

Université Pierre et Marie Curie
Master MEEF 1^{ère} année
2016 - 2017

Probabilités et statistiques

Raphaël Roux *

*. raphael.roux@upmc.fr

Table des matières

0	Quelques observations	5
0.1	Fréquences empiriques	5
0.1.1	Convergence de la moyenne	5
0.1.2	Écarts à la moyenne	6
0.2	Marches aléatoires	8
0.2.1	La marche aléatoire simple	8
0.2.2	La marche aléatoire biaisée	10
0.2.3	La marche aléatoire de Cauchy	12
1	Formalisme de la théorie des probabilités	15
1.1	Introduction	15
1.2	Espaces probabilisés	16
1.3	Variables aléatoires	19
1.4	Variables aléatoires réelles	20
1.4.1	Espérance	22
1.4.2	Théorème de transfert	27
1.4.3	Variance, moments d'ordres supérieurs	27
1.4.4	Inégalité de Markov et de Bienaymé-Tchebychev	30
2	Probabilités conditionnelles - indépendance	33
2.1	Probabilité conditionnelle	33
2.1.1	Définition	33
2.1.2	Arbres de probabilité	34
2.1.3	Formule des probabilités totales et formule de Bayes	36
2.2	Indépendance d'évènements	37
2.3	Variables aléatoires indépendantes	38
3	Variables aléatoires discrètes - dénombrement	43
3.1	Espaces probabilisés dénombrables (ou finis)	43
3.1.1	Indépendance	45
3.2	Lois discrètes usuelles	45
3.2.1	Loi uniforme	45
3.2.2	Loi de Bernoulli	46
3.2.3	Loi binomiale	46

3.2.4	Loi géométrique	47
3.2.5	Loi de Poisson	48
3.3	Dénombrement	48
3.3.1	Tirages ordonnés avec remise	48
3.3.2	Tirages ordonnés sans remise	49
3.3.3	Tirages non ordonnés sans remise	49
3.3.4	Tirage non ordonné avec remise	50
4	Variables aléatoires continues	51
4.1	Variables aléatoires réelles à densité	51
4.1.1	Définitions	51
4.2	Lois à densité usuelles	53
4.2.1	Loi uniforme	53
4.2.2	Loi de Gauss ou loi normale	53
4.2.3	Loi exponentielle	54
4.2.4	Loi de Cauchy	55
5	Suites de variables aléatoires - théorèmes limites	57
5.1	Suites des variables aléatoires	57
5.1.1	Convergence de suites de variables aléatoires réelles	57
5.2	Loi des grands nombres	59
5.3	Théorème limite central	59
5.4	Chaînes de Markov	61
6	Estimation	63
6.1	Estimation d'une proportion	63
6.2	Intervalle de confiance	65
6.2.1	Par l'inégalité de Bienaymé-Tchebychev	65
6.2.2	Dans le cas Gaussien	66
6.2.3	Par le théorème limite central	66

Chapitre 0

Quelques observations

Dans ce chapitre, nous n'allons pas faire de théorie sur le calcul des probabilités, mais plutôt décrire quelques "expériences", réalisables avec un ordinateur (ou avec un dé et beaucoup de patience), qui vont illustrer les principaux résultats qui seront formalisés rigoureusement dans les prochains chapitres.

Pour être précis, nous allons étudier la sortie de programmes représentant des lancers de dés. Nous allons supposer qu'un ordinateur est capable de reproduire fidèlement les lancers d'un dé équilibré, ce qui n'est en réalité le cas qu'approximativement ^(note 1).

0.1 Fréquences empiriques

On va tout d'abord s'intéresser à la proportion des différents résultats obtenus par plusieurs lancers de dés successifs.

0.1.1 Convergence de la moyenne

Considérons une suite infinie de lancers de dés. On note $(X_n)_{n \in \mathbb{N}}$ la suite (aléatoire) des valeurs obtenues. Une réalisation de la suite $(X_n)_{n \in \mathbb{N}}$ ressemble donc à

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

Si, par exemple, on s'intéresse à la proportion de "6" parmi les résultats obtenus, on observe ceci :

6 . . . 6 6 6 6 6 6 . 6

L'intuition nous dit alors que "un dés sur six va donner le résultat 6". Pour formaliser cette affirmation, on peut s'intéresser à la proportion de "6" parmi les n premiers tirages. Définissons la suite $\mathbf{1}_{X_n=6}$ par

$$\mathbf{1}_{X_n=6} = \begin{cases} 1 & \text{si } X_n = 6, \\ 0 & \text{sinon.} \end{cases}$$

La proportion de "6" parmi les n premiers tirages est alors donnée par $\frac{1}{n} \sum_{k=1}^n \mathbf{1}_{X_k=6}$. Traçons le graphe de $n \mapsto \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{X_k=6}$ pour une suite aléatoire $(X_n)_{n \in \mathbb{N}}$ fixée. Le résultat est donné en figure 0.1.

(note 1). Un ordinateur ne sait produire que des calculs déterministes. En revanche, il existe des algorithmes produisant des suites déterministes mais *ressemblant* à des suites aléatoires.

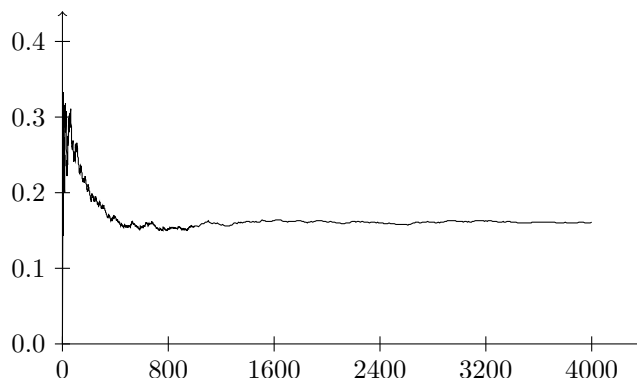


FIGURE 0.1 – Proportion de “6”, en fonction du nombre de lancers. En abscisse, le nombre de lancers, en ordonnée, la proportion de “6” (qui est dans $[0, 1]$).

À première vue, cette suite *semble* converger. Deux remarques :

- Cette convergence, si elle peut paraître intuitive, n’est *absolument pas évidente*, et il est difficile d’explicitier un argument prouvant la convergence. Sans analyse mathématiques plus poussée, on pourrait a priori croire que la suite $(\frac{1}{n} \sum_{k=1}^n \mathbf{1}_{X_k=6})_{n \in \mathbb{N}}$ oscille sans jamais se stabiliser.
- On observe ici *une* réalisation de la suite $(X_n)_{n \in \mathbb{N}}$, pour laquelle on obtient une certaine valeur limite λ . A priori, il n’est pas clair que la valeur de λ soit toujours la même. On pourrait imaginer que quelle que soit la suite $(X_n)_{n \in \mathbb{N}}$ obtenue en lançant des dés, la suite $(\frac{1}{n} \sum_{k=1}^n \mathbf{1}_{X_k=6})_{n \in \mathbb{N}}$ converge, mais que sa limite dépende de la suite de lancers de dés.

On verra plus tard que la suite est effectivement convergente, et que sa limite sera toujours la même. Ce résultat est la *loi des grands nombres*. Si l’on admet cela, on peut dans ce cas donner la valeur de la limite λ . En effet, par symétrie, les proportions de chacune des six valeurs $\{1, 2, 3, 4, 5, 6\}$ obtenues doivent toutes converger vers λ :

$$\text{pour } i = 1, \dots, 6, \quad \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{X_k=i} \xrightarrow{n \rightarrow \infty} \lambda.$$

De plus, la somme des six proportions est toujours égale à 1. Par conséquent, on a $6\lambda = 1$, d’où $\lambda = 1/6$. On retrouve l’intuition “on obtient le résultat 6 avec probabilité $1/6$ ”.

Si on affiche, pour plusieurs suites de lancers de dés, la convergence de la proportion de “6”, on obtient le graphe de la figure 0.2. On observe effectivement, dans tous les cas, la convergence vers la valeur fixée $1/6$ (ligne pointillée).

0.1.2 Écarts à la moyenne

On a remarqué dans la partie précédente que la proportion de 6 convergeait, quand le nombre de lancers tend vers l’infini, vers une valeur fixée. Toutefois, sur un nombre donné de lancers, il n’y a aucune raison que la proportion de 6 soit exactement $1/6$. On peut donc légitimement se poser la question suivante : pour un nombre n de lancers, à quelle écart peut-on s’attendre entre la valeur théorique et

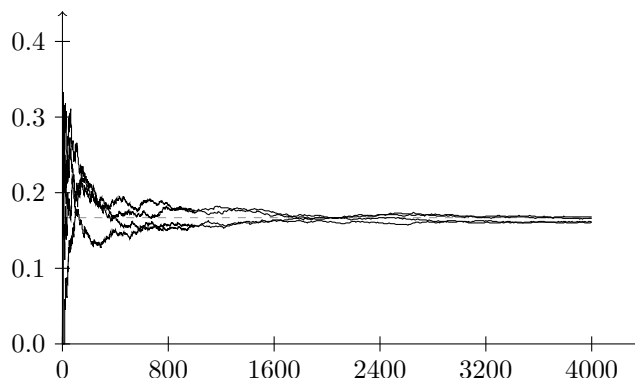


FIGURE 0.2 – Proportion de “6” sur quatre suites de lancers, avec comparaison à la valeur théorique $1/6$.

la valeur observée ? Plus précisément : la suite $\left| \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{X_k=6} - \frac{1}{6} \right|$ tend vers 0 quand $n \rightarrow \infty$, à quelle vitesse cette convergence a-t-elle lieu ?

Pour répondre à cette question, on fait l’expérience suivante : pour un certain nombre de valeurs de n (ici, $n = 8, 16, 32, 64, \dots, 4096$ — chaque nombre est le double du précédent), on calcule un grand nombre de fois l’écart $\left| \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{X_k=6} - \frac{1}{6} \right|$ (ici, 1000 fois), et on calcule la valeur moyenne de ces écarts, que l’on note ε_n . Autrement dit,

$$\varepsilon_n = \frac{1}{1000} \sum_{i=1}^{1000} \left| \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{X_k^{(i)}=6} - \frac{1}{6} \right|,$$

où pour chaque $1 \leq i \leq 1000$, $(X_n^{(i)})_{n \in \mathbb{N}}$ désigne une suite de lancers de dés. On trace alors la courbe $n \rightarrow \varepsilon_n$, et on obtient le graphique représenté en figure 0.3, à gauche.

L’erreur semble bien tendre vers 0. Toutefois, on aimerait un résultat plus quantitatif, par exemple avoir une décroissance de la forme $\varepsilon_n \simeq Cn^{-\alpha}$ pour certaines constantes C et α . Si ε_n présente une telle décroissance, on aura $\ln(\varepsilon_n) \simeq \ln C - \alpha \ln(n)$, de sorte que les couples $(\ln(n), \ln(\varepsilon_n))$ seront alignés sur la droite d’équation $y = \ln C - \alpha x$ (noter que α est le coefficient directeur de la droite). On a donc tracé sur le graphe de droite de la figure 0.3 les couples $(\ln(n), \ln(\varepsilon_n))$.

On obtient effectivement un tracé très proche d’une droite pour laquelle une régression linéaire donne l’équation $y = -1.204 - 0.501x$. En particulier, le coefficient directeur est très proche de $-1/2$, ce qui donnerait une décroissance de l’ordre de

$$\varepsilon_n \simeq \frac{C}{\sqrt{n}}.$$

On verra plus tard que l’erreur moyenne est en fait donnée par

$$\varepsilon_n \sim \sqrt{\frac{5}{18\pi n}}.$$

L’expression exacte de la constante $\sqrt{\frac{5}{18\pi}}$ n’a que peu d’intérêt. Le point important à noter ici est que l’erreur moyenne décroît comme $1/\sqrt{n}$.

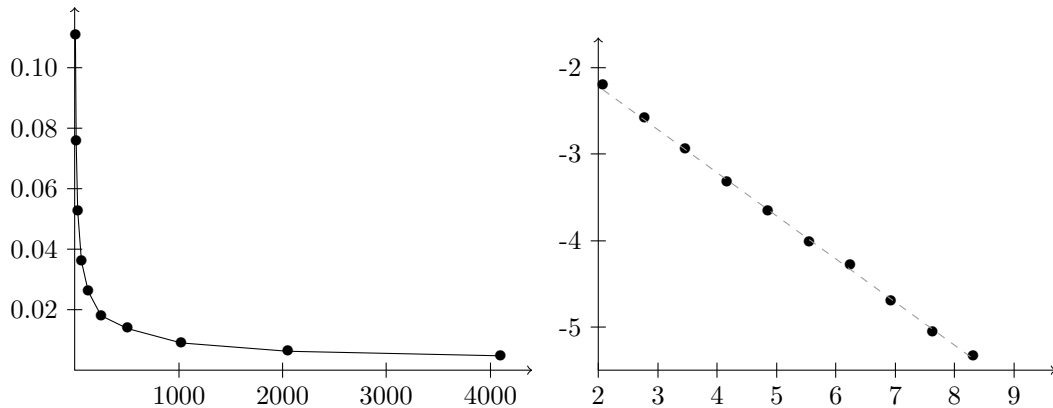


FIGURE 0.3 – Erreur moyenne ε_n entre la proportion de “6” obtenus sur n lancers et la valeur théorique $1/6$ (moyenne sur 1000 observations). À gauche : ε_n en fonction de n . À droite tracé logarithmique ($\ln(\varepsilon_n)$ en fonction de $\ln n$), avec en pointillé, la droite théorique $y = \ln\left(\sqrt{\frac{5}{18\pi}}\right) - \frac{1}{2}x$.

0.2 Marches aléatoires

Dans cette partie, on va s’intéresser à l’allure de différentes trajectoire obtenues à partir de la suite $(X_n)_{n \in \mathbb{N}}$ des lancers de dés.

0.2.1 La marche aléatoire simple

On définit ce qu’on appelle une *marche aléatoire* de la façon suivante. On définit $Z_0 = 0$, puis par récurrence,

$$Z_{n+1} = \begin{cases} Z_n + 1 & \text{si } X_n \in \{1, 2, 3\}, \\ Z_n - 1 & \text{si } X_n \in \{4, 5, 6\}. \end{cases}$$

Le variable n est à comprendre comme une variable de temps. On part à l’instant initial $n = 0$ de la position 0, puis à chaque unité de temps, on se déplace de $+1$ ou de -1 , en choisissant aléatoirement et symétriquement. Notamment, au bout de N pas, la marche se situera quelque part entre $-N$ et N . Un exemple de 20 pas d’une trajectoire est présenté en figure 0.4. En abscisse, on a représenté la variable de temps n et en ordonnée, la position Z_n . La grille grise représente l’ensemble des points qui sont atteignables par la marche. La marche part initialement de la pointe, à gauche de la grille. À chaque pas de temps, la marche se déplace vers la droite soit en montant, soit en descendant.

Une question naturelle est de savoir quelle est la position la plus probable pour la marche au bout de N pas, pour N fixé. Sur la figure 0.5, où l’on a tracé 10 réalisations des 20 premiers pas de la marche, on observe que la marche semble arriver plus souvent près de 0 que de $\pm N$. La raison en est qu’il y a beaucoup plus de trajectoires possibles qui finissent près de 0 que de trajectoires qui finissent près de $\pm N$. Par exemple, il n’existe qu’une seule trajectoire qui arrive à N (il s’agit de celle qui ne fait que monter) ou à $-N$ (celle qui ne fait que descendre).

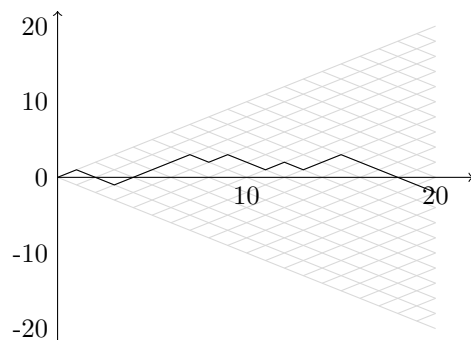


FIGURE 0.4 – Trajectoire d'une marche aléatoire.

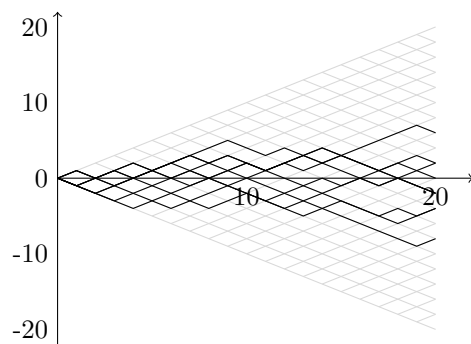


FIGURE 0.5 – Trajectoire de 10 marches aléatoires.

Analysons plus précisément ce phénomène : le tableau suivant donne la répartition des positions de 10000 marches au bout de $N = 20$ pas.

Position	-20	-18	-16	-14	-12	-10	-8	-6	-4	-2	0
Effectif	0	1	2	9	63	159	349	734	1183	1594	1752
Position	2	4	6	8	10	12	14	16	18	20	
Effectif	1606	1195	789	356	159	45	12	3	0	0	

Ce tableau est tracé sous forme d'histogramme sur la figure 0.6. On remarque d'une part que la trajectoire est beaucoup plus souvent proche de 0 que de $\pm N$, mais aussi que la probabilité de se trouver en position k semble dépendre de k de manière très régulière. Il s'agit là d'un premier aperçu du *théorème limite central*, que nous verrons au chapitre 5 (théorème 5.3.1).

On peut alors se poser la question suivante : quand n tend vers l'infini, quel est l'ordre de grandeur de Z_n ? Par exemple, existe-t-il une suite déterministe $(u_n)_{n \in \mathbb{N}}$ telle que Z_n/u_n admette une limite ? On a tracé à gauche en figure 0.7 la même chose qu'en figure 0.4, mais avec plus de pas : on est allé jusqu'à 400 pas. Par rapport au position atteignables extrêmes ± 400 , la trajectoire semble très proche de 0. Notamment, on peut s'attendre à ce que pour n grand, Z_n soit négligeable devant n .

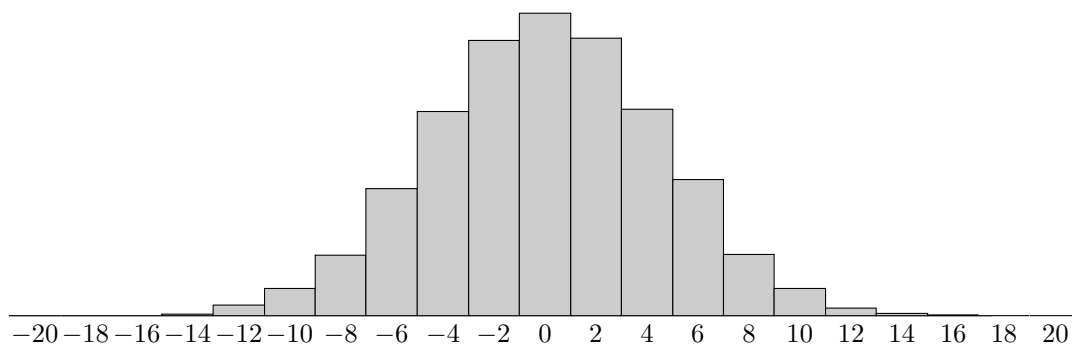
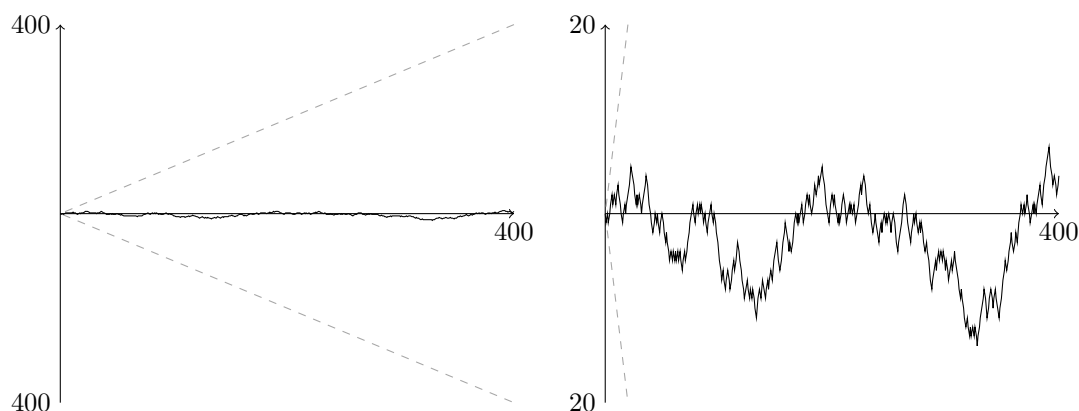


FIGURE 0.6 – Histogramme des points d'arrivée de 10000 marches aléatoires de 20 pas.

FIGURE 0.7 – Trajectoire de Z_n pour $n = 400$. En pointillés, la limite de la zone atteignable par la trajectoire.

L'explication est que pour n grand, Z_n n'est pas de l'ordre de n , mais de l'ordre de \sqrt{n} . Sur le graphe de droite de la figure 0.7, on a renormalisé les ordonnées, de sorte à afficher les positions entre $-\sqrt{n}$ et \sqrt{n} (on a $\sqrt{400} = 20$). À cette échelle-là on observe une trajectoire continue aléatoire. Autrement dit, sur une échelle de temps de l'ordre de n , les variations de Z_n sont de l'ordre de \sqrt{n} . On retiendra que pour n grand, $\sum_{k=1}^n Z_k$ a un comportement aléatoire de l'ordre de \sqrt{n} .

