THÉORIE DES PROBABILITÉS: LES BASES (2025)

JUHAN ARU

[1]

Table des matières

1

0 Introduction **3**

1 Cadre de base **4**
1.1 Espace probabilisé 4
1.2 Définition mathématique d'un espace de probabilité 6
1.3 Quelques propriétés de base des espaces de probabilité 8
1.4 Variables aléatoires 10

2 Probabilité conditionnelle et indépendance **14**
2.1 Probabilité conditionnelle 14
2.2 Indépendance d'événements 17
2.3 Indépendance de variables aléatoires 19
2.4 Indépendance et produits d'espaces de probabilité 20

3 Variables aléatoires et vecteurs aléatoires **26**
3.1 La fonction de répartition d'une variable aléatoire 26
3.2 Exemples de variables aléatoires discrètes 28
3.3 Variables aléatoires continues 32
3.4 Vecteurs aléatoires 35

4 Espérance mathématique **39**
4.1 Espérance d'une variable aléatoire discrète 39
4.2 Espérance d'une variable aléatoire arbitraire 43
4.3 Espérance d'une fonction d'une variable aléatoire 46
4.4 Variance et covariance 49
4.5 Moments d'une variable aléatoire 51
4.6 Fonction génératrice des moments et fonction caractéristique 54
4.7 ⋆ Preuves de quelques résultats auxiliaires (non examinable) ⋆ 56

1. Version 2025. Tout retour est apprécié, y compris pour les fautes (plus ou moins importantes) — juhan.aru@epfl.ch. Il s'agit d'une troisième version de ces notes. Pour la rédaction des versions précédentes, j'ai consulté les notes de I. Manolescu (Fribourg), Y. Velenik (Genève), A. Eberle (Bonn) (toutes disponibles sur leurs sites), ainsi que le livre de R. Dalang et D. Conus publié par EPFL Press.

5 Théorèmes limites **59**
5.1 Collections infinies d'événements et de variables aléatoires 59
5.2 Convergence de variables aléatoires 61
5.3 Théorème central limite 67

Section 0

Introduction

Ce cours porte sur la théorie des probabilités : le cadre mathématique qui permet de formaliser nos questions sur les phénomènes aléatoires, et leur étude mathématique.

Lorsque nous souhaitons décrire un phénomène aléatoire du monde réel, nous construisons un modèle mathématique. Choisir un modèle utile — c'est-à-dire un modèle qui nous dit réellement quelque chose sur le monde et d'autre coté est suffisaement simple pour étudier — implique de nombreuses simplifications bien choisies et des décisions réfléchies. Par exemple, pour modéliser un lancer de pièce, nous écartons habituellement la possibilité qu'elle tombe sur la tranche, ou, en l'absence d'informations supplémentaires, nous considérons pile et face équiprobables, bien que cela puisse ne pas être le cas, ne serait-ce qu'à cause d'une répartition des masses différente. Mais choisir le modèle qui correspond le mieux au monde observé n'est pas le sujet central de ce cours.

Dans ce cours, nous nous concentrerons plutôt sur la mise en place du cadre mathématique général pour l'étude des phénomènes aléatoires, sur la formulation de modèles probabilistes, puis sur les outils mathématiques nécessaires et utiles pour étudier de tels modèles. Nous espérons aussi avoir un peu de temps pour discuter de modèles intéressants et pertinents.

Section 1

Cadre de base

Dans ce chapitre, nous discutons quelques notions fondamentales de la théorie des probabilités :

— Espace probabilisé
— Variables aléatoires
— Indépendance

1.1 Espace probabilisé

Notre premier objectif est de motiver la notion moderne d'espace de probabilité, ou de modèle probabiliste. Pour cela, considérons deux exemples :

(1) Un nombre aléatoire à valeurs dans $\{1, 2, \ldots, 12\}$, par exemple celui issu d'une loterie.

(2) Décrire la météo à Lausanne le lendemain.

Pour décrire ces deux phénomènes aléatoires, nous utiliserons encore le vocabulaire et les intuitions de tous les jours. Ensuite, nous donnerons des définitions mathématiques qui fixeront le vocabulaire pour le reste du cours.

(1) Nombre aléatoire. Dans le cadre moderne, pour décrire mathématiquement un nombre aléatoire, nous utilisons trois ingrédients :

— L'ensemble de tous les résultats possibles : ici $\Omega = \{1, 2, 3, \ldots, 12\}$.
— La collection de questions oui/non auxquelles on peut répondre à propos du résultat effectif, c'est-à-dire de ce nombre aléatoire. Par exemple :
  — Ce nombre est-il égal à 3 ?
  — Ce nombre est-il pair ?
  — Ce nombre est-il inférieur à 4 ?

  À chacune de ces questions, nous associons le sous-ensemble des résultats pour lesquels la réponse est « oui » : respectivement $\{3\}$, $\{2, 4, 6, 8, 10, 12\}$ ou $\{1, 2, 3\}$. Nous appelons chacun de ces sous-ensembles un *événement*.
— Enfin, à chaque événement $E \subseteq \Omega$, nous souhaitons associer une valeur numérique $\mathbb{P}(E) \in [0, 1]$ que nous appelons *probabilité*. Elle doit correspondre à la proportion de fois où l'événement se produit si l'expérience est répétée un grand nombre de fois, par exemple si la loterie est rejouée de nombreuses fois. [2]

Ici, l'ensemble des résultats possibles était facile et directement donné par le problème. Il est aussi naturel de supposer que tout sous-ensemble $E \subseteq \Omega$ est un événement — autrement dit, pour tout E, on peut poser la question : « le nombre est-il dans E ? ». Cela signifie que nous pouvons prendre la collection des événements comme étant l'ensemble de toutes les parties de Ω.

La détermination de la probabilité dépend véritablement de ce que nous voulons modéliser — par exemple, si nous essayons de modéliser la loterie, nous pouvons supposer que tous les nombres sont équiprobables ; nous retrouvons alors le modèle peut-être déjà vu au gymnase

2. En réalité, on utilise aussi des modèles probabilistes pour décrire des phénomènes qui ne se produisent qu'une seule fois. Dans ce cas, la probabilité mesure en quelque sorte notre degré de croyance.

ou au lycée : on pose $\mathbb{P}(E) = |E|/|\Omega|$. En revanche, si nous voulions décrire la somme de deux dés, il faudrait choisir les nombres $\mathbb{P}(E)$ très différemment ![3]

Maintenant, si nous voulons que notre modèle corresponde à la notion intuitive de probabilité et prédise la fréquence lors d'expériences répétées, ces choix ne sont pas totalement libres — il faut ajouter certaines contraintes. Par exemple, nous ne pouvons pas définir une fonction $\mathbb{P}$ arbitraire : en effet, si l'on a deux événements $E_1 \subseteq E_2$, alors on devrait avoir $\mathbb{P}(E_1) \leq \mathbb{P}(E_2)$, puisque chaque fois que E_1 se réalise, E_2 se réalise aussi. On devrait aussi avoir $\mathbb{P}(\Omega) = 1$, puisque « quelque chose » arrive toujours, et en outre l'on doit avoir $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ si E et F sont disjoints (pourquoi ?). Bien sûr, toutes ces contraintes ne sont pas indépendantes — certaines peuvent en impliquer d'autres ; lorsque nous donnerons la définition d'un espace de probabilité ci-dessous, nous purifierons ces conditions et n'en choisirons que quelques-unes qui impliqueront mathématiquement toutes les autres.

(2) Météo à Lausanne le lendemain. Dans le cadre moderne que nous allons définir, nous voudrions à nouveau prendre trois décisions, mais ici la tâche est déjà plus difficile dès la première étape. Quel devrait être l'espace des états ? Un espace naturel pourrait être l'ensemble de tous les états microscopiques possibles de l'atmosphère jusqu'à 20 km d'altitude au-dessus de Lausanne… mais nous avons là de nombreux choix arbitraires — pourquoi 20 km, quelle largeur au-dessus du Léman, etc. ? Et, de toute façon, tout état « naturel » serait d'une complexité impossible !

Heureusement, nous n'avons pas vraiment besoin de nous en préoccuper — dans notre cadre, nous avons seulement à assigner des probabilités à tous les événements de la collection d'événements choisie ! Et nous avons une certaine liberté pour choisir cette collection d'événements — elle peut être déterminée par notre capacité à mesurer les états, par exemple nous pouvons mesurer la température à une certaine précision, ou la densité de molécules de CO_2 ou d'eau à une certaine précision ; cela détermine des sous-ensembles de l'espace d'états. De plus, si nous améliorons notre mesure, nous pouvons toujours agrandir notre modèle de manière incrémentale !

Cependant, comme pour la fonction de probabilité, il existe aussi des conditions naturelles de cohérence pour la collection d'événements : nous supposons que si l'on peut observer si l'événement E s'est produit, nous devrions aussi être capables de mesurer si son complémentaire E^c s'est produit. Ou encore, si nous sommes capables de dire si E est arrivé, ou si F est arrivé, nous devrions pouvoir dire si l'un des deux est arrivé — c'est-à-dire que $E \cup F$ devrait aussi être un événement. Et il s'avère que c'est tout ce dont nous avons besoin !

Naturellement, la mise en place des probabilités pour ce modèle est aussi horriblement compliquée — il n'y a pas d'hypothèses naturelles de symétrie comme celle utilisée pour la distribution uniforme. Même le meilleur physicien du monde ne sera pas capable de décrire la distribution de probabilité naturelle de tous les états microscopiques de l'atmosphère, d'autant plus qu'elle dépend fortement de ce qui s'est passé juste avant ! Ainsi, notre seul choix consiste essentiellement à essayer d'utiliser une combinaison de nos connaissances sur les processus atmosphériques et de nos observations passées pour établir des estimations pour le modèle ; puis naturellement nous essaierons de l'améliorer chaque jour suivant. Heureusement, cette tâche difficile ne nous incombe pas, mais revient aux services de météo et aux statisticiens !

3. Voir feuille d'exercices 1.

Remarque 1.1. *Enfin, avant de donner les définitions mathématiques, soulignons à nouveau que les trois composants du modèle — l'espace des états, l'ensemble des événements et leurs probabilités — sont des* entrées *que nous choisissons pour construire notre modèle. Lorsqu'on tente de modéliser un phénomène réel, on effectue généralement des simplifications pour chacun de ces choix. Par exemple, pour le lancer de pièce, nous n'utilisons que deux issues : pile et face, même si théoriquement la tranche est possible. Nous fixons aussi habituellement les probabilités à un demi, bien que cela ne soit pas exactement vrai non plus.*

1.2 Définition mathématique d'un espace de probabilité

Nous sommes maintenant prêts à passer par notre filtre mathématique et à donner une définition mathématique d'un espace de probabilité. En fait, nous commençons par présenter une définition mathématique dans un cadre restreint où Ω est un ensemble fini, puis nous la généralisons.

Définition 1.2 (Espace de probabilité fini, Kolmogorov 1933)**.** *Un espace de probabilité fini est un triplet $(\Omega, \mathcal{F}, \mathbb{P})$, où*

- *Ω est un ensemble fini, appelé l'espace d'états, espace d'échantillonnage ou l'univers.*
- *$\mathcal{F}$ est un ensemble de sous-ensembles de Ω, satisfaisant :*
  - *$\emptyset \in \mathcal{F}$;*
  - *si $A \in \mathcal{F}$, alors $A^c \in \mathcal{F}$;*
  - *si $A_1, A_2 \in \mathcal{F}$, alors aussi $A_1 \cup A_2 \in \mathcal{F}$.*

  $\mathcal{F}$ est appelée la collection d'événements et tout $A \in \mathcal{F}$ est un événement.
- *Enfin, on a une fonction $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfaisant $\mathbb{P}(\Omega) = 1$ et l'additivité pour les ensembles disjoints : si $A_1, A_2 \in \mathcal{F}$ sont disjoints, alors*

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

  Cette fonction $\mathbb{P}$ est appelée la probabilité.

Remarquez que certaines propriétés évoquées plus haut, comme le fait que pour des événements $E_1 \subseteq E_2$, on ait $\mathbb{P}(E_1) \leq \mathbb{P}(E_2)$, découlent directement de la définition. [4]

La plupart des phénomènes du monde réel peuvent être décrits par des ensembles finis simplement parce que nous ne pouvons mesurer les choses qu'à un niveau de précision fini. Cependant, de la même façon que la notion de fonction continue ou dérivable permet de simplifier nos descriptions mathématiques de la réalité et donc d'améliorer notre compréhension, les espaces de probabilité continus rendent aussi les descriptions mathématiques plus nettes, plus simples et, ainsi, facilitent l'étude des phénomènes aléatoires sous-jacents.

Quelques exemples naturels où des espaces d'échantillonnage infinis apparaissent :

- un point uniforme sur un segment (par exemple issu de la fracturation d'un bâton en plusieurs morceaux) ;
- la position dans la rue où tombe la première goutte de pluie de la journée ;
- l'espace de toutes les suites infinies de lancers de pièce.

Dans tous ces cas, l'espace d'états mathématiquement naturel est même non dénombrable. Des espaces d'états dénombrables apparaissent aussi : par exemple, si l'on modélise le premier instant où des lancers répétés d'une pièce donnent « pile », la valeur peut être $1, 2, 3$ ou, avec

4. Voir feuille d'exercices 1.

une probabilité très, très petite, aussi 10^{10} ; un espace d'états naturel contiendrait donc tous les entiers naturels.

Énonçons donc la définition générale :

Définition 1.3 (Espace de probabilité, Kolmogorov 1933)**.** *Un espace de probabilité est un triplet $(\Omega, \mathcal{F}, \mathbb{P})$, où*

- *Ω est un ensemble, appelé l'espace d'états, espace d'échantillonnage ou l'univers.*
- *$\mathcal{F}$ est un ensemble de sous-ensembles de Ω, satisfaisant :*
  - *$\emptyset \in \mathcal{F}$;*
  - *si $A \in \mathcal{F}$, alors $A^c \in \mathcal{F}$;*
  - *si $A_1, A_2, \cdots \in \mathcal{F}$, alors $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.*

  $\mathcal{F}$ est appelée la collection d'événements, ou une σ-algèbre, et tout $A \in \mathcal{F}$ est un événement.
- *Enfin, on a une fonction $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfaisant $\mathbb{P}(\Omega) = 1$ et l'additivité dénombrable pour des ensembles disjoints : si $A_1, A_2, \cdots \in \mathcal{F}$ sont deux à deux disjoints,*

$$\mathbb{P}\Big(\bigcup_{n\geq 1} A_n\Big) = \sum_{n\geq 1} \mathbb{P}(A_n).$$

  Cette fonction $\mathbb{P}$ est appelée la probabilité.

Notez les seules différences : 1) nous ne supposons pas Ω fini ; 2) nous supposons que l'ensemble des événements est stable par unions dénombrables ; 3) nous supposons aussi l'additivité de la probabilité sous unions dénombrables.

Exercice 1.1. *Montrer que tout espace de probabilité « élémentaire » est un espace de probabilité.*

En fait, les espaces de probabilité sont un exemple de la notion générale d'espaces mesurés — les espaces de probabilité ne sont que des espaces mesurés de masse totale égale à 1.

Définition 1.4 (Espace mesuré, Borel 1898, Lebesgue 1901–1903)**.** *Un espace mesuré est un triplet $(\Omega, \mathcal{F}, \mu)$, où*

- *Ω est un ensemble, appelé l'espace d'échantillonnage ou l'univers.*
- *$\mathcal{F}$ est un ensemble de sous-ensembles de Ω, satisfaisant :*
  - *$\emptyset \in \mathcal{F}$;*
  - *si $A \in \mathcal{F}$, alors $A^c \in \mathcal{F}$;*
  - *si $A_1, A_2, \cdots \in \mathcal{F}$, alors $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.*

  $\mathcal{F}$ est appelée une σ-algèbre ou tribu et tout $A \in \mathcal{F}$ un ensemble mesurable.
- *Enfin, on a une fonction $\mu : \mathcal{F} \to [0, \infty]$ satisfaisant $\mu(\emptyset) = 0$ et l'additivité dénombrable pour des ensembles disjoints : si $A_1, A_2, \cdots \in \mathcal{F}$ sont deux à deux disjoints,*

$$\mu\Big(\bigcup_{n\geq 1} A_n\Big) = \sum_{n\geq 1} \mu(A_n).$$

  Cette fonction μ est appelée une mesure. Si $\mu(\Omega) < \infty$, on appelle μ une mesure finie.

Géométriquement, nous interprétons :

- Ω comme notre espace de points,
- $\mathcal{F}$ comme la collection des sous-ensembles pour lesquels notre notion de volume peut être définie,

— μ comme notre notion de volume : elle donne à chaque ensemble mesurable son volume. Il est important d'établir ce lien avec la théorie de la mesure, car nombre de propriétés des espaces de probabilité en découlent directement. Il est toutefois bon de garder à l'esprit que la théorie des probabilités n'est pas *que* de la théorie de la mesure — comme l'a bien formulé M. Kac, « La théorie des probabilités, c'est la théorie de la mesure avec une âme », et nous adhérons à cette remarque philosophique.

Remarque 1.5. *Vous devriez comparer la définition d'un espace de probabilité / espace mesuré avec celle d'un espace topologique : là aussi, on utilise une collection de sous-ensembles, satisfaisant certaines propriétés, pour doter un ensemble d'une structure. Une question à se poser : pourquoi exigeons-nous exactement des unions et intersections* dénombrables *pour les événements, et non pas seulement finies, ou bien arbitraires ?*

1.3 Quelques propriétés de base des espaces de probabilité

Commençons par quelques remarques sur la définition d'un espace de probabilité :

Remarque 1.6. *Il vaut la peine de réfléchir aux raisons pour lesquelles on demande la stabilité dénombrable de la σ-algèbre ou l'additivité dénombrable de la probabilité. Même s'il s'agit davantage d'une question méta-mathématique, il est bon de la garder à l'esprit tout au long du cours. Contentons-nous ici de deux observations simples.*

Premièrement, les sommes dénombrables apparaissent naturellement lorsqu'on prend des limites de sommes finies. En fait, l'additivité dénombrable peut être vue comme équivalente à une certaine forme de continuité pour la probabilité (voir ci-dessous).

Deuxièmement, autoriser des unions arbitraires mène rapidement aux ensembles de parties, et des sommes non dénombrables de termes positifs ne peuvent pas être finies (voir la feuille d'exercices).

Exercice 1.2. *Montrer que l'additivité dénombrable dans les axiomes d'un espace de probabilité peut être remplacée par l'additivité finie plus l'énoncé suivant : pour toute suite décroissante d'événements $E_1 \supseteq E_2 \supseteq E_3 \dots$ telle que $\cap_{i\geq 1} E_i = \emptyset$, on a $\mathbb{P}(\cap_{i=1}^n E_i) \to 0$ lorsque $n \to \infty$.*
$\star$ *Cela vaut-il dans un espace mesuré général ?*

Considérons aussi un autre cadre qui explique bien l'utilité des σ-algèbres :

Remarque 1.7. *Dans la vie réelle, nous n'obtenons souvent des informations sur le monde que pas à pas, et si nous souhaitons continuer à travailler sur le même espace de probabilité (ce qui est avantageux car alors $\mathbb{P}$ n'aura besoin que d'être* étendue *et non redéfinie), nous pouvons considérer une suite de σ-algèbres $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \dots$ appelée* filtration *— chaque jour, nous pouvons poser davantage de questions oui/non, parce que, par exemple, nous savons déjà ce qui s'est passé la veille et avons peut-être appris quelque chose de nouveau. Toute l'information possible est contenue dans l'ensemble des parties $\mathcal{P}(\Omega)$.*

On classe habituellement les espaces de probabilité en deux types :

Définition 1.8 (Espaces de probabilité discrets et continus)**.** *Les espaces de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ dont l'espace d'états Ω est dénombrable sont appelés* discrets*, et ceux dont Ω est non dénombrable sont appelés* continus*.*

Dans ce cours, nous travaillerons principalement avec des espaces discrets, car ils sont techniquement plus simples. Cependant, les espaces continus apparaissent naturellement et nous ne pourrons pas non plus les éviter totalement.

Leur différence technique peut être résumée dans la proposition suivante, dont la preuve (non exigible) est laissée aux enthousiastes.

Proposition 1.9. *Soit Ω dénombrable et $\mathcal{F}$ une σ-algèbre sur Ω. Alors on peut trouver des événements disjoints $E_1, E_2, \cdots \in \mathcal{F}$ tels que, pour tout $E \in \mathcal{F}$, on ait la décomposition $E = \cup_{i \in I_E} E_i$.*

Essentiellement, cela dit que, pour tout espace de probabilité discret, il suffit de déterminer $\mathbb{P}(E_i)$ pour une collection dénombrable d'ensembles disjoints E_i, puis, pour tout autre ensemble E, on peut utiliser l'additivité dénombrable pour étendre $\mathbb{P}$. Notez que cela signifie qu'il est d'abord facile de vérifier si une fonction $\mathbb{P}$ donnée satisfait bien tous les axiomes et, plus important encore, il est facile de vérifier quand deux mesures de probabilité sont égales.

Pour les espaces continus, cela n'est pas nécessairement vrai — les σ-algèbres utiles sont généralement plus compliquées. Pour illustrer pourquoi on ne souhaite pas forcément utiliser l'ensemble des parties, considérons la proposition suivante (sa preuve est en annexe et repose sur l'axiome du choix) :

Proposition 1.10. *Il n'existe pas de probabilité $\mathbb{P}$ sur $([0,1], \mathcal{P}([0,1]))$ invariante par translation, c'est-à-dire telle que pour tout $A \in \mathcal{P}([0,1])$ et $\alpha \in [0,1)$, on ait $\mathbb{P}(A+\alpha \bmod 1) = \mathbb{P}(A)$, où $A + \alpha \bmod 1 := \{a + \alpha \bmod 1 : a \in A\}$ désigne l'ensemble obtenu en décalant A de α modulo 1.*

En fait, il s'avère que la seule manière de remédier à cette situation consiste à rendre la σ-algèbre pertinente plus petite. Nous souhaitons encore pouvoir répondre « oui » ou « non » à des questions telles que : « mon nombre aléatoire est-il égal à x ? » ou « est-il dans un intervalle (a, b) ? ». Grâce au fait que nous ne disposons que de l'additivité dénombrable, cela n'implique pas que notre σ-algèbre doive être l'ensemble des parties. Et grâce aux propriétés des σ-algèbres, on peut toujours construire au moins une σ-algèbre contenant tous nos ensembles favoris — voir la feuille d'exercices.

Énonçons maintenant quelques conséquences immédiates des définitions sur les σ-algèbres et les mesures de probabilité :

Lemme 1.11 (Stabilité de la σ-algèbre)**.** *Considérons un ensemble Ω muni d'une σ-algèbre $\mathcal{F}$.*

(1) Si $A_1, A_2, \cdots \in \mathcal{F}$, alors $\bigcap_{n \geq 1} A_n \in \mathcal{F}$.

(2) On a aussi $\Omega \in \mathcal{F}$ et, si $A, B \in \mathcal{F}$, alors $A \setminus B \in \mathcal{F}$.

(3) Pour tout $n \geq 1$, si $A_1, \ldots, A_n \in \mathcal{F}$, alors $A_1 \cup \cdots \cup A_n \in \mathcal{F}$ et $A_1 \cap \cdots \cap A_n \in \mathcal{F}$.

Preuve du Lemme 1.11. Par les lois de De Morgan, pour toute famille $(A_i)_{i \in I}$,

$$\bigcap_{i \in I} A_i = \big(\bigcup_{i \in I} A_i^c\big)^c.$$

Le point (1) en découle : si $A_1, A_2, \cdots \in \mathcal{F}$, alors, par définition d'une σ-algèbre, $A_1^c, A_2^c, \cdots \in \mathcal{F}$ et donc

$$\big(\bigcup_{i \geq 1} A_i^c\big)^c \in \mathcal{F}.$$

Pour (3), à nouveau par De Morgan, il suffit de montrer que $A_1 \cup \cdots \cup A_n \in \mathcal{F}$. Mais cela découle de la définition d'une σ-algèbre, car $A_1 \cup \cdots \cup A_n = \bigcup_{i\geq 1} A_i$ avec $A_k = \emptyset$ pour $k \geq n+1$.

Le point (2) est laissé en exercice. □

De même, les conditions de base sur la mesure donnent lieu à plusieurs propriétés naturelles :

Proposition 1.12 (Propriétés de base d'une probabilité)**.** *Considérons un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Soient $A_1, A_2, \cdots \in \mathcal{F}$. Alors*

(1) Pour tout $A \in \mathcal{F}$, on a $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(2) Pour tout $n \geq 1$, et $A_1, \ldots, A_n$ disjoints, on a l'additivité finie :

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

En particulier, si $A_1 \subseteq A_2$, alors $\mathbb{P}(A_1) \leq \mathbb{P}(A_2)$.

(3) Si, pour tout $n \geq 1$, $A_n \subseteq A_{n+1}$, alors, quand $n \to \infty$, on a $\mathbb{P}(A_n) \to \mathbb{P}(\bigcup_{k\geq 1} A_k)$.

(4) On a la sous-additivité dénombrable (ou union bound*) : $\mathbb{P}(\bigcup_{n\geq 1} A_n) \leq \sum_{n\geq 1} \mathbb{P}(A_n)$.*

(5) Si, pour tout $n \geq 1$, $A_n \supseteq A_{n+1}$, alors, quand $n \to \infty$, on a $\mathbb{P}(A_n) \to \mathbb{P}(\bigcap_{k\geq 1} A_k)$.

Démonstration. Les propriétés (1), (4) et la seconde partie de (2) figurent sur la feuille d'exercices 1. La première partie de (2) suit comme dans le lemme ci-dessus en prenant $A_{n+1} = A_{n+2} = \cdots = \emptyset$ et en utilisant l'additivité dénombrable.

Prouvons (3) : posez $B_1 = A_1$ et, pour $n \geq 2$, $B_n = A_n \setminus A_{n-1}$. Alors les B_n sont disjoints, $\bigcup_{n=1}^N B_n = A_N$ et $\bigcup_{n\geq 1} B_n = \bigcup_{n\geq 1} A_n$.

Par additivité dénombrable,

$$\mathbb{P}\Big(\bigcup_{i\geq 1} A_i\Big) = \mathbb{P}\Big(\bigcup_{i\geq 1} B_i\Big) = \sum_{i\geq 1} \mathbb{P}(B_i).$$

Mais $\mathbb{P}$ est positive, donc

$$\sum_{i\geq 1} \mathbb{P}(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i).$$

Par additivité dénombrable encore,

$$\sum_{i=1}^{n} \mathbb{P}(B_i) = \mathbb{P}\Big(\bigcup_{i=1}^{n} B_i\Big) = \mathbb{P}(A_n),$$

d'où (3). □

1.4 Variables aléatoires

En réalité, lorsque nous étudions un phénomène aléatoire, nous ne voulons certainement pas nous restreindre à des questions oui/non. Par exemple, dans notre modèle d'un nombre aléatoire parmi $\{1, 2, \ldots, 12\}$, la question naturelle n'est pas « est-ce que ce nombre est 5 ? » mais plutôt « quel est ce nombre ? ». De même, dans l'exemple sur la météo, il est plus naturel de demander « quelle est la température ? », « quelle quantité de pluie y aura-t-il l'après-midi ? ».

Ces observations numériques sur notre phénomène aléatoire seront formalisées sous le nom de *variables aléatoires*. En substance, elles associent un nombre à chaque état et sont donc des fonctions $X : \Omega \to \mathbb{R}$. Toutefois, nous ne souhaitons pas forcément autoriser toutes ces fonctions pour des raisons de cohérence. En effet, nous voulons pouvoir poser des questions oui/non à propos de nos quantités aléatoires, par exemple « la variable vaut-elle 3 ? », « la température dépasse-t-elle 18 ? ». Or, la réponse oui/non correspond à certains sous-ensembles d'états dans l'univers et, en tant que tels, ces sous-ensembles devraient être des événements de notre modèle. Il existe donc un lien entre la collection d'événements et la collection de fonctions pouvant jouer le rôle de variables aléatoires. Donnons sans plus tarder la définition générale :

Définition 1.13 (Variable aléatoire)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Une fonction $X : \Omega \to \mathbb{R}$ est appelée une* variable aléatoire *si, pour tout intervalle (a, b), l'ensemble $X^{-1}((a, b)) := \{\omega \in \Omega : X(\omega) \in (a, b)\}$ est un événement, c'est-à-dire appartient à $\mathcal{F}$.*

Il y a une simplification dans le cas des espaces de probabilité discrets :

Lemme 1.14 (Variables aléatoires sur les espaces discrets)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité discret. Alors $X : \Omega \to \mathbb{R}$ est une variable aléatoire si et seulement si, pour tout $y \in \mathbb{R}$, on a $X^{-1}(\{y\}) \in \mathcal{F}$.*

Démonstration. Ceci se vérifie soigneusement à partir des définitions et figurera sur la feuille d'exercices. □

Pour les esprits « structurés », la définition d'une variable aléatoire peut paraître quelque peu arbitraire. En effet, j'ai caché une information importante — la collection naturelle d'événements sur $\mathbb{R}$ esquissée un peu plus haut. Nous l'énoncerons directement sur $\mathbb{R}^n$.

