UNIVERSITÉ DE LIÈGE
FACULTÉ DES SCIENCES APPLIQUÉES

ELÉMENTS DU CALCUL DES PROBABILITÉS

Louis WEHENKEL

Janvier 2012. Version provisoire.

Table des matières

I Syllabus

Avant-propos et remerciements

Le calcul des probabilités, tel qu'il s'est cristallisé au fil des siècles, représente indéniablement une avancée remarquable de la science, tant sur le plan conceptuel et mathématique que du point de vue de son impact dans les applications. La nécessité de son enseignement dans la plupart des cursus universitaires est incontestée.

Comme la matière est subtile, et parfois contre-intuitive au premier abord, un premier cours sur ce sujet doit à la fois en introduire les notions fondamentales de façon rigoureuse et susciter l'intérêt en montrant combien ces notions sont riches et puissantes pour résoudre une gamme très large de problèmes.

C'est dans cette optique que ces notes ont été rédigées à destination des étudiants bacheliers ingénieurs et informaticiens. Le choix de la matière et le style de présentation résulte de discussions avec mes collègues responsables de l'enseignement de matières qui font appel au calcul des probabilités et de mes propres expériences au cours des années passées dans le cadre de mes activités d'enseignement et de recherche qui concernent principalement les applications du raisonnement probabiliste aux problèmes d'ingénierie électrique, en informatique, et dans les applications biomédicales.

La structure de ces notes est directement inspirée de la première partie du livre de Gilbert Saporta [Sap90] que je recommande vivement comme ouvrage de référence; il comporte, outre la présentation des bases du calcul de probabilités, une excellente présentation de la statistique et des outils d'analyse de données modernes. Je recommande aussi, en seconde lecture, l'ouvrage de David Williams [Wil01] qui est à la fois d'une très grande finesse dans sa façon de présenter les notions fondamentales du calcul de probabilités et un exemple à suivre en matière de pédagogie.

Lors de la préparation de ces notes j'ai énormément bénéficié de l'aide enthousiaste et du regard critique de plusieurs personnes : je remercie en ordre aléatoire François Schnitzler, Alejandro Marcos-Alvarez, Hélène Huaux, Kristel Van Steen, Pierre Lousberg, François Van Lishout, Arnaud Joly, et Didier Vigneron. Au fil des ans, mes réflexions ont aussi bénéficié de nombreuses discussions avec mes collègues et étudiants de l'Université de Liège, et avec mes collaborateurs scientifiques extérieurs; je les remercie de façon collective.

Louis Wehenkel

Janvier 2012

Guide de lecture

Symbole • : Certaines notions introduites dans ce syllabus sont plus ardues ou plus abstraites du point de vue mathématique, mais néanmoins très importantes pour la bonne compréhension du calcul de probabilités. Nous avons mis en évidence les sections qui introduisent ces notions en précédant leur intitulé par le symbole • . Nous conseillons au lecteur de lire attentivement ces sections lors d'un premier passage dans les notes, puis de leur consacrer une seconde lecture après avoir assimilé la suite.

Symbole ◦ : Certaines sections introduisent des notions fort utiles pour les applications ultérieures du calcul de probabilités (statistique, processus stochastiques), mais qui sont peu ou pas utilisées dans la suite de ce cours. Nous avons mis en évidence ces sections en précédant leur intitulé par le symbole ◦ . La lecture détaillée de ces sections peut donc être postposée au moment où ces notions se révèlent intéressantes.

Exemples et applications : Pour ce qui est de l'illustration et de l'applications des notions, nous faisons appel tour à tour à des exemples 'académiques' (pile-ou-face, lancer de dé, etc.) et à des exemples d'applications réelles en ingénierie. Cependant, les illustrations et applications de certaines idées sont réservées aux séances de répétitions et de travaux pratiques, ce que nous indiquons alors au lecteur là où il nous a semblé utile de le faire.

Notes de bas de page : Contrairement à un usage fréquemment adopté où on reproduit les notes en bas de page, nous avons adopté un style qui renvoie ces notes à la fin du chapitre courant. Dans le texte principal ces notes sont marquées par une indication de la forme $^{(i)}$; elles peuvent être ignorées en première lecture.

Références bibliographiques : Nous citons un nombre réduit d'ouvrages généraux dans lesquels on pourra trouver des compléments et des notions plus avancées, ainsi que des références plus circonstanciées aux travaux de base qui ont conduit au développement actuel du calcul de probabilités.

Remarques sur les notations utilisées dans ces notes

Le calcul des probabilités manipule un ensemble riche de notions, comme on le verra. Nous avons essayé dans ces notes d'être le plus précis et le plus cohérent possible dans l'utilisation des notations mathématiques. Les notations nouvelles introduites dans ce cours sont définies au fur et à mesure de leur apparition.

Notation logique vs notation ensembliste : Lorsqu'il est nécessaire d'en faire la distinction, nous utilisons la notation "$\neg A$" pour désigner la négation d'une proposition logique A et la notation "A^c" pour désigner le complémentaire $A \setminus \Omega$ d'un sous-ensemble A d'un ensemble de référence Ω. [(1)] Pour désigner l'intersection (respectivement l'union) d'ensembles nous utilisons la notation "$A \cap B$" (respectivement "$A \cup B$"), et pour désigner la conjonction (respectivement la disjonction) entre propositions logiques nous utiliserons la notation "$A \wedge B$" (respectivement "$A \vee B$"). Dans la plupart des cas, on peut cependant passer sans difficultés d'une formulation ensembliste à une formulation logique. Aussi, pour alléger les notations lorsque le contexte ne prête pas à confusion, nous utilisons souvent des écritures logiques du type "$f(\omega) \in X$" (ou "$f(\omega) = x_i$") à la place de leur écriture ensembliste "$\{\omega \in \Omega : f(\omega) \in X\}$" (ou "$\{\omega \in \Omega : f(\omega) = x_i\}$"). De façon symétrique, nous utilisons aussi parfois la notation ensembliste "A" à la place de la notation logique "$\omega \in A$".

Mesures, lois, fonctions de répartition, et densités de probabilité : Nous utilisons la notation P_Ω (ou P, quand le contexte ne prête pas à confusion) pour désigner une *mesure* de probabilité définie sur un ensemble Ω; cette mesure associe à tout sous-ensemble *mesurable* de Ω sa probabilité. Pour une variable aléatoire $\mathcal{X}$ définie sur Ω et à valeurs dans un ensemble $\Omega_\mathcal{X}$ nous utilisons la notation $P_\mathcal{X}$ pour désigner la *mesure* de probabilité qu'elle induit sur $\Omega_\mathcal{X}$ (c'est-à-dire la fonction qui associe à un sous-ensemble mesurable $X \subset \Omega_\mathcal{X}$ la probabilité d'observer la valeur de la variable aléatoire $\mathcal{X}$ dans cet ensemble). Pour une variable aléatoire discrète nous utilisons aussi $P_\mathcal{X}$ pour désigner sa *loi* de probabilité (c'est à dire la fonction qui associe à une valeur $x \in \Omega_\mathcal{X}$ de la variable aléatoire $\mathcal{X}$ la probabilité d'observer cette valeur). Lorsque la variable aléatoire est à valeurs réelles, nous utilisons $F_\mathcal{X}$ pour désigner sa *fonction de répartition* et, si elle est aussi continue, nous désignons par $f_\mathcal{X}$ sa *densité de probabilité*.

Notes

1. Pour éviter la confusion avec la notion de *fermeture* d'un sous-ensemble d'un espace topologique, nous n'utilisons pas la notation $\bar{A}$ pour désigner le complémentaire de A.

1 INTRODUCTION

"The true logic of this world lies in the calculus of probabilities."
- James Clerk Maxwell, 1831 - 1879

Ce chapitre a pour objectif de présenter notre motivation et notre méthode de travail dans ce cours. Certains termes techniques (par exemple "variable aléatoire", "densité de probabilité", "espérance mathématique", "indépendance conditionnelle", etc.) sont pour le moment utilisés sans en donner la définition précise; les chapitres suivants ont pour objet de faire cela de manière rigoureuse. Nous conseillons donc une première lecture rapide de ce chapitre avant d'aborder la suite, et une lecture plus approfondie après avoir assimilé l'ensemble de la matière du cours.

NB: Une indication du type $^{(i)}$ renvoie à la section des **Notes** située en fin de chapitre.

1.1 MOTIVATION

Si on s'interroge sur le rôle des futurs ingénieurs, on se rend compte qu'on leur demande un esprit critique et une capacité d'innovation de plus en plus grande, pour traiter des problèmes de plus en plus complexes. Les cursus traditionnels de formation des ingénieurs, mettent d'abord en avant la maîtrise du raisonnement *déductif* $^{(1)}$, la connaissance des "lois" générales de la physique, et la mise en oeuvre des méthodes mathématiques et numériques pour la résolution de problèmes techniques, à partir d'un modèle. Dans ce contexte, l'enseignement des *méthodes stochastiques* $^{(2)}$ a pour premier objectif de renforcer la capacité de raisonnement *inductif* $^{(3)}$, c'est-à-dire la capacité d'exploiter des données issues de l'observation d'un système et d'en tirer des conclusions intéressantes sur la nature et le comportement de ce système, afin d'en construire des modèles qui se prêtent à la mise en oeuvre des techniques déductives mathématiques et numériques.

On fait appel aux méthodes stochastiques lorsqu'on est en présence de phénomènes qu'il n'est pas possible ou peu pratique d'étudier de façon détaillée et déterministe. C'est notamment le cas lorsque les systèmes étudiés présentent une très grande complexité, ou lorsqu'on ne dispose que d'une connaissance partielle de leurs caractéristiques. Les méthodes stochastiques permettent alors d'étudier les comportements en moyenne, en modélisant de façon probabiliste les parties d'un système qu'on ne souhaite pas ou qu'on ne peut pas décrire en détails. Les méthodes stochastiques fournissent les outils nécessaires pour déduire les distributions de probabilités des grandeurs de sortie importantes, en fonction de celles des entrées et du modèle (déterministe ou non) du système. Elles permettent ensuite d'utiliser au mieux ces informations pour prendre des décisions appropriées.

Traditionnellement, trop peu de place était laissée aux méthodes stochastiques dans l'enseignement des ingénieurs. Ces méthodes sont cependant utilisées aujourd'hui dans toutes les disciplines scientifiques et techniques, et cela pour trois raisons principales : (i) la nécessité de traiter des systèmes de plus en plus complexes, décrits souvent par un nombre énorme de variables, (ii) la difficulté de prédire certains facteurs dimensionnants de manière exacte et/ou déterministe (notamment ceux caractérisant les effets environnementaux), (iii) l'impact très

important des technologies de l'information sur toutes les disciplines, par le fait qu'elles permettent de collecter des quantités énormes de données et offrent des outils efficaces pour les traiter de façon automatique.

Grâce à ces progrès, les méthodes stochastiques permettent en effet d'aborder efficacement la modélisation, l'analyse et la conception de systèmes de plus en plus complexes, comme ceux rencontrés en informatique, électricité, mécanique, chimie, aéronautique, etc. Ces méthodes jouent également un rôle croissant en économie et en finance, et dans les sciences naturelles, domaines auxquels les ingénieurs sont de plus en plus régulièrement confrontés. Par exemple, si on fait un inventaire des travaux de fin d'études des ingénieurs on se rend compte qu'une proportion croissante fait appel de façon directe ou indirecte aux méthodes stochastiques.

Le cours d'*Eléments de Probabilités* est le premier maillon d'une chaîne de trois cours faisant partie du tronc commun du programme de Bachelier Ingénieur civil et du Bachelier en Informatique. Avec les deux autres cours, à savoir respectivement le cours d'*Eléments de Statistique* et le cours d'*Introduction aux Processus Stochastiques*, ce cours a pour objectif de former les étudiants, toutes disciplines confondues, aux bases des méthodes stochastiques. Les cursus de Master Ingénieur civil et de Master en Informatique pourront ainsi s'appuyer sur ces compétences pour aborder des méthodes plus spécialisées faisant appel à ces connaissances de base, en fonction des besoins de chaque discipline.

1.1.1 Exemples de domaines de l'ingénieur faisant appel aux méthodes stochastiques

La **gestion des risques** industriels et technologiques fait appel aux méthodes stochastiques pour l'analyse et la maîtrise de la fiabilité et de la sécurité des systèmes. Les méthodes stochastiques sont notamment utilisées pour la conception des centrales nucléaires, la planification des réseaux d'énergie électrique, l'évaluation des risques des moyens de transport en commun (aéronautique, trains à grande vitesse), la maîtrise de la fiabilité des lanceurs spatiaux... Ces techniques permettent notamment d'identifier les sources de pannes les plus probables et de déterminer des parades à la fois efficaces et aussi économiques que possibles. Notons que l'étude des performances et la vérification des logiciels informatiques fait partie de ce domaine.

Les **arbitrages économiques** entre différents projets techniques visant à résoudre un problème (par exemple, pour justifier la construction d'une nouvelle route, d'une nouvelle usine, ou bien pour décider d'investir dans un changement technologique majeur comme les véhicules électriques ou la production d'énergie électrique éolienne) nécessitent de modéliser des scénarios micro- et macro-économiques sur des horizons de temps de plusieurs années, voire de plusieurs décennies, afin d'évaluer l'espérance mathématique des retours sur investissements. Ces arbitrages sont souvent stratégiques pour permettre l'adaptation aux changements globaux.

La **théorie des systèmes** est une discipline générale qui est utilisée pour l'étude et la conception d'une très grande diversité de systèmes, que ce soit en mécanique, en chimie, en électricité, ou encore en informatique. Une partie de la théorie des systèmes s'intéresse aux systèmes stochastiques qui sont notamment utilisés en estimation d'état et pour la conception de systèmes auto-adaptatifs. L'estimation d'état permet de tirer le meilleur profit des informations fournies par divers capteurs, notamment en filtrant les erreurs de mesures, en permettant la détection de fonctionnements anormaux de certains capteurs, et en utilisant de manière optimale les grandeurs mesurées pour inférer la valeur probable des grandeurs internes non directement mesurables qui caractérisent l'état du processus. Les systèmes auto-adaptatifs sont capables d'adapter leur stratégie de commande en fonction de changements des caractéristiques de l'environnement, du système piloté, et/ou de ses objectifs de réglage. Le **traitement du signal** (traitement de la parole, de signaux physiologiques, d'images) repose en grande partie sur des méthodes stochastiques. En imagerie spatiale, par exemple, ces techniques sont utilisées pour éliminer le bruit des informations brutes captées par les téléscopes et ainsi identifier automatiquement les corps stellaires. En médecine, elles permettent l'interprétation automatique de signaux physiologiques (électrocardiogrammes, électroencéphalogrammes, imagerie fonctionnelle) et facilitent ainsi le diagnostic et le traitement des maladies.

En **informatique** de nombreuses questions font appel aux méthodes stochastiques : l'optimisation des performances des systèmes, la compression de données, l'intelligence artificielle, l'analyse de données... L'analyse de données est par exemple utilisée par les constructeurs automobiles pour détecter les causes de certaines pannes répétitives. La compression de données est basée sur le codage de suites de symboles en fonction de leur probabilité d'apparition, les suites les plus fréquentes recevant les codes les plus courts; elle permet de réduire coûts de stockage et délais de transmission dans de nombreuses applications (stockage de masse, réseaux informatiques, disques compacts, multimédia, télévision digitale...).

L'ingénierie des **procédés chimiques** et de **la production mécanique** repose sur la mise en oeuvre des méthodes stochastiques pour la conception des usines, afin de minimiser les rapports coûts/performances des installations industrielles en fonction des sollicitations probables et des futurs coûts d'exploitation. Ces disciplines font aussi appel aux méthodes stochastiques pour la conception de systèmes de surveillance des installations, qui ont pour but de détecter en temps opportun les risques de pannes et la nécessité de programmer des opérations d'entretien, en fonction des mesures entachées d'erreurs et/ou incomplètes qui peuvent être faites sur ces systèmes.

En **aéronautique** et de façon plus générale dans le domaine des **transports**, les méthodes probabilistes sont utilisées pour contrôler les risques de pannes catastrophiques face aux aléas, notamment les erreurs humaines (de la part des concepteurs, opérateurs et utilisateurs de ces systèmes) et les risques naturels (engendrés par les orages, tempêtes, glissements de terrain etc.), auxquels les systèmes de transport sont régulièrement soumis. Les mêmes méthodes sont utilisées pour concevoir et exploiter les installations industrielles présentant un risque potentiel pour la population en cas d'accident (usines chimiques, centrales nucléaires, industrie du gaz, etc.).

En **géologie** les techniques probabilistes et statistiques sont de plus en plus utilisées pour aider à la localisation des lieux de sondage et de forage les plus prometteurs, en fonction des données partielles qui sont recueillies sur le terrain par des techniques de plus en plus sophistiquées. Dans le domaine de la **construction**, les techniques probabilistes sont utilisées pour dimensionner les structures de ponts face aux sollicitations aléatoires auxquelles ils sont soumis, pour étudier et concevoir de nouveaux revêtements plus performants en présence de pollution ou de variations de température exceptionnelles.

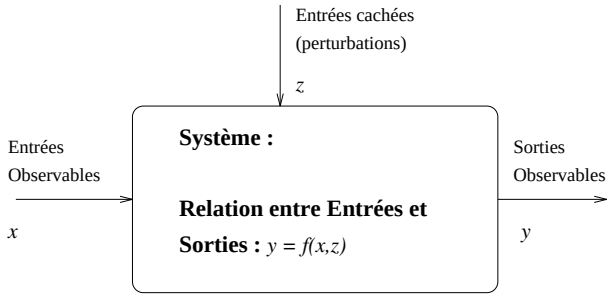
Les **télécommunications** reposent sur les méthodes stochastiques en ce qui concerne l'optimisation des performances, le codage de données et le filtrage du bruit. En particulier, les télécommunications font largement appel au traitement du signal, aux techniques d'optimisation des performances de systèmes informatiques distribués, à la compression et au codage de données. Par exemple, dans un réseau ATM chaque connexion se présente sous la forme d'une suite virtuellement unique de cellules transmises, qui peut être représentée comme la réalisation d'un processus stochastique. Les méthodes stochastiques permettent alors d'étudier les performances du système lorsqu'il est soumis à différents types de trafic (communications téléphoniques, transferts de données numériques, trafic multimédia…). Elles permettent aussi l'optimisation du codage des données dans le but de minimiser les pertes d'informations malgré l'effet du bruit et d'erreurs de transmission.

Dans le domaine de l'**ingénierie biomédicale**, les méthodes stochastiques sont utilisées pour exploiter les données biologiques pour identifier des gènes en relation avec les maladies complexes telles que le cancer, l'asthme et le diabète, pour modéliser les systèmes de régulation de gènes et comprendre la dynamique des systèmes biologiques (développement, différenciation des tissus) et pour concevoir des systèmes d'aide au diagnostic médical (analyse d'images et de données de séquençage, et de protéomique) ainsi que pour la conception de procédés de biotechnologie.

Disciplines méthodologiques de base

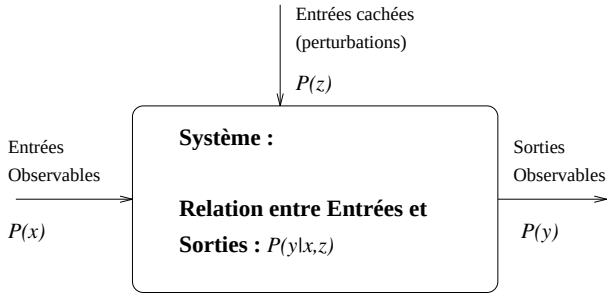
Le domaine des méthodes stochastiques est extrêmement riche. On peut le structurer de la manière suivante:

- **Calcul des probabilités et théorie de l'information :** formalisation du raisonnement sous incertitude, étude théorique de la notion d'incertitude et d'information; étude quantitative des performances de systèmes de collecte, de codage et de communication de l'information.
- **Statistique, analyse de données, et apprentissage automatique :** exploitation optimale de données issues de l'observation expérimentale ou de la simulation numérique pour l'inférence de modèles explicatifs et prédictifs et pour la vérification d'hypothèses.
- **Systèmes et processus stochastiques :** méthodes mathématiques et informatiques de description et de manipulation de signaux spatio-temporels et de systèmes dynamiques, caractérisés essentiellement par un très grand nombre de variables dont l'étude conjointe repose sur l'exploitation de propriétés globales de symétrie et/ou d'invariance, telles que la stationnarité et l'ergodicité.
- **Algorithmique stochastique et optimisation :** conception d'algorithmes pour résoudre des problèmes de codage de données, et d'analyse et d'optimisation de systèmes stochastiques; caractérisation des propriétés de ces algorithmes.



Disons que le système est une voiture : les entrées observables x seraient alors les actions prises par le conducteur (position du volant, de l'accélérateur, du frein, etc.), la sortie y serait la trajectoire du véhicule, et les entrées "cachées" z (i.e. les perturbations non observables) seraient par exemple les forces exercées par la route sur les pneus, et les mouvements des passagers.
La relation déterministe entrées-sorties $f(x, z)$ permet de calculer les sorties si on connait les entrées observables et cachées.

Figure 1.1: Représentation graphique d'un système (vision déterministe)



La loi $P(y|x, z)$ tient compte de la connaissance imparfaite du système (usure des pneus, température du moteur, etc.), la loi $P(z)$ exploite nos connaissances a priori (conditions hivernales, nombre de passagers) quant aux entrées cachées, et la loi $P(x)$ permet de modéliser les erreurs de mesure sur les entrées observables.
Les méthodes probabilistes permettent alors de calculer une loi de probabilité $P(y)$ sur l'ensemble des trajectoires possibles de la voiture étant données ces informations.

Figure 1.2: Représentation graphique d'un système (vision stochastique)

1.1.2 Notion de système stochastique ou de modèle stochastique d'un système

La figure 1.1 représente de façon graphique un système. Une telle représentation est en fait une abstraction (graphique) de la réalité physique; pour en faire l'étude on peut lui associer des objets mathématiques : espace d'entrée; espace de sortie; modèle entrée/sortie (équations différentielles, algébriques...). Si à des entrées fixées le modèle associe de manière unique les sorties, on dira que le modèle (et par extension le système) est déterministe. Un modèle non-déterministe est donc un modèle qui associe à une entrée donnée un ensemble de sorties possibles, c'est-à-dire compatibles avec le modèle. Notons d'emblée que *l'orientation* du modèle, c'est-à-dire le choix de ce qu'on convient d'appeler les entrées et les sorties est effectué en fonction des objectifs poursuivis. En particulier, un modèle déterministe peut devenir non-déterministe si on inverse le rôle joué par les entrées et les sorties. D'autre part, même dans le cas d'un modèle déterministe, il est souvent difficile ou impossible de déterminer toutes les entrées de manière précise (par exemple, certaines entrées sont qualifiées de *perturbations* inconnues); dans ce cas, on cherchera à déterminer les ensembles de sorties possibles, lorsque les entrées connues appartiennent à un sous-ensemble de l'espace d'entrée.

On conçoit que partant d'un système réel, on peut construire une hiérarchie de modèles comprenant des modèles très précis mais extrêmement complexes, voire impossibles à manipuler, ainsi que diverses approximations, plus simples à manipuler mais moins précises. Par exemple, en présence d'un simple circuit électrique (disons une "résistance" en série avec un "condensateur" [(4)], l'ingénieur qui s'intéresse à la relation entre le courant dans ce circuit et la tension à ses bornes dispose de toute une série de modèles : équations de Maxwell aux dérivées partielles, équations différentielles des circuits en éléments condensés (linéaires ou non-linéaires), équations algébriques, relations qualitatives... Même le plus sophistiqué de ces modèles reste une abstraction de la réalité, et possède un domaine de validité restreint, sinon bien cerné. C'est tout l'art du métier d'ingénieur que de choisir le modèle approprié, à la fois suffisamment précis et adapté aux besoins pratiques, et ce choix dépend évidemment du contexte. Par exemple, dans le cas de notre mini-circuit il dépendra de l'espace d'entrées (contenu fréquentiel des signaux à l'entrée, amplitude), ainsi que de la nature de l'environnement (perturbations thermiques ou électromagnétiques) dans lequel le circuit est censé fonctionner. Mais le choix de modélisation dépend également de l'information disponible : par exemple la mise en oeuvre des équations de Maxwell nécessite des informations détaillées sur la géométrie des conducteurs et diélectriques des composants utilisés, informations qui ne sont pas nécessairement disponibles.

La figure 1.2 schématise le point de vue adopté par les méthodes stochastiques pour l'étude des systèmes (en anticipant sur la suite de ces notes). Par rapport à la figure 1.1, ce modèle est complété par des hypothèses sur les distributions de probabilités des entrées $P(x), P(z)$, la relation entrée/sortie est représentée par une distribution conditionnelle $P(y|x,z)$, et les sorties sont caractérisées par une distribution de probabilités $P(y)$. Ces distributions de probabilités représentent un certain niveau de connaissance du système physique modélisé, et englobent comme cas particuliers les systèmes déterministes [5]. L'avantage principal de cette vision des choses est de mettre en évidence explicitement le degré de non-déterminisme des différentes parties du modèle. Les méthodes stochastiques visent à quantifier les incertitudes résiduelles de ces modèles en mettant en oeuvre le calcul des probabilités et les outils statistiques. Elles permettent de construire des modèles probabilistes des systèmes à partir de données mesurées, et ensuite de manipuler ces modèles (analyse mathématique, simulations numériques) pour prendre des décisions appropriées lors de la conception et de l'exploitation de ces systèmes.

1.2 PROBABILITÉS VERSUS STATISTIQUE

La théorie des probabilités est une branche des mathématiques qui vise à étudier les phénomènes aléatoires (c'est-à-dire où le "hasard" intervient). Cette théorie, qui a mis plusieurs siècles à se cristalliser sous sa forme actuelle, permet d'étudier et de résoudre un grand nombre de problèmes pratiques et aussi théoriques. Elle fournit également une approche pour la formalisation du "raisonnement logique" en présence d'informations partielles et/ou contradictoires. Comme toute théorie mathématique, la théorie des probabilités est une science déductive, qui se base sur un certain nombre d'axiomes et utilise les techniques usuelles en mathématiques pour la démonstration de théorèmes. On y déduit donc des propriétés spécifiques, à partir des hypothèses générales incarnées par les axiomes. Ses domaines d'application sont nombreux : la physique, l'intelligence artificielle, la théorie des systèmes, le traitement du signal, la statistique ... pour n'en citer que quelques uns.

La statistique, au sens le plus général, est une discipline qui consiste dans le recueil, le traitement et l'interprétation de données d'observations sur des systèmes physiques (réels ou simulés). En particulier, elle permet de construire des modèles (probabilistes ou non) qui représentent correctement la réalité mesurable du monde physique. Il s'agit d'une discipline faisant souvent appel au raisonnement inductif : à partir d'un certain nombre d'observations élémentaires on cherche à construire des lois générales qui "expliquent" ces observations. Etant donné ce caractère inductif, les résultats obtenus par la statistique peuvent être remis en question par de nouvelles observations, comme c'est le cas dans le domaine des sciences naturelles en général. Pour cette raison, une utilisation correcte de la statistique dans un domaine donné, va nécessairement de pair avec une bonne compréhension physique de ce domaine. Aussi, les résultats obtenus sont justifiés dans la mesure où ils sont opérationnels, et non pas parce qu'ils représenteraient la vérité absolue. Qu'on ne s'y trompe pas cependant, car l'utilisation des outils statistiques fait autant appel à la rigueur scientifique que les autres sciences expérimentales. Néanmoins, ces outils ne permettent de vérifier que la "plausibilité" (et non la "réalité") de la plupart des modèles utilisés par les ingénieurs et scientifiques de nombreuses disciplines. Etant donné la diversité des problèmes rencontrés en pratique, la statistique est un domaine extrêmement vaste dont l'appréhension d'ensemble nécessite du temps et de l'expérience pratique. Elle est surtout basée sur le calcul des probabilités, qui lui sert comme outil de raisonnement; elle fait cependant aussi appel à de nombreuses autres parties des mathématiques (analyse, algèbre, géométrie...). La statistique est aussi intimement liée à la philosophie des sciences, et plus particulièrement à l'étude des mécanismes de découverte et de révision des théories scientifiques.

Il y a donc une interdépendance forte entre probabilités et statistique, mais également une différence fondamentale dans leur approche : déductive pour le calcul des probabilités; inductive pour la statistique.

Ces notes de cours s'intéressent principalement au calcul des probabilités pour en établir les bases théoriques et les résultats les plus fondamentaux qui doivent être maîtrisés avant d'aborder la statistique et les autres méthodes stochastiques. Afin d'illustrer et de motiver notre propos, nous ferons appel à divers exemples pratiques, dont certains seront de nature académique (tel que le "double pile ou face") et d'autres seront des versions simplifiées de problèmes pratiques qui seront étudiés plus en détails dans d'autres cours. Les travaux pratiques viseront à aider à l'assimilation des notions de base, en combinant séances d'exercices et exercices sur ordinateur.

Pour une très bonne introduction aux probabilités et à la statistique nous recommandons vivement [Sap90] dont nous avons adopté la structure logique et la plupart des notations. Pour en savoir plus sur les fondements mathématiques du calcul des probabilités nous suggérons la lecture de la référence [Bil79].

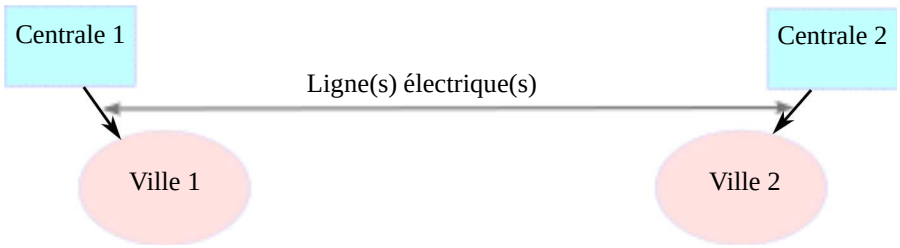


Figure 1.3: Problème de l'alimentation en électricité de deux villes. Le schéma représente de façon simplifiée la structure physique du système étudié

Pour conclure cette discussion, insistons sur le fait que la séparation probabilités/statistique que nous faisons volontairement n'est pas justifiée d'un point de vue fondamental, mais bien pour des raisons pédagogiques. Nous commençons en quelque sorte par analyser quelques arbres sans nous préoccuper de la forêt. Parmi les différentes possibilités qui s'offrent pour aborder un domaine comme celui-ci, le choix que nous avons fait est celui d'une rupture minimale (mais nécessaire) avec la façon habituelle de présenter probabilités et statistique dans un même cours. Nous souhaitons ainsi limiter la confusion entre les aspects déductifs et inductifs complémentaires du calcul des probabilités et de la statistique. Nous espérons qu'au fur et à mesure de sa familiarisation avec les méthodes stochastiques, l'étudiant prendra conscience comment ces deux disciplines couvrent les deux pans du *raisonnement en présence d'informations incomplètes*.

1.3 TROIS PROBLÈMES DE RAISONNEMENT SOUS INCERTITUDE

Dans cette section nous allons introduire trois problèmes caractéristiques de situations où le calcul des probabilités est un outil intéressant pour modéliser et puis résoudre le problème. Pour chaque problème, nous allons d'abord donner une version "intuitive" du problème, ensuite une version plus "mathématique" ou plus "ludique", ensuite quelques exemples de problèmes d'"ingénieurs" collant avec cette abstraction.

Différentes versions de ces problèmes seront exploitées dans la suite du cours pour illustrer l'intérêt des notions théoriques introduites du point de vue des applications.

1.3.1 Problème de prédiction

Il s'agit d'un problème où on cherche à établir les conséquences probables d'un certain nombre de choix, sachant que les conséquences peuvent être influencées par des facteurs exogènes aléatoires.

1.3.1.1 Exemple : évaluation de l'intérêt d'un investissement

Deux villes doivent être alimentées en électricité, grâce à un système électrique composé de deux centrales électriques, respectivement situées dans chacune des deux villes, et reliées par un réseau électrique (une ou plusieurs lignes à haute tension qui permettent d'acheminer de l'électricité d'une ville à l'autre, voir Figure 1.3). Il s'agit de dimensionner les centrales électriques et le réseau de façon à assurer un service de bonne qualité, c'est-à-dire de façon à pouvoir servir la demande en énergie des clients avec une très grande fiabilité et un coût minimal. En particulier, on souhaite savoir si le fait d'investir dans la construction de lignes pour interconnecter les deux villes est une option économiquement rentable.

1.3.1.2 Modélisation

Nous modélisons les aléas du problème à l'aide de deux variables aléatoires, qui représentent respectivement la demande d'électricité de chacune des deux villes (la demande varie d'un moment à l'autre, d'où la nécessité de la décrire par une variable mathématique, et comme ses variations ne sont pas prévisibles a priori, il s'agit de variables *aléatoires*). Ces deux variables aléatoires sont chacune influencées par d'autres variables aléatoires, qui représentent respectivement l'effet de la météo (effets qui sont essentiellement communs aux deux villes, celles-

ci étant supposées se trouver dans une même région géographique) et le choix des individus (indépendants d'un individu à l'autre) d'activer les différents appareils électriques dont ils disposent.

Dans une version simplifiée du problème, nous supposons que le réseau est de capacité non limitée, et totalement fiable, ce qui permet d'alimenter les deux villes grâce à la somme des capacités des deux centrales. Sous cette hypothèse, la capacité totale des centrales doit donc couvrir avec une très grande probabilité la somme des consommations des deux villes, à tout instant.

Le problème revient alors à déterminer la probabilité pour que la somme des consommations des deux villes soit supérieure à cette capacité totale de production, dans le scénario avec interconnexion, puis de comparer cela à la probabilité que pour chaque ville la consommation soit supérieure à la capacité de sa centrale (scénario sans interconnexion), et enfin de traduire ces informations en une recommandation, en prenant en compte le coût de l'interconnexion d'une part et le coût d'interruption de service d'autre part.

Ce problème de base peut évidemment être appliqué à différents scénarios (différentes capacités des centrales, et prise en compte de capacités limitées au niveau du réseau de transport) afin de choisir la meilleure configuration du système. Il est également possible de modéliser les probabilités de défaillance des centrales et des lignes du réseau, pour encore raffiner le modèle. Enfin, en pratique ce scénario se présente généralement dans une version avec n centrales, m villes, et k possibilités d'interconnexion.

1.3.1.3 Abstraction sous forme de problème jouet

Comme nous le verrons dans les chapitres suivants, ce problème revient essentiellement à déterminer la distribution d'une fonction (ici une somme) de variables aléatoires (ici les demandes d'électricité), à évaluer la probabilité pour que la valeur de cette fonction soit supérieure à un seuil (ici la capacité totale de production des deux centrales), puis à calculer l'espérance mathématique des valeurs de cette fonction selon différentes hypothèses (ici pour évaluer la qualité de service, en fonction de la quantité totale d'énergie non desservie).

1.3.1.4 Autres exemples d'applications pratiques

Tous les problèmes d'analyse de risque et d'évaluation des retours sur investissement se déclinent essentiellement de la même manière, avec des complications provenant de la complexité des systèmes étudiés (nombre de composants) et liées au fait qu'on étudie des événement rares mais auxquels sont associés des conséquences graves. On peut citer l'analyse de la fiabilité des centrales nucléaires, des systèmes de transport, et l'évaluation de leur valeur économique future.

1.3.2 Problème de diagnostic

Il s'agit d'un problème de raisonnement où on veut inférer les causes d'un phénomène observé à partir de l'observation de ses conséquences.

1.3.2.1 Exemple : diagnostic d'une panne

Ma voiture ne démarre pas; il y a une manifestement une panne. Je voudrais savoir quelle est l'origine de cette panne pour pouvoir la réparer. Il y a plusieurs origines possibles, certaines a priori plus probables que d'autres. Sachant que certaines origines peuvent sans doute être exclues étant donné les informations dont je dispose, dans le cas particulier qui me préoccupe, quelle est l'origine la plus probable de cette panne ?

1.3.2.2 Modélisation

Nous modélisons notre problème par une série de variables (aléatoires) binaires qui chacune décrivent une origine possible de la panne (batterie déchargée ? réservoir d'essence vide ? démarreur défaillant ? bougies encrassées ? faux contact dans l'électronique ? etc.) et une autre variable qui représente le fait que la voiture ne démarre pas. Nous ajoutons aussi des variables observables facilement, telles que 'la radio fonctionne', 'j'ai fait le plein hier soir' etc., qui représentent les informations dont nous disposons au moment du diagnostic.

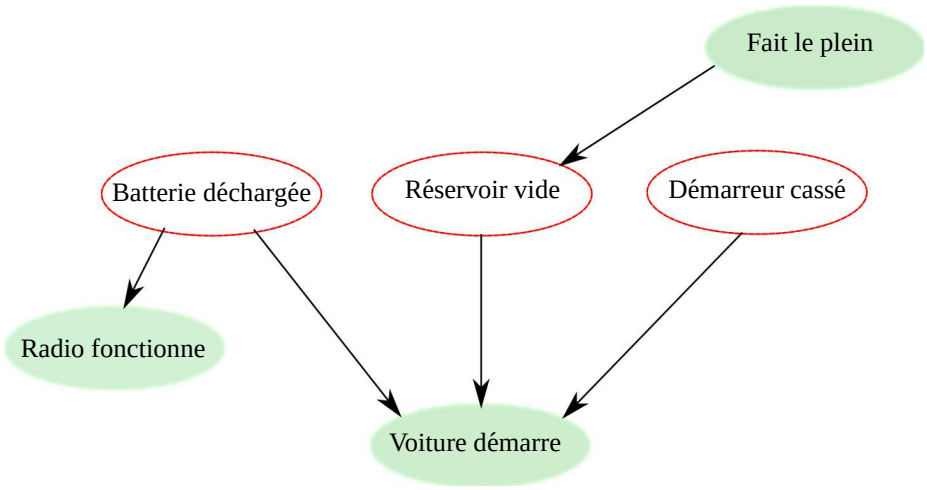


Figure 1.4: Problème de démarrage de la voiture. Le schéma représente les relations logiques entre les variables aléatoires servant à modéliser le problème. En vert plein, les variables observées; en rouge creux, les variables à diagnostiquer.

La variable ‘panne’ est dans notre modèle une fonction (booléenne) des variables ‘cause’ : si au moins une des variables ‘cause’ est vraie, la variable ‘panne’ est aussi ‘vraie’, sinon elle est ‘fausse’.

Nous supposons que les pannes sont indépendantes, et chacune caractérisée par une certaine probabilité a priori connue.

Le graphique de la Figure 1.4 représente intuitivement les relations entre les différentes variables modélisées, une flèche indiquant qu’une variable influence directement la valeur d’une autre.

Nous verrons comment ce modèle permet de répondre à la question posée dans les chapitres qui suivent.

1.3.2.3 Abstraction sous forme de problème jouet

Nous allons analyser ce problème de diagnostic, en nous basant sur un problème jouet, à savoir le double pile ou face. Dans ce problème, on considère une expérience qui consiste à lancer deux pièces (une d’un euro, et une de deux euros) simultanément. On suppose que chaque pièce a une certaine probabilité de tomber sur pile ou sur face, qui dans une situation idéalisée serait de 0.5. Par ailleurs, on suppose qu’un observateur de cette expérience peut vérifier si les deux pièces tombent du même côté; cet observateur n’est cependant pas entièrement fiable, et se trompe donc avec une certaine probabilité. On précise que les deux pièces se comportent de manière indépendante.

Connaissant les caractéristiques des pièces (leurs probabilités p_1 et p_2 de tomber sur ‘face’) et le taux d’erreur de l’observateur, disons α, on demande de définir une règle qui en fonction de l’information fournie par l’observateur détermine la configuration la plus probable des pièces. Nous utiliserons différentes versions de ce problème dans la suite pour illustrer le raisonnement des effets aux causes qui caractérise les problèmes de diagnostic.

1.3.2.4 Autres exemples d’applications pratiques

Le diagnostic médical combine le raisonnement à partir de symptômes (manifestation de douleurs, résultats d’analyses sanguines, radiographies etc.) et de facteurs qui prédisposent (fumer, alimentation, mode de vie, histoire familiale, etc.) vis-à-vis de certaines pathologies. En général, la recherche de “bugs” (failles de conception) dans un raisonnement, dans un programme informatique, dans une théorie, conduisent à des problèmes d’inférence de type “diagnostic”.

La statistique inférentielle a pour objet de déterminer la loi de probabilité qui explique les observations, par exemple la loi qui gouverne la durée de vie de composants électroniques issus d’une chaine de fabrication, les observations étant la durée de vie d’un échantillon de composants prélevés pour le contrôle de qualité. Ce problème

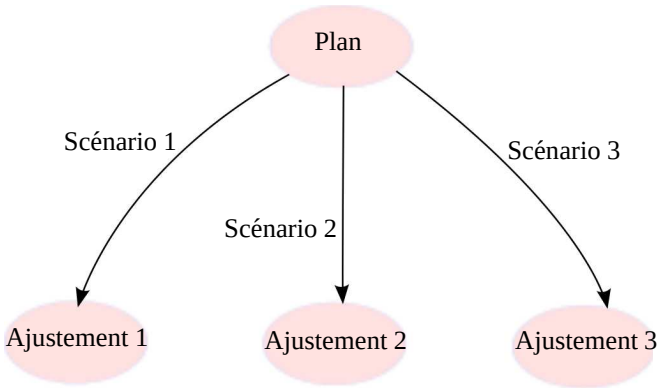


Figure 1.5: Arbre de décision pour la gestion de la production.

est un problème de “diagnostic” ayant pour objet de déterminer les paramètres de la population à partir des observations recueillies sur certains éléments de celle-ci.

1.3.3 Problème de décision séquentielle

Il s’agit d’un problème de raisonnement où on cherche à établir une stratégie permettant d’adapter son comportement à des informations qui seront collectées ultérieurement. En quelque sorte, il s’agit de se préparer à prendre des décisions futures et à mettre cette préparation à profit pour prendre une décision maintenant.

1.3.3.1 Exemple : gestion d’un parc de production

Un producteur d’électricité doit décider lesquelles de ses centrales doivent fonctionner le lendemain. Son objectif est de minimiser le coût de production, sachant qu’il s’est engagé à répondre à la demande de ses clients. Les clients ne sont pas parfaitement prévisibles, et il peut y avoir des pannes de fonctionnement de certaines centrales au cours de la journée du lendemain. Chaque centrale possède une certaine souplesse, qui lui permet d’ajuster rapidement son niveau de production, pour autant qu’elle soit en marche. Ce problème est fort différent du problème de décision concernant la construction d’une ligne électrique que nous avons présenté plus haut.

Ici, le problème du producteur est de choisir un sous-ensemble de centrales qui seront en fonctionnement le lendemain, et de faire en sorte qu’elles pourront ajuster leur niveau de production le lendemain à la demande d’électricité qui aura lieu. La décision en ce qui concerne les centrales à démarrer doit être prise la veille sur base des prévisions, alors que le niveau de production de chacune pourra être ajusté le lendemain en fonction de la demande réelle d’électricité de la part des consommateurs. Ensemble, ces deux décisions doivent conduire à maximiser le profit du producteur.

1.3.3.2 Modélisation

On peut représenter la situation par un arbre de décision tel que celui illustré à la Figure 1.5. Dans cet arbre, la racine correspond (indiquée en haut sur la figure) à la décision à prendre maintenant (ici, choisir le plan de démarrage des centrales pour le lendemain). Les branches (flèches) partant de la racine correspondent aux scénarios envisagés pour le lendemain (pannes ou bien variations de la demande par rapport aux prévisions). Ces branches conduisent à des noeuds successeurs de la racine (feuilles de l’arbre, représentées dans la partie inférieure de la figure) auxquels sont attachées des décisions de recours, c’est-à-dire des ajustements du plan de production qui permettraient de réagir de façon optimale à ces scénarios, compte tenu de la décision prise maintenant (ici, il s’agirait de modifier le niveau de production de chaque centrale démarrée, en fonction de la demande observée et/ou des centrales tombées en panne).

Nous illustrerons la notion d’arbre de décision et d’arbre de scénarios dans les chapitres suivants.

1.3.3.3 Abstraction sous forme de problème jouet

Dans le problème du “Monty Hall” un joueur participe à un jeu télévisuel. Dans ce jeu, le joueur doit choisir une porte parmi trois, sachant que derrière une des portes se trouve une voiture de grand luxe, et derrière les deux autres une portion de frites de mauvaise qualité (ou bien une chèvre rachitique, selon la version de l'histoire). Le joueur remportera le lot caché par la porte qu'il aura choisie.

Le jeu se passe en trois étapes, à savoir (i) le joueur désigne une porte; (ii) l'animateur choisit de révéler ce qui se trouve derrière une des deux portes non choisies par le joueur; (iii) le joueur peut remettre en question son choix initial en choisissant la porte non ouverte par l'animateur. Pour que les deux étapes du jeu soient intéressantes, on impose à l'animateur de ne pas ouvrir la porte derrière laquelle se trouve la voiture.

On demande de déterminer la stratégie optimale de décision en deux temps pour le joueur (c'est-à-dire qui maximise ses chances de remporter le lot intéressant); celle-ci doit combiner de façon optimale (i) le choix de la première porte à désigner, et (ii) une règle à utiliser pour choisir la seconde porte en fonction de la porte que l'animateur aura décidé d'ouvrir.

Ce problème sera traité en détail à la fin du chapitre suivant.

1.3.3.4 Autres exemples d'applications pratiques

En présence d'incertitudes, on souhaite en général retarder la prise de décisions coûteuses en se disant que l'information disponible plus tard sera meilleure et permettra de mieux justifier la décision à prendre; en même temps, le fait de retarder la prise de décision réduit souvent les marges de manoeuvre car certaines opportunités ne sont disponibles que pendant une période courte. De cela résulte un compromis qui s'articule de la manière suivante: quelle décision prendre maintenant, sachant que je peux ultérieurement réviser mes décisions en fonction des informations que j'aurai à ce moment à ma disposition, mais que ma capacité de révision sera soit facilitée soit empêchée par la décision que je m'apprête à prendre maintenant. De façon équivalente, le décideur se trouve en face d'une série de choix entre lesquels il est obligé d'arbitrer à un moment donné, sachant qu'il aura des opportunités de réagir ultérieurement en faisant d'autres choix à des instants postérieurs, la qualité et la disponibilité de ces choix étant potentiellement influencée par la décision qu'il doit prendre maintenant.

Les méthodes permettant d'accorder les décisions actuelles et les stratégies de révision relèvent de la programmation stochastique dynamique. Elles s'appliquent à tous les problèmes de décision séquentielle sous incertitude. Les exemples pour les ingénieurs concernent la conception de systèmes techniques dont les stratégies d'exploitation sont flexibles, et aussi la gestion journalière de tels systèmes.

Pratiquement, une des questions qui tombe dans ce créneau concerne la gestion de la maintenance d'une flotte de véhicules, sachant que la flotte doit être exploitée à tout moment pour maximiser les revenus et que la maintenance vise à limiter les coûts d'exploitation et à maximiser la disponibilité future des véhicules. A un moment particulier, il s'agira de décider si un véhicule particulier devrait être immobilisé pour maintenance, sachant que cela conduira à un manque à gagner à court terme, mais est susceptible d'augmenter l'espérance des bénéfices à plus long terme. Ce problème de gestion de la maintenance est évidemment générique.

De façon analogue, la question de la planification des investissements dans les outils de production se traduit par un problème de décision séquentielle en environnement incertain. Il s'agit ici de décider de la construction de nouvelles usines, à un moment donné, sachant que d'autres peuvent être construites plus tard ou que d'autres encore peuvent être revendues ou bien démantelées. De la même façon, pour une entreprise il est important de disposer d'une méthodologie lui permettant de décider rationnellement si elle doit investir en recherche sur les technologies qui pourraient lui être utiles dans le futur, et si oui de choisir au mieux les voies qu'elle décidera d'explorer. L'issue d'une recherche est souvent incertain mais il peut donner un avantage concurrentiel déterminant pour le développement d'une entreprise.

L'ensemble de ces problèmes relève de gestion de la production et de la stratégie d'entreprise, et il est donc fort important que les étudiants ingénieurs soient confrontés lors de leurs études aux fondements des méthodes permettant d'aborder ces problèmes.

1.4 ORGANISATION DU COURS

Le cours d'Eléments de Probabilités est composé de trois types d'activités :

- Le cours oral ex cathedra, qui aura pour objet de présenter les notions faisant l'objet de ces notes.
- Les séances d'exercices dirigés, qui ont pour objet d'approfondir certaines techniques et de les appliquer à divers problèmes concrets.
- Les travaux pratiques sur ordinateur, qui ont pour objet de former les étudiants en leur permettant d'apprendre à étudier certaines notions par le biais de simulations informatiques.

Les trois composantes sont nécessaires pour assimiler les notions et être en mesure de les mettre en oeuvre en pratique. L'évaluation portera donc sur ces trois parties:

- Examen écrit portant sur la théorie et les exercices.
- Evaluation des travaux pratiques sur ordinateur via la correction de rapports rédigés par les étudiants.

Le présent document constitue les notes du cours ex cathedra. La partie principale (désignée par le terme de syllabus) est complétée par quelques annexes qui constituent des rappels et des compléments d'information que nous mettons à la disposition des étudiants, mais qui ne feront pas l'objet de séances de répétition ni de travaux pratiques, ni d'une évaluation des connaissances.

Notes

1. En logique, la *déduction* est un processus de raisonnement qui part d'axiomes (ou plus généralement de modèles mathématiques) pour déduire des propriétés particulières intéressantes qui doivent être satisfaites dans toute situation qui respecte les axiomes.
2. Etymologiquement, l'adjectif *stochastique* fait référence à l'art de faire des conjectures ciblées relatives à un problème lorsque les informations disponibles ne permettent pas de donner une réponse précise et garantie comme correcte; le terme est souvent utilisé à la place des termes *aléatoire*, *incertain* ou *probabiliste*.
3. L'*induction* est un processus de raisonnement qui exploite des observations pour en inférer des lois générales qui sont des hypothèses explicatives de ce qui est observé; c'est le processus de base utilisé pour la construction des théories dans le domaine des sciences naturelles.
4. Nous utilisons ici ces termes pour désigner les composants physiques et non les abstractions correspondantes de la théorie des circuits.
5. Notons que dans la représentation probabiliste de la figure 1.2 nous avons fait implicitement l'hypothèse que les entrées observables et non-observables étaient indépendantes.

2 LE MODÈLE PROBABILISTE

"La théorie des probabilités n'est rien d'autre que le bon sens réduit sous forme de calcul."
- Pierre Simon Laplace, 1749 - 1827

Dans ce chapitre nous discutons la notion d'expérience aléatoire et nous présentons les bases mathématiques du calcul de probabilités en analysant les propriétés fondamentales d'une mesure de probabilité. Nous introduisons ensuite la notion de probabilité conditionnelle et étudions ses principales propriétés, fondamentales pour le raisonnement à partir d'un modèle probabiliste. Nous discutons aussi les différentes interprétations de la notion de probabilité.

2.1 NOTION DE PROBABILITE - INTERPRETATIONS

Dans cette section nous allons introduire, tout d'abord intuitivement puis plus formellement, la notion de probabilité. Ensuite nous discuterons très brièvement de ses différentes interprétations logiques et physiques.

2.1.1 Intuitivement

Comme mentionné au chapitre 1, le calcul des probabilités est un outil mathématique qui permet de représenter et de manipuler des situations/expériences dont l'issue est aléatoire et/ou au sujet desquelles on dispose de connaissances incomplètes/incertaines.

Une connaissance (c'est-à-dire une affirmation logique) est dite *incertaine* dans un contexte donné, si dans ce contexte il est impossible aussi bien de réfuter sa véracité que de la prouver. La notion de probabilité permet d'ordonner de telles connaissances par ordre de *plausibilité* croissante, et de remettre à jour cet ordonnancement lorsque de nouvelles informations deviennent disponibles.

2.1.1.1 Notion d'expérience aléatoire

Une *expérience* est qualifiée d'aléatoire si on ne peut pas prévoir par avance son résultat, et donc si, répétée dans des conditions apparemment identiques, elle pourrait donner lieu à des *résultats* différents. Le calcul des probabilités permet de modéliser et de simuler de telles expériences.

Pour étudier une telle expérience on s'intéresse tout d'abord à l'**univers** de tous les **résultats** (ou objets) possibles : on note usuellement Ω cet ensemble fondamental, et ω un élément particulier de Ω, c'est-à-dire un résultat particulier parmi ceux possibles.

Exemples illustratifs d'expériences aléatoires

1. Lancers de pièce. Nous considérerons dans la suite comme exemple simple de problème, mais néanmoins très riche, le lancer simple ou multiple d'une pièce. Si nous désignons par P le résultat d'un lancer de pièce correspondant à "pile" et par F celui correspondant à "face", nous pouvons considérer les expériences suivantes:

- **1.1 Lancer simple :** on lance une fois la pièce et on observe le résultat; on a
$$\Omega = \{P, F\}.$$
- **1.2 Triple lancer :** on lance trois fois de suite la pièce et on observe la suite des trois résultats; on a
$$\Omega = \{PPP, PPF, PFP, PFF, FPP, FPF, FFP, FFF\}.$$
- **1.3 Lancer jusqu'au double pile :** On lance la pièce autant de fois qu'il faut pour observer l'occurrence successive de deux fois "pile" puis observe la suite de résultats; on a
$$\Omega = \{PP, FPP, PFPP, FFPP, \ldots\},$$
qui est un ensemble infini mais dénombrable. Notons qu'il faut s'assurer que l'expérience est bien réalisable, en d'autres termes qu'on observera à coup sûr après un nombre fini de lancers l'occurrence d'un double "pile". En d'autres mots, il faut s'assurer que l'ensemble (en fait non dénombrable) des suites de longueur infinie dans lesquelles on n'observe pas le double "pile" est de probabilité égale à 0.

2. Roue de la fortune. On fait tourner une roue munie d'un repère, et lorsque la roue s'arrête on observe l'angle formé par le repère et le point situé à midi (en radians, et dans le sens des aiguilles d'une montre). Dans cet exemple $\Omega = [0, 2\pi[$, un ensemble non dénombrable. Un expérience analogue consiste à observer à un moment précis la position de la trotteuse des secondes d'une montre : si on mesure la position en secondes par rapport à "midi", l'ensemble $\Omega = [0, 60[$, c'est-à-dire un intervalle borné de la droite réelle $\mathbb{R}$.

3. Diagnostic médical. Par exemple, si on s'intéresse au diagnostic médical, l'expérimentateur pourrait être un médecin particulier, et l'expérience aléatoire pourrait concerner le premier diagnostic de l'année (disons, le 2 janvier au matin, en l'an 2020). Nous pourrions alors définir l'ensemble Ω pour ce problème comme l'ensemble de tous les patients que ce médecin est susceptible de diagnostiquer le 2 janvier au matin, en l'an 2020 (le médecin ne peut évidemment pas prévoir quel sera le patient particulier qui va se présenter devant lui).

4. Trafic internet. Un autre exemple intéressant concerne les réseaux informatiques. Plaçons nous en un noeud particulier du réseau Internet, et observons les messages par courrier électronique qui y transitent pendant une journée donnée. Un résultat particulier est alors la suite particulière de messages qui ont transité pendant la période d'observation. Avant d'avoir effectué l'expérience on ne peut évidemment pas prévoir quelle suite sera observée, et l'ensemble Ω est alors l'ensemble de toutes les suites de messages possibles pouvant transiter sur une journée, un ensemble certes très compliqué à caractériser mais néanmoins de taille finie en pratique.

Commentaires

Il faut noter que pour construire un modèle probabiliste, l'univers Ω est défini en fonction de l'objectif particulier poursuivi.

Ainsi, dans le troisième exemple ci-dessus on aurait pu définir Ω comme étant l'ensemble des maladies diagnostiquées par le médecin, ou encore comme étant l'ensemble des médicaments prescrits par celui-ci.

Dans le quatrième exemple, on aurait pu s'intéresser uniquement à l'expéditeur des messages et définir le résultat de l'expérience comme étant l'ensemble des adresses email des expéditeurs ayant envoyé au moins un message email pendant la journée : l'univers Ω serait alors l'ensemble de tous les sous-ensembles d'expéditeurs possibles de messages électroniques susceptibles de transiter par le noeud.

Le choix de Ω pour l'étude d'une expérience aléatoire est la première étape de modélisation probabiliste. En pratique ce choix doit souvent être progressivement raffiné, en fonction de l'avancement d'un projet.

2.1.1.2 Notion d'événement

Dans la terminologie du calcul des probabilités, un **événement** désigne une assertion logique vérifiable relative au résultat d'une expérience, et qui définit un **sous-ensemble** de Ω.

A tout événement on peut donc faire correspondre un sous-ensemble de Ω. En particulier, à tout $\omega \in \Omega$ on peut associer un **événement élémentaire** correspondant au singleton $\{\omega\}$.

Exemples d'événements

- Pour l'exemple 1.2 ci-dessus, l'assertion logique **"Pile" n'est observé au plus qu'une seule fois** définit un événement, correspondant au sous-ensemble $\{PFF, FPF, FFP, FFF\}$ de Ω.
- Pour l'exemple 3, l'assertion logique **Le premier patient qui se présentera le 2 janvier 2020 au matin est un étudiant qui devrait passer un examen le 3 janvier** définit un événement. Cette assertion logique est soit vraie soit fausse, et définit en fait un sous-ensemble de la "clientèle"; nous allons noter cet ensemble $A \subset \Omega$.

  Un autre événement pour cet exemple correspondrait à l'assertion logique **le premier patient ... a trop festoyé le jour du réveillon et souhaite un certificat médical**, à laquelle on peut également associer un sous-ensemble $B \subset \Omega$, a priori différent de A.

2.1.1.3 Notion de probabilité d'un événement

La notion de probabilité, qui sera formalisée ci-dessous, est une mesure de l'importance des événements : elle associe à un événement un nombre positif (entre 0 et 1) qui représente le degré de certitude qu'on peut associer *a priori* à la réalisation de celui-ci. Il traduit donc l'état de connaissance dans lequel on se trouve *avant* de réaliser une expérience.

Exemples de probabilités d'événements

- Pour l'expérience du triple lancer de pièce, décrite ci-dessus, on peut se demander quelle est la probabilité **d'observer au moins deux fois "pile"**; cet événement est la négation de l'événement **"Pile" n'est observé au plus qu'une seule fois** décrit plus haut et correspond donc au sous-ensemble de réalisations

  $$\Omega \setminus \{PFF, FPF, FFP, FFF\} = \{PPP, PPF, PFP, FPP\}.$$

  Si on pense que toutes les issues de l'expérience ont la même chance de se réaliser, on attribuerait à cet événement une probabilité $P(\{PPP, PPF, PFP, FPP\}) = 0.5$.

  De même, on attribuerait aussi une probabilité $P = 0.5$ à l'événement $\{PFF, FPF, FFP, FFF\}$, et une probabilité $P = 1$ à l'événement Ω correspondant à la réalisation de l'un ou l'autre de ces deux événements. On est en effet *certain* que l'un ou l'autre des deux événements doit se réaliser.

  On attribuerait aussi une probabilité $P = 0$ à l'événement décrit par **"pile" n'est observé qu'une seule fois et "pile" est observé au moins deux fois** puisqu'aucun résultat de l'expérience ne peut vérifier cette condition.
- Pour l'exemple de la roue de la fortune, on pourrait considérer que toutes les positions finales sont a priori possibles et qu'aucune n'a plus de chances de se produire que les autres. Dans ce cas, on attribuerait à un événement du type $\omega \in [\alpha, \beta]$ (voulant dire que la roue s'arrête entre la position α et la position β) une probabilité de $P(\omega \in [\alpha, \beta]) = (\beta - \alpha)/2\pi$.

  On retrouve alors $P(\Omega) = 1$.

  On aura aussi en général que $P([\omega \in A] \wedge [\omega \in B]) = P([\omega \in A \cap B])$; si A et B sont disjoints, cette probabilité vaudra naturellement 0.

  De même, on aura que $P([\omega \in A] \vee [\omega \in B]) = P([\omega \in A \cup B])$; si A et B sont disjoints, cette probabilité vaudra $P([\omega \in A]) + P([\omega \in B])$, et si leur union vaut $[0, 2\pi[$ elle vaudra 1.
- Pour les autres exemples d'expériences décrites ci-dessus, il est par contre plus difficile de donner une valeur fondée aux probabilités des événements que nous avons illustrés. Les difficultés pour le faire sont de deux natures différentes.

Par exemple pour le “lancer jusqu’au double pile” on doit s’assurer que les probabilités définies pour tous les événements sont cohérentes, et ne conduisent pas lorsqu’on combine ces valeurs à des nombres tombant en dehors de l’intervalle $[0, 1]$ ou bien à des valeurs différentes selon la manière dont on décrit un événement complexe en fonction d’autres événements. Pour les problèmes de diagnostic médical et de modélisation du trafic internet, vient s’ajouter une autre difficulté qui est de modéliser les probabilités de façon à ce qu’elles représentent correctement la réalité.

2.1.1.4 Evénements impossibles et événements certains

Notons que si l’univers Ω comprend un nombre fini d’éléments, les événements (qui définissent des sous-ensembles de Ω) sont forcément aussi des ensembles finis. Dans ce cas, un événement auquel on associe une probabilité égale à 1 est un événement dit certain (on est a priori certain qu’il se réalisera); symétriquement, un événement auquel on associe une probabilité nulle est un événement impossible : on est a priori certain qu’il ne se réalisera pas. Ces deux cas extrêmes sont les limites où le raisonnement probabiliste rejoint la logique classique : la partie intéressante concerne cependant tous les événements auxquels on associe des probabilités intermédiaires. La mesure de probabilité permet de trier l’ensemble de ces événements par ordre croissant de leur probabilité a priori et de mettre à jour cet ordonnancement en fonction des informations disponibles. Le calcul des probabilités permet de s’assurer que les raisonnements effectués à l’aide de ces nombres restent cohérents, et de mettre à jour de façon cohérente ces probabilités en fonction des informations disponibles.

2.1.1.5 Remarque sur la notion d’expérience reproductible

Certaines des expériences que nous avons illustrées ci-dessus ne peuvent pas en principe être répétées du tout : le 2 janvier 2020 au matin il n’y aura qu’un seul patient qui sera le premier. De même, au cours d’une journée donnée un seul ensemble de messages transitera en un noeud donné d’Internet.

Il est cependant souvent possible de supposer que les propriétés de certaines expériences ne changent pas au cours du temps : on peut alors envisager de répéter (éventuellement indéfiniment) cette expérience dans des conditions qui ne changent pas au fil du temps. Par exemple, on peut supposer qu’une pièce ne s’use pas au cours du temps et que le résultat d’une expérience de lancer de pièce ne dépend pas du temps, ou du nombre de fois qu’elle a déjà été lancée. Similairement, lorsqu’on répète un certain nombre de fois une opération de mesure, on peut supposer que l’ensemble des résultats possibles ne change pas d’une fois à la suivante et que les probabilités des événements restent constantes.

