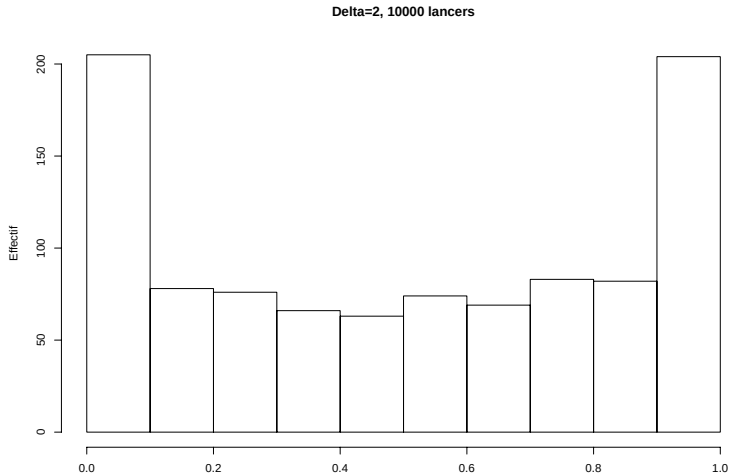
Avec 1000 simulations de 100 lancers, on obtient l'histogramme suivant pour la proportion de F.



Avec 1000 simulations de 1000 lancers :



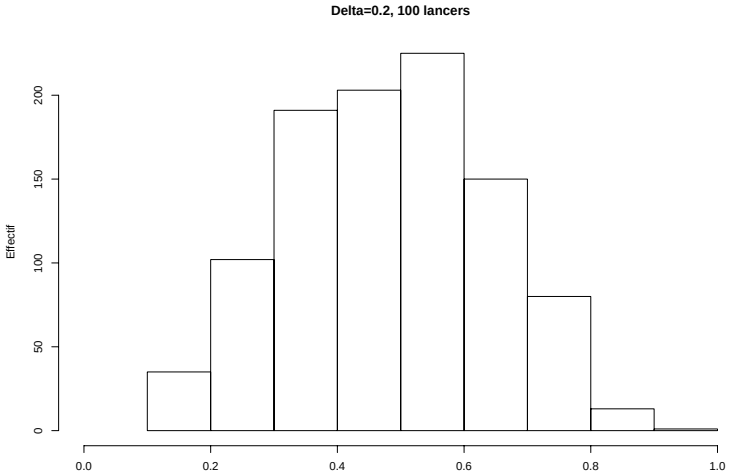
Avec 1000 simulations de 10000 lancers :



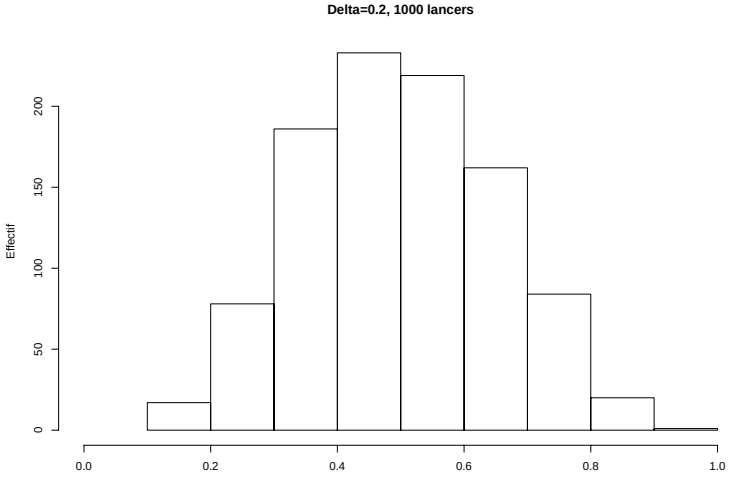
Manifestement, les proportions de F obtenues n'ont pas tendance à se concentrer autour d'une valeur fixée à mesure que N augmente, les histogrammes obtenus étant en gros identiques pour $N = 100, 1000, 10000$. La dépendance entre les lancers successifs met donc la loi des grands nombres en défaut, au moins d'après nos simulations.

Voici à présent des simulations effectuées avec $\Delta = 0, 2$, soit une dépendance plus faible des lancers vis-à-vis des résultats des lancers précédents.

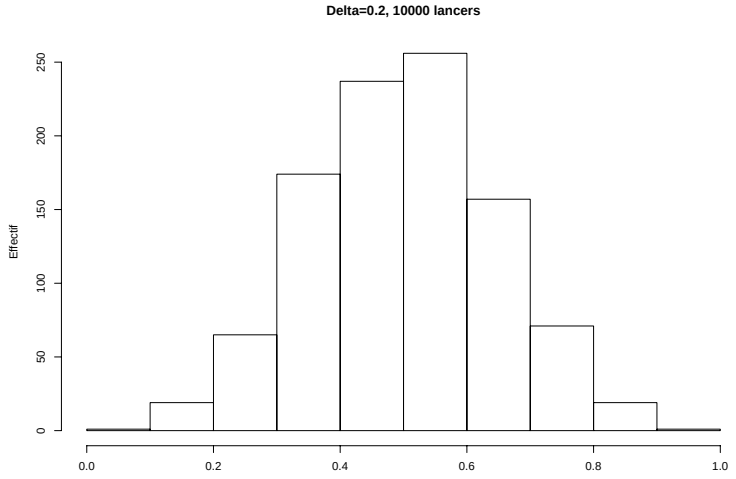
Avec 1000 simulations de 100 lancers :



Avec 1000 simulations de 1000 lancers :



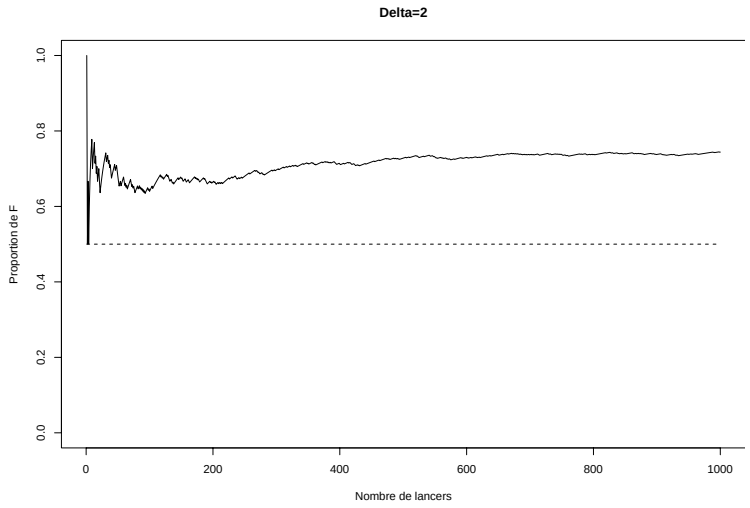
Avec 1000 simulations de 10000 lancers :

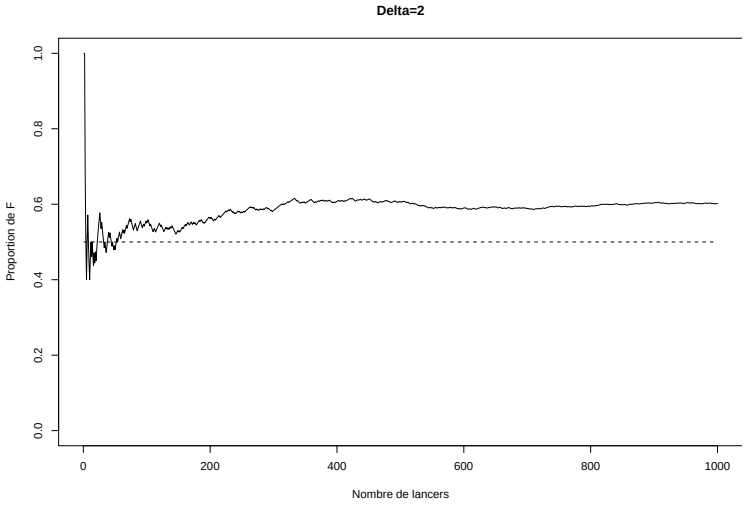


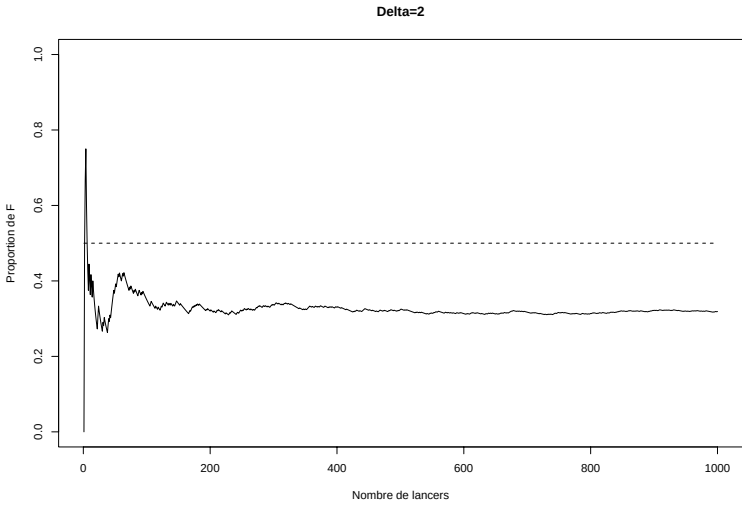
Les histogrammes obtenus sont plus resserrés autour de la valeur $1/2$ que dans le cas $\Delta = 2$, mais on n'observe, ici non plus, aucun resserrement lorsque la valeur de N croît.

Donnons à présent quelques représentations de T_i en fonction de i, dans l'esprit de la loi forte des grands nombres.

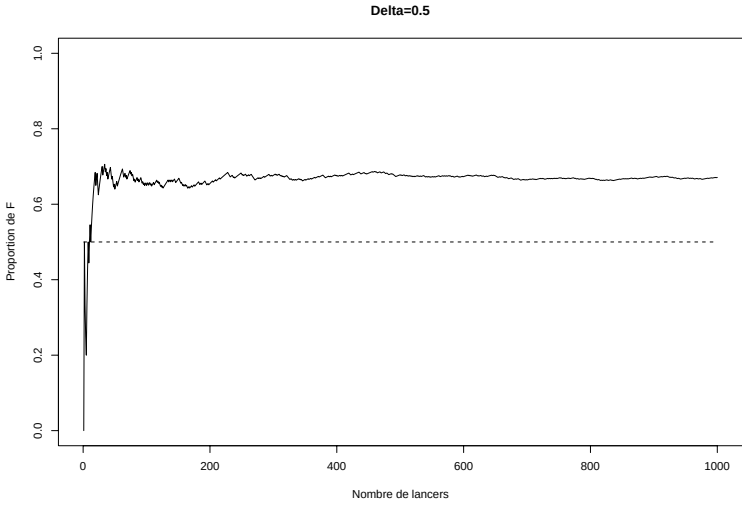
Voici trois courbes obtenues en simulant les lancers avec $\Delta = 2$.

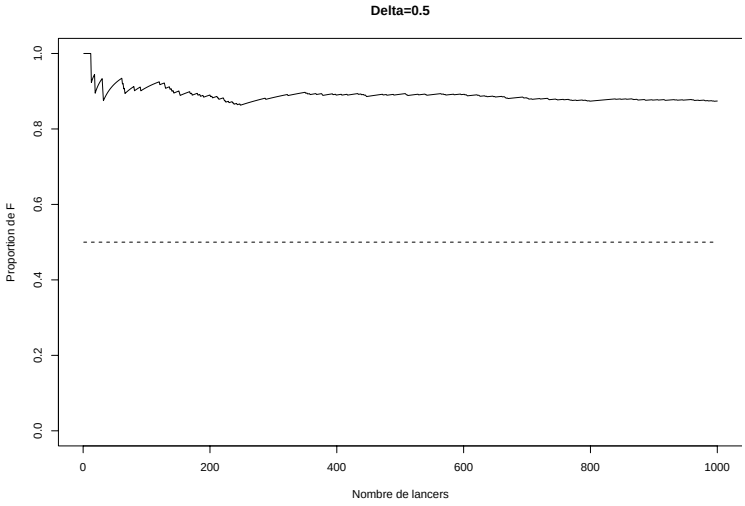


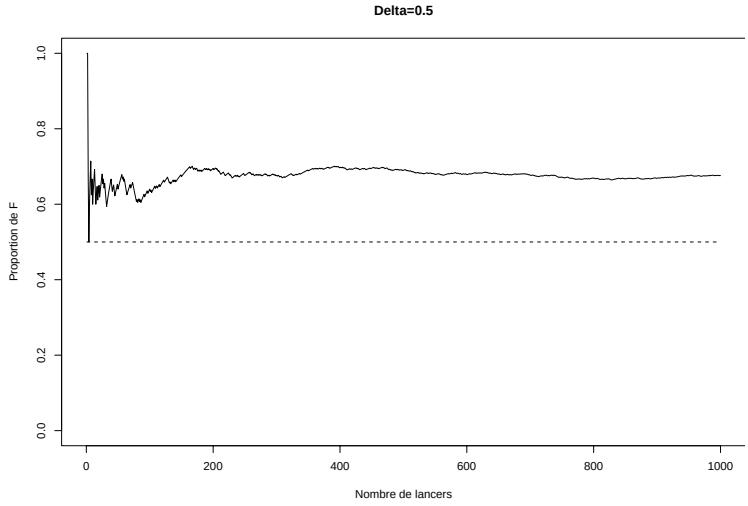




Voici trois courbes obtenues en simulant les lancers avec $\Delta = 0.5$.


Delta=0.5
Proportion de F
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Nombre de lancers


Delta=0.5
Proportion de F
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Nombre de lancers



De manière remarquable, ces graphiques suggèrent que la suite T_i converge effectivement lorsque i tend vers l'infini, mais que la limite est une variable aléatoire, ce qui ne correspond bien entendu pas au comportement décrit par la loi forte des grands nombres, qui énonce la convergence vers une valeur déterministe (non-aléatoire).

Nous vous suggérons de réexaminer ces résultats à la lueur de l'exercice 134, afin d'obtenir une compréhension théorique des phénomènes illustrés ici par simulation.

3.2.9 L'existence de l'espérance

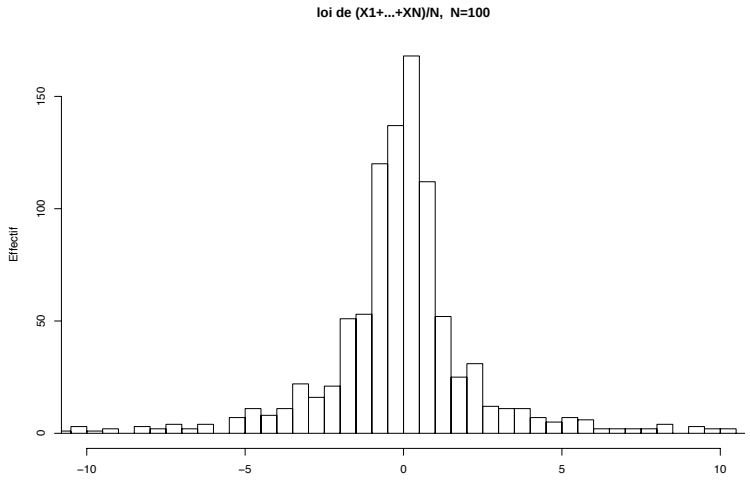
Dans les trois exemples précédents, nous avons considéré des sommes de variables aléatoires, certes dépendantes entre elles, mais ne pouvant prendre que les valeurs 0 et 1, et en fait toutes de loi de Bernoulli de paramètre $1/2$, ce qui assurait bien entendu l'existence de l'espérance.

Posons-nous à présent la question, dans le cas de répétitions indépendantes d'une variable aléatoire, de la robustesse de la loi des grands nombres vis-à-vis de l'existence de l'espérance. Un exemple classique de loi pour laquelle l'espérance n'est pas définie est la loi de Cauchy (voir le chapitre «Variables aléatoires»).

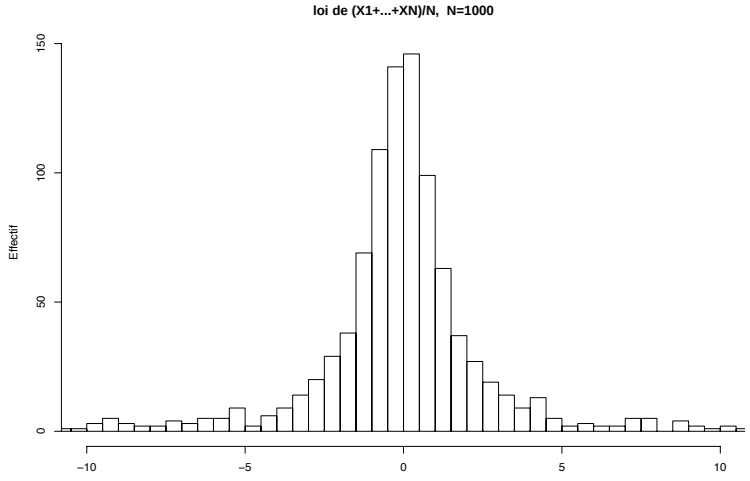
La densité de la loi étant symétrique par rapport à l'axe d'équation $x = 0$, on pourrait cependant s'attendre à ce que les valeurs positives et négatives prises par les X_i se compensent de manière à ce que les valeurs de $\frac{1}{N}(X_1 + \cdots + X_N)$ soient concentrées autour de la valeur 0. Les simulations suivantes illustrent le fait que ce n'est pas du tout ainsi que les choses se passent.

Voici les histogrammes (tronqués sur l'échelle horizontale) obtenus par simulation pour la loi de $\frac{1}{N}(X_1 + \cdots + X_N)$, les X_i étant des variables aléatoires mutuellement indépendantes de loi de Cauchy de paramètre $s = 1$ et $\ell = 0$.

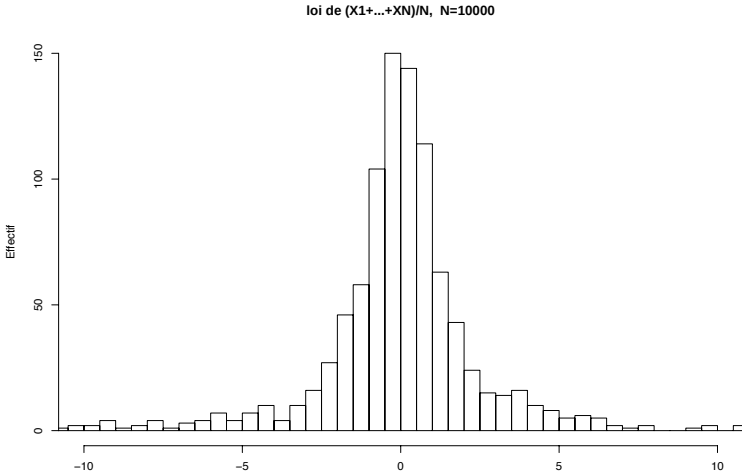
Avec 1000 simulations et $N = 100$.



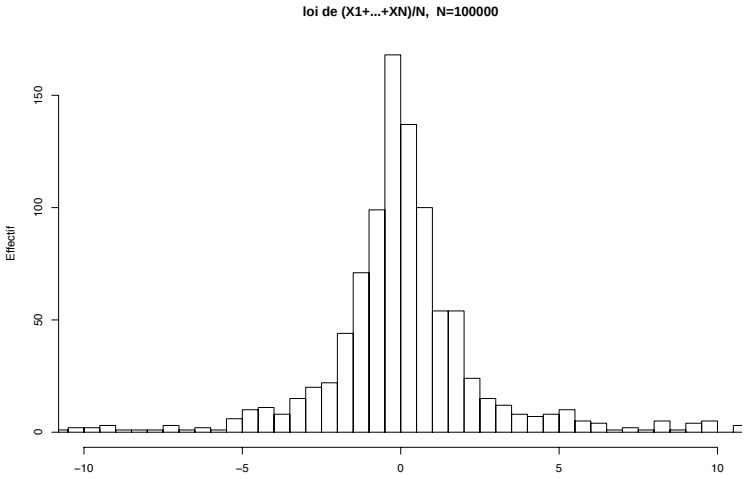
Avec 1000 simulations et $N = 1000$,



Avec 1000 simulations et $N = 10000$.

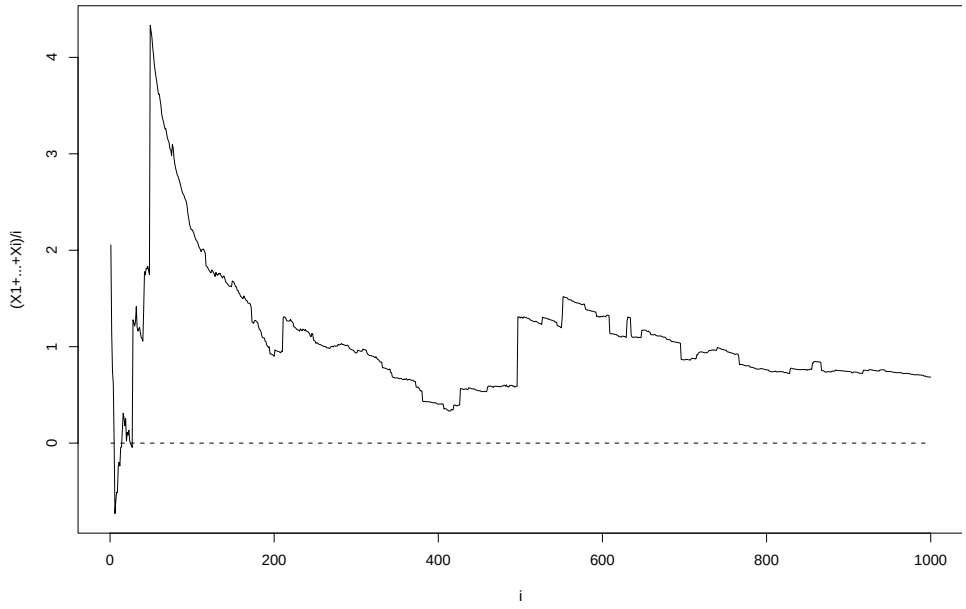


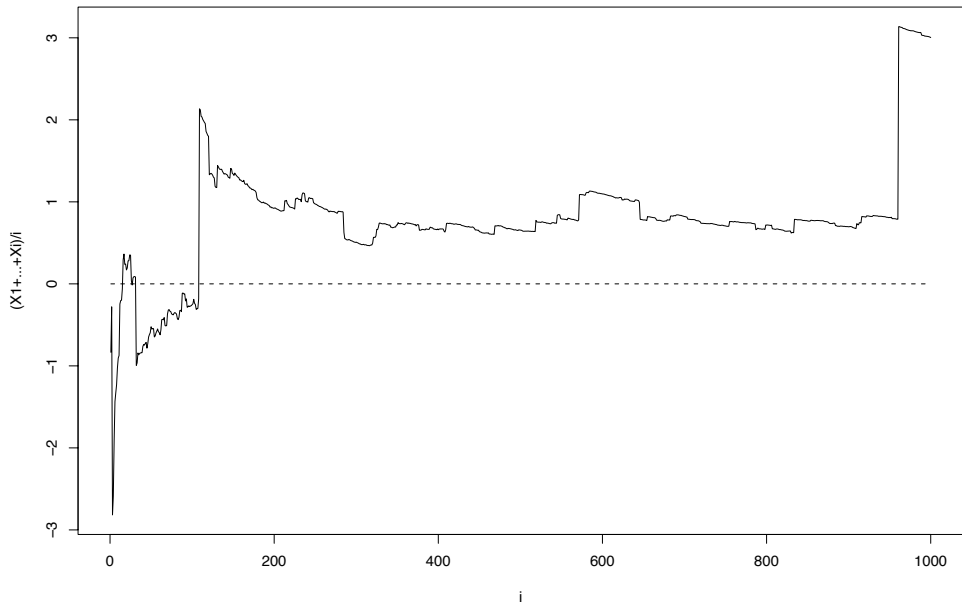
Avec 1000 simulations et $N = 100000$.

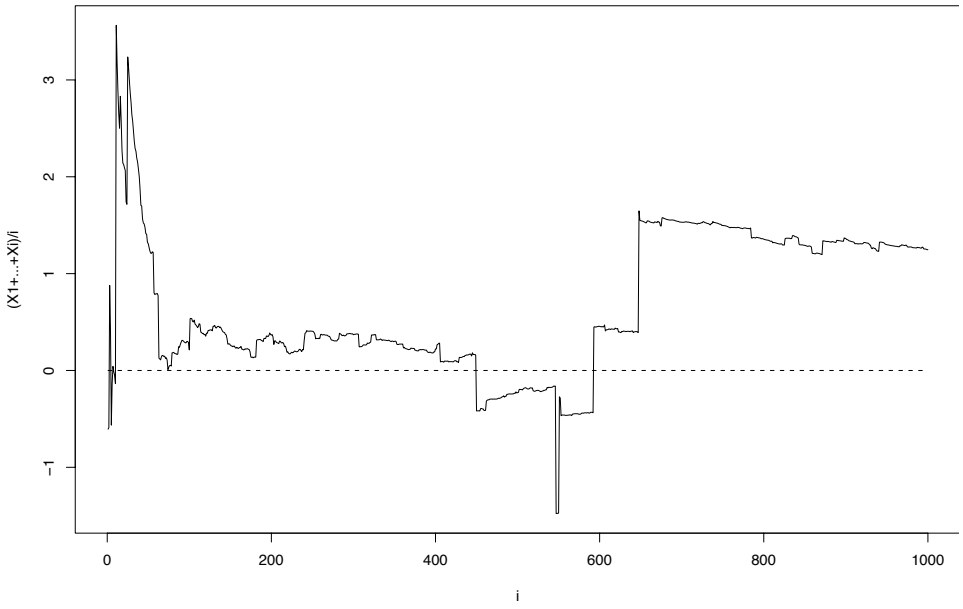


Certes, les histogrammes sont obtenus ont en gros la forme de pics symétriques par rapport à l'axe $x = 0$, mais l'on ne constate aucun phénomène de concentration de la loi autour d'une valeur fixée à mesure que N croît.

Si l'on trace l'évolution de $\frac{1}{i}(X_1 + \cdots + X_i)$ en fonction de i pour i variant de 1 à N, on obtient par exemple les courbes suivantes, avec $N = 1000$.


(X1+...+Xi)/i
0
1
2
3
4
0
200
400
600
800
1000
i


(X1+...+Xi)/i
−3
−2
−1
0
1
2
3
0
200
400
600
800
1000
i



Ces quelques exemples sont destinés à illustrer que la loi des grands nombres ne s'applique plus lorsque l'espérance n'est pas définie.

On notera la différence par rapport au comportement observé dans le troisième exemple de pièce obstinée : même si, dans le cas présent les histogrammes restent à peu près identiques lorsque N croît, les courbes $\frac{1}{i}(X_1 + \cdots + X_i)$ en fonction de i ne montrent, en revanche, aucun signe de convergence vers une valeur définie, fût-elle aléatoire. On notera également sur les trois derniers tracés le fait que quelques «sauts» de la courbe suffisent à apporter une contribution importante à sa position finale : quelques valeurs de X_i sont suffisament importantes pour chambouler la valeur moyenne obtenue sur un jeu pourtant important d'observations !

Il est naturel de se demander ce que signifie concrètement le fait qu'une quantité soit modélisée par une variable aléatoire dont l'espérance n'est pas définie : dans la plupart des situations réelles, les quantités que l'on considère sont en réalité bornées, même si les bornes correspondantes peuvent prendre des valeurs extrêmement élevées, si bien que l'espérance devrait toujours être définie. Disons, sans donner beaucoup plus de précision, qu'en pratique, l'amplitude des fluctuations des variables aléatoires que l'on ajoute peut être si importante que l'on ne peut pas s'attendre à observer un comportement du type décrit par la loi des grands nombres lorsque l'on considère des sommes d'un nombre raisonnable de telles variables aléatoires. Cette question (à partir de quelle valeur de N peut-on, dans un contexte donné, considérer que la loi

des grands nombres fournit une approximation raisonnable du comportement réel) est par exemple abordée dans le paragraphe sur les inégalités de concentration et le chapitre sur le théorème de la limite centrale du chapitre suivant. Elle rejoint la discussion du paragraphe «Qu'est-ce qu'un grand nombre ?».

3.2.10 Position de la loi des grands nombres

La loi des grands nombres - celle que nous avons énoncée, comme ses diverses extensions et variantes, donne souvent lieu à des interprétations ou des commentaires, parfois erronés, et souvent discutables, et ce, au moins depuis la parution de la première preuve formelle de ce résultat en 1713 dans l'*Ars Conjectandi* de Jacques Bernoulli.

Afin de préciser les choses, voici un bref commentaire visant à situer ce résultat, sa portée, et les interprétations que l'on peut en donner.

Nous nous limiterons, pour simplifier, à discuter le cas la loi des grands nombres pour les indicatrices d'événements, qui est presque le cas général, (voir l'exercice 91).

Soulignons tout d'abord que, dans le cadre de ce cours, la loi des grands nombres est un théorème de mathématiques, énonçant une propriété prouvée de certains objets mathématiques, à savoir les répétitions indépendantes de modèles probabilistes. Nous avons donné une preuve mathématique de ce résultat (moyennant l'hypothèse simplificatrice que les variables aléatoires considérées ont une variance définie, qui est en particulier toujours vérifiée dans le cas de fonctions indicatrices).

L'application de la loi des grands nombres à une situation réelle suppose donc que l'on puisse, de manière valable, représenter cette situation au moyen d'un modèle mathématique auquel la loi des grands nombres s'applique. A son tour, la question de la validité d'un tel modèle, et donc des conséquences qu'il est possible d'en tirer, n'a de sens qu'une fois précisée la traduction concrète de la notion de probabilité dans le contexte étudié.

Commençons par nous placer dans une interprétation rigoureusement fréquentielle de la probabilité. Pour interpréter ainsi la probabilité $\mathbb{P}^{\otimes N}$ sur Ω^N, on doit considérer un grand nombre de répétitions de la série des N expériences décrites par Ω^N. Insistons : **chaque répétition** dont il est question dans l'interprétation fréquentielle de la probabilité sur Ω^N **est constituée par** N **répétitions** de l'expérience décrite par Ω. La probabilité $\mathbb{P}^{\otimes N}$ sur Ω^N est alors censée décrire, dans la limite d'un grand nombre de répétitions des N expériences répétées décrites par Ω^N, la fréquence avec laquelle les différents éléments $(\omega_1, \ldots, \omega_N) \in \Omega^N$ surviennent.

Bien entendu, la façon dont sont répétées les séquences de N expériences doit être décrite précisément, et il faut disposer d'assurances concernant le fait que les fréquences se stabilisent effectivement autour d'une valeur limite lorsque l'on effectue des répétitions de cette façon. Ce qu'affirme la loi des grands nombres dans ce

contexte est que, si les fréquences limites avec lesquelles apparaissent les éléments de Ω^N sont données par $\mathbb{P}^{\otimes N}$, la fréquence limite avec laquelle on observe des séquences de N expériences vérifiant $\left|\frac{1}{N}(X_1 + \cdots + X_N) - \mathbb{E}(X)\right| \geq \epsilon$ tend vers zéro lorsque N tend vers l'infini, pour tout $\epsilon > 0$ fixé (ceci pouvant éventuellement être quantifié au moyen d'inégalités de déviation).

En pratique cependant, c'est rarement ainsi que l'on applique la loi des grands nombres : on considère généralement une seule répétition de N expériences, et l'on considère comme plausible le fait que

$$\left|\frac{1}{N}(X_1 + \cdots + X_N) - \mathbb{E}(X)\right| < \epsilon$$

si

$$\mathbb{P}^{\otimes N}\left(\left|\frac{1}{N}(X_1 + \cdots + X_N) - \mathbb{E}(X)\right| \geq \epsilon\right)$$

est suffisament petit, c'est-à-dire lorsque N est suffisament grand (ceci pouvant éventuellement être quantifié au moyen d'inégalités de déviation). Autrement dit, une telle application de la loi des grands nombres suppose que l'on interprète les fréquences de long terme (relatives à de nombreuses répétitions de séquences de N expériences) comme des mesures de plausibilités individuelles attachées aux différents tirages d'une telle séquence. Soulignons que le caractère rationnel et la pertinence de cette interprétation ne sont pas forcément assurés.

A ce stade, l'argument que nous venons d'exposer peut sembler sans objet. En effet, nous sommes partis de l'hypothèse selon laquelle, lorsque l'on répète un grand nombre de fois (dans des conditions contrôlées) une série de répétitions de N expériences, décrite par Ω^N, la fréquence de long terme des éléments de Ω^N est décrite par $\mathbb{P}^{\otimes N}$. Mais ceci entraîne automatiquement que, lors de la répétition (dans les mêmes conditions contrôlées) d'un grand nombre d'expériences décrites par Ω (et non plus d'une série de N telles expériences), la fréquence de long terme avec laquelle un élément $\omega \in \Omega$ apparaît doit être donnée par $\mathbb{P}(\omega)$. Nul besoin de loi des grands nombres pour parvenir à ce résultat : c'est une simple conséquence de notre hypothèse concernant l'interprétation fréquentielle de la probabilité. Qu'apporte alors la loi des grands nombres ? Selon nous, une réponse possible est que **la loi des grands nombres fait entrer l'étude des séries de N expériences dans le cadre de la modélisation probabiliste**, et peut ainsi non seulement confirmer le fait que la répétition de N expériences doit conduire, lorsque N est grand, à des fréquences limites données par $\mathbb{P}$, mais fournir des mesures quantitatives de la plausibilité d'un écart donné par rapport à une telle fréquence limite, par exemple au moyen d'inégalités de déviation.

Dans l'interprétation de la probabilité comme mesure de plausibilité attachée aux événements, la loi des grands nombres stipule simplement que, si l'on considère

une répétition de situations que l'on envisage comme indépendantes vis-à-vis des plausibilités qui en caractérisent les issues, chaque situation étant décrite par la même affectation de plausibilité aux différentes issues, on doit considérer comme fortement plausible, lorsque l'on effectue un grand nombre de répétitions, que la fréquence avec laquelle un événement se produit soit voisine de la plausibilité qu'on lui attribue. La portée de ce résultat dépend naturellement de la pertinence des affectations des plausibilités aux différentes issues.

Voici pour finir une petite liste commentée d'idées, vraies ou fausses, au sujet de la loi des grands nombres.

- «La loi des grands nombres est un théorème de mathématiques.» C'est vrai. Telle que nous l'avons énoncée, la loi des grands nombres est une propriété de certains objets mathématiques, les répétitions indépendantes de modèles probabilistes, et nous en avons donné une preuve (moyennant l'hypothèse simplificatrice que les variables aléatoires considérées ont une variance définie).
- «La loi des grands nombres est une loi de la Nature affirmant que, lors d'expériences répétées, la fréquence avec laquelle un événement se produit tend vers une valeur limite.» C'est faux dans le contexte qui est le nôtre ici : la loi des grands nombres est un résultat mathématique portant sur des modèles mathématiques de situations réelles, et non pas une loi au sens d'une loi de la Nature. L'application à une situation réelle de la loi des grands nombres que nous avons prouvée suppose que le modèle mathématique dont elle est déduite donne une description correcte de la situation considérée. Ceci suppose une interprétation concrète de la notion de probabilité, qui, la plupart du temps, contient déjà le fait que les fréquences limites se stabilisent, et n'a donc pas de rapport direct avec la loi des grands nombres que nous avons prouvée, et doit être établie sur d'autres bases. Cependant, on utilise parfois le terme de «loi des grands nombres» pour désigner cette propriété de stabilisation des fréquences. Rappelons que cette propriété de stabilité des fréquences lors d'un grand nombre de répétitions n'est en aucun cas une loi générale, et dépend du contexte et de la manière dont sont répétées les expériences.
- «La loi des grands nombres est un théorème qui prouve que, lors d'expériences répétées, la fréquence avec laquelle un événement se produit tend vers une valeur limite.» D'après ce que nous avons dit auparavant, certainement pas. Tout dépend de la validité du modèle dont est déduite la loi des grands nombres, et la validité de ce modèle suppose en général déjà que les fréquences se stabilisent autour d'une valeur limite.
- «La loi des grands nombres est une évidence.» Non, ou alors peut-être pour vous seul, car il a fallu les efforts de nombreux mathématiciens pour en apporter des preuves générales satisfaisantes. Considérer ce résultat comme évident peut résulter d'une confusion entre le contenu réel de la loi des grands nombres (un

théorème mathématique) et l'expérience concrète ou tout au moins l'intuition selon laquelle, par exemple, la fréquence d'apparition de pile et face doit se stabiliser au cours d'un grand nombre de lancers.

– La loi des grands nombres permet de donner une définition rigoureuse de la probabilité comme limite de la fréquence au cours d'un grand nombre d'expériences répétées. Eh, non ! La loi des grands nombres suppose définie la notion de probabilité (et ce qui l'accompagne : indépendance, variable aléatoire, etc...), et, à partir d'hypothèses formulées en termes de probabilités (répétitions indépendantes), prouve un résultat lui-même formulé en termes de probabilités. On ne peut pas définir la probabilité en supposant déjà connue la notion de probabilité !
– il est évident que si on lance une pièce de monnaie équilibrée de manière répétée la fréquence observée de face (et de pile) doit être voisine de $1/2$. Non (ou alors peut-être pour vous seulement). Avez-vous réalisé vous-même ce type d'expériences ? Que savez-vous de la physique du lancer d'une pièce de monnaie ? Vous pouvez consulter à ce sujet l'article de Diaconis et al. cité dans la bibliographie.
– Il est absurde de vouloir prouver la loi des grands nombres à partir du formalisme de la théorie des probabilités, alors que l'on s'est déjà appuyé, pour justifier ce formalisme, sur une définition de la probabilité comme limite de la fréquence au cours d'un grand nombre d'expériences répétées. C'est faux. Tout d'abord, le formalisme de la théorie des probabilités est également justifié par l'interprétation en termes de plausibilité, qui ne fait pas référence à la notion de fréquence, et dans laquelle la loi des grands nombres a parfaitement sa place. D'autre part, dans le cadre de l'interprétation en termes de fréquence, la loi des grands nombres a un sens bien précis, qui n'est pas redondant avec l'hypothèse de stabilité des fréquences, comme l'explique la discussion menée plus haut.
– la loi des grands nombres n'est qu'un théorème de mathématiques. C'est faux dans la mesure où les objets mathématiques dont elle traite servent de modèles pour décrire des situations réelles. La portée pratique de la loi des grand nombre est exactement celle des modèles auxquels elle peut s'appliquer.

3.3 Applications

Dans cette partie, nous présentons quelques applications concrètes de la loi des grands nombres, qu'il s'agisse exactement de celle que nous avons énoncée, ou plus largement de résultats entrant dans la même catégorie.

3.3.1 L'assurance et la mutualisation du risque

Le principe fondamental de l'assurance repose sur la loi des grands nombres. Considérons par exemple le risque associé aux dégâts qui peuvent être causés à un véhicule au cours d'une année (vol, accident, vandalisme,...) Pour un individu donné, la perte financière associée à un tel risque peut être représentée par une variable aléatoire. Avec une assez forte probabilité, le véhicule ne subit aucun dégât, et la variable aléatoire représentant la perte est donc nulle dans ce cas. Avec une probabilité très faible, le véhicule est volé ou complètement détruit, ce qui représente une perte financière importante, mais peu probable. De petits dégâts, représentant une perte moindre, possèderont une probabilité plus importante. Globalement, ceci se traduit par le fait que l'**espérance** de la perte possède une valeur faible en comparaison des pertes considérables qu'occasionnerait un dégât sérieux. Par exemple (en euros), la perte financière pourrait être modélisée par une variable aléatoire X de loi :

$$\mathbb{P}(X = 15000) = \frac{5}{1000}\ ,\ \mathbb{P}(X = 1000) = \frac{50}{1000},$$

$$\mathbb{P}(X = 200) = \frac{150}{1000}\ ,\ \mathbb{P}(X = 0) = \frac{795}{1000},$$

dont l'espérance est égale à :

$$\mathbb{E}(X) = \frac{5}{1000} \times 15000 + \frac{50}{1000} \times 1000 + \frac{150}{1000} \times 200 + \frac{795}{1000} \times 0 = 155,$$

et possède donc une valeur nettement plus faible que la plupart des pertes possibles.

Cependant, un individu isolé n'est confronté qu'à une seule réalisation de la variable aléatoire X, relative à son propre véhicule, et la valeur moyenne de X n'a que peu de sens pour cet individu pris isolément : avec une probabilité faible, mais non-négligeable, il doit accepter d'être confronté à l'éventualité d'une perte considérable, bien supérieure à 155 euros, que rien ne viendra compenser. Il est ainsi soumis à un risque individuel, aléatoire, et potentiellement important.

Le principe de l'assurance consiste à mutualiser les risques attachés à un grand nombre d'individus différents, de façon à éliminer complètement le risque aléatoire individuel, moyennant le versement d'une prime fixée à l'avance. Le montant total des pertes subies par N individus est égal à :

$$M = X_1 + \cdots + X_N,$$

où X_i désigne la perte subie par l'individu numéro i. Si chaque individu accepte de verser à l'avance à une compagnie d'assurance une somme légèrement supérieure à la perte moyenne, par exemple 160 euros, le montant total des sommes collectées par l'assurance s'élève à $N \times 160$.

En admettant que les pertes des différents individus sont indépendantes, la loi des grands nombres entraîne alors que, si N est suffisamment grand, le montant total M de la perte est inférieur au total des primes collectées : avec une probabilité très proche de 1,

$$\left| \frac{1}{N} (X_1 + \cdots + X_N) - 155 \right| < 5,$$

d'où le fait que :

$$M < 160 \times N$$

avec une très forte probabilité. Par conséquent, l´argent collecté auprès des N individus permet de compenser intégralement la perte aléatoire subie par chacun des N individus, et le risque individuel est ainsi annulé. C'est le principe de la mutualisation du risque : la somme des risques individuels associés à chaque individu donnant lieu à une valeur totale quasiment certaine, celle-ci peut donc être évaluée à l'avance, et chaque individu n'a qu'à payer de façon *certaine* une somme légèrement supérieure au risque moyen, pour être complètement couvert avec une quasi-certitude. (Bien entendu, les choses sont moins simples en pratique. Par exemple , il peut exister plusieurs types différents de couverture, les assurés peuvent être répartis en catégories correspondant à différents niveaux de risque, la question de l'aléa moral et des franchises à appliquer doit entrer en ligne de compte, ainsi que des considérations commerciales,..., mais le principe de base est bien celui de la loi des grands nombres.)

L'évaluation du risque moyen (c'est entre autres le métier des actuaires) est donc fondamentale pour les compagnies d'assurances, et fait également appel à la loi des grands nombres : en étudiant le montant total des pertes subies par un grand nombre d'individus, on peut évaluer précisément la valeur moyenne de la perte. La différence entre la prime versée et le risque moyen s'explique au moins par deux contributions distinctes : la nécessité de garantir que les pertes subies ne dépasseront le montant des primes collectées qu'avec une probabilité extrêmement faible (il s'agit donc de préciser le ϵ et le α), et, d'autre part, les frais de fonctionnement, salaires, provisions, etc... à la charge de la compagnie d'assurance (sans oublier les bénéfices s'il ne s'agit pas d'une mutuelle). Evaluer correctement les provisions nécessaires pour rendre suffisament faible le risque d'insolvabilité de la compagnie d'assurance est bien entendu une question importante en pratique !

Par ailleurs, il est clair que tous les risques ne se prêtent pas à une mutualisation de ce type : des phénomènes exceptionnels (tels que catastrophes naturelles, guerres, grandes crises économiques, épidémies,...), qui affectent simultanément un très grand nombre de personnes, voire la totalité d'une population, n'entreront pas forcément correctement dans le cadre décrit ci-dessus (des risques à l'impact suffisament limité et affectant suffisament peu de personnes en même temps).