Définition 1.15 (σ-algèbre borélienne)**.** *La plus petite σ-algèbre sur $\mathbb{R}^n$ contenant toutes les boîtes ouvertes de la forme $(a_1, b_1) \times \cdots \times (a_n, b_n)$ est appelée la* σ-algèbre borélienne. *On la note $\mathcal{F}_B$.*

Remarque 1.16. *Cette définition est en fait encore plus générale : étant donné un espace topologique (X, τ), la plus petite σ-algèbre contenant tous les ouverts est appelée σ-algèbre borélienne. Vous verrez sur la feuille d'exercices que cette définition plus générale se réduit à la précédente dans le cas de $\mathbb{R}^n$ muni de sa topologie euclidienne.*

Sur cette base, une définition équivalente, peut-être plus « structurelle », d'une variable aléatoire est la suivante : une fonction $X : \Omega \to \mathbb{R}$ est une variable aléatoire si la préimage de tout borélien de $\mathbb{R}$ par X est un événement. [5]

Une notion importante associée aux variables aléatoires est leur *loi* :

Lemme 1.17 (Loi d'une variable aléatoire)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $X : \Omega \to \mathbb{R}$ une variable aléatoire.*

Alors il existe une probabilité $\mathbb{P}_X$ induite sur $(\mathbb{R}, \mathcal{F}_B)$ en définissant $\mathbb{P}_X(F) := \mathbb{P}(X^{-1}(F))$ pour tout $F \in \mathcal{F}_B$. Cette probabilité $\mathbb{P}_X$ est appelée la loi *(ou* distribution*) de la variable aléatoire X.*

5. En théorie de la mesure, de telles fonctions sont appelées *mesurables* de $(\Omega, \mathcal{F})$ vers $(\mathbb{R}, \mathcal{F}_B)$; notez la similarité avec la définition de continuité en topologie.

C'est un lemme et non une simple définition, car il faut prouver que $\mathbb{P}_X$ est bien une probabilité sur $(\mathbb{R}, \mathcal{F}_B)$.

Preuve du Lemme. Il faut vérifier les axiomes d'une probabilité :
— $\mathbb{P}_X(\mathbb{R}) = \mathbb{P}(\Omega) = 1$;
— de même, $\mathbb{P}_X(F) = \mathbb{P}(X^{-1}(F)) \in [0,1]$ pour tout $F \in \mathcal{F}_B$;
— enfin, l'additivité dénombrable : si $F_1, F_2, \dots$ sont disjoints dans $\mathcal{F}_B$, alors

$$\mathbb{P}_X\Big(\bigcup_{i\geq 1} F_i\Big) = \mathbb{P}\Big(X^{-1}\big(\bigcup_{i\geq 1} F_i\big)\Big) = \mathbb{P}\Big(\bigcup_{i\geq 1} X^{-1}(F_i)\Big) = \sum_{i\geq 1} \mathbb{P}\big(X^{-1}(F_i)\big) = \sum_{i\geq 1} \mathbb{P}_X(F_i).$$

Ici, nous avons utilisé la définition aux première et dernière égalités, les propriétés des préimages à la seconde, et le fait que les $X^{-1}(F_i)$ sont disjoints, ainsi que l'additivité dénombrable, à la troisième.

□

En termes simples, nous avons montré que chaque variable aléatoire X induit une probabilité sur les réels en oubliant tout le contexte et en nous concentrant uniquement sur le nombre observé. Par exemple, dans le cas de la météo à Lausanne, la température fournit une variable aléatoire et, en ne regardant que sa valeur et rien d'autre, nous obtenons simplement un nombre réel aléatoire. Plus simplement encore, si nous lançons deux pièces équilibrées et comptons le nombre de piles, leur somme fournit une variable aléatoire à valeurs dans $\{0,1,2\}$. La notion de loi d'une variable aléatoire nous donne ainsi un moyen de comparer des quantités aléatoires provenant de contextes très différents.

Définition 1.18 (Égalité en loi)**.** *Soient X, Y deux variables aléatoires définies éventuellement sur des espaces de probabilité différents. Nous disons que X et Y sont* égales en loi *(ou* égales en distribution*), et nous notons $X \sim Y$, si pour tout $E \in \mathcal{F}_B$, on a $\mathbb{P}_X(E) = \mathbb{P}_Y(E)$.*

Nous soulignons que, lorsqu'on regarde la loi d'une variable aléatoire, le contexte est oublié — nous nous concentrons seulement sur la valeur numérique et l'espace de probabilité initial $(\Omega, \mathcal{F}, \mathbb{P})$ ne sert qu'à déterminer $\mathbb{P}_X$ mais ne joue plus de rôle par la suite. Cela permet de relier entre elles différentes situations aléatoires. Par exemple, les fonctions indicatrices de tout événement de probabilité p, quel que soit l'espace de probabilité où elles ont été définies, ont la même loi. Plus concrètement, par exemple, les variables aléatoires suivantes ont la même loi :
— le nombre de piles dans deux lancers indépendants ;
— le nombre de facteurs premiers lorsque l'on choisit uniformément un nombre dans $\{1,2,3,4\}$.

Dans une certaine mesure, une grande partie de ce cours consistera à étudier et décrire les lois de variables aléatoires.

Il existe également d'autres notions d'égalité pour des variables aléatoires :
— Nous disons que deux variables X, Y définies sur le même espace $(\Omega, \mathcal{F}, \mathbb{P})$ sont *partout égales* si, pour tout $\omega \in \Omega$, $X(\omega) = Y(\omega)$.
— Nous disons que deux variables X, Y définies sur le même espace $(\Omega, \mathcal{F}, \mathbb{P})$ sont *presque sûrement égales* si $\mathbb{P}(\{\omega : X(\omega) = Y(\omega)\}) = 1$. Ici, il faut bien sûr d'abord montrer que $\{\omega : X(\omega) = Y(\omega)\}$ est un événement, c'est-à-dire appartient à $\mathcal{F}$ — cela figure sur la feuille d'exercices.

Clairement, l'égalité partout implique l'égalité presque sûre, et la réciproque est fausse — par exemple, considérez les variables $X(\omega) := \omega \, \mathbf{1}_{\omega \neq 1/2}$ et $Y(\omega) := \omega$ sur l'espace $([0,1], \mathcal{F}_B, \mathbb{P}_U)$. Il est aussi clair que l'égalité en loi ne peut pas impliquer l'égalité p.s., même si les variables sont définies sur le même espace. Enfin,

Exercice 1.3. *Soient X, Y deux variables aléatoires définies sur le même espace de probabilité et presque sûrement égales. Alors elles sont aussi égales en loi.*

Section 2

Probabilité conditionnelle et indépendance

De manière générale, si nous apprenons quelque chose de nouveau sur notre phénomène aléatoire, cette connaissance influence — et modifie souvent — nos prédictions pour le reste du modèle.

— Par exemple, dans le cas d'un nombre uniformément aléatoire entre 1 et 12, si quelqu'un vous dit que ce nombre est pair, alors la probabilité d'observer 1 devient soudain 0, tandis que la probabilité d'observer 2 passe de 1/12 à 1/6.

— Dans le cas de la météo à Lausanne, si l'on vous dit qu'il pleut toute la journée, alors il est moins probable qu'il fasse aussi plus de 35 degrés.

Le but de cette section est de fixer le vocabulaire permettant de parler de la manière dont la connaissance d'un événement ou d'une variable aléatoire influence les probabilités que nous devrions attribuer à d'autres événements. Cela nous conduit à parler de probabilités conditionnelles, puis à discuter du cas où les événements ne s'influencent pas, donnant naissance à une notion importante en théorie des probabilités : l'indépendance.

2.1 Probabilité conditionnelle

Nous avons déjà considéré (dans le cours et sur les feuilles d'exercices) de nombreuses situations imprévisibles où plusieurs événements se produisent naturellement soit en même temps, soit successivement : une suite de lancers de pièce ou des pas successifs d'une marche aléatoire, ou encore des liens/arêtes différents dans un graphe aléatoire. Dans tous ces cas, le fait qu'un événement se soit produit peut facilement influencer les autres. Par exemple, si vous souhaitez modéliser les marchés financiers de demain, il semble plutôt avisé de tenir compte de ce qui s'est passé aujourd'hui. Pour parler du changement des probabilités lorsqu'on a observé quelque chose, on introduit la notion de probabilité conditionnelle :

Définition 2.1 (Probabilité conditionnelle)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $E \in \mathcal{F}$ avec $\mathbb{P}(E) > 0$. Alors, pour tout $F \in \mathcal{F}$, on définit la probabilité conditionnelle de l'événement F sachant E (c.-à-d. sachant que l'événement E a lieu) par*

$$\mathbb{P}(F|E) := \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(E)}.$$

Rappelez-vous que $E \cap F$ est l'événement « E et F se produisent ». Comme le dénominateur est toujours $\mathbb{P}(E)$, la probabilité conditionnelle sachant E est proportionnelle à $\mathbb{P}(E \cap F)$ pour tout événement F. Voici la justification de la division par $\mathbb{P}(E)$:

Lemme 2.2. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $E \in \mathcal{F}$ avec $\mathbb{P}(E) > 0$. Alors $P(\cdot|E)$ définit une probabilité sur $(\Omega, \mathcal{F})$, appelée la probabilité conditionnelle sachant E.*

Démonstration. Tout d'abord, remarquez que $\mathbb{P}(\cdot|E)$ est bien définie pour tout $F \in \mathcal{F}$. Ensuite, $\mathbb{P}(\emptyset|E) = \mathbb{P}(\emptyset)/\mathbb{P}(E) = 0$ et $\mathbb{P}(\Omega|E) = \mathbb{P}(E)/\mathbb{P}(E) = 1$. Il reste à vérifier l'additivité dénombrable.

Soient $F_1, F_2, \cdots \in \mathcal{F}$ disjoints. Alors $E \cap F_1, E \cap F_2, \ldots$ sont aussi disjoints. D'où

$$\mathbb{P}\Big(\bigcup_{i\geq 1} F_i \,\Big|\, E\Big) = \frac{\mathbb{P}\big((\bigcup_{i\geq 1} F_i) \cap E\big)}{\mathbb{P}(E)} = \frac{\mathbb{P}\big(\bigcup_{i\geq 1}(F_i \cap E)\big)}{\mathbb{P}(E)} = \sum_{i\geq 1} \frac{\mathbb{P}(F_i \cap E)}{\mathbb{P}(E)} = \sum_{i\geq 1} \mathbb{P}(F_i|E),$$

ce qui donne l'additivité dénombrable. □

On notera que la probabilité conditionnelle d'un événement peut parfois être proche de la probabilité initiale (nous y reviendrons très vite), mais elle peut aussi être très différente. Un exemple un peu « bête », mais instructif, est le suivant :

— La probabilité conditionnelle de E^c, conditionnée par E, est toujours nulle, quelle que soit la probabilité initiale ;
— de même, la probabilité conditionnelle de E, conditionnée par E, est toujours 1.

Ou, pour un exercice plus sensé, considérez ce qui suit :

Exercice 2.1 (Marche aléatoire et probabilités conditionnelles)**.** *Considérez la marche aléatoire simple de longueur n.*

— *Quelle est la probabilité que la marche soit au point n au temps n ? Maintenant, supposez que le premier pas ait été -1. Quelle est alors la probabilité d'être au point n au temps n ?*
— *Supposons n pair. Quelle est la probabilité que la marche soit au point 0 au temps n ? Maintenant, supposez que le premier pas ait été -1. Quelle est alors la probabilité d'être au point 0 au temps n ?*

Il faut aussi être très attentif au conditionnement exact, car deux conditionnements qui « se ressemblent » peuvent induire des probabilités conditionnelles très différentes. En général, il nous faut des informations supplémentaires sur la relation entre deux événements pour savoir comment la probabilité de l'un change quand on conditionne par l'autre.

Il existe néanmoins des cas où ces relations, et donc les probabilités conditionnelles, sont simples :

— Lorsque $E \subseteq F$, la probabilité conditionnelle de F sachant E vaut 1.
— Lorsque $F \subseteq E^c$, la probabilité conditionnelle de F sachant E vaut 0.
— Le troisième cas est lorsque F et E sont dits indépendants : dans ce cas $\mathbb{P}(F|E) = \mathbb{P}(F)$, essentiellement par définition (nous y revenons).

En général, il n'existe pas tant d'outils pour calculer des probabilités conditionnelles, mais il en existe un très utile : la formule de Bayes.

2.1.1 Formule de Bayes

Proposition 2.3 (Formule de Bayes)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et E, F deux événements de probabilité strictement positive. Alors*

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(F|E)\mathbb{P}(E)}{\mathbb{P}(F)}.$$

La preuve tient en une ligne : par définition de la probabilité conditionnelle,

$$\mathbb{P}(E|F)\mathbb{P}(F) = \mathbb{P}(E \cap F) = \mathbb{P}(F|E)\mathbb{P}(E).$$

Il s'agit pourtant d'une observation très utile — elle permet non seulement de calculer, mais se trouve aussi au cœur du cadre de la statistique bayésienne / de la pensée bayésienne des probabilités.

Analysons un exemple simple.

Exemple 2.4. *Considérez la situation avec trois pièces : l'une a pile sur les deux faces, l'une a face sur les deux faces, et l'une est équilibrée. Quelqu'un choisit, selon une certaine procédure, l'un des trois types de pièces, vous dit qu'elle a lancé la pièce et que le résultat a été pile. Quelle pièce a-t-elle lancée ?*

L'espace de probabilité pertinent qui contient les trois pièces et trois lancers est le suivant. L'espace d'états est le produit $\{C_h, C_t, C_f\} \times \{H, T\}$ *— la première coordonnée décrit le type de pièce, la seconde le résultat du lancer. Comme* σ*-algèbre, nous prenons l'ensemble des parties (on peut demander quelle face est sortie et quel était le type de pièce).*

Nous savons que, pour un ensemble fini avec l'ensemble des parties, il suffit de définir $\mathbb{P}$ *sur chaque singleton de l'espace d'états. D'après les hypothèses,* $\mathbb{P}(\{C_h, T\}) = \mathbb{P}(\{C_t, H\}) = 0$ *et* $\mathbb{P}(\{C_f, T\}) = \mathbb{P}(\{C_f, H\})$*. Si l'on pose* $p_f = \mathbb{P}(\{coin = C_f\})$*,* $p_h = \mathbb{P}(\{coin = C_h\})$*,* $p_t = \mathbb{P}(\{coin = C_t\})$*, on doit aussi avoir* $p_f + p_t + p_h = 1$*, ce qui laisse deux paramètres libres.*

Calculons maintenant les probabilités pertinentes. Clairement,

$$\mathbb{P}(\{coin = C_t\} \,|\, \{toss = H\}) = 0,$$

car la pièce « deux faces » (deux fois T) ne peut pas produire H. Pour les autres cas, on utilise Bayes :

$$\mathbb{P}(\{coin = C_h\} \,|\, \{toss = H\}) = \frac{\mathbb{P}(\{toss = H\} \,|\, \{coin = C_h\})\, \mathbb{P}(\{coin = C_h\})}{\mathbb{P}(\{toss = H\})} = \frac{\mathbb{P}(\{coin = C_h\})}{\mathbb{P}(\{toss = H\})}$$

et

$$\mathbb{P}(\{coin = C_f\} \,|\, \{toss = H\}) = \frac{\mathbb{P}(\{toss = H\} \,|\, \{coin = C_f\})\, \mathbb{P}(\{coin = C_f\})}{\mathbb{P}(\{toss = H\})} = \frac{\mathbb{P}(\{coin = C_f\})}{2\,\mathbb{P}(\{toss = H\})}.$$

Ainsi,

$$\frac{\mathbb{P}(\{coin = C_h\} \,|\, \{toss = H\})}{\mathbb{P}(\{coin = C_f\} \,|\, \{toss = H\})} = \frac{2\,\mathbb{P}(\{coin = C_h\})}{\mathbb{P}(\{coin = C_f\})} = 2p_h/p_f,$$

et, comme

$$\mathbb{P}(\{coin = C_h\} \,|\, \{toss = H\}) + \mathbb{P}(\{coin = C_f\} \,|\, \{toss = H\}) = 1,$$

on conclut

$$\mathbb{P}(\{coin = C_f\} \,|\, \{toss = H\}) = \frac{p_f}{p_f + 2p_h} \quad \textit{et} \quad \mathbb{P}(\{coin = C_h\} \,|\, \{toss = H\}) = \frac{2p_h}{p_f + 2p_h}.$$

Que peut-on conclure ? D'abord, sans connaissance des probabilités a priori *de chaque pièce, on ne peut pas dire grand-chose du résultat final, puisqu'il les contient ! Ce que nous supposons sur la probabilité initiale de chaque pièce compte beaucoup : si l'on estime que la pièce « deux piles » est très peu probable par rapport à la pièce équilibrée, disons* $p_h = 0{,}000001\, p_f$*, alors après avoir observé « pile », notre estimation donne* $\mathbb{P}(\{coin = C_f\} \,|\, \{toss = H\}) \approx 0{,}999999$*. En revanche, si nous n'avons aucune raison de croire qu'une pièce est plus probable qu'une autre (par exemple si la personne a choisi au hasard parmi les trois), alors* $p_f = p_h = p_t = 1/3$ *et la formule donne* $\mathbb{P}(\{coin = C_f\} \,|\, \{toss = H\}) = 1/3$ *et* $\mathbb{P}(\{coin = C_h\} \,|\, \{toss = H\}) = 2/3$*.*

Cependant, un point important est que, indépendamment des probabilités initiales, on peut dire comment les probabilités — ou plutôt le rapport *des probabilités — a changé : notre*

estimation que la pièce « deux piles » a été choisie augmente d'un facteur 2 par rapport à la pièce équilibrée. Et, comme vous le verrez sur la feuille d'exercices, si nous observions davantage de lancers, nous deviendrions de plus en plus confiants quant au type de pièce, indépendamment d'une estimation initiale éventuellement mauvaise. C'est aussi l'idée derrière l'approche bayésienne : nous ne connaissons pas forcément tous les paramètres au départ, mais nous pouvons les remplir par des hypothèses puis, à mesure que nous observons le monde, améliorer a posteriori ces hypothèses et raffiner nos modèles.

2.1.2 Formule des probabilités totales

Bien que les probabilités conditionnelles soient souvent délicates, elles sont indispensables — et utiles. Par exemple, elles permettent de décomposer l'espace de probabilité. En effet, le résultat suivant généralise l'idée intuitive : si l'on sait qu'exactement un des trois événements E_1, E_2, E_3 se produit toujours, alors, pour comprendre la probabilité de n'importe quel autre événement F, il suffit de connaître les probabilités conditionnelles $\mathbb{P}(F|E_i)$.

Proposition 2.5 (Formule des probabilités totales)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Soit I dénombrable et $(E_i)_{i\in I}$ une famille d'événements disjoints de probabilité strictement positive telle que $\Omega = \bigcup_{i\in I} E_i$. Alors, pour tout $F \in \mathcal{F}$,*

$$\mathbb{P}(F) = \sum_{i\in I} \mathbb{P}(F|E_i)\mathbb{P}(E_i).$$

Démonstration. Comme $\Omega = \bigcup_{i\in I} E_i$, on a $\mathbb{P}(F) = \mathbb{P}\Big(F \cap \big(\bigcup_{i\in I} E_i\big)\Big)$.

Or $F \cap (\bigcup_{i\in I} E_i) = \bigcup_{i\in I}(F \cap E_i)$. Comme les $(E_i)_{i\in I}$ sont disjoints, les $(F \cap E_i)_{i\in I}$ le sont aussi. Par additivité dénombrable,

$$\mathbb{P}(F) = \mathbb{P}\Big(\bigcup_{i\in I}(F \cap E_i)\Big) = \sum_{i\in I} \mathbb{P}(F \cap E_i).$$

Par définition, $\mathbb{P}(F \cap E_i) = \mathbb{P}(F|E_i)\mathbb{P}(E_i)$, d'où la formule. □

Remarque 2.6. *Presque la même preuve fonctionne si les E_i ne recouvrent pas tout l'espace, mais seulement à probabilité 1, i.e. si $\mathbb{P}\big(\Omega \setminus (\bigcup_i E_i)\big) = 0$. Cette généralisation est laissée en exercice.*

2.2 Indépendance d'événements

Les probabilités conditionnelles ne sont bien sûr difficiles pas du tout lorsque la probabilité d'un événement ne change pas sous conditionnement — i.e. lorsque $\mathbb{P}(E|F) = \mathbb{P}(E)$. De telles paires d'événements sont dites indépendantes. En fait, la définition rigoureuse est légèrement différente :

Définition 2.7 (Indépendance de deux événements)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. On dit que deux événements E, F sont indépendants si $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$.*

Observez que si $\mathbb{P}(F) > 0$, alors on retrouve l'énoncé intuitif d'indépendance : $\mathbb{P}(E|F) = \mathbb{P}(E)$. En effet, si E et F sont indépendants,

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \frac{\mathbb{P}(E)\mathbb{P}(F)}{\mathbb{P}(F)} = \mathbb{P}(E).$$

Nous avons choisi cette définition pour inclure automatiquement le cas $\mathbb{P}(F) = 0$.

Exemple 2.8. *Considérons le modèle d'un nombre uniforme parmi* $\{1, 2, \ldots, 12\}$ *et les événements* $E_1 := \{$*le nombre vaut* $1\}$, $E_2 := \{$*le nombre est divisible par* $2\}$, $E_3 := \{$*le nombre est divisible pa*
Lesquels sont indépendants ?

Calcul direct : $\mathbb{P}(E_1) = 1/12$, $\mathbb{P}(E_2) = 1/2$ *et* $\mathbb{P}(E_3) = 1/3$. *D'autre part,* $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1 \cap E_3) = 0$, *et* $\mathbb{P}(E_2 \cap E_3) = \mathbb{P}(\{$*divisible par* $6\}) = 1/6$. *On conclut que* E_2 *et* E_3 *sont indépendants, mais ni* E_1 *et* E_2, *ni* E_1 *et* E_3 *ne le sont.*

Déjà dans cet exemple, nous avions trois événements, et l'on peut se demander s'il existe une notion d'indépendance *conjointe* se généralisant à plusieurs événements. Il y en a en fait deux :

— l'indépendance mutuelle (ou conjointe) ;

— l'indépendance par paires.

La notion la plus forte et la plus importante est l'indépendance mutuelle.

Définition 2.9 (Indépendance mutuelle)**.** *Soit* $(\Omega, \mathcal{F}, \mathbb{P})$ *un espace de probabilité et* I *un ensemble d'indices. Les événements* $(E_i)_{i\in I}$ *sont dits mutuellement indépendants si, pour tout sous-ensemble fini* $I_1 \subseteq I$,

$$\mathbb{P}\Big(\bigcap_{i\in I_1} E_i\Big) = \prod_{i\in I_1} \mathbb{P}(E_i).$$

Parfois, on ne dispose pas de l'indépendance mutuelle (ou on ne sait pas qu'elle a lieu), mais on peut affirmer l'indépendance *par paires*. Il existe des notions analogues de k-indépendance.

Définition 2.10 (Indépendance par paires)**.** *Soit* $(\Omega, \mathcal{F}, \mathbb{P})$ *un espace de probabilité et* I *un ensemble d'indices. Les événements* $(E_i)_{i\in I}$ *sont dits indépendants par paires si, pour tout* $i \neq j \in I$, *les événements* E_i *et* E_j *sont indépendants.*

Il est important de noter que, bien que l'indépendance mutuelle implique l'indépendance par paires, la réciproque est fausse en général :

Exercice 2.2 (Indépendants par paires mais pas mutuellement)**.** *Considérez l'espace de probabilité de deux lancers de pièce indépendants. Soit* E_1 *l'événement « la première pièce est pile »,* E_2 *l'événement « la seconde pièce est pile » et* E_3 *l'événement « les deux pièces montrent la même face ». Montrer que* E_1, E_2, E_3 *sont indépendants par paires mais pas mutuellement indépendants.*

Enfin, on peut aussi parler d'indépendance de collections d'événements. Cela sera important pour généraliser l'indépendance d'événements à celle de variables aléatoires.

Définition 2.11 (Indépendance mutuelle de collections d'événements)**.** *Considérons deux collections d'événements* $(E_i)_{i\in I}$ *et* $(F_j)_{j\in J}$, *tous définis sur le même espace de probabilité. On dit qu'elles sont indépendantes si, pour tous* $i \in I, j \in J$ *:*

$$\mathbb{P}(E_i \cap F_j) = \mathbb{P}(E_i)\mathbb{P}(F_j).$$

Dans le cas de plusieurs collections $(E_{j,i})_{i\in I_j}$ *pour* $j = 1, 2, \ldots$, *on dit qu'elles sont mutuellement indépendantes si, pour tout sous-ensemble fini* $J_1 \subseteq J$ *et tout choix d'événements* E_{j,i_j} *avec* $j \in J_1$, *on a*

$$\mathbb{P}\Big(\bigcap_{j\in J_1} E_{j,i_j}\Big) = \prod_{j\in J_1} \mathbb{P}(E_{j,i_j}).$$

Autrement dit, on exige que toute sous-famille d'événements, en prenant au plus un événement dans chaque collection, soit mutuellement indépendante.

Avant de passer à l'indépendance des variables aléatoires, voici quelques propriétés de base de l'indépendance pour les événements :

Lemme 2.12 (Propriétés de base)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité.*
- *Si E est un événement avec $\mathbb{P}(E) = 1$, alors il est indépendant de tout autre événement.*
- *Si E, F sont indépendants, alors E^c et F le sont aussi. En particulier, tout événement de probabilité 0 est indépendant de tous les autres.*
- *Enfin, si un événement est indépendant de lui-même, alors $\mathbb{P}(E) \in \{0, 1\}$.*

Démonstration. Voir la feuille d'exercices. □

2.3 Indépendance de variables aléatoires

Nous formalisons maintenant la notion d'indépendance pour des quantités aléatoires, c'est-à-dire des variables aléatoires. Rappelez-vous que (la loi de) X est caractérisée par tous les événements $\{X \in (a, b)\}$. L'indépendance mutuelle de variables aléatoires est alors définie comme l'indépendance mutuelle de ces familles d'événements. Plus précisément,

Définition 2.13 (Variables aléatoires mutuellement indépendantes)**.** *Soit I un ensemble d'indices et $(X_i)_{i\in I}$ une famille de variables aléatoires définies sur le même espace $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que ces variables sont mutuellement indépendantes si, pour tout $J \subseteq I$ fini et toute collection d'intervalles $((a_j, b_j))_{j\in J}$, on a*

$$\mathbb{P}\Big(\bigcap_{j\in J}\{X_j \in (a_j, b_j)\}\Big) = \prod_{j\in J}\mathbb{P}\big(X_j \in (a_j, b_j)\big).$$

Remarque 2.14. *La définition plus « structurelle » utiliserait plutôt tous les boréliens $E_j \in \mathcal{F}_B$. C'est peu pratique, et il se trouve (par une théorie de la mesure non triviale) que c'est équivalent à la condition ci-dessus.*

Il existe naturellement d'autres conditions équivalentes. Par exemple, une condition utile (que nous verrons plus tard) est la suivante :

Exercice 2.3. *Soient $X_1, X_2, \ldots$ des variables aléatoires définies sur le même espace $(\Omega, \mathcal{F}, \mathbb{P})$. Alors $X_1, X_2, \ldots$ sont mutuellement indépendantes si et seulement si, pour tout $m \geq 2$ et tous $a_j \in \mathbb{R}$,*

$$\mathbb{P}\Big(\bigcap_{1\leq j\leq m}\{X_j \leq a_j\}\Big) = \prod_{1\leq j\leq m}\mathbb{P}(X_j \leq a_j).$$

Il y a aussi un critère particulièrement simple dans le cas discret.

Lemme 2.15 (Indépendance sur un espace discret)**.** *Soient $X_1, \ldots, X_n$ définies sur un espace de probabilité discret. Alors $X_1, \ldots, X_n$ sont mutuellement indépendantes si et seulement si, pour tous $s_1, \ldots, s_n \in \mathbb{R}$,*

$$\mathbb{P}\Big(\bigcap_{i=1}^{n}\{X_i = s_i\}\Big) = \prod_{i=1}^{n}\mathbb{P}(X_i = s_i).$$

Cela reste vrai plus généralement si $X_1, \dots, X_n$ sont définies sur n'importe quel espace de probabilité mais ne prennent qu'un nombre dénombrable de valeurs avec probabilité 1, c'est-à-dire s'il existe, pour chacune, un ensemble dénombrable S_i tel que $\mathbb{P}(X_i \in S_i) = 1$. [6]

Démonstration. Exercice. □

Comme vérification élémentaire, on voit alors facilement que, pour les espaces discrets (et en fait en général !), les indicatrices $1_E, 1_F$ de deux événements sont indépendantes si et seulement si E et F sont indépendants comme événements : en effet $\mathbb{P}(\{1_E = x\} \cap \{1_F = y\})$ vaut

$$1_{x=1}1_{y=1}\mathbb{P}(E)\mathbb{P}(F) + 1_{x=1}1_{y=0}\mathbb{P}(E)\mathbb{P}(F^c) + 1_{x=0}1_{y=1}\mathbb{P}(E^c)\mathbb{P}(F) + 1_{x=0}1_{y=0}\mathbb{P}(E^c)\mathbb{P}(F^c),$$

ce qui se réécrit

$$(1_{x=1}\mathbb{P}(E) + 1_{x=0}\mathbb{P}(E^c))(1_{y=1}\mathbb{P}(F) + 1_{y=0}\mathbb{P}(F^c)) = \mathbb{P}(\{1_E = x\})\mathbb{P}(\{1_F = y\}).$$

Exercice 2.4 (Marche aléatoire simple)**.** *Montrer que, pour une marche aléatoire simple de longueur n, tous les incréments $\Delta_i = S_i - S_{i-1}$ pour $i = 1, \dots, n$ sont mutuellement indépendants.*

La notion de variables aléatoires indépendantes est très importante et largement utilisée — souvent aussi parce que, sinon, il est très difficile de mener des calculs !

