Probabilités, Statistiques et Combinatoire

Jean-François Marckert

LaBRI – 2017

1 —— Introduction ——

Qu'est-ce que le hasard ? Comment évaluer les chances de tel ou tel événement ? Quelle information peut-on vraiment tirer d'un sondage ?

Ces sujets sont sources de controverses interminables, et nombre de gens ne croient ni à l'existence du hasard, ni à la possibilité de prédire quoi que ce soit, et encore moins à la pertinence des sondages (surtout après les succès des sondeurs lors des élections américaines, ou lors des primaires)... Bien sûr, la théorie des probabilités, et les statistiques, sont des sciences qui ont pour but d'avancer dans la compréhension de ces questions, de permettre des évaluations précises... Leurs succès, leur importance dans toutes les sciences et même dans tous les secteurs d'activité, laissent peu de doute quant à leur capacité à effectivement apporter le bon point de vue sur ces questions... Un des buts de ce cours est d'expliquer un peu ce qu'elles sont, et d'en appréhender aussi leurs limites.

Avant de commencer à définir des objets, démontrer des théorèmes, il est probablement utile de disserter un peu sur la nature des probabilités et statistique et discuter au passage le concept de modélisation.

La théorie des probabilités – comme d'autres sciences – est une science qui a pour but de comprendre les faits de la vie courante, et/ou d'établir certaines prédictions quantitatives quand à la réalisation d'événements à venir. Par ailleurs, les probabilités et les statistiques sont des branches des mathématiques, et en tant que telles, doivent respecter un ensemble de contraintes quant à leur rigueur... Une des contraintes, bien sûr, consiste à démontrer ce qu'on avance. Il n'y a pas d'axiomes additionnels dans la théorie des probabilités. Ce sont les mêmes axiomes que pour les mathématiques usuelles. Ainsi, les probabilités et les statistiques sont des sciences mathématiques, "et rien de plus".

Venons en à cette histoire de modélisation. Lorsqu'on utilise les probabilités (ou les stats) sur un problème réel, une phase de modélisation est nécessaire: on va supposer que la pièce que l'on lance est "juste", quelle a autant de chances de tomber de chaque côté, et que si on la lance 10 fois, les résultats obtenus sont indépendants. Ou alors, on va supposer que les 1000 personnes sondées ont bien été choisies uniformément et indépendamment dans la population, etc... Cette étape, qui consiste finalement à supposer quelque chose sur l'objet d'étude est une étape nécessaire pour passer au traitement mathématique; mais la modélisation possède une part d'arbitraire et "de croyance", et donc, cette étape n'offre pas toutes les garanties propres au traitement mathématique : Si un de vos interlocuteurs nie l'indépendance ou l'uniformité des résultats hebdomadaires du loto, ou déclare qu'il est impossible de choisir un échantillon uniforme de 1000 personnes en France pour faire un sondage, ou alors, pense que le hasard n'existe pas, alors, effectivement, il y a lieu à discussion, mais les mathématiques ne peuvent pas vraiment trancher ces controverses, même si elles peuvent aider à faire un

peu la part des choses pour chacune de ces questions.

Maintenant, discutons un peu la différence entre les deux sciences – souvent confondues – appelées respectivement "probabilités" et "statistiques". Elles sont différentes par nature, même si elles utilisent des outils mathématiques similaires.

- Dans la théorie des probabilités, le point de départ est un modèle idéal: on considère une certaine structure, munie d'une certaine loi de probabilité, et la question porte alors sur l'évaluation sous ce modèle de diverses quantités, comme la probabilité d'un événement, la moyenne, la variance, de telle variable aléatoire, ou l'étude de telle ou telle quantité limite. Par exemple, on va s'interroger sur la probabilité de gagner à la roulette au casino, si on utilise une certaine stratégie, en supposant que les résultats successifs sont indépendants, et que la boule choisit uniformément la case où elle s'arrête.
- Pour les statistiques, l'idée est plutôt d'évaluer les paramètres d'un modèle aléatoire. Cette fois, nous sont donnés les résultats d'une expérience, d'un sondage, et la question porte typiquement sur "la loi la plus vraisemblable de ces données", ou sur la vraisemblance que les données suivent une certaine loi. Par exemple, on va aller au casino, observer tous les résultats de la roulette pendant 15 jours, et essayer de voir à la vue des données recueillies, si la fameuse roulette ne possède pas un biais. Il faut qu'on puisse vraiment pouvoir quantifier ce qu'est un biais normal ou pas, si on veut bâtir une stratégie pour miser enfin (et gagner).

Pour les statistiques également, une part d'arbitraire et de modélisation jouent un rôle. Par exemple, dans l'exemple du casino, on doit supposer que les résultats successifs ont même loi, sont indépendants (ou au moins faire une hypothèse sur la manière dont ils dépendent les uns des autres), pour commencer à évaluer les paramètres du modèle.

Maintenant, en quoi l'informatique est-elle concernée par les probabilités et statistiques ?

Eh bien, les probabilités apparaissent partout en informatique ! Voici quelques domaines de l'informatique qui utilisent directement les probabilités et statistiques:

- En algorithmique en général: on utilise l'aléa pour casser des symétries dans des systèmes; par exemple, en algorithmique distribuée où on cherche à faire collaborer des ordinateurs ("des processus"). L'algorithmique est très souvent aléatoire, car les symétries éventuelles du système font que dans de nombreux cas, il n'existe tout simplement pas d'algorithme déterministe[1] pour effectuer certaines tâches. Il se passe la même chose dans la théorie des jeux, dans laquelle l'un des premiers résultats est que dans nombre de situations, les stratégies optimales sont aléatoires et non pas déterministes.
- L'aléa est utilisé également, lorsque l'on souhaite étudier la complexité de certains algorithmes. Lorsqu'on s'intéresse à la question "$P = NP$?", on s'intéresse à la complexité

[1] Déterministe= non aléatoire

des algorithmes dans le pire des cas. Mais, la question de la complexité typique des algorithmes elle, nécessite de modéliser les "données typiques" sur lesquelles travaille notre algorithme. Par exemple, lorsqu'on utilise un algorithme de tri, on va chercher à étudier le coût moyen, ou en distribution, de l'algorithme, si l'ordre des données est uniforme parmi tous les ordres possibles. Les algorithmes simples, par exemple, ceux qui travaillent sur les bases de données, construisent et parcourent des structures combinatoires. Et c'est donc en étudiant des structures combinatoires aléatoires que l'on parvient parfois, à en comprendre la complexité.

- L'apprentissage statistique est au coeur de tout un tas de recherches et d'applications ces dernières années. Il s'agit de trouver des stratégies pour construire des systèmes intelligents, capable de reconnaître un objet/visage dans une image, de trouver des stratégies pour faire les bonnes suggestions d'achat au client, de mieux jouer à tel ou tel jeu. Puisqu'on a une connaissance uniquement partielle de l'environnement, il s'agit de trouver le bon cadre statistique pour travailler.

2 ——— Probabilités discrètes ———

2.1 Introduction

Que vaut la probabilité d'un événement ? Qu'est-ce que le hasard ? On trouve dans les dictionnaires plusieurs définitions, tournant soit autour du concept "d'événement incertain, imprévisible", ou autour du concept de "cause imprévisible, sort, destin"... La théorie des probabilités ne définit pas le hasard, elle n'en a pas besoin, même si bien sûr, c'est bien le fait qu'une certaine incertitude joue un rôle qui est à l'origine de la modélisation d'un problème utilisant les outils probabilistes.

On pense par ailleurs généralement que la probabilité d'un événement E est donnée par la proportion asymptotique de succès si on recommence indéfiniment l'expérience dont E est une issue possible. Mais en fait, si cela correspond bien à la notion intuitive, ce point de vue ne permet pas de vraiment définir mathématiquement la probabilité... car comment savoir si cette proportion asymptotique existe ? Il faudrait commencer par le démontrer... avant même de pouvoir définir la probabilité d'un événement unique. Cela mène à des complications sans nom. Par ailleurs, il y a plein d'expériences aléatoires dans la vraie vie que l'on ne peut pas répéter indéfiniment... (probabilité de se noyer en traversant la Manche à la nage ?).

La théorie des probabilités utilise l'astuce suivante: on va considérer un ensemble des possibles Ω, appelé $univers\ de\ probabilité$, et on va associer à chaque partie E de Ω, un nombre, la probabilité de E. Bien sûr, la théorie construite colle avec ce que l'on souhaite faire; par exemple, on démontrera dans notre cadre formel que la probabilité d'un événement que l'on a défini correspond effectivement à la proportion asymptotique d'apparition de E, si on répète l'expérience dont E est une issue possible.

L'idée, qui consiste à faire reposer les fondements mathématiques de la théorie des probabilités sur la théorie de la mesure est due au grand mathématicien russe Kolmogorov, et date du début du 20ème siècle.

2.1.1 Univers - Évènements. Notion de probabilité

On ne s'intéresse pour l'instant qu'aux univers discrets, finis ou dénombrables.

Définition 2.1. On appelle univers de probabilité discret Ω tout ensemble fini ou dénombrable.
Une probabilité $\mathbb{P}$ sur Ω est une application

$$\begin{array}{rccc} \mathbb{P}: & \text{Parties}(\Omega) & \longrightarrow & [0,1] \\ & A & \longmapsto & \mathbb{P}(A) \end{array},$$

qui possède les propriétés suivantes:

- $\mathbb{P}(\Omega) = 1$,
- $\mathbb{P}$ est σ-additive: c'est-à-dire, si on prend des parties disjointes $A_1, A_2, \cdots$ de Ω, alors

$$\mathbb{P}\Big(\bigcup_i A_i\Big) = \sum_i \mathbb{P}(A_i), \tag{2.1}$$

formule valable pour un nombre fini (additivité) ou infini (σ-additivité) de A_i.
On appelle la paire $(\Omega, \mathbb{P})$ espace probabilisé.

La notation Parties(Ω) est utilisée pour désigner l'ensemble de tous les sous-ensembles de Ω, l'ensemble vide $\varnothing$ et l'ensemble complet Ω y compris.
Au lieu de *probabilité*, on dit aussi *loi de probabilité* ou *mesure de probabilité*.

On utilise bien sûr l'espace Ω pour modéliser le résultat d'une expérience aléatoire, et $\mathbb{P}$ pour décrire le poids relatif, c'est-à-dire la probabilité, de chaque résultat:

Exemple 2.1. (i) Lancer d'une pièce biaisée : $\Omega = \{pile, face\}$, $\mathbb{P}(\{pile\}) = 0.4$, $\mathbb{P}(\{face\}) = 0.6$

(ii) Lancer d'un dé juste : $\Omega = \{1,2,3,4,5,6\}$ avec $P(\{i\}) = 1/6$ pour tout $i \in \Omega$,

(iii) Loto: $\Omega = \big\{\{a_1, a_2, \ldots, a_6\}$ sous ensemble à 6 éléments de $\{1, \cdots, 49\}\big\}$, avec $\mathbb{P}(\{e\}) = 1/\#\Omega$, pour tout e dans Ω.

(iv) Nombre de lancers de pièces avant d'obtenir *face*: $\Omega = \{1,2,3,\cdots\}$, avec probabilité $P(\{i\}) = 1/2^i$ pour tout $i \in \Omega$.

Définition 2.2. Les éléments de Ω sont appelés événements élémentaires ou éventualités, et ceux de Parties(Ω), événements.

La définition d'une probabilité est un peu minimale... Voyons quelques conséquences immédiates:

Proposition 2.3. Soit $\mathbb{P}$ une probabilité sur un univers discret Ω. On a

(a) $\mathbb{P}(\varnothing) = 0$,

(b) Si A, B sont des parties de Ω et si $A \subset B$, alors $\mathbb{P}(A) \leq \mathbb{P}(B)$,

(c) Soit $A \in \text{Parties}(\Omega)$ et A^c son complémentaire dans Ω (c'est-à-dire, $A^c = \Omega \setminus A$): on a
$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

(d) Pour tout $A, B \in \text{Parties}(A)$, on a
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Intuitivement, ces propriétés sont celles que l'on attend d'une probabilité (ou encore d'une fonction qui associe un "poids" à chaque événement).

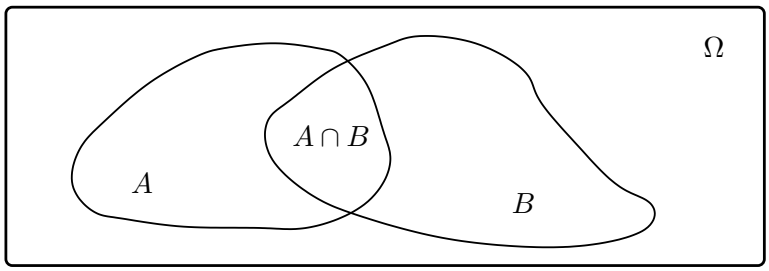


Figure 1: Décomposition de $A \cup B$.

Preuve. (a) On sait que $\varnothing$ et Ω sont disjoints et en plus que $\Omega \cup \varnothing = \Omega$. On a donc $\mathbb{P}(\Omega \cup \varnothing) = \mathbb{P}(\Omega) + \mathbb{P}(\varnothing) = \mathbb{P}(\Omega)$, d'où on déduit $\mathbb{P}(\varnothing) = 0$.

(b) Notons $C = B \setminus A$. On a alors $B = A \cup C$, avec A et C disjoints. Ainsi $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(C)$ et comme $\mathbb{P}$ prend ses valeurs dans $[0, 1]$, $\mathbb{P}(B) \geq \mathbb{P}(A)$.

(c) Il suffit de remarquer que $\Omega = A \cup A^c$ et que ces 2 ensembles sont disjoints.

(d) Il suffit de regarder la figure 1 ou de prendre l'argument suivant. Réécrivons l'identité $\mathbb{P}(A \cap B) + \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$, et remarquons que les ensembles $A \setminus B$, $B \setminus A$ et $A \cap B$ sont disjoints. Ainsi, on voit facilement que les deux membres de l'égalité valent $\mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + 2\mathbb{P}(A \cap B)$. □

Proposition 2.4. Soit $(\Omega, \mathbb{P})$ un espace de probabilité discret. La famille $(p_i, i \in \Omega)$ détermine entièrement la probabilité $\mathbb{P}$.
Réciproquement, à toute famille finie ou dénombrable $(p_i, i \in I)$ telle que

- $\forall i, \quad p_i \geq 0$,
- $\sum_{i \in I} p_i = 1$

on peut associer un univers $\Omega = \{\omega_i, i \in I\}$ et une probabilité $\mathbb{P}$ sur Ω telle que:

$$\mathbb{P}(\{\omega_i\}) = p_i.$$

Preuve. Cette propriété résulte du fait suivant: soit A appartenant à $\mathcal{P}(\Omega)$. Écrivons A comme union des événements élémentaires qui le compose (cette écriture est unique):

$$A = \bigcup_{\omega \in A} \{\omega\}.$$

D'après la définition d'une probabilité, puisque les $\{\omega\}$ sont disjoints,

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

La probabilité de A est donc bien définie.
La réciproque signifie que tout ensemble de nombres positifs, de somme 1, peut être considéré comme une probabilité sur un ensemble de cardinal égal. Ceci est clair puisqu'il suffit d'ailleurs de prendre $\Omega = I$ (l'ensemble qui indexe les p_i et poser $\mathbb{P}(\{i\}) = p_i$ pour $i \in I$). □

En fait, dans les exemples page 8, on a déjà utilisé sans le dire cette proposition: la probabilité de tout élément de Parties(Ω), est entièrement déterminée par la probabilité des événements élémentaires, "la probabilités des singletons". Cela ne sera plus valable lorsqu'on traitera des espaces de probabilité non discrets.

Exemple 2.2. Lorsqu'on lance un dé (équilibré), on utilise la modélisation suivante: $\Omega = \{1, 2, 3, 4, 5, 6\}$, Parties$(\Omega) = \big\{\{1\}, \cdots, \{6\}, \{1, 2\}, \{1, 3\}, \cdots, \{5, 6\}, \{1, 2, 3\}, \{1, 2, 4\} \cdots \{1, 2, 3, 4\}, \cdots \{3, 4, 5, 6\}, \{1, 2, 3, 4, 5\}, \cdots, \{2, 3, 4, 5, 6\}, \{1, 2, 3, 4, 5, 6\}, \varnothing\big\}$ avec $\mathbb{P}(\{i\}) = 1/6$ pour tout $i \in \Omega$. $A = \{1, 3, 5\}$ est un événement, correspondant au fait que le résultat du lancer est impair et $\mathbb{P}(A) = 3/6$.

Remarque 2.5. *Dans les cours plus abstraits de théorie des probabilités, on ne définit pas une probabilité sur l'ensemble des parties, mais plutôt sur ce qu'on appelle "une tribu", qui est un sous-ensemble de* Parties(Ω) *(qui contient Ω, qui est stable par union et intersection dénombrable). Lorsqu'on travaille sur des espaces finis ou dénombrables, fondamentalement, ces complications ne sont pas nécessaires car elles reviennent à supposer que certains éléments n'ont pas de probabilité propre.*

2.1.2 Équiprobabilité... et premières formules de combinatoire

Définition 2.6. On appelle équiprobabilité ou mesure uniforme sur un univers fini $\Omega = \{\omega_1, \cdots, \omega_N\}$, la mesure de probabilité accordant la même probabilité à chaque événement élémentaire:

$$\mathbb{P}(\omega_1) = \cdots = \mathbb{P}(\omega_N) = 1/N.$$

Sous cette équiprobabilité, pour tout événement $A \in$ Parties(Ω)

$$\mathbb{P}(A) = \#A/\#\Omega.$$

Cette formule n'est valable bien sûr que dans le cas de l'équiprobabilité. On dit parfois que les éléments de A sont les cas favorables et ceux de Ω les cas possibles.
Exemple: On tire une carte parmi 32 de façon équiprobable. Probabilité de tirer un coeur=8/32.

Combinaisons et Arrangements

On appelle k-uplet une suite finie $(x_1, \cdots, x_k)$ de longueur k; ainsi les 2-uplets correspondent aux couples. Deux k-uplets $(x_1, \cdots, x_k)$ et $(y_1, \cdots, y_k)$ sont différents s'il existe $i \in \{1, \cdots, k\}$ tel que $x_i \neq y_i$.

Définition 2.7. Soit A un ensemble à n éléments avec $n \geq 1$, et soit k tel que $0 \leq k \leq n$. On dit qu'un k-uplet $(x_1, \cdots, x_k)$ est un arrangement de k éléments de A, si les x_i sont dans A et différents 2 à 2.

Exemple 2.3. Les arrangements de 2 éléments de $\{1, 2, 3, 4\}$ sont $(1, 2)$, $(1, 3)$, $(1, 4)$, $(2, 1)$, $(2, 3)$, $(2, 4)$, $(3, 1)$, $(3, 2)$, $(3, 4)$, $(4, 1)$, $(4, 2)$, $(4, 3)$.

Proposition 2.8. Soit A un ensemble à n éléments avec $n \geq 1$, et soit k tel que $0 \leq k \leq n$. Le nombre A_n^k d'arrangements de k éléments de A est

$$A_n^k = \frac{n!}{(n-k)!} = n \times (n-1) \times \cdots \times (n-k+1). \tag{2.2}$$

Preuve. Si on cherche à compter le nombre de k-uplets différents $(x_1, \cdots, x_k)$, on voit que pour x_1 fixé, $(x_2, \cdots, x_k)$ parcourt l'ensemble des $k-1$-uplets de l'ensemble $A \setminus \{x_1\}$. On voit alors que puisqu'on a n possibilités pour x_1, qu'on a

$$A_n^k = nA_{n-1}^{k-1} = n(n-1)A_{n-2}^{k-2} = \cdots = n(n-1)\cdots(n-k)A_{n-(k-1)}^1;$$

en utilisant la formule évidente $A_m^1 = m$ pour tout m, (2.2) suit immédiatement. □

Exemple 2.4. "Toucher le tiercé dans l'ordre", signifie qu'on a deviné (et misé) sur l'ordre exact des 3 premiers chevaux dans une course. Il s'agit donc d'un arrangement. S'il y a 20 chevaux en tout, le nombre de tiercés possibles, en tout est $A_{20}^3 = 20!/17! = 20 \times 19 \times 18 = 6840$. Bien sûr, si on choisit uniformément au hasard ce que l'on joue, alors la proba de gain est $1/6840$.

Proposition 2.9. Soit A un ensemble à n éléments, $n \geq 0$. Pour $0 \leq k \leq n$, il y a

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2.3}$$

sous-ensembles de A à k éléments.

La notation $\binom{n}{k}$ est la notation utilisée, hors France. On parle ici de "combinaisons" de k éléments parmi n dans les cours élémentaires de probabilités... mais il serait probablement malin d'abandonner cette terminologie au profit de "sous-ensembles", notion plus claire.

Preuve. Prenons $k \geq 1$. Il suffit de voir que si on se donne un sous ensemble $\{x_1, \cdots, x_k\}$ (avec éléments distincts, donc) il y a $k!$-uplets $(y_1, \cdots, y_k)$ tels que $\{y_1, \cdots, y_k\} = \{x_1, \cdots, x_k\}$. Il y a donc $k!$ fois moins de sous-ensembles à k éléments que d'arrangements à k éléments. On peut voir que pour $k = 0$, la formule est valable également car, il y a un seul sous-ensemble de A à 0 éléments: $\varnothing$. □

Exemple 2.5. Nombre de grilles différentes au loto: on choisit 6 numéros parmi 49. Le choix est un sous-ensemble de 6 éléments parmi $\{1, \cdots, 49\}$. C'est-à-dire choisir $\{1, 5, 10, 15, 16, 42\}$ ou $\{1, 5, 10, 15, 42, 16\}$ c'est cocher les mêmes cases et c'est donc la même chose. Ainsi le nombre de choix est $C_{49}^6 = 49!/(43!6!) = 13983816$. Bien sûr, si on suppose que le résultat du loto est uniforme parmi les combinaisons possibles, la proba de gagner en jouant une seule grille est $1/13983816$.

Proposition 2.10. Les identités suivantes sont vraies:

(*i*) Pour tout $0 \leq p \leq n$, $C_n^p = C_n^{n-p}$.