Il est clair, néanmoins, que ce type de situation est une abstraction qui ne se réalise jamais parfaitement en pratique : il n’est pas possible d’observer un système physique sans le perturber. Néanmoins, cette abstraction est souvent vérifiée approximativement et est à la base d’une grande partie des statistiques. Nous allons pour le moment au moins admettre qu’une expérience peut être répétée. Nous discuterons à la section 2.1.3 plus finement pourquoi il n’en est pas toujours ainsi, et pourquoi il est néanmoins intéressant de se servir du calcul des probabilités lorsque ce n’est pas le cas.

2.1.2 Formellement

Dans cette section nous présentons la définition axiomatique du calcul de probabilités. Cette axiomatisation, qui a mis de nombreux siècles à se cristalliser, est garante de la cohérence logique de la théorie des probabilités. Il est par conséquent capital de bien l’assimiler.

Pour un ensemble universel Ω, la démarche consiste à définir d’abord une structure de σ-algèbre qui caractérise les événements dont on souhaite définir la probabilité, puis à définir une mesure de probabilité associant à chaque événement un nombre réel entre 0 et 1.

La notion de σ-algèbre assure qu’à partir de plusieurs événements on peut en construire d’autres par les opérations logiques de négation, conjonction (le “et” logique) et disjonction (le “ou” logique). Dans la suite nous utiliserons les notations ensemblistes (complément, intersection et union) pour spécifier ces propriétés.

La notion de mesure de probabilité est essentiellement définie de façon à ce que la probabilité associée à la disjonction d’événements qui ne peuvent pas se réaliser simultanément soit la somme des probabilités de ces événements.

2.1.2.1 Notion de σ-Algèbre d'événements définis sur un ensemble universel Ω

Notations. Dans la suite nous utiliserons

- des lettres minuscules grecques ($\alpha, \beta, \ldots$) pour désigner les éléments de Ω
- des lettres majuscules latines (p.ex. $A, B, \ldots$) pour désigner des sous-ensembles de Ω.

Par ailleurs, nous désignons par

- 2^Ω l'ensemble de tous les sous-ensembles de Ω
- des lettres rondes (p.ex. $\mathcal{A}, \mathcal{B}, \ldots$) pour désigner des sous-ensembles de 2^Ω, c'est-à-dire des ensembles de sous-ensembles de Ω.

Nous utiliserons également les notations

- $f(\cdot), g(\cdot), \ldots$ pour désigner une fonction définie sur (une partie de) Ω,
- $F(\cdot), G(\cdot), \ldots$ pour désigner des fonctions définies sur (une partie de) 2^Ω.

Enfin, nous désignerons par $\neg$[proposition logique] la négation d'une proposition logique s'appliquant à des réalisations et par A^c le complémentaire relatif à Ω d'un sous-ensemble A de Ω, c'est-à-dire que $A^c = \Omega \setminus A$.

2.1.2.2 Définition de la notion de σ-algèbre

Nous définissons la notion de σ-algèbre comme suit :

Définition de la notion de σ-algèbre sur Ω.

Une σ-algèbre $\mathcal{E}_\Omega$ d'événements, ou *tribu* [(1)], définie sur un univers Ω est une partie de 2^Ω (i.e. un ensemble de sous-ensembles de Ω) qui vérifie les propriétés suivantes :

T1. $\Omega \in \mathcal{E}_\Omega$;

T2. $A \in \mathcal{E}_\Omega \Rightarrow A^c \in \mathcal{E}_\Omega$;

T3. $\forall A_1, A_2, \ldots \in \mathcal{E}_\Omega$ (en nombre fini ou dénombrable [(2)]) : $\bigcup_i A_i \in \mathcal{E}_\Omega$.

Dans la suite, s'il n'y a pas de risque de confusion, nous utilisons $\mathcal{E}$ à la place de $\mathcal{E}_\Omega$ pour alléger les notations.

Les éléments de $\mathcal{E}$ sont désignés par le terme d'événements. Il s'agit de parties de Ω auxquelles nous conférons un statut particulier, à savoir qu'on peut parler de leur probabilité, comme nous le verrons ci-dessous. Les propriétés qui définissent la notion de σ-algèbre assurent que si nous pouvons parler de la probabilité de certains sous-ensembles de Ω caractérisés par des affirmations logiques, nous pouvons aussi parler de la probabilité de tout sous-ensemble décrit par une phrase logique combinant ces affirmations.

Remarques.

1. Les deux premières propriétés ci-dessus impliquent que l'ensemble vide (désigné par $\emptyset$) fait nécessairement partie de toute σ-algèbre d'événements.
2. La seconde et la troisième propriété impliquent également que $\bigcap_i A_i \in \mathcal{E}$.
3. $\{\emptyset, \Omega\}$ est une σ-algèbre d'événements : c'est la plus petite de toutes.
4. 2^Ω est une σ-algèbre d'événements : c'est la plus grande de toutes.
5. Si Ω est un ensemble fini, alors $\mathcal{E}$ l'est également.
6. Par contre, si Ω est infini (dénombrable ou non), $\mathcal{E}$ peut être non-dénombrable, dénombrable, et même finie.
7. Deux événements A et B sont dits *incompatibles* si $A \cap B = \emptyset$.

Exemples.

- Dans le cas du simple lancer de pièce, on peut définir $\mathcal{E} = \{\emptyset, \Omega, \{P\}, \{F\}\}$, c'est-à-dire l'ensemble de tous les sous-ensembles de $\Omega = \{P, F\}$.
- Dans le cas du triple lancer de pièce, on peut définir $\mathcal{E}$ comme étant l'ensemble de tous les sous-ensembles de $\Omega = \{PPP, PPF, PFP, PFF, FPP, FPF, FFP, FFF\}$. Dans ce cas, $\mathcal{E}$ comportera $2^8 = 256$ éléments.

  Cependant, on pourrait aussi choisir une σ-algèbre moins fine, par exemple

  $$\mathcal{E} = \{\emptyset, \Omega, \{PPP, PPF, PFP, PFF\}, \{FPP, FPF, FFP, FFF\}\},$$

  qui ne permettrait en fait de parler que de l'issue du premier lancer de pièce.

Nous verrons dans la suite que le calcul des probabilités nécessite de pouvoir considérer des σ-algèbres différentes pour un même problème. Par ailleurs, nous verrons aussi qu'en toute généralité, il est nécessaire d'imposer des conditions supplémentaires à la notion de σ-algèbre afin d'assurer la cohérence d'un modèle probabiliste.

Système complet d'événements

Définition de la notion de système complet d'événements.

$A_1, \ldots, A_n \in \mathcal{E}$ forment un *système complet d'événements* (on dit qu'ils forment une partition de Ω) si

- $\forall i \neq j \ : A_i \cap A_j = \emptyset$ (ils sont incompatibles deux à deux)
- et si $\bigcup_{i=1}^n A_i = \Omega$ (ils couvrent Ω).

NB: On supposera la plupart du temps que tous les A_i sont non-vides.

Exemple. Pour le triple lancer de pièce, on peut par exemple définir le système complet d'événements :

$$\begin{aligned} A_1 &= \{PPP, PPF, PFP, PFF\}, \\ A_2 &= \{FPP, FPF, FFP, FFF\}, \\ A_3 &= \emptyset. \end{aligned}$$

Remarque. Certains auteurs définissent une système complet d'événements de façon légèrement différente de celle que nous avons adpotée ci-dessus. Au lieu d'exiger que $\bigcup_{i=1}^n B_i = \Omega$ ils imposent à la place la condition plus faible que $P(\bigcup_{i=1}^n B_i) = 1$. Ce type de système, obéit également au théorème des probabilités totales. D'ailleurs, on peut évidemment compléter un tel système avec $B_{n+1} = (\bigcup_{i=1}^n B_i)^c$, avec $P(B_{n+1}) = 0$.

2.1.2.3 Notion de mesure de probabilité

Le statut particulier des événements est qu'il est possible de leur attribuer une probabilité, c'est-à-dire un nombre positif compris entre 0 et 1, qui doit répondre aux axiomes suivants.

Axiomes de Kolmogorov.

On appelle *mesure* (ou loi) de probabilité sur $(\Omega, \mathcal{E})$ une fonction $P_\Omega(\cdot)$ définie sur $\mathcal{E}$ telle que :

K1. $P_\Omega(A) \in [0, 1], \forall A \in \mathcal{E}$;

K2. $P_\Omega(\Omega) = 1$;

K3. $\forall A_1, A_2, \ldots \in \mathcal{E}$ (en nombre fini ou dénombrable) et incompatibles deux-à-deux on a : $P_\Omega(\bigcup_i A_i) = \sum_i P_\Omega(A_i)$ *(propriété essentielle de σ-additivité).*

Dans la suite, s'il n'y a pas de risque de confusion, nous utilisons P à la place de P_Ω pour alléger les notations.

Discussion. On voit que l'utilisation du calcul des probabilités passe par trois étapes successives de modélisation : définition de l'univers Ω, choix d'une σ-algèbre d'événements $\mathcal{E}$, et enfin quantification par le choix de la mesure de probabilité P. Les propriétés de base qui sont requises pour que l'ensemble soit cohérent sont le fait que $\mathcal{E}$ soit effectivement une σ-algèbre et que P satisfasse les axiomes de Kolmogorov. Un triplet $(\Omega, \mathcal{E}, P)$ qui vérifie ces conditions est appelé un *espace de probabilité*.

On constate également que le calcul des probabilités est compatible avec la logique classique (en tout cas, si Ω est fini). Il suffit de considérer le cas particulier où $P(\cdot)$ est définie sur $\{0, 1\}$, et associer à la valeur 1 la valeur de vérité "vrai" et à 0 la valeur "faux".

2.1.2.4 Propriétés remarquables

Des axiomes de Kolmogorov on peut déduire immédiatement les propriétés suivantes (qu'on démontrera à titre d'exercice, en se restreignant au cas où Ω est fini).

1. $P(\emptyset) = 0$.
2. $P(A^c) = 1 - P(A)$.
3. $A \subset B \Rightarrow P(A) \leq P(B)$.
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
5. $P(\bigcup_i A_i) \leq \sum_i P(A_i)$.
6. On a $A_i \downarrow \emptyset \Rightarrow \lim_{i\to\infty} P(A_i) = 0$
   (Nous écrivons $A_i \downarrow A$, pour désigner une suite d'ensembles $\{A_i\}_{i\in\mathbb{N}}$, telle que $A_{i+1} \subset A_i$ et $\bigcap_{i\in\mathbb{N}} A_i = A$.)

On a également :

1. $P(A) = 1 \Rightarrow P(A \cup B) = 1, \forall B \in \mathcal{E}$.
2. $P(A) = 1 \Rightarrow P(A \cap B) = P(B), \forall B \in \mathcal{E}$.
3. $P(A) = 0 \Rightarrow P(A \cap B) = 0, \forall B \in \mathcal{E}$.
4. $P(A) = 0 \Rightarrow P(A \cup B) = P(B), \forall B \in \mathcal{E}$.
5. et

Formule de Poincaré

Il s'agit de la généralisation de la formule $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ au cas de l'union d'un nombre fini quelconque d'événements. On a

$$\begin{aligned} P(\bigcup_{i=1}^{n} A_i) \quad = \quad & \sum_{\{i_1 : 1 \leq i_1 \leq n\}} P(A_{i_1}) \\ & - \sum_{\{(i_1,i_2) : 1 \leq i_1 < i_2 \leq n\}} P(A_{i_1} \cap A_{i_2}) \\ & + \sum_{\{(i_1,i_2,i_3) : 1 \leq i_1 < i_2 < i_3 \leq n\}} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) \\ & - \ldots \\ & + (-1)^{n-2} \sum_{\{(i_1,\ldots,i_{n-1}) : 1 \leq i_1 < \cdots < i_{n-1} \leq n\}} P(A_{i_1} \cap \cdots \cap A_{i_{n-1}}) \\ & + (-1)^{n-1} P(\bigcap_{i=1}^{n} A_i). \end{aligned}$$

La formule de Poincaré se démontre par récurrence.

2.1.2.5 Théorème des probabilités totales (version 1)

Le théorème des probabilités totales se formule comme suit.

Théorème des probabilités totales (version 1)

Soit $B_1, \ldots, B_n$ un système complet d'événements, alors $\forall A \in \mathcal{E} \ : P(A) = \sum_{i=1}^{n} P(A \cap B_i)$.

Remarque. Le théorème des probabilités totales est une conséquence directe du troisième axiome de Kolmogorov (**K3**). Il reste valable pour un nombre infini dénombrable d'ensembles B_i et peut également être étendu au cas où les B_i ne sont plus en nombre dénombrable.

2.1.3 • Différentes interprétations de la notion de probabilité

Il faut faire la distinction entre la formulation mathématique d'une théorie et l'utilisation que nous en faisons pour étudier des problèmes du monde réel qui nous entoure, c'est-à-dire son interprétation. Ceci est particulièrement vrai pour une théorie telle que le calcul des probabilités qui vise entre autres à modéliser une certaine forme du raisonnement humain, et qui s'adresse à des problèmes où l'incertitude joue un rôle fondamental, c'est-à-dire des problèmes où il pourrait être difficile de valider la théorie. En particulier, la théorie ne nous aide pas lorsqu'il s'agit de définir en pratique la loi de probabilité à associer aux événements choisis.

En réalité, depuis son origine, le calcul des probabilités a donné lieu à des débats intenses entre scientifiques, logiciens, physiciens, philosophes, en ce qui concerne la ou les interprétations physiques à donner à la notion même de probabilité. Ces débats sont encore d'actualité, et le resteront certainement encore longtemps; c'est la raison pour laquelle nous voulons mettre en évidence ici les différents points de vues qui s'opposent dans ce débat d'idées.

2.1.3.1 Le point de vue objectiviste

La vision classique. La vision classique est héritée des jeux de hasard. Dans cette vision Ω est fini, et on considère alors comme σ-algèbre d'événements l'ensemble 2^{Ω}, fini lui aussi.

La démarche adoptée pour alors définir la mesure de probabilité, consiste à attribuer des probabilités aux événements élémentaires $P(\{\omega\}); \forall w \in \Omega$; les probabilités des autres événements s'en déduisent par application des axiomes de Kolmogorov. En particulier, on aura en vertu du 3ème axiome de Kolmogorov que $P(\Omega) = \sum_{\omega \in \Omega} P(\{\omega\})$ ce qui en vertu du 2ème axiome de Kolmogorov impose évidemment que $\sum_{\omega \in \Omega} P(\{\omega\}) = 1$.

Sous cette contrainte, la vision classique impose des arguments de symétrie, en considérant que tous les événements élémentaires sont équiprobables. Par conséquent, $P(\{\omega\}) = |\Omega|^{-1}$ (où $|\Omega|$ désigne la taille finie de l'univers). C'est cette démarche qui conduit à associer aux 6 faces d'un dé "parfait", une probabilité de $\frac{1}{6}$.

La principale faiblesse de cette approche est qu'elle repose sur un postulat de symétrie (idéal, et irréalisable en pratique) et ne permet donc pas la remise en question des probabilités en fonction d'informations supplémentaires (obtenues par exemple en effectuant des expériences de lancer de dé). Une autre faiblesse est que cette approche ne s'étend pas au cas où Ω est infini (dénombrable ou non-dénombrable; voir à ce sujet la discussion au §2.1.4).

La vision fréquentiste. Elle repose sur la loi des grands nombres (voir fin du chapitre 3) et sur une autre idéalisation, à savoir la notion d'expérience indéfiniment reproductible dans les mêmes conditions. La loi des grands nombres assure en effet que dans une telle expérience la fréquence relative observée d'un événement converge vers la probabilité de celui-ci. La vision fréquentiste définit alors la probabilité d'un événement comme la limite de la proportion de cas favorables (la réalisation est compatible avec l'événement) au nombre total d'essais, quand on répète indéfiniment l'expérience aléatoire.

Notons que cette vision est aussi appelée la vision "orthodoxe" : tout comme dans la vision classique, la notion de probabilité est supposée définie de façon unique (c'est-à-dire indépendamment de l'observateur), et tous les observateurs doivent se soumettre à une expérience (théorique) similaire pour en déterminer la valeur.

Il est clair que la procédure expérimentale n'est pas pratiquement réalisable. D'autre part, elle n'autorise pas l'utilisation du calcul des probabilités pour raisonner sur des événements incertains mais non répétables (et en pratique, aucun événement n'est parfaitement répétable). Enfin, elle est basée sur un cercle vicieux logique : cette définition repose sur la loi des grands nombres qui elle-même suppose déjà défini le concept de probabilité.

2.1.3.2 Le point de vue subjectiviste

Les faiblesses des deux approches précédentes et le fait que d'un point de vue logique il soit souhaitable de permettre la remise en question de la probabilité d'un événement suite à l'obtention de nouvelles informations (par exemple, si nous apprenons que le dé est imparfait) conduisent à nier l'existence de la notion de probabilité "objective".

Ainsi, dans la conception subjectiviste on modélise l'état de connaissance d'un observateur. On peut alors argumenter que pour être cohérent avec lui-même, un observateur doit assigner des probabilités aux événements qui respectent les axiomes de Kolmogorov, mais, différents observateurs, ayant éventuellement des connaissances différentes, peuvent aboutir à des assignations différentes. De plus, un même observateur peut remettre à jour ces probabilités lorsque de nouvelles informations se présentent.

Mesure d'incertitude. La probabilité objective n'existe pas et n'est donc pas une grandeur mesurable; la probabilité subjective est simplement une mesure d'incertitude, qui peut varier avec les circonstances et avec l'observateur.

Illustration. On considère un expérimentateur qui doit lancer une pièce et une personne qui assiste au lancer.

- le maître du jeu montre la pièce et demande à l'expérimentateur et au spectateur de préciser la probabilité pour qu'elle tombe sur "pile" au prochain lancer. Tous deux répondent 0.5. Cependant le maître du jeu sait que la pièce est truquée, et pense qu'elle tombe en moyenne sur pile 70% du temps.

- l'expérimentateur lance la pièce et regarde l'issue du lancer. Il demande au spectateur et au maître du jeu de désigner maintenant la probabilité pour qu'elle soit tombée sur pile : réponses 0.5 et 0.7 respectivement. Ensuite il leur dit qu'il pense que la vraie valeur est 1, puisqu'il a vu que la pièce est tombée sur pile.

- le jeu est répété 100000 fois, et à chaque fois l'expérimentateur révèle l'issue; sur les 100000 lancers, la pièce est tombée 78000 fois sur pile; quelles seraient maintenant les estimations émises par les trois personnes pour le lancer suivant ?

Extension aux expériences non répétables. Puisque la notion d'expérience répétable n'est pas exploitée dans la vision subjectiviste, on peut étendre le domaine d'application du calcul de probabilités aux événements non répétables, c'est-à-dire au raisonnement en présence d'incertitudes (par exemple en intelligence artificielle, pour modéliser le raisonnement humain).

La vision bayesienne. Cette approche est développée dans des enseignements plus avancés. Pour le moment, contentons nous d'indiquer que cette approche consiste à attribuer des probabilités à tout ce qui est incertain. En particulier, cette approche consiste à attribuer des lois de probabilités aux probabilités des événements, si les informations disponibles ne sont pas suffisantes pour déterminer leurs valeurs exactes.

Ainsi, en présence d'un problème de jeu de "pile ou face", un bayesien va commencer par admettre qu'il ne connait pas suffisamment bien la pièce pour fixer a priori la probabilité de "pile". En d'autres mots, il admet avoir une incertitude sur la valeur de cette probabilité, qu'il va modéliser par une (méta-) loi de probabilités. Ensuite, il va utiliser cette loi de probabilités pour faire des prédictions, et si des expériences sont effectuées (par exemple des lancers de pièce) il va utiliser le calcul des probabilités (la formule de Bayes) pour remettre à jour la valeur des méta-probabilités en fonction de l'issue de l'expérience.

Il faut remarquer que cette approche n'est pas non plus entièrement satisfaisante puisqu'il reste une phase arbitraire qui consiste à choisir les méta-probabilités. Signalons simplement que certains arguments de symétrie et d'"esthétique" sont utilisés par les bayesiens pour fixer de façon "objective" les méta-probabilités...

2.1.4 • Ensembles universels finis, dénombrables, et non-dénombrables

Avant de nous attaquer au coeur du calcul des probabilités, nous voulons faire ici quelques remarques générales sur l'intérêt et la nécessité de pouvoir manipuler des lois de probabilités définies sur des univers de taille infinie et non-dénombrables.

2.1.4.1 Cas discret : Ω fini ou dénombrable

Lorsque l'ensemble Ω est fini, l'algèbre des événements l'est également. Par contre, lorsque Ω est infini, il est possible d'y définir des algèbres d'événements finies, dénombrables ou non-dénombrables. Par exemple, si Ω est infini mais dénombrable (il peut être mis en bijection avec l'ensemble $\mathbb{N}$ des entiers naturels), la σ-algèbre complète 2^Ω est non-dénombrable (elle peut être mise en bijection avec l'ensemble $\mathbb{R}$ des nombres réels). Notons néanmoins que le passage d'un univers fini à un univers infini dénombrable ne conduit pas à des difficultés mathématiques supplémentaires. Lorsque Ω est soit fini, soit dénombrable, nous dirons que l'espace est **discret**; nous pouvons alors utiliser la σ-algèbre maximale $\mathcal{E} = 2^\Omega$, et pour choisir une loi P compatible avec les axiomes de Kolmogorov, il suffit de définir cette loi pour tous les événements élémentaires, c'est-à-dire d'associer à tous les singletons $\{\omega\} \subset \Omega$, une valeur $P(\{\omega\}) \in [0,1]$ de telle façon que $\sum_{\omega\in\Omega} P(\{\omega\}) = 1$. Toute partie E de Ω étant finie ou dénombrable, le troisième axiome de Kolmogorov implique alors que $P(E) = \sum_{\omega\in E} P(\{\omega\})$, et le premier et le second axiome sont automatiquement vérifiés.

2.1.4.2 Cas non-discret : Ω non-dénombrable

Dans les applications il est très souvent utile de considérer le cas où Ω est un ensemble non-dénombrable (par exemple $\mathbb{R}$, $\mathbb{R}^p$ ou bien des parties de ces espaces). Cependant, l'application rigoureuse du calcul des probabilités aux ensembles infinis non-dénombrables conduit à un certain nombre de difficultés techniques. En particulier, on peut montrer que dans le cas non-dénombrable, l'utilisation d'une σ-algèbre trop grande, telle que l'ensemble 2^Ω de toutes les parties de Ω peut conduire à des contradictions logiques. Afin d'éviter ce type de problèmes, on est amené à exclure certaines parties de Ω de l'ensemble $\mathcal{E}$ en les décrétant comme "non-mesurables".

Par exemple, si nous prenons comme univers Ω l'intervalle $[0,1]$ de la droite réelle, nous pouvons le munir de la plus petite σ-algèbre contenant tous les intervalles du type $[0,x]$ ($x \in [0,1[$) [3] et lui associer une loi de probabilité en définissant une fonction croissante $F(x)$, telle que $F(0) = 0$, et $\lim_{x\to 1} F(x) = 1$ et en postulant que $P([0,x]) = F(x)$. En particulier, on peut le munir de la loi de probabilité uniforme qui associe à un intervalle $[0,x]$ la probabilité $F(x) = x$. On montre que ce choix conduit à un modèle cohérent, et en fait qu'il n'existe qu'une seule loi de probabilité sur $[0,1]$ qui soit telle que $P([0,x]) = x, \forall x \in [0,1[$. Notons que dans ce modèle, toutes les parties de $[0,1]$ qui peuvent s'exprimer comme une union dénombrable d'intervalles (ouverts, semi-ouverts ou fermés) sont mesurables, mais qu'il est possible de construire des parties non mesurables dont la probabilité n'est donc pas définie dans ce modèle. Ce modèle est cependant suffisamment riche pour les besoins pratiques, tout ensemble "pratiquement intéressant" y étant mesurable. Remarquons aussi que dans ce modèle tous les singletons sont mesurables et selon la loi uniforme de probabilité nulle, mais pas impossibles.

2.1.4.3 Discussion

Le langage mathématique associé à la manipulation rigoureuse et générale des ensembles mesurables relève de la *théorie de la mesure* (voir [Bil79, Rom75]). Il permet d'écrire de façon synthétique des propriétés qui sont vraies pour les ensembles finis, dénombrables, et qui le restent lors du passage aux cas non-dénombrables. Le traitement complet du calcul des probabilités via la théorie de la mesure dépasse le cadre de ce cours introductif. Nous nous limitons à en fournir quelques notions élémentaires à l'appendice B, en les discutant intuitivement. Notons que le fait de ne pas maîtriser ce langage ne remet pas en question la signification des propriétés fondamentales du calcul des probabilités qui sont toutes déjà présentes dans le cadre des univers discrets.

D'un point de vue pratique, on peut d'ailleurs adopter le point de vue que le monde tel qu'il est accessible à l'expérimentation physique est essentiellement fini (c'est d'ailleurs évident en ce qui concerne le monde de l'informatique digitale). On pourrait donc parfaitement justifier une approche qui consisterait à développer les théories sur base de modélisations par ensembles finis, et qui expliciterait les passages à la limite sur les résultats plutôt que sur les concepts de départ. On pourrait alors se débarrasser des difficultés engendrées par l'analyse

moderne (calcul infinitésimal, théorie de la mesure, des distributions...) au prix d'une lourdeur d'écriture accrue (et souvent excessive) d'un certain nombre de propriétés et de raisonnements.

Nous pensons que l'analyse mathématique est un outil mathématique non seulement intéressant du point de vue conceptuel, mais dont l'utilisation est pleinement justifiée par son caractère opérationnel. Cependant, la compréhension des principes de base importants dans le domaine du calcul de probabilités peut par contre très bien se faire sans y faire appel à tour de bras. En clair, nous suggérons aux étudiants d'effectuer leurs raisonnements dans le cadre d'univers finis, afin de bien assimiler la signification mathématique et physique des principales notions. Une fois bien maîtrisé le cas fini, ils pourront ensuite se poser la question de savoir ce qui se passe lors du passage aux univers infinis dénombrables et non-dénombrables.

2.2 ELEMENTS DE BASE DU CALCUL DE PROBABILITES

2.2.1 Probabilités conditionnelles et indépendance d'événements

Partons d'un espace de probabilité $(\Omega, \mathcal{E}, P)$, et supposons que l'on sache qu'un événement B est réalisé. Cherchons à savoir ce que devient alors la probabilité qu'un événement A quelconque soit réalisé, que nous allons noter $P(A|B)$.

Si A et B sont incompatibles il est clair que A ne peut se réaliser simultanément avec B et on a donc dans ce cas que $P(A|B) = 0$. Par contre, si $A \cap B \neq \emptyset$, alors la réalisation de A est possible, mais seule la partie de A qui est dans B est réalisable. Si $A \cap B = B$ alors nous sommes certains que A se réalisera : $P(A|B) = 1$ dans ce cas. Tout se passe donc comme si nous avions restreint notre univers à l'événement B et que nous nous intéressions uniquement aux probabilités relatives des parties des événements situées dans B.

2.2.1.1 Probabilité conditionnelle

Nous supposerons que B est de probabilité non-nulle (dans le cas où Ω est fini, cela n'a pas de sens d'envisager qu'un événement de probabilité nulle se soit réalisé), et nous *définissons* la probabilité conditionnelle de A sachant que B est réalisé comme suit.

Probabilité conditionnelle de A sachant B

$$P(A|B) \stackrel{\triangle}{=} \frac{P(A \cap B)}{P(B)}. \tag{2.1}$$

Notons que $A \supset B \Rightarrow P(A|B) = 1$, mais (attention) la réciproque est fausse! (*Suggestion: se convaincre que ces affirmations sont vraies.)*

Exemple. Dans le triple lancer de pièce, définissons l'événement B par la condition "on observe exactement deux fois pile". Si nous supposons que toutes les 8 réalisations de cette expérience aléatoire sont équiprobables et donc de probabilité $P = 1/8$, nous pouvons calculer la probabilité de l'événement A décrit par la condition "le premier lancer donne face" conditionnellement à B par $P(A|B) = P(A \cap B)/P(B)$. En prenant en compte le fait que $P(B) = 3/8$ et que $P(A \cap B) = 1/8$ cela donne $P(A|B) = 1/3$. On peut comparer cette valeur avec $P(A) = 1/2$. On constate que le fait de savoir que dans le lancer de pièce il y a deux "pile" rend moins probable le fait que le premier résultat soit "face".

2.2.1.2 Mesure de probabilité conditionnelle

L'équation (2.1) permet d'associer, à partir de la connaissance de la loi P et pour un événement B fixé (de probabilité non-nulle), un nombre réel à chaque élément de $\mathcal{E}$.

Cette nouvelle loi, notée $P(\cdot|B)$, est une loi de probabilité. En effet, elle vérifie les axiomes de Kolmogorov puisque :

- $A, B \in \mathcal{E} \Rightarrow A \cap B \in \mathcal{E}$ et donc $P(A|B)$ est bien définie sur $\mathcal{E}$.
- $P(A|B) \geq 0$.
- $A \cap B \subset B \Rightarrow P(A \cap B) \leq P(B) \Rightarrow P(A|B) \leq 1$.
- $P(\Omega|B) = 1$ (trivial, puisque $\Omega \supset B, \forall B \in \mathcal{E}$).
- $P\left(\bigcup_i A_i|B\right) = \frac{P\left(\left(\bigcup_i A_i\right)\cap B\right)}{P(B)} = \frac{P\left(\bigcup_i A_i \cap B\right)}{P(B)} = \sum_i \frac{P(A_i \cap B)}{P(B)} = \sum_i P(A_i|B)$, car si les A_i sont incompatibles les $A_i \cap B$ le sont également.

Discussion. Il est important de remarquer que la loi de probabilité conditionnelle $P(\cdot|B)$ est bien définie sur l'entièreté de la σ-algèbre $\mathcal{E}$, et ceci bien que ses valeurs ne dépendent en fait que de probabilités de sous-ensembles de B.

Pour deux événements donnés A et B non indépendants on peut avoir soit $P(A|B) < P(A)$ ou $P(A|B) > P(A)$. Un événement peut donc devenir plus ou moins probable lorsque on dispose d'informations nouvelles. (4)

Dans le cas d'un univers *fini ou dénombrable* tous les événements possibles sont toujours de probabilité strictement positive, et induisent par conséquent une loi de probabilité conditionnelle en appliquant notre définition. Dans cette même situation, les événements de probabilité nulle sont impossibles, et il n'est donc pas dérangeant que la notion de probabilité conditionnelle par rapport à ces événements ne soit pas définie.

Par contre, dans le cas d'un univers *infini non-dénombrable* (cf. l'exemple de l'intervalle $[0, 1]$ discuté plus haut), le fait qu'un événement se réalise n'implique pas nécessairement que sa probabilité a priori est non-nulle. Par exemple, on peut très bien imaginer que l'ensemble B est formé des puissances entières de $\frac{1}{2}$ et se poser la question de savoir quelle est sous cette condition la probabilité que $\omega \in [0, 0.1]$. Pour répondre à ce genre de questions, il est nécessaire de définir la notion de probabilité conditionnelle vis-à-vis d'événements a priori de probabilité nulle. Cette extension rigoureuse de la notion de loi de probabilité conditionnelle dans le contexte d'univers infinis non-dénombrables nécessite cependant le recours à la théorie de la mesure, ce qui dépasse le cadre de ce cours introductif.

2.2.1.3 Notion d'événements indépendants

A partir de la notion de probabilité conditionnelle on définit la notion d'événements indépendants, comme suit.

Evénements indépendants (première définition)

On dit que A est indépendant de B si $P(A|B) = P(A)$, c'est-à-dire si le fait de savoir que B est réalisé ne change en rien la probabilité de A. Pour dire que A est indépendant de B on utilisera la notation

$$A \perp B. \tag{2.2}$$

Notons que si $P(B) \in]0, 1[$, on a A indépendant de B si, et seulement si, $P(A|B) = P(A|B^c)$.
(*Suggestion : calculer alors $P(A)$ par le théorème des probabilités totales.*)

Exemple. Dans le triple lancer de pièce, définissons l'événement B par la condition "le second lancer donne pile" et l'événement A par la condition "le troisième lancer donne face". En faisant l'hypothèse que les 8 issues de l'expérience sont équiprobables, on calcule que $P(A) = P(A|B) = 1/2$. Les deux événements A et B sont donc indépendants sous cette hypothèse.

2.2.1.4 Indépendance conditionnelle

On peut étendre la notion d'indépendance à la notion d'indépendance conditionnelle, comme suit.

Conditionnement multiple et indépendance conditionnelle

Soient A, B, C trois événements avec $P(B \cap C) \neq 0$.
Alors, si $P(A|B \cap C) = P(A|C)$ on dit que A *est indépendant de* B *conditionnellement à* C, ce que l'on note par

$$A \perp B \mid C. \tag{2.3}$$

Exemple. Dans le triple lancer de pièce, définissons l'événement B par la condition "le second lancer donne pile" et l'événement A par la condition "le troisième lancer donne face" et l'événement C par " le premier lancer donne face". En postulant toujours que les 8 issues de l'expérience sont équiprobables, on calcule que $P(A|B, C) = P(A|C) = 1/2$. Les deux événements A et B sont donc conditionnellement indépendants sachant que C se réalise.

Discussion. Remarquons que

$$P(A|B \cap C) \stackrel{\triangle}{=} \frac{P(A \cap (B \cap C))}{P(B \cap C)} = \frac{P((A \cap B) \cap C)}{P(C)} \frac{P(C)}{P(B \cap C)} = \frac{P(A \cap B|C)}{P(B|C)}.$$

De la même façon, on peut se convaincre que

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

On constate ainsi que le conditionnement *simultané* par rapport à la réalisation conjointe des deux événements B et C est équivalent à deux opérations successives de conditionnement, d'abord sur l'un des événements en partant de la loi P, puis sur l'autre événement en partant de la loi conditionnelle induite par le premier. Cela justifie les notations suivantes:

$$P(A|B \cap C) = P(A|B, C) = P(A|C \cap B) = P(A|C, B),$$

que nous utiliserons dans la suite de cet ouvrage.

Intuitivement, cette propriété de symétrie correspond au fait que notre perception de la probabilité conditionnelle d'un événement (ici A) ne dépend pas de l'ordre dans lequel nous avons eu connaissance d'informations partielles sur le résultat de l'expérience (ici le fait que $\omega \in B$ et le fait qu'aussi $\omega \in C$).

2.2.2 Sur la notion d'indépendance

La notion d'indépendance est une notion centrale en théorie des probabilités. Aussi allons-nous détailler les diverses propriétés immédiates qui découlent de sa définition. Nous supposerons ci-dessous que $P(A) \in]0, 1[$ (resp. $P(B) \in]0, 1[$), et dans le cas contraire nous dirons que A (resp. B) est un événement trivial. Nous laissons au lecteur le soin de vérifier dans quels cas (et comment) ces conditions peuvent être relaxées à des événements triviaux.

Propriétés "positives". Nous demandons au lecteur de démontrer, à titre d'exercice immédiat, celles parmi les propriétés suivantes dont nous ne donnons pas la preuve.

- $\emptyset$ est indépendant de tout autre événement.
- Un événement de probabilité nulle est indépendant de tout autre événement : $P(A) = 0 \Rightarrow P(A \cap B) = 0 \Rightarrow P(A|B) = 0$.
- Tout événement est indépendant de Ω.
- Tout événement est indépendant de tout événement certain : $P(A) = 1 \Rightarrow P(A \cap B) = P(B)$ et donc $P(B|A) = P(B)$.

- "A indépendant de B" $\Leftrightarrow P(A \cap B) = P(A)P(B)$
  (conséquence directe de la définition, lorsque $P(B) > 0$; si $P(B) = 0$, la condition $P(A \cap B) = P(A)P(B)$ est encore vérifiée).

- "A indépendant de B" $\Rightarrow$ "B indépendant de A".

- "A indépendant de B" $\Rightarrow$ "A^c indépendant de B".

- "A indépendant de B" $\Rightarrow$ "A indépendant de B^c".

- "A indépendant de B" $\Rightarrow$ "A^c indépendant de B^c".

On peut donc utiliser en lieu et place de la définition de l'indépendance la définition suivante.

Evénements indépendants (seconde définition)

Deux événements A et B sont indépendants si et seulement si $P(A \cap B) = P(A)P(B)$.

Il est à noter que cette définition couvre le cas où $P(A)$ et/ou $P(B)$ sont nulles, la propriété étant trivialement vérifiée dans ces cas (car $0 \leq P(A \cap B) \leq \min\{P(A), P(B)\}$).

Propriétés "négatives". L'assimilation de celles-ci est au moins aussi importante pour la bonne compréhension de la notion d'indépendance que l'assimilation des propriétés positives.

Remarquons tout d'abord que des événements indépendants pour une loi de probabilité donnée peuvent très bien être non indépendants pour une autre loi de probabilité. En d'autres mots, la propriété d'indépendance dépend bien du choix de la loi de probabilité et pas seulement des propriétés ensemblistes.

Nous suggérons au lecteur de chercher des contre-exemples pour démontrer les propriétés négatives suivantes.

- Un événement quelconque non trivial n'est jamais indépendant de lui-même!

- A indépendant de B et B indépendant de $C \not\Rightarrow A$ indépendant de C.
  (Suggestion : à titre de contre-exemple, prendre A et B tous deux de probabilité non nulle et indépendants, et puis considérer la cas où $C = A$).

- A dépendant de B et B dépendant de $C \not\Rightarrow A$ dépendant de C.
  (Suggestion : prendre A et C indépendants, et $B = A \cap C$ en supposant que $P(B) > 0$.)

- A indépendant de $B \not\Rightarrow A$ indépendant de B conditionnellement à C.
  (Suggestion : prendre comme exemple le double "pile ou face" avec une pièce équilibrée, comme événements A "face au premier lancer", B "face au second lancer", C "même issue aux deux lancers").

Indépendance mutuelle de plusieurs événements. On peut étendre la seconde définition de l'indépendance au cas de n événements.

Indépendance mutuelle de n événements.

On dira que les événements $A_1, A_2, \ldots, A_n$ sont *mutuellement* indépendants si pour toute partie I de l'ensemble des indices allant de 1 à n on a :

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i). \tag{2.4}$$

Il est important de noter que l'indépendance mutuelle est une condition plus forte que l'indépendance deux à deux, qui elle est définie comme suit.

Indépendance deux à deux de n événements.

On dira que les événements $A_1, A_2, \ldots, A_n$ sont indépendants *deux à deux* si pour $\forall i \neq j$ on a :

$$P(A_i \cap A_j) = P(A_i)P(A_j). \tag{2.5}$$

Pour se convaincre que l'indépendance "deux à deux" n'implique pas l'indépendance "mutuelle", il suffit de reconsidérer notre double lancer de pile ou face ci-dessus. Dans cet exemple on a en effet, A indépendant de B, B indépendant de C, et C indépendant de A, alors que C n'est pas indépendant $A \cap B$.

Autres formules utiles.

$$P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C) = P(A|B,C)P(B|C)P(C) = P(A|C,B)P(C|B)P(B) \tag{2.6}$$

$$P(A \cap B|C) = P(A|C)P(B|C \cap A) = P(A|C)P(B|C,A) = P(A|C)P(B|A,C) \tag{2.7}$$

Simplification des notations. Dans la suite nous utiliserons de façon interchangeable les notations suivantes pour désigner l'occurrence simultanée de plusieurs événements :

- $A_1 \cap A_2 \cap \ldots \cap A_n$: la notation ensembliste (on insiste sur le fait que les A_i sont vus comme des ensembles).
- $A_1 \wedge A_2 \wedge \ldots \wedge A_n$: la notation logique (on insiste sur le fait que les A_i sont vus comme des formules logiques).
- $A_1, A_2, \ldots, A_n$: une notation plus compacte.

2.2.3 Formules de Bayes

Les formules de Bayes permettent d'exprimer $P(A|B)$ en fonction de $P(B|A)$.

Première formule de Bayes

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{2.8}$$

Il s'agit d'une conséquence immédiate de la définition de la notion de probabilité conditionnelle, sous l'hypothèse que $P(A)$ et $P(B)$ sont non nulles.

En exploitant la définition de la probabilité conditionnelle on peut aussi reformuler le théorème des probabilités totales comme suit.

Théorème des probabilités totales (version 2)

Si $B_1, B_2, \ldots, B_n$ est un système complet d'événements non triviaux alors le théorème des probabilités totales peut s'écrire sous la forme suivante

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i). \tag{2.9}$$

Il s'agit d'une conséquence immédiate de la définition de la notion de probabilité conditionnelle, sous l'hypothèse que les $P(B_i)$ sont non nulles, et de la première version du théorème des probabilités totales.

Deuxième formule de Bayes. Dès lors la formule de Bayes peut aussi s'écrire sous la forme suivante.

Deuxième formule de Bayes

Si $B_1, B_2, \ldots, B_n$ est un système complet d'événements non triviaux alors la première formule de Bayes peut s'écrire sous la forme suivante

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{k=1}^{n} P(A|B_k)P(B_k)}, \tag{2.10}$$

qui s'appelle aussi théorème sur la "probabilité des causes", car il permet de calculer les probabilités des causes possibles d'un événement sachant qu'une conséquence s'est réalisée, connaissant la probabilité de cette dernière sous l'hypothèse de chaque cause et connaissant la probabilité des causes a priori.

Exemple. Dans le triple lancer de pièce on souhaite calculer la probabilité de tomber exactement deux fois sur "pile" sachant que le premier lancer a donné "face". Nous supposons comme avant que les 8 issues sont équiprobables, et dans ce cas on peut a priori penser que la probabilité que nous cherchons à déterminer vaut $1/4$. Appliquons cependant ce que nous venons d'apprendre pour faire le calcul de façon à éviter de se tromper.

Nous désignons par A l'événement "le premier lancer donne face", par B_1 l'événement "on tombe exactement une fois sur pile", par B_2 le cas où "on tombe exactement 2 fois sur pile", par B_3 le cas où "on tombe trois fois sur pile", et par B_4 le cas où "on ne tombe aucune fois sur pile".

On a $P(A|B_1) = 2/3$ (les réalisations correspondant à B_1 sont $\{PFF, FPF, FFP\}$, dont deux sur trois correspondent à "face" au premier lancer), $P(A|B_2) = 1/3$ (cf le calcul fait ci dessus), $P(A|B_3) = 0$ (si on tombe trois fois sur pile, le premier lancer ne peut certainement pas donner face), $P(A|B_4) = 1$ (si on ne tombe aucune fois sur pile, le premier lancer donne certainement face).

On a aussi que $P(B_1) = 3/8$, $P(B_2) = 3/8$, $P(B_3) = 1/8$ et $P(B_4) = 1/8$. (Notez que la somme de ces 4 valeurs donne bien 1.)

Ces informations permettent donc de calculer $P(B_2|A)$ par

$$P(B_2|A) = \frac{P(A|B_2)P(B_2)}{\sum_{k=1}^{4} P(A|B_k)P(B_k)} = \frac{1/3 \times 3/8}{(2/3 \times 3/8) + (1/3 \times 3/8) + (0 \times 1/8) + (1 \times 1/8)} = 1/4.$$

On trouve ce que notre intuition nous dictait au départ, ce qui nous rassure. Cette intuition était en fait justifiée parce que les 8 issues de l'expérience aléatoire sont équiprobables (ce qui implique l'indépendance des trois résultats successifs de lancer).

Cependant, nous pouvons aussi entrevoir dans cette démarche comment on peut faire pour calculer la même probabilité, même si les 8 issues de notre expérience ne sont pas équiprobables, et/ou si les trois lancers ne sont pas indépendants.

Discussion. Le théorème de Bayes (on donne le nom de théorème de Bayes aux deux formules de Bayes) joue un rôle très important dans le cadre du calcul des probabilités. Il sert de fondement au raisonnement incertain probabiliste et est à la base de toute une branche de la statistique appelée *statistique bayesienne*.

Il permet de remettre à jour les probabilités d'un certain nombre d'alternatives B_i en fonction d'informations nouvelles (le fait que A soit réalisé). On utilise souvent le terme de *probabilités a priori* pour désigner les $P(B_i)$ et le terme de *probabilités a posteriori* pour désigner les $P(B_i|A)$.

Par exemple, dans le cadre du diagnostic médical cette formule permet à un médecin de remettre à jour la plausibilité de certaines maladies (désignées par les B_i) à partir des symptômes observés (désignés conjointement par A), partant d'une connaissance de la probabilité a priori d'observer les différentes maladies (obtenues par exemple en effectuant des statistiques) et une connaissance des probabilités d'observer les symptômes A pour chacune de ces maladies (obtenues également par application de méthodes statistiques). Ce type de raisonnement, appelé inférence probabiliste, est une extension de la logique classique au raisonnement en présence d'incertitudes.

Nous verrons que le théorème de Bayes peut aussi s'appliquer dans le cas où les causes B_i sont en nombre infini éventuellement non-dénombrable.

2.3 • ESPACES DE PROBABILITÉ PRODUITS

Nous introduisons ci-dessous quelques notions et terminologies qui seront utilisées et illustrées plus loin, notamment dans le cadre de l'étude des variables aléatoires dans les deux chapitres suivants.

2.3.1 Construction d'un espace produit à partir de modules plus simples

Etant donnés un nombre fini d'espaces de probabilité $(\Omega_i, \mathcal{E}_i, P_i)$ $(i = 1, \ldots, n)$ on peut définir un espace de probabilité produit $(\Omega, \mathcal{E}, P)$, de la manière suivante :

- **Univers produit :** $\Omega = \Omega_1 \times \ldots \times \Omega_n$, (produit cartésien classique) : un élément ω de Ω est un n-tuple $(\omega_1, \ldots, \omega_n)$ construit en combinant n éléments ω_i, avec ω_i choisi dans $\mathcal{E}_i$.
- **σ-algèbre produit :** $\mathcal{E}$ est l'ensemble des parties de Ω qui peuvent s'écrire sous la forme d'une union dénombrable d'ensembles disjoints $A_j \subset \Omega$ de la forme $A_j = A_{1,j} \times \ldots \times A_{n,j}$ avec $A_{i,j} \in \mathcal{E}_i, \forall i = 1, \ldots, n$. (Voir aussi l'appendice B.1 pour plus de précision à ce sujet).
- **Mesure de probabilité produit :** Pour un élément A_j de $\mathcal{E}$ qui s'écrit sous la forme $A_j = A_{1,j} \times \ldots \times A_{n,j}$, on définit $P(A_j) = \prod_{i=1}^{n} P_i(A_{i,j})$. Pour un élément A quelconque de $\mathcal{E}$ s'écrivant sous la forme d'une union finie ou dénombrable de tels ensembles disjoints (i.e. $A = \bigcup_j A_j$) on définit sa probabilité par $P(A) = \sum_j P(A_j)$. (Bien qu'un élément de $\mathcal{E}$ puisse s'exprimer de nombreuses façons différentes de cette façon, la probabilité calculée ainsi donne toujours la même valeur.)

On peut se convaincre que cette définition conduit bien à un espace de probabilité. On dira que les Ω_i sont les axes "orthogonaux" de l'espace produit et on parlera de la projection d'un événement A sur les axes pour désigner les A_i, et d'événements parallèles à un axe i si $A_i = \Omega_i$. On peut alors montrer que si deux événements de l'espace produit sont parallèles à des ensembles d'axes complémentaires, ils sont indépendants. En d'autres termes, si un événement ne spécifie rien selon un certain nombres d'axes, alors le fait de savoir qu'un événement soit réalisé qui ne spécifie que de l'information relative à ces axes ne fournit aucune information sur cet événement.

Exemple.

Partant d'un espace $(\Omega_1, \mathcal{E}_1, P_1)$ correspondant à un lancer d'un dé à six faces, et d'un espace $(\Omega_2, \mathcal{E}_2, P_2)$ correspondant au lancer d'une pièce de monnaie, on peut construire l'espace de probabilité produit $(\Omega, \mathcal{E}, P)$, avec $\Omega = \Omega_1 \times \Omega_2, \mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$ et $P = P_1 P_2$, correspondant à une expérience où on lance à la fois un dé et une pièce. Dans ce cas, l'ensemble $\mathcal{E}$ sera composé des sous-ensembles de $\Omega_1 \times \Omega_2$ qui se décrivent au moyen d'affirmations logiques (ici en nombre fini, vu que Ω est aussi fini) concernant les résultats des deux expériences. Le fait que le lancer de dé et de pièce n'interagissent pas physiquement justifie l'utilisation de la loi produit $P = P_1 P_2$ pour caractériser cette expérience.

2.3.2 Séries d'épreuves identiques et indépendantes

Un cas particulièrement intéressant en pratique d'espace produit est celui où tous les $(\Omega_i, \mathcal{E}_i, P_i(\cdot))$ sont identiques. Un tel type d'espace produit permet de modéliser les séries d'épreuves identiques et indépendantes, rencontrées en théorie de l'échantillonnage et à la base des statistiques.

Par exemple, partant de la définition d'une expérience relative au lancer d'une pièce, décrite par un espace de probabilité, il sera intéressant de considérer la modélisation d'une expérience consistant à lancer n fois successivement une pièce, en supposant que les lancers sont indépendants et interchangeables. L'étude de cette expérience "répétée" permettra de mettre en évidence des propriétés intéressantes qui concernent l'ensemble des réalisations d'un nombre croissant de lancer de pièces, et dont l'exploitation est fondamentalement à la base de la statistique et des techniques de simulation de Monte Carlo.

2.3.3 Factorisation d'un espace complexe sous forme de produit de facteurs simples

Dans certains cas il est possible de factoriser un espace de départ en effectuant l'opération inverse, c'est-à-dire de l'écrire sous la forme du produit cartésien d'espaces indépendants (non nécessairement identiques).

Dans les chapitres suivants de ces notes, nous verrons que la construction d'un modèle probabiliste sous la forme d'un produit de facteurs simples est l'approche essentielle de modélisation nécessaire pour mettre en oeuvre les méthodes probabilistes pour résoudre les problèmes de plus en plus complexes rencontrés en ingénierie.

(Suggestion : partir d'un espace fini dont on suppose que l'algèbre est engendré par deux événements indépendants A et B, et montrer qu'il peut se factoriser en deux axes correspondant à ces événements.)

2.3.4 Marginalisation

Partant d'un espace produit, il est possible de reconstituer les espaces produits correspondant à un sous-ensemble de ses axes par une opération de projection. En calcul de probabilité on dit qu'on "marginalise" les autres axes. Cette opération est intéressante dans un contexte où partant d'un modèle complexe, on souhaite en déduire un modèle plus simple ne faisant intervenir que certains aspects (les axes retenus) nécessaires pour la résolution d'un problème particulier.

Notons que cette opération de marginalisation peut s'effectuer sur un espace produit dont la loi de probabilité n'est pas factorisable de façon simple. Nous reviendrons sur cette opération de marginalisation dans les chapitres suivants.

2.4 LE PROBLÈME DU MONTY HALL

Dans cette section nous abordons un premier problème de raisonnement probabiliste, dans le but de mettre en évidence certaines techniques de résolution de base et attirer l'attention des étudiants sur la nécessité de bien préciser l'énoncé d'un problème avant de tenter de le résoudre. Le texte qui suit en italique est repris de la référence [LL04]. Nous conseillons de le lire attentivement et d'en méditer le contenu.

"In the September 9, 1990 issue of Parade magazine, the columnist Marilyn vos Savant responded to this letter:

> Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say number 1, and the host, who knows what's behind the doors, opens another door, say number 3, which has a goat. He says to you, "Do you want to pick door number 2?" Is it to your advantage to switch your choice of doors?
>
> Craig. F. Whitaker
> Columbia, MD

The letter roughly describes a situation faced by contestants on the 1970's game show Let's Make a Deal, hosted by Monty Hall and Carol Merrill. Marilyn replied that the contestant should indeed switch. But she soon received a torrent of letters - many from mathematicians - telling her that she was wrong. The problem generated thousands of hours of heated debate.

Yet this is is an elementary problem with an elementary solution. Why was there so much dispute? Apparently, most people believe they have an intuitive grasp of probability. (This is in stark contrast to other branches of mathematics; few people believe they have an intuitive ability to compute integrals or factor large integers!) Unfortunately, approximately 100% of those people are wrong. In fact, everyone who has studied probability at length can name a half-dozen problems in which their intuition led them astray? often embarassingly so.

The way to avoid errors is to distrust informal arguments and rely instead on a rigorous, systematic approach. In short: intuition bad, formalism good. If you insist on relying on intuition, then there are lots of compelling financial deals we'd love to offer you!"

2.4.1 Description précise du problème

Le texte qui précède nous avertit sur plusieurs aspects fondamentaux associés à la mise en oeuvre du calcul de probabilités :

- l'intuition est souvent de mauvais conseil, puisque des esprits brillants se sont fait "avoir" par un problème aussi élémentaire que le "Monty Hall";
- la modélisation précise du problème est absolument indispensable avant d'aborder toute tentative de résolution;

- la mise en oeuvre systématique du raisonnement déductif probabiliste est nécessaire et suffisante pour résoudre de tels problèmes.

Dans la suite nous allons illustrer ces idées, en montrant au passage que l'essentiel du raisonnement probabiliste est déjà contenu dans le raisonnement logique "classique".

Commençons par la modélisation du problème, en précisant suffisamment les hypothèses (en faisant cela, nous interprétons d'une manière précise, mais potentiellement erronée l'énoncé du problème; si notre interprétation est erronée, nous sommes prêts à la changer, mais ceci est une autre histoire). Nous considérons donc que

- lors du premier choix du joueur, celui-ci n'a aucune information autre que les règles du jeu (indiquées ci-dessous); il peut à ce stade choisir n'importe quelle porte;
- Les autres règles du jeu affirment que
  - seule une porte cache un lot intéressant (le lot intéressant est la voiture); a priori, aucune information n'est disponible pour le joueur qui lui permettrait de considérer qu'une des trois portes est plus susceptible de révéler la voiture;
  - suite au choix du joueur à la première étape du jeu, le présentateur est *obligé* de choisir une porte qui ne révèle pas la voiture et il ne peut pas non plus choisir d'ouvrir la porte désignée par le joueur au premier tour. (On suppose donc aussi que l'animateur connait la disposition des lots derrière les portes.)
  - si, compte tenu des règles qui précèdent, le présentateur dispose encore de la liberté de choisir n'importe laquelle des deux portes restantes, il fait ce choix au hasard.
  - après l'ouverture de la porte choisie par le présentateur, le joueur peut soit maintenir son premier choix, soit choisir la troisième porte (celle restant après son choix numéro un et le choix, numéro deux, du présentateur). A ce stade, il ne sait pas si la porte qu'il a choisie au départ cache le lot intéressant.

Etant données ces règles, la question posée est celle de la détermination d'une stratégie optimale pour le joueur; cette stratégie est composée d'une décision pour le choix de la porte à la première opportunité et d'une règle de décision pour choisir la seconde porte en fonction de la réaction de l'animateur. Elle serait optimale, si par rapport à toute autre stratégie de décision elle conduit avec une plus grande probabilité à choisir la bonne porte au deuxième coup.

2.4.2 Modélisation du problème au moyen d'un arbre de scénarios

Pour résoudre le problème du Monty Hall, nous allons considérer un espace de probabilité qui modélise les étapes successives du problème, à savoir

- **Etape 1.** Choix de disposition des lots derrière les 3 portes: il s'agit de choisir la porte $i \in \{1, 2, 3\}$ qui cache la voiture, et nous supposons que ce choix est fait aléatoirement. Nous désignons ce choix par la variable $v \in \{1, 2, 3\}$ (i.e. chaque porte à la même probabilité, $P(v = i) = 1/3$ de cacher la voiture). Seul le présentateur est au courant de ce choix.
- **Etape 2.** Désignation d'une première porte par le joueur; nous désignons ce choix par une variable j_1 pouvant prendre ses valeurs dans l'ensemble $\{1, 2, 3\}$.
- **Etape 3.** Choix d'une seconde porte par l'animateur (au hasard si les deux portes restantes ne cachent pas le gros lot, sinon choix de la seule porte restante ne cachant pas le gros lot). Nous désignons cette variable par a; ses valeurs possibles dépendent à la fois de v et de j_1.
- **Etape 4.** Désignation de la porte finale par le joueur, en fonction de son premier choix (j_1 à l'étape 2) et du choix de l'animateur (a à l'étape 3). Nous désignons par j_2 cette variable.

L'arbre de scénarios de la Figure 2.1 représente l'ensemble des possibilités résultant de la combinaison des quatre étapes du problème.

Comme nous allons le voir, l'analyse de ce modèle montre que si le joueur choisit la décision $j_1 = 1$ (choix de la première porte à la première étape du jeu), il a intérêt à changer de décision à l'étape suivante, quelle que soit la

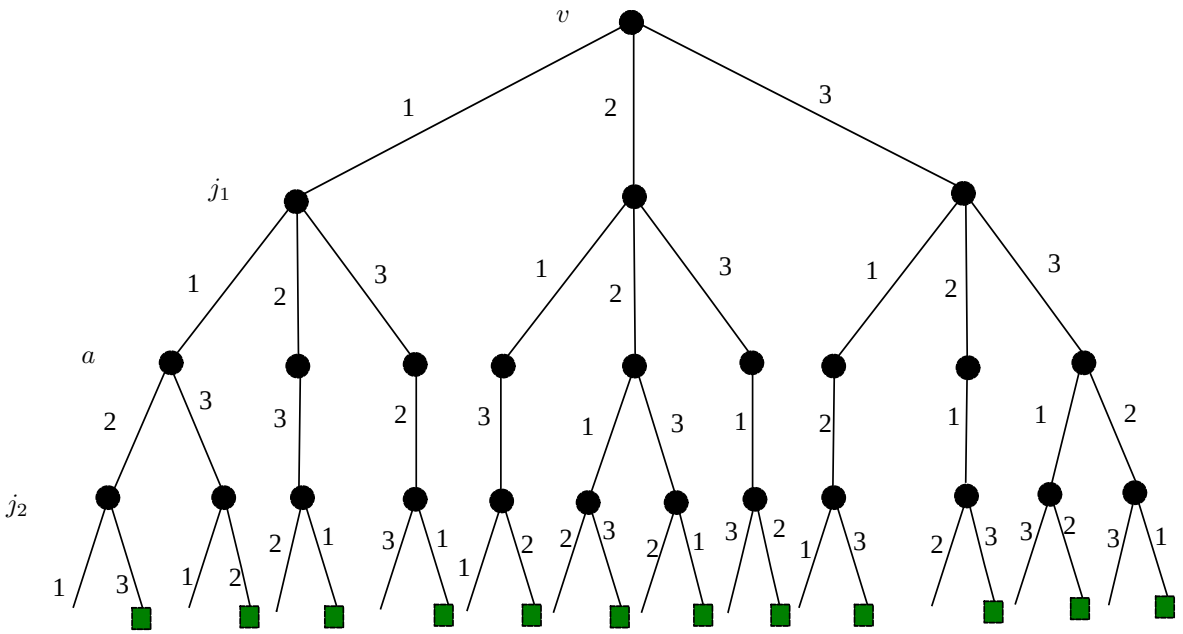


Figure 2.1: Arbre de scénarios pour le jeu du Monty Hall. A la dernière étape, il reste deux possibilités pour le joueur: soit il maintient son premier choix ($j_2 = j_1$), soit il révise son choix ($j_2 = \{1,2,3\} \setminus \{j_1, a\}$). La politique de décision j_2 qui consiste à systématiquement réviser le choix est indiquée par les feuilles de l'arbre marquées par un carré vert plein. Les scénarios gagnants sont ceux où $j_2 = v$.

réaction de l'animateur (supposée conforme aux règles du jeu). En effet, en changeant de décision il augmente la probabilité de choisir la bonne porte de $1/3$ à $2/3$. De même, s'il avait choisi la porte 2 à la première opportunité, ou bien la porte 3, la stratégie qui consiste à changer d'avis est encore gagnante avec une probabilité de 2/3. Enfin, si le joueur choisit au hasard la seconde fois, entre les deux possibilités qui s'offrent à lui, il gagnera avec une probabilité de 1/2 (ce qui est moins bon que de changer d'avis, mais meilleur que de rester sur son premier choix).

2.4.3 Evaluation des probabilités de chaque scénario en fonction de la stratégie du joueur

Nous allons successivement analyser trois stratégies, parmi celles-possibles pour le joueur. Chacune de ces stratégies donnera lieu à un modèle probabiliste différent. Dans la section suivante, nous calculerons pour chacun de ces modèles la probabilité de gagner.

2.4.3.1 Stratégie de jeu totalement aléatoire

Dans cette stratégie, le joueur choisit avec une probabilité de 1/3 une des trois portes à la première étape, puis avec une probabilité de 1/2 une des deux possibilités qui s'offrent à lui à la seconde étape.

La probabilité d'observer la séquence $(1,1,2,1)$ correspondant à la feuille de gauche de l'arbre de scénarios de la figure 2.1, s'obtient alors comme le produit de

$$P(v=1)P(j_1=1)P(a=2|v=1,j_1=1)P(j_2=1) = \frac{1}{3}\frac{1}{3}\frac{1}{2}\frac{1}{2} = \frac{1}{36}.$$

Le même raisonnement permet de se convaincre que les probabilités de chacune des trois feuilles suivantes, correspondant respectivement aux scénarios $(1,1,2,3)$, $(1,1,3,1)$ et $(1,1,3,2)$ sont aussi de $\frac{1}{36}$.

Les deux scénarios suivants, à savoir $(1,2,3,2)$ et $(1,2,3,1)$, sont quant à eux de probabilité égale à $\frac{1}{18}$, puisque dans ces scénarios l'animateur n'a finalement pas de choix réel possible, la seule possibilité pour lui étant d'ouvrir la porte numéro 3.