Remarque 2.16 (v.a. i.i.d.)**.** *On parle souvent d'une famille $(X_j)_{j\in J}$* i.i.d. *: cela signifie qu'elles sont mutuellement indépendantes (le premier « i ») et* identiquement distribuées *(le « i.d. »). Intuitivement, cela correspond à répéter exactement la même expérience aléatoire, encore et encore.*

Nous avons commencé le cours en construisant des espaces de probabilité, puis en y définissant des variables aléatoires. Mais il existe des cas naturels où l'on souhaite procéder dans l'autre sens : d'après l'observation ou l'expérience, nous voulons étudier un ensemble de variables aléatoires indépendantes — comment construire un espace de probabilité sur lequel elles « vivent » ? Cela peut sembler un peu trivial, mais mathématiquement la question n'est pas si facile ! Nous l'aborderons en partie dans la sous-section suivante.

2.4 Indépendance et produits d'espaces de probabilité

Bien que l'indépendance soit une notion probabiliste, elle est liée à une structure des espaces mesurés : le *produit*.

Exemple 2.17 (L'espace de n lancers de pièce équilibrée)**.** *Nous avons vu qu'on peut modéliser l'espace de n lancers de pièce équilibrée en prenant comme espace d'états Ω l'ensemble des n-uplets $\{x_1, \dots, x_n\}$ avec $x_i \in \{H, T\}$, puis $\mathcal{F}$ comme l'ensemble des parties, et enfin la probabilité de chaque singleton (chaque n-uplet) égale à 2^{-n}.*

Regardons cela ainsi :

- *Chaque n-uplet est un élément de l'espace produit $\{H, T\} \times \cdots \times \{H, T\}$, si bien qu'on peut prendre Ω comme espace produit. Notons $\Omega_0 = \{H, T\}$ l'espace d'un seul lancer.*
- *L'ensemble des parties de Ω est aussi la plus petite σ-algèbre contenant tous les ensembles de la forme $E_1 \times \cdots \times E_n$ avec E_i partie de $\{H, T\}$.*

6. De telles variables sont appelées *discrètes*, comme nous le verrons bientôt.

— *La probabilité uniforme sur Ω satisfait, par définition,*

$$\mathbb{P}(E_1 \times \cdots \times E_n) = \mathbb{P}_0(E_1) \cdots \mathbb{P}_0(E_n),$$

où $\mathbb{P}_0$ est la probabilité uniforme sur un seul lancer.

— *Enfin, le fait que les lancers soient indépendants revient à dire : pour tout i, les événements « la i-ème coordonnée est dans E_i », c.-à-d. $F_i = \Omega_0 \times \cdots \times E_i \times \cdots \times \Omega_0$, sont mutuellement indépendants. Le calcul est celui attendu :*

$$\mathbb{P}(F_i \cap F_j) = \mathbb{P}(\Omega_0 \times \cdots \times E_i \times \cdots \times E_j \times \cdots \times \Omega_0) = \mathbb{P}_0(E_i)\mathbb{P}_0(E_j) = \mathbb{P}(F_i)\mathbb{P}(F_j).$$

On voit donc que la structure de produit va de pair avec l'indépendance. En effet, c'est la règle générale : l'indépendance mutuelle de variables aléatoires est naturellement liée aux produits d'espaces de probabilité.

Poursuivons cela de façon mathématique, en parlant d'abord des espaces produits en général, puis de la construction d'espaces de probabilité pour des variables indépendantes.

2.4.1 Construction d'espaces produits

Considérons des espaces de probabilité $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ pour $i = 1, 2, \ldots$. Pour construire l'espace produit, on a besoin d'une σ-algèbre produit et d'une mesure produit.

(1) La σ-algèbre produit $\mathcal{F}_\Pi$: c'est la plus petite σ-algèbre contenant tous les ensembles $E_{i_1} \times \cdots \times E_{i_n}$ avec $E_{i_j} \in \mathcal{F}_{i_j}$ pour $j = 1, \ldots, n$, et $\{i_j\}_{j=1}^n$ un sous-ensemble fini de $\mathbb{N}$. On souligne qu'elle n'est pas égale à l'ensemble de *tous* les produits « rectangulaires » $E_{i_1} \times \cdots \times E_{i_n}$, même sur un produit fini. [7]

(2) La mesure produit $\mathbb{P}_\Pi$ de $\mathbb{P}_1, \mathbb{P}_2, \ldots$ sur $(\Pi_{i \geq 1}\Omega_i, \mathcal{F}_\Pi)$: elle est (intuitivement) la seule probabilité telle que

$$\mathbb{P}(E_{i_1} \times \cdots \times E_{i_n}) = \prod_{j=1}^{n} \mathbb{P}_{i_j}(E_{i_j})$$

pour tous les rectangles mesurables. La construction et l'unicité, même pour des produits finis, sont techniques en général et hors du cadre de ce cours.

Nous énonçons donc le théorème suivant sans preuve (voir mesure/proba avancée) :

Théorème 2.18 (Mesure produit // admis)**.** *Pour $i \in \mathbb{N}$, soit $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ des espaces de probabilité. Il existe une unique probabilité $\mathbb{P}_\Pi$ sur $(\Pi_{i\in\mathbb{N}}\Omega_i, \mathcal{F}_\Pi)$ telle que, pour tout sous-ensemble fini $J \subset \mathbb{N}$ et tout événement E de la forme $E = \Pi_{i\in\mathbb{N}}F_i$ avec $F_i = \Omega_i$ pour $i \notin J$ et $F_i = E_i \in \mathcal{F}_i$ pour $i \in J$, on ait*

$$\mathbb{P}_\Pi(E) = \prod_{i \in J} \mathbb{P}_i(E_i). \tag{2.1}$$

On appelle une telle mesure la mesure produit *de $((\Omega_i, \mathcal{F}_i, \mathbb{P}_i))_{i\geq 1}$.*

Remarque 2.19. *Nous l'utiliserons surtout lorsque $(\Omega_i, \mathcal{F}_i)$ sont $(\mathbb{R}, \mathcal{F}_B)$. Pour de tels produits finis, la preuve apparaît en Analyse IV et peut aussi se déduire de l'existence de la mesure de Lebesgue sur $([0,1], \mathcal{F}_B)$ (partie non exigible, feuille d'exercices).*

7. Même dans l'exemple ci-dessus avec 2 lancers, on peut vérifier que $\{(H,H),(T,T)\}$ n'est pas un « pavé ». Un phénomène analogue se produit pour la topologie produit.

Dans le cas discret et fini, l'existence et l'unicité sont faciles ; formulons-le ainsi :

Lemme 2.20 (Espaces produits discrets)**.** *Soient $(\Omega_i, \mathcal{P}(\Omega_i), \mathbb{P}_i)$, $i = 1, \dots, n$, des espaces de probabilité discrets. Alors la mesure produit $\mathbb{P}_\Pi$ sur $(\Pi_{i=1}^n \Omega_i, \mathcal{F}_\Pi)$ existe et est unique.*

Démonstration. Sur la feuille d'exercices. □

2.4.2 Espaces de probabilité pour variables indépendantes

Poursuivons l'idée annoncée :

— Si l'on connaît les lois de variables aléatoires et que l'on souhaite construire un espace commun sur lequel elles soient définies et *mutuellement indépendantes*, alors on utilisera des *produits*.

Nous donnons l'énoncé dans une généralité un peu plus large que la preuve.

Théorème 2.21 (Existence d'un espace portant des v.a. indépendantes // partiellement admis)**.** *Considérons des variables aléatoires $(X_i)_{i\geq 1}$. On peut trouver un espace commun $(\Omega, \mathcal{F}, \mathbb{P})$ et des variables $(\widetilde{X}_i)_{i\geq 1}$ définies dessus telles que*

— *pour tout $i \geq 1$, $\widetilde{X}_i$ a la même loi que X_i ;*

— *les variables $(\widetilde{X}_i)_{i\geq 1}$ sont mutuellement indépendantes.*

Exemple 2.22. *Supposez que vous ayez une pièce biaisée qui donne pile avec probabilité $p \in (0, 1)$. Comment modéliser une suite de n lancers indépendants ?*

L'hypothèse « toutes les suites ont la même probabilité » n'a plus de sens (par exemple quand p est proche de 1, la suite « tout pile » et « tout face » ne peuvent pas avoir la même probabilité). En revanche, l'hypothèse d'indépendance mutuelle et le lien avec les mesures produit sont utiles.

On définit l'espace comme suit :

— *on prend le produit de n copies de $(\{0, 1\}, \mathcal{P}(\{0, 1\}), \mathbb{P}_p)$, où $\mathbb{P}_p(\{1\}) = p$, $\mathbb{P}_p(\{0\}) = 1 - p$.*

Dans cet espace, la probabilité d'une suite fixée avec m piles et $n - m$ faces est exactement $p^m(1-p)^{n-m}$. Si l'on veut la probabilité d'avoir exactement m piles (positions quelconques), il faut sommer sur toutes les suites à m piles, ce qui donne $\binom{n}{m} p^m(1-p)^{n-m}$. Vérifiez que $\sum_{m=0}^n \binom{n}{m} p^m(1-p)^{n-m} = 1$!

Esquisse de preuve du Thm. 2.21 dans le cas discret et fini. Voir la feuille d'exercices pour les détails : on prend les espaces porteurs individuels et on forme le produit (Lemme 2.20), puis l'on définit $\widetilde{X}_i$ par projection : $\widetilde{X}_i(\omega_1, \dots, \omega_n) = X_i(\omega_i)$. On vérifie alors l'égalité en loi et l'indépendance par la propriété caractéristique de la mesure produit (2.1). □

Terminons cette section par un exemple important.

2.4.3 Graphe aléatoire d'Erdös–Rényi

Nous souhaitons décrire et étudier des graphes aléatoires. Les graphes sont des structures mathématiques simples qui aident à décrire des réseaux : réseaux sociaux, logistiques, ou le réseau de neurones dans le cerveau.

Définition 2.23 (Graphe simple)**.** *Soit $n \in \mathbb{N}$. Un graphe simple est une paire $G = (V, E)$ où $V = \{v_1, \dots, v_n\}$ est l'ensemble des sommets, et E est un sous-ensemble de $\{\{v_i, v_j\}$:*

$(v_i, v_j) \in V \times V, i \neq j\}$, *c'est-à-dire un ensemble de paires non ordonnées de sommets distincts, appelées arêtes.*

On peut représenter le graphe en plaçant les n sommets $v_1, \dots, v_n$ dans le plan et en traçant un segment entre v_i et v_j si et seulement si $\{v_i, v_j\} \in E$.

Si les réseaux sont très grands (cerveau, Facebook), il est impossible (et serait même peu pratique !) de les décrire en détail. De plus, ils ressemblent souvent à certains réseaux *aléatoires*. Pour comprendre leurs propriétés, on étudie des modèles simplifiés de réseaux aléatoires.

Le modèle le plus simple est le graphe d'Erdös–Rényi, où l'on inclut chaque arête avec probabilité $p > 0$.

Exemple 2.24 (Graphe d'Erdös–Rényi)**.** *Pour $n \in \mathbb{N}$, considérez un ensemble de sommets V de taille n, et E l'ensemble de toutes les arêtes non orientées possibles entre ces sommets.*

Le graphe aléatoire $G_{n,p}$ de taille n et paramètre d'arête $p \in [0,1]$ est défini en incluant chaque arête indépendamment *avec probabilité p.*

Pour définir l'espace de probabilité, on pose :

— *L'espace d'états contient tous les graphes possibles sur V. On peut l'encoder par les configurations d'arêtes : $\Omega = \{0,1\}^E$ (on interprète 1 comme « arête présente »).*

— *On suppose qu'on peut vérifier pour chaque arête si elle est présente ou non ; on prend donc $\mathcal{F} = \mathcal{P}(\Omega)$.*

— *Enfin, on place chaque arête indépendamment avec probabilité p. Pour $\omega \in \Omega$, on pose*

$$\mathbb{P}_p(\{\omega\}) := p^{|\omega|}(1-p)^{|E|-|\omega|},$$

où $|\omega|$ est le nombre d'arêtes présentes dans la configuration ω.

On identifie ω au graphe $G_{n,p}(\omega) = (V, E(\omega))$.

Quelles questions se poser ? Grossièrement : décrire l'aspect du graphe quand n est très grand ($n \to \infty$). On peut aussi s'intéresser à n petit, mais alors on peut tout expliciter.

Pour décrire l'aspect du graphe, on peut demander :

(1) Combien d'arêtes sont présentes ?

(2) Le graphe est-il connecté (pour tout v, w, existe-t-il un chemin $v = e_0 - e_1 - \cdots - e_k = w$) ?

(3) Si oui, quelle est la distance maximale entre deux sommets ?

(4) Si non, combien de composantes connectés y a-t-il ?

(5) Quelle est la plus grande composante connecté ?

(6) ...

Chacune de ces questions porte sur un graphe, i.e. une configuration ω. Dans le modèle, elles correspondent à un événement ou à une variable aléatoire, dont on peut étudier la probabilité ou la loi.

Par exemple, $N_E(\omega) := |\omega|$ (nombre d'arêtes) répond à (1). L'événement $F := \{\omega : \omega \text{ est connecté}\}$ répond à (2). Il y a des questions plus complexes lorsque l'on combine plusieurs questions.

On s'intéresse au comportement lorsque p est fixé et $n \to \infty$, mais aussi à la manière dont ce comportement change avec p. A priori, p peut dépendre de n : on peut considérer une suite $G_{n,p(n)}$.

L'étude de $G_{n,p}$ est un domaine très actif (des centaines / milliers d'articles). Nous n'en donnerons qu'un très bref aperçu.

Concentrons-nous sur la connexité et regardons des scénarios. Notez que si $p = 1$, le graphe est connecté p.s., et s'il $p = 0$, il n'est pas connecté p.s. Que se passe-t-il pour $p_n \in (0, 1)$ dépendant de n ?

Assertion 2.25. *Soit $p \in (0, 1)$ fixé. Lorsque $n \to \infty$, la probabilité que le graphe soit connecté converge vers* 1.

Ce n'est pas si surprenant : avec p fixé, on aura beaucoup d'arêtes — on s'attend à environ $p\, n(n-1)/2$ arêtes !

Démonstration. On montre que $\mathbb{P}_p(\{G_{n,p} \text{ non connecté}\}) \to 0$. D'abord,

$$\{G_{n,p} \text{ non connecté}\} = \bigcup_{v \neq w \in V} \{\ v, w \text{ non reliés par un chemin}\}.$$

Par l'union bound,

$$\mathbb{P}_p(\{G_{n,p} \text{ non connecté}\}) \leq \frac{1}{2} \sum_{v \neq w \in V} \mathbb{P}_p(\{\ v, w \text{ non reliés par un chemin}\}),$$

le facteur $1/2$ évitant le double comptage. Par symétrie, chaque paire (v, w) a la même probabilité ; le membre de droite vaut donc $n(n-1)/2 \cdot \mathbb{P}_p(\{\ v, w \text{ non reliés}\})$.

Bornons $\mathbb{P}_p(\{\ v, w \text{ non reliés}\})$. Regarder seulement l'arête $\{v, w\}$ ne suffit pas (elle est absente avec probabilité $1 - p$ qui ne tend pas vers 0). Mais il existe beaucoup d'autres chemins.

Considérons les chemins de longueur 2 via un sommet z : si v et w ne sont pas reliés, alors, pour tout z, on n'a pas simultanément $\{v, z\}$ et $\{z, w\}$ présentes. Donc

$$\begin{aligned}\mathbb{P}_p(\{\ v, w \text{ non reliés}\}) &\leq \prod_{z \in V \setminus \{v,w\}} \mathbb{P}_p\big(\{\{v, z\} \notin E\} \cup \{\{z, w\} \notin E\}\big) \\ &= \prod_z \big(1 - \mathbb{P}_p(\{v, z\} \in E,\ \{z, w\} \in E)\big) = (1 - p^2)^{n-2}.\end{aligned}$$

Cela tend clairement vers 0 quand $n \to \infty$. Donc deux sommets fixés sont reliés avec proba $\to 1$.

Revenons à la probabilité globale :

$$\mathbb{P}_p(\{G_{n,p} \text{ non connecté}\}) \leq \frac{n(n-1)}{2}\,(1 - p^2)^{n-2} \xrightarrow[n\to\infty]{} 0.$$

□

En regardant la preuve, on voit que l'énoncé reste vrai tant que $p = p(n)$ décroît « assez lentement ». La même preuve donne :

Assertion 2.26. *Soit $(p_n)_{n \geq 1}$ telle que $p_n \geq n^{-1/4}$. Alors, lorsque $n \to \infty$, la probabilité que le graphe soit connecté converge vers* 1.

Démonstration. Comme ci-dessus, $\frac{n(n-1)}{2}(1 - p_n^2)^{n-2} \to 0$ si $p_n \geq n^{-1/4}$. □

D'un autre côté :

Assertion 2.27. *Soit $(p_n)_{n\geq 1}$ telle que $p_n \leq n^{-2}$. Alors, lorsque $n \to \infty$, la probabilité que le graphe soit connecté converge vers* 0.

(Ceci figure sur la feuille d'exercices.) Notez le phénomène de *seuil*. Si p_n décroit très vite, la probabilité de connexité tend vers 0 ; s'il décroit lentement, elle tend vers 1. Pourquoi ne tend-elle pas vers une valeur entre 0 et 1 ? Où est le seuil exact ? Un théorème non trivial dit que le seuil est précisément $p_n = \frac{\log n}{n}$!

Section 3

Variables aléatoires et vecteurs aléatoires

Dans ce chapitre, nous regardons de plus près les variables aléatoires et les n-uplets de variables aléatoires, appelés *vecteurs aléatoires*.

3.1 La fonction de répartition d'une variable aléatoire

Rappelons que nous disons que deux variables sont égales en loi lorsque les probabilités qu'elles induisent sur $(\mathbb{R}, \mathcal{F}_B)$ sont égales — ceci permet de comparer des variables définies sur des espaces différents et issues de contextes différents.

Notre premier objectif est de voir comment classifier et comparer les variables plus facilement. Jusqu'ici, la loi d'une variable est décrite par la probabilité de *tous* les événements, ce qui est peu maniable.

Il s'avère que toute l'information sur la loi d'une variable peut être encodée de façon unique par la *fonction de répartition*.

Définition 3.1 (Fonction (cumulative) de répartition)**.** *On appelle* fonction (cumulative) de répartition *(f.c.r., ou c.d.f. en anglais) une fonction $F : \mathbb{R} \to [0, 1]$ satisfaisant :*

(1) F est croissante ;

(2) $F(x) \to 0$ lorsque $x \to -\infty$ et $F(x) \to 1$ lorsque $x \to +\infty$;

(3) F est continue à droite : pour tout $x \in \mathbb{R}$ et toute suite $(x_n)_{n\geq 1} \subset [x, \infty)$ telle que $x_n \to x$, on a $F(x_n) \to F(x)$.

Étant donnée une variable aléatoire X, on définit sa fonction de répartition comme suit :

Proposition 3.2 (Fonction de répartition d'une v.a.)**.** *Pour toute variable X (définie sur un espace $(\Omega, \mathcal{F}, \mathbb{P})$), la fonction $F_X(x) := \mathbb{P}_X((-\infty, x])$ est une fonction de répartition.*

Démonstration. On pose $F_X(x) = \mathbb{P}(X \in (-\infty, x])$. Comme $(-\infty, x] \subseteq (-\infty, y]$ pour $x \leq y$, (1) de la Prop. 1.12 donne que F est croissante.

Vérifions la continuité à droite. Soit $(x_n)_{n\geq 1} \subset [x, \infty)$, $x_n \to x$. Posons $A_n := \cap_{1\leq k\leq n}(-\infty, x_k]$; alors $\bigcap_{n\geq 1} A_n = (-\infty, x]$. Par continuité de $\mathbb{P}$ (point (5) de la Prop. 1.12), $\mathbb{P}_X(A_n) \to \mathbb{P}_X((-\infty, x])$. Comme $x_n \to x$, pour tout n assez grand, on a $\{-\infty, x_n\} \subseteq A_{m_n}$ pour un certain $m_n \to \infty$. Il s'ensuit $F_X(x) \leq F_X(x_n) \leq \mathbb{P}_X(A_{m_n})$, d'où $F_X(x_n) \to F_X(x)$.

Les deux dernières propriétés (bornes à $\pm\infty$) figurent sur la feuille d'exercices. □

Réciproquement, toute fonction de répartition donne naissance à une unique loi de variable aléatoire.

Théorème 3.3 (Les lois sont déterminées par la f.c.r. // admis)**.** *Toute fonction de répartition F correspond à une unique loi d'une variable X telle que $F_X(x) = \mathbb{P}_X((-\infty, x])$. Autrement dit, les f.c.r. sont en bijection avec les probabilités sur $(\mathbb{R}, \mathcal{F}_B)$.*

Admettons le théorème au niveau général, mais esquissons la construction à partir de l'uniforme sur $([0, 1], \mathcal{F}_B)$ — c'est aussi la base des simulateurs :

Esquisse : construire X à partir de l'uniforme. Soit F une f.c.r. L'idée est d'utiliser l'espace $((0,1], \mathcal{F}_B, \mathbb{P}_U)$ (mesure uniforme). On cherche une application $(0,1] \to \mathbb{R}$ envoyant la v.a. uniforme U vers la loi F.

Définissez $X_F : (0,1] \to \mathbb{R}$ par

$$X_F(x) := \inf\{y \in \mathbb{R} : F(y) \geq x\}.$$

Alors X_F est croissante et (exercice) mesurable de $((0,1], \mathcal{F}_B)$ vers $(\mathbb{R}, \mathcal{F}_B)$; c'est une v.a. [8]

On calcule

$$\mathbb{P}_U(X_F \in (-\infty, x]) = \mathbb{P}_U\big((0, \sup\{z \in (0,1] : z \leq F(x)\}]\big) = \mathbb{P}_U\big((0, F(x)]\big) = F(x),$$

d'où F est bien la f.c.r. de X_F. □

Exemple 3.4. *Calculons la f.c.r. de la variable de Bernoulli X qui vaut 1 avec probabilité p et 0 avec probabilité $1-p$. Notez que les indicatrices d'événements de probabilité p correspondent à ce cas.*

On a $F_X(x) = (1-p)\,\mathbf{1}_{\{x \geq 0\}} + p\,\mathbf{1}_{\{x \geq 1\}}$. Plus généralement, si X ne prend qu'un nombre fini de valeurs $x_1, \ldots, x_n$ avec probabilités $p_1, \ldots, p_n$, alors $F_X(x) = \sum_{i=1}^n p_i\,\mathbf{1}_{\{x \geq x_i\}}$ (pourquoi ?).

On voit que F_X encode naturellement le comportement de X. Regardons de plus près le lien entre F_X et X. On note $F(x^-)$ la limite de $F(x_n)$ pour des suites $(x_n) \to x$ avec $x_n < x$.

Lemme 3.5 (F.c.r. vs v.a.)**.** *Soit X une v.a. sur $(\Omega, \mathcal{F}, \mathbb{P})$ et F_X sa f.c.r. Alors, pour tous $x < y \in \mathbb{R}$:*

(1) $\mathbb{P}(X < x) = F(x^-)$;

(2) $\mathbb{P}(X > x) = 1 - F(x)$;

(3) $\mathbb{P}(X \in (x,y)) = F(y^-) - F(x)$;

(4) $\mathbb{P}(X = x) = F(x) - F(x^-)$.

Démonstration. Voir la feuille d'exercices. □

Exemple 3.6. *Voici la f.c.r. de l'uniforme U sur $[0,1]$: $F_U(x) = x\,\mathbf{1}_{\{0 \leq x \leq 1\}} + \mathbf{1}_{\{x > 1\}}$. Par le lemme ci-dessus, pour tout intervalle $(a,b) \subset [0,1]$, $\mathbb{P}(U \in (a,b)) = b - a$.*

"'tex On voit d'après ce qui précède que tous les sauts de F_X correspondent à des points x tels que $\mathbb{P}(X = x) > 0$. En fait, il ne peut y en avoir qu'un nombre dénombrable.

Lemme 3.7. *Une fonction de répartition F_X d'une variable aléatoire X a au plus un nombre dénombrable de sauts.*

Démonstration. Soit S_n l'ensemble des sauts qui sont plus grands que $1/n$ et $\widehat{S}_n$ un sous-ensemble fini quelconque de S_n. Alors $\widehat{S}_n$ est mesurable et $1 \geq \mathbb{P}(X \in S_n) \geq |\widehat{S}_n| n^{-1}$. Il s'ensuit donc que $|\widehat{S}_n| \leq n$. Comme ceci vaut pour tout sous-ensemble fini de S_n, on en déduit que $|S_n| \leq n$ et en particulier que S_n est fini.

Maintenant, l'ensemble de tous les sauts peut s'écrire comme une union $\bigcup_{n \geq 1} S_n$. Ainsi, comme chaque S_n est fini et qu'une union dénombrable d'ensembles finis est dénombrable, on conclut. □

8. On l'appelle parfois *fonction quantile généralisée* ; c'est « l'inverse à gauche » de F.

Ces sauts d'une c.d.f. F_X sont parfois appelés *atomes* de la loi de X. Plus précisément, on appelle $s \in \mathbb{R}$ un atome pour la loi de X si et seulement si $\mathbb{P}(X = s) > 0$.

Dans le cas extrême, F_X n'augmente que par sauts, c.-à-d. est constante par morceaux en ne changeant de valeur qu'un nombre au plus dénombrable de fois. Précisément :

Définition 3.8 (Constante par morceaux avec au plus un nombre dénombrable de sauts)**.** *On dit que $f : \mathbb{R} \to [0, \infty)$ est constante par morceaux avec un nombre dénombrable de sauts s'il existe un ensemble dénombrable S et des réels $c_s > 0$ pour $s \in S$ tels que $\sum_{s \in S} c_s < \infty$ et*

$$f(x) = \sum_{s \in S} c_s 1_{x \geq s}.$$

Remarquez que cet ensemble S peut être dense, comme l'ensemble des nombres rationnels, ce qui le rend difficile à imaginer comme une fonction en escalier !

À l'autre extrême, F_X pourrait aussi être continue partout. Ces observations motivent la distinction entre variables aléatoires discrètes et continues :

Définition 3.9 (Variables aléatoires discrètes et continues)**.** *Une variable aléatoire est dite discrète si sa c.d.f. F_X est constante par morceaux, en ne changeant de valeur qu'un nombre au plus dénombrable de fois. Elle est dite continue si sa c.d.f. F_X est continue.*

Ces définitions ont l'air un peu abstraites / peu parlantes du point de vue probabiliste. L'exercice suivant les reformule d'une autre manière :

Exercice 3.1 (Discrète vs variable aléatoire, ver 2)**.** *Considérez une variable aléatoire X. Montrez que*

- *X est discrète, i.e. sa fonction de répartition F_X est constante par morceaux, si et seulement s'il existe un ensemble dénombrable $S \subseteq \mathbb{R}$ tel que $\mathbb{P}(X \in S) = \mathbb{P}_X(S) = 1$.*
- *X est continue si et seulement si pour tout $y \in \mathbb{R}$, $\mathbb{P}(X = y) = \mathbb{P}_X(\{y\}) = 0$.*

Remarquez que toute variable aléatoire n'est pas nécessairement soit discrète soit continue : il peut aussi y avoir des mélanges des deux ; par exemple on pourrait imaginer une c.d.f. donnée par $F(x) = 0.5 \cdot 1_{x \geq 0} + 0.5 \cdot x \cdot 1_{x \in [0,1)} + 0.5 \cdot 1_{x \geq 1}$ (à quoi cela correspond-il ?).

3.2 Exemples de variables aléatoires discrètes

Il existe plusieurs familles de lois de variables aléatoires discrètes qui reviennent encore et encore. Comme nous le verrons, ces lois ont parfois aussi de très belles caractérisations mathématiques.

Rappelons que pour caractériser la loi d'une variable aléatoire, on peut soit donner la valeur de $\mathbb{P}_X(F)$ sur une famille suffisamment grande de F (p.ex. tous les intervalles), soit donner la c.d.f. Pour une variable aléatoire discrète X il suffit de déterminer le support S, i.e. le plus petit ensemble $S \subseteq \mathbb{R}$ tel que $\mathbb{P}(X \in S) = 1$, et de déterminer $\mathbb{P}(X = s)$ pour chaque $s \in S$ (pourquoi ?).

Variable aléatoire de Bernoulli

Comme déjà mentionné, une variable aléatoire qui ne prend que les valeurs $\{0, 1\}$, prenant la valeur 1 avec probabilité p, est appelée variable aléatoire de Bernoulli de paramètre p. Elle porte le nom du mathématicien suisse Jacob Bernoulli, qui pensait aussi que toutes les

sciences ont besoin des mathématiques, mais que les mathématiques n'ont besoin d'aucune. Je vous laisse juger ; voyons que ces exemples apparaissent très souvent.

En effet, sur tout espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, toute fonction indicatrice d'un événement, i.e. 1_E, donne une variable aléatoire de Bernoulli et le paramètre p est égal à la probabilité de l'événement. En effet, pour tout événement E dans un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, la fonction indicatrice $1_E : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F})$ est mesurable et donc une variable aléatoire. De plus, elle est à valeurs dans $\{0, 1\}$ par définition et $\mathbb{P}(\{1_E = 1\}) = \mathbb{P}(E) = p$.