0.2.2 La marche aléatoire biaisée

Dans cette partie, on va étudier une suite $(Y_n)_{n \in \mathbb{N}}$ dont la définition sera légèrement différente de celle de la suite $(Z_n)_{n \in \mathbb{N}}$. Plus précisément, on va poser $Y_0 = 0$ et

$$Y_{n+1} = \begin{cases} Y_n + 1 & \text{si } X_n \in \{1, 2, 3, 4\} \\ Y_n - 1 & \text{si } X_n \in \{5, 6\}. \end{cases}$$

Autrement dit, les trajectoires possibles de $(Y_n)_{n \in \mathbb{N}}$ sont les mêmes que celles de $(X_n)_{n \in \mathbb{N}}$, mais les probabilités entre les deux directions ne sont plus symétriques : on va choisir le déplacement $+1$ deux fois plus souvent. On a tracé une trajectoire de $(Y_n)_{n \in \mathbb{N}}$ sur la figure 0.8, jusqu'à $n = 400$. On voit clairement une différence avec la figure 0.7 : la trajectoire est très proche d'une trajectoire déterministe non-nulle, qui se trouve être la suite $\frac{n}{3}$.

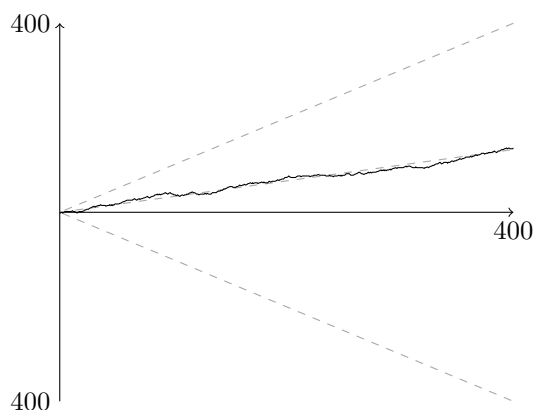


FIGURE 0.8 – Une trajectoire de la marche $(Y_n)_{n \in \mathbb{N}}$. En pointillés gris, les limites de la zone atteignable par la trajectoire, ainsi que la droite d'équation $y = x/3$.

Pour autant, la trajectoire Y_n est aléatoire, et présente des variations autour de $\frac{n}{3}$. Pour illustrer ces variations, il suffit de tracer la suite des points $Y_n - \frac{n}{3}$, ce qui est fait sur la figure 0.9.

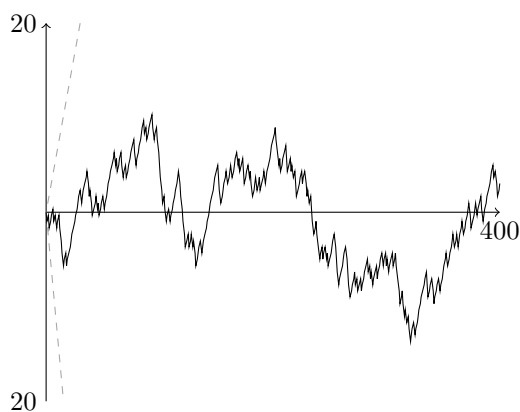


FIGURE 0.9 – Tracé de la trajectoire de $Y_n - \frac{n}{3}$.

On observe ici une trajectoire très similaire à celle de la figure 0.7. En particulier, $Y_n - \frac{n}{3}$ est de l'ordre de \sqrt{n} pour n grand. On est alors dans la situation suivante, qui se trouvera en fait être assez

générale :

$$Y_n = \left(\begin{array}{c} \text{partie déterministe} \\ \text{d'ordre } n \end{array} \right) + \left(\begin{array}{c} \text{partie aléatoire} \\ \text{d'ordre } \sqrt{n} \end{array} \right). \quad (0.1)$$

Ici, la partie déterministe est $n/3$. Pour X_n , la partie déterministe était en fait $0 \times n$, ce qui fait que le premier terme significatif était la partie aléatoire d'ordre \sqrt{n} .

0.2.3 La marche aléatoire de Cauchy

En équation (0.1), on donne une forme assez générale du comportement asymptotique des marches aléatoires. Toutefois, cette forme admet des exceptions, dont la marche que nous allons présenter dans cette partie. Cette marche est définie de la manière suivante : on pose encore $W_0 = 0$, mais W_{n+1} est construit à partir de W_n de sorte que l'angle Θ_n entre le vecteur $(1, W_{n+1} - W_n)$ et l'axe des abscisses soit choisi uniformément sur l'intervalle $[-\pi/2, \pi/2]$ (note 2). Pour être plus précis, remarquons que la tangente de l'angle entre le vecteur (x, y) et l'axe des abscisses est donné par $\frac{y}{x}$, de sorte que

$$\tan(\Theta_n) = \frac{W_{n+1} - W_n}{1}.$$

On va donc définir la suite $(W_n)_{n \in \mathbb{N}}$ par la formule

$$W_{n+1} = W_n + \tan(\Theta_n).$$

Un exemple de trajectoire obtenue est donné en figure 0.10.

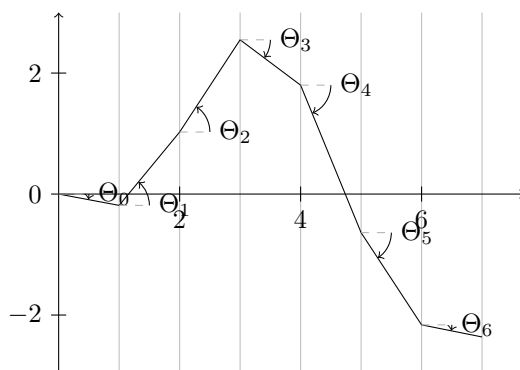
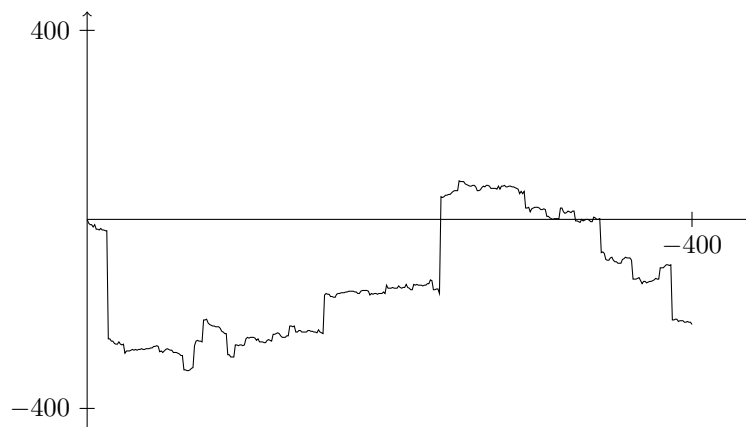


FIGURE 0.10 – Trajectoire de X_n , jusqu'à $n = 7$.

Traçons maintenant une trajectoire jusqu'à $n = 400$. Le résultat est donné sur la figure 0.11. Cette fois-ci, on observe que le comportement aléatoire de W_n est de l'ordre de n . Contrairement à la décomposition (0.1), la marche W_n présente directement de l'aléa à l'échelle n . Par ailleurs (mais on ne verra pas l'explication de ce phénomène), la forme de la trajectoire sur la figure 0.11 est différentes des trajectoires des figures 0.7 et 0.9 qui, elles, se ressemblent beaucoup. Par exemple, la trajectoire de la figure 0.11 n'est pas continue car elle présente des grands sauts.

(note 2). On n'a pas encore défini ce que signifiait "tirer un nombre uniformément dans un intervalle", et on ne peut pas construire Θ_n à partir de notre suite de lancers de dés, mais on va admettre qu'il y a un sens à "choisir une direction aléatoire sans privilégier de direction particulière".

FIGURE 0.11 – Trajectoire de X_n , jusqu'à $n = 400$.

Chapitre 1

Formalisme de la théorie des probabilités

1.1 Introduction

Le but de la théorie des probabilités est de donner un fondement mathématiques à la notion de hasard, dans le but de pouvoir définir et étudier précisément des modèles faisant intervenir de l'aléa. Le champ d'applications est ici très large : aujourd'hui des modèles probabilistes sont utilisés dans des domaines aussi variés que la physique, l'économie, la finance, la biologie ou la météorologie.

Le mot "hasard" désigne ici tout phénomène trop complexe pour être prévu précisément. De fait, il est souvent commode de considérer comme aléatoires des phénomènes qui sont pourtant parfaitement déterministes : si on lance un dé, la trajectoire du dé va entièrement être déterminée par les lois de la physique, du moins si l'on connaît très précisément la position initiale du dé, sa vitesse initiale, ainsi que la composition de l'air sur sa trajectoire et de la table sur laquelle il va tomber. Pour autant, prévoir ce mouvement exact est humainement impossible, et il est plus aisé de penser le résultat d'un dé comme étant le résultat du "hasard".

Il y a plusieurs manières de définir mathématiquement la notion de hasard, qui seront plus ou moins simples et plus ou moins puissantes. La manière qui fait consensus aujourd'hui parmi les mathématiciens est l'axiomatique de Kolmogorov, qui est basée sur la théorie de la mesure. Comme la théorie de la mesure n'est pas au programme du CAPES, nous ne passerons que rapidement sur les notions la faisant intervenir, pour nous concentrer sur la pratique du calcul des probabilités.

L'idée de l'axiomatique de Kolmogorov est la suivante : si X est une certaine quantité "aléatoire", elle peut être vue comme le résultat d'un processus faisant intervenir une source de "hasard". Formellement, on se donne donc un ensemble Ω , correspondant à tous les états possibles de la source de hasard (dans le cas du lancer de dé, Ω est donc l'ensemble des manières de lancer un dé — un ensemble très complexe !) et une fonction $X : \Omega \rightarrow E$ qui à un état de la source de hasard associe un résultat. Pour le lancer de dé, E est l'ensemble $\{1, 2, 3, 4, 5, 6\}$, et la fonction X est la fonction qui à une manière de lancer le dé associe le résultat (il s'agit également d'une fonction très complexe !).

L'axiomatique de Kolmogorov utilise aussi un autre objet, appelé *mesure de probabilité* et noté \mathbb{P} . Il s'agit d'une fonction qui à un sous-ensemble de Ω associe la probabilité que celui-ci se produise. Cet objet permet de décrire de quelle manière un élément de Ω est choisi "au hasard".

On retiendra de l'exemple du lancer de dé que l'ensemble Ω est très complexe à décrire. Pour autant, cela n'est nullement un problème, puisqu'en pratique *il est inutile de connaître l'ensemble Ω* . La seule chose importante est *la façon dont X transporte la mesure de probabilité \mathbb{P} de Ω vers E* . Cela correspond à la notion de *loi* d'une variable aléatoire.

1.2 Espaces probabilisés

Donnons maintenant des définitions plus précises.

Définition 1.2.1. On appelle espace probabilisé un couple (Ω, \mathbb{P}) tel que Ω soit un ensemble et \mathbb{P} est une application $\mathbb{P} : \mathcal{P}(\Omega) \rightarrow [0, 1]$ vérifiant ^(note 1) :

- $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$,
- si les ensembles A_n sont disjoints, alors

$$\mathbb{P}\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} \mathbb{P}(A_n). \quad (1.1)$$

On dit que \mathbb{P} est une mesure de probabilité sur Ω .

Quelques remarques :

- On a passé sous silence un point important de théorie de la mesure : dans de nombreux cas, il n'est pas possible de définir de manière cohérente une mesure \mathbb{P} sur tout $\mathcal{P}(\Omega)$. Dans ces cas-là, on devrait seulement supposer que \mathbb{P} est défini sur un sous-ensemble de $\mathcal{P}(\Omega)$ qui dispose d'une structure agréable.
On ne se préoccupera pas de cette subtilité dans ce cours, et on fera comme si toutes les probabilités étaient définies sur $\mathcal{P}(\Omega)$ entier. Cela est d'ailleurs toujours possible tant que Ω est *fini* ou *dénombrable*.
- On s'est volontairement limité aux familles $(A_n)_{n \in \mathbb{N}}$ *dénombrables*. En effet, dans le cas où Ω est indénombrable, il n'existe que très peu de fonctions de $\mathcal{P}(\Omega)$ dans $[0, 1]$ qui vérifierait (1.1) pour une famille quelconque.
- L'égalité (1.1) reste aussi vérifiée pour une famille *finie* $(A_n)_{n \in \{0, \dots, N\}}$ d'ensemble disjoints. Si on pose $A_n = \emptyset$ pour $n > N$, l'équation (1.1) se réécrit en effet

$$\mathbb{P}\left(\bigcup_{n=0}^N A_n\right) = \sum_{n=0}^N \mathbb{P}(A_n).$$

Notament, si A et B sont disjoints, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Donnons un peu de vocabulaire :

- Les sous-ensembles de Ω seront appelés les *événements* de l'espace probabilisé (Ω, \mathbb{P}) .
- Les éléments de Ω , souvent notés ω s'appellent des *issues*.

Puisque les événements sont des ensembles, on peut effectuer les opérations habituelles, avec la correspondance suivante entre les terminologies ensembliste et probabiliste.

(note 1). On note $\mathcal{P}(\Omega)$ l'ensemble des sous-ensembles de Ω .

Notation	Terminologie ensembliste	Terminologie probabiliste
Ω	ensemble entier	espace des états, évènement certain
\emptyset	ensemble vide	évènement impossible
ω	élément de Ω	issue
A	sous-ensemble de Ω	évènement
$\omega \in A$	ω appartient à A	A est réalisé si ω est le résultat de l'expérience
$A \subset B$	A est inclu dans B	si A alors B
$A \cup B$	réunion de A et B	A ou B
$A \cap B$	intersection de A et B	A et B
A^c	complémentaire de A	non A
$A \cap B = \emptyset$	A et B sont disjoints	A et B ne peuvent se produire simultanément

À partir de la seule définition 1.2.1, nous pouvons déjà donner un certain nombre de propriétés des mesures de probabilité.

Propriété 1.2.2. *Une mesure de probabilité \mathbb{P} sur Ω est une fonction croissante, au sens où si $A \subset B$, alors*

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

Démonstration. Les ensembles A et $B \setminus A$ sont disjoints, et l'inclusion $A \subset B$ donne l'égalité

$$A \cup (B \setminus A) = B.$$

On a donc $\mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(B)$. Comme \mathbb{P} est une mesure de probabilité, on a $\mathbb{P}(B \setminus A) \geq 0$, d'où $\mathbb{P}(A) \leq \mathbb{P}(B)$. \square

Propriété 1.2.3. *Soit (Ω, \mathbb{P}) un espace probabilisé. Si A et B sont deux évènements, alors :*

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;
- $\mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) = \mathbb{P}(A \cup B) - \mathbb{P}(A \cap B)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;

Démonstration. Pour le premier point, le fait que A et A^c soient disjoints montre que $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(A \cup A^c) = \mathbb{P}(\Omega) = 1$, d'où le résultat.

Pour le deuxième point, on utilise l'écriture

$$A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$$

où les trois ensembles $A \setminus B$, $A \cap B$ et $B \setminus A$ sont disjoints, de sorte que

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

Pour le troisième point, on remarque que $A = (A \setminus B) \cup (A \cap B)$, où l'union est disjointe, de sorte que $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$ (et de même $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$). On conclut en utilisant le deuxième point. \square

Propriété 1.2.4. *Une mesure de probabilité vérifie les relations suivantes :*

- (Continuité "par en dessous") *Si $B_n \subset B_{n+1}$ pour tout n , alors*

$$\mathbb{P}\left(\bigcup_{n \in \mathbb{N}} B_n\right) = \lim_n \mathbb{P}(B_n). \quad (1.2)$$

— (Continuité “par au dessus”) Si $C_{n+1} \subset C_n$ pour tout n , alors

$$\mathbb{P}\left(\bigcap_{n \in \mathbb{N}} C_n\right) = \lim_n \mathbb{P}(C_n). \quad (1.3)$$

En fait, si l’on suppose que pour des ensembles disjoints A et B on a $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$, alors les trois propriétés (1.1), (1.2) et (1.3) sont équivalentes, de sorte que n’importe laquelle des trois pourrait être utilisée comme définition d’une mesure de probabilité.

Démonstration. Montrons la propriété de continuité “par en dessous”. On considère donc une suite $(B_n)_{n \in \mathbb{N}}$ vérifiant $B_n \subset B_{n+1}$. Posons

$$A_n = \begin{cases} B_0 & \text{si } n = 0 \\ B_n \setminus B_{n-1} & \text{si } n \geq 1. \end{cases}$$

Cette définition est illustrée sur la figure 1.1.

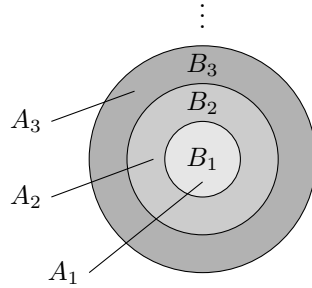


FIGURE 1.1 – Une vision schématique d’une suite croissante d’évènements $(B_n)_{n \geq 1}$. Les évènements $(A_n)_{n \geq 1}$ sont les anneaux successifs.

Avec ces notations, les ensembles A_n sont disjoints et vérifient $\bigcup_{n \in \mathbb{N}} B_n = \bigcup_{n \in \mathbb{N}} A_n$. Par conséquent, on a

$$\mathbb{P}\left(\bigcup_{n \in \mathbb{N}} B_n\right) = \mathbb{P}\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} \mathbb{P}(A_n) = \mathbb{P}(B_0) + \sum_{n=1} \mathbb{P}(B_n \setminus B_{n-1}).$$

Or $B_{n-1} \subset B_n$, de sorte que $\mathbb{P}(B_n \setminus B_{n-1}) = \mathbb{P}(B_n) - \mathbb{P}(B_{n-1})$. Par conséquent,

$$\mathbb{P}(B_0) + \sum_{n=1} \mathbb{P}(B_n \setminus B_{n-1}) = \mathbb{P}(B_0) + \lim_N \sum_{n=1}^N \mathbb{P}(B_n) - \mathbb{P}(B_{n-1}) = \mathbb{P}(B_0) + \lim_N \mathbb{P}(B_N) - \mathbb{P}(B_0) = \lim_N \mathbb{P}(B_N),$$

d’où le résultat.

La continuité “par au dessus” se déduit en passant au complémentaire. En effet, si $(C_n)_{n \in \mathbb{N}}$ est une suite décroissante d’ensembles, la suite $(C_n^c)_{n \in \mathbb{N}}$ est une suite croissante, de sorte que

$$\mathbb{P}\left(\bigcup_{n \in \mathbb{N}} C_n\right) = 1 - \mathbb{P}\left(\left(\bigcup_{n \in \mathbb{N}} C_n\right)^c\right) = 1 - \mathbb{P}\left(\bigcap_{n \in \mathbb{N}} (C_n^c)\right) = 1 - \lim_n \mathbb{P}(C_n^c) = \lim_n \mathbb{P}(C_n).$$

□

1.3 Variables aléatoires

Définition 1.3.1. Soit (Ω, \mathbb{P}) un espace probabilisé et E un ensemble. On appelle variable aléatoire à valeurs dans E une fonction de Ω dans E .

Comme on l'a déjà dit, la description exacte de l'ensemble Ω est en générale peu pertinente. En fait, les évènements "utiles" vont être ceux qui peuvent se décrire à partir des variables aléatoires définies sur Ω . On va introduire une notation commode pour manipuler ces évènements.

Définition 1.3.2. Soit $X : \Omega \rightarrow E$ une variable aléatoire, A un sous ensemble de E et x un élément de E . On note alors^(note 2)

$$\begin{aligned}\{X \in A\} &= X^{-1}(A) = \{\omega \in \Omega, X(\omega) \in A\}, \\ \{X = x\} &= X^{-1}(\{x\}) = \{\omega \in \Omega, X(\omega) = x\}.\end{aligned}$$

De manière générale, si \mathcal{P} est une certaine propriété, on notera

$$\{X \text{ vérifie } \mathcal{P}\} = \{\omega \in \Omega, X(\omega) \text{ vérifie } \mathcal{P}\}.$$

Par exemple, si $E = \mathbb{R}$, on utilisera souvent la notation

$$\{X \leq x\} = X^{-1}(]-\infty, x]) = \{\omega \in \Omega, X(\omega) \leq x\}$$

ainsi que les notations similaires $\{X < x\}$, $\{x \leq X \leq y\}$, etc.

La notation introduite en définition 1.3.2 permet de manipuler des évènements qui ont une interprétation pour le problème que l'on considère. En effet, si $X : \Omega \rightarrow E$ est une variable aléatoire, on n'a pas besoin de pouvoir représenter précisément l'ensemble Ω , mais seulement les évènements qui ont une signification par rapport à X : il s'agit des éléments de la forme $\{X \in A\}$.

Propriété 1.3.3. Si (Ω, \mathbb{P}) est un espace probabilisé, E est un ensemble et X est une variable aléatoire à valeurs dans E , alors la variable X induit une mesure de probabilité μ_X sur l'ensemble E , définie par

$$\mu_X(A) = \mathbb{P}(X \in A).$$

La mesure de probabilité μ_X est appelée loi de X .

Pour ne pas alourdir l'écriture, on a noté $\mathbb{P}(X \in A)$ pour $\mathbb{P}(\{X \in A\})$.

Démonstration. On a $\{X \in \emptyset\} = \emptyset$ et $\{X \in \Omega\} = E$. Par conséquent, on a bien

$$\mu_X(\emptyset) = \mathbb{P}(\emptyset) = 0 \text{ et } \mu_X(E) = \mathbb{P}(\Omega) = 1.$$

Ensuite, si $(A_n)_{n \in \mathbb{N}}$ est une suite de sous-ensembles disjoints de E , alors les ensembles $B_n = \{X \in A_n\}$ sont eux-aussi disjoints, d'où

$$\mu_X \left(\bigcup_{n \in \mathbb{N}} A_n \right) = \mathbb{P} \left(\bigcup_{n \in \mathbb{N}} B_n \right) = \sum_{n \in \mathbb{N}} \mathbb{P}(B_n) = \sum_{n \in \mathbb{N}} \mu_X(A_n).$$

□

(note 2). On rappelle que la notation $X^{-1}(A)$ désigne l'image réciproque de la partie A par l'application X . Cela ne sous-entend pas que X est bijective!

L'idée de la notion de loi de probabilité est la suivante : la mesure \mathbb{P} permet de "choisir au hasard" un ω dans Ω , c'est-à-dire, un état possible de la source de hasard. La variable aléatoire X associe à ce ω une observation $X(\omega) \in E$. D'une certaine manière, on a donc choisi, par le biais de X , un élément aléatoire de E . La mesure μ_X représente la manière dont cet élément aléatoire est tiré.