3.3.2 Sondages

Lorsque l'on décrit une expérience effectivement susceptible d'être répétée indépendamment un grand nombre de fois, la loi des grands nombres fait apparaître la probabilité comme un caractère physique de l'expérience, susceptible d'être mesuré : il suffit de répéter N fois l'expérience et de compter le nombre de fois où l'événement s'est réalisé pour évaluer sa probabilité, cette évaluation étant d'autant plus précise que N est grand. C'est le principe de base des sondages, qui reposent sur le fait qu'il suffit de sonder un échantillon de la population suffisament grand (mais très petit par rapport à la population totale, par exemple : 10 000 personnes pour une population de 60 millions d'individus) pour évaluer les proportions réelles au sein de la population totale.

3.3.3 Mécanique statistique

Un volume «ordinaire» de gaz à notre échelle (par exemple 10 litres), contient typiquement un très grand nombre de molécules identiques, de l'ordre de 10^{23}. Les paramètres physiques macroscopiques que l'on mesure, tels que la pression, ou la température, ne sont pas des caractéristiques d'une molécule du gaz en particulier, mais de l'ensemble du système de N molécules qui constitue le gaz, et apparaissent souvent comme de gigantesques moyennes associées à des caractéristiques physiques individuelles des molécules. Par exemple, dans le cas d'un gaz parfait (c'est-à-dire dans l'hypothèse où les molécules de gaz n'interagissent pas entre elles), l'énergie totale du gaz est :

$$U = \sum_{i=1}^{N} \frac{1}{2} m V_i^2,$$

où m désigne la masse d'une molécule de gaz, et V_i la vitesse de déplacement de la particule numéro i. Une modélistion probabiliste simple de ce système physique consiste à supposer que les vitesses V_i sont des variables aléatoires mutuellement indépendantes, puisque les particules n'interagissent pas entre elles. Dans le cadre de ce modèle, la loi des grands nombres explique pourquoi, bien que la vitesse d'une molécule donnée soit complètement aléatoire, ce que l'on peut effectivement observer en ne prenant en compte qu'un volume de gaz minuscule, la quantité physique macroscopique que l'on mesure prend une valeur bien déterminée, qui ne change pas aléatoirement à chaque nouvelle expérience : la somme des hasards individuels associés à la vitesse de chaque molécule concourt à former une valeur quasiment déterministe, du fait du très grand nombre de molécules en présence. La validité de la loi des grands nombres s'étend en fait bien au-delà de l'hypothèse très simplificatrice de l'indépendance entre les molécules, et joue un rôle fondamental en physique statistique.

3.3.4 Méthodes de Monte-Carlo

Le principe général des méthodes de Monte-Carlo est d'utiliser le hasard simulé pour évaluer des quantités qui apparaissent comme l'espérance d'une variable aléatoire, en calculant la moyenne d'un grand nombre de réalisations indépendantes de cette variable aléatoire. On peut appliquer ces méthodes, soit à des modèles stochastiques complexes (par exemple en épidémiologie, économie, physique) pour lesquels il n'est pas possible de mener des calculs explicites (ou même approchés), soit à des problèmes d'analyse numérique n'ayant *a priori* rien à voir avec l'aléatoire (système linéaires, équations aux dérivées partielles, intégrales) mais pour lesquels les méthodes de Monte-Carlo sont plus efficaces et/ou plus faciles à mettre en oeuvre que d'autres. Ces méthodes constituent un outil quasiment indispensable pour la modélisation et l'étude de systèmes complexes, et, pour cette raison, leur utilisation s'étend à la plupart des sciences qui font appel à la modélisation.

Voici deux exemples (simples) de leur utilisation :

Le problème de la percolation

Il s'agit d'un modèle stochastique destiné à étudier la possibilité pour un liquide de s'écouler à travers un matériau poreux (par exemple, l'eau à travers le café moulu...). On considère pour cela un cube du réseau $\mathbb{Z}^3$, centré en l'origine, de côté fixé. Chaque point du cube est initialement relié par une arête aux autres points du cube qui se trouvent à distance 1 de ce point. La structure des arêtes de ce cube est ensuite modifiée par une suppression aléatoire d'arêtes, chaque arête étant supprimée avec une probabilité $p \in]0,1[$, indépendamment des autres. On obtient ainsi un cube dans lequel un certain nombre d'arêtes ont disparu, et la question que l'on pose est la suivante : quelle est la probabilité pour que, dans un cube modifié par ce procédé, l'origine soit encore reliée à un sommet situé au bord du cube par une suite d'arêtes ? (Ce qui correspond, dans l'interprétation physique du modèle, à la possibilité pour un liquide de s'écouler de l'origine vers les bords.) On ne sait pas calculer explicitement cette probabilité, et l'une des possibilités pour l'évaluer numériquement consiste à générer un grand nombre de cubes auxquels on fait subir, aléatoirement et indépendamment des autres, la procédure de suppression des arêtes décrite ci-dessus. On vérifie, pour chaque cube, s'il existe un chemin menant de l'origine au bord du cube, et la proportion de cubes possédant cette propriété fournit, d'après la loi des grands nombres, une évaluation de la probabilité recherchée. Plus formellement, appelons $H_1, \ldots, H_N$ une suite de cubes ainsi générée, et, $l(H)$ la fonction qui vaut 1 si l'origine du cube H est reliée par un chemin au bord du cube, et 0 sinon. D'après la loi des grands nombres, la probabilité que nous recherchons est approximativement

égale, lorsque N est grand, à :

$$\frac{1}{N}\left(l(H_1)+\cdots+l(H_n)\right).$$

Evaluer un volume

Supposons que nous cherchions à évaluer le volume d'un objet tri-dimensionnel A, l'appartenance d'un point de l'espace A étant facile à tester algorithmiquement. Par exemple, l'ensemble des points de l'espace défini par :

$$A=\left\{(x,y,z)\in[-1,1]^3\ :\ x^2+3y^3-2xy^2\leq 2\ ,\ xy^5-7x^2\sin(y)\geq -1\right\}.$$

Déterminer le volume de A n'est pas *a priori* une tâche aisée, mais, en revanche, il est très facile de tester l'appartenance d'un point de coordonnées (x,y,x) à A, en vérifiant si oui ou non le triplet (x,y,z) vérifie les conditions qui définissent A. Pour évaluer le volume de A, une première étape consiste à discrétiser le cube $[-1,1]^3$ dans lequel A est inscrit en «petites» cellules, par exemple 10^{15} cellules cubiques, notées C_i, de côté $2/100000$. Une approximation du volume de A est alors fournie par la somme des volumes des cellules dont le centre se trouve dans A. En notant $g_A(C_i)$ la fonction qui vaut 1 lorsque le centre de C_i se trouve dans A, et 0 sinon, on a donc :

$$Vol(A)\approx\sum_{i=1}^{10^15} Vol(C_i)g_A(C_i).$$

Bien entendu, il est hors de question d'effectuer le calcul complet de cette somme, pour des raisons de temps d'exécution. L'utilisation de la méthode de Monte-Carlo repose sur le fait que l'égalité précédente peut se réécrire :

$$Vol(A)\approx\sum_{i=1}^{10^15} 10^{-15}g_A(C_i)=\sum_{i=1}^{10^15}\mathbb{P}(\mathcal{C}=C_i)g_A(C_i)=\mathbb{E}(g_A(\mathcal{C})),$$

où $\mathcal{C}$ désigne une variable aléatoire dont la loi est la loi uniforme sur l'ensemble des cellules C_i, chacune des 10^15 cellules ayant la même probabilité d'être choisie. On peut alors, d'après la loi des grands nombres, évaluer le volume de A en générant un grand nombre de réalisations indépendantes de $\mathcal{C}$, $\mathcal{C}_1,\ldots,\mathcal{C}_N$, et en calculant la moyenne empirique de g_A :

$$vol(A)\approx\mathbb{E}(g_A(\mathcal{C}))\approx\frac{1}{N}\sum_{j=1}^{N} g_A(\mathcal{C}_j).$$

Cette méthode s'applique également pour calculer une intégrale multiple dans le cas général, son principal intérêt par rapport aux autres procédés d'intégration

approchée étant qu'elle conserve la même forme quelle que soit la dimension de l'intégrale à évaluer, et que son application ne nécessite pas d'hypothèse sur la régularité (continuité, dérivabilité,...) de la fonction à intégrer. Les deux exemples d'utilisation de la méthode de Monte-Carlo que nous venons de présenter sont assez rudimentaires, mais illustrent le principe de base selon lequel une espérance est évaluée expérimentalement à l'aide de la loi des grands nombres. Des raffinements considérables ont été apportés à cette méthode, visant notamment à en améliorer la précision et la vitesse de convergence, ainsi qu'à mieux estimer le temps de calcul nécessaire, la méthode ne fournissant pas *a priori* de critère d'arrêt.

De la sociologie suicidaire ?

Enfin, la loi des grands nombres est parfois employée à des fins explicatives dans l'étude des phénomènes sociaux, avec tout ce que la modélisation peut avoir de problèmatique dans ce contexte. Elle explique pourquoi des quantités *a priori* aléatoires, et qui, dans le cadre d'une modélisation probabiliste, apparaissent comme des fréquences de réalisation d'un certain événement au cours d'un grand nombre d'expériences indépendantes, présentent une valeur approximativement constante. Par exemple, pourquoi le taux de suicide dans une région donnée reste-t-il à peu près fixe dans le temps, alors qu'il semble impossible d'admettre que les individus se concertent pour maintenir ce taux à une valeur constante ? La loi des grands nombres fournit une explication de ce phénomène qui a beaucoup intrigué les sociologues de la fin du XIXème siècle : en admettant que chaque individu a une probabilité fixe de se suicider, indépendamment des autres, la loi des grands nombres entraîne que le taux de suicide au sein d'une population nombreuse est une variable aléatoire approximativement constante. La somme des hasards individuels conduit à un résultat quasiment certain, du fait du grand nombre d'individus en présence.

3.4 Inégalités de déviation

Plus tard...

3.5 Convergence de la loi empirique

Plus tard...

3.5.1 Convergence des histogrammes

3.5.2 Le théorème de Glivenko-Cantelli

Plus tard...

3.6 Auto-évaluation

- Énoncez précisément les deux versions de la loi des grands nombres (hypothèses, et conclusion).
- En quoi la deuxième version entraîne-t-elle la première ?
- Quel lien la loi des grands nombres établit-elle entre loi et loi empirique ? Et entre moyenne théorique (espérance) et moyenne empirique ?
- En quoi la loi des grands nombres énonce-t-elle un comportement typique ? Quelle différence y a-t-il avec un comportement moyen ?
- En quoi la loi des grands nombres prouve-t-elle qu'une certaine quantité aléatoire est en fait essentiellement constante ?

3.7 Exercices

Exercice 154 *H. est passionné par la bourse, et consacre une grande partie de son temps à acheter et vendre des actions sur internet. Tous les mois, le montant de ses actifs se trouve multiplié par un coefficient aléatoire. On suppose que les coefficients associés aux mois successifs correspondent à des répétitions indépendantes d'une même variable aléatoire α, dont la loi est la suivante :*

$$\mathbb{P}(\alpha = 1,3) = 1/2 \ , \ \mathbb{P}(\alpha = 0,75) = 1/2.$$

Quelle est l'espérance de α ? Selon vous, comment la fortune de H. évolue-t-elle à long terme ?

Exercice 155 *Chez Jojo, dans le tiroir de la commode, se trouvent trois pièces de monnaie. Jojo se livre à l'expérience suivante : il ouvre le tiroir, choisit au hasard l'une des trois pièces, et effectue 10000 lancers. Il remet ensuite la pièce dans le tiroir, après avoir soigneusement noté la proportion de «face» obtenue. Il recommence l'expérience le lendemain, et obtient une valeur complètement différente pour la proportion de «face». Ces expériences contredisent-elles la loi des grands nombres ?*

Exercice 156 *M. C., marabout de son état, propose à ses clients de déterminer le sexe de leur enfant à naître dès sa conception. Pour gage de son talent, il propose même de rembourser les honoraires perçus, au cas où il se tromperait. Cette proposition engage-t-elle réellement la fiabilité de ses prédictions ? Justifiez.*

Exercice 157 *Toutes les dix secondes, Jojo peut (ou non) penser à envoyer un courrier électronique à son amie Hildegarde, de son lieu de travail. Celle-ci est extrêmement jalouse, et Jojo sait bien que si, par malheur, il s'écoulait une journée sans qu'il lui fît parvenir le moindre message, les conséquences en seraient incalculables...*

Sachant que les journées de travail de Jojo durent huit heures, et que, au cours des trente derniers jours, Jojo a envoyé en moyenne 2,3 messages par jour à son amie, pouvez-vous estimer la probabilité pour que l'irréparable se produise aujourd'hui ? Et au cours des trois prochains jours ?

Exercice 158 *Dans l'édition de demain du prestigieux journal «Jojo – Gazette», deux correcteurs différents ont relevé l'un 42 erreurs, l'autre 54 erreurs, seules 12 erreurs ayant été relevées à la fois par l'un et l'autre des deux correcteurs. Sur la base de ces données, pouvez-vous proposer une estimation du nombre total d'erreurs dans le journal ?*

Exercice 159 *On cherche à sonder la population au sujet d'un comportement d'ordre privé, et le caractère embarassant, au moins pour une minorité de personnes, de la question posée, amène à douter de la sincérité des réponses. En admettant que la question posée possède une réponse binaire de type oui/non, on se propose de procéder de la manière suivante. Avant d'être interrogée, chaque personne lance un dé à six faces, dont elle seule connaît le résultat. Si elle obtient un chiffre différent de six, elle doit répondre le contraire de la vérité lorsqu'elle est interrogée. Si elle obtient un six, elle doit en revanche répondre correctement. Comment, à partir des résultats ainsi obtenus, peut-on obtenir l'information souhaitée ? Pourquoi cette méthode devrait-elle inciter les personnes interrogées à répondre sincèrement ? Plus généralement, en admettant que l'on puisse fixer à une valeur arbitraire x (entre 0 et 1) la probabilité d'obtenir un six, comment peut-on fixer au mieux la valeur de x dans ce problème ?*

Exercice 160 *Prouvez, à partir de la loi des grands nombres que nous avons énoncée (pour des variables aléatoires à valeurs dans $\mathbb{R}$) un résultat analogue pour des variables aléatoires à valeurs dans $\mathbb{R}^d$.*

Exercice 161 *On considère une variable aléatoire X prenant la valeur N avec une probabilité de $1/N$, et la valeur 0 avec probabilité $1 - 1/N$. Quelle est l'espérance de X ? Considérons $X_1, \ldots, X_N$ des réalisations indépendantes de X. Est-il raisonnable de considérer la variable aléatoire $\frac{1}{N}(X_1 + \cdots + X_N)$ comme typiquement proche de cette espérance lorsque N est grand ?*

Chapitre 4

La courbe en cloche

4.1 Introduction

Ce chapitre est consacré à l'étude des lois de probabilité gaussiennes. L'intérêt de cette étude est tout sauf purement théorique, car les lois gaussiennes interviennent dans de très nombreux de modèles de situations concrètes.

Après avoir présenté les principales caractéristiques de cette famille de distributions, nous présenterons une classe de situations d'une importance fondamentale, et dans lesquelles les lois gaussiennes apparaissent de manière quasiment universelle, à savoir la description des fluctuations des sommes d'un grand nombre de variables aléatoires indépendantes.

Diverses illustrations et applications suivent, avant d'aborder la question plus complexe, mais très importante également, des lois gaussiennes mutli-dimensionnelles.

4.2 Les lois gaussiennes unidimensionnelles

On appelle courbe en cloche, ou plus correctement **courbe gaussienne** [1], toute fonction définie sur $\mathbb{R}$ de la forme :

$$\phi_{m,v}(x) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(x-m)^2}{2v}\right),$$

où $m \in \mathbb{R}$ et $v > 0$. Celle-ci définit une densité de probabilité sur $\mathbb{R}$, et la **loi gaussienne de paramètres** m **et** v est la loi (continue) de probabilité possédant la densité $\phi_{m,v}$ (on emploie parfois également pour la désigner le terme de loi **normale**, afin de souligner son importance, ou encore de loi de **Laplace-Gauss** [2]

1. Du nom de l'illustre mathématicien Carl Friedrich Gauss (1777–1855).
2. Du nom du non moins illustre mathématicien Pierre-Simon Laplace (1749–1827).

La gaussienne de paramètres $m = 0$ et $v = 1$ est appelée la gaussienne standard, ou encore gaussienne centrée réduite, du fait que, comme nous le verrons, elle possède une espérance égale à 0 et une variance égale à 1.

Pour vérifier qu'il s'agit bien d'une densité de probabilité (la positivité étant évidente), il convient de vérifier la condition de normalisation :

$$\int_{-\infty}^{+\infty} \phi_{m,v}(u)du = 1.$$

Cette égalité se ramène à l'autre, bien connue (voir votre cours d'analyse de premier cycle) :

$$\int_{-\infty}^{+\infty} e^{-x^2/2}dx = \sqrt{2\pi},$$

qui n'est autre que la condition de normalisation pour la gaussienne $\phi_{0,1}$. Moyennant un changement de variables décrit un peu plus loin, on peut en déduire la condition de normalisation pour toute gaussienne $\phi_{m,v}$ (à vous de faire la vérification !)

La courbe représentative d'une telle fonction présente effectivement l'aspect d'une cloche (en gros !), et les deux paramètres m et v déterminent précisément la forme de la cloche. On vérifie facilement que le point m est celui où $\phi_{m,v}$ prend son maximum, le «sommet» de la cloche : c'est celui où $(x-m)^2$ est minimal, car égal à 0. Qui plus est, on note que la cloche est **symétrique** par rapport à l'axe $x = m$, ce que l'on vérifie rigoureusement en établissant la relation (immédiate au vu de la définition de $\phi_{m,v}$) : pour tout $y \in \mathbb{R}$,

$$\phi_{m,v}(m + y) = \phi_{m,v}(m - y).$$

La valeur de m détermine ainsi la position horizontale de la cloche. La valeur de v étant fixée, on note que les courbes représentatives des fonctions $\phi_{a,v}$, a décrivant $\mathbb{R}$ se déduisent les unes des autres par des translations horizontales. Plus rigoureusement, on vérifie que, pour tout couple a_1, a_2, et tout $y \in \mathbb{R}$,

$$\phi_{a_2,v}(y) = \phi_{a_1,v}[y - (a_2 - a_1)].$$

On peut donc ramener, par translation horizontale, l'étude de la gaussienne $\phi_{m,v}$ à celle de la gaussienne $\phi_{0,v}$. Le paramètre v détermine, lui, la «taille» de la cloche. On établit précisément que, pour tous $v_1, v_2 > 0$, et tout $y \in \mathbb{R}$,

$$\phi_{0,v_2}(y) = \sqrt{v_1/v_2} \times \phi_{0,v_1}\left(\frac{y}{\sqrt{v_2/v_1}}\right).$$

Si l'on prend comme référence la courbe $\phi_{0,1}$, que l'on appelle **gaussienne standard** la courbe $\phi_{0,v}$ s'en déduit donc par une dilatation de coefficient $\sqrt{v}$ sur l'échelle horizontale, suivie d'une dilatation d'un facteur $1/\sqrt{v}$ sur l'échelle verticale. Ainsi,

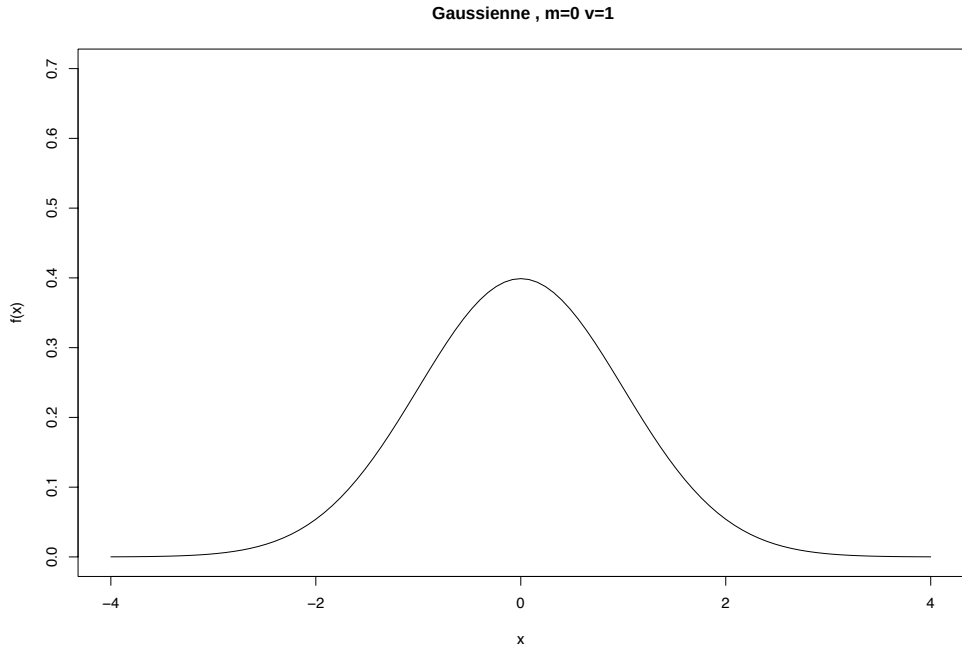
plus v est grand, plus la cloche est plate et étalée, plus v est petit, plus la cloche est haute et resserrée.

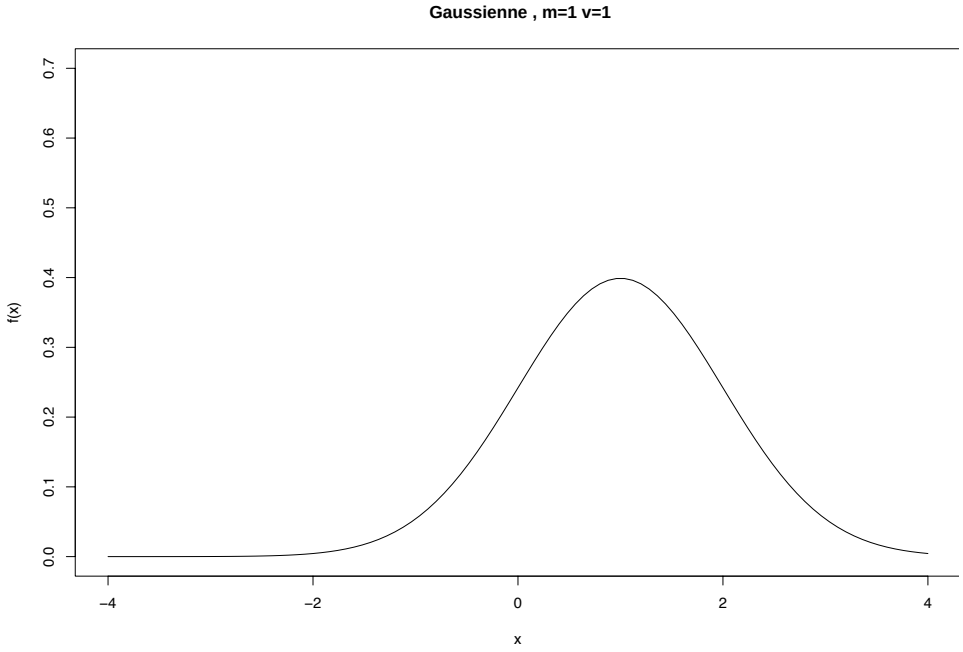
On note que la condition de normalisation

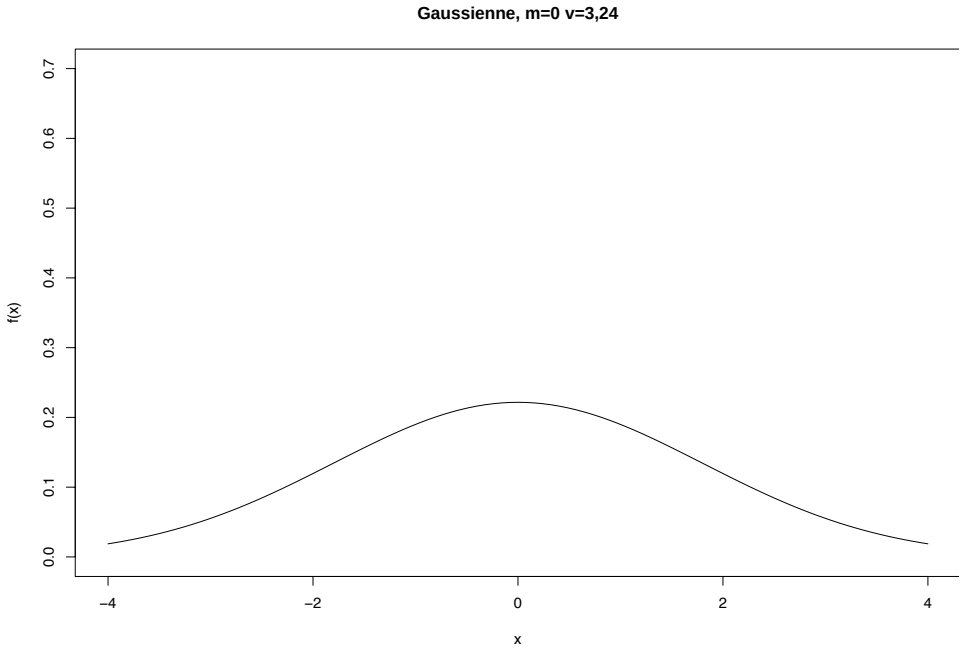
$$\int_{-\infty}^{+\infty} \phi_{m,v}(u)du = 1$$

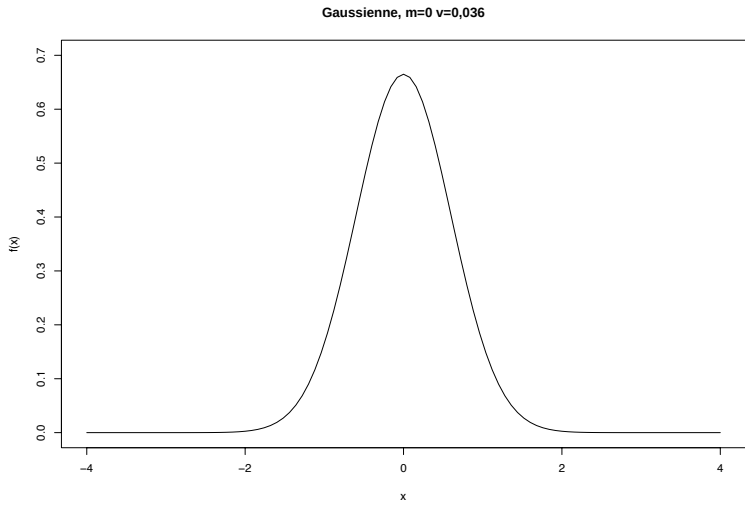
impose nécessairement que la cloche ne puisse pas diminuer de hauteur sans s'élargir, ou augmenter de hauteur sans devenir plus étroite.

Les pages suivantes contiennent quelques représentations graphiques destinées à illustrer ces propriétés.




Gaussienne , m=1 v=1
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−4
−2
0
2
4
x


Gaussienne, m=0 v=3,24
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−4
−2
0
2
4
x



Ces transformations géométriques permettant de passer d'une gaussienne à une autre ont également une interprétation très importante vis-à-vis des variables aléatoires dont les courbes gaussiennes représentent les densités. Nous retiendrons le résultat fondamental suivant : pour tous nombres réels $a \neq 0$ et b, **si X suit une loi gaussienne de paramètres m et v, alors la variable aléatoire :**

$$aX + b$$

suit une loi gaussienne de paramètres $m' = am + b$ et $v' = a^2 \times v$. En particulier, la variable aléatoire

$$\frac{X - m}{\sqrt{v}}$$

suit une loi gaussienne de paramètres 0 et 1 (gaussienne standard).

Cette propriété résulte de la formule de changement de variables donnée dans le chapitre «Variables aléatoires.»

Les paramètres m et v, quant à eux, n'ont pas seulement une signification géométrique en rapport avec les courbes gaussiennes, mais possèdent également une signification très importante vis-à-vis des variables aléatoires dont les courbes gaussiennes représentent les densités. De manière précise : **si X est une variable aléatoire continue de loi gaussienne de paramètres m et v, m est égal à l'espérance de X, et v à sa variance**. En d'autres termes :

$$\mathbb{E}(X) = m,\ \mathbb{V}(X) = v.$$

Pour prouver ces deux égalités, on utilise la formule donnant l'espérance d'une va-

riable aléatoire de loi continue, et l'on doit donc prouver que :

$$\int_{-\infty}^{+\infty} y \times \phi_{m,v}(y)dy = m$$

et

$$\int_{-\infty}^{+\infty} (y-m)^2 \times \phi_{m,v}(y)dy = v.$$

La preuve de ces égalités est laissée en exercice (pour l'espérance, la fonction à intégrer possède une primitive évidente, et, pour la variance, une intégration par parties permet de conclure).

En vertu de la symétrie de la cloche par rapport à son sommet (ce qui entraîne d'ailleurs facilement que $m = \mathbb{E}(X)$), une variable aléatoire gaussienne X possède donc la propriété remarquable selon laquelle $\mathbb{P}(X < m) = \mathbb{P}(X > m) = 1/2$ (en général l'espérance d'une variable aléatoire n'est pas sa médiane). Qui plus est, m correspond également au mode (le sommet de la cloche).

De plus, la loi d'une variable aléatoire gaussienne est entièrement déterminée par la connaissance de son espérance et sa variance (rappelons qu'en général, deux variables aléatoires possédant même espérance et même variance n'ont pas la même loi, voir l'exercice 77).

Notez que la signification géométrique des paramètres m et v s'accorde avec celle de l'espérance et de la variance d'une variable aléatoire en tant qu'indicateurs de position et de dispersion respectivement. Géométriquement, m indique la localisation de la cloche, donc la zone où la variable aléatoire a une probabilité non-négligeable de prendre ses valeurs. De même, plus v est grande, plus la cloche est large, et plus la variabilité de la variable aléatoire correspondante est élevée (à l'inverse, plus v est faible, plus la cloche est étroite, et plus les valeurs prises par la variable sont localisées à proximité de m). Les paramètres m et v jouent précisément le rôle de paramètres d'échelle et de position puisque, comme nous l'avons noté précédemment, si X suit une loi gaussienne de paramètres 0 et 1, $m + \sqrt{v}X$ suit une loi gaussienne de paramètres m et v.

La loi gaussienne centrée réduite sert donc de référence pour étudier la distribution des lois gaussiennes générales.

Par exemple, on a l'inégalité

$$\int_{-2}^{2} \phi_{0,1}(u)du \gtrsim 0,95.$$

En termes de variables aléatoires, ceci signifie qu'une variable aléatoire X qui suit une loi gaussienne standard vérifie :

$$\mathbb{P}(X \in [-2,2]) \gtrsim 0,95.$$

En utilisant la transformation permettant de passer d'une gaussienne de paramètres m et v à une gaussienne standard, ceci se traduit, pour une variable aléatoire Y gaussienne de paramètres m et v par :

$$\mathbb{P}(Y \in [m - 2\sqrt{v}, m + 2\sqrt{v}]) = \mathbb{P}\left(\frac{Y - m}{\sqrt{v}} \in [-2, 2]\right) = \mathbb{P}(X \in [-2, 2]) \gtrsim 0,95.$$

Par conséquent, avec une probabilité supérieure à 95%, une variable aléatoire suivant une loi gaussienne prend une valeur qui s'écarte de son espérance de moins de deux écarts-types.

À l'inverse, en utilisant l'inégalité :

$$\int_{-1}^{1} \phi_{0,1}(u)du \lesssim 0,7,$$

on obtient que

$$\mathbb{P}(X \notin [-1, 1]) \gtrsim 0,3$$

et que

$$\mathbb{P}(Y \in [m - \sqrt{v}, m + \sqrt{v}]) \lesssim 0,7.$$

Ainsi, avec une probabilité supérieure à 30%, une variable aléatoire suivant une loi gaussienne prend une valeur qui s'écarte de son espérance de plus d'un écart-type.

Ces deux inégalités ne sont donnés qu'à titre d'exemples, et parce qu'elles sont faciles à retenir, on peut en obtenir autant que l'on veut, pour trois écarts-types, un demi écart-type, etc...

Il est à noter qu'il n'existe pas de formule explicite en termes de fonctions élémentaires permettant de calculer, en fonction de a et b, les intégrales définissant

$$\mathbb{P}(X \in [a, b]) = \int_a^b \phi_{m,v}(u)du.$$

En revanche, on dispose de méthode numériques rapides et précises pour les calculer, ainsi que de tables collectant leurs valeurs.

Plus précisément, on dispose de moyens numériques permettant d'effectuer les opérations suivantes avec une précision satisfaisante :

– étant donné $x \in \mathbb{R}$, calculer l'intégrale

$$\int_{-\infty}^{a} \phi_{0,1}(u)du,$$

– étant donné $q \in]0, 1[$, trouver $x \in \mathbb{R}$ tel que

$$\int_{-\infty}^{x} \phi_{0,1}(u)du = q.$$

Rappelons pour finir le résultat très important de l'exercice 130 : **si X et Y sont deux variables aléatoires indépendantes, X suivant une loi gaussienne de paramètres m et v, Y suivant une loi gaussienne de paramètres m' et v', $X+Y$ suit une loi gaussienne de paramètres $m+m'$ et $v+v'$.**

Notez que la partie non-banale de ce résultat est que la loi de $X+Y$ est gaussienne (l'espérance de $X+Y$ est toujours égale à $m+m'$, et, X et Y étant supposées indépendantes, la variance de $X+Y$ est nécessairement égale à $v+v'$. Voir également à ce sujet l'exercice 169 et la remarque **??** sur le même sujet dans le paragraphe traitant des lois gaussiennes multidimensionnelles.

4.3 Le théorème de la limite centrale

4.3.1 Cadre et énoncé

On se place dans le même cadre que celui dans lequel nous avons énoncé la loi faible des grands nombres au chapitre précédent, que nous rappelons rapidement.

On considère donc un espace de probabilité $(\Omega, \mathbb{P})$, une variable aléatoire X définie sur Ω et à valeurs dans $\mathbb{R}$, l'espace de probabilité $(\Omega^N, \mathbb{P}^{\otimes N})$ décrivant N répétitions indépendantes de $(\Omega, \mathbb{P})$, et $X_1, \ldots, X_N$ les variables aléatoires correspondant à X dans chacune des réalisations successives.

De manière plus précise, les variables aléatoires X_i sont définies par Ω^N par $X_i((\omega_1, \ldots, \omega_N)) = X(\omega_i)$.

On remarque encore une fois que, partant de n'importe quel modèle probabiliste sur lequel est définie une famille de variables aléatoires $Y_1, \ldots, Y_N$ mutuellement indépendantes et possédant chacune la même loi, on peut se ramener à la situation décrite ci-dessus en considérant le modèle-image de $(Y_1, \ldots, Y_N)$.

Nous ferons l'hypothèse que $\mathbb{E}(X)$ **et** $\mathbb{V}(X)$ **sont définies,** ce qui constitue une restriction par rapport à l'énoncé de la loi des grands nombres, pour laquelle on supposait simplement l'existence de $\mathbb{E}(X)$. Nous supposerons également que $\mathbb{V}(X) \neq 0$, ce sans quoi les variables aléatoires considérées sont en fait constantes, et leur étude de peu d'intérêt ! Nous utiliserons dans la suite la notation $S_N = X_1 + \cdots + X_N$.

Le théorème de la limite centrale s'énonce alors de la manière suivante : **lorsque N tend vers l'infini, la loi de la variable aléatoire**

$$\frac{S_N - \mathbb{E}(S_N)}{\sqrt{\mathbb{V}(S_N)}}$$

tend vers une loi gaussienne centrée réduite ($m = 0$ et $v = 1$).

Nous n'avons pas défini précisément ce que signifie la convergence d'une suite de lois de probabilités. Un énoncé totalement précis du théorème de la limite centrale

est le suivant : pour tout intervalle $I \subset \mathbb{R}$, on a

$$\lim_{N\to+\infty} \mathbb{P}^{\otimes N}\left[\frac{S_N - \mathbb{E}(S_N)}{\sqrt{\mathbb{V}(S_N)}} \in I\right] = \int_I \phi_{0,1}(u)du,$$

où $\phi_{0,1}$ est la densité de la loi gaussienne centrée réduite, soit $\phi_{0,1}(u) = (2\pi)^{-1/2}\exp(-x^2/2)$.

Par un calcul déjà effectué au chapitre précédent, on vérifie que

$$\begin{cases} \mathbb{E}(S_N) = N \times \mathbb{E}(X) \\ \mathbb{V}(S_N) = N \times \mathbb{V}(X) \end{cases}$$

Le théorème de la limite centrale peut donc se réécrire sous la forme :

$$\lim_{N\to+\infty} \mathbb{P}^{\otimes N}\left[\frac{X_1 + \cdots + X_N - N \times \mathbb{E}(X)}{\sqrt{N \times \mathbb{V}(X)}} \in I\right] = \int_I \phi_{0,1}(u)du.$$

Dans la suite, nous utiliserons la notation

$$\gamma_N = \frac{S_N - \mathbb{E}(S_N)}{\sqrt{\mathbb{V}(S_N)}}.$$

Remarque 13 *On ne suppose pas dans les hypothèses du théorème que la variable X est elle-même de loi gaussienne : toute loi pour laquelle l'espérance et la variance sont définies fait l'affaire. Dans le cas particulier où X est de loi gaussienne, la propriété selon laquelle la loi d'une somme de variables gaussiennes indépendantes est elle-même gaussienne montre que, pour tout N, la loi de γ_N est exactement la loi gaussienne centrée réduite, alors que le théorème ci-dessus n'énonce qu'une convergence lorsque N tend vers l'infini. En ce sens, la loi gaussienne constitue en quelque sorte un point fixe pour le théorème de la limite centrale, et fournit un début d'explication au fait que, quelle que soit la loi initiale de X, la loi limite de γ_N est gaussienne.*

Avant tout commentaire, nous donnons dans ce qui suit quelques illustrations graphiques de ce résultat.

Remarque 14 *Dans ce chapitre, nous utilisons principalement la représentation graphique des fonctions de répartition, plutôt que les histogrammes, non pas parce que l'un de ces modes de représentation est, de manière générale, préférable à l'autre, mais pour éviter d'avoir à gérer la question du choix de la largeur des classes des histogrammes, ce qui ajouterait, selon nous, à la complexité de l'exposé, les questions d'échelle étant absolument cruciales dans ce chapitre et ne pouvant donc pas être traitées à la légère. De plus, cela permet de varier quelque peu les plaisirs...*

4.3.2 Des illustrations lorsque la loi de $X_1 + \cdots + X_N$ est connue explicitement

Lorsque la loi de la somme $X_1 + \cdots + X_N$ est connue de manière explicite, – ce qui n'est pas le cas en général, – on peut effectuer une comparaison directe entre la loi de $\frac{X_1+\cdots+X_N-N\times\mathbb{E}(X)}{\sqrt{N\times\mathbb{V}(X)}}$ et la loi gaussienne de paramètres $m = 0$ et $v = 1$.

Nous présentons dans ce qui suit quatre exemples classiques (loi de Bernoulli, loi de Poisson, loi exponentielle, carré de gaussienne) pour lesquels un tel calcul explicite est possible. Notons que dans le cas trivial où la loi de X est elle-même gaussienne, on vérifie immédiatement que la limite dans l'énoncé du théorème est en fait une égalité, valable pour tout N.

Les graphiques suivants représentent la fonction de répartition

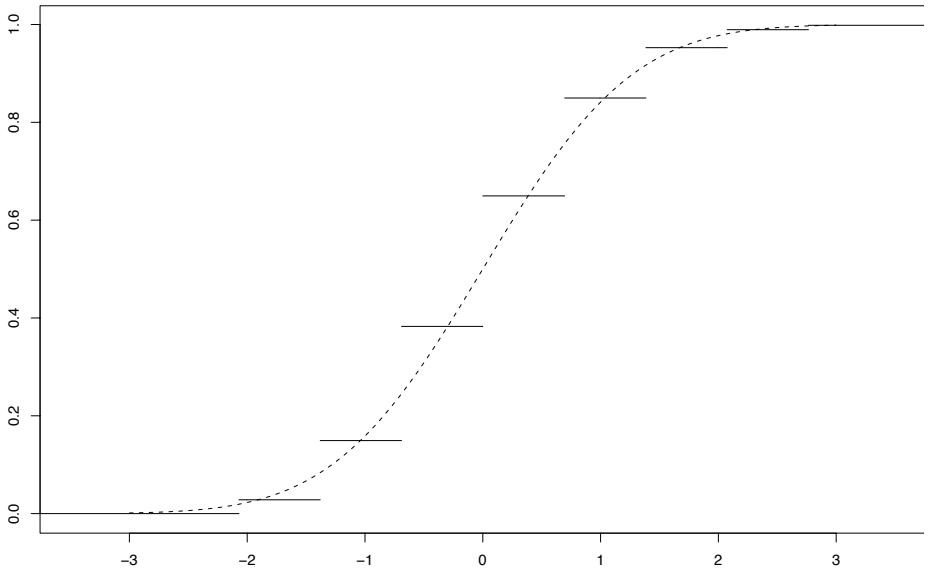
$$x \mapsto \mathbb{P}^{\otimes N}\left(\frac{X_1 + \cdots + X_N - N \times \mathbb{E}(X)}{\sqrt{N \times \mathbb{V}(X)}} \leq x\right)$$

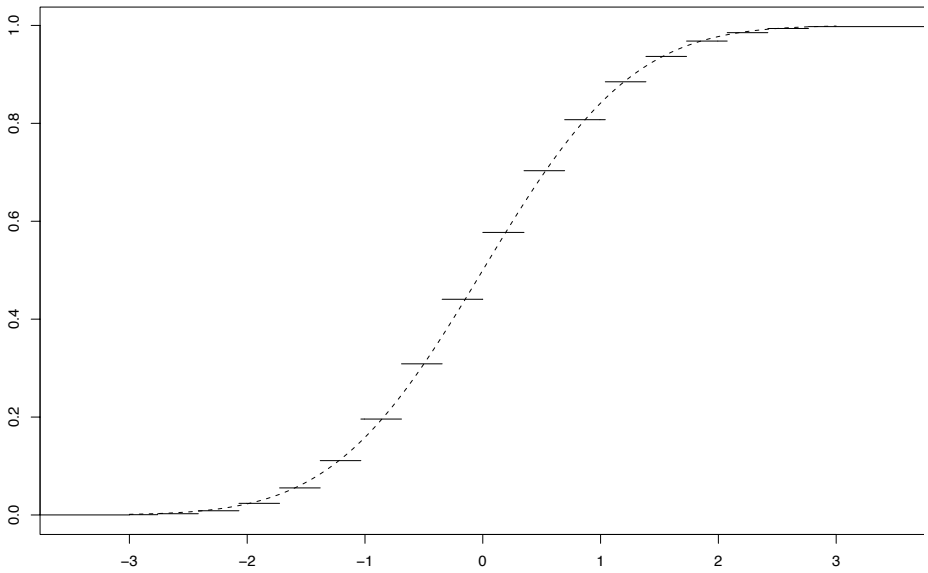
(en traits pleins) et celle de la gaussienne de paramètres $m = 0$ et $v = 1$, à savoir $x \mapsto \int_{-\infty}^{x} \phi_{0,1}(u)du$ (en traits intermittents), sur l'intervalle $[-3, +3]$. Cet intervalle représente pour la gaussienne $\phi_{0,1}$ une probabilité de présence de plus de $99,7\%$, et c'est typiquement pour des valeurs se trouvant dans cet intervalle que l'on utilise l'approximation fournie par le théorème de la limite centrale (voir la partie «Précision de l'approximation fournie par le théorème de la limite centrale» et «Attention à l'échelle» pour une discussion ce ce point).