Parfois on parle de variables aléatoires de Bernoulli plus généralement dès qu'il y a deux issues différentes, p.ex. aussi lorsque les valeurs sont $\{-1, 1\}$. On parle alors d'une variable aléatoire de Bernoulli à valeurs $\{-1, 1\}$.

Variable aléatoire uniforme

Toute variable aléatoire qui prend ses valeurs dans un ensemble fini $S = \{x_1, \ldots, x_n\}$, chacune avec probabilité égale $1/n$, est appelée variable aléatoire uniforme sur S. On appelle la loi de cette variable la loi uniforme. Sa c.d.f. est donnée simplement par $F_X(x) = n^{-1} \sum_{i=1}^{n} 1_{x \geq x_i}$.

Les exemples sont : un dé équilibré, le résultat de la roulette, tirer la carte du dessus d'un paquet bien mélangé, etc... Pour être concret, un exemple trivial : si l'on modélise un dé équilibré sur $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ et $\mathbb{P}(\{i\}) = 1/6$, alors la variable aléatoire $X(\omega) := \omega \in \mathbb{R}$ donne une variable aléatoire uniforme.

On utilise cette famille de variables aléatoires chaque fois qu'on n'a aucune raison a priori de préférer une issue plutôt qu'une autre. Une manière mathématique un peu plus « chic » de le dire serait : la loi uniforme est la seule loi de probabilité sur un ensemble fini qui soit invariante par permutations de cet ensemble. Nous verrons aussi sur la feuille d'exercices que c'est la distribution de probabilité d'entropie maximale à valeurs dans un ensemble fini S.

Variable aléatoire binomiale

Une variable aléatoire qui prend ses valeurs dans l'ensemble $\{0, 1, \ldots, n\}$, et qui prend chaque valeur k avec probabilité

$$p^k (1-p)^{n-k} \binom{n}{k}$$

est appelée variable aléatoire binomiale de paramètres $n \in \mathbb{N}$ et $0 \leq p \leq 1$ (pourquoi les probabilités somment-elles à 1 ?). On note la loi d'une telle variable aléatoire binomiale par $Bin(n, p)$.

Remarquez que pour $n = 1$, on retrouve la variable aléatoire de Bernoulli. La variable de Bernoulli apparaît naturellement dans des modèles de lancers de pièces indépendants, de graphes aléatoires, ou de modèles de marches aléatoires. La raison pour laquelle elle apparaît si souvent est qu'elle décrit toujours la situation suivante : on a une suite d'événements indépendants et indiscernables, et l'on compte le nombre de ceux qui se produisent. Autrement dit, la variable binomiale $Bin(n, p)$ peut être vue comme une somme de n variables aléatoires $Ber(p)$ indépendantes.

Exercice 3.2 (La v.a. binomiale est le nombre d'événements qui se produisent)**.** *Supposons que l'on ait n événements mutuellement indépendants $E_1, \ldots, E_k$ de probabilité p sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Considérez le nombre aléatoire d'événements qui se produisent : $X = \sum_{i=1}^{n} 1_{E_i}$. Montrez que X est une variable aléatoire et a la loi $Bin(n, p)$.*

Pour un exemple concret et vivant, revenons au graphe aléatoire d'Erdös–Rényi sur n sommets, où chaque arête est incluse indépendamment avec probabilité p. On peut alors fixer un sommet v et considérer la variable aléatoire M_v donnant le nombre de sommets adjacents à v, i.e. reliés à v par une arête. L'exercice ci-dessus montre que cette variable aléatoire a la loi $Bin(n-1,p)$.

Variable aléatoire géométrique

Une variable aléatoire qui prend ses valeurs dans l'ensemble $\mathbb{N}$, chaque valeur k avec probabilité $p(1-p)^{k-1}$ pour un certain $0 < p \leq 1$, est appelée variable aléatoire géométrique de paramètre p. On note la loi d'une variable aléatoire géométrique par $Geo(p)$. On devrait encore vérifier que ceci définit bien une variable aléatoire, en voyant que les probabilités somment à 1.

Une variable aléatoire géométrique décrit la situation suivante : on a des événements indépendants $E_1, E_2, \ldots$ chacun de probabilité de succès p et l'on demande le plus petit indice k tel que l'événement E_k se produise. Par exemple, $Geo(1/2)$ décrit le nombre de lancers nécessaires pour obtenir une première fois pile. Ceci sera rendu précis sur la feuille d'exercices.

Il y a aussi une jolie propriété qui caractérise la v.a. géométrique :

Lemme 3.10 (La v.a. géométrique est la seule variable sans mémoire à valeurs dans $\mathbb{N}$)**.** *On dit qu'une variable aléatoire X à valeurs dans $\mathbb{N}$ est sans mémoire si pour tous $k, l \in \mathbb{N}$ on a $\mathbb{P}(X > k+l | X > k) = \mathbb{P}(X > l)$. Toute variable aléatoire géométrique est sans mémoire, et en fait ce sont les seuls exemples de variables aléatoires sans mémoire sur $\mathbb{N}$.*

Démonstration. Commençons par montrer que la variable aléatoire géométrique satisfait la propriété sans mémoire. Tout d'abord, remarquons que si $\mathbb{P}(X = 1) = 1$, alors X est une variable aléatoire géométrique dégénérée avec $p = 1$. On peut donc supposer que l'on travaille dans le cas $\mathbb{P}(X > 1) > 0$.

La partie disant que la v.a. géométrique est sans mémoire est sur la feuille d'exercices.

Montrons maintenant que toute variable aléatoire satisfaisant la propriété sans mémoire a la loi d'une variable aléatoire géométrique. Là encore, si $\mathbb{P}(1) = 1$, c'est terminé. Sinon, on peut écrire

$$\mathbb{P}(X > 1 + l | X > 1)\mathbb{P}(X > 1) = \mathbb{P}(X > 1 + l).$$

Comme pour une variable sans mémoire on a $\mathbb{P}(X > l) = \mathbb{P}(X > 1 + l | X > 1)$, on obtient

$$\mathbb{P}(X > l)\mathbb{P}(X > 1) = \mathbb{P}(X > l + 1).$$

Ainsi, par récurrence, $\mathbb{P}(X > l) = \mathbb{P}(X > 1)^l$ et donc X est une variable aléatoire géométrique de paramètre $p = 1 - \mathbb{P}(X > 1)$. □

Variable aléatoire de Poisson

Poisson était un mathématicien français qui a dit célèbrement que la vie n'est bonne que pour deux choses : les mathématiques et enseigner les mathématiques. Ses variables aléatoires apparaissent assez souvent.

La variable aléatoire de Poisson est une variable aléatoire discrète à valeurs dans $\{0\} \cup \mathbb{N}$ et prenant la valeur k avec probabilité

$$e^{-\lambda}\frac{\lambda^k}{k!}$$

pour un certain $\lambda > 0$. On note cette distribution par $Poi(\lambda)$. Les variables aléatoires de Poisson décrivent des occurrences d'événements rares sur une certaine période de temps, où les événements se produisant dans deux périodes de temps consécutives sont indépendants. Par exemple, cela a été utilisé pour modéliser

— le nombre de visiteurs dans un petit musée hors des sentiers battus ;
— plus largement, le nombre d'étoiles dans une unité d'espace ;
— ou, plus sombrement, le nombre de soldats tués par des coups de sabot dans l'armée prussienne.

Une façon de voir apparaître la v.a. de Poisson est comme limite d'une distribution binomiale lorsque la probabilité de succès p est de l'ordre de $1/n$:

Lemme 3.11 (Variable aléatoire de Poisson comme limite de binomiales)**.** *Considérez la distribution binomiale $Bin(n, \lambda/n)$. Montrez que lorsque $n \to \infty$ elle converge vers $Poi(\lambda)$ au sens où, pour tout $k \in \{0\} \cup \mathbb{N}$, on a*

$$\mathbb{P}(Bin(n, \lambda/n) = k) \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

Démonstration. Par définition, pour tout $n \in \mathbb{N}$ fixé et tout $k \in \{0\} \cup \mathbb{N}$, on a

$$\mathbb{P}(Bin(n, \lambda/n) = k) = \binom{n}{k}\frac{\lambda^k}{n^k}\left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

En utilisant

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n(n-1)\cdots(n-k+1)}{k!},$$

on peut écrire

$$\mathbb{P}(Bin(n, \lambda/n) = k) = \frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^{n}\frac{n(n-1)\cdots(n-k+1)}{n^k}\left(1 - \frac{\lambda}{n}\right)^{-k}.$$

Mais lorsque $n \to \infty$,

$$\left(1 - \frac{\lambda}{n}\right)^{n} \to e^{-\lambda}.$$

De plus, pour tout $t > 0$ fixé, on a aussi $\frac{n-t}{n} \to 1$ lorsque $n \to \infty$ et donc

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1$$

et

$$\left(1 - \frac{\lambda}{n}\right)^{-k} = \left(\frac{n-\lambda}{n}\right)^{-k} \to 1,$$

ce qui prouve le lemme. □

Pour relier cela aux occurrences d'événements rares décrites plus haut, on peut raisonner ainsi. Supposons que l'on veuille modéliser le nombre d'arrivées sur une fenêtre de temps $[0,1]$, disons une année dans un lieu lointain. On découpe alors la fenêtre $[0,1]$ en n segments de temps égaux de longueur $1/n$ avec n grand, par exemple en 365 jours, de sorte que l'on puisse supposer que, dans chaque segment (chaque jour), il y a au plus une arrivée. Dans ce cas, on peut décrire l'arrivée ou la non-arrivée par $Ber(p)$ ou 1_E pour un certain événement E. Si l'on suppose en plus que tous les jours se ressemblent, on peut prendre ce paramètre p identique pour tous les segments de même longueur, p.ex. pour tous les jours. De plus, si l'on suppose qu'une arrivée dans un segment de temps n'influence pas les arrivées dans d'autres intervalles, on peut supposer que tous les événements E correspondant à des intervalles de temps différents sont mutuellement indépendants. Ainsi, le nombre total d'arrivées est le nombre d'événements indépendants qui se produisent, lorsque la probabilité de l'événement est p : on a vu plus haut que cela donne une variable aléatoire $Bin(n,p)$. Mais maintenant, si vous regardez attentivement la preuve ci-dessus, vous voyez que si p n'est pas de la forme λ/n pour un certain $\lambda > 0$, alors en fait le nombre d'événements va soit vers l'infini soit vers zéro — c.-à-d. que, pour obtenir une variable aléatoire non triviale à la limite $n \to \infty$, on est forcé de prendre $p = \lambda/n$.

Les variables aléatoires de Poisson se comportent aussi très bien lorsqu'on prend des copies indépendantes et lorsqu'on en prend des sous-ensembles aléatoires :

Exercice 3.3 (Variables aléatoires de Poisson)**.** *Soient $X_1 \sim Poi(\lambda_1)$ et $X_2 \sim Poi(\lambda_2)$ deux variables aléatoires indépendantes définies sur le même espace de probabilité.*

- *Montrez alors que $X_1 + X_2$ est aussi une variable aléatoire de Poisson, de paramètre $\lambda_1 + \lambda_2$.*
- *Soient maintenant $Y_1, Y_2, \ldots$ des variables aléatoires $Ber(p)$ indépendantes définies sur le même espace de probabilité. Montrez que $X := \sum_{i=1}^{X_1} Y_i$ a aussi la loi $Poi(p\lambda)$, que $X_1 - X$ a la loi $Poi((1-p)\lambda)$, et est indépendante de X.*

3.3 Variables aléatoires continues

Rappelons que l'on a appelé une variable aléatoire X continue si F_X était continue, i.e. sans aucun saut. D'après le Lemme 3.5, on a alors $\mathbb{P}(X = x) = 0$ pour tout $x \in \mathbb{R}$. Le plus souvent, les variables aléatoires continues apparaissent via ce qu'on appelle une fonction densité, et c'est aussi ainsi que nous les construirons en général.

Définition 3.12 (V.a. continue avec densité)**.** *Soit X une variable aléatoire et $f_X : \mathbb{R} \to \mathbb{R}$ une fonction non négative intégrable telle que $\int_{\mathbb{R}} f_X(x)\,dx = 1$. On dit alors que la v.a. X a pour densité f_X si pour tout $x \in \mathbb{R}$*

$$F_X(t) = \int_{-\infty}^{t} f_X(x)\,dx.$$

9

9. Vous avez peut-être déjà entendu dire qu'il existe plusieurs notions d'intégrale. Ici, l'intégrale naturelle à utiliser serait l'intégrale de Lebesgue car elle permet d'intégrer sur tous les boréliens, ce qui, comme vous l'avez peut-être vu, n'est pas possible avec l'intégrale de Riemann. Mais en fait, pour tous les exemples ici, penser à l'intégrale de Riemann est tout à fait suffisant.

Remarque 3.13. *Nous remarquons tout de suite qu'il existe aussi des variables aléatoires continues sans densité (voir la section étoilée des exercices).*

Regardons maintenant la définition de plus près. Tout d'abord, il est important de vérifier que la définition a bien un sens, i.e. que le F_X défini est effectivement une fonction de répartition :

Exercice 3.4. *Considérez une fonction f_X non négative Riemann-intégrable telle que $\int_{\mathbb{R}} f_X(x)\,dx = 1$. Définissez $F_X(x) := \int_{-\infty}^{x} f_X(x)\,dx$.*

- *Montrez que F_X est une fonction de répartition.*
- *Montrez que si deux variables aléatoires ont la même densité, elles ont la même loi.*
- *Montrez que, étant donné F_X, il existe au plus une fonction continue f_X telle que $F_X(t) := \int_{-\infty}^{t} f_X(x)\,dx$.*
- *Donnez des exemples montrant que f_X n'est toutefois pas déterminée de manière unique par F_X.*

Ensuite, regardons une interprétation. En utilisant le Lemme 3.5 et la remarque ci-dessus disant que $\mathbb{P}(X = x) = 0$, on peut aussi écrire

$$\mathbb{P}(X \in (a,b)) = \mathbb{P}(X \in [a,b]) = \int_a^b f_X(x)\,dx.$$

Il est important de remarquer que f_X ne vous donne pas la probabilité de $\{X = x\}$ en chaque point : on a déjà vu que pour les variables aléatoires continues cette probabilité vaut 0 pour tout $x \in \mathbb{R}$. Cependant, en prenant $b = a + \epsilon$, on peut tout de même obtenir une interprétation de f_X, ce qui explique pourquoi on l'appelle la densité. En effet, si par exemple f_X est continue, on peut écrire

$$\mathbb{P}(X \in (a, a+\epsilon)) = \int_a^{a+\epsilon} f_X(x)\,dx = \epsilon f_X(a) + o(\epsilon),$$

et l'on peut donc voir $\epsilon f_X(a)$ comme la probabilité d'être dans l'intervalle $(a, a+\epsilon)$. En particulier, remarquez que $\epsilon^{-1}\mathbb{P}(X \in (a, a+\epsilon)) \to f_X(a)$ lorsque $\epsilon \to 0$. Ceci est bien sûr lié au théorème fondamental de l'analyse, qui dans le cas d'une densité continue dit que $F_X'(x) = f_X(x)$.

Regardons maintenant quelques exemples. D'après l'exercice ci-dessus, on voit que pour décrire une variable aléatoire continue avec densité, il suffit de donner la densité : une fonction intégrable, non négative, d'intégrale totale 1.

Variable aléatoire uniforme sur $[a,b]$
Une variable aléatoire U de densité $f_U(x) = \frac{1}{b-a}1_{[a,b]}$ est appelée variable aléatoire uniforme sur l'intervalle $[a,b]$ et est parfois notée $U = U_{[a,b]}$. Nous avons déjà rencontré la variable aléatoire uniforme sur $[0,1]$ — comme prévu, sa loi $\mathbb{P}_U$ est égale à la mesure uniforme / de Lebesgue sur $[0,1]$, vue comme mesure de probabilité sur $\mathbb{R}$. Sa c.d.f. est donnée par $F_U(x) = 1_{0\le x} \min\{x, 1\}$. On peut aussi la voir comme la limite de variables aléatoires uniformes discrètes prenant leurs valeurs dans $\{i/n : i = 1 \dots n\}$ — nous avons vu une manière de le rendre précis sur la Feuille d'exercices 7.

Variable aléatoire exponentielle
Soit $\lambda > 0$. La variable aléatoire X de densité $f_X(x) = \lambda e^{-\lambda x}1_{x\ge 0}$ est appelée variable

aléatoire exponentielle de paramètre λ, et sa loi est parfois notée $Exp(\lambda)$. (Nous vérifierons sur la feuille d'exercices que la masse totale vaut bien 1.) Dans ce cas, on peut voir la variable exponentielle comme une amie continue de la variable géométrique, car elle satisfait elle aussi la propriété sans mémoire :

Exercice 3.5 (La v.a. exponentielle est la seule variable sans mémoire)**.** *On dit qu'une variable aléatoire continue X satisfaisant $\mathbb{P}(X > 0) = 1$ est sans mémoire si pour tous $x, y > 0$ on a $\mathbb{P}(X > x + y | X > y) = \mathbb{P}(X > x)$. Montrez que la variable aléatoire exponentielle est sans mémoire. De plus, montrez que toute variable aléatoire continue sans mémoire a la loi d'une variable aléatoire exponentielle.*

Comme pour les variables géométriques, les variables exponentielles sont liées à des temps d'attente ; simplement, le processus sous-jacent n'est plus en temps discret (comme une suite de lancers) mais en temps continu (comme attendre le prochain appel d'un(e) ami(e)). Nous pourrons faire des énoncés plus précis plus tard dans le cours.

Variable aléatoire gaussienne

L'exemple le plus important de variable aléatoire est peut-être celui d'une variable normale ou gaussienne. Étant donnés deux paramètres $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}$, on dit que N a la loi d'une variable normale de moyenne μ et de variance σ^2, notée $N \sim \mathcal{N}(\mu, \sigma^2)$, si sa densité est donnée par

$$f_N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big).$$

On appelle la loi $\mathcal{N}(0,1)$ la loi normale centrée réduite, ou la gaussienne standard. Les lois normales apparaissent partout à cause de ce qu'on appelle le théorème central limite. Une version faible pourrait être énoncée vaguement ainsi :

— Soient $X_1, X_2, \dots$ une suite de variables aléatoires i.i.d. telles que X_i ait la même loi que $-X_i$ et, de plus, que chaque X_i soit bornée au sens où il existe $C > 0$ tel que $\mathbb{P}(X_i < C) = 1$. Posons $S_n = \sum_{i=1}^n X_i$. Alors, à la limite $n \to \infty$, $\frac{S_n}{\sqrt{n}}$ devient une variable aléatoire normale : pour tout intervalle (a, b), on a $\mathbb{P}\big(\frac{S_n}{\sqrt{n}} \in (a,b)\big) \to \mathbb{P}(N \in (a,b))$, où N est une variable aléatoire gaussienne.

Par exemple, dans des expériences de physique, on s'attend rarement à obtenir la valeur « exacte », mais plutôt une valeur avec une erreur. On suppose que cette erreur est une somme de nombreuses petites erreurs indépendantes, et ainsi, sauf s'il y a un biais qui n'a pas été pris en compte, les valeurs observées ont une distribution normale autour de la valeur réelle.

Nous prouverons une version de ce théorème vers la fin du cours, après avoir développé davantage d'outils pour travailler avec des variables aléatoires. Il y en a une première version dans la section étoilée des exercices.

Il est courant de mentionner ici que, bien que la variable normale soit la plus utilisée, sa fonction de répartition — qui a sa propre notation Φ_{μ,σ^2} — donnée comme toujours par

$$\Phi_{\mu,\sigma^2}(x) = \mathbb{P}(N \leq t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big)\, dx$$

n'admet pas de formule plus explicite. Ainsi, autrefois, il fallait vraiment consulter une longue table de valeurs pour donner une réponse numérique à, disons, $\mathbb{P}(N > 12)$ ou $\mathbb{P}(|N| < 200)$. Je soupçonne qu'il existe aujourd'hui des méthodes plus modernes...

Un autre aspect important des gaussiennes est leur relation intime avec l'algèbre linéaire : les variables aléatoires gaussiennes et les vecteurs aléatoires se comportent extrêmement bien sous transformations linéaires, ce qui les rend déjà pour cette raison centrales dans de nombreux modèles probabilistes.

Voici un lemme simple dans cet esprit, donnant aussi un sens à μ et σ^2 comme translation et mise à l'échelle :

Lemme 3.14. *Soit X_{μ,σ^2} une variable aléatoire gaussienne. Soit aussi X une gaussienne standard. Alors $\sigma X + \mu$ a la même loi que X_{μ,σ^2}.*

Démonstration. Sur la feuille d'exercices. □

3.4 Vecteurs aléatoires

Nous avons déjà vu dans les notes et sur la feuille d'exercices que, souvent, plusieurs variables aléatoires apparaissent dans une même situation probabiliste et sont naturellement définies sur le même espace de probabilité. Jusqu'ici, nous regardions surtout leurs lois individuelles, ou bien la situation où elles étaient indépendantes. Mais ce n'est pas toujours le cas. Lorsque l'on s'intéresse au comportement conjoint de plusieurs variables aléatoires, il est parfois utile de penser en termes de vecteurs aléatoires :

Définition 3.15 (Vecteurs aléatoires et lois marginales)**.** *Considérez un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que $(X_1, X_2, \ldots, X_n)$ est un vecteur aléatoire si et seulement si chacun de $X_1, X_2, \ldots, X_n$ est une variable aléatoire. La loi $\mathbb{P}_{X_i}$ de chaque v.a. X_i s'appelle sa loi marginale.*

Les lois marginales ne sont rien d'autre que les lois individuelles des variables aléatoires X_i qui apparaissent comme composantes d'un vecteur aléatoire, et dont nous avons discuté jusqu'ici. Nous savons comment les décrire. Mais elles n'encodent pas la relation entre les variables aléatoires.

Par exemple :

— D'un côté, considérez (X_1, X_2), où X_1 et X_2 codent deux lancers de pièce équilibrée indépendants.

— D'un autre côté, considérez $(X_1, \widetilde{X}_2)$, où X_1 est un lancer de pièce équilibrée, mais $\widetilde{X}_2$ vaut pile lorsque X_1 vaut face et $\widetilde{X}_2$ vaut face lorsque X_1 vaut pile.

Alors les lois marginales des vecteurs (X_1, X_2) et $(X_1, \widetilde{X}_2)$ sont les mêmes (pourquoi ?), mais ils décrivent clairement des situations très différentes !

Comment alors encoder mathématiquement cette relation entre les variables aléatoires ? En fait, pour étudier les lois conjointes, il est plus naturel de voir $(X_1, \ldots, X_n)$ non pas simplement comme un vecteur de variables aléatoires à valeurs dans $\mathbb{R}$, mais plutôt comme une variable aléatoire à valeurs dans $\mathbb{R}^n$:

Lemme 3.16 (Loi conjointe des vecteurs aléatoires)**.** *Soit $\overline{X} = (X_1, \ldots, X_n)$ un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$. Alors $(X_1, \ldots, X_n)$, vu comme vecteur, est une variable aléatoire à valeurs dans $(\mathbb{R}^n, \mathcal{F}_B)$, i.e. l'application $\omega \mapsto (X_1(\omega), \ldots, X_n(\omega))$ est mesurable de*

$(\Omega, \mathcal{F})$ *vers* $(\mathbb{R}^n, \mathcal{F}_B)$*. En particulier, un vecteur aléatoire induit une mesure de probabilité* $\mathbb{P}_{\overline{X}}$ *sur* $(\mathbb{R}^n, \mathcal{F}_B)$*, appelée la loi conjointe du vecteur* $\overline{X}$*.*

Réciproquement, toute variable aléatoire à valeurs dans $(\mathbb{R}^n, \mathcal{F}_E)$ *donne naissance à un vecteur aléatoire au sens de la définition ci-dessus.*

Nous ne prouverons pas ce lemme, mais remarquons simplement que la question sous-jacente est celle de la mesurabilité : la mesurabilité de chaque composante comme fonction $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F}_E)$ garantit-elle la mesurabilité de la fonction $(\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{F}_E)$, et réciproquement ? Cela devrait vous rappeler votre cours de topologie et les fonctions continues à valeurs vectorielles [10].

Cette mise en place permet de prouver rapidement le résultat de base suivant :

Lemme 3.17. *Soit* $\overline{X}$ *un vecteur aléatoire dans* $\mathbb{R}^n$ *et* $\overline{a}$ *un vecteur fixé quelconque de* $\mathbb{R}^n$*. Alors* $\sum_{i=1}^n a_i X_i$ *est une variable aléatoire. De même,* $\Pi_{i=1}^n X_i$ *est une variable aléatoire.*

Montrer « à la main » que la somme de deux variables aléatoires X_1 et X_2 est une variable aléatoire demande pas mal de patience — p.ex. pensez au cas de variables de Bernoulli. Mais avec le lemme précédent, cela devient assez facile.

Démonstration. D'après ce qui précède, $\overline{X}$ est une application mesurable de $(\Omega, \mathcal{F})$ vers $(\mathbb{R}^n, \mathcal{F}_B)$. Mais maintenant $\Phi : \mathbb{R}^n \to \mathbb{R}$ donnée par $\Phi(\overline{x}) = \sum_{i=1}^n a_i x_i$ est continue de $(\mathbb{R}^n, \tau_B)$ vers $(\mathbb{R}, \tau_B)$, et en particulier elle est mesurable puisque, par définition de la continuité, l'image réciproque de tout ouvert est ouverte !

De plus, on vérifie directement que si $f_1 : (\Omega, \mathcal{F}) \to (\Omega_1, \mathcal{F}_1)$ et $f_2 : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ sont mesurables, alors leur composée $f_2 \circ f_1$ est mesurable de $(\Omega, \mathcal{F})$ vers $(\Omega_2, \mathcal{F}_2)$. Ainsi, $\sum_{i=1}^n a_i X_i = \Phi(\overline{X})$ est mesurable de $(\Omega, \mathcal{F})$ vers $(\mathbb{R}, \tau_E)$ et donc une variable aléatoire. □

3.4.1 Fonction de répartition conjointe

De manière analogue au cas d'une seule variable aléatoire, il est maintenant naturel de chercher des façons de caractériser les lois des vecteurs aléatoires. Par analogie avec le cas unidimensionnel, étant donné un vecteur aléatoire $\bar{X}$, on peut considérer les fonctions $F : \mathbb{R}^n \to [0, 1]$ données par $F_{\bar{X}}(t_1, \ldots, t_n) = \mathbb{P}_{\bar{X}}((-\infty, t_1] \times \cdots \times (-\infty, t_n])$. Ces fonctions s'appellent fonctions de répartition conjointes et sont définies comme suit.

Définition 3.18 (Fonction de répartition conjointe)**.** *Toute fonction* $F : \mathbb{R}^n \to [0, 1]$ *est appelée fonction de répartition conjointe (c.d.f.) si elle satisfait les conditions suivantes :*

(1) F *est non décroissante en chacune des coordonnées.*

(2) $F(x_1, \ldots, x_n) \to 1$ *lorsque tous les* $x_i \to \infty$*.*

(3) $F(x_1, \ldots, x_n) \to 0$ *lorsqu'au moins un des* $x_i \to -\infty$*.*

(4) F *est continue à droite, ce qui signifie que pour toute suite* $(x_1^m, \ldots, x_n^m)_{m \geq 1}$ *telle que, pour tout* $m \geq 1$*, on ait* $x_i^m \geq x_i$ *et* $x_i^m \to x_i$ *lorsque* $m \to \infty$*, on a* $F(x_1^m, \ldots, x_n^m) \to F(x_1, \ldots, x_n)$*.*

10. En effet, l'énoncé d'intérêt ici est le suivant. Si $(\Omega, \mathcal{F})$ et $((\Omega_i, \mathcal{F}_i))_{1 \leq i \leq n}$ sont des espaces mesurables, alors l'application $f : (\Omega, \mathcal{F}) \to (\Pi_{1 \leq i \leq n} \Omega_i, \mathcal{F}_\Pi)$ est mesurable si et seulement si pour tout $i = 1 \ldots n$ l'application $f_i = p_i \circ f$ envoyant (Ω, F) vers $(\Omega_i, \mathcal{F}_i)$ est mesurable. Comparez cela avec l'énoncé suivant de topologie : si $f_i : (X, \tau_X) \to (Y_i, \tau_{Y_i})$ sont continues, alors $f : (X, \tau_X) \to (Y_1 \times \cdots \times Y_n, \tau_\Pi)$ donnée par $f = (f_1, \ldots, f_n)$ l'est aussi.

(5) Pour chaque $\bar{y}^1 := (y_1^1, \dots, y_n^1)$ et $\bar{y}^0 := (y_1^0, \dots, y_n^0)$ avec $y_i^1 \leq y_i^0$ pour tout $i = 1 \dots n$, on a l'inégalité d'inclusion-exclusion suivante :

$$\sum_{S \subseteq \{1,\dots,n\}} (-1)^{|S|} F(y_1^{1_{1\in S}}, \dots, y_n^{1_{1\in S}}) \geq 0.$$

Remarquez que pour $n = 1$ on retrouve le cas des fonctions de répartition usuelles. De plus, si l'on envoie $n-1$ coordonnées vers $+\infty$, on obtient la c.d.f. de la coordonnée restante :

$$F_{X_i}(x_i) = F(\infty, \dots, \infty, x_i, \infty, \dots, \infty).$$

La condition (5) semble nouvelle et peut-être un peu effrayante au premier abord, mais elle garantit simplement que pour toute boîte B :)$[y_1^1, y_2^0] \times \cdots \times [y_n^1, y_n^0]$ on a $\mathbb{P}_{\bar{X}}(B) \geq 0$. Nous verrons sur la feuille d'exercices que cette condition est nécessaire pour obtenir la correspondance bijective souhaitée entre les c.d.f. conjointes et les mesures de probabilité. Dans le cas $n = 1$, elle est automatiquement satisfaite grâce à la monotonie.