La plupart du temps, quand on manipule des variables aléatoires, les seules hypothèses que l'on fera concerneront la loi de X . Nos résultats commenceront donc par "Soit X une variable aléatoire de loi μ ". Pour que tout cela ait un sens, on doit s'assurer que, étant donnée une mesure de probabilité μ sur un ensemble E , il existe une variable aléatoire dont la loi est μ . C'est ce que l'on énonce dans le résultat suivant.

Propriété 1.3.4. *Soit E un ensemble et μ une mesure de probabilité sur E . Alors, il existe un espace probabilisé (Ω, \mathbb{P}) sur lequel est définie une variable aléatoire X de loi μ .*

Démonstration. Il suffit de choisir $\Omega = E$, $\mathbb{P} = \mu$ et de prendre pour X la fonction identité. Dans ce cas, (Ω, \mathbb{P}) est bien un espace probabilisé et X est une variable aléatoire.

Par définition la loi de X est la mesure μ_X définie pour $A \subset X$ par $\mu_X(A) = \mathbb{P}(X^{-1}(A))$. Or, avec notre choix de \mathbb{P} et X , on a $\mathbb{P} = \mu$ et $X^{-1}(A) = A$. On a donc bien $\mu_X(A) = \mu(A)$. \square

1.4 Variables aléatoires réelles

On appellera variable aléatoire *réelle* toute variable aléatoire à valeurs dans \mathbb{R} .

Comme l'outil fondamental en théorie des probabilités est la mesure de probabilité, il est utile d'avoir des manières de caractériser une mesure. Plus précisément, si X et Y sont deux variables aléatoires, y a-t-il une façon économique de montrer qu'elles ont la même loi, c'est-à-dire de montrer $\mu_X(A) = \mu_Y(A)$ pour tout $A \subset E$? Dans le cas où les variables sont à valeurs dans \mathbb{R} , le théorème suivant répond à cette question :

Théorème 1.4.1. *Soient μ et ν deux mesures de probabilité sur \mathbb{R} . Si pour tout $t \in \mathbb{R}$ on a*

$$\mu(]-\infty, t]) = \nu(]-\infty, t]),$$

alors $\mu = \nu$.

Démonstration. Admis. On voudrait montrer que $\mu(A) = \nu(A)$ pour tout A . L'idée serait donc de montrer que tout sous-ensemble de \mathbb{R} peut se "décrire" à partir d'intervalles. \square

On est donc amenés à la définition suivante :

Définition 1.4.2. *Soit X une variable aléatoire de loi μ_X à valeurs dans \mathbb{R} . On appelle fonction de répartition de la loi μ_X ou encore, par abus, fonction de répartition de X , l'application F_X définie sur \mathbb{R} par :*

$$\forall t \in \mathbb{R}, F_X(t) = \mu_X(]-\infty, t]) = \mathbb{P}(X \leq t).$$

Avec la définition précédente, le théorème 1.4.1 revient à dire que si deux variables aléatoires X et Y ont la même fonction de répartition, alors elles ont la même loi.

Propriété 1.4.3. *Une fonction de répartition F_X satisfait les propriétés suivantes :*

— F_X est croissante ;

- F_X est continue à droite et admet une limite à gauche en tout point ;
- $\lim_{t \rightarrow -\infty} F_X(t) = 0$, $\lim_{t \rightarrow +\infty} F_X(t) = 1$.

Démonstration. Le premier point découle de la propriété de croissance des probabilités : si $s \leq t$, alors $] -\infty, s] \subset] -\infty, t]$, d'où $\mu_X(] -\infty, s]) \leq \mu_X(] -\infty, t])$.

Étant donné que F_X est croissante, pour montrer le deuxième point il suffit de montrer pour tout $t \in \mathbb{R}$ que $F_X(t + 1/n)$ tend vers $F_X(t)$ et que $F_X(t - 1/n)$ converge. Pour tout $n \geq 1$, on définit

$$A_n = \left] -\infty, t + \frac{1}{n} \right] \text{ et } B_n = \left] -\infty, t - \frac{1}{n} \right].$$

Alors $(A_n)_{n \geq 1}$ est une suite décroissante vérifiant $\bigcap_{n \geq 1} A_n =] -\infty, t]$, et la suite $(B_n)_{n \geq 1}$ est croissante avec $\bigcup_{n \geq 1} B_n =] -\infty, t[$. Donc, d'après la propriété 1.2.4, on sait que :

$$\lim_{n \rightarrow +\infty} \mu_X(A_n) = \mu_X(] -\infty, t]) \text{ et } \lim_{n \rightarrow +\infty} \mu_X(B_n) = \mu_X(] -\infty, t[),$$

d'où le résultat.

De même, pour démontrer le dernier point il suffit de montrer

$$\lim_{n \rightarrow \infty} F_X(-n) = 0 \text{ et } \lim_{n \rightarrow \infty} F_X(n) = 1.$$

La preuve est similaire à celle du troisième point, en utilisant les ensembles $C_n =] -\infty, -n]$ et $D_n =] -\infty, n]$. □

En fait, les propriétés listées dans la propriété 1.4.3 caractérisent les fonctions de répartition :

Théorème 1.4.4. *Toute application F de \mathbb{R} dans $[0, 1]$ vérifiant les trois propriétés listées dans la propriété 1.4.3 est la fonction de répartition d'une loi de probabilité sur \mathbb{R} .*

En vertu du théorème 1.4.1, la mesure de probabilité μ dont F est la fonction de répartition est unique.

Démonstration. Admis. Il s'agirait de construire une mesure de probabilité sur $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$. Or, on a déjà évoqué le fait que construire explicitement une mesure de probabilité sur un ensemble indénombrable (ici \mathbb{R}) pouvait être très complexe. □

Un exemple de fonction de répartition est donné sur la figure 1.2. D'après le théorème 1.4.1, la fonction F caractérise entièrement la mesure μ dont elle est la fonction de répartition. Il est donc en théorie possible de lire n'importe quelle valeur $\mu(A)$ sur le graphique de la figure 1.2. En pratique, certaines de ces valeurs sont très simples à exprimer : il s'agit du cas où A est un intervalle, pour lequel $\mu(A)$ peut s'exprimer comme un accroissement de F .

Propriété 1.4.5. *Soit (Ω, \mathbb{P}) un espace probabilisé, et X une variable aléatoire sur Ω de fonction de répartition F_X , alors :*

- $\mathbb{P}(X > x) = 1 - F_X(x)$,
- $\mathbb{P}(x < X \leq y) = F_X(y) - F_X(x)$,
- $\mathbb{P}(X < x) = F_X(x-)$.
- $\mathbb{P}(X = x) = F_X(x) - F_X(x-)$.

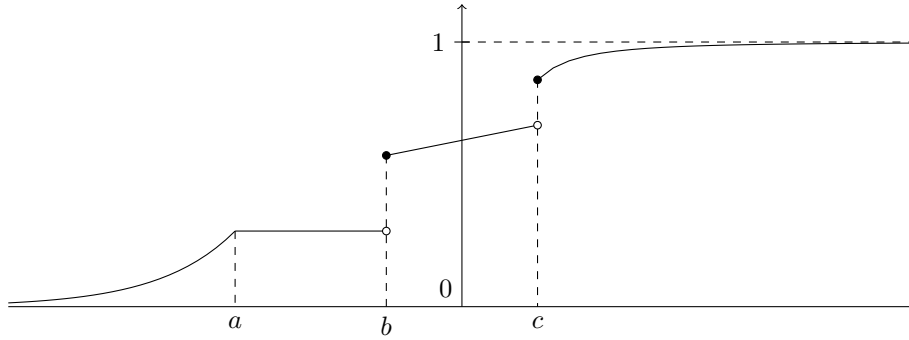


FIGURE 1.2 – Un exemple de fonction vérifiant les conditions de la propriété 1.4.3. Le théorème 1.4.1 assure qu’il existe une mesure μ dont c’est la fonction de répartition. Les valeurs de la fonction en b et c sont données par les points noirs (continuité à droite). Les points blancs représentent les limites à gauche en b et en c .

La notation $F_X(x-)$ désigne la limite à gauche de la fonction F_X au point x , dont l’existence a été donnée dans la propriété 1.4.3.

De la propriété 1.4.5, on peut retenir que les points de discontinuité de F ^(note 3) sont les x tels que $\mathbb{P}(X = x) > 0$. Dans ce cas, la valeur $\mathbb{P}(X = x)$ vaut $F(x) - F(x-)$: il s’agit de la taille du saut dans la fonction de répartition. Inversement, un “plat” dans la fonction de répartition correspond à un intervalle I tel que $\mathbb{P}(X \in I) = 0$. Par exemple, on voit que la mesure μ associée à la fonction de répartition tracée en figure 1.2 vérifie $\mu(\{b\}) > 0$ et $\mu(\{c\}) > 0$ (la fonction n’est pas continue en b et en c , ainsi que $\mu(]a, b]) = 0$ (la fonction est constante sur $]a, b]$).

1.4.1 Espérance

On veut donner un sens à la notion de valeur moyenne d’une variable aléatoire Y . Pour désigner cette valeur moyenne, on parlera de l’*espérance* de Y . Pour le cas d’une variable Y prenant un nombre fini de valeurs $\{y_1, \dots, y_k\}$, la définition est simple à donner. En effet, une telle variable Y peut s’écrire

$$Y = \sum_{k=1}^n y_k \mathbf{1}_{A_k}, \quad (1.4)$$

où l’on a noté

$$\forall \omega \in \Omega, \mathbf{1}_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A, \\ 0 & \text{sinon.} \end{cases}$$

La variable aléatoire $\mathbf{1}_A$ s’appelle l’*indicatrice* de A . Il est ensuite naturel de définir l’espérance de Y par la formule

$$\mathbb{E}Y = \sum_{k=1}^n y_k \mathbb{P}(A_k). \quad (1.5)$$

(note 3). Qui sont les x tels que $F(x-) \neq F(x)$, puisque F est continue à droite et admet une limite à gauche en tout point.

En effet, on fait la moyenne des valeurs prises par Y , pondérées par les probabilités de voir apparaître ces valeurs. Cette définition n'est pas complètement rigoureuse, car une fonction peut s'écrire de plusieurs manières sous la forme (1.4) (par exemple, pour A et B disjoints, on a $\mathbf{1}_{A \cup B} = \mathbf{1}_A + \mathbf{1}_B$) et il faudrait vérifier que l'expression (1.5) ne dépend pas de la décomposition que l'on a choisie.

L'équation (1.5) est illustrée sur la figure 1.3. On a représenté Ω comme étant l'axe des abscisses, et une fonction de Ω dans \mathbb{R} qui prend un nombre fini de valeurs peut se voir comme une fonction "en escalier". Dans ce cas, l'espérance de X est schématiquement "l'aire sous la courbe" : on ajoute les $y_i \times \mathbb{P}(A_i)$, où y_i est sur le dessin la hauteur du i ème rectangle, et $\mathbb{P}(A_i)$ sa largeur.

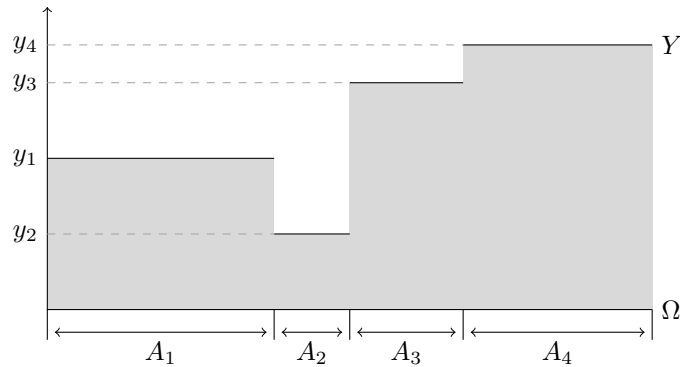


FIGURE 1.3 – Espérance d'une variable étagée.

On voudrait maintenant étendre la notion d'espérance à une classe plus large de variables aléatoires.

Le cas des variables positives

On remarque que l'espérance telle qu'on l'a définie est croissante, au sens où si deux variables aléatoires X et Y de la forme (1.4) vérifient $X \leq Y$, on a alors $\mathbb{E}X \leq \mathbb{E}Y$. En effet, on peut alors écrire

$$X = \sum_{i=1}^n x_i \mathbf{1}_{A_i} \text{ et } Y = \sum_{i=1}^n y_i \mathbf{1}_{A_i},$$

avec $x_i \leq y_i$, et on a bien

$$\mathbb{E}X = \sum_{i=1}^n x_i \mathbb{P}(A_i) \leq \sum_{i=1}^n y_i \mathbb{P}(A_i) = \mathbb{E}Y.$$

On aimerait donner un sens à $\mathbb{E}X$ pour plus de variables aléatoires X . Il semble normal de vouloir préserver la propriété de croissance. Par conséquent, on veut donner à $\mathbb{E}X$ un sens tel que, si Y est une variable de la forme (1.4) avec $Y \leq X$, alors $\mathbb{E}Y \leq \mathbb{E}X$. En particulier, on devrait avoir :

$$\sup_{Y \leq X} \mathbb{E}Y \leq \mathbb{E}X,$$

où la borne supérieure est prise sur l'ensembles des Y de la forme (1.4) avec $Y \leq X$. C'est cette formule qui va nous servir à *définir* de manière plus large l'espérance d'une variable aléatoire.

Définition 1.4.6. Soit X une variable aléatoire à valeurs dans $[0, \infty[$. On appelle espérance de X la valeur

$$\mathbb{E}X = \sup_{0 \leq Y \leq X} \mathbb{E}Y,$$

où la borne supérieure est prise sur l'ensembles des variables aléatoires Y vérifiant $0 \leq Y \leq X$ et prenant un nombre fini de valeurs. Pour un tel Y , $\mathbb{E}Y$ est défini par l'équation (1.5).

Quelques remarques :

- L'ensemble sur lequel on prend la borne supérieure est non-vide puisqu'il contient la fonction $Y = 0$. C'est pour cela que l'on a supposé X positive. Par conséquent, la borne supérieure est soit un réel, soit ∞ . Par ailleurs, pour Y vérifiant $0 \leq Y \leq X$, on a $\mathbb{E}Y \geq 0$, de sorte que dans le cas où la borne supérieure est un réel, elle est positive.
- On retiendra que l'espérance d'une variable *positive* a toujours un sens, même si on peut avoir $\mathbb{E}X = \infty$. En particulier, quelle que soit la variable aléatoire réelle X l'expression $\mathbb{E}|X|$ est bien définie.
- Graphiquement, la définition 1.4.6 est illustrée sur la figure 1.4. Comme sur la figure 1.3, on a placé schématiquement l'espace Ω sur l'axe des abscisses. L'espérance de X s'interpréterait comme l'aire sous le graphe de X . On a représenté trois variables aléatoires Y_1, Y_2, Y_3 vérifiant les condition de la définition 1.4.6. L'aire sous le graphe de X s'obtient en prenant des Y de plus en plus "proches" de X par en-dessous.

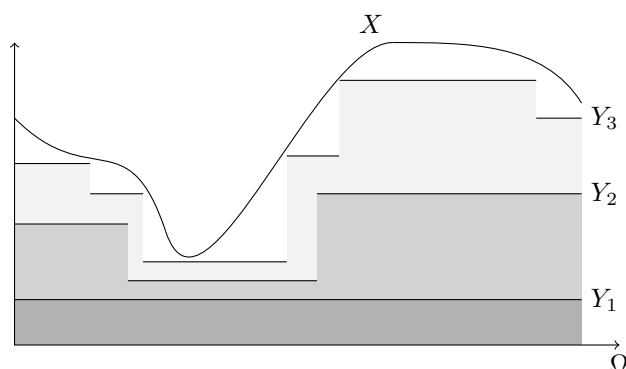


FIGURE 1.4 – Une variable aléatoire X , et plusieurs variables aléatoires Y_i inférieures à X et prenant un nombre fini de valeurs. L'espérance de X est la borne supérieure des $\mathbb{E}Y$ pour de tels Y .

L'espérance vérifie les propriétés suivantes de monotonie et de linéarité.

Propriété 1.4.7. Si X et \tilde{X} sont deux variables aléatoires positives avec $X \leq \tilde{X}$, alors $\mathbb{E}X \leq \mathbb{E}\tilde{X}$.

Démonstration. Si $0 \leq X \leq \tilde{X}$, alors toute variable aléatoire Y vérifiant $0 \leq Y \leq X$ vérifiera aussi $0 \leq Y \leq \tilde{X}$. Par conséquent, la borne supérieure apparaissant dans la définition de $\mathbb{E}\tilde{X}$ portera sur un ensemble plus grand que celle qui apparaît dans la définition de $\mathbb{E}X$. Cette borne supérieure sera donc plus grande. \square

Propriété 1.4.8. Soient X et \tilde{X} deux variables aléatoires positives et $a > 0$. On a

1. $\mathbb{E}(X + \tilde{X}) = \mathbb{E}X + \mathbb{E}\tilde{X}$, avec la convention $x + \infty = \infty$ pour tout x de $[0, \infty]$;
2. $\mathbb{E}(aX) = a\mathbb{E}X$, avec la convention $a \times \infty = \infty$.

Démonstration. Soient Y et \tilde{Y} deux variables de la forme (1.4) avec $Y \leq X$ et $\tilde{Y} \leq \tilde{X}$. Comme $Y + \tilde{Y} \leq X + \tilde{X}$, on a

$$\mathbb{E}Y + \mathbb{E}\tilde{Y} = \mathbb{E}(Y + \tilde{Y}) \leq \mathbb{E}(X + \tilde{X}).$$

En passant à la borne supérieure en Y et \tilde{Y} , on obtient

$$\mathbb{E}X + \mathbb{E}\tilde{X} \leq \mathbb{E}(X + \tilde{X}).$$

Pour obtenir l'inégalité inverse, on considère une variable Z de la forme (1.4) avec $Z \leq X + \tilde{X}$ et $\mathbb{E}Z \geq (1 - \frac{1}{n})\mathbb{E}(X + \tilde{X})$. Par définition, on a $\mathbb{E}Z \leq \mathbb{E}(X + \tilde{X})$. On peut alors construire deux variables aléatoires Y et \tilde{Y} prenant un nombre fini de valeurs et vérifiant $Y \leq X$, $\tilde{Y} \leq \tilde{X}$ et $Y + \tilde{Y} = (1 - \frac{1}{n})Z$. Par exemple, il suffit d'écrire

$$Y = Z \sum_{k=0}^{n-1} \frac{k}{n} \mathbf{1}_{\{X \in [\frac{k}{n}Z, \frac{k+1}{n}Z]\}} + Z \left(1 - \frac{1}{n}\right) \mathbf{1}_{\{X \geq Z\}}, \text{ ainsi que } \tilde{Y} = Z \left(1 - \frac{1}{n}\right) - Y.$$

On a donc

$$\left(1 - \frac{1}{n}\right)^2 \mathbb{E}(X + \tilde{X}) \leq \left(1 - \frac{1}{n}\right) \mathbb{E}Z = \mathbb{E}(Y + \tilde{Y}) = \mathbb{E}Y + \mathbb{E}\tilde{Y} \leq \mathbb{E}X + \mathbb{E}\tilde{X}.$$

Cette inégalité étant valable pour un n arbitraire, on a donc $\mathbb{E}(X + \tilde{X}) \leq \mathbb{E}X + \mathbb{E}\tilde{X}$.

Pour le deuxième point, on écrit

$$\begin{aligned} \mathbb{E}(aX) &= \sup_{0 \leq Y \leq aX} \mathbb{E}Y = \sup_{0 \leq \frac{Y}{a} \leq X} \mathbb{E}Y = \sup_{0 \leq Z \leq X} \mathbb{E}(aZ) \\ &= a \times \sup_{0 \leq Z \leq X} \mathbb{E}Z \\ &= a\mathbb{E}X. \end{aligned}$$

□

Le cas des variables intégrables

On va maintenant étendre la notion d'espérance à des variables qui ne sont pas nécessairement positives. On ne pourra pas forcément donner un sens à l'espérance de n'importe quelle variable aléatoire. Une des raisons est qu'il est possible de trouver des variables aléatoires X et Y positives avec $\mathbb{E}X = \mathbb{E}Y = \infty$ pour lesquelles $\mathbb{E}(X - Y)$ n'a pas de sens : en effet, on est en présence d'une forme indéterminée $\infty - \infty$.

Définition 1.4.9. Soit X une variable aléatoire réelle. On dit que X est intégrable si $\mathbb{E}|X| < \infty$. Dans ce cas on appelle espérance de X la quantité

$$\mathbb{E}X = \mathbb{E}(X\mathbf{1}_{\{X \geq 0\}}) - \mathbb{E}(-X\mathbf{1}_{\{X < 0\}}). \quad (1.6)$$

Une variable aléatoire d'espérance nulle est dite centrée.

Cette définition a bien un sens. En effet, d'une part, la variable $|X|$ est positive, de sorte que $\mathbb{E}|X|$ est bien définie. D'autre part, les deux variables $X\mathbf{1}_{\{X \geq 0\}}$ et $-X\mathbf{1}_{\{X < 0\}}$ sont positives et sont majorées par $|X|$, de sorte que

$$\mathbb{E}(X\mathbf{1}_{\{X > 0\}}) \leq \mathbb{E}|X| < \infty \text{ et } \mathbb{E}(-X\mathbf{1}_{\{X < 0\}}) \leq \mathbb{E}|X| < \infty,$$

ce qui fait que l'on n'est pas en présence d'une forme indéterminée $\infty - \infty$. Enfin, on remarquera que pour X positive, les deux définitions se rejoignent. La forme de l'expression (1.6) est motivée par l'égalité

$$X = X(\mathbf{1}_{\{X \geq 0\}} + \mathbf{1}_{\{X < 0\}}) = X\mathbf{1}_{\{X \geq 0\}} - (-X)\mathbf{1}_{\{X < 0\}}.$$

Graphiquement, sur la figure 1.5, l'équation (1.6) signifie que l'espérance de la variable aléatoire X correspond à la surface \mathcal{A}_+ moins la surface \mathcal{A}_- .