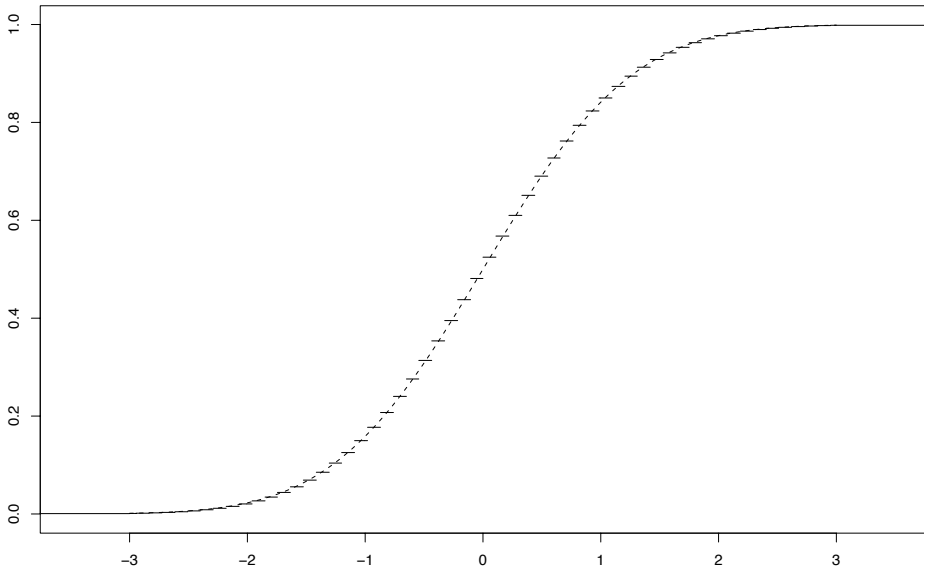
Variables aléatoires de loi de Bernoulli

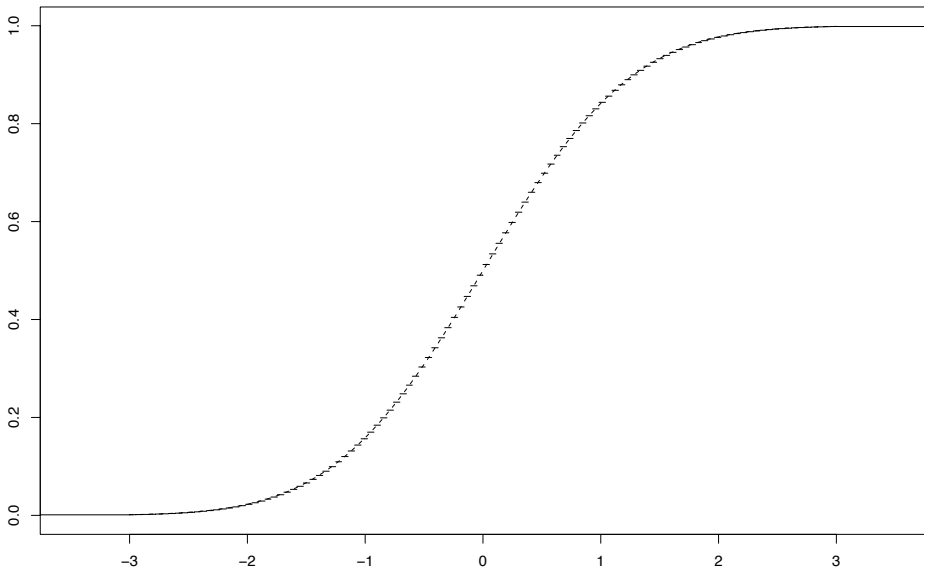
Dans cet exemple, on considère des variables aléatoires $X_1, \ldots, X_N$ indépendantes, possédant toutes la loi de Bernoulli de paramètre p. La loi de $X_1 + \cdots + X_N$ est naturellement connue dans ce cas : il s'agit d'une loi binomiale de paramètres n et p.

Les quatre graphiques suivants correspondent à $p = 0,3$ et N successivement égal à 10, 40, 400 et 1600.


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

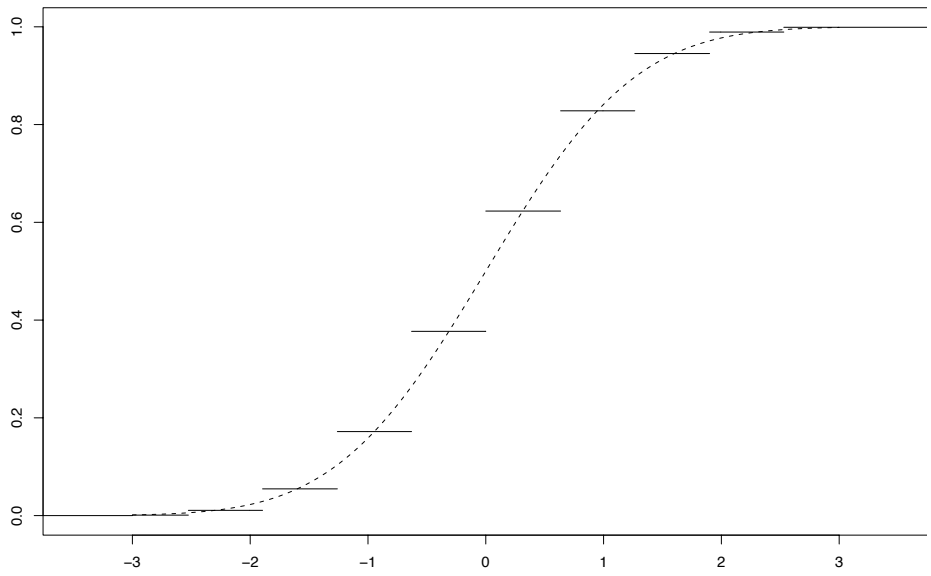

0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

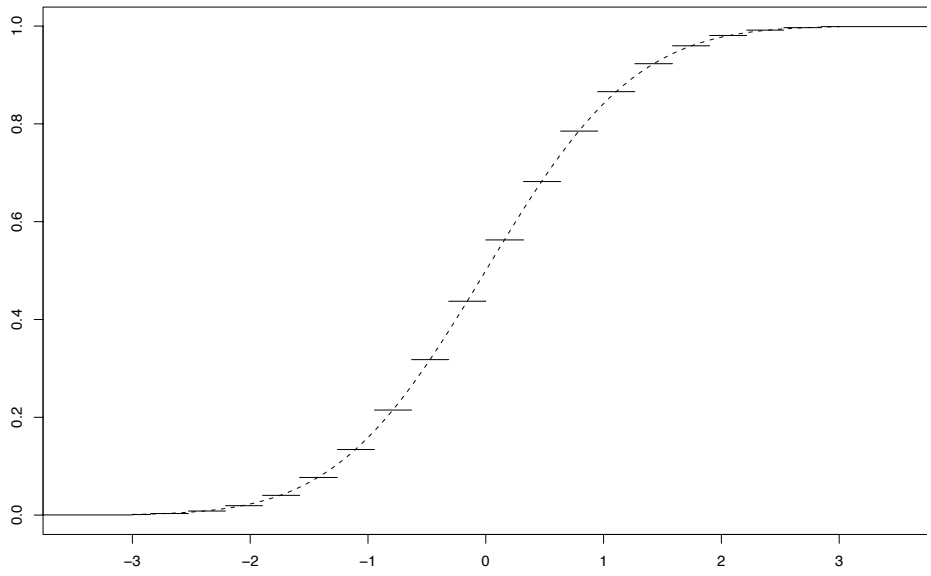


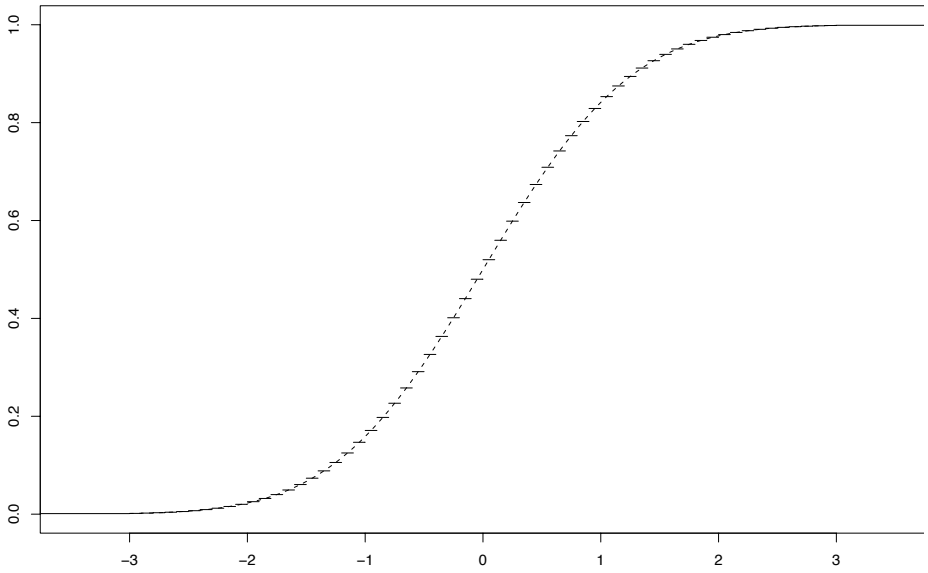


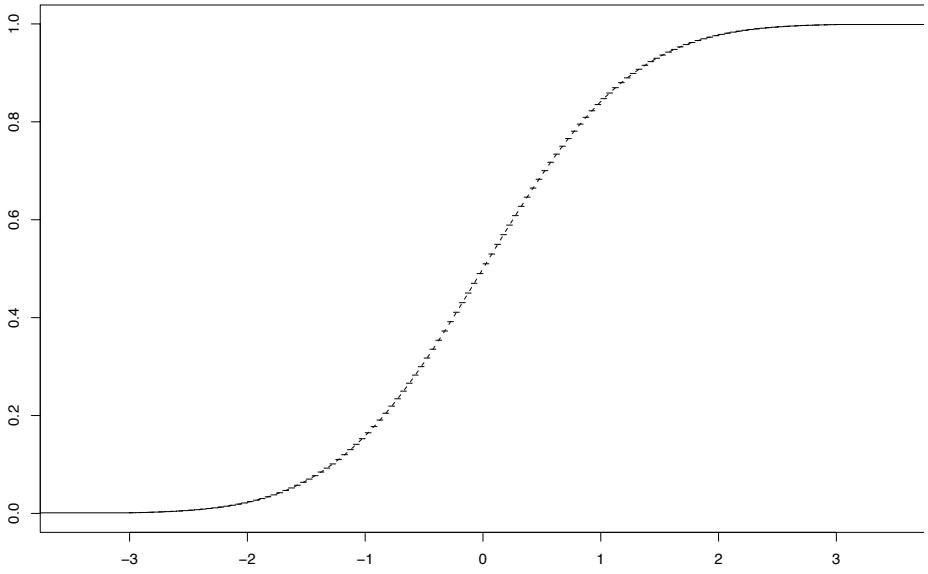
Voici à présent quatre graphiques correspondant à $p = 0,5$ et N successivement

égal à 10, 40, 400 et 1600.



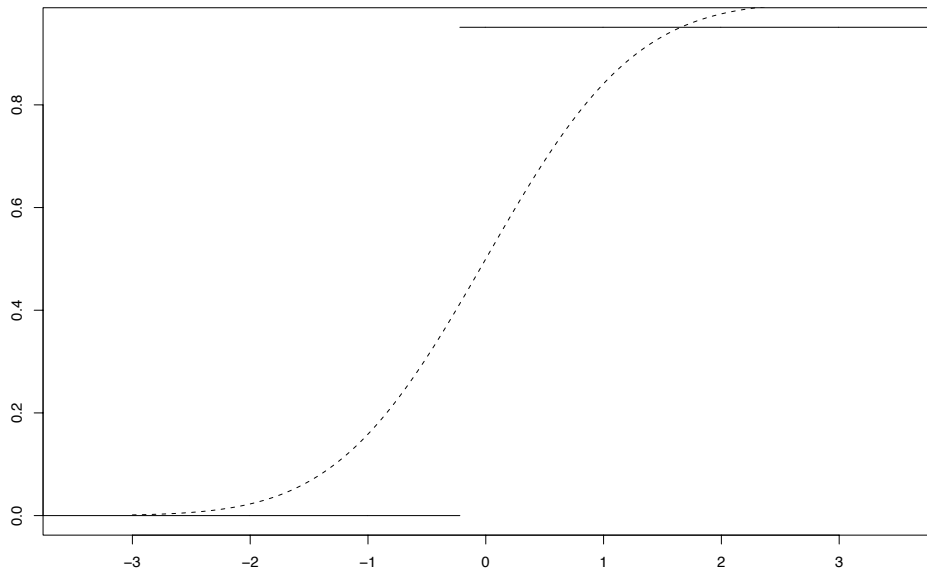


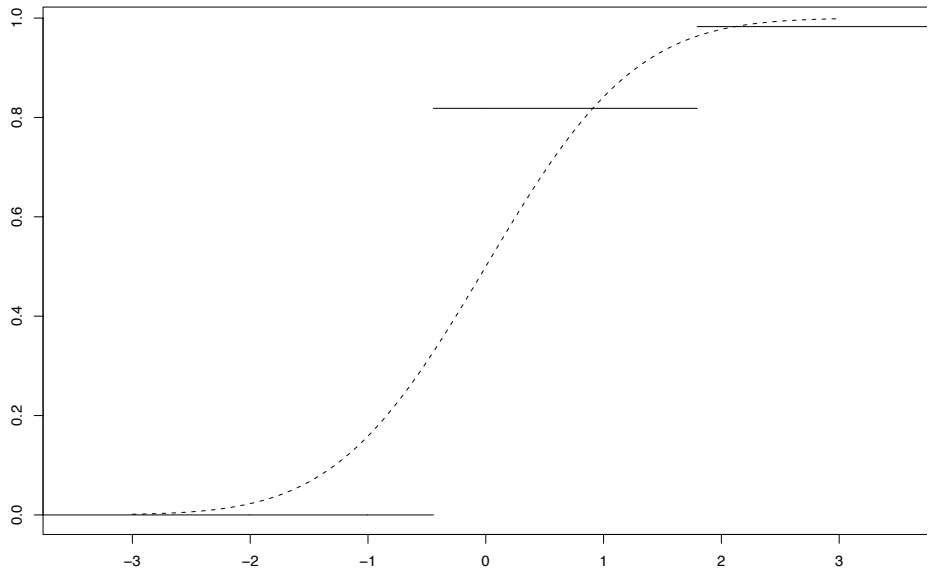


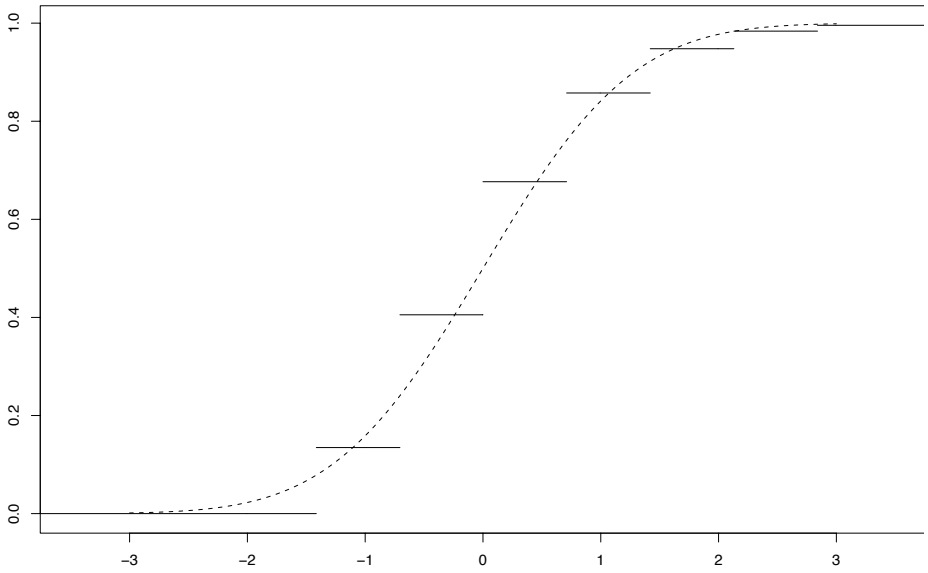


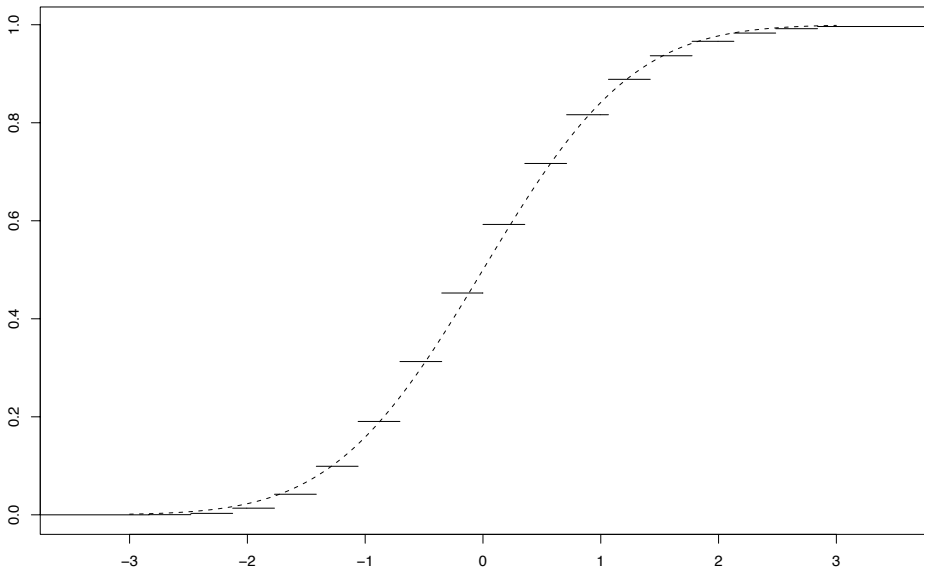
Voici enfin quatre graphiques correspondant à $p = 0,005$ et N successivement

égal à 10, 40, 400 et 1600.









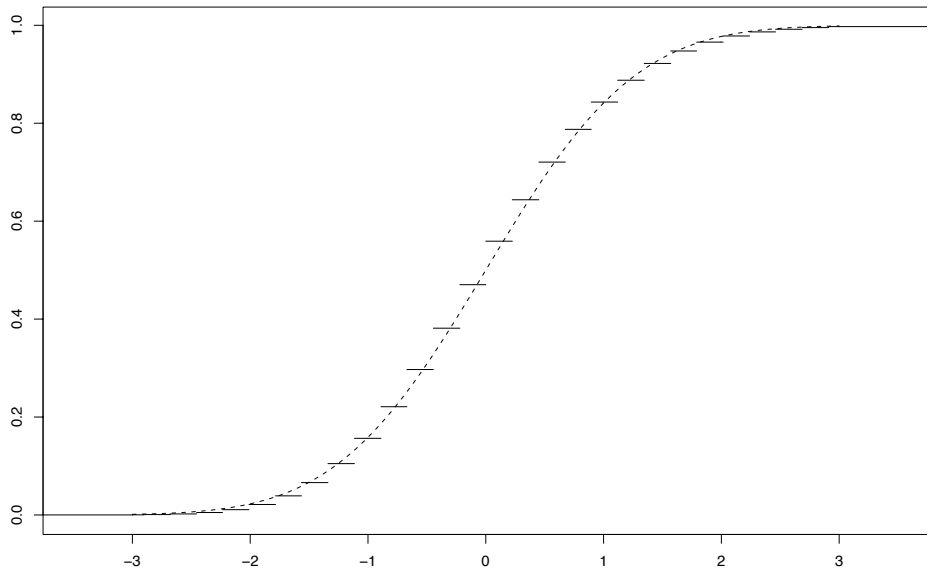
On constate visuellement la convergence énoncée par le théorème de la limite

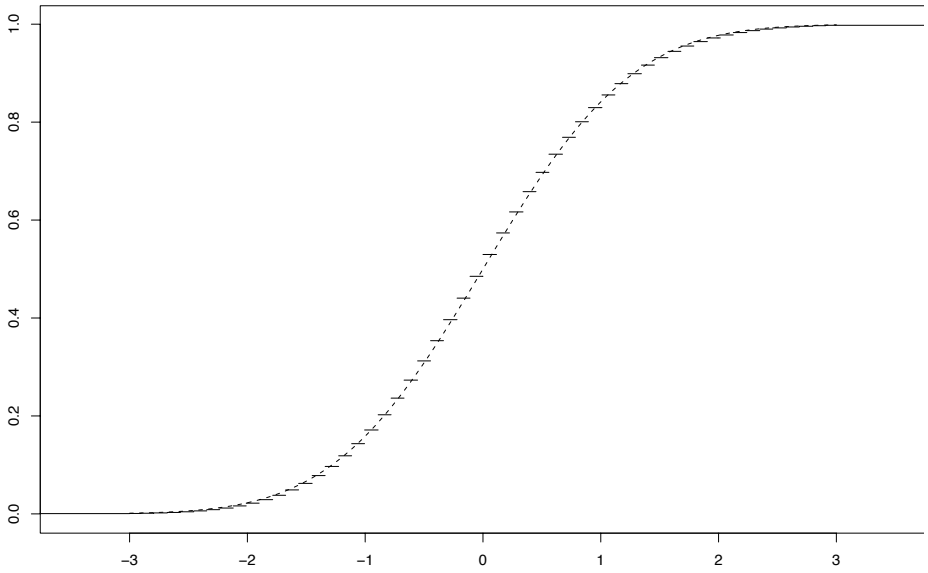
centrale, ainsi que le fait que la rapidité de celle-ci dépend manifestement de la loi de X.

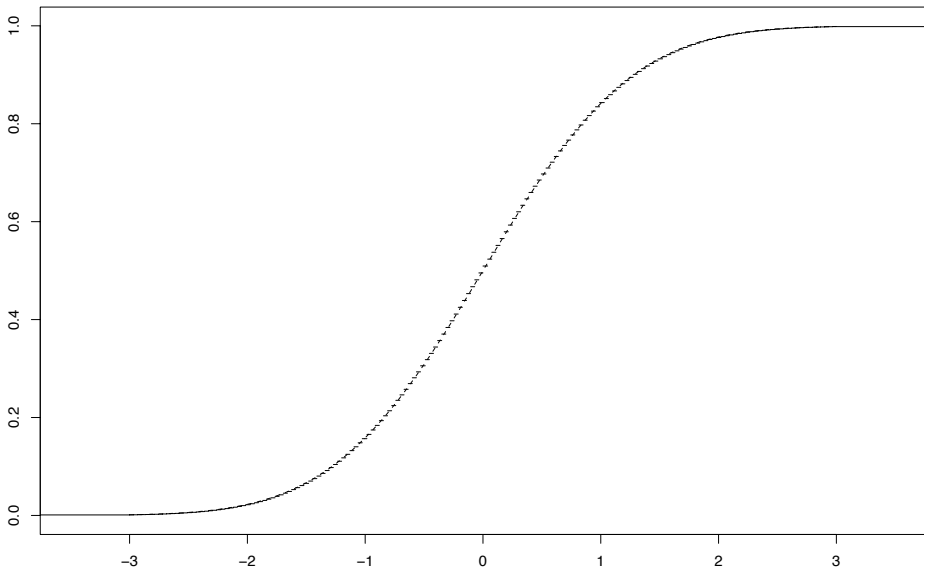
Variables aléatoires de loi de Poisson

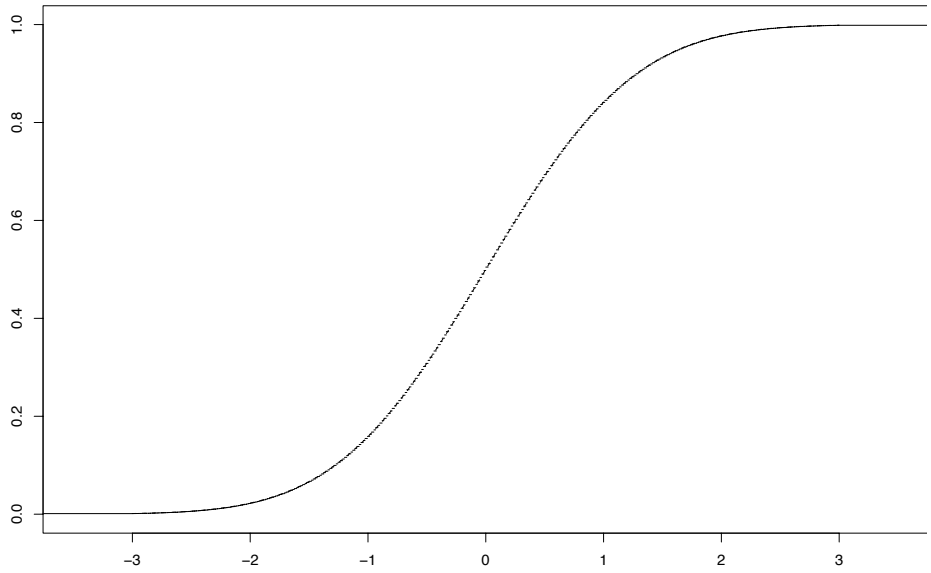
Dans cet exemple, on considère des variables aléatoires $X_1, \ldots, X_N$ indépendantes possédant toutes une loi de Poisson de paramètre λ. La loi de $X_1 + \cdots + X_N$ est connue dans ce cas : il s'agit d'une loi de Poisson de paramètre $N \times \lambda$ (voir l'exercice 117).

Les quatre graphiques suivants correspondent à $\lambda = 2$ et N successivement égal à 10, 40, 400 et 1600.

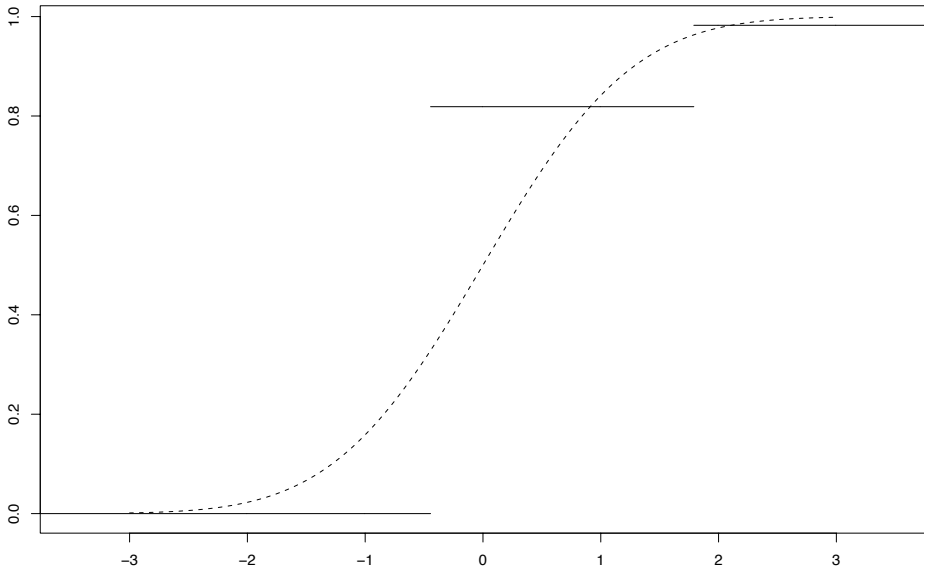


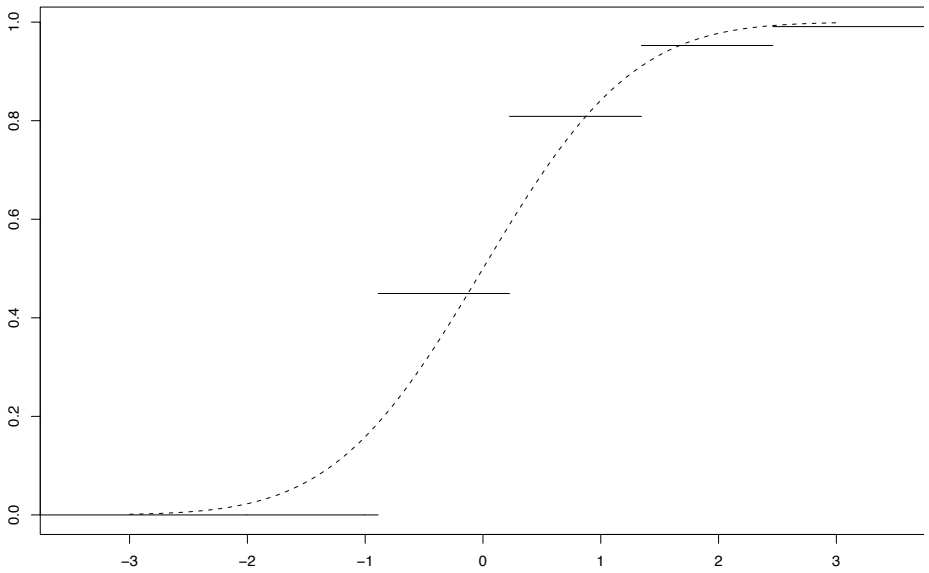

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

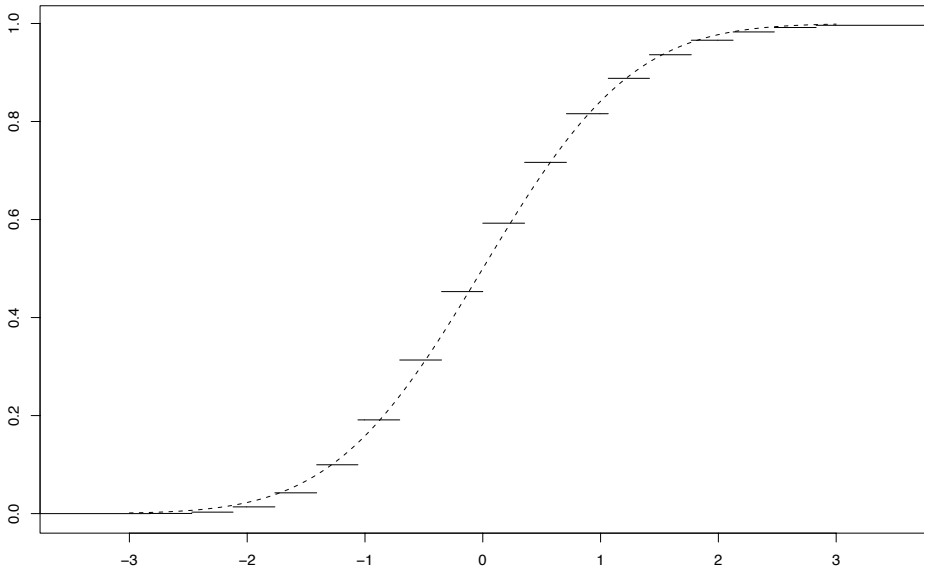

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

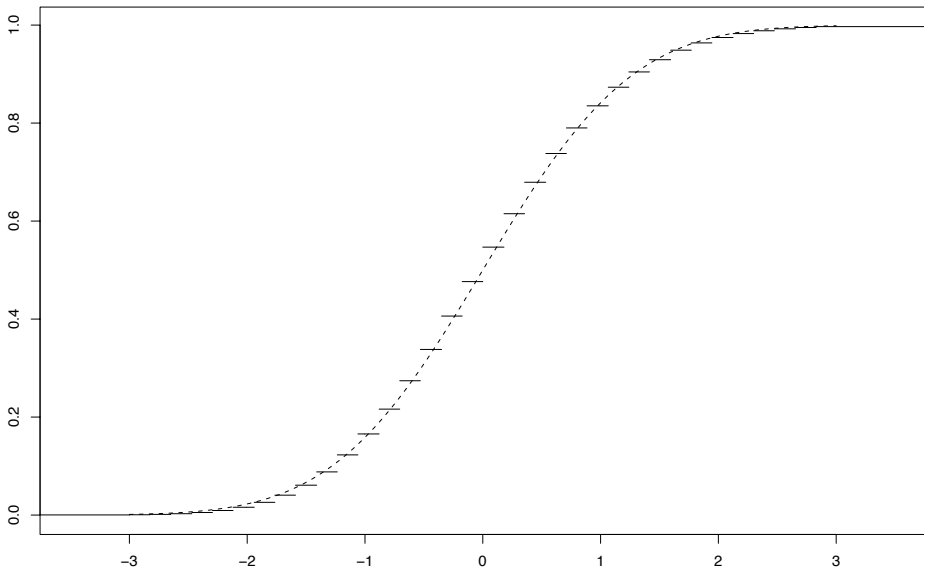


Les quatre graphiques suivants correspondent à $\lambda = 0,002$ et N successivement égal à 10, 40, 400 et 1600.


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

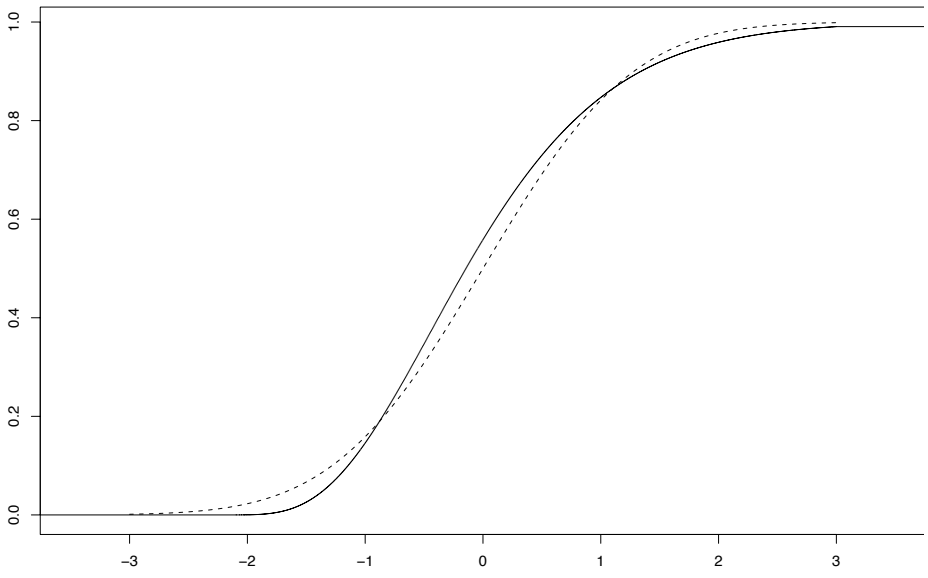

0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

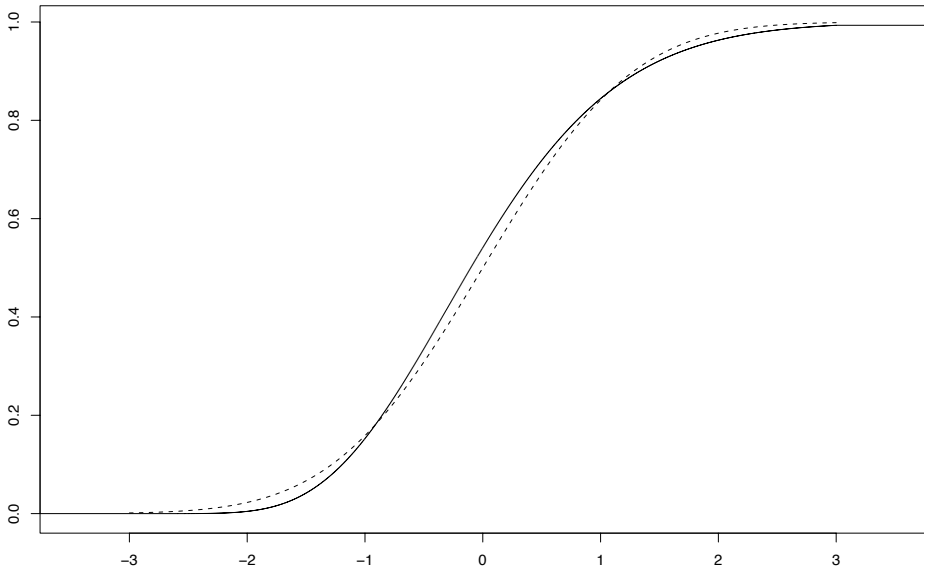

0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

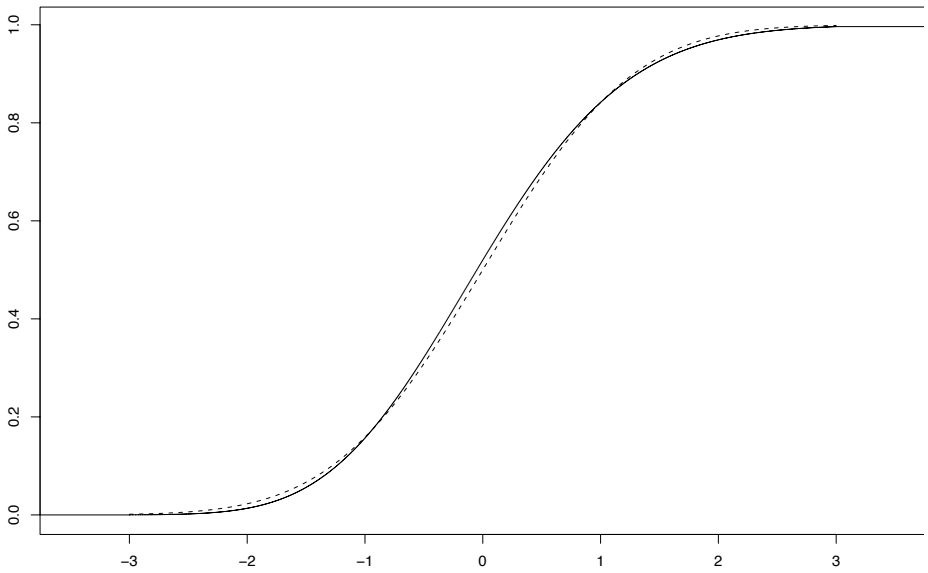
Variables aléatoires de loi exponentielle

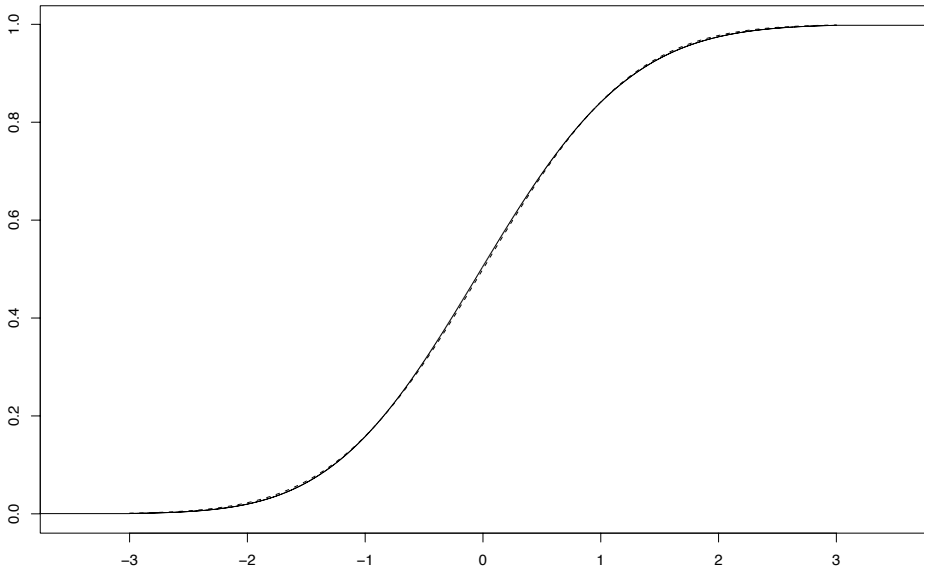
Dans cet exemple, on considère des variables aléatoires $X_1, \ldots, X_N$ indépendantes et possédant toutes la loi exponentielle de paramètre λ. La loi de $X_1 + \cdots + X_N$ est connue dans ce cas : il s'agit d'une loi dite Gamma (voir exercice 128) de paramètres $a = n$ et $s = \lambda$

Les quatre graphiques suivants illustrent le cas $\lambda = 3$ et N successivement égal à 5, 10, 40 et 400.