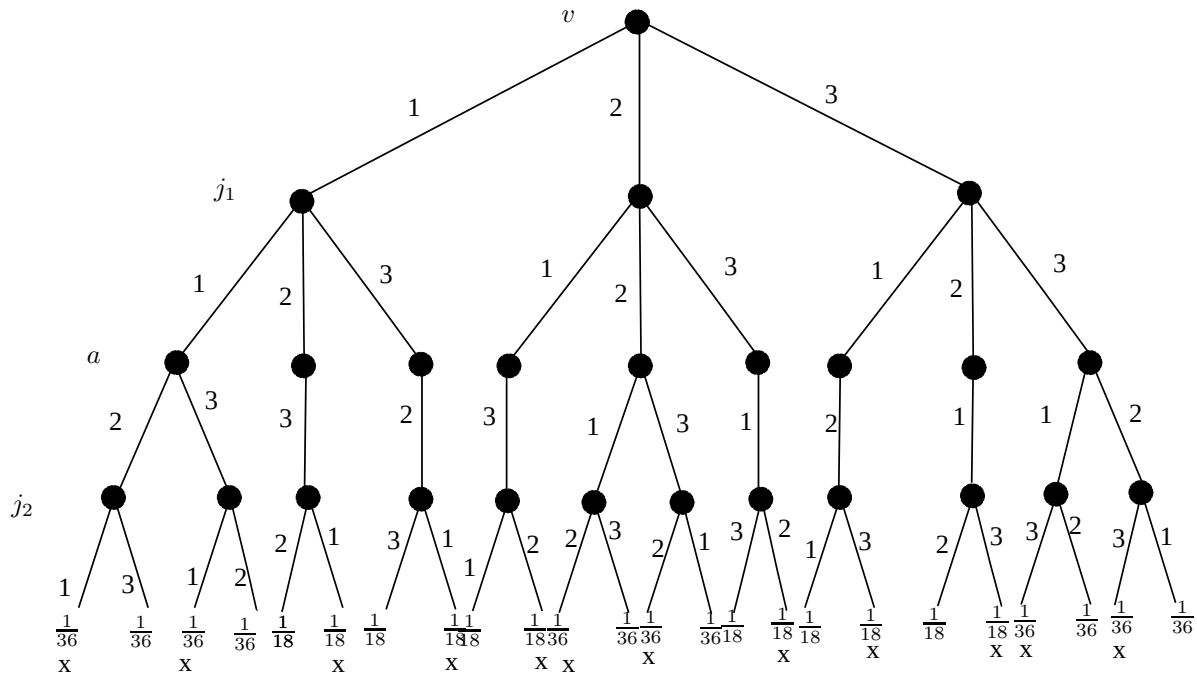


Figure 2.2: Probabilités associées aux feuilles de l'arbre de scénarios pour le jeu du Monty Hall, lorsque le joueur applique la stratégie totalement aléatoire. Les scénarios gagnants sont ceux où $j_2 = v$, et sont indiqués par des croix. La probabilité de remporter le lot vaut $\frac{6}{36} + \frac{6}{18} = \frac{1}{2}$.

Poursuivant ces calculs, pour l'ensemble des feuilles de l'arbre, nous obtenons la loi de probabilité pour l'ensemble des scénarios possibles. Ceci est représenté à la Figure 2.2.

2.4.3.2 Stratégie de jeu têtue

Dans la stratégie de jeu têtue, le joueur se fixe une fois pour toutes (c'est-à-dire, dès l'étape j_1) sur le choix de la porte et ne change plus d'avis par après. De plus, nous supposons qu'il choisit toujours la porte numéro 1.

Sous cette hypothèse, la probabilité d'observer la séquence $(1, 1, 2, 1)$ correspondant à la feuille de gauche de l'arbre de scénarios de la figure 2.1, s'obtient alors comme le produit de

$$P(v = 1)P(j_1 = 1)P(a = 2|v = 1, j_1 = 1)P(j_2 = 1) = \frac{1}{3}1\frac{1}{2}1 = \frac{1}{6}.$$

La feuille suivante $(1, 1, 2, 3)$ est quant à elle de probabilité nulle.

Poursuivant ces calculs, pour l'ensemble des feuilles de l'arbre, nous obtenons la loi de probabilité pour l'ensemble des scénarios possibles. Ceci est représenté à la Figure 2.3, où nous avons omis de représenter les probabilités de valeur nulle.

2.4.3.3 Stratégie de jeu versatile

Dans cette stratégie le joueur change systématiquement d'avis au second coup. Toujours sous l'hypothèse où il joue systématiquement la porte numéro 1 au premier coup, on obtient l'arbre de probabilités de la Figure 2.4, où nous avons aussi omis de représenter les probabilités de valeur nulle.

2.4.4 Calcul de la probabilité de remporter le lot selon une stratégie de jeu donnée

Pour obtenir la probabilité de remporter le lot, pour une stratégie donnée, il suffit de calculer la somme des probabilités associées par cette stratégie aux feuilles de l'arbre qui correspondent au gain du lot, c'est-à-dire correspondant aux scénarios (v, j_1, a, j_2) tels que $j_2 = v$.

2.22

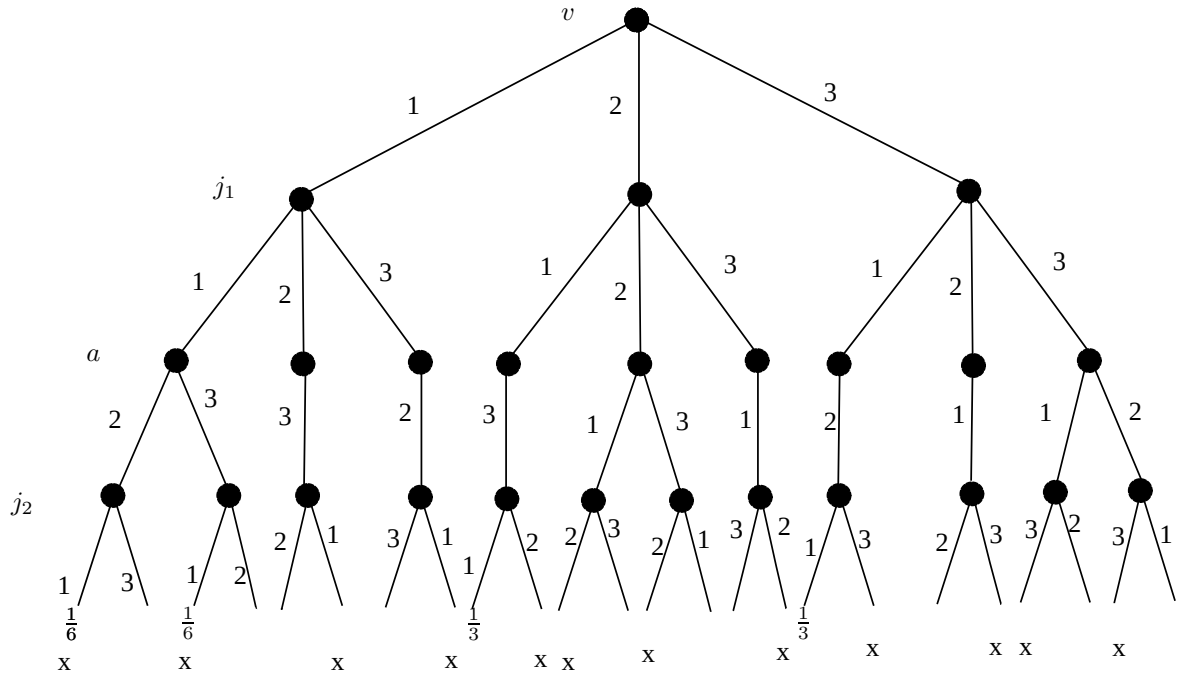


Figure 2.3: Probabilités associées aux feuilles de l'arbre de scénarios pour le jeu du Monty Hall, lorsque le joueur applique la stratégie têtue. Les scénarios gagnants sont ceux où $j_2 = v$, et sont indiqués par des croix. La probabilité de remporter le lot vaut $\frac{2}{6} = \frac{1}{3}$.

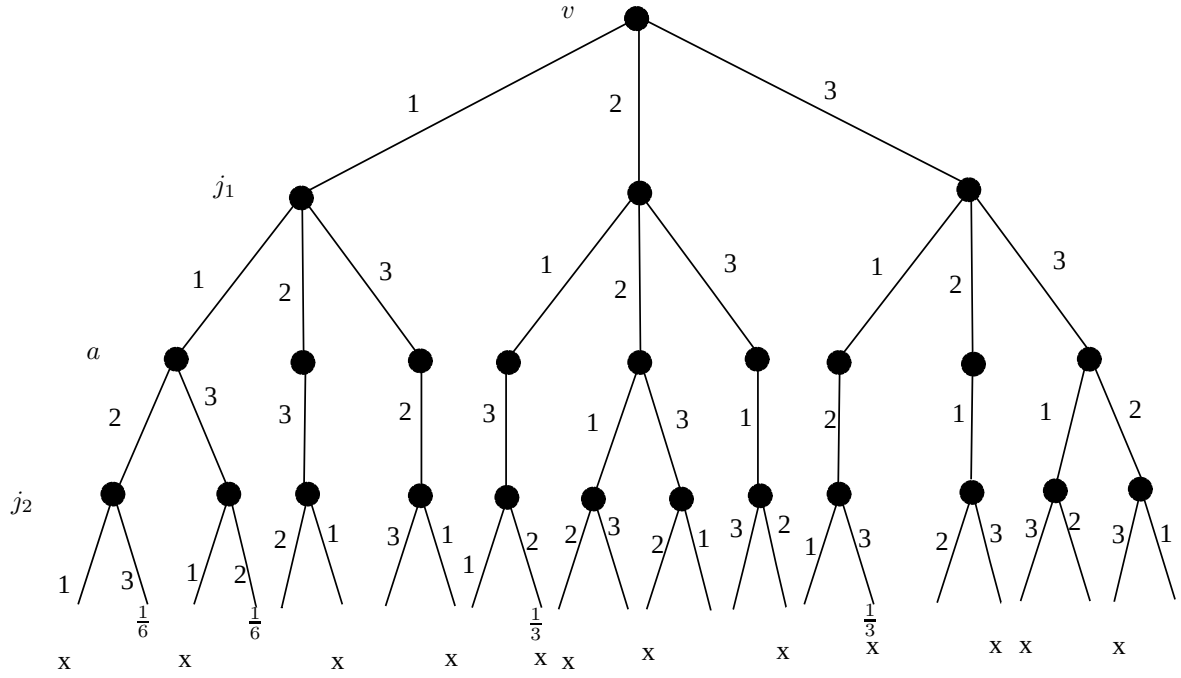


Figure 2.4: Probabilités associées aux feuilles de l'arbre de scénarios pour le jeu du Monty Hall, lorsque le joueur applique la stratégie versatile. Les scénarios gagnants sont ceux où $j_2 = v$, et sont indiqués par des croix. La probabilité de remporter le lot vaut $\frac{2}{3}$.

Comme explicité dans les légendes des trois figures, nous obtenons

- $P(j_2 = v) = \frac{1}{2}$ pour la stratégie totalement aléatoire,

- $P(j_2 = v) = \frac{1}{3}$ pour la stratégie têtue,
- $P(j_2 = v) = \frac{2}{3}$ pour la stratégie versatile.

Notons également (nous demandons au lecteur de s'en convaincre, en construisant les arbres de probabilités correspondants) que pour les stratégies têtue et versatile, la probabilité de gagner reste la même si le joueur décide d'abord de choisir la porte numéro deux, ou bien la porte numéro trois, et donc également s'il décide de tirer la valeur de j_1 au hasard, selon une loi de probabilité arbitraire.

Parmi toutes les stratégies de jeu que nous avons discutées, c'est donc bien la stratégie versatile qui est la meilleure, et la stratégie têtue qui est la moins bonne. Comme les stratégies que nous avons analysées couvrent essentiellement toutes les stratégies possibles pour le joueur qui respecte les règles du jeu, nous pouvons donc aussi affirmer que la stratégie globalement optimale est la stratégie versatile.

2.4.5 Discussion

Partant de l'énoncé du problème, nous avons tout d'abord clarifié quelques hypothèses qui étaient implicites dans le texte publié dans le journal, et c'est sans doute le caractère implicite de ces hypothèses qui a conduit aux nombreux débats qui ont suivi la publication. Ensuite nous avons appliqué une méthode systématique pour modéliser notre problème et puis le résoudre. Cette méthode comporte les quatre étapes suivantes:

1. Déterminer l'ensemble des résultats possibles de l'expérience aléatoire (c'est-à-dire Ω), ce que nous avons fait au moyen d'un arbre de scénarios pouvant couvrir un ensemble suffisamment large de stratégies de jeu.
2. Déterminer le (ou les) sous-ensembles de Ω dont on souhaite calculer la probabilité. Ici, il s'agissait de l'ensemble des scénarios conduisant au gain du joueur après les deux étapes de jeu.
3. Déterminer, en fonction du choix de la stratégie de jeu, les probabilités associées aux différents résultats possibles de l'expérience, correspondant aux feuilles de l'arbre. Ici nous avons fait explicitement cette opération pour les trois stratégies de jeu qu'il nous semblait intéressant d'évaluer.
4. Calculer la probabilité de l'événement d'intérêt en faisant la somme des probabilités des résultats qui appartiennent à cet événement (ici les scénarios gagnants).

Le respect de cette démarche a conduit à la bonne réponse, et ceci d'une façon convaincante qui ne prête pas le flanc à la critique.

Notes

1. Le terme consacré est en réalité "σ-algèbre de Boole" ou "tribu". Le terme "algèbre" est normalement réservé au cas où la troisième propriété est relaxée à l'union finie. Cependant, dans la suite nous utiliserons la plupart du temps simplement le terme "algèbre" étant entendu que dans le cas infini il faut comprendre σ-algèbre.

2. Dorénavant nous utiliserons la notation $A_1, A_2, \ldots$ pour désigner une suite dénombrable (éventuellement finie) d'ensembles.

3. C'est-à-dire la tribu borélienne sur $[0, 1]$.

4. Cependant, dans le cours de théorie de l'information on montrera qu'en moyenne l'incertitude concernant une expérience aléatoire diminue, lorsqu'on utilise de l'information complémentaire.

3 VARIABLES ALÉATOIRES

Dans ce chapitre nous introduisons la notion fondamentale de variable aléatoire, et les notions associées telles que loi de probabilité induite, fonction de répartition, densité, espérance mathématique, variance, etc., et aussi la notion de fonction d'une variable aléatoire ou d'un ensemble de variables aléatoires indépendantes.

3.1 NOTION DE VARIABLE ALÉATOIRE

3.1.1 Discussion intuitive

Intuitivement, la notion de variable aléatoire (nous utiliserons souvent l'abréviation *v.a.*) modélise de façon mathématique la notion *d'instrument de mesure* dans le contexte d'une expérience aléatoire modélisée par un espace de probabilité $(\Omega, \mathcal{E}, P)$. Avant de pouvoir définir la notion, il est nécessaire de définir de manière précise l'ensemble Ω des résultats possibles de cette expérience aléatoire (**intervention de la partie Ω du modèle probabiliste** $(\Omega, \mathcal{E}, P)$) : une fois que l'univers Ω est défini de manière précise, une variable aléatoire est essentiellement une *fonction* définie sur Ω dont on peut dans un certain contexte observer la valeur.

Par exemple, dans le cas d'une expérience de double lancer de dés, on peut [(1)] définir l'univers Ω comme étant l'ensemble des 36 couples $\{(1,1),(1,2),\ldots,(6,5),(6,6)\}$, puis on pourrait définir sur cet ensemble différentes fonctions, par exemple une fonction $\mathcal{X}$ qui calcule la somme des deux faces supérieures, dont l'ensemble de valeurs possibles est donc $\{2,3,4,5,6,7,8,9,10,11,12\}$, ou bien une fonction $\mathcal{Y}$ dont la valeur est le nombre lu sur la face supérieure du premier dé, dont l'ensemble de valeurs est donc $\{1,2,3,4,5,6\}$, ou bien encore une fonction $\mathcal{Z}$ dont la valeur est 1 si les deux dés sont tombés sur la même face et 0 sinon, dont l'ensemble de valeurs possibles est donc $\{0,1\}$.

Pour un observateur, le fait de connaître la valeur prise par une variable aléatoire lui fournit une information (en général partielle) sur le résultat de l'expérience aléatoire : par exemple, dans le cas de notre double lancer de dés, s'il sait seulement que la valeur de la v.a. $\mathcal{Z}$ vaut 1, il peut en déduire que le résultat doit se trouver dans le sous-ensemble $\{(1,1),(2,2),\ldots,(5,5),(6,6)\}$ de Ω, ou bien s'il sait seulement que la valeur de la variable $\mathcal{X}$ vaut 8, il peut en tirer que le résultat doit appartenir au sous-ensemble $\{(2,6),(3,5),(4,4),(5,3),(6,2)\}$ de Ω; dans les deux cas il s'agit d'une information partielle, puisque l'ensemble de réalisations compatibles avec l'observation n'est pas un singleton. De façon plus générale, le fait de savoir que la valeur d'une variable aléatoire se situe dans un *sous-ensemble* particulier de son ensemble de valeurs possibles, permet à l'observateur de situer le résultat de l'expérience aléatoire dans un sous-ensemble de Ω.

Dans le cadre d'une expérience aléatoire, les informations fournies par l'observation simultanée des valeurs de plusieurs variables aléatoires peuvent être redondantes ou bien complémentaires. En effet, dans notre exemple on peut déduire de la valeur de la variable $\mathcal{X}$ la valeur que doit prendre la variable $\mathcal{Z}$: l'observation de la valeur de $\mathcal{Z}$ ne peut donc en aucun cas apporter à notre observateur une information plus riche en ce qui concerne le résultat de l'expérience aléatoire que celle fournie par l'observation de la valeur de la variable $\mathcal{X}$; on peut dire que $\mathcal{Z}$ est redondante par rapport à $\mathcal{X}$ en termes d'information sur le résultat de l'expérience aléatoire. A contrario, s'il dispose simultanément de la valeur de $\mathcal{X}$ et de celle de $\mathcal{Y}$, il peut déterminer de façon exacte le résultat de l'expérience aléatoire, ce qu'il n'est pas en mesure de faire en observant la valeur d'une seule de ces deux variables; on peut dire que les deux variables se complètent en termes d'information sur Ω.

On voit que la notion de variable aléatoire permet de modéliser un processus de collecte d'information partielle quant à l'issue d'une expérience aléatoire, sous la forme d'une fonction définie sur Ω dont l'observation des valeurs permet de spécifier des sous-ensembles de Ω qui doivent contenir le résultat de l'expérience dans un certain contexte. On voit aussi qu'il est possible de raisonner de façon abstraite sur les relations entre variables aléatoires, au sujet de leur redondance, ou bien de leur complémentarité. Nous verrons de nombreux exemples, dans ce chapitre et dans les suivants, qui reposent sur l'étude des propriétés des variables aléatoires, et de leurs relations.

Pour que la notion de variable aléatoire soit exploitable dans le cadre du raisonnement probabiliste, il faut qu'elle soit compatible avec la structure d'événements $\mathcal{E}$ définie sur Ω, puisque la loi de probabilité est seulement définie pour les éléments de $\mathcal{E}$. Plus précisément, il faut que les sous-ensembles de Ω qui correspondent aux sous-ensembles de valeurs intéressants de la variable aléatoire soient des événements, afin qu'on puisse en déterminer la probabilité étant donnée la mesure P définie sur $\mathcal{E}$. Au chapitre précédent, nous avons fait la remarque que pour que le calcul des probabilités donne lieu à une théorie mathématique cohérente, il est parfois nécessaire de restreindre la structure de σ-algèbre à un sous-ensemble propre de l'ensemble 2^{Ω} des parties de Ω; si c'est le cas, il faut aussi restreindre la notion de variable aléatoire, en assurant qu'elle est *une fonction mesurable* (**intervention de la partie $\mathcal{E}$ du modèle probabiliste** $(\Omega, \mathcal{E}, P)$). Cette condition est une conséquence technique, mais nécessaire afin de pouvoir bâtir la suite du calcul de probabilités de façon cohérente.

3.1.2 Définition mathématique

Pour qu'une fonction définie sur Ω soit intéressante dans le contexte du raisonnement probabiliste, il est nécessaire que toutes les informations qu'il est souhaitable de pouvoir exprimer sur base de valeurs de la variable aléatoire correspondent à des sous-ensembles de Ω dont la probabilité est bien définie, c'est-à-dire à des ensembles mesurables de Ω.

Partant d'une fonction $\mathcal{X}(\cdot)$ définie sur Ω et à valeurs dans un ensemble $\Omega_{\mathcal{X}}$, on commence donc par choisir l'ensemble des informations qu'on souhaite exprimer à l'aide de cette fonction, en définissant une structure de σ-algèbre $\mathcal{E}_{\mathcal{X}}$ sur $\Omega_{\mathcal{X}}$. Cette structure définit les ensembles de valeurs de $\mathcal{X}$ qu'on souhaite pouvoir *mesurer*, c'est-à-dire ceux qu'on souhaite manipuler dans le raisonnement probabiliste faisant intervenir la variable $\mathcal{X}$.

Une fois cela fait, on doit vérifier que la structure $\mathcal{E}_{\mathcal{X}}$ définie sur $\Omega_{\mathcal{X}}$ est bien compatible avec la structure $\mathcal{E}$ définie sur Ω : on exige que chaque sous-ensemble défini sur Ω en spécifiant que la valeur de la variable aléatoire appartient à l'un des éléments de $\mathcal{E}_{\mathcal{X}}$ soit aussi un élément de $\mathcal{E}$. [(2)]

Cela se traduit par la définition générale suivante de la notion de variable aléatoire :

Notion générale de variable aléatoire

Soient un espace de probabilité $(\Omega, \mathcal{E}, P)$ et une fonction $\mathcal{X}(\cdot)$ définie sur Ω et à valeurs dans un ensemble $\Omega_{\mathcal{X}}$ muni d'une $\sigma-$algèbre $\mathcal{E}_{\mathcal{X}}$. La fonction $\mathcal{X}(\cdot)$ est une variable aléatoire si

$$\forall A' \in \mathcal{E}_{\mathcal{X}} \ : \{w \in \Omega | \mathcal{X}(w) \in A'\} \in \mathcal{E}. \tag{3.1}$$

On dit qu'une fonction $\mathcal{X}(\cdot)$ qui vérifie (3.1) est $(\mathcal{E}, \mathcal{E}_{\mathcal{X}})$-**mesurable** (ou simplement **mesurable**).

Dans la suite nous utiliserons $\mathcal{X}$ pour désigner la variable aléatoire et la notation $\mathcal{X}^{-1}(A')$ pour désigner l'ensemble $\{w \in \Omega | \mathcal{X}(w) \in A'\}$.

3.1.2.1 Mesure de probabilité $P_{\mathcal{X}}$ induite sur $(\Omega_{\mathcal{X}}, \mathcal{E}_{\mathcal{X}})$ par la v.a. $\mathcal{X}$

La propriété de *mesurabilité* assure que toute proposition logique relative à la valeur de la variable aléatoire et qui s'exprime à partir des sous-ensembles de valeurs faisant partie de $\mathcal{E}_{\mathcal{X}}$ se traduit aussi par une proposition logique en ce qui concerne la réalisation de l'expérience aléatoire et qui définit un événement de $\mathcal{E}$ pour lequel la mesure P nous permet d'évaluer sa probabilité.

Cela veut aussi dire qu'on peut associer des probabilités aux éléments de $\mathcal{E}_{\mathcal{X}}$ de la manière suivante :

Mesure de probabilité $P_{\mathcal{X}}$ induite sur $(\Omega_{\mathcal{X}}, \mathcal{E}_{\mathcal{X}})$ par la variable $\mathcal{X}$

$$\forall A' \in \mathcal{E}_{\mathcal{X}} : P_{\mathcal{X}}(A') \stackrel{\triangle}{=} P\left(\mathcal{X}^{-1}(A')\right). \tag{3.2}$$

La condition de "mesurabilité" de la fonction $\mathcal{X}(\cdot)$ nous assure que la mesure $P_{\mathcal{X}}$ ainsi induite sur $\Omega_{\mathcal{X}}$ est bien définie pour tout élément de $\mathcal{E}_{\mathcal{X}}$. Montrons qu'elle vérifie aussi les axiomes de Kolmogorov :

- **K1:** on a $\forall A' \in \mathcal{E}_{\mathcal{X}} : P_{\mathcal{X}}(A') = P(\mathcal{X}^{-1}(A')) \in [0,1]$.
- **K2:** on a $\Omega = \mathcal{X}^{-1}(\Omega_{\mathcal{X}})$ et par conséquent $P_{\mathcal{X}}(\Omega_{\mathcal{X}}) = P(\Omega) = 1$.
- **K3:** $\forall A'_1, A'_2 \ldots \in \mathcal{E}_{\mathcal{X}}$ incompatibles, les ensembles $A_i \stackrel{\triangle}{=} \mathcal{X}^{-1}(A'_i)$ sont des événements de $\mathcal{E}$ incompatibles et on a $\mathcal{X}^{-1}(\bigcup_i A'_i) = \bigcup_i A_i$; par conséquent $P_{\mathcal{X}}(\bigcup_i A'_i) = P(\mathcal{X}^{-1}(\bigcup_i A'_i)) = P(\bigcup_i A_i) = \sum_i P(A_i) = \sum_i P_{\mathcal{X}}(A'_i)$.

Le triplet $(\Omega_{\mathcal{X}}, \mathcal{E}_{\mathcal{X}}, P_{\mathcal{X}})$ forme donc un nouvel espace de probabilité auquel on peut appliquer l'ensemble des résultats du calcul de probabilités, en "oubliant" le mécanisme qui a donné lieu à sa naissance. Faire cet oubli est cependant une erreur souvent commise que nous décourageons vivement, car cela conduit à se déconnecter du modèle $(\Omega, \mathcal{E}, P)$ de base au sujet duquel on veut raisonner, ce qui est fort contre-productif lorsqu'il s'agit de manipuler plusieurs variables aléatoires qui apportent de l'information complémentaire au sujet de l'expérience aléatoire, comme c'est le cas dans toutes les applications pratiques.

3.1.2.2 σ-algèbre $\mathcal{E}_{\Omega/\mathcal{X}}$ induite sur Ω par la v.a. $\mathcal{X}$

Une variable aléatoire induit aussi une structure de σ-algébre sur Ω, de la façon suivante

σ-algébre $\mathcal{E}_{\Omega/\mathcal{X}}$ induite sur Ω par la variable $\mathcal{X}$

$$\mathcal{E}_{\Omega/\mathcal{X}} \stackrel{\triangle}{=} \left\{A \subset \Omega : \left(\exists A' \in \mathcal{E}_{\mathcal{X}} : A = \mathcal{X}^{-1}(A')\right)\right\}. \tag{3.3}$$

Etant donnée la condition de mesurabilité, on a $\mathcal{E}_{\Omega/\mathcal{X}} \subset \mathcal{E}$. Par ailleurs, on peut se convaincre que l'ensemble $\mathcal{E}_{\Omega/\mathcal{X}}$ de parties de Ω est bien une σ-algèbre. En effet

- **T1:** $\Omega \in \mathcal{E}_{\Omega/\mathcal{X}}$, puisque $\Omega_{\mathcal{X}} \in \mathcal{E}_{\mathcal{X}}$ et que $\Omega = \mathcal{X}^{-1}(\Omega_{\mathcal{X}})$.
- **T2:** Si $A \in \mathcal{E}_{\Omega/\mathcal{X}}$ alors il peut s'écrire sous la forme $A = \mathcal{X}^{-1}(A')$ avec $A' \in \mathcal{E}_{\mathcal{X}}$; dans ce cas on a $A^c = \mathcal{X}^{-1}(A'^c)$ et $A'^c \in \mathcal{E}_{\mathcal{X}}$; par conséquent $A^c \in \mathcal{E}_{\Omega/\mathcal{X}}$.
- **T3:** Soient $A_1, A_2 \ldots \in \mathcal{E}_{\Omega/\mathcal{X}}$ et $A = \bigcup_i A_i$, et soient les ensembles correspondants $A'_i \in \mathcal{E}_{\mathcal{X}}$, avec $A_i = \mathcal{X}^{-1}(A'_i)$. On a $A' = \bigcup_i A'_i \in \mathcal{E}_{\mathcal{X}}$ (car $\mathcal{E}_{\mathcal{X}}$ est une σ-algèbre), et comme $A = \mathcal{X}^{-1}(A')$ on a aussi que $A \in \mathcal{E}_{\Omega/\mathcal{X}}$.

Une variable aléatoire induit donc une σ-algèbre $\mathcal{E}_{\Omega/\mathcal{X}}$ comprise dans $\mathcal{E}$ (et en général moins fine que $\mathcal{E}$) : les sous-ensembles de $\mathcal{E}$ qui peuvent se décrire par des propositions logiques concernant la valeur prise par la variable aléatoire sont tous des événements de $\mathcal{E}$ mais la réciproque n'est pas vraie en général; certains éléments de $\mathcal{E}$ pourraient ne pas être exprimables au moyen d'une proposition logique ne portant que sur les valeurs d'une variable aléatoire particulière.

3.1.2.3 Discussion, interprétation, notations et exemples

Une variable aléatoire est une fonction définie sur un espace de probabilité qui est compatible avec les algèbres d'événements définies sur ses espaces d'origine et de destination. Elle induit une loi de probabilité sur l'espace de destination et une σ-algèbre sur l'espace de départ. Cela est illustré à la Figure 3.1.

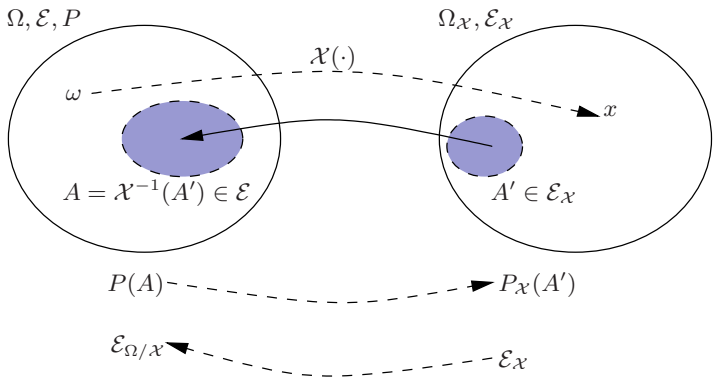


Figure 3.1: Notion de variable aléatoire. On a $x \in A' \Leftrightarrow \omega \in A = \mathcal{X}^{-1}(A')$, et $A' \in \mathcal{E}_{\mathcal{X}} \Rightarrow A \in \mathcal{E}$. On a $P_{\mathcal{X}}(A') = P(A)$. $\mathcal{E}_{\Omega/\mathcal{X}}$ est l'ensemble des parties de Ω pouvant s'exprimer comme $\mathcal{X}^{-1}(A')$ avec $A' \in \mathcal{E}_{\mathcal{X}}$.

Notons d'emblée que si les deux univers Ω et $\Omega_{\mathcal{X}}$ sont finis ou dénombrables et munis des algèbres maximales, alors toute fonction de Ω vers $\Omega_{\mathcal{X}}$ est mesurable et définit par conséquent une variable aléatoire. De fait, si nous avons pris la précaution de formuler les restrictions de *mesurabilité* ci-dessus, c'est que nous voulons appliquer le concept de variable aléatoire dans des situations où l'espace de départ est non-dénombrable (p.ex. $\Omega = \mathbb{R}^n$ muni de sa tribu borélienne, voir Appendice B).

La condition de mesurabilité (3.1) impose que l'algèbre $\mathcal{E}_{\Omega/\mathcal{X}}$ des événements induite par la fonction $\mathcal{X}(\cdot)$ à partir de $\mathcal{E}_{\mathcal{X}}$ sur Ω est inclue dans l'algèbre $\mathcal{E}$ pour laquelle la loi P est définie. Observer la valeur de la variable aléatoire situe le résultat de l'expérience aléatoire relativement aux événements de $\mathcal{E}_{\Omega/\mathcal{X}}$, ce qui donne une information en général moins précise que la localisation par rapport aux événements de $\mathcal{E}$ (généralement plus nombreux). La variable aléatoire opère donc sur l'espace de probabilité en condensant l'information. Il s'agit bien là du sens profond de la notion de variable aléatoire : l'observation d'une variable aléatoire fournit une information généralement partielle sur la réalisation des issues possibles d'une expérience aléatoire. La condition de mesurabilité assure que cette information puisse être exploitée pour le raisonnement probabiliste.

Notations. Nous utiliserons des lettres "majuscules calligraphiées" ($\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$) pour désigner des variables aléatoires, des lettres minuscules ($x, y, z, \ldots$) pour désigner une valeur particulière d'une variable aléatoire, et des lettres majuscules ($X, Y, Z, \ldots$) pour désigner des sous-ensembles particuliers de $\Omega_{\mathcal{X}}$, $\Omega_{\mathcal{Y}}$, $\Omega_{\mathcal{Z}}, \ldots$.

Exemple 1(a) : simple lancer de dé. Dans le lancer de dé on peut choisir que Ω est l'ensemble des valeurs numériques $\{1, 2, 3, 4, 5, 6\}$ pouvant être lues sur la face supérieure du dé une fois au repos. On peut par exemple définir une variable aléatoire $\mathcal{X}$ telle que $\mathcal{X}(1) = \mathcal{X}(3) = \mathcal{X}(5) =$ "vrai" et $\mathcal{X}(2) = \mathcal{X}(4) = \mathcal{X}(6) =$ "faux". C'est une variable aléatoire qui mesure le fait que le résultat du lancer de dé est pair. Observer sa valeur, permet (seulement) de situer le résultat de l'expérience soit dans l'ensemble $\{1, 3, 5\}$ soit dans l'ensemble $\{2, 4, 6\}$. Si le dé est équilibré, alors nous aurons que $P_{\mathcal{X}}(\{\text{"vrai"}\}) = P_{\mathcal{X}}(\{\text{"faux"}\}) = \frac{1}{2}$.

Suggestion : se convaincre que $P_{\mathcal{X}}(\{\text{"vrai"}\}) = \frac{1}{2}$ est bien vrai.

Exemple 1(b) : double lancer de dé. Considérer ensuite le problème du lancer simultané d'un dé rouge et d'un dé noir (chacun à six faces). Supposer que les 36 combinaisons de faces supérieures sont équiprobables. Décrire l'ensemble Ω des résultats possibles, et définir une variable aléatoire $\mathcal{X}$ dont la valeur est la somme des deux nombres lus sur la face supérieure des deux dés lorsqu'ils sont au repos. Décrire l'ensemble $\Omega_{\mathcal{X}}$ des valeurs possibles de cette variable et la loi de probabilité $P_{\mathcal{X}}$ qu'elle induit sur cet ensemble. Décrire l'algèbre induite $\mathcal{E}_{\Omega/\mathcal{X}}$ et donner quelques exemples d'éléments de $\mathcal{E}$ qui ne font par partie de cette algèbre induite.

Exemple 2: double pile ou face. Nous considérons une pièce d'un Euro et une pièce de deux Euros, qui sont lancées simultanément. L'ensemble Ω est défini comme l'ensemble des couples de résultats (pile ou face), i.e. $\Omega = \{PP, PF, FP, FF\}$, et on peut y définir par exemple les trois variables aléatoires suivantes

- $\mathcal{X}_1$ telle que $\mathcal{X}_1(PP) = \mathcal{X}_1(PF) = P$ et $\mathcal{X}_1(FP) = \mathcal{X}_1(FF) = F$,
- $\mathcal{X}_2$ telle que $\mathcal{X}_2(PP) = \mathcal{X}_2(FP) = P$ et $\mathcal{X}_2(PF) = \mathcal{X}_2(FF) = F$,
- et $\mathcal{Z}$ telle que $\mathcal{Z}(PP) = \mathcal{Z}(FF) =$ "vrai" et $\mathcal{Z}(PF) = \mathcal{Z}(FP) =$ "faux".

Suggestion : Dans l'hypothèse où les quatre résultats élémentaires de Ω sont équiprobables, calculer les lois de probabilité induites par ces trois variables aléatoires. Refaire le calcul en supposant que $P(\{PP\}) = 0.04, P(\{PF\}) = 0.36, P(\{FP\}) = 0.06, P(\{FF\}) = 0.54$.

3.1.2.4 • Petite digression sur l'étude simultanée de plusieurs variables aléatoires

Sur un espace de probabilité $(\Omega, \mathcal{E}, P)$ on peut évidemment définir un grand nombre (une infinité, en réalité) de variables aléatoires. Par exemple, dans le cas du double lancer de dés, on peut définir la variable $\mathcal{X}$ qui désigne la somme des deux faces, la variable $\mathcal{Y}$ qui désigne la face sur laquelle est tombé le premier dé, la variable $\mathcal{Z}$ qui désigne la face sur laquelle est tombé le second dé, la variable $\mathcal{W}$ qui indique si les deux dés sont tombés sur la même face, etc. etc.

On peut étudier ces variables isolément les unes des autres, et c'est essentiellement le but du présent chapitre de développer les outils nécessaires pour ce genre d'étude *isolée*.

Cependant, la vraie richesse du calcul de probabilités se situe dans les outils qu'il fournit pour étudier de façon *conjointe* des ensembles de variables aléatoires.

Par exemple, dans le domaine de la sociologie, on peut étudier le comportement des personnes ou groupes de personnes, au moyen de différents indicateurs (des variables aléatoires, telles que statut civil, âge, niveau de formation, occupation, revenu, qualité de vie, intentions de vote, etc. etc.), et les études les plus intéressantes sont celles qui essayent de comprendre les relations entre ces différents indicateurs.

De même, dans un contexte technique ou technico-économique, on peut étudier les performances d'une installation industrielle, en termes de coûts économiques, revenus engendrés, emplois assurés, qualité des produits, empreinte écologique etc. etc. Cependant l'étude isolée de chacun de ces facteurs n'apporte généralement pas beaucoup de nouvelles connaissances, alors que leur étude conjointe peut révéler des choses fort intéressantes lorsqu'il s'agit de prendre des décisions d'investissement dans une nouvelle usine.

C'est le but du chapitre 4 de développer les outils pour l'étude conjointe d'ensembles de variables aléatoires. Cependant, à ce stade de la présentation, il nous paraît utile de déjà anticiper sur cette étude en mettant en évidence ce qu'elle a de particulier par rapport à l'étude isolée des variables aléatoires, car cela est important pour motiver une partie des idées introduites dans le présent chapitre, et en particulier la définition même de la notion de variable aléatoire.

Prenons un exemple médical où nous voulons étudier l'impact de certaines habitudes (disons le tabagisme) sur l'espérance de vie d'une personne issue d'une certaine population. Dans un premier temps, on peut modéliser ce problème en définissant l'ensemble Ω des personnes (la population), et deux variables aléatoires, à savoir $\mathcal{X}$ le nombre total de cigarettes fumées par une personne et $\mathcal{Y}$ son âge. Ce qui nous intéresse, est de savoir si le fait d'avoir fumé beaucoup de cigarettes réduit l'espérance de vie. Si c'est le cas, on doit s'attendre à ce que les personnes qui ont peu ou pas fumé forment un sous-ensemble de Ω dans lequel on trouve une plus grande proportion de personnes âgées que parmi les fumeurs. Pour répondre à ce genre de questions, il faut donc pouvoir calculer la probabilité d'événements (nombre de personnes d'un certain profil) dont la description fait intervenir des affirmations logiques qui portent à la fois sur les valeurs de la variable $\mathcal{X}$ et de la variable $\mathcal{Y}$, c'est-à-dire des grandeurs qui ne peuvent pas être le fruit de l'étude isolée des deux variables aléatoires. Evidemment, si nous voulions réellement faire un étude sérieuse sur le sujet, nous serions fort probablement amenés à introduire d'autres informations intéressantes, c'est-à-dire d'autres variables aléatoires, telles que risque génétique, autres habitudes (alimentaires, stress professionnel, niveau d'activités physiques, etc. etc.) afin d'examiner différentes autres hypothèses. Dans certains cas, certaines variables aléatoires sont des fonctions d'autres variables aléatoires, par exemple on pourrait imaginer que le stress a pour conséquence de pousser au tabagisme, et que le stress réduit

l'espérance de vie; une étude seule du tabagisme pourrait conduire à la conclusion que les habitudes de tabagisme réduisent l'espérance de vie, alors qu'une étude conjointe du stress et du tabagisme pourrait montrer que c'est en fait le stress qui réduit l'espérance de vie.

Pour caricaturer l'exemple précédent dans un contexte moins dramatique et plus directement accessible à l'expérience, si on étudie le nombre de passants munis d'un parapluie dans les rues de Liège, et le débit de la Meuse, on pourrait conclure que lorsque les passants se munissent de parapluies alors le débit de la Meuse va fort probablement monter dans les heures qui suivent. Dans cet exemple, le bon sens nous dit néanmoins que si on demande la veille aux promeneurs de ne pas se munir de parapluies, on n'évitera pas les inondations du côté de Visé s'il pleut beaucoup. Si par contre l'analyse est faite par un ordinateur, que nous supposons privé de bon sens, il faudrait lui suggérer de prendre aussi en compte dans son raisonnement le fait qu'il pleut à un moment donné. En effet, s'il suit ce conseil, il pourra soupçonner la bonne explication en calculant, qu'à niveau de pluie fixé, le fait qu'il y a plus ou moins de porteurs de parapluie ne permet pas de dire grand chose d'intéressant en ce qui concerne la dérivée du débit de la Meuse et le risque d'inondations futures à Visé. En jargon probabiliste, nous disons que la variable "inondation" est **conditionnellement indépendante** de la variable "parapluie" étant donnée la connaissance de la valeur de la variable "pluie", bien que la variable "parapluie" ne soit pas indépendante de la variable "inondation" (ni d'ailleurs de la variable "pluie", mais ceci est une autre histoire).

Les questions évoquées ci-dessus se ramènent pour l'essentiel à comparer des probabilités associées aux événements de Ω dont la description (c'est-à-dire la spécification) fait intervenir les valeurs de plusieurs variables aléatoires. Par exemple, pour le double lancer de dés on peut se demander quelle est la probabilité d'observer 8 comme total, sachant que le premier dé tombe sur la même face que le second, ou bien sachant que le premier dé tombe sur une face impaire etc. etc. On peut aussi se demander si le fait de révéler la valeur du premier dé apporte de l'information sur celle du second dé, ou bien si cela conduit à changer d'avis en ce qui concerne la valeur la plus probable de la somme des deux faces (par rapport à une situation où nous sommes en l'absence de cette information). Répondre à ce genre de questions à partir de la spécification d'un espace de probabilité et de plusieurs variables aléatoires définies sur cet espace s'appelle faire de l'**inférence probabiliste**.

Comme nous le verrons au chapitre 4, dans le domaine de l'inférence probabiliste, les notions centrales sont l'**indépendance entre variables aléatoires**, le **conditionnement** par rapport aux valeurs de variables aléatoires, et la notion d'**indépendance conditionnelle**. Dans le présent chapitre, nous allons seulement introduire (une première fois) la notion d'indépendance entre variables aléatoires, car elle est importante pour comprendre les propriétés de base des notions que nous introduisons pour l'étude isolée de variables aléatoires. De fait, si deux variables aléatoires sont **indépendantes**, leur étude isolée successive révèle essentiellement les mêmes connaissances que leur étude simultanée : on peut donc se contenter de leur caractérisation *indépendante*.

3.2 TYPES DE V.A. ET CARACTÉRISATION DE LEUR MESURE INDUITE

Dans cette section nous considérons les variables aléatoires discrètes à valeurs quelconques et celles à valeurs réelles, et nous introduisons leurs σ-algèbres naturelles et des fonctions définies sur Ω (densités et fonction de répartition) qui caractérisent entièrement leur mesure de probabilité .

3.2.1 Variables aléatoires discrètes à valeurs quelconques

Une variable aléatoire $\mathcal{X}$ est discrète si l'ensemble de ses valeurs possibles $\Omega_{\mathcal{X}} = \{x_1, x_2, \ldots\}$ est soit fini soit dénombrable. (3) Nous prenons comme σ-algèbre naturelle de ce type de variable l'ensemble $2^{\Omega_{\mathcal{X}}}$ de *toutes* les parties de $\Omega_{\mathcal{X}}$. Nous définissons la densité de probabilité sur $\Omega_{\mathcal{X}}$ par

$$p_{\mathcal{X}}(x) \triangleq P_{\mathcal{X}}(\{x\}). \tag{3.4}$$

Tout sous-ensemble $X \in 2^{\Omega_{\mathcal{X}}}$ étant dénombrable, on peut donc calculer sa mesure de probabilité par

$$P_{\mathcal{X}}(X) = \sum_{x \in X} p_{\mathcal{X}}(x), \tag{3.5}$$

à partir de la seule connaissance de la densité de probabilité $p_{\mathcal{X}}$. Dans la suite de ce cours nous utiliserons indifféremment la notation $P_{\mathcal{X}}$ (P majuscule) et $p_{\mathcal{X}}$ (p minuscule) pour désigner la "densité" d'une variable aléatoire discrète. (4)

3.2.2 Variables aléatoires à valeurs réelles

Une v.a. $\mathcal{X}$ est dite à valeurs réelles (ou simplement réelle) si $\Omega_{\mathcal{X}} \subset \mathbb{R}$ muni de la tribu borélienne $\mathcal{B}_{\mathbb{R}}$.

(*Suggestion : lire l'appendice B pour la définition de la notion de tribu borélienne.*).

Exemple 3. Consommation électrique d'un bâtiment. Considérons la puissance électrique instantanée (en kW) consommée par un bâtiment tel que l'Institut Montefiore. Cette puissance varie de façon relativement aléatoire d'un moment à l'autre, même si physiquement elle est contrainte par le dimensionnement de l'installation électrique et la puissance totale p_{m} de tous les équipements installés dans l'immeuble. Supposons que nous disposions d'un enregistrement complet de cette puissance au cours de l'année précédente, et formulons une expérience aléatoire qui consiste à choisir au hasard un moment (un instant $t \in [t_{\min}, t_{\max}] \subset \mathbb{R}$, cet intervalle couvrant toute la période considérée), et à observer la consommation instantanée à cet instant t. Cette expérience définit une variable aléatoire $\mathcal{X}$ à valeurs réelles, avec $\Omega_{\mathcal{X}} = [0, p_{\mathrm{m}}] \subset \mathbb{R}$.

Exemple 4. Fonction caractéristique d'un événement. Soit un espace de probabilité $(\Omega, \mathcal{E}, P)$ et soit $A \in \mathcal{E}$ un événement. On définit la fonction caractéristique de l'ensemble A, notée $1_A(\cdot)$ par

$$\begin{cases} 1_A(\omega) = 1, \forall \omega \in A, \\ 1_A(\omega) = 0, \forall \omega \in A^c. \end{cases} \tag{3.6}$$

Cette fonction à valeurs réelles est mesurable et définit par conséquent une variable aléatoire réelle (et discrète) sur Ω. La σ-algèbre induite par cette variable sur Ω est $\{\Omega = A \cup A^c, \emptyset = A \cap A^c, A, A^c\}$.

3.2.2.1 Définition de la notion de fonction de répartition

Par définition, la fonction de répartition $F_{\mathcal{X}}$ d'une v.a. réelle $\mathcal{X}$ est la fonction de $\mathbb{R}$ dans $[0, 1]$ définie par

$$F_{\mathcal{X}}(x) = P(\mathcal{X}(\omega) \in]-\infty, x[). \tag{3.7}$$

Elle peut donc en principe avoir des discontinuités (à droite), si certaines des valeurs sont de probabilité non-nulle. Elle est monotone croissante, et $F_{\mathcal{X}}(-\infty) = 0$ et $F_{\mathcal{X}}(+\infty) = 1$.

Cette fonction *caractérise* la loi de probabilité induite par la variable aléatoire et permet de calculer la probabilité de tout intervalle de $\mathbb{R}$ par

$$P(a \leq \mathcal{X}(\omega) < b) = P_{\mathcal{X}}([a, b[) = F_{\mathcal{X}}(b) - F_{\mathcal{X}}(a). \tag{3.8}$$

Elle permet aussi de calculer la probabilité de tout ensemble $B \in \mathcal{B}_{\mathbb{R}}$, en exprimant cet ensemble à partir d'une collection dénombrable de semi-intervalles ouverts à droite (voir appendice B).

Exemple 3 (suite). Consommation électrique d'un bâtiment. La valeur de la fonction de répartition peut s'interpréter dans cet exemple. En effet, $F_{\mathcal{X}}(x)$ est égale à la proportion du temps où la consommation électrique du bâtiment est inférieure à x. On a $\forall x \leq 0 : F_{\mathcal{X}}(x) = 0$ et $\forall x > p_{\mathrm{m}} : F_{\mathcal{X}}(x) = 1$.

Remarque. *On peut tout aussi bien définir la fonction de répartition de façon à ce qu'elle soit continue à droite, on a alors*

$$F_{\mathcal{X}}(x) = P(\mathcal{X}(\omega) \in]-\infty, x]). \tag{3.9}$$

C'est la convention qu'on trouve généralement dans la littérature anglo-saxonne.

3.2.2.2 Variable aléatoire (réelle) discrète et sa fonction de répartition

Une variable aléatoire est discrète si son ensemble de valeurs $\Omega_{\mathcal{X}}$ est dénombrable. Evidemment si Ω est lui-même un ensemble dénombrable, toute variable aléatoire réelle sera nécessairement aussi discrète.

Lorsque une v.a. réelle est discrète, il n'y a donc qu'un nombre dénombrable de points de $\mathbb{R}$ de probabilité non-nulle. La fonction de répartition prend alors l'allure indiquée à la figure 3.2.

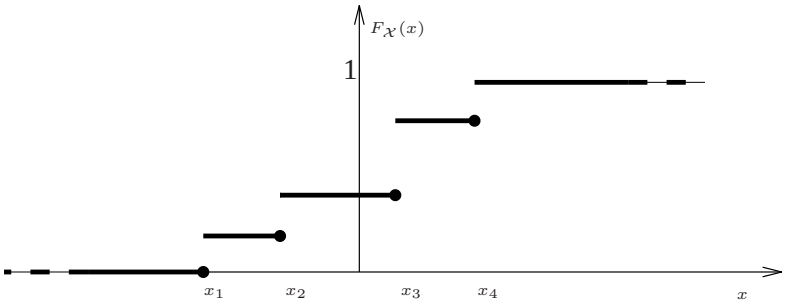


Figure 3.2: Exemple de fonction de répartition d'une v.a. réelle discrète. Les discontinuités correspondent aux valeurs possibles x_i de la variable aléatoire (ici au nombre de 4).

Exemple 3 (suite). Consommation électrique d'un bâtiment. Si nous supposons que les appareils électriques du bâtiment consomment tous une puissance fixe, une fois qu'ils sont branchés, alors la variable consommation totale instantanée sera une variable discrète, les paliers étant définis par les appareils qui sont branchés à un moment particulier. Par exemple, si nous avons trois appareils de puissance respective $p_1 = 100$, $p_2 = 150$ et $p_3 = 250$ les valeurs possibles pour $\mathcal{X}$ seraient respectivement de $0, 100, 150, 250, 350, 400$ et 500.

Exemple 4 (suite). Fonctions caractéristiques. La fonction de répartition d'une v.a. définie par la fonction caractéristique d'un événement A est une fonction en escalier, prenant au plus trois valeurs différentes, à savoir 0, $1 - P(A)$ et 1. Nous suggérons de la dessiner, et puis de dessiner la fonction de répartition d'un autre événement B indépendant de A, et puis de dessiner la fonction de répartition de la variable aléatoire qui est la somme de ces deux fonctions caractéristiques.

3.2.2.3 Variable aléatoire (réelle) continue et sa densité de probabilité

Une variable aléatoire réelle $\mathcal{X}$ est dite continue [(5)] si elle admet une densité, c'est-à-dire s'il existe une fonction $f_{\mathcal{X}}(\cdot)$ définie sur $\mathbb{R}$ telle que $\forall a \leq b$ on a

$$P(\mathcal{X} \in]a, b[) = P(\mathcal{X} \in [a, b]) = P(\mathcal{X} \in [a, b[) = P(\mathcal{X} \in]a, b]) = F(b) - F(a) = \int_a^b f_{\mathcal{X}}(x)\,\mathrm{d}x. \qquad (3.10)$$

Dans ce cas, $F_{\mathcal{X}}(\cdot)$ est dérivable (et donc continue) et admet $f_{\mathcal{X}}(\cdot)$ comme dérivée. La densité $f_{\mathcal{X}}(\cdot)$ est alors positive et d'intégrale sur $\mathbb{R}$ égale à 1. La figure 3.3 représente graphiquement la fonction de répartition de ce type de variable aléatoire.

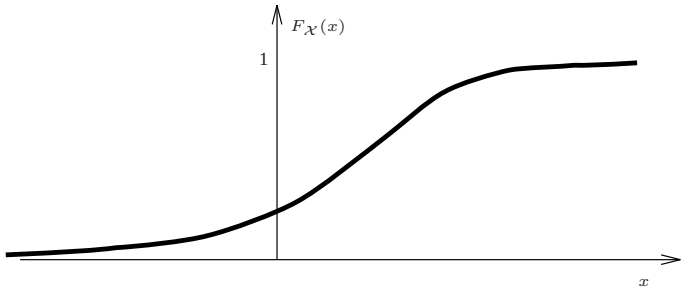


Figure 3.3: Exemple de fonction de répartition d'une v.a. réelle continue

Dans la suite nous utiliserons la notation $\mathcal{X} \sim F_{\mathcal{X}}(x)$ (respectivement $\mathcal{X} \sim f_{\mathcal{X}}(x)$) pour indiquer qu'une v.a. possède une certaine fonction de répartition (respectivement possède une certaine densité).

Exemple 3 (suite). Consommation électrique d'un bâtiment. En pratique, la puissance consommée par un appareil électrique varie continûment : par exemple celle d'un moteur (disons d'un ascenseur) varie en fonction de sa charge, de son accélération; celle d'une ampoule électrique ou d'une résistance de

chauffage varie aussi en fonction de la tension du réseau électrique (et éventuellement de la fréquence) qui varient en pratique d'un moment à l'autre; enfin la puissance électrique consommée par un ordinateur portable connecté au réseau varie en fonction de la nature des tâches qu'il est en train d'effectuer, et souvent en fonction des conditions d'éclairage de son environnement. Aussi le nombre d'appareils électriques dans un bâtiment de la taille de l'Institut Montefiore est en réalité trés grand (quelques dizaines de milliers d'appareils). Compte tenu de ces remarques, il est souvent plus réaliste et aussi plus pratique de considérer que la puissance consommée par le bâtiment varie continument au cours du temps, et donc de modéliser $\mathcal{X}$ sous la forme d'une variable aléatoire continue. $F_{\mathcal{X}}(x)$ peut alors en principe être déterminée pour tout x en vérifiant la proportion du temps où $\mathcal{X} < x$. Nous verrons plus loin dans ce cours qu'en pratique une bonne approximation consiste souvent à supposer que $\mathcal{X}$ est une variable aléatoire Gaussienne, dont la distribution peut-être caractérisée de façon très simple.

3.2.2.4 Cas général de la variable aléatoire réelle : fonction de répartition et densité

Une variable aléatoire réelle peut être ni discrète, ni continue. Une fonction de répartition ne peut cependant avoir au plus qu'un ensemble dénombrable de points de discontinuité. Dans le cas général on peut séparer la v.a. réelle en la somme d'une composante continue et d'une composante discrète [(6)]. Notons qu'on peut dans le cas général aussi faire appel à la théorie des distributions pour définir cette fois la densité comme une *distribution* qui s'écrit sous la forme d'une combinaison linéaire (convexe) d'une fonction et d'une série d'impulsions de Dirac. Cette situation est schématisée graphiquement à la figure 3.4 (voir appendices, pour plus de détails).

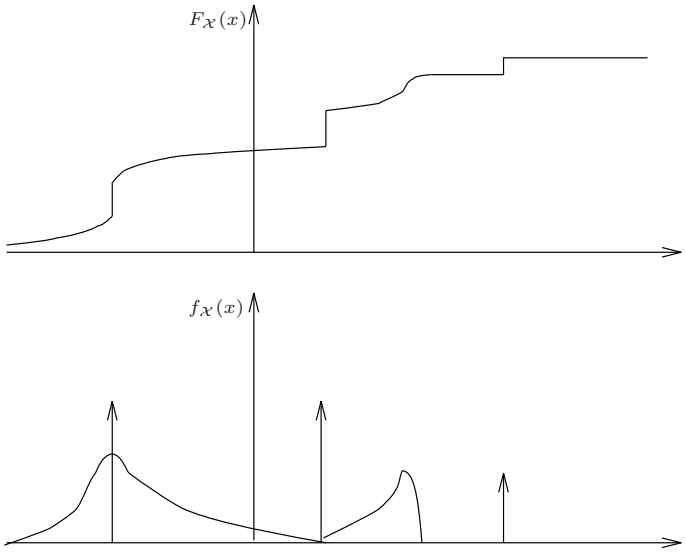


Figure 3.4: Fonction de répartition (graphique du dessus) et distribution de probabilité (graphique du dessous). La distribution de probabilité est composée d'une somme d'une densité et d'une série d'impulsions de Dirac. Une impulsion de Dirac est un objet mathématique qui représente la limite d'une suite de fonctions qui sont d'intégrale constante et dont le support tend vers un singleton. Cette notion est vue plus en détails dans d'autres cours.

Exemple 3 (suite). Consommation électrique d'un bâtiment. Pour rendre notre modèle encore plus réaliste, il faut tenir compte des interruptions de fourniture d'électricité et des pannes internes qui arrivent de temps en temps et conduisent à une coupure momentanée de l'alimentation électrique du bâtiment, coupure qui peut avoir une durée variable. En termes de consommation totale, cela se traduit par une proportion du temps, non nulle, où la consommation totale est nulle, i.e. par une probabilité $P_{\mathcal{X}}(0)$ strictement supérieure à zéro. La variable $\mathcal{X}$ est dans ce cas ni discrète ni continue. Sa densité sera composée d'une somme de deux termes, le premier étant une impulsion de Dirac à l'origine de hauteur $P_{\mathcal{X}}(0)$ et et le second étant une densité qui représente la loi associée aux fonctionnements normaux (d'intégrale égale à $1 - P_{\mathcal{X}}(0)$).

3.2.3 ○ Variables aléatoires complexes

Tout ce qui vient d'être dit concernant les variables aléatoires réelles peut, à peu de choses près, être appliqué aux variables aléatoires à valeurs complexes. D'ailleurs, on peut séparer toute fonction complexe en ses parties réelle et imaginaire qui sont des fonctions réelles. Une variable aléatoire complexe est donc de ce point de vue équivalente à un couple de variables aléatoires réelles. Nous ne faisons pas dans ces notes de traitement particulier des variables aléatoires à valeurs complexes.

Exemple 3 (suite). Consommation électrique d'un bâtiment. On obtient un exemple de variable aléatoire complexe, si on considère au lieu de la puissance active seulement, la variation de la puissance complexe (dont la composante imaginaire est la puissance réactive).

3.3 FONCTION D'UNE VARIABLE ALÉATOIRE

Soit une variable aléatoire $\mathcal{X}(\cdot)$ définie sur Ω et à valeurs dans $\Omega_{\mathcal{X}}$, et une certaine fonction $\phi(\cdot)$ définie sur $\Omega_{\mathcal{X}}$ et à valeurs dans Ω_{ϕ}. Alors, si $\phi(\cdot)$ a le statut de variable aléatoire sur $\Omega_{\mathcal{X}}$ (compatibilité de $\mathcal{E}_{\mathcal{X}}$ et $\mathcal{E}_{\phi}$), la fonction composée $\phi \circ \mathcal{X}(\cdot) = \phi(\mathcal{X}(\cdot))$ définit également une v.a. sur Ω. Ceci est illustré à la figure 3.5. Notons que l'observation des valeurs de la fonction composée apporte en général moins d'information sur la réalisation de l'expérience dans Ω que l'observation des valeurs de la variable $\mathcal{X}$.

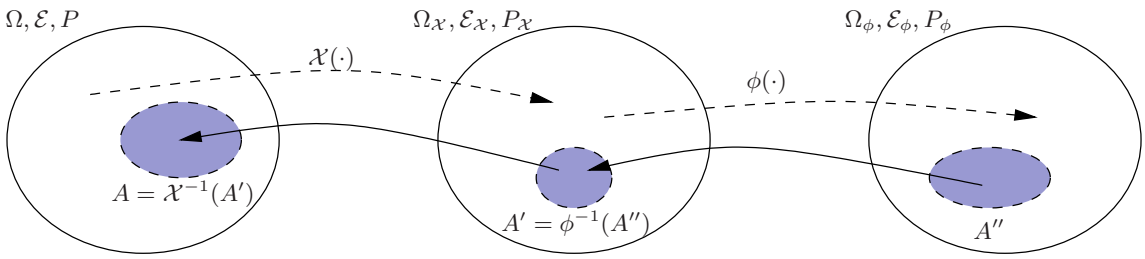


Figure 3.5: Illustration de la notion de fonction d'une variable aléatoire

Nous discuterons ci-dessous le cas particulièrement fréquent en pratique où les deux fonctions sont des v.a. à valeurs réelles.

3.3.1 Fonction de répartition et densité d'une fonction à valeurs réelles d'une v.a. réelle

On suppose que la variable aléatoire $\mathcal{X}$ est continue, et nous désignons par $F_{\mathcal{X}}$ sa fonction de répartition et par $f_{\mathcal{X}}$ sa densité. Nous supposons que la fonction ϕ est dérivable et notons par ϕ' sa dérivée. Nous allons calculer la densité $f_{\mathcal{Y}}$ et la fonction de répartition $F_{\mathcal{Y}}$ de la variable $\mathcal{Y} = \phi(\mathcal{X})$.

3.3.1.1 Cas où la fonction ϕ est bijective

Si la fonction ϕ est bijective elle est soit monotone croissante, soit monotone décroissante.

Si ϕ est monotone croissante, nous avons $\mathcal{X} < x \Leftrightarrow \mathcal{Y} < \phi(x)$. Par conséquent, $F_{\mathcal{Y}}(y) = F_{\mathcal{X}}(\phi^{-1}(y))$, et donc $f_{\mathcal{Y}}(y) = \frac{f_{\mathcal{X}}(\phi^{-1}(y))}{\phi'(\phi^{-1}(y))}$.

Si ϕ est monotone décroissante, nous avons $\mathcal{X} < x \Leftrightarrow \mathcal{Y} > \phi(x)$. Par conséquent, $F_{\mathcal{Y}}(y) = 1 - F_{\mathcal{X}}(\phi^{-1}(y))$, et donc $f_{\mathcal{Y}}(y) = -\frac{f_{\mathcal{X}}(\phi^{-1}(y))}{\phi'(\phi^{-1}(y))}$.

Puisque lorsque ϕ est monotone croissante on a $\phi' > 0$, et que lorsque ϕ est monotone décroissante on a $\phi' < 0$, les deux formules se résument par

Densité d'une fonction bijective d'une v.a. réelle continue

$$\mathcal{Y} = \phi(\mathcal{X}) \text{ avec } \phi \text{ bijective et dérivable} \Rightarrow f_{\mathcal{Y}}(y) = \frac{f_{\mathcal{X}}(\phi^{-1}(y))}{|\phi'(\phi^{-1}(y))|}. \tag{3.11}$$

Exemples importants. Appliquons l'équation (3.11) au cas où $\mathcal{Y} = F_{\mathcal{X}}(x)$ ($\phi = F_{\mathcal{X}}$, et $\phi' = f_{\mathcal{X}}$). On obtient que $f_{\mathcal{Y}}(y) = 1$, et on en déduit que la variable $\mathcal{Y}$ possède une densité constante sur l'intervalle $[0, 1]$; il s'agit d'une variable uniforme.