(*ii*) Pour tout $1 \leq p \leq n-1$, $C_n^p = C_{n-1}^{p-1} + C_{n-1}^p$.

(*iii*) Pour tout $n \geq 0$

$$2^n = \sum_{k=0}^{n} C_n^k, \qquad \sum_{k=0}^{n} C_n^k (-1)^k = 0$$

Preuve. (*i*) Évident.

(*ii*) Se vérifie en factorisant le membre de droite

(*iii*) C'est une conséquence de la formule du binôme de Newton:

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}. \tag{2.4}$$

Prendre $a = b = 1$ pour la première, et $a = -1$, $b = 1$ pour la seconde. □

2.1.3 Probabilité conditionnelle - Indépendance

Les probabilités conditionnelles ont pour but d'évaluer "le changement de probabilité" dû à l'acquisition d'informations. Par exemple, si l'on dispose d'un dé juste, la probabilité d'obtenir un 1 est 1/6. Si quelqu'un lance le dé pour nous et nous donne l'information suivante: "le résultat est impair". On peut écarter les événements $\{2\}, \{4\}, \{6\}$ et en déduire que maintenant, le résultat est 1 avec probabilité 1/3. Formalisons tout cela...

Définition 2.11. Soit $(\Omega, \mathbb{P})$ un espace probabilisé et $B \in \text{Parties}(\Omega)$ tel que $\mathbb{P}(B) > 0$; soit A un élément de Parties(Ω). La probabilité conditionnelle de A sachant B est définie par:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{2.5}$$

Remarque 2.12. *Les nombres $\mathbb{P}(B \mid A)$ et $\mathbb{P}(A \mid B)$ sont (en général) différents. Si $\mathbb{P}(A) > 0$ et $\mathbb{P}(B > 0)$ la formule (2.5) indique que*

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

Ce simple jeu d'écriture est souvent utilisé, pour "retourner des probabilités conditionnelles: on a ainsi $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)\mathbb{P}(A)/\mathbb{P}(B)$, formule qui découle directement de (2.5).

Le théorème suivant est important: il explique comment et pourquoi la "probabilité conditionnelle à l'événement B" est bien une probabilité qui vérifiera donc toutes les propriétés propres aux probabilités établies plus haut.

Théorème 2.13. Soit $(\Omega, \mathbb{P})$ un espace probabilisé et $B \in \text{Parties}(\Omega)$ tel que $\mathbb{P}(B) > 0$. L'application

$$\begin{aligned} \mathbb{P}(. \mid B) : \text{Parties}(\Omega) &\longrightarrow [0,1] \\ A &\longrightarrow \mathbb{P}(A \mid B) \end{aligned}$$

est une probabilité sur Ω (pour laquelle $\mathbb{P}(B \mid B) = 1$).

Preuve. On a immédiatement $\mathbb{P}(\varnothing \mid B) = 0, \mathbb{P}(\Omega \mid B) = 1$, mais aussi $\mathbb{P}(B \mid B) = \mathbb{P}(B \cap B)/\mathbb{P}(B) = 1$. Prenons maintenant une famille $(A_i, i \in I)$ de parties de Ω disjointes, et vérifions que $\mathbb{P}(\ \cdot\ |B)$ est bien σ-additive. On a

$$\mathbb{P}\left(\bigcup_{i\in I} A_i \,\middle|\, B\right) = \frac{\mathbb{P}((\bigcup_{i\in I} A_i) \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}\big(\bigcup_{i\in I}(A_i \cap B)\big)}{\mathbb{P}(B)}$$

comme les ensembles $A_i \cap B$ sont disjoints (c'est-à-dire $(A_i \cap B) \cap (A_j \cap B) = \varnothing$ si $i \neq j$), donc

$$\mathbb{P}\left(\bigcup_{i\in I} A_i \,\middle|\, B\right) = \sum_{i\in I} \frac{\mathbb{P}\big(A_i \cap B\big)}{\mathbb{P}(B)} = \sum_{i\in I} \mathbb{P}(A_i \mid B),$$

ce qui montre la σ-additivité. □ □

La formule des probabilités conditionnelles correspond tout à fait au changement de probabilité intuitif. Le référent n'est plus Ω mais B. Ainsi (2.5) traduit le fait que les cas possibles sont dans B; la probabilité de B sachant B vaut donc 1.

Exemple 2.6. Supposons que $\Omega = \{1,2,3,4,5,6\}$ muni de la probabilité suivante:

$$\mathbb{P}(\{i\}) = i/21. \tag{2.6}$$

Autrement dit, la probabilité de tomber sur i est proportionnelle à i. Supposons $B = \{1,2,3,4\}$. On voit que $\mathbb{P}(B) = 10/21$. Intéressons nous aux singletons: $\{i\}$ est dans B sssi $1 \leq i \leq 4$. On voit alors que

$$\mathbb{P}(\{i\}|B) = \frac{\mathbb{P}(\{i\} \cap B)}{\mathbb{P}(B)} = \begin{cases} \frac{i/21}{10/21} = i/10 & \text{si } i \in \{1,2,3,4\} \\ 0 & \text{si } i \in \{5,6\} \end{cases}.$$

- Puisque $\mathbb{P}(. \mid B)$ est une probabilité, pour tout $A \in \text{Parties}(\Omega)$, $\mathbb{P}(A \mid B) = \sum_{j\in A} \mathbb{P}(\{j\} \mid B)$.

- La probabilité $\mathbb{P}(. \mid B)$ ne charge que B, et on a $\mathbb{P}(B \mid B) = \mathbb{P}(B \cap B)/\mathbb{P}(B) = 1$.
- On remarque aussi que la probabilité $\mathbb{P}(. \mid B)$ est proportionnelle pour les événements inclus dans B à leur valeur initiale sous la probabilité $\mathbb{P}$ (dans notre exemple, $\mathbb{P}(\{i\} \mid B)$ est proportionnel à i si i est dans B. C'est un point qui doit être intuitif... Recevoir l'information que l'on est dans B, ne doit pas changer les rapports de probabilité entre les événements inclus dans B; ceux totalement à l'extérieur de B voient leur proba tomber à 0....

On va maintenant énoncer la fameuse formule de Bayes. Avant cela, rappelons la notion de partition.

Définition 2.14. Soit A un ensemble quelconque. On appelle partition de A toute famille d'ensembles $(A_i, i \in I)$ (où I est un ensemble d'indices) telle que les A_i sont disjoints, d'union A; en d'autres termes :

$$\begin{cases} i) & \bigcup_{i \in I} A_i = A, \\ ii) & A_i \cap A_j = \varnothing, \forall i, j \in I, \quad \text{si } i \neq j \end{cases} \tag{2.7}$$

La formule des probabilités totales est une formule évidente, que l'on utilise très souvent en probabilité, parfois même sans s'en rendre compte...

Proposition 2.15. [Formule des probabilités totales] Soit $(\Omega, \mathbb{P})$ un espace probabilisé et $(A_i, i \in I)$ une partition de Ω. On a, pour tout $B \in$ Parties(Ω)

$$\mathbb{P}(B) = \mathbb{P}(B \cap \Omega) = \mathbb{P}\left(B \cap (\bigcup_{i \in I} A_i)\right) = \sum_{i \in I} \mathbb{P}\left(B \cap A_i\right).$$

Preuve. Toutes les égalités sont évidentes. La 3ème se comprend aisément sur la figure 2, et si on veut une preuve plus classique, on utilise le fait que comme les ensembles A_i sont disjoints,

$$B \cap \left(\bigcup_{i \in I} A_i\right) = \bigcup_{i \in I} (A_i \cap B),$$

avec les ensembles $A_i \cap B$ disjoints. □

Exemple 2.7. Munissons un ensemble d'individus de l'équiprobabilité. Notons R_i le sous-ensemble des individus qui habitent dans la région i pour i allant de 1 à 12 (chaque individu habite dans une seule région); notons V le sous-ensemble de ceux qui sont végétariens. On a

$$\mathbb{P}(V) = \sum_{i=1}^{12} \mathbb{P}(V \cap R_i),$$

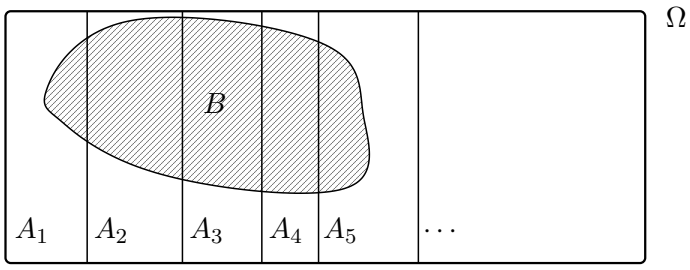


Figure 2: Intersection de B avec une partition...

c'est-à-dire, la somme des proportions des individus qui sont végétariens et dans la région i. Notez bien la présence du signe $\cap$, que l'on ne peut pas remplacer par $\mid$. Si on veut passer aux probas conditionnelles, on peut le faire comme suit, par exemple,

$$\mathbb{P}(V) = \sum_{i=1}^{12} \mathbb{P}(V \cap R_i) = \sum_{i=1}^{12} \mathbb{P}(V \mid R_i)\mathbb{P}(R_i),$$

où on voit qu'il faut pondérer la probabilité conditionnelle $\mathbb{P}(V \cap R_i)$ par la proportion $\mathbb{P}(R_i)$ du nombre d'habitants dans i.

Énonçons maintenant la formule de Bayes, utilisée pour "inverser" des probabilités conditionnelles. Elle est très facile à retrouver !

Proposition 2.16. [Formule de Bayes] Soit $(\Omega, \mathbb{P})$ un espace probabilisé et $(A_i, i \in I)$ une partition de Ω. Si $\mathbb{P}(B) > 0$ et $\mathbb{P}(A_i) > 0$ pour tout i, alors:

$$\mathbb{P}(A_j \mid B) = \frac{\mathbb{P}(A_j \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A_j)\,\mathbb{P}(A_j)}{\sum_{i\in I} \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}. \tag{2.8}$$

Preuve. On sait que $\mathbb{P}(B \cap A_j) = \mathbb{P}(B|A_j)\mathbb{P}(A_j) = \mathbb{P}(A_j \mid B)\mathbb{P}(B)$. On voit qu'on peut donc s'amuser à conditionner par A_j ou par B et que toute cela n'est qu'un jeu d'écriture. Dans la formule de Bayes, on veut conditionner par B et retourner le conditionnement... pour montrer (2.8), on voit que seul le dénominateur reste à comprendre... La formule des probabilités totales dit que $\mathbb{P}(B) = \mathbb{P}(B \cap (\bigcup_{i\in I} A_i)) = \sum_{i\in I} \mathbb{P}(B \cap A_i)$; or ceci est égal à $\sum_{i\in I} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)$. □

Exemple 2.8. Supposons que dans une classe 34% des élèves sont des filles, 66% des garçons, et que 30% des filles fument, 25% des garçons fument. On écrit cela sous la forme $\mathbb{P}(F) = 0.34$, $\mathbb{P}(G) = 0.66$, et $\mathbb{P}(f \mid F) = 0.30$, $P(f \mid G) = 0.25$. La formule de Bayes permet d'"inverser" des probabilités conditionnelles, et donc de calculer la probabilité d'être une fille sachant qu'on fume:

$$\mathbb{P}(F \mid f) = \frac{\mathbb{P}(F \cap f)}{\mathbb{P}(f)} = \frac{\mathbb{P}(f \mid F)\mathbb{P}(F)}{\mathbb{P}(f|F)\mathbb{P}(F) + \mathbb{P}(f|G)\mathbb{P}(G)} = \frac{0.3 \times 0.34}{0.3 \times 0.34 + 0.25 \times 0.66}.$$

On définit maintenant l'une des notions centrales de la théorie des probabilités: la notion d'indépendance.

Définition 2.17. Soit $(\Omega, \mathbb{P})$ un espace probabilisé. On dit que deux événements A et B sont indépendants si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{2.9}$$

Plus généralement, si $(A_i, 1 \leq i \leq K)$ est une famille d'événements, il sont dits indépendants si pour toute partie I de $\{1, \cdots, K\}$,

$$\mathbb{P}\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} \mathbb{P}(A_i). \tag{2.10}$$

Remarque 2.18. *Notez que (2.10) n'est pas équivalent à* $\mathbb{P}\left(\bigcap_{i=1}^K A_i\right) = \prod_{i=1}^K \mathbb{P}(A_i)$*; c'est une condition nettement plus restrictive.*

L'indépendance est une notion primordiale en probabilité comme on va le voir par la suite. Si A et B sont indépendants, par (2.5) on a:

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

(Ceci bien sûr si $\mathbb{P}(B) \neq 0$.) L'interprétation est la suivante: si A et B sont indépendants, savoir que B est réalisé ne modifie en rien la probabilité de réalisation (ou non) de A.

Exemple 2.9. Si on lance un dé juste, et si on note $A = \{1, 3, 5\}$ (le résultat est impair), $B = \{2, 4, 6\}$ (le résultat est pair), $C = \{3, 6\}$ (le résultat est divisible par 3), alors A et C sont indépendants, et aussi B et C sont indépendants car

$$\mathbb{P}(A \cap C) = \mathbb{P}(\{3\}) = \frac{1}{6} = \mathbb{P}(A)\mathbb{P}(C) = \frac{3}{6} \times \frac{2}{6}$$

(même calcul pour B et C) mais A et B ne le sont pas car

$$\mathbb{P}(A \cap B) = \mathbb{P}(\varnothing) = 0 \neq \mathbb{P}(A)\mathbb{P}(B) = \left(\frac{3}{6}\right)^2.$$

Intuitivement : A et B ne sont pas indépendants, car si A est réalisé alors on sait que B ne l'est pas. Donc, on a acquis une information qui change la probabilité de la réalisation de B. Maintenant, A et C sont indépendants: que A soit réalisé ou pas, C a une chance sur 3 de se réaliser... que l'on nous donne ou pas d'information sur la réalisation de C, A a une chance sur 2 de se réaliser.

"Dans la vraie vie", l'indépendance d'événements est souvent une question cruciale; en médecine par exemple, à cause de l'effet placebo et aussi, du fait que les traitements médicamenteux ne marchent souvent pas instantanément, ni à 100%, on se pose naturellement la question de savoir si la guérison est indépendante ou non de la prise d'un médicament. Seules des études statistiques permettent d'avancer sur ces questions.

Plus ou moins naïvement, on peut affirmer qu'une personne superstitieuse se distingue d'une non superstitieuse en ce qu'elle perçoit de la dépendance entre deux événements, là où la seconde n'en voit pas. Ainsi, le non superstitieux ne pense pas qu'il ne risque rien s'il croise un chat noir, mais pense que la probabilité qu'il lui arrive quelque chose de malheureux (disons dans la semaine qui suit), n'est pas modifiée par cette rencontre.

Remarque 2.19. *Si on se donne 3 événements A_1, A_2, A_3, ceux-ci peuvent être indépendants 2 à 2, sans être indépendants. Trouver un tel exemple de 3 événements est un bon exercice...*

2.2 Variables aléatoires

2.2.1 Introduction

Définition 2.20. Soit $(\Omega, \mathbb{P})$ un espace probabilisé. On appelle variable aléatoire sur Ω toute application

$$\begin{aligned} X : \Omega &\longrightarrow \Omega' \\ \omega &\longmapsto X(\omega) \end{aligned}$$

où Ω' est un autre ensemble.

Exemple 2.10. Par exemple, si $\Omega = \{1, 2, 3, 4, 5, 6\}$, l'application X définie par $X(i) = p$ si i est pair et $X(i) = f$ si i est impair, est une variable aléatoire. L'espace Ω' dans ce cas est $\{p, f\}$, ou, plus généralement, tout ensemble qui contient p et f.

Remarque 2.21. *Alors ici il faut quand même noter que la dénomination "variable aléatoire" n'est pas particulièrement bien choisie... En fait, même si la définition des variables aléatoires est un plus compliquée que celle présentée ici lorsque l'on travaille avec des espaces non discrets, on voit bien qu'une variable aléatoire est une fonction en non pas une variable... Elle n'est pas non plus aléatoire, elle ne dépend pas de $\mathbb{P}$. Par la suite, on verra tout de même pourquoi on peut comprendre un peu le choix de cette dénomination, source de nombreuses confusions.*

Avant même d'aller plus loin, regardons de près un exemple:

Exemple 2.11. Prenons $\Omega = \{1, 2, 3, 4, 5, 6\}$ munie de l'équiprobabilité, $1/6$ pour chaque élément. Soit X la variable aléatoire définie par $X(a) = (a-4)(a-1)$. On voit alors que

$$X(1) = 0, X(2) = -2, X(3) = -2, X(4) = 0, X(5) = 4, X(6) = 10.$$

Autrement dit : Alice et Bob jouent au dé, et pour une raison non divulguée ici, il a été décidé que si le dé tombait sur a, alors Alice donnerait $X(a)$ euros à Bob. Maintenant, on voit que "l'aléa initial" qui concernait le résultat du dé est transporté par X sur l'aléa de la dette d'Alice.

L'espace image

$$X(\Omega) = \{X(i), i \in \Omega\} = \{0, -2, 4, 10\} \tag{2.11}$$

est l'ensemble des dettes possibles. Maintenant, il est clair que ces 4 dettes ne sont pas équiprobables, puisque 0 et -2 sont chacun image de 2 éléments de Ω; ainsi, la probabilité que X prennent la valeur 0 est $2/6$, la valeur -2, $2/6$ aussi, et enfin, la probabilité que X prenne la valeur 4 est $1/6$, et même chose pour 10. Bien sûr, si le dé n'avait pas été juste, il aurait fallu modifier le calcul en tenant compte de la probabilité de chaque face du dé.

Bon, voilà, on va maintenant formaliser le cas général, mais toute l'idée est là: une variable aléatoire est définie sur un espace de probabilité; on s'intéresse à l'image de cette variable aléatoire, et surtout à ce qu'on appelle la loi de cette variable: la probabilité sur l'espace image que X prenne telle ou telle valeur. Cette loi est une loi de probabilité, image de la probabilité initiale au sens suivant:

Définition 2.22. Soit $(\Omega, \mathbb{P})$ un espace discret probabilisé, et $X : \Omega \to \Omega'$ une variable aléatoire définie sur cet espace. L'ensemble image

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

est fini ou dénombrable. On appelle loi de X, la mesure de probabilité $\mathbb{P}_X$ définie sur $X(\Omega)$ par pour tout $A \in \text{Parties}(X(\Omega))$,

$$\mathbb{P}_X(A) = \mathbb{P}(\{\omega, X(\omega) \in A\}) = \mathbb{P}\left(X^{-1}(A)\right) = \sum_{\omega \in \Omega \ : X(\omega) \in A} \mathbb{P}(\{\omega\}). \tag{2.12}$$

Remarque 2.23.

- *La notation $\mathbb{P}_X$ est classique; il s'agit juste de donner un nom à cette loi image (on appelle parfois $\mathbb{P}_X$ la probabilité image de $\mathbb{P}$ par X). On peut l'appeler $\mathbb{P}'$ ou $\mathbb{Q}$ si on préfère éviter ce "X" en indice qui ne joue aucun rôle particulier.*
- *Au lieu de noter $\mathbb{P}_X(A)$ il est classique de noter $\mathbb{P}(X \in A)$. De même si $A = \{a\}$, un simple élément, on écrit $\mathbb{P}_X(\{a\}) = \mathbb{P}(X = a)$. On note aussi $\mathbb{P}(X \leq x)$ au lieu*

de $\mathbb{P}(X \in (-\infty, x])$, etc. Du coup, c'est vrai qu'ici, dans les notations, ou oublie un peu que X est une fonction, et on fait comme si X était une variable. Mais bon.

- *Remarque un peu subtile: Puisque X est une fonction de Ω dans Ω', on peut tout à fait voir $\mathbb{P}_X$ comme une loi de probabilité sur Ω' plutôt que sur $X(\Omega)$. Par exemple, dans le jeu d'Alice et Bob, on pourrait dire que X est une fonction à valeurs dans $\mathbb{Z}$, ou dans $\mathbb{R}$. Ça ne change rien au fait que $\mathbb{P}_X$ n'attribue une masse positive qu'aux éléments de $X(\Omega) = \{-2, 0, 4, 10\}$, et que la loi se décrit "de manière économique" sur $X(\Omega)$ plutôt. Mais si on travaille sur Ω' au lieu de $X(\Omega)$ on n'est plus forcément dans un cadre dénombrable et les ensembles Ω' et* Parties(Ω') *peuvent-être titanesques. De nouveau, ça ne pose pas ici de problème particulier. Dans le cas non discret, il faudra faire un peu plus attention.*

2.2.2 Gommage de l'espace de probabilité

Il s'agit ici d'une réflexion un peu abstraite, mais qui permet de comprendre pourquoi on se permet parfois de parler de variables aléatoires, de leur loi, sans définir correctement l'espace initial $(\Omega, \mathbb{P})$.

Au début, nous avons discuté d'une mesure $\mathbb{P}$ définie sur un ensemble Ω. Maintenant, on voit que la loi $\mathbb{P}_X$ de la variable aléatoire X est également une loi de probabilité. On peut se demander quelle est la différence de nature entre une loi de probabilité sur un espace Ω, et la loi $\mathbb{P}_X$ d'une variable X, mettons prenant ses valeurs dans un espace Ω'... En fait, il n'y a pas de différence de nature, ni de degré de généralité. Il y a plusieurs façons de voir cela. La plus simple, mais un peu abstraite, consiste à considérer la variable aléatoire identité, définie par $X(\omega) = \omega$ définie sur Ω et donc à valeurs dans Ω. Si on prend comme loi initiale $\mathbb{P}$, alors on voit que la loi $\mathbb{P}_X$ est égale à $\mathbb{P}$: en effet pour tout A,

$$\mathbb{P}_X(A) = \mathbb{P}(\{\omega : X(\omega) \in A\}) = \mathbb{P}(\{\omega : \omega \in A\}) = \mathbb{P}(A).$$

Ainsi, on peut énoncer le résultat suivant, valable en toute généralité:

Lemme 2.24. Toute mesure de probabilité est la loi d'une variable aléatoire.