1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

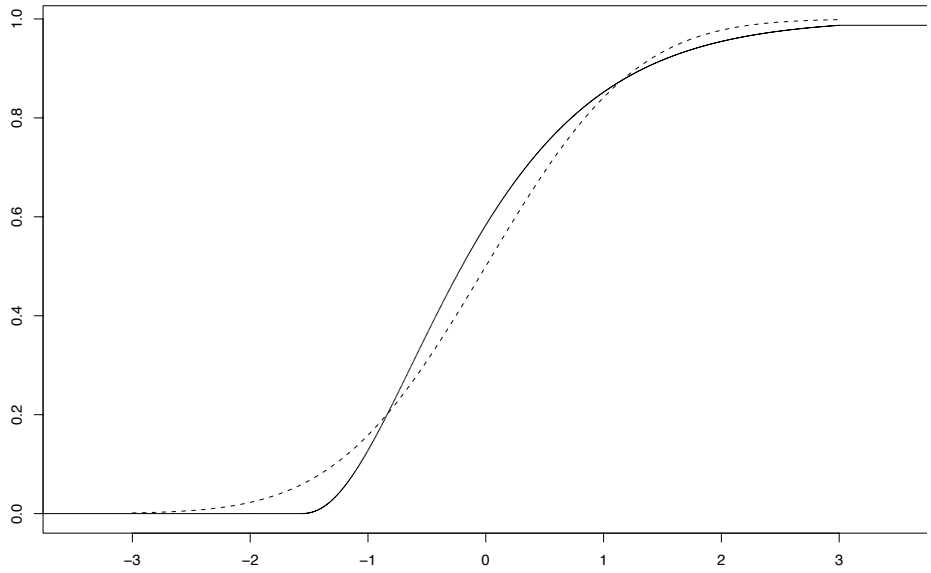

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

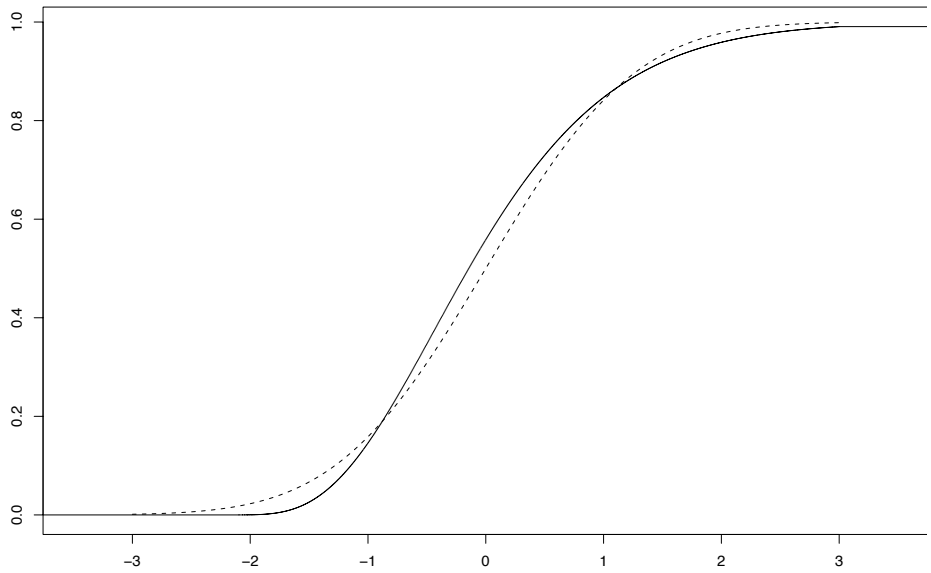


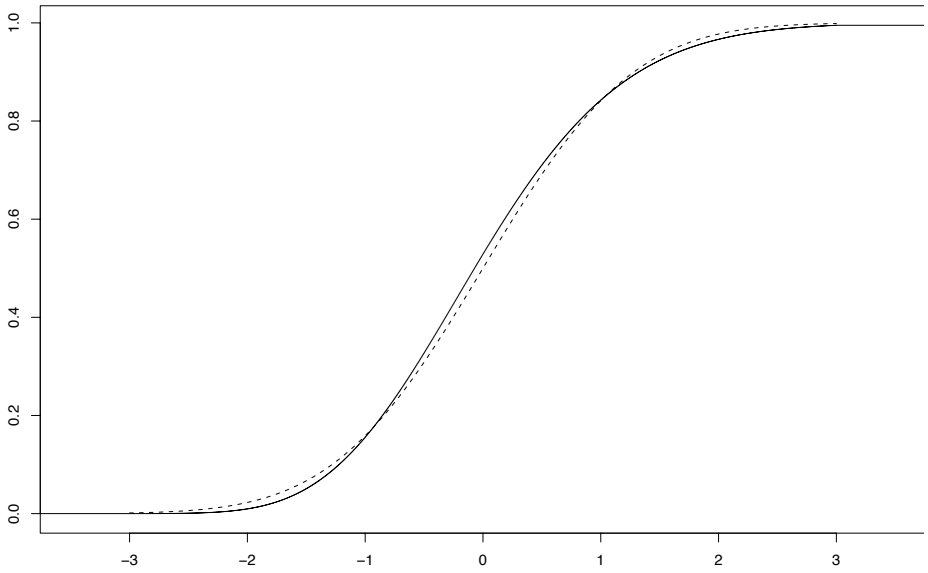
Carrés de variables aléatoires gaussiennes

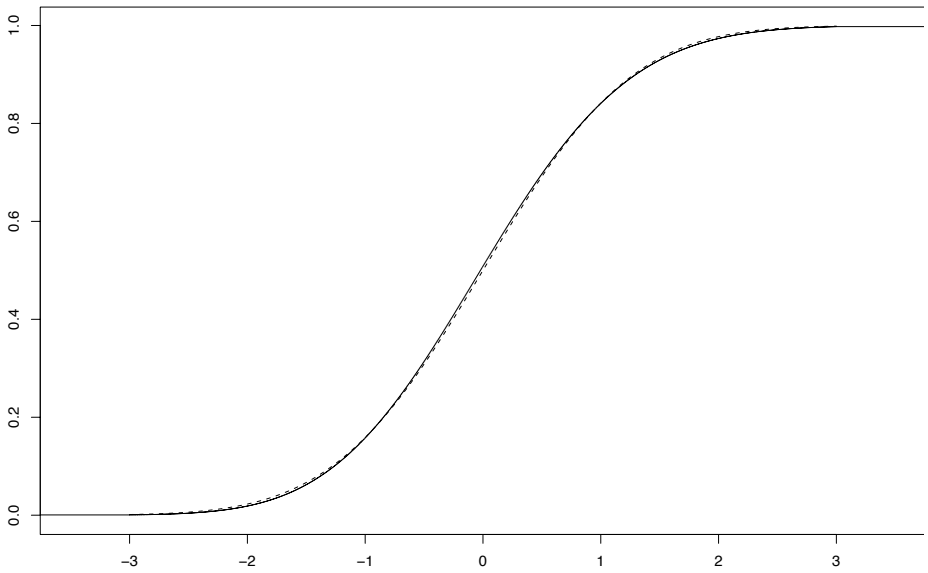
Dans cet exemple, on part de variables aléatoires $G_1, \ldots, G_N$ indépendantes et possédant toutes la loi gaussienne de paramètres $m = 0$ et $v = 1$, et l'on pose $X_1 = G_1^2, X_2 = G_2^2, \ldots, X_N = G_N^2$. La loi de $X_1 + \cdots + X_N$ est connue dans ce cas : il s'agit d'une loi du chi-deux à n degrés de liberté (voir l'exercice 129).

Les quatre graphiques suivants illustrent les cas où N est successivement égal à 5, 10, 40 et 400.


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

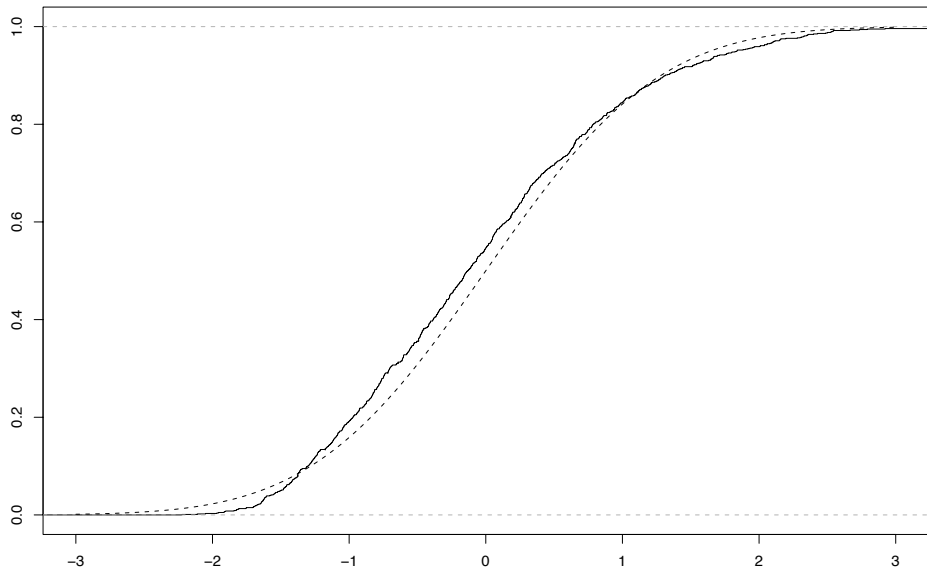
4.3.3 Des illustrations lorsque la loi de $X_1+\cdots+X_N$ n'est pas connue explicitement

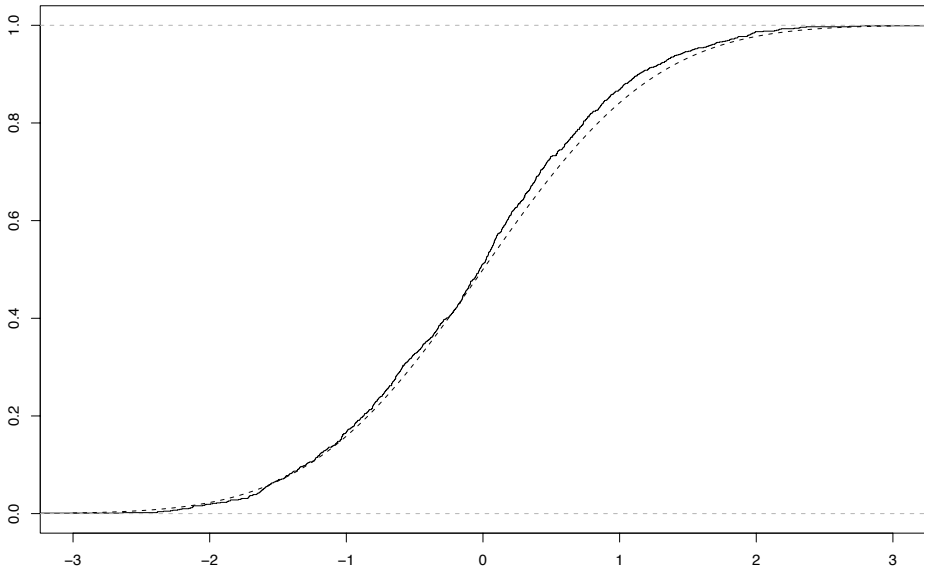
Lorsque la loi de $X_1+\cdots+X_N$ n'est pas connue de manière explicite, on peut par exemple y avoir accès par simulation, en effectuant un grand nombre de simulations consistant chacune à tirer N variables aléatoires indépendantes $X_1, \ldots, X_N$ de même loi que X. On peut alors comparer la loi empirique de $\frac{X_1+\cdots+X_N-N\times\mathbb{E}(X)}{\sqrt{N\times\mathbb{V}(X)}}$ à la loi limite gaussienne énoncée par le théorème de la limite centrale.

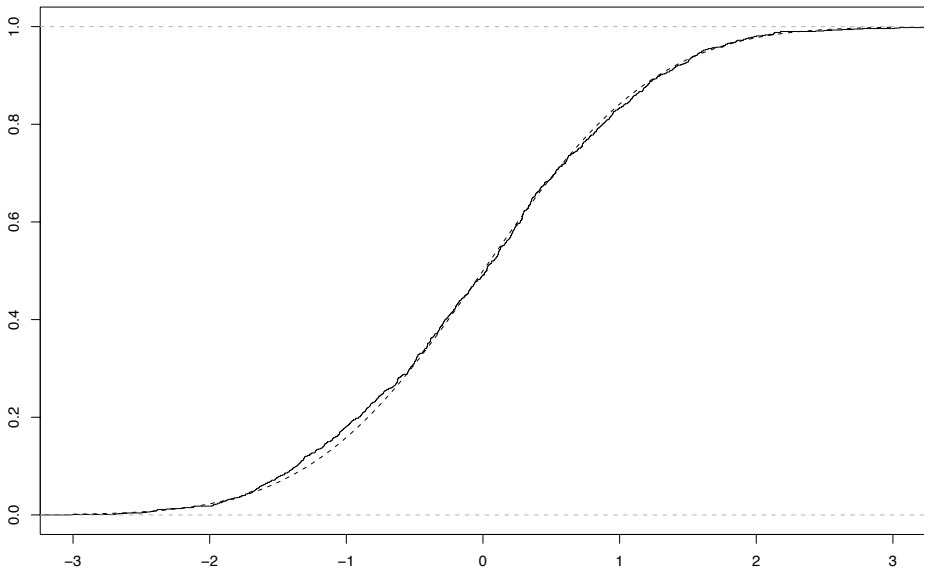
Puissances de variables aléatoires de loi uniforme sur $[0,1]$.

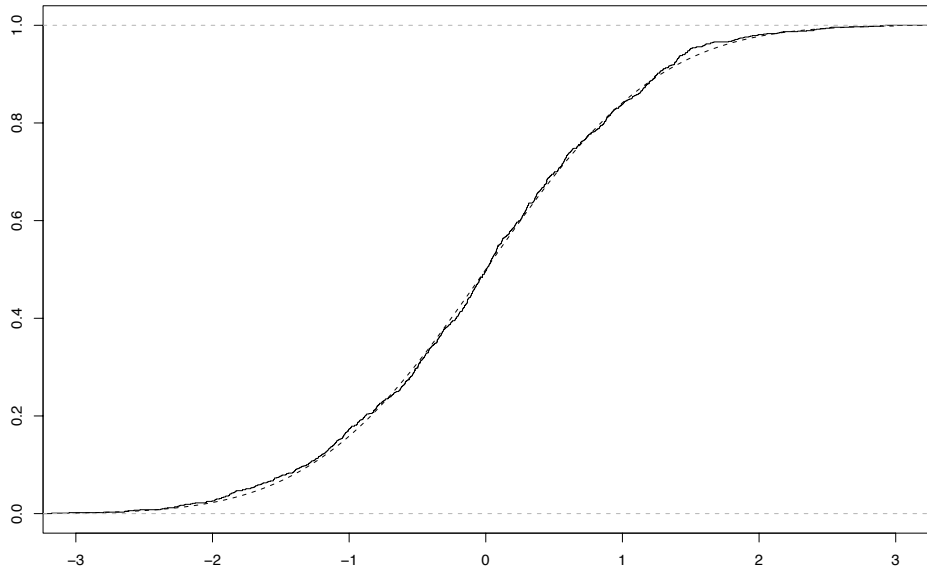
Dans cet exemple, on considère des variables aléatoires $U_1, \ldots, U_N$ indépendantes et possédant toutes la loi uniforme sur $[0,1]$, et l'on pose $X_1 = U_1^4, X_2 = U_2^4, \ldots, x_N = U_N^4$.

Les graphiques suivants illustrent l'approximation de la fonction de répartition de la loi gaussienne réalisée par la fonction de répartition associée à la loi empirique obtenue en effectuant 1000 simulations de γ_N, où N est successivement égal à 10, 40, 400 et 1600.




1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3



A priori, dans l'appréciation de la proximité entre les fonctions de répartitions empiriques présentées et la fonction de répartition de la loi gaussienne, il convient de séparer les deux sources d'écarts que peuvent être l'erreur d'approximation entre la loi théorique et la loi empirique de $\frac{X_1+\cdots+X_N-N\times\mathbb{E}(X)}{\sqrt{N\times\mathbb{V}(X)}}$, d'une part, et l'erreur d'approximation entre la loi théorique de $\frac{X_1+\cdots+X_N-N\times\mathbb{E}(X)}{\sqrt{N\times\mathbb{V}(X)}}$ et la gaussienne de paramètres $m=0$ et $v=\mathbb{V}(X)$ d'autre part. Nous vous renvoyons notamment à la discussion du chapitre précédent sur le théorème de Glivenko-Cantelli pour la question de l'approximation entre fonction de répartition empirique et théorique.

4.3.4 Deux erreurs fréquentes

Une première erreur fréquente dans l'utilisation du théorème de la limite centrale consiste à oublier le fait qu'il ne s'agit que d'un résultat asymptotique, et à affirmer que $\frac{S_N-\mathbb{E}(S_N)}{\sqrt{\mathbb{V}(S_N)}}$ suit **exactement** une loi gaussienne centrée réduite, alors que, sauf dans le cas où les X_i sont elles-mêmes de loi gaussienne, ceci n'est vrai qu'à une certaine approximation près, d'autant meilleure que N est grand. Remplacer directement dans un raisonnement ou un calcul la loi de $\frac{S_N-\mathbb{E}(S_N)}{\sqrt{\mathbb{V}(S_N)}}$ par une loi gaussienne sans tenir compte de l'approximation ainsi commise peut avoir des conséquences parfois délétères sur la validité de celui-ci, et il est indispensable de s'interroger sur la qualité de l'approximation fournie par l'utilisation du théorème de la limite cen-

trale lorsque l'on considère une valeur définie de N et non pas seulement une limite lorsque N tend vers l'infini. Ce point sera rediscuté dans les paragraphes «Attention à l'échelle» et «Quantification de la convergence».

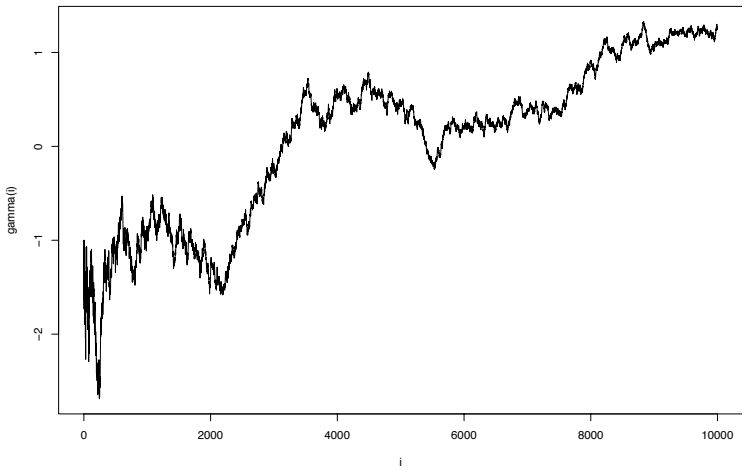
Une second erreur consiste à interpréter le résultat fourni par le théorème de la limite centrale comme signifiant que, avec une probabilité égale à 1, on a

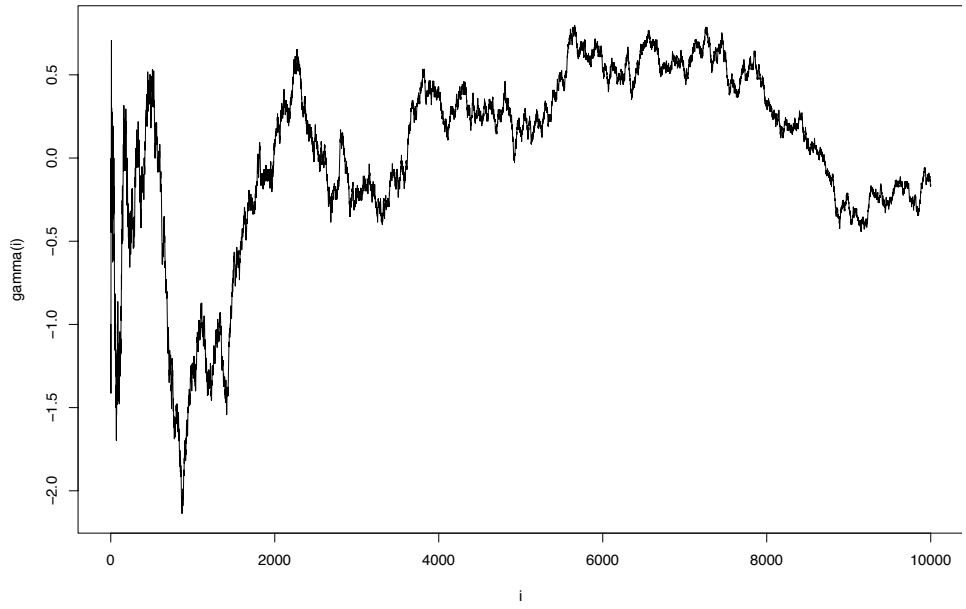
$$\lim_{N\to+\infty} \frac{S_N - \mathbb{E}(S_N)}{\sqrt{\mathbb{V}(S_N)}} = G, \tag{4.1}$$

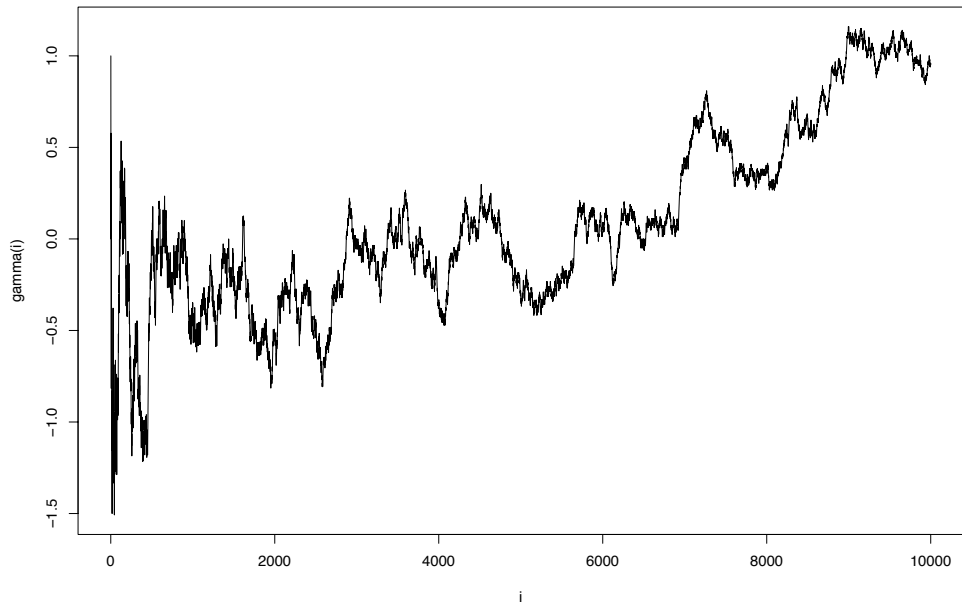
G étant une variable aléatoire de loi gaussienne centrée réduite. Une telle propriété n'est pas *a priori* en contradiction avec le théorème de la limite centrale, mais elle est néanmoins totalement fausse, car la suite de variables aléatoires $\frac{S_N-\mathbb{E}(S_N)}{\sqrt{\mathbb{V}(S_N)}}$ n'a pas de limite lorsque N tend vers l'infini. Avant d'expliquer ce point, vous pouver noter l'analogie existant entre les énoncés du théorème de la limite centrale et de la loi faible des grands nombres : ils énoncent tous les deux une propriété de la loi jointe de $(X_1, \ldots, X_N)$ lorsque N tend vers l'infini, tandis que l'énoncé de la loi forte des grands nombres et l'énoncé (4.1) – faux, répétons-le, dans notre contexte – se réfèrent à la loi de toute la suite infinie $(X_1, X_2, \ldots)$ (qui ne peut d'ailleurs pas vraiment être définie dans le cadre des espaces de probabilité discrets, comme nous l'avons déjà noté au chapitre précédent).

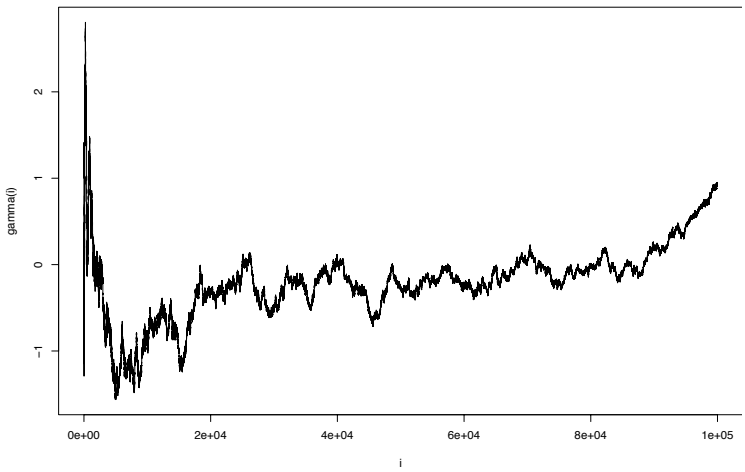
Tout d'abord, voici quelques simulations représentant $\gamma_i = \frac{S_i-\mathbb{E}(S_i)}{\sqrt{\mathbb{V}(S_i)}}$ en fonction de i. Le moins que l'on puisse dire est qu'elles ne suggèrent pas qu'il y ait convergence vers une valeur déterminée lorsque i tend vers l'infini. Les graphiques suivant correspondent à la situation où X suit une loi de Bernoulli de paramètre $1/2$.

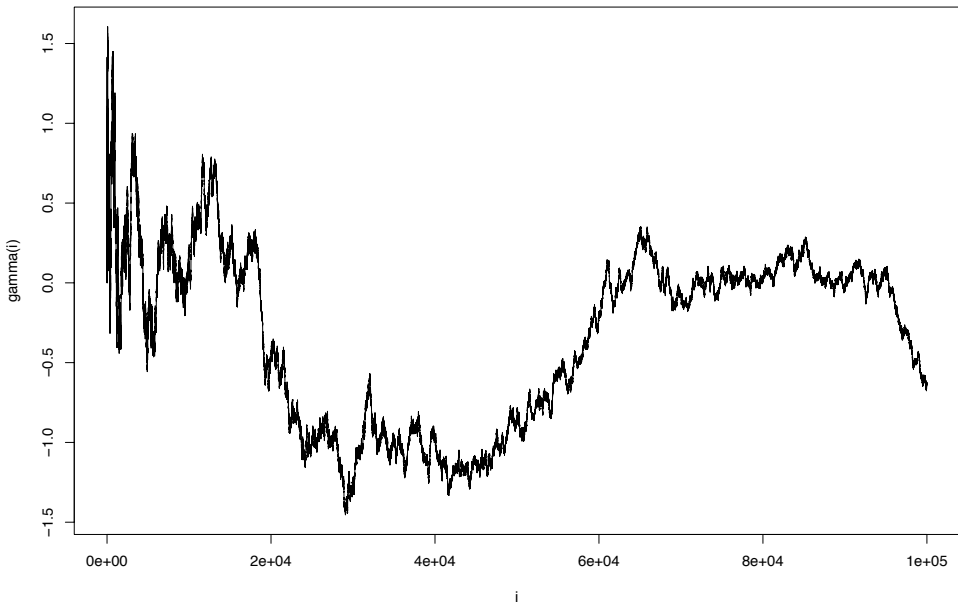
Les trois premiers montrent des simulations pour i variant de 1 à 10000, les trois suivants pour i variant de 1 à 100000.

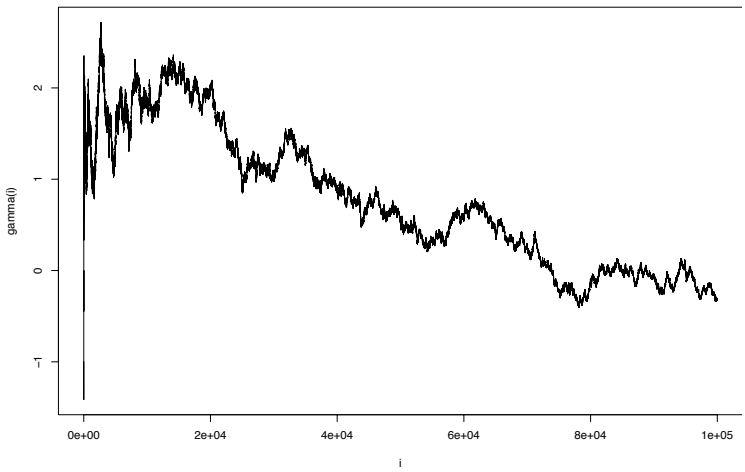



gamma(i)
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
2000
4000
6000
8000
10000
i


gamma(i)
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
2000
4000
6000
8000
10000
i


gamma(i)
2
1
0
−1
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
i


gamma(i)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
i



Il est également facile de comprendre en théorie pourquoi un énoncé tel que (4.1) ne peut pas être valable, au moyen du raisonnement suivant (que nous présentons de manière informelle, mais qu'il est possible de rendre parfaitement rigoureux). Si (4.1) était valable, on devrait avoir le fait que $\frac{S_m - m\times\mathbb{E}(S_m)}{\sqrt{m\times\mathbb{V}(X)}} \approx G$ pour toute valeur de m suffisament grande. Par conséquent, en choisissant N suffisament grand, on devrait avoir le fait que

$$\frac{S_{2N} - 2N\mathbb{E}(X)}{\sqrt{2N\mathbb{V}(X)}} \approx \frac{S_N - N\mathbb{E}(X)}{\sqrt{N\mathbb{V}(X)}}. \tag{4.2}$$

En écrivant le fait que $S_{2N} = S_{2N} - S_N + S_N$, on en déduirait, après un petit calcul (faites-le !) que

$$\frac{S_{2N} - S_N - N\mathbb{E}(X)}{\sqrt{2N\mathbb{V}(X)}} \approx (1 - 1/\sqrt{2})\frac{S_N - N\mathbb{E}(X)}{\sqrt{N\mathbb{V}(X)}}. \tag{4.3}$$

D'après le théorème de la limite centrale, la loi de $\frac{S_N - N\mathbb{E}(X)}{N\sqrt{\mathbb{V}(X)}}$ est approximativement une loi gaussienne centrée réduite. D'autre part, en notant que $S_{2N} - S_N$ est également une somme de N variables aléatoires indépendantes et de même loi que X, le théorème de la limite centrale entraîne que la loi de $\frac{S_{2N} - S_N - N\mathbb{E}(X)}{\sqrt{N\mathbb{V}(X)}}$ est aussi approximativement une loi gaussienne centrée réduite. Or $\frac{S_N - N\mathbb{E}(X)}{\sqrt{N\mathbb{V}(X)}}$ et $\frac{S_{2N} - S_N - N\mathbb{E}(X)}{\sqrt{N\mathbb{V}(X)}}$ sont deux variables aléatoires indépendantes (la première s'exprime en fonction de $X_{N+1}, \ldots, X_{2N}$, et la deuxième en fonction de $X_1, \ldots, X_N$). Ceci est clairement en contradiction avec la relation (4.3), qui exprime (à une approximation près) ces deux variables en fonction l'une de l'autre (deux variables aléatoires indépendantes ne peuvent s'exprimer en fonction l'une de l'autre, sauf à être constantes, ce qui n'est

pas le cas ici puisque les variables aléatoires considérées possèdent des lois approximativement gaussiennes).

4.3.5 Preuve du théorème de la limite centrale

Donner une preuve rigoureuse du théorème de la limite centrale dépasse le niveau mathématique de ce cours. Nous pouvons néanmoins donner une idée de preuve, quitte à admettre un certain nombre de points techniques.

Pas encore fait.

Nous vous renvoyons aux différents ouvrages d'introduction à la théorie mathématique des probabilités pour des preuves de ce résultat.

4.3.6 Le théorème de la limite centrale et la loi des grands nombres

La loi des grands nombres entraîne que, sous les hypothèses du théorème limite central, on peut écrire

$$\frac{X_1 + \cdots + X_N}{N} = \mathbb{E}(X) + \epsilon_N,$$

où le terme d'erreur ϵ_N est une quantité aléatoire, mais «petite», au sens où, lorsque N est grand, ϵ_N est «petit» avec une très forte probabilité Une question naturelle est alors de déterminer précisément quel est l'ordre de grandeur exact de ce «petit» terme correctif ϵ_N lorsque N est grand, et c'est justement ce que fait le théorème de la limite centrale. En reprenant les notations nous ayant servi à énoncer celui-ci, on constate que l'identité ci-dessus se réécrit :

$$\epsilon_N = \frac{S_N - N\mathbb{E}(X)}{N}.$$

On voit ainsi, en reprenant les notations de la partie précédentes, que

$$\epsilon_N = \sqrt{\frac{\mathbb{V}(X)}{N}}\gamma_N.$$

Le théorème de la limite centrale affirme donc que la variable aléatoire

$$\epsilon_N \times \sqrt{\frac{N}{\mathbb{V}(X)}}$$

est, lorsque N est grand, approximativement distribuée selon une loi gaussienne centrée réduite i.e. de paramètres $m = 0$ et $v = 1$. En un sens un peu vague, on peut affirmer que les valeurs de γ_N restent, lorsque N est grand, de l'ordre de l'unité : quoiqu'aléatoires, ces valeurs sont approximativement distribuées suivant une loi de probabilité qui ne dépend ni de N, ni de la loi de X. Toujours en restant assez vague, on peut donc affirmer que **l'ordre de grandeur des valeurs prises par ϵ_N lorsque**

N **est grand est** $\sqrt{\frac{\mathbb{V}(X)}{N}}$. Le terme en $\sqrt{N}$ au dénominateur quantifie l'influence de N sur la dispersion autour de zéro des valeurs que peut prendre ϵ_N lorsque N est grand. Toujours de manière vague, on peut donc dire que, vis-à-vis de N, la vitesse de convergence dans la loi des grands nombres est de l'ordre de $\frac{1}{\sqrt{N}}$. (Et l'on peut noter au passage qu'une telle vitesse de convergence est habituellement considérée comme médiocre dans un contexte numérique où l'on souhaite, autant que possible, avoir une vitesse de convergence au moins exponentielle en le nombre d'itérations effectuées). Le terme en $\sqrt{\mathbb{V}(X)}$ illustre, quant à lui, le fait que la convergence dans la loi des grands nombres a lieu d'autant plus lentement que les fluctuations de X, telles que mesurées par sa variance, sont importantes, ce que nous avions déjà observé empiriquement dans les simulations effectuées au chapitre précédent.

Insistons bien sur le fait que, même si nous avons utilisé l'identité $\epsilon_N = \sqrt{\frac{\mathbb{V}(X)}{N}}\gamma_N$ pour affirmer que l'ordre de grandeur des valeurs prises par ϵ_N sont de l'ordre de $\sqrt{\frac{\mathbb{V}(X)}{N}}$, les valeurs de γ_N sont aléatoires, et peuvent parfois s'éloigner considérablement de 1 (en valeur absolue), – si bien que ϵ_N peut être en réalité beaucoup plus grand, en valeur absolue, que $\sqrt{\frac{\mathbb{V}(X)}{N}}$ –, mais elles ne peuvent le faire qu'avec une faible probabilité, car la loi de γ_N est approximativement une loi gaussienne centrée réduite.

Par exemple, lorsque N est suffisament grand, la probabilité pour que γ_N soit compris entre -2 et 2 est d'environ 95%, d'environ 97,5 % pour que γ_N soit compris entre -3 et 3, d'environ 68% pour que γ_N soit compris entre -1 et 1.

Voici, pour fixer les idées, dix valeurs simulées (tronquées à 8 décimales) d'une variable aléatoire gaussienne centrée réduite, c'est-à-dire, dans notre contexte, dix valeurs que l'on pourrait obtenir pour γ_N lorsque N est grand : 0,15452532 ; 1,41194894 ; 0,08843478 ; -1,24517492 ; -0,07274697 ; 1,41970892 ; -0,60299238 ; -1,09537318 ; 0,70421432 ; 0,04185794.

Illustrons notre propos par un exemple simulé, en simulant, par exemple, 1000 variables aléatoires indépendantes $X_1, \ldots, X_{1000}$ de loi de Poisson de paramètre $\lambda = 2$. Rappelons que l'on a alors $\mathbb{E}(X) = \mathbb{V}(X) = \lambda = 2$.

– Expérience 1 : on trouve $S_{1000} = X_1 + \ldots + X_{1000} = 2042$. On a donc

$$\epsilon_{1000} = \frac{2042}{1000} - 2 = 0,042 \ , \ \gamma_{1000} \approx 0,94.$$

– Expérience 2 : on trouve cette fois $S_{1000} = X_1 + \ldots + X_{1000} = 1936$. On a donc

$$\epsilon_{1000} = \frac{1936}{1000} - 2 = -0,064 \ , \ \gamma_{1000} \approx -1,43.$$

– Expérience 3 : on trouve cette fois $X_1 + \ldots + X_{1000} = 2075$. On a donc

$$\epsilon_{1000} = \frac{2075}{1000} - 2 = 0,075 \ , \ \gamma_{1000} \approx 1,68.$$

Reprenons l'expérience, mais avec cette fois une somme de 100000 variables aléatoires au lieu de 1000.

– Expérience 4 : on trouve $S_{10000} = X_1 + \ldots + X_{100000} = 200972$. On a donc

$$\epsilon_{100000} = \frac{200972}{100000} - 2 = 0,00972 \ , \ \gamma_{10000} \approx 1,69.$$

– Expérience 5 : on trouve cette fois $X_1 + \ldots + X_{100000} = 200645$. On a donc

$$\epsilon_{100000} = \frac{200645}{100000} - 2 = 0,00645 \ , \ \gamma_{100000} \approx 0,46.$$

– Expérience 6 : on trouve cette fois $X_1 + \ldots + X_{100000} = 199551$. On a donc

$$\epsilon_{100000} = \frac{199551}{100000} - 2 = -0,00449 \ , \ \gamma_{100000} \approx -0,31.$$

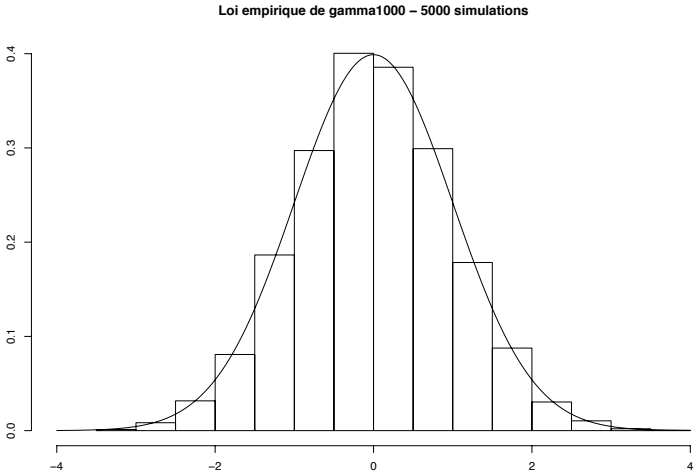
On constate que, dans ces six expériences, la valeur absolue de ϵ_N est relativement petite. Conformément à la loi des grands nombres, dans chacune des expériences

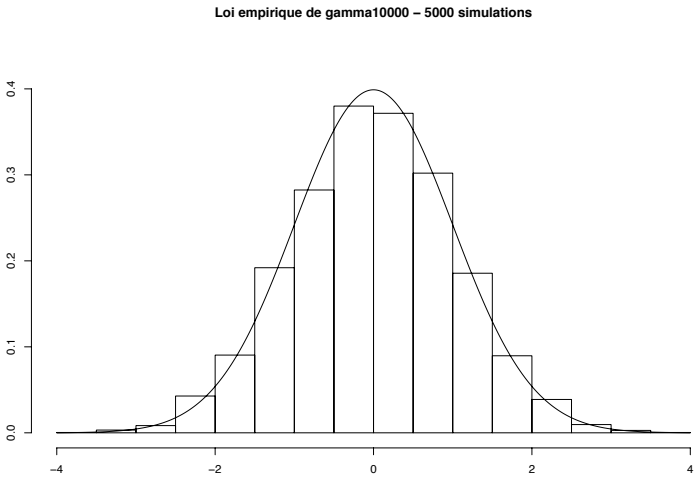
$$\frac{X_1 + \cdots + X_N}{N} \approx \mathbb{E}(X) = 2,$$

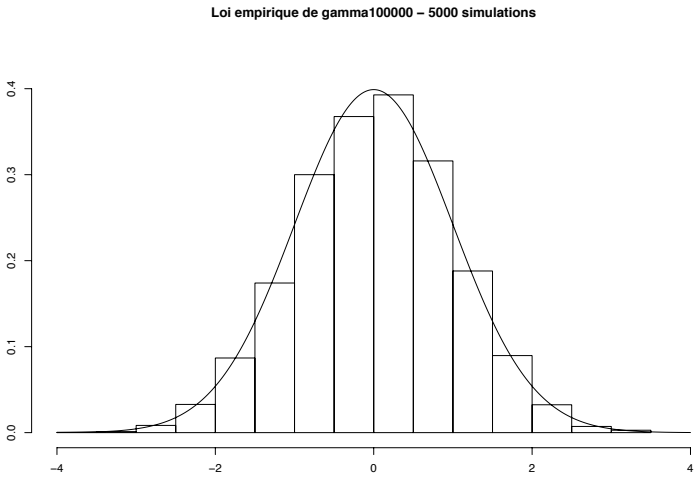
l'approximation étant d'autant meilleure que N est grand. Lorsque l'on observe à la loupe de grossissement $\sqrt{\frac{N}{\mathbb{V}(X)}}$ les petites valeurs prises par ϵ_N, on obtient des valeurs qui sont toujours de l'ordre de l'unité, que $N = 1000$ ou que $N = 100000$, mais qui diffèrent complètement d'une expérience à l'autre. Ceci reflète le caractère aléatoire de γ_N, c'est-à-dire de ϵ_N, observé avec le grossissement $\sqrt{\frac{N}{\mathbb{V}(X)}}$.

Si l'on effectue maintenant non plus deux ou trois expériences, mais un grand nombre d'expériences, par exemple 5000, on met en évidence non seulement le fait que les valeurs de γ_N, bien qu'aléatoires, restent de l'ordre de grandeur de l'unité lorsque N est grand, mais également le fait que celles-ci sont distribuées, au moins approximativement, selon une loi gaussienne centrée réduite.

Voici donc les histogrammes (gradués verticalement en densité) des valeurs simulées de γ_{1000}, γ_{10000} et γ_{100000}, avec 5000 simulations effectuées pour chaque histogramme, et en superposition la densité gaussienne $\phi_{0,1}$ (pour changer un peu des fonctions de répartition !).


Loi empirique de gamma1000 – 5000 simulations
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4


Loi empirique de gamma10000 – 5000 simulations
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4



En conclusion, le théorème de la limite centrale permet de décrire le comportement de l'écart entre S_N/N et $\mathbb{E}(X)$ dans la loi des grands nombres. L'ordre de grandeur (vis-à-vis de N et de la loi de X) de cet écart est, lorsque N est grand, $\sqrt{\frac{\mathbb{V}(X)}{N}}$. Observé à l'échelle de cet ordre de grandeur, cet écart est aléatoire et sa loi est approximativement une loi gaussienne centrée réduite.

4.3.7 Attention à l'échelle

Le théorème de la limite centrale décrit le comportement de la variable aléatoire $S_N = X_1 + \cdots + X_N$ ramenée à ce que l'on pourrait appeler l'échelle naturelle de ses fluctuations autour de son espérance (nous parlerons simplement par la suite d'«échelle naturelle des fluctuations», ou tout simplement d'«échelle naturelle») : S_N est **centrée**, de manière à pouvoir étudier les fluctuations qu'elle présente autour de son espérance, puis **réduite**, c'est-à-dire divisée par son écart-type de manière à ramener l'écart-type de ces fluctuations à 1, quelle que soit la valeur de N. On peut noter qu'une transformation affine appliquée à S_N, ou, ce qui revient au même, aux variables X_i, ne modifie pas la valeur de $\gamma_N = \frac{S_N - \mathbb{E}(S_N)}{\sqrt{N \times \mathbb{V}(X)}}$, qui demeure donc inchangée par ce type d'opérations.

Remarque 15 *En fait, il n'est pas évident a priori que l'opération consistant à centrer puis réduire S_N ramène celle-ci sur une échelle naturelle pour étudier sa loi, c'est-à-dire la transforme en une variable aléatoire dont la dispersion est de l'ordre de l'unité. Si l'écart-type de S_N donnait une indication complètement erronée de l'ordre de grandeur des valeurs de $S_N - \mathbb{E}(S_N)$, ou encore, si $\mathbb{E}(S_N)$ donnait une indication totalement erronée de la localisation des valeurs de S_N, – et nous savons,*

d'après le chapitre «Variables aléatoires» que ceci peut se produire dans certains cas – considérer $\frac{S_N - \mathbb{E}(S_N)}{\sqrt{N \times \mathbb{V}(X)}}$ *n'aurait en fait rien de pertinent. (Nous vous invitons de plus à consulter à ce sujet le paragraphe consacre à la non-robustesse du théorème de la limite centrale lorsque les variables aléatoires considérées ne possèdent plus de variance.) Une conséquence importante du théorème de la limite centrale est justement que ces deux indicateurs :* $\mathbb{E}(S_N)$ *et* $\mathbb{V}(S_N)$ *fournissent des indications fiables, au moins dans la limite où* N *tend vers l'infini, lorsque* S_N *est une somme de variables aléatoires indépendantes et de même loi (pour laquelle espérance et variance sont définies).*

Le théorème de la limite centrale affirme donc que, S_N, une fois ramenée à son échelle naturelle, suit approximativement une loi gaussienne centrée réduite lorsque N est grand. Le caractère gaussien de la loi d'une variable aléatoire étant conservé par changement d'échelle affine, on pourrait donc s'attendre à ce que S_N, observée sur n'importe quelle échelle, possède une loi approximativement gaussienne. Cependant, le fait que la loi de $\frac{S_N - \mathbb{E}(S_N)}{\sqrt{N \times \mathbb{V}(X)}}$ ne soit qu'**approximativement** gaussienne pour de grandes valeurs de N, et non pas exactement (même si cette approximation est d'autant meilleure que N est grand) limite fortement la portée de cette remarque.