Réciproquement, si $\mathcal{X}$ est une variable de densité uniforme sur $[0, 1]$, la fonction $\mathcal{Y} = F_{\mathcal{Y}}^{-1}(\mathcal{X})$ possède la fonction de répartition $F_{\mathcal{Y}}$ et la densité $f_{\mathcal{Y}} = F'_{\mathcal{Y}}$ si celle-ci existe.

La dernière formule est utile en pratique, pour générer dans les simulations informatiques des variables aléatoires de distribution donnée, à partir d'un générateur de nombres aléatoires uniformes sur l'intervalle $[0, 1]$.

3.3.1.2 Cas où la fonction ϕ est quelconque

Le principe consiste toujours à identifier la valeur de la fonction de répartition $F_{\mathcal{Y}}(y)$ en recherchant la condition sur $\mathcal{X}$ qui correspond à l'événement $\mathcal{Y} < y = \phi(x)$.

Par exemple, si $\mathcal{Y} = \mathcal{X}^2$, on a $\mathcal{Y} < y \Leftrightarrow -\sqrt{y} < \mathcal{X} < \sqrt{y}$, et par conséquent la fonction de répartition de $\mathcal{Y}$ est obtenue par $F_{\mathcal{Y}}(y) = F_{\mathcal{X}}(\sqrt{y}) - F_{\mathcal{X}}(-\sqrt{y})$, et sa densité ensuite par $f_{\mathcal{Y}}(y) = \frac{1}{2\sqrt{y}}(f_{\mathcal{X}}(\sqrt{y}) - f_{\mathcal{X}}(-\sqrt{y}))$.

3.4 INDÉPENDANCE DE DEUX VARIABLES ALÉATOIRES

Dans cette section nous anticipons sur des notions qui seront étudiées en détails au chapitre 4, en définissant la notion d'indépendance de deux variables aléatoires. Au chapitre 4 nous étendrons notamment cette notion à la notion d'indépendance *mutuelle* de plusieurs variables aléatoires et a la notion d'indépendance *conditionnelle* entre ensembles de variables aléatoires.

3.4.1 Définition générale

Deux variables aléatoires $\mathcal{X}$ et $\mathcal{Y}$ définies sur un même espace de probabilité $(\Omega, \mathcal{E}, P)$ sont indépendantes si et seulement si, $\forall A' \in \mathcal{E}_{\mathcal{X}}, \forall B' \in \mathcal{E}_{\mathcal{Y}}$ on a

$$P(\mathcal{X}^{-1}(A') \cap \mathcal{Y}^{-1}(B')) = P(\mathcal{X}^{-1}(A'))P(\mathcal{Y}^{-1}(B')). \tag{3.12}$$

Nous dirons que la loi induite par la v.a. $\mathcal{Z}(\cdot) = (\mathcal{X}(\cdot), \mathcal{Y}(\cdot))$ sur l'espace produit $\Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$ est factorisable *en un produit des lois marginales de $\mathcal{X}$ et de $\mathcal{Y}$* si et seulement si les variables $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes. On a en effet dans ce cas

$$\forall A' \in \mathcal{E}_{\mathcal{X}}, \forall B' \in \mathcal{E}_{\mathcal{Y}} : P_{\mathcal{X},\mathcal{Y}}((x, y) \in A' \times B') = P_{\mathcal{X}}(x \in A')P_{\mathcal{Y}}(y \in B'). \tag{3.13}$$

Nous notons par $\mathcal{X} \perp \mathcal{Y}$ le fait que les variables $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes.

Exemple 2 (suite): double pile ou face (voir page 3.5). Dans le cas où les résultats PP, PF, FP, FF sont équiprobables, on peut se convaincre que les variables $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Z}$ sont deux à deux indépendantes. On a en effet (vérifier cela explicitement) que $\mathcal{X}_1 \perp \mathcal{X}_2$, $\mathcal{X}_1 \perp \mathcal{Z}$, et $\mathcal{X}_2 \perp \mathcal{Z}$.

3.4.2 Cas de variables aléatoires réelles

Dans le cas où les variables aléatoires sont réelles cette condition se traduit par

$$F_{\mathcal{X},\mathcal{Y}}(x, y) \stackrel{\triangle}{=} P(\mathcal{X}(\omega) < x \wedge \mathcal{Y}(\omega) < y) = P(\mathcal{X}(\omega) < x)P(\mathcal{Y}(\omega) < y) = F_{\mathcal{X}}(x)F_{\mathcal{Y}}(y), \tag{3.14}$$

où $F_{\mathcal{X}}(\cdot)$ et $F_{\mathcal{Y}}(\cdot)$ sont les fonctions de répartition respectivement de $\mathcal{X}$ et $\mathcal{Y}$. Si de plus $\mathcal{X}$ et $\mathcal{Y}$ admettent les densités $f_{\mathcal{X}}(\cdot)$ et $f_{\mathcal{Y}}(\cdot)$, alors il en est de même pour le couple, dont la densité est alors le produit de ces densités :

$$f_{\mathcal{X},\mathcal{Y}}(x,y) \triangleq \frac{\partial^2 F_{\mathcal{X},\mathcal{Y}}(x,y)}{\partial x \partial y} = f_{\mathcal{X}}(x) f_{\mathcal{Y}}(y).$$

Notons que nous pouvons étendre ces notions et propriétés par induction au cas d'un nombre fini quelconque de v.a. On parle alors de vecteurs aléatoires et le cas particulier intéressant est celui où celui-ci appartient à $\mathbb{R}^n$. Ces idées seront explorées plus en détails dans les chapitres suivants.

3.4.3 Indépendance de fonctions de variables aléatoires indépendantes

Si les variables $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes, alors quelles que soient les fonctions $\phi_{\mathcal{X}}$ et $\phi_{\mathcal{Y}}$ mesurables par rapport à $\mathcal{E}_{\mathcal{X}}$ et $\mathcal{E}_{\mathcal{Y}}$, les variables aléatoires $\phi_{\mathcal{X}} \circ \mathcal{X}$ et $\phi_{\mathcal{Y}} \circ \mathcal{Y}$ sont aussi indépendantes. C'est assez immédiat de s'en convaincre : en effet, soient deux ensembles A'' et B'' choisis respectivement dans $\mathcal{E}_{\phi_{\mathcal{X}}}$ et $\mathcal{E}_{\phi_{\mathcal{Y}}}$: à partir de ces ensembles les fonctions $\phi_{\mathcal{X}}$ et $\phi_{\mathcal{Y}}$ induisent des événements A' et B' de $\mathcal{E}_{\mathcal{X}}$ et $\mathcal{E}_{\mathcal{Y}}$ respectivement, et via ceux-ci des événements A et B de $\mathcal{E}$ par le biais des variables $\mathcal{X}$ et $\mathcal{Y}$: ces ensembles sont tels que $P(A \cap B) = P(A)P(B)$, vue l'indépendance des variables $\mathcal{X}$ et $\mathcal{Y}$.

3.5 ESPÉRANCE MATHÉMATIQUE D'UNE V.A. RÉELLE

La notion d'espérance mathématique est une notion centrale du calcul de probabilités. Pour cette raison, nous prenons un soin particulier pour l'introduire de façon progressive et rigoureuse. Nous commençons par une série de *premières définitions* qui s'appliquent dans des situations particulières, puis nous en donnons la définition tout à fait générale et rigoureuse d'un point de vue mathématique qui en induit toutes les propriétés qui seront présentées plus loin dans le chapitre présent et les deux suivants.

3.5.1 Premières définitions de la notion d'espérance mathématique

3.5.1.1 Cas où Ω est fini

Espérance mathématique d'une v.a. réelle définie sur un espace $(\Omega, \mathcal{E}, P)$ fini

Pour une variable aléatoire réelle définie sur un espace Ω fini, on définit son espérance mathématique (on dit aussi sa moyenne) par

$$E\{\mathcal{X}\} = \mu_{\mathcal{X}} \triangleq \sum_{\omega \in \Omega} \mathcal{X}(\omega) P(\omega) = \sum_{x_k \in \mathcal{X}} x_k P_{\mathcal{X}}(x_k), \tag{3.15}$$

où la seconde somme porte sur l'ensemble des valeurs possibles de la variable aléatoire.

Etant donné le caractère fini de Ω, cette grandeur est toujours finie.

Exemple et interprétation. Prenons l'exemple du double lancer de dés. Supposons que les deux dés soient équilibrés et indépendants et définissons une variable aléatoire qui lors d'un double lancer vaut la somme des deux faces supérieures. L'expérience correspond à un espace Ω avec $36 = 6 \times 6$ éléments possibles, chacun de probabilité $1/36$. L'espérance de cette variable aléatoire, selon la définition ci-dessus vaut 7.

Si nous répétons un très grand nombre de fois cette expérience disons $N \gg 1000$ fois, nous pensons intuitivement que chacun des 36 résultats possibles devrait se produire à peu près $N/36$ fois; si cela est vrai, alors la valeur moyenne de la variable aléatoire sur ces N réalisations devrait être proche de 7. Vu sous cet angle, l'espérance mathématique d'une variable aléatoire représente donc une valeur théorique moyenne qui serait observée dans ce genre d'expérience répétée.

Nous reviendrons à la fin de ce chapitre sur cette interprétation intuitive de la notion d'espérance, lorsque nous parlerons des notions de convergence de suites de variables aléatoires. Pour le moment, nous allons poursuivre notre étude des propriétés de cette notion du point de vue mathématique.

3.5.1.2 Cas où Ω est infini et que la variable aléatoire est discrète

Lorsque Ω est infini mais que la variable aléatoire ne prend quand même qu'un nombre fini de valeurs différentes, disons $\mathcal{X}(\omega) \in \{x_1, x_2, \ldots, x_n\}$, son espérance est définie par

$$E\{\mathcal{X}\} = \mu_{\mathcal{X}} = \sum_{k=1}^{n} x_k P_{\mathcal{X}}(x_k), \tag{3.16}$$

qui est équivalent au membre de droite de l'équation (3.15). Notons que cette valeur est indépendante de l'ordre choisi pour énumérer les x_k, puisqu'ils sont en nombre fini.

L'extension de cette formule au cas où la variable aléatoire prend un nombre infini dénombrable de valeurs différentes donne la définition candidate suivante

$$E\{\mathcal{X}\} = \mu_{\mathcal{X}} = \sum_{k=1}^{\infty} x_k P_{\mathcal{X}}(x_k). \tag{3.17}$$

Cette définition n'a d'utilité pratique que si cette série se comporte de la même manière quel que soit l'ordre choisi pour énumérer les valeurs x_k. Par conséquent, pour qu'une variable aléatoire discrète possède une espérance mathématique finie, il faut imposer que la série

$$\sum_{k=1}^{\infty} |x_k P_{\mathcal{X}}(x_k)| = \sum_{k=1}^{\infty} |x_k| P_{\mathcal{X}}(x_k) \tag{3.18}$$

converge. Dans ce cas la limite de (3.19) est finie et est indépendante de l'ordre choisi pour sommer.

On en tire la définition suivante de l'espérance d'une variable aléatoire discrète quelconque.

Espérance mathématique d'une v.a. réelle discrète quelconque

Pour une variable aléatoire discrète quelconque, on définit son espérance mathématique (on dit aussi sa moyenne) par

$$E\{\mathcal{X}\} = \mu_{\mathcal{X}} \stackrel{\triangle}{=} \sum_{k=1}^{\infty} x_k P_{\mathcal{X}}(x_k), \tag{3.19}$$

lorsque cette série converge absolument.

Il existe évidemment de nombreux exemples de variables aléatoires discrètes dont l'espérance n'existe pas. Il existe aussi des cas où la série (3.19) tend vers l'infini, indépendamment de l'ordre des termes; dans ce cas on dit que l'espérance est infinie. C'est par exemple le cas lorsque tous les termes sont de même signe et que la série ne converge pas absolument.

3.5.1.3 Cas où la variable aléatoire est continue

Remarquons tout d'abord que pour qu'une variable aléatoire puisse être continue, il est nécessaire que l'espace Ω soit infini non-dénombrable.

Espérance mathématique d'une v.a. réelle continue

Pour une variable aléatoire à valeurs réelles continue on définit son espérance par

$$E\{\mathcal{X}\} = \mu_{\mathcal{X}} = \int_{\mathbb{R}} x f_{\mathcal{X}}(x)\, \mathrm{d}x, \tag{3.20}$$

lorsque cette intégrale converge absolument.

Il faut souligner, même si c'est évident, que cette intégrale n'est pas toujours définie. Il existe en effet des variables aléatoires continues dont l'espérance mathématique calculée par cette formule n'est pas finie.

Par exemple, une variable aléatoire de *Cauchy* (voir plus loin), dont la densité est donnée par

$$f_{\mathcal{X}}(x) = \frac{1}{\pi(1+x^2)}, \tag{3.21}$$

n'admet pas d'espérance selon la formule (3.20).

Cependant, toute fonction continue et à support compact étant intégrable, toute variable aléatoire réelle continue et bornée admet une espérance mathématique. C'est donc le cas aussi pour la variable aléatoire de notre Exemple 3, et aussi en principe pour toute quantité physique qu'on peut supposer être bornée.

3.5.1.4 Cas général

Dans le cas général on peut combiner les deux formules ci-dessus en exploitant le fait que la fonction de répartition peut se décomposer en une partie discrète et une partie continue.

L'écriture générale, qu'on rencontre dans de nombreux ouvrages de référence, est la suivante

$$E\{\mathcal{X}\} = \int_{\Omega} \mathcal{X}(\omega)\, \mathrm{d}P(\omega), \tag{3.22}$$

où le dP indique que l'intégrale est prise par rapport à la mesure P définie sur l'espace de départ Ω, ce qui est équivalent à

$$E\{\mathcal{X}\} = \int_{\mathbb{R}} x\, \mathrm{d}P_{\mathcal{X}}(x), \tag{3.23}$$

où le $\mathrm{d}P_{\mathcal{X}}$ indique que l'espérance est calculée par rapport à la loi de probabilité induite sur l'espace d'arrivée.

La section suivante explique ces notations et précise les conditions générales d'existence de l'espérance.

3.5.2 • Définition mathématique rigoureuse de la notion d'espérance mathématique

Nous considérons un espace de probabilité quelconque $(\Omega, \mathcal{E}, P)$ et nous allons donner la définition mathématique de la notion générale d'espérance mathématique d'une variable aléatoire réelle quelconque définie sur cet espace. Cette définition suit une démarche en trois étapes, similaire à celle de la définition de la notion d'intégrale au sens de Lebesgue vue au cours d'analyse, lorsqu'on remplace la notion de longueur (ou en général de volume) d'un intervalle par la probabilité d'un événement. La démarche fournit les conditions d'existence, explique les notations utilisées dans les formules (3.22-3.23), et révèle les propriétés fondamentales de l'espérance mathématique.

3.5.2.1 Variable aléatoire non-négative simple

Une v.a. non-négative simple est une fonction définie sur Ω qui peut s'écrire sous la forme

$$\mathcal{Y}(\omega) = \sum_{k=1}^{n} y_k 1_{A_k}(\omega), \tag{3.24}$$

où les $y_k \in [0, \infty[$ sont des nombres réels non-négatifs, où les ensembles A_k sont des événements, et où les fonctions $1_{A_k}(\omega)$ sont les fonctions caractéristiques des événements A_k (cette fonction vaut 1 si $\omega \in A_k$, 0 sinon). On peut vérifier que ce type de fonction est nécessairement $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurable, puisque $A_k \in \mathcal{E}, \forall k$.

Pour un espace mesurable $(\Omega, \mathcal{E})$ donné, nous désignons par L_{Ω}^{s+} l'ensemble des variables aléatoires réelles simples qu'on peut y définir.

L'espérance mathématique d'une variable aléatoire non-négative simple est définie par

$$E\{\mathcal{Y}\} \triangleq \sum_{k=1}^{n} y_k P(A_k). \tag{3.25}$$

NB: L'écriture (3.24) n'est pas unique pour une v.a. simple, mais on peut montrer que la valeur (3.25) est indépendante de la façon d'exprimer $\mathcal{Y}$ sous la forme (3.24). Cette définition est évidemment conforme à la définition (3.15) lorsque Ω est fini, et aussi dans le cas où Ω n'est pas fini mais que la variable aléatoire ne prend qu'un nombre fini de valeurs différentes.

3.5.2.2 Variable aléatoire non-négative quelconque

Une variable aléatoire $\mathcal{X}$ est non-négative si $\forall\omega \in \Omega : \mathcal{X}(\omega) \in [0, \infty]$; nous tolérons donc que la variable prenne éventuellement la valeur ∞ en certains points de Ω.

Pour deux v.a. $\mathcal{X}$ et $\mathcal{Y}$ nous écrivons $\mathcal{Y} \leq \mathcal{X}$, si $\mathcal{Y}(\omega) \leq \mathcal{X}(\omega), \forall\omega \in \Omega$.

L'espérance mathématique d'une variable aléatoire non-négative quelconque est alors définie par

$$E\{\mathcal{X}\} \triangleq \int_\Omega \mathcal{X}(\omega)\, \mathrm{d}P(\omega) \triangleq \sup\{E\{\mathcal{Y}\} : \mathcal{Y} \in L_\Omega^{s+}, \mathcal{Y} \leq \mathcal{X}\} \leq \infty. \tag{3.26}$$

La borne supérieure qui définit l'espérance peut être finie ou bien infinie. Dans le cas où $E\{\mathcal{X}\} < \infty$ on dit que $\mathcal{X}$ est *P-intégrable* (nous dirons simplement qu'elle est *intégrable* dans ce qui suit).

Une variable aléatoire peut-être intégrable même si elle vaut ∞ par endroits, et aussi elle peut-être non-intégrable même si elle est finie partout sur Ω.

Notons que la définition (3.26) appliquée à une v.a. non-négative simple est équivalente à (3.25).

Conséquences importantes.

- si $\mathcal{X} \geq 0$ alors $[E\{\mathcal{X}\} = 0] \Leftrightarrow [P(\mathcal{X} = 0) = 1]$, et $[E\{\mathcal{X}\} < \infty] \Rightarrow [P(\mathcal{X} = \infty) = 0]$.
- la somme de deux variables aléatoires non-négatives intégrables est évidemment encore une variable aléatoire non-négative intégrable.
- si $\mathcal{X} \geq \mathcal{Y} \geq 0$ et que $\mathcal{X}$ est intégrable, alors $\mathcal{Y}$ l'est aussi.

3.5.2.3 Variable aléatoire réelle quelconque

Nous décomposons $\mathcal{X}$ en sa partie positive $\mathcal{X}^+$ et sa partie négative $\mathcal{X}^-$, avec

$$\mathcal{X}^+(\omega) = \begin{cases} \mathcal{X}(\omega) & \text{si} \quad \mathcal{X}(\omega) \geq 0, \\ 0 & \text{si} \quad \mathcal{X}(\omega) < 0, \end{cases} \quad \text{et } \mathcal{X}^- = (-\mathcal{X})^+. \tag{3.27}$$

On a bien sûr que $\mathcal{X} = \mathcal{X}^+ - \mathcal{X}^-$, et les deux variables $\mathcal{X}^+$ et $\mathcal{X}^-$ sont non-négatives. Nous disons que $\mathcal{X}$ est P-intégrable si $E\{\mathcal{X}^+\}$ et $E\{\mathcal{X}^-\}$ sont finies.

Lorsque $\mathcal{X}$ est intégrable son espérance mathématique est définie par

$$E\{\mathcal{X}\} \triangleq E\{\mathcal{X}^+\} - E\{\mathcal{X}^-\}, \tag{3.28}$$

et est finie.

Notons que cette définition est évidemment équivalente à (3.26) lorsque la variable aléatoire est non-négative.

Conséquences importantes.

- $\mathcal{X}$ est intégrable si, et seulement si $|\mathcal{X}| = \mathcal{X}^+ + \mathcal{X}^-$ est intégrable. On désigne par L_Ω^1 l'ensemble des variables aléatoires réelles intégrables pouvant être définies sur Ω.
- $|E\{\mathcal{X}\}| = |E\{\mathcal{X}^+\} - E\{\mathcal{X}^-\}| \leq |E\{\mathcal{X}^+\}| + |E\{\mathcal{X}^-\}| = E\{|\mathcal{X}|\}$ (et donc $E\{\mathcal{X}\} \leq E\{|\mathcal{X}|\}$)
- Si $\mathcal{X}$ et $\mathcal{Y}$ sont intégrables, alors $\mathcal{Z} = \alpha\mathcal{X} + \beta\mathcal{Y}$ est intégrable, $\forall\alpha, \beta \in \mathbb{R}$ et on a $E\{\mathcal{Z}\} = \alpha E\{\mathcal{X}\} + \beta E\{\mathcal{Y}\}$. L'ensemble L_Ω^1 est donc un espace vectoriel linéaire et l'opérateur d'espérance $E\{\cdot\}$ **est un opérateur linéaire** défini sur cet espace.
- Si $P(\mathcal{X} \neq \mathcal{Y}) = 0$ alors $E\{\mathcal{X}\} = E\{\mathcal{Y}\}$.
- Si $\mathcal{Y}$ peut s'écrire comme une fonction ϕ de $\mathcal{X}$, alors $E\{\mathcal{Y}\} = \int_{\mathbb{R}} \phi(x)\, \mathrm{d}P_{\mathcal{X}}(x)$. C'est le théorème de la mesure image qui justifie l'écriture (3.23) dans le cas particulier où $\phi(x) = x$.

3.5.2.4 Conditionnement et première version du théorème de l'espérance totale

Soit $B \in \mathcal{E}$, un événement de probabilité $P(B) > 0$ et $\mathcal{X}$ une v.a. intégrable. Nous désignons l'espérance conditionnelle de $\mathcal{X}$ sachant que l'événement B est réalisé par $E\{\mathcal{X}|B\}$. Il s'agit de l'espérance conditionnelle de $\mathcal{X}$ définie selon le schéma qui précède où on remplace la loi $P(\cdot)$, par la loi conditionnelle $P(\cdot|B)$.

On peut se convaincre que si $\mathcal{X}$ est intégrable alors la variable $\mathcal{Z} \stackrel{\triangle}{=} \mathcal{X}1_B$, égale à $\mathcal{X}$ sur B et nulle ailleurs, est aussi intégrable (on a en effet $|\mathcal{Z}| \leq |\mathcal{X}|$). On montre alors que

$$E\{\mathcal{X}|B\} = \frac{E\{\mathcal{X}1_B\}}{P(B)}.$$

Suggestion : faire le raisonnement pour le cas où $\mathcal{X}$ est une v.a. non-négative simple.

Si aussi $P(B^c) > 0$, alors

$$E\{\mathcal{X}\} = P(B)E\{\mathcal{X}|B\} + P(B^c)E\{\mathcal{X}|B^c\},$$

puisque $\mathcal{X} = (1_B + 1_{B^c})\mathcal{X}$.

Cette dernière égalité constitue en réalité une première version d'un résultat fondamental du calcul de probabilités, à savoir le **théorème de l'espérance totale**, sur la formulation générale duquel nous reviendrons plus en détails au chapitre 4.

3.5.3 Inégalité de Markov

Cette inégalité est fondamentale. Elle s'énonce comme suit:

Inégalité de Markov

Si $\mathcal{X}$ est une variable aléatoire positive (ou nulle) et d'espérance $\mu_{\mathcal{X}}$ finie (et donc aussi positive), alors $\forall c > 0$ on a

$$P(\mathcal{X} \geq c\mu_{\mathcal{X}}) \leq \frac{1}{c}. \tag{3.29}$$

Cette inégalité (que nous ne démontrerons pas) nous indique qu'une variable aléatoire positive ne peut dévier très au dessus de son espérance que très rarement.

Exemple. Si la durée de vie moyenne d'une batterie de voiture est de 3 ans, alors au moins 50% des batteries de voiture auront cessé de fonctionner après 6 ans, et au moins 75% auront cessé de fonctionner après 12 ans.

Nota Bene. Si la variable aléatoire est négative ou nulle, on obtient une borne similaire en appliquant l'inégalité de Markov à $\mathcal{Y} = -\mathcal{X}$. Si la variable est ni négative ni positive, mais bornée supérieurement ou inférieurement, des variantes peuvent être obtenues, en appliquant l'inégalité de Markov à des variables aléatoires "translatées" de façon adéquate.

3.5.4 Espérance mathématique d'une fonction d'une variable aléatoire

L'espérance d'une fonction $\phi(\cdot)$ à valeurs réelles d'une variable aléatoire discrète $\mathcal{X}$ (pas nécessairement à valeurs réelles) est obtenue par

Espérance d'un fonction à valeurs réelles d'une variable aléatoire discrète

$$E\{\phi(\mathcal{X})\} = \sum_{k=1}^{\infty} \phi(x_k)P(\mathcal{X} = x_k) = \sum_{k=1}^{\infty} \phi(x_k)P_{\mathcal{X}}(x_k), \tag{3.30}$$

pour autant que la série converge.

Similairement, dans le cas où la variable $\mathcal{X}$ est réelle et continue on a

Espérance d'un fonction à valeurs réelles d'une variable aléatoire réelle continue

$$E\{\phi(\mathcal{X})\} = \int_{\mathbb{R}} \phi(x) f_{\mathcal{X}}(x)\,\mathrm{d}x, \tag{3.31}$$

pour autant que cette intégrale soit définie.

Ces deux formules sont très utiles en pratique, puisqu'elles permettent le calcul de l'espérance d'une fonction d'une variable aléatoire sans devoir recourir au calcul de la loi de probabilité de cette fonction.

Cas particuliers importants :

1. Fonction constante ($\phi(x) = a$) : $E\{\phi\} = a$.
2. Produit par une constante ($\phi(x) = ax$) : $E\{\phi\} = aE\{\mathcal{X}\}$.
3. Somme avec une constante ($\phi(x) = x + a$) : $E\{\phi\} = E\{\mathcal{X}\} + a$.
4. Fonction additive ($\phi(x) = \sum_{i=1}^{n} \phi_i(x)$) : $E\{\phi\} = \sum_{i=1}^{n} E\{\phi_i\}$.

Ces propriétés résultent de la linéarité de l'opérateur d'intégration: l'intégrale d'une combinaison linaire de fonctions est la combinaison linéaire des intégrales (pour autant que toutes les intégrales soient bien définies). Elles seront illustrées lors des séances de répétitions et dans le cadre des travaux pratiques.

3.5.4.1 ◦ Fonctions convexes, concaves et inégalité de Jensen

Définition de la notion de convexité.

- **Ensemble convexe.** On dit qu'un sous-ensemble C de $\mathbb{R}^n$ est convexe, si et seulement si

$$[x, y \in C] \Rightarrow \lambda x + (1-\lambda)y \in C, \forall \lambda \in]0,1[. \tag{3.32}$$

  Les sous-ensembles convexes de $\mathbb{R}$ sont les intervalles, semi-intervales, bornés ou non, et $\mathbb{R}$ lui-même.

- **Fonction convexe.** Une fonction f définie sur un sous-ensemble convexe C de $\mathbb{R}^n$ est dite convexe, si et seulement si

$$[x, y \in C] \Rightarrow f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \forall \lambda \in]0,1[. \tag{3.33}$$

  Si l'inégalité est stricte $\forall x, y \in C, \forall \lambda \in]0,1[$, on dit que la fonction est strictement convexe. Lorsque la fonction est définie sur un ensemble convexe de $\mathbb{R}$ et qu'elle est deux fois dérivable, elle est convexe si et seulement sa dérivée seconde est non-négative, et strictement convexe si et seulement si sa dérivée seconde est strictement positive sauf peut-êtte en un nombre fini de points où elle peut être nulle.

- **Fonction concave.** Une fonction f est dite concave (respectivement strictement concave) sur C si et seulement si la fonction $g(x) = -f(x)$ est convexe (respectivement strictement convexe).

Exemples de fonctions convexes.

- La fonction $f(x) = x^2$ est strictement convexe sur tout sous-ensemble convexe de $\mathbb{R}$.
- Pour toute valeur non-nulle $\lambda \in \mathbb{R}$, la fonction $g(x) = \exp(\lambda x)$ est strictement convexe sur tout sous-ensemble convexe de $\mathbb{R}$.
- La fonction $h(x) = \frac{1}{x}$ est strictement convexe sur tout ensemble convexe de réels positifs et strictement concave sur tout ensemble convexe de réels négatifs.

- La fonction $\log x$ est strictement concave sur tout ensemble convexe de réels positifs.
- La fonction $|x|$ est convexe sur $\mathbb{R}$ mais pas strictement convexe.

Inégalité de Jensen.

Cette inégalité relie l'espérance d'une fonction convexe à la valeur de cette fonction appliquée à l'espérance.

Inégalité de Jensen

Si $\phi(\cdot)$ est une fonction convexe sur un sous-ensemble C de $\mathbb{R}$, et si $\mathcal{X}$ est une v.a. réelle à valeurs dans C, alors

$$E\{\phi(\mathcal{X})\} \geq \phi(E\{\mathcal{X}\}), \tag{3.34}$$

et si la fonction est strictement convexe, alors l'égalité implique $\mathcal{X} = cnste$, sauf éventuellement sur un ensemble de probabilité nulle (nous disons qu'elle est presque sûrement (p.s.) constante).

L'inégalité de Jensen est très utile en pratique. Elle permet par exemple de déduire que

$$E\{\mathcal{X}^2\} \geq (E\{\mathcal{X}\})^2, \text{ l'égalité n'étant vérifiée que si } \mathcal{X} \text{ est p.s. constante,}$$

que

$$E\{\mathcal{X}^2\} = E\{|\mathcal{X}|^2\} \geq (E\{|\mathcal{X}|\})^2, \text{ l'égalité n'étant vérifiée que si } \mathcal{X} \text{ est p.s. constante,}$$

et que

$$E\{|\mathcal{X}|\} \geq |E\{\mathcal{X}\}|, \text{ l'égalité n'étant vérifiée que si } \mathcal{X} \text{ est p.s. de signe constant.}$$

3.5.5 Espérance mathématique d'une fonction de deux ou plusieurs variables aléatoires

Soit $(\Omega, \mathcal{E}, P)$ un espace de probabilité, et soient $\mathcal{X}$ et $\mathcal{Y}$ deux variables aléatoires quelconques définies sur cet espace (avec leur σ-algèbre supposée telles que ces fonctions sont bien $(\mathcal{E}, \mathcal{E}_\mathcal{X})$ et $(\mathcal{E}, \mathcal{E}_\mathcal{Y})$ mesurables). Une fonction $\phi(\mathcal{X}, \mathcal{Y})$ à valeurs réelles mesurable par rapport à la σ-algèbre produit $\mathcal{E}_{\mathcal{X},\mathcal{Y}} = \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$ définit alors une variable aléatoire réelle sur $(\Omega, \mathcal{E}, P)$.

3.5.5.1 Cas général

Espérance mathématique d'une fonction de deux v.a. discrètes

Si les deux variables $\mathcal{X}$ et $\mathcal{Y}$ sont discrètes, le couple $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ est encore une variable aléatoire discrète et l'espérance de la variable ϕ se déduit donc de ce qui précède, et est donnée par la formule

$$E\{\phi(\mathcal{X}, \mathcal{Y})\} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \phi(x_i, y_j) P_{\mathcal{X},\mathcal{Y}}(x_i, y_j), \tag{3.35}$$

pour autant que la série converge absolument, et où $P_{\mathcal{X},\mathcal{Y}}(x_i, y_j)$ désigne $P(\mathcal{X} = x_i \wedge \mathcal{Y} = y_j)$.

Notons que si $\Omega_\mathcal{X}$ et $\Omega_\mathcal{Y}$ sont dénombrables, il en est de même de leur produit cartésien $\Omega_\mathcal{X} \times \Omega_\mathcal{Y}$, et donc le couple $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ est une variable aléatoire discrète dans ces conditions.

Espérance mathématique d'une fonction de deux v.a. conjointement continues

Si les deux variables sont continues et possèdent une densité conjointe (voir chapitre 4), l'espérance de la variable ϕ est donnée par la formule

$$E\{\phi(\mathcal{X}, \mathcal{Y})\} = \int_{\mathbb{R}} \int_{\mathbb{R}} \phi(x, y) f_{\mathcal{X},\mathcal{Y}}(x, y) \, \mathrm{d}x \, \mathrm{d}y, \tag{3.36}$$

pour autant que l'intégrale double converge.

Notons que si la fonction ϕ ne dépend en réalité que de l'une des deux variables aléatoires (disons qu'elle peut s'écrire sous la forme $\phi(\mathcal{X})$), alors ces deux formules donnent respectivement (voir chapitre 4)

$$E\{\phi(\mathcal{X},\mathcal{Y})\} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\phi(x_i)P_{\mathcal{X},\mathcal{Y}}(x_i,x_j) = \sum_{i=1}^{\infty}\phi(x_i)\left(\sum_{j=1}^{\infty}P_{\mathcal{X},\mathcal{Y}}(x_i,x_j)\right) = \sum_{i=1}^{\infty}\phi(x_i)P_{\mathcal{X}}(x_i)$$

et

$$E\{\phi(\mathcal{X},\mathcal{Y})\} = \int_{\mathbb{R}}\int_{\mathbb{R}}\phi(x)f_{\mathcal{X},\mathcal{Y}}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathbb{R}}\phi(x)\left(\int_{\mathbb{R}}f_{\mathcal{X},\mathcal{Y}}(x,y)\,\mathrm{d}y\right)\,\mathrm{d}x = \int_{\mathbb{R}}\phi(x)f_{\mathcal{X}}(x)\,\mathrm{d}x,$$

c'est-à-dire les formules déjà présentées plus haut.

Par ailleurs, si la fonction ϕ est une somme de deux (ou plusieurs) fonctions, son espérance est la somme des espérances de ces fonctions (pour autant que les espérances en question soient finies), toujours à cause de la linéarité de l'opérateur d'intégration/sommation:

$$E\{\sum_{i=1}^{n}\phi_i(\mathcal{X},\mathcal{Y})\} = \sum_{i=1}^{n}E\{\phi_i(\mathcal{X},\mathcal{Y})\}. \tag{3.37}$$

Enfin, ces idées peuvent être généralisées, en considérant un nombre fini quelconque de variables aléatoires $\mathcal{X}_i$ et un nombre quelconque fini de fonctions ϕ_j de ces variables aléatoires, de la manière suivante:

$$E\{\sum_{i=1}^{n}\phi_i(\mathcal{X}_1,\ldots,\mathcal{X}_m)\} = \sum_{i=1}^{n}E\{\phi_i(\mathcal{X}_1,\ldots,\mathcal{X}_m)\}. \tag{3.38}$$

On en déduit le théorème extrêmement important suivant (voir section 3.5.2) :

Linéarité de l'espérance mathématique

$$E\{\sum_{i=1}^{n}\lambda_i\mathcal{X}_i)\} = \sum_{i=1}^{n}\lambda_iE\{\mathcal{X}_i\}, \tag{3.39}$$

qui exprime le fait que l'espérance mathématique d'une combinaison linéaire de variables aléatoires d'espérance finie est la combinaison linéaire correspondante des espérances mathématiques de ces variables (**sans hypothèse d'indépendance entre les variables aléatoires en question**). Ce théorème reste vrai même si certaines des variables aléatoires sont ni continues ni discrètes, pour autant qu'elles soient toutes d'espérance finie.

3.5.5.2 Espérance mathématique d'un produit de deux variables aléatoires

Définissant une fonction ϕ, produit de deux v.a. ($\phi(\mathcal{X},\mathcal{Y}) = \mathcal{X}\mathcal{Y}$), son espérance est définie de manière générique par (*Suggestion : écrire cette formule lorsque les variables $\mathcal{X}$, $\mathcal{Y}$ sont conjointement continues, ou bien discrètes.*)

$$E\{\mathcal{X}\mathcal{Y}\} = \int_{\mathbb{R}^2}xy\,\mathrm{d}P_{\mathcal{X}\mathcal{Y}}(x,y).$$

Lorsque $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes, la mesure $dP_{\mathcal{X}\mathcal{Y}}(x,y)$ se factorise et l'intégrale double peut se décomposer en produit des deux intégrales simples :

$$E\{\mathcal{X}\mathcal{Y}\} = \int_{\mathbb{R}}x\,\mathrm{d}P_{\mathcal{X}}(x)\int_{\mathbb{R}}y\,\mathrm{d}P_{\mathcal{Y}}(y) = E\{\mathcal{X}\}E\{\mathcal{Y}\}.$$

La réciproque n'est pas vraie.

Les travaux pratiques et séances de répétition illustreront cette dernière formule.

3.6 VARIANCE, ÉCART-TYPE, COVARIANCE

3.6.1 Définition

Lorsque l'espérance $\mu_{\mathcal{X}} = E\{\mathcal{X}\}$ est finie, la **variance** de la variable aléatoire $\mathcal{X}$ est définie par

$$V\{\mathcal{X}\} = \sigma_{\mathcal{X}}^2 = E\{(\mathcal{X} - \mu_{\mathcal{X}})^2\}, \tag{3.40}$$

lorsque cette grandeur est finie.

L'**écart-type**, désigné par $\sigma_{\mathcal{X}}$, est la racine carrée positive de la variance $V\{\mathcal{X}\}$.

Notons que si $E\{\mathcal{X}^2\}$ est finie, alors l'espérance $E\{\mathcal{X}\}$ l'est aussi et aussi la variance $V\{\mathcal{X}\}$.

On définit la **covariance** de deux variables aléatoires réelles, $\mathcal{X}$ et $\mathcal{Y}$, par

$$\text{cov}\{\mathcal{X};\mathcal{Y}\} \triangleq E\left\{(\mathcal{X} - E\{\mathcal{X}\})(\mathcal{Y} - E\{\mathcal{Y}\})\right\} = E\{\mathcal{X}\mathcal{Y}\} - E\{\mathcal{X}\}E\{\mathcal{Y}\}. \tag{3.41}$$

Nous reviendrons au chapitre 4 sur les conditions d'existence et les relations entre ces grandeurs.

3.6.2 Propriétés de base

On a (a étant un nombre réel constant)

$$E\{(\mathcal{X} - a)^2\} = V\{\mathcal{X}\} + (E\{\mathcal{X}\} - a)^2, \tag{3.42}$$

et par conséquent, la variance est la valeur minimale de $E\{(\mathcal{X} - a)^2\}$, et $a = E\{\mathcal{X}\}$ minimise $E\{(\mathcal{X} - a)^2\}$. Cette propriété est exploitée très largement en statistiques, dans le domaine de l'estimation au sens des moindres carrés. Elle est démontrée à la section 4.3.4.1.

On en déduit, en prenant $a = 0$ que

$$V\{\mathcal{X}\} = E\{X^2\} - (E\{\mathcal{X}\})^2. \tag{3.43}$$

Par ailleurs, on a

- $V\{\mathcal{X} + a\} = V\{\mathcal{X}\}$.
- $V\{a\mathcal{X}\} = a^2 V\{\mathcal{X}\}$.
- $V\{\mathcal{X} + \mathcal{Y}\} = V\{\mathcal{X}\} + V\{\mathcal{Y}\} + 2\text{cov}\{\mathcal{X};\mathcal{Y}\}$.
  Si les v.a. sont indépendantes, on a $E\{\mathcal{X}\mathcal{Y}\} = E\{\mathcal{X}\}E\{\mathcal{Y}\}$ et donc $\text{cov}\{\mathcal{X};\mathcal{Y}\} = 0$. Dans ce cas
  $$V\{\mathcal{X} + \mathcal{Y}\} = V\{\mathcal{X}\} + V\{\mathcal{Y}\}.$$
  La réciproque n'est pas vraie. (7)

Ces notions seront illustrées et étudiées dans le cadre des répétitions et travaux pratiques.

3.6.3 Inégalité de Bienaymé-Tchebyshev

L'espérance et la variance sont reliées par *l'inégalité de Bienaymé-Tchebyshev* :

Inégalité de Bienaymé-Tchebyshev

Si $\mathcal{X}$ est une variable aléatoire d'espérance $\mu_{\mathcal{X}}$ et d'écart-type $\sigma_{\mathcal{X}}$, on a $\forall c > 0$:

$$P(|\mathcal{X} - \mu_{\mathcal{X}}| \geq c\sigma_{\mathcal{X}}) \leq \frac{1}{c^2}. \tag{3.44}$$

On déduit de cette inégalité que si $\sigma_{\mathcal{X}} = 0$ la v.a. est presque sûrement égale à son espérance $\mu_{\mathcal{X}}$, c'est-à-dire presque sûrement constante. La variance mesure donc bien le caractère aléatoire d'une v.a. du point de vue de ses écarts possibles par rapport à son espérance.

Notons que cette inégalité est une conséquence directe de l'inégalité de Markov appliquée à la variable aléatoire positive $\mathcal{Y} = (\mathcal{X} - \mu_{\mathcal{X}})^2$ dont l'espérance vaut $\sigma^2_{\mathcal{X}}$ et est donc finie lorsque la variance de $\mathcal{X}$ l'est.

L'inégalité de Markov est cependant plus générale, puisqu'elle s'applique à toute variable aléatoire positive d'espérance finie, et qu'il existe des variables aléatoires positives d'espérance finie mais dont la variance n'est pas finie. *(Suggestion : trouver un exemple de variable aléatoire positive, d'espérance finie et de variance infinie.)*

3.7 AUTRES MOMENTS

On définit, s'ils existent, les moments *centrés* d'ordre k par

$$\mu_k = E\left\{(\mathcal{X} - \mu_{\mathcal{X}})^k\right\}. \tag{3.45}$$

On a évidemment $\mu_1 = 0$ et $\mu_2 = \sigma^2_{\mathcal{X}}$. Si la distribution de la variable est symétrique par rapport à sa moyenne on a $\mu_{2k+1} = 0, \forall k$.

Le moment *non-centré* d'ordre k est quant à lui simplement défini par $E\left\{\mathcal{X}^k\right\}$. La moyenne est donc le moment non-centré d'ordre 1.

Les moments caractérisent ensemble (sauf exception rare) la loi de probabilité de la variable aléatoire (cf la discussion sur les fonctions caractéristiques et génératrices dans les sections qui suivent).

3.8 LOIS DE PROBABILITE D'USAGE COURANT

Dans cette section nous décrivons quelques lois de probabilités usuellement rencontrées et nous énonçons, sans les démontrer, leurs propriétés principales. Pour plus de détails nous renvoyons le lecteur intéressé à l'ouvrage de référence [Sap90].

3.8.1 Lois de variables discrètes

3.8.1.1 Loi uniforme

Il s'agit de la loi d'une variable aléatoire $\mathcal{X}$ définie sur $\{1, 2, \ldots, n\}$ et telle que chacune de ses n valeurs possibles $i \in \{1, 2, \ldots, n\}$ soit de probabilité $P(\mathcal{X}(\omega) = i) = \frac{1}{n}$.

On a donc

$$P_{\mathcal{X}}(i) = \frac{1}{n}, \forall i \in \{1, 2, \ldots, n\}, P_{\mathcal{X}}(i) = 0, \forall i \notin \{1, 2, \ldots, n\},$$

$$E\{\mathcal{X}\} = \sum_{i=1}^{n} i\frac{1}{n} = \frac{n+1}{2}$$

et

$$V\{\mathcal{X}\} = \sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2 \frac{1}{n} = \frac{n^2-1}{12}.$$

3.8.1.2 Loi de Bernoulli

C'est une loi d'une v.a. $\mathcal{X}$ ne pouvant prendre que deux valeurs possibles 1 ou 0, avec les probabilités p et $1-p$. En d'autres termes, $\mathcal{X}$ est la fonction indicatrice d'un événement A de probabilité $P(A) = p$.

On a

$$E\{\mathcal{X}\} = 1 * p + 0 * (1-p) = p$$

et

$$V\{\mathcal{X}\} = (1-p)^2 p + (0-p)^2(1-p) = p(1-p).$$

Exemple 3 (suite). Consommation électrique d'un bâtiment. Supposons qu'un appareil (disons une lampe) ait une probabilité $p = 10^{-1}$ d'être branché à un moment particulier, et consomme à ce moment une puissance fixe (disons de $100W$). La consommation de cet appareil peut alors être modélisée par une variable $\mathcal{X}$ qui est le produit de la constante 100 et d'une variable de Bernoulli. On a donc que $E\{\mathcal{X}\} = 100p = 10W$, et que $V\{\mathcal{X}\} = 100^2 p(1-p) = 900$. Donc $\sigma_{\mathcal{X}} = 30W$.

3.8.1.3 Loi binomiale

Supposons qu'on répète n fois une expérience de Bernoulli, et qu'on compte le nombre de fois sur n que l'événement A est réalisé. Désignons par $\mathcal{X}$ la variable aléatoire qui désigne le compte. $\mathcal{X}$ est la somme de n v.a. indépendantes et identiquement distribuées (i.i.d.)

$$\mathcal{X} = \sum_{i=1}^{n} \mathcal{X}_i.$$

La loi de cette v.a. est par définition la loi binomiale $\mathcal{B}(n,p)$. Les valeurs possibles de $\mathcal{X}$ sont $\{0,1,\ldots,n\}$

On a $E\{\mathcal{X}\} = np$, et $V\{\mathcal{X}\} = np(1-p)$. D'autre part, on a

$$P(\mathcal{X} = k) = C_n^k p^k (1-p)^{(n-k)}.$$

En effet, une réalisation particulière d'un tirage $x_1,\ldots,x_n$ qui satisfait $\sum_i x_i = k$ est de probabilité $p^k(1-p)^{n-k}$, et il y a au total autant de réalisations différentes $x_1,\ldots,x_n$ qui satisfont $\sum_i x_i = k$, qu'il y a de façons de choisir un sous-ensemble de taille k dans un ensemble de taille n (c'est-à-dire C_n^k, voir appendice A.2).

On a la propriété importante (et évidente) suivante : soient $\mathcal{X} \sim \mathcal{B}(n_1,p)$ et $\mathcal{Y} \sim \mathcal{B}(n_2,p)$ indépendantes, alors

$$\mathcal{Z} = \mathcal{X} + \mathcal{Y} \sim \mathcal{B}(n_1+n_2,p).$$

En effet, faire deux expériences successives indépendantes avec respectivement n_1 et n_2 essais et compter la somme des succès, revient manifestement à compter le nombre de succès dans une seule expérience avec n_1+n_2 essais. La loi binomiale permet la modélisation des tirages successifs avec remise.

Si au lieu de compter le nombre total de succès, on considère la proportion de succès $\mathcal{Z} = \frac{\mathcal{X}}{n}$, on déduit que $E\{\mathcal{Z}\} = p$ et $V\{\mathcal{Z}\} = \frac{p(1-p)}{n}$. On constate que la variance de cette variable tend vers zéro lorsque le nombre d'essais tend vers l'infini. L'inégalité de Bienaymé-Tchebyshev permet donc d'affirmer que la proportion $\mathcal{Z}$ de succès tend (en un certain sens, voir plus loin) vers la probabilité de succès p lorsque n est suffisamment grand.

Exemple 3 (suite). Consommation électrique d'un bâtiment. Supposons qu'il y a n lampes (disons $n = 100$) chacune ayant une probabilité $p = 10^{-1}$ d'être branchée à un moment particulier, et consomme à ce moment une puissance fixe (disons de $100W$). Si nous supposons que les 100 lampes sont allumées par des personnes agissant de manière indépendante, le nombre de lampes branchées à tout moment est une variable binomiale $\mathcal{B}(n,p)$. La consommation totale de ces lampes peut donc être modélisée par une variable $\mathcal{X}$ qui est le produit de la constante 100 et d'une variable de Bernoulli $\mathcal{B}(n,p)$. On a donc que $E\{\mathcal{X}\} = 100np = 1000W$, et que $V\{\mathcal{X}\} = 100^2 np(1-p) = 90000W^2$. Donc $\sigma_{\mathcal{X}} = 300W$. On constate que la distribution se concentre autour de la moyenne.

3.8.1.4 Loi de Poisson

La loi de Poisson $\mathcal{P}(\lambda)$ est la loi d'une v.a. entière positive ou nulle qui satisfait à

$$P_{\mathcal{X}}(x) = \exp(-\lambda)\frac{\lambda^x}{x!}.$$

On a $E\{\mathcal{X}\} = \lambda$, et $V\{\mathcal{X}\} = \lambda$. On montre que si $\mathcal{X}_n \sim \mathcal{B}(n,p)$ est une suite de v.a. binomiales telle que

$$\lim_{n\to\infty} np = \lambda,$$

alors $\mathcal{X}_n$ converge en loi (voir ci-dessous) vers $\mathcal{P}(\lambda)$.

Exemple 3 (suite). Consommation électrique d'un bâtiment. Dans notre exemple ci-dessus on a $np = 10$, $p = 10^{-1}$, $n = 100$. On demande de comparer la loi $\mathcal{B}(n, p)$ avec la loi $\mathcal{P}_{np}$, avec MatLab.

Applications de la loi de Poisson. La loi de Poisson est utilisée dans des nombreuses applications techniques pour modéliser le nombre d'événements d'un certain type survenant pendant une période de temps fixée (par exemple, nombre de pannes par an, nombre d'arrivées dans une file d'attente, nombre de pièces défectueuses dans un lot de production etc.). Nous verrons plus loin que cette loi est très liée aux processus de Poisson.

3.8.2 Lois de variables continues

3.8.2.1 Loi uniforme

La loi uniforme sur $[0, a]$, notée $\mathcal{U}_{[0,a]}$ est définie par la densité uniforme $u_{[0,a]}(x) = \frac{1}{a}$ sur $[0, a]$, et 0 ailleurs.

On a $E\{\mathcal{X}\} = \frac{a}{2}$, et $V\{\mathcal{X}\} = \frac{a^2}{12}$.

La somme de deux v.a. uniformes indépendantes est une loi triangulaire (cf Figure 3.6).

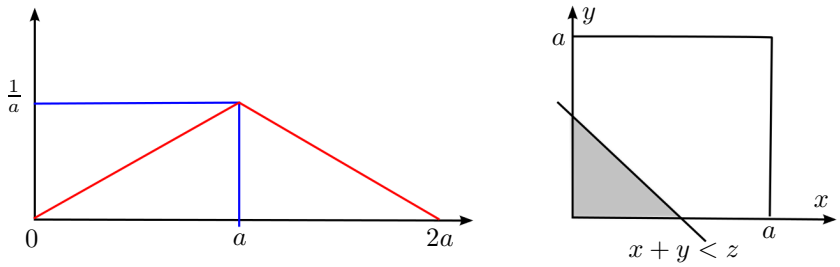


Figure 3.6: A gauche : densité de probabilité de la loi uniforme (en bleu) et de la loi triangulaire de la somme de deux variables uniformes indépendantes (en rouge). A droite : représentation géométrique de la probabilité $P(x + y < z)$ avec $\mathcal{X}, \mathcal{Y}$ uniformes sur $[0, a]$ et indépendantes

Le graphique de droite de la Figure 3.6 illustre géométriquement la probabilité que la somme de deux variables aléatoires uniformes sur $[0, a]$ soit inférieure à une certaine valeur z: il s'agit de l'aire hachurée, multipliée par la densité produit $f_{\mathcal{X}} f_{\mathcal{Y}} = \frac{1}{a^2}$. On a par conséquent, si $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$, que

$$F_{\mathcal{Z}}(z) = P(\mathcal{X} + \mathcal{Y} < z) = \frac{z^2}{2a^2}, \text{ lorsque } z \in [0, a] \text{ et}$$

$$F_{\mathcal{Z}}(z) = 1 - \frac{(2a - z)^2}{2a^2}, \text{ lorsque } z \in [a, 2a].$$

On en déduit que la densité $f_{\mathcal{Z}}(z) = F'_{\mathcal{Z}}(z)$ vaut

$$f_{\mathcal{Z}}(z) = \frac{z}{a^2}, \text{ lorsque } z \in [0, a] \text{ et}$$

$$f_{\mathcal{Z}}(z) = \frac{(2a - z)}{a^2}, \text{ lorsque } z \in [a, 2a].$$

Exemple 4. Intérêt d'un réseau électrique reliant deux villes. Nous considérons une version très simplifiée du problème mentionné à la Section 1.3.1.1. Deux villes doivent être alimentées en électricité, grâce à un système électrique composé de deux centrales électriques, respectivement situées dans chacune des deux villes. Les demandes instantanées en puissance électrique des deux villes sont des variables aléatoires $\mathcal{X}_1$ et $\mathcal{X}_2$, que nous supposons uniformément réparties sur l'intervalle $[0, P_{\max}]$ et indépendantes. Chaque ville dispose d'une centrale électrique de puissance $0.9P_{\max}$, qui peut ajuster sa production à la demande locale pour en couvrir au moins 90% dans le pire cas. Nous considérons d'abord le scénario où les deux villes vivent en autarcie, chacune utilisant sa centrale pour fournir sa demande; ensuite nous considérons une situation où les

deux villes se mutualisent pour construire une ligne électrique de capacité $0.1P_{\max}$, afin de partager la capacité des deux centrales pour couvrir la demande des deux villes. On demande de calculer, dans les deux scénarios, la probabilité p_{def} que la demande d'au moins une des deux villes ne soit pas desservie à un moment donné (probabilité de défaillance), et l'espérance mathématique de la quantité totale de puissance non-desservie $\mathcal{Y}$.

Nous esquissons ci-dessous la solution (en invitant le lecteur à vérifier les raisonnements et les calculs de façon aussi scrupuleuse que possible) :

- **Fonctionnement en autarcie.**
  - Raisonnons d'abord pour la première ville :
    - La probabilité p_{def_1} que la demande cette ville ne soit pas complètement desservie vaut bien entendu $P(\mathcal{X}_1 > 0.9P_{\max}) = 1 - F_{\mathcal{X}_1}(0.9P_{\max})$; puisque $\mathcal{X} \sim \mathcal{U}[0, P_{\max}]$ on a $p_{\text{def}_1} = 0.1$;
    - désignons par $\mathcal{Y}_1 = \max(0, \mathcal{X}_1 - 0.9P_{\max})$ la variable qui mesure la quantité de charge ne pouvant être desservie; $\mathcal{Y}_1$ n'est ni discrète ni continue, car elle prend une valeur nulle avec une probabilité de $1 - p_{\text{def}_1} = 0.9$ et avec une probabilité de 0.1 est distribuée de façon uniforme sur $[0, 0.1P_{\max}]$.
      On calcule que $E\{\mathcal{Y}_1\} = 0.9 * 0.0 + 0.1 * 0.05P_{\max} = 0.005P_{\max}$.
  - De même, pour la seconde ville on a $p_{\text{def}_2} = 0.1$, et $E\{\mathcal{Y}_2\} = 0.005P_{\max}$.
  - La probabilité d'une défaillance dans au moins une des deux villes vaut donc $p_{\text{def}} = 0.1 + 0.1 - (0.1)^2 = 0.19$, puisque les demandes dans les deux villes varient de manière indépendante (et que donc les deux types de défaillances sont des événements indépendants).
  - La quantité totale $\mathcal{Y}$ de la demande des deux villes non desservie vaut $= \mathcal{Y}_1 + \mathcal{Y}_2$. Son espérance vaut donc $E\{\mathcal{Y}\} = E\{\mathcal{Y}_1\} + E\{\mathcal{Y}_2\} = 0.01P_{\max}$.
- **Fonctionnement mutualisé.**
  - Notons que la capacité $0.1P_{\max}$ de la ligne est juste suffisante pour acheminer la puissance encore disponible dans l'une ou l'autre des villes et qui pourrait être nécessaire pour suppléer la capacité de production locale de l'autre.
  - Nous considérons donc la capacité totale de production des deux villes (qui est de $1.8P_{\max}$) face à leur demande totale $\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2$, qui est de loi triangulaire (sur l'intervalle $[0, 2P_{\max}]$) .
  - La probabilité p_{def} vaut donc $1 - F_{\mathcal{X}}(1.8P_{\max})$, soit 0.02.
  - La quantité totale de puissance non-desservie vaut quant à elle $\mathcal{Y} = \max(0, \mathcal{X} - 1.8P_{\max})$ et est distribuée sur l'intervalle $[0, 0.2P_{\max}]$; elle vaut 0.0 avec une probabilité de 0.98 et possède ensuite une densité qui décroit de manière linéaire de 0.0 à $0.2P_{\max}$.
    On trouve $E\{\mathcal{Y}\} = 0.98 * 0.0 + 0.02 * 0.066666P_{\max} = 0.0013333P_{\max}$.

Cet exercice met en évidence l'intérêt potentiel de la construction d'une ligne électrique de façon à permettre le partage de ressources face à des événements indésirables. Afin d'en évaluer l'intérêt réel, il faudrait le compléter par une analyse économique comprenant une évaluation des coûts associés aux pannes et des coûts d'investissement et d'exploitation de la ligne. Ce type d'analyse économique très fréquent dans les problèmes d'arbitrages techniques sort cependant du cadre de ce cours.

Notons aussi que notre modèle très simpliste peut être enrichi de beaucoup de façons différentes, par exemple en prenant en compte les probabilités de pannes des centrales électriques (elles aussi en général indépendantes), ou les dépendances partielles entre les comportements des consommateurs (en partie liés à des effets communs tels que les cycles journaliers, hebdomadaires, et saisonniers, et la météo), et des lois de probabilité plus réalistes que les lois uniformes. Nous reviendrons plus loin dans ces notes sur ce type de considérations.

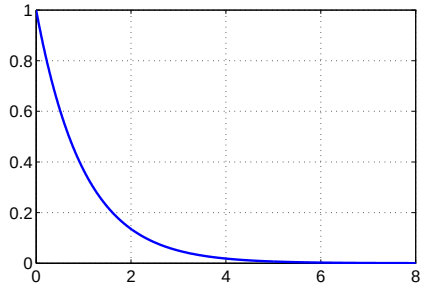
3.8.2.2 Loi exponentielle

La densité de la loi exponentielle de paramètre λ est

$$f(x) = \lambda \exp(-\lambda x)$$

si $x > 0$, 0 ailleurs.

On a $E\{\mathcal{X}\} = \frac{1}{\lambda}$, et $V\{\mathcal{X}\} = \frac{1}{\lambda^2}$, et $F_{\mathcal{X}}(x) = 1 - \exp(-\lambda x)$.




```
x = 0:0.01:8;
lambda = 1;
y = lambda * exp(-lambda * x);
plot(x, y, 'LineWidth', 2);
```

Figure 3.7: Densité de probabilité de la loi exponentielle avec $\lambda = 1$, avec le code MATLAB permettant de générer la figure

Applications. La loi exponentielle est souvent utilisée dans les études de fiabilité pour représenter la durée de vie d'un composant, notamment dans le domaine de l'électronique. Dans ce contexte l'espérance $1/\lambda$ est souvent appelée le *Mean Time Between Failures (MTBF)* et λ le taux instantané de défaillance.

En effet, si $\mathcal{X}$ désigne une variable mesurant la durée de vie d'un composant, on a que $1 - F_{\mathcal{X}}(x)$ mesure la probabilité que le composant soit encore en vie au moment x, et $f_{\mathcal{X}}(x)dx$ mesure la probabilité que ce composant tombe en panne juste après ce moment. On a donc que la probabilité conditionnelle que le composant tombe en panne juste après x sachant qu'il a survécu jusque là vaut

$$\frac{f_{\mathcal{X}}(x)dx}{1 - F_{\mathcal{X}}(x)}.$$

La grandeur $h_{\mathcal{X}}(x) = \frac{f_{\mathcal{X}}(x)}{1-F_{\mathcal{X}}(x)}$ désigne le taux instantané de défaillance. Dans le cas d'une loi exponentielle on calcule que $h_{\mathcal{X}}(x) = \lambda$; il est donc constant ici, ce qui peut s'interpréter en disant que le risque de tomber en panne ne dépend pas de l'âge du composant. C'est pourquoi on utilise cette loi pour modéliser des processus de défaillance de composants qui ne souffrent pas de vieillissement sur leur période d'exploitation.

3.8.2.3 Loi Gaussienne (ou normale)

$\mathcal{X}$ suit une loi Gaussienne de moyenne μ et de variance σ^2, notée $\mathcal{G}(\mu, \sigma^2)$ ou $\mathcal{N}(\mu, \sigma^2)$, si sa densité est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

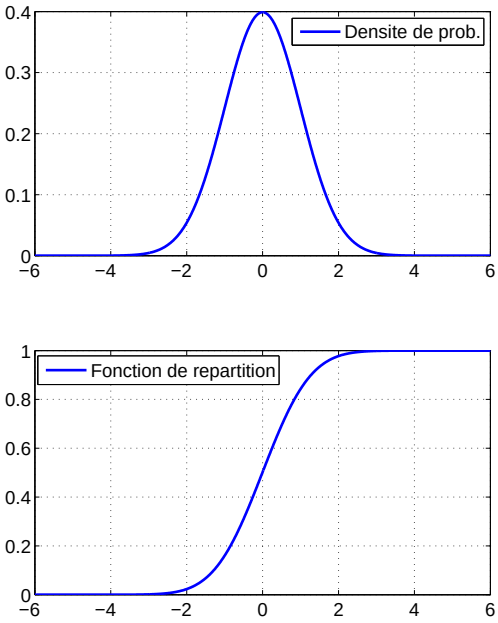
On a $E\{\mathcal{X}\} = \mu$, et $V\{\mathcal{X}\} = \sigma^2$. Si $\mu = 0$ on dit que la loi est centrée. Si $\sigma = 1$ on dit qu'elle est réduite. La figure 3.8 montre l'allure de cette loi.

On montre que la loi de Gauss peut être obtenue en transformant deux variables $\mathcal{X}_1$ et $\mathcal{X}_2$ uniformes $\mathcal{U}_{[0,1]}$ et indépendantes. En effet les variables $\mathcal{Y}_1$ et $\mathcal{Y}_2$ obtenues par

$$\mathcal{Y}_1 = -\sqrt{2\ln \mathcal{X}_1}\cos(2\pi\mathcal{X}_2) \quad (3.46)$$
$$\mathcal{Y}_2 = -\sqrt{2\ln \mathcal{X}_1}\sin(2\pi\mathcal{X}_2), \quad (3.47)$$

sont alors indépendantes et suivent chacune une loi normale centrée réduite $\mathcal{N}(0;1)$. Cette propriété est utile pour obtenir des variables aléatoires Gaussiennes à partir d'un générateur de nombres aléatoires (cf. exercices MATLAB).

Exemple 3 (suite). Consommation électrique d'un bâtiment. Dans notre exemple ci-dessus on a $np = 10$, $p = 10^{-1}$, $n = 100$. On demande de comparer la loi $\mathcal{B}(n,p)$ avec la loi $\mathcal{N}(np, np(p-1))$. Un exercice qu'on pourra faire avec MatLab. Refaire le même exercice en considérant $n = 10000$.




```
x = -6:0.01:6;
y = 1 / (sqrt(2 * pi)) * exp(-0.5 * x .^ 2);
plot(x, y, 'LineWidth', 2);
hold on;
```



```
z = zeros(size(x));
for i=1:length(x)
    z(i) = sum(y(1:i));
end
z = z * (x(2) - x(1));
plot(x, z, 'LineWidth', 2);
```

Figure 3.8: (a) Densité de probabilité de la loi normale centrée réduite $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$. (b) Fonction de répartition $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z$. (Avec le code MATLAB permettant de générer ces courbes).