Cette connexion intervient dans deux résultats qui suivent. Un sens est à nouveau facile :

Proposition 3.19 (Fonctions de répartition conjointes de vecteurs aléatoires)**.** *Soit $\overline{X} := (X_1, \dots, X_n)$ un vecteur aléatoire défini sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Alors*

$$F_{\overline{X}}(x_1, \dots, x_n) := \mathbb{P}_{\overline{X}}(X_1 \leq x_1, \dots, X_n \leq x_n)$$

définit une fonction de répartition conjointe.

Démonstration. Ceci est laissé en exercice. □

Cependant, la partie existence et unicité à partir de la c.d.f. conjointe est technique et donc admise.

Théorème 3.20 (Existence et unicité de vecteurs aléatoires via la c.d.f. conjointe (admis))**.** *Toute c.d.f. conjointe donne lieu à une loi conjointe unique d'un vecteur aléatoire.*

Encore une fois, les vecteurs aléatoires nous donnent surtout une façon plus claire de regarder les choses. On peut par exemple maintenant reformuler l'indépendance :

Lemme 3.21 (Indépendance via la c.d.f. conjointe)**.** *Considérez un vecteur aléatoire $\overline{X} = (X_1, \dots, X_n)$ défini sur un espace de probabilité. Alors $X_1, \dots, X_n$ sont mutuellement indépendantes si et seulement si $F_{\overline{X}}(x_1, \dots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n)$ pour tout $\overline{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$.*

Beaucoup d'exemples pertinents proviennent en fait de lois conjointes, où chaque loi marginale est différente. Cependant, le cas des vecteurs gaussiens est très répandu en apprentissage automatique / statistiques et ailleurs. Pour l'énoncer, nous définissons d'abord la notion de densité pour les vecteurs aléatoires.

Définition 3.22 (Vecteurs aléatoires avec densité)**.** *Soit $\overline{X} = (X_1, \dots, X_n)$ un vecteur aléatoire et soit $f_{\overline{X}}$ une fonction intégrable non négative*[11] *de $\mathbb{R}^n \to [0, \infty)$ d'intégrale totale égale à 1. On dit que $f_{\overline{X}}$ est la densité conjointe de $\overline{X}$ si et seulement si, pour toute boîte $(a_1, b_1] \times \dots (a_n, b_n]$,*

$$\mathbb{P}_{\bar{X}}(X_1 \in (a_1, b_1], \dots, X_n \in (a_n, b_n]) = \int_{(a_1,b_1]\times\cdots\times(-a_n,b_n]} f_{\overline{X}}(\bar{x}) d\bar{x}. \tag{3.1}$$

11. Là encore, vous pouvez supposer que l'on utilise l'intégrale de Riemann. On pourrait donner une définition plus naturelle via l'intégrale de Lebesgue, mais celle-ci marche très bien aussi.

Comme dans le cas unidimensionnel, on peut aussi interpréter cette densité comme représentant la probabilité d'être dans un voisinage infinitésimal autour d'un point $\overline{t} = (t_1, \ldots, t_n)$. En effet, si $f_{\overline{X}}$ est continue, alors on peut vérifier que

$$\mathbb{P}_{\overline{X}}((X_1, \ldots, X_n) \in (t_1, \ldots, t_n) + [-\epsilon/2, \epsilon/2]^n) = f_{\overline{X}}(t_1, \ldots, t_n)\epsilon^n + o(\epsilon^n). \tag{3.2}$$

De plus, en faisant tendre $a_i \to -\infty$, pour tout $(t_1, \ldots, t_n) \in \mathbb{R}^n$ on peut poser

$$F_{\bar{X}}(t_1, \ldots, t_n) := \int_{(-\infty, t_1] \times \cdots \times (-\infty, t_n]} f_{\overline{X}}(\bar{x}) d\bar{x}$$

et vérifier que cela définit bien une c.d.f. Ainsi, comme les c.d.f. conjointes caractérisent la loi conjointe des variables aléatoires, on peut définir des lois de vecteurs aléatoires via leur densité.

On peut maintenant énoncer l'exemple clé :

Vecteur aléatoire gaussien. Le vecteur (ou aussi normal) gaussien est noté $\mathcal{N}(\overline{\mu}, C)$, où $\overline{\mu}$ est un vecteur de $\mathbb{R}^n$ et C une matrice symétrique définie positive de taille $n \times n$. On appellera $\overline{\mu}$ la moyenne du vecteur gaussien, et la matrice C la matrice de covariance — nous verrons dans quelques cours les raisons de ce vocabulaire. La densité du vecteur gaussien est donnée par :

$$f_{\overline{X}}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(C)}} \exp\Big(-\frac{1}{2}(\overline{x} - \overline{\mu})^T C^{-1}(\overline{x} - \overline{\mu})\Big).$$

Lorsque $\overline{\mu} = 0$ et C est la matrice identité I_n de taille $n \times n$, on appelle la loi $\mathcal{N}(0, I_n)$ la gaussienne standard dans $\mathbb{R}^n$.

Les vecteurs gaussiens s'accordent bien avec l'algèbre linéaire : si l'on applique une transformation linéaire à un vecteur gaussien, alors il reste gaussien. En fait, tous les vecteurs gaussiens dans $\mathbb{R}^n$ s'obtiennent simplement comme transformations linéaires de la gaussienne standard.

Section 4

Espérance mathématique

Nous allons continuer à travailler avec des variables aléatoires et commencer à regarder plusieurs caractéristiques ou propriétés différentes de leur loi, basées sur le concept d'espérance mathématique. À bien des égards, l'espérance mathématique d'une distribution de probabilité est le nombre qu'il faut donner si l'on demande un seul nombre pour décrire la distribution.

L'espérance mathématique, ou simplement « espérance », ou « valeur attendue », ou « moyenne », est un nom un peu sophistiqué pour désigner la moyenne dans le contexte des mesures de probabilité. Son introduction aux débuts de la théorie des probabilités était grossièrement motivée par une question très simple :

— Supposons qu'on vous propose le marché suivant : on lance un dé et vous recevez autant de francs qu'il apparaît de points sur la face supérieure ; mais en retour vous devez payer n francs indépendamment du résultat. Combien de francs devriez-vous accepter de payer ?

Alors même que la « bonne » réponse dépend encore de certaines conditions et hypothèses supplémentaires. Cependant, le résultat mathématique suivant, énoncé vaguement, donne une certaine intuition du problème (et était utilisé à l'époque des jeux d'argent !) :

— Soient $X_1, X_2, \dots$ des lancers indépendants d'un dé. Posons $S_n = \sum_{i=1}^{n} X_i$. Alors, à la limite $n \to \infty$, $\frac{S_n}{n}$ converge vers $\frac{1+2+3+4+5+6}{6} = 3.5$.

Ce résultat est un cas particulier de la loi des grands nombres, et il vous dit que le gain moyen d'un lancer de dé est 3.5. Est-ce que cela signifie que vous devriez proposer n'importe quoi en dessous de 3.5 francs ? En méditant sur ce problème très terrestre, plongeons-nous dans la théorie mathématique. “‘ “‘tex

4.1 Espérance d'une variable aléatoire discrète

Nous commençons par le cas discret afin de poser des bases claires. Le cas général peut être vu comme une extension de celui-ci :

Définition 4.1 (Espérance d'une variable aléatoire discrète)**.** *Soit X une variable aléatoire discrète définie sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et de support S. On dit que X admet une espérance, ou que X est intégrable, si $\sum_{x \in S} |x| \mathbb{P}(X = x) < \infty$.*

Pour une variable aléatoire intégrable X, l'espérance de X, notée $\mathbb{E}(X)$, est définie par

$$\mathbb{E}(X) = \sum_{x \in S} x \mathbb{P}(X = x).$$

Remarque 4.2. *Observons ce qui suit :*

— *La condition d'intégrabilité est celle de sommabilité absolue : sinon, l'ordre dans la somme importerait, et il n'y aurait pas de valeur unique pour l'espérance. On a que X est intégrable si et seulement si $|X|$ l'est.*

— *L'espérance ne dépend que de la loi $\mathbb{P}_X$ de la variable aléatoire et non de l'espace de probabilité sous-jacent.*

— *Les variables aléatoires discrètes à support fini sont toujours intégrables.*

Avant de démontrer certaines propriétés qui rendent l'espérance extrêmement utile, regardons quelques exemples :

Variable aléatoire déterministe
Si une variable aléatoire X prend une valeur $x \in \mathbb{R}$ avec probabilité 1, alors son espérance est clairement égale à x.

Variable aléatoire de Bernoulli
Soit E un événement sur un espace de probabilité, et considérons la variable aléatoire 1_E. Comme son support est fini, elle est intégrable. Par définition de l'espérance, on a directement $\mathbb{E}(1_E) = \mathbb{P}(E)$. Ainsi, en particulier, si X est une variable $Ber(p)$, alors son espérance vaut simplement $\mathbb{E}(X) = p$.

Variable aléatoire uniforme
Considérons la variable aléatoire uniforme U_n sur $\{1, 2, \ldots, n\}$. Là encore, comme elle ne prend qu'un nombre fini de valeurs, elle est intégrable. Son espérance vaut

$$\mathbb{E}(U_n) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}.$$

Variable aléatoire de Poisson
Considérons une variable aléatoire de Poisson P de paramètre $\lambda > 0$. Le support d'une variable de Poisson n'est pas fini ; il faut donc vérifier qu'elle est intégrable. En fait, le même calcul donne aussi son espérance :

$$\mathbb{E}(P) = \sum_{n\geq 0} n\mathbb{P}(P = n) = \sum_{n\geq 1} n\frac{e^{-\lambda}\lambda^n}{n!} = \lambda e^{-\lambda}\sum_{m\geq 0}\frac{\lambda^m}{m!} = \lambda.$$

Ainsi, même si une variable aléatoire peut prendre des valeurs arbitrairement grandes, son espérance peut être finie. Ce n'est cependant pas toujours le cas. Par exemple :

— Considérons une variable aléatoire X telle qu'elle prenne la valeur 2^n avec probabilité 2^{-n}. Alors clairement $\mathbb{E}(X) = \infty$ et X n'est pas intégrable.

Si une variable aléatoire est non négative, son espérance n'existe que lorsqu'elle est "trop grande", c.-à-d. infinie. On définit parfois malgré tout l'espérance pour toute variable aléatoire positive, en disant simplement que $\mathbb{E}(X) = \infty$ dans le cas où elle est infinie.

Vous verrez plus d'exemples sur la feuille d'exercices :

Exercice 4.1 (Espérances de variables aléatoires discrètes)**.** *Montrer que l'espérance d'une variable aléatoire binomiale $Bin(n, p)$ est égale à np. Montrer aussi que l'espérance d'une variable aléatoire géométrique de paramètre p est égale à $1/p$.*

Comme mentionné, l'espérance est en un sens le meilleur nombre unique pour décrire une distribution de probabilité. Il y a plusieurs raisons de le dire, et la première est la suivante : elle minimise l'erreur attendue que l'on fait en estimant la valeur de X par un seul nombre déterministe, lorsqu'on mesure l'erreur en termes de moyenne des carrés des écarts.

Lemme 4.3. *Soit X une variable aléatoire discrète intégrable de support S. Supposons que X^2 soit aussi intégrable. Alors $c = \mathbb{E}(X)$ minimise l'expression $g(c) := \sum_{x\in S}(x-c)^2\mathbb{P}(X = x)$.*

De plus, montrer à partir de la définition que la valeur $g(\mathbb{E}(X))$ peut s'écrire $\mathbb{E}((X-\mathbb{E}(X))^2)$. On appelle cela la variance de X.

Démonstration. Ceci est sur la feuille d'exercices.

□

Une autre bonne raison d'aimer l'espérance est le fait qu'elle est un opérateur linéaire sur les variables aléatoires. Avec cela, vérifions aussi quelques autres propriétés simples.

Proposition 4.4. *Soient X, Y deux variables aléatoires discrètes intégrables définies sur le même espace de probabilité. Alors l'espérance satisfait les propriétés suivantes :*

- *Elle est linéaire : $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$ pour tout $\lambda \in \mathbb{R}$. De plus, $X+Y$ est intégrable et $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.*
- *Si $X \geq 0$, i.e. $\mathbb{P}(X \geq 0) = 1$, alors $\mathbb{E}(X) \geq 0$.*
- *Si $X \geq Y$, i.e. $\mathbb{P}(X \geq Y) = 1$, alors $\mathbb{E}(X) \geq \mathbb{E}(Y)$. En déduire que si $\mathbb{P}(c \leq X \leq C) = 1$, alors $c \leq \mathbb{E}(X) \leq C$.*
- *On a $\mathbb{E}(|X|) \geq |\mathbb{E}(X)|$.*

Démonstration. Le fait que $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$ découle directement de la définition. Montrons ensuite que $X+Y$ est intégrable et que $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$. Notons S_X, S_Y les supports de X et Y respectivement. Notons S_{X+Y} le support de $X+Y$. Remarquons que

$$\mathbb{P}(X+Y=s) = \sum_{x \in S_X} \sum_{y \in S_Y} \mathbb{P}(X=x, Y=y) 1_{x+y=s}.$$

On peut donc écrire

$$\sum_{s \in S_{X+Y}} |s| \mathbb{P}(X+Y=s) = \sum_{s \in S_{X+Y}} \sum_{x \in S_X} \sum_{y \in S_Y} |x+y| \mathbb{P}(X=x, Y=y) 1_{x+y=s}.$$

Par l'inégalité triangulaire, on peut majorer $|x+y| \leq |x| + |y|$ et obtenir

$$\sum_{s \in S_{X+Y}} |s| \mathbb{P}(X+Y=s) \leq \sum_{s \in S_{X+Y}} \sum_{x \in S_X} \sum_{y \in S_Y} (|x|+|y|) \mathbb{P}(X=x, Y=y) 1_{x+y=s}. \tag{4.1}$$

Observons maintenant que, pour x et y fixés, on a soit $\mathbb{P}(X=x, Y=y) = 0$, soit $x+y \in S_{X+Y}$, et dans ce cas

$$\mathbb{P}(X=x, Y=y) = \mathbb{P}(X=x, Y=y) \sum_{s \in S_{X+Y}} 1_{x+y=s}.$$

De plus, pour x fixé, la loi des probabilités totales donne

$$\sum_{y \in S_Y} \mathbb{P}(X=x, Y=y) = \mathbb{P}(X=x).$$

Ainsi, comme tout dans l'Équation (4.1) est positif, on peut maintenant permuter l'ordre des sommes, et reconnaître le membre de droite comme la somme de

$$\sum_{x \in S_X} \sum_{y \in S_Y} \sum_{s \in S_{X+Y}} |x| \mathbb{P}(X=x, Y=y) 1_{x+y=s} = \sum_{x \in S_X} |x| \mathbb{P}(X=x)$$

et

$$\sum_{y \in S_Y} \sum_{x \in S_X} \sum_{s \in S_{X+Y}} |y| \mathbb{P}(X=x, Y=y) 1_{x+y=s} = \sum_{y \in S_Y} |y| \mathbb{P}(Y=y).$$

On obtient donc la majoration

$$\sum_{s\in S_{X+Y}} |s|\mathbb{P}(X+Y=s) \leq \sum_{x\in S_X} |x|\mathbb{P}(X=x) + \sum_{y\in S_Y} |y|\mathbb{P}(Y=y),$$

ce qui donne l'intégrabilité. Ensuite, la même séparation des sommes donne aussi $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Le reste est sur la feuille d'exercices.

$$\mathbb{E}(X) = \sum_{x\in S_X} x\mathbb{P}(X=x) \leq \sum_{x\in S_X} |x|\mathbb{P}(X=x) = \mathbb{E}(|X|).$$

□

Une preuve très semblable montre le fait suivant.

Exercice 4.2. *Soient X, Y deux variables aléatoires discrètes indépendantes et intégrables. Alors XY est intégrable et $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

Cela nous permet d'arriver à l'autre propriété fondamentale de l'espérance : la moyenne empirique converge vers l'espérance mathématique, ce qui permet de justifier pourquoi l'on devrait peut-être être content de payer moins de 3.5 francs pour pouvoir rejouer indéfiniment au jeu du dé ci-dessus...

Théorème 4.5 (Une version de la loi des grands nombres)**.** *Soient $X_1, X_2, \ldots$ des variables aléatoires discrètes i.i.d. intégrables telles que X_1^2 soit aussi intégrable. Alors pour tout $\epsilon > 0$,*

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\Big| > \epsilon\Big) \to 0$$

lorsque $n \to \infty$.

Grossièrement, cette loi des grands nombres dit que si l'on répète la même expérience aléatoire de manière indépendante n fois, donnant des variables aléatoires i.i.d. $X_1, X_2, \ldots, X_n$, alors lorsque $n \to \infty$ la moyenne des X_i converge vers l'espérance de X_1. Il est assez remarquable que la distribution des variables ne joue pas de rôle plus important dans cette limite : seule l'intégrabilité et l'espérance comptent. Ces théorèmes sont liés aux théorèmes ergodiques, qui relient grossièrement les moyennes temporelles (ici n) et spatiales (ici $\mathbb{E}$).

Nous avons besoin d'un dernier ingrédient avant de prouver ceci :

Proposition 4.6 (Markov)**.** *Soit X une variable aléatoire discrète non négative et intégrable. Alors $\mathbb{P}(X \geq t) \leq t^{-1}\mathbb{E}(X)$.*

Remarque 4.7. *Cette proposition, ainsi que l'énoncé sur l'indépendance, restent bien sûr vrais pour des variables aléatoires générales ; il faut simplement d'abord définir leur espérance !*

Preuve du Théorème. Par hypothèse il existe C tel que $\mathbb{E}X_1^2 < C$. Posons $S_n = n^{-1}\sum_{i=1}^n X_i$.

Notre objectif est d'utiliser l'inégalité de Markov. Cependant, comme la valeur absolue est pénible à manipuler, nous l'appliquerons plutôt au carré, ce qui se marie bien avec la linéarité de l'espérance et la propriété d'indépendance ci-dessus :

$$\mathbb{P}(|S_n - \mathbb{E}(X_1)| > \epsilon) = \mathbb{P}((S_n - \mathbb{E}(X_1))^2 > \epsilon^2) \leq \mathbb{E}((S_n - \mathbb{E}(X_1))^2)/\epsilon^2.$$

Calculons donc $\mathbb{E}((S_n - \mathbb{E}X_1)^2)$. D'abord, en développant S_n, en ouvrant les parenthèses à l'intérieur de l'espérance, puis en utilisant la linéarité de l'espérance, on obtient

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^2) = \sum_{i,j \leq n} n^{-2} \mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)\right].$$

On a $\mathbb{E}X_j = \mathbb{E}X_1$. Ainsi, par linéarité,

$$\mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)\right] = \mathbb{E}(X_i X_j) + (\mathbb{E}(X_1))^2 - 2(\mathbb{E}(X_1))^2 = \mathbb{E}(X_i X_j) - (\mathbb{E}(X_1))^2.$$

Mais pour $i \neq j$, l'indépendance donne $\mathbb{E}(X_i X_j) = \mathbb{E}(X_i)\mathbb{E}(X_j) = (\mathbb{E}(X_1))^2$, donc

$$\mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)\right] = 0 \qquad \text{pour } i \neq j.$$

Par conséquent,

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^2) = n^{-2} \sum_{i=1}^{n} \left(\mathbb{E}(X_i^2) - (\mathbb{E}(X_1))^2\right) \leq n^{-2}\, n\, C = n^{-1} C \to 0$$

lorsque $n \to \infty$. On en déduit donc

$$\mathbb{P}(|S_n - \mathbb{E}X_1| > \epsilon) \leq \epsilon^{-2} n^{-1} C \to 0,$$

et le théorème est démontré. □

Il nous reste à prouver l'énoncé de l'Exercice 4.2 ainsi que l'inégalité de Markov. Le premier sera sur la prochaine feuille d'exercices ; l'inégalité de Markov vient maintenant :

Preuve de l'inégalité de Markov : Soit X une variable aléatoire discrète non négative et intégrable. Alors $Y_t = X 1_{X \geq t}$ est aussi une variable aléatoire discrète non négative et intégrable car $Y_t \leq X$. Or on observe que $Y_t \geq t 1_{X \geq t}$, et donc

$$\mathbb{E}(X) \geq \mathbb{E}(Y_t) \geq \mathbb{E}(t 1_{X \geq t}).$$

Mais $\mathbb{E}(t 1_{X \geq t}) = t\mathbb{P}(X \geq t)$ par linéarité et le fait que 1_E est une variable de Bernoulli. On obtient donc $\mathbb{E}(X) \geq t\mathbb{P}(X \geq t)$, comme voulu. □

J'espère que vous êtes convaincu(e) que la notion d'espérance mathématique est très utile. Nous allons maintenant voir comment la généraliser à des variables aléatoires arbitraires, pas nécessairement discrètes.

4.2 Espérance d'une variable aléatoire arbitraire

L'idée pour définir l'espérance d'une variable aléatoire générale X est de l'approximer par des variables aléatoires discrètes. Plus précisément, étant donné X, on définit les discrétisations de X par :

$$X_n(\omega) = 2^{-n} \lfloor 2^n X(\omega) \rfloor = \sum_{k \in \mathbb{Z}} k 2^{-n} 1_{X(\omega) \in [k2^{-n}, (k+1)2^{-n})}.$$

Remarquez que X_n est bien une variable aléatoire discrète : c'est une fonction croissante de X, donc c'est une variable aléatoire, et elle ne prend qu'un nombre dénombrable de valeurs, donc elle est discrète. L'exercice suivant dit que ces discrétisations approchent vraiment très bien la variable aléatoire initiale.

Exercice 4.3 (Les discrétisations sont sympathiques)**.** *Soit X une variable aléatoire définie sur $(\Omega, \mathcal{F}, \mathbb{P})$, et $(X_n)_{n\geq 1}$ ses discrétisations $X_n = 2^{-n}\lfloor 2^n X \rfloor = \sum_{k\in\mathbb{Z}} k2^{-n} 1_{X\in[k2^{-n},(k+1)2^{-n})}$.*

Montrer que, pour tout $\omega \in \Omega$, on a $X_n(\omega) \leq X(\omega) \leq X_n(\omega) + 2^{-n}$ et donc que la suite $(X_n(\omega))_{n\geq 1}$ converge vers $X(\omega)$.

On peut maintenant utiliser la définition de l'espérance $\mathbb{E}(X)$ pour les variables aléatoires discrètes X afin de définir l'espérance d'une variable aléatoire arbitraire :

Proposition 4.8 (Espérance d'une variable aléatoire)**.** *Soit X une variable aléatoire définie sur un espace de probabilité. Si $\mathbb{E}(|X_m|) < \infty$ pour un certain m, alors $\mathbb{E}(|X_n|) < \infty$ pour tout n et on dit que X est intégrable. L'espérance de X est alors définie par*

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n).$$

Remarque 4.9. *Observons à nouveau que l'espérance ne dépend que de la loi de X et non de l'espace de probabilité sous-jacent : c'est clair dans le cas discret, mais maintenant remarquez que si X et Y ont la même loi, alors leurs discrétisations X_n et Y_n ont aussi la même loi.*

Remarque 4.10. *Un petit aperçu du futur : si l'on considère $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{F}_L, \mathbb{P}_U)$ où $\mathcal{F}_L$ est la tribu de Lebesgue et $\mathbb{P}_U$ la mesure de Lebesgue (que l'on a aussi appelée mesure uniforme), alors pour toute variable aléatoire intégrable X, qui est simplement une fonction mesurable de $([0,1], \mathcal{F}_L)$ vers $(\mathbb{R}, \mathcal{F}_B)$, $\mathbb{E}X$ est son intégrale de Lebesgue. Vous verrez une construction plus générale en Analyse IV à l'aide d'approximations monotones générales.*

L'idée pour démontrer cette proposition est simplement de montrer que la suite $\mathbb{E}(X_n)$ est de Cauchy.

Démonstration. Remarquez que, d'après l'Exercice 4.3 ci-dessus, on a $X_1 - 1 \leq X_n \leq X_1 + 1$ et donc $|X_n| \leq |X_1| + 1$. Ainsi, la Proposition 4.4 implique que $\mathbb{E}(|X_n|) < \infty$ si et seulement si $\mathbb{E}(|X_1|) < \infty$, ce qui donne la première affirmation.

Nous affirmons maintenant que $\mathbb{E}(X_n)$ est une suite de Cauchy. Soient donc $m \geq n$. Alors, d'après la Proposition 4.4,

$$|\mathbb{E}(X_n) - \mathbb{E}(X_m)| = |\mathbb{E}(X_n - X_m)| \leq \mathbb{E}(|X_n - X_m|).$$

Mais on peut majorer $|X_n - X_m| \leq 2^{-n}$ grâce à l'Exercice 4.3. Ainsi, $|\mathbb{E}(X_n) - \mathbb{E}(X_m)| \leq \mathbb{E}(2^{-n}) = 2^{-n}$. Il s'ensuit que la suite $(\mathbb{E}(X_n))_{n\geq 1}$ est de Cauchy et donc converge vers une limite unique lorsque $n \to \infty$. □

Une vérification simple mais importante est que cette définition coïncide bien avec la définition précédente pour les variables aléatoires discrètes, i.e. que la Définition 4.1 de $\mathbb{E}(X)$ et la définition de $\mathbb{E}(X)$ via la Proposition 4.8 coïncident pour toute variable aléatoire discrète X. Ceci est sur la feuille d'exercices.

De plus, on peut aussi vérifier que toutes les propriétés valables pour l'espérance d'une variable aléatoire discrète restent vraies en général :

Proposition 4.11. *Soient X, Y deux variables aléatoires intégrables définies sur le même espace de probabilité. Alors l'espérance satisfait les propriétés suivantes :*

- *Elle est linéaire : $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$ pour tout $\lambda \in \mathbb{R}$. De plus, $X + Y$ est intégrable et $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.*
- *Si $X \geq 0$, i.e. $\mathbb{P}(X \geq 0) = 1$, alors $\mathbb{E}(X) \geq 0$.*

— *Si $X \geq Y$, i.e. $\mathbb{P}(X \geq Y) = 1$, alors $\mathbb{E}(X) \geq \mathbb{E}(Y)$. En déduire que si $\mathbb{P}(c \leq X \leq C) = 1$, alors $c \leq \mathbb{E}(X) \leq C$.*
— *On a $\mathbb{E}(|X|) \geq |\mathbb{E}(X)|$.*
— *Si X, Y sont indépendantes, alors XY est aussi intégrable et $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*
De plus, l'inégalité de Markov est aussi vraie.

Démonstration. Tous ces points se déduisent de la Proposition 4.4 via les discrétisations et l'Exercice 4.3. C'est une vérification un peu fastidieuse que je vous laisse. La partie sur les produits est sur la feuille d'exercices.

L'inégalité de Markov peut se prouver soit par discrétisation, soit en fait exactement par la même preuve que celle donnée plus haut. □

Voyons maintenant que, dans le cas des variables aléatoires admettant une densité, on peut utiliser l'intégrale de Riemann et la densité pour calculer l'espérance.

Proposition 4.12 (Espérance pour une v.a. avec densité)**.** *Soit X une variable aléatoire de densité f_X. Alors X est intégrable si et seulement si $\int_{\mathbb{R}} |x| f_X(x)dx < \infty$ et l'on a*

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x)dx.$$

Démonstration. Considérons les discrétisations $X_n = 2^{-n}\lfloor 2^n X \rfloor$. Remarquons que

$$\mathbb{P}(X_n \in [k2^{-n}, (k+1)2^{-n})) = \int_{k2^{-n}}^{(k+1)2^{-n}} f_X(x)dx,$$

et donc

$$\mathbb{E}(|X_1|) = \sum_{k \geq 0} k2^{-1} \int_{k2^{-1}}^{(k+1)2^{-1}} f_X(x)dx + \sum_{k \geq 1} k2^{-1} \int_{-k2^{-1}}^{(-k+1)2^{-1}} f_X(x)dx.$$

Or, si $|x| \in [k2^{-1}, (k+1)2^{-1})$, alors $k2^{-1} \leq |x| \leq k2^{-1} + 2^{-1}$. En utilisant le fait que $\int_{\mathbb{R}} f_X(x)dx = 1$ et que f_X est non négative, on conclut que

$$-1 + \int_{\mathbb{R}} |x| f_X(x)dx \leq \mathbb{E}(|X_1|) \leq 1 + \int_{\mathbb{R}} |x| f_X(x)dx.$$

Ainsi, X est intégrable si et seulement si $\int_{\mathbb{R}} |x| f_X(x)dx < \infty$.