Illustrons ceci dans la situation où X suit une loi de Bernoulli de paramètre $p = 1/2$, et donc où S_N suit la loi binomiale de paramètres N et $1/2$. Le théorème de la limite centrale nous permet de nous attendre à ce que, par exemple, la loi de $\gamma_{10000} = \frac{S_{10000} - 5000}{50}$ soit approximativement une loi gaussienne centrée réduite. Numériquement, on peut par exemple calculer que

$$\mathbb{P}^{\otimes 10000}\left[0,5 \leq \gamma_{10000} \leq 1,5\right] \approx 0,247 \text{ tandis que } \int_{0,5}^{1,5} \phi_{0,1}(u)du \approx 0,242,$$

ou

$$\mathbb{P}^{\otimes 10000}\left[-0,9 \leq \gamma_{10000} \leq -0,5\right] \approx 0,131 \text{ tandis que } \int_{-0,9}^{-0,5} \phi_{0,1}(u)du \approx 0,124,$$

ou encore

$$\mathbb{P}^{\otimes 10000}\left[\gamma_{10000} \leq -1,2\right] \approx 0,117 \text{ tandis que } \int_{-\infty}^{-1,2} \phi_{0,1}(u)du \approx 0,115.$$

En appliquant le changement d'échelle affine $x \mapsto 100x$ à γ_{10000}, on obtient que $100 \times \gamma_{10000}$ devrait possèder approximativement une loi gaussienne de paramètres $m = 0$ et $v = 100^2 = 10000$. Numériquement, on peut calculer que

$$\mathbb{P}^{\otimes 10000}\left[0,5 \leq 100 \times \gamma_{10000} \leq 1,5\right] = 0 \text{ tandis que } \int_{0,5}^{1,5} \phi_{0,100}(u)du \approx 0,004,$$

ou

$$\mathbb{P}^{\otimes 10000}\left[-0,9 \leq 100 \times \gamma_{10000} \leq -0,5\right] = 0 \text{ tandis que } \int_{-0,9}^{-0,5} \phi_{0,100}(u)du \approx 0,0016,$$

ou encore

$$\mathbb{P}^{\otimes 10000}\left[100 \times \gamma_{10000} \leq -1,2\right] \approx 0,496 \text{ tandis que } \int_{-\infty}^{-1,2} \phi_{0,100}(u)du \approx 0,495.$$

Les probabilités calculées pour la loi exacte de $100 \times \gamma_{10000}$ et pour une loi gaussienne de paramètres $m = 0$ et $v = 100^2 = 10000$ sont certes voisines, mais on constate que, dans les deux premiers cas, il serait catastrophique d'utiliser l'approximation par une loi gaussienne comme une estimation fiable de l'ordre de grandeur des probabilités auxquelles on s'intéresse : elles valent exactement 0, et non pas 0,004 ou 0,0016. Tout simplement, dans notre exemple, le changement d'échelle effectué fait apparaître le caractère discret de la variable aléatoire S_N, qui ne peut prendre que des valeurs entières. A une échelle où ce caractère discret est visible, il est clairement absurde d'assimiler la loi de S_N à une loi continue gaussienne. Si l'on en revient à la variable aléatoire γ_{10000}, les probabilités que nous venons de calculer se réécrivent : $\mathbb{P}^{\otimes 10000}\left[0,005 \leq \gamma_{10000} \leq 0,015\right]$, $\mathbb{P}^{\otimes 10000}\left[-0,009 \leq \gamma_{10000} \leq -0,005\right]$, et enfin $\mathbb{P}^{\otimes 10000}\left[\gamma_{10000} \leq -0,012\right]$. Les deux premières probabilités correspondent à des intervalles de très petite taille, et font donc intervenir la loi de γ_{10000} à une échelle trop fine pour que l'approximation par une loi gaussienne centrée réduite produise des résultats fiables (par exemple au sens d'une faible erreur relative sur le calcul des probabilités de la forme $\mathbb{P}^{\otimes N}(\gamma_N \in I)$).

Bien entendu, le théorème de la limite centrale, qui est un résultat asymptotique, énonce le fait que, pour tout intervalle $I \subset \mathbb{R}$, on a

$$\lim_{N \to +\infty} \mathbb{P}^{\otimes N}\left[\gamma_N \in I\right] = \int_I \phi_{0,1}(u)du,$$

sans faire aucune différence entre un intervalle tel que $[0,5; 1,5]$ et $[0,005; 0,0015]$. Le calcul ci-dessus suggère simplement que, si l'on cherche à extrapoler à des valeurs grandes mais finies de N le résultat asymptotique valable lorsque $N \to +\infty$ énoncé par le théorème de la limite centrale, l'approximation par une loi gaussienne peut nécessiter, pour être fiable, des valeurs de N plus importantes pour des intervalles de petite taille que pour des intervalles dont la largeur est de l'ordre de l'unité. On peut chercher à rendre compte de ce fait dans un cadre asymptotique en étudiant le comportement lorsque N tend vers l'infini de probabilités de la forme $\mathbb{P}^{\otimes N}\left[\gamma_N \in I_N\right]$, où la taille de l'intervalle I_N peut donc varier avec N. Pour systématiser l'exemple précédent, dans lequel X suit une loi de Bernoulli de paramètre $p = 1/2$, on voit facilement que l'intervalle $I_N = [0,2 \times (N/2)^{-1/2}; 0,4 \times (N/2)^{-1/2}]$ est tel que $\mathbb{P}^{\otimes N}\left[\gamma_N \in I_N\right] = 0$ du fait que S_N ne peut prendre que des valeurs

entières, tandis que $\int_{I_N} \phi_{0,1}(u)du = \Theta(N^{-1/2})$. La meilleure manière d'aborder correctement cette question est d'étudier de manière quantitative la convergence vers la gaussienne dans le théorème de la limite centrale, ce qui fait l'objet du paragraphe suivant.

4.3.8 Quantification de la convergence dans le théorème de la limite centrale

Tel que nous l'avons formulé, le théorème de la limite centrale est un résultat asymptotique, valable seulement dans la limite où N tend vers l'infini. Cependant, en pratique, on l'utilise avec des valeurs finies et supposées «grandes» de N, pour approcher la loi de S_N par une loi gaussienne. Les illustrations précédentes montrent que, pour une valeur donnée de N, la précision de l'approximation par une loi gaussienne dépend de la loi de X, et, de même que pour la loi des grands nombres, il n'existe pas de nombre N «grand» dans l'absolu, et qui permettrait de garantir une certaine qualité d'approximation pour toutes les lois de X possibles. Pour préciser cette question, commençons par énoncer un résultat : **sous les mêmes hypothèses que celles de l'énoncé du théorème de la limite centrale, et en ajoutant le fait que l'espérance $\mathbb{E}\left(|X-\mathbb{E}(X)|^3\right)$ existe, on a l'inégalité suivante** (appelée inégalité de Berry-Esséen, nous renvoyons par exemple à l'ouvrage de Shyriaev cité dans la bibliographie pour une preuve de ce résultat).

$$\forall x \in \mathbb{R},\ \left|\mathbb{P}^{\otimes N}\left[\frac{S_N - \mathbb{E}(S_N)}{\sqrt{\mathbb{V}(S_N)}} \leq x\right] - \int_{-\infty}^{x} \phi_{0,1}(u)du\right| \leq \frac{0,8}{\sqrt{N}} \times \frac{\mathbb{E}\left(|X-\mathbb{E}(X)|^3\right)}{\mathbb{V}(X)^{3/2}}. \tag{4.4}$$

Au prix d'une (petite) hypothèse supplémentaire par rapport à l'énoncé du théorème de la limite centrale, on peut donc obtenir une borne non-asymptotique et explicite concernant l'approximation de la loi de γ_N par une loi gaussienne.

Une borne sur les probabilités du type $\mathbb{P}^{\otimes N}(\gamma_N \in]a,b])$ s'en déduit immédiatement, en écrivant que $\mathbb{P}^{\otimes N}(\gamma_N \in]a,b]) = \mathbb{P}^{\otimes N}(\gamma_N \leq b) - \mathbb{P}^{\otimes N}(\gamma_N \leq a)$.

Remarquons tout de suite que la borne supérieure (4.4) ne fait pas intervenir x, et qu'elle constitue donc une borne sur le «pire» écart, c'est-à-dire

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}^{\otimes N}(\gamma_N \leq x) - \int_{-\infty}^{x} \phi_{0,1}(u)du\right|.$$

Clairement, N étant fixé, on peut toujours, en choisissant x suffisament petit, faire en sorte que les probabilités $\mathbb{P}^{\otimes N}(\gamma_N \leq x)$ et $\int_{-\infty}^{x} \phi_{0,1}(u)du$ soit aussi petites qu'on le souhaite, et donc que la borne (4.4), qui ne fait pas intervenir x, soit arbitrairement imprécise. Pour n'être pas précise pour toutes les valeurs de x, cette borne fournit néanmoins une borne supérieure quant à l'ordre de grandeur des valeurs

de N nécessaires à l'obtention d'une approximation donnée de $\int_{-\infty}^{x} \phi_{0,1}(u)du$ par $\mathbb{P}^{\otimes N}(\gamma_N \leq x)$, et, qui plus est, pour des valeurs de x demeurant de l'ordre de l'unité, cette borne fournit en général le bon ordre de grandeur. De nombreuses améliorations de cette borne existent, incluant des développements asymptotiques précis de $\left|\mathbb{P}^{\otimes N}(\gamma_N < x) - \int_{-\infty}^{x} \phi_{0,1}(u)du\right|$ par rapport à x et N, mais il s'agit de questions trop avancées pour que nous les abordions ici. Nous vous renvoyons, par exemple, à l'ouvrage de Feller (Tome 2) cité dans la bibliographie, pour en apprendre davantage à ce sujet. Nous vons invitons également à traiter l'exercice 168.

4.3.9 Robustesse du théorème de la limite centrale

Nous avons énoncé le théorème de la limite centrale dans un cadre identique à celui de la loi faible des grands nombres du chapitre précédent, en ajoutant l'hypothèse que la loi de X devait posséder une variance.

Comme dans le cas de la loi des grands nombres, le théorème de la limite centrale reste valable à condition que les variables aléatoires X_i restent approximativement indépendante, et que l'ordre de grandeur des valeurs qu'elles peuvent prendre soit suffisament bien contrôlé (ce qui correspond dans notre énoncé à l'existence de l'espérance et de la variance de X).

Nous ne tenterons pas plus que dans le chapitre sur la loi des grands nombres de formuler précisément ce que peuvent être ces conditions, mais nous nous contenterons d'illustrer sur quelques exemples la robustesse, ou la non-robustesse, du résultat énoncé par le théorème de la limite centrale.

L'hypothèse de répétition indépendante

Nous reprendrons ici les trois (plus une normale) pièces obstinées du chapitre précédent. Pour éviter de pénibles renvois au chapitre précédent, et quitte à nous répéter, nous reprenons en détail les descriptions de chacune des pièces considérées.

Une pièce normale

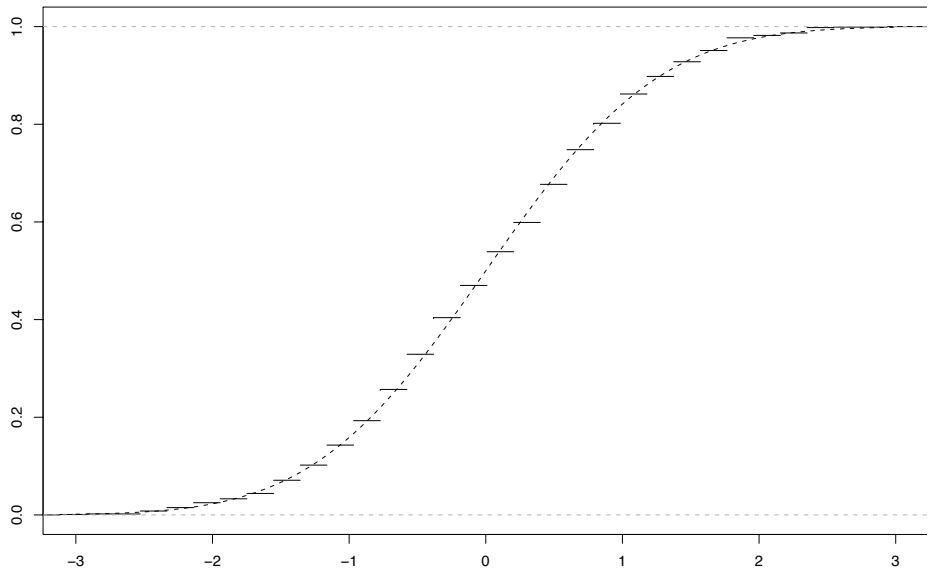
A titre de comparaison pour la suite, voici ce que l'on obtient avec une pièce «normale», dont les lancers sont indépendants et suivent une loi de Bernoulli de paramètre $1/2$: $\mathbb{P}(X_i = \mathrm{F}) = 1/2$. La nombre total de F obtenu au cours des N premiers lancers peut s'écrire

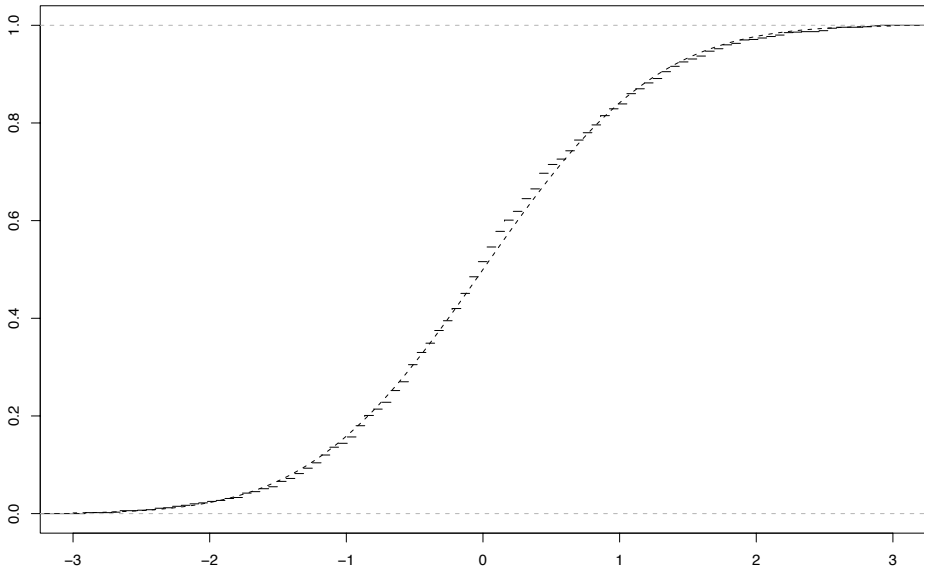
$$S_N = f(X_1) + \cdots + f(X_N),$$

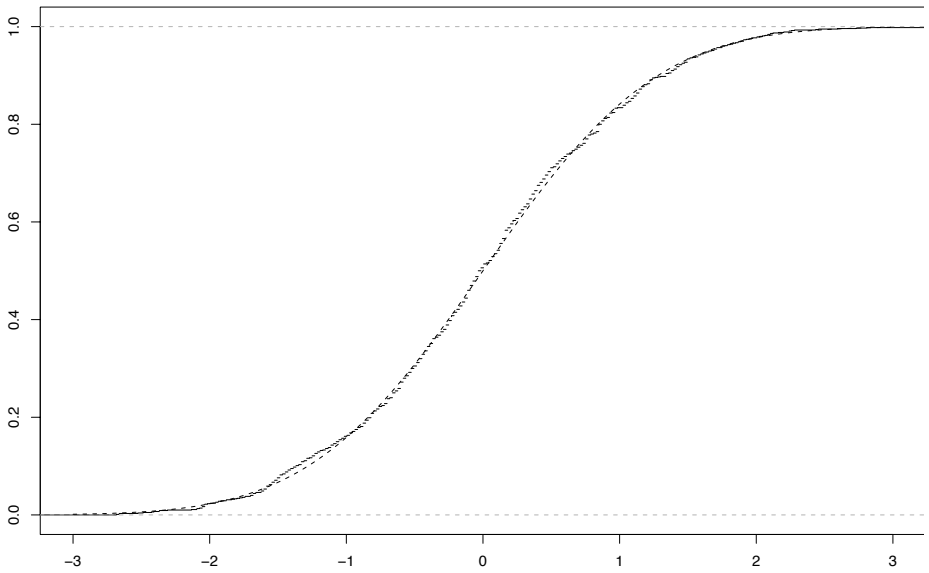
en posant $f(\mathrm{F}) = 1$ et $f(\mathrm{P}) = 0$.

Chacun des trois graphiques suivant correspond à la loi empirique obtenue avec 1000 simulations de N lancers, avec successivement $N = 100$, $N = 1000$, $N =$

10000, centrée puis réduite, avec en superposition la fonction de répartition de la loi gaussienne standard.




1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

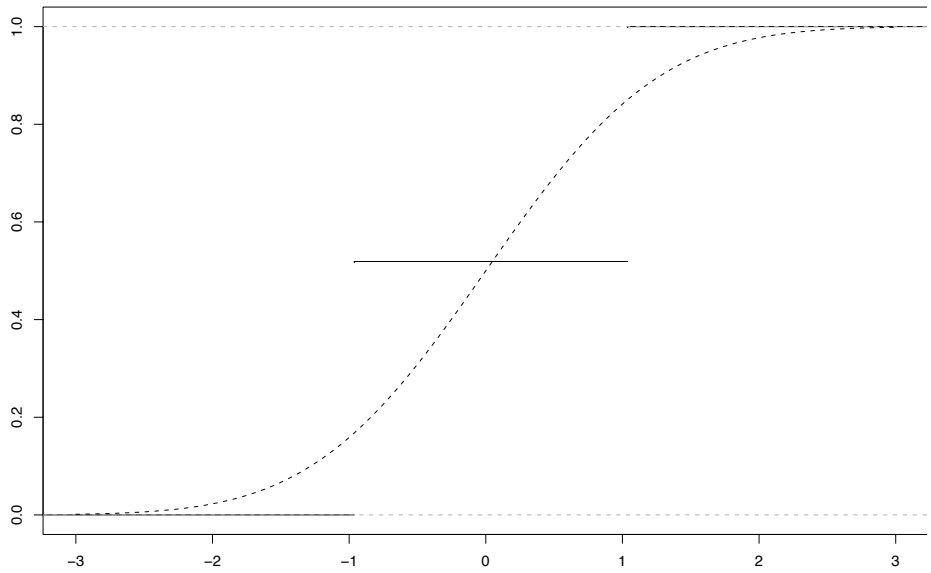

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

Une pièce de monnaie obstinée

On suppose que l'on a affaire à une pièce de monnaie obstinée possédant la propriété suivante : une fois la pièce sortie de sa boîte, le premier lancer est effectivement aléatoire, pouvant donner pile ou face avec une probabilité égale à $1/2$, mais, au cours de tous les lancers suivants, la pièce se souvient du résultat de son premier lancer, et s'arrange toujours pour retomber exactement du même côté. Si l'on note $X_1, \ldots, X_N$ les résultats des N premiers lancers de la pièce, on se trouve ici dans un cas extrême de non-indépendance : la valeur de X_{i+1} est toujours égale à la valeur de X_i. En revanche, les lancers sont tous décrits individuellement par une loi de Bernoulli de paramètre $1/2$: $\mathbb{P}(X_i = \mathrm{P}) = \mathbb{P}(X_i = \mathrm{F}) = 1/2$. La nombre total de F obtenu au cours des N premiers lancers peut s'écrire

$$S_N = f(X_1) + \cdots + f(X_N),$$

en posant $f(\mathrm{F}) = 1$ et $f(\mathrm{P}) = 0$. Pas plus que la loi des grands nombres, le théorème de la limite centrale ne peut s'appliquer à S_N, qui prend la valeur 0 avec probabilité $1/2$, et N avec probabilité $1/2$. Par exemple, le graphique ci-dessous représente la fonction de répartition de la loi empirique de l'échantillon obtenu en effectuant 1000 simulations de S_{10000}, centrée et réduite. En pointillés, la fonction de répartition de la loi gaussienne standard.



Une pièce moins obstinée

Considérons à présent une autre pièce obstinée, conservant également la mémoire de ses lancers passés, mais de manière moins stricte que la précédente. Spécifiquement, une fois la pièce sortie de sa boîte, le premier lancer effectué est aléatoire, donnant pile ou face avec une probabilité égale à $1/2$. Ensuite, pour tout $i \geq 1$, étant donnés les résultats des i premier lancers, le $i+1-$ème lancer se déroule de la façon suivante : la pièce reproduit le résultat du $i-$ème lancer avec une probabilité p fixée, et produit le résultat inverse avec une probabilité $1-p$. Si p est égal à 1, la pièce se comporte comme celle étudiée dans le paragraphe précédent. Si $p = 0$, on obtient une alternance stricte de P et de F. Nous supposerons dans la suite que $0 < p < 1$. Si $1/2 < p < 1$, la pièce conserve sa tendance à redonner lors d'un lancer la valeur obtenue à l'issue du lancer précédent, mais de manière moins stricte que dans le cas précédent. Si $p = 1/2$, on retrouve une suite de répétitions indépendantes de lancers de Bernoulli. Enfin, si $0 < p < 1/2$, la pièce a tendance à produire lors d'un lancer un résultat inversé par rapport au lancer précédent.

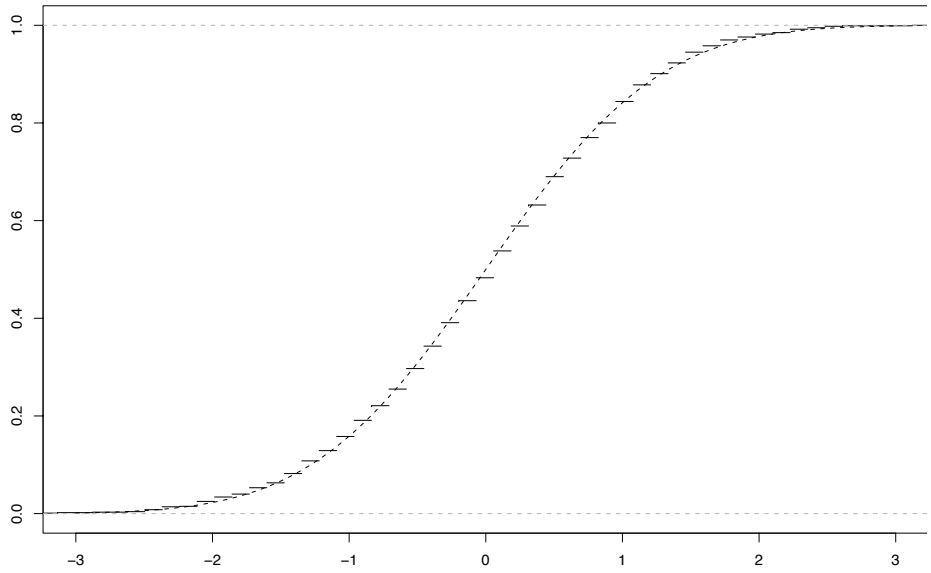
Les variables aléatoires $X_1, \ldots, X_N$ ne sont donc pas indépendantes lorsque $p \neq 1/2$, puisque le résultat obtenu au cours d'un lancer affecte la loi de probabilité attachée au lancer suivant. Cependant, il semble clair que, si k est suffisament grand, le résultat du lancer $i+k$ doit être relativement indépendant du résultat du lancer i, car la mémoire du résultat du lancer i est de plus en plus brouillée au fur et à mesure que les lancers se répètent (voir à ce sujet l'exercice 65). Il existe donc une certaine forme d'indépendance approchée entre les résultats suffisament éloignés dans la séquence des lancers.

On peut par ailleurs facilement vérifier que, pris de manière individuelle, les lancers sont décrits par une loi de Bernoulli de paramètre $1/2$: $\mathbb{P}(X_i = \mathrm{P}) = \mathbb{P}(X_i = \mathrm{F}) = 1/2$.

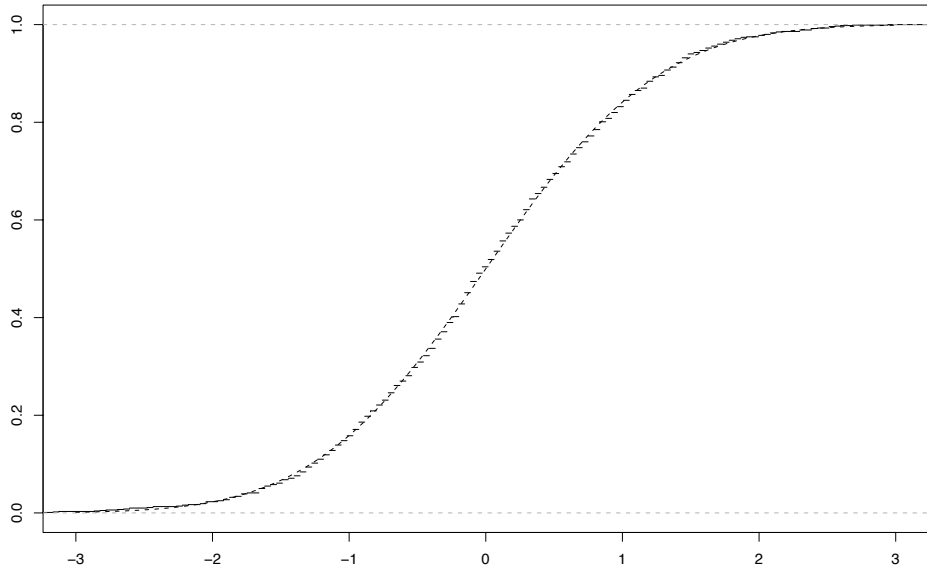
Il se trouve que, quelle que soit la valeur de $p \in]0,1[$ dans ce modèle, le théorème de la limite centrale est effectivement vérifié par le nombre de P obtenu après N lancers, que nous notons S_N comme dans le paragraphe précédent.

Pour l'illustrer, nous présentons des graphiques représentant – pour diverses valeurs de p et de N – la fonction de répartition de la loi empirique de l'échantillon obtenu en effectuant 1000 simulations de S_N, centrée et réduite. En pointillés, la fonction de répartition de la loi gaussienne standard.

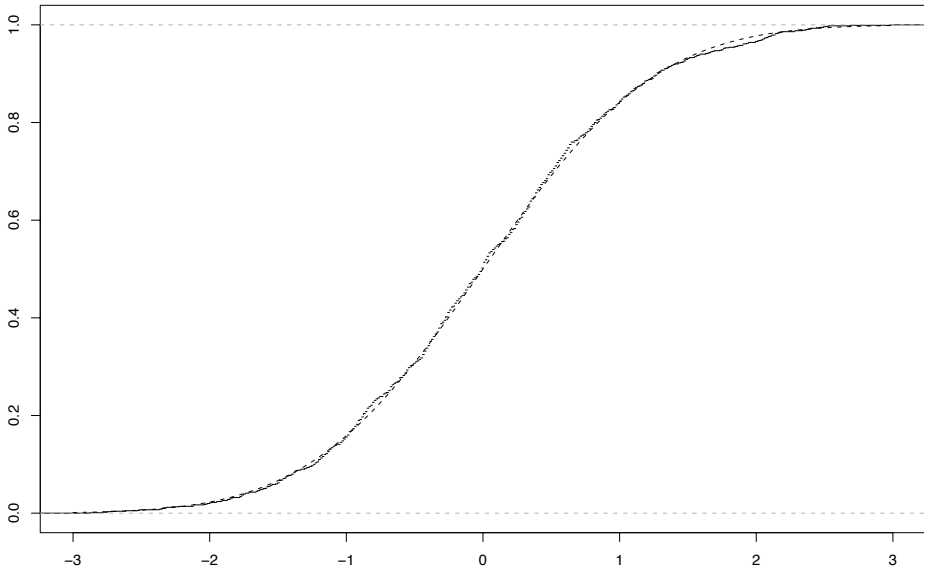
Prenons par exemple $p = 0,7$ et $N = 100$.



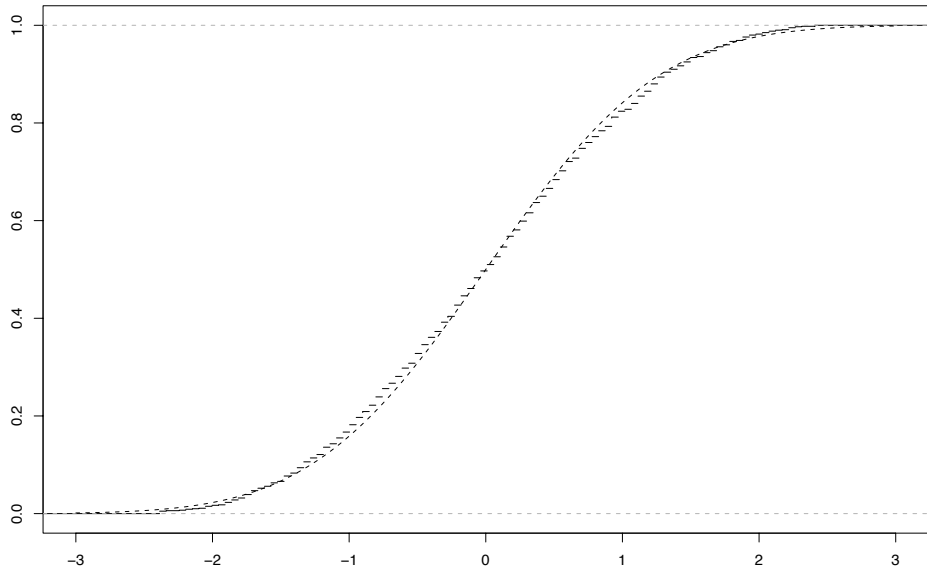
Puis $p = 0,7$ et $N = 1000$.



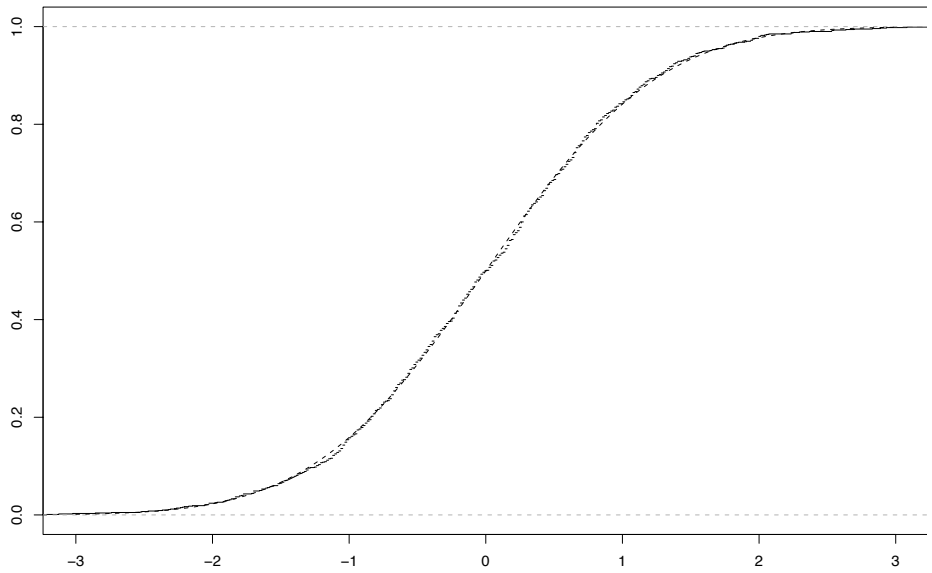
Et enfin $p = 0,7$ et $N = 10000$.



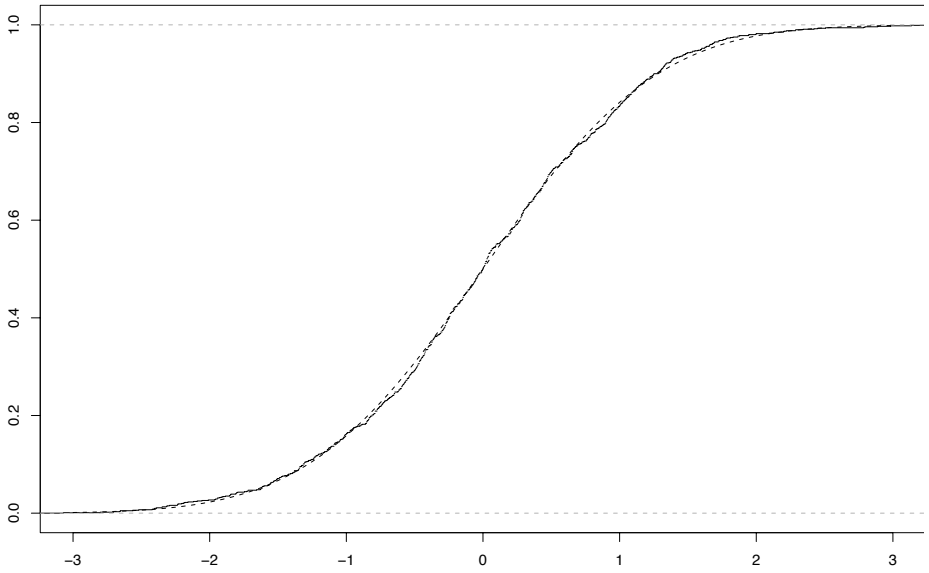
Avec maintenant $p = 0,95$ et $N = 100$.



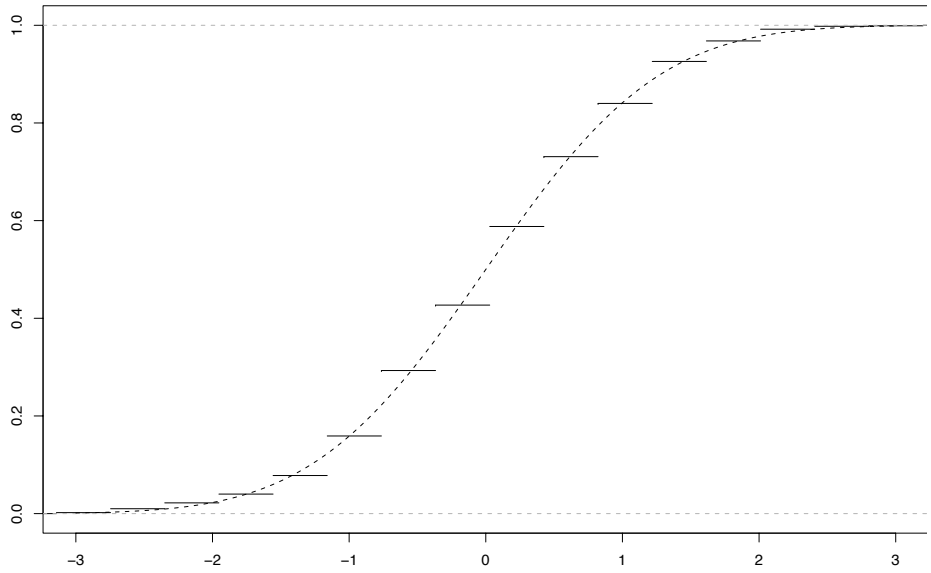
Puis $p = 0,95$ et $N = 1000$.



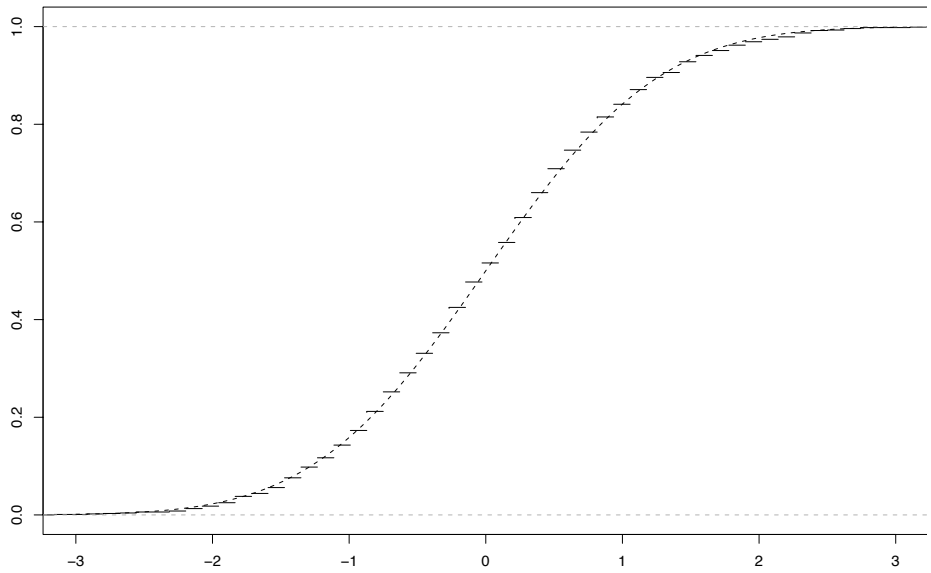
Et enfin $p = 0,95$ et $N = 10000$.



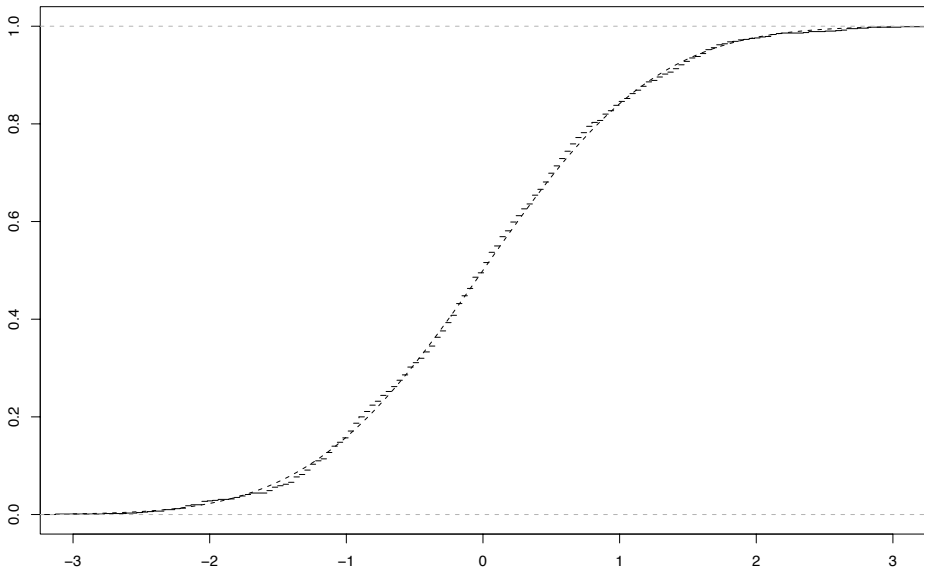
A présent $p = 0,2$ et $N = 100$.



Puis $p = 0,2$ et $N = 1000$.



Et enfin $p = 0,2$ et $N = 10000$.



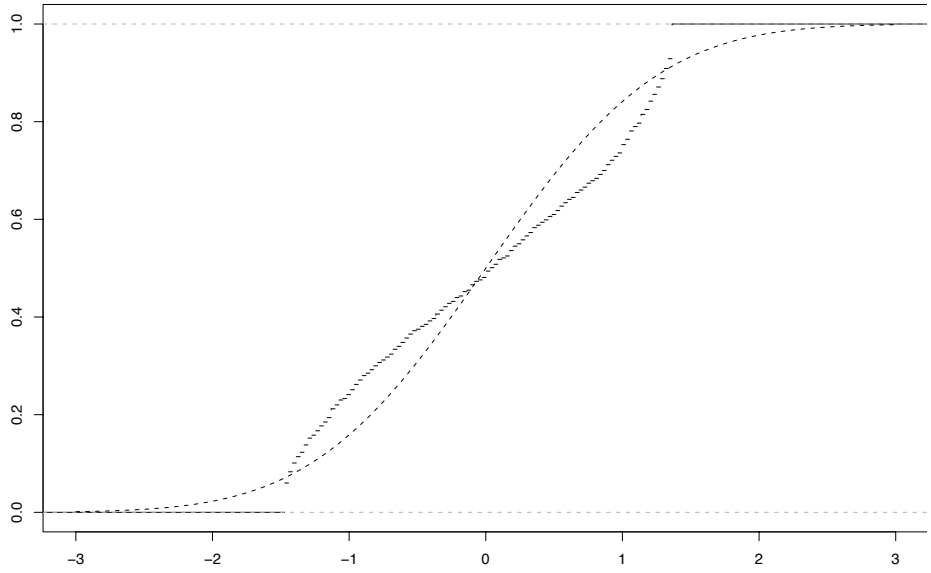
Encore une pièce obstinée

Considérons à présent une pièce dont les lancers successifs sont reliées de la manière suivante. Une fois la pièce sortie de sa boîte, le premier lancer effectué est aléatoire, donnant pile ou face avec une probabilité égale à $1/2$. Ensuite, pour tout $i \geq 1$, étant donnés les résultats des i premier lancers, le $i+1-$ème lancer se déroule de la façon suivante : la pièce accorde à P une probabilité proportionnelle à $1 + \Delta N_i(P)$ et à F une probabilité proportionnelle à $1 + \Delta N_i(F)$, $N_i(P)$ et $N_i(F)$ désignant respectivement les nombres de fois où P et F sont sortis au cours des i premiers lancers, et $\Delta > 0$ désignant un paramètre. En d'autres termes, chaque nouveau lancer donnant lieu à un F renforce d'une valeur égale à Δ le poids accordé à F dans les futurs lancers, et il en va de même pour P. On peut vérifier facilement que, pris de manière individuelle, les lancers sont décrits par une loi de Bernoulli de paramètre $1/2$: $\mathbb{P}(X_i = \mathrm{P}) = \mathbb{P}(X_i = \mathrm{F}) = 1/2$.

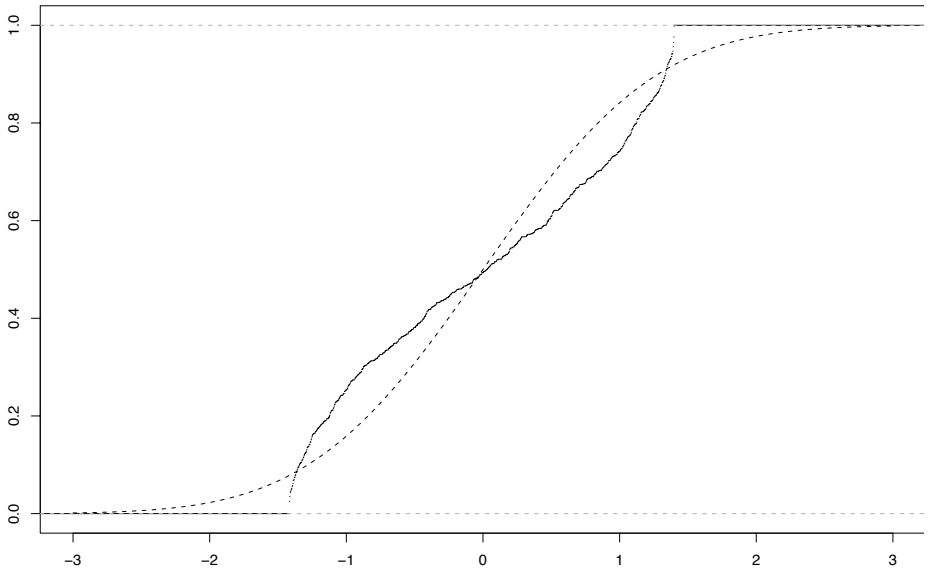
Comme précédemment nous présentons des graphiques représentant – pour diverses valeurs de Δ et de N – la fonction de répartition de la loi empirique de l'échantillon obtenu en effectuant 1000 simulations de S_N, centrée et réduite. En pointillés, la fonction de répartition de la loi gaussienne standard.

Voici quelques exemples de simulations effectuées avec $\Delta = 2$.