Additivité. Si $\mathcal{X} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{Y} \sim \mathcal{N}(\mu_2, \sigma_2^2)$ sont deux variables aléatoires **indépendantes**, alors leur somme suit encore une loi normale et on a

$$\mathcal{Z} = \mathcal{X} + \mathcal{Y} \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

La moyenne de n v.a. normales centrées réduites indépendantes est une variable normale centrée de variance $\frac{1}{\sqrt{n}}$.

La loi Gaussienne joue un rôle très important, notamment à cause du théorème central-limite qui permet d'affirmer que la loi est d'application dans de nombreuses situations pratiques. Nous allons en voir au chapitre 5 la généralisation aux vecteurs aléatoires de dimension n.

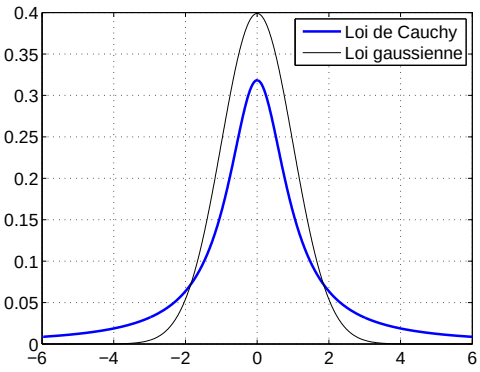
Commentaires historiques (voir [Jay03]). La découverte de la loi normale est attribuée au mathématicien de Moivre (1733) qui tomba sur cette loi de façon accidentelle sans y accorder de l'importance. Les mathématiciens Laplace et Bernoulli l'étudièrent au 18ème siècle, sous différents angles, mais ce fut Gauss, qui en 1809 en fournira une justification fondamentale dans le contexte de la modélisation des erreurs de mesures en astronomie. Cette justification repose sur un argument de symétrie, qui pose que la distribution des erreurs de mesure doit être indépendante du repère; cette simple hypothèse conduit obligatoirement à la forme mathématique de la loi normale; c'est encore Laplace qui reconnut rapidement le trait de génie de Gauss, et suite à cela la loi porte le nom de loi Gaussienne. Son caractère "normal" est à la fois discutable et justifié par son ubiquité en pratique. A l'époque moderne, cette loi est utilisée dans de nombreux contextes en sciences appliquées et en physique; une façon de la caractériser est de dire que c'est la loi d'une variable continue dont on a spécifié la moyenne et l'écart-type, sans rien imposer de plus en ce qui concerne ses moments d'ordre supérieur. Son invariance par combinaison linéaire, que nous étudierons dans les chapitres suivants, la rend à la fois crédible comme outil de modélisation et pratique du point de vue de sa manipulation algorithmique.

3.8.2.4 Loi de Cauchy

La distribution de *Cauchy* (voir Figure 3.9) est définie par

$$f_{\mathcal{X}}(x) = \frac{1}{\pi(1+x^2)}. \tag{3.48}$$

Cette loi n'admet pas d'espérance (la densité décroit trop lentement pour que l'intégrale converge), ni a fortiori de variance.




```
x = -6:0.01:6;
g = 1 / (sqrt(2 * pi)) * exp(-0.5 * x .^2);
c = 1 ./ (pi * (1 + x .^2));
plot(x, c, 'LineWidth', 2);
hold on;
plot(x, g, 'k');
```

Figure 3.9: Densité de probabilité de la loi de Cauchy comparée à celle de la loi normale centrée réduite, avec le code MATLAB permettant de générer ces courbes

On montre que la loi de Cauchy est la loi du rapport de deux variables aléatoires $\mathcal{N}(0;1)$ indépendantes.

3.9 ○ CONVOLUTION, FONCTIONS CARACTÉRISTIQUES ET FONCTIONS GÉNÉRATRICES

Dans cette section nous développons quelques notions utiles pour manipuler efficacement les fonctions de densités de manière analytique. Du point de vue mathématique, ces notions anticipent sur des matières qui seront vues plus en profondeur dans d'autres enseignements. Nous ne les utiliserons que marginalement dans la suite de ces notes.

3.9.1 Convolution

Un problème courant consiste à trouver la loi de probabilité (densité ou distribution) d'une variable aléatoire qui s'exprime comme la somme d'un certain nombre de variables aléatoires à valeurs réelles et indépendantes, et dont les lois sont connues. Le produit de convolution permet d'effectuer cette opération.

Considérons le cas de deux variables aléatoires réelles continues et indépendantes de densités $f_{\mathcal{X}}$ et $f_{\mathcal{Y}}$, et soit $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$. Calculons la fonction de répartition de la variable $\mathcal{Z}$, par [8]

$$F_{\mathcal{Z}}(z) = \int_{\mathbb{R}} \int_{\mathbb{R}} 1(x+y<z) f_{\mathcal{X}}(x) f_{\mathcal{Y}}(y)\,\mathrm{d}x\,\mathrm{d}y.$$

Un changement de variables $x+y=v$ et $x=u$ donne alors

$$F_{\mathcal{Z}}(z) = \int_{\mathbb{R}} \int_{\mathbb{R}} 1(v<z) f_{\mathcal{X}}(u) f_{\mathcal{Y}}(v-u)\,\mathrm{d}u\,\mathrm{d}v = \int_{\mathbb{R}} 1(v<z)\,\mathrm{d}v \int_{\mathbb{R}} f_{\mathcal{X}}(u) f_{\mathcal{Y}}(v-u)\,\mathrm{d}u.$$

Par conséquent, la fonction

$$f_{\mathcal{Z}}(v) = \int_{\mathbb{R}} f_{\mathcal{X}}(u) f_{\mathcal{Y}}(v-u)\,\mathrm{d}u \tag{3.49}$$

est bien telle que

$$F_{\mathcal{Z}}(z) = \int_{\mathbb{R}} 1(v < z) f_{\mathcal{Z}}(v)\,\mathrm{d}v$$

et il s'agit donc bien de la densité de la variable $\mathcal{Z}$.

L'opération qui prend deux fonctions f et g et calcule la fonction $f * g$ définie par

$$(f * g)(v) = \int_{\mathbb{R}} f(u) g(v-u)\,\mathrm{d}u$$

est appelée **produit de convolution**. Cette opération est commutative et jouit d'une série d'autres propriétés intéressantes (elle seront analysées plus en détails dans d'autres enseignements). Dans le contexte qui nous intéresse, nous pouvons l'utiliser pour construire la densité d'une somme d'un nombre quelconque de variables aléatoires indépendantes en l'appliquant plusieurs fois de suite, et en faisant dans un ordre quelconque les produits de convolution de leurs densités.

3.9.1.1 Exemples

L'intérêt du produit de convolution sera mis en évidence lors des séances d'exercices, notamment pour montrer que la convolution de deux lois uniformes donne bien une loi triangulaire et que la convolution de deux lois gaussiennes donne bien une loi gaussienne.

3.9.2 Fonctions caractéristiques

La fonction caractéristique d'une variable aléatoire à valeurs réelles et continue est la transformée de Fourier (la convention de signe adoptée en calcul des probabilités est opposée à celle utilisée en traitement du signal) de sa densité de probabilité [(9)]. On a ($t \in \mathbb{R}$ et $i = \sqrt{-1}$)

$$\phi_{\mathcal{X}}(t) = \int_{\mathbb{R}} e^{itx} f_{\mathcal{X}}(x)\,\mathrm{d}x. \tag{3.50}$$

Cette fonction caractéristique existe toujours [(10)] et est une fonction continue de son argument t.

3.9.2.1 Propriétés des fonctions caractéristiques

L'intérêt de la notion de fonction caractéristique est directement lié à ses propriétés mathématiques. Nous les énonçons sans les démontrer, cela faisant l'objet d'autres enseignements.

1. La fonction caractéristique de la variable $a\mathcal{X}$ est
$$\phi_{a\mathcal{X}}(t) = \phi_{\mathcal{X}}(at).$$
2. La fonction caractéristique de la variable $a + \mathcal{X}$ est
$$\phi_{a+\mathcal{X}}(t) = e^{iat}\phi_{\mathcal{X}}(t).$$
3. La fonction caractéristique d'une densité qui s'exprime comme une convolution de deux densités est le produit des fonctions caractéristiques. En particulier, pour deux variables indépendantes $\mathcal{X}, \mathcal{Y}$ on a
$$\phi_{\mathcal{X}+\mathcal{Y}}(t) = \phi_{\mathcal{X}}(t)\phi_{\mathcal{Y}}(t).$$
Cela s'étend donc aussi à une somme d'un nombre quelconque de variables aléatoires indépendantes.
4. Dérivées à l'origine de la fonction caractéristique:
$$\left.\frac{\partial^k \phi_{\mathcal{X}}(t)}{(\partial t)^k}\right|_{t=0} = i^k E(\mathcal{X}^k).$$
5. Inversion :
$$f_{\mathcal{X}}(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \phi_{\mathcal{X}}(t) e^{-itx}\,\mathrm{d}t.$$

3.9.3 Fonctions caractéristiques des lois usuelles

3.9.4 Lois discrètes

Pour les lois discrètes la transformé de Fourier est obtenue par la formule

$$\phi_{\mathcal{X}}(t) = \sum_{i=1}^{\infty} e^{itx_i} P_{\mathcal{X}}(x_i). \tag{3.51}$$

- La loi de Bernoulli: $\phi_{\mathcal{X}}(t) = (pe^{it} + q)$.
- La loi binomiale: $\phi_{\mathcal{X}}(t) = (pe^{it} + q)^n$.
- La loi de Poisson: $\phi_{\mathcal{X}}(t) = e^{\lambda(e^{it}-1)}$.

3.9.5 Lois continues

- Loi uniforme sur $[-a, a]$: $\phi_{\mathcal{X}}(t) = \frac{\sin at}{at}$.
- Loi Gaussienne centrée réduite $\mathcal{N}(0; 1)$: $\phi_{\mathcal{X}}(t) = e^{-t^2/2}$.
- Loi Gaussienne quelconque $\mathcal{N}(\mu; \sigma^2)$: $\phi_{\mathcal{X}}(t) = e^{\left(i\mu t - \frac{\sigma^2}{2}t^2\right)}$.

On déduit de cette dernière formule que la somme de deux variables Gaussienne indépendantes est encore une Gaussienne, et plus généralement que toute combinaison affine de variables Gaussiennes indépendantes est encore une variable Gaussienne.

3.9.6 Fonctions génératrices des moments

On définit la fonction génératrice d'une variable aléatoire réelle $\mathcal{X}$ par la transformé de Laplace de la loi de $-\mathcal{X}$:

$$M_{\mathcal{X}}(t) = E\{e^{t\mathcal{X}}\} = \int_{\mathbb{R}} e^{tx} f_{\mathcal{X}}(x)\,\mathrm{d}x. \tag{3.52}$$

On a la propriété importante suivante: $E(\mathcal{X}^n) = M_{\mathcal{X}}^{(n)}(0)$: les moments non centrés de $\mathcal{X}$ sont obtenus à partir des dérivées à l'origine de la fonction génératrice des moments.

Ces fonctions peuvent être exploitées pour calculer les moments de sommes de variables aléatoires indépendantes (dont les fonctions génératrices sont obtenues comme produits des fonctions génératrices individuelles).

3.10 • SUITES DE V.A. ET NOTIONS DE CONVERGENCE

Il existe différentes façons de définir la notion de convergence de suites de variables aléatoires. Nous les définissons brièvement ci-dessous en indiquant les relations qui existent entre ces notions. Ces notions sont importantes pour étudier les estimateurs définis en statistiques et pour caractériser les processus aléatoires. Elles sont introduites ici car elles interviennent dans la formulation des théorèmes de convergence de la section 3.11.

3.10.1 Convergence en probabilité

La suite $(\mathcal{X}_n)$ de v.a. réelles **converge en probabilité** vers la constante a; si $\forall\epsilon$ et η (arbitrairement petits), $\exists n_0$ tel que $n > n_0$ entraîne

$$P(|\mathcal{X}_n - a| > \epsilon) < \eta. \tag{3.53}$$

On note alors $(\mathcal{X}_n) \xrightarrow{P} a$.

On définit la convergence en probabilité d'une suite de v.a. $(\mathcal{X}_n)$ vers une v.a. $\mathcal{X}$ comme la convergence vers la constante 0 de la suite $(\mathcal{X}_n - \mathcal{X})$.

3.10.2 Convergence presque sûre ou convergence forte

La suite $(\mathcal{X}_n)$ de v.a. réelles **converge presque sûrement** vers $\mathcal{X}$ si :

$$P(\{\omega | \lim_{n\to\infty} \mathcal{X}_n(\omega) \neq \mathcal{X}(\omega)\}) = 0. \tag{3.54}$$

On note alors $(\mathcal{X}_n) \xrightarrow{p.s.} \mathcal{X}$.

La convergence presque sûre implique la convergence en probabilité, c'est pourquoi on l'appelle aussi convergence forte.

3.10.3 Convergence en moyenne d'ordre p

Si $E\{(\mathcal{X}_n - \mathcal{X})^p\}$ existe $\forall n$, alors on a
$(\mathcal{X}_n) \to \mathcal{X}$ **en moyenne d'ordre** p si $E\{(\mathcal{X}_n - \mathcal{X})^p\} \to 0$.

Le cas pratique usuel est la moyenne quadratique ($p = 2$).

La convergence en moyenne d'ordre p implique la convergence en probabilité.

3.10.4 Convergence en loi

La suite $(\mathcal{X}_n)$ de v.a. réelles **converge en loi** vers $\mathcal{X}$ de fonction de répartition $F(\cdot)$ si pour tout point de continuité x de $F(\cdot)$, la suite $(F_n(x))$ converge vers $F(x)$. On note

$$(\mathcal{X}_n) \xrightarrow{\mathcal{L}} \mathcal{X}. \tag{3.55}$$

Il s'agit de la convergence la plus faible. En particulier, la convergence en probabilité implique la convergence en loi. Cette dernière est très utilisée en pratique car elle permet d'approximer la fonction de répartition de $(\mathcal{X}_n)$ par celle de $\mathcal{X}$, et réciproquement.

On montre que si $F(\cdot)$ est continue alors la convergence est uniforme (il s'agit d'un comportement plus régulier que la convergence ponctuelle, vu au cours d'analyse mathématique). De plus, si les $F_n(\cdot)$ admettent des densités alors la convergence en loi implique la convergence ponctuelle des densités.

3.11 THEOREMES DE CONVERGENCE

3.11.1 Théorème de Moivre-Laplace

Il dit que, si $(\mathcal{X}_n)$ forme une suite de v.a. binomiales $\mathcal{B}(n, p)$, alors

$$\frac{\mathcal{X}_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \tag{3.56}$$

Ce théorème utile en statistiques, permet d'approximer une loi binomiale par une loi Gaussienne.

3.11.2 Théorème central-limite

Il dit que, si $(\mathcal{X}_n)$ forme une suite de v.a. i.i.d. de moyenne μ et d'écart-type σ (ces deux moments sont donc supposés exister), alors

$$\left(\frac{\sum_{i=1}^{n}\mathcal{X}_i - n\mu}{\sigma\sqrt{n}}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{3.57}$$

Ce théorème établit la convergence en loi vers la loi normale d'une somme de v.a. *indépendantes, et identiquement distribuées* (nous dirons "i.i.d.") sous des hypothèses très peu contraignantes.

On retrouve comme cas particulier le théorème de Moivre-Laplace, en prenant des variables de Bernouilli.

Contre-exemple : loi de Cauchy.

3.11.3 Lois des grands nombres

3.11.3.1 Loi faible des grands nombres

Soient $\mathcal{X}_i, \forall i = 1, \ldots n$ indépendantes d'espérance μ_i finies et de variances σ_i finies, alors
Si $\frac{1}{n}\sum_{i=1}^{n}\mu_i \to \mu$ *et* $\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \to 0$, *alors* $\overline{\mathcal{X}}_n \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathcal{X}_i$ *est telle que*

$$(\overline{\mathcal{X}}_n) \xrightarrow{P} \mu. \tag{3.58}$$

Cas particulier : les v.a. $\mathcal{X}_i$ sont i.i.d. μ, σ. On a alors $\frac{1}{n}\sum_{i=1}^{n}\mu_i = \mu$ et $\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 = \frac{\sigma^2}{n}$.

3.11.3.2 Loi forte des grands nombres

Soient $\mathcal{X}_i, \forall i = 1, \ldots n$ indépendantes d'espérance μ_i finies et de variances σ_i finies, *si de plus* $\sum_{i=1}^{n}\frac{\sigma_i^2}{i^2} \to a$, *alors*

$$(\overline{\mathcal{X}}_n) \xrightarrow{p.s.} \mu. \tag{3.59}$$

Cas particulier : les v.a. $\mathcal{X}_i$ sont i.i.d. μ, σ. On a alors $\frac{1}{n}\sum_{i=1}^{n}\mu_i = \mu$ et $\sum_{i=1}^{n}\frac{{\sigma_i}^2}{i^2} = \sigma^2\sum_{i=1}^{n}\frac{1}{i^2}$, qui converge.

Remarque. Les deux lois des grands nombres que nous avons formulées ci-dessus (avec leurs hypothèses) correspondent à la formulation habituelle de ces lois, telles qu'on les trouve dans les ouvrages introductifs. Cependant, des formulations alternatives, reposant sur des hypothèses différentes (ni plus générales ni plus spécifiques) existent également. En particulier, si les variables $\mathcal{X}_n$ sont i.i.d. (cas très fréquent en statistiques), on montre que la loi forte (et donc aussi la loi faible)) ne repose pas sur l'existence des moments du second ordre, et que l'existence du premier moment (espérance finie) est dans ce cas une condition non seulement suffisante mais aussi nécessaire de convergence (version "Kolmogorov" de la loi forte des grands nombres). Par ailleurs, il existe également des versions de ces lois qui concernent des suites de variables non-nécessairement indépendantes; leur étude fait l'objet du large domaine des processus stochastiques et repose sur la notion d'ergodicité, qui a permis de développer de nombreux algorithmes modernes de traitement de l'information, notamment dans le domaine de la compression de données et des codes correcteurs d'erreurs, et des techniques de sondage intelligents. Ces notions, qui font l'objet d'enseignements plus avancés, seront évoquées au chapitre 5.

3.12 PROBLÈMES ET APPLICATIONS

3.12.1 Problèmes d'ingénieurs faisant appel aux notions introduites dans ce chapitre

Le problèmes d'ingénieurs que nous discutons ci-dessous sont des problèmes de *prédiction* discutés à la section 1.3.1. Cependant, les notions vues dans le présent chapitre sont aussi intensivement utilisées en statistique pour résoudre des problèmes de traitement de données rencontrés dans les applications; nous laissons le soin au cours de statistique de les mettre en oeuvre dans ce contexte.

3.12.1.1 Evaluation de la fiabilité d'un système technique

Beaucoup de problèmes d'ingénieurs impliquent la détermination des propriétés d'une ou de plusieurs variables, qui sont des fonctions plus ou moins compliquées d'autres variables.

Par exemple, considérons un système mécanique, disons un panneau routier, dont on veut déterminer les déplacements possibles suite à des perturbations extérieures, disons la force appliquée par le vent sur le panneau. En fonction des paramètres de dimensionnement et des matériaux utilisés, les méthodes d'ingénieur nous permettent d'écrire un modèle mathématique qui relie le déplacement mesuré en un point du panneau, dans une direction donnée, en fonction de la force du vent appliquée selon une autre direction. De façon générique, on peut écrire ce modèle sous la forme

$$y = f(x, \lambda_1, \ldots, \lambda_K), \tag{3.60}$$

où x représente la variable d'entrée (ici la force du vent), les λ_i sont des paramètres constants qui décrivent la structure et son matériau, et y désigne la variable de sortie, ici le déplacement qu'on veut surveiller.

Le système devant être dimensionné pour une gamme de sollicitations prescrite, on cherchera à déterminer si compte tenu de cette gamme, disons $\forall x \in X$, on peut garantir que $y \in Y_{\text{tol}}$, où Y_{tol} décrit l'ensemble des déplacements tolérables (typiquement un intervalle, lorsque $y \in \mathbb{R}$ comme dans notre exemple.

Sous certaines conditions, (par exemple si la fonction f est monotone, et lorsque X est borné) il est possible de déterminer de façon numérique ou analytique des bornes sur y, et de vérifier ainsi si le système répond aux spécifications. Le processus de conception du système pourrait alors viser à choisir la combinaison "la moins coûteuse" des paramètres λ_i telle que cette condition soit respectée (ce qui se traduit en un problème d'optimisation cherchant à trouver la valeur minimale du coût $h(\lambda_1, \ldots, \lambda_K)$, dont la résolution fait l'objet d'autres enseignements).

Cette approche déterministe souffre cependant de certaines difficultés intrinsèques. En effet, soit on choisit de définir l'ensemble X suffisamment vaste pour couvrir avec une très grande probabilité l'ensemble des sollicitations possibles, et le design conduira à une structure très solide mais économiquement très coûteuse, soit on choisit X trop petit, et la structure peut présenter un risque élevé de ne pas survivre fort longtemps. Afin d'évaluer ces risques, et de finalement pouvoir faire un choix de design offrant le bon rapport coût/fiabilité, il est nécessaire de disposer d'une information supplémentaire sur la nature des perturbations, à savoir leur loi de probabilité $P_{\mathcal{X}}$, et de combiner cette information avec le modèle pour en déduire la loi $P_{\mathcal{Y}}$ régissant les effets sur le système.

En pratique, pour une valeur donnée des paramètres λ_i et pour une distribution des perturbations (donnée sous la forme $P_{\mathcal{X}}$, $F_{\mathcal{X}}$ ou $f_{\mathcal{X}}$), on cherchera alors à déterminer la valeur de $P_{\mathcal{Y}}(y \in Y_{\text{tol}})$ et si cette valeur est suffisamment proche de 1 (disons ≥ 0.99999), on acceptera le design. Les méthodes de ce chapitre permettent en principe d'évaluer ces probabilités, soit de façon analytique soit de façon numérique.

3.12.1.2 Evaluation du coût d'exploitation d'un système

Contrairement au cas de notre panneau dont le seul coût est un coût de conception (et d'installation), de nombreux systèmes techniques sont également sujet à des coûts d'exploitation. Par exemple une voiture coûte à l'achat puis à l'utilisation; il en est de même pour une usine de production sidérurgique, pour les routes et les bâtiments, et même pour les systèmes informatiques. Souvent il est nécessaire d'arbitrer entre coût d'investissement et coût d'exploitation, par exemple acheter une voiture neuve coûte plus cher que d'acheter une occasion, mais généralement se traduit par des coûts d'exploitation plus faibles.

Lors de l'investissement, il est rarement possible de déterminer de façon précise comment le système sera utilisé au cours de sa période d'exploitation (souvent plusieurs années, et pour les investissements lourds souvent plusieurs dizaines d'années).

L'approche stochastique consiste alors à modéliser les conditions futures d'exploitation par une (ou plusieurs) variables aléatoires $\mathcal{X}$, et à utiliser un modèle du système, sous la forme

$$y = g(x, \lambda_1, \ldots, \lambda_K), \tag{3.61}$$

où la sortie représente cette fois le coût d'exploitation associé aux conditions d'exploitation x (qui représentent ici les entrées), où les λ_i représentent toujours les paramètres de conception.

On cherche ensuite à déterminer la loi de $\mathcal{Y}$, et notamment son espérance mathématique $\mu_{\mathcal{Y}} = E\{\mathcal{Y}\}$ et son écart-type $\sigma_{\mathcal{Y}}$, qui représentent de manière compacte la distribution des coûts d'exploitation.

Ayant ces grandeurs, on peut évaluer le coût total $c(\lambda_1, \ldots, \lambda_K) = h(\lambda_1, \ldots, \lambda_K) + \mu_{\mathcal{Y}}(\lambda_1, \ldots, \lambda_K)$ afin de faire le meilleur choix d'investissement, compte tenu du coût d'exploitation moyen. La valeur de l'écart-type permet d'évaluer dans quelle mesure on peut se fier à cette estimation, compte tenu de la variabilité des conditions d'exploitation. En particulier, si cet écart-type est suffisamment faible on aura bonne confiance dans notre choix optimal, sinon on peut éventuellement essayer de minimiser un critère numérique qui réalisera un autre compromis entre coût moyen et risque de surcoût.

3.12.1.3 Discussion

Les deux problèmes génériques décrits ci-dessus se compliquent dans la pratique par le fait que généralement les systèmes étudiés sont complexes et sont influencés par de nombreux facteurs aléatoires. Les méthodes des 2 chapitres suivants visent précisément à aborder ce genre de situation de manière systématique.

3.12.2 Méthode de Monte-Carlo

De nombreux problèmes d'ingénieurs se traduisent par l'évaluation d'une intégrale d'une fonction plus où moins compliquée. Les méthodes analytiques (vues aux cours d'analyse) et numériques (vues au cours d'analyse numérique) permettent d'en calculer de nombreuses de manière efficace. Ces méthodes sont cependant inutilisables dans certaines situations, en particulier lorsque il s'agit d'une intégrale multiple d'une fonction dépendant d'un grand nombre de variables et qui ne se factorise pas sous la forme de produits simples.

3.12.2.1 Evaluation d'une intégrale simple

Nous reviendrons sur la généralisation de la méthode de Monte-Carlo pour l'intégration multiple au chapitre suivant; pour le moment contentons-nous d'en expliquer le principe pour calculer une intégrale simple de la forme

$$I_g = \int_0^1 g(x)\,\mathrm{d}x. \tag{3.62}$$

Remarquons que toute intégrale simple peut toujours être ramenée, via un changement de variables, à une intégrale sur l'intervalle $[0, 1]$.

Remarquons ensuite que le calcul de (3.62) est équivalent à la détermination de l'espérance mathématique d'une variable aléatoire $\mathcal{Y} = g(\mathcal{X})$ où $\mathcal{X} \sim \mathcal{U}[0, 1]$.

Si nous supposons que $\sigma_{\mathcal{Y}}$ est finie (c'est le cas, si la fonction $g(x)$ est continue), et que nous disposons d'un échantillon i.i.d. de valeurs y_i de $\mathcal{Y}$, nous pouvons estimer cette espérance par

$$\hat{I}_g = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{3.63}$$

l'écart-type de l'estimateur étant donné par

$$\sigma_{\hat{I}_g} = \frac{1}{\sqrt{n}}\sigma_{\mathcal{Y}}. \tag{3.64}$$

En pratique, on peut générer un échantillon de valeurs de $\mathcal{Y}$ en utilisant un générateur de nombres aléatoires $x_i \sim \mathcal{U}[0,1]$, et en calculant $y_i = g(x_i)$.

Remarquons que la variance de l'estimateur est proportionnelle à la variance de $\mathcal{Y}$; le nombre d'échantillons nécessaires pour atteindre une certaine précision croit linéairement avec cette variance.

3.12.2.2 Echantillonnage d'importance

Supposons que nous soyons en mesure de générer un échantillon de valeurs x_i distribuées selon une certaine densité $f_{\mathcal{X}}$ définie sur $[0,1]$ et non nécessairement uniforme, et soit h une fonction. La grandeur

$$\hat{I}_h^{f_{\mathcal{X}}} = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

est alors un estimateur de l'intégrale

$$I_h^{f_{\mathcal{X}}} = \int_0^1 h(x) f_{\mathcal{X}}(x)\,\mathrm{d}x.$$

dont la variance est cette fois liée à la variance de la variable $\mathcal{Z} = h(\mathcal{X})$.

Si nous prenons $h(x) = \frac{g(x)}{f_{\mathcal{X}}(x)}$ nous obtenons un nouvel estimateur de l'intégrale (3.62), donné par

$$\hat{I}_{g/f_{\mathcal{X}}}^{f_{\mathcal{X}}} = \frac{1}{n}\sum_{i=1}^{n} \frac{g(x_i)}{f_{\mathcal{X}}(x_i)},$$

dont la variance sera d'autant plus faible que le rapport $\frac{g(x)}{f_{\mathcal{X}}(x)}$ est proche d'une constante.

Notons que dans le cas idéal où ce rapport serait constant, nous aurions

$$\frac{g(x)}{f_{\mathcal{X}}(x)} = I_g,$$

puisque $\int_0^1 f_{\mathcal{X}}(x)\,\mathrm{d}x = 1$, ce qui veut dire qu'on connait déjà la valeur de I_g dès lors qu'on connait la fonction idéale d'échantillonnage d'importance.

En pratique, on choisira une densité d'échantillonnage d'importance dont l'allure est voisine de celle de la fonction qu'on souhaite intégrer.

3.12.2.3 Variable de contrôle

Supposons que nous disposions d'une fonction $h(x)$ dont nous connaissons déjà l'intégrale I_h avec grande précision. Nous pouvons alors calculer I_g par

$$I_g = I_h + \int_0^1 (g(x) - h(x))\,\mathrm{d}x,$$

Si $g(x) - h(x)$ varie peu par rapport à h sur l'intervalle d'intégration, on pourra appliquer la méthode de Monte-Carlo à $g - h$, ce qui nécessitera un nombre plus faible d'échantillons, à précision égale, mais une double évaluation de fonction pour chaque échantillon.

Cette méthode peut se révéler très intéressante dans les problèmes d'ingénieurs (tels ceux mentionnés ci-dessus) où on est amené à recalculer des intégrales d'une même fonction pour différentes valeurs de paramètres. Plus généralement, elle permet de tirer profit de modèles approchés plus simples à manipuler analytiquement, tels que des approximations linéaires, et de réserver le travail de la méthode de Monte-Carlo pour estimer les corrections à apporter aux valeurs approchées calculées à partir de ces modèles.

3.12.2.4 Discussion

La méthode de Monte-Carlo, dont nous n'avons fait qu'effleurer les nombreuses facettes, est devenue un outil incontournable dans les diverses disciplines des ingénieurs. Dans ce contexte, elle offre en effet des approches

versatiles pour exploiter les puissances de calcul dans le domaine de l'évaluation et de la conception des systèmes techniques. Notons en particulier que les évaluations des fonctions g, h etc, peuvent être effectuées de façon parallèle en faisant appel aux grilles de calcul de plus en plus puissantes à notre disposition.

Nous reviendrons sur cette méthode à la fin du chapitre suivant.

Notes

1. On peut définir l'ensemble Ω de beaucoup de manières différentes pour une expérience donnée du monde réel, cependant la définition et l'étude d'un ensemble de variables aléatoires nécessite en principe de fixer ce choix une fois pour toutes. Voir cependant à ce sujet la discussion de la section 4.4, à la fin du chapitre suivant.
2. Les mêmes restrictions s'appliquent à $\mathcal{E}_\mathcal{X}$ que celles que nous avons évoquées au chapitre 2 en ce qui concerne la structure de $\mathcal{E}$ définie sur Ω. Par exemple, si $\Omega_\mathcal{X}$ est égal à la droite réelle $\mathbb{R}$ on sera amené à considérer la tribu boréllienne plutôt que $2^{\Omega_\mathcal{X}}$ comme σ-algèbre $\mathcal{E}_\mathcal{X}$. Cependant, cette condition n'est pas réellement restrictive : si Ω est au plus dénombrable il en est de même pour $\Omega_\mathcal{X}$, et sinon, la condition de mesurabilité de la variable aléatoire implique de toutes façons que sa σ-algèbre n'est pas plus fine que celle de Ω.
3. Notons que si l'espace Ω est fini ou dénombrable, alors toute variable aléatoire est nécessairement discrète.
4. Le terme de "densité" est le terme consacré en théorie de la mesure, pour indiquer une fonction qui permet de calculer la valeur de la mesure de tout ensemble mesurable par "intégration" d'une fonction le long d'une autre mesure. On dit qu'une mesure est absolument continue par rapport à une autre mesure, s'il existe une densité qui permet de calculer la première mesure par intégration de cette densité selon la seconde mesure. De ce point de vue la densité d'une variable discrète, est la "densité" par rapport à la mesure de "comptage", c'est-à-dire la mesure qui compte simplement le nombre d'éléments d'un ensemble.
5. Pour une variable continue, la mesure induite sur $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ est absolument continue par rapport à la mesure de Lebesgue (celle qui définit la longueur des intervalles), et il existe donc une densité qui permet de calculer la probabilité d'un intervalle par intégration au sens de Lebesgue de cette densité sur l'intervalle.
6. En toute généralité, on peut montrer que toute fonction de répartition peut se décomposer en une somme de trois termes ($F(x) = F_c(x) + F_d(x) + F_s(x)$) tels que F_c soit absolument continue (continue et dérivable), F_d est discrète, et F_s (composante singulière) est continue mais ne possède pas de dérivée. Nous supposons que $F_s = 0$.
7. Une forme réciproque de ce théorème s'énonce comme suit : si toutes les fonctions mesurables $\phi_\mathcal{X}$ et $\phi_\mathcal{Y}$ de respectivement $\mathcal{X}$ et $\mathcal{Y}$ sont de covariance nulle alors les variables $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes.
8. La fonction $1(\cdot)$ vaut 1 si son argument est vrai, sinon elle vaut 0.
9. Cette notion peut être généralisée au moyen de la formule $\phi_\mathcal{X}(t) = \int_\mathbb{R} e^{itx}\,\mathrm{d}P_\mathcal{X}(x)$.
10. L'intégrale converge pour tout $t \in \mathbb{R}$ et toute densité $f_\mathcal{X}$, étant donné le caractère intégrable de $f_\mathcal{X}$ et le fait que $|e^{itu}| = 1$.

4 ENSEMBLES DE VARIABLES ALÉATOIRES ET CONDITIONNEMENT

Dans ce chapitre nous abordons l'étude simultanée de plusieurs variables aléatoires dépendantes. Nous commençons par considérer l'étude de couples de variables aléatoires, et des opérations de conditionnement. Ensuite nous présentons une formalisation mathématique de l'ensemble de variables aléatoires à valeurs réelles pouvant être définies dans le cadre d'un problème, en montrant que cet ensemble possède une structure géométrique analogue à celle de l'espace euclidien $\mathbb{R}^n$. Enfin, nous étudions les ensembles de variables aléatoires sous l'angle de la modélisation de leur loi de probabilité conjointe, et montrons dans ce contexte l'importance de la notion d'indépendance conditionnelle.

4.1 COUPLES DE V.A. DISCRÈTES ET CONDITIONNEMENT

Nous étudions dans cette section des couples de variables aléatoires discrètes définies sur un même espace de probabilité $(\Omega, \mathcal{E}, P)$ et non nécessairement indépendantes.

Nous étudions ici les couples $(\mathcal{X}, \mathcal{Y})$ de v.a. tels que $\mathcal{X}$ et $\mathcal{Y}$ prennent leurs valeurs dans un ensemble fini désigné respectivement par $\Omega_{\mathcal{X}} = \{x_1, \ldots, x_k\}$ et $\Omega_{\mathcal{Y}} = \{y_1, \ldots, y_l\}$ munis de leur $\sigma-$algèbre maximale. Dans ce cas, le couple prend ses valeurs dans $\Omega_{\mathcal{X},\mathcal{Y}} = \Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$ muni de la $\sigma-$algèbre produit, qui est également maximale. On doit supposer que la fonction ainsi induite de Ω dans $\Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$ est bien $\mathcal{E}$-mesurable, ce qui est le cas si et seulement si les deux fonctions $\mathcal{X}$ et $\mathcal{Y}$ sont elles-mêmes $\mathcal{E}$-mesurables.

Toutes les notions définies dans cette section peuvent cependant se généraliser au cas où les variables aléatoires prennent un nombre infini mais au plus dénombrable de valeurs différentes, à condition de prendre la précaution de supposer que les séries infinies définissant les espérances et les variances convergent absolument.

4.1.1 Cas où les variables aléatoires sont à valeurs quelconques

4.1.1.1 Loi (con)jointe

La loi de probabilité conjointe $P_{\mathcal{X},\mathcal{Y}}$ du couple $(\mathcal{X}, \mathcal{Y})$ est déterminée complètement par la connaissance des kl nombres

$$P_{\mathcal{X},\mathcal{Y}}(x_i, y_j) \stackrel{\triangle}{=} P([\mathcal{X} = x_i] \wedge [\mathcal{Y} = y_j]), \forall i = 1, \ldots, k, \forall j = 1, \ldots, l. \tag{4.1}$$

$P_{\mathcal{X},\mathcal{Y}}(x_i, y_j)$ désigne donc la probabilité de l'événement $\{\omega \in \Omega : \mathcal{X}(\omega) = x_i \text{ et } \mathcal{Y}(\omega) = y_j\}$. Une fois

	y_1	$\cdots$	y_j	$\cdots$	y_l	$\sum$
x_1						
$\vdots$			$\vdots$			
x_i		$\cdots$	$P_{\mathcal{X},\mathcal{Y}}(x_i,y_j)$	$\cdots$		$P_{\mathcal{X}}(x_i)$
$\vdots$			$\vdots$			
x_k						
$\sum$			$P_{\mathcal{Y}}(y_j)$			1

Figure 4.1: Table de contingence

qu'on connait la loi $P_{\mathcal{X},\mathcal{Y}}(x_i,y_j)$ pour toutes les valeurs de ses arguments, on peut en déduire la probabilité de tout événement qui peut se décrire par une affirmation logique faisant intervenir les valeurs des deux variables aléatoires.

On a bien entendu que $\sum_{i=1}^{k}\sum_{j=1}^{l} P_{\mathcal{X},\mathcal{Y}}(x_i,y_j)=1$.

4.1.1.2 Lois marginales

La loi de $\mathcal{X}$ est obtenue à partir de la loi jointe par l'opération de **marginalisation** :

$$P_{\mathcal{X}}(x_i) \triangleq P(\mathcal{X}=x_i)=\sum_{j=1}^{l} P_{\mathcal{X},\mathcal{Y}}(x_i,y_j), \forall i=1,\dots k. \tag{4.2}$$

De même, la loi de $\mathcal{Y}$ est obtenue par

$$P_{\mathcal{Y}}(y_j) \triangleq P(\mathcal{Y}=y_j)=\sum_{i=1}^{k} P_{\mathcal{X},\mathcal{Y}}(x_i,y_j), \forall j=1,\dots l. \tag{4.3}$$

C'est pour cette raison, qu'on désigne souvent ces lois par le terme de **lois marginales**.

On représente souvent un couple de v.a. à l'aide d'une *table de contingences*, comme illustré à la Figure 4.1.

4.1.1.3 Lois conditionnelles

La loi conditionnelle de $\mathcal{X}$ connaissant $\mathcal{Y}$ est définie par

$$P_{\mathcal{X}|\mathcal{Y}}(x_i|y_j) \triangleq P(\mathcal{X}=x_i|\mathcal{Y}=y_j)=\frac{P_{\mathcal{X},\mathcal{Y}}(x_i,y_j)}{P_{\mathcal{Y}}(y_j)}, \forall i=1,\dots k, \forall j=1,\dots l. \tag{4.4}$$

Cette loi conditionnelle n'est ainsi définie que pour les valeurs de y_j de probabilité marginale non-nulle. Cependant, la valeur précise qu'on lui attribue lorsque $P_{\mathcal{Y}}(y_j)=0$ (auquel cas on a aussi $P_{\mathcal{X},\mathcal{Y}}(x_i,y_j)=0$) n'a pas réellement d'importance ici. Nous analyserons de plus près à la section 4.4 cette question.

La loi conditionnelle de $\mathcal{Y}$ connaissant $\mathcal{X}$ est définie de façon analogue par

$$P_{\mathcal{Y}|\mathcal{X}}(y_j|x_i) \triangleq P(\mathcal{Y}=y_j|\mathcal{X}=x_i)=\frac{P_{\mathcal{X},\mathcal{Y}}(x_i,y_j)}{P_{\mathcal{X}}(x_i)}, \forall i=1,\dots k, \forall j=1,\dots l. \tag{4.5}$$

On a donc $P_{\mathcal{X},\mathcal{Y}}(x_i,y_j)=P_{\mathcal{X}|\mathcal{Y}}(x_i|y_j)P_{\mathcal{Y}}(y_j)=P_{\mathcal{Y}|\mathcal{X}}(y_j|x_i)P_{\mathcal{X}}(x_i), \forall i=1,\dots k, \forall j=1,\dots l$, ce que nous écrivons de façon plus synthétique par

Factorisation de la loi jointe

$$P_{\mathcal{X},\mathcal{Y}} = P_{\mathcal{X}|\mathcal{Y}} P_{\mathcal{Y}} = P_{\mathcal{Y}|\mathcal{X}} P_{\mathcal{X}}.$$

Remarquons que lorsque $\mathcal{X} \perp \mathcal{Y}$, on a $P_{\mathcal{X}|\mathcal{Y}} = P_{\mathcal{X}}$, $P_{\mathcal{Y}|\mathcal{X}} = P_{\mathcal{Y}}$, et $P_{\mathcal{X},\mathcal{Y}} = P_{\mathcal{X}} P_{\mathcal{Y}}$.

Cas particulier: une des variables est une fonction de l'autre. Supposons que $\mathcal{Y}$ est une fonction de la variable $\mathcal{X}$. La table de contingences comporte alors sur chaque ligne x_i une seule entrée de probabilité non-nulle, correspondant à la valeur $y = \phi(x_i)$. On a donc $\forall i$

$$P_{\mathcal{X},\mathcal{Y}}(x_i, \phi(x_i)) = P_{\mathcal{X}}(x_i) \quad (4.6)$$
$$P_{\mathcal{Y}|\mathcal{X}}(\phi(x_i)|x_i) = 1, \quad (4.7)$$

et $\forall y_j \neq \phi(x_i)$

$$P_{\mathcal{X},\mathcal{Y}}(x_i, y_j) = 0 \quad (4.8)$$
$$P_{\mathcal{Y}|\mathcal{X}}(y_j|x_i) = 0. \quad (4.9)$$

Si à la fois $\mathcal{Y}$ est une fonction de $\mathcal{X}$ et $\mathcal{X}$ est une fonction de $\mathcal{Y}$, alors la table de contingences est forcément carrée et comporte une seule entrée non-nulle sur chaque ligne et chaque colonne. Un réarrangement de l'ordre des lignes (valeurs de $\mathcal{X}$) ou des colonnes (valeurs de $\mathcal{Y}$) donne alors une table diagonale. La connaissance de la valeur l'une des variables détermine la valeur de l'autre, et réciproquement. Tout se passe comme si les deux variables aléatoires étaient identiques : elles induisent la même $\sigma-$algèbre sur Ω.

4.1.2 Cas où $\mathcal{Y}$ est réelle et $\mathcal{X}$ quelconque

Supposons que $\mathcal{Y}$ soit une v.a. discrète finie réelle et $\mathcal{X}$ une variable aléatoire discrète finie quelconque.

4.1.2.1 Espérance conditionnelle : définition et propriétés

Considérons la variable $\mathcal{Z} = \phi(\mathcal{X})$ définie par

$$z(x) \stackrel{\triangle}{=} \sum_{j=1}^{l} y_j P_{\mathcal{Y}|\mathcal{X}}(y_j|x). \quad (4.10)$$

Il s'agit d'une variable aléatoire réelle discrète dont le nombre de valeurs différentes est au plus égal au nombre de valeurs différentes de $\mathcal{X}$. Pour une valeur fixée x_i de $\mathcal{X}$, la valeur $z(x_i)$ est l'espérance de la variable $\mathcal{Y}$ calculée par rapport la loi conditionnelle $P(\cdot|X_i)$ induite sur Ω par l'événement $X_i = \{\omega \in \Omega : \mathcal{X}(\omega) = x_i\}$.

Cette expression définit la valeur de l'espérance conditionnelle, $E\{\mathcal{Y}|x\}$, de $\mathcal{Y}$ étant donné que $\mathcal{X} = x$.

Espérance conditionnelle

$$E\{\mathcal{Y}|x\} \stackrel{\triangle}{=} \sum_{j=1}^{l} y_j P_{\mathcal{Y}|\mathcal{X}}(y_j|x). \quad (4.11)$$

$E\{\mathcal{Y}|x\}$ est donc une fonction réelle de la variable aléatoire $\mathcal{X}$ définie sur $(\Omega, \mathcal{E}, P)$, aussi appellée **fonction de régression** de $\mathcal{Y}$ en x. Comme $\mathcal{X}$ est une v.a. cette fonction définit aussi une v.a. réelle sur $(\Omega, \mathcal{E}, P)$.

Cette variable aléatoire est notée par $E\{\mathcal{Y}|\mathcal{X}\}$ et présente des propriétés remarquables. Ces propriétés se démontrent aisément dans le cas "discret-fini" que nous considérons dans cette section, et elles restent valides, sous-réserve de l'existence des notions impliquées, dans les cas plus généraux discutés dans la suite.

Linéarité de l'espérance conditionnelle
Soit $\mathcal{Y} = \alpha\mathcal{Y}_1 + \beta\mathcal{Y}_2$ une combinaison linéaire de deux variables aléatoires discrètes finies et réelles, et $\mathcal{X}$ une variable aléatoire discrète finie quelconque, définies sur $(\Omega, \mathcal{E}, P)$. On a

$$E\{\mathcal{Y}|\mathcal{X}\} = \alpha E\{\mathcal{Y}_1|\mathcal{X}\} + \beta E\{\mathcal{Y}_2|\mathcal{X}\}. \tag{4.12}$$

La démonstration est assez immédiate, en exploitant la linéarité de la notion de base d'espérance mathématique.

Dans le contexte de cette section, on a la version suivante du théorème de l'espérance totale, que nous avons introduit au chapitre précédent.

Théorème de l'espérance totale (ou de la moyenne conditionnelle)
Soit $\mathcal{X}$ une variable aléatoire discrète finie quelconque et $\mathcal{Y}$ une variable aléatoire discrète finie à valeurs réelles, définies sur $(\Omega, \mathcal{E}, P)$. On a

$$E\{E\{\mathcal{Y}|\mathcal{X}\}\} = E\{\mathcal{Y}\}. \tag{4.13}$$

En effet, on peut calculer l'espérance mathématique de la variable $E\{\mathcal{Y}|\mathcal{X}\}$ de la façon suivante :

$$\begin{aligned} E\{E\{\mathcal{Y}|\mathcal{X}\}\} &= \sum_{i=1}^{k} E\{\mathcal{Y}|x_i\} P_{\mathcal{X}}(x_i) \\ &= \sum_{i=1}^{k} P_{\mathcal{X}}(x_i) \sum_{j=1}^{l} y_j P_{\mathcal{Y}|\mathcal{X}}(y_j|x_i) \\ &= \sum_{j=1}^{l} y_j \sum_{i=1}^{k} P_{\mathcal{X}}(x_i) P_{\mathcal{Y}|\mathcal{X}}(y_j|x_i) = \sum_{j=1}^{l} y_j P_{\mathcal{Y}}(y_j) = E\{\mathcal{Y}\}. \end{aligned}$$

Lorsque les deux variables sont indépendantes, on a la propriété suivante :

Indépendance $\Rightarrow$ espérance conditionnelle constante
Soit $\mathcal{X}$ une variable aléatoire discrète finie quelconque et $\mathcal{Y}$ une variable aléatoire discrète finie à valeurs réelles, définies sur $(\Omega, \mathcal{E}, P)$.
Si $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes, alors l'espérance conditionnelle est constante, et $E\{\mathcal{Y}|x\} = E\{\mathcal{Y}\}, \forall x$.
La réciproque de cette propriété n'est pas vraie.

En effet, dans le second membre de l'équation (4.10), on peut alors remplacer $P_{\mathcal{Y}|\mathcal{X}}(y_j|x)$ par $P_{\mathcal{Y}}(y_j)$, et on obtient bien $z(x_i) = E\{\mathcal{Y}\}, \forall i = 1, \ldots, k$.

Lorsque $\mathcal{Y}$ est une fonction de $\mathcal{X}$ on a la propriété suivante.

Espérance conditionnelle d'une fonction de $\mathcal{X}$
Lorsque $\mathcal{Y}$ est une fonction de $\mathcal{X}$, on a $E\{\mathcal{Y}|\mathcal{X}\} = \mathcal{Y}$. En particulier, on a $E\{\mathcal{Y}|\mathcal{Y}\} = \mathcal{Y}$.

En effet, le second membre de l'équation (4.10) devient alors pour une valeur donnée x_i,

$$E\{\mathcal{Y}|x_i\} = \sum_{j=1}^{l} y_j P_{\mathcal{Y}|\mathcal{X}}(y_j|x_i) = \sum_{j=1}^{l} y_j \delta_{y_j, \phi(x_i)} = \phi(x_i),$$

où δ est le symbole de Kronecker. Autrement dit, $E\{\mathcal{Y}|x_i\} = \phi(x_i)$, et donc $E\{\mathcal{Y}|\mathcal{X}\} = \phi(\mathcal{X}) = \mathcal{Y}$. Ce théorème, dans sa version générale, a une réciproque que nous exprimerons à la section 4.3.

4.1.2.2 Variance conditionnelle : définition et propriétés

On définit, la variance conditionnelle de $\mathcal{Y}$ sachant $\mathcal{X}$ comme suit.

Variance conditionnelle

$$V\{\mathcal{Y}|x\} \triangleq \sum_{j=1}^{l} (y_j - E\{\mathcal{Y}|x\})^2 P_{\mathcal{Y}|\mathcal{X}}(y_j|x). \tag{4.14}$$

Comme pour l'espérance conditionnelle, cette formule définit une nouvelle variable aléatoire, notée $V\{\mathcal{Y}|\mathcal{X}\}$, et qui est une fonction de la variable aléatoire $\mathcal{X}$.

On a le théorème de la variance totale suivant :

Théorème de la variance totale (ou de la variance conditionnelle)

$$V\{\mathcal{Y}\} = E\left\{V\{\mathcal{Y}|\mathcal{X}\}\right\} + V\left\{E\{\mathcal{Y}|\mathcal{X}\}\right\}. \tag{4.15}$$

Nous reviendrons sur ce théorème dans sa version générale à la section 4.3.

Lorsque les deux variables sont indépendantes, on a la propriété suivante :

Indépendance $\Rightarrow$ variance conditionnelle constante
Si $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes, alors la variance conditionnelle est constante, et $V\{\mathcal{Y}|x\} = V\{\mathcal{Y}\}, \forall x$.
La réciproque n'est pas vraie.

Ce résultat s'obtient directement en remplaçant $E\{\mathcal{Y}|x\}$ par $E\{\mathcal{Y}\}$ et $P_{\mathcal{Y}|\mathcal{X}}(y_j|x)$ par $P_{\mathcal{Y}}(y_j)$ dans (4.14).

Enfin, lorsque $\mathcal{Y}$ est une fonction de $\mathcal{X}$ on a la propriété suivante.

Variance conditionnelle d'une fonction de $\mathcal{X}$
Lorsque $\mathcal{Y}$ est une fonction de $\mathcal{X}$, on a $V\{\mathcal{Y}|\mathcal{X}\} = 0_\Omega$, où nous désignons par 0_Ω une variable aléatoire qui est identiquement nulle sur Ω. En particulier $V\{\mathcal{Y}|\mathcal{Y}\} = 0_\Omega$.

En effet, le second membre de l'équation (4.14) devient alors pour une valeur donnée x_i :

$$\begin{aligned} V\{\mathcal{Y}|x_i\} &\triangleq \sum_{j=1}^{l} (y_j - E\{\mathcal{Y}|x_i\})^2 P_{\mathcal{Y}|\mathcal{X}}(y_j|x_i) \\ &= \sum_{j=1}^{l} (y_j - \phi(x_i))^2 P_{\mathcal{Y}|\mathcal{X}}(y_j|x_i) \\ &= \sum_{j=1}^{l} (y_j - \phi(x_i))^2 \delta_{y_j,\phi(x_i)} = 0. \end{aligned}$$

4.1.3 Cas où $\mathcal{X}$ et $\mathcal{Y}$ sont à valeurs réelles

Lorsque les deux variables sont réelles, on définit aussi leur fonction de répartition *conjointe* $F_{\mathcal{X},\mathcal{Y}}$ sur $\mathbb{R}^2$ par

$$F_{\mathcal{X},\mathcal{Y}}(x,y) \triangleq \sum_{x_i<x} \sum_{y_j<y} P_{\mathcal{X},\mathcal{Y}}(x_i,y_j). \tag{4.16}$$

On retrouve les fonctions de répartitions marginales par

$$F_{\mathcal{X}}(x) = F_{\mathcal{X},\mathcal{Y}}(x,+\infty), \tag{4.17}$$
$$F_{\mathcal{Y}}(y) = F_{\mathcal{X},\mathcal{Y}}(+\infty,y). \tag{4.18}$$

Si $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes, alors $F_{\mathcal{X},\mathcal{Y}}(x,y) = F_{\mathcal{X}}(x)F_{\mathcal{Y}}(y), \forall (x,y) \in \mathbb{R}^2$, et la réciproque est vraie.

4.1.4 Espace de variables aléatoires à valeurs réelles définies sur un espace de probabilité fini

A ce stade, il est déjà intéressant de faire une analogie entre la notion de variable aléatoire réelle et la notion de vecteur réel (un point de $\mathbb{R}^n$), analogie sur laquelle nous allons revenir plus en profondeur dans le cadre de la section 4.3.

Supposons que Ω soit fini, et soit $\Omega = \{\omega_1, \ldots, \omega_n\}$ où nous avons choisi un ordre particulier des résultats possibles de l'expérience aléatoire, et supposons aussi que $\forall \omega_i : P(\omega_i) > 0$.

Une variable aléatoire réelle $\mathcal{X}$ est alors définie par les n valeurs, $x_i \triangleq \mathcal{X}(\omega_i) \in \mathbb{R}$, et réciproquement la donnée d'une variable aléatoire réelle définit un point $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Par conséquent, l'espace de variables aléatoires qu'on peut définir sur $(\Omega, 2^\Omega, P)$ est en bijection avec l'espace euclidien $\mathbb{R}^n$; nous noterons cet espace de variables aléatoires par $\mathcal{F}_\Omega$, et nous allons distinguer un élément dans cet ensemble, à savoir la variable aléatoire (constante) qui vaut 1 partout sur Ω; nous la noterons par 1_Ω.

L'espace $\mathbb{R}^n$ est un espace linéaire (sur lequel on peut effectuer des opérations de combinaisons linéaires de vecteurs), et cette structure linéaire se transpose à l'espace de variables aléatoires $\mathcal{F}_\Omega$. On peut aussi définir sur cet espace de variables aléatoires la notion de produit scalaire, par la formule suivante:

Produit scalaire de deux variables aléatoires et orthogonalité

$$\langle \mathcal{X}, \mathcal{Y} \rangle \triangleq \sum_{i=1}^{n} \mathcal{X}(\omega_i)\mathcal{Y}(\omega_i)P(\omega_i). \tag{4.19}$$

On dit que deux variables aléatoires sont **orthogonales** si leur produit scalaire est nul.

Notons qu'on a donc

$$\langle \mathcal{X}, \mathcal{Y} \rangle = E\{\mathcal{X}\mathcal{Y}\}.$$

Ce produit scalaire se réduit au produit scalaire usuel dans $\mathbb{R}^n$ (divisé par n) si la loi P est uniforme. Il est symétrique et compatible avec la structure d'espace linéaire, c'est-à-dire si $\mathcal{Z} = a\mathcal{X}_1 + b\mathcal{X}_2$, avec $a, b \in \mathbb{R}$ on a

$$\langle \mathcal{Z}, \mathcal{Y} \rangle = \langle \mathcal{Y}, \mathcal{Z} \rangle = a\langle \mathcal{X}_1, \mathcal{Y} \rangle + b\langle \mathcal{X}_2, \mathcal{Y} \rangle, \tag{4.20}$$

et pour toute variable aléatoire on a

$$E\{\mathcal{X}\} = \langle \mathcal{X}, 1_\Omega \rangle.$$

Le produit scalaire induit aussi une norme et puis une distance sur l'ensemble des variables aléatoires. Ces notions sont définies comme suit.

Norme d'une variable aléatoire

$$\|\mathcal{X}\| \triangleq \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle} = \sqrt{E\{\mathcal{X}^2\}}, \tag{4.21}$$

Distance de deux variables aléatoires

$$d(\mathcal{X}, \mathcal{Y}) \triangleq \|\mathcal{X} - \mathcal{Y}\|. \tag{4.22}$$

En particulier, deux variables aléatoires sont identiques, si et seulement si leur distance est nulle. On voit également que $\|1_\Omega\| = 1$.

4.1.4.1 Propriétés géométriques de l'espace $\mathcal{F}_\mathcal{X}$

A partir d'une variable aléatoire réelle $\mathcal{X}$, on peut définir l'ensemble des variables aléatoires réelles qui peuvent s'écrire comme une fonction de $\mathcal{X}$; désignons cet ensemble par $\mathcal{F}_\mathcal{X}$.

On a les propriétés géométriques suivantes

- $\mathcal{F}_{1_\Omega}$ est l'ensemble des variables constantes. C'est un sous-espace linéaire de $\mathcal{F}_\Omega$, car il contient l'ensemble de ses combinaisons linéaires (toutes constantes), et en particulier la variable 0_Ω qui est nulle partout sur Ω.

- L'ensemble des variables aléatoires orthogonales à (de produit scalaire nul avec) 1_Ω est identique à l'ensemble de variables aléatoires d'espérance nulle. C'est aussi un sous-espace linéaire de $\mathcal{F}_\Omega$.

- Pour toute variable aléatoire $\mathcal{X}$, l'ensemble de ses fonctions $\mathcal{F}_\mathcal{X}$ est un sous-espace linéaire de $\mathcal{F}_\Omega$, car toute combinaison linéaire de deux fonctions de $\mathcal{X}$ est encore une fonction de $\mathcal{X}$. Cet espace contient 0_Ω, et il contient aussi le sous-espace $\mathcal{F}_{1_\Omega}$, puisque toute v.a. constante est aussi une fonction de $\mathcal{X}$.

- Même si $\mathcal{X}$ est de moyenne nulle, tous les éléments de $\mathcal{F}_\mathcal{X}$ ne sont pas de moyenne nulle, et $\mathcal{F}_\mathcal{X}$ n'est donc pas un sous-espace linéaire orthogonal à $\mathcal{F}_{1_\Omega}$. En fait il contient cet espace comme nous venons de l'indiquer au point précédent.

- Cependant, toute variable aléatoire peut s'exprimer de façon unique comme la somme d'une variable constante et d'une variable de moyenne nulle:

$$\mathcal{X} = E\{\mathcal{X}\}1_\Omega + (\mathcal{X} - E\{\mathcal{X}\}1_\Omega).$$

- La variable

$$\mathcal{Z} = E\{\mathcal{Y}\}1_\Omega = \langle \mathcal{X}, 1_\Omega \rangle 1_\Omega$$

  est l'élément de $\mathcal{F}_{1_\Omega}$ le plus proche de $\mathcal{Y}$ au sens de la mesure de distance que nous avons définie entre variables aléatoires sur base du produit scalaire. On dit que $E\{\mathcal{Y}\}1_\Omega$ est la *projection orthogonale* de $\mathcal{Y}$ sur $\mathcal{F}_{1_\Omega}$.

- Plus généralement, la variable $\mathcal{Z} = E\{\mathcal{Y}|\mathcal{X}\}$ est l'élément dans $\mathcal{F}_\mathcal{X}$ le plus proche de $\mathcal{Y}$. On dit que $E\{\mathcal{Y}|\mathcal{X}\}$ est la *projection orthogonale* de $\mathcal{Y}$ sur $\mathcal{F}_\mathcal{X}$. Cette propriété est démontrée à la section 4.3.

L'opérateur d'espérance conditionnelle peut donc s'interpréter comme la projection d'une variable aléatoire sur le sous-espace des variables aléatoires qui peuvent s'exprimer comme une fonction de la variable de conditionnement. Remarquons que nous avons déjà vu que si $\mathcal{Y}$ est bien une fonction de $\mathcal{X}$, alors $\mathcal{Y} = E\{\mathcal{Y}|\mathcal{X}\}$, en d'autres termes si $\mathcal{Y}$ appartient à $\mathcal{F}_\mathcal{X}$ alors elle est égale à sa projection sur cet espace, conformément à l'intuition.

4.1.4.2 Orthogonalité vs indépendance

NB : pour désigner l'**orthogonalité** de $\mathcal{X}$ et $\mathcal{Y}$ nous écrivons $\langle \mathcal{X}, \mathcal{Y} \rangle = 0$ alors que nous réservons la notation $\mathcal{X} \perp \mathcal{Y}$ pour indiquer que ces variables sont **indépendantes**.

Ces propriétés sont à rapprocher de la notion d'indépendance de deux variables aléatoires. On peut en effet montrer que (suite à la définition de la notion d'indépendance de v.a.) :

- Tout élément de $\mathcal{F}_{1_\Omega}$ est indépendant (au sens probabiliste du terme) de tout élément de $\mathcal{F}_\Omega$ (y compris de lui-même). En particulier, 0_Ω est indépendante de toute variable aléatoire.

- Si $\mathcal{Y} \perp \mathcal{X}$ alors $\forall \mathcal{Z} \in \mathcal{F}_\mathcal{X}$, on a $\mathcal{Y} \perp \mathcal{Z}$.

- Si $\mathcal{Y}$ ou $\mathcal{X}$ est de moyenne nulle, alors $[\mathcal{Y} \perp \mathcal{X} \Rightarrow \langle \mathcal{X}, \mathcal{Y} \rangle = 0]$. Cependant, la réciproque n'est pas vraie.

- $\mathcal{Y} \perp \mathcal{X} \Leftrightarrow (\mathcal{Y} - E\{\mathcal{Y}\}) \perp (\mathcal{X} - E\{\mathcal{X}\})$.

- $\mathcal{Y} \perp \mathcal{X}$, si et seulement si $\langle f(\mathcal{X}), g(\mathcal{Y}) \rangle = 0$ quelles que soient les fonctions f et g de moyennes nulles.

On voit donc que la notion d'orthogonalité est intimement liée à celle d'indépendance probabiliste. Cependant, on voit aussi que la notion d'indépendance probabiliste n'est pas équivalente à la notion d'orthogonalité ni à celle d'indépendance linéaire de vecteurs dans $\mathbb{R}^n$.

Dans le cas considéré ici (Ω fini), toutes ces propriétés énoncées ci-dessus peuvent être démontrées assez facilement à partir des notions connues à ce stade. Nous ne le faisons cependant pas ici, car nous reviendrons à la section 4.3 sur ces propriétés dans le cas général d'un espace $(\Omega, \mathcal{E}, P)$ quelconque.