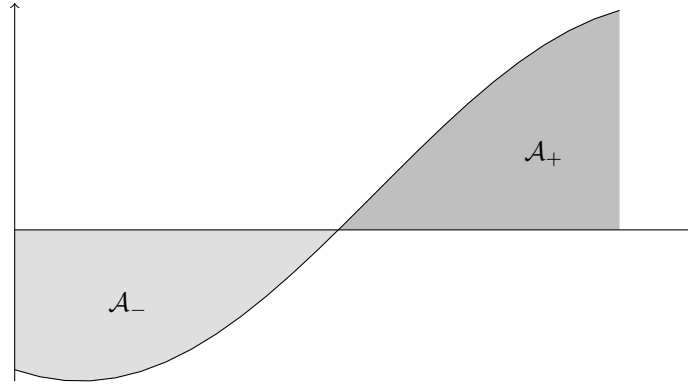


FIGURE 1.5 – Espérance d'une variable intégrable.

On notera qu'une variable aléatoire bornée est *toujours* intégrable (si $-a < X < a$, alors $|X| < a$, donc $\mathbb{E}|X| < a < \infty$).

Les résultats énoncés dans les propriétés 1.4.7 et 1.4.8 s'adaptent à des variables aléatoires intégrables.

Propriété 1.4.10. Soient X et Y deux variables aléatoires intégrables et $a > 0$.

- Si $X \leq Y$, alors $\mathbb{E}X \leq \mathbb{E}Y$. En particulier, on a $|\mathbb{E}X| \leq \mathbb{E}|X|$.
- Les variables aléatoires $X + Y$ et aX sont intégrables et on a

$$\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y \text{ et } \mathbb{E}(aX) = a\mathbb{E}X.$$

Démonstration. Pour la première propriété, on remarque que si $X \leq Y$, alors $X\mathbf{1}_{\{X \geq 0\}} \leq Y\mathbf{1}_{\{Y \geq 0\}}$ et $(-X)\mathbf{1}_{\{X > 0\}} \geq (-Y)\mathbf{1}_{\{Y > 0\}}$.

Pour la deuxième propriété, on voit que $X + Y$ est intégrable car $|X + Y| \leq |X| + |Y|$, d'où

$$\mathbb{E}|X + Y| \leq \mathbb{E}|X| + \mathbb{E}|Y| < \infty.$$

Pour aX , cela découle de $\mathbb{E}|aX| = |a|\mathbb{E}|X| < \infty$. □

1.4.2 Théorème de transfert

Les définitions 1.4.6 et 1.4.9 *ne sont pas* les définitions qui seront utilisées *en pratique* pour calculer une espérance. Le théorème 1.4.11 ci-dessous est une première étape pour donner une expression plus explicite de l'espérance d'une variable aléatoire. Il sera précisé dans des cas particuliers par les théorèmes 3.1.2 et 4.1.5.

Théorème 1.4.11 (Théorème de transfert, version 1). *Soit X une variable aléatoire à valeurs dans E et f une fonction de E dans \mathbb{R} . Si X prend un nombre fini de valeurs $\{x_1, \dots, x_n\}$, alors*

$$\mathbb{E}f(X) = \sum_{k=1}^n f(x_k)\mathbb{P}(X = x_k) = \sum_{k=1}^n f(x_k)\mu_X(\{x_k\}).$$

Démonstration. Si la variable X ne prend qu'un nombre fini de valeurs x_1, \dots, x_n , alors c'est également le cas de $f(X)$ et on peut écrire

$$f(X) = \sum_{k=1}^n f(x_k)\mathbf{1}_{\{X=x_k\}}.$$

Le résultat se déduit alors de (1.5) □

On remarque en particulier que $\mathbb{E}f(X)$ ne dépend que de f et de la loi μ_X de X : si X et Y sont deux variables aléatoires de même loi, alors $\mathbb{E}f(X) = \mathbb{E}f(Y)$.

1.4.3 Variance, moments d'ordres supérieurs

Certaines propriétés intéressantes des variables aléatoires s'expriment à partir des espérances des puissances de X , qui s'appellent les *moments* de X :

Définition 1.4.12. *Pour $n \in \mathbb{N}^*$, si X^n est intégrable, la quantité $\mathbb{E}(X^n)$ est appelée moment d'ordre n de la variable X . On dira alors que X admet un moment d'ordre n .*

En particulier, pour $n = 2$, si X^2 est intégrable, on dira que X est de carré intégrable.

Le fait d'admettre un moment d'ordre n est d'autant plus restrictif que n est grand :

Propriété 1.4.13. *Soit X une variable aléatoire réelle et $n < m$ deux entiers. Si X admet un moment d'ordre m , alors elle admet un moment d'ordre n .*

En particulier une variable aléatoire de carré intégrable est intégrable.

Démonstration. Soit $x \geq 0$. Si $x \leq 1$, alors $x^n \leq 1 \leq 1 + x^m$. Si $1 < x$, alors $1 < x^{m-n}$, d'où $x^n < x^m < 1 + x^m$. On a donc dans tous les cas $x^n \leq 1 + x^m$. Par conséquent :

$$\mathbb{E}(|X|^n) \leq \mathbb{E}(1 + |X|^m) = 1 + \mathbb{E}(|X|^m) < \infty.$$

□

Propriété 1.4.14 (Inégalité de Cauchy-Schwarz). *Si X et Y sont deux variables aléatoires réelles de carré intégrable, alors*

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

En particulier, pour $Y = 1$, on trouve :

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

Cette dernière égalité est une autre manière de voir qu'une variable aléatoire de carré intégrable est intégrable.

Démonstration. Par linéarité de l'espérance, on a

$$\mathbb{E}((|X| + \lambda|Y|)^2) = \mathbb{E}(X^2) + 2\lambda\mathbb{E}(|XY|) + \lambda^2\mathbb{E}(Y^2).$$

Vue comme une fonction de λ , cette expression est un polynôme de degré 2 qui ne prend que des valeurs positives. Par conséquent, son discriminant est négatif, ce qui s'écrit

$$4\mathbb{E}(|XY|)^2 - 4\mathbb{E}(Y^2)\mathbb{E}(X^2) \leq 0.$$

Cette inégalité n'est qu'une réécriture de celle que l'on cherchait à démontrer. \square

Corollaire 1.4.15. *Si deux variables aléatoires réelles X et Y sont de carré intégrable, alors $X + Y$ aussi.*

Démonstration. On a $(X + Y)^2 = X^2 + 2XY + Y^2 \leq X^2 + 2|XY| + Y^2$. Par conséquent

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

\square

Si X et Y sont deux variables aléatoires réelles de carré intégrable, on peut définir une "distance" entre elles par la formule

$$\delta(X, Y) = \mathbb{E}((X - Y)^2).$$

Par exemple, on a $\delta(X, Y) = 0$ si et seulement si $\mathbb{P}(X = Y) = 1$.

Pour une variable aléatoire réelle X de carré intégrable donnée, on peut se demander quelle est la constante dont elle est la plus proche, au sens de δ . C'est-à-dire, pour quelle(s) valeur(s) λ_0 a-t-on

$$\delta(X, \lambda_0) = \inf_{\lambda \in \mathbb{R}} \delta(X, \lambda).$$

En développant, on voit que $\delta(X, \lambda)$ est polynomiale par rapport à λ :

$$\delta(X, \lambda) = \mathbb{E}((X - \lambda)^2) = \mathbb{E}(X^2) - 2\lambda\mathbb{E}X + \lambda^2.$$

Ce polynôme atteint son minimum en $\lambda = \mathbb{E}X$. Autrement dit, *l'espérance de X peut s'interpréter comme la variable constante la plus "proche" de X* . La quantité $\delta(X, \mathbb{E}X) = \mathbb{E}((X - \mathbb{E}X)^2)$ mesure donc à quel point X est proche d'être constante. Cela motive la définition suivante :

Définition 1.4.16. *Si X est une variable aléatoire de carré intégrable, on définit sa variance, notée $\mathbb{V}(X)$, par :*

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}X)^2].$$

L'écart-type, noté $\sigma(X)$, est défini par $\sqrt{\mathbb{V}(X)}$.

Une variable aléatoire de variance égale à 1 est dite réduite.

Remarquer que si la variable aléatoire X a une dimension, alors l'écart-type a la même dimension, tandis que la variance a la dimension au carré, d'où l'intérêt dans la pratique de travailler avec l'écart-type. Par exemple, si X représente la taille (en cm) d'un individu pris au hasard dans une population, alors l'espérance $\mathbb{E}X$ s'exprime en centimètres, de même que l'écart-type $\sigma(X)$, alors que la variance s'exprime en centimètres carrés. On verra dans la propriété 1.4.19 que l'écart-type correspond à un écart caractéristique à la moyenne. Par exemple, sur la figure 1.6 qui représente les courbes de croissances des enfants jusqu'à trois ans^(note 4), on a tracé les graphes de la moyenne à laquelle est ajouté ou soustrait l'écart-type ou son double. On verra plus tard que l'on peut donner des résultats précis sur la proportion d'enfants qui se situent entre ces courbes.

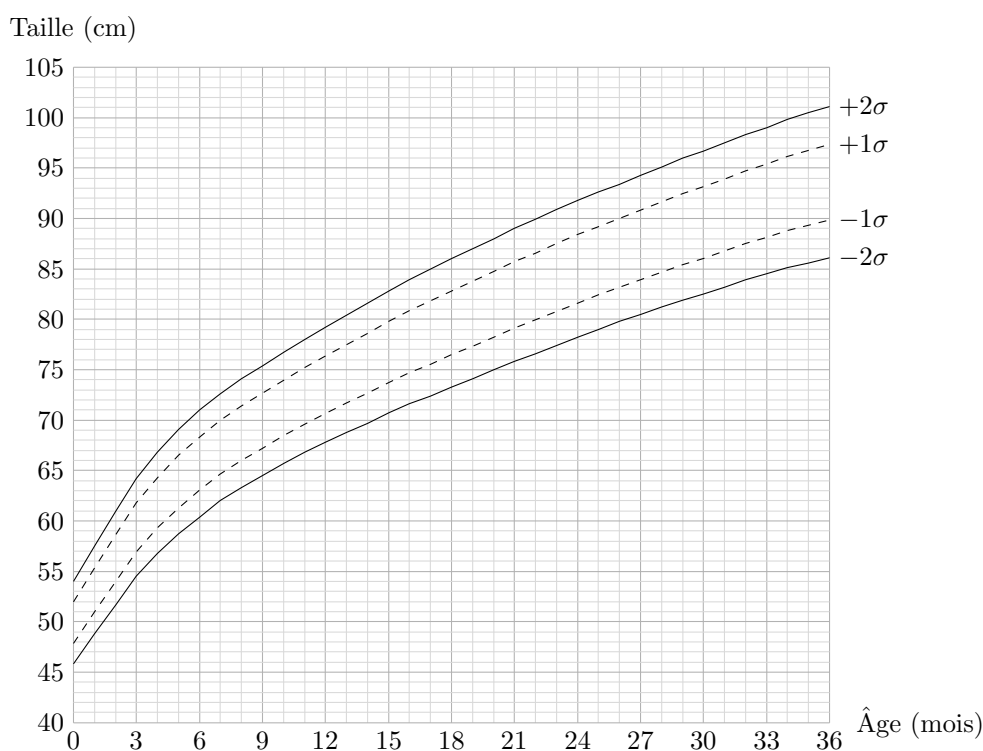


FIGURE 1.6 – Courbe de croissance des enfants de 0 à 36 mois.

Donnons quelques propriétés algébriques de la variance :

Propriété 1.4.17. Soit X une variable aléatoire de carré intégrable, alors :

- $\mathbb{V}(X) \geq 0$,
- La variance est invariante par translation : pour tout réel a , on a $\mathbb{V}(X + a) = \mathbb{V}(X)$.
- La variance est quadratique : pour tout réel a , on a $\mathbb{V}(aX) = a^2\mathbb{V}(X)$,

(note 4). Ce sont les graphiques que l'on retrouve dans les carnets de santé, avec les mêmes courbes.

— On a une autre expression de la variance, qui simplifie souvent certains calculs :

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2,$$

— La variance de X est nulle si et seulement si X est une variable aléatoire constante.

Démonstration. La positivité de $\mathbb{V}(X)$ découle de la monotonie de l'espérance et du fait que $(X - \mathbb{E}(X))^2 \geq 0$.

En utilisant la définition de la variance et la linéarité de l'espérance, on a, pour l'invariance par translation :

$$\mathbb{V}(X + a) = \mathbb{E}[(X + a - \mathbb{E}(X + a))^2] = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{V}(X),$$

et pour le caractère quadratique :

$$\mathbb{V}(aX) = \mathbb{E}[(aX - \mathbb{E}(aX))^2] = \mathbb{E}[a^2(X - \mathbb{E}(X))^2] = a^2\mathbb{V}(X).$$

La deuxième expression se prouve en développant :

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2 - 2\mathbb{E}(X)X + \mathbb{E}(X)^2] \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2. \end{aligned}$$

Pour le dernier point, on considère une variable X de variance 0. On va montrer $\mathbb{P}(X = \mathbb{E}X) = 1$. On peut écrire

$$\{X = \mathbb{E}X\} = \bigcap_{n=1}^{\infty} \left\{ |X - \mathbb{E}X| \leq \frac{1}{n} \right\},$$

où les ensembles $A_n = \{|X - \mathbb{E}X| \leq \frac{1}{n}\}$ constituent une suite décroissante. Par la propriété 1.2.4, on a donc

$$\mathbb{P}(X = \mathbb{E}X) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_n \mathbb{P}(A_n).$$

Or, par l'inégalité de Bienaymé-Tchebychev 1.4.19,

$$\mathbb{P}(A_n) = 1 - \mathbb{P}\left(|X - \mathbb{E}X| > \frac{1}{n}\right) \geq 1 - n^2\mathbb{V}X = 1.$$

On a donc $\mathbb{P}(X = \mathbb{E}X) = 1$. □

1.4.4 Inégalité de Markov et de Bienaymé-Tchebychev

On présente dans cette partie deux inégalités permettant de “localiser” les valeurs prises par une variable aléatoire, en fonction de sa moyenne et de sa variance.

La première inégalité formalise l'intuition suivante : si une variable aléatoire est intégrable alors elle “n'est pas trop grande”, et dans ce cas, sa probabilité de prendre des “grandes” valeurs est “petite”.

Propriété 1.4.18 (Inégalité de Markov). *Soit X une variable aléatoire positive. Alors,*

$$\forall a > 0, \quad \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Plus généralement, si X est une variable aléatoire admettant un moment d'ordre $n \in \mathbb{N}^$ et si $a > 0$, alors*

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

Démonstration. On a l'inégalité $a\mathbf{1}_{\{X \geq a\}} \leq X$, comme on peut le voir sur la figure 1.7.

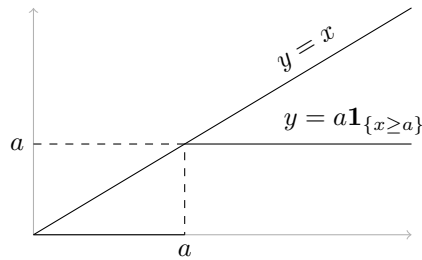


FIGURE 1.7 – L'inégalité $a\mathbf{1}_{x \geq a} \leq x$.

Par conséquent,

$$\mathbb{P}(X \geq a) = \mathbb{E}\mathbf{1}_{\{X \geq a\}} \leq \mathbb{E}\left(\frac{X}{a}\right) = \frac{\mathbb{E}X}{a}.$$

Pour la seconde inégalité, il suffit de remarquer l'égalité entre évènement $\{|X| \geq a\} = \{|X|^n \geq a^n\}$ et d'appliquer la première égalité à $|X|^n$ et a^n . \square

On peut aussi lire l'inégalité de Markov comme “une variable aléatoire positive a peu de chances d'être beaucoup plus grande que sa moyenne”. En effet, en prenant $a = \alpha\mathbb{E}X$, l'inégalité se récrit

$$\mathbb{P}(X \geq \alpha\mathbb{E}X) \leq \frac{1}{\alpha}.$$

La deuxième inégalité formalise l'intuition suivante : la variance mesure l'écart entre une variable aléatoire et sa moyenne. Par conséquent si une variable a une “petite” variance, elle aura une “grande” probabilité de se trouver “près” de sa moyenne.

Propriété 1.4.19 (Inégalité de Bienaymé-Tchebychev). *Soient X une variable aléatoire de carré intégrable et $a > 0$. Alors,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathbb{V}(X)}{a^2}.$$

Démonstration. L'inégalité de Bienaymé-Tchebychev est une conséquence de l'inégalité de Markov. La variable aléatoire X étant de carré intégrable, il en est de même pour la variable aléatoire $X - \mathbb{E}X$. On applique alors l'inégalité de Markov avec la variable aléatoire $Y = |X - \mathbb{E}X|^2$: pour tout $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) = \mathbb{P}(|X - \mathbb{E}X|^2 \geq a^2) \leq \frac{\mathbb{E}[|X - \mathbb{E}X|^2]}{a^2} = \frac{\mathbb{V}(X)}{a^2}.$$

\square

En appliquant la propriété précédente à $a = 2\sigma(X)$ et $a = 3\sigma(X)$, on obtient, pour toute variable aléatoire réelle X ,

$$\mathbb{P}(|X - \mathbb{E}X| \leq 2\sigma(X)) \geq \frac{3}{4} = 0.75$$
$$\mathbb{P}(|X - \mathbb{E}X| \leq 3\sigma(X)) \geq \frac{8}{9} = 0.888\dots$$

Ces propriétés sont encore une illustration du fait que l'écart-type mesure la dispersion de la variable aléatoire autour de sa moyenne. Par exemple, la première nous dit qu'au moins les trois quarts des enfants ont leur taille comprise entre les deux courbes en trait plein de la figure 1.6.

Chapitre 2

Probabilités conditionnelles - indépendance

2.1 Probabilité conditionnelle

Dans de nombreuses situations, on observe un phénomène aléatoire dont on connaît une information partielle. Par exemple, pendant une partie de cartes à jouer, la distribution des cartes est faite aléatoirement, mais chaque joueur connaît ses propres cartes et dispose donc d'une information sur le jeu des autres.

On voudrait donc, à partir d'un premier espace probabilisé (Ω, \mathbb{P}) , définir une nouvelle mesure de probabilité sur Ω ne portant que sur un événement B : plus spécifiquement, on veut donner une probabilité nulle aux événements disjoints de B . Cette idée est formalisée dans la notion de *probabilité conditionnelle*.

2.1.1 Définition

À partir d'un espace probabilisé (Ω, \mathbb{P}) , on veut donc construire une nouvelle probabilité qui "ressemble" à \mathbb{P} , mais telle que les sous-ensembles de B^c aient une probabilité nulle. Une manière de faire pourrait être de poser $\mathbb{P}_B(A) = \mathbb{P}(A \cap B)$, mais l'objet obtenu ne serait pas une mesure de probabilité, puisqu'on aurait $\mathbb{P}_B(\Omega) = \mathbb{P}(B) \neq 1$ (a priori). La solution consiste à renormaliser par la quantité $\mathbb{P}(B)$.

Définition 2.1.1. Soit (Ω, \mathbb{P}) un espace probabilisé, et B un événement tel que $\mathbb{P}(B) > 0$. Pour tout $A \subset \Omega$, on définit la probabilité conditionnelle de A sachant B , notée $\mathbb{P}_B(A)$ ou $\mathbb{P}(A|B)$, par :

$$\mathbb{P}_B(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \quad (2.1)$$

La condition $\mathbb{P}(B) > 0$ permet de diviser par $\mathbb{P}(B)$. En pratique, on a $0 < \mathbb{P}(B) < 1$. En effet, le cas $\mathbb{P}(B) = 1$ a peu d'intérêt, puisqu'on obtient alors $\mathbb{P}_B = \mathbb{P}$.

La notation $\mathbb{P}(A|B)$ est la plus couramment utilisée ; en revanche, la notation figurant au programme du lycée est $\mathbb{P}_B(A)$. L'intérêt de la notation $\mathbb{P}_B(A)$ est qu'elle souligne bien le fait que \mathbb{P}_B est une mesure de probabilité que l'on applique à l'évènement A . En outre, la notation $\mathbb{P}(A|B)$ pourrait faire

croire que l'on calcule la probabilité d'un évènement noté $A|B$. Or, on ne peut pas donner de sens à un "évènement $A|B$ ".

Propriété 2.1.2. *L'application \mathbb{P}_B est une probabilité sur Ω .*

Démonstration. Il faut montrer que \mathbb{P}_B satisfait les trois axiomes caractérisant une probabilité.

- Par hypothèse, on a $\mathbb{P}(B) > 0$. De plus, pour tout $A \subset \Omega$, on a $A \cap B \subset B$, d'où l'inégalité $0 \leq \mathbb{P}(A \cap B) \leq \mathbb{P}(B)$, et donc :

$$0 \leq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \leq 1.$$

- Comme $\Omega \cap B = B$, et $\Omega \cap \emptyset = \emptyset$, on a bien $\mathbb{P}_B(\Omega) = 1$ et $\mathbb{P}_B(\emptyset) = 0$
- Soit $(A_n)_{n \geq 1}$ une famille dénombrable d'évènements deux à deux disjoints. Alors on a l'égalité $(\bigcup_{n \geq 1} A_n) \cap B = \bigcup_{n \geq 1} (A_n \cap B)$. Comme les évènements $(A_n)_{n \geq 1}$ sont deux à deux disjoints, il en est de même pour les évènements $(A_n \cap B)_{n \geq 1}$. Par conséquent, on a

$$\mathbb{P} \left(\bigcup_{n \geq 1} (A_n \cap B) \right) = \sum_{n \geq 1} \mathbb{P}(A_n \cap B),$$

et on conclut :

$$\mathbb{P}_B \left(\bigcup_{n \geq 1} A_n \right) = \frac{\mathbb{P} \left((\bigcup_{n \geq 1} A_n) \cap B \right)}{\mathbb{P}(B)} = \sum_{n \geq 1} \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_{n \geq 1} \mathbb{P}_B(A_n).$$

□

La définition de la probabilité conditionnelle peut se récrire $\mathbb{P}(A \cap B) = \mathbb{P}_B(A)\mathbb{P}(B)$. Autrement dit, on peut calculer la probabilité d'une intersection en conditionnant par un des évènements. En appliquant plusieurs fois la même égalité, on arrive à l'équation suivante :

$$\mathbb{P} \left(\bigcap_{k=1}^n A_k \right) = \mathbb{P}(A_1)\mathbb{P}_{A_1}(A_2)\mathbb{P}_{A_1 \cap A_2}(A_3) \cdots \mathbb{P}_{A_1 \cap \cdots \cap A_{n-1}}(A_n), \quad (2.2)$$

où les $(A_k)_{1 \leq k \leq n}$ sont des évènements tels que $\mathbb{P} \left(\bigcap_{k=1}^{n-1} A_k \right) > 0$. C'est l'égalité (2.2) qui justifie la manipulation des arbres de probabilité.