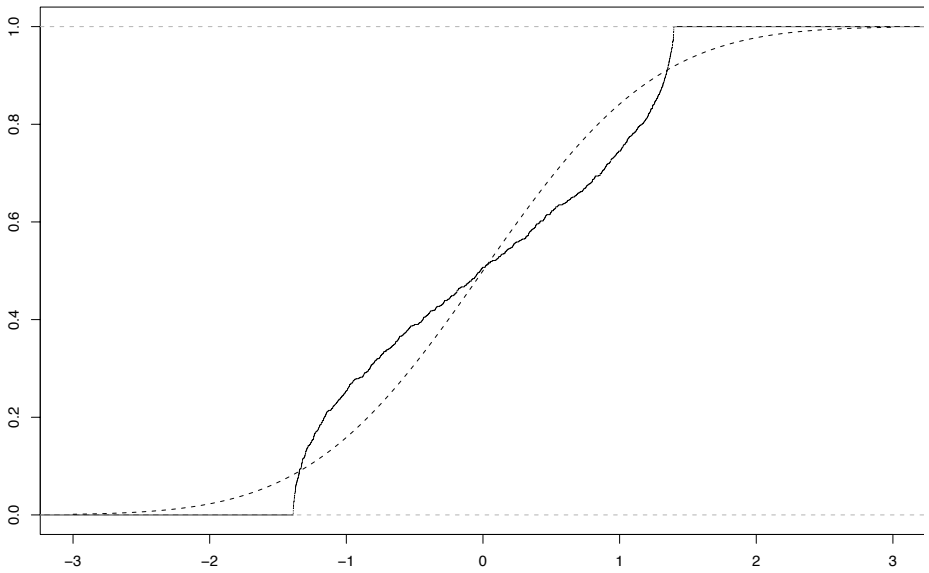
Pour $N = 100$.



Pour $N = 1000$.

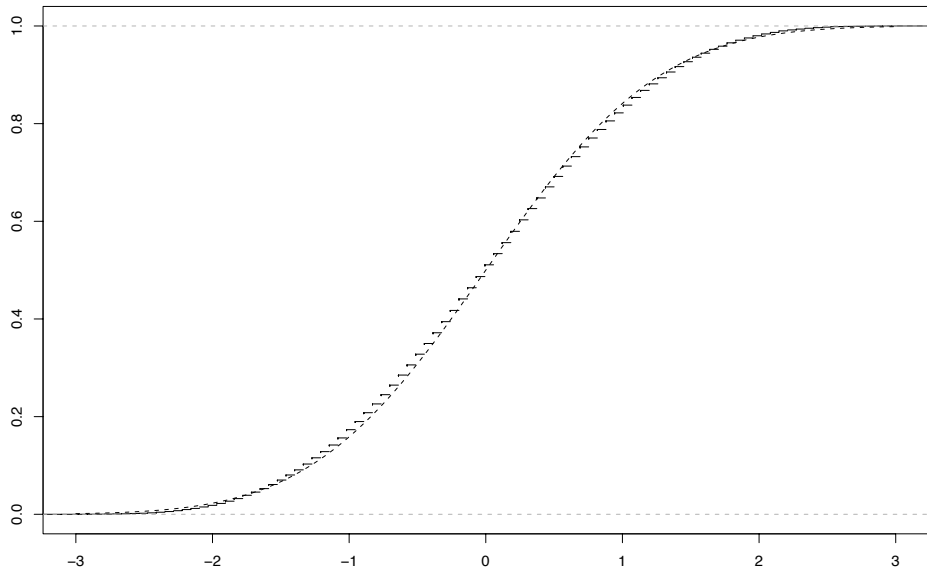


Pour $N = 10000$.

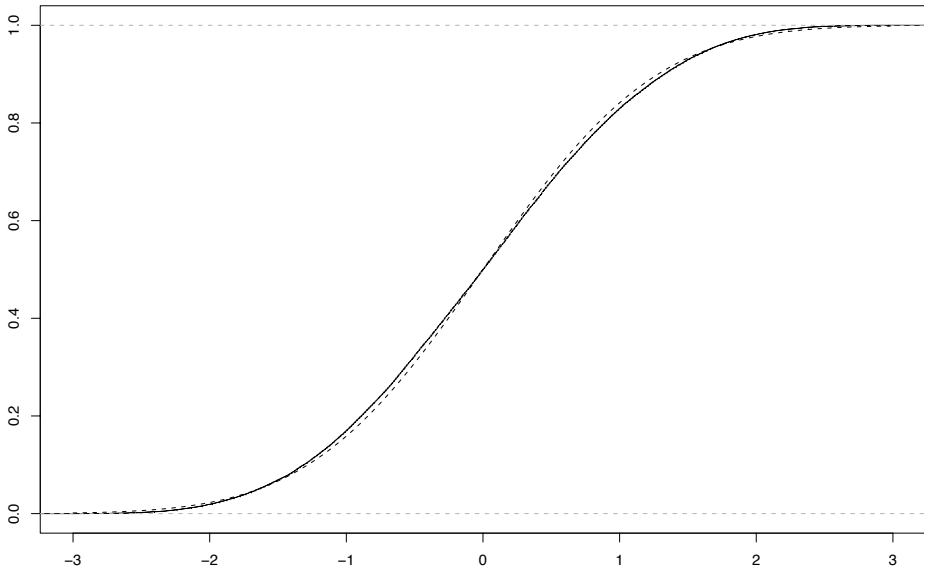


Voici à présent des simulations effectuées avec $\Delta = 0,2$, soit une dépendance plus faible des lancers vis-à-vis des résultats des lancers précédents. La différence avec une loi gaussienne, quoique réelle, étant plus difficile à observer, nous simulons cette fois des échantillons de taille 50000 plutôt que de taille 1000.

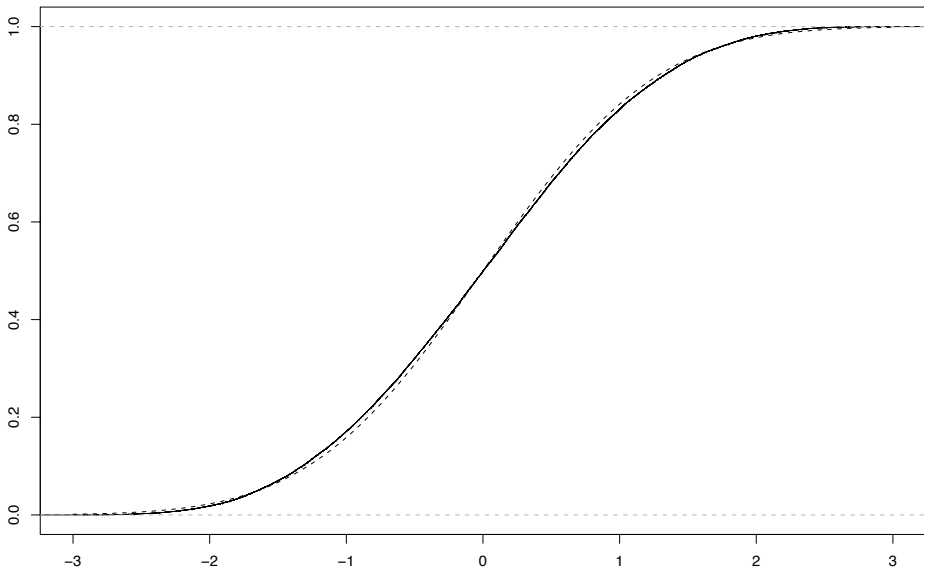
Pour $N = 100$.



Pour $N = 1000$.



Pour $N = 10000$.



L'existence de la variance

Dans les trois exemples précédents, nous avons considéré des sommes de variables aléatoires, certes dépendantes entre elles, mais ne pouvant prendre que les valeurs 0 et 1, et en fait toutes de loi de Bernoulli de paramètre $1/2$, ce qui assurait bien entendu l'existence de l'espérance et de la variance.

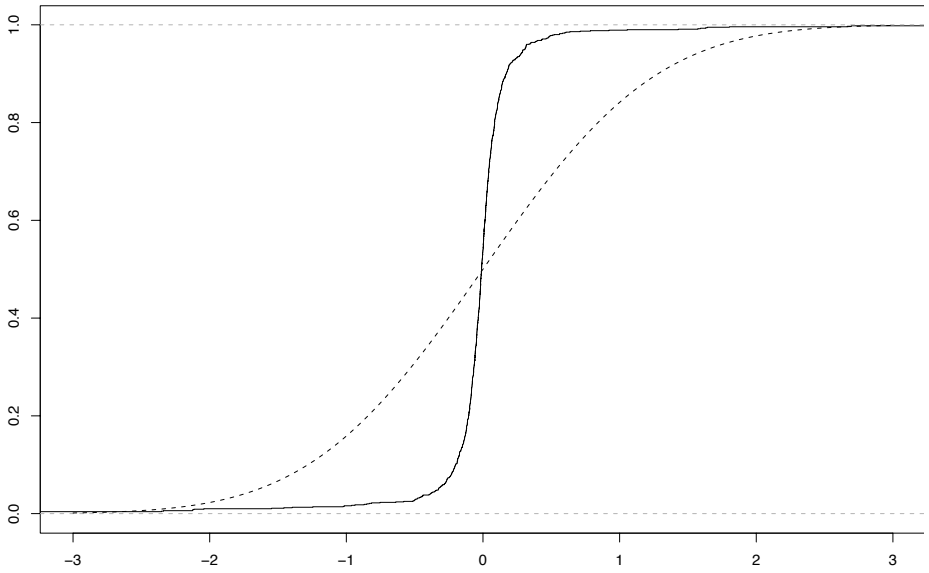
Posons-nous à présent la question, dans le cas de répétitions indépendantes d'une variable aléatoire, de la robustesse du théorème de la limite centrale vis-à-vis de l'existence de la variance de X. Du fait que $\mathbb{V}(X_1+\cdots+X_N)=+\infty$ si $\mathbb{V}(X)=+\infty$, l'énoncé selon lequel $\frac{S_N-\mathbb{E}(S_N)}{\sqrt{\mathbb{V}(S_N)}}$ suit approximativement une loi gaussienne lorsque N est grand, n'a plus de sens, et l'on ne peut plus ramener S_N à une échelle naturelle pour ses fluctuations en la centrant et en la réduisant comme nous l'avions fait jusqu'à présent. Il existe néanmoins une telle échelle naturelle, définie différemment, et l'observation de S_N sur cette échelle ne conduit pas à une loi gaussienne, mais à une loi dont la variance n'existe pas.

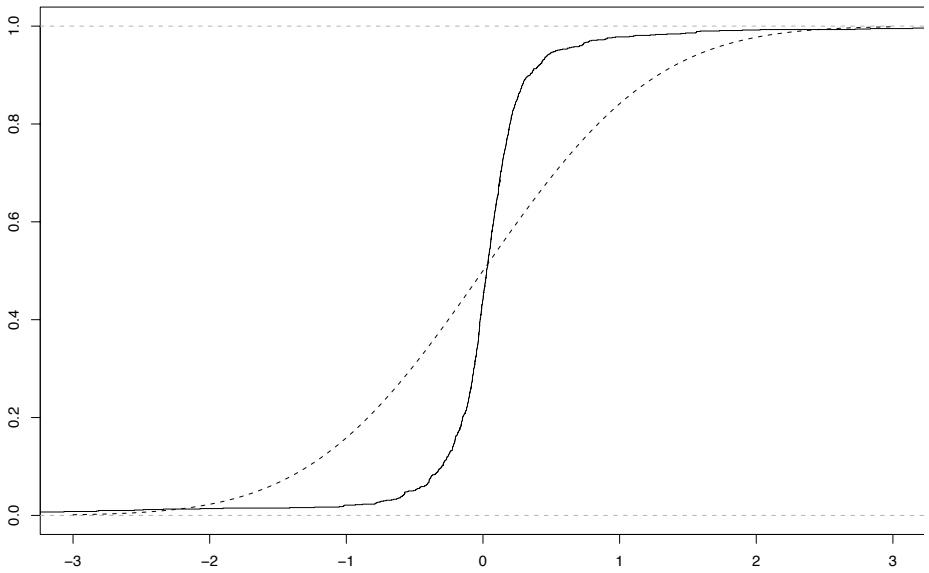
Nous vous renvoyons aux ouvrages d'introduction à la théorie des probabilités pour en apprendre plus sur le comportement de S_N en l'absence de variance définie.

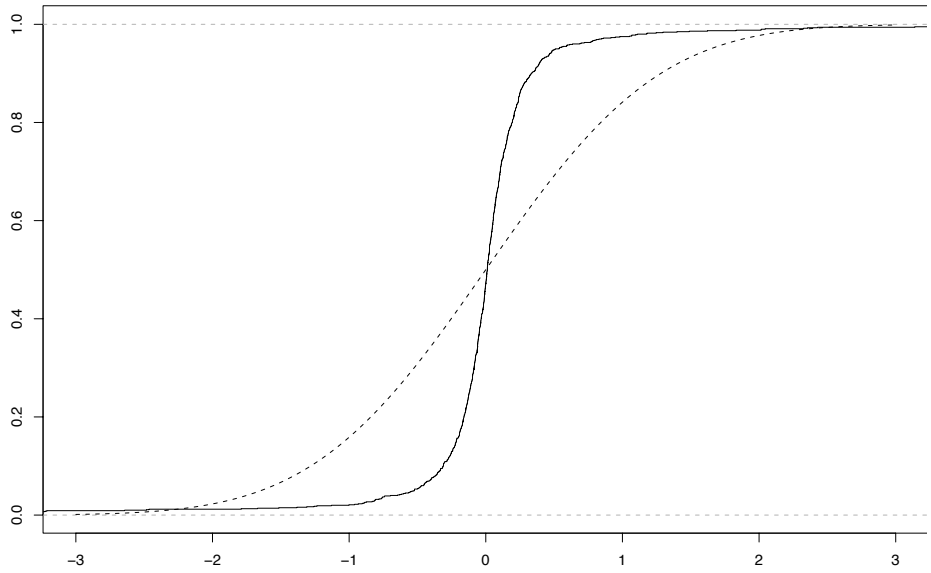
Pour prendre un exemple, si l'on choisit X de la forme $X=\text{signe}(C)\times|C|^{0,8}$, où C suit une loi de Cauchy de paramètre $s=1$ et $\ell=0$, on se trouve dans le cas où $\mathbb{E}(X)$ est définie, mais sans que $\mathbb{V}(X)$ le soit. Il est toujours possible, étant donné un échantillon simulé de valeurs de S_N, de centrer et de réduire la loi empirique associée à cet échantillon (puisque les valeurs de l'échantillon sont en nombre fini, l'espérance et la variance de cette loi sont toujours définies).

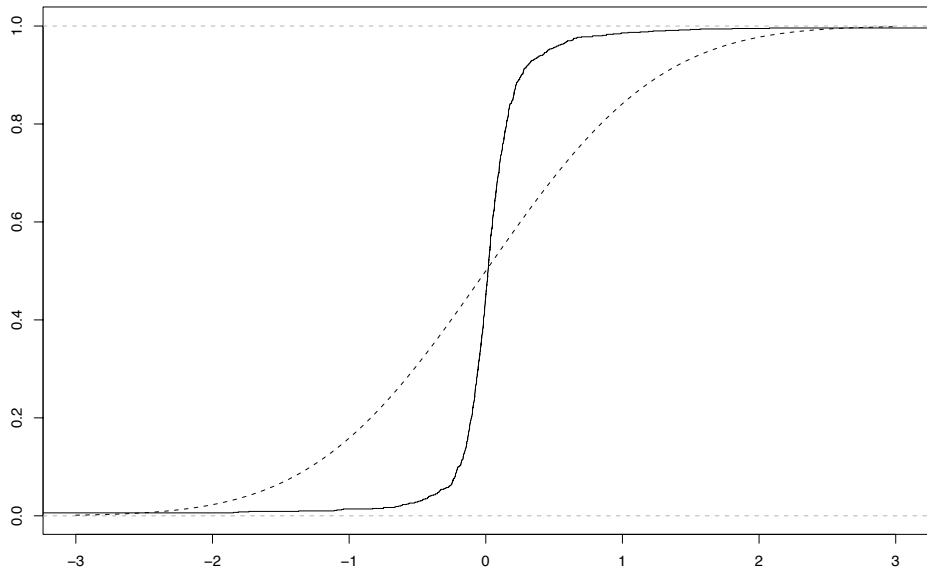
Voici quelques exemples de ce que l'on obtient en procédant de cette manière, avec des échantillons de taille 1000 :

Pour $N=100$, $N=1000$, $N=10000$ puis $N=100000$.


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3





On constate que les lois obtenues ne correspondent manifestement pas à des lois

gaussiennes centrées réduites, contrairement à la situation qui prévalait dans le cas où la variance était finie. Les graphes ci-dessus suggèrent en fait qu'il pourrait y avoir convergence, mais vers une loi différente d'une loi gaussienne. La situation est plus compliquée qu'il n'y paraît, car les valeurs de la variance empirique par lesquelles nous avons normalisé nos variables fluctuent considérablement d'une simulation à l'autre, ce qui ne permet pas de parler d'une échelle déterminée à laquelle on ramènerait la loi étudiée en procédant comme nous l'avons fait. Voici par exemple quatre valeurs successivement obtenues pour la variance de la loi empirique de S_N en effectuant 1000 simulations, et pour $N = 1000$: 2200,527 ; 1472,161 ; 2992,175 ; 11496,97.

Augmenter la taille des échantillons dans l'espoir de mieux stabiliser cette valeur n'est guère efficace. Toujours avec $N = 1000$, mais cette fois avec 20000 simulations, voici encore quatre valeurs successivement obtenues pour la variance de la loi empirique de S_N : 4204,347 ; 7855,702 ; 47786,29 ; 7700,53... Ceci est bien entendu lié au fait que la variance (théorique) de S_N n'est pas définie, et que la variance de la loi empirique ne peut se stabiliser autour de quelque valeur finie que ce soit.

Comme dans le chapitre précédent, on peut noter que, concrètement, le fait que la variance d'une variable aléatoire (dont on peut presque toujours prouver qu'elle est en fait bornée, même si les bornes sont gigantesques) n'est pas définie signifie plutôt ici que l'on ne peut observer un comportement du type de celui décrit par le théorème de la limite centrale pour un nombre raisonnable de termes dans la somme que l'on considère.

4.3.10 Le théorème de la limite centrale et le caractère universel (?) de la loi gaussienne

Le théorème de la limite centrale stipule que, observées sur leur échelle naturelle, les sommes d'un grand nombre de variables aléatoires indépendantes et de même loi possédant une espérance et une variance, présentent une distribution de probabilité approximativement gaussienne.

Il est tout-à-fait remarquable que ce soit toujours la loi gaussienne qui intervienne dans ce résultat, **quelle que soit** la loi des variables aléatoires que l'on additionne (pour peu qu'elle possède une espérance et une variance), même si celle-ci n'a au départ aucun rapport avec une loi gaussienne. La loi gaussienne possède donc, en ce sens, un caractère universel, car elle intervient systématiquement lorsque l'on a affaire à des sommes d'un grand nombre de variables aléatoires indépendantes et de même loi. [3]

3. Notons au passage, pour les lecteurs que l'intérêt d'une preuve mathématique de la loi des grands nombres, par rapport au recours à la simple intuition, aurait laissés sceptiques, que l'on obtient ici, en poursuivant l'étude mathématique du comportement asymptotique des sommes de variables aléatoires indépendantes, un résultat d'une grande portée, que l'intuition seule et non formalisée serait bien en peine d'atteindre.

Or, dans une grande variété de situations concrètes, on peut s'attendre à ce que les quantités que l'on étudie se présentent effectivement comme le résultat de l'addition d'un grand nombre de termes aléatoires, approximativement indépendants et du même ordre de grandeur.

Par conséquent, en prenant en compte la robustesse du théorème de la limite centrale (voir le paragraphe «Robustesse du théorème de la limite centrale» sur cette question), et le fait qu'il n'est souvent pas nécessaire que le nombre de variables mises en jeu soit très élevé pour que l'on observe une assez bonne approximation par une loi gaussienne, il est naturel de s'attendre à ce qu'un grand nombre de quantités présentent une distribution de probabilité décrite, au moins approximativement, par une loi gaussienne.

De fait, dans de nombreux domaines, il est très courant de modéliser – au moins en première approximation –, des variables quantitatives continues sous la forme d'un terme constant auquel s'ajoute un terme de fluctuation gaussien décrivant la variabilité de cette quantité.

Précisons un peu le rôle que peut jouer le théorème de la limite centrale dans ce contexte.

- un rôle de suggestion : si, sans pour autant disposer d'un grand nombre de données, ou de connaissances préalables approfondies, on peut raisonnablement penser que la quantité étudiée apparaît comme la somme d'un grand nombre de variables aléatoires approximativement indépendantes et de même loi possédant une espérance et une variance, alors le théorème de la limite centrale suggère qu'il peut être pertinent, au moins en première approximation, de tenter de modéliser la distribution de cette quantité au moyen d'une loi gaussienne. Bien entendu, cette modélisation doit, autant que possible, être ensuite confrontée avec les données recueillies, et plus généralement les connaissances acquises, sur la quantité considérée. Par ailleurs, il ne s'agit pas de la seule raison pouvant suggérer l'utilisation d'une loi gaussienne dans un modèle, d'autres propriétés de cette loi [4] de celle-ci pouvant conduire à la sélectionner dans certains contextes.
- un rôle d'explication : si la distribution observée d'une quantité apparaît, au moins approximativement, comme gaussienne, le théorème de la limite centrale suggère comme une explication possible le fait que cette quantité résulte de l'addition d'un grand nombre de variables aléatoires approximativement indépendantes et de même loi possédant une espérance et une variance. Ce n'est bien entendu pas la seule explication possible, et il ne s'agit que d'une sugges-

4. Par exemple ses propriétés de maximisation d'entropie, ou d'isotropie spatiale, voir les exercices 166 et 167. Ou encore, la possibilité qu'elle offre de mener explicitement un certain nombre de calculs, ce qui, avant l'avènement des ordinateurs modernes et de leurs puissantes capacités de calcul, la rendaient parfois la seule utilisable en pratique.

tion d'explication tant qu'elle n'a pas été effectivement validée par la mise en évidence de ces variables aléatoires et la vérification des propriétés qu'on leur prête.

Avant de donner des illustrations concrètes d'apparition de la loi gaussienne, mentionnons le fait qu'elle a été considérée comme tellement omniprésente qu'on lui a également attribué le nom de «loi normale». Pour citer une boutade attribuée à Henri Poincaré [5] à propos de l'utilisation de la loi gaussienne : «Tout le monde y croit cependant, car les expérimentateurs s'imaginent que c'est un théorème mathématique, et les mathématiciens que c'est un fait expérimental.»

4.4 Des exemples concrets

Pour illustrer le fait que la loi gaussienne apparaît effectivement dans certaines situations réelles, mais qu'elle n'apparaît pas non plus de manière systématique, nous présentons dans ce qui suit plusieurs jeux de données réelles (tous issus de la base de données `MASS` du logiciel `R`).

Dans tous les exemples qui suivent, il est clair qu'une grande quantité de facteurs interviennent dans la formation des quantités étudiées. Quant à déterminer si c'est effectivement le théorème de la limite centrale qui explique l'apparition de la loi gaussienne dans les exemples où celle-ci est observée, ou quel écart par rapport aux hypothèses de ce théorème pourrait expliquer le caractère non-gaussien des autres exemples, nous nous en remettons principalement à votre propre sagacité.

Précisément, nous comparons dans ce qui suit les lois empiriques associées à des échantillons de valeurs mesurées à la loi gaussienne centrée réduite, après les avoir centrées, puis réduites. Dans l'interprétation fréquentielle de la probabilité, qui n'est pas *a priori* garantie dans nos exemples – il faudrait en savoir bien davantage sur la manière dont les données ont été collectées –, mais qui constitue le cadre dans lequel nous nous placerons par défaut, la loi empirique associée à un grand échantillon fournit une approximation de la loi théorique de la quantité sur laquelle portent les données mesurées. Pour juger à quel point l'écart observé entre une telle loi empirique et la loi gaussienne peut être attribué à un écart entre la loi théorique (centrée réduite) et la loi gaussienne centrée réduite, ou plutôt à des fluctuations d'échantillonnage, liées au fait que l'on ne considère que des échantillons comportant un nombre fini de données, et se traduisant par le fait qu'il existe presque toujours un écart entre la loi empirique associée aux données et la loi théorique, il est nécessaire de faire des hypothèses supplémentaires sur le processus d'échantillonnage (qu'il faudrait elles-mêmes valider) : par exemple supposer que les valeurs mesurées peuvent être considérées comme issues de réalisations indépendantes de la loi théorique. Ce type de question

5. Henri Poincaré (1854–1912).

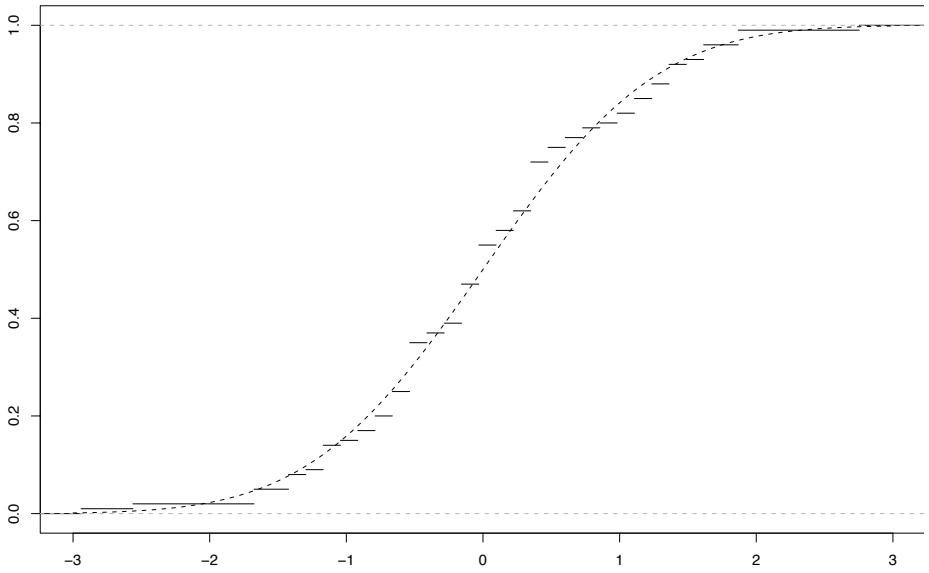
s'aborde normalement dans le cadre méthodologique des tests statistiques, qui sera décrit dans le chapitre «Statistique». Vous pouvez également consulter avec profit le paragraphe traitant du théorème de Glivenko-Cantelli dans le chapitre précédent. Nous nous contenterons, à titre d'illustration, de comparer succintement et graphiquement les écarts observés entre les lois empiriques associées aux données et la loi gaussienne, à des écarts observés entre les lois empiriques associées à des échantillons de même taille que les échantillons de données, mais constitués de simulations de variables aléatoires indépendantes et de loi gaussienne.

4.4.1 Des exemples approximativement gaussiens

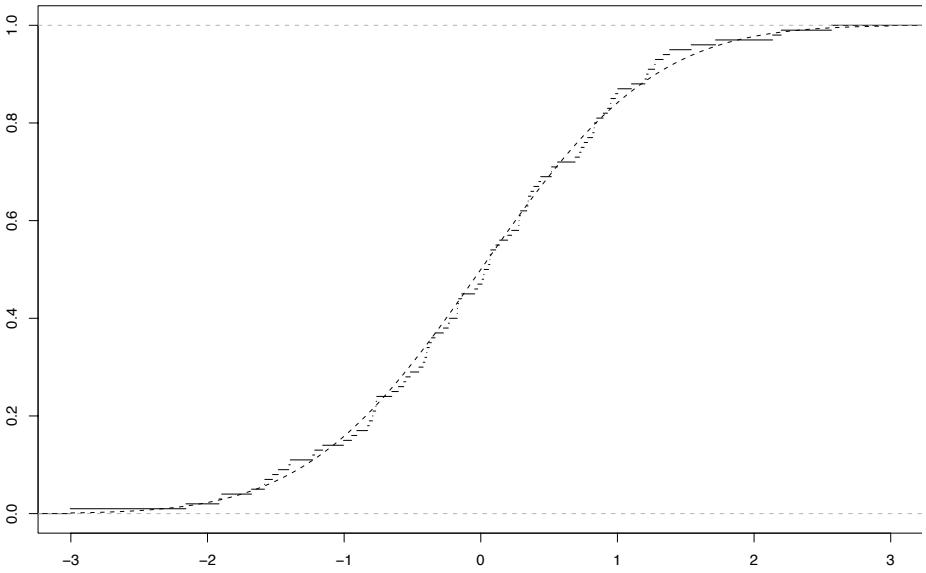
À la vitesse de la lumière !

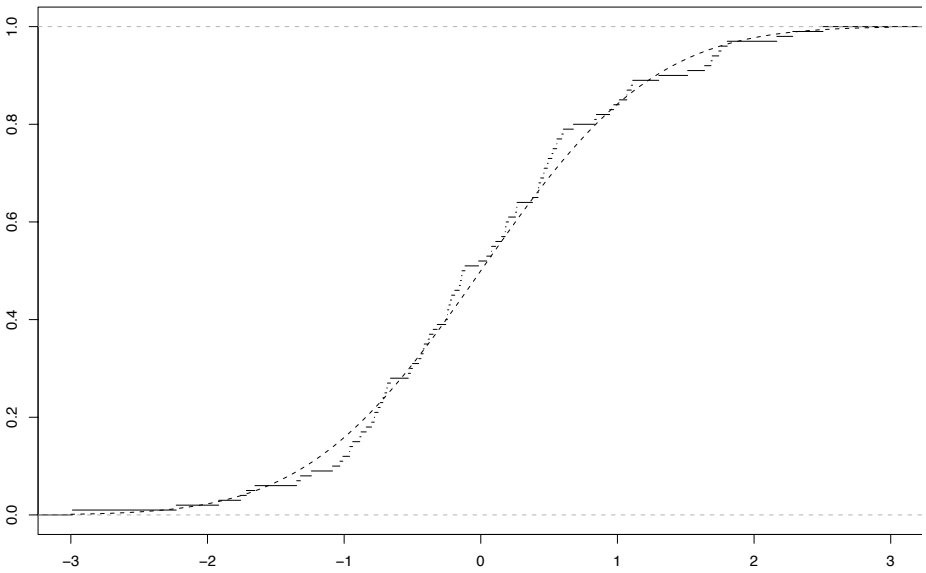
Pour commencer, un exemple historique : une liste de 100 mesures de la vitesse de la lumière dans l'air (en km/s), relevées au cours de la célèbre expérience de Michelson (1879), qui permit de montrer que, contrairement aux prédictions de la mécanique newtonienne, la vitesse de la lumière était la même dans tous les référentiels, ouvrant ainsi la voie à la théorie de la relativité restreinte d'Einstein. (Ces données proviennent d'articles de A. Weekes et S. Stigler repris dans la base de données `MASS` du logiciel `R`).

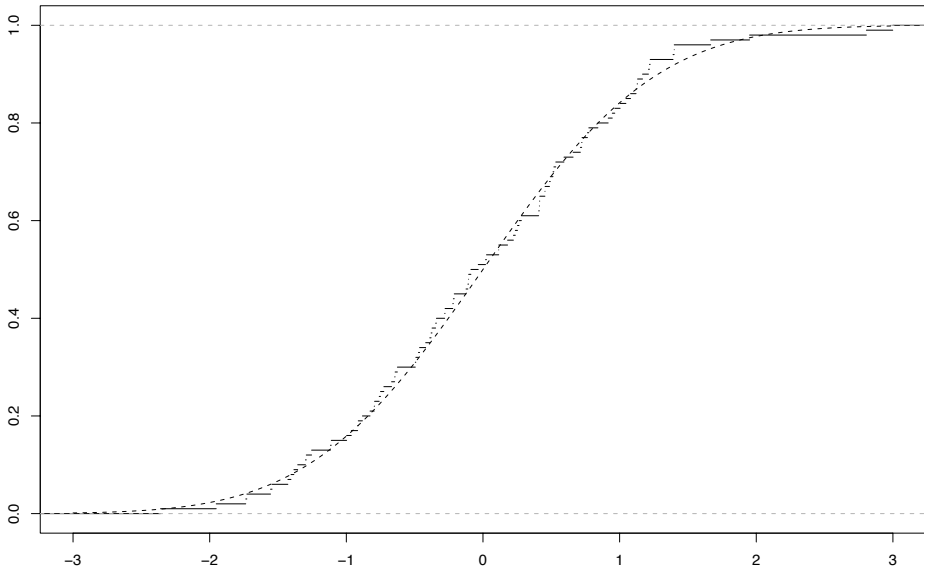
Le graphique ci-dessous représente la fonction de répartition de la loi empirique de l'échantillon constitué par les 100 mesures, centrée et réduite. En pointillés, la fonction de répartition de la loi gaussienne standard.



Les deux fonctions de répartition semblent raisonnablement proches, ce qui accrédite le fait que les mesures sont approximativement distribuées selon une loi gaussienne. A titre de comparaison, voici plusieurs graphiques obtenus en appliquant le même traitement à un échantillon simulé de 100 variables aléatoires gaussiennes centrées réduites indépendantes.

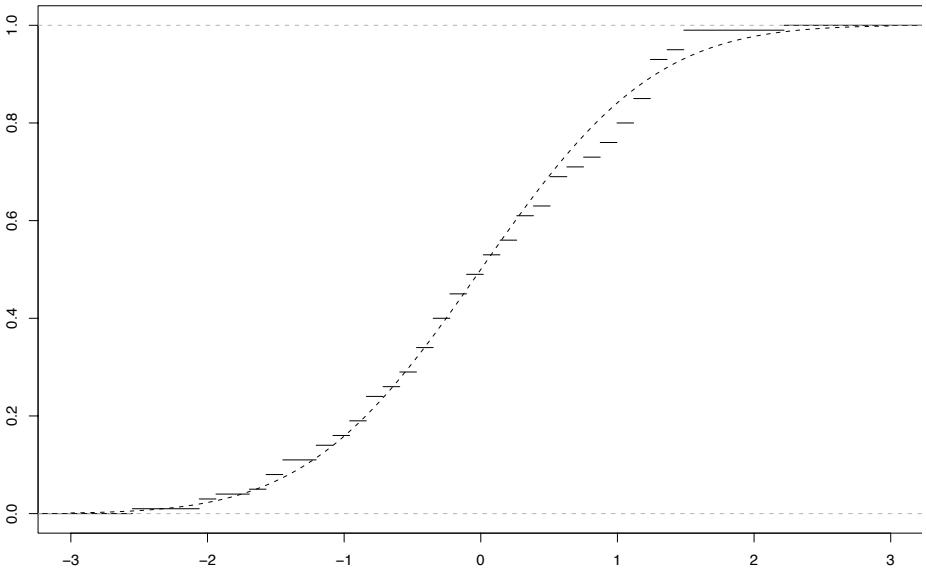

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

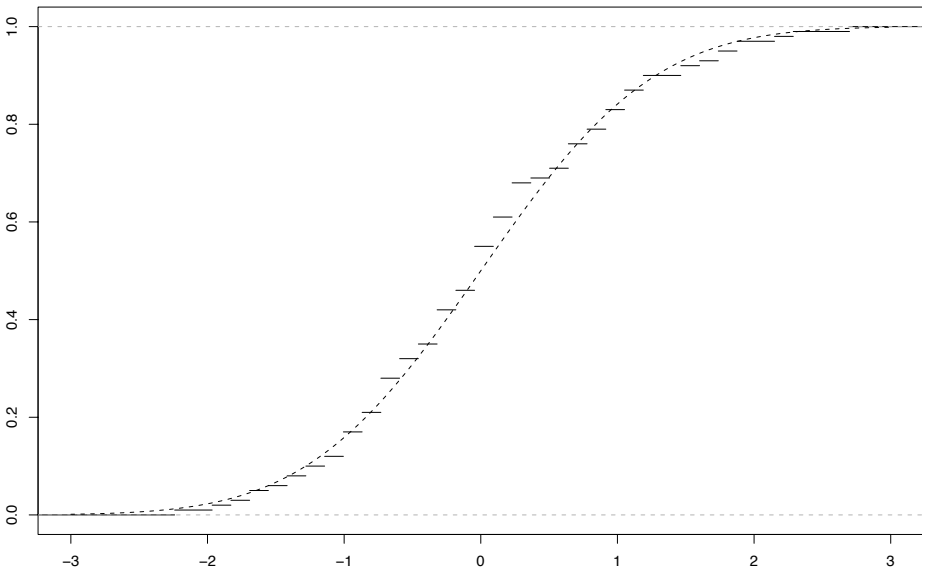

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

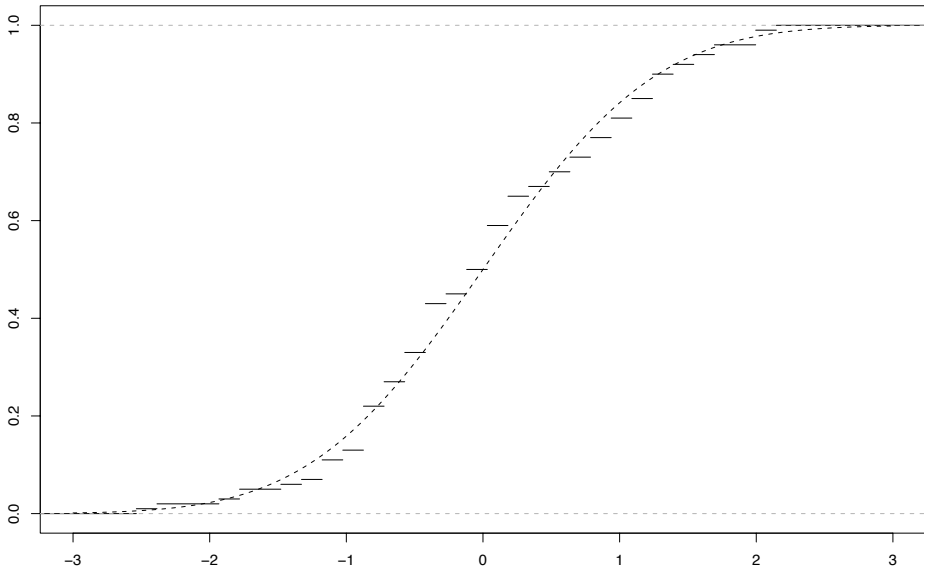


Ces trois graphiques sont bien entendu différents les uns des autres, ce qui illustre les variations de la loi empirique lorsque l'on considère des échantillons de taille modérée. Cependant, chacun de ces graphiques semble plus proche des deux autres qu'il ne l'est de celui associé aux données mesurées, ce qui peut amener à douter du fait que celles-ci puissent être exactement modélisées au moyen d'une loi gaussienne. Pour bien faire, il faudrait naturellement simuler un **grand nombre** de tels graphiques, afin de vérifier si le graphique obtenu avec nos données mesurées est réellement atypique par rapport à l'ensemble de ceux-ci, alors que nous nous sommes contentés de trois exemples. C'est exactement le principe des tests statistiques, qui nécessite bien entendu une définition plus précise et quantitative de l'écart que le simple fait que nos yeux (et notre cerveau) nous suggèrent une différence. Nous nous restreindrons cependant ici à ces trois exemples, en renvoyant au chapitre «Statistique» pour un traitement plus abouti de ce type de question.

Revenant à nos données, on constate qu'une petite remarque permet de mieux comprendre ce qui se passe : en fait, les valeurs mesurées dont nous disposons (en km/s) ont manifestement été arrondies à la dizaine. Si l'on simule 100 variables aléatoires gaussiennes de même espérance et de même variance que la loi empirique associée aux 100 mesures, et qu'on leur fait subir le même type d'arrondi, on obtient les graphiques suivants :


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3



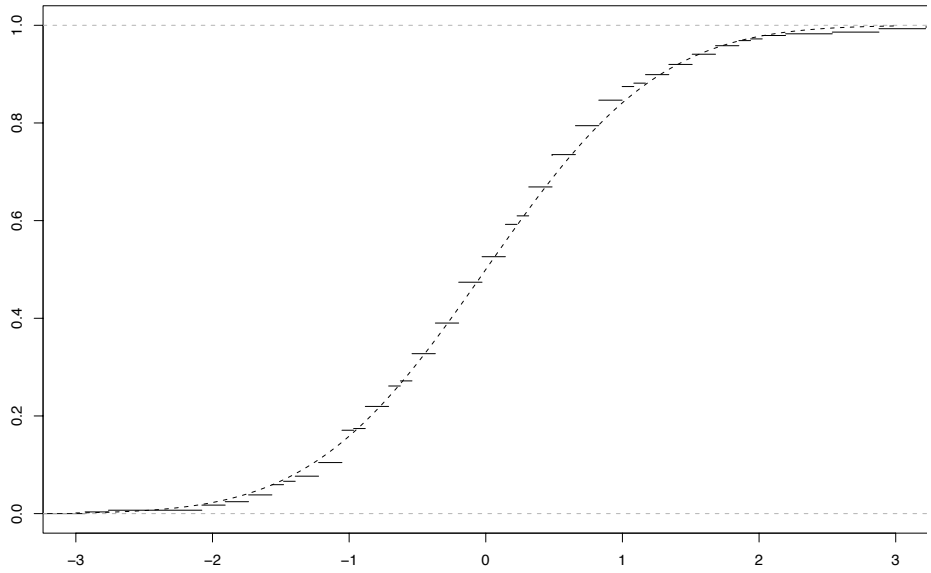
qui ressemblent beaucoup plus que les trois graphiques précédents au graphique obtenu avec les valeurs mesurées [6].

Nous verrons, dans le chapitre «Statistique», des moyens systématiques de tester l'adéquation entre des valeurs mesurées et un modèle en tenant compte des variations possibles de la loi empirique résultant de l'échantillonnage. Nous nous sommes ici contentés d'un traitement on ne peut plus informel de cette question. Par ailleurs, cet exemple fait apparaître le caractère crucial de la qualité des données (et, en particulier, du traitement qu'elles peuvent avoir subi).

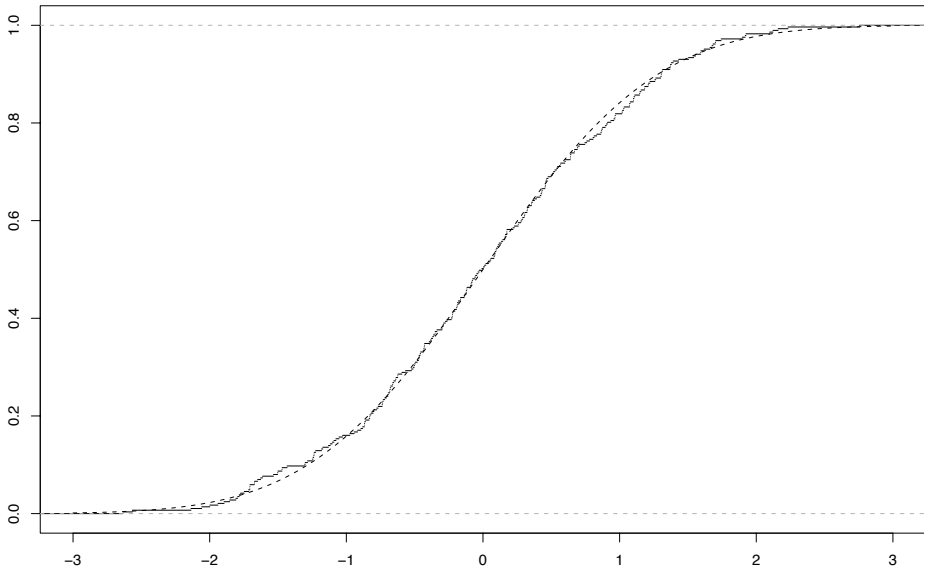
Des Indiennes

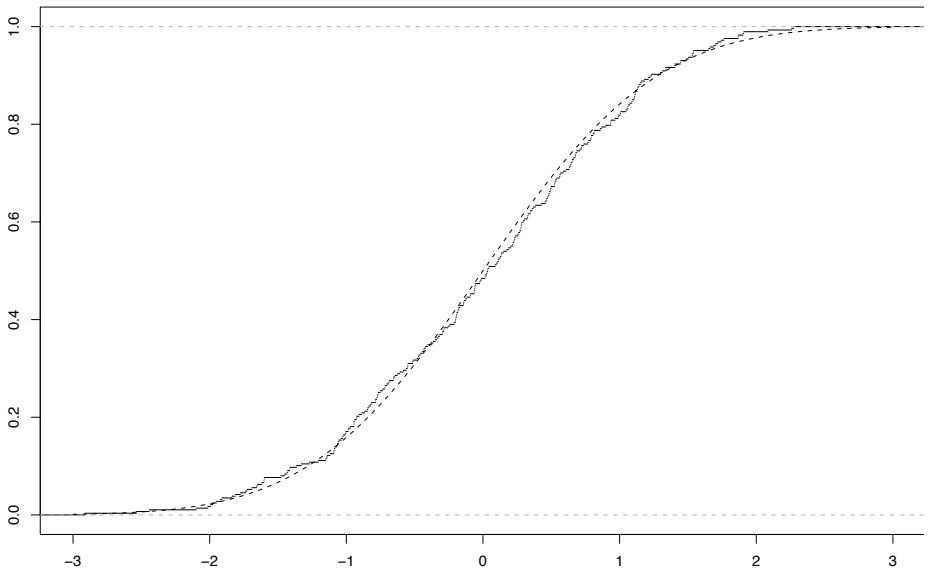
Prenons un autre exemple : des mesures de la pression sanguine diastolique (en millimètres de mercure) chez 287 femmes de la tribu indienne Pima, âgées de plus de 21 ans et vivant près de la ville de Phoenix, Arizona, États-Unis d'Amérique. (Ces données proviennent d'un article de J. Smith et coll. repris dans la base de données `MASS` du logiciel `R`).