Ensuite, comme

$$\mathbb{E}(X_n) = \sum_{k \in \mathbb{Z}} k2^{-n} \int_{k2^{-n}}^{(k+1)2^{-n}} f_X(x)dx,$$

on voit de même que

$$\mathbb{E}(X_n) \leq \int_{\mathbb{R}} x f_X(x)dx \leq \mathbb{E}(X_n) + 2^{-n}.$$

Mais $\mathbb{E}(X_n) \to \mathbb{E}(X)$ lorsque $n \to \infty$, et la proposition s'ensuit en faisant tendre $n \to \infty$. □

Calculons l'espérance pour quelques variables aléatoires bien connues :

Variable aléatoire uniforme sur $[a, b]$
Considérons une variable aléatoire uniforme U sur $[a, b]$. Rappelons que sa densité est donnée par $f_U(x) = (b-a)^{-1}1_{x\in[a,b]}$. Remarquons d'abord que U est bornée et donc intégrable. Ainsi,

$$\mathbb{E}(U) = (b-a)^{-1}\int_{\mathbb{R}} x1_{x\in[a,b]}dx = (b-a)^{-1}\int_a^b xdx = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

Variable aléatoire gaussienne
Considérons une variable normale centrée réduite $N \sim \mathcal{N}(0, 1)$. On note d'abord que

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} |x| \exp(-\frac{x^2}{2})dx = \frac{2}{\sqrt{2\pi}}\int_0^\infty x \exp(-\frac{x^2}{2})dx = \frac{2}{\sqrt{2\pi}} < \infty.$$

Ainsi, N est intégrable. On remarque ensuite que

$$\mathbb{E}(N) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} x \exp(-\frac{x^2}{2})dx = \mathbb{E}(-N),$$

car la densité de $-N$ est la même que celle de N. La Proposition 4.11 implique donc que $\mathbb{E}(N) = 0$.

Considérons maintenant une variable gaussienne générale $N_{\mu,\sigma^2} \sim \mathcal{N}(\mu, \sigma^2)$. Rappelons que l'on peut écrire $N_{\mu,\sigma^2} \sim \sigma N + \mu$; ainsi N_{μ,σ^2} est intégrable. De plus, on peut utiliser encore une fois la Proposition 4.11 pour en déduire que $\mathbb{E}N_{\mu,\sigma^2} = \sigma\mathbb{E}(N) + \mu = \mu$. C'est la raison pour laquelle μ est appelée la moyenne de la variable aléatoire gaussienne.

D'autres exemples seront sur la feuille d'exercices.

4.3 Espérance d'une fonction d'une variable aléatoire

Il se trouve que l'espérance, bien que ce ne soit qu'un nombre, est un outil très utile pour décrire une variable aléatoire. Souvent, on ne s'intéresse pas à l'espérance de certaines variables aléatoires données, mais à celle de certaines fonctions de ces variables. Par exemple, nous avons déjà vu que, pour une v.a. X, on peut s'intéresser à $\mathbb{E}\left((X - \mathbb{E}X)^2\right)$, ou, pour X, Y, à $\mathbb{E}(XY)$. En fait, comme nous le verrons, si l'on connaît $\mathbb{E}g(X)$ pour suffisamment de fonctions g, cela détermine la variable aléatoire elle-même !

Pour commencer, regardons la proposition suivante qui généralise l'exercice montrant que, pour une variable aléatoire discrète, $\mathbb{E}\left((X - s)^2\right) = \sum_{x\in S_X}(x - s)^2\mathbb{P}(X = x)$, c.-à-d. qui donne une façon agréable de calculer les espérances de fonctions d'une v.a. :

Proposition 4.13. *Soit $\overline{X} = (X_1, \ldots, X_n)$ un vecteur aléatoire défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et ϕ une fonction mesurable de $(\mathbb{R}^n, \mathcal{F}_E)$ vers $(\mathbb{R}, \mathcal{F}_E)$, de sorte que $\phi(\overline{X})$ soit une variable aléatoire.*

— *Si $X_1, \ldots, X_n$ sont toutes discrètes et que $\phi(\overline{X})$ est intégrable, alors*

$$\mathbb{E}(\phi(\overline{X})) = \sum_{\overline{x}\in S_{\overline{X}}} \phi(\overline{x})\mathbb{P}(\overline{X} = \overline{x}),$$

où $S_{\overline{X}} \subseteq \mathbb{R}^n$ est le support du vecteur aléatoire $\overline{X}$, i.e. l'ensemble des $\overline{s} = (s_1, \ldots, s_n) \in \mathbb{R}^n$ tels que $\mathbb{P}(\overline{X} = \overline{s}) > 0$ pour tout $\overline{x} \in S_{\overline{X}}$ et $\mathbb{P}(\overline{X} \in S_{\overline{X}}) = 1$.

— *Si $\overline{X}$ est un vecteur aléatoire admettant une densité, si $\phi(\overline{X})$ est intégrable et si ϕ est suffisamment "sympathique" — c.-à-d. si $\phi^{-1}([a, b))$ est Riemann-mesurable pour tout*

intervalle $[a, b)$ *— alors*

$$\mathbb{E}(\phi(\overline{X})) = \int_{\mathbb{R}^n} \phi(\overline{x}) f_{\overline{X}}(\overline{x})\, d\bar{x}.$$

La condition "suffisamment sympathique" n'est bien sûr pas entièrement naturelle. C'est encore une conséquence du fait que l'intégration de Riemann et la mesurabilité au sens borélien (ou lebesguien) ne s'accordent pas parfaitement. Après l'Analyse IV au semestre prochain, vous pourrez revisiter ces résultats et les reformuler de manière plus naturelle, si cela vous intéresse. Remarquez néanmoins que la condition est satisfaite pour beaucoup de fonctions naturelles comme x^n ou $\exp(x)$.

Démonstration. Seul le cas discret est examinable.

Notons S_ϕ le support de $\phi(\overline{X})$. Par définition, $\phi(\overline{X})$ est intégrable si et seulement si

$$\sum_{s \in S_\phi} |s| \mathbb{P}(\phi(\overline{X}) = s) < \infty,$$

et alors

$$\mathbb{E}(\phi(\overline{X})) = \sum_{s \in S_\phi} s \mathbb{P}(\phi(\overline{X}) = s).$$

Par la loi des probabilités totales, on peut écrire $\mathbb{P}(\phi(\overline{X}) = s) = \sum_{\overline{x} \in S_{\overline{X}}} \mathbb{P}\big(\overline{X} = \overline{x}\big) \cdot 1_{\phi(\overline{x})=s}$, et donc l'expression entière s'écrit

$$\sum_{s \in S_\phi} s \sum_{\overline{x} \in S_{\overline{X}}} 1_{\phi(\overline{x})=s} \mathbb{P}(\overline{X} = \overline{x}) = \sum_{\overline{x} \in S_{\overline{X}}} \mathbb{P}(\overline{X} = \overline{x}) \sum_{s \in S_\phi} s 1_{\phi(\overline{x})=s},$$

où l'on a pu permuter l'ordre des sommes car la série est absolument sommable. Enfin, pour tout $\overline{x} \in \mathbb{R}^n$ fixé, on a $\sum_{s \in S_\phi} s 1_{\phi(\overline{x})=s} = \phi(\overline{x})$, ce qui conclut.

[* Début de la partie non examinable *]

Pour démontrer le cas des variables avec densité, on utilise les discrétisations : on pose $\phi_n(\overline{x}) = 2^{-n} \lfloor \phi(\overline{x}) 2^n \rfloor$. Alors, sous l'hypothèse d'intégrabilité, on a

$$\mathbb{E}(\phi_n(\overline{X})) = \sum_{k \in \mathbb{Z}} k 2^{-n} \mathbb{P}(\phi_n(\overline{X}) = k 2^{-n}).$$

Comme $\phi^{-1}([a, b))$ est Riemann-mesurable, on peut écrire

$$k 2^{-n} \mathbb{P}(\phi_n(\overline{X}) = k 2^{-n}) = \int_{\mathbb{R}^n} 1_{\overline{x} \in \phi^{-1}([k2^{-n}, (k+1)2^{-n}))} k 2^{-n} f_{\overline{X}}(\overline{x})\, d\bar{x}.$$

À nouveau, par sommabilité absolue [12] on peut permuter somme et intégrale :

$$\mathbb{E}(\phi_n(\overline{X})) = \int_{\mathbb{R}^n} f_{\overline{X}}(\overline{x}) \sum_{k \in \mathbb{Z}} 1_{\overline{x} \in \phi^{-1}([k2^{-n}, (k+1)2^{-n}))} k 2^{-n}\, d\bar{x}.$$

12. Plus précisément, on utilise le fait que si $\sum_{n \geq 1} \int_{\mathbb{R}} |f_n(x)| dx < \infty$ ou $\int_{\mathbb{R}} \sum_{n \geq 1} |f_n(x)| dx < \infty$, alors $\int_{\mathbb{R}} \sum_{n \geq 1} f_n(x) dx = \sum_{n \geq 1} \int_{\mathbb{R}} f_n(x) dx$. Vous avez rencontré le résultat analogue pour permuter deux sommes $\sum_{k \geq 1} \sum_{n \geq 1}$, et la preuve est essentiellement la même.

Comme précédemment, pour $\overline{x}$ fixé, le terme $1_{\overline{x}\in\phi^{-1}([k2^{-n},(k+1)2^{-n}))}$ vaut 1 pour une unique valeur de k, et donc, par définition de ϕ_n,

$$\sum_{k\in\mathbb{Z}} 1_{\overline{x}\in\phi^{-1}([k2^{-n},(k+1)2^{-n}))} k2^{-n} = \phi_n(\overline{x}).$$

Ainsi,

$$\mathbb{E}(\phi_n(\overline{X})) = \int_{\mathbb{R}^n} \phi_n(\overline{x}) f_{\overline{X}}(\overline{x})\, d\bar{x}.$$

On conclut ensuite comme dans la Proposition 4.12.

[* Fin de la partie non examinable *]

□

Regarder les espérances de fonctions d'une variable aléatoire est en fait très puissant :

Proposition 4.14. *Soient X, Y deux variables aléatoires. Alors X et Y sont égales en loi si et seulement si, pour toute fonction continue et bornée $g : \mathbb{R} \to \mathbb{R}$, on a $\mathbb{E}g(X) = \mathbb{E}g(Y)$.*

Démonstration. Si X et Y ont la même loi, alors $g(X)$ et $g(Y)$ ont aussi la même loi pour toute fonction g continue et bornée. Ainsi, comme les fonctions bornées sont intégrables et que l'espérance ne dépend que de la loi, on a bien $\mathbb{E}g(X) = \mathbb{E}g(Y)$.

Dans l'autre sens, notre objectif est de montrer que, pour tout $t \in \mathbb{R}$, $F_X(t) = F_Y(t)$. Pour cela, rappelons que $F_X(t) = \mathbb{P}(X \leq t) = \mathbb{E}(1_{X\leq t})$, donc notre but sera de considérer des approximations continues $g_{t,n}$ de la fonction indicatrice $1_{x\leq t}$, définies comme suit. Fixons $t \in \mathbb{R}$ et posons $g_{t,n}(x) = 1$ si $x \leq t$, $g_{t,n}(x) = 0$ si $x \geq t + 2^{-n}$, et $g_{t,n}(x) = 1 - 2^n(x - t)$ sur l'intervalle $(t, t + 2^{-n})$.

Alors, d'une part,

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{E}(1_{X\leq t}) \leq \mathbb{E}(g_{t,n}(X)),$$

et d'autre part,

$$\mathbb{E}(g_{t,n}(X)) \leq \mathbb{E}(1_{X\leq t+2^{-n}}) = \mathbb{P}(X \leq t + 2^{-n}) = F_X(t + 2^{-n}).$$

Ainsi, par continuité à droite de $F_X(t)$, on voit que $\mathbb{E}(g_{t,n}(X))$ converge vers $F_X(t)$ lorsque $n \to \infty$. De même, $\mathbb{E}(g_{t,n}(Y))$ converge vers $F_Y(t)$ lorsque $n \to \infty$. Comme par hypothèse $\mathbb{E}(g_{t,n}(X)) = \mathbb{E}(g_{t,n}(Y))$, on conclut la proposition. □

Un argument très similaire donne le résultat utile suivant : X, Y sont indépendantes si et seulement si l'espérance se factorise pour toutes fonctions continues !

Proposition 4.15. *Soient X, Y deux variables aléatoires. Alors*

— *Si, pour toutes fonctions $g : \mathbb{R} \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$ continues et bornées, on a*

$$\mathbb{E}\left(g(X)h(Y)\right) = \mathbb{E}g(X)\mathbb{E}h(Y), \tag{4.2}$$

alors X et Y sont indépendantes.

— *Réciproquement, si X et Y sont indépendantes, alors pour toutes fonctions mesurables $g, h : \mathbb{R} \to \mathbb{R}$ telles que $g(X)$ et $h(Y)$ soient intégrables, l'Équation (4.2) est vérifiée.*

Démonstration. La première partie se démontre comme la proposition précédente.

D'après le Lemme 3.21, pour montrer que X et Y sont indépendantes, il suffit de prouver que pour tous $s, t \in \mathbb{R}$,

$$F_{(X,Y)}(s,t) = F_X(s)F_Y(t).$$

Rappelons ensuite que

$$F_{(X,Y)}(s,t) = \mathbb{E}1_{X\leq s,\, Y\leq t} = \mathbb{E}(1_{X\leq s}1_{Y\leq t}).$$

On suit la stratégie de la Proposition 4.14. En effet, considérons les mêmes fonctions continues $g_{t,n}(x)$ satisfaisant $1_{x\leq t} \leq g_{t,n}(x) \leq 1_{x\leq t+2^{-n}}$.

En utilisant l'expression de $F_{(X,Y)}$ ci-dessus, la définition de $g_{t,n}$ et les propriétés de l'espérance, on peut majorer

$$F_{(X,Y)}(s,t) \leq \mathbb{E}\big(g_{s,n}(X)g_{t,n}(Y)\big) \leq F_{(X,Y)}(s+2^{-n}, t+2^{-n}).$$

Par hypothèse,

$$\mathbb{E}\big(g_{s,n}(X)g_{t,n}(Y)\big) = \mathbb{E}g_{s,n}(X)\,\mathbb{E}g_{t,n}(Y).$$

Par continuité à droite de $F_{(X,Y)}$, on sait que $F_{(X,Y)}(s+2^{-n}, t+2^{-n})$ converge vers $F_{(X,Y)}(s,t)$, et donc $\mathbb{E}(g_{s,n}(X)g_{t,n}(Y))$ aussi. De plus, on a vu que $\mathbb{E}g_{s,n}(X)$ converge vers $F_X(s)$, et de même $\mathbb{E}g_{t,n}(Y)$ converge vers $F_Y(t)$. On conclut ainsi que $F_{(X,Y)}(s,t) = F_X(s)F_Y(t)$, comme voulu.

Pour l'autre direction, observons d'abord ce qui suit (ce sera sur la feuille d'exercices) :

Exercice 4.4. *Montrer que si X, Y sont indépendantes, alors $g(X)$ et $h(Y)$ le sont aussi.*

Étant donné cela, le deuxième point suit dès qu'on montre que pour deux variables aléatoires intégrables et indépendantes X, Y, on a $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, ce qui était sur la feuille d'exercices. □

4.4 Variance et covariance

À côté de la valeur moyenne ou espérance, un paramètre (ou caractéristique) essentiel d'une variable aléatoire est sa variance (et son écart-type, qui n'est que la racine carrée de la variance).

Elle mesure l'écart par rapport à la moyenne, et nous l'avons déjà rencontrée en caractérisant l'espérance comme minimiseur de déviation :

Définition 4.16 (Variance d'une variable aléatoire)**.** *Soit X une variable aléatoire intégrable. Si $\mathbb{E}(|X|^2) < \infty$, on dit que X admet un second moment fini et on définit sa variance par*

$$Var(X) := \mathbb{E}\big((X - \mathbb{E}X)^2\big) \geq 0.$$

L'écart-type est défini par $\sigma(X) := \sqrt{VarX}$.

Remarquons que $(X-\mathbb{E}X)^2$ est bien intégrable lorsque $|X|^2$ l'est, puisque l'on peut écrire $(X-\mathbb{E}X)^2 \leq 2|X|^2+2(\mathbb{E}X)^2$. Un outil utile pour calculer la variance est de remarquer qu'en développant le carré, on obtient

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \mathbb{E}X)^2\right) = \mathbb{E}(X^2) - 2\mathbb{E}(X\mathbb{E}X) + (\mathbb{E}X)^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.$$

Calculons maintenant quelques variances à l'aide de cette formule :

— La variance d'une variable aléatoire de Bernoulli $X \sim Ber(p)$ est $\mathbb{E}(X^2) - (\mathbb{E}X)^2 = p - p^2 = p(1-p)$. Pourquoi est-ce raisonnable ?

— De même, on peut calculer la variance d'une variable exponentielle $X \sim Exp(\lambda)$. En effet, comme x^2 satisfait les conditions de la Proposition 4.13, on peut écrire

$$\mathbb{E}X^2 = \lambda \int_0^\infty x^2 \exp(-\lambda x)\, dx.$$

On calcule alors, en intégrant par parties deux fois,

$$\lambda \int_0^\infty x^2 \exp(-\lambda x)\, dx = 2 \int_0^\infty x \exp(-\lambda x)\, dx = 2\lambda^{-1}\mathbb{E}X = 2\lambda^{-2}.$$

Ainsi $\mathrm{Var}(X) = \lambda^{-2}$.

La variance indique à quel point la variable aléatoire fluctue (ou s'écarte) autour de sa moyenne, comme l'illustre par exemple le lemme suivant, dont la preuve était sur la feuille d'exercices.

Lemme 4.17 (Inégalité de Chebyshev)**.** *Soit X une variable aléatoire intégrable de variance finie. Alors*

$$\mathbb{P}(|X - \mathbb{E}X| > t) \leq \frac{Var(X)}{t^2}.$$

4.4.1 Covariance et corrélation

Comme discuté, on s'intéresse souvent à la façon dont deux variables aléatoires sont reliées. Nous avons déjà vu la notion d'indépendance : les variables sont indépendantes si elles ne s'influencent pas du tout. À l'autre extrême, il y a le cas où elles sont égales, i.e. $\mathbb{P}(X = Y) = 1$, auquel cas on dit que $X = Y$ presque sûrement. Ces deux notions sont très fortes. La relation précise entre deux variables est encodée dans leur loi jointe, mais cela peut être assez compliqué.

Nous introduisons ici une mesure plus simple et plus faible de la relation entre deux variables, et une manière de quantifier (dans une certaine mesure) leur niveau de dépendance.

Définition 4.18 (Covariance et corrélation)**.** *Supposons que X, Y soient deux variables aléatoires intégrables de variance finie, définies sur le même espace de probabilité. La covariance de X et Y, notée $Cov(X,Y)$, est définie par*

$$Cov(X,Y) = Cov(Y,X) = \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\right) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y.$$

Si ni X ni Y n'est presque sûrement constante, alors la corrélation $\rho(X,Y)$ est définie par

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}.$$

Une première question est : pourquoi la covariance est-elle bien définie ? i.e. pourquoi $\mathbb{E}(XY)$ est-elle finie lorsque X et Y ont une variance finie ? Cela découle de l'inégalité de Cauchy-Schwarz, que vous avez probablement déjà vue sous une certaine forme. Vous trouverez une preuve non examinable à la fin de la section.

Théorème 4.19 (Inégalité de Cauchy-Schwarz)**.** *Soient X, Y deux variables aléatoires sur $(\Omega, \mathcal{F}, \mathbb{P})$ telles que X^2 et Y^2 soient intégrables. Alors $|XY|$ est aussi intégrable, et de plus*

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

De plus, il y a égalité si et seulement si $|X| = \lambda|Y|$ presque sûrement pour un certain $\lambda > 0$.

Remarquez qu'en particulier il s'ensuit que

$$\mathbb{E}(XY) \leq |\mathbb{E}(XY)| \leq \mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Les cas d'égalité pertinents peuvent aussi être déterminés.

En utilisant cette inégalité, on voit que non seulement la covariance et la corrélation sont bien définies, mais aussi que la corrélation de valeur absolue 1 signifie que les variables sont (presque sûrement) liées de façon affine.

Lemme 4.20 (Covariance et dépendance)**.** *Soient X, Y deux variables aléatoires de variance finie strictement positive définies sur le même espace de probabilité.*

- *Alors la corrélation $\rho(X, Y)$ appartient à $[-1, 1]$. De plus, elle vaut 1 si et seulement s'il existe $\lambda > 0$ et $c \in \mathbb{R}$ tels que $X = \lambda Y + c$ presque sûrement ; elle vaut -1 si et seulement s'il existe $\lambda > 0$ et $c \in \mathbb{R}$ tels que $X = -\lambda Y + c$ presque sûrement.*
- *De plus, si X, Y sont indépendantes, intégrables et de variance finie, alors leur covariance est nulle.*

Démonstration. Le premier point découle de l'inégalité de Cauchy-Schwarz.

Pour le second point, on calcule :

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Or, par indépendance de X et Y, on sait que $\mathbb{E}(XY) = \mathbb{E}X\,\mathbb{E}Y$, ce qui conclut. □

Étant donné un vecteur aléatoire, il est souvent utile de définir la covariance entre chaque paire de composantes.

Définition 4.21 (Matrice de covariance)**.** *Soit $\overline{X} = (X_1, \ldots, X_n)$ un vecteur aléatoire tel que toutes ses composantes aient une variance finie. La matrice de covariance Σ est définie par*

$$\Sigma_{i,j} = Cov(X_i, X_j).$$

En fait, nous avons déjà rencontré une matrice de covariance ! En effet, pour un vecteur aléatoire gaussien $\mathcal{N}(\overline{\mu}, C)$, la matrice symétrique définie positive C est la matrice de covariance et $\overline{\mu} = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)$:

Exercice 4.5 (Indépendance et gaussiennes)**.** *Montrer que pour un vecteur gaussien $\bar{X} \sim \mathcal{N}(\overline{\mu}, C)$, la matrice C est la matrice de covariance et $\overline{\mu} = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)$. Montrer que dans le cas d'un vecteur gaussien, si $Cov(X_i, X_j) = 0$, alors X_i et X_j sont indépendantes.*

Remarquez que cela signifie en particulier qu'un vecteur gaussien est entièrement déterminé par sa moyenne et sa covariance, ce qui est très agréable !

4.5 Moments d'une variable aléatoire

Nous avons vu que $\mathbb{E}(X)$ et $\mathbb{E}((X - \mathbb{E}X)^2)$ contiennent des informations précieuses sur une variable aléatoire X. De plus, nous avons vu que si l'on considère $\mathbb{E}g(X)$ pour toutes les fonctions g continues et bornées, alors cela détermine entièrement la loi de X. Mais c'est déjà beaucoup d'information ! Un objectif intermédiaire serait de demander $\mathbb{E}X^n$ pour tous les $n \geq 1$. Est-ce que connaître cela détermine la variable aléatoire ?

Définition 4.22 (Moments d'une v.a.)**.** *Soit X une variable aléatoire et $n \in \mathbb{N}$. Si $\mathbb{E}|X|^n < \infty$, on dit que X admet un n-ième moment. On appelle $\mathbb{E}X^n$ le n-ième moment de X.*

Pour comprendre la relation entre différents moments, rappelons l'inégalité de Jensen. Une fonction $\phi : \mathbb{R} \to \mathbb{R}$ est dite convexe si pour tous x, y et tout $\lambda \in [0, 1]$ on a

$$\phi(\lambda x + (1 - \lambda)y) \leq \lambda\phi(x) + (1 - \lambda)\phi(y).$$

On appelle $\lambda x + (1 - \lambda)y$ une combinaison convexe de x et y. En utilisant ce vocabulaire, l'inégalité de Jensen peut être reformulée en disant que l'image par ϕ d'une combinaison convexe de deux points est toujours plus petite que la combinaison convexe des images des deux points par ϕ. (Que signifie cela géométriquement ?)

L'inégalité de Jensen dans un cadre probabiliste s'énonce comme suit :

Théorème 4.23 (Inégalité de Jensen)**.** *Soit X une variable aléatoire intégrable et ϕ une fonction convexe telle que $\phi(X)$ soit aussi intégrable*[13]*. Alors*

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

Remarquez la similarité avec la propriété définissant la convexité : $\mathbb{E}X$ peut être vu comme une combinaison convexe des valeurs possibles de X. Ainsi, par exemple, si X ne prend que deux valeurs x, y avec probabilités λ et $1 - \lambda$, alors l'inégalité de Jensen n'est qu'une reformulation de la propriété de convexité.

J'espère que vous avez vu (et verrez) de nombreuses preuves de cette belle inégalité. Il y en a néanmoins une en annexe à cette section, pour complétude.

Comme corollaire, on a le lemme simple suivant, qui dit que l'existence de moments d'ordre élevé implique l'existence de moments d'ordre plus faible :

Lemme 4.24. *Soit X une variable aléatoire définie sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ admettant un n-ième moment. Alors X admet aussi un m-ième moment pour tout $m \leq n$, et de plus*

$$\mathbb{E}|X|^n \geq (\mathbb{E}(|X|^m))^{n/m}.$$

Démonstration. Soit $m \leq n$. Remarquons d'abord que si $|X|^n$ est intégrable, alors $|X|^m$ l'est aussi pour $m \leq n$. En effet, on peut majorer

$$|X(\omega)|^m \leq \max(|X(\omega)|^n, 1) \leq |X(\omega)|^n + 1,$$

et ainsi l'intégrabilité de $|X|^m$ découle de celle de $|X|^n$.

Maintenant, pour $n \geq m$, considérons $\phi(x) = |x|^{n/m}$. C'est une fonction convexe. Ainsi, comme $|X|^m$ et $|X|^n = \phi(|X|^m)$ sont intégrables, on peut appliquer Jensen à ϕ et $|X|^m$ et obtenir

$$\mathbb{E}|X|^n = \mathbb{E}(\phi(|X|^m)) \geq \phi(\mathbb{E}|X|^m) = (\mathbb{E}(|X|^m))^{n/m},$$

ce qui conclut. □

En particulier, si le second moment de X existe, alors X est intégrable et de variance finie. Beaucoup de variables aléatoires en statistiques ou en numérique auront une variance finie, donc il est utile d'avoir une bonne condition pour cela. Vous verrez sur la feuille d'exercices que la réciproque est fausse : il y aura des exemples de variables intégrables de variance infinie, etc.

13. Rappel : une fonction convexe est continue, et donc si X est une variable aléatoire, alors $\phi(X)$ en est aussi une.

L'existence de moments influence directement le comportement des queues de distribution. En effet, par l'inégalité de Markov, si $\mathbb{E}|X|^n < \infty$, alors

$$\mathbb{P}(X > t) \leq \mathbb{P}(|X|^n > t^n) \leq \frac{\mathbb{E}|X|^n}{t^n},$$

c.-à-d. que la queue se comporte comme $O(t^{-n})$. Dans le cas de la variance finie, on ne savait par exemple que la queue se comportait comme $O(t^{-2})$. En mots simples : avoir des moments plus élevés signifie que des valeurs très grandes sont prises avec une probabilité plus petite.

Venons-en maintenant à la question intéressante : les moments déterminent-ils de manière unique la distribution ? Ceci est vrai dans une assez grande généralité, mais pas toujours. Nous allons ici démontrer un résultat partiel :

Proposition 4.25. *Soient X, Y deux variables aléatoires presque sûrement bornées, i.e. telles que presque sûrement $X \in [-A, A]$ et $Y \in [-A, A]$ pour un certain $A > 0$. Supposons de plus que $\mathbb{E}X^n = \mathbb{E}Y^n$ pour tout $n \in \mathbb{N}$. Alors X et Y ont la même loi.*

Avant de commencer la preuve, remarquons que, trivialement, pour des variables bornées tous les moments existent : si X est bornée alors chaque $|X|^n$ l'est aussi. La preuve que nous donnons s'appuie sur le résultat suivant, qui dit que l'on peut approximer toute fonction continue sur un intervalle compact arbitrairement bien par des polynômes :

Théorème 4.26 (Stone-Weierstrass)**.** *Soit f une fonction continue sur un intervalle $I = [-A, A]$. Alors f peut être approximée uniformément par des polynômes : i.e. il existe une suite de polynômes $(P_n)_{n\geq 1}$ telle que $(P_n)_{n\geq 1}$ converge vers f dans $(C(I, \mathbb{R}), d_\infty)$, où comme d'habitude $d_\infty(f, g) = \sup_{x\in I} |f(x) - g(x)|$.*

Vous verrez très probablement la preuve de ce théorème dans plusieurs cours, sous plusieurs points de vue. Il y a une preuve probabiliste courte, mais non examinable, à la fin de la sous-section. Voyons ici comment cela implique la proposition.

Preuve de la Proposition 4.25. La proposition découle assez facilement du théorème de Stone-Weierstrass. En effet, par hypothèse et par linéarité de l'espérance, on a $\mathbb{E}P(X) = \mathbb{E}P(Y)$ pour tout polynôme P.

Notre but est d'utiliser la Proposition 4.14, i.e. de montrer que $\mathbb{E}\widehat{g}(X) = \mathbb{E}\widehat{g}(Y)$ pour toute fonction continue bornée $\widehat{g}$. Remarquons qu'une telle $\widehat{g}$ induit une fonction continue $g : [-A, A] \to \mathbb{R}$ par restriction. De plus, comme $X, Y \in [-A, A]$ presque sûrement, on a $\mathbb{E}\widehat{g}(X) = \mathbb{E}g(X)$, et il suffit donc de montrer $\mathbb{E}g(X) = \mathbb{E}g(Y)$ pour toute fonction continue sur $[-A, A]$.