2.1.2 Arbres de probabilité

Les arbres de probabilité sont un outil permettant de calculer simplement diverses probabilités concernant une même expérience aléatoire.

Nous avons d'abord besoin de définir la notion d'arbre.

Définition 2.1.3. *Un arbre orienté est un ensemble \mathcal{A} , dont un des éléments x est spécifié. L'élément x est appelé racine de l'arbre.*

À chaque élément de \mathcal{A} (on parlera aussi de sommet) correspond un sous-ensemble de \mathcal{A} qui est l'ensemble de ses enfants. On suppose que chaque sommet est l'enfant d'un seul autre sommet, à l'exception de la racine, qui n'a pas de père.

Par ailleurs, on suppose qu'en partant d'un sommet puis en allant à son père, puis au père de son père, et ainsi de suite, on ne retombera jamais sur le sommet initial. En particulier, on retombera toujours sur la racine de l'arbre.

Graphiquement, on représentera les sommets d'un arbre en les reliant d'une flèche à leurs enfants. Sur la figure 2.1, le premier graphe est un arbre dont la racine est R . Les enfants de R sont A et B . Les enfants de B sont C , D et E . Les sommets A , C , D et E n'ont pas d'enfants. Le deuxième graphe n'est pas un arbre car le sommet C est l'enfant de plusieurs sommets.

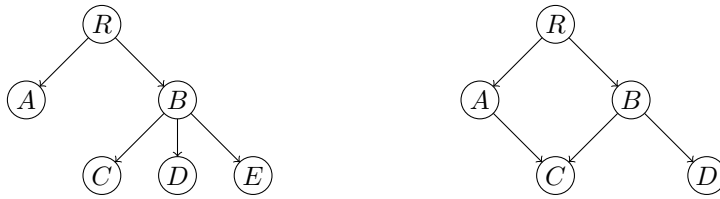


FIGURE 2.1 – Exemple et contre-exemple d'arbres.

Définition 2.1.4. *Un arbre de probabilité est un arbre orienté dont chaque sommet est un événement de probabilité non nulle tel que :*

- la racine de l'arbre est l'évènement Ω ;
- si B est un sommet de l'arbre, alors l'ensemble des enfants de B constitue une partition^(note 1) de B .

À chaque arête reliant un sommet B à un de ses enfants, B , on associe la probabilité $\mathbb{P}_B(A)$.

Dans de nombreux cas, les enfants d'un sommet A sont de la forme $A \cap B_i$, où les B_i constituent une partition de Ω . On a alors bien $\bigcup_i A \cap B_i = A \cap \bigcup_i B_i = A \cap \Omega = A$, avec une union disjointe.

La somme des probabilités des branches issues d'un même sommet est 1. En effet, si \mathcal{E} est l'ensemble des enfants d'un sommet A , on a par hypothèse $\bigcup_{B \in \mathcal{E}} B = A$, où l'union est disjointe. Comme \mathbb{P}_A est une mesure de probabilité, on a donc

$$\sum_{B \in \mathcal{E}} \mathbb{P}_A(B) = \mathbb{P}_A\left(\bigcup_{B \in \mathcal{E}} B\right) = \mathbb{P}_A(A) = 1.$$

On donne un exemple d'arbre de probabilité en figure 2.2.

Dans l'arbre de la figure 2.2, on a les décompositions $\Omega = A_1 \cup A_2$, ainsi que $A_2 = A_{21} \cup A_{22} \cup A_{23}$ et $A_{21} = A_{211} \cup A_{212}$ où les unions sont disjointes. Comme les sommes des probabilités partant d'un sommet doivent valoir 1, on a aussi $a + b = c + d + e = f + g = 1$.

En effectuant le produit des probabilités le long d'une branche, on retrouve l'équation (2.2) :

Propriété 2.1.5. *Soit (A_0, A_1, \dots, A_n) un chemin de l'arbre, au sens où A_i est toujours le père de A_{i+1} , avec $A_0 = \Omega$. Alors $\mathbb{P}(A_n)$ est donné par le produit des probabilités sur les branches formant le chemin.*

(note 1). On rappelle qu'une *partition* de B est une famille $(A_i)_{i \in I}$ d'ensembles *disjoints* tels que $\bigcup_{i \in I} A_i = B$.

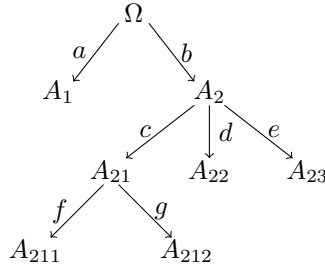


FIGURE 2.2 – Un exemple d'arbre de probabilité.

Démonstration. C'est une récurrence sur la longueur du chemin : si le chemin est de longueur 0, il est réduit à $A_0 = \Omega$, et alors $\mathbb{P}(A_0) = \mathbb{P}(\Omega) = 1$. Or, il n'y a pas de branche sur le chemin, et un produit vide vaut 1.

Ensuite, on suppose la propriété vraie pour les chemins de longueur n . Si (A_0, \dots, A_{n+1}) est un chemin de longueur $n + 1$, alors $\mathbb{P}(A_{n+1}) = \mathbb{P}_{A_n}(A_{n+1})\mathbb{P}(A_n)$. Or, par hypothèse de récurrence, $\mathbb{P}(A_n)$ est le produit des n premières probabilités présentes sur les branches du chemin, et $\mathbb{P}_{A_n}(A_{n+1})$ est la $n + 1$ ème. \square

Pour illustration, sur la figure 2.2, on trouve par exemple $\mathbb{P}(A_{211}) = bcf$, $\mathbb{P}(A_{23}) = be$ ou encore $\mathbb{P}(A_{21}) = bc$.

2.1.3 Formule des probabilités totales et formule de Bayes

Dans cette partie, on présente deux formules importantes faisant intervenir des probabilités conditionnelles.

La première formule permet de calculer la probabilité d'un évènement à partir d'une partition de Ω .

Théorème 2.1.6 (Formule des probabilités totales). *Soit $(B_n)_{n \in \mathbb{N}}$ une partition de Ω , telle que pour tout $n \in \mathbb{N}$, $\mathbb{P}(B_n) > 0$, et soit $A \subset \Omega$. Alors :*

$$\mathbb{P}(A) = \sum_{n \in \mathbb{N}} \mathbb{P}(A \cap B_n) = \sum_{n \in \mathbb{N}} \mathbb{P}_{B_n}(A)\mathbb{P}(B_n).$$

La formule reste valide si on remplace la partition infinie dénombrable $(B_n)_{n \in \mathbb{N}}$ par une partition finie $(B_n)_{n \in \{1, \dots, N\}}$.

Démonstration. Comme $(B_n)_{n \in \mathbb{N}}$ est une partition de Ω , on a :

$$A = A \cap \Omega = A \cap \left(\bigcup_{n \in \mathbb{N}} B_n \right) = \bigcup_{n \in \mathbb{N}} (A \cap B_n),$$

où les évènements $(A \cap B_n)_{n \in \mathbb{N}}$ sont deux à deux disjoints. D'où le résultat. \square

La deuxième formule permet d'exprimer la probabilité conditionnelle de A sachant B à l'aide de la probabilité conditionnelle de B sachant A .

Théorème 2.1.7 (Formule de Bayes). *Soient A et B deux évènements, tels que $\mathbb{P}(A) > 0$ et $\mathbb{P}(B) > 0$, alors :*

$$\mathbb{P}_B(A) = \frac{\mathbb{P}_A(B)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Démonstration. Par définition de la probabilité conditionnelle, on a :

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

□

2.2 Indépendance d'évènements

Définition 2.2.1. *Deux évènements A et B sont dits indépendants, si*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \quad (2.3)$$

Si A est un évènement de probabilité non nulle, l'équation (2.3) peut se récrire $\mathbb{P}(B) = \mathbb{P}_A(B)$. Cette équation se comprend comme le fait que la connaissance de A ne modifie pas la probabilité de B .

Par exemple, les évènements \emptyset et Ω sont indépendants de n'importe quel autre évènement A . En effet,

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

de même

$$\mathbb{P}(\Omega \cap A) = \mathbb{P}(A) = 1 \times \mathbb{P}(A) = \mathbb{P}(\Omega)\mathbb{P}(A).$$

Propriété 2.2.2. *Si les évènements A et B sont indépendants, alors il en est de même des évènements A^c et B , A et B^c , A^c et B^c .*

Démonstration. On démontre seulement que A^c et B sont indépendants, les autres cas s'en déduisent. Comme $B = (A \cap B) \cup (A^c \cap B)$ avec une union disjointe, on a

$$\begin{aligned} \mathbb{P}(A^c \cap B) &= \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) \\ &= \mathbb{P}(B)(1 - \mathbb{P}(A)) \\ &= \mathbb{P}(B)\mathbb{P}(A^c). \end{aligned}$$

□

Quand on a plus que deux évènements, il y a deux généralisations de la notion d'indépendance, qui ne sont pas équivalentes.

Définition 2.2.3. *Soient $(A_n)_{n \in \mathbb{N}}$ une suite d'évènements*

- *Les A_n sont dits mutuellement indépendants, si pour tout entier $k \geq 1$ et tout k -uplet d'entiers $1 \leq i_1 < \dots < i_k$, on a*

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

- *Les A_n sont dits deux à deux indépendants, si pour tous entiers $1 \leq i_1 < i_2$, on a :*

$$\mathbb{P}(A_{i_1} \cap A_{i_2}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}).$$

Propriété 2.2.4. *Des évènements mutuellement indépendants sont deux à deux indépendants.*

Démonstration. La définition de l'indépendance deux à deux est un cas particulier de l'indépendance mutuelle, pour $k = 2$. \square

Il existe toutefois des familles d'évènements qui sont indépendants deux à deux sans être mutuellement indépendants. Par exemple, si on se donne deux variables aléatoires X et Y telles que $\mathbb{P}(X = \pm 1, Y = \pm 1) = \frac{1}{4}$ (pour tous les choix de signes possibles), alors les évènements $\{X = 1\}$, $\{Y = 1\}$ et $\{XY = 1\}$ sont indépendants deux à deux, mais pas mutuellement indépendants. En effet, dans ce cas, on a

$$\mathbb{P}(X = 1) = \mathbb{P}(X = 1, Y = 1) + \mathbb{P}(X = 1, Y = -1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

On a de même $\mathbb{P}(Y = 1) = \frac{1}{2}$. On a donc bien

$$\mathbb{P}(X = 1, Y = 1) = \frac{1}{4} = \mathbb{P}(X = 1)\mathbb{P}(Y = 1).$$

On a également

$$\mathbb{P}(XY = 1) = \mathbb{P}(X = 1, Y = 1) + \mathbb{P}(X = -1, Y = -1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

d'où

$$\mathbb{P}(X = 1, XY = 1) = \mathbb{P}(X = 1, Y = 1) = \frac{1}{4} = \mathbb{P}(X = 1)\mathbb{P}(XY = 1),$$

et de même, on a $\mathbb{P}(Y = 1, XY = 1) = \mathbb{P}(Y = 1)\mathbb{P}(XY = 1)$. L'indépendance deux à deux est donc bien vérifiée. En revanche, ce n'est pas le cas de l'indépendance mutuelle puisque :

$$\mathbb{P}(X = 1, Y = 1, XY = 1) = \mathbb{P}(X = 1, Y = 1) = \frac{1}{4} \neq \mathbb{P}(X = 1)\mathbb{P}(Y = 1)\mathbb{P}(XY = 1).$$

On peut expliquer cet exemple de la manière suivante : X et Y valent 1 ou -1 avec même probabilité, et connaître X ne donne aucune information sur la valeur de Y (et inversement). On remarque aussi que connaître X ne donne aucune information sur la valeur de XY (et de même, connaître Y ne donne aucune information sur la valeur de XY). En revanche, connaître X et Y nous donne la valeur de XY .

2.3 Variables aléatoires indépendantes

Dans l'exemple de la fin de la partie 2.2, les évènements indépendants étaient obtenus à partir de variables aléatoires. Cela nous amène à la notion de *variables aléatoires indépendantes*.

Définition 2.3.1. *Deux variables aléatoires X et Y à valeurs dans E sont dites indépendantes, si pour tous sous-ensembles A et B de E les évènements $\{X \in A\}$ et $\{Y \in B\}$ sont indépendants :*

$$\mathbb{P}(\{X \in A\} \cap \{Y \in B\}) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B). \quad (2.4)$$

Pour le cas de plusieurs variables on a la même généralisation que dans le cas des évènements.

Définition 2.3.2. *Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires à valeurs dans E .*

- On dit que les X_n sont deux à deux indépendantes si, pour tout n et m , les variables X_n et X_m sont indépendantes.
- On dit que les variables X_n sont mutuellement indépendantes si pour toute suite $(A_n)_{n \in \mathbb{N}}$ de sous-ensembles de E , les évènements $\{X_n \in A_n\}$ sont mutuellement indépendants.

Reprenons l'exemple de la fin de la partie 2.2, où l'on avait deux variables X et Y vérifiant $\mathbb{P}(X = \pm 1, Y = \pm 1) = \frac{1}{4}$. Dans ce cas, les trois variables X , Y et XY sont indépendantes deux à deux, mais pas mutuellement.

On peut récrire la définition de l'indépendance en utilisant des espérances. En effet, l'équation (2.4) peut se récrire

$$\mathbb{E}(\mathbf{1}_A(X)\mathbf{1}_B(Y)) = \mathbb{E}(\mathbf{1}_A(X))\mathbb{E}(\mathbf{1}_B(Y)).$$

Par conséquent, si on applique cette relation à plusieurs sous-ensembles $(A_i)_{i=1, \dots, I}$ et $(B_j)_{j=1, \dots, J}$ de E , on obtient la relation suivante :

$$\mathbb{E} \left(\sum_{i=1}^I \alpha_i \mathbf{1}_{A_i}(X) \sum_{j=1}^J \beta_j \mathbf{1}_{B_j}(Y) \right) = \mathbb{E} \left(\sum_{i=1}^I \alpha_i \mathbf{1}_{A_i}(X) \right) \mathbb{E} \left(\sum_{j=1}^J \beta_j \mathbf{1}_{B_j}(Y) \right)$$

Si on définit alors deux fonctions f et g de E dans \mathbb{R} par :

$$f(x) = \sum_{i=1}^I \alpha_i \mathbf{1}_{A_i}(x) \text{ et } g(y) = \sum_{j=1}^J \beta_j \mathbf{1}_{B_j}(y), \quad (2.5)$$

on a montré la relation

$$\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y)),$$

pour toutes fonctions de la forme (2.5), c'est-à-dire, toutes fonctions ne prenant qu'un nombre fini de valeurs. Cette relation se généralise en fait à toutes fonctions f et g .

Propriété 2.3.3. Soit X et Y deux variables aléatoires. Les assertions suivantes sont équivalentes.

1. Les variables aléatoires X et Y sont indépendantes ;
2. Pour toutes fonctions bornées f et g de E dans \mathbb{R} , on a :

$$\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y)) ;$$

3. Pour toutes fonctions positives f et g sur E , on a :

$$\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y)).$$

4. Pour toutes fonctions f et g de E dans \mathbb{R} telles que $f(X)$, $g(Y)$ et $f(X)g(Y)$ soient intégrables, on a :

$$\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y)).$$

Les conditions sur f et g permettent de s'assurer que les espérances sont bien définies : dans le cas 2, il n'y a que des variables aléatoires bornées, donc intégrables, dans le cas 3, il n'y a que des variables positives, dont l'espérance est donc bien définie.

Démonstration. L'indépendance se déduit de l'égalité entre les espérance en choisissant $f = \mathbf{1}_A$ et $g = \mathbf{1}_B$.

La réciproque est plus difficile et sera admise. \square

On déduit de la propriété 2.3.3 le fait suivant : si les variables aléatoires X et Y sont indépendantes, alors pour toutes fonctions f et g , les variables aléatoires $f(X)$ et $g(Y)$ sont aussi indépendantes.

Propriété 2.3.4. Soient X et Y deux variables aléatoires de carré intégrable. Alors, si X et Y sont indépendantes,

1. $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$,
2. $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$.

Démonstration. La première égalité découle de la propriété 2.3.3 appliquée à $f(x) = x$ et $g(y) = y$.

Pour la deuxième, on suppose, quitte à changer X en $X - \mathbb{E}X$, que les variables sont centrées. On a alors

$$\begin{aligned} \mathbb{V}(X + Y) &= \mathbb{E}((X + Y)^2) - (\mathbb{E}(X + Y))^2 \\ &= \mathbb{E}(X^2) + \mathbb{E}(Y^2) + 2\mathbb{E}(XY) - (\mathbb{E}X)^2 - (\mathbb{E}Y)^2 - 2\mathbb{E}X\mathbb{E}Y \\ &= \left[\mathbb{E}(X^2) - (\mathbb{E}X)^2 \right] + \left[\mathbb{E}(Y^2) - (\mathbb{E}Y)^2 \right] + \left[2\mathbb{E}(XY) - 2\mathbb{E}X\mathbb{E}Y \right] \\ &= \mathbb{V}(X) + \mathbb{V}(Y) + 2 \times 0. \end{aligned}$$

\square

Attention, l'égalité $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$ n'implique pas nécessairement que les variables aléatoires X et Y sont indépendantes. Par exemple, soient X et Y deux variables indépendantes avec $\mathbb{P}(X = \pm 1) = \frac{1}{2}$ et $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = \frac{1}{2}$. Par conséquent, $\mathbb{E}X = \frac{1}{2} \times 1 + \frac{1}{2} \times (-1) = 0$. On a donc

$$\mathbb{E}(Y \times (XY)) = \mathbb{E}(Y^2 \times X) = \mathbb{E}(Y^2)\mathbb{E}X = \mathbb{E}(Y^2) \times 0 = 0$$

et

$$\mathbb{E}Y\mathbb{E}(XY) = (\mathbb{E}Y)^2\mathbb{E}X = 0.$$

On a donc bien $\mathbb{E}(Y \times (XY)) = \mathbb{E}Y \times \mathbb{E}(XY)$, or les deux variables Y et XY ne sont pas indépendantes. En effet, $\mathbb{P}(Y = 1, XY = 2) = 0$, alors que $\mathbb{P}(Y = 1)\mathbb{P}(XY = 2) = \frac{1}{2} \times \frac{1}{4}$.

Nous finissons ce chapitre sur un résultat semblable dans l'esprit à la propriété 1.3.4. On va souvent donner des hypothèses sous la forme "Soient X et Y deux variables aléatoires indépendantes de loi (...)". En toute rigueur, on doit s'assurer qu'il *existe* un espace probabilisé sur lequel on puisse définir de telles variables aléatoires, sans quoi le résultat énoncé serait vide. C'est le but du théorème suivant :

Théorème 2.3.5. Soient μ et ν deux mesures de probabilité sur un ensemble E . Alors il existe un espace probabilisé (Ω, \mathbb{P}) et deux variables aléatoires X et Y sur (Ω, \mathbb{P}) telles que X et Y sont indépendantes et de loi respectives μ et ν .

Démonstration. On pose $\Omega = E \times E$, et on définit

$$\left\{ \begin{array}{l} X : E \times E \rightarrow E, \\ (x, y) \mapsto x, \end{array} \right. \text{ et } \left\{ \begin{array}{l} Y : E \times E \rightarrow E, \\ (x, y) \mapsto y. \end{array} \right.$$

On doit alors définir une probabilité sur $\Omega = E \times E$ telle que pour tout A et B inclus dans E , on ait

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(\{(x, y) \in E \times E, x \in A, y \in B\}) = \mu(A)\nu(B).$$

Dans le cas où E est indénombrable cette construction est complexe et nous ne la ferons pas. Dans le cas où E est dénombrable, on peut remarquer que $E \times E$ est également dénombrable. On verra au chapitre 3 (théorème 3.1.1) qu'il suffit alors de définir $\mathbb{P}(\{(x, y)\})$ pour tout élément (x, y) de $E \times E$. Ici le choix est nécessairement $\mathbb{P}(\{(x, y)\}) = \mu(\{x\})\nu(\{y\})$. \square

En appliquant plusieurs fois le résultat précédent, on peut construire plusieurs variables mutuellement indépendantes de lois données.

Chapitre 3

Variables aléatoires discrètes - dénombrément

3.1 Espaces probabilisés dénombrables (ou finis)

Dans ce chapitre, nous allons considérer des variables aléatoires à valeurs dans un ensemble E dénombrable ou fini. Les exemples les plus courants seront $E = \{1, \dots, N\}$ ou $E = \mathbb{N}$. Il se trouve que dans le cas d'un ensemble E dénombrable, il est simple de décrire l'ensemble des mesures de probabilité sur E .

Théorème 3.1.1. *Soit μ une mesure de probabilité sur un ensemble E dénombrable. Alors pour tout sous-ensemble A de E , on a*

$$\mu(A) = \sum_{x \in A} \mu(\{x\}).$$

Inversement, si $(p_x)_{x \in E}$ est une famille d'éléments de $[0, 1]$ avec $\sum_{x \in E} p_x = 1$, alors il existe une unique mesure de probabilité μ sur E telle que

$$\mu(\{x\}) = p_x.$$

Les sommes qui apparaissent dans ce théorème ont bien un sens car il s'agit de sommes dénombrables de termes positifs.