Dans nombre de situations on prendra directement une variable aléatoire dont on spécifiera la loi, plutôt que de définir d'abord un espace de probabilité. On se permettra donc de dire directement: "soit X une variable aléatoire uniforme sur $\{1, 2, 3, 4, 5, 6\}$", sans spécifier que X est définie sur un espace $(\Omega, \mathbb{P})$, et surtout, sans spécifier Ω...

2.2.3 Variable aléatoire réelle

Définition 2.25. On appelle variable aléatoire réelle, une variable aléatoire définie sur un espace de probabilité $(\Omega, \mathbb{P})$ et prenant ses valeurs dans $\mathbb{R}$, c'est-à-dire, $X : \Omega \to \mathbb{R}$.

Exemple 2.12. Dans le jeu avec Alice et Bob ci-dessus (Exemple 2.11), X est une variable aléatoire réelle, mais ce n'est pas une variable réelle dans l'Exemple 2.10.

Les variables aléatoires réelles sont intéressantes en cela que dans nombre d'applications, ce sont effectivement des données aléatoires quantitatives qui apparaissent, mais aussi, car avec des nombres, on peut faire des calculs.

Pour une variable aléatoire discrète réelle X, la loi de X est entièrement décrite par l'ensemble des valeurs $\mathbb{P}(X = x)$ pour x décrivant $\mathbb{R}$. Bien sûr, pour au plus un nombre dénombrable de x, $\mathbb{P}(X = x)$ peut être strictement positif. Ces "x" qui ont une probabilité positive sont appelés *atomes* de la loi $\mathbb{P}_X$.

Définition 2.26. La fonction de répartition de X est la fonction définie par:

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0,1] \\ x &\longmapsto F(x) = \mathbb{P}(X \leq x). \end{aligned}$$

On voit clairement que la fonction de répartition est croissante. Ses limites sont 0 en $-\infty$ et 1 en $+\infty$. La fonction de répartition est constante entre les atomes, et saute de $\mathbb{P}(X = x)$ en un atome x.

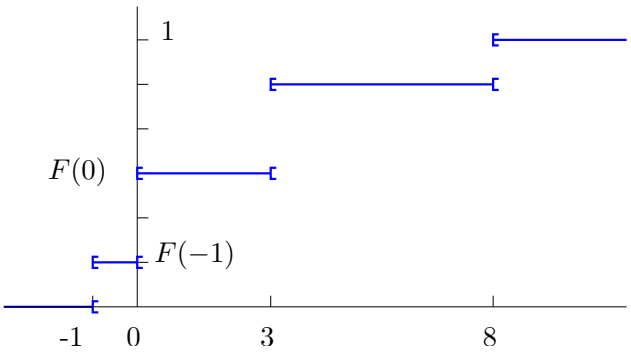


Figure 3: Exemple d'une fonction de répartition

Définition 2.27. Soit X une variable aléatoire réelle dont l'ensemble des atomes est $\{x_i, i \in I\}$ (pour un ensemble d'indice I fini ou dénombrable). L'espérance mathématique de X, également appelée moyenne de X, est le nombre

$$\mathbb{E}(X) = \sum_{i\in I} x_i \,\mathbb{P}(X = x_i) \underset{(a)}{=} \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\{\omega\}),$$

lorsque cette dernière quantité existe.

L'égalité (a) est juste la retranscription du fait que $\mathbb{P}_X$ est la loi image de $\mathbb{P}$ par X.

Exemple 2.13.

- Dans l'histoire d'Alice est Bob, la dette moyenne d'Alice vaut donc

$$\mathbb{E}(X) = 0 \times \left(\frac{2}{6}\right) + (-2) \times \left(\frac{2}{6}\right) + 4 \times \left(\frac{1}{6}\right) + 10 \times \left(\frac{1}{6}\right) = \frac{10}{6}.$$

- Lorsque l'espace est dénombrable, la somme comporte un nombre infini de termes, et l'espérance peut diverger, "ne pas exister": par exemple, si $\mathbb{P}(X = i) = \frac{6}{\pi^2 i^2}$ pour $i \in \{1, 2, \cdots\}$. Alors

$$\mathbb{E}(X) = \sum_{i\geq 1} i\frac{6}{\pi^2 i^2} = \frac{6}{\pi^2}\sum_{i\geq 1}\frac{1}{i}$$

et cette somme est connue pour diverger; dans ce cas on peut s'accorder pour dire que la moyenne vaut $+\infty$.

- Dans la théorie des séries, on apprend que certaines séries ne convergent pas, n'ont pas de valeurs. Par exemple, si la loi de X est donnée par $\mathbb{P}(X = i) = \frac{3}{\pi^2 i^2}$ pour $i \in \mathbb{Z}^\star$, alors X ne possède pas de moyenne (la condition de non existence est le fait que $\sum_{i:x_i\geq 0} p_i x_i = +\infty$ et $\sum_{i:x_i<0} p_i x_i = -\infty$, si les x_i ont proba. p_i).

Proposition 2.28. Soit X une v.a.r sur $(\Omega, \mathbb{P})$ et soit Ψ une fonction de $\mathbb{R}$ dans $\mathbb{R}$. Alors $\Psi(X)$ est une v.a.r. sur $(\Omega, \mathbb{P})$ et

$$\mathbb{E}(\Psi(X)) = \sum_{\omega\in\Omega} \Psi(X(\omega))\,\mathbb{P}(\omega) \tag{2.13}$$

$$= \sum_{x\in X(\Omega)} \Psi(x)\mathbb{P}(X = x) \tag{2.14}$$

(lorsque cette quantité existe).

Exemple 2.14. Dans l'affaire de la dette d'Alice X. Supposons qu'elle ait promis qu'elle donnerait $\Phi(X) = X^2 - 3X - 4$ à sa grande tante, si X est sa dette après la partie avec Bob. On voit qu'on peut tout ré-exprimer en fonction de a sur Ω puisque $X(a) =$

$(a-4)(a-1)$. Ainsi on voit que $\Phi(X)$ (on devrait écrire $a \mapsto \Phi(X(a))$) définie par $\Phi(X(a)) = ((a-4)(a-1))^2 - 3(a-4)(a-1) - 4$ est une variable aléatoire sur l'espace initial, et c'est cela qui donne la première formule: on peut sommer sur toutes les possibilités initiales, et pondérer par les probas initiales. Comme $\Phi(X)$ est une fonction de X dont on connaît aussi la loi, on peut à la place, travailler sous la loi de X ce qui donne la 2ème formule. Finalement, on trouve

$$\mathbb{E}(\Phi(X)) = \sum_{a=1}^{6} \frac{1}{6}\Big(((a-4)(a-1))^2 - 3(a-4)(a-1) - 4\Big),$$

qui vaut aussi, donc

$$\mathbb{E}(\Phi(X)) = \sum_{b\in\{-2,0,4,10\}} \mathbb{P}(X=b)\Phi(b)$$

avec $\mathbb{P}(X=-2) = \mathbb{P}(X=0) = 2/6, \mathbb{P}(X=4) = 1/6, \mathbb{P}(X=10) = 1/6$ comme expliqué dans Exemple 2.11 page 19.

Preuve. La preuve suit l'exemple ci-dessus. Il faut d'abord remarquer que $\Psi(X)$ est une variable aléatoire. C'est pour cela qu'elle possède une espérance. Il s'agit de l'application $\begin{array}{rccc} f: & \Omega & \longrightarrow & \mathbb{R} \\ & \omega & \longmapsto & \Psi(X(\omega)) \end{array}$; la formule (2.13) suit directement cette observation. Maintenant (lorsque la somme converge absolument[2]), on peut rassembler les termes qui ont même image par X:

$$\begin{aligned} \sum_{\omega\in\Omega} \Psi(X(\omega))\,\mathbb{P}(\omega) &= \sum_{x\in\mathbb{R}} \sum_{\omega\in\Omega:X(\omega)=x} \Psi(X(\omega))\,\mathbb{P}(\omega) \\ &= \sum_{x\in\mathbb{R}} \Psi(x) \sum_{\omega\in\Omega:X(\omega)=x} \mathbb{P}(\omega) = \sum_{x\in\mathbb{R}} \Psi(x)\mathbb{P}(X=x). \end{aligned}$$

Une fois encore, dans la dernière somme, au plus un nombre dénombrable de x contribuent. □

Proposition 2.29. L'espérance est linéaire: si a et b sont des réels et X une variable aléatoire réelle qui possède une espérance, alors

$$\mathbb{E}(aX+b) = a\mathbb{E}(X) + b.$$

Preuve. Il suffit d'utiliser Proposition 2.28 avec $\Psi(X) = aX + b$. □

[2]Il faut justifier le fait qu'on a réarrangé l'ordre de sommation, car ce n'est pas une opération toujours valide: un critère dû à Fubini, dit qu'on peut réordonner si les éléments de la somme sont positifs ou si les sommes convergent lorsqu'on ajoute une valeur absolue autour de $|\Psi(X)|$.

Remarque 2.30. *(i) Attention, en général* **on n'a pas** $\mathbb{E}(\Psi(X)) = \Psi(\mathbb{E}(X))$.

(ii) En prenant $a = 0$ *dans cette formule, on voit que pour une constante* b*, on a* $\mathbb{E}(b) = b$*, formule que l'on a déjà utilisée plusieurs fois.*

Définition 2.31. La variance de X (ou de la loi de X) est définie par:

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

(lorsque cette quantité existe).

Une fois encore, $\mathbb{E}(X)$ peut être non définie, donc dans ce cas la variance n'existe pas. Lorsque $\mathbb{E}(X)$ existe, $\mathsf{Var}(X)$ existe et peut être infinie.

La variance "mesure" la dispersion de X autour de sa moyenne. Notez qu'il s'agit des écarts quadratiques moyens... Plus la variance est grande, plus la dispersion l'est également. La variance peut être infinie.

Définition 2.32. L'écart type de X est défini par

$$\sigma_X = \sqrt{\mathsf{Var}(X)}$$

L'écart type semble plus naturel que la variance pour mesurer les écarts à la moyenne, car dans la variance il y a un carré... Il y a plus naturel encore, comme l'écart absolu moyen $\mathbb{E}(|X - \mathbb{E}(X)|)$ (notez que l'écart moyen $\mathbb{E}(X - \mathbb{E}(X))$, sans valeur absolue, vaut 0 et ne mesure rien du tout, puisque $\mathbb{E}(X)$ étant une constante, par la Proposition 2.29, $\mathbb{E}(X - \mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(X)$). Seulement, c'est bien la variance que la tradition probabiliste a consacrée. La principale raison tient au fait que la variance se calcule bien et apparaît comme étant un paramètre naturel dans de nombreux modèles...

Proposition 2.33. [Propriétés de la variance]
Soit X est une v.a.r. sur $(\Omega, \mathbb{P})$, a et b deux réels.

(i) $\mathsf{Var}(X) = \sum_x (x - \mathbb{E}(X))^2 \mathbb{P}(X = x)$

(ii) $\mathsf{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$

(iii) $\mathsf{Var}(aX + b) = a^2 \mathsf{Var}(X)$

Preuve. (i) On applique la Proposition 2.28: $\Psi(X) = (X - m)^2$ avec $m = \mathbb{E}(X)$.

(ii) On développe la formule dans définition 2.31: $\mathbb{E}((X - m)^2) = \mathbb{E}(X^2 + m^2 - 2Xm)$. On utilise encore la linéarité de l'espérance ici (il suffit encore d'utiliser (2.14) pour montrer qu'on a bien le droit): on obtient $\mathbb{E}((X - m)^2) = \mathbb{E}(X^2) + m^2 - 2m \times m = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

(*iii*) Tout d'abord, on sait que $\mathbb{E}(aX+b) = a\mathbb{E}(X)+b$. Ainsi, $\mathbb{E}((aX+b-\mathbb{E}(aX+b))^2) = \mathbb{E}((aX - a\mathbb{E}(X))^2)$ et on conclut aisément. □

Exemple 2.15. Si X est une variable de loi $\mathbb{P}(X = 1) = 1/3$, $\mathbb{P}(X = 2) = 1/6$, $\mathbb{P}(X = 4) = 1/2$, alors

$$\mathbb{E}(X) = \frac{1}{3} \times 1 + \frac{1}{6} \times 2 + \frac{1}{2} \times 4 = 8/3,$$

et

$$\mathbb{E}(X^2) = \frac{1}{3} \times 1^2 + \frac{1}{6} \times 2^2 + \frac{1}{2} \times 4^2 = 9,$$

donc

$$\mathsf{Var}(X) = 9 - (8/3)^2 = 17/9,$$

et l'écart type est $\sigma_X = \sqrt{17/9}$.

Proposition 2.34. Si X est une variable aléatoire telle que $\mathsf{Var}(X) = 0$ alors $\mathbb{P}(X = \mathbb{E}(X)) = 1$ (on dit que X est égale à son espérance ou, en jargon probabiliste, presque sûrement égale à son espérance).

Preuve. Utilisons la première formule de la Proposition 2.33: puisque $\mathsf{Var}(X) = \sum_x (x - \mathbb{E}(X))^2 \mathbb{P}(X = x)$ est une moyenne pondérée de nombres clairement positifs, si cette moyenne est nulle, c'est que les éléments de la somme sont nuls. Puisque $\mathbb{E}(X)$ est constante, ça veut dire que ou bien $x = \mathbb{E}(X)$, ou bien $\mathbb{P}(X = x) = 0$. Donc, il y a un seul x pour lequel la proba $\mathbb{P}(X = x)$ est non nul, c'est $x = \mathbb{E}(X)$. □

2.3 Deux inégalités...

2.3.1 Inégalité de Markov

Idée intuitive: si X est une variable positive de moyenne finie. La probabilité que X soit grand est petite.

Proposition 2.35. [Inégalité de Markov] Soit X une variable aléatoire positive sur Ω. Pour tout $x \in \mathbb{R}^+$, on a:

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}$$

Preuve. Il s'agit d'une inégalité facile à démontrer:

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) \geq \sum_{\omega \in \Omega : X(\omega) \geq x} X(\omega)\mathbb{P}(\omega) \geq \sum_{\omega \in \Omega : X(\omega) \geq x} x\,\mathbb{P}(\omega) = x\,\mathbb{P}(X \geq x).$$

□

Remarque 2.36. *Il y a des cas particuliers où on peut calculer directement* $\mathbb{P}(X \geq x)$, *par exemple, si* X *est une variable géométrique de paramètre* p, *alors*

$$\mathbb{P}(X \geq x) = (1-p)^{x-1}$$

(car ceci correspond à ce que toutes les premières tentatives ont échouées...). L'inégalité de Markov, dans ce cas donne juste

$$\mathbb{P}(X \geq x) \leq \mathbb{E}(X)/x = 1/(px)$$

qui est une inégalité de mauvaise qualité.

Exemple 2.16. Prenons une variable binomiale Y de paramètre (n,p), donc de moyenne np,

$$\mathbb{P}(Y \geq x) \leq np/x.$$

Notez que cette inégalité n'a aucune utilité lorsque $x \leq np$. Ainsi, si $n = 200$ (200 expériences), et $p = 1/2$ (une chance sur 2 de succès)

$$\mathbb{P}(Y \geq x) \leq 100/x$$

et l'inégalité devient non triviale pour $x \geq 100$.

2.3.2 Inégalité de Bienaymé-Tchebichev

Idée intuitive: X a peu de chance d'être loin de sa moyenne. De plus, la dispersion est d'autant plus petite que la variance l'est.

Proposition 2.37. [Inégalité de Bienaymé-Tchebichev] Soit X une v.a.r. et $x > 0$ un réel. On a:

$$\mathbb{P}(\mid X - \mathbb{E}(X) \mid \geq x) \leq \frac{\mathsf{Var}(X)}{x^2}.$$

Cette inégalité montre l'intérêt de la variance pour mesurer la dispersion d'une v.a..

Preuve.

$$\begin{aligned} \mathbb{P}(\mid X - \mathbb{E}(X) \mid \geq x) &= \mathbb{P}\left(\mid X - \mathbb{E}(X) \mid^2 \geq x^2\right) \\ &\underset{Markov}{\leq} \frac{\mathbb{E}\left(\mid X - \mathbb{E}(X) \mid^2\right)}{x^2} = \frac{\mathsf{Var}(X)}{x^2}. \end{aligned}$$

□

Exemple 2.17. Reprenons X suivant la loi binomiale (n,p): l'inégalité de Bienaymé-Tchebichev dit que, pour tout $x>0$,

$$\mathbb{P}\left(|X-np|\geq x\right)\leq\frac{\mathsf{Var}(X)}{x^2}=\frac{np(1-p)}{x^2}$$

et donc, cette fois, dès que $x\geq\sqrt{np(1-p)}$, on a une info non triviale. Toujours pour $(n,p)=(1000,1/2)$, on obtient, par exemple, pour $x=50$,

$$\mathbb{P}(|X-500|\geq 50)\leq 250/50^2=1/10.$$

La vraie proba est 0.00173... donc, une fois encore, cette inégalité n'est pas super bonne... Si on prend $x=20$, on trouve $\mathbb{P}(|X-500|\geq 20)\leq 250/20^2=0.625$ alors que $\mathbb{P}(|X-500|\geq 20)=0.217....$ L'inégalité est plus performante ici sans être super précise non plus.

2.4 Lois discrètes importantes

2.4.1 loi de Bernoulli

X est une variable de Bernoulli de paramètre $p\in[0,1]$ si

$$\mathbb{P}(X=1)=p,\quad\mathbb{P}(X=0)=1-p.$$

On note $X\sim\mathcal{B}(p)$.

$$\mathbb{E}(X)=p,\quad\mathsf{Var}(X)=p(1-p).$$

Pour le calcul de la variance, il faut remarquer que

$$\mathbb{E}(X)=p\times 1+(1-p)\times 0=p,\qquad\mathbb{E}(X^2)=p\times 1^2+(1-p)\times 0^2=p$$

et donc

$$\mathsf{Var}(X)=\mathbb{E}(X^2)-\mathbb{E}(X)^2=p-p^2=p(1-p).\tag{2.15}$$

Exemple 2.18. Toto lance une pièce où il y a une face 1, et une face 0. Chaque fois, il fait pile ou face avec proba $1/2$, et les résultats des lancers sont indépendants. Notons P le produit de 12 résultats successifs: Chaque lancer est une $\mathcal{B}(1/2)$, et $P\sim\mathcal{B}(1/2^{12})$.

2.4.2 loi binomiale

X est une variable binomiale de paramètre (n,p), $p\in[0,1]$ si

$$\mathbb{P}(X=k)=C_n^k p^k(1-p)^{n-k}\text{ pour }k\in\{0,1,\cdots,n\}\text{ et }0\text{ sinon.}\tag{2.16}$$

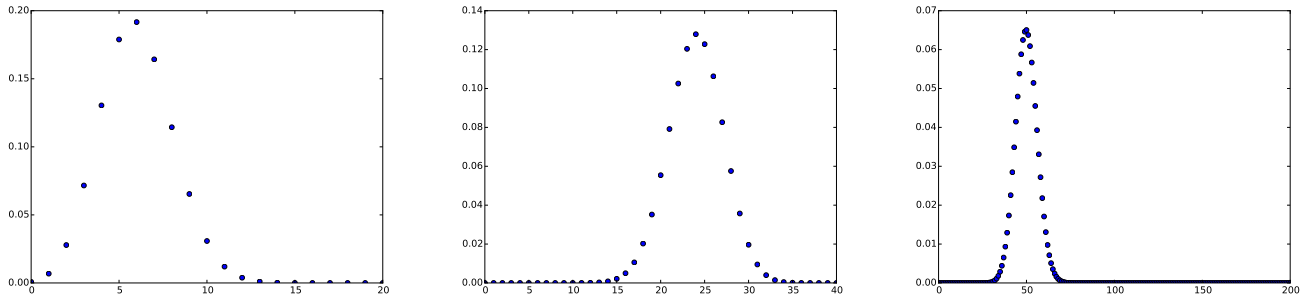


Figure 4: Distribution binomial paramètres (20,0.3), (40,0.6) et (200,0.25)

Lemme 2.38. Si $B_1, \cdots, B_n$ sont des variables de Bernoulli de paramètre p indépendantes alors la loi de $B_1 + \cdots + B_n$ est la loi binomiale de paramètre (n, p).

Preuve. Calculons $\mathbb{P}(B_1 + \cdots + B_n = k)$ pour un k dans $\{0, 1, \cdots, n\}$. On voit que parmi les B_i, k doivent valoir 1, et $n - k$ doivent valoir 0. Puisqu'il y a $\binom{n}{k}$ sous ensembles d'indices de cardinal n, et que la probabilité que k B_i's données valent 1 est p^k et que la proba pour les autres de valoir 0 est $(1-p)^{n-k}$, on a bien $\mathbb{P}(B_1 + \cdots + B_n = k) = C_n^k p^k (1-p)^{n-k}$. □

On en déduit que pour $X \sim B(n, p)$.

$$\mathbb{E}(X) \;=\; \mathbb{E}(B_1 + \cdots + B_n) = np \tag{2.17}$$

et

$$\mathsf{Var}(X) \;=\; \mathsf{Var}(B_1 + \cdots + B_n) = \mathsf{Var}(B_1) + \cdots + \mathsf{Var}(B_n) = np(1-p) \tag{2.18}$$

puisque les Bernoulli sont indépendantes.

Remarque 2.39. *Les variables $B(n, p)$ sont des variables qui "comptent" le nombre de succès lors de n expériences indépendantes de probabilité de succès p.*

2.4.3 loi uniforme

X est une variable uniforme sur $\{1, \cdots, n\}$ si

$$\mathbb{P}(X = i) = \frac{1}{n} \text{ pour tout } i \in \{1, \cdots, n\}$$

On note $X \sim \mathcal{U}(\{1, \cdots, n\})$. On a

$$\mathbb{E}(X) = \frac{n+1}{2} \quad , \text{ et } \quad \mathsf{Var}(X) = \frac{n^2 - 1}{12}. \tag{2.19}$$

Preuve.

$$\mathbb{E}(X) = \sum_{k=1}^{n} \frac{k}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

et

$$\mathbb{E}(X^2) = \sum_{k=1}^{n} \frac{k^2}{n} = \frac{n(n+1)(2n+1)}{6n} = \frac{(n+1)(2n+1)}{6}$$

d'où

$$\mathsf{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}.$$

□

2.4.4 loi géométrique

X est une variable géométrique de paramètre $0 < p < 1$ si

$$\mathbb{P}(X = k) = p(1-p)^{k-1} \text{ pour } k \geq 1.$$

On notera $X \sim \mathsf{Geo}(p)$ pour dire que G suit la loi géométrique de paramètre p.