Soit donc g une telle fonction. Par Stone-Weierstrass, pour tout $\epsilon > 0$, il existe un polynôme P_ϵ tel que $d_\infty(g, P_\epsilon) < \epsilon$. Comme $\mathbb{E}P_\epsilon(X) = \mathbb{E}P_\epsilon(Y)$, on peut écrire

$$|\mathbb{E}g(X) - \mathbb{E}g(Y)| = |\mathbb{E}g(X) - \mathbb{E}P_\epsilon(X) + \mathbb{E}P_\epsilon(Y) - \mathbb{E}g(Y)|,$$

et majorer cela, par l'inégalité triangulaire, par

$$|\mathbb{E}\left(g(X) - P_\epsilon(X)\right)| + |\mathbb{E}\left(g(Y) - P_\epsilon(Y)\right)|.$$

De plus, $|\mathbb{E}\left(g(X) - P_\epsilon(X)\right)| \leq \mathbb{E}|g(X) - P_\epsilon(X)|$. Or, comme $X \in [-A, A]$ presque sûrement et que $|g(x) - P_\epsilon(x)| < \epsilon$ pour $x \in [-A, A]$, on a $|g(X) - P_\epsilon(X)| < \epsilon$ presque sûrement, et donc (par les propriétés de l'espérance) $\mathbb{E}|g(X) - P_\epsilon(X)| \leq \epsilon$.

On conclut donc que $|\mathbb{E}g(X) - \mathbb{E}g(Y)| \leq 2\epsilon$, et comme $\epsilon > 0$ était arbitraire, on obtient $\mathbb{E}g(X) = \mathbb{E}g(Y)$. Comme g était arbitraire, la proposition découle alors de la Proposition 4.14. □

Pour des variables qui ne sont pas presque sûrement bornées, cette caractérisation peut échouer pour plusieurs raisons. D'abord, bien sûr, tous les moments peuvent ne pas exister, et dans ce cas les quelques moments existants peuvent ne pas caractériser la loi. Ensuite, même si tous les moments existent, ils peuvent croître trop vite pour caractériser la distribution :

Exercice 4.6 (Problème des moments)**.** *Soit X une variable normale centrée réduite. Montrer que $W = \exp(X)$ admet tous les moments et calculer ces moments. Soit $a > 0$, et considérons une variable aléatoire discrète Y_a de support*

$$S_a = \{ae^m : m \in \mathbb{Z}\}$$

et définie par

$$\mathbb{P}(Y_a = ae^m) = \frac{1}{Z} a^{-m} e^{-m^2/2}$$

où $Z = \sum_{m\in\mathbb{Z}} a^{-m}e^{-m^2/2}$ (pourquoi est-ce fini ?). Montrer que Y_a admet tous les moments et que, de plus, pour tout $n \in \mathbb{N}$,

$$\mathbb{E}W^n = \mathbb{E}\exp(nX) = \mathbb{E}Y_a^n.$$

4.6 Fonction génératrice des moments et fonction caractéristique

Nous avons considéré les moments et vu qu'ils peuvent donner une collection dénombrable de nombres qui caractérisent la variable aléatoire sous-jacente. Mais si, au lieu des moments, on regarde une autre famille de fonctions $g(X)$ et leurs espérances ?

Les fonctions les plus populaires ensuite sont peut-être $x \mapsto \exp(ax)$. Et en effet, regarder leurs espérances s'avère très utile !

Définition 4.27 (Fonction génératrice des moments)**.** *Si X est une variable aléatoire telle que $\exp(tX)$ soit intégrable sur un intervalle $I = (-c, c)$ autour de 0, on dit que X admet une fonction génératrice des moments (MGF) dans un voisinage de 0 et on note*

$$M_X(t) = \mathbb{E}\exp(tX) \qquad \text{pour } t \in I.$$

Le nom vient du fait que, lorsque $M_X(t)$ existe sur un petit intervalle, on peut écrire

$$M_X(t) = \mathbb{E}(\exp(tX)) = \mathbb{E}\Big(\sum_{n\geq 1} \frac{t^n X^n}{n!}\Big).$$

En vérifiant que l'on peut échanger la somme et l'espérance (sur la feuille d'exercices), on obtient

$$M_X(t) = \sum_{n\geq 1} \frac{t^n}{n!} \mathbb{E}X^n.$$

En particulier, si l'on regarde $M_X(t)$ comme une fonction de t, alors $\frac{d^n}{dt^n} M_X(t)$ évalués en $t = 0$ donnent le n-ième moment. Nous ne ferons pas ce calcul, qui n'est pas examinable. Pour $t < 0$, la fonction génératrice des moments est aussi appelée la transformée de Laplace.

Définition 4.28 (Fonction caractéristique)**.** *Soit X une variable aléatoire quelconque. Alors*

$$\psi_X(t) = \mathbb{E}e^{itX} = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX)$$

s'appelle la fonction caractéristique de X.

Ce qui est agréable, c'est que la fonction caractéristique existe pour tout $t \in \mathbb{R}$ puisque $\cos(tX)$ et $\sin(tX)$ sont bornées et continues, donc intégrables. De plus, dans le cas des variables à densité, elle correspond à la transformée de Fourier de la densité.

Il se trouve que la fonction génératrice des moments et la fonction caractéristique déterminent aussi la loi. Nous énonçons ce résultat, que vous pouvez utiliser librement, bien que la preuve soit hors du cadre de ce cours :

Théorème 4.29 (MGF / CF déterminent la loi (admis))**.** *Soient X, Y deux variables aléatoires.*

- *Supposons que $M_X(t)$ et $M_Y(t)$ existent dans un intervalle ouvert autour de 0, et de plus que $M_X(t) = M_Y(t)$ sur cet intervalle. Alors X et Y ont la même loi.*
- *Supposons que $\psi_X(t) = \psi_Y(t)$ pour tout $t \in \mathbb{R}$. Alors X et Y ont la même loi.*

En fait, les définitions et caractérisations se généralisent joliment aux vecteurs aléatoires.

Théorème 4.30 (MGF pour les vecteurs aléatoires (admis))**.** *Soit $\overline{X}$ un vecteur aléatoire à valeurs dans $\mathbb{R}^n$ tel que $\mathbb{E}e^{\langle \bar{t}, \overline{X}\rangle} < \infty$ pour $\bar{t}$ dans un voisinage ouvert de 0.*[14] *On appelle alors*

$$M_{\overline{X}}(\bar{t}) = \mathbb{E}e^{\langle \bar{t}, \overline{X}\rangle}$$

la fonction génératrice des moments de $\overline{X}$. De plus, si les MGF de deux vecteurs aléatoires $\overline{X}$ et $\overline{Y}$ coïncident dans un voisinage de 0, alors $\overline{X}$ et $\overline{Y}$ ont la même loi.

Théorème 4.31 (Fonction caractéristique des vecteurs aléatoires (admis))**.** *Soit $\overline{X}$ un vecteur aléatoire à valeurs dans $\mathbb{R}^n$. On appelle*

$$\psi_{\overline{X}}(\bar{t}) = \mathbb{E}e^{i\langle \bar{t}, \overline{X}\rangle}$$

la fonction caractéristique de $\overline{X}$. Si les fonctions caractéristiques de deux vecteurs aléatoires $\overline{X}$ et $\overline{Y}$ coïncident pour tout $\bar{t} \in \mathbb{R}^n$, alors $\overline{X}$ et $\overline{Y}$ ont la même loi.

Ces deux résultats sont extrêmement utiles. Par exemple, ils permettent de déterminer l'indépendance :

Lemme 4.32 (Indépendance et MGF)**.** *Soient X, Y deux variables aléatoires.*

- *Supposons qu'il existe un intervalle ouvert $I \subset \mathbb{R}$ contenant 0 tel que $M_X(t)$ et $M_Y(t)$ existent pour tout $t \in I$. Alors X et Y sont indépendantes si et seulement si, pour tous $t, s \in I$,*

$$M_X(t)M_Y(s) = M_{(X,Y)}((t,s)).$$

- *X et Y sont indépendantes si et seulement si, pour tous $t, s \in \mathbb{R}$,*

$$\psi_X(t)\psi_Y(s) = \psi_{(X,Y)}((t,s)).$$

Je n'ai pas eu le temps de faire cette preuve en cours, donc elle est admise. Je donne tout de même la preuve de la première partie ici ; la seconde est exactement la même.

14. Ici $\langle \cdot, \cdot \rangle$ désigne le produit scalaire sur $\mathbb{R}^n$.

Preuve non examinable. D'abord, si X et Y sont indépendantes, la condition découle directement de la Proposition 4.15. En effet, pour $t, s \in I$, on prend $g(x) = \exp(tx)$ et $h(y) = \exp(sy)$. Alors $M_X(t) = \mathbb{E}g(X)$ et $M_Y(s) = \mathbb{E}h(Y)$, et par hypothèse ces quantités sont finies. La proposition donne alors

$$M_X(t)M_Y(s) = \mathbb{E}\exp(tX + sY) = M_{(X,Y)}((t, s)).$$

Dans l'autre sens, supposons que (X, Y) soit un couple de variables aléatoires tel que pour tous $t, s \in I$,

$$M_X(t)M_Y(s) = M_{(X,Y)}((t, s)).$$

Soit $(\tilde{X}, \tilde{Y})$ un couple de variables aléatoires indépendantes tel que $\tilde{X}$ ait la loi de X et $\tilde{Y}$ la loi de Y. Alors $M_X(t) = M_{\tilde{X}}(t)$ et $M_Y(s) = M_{\tilde{Y}}(s)$ pour tous $t, s \in I$.

Par le premier point, on a $M_{\tilde{X}}(t)M_{\tilde{Y}}(s) = M_{(\tilde{X},\tilde{Y})}((t, s))$. On conclut donc que

$$M_{(X,Y)}((t, s)) = M_{(\tilde{X},\tilde{Y})}((t, s)) \quad \text{pour tous } t, s \in I.$$

D'après le Théorème 4.30, cela implique que (X, Y) et $(\tilde{X}, \tilde{Y})$ ont la même loi jointe. En particulier, X et Y sont indépendantes. □

Deuxièmement, ces outils simplifient vraiment certaines choses, en particulier les calculs avec des gaussiennes. Ici on utilise les MGF, mais on aurait tout aussi bien pu utiliser les fonctions caractéristiques.

Exercice 4.7. *Montrer que $\overline{X}$ est un vecteur gaussien de moyenne $\overline{\mu}$ et de covariance C si et seulement si*

$$M_{\overline{X}}(\overline{t}) = \exp\left(\langle \overline{t}, \overline{\mu}\rangle + \frac{1}{2}\langle \overline{t}, C\overline{t}\rangle\right).$$

En déduire que

- *Si X est une gaussienne standard sur $\mathbb{R}^n$, alors OX l'est aussi pour toute matrice orthogonale $n \times n$.*
- *Le vecteur gaussien de moyenne $\overline{\mu}$ et de covariance C sur $\mathbb{R}^n$ peut s'écrire $A\overline{Y} + \overline{\mu}$, où $\overline{Y}$ est la gaussienne standard sur $\mathbb{R}^n$ et $C = \sqrt{AA^T}$ (vous pouvez admettre qu'une telle matrice A existe, mais vous l'avez vue en algèbre linéaire !).*

Ainsi, les MGF et les fonctions caractéristiques peuvent vraiment simplifier et réduire des calculs.

L'inconvénient des fonctions génératrices des moments est qu'elles n'existent pas toujours.

Exercice 4.8. *Considérer la variable log-normale, i.e. $Z = \exp(X)$ où X est une gaussienne standard. Montrer qu'il n'existe pas de voisinage ouvert de 0 sur lequel $M_Z(t)$ existe.*

L'inconvénient de la fonction caractéristique est que son lien avec l'existence des moments est bien moins évident.

4.7 ⋆ Preuves de quelques résultats auxiliaires (non examinable) ⋆

[⋆ début de la section non examinable ⋆]

Dans cette section non examinable, nous présentons des preuves de quelques résultats auxiliaires. Je recommande particulièrement la preuve probabiliste du théorème de Stone-Weierstrass : c'est un petit bijou !

Commençons par prouver l'inégalité de Cauchy-Schwarz :

Preuve de l'inégalité de Cauchy-Schwarz. Définissons $\widehat{Y}$ et $\widehat{X}$ par $\widehat{Y} = \frac{Y}{\sqrt{\mathbb{E}(Y^2)}}$ et $\widehat{X} = \frac{X}{\sqrt{\mathbb{E}(X^2)}}$. Ceci est possible puisque X^2 et Y^2 sont intégrables. Remarquons que par définition $\mathbb{E}(\widehat{Y}^2) = \mathbb{E}(\widehat{X}^2) = 1$. De plus, l'inégalité de Cauchy-Schwarz est alors équivalente à

$$\mathbb{E}(|\widehat{X}\widehat{Y}|) \leq 1. \tag{4.3}$$

Or, pour tout $\omega \in \Omega$, on a $|\widehat{X}(\omega)\widehat{Y}(\omega)| \leq \frac{1}{2}(\widehat{X}^2(\omega) + \widehat{Y}^2(\omega))$. On en déduit que $|XY|$ est intégrable et, par les propriétés de l'espérance,

$$\mathbb{E}(|\widehat{X}\widehat{Y}|) \leq \frac{1}{2}\mathbb{E}(\widehat{X}^2 + \widehat{Y}^2) = 1,$$

ce qui donne (4.3).

Il y a égalité si et seulement si $|\widehat{X}\widehat{Y}| = \frac{1}{2}(\widehat{X}^2 + \widehat{Y}^2)$ presque sûrement, ce qui équivaut à $|\widehat{X}| = |\widehat{Y}|$ presque sûrement. Comme $\widehat{X}, \widehat{Y}$ sont des versions normalisées de X, Y, cela revient à dire qu'il existe $\lambda > 0$ tel que $|X| = \lambda|Y|$ presque sûrement. □

Ensuite, il est temps de prouver l'inégalité de Jensen. Nous allons le faire à l'aide de la caractérisation suivante des fonctions convexes :

— $\phi : \mathbb{R} \to \mathbb{R}$ est convexe si et seulement si, pour tout $x \in \mathbb{R}$, il existe un $c = c(x) \in \mathbb{R}$ tel que pour tout $y \in \mathbb{R}$,

$$\phi(x + y) \geq \phi(x) + c_x y.$$

Preuve de l'inégalité de Jensen. Posons $x = \mathbb{E}X$ et $y = X - \mathbb{E}X$. En injectant ceci dans la formulation de convexité ci-dessus, on obtient

$$\phi(X) \geq \phi(\mathbb{E}X) + c\,(X - \mathbb{E}X)$$

presque sûrement. En prenant l'espérance et en utilisant $\mathbb{E}(X - \mathbb{E}X) = 0$, on déduit

$$\mathbb{E}\phi(X) \geq \phi(\mathbb{E}X),$$

comme annoncé. □

Et finalement, la jolie preuve probabiliste du théorème de Stone-Weierstrass :

Preuve du Théorème 4.26. Par translation et changement d'échelle, il est facile de voir qu'il suffit de prouver le théorème pour $I = [0, 1]$ et f continue sur $[0, 1]$. Pour tout $x \in [0, 1]$ et $n \in \mathbb{N}$, soit $X_{n,x}$ une variable binomiale de paramètres (n, x). On définit

$$P_n(x) = \mathbb{E}f(X_{n,x}/n).$$

Par la Proposition 4.13, on a alors

$$P_n(x) = \sum_{k=0}^{n} f(k/n)\binom{n}{k}x^k(1 - x)^{n-k},$$

et donc $P_n(x)$ est un polynôme (de degré au plus n) en x.

Nous affirmons que P_n converge uniformément vers f. Comme f est continue sur $[0,1]$, elle est bornée par une constante M, et uniformément continue : pour tout $\epsilon > 0$, il existe $\delta_\epsilon > 0$ tel que si $|x-y| < \delta_\epsilon$, alors $|f(x) - f(y)| < \epsilon$.

Écrivons

$$|P_n(x) - f(x)| = \left|\mathbb{E}(f(X_{n,x}/n)) - f(x)\right| \leq \mathbb{E}|f(X_{n,x}/n) - f(x)|.$$

L'idée clé est quelque chose que nous avons déjà vu : $X_{n,x}$ est très proche de son espérance nx pour n grand. En effet, par Chebyshev et le fait que $\mathrm{Var}(X_{n,x}) = nx(1-x)$,

$$\mathbb{P}(|X_{n,x}/n - x| > t/n) = \mathbb{P}(|X_{n,x} - nx| > t) \leq \frac{\mathrm{Var}(X_{n,x})}{t^2} = \frac{nx(1-x)}{t^2}.$$

En particulier, en choisissant $t = n^{2/3}$, on obtient $\mathbb{P}(|X_{n,x}/n - x| > n^{-1/3}) \leq n^{-1/3}$.

Pour exploiter cela, on écrit :

$$\begin{aligned}\mathbb{E}|f(X_{n,x}/n) - f(x)| = {} & \mathbb{E}\left(|f(X_{n,x}/n) - f(x)|\, 1_{\{|X_{n,x}/n - x| > n^{-1/3}\}}\right) \\ & + \mathbb{E}\left(|f(X_{n,x}/n) - f(x)|\, 1_{\{|X_{n,x}/n - x| < n^{-1/3}\}}\right).\end{aligned}$$

Comme $|f| \leq M$ sur $[0,1]$, le premier terme est majoré par

$$2M\,\mathbb{P}(|X_{n,x}/n - x| > n^{-1/3}) \leq 2Mn^{-1/3}.$$

Fixons $\epsilon > 0$ et prenons n assez grand pour que $n^{-1/3} < \delta_\epsilon$. Alors sur l'événement $\{|X_{n,x}/n - x| < n^{-1/3}\}$ on a $|f(X_{n,x}/n) - f(x)| < \epsilon$, donc le second terme est majoré par ϵ. En imposant en plus $n^{-1/3} < \epsilon$, on obtient

$$\mathbb{E}|f(X_{n,x}/n) - f(x)| \leq 2Mn^{-1/3} + \epsilon \leq (2M+1)\epsilon.$$

Cette borne est uniforme en x, et comme ϵ est arbitraire, on conclut que $P_n \to f$ uniformément. □

[⋆ fin de la section non examinable ⋆] “‘ “‘tex

Section 5

Théorèmes limites

Dans cette section, nous allons considérer des suites infinies d'événements et des suites infinies de variables aléatoires. Parmi les questions qui nous intéressent :

— Quand peut-on être sûr qu'au moins un des événements $A_1, A_2, \ldots$ se produit ? Par exemple, sous quelles conditions peut-on garantir qu'on finira par gagner à une loterie ou par obtenir un 6 à un examen ? Ou encore : on démarre une marche aléatoire à Manhattan — à chaque carrefour on choisit uniformément une des quatre directions. Reviendra-t-on un jour à son hôtel ?

— Sous quels critères peut-on assurer que seuls un nombre fini des événements $A_1, A_2, \ldots$ d'une suite se produisent ? Cela pourrait par exemple modéliser le fait qu'une maladie infectieuse n'ait qu'une propagation limitée.

— Quand peut-on dire quelque chose sur la limite d'une suite de variables aléatoires $X_1, X_2, \ldots$? En quels sens peut-on parler de convergence ? Nous avons déjà vu des affirmations un peu informelles du type : $Bin(n, \lambda/n)$ converge vers une loi de Poisson, ou bien la moyenne empirique de v.a. i.i.d. converge vers son espérance. Quelles sont les notions et les énoncés mathématiques corrects ?

5.1 Collections infinies d'événements et de variables aléatoires

Avant d'énoncer quelques théorèmes limites intéressants, commençons par formaliser certaines notions de limite dans le contexte des événements. Fixons un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et une suite d'événements $E_1, E_2, \ldots$, qui pourraient par exemple être des répétitions d'une même situation aléatoire, comme des lancers de pièce indépendants. [15]

Rappelons que dire que E_i est un événement signifie $E_i \subseteq \Omega$ et $E_i \in \mathcal{F}$. Chaque ω représente un état possible de l'univers, et $\omega \in E_i$ si l'événement E_i se produit dans cet état.

Nous pouvons alors poser :

— D'abord, on peut demander si au moins un événement de la suite $(E_n)_{n\geq 1}$ se produit. Par définition,

$$\{\omega \in \Omega : \omega \in E_i \text{ pour au moins un } i\} = \bigcup_{n\geq 1} E_n.$$

On dit parfois que "E_i se produit finalement" (eventually). Un exemple est celui d'une feuille d'exercices précédente : en lançant des pièces indépendantes, on obtient finalement pile avec probabilité 1 (cela découle aussi du lemme juste ci-dessous). Remarquez qu'il existe une suite de lancers qui ne donne jamais pile — la suite $TTTTT\ldots$ — mais comme elle a probabilité 0, cela ne compte pas.

— Ensuite, on peut demander si une infinité des événements E_i se produisent. Formulons cela : on vérifie que

$$\{\omega \in \Omega : \omega \in E_i \text{ pour une infinité de } i\} = \bigcap_{m\geq 1} \bigcup_{n\geq m} E_n.$$

15. Comme discuté, il n'est pas totalement immédiat de construire un espace de probabilité sur lequel on dispose d'une suite infinie de lancers de pièce indépendants. Ici, nous l'admettons ; voir aussi l'appendice de cette section pour une preuve (en admettant l'existence de la mesure de Lebesgue).

Cet événement est aussi noté $\limsup_{n\geq 1} E_n$. Dans le cas des lancers de pièce, E_i peut signifier : "le i-ième lancer donne pile" ; et nous avons vu que pour des pièces indépendantes, alors E_i se produit une infinité de fois avec probabilité 1.

— Enfin, on peut demander si tous les E_i sauf un nombre fini se produisent. On vérifie (sur la feuille d'exercices) que

$$\{\omega \in \Omega : \omega \in E_i \text{ pour tout } i \text{ sauf un nombre fini}\} = \bigcup_{m\geq 1} \bigcap_{n\geq m} E_n.$$

Cet événement est noté $\liminf_{n\geq 1} E_n$. Un exemple : vous commencez avec 10 CHF, et tant qu'il vous reste de l'argent, vous pariez avec la Banque centrale européenne (qui peut toujours imprimer plus d'argent si besoin !) sur le résultat de lancers de pièce indépendants. Le gagnant reçoit 1 CHF, le perdant perd 1 CHF. C'est un fait mathématique qu'avec probabilité 1, après un nombre fini de paris vous êtes ruiné (vous avez 0 CHF). Si l'on note E_i l'événement "après i paris, vous êtes ruiné", alors cet événement échoue seulement un nombre fini de fois.

Voici quelques critères utiles pour étudier ces événements. D'abord, un critère très naïf :

Lemme 5.1. *Soient $E_1, E_2, \ldots$ des événements indépendants de probabilités p_i. Alors $\mathbb{P}(\bigcup_{i\geq 1} E_i) = 1$ si et seulement si $\prod_{i=1}^{n}(1-p_i) \to 0$ lorsque $n \to \infty$.*

Démonstration. Ceci est sur la feuille d'exercices. □

Par exemple, si chaque événement se produit avec la même probabilité p, alors $\prod_{i=1}^{n}(1-p) = (1-p)^n$, qui tend clairement vers zéro. Ainsi, même si vous lancez une pièce qui donne pile avec probabilité 0,00001, vous verrez finalement pile.

Un critère très utile pour vérifier qu'un événement ne peut se produire qu'un nombre fini de fois est donné par le premier lemme de Borel-Cantelli :

Lemme 5.2 (Borel-Cantelli I)**.** *Soit $(E_n)_{n\geq 1}$ une suite quelconque d'événements sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Si $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, alors presque sûrement, seuls un nombre fini des événements E_i se produisent, i.e.*

$$\mathbb{P}\left(\bigcap_{m\geq 1} \bigcup_{n\geq m} E_n\right) = 0.$$

Remarquez que nous ne supposons rien sur la dépendance ou l'indépendance des E_i ! Et ce lemme ne dit pas qu'il existe un nombre fixe (comme 1000) d'événements qui se produisent : combien et lesquels se produisent dépend de $\omega \in \Omega$.

Par exemple, considérons une suite de pièces biaisées avec probabilité de pile au n-ième lancer égale à n^{-2}. Si E_n désigne l'événement "pile au n-ième lancer", alors $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$. Donc, par le lemme, on obtient que presque sûrement, on n'observe qu'un nombre fini de piles sur une suite infinie de lancers. Cependant, obtenir 10 piles ou même 100 piles dépend de la suite exacte des lancers, i.e. de l'aléa encodé par ω.

Démonstration. Fixons $\epsilon > 0$. Comme $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, il existe $n_0 \in \mathbb{N}$ tel que $\sum_{n\geq n_0} \mathbb{P}(E_n) < \epsilon$. Or comme $\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$,

$$\mathbb{P}\left(\bigcap_{m\geq 1} \bigcup_{n\geq m} E_n\right) \leq \mathbb{P}\left(\bigcup_{n\geq n_0} E_n\right) \leq \sum_{n\geq n_0} \mathbb{P}(E_n) < \epsilon,$$

où l'on utilise la borne par union. Comme ϵ était arbitraire, le résultat suit. □

La brièveté de la preuve peut rendre ce résultat suspect, mais nous verrons bientôt qu'il est très utile, par exemple pour obtenir des convergences de variables aléatoires.

Ceci est en partie complété par le second lemme de Borel-Cantelli, qui donne une condition pour que des événements se produisent une infinité de fois. Ici, on suppose à nouveau l'indépendance.

Lemme 5.3 (Borel-Cantelli II)**.** *Soit $(E_n)_{n\geq 1}$ une suite d'événements indépendants sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Si $\sum_{n\geq 1} \mathbb{P}(E_n) = \infty$, alors presque sûrement, une infinité des E_i se produisent, i.e.*

$$\mathbb{P}\left(\bigcap_{m\geq 1}\bigcup_{n\geq m} E_n\right) = 1.$$

Démonstration. Sur la feuille d'exercices. □

Ces lemmes ont l'air innocents, mais ils ont de jolies applications (nous en verrons). Un corollaire simple dit que des événements indépendants se produisent infiniment souvent avec probabilité 1 ou 0 — ce qui est remarquable, car a priori on pourrait croire qu'on peut obtenir n'importe quelle probabilité (comme dans l'exemple ci-dessus). On voit ainsi comment l'hypothèse d'indépendance, "simple" en apparence, peut changer radicalement les choses :

Corollaire 5.4. *Soient $E_1, E_2, \ldots$ des événements mutuellement indépendants sur un même espace de probabilité. Alors*

$$\mathbb{P}\left(\bigcap_{m\geq 1}\bigcup_{n\geq m} E_n\right) \in \{0, 1\},$$

i.e. "E_i se produit une infinité de fois" a probabilité 0 ou 1.

Démonstration. Cela découle immédiatement des lemmes de Borel-Cantelli : soit $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, soit $\sum_{n\geq 1} \mathbb{P}(E_n) = \infty$. □

En fait, c'est un cas particulier de la loi 0-1 de Kolmogorov, que nous ne verrons que dans la section non examinable cette année.

5.2 Convergence de variables aléatoires

Quand on passe des événements aux suites de variables aléatoires $X_1, X_2, \ldots$, la première question est à nouveau : quelles questions peut-on poser ?

Par exemple :

— Une valeur $\geq k$ est-elle atteinte par la suite de variables aléatoires ?

— Tous les X_i sauf un nombre fini sont-ils positifs ?

— La suite (X_i) est-elle bornée en valeur absolue ?

— Converge-t-elle ?

Pour la première question, la mesurabilité est claire, car on peut écrire l'événement comme l'union $\bigcup_{i\geq 1}\{X_i \geq k\}$, et de même pour la seconde. Pour la troisième, il faut déjà réfléchir un peu : l'événement "la suite est bornée en valeur absolue par $M \in \mathbb{N}$" est donné par $E_M := \bigcap_{i\geq 1}\{|X_i| \leq M\}$. Mais on veut autoriser des bornes différentes selon les suites. On

doit donc prendre aussi l'union sur M, ce qui donne $\bigcup_{M\in\mathbb{N}} E_M$, et cela montre encore que la question a un sens. Pour la quatrième, on énonce un lemme (facile à vérifier) :

Lemme 5.5. *Soient $X, X_1, X_2, \dots$ des variables aléatoires sur un même espace de probabilité. Montrer que les ensembles*

$$E := \{\omega : (X_i(\omega))_{i\geq 1} \text{ converge}\} \quad et \quad E_\infty := \{\omega : (X_i(\omega))_{i\geq 1} \text{ converge vers } X(\omega)\}$$

sont des événements, i.e. sont mesurables.

Démonstration. Sur la feuille d'exercices. □

5.2.1 Convergence presque sûre et loi des grands nombres

L'exercice précédent introduit aussi ce qui est peut-être la notion la plus naturelle de convergence pour des variables aléatoires définies sur un même espace de probabilité. Le cadre est le suivant : on a des variables aléatoires $X_1, X_2, \dots$ définies sur le même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, et l'on s'intéresse à l'événement $\{\omega \in \Omega : X_n(\omega) \text{ converge}\}$. Par exemple, pour les lancers de pièce : on lance la pièce 100 fois et on prend la moyenne, puis 1000 fois et on prend la moyenne. Ces moyennes convergent-elles ?

Définition 5.6 (Convergence presque sûre)**.** *Soient $X, X_1, X_2, \dots$ des variables aléatoires définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Si*

$$\mathbb{P}\big(\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}\big) = 1,$$

alors on dit que la suite $(X_n)_{n\geq 1}$ converge presque sûrement vers X.