Démonstration. L'expression de $\mu(A)$ découle du fait que $A = \bigcup_{x \in A} \{x\}$, où l'union est disjointe.

Montrons maintenant qu'une famille $(p_x)_{x \in E}$ vérifiant les conditions de l'énoncé définit bien une mesure de probabilité. On pose pour tout $A \subset E$

$$\mu(A) = \sum_{x \in A} p_x.$$

On va montrer que μ est bien une probabilité. Comme $0 \leq p_x$ et $\sum_{x \in E} p_x = 1$, on a bien $\mu(A) \in [0, 1]$. Ensuite, si $(A_n)_{n \in \mathbb{N}}$ est une suite d'ensembles disjoints on a

$$\mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{x \in \bigcup_{n \in \mathbb{N}} A_n} p_x = \sum_{n \in \mathbb{N}} \sum_{x \in A_n} p_x = \sum_{n \in \mathbb{N}} \mu(A_n)$$

(le fait que les A_n sont disjoints est utilisé dans la deuxième égalité). \square

Pour une mesure sur un ensemble dénombrable, l'espérance peut se mettre sous la forme d'une somme.

Théorème 3.1.2 (Théorème de transfert, version discrète). *Soit X est une variable à valeurs dans un ensemble dénombrable $E = \{x_n, n \in \mathbb{N}\}$ et soit f une fonction de E dans \mathbb{R} .*

1. *Si f est à valeurs positives, alors*

$$\mathbb{E}f(X) = \sum_{n \in \mathbb{N}} f(x_n) \mathbb{P}(X = x_n).$$

2. *Dans le cas général, $f(X)$ est intégrable si et seulement si*

$$\sum_{n \in \mathbb{N}} |f(x_n)| \mathbb{P}(X = x_n) < \infty.$$

Dans ce cas, on a alors

$$\mathbb{E}f(X) = \sum_{n \in \mathbb{N}} f(x_n) \mathbb{P}(X = x_n).$$

Démonstration. Commençons par le cas $f \geq 0$. On doit revenir à la définition 1.4.6 La variable $Y_N = \sum_{n=1}^N f(x_n) \mathbf{1}_{\{X=x_n\}}$ est une variable prenant un nombre fini de valeurs et vérifiant $Y_N \leq f(X)$. Par conséquent,

$$\mathbb{E}f(X) \geq \sup_N \mathbb{E}Y_N = \sup_N \sum_{n=1}^N f(x_n) \mathbb{P}(X = x_n) = \sum_{n=1}^{\infty} f(x_n) \mathbb{P}(X = x_n).$$

Pour obtenir la majoration inverse, on considère une variable aléatoire Y vérifiant $Y \leq f(X)$ et prenant un nombre fini de valeurs $\{y_1, \dots, y_M\}$. On a alors

$$\begin{aligned} \mathbb{E}Y &= \sum_{k=1}^M y_k \mathbb{P}(Y = y_k) = \sum_{k=1}^M y_k \sum_{n=1}^{\infty} \mathbb{P}(Y = y_k, X = x_n) = \sum_{n=1}^{\infty} \sum_{k=1}^M y_k \mathbb{P}(Y = y_k, X = x_n) \\ &\leq \sum_{n=1}^{\infty} \sum_{k=1}^M f(x_n) \mathbb{P}(Y = y_k, X = x_n) \\ &= \sum_{n=1}^{\infty} f(x_n) \mathbb{P}(X = x_n), \end{aligned}$$

où l'inégalité vient du fait que l'hypothèse $Y \leq f(X)$ implique que $y_k \leq f(x_n)$ si $\mathbb{P}(X = x_n, Y = y_k)$ est non-nul.

La caractérisation de l'intégrabilité de $f(X)$ vient de l'égalité

$$\mathbb{E}|f(X)| = \sum_{n \in \mathbb{N}} |f(x_n)| \mathbb{P}(X = x_n),$$

qui se déduit du cas positif. L'expression de l'espérance vient alors de la décomposition de $f(X)$ en partie positive et partie négative. \square

3.1.1 Indépendance

Si X et Y sont deux variables aléatoires à valeurs dans un ensemble dénombrable, alors le couple (X, Y) prend aussi ses valeurs dans un ensemble dénombrable. Dans ce cas, on peut alors caractériser simplement l'indépendance des variables X et Y .

Propriété 3.1.3. *Soit E un ensemble dénombrable et soient X et Y deux variables aléatoires à valeurs dans E . Alors, X et Y sont indépendantes si et seulement si pour tous éléments x et y de E on a*

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y). \quad (3.1)$$

Démonstration. L'implication directe découle de la définition de l'indépendance.

Pour la réciproque, on suppose que X et Y vérifient (3.1). Soient A et B deux sous-ensembles de E . On a

$$\begin{aligned} \mathbb{P}(X \in A, Y \in B) &= \sum_{x \in A} \sum_{y \in B} \mathbb{P}(X = x, Y = y) = \sum_{x \in A} \sum_{y \in B} \mathbb{P}(X = x)\mathbb{P}(Y = y) \\ &= \left(\sum_{x \in A} \mathbb{P}(X = x) \right) \left(\sum_{y \in B} \mathbb{P}(Y = y) \right) \\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B). \end{aligned}$$

Les variables X et Y sont donc bien indépendantes. □

3.2 Lois discrètes usuelles

Dans cette partie on donne quelques propriétés de variables aléatoires discrètes les plus courantes.

3.2.1 Loi uniforme

Définition 3.2.1. *Soit E un ensemble fini. On appelle loi uniforme sur E la mesure de probabilité définie par*

$$\mu(A) = \frac{\text{card}A}{\text{card}E}.$$

Autrement dit, on a $\mu(\{x\}) = \frac{1}{\text{card}E}$, pour tout $x \in E$.

Dans le cas de la loi uniforme, un calcul de probabilité se ramène donc à un calcul de cardinal, ce qui sera le sujet de la partie 3.3.

Lorsqu'une variable aléatoire suit la loi uniforme sur un ensemble E , on dit aussi que l'on est en situation d'*équiprobabilité*. Noter aussi que dire qu'un élément de E est tiré "au hasard" pour signifier qu'il suit une loi uniforme sur E peut être ambigu.

Dans le cas de la loi uniforme sur $\{0, \dots, n\}$, on peut donner les résultats suivants :

Propriété 3.2.2. *Soit X une variable aléatoire de loi uniforme sur $\{0, \dots, n\}$. Alors X est de carré intégrable et vérifie*

$$\mathbb{E}X = \frac{n}{2} \text{ et } \mathbb{V}X = \frac{n(n+1)}{12}.$$

3.2.2 Loi de Bernoulli

Soit $0 < p < 1$ et une variable aléatoire $X : \Omega \rightarrow \{0, 1\}$ de loi de probabilité :

$$\mu_X(\{1\}) = p, \mu_X(\{0\}) = 1 - p.$$

Alors la loi de X est appelée *loi de Bernoulli de paramètre p* .

Les variables aléatoires de Bernoulli sont exactement les variables aléatoires de la forme $\mathbf{1}_A$. En effet, si X suit une loi de Bernoulli, alors on peut partitionner Ω en $\Omega = A \cup A^c$ où $A = \{X = 1\}$ et $A^c = \{X = 0\}$. La variable $\mathbf{1}_A$ est donc une variable de Bernoulli de paramètre $\mathbb{P}(A)$.

Propriété 3.2.3. Une variable aléatoire X de loi de Bernoulli de paramètre p vérifie

$$\mathbb{E}X = p \text{ et } \mathbb{V}(X) = p(1 - p).$$

3.2.3 Loi binomiale

On veut étudier la répétition de plusieurs expériences identiques et indépendantes. Considérons donc n variables aléatoires X_1, \dots, X_n indépendantes de loi de Bernoulli de même paramètre p . On note S_n la somme de ces n variables : $S_n = \sum_{k=1}^n X_k$. Autrement dit, on considère une expérience qui réussit avec probabilité p , et on réalise cette expérience n fois, de manière indépendante. La variable S_n compte le nombre de succès parmi les n . Notamment, S_n prend ses valeurs dans $\{0, \dots, n\}$.

Propriété 3.2.4. On considère la variable S_n définie ci-dessus. Pour k dans $\{0, \dots, n\}$ on a

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

(la notation $\binom{n}{k}$ est définie en partie 3.3.3). La loi de S_n est appelée *loi binomiale* de paramètres n et p .

Démonstration. On a

$$\{S_n = k\} = \left\{ \text{card}\{i, X_i = 1\} = k \right\} = \bigcup_{\substack{I \subset \{1, \dots, n\} \\ \text{card} I = k}} \left\{ \{i, X_i = 1\} = I \right\},$$

où la dernière union est disjointe. Par conséquent, par indépendance des X_i

$$\begin{aligned} \mathbb{P}(S_n = k) &= \sum_{\substack{I \subset \{1, \dots, n\} \\ \text{card} I = k}} \mathbb{P}(\{i, X_i = 1\} = I) = \sum_{\substack{I \subset \{1, \dots, n\} \\ \text{card} I = k}} \prod_{i \in I} p \prod_{i \notin I} (1 - p) = \sum_{\substack{I \subset \{1, \dots, n\} \\ \text{card} I = k}} p^k (1 - p)^{n-k} \\ &= \binom{n}{k} p^k (1 - p)^{n-k}. \end{aligned}$$

□

Propriété 3.2.5. Une variable aléatoire X de loi binomiale de paramètres n et p vérifie :

$$\mathbb{E}X = np \text{ et } \mathbb{V}X = np(1 - p).$$

Démonstration. Par définition, X à la même loi que $\sum_{k=1}^n Y_k$, où les Y_k sont indépendantes et de loi de Bernoulli de paramètre p . On a donc $\mathbb{E}X = \sum_{k=1}^n \mathbb{E}Y_k = n\mathbb{E}Y_1 = np$. Par ailleurs, l'indépendance des Y_k donne $\mathbb{V}X = \sum_{k=1}^n \mathbb{V}Y_k = n\mathbb{V}Y_1 = np(1 - p)$. □

Propriété 3.2.6. Si X et Y sont deux variables aléatoires indépendantes de loi binomiale de paramètres (n, p) et (m, p) , alors $X + Y$ est suit la loi binomiale de paramètre $(n + m, p)$.

3.2.4 Loi géométrique

On répète indéfiniment une expérience ayant une probabilité p de succès. Au bout de combien de temps va-t-on observer le premier succès? Pour modéliser cela, on considère $(X_n)_{n \in \mathbb{N}}$ une suite infinie^(note 1) de variables aléatoires indépendantes de loi de Bernoulli de paramètre p . On pose $T = \inf\{n \in \mathbb{N}, X_n = 1\}$.

Propriété 3.2.7. On a $\mathbb{P}(T = n) = p(1 - p)^n$, pour tout $n \in \mathbb{N}$.

Démonstration. L'évènement $\{T = n\}$ peut se récrire

$$\{T = n\} = \{X_n = 1\} \cap \bigcap_{i=0}^{n-1} \{X_i = 0\}.$$

Sa probabilité est donc, par indépendance,

$$\mathbb{P}(T = n) = \mathbb{P}\left(\{X_n = 1\} \cap \bigcap_{i=0}^{n-1} \{X_i = 0\}\right) = \mathbb{P}(X_n = 1) \times \prod_{i=0}^{n-1} \mathbb{P}(X_i = 0) = p(1 - p)^n.$$

□

Propriété 3.2.8. Une variable aléatoire T de loi géométrique de paramètre p est de carré intégrable et

$$\mathbb{E}T = \frac{1}{p}, \quad \mathbb{V}(T) = \frac{1 - p}{p^2}.$$

La loi géométrique vérifie la propriété suivante, dit d'absence de mémoire.

Propriété 3.2.9. Soient T une variable aléatoire de loi géométrique, et n et m deux entiers naturels. Alors

$$\mathbb{P}_{\{T \geq n\}}(T \geq n + m) = \mathbb{P}(T \geq m).$$

Démonstration. On peut commencer par remarquer que

$$\mathbb{P}(T \geq n) = \sum_{k=n}^{\infty} p(1 - p)^k = p(1 - p)^n \frac{1}{1 - (1 - p)} = (1 - p)^n.$$

On a donc

$$\mathbb{P}_{\{T \geq n\}}(T \geq n + m) = \frac{\mathbb{P}(\{T \geq n\} \cap \{T \geq n + m\})}{\mathbb{P}(T \geq n)} = \frac{\mathbb{P}(\{T \geq n + m\})}{\mathbb{P}(T \geq n)} = \frac{(1 - p)^{n+m}}{(1 - p)^n} = (1 - p)^m.$$

□

Propriété 3.2.10. Si X et Y sont deux lois géométriques de paramètres respectifs p et q , alors la variable $\min(X, Y)$ suit la loi géométrique de paramètre $1 - (1 - p)(1 - q) = p + q - pq$.

Démonstration. On écrit

$$\mathbb{P}(\min(X, Y) \geq n) = \mathbb{P}(X \geq n, Y \geq n) = (1 - p)^n (1 - q)^n = (1 - p - q + pq)^n.$$

Pour conclure, on remarque que

$$\mathbb{P}(\min(X, Y) = n) = \mathbb{P}(\min(X, Y) \geq n) - \mathbb{P}(\min(X, Y) \geq n + 1).$$

□

(note 1). On explique brièvement qu'une telle suite existe au début du chapitre 5.

3.2.5 Loi de Poisson

Soit θ un réel strictement positif. Si X est une variable aléatoire de loi binomiale de moyenne θ , alors ses paramètres sont nécessairement n et $\frac{\theta}{n}$. Quand n tend vers l'infini, on est donc en train de regarder un grand nombre de fois un phénomène qui a une chance très faible de se produire. On peut par exemple penser au nombre de cas d'une maladie rare dans un pays : on a une grande population, dont chaque individu est touché par la maladie avec une probabilité faible. En conséquence, on observe seulement quelques cas sur tout le territoire (avec les notations ci-dessus, en moyenne θ cas).

La probabilité d'observer k cas est :

$$\binom{n}{k} \left(\frac{\theta}{n}\right)^k \left(1 - \frac{\theta}{n}\right)^{n-k} = \frac{n(n-1)\dots(n-k+1)\theta^k}{k!n^k} \exp\left((n-k)\ln\left(1 - \frac{\theta}{n}\right)\right). \quad (3.2)$$

Le quotient $\frac{n(n-1)\dots(n-k+1)}{n^k}$ tend vers 1 quand n tend vers ∞ . Par ailleurs, on a le développement asymptotique en $n \rightarrow \infty$:

$$\exp\left((n-k)\ln\left(1 - \frac{\theta}{n}\right)\right) = \exp\left(-\left(n-k\right)\left(\frac{\theta}{n} + o\left(\frac{1}{n}\right)\right)\right) = \exp(-\theta + o(1)).$$

Par conséquent l'expression en (3.2) tend vers $\frac{\theta^k}{k!}e^{-\theta}$ quand $n \rightarrow \infty$. On est donc amené à la définition suivante.

Définition 3.2.11. Soit $\theta > 0$. On dit qu'une variable aléatoire X à valeurs dans \mathbb{N} suit la loi de Poisson ^(note 2) de paramètre θ si pour tout entier k , on a

$$\mathbb{P}(X = k) = e^{-\theta} \frac{\theta^k}{k!}.$$

Propriété 3.2.12. Si X et Y sont deux variables de loi de Poisson de paramètres θ et η , alors $X + Y$ suit une loi de Poisson de paramètre $\theta + \eta$.

Propriété 3.2.13. Si X suit la loi de Poisson de paramètre θ , alors $\mathbb{E}X = \theta$ et $\mathbb{V}X = \theta$.

3.3 Dénombrement

Le dénombrement consiste à calculer le cardinal d'un ensemble. La plupart du temps, cela revient à trouver une bijection entre l'ensemble dont on cherche le cardinal et un ensemble dont le cardinal est connu. En effet, deux ensembles en bijection ont même cardinal.

Nous présentons dans cette partie les cardinaux de divers ensembles "de référence".

3.3.1 Tirages ordonnés avec remise

Propriété 3.3.1. L'ensemble des k -uplets de $\{1, \dots, n\}$. a n^k éléments.

Démonstration. C'est une conséquence du fait que le cardinal de $E \times F$ est le produit des cardinaux de E et de F . □

(note 2). "Poisson" prend une majuscule, puisqu'il s'agit de Siméon Denis Poisson.

Propriété 3.3.2. *L'ensemble des fonctions de $\{1, \dots, k\}$ dans $\{1, \dots, n\}$ est en bijection avec l'ensemble des k -uplets de $\{1, \dots, n\}$.*

Démonstration. La bijection est donnée par la fonction qui, à un k -uplet (x_1, \dots, x_k) associe la fonction $i \mapsto x_i$. \square

3.3.2 Tirages ordonnés sans remise

Propriété 3.3.3. *L'ensemble des injections de $\{1, \dots, k\}$ dans $\{1, \dots, n\}$ a pour cardinal $\frac{n!}{(n-k)!}$.*

Démonstration. Par récurrence sur k . Pour $k = 1$, une injection de $\{1\}$ dans $\{1, \dots, n\}$ peut s'assimiler à un élément de $\{1, \dots, n\}$, on a donc bien $n = \frac{n!}{(n-1)!}$ éléments.

On suppose le résultat vrai pour un k fixé avec $1 \leq k < n$. Se donner une injection de $\{1, \dots, k+1\}$ dans $\{1, \dots, n\}$ revient à se donner une injection de $\{1, \dots, k\}$ dans $\{1, \dots, n\}$ et se donner l'image de $k+1$, qui doit être différente des k éléments déjà choisis. Il reste donc $n-k$ choix pour l'image de $k+1$, et on a donc $\frac{n!}{(n-k)!} \times (n-k) = \frac{n!}{(n-k-1)!}$ éléments. \square

Le cas particulier $k = n$ de la proposition 3.3.3 se récrit comme suit.

Corollaire 3.3.4. *L'ensemble des bijections de $\{1, \dots, n\}$ dans lui-même a $n!$ éléments.*

Propriété 3.3.5. *L'ensemble des injections de $\{1, \dots, k\}$ dans $\{1, \dots, n\}$ est en bijection avec l'ensemble des k -uplets d'éléments tous distincts de $\{1, \dots, n\}$.*

3.3.3 Tirages non ordonnés sans remise

Propriété 3.3.6. *Soient k et n deux entiers avec $0 \leq k \leq n$. L'ensemble des sous-ensembles à k éléments de $\{1, \dots, n\}$ a pour cardinal $\frac{n!}{k!(n-k)!}$. Cette valeur est appelée coefficient binomial et est noté $\binom{n}{k}$ ou C_n^k .*

Démonstration. On va montrer par récurrence sur n la propriété "Pour tout entier k avec $0 \leq k \leq n$, le nombre d'éléments est $\frac{n!}{k!(n-k)!}$ ".

Pour $n = 0$, on a un sous-ensemble à 0 éléments. Pour $n = 1$, on a un sous-ensemble à 0 éléments, et un sous-ensemble à un élément.

Supposons le résultat vrai pour un n fixé. Un sous-ensemble à k éléments de $\{1, \dots, n+1\}$ est :

- soit un sous-ensemble de $\{1, \dots, n\}$ à k éléments ;
- soit l'union de $\{n+1\}$ et d'un sous-ensemble de $\{1, \dots, n\}$ à $k-1$ éléments.

Le nombre d'éléments est donc

$$\begin{aligned} \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-k+1)!} &= \frac{n!}{(k-1)!(n-k)!} \left(\frac{1}{k} + \frac{1}{n-k+1} \right) \\ &= \frac{n!}{k!(n-k+1)!} \times (n+1) \\ &= \frac{(n+1)!}{k!(n-k+1)!}, \end{aligned}$$

ce qui achève la preuve. \square

Partitionnement

Propriété 3.3.7. Soit n un entier positif, et r_1, \dots, r_k des entiers tels que $r_1 + \dots + r_k = n$. L'ensemble des partitions de $\{1, \dots, n\}$ en k ensembles de cardinaux respectifs r_1, \dots, r_k , a pour cardinal

$$\frac{n!}{r_1! \cdots r_k!}.$$

On remarque que cas particulier $k = 2$ correspond à la propriété 3.3.6. En effet, choisir un sous-ensemble de $\{1, \dots, n\}$ à k éléments revient à choisir un sous-ensemble à k élément et un autre sous-ensemble à $n - k$ éléments (le complémentaire de l'ensemble initial).

Démonstration. On raisonne par récurrence sur k . Pour $k = 1$, on a alors $r_1 = n$ et il n'y a qu'un seul choix possible, or $1 = \frac{n!}{r_1!}$.

Supposons le résultat vrai pour un certain $k \geq 1$. On se donne $k + 1$ entiers vérifiant

$$r_1 + \dots + r_k + r_{k+1} = n.$$

Partitionner $\{1, \dots, n\}$ en $k + 1$ ensembles de cardinaux r_1, \dots, r_{k+1} revient à partitionner $\{1, \dots, n\}$ en k ensembles de cardinaux $r_1, \dots, r_{k-1}, (r_k + r_{k+1})$, puis à partitionner le dernier ensemble (à $r_k + r_{k+1}$ éléments) en un ensemble à r_k éléments et un à r_{k+1} éléments. Le nombre d'éléments est donc

$$\frac{n!}{r_1! \cdots r_{k-1}!(r_k + r_{k+1})!} \times \frac{(r_k + r_{k+1})!}{r_k!r_{k+1}!} = \frac{n!}{r_1! \cdots r_{k-1}!r_k!r_{k+1}!}.$$

□

3.3.4 Tirage non ordonné avec remise

Propriété 3.3.8. L'ensemble des n -uplets d'entiers (positifs ou nuls) dont la somme vaut k a $\binom{k+n-1}{k}$ éléments.