Lemme 2.40. Soit $(B_1, B_2, \cdots)$ une suite de variables de Bernoulli de paramètre p indépendantes. Soit $G = \min\{k \ : B_k = 1\}$ l'indice de la première Bernoulli qui vaut 1. On a

$$G \sim \mathsf{Geo}(p).$$

Preuve. On voit bien que G vaut k sssi $B_1 = \cdots = B_{k-1} = 0$ et $B_k = 1$. La probabilité de cet événement est $(1-p)^{k-1}p$. □

On a

$$\mathbb{E}(X) = 1/p, \quad \text{et} \quad \mathsf{Var}(X) = (1-p)/p^2 \tag{2.20}$$

Preuve. 1ere preuve: On remarque que G fait 1 avec proba p, et en cas d'échec a même loi que $G' = 1 + G$ puisqu'il faut recommencer, mais on a déjà fait un lancer... On voit alors que

$$\mathbb{E}(G) = p + (1-p)\mathbb{E}(G') = p + (1-p) + (1-p)\mathbb{E}(G).$$

On résout et on trouve $\mathbb{E}(G) = 1/p$. Même chose, pour

$$\begin{aligned} \mathbb{E}(G^2) = p + (1-p)\mathbb{E}((1+G)^2) &= p + (1-p)(\mathbb{E}(G)^2 + 2\mathbb{E}(G) + 1) \\ &= p + (1-p)(\mathbb{E}(G)^2 + 2/p + 1) \end{aligned}$$

On résout, on trouve $\mathbb{E}(G^2) = (2-p)/p^2$, puis on utilise $\mathsf{Var}(G) = \mathbb{E}(G^2) - \mathbb{E}(G)^2 = (1-p)/p^2$.
2ème preuve: On peut faire le calcul en procédant comme suit:

$$\mathbb{E}(X) \;=\; p\sum_{k\geq 1} k(1-p)^{k-1}$$

Il s'agit de p fois une somme infinie; pour nous en sortir, il faut reconnaître dans cette somme une identité. Il faut oublier que p est une constante, et remplacer p par x et voir le résultat comme une fonction de x (une série entière): On écrit

$$\sum_{k\geq 1} k(1-x)^{k-1} \;=\; \sum_{k\geq 0} \big(-(1-x)^k\big)' = -\big(1/x\big)' = 1/x^2$$

et maintenant, on voit donc que $\mathbb{E}(X) = p/p^2 = 1/p$. Pour la variance, même chose:

$$\mathsf{Var}(X) = \mathbb{E}(X(X-1)) + \mathbb{E}(X) - \mathbb{E}(X)^2. \tag{2.21}$$

on voit qu'on a besoin de calculer

$$\mathbb{E}(X(X-1)) \;=\; p(1-p)\sum_{k\geq 2} k(k-1)(1-p)^{k-2}$$

De nouveau, on introduit

$$\begin{aligned}\sum_{k\geq 0} k(k-1)(1-x)^{k-2} \;&=\; \sum_{k\geq 0}\big((1-x)^k\big)''\\ &=\; (1/x)'' = \frac{2}{x^3}\end{aligned}$$

d'où $\mathbb{E}(X(X-1)) = 2p(1-p)/p^3$ et enfin[3]

$$\mathsf{Var}(X) = 2(1-p)/p^2 + 1/p - 1/p^2 = (1-p)/p^2.$$

□

2.4.5 loi de Poisson

X suit une loi de Poisson de paramètre λ, (on note $X \sim \mathsf{Poisson}(\lambda)$) si

$$\mathbb{P}(X=k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ pour } k \in \mathbb{N} \tag{2.22}$$

On a

$$\mathbb{E}(X) = \lambda, \quad \mathsf{Var}(X) = \lambda. \tag{2.23}$$

L'importance de cette loi apparaîtra plus tard.

[3]Pour rendre cette preuve rigoureuse, il faut bien sûr justifier le fait qu'on ait dérivé (puis intégré) ces séries terme à terme. On a "bien le droit", comme on l'apprend dans les cours sur les séries de fonctions, car on a affaire à des séries entières. La convergence est uniforme dans le disque de convergence ($\mathcal{C}(0,1)$ ici) aussi bien pour $\sum_{k\geq 1} k(1-x)^{k-1}$ que pour ses dérivées.

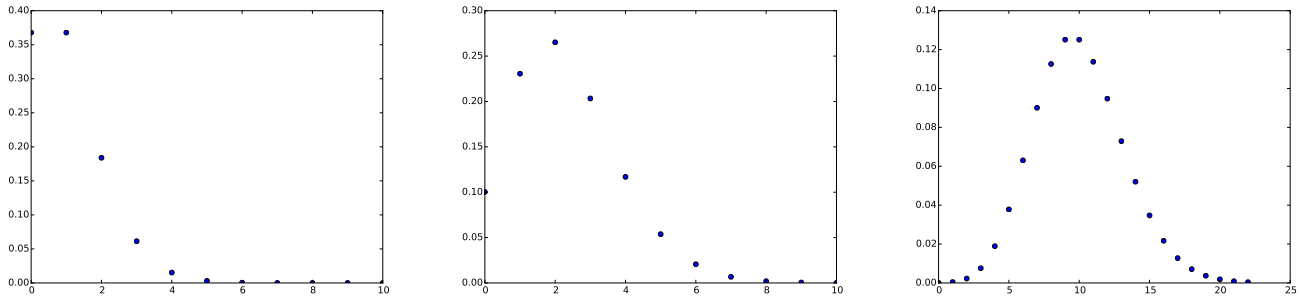


Figure 5: Distribution de Poisson de paramètres 1, 2.3 et 10.

Preuve. Tout d'abord, rappelons l'identité importante:

$$\exp(a) = \sum_{k=0}^{+\infty} a^k/k!, \text{ pour tout } a \in \mathbb{R}.$$

On a donc

$$\mathbb{E}(X) = \sum_{k\geq 0} \frac{ke^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\lambda \sum_{k\geq 1} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

Ensuite, on écrit,

$$\mathsf{Var}(X) = \mathbb{E}(X(X-1)) + \mathbb{E}(X) - \mathbb{E}(X)^2.$$

Et donc, il reste à calculer

$$\mathbb{E}(X(X-1)) = e^{-\lambda} \sum_{k\geq 2} \frac{k(k-1)\lambda^k}{k!} = e^{-\lambda}\lambda^2 \sum_{k\geq 2} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2$$

d'où $\mathsf{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$. □

2.5 Loi d'un couple de v.a.

Prenons X et Y deux variables aléatoires définies sur un même espace de probabilité discret $(\Omega, \mathbb{P})$ et à valeurs respectivement dans Ω' et Ω''. On s'intéresse à la paire (X, Y), que l'on considère comme une variable aléatoire définie sur Ω est à valeurs dans $\Omega' \times \Omega''$: c'est juste l'application

$$\begin{array}{rccc} (X,Y): & \Omega & \longrightarrow & \Omega' \times \Omega'' \\ & \omega & \longmapsto & (X(\omega), Y(\omega)) \end{array}.$$

Exemple 2.19. Par exemple, dans le jeu du dé on peut imaginer qu'en plus d'Alice et Bob, 2 autres joueurs, Christophe et Denis ont décidé que si le dé tombait sur a alors Christophe donnerait $Y(a) = (a-1)^2(a-4)$ à Denis. Maintenant, on voit que la paire (X, Y) est une

variable aléatoire de $\{1,2,3,4,5,6\}$ à valeurs dans $\mathbb{R}^2$.

	1	2	3	4	5	6
X	0	-2	-2	0	4	10
Y	0	-2	-4	0	16	50

L'ensemble des valeurs prises par X et Y est discret, et donc celles prises par (X,Y) aussi; dans cet exemple, il s'agit de $\{(0,0),(-2,-2),(-2,-4),(4,16),(10,50)\}$. La loi de la paire décrit une fois encore, pour chaque partie de l'ensemble image, la probabilité que (X,Y) soit dans cet ensemble. On a ici, par exemple:

$$\begin{aligned}\mathbb{P}((X,Y)=(0,0)) &= 2/6,\\ \mathbb{P}((X,Y)=(4,16)) &= 1/6,\\ \mathbb{P}((X,Y)=(10,-2)) &= 0.\end{aligned}$$

Définissons le cadre général:

Définition 2.41. Soient X et Y deux variables aléatoires définies sur un même espace de probabilité discret $(\Omega,\mathbb{P})$ et à valeurs respectivement dans Ω' et Ω''. La loi du couple (X,Y) est la mesure de probabilité sur $\Omega'\times\Omega''$ définie par

$$\mathbb{P}((X,Y)\in A)=\mathbb{P}(\{\omega:(X(\omega),Y(\omega))\in A\}),$$

pour tout $A\in$ Parties$(X(\Omega)\times Y(\Omega))$. Puisque (X,Y) est une variable aléatoire discrète, la loi de (X,Y) est caractérisée par $\mathbb{P}((X,Y)=(x,y))$ pour tout $(x,y)\in X(\Omega)\times Y(\Omega)$.

Si l'on connaît la loi de jointe du couple (X,Y) alors pour tout $a\in\Omega'$ et tout $b\in\Omega''$, on a, par la formule des probabilités totales:

$$\begin{aligned}\mathbb{P}(X=a) &= \sum_{y\in Y(\Omega)}\mathbb{P}((X,Y)=(a,y))\\ \mathbb{P}(Y=b) &= \sum_{x\in X(\Omega)}\mathbb{P}((X,Y)=(x,b)).\end{aligned}$$

Donnée la loi du couple (X,Y), les lois de X et de Y s'appellent lois marginales. Ainsi, la loi du couple (X,Y), détermine la loi de X et de Y, mais l'inverse est faux.

Tableau de distribution

Pour deux variables discrètes, on peut représenter la loi de (X,Y) à l'aide d'une matrice potentiellement infinie $\mathbb{P}((X,Y)=(x_i,y_j))$, où x_i parcourt $X(\Omega)$ et y_j parcourt $Y(\Omega)$. Il y

a des choses que l'on peut bien voir dans ce tableau. Voici un exemple:

$X\backslash Y$	4	9	12	15	*loi de* X
0	0.1	0.05	0.02	0.12	0.29
1	0.07	0.1	0	0.08	0.25
2	0.02	0.06	0.14	0.04	0.26
3	0.05	0.1	0.04	0.01	0.20
loi de Y	0.24	0.31	0.20	0.25	1

Sommer sur les lignes, resp. sur les colonnes, donne la marginale loi de X, et celle de Y.

Définition 2.42. Soient X et Y deux variables discrètes. Les variables X et Y sont dites indépendantes, si pour tout $A \in \text{Parties}(X(\Omega))$, pour tout $B \in \text{Parties}(Y(\Omega))$,

$$\mathbb{P}((X,Y) \in A \times B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B). \tag{2.24}$$

Ainsi, les variables X et Y sont dites indépendantes, si tout événement qui concerne la première est indépendant de tout événement qui concerne la seconde: pour tout A et tout B inclus respectivement dans Ω' et Ω'', l'événement $\{\omega : X(\omega) \in A\}$ est indépendant de $\{\omega : Y(\omega) \in B\}$.

Exemple 2.20. En regardant un peu l'exemple 2.9, on voit que si on définit sur $\Omega = \{1,2,3,4,5,6\}$ les variables aléatoires $X(a) = a \bmod 2$ et $Y(a) = a \bmod 3$, alors si Ω est muni de l'équiprobabilité, X et Y sont indépendantes.

Proposition 2.43. X et Y sont indépendantes si et seulement si pour tout $(a,b) \in X(\Omega) \times Y(\Omega)$

$$\mathbb{P}((X,Y) = (a,b)) = \mathbb{P}(X = a)\mathbb{P}(Y = b). \tag{2.25}$$

Preuve.

- D'abord, on voit que si on prend A et B réduits respectivement à $\{a\}$ et $\{b\}$, la formule (2.24) implique (2.25).
- Maintenant, pour la réciproque, notons que

$$\mathbb{P}((X,Y) \in A \times B) = \sum_{(a,b)\in A\times B} \mathbb{P}((X,Y) = (a,b)) = \sum_{a\in A}\sum_{b\in B} \mathbb{P}(X = a)\mathbb{P}(Y = b) \tag{2.26}$$

$$= \left(\sum_{a\in A} \mathbb{P}(X = a)\right)\left(\sum_{b\in B} \mathbb{P}(Y = b)\right) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \tag{2.27}$$

□

Dans la représentation de la loi par tableau de distribution, on voit qu'il y a indépendance si la masse d'une case égale le produit de la masse de la ligne et de la colonne qui la contiennent.

Loi de la somme

Soit X et Y deux variables aléatoires réelles sur un même espace de probabilité $(\Omega, \mathbb{P})$. Notons Z la variable aléatoire définie par

$$Z = X + Y.$$

Alors pour tout $c \in \mathbb{R}$

$$\mathbb{P}(Z = c) = \sum_{x \in X(\Omega)} \mathbb{P}((X, Y) = (x, c - x)) \tag{2.28}$$

Si X et Y sont indépendantes, on a:

$$\mathbb{P}(Z = c) = \sum_{x \in X(\Omega)} \mathbb{P}(X = x)\mathbb{P}(Y = c - x).$$

La loi de Z est donnée par le produit de convolution de la loi de X par celle de Y. Bien sûr, on peut définir de la même manière la loi de n'importe quelle fonction de 2 variables aléatoires:

$$\mathbb{P}(f(X, Y) = k) = \sum_{(x,y): f(x,y)=k} \mathbb{P}((X, Y) = (x, y))$$

et aussi l'espérance de la somme de deux variables:

$$\mathbb{E}(f(X, Y)) = \sum_{(x,y)} f(x, y)\mathbb{P}((X, Y) = (x, y)). \tag{2.29}$$

Proposition 2.44. (Propriétés de l'espérance mathématique)
Soient X et Y deux v.a.r. définie sur $(\Omega, \mathbb{P})$, et λ et c deux réels: L'espérance est linéaire: si X et Y ont tous deux une moyenne,

(i) On a $\mathbb{E}(\lambda X) = \lambda\mathbb{E}(X)$

(ii) et $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

(iii) Si X et Y sont indépendantes alors, si ces quantités existent

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y). \tag{2.30}$$

Preuve. (i) Multiplier par λ commute avec la somme.

(ii)

$$\mathbb{E}(X+Y) = \sum_{(x,y)}(x+y)\mathbb{P}((X,Y)=(x,y)) = \sum_x \left(\sum_y (x+y)\mathbb{P}(X=x,Y=y) \right) \quad (2.31)$$

$$= \sum_x \left(\sum_y x\mathbb{P}(X=x,Y=y) + \sum_y y\mathbb{P}(X=x,Y=y) \right) \quad (2.32)$$

$$= \sum_x x\mathbb{P}(X=x) + \sum_y y\mathbb{P}(Y=y) = \mathbb{E}(X)+\mathbb{E}(Y) \quad (2.33)$$

(iii) On écrit, une fois encore $\mathbb{E}(XY) = \sum_{x,y} xy\mathbb{P}((X,Y)=(x,y)) = \sum_{x,y} xy\mathbb{P}(X=x)\mathbb{P}(Y=y) = (\sum_x x\mathbb{P}(X=x))(\sum_y y\mathbb{P}(y=y)) = \mathbb{E}(X)\mathbb{E}(Y)$. □

Bon, on a mis un peu de poussière sous le tapis ici, disons que c'est comme cela que ça marche formellement, mais lorsque les sommes sont infinies, des problèmes peuvent apparaître. Il faut que les sommes convergent absolument (en ajoutant une valeur absole sous le signe somme) pour s'assurer de la validité du réarrangement des sommes.

2.5.1 Probabilité conditionnelle. Indépendance de v.a.r.

Définition 2.45. Soient X et X deux v.a.r. définies sur le même espace de probabilité Ω et à valeurs dans Ω' et Ω''. Soient x et y deux éléments de Ω' et Ω''. La probabilité conditionnelle de $X=x$ sachant $Y=y$ est définie pour $(x,y)\in\Omega'\times\Omega''$ tels que $\mathbb{P}(Y=y)>0$ par :

$$\mathbb{P}(X=x \mid Y=y) = \frac{\mathbb{P}\big((X=x)\cap(Y=y)\big)}{\mathbb{P}(Y=y)}.$$

Remarque 2.46. *Notons que si X et Y sont indépendantes, alors, si $\mathbb{P}(Y=y)>0$, on a:*

$$\mathbb{P}(X=x \mid Y=y) = \frac{\mathbb{P}\big((X=x)\cap(Y=y)\big)}{\mathbb{P}(Y=y)} = \frac{\mathbb{P}(X=x)\mathbb{P}(Y=y)}{\mathbb{P}(Y=y)} = \mathbb{P}(X=x)$$

Une fois encore, en cas d'indépendance, on n'interprète les choses comme suit: quelle que soit l'information qu'on apprend concernant Y, la probabilite de n'importe quel événement concernant X (c'est-à-dire, $X\in A$) est non modifiée.

Proposition 2.47. si X et Y sont deux v.a. de variance finie alors:

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathrm{cov}(X,Y) \tag{2.34}$$

où $\mathrm{cov}(X,Y)$ est la covariance de X et Y,

$$\begin{aligned} \mathrm{cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) && (2.35)\\ &= \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))), && (2.36) \end{aligned}$$

et dans le cas où X et Y sont indépendants, la covariance est nulle et

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y).$$

Preuve. On écrit tout d'abord

$$\begin{aligned} \mathsf{Var}(X+Y) &= \mathbb{E}\big((X+Y)^2\big) - (\mathbb{E}(X+Y))^2\\ &= \mathbb{E}(X^2+2XY+Y^2) - \mathbb{E}(X)^2 - \mathbb{E}(Y)^2 - 2\mathbb{E}(X)\mathbb{E}(Y)\\ &= \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\big(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\big). \end{aligned}$$

où on identifie bien la formule (2.34) avec la première formule de la covariance. Pour la 2ème formule de covariance, il suffit de développer $\mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))) = \mathbb{E}(XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y))$ et utiliser la linéarité de l'espérance pour voir que cela $= \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y)))$.

Maintenant, dans le cas où X et Y sont indépendants.

$$\begin{aligned} \mathbb{E}(XY) &= \sum_{\Omega} \mathbb{P}(\omega)X(\omega)Y(\omega)\\ &= \sum_{(x,y)\in X(\Omega)\times Y(\Omega)} \mathbb{P}(X=x \cap Y=y)xy\\ &= \left(\sum_{x\in X(\Omega)} \mathbb{P}(X=x)x\right)\left(\sum_{y\in Y(\Omega)} \mathbb{P}(Y=y)y\right) = \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

et donc $\mathrm{cov}(X,Y) = 0$ dans ce cas. □

En général, $\mathsf{Var}(X+Y) \neq \mathsf{Var}(X) + \mathsf{Var}(Y)$.

Exemple 2.21. Reprenons l'exemple 2.19 du jeu de dé, avec les 2 variables X et Y. On a

$$\mathbb{E}(X+Y) = \frac{2}{6}(0+0) + \frac{1}{6}(-2-2) + \frac{1}{6}(-2-4) + \frac{1}{6}(4+16) + \frac{1}{6}(10+50) = \frac{35}{3}$$

et

$$\mathbb{E}((X+Y)^2) = \frac{2}{6}(0+0)^2 + \frac{1}{6}(-2-2)^2 + \frac{1}{6}(-2-4)^2 + \frac{1}{6}(4+16)^2 + \frac{1}{6}(10+50)^2 = \frac{2026}{3}$$

et donc

$$\mathrm{Var}(X+Y) = \mathbb{E}((X+Y)^2) - \mathbb{E}(X+Y)^2 = \frac{2026}{3} - (\frac{35}{3})^2 = 4853/9 = 539.222...$$

Par ailleurs, on peut calculer $\mathbb{E}(X) = 5/3$, $\mathbb{E}(X^2) = 62/3$, $\mathbb{E}(Y^2) = 1388/3$, $\mathbb{E}(Y) = 10$, $\mathbb{E}(XY) = \frac{2}{6}(0\times 0) + \frac{1}{6}(-2\times(-2)) + \frac{1}{6}(-2\times(-4)) + \frac{1}{6}(4\times 16) + \frac{1}{6}(10\times 50) = 96$. La covariance est donc

$$\mathrm{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 96 - 10\times\frac{5}{3} = \frac{238}{3}$$

et on peut maintenant vérifier que le calcul de la variance $\mathrm{Var}(X+Y)$ pouvait se faire à l'aide de la covariance:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{cov}(X+Y) = (62/3-(5/3)^2)+(1388/3-10^2)+2\frac{238}{3} = 4853/9$$

Le signe de la covariance est important. Lorsque le signe est positif, X et Y ont "tendance" à être du même côté de leur moyenne en même temps. On dit que X et Y sont corrélées positivement (négativement si $\mathrm{cov}(X,Y) < 0$). Lorsque $\mathrm{cov}(X,Y) = 0$, on dit que X et Y ne sont pas corrélées (il peut arriver que $\mathrm{cov}(X,Y) = 0$ sans que X et Y soient indépendantes).