4.2 VARIABLES ALÉATOIRES CONTINUES ET CONDITIONNEMENT

4.2.1 Une des deux variables est continue et l'autre est discrète

On peut directement étendre ce qui précède au cas où $\mathcal{Y}$ est une variable continue ($\mathcal{X}$ étant discrète) en remplaçant les probabilités par les fonctions de répartition ou des densités. On note

$$F_{\mathcal{Y}|\mathcal{X}}(y|x) \stackrel{\triangle}{=} P(\mathcal{Y}(\omega) < y|\mathcal{X}(\omega) = x), \tag{4.23}$$

puis si elle existe, la densité conditionnelle $f_{\mathcal{Y}|\mathcal{X}}(y|x)$ est la dérivée de $F_{\mathcal{Y}|\mathcal{X}}(y|x)$ en y.

La fonction de répartition marginale s'écrit

$$F_{\mathcal{Y}}(y) \stackrel{\triangle}{=} \sum_{i=1}^{k} P_X(x_i) F_{\mathcal{Y}|\mathcal{X}}(y|x_i) \tag{4.24}$$

qui dérivée terme à terme donne la densité marginale

$$f_{\mathcal{Y}}(y) \stackrel{\triangle}{=} \sum_{i=1}^{k} P_X(x_i) f_{\mathcal{Y}|\mathcal{X}}(y|x_i). \tag{4.25}$$

Les théorèmes de l'espérance et de la variance totales restent également d'application.

On peut également écrire

$$P(\mathcal{X} = x|\mathcal{Y} < y) = \frac{F_{\mathcal{Y}|\mathcal{X}}(y|x) P_{\mathcal{X}}(x)}{F_{\mathcal{Y}}(y)}, \tag{4.26}$$

mais nous ne pouvons pas pour le moment écrire

$$P(\mathcal{X} = x|\mathcal{Y} = y) = \frac{f_{\mathcal{Y}|\mathcal{X}}(y|x) P_{\mathcal{X}}(x)}{f_{\mathcal{Y}}(y)}, \tag{4.27}$$

car la condition $\mathcal{Y} = y$ désigne un événement de probabilité nulle par rapport auquel on ne peut pas en principe conditionner. Nous allons indiquer ci-dessous sous quelles conditions un conditionnement de ce type est permis.

Illustration. Un exemple pratique important où on considère les dépendances entre variables continues et discrètes est fourni par la théorie de la décision, qui intervient dans les problèmes de classification en apprentissage automatique, et également dans les problèmes de transmission de données numériques à l'aide de signaux analogiques et dans les problèmes de diagnostic (notamment en médecine).

Prenons par exemple, le problème de l'allocation de crédit bancaire qui se ramène à celui de l'étude des relations entre variables numériques (montant du crédit souhaité, niveau de salaire, endettement, âge ...) et discrètes décrivant la situation financière et sociale d'un demandeur de crédit (son état civil, le fait d'être propriétaire de son lieu d'habitation, son statut professionnel...), et la décision de la banque (accord ou non du crédit).

Du point de vue du banquier non altruiste, la décision optimale est celle qui maximise l'espérance mathématique du bénéfice de la banque. Si le crédit est accordé, ce bénéfice dépendra du fait que le demandeur sera capable de rembourser les mensualités ou non. Si le crédit n'est pas accordé, le bénéfice est nul. Le banquier fera donc appel à un logiciel qui déterminera, sur base des informations fournies par le demandeur, la probabilité de remboursement complet du crédit (disons $P(\mathcal{R} = V|\mathcal{I})$, où $\mathcal{R}$ désigne une variable qui vaut V s'il y a remboursement et F sinon, et $\mathcal{I}$ symbolise les informations propres au demandeur), à partir de laquelle on pourra déterminer l'espérance mathématique du bénéfice par la formule de l'espérance totale (conditionnée par l'information fournie par le demandeur)

$$E\{\mathcal{B}|\mathcal{I}\} = E\{\mathcal{B}|\mathcal{R} = V, \mathcal{I}\} P(\mathcal{R} = V|\mathcal{I}) + E\{\mathcal{B}|\mathcal{R} = F, \mathcal{I}\} P(\mathcal{R} = F|\mathcal{I}). \tag{4.28}$$

Dans cette formule, on a évidemment

$$P(\mathcal{R} = F|\mathcal{I}) = 1 - P(\mathcal{R} = V|\mathcal{I}),$$

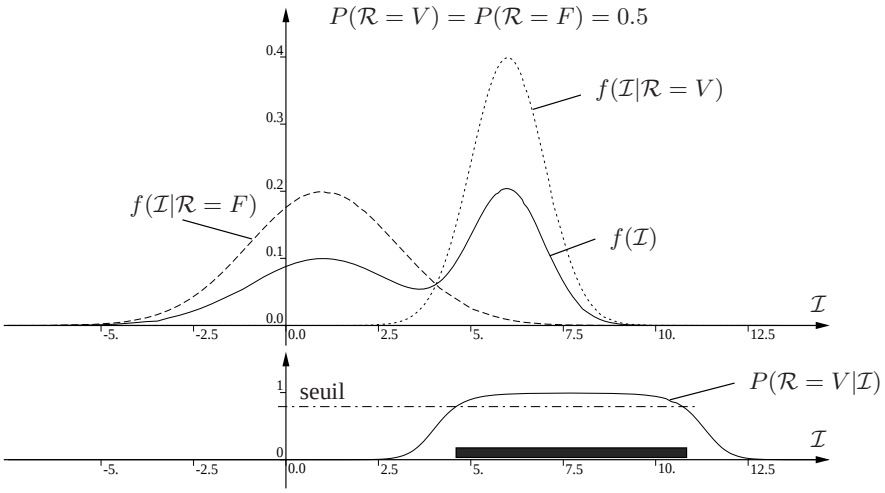


Figure 4.2: Illustration des densités conditionnelles

et le chiffre $E\{\mathcal{B}|\mathcal{R}=V,\mathcal{I}\}$ correspond au gain de la banque calculé au moyen de formules d'actualisation tenant compte des conditions du crédit (intérêt, type de remboursement, ...), du coût de l'argent immobilisé que la banque doit assumer, et est évidemment proportionnel au montant du crédit. D'autre part, le terme $E\{\mathcal{B}|\mathcal{R}=F,\mathcal{I}\}$ est quant à lui un "bénéfice" négatif.

Par conséquent, le crédit sera alloué si

$$E\{\mathcal{B}|\mathcal{I}\}>0 \Leftrightarrow P(\mathcal{R}=V|\mathcal{I})>\frac{-E\{\mathcal{B}|\mathcal{R}=F,\mathcal{I}\}}{E\{\mathcal{B}|\mathcal{R}=V,\mathcal{I}\}-E\{\mathcal{B}|\mathcal{R}=F,\mathcal{I}\}}, \tag{4.29}$$

et le problème se ramène donc essentiellement au calcul de $P(\mathcal{R}=V|\mathcal{I})$ et à la comparaison de celle-ci à un certain seuil, $\mathcal{R}$ étant une variable discrète et $\mathcal{I}$ un ensemble de variables dont en général certaines sont discrètes et d'autres continues. Nous verrons au cours d'apprentissage automatique que les méthodes utilisées par les banquiers se fondent essentiellement sur une approximation de $P(\mathcal{R}=V|\mathcal{I})$ obtenue à partir de bases de données des clients antérieurs de la banque et grâce aux méthodes d'apprentissage.

Notons que nous avons utilisé la notation explicite $\mathcal{R}=V$ ou $\mathcal{R}=F$ pour bien mettre en évidence le conditionnement sur des valeurs prises par la v.a. discrète $\mathcal{R}$. Selon notre convention, la notation $f(\mathcal{I}|\mathcal{R})$ désigne en effet une fonction à deux arguments définie par

$$f(\mathcal{I}|\mathcal{R}) \quad = \quad \begin{cases} f(\mathcal{I}|\mathcal{R}=V) \text{ si } \mathcal{R}=V \\ f(\mathcal{I}|\mathcal{R}=F) \text{ si } \mathcal{R}=F \end{cases}, \tag{4.30}$$

et $f(\mathcal{I},\mathcal{R})$ est définie par

$$f(\mathcal{I}|\mathcal{R})P(\mathcal{R}) \tag{4.31}$$

où $\mathcal{R}$ peut désigner soit la valeur V soit la valeur F.

Cette remarque étant faite, illustrons ces idées graphiquement pour un cas simple où $\mathcal{I}$ se réduit à une seule variable numérique (disons un chiffre magique obtenu en combinant les différentes informations selon une formule pré-établie) et faisons l'hypothèse que cette variable est continue. La figure 4.2 représente graphiquement la situation, en terme des densités de probabilité $f(\mathcal{I})$, $f(\mathcal{I}|\mathcal{R}=V)$, $f(\mathcal{I}|\mathcal{R}=F)$ et la probabilité conditionnelle $P(\mathcal{R}=V|\mathcal{I})$.

Notons qu'à la figure 4.2 les distributions conditionnelles $f(\mathcal{I}|\mathcal{R}=V)$ et $f(\mathcal{I}|\mathcal{R}=F)$ sont gaussiennes, de moyennes et de variances différentes. On suppose donc que les bons et les mauvais clients présentent des valeurs assez différentes de notre variable "magique". On a également supposé qu'a priori dans la population qui s'adresse aux banques pour obtenir des crédits on a la même proportion de bons et de mauvais clients, ce qui se traduit par l'égalité des probabilités a priori $P(\mathcal{R}=V)$ et $P(\mathcal{R}=F)$. On a,

$$f(\mathcal{I})=f(\mathcal{I}|\mathcal{R}=V)P(\mathcal{R}=V)+f(\mathcal{I}|\mathcal{R}=F)P(\mathcal{R}=F), \tag{4.32}$$

et la probabilité a posteriori $P(\mathcal{R} = V|\mathcal{I})$ représentée sur la partie inférieure de la figure 4.2 est obtenue par la formule de Bayes

$$P(\mathcal{R} = V|\mathcal{I}) = \frac{f(\mathcal{I}|\mathcal{R} = V)P(\mathcal{R} = V)}{f(\mathcal{I})}, \tag{4.33}$$

et on voit qu'elle vaut 0.5 au point de croisement des trois courbes du haut, c'est-à-dire au point où,

$$f(\mathcal{I}) = f(\mathcal{I}|\mathcal{R} = V) = f(\mathcal{I}|\mathcal{R} = F), \tag{4.34}$$

parce que les classes sont a priori équiprobables.

Sur la partie inférieure de la figure on a illustré la règle de décision du banquier par un seuil supposé indépendant de $\mathcal{I}$ (ce qui n'est pas nécessairement vrai en pratique comme il ressort des formules générales indiquées ci-dessus). L'ensemble des valeurs de $\mathcal{I}$ pour lesquelles le crédit est alloué est l'intervalle indiqué sur la figure.

(Suggestion : trouver l'expression générale en fonction de $P(\mathcal{R} = V)$, de la relation entre $f(\mathcal{I}|\mathcal{R} = V)$ et $f(\mathcal{I}|\mathcal{R} = F)$ au point où $P(\mathcal{R} = V|\mathcal{I}) = seuil$.)

4.2.2 Deux variables $\mathcal{X}$ et $\mathcal{Y}$ conjointement continues

Nous disons que les variables aléatoires $\mathcal{X}$ et $\mathcal{Y}$ sont conjointement continues, s'il existe une fonction $f_{\mathcal{X},\mathcal{Y}}$ (appelée densité conjointe) telle que $\forall a, b \in \mathbb{R} : F_{\mathcal{X},\mathcal{Y}}(a, b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{\mathcal{X},\mathcal{Y}}(x, y)\,\mathrm{d}x\,\mathrm{d}y$. Dans ce cas, les deux variables sont continues, et leurs densités marginales et conditionnelles existent aussi et s'obtiennent par analogie au cas discret, de même que leurs espérances et variances conditionnelles.

En particulier on a (en évitant de diviser par zéro)

$$f_{\mathcal{X}}(x) = \int_{\mathbb{R}} f_{\mathcal{X},\mathcal{Y}}(x, y)\,\mathrm{d}y \qquad \text{et} \qquad f_{\mathcal{Y}}(y) = \int_{\mathbb{R}} f_{\mathcal{X},\mathcal{Y}}(x, y)\,\mathrm{d}x, \tag{4.35}$$

$$f_{\mathcal{Y}|\mathcal{X}}(y|x) = \frac{f_{\mathcal{X},\mathcal{Y}}(x, y)}{f_{\mathcal{X}}(x)} \qquad \text{et} \qquad f_{\mathcal{X}|\mathcal{Y}}(x|y) = \frac{f_{\mathcal{X},\mathcal{Y}}(x, y)}{f_{\mathcal{Y}}(y)}, \tag{4.36}$$

$$E\{\mathcal{Y}|x\} = \int y f_{\mathcal{Y}|\mathcal{X}}(y|x)\,\mathrm{d}y \qquad \text{et} \qquad V\{\mathcal{Y}|x\} = \int (y - E\{\mathcal{Y}|x\})^2 f_{\mathcal{Y}|\mathcal{X}}(y|x)\,\mathrm{d}y, \tag{4.37}$$

$$E\{\mathcal{X}|y\} = \int x f_{\mathcal{X}|\mathcal{Y}}(x|y)\,\mathrm{d}x \qquad \text{et} \qquad V\{\mathcal{X}|y\} = \int (x - E\{\mathcal{X}|y\})^2 f_{\mathcal{X}|\mathcal{Y}}(x|y)\,\mathrm{d}x, \tag{4.38}$$

et les formules de Bayes

$$f_{\mathcal{Y}|\mathcal{X}}(y|x) = \frac{f_{\mathcal{X}|\mathcal{Y}}(x|y)f_{\mathcal{Y}}(y)}{f_{\mathcal{X}}(x)} \qquad \text{et} \qquad f_{\mathcal{X}|\mathcal{Y}}(x|y) = \frac{f_{\mathcal{Y}|\mathcal{X}}(y|x)f_{\mathcal{X}}(x)}{f_{\mathcal{Y}}(y)}. \tag{4.39}$$

Conditions d'existence. Dans le cadre de ce cours, on peut se contenter de l'information suivante : si $\mathcal{Y}$ est une variable aléatoire réelle, et si $\mathcal{X}$ est une variable aléatoire soit discrète, soit à valeurs dans $\mathbb{R}^p$, alors il est permis de conditionner Ω et donc $\mathcal{Y}$ par rapport à $\mathcal{X}$ localement. De plus, si $E\{\mathcal{Y}\}$ existe alors il existe une v.a. aléatoire "espérance conditionnelle" qui satisfait au théorème de l'espérance totale. Enfin, si $V\{\mathcal{Y}\}$ existe, alors la variance conditionnelle existe aussi, et cette v.a. satisfait alors aussi au théorème de la variance totale.

Cas général. Pour une présentation rigoureuse de la notion générale de densité de probabilité conditionnelle, et d'espérance et de variance conditionnelle, et de leurs conditions d'existence, nous renvoyons le lecteur à l'ouvrage [Bil79]. La référence [Sap90] en discute les implications utiles dans certaines applications, lorsqu'on est amené appliquer le conditionnemment à des v.a. quelconques, ni continues ni discrètes.

4.2.3 Covariance, coefficient de corrélation, et régression linéaire au sens des moindres carrés

Rappelons qu'on définit la covariance de deux variables aléatoires réelles par

$$\mathrm{cov}\{\mathcal{X}; \mathcal{Y}\} = E\{(\mathcal{X} - E\{\mathcal{X}\})(\mathcal{Y} - E\{\mathcal{Y}\}) = E\{\mathcal{X}\mathcal{Y}\} - E\{\mathcal{X}\}E\{\mathcal{Y}\}.$$

Notons que si la variance de l'une au moins des deux variables est nulle, alors $\text{cov}\{\mathcal{X};\mathcal{Y}\} = 0$ (voir section 4.3.4).

Lorsque les variances des deux variables sont strictement positives, on définit le coefficient de corrélation linéaire de deux variables de la manière suivante.

Coefficient de corrélation linéaire

$$\rho_{\mathcal{X};\mathcal{Y}} \triangleq \frac{\text{cov}(\mathcal{X};\mathcal{Y})}{\sigma_{\mathcal{X}}\sigma_{\mathcal{Y}}}.$$

Le terme de coefficient de corrélation linéaire utilisé pour désigner cette grandeur est expliqué dans la suite de cette section. Il s'agit d'une grandeur symétrique tout comme la covariance : on a $\rho_{\mathcal{X};\mathcal{Y}} = \rho_{\mathcal{Y};\mathcal{X}}$. On a $|\rho_{\mathcal{X};\mathcal{Y}}| \leq 1$, et si $\mathcal{X} = \mathcal{Y}$, $\rho_{\mathcal{X};\mathcal{Y}} = 1$ (voir 4.3.4).

4.2.3.1 Régression linéaire au sens des moindres carrés

Considérons deux variables aléatoires réelles conjointement continues et de densité conjointe $f_{\mathcal{X},\mathcal{Y}}$, et posons le problème de prédire la valeur de $\mathcal{Y}$ à l'aide d'une fonction affine de $\mathcal{X}$.

Il s'agit donc de trouver les constantes a et b telles que la variable $\mathcal{Z} = a + b\mathcal{X}$ soit en un certain sens la plus proche possible de la variable $\mathcal{Y}$. En d'autres termes, nous souhaitons que les valeurs de la variable $\mathcal{Y} - \mathcal{Z}$ soient aussi petites que possible en valeur absolue.

Pour préciser cet objectif, définissons l'écart quadratique moyen entre la variable $\mathcal{Z}$ et la variable $\mathcal{Y}$ en fonction de a et b par

$$MSE(a,b) \triangleq \int_{\mathbb{R}}\int_{\mathbb{R}} (y-a-bx)^2 f_{\mathcal{X},\mathcal{Y}}(x,y)\,\mathrm{d}x\,\mathrm{d}y = E\{(\mathcal{Y}-\mathcal{Z})^2\}. \tag{4.40}$$

Notons que cette fonction est bien définie quelles que soient les valeurs de a et b si et seulement si à la fois $\mathcal{X}$ et $\mathcal{Y}$ sont de variance finie. Nous allons donc supposer que c'est le cas dans ce qui suit.

La fonction $MSE(a,b)$ est positive ou nulle mais non bornée supérieurement et continûment dérivable par rapport à ses arguments. Elle admet par conséquent un minimum global. Nous souhaitons alors trouver des valeurs a et b dans $\mathbb{R}$ qui réalisent ce minimum global. A l'optimum on doit avoir

$$\frac{\partial MSE(a,b)}{\partial a} = 0, \tag{4.41}$$

$$\frac{\partial MSE(a,b)}{\partial b} = 0. \tag{4.42}$$

On a

$$\frac{\partial (y-a-bx)^2}{\partial a} = -2(y-a-bx), \tag{4.43}$$

$$\frac{\partial (y-a-bx)^2}{\partial b} = -2(y-a-bx)x, \tag{4.44}$$

et en passant la dérivée sous le signe d'intégration on doit donc avoir:

$$\frac{\partial MSE(a,b)}{\partial a} = -2(E\{\mathcal{Y}\} - a - bE\{\mathcal{X}\}) = 0, \tag{4.45}$$

$$\frac{\partial MSE(a,b)}{\partial b} = -2(E\{\mathcal{Y}\mathcal{X}\} - aE\{\mathcal{X}\} - bE\{\mathcal{X}^2\}) = 0. \tag{4.46}$$

On tire, après quelques manipulations

$$b = \frac{E\{\mathcal{X}\mathcal{Y}\} - E\{\mathcal{X}\}E\{\mathcal{Y}\}}{E\{\mathcal{X}^2\} - (E\{\mathcal{X}\})^2} = \frac{\text{cov}(\mathcal{X};\mathcal{Y})}{V\{\mathcal{X}\}}, \tag{4.47}$$

$$a = E\{\mathcal{Y}\} - bE\{\mathcal{X}\} = E\{\mathcal{Y}\} - \frac{\text{cov}(\mathcal{X};\mathcal{Y})}{V\{\mathcal{X}\}}E\{\mathcal{X}\}. \tag{4.48}$$

Ces valeurs sont bien définies de façon univoque, puisque que la variance de la variable $\mathcal{X}$ est strictement non nulle et finie. En effet, par hypothèse $\mathcal{X}$ est de variance finie; comme de plus nous avons supposé que $\mathcal{X}$ et $\mathcal{Y}$ sont conjointement continues, $\mathcal{X}$ est aussi continue et par conséquent sa variance ne peut être nulle.

Discussion. Lorsque les variables $\mathcal{X}$ et $\mathcal{Y}$ sont d'espérances nulles, ces formules se simplifient comme suit:

$$b = \frac{E\{\mathcal{X}\mathcal{Y}\}}{E\{\mathcal{X}^2\}} = \frac{\text{cov}(\mathcal{X};\mathcal{Y})}{V\{\mathcal{X}\}}, \tag{4.49}$$

$$a = 0. \tag{4.50}$$

Lorsque les variables sont aussi de variance égale à 1, on a de plus

$$b = \rho_{\mathcal{X};\mathcal{Y}}. \tag{4.51}$$

On déduit de ce qui précède que

$$\mathcal{Z} = \mu_{\mathcal{Y}} + \rho_{\mathcal{X};\mathcal{Y}}\sigma_{\mathcal{Y}}\frac{(\mathcal{X} - \mu_{\mathcal{X}})}{\sigma_{\mathcal{X}}} \tag{4.52}$$

est la meilleure approximation au sens des moindres carrés de $\mathcal{Y}$ par une fonction affine de $\mathcal{X}$.

Comme la meilleure approximation au sens des moindres carrés de $\mathcal{Y}$ par une fonction quelconque de la variable $\mathcal{X}$ est la variable $\mathcal{Z} = E\{\mathcal{Y}|\mathcal{X}\}$, on déduit que si $E\{\mathcal{Y}|x\}$ varie de façon affine avec x, alors la formule (4.52) exprime aussi la variable $E\{\mathcal{Y}|\mathcal{X}\}$.

Exemple 5. Scores simultanés aux examens de deux cours. Nous considérons l'ensemble des étudiants qui sont amenés à suivre pendant un même semestre deux cours à l'Université, dans le cadre des études d'ingénieur en deuxième année du bachelier. Le premier cours est un cours plutôt théorique et le second est un cours plutôt pratique. Nous désignons par $\mathcal{X}$ (respectivement $\mathcal{Y}$) la note obtenue par un de ces étudiants dans le premier cours (respectivement dans le second cours); ces notes sont comprises dans l'intervalle $[0, 20]$. L'ensemble Ω est composé des étudiants inscrits aux deux cours pour une année académique donnée, disons qu'il y en a 200. Nous voudrions étudier les relations entre les deux cotes obtenues par un étudiant.

Nous proposons le modèle suivant pour étudier ce problème. Chaque étudiant décide d'allouer une quantité de travail totale $\mathcal{T}$ pour préparer les examens des deux cours: $\mathcal{T}$ est un entier situé entre 1 et 10 (inclus) et nous supposons que la distribution de cette variable est uniforme. Il y a quatre types d'étudiants, à savoir ceux qui ont bien aimé le cours théorique mais pas le cours pratique (type A), ceux qui ont bien aimé le cours pratique mais pas le cours théorique (B), ceux qui ont bien aimé les deux (C), et ceux qui n'ont aimé aucun des deux (D). Nous considérons, de façon idéalisée, que le fait d'avoir aimé ces deux cours ou non n'a pas d'influence sur la quantité totale $\mathcal{T}$ de travail allouée par un étudiant pour préparer les examens des deux cours, mais que cela joue sur l'allocation de cette quantité totale de travail entre les deux cours. Nous désignons par $\mathcal{T}_{\mathcal{X}}$ la partie allouée pour le premier cours et $\mathcal{T}_{\mathcal{Y}}$ celle allouée au second, avec donc $\mathcal{T}_{\mathcal{X}} + \mathcal{T}_{\mathcal{Y}} = \mathcal{T}$, et nous supposons qu'un étudiant qui aime bien un cours mais pas un autre, consacrera une unité de travail en moins pour le cours qu'il n'aime pas que pour le cours qu'il aime bien, et qu'un étudiant qui n'a pas de préférence consacrera la même quantité de travail aux deux cours. Enfin, nous supposons que la cote obtenue par un étudiant dans un cours (disons la valeur de $\mathcal{X}$, s'il s'agit du premier cours) est directement liée à la quantité de travail qu'il a allouée pour le préparer (et ne dépend donc pas du fait qu'il l'aime ou pas, une fois que cette quantité de travail allouée est connue), ce que nous exprimons par les formules $\mathcal{X} = 7 + 2\mathcal{T}_{\mathcal{X}} + \mathcal{Z}_1$ et $\mathcal{Y} = 7 + 2\mathcal{T}_{\mathcal{Y}} + \mathcal{Z}_2$ où $\mathcal{Z}_1$ et $\mathcal{Z}_2$ sont des variables aléatoires entières uniformes sur $[-1, 1]$ indépendantes et indépendantes des autres variables aléatoires introduites ($\mathcal{X}, \mathcal{Y}, \mathcal{T}$) et du profil de l'étudiant (A, B, C ou D).

On considère l'expérience aléatoire qui consiste à choisir un étudiant ω au hasard parmi les 200 inscrits, $\mathcal{X}(\omega)$ désignant sa cote obtenue dans le premier cours et $\mathcal{Y}(\omega)$ celle qu'il aura obtenue dans le second cours, $\mathcal{T}(\omega)$ désignant la quantité de travail qu'il a allouée pour l'étude des deux cours, et $\mathcal{P}(\omega)$ son profil de préférences (A, B, C ou D). Dans cette expérience, tous les étudiants inscrits aux deux cours ont la même probabilité d'être choisis.

On demande de calculer les valeurs moyennes et les variances des variables $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{T}$, les espérances conditionnelles de $\mathcal{X}$ étant donnée $\mathcal{Y}$ (et de $\mathcal{Y}$ étant donnée $\mathcal{X}$), la covariance de $\mathcal{X}$ et de $\mathcal{Y}$ et la régression linéaire de $\mathcal{X}$ en fonction de $\mathcal{Y}$ (et même chose pour $\mathcal{Y}$ en fonction de $\mathcal{X}$) et de discuter les résultats obtenus. En particulier nous voudrions savoir si la régression linéaire est dans ce problème une bonne approximation de l'espérance conditionnelle.

On suppose ensuite que l'on peut aussi observer la variable $\mathcal{T}$, peut-être en demandant aux étudiants combien de temps au total ils ont consacrés à l'étude des deux cours. Dans ce cas, il est intéressant d'étudier la relation entre la variable $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$ et la variable $\mathcal{T}$. Que peut-on conclure de cette étude ?

Enfin, si on connaissait aussi le profil de l'étudiant (type A, B, C ou D) que pourrait-on dire de l'espérance conditionnelle de $\mathcal{X} - \mathcal{Y}$ étant donnée $\mathcal{T}$, pour un de ces profils ?

La résolution de cet exercice fait appel à plusieurs notions introduites précédemment et peut notamment servir à vérifier que ces notions sont assimilées à ce stade. Nous avons choisi volontairement de le formuler dans le cas discret, de façon à éviter des embûches potentielles liées à la manipulation de variables continues.

4.3 • SYNTHÈSE GÉOMÉTRIQUE DU PROBLÈME DE RÉGRESSION

Dans cette section nous présentons de manière très générale et plus rigoureuse les idées introduites une première fois à la section 4.1.4 dans le cas d'un espace Ω fini. Nous devons pour ce faire introduire un certain nombre de notions mathématiques nouvelles, dont l'assimilation peut être difficile au premier abord. Comme ces notions sont importantes pour la suite, nous recommandons une première lecture approfondie, puis une seconde lecture après avoir assimilé les autres notions introduites dans ce cours.

4.3.1 Espace de variables aléatoires à valeurs réelles

Partant d'un espace de probabilité $(\Omega, \mathcal{E}, P)$, on définit l'ensemble $\mathcal{F}_\Omega$ comme suit.

$\mathcal{F}_\Omega$ est l'ensemble des fonctions $(\mathcal{E}, \mathcal{B}_\mathbb{R})$ mesurables.

Dans cet ensemble $\mathcal{F}_\Omega$ on définit la notion "d'égalité presque sûrement" par

Egalité "presque sûrement"

$$\mathcal{X} \overset{p.s.}{=} \mathcal{Y} \Leftrightarrow P(\mathcal{X}(\omega) \neq \mathcal{Y}(\omega)) = 0. \tag{4.53}$$

Cette relation est une relation d'équivalence, c'est-à-dire qu'elle est

- partout définie sur $\mathcal{F}_\Omega \times \mathcal{F}_\Omega$,
- réflexive : $\forall \mathcal{X} \in \mathcal{F}_\Omega : \left(\mathcal{X} \overset{p.s.}{=} \mathcal{X}\right)$,
- symétrique : $\forall \mathcal{X}, \mathcal{Y} \in \mathcal{F}_\Omega : \left(\mathcal{X} \overset{p.s.}{=} \mathcal{Y}\right) \Rightarrow \left(\mathcal{Y} \overset{p.s.}{=} \mathcal{X}\right)$,
- et transitive. $\forall \mathcal{X}, \mathcal{Y}, \mathcal{Z} \in \mathcal{F}_\Omega : \left[\left(\mathcal{X} \overset{p.s.}{=} \mathcal{Y}\right) \wedge \left(\mathcal{Y} \overset{p.s.}{=} \mathcal{Z}\right)\right] \Rightarrow \left(\mathcal{X} \overset{p.s.}{=} \mathcal{Z}\right)$.

Pour être rigoureux, nous devons considérer dans la suite l'ensemble des classes d'équivalence induites par la relation d'équivalence $\overset{p.s.}{=}$ sur $\mathcal{F}_\Omega$. Une telle classe d'équivalence est un sous-ensemble de taille maximale de $\mathcal{F}_\Omega$ tel que tous ses éléments soient $\overset{p.s.}{=}$ deux à deux. L'ensemble des classes d'équivalence définies par une relation d'équivalence sur un ensemble définit une partition de cet ensemble en sous-ensembles. On appelle l'ensemble de sous-ensembles résultant le **quotient** de l'ensemble de départ par la relation d'équivalence. En toute rigueur, nous devrions le noter par $\mathcal{F}_{\Omega/\overset{p.s.}{=}}$ afin de le distinguer formellement de l'ensemble $\mathcal{F}_\Omega$. Pour alléger les notations, nous utiliserons $\mathcal{F}_\Omega$ indifféremment pour désigner à la fois l'ensemble de départ et sont quotient par la relation d'équivalence, et nous insisterons sur le fait que nous considérons ces classes d'équivalence de variables aléatoires "presque sûrement" égales, aux endroits où cela est approprié.

Notons que deux variables presque sûrement égales sont indistinguables d'un point de vue pratique.

4.3.2 Espace de Hilbert des variables aléatoires de carré intégrable sur Ω

Mathématiquement, la notion d'espace de Hilbert est une généralisation de la notion d'espace euclidien (i.e. similaire à $\mathbb{R}^n$). Il s'agit d'une structure mathématique qui marie intimement la notion d'espace vectoriel linéaire avec la notion de projection orthogonale et de convergence des suites. Nous allons montrer comment cette structure est construite pour étudier les variables aléatoires à valeurs réelles définies sur un espace de probabilité quelconque. Nous recommandons la référence [Rom75] au lecteur intéressé par la construction de structures mathématiques telles que les espaces de Hilbert.

4.3.2.1 Espace linéaire L^2_Ω des variables aléatoires de carré intégrable

Nous considérons maintenant le sous-ensemble de $\mathcal{F}_\Omega$ d'éléments $\mathcal{X}$ tels que $E\{\mathcal{X}^2\}$ soit finie (nous dirons $\mathcal{X}$ est de carré intégrable, ce qui implique que $E\{\mathcal{X}\}$ est aussi finie), et désignons par L^2_Ω le sous-ensemble suivant :

L'espace linéaire L^2_Ω est l'ensemble des classes d'équivalence induites par la relation $\overset{p.s.}{=}$ sur le sous-ensemble de variables aléatoires de de carré intégrable définies sur $(\Omega, \mathcal{E}, P)$.

Notons que si $\mathcal{Z}{=}a\mathcal{X} + b\mathcal{Y}$ et que $\mathcal{X}$ et $\mathcal{Y}$ sont de carré intégrable, alors $\mathcal{Z}$ l'est aussi, $\forall a, b \in \mathbb{R}$; de plus il en est alors de même de toutes les v.a. qui sont "presque sûrement" égales à $\mathcal{Z}$ et toutes ces variables partagent la même valeur de leur espérance et de l'espérance de leur carré. Ces fonctionnelles (espérance et variance) peuvent donc être appliquées indifféremment aux éléments de L^2_Ω ou à l'une quelconque de leurs variables aléatoires.

L'ensemble L^2_Ω est bien donc un *sous-espace linéaire* de l'espace de variables aléatoires définies sur Ω. Nous restreignons la suite de notre discussion à ce sous-espace.

Le **vecteur nul** de L^2_Ω correspond à l'ensemble de variables aléatoires presque sûrement nulles; il est désigné par 0_Ω.

On a évidemment $E\{(0_\Omega)^2\} = E\{(0_\Omega)\} = 0$, et $\forall \mathcal{X} \in L^2_\Omega : \mathcal{X} = \mathcal{X} + 0_\Omega$.

4.3.2.2 Droite des constantes $L^2_{1_\Omega}$

Le sous-ensemble de L^2_Ω des variables constantes ($\overset{p.s.}{=}$ à une variable aléatoire constante) est un sous-espace linéaire de L^2_Ω de dimension égale à 1: cet ensemble contient toute combinaison linéaire de ses éléments, et tous ses éléments sont $\overset{p.s.}{=}$ à un multiple de la variable aléatoire 1_Ω qui vaut 1 partout sur Ω. 1_Ω constitue donc une *base* de dimension 1 de cet ensemble. De fait, si $\mathcal{X}$ est p.s. constante, alors $\mathcal{X} \overset{p.s.}{=} E\{\mathcal{X}\}1_\Omega$.

On note ce sous-espace par $L^2_{1_\Omega}$, et on l'appelle "droite des constantes".

Evidemment, $0_\Omega \in L^2_{1_\Omega}$.

La figure 4.3 représente sous la forme d'un diagramme de Venn les relations entre les sous-ensembles de variables aléatoires qui viennent d'être introduits.

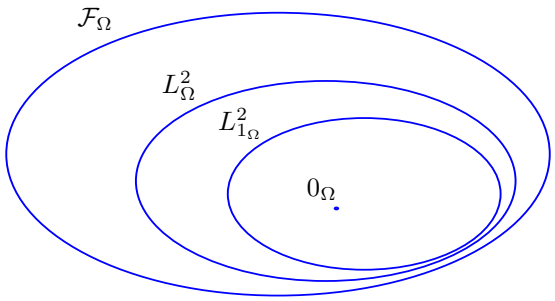


Figure 4.3: Relations entre les sous-ensembles de variables aléatoires $\mathcal{F}_\Omega, L^2_\Omega, L^2_{1_\Omega}$ et l'élément nul 0_Ω.

4.3.2.3 Sous-espace linéaire $L^2_\mathcal{X}$ des fonctions d'une variable $\mathcal{X}$

Plus généralement, si $\mathcal{X} \in L^2_\Omega$, nous désignons par $L^2_\mathcal{X}$ le sous-ensemble de L^2_Ω des éléments qui peuvent s'écrire sous la forme d'une fonction de la variable $\mathcal{X}$. Il s'agit de toutes les fonctions de $\mathcal{X}$ de carré intégrable.

$L^2_\mathcal{X}$ est un sous-espace linéaire de L^2_Ω, car il contient toutes ses combinaisons linéaires, puisque la combinaison linéaire de fonctions de carré intégrable de $\mathcal{X}$ est aussi une fonction de carré intégrable de $\mathcal{X}$.

Notons qu'en particulier, $L^2_\mathcal{X}$ contient (quelle que soit $\mathcal{X} \in L^2_\Omega$) toutes les variables aléatoires constantes (puisque toute variable aléatoire constante peut-être décrite comme une fonction de $\mathcal{X}$). On a donc $L^2_{1_\Omega} \subset L^2_\mathcal{X}$.

Si $\mathcal{X}$ est une variable aléatoire constante, on a $L^2_{\mathcal{X}} = L^2_{1_\Omega}$, de dimension 1.

Plus généralement, si $\mathcal{X}$ est une variable aléatoire discrète prenant un nombre fini (disons m) de valeurs différentes sur Ω, alors $L^2_{\mathcal{X}}$ est de dimension m. En effet, si $\{x_1, \ldots, x_m\}$ sont les m valeurs de $\mathcal{X}$, supposées de probabilité non-nulle, on peut écrire $\mathcal{X}$ sous la forme

$$\mathcal{X}(\omega) = \sum_{i=1}^{m} x_i 1_{X_i}(\omega), \tag{4.54}$$

avec $X_i = \{\omega : \mathcal{X}(\omega) = x_i\}$ et $P(X_i) > 0, \forall i$. On voit que toute fonction de $\mathcal{X}$ peut alors s'écrire sous la forme

$$\mathcal{Y}(\omega) = \sum_{i=1}^{m} y_i 1_{X_i}(\omega), \tag{4.55}$$

et que toute variable qui s'écrit sous cette forme est une fonction de $\mathcal{X}$. Il en résulte que, puisque les x_i sont tous différents et que $\sum_{i=1}^{m} P(X_i) = 1$, on a

$$\mathcal{Y} \overset{p.s}{=} 0_\Omega \Leftrightarrow \forall i = 1, \ldots m : y_i = 0, \tag{4.56}$$

autrement dit, les variables aléatoires $\{1_{X_1}, \ldots, 1_{X_m}\}$ sont linéairement indépendantes et forment une base de dimension m de l'espace $L^2_{\mathcal{X}}$. Evidemment, si $m = 1$, $\mathcal{X}$ est une constante et on retombe sur $L^2_{1_\Omega}$.

Réciproquement, si $L^2_{\mathcal{X}}$ est de dimension finie, alors $\mathcal{X}$ est p.s. égale à une variable aléatoire ne prenant qu'un nombre fini de valeurs différentes sur Ω.

Dans tous les autres cas, $L^2_{\mathcal{X}}$ est donc de dimension infinie. La dimension de $L^2_{\mathcal{X}}$ résume la richesse de l'information fournie par $\mathcal{X}$ sur ω. Par exemple, si $\mathcal{Y}$ est une fonction de $\mathcal{X}$ alors la dimension de $L^2_{\mathcal{Y}}$ est au plus égale à la dimension de $L^2_{\mathcal{X}}$.

On peut étendre ces notions en définissant les sous-espaces $L^2_{\mathcal{X},\mathcal{Y}}$, $L^2_{\mathcal{X},\mathcal{Y},\mathcal{Z}}$ etc. de L^2_Ω composés des variables aléatoires (de carré intégrable) qui peuvent s'écrire comme une fonction de deux, trois, etc., variables aléatoires. On a en général $dim(L^2_{\mathcal{X}}) \leq dim(L^2_{\mathcal{X},\mathcal{Y}}) \leq dim(L^2_{\mathcal{X},\mathcal{Y},\mathcal{Z}}) \leq$ etc.

Enfin, terminons cette discussion sur la dimension des espaces de variables aléatoires, en remarquant que si Ω est un ensemble fini, de *cardinalité* égale à m, alors toutes ces dimensions sont inférieures ou égales à m, conformément aux idées introduites à la section 4.1.4.

4.3.2.4 Sous-espace linéaire $L^2_{\text{aff}(\mathcal{X})}$ des fonctions affines d'une variable $\mathcal{X}$

L'ensemble de variables aléatoires qui peuvent s'écrire sous la forme d'une fonction affine d'une variable $\mathcal{X}$ (i.e. sous la forme $\mathcal{Y} = a + b\mathcal{X}$, avec $a, b \in \mathbb{R}$) est aussi un sous-espace linéaire de L^2_Ω. Nous notons ce sous-espace par $L^2_{\text{aff}(\mathcal{X})}$. Il contient la droite des constantes et est inclus dans $L^2_{\mathcal{X}}$.

La figure 4.4 indique les relations entre les ensembles de variables aléatoires que nous venons d'introduire.

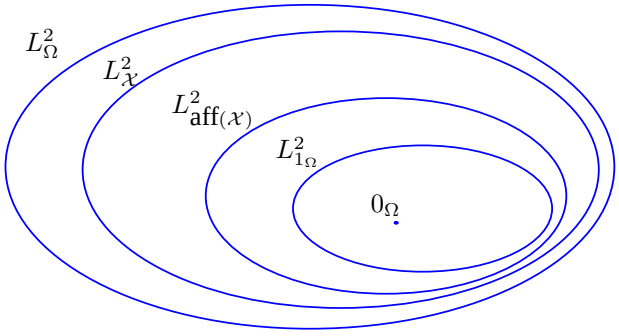


Figure 4.4: Relations entre les sous-ensembles de variables aléatoires $L^2_\Omega, L^2_{\mathcal{X}}, L^2_{\text{aff}(\mathcal{X})}, L^2_{1_\Omega}$ et l'élément nul 0_Ω.

Si $\mathcal{X}$ est une variable aléatoire constante, alors $L^2_{\text{aff}(\mathcal{X})} = L^2_{1_\Omega}$ (de dimension égale à 1). Si $\mathcal{X}$ est une variable aléatoire de variance non-nulle, $L^2_{\text{aff}(\mathcal{X})}$ est de dimension deux, et est composé de l'ensemble des combinaisons linéaires des vecteurs 1_Ω et $\mathcal{X}$ qui sont alors linéairement indépendants. Ceci est graphiquement illustré à la Figure 4.5 (NB: cette figure anticipe sur la notion de projection orthogonale définie un peu plus loin).

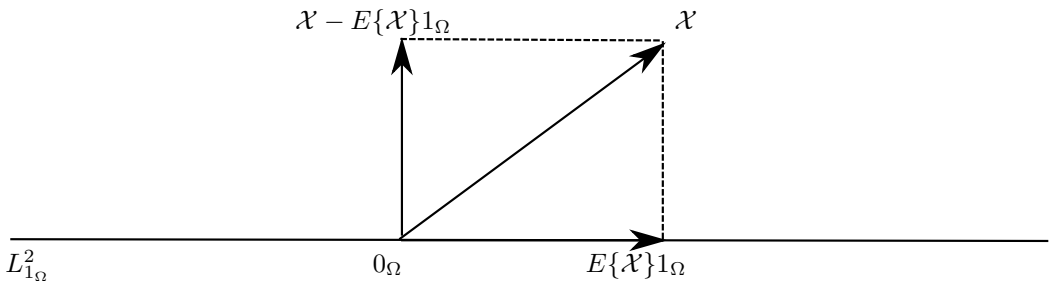


Figure 4.5: Projection d'une variable $\mathcal{X}$ de variance non-nulle sur la droite des constantes $L^2_{1_\Omega}$. Le plan formé par la droite des constantes $L^2_{1_\Omega}$ et la variable $\mathcal{X}$ constitue l'espace $L^2_{\text{aff}(\mathcal{X})}$. Si $V\{\mathcal{X}\} = 0$, $L^2_{\text{aff}(\mathcal{X})}$ se réduit à $L^2_{1_\Omega}$

Nous considérons aussi dans la suite le sous-espace linéaire de multiples de la variable $\mathcal{X}$, c'est-à-dire le sous-ensemble noté $L^2_{\text{lin}(\mathcal{X})}$ de L^2_Ω des variables $\mathcal{Y}$ s'écrivant sous la forme $\mathcal{Y} = \lambda\mathcal{X}$. Il s'agit d'un sous-espace linéaire de dimension 1 de L^2_Ω pour autant que $\mathcal{X} \neq 0_\Omega$.

4.3.2.5 Produit scalaire, norme, orthogonalité et convergence des suites, complétude

On définit par

$$\langle \mathcal{X}, \mathcal{Y} \rangle \stackrel{\triangle}{=} E\{\mathcal{X}\mathcal{Y}\}, \tag{4.57}$$

le **produit scalaire** entre deux variables aléatoires.

Cette quantité est bien définie (et finie) sur L^2_Ω, et vérifie **l'inégalité de Cauchy-Schwarz**

$$\langle \mathcal{X}, \mathcal{Y} \rangle^2 \leq \langle \mathcal{X}, \mathcal{X} \rangle \langle \mathcal{Y}, \mathcal{Y} \rangle, \tag{4.58}$$

l'égalité étant vérifiée si et seulement si $\mathcal{X}$ et $\mathcal{Y}$ sont linéairement dépendantes, c'est-à-dire si l'une des deux peut s'écrire sous la forme d'un multiple de l'autre.

Le produit scalaire ainsi défini, est une forme symétrique, c'est-à-dire que

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \langle \mathcal{Y}, \mathcal{X} \rangle \tag{4.59}$$

et bi-linéaire, c'est-à-dire que

$$\langle \mathcal{X}, \alpha\mathcal{Y}_1 + \beta\mathcal{Y}_2 \rangle = \alpha\langle \mathcal{X}, \mathcal{Y}_1 \rangle + \beta\langle \mathcal{X}, \mathcal{Y}_2 \rangle \tag{4.60}$$

et aussi

$$\langle \alpha\mathcal{X}_1 + \beta\mathcal{X}_2, \mathcal{Y} \rangle = \alpha\langle \mathcal{X}_1, \mathcal{Y} \rangle + \beta\langle \mathcal{X}_2, \mathcal{Y} \rangle. \tag{4.61}$$

Par définition deux éléments de L^2_Ω sont **orthogonaux** si leur produit scalaire vaut zéro.

La Figure 4.5, ci dessus, montre la décomposition d'une variable $\mathcal{X}$ en deux composantes, respectivement parallèle et orthogonale à 1_Ω. On a en effet

$$\langle 1_\Omega, \mathcal{X} - E\{\mathcal{X}\}1_\Omega \rangle = 0. \tag{4.62}$$

Par extension, on dit que deux sous-ensembles (ou bien deux sous-espaces) A et B de L^2_Ω sont orthogonaux, s'ils sont composés d'éléments deux à deux orthogonaux : i.e. si $\forall(\mathcal{X}, \mathcal{Y}) \in A \times B : \langle \mathcal{X}, \mathcal{Y} \rangle = 0$. Par exemple, la Figure 4.5 suggère que les sous-espaces $L^2_{1_\Omega}$ et l'ensemble des variables aléatoires multiples de $\mathcal{X} - E\{\mathcal{X}\}1_\Omega$ sont orthogonaux, ce qu'on peut vérifier en constatant que $\langle \mathcal{Y}_1, \mathcal{Y}_2 \rangle = 0, \forall\mathcal{Y}_1 = \lambda 1_\Omega, \forall\mathcal{Y}_2 = \mu(\mathcal{X} - E\{\mathcal{X}\}1_\Omega)$.

Le produit scalaire $\langle \cdot, \cdot \rangle$ induit la notion de norme sur L^2_Ω par

$$\|\mathcal{X}\| = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle} = \sqrt{E\{\mathcal{X}^2\}} = \sqrt{\sigma^2_{\mathcal{X}} + \mu^2_{\mathcal{X}}}. \tag{4.63}$$

On a $\|\mathcal{X}\| \geq 0$, l'égalité étant obtenue si et seulement si $\mathcal{X} \overset{p.s.}{=} 0_\Omega$; $\|1_\Omega\| = 1$, et $\|0_\Omega\| = 0$.

Sur la Figure 4.5 on voit que le carré de l'hypoténuse du triangle rectangle formé par $0_\Omega, \mathcal{X}$ et $E\{\mathcal{X}\}1_\Omega$ vaut la somme des carrés de ses côtés, c'est-à-dire $E\{\mathcal{X}^2\} = \mu^2_{\mathcal{X}} + \sigma^2_{\mathcal{X}}$.

La norme $\|\cdot\|$ induit la notion de distance entre deux éléments de L^2_Ω par

$$d(\mathcal{X}, \mathcal{Y}) = \|\mathcal{X} - \mathcal{Y}\|. \tag{4.64}$$

Cette distance vérifie l'inégalité triangulaire (conséquence de l'inégalité de Schwarz), c'est-à-dire

$$d(\mathcal{X}, \mathcal{Y}) \leq d(\mathcal{X}, \mathcal{Z}) + d(\mathcal{Z}, \mathcal{Y}). \tag{4.65}$$

La notion de distance permet de définir une topologie sur L^2_Ω, par la notion de boule ouverte centrée sur un élément quelconque de L^2_Ω qui permet à son tour de définir la notion de sous-ensemble ouvert L^2_Ω (un ensemble qui contient une boule ouverte centrée autour de chacun de ses éléments) et de sous-ensemble fermé de L^2_Ω (ceux qui sont les complémentaires dans L^2_Ω d'un sous-ensemble ouvert). La notion de distance permet aussi de définir la notion de convergence des suites d'éléments de L^2_Ω.

On démontre que dans L^2_Ω toute suite de variables aléatoires qui est de Cauchy, converge vers une variable aléatoire de L^2_Ω (une suite de Cauchy est une suite dont les éléments successifs deviennent de plus en plus proches). Un espace qui possède cette propriété est appelé "espace complet". Si cet espace complet est un espace linéaire normé on dit que c'est un espace de Banach. Si la norme de l'espace de Banach est induite par un produit scalaire, on dit que c'est un **espace de Hilbert**. L^2_Ω est donc un espace de Hilbert. Les propriétés des éléments des espaces de Hilbert sont le fruit de l'étude conjointe des notions de combinaison linéaire, de projection orthogonale, et de convergence.

De façon essentielle, le fait que l'espace L^2_Ω est un espace de Hilbert implique que si la suite $\mathcal{X}_n$ de variables aléatoires de L^2_Ω est telle que la série des normes converge, i.e.

$$\sum_{k=1}^{\infty} \|\mathcal{X}_k\| < \infty, \tag{4.66}$$

alors la suite des sommes partielles définie par

$$\mathcal{Y}_n \overset{\triangle}{=} \sum_{k=1}^{n} \mathcal{X}_k \tag{4.67}$$

converge vers un élément de L^2_Ω.

4.3.3 Notion de projection orthogonale

Nous définissons la notion de **projection orthogonale** d'un élément $\mathcal{Y} \in L^2_\Omega$ sur un sous-espace linéaire L de L^2_Ω comme étant l'élément de L le plus proche de $\mathcal{Y}$ au sens de la mesure de distance d induite par le produit scalaire.

On montre, conformément à l'intuition, que si $\mathcal{Z}$ est la projection orthogonale de $\mathcal{Y}$ sur un sous-espace linéaire L de L^2_Ω , alors $\langle \mathcal{X}, \mathcal{Y} - \mathcal{Z} \rangle = 0$ quelque soit $\mathcal{X} \in L$ (et en particulier $\mathcal{Z}$).

On a les résultats généraux suivants :

- La variable

$$\mathcal{Z} \overset{\triangle}{=} E\{\mathcal{Y}\}1_\Omega = \langle \mathcal{Y}, 1_\Omega \rangle 1_\Omega \tag{4.68}$$

est la projection orthogonale de $\mathcal{Y}$ sur $L^2_{1_\Omega}$, dont le vecteur de base est 1_Ω; on a bien

$$\begin{aligned}\langle \mathcal{Z}, \mathcal{Y} - \mathcal{Z}\rangle &= \langle E\{\mathcal{Y}\}1_\Omega, \mathcal{Y}\rangle - \langle E\{\mathcal{Y}\}1_\Omega, E\{\mathcal{Y}\}1_\Omega\rangle &(4.69)\\ &= E\{\mathcal{Y}\}\langle 1_\Omega, \mathcal{Y}\rangle - (E\{\mathcal{Y}\})^2\langle 1_\Omega, 1_\Omega\rangle &(4.70)\\ &= E\{\mathcal{Y}\}^2 - E\{\mathcal{Y}\}^2 = 0. &(4.71)\end{aligned}$$

La formule de König-Huyghens démontrée ci-dessous implique qu'il n'y a en effet pas d'autre v.a. constante qui soit plus proche de $\mathcal{Y}$.

- Si $\mathcal{X}$ est de variance strictement positive, la variable

$$\mathcal{Z} \triangleq \langle \mathcal{Y}, 1_\Omega\rangle 1_\Omega + \left\langle \mathcal{Y}, \frac{\mathcal{X} - \langle \mathcal{X}, 1_\Omega\rangle 1_\Omega}{\|\mathcal{X} - \langle \mathcal{X}, 1_\Omega\rangle 1_\Omega\|}\right\rangle \frac{\mathcal{X} - \langle \mathcal{X}, 1_\Omega\rangle 1_\Omega}{\|\mathcal{X} - \langle \mathcal{X}, 1_\Omega\rangle 1_\Omega\|} \tag{4.72}$$

est bien définie et représente la projection de $\mathcal{Y}$ sur $L^2_{\text{aff}(\mathcal{X})}$, car les v.a. 1_Ω et $\frac{\mathcal{X} - \langle \mathcal{X}, 1_\Omega\rangle 1_\Omega}{\|\mathcal{X} - \langle \mathcal{X}, 1_\Omega\rangle 1_\Omega\|}$ forment alors une base orthonormée de $L^2_{\text{aff}(\mathcal{X})}$ obtenue par la procédure de Gram-Schmidt appliquée au couple de v.a. $\{1_\Omega, \mathcal{X}\}$ (comme aussi suggéré à la Figure 4.5).

- $\mathcal{Z} = E\{\mathcal{Y}|\mathcal{X}\}$ est bien définie $\forall \mathcal{X}, \mathcal{Y} \in L^2_\Omega$ et représente la projection orthogonale de $\mathcal{Y}$ sur $L^2_{\mathcal{X}}$; on a donc $\langle \phi(\mathcal{X}), \mathcal{Y} - E\{\mathcal{Y}|\mathcal{X}\}\rangle = 0$ pour toute fonction ϕ de carré intégrable de $\mathcal{X}$. En particulier on a $\langle 1_\Omega, \mathcal{Y} - E\{\mathcal{Y}|\mathcal{X}\}\rangle = 0$ ce qui implique que $\mathcal{Y} - E\{\mathcal{Y}|\mathcal{X}\}$ est de moyenne nulle, et donc que les variables $\mathcal{Y}$ et $E\{\mathcal{Y}|\mathcal{X}\}$ ont la même espérance (théorème de l'espérance totale; démonstrations à la section 4.3.4.3).

4.3.4 Géométrie de L^2_Ω et de ses sous-espaces

4.3.4.1 Formule de König-Huyghens

Cette formule s'exprime comme suit.

Formule de König-Huyghens

$$E\{(\mathcal{X} - a)^2\} = V\{\mathcal{X}\} + (E\{\mathcal{X}\} - a)^2. \tag{4.73}$$

Elle s'interprète comme le théorème de Pythagore appliqué au triangle rectangle $\mathcal{X}, a1_\Omega, E\{\mathcal{X}\}1_\Omega$ de la figure 4.6. La variable $E\{\mathcal{X}\}1_\Omega$ est en effet orthogonale à la variable $\mathcal{X}$ et colinéaire avec la constante $a1_\Omega$.

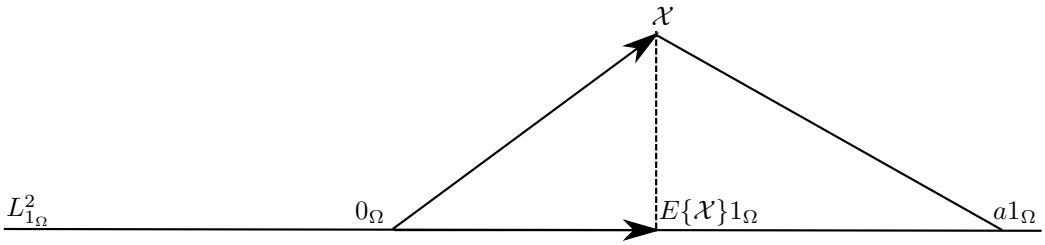


Figure 4.6: Formule de König-Huyghens : $E\{(\mathcal{X} - a)^2\} = V\{\mathcal{X}\} + (E\{\mathcal{X}\} - a)^2$

On en déduit que $E\{\mathcal{X}\}$ est la constante qui est la plus proche de $\mathcal{X}$ au sens de la norme de L^2_Ω, puisque si on la substitue à a dans l'équation précédente elle conduit au minimum absolu (égal à la variance de $\mathcal{X}$).

4.3.4.2 Coefficient de corrélation linéaire

L'inégalité de Cauchy-Schwarz appliquée aux variables $\mathcal{X} - E\{\mathcal{X}\}$ et $\mathcal{Y} - E\{\mathcal{Y}\}$ donne

$$\langle \mathcal{X} - E\{\mathcal{X}\}, \mathcal{Y} - E\{\mathcal{Y}\}\rangle^2 \leq \|\mathcal{X} - E\{\mathcal{X}\}\|^2 \|\mathcal{Y} - E\{\mathcal{Y}\}\|^2 \tag{4.74}$$

c'est-à-dire

$$|\text{cov}(\mathcal{X}; \mathcal{Y})| \leq \sigma_{\mathcal{X}}\sigma_{\mathcal{Y}}. \tag{4.75, 4.76}$$

L'égalité a lieu ssi $\mathcal{X} - E\{\mathcal{X}\} = \lambda(\mathcal{Y} - E\{\mathcal{Y}\})$, et se traduit par un coefficient de corrélation linéaire

$$\rho_{\mathcal{X};\mathcal{Y}} = \frac{\text{cov}(\mathcal{X};\mathcal{Y})}{\sigma_{\mathcal{X}}\sigma_{\mathcal{Y}}} \tag{4.77}$$

qui vaut ± 1. Les variables $\mathcal{X}$ et $\mathcal{Y}$ sont dites (linéairement) non-corrélées si $\rho_{\mathcal{X};\mathcal{Y}} = 0$, ce qui est équivalent à l'orthogonalité des variables $\mathcal{X} - E\{\mathcal{X}\}$ et $\mathcal{Y} - E\{\mathcal{Y}\}$. Cette vision géométrique est schématisée à la figure 4.7.

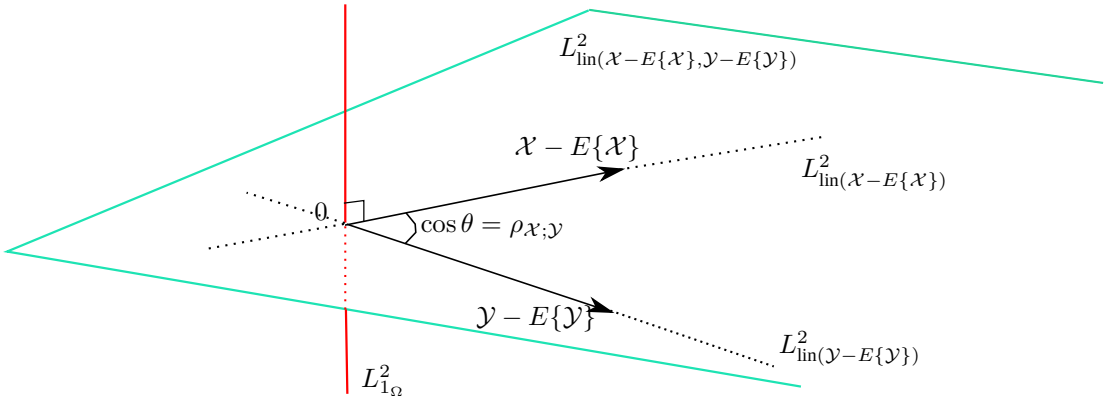


Figure 4.7: Le coefficient de corrélation linéaire est le cosinus de l'angle formé par les variables $\mathcal{X} - E\{\mathcal{X}\}$ et $\mathcal{Y} - E\{\mathcal{Y}\}$. Le plan schématisé sur la figure (orthogonal à L^2_Ω) représente ici l'ensemble des variables aléatoires $\mathcal{Z}$ qui s'écrivent sous la forme $\mathcal{Z} = \alpha(\mathcal{X} - E\{\mathcal{X}\}) + \beta(\mathcal{Y} - E\{\mathcal{Y}\})$. Ce sous-espace linéaire est noté $L^2_{\text{lin}(\mathcal{X}-E\{\mathcal{X}\},\mathcal{Y}-E\{\mathcal{Y}\})}$

4.3.4.3 Espérance conditionnelle et théorèmes de l'espérance et la variance totale

Dans le contexte qui nous occupe dans cette section, nous définissons l'espérance conditionnelle $E\{\mathcal{Y}|\mathcal{X}\}$ d'une variable aléatoire $\mathcal{Y} \in L^2_\Omega$ par rapport à une variable aléatoire $\mathcal{X} \in L^2_\Omega$ comme la projection orthogonale de $\mathcal{Y}$ sur l'espace $L^2_\mathcal{X}$, c'est-à-dire la fonction de $\mathcal{X}$ la plus proche de $\mathcal{Y}$ au sens de la norme définie sur L^2_Ω.

On montre que cette v.a. est bien définie, c'est-à-dire qu'il existe bien un élément de $L^2_\mathcal{X}$ qui réalise le minimum de la distance avec $\mathcal{Y}$. (Ceci est une conséquence du fait que $L^2_\mathcal{X}$ est un sous-espace fermé et contient donc les limites de toutes ses suites convergentes.)

On montre également que cette notion est conforme aux définitions que nous avons données antérieurement, à savoir dans le cas où les deux variables sont discretes

$$E\{\mathcal{Y}|\mathcal{X}\}(\omega) = \sum_j y_j P_{\mathcal{Y}|\mathcal{X}}(y_j|x(\omega)), \tag{4.78}$$

et dans le cas ou elles sont conjointement continues

$$E\{\mathcal{Y}|\mathcal{X}\}(\omega) = \int_{\mathbb{R}} y f_{\mathcal{Y}|\mathcal{X}}(y|x(\omega))\,\mathrm{d}y. \tag{4.79}$$

La définition générale de la notion d'espérance conditionnelle dépasse le cadre de ce cours introductif.

La figure 4.8 illustre graphiquement la notion d'espérance conditionnelle dans L^2_Ω. Sur cette figure on constate que la projection de $E\{\mathcal{Y}|\mathcal{X}\}$ sur $L^2_{1_\Omega}$ est $E\{\mathcal{Y}\}$, en d'autres termes, on a

$$E\{E\{\mathcal{Y}|\mathcal{X}\}\} = E\{\mathcal{Y}\}, \tag{4.80}$$

c'est-à-dire le **théorème de l'espérance totale**.