Démonstration. Fixons n et k . On va représenter l'entier k par une suite de k étoiles. Par exemple, pour $k = 11$

Pour représenter les entiers dont la somme fera k , on va séparer cet ensemble d'étoiles à l'aide des barres verticales :

*|***||**|****|*|

Ici, on a représenté la décomposition $1 + 3 + 0 + 2 + 4 + 1 + 0 = 11$. Dans cet exemple, on a tracé 6 barres verticales, pour obtenir une décomposition en somme de $n = 7$ entiers. Chaque disposition de k étoiles et $n - 1$ barres fournit une décomposition de k en somme de n entiers. On a donc $k + n - 1$ objets à placer, k étoiles et $n - 1$ barres. La propriété 3.3.6 nous dit qu'il y a $\frac{(k+n-1)!}{k!(n-1)!}$ tracés possibles. □

La situation de la propriété 3.3.8 correspond à un tirage non ordonné avec remise de k éléments parmi $\{1, \dots, n\}$. Autrement dit, on peut tirer un même élément plusieurs fois, et on ne s'intéresse qu'au nombre de fois où chaque élément a été tiré, et pas à l'ordre dans lequel ils ont été tirés.

En effet, le k -uplet (r_1, \dots, r_k) correspond au tirage non ordonné où l'on a tiré r_i fois le nombre i .

Chapitre 4

Variables aléatoires continues

4.1 Variables aléatoires réelles à densité

4.1.1 Définitions

On rappelle que l'ensemble des mesures de probabilité sur \mathbb{R} est en bijection avec l'ensemble des fonctions de répartition. Comme l'ensemble des fonctions de répartition est essentiellement l'ensemble des fonctions croissantes, les fonctions de répartition dérivable ont une dérivée positive. Plus précisément on peut énoncer la propriété suivants :

Propriété 4.1.1. *Soit F une fonction de répartition. Si il existe une fonction $p : \mathbb{R} \rightarrow \mathbb{R}$ telle que pour tout réel x on ait*

$$F(x) = \int_{-\infty}^x p(t)dt, \quad (4.1)$$

alors la fonction p est positive et vérifie $\int_{\mathbb{R}} p(t)dt = 1$.

Inversement, si $p : \mathbb{R} \rightarrow \mathbb{R}$ est une fonction positive d'intégrale 1, alors, la fonction F définie par la relation (4.1) est une fonction de répartition.

On a donc caractérisé un sous-ensemble de l'ensemble des mesures sur \mathbb{R} , que l'on appellera l'ensemble des mesures à densité :

Définition 4.1.2. *Une variable aléatoire X est dite à densité, s'il existe une fonction p_X telle que la loi de X vérifie :*

$$\mathbb{P}(X \leq x) = \int_{-\infty}^x p_X(t)dt.$$

La fonction p_X est alors appelée densité de X .

Propriété 4.1.3. *Pour une variable aléatoire X à densité, on a les résultats suivants :*

1. La fonction de répartition F_X est continue sur \mathbb{R} . En particulier $\mathbb{P}(X = x) = 0$, pour tout réel x ;
2. Pour tout intervalle I de \mathbb{R} , on a

$$\mathbb{P}(X \in I) = \int_I p_X(x)dx ;$$

3. La fonction de répartition F_X est dérivable partout où la densité p_X est continue et, en ces points, $F'_X = p_X$.

En pratique, la relation à retenir est $F'_X = p_X$. On a notamment le résultat suivant :

Propriété 4.1.4. Soit X une variable aléatoire de fonction de répartition F_X . Si l'application F_X est continue sur \mathbb{R} et \mathcal{C}^1 par morceaux, alors X est une variable aléatoire à densité, de densité égale à F'_X partout où elle est dérivable.

La densité d'une variable aléatoire à densité peut également se caractériser à partir de calculs d'espérance. En effet, une fonction f en escaliers peut s'écrire

$$f(x) = \sum_{k=1}^n \alpha_k \mathbf{1}_{]a_{k-1}, a_k[},$$

avec $-\infty = a_0 < a_1 < a_2 \dots < a_{n-1} < a_n = \infty$. On a alors

$$\mathbb{E}f(X) = \sum_{k=1}^n f(\alpha_k) \mathbb{P}(a_{k-1} < X < a_k) = \sum_{k=1}^n f(\alpha_k) \int_{a_{k-1}}^{a_k} p(x) dx = \int_{\mathbb{R}} f(x) p(x) dx.$$

On admettra que ce résultat s'étend à un f quelconque.

Théorème 4.1.5 (Théorème de transfert, version continue). Si f est une fonction positive ou bornée, alors

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

Démonstration. Admis. □

On a même une réciproque au théorème 4.1.5.

Propriété 4.1.6. Soit X une variable aléatoire telle qu'il existe une fonction $p : \mathbb{R} \rightarrow \mathbb{R}$ telle que pour toute fonction f positive, on ait

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

Alors X admet p pour densité.

Démonstration. En prenant $f = \mathbf{1}_{]-\infty, t]}$, on trouve

$$\mathbb{P}(X \leq t) = \mathbb{E}f(X) = \int_{-\infty}^t p(x) dx,$$

ce qui correspond à la définition d'une variable à densité. □

On peut également caractériser l'indépendance à partir de la densité de variables aléatoires.

Propriété 4.1.7. Deux variables X et Y de densités respectives p_X et p_Y sont indépendantes si et seulement si pour toute fonction f positive (ou telle que $f(X, Y)$ est intégrable) définie sur \mathbb{R}^2 , on a

$$\mathbb{E}f(X, Y) = \int_{\mathbb{R}^2} f(x, y) p_X(x) p_Y(y) dx dy. \quad (4.2)$$

Démonstration. Pour montrer l'indépendance à partir de la relation (4.2), il suffit de choisir $f(x, y) = \mathbf{1}_A(x)\mathbf{1}_B(y)$, auquel cas on a

$$\begin{aligned} \mathbb{P}(X \in A, Y \in B) &= \mathbb{E}f(X, Y) = \int_{\mathbb{R}^2} \mathbf{1}_A(x)\mathbf{1}_B(y)p_X(x)p_Y(y)dxdy = \int_{\mathbb{R}} \mathbf{1}_A(x)p_X(x)dx \int_{\mathbb{R}} p_Y(y)\mathbf{1}_B(y)dy \\ &= \mathbb{P}(X \in A)\mathbb{P}(X \in B). \end{aligned}$$

La réciproque est admise. \square

4.2 Lois à densité usuelles

Dans cette partie on présente les lois à densité les plus usuelles, avec certaines de leurs propriétés.

4.2.1 Loi uniforme

Définition 4.2.1. Soit $a < b$ deux réels. On dit qu'une variable aléatoire suit la loi uniforme sur l'intervalle $[a, b]$ si elle admet pour densité

$$\frac{1}{b-a}\mathbf{1}_{[a,b]}.$$

Le nom de loi *uniforme* vient du fait que la probabilité que X appartienne à un sous-intervalle I de $[a, b]$ est proportionnelle à la longueur de J .

Propriété 4.2.2. Une variable de loi uniforme sur $[a, b]$ est bornée et vérifie

$$\mathbb{E}X = \frac{a+b}{2} \text{ et } \mathbb{V}X = \frac{(b-a)^2}{12}.$$

4.2.2 Loi de Gauss ou loi normale

Définition 4.2.3. Une variable aléatoire réelle X suit une loi de Gauss ou loi normale centrée réduite si elle admet pour densité l'application p_X définie sur \mathbb{R} par :

$$\forall x \in \mathbb{R}, \quad p_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

Plus généralement, on définit la loi normale de moyenne m et de variance $\sigma^2 > 0$, notée $\mathcal{N}(m, \sigma^2)$, comme la loi dont la densité est donnée par

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

On verra au chapitre 5 le théorème limite central 5.3.1, qui fait apparaître très naturellement la loi normale.

Propriété 4.2.4. Une variable aléatoire de loi normale de paramètres μ et σ^2 est de carré intégrable et on a

$$\mathbb{E}X = \mu \text{ et } \mathbb{V}X = \sigma^2.$$

Propriété 4.2.5. Soient X et Y deux variables aléatoires indépendantes de loi normales de paramètres respectifs (m_X, σ_X^2) et (m_Y, σ_Y^2) , et soient a et b deux réels. Alors

- la variable aléatoire $aX + b$ est de loi normale de moyenne $am_X + b$ et de variance $a^2\sigma_X^2$.
- la variable aléatoire $X + Y$ est de loi normale de moyenne $m_X + m_Y$ et de variance $\sigma_X^2 + \sigma_Y^2$.

Propriété 4.2.6. Soit X une variable aléatoire de loi normale centrée réduite et n un entier positif. Alors :

$$\mathbb{E}X^n = \begin{cases} 0 & \text{si } n \text{ est impair,} \\ (n-1) \times (n-3) \dots \times 1 = \frac{(2k)!}{2^k k!} & \text{si } n = 2k \text{ est pair.} \end{cases}$$

Démonstration. Le cas n vient du fait que $\mathbb{E}X^n$ est donné par l'intégrale sur \mathbb{R} d'une fonction impaire. Pour le cas n pair, on procède par récurrence, en intégrant par parties (on dérive x^{n+1} et on primitive $xe^{-x^2/2}$) :

$$\sqrt{2\pi}\mathbb{E}(X^{n+2}) = \int_{\mathbb{R}} x^{n+2} e^{-x^2/2} dx = \left[-x^{n+1} e^{-x^2/2} \right]_{-\infty}^{\infty} + \int_{\mathbb{R}} (n+1)x^n e^{-x^2/2} dx = (n+1)\mathbb{E}(X^n).$$

□

4.2.3 Loi exponentielle

Définition 4.2.7. La variable aléatoire X suit une loi exponentielle de paramètre $\lambda > 0$, si elle admet pour densité l'application p_X définie par :

$$\forall x \in \mathbb{R}, p_X(x) = \mathbf{1}_{[0, +\infty[}(x) \lambda e^{-\lambda x}.$$

En vertu de la propriété suivante, on dit que la loi exponentielle est une loi “sans mémoire”.

Propriété 4.2.8. Soit X une variable aléatoire de loi exponentielle (de paramètre quelconque). Alors, pour tout $t_0 > 0$ et $t > 0$, on a

$$\mathbb{P}(X > t_0 + t | X > t_0) = \mathbb{P}(X > t).$$

Démonstration. On remarque tout d'abord que

$$\mathbb{P}(X > t) = \int_t^{\infty} \lambda e^{-\lambda s} ds = [-e^{-\lambda s}]_t^{\infty} = e^{-\lambda t}.$$

On a donc

$$\mathbb{P}(X > t_0 + t | X > t_0) = \frac{\mathbb{P}(X > t_0 + t)}{\mathbb{P}(X > t_0)} = \frac{e^{-\lambda(t_0+t)}}{e^{-\lambda t_0}} = e^{-\lambda t},$$

d'où le résultat. □

On remarquera que la propriété 4.2.8 ainsi que sa preuve sont presque des recopies de la propriété 3.2.9 qui concernait la loi géométrique. En fait, la loi exponentielle est un équivalent continu de la loi géométrique, et ces deux lois partagent de nombreuses propriétés.

Propriété 4.2.9. Si X et Y sont deux variables aléatoires de loi exponentielle de paramètre respectifs λ et μ , alors $\min(X, Y)$ suit la loi exponentielle de paramètre $\lambda + \mu$.

Démonstration. On calcule, pour $t \geq 0$:

$$\mathbb{P}(\min(X, Y) > t) = \mathbb{P}(X > t, Y > t) = \mathbb{P}(X > t)\mathbb{P}(Y > t) = e^{-\lambda t}e^{-\mu t} = e^{-(\lambda+\mu)t}.$$

La fonction de répartition est alors $F(t) = 1 - e^{-(\lambda+\mu)t}$ (pour $t \geq 0$, et nulle pour $t < 0$). Par la propriété 4.1.4 montre alors le résultat. \square

En raisonnant par récurrence on montrerait que les X_i , $1 \leq i \leq n$ sont des variables aléatoires indépendantes de loi exponentielle de paramètres $\lambda_1, \dots, \lambda_n$ respectivement, alors $\min(X_1, \dots, X_n)$ suit la loi exponentielle de paramètre $\lambda_1 + \dots + \lambda_n$.

Propriété 4.2.10. *Une variable aléatoire de loi exponentielle de paramètre λ est de carré intégrable et on a*

$$\mathbb{E}X = \frac{1}{\lambda} \text{ et } \mathbb{V}X = \frac{1}{\lambda^2}.$$

Propriété 4.2.11. *Si U est une variable aléatoire de loi uniforme sur $[0, 1]$, alors la variable $-\frac{1}{\lambda} \ln U$ suit la loi exponentielle de paramètre λ .*

Propriété 4.2.12. *Soit X une variable aléatoire de loi exponentielle de paramètre λ et n un entier naturel. Alors*

$$\mathbb{E}(X^n) = \frac{n!}{\lambda^n}.$$

4.2.4 Loi de Cauchy

Définition 4.2.13. *La variable aléatoire X suit une loi de Cauchy standard si elle admet pour densité l'application p_X définie sur \mathbb{R} par*

$$p_X(x) = \frac{1}{\pi(1+x^2)}.$$

On parle également de loi de Cauchy de paramètre c si la densité est donnée par

$$p_X(x) = \frac{c}{\pi(c^2+x^2)}.$$

Le paramètre de la loi de Cauchy correspond à changement d'échelle :

Propriété 4.2.14. *Si X est une variable aléatoire de loi de Cauchy de paramètre c , alors λX suit la loi de Cauchy de paramètre λc .*

Si X et Y sont deux variables aléatoires de loi de Cauchy de paramètre respectifs c et c' . Alors $X + Y$ suit une loi de Cauchy de paramètre $c + c'$.

La loi de Cauchy n'est pas particulièrement utile, mais elle a la particularité suivante :

Propriété 4.2.15. *Une variable aléatoire de loi de Cauchy n'est pas intégrable.*

Démonstration. Si X suit la loi de Cauchy (de paramètre 1), alors $\mathbb{E}|x| = \int_{\mathbb{R}} \frac{|x|}{1+x^2} dx$. Or, $\frac{|x|}{1+x^2}$ est équivalent à ∞ quand x tend vers $\pm\infty$. Par conséquent cette fonction n'est pas intégrable et $\mathbb{E}|X| = \infty$. \square

Une manière "naturelle" de faire apparaître la loi de Cauchy est la suivante :

Propriété 4.2.16. Soit Θ une variable aléatoire de loi uniforme sur $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Alors $\tan(\Theta)$ suit la loi de Cauchy.

La propriété précédente s'interprète géométriquement de la manière suivante : sur la figure 4.1, le point C est situé à une distance 1 de la droite \mathcal{D} . Le point O est le projeté orthogonal de C sur \mathcal{D} . On tire un angle Θ uniformément sur $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Il existe un unique point X de \mathcal{D} tel que l'angle (\vec{CO}, \vec{CX}) soit égal à Θ . La distance algébrique OX suit alors une loi de Cauchy.

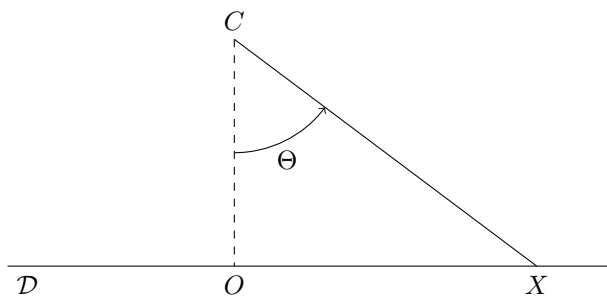


FIGURE 4.1 – Interprétation géométrique de la loi de Cauchy.

Chapitre 5

Suites de variables aléatoires - théorèmes limites

5.1 Suites des variables aléatoires

On voudrait étudier la situation où un même phénomène aléatoire se répète indéfiniment, et de manière indépendante à chaque fois. Pour cela, il faut construire une suite (infinie) de variables de même loi indépendante les unes des autres. Il n'est pas évident a priori qu'une telle suite de variables puisse exister. D'après, le théorème suivant, c'est effectivement le cas

Théorème 5.1.1. *Soit $(\mu_n)_{n \in \mathbb{N}}$ une suite de mesures de probabilité sur un ensemble E . Alors il existe un espace probabilisé (Ω, \mathbb{P}) et une suite de variables aléatoires $(X_n)_{n \in \mathbb{N}}$ à valeurs dans E telles que les X_n soient indépendantes et que pour tout n , X_n ait pour loi μ_n .*

Démonstration. Admis. Il s'agirait de construire une mesure sur l'ensemble indénombrable $E^{\mathbb{N}}$ (ensemble des suites à valeurs dans E). Comme il a déjà été dit, construire une mesure sur un ensemble indénombrable peut être très complexe, et on ne le fera pas ici. \square

5.1.1 Convergence de suites de variables aléatoires réelles

Dans cette partie, on considère une suite de variables aléatoires réelles $(X_n)_{n \in \mathbb{N}}$, et on veut donner un sens à la phrase " X_n converge". Il existe en fait plusieurs sens possibles à donner à cette affirmation.

La définition la plus simple a priori revient à dire que toute réalisation de cette suite est une suite convergente. C'est par exemple ce que l'on observait sur la figure 0.2.

Définition 5.1.2. *On dit que la suite $(X_n)_{n \in \mathbb{N}}$ converge presque sûrement vers une variable aléatoire X si*

$$\mathbb{P}(X_n \text{ converge vers } X) = 1.$$

Ici, $\{X_n \text{ converge vers } X\}$ désigne un évènement à savoir $\{\omega \in \Omega, X_n(\omega) \text{ converge vers } X(\omega)\}$.

Toutefois, cette notion est souvent difficile à prouver, et on va plutôt utiliser des notions plus faibles. La seconde revient à dire que, pour n grand, X_n a de grandes chances d'être proche de X .

Définition 5.1.3. On dit que $(X_n)_{n \geq 1}$ converge en probabilité vers X si pour tout $\varepsilon > 0$, on a

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X - X_n| < \varepsilon) = 1.$$

Le résultat suivant relie les deux notions précédentes.

Propriété 5.1.4. Si $(X_n)_{n \in \mathbb{N}}$ converge presque sûrement vers X , alors elle converge aussi en probabilité vers X .

Démonstration. L'évènement $\{X_n \text{ converge vers } X\}$ peut se récrire

$$\{X_n \text{ converge vers } X\} = \{\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |X_n - X| < \varepsilon\} = \bigcap_{\varepsilon > 0} \bigcup_{N \in \mathbb{N}} \bigcap_{n \geq N} \{|X_n - X| < \varepsilon\}.$$

Comme cet évènement est de probabilité 1, pour tout $\varepsilon > 0$, l'évènement plus grand

$$\bigcup_{N \in \mathbb{N}} \bigcap_{n \geq N} \{|X_n - X| < \varepsilon\}$$

est aussi de probabilité 1. Or, comme il s'agit de l'union d'une famille croissante d'évènements, on a

$$1 = \mathbb{P} \left(\bigcup_{N \in \mathbb{N}} \bigcap_{n \geq N} \{|X_n - X| < \varepsilon\} \right) = \lim_N \mathbb{P} \left(\bigcap_{n \geq N} \{|X_n - X| < \varepsilon\} \right).$$

Or $\mathbb{P} \left(\bigcap_{n \geq N} \{|X_n - X| < \varepsilon\} \right) \leq \mathbb{P}(|X_n - X| < \varepsilon)$, ce qui montre que $\mathbb{P}(|X_n - X| < \varepsilon)$ tend vers 1 pour tout ε . \square

La réciproque de la propriété 5.1.4 n'est pas vraie, comme le montre le contre-exemple que nous allons construire maintenant. On considère une suite de variables aléatoires indépendantes $(X_n)_{n \in \mathbb{N}}$ telle que X_n suive la loi uniforme sur $\{2^n, \dots, 2^{n+1} - 1\}$. On définit alors, pour tout entier k ,

$$Y_k = \begin{cases} 1 & \text{si } X_n = k \text{ pour un certain } n, \\ 0 & \text{sinon.} \end{cases}$$

Expliquons un peu cette construction : le choix des $(X_n)_{n \in \mathbb{N}}$ revient à choisir un et un seul élément uniformément dans chacun des ensembles $\{1\}$, $\{2, 3\}$, $\{4, 5, 6, 7\}$, $\{8, \dots, 15\}$, ..., $\{2^n, \dots, 2^{n+1} - 1\}$, etc, qui constituent une partition de \mathbb{N} .

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
X_n	×		×		×					×						...
Y_k	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0	

Dans le tableau ci-dessus, les croix symbolisent les valeurs qui ont été choisies pour les X_n . Ici, $X_0 = 1$, $X_1 = 3$, $X_2 = 5$, $X_3 = 10$, etc. La suite $(Y_k)_{k \in \mathbb{N}}$ est la suite de 0 et de 1 en dernière ligne. On remarque alors que :

- La suite $(Y_k)_{k \in \mathbb{N}}$ ne converge pas presque sûrement vers 0. En effet, pour tout n , il existe toujours une valeur k de $\{2^n, \dots, 2^{n+1} - 1\}$ telle que $Y_k = 1$.
- La suite $(Y_k)_{k \in \mathbb{N}}$ converge en probabilité vers 0. En effet, soit $\varepsilon > 0$. Si $\varepsilon \geq 1$, on a $\mathbb{P}(|Y_k - 0| > \varepsilon) = 0$. Si, $\varepsilon < 1$, pour k un entier fixé, si n est l'entier tel que $2^n \leq k < 2^{n+1}$, alors

$$\mathbb{P}(|Y_k - 0| > \varepsilon) = \mathbb{P}(Y_k = 1) = 2^{-n} \xrightarrow{k \rightarrow \infty} 0.$$

Dans tous les cas $\mathbb{P}(|Y_k - 0| > \varepsilon)$ tend vers 0.

5.2 Loi des grands nombres

Le théorème suivant est la justification théorique des observations faites au chapitre 0 sur la figure 0.2.

Théorème 5.2.1 (Loi (forte) des grands nombres). *Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires indépendantes, de même loi et intégrables. La suite $\frac{1}{n} \sum_{k=1}^n X_k$ converge presque sûrement vers $\mathbb{E}X_1$.*

Remarquons que l'hypothèse " X_n intégrable" est tout ce dont on a besoin pour donner un sens à $\mathbb{E}X_1$. La preuve de ce théorème est complexe, nous ne démontrons que la version plus faible suivante.