Nous avons vu que ces ensembles sont bien mesurables. Un critère utile pour la convergence presque sûre vient de Borel-Cantelli I :

Lemme 5.7. *Soient $X, X_1, X_2, \dots$ des variables aléatoires définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. S'il existe une suite décroissante $(a_n)_{n\geq 1}$ de réels ≥ 0 tendant vers 0 telle que, pour les événements*

$$E_n := \{\omega : |X_n(\omega) - X(\omega)| > a_n\},$$

on ait $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, alors X_n converge presque sûrement vers X.

Démonstration. Par Borel-Cantelli I, presque sûrement seuls un nombre fini des E_n se produisent, i.e. $\mathbb{P}(\cap_{m\geq 1} \cup_{k\geq m} E_k) = 0$. Or

$$\{\omega : X_n(\omega) \text{ ne converge pas vers } X(\omega)\} \subseteq \cap_{m\geq 1} \cup_{k\geq m} E_k.$$

En effet, si $X_n(\omega)$ ne converge pas vers $X(\omega)$, alors il existe $\epsilon > 0$ et une sous-suite $(n_\ell)_{\ell\geq 1}$ telle que $|X_{n_\ell}(\omega) - X(\omega)| > \epsilon$. Si l'on prend k tel que $a_k < \epsilon$, alors pour tout $n_\ell > k$ on a $\omega \in E_{n_\ell}$, ce qui conclut. □

Un des exemples les plus importants de convergence presque sûre est la loi des grands nombres, dont nous avons déjà vu une version dans le Théorème 4.5.

5.2.2 Loi des grands nombres

Commençons par reformuler une version plus générale de la loi faible des grands nombres, i.e. le Théorème 4.5.

Théorème 5.8 (Loi faible des grands nombres (WLLN))**.** *Soient $X_1, X_2, \dots$ des v.a. i.i.d. intégrables définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. Alors, lorsque $n \to \infty$, pour tout $\epsilon > 0$,*

$$\mathbb{P}\Big(\Big|\frac{\sum_{i=1}^n X_i}{n} - \mathbb{E}X_1\Big| > \epsilon\Big) \to 0.$$

En fait, cette notion de convergence porte un nom : la convergence en probabilité. La définition est :

Définition 5.9 (Convergence en probabilité)**.** *Soient $X_1, X_2, \dots$ des variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. S'il existe une variable aléatoire X_0 définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ telle que, pour tout $\epsilon > 0$, lorsque $n \to \infty$,*

$$\mathbb{P}(|X_n - X_0| > \epsilon) \to 0,$$

alors on dit que $(X_n)_{n\geq 1}$ converge en probabilité vers X_0.

Comme mentionné, la preuve dans le cas de variables de variance finie est exactement comme dans la preuve du Théorème 4.5. On peut la renforcer en loi forte des grands nombres, en remplaçant la convergence en probabilité par la convergence presque sûre.

Théorème 5.10 (Loi forte des grands nombres (SLLN))**.** *Soient $X_1, X_2, \dots$ des v.a. i.i.d. intégrables définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. Alors $\frac{\sum_{i=1}^n X_i}{n}$ converge presque sûrement vers $\mathbb{E}X_1$, i.e.*

$$\mathbb{P}\Big(\frac{\sum_{i=1}^n X_i}{n} \text{ converge vers } \mathbb{E}X_1\Big) = 1.$$

Grossièrement, les deux théorèmes disent que si l'on répète indépendamment une même expérience aléatoire n fois, obtenant des v.a. i.i.d. $X_1, \dots, X_n$, alors la moyenne $\frac{1}{n}\sum_{i=1}^n X_i$ converge, lorsque $n \to \infty$, vers $\mathbb{E}X_1$.

C'est remarquable : la distribution précise des X_i ne joue presque aucun rôle dans la limite — seule l'intégrabilité (et donc l'espérance) compte. Ces théorèmes sont reliés aux théorèmes ergodiques, qui relient en gros les moyennes temporelles (ici l'indice n) aux moyennes spatiales (ici $\mathbb{E}$). Mais quelle est la différence entre ces deux résultats ?

- La loi faible dit que si vous faites des expériences indépendantes $X_1, X_2, \dots$ et regardez la moyenne empirique des n premières, alors la variable aléatoire obtenue est très proche de la constante $\mathbb{E}X_1$. En effet, pour tout $\epsilon > 0$ et tout $\delta > 0$, si vous faites assez d'expériences, la probabilité que la moyenne diffère de $\mathbb{E}X_1$ de plus de ϵ est $< \delta$. La WLLN ne dit cependant pas comment se comportent, le long d'une trajectoire fixée, les moyennes empiriques successives.
- La loi forte dit précisément qu'avec probabilité 1, pour (presque) toute suite de réalisations, si l'on regarde la moyenne des n premiers résultats et que l'on augmente n, alors ces moyennes convergent vers $\mathbb{E}X_1$. La SLLN ne regarde pas seulement des "instantanés" à n fixé : elle décrit l'évolution des moyennes le long d'une trajectoire.

Dans les deux résultats, l'intégrabilité et l'indépendance sont importantes. Vous réfléchirez au rôle de l'intégrabilité sur la feuille d'exercices ; pour voir qu'une certaine indépendance est nécessaire, pensez au cas $X_1 = X_2 = \dots$. Alors la moyenne de $X_1, \dots, X_n$ vaut X_1 et

n'a aucune raison de converger vers une constante. En général, des lois des grands nombres existent aussi sous des hypothèses de dépendance plus faibles, comme nous le commenterons dans la preuve.

Pour complétude, nous redonnons la preuve du cas particulier de la WLLN (en soulignant que c'est la même preuve que celle du Théorème 4.5).

Preuve de la WLLN pour des v.a. i.i.d. de variance bornée. Supposons que $\mathbb{E}X_1^2 < C$. Alors $\mathbb{E}(|S_n - \mathbb{E}X_1|^2)$ est bien définie et l'on peut écrire

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^2) = \sum_{i,j \leq n} n^{-2} \mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)\right].$$

Or $X_1, X_2, \ldots$ sont mutuellement indépendantes et $\mathbb{E}X_j = \mathbb{E}X_1$. Ainsi, si $i \neq j$, on a $\mathbb{E}[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)] = 0$. Donc

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^2) = n^{-2} \sum_{i=1}^{n} \mathrm{Var}(X_i) \leq n^{-1} C \to 0$$

lorsque $n \to \infty$. Par l'inégalité de Chebyshev,

$$\mathbb{P}(|S_n - \mathbb{E}X_1| > \epsilon) \leq \frac{\mathbb{E}(|S_n - \mathbb{E}X_1|^2)}{\epsilon^2} \leq \frac{C}{\epsilon^2} \frac{1}{n} \to 0,$$

et la WLLN est démontrée dans ce cas. □

Remarque : dans cette preuve, on n'utilise pas l'indépendance complète, seulement le fait que $\mathrm{Cov}(X_i, X_j) = 0$ pour $i \neq j$! De plus, on n'utilise pas que les variables sont i.i.d., seulement que $\mathbb{E}X_i^2 < C$ pour tout i, i.e. que les variances sont uniformément bornées.

Nous prouvons la SLLN sous une hypothèse encore plus forte. La preuve commence de façon similaire, puis on utilise le corollaire de Borel-Cantelli via le Lemme 5.7.

Preuve de la SLLN pour des v.a. i.i.d. avec $\mathbb{E}X_i^4 < C$. Supposons que pour une constante $C > 0$ on ait $\mathbb{E}X_i^4 < C$. En augmentant C (mais pas le nombre de notations !), on peut supposer aussi que $\mathbb{E}(X_i - \mathbb{E}X_i)^4 < C$ (pourquoi ?). Alors $\mathbb{E}(|S_n - \mathbb{E}X_1|^4)$ est bien définie et

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^4) = \sum_{i,j,k,l \leq n} n^{-4} \mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)(X_k - \mathbb{E}X_1)(X_l - \mathbb{E}X_1)\right].$$

Si un indice apparaît une seule fois (par exemple $i = 1$ et $j = k = l = 2$), alors, comme dans la preuve de la WLLN,

$$\mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)(X_k - \mathbb{E}X_1)(X_l - \mathbb{E}X_1)\right] = 0,$$

par indépendance et parce que $\mathbb{E}(X_i - \mathbb{E}X_1) = 0$. Ainsi,

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^4) = n^{-4} \sum_{i,j \leq n} \mathbb{E}\left[(X_i - \mathbb{E}X_1)^2 (X_j - \mathbb{E}X_1)^2\right].$$

Par Cauchy-Schwarz,

$$\mathbb{E}\left[(X_i - \mathbb{E}X_1)^2 (X_j - \mathbb{E}X_1)^2\right] \leq \sqrt{\mathbb{E}(X_i - \mathbb{E}X_1)^4 \, \mathbb{E}(X_j - \mathbb{E}X_1)^4} \leq C.$$

Donc

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^4) \leq C n^{-2}.$$

Appliquons maintenant le Lemme 5.7 avec $a_n = n^{-1/8}$. Par l'inégalité de Markov,

$$\begin{aligned}\mathbb{P}(E_n) &= \mathbb{P}(|S_n - \mathbb{E}X_1| > n^{-1/8}) = \mathbb{P}(|S_n - \mathbb{E}X_1|^4 > n^{-1/2}) \\ &\leq \frac{\mathbb{E}|S_n - \mathbb{E}X_1|^4}{n^{-1/2}} \leq Cn^{-3/2}.\end{aligned}$$

Ainsi $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, et le Lemme 5.7 donne la convergence presque sûre de S_n vers $\mathbb{E}X_1$. □

Remarque 5.11. *Remarquez encore une fois que dans cette preuve on n'utilise pas que les X_i soient identiquement distribuées : on utilise seulement que $\mathbb{E}X_i^4 < C$. Posez-vous la question : pourquoi a-t-on besoin ici du 4ème moment, alors que dans la WLLN le 2ème moment suffit ?*

Ces deux théorèmes sont à la base de l'approche fréquentiste des probabilités. On obtient immédiatement le corollaire suivant (rappelez-vous comme c'était pénible à démontrer sur la première feuille !) :

Corollaire 5.12. *Soient $E_1, E_2, \ldots$ des événements indépendants avec $\mathbb{P}(E_i) = p$. Alors*

$$\frac{\#\{i \leq n : E_i \text{ se produit}\}}{n}$$

converge presque sûrement vers p.

Démonstration. Ceci découle directement de la SLLN en remarquant que $1_{E_1}, 1_{E_2}, \ldots$ sont des v.a. i.i.d. intégrables d'espérance p. □

Par exemple, si vous avez une pièce dont la probabilité p de pile est inconnue, vous la lancez et regardez la proportion de piles parmi les n premiers lancers. La SLLN dit qu'avec probabilité 1 ces proportions convergent vers p. Une question naturelle est : à quelle vitesse ? i.e. à quel point connaît-on p après 25 ou 100 lancers ? Répondre précisément dépasse le cadre du cours, mais nous abordons maintenant le Théorème central limite, qui décrit les fluctuations de la moyenne autour de sa valeur moyenne.

Avant cela, mettons en perspective la relation entre loi faible et loi forte : elle est liée à la relation entre convergence presque sûre et convergence en probabilité.

Lemme 5.13. *Si $X_1, X_2, \ldots$ converge presque sûrement vers X, alors la suite converge aussi en probabilité : pour tout $\epsilon > 0$,*

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \quad \textit{lorsque } n \to \infty.$$

Démonstration. Fixons $\epsilon > 0$. Alors

$$\{(X_n)_{n\geq 1} \to X\} \subseteq \{|X_n - X| < \epsilon \text{ pour tout } n \text{ assez grand}\} = \bigcup_{m\geq 1} E_m,$$

[16] où $E_m = \{\forall n \geq m : |X_n - X| < \epsilon\}$. Ces événements sont emboîtés : $E_m \subseteq E_{m+1}$. Comme $\mathbb{P}(\{(X_n) \to X\}) = 1$, on a

$$1 = \mathbb{P}\Big(\bigcup_{m\geq 1} E_m\Big) = \lim_{m\to\infty} \mathbb{P}(E_m).$$

16. Si cette inclusion vous semble opaque, je recommande d'écrire tout explicitement avec ω, i.e. $\{\omega : (X_n(\omega)) \to X(\omega)\} \subseteq \{\omega : |X_n(\omega) - X(\omega)| < \epsilon \text{ pour tout } n \geq n(\omega)\}$, etc.

Or $\mathbb{P}(|X_n - X| > \epsilon) \leq 1 - \mathbb{P}(E_n)$, donc $\mathbb{P}(|X_n - X| > \epsilon) \to 0$. □

Dans l'autre sens, la convergence en probabilité est plus faible que la convergence presque sûre, mais elle implique une convergence presque sûre le long d'une sous-suite (potentiellement très clairsemée). Ceci repose sur le Lemme 5.7 et est exploré sur la feuille d'exercices.

5.2.3 Convergence en loi

La troisième notion importante est celle de convergence en loi. Elle concerne uniquement les lois des variables aléatoires, et s'applique donc aussi à des suites de variables aléatoires définies sur des espaces de probabilité différents. Géométriquement, on peut la voir comme la convergence des fonctions de répartition, ou (plus visuellement) celle des histogrammes. Nous l'avons déjà rencontrée, par exemple lorsque nous avons montré que $Bin(n, \lambda/n)$ converge vers $Poiss(\lambda)$, ou lorsque nous avons discuté la convergence de lois uniformes sur $[0,1] \cap n^{-1}\mathbb{Z}$ vers la loi uniforme $U_{[0,1]}$.

On peut aussi imaginer le conte suivant : votre but dans la vie est de devenir un lanceur de pièce parfait, c'est-à-dire que, pour une pièce équilibrée, vous obtenez vraiment pile avec probabilité 1/2. Au début, vos lancers sont biaisés : la pièce fait une seule rotation et retombe souvent avec la face qui était en bas au départ. Vous modélisez vos lancers par une loi $Ber(p)$ avec $p \neq 1/2$. En vous entraînant, vous vous rapprochez d'une pièce parfaite, donc de $Ber(1/2)$. À différents stades, vos lancers ont des distributions différentes, potentiellement sur des espaces de probabilité différents. Si X_n décrit vos lancers à l'année n, on espère idéalement que X_n converge vers $Ber(1/2)$ lorsque $n \to \infty$. Il semble qu'en pratique, la perfection demande un vieillissement infini.

Nous voulons donc dire que les lois $\mathbb{P}_{X_n}$ (mesures de probabilité sur $\mathbb{R}$) convergent.

Le premier réflexe serait de demander $\mathbb{P}_{X_n}(E) \to \mathbb{P}_X(E)$ pour tout borélien E. Cela définit bien une notion de convergence de mesures [17], mais elle est trop forte pour être utile en probabilité, comme le montre l'exemple suivant.

Exemple 5.14. *Considérons des variables déterministes X_n prenant la valeur $1/n$. Intuitivement, on veut dire que X_n converge vers la variable déterministe X qui vaut 0 presque sûrement. En effet, sur un même espace de probabilité, X_n converge vers X presque sûrement et en probabilité ; donc il faut aussi que X_n converge en loi vers X. Pourtant, $\mathbb{P}_{X_n}(\{0\}) = 0$ pour tout $n \geq 1$, alors que $\mathbb{P}_X(\{0\}) = 1$.*

On cherche donc une notion plus faible. Il existe plusieurs définitions équivalentes ; en voici deux classiques :

Définition 5.15 (Convergence en loi I)**.** *On dit que $X_1, X_2, \ldots$ converge en loi (ou : en distribution) vers X si, pour toute fonction continue bornée $g : \mathbb{R} \to \mathbb{R}$, on a*

$$\mathbb{E}g(X_n) \to \mathbb{E}g(X).$$

Définition 5.16 (Convergence en loi II)**.** *On dit que $X_1, X_2, \ldots$ converge en loi (ou : en distribution) vers X si*

$$F_{X_n}(t) \to F_X(t)$$

pour tout t qui est un point de continuité de F_X, i.e. tel que $\mathbb{P}(X = t) = 0$.

17. On appelle cela la convergence ensembliste (setwise) des mesures.

L'exclusion des points de discontinuité renvoie exactement à l'exemple ci-dessus. Remarquez toutefois que si la variable limite est continue, on demande bien la convergence ponctuelle de la fonction de répartition en tout point.

Soulignons aussi que, dans ces deux définitions, on ne suppose pas que $X_1, X_2, \dots$ soient définies sur un même espace de probabilité : ce n'est pas nécessaire, puisque nous ne regardons que leurs lois $\mathbb{P}_{X_i}$, déterminées par F_{X_i}.

Nous ne donnerions pas deux définitions si elles n'étaient pas équivalentes :

Proposition 5.17. *Une suite $X_1, X_2, \dots$ converge en loi vers X au sens de la Définition 5.15 si et seulement si elle converge en loi vers X au sens de la Définition 5.16.*

Démonstration. Montrons une direction (celle que nous réutiliserons plus tard). L'autre direction est non examinable.

Supposons donc que X_n converge vers X au sens de la Définition 5.15, et montrons la convergence au sens de la Définition 5.16.

Comme dans la preuve de la Proposition 4.14, on considère des approximations continues de l'indicatrice $1_{\{x \le t\}}$ par au-dessus et par en-dessous. L'approximation par au-dessus est définie par : $g_{t,\epsilon}(x) = 1$ pour $x \le t$, $g_{t,\epsilon}(x) = 0$ pour $x \ge t + \epsilon$, et $g_{t,\epsilon}$ est linéaire sur $[t, t+\epsilon]$. L'approximation par en-dessous est $f_{t,\epsilon}(x) := g_{t-\epsilon,\epsilon}(x)$.

Par hypothèse, pour tout $\epsilon > 0$, on a $\mathbb{E}(f_{t,\epsilon}(X_n)) \to \mathbb{E}(f_{t,\epsilon}(X))$ et $\mathbb{E}(g_{t,\epsilon}(X_n)) \to \mathbb{E}(g_{t,\epsilon}(X))$. De plus, par monotonie de l'espérance,

$$\mathbb{E}(f_{t,\epsilon}(X_n)) \le F_{X_n}(t) \le \mathbb{E}(g_{t,\epsilon}(X_n)).$$

En passant à lim inf et lim sup, on obtient

$$\mathbb{E}(f_{t,\epsilon}(X)) \le \liminf_n F_{X_n}(t) \le \limsup_n F_{X_n}(t) \le \mathbb{E}(g_{t,\epsilon}(X)).$$

Si t est un point de continuité de F_X, alors lorsque $\epsilon \to 0$ on a à la fois $\mathbb{E}(g_{t,\epsilon}(X)) \to F_X(t)$ et $\mathbb{E}(f_{t,\epsilon}(X)) \to F_X(t)$, ce qui force $\liminf_n F_{X_n}(t) = \limsup_n F_{X_n}(t) = F_X(t)$, i.e. $F_{X_n}(t) \to F_X(t)$. □

Le plus important exemple de convergence en loi est le théorème central limite.

5.3 Théorème central limite

Le résultat final du cours est le Théorème central limite (TCL).

Théorème 5.18 (Théorème central limite)**.** *Soient $X_1, X_2, \dots$ des v.a. i.i.d. de variance finie σ^2 définies sur un même espace de probabilité. Alors*

$$n^{-1/2} \sum_{i=1}^{n} (X_i - \mathbb{E}X_i)$$

converge en loi vers $N(0, \sigma^2)$.

C'est un résultat remarquable : en additionnant des variables indépendantes de variance finie, on obtient toujours la même distribution limite — la loi gaussienne ! C'est une des raisons pour laquelle (au moins heuristiquement) les erreurs de mesure en physique ressemblent à des gaussiennes : elles sont des sommes de petites contributions indépendantes ; ou encore pourquoi des gaussiennes apparaissent quand on regarde des caractéristiques comme les

tailles dans une population. Le phénomène suivant lequel les propriétés fines des X_i n'influencent la loi limite que via quelques paramètres (l'espérance, la variance) est parfois appelé *universalité*.

Dans le TCL, les hypothèses de variance finie et d'indépendance sont cruciales : vous verrez un exemple sur les conditions de moments sur la feuille d'exercices. Pour voir qu'en l'absence d'indépendance le TCL peut échouer, pensez au cas $X_1 = X_2 = \dots$. Alors $n^{-1/2}\sum_{i=1}^n X_i = n^{1/2}X_1$, ce qui ne converge certainement pas et n'a aucune raison d'être gaussien. Même si l'indépendance peut être un peu affaiblie, il faut une quantité substantielle "d'autonomie" pour garantir que l'effet de chaque X_i sur la somme soit négligeable.

On peut par exemple en déduire très facilement le résultat non trivial suivant :

Corollaire 5.19. *Soit $X_n \sim Bin(n,p)$. Alors*

$$\frac{X_n - np}{\sqrt{n}}$$

converge en loi vers une gaussienne de variance $\sigma^2 = p(1-p)$.

Démonstration. On peut écrire $X_n - np = \sum_{i=1}^n (Y_i - \mathbb{E}Y_i)$, où Y_i sont des v.a. i.i.d. $Ber(p)$. Alors, par le TCL,

$$\frac{X_n - np}{\sqrt{n}} = \frac{\sum_{i=1}^n (Y_i - \mathbb{E}Y_i)}{\sqrt{n}}$$

converge en loi vers une gaussienne de variance $\mathrm{Var}(Y_i) = p(1-p)$. □

De plus, si l'on considère des v.a. à valeurs ± 1, alors $\sum X_i$ est exactement le nombre de $+1$ obtenus moins le nombre de -1 obtenus. La loi des grands nombres dit que ce nombre vaut typiquement $\approx n(p-1/2)$, où p est la probabilité d'obtenir 1 ; et le TCL décrit les fluctuations autour de cette valeur.

Plusieurs aspects du résultat sont intéressants :

— Le facteur d'échelle $1/\sqrt{n}$. Il s'explique par un calcul de variance :

$$\mathrm{Var}\Big(c_n \sum_{i=1}^n (X_i - \mu)\Big) = c_n^2\, n\, \mathrm{Var}(X_1),$$

ce qui force $c_n = 1/\sqrt{n}$ si l'on espère obtenir une quantité d'ordre $O(1)$.

— Pourquoi une gaussienne ? On observe que si $X_1, X_2, \dots$ sont des gaussiennes centrées indépendantes de variance σ^2, alors $n^{-1/2}\sum_{i=1}^n X_i$ est aussi gaussienne centrée de variance σ^2 ! En fait, les gaussiennes sont les seules lois à posséder cette propriété.

Nous allons démontrer le TCL sous une hypothèse supplémentaire, à savoir $\mathbb{E}|X_i|^3 < \infty$. Il existe de nombreuses preuves du TCL, chacune mettant en lumière un aspect différent. Celle que nous suivons repose sur l'idée suivante : pour des gaussiennes, le résultat est vrai. Pour des variables générales Y_i, on va tenter de les remplacer une par une par des gaussiennes de même moyenne et variance. On fait à chaque fois une erreur, mais si l'on contrôle l'erreur cumulée, on a gagné. C'est exactement ce que nous allons faire.

5.3.1 Preuve du TCL

L'étape clé ci-dessus est encapsulée dans la proposition suivante.

Proposition 5.20 (Principe d'échange de Lindeberg)**.** *Soient $X_1, X_2, \dots$ des v.a. i.i.d. de moyenne nulle et variance* 1*, avec $\mathbb{E}|X_i|^3 < \infty$. Soit Y une gaussienne standard. Posons*

$$S_n := n^{-1/2} \sum_{i=1}^{n} X_i.$$

Alors, pour toute fonction $f : \mathbb{R} \to \mathbb{R}$ lisse dont les dérivées jusqu'à l'ordre 3 *sont uniformément bornées, on a*

$$|\mathbb{E}f(S_n) - \mathbb{E}f(Y)| \to 0 \quad \textit{lorsque } n \to \infty.$$

Avant de prouver cette proposition, voyons comment en déduire le TCL. Nous avons déjà vu que connaître $\mathbb{E}g(X)$ pour toutes les fonctions continues bornées g détermine la loi de X. En fait, cela reste vrai si l'on ne considère que des g lisses (suffisamment riches). De plus, la convergence en loi peut aussi être obtenue à partir de la convergence $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ pour toutes les fonctions g lisses et bornées, sous la condition que les dérivées soient bornées. L'idée est la même que dans la Proposition 4.14 : on approxime les indicatrices $1_{\{x \le t\}}$ par des fonctions lisses.

Lemme 5.21. *Soient $X, X_1, X_2, \dots$ des variables aléatoires. Si pour toute fonction g lisse et bornée, dont les dérivées jusqu'au* 3*e ordre sont uniformément bornées, on a $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ lorsque $n \to \infty$, alors X_n converge en loi vers X.*

Démonstration. On procède en approchant $F_X(t)$ et $F_{X_n}(t)$ par $\mathbb{E}g_t(X)$ et $\mathbb{E}g_t(X_n)$ pour des fonctions g_t bien choisies, exactement comme dans la preuve de la Proposition 5.17. Comme la limite est une gaussienne, sa fonction de répartition est continue, et il faut donc montrer la convergence pour tout $t \in \mathbb{R}$. □

Preuve du TCL. Si $\mathrm{Var}(X_i) = \sigma^2$, alors

$$\widehat{X}_i := \frac{X_i - \mathbb{E}X_i}{\sigma}$$

a moyenne 0 et variance 1. On peut donc appliquer la Proposition 5.20 et le Lemme 5.21 pour obtenir que $n^{-1/2} \sum_{i=1}^{n} \widehat{X}_i$ converge en loi vers une gaussienne standard. En multipliant par σ, on obtient le TCL. □

Il reste à prouver la proposition. Ceci est non examinable cette année (nous avons dû aller un peu vite).

Preuve du principe d'échange de Lindeberg (non examinable). Soient $Y_1, Y_2, \dots$ des gaussiennes standards i.i.d. Pour $k \ge 1$, posons

$$S_{n,k} := \frac{\sum_{i=1}^{k-1} X_i + \sum_{i=k}^{n} Y_i}{n^{1/2}}.$$

Remarquez que $S_{n,n+1} = S_n$ et $S_{n,1} = n^{-1/2} \sum_{i=1}^{n} Y_i \sim N(0,1)$. On peut donc écrire

$$f(S_n) - f(Y) = \sum_{k=1}^{n} \big(f(S_{n,k+1}) - f(S_{n,k}) \big). \tag{5.1}$$

Notre but est de contrôler chaque terme.

Introduisons aussi

$$S^0_{n,k} := \frac{\sum_{i=1}^{k-1} X_i + \sum_{i=k+1}^{n} Y_i}{n^{1/2}},$$

où le k-ième terme est omis.

Par la formule de Taylor à l'ordre 3, on peut écrire p.s.

$$f(S_{n,k+1}) = f(S^0_{n,k}) + \frac{X_k}{n^{1/2}} f'(S^0_{n,k}) + \frac{X_k^2}{2n} f''(S^0_{n,k}) + \frac{X_k^3}{6n^{3/2}} f'''(x_1),$$

où x_1 est entre $S_{n,k+1}$ et $S^0_{n,k}$; de même

$$f(S_{n,k}) = f(S^0_{n,k}) + \frac{Y_k}{n^{1/2}} f'(S^0_{n,k}) + \frac{Y_k^2}{2n} f''(S^0_{n,k}) + \frac{Y_k^3}{6n^{3/2}} f'''(x_2),$$

où x_2 est entre $S_{n,k}$ et $S^0_{n,k}$.

En prenant l'espérance, et en utilisant que X_k est indépendant de $S^0_{n,k}$, on obtient

$$\mathbb{E}f(S_{n,k+1}) = \mathbb{E}f(S^0_{n,k}) + \frac{\mathbb{E}X_k}{n^{1/2}} \mathbb{E}f'(S^0_{n,k}) + \frac{\mathbb{E}X_k^2}{2n} \mathbb{E}f''(S^0_{n,k}) + \mathbb{E}\left(\frac{X_k^3}{6n^{3/2}} f'''(x_1) \right).$$

Comme $\mathbb{E}X_k = 0$ et $\mathbb{E}X_k^2 = 1$, cela devient

$$\mathbb{E}f(S_{n,k+1}) = \mathbb{E}f(S^0_{n,k}) + \frac{1}{2n} \mathbb{E}f''(S^0_{n,k}) + E_r,$$

avec

$$|E_r| \leq \frac{1}{6n^{3/2}} \mathbb{E}\big(|X_k|^3\big) \sup_{x \in \mathbb{R}} |f'''(x)| = O(n^{-3/2}).$$

De même, comme $\mathbb{E}Y_k = 0$ et $\mathbb{E}Y_k^2 = 1$ (et $\mathbb{E}|Y_k|^3 < \infty$),

$$\mathbb{E}f(S_{n,k}) = \mathbb{E}f(S^0_{n,k}) + \frac{1}{2n} \mathbb{E}f''(S^0_{n,k}) + \widehat{E}_r,$$

avec $|\widehat{E}_r| = O(n^{-3/2})$.

Donc

$$|\mathbb{E}f(S_{n,k+1}) - \mathbb{E}f(S_{n,k})| = O(n^{-3/2}).$$

En sommant sur $k = 1, \ldots, n$ dans (5.1) et en utilisant l'inégalité triangulaire, on obtient

$$|\mathbb{E}f(S_n) - \mathbb{E}f(Y)| \leq \sum_{k=1}^{n} O(n^{-3/2}) = O(n^{-1/2}),$$

ce qui prouve la proposition. □

Je voudrais qu'il y en ait plus. . .mais c'est tout !