On voit également, sur la figure 4.8, que le triangle formé par les v.a. $\mathcal{Y}$, $E\{\mathcal{Y}\}$ et $E\{\mathcal{Y}|\mathcal{X}\}$ est un triangle rectangle dont l'hypoténuse est le vecteur $\mathcal{Y} - E\{\mathcal{Y}\}$. On doit donc avoir

$$\begin{aligned}
\|\mathcal{Y} - E\{\mathcal{Y}\}\|^2 = V\{\mathcal{Y}\} &= \|E\{\mathcal{Y}|\mathcal{X}\} - E\{\mathcal{Y}\}\|^2 + \|\mathcal{Y} - E\{\mathcal{Y}|\mathcal{X}\}\|^2, &(4.81)\\
&= \|E\{\mathcal{Y}|\mathcal{X}\} - E\{E\{\mathcal{Y}|\mathcal{X}\}\}\|^2 + E\{(\mathcal{Y} - E\{\mathcal{Y}|\mathcal{X}\})^2\}, &(4.82)\\
&= V\{E\{\mathcal{Y}|\mathcal{X}\}\} + E\{V\{\mathcal{Y}|\mathcal{X}\}\}, &(4.83)
\end{aligned}$$

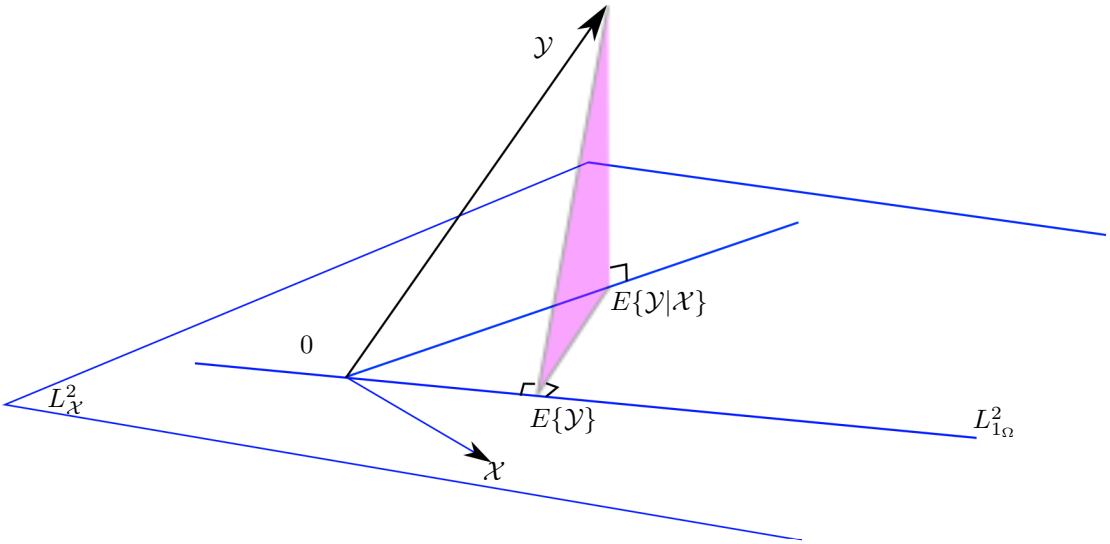


Figure 4.8: L'espérance conditionnelle $E\{\mathcal{Y}|\mathcal{X}\}$ est la projection orthogonale de $\mathcal{Y}$ sur $L^2_{\mathcal{X}}$ et la droite $(\mathcal{Y}, E\{\mathcal{Y}|\mathcal{X}\})$ est donc orthogonale à $L^2_{\mathcal{X}}$ et donc aussi à $L^2_{1_\Omega} \subset L^2_{\mathcal{X}}$. L'espérance $E\{\mathcal{Y}\}$ est la projection orthogonale de $\mathcal{Y}$ sur $L^2_{1_\Omega}$, et la droite $(\mathcal{Y}, E\{\mathcal{Y}\})$ est donc orthogonale à $L^2_{1_\Omega}$. Le plan indiqué en rose engendré par ces deux droites est donc aussi orthogonal à $L^2_{1_\Omega}$. Par conséquent, la projection de $E\{\mathcal{Y}|\mathcal{X}\}$ sur $L^2_{1_\Omega}$ est $E\{\mathcal{Y}\}$

ce qui est le **théorème de la variance totale**.

4.3.5 Indépendance de variables aléatoires et rapport de corrélation

Puisque $V\{\mathcal{Y}\} \leq V\{E\{\mathcal{Y}|\mathcal{X}\}\}$ on peut définir la grandeur, appelée rapport de corrélation, suivante :

Rapport de corrélation

$$\eta^2_{\mathcal{X};\mathcal{Y}} \triangleq \frac{V\{E\{\mathcal{Y}|\mathcal{X}\}\}}{V\{\mathcal{Y}\}} = \frac{V\{E\{\mathcal{Y}|\mathcal{X}\}\}}{V\{E\{\mathcal{Y}|\mathcal{X}\}\} + E\{V\{\mathcal{Y}|\mathcal{X}\}\}}. \tag{4.84}$$

Ce rapport est le carré du cosinus de l'angle formé par la variable $\mathcal{Y} - E\{\mathcal{Y}\}$ et le sous-espace $L^2_{\mathcal{X}}$. On a bien entendu

$$0 \leq \eta^2_{\mathcal{X};\mathcal{Y}} \leq 1. \tag{4.85}$$

Si $\eta^2_{\mathcal{X};\mathcal{Y}} = 1$, alors $E\{V\{\mathcal{Y}|\mathcal{X}\}\} = 0$ (et réciproquement). Cela implique que $\mathcal{Y}$ est (p.s.) égale à une fonction de $\mathcal{X}$.

Si $\eta^2_{\mathcal{X};\mathcal{Y}} = 0$, cela implique que $V\{E\{\mathcal{Y}|\mathcal{X}\}\} = 0$, ce qui veut dire que $E\{\mathcal{Y}|\mathcal{X}\}$ est (p.s.) égale à une constante. Dans ce cas on dit que $\mathcal{Y}$ et $\mathcal{X}$ sont non corrélées (ou indépendantes en moyenne), ce qui est en particulier la cas si elles sont indépendantes au sens probabiliste du terme, mais la réciproque n'est pas vraie.

On démontre que l'indépendance de deux variables aléatoires de L^2_Ω est équivalente à la non-corrélation de toutes leurs fonctions (voir Figure 4.9, et lire la légende).

Au chapitre suivant, nous verrons que les variables distribuées conjointement de façon Gaussienne sont indépendantes si et seulement si elles sont non-corrélées linéairement. C'est un cas particulier très important, mais il est au moins aussi important de noter que cela n'est qu'un cas particulier.

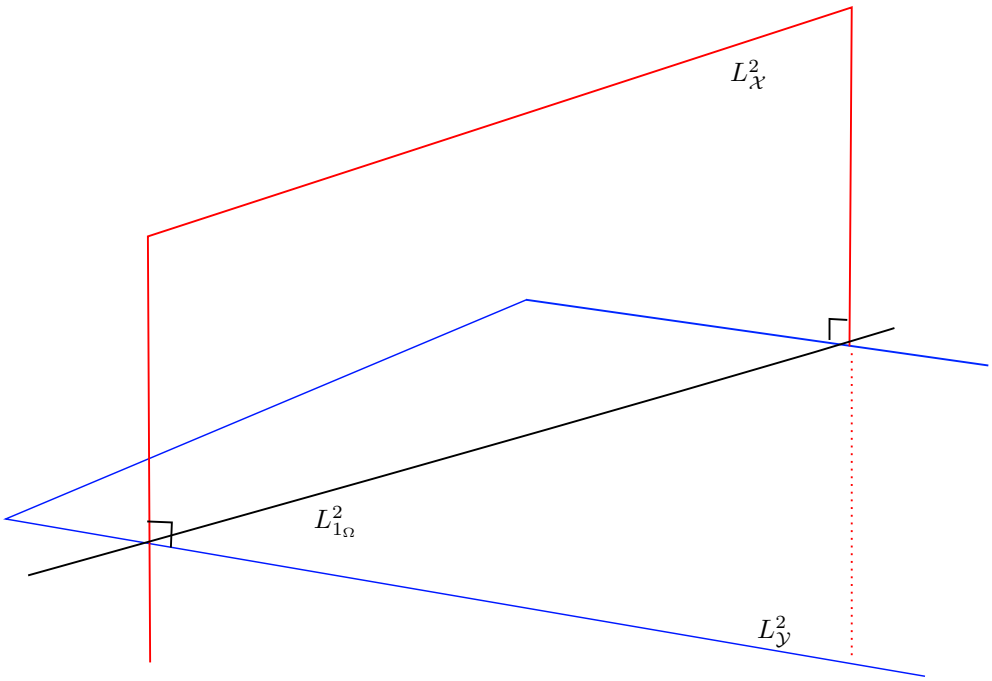


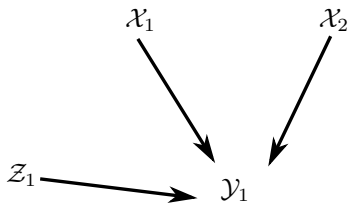
Figure 4.9: Les variables $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes si et seulement si les espaces de fonctions $L^2_{\mathcal{X}}$ et $L^2_{\mathcal{Y}}$ sont "orthogonaux le long de la droite des constantes $L^2_{1_\Omega}$". $L^2_{\mathcal{X}} \cap L^2_{\mathcal{Y}}$ est alors la droite des constantes $L^2_{1_\Omega}$ et la projection d'un point de $L^2_{\mathcal{X}}$ sur $L^2_{\mathcal{Y}}$ appartient alors à $L^2_{1_\Omega}$ de même que la projection d'un point de $L^2_{\mathcal{Y}}$ sur $L^2_{\mathcal{X}}$. La projection d'un point $\mathcal{Z}$ de $L^2_{\mathcal{X}}$ sur $L^2_{\mathcal{Y}}$ (et aussi d'un point $\mathcal{V}$ de $L^2_{\mathcal{Y}}$ sur $L^2_{\mathcal{X}}$) est donc une v.a. (constante) appartenant à $L^2_{1_\Omega}$ et la variable $\mathcal{Z} - E\{\mathcal{Z}\}$ est orthogonale à $L^2_{\mathcal{Y}}$. Pour tout couple $(\mathcal{Z}, \mathcal{V}) \in L^2_{\mathcal{X}} \times L^2_{\mathcal{Y}}$ on a donc $\text{cov}(\mathcal{Z}; \mathcal{V}) = 0$

4.4 ENSEMBLES DE VARIABLES ALÉATOIRES, CONSTRUCTION ET EXPLOITATION DE MODÈLES PROBABILISTES

Dans les sections précédentes nous avons supposé que la construction d'un modèle probabiliste est effectuée en définissant dans l'ordre un espace de probabilité $(\Omega, \mathcal{E}, P)$, puis en introduisant les variables aléatoires (quantités observables et quantités dont on veut déduire des propriétés) comme des fonctions (mesurables) définies sur Ω. Un fois que ces deux étapes sont réalisées, on peut alors appliquer les techniques d'inférence probabiliste qui permettent de déterminer, lois marginales et conjointes et relations entre ces variables aléatoires, sous la forme de lois conditionnelles, et de calculer des approximations de certaines variables aléatoires en fonction d'autres.

En pratique cette approche globale est cependant souvent difficile voire impossible à mettre en oeuvre. En effet, la situation plus courante est qu'on ne sait pas à l'avance quelles seront les propriétés du système étudié qu'on devra exploiter pour résoudre le problème considéré, et en particulier on ne sait souvent pas à l'avance quelles seront les variables aléatoires qui seront utiles pour sa résolution; comme le choix de $(\Omega, \mathcal{E}, P)$ dépend des quantités d'intérêt à modéliser, il n'est alors pas non plus possible de faire ce choix à l'avance. De manière plus embêtante encore, même si on peut déterminer pour un niveau de modélisation donné l'ensemble fondamental Ω (et une structure $\mathcal{E}$), définir une mesure de probabilité P qui obéit à la fois aux axiomes de Kolmogorov et qui est pratiquement réaliste est un problème difficile.

Nous allons expliquer dans cette section une démarche systématique de construction de modèle probabiliste qui construit la loi conjointe des variables aléatoires du problème, sans passer par la spécification explicite d'un ensemble fondamental Ω. Nous développons ces idées dans le cas particulier où toutes les variables aléatoires sont discrètes et finies, afin d'alléger les notations et de simplifier certains calculs.



(a) $\mathcal{Y}_1$ est une fonction des variables $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Z}_1$

$\mathcal{X}_1$	$\mathcal{X}_2$	$\mathcal{Z}_1$	$\mathcal{Y}_1$
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0

(b) Valeur de $\mathcal{Y}_1 = f_1(\mathcal{X}_1, \mathcal{X}_2, \mathcal{Z}_1)$

Figure 4.10: Double pile-ou-face bruité : (a) relations entre les variables aléatoires et (b) fonction f_1 qui définit $\mathcal{Y}_1$

4.4.1 Illustration: "Double pile-ou-face bruité"

Le problème du double pile-ou-face bruité est, comme nous le verrons, une abstraction du problème général de diagnostic (diagnostic médical, diagnostic de pannes, etc.). Nous illustrons sur base de ce problème la construction de modèles probabilistes.

On suppose qu'on lance deux pièces, dont on ne sait observer le résultat qu'au travers d'un certain nombre d'indications partielles et possiblement entachées d'erreurs.

La modélisation du problème nécessitera bien entendu l'introduction de deux variables aléatoires binaires disons $\mathcal{X}_1$ et $\mathcal{X}_2$, relatives aux deux pièces, ainsi que d'autant de variables aléatoires qu'il y a de quantités observables, disons $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$. Nous supposons que chacune de ces variables aléatoires $\mathcal{Y}_j$ est binaire et dépend de l'une, ou de l'autre, ou des deux variables $\mathcal{X}_i$ ainsi que d'une variable de chance $\mathcal{Z}_j$ qui vaudra 1 s'il y a erreur dans l'observation et 0 sinon. On suppose de plus que les deux pièces sont indépendantes, et que les erreurs d'observations le sont aussi et sont aussi indépendantes du lancer de pièces; cela se traduit par l'indépendance mutuelle des variables aléatoires $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Z}_1, \ldots, \mathcal{Z}_p$, et par le fait que chaque variable $\mathcal{Y}_j$ est une certaine fonction f_j de $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Z}_j$.

Nous commençons par construire le modèle dans une situation aussi simple que possible, c'est-à-dire lorsqu'il n'y a qu'une seule observation bruitée.

4.4.1.1 Construction, vérification et exploitation du modèle $(\Omega, \mathcal{E}, P)$

Pour illustrer la situation, supposons que les pièces sont équilibrées et que nous ne disposions dans un premier temps que d'une seule observation, à savoir une indication qui nous dit si les deux pièces sont tombées du même côté, avec une certaine probabilité d'erreur $p = 0.1$. Cela peut se schématiser par le graphique de relations et la table de vérité indiqués à la Figure 4.10. Dans ce modèle, lorsque $\mathcal{Z}_1 = 0$, la valeur $\mathcal{Y}_1 = 1$ (resp. $\mathcal{Y}_1 = 0$) indique correctement que les deux pièces sont tombées du même côté (resp. ne sont pas tombées du même côté). Par contre, lorsque $\mathcal{Z}_1 = 1$, la valeur de la variable $\mathcal{Y}_1$ est erronée. On a donc $P(\mathcal{Z}_1 = 1) = p = 0.1$.

Pour modéliser ce problème d'un point de vue probabiliste, on peut choisir un espace Ω qui permet de définir la probabilité de tous les événements pouvant être décrits au moyen d'affirmations relatives aux valeurs des quatre variables $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Z}_1, \mathcal{Y}_1$. Les variables étant binaires, elles peuvent prendre au total $2^4 = 16$ configurations conjointes, et Ω comportera donc 16 éléments (un élément pour chacune de ces configurations). Ensuite, après avoir associé à chacun de ces éléments une des configurations, il faut définir les probabilités de chaque élément de Ω, tout en respectant les hypothèses du problème, à savoir les indépendances mutuelles des variables $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Z}_1$, et le fait que la fonction f_1 définit $\mathcal{Y}_1$ en fonction des trois autres, le fait que $\mathcal{Z}_1$ vaut 1 avec une probabilité p, et que les pièces sont équilibrées. Un tel modèle est décrit à la Table 4.1; nous expliquerons plus loin comment les valeurs $P(\omega)$ ont été calculées de manière systématique.

ω	$P(\{\omega\})$	$\mathcal{X}_1(\omega)$	$\mathcal{X}_2(\omega)$	$\mathcal{Z}_1(\omega)$	$\mathcal{Y}_1(\omega)$
ω_1	0.000	0	0	0	0
ω_2	0.225	0	0	0	1
ω_3	0.025	0	0	1	0
ω_4	0.000	0	0	1	1
ω_5	0.225	0	1	0	0
ω_6	0.000	0	1	0	1
ω_7	0.000	0	1	1	0
ω_8	0.025	0	1	1	1
ω_9	0.225	1	0	0	0
ω_{10}	0.000	1	0	0	1
ω_{11}	0.000	1	0	1	0
ω_{12}	0.025	1	0	1	1
ω_{13}	0.000	1	1	0	0
ω_{14}	0.225	1	1	0	1
ω_{15}	0.025	1	1	1	0
ω_{16}	0.000	1	1	1	1

Table 4.1: Espace de probabilité et valeurs des variables aléatoires pour le double pile-ou-face bruité

On peut se convaincre que le modèle de la Table 4.1 respecte bien l'ensemble des hypothèses du problème. Par exemple, on peut vérifier que les variables respectent bien les lois de probabilités postulées et la relation déterministe entre les valeurs de $\mathcal{Y}_1$ et les trois autres variables:

- $P(\mathcal{X}_1(\omega) = 0) = 0.5 =$ somme des probabilités de ω_1 à ω_8 (pièce équilibrée).
- $P(\mathcal{X}_2(\omega) = 0) = 0.5$ (ω_1 à ω_4 et ω_9 à ω_{12}, pièce équilibrée).
- $P(\mathcal{Z}_1(\omega) = 0) = 0.9$ (somme des probabilités des réalisations $\omega_1, \omega_2, \omega_5, \omega_6, \omega_9, \omega_{10}, \omega_{13}, \omega_{14}$ qui correspondent bien aux cas où la valeur de $\mathcal{Y}_1$ reflète correctement si oui ou non les deux pièces sont tombées du même côté).
- Les configurations incompatibles avec la table de la Figure 4.10(b) sont toutes de probabilité nulle ($\omega_1, \omega_4, \omega_6$, $\omega_7, \omega_{10}, \omega_{11}, \omega_{13}, \omega_{16}$).

On peut également vérifier que les relations d'indépendance mutuelle postulées entre les variables $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Z}_1$ sont bien respectées par le modèle de la Table 4.1:

- $P(\mathcal{X}_1(\omega) = 0 \wedge \mathcal{X}_2(\omega) = 0) = 0.250 = P(\mathcal{X}_1(\omega) = 0)P(\mathcal{X}_2(\omega) = 0)$ (la première égalité étant obtenue en sommant les probabilités de ω_1 à ω_4).
- $P(\mathcal{X}_1(\omega) = 1 \wedge \mathcal{X}_2(\omega) = 0) = 0.250 = P(\mathcal{X}_1(\omega) = 1)P(\mathcal{X}_2(\omega) = 0)$, etc. etc.
- $P(\mathcal{X}_1(\omega) = 0 \wedge \mathcal{Z}_1(\omega) = 0) = 0.450 = P(\mathcal{X}_1(\omega) = 0)P(\mathcal{Z}_1(\omega) = 0)$, etc. etc.
- $P(\mathcal{X}_1(\omega) = x_1 \wedge \mathcal{X}_2(\omega) = x_2 \wedge \mathcal{Z}_1(\omega) = z_1) = P(\mathcal{X}_1(\omega) = x_1)P(\mathcal{X}_2(\omega) = x_2)P(\mathcal{Z}_1(\omega) = z_1)$ quelles que soient les valeurs $x_1 \in \{0, 1\}, x_2 \in \{0, 1\}, z_1 \in \{0, 1\}$.

A partir du modèle probabiliste de la Table 4.1, ont peut calculer la probabilité d'un événement quelconque défini au moyen d'une expression faisant intervenir les valeurs des variables aléatoires, simplement en sommant les probabilités des lignes pour lesquelles cette expression est vérifiée. Par exemple on calcule:

- $P(\mathcal{Y}_1 = 0) = 0.5 = P(\mathcal{Y}_1 = 1)$,
- $P(\mathcal{Y}_1 = 0, \mathcal{X}_1 = 0) = 0.250 = P(\mathcal{Y}_1 = 1, \mathcal{X}_1 = 0)$,
- $P(\mathcal{Y}_1 = 0, \mathcal{X}_1 = 1) = 0.250 = P(\mathcal{Y}_1 = 1, \mathcal{X}_1 = 1)$,

ce qui nous indique aussi que $\mathcal{Y}_1 \perp \mathcal{X}_1$; on peut vérifier de la même manière que $\mathcal{Y}_1 \perp \mathcal{X}_2$.

Enfin, on obtient les probabilités conditionnelles d'un événement conditionnellement à un autre événement en calculant la probabilité conjointe des deux événements et celle du second, et en divisant ensuite la première par la seconde. Par exemple, on obtient ainsi

- $P(\mathcal{Y}_1 = 0 | \mathcal{X}_1 = 0 \wedge \mathcal{X}_2 = 0) = \frac{0.025}{0.250} = 0.100$,
- $P(\mathcal{Y}_1 = 0 | \mathcal{X}_1 = 0 \wedge \mathcal{X}_2 = 1) = \frac{0.225}{0.250} = 0.900$,
- $P(\mathcal{Y}_1 = 0 | \mathcal{X}_1 = 1 \wedge \mathcal{X}_2 = 0) = \frac{0.225}{0.250} = 0.900$,
- $P(\mathcal{Y}_1 = 0 | \mathcal{X}_1 = 1 \wedge \mathcal{X}_2 = 1) = \frac{0.025}{0.250} = 0.100$,

ce qui reflète bien la probabilité d'erreur $p = 0.1$.

4.4.1.2 Modélisation directe et exploitation de la loi conjointe $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}$

L'information contenue dans la Table 4.1 est équivalente à la description de la loi conjointe $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}$, c'est-à-dire la donnée des 16 nombres $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}(x_1, x_2, z_1, y_1)$ avec $x_1, x_2, z_1, y_1 \in \{0, 1\}$.

Sans passer par la définition de l'espace $(\Omega, \mathcal{E}, P)$ on peut construire directement cette loi en introduisant progressivement les hypothèses du problème, de la manière suivante.

Premièrement, on exploite les informations *structurelles* du problème:

$$\begin{aligned} P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}(x_1, x_2, z_1, y_1) &= P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1}(x_1, x_2, z_1) P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{Z}_2,\mathcal{Z}_1}(y_1|x_1, x_2, z_1) && (4.86) \\ &= P_{\mathcal{X}_1}(x_1) P_{\mathcal{X}_2}(x_2) P_{\mathcal{Z}_1}(z_1) P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{Z}_2,\mathcal{Z}_1}(y_1|x_1, x_2, z_1) && (4.87) \end{aligned}$$

où nous avons tiré profit de l'indépendance mutuelle des variables $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Z}_1$, pour remplacer leur loi conjointe $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1}(x_1, x_2, z_1)$ par le produit $P_{\mathcal{X}_1}(x_1)P_{\mathcal{X}_2}(x_2)P_{\mathcal{Z}_1}(z_1)$ de leurs lois marginales.

Deuxièmement, on exploite les informations *quantitatives* du problème:

- $P_{\mathcal{X}_1}(x_1) = 0.5, \forall x_1 \in \{0, 1\}$ (première pièce équilibrée)
- $P_{\mathcal{X}_2}(x_2) = 0.5, \forall x_2 \in \{0, 1\}$ (seconde pièce équilibrée)
- $P_{\mathcal{Z}_1}(0) = 0.9$ et $P_{\mathcal{Z}_1}(1) = 0.1$ (probabilité d'erreur $p = 0.1$ associée à l'observation $\mathcal{Y}_1$)
- $P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{Z}_2,\mathcal{Z}_1}(y_1|x_1, x_2, z_1) = 1$ si $y_1 = f_1(x_1, x_2, z_1)$ et $P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{Z}_2,\mathcal{Z}_1}(y_1|x_1, x_2, z_1) = 0$ sinon (f_1 étant exprimée à la Figure 4.10(b)).

Mises ensemble, ces informations structurelles et quantitatives permettent de construire sans difficultés la Table 4.2 exprimant la loi conjointe, et qui contient essentiellement la même information que la Table 4.1.

La loi conjointe ainsi obtenue est unique et obéit par construction aux hypothèses du problème. On peut l'exploiter pour obtenir les lois conjointes de n'importe quel sous-ensemble de variables modélisées. Par exemple, on peut obtenir la loi conjointe $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$ en éliminant par marginalisation la variable $\mathcal{Z}_1$ de $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}$:

$$P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(x_1, x_2, y_1) = \sum_{z_1 \in \{0,1\}} P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}(x_1, x_2, z_1, y_1). \tag{4.88}$$

Cette loi conjointe est représentée à la Table 4.3. La dernière colonne est calculée par marginalisation, et l'avant dernière colonne est donc donnée par

$$P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{X}_2}(y_1|x_1, x_2) = \frac{P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(x_1, x_2, y_1)}{P_{\mathcal{X}_1,\mathcal{X}_2}(x_1, x_2)} = \frac{P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(x_1, x_2, y_1)}{P_{\mathcal{X}_1}(x_1)P_{\mathcal{X}_2}(x_2)}. \tag{4.89}$$

x_1	x_2	z_1	y_1	$P_{\mathcal{X}_1}$	$P_{\mathcal{X}_2}$	$P_{\mathcal{Z}_1}$	$P_{\mathcal{Y}_1\|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1}$	$P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}$
0	0	0	0	0.5	0.5	0.9	0.0	0.000
0	0	0	1	0.5	0.5	0.9	1.0	0.225
0	0	1	0	0.5	0.5	0.1	1.0	0.025
0	0	1	1	0.5	0.5	0.1	0.0	0.000
0	1	0	0	0.5	0.5	0.9	1.0	0.225
0	1	0	1	0.5	0.5	0.9	0.0	0.000
0	1	1	0	0.5	0.5	0.1	0.0	0.000
0	1	1	1	0.5	0.5	0.1	1.0	0.025
1	0	0	0	0.5	0.5	0.9	1.0	0.225
1	0	0	1	0.5	0.5	0.9	0.0	0.000
1	0	1	0	0.5	0.5	0.1	0.0	0.000
1	0	1	1	0.5	0.5	0.1	1.0	0.025
1	1	0	0	0.5	0.5	0.9	0.0	0.000
1	1	0	1	0.5	0.5	0.9	1.0	0.225
1	1	1	0	0.5	0.5	0.1	1.0	0.025
1	1	1	1	0.5	0.5	0.1	0.0	0.000

Table 4.2: Construction de $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}$ produit de $P_{\mathcal{X}_1}, P_{\mathcal{X}_2}, P_{\mathcal{Z}_1}$ et $P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1}$

x_1	x_2	y_1	$P_{\mathcal{X}_1}$	$P_{\mathcal{X}_2}$	$P_{\mathcal{Y}_1\|\mathcal{X}_1,\mathcal{X}_2}$	$P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$
0	0	0	0.5	0.5	0.9	0.225
0	0	1	0.5	0.5	0.1	0.025
0	1	0	0.5	0.5	0.9	0.225
0	1	1	0.5	0.5	0.1	0.025
1	0	0	0.5	0.5	0.9	0.225
1	0	1	0.5	0.5	0.1	0.025
1	1	0	0.5	0.5	0.9	0.225
1	1	1	0.5	0.5	0.1	0.025

Table 4.3: Loi $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$ obtenue par marginalisation de $\mathcal{Z}_1$ dans $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}$

Un autre façon de calculer $P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{X}_2}$, directement à partir de la Table 4.2 est la suivante:

$$P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{X}_2}(y_1|x_1,x_2) = \sum_{z_1\in\{0,1\}} P_{\mathcal{Z}_1,\mathcal{Y}_1|\mathcal{X}_1,\mathcal{X}_2}(z_1,y_1|x_1,x_2) \quad (4.90)$$

$$= \sum_{z_1\in\{0,1\}} P_{\mathcal{Z}_1|\mathcal{X}_1,\mathcal{X}_2}(z_1|x_1,x_2) P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1}(y_1|x_1,x_2,z_1) \quad (4.91)$$

$$= \sum_{z_1\in\{0,1\}} P_{\mathcal{Z}_1}(z_1) P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1}(y_1|x_1,x_2,z_1). \quad (4.92)$$

4.4.1.3 Enrichissement progressif du modèle

Après avoir construit la loi conjointe $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Z}_1,\mathcal{Y}_1}$ puis inféré la loi $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$ par marginalisation, voyons comment enrichir cette dernière en introduisant de nouvelles variables, si possible sans recommencer tout à zéro.

Pour illustrer cela, supposons que nous pouvons maintenant aussi disposer d'une seconde observation, que nous désignons par $\mathcal{Y}_2$, et qui nous donne une information (bruitée, avec un taux d'erreur $p' = 0.2$) en ce qui concerne le côté sur lequel est tombée la première pièce. Nous supposons que l'erreur de mesure ($\mathcal{Z}_2$) est indépendante des autres variables du problème et que l'observation ne perturbe pas les relations entre celles-ci. Pour aller directement au but, nous ne souhaitons pas cette fois-ci introduire explicitement dans le modèle les variables $\mathcal{Z}_1$ et $\mathcal{Z}_2$ indiquant s'il y a erreur de mesure au niveau de $\mathcal{Y}_1$ et/ou $\mathcal{Y}_2$.

x_1	x_2	y_1	y_2	$P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$	$P_{\mathcal{Y}_2\|\mathcal{X}_1}$	$P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}$
0	0	0	0	0.025	0.800	0.020
0	0	0	1	0.025	0.200	0.005
0	0	1	0	0.225	0.800	0.180
0	0	1	1	0.225	0.200	0.045
0	1	0	0	0.225	0.800	0.180
0	1	0	1	0.225	0.200	0.045
0	1	1	0	0.025	0.800	0.020
0	1	1	1	0.025	0.200	0.005
1	0	0	0	0.225	0.200	0.045
1	0	0	1	0.225	0.800	0.180
1	0	1	0	0.025	0.200	0.005
1	0	1	1	0.025	0.800	0.020
1	1	0	0	0.025	0.200	0.005
1	1	0	1	0.025	0.800	0.020
1	1	1	0	0.225	0.200	0.045
1	1	1	1	0.225	0.800	0.180

Table 4.4: Construction de la loi conjointe $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}$ produit des lois $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$ et $P_{\mathcal{Y}_2|\mathcal{X}_1}$

Nous voulons donc construire la loi $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}$ à partir de la loi $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$ et des nouvelles hypothèses concernant $\mathcal{Y}_2$. De façon générale on a évidemment

$$P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}(x_1,x_2,y_1,y_2) = P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(x_1,x_2,y_1)P_{\mathcal{Y}_2|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(y_2|x_1,x_2,y_1). \tag{4.93}$$

Comme notre prise de mesure n'influence pas le comportement des variables déjà modélisées, nous pouvons dans le second membre de cette équation réutiliser la loi $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$ précédemment déterminée. Il ne reste donc qu'à construire la loi conditionnelle $P_{\mathcal{Y}_2|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$.

Montrons d'abord que suite aux hypothèses concernant la mesure $\mathcal{Y}_2$, on a

$$P_{\mathcal{Y}_2|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(y_2|x_1,x_2,y_1) = P_{\mathcal{Y}_2|\mathcal{X}_1}(y_2|x_1). \tag{4.94}$$

En effet, par hypothèse on doit avoir

$$\begin{aligned}
P_{\mathcal{Y}_2|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(0|0,x_2,y_1) &= 0.8, &(4.95)\\
P_{\mathcal{Y}_2|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(0|1,x_2,y_1) &= 0.2, &(4.96)\\
P_{\mathcal{Y}_2|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(1|0,x_2,y_1) &= 0.2, &(4.97)\\
P_{\mathcal{Y}_2|\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(1|1,x_2,y_1) &= 0.8, &(4.98)
\end{aligned}$$

quelles que soient les valeurs x_2 de $\mathcal{X}_2$ et y_2 de $\mathcal{Y}_1$ ce qui implique que l'identité (4.94) est bien vérifiée.

La loi conjointe $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}$ peut donc être obtenue directement à partir de $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$ et de $P_{\mathcal{Y}_2|\mathcal{X}_1}$ comme indiqué à la Table 4.4. La connaissance de la loi conjointe $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}$ permet de répondre à de nouvelles questions.

Par exemple, sachant que nous observons $\mathcal{Y}_1 = 1$ (les deux pièces sont probablement tombées du même côté) et que $\mathcal{Y}_2 = 0$ (la première pièce est probablement tombée sur pile, i.e. $\mathcal{X}_1 = 0$), quelle est la probabilité que la seconde pièce soit tombée sur pile, i.e. quelle est la valeur de $P_{\mathcal{X}_2|\mathcal{Y}_1,\mathcal{Y}_2}(0|1,0)$?

On calcule à partir de la Table 4.4 que

$$P_{\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}(0,1,0) = 0.180 + 0.005 = 0.185,$$

et que

$$P_{\mathcal{Y}_1,\mathcal{Y}_2}(1,0) = 0.180 + 0.020 + 0.005 + 0.045 = 0.250,$$

et on en déduit que

$$P_{\mathcal{X}_2|\mathcal{Y}_1,\mathcal{Y}_2}(0|1,0) = \frac{0.185}{0.250} = 0.720.$$

4.4.1.4 Synthèse

La construction de modèles probabilistes utiles dans les applications nécessite souvent la manipulation d'un nombre important de variables aléatoires, et bien souvent il est nécessaire de faire évoluer le modèle au fur et à mesure en introduisant de nouvelles variables aléatoires, et/ou en mettant à jour les valeurs numériques postulées à un moment, telles que la probabilité de tomber sur pile d'une pièce ou la probabilité d'erreur d'une observation particulière dans notre exemple jouet ci-dessus.

Une façon systématique de construire un modèle probabiliste est la suivante:

- On se fixe pour objectif de modéliser la loi conjointe des variables aléatoires d'intérêt.
- On introduit progressivement les variables aléatoires dans le modèle, en exploitant à chaque étape lorsqu'on introduit une nouvelle variable aléatoire les indépendances (connaissances structurelles) entre la nouvelle variable et celles déjà introduites dans le problème, pour ajouter un nouveau facteur aussi simple que possible à multiplier avec la loi conjointe des variables déjà traitées.
- On introduit, au moment où l'on veut commencer à faire des inférences probabilistes, les données numériques et connaissances de relations fonctionnelles entre variables, pour définir précisément les lois élémentaires qui interviennent comme facteurs dans la loi conjointe des variables qui sont concernées par l'inférence probabiliste qu'on souhaite effectuer.
- On exploite ensuite les opérations de marginalisation pour simplifier les lois conjointes afin d'en éliminer des variables, et pour calculer les probabilités associées aux valeurs des variables d'intérêt.
- On exploite la formule de Bayes pour établir des lois conditionnelles entre variables qui ne peuvent pas directement être déduites des hypothèses décrivant le problème, et pour répondre à des questions spécifiques d'inférence probabiliste.

Cette démarche est complétée et très fortement enrichie par la théorie des modèles d'indépendances, et plus spécifiquement les modèles probabilistes graphiques, qui permettent de formaliser les indépendances conditionnelles qui peuvent et doivent coexister dans un modèle probabiliste donné, et ainsi de formuler des modèles cohérents et d'inférer de façon systématique les autres indépendances conditionnelles valides dans un modèle donné [Pea88].

Les modèles probabilistes graphiques sont également le point de départ pour développer des algorithmes informatiques efficaces permettant d'automatiser sur ordinateur l'inférence probabiliste, et sont encore aujourd'hui un domaine de recherche très actif mariant raisonnement probabiliste, statistiques, algorithmique et optimisation, notamment dans le cadre de l'avalanche de données rendues disponibles grâce à Internet.

Cette matière fait l'objet d'enseignements plus avancés. Dans les sections qui suivent, nous en mettons en évidence les idées principales dans le contexte où toutes les variables aléatoires sont discrètes, la généralisation aux variables continues pouvant se faire en principe avec quelques précautions mathématiques mais sans perturber l'intuition que nous souhaitons mettre en avant.

4.4.2 Marginalisation et conditionnement de lois de probabilités conjointes

Soit $\{\mathcal{X}_1, \ldots \mathcal{X}_p\}$ un ensemble de variables aléatoires discrètes définies sur un même espace de probabilité $(\Omega, \mathcal{E}, P)$ et soit $P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}(x_1, \ldots, x_p)$ la loi de probabilité conjointe de ces variables définie pour toute configuration $x_1, \ldots, x_p$ des valeurs des variables $\mathcal{X}_1, \ldots, \mathcal{X}_p$.

4.4.2.1 Elimination de variables par marginalisation

L'opération de marginalisation est une opération élémentaire qui consiste à éliminer une variable, disons $\mathcal{X}_k$ de $\{\mathcal{X}_1, \ldots \mathcal{X}_p\}$ afin de déduire la loi conjointe des autres variables. Nous noterons $\{\mathcal{X}_1, \ldots, [\mathcal{X}_k], \ldots, \mathcal{X}_p\}$ l'ensemble des autres variables.

La loi conjointe des variables $\{\mathcal{X}_1, \ldots, [\mathcal{X}_k], \ldots, \mathcal{X}_p\}$ s'obtient par le théorème des probabilités totales:

$$P_{\mathcal{X}_1,\ldots,[\mathcal{X}_k],\ldots,\mathcal{X}_p}(x_1,\ldots,[x_k],\ldots,x_p) = \sum_{x_k \in \mathcal{X}_k} P_{\mathcal{X}_1,\ldots,\mathcal{X}_k,\ldots,\mathcal{X}_p}(x_1,\ldots,x_k,\ldots,x_p), \tag{4.99}$$

où la somme porte sur toutes les valeurs possibles de la variable $\mathcal{X}_k$.

Partant de la description complète de la loi $P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}(x_1,\ldots,x_p)$, on peut appliquer l'opération de marginalisation successivement sur plusieurs variables, pour déduire la loi conjointe d'un quelconque sous-ensemble des variables $\{\mathcal{X}_1,\ldots\mathcal{X}_p\}$. L'ordre dans lequel on élimine les variables n'a pas d'importance étant donnée la commutativité de l'opération de sommation.

Retenons que l'opération de marginalisation permet d'obtenir à partir de $P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}(x_1,\ldots,x_p)$ la loi de probabilité conjointe de n'importe quel sous-ensemble de variables de $\{\mathcal{X}_1,\ldots\mathcal{X}_p\}$, et que nous pouvons choisir l'ordre d'élimination des variables qui nous arrange le mieux.

Par exemple, l'opération de marginalisation des variables $\mathcal{X}_1, \mathcal{X}_2$ sur la loi conjointe de $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}$ de la Table 4.4 nous fournit la loi conjointe $P_{\mathcal{Y}_1,\mathcal{Y}_2}$. On vérifiera que

$$\forall y_1, y_2 \in \{0,1\} : P_{\mathcal{Y}_1,\mathcal{Y}_2}(y_1,y_2) = 0.250.$$

Partant de $P_{\mathcal{Y}_1,\mathcal{Y}_2}$ on en déduit ensuite que $P_{\mathcal{Y}_1}(0) = P_{\mathcal{Y}_1}(1) = P_{\mathcal{Y}_2}(0) = P_{\mathcal{Y}_2}(1) = 0.5$, et aussi que les deux variables $\mathcal{Y}_1$ et $\mathcal{Y}_2$ sont indépendantes, puisque $\forall y_1, y_2 \in \{0,1\}$ on a

$$P_{\mathcal{Y}_1,\mathcal{Y}_2}(y_1,y_2) = P_{\mathcal{Y}_1}(y_1)P_{\mathcal{Y}_2}(y_2).$$

4.4.2.2 Construction de lois conditionnelles

La loi conditionnelle des variables $\{\mathcal{X}_1, \ldots, [\mathcal{X}_k], \ldots, \mathcal{X}_p\}$ étant donnée la variable $\mathcal{X}_k$ est définie par

$$P_{\mathcal{X}_1,\ldots,[\mathcal{X}_k],\ldots,\mathcal{X}_p|\mathcal{X}_k}(x_1,\ldots,[x_k],\ldots,x_p|x_k) \triangleq \frac{P_{\mathcal{X}_1,\ldots,\mathcal{X}_k,\ldots,\mathcal{X}_p}(x_1,\ldots,x_k,\ldots,x_p)}{P_{\mathcal{X}_k}(x_k)}. \tag{4.100}$$

Son obtention se résume par conséquent à une opération de marginalisation (les variables $\{\mathcal{X}_1, \ldots, [\mathcal{X}_k], \ldots, \mathcal{X}_p\}$, pour calculer $P_{\mathcal{X}_k}(x_k)$) et ensuite une division. Notons que la loi $P_{\mathcal{X}_1,\ldots,[\mathcal{X}_k],\ldots,\mathcal{X}_p|\mathcal{X}_k}(x_1,\ldots,[x_k],\ldots,x_p|x_k)$ n'est définie que pour x_k de probabilité $P_{\mathcal{X}_k}(x_k) > 0$.

On peut conditionner sur plusieurs variables, soit en une seule opération, soit en effectuant successivement cette opération dans un ordre quelconque. Cependant, lorsqu'on conditionne une loi conditionnelle sur une nouvelle variable, il faut le faire en divisant par la loi *conditionnelle* de cette variable par rapport aux variables préalablement dans le conditionnement. On a par exemple,

$$P_{\mathcal{X}_1,\ldots,[\mathcal{X}_{k_1},\mathcal{X}_{k_2}],\ldots\mathcal{X}_p|\mathcal{X}_{k_1},\mathcal{X}_{k_2}}(x_1,\ldots,[x_{k_1},x_{k_2}],\ldots,x_p|x_{k_1},x_{k_2}) \triangleq \frac{P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}(x_1,\ldots,x_p)}{P_{\mathcal{X}_{k_1},\mathcal{X}_{k_2}}(x_{k_1},x_{k_2})}. \tag{4.101}$$

Cette loi peut-être obtenue en conditionnnant d'abord sur $\mathcal{X}_{k_1}$ dans $P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}$, puis en conditionnant sur $\mathcal{X}_{k_2}$ dans $P_{\mathcal{X}_1,\ldots,[\mathcal{X}_{k_1}],\ldots,\mathcal{X}_p|\mathcal{X}_{k_1}}$, ou bien dans l'autre ordre. En effet, on a

$$\frac{P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}(x_1,\ldots,x_p)}{P_{\mathcal{X}_{k_1},\mathcal{X}_{k_2}}(x_{k_1},x_{k_2})} = \frac{P_{\mathcal{X}_1,\ldots,[\mathcal{X}_{k_1}],\ldots,\mathcal{X}_p|\mathcal{X}_{k_1}}(x_1,\ldots,[x_{k_1}],\ldots,x_p|x_{k_1})}{P_{\mathcal{X}_{k_2}|\mathcal{X}_{k_1}}(x_{k_2}|x_{k_1})} \tag{4.102}$$

$$= \frac{P_{\mathcal{X}_1,\ldots,[\mathcal{X}_{k_2}],\ldots,\mathcal{X}_p|\mathcal{X}_{k_2}}(x_1,\ldots,[x_{k_2}],\ldots,x_p|x_{k_2})}{P_{\mathcal{X}_{k_1}|\mathcal{X}_{k_2}}(x_{k_1}|x_{k_2})}, \tag{4.103}$$

puisque $P_{\mathcal{X}_{k_1},\mathcal{X}_{k_2}}(x_{k_1},x_{k_2}) = P_{\mathcal{X}_{k_2}|\mathcal{X}_{k_1}}(x_{k_2}|x_{k_1})P_{\mathcal{X}_{k_1}}(x_{k_1}) = P_{\mathcal{X}_{k_1}|\mathcal{X}_{k_2}}(x_{k_1}|x_{k_2})P_{\mathcal{X}_{k_2}}(x_{k_2})$.

4.4.2.3 Récapitulation

En résumé, les opérations de marginalisation et de conditionnement permettent de déduire de la loi jointe $P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}$ les lois marginales et conditionnelles portant sur des sous-ensembles quelconques de variables de $\{\mathcal{X}_1,\ldots\mathcal{X}_p\}$.

Notons qu'on peut effectuer les opérations de marginalisation et de conditionnement dans un ordre quelconque, les propriétés du calcul de probabilités assurant que le résultat final est indépendant de l'ordre choisi.

4.4.3 Exploitation de la notion d'indépendance conditionnelle

Dans cette section nous développons la notion très importante en pratique d'indépendance conditionnelle entre (ensembles de) variables aléatoires.

4.4.3.1 Indépendance conditionnelle de deux variables étant donnée une troisième

On dit que la variable $\mathcal{X}$ est indépendante de la variable $\mathcal{Y}$ conditionnellement à la variable $\mathcal{Z}$, ce que l'on note par $\mathcal{X} \perp \mathcal{Y}|\mathcal{Z}$, si

$$\forall x, y : P_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}(x, y|z) = P_{\mathcal{X}|\mathcal{Z}}(x|z)P_{\mathcal{Y}|\mathcal{Z}}(y|z), \tag{4.104}$$

pour toute valeur de z telle que $P_{\mathcal{Z}}(z) > 0$. (Les lois conditionnelles ne sont pas définies lorsque $P_{\mathcal{Z}}(z) = 0$.)

Lorsque $\mathcal{X} \perp \mathcal{Y}|\mathcal{Z}$ on a donc $P_{\mathcal{X},\mathcal{Y},\mathcal{Z}}(x, y, z) = P_{\mathcal{Z}}(z)P_{\mathcal{X}|\mathcal{Z}}(x|z)P_{\mathcal{Y}|\mathcal{Z}}(y|z)$, si $P_{\mathcal{Z}}(z) > 0$ et cette identité reste valable pour un choix arbitraire de $P_{\mathcal{X}|\mathcal{Z}}(x|z)$ et $P_{\mathcal{Y}|\mathcal{Z}}(y|z)$ lorsque $P_{\mathcal{Z}}(z) = 0$. Cela nous permet d'écrire que

$$\mathcal{X} \perp \mathcal{Y}|\mathcal{Z} \Leftrightarrow P_{\mathcal{X},\mathcal{Y},\mathcal{Z}} \overset{p.s.}{=} P_{\mathcal{Z}}P_{\mathcal{X}|\mathcal{Z}}P_{\mathcal{Y}|\mathcal{Z}}, \tag{4.105}$$

où l'égalité est interprétée de façon fonctionnelle au sens "presque sûrement", et reste valable quels que soient les choix arbitraires concernant les lois conditionnelles pour les valeurs de z telles que $P_{\mathcal{Z}}(z) = 0$.

Notons aussi que la définition (4.104) implique que si $\mathcal{X} \perp \mathcal{Y}|\mathcal{Z}$ alors

$$\forall x : P_{\mathcal{X}|\mathcal{Y},\mathcal{Z}}(x|y, z) = P_{\mathcal{X}|\mathcal{Z}}(x|z) \tag{4.106}$$

pour tout couple de valeurs y, z tel que $P_{\mathcal{Y},\mathcal{Z}}(y, z) > 0$, ainsi que

$$\forall y : P_{\mathcal{Y}|\mathcal{X},\mathcal{Z}}(y|x, z) = P_{\mathcal{Y}|\mathcal{Z}}(y|z) \tag{4.107}$$

pour tout couple de valeurs x, z tel que $P_{\mathcal{X},\mathcal{Z}}(x, z) > 0$.

Il s'en suit que lorsque $\mathcal{X} \perp \mathcal{Y}|\mathcal{Z}$, l'on a aussi $P_{\mathcal{X},\mathcal{Y},\mathcal{Z}} \overset{p.s.}{=} P_{\mathcal{Y}}P_{\mathcal{Z}|\mathcal{Y}}P_{\mathcal{X}|\mathcal{Z}}$ et $P_{\mathcal{X},\mathcal{Y},\mathcal{Z}} \overset{p.s.}{=} P_{\mathcal{X}}P_{\mathcal{Z}|\mathcal{X}}P_{\mathcal{Y}|\mathcal{Z}}$.

En résumé,

$$\begin{aligned}
\mathcal{X} \perp \mathcal{Y}|\mathcal{Z} \quad &\Leftrightarrow \\
\forall x, y, z : P_{\mathcal{X},\mathcal{Y},\mathcal{Z}}(x, y, z) \quad &= \quad P_{\mathcal{Z}}(z)P_{\mathcal{X}|\mathcal{Z}}(x|z)P_{\mathcal{Y}|\mathcal{Z}}(y|z) && (4.108)\\
&= \quad P_{\mathcal{Y}}(y)P_{\mathcal{Z}|\mathcal{Y}}(z|y)P_{\mathcal{X}|\mathcal{Z}}(x|z) && (4.109)\\
&= \quad P_{\mathcal{X}}(x)P_{\mathcal{Z}|\mathcal{X}}(z|x)P_{\mathcal{Y}|\mathcal{Z}}(y|z). && (4.110)
\end{aligned}$$

L'équation (4.108) est obtenue en multipliant les deux membres de (4.104) par $P_{\mathcal{Z}}(z)$, puis le passage à (4.109) est obtenu en observant que $P_{\mathcal{Z}}(z)P_{\mathcal{Y}|\mathcal{Z}}(y|z) = P_{\mathcal{Y},\mathcal{Z}}(y, z) = P_{\mathcal{Y}}(y)P_{\mathcal{Z}|\mathcal{Y}}(z|y)$, et enfin le passage à (4.110) se déduit de $P_{\mathcal{Z}}(z)P_{\mathcal{X}|\mathcal{Z}}(x|z) = P_{\mathcal{X},\mathcal{Z}}(x, z) = P_{\mathcal{X}}(x)P_{\mathcal{Z}|\mathcal{X}}(x|y)$.

Notons que si $\mathcal{Z}$ est une variable constante, alors elle est indépendante de tout autre ensemble de variables, et l'indépendance conditionnelle $\mathcal{X} \perp \mathcal{Y}|\mathcal{Z}$ équivaut dans ce cas à l'indépendance simple $\mathcal{X} \perp \mathcal{Y}$.

4.4.3.2 Indépendance conditionnelle entre ensembles de variables

La notion d'indépendance conditionnelle se généralise à des ensembles $\{\mathcal{X}_i\}_1^p, \{\mathcal{Y}_j\}_1^q$ et $\{\mathcal{Z}_k\}_1^r$ de variables aléatoires. On dit que les variables $\{\mathcal{X}_i\}_1^p$ sont (conjointement) indépendantes des variables $\{\mathcal{Y}_j\}_1^q$ condition-

nellement aux variables de $\{\mathcal{Z}_k\}_1^r$ (noté par $\{\mathcal{X}_i\}_1^p \perp \{\mathcal{Y}_j\}_1^q | \{\mathcal{Z}_k\}_1^r$), si $\forall x_1, \ldots, x_p, y_1, \ldots, y_q$ on a

$$P_{\mathcal{X}_1,\ldots,\mathcal{X}_p,\mathcal{Y}_1,\ldots,\mathcal{Y}_q|\mathcal{Z}_1,\ldots,\mathcal{Z}_r}(x_1,\ldots,x_p,y_1,\ldots,y_q|z_1,\ldots,z_r)$$
$$=$$
$$P_{\mathcal{X}_1,\ldots,\mathcal{X}_p|\mathcal{Z}_1,\ldots,\mathcal{Z}_r}(x_1,\ldots,x_p|z_1,\ldots,z_r)P_{\mathcal{Y}_1,\ldots,\mathcal{Y}_q|\mathcal{Z}_1,\ldots,\mathcal{Z}_r}(y_1,\ldots,y_q|z_1,\ldots,z_r),$$

pour toute configuration $z_1, \ldots, z_r$ telle que $P_{\mathcal{Z}_1,\ldots,\mathcal{Z}_r}(z_1,\ldots,z_r) > 0$.

Il est important de noter les propriétés suivantes:

- Si $\{\mathcal{X}_i\}_1^p \perp \{\mathcal{Y}_j\}_1^q | \{\mathcal{Z}_k\}_1^r$, alors $\forall i = 1, \ldots, p, j = 1, \ldots, q$ on a aussi $\mathcal{X}_i \perp \mathcal{Y}_j | \{\mathcal{Z}_k\}_1^r$, et plus généralement tout sous-ensemble de variables de $\{\mathcal{X}_i\}_1^p$ est indépendant de tout sous-ensemble de variables de $\{\mathcal{Y}_j\}_1^q$ conditionnellement à toutes les variables de $\{\mathcal{Z}_k\}_1^r$.
- Si $\{\mathcal{X}_i\}_1^p \perp \{\mathcal{Y}_j\}_1^q | \{\mathcal{Z}_k\}_1^r$, alors tout ensemble de fonctions des $\{\mathcal{X}_i\}_1^p$ est indépendant de tout ensemble de fonctions des $\{\mathcal{Y}_j\}_1^q$ conditionnellement à toutes les variables de $\{\mathcal{Z}_k\}_1^r$.

Par contre, il est tout aussi important de garder à l'esprit que $\{\mathcal{X}_i\}_1^p \perp \{\mathcal{Y}_j\}_1^q | \{\mathcal{Z}_k\}_1^r$ n'implique pas que $\{\mathcal{X}_i\}_1^p \perp \{\mathcal{Y}_j\}_1^q | \mathcal{Z}_k$ (ou plus généralement conditionnellement à un sous-ensemble propre de $\{\mathcal{Z}_k\}_1^r$, ou à ensemble quelconque de fonctions des variables de $\{\mathcal{Z}_k\}_1^r$).

En fait, "conditionner sur une ou plusieurs variables" peut rendre indépendantes des variables qui ne sont pas indépendantes, ou bien rendre dépendantes des variables qui sont indépendantes.

4.4.3.3 Expression de la loi jointe comme produit de facteurs simples

La loi de probabilité conjointe d'un certain nombre de variables aléatoires, disons $\{\mathcal{X}_i\}_1^p$ peut se **factoriser** de manière générique sous la forme

$$P_{\mathcal{X}_1,\ldots,\mathcal{X}_p} = P_{\mathcal{X}_1} P_{\mathcal{X}_2|\mathcal{X}_1} P_{\mathcal{X}_3|\mathcal{X}_1,\mathcal{X}_2} \cdots P_{\mathcal{X}_p|\mathcal{X}_1,\ldots,\mathcal{X}_{p-1}}, \tag{4.111}$$

et cette factorisation, dont les facteurs dépendent de l'ordre dans lequel on prend les variables, est valide quel que soit cet ordre.

Pour un ordre fixé de factorisation, on peut exploiter les relations d'indépendances conditionnelles entre les variables du problème, afin de simplifier cette expression. Par exemple, si on peut assurer que $\mathcal{X}_i \perp \mathcal{X}_1 | \{\mathcal{X}_i\}_2^{i-1}$ on peut remplacer le facteur $P_{\mathcal{X}_i|\mathcal{X}_1,\ldots,\mathcal{X}_{i-1}}$ par le facteur $P_{\mathcal{X}_i|\mathcal{X}_2,\ldots,\mathcal{X}_{i-1}}$ dans cette formule.

Afin de simplifier au maximum, on peut essayer de déterminer pour chaque facteur $P_{\mathcal{X}_i|\mathcal{X}_1,\ldots,\mathcal{X}_{i-1}}$, un sous-ensemble de taille minimale de variables de l'ensemble $\{\mathcal{X}_i\}_1^{i-1}$, disons $Pa(\mathcal{X}_i, \{\mathcal{X}_1\}_1^{i-1})$ tel que

$$\mathcal{X}_i \perp (\{\mathcal{X}_1\}_1^{i-1} \setminus Pa(\mathcal{X}_i, \{\mathcal{X}_1\}_1^{i-1})) | Pa(\mathcal{X}_i, \{\mathcal{X}_1\}_1^{i-1}).$$

Une fois qu'on a déterminé pour chaque variable $\mathcal{X}_i$ un tel ensemble $Pa(\mathcal{X}_i, \{\mathcal{X}_1\}_1^{i-1})$, on peut représenter la loi jointe sous la forme suivante:

$$P_{\mathcal{X}_1,\ldots,\mathcal{X}_p} = P_{\mathcal{X}_1} P_{\mathcal{X}_2|Pa(\mathcal{X}_2,\{\mathcal{X}_i\}_1^1)} P_{\mathcal{X}_3|Pa(\mathcal{X}_3,\{\mathcal{X}_i\}_1^2)} \cdots P_{\mathcal{X}_p|Pa(\mathcal{X}_p,\{\mathcal{X}_i\}_1^{p-1})}. \tag{4.112}$$

Par exemple, dans le cas du double pile-ou-face bruité avec deux observations, on peut se convaincre qu'en prenant les variables dans l'ordre $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}_1, \mathcal{Y}_2$ on peut factoriser la loi conjointe comme suit:

$$P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2} = P_{\mathcal{X}_1} P_{\mathcal{X}_2} P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{X}_2} P_{\mathcal{Y}_2|\mathcal{X}_1}, \tag{4.113}$$

puisque $\mathcal{X}_1 \perp \mathcal{X}_2$ et que $\mathcal{Y}_2 \perp \{\mathcal{X}_2, \mathcal{Y}_1\} | \mathcal{X}_1$. De plus, il n'est pas possible de simplifier plus encore les différents facteurs de cette loi, puisque (nous conseillons au lecteur de vérifier)

- le premier et le second facteur ne peuvent pas être plus simplifiés en exploitant des indépendances, puisqu'il ne font intervenir chacun qu'une seule variable;

- le troisième facteur ne peut pas non plus être simplifié, puisque $\mathcal{Y}_1 \not\perp \mathcal{X}_1|\mathcal{X}_2$, $\mathcal{Y}_1 \not\perp \mathcal{X}_2|\mathcal{X}_1$, et $\mathcal{Y}_1 \not\perp \{\mathcal{X}_1, \mathcal{X}_2\}$;
- enfin, le quatrième facteur est lui aussi minimal en termes de nombre de variables retenues, puisque $\mathcal{Y}_2 \not\perp \mathcal{X}_1$.

D'autre part, si pour ce même problème nous choisissons de factoriser la loi jointe dans l'ordre $\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{X}_2, \mathcal{X}_1$ puis de la simplifier nous obtenons (nous conseillons au lecteur de s'en convaincre):

$$P_{\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{X}_2,\mathcal{X}_1} = P_{\mathcal{Y}_1} P_{\mathcal{Y}_2} P_{\mathcal{X}_2|\mathcal{Y}_1,\mathcal{Y}_2} P_{\mathcal{X}_1|\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{X}_2}. \tag{4.114}$$

Cette factorisation conduit à une expression plus compliquée que la précédente, puisque le quatrième facteur fait intervenir 4 variables au lieu de 2. On constate donc que l'ordre dans lequel les variables sont considérées pour construire une factorisation minimale a une influence sur la complexité des facteurs.

4.4.3.4 Indépendances numériquement instables

Lorsqu'on modélise un problème en construisant la loi de probabilité conjointe des variables d'intérêt, comme nous l'avons fait pour notre exemple du double pile-ou-face bruité, certaines indépendances (simples ou conditionnelles) sont la conséquence directe des hypothèses structurelles du problème, alors que d'autres peuvent être le fruit du choix des paramètres numériques (p.ex. les probabilités de tomber sur pile ou face d'une pièce, dans notre exemple). Ce deuxième type d'indépendances ne résiste pas à une petite perturbation des valeurs numériques, et on peut donc les qualifier de numériquement instables. Nous allons illustrer cela dans le cas de notre exemple.

Rappelons que nous avons montré plus haut que si les pièces sont équilibrées, nous avons $\mathcal{Y}_1 \perp \mathcal{X}_1$ et $\mathcal{Y}_1 \perp \mathcal{X}_2$. Analysons ceci de plus près, en supposant que les pièces ne sont plus parfaitement équilibrées. En d'autres termes, supposons que $P_{\mathcal{X}_1}(0) = 0.5 + \epsilon_1$ et $P_{\mathcal{X}_2}(0) = 0.5 + \epsilon_2$. Supposons par ailleurs que la valeur de la probabilité d'erreur p associée à l'observation de $\mathcal{Y}_1$ vaut $0.1 + \epsilon_3$. Pour fixer les idées prenons $\epsilon_1 = \epsilon_2 = \epsilon_3 = 0.05$.

Avec ces valeurs nous obtenons que

$$P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(0,0,0) = (0.5+\epsilon_1)(0.5+\epsilon_2)(0.1+\epsilon_3) = 0.045375 \tag{4.115}$$

et

$$P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}(0,1,0) = (0.5+\epsilon_1)(0.5-\epsilon_2)(0.9-\epsilon_3) = 0.210375, \tag{4.116}$$

et donc

$$P_{\mathcal{X}_1,\mathcal{Y}_1}(0,0) = (0.5+\epsilon_1)\left((0.5+\epsilon_2)(0.1+\epsilon_3) + (0.5-\epsilon_2)(0.9-\epsilon_3)\right) = 0.25575. \tag{4.117}$$

On calcule de même que

$$P_{\mathcal{X}_1,\mathcal{Y}_1}(1,0) = (0.5-\epsilon_1)\left((0.5+\epsilon_2)(0.9-\epsilon_3) + (0.5-\epsilon_2)(0.1+\epsilon_3)\right) = 0.24075. \tag{4.118}$$

En sommant ces deux probabilités on obtient que $P_{\mathcal{Y}_1}(0) = 0.4965$. On en déduit que $\mathcal{Y}_1 \not\perp \mathcal{X}_1$, puisque $P_{\mathcal{Y}_1|\mathcal{X}_1}(0|1) = \frac{P_{\mathcal{X}_1,\mathcal{Y}_1}(1,0)}{P_{\mathcal{X}_1}(1)} = \frac{0.24075}{0.5-0.05} = 0.535$, alors que $P_{\mathcal{Y}_1}(0) = 0.4965$. On peut se convaincre que pour presque toutes les valeurs de $P_{\mathcal{X}_1}(0)$, $P_{\mathcal{X}_2}(0)$ et p on a $\mathcal{Y}_1 \not\perp \mathcal{X}_1$.

On peut montrer, de la même façon, que dans la version étendue du problème on a aussi $\mathcal{Y}_1 \not\perp \mathcal{Y}_2$, pour presque toutes les valeurs de $P_{\mathcal{X}_1}(0)$, $P_{\mathcal{X}_2}(0)$ et p. Il s'en suit que la factorisation de l'équation (4.114) est fortuitement liée au choix des paramètres (pièces équilibrées du problème), et que si on impose l'ordre $\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{X}_2, \mathcal{X}_1$ la seule factorisation stable vis-à-vis du choix de ces paramètres est la factorisation

$$P_{\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{X}_2,\mathcal{X}_1} = P_{\mathcal{Y}_1} P_{\mathcal{Y}_2|\mathcal{Y}_1} P_{\mathcal{X}_2|\mathcal{Y}_1,\mathcal{Y}_2} P_{\mathcal{X}_1|\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{X}_2}, \tag{4.119}$$

qui ne met en évidence aucune indépendance conditionnelle.

Par contre, la factorisation de l'équation (4.113) reste valable quels que soient les choix des paramètres $P_{\mathcal{X}_1}(0)$, $P_{\mathcal{X}_2}(0)$ et p, car elle ne repose que sur les hypothèses structurelles du problème.