Proposition 2.48. Lorsque ces quantités existent, on a:

(i) $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$

(ii) $\mathrm{cov}(X,X) = \mathrm{Var}(X)$

(iii) $\mathrm{cov}(aX+b,Y) = a\,\mathrm{cov}(X,Y)$

Preuve. Les 3 propriétés sont très simples à vérifier. □

2.6 Plus de 2 variables ?

On peut bien sûr définir la loi de k variables aléatoires définie sur le même espace comme la mesure de probabilité $(\Omega,\mathbb{P})$,

$$\mathbb{P}((X_1,\cdots,X_k)\in(A_1\times\cdots\times A_k)) = \sum_{\omega\in\Omega}\mathbb{P}((X_1(\omega),\cdots,X_k(\omega))\in(A_1\times\cdots\times A_k)).$$

Une fois encore, cette loi est déterminée par les quantités $\mathbb{P}((X_1,\cdots,X_k)=(x_1,\cdots,x_k))$ pour tout $(x_1,\cdots,x_k)$ de l'ensemble image. On définit encore l'indépendance de la famille par la propriété,

$$\mathbb{P}((X_{n_1},\cdots,X_{n_j})\in(A_{n_1}\times\cdots\times A_{n_k})) = \prod_{\ell=1}^{j}\mathbb{P}(X_{n_\ell}\in A_{n_\ell})$$

pour toute sous suite $n_1<\cdots<n_j$, et toutes parties $A_{n_\ell}\in\mathrm{Parties}(X_{n_\ell}(\Omega))$.

Par contre, on n'étend pas la notion de covariance à plus de 2 variables.

3 ——— Éléments de combinatoire ———

La combinatoire est la science des structures discrètes, comme par exemple, les permutations, les arbres, les graphes, les partitions, les triangulations, les suites de lettres satisfaisant certaines contraintes, etc.

L'algorithmique fait une grande utilisation de structures combinatoires, en particulier autour des structures de données, qui sont souvent des tableaux, des arbres, ou éventuellement des graphes. Ainsi de nombreux algorithmes travaillent explicitement sur des structures combinatoires, et connaître ces structures permet de mieux programmer, et également de comprendre la performance des algorithmes qui les utilisent.

Étudier les structures combinatoires consiste à les décomposer, "les compter" (combien d'objets de taille n ?), comprendre leur comportement typique (quelle est la hauteur moyenne d'un arbre binaire à 100 000 noeuds internes ?)

On va dans un premier temps expliquer un peu les méthodes utilisées en combinatoire pour décomposer les objets, et compter ceux de taille n. On verra ensuite que les méthodes probabilistes permettent de comprendre le comportement typique de certains paramètres de ces structures combinatoires. Les simulations informatiques permettront également de percevoir certains de ces aspects, et on verra comment engendrer exhaustivement certaines structures élémentaires, ou bien, comment calculer explicitement, le nombre d'objets de taille n à l'aide de petits programmes.

3.1 Principes généraux

Définition 3.1. On appelle classe combinatoire $\mathcal{A}$ un ensemble muni d'une fonction taille

$$\begin{array}{rccc} \text{taille}: & \mathcal{A} & \longrightarrow & \mathbb{N} \\ & a & \longmapsto & \text{taille}(a) \end{array}, \tag{3.1}$$

telle que, pour tout $n \in \mathbb{N}$, l'ensemble des objets de taille n, c'est-à-dire

$$\mathcal{A}_n := \{a \in \mathcal{A} : \text{taille}(a) = n\}$$

est de cardinal fini.

On notera la taille de x simplement $|x|$ au lieu de taille(x). La fonction génératrice de la classe $\mathcal{A}$ est la série formelle $G_{\mathcal{A}}$ définie par

$$G_{\mathcal{A}}(x) = \sum_{a \in \mathcal{A}} x^{|a|} = \sum_{n \geq 0} \#\mathcal{A}_n x^n.$$

Autrement dit, il s'agit de la série dont le coefficient de x^n est $\#\mathcal{A}_n$.

Exemple 3.22. ▪ Par exemple, l'ensemble $\mathcal{A} = \{\varepsilon, a, b, c, aa, ab, ac, ba, bb, bc, ca, cc, cc, aaa, \cdots\}$ des mots sur l'alphabet de 3 lettres $\{a, b, c\}$ est une classe combinatoire (où ε est le mot vide): il y a 3^n mots avec n lettres. Ansi, si on définit la taille d'un mot comme étant simplement son nombre de lettres, la fonction génératrice est

$$G_1(x) = \sum_{k\geq 0} 3^k x^k = \frac{1}{1-3x}.$$

Avec la même notion de taille, l'ensemble des mots sur l'alphabet à deux lettres $\{a, b\}$, qui contiennent le même nombre de lettres a et b est une classe combinatoire. On a

$$G_2(x) = \sum_{n\geq 0} x^{2n} \binom{2n}{n}.$$

▪ Si on prend l'ensemble $\mathbb{N}$, et qu'on dit que la taille de n est n, alors $\mathbb{N}$ muni de cette taille est une classe combinatoire. Il y a un objet de chaque taille, donc, sa fonction génératrice est $G_3(x) = \sum_{k=0}^{+\infty} x^k$.

▪ L'ensemble des paires d'entiers relatifs (a, b) munis de la taille $|(a, b)| = |a + b|$ n'est pas une classe combinatoire, car il y a un nombre infini de paires (a, b) tels que $|a + b| = n$, pour tout n. Si, maintenant on considère l'ensemble des paires d'entiers > 0 avec la même notion de taille, il s'agit cette fois d'une classe combinatoire.

Nous allons voir maintenant que si une classe combinatoire peut-être décomposée dans un certain sens alors on obtient automatiquement une formule pour la fonction génératrice de la classe en question... et avec un petit programme on peut calculer les valeurs des premiers $\#A_n$ en un temps raisonnable.

Remarque 3.2. *Certaines classes combinatoires ne semblent pas décomposables, et aucune formule pour leur fonction génératrice n'est connue. Par exemple, on ne sait pas combien il y a de chemins commençant en $(0, 0)$ avec pas $(0, 1)$ ou $(1, 0)$ ou $(0, -1)$ ou $(-1, 0)$, si on demande à ce que ces chemins ne passent pas 2 fois au même endroit. En l'absence de décomposition "même bancale", on n'a d'autre choix de compter les objets (presque) un par un.*

3.1.1 Principes de décomposition

1. [**Union disjointe**] Supposons que $\mathcal{A}$ et $\mathcal{B}$ sont deux classes disjointes de fonction taille $|.|_{\mathcal{A}}$ et $|.|_{\mathcal{B}}$. Si $\mathcal{C}$ est la classe combinatoire formée par $\mathcal{A} \dot\cup \mathcal{B}$ et avec fonction taille $|.|_{\mathcal{C}}$ qui coïncide avec $|.|_{\mathcal{A}}$ sur $\mathcal{A}$, et avec $|.|_{\mathcal{B}}$ sur $\mathcal{B}$, alors

$$G_{\mathcal{C}}(x) = G_{\mathcal{A}}(x) + G_{\mathcal{B}}(x).$$

En effet, écrivons juste $G_{\mathcal{C}}(x) = \sum_n x^n \#\mathcal{C}_n$ et décomposons $\#\mathcal{C}_n = \#\mathcal{A}_n + \#\mathcal{B}_n$.

Exemple 3.23. Si $\mathcal{A}$ est la classe combinatoire des mots sur l'alphabet $\{a, b\}$, et $\mathcal{B}$ celle des mots sur l'alphabet $\{c, d, e\}$, alors $\mathcal{C} := \mathcal{A} \dot\cup \mathcal{B}$ a pour fonction génératrice

$$G_{\mathcal{C}}(x) = \frac{1}{1-2x} + \frac{1}{1-3x} - 1,$$

(il y a un -1 pour éviter le double comptage du mot vide). De là on déduit $\#\mathcal{C}_n = 2^n + 3^n$ pour $n \geq 1$ (et $\#\mathcal{C}_0 = 1$).

2. [**Produit cartésien**] Si $\mathcal{C}$ est la classe combinatoire $\mathcal{A} \times \mathcal{B}$ (formée donc des paires (a, b) avec a dans $\mathcal{A}$ et b dans $\mathcal{B}$) et muni de la fonction taille vérifiant pour $c = (a, b) \in \mathcal{A} \times \mathcal{B}$, $|c|_{\mathcal{C}} = |a|_{\mathcal{A}} + |b|_{\mathcal{B}}$, alors

$$G_{\mathcal{C}}(x) = G_{\mathcal{A}}(x) G_{\mathcal{B}}(x). \tag{3.2}$$

La preuve de (3.2) est comme suit: les éléments de $\mathcal{C}$ de taille n sont les paires (a, b) avec $|a|_{\mathcal{A}} = k$ et $|b|_{\mathcal{B}} = n - k$ pour un k quelconque: ainsi

$$\begin{aligned} G_{\mathcal{C}}(x) &= \sum_{n\geq 0} \#\mathcal{C}_n x^n = \sum_{n\geq 0} \sum_{k=0}^{n} \#\mathcal{A}_k \#\mathcal{B}_{n-k} x^k x^{n-k} \\ &= \left(\sum_{k\geq 0} \#\mathcal{A}_k x^k \right) \left(\sum_{k\geq 0} \#\mathcal{B}_k x^k \right) = G_{\mathcal{A}}(x) G_{\mathcal{B}}(x). \end{aligned}$$

Si on prend $\mathcal{A}$ et $\mathcal{B}$ comme dans l'exemple précédent, pour $\mathcal{D} := \mathcal{A} \times \mathcal{B}$, alors

$$G_{\mathcal{D}}(x) = \frac{1}{1-2x} \frac{1}{1-3x} = \frac{-2}{1-2x} + \frac{3}{1-3x}$$

et donc $\#\mathcal{D}_n = -2^{n+1} + 3^{n+1}$ qui n'est pas si évident, cette fois...! Par exemple pour $n = 2$, on trouve $\#\mathcal{D}_2 = -2^3 + 3^3 = 19$ qui peut-être vérifié à la main. Observons que l'on a considéré le mot vide comme un mot possible dans $\mathcal{A}$ et dans $\mathcal{B}$. Si finalement, on ne veut pas les autoriser, il suffit de modifier un peu $\mathcal{A}$ et $\mathcal{B}$, et prendre $G_{\mathcal{A}}(x) = 2x/(1-2x)$ et $G_{\mathcal{B}}(x) = 3x/(1-3x)$, pour des raisons qu'on est invité à deviner...

3. [**Suites**] Si la classe combinatoire $\mathcal{C}$ est formée par les suites d'éléments de la classe $\mathcal{A}$ (avec répétition possible), c'est-à-dire si $\mathcal{C} = \varepsilon \cup \bigcup_{n\geq 1} \mathcal{A}^n$ (on écrit simplement $\mathcal{C} = \mathsf{SEQ}(\mathcal{A})$) avec fonction taille $|(a_1, \cdots, a_k)|_{\mathcal{C}} = \sum_{j=1}^{k} |a_j|_{\mathcal{A}}$, alors

$$G_{\mathsf{SEQ}(A)}(x) = \sum_{k\geq 0} G_{\mathcal{A}}(x)^k = \frac{1}{1 - G_{\mathcal{A}}(x)}.$$

Attention, cette formule n'est pas valide si des éléments de $\mathcal{A}$ ont pour taille 0 (auquel cas, il existerait un nombre infini d'éléments de la classe $\mathcal{C}$ de chaque taille). La raison est la suivante... L'élément $G_{\mathcal{A}}(x)^k$ compte les éléments de la classe $\mathsf{SEQ}(\mathcal{A})$ formés de k-uplet d'éléments de $\mathcal{A}$.

Exemple 3.24. Si $\mathcal{A}$ est la classe combinatoire des nombres $\{1, 2, 4\}$ dont les tailles sont simplement $|a| = a$, alors

$$G_{\mathsf{SEQ}(\mathcal{A})}(x) = \frac{1}{1 - x^1 - x^2 - x^4}.$$

Le coefficient de x^n dans $G_{\mathsf{SEQ}(A)}(x)$ est le nombre de manière d'écrire n comme somme de nombre égaux à 1, 2, ou 4... Par exemple pour $n = 5$, les 10 possibilités sont $(1,1,1,1,1)$, $(1,1,1,2)$,$(1,1,2,1)$, $(1,2,1,1)$, $(2,1,1,1)$, $(1,2,2)$,$(2,1,2)$,$(2,2,1)$,$(4,1)$,$(1,4$
En développant $G_{\mathsf{SEQ}(\mathcal{A})}(x)$ en série, on trouve

$$G_{\mathsf{SEQ}(\mathcal{A})}(x) = 1+x+2x^2+3x^3+6x^4+10x^5+18x^6+31x^7+55x^8+96x^9+169x^{10}+296x^{11}+\cdots$$

et le coefficient de x^5 est 10 comme attendu...

4. [**Multisets.**] Un multiset est, informellement, un ensemble dans lequel la répétition d'éléments est autorisée, et bien sûr, différentes multiplicités donne des multisets différents . Prenons l'ensemble des multisets $\mathcal{C} = \mathsf{MSET}(\mathcal{A})$ formés d'éléments de la classe combinatoire $\mathcal{A}$. On a

$$G_{\mathsf{MSET}(\mathcal{A})}(x) = \prod_{a\in\mathcal{A}} \frac{1}{1 - x^{|a|}} \tag{3.3}$$

$$= \prod_{n\geq 1} \frac{1}{(1 - x^n)^{|\mathcal{A}_n|}} \tag{3.4}$$

De nouveau cette formule n'est valide que lorsqu'il n'y a pas d'élément de $\mathcal{A}$ qui possèdent la taille 0. La preuve est simple: $\frac{1}{1-x^{|a|}}$ est la série génératrice de suites d'éléments égaux à a, et le produit $\prod_{a\in\mathcal{A}}$ signifie qu'on s'intéresse aux suites indexées par a: or un multiset est justement une suite indexée par a, de suites finies d'éléments égaux à a.

Exemple 3.25. Si l'ensemble $\mathcal{A}$ contient uniquement 2 éléments $\{a, b\}$ tous deux de taille 1, on trouve

$$G_{\mathsf{MSET}(\mathcal{A})}(x) = \frac{1}{(1 - x)^2} = \left(\sum_{k\geq 0} \frac{1}{1 - x}\right)' = \sum_{k\geq 0}(k + 1)x^k.$$

La raison est que les objets de taille k sont formés de j a's et $(k-j)$ b's pour un j allant de 0 à k (donc $k + 1$ possibilités).
Si $\mathcal{A}$ contient 2 éléments de taille 3 et 1 élément de taille 7, alors

$$G_{\mathsf{MSET}(\mathcal{A})}(x) = \left(\frac{1}{1 - x^3}\right)^2 \frac{1}{1 - x^7}.$$

Le coefficient de x^{567} dans ce produit est le nombre de manières d'écrire 567 sous $3k + 3\ell + 7m$ pour des (k, ℓ, m) dans $\mathbb{N}^3$; l'écriture d'un petit programme nous dit qu'il y a 2674 façons de faire.

Remarque 3.3. *La notion de sous-ensemble de k éléments d'un ensemble E (muni d'un ordre total) est équivalent à la notion de sous-suite strictement croissante de k termes de E, puisqu'on peut associer de manière bijective à chaque ensemble, la liste de ses éléments ordonnés. La notion de multiset de k éléments pris dans E coincide avec la notion de suite croissante (au sens large) de k éléments de E.*

3.2 Applications

3.2.1 Arbres binaires

Un arbre binaire enraciné est un arbre qui possède une racine donc, et dans lequel chaque noeud possède 0 enfants, ou 2 enfants (un droit et un gauche). Ainsi, un arbre binaire est, ou bien réduit à sa racine, ou est constitué d'une racine ayant 2 enfants, chacun d'eux étant racine d'un arbre binaire. On a donc une décomposition claire des arbres binaires. Si on veut

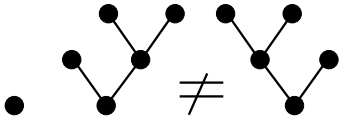


Figure 6: 3 arbres binaires différents. Le premier a un nœud, les 2 autres, 5.

tourner cela en série génératrice, alors, il faut écrire une formule sur les classes combinatoires, et tenir compte de la notion de taille qui doit coller à la décomposition (par exemple, si on écrit que $C = A \times B$, il faut en plus que la notion de taille naturelle sur C soit la somme de la taille sur A et de celle sur B pour que cela soit utile. Il faut donc réfléchir pas mal avant d'écrire une décomposition. Plusieurs solutions sont possibles.

Par exemple, on peut énumérer les arbres en terme du nombre de noeuds internes (ceux qui ont 2 enfants). Le nombre total de noeuds est $2n + 1$ s'il y a n noeuds internes, donc, on peut passer de la notion de taille, nombre de noeuds internes, à celle, nombre de noeuds, par une simple manipulation.

Notons ε celui réduit à sa racine; il est de taille 0. Maintenant, un arbre non réduit à sa racine possède une racine qui est donc un noeud interne, celle-ci a 2 enfants qui sont racine d'arbres potentiellement réduit à leur racine (de taille 0) ou de taille plus grande.

On traduit cela par la décomposition suivante:

$$B = \{\varepsilon\} \bigcup \{r\} \times B \times B. \tag{3.5}$$

Ici ε est un arbre vide, de taille 0,
r est une racine de taille 1,
et donc, on compte globalement la taille d'un arbre par son nombre de noeuds internes.

Il faut faire super attention ici; il s'agit d'une définition récursive des arbres. La moindre erreur entraîne une erreur gigantesque in fine. La série génératrice des arbres comptés selon

le nombre de noeuds internes, G satisfait donc l'équation

$$G(x) = 1 + xG(x)^2, \tag{3.6}$$

ou autrement dit si C_k donne le nombre d'arbres de taille k

$$C_n = \begin{cases} 1 & \text{si } n = 0 \\ \sum_{j=0}^{n-1} C_j C_{n-1-j} & \text{si } n > 0 \end{cases}. \tag{3.7}$$

En effet, l'équation (3.5) porte sur une union disjointe. La série génératrice de $\{\varepsilon\}$ est $x^0 = 1$ car cet ensemble ne contient qu'un objet de taille 0, et enfin, $\{r\} \times B \times B$, est un produit cartésien, et donc sa série génératrice est le produit des séries génératrices des ensembles concernés: la série génératrice de $\{r\}$ est $x^1 = x$, et celle de B est G.

On voit que G est solution d'un polynôme $g^2 - g/x + 1/x = 0$ du second degré en g. On peut le résoudre facilement en calculant le discriminant: on trouve

$$g = \frac{1/x \pm \sqrt{1/x^2 - 4/x}}{2}.$$

En développant en séries, on trouve que

$$G(x) = \frac{1/x - \sqrt{1/x^2 - 4/x}}{2} \tag{3.8}$$

(car l'autre n'est pas une série entière), et en développant en séries, on trouve $g(x) = 1 + x + 2\,x^2 + 5\,x^3 + 14\,x^4 + 42\,x^5 + 132\,x^6 + 429\,x^7 + 1430\,x^8 + 4862\,x^9 + 16796\,x^{10} + 58786\,x^{11} + 208012\,x^{12} + 742900\,x^{13} + ...$ et on trouve qu'il y a donc 742900 arbres avec 13 noeuds internes.

On peut compter aussi les arbres en fonction du nombre d'arêtes. L'arbre réduit à sa racine a 0 arête, et la décomposition cette fois est

$$C = \{\varepsilon\} \bigcup \{p\} \times C \times C$$

où p désigne la paire d'arête entre la racine et ses 2 enfants, donc un objet de taille 2. Les sous-arbres dessous peuvent possèder 0 arrêtes... La décomposition récursive marche. En termes de série génératrice, ça donne

$$H(x) = 1 + x^2 H(x)^2.$$

On peut encore résoudre l'équation de la même façon.

Finalement, si on note C_n le nombre d'arbres binaires à n noeuds internes, on trouve

$$C_n = \binom{2n}{n}/(n+1), \quad \forall n \geq 0 \tag{3.9}$$

ce qu'on peut démontrer soit par récurrence en partant de (3.7) soit en utilisant la formule (3.8), à l'aide d'un développement de Taylor.

3.2.2 Partitions / Compositions

Dans la Section 3.1, on a introduit les séries génératrices des suites et des multisets d'une classe combinatoire $\mathcal{A}$.

Définition 3.4. Prenons $n \in \mathbb{N}$.
- On appelle partition de n une suite finie croissante (au sens large) d'entiers > 0, $(k_1, \cdots, k_m)$ telle que $\sum_{j=1}^{m} k_j = n$. Par exemple l'ensemble des partitions de 5 est $\{(5), (4,1), (3,2), (3,1,1), (2,2,1), (2,1,1,1), (1,1,1,1,1)\}$.
- Une composition de n est une suite d'entiers > 0 $(k_1, \cdots, k_m)$ telle que $\sum_{j=1}^{m} k_j = n$. Par exemple, l'ensemble des compositions de 4 est $\{(4), (1,3), (3,1), (2,2), (2,1,1), (1,2,1), (1,1,2), (1,1,1,1)\}$.

Les seules partitions et compositions de 0 sont les suites vides. Pour une partition (ou composition) $(x_1, \cdots, x_k)$ de n, x_i est une part, k est appelé nombre de parts.

Théorème 3.5. La série génératrice des partitions est

$$G_{Part}(x) = \prod_{j=1}^{+\infty} \frac{1}{1-x^j}.$$

Preuve. La notion de partition coincide avec celle de multiset d'entiers strictement positif, pour une notion de taille de partition qui coincide avec la somme des entiers qui la compose. Ainsi, on est amené à poser que la taille de l'entier k est k. Maintenant, par la Formula (3.4), la fonction génératrice des partitions est $\prod_{n\geq 1} 1/(1-x^n)^1$. □

Remarque 3.6. *L'ensemble des partitions de n est en bijection avec l'ensemble des suites décroissantes (au sens large) d'entiers > 0 de somme n.*

Exercice 1. [Difficile] Un théorème dû à Euler dit qu'il y a autant de partitions de n dont les parts sont distinctes 2 à 2 que de partitions de n dont les parts sont toutes impaires (et ce, pour tout $n \geq 1$). Démontrez ce fait.

3.3 Permutations

3.3.1 Rappels

Il y a plusieurs manières de voir les permutations. On adopte la suivante: on appelle permutation d'un ensemble fini E, toute bijection de E dans E. Le nombre de bijections de E dans E, est simplement $\#E!$.