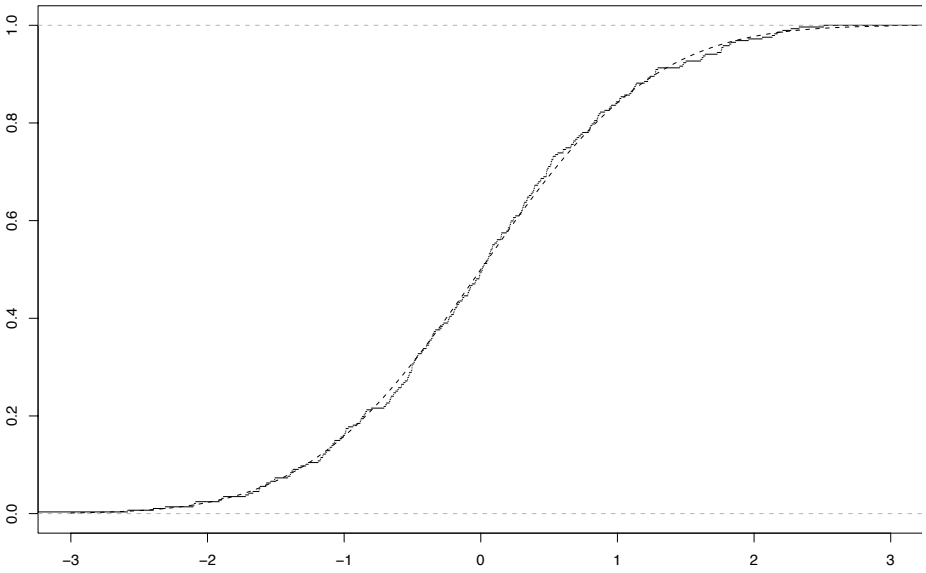
6. Des histogrammes dont la largeur des classes serait macroscopique devant l'ordre de grandeur des arrondis effectués ne feraient pas apparaître la différence que nous avons constatée sur les fonctions de répartition. C'est à la fois un avantage (cette question d'arrondi n'est pas forcément pertinente à l'échelle à laquelle on entend décrire la vitesse) et un inconvénient, car l'utilisation exclusive d'histogrammes, sans examen des données elles-mêmes pourrait nous faire passer à côté de cette propriété des données.

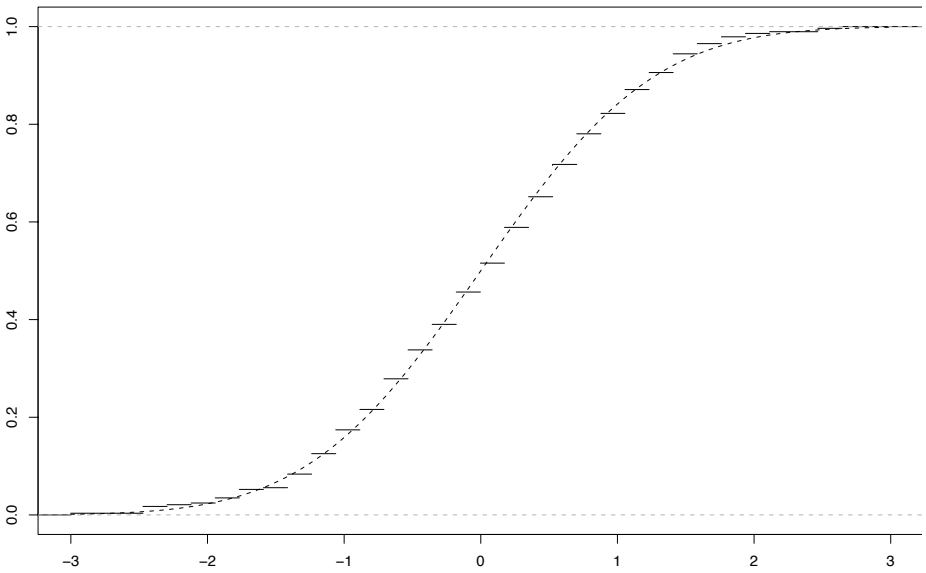


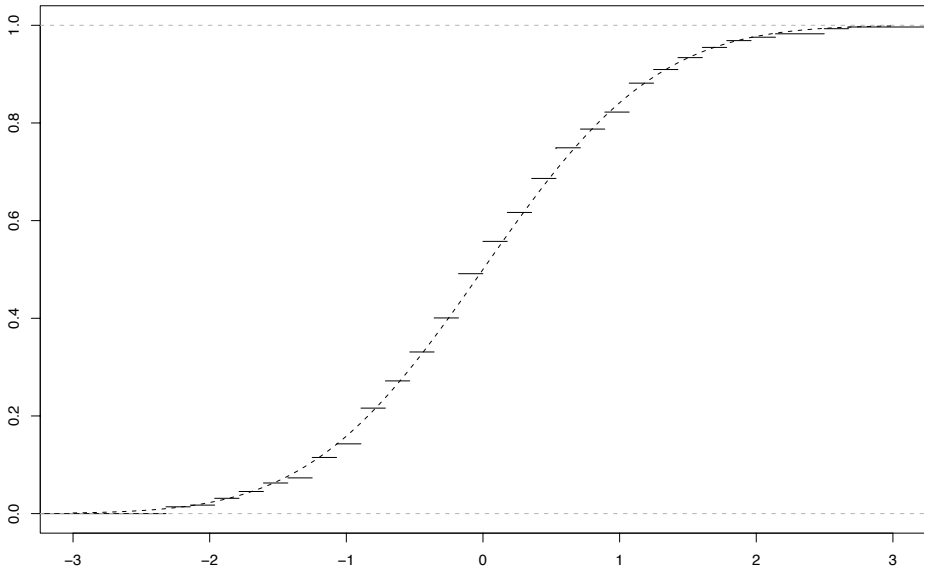
Ici encore, les données ont manifestement été arrondies, car les 287 mesures sont toutes des nombres entiers (à deux chiffres), dont 273 sont des nombres pairs. Le même type de remarque que précédemment s'applique donc. Voici les six graphiques correspondants : les trois premiers associées à des échantillons de 287 variables aléatoires gaussiennes simulées, les trois suivants à des échantillons de 287 variables aléatoires gaussiennes simulées et convenablement arrondies.

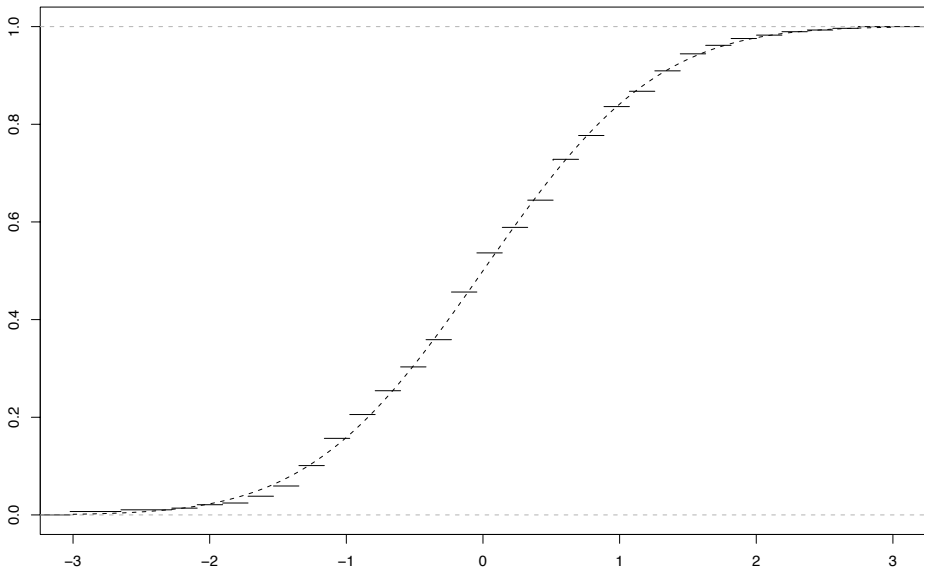

0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

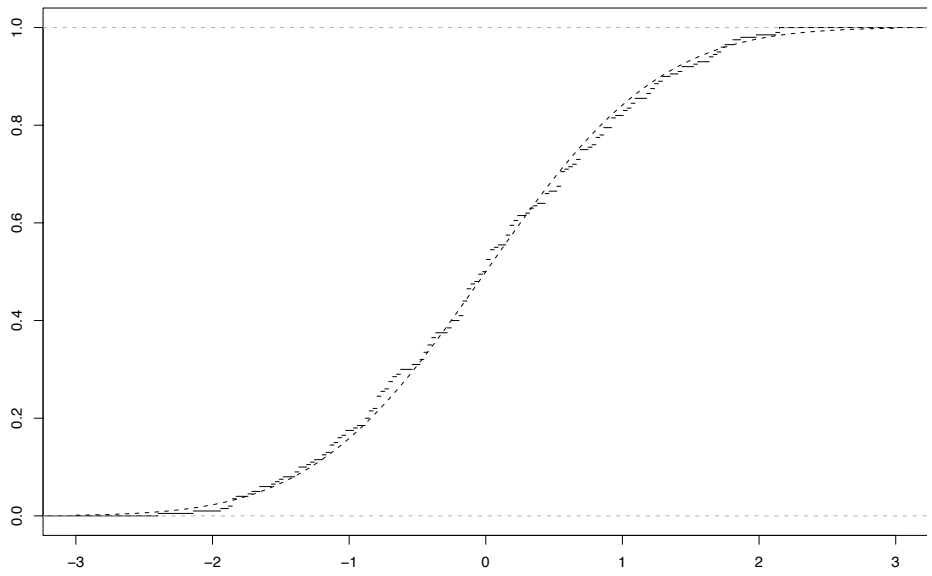

0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

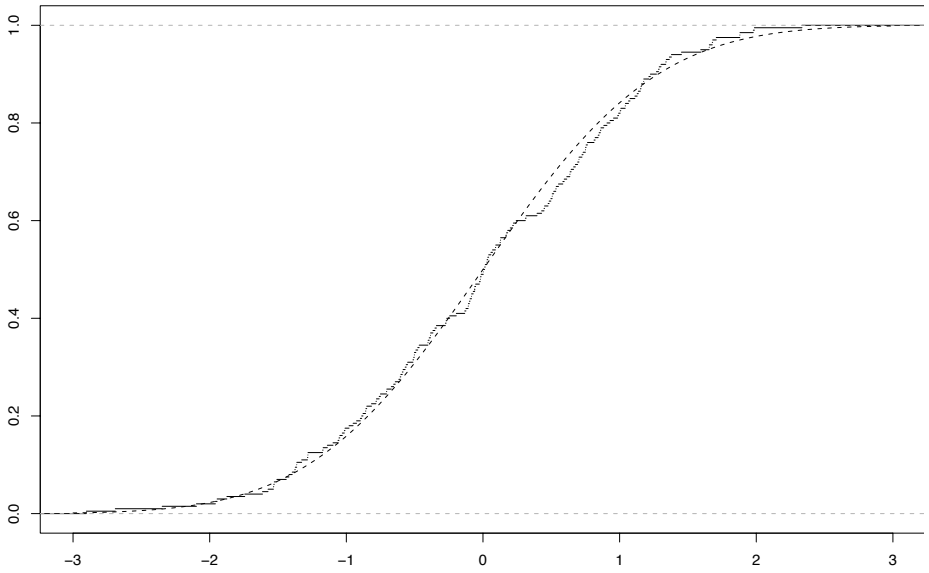
Des crabes

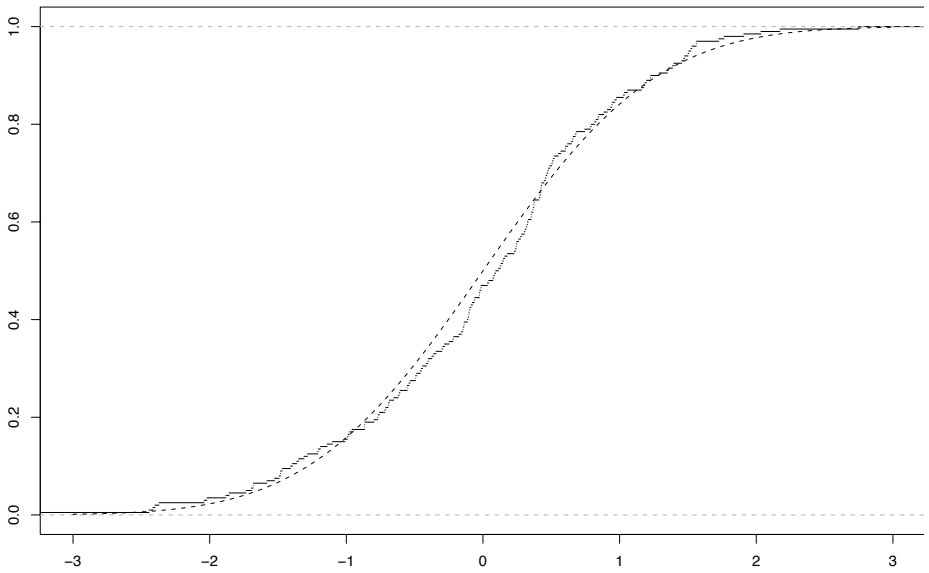
Encore un exemple : des mesures de la taille du lobe frontal (exprimée en millimètres) chez le crabe *Leptograpsus variegatus*, effectuées sur 200 spécimens.

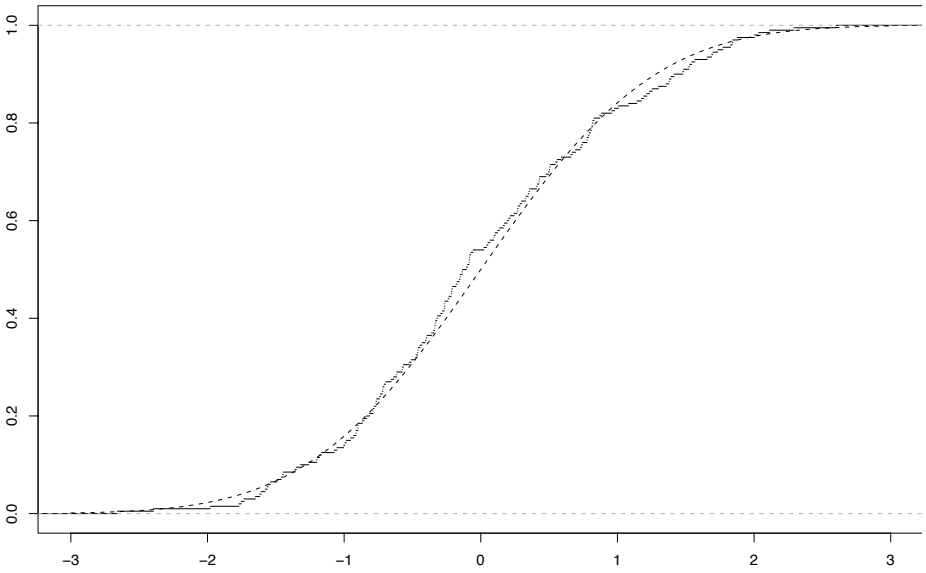
(Ces données proviennent d'un l'article de N. Campbell et J. Mahon repris dans la base de données `MASS` du logiciel `R`).

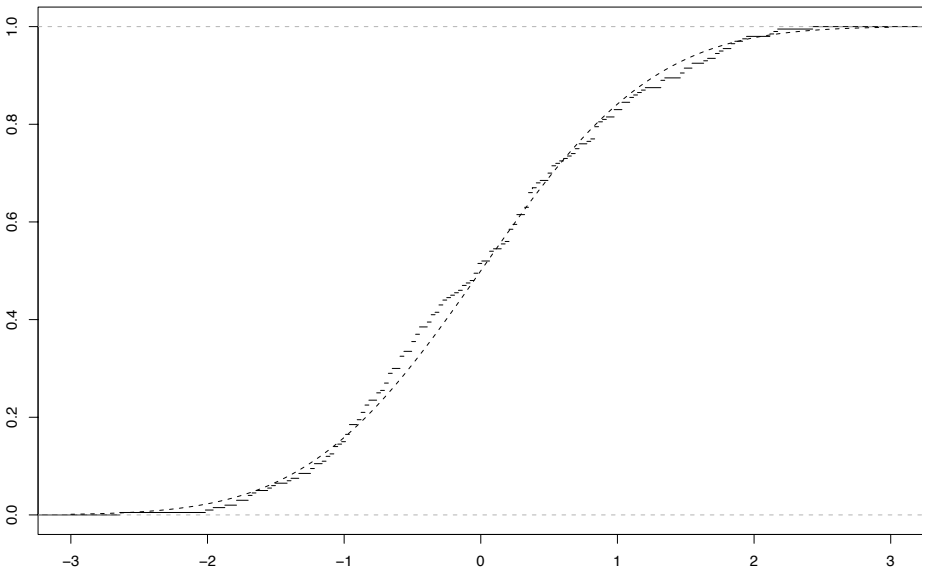


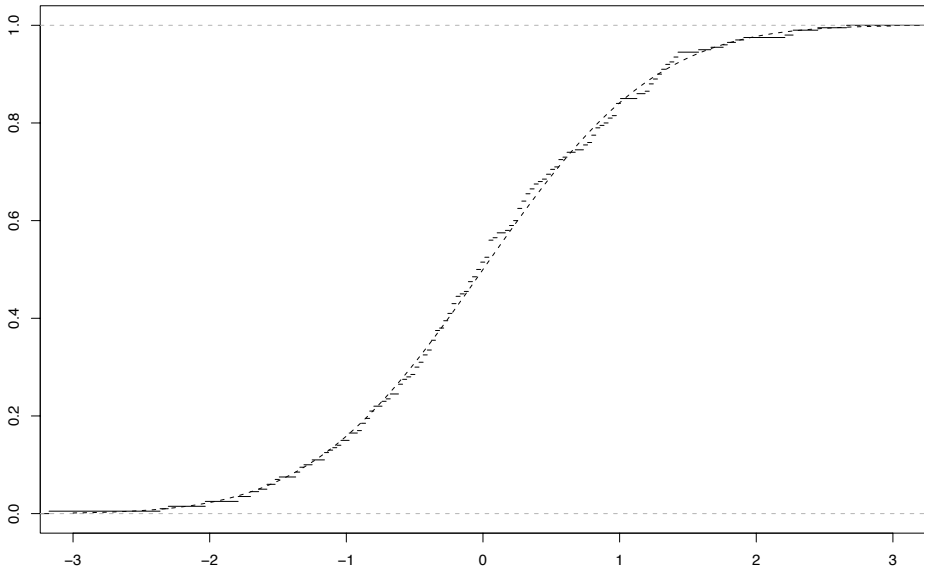
Ici encore, les données ont été manifestement été arrondies à la première décimale. Le même type de remarque que précédemment s'applique donc. Voici les six graphiques correspondants : les trois premiers associées à des échantillons de 200 variables aléatoires gaussiennes simulées, les trois suivants à des échantillons de 200 variables aléatoires gaussiennes simulées et convenablement arrondies.

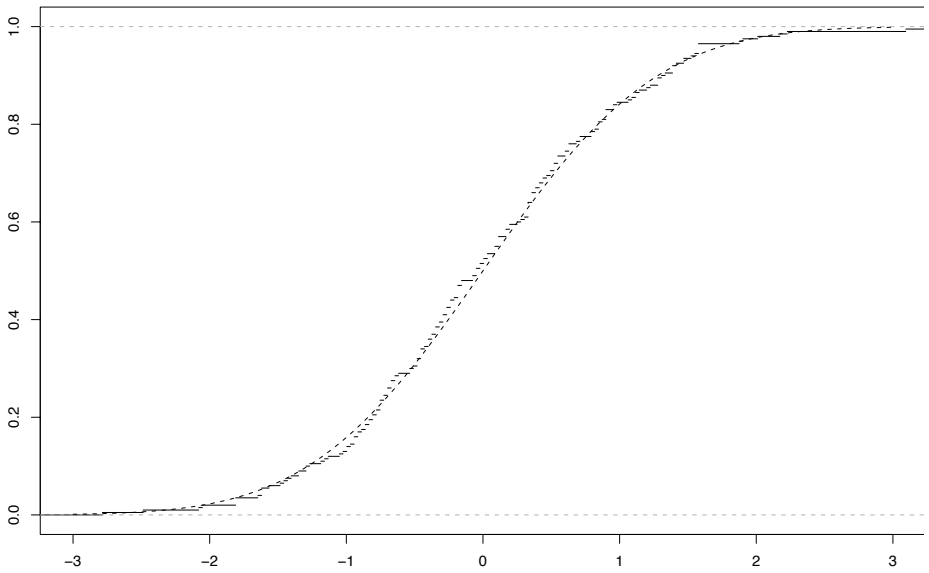

0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

4.4.2 Des exemples non gaussiens, même approximativement

Nous donnons dans ce qui suit des exemples de données dans lesquelles la loi empirique des données diffère grossièrement d'une gaussienne.

Tout d'abord, notons qu'un grand nombre de quantités ne peuvent évidemment pas posséder une distribution gaussienne, notamment toutes les quantités possédant un caractère discret à l'échelle où on les observe. Dans la suite, nous donnons des exemples de quantités continues dont la distribution, une fois ramenée à son échelle naturelle, pourrait *a priori* être décrite par une gaussienne, mais ne l'est manifestement pas.

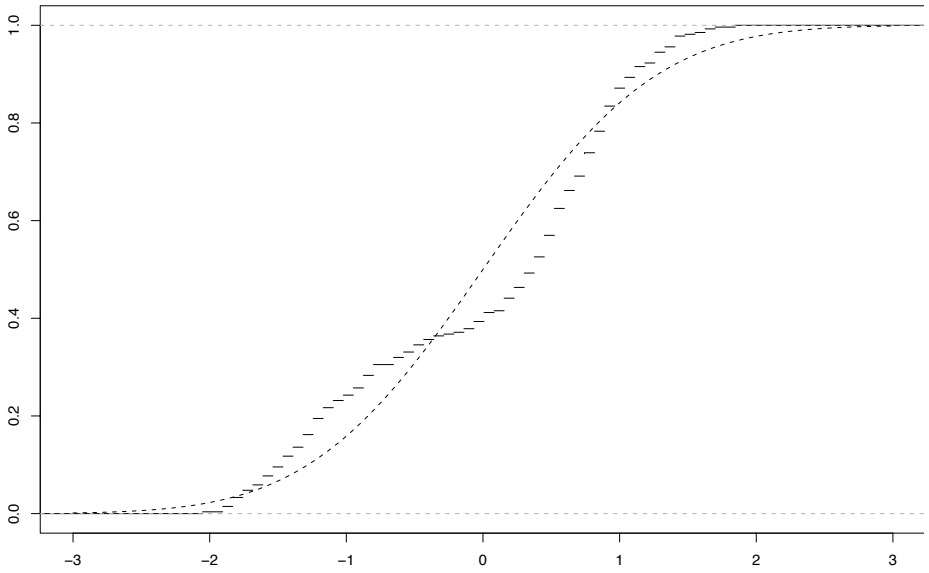
Une loi de probabilité peut différer d'une loi gaussienne de bien des manières, mais il n'est pas inutile de caractériser, même grossièrement, le type de propriété d'une loi gaussienne qui n'est pas satisfaite par les données. Trois propriétés fondamentales de la loi gaussienne sont par exemple : son caractère unimodal, son caractère symétrique, et, si les deux précédentes propriétés sont vérifiées, la forme précise de la fonction qui délimite la «cloche».

Un geyser fidèle

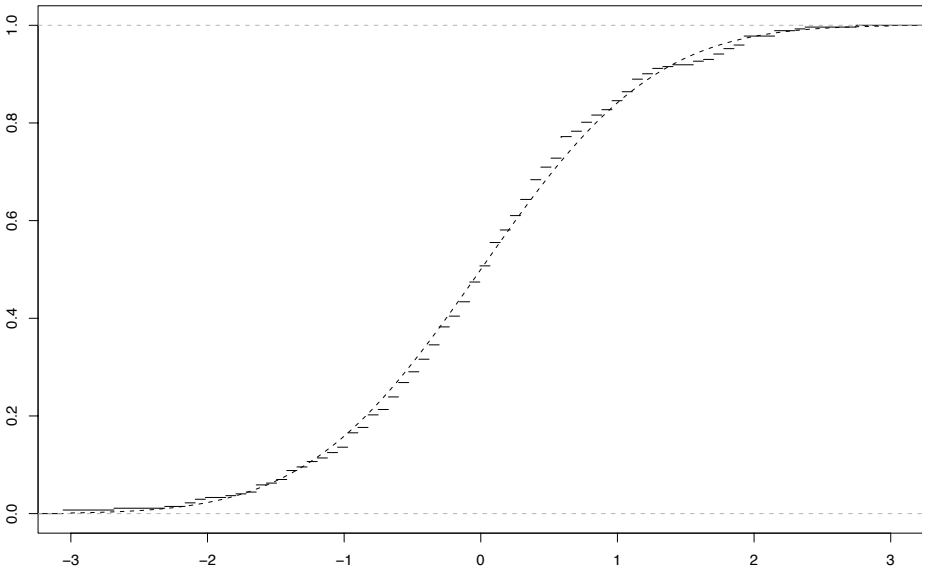
Cet exemple est constitué par une liste de mesures des durées inter-éruptions (en minutes) du geyser dénommé «The Old Faithful» dans le parc du Yellowstone aux États-Unis, réalisées en continu pendant deux semaines au mois d'août 1985. Cette liste comporte 272 mesures.

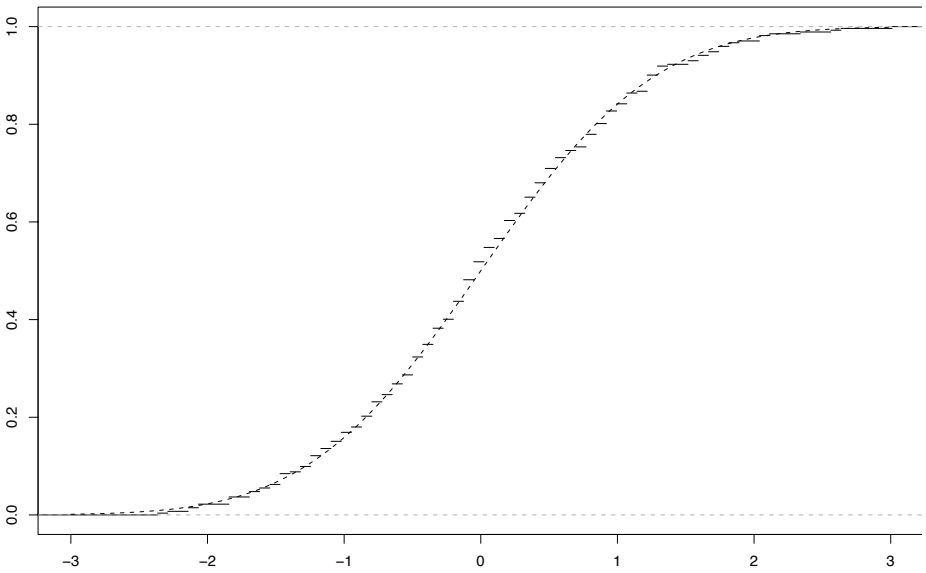
(Ces données proviennent d'un article de W. Härdle repris dans la base de données `MASS` du logiciel `R`).

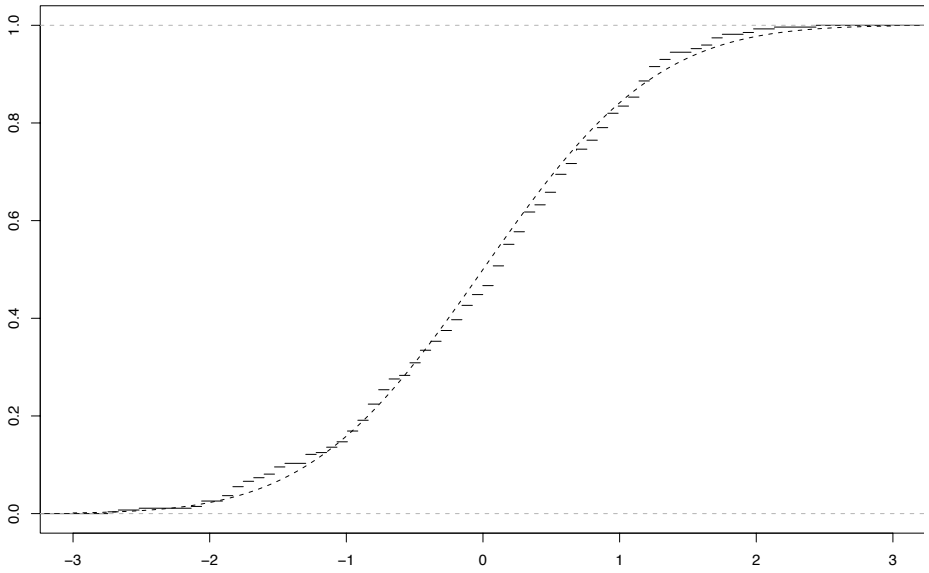
Le graphique ci-dessous représente la fonction de répartition de la loi empirique de l'échantillon constitué par les 272 mesures, centrée et réduite. En pointillés, la fonction de répartition de la loi gaussienne standard.



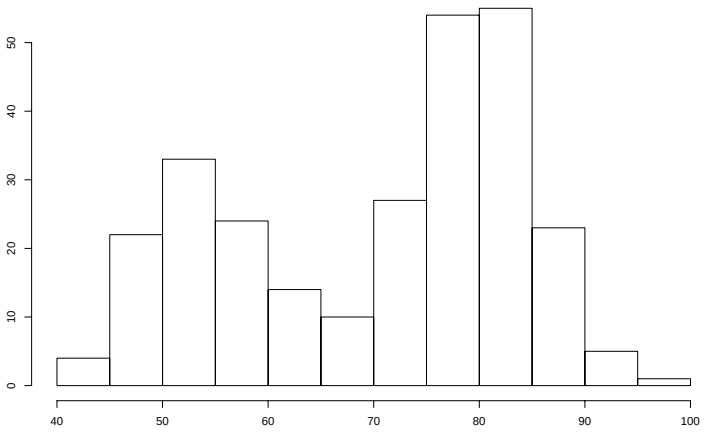
Cette fois, même en tenant compte du fait que les données ont manifestement été arrondies (les durées en minutes sont toutes des nombres entiers), il ne semble pas que l'écart observé entre la fonction de répartition gaussienne et celle des données puisse être mis sur le compte de fluctuations de la loi empirique associée à un échantillonnage de taille finie. Une méthode pour quantifier ce fait de manière correcte et précise serait d'effectuer un test statistique, mais nous nous contenterons ici, comme dans les exemples précédents, de comparer avec trois graphiques correspondant à 272 variables aléatoires gaussiennes simulées et arrondies d'une manière comparable. Les données présentées étant structurées, il semble malgré tout moins pertinent que dans les exemples précédents de simplement comparer nos données avec des échantillons de simulations **indépendantes** de variables aléatoires gaussiennes. Tenter de tenir compte correctement du caractère structuré des données, et de son éventuelle influence, pour aborder cette question dépasse de loin le niveau de ce que nous souhaitons présenter ici, mais il n'est certainement pas inutile de mentionner ce point, afin au moins de souligner que, de manière générale, des méthodes générales et standardisées ignorant une partie de la structure sous-jacentes à un phénomène que l'on étudie, ne sont pas forcément les plus pertinentes.


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3


0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3



En représentant un histogramme des données relatives au geyser, la situation s'éclaire quelque peu.

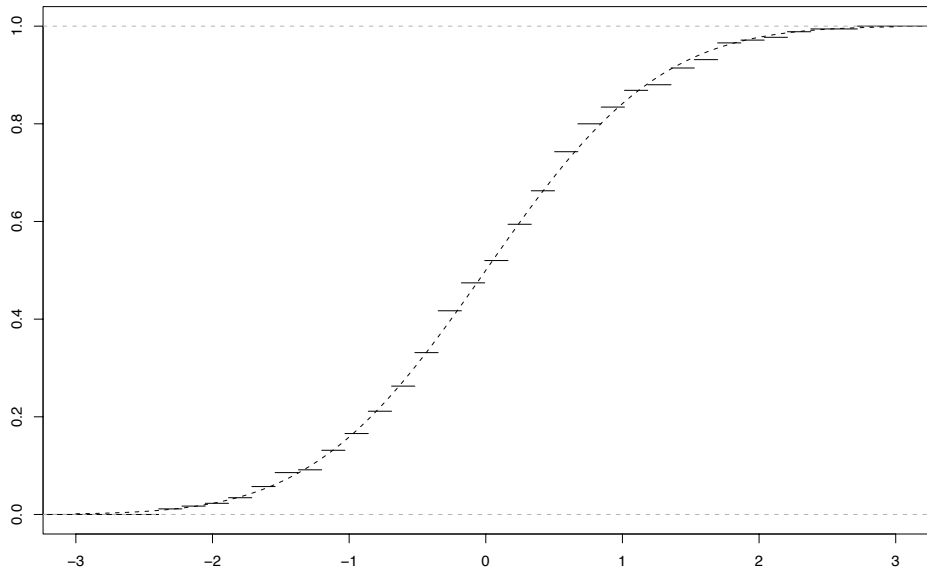


En effet, on constate que la distribution des données est grossièrement bimodale, et viole donc le caractère unimodal de la loi gaussienne.

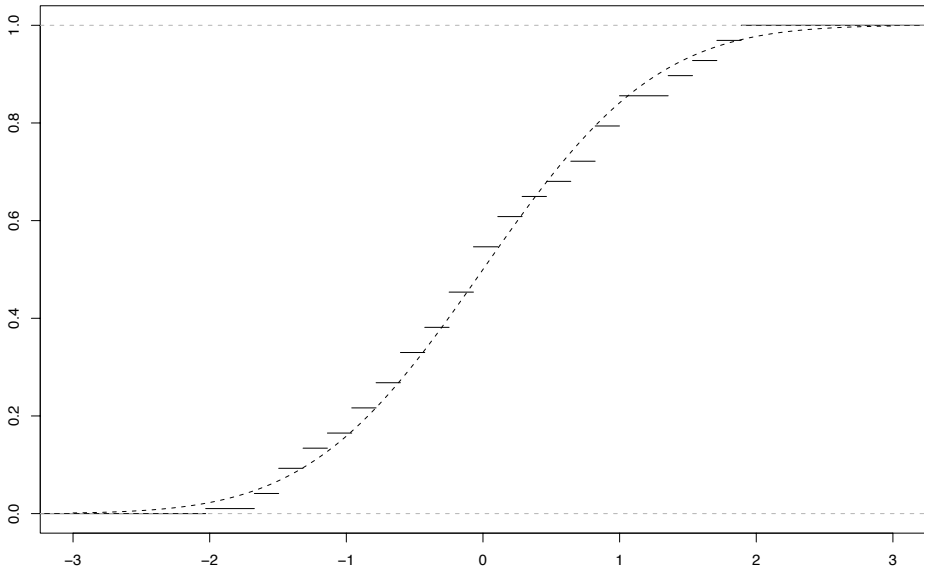
En revanche, si l'on sépare les données en deux groupes, selon que les durées sont

supérieures ou inférieures à 65 minutes, et que l'on représente les lois empiriques – centrées et réduites – associées à chacun de ces deux groupes de données, on obtient les deux graphiques suivants.

Pour le groupe des 175 données supérieures (strictement) à 65 minutes :



Pour le groupe des 97 données inférieures (ou égales) à 65 minutes :



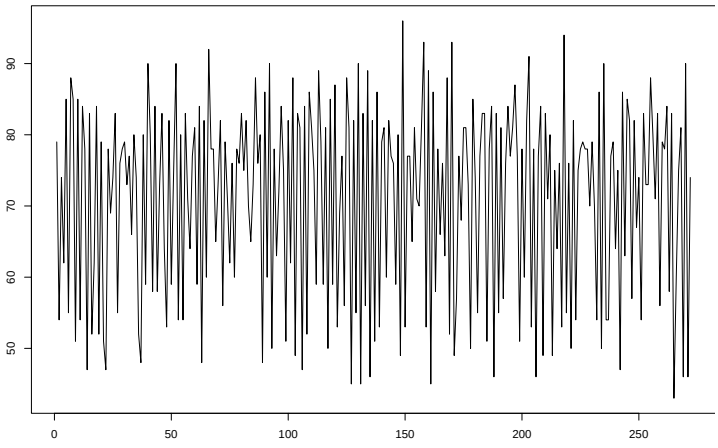
On constate qu'il n'apparaît pas *a priori* déraisonnable de représenter la distribution des données comme un mélange de deux lois gaussiennes de paramètres différents, correspondant à une loi gaussienne pour les faibles durées, et à une autre loi gaussienne pour les fortes durées. Il n'y a pas de raison *a priori* pour qu'une loi multimodale puisse toujours être représentée comme un mélange de lois gaussiennes, mais cette propriété remarquable de notre exemple méritait d'être soulignée. Qui plus est, elle suggère une raison générale permettant de s'attendre à ce qu'une quantité ne soit pas décrite par une loi gaussienne : l'existence de plusieurs sous-populations au sein de la population échantillonnée, la quantité étudiée étant effectivement décrite par une loi gaussienne au sein de chaque sous-famille, ces gaussiennes n'ayant pas les mêmes paramètres d'une famille à l'autre.

Pour prendre un exemple familier, la répartition de quantités morphologiques telles que le poids ou la taille dans les populations humaines doit clairement être bimodale du fait de l'existence de deux groupes bien distincts quant à leur morphologie : les hommes et les femmes.

Par ailleurs, réinsistons sur le fait que les données que nous avons utilisées dans cet exemple sont des données structurées, car elles correspondent à des mesures qui se succèdent dans le temps. Cette structure n'est pas prise en compte lorsque que l'on ne considère que la loi empirique, et une analyse plus poussée de ces données devrait obligatoirement faire appel aux techniques spécifiques permettant d'analyser

des séries chronologiques – sujet fort intéressant mais qui dépasse le niveau de ce cours.

Voici, juste pour le plaisir, le tracé des durées inter-éruptions dans l'ordre de leur succession (les valeurs successives ont été reliées entre elles par des segments de droite).

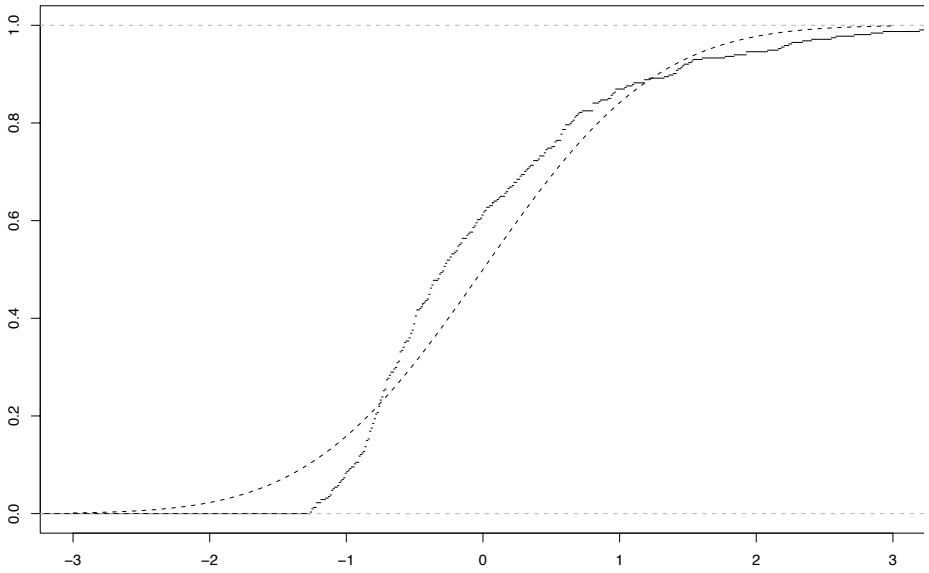


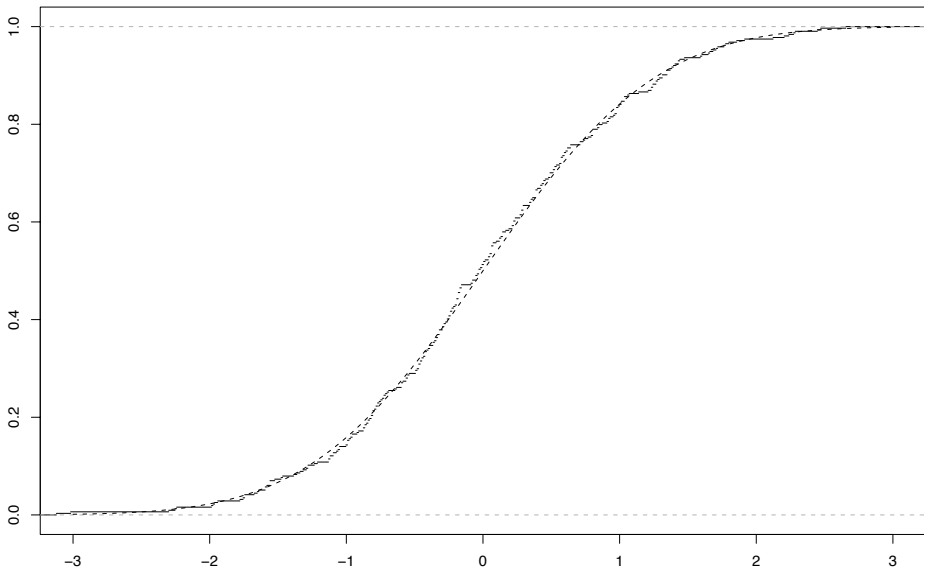
Analyses d'urine

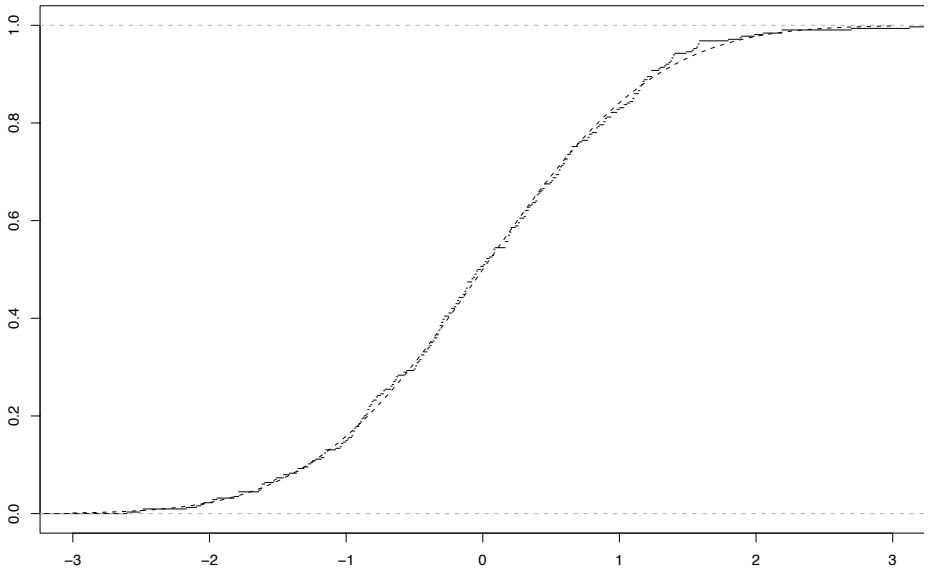
Cette fois, un exemple de mesures de la concentration de GAG (un composé chimique) dans les urines de 314 enfants âgés de 0 à 17 ans (effectuées dans un but d'étalonnage : pouvoir déterminer, par comparaison, le caractère normal ou non d'une concentration mesurée chez un enfant donné.)