(a) $P_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_1} = P_{\mathcal{X}_1} P_{\mathcal{X}_2} P_{\mathcal{Y}_1|\mathcal{X}_1,\mathcal{X}_2} P_{\mathcal{Y}_2|\mathcal{X}_1}$ (b) $P_{\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{X}_2,\mathcal{X}_1} = P_{\mathcal{Y}_1} P_{\mathcal{Y}_2|\mathcal{Y}_1} P_{\mathcal{X}_2|\mathcal{Y}_1,\mathcal{Y}_2} P_{\mathcal{X}_1|\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{X}_2}$

Figure 4.11: Deux graphes de factorisation de la loi jointe pour le double pile-ou-face bruité

4.4.3.5 Graphes de factorisation et réseaux bayesiens

La factorisation d'une loi de probabilité peut se représenter de façon graphique, de la manière suivante:

- on associe à chaque variable un noeud dans le graphe de représentation,
- puis on choisit un ordre de factorisation,
- enfin on associe à chaque noeud représentant une variable les noeuds parents qui correspondent aux variables de conditionnement qui sont retenues dans le facteur associé à cette variable, compte tenu de l'ordre de factorisation choisi.

Il est préférable de représenter la structure du graphe en ne se basant que sur les hypothèses structurelles du problème, de façon à garantir sa stabilité par rapport à différents choix possibles des paramètres numériques.

La Figure 4.11 illustre cette idée pour notre exemple du double pile-ou-face pour les deux ordres de factorisation analysés ci-dessus. Dans le graphe de droite, nous avons mis en évidence la flèche qui disparaîtrait si les pièces étaient parfaitement équilibrées. Dans le graphe de gauche, on voit que les indépendances structurelles se traduisent par l'absence d'un certain nombre de flèches.

Ces représentations graphiques de la factorisation d'une loi de probabilité s'appellent des "réseaux bayésiens". Il s'agit d'outils de modélisation très intéressants en pratique, et qui offrent un support très riche pour effectuer des opérations d'inférence probabiliste de manière automatique [Pea88]. Leur étude plus approfondie fait l'objet d'autres enseignements.

4.4.4 Extension au cas des variables continues

Les idées qui ont été discutées dans cette section peuvent évidemment être étendues au cas où on souhaite modéliser la densité conjointe d'un ensemble de variables continues, en exploitant les propriétés d'indépendance conditionnelle pour représenter cette densité conjointe sous la forme d'un produit de facteurs faisant chacun intervenir un nombre minimal de variables et se prêtant à l'inférence probabiliste (marginalisation et conditionnement).

Par ailleurs, dans de nombreux problèmes on est amené à modéliser les interactions entre variables aléatoires dont certaines sont discrètes et d'autres continues. Dans ce cas encore, les mêmes idées s'appliquent pour modéliser des "densités" conjointes, portant à la fois sur des variables continues et discrètes.

4.5 PROBLÈMES ET APPLICATIONS

Nous re-discutons dans cette section les grandes classes de problèmes types d'inférence probabiliste qui font appel aux notions introduites dans ce chapitre, en les illustrant par des problèmes génériques rencontrés dans la pratique, puis nous revenons brièvement sur la méthode de Monte-Carlo.

4.5.1 Problèmes types d'inférence probabiliste

4.5.1.1 Analyse de la fiabilité (problème type de prédiction)

A la section 3.12.1 nous avons introduit les problèmes d'analyse de fiabilité et d'évaluation économique de solutions techniques, en les modélisant à partir d'une relation entrée-sortie faisant intervenir une seule variable

d'entrée et une seule variable de sortie. Les notions introduites dans le présent chapitre permettent d'étendre ce type d'analyse au cas plus réaliste où plusieurs variables d'entrée doivent être prises en compte, et lorsqu'on souhaite analyser plusieurs variables de sortie.

Par exemple, en analyse de la fiabilité, on souhaite généralement considérer l'ensemble des causes possibles d'une panne, ce qui se traduit par un besoin de modéliser la loi conjointe de plusieurs variables aléatoires "d'entrée". Par exemple, dans un système technique (usine chimique, réseau électrique, moyen de transport) on souhaitera modéliser les effets externes (tremblements de terre, tempêtes, variations de la température, etc.) aussi bien que les phénomènes internes (érosion, fatigue, ruptures de composants, etc.). Par ailleurs, dans une étude réelle, on souhaite généralement étudier le comportement du système par le biais de plusieurs "variables de sortie", par exemple la durée et l'ampleur de la panne et ses impacts écologiques, sociaux et économiques.

La résolution de ces problèmes de prédiction peut-être effectuée en trois étapes :

- Modélisation de la loi conjointe des variables d'entrée $\mathcal{X}_1, \ldots, \mathcal{X}_p$.
- Pour chaque variable de sortie $\mathcal{Y}_i, \forall i = 1, \ldots, n$, modélisation de la relation fonctionnelle

$$y_i = f_i(x_1, \ldots, x_p, \lambda_1, \ldots, \lambda_K).$$

- Détermination des lois marginales des variables $\mathcal{Y}_i$, et éventuellement de leur loi conjointe à partir des informations précédentes.

La résolution du problème pour différentes valeurs des variables de design $\lambda_1, \ldots, \lambda_K$, permet de comparer plusieurs designs alternatifs, et en principe de choisir la combinaison qui optimise les performances moyennes du système selon le critère de performances choisi.

4.5.1.2 Estimation d'état (problème type de diagnostic)

Le problème de l'estimation d'état est un problème de diagnostic qu'on rencontre de façon générique dans de nombreuses applications. Ce problème peut se formuler de la manière suivante : étant donné un système sur lequel on peut effectuer un série de mesures, comment déterminer les valeurs de certaines grandeurs importantes (et non directement mesurables) à partir des mesures possibles.

Par exemple, les réacteurs chimiques sont équipés de nombreux capteurs, permettant de mesurer le débit, la pression et la température à certains endroits accessibles. Ces instruments fournissent des informations (généralement un peu erronées) de certaines grandeurs physiques qui sont liées à l'état "interne" du réacteur (équilibre, température interne, rendement, etc.). Bien qu'on ne sache pas mesurer directement ces grandeurs internes, on souhaite exploiter les mesures pour en estimer les valeurs les plus probables.

Mathématiquement, ce problème peut être abordé de la manière suivante :

- On établit le modèle du système, qui relie les valeurs des variables externes, $\mathcal{Y}_i, \forall i = 1, \ldots, n$, aux valeurs des variables internes $\mathcal{X}_1, \ldots, \mathcal{X}_p$, sous la forme

$$y_i = h_i(x_1, \ldots, x_p, \lambda_1, \ldots, \lambda_K).$$

- On décrit les systèmes de mesure des variables externes en modélisant les erreurs de mesure, c'est-à-dire en écrivant pour chaque variable externe un modèle du type

$$\mathcal{Y}_{i,\text{obs}} = \mathcal{Y}_i + \mathcal{Z}_i,$$

  où $\mathcal{Z}_i$ est une variable aléatoire qui représente les erreurs de mesure de cette grandeur.

- On définit une loi de probabilité $P_{\mathcal{Z}_1,\ldots,\mathcal{Z}_n}$ pour les erreurs de mesure (généralement, en supposant que celles-ci sont indépendantes, et gaussiennes et de moyenne nulle).
- On définit une loi de probabilité a priori $P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}$ relative à l'état interne du système, par exemple une loi uniforme sur le domaine de fonctionnement possible du système.

- On construit un estimateur, c'est-à-dire un algorithme qui calcule les valeurs estimées

$$\hat{x}_i = \text{Algo}(y_{1,\text{obs}}, \ldots, y_{n,\text{obs}}; P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}; P_{\mathcal{Z}_1,\ldots,\mathcal{Z}_n}; h_1, \ldots, h_n),$$

  à partir des valeurs mesurées $y_{1,\text{obs}}, \ldots, y_{n,\text{obs}}$ et du modèle probabiliste (spécifié de façon conjointe par $P_{\mathcal{X}_1,\ldots,\mathcal{X}_p}$, $P_{\mathcal{Z}_1,\ldots,\mathcal{Z}_n}$ et $h_1, \ldots, h_n$).

Une façon de spécifier un algorithme d'estimation d'état consiste à définir un critère de précision numérique : par exemple on pourrait (et on le fait très souvent en pratique) vouloir que l'erreur quadratique moyenne entre les valeurs estimées $\hat{x}_i$ et les vraies valeurs x_i soit minimale.

Partant des notions introduites dans ce chapitre on peut montrer qu'un algorithme qui calcule $\hat{x}_i$ sous la forme de l'espérance conditionnelle de $\mathcal{X}_i$ étant données les observations $\mathcal{Y}_{1,\text{obs}}, \ldots, \mathcal{Y}_{n,\text{obs}}$ est la solution à ce problème. Cet algorithme réalise essentiellement une projection des variables $\mathcal{X}_i$ sur l'espace des fonctions des grandeurs observables $L^2_{\mathcal{Y}_{1,\text{obs}},\ldots,\mathcal{Y}_{n,\text{obs}}}$.

La capacité de résoudre ce problème d'estimation d'état permet d'étudier la précision moyenne des valeurs estimées des grandeurs internes, et par là de comparer plusieurs systèmes de capteurs alternatifs en termes de performances moyennes. C'est donc aussi un outil intéressant pour la conception des systèmes de capteurs.

4.5.1.3 Planification de la production (problème type de prise de décisions séquentielles)

En planification de la production, on est amené à coordonner des décisions qui doivent être prises successivement dans le temps, d'où le terme de *prise de décisions séquentielles*.

Par exemple, pour optimiser le bénéfice qu'on peut tirer d'une chaîne de production de voitures, il faut s'assurer de l'approvisionnement en matières premières en établissant les contrats avec les fournisseurs (ce qui nécessite une anticipation sur les ventes futures sur une certaine période), il faut choisir la politique de marketing (qui peut s'adapter en prenant en compte la manière dont les consommateurs réagissent), et puis il faut aussi gérer la chaîne de fabrication (en fonction des stocks et des commandes) de façon à maximiser la valeur économique des ventes réalisables sur une certaine période (seulement les voitures effectivement livrées donnant lieu au payement de la facture par le client).

Comme une partie des facteurs qui influencent la qualité d'une stratégie de décision (combinant les décisions aux différentes étapes) sont aléatoires, il est naturel de formuler ce type de problème en faisant appel aux techniques probabilistes. En particulier, lorsqu'on doit décider des contrats avec les fournisseurs, on voudra prendre en compte les incertitudes qu'on a en ce qui concerne le comportement futur du marché tout en exploitant le fait qu'on pourra ajuster sa politique de marketing et en aval sa politique de gestion de l'usine.

La modélisation de problèmes de décisions séquentielles passe par trois étapes essentielles :

- Identification des facteurs exogènes (aléas), et des degrés de liberté endogènes sur lesquels on peut agir.

- Modélisation de la loi de probabilité des facteurs exogènes, typiquement sous la forme d'une suite de variables aléatoires faisant référence aux instants de temps futurs successifs (processus aléatoire, voir chapitres suivants).

- Formulation d'un problème d'optimisation, faisant intervenir une première série de variables relatives aux décisions à prendre maintenant, puis une suite de variables dont les valeurs seront contingentes aux réalisations futures des variables exogènes (à choisir par une *politique de décision* qui doit aussi être déterminée).

Lorsque le problème peut-être modélisé de façon discrète, comme c'était le cas du problème du "Monty Hall", on peut se servir d'un arbre de scénarios pour déterminer les décisions optimales, pour chaque scénario possible. Dans les situations pratiques plus complexes, on peut utiliser des techniques de discrétisation qui regroupent *a priori* des scénarios similaires, de façon à rendre le problème abordable.

La modélisation et la résolution de problèmes de prise de décisions séquentielles en environnement incertain repose directement sur les notions introduites dans ce chapitre et fait l'objet d'enseignements plus avancés dans les cursus de Master ingénieur et informaticien.

4.5.2 Méthode de Monte-Carlo

Au chapitre précédent nous avons introduit la méthode de Monte-Carlo pour estimer une intégrale simple du type $I_g = \int_0^1 g(x)\,\mathrm{d}x$ au moyen de la formule $\hat{I}_g = \frac{1}{n}\sum_{i=1}^n g(x_i)$ où les valeurs x_i sont obtenues par tirage uniforme et indépendant dans l'intervale $[0,1]$. L'écart-type de cet estimateur est donné par $\sigma_{\hat{I}_g} = \frac{1}{\sqrt{n}}\sigma_{g(\mathcal{X})}$.

Cette technique peut directement s'appliquer au calcul de l'intégrale multiple

$$I_g = \int_0^1 \cdots \int_0^1 g(x_1,\ldots,x_p)\,\mathrm{d}x_1 \cdots \mathrm{d}x_p \tag{4.120}$$

au moyen de l'estimateur

$$\hat{I}_g = \frac{1}{n}\sum_{i=1}^n g(x_1,\ldots,x_p), \tag{4.121}$$

où chaque terme est calculé en appliquant la fonction g à p nombres aléatoires $x_1,\ldots,x_p$ tirés indépendamment dans $[0,1]$. L'écart-type de cet estimateur est encore donné par

$$\sigma_{\hat{I}_g} = \frac{1}{\sqrt{n}}\sigma_{g(\mathcal{X}_1,\ldots,\mathcal{X}_p)}. \tag{4.122}$$

De même on peut approximer

$$I_g^f = \int_0^1 \cdots \int_0^1 g(x_1,\ldots,x_p) f_{\mathcal{X}}(x_1,\ldots,x_p)\,\mathrm{d}x_1 \cdots \mathrm{d}x_p, \tag{4.123}$$

par

$$\hat{I}_g^f = \frac{1}{n}\sum_{i=1}^n g(x_1,\ldots,x_p), \tag{4.124}$$

à condition que les p-tuples $x_1,\ldots,x_p$ soient obtenus par tirage dans la loi jointe $f_{\mathcal{X}}(x_1,\ldots,x_p)$.

Si on dispose d'une factorisation de la loi jointe $f(x_1,\ldots,x_p)$, ces tirages peuvent se faire en exploitant cette factorisation. Par exemple si les variables $\mathcal{X}_i$ sont indépendantes, le tirage d'un p-tuple $x_1,\ldots,x_p$ peut être fait en tirant chaque composante x_i indépendamment des autres selon sa loi marginale $f_{\mathcal{X}_i}$. De façon plus générale on tirera, dans l'ordre de la factorisation, les valeurs de x_i selon la loi $f_{\mathcal{X}_i|\mathcal{X}_1,\ldots,\mathcal{X}_{i-1}}$, en prenant en compte les valeurs $x_1,\ldots,x_{i-1}$ des variables dont a déjà tiré les valeurs.

Il est **remarquable** que la précision de cet estimateur de Monte-Carlo de l'intégrale multiple ne dépend pas directement de la dimension p de l'espace sur lequel on intègre, mais seulement de la variance de la fonction intégrée σ_g. Puisque les techniques de quadrature (vues par exemple en analyse numérique) sont de complexité croissante exponentiellement avec la dimension p, la méthode de Monte-Carlo est une alternative qui est souvent très intéressante d'un point de vue computationnel dans les problèmes de grande dimension.

4.5.2.1 Sondage stratifié

Le sondage stratifié, dans sa version simple, consiste à remplacer l'estimation directe par la méthode de Monte-Carlo de $E\{g(\mathcal{X})\}$ par la procédure suivante :

- On suppose qu'on dispose d'une variable binaire auxiliaire $\mathcal{Z}$ et qu'on est capable de tirer des échantillons selon les distributions conditionnelles $f_{\mathcal{X}|\mathcal{Z}}$ (pour chacune des valeurs de $\mathcal{Z} \in \{0,1\}$).
- On estime, $E\{g(\mathcal{X})|\mathcal{Z}\}$ pour chacune des valeurs de $\mathcal{Z} \in \{0,1\}$, par Monte-Carlo.
- On combine ces valeurs par la formule de l'espérance totale

$$E\{g(\mathcal{X})\} = P_{\mathcal{Z}}(0)E\{g(\mathcal{X})|\mathcal{Z}=1\} + P_{\mathcal{Z}}(0)E\{g(\mathcal{X})|\mathcal{Z}=1\}.$$

Pour un même nombre total d'évaluations de la fonction g, la variance de l'estimateur basé sur le sondage stratifié peut être nettement plus faible que celle de l'estimateur de Monte-Carlo de base. De fait, cette variance dépend

des variances conditionnelles de g sachant que $\mathcal{Z} = 0$ et sachant que $\mathcal{Z} = 1$, et du nombre n_0 et n_1 d'échantillons utilisés pour estimer les valeurs de $E\{g(\mathcal{X})|\mathcal{Z} = 0\}$ et $E\{g(\mathcal{X})|\mathcal{Z} = 1\}$.

Connaissant les variances conditionnelles de la variable g, on peut d'ailleurs déterminer la manière optimale d'allouer un nombre donné n d'évaluations de la fonction g entre les deux sous-populations correspondant respectivement aux deux valeurs possibles de la variable $\mathcal{Z}$. Bien entendu, ce raisonnement peut s'appliquer (récursivement) au cas où l'information auxiliaire $\mathcal{Z}$ est fournie par une variable discrète prenant un nombre fini quelconque de valeurs.

Exemple : enquêtes d'opinion. Dans le domaine des enquêtes d'opinion, on utilise très souvent la technique du sondage stratifié.

Par exemple, si on souhaite établir le nombre moyen d'heures par jour passés dans les transports en commun dans un pays, on peut postuler a priori que ce nombre est assez différent selon qu'on se trouve en zone rurale, ou bien dans une grande agglomération.

Intuitivement, on peut penser que les habitants d'une zone rurale utilisent très peu les transports en commun, l'espérance conditionnelle de la variable d'intérêt étant donc proche de 0 et la variance conditionnelle aussi; il faudrait donc sonder un nombre relativement faible d'habitants de zones rurales pour établir avec une bonne précision le nombre d'heures que ce type d'habitant passe en moyenne dans les transports en commun. De façon symétrique, les habitants des grandes agglomérations ont évidemment tendance à passer bien plus de temps dans les transports en commun, avec une variance sans doute plus élevée que celle des habitants de zones rurales, mais néanmoins plus faible que celle de la population générale.

Si le sondeur dispose d'un budget de n personnes qu'il souhaite interviewer, connaît la taille et a une bonne idée de la variance des ces deux sous-populations, il peut allouer son effort de manière "optimale" , en attribuant à chacune de sous-populations un nombre de sondés d'autant plus élevé que la variance dans cette population est élevée, et d'autant plus faible que la taille de cette sous-population est faible.

Suggestion: déterminer l'allocation optimale de n sondages sur deux sous-populations de citoyens, de façon à minimiser la variance de l'estimateur composite de l'espérance d'une variable aléatoire, en fonction des tailles N_1 et N_2 des sous-populations et des variances conditionnelles de la variable étudiée dans chacune des sous-populations.

4.5.2.2 Combinaison de la méthode de Monte-Carlo et des techniques de régression

Les techniques de régression permettent de construire des fonctions de certaines variables aléatoires qui sont en un certain sens les plus proches possibles d'une variable cible (cf les sections précédentes). La statistique et l'apprentissage automatique fournissent un panel de méthodes très larges pour construire ce type de fonctions à partir d'observations.

Cela veut dire qu'en pratique, lorsqu'on étudie le comportement d'un système complexe, modélisé par une fonction $y = f(x)$ (y et x étant potentiellement de dimension élevée), on dispose souvent de modèles approchés, $\hat{y} = \hat{f}(x)$ qui conduisent à une estimation assez précise des sorties y étant donné x, dans le sens que $E\{(\mathcal{Y} - \hat{\mathcal{Y}})^2\}$ est proche de zéro.

Si le modèle approché $\hat{f}(x)$ est utilisable en simulation informatique, il est peu couteux de déterminer les valeurs et aussi l'espérance de la variable $\hat{\mathcal{Y}}$, alors que la détermination de la variable $\mathcal{Y}$ peut nécessiter le recours à des expériences coûteuses en temps et en argent.

Dans ce cas, une bonne stratégie consiste à utiliser les connaissances disponibles pour construire un modèle informatique approché de la relation entrée-sortie, puis à utiliser la technique de Monte-Carlo pour estimer l'écart moyen entre la sortie du modèle et la vraie valeur $\mathcal{Y}$, et séparément pour déterminer la valeur de $E\{\hat{\mathcal{Y}}\}$.

Si le modèle est suffisamment précis, l'écart $\mathcal{Y} - \hat{\mathcal{Y}}$ sera faible et donc aussi de faible variance, ce qui se traduit alors par une réduction importante du nombre de mesures nécessaires de la variable $\mathcal{Y}$ pour estimer son espérance avec la précision souhaitée, sachant que l'essentiel de l'information est obtenue au moyen de l'estimation de $E\{\hat{\mathcal{Y}}\}$ par le biais de simulations informatiques peu coûteuses.

5 VECTEURS ALÉATOIRES ET PROCESSUS ALÉATOIRES GAUSSIENS

NB: La version complète de ce chapitre sera rendu disponible en cours de semestre.

5.1 VECTEURS ALEATOIRES

Nous nous intéressons ici aux v.a. à valeurs dans l'espace euclidien $\mathbb{R}^p$. Nous introduisons d'abord quelques notations et propriétés générales de telles variables aléatoires, puis nous nous focaliserons sur l'étude des vecteurs aléatoires gaussiens.

Ci-dessous nous indiquerons en gras les vecteurs (colonnes) et matrices. Etant donné un vecteur ou une matrice $\boldsymbol{V}$ nous noterons par $\boldsymbol{V}^T$ le vecteur ou la matrice transposée. Etant donnée une matrice $\boldsymbol{M}$ nous noterons par $|\boldsymbol{M}|$ son déterminant.

5.1.1 Généralités sur les v.a. vectorielles

Une v.a. vectorielle ou vecteur (colonne) aléatoire $\boldsymbol{\mathcal{X}}$ est une application mesurable de $(\Omega, \mathcal{E}, P)$ dans $\mathbb{R}^p$ muni de sa σ-algèbre borélienne (produit cartésien de p σ-algèbres boréliennes sur $\mathbb{R}$).

La fonction de répartition d'un vecteur aléatoire est une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$ définie par

$$F_{\boldsymbol{\mathcal{X}}}(x_1, x_2, \ldots, x_p) \stackrel{\triangle}{=} P(\mathcal{X}_1 < x_1, \ldots, \mathcal{X}_p < x_p), \tag{5.1}$$

où $\mathcal{X}_i$ désigne la i-ème composante de $\boldsymbol{\mathcal{X}}$.

Si la densité existe, elle est définie par

$$f_{\boldsymbol{\mathcal{X}}}(x_1, x_2, \ldots, x_p) \stackrel{\triangle}{=} \frac{\partial^p F_{\boldsymbol{\mathcal{X}}}}{\partial x_1 \ldots \partial x_p}. \tag{5.2}$$

On note par $\boldsymbol{\mu}$ (ou $\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}$, si nécessaire; certains auteurs utilisent la notation $\overline{\boldsymbol{x}}$) le vecteur colonne

$$\boldsymbol{\mu} \stackrel{\triangle}{=} E\{\boldsymbol{\mathcal{X}}\} = \begin{bmatrix} E\{\mathcal{X}_1\} \\ \vdots \\ E\{\mathcal{X}_p\} \end{bmatrix}, \tag{5.3}$$

dont les composantes sont les espérances mathématiques des p composantes de $\boldsymbol{\mathcal{X}}$. Dans ce qui suit, nous supposons que ces quantités existent (et sont donc finies).

On note par $\boldsymbol{\Sigma}$ (ou $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{X}}}$, si nécessaire) la matrice $p \times p$ de variance-covariance définie par

$$\boldsymbol{\Sigma} \triangleq E\{(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})^T\}, \tag{5.4}$$

où l'opérateur d'espérance est appliqué élément par élément à la matrice $(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})^T$ dont l'élément i, j est la variable aléatoire $(\mathcal{X}_i - \mu_i)(\mathcal{X}_j - \mu_j)$. Dans ce qui suit, nous supposons que toutes ces grandeurs existent (et sont donc finies).

L'élément i, j de $\boldsymbol{\Sigma}$ vaut donc

$$\boldsymbol{\Sigma}_{i,j} = \text{cov}(\mathcal{X}_i; \mathcal{X}_j). \tag{5.5}$$

En particulier, on a $\boldsymbol{\Sigma}_{i,i} = V\{\mathcal{X}_i\}$, puisque $\text{cov}(\mathcal{X}_i; \mathcal{X}_i) = V\{\mathcal{X}_i\}$.

Il est important de noter que la matrice $\boldsymbol{\Sigma}$ est par définition symétrique et semi-définie positive. On a en effet $\forall \boldsymbol{y} \in \mathbb{R}^p$ que

$$\begin{aligned}
\boldsymbol{y}^T \boldsymbol{\Sigma} \boldsymbol{y} &\triangleq \boldsymbol{y}^T E\{(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})^T\} \boldsymbol{y} && (5.6)\\
&= E\{\boldsymbol{y}^T (\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})^T \boldsymbol{y}\} && (5.7)\\
&= E\{\|\boldsymbol{y}^T (\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})\|^2\} && (5.8)\\
&\geq 0, && (5.9)
\end{aligned}$$

puisque l'espérance d'une variable positive (équation (5.8)) doit être positive. Notons que le passage de (5.6) à (5.7) est autorisé, car l'espérance d'une combinaison linéaire de variables est égale à la combinaison linéaire des espérances de ces variables (pour autant que ces dernières soient finies).

Remarquons aussi que

$$\boldsymbol{\Sigma} = E\{\boldsymbol{\mathcal{X}} \boldsymbol{\mathcal{X}}^T\} - \boldsymbol{\mu} \boldsymbol{\mu}^T. \tag{5.10}$$

En effet, $\boldsymbol{\Sigma}_{i,j} = \text{cov}(\mathcal{X}_i; \mathcal{X}_j) = E\{(\mathcal{X}_i - \mu_i)(\mathcal{X}_j - \mu_j)\}$ et $E\{(\mathcal{X}_i - \mu_i)(\mathcal{X}_j - \mu_j)\} = E\{\mathcal{X}_i \mathcal{X}_j\} - \mu_i \mu_j$, puisque $E\{(\mathcal{X}_i - \mu_i)\mu_j)\} = 0$, que $E\{\mu_i(\mathcal{X}_j - \mu_j)\} = 0$, et que $E\{\mu_i \mu_j\} = \mu_i \mu_j$.

5.1.1.1 Fonction caractéristique

On appelle fonction caractéristique du vecteur aléatoire $\boldsymbol{\mathcal{X}} \in \mathbb{R}^p$ la fonction définie $\forall \boldsymbol{a} \in \mathbb{R}^p$ par

$$\phi_{\boldsymbol{\mathcal{X}}}(\boldsymbol{a}) = E\{\exp(i \boldsymbol{a}^T \boldsymbol{\mathcal{X}})\}. \tag{5.11}$$

La fonction caractéristique porte bien son nom, puisque sa connaissance est nécessaire et suffisante pour caractériser complètement le comportement conjoint de l'ensemble des variables aléatoires formant le vecteur $\boldsymbol{\mathcal{X}}$.

On démontre que les composantes $\mathcal{X}_i$ du vecteur aléatoire $\boldsymbol{\mathcal{X}}$ sont mutuellement indépendantes si et seulement si la fonction caractéristiques se factorise comme suit:

$$\phi_{\boldsymbol{\mathcal{X}}}(\boldsymbol{a}) = \prod_{i=1}^{p} \phi_{\mathcal{X}_i}(a_i). \tag{5.12}$$

5.1.1.2 Transformations linéaires

Soit $\boldsymbol{A}$ une matrice $r \times p$ et $\boldsymbol{\mathcal{X}}$ un v.a. de $\mathbb{R}^p$. Alors $\boldsymbol{\mathcal{Y}} = \boldsymbol{A}\boldsymbol{\mathcal{X}}$ est un vecteur aléatoire de $\mathbb{R}^r$.

On a $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}} = \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}$ et $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Y}}} = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{X}}}\boldsymbol{A}^T$.

Soit $\boldsymbol{b}$ un vecteur de $\mathbb{R}^p$ et $\boldsymbol{\mathcal{X}}$ un v.a. de $\mathbb{R}^p$. Alors $\boldsymbol{\mathcal{Z}} = \boldsymbol{b} + \boldsymbol{\mathcal{X}}$ est un vecteur aléatoire de $\mathbb{R}^p$.

On a $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Z}}} = \boldsymbol{b} + \boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}$ et $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Z}}} = \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{X}}}$.

5.1.1.3 Théorème de Cramer-Wold

La loi de $\boldsymbol{\mathcal{X}}$ est entièrement déterminée par celles de toutes les combinaisons linéaires de ses composantes $\boldsymbol{a}^T\boldsymbol{\mathcal{X}}$, $\forall \boldsymbol{a} \in \mathbb{R}^p$.

En effet, si nous connaissons la loi de $\mathcal{Y} = \boldsymbol{a}^T\boldsymbol{\mathcal{X}}$, nous connaissons aussi la fonction caractéristique de $\mathcal{Y}$, c'est-à-dire $\forall t$ la valeur de la fonction

$$\phi_{\mathcal{Y}}(t) = E\{\exp(it\mathcal{Y}\} = E\{\exp(it\boldsymbol{a}^T\boldsymbol{\mathcal{X}}\}. \tag{5.13}$$

Par conséquent, posant $t = 1$, nous connaissons $\forall \boldsymbol{a}$ la valeur $E\{\exp(i\boldsymbol{a}^T\boldsymbol{\mathcal{X}}\} = \phi_{\boldsymbol{\mathcal{X}}}(\boldsymbol{a})$, ce qui détermine la loi de $\boldsymbol{\mathcal{X}}$.

5.1.1.4 Décorrélation

Puisque la matrice $\boldsymbol{\Sigma}$ est s.d.p., elle peut être diagonalisée au moyen d'une transformation orthogonale. Le résultat de cette transformation donne un vecteur aléatoire dont les composantes sont *décorrélées*, mais pas nécessairement indépendantes. Cette opération se fait en pre-multipliant $\boldsymbol{\mathcal{X}}$ par une matrice $\boldsymbol{A}$ (de dimension $p \times p$) dont les lignes sont les vecteurs propres de $\boldsymbol{\Sigma}$. Les composantes du vecteur $\boldsymbol{\mathcal{Y}} = \boldsymbol{A}\boldsymbol{\mathcal{X}}$ sont alors les projections de $\boldsymbol{\mathcal{X}}$ sur les vecteurs propres de la matrice $\boldsymbol{\Sigma}$. La matrice $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Y}}}$ est diagonale; elle comportera des termes diagonaux nuls, si et seulement si, la matrice $\boldsymbol{\Sigma}$ est de rang inférieur à p.

5.1.2 Vecteurs aléatoires gaussiens

Définition. $\boldsymbol{\mathcal{X}} \in \mathbb{R}^p$ est (par définition) un vecteur aléatoire gaussien à p dimensions (on note $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, où $\boldsymbol{\Sigma}$ est la matrice de variance-covariance de $\boldsymbol{\mathcal{X}}$, et $\boldsymbol{\mu}$ sa moyenne), si toute combinaison linéaire de ses composantes $\boldsymbol{a}^T\boldsymbol{\mathcal{X}}$, $\forall \boldsymbol{a} \in \mathbb{R}^p$, suit une loi de Laplace-Gauss $\mathcal{N}(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})$.

La propriété d'être gaussien est donc invariante vis-à-vis de toute transformation linéaire (rotation, dilatation, translation,...) de l'espace $\mathbb{R}^p$. Cette propriété implique en particulier que toutes les composantes suivent des lois gaussiennes (mais la réciproque est fausse).

5.1.2.1 Propriétés fondamentales

On a les propriétés fondamentales suivantes :

- En général, si $\boldsymbol{A}$ est une matrice $r \times p$ et $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ un vecteur aléatoire gaussien de $\mathbb{R}^p$, alors $\boldsymbol{\mathcal{Y}} = \boldsymbol{A}\boldsymbol{\mathcal{X}}$ est un vecteur aléatoire gaussien de $\mathbb{R}^r$, et on a $\boldsymbol{\mathcal{Y}} \sim \mathcal{N}_r(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.
  (Suggestion : montrer que $\boldsymbol{\mathcal{Y}}$ est bien un vecteur aléatoire gaussien, en prouvant que toutes ses projections sont bien des variables aléatoires gaussiennes.)

- Donc, si $\mathcal{Y} = \boldsymbol{a}^T\boldsymbol{\mathcal{X}}$ alors $E\{\mathcal{Y}\} = \boldsymbol{a}^T\boldsymbol{\mu}$, et $V\{\mathcal{Y}\} = \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}$, donc $\mathcal{Y} \sim \mathcal{N}(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})$.

- On déduit de la propriété précédente que les distributions marginales (des composantes de $\boldsymbol{\mathcal{X}}$) sont les suivantes : $\mathcal{X}_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$ conformément à l'intuition.
  (Suggestion : appliquer la propriété précédente au vecteur $\boldsymbol{a}$ de composantes $a_j = \delta_{i,j}$.)

- Les composantes de $\boldsymbol{\mathcal{X}}$ sont mutuellement indépendantes si, et seulement si, $\boldsymbol{\Sigma}$ est une matrice diagonale, c'est-à-dire si les composantes sont décorrélées deux à deux.
  (Suggestion : montrer que si les variables sont indépendantes alors la matrice $\boldsymbol{\Sigma}$ est bien diagonale; la démonstration de la réciproque exploite les propriétés de la fonction caractéristique.)

- Lorsque $\boldsymbol{\Sigma}$ est régulière, et seulement dans ce cas, la densité existe et vaut

$$f(\boldsymbol{\mathcal{X}}) = \frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(\frac{1}{2}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})\right). \tag{5.14}$$

5.1.2.2 Distributions conditionnelles

Dans ce qui suit, nous supposons que la matrice $\mathbf{\Sigma}$ est non singulière, afin d'éviter des complications dans nos formulations. Nous invitons le lecteur à regarder comment certaines formules devraient être mises à jour si la matrice $\mathbf{\Sigma}$ est singulière.

Si on partitionne $\boldsymbol{\mathcal{X}}$ en deux sous-vecteurs $\boldsymbol{\mathcal{X}}_1$ et $\boldsymbol{\mathcal{X}}_2$ à k et $p-k$ composantes, respectivement de moyennes $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$:

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} \boldsymbol{\mathcal{X}}_1 \\ \boldsymbol{\mathcal{X}}_2 \end{bmatrix}, \tag{5.15}$$

la moyenne se partitionne selon

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \tag{5.16}$$

et la matrice de variance-covariance se partitionne selon

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{1,1} & \mathbf{\Sigma}_{1,2} \\ \mathbf{\Sigma}_{2,1} & \mathbf{\Sigma}_{2,2} \end{bmatrix} \tag{5.17}$$

Notons que si $\mathbf{\Sigma}$ est non singulière, alors il doit en être de même de $\mathbf{\Sigma}_{2,2}$.

La loi conditionnelle de $\boldsymbol{\mathcal{X}}_1$ lorsque $\boldsymbol{\mathcal{X}}_2$ est connu est alors une Gaussienne à k dimensions

- d'espérance conditionnelle
$$E\{\boldsymbol{\mathcal{X}}_1|\boldsymbol{\mathcal{X}}_2\} = \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{1,2}\mathbf{\Sigma}_{2,2}^{-1}(\boldsymbol{\mathcal{X}}_2 - \boldsymbol{\mu}_2). \tag{5.18}$$
- de matrice de variance-covariance conditionnelle
$$\mathbf{\Sigma}_{1,1|2} = \mathbf{\Sigma}_{1,1} - \mathbf{\Sigma}_{1,2}\mathbf{\Sigma}_{2,2}^{-1}\mathbf{\Sigma}_{2,1}. \tag{5.19}$$

On constate donc que l'espérance conditionnelle est une fonction affine de $\boldsymbol{\mathcal{X}}_2$ et que la matrice de variance-covariance conditionnelle ne dépend pas de la valeur de $\boldsymbol{\mathcal{X}}_2$.

5.1.2.3 Cas particulier : $p = 2$

Dans le cas particulier où $p = 2$ on a

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{5.20}$$

où

$$\rho \triangleq \frac{\text{cov}\{\mathcal{X}_1, \mathcal{X}_2\}}{\sigma_1\sigma_2}$$

est le coefficient de corrélation linéaire.

La distribution conditionnelle de $\mathcal{X}_1$ étant donné $\mathcal{X}_2$ est alors

$$f(\mathcal{X}_1|\mathcal{X}_2) \sim \mathcal{N}(\mu_1 + \rho\sigma_1\frac{\mathcal{X}_2 - \mu_2}{\sigma_2}, \sigma_1\sqrt{1-\rho^2}). \tag{5.21}$$

La densité n'existe que si $|\rho| < 1$ dans ce cas particulier.

5.1.2.4 Remarques et interprétations

On voit que la distribution Gaussienne est fortement liée à la notion de linéarité. Une distribution Gaussienne est en effet une distribution qui garde sa structure Gaussienne lorsqu'on effectue des transformations linéaires. D'autre part, pour des variables conjointement Gaussiennes, l'espérance conditionnelle est une fonction linéaire, et la matrice de variance-covariance conditionnelle est indépendante de la valeur de la variable qui conditionne. Enfin, il est possible de diagonaliser la matrice de variance-covariance au moyen d'une transformation linéaire (orthogonale). Une fois diagonalisée, les composantes sont indépendantes, ce qui veut dire que dans le cas

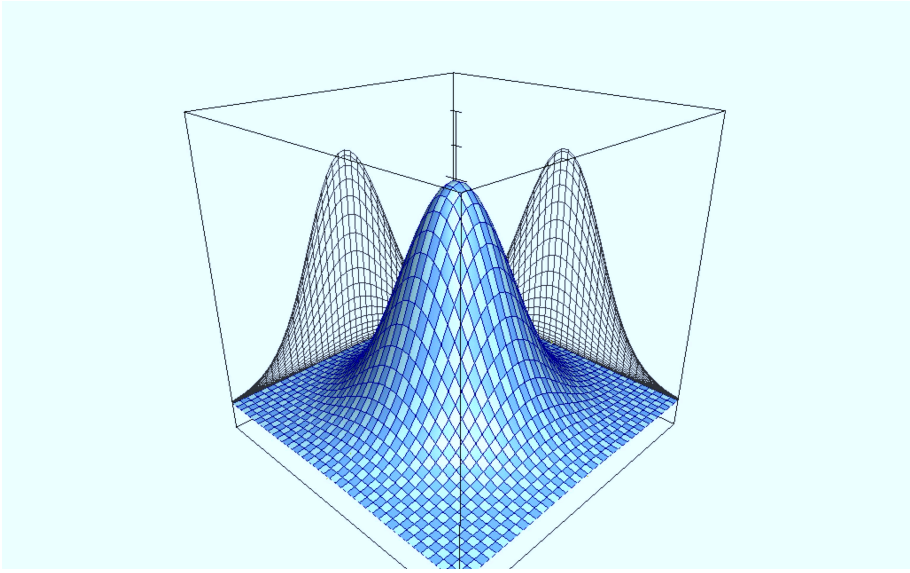


Figure 5.1: Loi normale dans $\mathbb{R}^2$ avec $\mu = (0\,0)^T$ et $\Sigma = \text{Diag}(1\,1)$: $f_{\mathcal{X}_1\mathcal{X}_2}(x_1, x_2) = \frac{1}{2\pi} e^{\left(-\frac{x_1^2+x_2^2}{2}\right)}$

de distributions Gaussiennes la notion de dépendance probabiliste et celle de dépendance linéaire coïncident essentiellement.

Enfin, on voit que pour un couple de v.a. conjointement Gaussiennes, le coefficient de corrélation linéaire ρ mesure la dépendance entre celles-ci. Il est nul si, et seulement si, les v.a. sont statistiquement indépendantes; il vaut 1 si, et seulement si, l'une des variables est une fonction linéaire de l'autre. Enfin, il prend une valeur non triviale si, et seulement si, les deux variables peuvent s'exprimer sous la forme de deux combinaisons linéaires **linéairement indépendantes** de deux v.a. gaussiennes indépendantes.

Nous verrons dans l'annexe sur les statistiques que les distributions Gaussiennes jouent un rôle très important en estimation statistique, notamment à cause des fortes propriétés mathématiques qui les caractérisent. Pour terminer, signalons que le théorème central-limite formulé ci-dessous s'applique également au cas des v.a. Gaussiennes de $\mathbb{R}^p$.

5.1.3 Illustrations et applications

5.2 FONCTIONS ALÉATOIRES ET PROCESSUS STOCHASTIQUES

5.2.1 Notion de processus stochastique

5.2.2 Processus gaussiens

5.2.3 Illustrations et applications

II Rappels et compléments

Appendice A

Théorie des ensembles et analyse combinatoire

Cette appendice a pour objet de rappeler quelques notions élémentaires de la théorie des ensembles et d'analyse combinatoire.

A.1 LOIS DE DE MORGAN

Les lois de *de Morgan* existent sous deux variantes équivalentes, à savoir une version ensembliste et une version logique.

On a, dans la version logique (A, B désignent des propositions logiques quelconques)

$$\neg(A \wedge B) \equiv (\neg A \vee \neg B) \text{ et } \neg(A \vee B) \equiv (\neg A \wedge \neg B). \tag{A.1}$$

On a, dans la version ensembliste (A, B désignent des sous-ensembles quelconques d'un univers Ω)

$$(A \cap B)^c = (A^c \cup B^c) \text{ et } (A \cup B)^c = (A^c \cap B^c). \tag{A.2}$$

Ces lois restent vraies pour un nombre quelconque d'ensembles (propositions). Par exemple, pour une collection dénombrables de parties A_i de Ω on a

$$\left(\bigcup_i A_i\right)^c = \bigcap_i A_i^c. \tag{A.3}$$

A.2 CARDINALITÉS ET DÉNOMBREMENTS

A.2.1 Cardinalités

Dans cette section nous rappelons la notion de cardinalité d'un ensemble et nous dérivons les formules utiles d'analyse combinatoire.

Nous désignons par $\mathbb{N}$ l'ensemble des nombres naturels (y compris 0) et par $\mathbb{N}_0$ l'ensemble des naturels strictement positifs.

Un ensemble A est fini, si et seulement si il est vide ou peut être mis en bijection avec un ensemble du type $\{1, \ldots, n\}$ ($n \in \mathbb{N}_0$). On note $|A|$ sa cardinalité; dans le premier cas elle vaut 0 dans le second cas elle vaut n.

Un ensemble est dénombrable, ssi il peut être mis en bijection avec une partie de $\mathbb{N}$. S'il peut être mis en bijection avec $\mathbb{N}$ nous dirons qu'il est infini dénombrable; dans ce cas il n'est pas fini.

On montre que toute union dénombrable d'ensembles dénombrables est elle aussi dénombrable; en particulier l'ensemble $\mathbb{Q}$ des rationnels est dénombrable.

Soit A un ensemble. On désigne par $\mathcal{P}(A)$ l'ensemble des parties (ou encore sous-ensembles) de A. Si A est fini, $\mathcal{P}(A)$ l'est aussi et on a $|\mathcal{P}(A)| = 2^{|A|}$. Par extension on utilise aussi la notation 2^A pour désigner l'ensemble des parties d'un ensemble quelconque (pas nécessairement fini), et on utilise aussi la notation $|A|$ pour désigner la *cardinalité* de A: deux ensembles sont dits de même cardinalité, s'ils peuvent être mis en bijection.

On a $|2^{\mathbb{N}}| = |\mathbb{R}|$, et on montre qu'en toute généralité on a

$$|A \cup B| + |A \cap B| = |A| + |B|.$$

Si A et B sont finis, alors

$$|A \times B| = |A| \cdot |B|.$$

A.2.2 Dénombrements

Dans cette section nous considérons la cardinalité d'ensembles finis construits à partir d'un ensemble fini A de cardinalité $n = |A|$; k désigne un nombre entier positif ou nul.

A.2.2.1 Tirages avec remise

Un k-tirage avec remise dans A est un élément (un k-tuple) de A^k.

On a donc $|A^k| = |A|^k = n^k$ tirages sans remise différents.

En effet, pour $i = 1, \ldots, k$ on peut choisir un élément quelconque parmi les n de A.

On appelle aussi un tel tuple un arrangement avec répétitions (éventuelles).

Notons qu'un k-tuple de longueur nulle (i.e. avec $k = 0$) s'identifie à la liste vide; par conséquent $|A^0| = 1$.

A.2.2.2 Tirages sans remise

Un k-tirage sans remise dans A est un élément (un k-tuple) de A^k tel que toutes ses composantes soient différentes. On appelle aussi ce type de tuple, un arrangement sans répétitions, et on désigne par $\mathcal{A}_n^k$ le nombre de tels arrangements.

Si on a $1 \leq k \leq n$, on a

$$\mathcal{A}_n^k = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

En effet, pour $i = 1, \ldots, k$ on peut choisir un élément quelconque parmi les $n - (i-1)$ de A non encore choisis.

Par ailleurs, pour $k = 0$ on a $\mathcal{A}_n^k = 1$ et pour $k > n$ on a $\mathcal{A}_n^k = 0$.

A.2.2.3 Permutations (sans répétitions)

Une permutation de A sans répétition est un tirage sans remise de $|A|$ éléments de A.

Il y a donc $\mathcal{A}_n^n = n!$ permutations différentes de A.

A.2.2.4 Permutations avec répétitions

Il s'agit de choisir un tuple de longueur $\sum_{i=1}^{n} n_i$, composé de n_i copies de chaque élément a_i de A.

Le nombre total de choix possibles est

$$\frac{(\sum_{i=1}^{n} n_i)!}{\prod_{i=1}^{n}(n_i!)}.$$

On démontre cette formule en considérant les permutations d'un ensemble composé de n_i copies de chaque élément de A, puis en remarquant que ces permutations forment des tuples invariants par permutation des différents groupes de copies des éléments de A.

A.2.2.5 Combinaisons sans répétition

Une combinaison sans répétitions de k éléments de A est un sous-ensemble de taille k de A. On obtient le nombre $\mathcal{C}_n^k$ de combinaisons en remarquant que les arrangements sans répétitions sont obtenus en permutant les combinaisons.

Le nombre total de combinaisons sans répétition vaut donc

$$\mathcal{C}_n^k = \frac{\mathcal{A}_n^k}{k!} = \frac{n!}{(n-k)!k!}.$$

Notons que $\mathcal{C}_n^n = 1$ (la seule combinaison étant l'ensemble A lui-même); de même, $\mathcal{C}_n^0 = 1$ (la seule combinaison étant l'ensemble vide $\emptyset$).

Le nombre de combinaisons de i éléments pris parmi j est égal à la somme du nombre de combinaisons de $i-1$ éléments parmi $j-1$ et du nombre de combinaisons de i éléments parmi $j-1$:

Règle de Pascal

$$\forall 0 < i < j : \mathcal{C}_j^i = \mathcal{C}_{j-1}^{i-1} + \mathcal{C}_{j-1}^i.$$

(**Suggestion :** se convaincre que cette formule est bien correcte.)

Notons enfin que puisque les sous-ensembles de A sont en bijection avec les combinaisons de taille quelconque de A, on en déduit que

$$\sum_{i=0}^{n} \mathcal{C}_n^i = 2^n.$$

Appendice B
Notion de tribu borélienne

Cet appendice a pour objet de définir de manière rigoureuse la notion de tribu (ou $\sigma-$algèbre) borélienne et d'énoncer les propriétés les plus importantes des fonctions mesurables à valeurs réelles qui assurent la cohérence de la notion de variable aléatoire.

B.1 σ-ALGÈBRES

Définition: $\sigma-$algèbre engendrée
Soit Ω un ensemble et soit $\mathcal{B}$ une collection de parties de Ω (en nombre quelconque), la $\sigma-$algèbre engendrée par $\mathcal{B}$ est par définition la plus petite $\sigma-$algèbre de Ω qui contient tous les éléments de $\mathcal{B}$. (Elle existe toujours et est unique). On la note $\sigma(\mathcal{B})$.

Il s'agit de l'ensemble des parties de Ω qui peuvent être construites à partir des ensembles $B_i \in \mathcal{B}$ en appliquant les opérations de complémentation, d'union et d'intersection au plus un nombre dénombrable de fois.

Définition: $\sigma-$algèbre produit
Soient $(\Omega_1, \mathcal{E}_1), \ldots, (\Omega_n, \mathcal{E}_n)$, n espaces mesurables, et soit le produit cartésien $\Omega = \Omega_1 \times \cdots \times \Omega_n$ composé des n-tuples $\omega = (\omega_1, \ldots, \omega_n)$ tels que $\forall i = 1, \ldots, n : \omega_i \in \Omega_i$.
La $\sigma-$algèbre produit des $\mathcal{E}_i$ est définie comme la $\sigma-$algèbre engendrée sur Ω par l'ensemble

$$\mathcal{B} = \{B = B_1 \times \cdots \times B_n : (\forall i = 1, \ldots, n : B_i \in \mathcal{E}_i)\}.$$

La notion est graphiquement illustrée sur la figure B.1

L'espace mesurable produit des $(\Omega_i, \mathcal{E}_i)$ est l'espace mesurable $(\Omega, \sigma(\mathcal{B}))$.

B.2 TRIBU BORÉLIENNE SUR LA DROITE RÉELLE

$\mathcal{B}_\mathbb{R}$, la tribu borélienne dans $\mathbb{R}$ est la plus petite tribu contenant l'ensemble des intervalles et semi-intervalles (ouverts et fermés, à gauche et à droite).

On peut montrer que c'est aussi la $\sigma-$algèbre engendrée par l'ensemble des intervales qui peuvent s'écrire sous la forme $]q, +\infty[$, où q est un nombre rationnel.

La tribu borélienne dans $\mathbb{R}$ contient donc aussi tous les singletons, toutes les parties finies, et toutes les parties dénombrables de $\mathbb{R}$.

Nous désignerons par $\mathcal{B}_\mathbb{R}$ cette tribu.

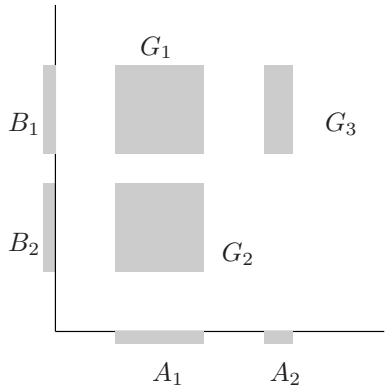


Figure B.1: Un élément G de la $\sigma-$algèbre produit des algèbres $\mathcal{A}$ et $\mathcal{B}$

B.3 TRIBU BORÉLIENNE SUR UN ESPACE EUCLIDIEN

$\mathcal{B}_{\mathbb{R}^n}$, la tribu borélienne sur $\mathbb{R}^n$ est la la tribu produit (n fois) de tribus boréliennes de $\mathbb{R}$.

On peut aussi dire qu'il s'agit de la tribu engendrée par l'ensemble de tous les hyper-rectangles de $\mathbb{R}^n$. Elle contient également toutes les parties dénombrables de $\mathbb{R}^n$.

On démontre que la tribu borélienne dans $\mathbb{R}^n$ est aussi la tribu engendrée par l'ensemble de boules ouvertes de centre et de rayon rationnels.

Nous désignerons par $\mathcal{B}_{\mathbb{R}^n}$ cette tribu.

Notons que $\mathcal{B}_{\mathbb{R}^n}$ forme un ensemble de parties très riches de $\mathbb{R}^n$. Néanmoins, on peut démontrer qu'il existe des parties de $\mathbb{R}^n$ qui n'appartiennent pas à la tribu borélienne $\mathcal{B}_{\mathbb{R}^n}$.

B.4 FONCTIONS MESURABLES À VALEURS RÉELLES

Soit $(\Omega, \mathcal{E})$ un espace mesurable, et soient f et g des fonctions à valeurs réelles définies sur Ω. On a les résultats suivants:

- f est $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurable, ssi $\forall x \in \mathbb{R}$, l'ensemble $\{\omega \in \Omega : f(\omega) < x\} \in \mathcal{E}$.
- f constante sur Ω implique que f est $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurable
- 1_A (fonction caractéristique de l'ensemble A) avec $A \subset \Omega$ est $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurable, si et seulement si $A \in \mathcal{E}$
- f et g $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurables, implique que
  - $f + g$ est $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurable
  - fg est $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurable
  - $\max(f, g)$ et $\min(f, g)$ sont $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurables
- Si $(\Omega, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$, alors f continue implique f mesurable.

B.4.1 Fonctions à valeurs vectorielles

Soient $f_i, i = 1, \ldots, n$ des fonctions $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurables.

Alors on montre que la fonction $f = (f_1, \ldots, f_n)$ est $(\mathcal{E}, \mathcal{B}_{\mathbb{R}^n})$-mesurable.

Réciproquement, si une fonction g est $(\mathcal{E}, \mathcal{B}_{\mathbb{R}^n})$-mesurable, alors les fonctions g_i qui correspondent aux composantes de g sont toutes $(\mathcal{E}, \mathcal{B}_{\mathbb{R}})$-mesurables.

D'ailleurs g est $(\mathcal{E}, \mathcal{B}_{\mathbb{R}^n})$-mesurable, ssi $\forall (x_1, \ldots, x_n) \in \mathbb{R}^n$, l'ensemble $\{\omega \in \Omega : g_i(\omega) < x_i\} \in \mathcal{E}$.

B.4.2 Suites de fonctions mesurables

Si la suite f_n de fonctions mesurables converge ponctuellement vers une fonction f, alors f est aussi mesurable (on suppose que les fonctions sont à valeurs dans $\mathbb{R}^n$). La convergence ponctuelle signifie que $\forall \omega \in \Omega$, la suite $f_n(\omega)$ de $\mathbb{R}^n$ converge vers $f(\omega)$.

Appendice C
Petite histoire du calcul des probabilités

Le développement du calcul des probabilités est une des plus belles aventures de l'histoire de la science.

Nous reproduisons ci-dessous des extraits de l'étude de l'histoire du calcul des probabilités, depuis les origines jusqu'à notre temps, fournis par différents auteurs.

Nous recommandons aussi au lecteur intéressé la référence [Jay03] pour une lecture critique et intéressante des travaux sur le sujet.

Texte de Wikipedia, extrait de http://fr.wikipedia.org/wiki/Probabilité.

La probabilité (du latin probabilitas) est une évaluation du caractère probable d'un évènement. En mathématiques, l'étude des probabilités est un sujet de grande importance donnant lieu à de nombreuses applications.

La probabilité d'un événement est un nombre réel compris entre 0 et 1. Plus ce nombre est grand, plus le risque (ou la chance, selon le point de vue) que l'événement se produise est grand. Si on considère que la probabilité qu'un lancer de pièce donne pile est égale à 1/2, cela signifie que, si on lance un très grand nombre de fois cette pièce, la fréquence des piles va très probablement tendre vers 1/2, sans préjuger de la régularité de leur répartition. Cette notion empirique sera définie plus rigoureusement dans le corps de cet article.

Contrairement à ce que l'on pourrait penser de prime abord l'étude scientifique des probabilités est relativement récente dans l'histoire des mathématiques. D'autres domaines tels que la géométrie, l'arithmétique, l'algèbre ou l'astronomie faisaient l'objet d'étude mathématique durant l'Antiquité mais on ne trouve pas de trace de textes mathématiques sur les probabilités. L'étude des probabilités a connu de nombreux développements au cours des trois derniers siècles en partie grâce à l'étude de l'aspect aléatoire et en partie imprévisible de certains phénomènes, en particulier les jeux de hasard. Ceux-ci ont conduit les mathématiciens à développer une théorie qui a ensuite eu des implications dans des domaines aussi variés que la météorologie, la finance ou la chimie. Cet article est une approche simplifiée des concepts et résultats d'importance en probabilité ainsi qu'un historique de l'usage du terme "probabilité" qui a eu plusieurs autres sens avant celui qu'on lui connaît aujourd'hui en mathématiques.

(. . .)

À l'origine, dans les traductions d'Aristote, le mot "probabilité" ne désigne pas une quantification du caractère aléatoire d'un fait mais l'idée qu'une idée est communément admise par tous. Ce n'est qu'au cours du Moyen Âge puis de la Renaissance autour des commentaires successifs et des imprécisions de traduction de l'oeuvre d'Aristote que ce terme connaîtra un glissement sémantique pour finir par désigner la vraisemblance d'une idée. Au 16ème siècle puis au 17ème siècle c'est ce sens qui prévaut en particulier dans le probabilisme en théologie morale. C'est dans la deuxième moitié du 17ème siècle, à la suite des travaux de Blaise Pascal, Pierre de Fermat et Christian Huygens sur le problème des partis que ce mot prend peu à peu son sens actuel avec les développements du traitement mathématique du sujet par Jakob Bernoulli. Ce n'est alors qu'au 19ème siècle qu'apparaît ce qui peut être considéré comme la théorie moderne des probabilités en mathématiques.

(. . .)

La théorie de la probabilité classique ne prend réellement son essor qu'avec les notions de mesure et d'ensembles mesurables qu'Émile Borel introduit en 1897. Cette notion de mesure est complétée par Henri Léon Lebesgue et sa théorie de l'intégration. La première version moderne du théorème de la limite centrale est donné par Alexandre Liapounov en 1901 et la première preuve du théorème moderne est donnée par Paul Lévy en 1910. En 1902, Andrei Markov introduit les chaînes de Markov pour entreprendre une généralisation de la loi des grands nombres pour une

suite d'expériences dépendant les unes des autres. Ces chaînes de Markov connaîtront de nombreuses applications entre autres pour modéliser la diffusion ou pour l'indexation de sites internet sur Google.

Il faudra attendre 1933 pour que la théorie des probabilités sorte d'un ensemble de méthodes et d'exemples divers et devienne une véritable théorie, axiomatisée par Kolmogorov.

Kiyoshi Itô met en place une théorie et un lemme qui porte son nom dans les années 1940. Ceux-ci permettent de relier le calcul stochastique et les équations aux dérivées partielles faisant ainsi le lien entre analyse et probabilités. Le mathématicien Wolfgang Doeblin avait de son côté ébauché une théorie similaire avant de se suicider à la défaite de son bataillon en juin 1940. Ses travaux furent envoyés à l'Académie des sciences dans un pli cacheté qui ne fut ouvert qu'en 2000.

Texte de Wikipedia, extrait de http://en.wikipedia.org/wiki/Timeline_of_probability_and_statistics.

Timeline of probability and statistics

Before 1600

- 9th Century - Al-Kindi was the first to use statistics to decipher encrypted messages and developed the first code breaking algorithm in the House of Wisdom in Baghdad, based on frequency analysis. He wrote a book entitled "Manuscript on Deciphering Cryptographic Messages", containing detailed discussions on statistics,
- 1560s (published 1663) - Cardano's Liber de ludo aleae attempts to calculate probabilities of dice throws,
- 1577 - Bartolomé de Medina defends probabilism, the view that in ethics one may follow a probable opinion even if the opposite is more probable.

17th century

- 1654 - Pascal and Fermat create the mathematical theory of probability,
- 1657 - Huygens's De ratiociniis in ludo aleae is the first book on mathematical probability,
- 1662 - Graunt's Natural and Political Observations Made upon the Bills of Mortality makes inferences from statistical data on deaths in London,
- 1693 - Halley prepares the first mortality tables statistically relating death rate to age.

18th century

- 1710 - Arbuthnot argues that the constancy of the ratio of male to female births is a sign of divine providence,
- 1713 - Posthumous publication of Jacob Bernoulli's Ars Conjectandi, containing the first derivation of a law of large numbers,
- 1724 - Abraham de Moivre studies mortality statistics and the foundation of the theory of annuities in Annuities on Lives,
- 1733 - Abraham de Moivre introduces the normal distribution to approximate the binomial distribution in probability,
- 1739 - Hume's Treatise of Human Nature argues that inductive reasoning is unjustified,
- 1761 - Thomas Bayes proves Bayes' theorem,
- 1786 - Playfair's Commercial and Political Atlas introduces graphs and bar charts of data.

19th century

- 1801 - Gauss predicts the orbit of Ceres using a line of best fit,
- 1805 - Adrien-Marie Legendre introduces the method of least squares for fitting a curve to a given set of observations,
- 1814 - Laplace's Essai philosophique sur les probabilités defends a definition of probabilities in terms of equally possible cases, introduces generating functions and Laplace transforms, uses conjugate priors for exponential families, proves an early version of the Bernstein - von Mises theorem on the asymptotic irrelevance of prior distributions on the limiting posterior distribution and the role of the Fisher information on asymptotically normal posterior modes,
- 1835 - Quetelet's Treatise on Man introduces social science statistics and the concept of the "average man",
- 1866 - Venn's Logic of Chance defends the frequency interpretation of probability,
- 1877 - 1883 - Charles Sanders Peirce outlines frequentist statistics, emphasizing the use of objective randomization in experiments and in sampling. Peirce also invented an optimally designed experiment for regression,
- 1880 - Thiele gives a mathematical analysis of Brownian motion, introduces the likelihood function, and invents cumulants,

- 1888 - Galton introduces the concept of correlation.

20th century

- 1900 - Bachelier analyzes stock price movements as a stochastic process,
- 1908 - Student's t-distribution for the mean of small samples published in English (following earlier derivations in German),
- 1921 - Keynes' Treatise on Probability defends a logical interpretation of probability. Wright develops path analysis,
- 1928 - Tippett and Fisher's introduce extreme value theory,
- 1933 - Andrey Nikolaevich Kolmogorov publishes his book Basic notions of the calculus of probability (Grundbegriffe der Wahrscheinlichkeitsrechnung) which contains an axiomatization of probability based on measure theory,
- 1935 - R. A. Fisher's Design of Experiments (1st ed),
- 1937 - Neyman introduces the concept of confidence interval in statistical testing,
- 1946 - Cox's theorem derives the axioms of probability from simple logical assumptions,
- 1948 - Shannon's Mathematical Theory of Communication defines capacity of communication channels in terms of probabilities,
- 1953 - Nicholas Metropolis introduces the idea of thermodynamic simulated annealing methods.

Bibliographie

[Bil79] P. Billingsley, *Probability and measure*, John Wiley and Sons, 1979. 1.5, 2.10, 4.10

[Jay03] E. T. Jaynes, *Probability Theory: the Logic of Science*, Cambridge University Press, 2003. 3.26, C.1

[LL04] E. Lehman, T. Leighton, *Mathematics for Computer Science*, 2004 2.18

[Pea88] J. Pearl, *Probabilistic reasoning in intelligent systems - Networks of plausible inference*, Morgan Kauffman, 1988. 4.27, 4.32

[Rom75] P. Roman, *Some modern mathematics for physicists and other outsiders*, Pergamon Press, 1975. 2.10, 4.13

[Sap90] G. Saporta, *Probabilités, analyse des données et statistique*, 2ème édtition, Technip, 2006. 3, 1.5, 3.21, 4.10

[Wil01] D. Williams, *Wighting the odds. A course in probability and statistics*, Cambridge University Press, 2001. 3