Théorème 5.2.2 (Loi (faible) des grands nombres). *Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires indépendantes, de même loi et de carrés intégrables. La suite $\frac{1}{n} \sum_{k=1}^n X_k$ converge en probabilité vers $\mathbb{E}X_1$.*

Remarquons que le théorème 5.2.2 est plus faible que le théorème 5.2.1 à deux titres :

- ses hypothèses sont plus fortes : toute variable de carré intégrable est intégrable ;
- sa conclusion est plus faible : la convergence presque sûre entraîne la convergence en probabilité.

Par conséquent, le théorème 5.2.2 n'a en fait que peu d'intérêt par rapport au théorème 5.2.1, si ce n'est d'avoir une démonstration beaucoup plus simple.

Démonstration. On remarque que la moyenne de la variable $M_n = \frac{1}{n} \sum_{k=1}^n X_k$ est $\mathbb{E}X_1$. En effet,

$$\mathbb{E} \left(\frac{1}{n} \sum_{k=1}^n X_k \right) = \frac{1}{n} \sum_{k=1}^n \mathbb{E}X_k = \frac{1}{n} \sum_{k=1}^n \mathbb{E}X_1 = \mathbb{E}X_1.$$

De plus, la variance de M_n est

$$\mathbb{V} \left(\frac{1}{n} \sum_{k=1}^n X_k \right) = \frac{1}{n^2} \sum_{k=1}^n \mathbb{V}X_k = \frac{1}{n^2} \sum_{k=1}^n \mathbb{V}X_1 = \frac{\mathbb{V}(X_1)}{n}$$

(la variance de la somme est ici la somme des variance grâce à l'*indépendance* des X_i).

Soit $\varepsilon > 0$; l'inégalité de Bienaymé-Tchebychev nous donne

$$\mathbb{P}(|M_n - \mathbb{E}X_1| > \varepsilon) \leq \frac{\mathbb{V}(X_1)}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0,$$

ce qui achève la démonstration. □

5.3 Théorème limite central

Au vu de l'énoncé de la loi des grands nombres, une question légitime est maintenant de savoir à quelle vitesse la convergence de $\frac{1}{n} \sum_{k=1}^n X_k$ vers $\mathbb{E}X_1$ a lieu. La réponse à cette question est donnée par le théorème suivant :

Théorème 5.3.1 (Théorème limite central ^(note 1)). *Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires de carrés intégrables, indépendantes, de même loi. On suppose également que la variance des X_n est non*

(note 1). Le nom de "théorème limite central" est une traduction de l'allemand "Zentraler Grenzwertsatz". Le terme "Grenzwertsatz" désigne un "théorème limite", c'est-à-dire un résultat de convergence d'une suite de variables aléatoires. Le nom de théorème limite *central* signifie ici que l'on a affaire à un théorème limite particulièrement important. On trouve parfois d'autres traductions, comme "théorème central limite" ou "théorème de la limite centrée".

nulle. On pose $M_n = \frac{1}{n} \sum_{k=1}^n X_k$. Alors pour tout intervalle I , on a

$$\mathbb{P} \left(\sqrt{\frac{n}{\mathbb{V}(X)}} (M_n - \mathbb{E}X_1) \in I \right) \xrightarrow{n \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_I e^{-x^2/2} dx = \mathbb{P}(G \in I),$$

où G est une variable de loi normale centrée réduite.

Ce théorème a d'abord été démontré lorsque la loi commune des variables est une loi de Bernoulli de paramètre p (le théorème est dans ce cas connu sous le nom de théorème de Moivre-Laplace^(note 2)). Dans ce cas, la variable $\sum_{k=1}^n X_k$ suit une loi binomiale de paramètres n et p . Cela explique que la distribution vue en figure 0.6 présente un profil proche de celui d'une Gaussienne.

Démonstration. Admis. On peut toutefois remarquer que la normalisation

$$\sqrt{\frac{n}{\mathbb{V}(X)}} (M_n - \mathbb{E}X_1)$$

est la bonne pour avoir une variable de moyenne 0 et de variance 1, comme la variable G qui apparaît sur le membre de droite. \square

Ce théorème dit que la variable $\sqrt{\frac{n}{\mathbb{V}(X)}} (M_n - \mathbb{E}X_1)$ a "presque la même loi" qu'une loi normale centrée réduite. On parle en fait de convergence *en loi*. On peut également le récrire sous la forme suivante, non rigoureuse mais plus intuitive :

$$\frac{1}{n} \sum_{k=1}^n X_k \simeq \mathbb{E}X_1 + \frac{1}{\sqrt{n}} \times \sigma(X)G. \quad (5.1)$$

Autrement dit, la moyenne empirique est égale à l'espérance, avec une erreur à peu près Gaussienne de l'ordre de $\frac{1}{\sqrt{n}}$.

Si on multiplie (5.1) par n , on a l'expression

$$\sum_{k=1}^n X_k \simeq n \times \mathbb{E}X_1 + \sqrt{n} \times \sigma(X)G,$$

qui signifie qu'une somme de variables aléatoires de carré intégrable et de même loi se décompose en une partie déterministe d'ordre n , et d'une partie aléatoire d'ordre \sqrt{n} . Ceci est conforme à ce qu'on avait observé au chapitre 0, voir par exemple l'équation (0.1).

Dans le chapitre 0, on avait toutefois observé le cas de la marche aléatoire de Cauchy qui ne vérifiait pas cette décomposition. La raison en est que pour la marche aléatoire de Cauchy, les variables indépendantes et de même loi qui apparaissent sont de loi de Cauchy, et ne sont donc pas intégrables. Le comportement de l'ordre de n que l'on avait observé est dû au fait que la somme de n variables de loi de Cauchy de paramètre 1 indépendantes suit la loi de Cauchy de paramètre n (voir propriété 4.2.14).

(note 2). D'après *Abraham de Moivre*, il s'agit donc bien du théorème de Moivre-Laplace et non du théorème de Moivre-Laplace.

5.4 Chaînes de Markov

Dans cette partie, nous allons effleurer la très riche théorie des phénomènes aléatoires dit *Markoviens*. Le principe est le suivant : une chaîne de Markov est une suite $(X_n)_{n \in \mathbb{N}}$ de variables aléatoires, qui n'est pas composée de variables indépendantes, mais telle que "la loi de X_{n+1} ne dépend que de la valeur de X_n ". Donnons une définition un peu plus précise :

Définition 5.4.1. Soit E un ensemble fini. Une chaîne de Markov sur E est une suite de variables aléatoires $(X_n)_{n \in \mathbb{N}}$ telle que pour tous x_0, \dots, x_n de E , on a

$$\mathbb{P}(X_n = x_n | X_0 = x_0, \dots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}).$$

On résume parfois cette définition par la phrase (pas forcément très parlante) "conditionnellement au présent, le passé et le futur sont indépendants". On parle aussi de phénomènes *sans mémoire*.

Une conséquence de la définition est que la loi de la suite $(X_n)_{n \in \mathbb{N}}$ est entièrement caractérisée par la donnée des $\mathbb{P}(X_n = y | X_{n-1} = x)$ et par la loi de X_0 . En effet :

Propriété 5.4.2. Soit x_0, \dots, x_n des éléments de E . Alors

$$\mathbb{P}(X_n = x_n, \dots, X_0 = x_0) = \mathbb{P}(X_0 = x_0) \prod_{k=1}^n \mathbb{P}(X_k = x_k | X_{k-1} = x_{k-1}).$$

C'est de ce résultat que vient le nom de *chaîne* de Markov : les relations entre X_n et X_{n+1} caractérisent la loi de toute la suite, comme la solidité d'une chaîne vient des jonctions entre deux maillons consécutifs.

Démonstration. C'est une récurrence sur n . Pour $n = 0$, il n'y a rien à démontrer.

Si le résultat est vrai pour un n donné, alors

$$\begin{aligned} \mathbb{P}(X_{n+1} = x_{n+1}, \dots, X_0 = x_0) &= \mathbb{P}(X_n = x_n, \dots, X_0 = x_0) \times \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0) \\ &= \mathbb{P}(X_0 = x_0) \prod_{k=1}^n \mathbb{P}(X_k = x_k | X_{k-1} = x_{k-1}) \times \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n), \end{aligned}$$

d'où le résultat. □

Dans la suite, on se limitera au cas où $\mathbb{P}(X_n = y | X_{n-1} = x)$ ne dépend pas de n . On notera alors

$$P_{x,y} = \mathbb{P}(X_1 = y | X_0 = x).$$

P peut être vu comme une matrice de taille $N \times N$, où N est le nombre d'éléments de E . On l'appelle la *matrice de transition* de la chaîne de Markov. Cette matrice permet de calculer la loi de chacun des X_n .

Propriété 5.4.3. Si on note $\mu_x = \mathbb{P}(X_0 = x)$ la loi de X_0 , et si on voit μ comme un vecteur ligne, alors la loi de X_n est donnée par

$$\mathbb{P}(X_n = x) = (\mu P^n)_x.$$

On a noté μP le produit d'un vecteur ligne par une matrice :

$$(\mu P)_x = \sum_{y \in E} \mu_y P_{y,x}.$$

Démonstration. Comme on a supposé que $\mathbb{P}(X_{n+1} = y | X_n = x)$ ne dépendait pas de n , il suffit de montrer que la loi de X_1 est donnée par μP . Cela se prouve en utilisant la propriété 5.4.2 :

$$\mathbb{P}(X_1 = x) = \sum_{y \in E} \mathbb{P}(X_0 = y, X_1 = x) = \sum_{y \in E} \mathbb{P}(X_0 = y) \mathbb{P}(X_1 = x | X_0 = y) = \sum_{y \in E} \mu_y P_{y,x} = \mu P.$$

□

Le théorème suivant donne de nombreuses informations sur le comportement asymptotique d'une chaîne de Markov.

Théorème 5.4.4. *Soit P une matrice de transition sur un ensemble E fini, μ une probabilité sur E et $(X_n)_{n \in \mathbb{N}}$ une chaîne de Markov de matrice de transition P telle que X_0 soit de loi μ .*

1. *Si μ vérifie $\mu P = \mu$, alors X_n est de loi μ pour tout n . On dit d'un tel μ qu'il s'agit d'une mesure invariante.*
2. *Une chaîne de Markov admet au moins une mesure invariante.*
3. *Si il existe une unique mesure invariante $\bar{\mu}$, alors $\frac{1}{n} \sum_{k=1}^n \mathbf{1}_{\{X_k=x\}}$ converge presque sûrement vers $\bar{\mu}_x$.*
4. *Si 1 est la seule valeur propre de module 1 de P , alors il existe une unique mesure invariante $\bar{\mu}$ et la loi de $\mathbb{P}(X_n = x)$ converge vers $\bar{\mu}_x$ pour tout x .*

Démonstration. Admis.

□

Chapitre 6

Estimation

6.1 Estimation d'une proportion

On cherche à prévoir le résultat d'une élection : on a une population de N individus à laquelle sont proposés plusieurs candidats. Monsieur Michu est candidat à cette élection et veut avoir une estimation de son score $p \in [0, 1]$ (la proportion des N individus qui vont voter pour lui). Une méthode pour obtenir une telle estimation est d'interroger n personnes choisies aléatoirement dans la population, et de calculer la proportion \hat{p}_n de personnes qui vont voter pour monsieur Michu parmi ces n personnes.

Formalisons un peu le problème : pour $1 \leq k \leq n$, on va définir le réel X_k par $X_k = 1$ si la k ème personne choisie aléatoirement va voter pour monsieur Michu, et $X_k = 0$ dans le cas contraire. Si les personnes sont choisies indépendamment et uniformément dans la population ^(note 1), les variables X_k sont alors des variables de Bernoulli indépendantes de paramètres p .

L'estimation de la proportion p est donnée par

$$\hat{p}_n = \frac{1}{n} \sum_{k=1}^n X_k.$$

On dit que \hat{p}_n est un *estimateur* de p . On reconnaît la moyenne empirique d'une suite variables aléatoires. D'après la loi des grands nombres, on a convergence de \hat{p}_n vers p .

Théorème 6.1.1. *L'estimateur \hat{p}_n est convergent, au sens où \hat{p}_n converge presque sûrement et en probabilité vers p :*

$$\mathbb{P}(\hat{p}_n \rightarrow p) = 1 \text{ et } \forall \varepsilon > 0, \mathbb{P}(|\hat{p}_n - p| > \varepsilon) \rightarrow_n 0.$$

Toutefois, le résultat donné par \hat{p}_n n'a aucune raison d'être *égal* à p , et une question naturelle que monsieur Michu va se poser est "Quelle fiabilité puis-je accorder à ce résultat?". La réponse est une conséquence du théorème limite central :

Théorème 6.1.2. *Soit $c > 0$. Alors $\mathbb{P}(|\hat{p}_n - p| \leq \frac{c}{2\sqrt{n}})$ converge et on a*

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(|\hat{p}_n - p| \leq \frac{c}{2\sqrt{n}}\right) \geq \mathbb{P}(|G| \leq c) = \sqrt{\frac{2}{\pi}} \int_0^c e^{-x^2/2} dx,$$

(note 1). Cela implique en particulier qu'il serait possible d'interroger deux fois la même personne. Toutefois, cela n'arrive que rarement si le nombre de personnes interrogées est petit devant la taille de la population ($n \ll N$).

(G désigne une variable aléatoire de loi normale centrée réduite).

Démonstration. Les X_i sont ici des variables de loi de Bernoulli de paramètre p , dont la moyenne et la variance sont donc respectivement p et $p(1-p)$. D'après le théorème limite central, on a donc

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P} \left(|\hat{p}_n - p| \leq \frac{c}{2\sqrt{n}} \right) &= \lim_{n \rightarrow \infty} \mathbb{P} \left(\sqrt{\frac{n}{p(1-p)}} |\hat{p}_n - p| \leq \frac{c}{2\sqrt{p(1-p)}} \right) = \mathbb{P} \left(|G| \leq \frac{c}{2\sqrt{p(1-p)}} \right) \\ &\geq \mathbb{P}(|G| \leq c) \\ &= \sqrt{\frac{2}{\pi}} \int_0^c e^{-x^2/2} dx. \end{aligned}$$

L'inégalité découle ici du fait que $2\sqrt{p(1-p)} \leq 1$ pour tout $p \in [0, 1]$. □

Pour un c fixé, on sait donc que, asymptotiquement, la probabilité que p soit dans l'intervalle

$$\left[\hat{p}_n - \frac{c}{2\sqrt{n}}, \hat{p}_n + \frac{c}{2\sqrt{n}} \right]$$

est d'au moins $\sqrt{\frac{2}{\pi}} \int_0^c e^{-x^2/2} dx$.

En pratique, on se fixe une valeur $\alpha \in]0, 1[$, appelé le niveau de l'intervalle de confiance, et on cherche un $c > 0$ tel que

$$\sqrt{\frac{2}{\pi}} \int_0^c e^{-x^2/2} dx = \alpha.$$

Comme la fonction $c \mapsto \sqrt{\frac{2}{\pi}} \int_0^c e^{-x^2/2} dx$ est strictement croissante, envoie 0 sur 0 et converge vers 1 quand c tend vers $+\infty$, un tel c existe et est unique. Un choix courant est $\alpha = 0.95$, pour lequel $c = 1.96$ est une valeur presque exacte ^(note 2). Pour cette valeur de c , on peut notamment légèrement majorer $c/2$ par 1, et on obtient le résultat suivant :

Propriété 6.1.3. *La quantité*

$$\mathbb{P} \left(|p - \hat{p}_n| \leq \frac{1}{\sqrt{n}} \right)$$

converge quand n tend vers ∞ , et sa limite est supérieure à 0.95.

Le paragraphe suivant est extrait du programme de classe de seconde (ce que nous notons \hat{p}_n y est noté f) :

Le professeur peut indiquer aux élèves le résultat suivant, utilisable dans la pratique pour des échantillons de taille $n \geq 25$ et des proportions p du caractère comprises entre 0.2 et 0.8 : si f désigne la fréquence du caractère dans l'échantillon, f appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$ avec une probabilité d'au moins 0.95. Le professeur peut faire percevoir expérimentalement la validité de cette propriété mais elle n'est pas exigible.

À proprement parler, le résultat donné est *faux*. Par exemple, pour $n = 27$ et $p = 0.5$, on a l'équivalence (en posant $Y_n = \sum_{k=1}^n X_k$)

$$\left| \frac{Y_n}{n} - p \right| \leq \frac{1}{\sqrt{n}} \Leftrightarrow Y_n \in \{9, 10, \dots, 28\}.$$

(note 2). Avec plus de décimales, on a $c = 1.95996398\dots$

Or, le calcul donne

$$\mathbb{P}(Y_n \in \{9, \dots, 28\}) = 0.9478\dots$$

De même, pour $n = 55$ et $p = 0.5$, on a l'équivalence

$$\left| \frac{Y_n}{n} - p \right| \leq \frac{1}{\sqrt{n}} \Leftrightarrow Y_n \in \{21, 22, \dots, 34\},$$

mais on peut montrer

$$\mathbb{P}(Y_n \in \{21, \dots, 34\}) = 0.9419\dots$$

Il faut toutefois bien voir que le résultat cité par le programme est *presque vrai*, la version rigoureuse en étant la propriété 6.1.3.

6.2 Intervalles de confiance

Dans cette partie, nous allons généraliser la situation de la partie précédente à des variables aléatoires qui ne suivent pas forcément des lois de Bernoulli. On considère une suite de variables aléatoires de carré intégrable $(X_n)_{n \in \mathbb{N}}$ de même loi, dont la moyenne $m = \mathbb{E}X_1$ est inconnue, mais la variance σ^2 est connue.

On appellera *intervalle de confiance* de niveau α sur m un intervalle *aléatoire* \hat{I}_n tel que

$$\mathbb{P}(m \in \hat{I}_n) \geq \alpha.$$

Par l'expression "intervalle aléatoire", on désigne simplement un intervalle dont les bornes sont des variables aléatoires. Dans l'expression $\mathbb{P}(m \in \hat{I}_n)$, on notera bien que la valeur m est parfaitement déterministe, c'est l'intervalle \hat{I}_n qui est aléatoire.

Par exemple, la propriété 6.1.3 énonce en fait que l'intervalle $\hat{I}_n = \left[\hat{p}_n - \frac{1}{\sqrt{n}}, \hat{p}_n + \frac{1}{\sqrt{n}} \right]$ est asymptotiquement (note 3) un intervalle de confiance pour p de niveau 0.95.

Si on revient à l'exemple de l'élection, les échantillons habituellement utilisés lors des sondages en France sont de l'ordre de 1000 personnes. La quantité $\frac{1}{\sqrt{n}}$ correspond alors environ à 3 points de pourcentage : un sondage sur 1000 personnes donnera un score en pourcentage correct à $\pm 3\%$.

6.2.1 Par l'inégalité de Bienaymé-Tchebychev

En appliquant l'inégalité de Bienaymé-Tchebychev à la variable aléatoire $\hat{m}_n = \frac{1}{n} \sum_{k=1}^n X_k$, dont la moyenne est m et la variance est $\frac{\sigma^2}{n}$, on obtient :

$$\mathbb{P}(|\hat{m}_n - m| \leq a) \geq 1 - \frac{\sigma^2}{a^2 n}.$$

En posant $a = c \frac{\sigma}{\sqrt{n}}$, on a

$$\mathbb{P}\left(|\hat{m}_n - m| \leq \frac{c\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{c^2}.$$

Cette équation peut se récrire comme le fait que l'intervalle

$$\hat{I}_n = \left[\hat{m}_n - \frac{c\sigma}{\sqrt{n}}, \hat{m}_n + \frac{c\sigma}{\sqrt{n}} \right]$$

(note 3). Car la propriété 6.1.3 énonce une *convergence* quand n tend vers ∞ et non pas une *égalité*.

est un intervalle de confiance de niveau $\alpha = 1 - \frac{1}{c^2}$ pour m . Par exemple, pour $\alpha = 0.95$, on trouve $c \simeq 4.47$.

6.2.2 Dans le cas Gaussien

Dans le cas où les X_i sont des variables aléatoires indépendantes de loi Gaussienne, alors \hat{m}_n suit la loi Gaussienne de moyenne m et de variance $\frac{\sigma^2}{n}$. Par conséquent, on a

$$\mathbb{P}\left(|\hat{m}_n - m| \leq \frac{c\sigma}{\sqrt{n}}\right) = \mathbb{P}(|G| \leq c),$$

où G suit la loi normale centrée réduite. Autrement dit, l'intervalle

$$\hat{I}_n = \left[\hat{m}_n - \frac{c\sigma}{\sqrt{n}}, \hat{m}_n + \frac{c\sigma}{\sqrt{n}} \right]$$

est un intervalle de confiance de niveau $\mathbb{P}(|G| \leq c)$. Par exemple, pour un intervalle de niveau $\alpha = 0.95$, on trouve $c \simeq 1.96$.

On remarque que pour un même niveau, l'intervalle obtenu ici est beaucoup plus étroit que celui obtenu par l'inégalité de Bienaymé-Tchebychev (pour $\alpha = 0.95$, on a $c = 1.96$ au lieu de $c = 4.47$, soit un intervalle plus que deux fois plus fin).

6.2.3 Par le théorème limite central

Dans le cas général où les X_i ne sont pas Gaussiens, on a quand même, par le théorème limite central :

$$\mathbb{P}\left(|\hat{m}_n - m| < \frac{c\sigma}{\sqrt{n}}\right) \xrightarrow{n \rightarrow \infty} \mathbb{P}(|G| < c).$$

Autrement dit, l'intervalle

$$\hat{I}_n = \left[\hat{m}_n - \frac{c\sigma}{\sqrt{n}}, \hat{m}_n + \frac{c\sigma}{\sqrt{n}} \right]$$

est un intervalle de confiance asymptotique de niveau $\mathbb{P}(|G| \leq c)$. Noter que l'intervalle de confiance est le même que dans la partie 6.2.2, avec le même niveau. La différence avec le cas précédent est qu'ici l'égalité n'est vraie que pour $n \rightarrow \infty$, d'où le terme "intervalle de confiance *asymptotique*".