La structure de l'ensemble des bijections d'un ensemble ne dépend que du cardinal de l'ensemble... et donc, toute la structure des permutations d'un ensemble de taille n se comprend sur les permutations de l'ensemble $E_n = \{1, \cdots, n\}$ et la tradition est donc de travailler sur les permutations de E_n; on appelle $\mathcal{S}_n$, l'ensemble des permutations de E_n. Une permutation $\sigma \in \mathcal{S}_n$ est clairement donnée par la suite finie des images successives des éléments de E_n par σ:

$$(\sigma(1), \cdots, \sigma(n)).$$

Évidemment, on reconnaît une permutation au premier coup d'oeil, puisque la suite $(\sigma(1), \cdots, \sigma(n)$ doit contenir chaque élément de E_n exactement une fois.

Exemple 3.26. Les permutations de $\mathcal{S}(3)$ se représentent donc juste par les vecteurs d'images successifs, qui sont ici, $(1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)$.

3.3.2 Représentation matricielle

Une manière naturelle de représenter une permutation est la suivante: on associe à $\sigma \in \mathcal{S}_n$ la matrice

$$M_\sigma = \left[1_{\sigma(j)=i}\right]_{1 \leq i,j \leq n}$$

Autrement dit: dans la matrice il n'y a que des 0 et des 1, et il y a un 1 dans la case i, j, sssi $\sigma(i) = j$. Comme $\sigma(i) \neq \sigma(i')$ lorsque $i \neq i'$ (car σ) est une bijection, la matrice M_σ, possède exactement un 1 dans chaque ligne et un 1 dans chaque colonne...

Les matrices correspondant aux permutations données en (3.26) sont

Exemple 3.27.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

En fait, cela revient à faire le graphe de la fonction discrète σ (sauf que l'ordonnée croît "vers le bas").

3.3.3 Permutation aléatoire uniforme

Il y a $n!$ permutations de taille n... la loi uniforme sur l'ensemble $\mathcal{S}_n$ attribue donc une masse de

$$\frac{1}{n!} = \frac{1}{n} \times \frac{1}{n-1} \times \cdots \times \frac{1}{2} \times \frac{1}{1}$$

à chacune de ces permutations.

Ce nombre est simple... $n!$, c'est aussi le nombre d'éléments dans $E_n \times E_{n-1} \times \cdots \times E_2 \times E_1$, pour, $E_n = \{1, \cdots, n\}$, $E_{n-1} = \{1, \cdots, n-1\}$,...$E_2 = \{1, 2\}$, $E = \{1\}$. Donc, il y a une bijection entre $\mathcal{S}_n$ et $E_n \times E_{n-1} \times \cdots \times E_2 \times E_1$ qui associe à toute bijection σ une liste de n nombres $(u_n, \cdots, u_1)$ avec $1 \leq u_i \leq i$ pour tout i (et réciproquement).

Pour engendrer une permutation uniforme, il suffit donc de savoir construire une application Φ de $E_n \times E_{n-1} \times \cdots \times E_2 \times E_1$ vers $\mathcal{S}_n$ qui soit bijective.

– On tire alors des variables U_i indépendantes, pour i allant de n à 1, avec U_i uniforme dans $E_i = \{1, \cdots, i\}$. Ainsi, la proba. de tomber sur un élément quelconque $(u_n, \cdots, u_1)$ de $E_n \times E_{n-1} \times \cdots \times E_2 \times E_1$ est $1/n!$, et puisque Φ est bijective, $\Phi(U_n, \cdots, U_i)$ est une permutation uniforme.

Exercice 2. Trouver une bijection entre $E_n \times E_{n-1} \times \cdots \times E_2 \times E_1$ et $\mathcal{S}_n$. [Il en existe des dizaines !]

3.4 Arbres binaires de recherche

Tout d'abord, parlons un peu de l'algorithme de tri célèbre nommé Quicksort (algorithme de tri rapide). L'algorithme fonctionne comme suit: supposons qu'on ait à trier $x_1, \cdots, x_n$ des éléments de $\mathbb{R}$, ou d'un ensemble possédant une relation d'ordre total $<$.

Il s'agit d'un algorithme récursif: on définit $Q(x_1, \cdots, x_n)$ comme suit:
– si $n \leq 1$, la suite est triée: on renvoie la liste elle même ($Q(x_1) = x_1$ si $n = 1$ et $Q(\varnothing) = \varnothing$ is $n = 0$).
– si $n > 1$, on compare tous les éléments de $x_2, \cdots, x_n$ à x_1. On fait deux paquets, P_- et P_+ des éléments qui sont inférieurs et supérieurs à x_1 (en gardant leurs ordres relatifs dans la liste de départ) et on renvoie

$$Q(x_1, \cdots, x_n) = Q(P_-), x_1, Q(P+).$$

L'arbre binaire de recherche (ABR) est la structure de donnée associée à cet algorithme; il s'agit d'un arbre binaire incomplet (cela signifie que les noeuds ont ou bien 0 enfant, 2 enfants, ou un enfant qui est soit un fils gauche, soit un fils droit).

Il est construit récursivement comme suit: on part de l'arbre vide. On insère successivement les données $x_1, x_2, \cdots$ et on construit un arbre binaire, comme suit: on compare la donnée x_k à celle stockée à la racine r de l'arbre:

- s'il n'y a pas de racine, on la crée, et on y stocke la valeur x_k,
- Si la racine r existe, elle possède alors une donnée x. Si $x_k < x$, on insère x_k dans le sous arbre droit de l'arbre enraciné en r, si $x_k > x$, on insère x_k dans le sous arbre gauche.

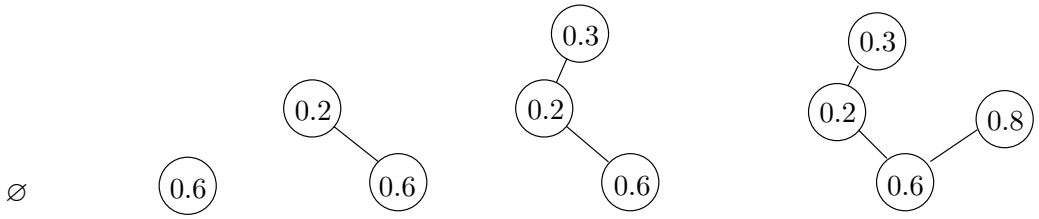


Figure 7: Insertion successive de 0.6, 0.2, 0.3, 0.8.

Exercice 3. Comprendre le lien entre quicksort et l'arbre binaire de recherche.

On voit aisément que ce qui détermine entièrement la géométrie de l'arbre est uniquement l'ordre respectif des données, et non pas leur valeur précise. En d'autres termes, si les données sont $x_1, \cdots, x_k$ ce qui détermine la forme de l'arbre c'est la permutation σ telle que

$$x_{\sigma(1)} < \cdots < x_{\sigma(k)}.$$

Si on choisit une permutation uniforme de n éléments, par exemple de $\{1, \cdots, n\}$ pour construire l'ABR, si on regarde la première donnée, son rang est uniforme dans $\{1, \cdots, n\}$. Cela signifie que le sous arbre droit contient un nombre de noeuds M qui est aléatoire et qui est uniforme entre 0 et $n-1$. Par ailleurs, les données inférieures à la valeur de la racine v sont toutes les valeurs entre 1 et $v-1$ et leur ordre est encore uniforme parmi tous les ordres possibles. Ainsi, le sous arbre droit (ou gauche) conditionnellement à sa taille, a la même distribution qu'un ABR de cette taille: pour être plus juste, il faut considérer que la "forme" de l'arbre, et non pas les étiquettes, qui elles sont différentes dans l'arbre gauche et dans l'arbre droit:

Proposition 3.7. Soit T un ABR sous le modèle de permutation uniforme, à n noeuds, et soit $F(T)$ sa forme (c'est-à-dire, $F(T)$ est obtenu de T en effaçant ses étiquettes). On a :
– le sous arbre gauche T_g de T est de taille U, avec U uniforme entre 0 et $n-1$ (le sous arbre droit T_d de taille $n-1-U$),
– conditionnellement à $U = u$, les formes $F(T_g)$ et $F(T_d)$ sont indépendantes, et ont même loi respectivement que des formes d'ABR de taille u et $n-1-u$.

Remarque 3.8. *On peut utiliser cette propriété pour démontrer que la hauteur des arbres binaires de recherche construits sous n données (sous le modèle de permutation aléatoire) est de hauteur de l'ordre de* $4.311... \log(n)$ *(et une toute petite variance). Ce type de résultat est très difficile à prouver.*

4 ——— Théorèmes limites en probabilité ———

Table des matières

Les théorèmes limites en probabilité ont pour but de décrire le comportement asymptotique d'une suite de variables aléatoires $(X_n, n \geq 1)$, ou alors, le comportement asymptotique de la loi de X_n (par exemple, de sa fonction de répartition).

Il est maintenant bien connu que lorsque l'on reproduit une expérience aléatoire, comme par exemple, lancer une pièce, alors, un phénomène de régularisation apparaît: si on note p_n la proportion de pile à l'instant n, alors la suite aléatoire p_n converge... Ce phénomène est si bien intégré, que l'on pense souvent que la définition même de la probabilité d'avoir pile est la limite en question. On a déjà discuté de la difficulté de définir la probabilité en passant par là, nous n'y reviendrons pas. Mais, ces phénomènes de régularisation sont nettement plus généraux que cela, et concernent, pour ainsi dire tous les phénomènes aléatoires:
– par exemple le nombre de naissances en France, par année, de 2010 à 2014: 830 000 puis 823 400, 821 000, 811 500, puis 818 600. C'est très régulier, non ? Et de même, le nombre de bacheliers fluctue peu, le nombre d'accidents domestiques ou de la route, etc.
– si on vide du sable sur le sol, les milliers de grain qui tombent aléatoirement vont former un joli tas régulier... et d'ailleurs, c'est tellement régulier, qu'on a inventé des sabliers pour mesurer le temps,
– les phénomènes de désintégration nucléaire de l'uranium, totalement aléatoires, le sont globalement nettement moins quand on en amasse des centaines de kilos... On peut alors prédire les phénomènes, les contrôler (le plus souvent...), et fabriquer des centrales nucléaires.

Lorsqu'il a été remarqué au 18ème siècle que la mortalité annuelle fluctuait peu, et que plus précisément, la mortalité par âge fluctuait peu, des produits financiers ont été créés: les assurances décès. C'est semble-t-il, l'une des premières études statistiques publiées, qui a fait grand bruit à l'époque, car cette régularité n'avait pas été perçue auparavant, et a donc beaucoup surpris.

Les modèles évoqués plus haut sont des phénomènes complexes, et bien sûr, on ne va pas les étudier ici. Par contre, ce qu'on va voir rapidement dans cette section c'est, dans des cas plus simples, ce que sont ces phénomènes de régularisation, appelés, "théorèmes limites" en

théorie des probabilités. Il y en a de deux types[4], et c'est important, conceptuellement, de les différentier.

• Le premier type de convergence, est la convergence en probabilité: il s'agit de la convergence d'une suite aléatoire X_n vers une limite, qui peut être aléatoire. La définition est donnée plus bas,

• Le deuxième type de convergence est la convergence en loi; on devrait dire, convergence de la loi, pour éviter certaines confusions. L'idée est différente: on dit qu'une suite X_n converge en loi, lorsque sa loi, par exemple, sa fonction de répartition converge vers une limite (dans un certain sens). Ainsi, ça ne veut pas dire que la suite (X_n) se régularise, mais que la suite de fonction de répartitions associées (F_n) se régularise.

4.1 Convergence en probabilité

Définition 4.1. Soit $X, X_1, X_2, \cdots$ une suite de variables aléatoire définies sur un espace de probabilité. On dit que la suite de v.a.r. (X_n) converge en probabilité vers la variable aléatoire X, si pour tout $\varepsilon > 0$

$$\mathbb{P}(\mid X_n - X \mid \geq \varepsilon) \underset{n \longrightarrow +\infty}{\longrightarrow} 0.$$

On note $X_n \overset{proba}{\longrightarrow} X$.

Exemple 4.28. Prenons une suite $(B_n, n \geq 1)$ de v.a. indépendante de Bernoulli $B(p)$ (pour le p de votre choix). On construit pour tout n, $X_n = \sum_{j=1}^{n} B_j/2^j$, et X la somme $X = \sum_{j \geq 1} B_j/2^j$. Autrement dit, en base 2, X possède des chiffres après la virgule aléatoires. On obtient X_n en ne conservant que les n premiers chiffres. On voit alors que X_n converge vers X puisqu'on a $|X_n - X| \leq 1/2^n$, de sorte que pour $\varepsilon > 0$ fixé, $\mathbb{P}(|X_n - X| > \varepsilon) = 0$ pour n assez grand. (On peut voir autrement que la suite $(X_n, n \geq 0)$ converge: elle est croissante, bornée par 1, et donc, elle converge). Ainsi la suite (X_n) converge en probabilité, et dans ce cas, la limite est aléatoire... c'est X.

Lemme 4.2. Soit $X, X_1, X_2, \cdots$ une suite de v.a. définies sur le même espace de probabilité telles que,

$$\lim_n \mathbb{E}((X_n - X)^2) = 0$$

alors

$$X_n \overset{proba}{\longrightarrow} X.$$

[4] il y en a en fait davantage, mais on peut toujours les classer dans les 2 familles exposées ici

Preuve. Par Markov, on a:

$$\mathbb{P}(\mid X_n - X \mid \geq \varepsilon) = \mathbb{P}(\mid X_n - X \mid^2 \geq \varepsilon^2) \leq \frac{1}{\varepsilon^2}\mathbb{E}((X_n - X)^2).$$

Donc, puisque par hypothèse le terme de droite tend vers 0, celui de gauche aussi... □

4.2 Loi faible des grands nombres

Théorème 4.3. (loi faible des grands nombres) Soit $(X_n, n \geq 1)$ une suite de v.a. indépendantes et de même loi, et de moyenne m. On a

$$\frac{1}{n}\sum_{k=1}^{n} X_k \underset{n\to+\infty}{\overset{proba}{\longrightarrow}} \mathbb{E}(X_1) = m. \tag{4.1}$$

Exemple 4.29. Par exemple, si les X_j sont des variables aléatoires de Bernoulli indépendantes $B(p)$, alors $\mathbb{E}(X_1) = p$. On obtient

$$\frac{1}{n}\sum_{k=1}^{n} X_k \overset{proba}{\longrightarrow} p.$$

Puisque $\overline{X}_n$ est la proportion de Bernoulli qui valent 1, c'est la loi faible des grands nombres qui affirme que la proportion de *pile* dans une suite de pile ou face tend vers $1/2$. C'est donc la loi des grands nombres qui fait le lien entre la probabilité "abstraite" P d'un événement A que l'on a définie au début du cours, et la proportion empirique asymptotique des événements A obtenus dans une suite d'expériences aléatoires.

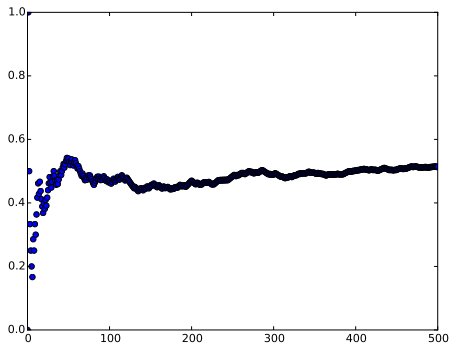


Figure 8: Traçage du graphe $n \mapsto p_n$, avec p_n proportion à l'instant n du nb de piles dans une suite de pile-face aléatoires indépendants, avec proba de pile valant 1/2. On perçoit sur le dessin la convergence annoncée.

Preuve. On donne une preuve sous la condition supplémentaire de l'existence d'une variance finie: On suppose donc que $\mathrm{var}(X_1) = \sigma^2$, et on note $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. On a

$\mathbb{E}(\overline{X}_n) = m$ et $\mathsf{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$. Ainsi

$$\mathbb{E}((\overline{X}_n - M)^2) = \mathsf{Var}(\overline{X}_n) \underset{n\longrightarrow+\infty}{\longrightarrow} 0.$$

Ce qui montre que $\overline{X}_n \overset{proba}{\longrightarrow} m$ d'après le lemme précédent. □

4.3 Convergence en loi

Définition 4.4. Soit $(X_n)_n$ une suite de v.a.r. de fonctions de répartition respectives F_n. On dit que la suite $(X_n)_n$ converge en loi vers la v.a.r. X (de fonction de répartition F) si:

$$\lim_n F_n(a) = F(a)$$

pour tout a, point de continuité de F (i.e. partout si F est la fonction de répartition d'une v.a. à densité).

Il s'agit donc de la convergence simple de la suite de fonctions de répartition F_n (sauf au plus sur l'ensemble des points de discontinuité de F).
Une conséquence de cette définition est la suivante:

Théorème 4.5. Si $(X_n)_n$ est une suite de variables discrètes, prenant ses valeurs sur $\mathbb{N}$,

$$X_n \overset{loi}{\longrightarrow} X$$

si pour tout $k \in \mathbb{N}$,

$$\mathbb{P}(X_n = k) \longrightarrow \mathbb{P}(X = k).$$

Proposition 4.6. (Convergence de la binomiale vers la loi de Poisson)
Soit X_n une suite de v.a. de loi binomiale de paramètre $\mathcal{B}(n, \lambda/n)$ et X une v.a. de Poisson de paramètre λ. On a:

$$X_n \overset{loi}{\longrightarrow} X.$$

Preuve. Soit k fixé.

$$\begin{aligned}\mathbb{P}(X_n = k) &= = \frac{n!}{(n-k)!k!}\Big(\frac{\lambda}{n}\Big)^k\Big(1-\frac{\lambda}{n}\Big)^{n-k} \\ &= \frac{\lambda^k}{k!}\frac{n!}{(n-k)!n^k}\Big(1-\frac{\lambda}{n}\Big)^{n-k}\end{aligned}$$

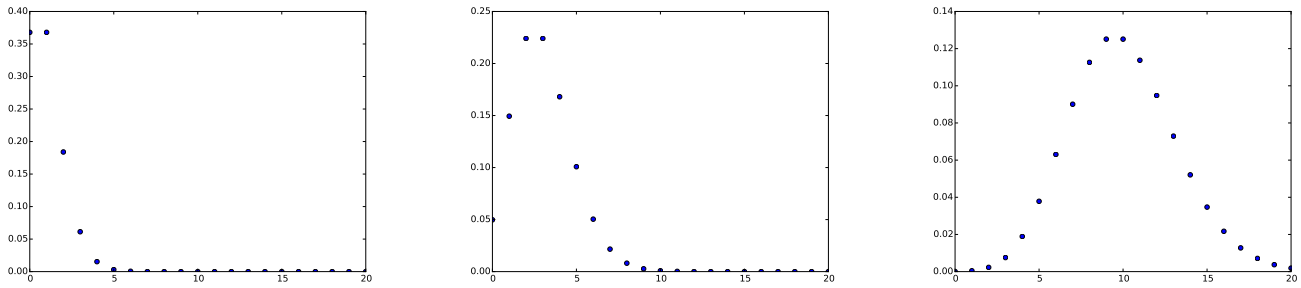


Figure 9: Distribution de Poisson de paramètres 1, 3 et 10

Le second terme du produit tend vers 1. Pour le troisième on fait un dl. On obtient:

$$\left(1-\frac{\lambda}{n}\right)^{n-k} = e^{(n-k)\ln(1-\frac{\lambda}{n})} = e^{(n-k)(-\frac{\lambda}{n}-o(1/n))} = e^{-\lambda+o(1)}$$

d'où

$$\mathbb{P}(X_n = k) \longrightarrow \frac{e^{-\lambda}\lambda^k}{k!}$$

□

En pratique, lorsque que n est grand, on "approxime" $\mathbb{P}(X_n = k)$ par $\mathbb{P}(X = k)$.

Théorème 4.7. (Théorème de la limite centrale) Soit (X_n) une suite de v.a.r. indépendantes et de même loi, d'espérance finie m et de variance finie et non nulle σ^2. Notons $S_n = X_1 + \cdots + X_n$ la somme des n premières variables. On a, pour tout $x < y$ (finis ou pas):

$$\mathbb{P}\left(x \leq \frac{S_n - nm}{\sigma\sqrt{n}} \leq y\right) \underset{n\to+\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_x^y \exp(-t^2/2)dt. \qquad (4.2)$$

On dit aussi, théorème central limite, TCL, ou TLC. La fonction $t \mapsto \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$ est appelée, courbe en cloche, courbe de Gauss, densité normale, ou densité Gaussienne.
En pratique, "lorsque n est grand", on approxime le membre de gauche dans (4.2) par celui de droite.

Exemple 4.30. Si on prend des B_i Bernoulli$(1/2)$ indépendantes, alors pour $S_n = B_1 + \cdots + B_n$, le TCL dit que (pour $p = 1/2$ ici)

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq y\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y \exp(-t^2/2)dt.$$

Si on trace le graphe $n \mapsto \frac{S_n - np}{\sqrt{np(1-p)}}$, alors on ne perçoit pas de régularisation: il n'y en a pas. C'est la fonction de répartition de $\frac{S_n - np}{\sqrt{np(1-p)}}$ qui converge; pour le voir, il faut calculer la loi de cette variable... éventuellement, on peut le voir en faisant des milliers de simulations de S_n pour approcher sa fonction de répartition empirique, mais en tout cas, pas avec une seule.

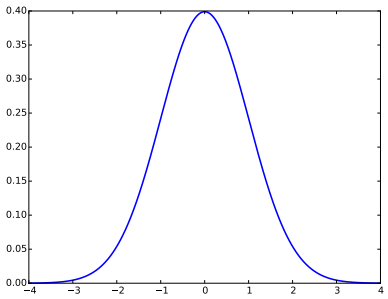


Figure 10: Fonction $x \mapsto \exp(-x^2/2)$ (bien sûr elle ne fait pas 0 à partir de 4...)

.