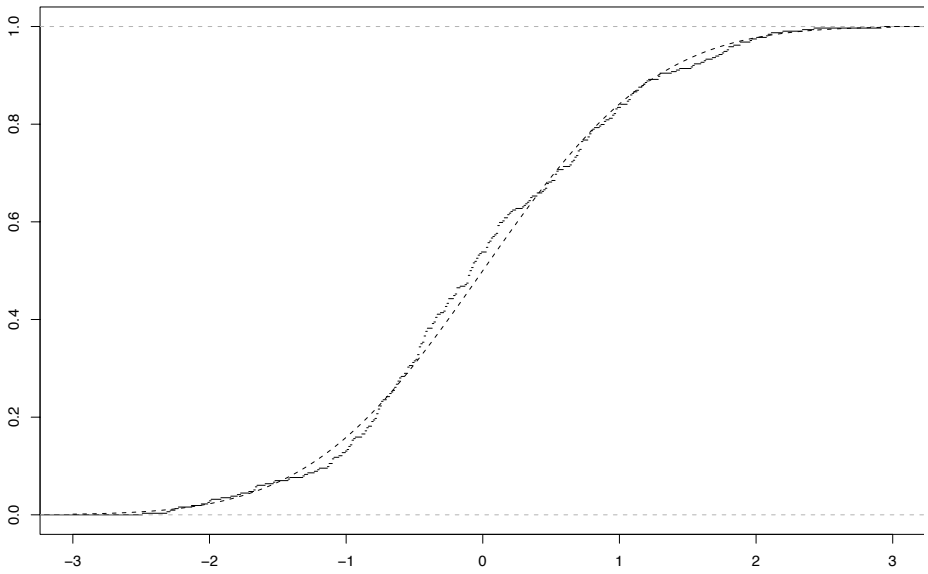
(Ces données proviennent d'un article de S. Prosser repris dans la base de données `MASS` du logiciel `R`).

Le graphique ci-dessous représente la fonction de répartition de la loi empirique de l'échantillon constitué par les 314 mesures, centrée et réduite. En pointillés, la fonction de répartition de la loi gaussienne standard. Suivent trois exemples de graphiques obtenus par simulation de 314 variables gaussiennes arrondies de manière comparable aux données.


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

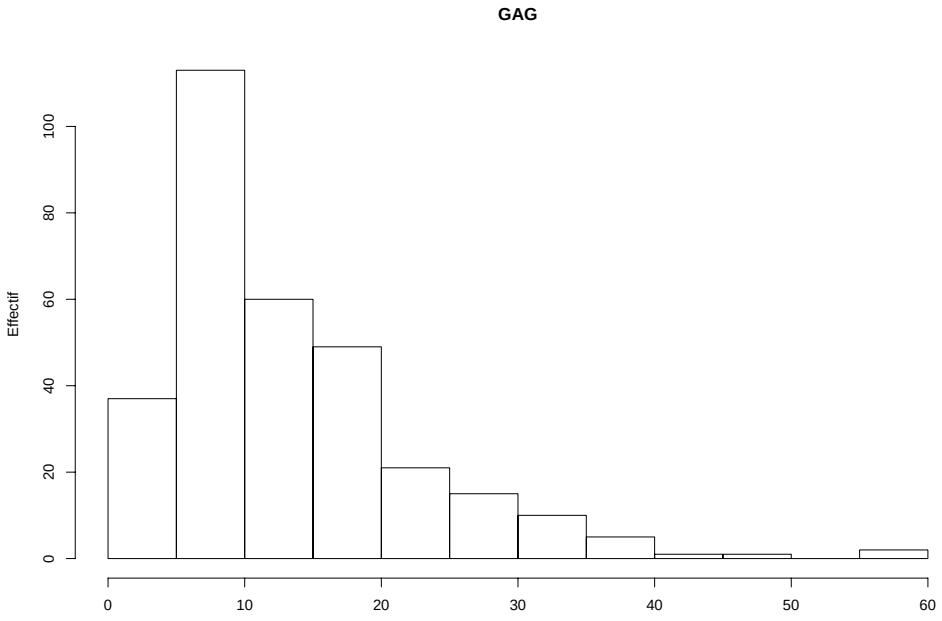

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3



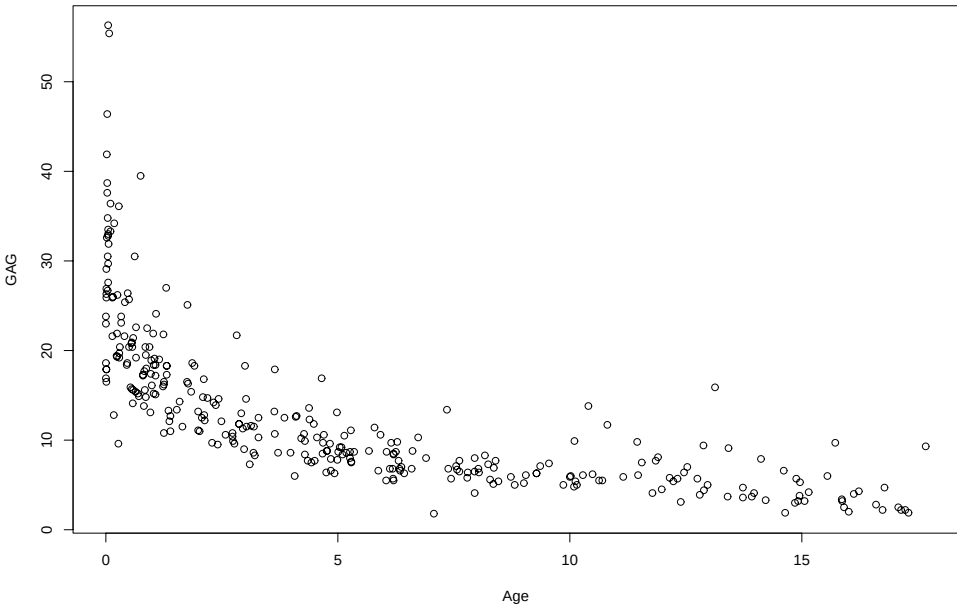


Sans commentaire ! Voici maintenant un histogramme associé aux données, qui

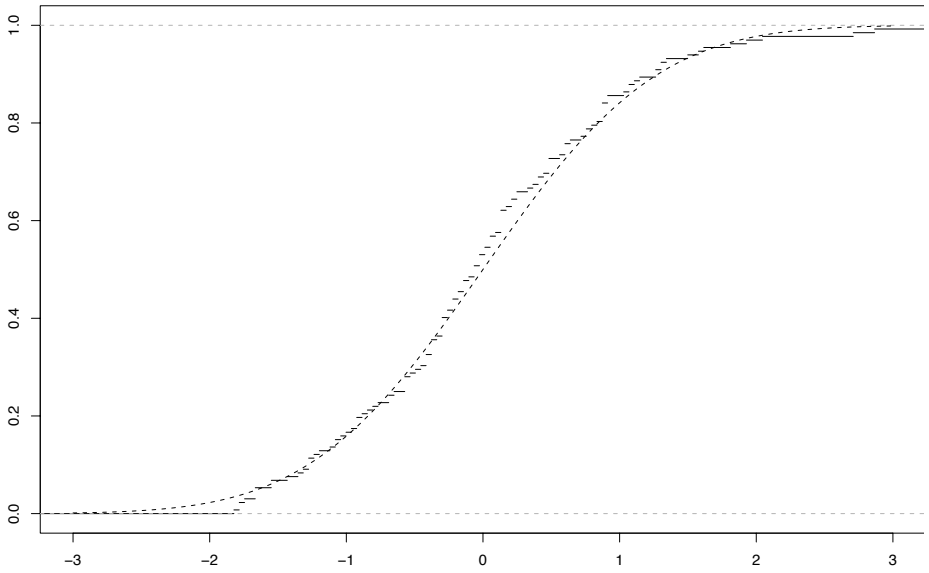
fait très clairement apparaître la violation du caractère symétrique de la loi gaussienne.



Notons que, dans cet exemple, nous disposons de la donnée de l'âge des enfants, en plus de la valeur mesurée de la concentration de GAG, et il existe clairement une forte association entre ces deux quantités, comme le montre le graphique suivant, qui représente les 314 paires (âge en années, concentration en GAG).



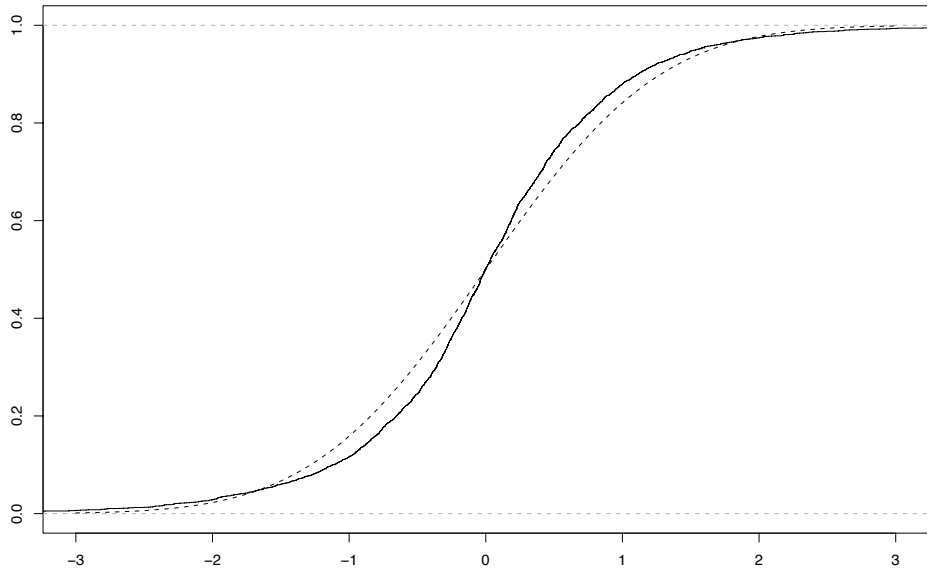
Si l'on se restreint, par exemple, aux 132 mesures de GAG effectuées sur des enfants de strictement plus de 5 ans, pour lesquels une certaine homogénéité dans la distribution de la concentration en GAG est suggérée par le graphique ci-dessus, on obtient le graphique suivant :



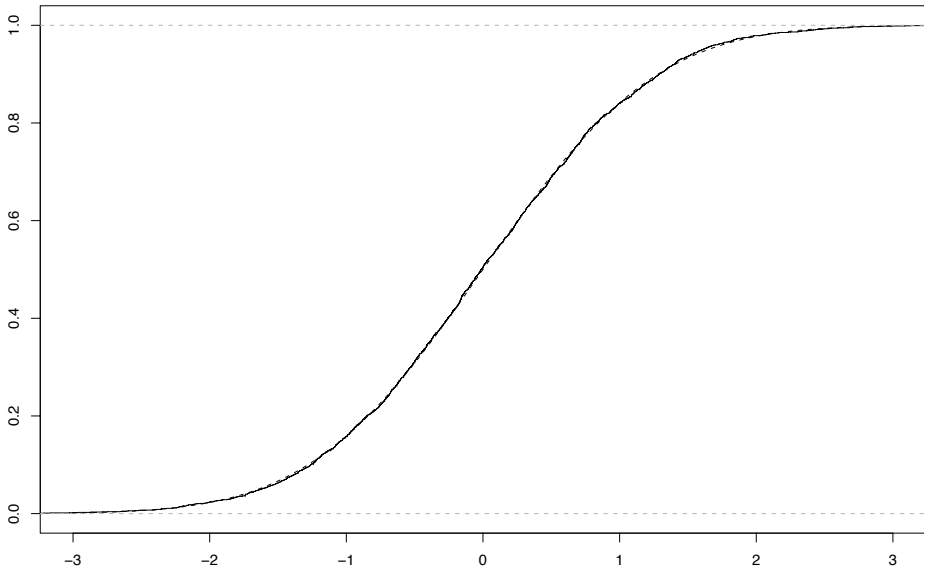
4.4.3 Phynances !

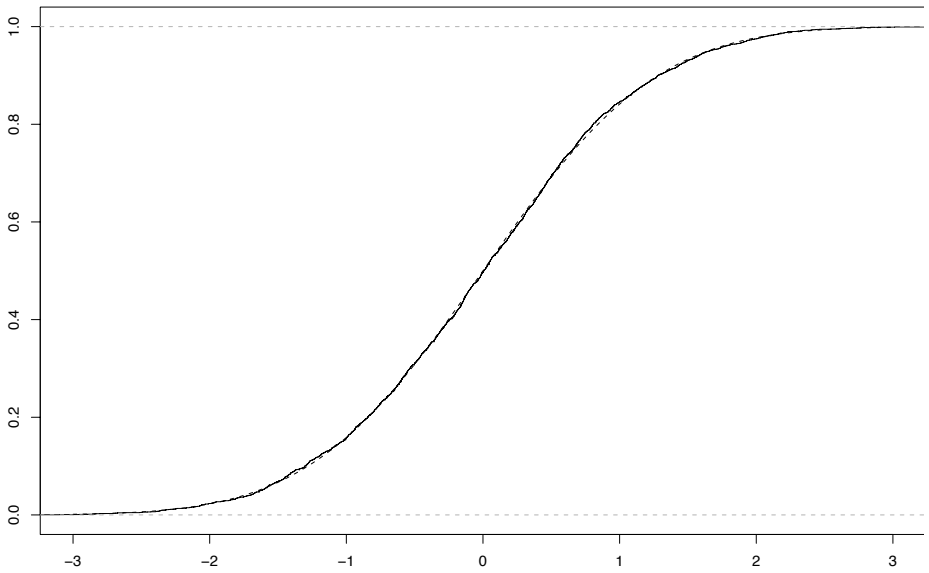
Nous nous attaquons maintenant à une liste de 2780 données correspondant aux variations quotidiennes de l'indice Standard and Poors 500 au cours des années 1990 à 1999 (restreintes au jours d'ouverture des marchés).

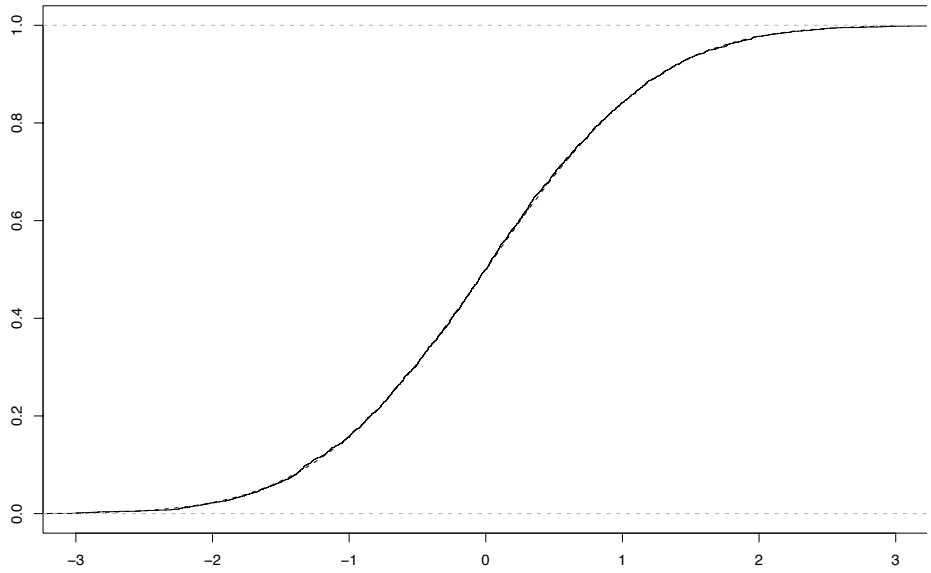
Comme précédemment, la fonction de répartition empirique suivie de trois comparaisons avec des simulations de 2780 variables aléatoires gaussiennes indépendantes (sans arrondi cette fois, les données comportant 9 décimales !).



On note la grande régularité des courbes obtenues, à mettre en rapport avec le grand nombre de données disponibles.

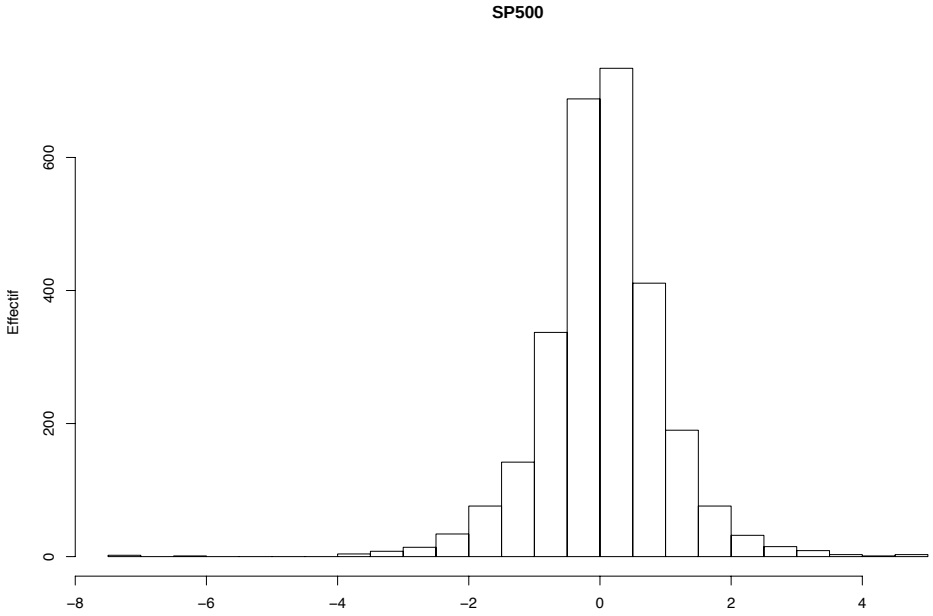

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3



Ici encore, du fait que les données sont structurées en temps, la comparaison avec des gaussiennes simulées indépendantes perd bien entendu un peu de sa pertinence.

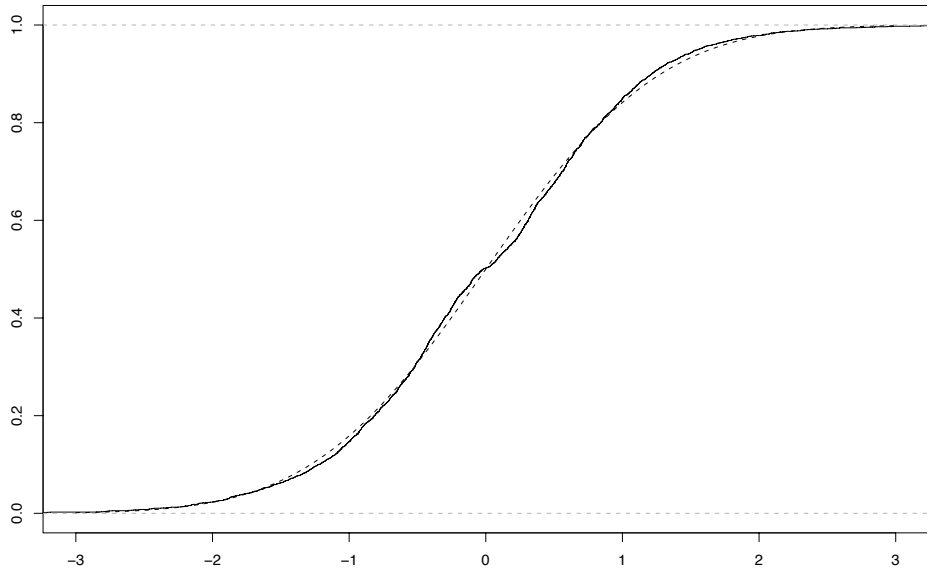
Voici un histogramme associé aux valeurs mesurées.



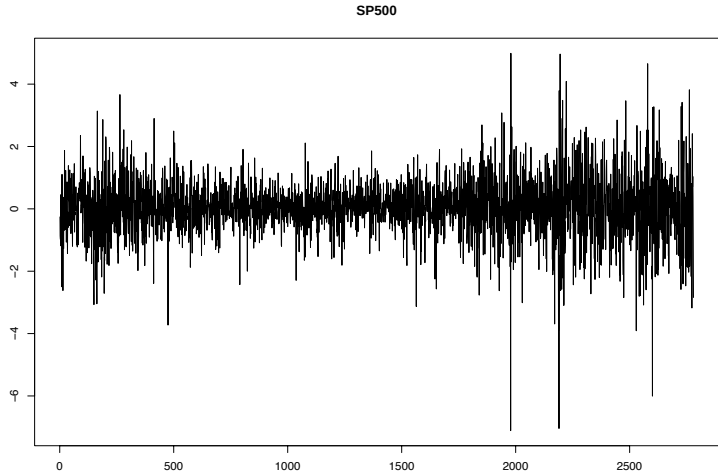
L'histogramme illustre ce que l'on pouvait déjà observer sur le premier graphique, à savoir que, même si la distribution des valeurs, une fois centrée et réduite, est clairement unimodale et symétrique, au moins approximativement, la forme de la «cloche» qui lui est associée ne correspond pas à une courbe gaussienne, mais à une cloche plus resserrée.

Cependant dans ce cas, et c'est une observation d'une portée certaine dans de nombreux exemples de modélisation à partir de données réelles, on peut obtenir une distribution qui se rapproche bien davantage d'une gaussienne en effectuant une transformation non-linéaire bien choisie sur les données. Voici par exemple le graphique obtenu en considérant la loi empirique des données centrées, puis transformées par l'application $x \mapsto \text{signe}(x) \times |x|^{0.76}$, puis réduites. La distribution ainsi obtenue ne s'identifie pas parfaitement à une loi gaussienne, mais n'en est pas si éloignée, surtout si l'on prend en compte le nombre élevé de données disponibles. Dans les faits, il est rare que, lorsque l'on dispose d'un nombre de données suffisant pour estimer très précisément la loi associée aux données, cette loi coïncide exactement avec une gaussienne, même si elle en est très proche (lorsque l'on ne dispose que d'un nombre limité de données, on ne peut estimer très précisément la loi associée aux données, et l'on ne peut donc pas aussi facilement la différencier d'une gaussienne lorsqu'un écart entre ces deux lois existe). Ceci appelle encore une remarque : en pratique, nous ne sommes pas forcément intéressés par le fait que les variables étu-

diées suivent exactement une loi gaussienne, notre but étant plutôt d'utiliser cette loi comme une bonne approximation pour évaluer les quantités (typiquement, des probabilités) d'intérêt. Par conséquent, un écart réel entre la loi gaussienne et les variables étudiées ne signifie pas forcément qu'il faut renoncer à utiliser la loi gaussienne pour décrire celles-ci. Simplement, il faut s'assurer que l'écart est suffisament faible pour que les quantités auxquelles on s'intéresse n'en soient pas significativement affectées. Bien entendu, aucun n'écart n'est faible ou important dans l'absolu : tout dépend des quantités que l'on cherche à calculer à partir d'une approximation par une loi gaussienne. Nous vous renvoyons également à la discussion donnée dans la partie «Attention à l'échelle».



Nous ne résistons pas au plaisir d'ajouter un tracé des données dans l'ordre de leur succession (les valeurs successives ont été reliées entre elles par des segments de droite).



4.5 Quelques applications du TCL

4.5.1 Sondages

Le fait que le théorème de la limite centrale permette de préciser l'ordre de grandeur de l'erreur dans la loi des grands nombres fait qu'il intervient de manière systématique lorsque la loi des grands nombres est employée pour estimer une certaine quantité. Pour prendre un exemple très simple, supposons que l'on sonde la population pour déterminer la proportion p d'individus ayant telle caractéristique particulière (par exemple : la proportion de personnes utilisant de préférence les transports en commun à des moyens de transport individuel pour se rendre sur leur lieu de travail). Un modèle très simple de sondage est le suivant : on interroge N personnes, choisies aléatoirement selon la loi uniforme dans la population étudiée. En appelant X_i la variable aléatoire prenant la valeur 1 lorsque la réponse à la question posée est «oui», et 0 lorsque celle-ci est «non», les variables $X_1, \ldots, X_N$ sont alors des variables aléatoires indépendantes, possédant toutes la loi de Bernoulli de paramètre p.

L'estimation de p obtenue par le sondage est alors égale à $N^{-1}(X_1 + \cdots + X_N)$.

D'après la loi des grands nombres, on s'attend à ce que, typiquement, $N^{-1}(X_1 + \cdots + X_N) \approx p$ lorsque N est grand, et le théorème de la limite centrale affirme que l'erreur d'estimation, soit $N^{-1}(X_1 + \cdots + X_N) - p$, est de la forme $\sqrt{\frac{p(1-p)}{N}} \times \gamma_N$, où γ_N suit approximativement une loi gaussienne centrée résuite lorsque N est grand. En supposant par exemple que $p = 1/2$ et $N = 500$, on obtient que $\sqrt{\frac{p(1-p)}{N}} \approx 2,2\%$. Par conséquent, on peut s'attendre à ce que, avec une probabilité d'environ

95%, l'erreur d'estimation soit comprise entre $-4,4\%$ et $4,4\%$. De même, on peut s'attendre à ce que, avec une probabilité de plus de 30%, l'erreur d'approximation soit supérieure (en valeur absolue) à $2,2\%$. (Ce ne sont là que deux exemples particuliers, l'approximation par la loi gaussienne fournissant une approximation de la totalité de la loi de probabilité de l'erreur.)

Si l'on doublait le nombre de personnes interrogées, les intervalles calculés verraient leur taille divisée par $\sqrt{2}$, soit environ 95% de probabilité d'avoir une erreur comprise entre $-3,1\%$ et $3,1\%$, et plus de 30% de probabilité d'avoir une erreur comprise entre $-1,5\%$ et $1,5\%$.

(Au passage, vous pouvez comparer cette incertitude, inévitable du fait du principe même du sondage, avec l'amplitude des variations qui sont systématiquement commentées et interprétées par les médias, dans les sondages d'opinion).

Plusieurs remarques :

- si l'on disposait de plusieurs sondages indépendants, on pourrait obtenir une estimation plus précise en regroupant entre eux les différents résultats obtenus (comment ?) ;
- le calcul ci-dessus fait intervenir de manière cruciale la taille de l'échantillon étudié (N), mais pas la taille de la population sondée ;
- en pratique, les sondages effectués par les instituts menant des études d'opinion ne s'accordent pas avec le modèle de tirage aléatoire employé ici (d'autre méthodes, moins difficiles et/ou moins coûteuses à mettre en pratique sont employées, telle la méthode des quotas, qui consiste à s'assurer que l'échantillon utilisé comprend des quotas d'individus possédant diverses caractéristiques, par exemple : quotas d'hommes et de femmes, de travailleurs salariés, de personnes de plus de 60 ans, etc...), et le calcul d'erreur que nous avons mené ici ne s'applique donc pas directement ; cependant, l'échantillonnage aléatoire est, en gros, le seul pour lequel on puisse obtenir des estimations rigoureuses de l'erreur ;
- nous avons supposé ici un sondage avec uniquement deux réponses possibles (oui ou non), mais la méthode peut bien entendu se généraliser à un sondage comportant un nombre quelconque de modalités de réponse ; soulignons que la méthode qui consisterait à ignorer les non-réponses pour ne considérer dans l'estimation fournie que les réponses effectivement formulées est grossièrement erronnée, car elle ignore la dépendance pouvant exister entre le fait de ne pas répondre et la caractéristique étudiée par le sondage, ce qui fait que l'on ne peut plus utiliser le modèle selon lequel les réponses utilisées pour l'estimation sont décrites par des variables aléatoires indépendantes de Bernoulli de paramètre égal à la proportion recherchée.

4.5.2 Méthodes de Monte-Carlo

Pas encore écrit ici...

4.6 Lois gaussiennes multidimensionnelles – Vecteurs aléatoires gaussiens

A faire

4.6.1 Vecteurs gaussiens et régression linéaire

Exemple historique de Galton.

4.6.2 Le principe du test du chi-deux

4.7 Exercices

Exercice 162 *(Marche au hasard)*

Un ivrogne se promène en titubant dans une ruelle étroite...

1) On modélise ses déplacements de la manière suivante : chaque pas est effectué vers l'avant avec probabilité $1/2$*, vers l'arrière avec probabilité* $1/2$*, indépendamment des autres pas, et l'on suppose que la taille des pas est constante (par exemple 80cm). Que pouvez vous dire de la position de l'ivrogne après un grand nombre de pas ? À quelle distance se trouve-t-il de son point de départ ?*

2) On suppose à présent qu'un vent violent balaye la rue, soufflant toujours dans la même direction, ce qui fait que la probabilité d'effectuer un pas contre le vent est maintenant de $0,4$*, et celle d'effectuer un pas dans le sens du vent est de* $0,6$*. Comment le résultat précédent est-il modifié ?*

Exercice 163 *La compagnie aérienne Air-Jojo pratique, comme nombre de ses concurrentes, la surréservation, c'est-à-dire que, pour un vol donné, le nombre de places vendues est supérieur au nombre total de places disponibles dans l'avion, la compagnie comptant sur le fait qu'un certain nombre de passagers annulent finalement leur départ, et souhaitant remplir au maximum ses avions. En supposant par exemple qu'un vol dispose de 300 places, et que chaque passager a, indépendamment des autres, une probabilité de* $0,1$ *d'annuler son départ, pouvez-vous estimer le nombre maximum* K *de places que la compagnie peut vendre pour que le nombre de passagers présents au départ de l'avion soit inférieur ou égal au nombre total de places disponibles avec une probabilité de plus de 90%. Quelle est alors la probabilité que plus de 10 passagers ne puissent pas monter dans l'avion ?*

Exercice 164 *Des bits d'information sont transmis le long d'une ligne téléphonique, chaque bit ayant une (faible) probabilité p d'être mal transmis et inversé, indépendament des autres. Si le nombre total de bits transmis est N, quelle est la loi de la variable aléatoire X comptant le nombre de bits mal transmis ? Que peut-on dire de la loi de X lorsque N est grand ? Qu'en est-il dans les exemples suivants :*

- $N = 10^6$ *et* $p = 1/10$ *;*
- $N = 10^7$ *et* $p = 1/100$ *;*
- $N = 10^6$ *et* $p = 10^{-6}$ *;*
- $N = 10^6$ *et* $p = 10^{-7}$ *;*
- $N = 10$ *et* $p = 1/10$ *;*
- $N = 10$ *et* $p = 10^{-6}$ *;*
- $N = 100$ *et* $p = 1/10$.

Exercice 165 *On effectue des lancers avec une pièce de monnaie, supposée honnête. Appelons X le nombre de face obtenu après 1000 lancers. Quelle doit être approximativement la valeur de $X/1000$. À quel écart par rapport à cette valeur peut-on s'attendre ?*

Exercice 166 *On considère une variable aléatoire $Y = (X_1, \ldots, X_d)$ sur $\mathbb{R}^d$, où $d \geq 2$, possédant une loi continue définie par une densité $f : \mathbb{R}^d \to \mathbb{R}_+$. On suppose que les coordonnées $X_1, \ldots, X_d$ de Y forment une famille de variables aléatoires mutuellement indépendantes, que f est continue et peut se mettre sous la forme $f(x) = g(||x||)$. Dans quel contexte ces hypothèses peuvent-elles intervenir ? Montrer que, sous ces hypothèses, les X_i sont chacune distribuées selon une même loi gaussienne, vérifiant en outre $m = 0$.*

Exercice 167 *Si S est un sous-ensemble fini de $\mathbb{R}$, quelle est la loi de probabilité sur S possédant la plus grande entropie ? Même question en se restreignant aux lois de probabilité possédant une espérance nulle et une variance unité. En déduire (au moins heuristiquement) une caractérisation de la loi gaussienne.*

Exercice 168 *Supposons que X suive une loi de Bernoulli de paramètre $p = 1/2$, et $X_1, \ldots, X_N$ des variables aléatoires indépendantes de même loi que X. Pour N fixé, que pouvez-vous dire du comportement de $\mathbb{P}^{\otimes N}(\gamma_N \leq -\epsilon)$ lorsque ϵ tend vers zéro. Même question avec $\mathbb{P}^{\otimes N}(\gamma_N \geq \epsilon)$? Pouvez-vous donner une borne inférieure sur $\left|\mathbb{P}^{\otimes N}(\gamma_N \leq -\epsilon) - \int_{-\infty}^{-\epsilon} \phi_{0,1}(u)du\right|$? Comparez celle-ci avec la borne supérieure fournie par l'inégalité de Berry-Esséen.*

Exercice 169 *Montrez sans calcul, mais en vous appuyant sur le théorème de la limite centrale, que la somme de deux variables aléatoires indépendantes et suivant chacune une loi gaussienne, possède elle-même une loi gaussienne.*

Exercice 170 *(La taille de l'empereur de Chine) Il était une fois... un tailleur ayant eu l'honneur d'être choisi pour confectionner un habit destiné à l'empereur de Chine. Seul problème : pour des raisons d'étiquette, il était absolument impossible que l'empereur se laisse mesurer par quiconque, et encore moins par un tailleur. La solution choisie fut la suivante : plutôt que de mesurer directement l'empereur, on demanda à un grand nombre de ses sujets quelle était la taille qu'ils estimaient être celle de l'empereur, et l'on prit la moyenne des réponses obtenues. Un modèle simple et classique (signal + bruit gaussien centré) pourrait être le suivant : la taille de l'empereur estimée par une personne donnée est égale à la véritable taille de l'empereur, plus une erreur dont la loi est supposée gaussienne, d'espérance nulle, et de variance v inconnue.*

En supposant que $\sqrt{(v)} = 10cm$ *et que l'on interroge 100 millions de personnes, quelle est la précision avec laquelle on peut connaître la taille de l'empereur ?*

Ce résultat vous semble-t-il pertinent ?

Exercice 171 *Revenons sur le problème de Galton (voir Exercice 141).*

Il apparaît que les couples (taille du père, taille du fils) considérés par Galton pouvaient être décrits par une loi gaussienne bidimensionnelle avec une covariance non-nulle.

- ellipse (interprétation géométrique de la loi gaussienne bidimensionnelle)

- pourquoi une telle loi pourrait-elle apparaître dans ce domaine

- en quoi ceci éclaire-t-il les propriétés constatées dans l'exemple simulé ?

Chapitre 5

Bibliographie

Cette bibliographie compte plusieurs types d'entrées : les ouvrages dont la lecture est recommandée pour travailler ce cours (ouvrages d'introduction et/ou de vulgarisation), les ouvrages ou articles de référence, plus spécialisés, cités sur des points précis en rapport avec le cours, et/ou pouvant être utilisés pour un vaste approfondissement, et enfin les ouvrages n'appartenant pas aux deux catégories précédentes, mais néanmoins utilisés pour élaborer le cours.

Nous citons entre autres parmi les références quelques bons ouvrages d'introduction à la théorie mathématique des probabilités, sans prétention à l'exhaustivité.

5.1 Ouvrages recommandés pour travailler ce cours.

H. Tijms. "Understanding probability".

Scientific reasoning : the bayesian approach, C. Howson et P. Urbach. Open Court, 1993. (*Ouvrage sur l'approche bayésienne et le raisonnement en univers incertain.*)
L'ouverture au probable, I. Hacking, M. Dufour, Armand Colin, 2001. (*Ouvrage traitant des différents aspects du raisonnement probabiliste.*)
Chemins de l'aléatoire, D. Dacunha-Castelle. Flammarion, 1996. (*Ouvrage de vulgarisation sur les probabilité et leurs applications.*)
Hasard et chaos, D. Ruelle. Odile Jacob, 1991. (*Ouvrage de vulgarisation davantage tourné vers la physique.*)
Le Jeu de la science et du hasard : la statistique et le vivant, D. Schwartz. Flammarion, 1999. (*Comme son nom l'indique, ouvrage de vulgarisation plus concerné par les applications au vivant.*)
Probabilités, N. Boccara. Ellipses, 1998. (*Ouvrage d'introduction aux probabilités.*)
Une initiation aux probabilités, R. Isaac. Vuibert, 2005 (*Ouvrage d'introduction aux probabilités.*)

Contes et décomptes de la statistique, C. Robert. Vuibert, 2003. (*Ouvrage d'introduction à la statistique.*)
How to lie with Statistics, D. Huff. W. W. Norton, 1993. (*Comme son nom l'indique... Un must..*)
Flaws and Fallacies in Statistical Thinking, S. Campbell. Dover publications, 2004. (*Les pièges des statistiques : exemples concrets.*)
Introduction to Probability, C.M. Grinstead et J.L. Snell. Disponible à l'adresse `http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html` (*Excellent ouvrage d'introduction aux probabilités, dans l'esprit de ce cours.*)
An introduction to probabilistic modeling, P. Brémaud. Springer. (*Ouvrage d'introduction aux probabilités, plus mathématique que les précédents.*)
Chance News, Adresse : `http://www.dartmouth.edu/~chance/chance_news/news.html` (*Un site consacré au raisonnement probabiliste. Innombrables exemples concrets..*)
En passant par hasard... Les probabilités de tous les jours, G. Pagès, C. Bouzitat, F. Carrance, F. Petit. Vuibert. (*Des exemples concrets, étudiés à la lumière de modèles probabilistes simples et traités de manière détaillée.*)
Modèles aléatoires. Applications aux sciences de l'ingénieur et du vivant, J.-F. Delmas, B. Jourdain. Springer. (*Des exemples d'utilisation de modèles probabilistes dans des applications d'un niveau plus avancé que celles abordées dans ce cours, traités de manière détaillée.*)

5.2 Ouvrages et articles de référence.

Probability, A.N. Shiryaev. Springer, 1996. (*Un ouvrage classique d'introduction à la théorie mathématique des probabilités.*)

An Introduction to Probability Theory and Its Applications. Vol. 1–2, W. Feller. Wiley. (*Un double ouvrage classique d'introduction à la théorie mathématique des probabilités.*)

Probability and measure, P. Billingsley. Wiley. (*Un ouvrage classique d'introduction à la théorie mathématique des probabilités.*)

Probability : Theory and Examples, R. Durrett. Duxbury Press. (*Un ouvrage classique d'introduction à la théorie mathématique des probabilités.*)

Probability and Random Processes, G.R. Grimmett et D.R. Stirzaker.Clarendon Press. (*Un ouvrage classique d'introduction à la théorie mathématique des probabilités.*)

Probability with martingales, D. Williams. Cambridge Mathematical Textbooks. (*Un ouvrage classique d'introduction à la théorie mathématique des probabilités.*)

Probabilités en vue des applications (2 tomes), V. Girardin et N. Limnios, Vuibert. (*Un ouvrage en langue française d'introduction à la théorie mathématique*

des probabilités et de la statistique.)

Simulation modeling and analysis, M. Law, W. Kelton. Mc Graw Hill, 2000. (*Un ouvrage classique sur la simulation et la modélisation.*)

Randomized Algorithms, R. Motwani et P. Raghavan, Cambridge University Press, 1995. (*Une référence excellente et accessible sur l'algorithmique randomisée.*)

The art of computer programming. Vol. 2 : Seminumerical algorithms, D. Knuth. Addison-Wesley, 1998. (*Une référence incontournable en la matière (en plusieurs volumes). Si vous n'avez jamais feuilleté ces ouvrages, il n'est que temps de le faire !*)

Constructing a logic of plausible inference : a guide to Cox's Theorem, K.S. Van Horn, International Journal of Approximate Reasoning 34, no. 1 (Sept. 2003), pp. 3-24. Disponible à l'adresse `http://leuther-analytics.com/bayes/papers.html` (*Un article expliquant comment retrouver les règles du calcul des probabilités à partir de considérations très générales sur la cohérence du raisonnement en univers incertain.* .)

Sum the odds to one and stop, Thomas Bruss, The Annals of Probability 28, no. 3 (2000), pp. 1384-1391. (*Un article décrivant la solution d'une classe générale de problèmes d'arrêt optimal.*)

Dynamical Bias in the Coin Toss, P. Diaconis, S. Holmes, R. Montgomery, SIAM Review 49, no. 2 (2007), pp. 211-235. Cet article est disponible à l'adresse `http://stat.stanford.edu/~cgates/PERSI/papers/headswithJ.pdf`

Dynamics of coin tossing is predictable, J. Strzałko, J. Grabski, A. Stefański, P. Perlikowski, T. Kapitaniak, Physics Reports 469 (2008), pp. 59-92. (*Deux articles étudiant le processus physique du lancer de pièce de monnaie.*)

Le suffrage universel inachevé, M. Balinski. Belin, 2004. (*Un ouvrage sur les questions d'arithmétique électorale.*)

Initiation à la physique quantique, V. Scarani. Vuibert, 2003. (*Comme son nom l'indique. Sans formalisme.*)

The MacTutor History of Mathematics archive, L'une des adresses est `http://www-groups.dcs.st-and.ac.uk/~history/` (*Un site de référence pour, entre autres, les biographies des mathématiciens célèbres rencontrés dans ce cours.*)

The first-digit phenomenon, T. Hill, American Scientist, 86, 358–363, 1998. Disponible à l'adresse `http://www.math.gatech.edu/~hill/publications/cv.dir/1st-fig.pdf` (*Un article de référence sur la loi de Benford.*)

Elements of Information Theory, T. Cover et J. Thomas, Wiley, 1991. (*Un ouvrage de référence sur la théorie de l'information.*)

Game Theory Text, T. Ferguson. Ouvrage en ligne disponible à l'adresse `http://www.math.ucla.edu/~tom/Game_Theory/Contents.html` (*Un cours d'introduction à la théorie des jeux.*)

The Elements of Statistical Learning, T. Hastie, R. Tibshirani et J. Friedman, Springer, 2001. (*Un ouvrage de référence sur l'apprentissage et la régression statistiques.*)

Judgment Under Uncertainty : Heuristics and Biases, D. Kahneman, P. Slovic, A. Tversky, Cambdrige University Press, 1982.

Choices, Values, and Frames. D. Kahneman, A. Tversky, Cambridge University Press, 2000. (*Deux ouvrages de références sur les biais psychologiques affectant, entre autres, notre perception intuitive des probabilités.*)

From association to causation via regression, D. Freedman. Technical Report No. 408, Statistics Department, Univ. of California, Berkeley, 1994. (*Un article de réflexion critique sur l'utilisation de la régression pour inférer des relations de cause à effet.*)