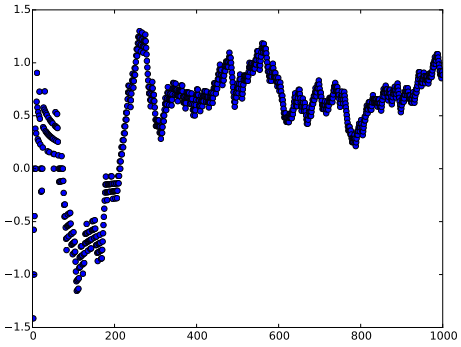


Figure 11: Représentation de $n \mapsto \frac{S_n - np}{\sqrt{np(1-p)}}$

Exemple 4.31. John joue chaque semaine à un jeu, où il mise 10 euros. Il peut gagner 0 avec proba 0.99, 10 euros, avec proba 0.005, 100 euros avec proba 0.0025 et 2000 euros avec proba 0.0025. Il joue cela pendant 20 ans, toutes les semaines: finalement, il joue 1000 fois. Quel est le bilan financer typique de cet acharnement ?

Bon, alors, on ne peut pas savoir combien John a gagné ou pas, mais on peut calculer la loi approximative de son gain grâce au TCL. On calcule la moyenne et la variance du gain:

$$\mathbb{E}(G') = 0 \times 0.99 + 10 \times 0.995 + 100 \times 0.0025 + 2000 \times 0.0025 = 5.3$$

mais son vrai gain moyen est $\mathbb{E}(G) = 5.3 - 10 = -4.7$ car, il ne faut pas oublier dans le bilan financier, sa mise. Donc, en fait son gain est $G = G' - 10$; la variance de G' et de G sont égales: on calcule

$$\mathbb{E}(G'^2) = 0 \times 0.99 + 10^2 \times 0.995 + 100^2 \times 0.0025 + 2000^2 \times 0.0025 = 10025.5$$

donc, $\mathrm{var}(G') = 100025.5 - 5.3^2 = 9997.41$. Ces données sont typiques des jeux de grattage de la française des jeux: en moyenne la moitié de la mise est perdue, et on a une distribution des gains du type de ceux évoqués. Le théorème de la limite centrale, dit que si on note B_n

le bilan financier après n semaines, alors

$$\mathbb{P}\left(\frac{B_n + 4.7n}{\sqrt{n\,\mathrm{var}(G)}} \leq x\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)dt \tag{4.3}$$

le membre de gauche coïncide avec $\mathbb{P}\left(B_n \leq x\sqrt{n\,\mathrm{var}(G)} - 4.7n\right)$. Comment comprendre ce résultat ? Tout d'abord, la première formule dit que B_n se situe aux alentours de $-4.7n$ avec des fluctuations typiques de $\sqrt{n\,\mathrm{var}\,(G)} \sim 100\sqrt{n}$. Pour $n = 1000$, $\sqrt{n} = 31.62...$, les fluctuations typiques sont donc, de l'ordre de 3100, et comme la moyenne de la perte est aux alentours de -4700, ces quantités sont du même ordre (ce ne serait plus vrai si on prenait $n = 100000$, par exemple, car la moyenne évolue linéairement avec n, alors que $\sqrt{n\,\mathrm{var}(G)}$ évolue comme la fonction $x \mapsto \sqrt{x}$ qui tend vers l'infini moins vite). Si on reprend la formule (4.3), avec $n = 1000$, on obtient

$$\mathbb{P}\left(\frac{B_n + 4700}{3162} \leq x\right) \sim \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)dt \tag{4.4}$$

Quelle est la proba qu'il soit dans les négatifs ? $\frac{B_n+4700}{3162} \leq x \Leftrightarrow B_n \leq 3162x - 4700$. C'est en prenant $x = 4700/3162 = 1.486...$ qu'on peut évaluer $\mathbb{P}(B_n \leq 0)$, car on vient de voir que

$$\mathbb{P}\left(\frac{B_n + 4700}{3162} \leq 1.486\right) = \mathbb{P}(B_n \leq 0).$$

Le résultat est donc par le TCL proche[5] de $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1.486} \exp(-t^2/2)dt = 0.9313$ (calcul par ordi). Bref, John a perdu de l'argent avec proba 0.9313. Puisque la limite est symétrique autour de 0, cela dit que l'on sait que John a grossièrement une chance sur 2 d'avoir perdu davantage que 4700 euros, et une chance sur 2 d'avoir perdu moins que cela.... La densité gaussienne permet d'aller plus loin dans l'analyse, en prenant le x de notre choix.

Remarque 4.8. **Pourquoi le TCL est-il un théorème important dans toutes les sciences ?**

La plupart des sciences "ont une connaissance du monde de nature statistique": elles observent d'abord la nature, font des mesures de telles ou telles quantités, et en déduisent des principes, des lois. Or, qui dit "mesure", dit, "erreur de mesure". Pour avancer, on doit donc quantifier quelle erreur on fait en faisant la moyenne des mesures obtenues par exemple. Eh bien, le TCL énonce que sous des hypothèses assez faibles (existence d'une variance, expériences indépendantes), la fonction de répartition de la variable $(S_n - nm)/\sqrt{n\,\mathrm{var}(X_1)}$ *converge. Or, la moyenne empirique dont dispose l'expérimentateur c'est* S_n/n *et la quantité recherchée est* m*. On écrit*

$$\frac{S_n - nm}{\sqrt{n\,\mathrm{var}(X_1)}} = \left(\frac{S_n}{n} - m\right) \frac{\sqrt{n}}{\sqrt{\mathrm{var}(X_1)}},$$

[5]En fait, en toute rigueur, il faut déjà démontrer que l'approximation de $\mathbb{P}\left(\frac{B_n+4700}{3162} \leq x\right)$ par sa limite est valide, ce qu'on admet

puis donc

$$\mathbb{P}\left(x \leq \frac{S_n - nm}{\sqrt{n\,\mathrm{var}(X_1)}} \leq y\right) = \mathbb{P}\left(x\frac{\sqrt{\mathrm{var}(X_1)}}{\sqrt{n}} \leq \frac{S_n}{n} - m \leq y\frac{\sqrt{\mathrm{var}(X_1)}}{\sqrt{n}}\right);$$

comme ceci "converge" pour x et y fixés dans le sens dit plus haut, ça dit que $\frac{S_n}{n} - m$ est de l'ordre de $1/\sqrt{n}$: on connait donc la vitesse de convergence de la moyenne empirique, c'est $1/\sqrt{n}$ quelque soit l'expérience !! et donc, la moyenne inconnue m vaut S_n/n plus un terme d'erreur qui est de l'ordre de $1/\sqrt{n}$. Par dessus le marché, la loi de proba de l'erreur faite converge et ne dépend pas de la loi des expériences, mais seulement de la variance de celles-ci (qui peut d'ailleurs être également évaluée). Ce résultat est de nature très inhabituelle, car il révèle un principe d'invariance très général.

Si on note $F(t) = P(X \leq t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} \exp\left(-x^2/2\right) dx$, on voit alors, en utilisant les symétries de la densité gaussienne que que $F(-t) = 1 - F(t)$. Il se trouve que $\exp\left(-x^2/2\right)$ ne possède pas de primitive qui "s'exprime simplement à l'aide de fonctions classiques" et donc, il n'existe pas de formule plus simple pour $F(t)$ que l'écriture de cette intégrale. Par contre, on peut calculer avec des méthodes d'analyse numérique ces intégrales pour un t donné. C'est crucial de pouvoir faire cela pour les applications statistiques; on fabrique alors des tableaux, comme le suivant:

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

La case rouge du tableau: doit être comprise comme suit: $\mathbb{P}(X \leq 1.35) = 0.9115$, et oui, il s'agit bien d'une approximation à 10^{-4} près de la vérité; ce n'est pas vraiment égal donc, mais on fait "comme si", dans les applications.

4.4 Complément sur les lois continues (non exigible)

On a vu jusque maintenant des variables aléatoires discrètes sur $\mathbb{R}$. La loi de telles variables aléatoires X est donnée par la probabilité que $\mathbb{P}(X = x)$ pour tout x dans $\mathbb{R}$, un nombre au plus dénombrable de tels x possédant une probabilité strictement positive. On a vu également que la fonction de répartition $x \mapsto F(x) = \mathbb{P}(X \leq x)$ caractérisait la loi de X.

Si on réfléchit en terme de fonction de répartition, on voit que cette fonction de répartition permet de calculer $\mathbb{P}(X \in]a, b]) = F(b) - F(a)$ pour tout intervalle $]a, b]$.

Maintenant, prenons une fonction F quelconque, croissante, valant 0 en $-\infty$ et 1 en $+\infty$. Supposons aussi que F soit continue à droite (c'est juste pour tenir compte du fait que lorsque F présente un saut, c'est qu'il y a un atome en x). Avec cette fonction F, on peut définir une loi de proba sur $\mathbb{R}$, en décrétant, que $\mathbb{P}(X \leq x) = F(x)$.

C'est plus général que le cas discret, car F peut très bien ne pas présenter de sauts. Toute mesure de probabilité sur $\mathbb{R}$ peut-être caractérisée de cette manière: donc, on se donne une fonction de répartition d'abord, et on s'en sert pour construire une mesure de probabilité sur $\mathbb{R}$, qui va non plus déterminer la probabilité de valoir une valeur, mais une probabilité de tomber dans un intervalle.

Il est facile de voir que si on se donne une fonction f positive ou nulle, d'intégrale 1, alors si on pose

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

alors F possède toutes les caractéristiques d'une fonction de répartition. Nous allons expliquer pourquoi cela détermine entièrement la loi d'une variable aléatoire X dont la fonction de répartition est F. La fonction f sera appelée densité de la loi de X, ou plus simplement densité de X.

Définition 4.9. Une fonction f continue par morceaux,

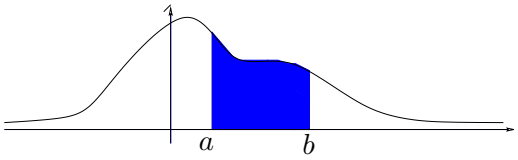
$$\begin{array}{rcl} f : \mathbb{R} & \longrightarrow & \mathbb{R}^+ \\ x & \longmapsto & f(x) \end{array}$$

est une densité de probabilité si $\int_{\mathbb{R}} f(x)dx = 1$ (on dit aussi, directement qu'une loi $\mathbb{P}_X$ possède une densité si pour tout intervalle A, $\mathbb{P}_X(A) = \int_A f(x)dx$).

Remarque 4.10. *Toutes les distributions sur $\mathbb{R}$ ne possèdent pas forcément ou bien une densité, ou bien une loi discrète, mais cela nous ferait sortir du cadre de ce cours de décrire des exemples différents ici. Par ailleurs, la condition "continue par morceaux" n'est pas non plus nécessaire, mais donner un sens aux densités qui ne sont pas de ce type sort également au cadre de ce cours.*

Définition 4.11. On dit qu'une v.a. X a pour densité f si:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx \text{ pour tout } -\infty \leq a \leq b \leq +\infty.$$



Ainsi la densité f détermine entièrement la loi de X. De plus, deux densités f_1 et f_2 déterminant la même loi sont égales ("presque partout").

Proposition 4.12. La fonction de répartition F de la v.a. X et sa densité sont liées par la formule

$$F'(x) = f(x).$$

Preuve. (pour être précis, ce résultat n'est vrai que là où f est continue.)

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)du$$

d'où le résultat. □

Remarque 4.13.
- *Lorsqu'une variable aléatoire admet une densité alors comme l'intégrale $\int_a^a f(x)dx = 0$, on voit que $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X < b)$ et donc, la probabilité que X vaill précisément un certain a est 0 pour tout a. Ainsi, cette fois, la loi n'est pas caractérisée par la probabilité de valoir quelque a que ce soit, contrairement au cas discret.*
- *On sait calculer la probabilité d'un intervalle. On voit bien qu'on peut calculer aussi la probabilité de l'union de deux intervalles, ou de l'union d'un nombre quelconque d'entre eux. Par contre, il y a une impossibilité générale d'affecter une probabilité à tous sous ensembles de $\mathbb{R}$. C'est pour cela que dans les cours avancés de probabilité, on fait appel à la notion de tribu et de Borélien pour désigner les ensembles pour lesquels on peut définir la probabilité. Cela introduit une complication technique importante, même si cela influe peu sur ce qu'on fait in fine. Ici, on va se contenter de calculer des probabilités d'intervalles, pour lesquels, aucun problème n'apparaît.*

Espérance et Variance

Les sommes du cas discret sont remplacées par des intégrales:

$$\begin{aligned} \mathbb{E}(X) &= \int_{\mathbb{R}} uf(u)du \\ \mathbb{E}(X^2) &= \int_{\mathbb{R}} u^2 f(u)du \\ \mathbb{E}(g(X)) &= \int_{\mathbb{R}} g(u)f(u)du \end{aligned}$$

Toutes les formules et propriétés du premier chapitre s'étendent ici.

4.5 Exemple de la loi normale

Définition 4.14. Soit $m \in \mathbb{R}$ et $\sigma^2 > 0$. On dit que X suit une loi normale de paramètre m et σ^2 si sa densité de probabilité est:

$$\begin{aligned} f : \mathbb{R} &\longrightarrow \mathbb{R}^+ \\ x &\longmapsto \frac{e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}}{\sqrt{2\pi\sigma^2}}. \end{aligned}$$

On note $X \sim \mathcal{N}(0,1)$.

Ainsi définie f est bien une densité puisque elle est positive, continue et que son intégrale est 1. La preuve de ce dernier point est difficile à trouver seul(e)! Tout d'abord, on fait un changement de variable, et on pose $t = (x-m)/\sigma$, on trouve $\int_{\mathbb{R}} f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-t^2/2}dt$. Ensuite, on calcule le carré de l'intégrale, et on utilise un changement en

coordonnée polaire :

$$\begin{aligned}\left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-t^2}dt\right)^2 &= \left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-x^2}dx\right)\left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-y^2}dy\right) = \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-(x^2+y^2)/2}dx\,dy \\ &= \frac{1}{2\pi}\int_0^{+\infty}\int_0^{2\pi}\rho e^{-\rho^2/2}d\theta\,d\rho = \int_{\rho\in\mathbb{R}^+}\rho e^{-\rho^2/2}d\rho = \left[-\exp(-\rho^2/2)\right]_0^{+\infty} = 1\end{aligned}$$

Proposition 4.15. Soit $X \sim \mathcal{N}(m,\sigma^2)$ et $Y \sim \mathcal{N}(0,1)$:
i) $\mathbb{E}(X) = m,\ \mathsf{Var}(X) = \sigma^2$
ii) $X \overset{loi}{=} \sigma Y + m$

Preuve:

$$i)\ \ \mathbb{E}(X) = \int_{\mathbb{R}} x\frac{e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}}{\sqrt{2\pi\sigma^2}}dx = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}(\sigma t+m)e^{-t^2/2}dt = m$$

par le changement de variable $t = (x-m)/\sigma$.

$$\begin{aligned}\mathsf{Var}(X) &= \int_{\mathbb{R}}(x-m)^2\frac{e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}}{\sqrt{2\pi\sigma^2}}dx \\ &= \frac{\sigma^2}{\sqrt{2\pi}}\int_{\mathbb{R}} t^2e^{-t^2/2}dt \\ &= \frac{\sigma^2}{\sqrt{2\pi}}\left(\Big[-e^{-t^2/2}t\Big]_{-\infty}^{+\infty} + \int_{\mathbb{R}} e^{-t^2/2}dt\right) \\ &= \sigma^2\end{aligned}$$

ii) Pour prouver ce deuxième point on montre que les fonctions de répartitions de deux variables X et $\sigma Y + m$ sont égales. Or, on a vu plus haut que ces f.r. déterminaient la loi.

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x}\frac{e^{-\frac{1}{2}\left(\frac{u-m}{\sigma}\right)^2}}{\sqrt{2\pi\sigma^2}}du.$$

$$\begin{aligned}F_{\sigma Y+m}(x) &= \mathbb{P}(\sigma Y + m \le x) = \mathbb{P}(Y \le \frac{x-m}{\sigma}) \\ &= \int_{-\infty}^{\frac{x-m}{\sigma}}\frac{e^{-t^2/2}}{\sqrt{2\pi}}dt\end{aligned}$$

Le changement de variable $t = \frac{u-m}{\sigma}$ dans cette dernière intégrale donne le résultat escompté.

Corollaire 4.16. Si $X \sim \mathcal{N}(m,\sigma^2)$ alors:

$$\frac{X-m}{\sigma} \sim \mathcal{N}(0,1).$$

Ainsi tout calcul impliquant des normales se ramène à la loi normale centrée réduite.
La fonction de répartition de la loi normale ne possède pas de formule plus simple que

$$F(x) = \int_{-\infty}^{x}\frac{e^{-t^2/2}}{\sqrt{2\pi}}dt.$$

On est donc "réduit" à utiliser des tables pour connaître F.

Le théorème central limite reformulé...

En fait, dans le théorème de la limite centrale, la limite fait appel à la densité de la gaussienne réduite centré. On exprime donc en général le théorème de la limite centrale comme suit:

Théorème 4.17. (Théorème de la limite centrale) Soit (X_n) une suite de v.a.r. indépendantes et de même loi, d'espérance finie m et de variance finie et non nulle σ^2. Notons $S_n = X_1 + \cdots + X_n$ la somme des n premières variables. On a, pour tout $x < y$:

$$\mathbb{P}\left(x \leq \frac{S_n - nm}{\sigma\sqrt{n}} \leq y\right) \longrightarrow \mathbb{P}(x \leq N \leq y)$$

où N est une variable aléatoire de loi gaussienne réduite centrée $(\mathcal{N}(0,1))$.

D'autres lois continues

Il existe bien sûr autre chose que la loi gaussienne, dans la vie, et on peut citer:

(i) la loi uniforme sur $[0,1]$ dont la densité est simplement $f(x) = 1_{[0,1]}(x)$, c'est à dire, la fonction valant 1 pour $x \in [0,1]$ et 0 sinon. On a alors pour $0 \leq a \leq b \leq 1$, pour U une v.a. suivant cette loi

$$\mathbb{P}(a \leq U \leq b) = \int_a^b 1dx = b - a.$$

C'est une variable de cette loi que les générateurs aléatoires des langages de programmation tentent de reproduire.

(ii) La loi exponentielle: il s'agit de la loi de densité $f(x) = e^{-x}1_{[0,+\infty[}(x)$. C'est en quelque sorte, la version continue de la loi géométrique.

5 ——— Quelques éléments de statistiques ———

5.1 Introduction

Dans la théorie des probabilités, le point de départ suppose la pré-connaissance du modèle: on a calculé des lois, des espérances, on a donné des théorèmes limites, et précédemment, à chaque fois, on a supposé connues les lois des variables aléatoires en présence. Dans la $vraie$ vie le problème est souvent inverse: on observe un phénomène aléatoire et on essaie d'établir ses caractéristiques, sa loi.

On a vu dans ce cours que de nombreux phénomènes aléatoires possédaient des phénomènes de régularisation asymptotique. Par exemple, la loi faible des grands nombres montre que la moyenne empirique converge en probabilité vers l'espérance mathématique; si la variance existe, le théorème de la limite centrale donne même une "vitesse de convergence". Il est donc vrai que l'on va pouvoir apprendre des choses grâces aux informations observées.

5.2 Quelques réflexions liminaires

D'abord, même dans le cas "d'une expérience parfaite", on ne peut être sûr à 100% de rien: si on a lancé une pièce 1000 fois et observé 607 fois $piles$, on ne peut pas être sûr que la probabilité p que la pièce tombe sur pile est à peu près 0,6... Quelle que soit la probabilité $p \in]0,1[$ on peut tout à fait observer 607 piles, même si cela est plus plausible pour $p = 0,607$ que pour $p = 0,0001$. Les statistiques ont pour objet de définir des concepts et des outils mathématiques permettant de mieux appréhender ces questions.

Avant d'entrer dans le cœur du sujet, il est utile de discuter d'abord un peu l'utilisation des statistiques dans la société et dans les médias, car savoir calculer sans bien réfléchir à ce qu'on calcule est inutile.

Il est naturel dans presque toutes les situations réelles de faire une étude statistique avant de prendre des décisions d'ampleur. Par exemple, avant de créer une entreprise, on doit évaluer combien on aura de clients : on fait un sondage. L'état pour dimensionner ses politiques publiques a besoin de connaître ses citoyens, leur âge, leur revenus, leurs besoins et envies: on fait un recensement.

5.2.1 La collecte des données

On doit commencer par recueillir des données. On a plusieurs façons de faire.

(a) On peut essayer de récolter de manière exhaustive "toutes les données". C'est ce que fait l'état lors du recensement, et plus généralement, l'INSEE étudie un nombre impressionnant de données concernant tous les secteurs d'activité (allez voir leur site web, c'est impressionnant !).

(b) On peut procéder par sondage, tirage au sort; on choisit 25 pièces qui viennent de sortir de la machine, et on regarde combien sont défectueuses, on appelle 1000 citoyens au téléphone, et on leur demande pour qui ils votent, ou s'ils consomment telle ou telle marque de cassoulet.

5.2.2 La qualité des données

Une fois que les données sont recueillies, on peut s'interroger sur leur qualité: par exemple,

(a) pour le recensement: Il y a des gens qui refusent de répondre, mentent, ou se cachent, qui déménagent juste avant ou juste après le recensement, certains sont comptés 2 fois, certains meurent juste après, etc. On sait qu'on n'aura pas un résultat exact, mais pour ce qu'on en fait, c'est suffisant.

(b) Les gens victimes d'agression physique portent peu plainte et c'est par des sondages géants que l'INSEE parvient à mesurer les taux d'agressions. Les données exhaustives de la police/gendarmerie ne portant que sur les plaines déposées ne sont pas fiables du tout pour connaître les chiffres de l'insécurité. Les sondages sont nettement plus précis !

(c) Le cas des sondages électoraux est assez intéressant. On choisit "au hasard" des gens dans la population et on leur demande ce qu'ils vont voter. Dans un modèle idéal, les gens sondés seraient vraiment tirés au sort selon la loi uniforme, serait obligés de répondre, aurait déjà réfléchi pour qui ils vont voter (ou auraient déjà décidé de s'abstenir), et diraient la vérité au sondeur. Il se trouve qu'aucune de ces conditions n'a lieu: on ne peut pas tirer au hasard des individus dans la population, car personne ne dispose d'une telle liste (en tout cas, pas les sondeurs). Les gens sollicités au téléphone ou dans la rue refusent de répondre pour une bonne partie (et ce refus n'est pas indépendant de leur vote). Il s'agit d'un biais dit "de sélection". Pire encore, ils mentent aux sondeurs et une fois encore, ce n'est pas indépendant du vote, car on ment davantage si on n'est pas fier de notre vote.

Pour établir "un sondage valable", les sondeurs utilisent des recettes qui consistent à changer le poids des données recueillies:
– en tenant compte du poids relatif dans la population de la catégorie socio-professionnelle des sondés. Par exemple, si on sait qu'il y a 13% d'ouvriers dans la population, mais que dans notre sondage seuls, 6.5% des sondés étaient des ouvriers, on compte 2 fois chaque vote ouvrier recueilli,
– on essaie de corriger le taux "d'erreur par parti" (dûs aux mensonges, en particulier), en comparant les sondages passés aux résultats passés; par exemple, si lors de la dernière élection, lors d'un sondage on trouvait 10% des votants pour A, mais que 20% des

électeurs votaient finalement pour A, alors on sait, qu'en gros il faut multiplier par 2 les résultats du sondage en ce qui concerne le candidat A.

Ces recettes marchent mal et créent des biais qu'il est impossible à évaluer. Lors des primaires du parti LR, d'énormes différences entre les sondages et les résultats ont été observés (plus de 10%). Les sondeurs ont justifié leur échec par la dynamique de campagne, par les opinions volatiles... Mais des explications beaucoup plus terre à terre expliquent cette échec, particulièrement pour les primaires:
– les sondeurs n'ont pas vus que les retraités et les gens de "la manif pour tous" seraient sur-représentés parmi les votants. Ils n'ont donc pas pu utiliser "la bonne recette".
– Par ailleurs, les sociologues ont montré que dans tous les sondages, un certain pourcentage des sondés répondent n'importe quoi: soit il ne comprennent pas la question, soit ils répondent ce qui leur semble être la réponse la plus consensuelle, alors qu'en fait ils n'ont pas d'avis. Par exemple, ils vont dire au sondeur qu'ils vont aller voter aux primaires alors qu'il n'en est rien, et qu'ils vont voter A car A leur semble plus connu/sage que B, alors qu'en fait, ils ne le feront pas.

Ce biais, qui peut ne pas être important pour des sondages classiques, l'est nettement plus, pour les primaires, car moins de 10% du corps électoral vote aux primaires. On doit donc sonder 10000 personnes pour en trouver 1000 qui déclarent qu'ils vont voter. Seulement, si 2% des 10 000 répondent n'importe quoi, on se retrouve avec 200 réponses sur les 1000 qui ne valent rien... soit 20% de mauvaises données... Dans ce cas de figure, un écart de 10% entre un sondage et le vote, c'est tout à fait normal, et c'est même une performance si on ajoute dans la balance les autres biais évoqués plus haut! C'est un cas où il n'est pas clair qu'on puisse faire des sondages fiables.

5.3 L'utilisation des données

Il est encore bon de réfléchir un peu à l'utilisation des données/statistiques dans les média, car il n'est pas rare du tout que des raisonnements totalement fallacieux reposent sur des statistiques. Voici quelques erreurs typiques, mais il y en a des palanquées!
– L'absence de comparaison: On entend souvent des gens qui déclarent qu'il faut faire très attention en voiture quand on roule près de chez soi, car 85% des accidents de la route ont lieu à moins de 15 km de son propre domicile. Supposons que ces données soient vraies. La déduction elle est mauvaise car... il faudrait savoir quelle proportion p de km sont effectués à moins de 15km du domicile: si $p > 85\%$, alors, il est plus risqué d'être loin... et si c'est $p \leq 85\%$ alors c'est vrai. En l'absence de ce second élément de comparaison, on ne peut pas savoir. De la même façon, savoir que 80% de gens qui ont pris le produit A ont été guéris de leur rhume en moins de 6j, ne permet pas de savoir si le produit est utile... si on ne sait

pas combien ont été guéris parmi ceux qui ne l'ont pas pris. Il faut comprendre que seules des statistiques fiables peuvent trancher l'efficacité de tel médicament, ou de telle médecine parallèle ou non (en particulier, on peut considérer que tout témoignage unique concernant l'efficacité d'un remède n'est d'aucune utilité).

– Les biais de confirmation (très utilisés par les politiques): on ne parle que des stats qui vont dans le sens qui nous arrangent, ou pire, on ne retient que les données qui nous arrangent. C'est aussi la base des superstitions et des théories du complot: on ne retient que les événements marquant qui ont lieu les vendredi 13, et on oublie les vendredi 13 où il ne se passe rien, et les mardi 24 s'il s'y passe quelque chose.

5.4 Estimation ponctuelle

On se trouve dans une situation où des variables aléatoires $(X_1, \cdots, X_n)$ indépendantes et de même loi inconnue $\mathbb{P}_\theta$ ont été tirées, et on a observé l'échantillon $x_1, \cdots, x_n$. Typiquement $\mathbb{P}_\theta$ est une loi à paramètres (comme la loi Bernoulli $B(\theta)$), mais souvent la question est plus générale, et on ne se pose pas cela, mais juste que la loi a une moyenne ou une variance.

Définition 5.1. On appelle n échantillon un n-uplet $(X_1, \cdots, X_n)$ où les v.a. X_i's sont indépendantes et de même loi. Une réalisation $(x_1, \cdots, x_n)$ de l'échantillon est un n-uplet de valeurs prises par l'échantillon (c'est-à-dire, de fait, ce qu'on observe).

Définition 5.2. On appelle estimateur d'un paramètre f d'un modèle, toute variable aléatoire du type $\Phi(X_1, \cdots, X_n)$ pour une fonction $\Phi : \mathbb{R}^n \to \mathbb{R}$ (qui va associer un paramètre à nos variables aléatoires). Une estimation de θ est la valeur prise par $\Phi(x_1, \cdots, x_n)$ où $(x_1, \cdots, x_n)$ est une réalisation.

Le paramètre du modèle peut-être par exemple le paramètre θ de la loi de Bernoulli $B(\theta)$, où de celui de Poisson $\mathcal{P}(\lambda)$ mais ça peut aussi être par exemple la moyenne de la loi, ou sa variance.

Notez que la définition d'un estimateur ne contient aucune notion de qualité; ce sera l'objet des définitions suivantes.

Exemple 5.32. (i) Si on se donne un échantillon $X_1, \cdots, X_n$, alors typiquement il est raisonnable de prendre la moyenne empirique

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

comme estimateur de la moyenne. Mais comme tout est autorisé $\sum_{j=1}^n \sum_{i=1}^n (X_i - X_j)^{17}/n^{28}$ est aussi un estimateur. Il est très mauvais, voilà tout.

(ii) Pour des variables de Bernoulli indépendantes $(B_i, i \geq 1)$ de paramètre θ (inconnu), un estimateur de θ est donné par

$$\theta_n = (\sum_{j=1}^{n} B_j)/n.$$

(iii) Pour des variables $(U_i, i \geq 1)$ indépendantes et uniformes sur $\{0, M\}$, un estimateur de M est donné par exemple par

$$M_n = \max\{U_i, 1 \leq i \leq n\}$$

ou

$$M'_n = 2(\sum_{i=1}^{n} U_i)/n$$

la moyenne de U_1 est $M/2$.

Introduisons maintenant les concepts nécessaires pour "quantifier la qualité d'un estimateur":

Définition 5.3. On dit qu'un estimateur T_n d'un paramètre f est sans biais si $\mathbb{E}(T_n) = f$.
On dit qu'il est asymptotiquement sans biais si $\mathbb{E}(T_n) \to f$.
On dit qu'il est convergent si $T_n \xrightarrow[n]{proba} f$.

Exemple 5.33. Soit $\mathbb{P}$ une loi de proba possédant une moyenne, et $X_1, \cdots, X_n$ des variables aléatoires indépendantes de loi $\mathbb{P}$. Alors la moyenne empirique $\overline{X_n}$ est un estimateur sans biais de la moyenne m de la loi $\mathbb{P}$, et est convergent. En effet, on sait que $\mathbb{E}(\overline{X}_n) = m$, donc l'estimateur est sans biais, et par la loi faible des grands nombres, $\overline{X}_n \xrightarrow[n]{proba} m$, donc il est convergent.

Exemple 5.34. Estimateur de la variance: supposons que $\mathbb{P}$ possède une moyenne m est une variance σ^2. Un estimateur naturel pour la variance est le suivant

$$S_n^{(2)} = \frac{\sum_{j=1}^{n}(X_j - \overline{X}_n)^2}{n}$$

avec $\overline{X_n} = (X_1 + \cdots + X_n)/n$. En effet, il s'agit des écarts quadratiques moyens, à la moyenne empirique. Regardons si $S_n^{(2)}$ est sans biais: On calcule $\mathbb{E}\left(S_n^{(2)}\right)$. Pour cela on voit que les variables $X_i - \overline{X_n}$ ont même loi et donc

$$\begin{aligned}\mathbb{E}(S_n^{(2)}) &= n\mathbb{E}((X_1 - \overline{X}_n)^2)/n = \mathbb{E}\left(((X_1 - m) - (\overline{X_n} - m))^2\right) \\ &= \mathbb{E}\left(\left((X_1 - m)\frac{n-1}{n} + \sum_{k=2}^{n}\frac{X_k - m}{n}\right)^2\right).\end{aligned}$$

Comme les variables

$$Y_1 := (X_1 - m)(n-1)/n, \tag{5.1}$$
$$Y_k := (X_k - m)/n \quad \text{pour} \quad k \geq 2 \tag{5.2}$$

sont de moyenne 0, on voit qu'on est en train de faire un calcul du type $\mathbb{E}((Y_1+Y_2+\cdots+Y_n)^2)$ pour des variables de moyenne 0. Ceci c'est identique à la variance de $Var(Y_1 + \cdots + Y_n)$ (car la moyenne est 0). Comme la variance d'une somme de v.a. indépendantes est la somme des variances, ça donne pour σ^2 la variance de X_1,

$$\mathbb{E}(S_n^{(2)}) = \frac{(n-1)^2}{n^2}\sigma^2 + (n-1)\frac{\sigma^2}{n^2} = \sigma^2\frac{n-1}{n}.$$

Ainsi, $S_n^{(2)}$ n'est pas un estimateur sans biais de la variance ! Par contre

Lemme 5.4.

$$\sigma_n^2 = \frac{n}{n-1}S_n^{(2)} \tag{5.3}$$

est un estimateur sans biais de la variance.

Exemple 5.35. Soit $B_1, \cdots, B_n$ des variables de Bernoulli de paramètre θ. Si on note $\overline{B}_n = (B_1 + \cdots + B_n)/n$, on voit que $\mathbb{E}(\overline{B}_n) = \theta$ et donc $\overline{B}_n$ est un estimateur sans biais de θ. Par la loi faible des grands nombres $\overline{B}_n \xrightarrow[n]{proba} \theta$, et donc $\overline{B}_n$ est un estimateur convergent.

Remarque 5.5. *(Non exigible) Pour comparer 2 estimateurs T_n et T'_n d'un paramètre θ de notre distribution, on compare ce qu'on appelle le risque quadratique moyen qui sont les nombres $\mathbb{E}\left((T_n - \theta)^2\right)$ et $\mathbb{E}\left((T'_n - \theta)^2\right)$ donc, les écarts quadratiques moyens au paramètre. Plus cet écart est faible, meilleur est jugé l'estimateur.*

5.5 Intervalles de confiance

Dans la section précédente, on a parlé d'estimateur θ_n pour un paramètre f. L'estimateur est typiquement aléatoire et donc, lorsqu'on a θ_n, mettons un estimateur convergent, il est raisonnable de penser que θ_n est proche de f, mais vraisemblablement pas égal à f. La proximité de f à θ_n s'exprime en disant typiquement que $|f - \theta_n| \leq a$ pour un certain a. On voit alors que les deux assertions suivantes sont équivalentes:

$$f \in [\theta_n - a, \theta_n + a] \tag{5.4}$$

$$\theta_n \in [f - a, f + a]. \tag{5.5}$$

Cette remarque a l'air stupide, mais elle ne l'est pas, et même elle revêt une importance capitale en statistique, car f, le paramètre du modèle est non aléatoire, alors que θ_n lui l'est. Lorsqu'on se donne le modèle, f est connu, et à l'aide des théorèmes probabilistes que l'on a (en particulier le TCL), on peut souvent évaluer

$$q = \mathbb{P}(\theta_n \in [f - a, f + a]).$$

Grâce à (5.4), on va pouvoir en déduire

$$\mathbb{P}(f \in [\theta_n - a, \theta_n + a]) = q.$$

En d'autres termes, on pourra évaluer la probabilité que le paramètre inconnu f appartienne à un intervalle (qui dépend des observations, en pratique). Si on se donne $q = 0.95$ (ou un autre nombre proche de 1), trouver un a qui convient revient à dire que la probabilité que f soit dans l'intervalle $[\theta_n - a, \theta_n + a]$ est 0.95... On parle alors d'un intervalle de confiance pour un risque de 0.05. Formalisons un peu tout cela.

Définition 5.6. Soit $\alpha \in (0, 1)$ un réél (le niveau de risque). Un intervalle de confiance d'un paramètre f de niveau de confiance $1 - \alpha$ est un intervalle I, dépendant de l'observation mais pas du paramètre, tel que

$$\mathbb{P}(f \in I) \geq 1 - \alpha.$$

Les valeurs usuelles pour α sont 0.01, 0.05, 0.1.

5.6 Intervalle de confiance pour une Bernoulli

Il s'agit ici du cas qui correspond aux sondages: on a une question binaire (oui/non), (0/1). Il existe en gros 2 méthodes pour fabriquer des intervalles de confiance: le théorème de la limite centrale, et l'inégalité de Bienaymé Tchebitchev (d'autres inégalités non traitées dans le cours peuvent également être utilisées). Celles fournies par le théorème de la limite centrale sont nettement meilleures.

On se donne des Bernoulli $(B_i, 1 \leq i \leq 1000)$ indépendantes, de paramètre θ inconnu. Mettons que pour l'application, on a observé un échantillon $b_1, \cdots, b_{1000}$ et que 317 de ces valeurs valaient 1. On cherche un intervalle de confiance pour θ de niveau de risque 0.05.

5.6.1 Par Bienaymé-Tchebichev

On se souvient que $\mathbb{P}(|X - \mathbb{E}(X)| \geq x) \leq \mathrm{Var}(X)/x^2$. Un estimateur convergent pour le paramètre θ est

$$\theta_n = \frac{B_1 + \cdots + B_n}{n}.$$

En fait $B_1 + \cdots + B_n$ est une binomiale $B(n, \theta)$ de moyenne $n\theta$, et de variance $n\theta(1-\theta)$, de sorte que $\mathbb{E}(\theta_n) = \theta$, $\mathrm{Var}(\theta_n) = \theta(1-\theta)/n$. Ici l'échantillon observé fait que l'on dispose de θ_{1000}. Par Bienaymé-Tchebichev on a

$$\mathbb{P}(|\theta_n - \theta| \geq x) \leq \frac{\theta(1-\theta)}{nx^2}$$

ce qu'on peut réécrire

$$1 - \mathbb{P}(|\theta_n - \theta| < x) \leq \frac{\theta(1-\theta)}{nx^2}$$

et donc

$$\mathbb{P}(|\theta_n - \theta| < x) = \mathbb{P}(\theta \in]\theta_n - x, \theta_n + x[) \geq 1 - \frac{\theta(1-\theta)}{nx^2}$$

ça ressemble à ce qu'on cherche, sauf qu'on cherche un intervalle qui ne doit pas dépendre de θ, donc x ne doit pas dépendre de θ, et le niveau $1-\alpha$ ne doit pas dépendre non plus de θ. Bref, il faut se débarrasser du θ dans le membre de droite. On utilise le fait que pour tout $\theta \in (0,1)$, $\theta(1-\theta) \leq 1/4$. On a donc

$$\mathbb{P}(\theta \in]\theta_n - x, \theta_n + x[) \geq 1 - \frac{1/4}{nx^2}.$$

On revient à notre problème: ici $n = 1000$, on a observé $\theta_n = 317/1000$, on cherche un intervalle de confiance de niveau 0.95 (de risque 0.05). On souhaite donc que

$$1 - \frac{1/4}{1000x^2} = 0.95.$$

Ca donne $x = 0.0707$. Autrement dit, on a une confiance à 95% que θ soit dans l'intervalle $[0.317 - 0.07, 0.317 + 0.07]$.

5.6.2 Par le théorème de la limite centrale

On reprend le calcul précédent, mais au lieu d'utiliser l'inégalité donnée par Bienaymé-Tchebitchev, on va utiliser l'approximation donnée par le théorème de la limite centrale.

L'idée est donc de dire que comme $\theta_n - \theta = \frac{\sum_{i=1}^n (B_i - \theta)}{n} = \frac{\sum_{i=1}^n (B_i-\theta)}{\sqrt{n\theta(1-\theta)}} \times \frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}}$ avec bien sûr, le premier terme dans ce produit, est le terme qui est contrôlé par le TCL. On écrit alors

$$\mathbb{P}(|\theta_n - \theta| \leq x) \;=\; \mathbb{P}\left(\left| \frac{\sum_{i=1}^n (B_i - \theta)}{\sqrt{n\theta(1-\theta)}} \right| \leq x \frac{\sqrt{n}}{\sqrt{\theta(1-\theta)}} \right) \tag{5.6}$$

L'idée ici, est de ne pas utiliser la borne $\theta(1-\theta)$ (on pourrait, mais c'est moins bon), mais plutôt d'utiliser l'approximation $\theta(1-\theta) \sim \theta_n(1-\theta_n)$. Et donc on écrit que

$$\mathbb{P}(|\theta_n - \theta| \leq x) \;=\; \mathbb{P}\left(\left| \frac{\sum_{i=1}^n (B_i - \theta)}{\sqrt{n\theta(1-\theta)}} \right| \leq x \frac{\sqrt{n}}{\sqrt{\theta_n(1-\theta_n)}} \right) \sim \int_{-y}^{y} e^{-t^2/2}/\sqrt{2\pi} dt = \Psi(y) \tag{5.7}$$

pour $y = x\frac{\sqrt{n}}{\sqrt{\theta_n(1-\theta_n)}}$. Ici, on a utilisé le TCL. Maintenant, la table donnée page 56, nous dit combien vaut $F(a) = \int_{-\infty}^a e^{-t^2/2}/\sqrt{2\pi} dt$. Comme la fonction $t \mapsto e^{-t^2/2}$ est paire et que $\lim_{a\to+\infty} F(a) = 1$ on voit que pour $R(y) = \int_y^{+\infty} e^{-t^2/2}/\sqrt{2\pi}$ (ce qu'il manque à $F(y)$ pour faire 1), on a: $R(y) = 1 - F(y)$,

$$\Psi(y) = 1 - 2R(y) = 1 - 2(1 - F(y)) = 2F(y) - 1.$$

Si on se donne le niveau de risque α, alors, on cherche x et donc y t.q.

$$\mathbb{P}(|\theta_n - \theta| \leq x) = \Psi(y) = 1 - \alpha$$

ce qui revient à dire que $2F(y) - 1 = 1 - \alpha$,

$$F(y) = 1 - \frac{\alpha}{2}.$$

Ça dit que y doit être le nombre u_α correspond à la case $1 - \alpha/2$, et donc, on doit avoir

$$x\frac{\sqrt{n}}{\sqrt{\theta_n(1-\theta_n)}} = u_\alpha \Leftrightarrow x = u_\alpha \frac{\sqrt{\theta_n(1-\theta_n)}}{\sqrt{n}}$$

ainsi l'intervalle de confiance de niveau de risque α que l'on donne est

$$I_\alpha = \left[\theta_n - u_\alpha \frac{\sqrt{\theta_n(1-\theta_n)}}{\sqrt{n}}, \theta_n + u_\alpha \frac{\sqrt{\theta_n(1-\theta_n)}}{\sqrt{n}} \right]. \tag{5.8}$$

Voyons notre application numérique: $n = 1000$, $\theta_n = 317/1000$, $\alpha = 0.05$. On cherche dans la table la valeur de u_α correspondant à la case $1 - 0.05/2 = 0.975$. On voit que

$$u_\alpha = 1.96.$$

Notre intervalle est donc

$$I_{0,05} = \left[0.317 - 1.96\frac{\sqrt{0.317(1-0.317)}}{\sqrt{1000}}, 0.317 + 1.96\frac{\sqrt{0.317(1-0.317)}}{\sqrt{1000}} \right] = [0.288, 0.345].$$

Typiquement, si on fait un sondage dans une population, est que l'on trouve que 31.7% des électeurs sondés vont voter pour A, le vrai pourcentage se trouve dans $[0.288, 0.345]$ avec proba 0.95. Il faudrait donc que les sondeurs et les médias publient ces intervalles de confiance pour informer correctement les citoyens (enfin, à cause des recettes d'ajustements évoquées plus haut, c'est presque impossible).

Remarque 5.7. *Comme expliqué, on peut retourner l'énoncé: si dans la population générale, il y a 31.7% d'électeurs pour A, alors, 95% des sondages tomberont dans l'intervalle* $[0.288, 0.345]$*... et donc, 5% des sondages se tromperont de plus de 3% à la hausse ou à la baisse. Pourtant, lorsque des écarts entre des sondages successifs apparaissent lors d'une campagne électorale, on parle de dynamique de campagne pour expliquer ces évolutions, même pour des écarts de 1% ou de 0.5%! Or tant qu'on reste dans des fluctuations de moins de 2% (et même 3%) pour des sondages portant sur 1000 individus (en dehors bien sûr de période de scandales ou d'actualité marquante) commenter de la sorte des petites fluctuations dans les sondages relève de la malhonnêteté intellectuelle, car l'aléa propre à cette technique suffit amplement à expliquer de telles fluctuations... et les sondeurs ne peuvent l'ignorer. Parfois les sondeurs suivent les mêmes sondés sur plusieurs semaines pour mesurer "la vraie évolution". Mais le fait d'être "surveillé", forcément change le comportement des sondés, et ça ne donne pas forcément des renseignements beaucoup plus précis.*

Table des matières